Improved Integration of Secondary Data using Self-Healing Sequential Gaussian Simulation

Stefan Zanon[1], *Ty Faechner*[1], *Clayton V. Deutsch*[1]

ABSTRACT

Secondary data are important in geostatistical simulation of continuous variables. Geophysical data and geological trends are used for modeling porosity and permeability. Multiple mineral grades must often be modeled for mining applications. Sequential Gaussian simulation (SGS) is often used because of its relative simplicity and robustness. The two most common approaches to integrate secondary data in Gaussian simulation are with (1) locally varying mean, or (2) collocated cokriging. A significant problem with both of these techniques is *variance inflation*, that is, the variance of the resulting simulated values is too high because of an inappropriate decision of stationarity or an artifact of choosing a single secondary data in presence of many. We introduce a *self-healing* procedure for dynamic correction of this problem. The dynamic correction is different for the locally varying mean approach and for collocated cokriging since there are different reasons why each of these methods causes variance inflation. The reasons for variance inflation are discussed and the self-healing is applied to a number of data sets.

INTRODUCTION

Sequential Gaussian simulation (SGS) is remarkably robust and flexible for the generation of geostatistical realizations (Isaaks, 1990). SGS is arguably the most powerful and commonly used geostatistical simulation technique at the present time; the sgsim program in GSLIB is a common implementation (Deutsch and Journel, 1998). The key features of SGS are that input data are honored at their locations and the global histogram and variogram are reproduced within ergodic fluctuations.

The histogram and variogram 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 simulated values were exactly normal with a mean of 0, variance of 1, and correct shape. Statistical or ergodic fluctuations between realizations are expected. In practice, these fluctuations appear reasonable when the input data are consistent with the stationary histogram and variogram and there are no secondary data.

Unreasonable results have been observed in presence of secondary data. A full block cokriging approach has the theoretical foundation to work correctly, but it is avoided because of significant inference and CPU time. The locally varying mean (LVM) approach is widely used to account for geological trends, but the LVM violates the implicit assumption of stationarity in SGS. The collocated cokriging (CLK) is also widely used because of its simplicity, but the CLK approach does not consider highly correlated secondary data near the location being estimated. The consequence of these theoretical violations is often an inflated variance in normal space.

The dispersion variance $D^2(v,A)$, where v is the modeling scale and A is the domain being simulated, is the expected variance of the simulated values. This variance is very close to 1.0 in normal space. The variance of the simulated values when using an LVM or CLK can systematically be in the range of 1.3 to 2.0, which translates to a significant bias and error on back transformation to the original Z data units.

[1] University of Alberta, Edmonton, Alberta, CANADA

Some empirical corrections have been used. The variance of the secondary data can be reduced, that is, the normal score transforms of the secondary values are multiplied by a factor less than one. Alternatively, the kriging variance can be reduced by a factor less than one during the sequential simulation procedure. Both of these correction factors require some iteration to set correctly. The correction factors depend on the variograms of the primary and secondary data, on the correlation between the variables, and on many other implementation decisions such as search neighborhood. It is intractable to iteratively find the required correction factor for each combination of input parameters. An automatic procedure to correct the simulated values would remove the need for these iterations.

We propose a method to dynamically correct the simulated realizations as the simulation proceeds. The method is called *self-healing* because the correction is only administered when the results start going wrong and the size of the correction depends on the magnitude of the problem. Self-healing will be described in detail with examples.

INTEGRATION OF SECONDARY DATA

Sequential simulation is described in many sources (Isaaks, 1990; Deutsch and Journel, 1998) and will not be repeated here. Of particular concern to us is the presence of secondary data. Secondary data come from a variety of sources:

- Geologic trend mapping in the horizontal plane or vertical direction provides critical information on large-scale variations in the variable. There are good reasons to expect the average value to depend on location; virtually all of our study areas are a type of geologic anomaly.
- Geophysical data can often provide large-scale information related to the variable we are modeling. The geophysical data and the variable under consideration are often correlated with a correlation coefficient in the range of 0.5 to 0.8.
- Dynamic flow or historical production data can also provide valuable secondary data related to the variable we are modeling.
- Multiple variables are often constructed sequentially, so the first variable modeled provides a secondary variable for the next, and so on.

