

University of Alberta

Library Release Form

Name of Author: Stefan D. J. Zanon

Title of Thesis: Advanced Aspects of Sequential Gaussian Simulation

Degree: Master of Science

Year this Degree Granted: 2004

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Stefan D. J. Zanon
RR 4 1610 2nd Avenue
Invermere, B.C.
Canada, V0A 1K4

Date: _____

University of Alberta

ADVANCED ASPECTS OF SEQUENTIAL GAUSSIAN SIMULATION

by

Stefan D. J. Zanon

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of **Master of Science**.

in

Mining Engineering

Department of Civil and Environmental Engineering

Edmonton, Alberta
Spring 2004

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Advanced Aspects of Sequential Gaussian Simulation** submitted by Stefan D. J. Zanon in partial fulfillment of the requirements for the degree of **Master of Science** in *Mining Engineering*.

Dr. Clayton V. Deutsch (Supervisor)

Dr. Ben Rostron (External)

Dr. Samuel Frimpong

Dr. Tim Joseph

Date: _____

Abstract

Sequential Gaussian simulation (SGS) is a simulation technique used throughout the natural resources industry to construct multiple equiprobable numerical models. The SGS methodology is straightforward, but there are complex interdependent implementation decisions in the transformation from theory to practice.

Some of the implementation issues include: simulation path, search strategy, number of conditioning data, and secondary data affects. The simulation path is random to avoid artifacts. At every unsampled location along this path, the local data is identified by a search and their covariances calculated. This data is then used as input to condition the simulation process. To produce the best possible estimate, the conditioning data should be representative of the surrounding region up to a maximum number of data to avoid excessive CPU time. Secondary data can be considered in this process, but an understanding of how this data will affect the results and problems associated with different techniques will allow the practitioner to make a balanced decision.

Acknowledgements

I would like to extend my thanks and appreciation to the members of the Centre for Computational Geostatistics. I would especially like to thank the director, and my advisor, Dr. Clayton V. Deutsch. It was through their help and support that I was able to complete this thesis.

I am dedicating this thesis to my family for their motivation and support of my academic career. To my parents, Bob and Shannon, I thank you for providing the encouragement that kept me focused on my goal. To my sister Kathleen (Kate), I thank you for keeping life interesting and being the special person you are.

Finally, I would like to thank all others who have touched my life and helped form the person I am today.

Contents

1	Background / Theory	1
1.1	Linear Estimation	2
1.2	Estimation Variance	3
1.3	Minimized Estimation Variance - Kriging	4
1.4	Covariance of Kriged Estimates	5
1.5	Missing Variance	6
1.6	Addition of a Random Component	6
1.7	Gaussian Simulation	7
2	Sequential Gaussian Simulation Implementation Details	9
2.1	Data Transformation	9
2.2	Assign Data to Grid Nodes	11
2.3	Local Data	11
2.3.1	Maximum Number of Data	13
2.3.2	Searching for Data	18
2.4	Simulation Path	23
2.4.1	Regular Path	24
2.4.2	Spiral Path	26
2.4.3	Random Path	26
2.4.4	Multiple Grids	26
3	Non Stationary Parameters	29
3.1	Mean	29
3.2	Variance	31
3.3	Correlation	34
3.4	Angles	34
3.5	Rock Types and Soft Boundaries	37
4	Self-healing	39
4.1	Collocated Cokriging	39
4.2	Locally Varying Mean	40
4.3	Fixing the Problem	41
4.4	Example	45

5	Alternative Secondary Data Integration	56
5.1	Full Cokriging	56
5.2	Stepwise Conditional Transformation	58
6	Final Comments	60
6.1	Conclusions	60
6.2	Future Work	62
	References	64

List of Figures

2.1	Data transformation from original units to normal space.	10
2.2	Assigning data to the grid system.	12
2.3	Change in CPU time as the number of conditioning data increases. . .	14
2.4	Example using Gandin's method of predicting the bounds for variance reduction with the addition of a single data point.	15
2.5	Location of the input data and test points in the example.	17
2.6	Change in the kriging estimate and variance as the number of condi- tioning data increases.	19
2.7	Spiral search strategy.	20
2.8	Super block search gridding and template.	21
2.9	Super block search indexing.	22
2.10	CPU cost of a two part search.	23
2.11	The kriging weight template used in a regular simulation path.	24
2.12	Variogram reproduction between the random and regular path.	25
2.13	Spiral simulation path.	27
2.14	Multiple grid layout.	28
3.1	Locally varying mean example.	31
3.2	Histograms for locally varying mean example.	32
3.3	Calculation of the variance scaling factor.	33
3.4	Rotation angles and ratios describing anisotropy.	35
3.5	Geological setting for locally varying angles.	36
4.1	Illustration of variance inflation for locally varying mean.	41
4.2	The self-healing correction factor used with collocated cokriging. . . .	43
4.3	The link between the kriged mean and variance for LVM.	44
4.4	Plots of the CCK self-healing example.	47
4.5	Histograms of the CCK self-healing example.	48
4.6	Scatter plots of the CCK self-healing example.	49
4.7	Variograms of the CCK self-healing example.	50
4.8	Plots of the LVM self-healing example.	52
4.9	Histograms of the LVM self-healing example.	53
4.10	Scatter plots of the LVM self-healing example.	54
4.11	Variograms of the LVM self-healing example.	55

List of Tables

3.1 Logic matrix used to describe rock type boundaries. 38

Chapter 1

Background / Theory

The natural resources industry is reliant on modelling techniques to estimate rock properties at unsampled locations. Linear estimators are commonly used to predict these properties by weighting local data. Kriging is a common estimator that finds the set of weights that minimizes the error variance (Deutsch 2002a). Sequential Gaussian simulation (SGS) is built on the foundation of kriging. SGS incorporates many of the benefits of kriging, but it also accounts for heterogeneity and uncertainty.

Sequential Gaussian simulation is a robust algorithm, but implementation requires an understanding of how the algorithm operates. The SGS theory has been documented in different sources (Isaaks 1990, Deutsch 1998, 2002a), but will be summarized below to set the platform for the rest of this thesis. This chapter starts with a look at the development of the simple kriging equations and then extends this to simulation. Some of the mathematical properties of these techniques are discussed.

The details behind transforming SGS from theory to application is the focus of Chapter 2. To start, an explanation of how non-Gaussian input data can undergo normal score transformation resulting in a Gaussian distribution. SGS requires a number of implementation decisions regarding the assignment of data to the grid nodes, simulation path, data search method, and maximum number of conditioning data. The different options available for each decision are presented and some guidance is provided. In regards to the maximum number of conditioning data, some earlier research is explained. This chapter presents some of the different options available when applying SGS and was designed to help the practitioner make more informed decisions.

In some cases, the assumption of stationarity used in SGS can be violated by the data. For these non stationary variables, Chapter 3 presents some alternative formulations of SGS to try and incorporate extra information to account for the changes in these variables. The most common non stationarity is a location-dependant mean. The theory of how to incorporate a location-dependant mean into SGS is presented

along with an example. Other non stationary variables include the variance, correlation, and angles of continuity. Limited research has been done on these topics, but the groundwork is presented as a starting point and to guide further research. The final violation to stationarity is the presence of different rock types inside of the modelling area. The problems associated with incorporating rock types into a model are presented along with some ideas on how correlations between the rock types can be utilized.

In Chapter 4, an algorithm is presented on how to deal with global variance inflation problems associated with the collocated cokriging (Xu 1992) and locally varying mean forms of SGS. The cause for variance inflation is explained specific to each technique. To correct this problem, an algorithm termed *self-healing* (Zanon 2002) was developed by the author. Self-healing dynamically alters the kriging variances used in SGS to correct the global variance. The application of self-healing, specific to each technique, is outlined along with the limitations. An example for each method is provided.

For some of the methods presented there are alternative techniques that can be used. Full cokriging and Stepwise conditional transformation (Leuangthong 2003a, 2003b) are two alternatives that are discussed in Chapter 5. Full cokriging is another way to incorporate secondary information into the simulation process. The theory behind full cokriging is reviewed along with the advantages and difficulties associated with its application. Stepwise conditional transformation is another alternative technique that has recently gained more notice. This technique is a form of normal score transformation applied to multiple variables where the resulting distributions are statistically independent. This allows all the variables to be simulated using basic SGS techniques and the correlation between the variables is preserved in the transformation process.

Chapter 6 is the final chapter in this thesis. The conclusions from all the previous chapters are presented along with any other final comments. Many of the techniques discussed in this thesis require more research. These areas of future research are presented along with the some ideas on the direction this research should take.

The programs used throughout this thesis were from the publicly available geo-statistical software library - GSLIB (Deutsch 1998). The primary program that incorporates the SGS algorithm is SGSIM. In many cases, SGSIM was altered to explore specific implementations.

1.1 Linear Estimation

We start with a random function (RF) model $Z(\mathbf{u}), \forall \mathbf{u} \in A$, that is, a mathematical representation of the petrophysical properties, z-values, at all locations of interest, \mathbf{u} , within the study area A . The random function is assumed to be stationary (Journel 1978). Under this assumption, the expected value at all locations is equal to the global mean, $E\{Z(\mathbf{u})\} = m, \forall \mathbf{u} \in A$, and the variability at all locations is

the global variance, $\sigma^2(\mathbf{u}) = \sigma^2, \forall \mathbf{u} \in A$. We subtract the mean at all locations to get a residual value, $y(\mathbf{u}) = z(\mathbf{u}) - m$. This process can be reversed by adding the mean back to the residuals, returning the data to its original units.

$$z(\mathbf{u}) = y(\mathbf{u}) + m \quad (1.1)$$

where the expected value of the residuals is zero, $E\{Y(\mathbf{u})\} = 0$. Note that it is conventional to use upper case notation for a random variable and lower case notation for actual numbers.

A linear estimate, y^* , is commonly used to estimate the value at an unsampled location:

$$y^*(\mathbf{u}) = \sum_{\alpha=1}^n \lambda_{\alpha} \cdot y(\mathbf{u}_{\alpha}), \quad (1.2)$$

This estimate is a weighted linear combination of n nearby data variables at locations $\mathbf{u}_{\alpha}, \alpha = 1, \dots, n$. The weights, $\lambda_{\alpha}, \alpha = 1, \dots, n$, can take any values, but common practice consists of defining an estimation variance and determining the weights that minimize this variance.

1.2 Estimation Variance

The estimate at an unsampled location is unlikely to be the same as the true value at that location. The error in the estimate can be calculated in expected value. The estimation variance is written:

$$\sigma_e^2(\mathbf{u}) = E\{[Y^*(\mathbf{u}) - Y(\mathbf{u})]^2\} \quad (1.3)$$

Note that we are considering both the estimate and the truth to be random variables. The right hand side can be expanded into the following form:

$$\sigma_e^2(\mathbf{u}) = E\{[Y^*(\mathbf{u})]^2\} - 2 \cdot E\{Y^*(\mathbf{u}) \cdot Y(\mathbf{u})\} + E\{[Y(\mathbf{u})]^2\}$$

The estimator (Equation 1.2) is expressed in expected value terms and substituted into this equation to get:

$$\sigma_e^2(\mathbf{u}) = \sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_{\alpha} \lambda_{\beta} E\{Y(\mathbf{u}_{\alpha}) \cdot Y(\mathbf{u}_{\beta})\} - 2 \cdot \sum_{\alpha=1}^n \lambda_{\alpha} E\{Y(\mathbf{u}) \cdot Y(\mathbf{u}_{\alpha})\} + E\{[Y(\mathbf{u})]^2\}$$

To simplify further, the covariance must be introduced. The covariance of a single variable is a measure of the linear relationship between two values a specified distance apart:

$$C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta}) = C(\mathbf{h}_{\alpha, \beta}) = E\{(Y(\mathbf{u}_{\alpha}) - m) \cdot (Y(\mathbf{u}_{\beta}) - m)\}$$

Since we are working with residuals, the mean is zero and the covariance simplifies to:

$$C(\mathbf{h}_{\alpha, \beta}) = E\{Y(\mathbf{u}_{\alpha}) \cdot Y(\mathbf{u}_{\beta})\} \quad (1.4)$$

The covariance can now be used to replace the expected value terms.

$$\sigma_e^2(\mathbf{u}) = \underbrace{\sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_{\alpha} \lambda_{\beta} C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta})}_{\text{part 1}} - 2 \cdot \underbrace{\sum_{\alpha=1}^n \lambda_{\alpha} C(\mathbf{u}, \mathbf{u}_{\alpha})}_{\text{part 2}} + \underbrace{C(0)}_{\text{part 3}} \quad (1.5)$$

where the covariance at $\mathbf{h} = 0$ is equal to the global variance, $C(0) = \sigma^2$.

This estimation variance consists of three parts, each accounting for a different aspect of the error in the estimate. The first part of the equation accounts for the redundancy between the data, $C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta})$. The greater the redundancy in the data, the higher the covariance between the data and the higher the error variance. The next part of the equation accounts for the closeness of the input data to the point of estimation, $C(\mathbf{u}, \mathbf{u}_{\alpha})$. A negative sign is attached to this part of the equation, causing a reduction in the error variance. The closer the data are to the location being estimated, the larger the covariance and the greater the reduction. The third part of the equation is the global variance. When no local data are available, the estimate will be the local mean value and the error variance will be the global variance (Deutsch 2002a).

1.3 Minimized Estimation Variance - Kriging

(Equation 1.5) allows the variance to be calculated for any set of weights, $\lambda_{\alpha}, \alpha = 1, \dots, n$. It is desirable to find the set of weights that minimize this error variance. The optimal weights can be found by taking the partial derivatives of (Equation 1.5) with respect to each of the weights $\lambda_{\alpha}, \alpha = 1, \dots, n$:

$$\frac{\partial[\sigma_e^2(\mathbf{u})]}{\partial \lambda_{\alpha}} = 2 \cdot \sum_{\beta=1}^n \lambda_{\beta} \cdot C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta}) - 2 \cdot C(\mathbf{u}, \mathbf{u}_{\alpha}), \quad \alpha = 1, \dots, n \quad (1.6)$$

We are looking for the minimum when this derivative is set to zero. (Equation 1.6) is sure to be a minimum because of the positivity property of the covariance function and the fact that the mean square error has no upper limit (Chilès 1999). This will be a unique solution.

Setting the n-derivatives equal to zero and moving $C(\mathbf{u}, \mathbf{u}_{\alpha})$ to the left hand side leads to the simple kriging (SK) equations.

$$\sum_{\beta=1}^n \lambda_{\beta} \cdot C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta}) = C(\mathbf{u}, \mathbf{u}_{\alpha}), \quad \alpha = 1, \dots, n \quad (1.7)$$

Solving this system of equations will yield the set of weights that minimize the error variance. The kriging equations can be substituted back into (Equation 1.5) to simplify the estimation variance specific to SK:

$$\sigma_{sk}^2(\mathbf{u}) = C(0) - \sum_{\alpha=1}^n \lambda_{\alpha} C(\mathbf{u}, \mathbf{u}_{\alpha}) \quad (1.8)$$

This is a classical minimization approach. Note that the simple kriging weights only depend on the data configuration and the covariance model. If the location being estimated falls on a data location, the collocated data will receive a weight of one and all other data will have weights of zero; the location being estimated will exactly reproduce the collocated data.

The covariance used in kriging is linked to another function called the variogram. The variogram is a measure of the variability between two values that are separated by a distance \mathbf{h} . The variogram is defined as:

$$2\gamma(\mathbf{h}) = E\{[Y(\mathbf{u}) - Y(\mathbf{u} + \mathbf{h})]^2\}$$

In practice, the variogram is calculated for a limited number of separation distances and directions, \mathbf{h} . The experimental points are then fitted with a model that allows the variogram to be calculated for any direction and distance.

The covariance is linked to the variogram through the following expression resulting from the assumption of stationarity:

$$C(\mathbf{h}) = \sigma^2 - \gamma(\mathbf{h})$$

The covariance is calculated from the variogram model.

1.4 Covariance of Kriged Estimates

A kriging estimate can be calculated for any location and any configuration of input data; however, the kriged values will be smoother, have less variance, than the original data. Nevertheless, kriging has the very interesting property that the covariance between the estimates and the data are correct. This can be demonstrated by starting with the definition for the covariance (Equation 1.4) and then substituting in the expression for the $Y^*(\mathbf{u})$ estimator (Equation 1.2):

$$\begin{aligned} Cov\{Y^*(\mathbf{u}), Y(\mathbf{u}_{\alpha})\} &= E\{Y^*(\mathbf{u}) \cdot Y(\mathbf{u}_{\alpha})\} \\ &= E\left\{\sum_{\beta=1}^n \lambda_{\beta} \cdot Y(\mathbf{u}_{\beta}) \cdot Y(\mathbf{u}_{\alpha})\right\} \\ &= \sum_{\beta=1}^n \lambda_{\beta} \cdot E\{Y(\mathbf{u}_{\alpha}) \cdot Y(\mathbf{u}_{\beta})\} \\ &= \sum_{\beta=1}^n \lambda_{\beta} \cdot C(\mathbf{u}_{\alpha}, \mathbf{u}_{\beta}) \end{aligned}$$

The last line is equal to the left hand side of the simple kriging equations (Equation 1.7) and can be replaced by the right hand side:

$$Cov\{Y^*(\mathbf{u}) \cdot Y(\mathbf{u}_{\alpha})\} = C(\mathbf{u}, \mathbf{u}_{\alpha})$$

This final covariance, $C(\mathbf{u}, \mathbf{u}_{\alpha})$, is the correct model covariance. Thus, kriging reproduces the covariance between the data, but the smoothing effect changes the variance and the variogram is not reproduced.

1.5 Missing Variance

Under stationarity, the expected value of the variance at every location should be the global variance $\sigma^2 = C(0)$. The variance of the kriging estimate can be calculated:

$$Var\{Y^*(\mathbf{u})\} = E\{[Y^*(\mathbf{u}) - m]^2\}$$

(Equation 1.2) is substituted into the quadratic term. Note that the expected value of the estimate is zero.

$$\begin{aligned} Var\{Y^*(\mathbf{u})\} &= \sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_{\alpha} \lambda_{\beta} E\{Y(\mathbf{u}_{\alpha}) \cdot Y(\mathbf{u}_{\beta})\} - 0^2 \\ &= \sum_{\alpha=1}^n \lambda_{\alpha} C(\mathbf{u}, \mathbf{u}_{\alpha}) \end{aligned}$$

If these results are substituted into (Equation 1.8) and rearranged:

$$Var\{Y^*(\mathbf{u})\} = C(0) - \sigma_{sk}^2(\mathbf{u}) \quad (1.9)$$

Note that the variance of the estimate is only correct when the kriging variance is equal to zero. This will only occur at the data locations. At all other locations the variance will be too small and the reduction in variance is equal to the kriging variance.

1.6 Addition of a Random Component

To correct the smoothing effect of kriging, the missing variance must be added back into the estimate without changing the expected value of the estimate or the covariance between the estimate and the data values. This is done by Monte Carlo simulation (Journal 1989). The simulated residual is added to the kriging estimate resulting in the simulated value.

$$y_s(\mathbf{u}) = y^*(\mathbf{u}) + r(\mathbf{u})$$

The simulated residual, $r(\mathbf{u})$, is drawn from a distribution with a mean of zero and a variance equal to the kriging or missing variance. This residual is added to the kriged estimate to produce the final simulated value retained at location \mathbf{u} . A different outcome is obtained if the process is repeated. Since the expected value of the residual is zero, $E\{R(\mathbf{u})\} = 0$, the simulated value is equal to the kriged estimate in expected terms, $E\{Y_s(\mathbf{u})\} = E\{Y^*(\mathbf{u})\}$. The variance of the simulated value will be increased due to the addition of the residual.

$$\begin{aligned} Var\{Y_s(\mathbf{u})\} &= Var\{Y^*(\mathbf{u})\} + Var\{R(\mathbf{u})\} + 2 \cdot Cov\{Y^*(\mathbf{u}), R(\mathbf{u})\} \\ &= [C(0) - \sigma_{sk}^2(\mathbf{u})] + \sigma_{sk}^2(\mathbf{u}) + 0 \\ &= C(0) \end{aligned}$$

The variance of $Y_s(\mathbf{u})$ is now equal to the expected variance and the expected value is unchanged. Furthermore, the covariance remains unchanged:

$$\begin{aligned}
Cov\{Y_s^*(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} &= E\{Y_s^*(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} \\
&= E\left\{\left[\sum_{\beta=1}^n \lambda_\beta \cdot Y(\mathbf{u}_\beta) + R(\mathbf{u})\right] \cdot Y(\mathbf{u}_\alpha)\right\} \\
&= \sum_{\beta=1}^n \lambda_\beta \cdot E\{Y(\mathbf{u}_\alpha) \cdot Y(\mathbf{u}_\beta)\} + E\{R(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} \\
&= \sum_{\beta=1}^n \lambda_\beta \cdot C(\mathbf{u}_\alpha, \mathbf{u}_\beta) \\
&= C(\mathbf{u}, \mathbf{u}_\alpha)
\end{aligned}$$

Where $E\{R(\mathbf{u}) \cdot Y(\mathbf{u}_\alpha)\} = 0$ since the random component is independent of the kriged estimator; therefore, the covariance between them is zero. The covariance has not been changed with the addition of the residual; it is still the right covariance.

