University of Alberta

Uncertainty in the Global Mean for Improved Geostatistical Modeling

by

Martha Emelly Villalba Matamoros

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

Winter 2011
Edmonton, Alberta

Examining Committee

Dr. Clayton V. Deutsch (Supervisor), Civil and Environmental Engineering

Dr. Jeffery B. Boisvert (Chair and Examiner), Civil and Environmental Engineering

Dr. Peng Zhang (Examiner), Math and Statistical Sciences

To my father, Jose Daniel

Abstract

Analysis of uncertainty in ore reserves impacts investment decisions, mine planning and sampling. Uncertainty is evaluated by geostatistical simulation and is affected by the amount of data and the modeling parameters. Incomplete uncertainty is given because the parameter uncertainty is ignored. Also, greater spatial continuity leads to more uncertainty. This increase is unreasonable in earth science. To address these problems, two approaches are proposed. The first approach is based on multiGaussian simulation where many realizations are performed at translated and/or rotated configurations and conditioned to the data. Variable configurations give different mean values that define uncertainty. The second approach is based on a stochastic trend; this approach randomizes the trend coefficients accounting for the fitted coefficients correlation. Variable set of coefficients provide different mean values. Furthermore, a methodology to account for parameter uncertainty is proposed. The uncertainty in the mean is transferred through simulation to deliver a more complete uncertainty.

Acknowledgements

I want to thank my supervisor Dr. Clayton Deutsch for his advice, guidance, patience and friendship. He is a source of new ideas and a great mentor.

I thank the Natural Sciences and Engineering Research Council of Canada as well as the industrial sponsors of the Centre of Computational Geostatistics (CCG) for their financial funding. Please keep supporting the research made in the CCG.

My parents, brothers and sister have been a source of stability and I would like to specially thank my mother, Emilia, for her constant encouragement to reach my goals. She is my inspiration.

My greatest thanks go to Serge, his patience and love is invaluable, Serge is my emotional support to stay positive every day.

Finally, I would like to thank everyone who helped me through my professional journey.

Table of Contents

List of Tables

List of Figures

Nomenclature

A	Area or domain of interest
a_l	lth regression coefficient
$\hat{a}_l$	Least square estimates of lth regression parameter a_l
$C(\mathbf{h})$	Covariance between two random variables separated by vector $\mathbf{h}$
$C(\mathbf{u}_1, \mathbf{u}_2)$	Covariance between random variables $Z(\mathbf{u}_1)$ and $Z(\mathbf{u}_2)$ of two different locations $\mathbf{u}_1$ and $\mathbf{u}_2$
$C(\hat{\mathbf{a}})$	Covariance matrix between the regression coefficients
$f(z)$	Probability density function of random variable Z (pdf)
$f_l(\mathbf{u})$	Function of the coordinates used in a trend model $m(\mathbf{u})$
$F(z)$	Cumulative distribution function of random variable Z (cdf)
$F(\mathbf{u}; z)$	Non-stationary cumulative distribution function of random variable $Z(\mathbf{u})$
$F(\mathbf{u}_1, \dots, \mathbf{u}_n; z_1, \dots, z_n)$	n variate cumulative distribution function of the n random variables $Z(\mathbf{u}_1), \dots, Z(\mathbf{u}_n)$
$F^{-1}(p)$	Quantile function or inverse cumulative distribution for the probability value, $p \in [0,1]$
$G(y)$	Standard normal or Gaussian cumulative distribution function
$G^{-1}(p)$	Standard normal quantile function or inverse cumulative distribution for the probability value, $p \in [0,1]$
$\mathbf{h}$	Separation vector distance between two points
n	Number of available data
N	The domain A is discretizated in N nodes
m	Stationary mean of the random variable $Z(\mathbf{u})$
$m(\mathbf{u})$	Mean at location $\mathbf{u}$, expected value of random variable $Z(\mathbf{u})$, or trend component model
p	Probability value
$\mathbf{u}$	A location in the space (1, 2, or 3D)
w	Independent Gaussian value

$w(\mathbf{u})$	Weight assigned to a datum at location $\mathbf{u}$ (often from declustering)
$Y(\mathbf{u})$	Continuous random variable at location $\mathbf{u}$ in normal score scale
Z	Generic random variable
z	A particular outcome of the random variable Z
$Z(\mathbf{u})$	Continuous random variable at location $\mathbf{u}$
$z(\mathbf{u})$	Realization from a random variable $Z(\mathbf{u})$
$Z^*(\mathbf{u})$	Estimation at location $\mathbf{u}$
$\gamma(\mathbf{h})$	Semivariogram between two random variables separated by vector $\mathbf{h}$
σ^2	Variance
σ_{al}	Standard deviation of *l*th regression parameter
ε	Deviation of the data from the regression model
λ_α	Kriging weight applied to sample α
ρ	Correlation (standardized covariance) coefficient between two variables
$\rho(\hat{\mathbf{a}})$	Correlation matrix between the regression coefficients
CB	Conventional Bootstrap
CFD	Conditional Finite Domain
SB	Spatial Bootstrap
ST	Stochastic Trend
$UMID$	Uncertainty in the Mean of the Input Distribution

Chapter 1

Introduction

Mineral deposits, petroleum reservoirs and environmental sites are uncertain because data scarcity makes local estimation a challenge. Geostatistical simulation constructs a set of realizations that provide an assessment of uncertainty. Input parameters are required in addition to the local data; however, the mean of the input distribution is perhaps the most important because the mean has a direct influence on resources and reserves: quantity of metal, hydrocarbon and contaminant. Uncertainty in the input mean could be transferred though the simulation to provide a more complete measure of uncertainty.

A realistic evaluation of uncertainty is important for mine planning. The priority of mining is given to zones of ore with low uncertainty and building access in zones of waste with low uncertainty. An improved evaluation of uncertainty could avoid some mistakes in mine planning, reduce problems and potentially increase profit during production.

The assessment of uncertainty is also used to identify zones of interest for sampling. Sampling areas unnecessarily or leaving areas unsampled would be suboptimal. A realistic evaluation of uncertainty permits the risk to be understood more accurately and improved decisions on sampling will be made.

1.1 Problem Setting

Analysis of uncertainty in resources and reserves impacts investment decisions. Consequently, the assessment of uncertainty is important for making the correct decision. Global evaluations are used in the early stages of a project, where there are too few data to perform reliable local evaluations. In the later stages, geostatistical simulation is used

to evaluate local and global uncertainty. The simulation often considers fixed input parameters, that is, parameter uncertainty is often ignored in simulation. Then, the global uncertainty may be underestimated.

Parameter uncertainty could be evaluated by bootstrap techniques. A shortcoming of these techniques is that conditioning data are not considered; the locations and the outcomes are randomized. Also, the bootstrap relies on the assumption that the data are representative.

Evaluations in large deposits with few drill holes show that local fluctuations cancel out because the range of correlation become small respect to the domain size. As this relation is increased, more locations in the domain obtain an expected value equal to the stationary mean m(**u**). As a result, the fluctuations in the global mean from simulation are unrealistically small.

1.2 Objectives of the Thesis

The principal objective is to improve the evaluation of uncertainty in resources and reserves. Both the local and the global uncertainty would be improved with the transference of the uncertainty in the mean through the process of simulation. There are techniques to evaluate parameter uncertainty; however they have limitations. The thesis reviews the available bootstrap techniques. New approaches are devised that overcome some of the pitfalls of traditional bootstrap techniques. The new approaches evaluate the uncertainty in the mean accounting for the domain limits and the conditioning data. The assumption of stationarity is relaxed by the use of trend equation. To accomplish the principal objective, a methodology is developed to transfer the uncertainty in the mean through simulation; a more complete evaluation of uncertainty is provided.

1.3 Proposed Approach

The bootstrap and spatial bootstrap are traditional tools to evaluate the uncertainty in statistical parameters; however, those techniques do not consider the domain limits and are not conditional to the available data. Two new approaches to evaluate the uncertainty in the mean of the input univariate distribution are proposed. The first one, the

implementation of conditional finite domain based on LU simulation, where the simulation is performed only at the points to be sampled conditioned to the data, the configuration of those sampled points follows the original strategy of sampling; the sampling of many simulated configurations gives different mean values that define uncertainty. The second approach is based on a stochastic trend. The use of a trend equation relaxes the assumption of stationarity and defines the mean dependent on the sample location within the domain. The stochastic trend approach randomizes the trend coefficients accounting for the correlation of the original fitted coefficients. Different coefficients provide different mean values that are combined to a distribution of uncertainty in the mean.

The uncertainty in the mean is transferred through simulation and the thesis demonstrates the use of a simple methodology to account for parameter uncertainty. Multiple distributions are constructed and used in simulation as reference distributions. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted for by changing the reference distribution for transformation.

1.4 Dissertation Online

Chapter Two presents some introductory notations, introduces the current analytical approach of uncertainty in the mean and reviews the theory of the bootstrap.

Chapter Three proposes the conditional finite domain technique that is based on LU conditional simulation that is performed only at sampled locations. Those locations are defined by the set of configurations that honour the original strategy of sampling and are inherent in the domain of interest. A sensitivity analysis demonstrates the robustness and reasonableness of this approach in scenarios where other techniques struggle with unrealistic uncertainty.

Chapter Four proposes a methodology to evaluate the uncertainty in the input parameter relaxing the assumption of stationarity. The mean of the random function is not taken to be a constant and is calculated as a function of location in the domain. The methodology

is implemented in a simple scenario to explain the methodology and a sensitivity analysis demonstrates the robustness of this approach in real scenarios.

Chapter Five presents a simple process to transfer of the uncertainty of the mean through the simulation. An example illustrates the methodology of the process, discusses and compares the results of simulation with parameter uncertainty.

Chapter Six gives the conclusions; summarizes the contributions of the thesis and presents future works.

An Appendix presents equations that quantify the fluctuations due to a finite domain size in presence of conditioning data; the analytical model is validated by the numerical uncertainty of many realizations. Many programs were developed through the thesis, the parameter code of those program are explained.

Chapter 2

Literature Review

The use of geostatistics is still growing since G. Matheron first introduced the methodology in 1962. Since then, many researchers have contributed to the developed of this science, such as (Journel & Huijbregts, 1978) and (David, 1977). The interest in geostatistics is high because crucial decisions are taken based on estimates of the resources and reserves. The resources and reserves evaluation process follows the stages of a drilling, quality assurance and quality control (QA/QC) of the data, a deterministic geological model, an evaluation of the grade and density, and considerations of economic and engineering factors. Relatively few data lead to inevitable uncertainty at every stage in the project evaluation.

The evaluation differs for various phases in the project, such as a scoping/conceptual study, a pre-feasibility/preliminary study and a feasibility/definitive study. Global evaluations are often used to make decisions in the first two studies. Afterwards, local evaluations are also required to make decisions. The difference of requirements is due to the nature of the decisions and a change in the data density, starting with sparse sampling in early studies and finishing with closer data to define local evaluations (Dominy, Noppé, & Annels, 2002).

Although a global mean is required in all the stages of resource and reserve evaluation, the objective of this thesis is to improve the evaluation of uncertainty in the grade evaluation stage of the process. Locations in the geologic block model are classified in decreasing order of confidence such as, measured, indicated, and inferred for the resource. The criteria of classification depend on the type of deposit and the expertise of the competent person. Classification may be based on the kriging variance or the local

variance of the weighted average (Arik, 1999). Another way to evaluate uncertainty is with conditional simulation (Journel & Huijbregts, 1978), this procedure provides realizations of the possible grades at unsampled locations.

This uncertainty obtained from geostatistical simulation depends on many input parameters, such us the variogram and input distribution of the data. Variability of the input parameters is usually ignored and a prediction error could be wrongly estimated (Wang & Wall, 2003). All input parameters have uncertainty; however, the distribution of the data is considered the most important in the simulation because this is the target distribution that simulation tries to reproduce (Babak & Deutsch, 2008). The simulation process assumes that the input univariate distribution is representative of the orebody. Preferential sampling is handled by declustering techniques, but the uncertainty in the distribution must also be considered.

The conventional bootstrap (CB) procedure assesses uncertainty in statistical parameters, but considers the data to be independent. The bootstrap is an application of Monte Carlo simulation where the samples are drawn with replacement (Efron, 1979). A set of simulated realizations are generated to define the model of uncertainty. The bootstrap approach assumes an arbitrary randomization and data independence that is not suitable to geological data. The independence assumption of the CB technique is relaxed with the use of spatial correlation in the sampling. This leads to a technique called the spatial bootstrap (SB) and another technique called the conditional finite domain (CFD) technique. These will be described after some notation is presented.

2.1 Some Introductory Notation

Scarce data is a usual feature of reservoir data and uncertainty arises because sampling is not complete until actual mining takes place. Even at the time of mining, production drilling does not completely define complex deposits. Another feature that increases uncertainty is preferential sampling in zones with greater economic value. Data are often expensive to collect and it is neither optimal nor feasible to collect data uniformly over the entire site being characterized (Leuangthong, Khan, & Deutsch, 2008). Set of unsampled locations is evaluated given original sparse data.

The set of unsampled locations in a deposit are defined as a set of spatially dependent random variables where the local uncertainty at any particular location **u** is modeled by the random variable at that location. A location is defined by the vector **u** that defines the East, North and Elevation coordinates. The available data are viewed as particular outcomes z of the random variables Z at the locations that have been sampled. We are mainly concerned with metal concentrations that are continuous random variables, $Z(\mathbf{u})$. These variables are characterized by their cumulative distribution fuction (cdf). This function specifies the probability that the variable Z at location **u** is no greater than any given threshold z and the cfd is a non-decreasing function of z (Goovaerts, 1997):

$$F(\mathbf{u};z) = \text{Prob}\{Z(\mathbf{u}) \leq z\} \quad \in [0,1] \tag{2.1}$$

The probability density function (pdf) is the derivative of the cumulative distribution function if it exists:

$$f(\mathbf{u};z) = F'(\mathbf{u};z) \tag{2.2}$$

The probability density function $f(\mathbf{u};z)$ and the histogram have a similar shape, but the frequency of samples in a discrete histogram class is not the same as the continuous pdf curve.

2.1.1 Random Function

A random function (RF) is defined as a set of random variables at many spatial locations **u** in an study area A, $\{Z(\mathbf{u}), \forall\ \mathbf{u} \in A\ \}$. A characteristic of mineral deposits is their spatial correlation between the random variables. A set of n locations could be defined by their respective vectors $\mathbf{u}_i$ where $i = 1,\ldots,n$. Thus, n random variables $\{Z(\mathbf{u}_1),\ldots,Z(\mathbf{u}_n)\}$ are characterized by an n point cdf or multivariate cdf (2.3). This defines the joint uncertainty about the n actual values $z(\mathbf{u}_1),\ldots,z(\mathbf{u}_n)$. This is also sometimes referred to as the spatial law of the random function $Z(\mathbf{u})$.

$$F(\mathbf{u}_1,\ldots,\mathbf{u}_n;z_1,\ldots,z_n) = \text{Prob}\{Z(\mathbf{u}_1) \leq z_1,\ldots,Z(\mathbf{u}_n) \leq z_n\} \tag{2.3}$$

In practical geostatistics application, it is impossible to use the complete spatial law because of incomplete sampling. Often, the first two moments provide reasonable

solutions (Journel & Huijbregts, 1978). The first order moment is the expected value of the distribution function of $Z(\mathbf{u})$ that is, in general, a function of location **u.**

$$\mathrm{E}\{Z(\mathbf{u})\} = m(\mathbf{u}) \tag{2.4}$$

The variance, covariance and variogram are known as second order moments. The second order moment about the expectation of the random variable is defined as follows. It may also be a function of location **u**.

$$\mathrm{Var}\{Z(\mathbf{u})\} = \mathrm{E}\left\{\left[Z(\mathbf{u}) - m(\mathbf{u})\right]^2\right\} \tag{2.5}$$

The covariance is defined between two random variables $Z(\mathbf{u}_1)$ and $Z(\mathbf{u}_2)$ of two different locations $\mathbf{u}_1$ and $\mathbf{u}_2$.

$$C(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{E}\left\{\left[Z(\mathbf{u}_1) - m(\mathbf{u}_1)\right] \times \left[Z(\mathbf{u}_2) - m(\mathbf{u}_2)\right]\right\} \tag{2.6}$$

The variogram function is defined as the variance of the increments or difference between two random variables at locations $\mathbf{u}_1$ and $\mathbf{u}_2$. In practice, the semivariogram is identified with the function $\gamma(\mathbf{u}_1,\mathbf{u}_2)$

$$2\gamma(\mathbf{u}_1, \mathbf{u}_2) = \mathrm{Var}\{Z(\mathbf{u}_1) - Z(\mathbf{u}_2)\} \tag{2.7}$$

The multivariate distribution of *n* locations or the summary first and second order moments can only be calculated in practice by combing a sufficient number of data together.

2.1.2 Stationarity

The decision of stationarity permits the use of samples at different locations to infer a model of the probability distribution. The decision to put together some samples is based on the idea that they belong to a zone of homogeneous mineralization. Where, the correlation of two data values $z(\mathbf{u}_1)$ and $z(\mathbf{u}_2)$ does not depend on their locations within the study area but rather on the **h** vector between these data (Journel & Huijbregts, 1978).

The stationarity of first order (2.8) declares that the mean of the random function is constant over the domain and the stationarity of second order (2.9) declares that the

covariance is constant under translation over the area of interest A. The covariance depends of the **h** vector between two random functions $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$. Where **h** represent a vector in the three dimensional space.

$$\mathrm{E}\big(Z(\mathbf{u})\big) = m, \quad \forall \mathbf{u} \in A \tag{2.8}$$

$$\mathrm{Cov}(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})) = C(\mathbf{h}), \quad \forall \mathbf{u} \in A \tag{2.9}$$

A decision of stationarity is required in practice; otherwise, no inference beyond the data values is possible. Although the data belong to a homogeneous geological domain, they are also related together in important ways.

2.1.3 Dependence of the Random Variables

Two random variables $Z(\mathbf{u})$ and $Z(\mathbf{u}')$ are dependent if the probability distribution of either one is affected by the knowledge about the other one, then, a conditional cumulative distribution function (ccdf) is the cumulative distribution function of $Z(\mathbf{u})$ given knowledge about $Z(\mathbf{u}')$ (Goovaerts, 1997):

$$\begin{aligned} F\left(\mathbf{u}; z \middle| Z(\mathbf{u}') \le z'\right) &= \mathrm{Prob}\left\{Z(\mathbf{u}) \le z \middle| Z(\mathbf{u}') \le z'\right\} = \frac{\mathrm{Prob}\left\{Z(\mathbf{u}) \le z, Z(\mathbf{u}') \le z'\right\}}{\mathrm{Prob}\left\{Z(\mathbf{u}') \le z'\right\}} \\ &= \frac{F\left(\mathbf{u}, \mathbf{u}'; z, z'\right)}{F\left(\mathbf{u}'; z'\right)} \end{aligned} \tag{2.10}$$

The independence of random variables is when the ccdf of one random variable given other random variable is equal to the cumulative distribution of the first random variable:

$$\begin{aligned} F\left(\mathbf{u}; z \middle| Z(\mathbf{u}') \le z'\right) &= F\left(\mathbf{u}; z\right) \quad \forall z' \\ F\left(\mathbf{u}'; z' \middle| Z(\mathbf{u}) \le z\right) &= F\left(\mathbf{u}'; z'\right) \quad \forall z \end{aligned} \tag{2.11}$$

From the last expression of (2.10) and first expression of (2.11) it is seen that the bivariate distribution between two independent random variables is simply the product of the lower order distributions.

$$F\left(\mathbf{u}, \mathbf{u}'; z, z'\right) = F\left(\mathbf{u}; z\right) F\left(\mathbf{u}'; z'\right) \ \forall z, z' \tag{2.12}$$

As mentioned, in practice, grades at different locations within a mineral deposit are often dependent on each other.