Geostatistical models must consider such secondary information and deterministic trends. The two commonly used procedures to integrate such secondary data are the locally varying mean (LVM) or collocated cokriging (CLK). The external drift formalism is infrequently used in simulation because of the lack of explicit control over the magnitude of the correlation. Full cokriging formalism is also infrequently used because of the additional inference and computational effort required.

Locally Varying Mean (LVM)

The LVM kriging estimate, system of equations, and kriging variance may be written:

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

The approach simply amounts to estimating the residual from the local mean at **u** using residuals of the data from their local mean values $\mathbf{u}_i$, $i=1,\ldots,n$. The stationary covariance $C(\mathbf{h}=\mathbf{u}_1-\mathbf{u}_2)$ is often inferred from variogram calculation using the original data ($C(\mathbf{h})=1-\gamma_y(\mathbf{h})$). Of course, we note immediately that the *right* covariance is the residual covariance, but that is problematic since there are no true residual data and, inevitably, the mean values $m(\mathbf{u})$ are correlated with the data values $y(\mathbf{u})$.

The locally varying mean values $\{m(\mathbf{u}), \mathbf{u} \in A\}$ come from the secondary data. The original local mean data are in the units of the original Z data variable. These values, however, must be transformed to standard normal space for use as an LVM. In practice, the normal score transform from the Z variable is used, e.g.,

$$m(\mathbf{u})=G^{-1}(F_Z(m_z(\mathbf{u}))) \qquad (2)$$

where $m_z(\mathbf{u})$ is the local mean coming directly from geological mapping or some other source, $F_z()$ is the stationary CDF of the Z variable, and $G()$ is the standard normal CDF. As the variation in the local mean values increases there is more potential for variance inflation in the final simulated values. One correction procedure is to scale the local mean values by a correction factor, f, that is less than one and determined iteratively: $m^*(\mathbf{u}) = f \bullet m(\mathbf{u})$. There is no guarantee that the local mean values will be reproduced in the final simulated realizations. We could, of course, check reproduction of the local mean values by generating multiple realizations and calculating the resultant local mean values from the simulated realizations.

Collocated Cokriging (CLK)

The collocated cokriging (CLK) estimate, system of equations, and kriging variance may be written (Xu et.al., 1992):

$$y^*_{CLK}(\mathbf{u})=\sum_{i=1}^n \lambda_i \cdot y(\mathbf{u}_i)+\mu \cdot y_2(\mathbf{u})$$

$$\begin{cases} \sum_{j=1}^n \lambda_j \cdot C(\mathbf{u}_j-\mathbf{u}_i)+\rho \cdot \mu \cdot C(\mathbf{u}-\mathbf{u}_i)=C(\mathbf{u}-\mathbf{u}_i), & i=1,\dots,n \\ \sum_{j=1}^n \lambda_j \cdot \rho \cdot C(\mathbf{u}_j-\mathbf{u}_i)+\mu=\rho \end{cases} \qquad (3)$$

$$\sigma_K^2(\mathbf{u})=\sigma^2-\sum_{i=1}^n \lambda_i \cdot C(\mathbf{u}-\mathbf{u}_i)-\mu \cdot \rho$$

The collocated secondary data at the location being estimated, $y_2(\mathbf{u})$, is needed at each location being estimated and the correlation coefficient, ρ, between collocated y/y_2 data is also required. An equivalent Bayesian updating formalism could be written.

The secondary data do not come in standard Gaussian units, but the variable is independently transformed to Gaussian distribution using normal scores transformation.

The practical problem that arises with CLK is an overstatement of the kriging variance, $\sigma^2_k(\mathbf{u})$, because only one of many nearby secondary data are used. The secondary data are usually smooth so including more secondary data would not significantly change the estimate, but the kriging variance would go down. The amount it would go down depends on the spacing of the data, the variogram of the primary variable, the variogram of the secondary variable, and the correlation between them, and other implementation details of the search. Efforts to determine how much the variance is overstated have largely failed due to the interrelationships between these variables. A full cokriging formalism (Verly, 1993) would solve this problem; however a full model of coregionalization would be required.