The process so far has described how a value at a specified location is simulated conditional to the input data. To populate a whole model, every unsampled location will undergo the same process in a sequential order. Starting with a limited number of input data, an unsampled location is chosen and the simulated value is calculated. Moving to the next location, the previously simulated value is added to the conditioning data and the process is repeated. This continues until every location in the model is populated. Note that including the previously simulated values as conditioning data will produce the right covariance between the simulated values.

1.7 Gaussian Simulation

Sequential simulation is designed to create multiple realizations of the underlying geostatistical model. This is possible since there is an infinite number of values that can be drawn from the distribution at each location. The global distribution for every realization should reproduce the input distribution within statistical or ergodic fluctuations. Simulation will reproduce the global mean and variance, but there is no control over the shape of the resulting distribution. This final shape is a function of the the input distribution, the shape of the residual distributions and the shape created by combining the average of multiple simulated locations.

The input distribution is based on the input data or a reference distribution from another source. This distribution is what is trying to be reproduced, however, it can have any shape. Data transformation (Section 2.1) provides a way to transform the input distribution to any other distribution and back again. If the shape, mean, and variance of the transformed distribution are maintained, backtransformation will return the data to their original distribution and units.

The simulated value at each location is a single sample taken from the distribution of possible values at that location. The final distribution is then calculated from the collection of simulated values. The averaging process of kriging invokes the central limit theorem (Scheaffer 1995). Under this theorem, the sum of a large

number of independent and identically distributed samples taken from any distribution will produce a distribution that is Gaussian in shape. Even though there may be relatively too few variables and they may be correlated, the results of kriging tend towards a Gaussian distribution. The Gaussian distribution is a "bell" shaped curve described by the Gaussian density function (Deutsch 2002):

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(z-m)^2/\sigma^2}$$

where m is the desired mean and σ^2 is the desired variance for the Gaussian distribution.

The Gaussian distribution has the unique property that when multiple Gaussian distributions are combined, the resulting distribution will maintain the Gaussian shape. Since kriging tends towards Gaussianity, this shape should be maintained throughout the simulation process. Data transformation allows us to transform the input data to be Gaussian. The distribution of the residuals is chosen to be Gaussian and the simulated values will be approximately Gaussian in shape. Therefore, maintaining the Gaussian shape throughout the simulation process will preserve this shape in the output distribution and backtransformation can be used to return the data to its original distribution. The expected distribution of simulated values from SGS is Gaussian in shape with a mean of zero and variance of one. Deviating from these results indicates a problem and must be investigated.

Chapter 2

Sequential Gaussian Simulation Implementation Details

Sequential Gaussian simulation is a common method used to create multiple equiprobable numerical models based on a set of conditioning data (Isaaks 1990). Since the early 1990's, SGS has gained in popularity due to its simplicity, flexibility, and reasonable CPU time (Deutsch 2002a). The basic methodology is not difficult, but there are some important implementation details.

2.1 Data Transformation

A key component to SGS is the Gaussian distribution. It is rare to find real data that follows a Gaussian distribution exactly. To correct this problem, the data are transformed to follow a standard Gaussian or normal distribution. This process is reversible so the data can be returned to their original distribution and units.

Normal score transformation programs, like `nscore` (Deutsch 1998), will transform any set of z -data to a standard normal distribution $G(y)$ (Figure 2.1). Starting with the z -data's cumulative distribution function (cdf), $F_Z(z)$, the transformation is achieved by:

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

where z is a data value in original units and y is the corresponding normal score transformed value. The transformation is achieved by linking the two cdfs through their quantiles. Reversing this process will yield the original z -value:

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

SGS is designed to reproduce the input distribution within statistical fluctuations. Typically, the cdf used in transformation is based on the input data, however, data collection is not usually preformed to define the underlying distribution (Deutsch 2002a, p50). This will produce realizations that honour the input data

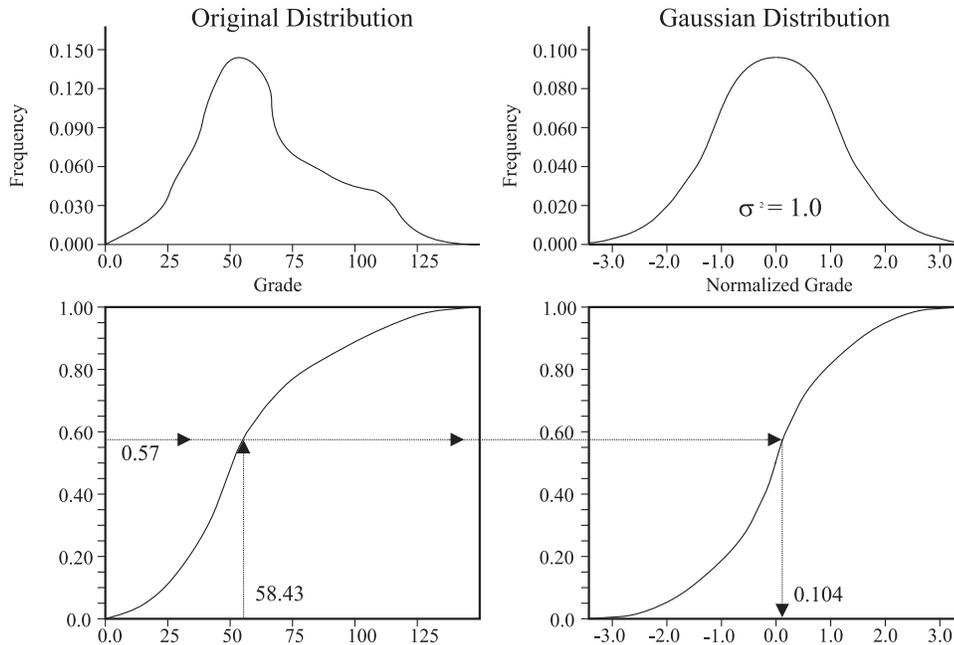


Figure 2.1: Starting with the z-data's cdf, the quantile for any z-value can be found. This quantile is then transferred to the standard normal cdf and the corresponding y-value is found.

and not the desired underlying distribution. To fix this problem, the input data must be corrected to be more representative or more information is required about the true global distribution.

Preferably, an alternative source can be used to define the global distribution and this reference distribution is then used to transform the input data. In most situations, the input data is the only available information. One way to correct for the sampling process is to decluster the data. Declustering is a process that weights the data based on their area of influence relative to surrounding data. If many samples are taken close together, they are all sampling the same part of the underlying distribution. To correct for this over-sampling, the data will be weighted to reduce their influence on the global distribution. Alternatively, a single sample far away from other data provides all of the information for a large area. To correct for this under-sampling, the weight applied to this data will be large, increasing its influence on the global distribution. To try to correct for the sampling process, the weights are applying to the data in the construction of the global cdf.

One problem that can occur in the transformation is caused by multiple data with the exact same value. These data cause a spike or vertical jump in the cdf. This spike of a single value will cover a range of quantiles in the original cdf, and this causes problems in the transformation. To correct this problem, a despiking process is used to remove the vertical jump in the cdf and transformation can proceed.

For multiple variables, it is common practice to transform every variable using a normal score transform. This collection of independent transformations will not ensure multivariate Gaussianity. To ensure multi-Gaussianity, a different transformation, like Stepwise Conditional Transformation (Leuangthong 2003b), Section 5.2, should be considered.

2.2 Assign Data to Grid Nodes

SGS can consider data at their original locations or relocate them to the nearest grid node. Since the model created by SGS is typically based on a cell centred grid network, it is computationally efficient to move the data to grid nodes. For each grid cell containing data (Figure 2.2) only the data inside of the cell will be considered. The data point closest to the centre of the cell is retained and this data is exactly reproduced in every realization. Note that data cannot cross the cell boundaries. The other data are only used in the definition of the global distribution $F_Z(z)$.

Assigning the data to grid nodes reduces covariance calculations in the simulation process. Since the data fall on a regular grid system, the covariance between the nodes can be calculated and stored in a look-up table. The covariance between two nodes is easily found by calculating the number of nodes separating them in each direction and then use the look-up table to find the corresponding covariance.

There are two problems that can arise when assigning the data to grid nodes: data outside the modelling area and large shifts in location. In some studies, only a portion of the sampled area is to being simulated, but all of the data is included to better define the input distribution. Historically, the data outside the modelling area was shifted to the edge of the grid system, potentially over-populating the edge nodes and moving data large distances. To prevent this problem, only the data inside of the modelling area are considered when assigning data to the grid. Large shifts in location can occur when the grid spacing is much larger than the sample spacing. If the shifts are small, the error in the covariance will be small and average out over the modelling area. Working with a grid that is much smaller than the sample spacing will help to prevent large shifts and reduce the number of data lost when assigning the data to the nodes.

2.3 Local Data

The theory behind SGS is based on using all data and previously simulated values to create an estimate at location \mathbf{u} . As simulation progresses, the addition of each new simulated value will continuously increase CPU demands. The last estimate in a model could potentially use $M - 1$ data in the kriging step, where M is the number of grid nodes in the model. If the data are not assigned to the grid nodes, this number is even larger. This require the inversion of a very large matrix and is

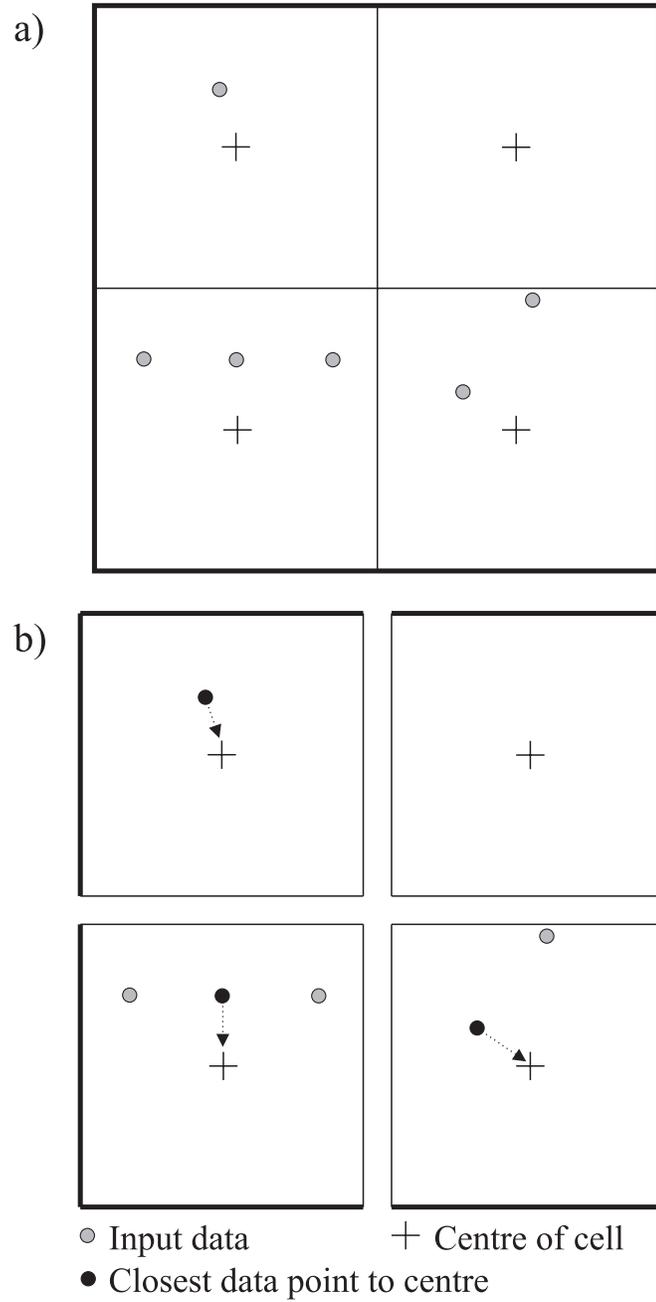


Figure 2.2: a) four grid cells with data and b) assigning the closest data to the grid node at the centre of the cell.

impractical. To make SGS practical, only the nearby or local data are used. This is reasonable since the data closest to the point being estimated heavily screen the data further away; however, choosing what data to use, the number of data, and efficiently searching for this data are important implementation decisions.

2.3.1 Maximum Number of Data

The number of conditioning data, n , used in the kriging system must balance two factors: CPU time and improvement to the estimate. As n increase, the improvement in the estimate is diminishing and the CPU demand is growing exponentially. For this reason, a maximum number of conditioning data are used.

To solve for the n kriging weights, (Equation 1.7) must be solved. This is expressed in matrix form:

$$\begin{bmatrix} C(\mathbf{u}_1, \mathbf{u}_1) & \cdots & C(\mathbf{u}_1, \mathbf{u}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n, \mathbf{u}_1) & \cdots & C(\mathbf{u}_n, \mathbf{u}_n) \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C(\mathbf{u}, \mathbf{u}_1) \\ \vdots \\ C(\mathbf{u}, \mathbf{u}_n) \end{bmatrix}$$

As the number of data increases, two things will require more CPU time: the search and solving for the kriging weights. As n becomes larger, the demand on the search increases by some multiple of n depending on the density of the surrounding data. The covariances from the n data will create an $n \times n$ matrix or a set of n linearly independent equations. The kriging weights are calculated by solving these equations simultaneously. The cost of this process is proportional to n^3 . The total change in CPU time will initially be a combination of the two factors, $n + n^3$ but as n becomes sufficiently large, solving for the kriging weights will dominate the CPU time.

The change in CPU time is shown with an example. A 100 x 100 grid was kriged using 300 randomly placed input data. The maximum number of conditioning data was varied between 5 and 300. The search radius and variogram range was set at 200 so that the maximum number would be found for every location. A 2.27 GHz Pentium 4 processor with 1.0 GB of RAM was used to calculate the CPU time (Figure 2.3). The figure shows that as the number of data increases from 5 to 50, the curve is continuously increasing. When more than 50 data are used, the N^3 term is dominating the CPU time and the slope becomes constant. The solid line on the figure represents the N^3 slope where one order of magnitude change in the X direction will cause a three order of magnitude change in the Y.

The second reason to limit the number of data used in SGS is that the closer data screens the data further away. All of the available input data and previously simulated value supply some degree of information; however, as more data are included, the distant data are screened by the closer data. This screening reduces the contribution supplied by the data and eventually the increase in information will be small enough to ignore (Isaaks 1990, p16). The decision to consider only nearby data is a Markov-type screening approximation, where the closest data in

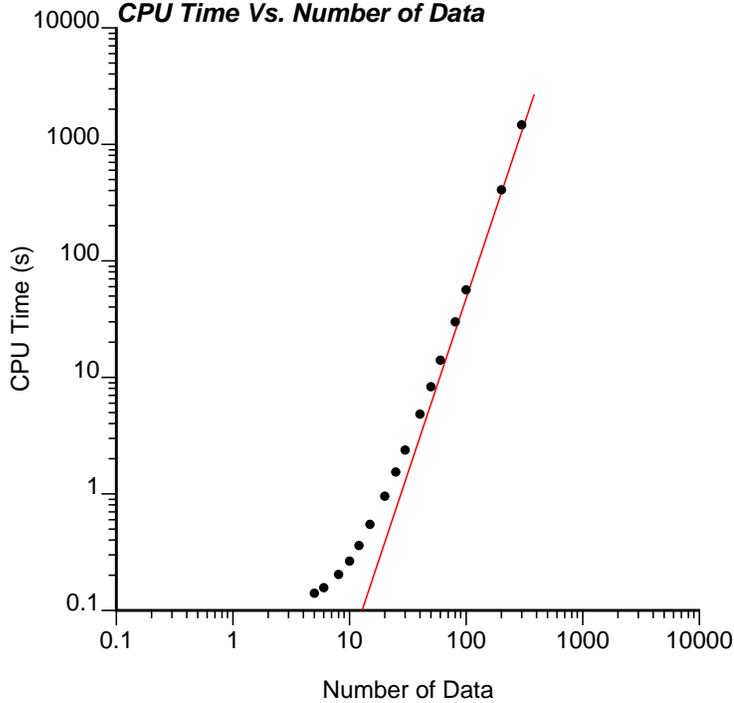


Figure 2.3: The change in CPU time as the number of conditioning data increases.

each direction provide the same information as including the more distant data in the same direction. These data will have large kriging weights associated with them to reflect their importance in determining the kriging estimate and variance. The data from further away will be overshadowed by the closer data and their kriging weights are small to reflect the limited increase in information.

The kriging weights account for two things, the distance from the point of estimation to the conditioning data and the redundancy between the conditioning data. The closer this weight is to zero (regardless of the sign) the less information that point supplies to the kriging system. Therefore, the point with the smallest absolute kriging weight supplies the least amount of information to the estimate. (Gandin 1965, p70) calculated the bounds on how the kriging variance would increase if this least informative point was excluded. The calculation requires three sets of kriging weights (Figure 2.4). To start, kriging is performed using n data to calculate the kriging weights, λ_α $\alpha = 1, \dots, n$, and a kriging variance, $\sigma_{sk,n}^2$. In the example there are five data that produce a kriging variance of 0.3097. The location that supplies the least information, k , is found by squaring the kriging weights and taking the location with the minimum value.

$$k = \min\{\lambda_\alpha^2\}, \alpha = 1, \dots, n$$

In the example, location 5 has the smallest squared weight at 0.0064.

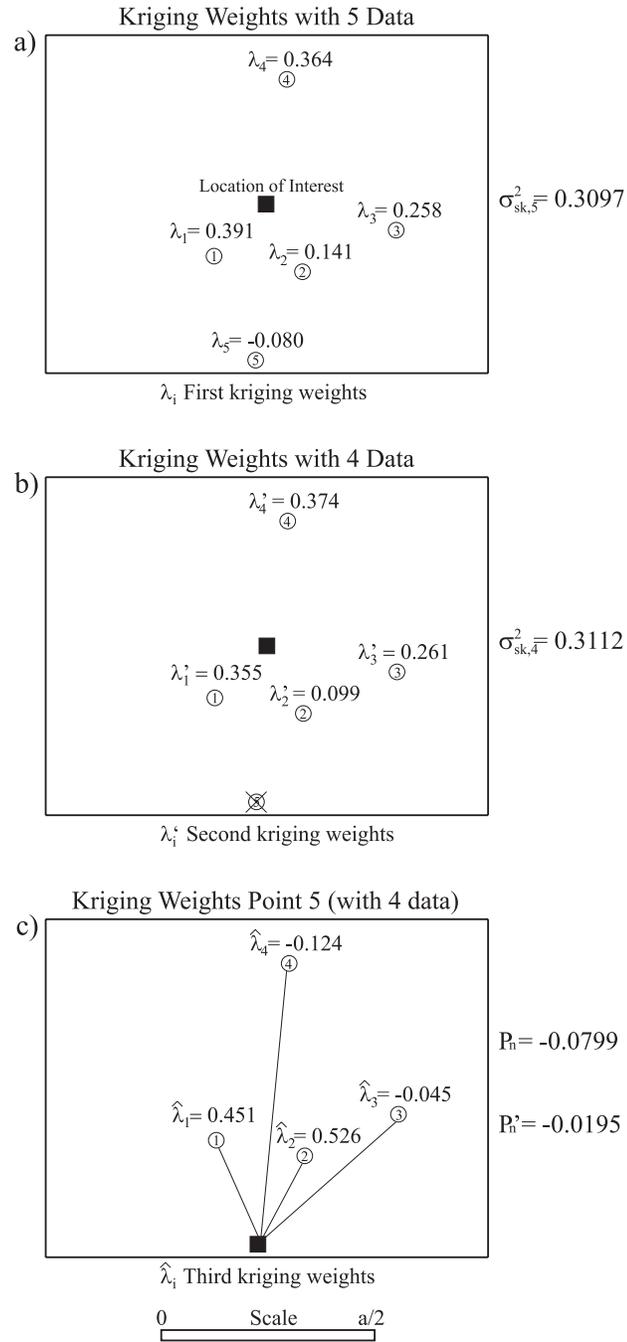


Figure 2.4: a) the configuration of the five input data along with the kriging weights and variance, b) the new kriging weights and variance when the 5th data point is not used, and c) changing the location of interest and performing a third kriging to obtain another set of weights and a new variance. Note a is the range of the variogram.