2.1.4 Uncertainty in the Mean of the Univariate Distribution

Many realizations $z(\mathbf{u}^{(1)}),\ldots,z(\mathbf{u}^{(L)})$ of $Z(\mathbf{u})$ are required to infer their probability law. Similarly, many realizations of the univariate distribution are required to evaluate the uncertainty in the mean of the univariate distribution. There are different approaches to evaluate uncertainty in the mean that involve assumptions such as (1) the data are independent or spatially correlated, (2) the realizations are limited to some domain or not, (3) the realizations are conditioned to the original data or not, and (4) the mean could be calculated based on a trend equation. A process of sampling is often used to assemble set of histograms. Some approaches sample from the original data and others sample from the map of estimations using the original spatial strategic of sampling.

2.2 Traditional Techniques to Evaluate Uncertainty in the Input Parameter

Geological models are uncertain because relatively sparse data are used to evaluate grade and tonnes. Geostatistical simulation is used to build multiple realizations that will help assess uncertainty. Input parameters are required in addition to the local data; input parameters include the histogram, variogram, search distances and other implementation decisions. The mean of the histogram is perhaps the most important because it has a direct influence on the evaluations. Following is a review of techniques to calculate uncertainty in the mean.

2.2.1 Analytical

Consider n random variables $\{Z_i,\ i = 1,\ldots,n\}$ that are taken as samples from a stationary statistical population. Stationary entails that the mean and variance are the same under translation:

$$\mathrm{E}\{Z_i\} = m \quad \forall\ i = 1,\ldots,n$$
$$Var\{Z_i\} = \mathrm{E}\{(Z_i - m)^2\} = \sigma^2 \quad \forall\ i = 1,\ldots,n$$

Consider that the multiple random variables are equally weighted to obtain the arithmetic mean.

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \tag{2.13}$$

Consider the mean itself to be a random variable, then, the expected value and the variance of these arithmetic means are given by:

$$\begin{aligned} \mathrm{E}\left\{\overline{Z}\right\} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{Z_i\right\} \\ \mathrm{Var}\left\{\overline{Z}\right\} &= \mathrm{E}\left\{\left[\overline{Z}-m\right]^2\right\} \end{aligned} \tag{2.14}$$

The variance of these means require knowledge of the covariance between the Z_i RVs. As mentioned above, the covariance of two random variables is given by the expected value of the product of the difference of both variables and their respective means.

$$\mathrm{Cov}\left(Z_i, Z_j\right) = \mathrm{E}\left\{\left[Z_i - m\right]\times\left[Z_j - m\right]\right\} \tag{2.15}$$

The previous covariance expression is expanded and simplified:

$$\begin{aligned} \mathrm{Cov}\left(Z_i, Z_j\right) &= \mathrm{E}\left\{Z_i Z_j - Z_i m - Z_j m + m^2\right\} \\ &= \mathrm{E}\left\{Z_i Z_j\right\} - \mathrm{E}\left\{Z_i\right\} m - \mathrm{E}\left\{Z_j\right\} m + m^2 \\ &= \mathrm{E}\left\{Z_i Z_j\right\} - m^2 \end{aligned} \tag{2.16}$$

The variance of the mean is similarly expanded and simplified:

$$\begin{aligned} \mathrm{Var}\left\{\overline{Z}\right\} &= \mathrm{E}\left\{\left[\overline{Z}-m\right]^2\right\} \\ &= \mathrm{E}\left\{\overline{Z}^2 - 2\overline{Z}m + m^2\right\} \\ &= \mathrm{E}\left\{\overline{Z}^2\right\} - 2m\mathrm{E}\left\{\overline{Z}\right\} + m^2 \\ &= \mathrm{E}\left\{\overline{Z}^2\right\} - m^2 \end{aligned} \tag{2.17}$$

Equation (2.13) is replaced in the previous equation to obtain the variance of the mean.

$$\begin{aligned}\mathrm{Var}\left\{\overline{Z}\right\} &= \mathrm{E}\left\{\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]^2\right\} - m^2 \\ &= \mathrm{E}\left\{\left[\frac{1}{n}\sum_{i=1}^{n} Z_i\right]\left[\frac{1}{n}\sum_{j=1}^{n} Z_j\right]\right\} - m^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}\left\{Z_i Z_j\right\} - m^2\end{aligned} \tag{2.18}$$

The expected value of the product of random variables is replaced by the derived term of Equation (2.16), then, the equation of the variance of the means is simplified to the average of the covariances between n multiple random variables. The stationarity assumption of first order and second order are inherent to infer the next expression.

$$\begin{aligned}\mathrm{Var}\left\{\overline{Z}\right\} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mathrm{Cov}\left(Z_i, Z_j\right) + m^2\right] - m^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(Z_i, Z_j\right) + \frac{1}{n^2}n^2 m^2 - m^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(Z_i, Z_j\right)\end{aligned} \tag{2.19}$$

The expression of the average of the covariance between multiple random variables involves $n \times n$ covariance values. The equation of the variance may be rewritten by separating out the covariance of each value with itself, that is, the covariances $\mathrm{Cov}(Z_i, Z_i) = \mathrm{Var}(Z_i)$.

$$\mathrm{Var}\left\{\overline{Z}\right\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(Z_i, Z_j\right) + \frac{1}{n^2}\sum_{i=1}^{n}\sigma_{Z_i}^2$$

$$\mathrm{Var}\left\{\overline{Z}\right\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}\left(Z_i, Z_j\right) + \frac{\sigma_Z^2}{n} \tag{2.20}$$

where $i \neq j$

From the previous equation, a lack of correlation between random variables causes the first term to be equal to zero, then, the variance of the mean is equal to the variance of samples inversely proportional to the number of samples. The variance of the mean assumming independence obtains less variance that using spatial correlation because the term of correlation between no equal samples become zero.

$$\mathrm{Var}\left\{\overline{Z}\right\} = \frac{\sigma_Z^2}{n} \tag{2.21}$$

The previous equation is used to derive Equation (2.22) of the *effective* number of data. This is a relation between the variance of the samples and the uncertainty of the mean.

$$n_{eff} = \frac{\sigma_Z^2}{\mathrm{Var}\left\{\overline{Z}\right\}} \tag{2.22}$$

The effective number of independent data is equal to the number of input data (n) when the evaluation of uncertainty in the input parameter is based on independent data. n_{eff} defines how many independent data are really available (Kitanidis, 1997).The effective number of independent data is less than n.

Complex statistics such as uncertainty in the mean of the fraction of material above a cutoff are not necessarily amenable to such simple analytical calculations showed before and simulation procedures are required.

2.2.2 Conventional Bootstrap

The conventional bootstrap is an application of Monte Carlo simulation and is a statistical re-sampling technique that permits the quantification of uncertainty in statistics by drawing from the original data (Efron, 1979). The assumption of independence is the main feature of the conventional bootstrap technique. The samples could be drawn multiple times since the process of sampling is executed with replacement from the input distribution. This bootstrap may be useful to measure the uncertainty in the mean in the early stage of a project, where the data are widely spaced (Deutsch, 2002). The methodology is as follows:

- Assemble the representative histogram of the n available data and the cumulative distribution function F(z) of this data. Declustering methods could be used for irregular preferential sampling and debiasing if required.

- Draw n uniformly distributed random numbers between 0 and 1. Those values characterize the cdf values on the vertical axes of Figure 2.1. Therefore, p_i, i = 1,...,n.

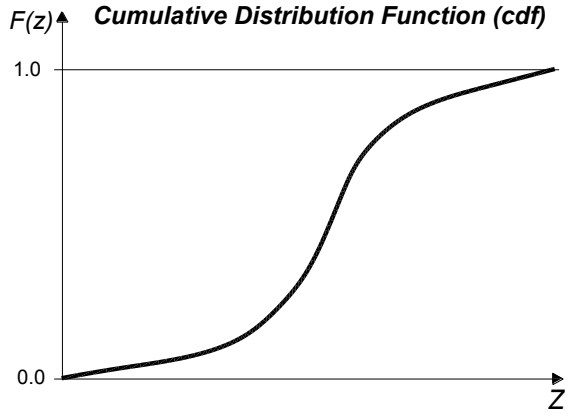


Figure 2.1: Illustration of the cumulative distribution function of the input available data, the sampling is drawn from this function to obtain L realizations of the input univariate distribution.

- Read the equivalent quantile values $F^{-1}(p_i)$, $i = 1,...,n$.
- The drawn samples z_i, $i = 1,...,n$ represent a realization of the original univariate distribution. The statistics of interest (the mean) is calculated from the realization.
- The previous three steps are repeated many times for L realizations.
- The distribution of the L possible statistics (mean values) provides a model of uncertainty in the statistic.

The uncertainty does not consider correlation between the samples and could be considered in the early stages of a project, when the data are widely spaced.

2.2.3 Spatial Bootstrap

The limitation of the conventional bootstrap is the assumption of independence, then, spatial correlation could be incorporated in the technique. The Spatial Bootstrap (SB) is a generalization of the bootstrap concept (Deutsch, 2004). The SB uses unconditional simulation and the covariance function. An unconditional simulation of the random function is a realization of $Z(\mathbf{u})$. Its construction requires knowledge of the spatial distribution of the random function $Z(\mathbf{u})$ (Chilès & Delfinier, 1999).

The spatial distribution of the random functions is described by the variogram also called structural analysis of a regional phenomenon (Journel & Huijbregts, 1978). The structural

analysis consists of three steps, the construction of experimental variogram, understanding of the results, and fitting an acceptable model.

The variogram is defined again in Expression (2.23). This equation shows the variability between $Z(\mathbf{u})$ and $Z(\mathbf{u}+\mathbf{h})$ separated by the distance vector **h.**

$$2\gamma(\mathbf{h}) = \mathrm{E}\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]^2\right\} \tag{2.23}$$

The **h** vector distance is used plus and minus some distance tolerance to get more samples in the evaluation of variability. As more pairs are used in every lag distance evaluation, there is more chance of having a robust description of the structure. Different lag distances are evaluated; the plot of the experimental variogram shows the variability at different lag distances simultaneously.

The variability of the grade is investigated by direction. Many directions are evaluated to find the ones with the greatest and least continuity. The experimental variograms must honour the conceptual geological model of the deposit.

The last step is to fit the experimental variograms with valid functions such as the spherical, exponential, or Gaussian function. The variogram function must have the mathematical property of positive definiteness for the respective covariance model and, thus, the covariance model necessarily has a strictly positive determinant. This ensures that the kriging equations can be solved and obtain a positive kriging variance (Journel & Huijbregts, 1978). The variogram could be modeled as a positive sum of variogram functions to fit better the experimental variogram (2.24). The variogram model converges to a value that is called the sill and the distance at which the variogram reaches this plateau is called the range.

$$\gamma(\mathbf{h}) = \sum_{i=0}^{nst} C_i \Gamma_i(\mathbf{h}) \tag{2.24}$$

The values at adjacent locations are expected to have low variability; however, unexpected variability is given because some mistakes are present during the taking of samples or the deposit shows complex geological features at small scale. This variability is modeled by an independent structure called the nugget effect (C_o).

Many directions are evaluated to capture the behaviour of the continuity in three dimensional space and to identify the anisotropy of the phenomena under study. In mining, geometric anisotropy is common, where more continuity is observed in one direction and less continuity is seen in the perpendicular or vertical direction. The variability of these directions often converges to the same sill. The sills could change in each direction (Isaaks & Srivastava, 1989).

The spatial bootstrap considers the spatial correlation of the samples. The basic approach is explained by the following steps:

- Assemble the representative distribution of the Z random variable $F_Z(z)$; consider the use of declustering method for irregular grid and debiasing method if appropriate.
- Transform the $F_Z(z)$ to normal score $G(y)$ and compute their 3D variogram model $\gamma(\mathbf{h})$.
- Using $\gamma(\mathbf{h})$, calculate the covariance matrix n × n among input location samples. Then, calculate its Cholesky decomposition. $\mathbf{C} = \mathbf{L}\ \mathbf{U}$, where $\mathbf{L}$ is the lower matrix and $\mathbf{U}$ the upper matrix.
- Draw n values from uniformly distribution between 0 and 1. These independent values are transformed to Gaussian values w_i, $i = 1,...,n$
- Correlate w_i independent values by multiplying with the lower matrix n × n of the Cholesky decomposition, $\mathbf{y} = \mathbf{L}\,\mathbf{w}$
- $\mathbf{y}$ is the result $n \times 1$ vector of values with correlation. These Gaussian values are transformed to probability values to locate in the representative distribution. $p_i = G(y_i)$, $i = 1,...,n$
- The z-values are calculated as: $z_i = F_Z^{-1}(p_i)$, $i = 1,...,n$. The mean of z_i values is stored for further global evaluation of the mean.
- The last four steps are repeated L realizations. In summary, the generation of L realizations can be calculated simply as: $\mathbf{z}^{(l)} = F_Z^{-1}(G(\mathbf{L}\,\mathbf{w}^{(l)}))$, $l = 1,...,L$.

Assemble the distribution of L possible means to provide a model of uncertainty in the mean of the input parameter.

2.2.4 Comments on the Current Methodologies

Considering spatial correlation generates more uncertainty than assuming independence of the data. The analytical solution illustrates this when the covariance between two different data is non-zero.

A shortcoming of these techniques is that conditioning data are not considered; the locations and the outcomes are randomized. Also, the representativeness of the data may be questionable because preferential sampling is a common feature of geological data. An alternative technique uses a different pattern of randomization to define the error of the distribution. The pattern consists on sampling from multiple conditional simulations using the original sampling strategy (Journel A. G., 1994). This concept is the general idea of conditional finite domain (CFD). Where the sampling is done from multiple translate and/or rotated configurations after being simulated conditioned to the original data. The sampling is limited to the domain of interest. (Babak & Deutsch, 2008).

Chapter 3

Conditional Finite Domain

The Conditional Finite Domain (CFD) technique quantifies the uncertainty in the mean. This technique samples from translate and/or rotated configurations of the data to obtain different mean values and assembles a distribution of them. The CFD technique may be considered as an extension of the conventional bootstrap and spatial bootstrap techniques. CFD starts by creating new configurations by random translation and/or random rotation of the data locations throughout the domain. An order k_i, $i = 1,..,K$ of the CFD approach represents sets of realizations or configurations that use reference distributions from the previous order (k_{i-1}). The configurations are simulated with different reference distributions and the realizations are conditioned to the original data. The sampling of many simulated configurations gives different mean values that define a mean and uncertainty for each order. The uncertainty is stabilized after many configurations and orders are performed.

The total number of realizations is equal to the number of orders required multiplied by the number of configurations. Each realization has n conditioning data and n locations to simulate. The original CFD technique was implemented using sequential Gaussian simulation (sgsim) by (Babak & Deutsch, 2008). The use of sgsim is relatively inefficient because the search strategy and covariance lookup table do not allow simulation only at some locations because it uses the full grid; otherwise, the LU algorithm permits simulation only at the n locations to be sampled which is more efficient than sgsim.

Simulation is done in Gaussian units, then, a back transformation is performed. The normal scores transformation of the data becomes sensitive to the extrapolation options in

the lower and upper tail as the order of simulation increases. Reasonable values must be chosen in the implementation.

3.1 LU Simulation in CFD

Simulation through LU decomposition of the covariance matrix is a well established multiGaussian simulation technique (Goovaerts, 1997). The LU decomposition provides a fast solution provided there are not too many locations. The CFD technique simulates only at the locations to be sampled, therefore, the CFD technique based on LU decomposition is reasonable. The simulation of a continuous attribute z at N nodes $\mathbf{u}^{(j)}$ conditioned to the data set $\{z(\mathbf{u}_i),\ i = 1,\ldots,n\}$. The LU simulation approach is strictly made in Gaussian space. The original z values must be transformed to normal scores and the variogram must be calculated with these normal score values.

The covariance matrix between n input data values and N nodes to simulate is as follow:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1) & \cdots & C(\mathbf{u}_1 - \mathbf{u}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n - \mathbf{u}_1) & \cdots & C(\mathbf{u}_n - \mathbf{u}_n) \end{bmatrix} & \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}^{(1)}) & \cdots & C(\mathbf{u}_1 - \mathbf{u}^{(N)}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n - \mathbf{u}^{(1)}) & \cdots & C(\mathbf{u}_n - \mathbf{u}^{(N)}) \end{bmatrix} \\ \begin{bmatrix} C(\mathbf{u}^{(1)} - \mathbf{u}_1) & \cdots & C(\mathbf{u}^{(1)} - \mathbf{u}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}^{(N)} - \mathbf{u}_1) & \cdots & C(\mathbf{u}^{(N)} - \mathbf{u}_n) \end{bmatrix} & \begin{bmatrix} C(\mathbf{u}^{(1)} - \mathbf{u}^{(1)}) & \cdots & C(\mathbf{u}^{(1)} - \mathbf{u}^{(N)}) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}^{(N)} - \mathbf{u}^{(1)}) & \cdots & C(\mathbf{u}^{(N)} - \mathbf{u}^{(N)}) \end{bmatrix} \end{bmatrix}$$

The covariance $\mathbf{C}$ is symmetric since $\mathbf{C}_{21} = \mathbf{C}_{12}^{T}$. Also $\mathbf{C}$ is positive definite. A symmetric positive definite matrix has a Cholesky $\mathbf{LU}$ decomposition $\mathbf{C} = \mathbf{L}\,\mathbf{U} = \mathbf{L}\,\mathbf{L}^{T}$. Where the lower triangular matrix has all the elements above the diagonal as null and the upper triangular matrix has all the elements below the diagonal as null. Then, the decomposition of the large covariance matrix $\mathbf{C}$ is defined as follow:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{11} & 0 \\ \mathbf{A}_{21} & \mathbf{L}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{11} & \mathbf{B}_{12} \\ 0 & \mathbf{U}_{22} \end{bmatrix}$$

Both $\mathbf{L}_{11}$ and $\mathbf{L}_{22}$ are lower matrices; while, the upper triangular matrices are $\mathbf{U}_{11}$ and $\mathbf{U}_{22;}$ otherwise, the matrices $\mathbf{A}_{21}$ and $\mathbf{B}_{12}$ are not triangular matrices. From the previous expression, the covariance matrix between the data locations is determined by LU decomposition.

$$\mathbf{C}_{11} = \mathbf{L}_{11}\mathbf{U}_{11}$$

Both matrices $\mathbf{B}_{12}$ and $\mathbf{A}_{21}$ are expressed in function of their covariances.

$$\mathbf{C}_{12} = \mathbf{L}_{11}\mathbf{B}_{12} \Rightarrow \mathbf{B}_{12} = \mathbf{L}_{11}^{-1}\mathbf{C}_{12}$$

One may notice that $(\mathbf{U}^{-1})^{T} = (\mathbf{L}^{-1})$ or $(\mathbf{L}^{-1})^{T} = (\mathbf{U}^{-1})$ because they are symmetric matrices. Then, the transpose of $\mathbf{A}_{21}$ matrix is equal to the $\mathbf{B}_{12}$ matrix. These matrices are equal to the product of triangular matrices $\mathbf{L}_{11}^{-1}$ and $\mathbf{C}_{12}$.