One correction procedure is to scale the local kriging variance values by a correction factor, f, that is less than one and determined iteratively: $\sigma^{2*}_k(\mathbf{u}) = f \bullet \sigma^2_k(\mathbf{u})$. This has been reasonably successful in practice since the underlying problem is a consistent overstatement of the kriging variance. The factor must be determined iteratively and it is very sensitive to changes in the variogram structure and correlation coefficient. The dependence is highly non linear and even non monotonic. For example, the f factor should be 1 for ρ=0 or $|\rho|$=1, but something less than one in between. The sgsim program allows specification of f and the user must determine what it should be by repeatedly running the program.

METHODOLOGY FOR SELF HEALING

The general idea of *self-healing* is to dynamically fix the simulated values as simulation proceeds. This is convenient since most people find it annoying to iteratively tweak a parameter until the global histogram is reproduced. Leaving the variance reduction parameter at the default amounts to accept realizations that do not match the global histogram. The mechanism of variance inflation is different for LVM and CLK; therefore, a different dynamic correction is applied in both cases. The general idea is the same though:

- The mean and variance of all previously simulated values (including original data) is kept up to date during the simulation, n, m, s^2.
- A correction is considered when s^2 exceeds the theoretical variance $D^2(v,A)$, that is, when the ratio $s^2/D^2(v,A)$ exceeds one. The amount of the correction depends on how large $(s^2/D^2(v,A)-1)$ becomes.
- The mean, m, and variance, s^2, are not reliably informed until at some number of nodes (say n>200) are simulated. The first $n^{'}$ nodes are simulated with no modification; then, healing is considered. The first $n^{'}$ nodes are revisited after the entire grid has been populated to avoid artifacts due to excess variance at the first grid node locations.

The correction mechanism for both LVM and CLK is to reduce the kriging variance by a multiplicative constant:

$$\sigma_k^{2*}(\mathbf{u}) = f \cdot \sigma_k^2(\mathbf{u}) \tag{4}$$

The correction factor, f, is 1.0 when no correction (or healing) is required. Correcting the kriging variance and not the kriging estimate ensures that local conditioning data are reproduced exactly. Of course, changing the kriging variance does change the final variance, which is our goal, and it changes the final variogram reproduction by modification of the variance. Note that the covariance reproduction is not changed, that is, the covariance between the simulated values and the data values is correct; however, the variance *and* variogram are changed. This is the price of fixing the histogram. We discuss this more later.

As mentioned above, the mechanism of variance inflation in LVM and CLK is different; therefore, a different prescription is used for each method.

Correction for Locally Varying Mean (LVM)

The mechanism for variance inflation in LVM is a too high probability for extreme values in regions where the local mean is high or low and the estimation variance is high. Figure 1 shows an illustration of a locally varying mean (solid line) versus location, **u**, and three conditional distributions. The shaded regions of the two outside distributions cause variance inflation because there is a too high probability to draw large and small values. The simulation of a few high and low values cause even more problems as the sequential simulation proceeds; there is a need for dynamic correction of such extreme values.

The kriging estimate, $y^*_{LVM}(\mathbf{u})$ cannot be far from zero when the kriging variance, $\sigma^2_k(\mathbf{u})$, is large because the kriging weights in such a case must all be small (refer back to equation (1)). The presence of data that would make the kriging estimate different from zero also makes the kriging variance small. Variance inflation in LVM is caused by locations where the kriging estimate is large or small and the kriging variance is large.

Our prescription for LVM is to link the f factor to the bivariate relation between the kriging estimate, $y^*_{LVM}(\mathbf{u})$ and the kriging variance, $\sigma^2_k(\mathbf{u})$. Figure 2 illustrates correction factor on a cross plot of the kriging variance (vertical axis) and absolute value of the kriging estimate (horizontal axis). When the points fall in an allowable region (hatched region) the f factor is set to 1. The f factor is set to zero if the points fall outside (the spotted region). There is an intermediate region where the f factor is set between 0 and 1.