The data point at location k is removed and kriging is performed with the remaining $n - 1$ data (Figure 2.4b):

$$\sum_{\alpha=1, \alpha \neq k}^n \lambda'_\alpha \cdot C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}, \mathbf{u}_\beta), \beta = 1, \dots, n, \beta \neq k \quad (2.1)$$

This second set of kriging weights, λ'_α , will produce a slightly higher variance than before, $\sigma_{sk, n-1}^2$ or 0.3112.

Using the same $n - 1$ data but moving to location k , a third kriging is performed (Figure 2.4c):

$$\sum_{\alpha=1, \alpha \neq k}^n \hat{\lambda}_\alpha \cdot C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}_k, \mathbf{u}_\beta), \beta = 1, \dots, n, \beta \neq k \quad (2.2)$$

Note that the right hand side covariances are in terms of \mathbf{u}_k . This provides the set of kriging weights, $\hat{\lambda}_\alpha$, that would be used to find an estimate at location k . The λ'_α and $\hat{\lambda}_\alpha$ kriging weights are related using the following relationship:

$$P_n = \frac{P'_n}{C(\mathbf{u}_k, \mathbf{u}_k) - \sum_{\alpha=1, \alpha \neq k}^n \hat{\lambda}_\alpha \cdot C(\mathbf{u}_k, \mathbf{u}_\alpha)} \quad (2.3)$$

where $C(\mathbf{u}_k, \mathbf{u}_k) = 1$ for a standardized variable and P'_n is equal to:

$$P'_n = C(\mathbf{u}, \mathbf{u}_k) - \sum_{\alpha=1, \alpha \neq k}^n \lambda'_\alpha \cdot C(\mathbf{u}_k, \mathbf{u}_\alpha) \quad (2.4)$$

The three sets of weights are linked by the following relationship:

$$\lambda_\alpha = \lambda'_\alpha - \hat{\lambda}_\alpha \cdot P_n, \alpha = 1, \dots, n, \alpha \neq k \quad (2.5)$$

The change in variance by adding the data point at location k will be bounded by two inequalities:

$$\sigma_{sk, n-1}^2(\mathbf{u}) - \sigma_{sk, n}^2(\mathbf{u}) \leq (P_n)^2 \quad (2.6)$$

and

$$\sigma_{sk, n-1}^2(\mathbf{u}) - \sigma_{sk, n}^2(\mathbf{u}) \geq (P'_n)^2 \quad (2.7)$$

Using the results from the example, the upper bound (Equation 2.6) is 0.0064 and the lower bound (Equation 2.7) is 0.0004. The true change in the variance is 0.0015, which falls within the the bounding limits.

(Gandin 1965) has provided a way to determine bounds on how the kriging variance changes when the least informative conditioning data is removed; however, these bound are specific to the data configuration and would change if the location of the input data were different. Theoretically, this process could be implemented with SGS to find the optimum number of conditioning data based on the maximum

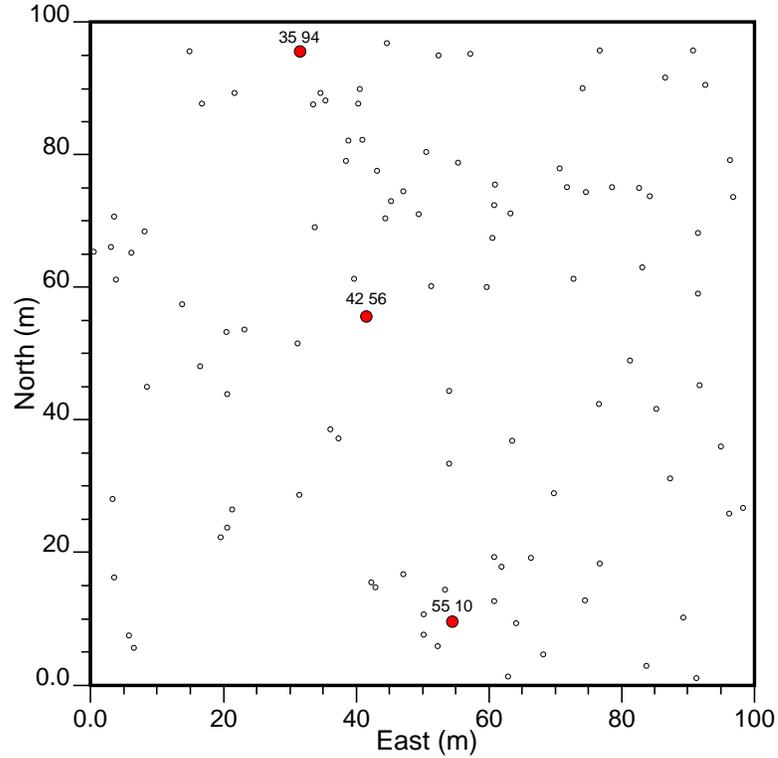


Figure 2.5: The input data is represented as small circles and the three test points as large dots.

possible change in variance. This would be very CPU intensive and the change in variance could be calculated directly from the weights. A more practical application would be to study several locations in the model area before simulation. The chosen locations should cover a range of data orientations from many nearby data to only a few distant data. At each location, several increments of conditioning could be looked at. For example, the change in variance between the 5 and 6 data, 10 and 11 data, 15 and 16 data, and so on. Based on the spread of the bounding limits and the maximum possible change in the variance, the maximum number of conditioning data would be chosen.

A 2-D example is used to better understand how the kriging estimate and variance change as the number of data increases. The model area (Figure 2.5) has 100 randomly placed input data and three locations were chosen to track how the kriging estimate and variance change as the number of input data is varied.

For the three chosen locations, the maximum number of conditioning data was varied between 1 and 100. The kriging estimates and variances were determined and plotted (Figure 2.6). The change in the kriging estimate for each location, a), c), and e), shows a large degree of fluctuation for the first 10 conditioning data. The initial conditioning data can all have very different values with relatively large

weights. Since the kriging estimate is a weighted combination of the conditioning data, the more variability in the conditioning data the larger the fluctuations. Once more than 10 data are used, there is almost no change in the estimate and any fluctuations are centred on the best estimate.

The kriging variance, b), d), and f), continuously decreasing with the addition of more data. Initially, the change in the variance is large, but continuously approaching the minimum value. Once about 7 to 10 conditioning data are used, the change in the variance is close to the minimum and the slope is almost flat. Sensitivity of these results to the variogram and nugget effected were performed, but the rate of convergence stayed the same.

In general, the conditioning data should be representative of the surrounding area. As the data becomes more redundant and/or distant, the kriging weights will become smaller. Once more than 10 conditioning data are used, the kriging estimate and variance are mostly fixed. Even when the data point comes from a previously unrepresented direction, there is almost no change in the kriging mean and variance.

In determining the maximum number of conditioning data, the CPU requirements and the accuracy of the kriging estimate and variance must be considered. The lower limit for the maximum number of data would be 8 to 10 data. At this number, the kriging estimate and variance are close to convergence and the CPU time is still small relative to using more data. If possible, more data should be used depending on the size of the model, the number of realizations and the limitations placed on CPU time. If zonal anisotropy or periodicity have been observed in the data, this number should increase to help capture these features in the model and to improve variogram reproduction.

2.3.2 Searching for Data

To preform kriging at an unsampled location, the surrounding data must be identified. A local neighbourhood is defined by specifying a search ellipsoid that corresponds to the principle ranges and directions of continuity based on the variogram. Data beyond these ranges will have a limited number of pairs in calculating the variogram and the covariance will be poorly understood (Deutsch 1998). Incorporating this data would provide little to no extra information. The presence of zonal anisotropy or periodicity in the data is an exception to this rule and the search ellipsoid may need to be increased.

An exhaustive search can be performed to check every location for input data, or previously simulated values, and their covariance calculated to the point of interest. These covariances could then be sorted into descending order and the data with the largest covariances, up to the maximum, are used in the kriging system. This process is not computationally efficient and is only practical for small grids.

The spiral search (Deutsch 1998) is an efficient way to find data on a regular grid. Starting at the location being estimated, the search spirals towards less correlated,

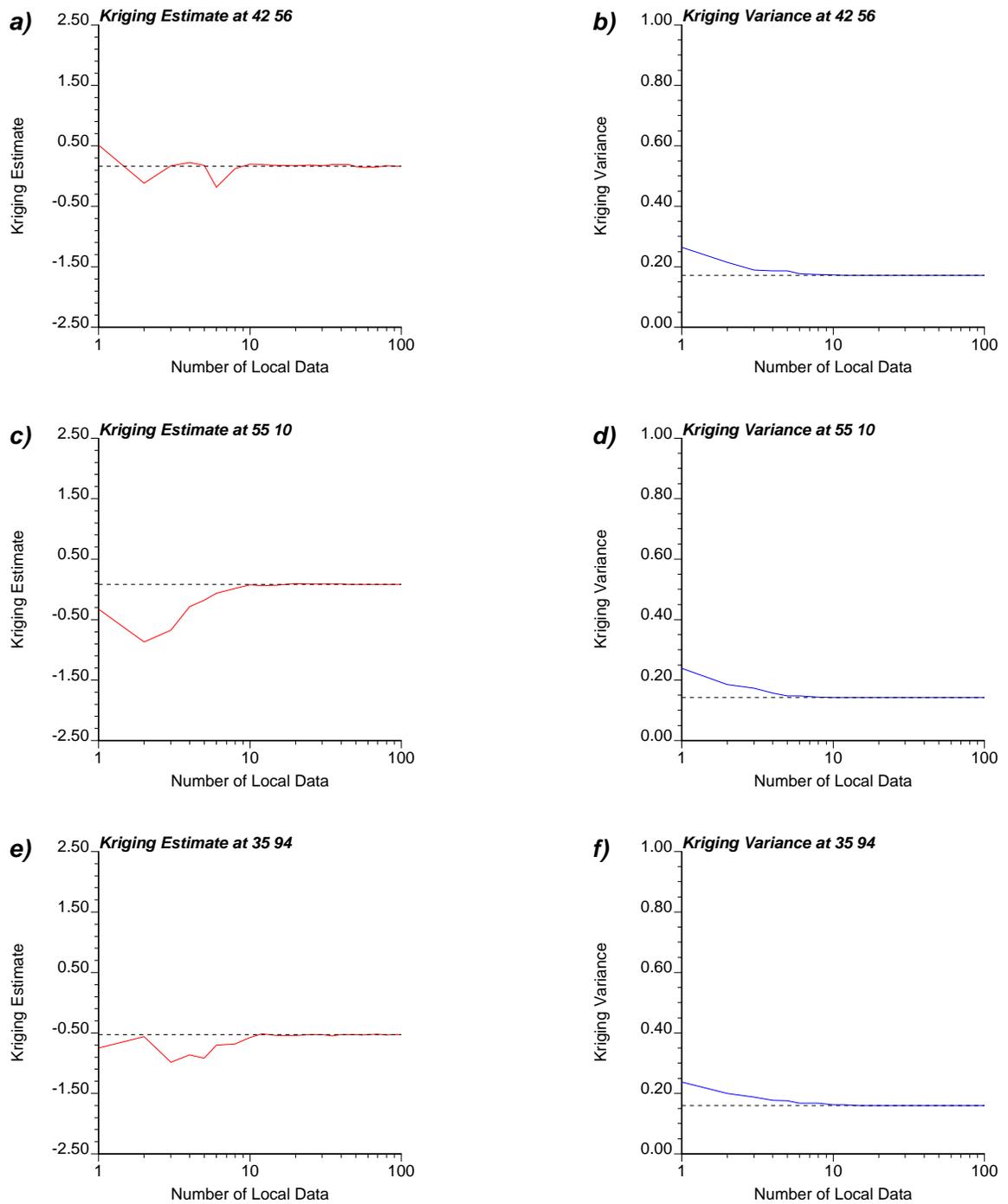


Figure 2.6: The kriging estimates, a), c), and e), and variances, b), d), and f), for the three locations of interest. The dotted lines on each plot is the best possible estimate when 100 conditioning data area used.

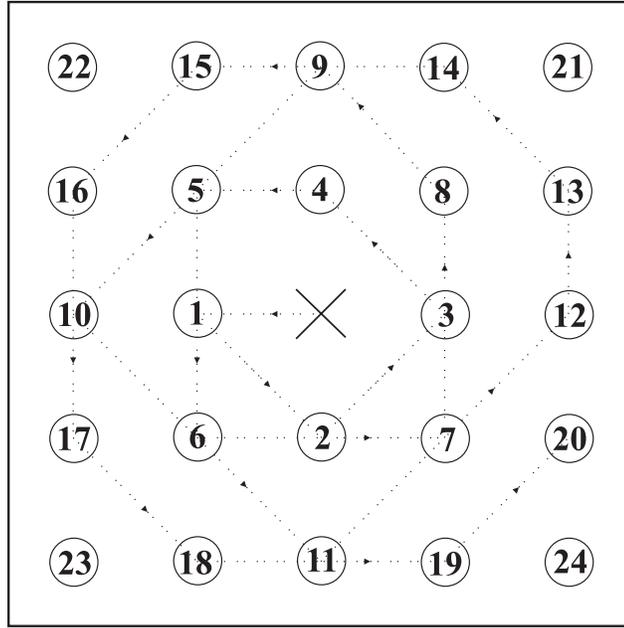


Figure 2.7: The spiral search starts at the closest nodes from the central point of interest and spirals towards increasingly less correlated nodes. The first 24 nodes are shown in this 2-D example for an isotropic variogram.

or increasingly distant, nodes based on the defined grid system, and the variogram (Figure 2.7). The search continues until the maximum number of data have been located or the search radius is reached. If the data shows some degree of anisotropy, the search will become elliptical to account for the different ranges of continuity. In 3-D, the nodes above and below the 2-D plane are included in the search path. The search will simply start at the node with the highest covariance, regardless of direction, and then continue through nodes with decreasing covariance. The spiral search requires a minimum of CPU time since as few as possible locations are searched and no ordering is required.

The super block search (Deutsch 1998) is designed to efficiently find non-gridded data in a local neighbourhood when the number of input data is constant. To start, an independent grid is placed over the model area to create blocks (Figure 2.8). A template is then constructed based on the number of super blocks that represent the search ellipsoid. The template is centred on the super block that contains the point of interest. As simulation progresses, the template is moved around the map to quickly identify all the surrounding super blocks inside the local neighbourhood. These super blocks are then searched to locate the conditioning data.

The speed advantage of the super block search comes from the way it sorts and indexes the data. Starting at one corner of the map, the super blocks and data are numbered (Figure 2.9). In ascending order, each super block is searched and the

Search Ellipsoid Super Block Template

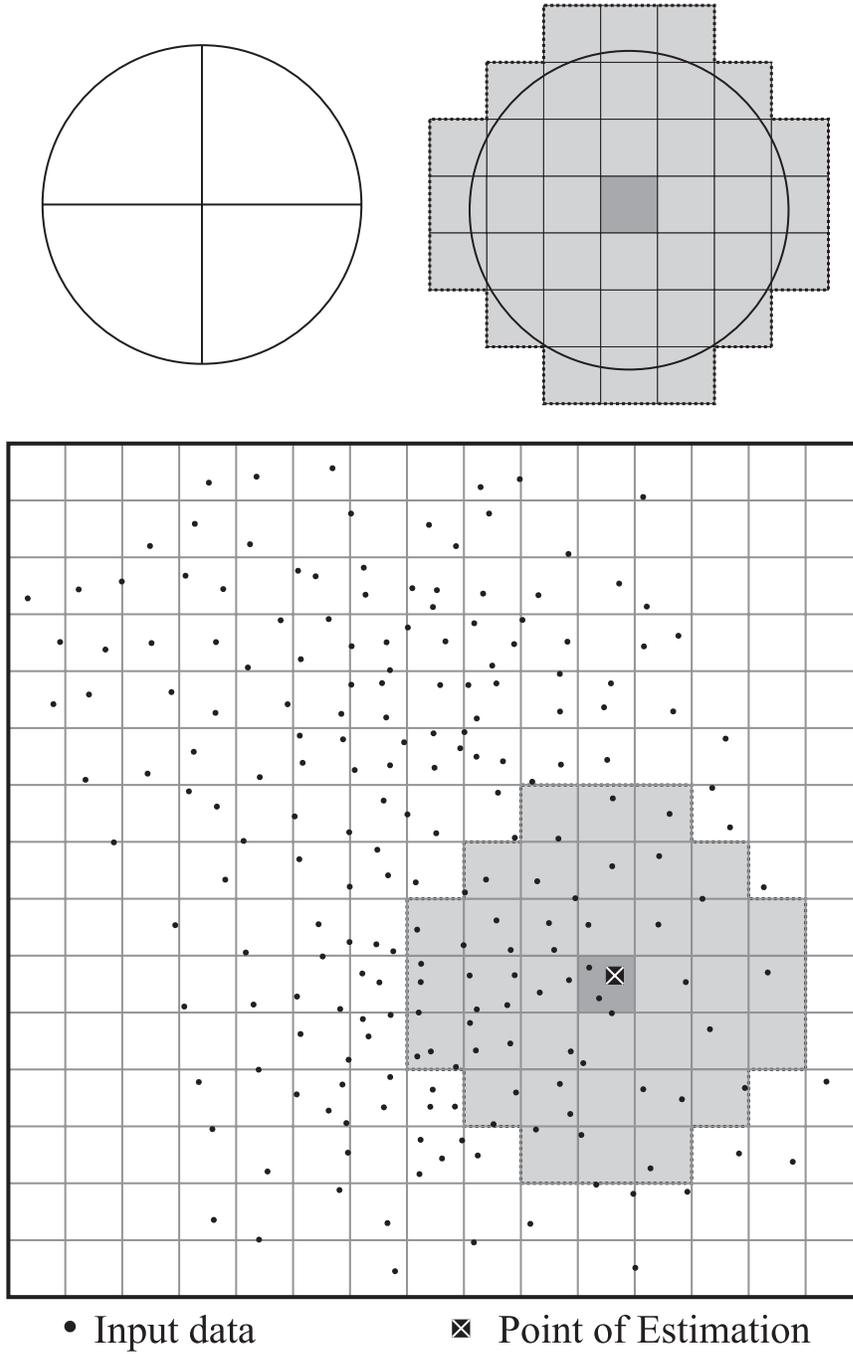


Figure 2.8: A 2-D illustration of how the super block search divides the modelling area into blocks and constructs a template based on the search ellipsoid.

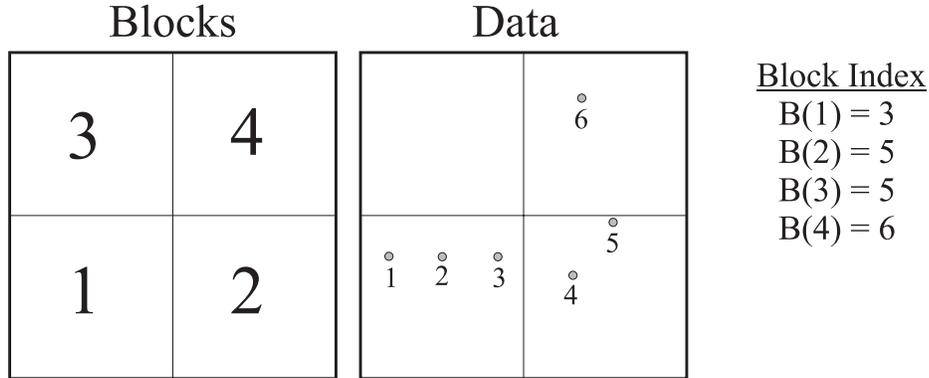


Figure 2.9: The super block search numbers all super blocks and then identifies the data in each block. The cumulative number of data is stored in an array and the data is indexed for quick retrieval.

cumulative number of data is stored in an array:

$$Block(i) = \sum_{j=1}^i dat(j), \quad i = 0, \dots, n$$

where $Block(0) = 0$, $dat(i)$ is the cumulative number of data for the first i super blocks, and there are m super blocks in total. The number of data inside of any super block is found by subtracting the previous super block value from the super block of interest, $Block(j) - Block(j - 1)$. The index location of data inside of this super block ranges between $Block(j - 1) + 1$ to $Block(j)$. Using the example in (Figure 2.9), super block 2 has $5 - 3$, or 2, data and their index locations are 4 and 5, respectively.

Working with non-gridded data in simulation requires a two part search. The spiral search locates all previously simulated nodes on the grid system and the super block search finds the non-gridded input data. Keeping the two data types separate allows input data to be preferentially used over previously simulated values.