$$\mathbf{C}_{21} = \mathbf{A}_{21}\mathbf{U}_{11} \Rightarrow \mathbf{A}_{21} = \mathbf{C}_{21}\mathbf{U}_{11}^{-1} \Rightarrow \mathbf{A}_{21}^{T} = \left(\mathbf{U}_{11}^{-1}\right)^{T}\mathbf{C}_{12} = \mathbf{L}_{11}^{-1}\mathbf{C}_{12}$$
$$\therefore \mathbf{A}_{21}^{T} = \mathbf{B}_{12} = \mathbf{L}_{11}^{-1}\mathbf{C}_{12}$$

The covariance matrix between estimated node locations $\mathbf{C}_{22}$ is equal to the **LU** decomposition $\mathbf{L}_{22}\mathbf{U}_{22}$ only if the covariance matrix between nodes and data $\mathbf{C}_{21}$ is equal to zero. Otherwise, the covariance $\mathbf{C}_{22}$ is derived from the following:

$$\mathbf{C}_{22} = \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{L}_{22}\mathbf{U}_{22}$$

The product of $\mathbf{A}_{21}$ and $\mathbf{B}_{12}$ is replaced by their equivalent equations derived from the large covariance matrix **C**. Then, the product of the lower triangular matrix $\mathbf{L}_{22}$ and the upper triangular matrix $\mathbf{U}_{22}$ is expressed as follow:

$$\textit{if } \mathbf{A}_{21}\mathbf{B}_{12} = \mathbf{C}_{21}\mathbf{U}_{11}^{-1}\mathbf{L}_{11}^{-1}\mathbf{C}_{12} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$$
$$\Rightarrow \mathbf{L}_{22}\mathbf{U}_{22} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$$

The equations for simple kriging and the kriging variance are used to demonstrate that $\mathbf{L}_{22}\mathbf{U}_{22}$ corresponds to matrix of error covariances $\mathbf{K}_{22}$ (Deutsch, 2000). Simple kriging is written as a function of covariances as:

$$Z^{*} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}Z_{\alpha}$$
$$\textit{if sk system } \lambda = \mathbf{C}_{12}\mathbf{C}_{11}^{-1} \Rightarrow \lambda^{T} = \mathbf{C}_{21}\mathbf{C}_{11}^{-1}$$
$$\therefore Z^{*} = \lambda^{T}Z_{\alpha}$$

The matrix of error covariances $\mathbf{K}_{22}$ is deduced as follow:

$$E\left\{\left[Z\left(\mathbf{u}^{(i)}\right)-Z^{*}\left(\mathbf{u}^{(i)}\right)\right]\left[Z\left(\mathbf{u}^{(j)}\right)-Z^{*}\left(\mathbf{u}^{(j)}\right)\right]\right\}\forall(i),(j)=1,...,N$$

$$E\left\{\left[Z\left(\mathbf{u}^{(i)}\right)Z\left(\mathbf{u}^{(j)}\right)-Z\left(\mathbf{u}^{(i)}\right)Z^{*}\left(\mathbf{u}^{(j)}\right)-Z^{*}\left(\mathbf{u}^{(i)}\right)Z\left(\mathbf{u}^{(j)}\right)+Z^{*}\left(\mathbf{u}^{(i)}\right)Z^{*}\left(\mathbf{u}^{(j)}\right)\right]\right\}$$

The simple kriging system $Z^{*}=\lambda^{\mathrm{T}}Z_{\alpha}$ is replaced in the previous extended expression.

$$E\left\{\left[Z\left(\mathbf{u}^{(i)}\right)Z\left(\mathbf{u}^{(j)}\right)-\lambda^{T}Z_{\alpha}Z\left(\mathbf{u}^{(i)}\right)-\lambda^{T}Z_{\alpha}Z\left(\mathbf{u}^{(j)}\right)+\lambda^{T}Z_{\alpha}\lambda^{T}Z_{\alpha}\right]\right\}$$

The extended equation of the covariance is split in four terms resulting in the covariance between estimated node locations minus the transpose of the weights matrix times the covariance between data locations and estimated node locations. This result corresponds to $\mathbf{L}_{22}\mathbf{U}_{22}$ and is the decomposition of the $\mathbf{K}_{22}$ matrix of error covariances. The lower triangular matrix $\mathbf{L}_{22}$ is used in algorithms of conditional or non conditional simulation.

$$\begin{aligned}
&E\left\{Z\left(\mathbf{u}^{(i)}\right)Z\left(\mathbf{u}^{(j)}\right)\right\}=\mathbf{C}_{22}\\
&-\lambda^{T}E\left\{Z_{\alpha}Z\left(\mathbf{u}^{(i)}\right)\right\}=-\lambda^{T}\mathbf{C}_{12}\\
&-\lambda^{T}E\left\{Z_{\alpha}Z\left(\mathbf{u}^{(j)}\right)\right\}=-\lambda^{T}\mathbf{C}_{12}\\
&+\lambda^{T}E\left\{Z_{\alpha}Z_{\alpha}^{T}\right\}\lambda=+\lambda^{T}\mathbf{C}_{11}\lambda
\end{aligned}$$

$$\Rightarrow\mathbf{C}_{22}-2\lambda^{T}\mathbf{C}_{12}+\lambda^{T}\mathbf{C}_{11}\lambda=\mathbf{C}_{22}-2\lambda^{T}\mathbf{C}_{12}+\lambda^{T}\mathbf{C}_{11}\mathbf{C}_{12}\mathbf{C}_{11}^{-1}\therefore\mathbf{K}_{22}=\mathbf{C}_{22}-\lambda^{T}\mathbf{C}_{12}$$

Then, a conditional simulation is performed by the product of the **L** matrix by **w** matrix of normal deviates. Also, Y_1 is the column matrix of n normal score conditioning data and Y_2 represents the column matrix of the N conditional simulated values (Deutsch & Journel, 1998).

$$\begin{bmatrix}Y_1\\Y_2\end{bmatrix}=\mathbf{L}\mathbf{w}=\begin{bmatrix}\mathbf{L}_{11}&0\\\mathbf{A}_{21}&\mathbf{L}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{w}_1\\\mathbf{w}_2\end{bmatrix}$$

$$Y_1=\mathbf{L}_{11}\mathbf{w}_1\Rightarrow\mathbf{w}_1=\mathbf{L}_{11}^{-1}Y_1$$

The vector of the conditional simulation Y_2 is shown as follow. Where, $\mathbf{w}_1$ is replaced by its equivalent expression. The $\mathbf{L}_{22}\mathbf{w}_2$ term symbolizes the error vector and represents an unconditional simulation. Otherwise, the first term is linked with the expression of kriging.

$$Y_2 = \mathbf{A}_{21}\mathbf{w}_1 + \mathbf{L}_{22}\mathbf{w}_2 = \mathbf{A}_{21}\mathbf{L}_{11}^{-1}Y_1 + \mathbf{L}_{22}\mathbf{w}_2$$

Conditional simulations are generated by drawing set of error vectors $\mathbf{L}_{22}\mathbf{w}_2$, that is, by drawing different random number vectors $\mathbf{w}_2$ (Deutsch, 2000). These vectors $\mathbf{w}_2$ are independent normal (0,1) random values that are correlated through the product with $\mathbf{L}_{22}$. As explained before, $\mathbf{L}_{22}$ is the lower triangular matrix of the decomposition of $\mathbf{K}_{22} = \mathbf{L}_{22}\mathbf{U}_{22} = \mathbf{C}_{22} - \mathbf{A}_{21}\mathbf{B}_{12}$, that is, $\mathbf{K}_{22}$ matrix is positive definite and symmetric and their **LU** decomposition is always possible.

3.2 Methodology

The uncertainty in the mean converges to a constant value after many orders of simulation are performed. Uncertainty in the mean relies on the series of simulations and stochastic algorithms. Thus, similar to the Markov Chain Monte Carlo approach, a period of "burn-in" is followed by a period of stabilization where similar fluctuations of the uncertainty around some constant value are observed. The algorithm of CFD is summarized as follows:

- Assemble the representative histogram of the input data; one must use declustering method for irregular grid and debiasing method if the scenario requires.
- Set L number of configurations through translation and rotation of the data locations.
- k_i, $i = 0$, apply **LU** conditional simulation with the distribution of the original data as reference. Simulation of l_j, $j = 1,...,L$ equiprobable realizations of the variable of interest in L new configurations.
- Sample simulated values of each l_j configuration. New reference distributions are assembled for the next step.
- k_i, $i = 1$, apply LU conditional simulation with reference distributions established in k_{i-1} to create l_j, $j = 1,...,L$ equiprobable realizations of the variable of interest in L configurations and conditioned to the original data.
- Sample simulated values of each l_j configuration, calculate and store the mean of the configuration, the mean of the order and standard deviation of the means.

- Assemble new l_j, j = 1,...,L reference distributions with sampled values of the configurations.

- The last three steps are repeated as a loop until the desired number of K orders is reached.

Through the whole process, the mean and variance are stored for every l_j realization and k_i order. The uncertainty in the mean is evaluated from the distribution of the possible means in L configurations and K orders. Also, the uncertainty of K orders could be illustrated in a plot of the standard deviation against the order number to verify convergence. The process of sampling is done from L configurations within the study area; hence the Finite Domain part of the name.

Figure 3.1 shows a small example that illustrates the methodology, where the original distribution in blue is used like reference distributions only in the order zero. Four samples $z(\mathbf{u}_i)$ for i = 1,...,4 are located in a domain of four by four units. The reference distributions that are updated after every order are illustrated in orange.

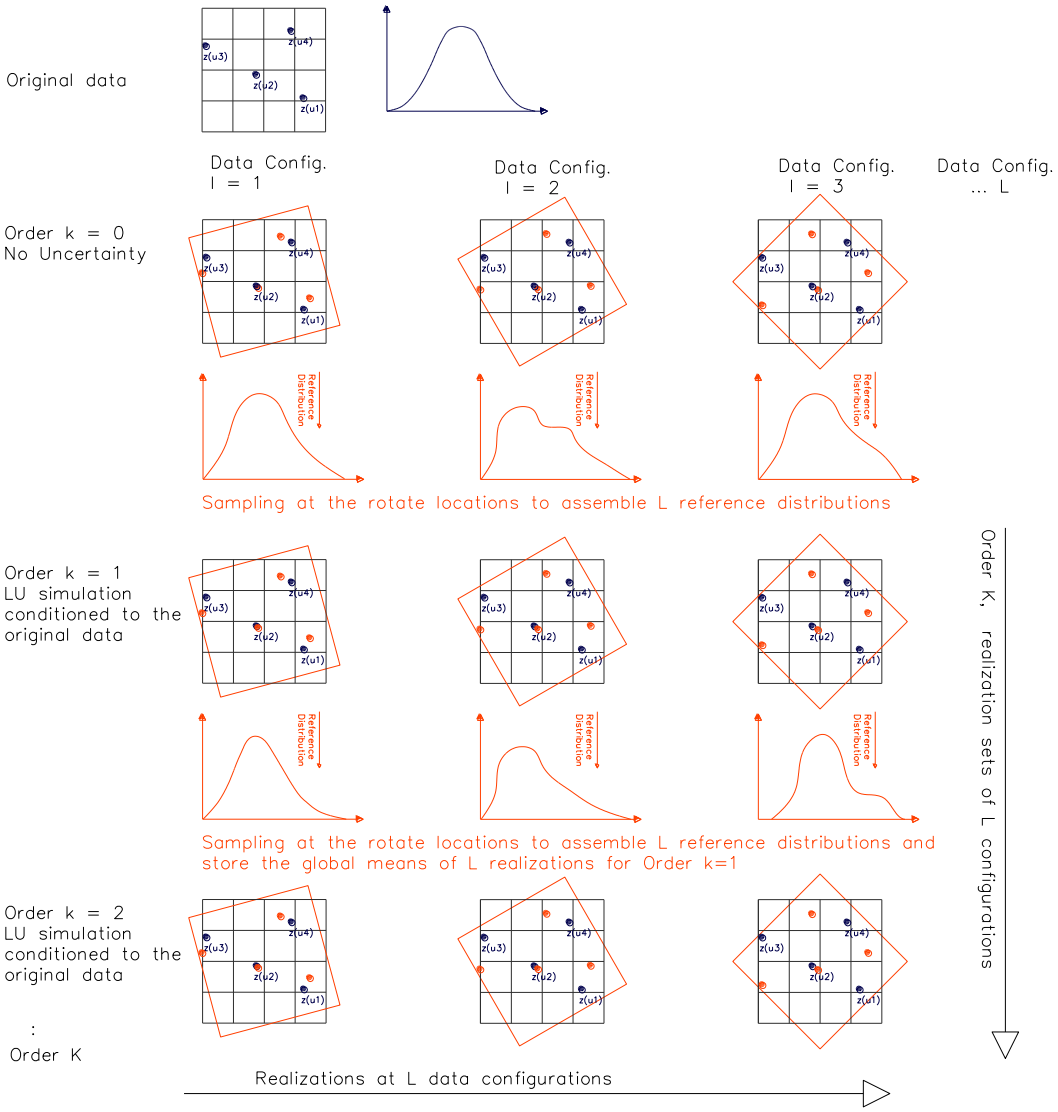


Figure 3.1: Sketch of the conditional finite domain process, blue points are locations of the conditioning data, orange points are set of new configurations and reference distribution is illustrated by each realization.

The simulations are done at l_j, $j = 1,...,L$ configurations locations conditioned to the original data. The L configurations are the same for every order and CFD technique sample and evaluate uncertainty of the mean in base on the global means of $L \times K$ conditional simulations.

3.3 Implementation

The data are not equally spaced and preferential sampling is often observed, that is, the histogram should use declustering weights. The discretization of the domain during evaluation ensures a representative global mean; however, the simulation in CFD is performed only at the points to be sampled and conditioned to the original data. Then, the declustering weights are relevant in the process. The theory about declustering methods is found in (Isaaks & Srivastava, 1989, pp. 235-248) among others.

Normal transformation of the original values is performed $y_i = G^{-1}(p_i)$, $i = 1,...,n$. Specification of the distribution tails are important to obtain a reasonable uncertainty in the input parameter.

The rotated and translated configurations of the original data should sample the whole domain. A representative sampling of the domain is relevant in the process to obtain realistic uncertainty in the mean. Then, this result is transferred to the simulation to improve a posterior evaluation of the global uncertainty and the decision making. The configuration of the data versus the shape of the domain must be checked before implementing CFD. The shape of the domain should permit representative re-sampling. Specific strategies of re-sampling should be done in some scenarios. For instance, sometimes irregular domain shape is observed in some complex deposits. Then, the translation and rotation of the sampling configurations could be restricted to a narrow space between sample locations and the boundary of the domain. Figure 3.2 shows a graphic of an irregular domain, where the samples $z(\mathbf{u}_i)$, $i = 1,...,6$ are located close to the boundary of the domain.

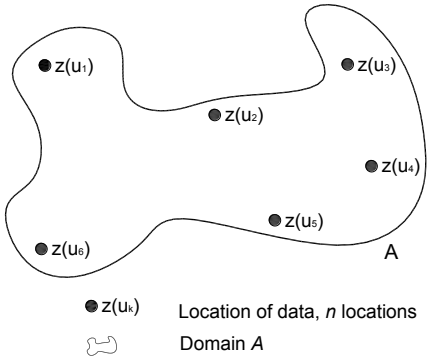


Figure 3.2: Example of domain that could truncate rotated and translated configurations.

The new configurations are truncated to the domain A. Then, all the new strategies of sampling will be close to one another. The uncertainty of the mean would be very small and not realistic.

The uncertainty in the mean using CFD often is greater than conventional bootstrap and spatial bootstrap; however, sometimes conditioning leads to less uncertainty than the SB. The CFD results are more reasonable because of the conditioning to original samples. An example of low uncertainty is observed in a domain that shows low variability or the values do not show strong anisotropy. Also, the size of the domain influences on the uncertainty in the mean. Applications of the methodology show that the uncertainty in the mean increases as the size of the domain is expanded with the same available data.

3.4 Application and Challenges

The methodology is applied and comparison with traditional techniques is evaluated. The influence of the size of domain on the uncertainty in the mean is verified. The data set that we use for this evaluation is *red data*; this file is available in the CCG network. The data have 68 samples of gold, silver, copper and zinc and are illustrated in Figure 3.3.

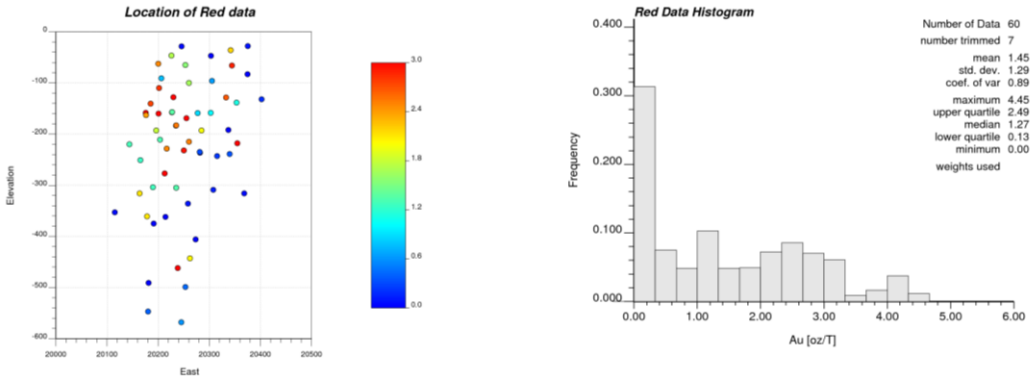


Figure 3.3: The spatial location of the Red data on the left side and the distribution of the data on the right side.

The thickness of the samples is between 0.13 meter and 18.86 meters. Eight samples with no metal no thickness values were removed. The spatial distance between samples are about 30 meters, this value is used to define the cell size of the declustering process. A variogram model of gold values in normal score units is required because the spatial

bootstrap and conditional finite domain work in Gaussian space. Figure 3.4 shows the experimental and variogram model.

$$\gamma(\mathbf{h}) = 0.44 Exp_{\substack{ah1=100 \\ ah2=90}}(\mathbf{h}) + 0.56 Exp_{\substack{ah1=250 \\ ah2=95}}(\mathbf{h})$$

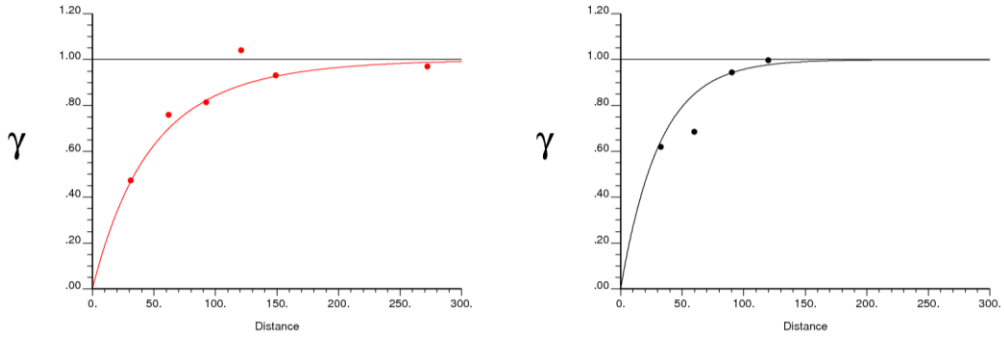


Figure 3.4: The left exponential variogram model corresponds to the major axis and the right one corresponds to the minor axis.