The $A^{'}$ distance parameter (in units of the kriging estimate) on Figure 2 is somewhat subjective; however, we know that it should be large when there are no problems and shorter when the variance starts inflating. Although it seems to be getting complicated, we link the $A^{'}$ parameter to the variance inflation $s^2/D^2(v,A)$, see Figure 3. Some numerical experimentation was done to arrive at universally robust settings of these parameters. In practice, the user does not need to concern themselves with these details; they just "check the self-healing box."

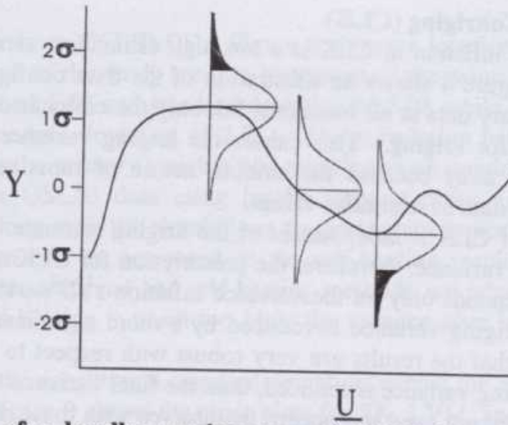

Figure 1 Illustration of a locally varying mean (solid line) versus location, u, and three conditional distributions. The shaded regions of the two outside distributions cause variance inflation because there is a too high probability to draw large and small values.

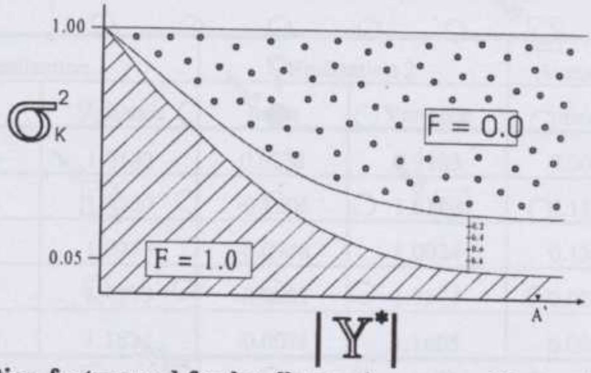

Figure 2 The correction factor used for locally varying mean. The region of F=1.0 is where no correction is applied to the kriging variance. The region of F=0.0 is where the kriging variance is reduced to zero. There is a region where a factor between 1 and 0 is used.

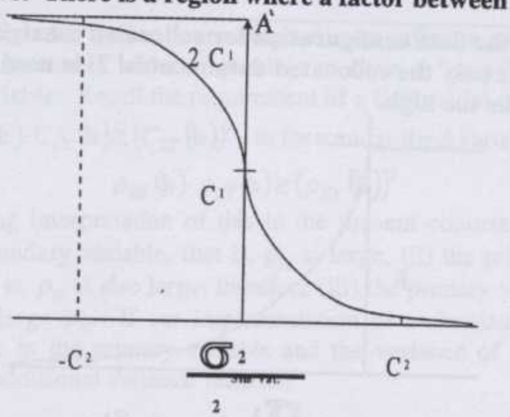

Figure 3 The correction factor for locally varying mean. The horizontal axis is the ratio of the actual variance to the theoretically expected variance. The vertical axis is the "range" parameter for the correction scheme (see Figure 2). As the variance becomes too large, this range becomes short. The range is longer when the variance is too small.

Correction for Collocated Cokriging (CLK)

The mechanism for variance inflation in CLK is a too high estimation variance because too few secondary data are used. Figure 4 shows an illustration of the data configuration for collocated cokriging. There are secondary data at all locations, but only the collocated data (central location marked with a "?") is used for kriging. This causes the kriging variance to be too high. The kriging estimate is typically okay because the smooth nature of most secondary data sources means that nearby secondary data have similar values.

The variance inflation of CLK is independent of the kriging estimate. There is a systematic overstatement of the kriging variance; therefore, the prescription for CLK is to reduce the kriging variance by a factor *f* that depends only on the variance inflation $s^2/D^2(v,A)$, see Figure 5. As the variance gets inflated, the kriging variance is reduced by a more significant *f* factor. Numerical experimentation has shown that the results are very robust with respect to the details of how *f* is reduced. As long as the kriging variance is reduced, then the final variance will be reproduced.