The cost of using the two part search is CPU time due to the setup and operation of the searches. To understand the CPU cost, the same time example as (Section 2.3.1) will be used with some minor changes. The number of previously simulated values will be set to zero so only the non-gridded input data will be used. The grid will be reduced to 97 X 100 since the previous example did not simulate the 300 gridded input data. The CPU time (Figure 2.10) for the two part search, dots, is higher than using just the spiral search, black line. Initially, the two part search can take over an order of magnitude more time than just the spiral search. Since both methods use the spiral search, this increase in time is due to the setup and execution of the super block search. As the number of data increases, the difference between the search becomes less important since matrix inversion takes most of the

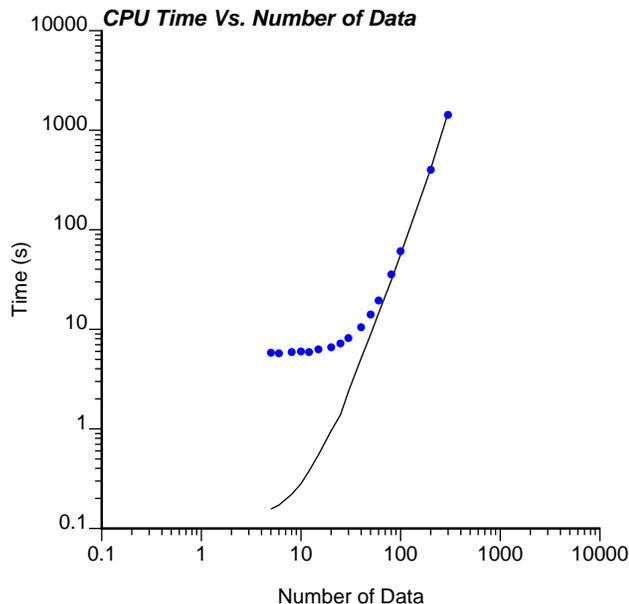


Figure 2.10: The CPU time using a two part search is shown by the dots. The solid black line shows the change in CPU time when only the spiral search is used.

CPU time. Note, once the super block search is set up, the CPU time should not change since the same blocks are checked every time.

Drillhole data is typically densely sampled in the vertical direction and sparsely sampled in the horizontal. Working near a drillhole can be problematic since the maximum number of conditioning data is quickly reached from a single drillhole, causing the horizontal data to be poorly represented. To force data to come from different directions, an octant search (Deutsch 1998) can be used. The octant search divides the search ellipsoid into octants in 3-D and quarters in 2-D. A maximum number of data are retained from each octant. Since a straight drill hole typically only goes through two octants, a minimum of three octants with data should be used to ensure the data comes from more than one drillhole.

2.4 Simulation Path

Simulation does not distinguish between input data and previously simulated values. After a location has been simulated, it will be added to the set of possible conditioning data and is used to help inform surrounding locations. The influence of previously simulated values will extend to the surrounding unsampled values, causing the order in which the nodes are visited to be important. If a poor choice of path is used then artifacts can be added to the model. This should be avoided so

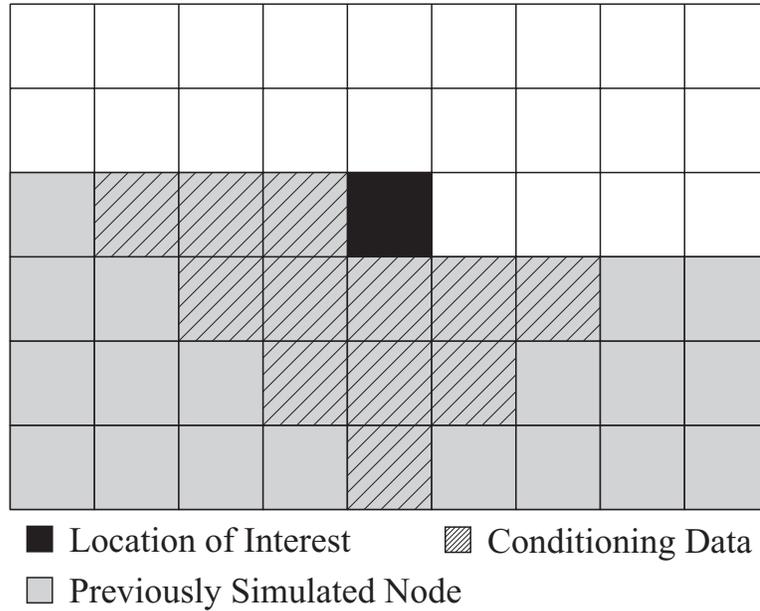


Figure 2.11: In areas away from the edge of the model, the same relative nodes will be used in the kriging system for a regular simulation path. All locations that fit this template can use the same kriging weights.

that any structure in the model is based on the input data and the model variogram.

2.4.1 Regular Path

Following the grid system in a regular fashion is one possible simulation path. The GSLIB (Deutsch 1998) grid format starts at the bottom left hand corner of the lowest level. The grid then cycles in the X direction first, followed by the Y and then Z directions. The simulation path could follow the same sequence.

This regular path would reduce the number of matrices that must be solved when kriging. As the path moves away from the edges of the model, assuming there is no input data, the same data locations relative to the point of estimation will be used. Therefore, the same kriging weights can be used repeatedly, removing the need to solve the same kriging matrix. The number of locations that conform to this template depends on the search radius and the presence of input data. If a small search radius is used and there are no input data, then up to 90% or more of the model will use the template weights.

Although this choice of path may be computationally fast, the input variogram will be poorly reproduced (McLennan 2002) (Figure 2.12). Variogram reproduction, within ergodic fluctuations, is important when creating a simulated model to ensure the proper geological structure is maintained. This is enough to discourage the use of this method.

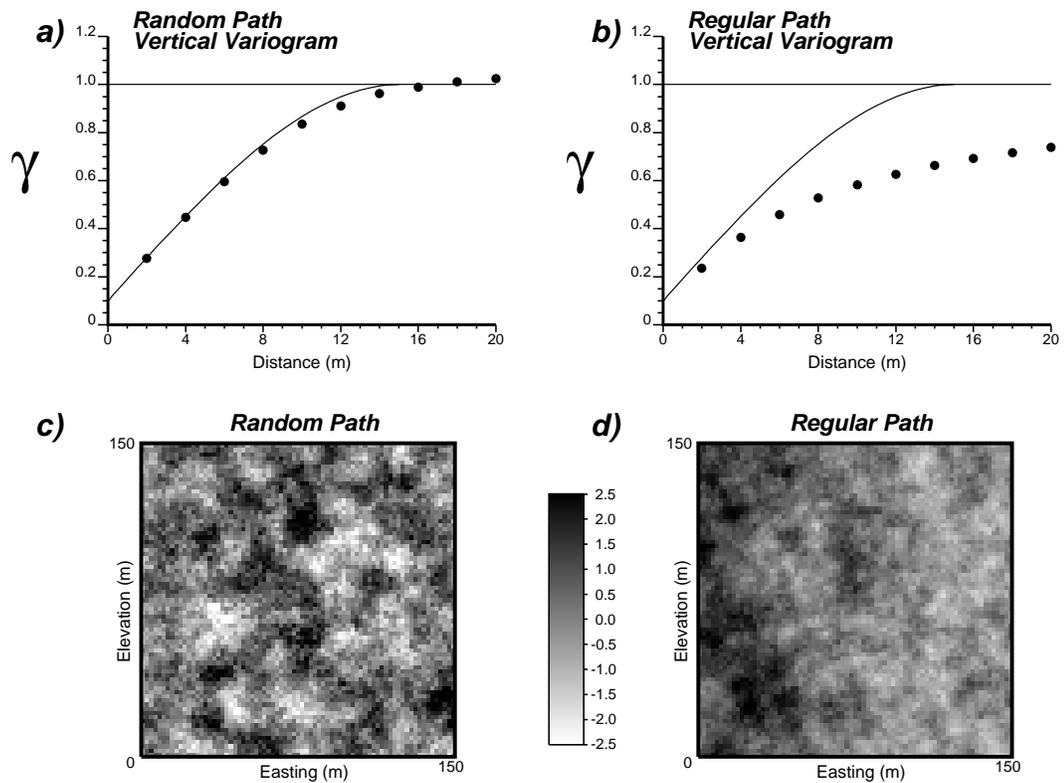


Figure 2.12: The average North variogram reproduction, dots, for 100 realizations using a random simulation path, a), and a regular simulation path, b). The solid line is the input variogram. One realization for the random path, c), and the regular path, d), are shown. The units have been transformed to follow the Gaussian distribution. This figure is taken from (McLennan 2002).

2.4.2 Spiral Path

A spiral path, moving away from input data, could propagate the influence of the input data over more of the modelling area (McLennan 2002). The path is created by finding the distance between each node and the closest input data. These distances are stored and sorted in ascending order. Simulation proceeds by visiting the nodes in this order (Figure 2.13). To break the ties caused by a regular grid, a small random component can be added to the distances.

This method will preferentially travel through all the locations closest to the input data. The problem is that as simulation progresses, each location will be next to several previously simulated values. This causes most of the conditioning data to come from the same direction, with only small separation distances. Using only the closest data will cause poor variogram reproduction in the large-scale features. This lack of variogram reproduction is enough to discourage the use of this method.

2.4.3 Random Path

Practice has shown that it is best to proceed randomly to prevent any influence of the path on the model (Deutsch 1998, Isaaks 1990, McLennan 2002). A random number is assigned to each grid node and then these values are sorted. The starting point and sequence of this path will change for every realization. Over multiple realizations, the structure found in the model will be based on the input data and model variogram. It will not be an artifact of the simulation path.

2.4.4 Multiple Grids

Using a fully random path and only nearby conditioning data may not reproduce the variogram structure at large distances in the model. To correct this problem, the simulation path can be performed on multiple grids (Tran 1994) where a random path is used for each grid. To start, a coarse grid is simulated to help capture the large-scale features in the model by working with conditioning data over long distances. This grid is then reduced in several steps until the final model grid size is reached. (Figure 2.14) shows an example of how the multiple grids is created using three different grid sizes. Incorporating multiple grids into the random path will improve variogram reproduction at large distances without compromising the benefits of the random path.

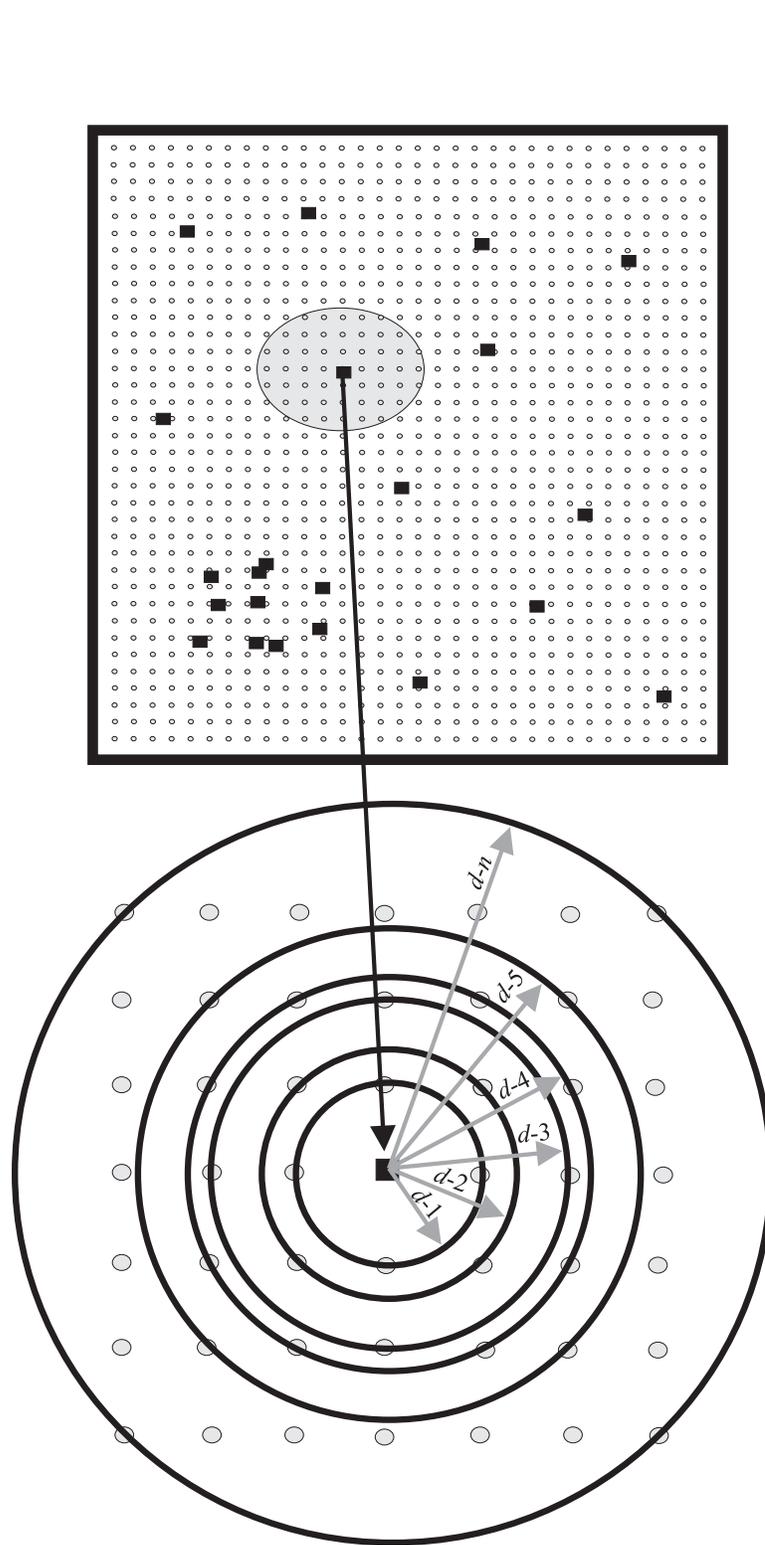


Figure 2.13: The spiral path first simulates all the nodes closest to the input data followed by increasingly more distant nodes. This figure is taken from (McLennan 2002).

Chapter 3

Non Stationary Parameters

Stationarity is a basic assumption required in geostatistics. It assumes all the data belong to the same statistical population and they share the same limit statistics (Journel 1978). In most situations, local trends are observed in the data that deviate from the global statistics. Ignoring these trends does not affect reproduction of the global histogram, but local structure in the models will not be reproduced. Trends can be observed in the mean, variance, and correlation with secondary variables. In structurally controlled areas, these trends can even be observed in the angles and anisotropy ratios that describe the spatial variability.

Rock types are another form of secondary information. Separating the data by rock type removes the need for a global assumption of stationarity; stationarity is assumed inside each rock type. The limiting factor in dividing the data is that there must be a minimum amount of data in each rock type to: inform the required statistics, transform the data, and model the variogram. In some cases, the data are correlated between rock types and this information should be allowed to cross the rock type boundaries. These soft boundaries allow data from either rock type to be used in SGS, but this can be problematic.

3.1 Mean

The most common violation of stationarity is a location-dependent mean. In many geostatistical studies, the modelling area will contain large regions of high and low values; thus, the local mean changes throughout the modelling area. Ideally, these regions would be modelled separately. This may not be practical due to the limited number of data, and the uncertainty in the boundaries between the regions due to the gradational nature of these changes. One approach is to group data based on similar characteristics and use a trend to describe local departures from the global average. The model that describes the local mean is constructed based on some form of secondary information, expert knowledge, or large-scale contouring of the data.

The local mean is built at the same scale as the simulated model to provide a mean value at every location. Under SK, the mean is subtracted from the data and the stationarity assumption is applied to the resulting residual values. Simple kriging with a locally varying mean (LVM) is written:

$$z_{LVM}^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{\alpha=1}^n \lambda_{\alpha} \cdot [z(\mathbf{u}_{\alpha}) - m(\mathbf{u}_{\alpha})] \quad (3.1)$$

where $m(\mathbf{u})$ and $m(\mathbf{u}_{\alpha})$ are the location-dependant means.

In this form, the covariances should come from the residual values. In practice, the variogram of residuals can be difficult to obtain, so the original data variogram is often used. This is reasonable at short distances since the trend most significantly affects the large-scale variogram. In this form, kriging provides an estimate of the residual value or the degree of fluctuation from the local mean. To return to original units, the calculated residual is added to the local mean. The kriging variance is expressed as:

$$\sigma_{LVM}^2(\mathbf{u}) = C(0) - \sum_{\alpha=1}^n \lambda_{\alpha} \cdot C(\mathbf{u}, \mathbf{u}_{\alpha}) \quad (3.2)$$

A 2-D synthetic example was created to show the affect of local means on kriging. To start, 30 primary data points were placed in a 50m x 50m area. To create the map of local means, the 30 primary data had a small random component added to them, reducing the correlation from 1.0. Some more data was added to the local means to create structure not captured by the primary data. The local mean data was block kriged produce a 50 x 50 cell map to cover the whole area. At the primary data locations, the block kriged local mean values were extracted from the map and a correlation of 0.750 was calculated between the primary data and local means. The primary data and local mean values (Figure 3.1a,d) were then used to create kriged maps using SK, b), and LVM, e). The SK map and LVM map have some similar properties and both honour the input. To see the differences, the SK values were subtracted from the LVM values to create a difference map, c). The scale of this map goes from -25.0, white, to +25.0, black, with the grey areas close to 0.0. The same scale was used for the LVM minus local means difference map, f). It is easy to see from the difference maps that LVM is similar to the local means in most areas, but shows much more deviation when compared the SK results.

To look at the affects of the local mean on the global distribution, the histograms for the primary data, SK values, LVM values, and the local mean values were created (Figure 3.2). The shape of the input data histogram is approximately lognormal with mean and standard deviation of 30.29 and 38.17, respectively. The SK run maintains the lognormal shape and almost the same mean value at 30.39, but the standard deviation decreased to 18.00. This decrease in standard deviation is caused by the limited number of high values as seen by the small tail on the distribution. The LVM distribution is a combination of the input data and the distribution of

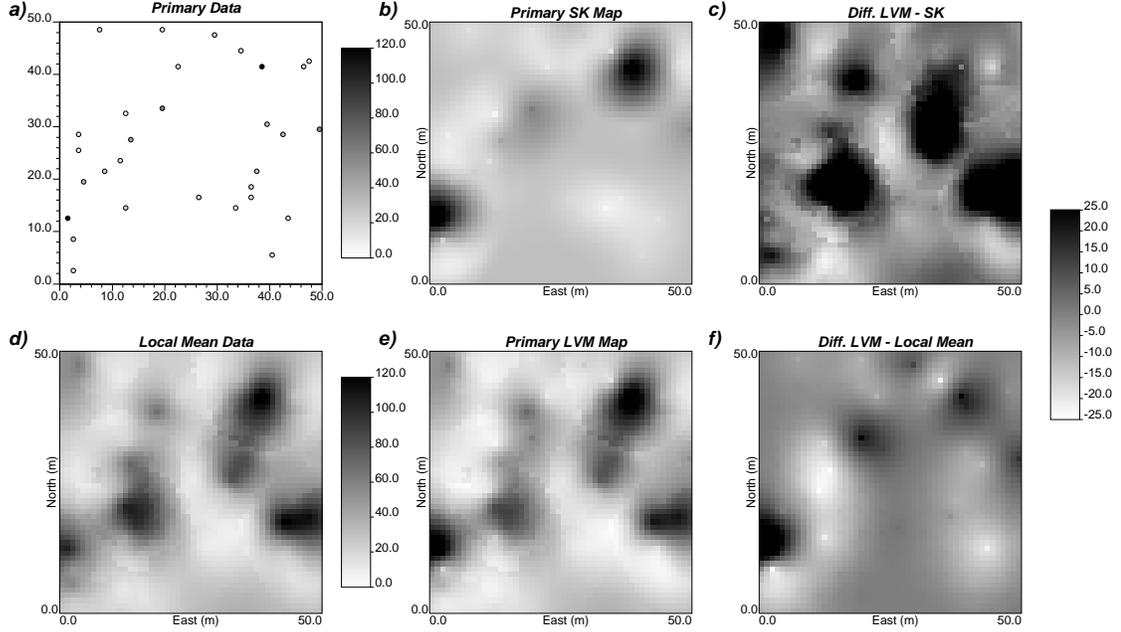


Figure 3.1: The 30 primary data locations, a), and local means, d), were used to create a SK map, b), and a LVM, e) for permeability in mD. Difference maps between LVM - SK, c), and LVM - local means, f), can be used to compare the results.

local means. The LVM mean at 35.26 is between the lower bound from the input data, 30.29, and the upper bound of the local mean, 38.05. This is also true for the standard deviation at 25.44 where the lower bound is 23.45 from the local means and the upper bound is 38.17 from the input data. Note that the local means used in this example may show more local features than usual. Changes in the local mean map are typically restricted to large-scale features.