Two structures were required to model the experimental variogram, *ah1* is the major range in the 15° azimuth and *ah2* is the minor range in perpendicular direction to *ah1*. The program *cfdlu* was developed to apply the methodology. One hundred *L* new configurations are used to sampling the domain. The domain could be defined by polygons in each plane or by specified the space where the centroid of the data spatial location could be displaced, more information about that space or window is written in the Appendix of the thesis.

One hundred orders are performed for this application. For each k_i order l_j, j = 1,...,100 realizations are performed at 60 points conditioned to original data. One thousand realizations are performed, each one provides a global mean and the deviation of these means is the uncertainty in the mean of the univariate input distribution. Conventional Bootstrap and Spatial Bootstrap are set with 10000 realizations as well. Figure 3.5 shows the shape of the uncertainty in the mean generated by traditional techniques and CFD against the univariate input distribution shape.

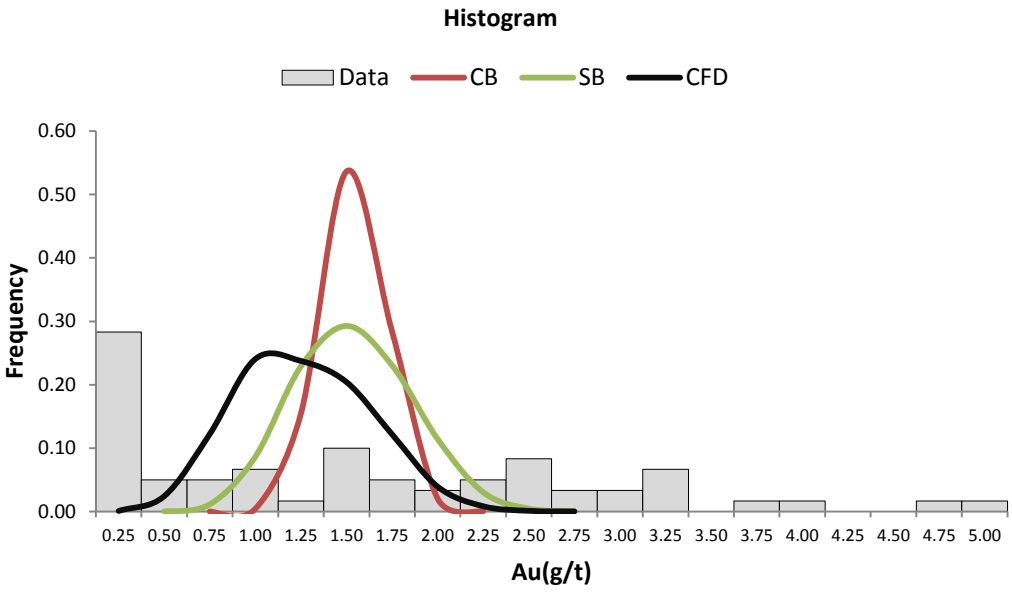
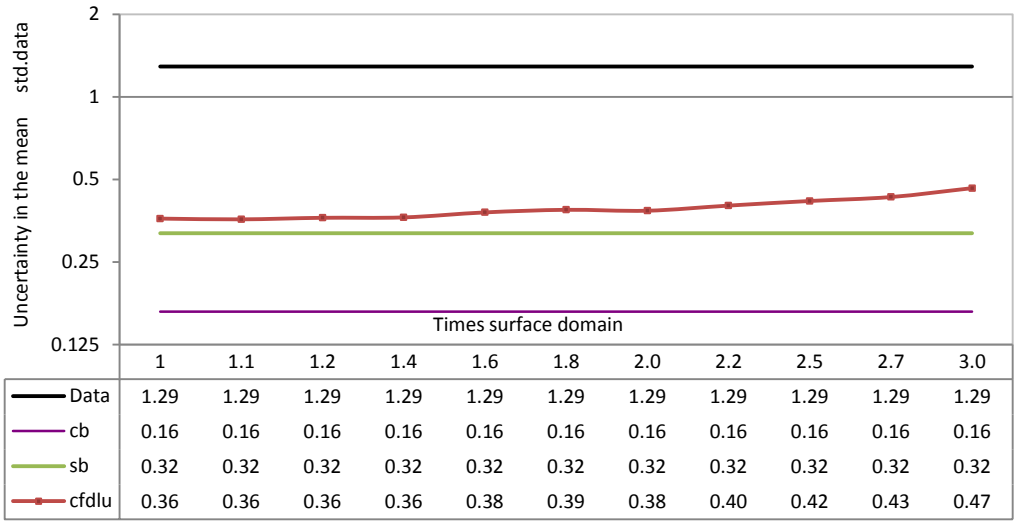


Figure 3.5: The Histogram of the original Red data is overlapped with three distributions of 10000 means that define the uncertainty in its mean. Distribution in red line is the results of CB, distribution in green line is the results of SB and distribution in black line is the results of CFD.

The weighted mean of the input distribution is 1.45 g/t and the standard deviation is 1.29, conventional bootstrap ($1.29/\sqrt{60}$) obtain a uncertainty in this mean 0.16 equivalent to 10% of the standard deviation of the data, the spatial bootstrap that consider the spatial correlation between the 60 values obtain an uncertainty in the mean 0.32 equivalent to 20% of the standard deviation of the data and finally the methodology of the chapter obtain 0.36 equivalent to the 30% of the standard deviation of the data. The means of 10000 realizations of both CB and SB are almost the same to the mean of the univariate input distribution; however, the mean of 10000 realizations in CFD is a little less because the domain contain almost 30% of the data between the range 0 g/t - 0.25 g/t and those values are located in the limits of the configuration of the data. Extrapolations of these values are made until the limit of the domain is reached.

For the example, we assume the domain *A* of the Red data as a bounded polygon around the data. This domain is expanded 10 times, that is, the 100 configurations will have more area to sampling simulated points conditioned to the original data. Since the sampling configuration is more distant from the conditioning data, an increase of uncertainty is

expected because the sampled points are less correlated. Figure 3.6 shows the impact of change the size of the domain *A* on the evaluation of uncertainty in the mean.



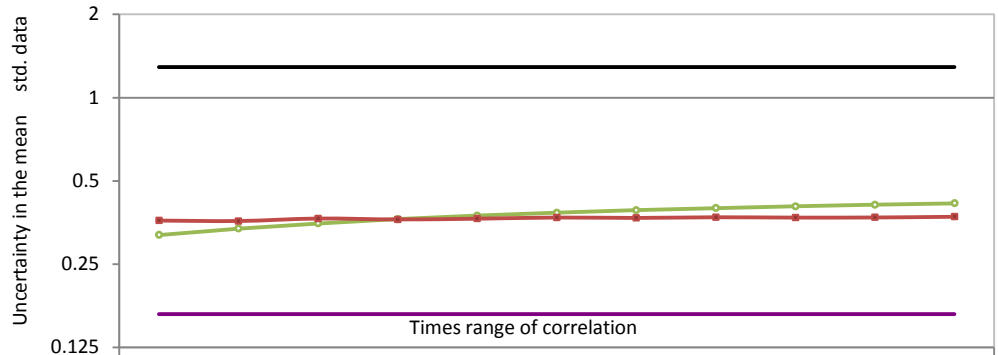
	1	1.1	1.2	1.4	1.6	1.8	2.0	2.2	2.5	2.7	3.0
Data	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29
cb	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
sb	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32	0.32
cfdlu	0.36	0.36	0.36	0.36	0.38	0.39	0.38	0.40	0.42	0.43	0.47

Figure 3.6: Sensitivity analysis of the uncertainty (evaluated by different techniques) with respect to the change of domain size. The standard deviation of the data is illustrated to compare to the uncertainty of the means by CB, SB and CFD techniques.

The uncertainty in the mean (calculated by CFD methodology) increases from 0.36 to 0.47 when the domain is expanded three times. As expected, the CB and SB do not change as the size of the domain is increased. The sampling assuming independence or accounting for the spatial correlation among samples are not enough when the domain is expanded without support of new data. All the scenarios require a realistic evaluation of uncertainty and the CFD technique is more robust in domains where the data support only a small percentage of the domain. In practice, zones without support of data are known like potential resources.

CFD techniques looks more robust to the sensitivity analysis of size of domain and range of correlation, the range of correlation should impact the uncertainty in the mean, more correlated samples should give less uncertainty; however, the SB shows opposite results because as the range of correlation become similar to and bigger than the size of the domain, ergodic fluctuations take place. The relation between size of the domain and range of correlation should be at least more than 10 times to avoid ergodic fluctuations.

More details about ergodic fluctuations are explained in Appendix A of the thesis. Figure 3.7 shows the influence of the range of correlation on the uncertainty in the mean.



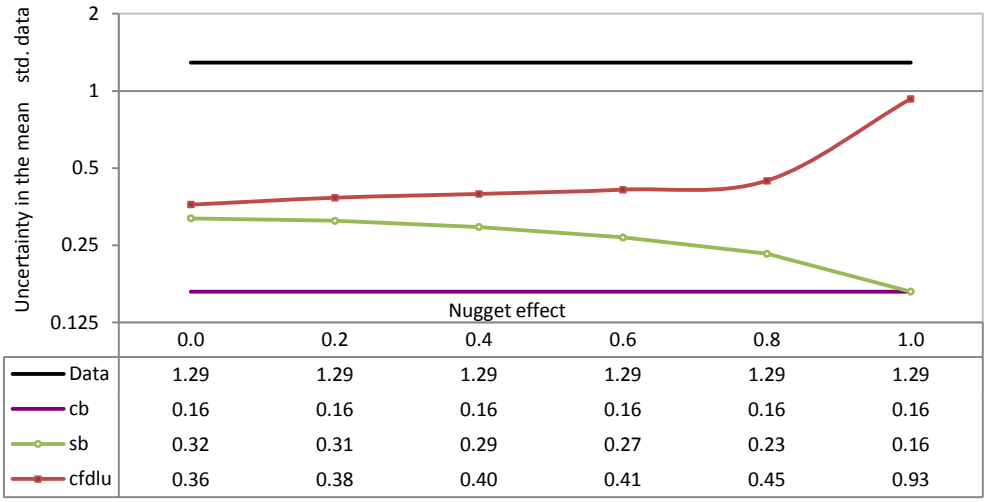
	1	1.1	1.2	1.4	1.6	1.8	2.0	2.2	2.5	2.7	3.0
Data	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29	1.29
cb	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
sb	0.32	0.34	0.35	0.36	0.37	0.38	0.39	0.40	0.40	0.41	0.41
cfdlu	0.36	0.36	0.37	0.36	0.36	0.37	0.37	0.37	0.37	0.37	0.37

Figure 3.7: Sensitivity analysis of the uncertainty with respect to the change of the correlation range. The evaluation of uncertainty (std) in the mean of the input univariate distribution is made by different techniques.

The conventional bootstrap does not change because only the data locations are used to sampling without taking account the spatial correlation among data, their value of uncertainty in the mean 0.16 is the same for all the evaluation of sensitivity analysis showed before. By the other hand, the SB execute 10000 realizations, sample 60 values from the cdf of the input distribution and use the covariance matrix between data to correlated the sampled values. The increase of range of correlation in SB leads to increase in uncertainty from 0.32 to 0.41 because ergodic fluctuations make the results no representatives. The uncertainty with CFD is almost the same from 0.36 to 0.37 when the range of correlation increase until it reaches 3 times the original range.

Sometimes, the variogram will not tend to zero as **h** does. It means that in small distance **h** the result of sampling could be different. Poor analytical precision, poor sample preparation or highly erratic mineralization at low scale could lead to nugget effect C_o (David, 1977). This C_o component of the variogram measure the independence of the data, for instance, when the nugget effect is equal to the variance or to one in variogram

with normal score units, the spatial bootstrap results 0.16 equal to conventional bootstrap because the data become independent. The increase of nugget effect indicates less correlation between samples, and then an increase of uncertainty in the mean is expected. The conditional finite domain gives an uncertainty that rises as the samples become less correlated. Figure 3.8 shows the sensitivity analysis of the uncertainty in the mean with respect to the change of nugget effect.



	0.0	0.2	0.4	0.6	0.8	1.0
Data	1.29	1.29	1.29	1.29	1.29	1.29
cb	0.16	0.16	0.16	0.16	0.16	0.16
sb	0.32	0.31	0.29	0.27	0.23	0.16
cfdlu	0.36	0.38	0.40	0.41	0.45	0.93

Figure 3.8: Sensitivity analysis of the uncertainty with respect to the change of the nugget effect. The evaluation of uncertainty (std) in the mean of the input univariate distribution is made by different techniques.

The CFD shows more robustness in many scenarios; however, many parameters are required to set the methodology. Also, many realizations at different configurations and orders could be computationally intensive with big set of data even if CFD simulate only in the points to be sampled. An evaluation of uncertainty in the mean should be simple as CB and SB and robust as CFD. Another technique is proposed in the next chapter where the uncertainty in the mean is acquired by the use of a stochastic trend. The use of a trend equation relaxes the assumption of stationarity and defines the mean dependent on the sample location at the domain.

Chapter 4

Stochastic Trend Approach

Estimates are made under uncertainty because few data are present in the evaluations and many input parameters are required in the estimates. The uncertainty in the input parameters must be transferred into the estimates to provide a more realistic assessment of uncertainty. Decision making often considers the uncertainty. The chance of success/failure, potential gain and potential loss are some considerations in risk analyses. These considerations rely on the estimates and their uncertainty (Rose, 2001). We are motivated to accurately evaluate the uncertainty of our estimates.

The conditional finite domain requires the setting of many parameters that affect the distribution of uncertainty in a non-intuitive and non-transparent manner. Relatively minor changes in geostatistical parameters could have a large effect on probabilistic estimates (Deutsch, Leuangthong, & Ortiz C., 2006). Sensitivity analysis shows that an increase in the nugget effect, a reduction of the range of correlation or an increase in the size of the domain leads to less uncertainty in large scale averages. This makes statistical sense because random variations in the variable average out.

A domain is an area or volume where samples follow a known distribution (e.g., normal or lognormal distribution). The domain includes samples that are not often uniformly spaced; preferential sampling is common in places of high grade. Geostatistics techniques often lead to less uncertainty for larger domains even if the number of data stays the same. An alternative to evaluate uncertainty in these domains could be to relax the assumption of stationarity. Stationarity permits calculation of the mean, covariance and semi variogram by pooling the data within the chosen domain (Goovaerts, 1997). The use of a trend equation relaxes this assumption of stationarity and defines the mean dependent

on the sample location within the domain, see Equation (4.1). L symbolizes the number of drift or trend terms, a_l represents the unknown coefficients and the f_l (**u**) terms are functionals that represent the shape of the trend. The functional may be linear, quadratic, sine/cosine, drawn by hand or specified arbitrarily.

$$m(\mathbf{u}) = \sum_{l=0}^{L} a_l f_l(\mathbf{u}) \tag{4.1}$$

Although the linear and the quadratic equations are considered in the implementation here, different equations could be used if required. A FORTRAN code called *uncregcoef.for,* which is a modification of the *correlate* program, was developed to implement the stochastic trend technique. The trend model could be used to predict the value at unsampled locations, that is, nodes or locations in the domain are evaluated with a polynomial with coefficients computed by least square method. A global mean is then evaluated with the set of coefficients evaluated at all locations within the domain. The stochastic trend (ST) approach developed below proposes to randomize these coefficients taking account the correlation of the original fitted coefficients. Many sets of coefficients provide different mean values. Then, the uncertainty in the input parameter is calculated from the distribution of these means.

4.1 Methodology

The coefficients of the trend are calculated based on the well-known linear regression theory. The sum of the squared difference between the estimates and the available data are as small as possible. This difference is not zero because the estimated value fluctuates about its expected value, that is, the method of least square selects the regression coefficients with the criteria of minimizing the sum square of these fluctuations (Carl Friedrich Gauss, 1794), (Johnson & Wichern, 2007)

The coefficients of the trend and their individual variances are calculated by the theory of multiple regression models. The linear regression model is defined by the equation matrix (4.2). The dependent variable is denoted as **Z** and the independent variables **X**. The dependent variable in ST is the data values and the independent variables are the coordinates of the data. The model could be written in matrix notation as:

$$\mathbf{Z} = \mathbf{X}\mathbf{a} + \boldsymbol{\varepsilon} \tag{4.2}$$

The dependent variable **Z** represents the ($n \times 1$) vector of the data, where n is the number of data available. **X** is the ($n \times L$) matrix of the levels of the independent variables, where L is the number of parameters, explanatory variables or regressor predictor that define the fitted equation of the model equation. For instance, if the data have coordinates at East, North and Elevation, the (n , 2), (n , 3) and (n , 4) of matrix **X** may correspond to the coordinate data. Then, (n , 5 - L) columns are interaction of the coordinates to better define the shape of the trend with more parameters. **a** corresponds to the ($L \times 1$) vector of the coefficients or vector of regression parameters. The last element **ε** regards to an error ($n \times 1$) vector, this error vector is assumed as a random part of the model equation that has a distribution of mean zero and unknown variance s^2. The next equation shows the explained variables:

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1L} \\ 1 & x_{21} & x_{22} & \cdots & x_{2L} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nL} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_L \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{4.3}$$

$$\mathbf{Z} \quad = \quad \mathbf{X} \qquad\qquad \mathbf{a} \quad \boldsymbol{\varepsilon}$$

The location of the available data is translated to the **X** matrix based on the specified functional polynomials or trend model. The method of least squares is commonly used to estimate the regression coefficients in a multiple linear regression models, the term *linear* because the model is a linear function of the unknown parameters a_0, a_1,...,a_L. The model describes a hyper-plane in L-dimensional space of the regressor variables $\{x_l\}$ (Montgomery, 2000). The coefficients selected by the least square are called least squared estimates of the regression parameter **a**. They are denoted by **â** because they are estimates of **a**.

$$\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} \tag{4.4}$$

Then, the fitted regression equation is given in matrix notation. This fitted model predicts the value at unsampled locations as a function of its coordinates:

$$\hat{\mathbf{Z}} = \mathbf{X}\hat{\mathbf{a}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} \tag{4.5}$$

The residual describes the error in the fitting of the trend model to the n samples z_i, i = 1,...,n. The vector of residuals is the difference between the original value samples and the ones calculated by the fitted equation at known sample locations. The variance of the residuals (4.7) is the sum of the squared residuals divided by the (n - (L+1)) degree of freedom, where (L+1) is the number of parameters or coefficients. Notice that the number of data must be greater than number of coefficients.