Once again, the user does not need to concern themselves with these details; they just "check the self-healing box."

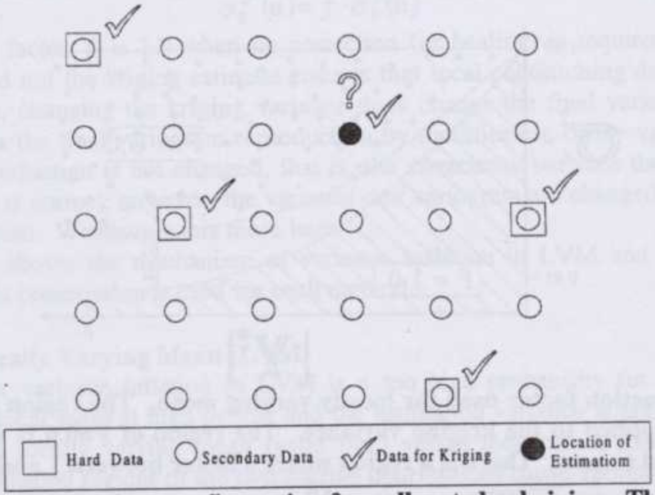

Figure 4 Illustration of the data configuration for collocated cokriging. There are secondary data at all locations, but only the collocated data (central ?) is used for kriging. This causes the kriging variance to be too high.

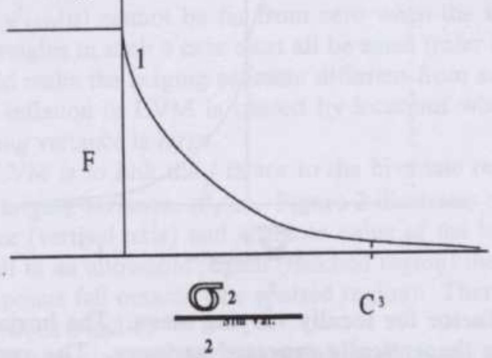

Figure 5 The correction factor for collocated cokriging. The horizontal axis is the ratio of the actual variance to the theoretically expected variance. The vertical axis is the correction factor. No change is made when the variance is too small, but an increasingly severe correction is made as the variance becomes too large.

EXAMPLE

The data set was taken from GSLIB CD. Figure 6 shows a location map (note the area of high values in the NE portion of the study area), a histogram of the primary hard data variable for the 29 data (note the mean of 3.38 and standard deviation of 5.08 while the secondary variable has a mean of 2.32 and a standard deviation of 2.71). The correlation between these two variables is 0.774. A variogram for the primary variable was modeled as an omnidirectional variogram.

Simulations of the GSLIB data using locally varying mean and collocated cokriging are shown in Figure 7. Along with the simulations the variogram reproduction and probability plots are also shown. These can be compared to the self-healing results shown in Figure 8. The differences between the traditional and self-healing methods are tabulated in Table 1. Note that the variance before self-healing is much too high; the varance after healing is much closer to the correct value of 1.0.

The cross plot of the conditional standard deviation versus the conditional mean values also reveals the problem. Figure 9 shows the cross plots for SK, LVM, and CLK. Note that LVM and CLK leads to values that are more spread out than SK. The values on the *X*-axis are the conditioning data points (standard deviation of 0.0). The nugget effect causes the standard deviation to approximately greater than 0.4.

Technique	Realization 1		Realization 2		Average of 5 Realizations	
	mean	Variance	mean	Variance	mean	variance
SK	0.5670	1.0182	0.0238	0.9283	0.0053	0.9489
LVM	-0.2044	1.8260	-0.1006	1.8188	0.1311	1.8402
LVM-SH	-0.2321	1.0275	-0.0546	1.0024	0.1368	1.0284
CCK	0.0002	1.4273	-0.0083	1.3574	0.0220	1.3764
CCK-SH	-0.0197	1.1815	-0.0075	1.1605	0.0253	1.1111

Table 1 Summary of results for modeling the GSLIB data set.