LVM is a simple way to include information about a location-dependent mean in a one-step kriging process. In a simulation context, this process is not as simple. The addition of a random component when simulating with LVM can cause global variance inflation problems. This problem is further explained and a correction is proposed in Chapter 4.

3.2 Variance

Variability is another parameter that can be non stationary. When the change in variance is linked to the magnitude of the variable it is known as heteroscedasticity or the proportional effect (Isaaks 1989, Oz 2002). The variance is then a function of the local mean, $\sigma^2(\mathbf{u}) = f(m(\mathbf{u}))$, and is assumed to be pseudo-stationary or locally accurate. This property can be very pronounced in original units and is of

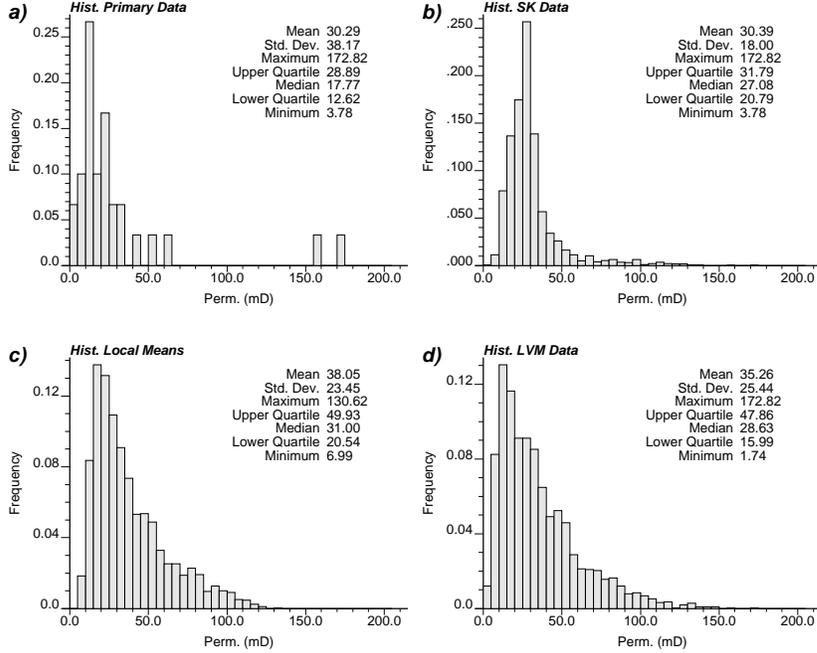


Figure 3.2: a) primary data histogram, b) SK histogram, c) local means histogram, and d) LVM histogram.

great concern in any direct sequential simulation technique. When a normal score transformation is performed, the proportional effect is assumed to be removed from the data (Deutsch 2002a, p167). For most cases this is reasonable, but sometimes a remnant of the problem can be seen in normal space. In these special cases, this change in variability should be accounted for in the simulation process.

The relative variogram (Olea 1991), $\gamma_R(\mathbf{h})$, is one solution that can be used to account for heteroscedasticity in the data. A semivariogram, $\gamma(\mathbf{h})$, is modelled from the data and is then scaled by a mean-dependant function, $f(m(\mathbf{u}))$, to obtain the relative variogram:

$$\gamma_R(\mathbf{h}) = \frac{\gamma(\mathbf{h})}{f(m(\mathbf{u}))} \quad (3.3)$$

where the scaling function is commonly the square of the mean, $f(m(\mathbf{u})) = m(\mathbf{u})^2$. This will adequately remove the effects of heteroscedasticity for application in SGS.

Alternatively, a scaling factor could be applied to the kriging variance to account for local departures from the global variance (Lyll 2000). To start, a scatter plot is created to compare the moving window means and variances for normal score data (Figure 3.3). The data is then divided into bins where there should be at least 10 data per bin and from 5 to 20 bins depending on the number of data. The bin width can either be constant, as shown, or vary so that each bin has the same number of data. The variance for each bin is calculated and the correction factor, F , is the bin

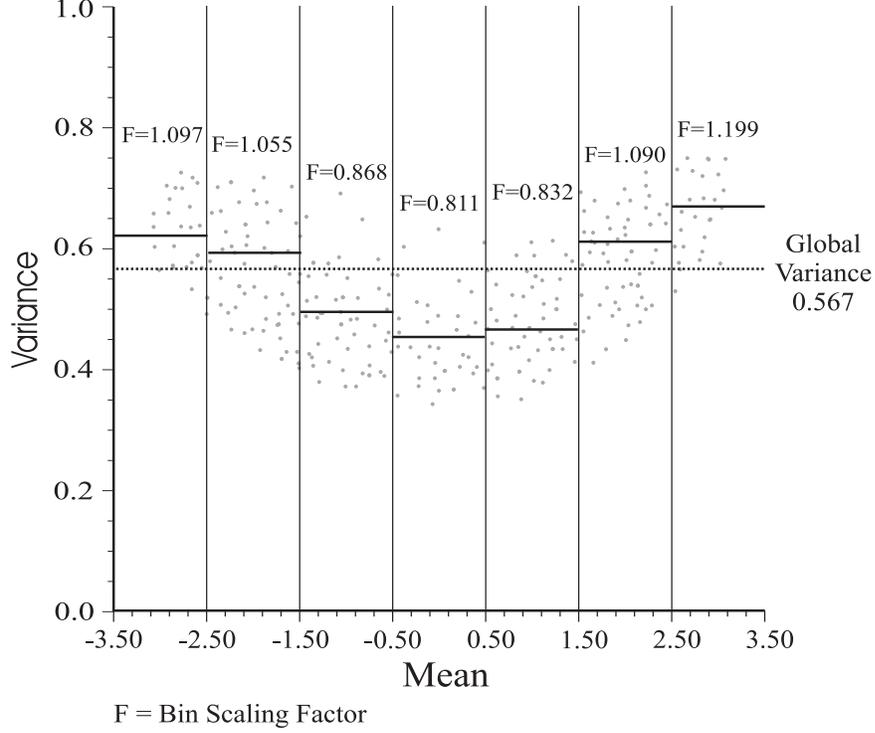


Figure 3.3: The variance scaling factor is calculated per bin on the scatter plot by dividing the bin variance by the global variance. The mean is in the Gaussian transform of the ordinal units. Note, the heteroscedasticity is overstated for illustrative purposes.

variance divided by the global variance:

$$F_{bin} = \frac{\sigma_{bin}^2}{\sigma_{Global}^2} \quad (3.4)$$

During simulation, the kriging estimate is calculated and is matched to a bin. The bin specific correction factor is applied to the kriging variance and simulation continues. This will not affect data reproduction since the kriging variance is zero at data locations. Correcting the variance locally will affect the global histogram. The input histogram will have poor reproduction, especially for the global variance.

The change in variance can be linked to regional fluctuations and changes depending on location. To describe how the variance changes, a trend map is constructed and the correction factor is calculated from this map. The location dependent correction factor, $F(\mathbf{u})$, is applied in the same way as above.

3.3 Correlation

Collocated cokriging (Xu 1992) (Section 4.1) incorporates secondary information into the estimate through the correlation between the two variables. This correlation is most often assumed constant, however, under certain geological conditions it can vary over the model area. Accounting for this change is simply a matter of calculating the local or regional correlation and applying this correlation to the cokriging equations. It is the geological context of how the correlation changes that needs to be established.

In the natural resources industries, certain geological phenomena can cause changes in the correlation between variables of interest. In mining, the correlation between the minerals can change throughout the deposit. Not accounting for these changes could cause misclassification of the material, especially in areas near the cut-off grade. In the petroleum industry, the correlation between static properties, or between the seismic response and static properties, may change throughout the reservoir. These changes in correlation can have an impact on the calculated reserves and predicted recovery factor. In both industries, even small changes in the resource model can have a large financial impact on projects.

3.4 Angles

The angles associated with the search radius and variogram modelling can be non stationary. It is possible to have preferential directions of continuity in some geological formations and other naturally occurring phenomena. If these directions of continuity change over the area of interest, they should be accounted for in the modelling process. A prior coordinate transformation is sometimes used to remove these changes. In complex 3-D cases, this transformation can be difficult. In these situations, the angles of continuity may become location-dependent.

Continuity is described by three angles and two anisotropy ratios for a 3-D area (Figure 3.4). Following the GSLIB (Deutsch 1998) convention, the three angles relate to the rotation around the X, Y and Z axis and the anisotropy ratios relate to the differences in the range of continuity in the X-Y, horizontal, and Y-Z, vertical, directions. Angle one is the degree of shift to create a new X' and Y' in the X-Y plane. The second angle is the tilt of the Z-Y' plane around the X' axis creating new Y'' and Z' directions. The third angle is the degree of rotation around the Y'' axis that shifts the X'-Z' plane to create new X'' and Z'' directions. The first anisotropy value is the ratio of the X'' range over the Y'' range and the second anisotropy value is the ratio described by the Z'' range over the Y'' range.

In most cases, the change in the principle directions and anisotropy ratios over the modelling area are small or non existent. The global averages for these parameters are the best choice. In complex 3-D cases, the average would produce poor results since the local neighbourhood would not be properly defined (Figure 3.5),

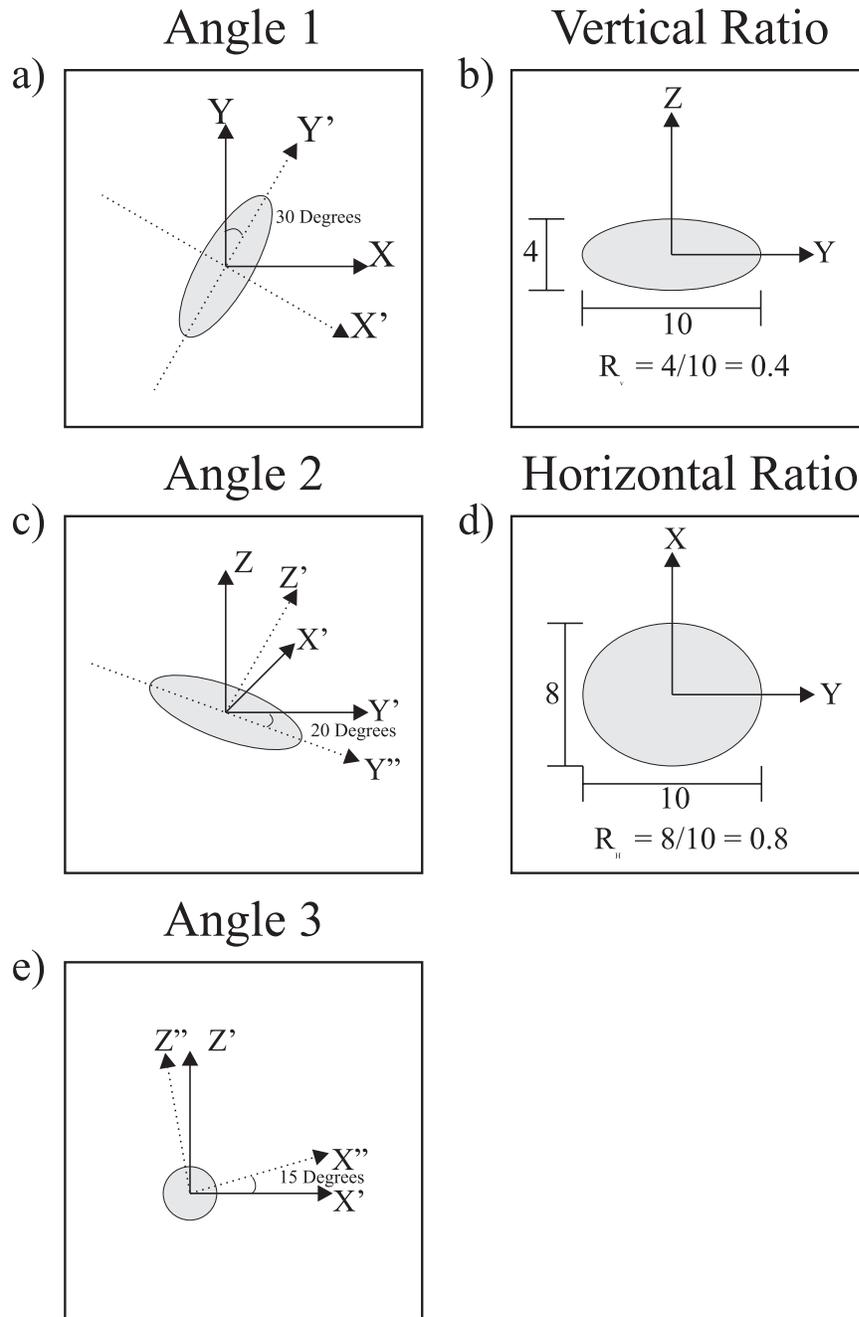


Figure 3.4: Rotation in 3-D require the definition of three rotation angles, plots a), c), and e). The anisotropy is described by two anisotropy ratios, plots b) and d).

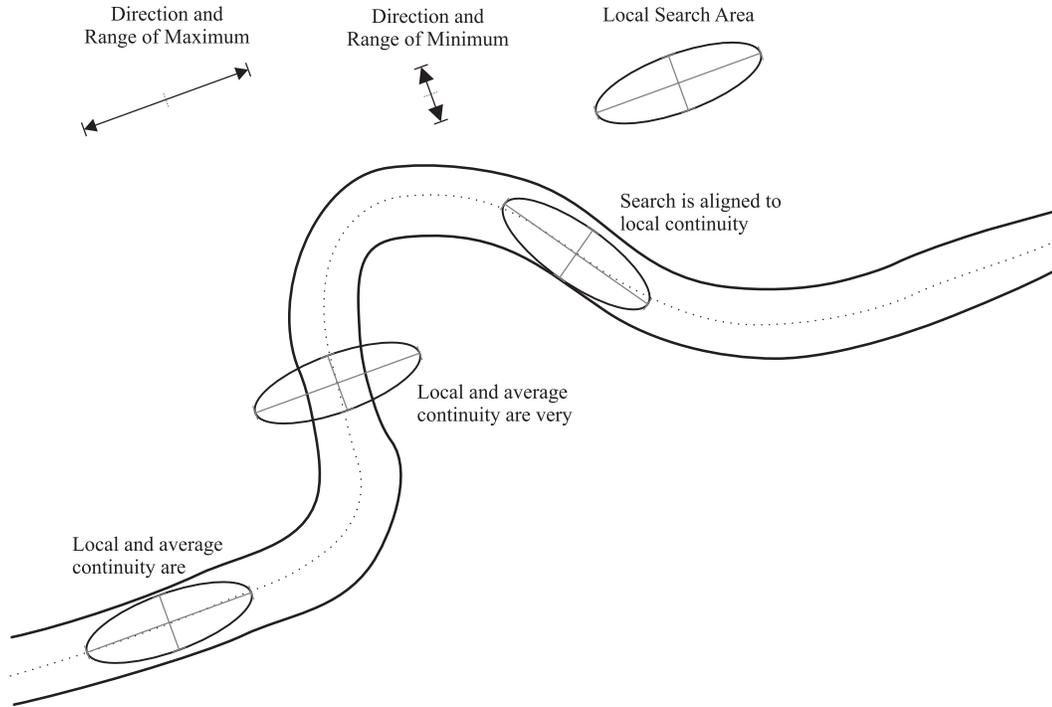


Figure 3.5: Geological structures with complex geometry cannot use average directions of continuity.

and the covariances would be incorrect. To correct this problem, the angles of continuity can vary locally over the modelling area. The changes in the angles can be found deterministically by rock type or hand contouring, or stochastically through surface and/or fault models. The anisotropy ratios are often assumed constant; however, location-dependent anisotropy ratio models could be used by scaling the variogram ranges (Deutsch 1992).

A difficulty in dealing with locally varying angles (LVA) is the calculation of the distance between two locations. The distance measurement should account for the local directions of continuity along the path between two points and not just the Euclidian distance. Difficulty in calculating the separation distance translates to uncertainty in the covariance between any two locations. This uncertainty increases over larger distances and with more complex structure.

To implement LVA, the varying parameters must be modelled at the same scale as the primary variable model. The parameters can change at every location or within regions. The variogram model should account for these angle changes, but due to the uncertainty in the separation distances, calculating the lags would be almost impossible.

Once variogram modelling is completed and the trends in the angles are mapped, simulation can be performed. At an unsampled location, the search ellipsoid and

variogram is oriented to local angles of continuity and anisotropy ratios. Local data are identified inside of the shifted search ellipsoid and the covariances are calculated based on the shifted variogram. The rest of the steps to create a simulated value will be the same as the standard SGS method.

3.5 Rock Types and Soft Boundaries

In a geostatistical study, the variable(s) of interest will typically cover the whole modelling area. In some cases, this area could contain two or more different rock types, all associated with the variable(s) of interest. The problem is that the spatial correlation and histogram can be rock type dependent, globally violating stationarity. To capture these features, the rock types need to be treated separately.

For each rock type, variograms and statistics are calculated based only on the data contained in that rock type. Simulation is performed independently by rock type. Usually, the boundaries between the rock types are *hard*, that is no information can cross them. In some cases, there is correlation between the rock types and this must be accounted for in the modelling process. The boundaries between the correlated rock types will be *soft* or information can pass from one side to the other. There can be two kinds of soft boundaries: one-way and two-way boundaries. In most cases, the correlation between two variables will go both ways. Rock type *A* is correlated to rock type *B* and rock type *B* is correlated to rock type *A*. This will create a two-way boundary where information can move in either direction. An example of a two-way boundary is where there is a transitional zone between the two rock types. Near the boundary, the rock type properties would be similar and a two-way boundary would be practical. Alternatively, a one-way boundary is when rock type *B* can inform *A*, but *A* provides no information to *B*. An example of a one-way boundary is a high grade rock type can use data from an adjacent low grade rock type, but the low grade rock type can only use internal data. If the reverse was allowed, the high grade values would over influence the low grade results.

To control these boundaries, a logic matrix is required where 0 is a hard boundary, no data is used, and 1 is a soft boundary, data can be used (Deutsch 2002b) (Table 3.1). If all the boundaries are set to zero, including the principle diagonal, this rock type will be ignored in the simulation process. This allows the matrix to be used to key out areas of the model.

Each rock type is independently transformed to normal space for use in SGS. This can cause problems when using soft boundaries since a normal value in rock type *A* is not equal to the same value in rock type *B* after back transformation. If the distributions in original space are similar, these effects will be small and this problem can be ignored. If the distributions are different, these effects can cause problems. For example, a gold grade of 10g/tonne may be equal to the 0.68 quantile for rock type *A* and the 0.54 quantile for rock type *B*. This will cause the same value in original space to be unequal in normal space. To correct this, any value crosses a

Can -- be used to estimate rock type -- Yes = 1 No = 0	Rock Type			
	A	B	C	...
A	1	0	0	
B	1	1	0	
C	0	0	1	
⋮		⋮		⋱

Table 3.1: The logic matrix used to describe hard, soft one-way, and soft two-way boundaries. A 0 is a hard boundary and a 1 is a soft boundary.

boundary must first be transformed back into original space and then transformed into normal space specific to the rock type being estimated.

Chapter 4

Self-healing

New forms of SGS were developed to incorporate secondary information. Collocated cokriging (Xu 1992) and simple kriging with a locally varying mean (Section 3.1) are two such techniques that are commonly used in practice due to their simplicity; however, the cost of this simplicity is often an inflation to the global variance. This variance inflation is problematic in the backtransformation to original units, causing the input histogram not to be reproduced. An algorithm was developed to correct the variance inflation problem based on the underlying cause for each technique. This algorithm is referred to as *self-healing* (Zanon 2002) since it is only applied when variance inflation is detected during simulation.