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Z} - \hat{\mathbf{Z}} \tag{4.6}$$

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}}{n - (L+1)} \tag{4.7}$$

$$E\{s^2\} = \sigma^2 \tag{4.8}$$

The method of least squares produces an unbiased estimator of the parameters **a** in the multivariate linear regression model. Properties of the **â** estimators are defined below:

$$E\{\hat{\mathbf{a}}\} = \mathbf{a} \tag{4.9}$$

$$Cov(\hat{\mathbf{a}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \tag{4.10}$$

The next matrix shows the covariance between the regression coefficients. Then, the variances of the coefficients are acquired from the diagonal of the symmetric matrix because $C(a_l, a_l) = \sigma^2_{al}$

$$Cov(\hat{\mathbf{a}}) = \begin{bmatrix} c_{a_{00}} & c_{a_{01}} & \cdots & c_{a_{0L}} \\ c_{a_{10}} & c_{a_{11}} & \cdots & c_{a_{1L}} \\ \vdots & \vdots & \ddots & \vdots \\ c_{a_{L0}} & c_{a_{L1}} & \cdots & c_{a_{LL}} \end{bmatrix} = \begin{bmatrix} \sigma^2_{a_{00}} & c_{a_{01}} & \cdots & c_{a_{0L}} \\ c_{a_{10}} & \sigma^2_{a_{11}} & \cdots & c_{a_{1L}} \\ \vdots & \vdots & \ddots & \vdots \\ c_{a_{L0}} & c_{a_{L1}} & \cdots & \sigma^2_{a_{LL}} \end{bmatrix}$$

Now each term of the covariance is divided by their respective combined standard deviations to obtain the correlation matrix between the coefficients.

$$\rho(\hat{\mathbf{a}}) = \begin{bmatrix} 1/\sigma_{a_{00}} & 0 & 0 & 0 \\ 0 & 1/\sigma_{a_{11}} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_{a_{LL}} \end{bmatrix} \times \begin{bmatrix} \sigma^2_{a_{00}} & c_{a_{01}} & \cdots & c_{a_{0L}} \\ c_{a_{10}} & \sigma^2_{a_{11}} & \cdots & c_{a_{1L}} \\ \vdots & \vdots & \ddots & \vdots \\ c_{a_{L0}} & c_{a_{L1}} & \cdots & \sigma^2_{a_{LL}} \end{bmatrix} \times \begin{bmatrix} 1/\sigma_{a_{00}} & 0 & 0 & 0 \\ 0 & 1/\sigma_{a_{11}} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1/\sigma_{a_{LL}} \end{bmatrix}$$

$$\rho(\hat{\mathbf{a}}) = \begin{bmatrix} 1 & \rho_{a_{01}} & \cdots & \rho_{a_{0L}} \\ \rho_{a_{10}} & 1 & \cdots & \rho_{a_{1L}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{a_{L0}} & \rho_{a_{L1}} & \cdots & 1 \end{bmatrix} \tag{4.11}$$

Many sets of coefficients could be generated by Monte Carlo simulation to evaluate uncertainty in the fitted trend. The correlation between the regression coefficients will be preserved. Each coefficient $\hat{a}_l$, $l = 0, 1,...,L$ has a distribution defined by its mean and its variance. Then, many sets of coefficients are drawn from their distributions. Those sets take account the correlation of the original regression coefficients. Otherwise, different techniques like spatial bootstrap use the covariance matrix of the sample locations ($\mathbf{C}_{11}$) to conditional the sampling and evaluate uncertainty. Both matrices are symmetric and positive definite. The steps of the proposed stochastic trend (ST) approach are as follows:

- Define the number of parameters or terms L for the trend equation, the model could be linear or quadratic.
- Compute the regression coefficients of the trend by least square method $\hat{a}_l$, $l = 0, 1,...,L$ where the best fit minimizes the sum of squared residuals.
- Define the covariance matrix and deduce the variance of the coefficients σ_{al}, $l = 0, 1,...,L$.
- Define the correlation matrix for the regression coefficients.
- Perform Cholesky decomposition of the correlation matrix $\boldsymbol{\rho} = \mathbf{LU}$
- Sample independent normal Gaussian score values $w = G^{-1}(p)$ and correlate them by the use of the lower decomposed correlation matrix $\mathbf{L}$, $\mathbf{Y} = \mathbf{L}\,\mathbf{w}$.
- Non-standardize these $\mathbf{Y}$ values by the use of their respective fitted regression coefficient a_l and their respective standard deviation σ_{al}.

$$a_l^k = \hat{a}_l + \sigma_{a_l} \times y_l^k, \quad l = 0,\ 1,...,L \tag{4.12}$$

- Use the set of coefficients in the polynomial to evaluate the fitted mean $Z(\mathbf{u})$ at all locations within the domain. The mean of the values at these locations is the mean of the first realization m_k, $k = 1,...,K$ (where K is the number of times the workflow is repeated)

The distribution of m_k means can be assembled to model the uncertainty in the mean.

4.2 Criteria in the Implementation

All techniques to calculate the uncertainty in the mean require implementation choices. The stochastic trend approach does not require as many input parameters as most techniques. This approach only requires the form of the trend model. One should choose the polynomial representation that appears reasonable given the data. The uncertainty in the mean will be sensitive to the number of terms.

The uncertainty in the mean using the stochastic trend approach does not require a variogram; the variogram is required for other techniques like the spatial bootstrap and the conditional finite domain.

The fitted regression equation or trend model is used to predict the mean at all locations. The linear relationship may not be necessarily valid for extrapolation purposes (Montgomery & Runger, 2006). Uncertainty in the regression model is used specifically to calculate the uncertainty in the mean. The generation of the stochastic trends is explained with a simple example. A synthetic data set of 11 values is located in one dimension. The mean of the values is 2.409 and the variance 0.695.

$$\begin{bmatrix} 2.0 \\ 1.8 \\ 1.4 \\ 1.2 \\ 1.8 \\ 2.4 \\ 2.6 \\ 2.8 \\ 3.0 \\ 3.5 \\ 4.0 \end{bmatrix} = \begin{bmatrix} 1 & 5 \\ 1 & 10 \\ 1 & 16 \\ 1 & 22 \\ 1 & 32 \\ 1 & 45 \\ 1 & 60 \\ 1 & 70 \\ 1 & 80 \\ 1 & 85 \\ 1 & 95 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{11} \end{bmatrix}$$

$$\mathbf{Z} = \mathbf{X}\ \mathbf{a}\ \boldsymbol{\varepsilon}$$

The trend of the data is model with a linear trend that considers only two coefficients a_0 and a_1. The location of the data is transferred to matrix $\mathbf{X}(11 \times 2)$ according the linear trend model, where the East coordinate of the data correspond to the column two. The vector $\mathbf{Z}$ with dimension (11 , 1) represent the values of the data set. The least squares method gives the estimates $\hat{a}_0$ and $\hat{a}_1$ of the parameters a_0 and a_1 through the simplified operations of matrices. Then, the first coefficient is 1.2403 and second coefficient is 0.0247.

$$\begin{aligned} \hat{\mathbf{a}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z} \\ &= \begin{bmatrix} 1.2403 \\ 0.0247 \end{bmatrix} \end{aligned}$$

The trend model is defined by the equation that evaluates $Z(\mathbf{u})$ as a function of location in one dimension.

$$Z(\mathbf{u}) = 1.2403 + 0.0247x$$

The data location is replaced in the regression line to calculate the $\boldsymbol{\varepsilon}$ residual or deviation of the data from the estimated regression model. Then, the variance of the residuals σ^2 is as follow:

$$\sigma^2 = \frac{\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}}{n-p} = \frac{1.21}{11-2} = 0.134$$

The covariance of the coefficients is defined and the variance of the regression coefficients are extracted from the diagonal of this $Cov(\hat{\mathbf{a}})$.

$$Cov(\hat{\mathbf{a}}) = \sigma^2(\mathbf{X}\mathbf{X}^T)^{-1} = \begin{bmatrix} 4.19E-02 & -6.22E-04 \\ -6.22E-04 & 1.32E-05 \end{bmatrix}$$

Then, each term of the covariance matrix is divided by their respective standard deviations to obtain the correlation matrix.

$$\rho = \begin{bmatrix} \dfrac{C_{a_{00}}}{\sigma_{a_0}\sigma_{a_0}} & \dfrac{C_{a_{01}}}{\sigma_{a_0}\sigma_{a_1}} \\ \dfrac{C_{a_{10}}}{\sigma_{a_1}\sigma_{a_0}} & \dfrac{C_{a_{11}}}{\sigma_{a_1}\sigma_{a_1}} \end{bmatrix} = \begin{bmatrix} \rho_{a_{00}} & \rho_{a_{01}} \\ \rho_{a_{10}} & \rho_{a_{11}} \end{bmatrix} = \begin{bmatrix} 1 & -0.837 \\ -0.837 & 1 \end{bmatrix}$$

The correlation matrix is symmetric and is positive-definite, then, the Cholesky decomposition is possible. The lower triangular matrix helps to correlate the independent normal values w_0 and w_1.

$$\begin{bmatrix} y_0^k \\ y_1^k \end{bmatrix} = \underbrace{\begin{bmatrix} \rho_{a_{00}} & 0 \\ \rho_{a_{10}} & \sqrt{1-\rho_{a_{10}}^2} \end{bmatrix}}_{\textit{Lower matrix}} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -0.837 & \sqrt{1-(-0.837)^2} \end{bmatrix} \begin{bmatrix} 0.772 \\ 0.279 \end{bmatrix}$$

The k realization of y_0^k and y_1^kcoefficients are in Gaussian units. These are non-standardized to get a_0^k and a_1^k in original units.

$$a_0^k = a_0 + \sigma_{a_0} y_0^k = 1.2403 + 0.2048 \times 0.772$$
$$a_1^k = a_1 + \sigma_{a_1} y_1^k = 0.0247 + 0.0036 \times (-0.493)$$

The values of the first realization k = 1 correspond to the set of coefficients a_0^1 = 1.398 and a_1^1 = 0.023. Then, one hundred times k = 1,...,100 are sampled from the distribution of regression coefficients with mean $\hat{a}_0$ and $\hat{a}_1$ and standard deviation σ_{a0} and σ_{a1}. The original correlation of the coefficients is -0.837, and then the correlation between the two coefficients after one hundred realizations is preserved. The scatter plot of the first stochastic trend coefficients is illustrated in a red dot and the next ones are illustrated in black small dots see Figure 4.1.

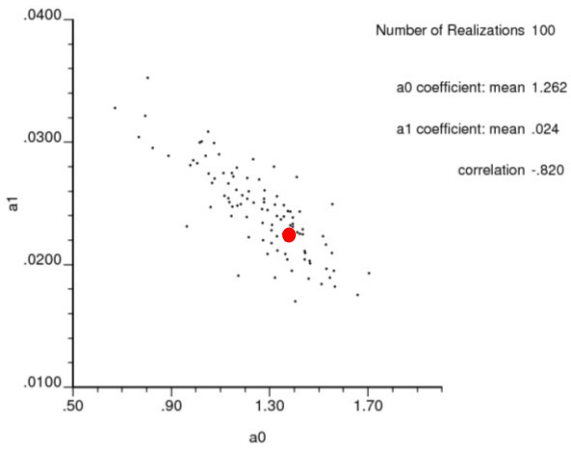


Figure 4.1: Verification whether the correlation of the 100 realizations of $\hat{a}_0$ and $\hat{a}_1$ reproduce the input correlation -0.837 between the regression coefficients, the red dot correspond to the first realization developed before.

One hundred sets of coefficients are used in the trend model. Then, one hundred stochastic trends are defined. A trend model is used to evaluate the value at all locations and the corresponding global mean. The fluctuations of the trend provide one hundred means. The fluctuations of the trend are illustrated in Figure 4.2.

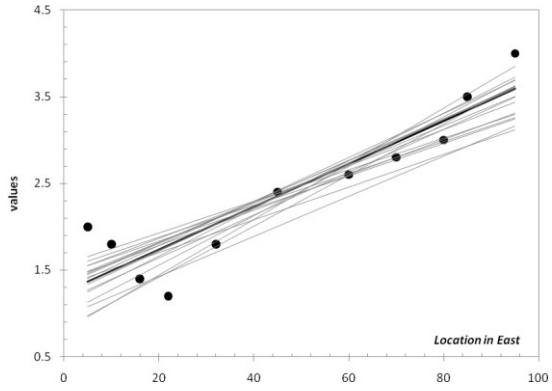


Figure 4.2: Stochastic trend model, where the black points are the data; the solid black line is the linear trend model with two parameters; and the grey lines are the realizations.

One hundred stochastic trend models provide an uncertainty of 0.114. This value represent 45 % of the uncertainty assuming independence of the data $\sigma/\sqrt{n} = 0.251$.

4.3 Application and Challenges

The stochastic trend is applied to a 3-D data set. Also, the influence of variable domains on the uncertainty and the complexity of the model in the uncertainty are evaluated. The gold values are located in a domain of 200 meters by 200 meters in the horizontal direction and 20 meters in the vertical direction. The mean of the gold values is 1.10 g/t and the standard deviation is 1.4. The evaluations consider three different domains:

- Pessimistic criteria, the domain which limits are less than half distance between samples.
- Normal criteria, the reasonable domain which limits are around the half of the median distance between samples.
- Optimistic criteria, the domain which limits are excessively far from the sample, it means almost twice or more than the median distance between samples.

The uncertainty of the gold values is calculated by a quadratic trend of 10 coefficients and a linear trend of 4 coefficients.

$$Z(\mathbf{u}) = a_0 + a_1 x + a_2 y + a_3 z + a_4 x^2 + a_5 y^2 + a_6 z^2 + a_7 xy + a_8 xz + a_9 yz$$
$$Z(\mathbf{u}) = a_0 + a_1 x + a_2 y + a_3 z$$

The variables x, y and z are the coordinates East, North and Elevation of the vector location $\mathbf{u}$. These values are transferred to the matrix $\mathbf{X}$ to obtain the regression coefficients, $\hat{\mathbf{a}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$. The vector $\mathbf{Z}$ contains the 49 gold values. The regression coefficients are computed for both trend models.

$$Z(\mathbf{u}) = 8.554 - 0.0029x + 0.0685y - 0.7056z + 0.0001x^2 + 0.0003y^2 + 0.0092z^2 - 0.0008xy + 0.0024xz - 0.0017yz$$
$$Z(\mathbf{u}) = 5.403 - 0.0008x + 0.004y - 0.153z$$

Each trend model requires a covariance matrix and a correlation matrix of the regression coefficients. One hundred realizations of the coefficients are simulated with the LU methodology described above. Sets of the coefficients $\hat{\mathbf{a}}$ are replaced in their trend models to estimate the gold value at each node of the domain. The uncertainty with the quadratic trend results greater than the linear trend for the three domains.

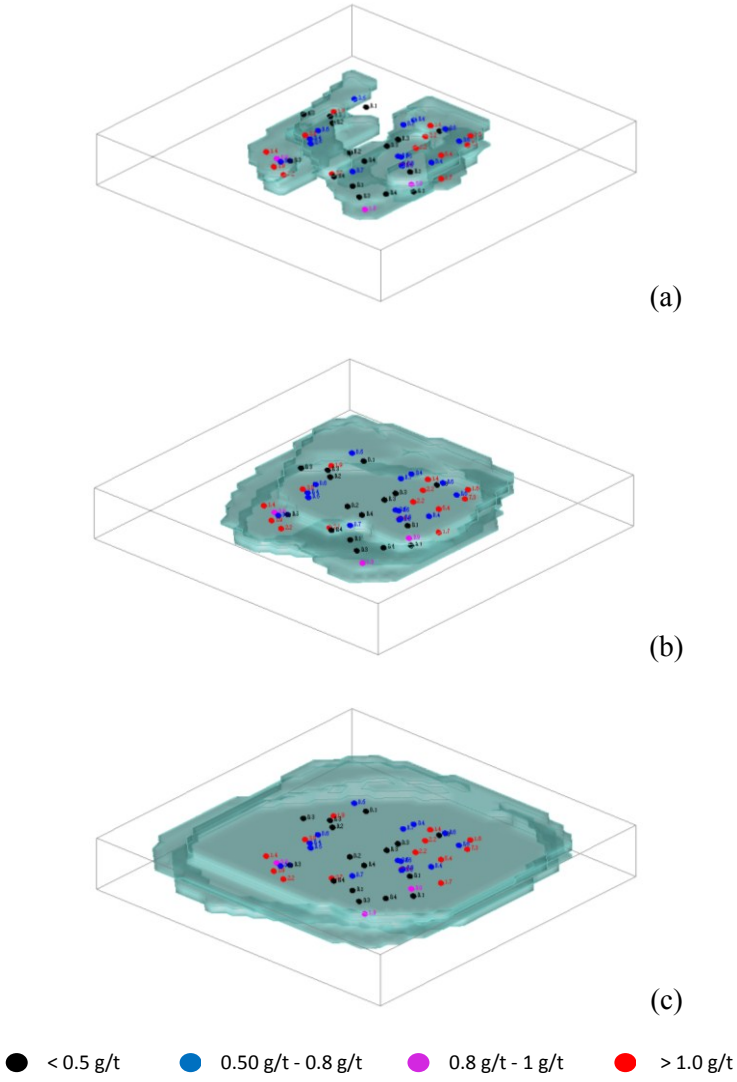


Figure 4.3: The data are illustrated with points, the first graphic (a) shows the pessimistic criteria used to model the domain, the second graphic (b) shows the reasonable criteria and the third graphic (c) shows the optimistic criteria.

Forty-nine data of gold values located in a pessimistic domain give an uncertainty of 0.227 (std) using a linear trend and 0.515 (std) using a quadratic trend. Those values are greater than $1.4/\sqrt{49} = 0.20$ (std) with CB that assumes independence. As expected, the uncertainty increases as the domain increases. The optimistic (large) domain provides an uncertainty of 0.234 using linear trend and 0.721 using quadratic trend. The sudden increase of uncertainty as the domain increases without support of data is an advantage of

the ST approach because traditional geostatistics techniques struggle with less uncertainty in this scenario. Table 4.1 shows the increase of uncertainty in the mean.

	Lineal Model	*Quadratic Model*
Pessimistic	0.227	0.515
Reasonable	0.229	0.531
Optimistic	0.234	0.721

Table 4.1: Uncertainty of the mean of the input distribution using different trend models.

The technique is simple to apply because it does not require a covariance model of the data and does not need to be set many parameters during the implementation. This technique considers the size of the domain and assume that the mean dependent of the location data; however, the specific form of the trend is an important choice. Moreover, the model parameters are assumed to be multivariate Gaussian.

Chapter 5

Transference of Uncertainty

Geostatistical simulation is usually performed with a fixed input distribution; this fixed univariate distribution comes from the input data and assumes a mean without uncertainty. An important aspect of this thesis is that uncertainty in the mean of the input distribution (UMID) must be transferred through simulation for a more complete understanding of uncertainty. Uncertainty in the mean (UMID) is evaluated with techniques introduced in Chapters 3 and 4. The techniques of conditional finite domain (CFD) and stochastic trend (ST) provide the UMID. Multiple distributions could be constructed and used in geostatistical simulation as reference distributions.

Simulation is performed in Gaussian space because a consistent multivariate distribution is required and the multivariate Gaussian is the only known practical multivariate distribution. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted by changing the reference distribution for transformation. A sequential Gaussian simulation (SGS) algorithm is adopted in this thesis; however, any Gaussian algorithm for simulation could be used. SGS is used because it is simple, flexible, and reasonably efficient (Deutsch, 2002).

A change in local and global uncertainty is expected when UMID is transferred through the simulation process. A measure of local uncertainty is available at every location by generating a set of L realizations:

$$z^{(l)}(\mathbf{u}), l = 1, \ldots$$

$$F(\mathbf{u}; z \mid (n)) = \text{Prob}\{Z(\mathbf{u}) \leq z \mid (n)\} \tag{5.1}$$

Local uncertainty could be used for planning and decision making; however, some applications require the uncertainty of more than one location simultaneously, a measure of the joint uncertainty about attribute values at several locations taken together. This spatial uncertainty is modeled by generating multiple realizations of the joint distribution of the attribute value (Goovaerts, 1997). Those realizations should reasonably reproduce the sample histogram and the semivariogram model. The set of simulated maps is generated by sampling the *N*-variate ccdf that models the joint uncertainty at the *N* locations $\mathbf{u}'_j$:

$$\left\{z^{(l)}\left(\mathbf{u}'_j\right), j=1,\ldots,N\right\}, l=1,\ldots,L$$

$$F\left(\mathbf{u}'_1,\ldots,\mathbf{u}'_N;z_1,\ldots,z_N\,|\,(n)\right)=\text{Prob}\left\{Z\left(\mathbf{u}'_1\right)\le z_1,\ldots,Z\left(\mathbf{u}'_N\right)\le z_N\,|\,(n)\right\} \quad (5.2)$$

Spatial uncertainty is a result of our incomplete knowledge of the spatial distribution of the variable of interest.