DISCUSSION

The self-healing appears to work well at controlling variance inflation at the expense of variogram reproduction. This was inevitable. It is impossible to have a "noisy" primary variable highly correlated to a smooth variable. Recall the requirement of a licit model of coregionalization:

$$C_{ZZ}(\mathbf{h})\cdot C_{YY}(\mathbf{h})\geq (C_{ZY}(\mathbf{h}))^2, \text{ or for standardized variables:} \quad (5)$$

$$\rho_{ZZ}(\mathbf{h})\cdot \rho_{YY}(\mathbf{h})\geq (\rho_{ZY}(\mathbf{h}))^2$$

There is an interesting interpretation of this in the present context. Consider the following scenario (i) a smooth secondary variable, that is, ρ_{yy} is large, (ii) the primary and secondary data are highly correlated, that is, ρ_{zy} is also large, therefore (iii) the primary variable must have smooth spatial structure, that is, large ρ_{zz}. If our implementation of a simulation algorithm is trying to enforce greater variability in the primary variable and the variance of the secondary variable is fixed, then there must be additional variance inflation.

CONCLUSIONS

No objective criterion has been developed to judge when the excess variance is truly unacceptable. It is unlikely such an objective criterion could ever be determined. Nevertheless, additional study is warranted to provide some check that the algorithm is giving acceptable results.

We encounter a paradox as more data becomes available. The additional data provide the means to identify a more reliable trend model, but they also make the trend model less important. That is, the conditioning of geostatistical models to many data enforces both the deterministic and stochastic variations of the variable of interest. Clearly, the usage of trend models is important in presence of sparse data. Additional work is warranted to provide guidelines on when trend models should be considered.

ACKNOWLEDGMENTS

We are grateful to the industry sponsors of the Centre for Computational Geostatistics at the University of Alberta for supporting this research.

REFERENCES

C. V. Deutsch, A.G. Journel. 1998. *GSLIB: Geostatistical Software Library: and User's Guide*. Oxford University Press, New York, 2nd Ed.

P. Goovaerts. 1997. *Geostatistics for natural resources evaluation*, Oxford University Press, New York.

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

E. H. Isaaks and R. M. Srivastava. 1989. *An introduction to applied geostatistics*, Oxford University Press, New York.

W. Xu, T. T. Tran, R. M. Srivastava and A. G. Journel. 1992. Integrating Seismic Data in Reservoir Modeling: the Collocated Cokriging Alternative, 1992 SPE Annual Technical Conference and Exhibition, Washington, DC, SPE Paper Number 24742, pp 833-842.

G. Verly, 1993. Sequential Gaussian Cosimulation: A Simulation Method Integrating Several Types of Information, Geostatistics Troia '92, vol 1., ed. A. Soares, Kluwer Academic Publishers, pp. 543-554.

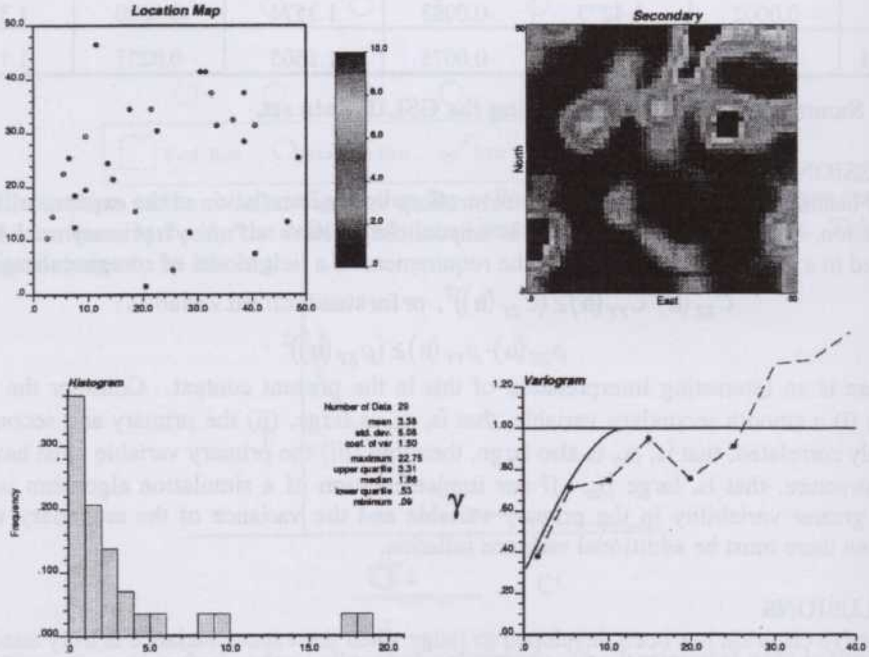

Figure 6 GSLIB data: upper left – location map of the 29 hard data, upper right – color scale map of the secondary data, lower left – histogram of the data, and lower right – normal score variogram from the hard data.