4.1 Collocated Cokriging

Collocated cokriging (CCK) was developed as an alternative to the much more difficult to apply full cokriging (Section 5.1). In CCK, only the collocated secondary variable is used due to a Markov-type screening assumption (Xu 1992). Under this assumption, the collocated secondary data carries the same information as all the surrounding data. This will result in the following equation:

$$y_{CLK}^*(\mathbf{u}) = \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1} \cdot y(\mathbf{u}_{\alpha_1}) + \lambda'_2 \cdot y_2(\mathbf{u}) \quad (4.1)$$

where there are n_1 primary data and λ'_2 is the weight applied to the single collocated secondary variable, $y_2(\mathbf{u})$. To solve for the weights, a $n_1 + 1$ covariance matrix is created. The first n_1 equations in the matrix follow the form:

$$\sum_{\beta_1=1}^{n_1} \lambda_{\alpha_1} \cdot C_{11}(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\beta_1}) + \lambda'_2 \cdot C_{12}(\mathbf{u}_{\alpha_1}, \mathbf{u}) = C_{11}(\mathbf{u}, \mathbf{u}_{\alpha_1}), \quad \alpha = 1, \dots, n_1 \quad (4.2)$$

and the last equation is then:

$$\sum_{\beta_1=1}^{n_1} \lambda_{\alpha_1} C_{21}(\mathbf{u}, \mathbf{u}_{\beta_1}) + C_{22}(0) = C_{12}(0) \quad (4.3)$$

where the $C_{12}(0)$ is equal to $\rho_{12}(0)$ and the $C_{12}(\mathbf{u}_{\alpha_1}, \mathbf{u})$ term is approximated by $\rho_{12}(0) \cdot C_{11}(\mathbf{u}, \mathbf{u}_{\alpha_1})$, $\forall \mathbf{u}, \mathbf{u}_{\alpha_1}$ for normal score variables (Deutsch 2002a). The corresponding kriging variance for CCK will then be:

$$\sigma_{CLK}^2(\mathbf{u}) = C(0) - \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1} \cdot C(\mathbf{u} - \mathbf{u}_{\alpha_1}) - \lambda'_2 \cdot \rho_{12}(0) \quad (4.4)$$

CCK provides the means to incorporate secondary data by only slightly increasing the size of the kriging matrix and with only the knowledge of the correlation coefficient, $\rho_{12}(0)$. The cost of this simplification is that the global variance has a tendency to exceed the expected value of 1.0 in normal space. This causes problems in the backtransformation and poor reproduction of the global histogram. The cause for this global variance inflation is linked to the Markov approximation. Restricting the secondary data to only the collocated variable yields a close approximation to the full cokriging estimate; however, this limitation causes the kriging variance to be larger. In (Equation 4.4), the secondary data acts to reduce the kriging variance. Including more data would decrease the variance even further. This local increase in variance affects every unsampled location in the model area, thus increasing the global variance.

4.2 Locally Varying Mean

Locally varying mean (Section 3.1) is an extension of SK that utilizes secondary data in the form of a geological trend in the local mean. Accounting for this trend can help to capture high and low regions in the model area but this departure from stationarity can cause inflation in the global variance. This increase in the global variance is linked to a relationship between the kriged mean and variance.

Simulation uses kriging to predict an estimate and variance at every unsampled location. Based on these values, a Gaussian distribution is created and the simulated value is drawn. In SK, the expected mean is zero and the highest possible kriging variance is one. The probability of drawing an extreme value from this distribution is very small. The mean can deviate away from zero but this would be driven by nearby local data and the corresponding variance would be small. This will maintain a low probability of drawing an extreme value.

In the case of LVM, departures from the expected value of zero are based on the secondary data and not the local data. Without the supporting local data, the kriging variance will not be reduced and the probability of drawing an extreme value

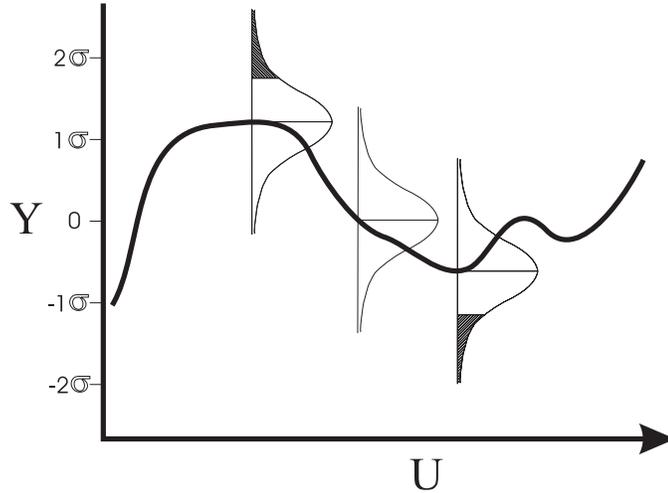


Figure 4.1: Illustration of a locally varying mean (solid line) versus location, \mathbf{u} , and three conditional distributions. The shaded regions of the two outside distributions cause variance inflation because there is a too high probability to draw extreme values.

in simulation is much higher. These extreme values will then be used as conditioning data, compounding the problem. This will result in an inflated global variance. A graphical example of this is shown (Figure 4.1). Three local distributions using the same variance with different mean values can be seen. When the mean deviates away from zero, left and right distributions, a large variance causes the tail on the extreme size to be too large. This same variance would not cause this problem if the mean was close to zero, centre distribution.

4.3 Fixing the Problem

To fix the global variance inflation problem, the kriging variance for both cases must be reduced. The simplest historical fix to this problem is to use a global correction factor to reduce the variance, $\hat{\sigma}_{sk}^2 = f \cdot \sigma_{sk}^2$. This single value is then applied at every location. A global correction fix was problematic in two ways, the correction factor was found iteratively and was applied at every location. Testing a suite of correction factors and choosing the one that closest reproduces a global variance of 1.0 is slow and does not allow for automation of the simulation process. Applying the correction factor everywhere will reduce the local variability in areas that are not a problem.

To overcome these limitations, an algorithm was developed that dynamically corrects the variance when variance inflation is detected. The algorithm in essence allows the program to heal the problem itself, hence the name *self-healing*. To start,

the first twenty locations in the model are simulated without restriction to develop an estimate of the global variance. The global variance is then tracked throughout the simulation process. If a simulated value causes the global variance to exceed the accepted maximum, the kriging variance is corrected and a new estimate is calculated at that location. This algorithm is an improvement in three ways: the degree of correction can change at every location, it is only applied to problem locations, and all calculations are performed internally in the program.

For CCK, the correction factor is calculated based on the global variance. Initially, a value is simulated at a given location and the temporary global variance is calculated. If the temporary variance is greater than 1.02 then self-healing is applied. A value of 1.02 allows for some deviation from the expected variance of 1.00. Values larger than 1.02 were tested, but it was found that self-healing was applied more often and allowed a higher global variance.

When self-healing is applied, a corrected variance is calculated by dividing the temporary global variance by the dispersion variance:

$$\sigma_{cor.}^2 = \frac{\sigma_{sim. global}^2}{D^2(v, A)} \quad (4.5)$$

where the dispersion variance is close to one for point estimates. The corrected variance is then used to find the correction factor that will be applied to the local variance:

$$factor = \frac{1}{e^{\left(\frac{3 \cdot (\sigma_{cor.}^2 - 1)}{ACC}\right)}} \quad (4.6)$$

where ACC is range function that was optimized. This equation is shown graphically (Figure 4.2).

To determine the best value to use for ACC , self-healing was applied to a few test cases. For each case, the global mean, variance, correlation, and variogram were tracked for five realizations as ACC was varied between 0.002 and 0.500. There was little change from 0.002 to 0.010 for all the cases but beyond this value an increase in the variance and/or correlation was observed. For all values of ACC , there was only minor changes in the means and almost no change in the variogram. A final value of 0.01 was chosen for ACC since this will provide the proper degree of correction while limiting the possibility of over-correction.

For LVM, the cause for variance inflation is different and the maximum acceptable kriging variance is calculated based on a theoretical function. A relationship between local mean and variance was approximated using an exponential function. This expression is used to calculate the maximum acceptable variance based on a specified mean:

$$\sigma_{LVM}^2 = e^{-|y(\mathbf{u})| \cdot L} \quad (4.7)$$

where $|y(\mathbf{u})|$ is the absolute value of the kriged mean at location \mathbf{u} , σ_{LVM}^2 is the maximum acceptable variance for the given mean, and L is a positive scaling factor.

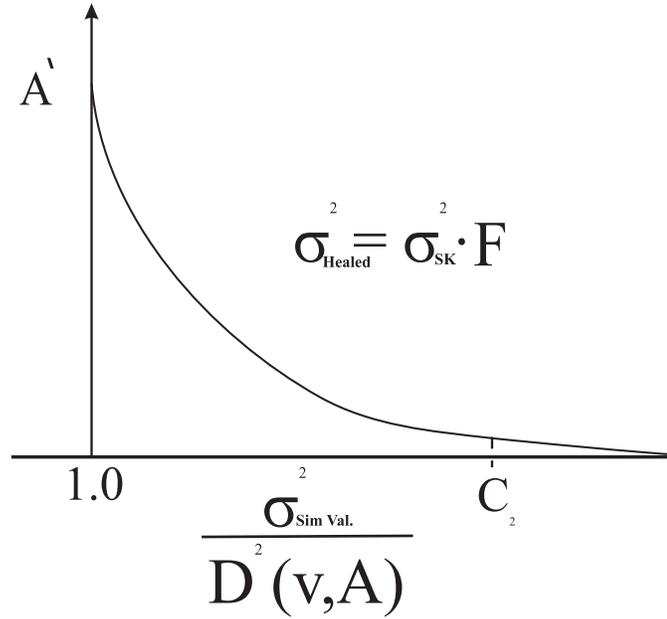


Figure 4.2: The calculation of the self-healing correction factor used in CCK.

Working with the example data used in the next section, SK was performed and the means versus variances were plotted (Figure 4.3). The lower limit of the variance is linked to the variogram used in the example, specifically the nugget effect. The figure shows three sets of lines that correspond to the maximum acceptable variance based on (Equation 4.7) and three different values for L equal to 0.5, 1.0, and 2.0. The $L = 0.5$ line provides the widest limits and allows a much higher variance for non zero means. This line contains most of the points from the example data, but when implemented, the global variance was well above the expected value of 1.0. Increasing L to 1.0 and 2.0 will greatly improve the global statistics. A final value of 1.0 for L provides the desired reduction in the global variance while providing an output correlation closest to the input value.

To apply this calculated variance some specific criteria must be satisfied. To start, the temporary global variance must again exceed 1.02 to even consider self-healing. When a point has been identified as a candidate for self healing, the maximum variance is calculated from (Equation 4.7). If the kriging variance is below the calculated maximum variance, then the kriging variance is retained; otherwise, the maximum acceptable variance is used as the kriging variance. If the kriging estimate is positive, the new kriging variance is applied to the right hand side of the residual distribution. In this way, simulated values drawn from the right half of the distribution will be limited, reducing the probability of drawing extreme values, but the full variability is accepted on the left hand side. This is reversed for negative kriging estimates. An earlier version of self-healing applied the calculated kriging

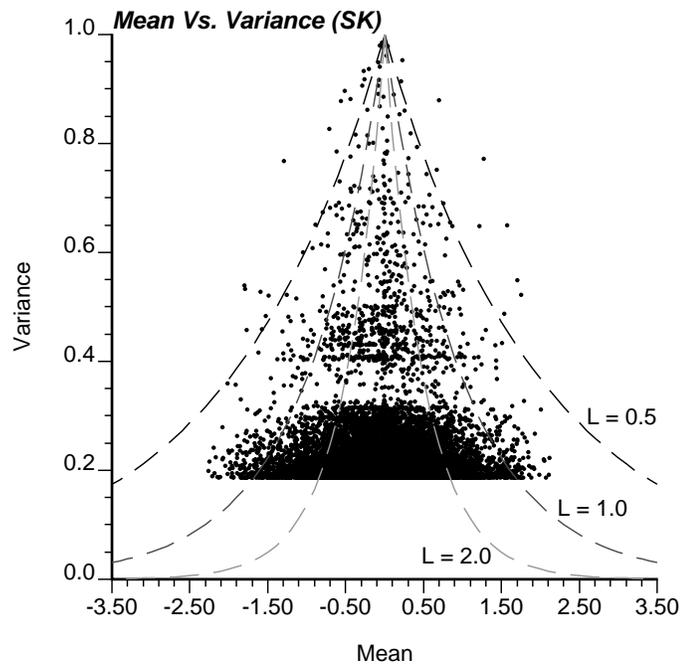


Figure 4.3: The points on this graph show the link between the SK mean and variance for the example data set used later in this chapter. The dashed lines represent the variance limit used for self-healing for different values of L . The mean is in the Gaussian transform of the ordinal units.

variance to the whole distribution, but the model was overly smoothed. Limiting the correction to one side of the distribution allows for more variability and improved variogram reproduction.

Altering the variance will have an affect on the licit model of coregionalization, however, these effects are not clearly understood. To satisfy the licit model of coregionalization, the following equality must be checked:

$$C_{11}(\mathbf{h}) \cdot C_{22}(\mathbf{h}) \geq (C_{12}(\mathbf{h}))^2 \quad (4.8)$$

where $C_{11}(\mathbf{h})$ is the covariance of the primary variable, $C_{22}(\mathbf{h})$ is the covariance of the secondary variable, and $C_{12}(\mathbf{h})$ is the covariance between the variables. For standardized variables this is re-written:

$$\rho_{11}(\mathbf{h}) \cdot \rho_{22}(\mathbf{h}) \geq (\rho_{12}(\mathbf{h}))^2 \quad (4.9)$$

where the covariances have been replaced with correlations. When the correlation varies locally or self-healing is used, this restriction comes into effect. For example, a smooth secondary variable will have a large correlation, ρ_{22} is large, and if this is highly correlated with the primary variable, ρ_{12} is also large, then ρ_{11} must be large to satisfy this requirement. A large correlation for the primary variable will result in a smooth model. The implications of this restriction needs to be looked into further and the resulting models should be checked to ensure that it is not violated.

Self-healing is an effective way to correct the variance inflation caused by CCK and LVM; however, there is no clear definition of when the variance is considered to be inflated. The boundary between acceptable fluctuations in the variance and a problematic variance needs to be studied. This boundary will most likely change depending on the data, setting, and context of the study.

4.4 Example

The porosity of a petroleum reservoir will be used with seismic data as an example for the application of self-healing for CCK and LVM. This reservoir will be modelled using 50m x 50m cells with a range of 4000m in the X direction and 8200m in the Y direction. The isotropic variogram model has a nugget of 0.1 and two spherical structures with a range of 650 and 1000 with corresponding contributions of 0.50 and 0.40, respectively. The variogram sill is at 1.0 to reflect the expected variability of the data. All of the results are for the normalized variables.

The input data for this example consists of 100 primary data and a map of secondary data at the same scale as the model (Figure 4.4 a,b). Note that the variability of the secondary variable may be larger than usual. One realization for CCK was produced using no self-healing, c), and with self-healing d). The self-healing map shows a decrease in the solid black and white areas, corresponds to a decrease in extreme values. This assessment is reflected in the difference map where

the no self-healing results were subtracted from the self-healing results, e). The difference map shows negative values, white, in the high value areas and positive values, black, in the low value areas confirming the reduction in extreme values.

Self-healing affects the global distribution of the data, as seen by the histogram (Figure 4.5). Histograms were created for two realizations of CCK without self healing, a) and c), and with self-healing, b) and d). When self-healing is not used, the histograms show an increase in the spread of the distributions and variances are close to 1.50. Self-healing reduces this spread, along with decreasing the variance to 1.00. All of the histograms have means of approximately zero.

To look at how the correlation was reproduced, scatter plots between the simulated results and secondary data were created (Figure 4.6). The two realizations using no self-healing, a) and c), show an increase in correlation from the input value of 0.77 to 0.913 and 0.908, respectively. This is reflected in the narrow range of points on the scatter plots. Self-healing, b) and d), did not correct this increase in correlation, but there was some improvement at 0.874 and 0.864, respectively. The significant of this improvement is not clear.

The final comparison is a look at the variogram reproduction for five realizations, dotted lines, against the input variogram, solid line (Figure 4.7). The results for CCK without self healing were plotted in the X direction, a), and the Y direction, b). In both directions, the variogram reproduction is fairly good until the input variogram range is reached at 1000m. The variograms continue to increase beyond the expected sill of 1.0 and reach maximum variability at over 1.50 with a range of about 1600m. Self-healing, b) and d), show reproduction of the nugget in both directions, but as the distance increases, there is significant deviation away from the input variogram. The increase to the range of continuity causes the variogram to reach the expected sill at about 1500m and this corresponds to a maximum variability of close to 1.0. This smoothing effect is the cost of applying the self-healing algorithm. It is interesting to note that the self-healing variogram range is similar to the range of maximum continuity when self-healing is not used.

The same data set was used to show the changes to LVM when self-healing is applied (Figure 4.8a,b). The LVM realization without self-healing, c), shows an increase in extreme values compared to the self-healing realization, d). Subtracting the no self-healing results from the self-healing results provides the difference map for LVM, e). This map is much smoother than the CCK difference map, but it again shows that the extreme values have been reduced by using self-healing.

The histograms for all realizations show some deviation from the normal shape, but this deviation is smaller in the self-healing results (Figure 4.9b,c). For both realizations, the variance is at 1.02. The histograms when self-healing is not used, a) and c), have much larger variances exceeding 1.60. This is reflected in the spread of the histograms and the increased number of extreme values. The mean of the histogram are all close to zero. In (Figure 4.9c), the last bin on the right is the collection of all values above the maximum of 3.50.

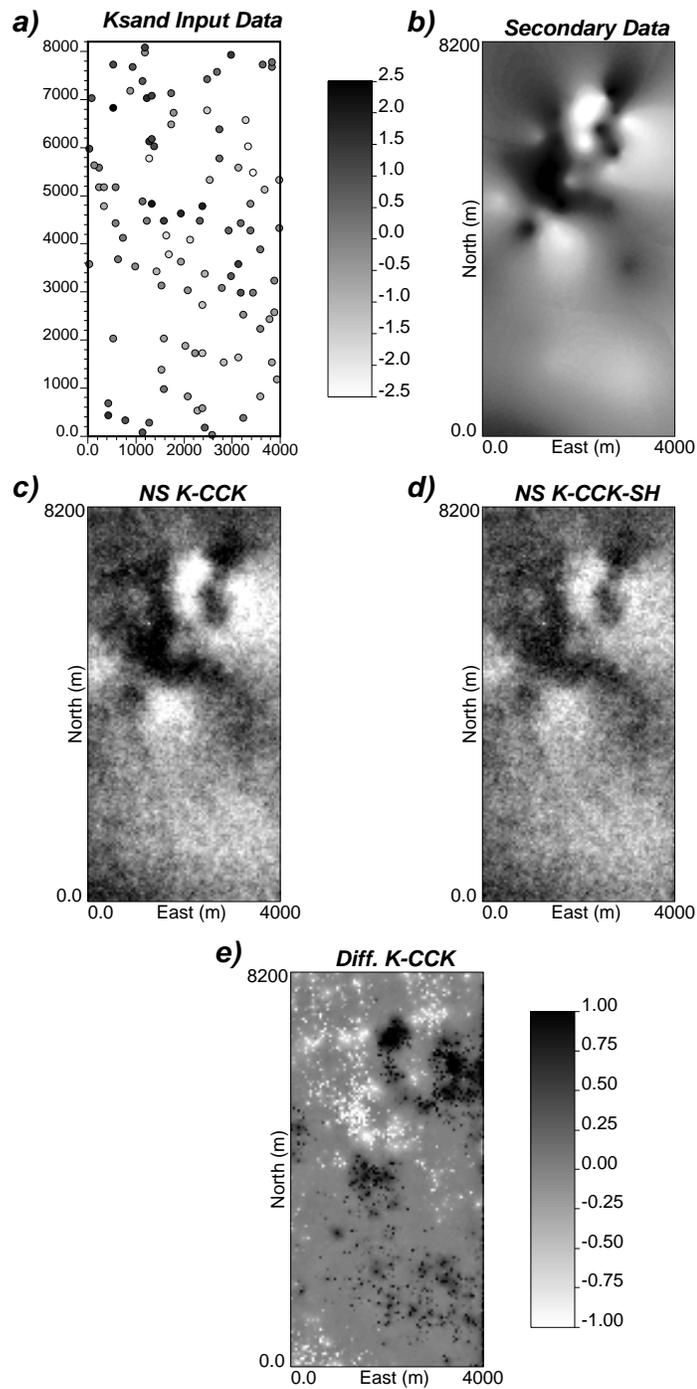


Figure 4.4: The input data, a), and secondary data, b), were used to create CCK models without self-healing, c), and with self-healing, d). A difference map, e), provides a comparison between the models. The plots show the Gaussian transform of porosity.