5.1 Methodology

The probability distributions of continuous data are often summarized by a central value such as the mean (Deutsch, 2000). The mean of the distribution is a fixed parameter in the simulation process. Where there are *n* data values $z(\mathbf{u}_i)$, $i = 1,...,n$ with different weights $w(\mathbf{u}_i)$, $i = 1,...,n$:

$$m_r=\frac{\sum_{i=1}^{n} z\left(\mathbf{u}_i\right)w\left(\mathbf{u}_i\right)}{\sum_{i=1}^{n} w\left(\mathbf{u}_i\right)}$$

The uncertainty in the mean of the original distribution (σ_m) comes from one of the techniques developed earlier such as the CFD or ST. The calculated fluctuations of the mean are summarized by distributions that have different mean values. The number of distributions or reference distributions in the simulation could be defined by *L* equally spaced quantiles:

$$p_l=\frac{l-0.5}{L},\quad l=1,...,L$$

The specific mean values corresponding to these quantiles are computed from a non standard Gaussian distribution computed to represent the UMID like standard deviation.

$$y_l = G^{-1}(p_l)$$
$$m_l = y_l \times \sigma_m + m_r$$

Figure 5.1 shows correspondence sketch of a quantile and the respective mean.

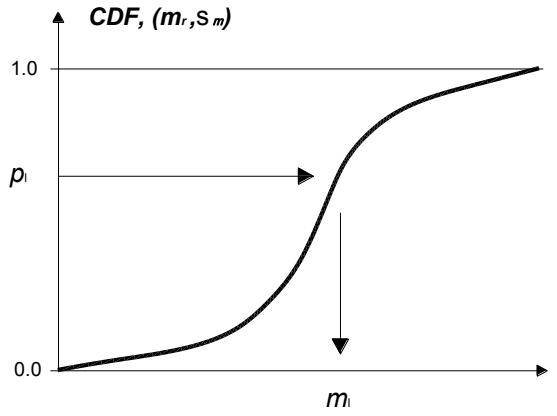


Figure 5.1: Sketch of the cumulative distribution function of distribution (m_r, σ_m), where $\sigma_m = S_m$.

Once the m_l values are calculated, the relation of these values and the mean of the original distribution (m_r) provide factors that are multiplied by each value of the original distribution. The factors (m_l / m_r) are ordered values because m_l corresponds to each quantile in the cdf. Then, the variable distributions have values that have the lower mean when $l = 1$ and the biggest mean when $l = L$.

$$z\left(\mathbf{u}_i^l\right) = z\left(\mathbf{u}_i\right) \times \frac{m_l}{m_r}, \quad i = 1,...,n$$

This factor is applied to strictly positive variables. Almost all data in the earth science are positive values including mineral grades, porosity and contaminant concentrations. The variance is not preserved in this transformation, but the most important statistic is the mean. The declustered weights of the variable distributions are the same as the original data because their spatial locations are the same.

5.2 Implementation

Simulation requires the original z-data to be transformed into y-values with a standard normal histogram, the normal score transform function can be derived through a graphical correspondence between the cumulative distribution of the original and standard normal variables (Goovaerts, 1997). The transformation process often uses the fixed ccdf of the original data; however, a different reference ccdf could be used. A simple way to transfer uncertainty in the mean through simulation is to use different reference distributions. An increase in global and local uncertainty is expected. A simple scenario explains the methodology, where a spherical semivariogram model is assumed, the distance between the $\mathbf{u}_{\square}$ location to be simulated and the sample $z(\mathbf{u}_1)$ corresponds to half of the range of the semivariogram ($a/2$). Since there is only one conditioning data, the conditional mean and variance is simplified to:

$$m_{c_{\square}} = \rho(\mathbf{h}) \cdot y(\mathbf{u}_1)$$
$$\sigma_{c_{\square}} = 1 - \rho(\mathbf{h})^2$$

Two scenarios are evaluated; the first scenario considers a fixed global distribution. The simulation uses the parameters of the original fixed global distribution to standardize its datum of 2.5 original units into normal scores. Then, the conditional mean kriging (0.313) and conditional variance kriging (0.902) are predicted for the unsampled location. An independent residual that follows a normal distribution with mean of zero and the conditional variance is drawn with classical Monte Carlo simulation. The simulated value is the addition of the conditional mean kriging and the residual for that location (Deutsch, 2002). Figure 5.2 shows the result of this simulation, where the output distribution for the unsampled location is illustrated.

The second scenario account for the simulation that uses a different reference distribution for the transformation of original values into normal scores and vice versa. The uncertainty in the mean of the input univariate distribution (UMID) has a standard deviation of 0.2. The sampled location $z(\mathbf{u}_1)$ will have different ccdfs (parameters) to transform to normal scores, the means of these ccdfs fluctuate equal to the UMID ; $y(\mathbf{u}_1)$ in normal score unit takes values from 0.66 to 1.34. The product of those values and the weight kriging (0.313) gives different mean values and a constant variance kriging

(0.902). Those values are sampled many times and back transformed into original units. The back transformation should use their respective transformation table matching the forward transform. Figure 5.3 shows the result of this simulation, where the output distribution for the $\mathbf{u}_{\square}$ unsampled location is wider than the previous simulation with fixed distribution.

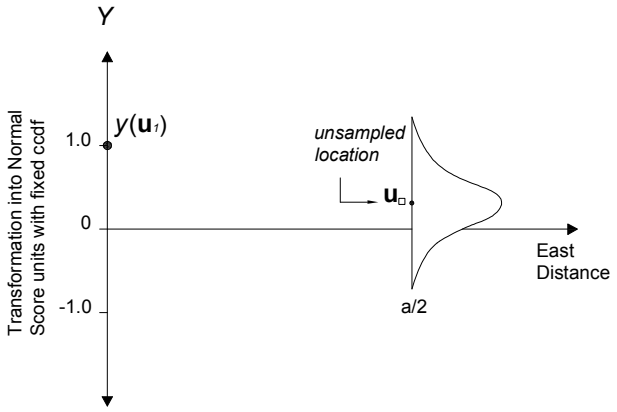


Figure 5.2: Sketch of simulation at one node using fixed ccdf $(1,1.5^2)$ like input parameter.

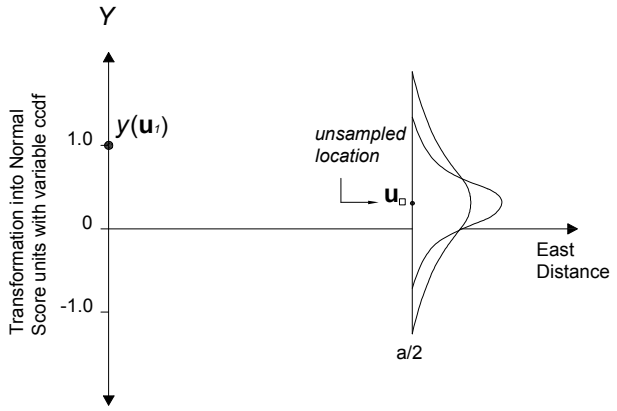


Figure 5.3: Sketch of simulation at one node using variable ccdf to transfer the uncertainty in the original distribution 0.2 to the simulation process.

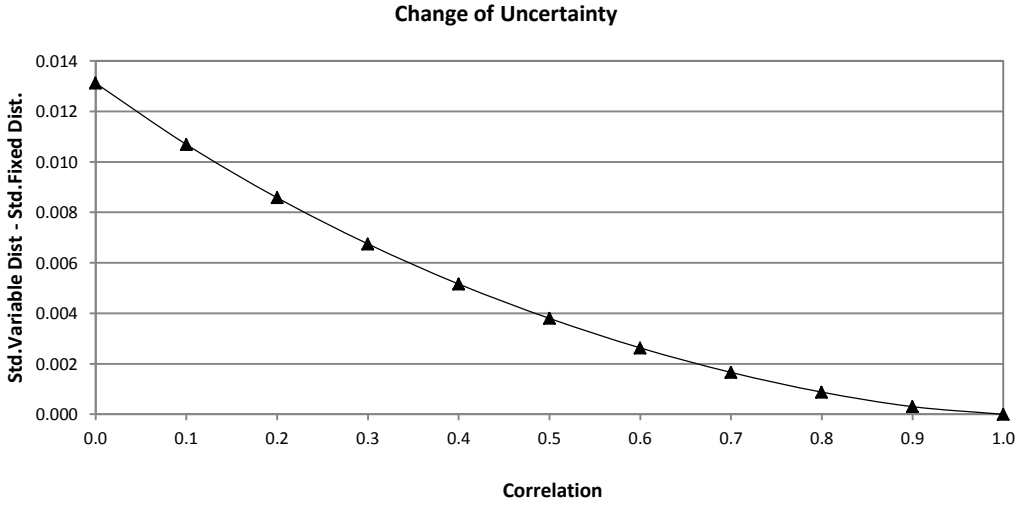
The number of reference distributions is denoted with the letter *L* and the number of realizations of every reference distribution is denoted with the letter *K*. The resulting mean and variance:

$$m_z = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K} Z_k^l(\mathbf{u}_{\square}) = 1.47 \quad \sigma^2_{z_{unc}} = \frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K}\left(Z_k^l(\mathbf{u}_{\square}) - m_z\right)^2$$

As expected, the distribution of the output mean using uncertainty in the input parameter is wider than using a fixed input parameter. Uncertainty in the sampled location with fixed reference distribution is 1.41 and with variable reference distributions is 1.55.

5.3 Sensitivity Analysis

One thousand simulations are performed with parameter uncertainty. Each simulation considers variable transforms into normal scores. One thousand quantiles are used to draw the residuals. The change of the correlation between the conditioning data and the unsampled location are evaluated. As expected, the results show that the uncertainty goes down when the correlation increases. Figure 5.4 shows the less increase in uncertainty as the spatial correlation increases.



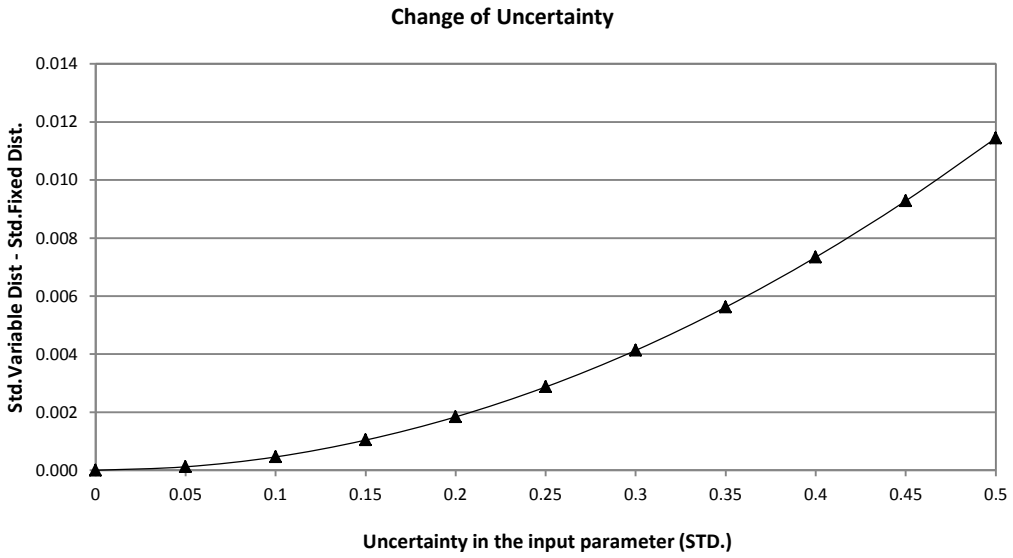
Data mean	1.00
Data STD	1.50
Uncertainty in mean of Input Distribution (STD.)	0.20
Conditioning value	1.00

Figure 5.4: Sensitivity analysis of the uncertainty with respect to the change of correlation, 1000 realizations are generated with fixed ccdf and with 100 variable ccdfs, the table shows the parameter used in the simulation where global mean and conditioning value are in original units.

The simulation with variable ccdfs and zero range of correlation show uncertainty of the node equal to 1.512; the simulation with fixed mean gives an uncertainty of 1.499.

Moreover the change of the conditioning sample value does not change the uncertainty when the simulation uses a fixed distribution. This is expected because, under a Gaussian model, errors are independent of the data values and dependent only on the data configuration (Goovaerts, 1997); however, a change of uncertainty in the node is observed when a lognormal distribution is used.

As expected, the uncertainty at the unsampled location increases as the uncertainty in the input parameter increases. Figure 5.5 shows the increase in uncertainty at the unsampled location as the input parameter uncertainty increases.



Data mean	1.00
Data STD	1.50
Conditioning value	1.00
Correlation	0.68

Figure 5.5: Sensitivity analysis of the uncertainty with respect to the change of UMID, 1000 realizations are generated with fixed cdf and with 100 variable cdfs, the table shows the parameters used in the simulation where global mean and conditioning value are in original units.

The same scenario of one conditioning sample is expanded to a grid of five nodes in the east direction and five nodes in the north direction. The size of the nodes is one unit. The change of the local uncertainty and global uncertainty is evaluated for this scenario. The uncertainty in the mean of the univariate distribution is accounted for in the generation of

1000 variables means. The sequential Gaussian simulation approach is applied. The conditioning data is located in the center of the domain; an exponential variogram model is used with range of 7.779 units because the covariance between the conditioning data and the closest node was set to 0.68. Figure 5.6 shows the change of the distribution of the global means when uncertainty in the mean of the distribution is incorporated to the process of simulation.

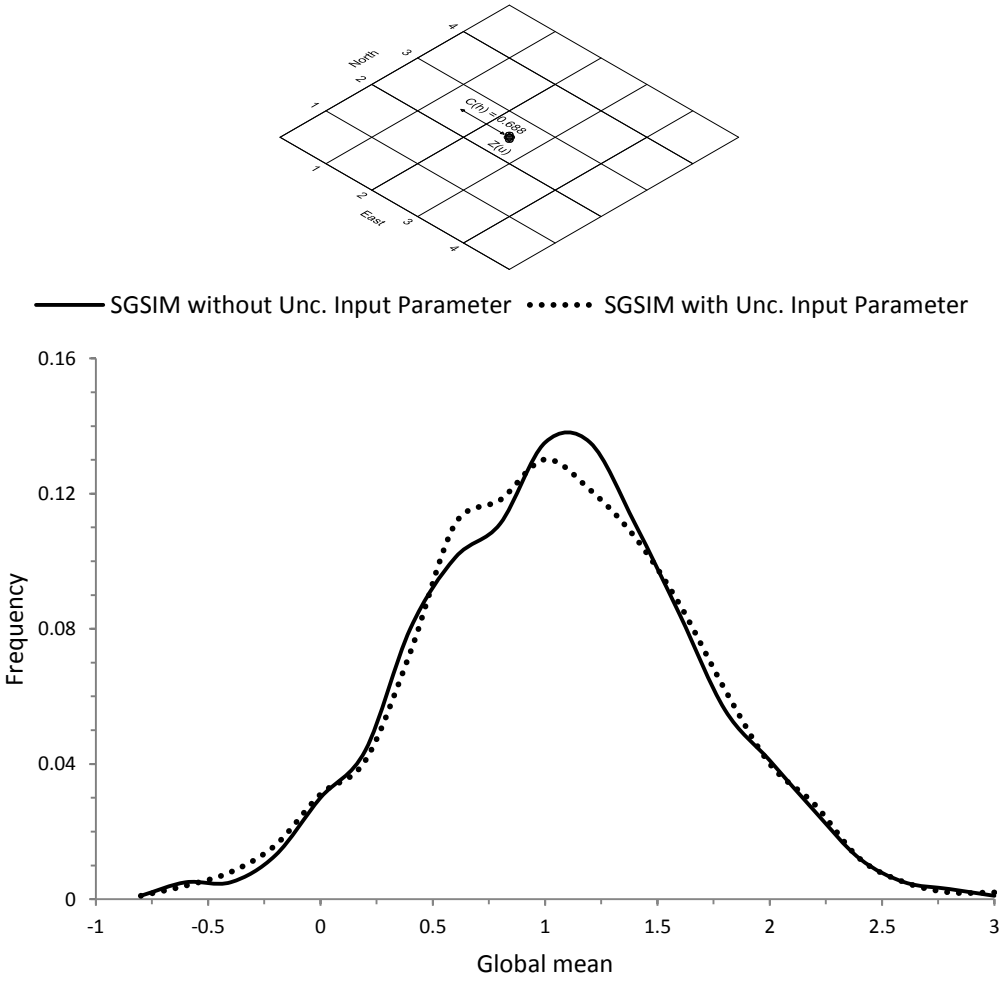


Figure 5.6: Spatial location of the conditioning sample $z(\mathbf{u})$ in the domain, where the covariance $z(\mathbf{u})$ to the nearest node is 0.688. The distribution of global means from SGSIM that use parameter uncertainty is compared with the one without parameter uncertainty.

Twenty five nodes are evaluated in a 2D map. Just as in the case of one node, two scenarios are evaluated; both of them run with the same random number seed. The increase of global uncertainty (std.) is from 0.60 to 0.61. Also, the increase in uncertainty

at all the nodes is observed when the distribution of the residuals is drawn with quantiles instead of random numbers. Conversely, the sampling of the residuals with random numbers shows two nodes with a very small reduction of local uncertainty.

5.4 Practical Considerations

Many techniques to evaluate uncertainty in the input parameters were presented in Chapters 2, 3 and 4. All of them give reasonable output; however, it is important to keep in mind that some scenarios or phases of a project development require more parameters than other. Parameters like spatial correlation and finite domain must be taken account when the project has enough data to define the domain.

The input distribution should be representative of the domain or volume to be evaluated. The limits in the tails of the distribution should be carefully defined. A wrong definition of the tails value could generate some artifact in the transformation of the values into Gaussian units. The tail values may need to be chosen separately for each variable ccdf.

The red data file is available in the CCG network and has 68 samples through a vein. There are samples of gold, silver, copper and zinc. The thickness of the samples is between 0.13 and 18.86 meters. The spatial distance between the samples is about 30 meters. The gold value is evaluated. A semivariogram model of the gold values in normal score units is required for the simulation-based approaches.

$$\gamma(\mathbf{h}) = 0.44 Exp_{\substack{ah1=100 \\ ah2=90}}(\mathbf{h}) + 0.56 Exp_{\substack{ah1=250 \\ ah2=95}}(\mathbf{h})$$

Two structures were required to model the experimental variogram, which *ah1* is the mayor range in the 15° azimuth and *ah2* is the minor range in perpendicular direction to *ah1*. The evaluations are done on a domain of size 500 meters × 600 meters using a discretization of 5 meters × 5 meters blocks. The mean of the input distribution is 1.415 ppm with a standard deviation of 1.288.

Uncertainty in the mean of the input distribution is evaluated for the Red data. The stochastic trend methodology uses a polynomial equation of second order to describe the shape of the trend. This polynomial fit reasonably the Red data to ensure the reproduction of the input parameter distribution. 10000 stochastic trends are used to evaluate the gold

value at the locations within the domain. The deviation standard of 10000 global means is 0.23.

$$Z(\mathbf{u}) = -113.60 + 5.62\text{E}03x + 0.64y - 3.70\text{E}03y^2 + 3.11\text{E}05xy$$

The same as stochastic trend, 10000 realizations are run with conventional bootstrap, spatial bootstrap and conditional finite domain techniques. The fluctuations of the mean using these techniques are illustrated in Figure 5. 7. The theories about parameter uncertainty techniques were developed in the previous Chapters.