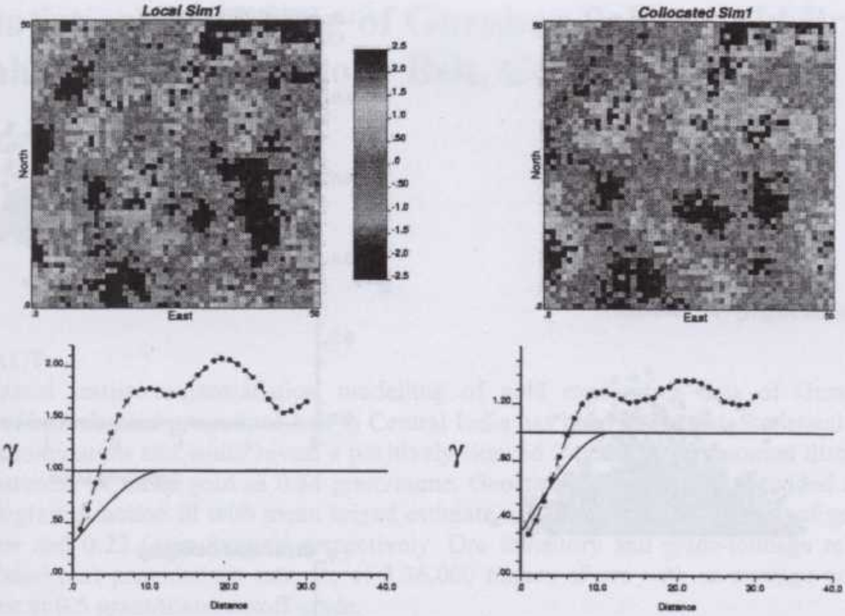

Figure 7 GSLIB data: left side – first realization with locally varying mean and variogram of simulated values; right side – first realization of collocated cokriging and variogram of simulated values. All results are shown in normal space.

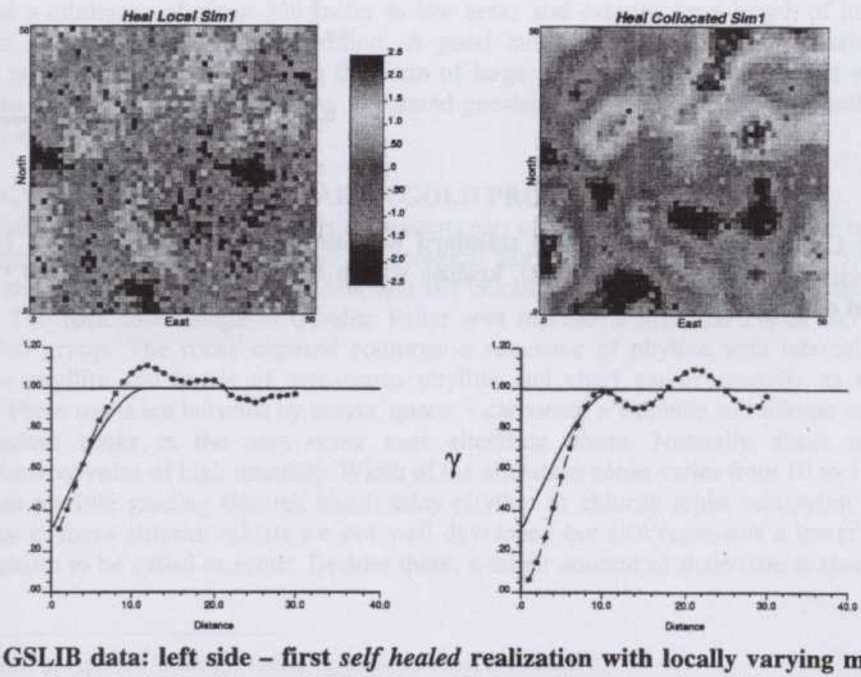

Figure 8 GSLIB data: left side – first *self healed* realization with locally varying mean and variogram of simulated values; right side – first realization of *self healed* collocated cokriging and variogram of simulated values. All results are shown in normal space.