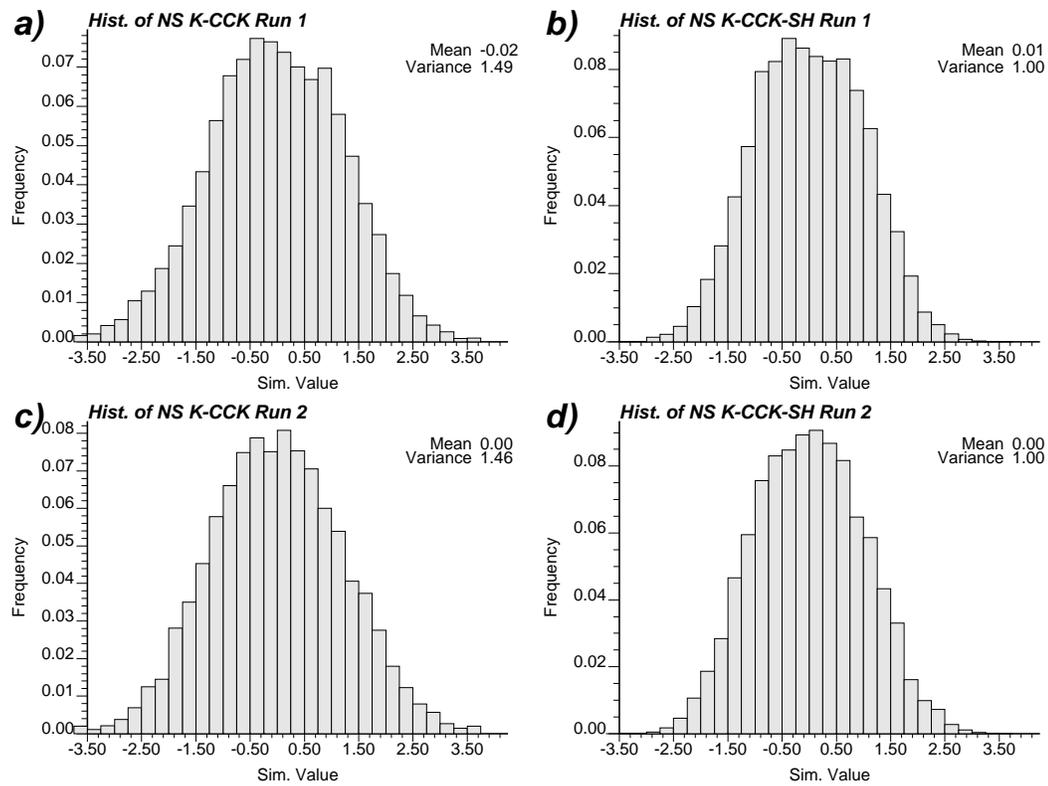


Figure 4.5: Histograms for two realizations of CCK without self-healing, a) and c), and with self-healing, b) and d). The histograms are in the Gaussian transform of porosity.

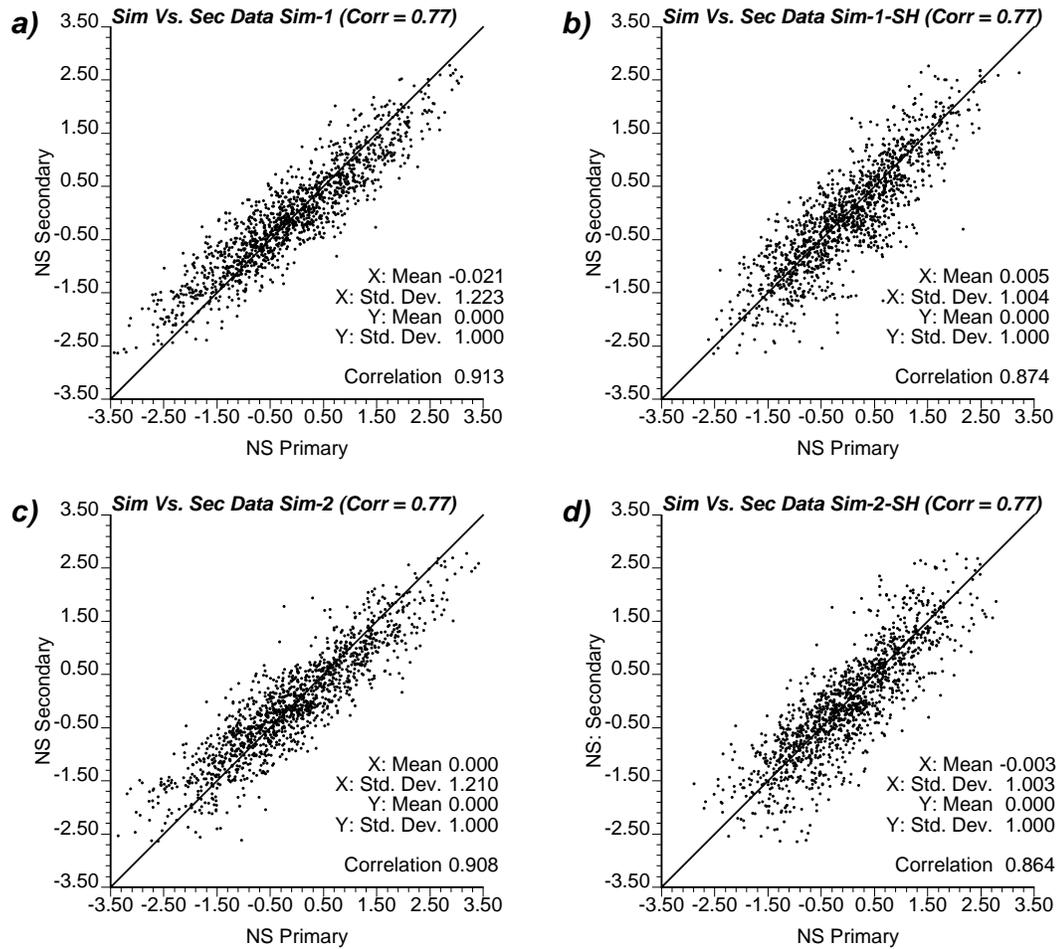


Figure 4.6: Scatter plots for two realizations of CCK without self-healing, a) and c), and and with self-healing, b) and d). The primary and secondary variables are Gaussian transforms of porosity.

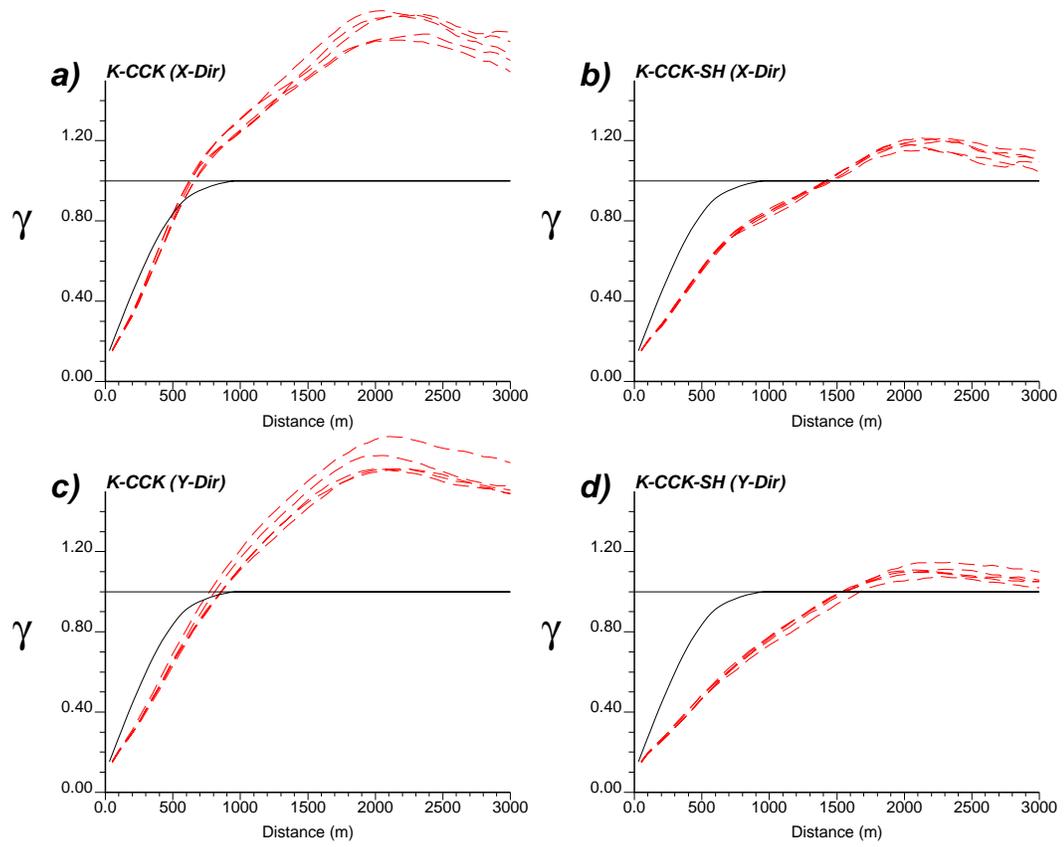


Figure 4.7: A look at variogram reproduction for 5 realizations, dotted lines, of CCK without self-healing, a) and c), and with self-healing, b) and d). The solid line is the input variogram.

The scatter plots (Figure 4.10) show the relationship between the kriged values and local means. The realizations when self-healing is not used, a) and c), show more spread in the data on the x axis compared the self healing results, b) and d). This change in spread is caused by the decrease in extreme values. Self-healing does show some banding the the scatter plots. The cause for this banding is most likely linked to the calculation of the maximum acceptable variance, but the relationship is not clear.

The variogram reproduction (Figure 4.11) for LVM shows different results from that of CCK. The variograms for LVM without self-healing, a) and c), have the correct nugget but they increase much faster than the model variogram, reaching a variability of 1.0 in about 400m. The variograms continue to increase until they reach maximum variability at over 1.5. The self-healing results, b) and d), reproduce the nugget and input variogram up to a variability of about 0.5. Above 0.5, the variograms do not increase as quickly as the model variogram and maximum variability is not reached until about 1200m. The self-healing variograms show that the global variance has been reduced, while maintaining partial variogram reproduction.

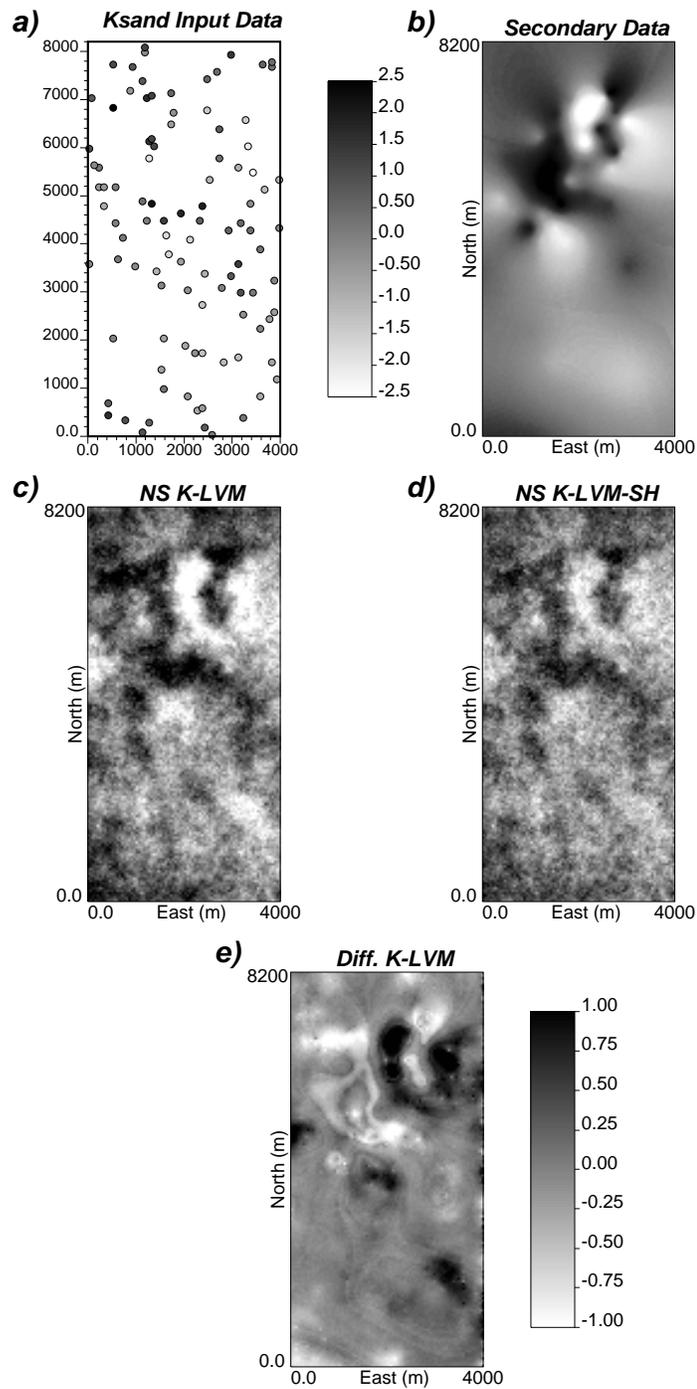


Figure 4.8: The input data, a), and secondary data, b), were used to create LVM models without self-healing, c), and with self-healing, d). A difference map, e), provides a comparison between the models. The plots show the Gaussian transform of porosity.

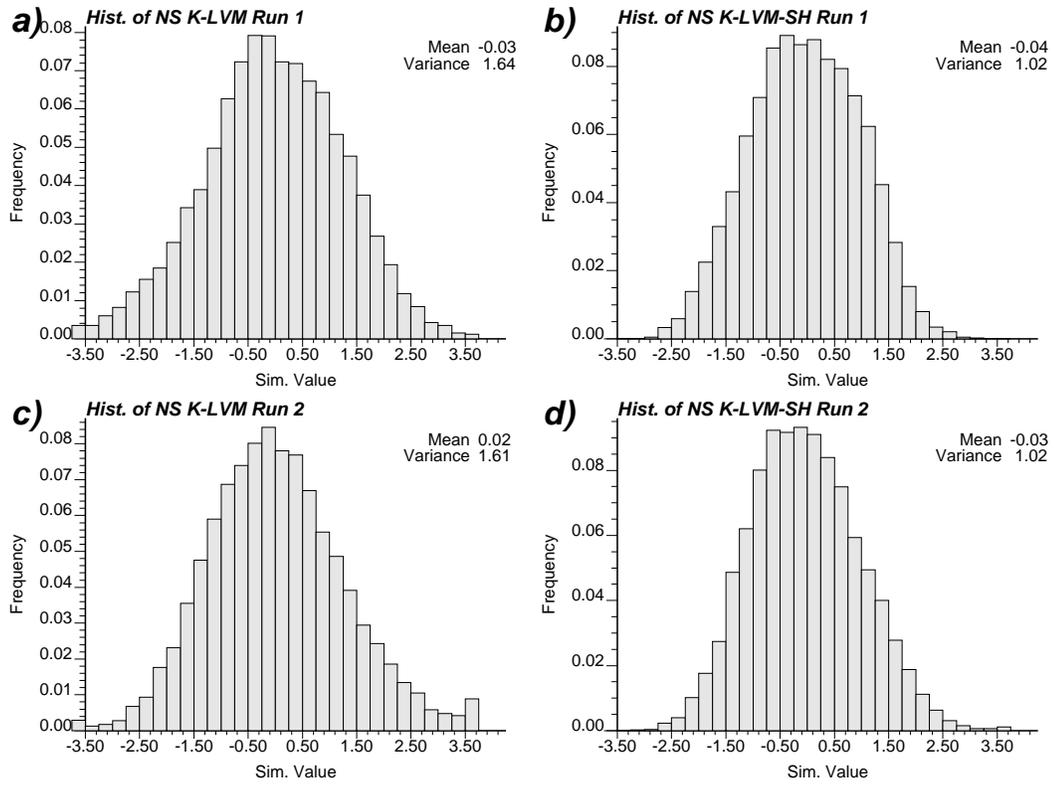


Figure 4.9: Histograms for two realizations of LVM without self-healing, a) and c), and with self-healing, b) and d). The histograms are in the Gaussian transform of porosity.

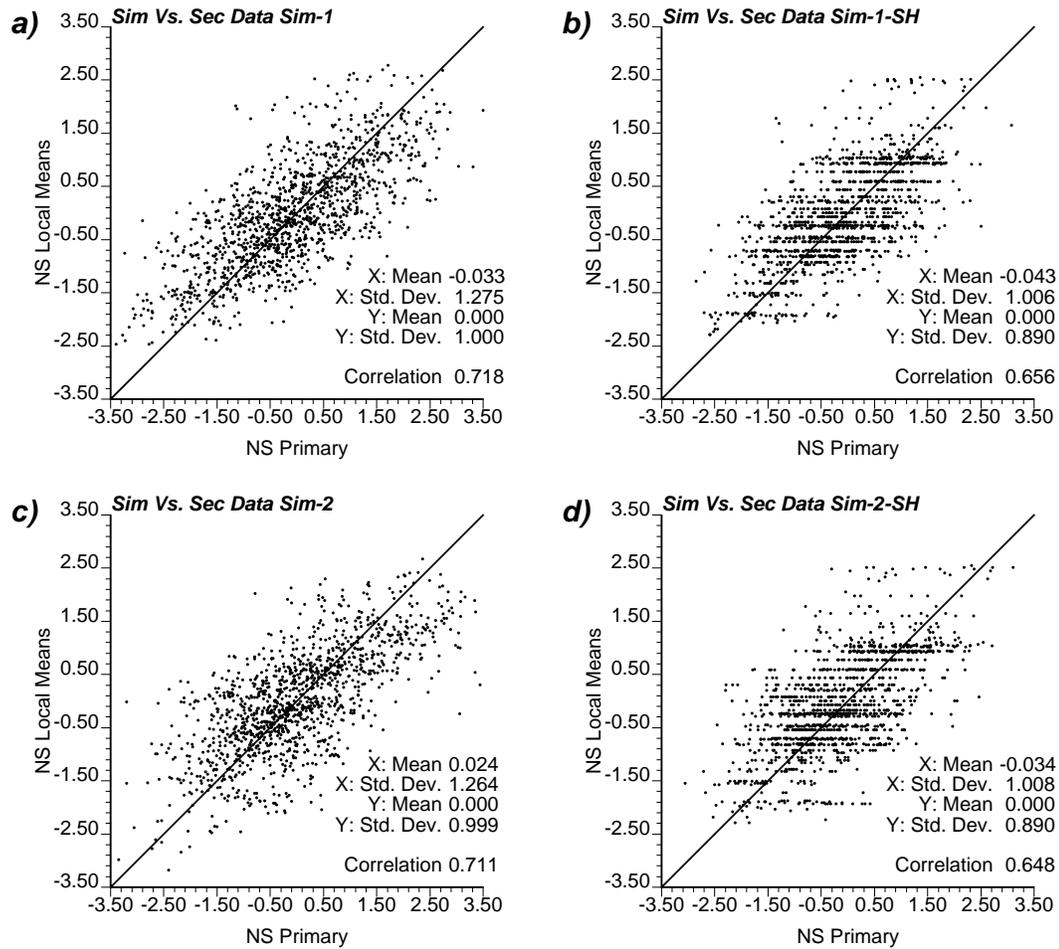


Figure 4.10: Scatter plots for two realizations of LVM without self-healing, a) and c), and and with self-healing, b) and d). The primary and secondary variables are Gaussian transforms of porosity.

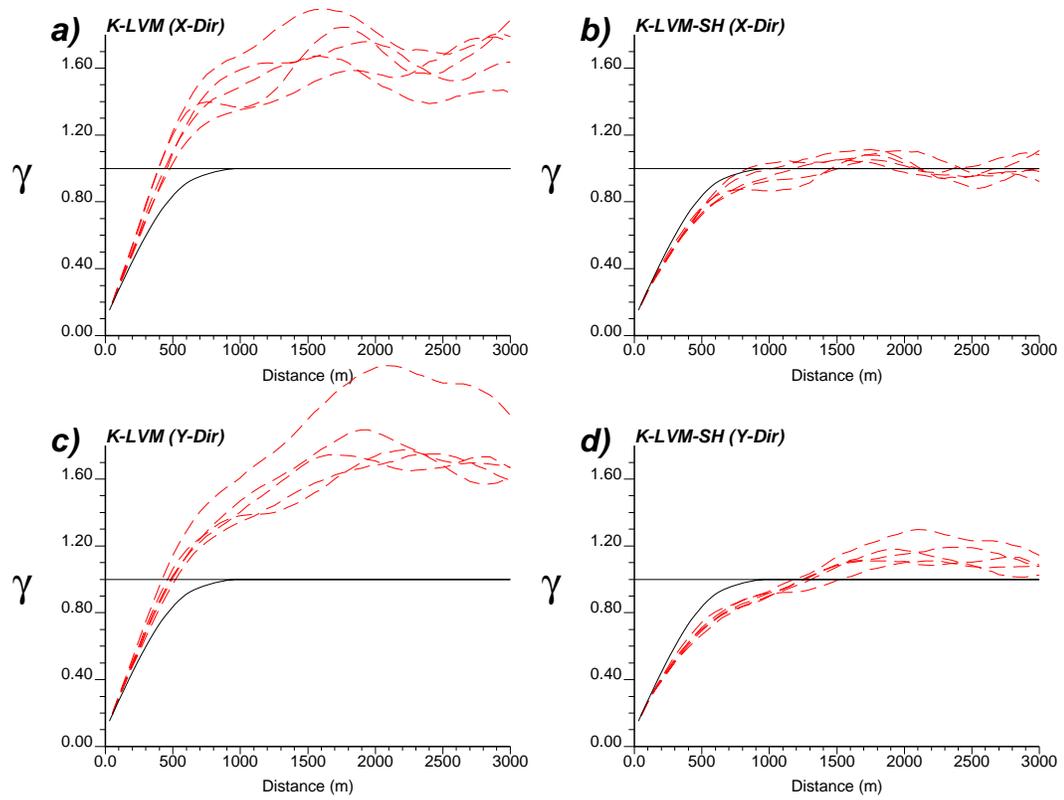


Figure 4.11: A look at variogram reproduction for 5 realizations, dotted lines, of LVM without self-healing, a) and c), and with self-healing, b) and d). The solid line is the input variogram.