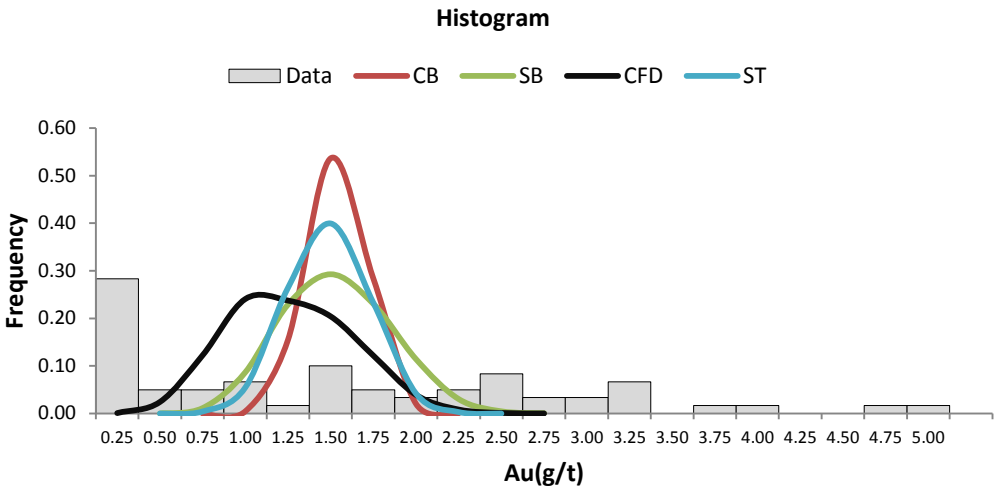


Figure 5. 7: Distribution of the uncertainty in the mean by the different techniques, CB, SB, CFD and ST are overlapped with the original distribution.

The uncertainty in the mean of the input distribution is the standard deviation of 0.360 using conditional finite domain. This uncertainty is transferred through the simulation, a non standard Gaussian distribution is defined with the mean similar to the original input distribution and the standard deviation from CFD. One hundred means are drawn from this non standard distribution, the relation of those means and the mean of the original distribution provide 100 factors. These factors are multiplied to the original distribution to generate 100 variable cdfs.

One hundred variable cdfs are created with a *genrefdist.for* program that was developed for this approach. One hundred sets of simulations are performed with their respective variable ccdf. Those ccdfs are used as reference ditribution in the program *sgsim.for*. The

output files are gathered with a program *mixsim.for*. The same number of simulations is executed for the fixed ccdf. The two sets of simulations are compared and the increase in uncertainty is illustrated in a 2D map. The increase of the local uncertainty is visible in zones where the samples show high spatial variability. For instance, from the Figure 5.8 the samples that have 0.001 ppm the lower quantity of gold are neighbours with samples with thousand times more high values.

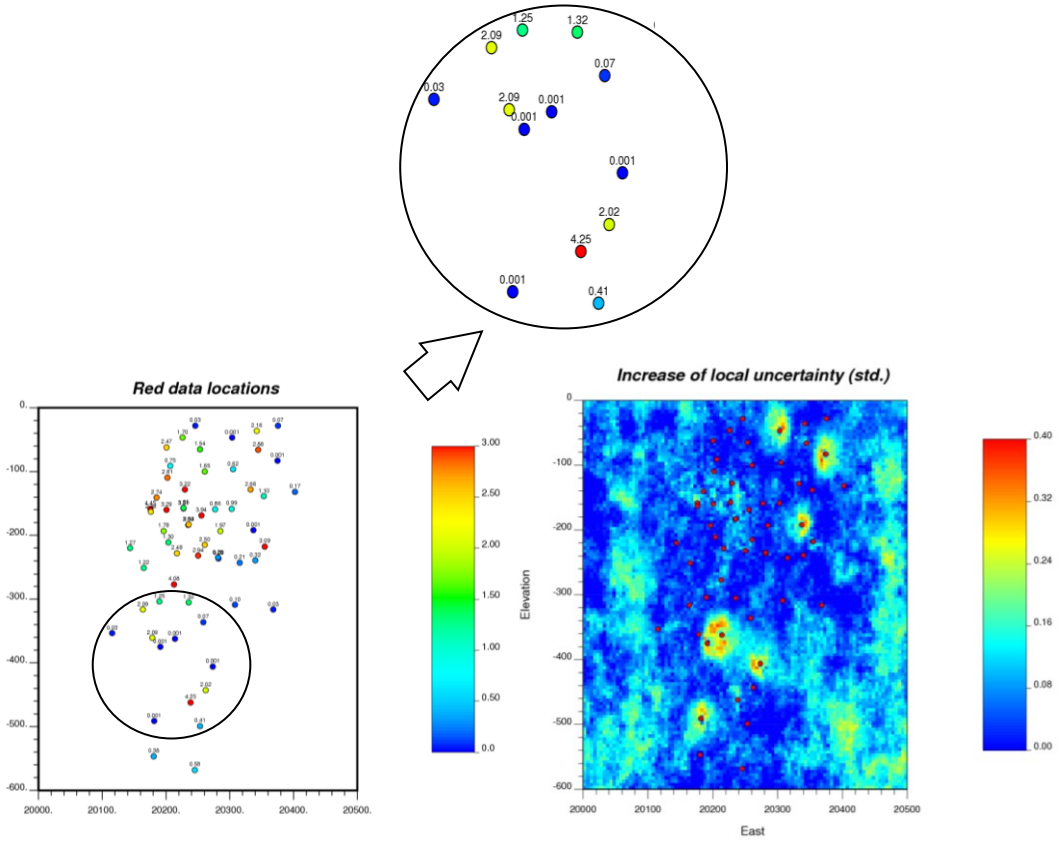


Figure 5.8: Location of the red data and map of the increase in local uncertainty because the simulation consider parameter uncertainty. The map contains red dots that represent data location.

Besides those zones, the other zones that present a considerable increase in uncertainty is in zones that are located far from the conditioning data.

The local uncertainty at the nodes using a fixed reference distribution is compared to the one using different reference distributions by scatter graphic. Positive correlation is observed in Figure 5.9.

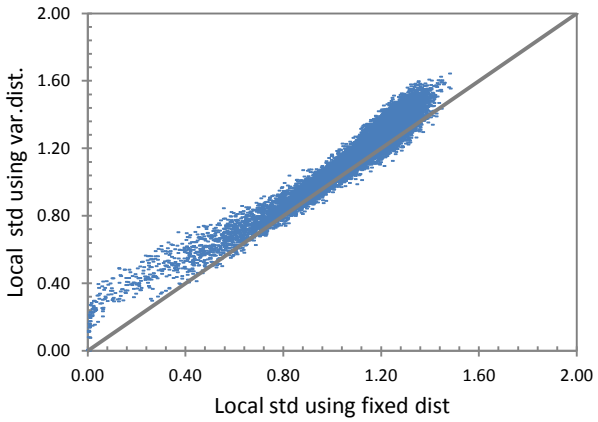


Figure 5.9: Increase in uncertainty at each node after being simulated with different reference distributions.

Every realization or map gives a unique mean that change through the realizations. The standard deviation of these means is defined as a global uncertainty. The global uncertainty with fixed ccdf is 0.12 and a narrow shape of the global means is given between 1.03 ppm and 1.64 ppm. The second scenario when the simulation take account the uncertainty in the mean of the input univariate distribution, the deviation of the global means increase to 0.24, that is, the tails of the distributions of the global means is expanded from 0.77 ppm to 2.30 ppm. Figure 5.10 shows the increase in uncertainty when is transferred the uncertainty of the input parameter to the simulation.

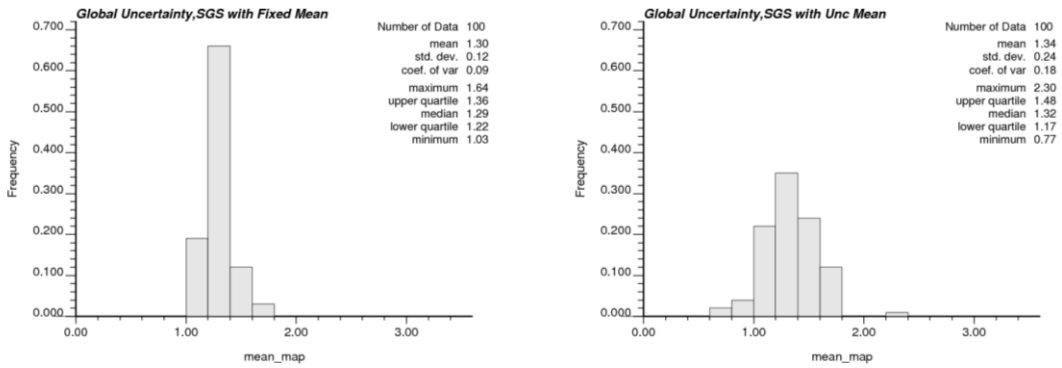


Figure 5.10: Change of global uncertainty, fixed ccdf right histogram and variable ccdfs left histogram. Histograms is with results in original units.

The example shows that an increase in uncertainty is observed in local scale or zones close to the data and in long scale or zones that are far from the data. Also, the

uncertainty at the locations of the data is zero with fixed and variable ccdf. The narrow uncertainty of the global mean is the optimal scenario provided that this uncertainty is accurate; however, uncertainty in the input parameter should not be ignored in the simulation process.

Chapter 6

Conclusions

Uncertainty in ore reserves is often evaluated by geostatistical simulation of multiple realizations. The uncertainty is affected by the amount of local data and uncertainty in the modeling parameters such as mean, univariate distribution and variogram. Parameter uncertainty is often ignored in geostatistical simulation. Then, the global uncertainty is underestimated. Evaluations in large deposits with few drill holes show that local fluctuations cancel out. As a result, the fluctuations in the global mean are very small. Accounting for uncertainty in the modeling parameters, especially the mean, is considered important for a realistic assessment of uncertainty.

Many data are required to evaluate the mean of a univariate distribution with negligible uncertainty. There are different approaches to evaluate uncertainty in the mean that involve assumptions such as (1) the data are independent or spatially correlated, (2) the realizations are limited within the domain limits or not, (3) the realizations are conditioned to the original data or not, and (4) the mean could be calculated based on a trend equation.

Traditional techniques, such as the conventional bootstrap and spatial bootstrap, sample from the distribution of the data without considering other possible values. A shortcoming of these techniques is that conditioning data are not considered; the locations and the outcomes are randomized. Also, the representativeness of the data may be questionable because preferential sampling is a common feature of geological data. The conditional finite domain technique uses a different pattern of randomization to define uncertainty in the distribution. The pattern consists on sampling from multiple conditional simulations using the original sampling strategy.

The Conditional Finite Domain (CFD) technique samples from rotated and translated configurations of the data to obtain different mean values. The configurations are simulated with different reference distributions, but are conditioned to the same original data. The sampling of many simulated configurations gives different mean values that define uncertainty for every iteration level or order. The uncertainty is stabilized after many configurations and orders are performed. A sensitivity analysis demonstrated robustness and reasonableness of this approach in scenarios where other techniques struggle with unrealistic uncertainty. An increase in the nugget effect, a reduction of the range of correlation or an increase in the size of the domain leads to more uncertainty, which is reasonable.

Geostatistics techniques often lead to less uncertainty for larger domains even if the number of data stays the same. An alternative to evaluate uncertainty in these domains could be to use a trend equation and define the mean dependent on the location within the domain. The trend model could be used to predict the value at unsampled locations. A global mean is evaluated with the set of coefficients evaluated at all locations within the domain. The stochastic trend approach proposes to randomize these coefficients considering the correlation of the original fitted coefficients. Different coefficients provide different mean values that are combined to a distribution of uncertainty in the mean.

This thesis demonstrates how uncertainty in the mean of the input distribution is transferred clearly through geostatistical simulation for a more complete understanding of uncertainty. Geostatistical simulation is usually performed with a fixed input distribution; this fixed univariate distribution comes from the input data and assumes a mean without uncertainty Multiple distributions could be constructed and used in geostatistical simulation as reference distributions. Original values are transformed into Gaussian units according to a specified reference distribution. The uncertainty in the mean of the univariate distribution is accounted by changing the reference distribution for transformation. Any Gaussian algorithm for simulation could be used. A change in local and global uncertainty is expected when the uncertainty is transferred through the simulation process.

6.1 Summary of Contributions

The original conditional finite domain technique was implemented using sequential gaussian simulation (sgsim) by (Babak & Deutsch, 2008). The use of sgsim is relatively inefficient because the search and covariance lookup table use a full grid. This thesis proposes the use of LU simulation. The LU algorithm simulates only at the *n* locations to be sampled which is more efficient than sgsim. The *n* locations correspond to configurations that are created by random translation and rotation of the data locations limited to some domain The *order* in CFD is defined as the series of simulation for each of the configurations. LU simulation is executed for each configuration and the reference distribution is conditioned to the original data. The reference distribution is taken from the previous order for each configuration. The uncertainty in the mean is taken from the final realizations. A program *cfdlu.for* was developed to implement this technique and incorporate the conventional and spatial bootstrap algorithms. A summary table *report.out* is provided to compare the uncertainty in the mean of the input distribution by these techniques.

Another contribution of the thesis is the stochastic trend methodology to evaluate uncertainty in the mean. The use of a trend equation relaxes the assumption of stationarity and defines the mean dependent on the sample location within the domain. The stochastic trend approach randomizes the trend coefficients taking account the correlation of the original fitted coefficients. The uncertainty in the input parameter is calculated from the distribution of the mean values. A program *uncrefcorr.for* was developed to implement this technique.

A simple methodology to transfer the uncertainty in the mean of the input distribution through simulation is another contribution of the thesis. Geostatistical simulation is often performed with a fixed reference distribution. The uncertainty in the mean of the univariate distribution is accounted for by changing the reference distribution for transformation. Any Gaussian algorithm for simulation could be used. The local and global uncertainty is improved. A program *genrefdis.for* was developed to generate multiple distributions.

Greater spatial continuity (larger range and/or smaller nugget effect) leads to more uncertainty with some geostatistical evaluations. This unrealistic increase in uncertainty makes statistical sense. The developed algorithms in this thesis generate more realistic results.

An Appendix to this thesis presents equations that quantify the fluctuations due to a finite domain size in presence of conditioning data. These statistical fluctuations are part of the uncertainty in the mean. These variations depend of the size of the domain and the range of correlation. When the size of the domain becomes in the order of 10 times the range of correlation the fluctuations converge to zero. The fluctuations are less when there are more conditioning data. The decrease of expected fluctuations as the size of the domain increase is reproduced analytically. The analytical model is validated by the numerical uncertainty of many realizations. A program *ulusim.for* was developed, which is a modification of the program *lusim* Alabert (1986), Deutsch (1999) to evaluate the uncertainty with the analytical model.

6.2 Future Work

During the implementation of the stochastic trend technique, the adequacy of the regression model should be verified. The examples presented in Chapter 4 were restrained to first or second order model with interaction between coordinates, this model is capable to account for a wide variety of shapes; however more orders could be considered. There are different approaches for selecting the regressors x_i, $i = 1,...,L$ in a multiple linear regression model and the purpose of them is to identify which regressors contribute significantly to the model. It is recommended to validate the regression model before implementing the stochastic trend technique.

A more complete understanding of uncertainty was the main goal of the thesis. The uncertainty in the mean is transferred through the geostatistical simulation to improve the characterization of the uncertainty. These results constitute an essential input in mine planning. Many tasks are carried out in mine planning; another future work is the incorporation of the uncertainty into the cutoff optimization task. The definition of ore and waste is a function of many variables. The objective of the optimization could be to find the best net present value related to variable cutoffs. Some variables such as metal

price cannot be controlled directly; however, there are others that can be handled and their uncertainty analysis is relevant. The transference of uncertainty in the cutoff may help to improve mine plans.

Chapter 7

Bibliography

Arik, A. (1999). An Alternative Approach to resource Classification. *APCOM Proceeding, Computer Applications in the Mineral Industries* , 45-33.

Babak, O., & Deutsch, C. V. (2008). Reserves Uncertainty Calculation Accounting for Parameter Uncertainty. *Canadian Petroleum Technology , 8*, 41-49.

Chilès, J.-P., & Delfinier, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* New York: John Wiley & Sons.

David, M. (1977). *Geostatistical Ore Reserve Estimation.* Netherland: Elsevier Scientific Publishing Company.

Deutsch, C. V. (2004). A statistical Resampling Program for Correlated Data, Spatial Boostrap. *Six Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil & Environmental Engineering University of Alberta.

Deutsch, C. V. (2002). *Geostatistical Reservoir Modeling.* New York: Oxford University Press, Inc.

Deutsch, C. V. (2000). *Special Topics in Geostatistics.* Alberta: University of Alberta.

Deutsch, C. V., & Begg, S. H. (2001). How Many realizations Do We Need? Alberta: Department of Civil & Environmental Engineering University of Alberta.

Deutsch, C. V., & Journel, A. (1998). *GSLIB: Geostatistical Softare Library and User's Guide* (2nd Edition ed.). New York: Oxford University Press.

Deutsch, C. V., Leuangthong, O., & Ortiz C., J. (2006). A Case for Geometric Criteria in Resources and Reserves Classification. *Eight Annual Report of the Centre for Computational Geostatistics* (p. 21). Edmonton: Department of Civil & Environmental Engineering-University of Alberta; Department of Mining Engineering-University of Chile.

Deutsch, C. V., Leuangthong, O., & Ortiz, J. M. (December 2007). A Case for Geometric Criteria in Resources and Reserves Classifications. *SME , 322*, 11 pages.

Dominy, S. C., Noppé, M. A., & Annels, A. E. (2002). Errors and Uncertainty in Mineral Resource and Ore Reserve Estimation: The Importance of Getting it Right. *Exploration and Mining Geology , 11*, 77-98.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics , 7*, 1-26.

Emery, X. (2008). Uncertainty modeling and spatial prediction by multi-Gaussian kriging: accounting for an unknown mean value. *Computer & Geosciences 34* , 1431-1442.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* New York: Oxford University Press.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to Applied Geostatistics.* New York: Oxford University Press.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statisticsl Analysis.* Toronto: Pearson Education, Inc.

Journel, A. G. (1994). Resampling from Stochastic Simulation. *Environmental and Ecological Statistics , 1*, 63-91.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics.* New Jersey: The Blakburn Press.

Journel, A. (2004). Roadblocks to the Evaluation of Ore Reserve, The simulation Overpass and Putting More Geology into Numerical Model of Deposits. In R.

Dimitrakopoulos (Ed.), *Ore Modeling and Strategic Mine Planning - Uncertainty and Risk Management Models* (pp. 17-20). Perth: AusIMM.

Kitanidis, P. (1997). *Introduction to Geostatistics, Applications in Hidrogeology.* New York: Cambridge University Press.

Lane, K. F. (1991). *The Economic Definition of Ore: Cut Off Grades in Theory and Practice.* London: Mining Journal Books.

Leuangthong, O., Khan, K. D., & Deutsch, C. V. (2008). *Solved Problem in Geostatistics.* Hoboken, New Jersey.: John Wiley & Sons, Inc.

Leuangthong, O., McLennan, J., & Deutsch, C. V. (2005). Acceptable Ergodic Fluctuations and Simulation of Skewed Distributions. *Seventh Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil&Environmental Engineering-University of Alberta.