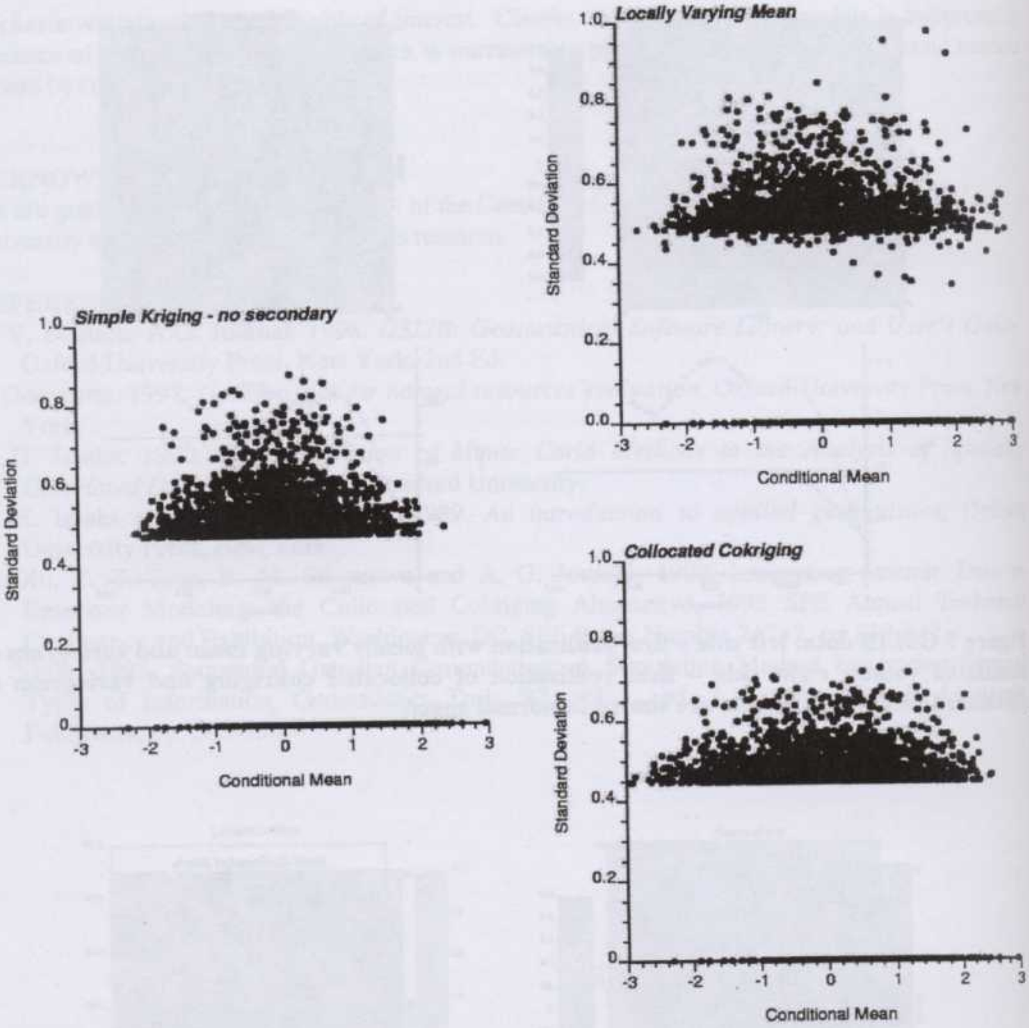

Figure 9 Cross plots of conditional standard deviation and conditional mean for simple kriging with no secondary data (left), kriging with a locally varying mean (top right) and collocated cokriging (bottom right).