Chapter 5

Alternative Secondary Data Integration

Collocated cokriging and locally varying mean suffer from variance inflation problems due to simplifications in their theory. There are alternative simulation techniques that can be applied that do not have these same limitations. Full cokriging and stepwise conditional transformation are two such techniques (Leuangthong 2003a, 2002b). Full cokriging is theoretically correct in a multiGaussian setting. It utilizes all local secondary data in the kriging estimate and this improves variance reproduction. Stepwise transformation accounts for the correlation between the variables in the transformation process. The transformation leads to variables that can be simulated independently and their correlation will be injected in the backtransformation process. These techniques can improve on the previously presented techniques, however, they require the user to perform more work and do have some limitations.

5.1 Full Cokriging

Full cokriging is an extension of the kriging formalism (Equation 1.2) where secondary information is included in the estimate. The full cokriging equation is expressed using residuals, Y_i , where the mean, m_i , specific to each variable, is subtracted from all the data in variables k . In this form, Y_1 is the primary variable and $Y_{k>1}$ is any number of secondary variables:

$$y_{FC}^*(\mathbf{u}) = \sum_{i=1}^k \sum_{\alpha_i=1}^{n_i} \lambda_{\alpha_i} \cdot y_i(\mathbf{u}_{\alpha_i}) \quad (5.1)$$

where all k variables can have a different number of data, n_i , at different location, \mathbf{u}_{α_i} , and each variable will have its own set of weights, λ_{α_i} , to help predict the full cokriging estimate, $y_{FC}^*(\mathbf{u})$.

Kriging amounts to minimizing the error variance of the linear estimate (Equation 5.1). This is done by taking the partial derivatives with respect to the different kriging weights and setting them to zero (Section 1.3). This leads to the cokriging equations:

$$\sum_{i=1}^k \sum_{\beta_i=1}^{n_i} \lambda_{\beta_i} \cdot C_{ij}(\mathbf{u}_{\alpha_j}, \mathbf{u}_{\beta_i}) = C_{j1}(\mathbf{u}, \mathbf{u}_{\alpha_j}), \quad \alpha_j = 1, \dots, n_j, \quad j = 1, \dots, k \quad (5.2)$$

where $C_{ij}(\mathbf{u}_{\alpha_j}, \mathbf{u}_{\beta_i})$ is the covariance when $i = j$ and the cross-covariance when $i \neq j$. The corresponding full cokriging variance is expressed:

$$\sigma_{FC}^2 = C(0) - \sum_{i=1}^k \sum_{\alpha_i=1}^{n_i} \lambda_{\alpha_i} \cdot C_{i1}(\mathbf{u}, \mathbf{u}_{\alpha_i}) \quad (5.3)$$

where $C_{i1}(\mathbf{u}, \mathbf{u}_{\alpha_i})$ is the covariance for $i = 1$ or the cross-covariance, when $i > 1$, between the primary variable at the point being estimated and any other variable at distance \mathbf{h} away.

To obtain the covariances needed to solve these equations, the (cross)variograms for each variable and between every variable must be modelled. A requirement for this modelling process is that the $(n_1 + n_2 + \dots + n_k)^2$ matrix must be (semi)positive definite (Journal 1978, p35) and to ensure this all the variograms must be modelled using a linear model of coregionalization (LMC) (Goovaerts 1997, p208). This requires the joint modelling of $\frac{k \cdot (k+1)}{2}$ (cross)variograms if $C_{\alpha\beta}(\mathbf{h}) = C_{\beta\alpha}(\mathbf{h})$ is assumed. This is done by following the rules outlines in (Deutsch 2002a, p144) or (Goovaerts 1997, p112). Due to the difficulty in obtaining these (cross)variograms and the size of the resulting matrix, (Equation 5.1) is typically only expressed for a single secondary variable:

$$y_{FC}^*(\mathbf{u}) = \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1} \cdot y_1(\mathbf{u}_{\alpha_1}) + \sum_{\alpha_2=1}^{n_2} \lambda_{\alpha_2} \cdot y_2(\mathbf{u}_{\alpha_2}) \quad (5.4)$$

In this form, only three covariance functions must be known.

$$\begin{aligned} C_{11}(\mathbf{h}) &= Cov\{Y_1(\mathbf{u}), Y_1(\mathbf{u} + \mathbf{h})\} \\ C_{12}(\mathbf{h}) &= Cov\{Y_1(\mathbf{u}), Y_2(\mathbf{u} + \mathbf{h})\} \\ C_{22}(\mathbf{h}) &= Cov\{Y_2(\mathbf{u}), Y_2(\mathbf{u} + \mathbf{h})\} \end{aligned}$$

where $C_{12}(\mathbf{h})$ is assumed identical to $C_{21}(\mathbf{h})$ and the covariances are defined between any two data located at \mathbf{u} and $\mathbf{u} + \mathbf{h}$.

Working with only a single secondary variable reduces (Equation 5.2) to only $(n_1 + n_2)$ equations. The first n_1 equations will follow the form:

$$\sum_{\beta_1=1}^{n_1} \lambda_{\beta_1} \cdot C(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\beta_1}) + \sum_{\beta_2=1}^{n_2} \lambda_{\beta_2} \cdot C(\mathbf{u}_{\alpha_1}, \mathbf{u}_{\beta_2}) = C(\mathbf{u}, \mathbf{u}_{\alpha_1}), \quad \alpha_1 = 1, \dots, n_1 \quad (5.5)$$

and the final n_2 equations follow the form:

$$\sum_{\beta_1=1}^{n_1} \lambda_{\beta_1} \cdot C(\mathbf{u}_{\alpha_2}, \mathbf{u}_{\beta_1}) + \sum_{\beta_2}^{n_2} \lambda_{\beta_2} \cdot C(\mathbf{u}_{\alpha_2}, \mathbf{u}_{\beta_2}) = C(\mathbf{u}, \mathbf{u}_{\alpha_2}), \quad \alpha_2 = 1, \dots, n_2 \quad (5.6)$$

Full cokriging has several advantages over the simplified alternatives, but there are some significant difficulties associated with its implementation that prevent its wide spread use. There is a high demand on the number of data required to properly model the LMC and for k variables, $\frac{k \cdot (k+1)}{2}$ variograms must be modelled. Automatic variogram fitting programs can help to alleviate the difficulties and time requirements associated with manually fitting the LMC for all variograms. Also, full cokriging can suffer from matrix instability when many highly redundant secondary data are used (Xu 1992). The above reasons have caused most users to lean towards the simplified alternatives when incorporating secondary data.

The full cokriging option will more closely match the global histogram than CCK (Section 4.1) since the variance at each location will be smaller by incorporating more secondary data. A variable dependent search is used to define the local neighbourhood and any secondary data, gridded or non-gridded, can be used in the estimate. There is no requirement that a secondary variable covers the whole modelling area.

5.2 Stepwise Conditional Transformation

An inherent assumption in performing SGS with secondary data is that all the variables are multivariate Gaussian (Deutsch 1998). Conventionally, all variables are independently transformed through normal scores transformation (Section 2.1). This will create univariate Gaussian distributions, but there is no control over the higher order multivariate distributions. Stepwise conditional transformation (Leuangthong 2003a, 2002b) is one way to ensure multivariate Gaussianity of collocated variables in the transformation process.

Stepwise conditional transformation follows the same procedure as normal score transformation, but is done in a hierarchical order so that each transformation is conditional to the transformations already performed. For k variables the k^{th} transformation is conditioned by the previous $k - 1$ transformations:

$$\begin{aligned} y'_1 &= G^{-1}[F_1(z_1)] \\ y'_2 &= G^{-1}[F_{2|1}(z_2|z_1)] \\ &\vdots \\ y'_k &= G^{-1}[F_{k|1,\dots,k-1}(z_k|z_1, \dots, z_{k-1})] \end{aligned}$$

where $y'_i, i = 1, \dots, k$ are independent multivariate Gaussian distributions and the (\prime) indicates that stepwise transformation was used. Note that the transformation of the first variable is identical to normal score transformation.

The resulting distributions of the k variables will be multivariate Gaussian with correlations of zero at a lag distance of zero, $\mathbf{h} = 0$. Since the variables are independent, each variable can be simulated independently using SGS. This removes the need to model or assume an LMC. The model of coregionalization is implicit in the stepwise transformation (Leuangthong 2003b). Once independent realizations have been created for all k variables, conditional backtransformation is performed.

The ordering of the variables will have an affect on the final results. Since the first transformation is not conditioned to any other variable, simulated results will be the same regardless of what transformation is used. Later variables are conditioned to the previously transformed variables and this will cause changes in the resulting statistics and variogram reproduction. To limit these effects, the most continuous variable should be the primary variable, followed by the second most continuous, and so on (Leuangthong 2003b).

A limitation of this method, and virtually any method, is that sparse data can cause erratic and nonrepresentative distributions. It has been found that for k variables there should be 10^k to 20^k data to properly inform all distributions. The data must be isotopically sampled to infer the relationships between all of the variables. This will permit each distribution to be discretized into 10-20 classes with 10-20 data in each class. If there are fewer than this number of data, smoothing algorithms can supplement the data by filling in gaps in the distributions (Leuangthong 2003b).

Chapter 6

Final Comments

This thesis has looked into several application details associated with SGS. These details were presented to explain the SGS algorithm and the reasoning behind some of the conventional choices. This chapter provides a summary of the main points from the thesis and areas of future research.

6.1 Conclusions

The transformation of SGS from theory to application requires implementation details and simplifications to make the process practical. Once the data has been transformed into normal space, the data can stay at their original locations or be assigned to the grid nodes. Assigning the data to the grid nodes allows data reproduction in the final model with only a minor change to the covariances; provided the grid spacing is much smaller than the data spacing. Searching for local data is performed using the efficient spiral search; however, input data and previously simulated values are treated equally. Keeping the input data at their original locations allows the input data to be used preferentially over previous simulated values. Searching for conditioning data is performed in two steps. The spiral search is used to locate previously simulated gridded data and the super block search locates non-gridded input data. Performing two searches increases the CPU cost.

Working under the Markov screening assumption allows a limited number of conditioning data to be used. The choice of this number is very important, but several factors must be balanced. To achieve reasonably accurate kriging estimates and variances, more than 8 to 10 data should be used. This data should come from all directions to achieve a representative sample of the surrounding area. The octant search will help to achieve this when samples are aligned along drillholes. Working with more than 10 data will cause only minor decreases in the kriging variance since the kriging weights will approach zero for the more distant data. The CPU requirements will increase rapidly as more data is used. The extra cost in CPU time may be required if complex variogram features are present in the model. To fully

capture zonal anisotropy and periodicity, more data are required. This is reflected in the long-range variogram reproduction and is directly related to the more distant data even though they receive small kriging weights.

The order in which the unsampled locations are visited can have an effect on the final SGS model. The regular and spiral search paths were looked at due to their perceived improvement to reduce CPU time or extend the influence of the input data. Both of these techniques were rejected due to the poor reproduction of the variogram. Practice has shown that the random path is the best choice since it does not influence the model over multiple realizations. A fully random path may poorly reproduce the long-range variogram when only nearby data are used for conditioning. To improve variogram reproduction, a multiple grid search is used in conjunction with the random path. The simulation path starts on a coarse grid that is much larger than the model resolution. This grid is then refined in several increments until the specified resolution has been reached. The path for each grid is still defined randomly. The multiple grid search retains all of the benefits of the random path while improving the long-range variogram reproduction.

One complication to SGS is the presence of non-stationary parameters. The most common of these is a location-dependant mean. To apply a locally varying mean, it must be defined everywhere and at the same scale as the final model. SGS allows the mean to change with location, but practice has shown variance inflation can occur in the global variance.

The variance can change locally as well. Variability in the variance can be related to the heteroscedasticity of the data or simply due to regional fluctuations. Heteroscedasticity is typically removed through normal score transformation, but sometimes a residual amount is left in the data. To correct this, the data is divided into groups or bins based on their mean in normal space. The variance in each bin is calculated and correction factors are determined by dividing the bin variance by the global variance. During simulation, the kriging estimate at every location is matched to a bin and the corresponding correction factor is applied to the kriging variance. Alternatively, regional changes in variability are accounted for by constructing a trend map. From this map, the location-dependant correction factor is calculated.

Incorporating secondary data into SGS requires a different type of kriging. Collocated cokriging is the easiest to apply since it only requires the collocated correlation between the two variables. This correlation can vary with location in specific geological settings. Identifying the geological setting where this variability can occur is important. A correlation trend map could be constructed and at every location the corresponding correlation is read from the trend map and applied to the CCK system.

Additional non-stationarity parameters include the local ranges and directions of continuity. Folding and faulting can distort these parameters so that they change locally. Coordinate transformation can reverse these effects but complex 3-D cases are difficult to unravel. The angles describing local continuity must be calculated

at each location. This can be extended to changes in the anisotropy ratios as well. These local parameters can then be applied during simulation. The most difficult part of this process is the construction of the variogram.

In the presence of multiple rock types, separating the data by rock type can help to avoid some non stationary parameters. SGS can be performed for each rock type independently. This requires variograms modelling by rock type. In some cases, the rock types can be correlated and conditioning data is allowed to cross soft rock type boundaries. Since each rock type can have a different distribution in original space, data crossing soft boundaries must be transformed into original units and then back into normal space specific to the rock type being simulated. Soft boundaries can allow data to travel in one or two directions.

When secondary information is used in SGS, under CCK or LVM, inflation in the global variance can occur. To correct this problem, an algorithm was developed that dynamically changes the kriging variances when inflation in the global variance is detected. This algorithm is called self-healing. During simulation, the global variance is tracked. When a location being simulated causes this variance to exceed a specified limit, self-healing is applied. For CCK, a correction factor is calculated based on the global variance and this factor is applied to the kriging variance. The location is then re-simulated using the corrected kriging variance. The LVM approach is to determine the maximum acceptable variance based on a function. This function uses the kriging estimate to calculate the maximum variance. The location causing variance inflation are re-simulated, where the calculated variance is only applied to the problem side of the residual distribution. Both of these techniques have been effective in controlling the global variance. An improved global variance translates to improved histogram reproduction; however, the variogram will typically be reproduced worse.

Alternative techniques to CCK and LVM are available that do not suffer from the same variance inflation problems. Full cokriging does not limit the secondary data to only the collocated point. The increased number of secondary data will reduce the kriging variance, but this requires the modelling of all variograms and cross-variograms. The variogram modelling process can be made easier with an automatic fitting program. Stepwise conditional transformation accounts for the correlation between multiple variables in the transformation process. After transformation, these variables will be independent and can be simulated separately. The correlation between the variables is re-injected in the backtransformation step. To properly account for the relationships between the variables, at least 10^k data must be present for k variables. The data must be sampled at the same locations.

6.2 Future Work

This thesis has considered some of the details associated with SGS, but more work is required in several areas. When working with non stationary parameters, trend

modelling is used to understand how the parameters change with location. The methods used to perform trend modelling are not well understood. More flexible techniques may need to be developed with an understanding of how uncertainty in the trend model affects the simulation process.

The research into non stationary parameters has only just started. An understanding of their full flexibility is required. Specifically, LVA is limited by the quality of the variogram used for simulation. Improved variogram modelling methods specific to LVA must be looked at to help inject the proper structure into the models.

Variogram reproduction is another area where more work is required. Applying self-healing to CCK leads to poor variogram reproduction. Modifications to the self-healing algorithm may provide the required variance reduction without such a large impact on the variogram.

For the maximum number of conditioning data, it has been shown that once more than 10 data are used, there is little change in the kriging mean and variance. This assumes the data is representative of the surrounding area. This minimal change is due to the small kriging weights assigned to the more distant and/or the redundancy in the data; however, these small weights appear to have an impact on the variogram reproduction. Understanding the importance of these small weights and determining how many data are required for adequate variogram reproduction is another area of future research.

The application of Gandin's bounds on the kriging variance needs to be considered. Determining where these bounding limits can be applied is important to justify the computational effort required to obtain them.

Finally, full cokriging and stepwise conditional transformation have been presented as alternative techniques to CCK and LVM. A comparative study is required to show where each of these methods is best applied and their (dis)advantages. This study should include the application of self-healing to correct for variance inflation problems in CCK and LVM. Detailed guidelines on where and when to apply these techniques would be beneficial to most practitioners of geostatistics.

References

- [1] J. Chilès and P. Delfiner. *Geostatistics - Modeling Spatial Uncertainty*. A Wiley Interscience Publication, New York, 1999.
- [2] C. V. Deutsch. *Geostatistical Reservoir Modeling*. Oxford University Press, New York, 2002.
- [3] C. V. Deutsch and A. G. Journel. *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, New York, second edition, 1998.
- [4] C. V. Deutsch and R. W. Lewis. Advances in the practical implementation of indicator geostatistics. In *Proceedings of the 23rd International APCOM Symposium*, pages 133–148, Tucson, AZ, April 1997. Society of Mining Engineers, Oxford University Press.
- [5] C. V. Deutsch and S. D. J. Zanon. UltimateSGSIM: Non-stationary sequential gaussian cosimulation by rock type. In *Centre for Computational Geostatistics Report Four*, Edmonton, Alberta, March 2002. University of Alberta.
- [6] L. S. Gandin. *Objective Analysis of Meteorological Fields*. Leningrad, 1963. Translated from Russian: Israel Program for Scientific Translations, Jerusalem, Israel, 1965.
- [7] P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.
- [8] E. H. Isaaks. *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*. PhD thesis, Stanford University, Stanford, CA, USA, October 1990.
- [9] E. H. Isaaks and R. M. Srivastava. *An Introduction to Applied Geostatistics*. Oxford University Press, New York, 1989.
- [10] A. G. Journel. *Fundamentals of Geostatistics in Five Lessons*. Volume 8 Short Course in Geology. American Geophysical Union, Washington, DC, 1989.
- [11] A. G. Journel and Ch. J. Huijbregts. *Mining Geostatistics*. Academic Press, New York, 1978.

- [12] O. Leuangthong. *Stepwise Conditional Transformation for Multivariate Geostatistical Simulation*. PhD thesis, University of Alberta, Edmonton, AB, Canada, June 2003.
- [13] O. Leuangthong and C. V. Deutsch. Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, 35:155–173, 2003.
- [14] G. Lyall and C. V. Deutsch. Geostatistical modeling of multiple variables in presence of complex trends and mineralogical constraints. In *Geostatistics 2000 Cape Town - Proceedings of the 6th International Geostatistics Congress*, Cape Town, South Africa, April 2000. Geostatistical Association of South Africa.
- [15] J. A. McLennan. The effect of the simulation path in sequential gaussian simulation. In *Centre for Computational Geostatistics Report Four*, Edmonton, Alberta, March 2002. University of Alberta.
- [16] R. A. Olea, editor. *Geostatistical Glossary and Multilingual Dictionary*. Oxford University Press, New York, 1991.
- [17] B. Oz and C. V. Deutsch. A short note on the proportional effect and direct sequential simulation. In *Centre for Computational Geostatistics Report Four*, Edmonton, Alberta, March 2002. University of Alberta.
- [18] R. L. Scheaffer and J. T. McClave. *Probability and Statistics for Engineers*. Duxbury Press, Belmont, California, fourth edition, 1995.
- [19] T. T. Tran. Improving variogram reproduction on dense simulation grids. *Computers and Geosciences*, 20(7/8):1161–1168, 1994.
- [20] W. Xu, T. T. Tran, R. M. Srivastava, and A. G. Journel. Integrating seismic data in reservoir modeling: The collocated cokriging alternative. *Society of Petroleum Engineers*, 1992. SPE 24742.
- [21] S. Zanon, T. Faechner, and C. V. Deutsch. Improved integration of secondary data using self-healing sequential Gaussian simulation. In *30th International Symposium on Computer Applications in the Mineral Industries (APCOM)*, Phoenix, Arizona, February 2002. SME.