Montgomery, D. C. (2000). *Design and Analysis of Experiments Fifth Edition.* New York: John Wiley & Sons, Inc.

Montgomery, D. C., & Runger, G. C. (2006). *Applied Statistics and Probability for Engineers.* Arizona: John Wiley & Sons, Inc.

Neufeld, C. T., Ortiz, J. M., & Deutsch, C. V. (2005). A Non Stationary Correction of the Probability Field Covariance Bias. *Seventh Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil & Environmental Engineering-University of Alberta; Department of Mining Engineering-University of Chile.

Pyrcz, M., & Deutsch, C. (2002). Two artifacts of probability field simulation. *Mathematical Geology , 33*, 775-800.

Ren, W., & Deutsch, C. V. (2006). Bayesian Unpdating with Local Varying Correlation. *Eight Annual Report of the Centre for Computational Geostatistics.* Edmonton: Department of Civil & Environmental Engineering University of Alberta.

Rose, P. R. (2001). *Risk Analysis and Management of Petroleum Exploration Ventures.* Tulsa, Oklahoma: The American Association of Petroleum Geologist.

Strunk Jr., W., & White, E. (2009). *The Elements of Style* (Fiftieth Anniversary ed.). New York, United States: Pearson Education, Inc.

Wang, F., & Wall, M. M. (2003). Incorporating Parameter Uncertainty into Prediction Intervals for Spatial Data Modeled via Parametric Variogram. *Agricultural, Biological and Environmental Statistics.* , 296-309.

Appendix A

A.1 Ergodicity

Geostatistical techniques for resource evaluation, such as Kriging and Simulation, require two assumptions. The first assumption of stationarity states that all multivariate distributions are invariant by translation over the study domain. Multivariate distributions are summarized by the mean vector and covariance matrix for all locations. The second assumption of ergodicity states that the spatial average (A.1) of a random stationary function (RF) $Z(\mathbf{u})$ over a domain A converges to the expected value $m=E\{Z(\mathbf{u})\}$ when A tends to infinity (A.2) (Chilès & Delfinier, 1999).

$$\overline{Z}_A = \frac{1}{|A|}\int_A Z(\mathbf{u})du \tag{A.1}$$

$$\lim_{A\to\infty} \overline{Z}_A = m \tag{A.2}$$

When the domain tends to infinity, the variance of the spatial average is expected to be zero. In practice A is finite and the spatial average Z_A will be variable when A is finite. Figure A.1 shows the change of the spatial average variance as a function of the size of A.

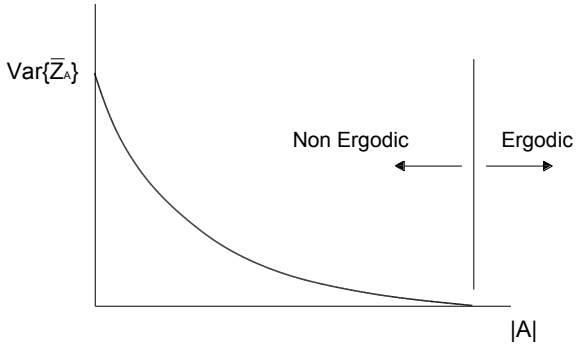


Figure A.1: The variance of spatial average versus A. When this variance is significantly greater than zero, the domain is called non ergodic.

Simulation algorithms are often based on the multivariate Gaussian RF model. This parametric model is the most widely used with extremely congenial properties (Goovaerts, 1997). The simulation of standard normal values should reproduce the Gaussian distribution with mean zero and variance one; however, ergodic fluctuations make the results different from zero and one.

A study on acceptable ergodic fluctuations (Leuangthong, McLennan, & Deutsch, 2005) shows significant statistical fluctuations for three examples with variogram range of 20%, 50%, and 100% of the domain. Even when the domain becomes relatively large compared to the range of correlation, these statistical fluctuations are considerable and are a part of the global uncertainty. The magnitude of the statistical fluctuations can be quantified by performing non conditional simulation. The expected fluctuations in the mean are derived below in presence of conditioning and verified by numerical examples.

A.2 Expected Fluctuations in the Mean

The variance of the spatial average in the domain A is a measure of the expected statistical fluctuations. The domain could be discretized by N nodes, these are defined by the variable function $Z(\mathbf{u}^{(i)})$, where the location of every node is $\mathbf{u}^{(i)}$, $i=1,...,N$. The available data are defined by $z(\mathbf{u}_k)$, where the location of each data value is $\mathbf{u}_k$, $k = 1,..,n$. These n available data values and N nodes define the domain A. Values at every node are estimated conditioned to the available n values.

The covariance of the RF $Z(\mathbf{u})$ should be constant over the domain; however, a non stationary covariance is observed in the presence of conditioning data. The covariance near the conditioning data is a function of the input "ergodic" covariance model and the location of the conditioning data. Figure A.2 shows a domain A that has $z(\mathbf{u}_k)$, $k = 1,...,6$ conditioning data. The discretization of the domain is with 100 nodes. As expected, the covariance between adjacent node location $\mathbf{u}^{39}$ and $\mathbf{u}^{40}$ will be different than the covariance between the node locations $\mathbf{u}^{71}$ and $\mathbf{u}^{72}$. This difference is because $C(\mathbf{u}^{71},\mathbf{u}^{72})$ has node locations that are near the conditioning data; $C(\mathbf{u}^{39},\mathbf{u}^{40})$ has node locations that are far from the conditioning data. That is, the distance to the conditioning data matters in the evaluation of covariances.

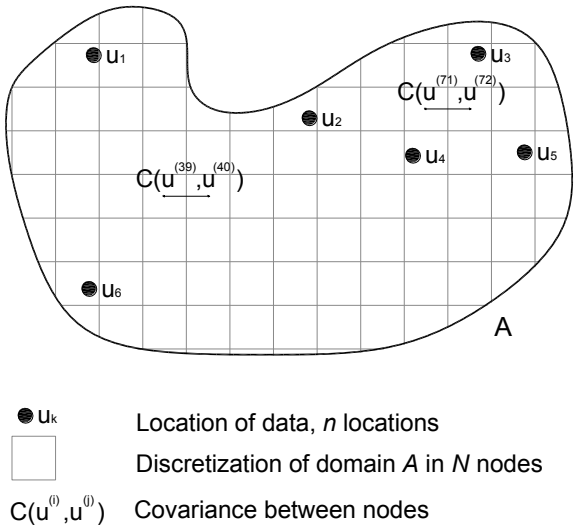


Figure A.2: The covariance $C(\mathbf{u}^{39},\mathbf{u}^{40})$ that is far from the conditioning data is different to the covariance $C(\mathbf{u}^{71},\mathbf{u}^{72})$ that is near from the conditioning data.

This non stationary covariance is correctly reproduced in sequential Gaussian simulation because the previous simulated nodes are used in the estimation of subsequent nodes (Neufeld, Ortiz, & Deutsch, 2005). These conditional covariances are required to compute the variance of the spatial average, which is given by:

$$Var\left\{\overline{Z}_A\right\} = \frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N} E\left\{Z\left(\mathbf{u}^i\right)Z\left(\mathbf{u}^j\right)\right\} - \left[E\left\{\overline{Z}_A\right\}\right]^2 \tag{A.3}$$

The variance of the spatial average is expanded below. The first term is equivalent to the expected value of the conditional covariance between nodes plus the quadratic of the expected value of the spatial average, and the second term is the quadratic of the expected value of the spatial average.

$$Var\left\{\overline{Z}_A\right\} = \frac{1}{N^2}\left[\sum_{i}^{N}\sum_{j}^{N} Cov\left\{Z\left(\mathbf{u}^i\right)Z\left(\mathbf{u}^j\right)\right\} + \left[E\left\{\overline{Z}_A\right\}\right]^2\right] - \left[E\left\{\overline{Z}_A\right\}\right]^2$$

The previous equation is simplified and the quadratic of the expected value of the spatial averages are canceled out. Where, $Cov\{Z(\mathbf{u}^i)Z(\mathbf{u}^j)\}$ corresponds to the covariance between two random variables conditioned to the available data.

$$Var\left\{\overline{Z}_A\right\} = \frac{1}{N^2}\sum_{i}^{N}\sum_{j}^{N} Cov\left\{Z\left(\mathbf{u}^i\right)Z\left(\mathbf{u}^j\right)\right\} \tag{A.4}$$

The conditional covariance could be developed by the following steps: Define an $n \times n$ covariance matrix between available n data as C_{11}.

$$C_{11} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}_1)\ldots C(\mathbf{u}_1 - \mathbf{u}_n) \\ \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ C(\mathbf{u}_n - \mathbf{u}_1)\ldots C(\mathbf{u}_n - \mathbf{u}_n) \end{bmatrix}$$

Define the covariance matrix between n data and N locations of the discretized domain as C_{12}. Also the notation $Z(\mathbf{u})$ will be simplified in the expressions by just vector **u.**

$$C_{12} = \begin{bmatrix} C(\mathbf{u}_1 - \mathbf{u}^{(1)})\ldots C(\mathbf{u}_1 - \mathbf{u}^{(N)}) \\ \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ C(\mathbf{u}_n - \mathbf{u}^{(1)})\ldots C(\mathbf{u}_n - \mathbf{u}^{(N)}) \end{bmatrix}$$

Define the covariance matrix between N locations of the discretized domain in nodes as C_{22}.

$$C_{22} = \begin{bmatrix} C(\mathbf{u}^{(1)} - \mathbf{u}^{(1)})\ldots C(\mathbf{u}^{(1)} - \mathbf{u}^{(N)}) \\ \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ C(\mathbf{u}^{(N)} - \mathbf{u}^{(1)})\ldots C(\mathbf{u}^{(N)} - \mathbf{u}^{(N)}) \end{bmatrix}$$

The expression for the conditional covariance matrix of N node locations given n conditioning values is given after combining covariances matrices C_{11}, C_{12} and C_{22}. The calculations of all the covariances use an input model covariance.

$$C_{\left(u^1,\ldots,u^N \mid u_1,\ldots,u_n\right)} = C_{22} - C_{12}^T C_{11}^{-1} C_{12} \tag{A.5}$$

The kriging system is given in Equation (A.6). This term is observed in the conditional covariance equation.

$$[\lambda] = C_{11}^{-1} C_{12} \tag{A.6}$$

The covariance matrix between the n data and the N nodes is transposed and multiplied by the kriging weights.

$$C_{12}^{T}[\lambda]=\begin{bmatrix}\sum_{k=1}^{n}\lambda_k^{(1)}C(\mathbf{u}_k-\mathbf{u}^{(1)})\ldots\sum_{k=1}^{n}\lambda_k^{(N)}C(\mathbf{u}_k-\mathbf{u}^{(1)})\\ \vdots \qquad\qquad \ddots \qquad \vdots \\ \sum_{k=1}^{n}\lambda_k^{(1)}C(\mathbf{u}_k-\mathbf{u}^{(N)})\ldots\sum_{k=1}^{n}\lambda_k^{(N)}C(\mathbf{u}_k-\mathbf{u}^{(N)})\end{bmatrix} \tag{A.7}$$

The previous matrix is substituted by the outcome of minimizing the kriging variance Equation (A.8). The covariance between random variables in the presence of conditioning data is deduced (A.9) (Neufeld, Ortiz, & Deutsch, 2005).

$$\sum_{k'=1}^{n}\lambda_{k'}C_{kk'}=C_{k0}\ \ where\ k=1,\ldots,n \tag{A.8}$$

$$C_{\left(u^1,\ldots,u^N\mid u_1,\ldots,u_n\right)}=C_{22}-\begin{bmatrix}\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(1)}\lambda_{k'}^{(1)}C(\mathbf{u}_k-\mathbf{u}_{k'})\ldots\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(N)}\lambda_{k'}^{(1)}C(\mathbf{u}_k-\mathbf{u}_{k'})\\ \vdots \qquad\qquad \ddots \qquad\qquad \vdots \\ \sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(1)}\lambda_{k'}^{(N)}C(\mathbf{u}_k-\mathbf{u}_{k'})\ldots\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(N)}\lambda_{k'}^{(N)}C(\mathbf{u}_k-\mathbf{u}_{k'})\end{bmatrix} \tag{A.9}$$

The simplified equation of the previous matrix is a function of the covariances between node locations, the set of weights and the covariances between conditioning data. Where no conditioning data are present, the covariances are identical to the input covariance model.

$$C_{\left(\mathbf{u}^{(i)},\mathbf{u}^{(j)}\mid \mathbf{u}_1,\ldots,\mathbf{u}_n\right)}=C_{ij}-\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(j)}\lambda_{k'}^{(i)}C_{kk'} \tag{A.10}$$

The equation for the conditional covariance between random variables is replaced in Equation (A.4) to obtain the conditional variance of the spatial average.

$$Var\left\{\overline{Z}_A\right\}=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(C_{ij}-\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(j)}\lambda_{k'}^{(i)}C_{kk'}\right)$$

The variance of the spatial average is expanded. The two terms show their influence on the total $Var\{\bar{Z}_A\}$. That equation accounts for the covariances of all the nodes that are inside the domain A. The non stationary covariance is reproduced in the presence of conditioning data. The first term is the average of the $N \times N$ covariances between nodes

that belong to domain A, and the second term is the average of $N \times N$ nodes combinations of redundancy measures of the available data regard to the nodes.

$$Var\left\{\overline{Z}_A\right\} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(C_{ij}\right) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{k=1}^{n}\sum_{k'=1}^{n}\lambda_k^{(j)}\lambda_{k'}^{(i)}C_{kk'}\right)$$

The expected value of the spatial average is represented by:

$$E\left\{\overline{Z}_A\right\} = E\left\{\frac{1}{N}\sum_{i=1}^{N}z_i^*\right\} = \frac{1}{N}\sum_{i=1}^{N}E\left\{z_i^*\right\} = \frac{1}{N}\sum_{i=1}^{N}E\left\{\sum_{k=1}^{n}\lambda_k^{(i)}z_k\right\}$$

$$assume\ \ E\left\{z_k\right\} = z_k \Longrightarrow E\left\{\overline{Z}_A\right\} = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{n}\lambda_k^{(i)}z_k$$

The previous two equations permit the evaluation of fluctuations due to a lack of ergodicity in presence of conditioning. These statistical fluctuations are part of the uncertainty in the mean for the domain A. These variations depend of the size of the domain and the range of correlation. When the size of the domain becomes in the order of 10 times the range of correlation the fluctuations converge to zero. The fluctuations are less when there are more conditioning data. These results are verified numerically.

A.3 Application

A.3.1 Verification of Non Stationary Covariance

A simple scenario is used to show the influence of the conditioning data on the evaluation of the covariance. Three samples are located in an area of 150 meters × 150 meters. These samples are standard normal Gaussian. The variogram model is spherical and isotropic.

$$\gamma(\mathbf{h}) = 0.2 + 0.8 \cdot sph_{a=150}(\mathbf{h})$$

The area of study is discretized by nine nodes $\mathbf{u}^{(i)}$, i = 1,...,9. The covariance between node $\mathbf{u}^{(2)}$ and node $\mathbf{u}^{(3)}$ given three conditioning data $\mathbf{u}_k$, k = 1,...,3 requires the kriging weights for every node σ_k, k=1,...,3, the covariance between samples locations $C_{kk'}$ and the covariance between $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$. For the node $\mathbf{u}^{(2)}$ the kriging weights result -0.063, 0.109 and 0.508 and for the node $\mathbf{u}^{(3)}$ the kriging weights result -0.094, 0.208 and 0.192.

As expected, the kriging weights are proportional to the distance between nodes and samples. Furthermore, the covariance matrix between the sample locations is as follow:

$$C_{kk'} = \begin{bmatrix} 1 & 0.373 & 0.250 \\ 0.373 & 1 & 0.276 \\ 0.250 & 0.276 & 1 \end{bmatrix}$$

From Figure A.3, the size of every node is 50 meters × 50 meters, then, the covariance $C(\mathbf{u}^{(2)}\ \mathbf{u}^{(3)})$ between adjacent evaluated nodes has the lag distance **h** equal to 50 meters. The covariance between $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$ given three conditioning samples is solved in the next equation:

$$\begin{aligned} C_{\left(\mathbf{u}^{(2)},\mathbf{u}^{(3)}\mid \mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\right)} &= C_{23} - \sum_{k=1}^{3}\sum_{k'=1}^{3} \lambda_k^{(3)} \lambda_{k'}^{(2)} C_{kk'} \\ &= 0.41 - (-0.094 \cdot -0.063 \cdot 1 + -0.094 \cdot 0.109 \cdot 0.373 + -0.094 \cdot 0.508 \cdot 0.250; \\ &\quad + 0.208 \cdot -0.063 \cdot 0.373 + 0.208 \cdot 0.109 \cdot 1 + 0.208 \cdot 0.508 \cdot 0.276; \\ &\quad + 0.192 \cdot -0.063 \cdot 0.250 + 0.192 \cdot 0.109 \cdot 0.276 + 0.192 \cdot 0.508 \cdot 1) \\ &= 0.41 - 0.138 = 0.272 \end{aligned}$$

To verify the non stationary covariance in the presence of conditioning data, two other nodes with the same vector lag distance **h** are evaluated, namely the covariance between $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ that are near to the conditioning data. Where, the kriging weights for the node $\mathbf{u}^{(4)}$ result 0.258, 0.004 and 0.457; for the node $\mathbf{u}^{(5)}$ result 0.141, 0.363 and 0.387. Like the previous evaluation, the equation of the covariance conditioned to the data is as follows:

$$\begin{aligned} C_{\left(\mathbf{u}^{(4)},\mathbf{u}^{(5)}\mid \mathbf{u}_1,\mathbf{u}_2,\mathbf{u}_3\right)} &= C_{45} - \sum_{k=1}^{3}\sum_{k'=1}^{3} \lambda_k^{(5)} \lambda_{k'}^{(4)} C_{kk'} \\ &= 0.41 - (0.141 \cdot 0.258 \cdot 1 + 0.141 \cdot 0.004 \cdot 0.373 + 0.141 \cdot 0.457 \cdot 0.250; \\ &\quad + 0.363 \cdot 0.258 \cdot 0.373 + 0.363 \cdot 0.004 \cdot 1 + 0.363 \cdot 0.457 \cdot 0.276; \\ &\quad + 0.387 \cdot 0.258 \cdot 0.250 + 0.387 \cdot 0.004 \cdot 0.276 + 0.387 \cdot 0.457 \cdot 1) \\ &= 0.41 - 0.337 = 0.073 \end{aligned}$$

The data of the example are illustrated in Figure A.3. The covariance C_{23} and C_{45} without conditioning data are equal to 0.41 because the vector distances in both cases are the same; however, in the presence of conditioning data, the conditional covariances over the domain become non stationary and depend on the distance to the conditioning data.

The conditional covariance $C_{23|1,2,3}$ is farther from the conditioning data than the other conditioning covariance $C_{45|1,2,3}$, then, the example shows that $C_{23|1,2,3}$ is greater than $C_{45|1,2,3}$ because the second term depends on kriging weights and is subtracted from the constant value 0.41 to get the conditional covariance of these nodes distant in 50 meters. For instance, the nodes $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ obtain greater kriging weights than nodes $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$ because they are located close to the conditioning data. As a result, the second term is 0.337 for nodes $\mathbf{u}^{(4)}$ and $\mathbf{u}^{(5)}$ greater than 0.138 for nodes $\mathbf{u}^{(2)}$ and $\mathbf{u}^{(3)}$

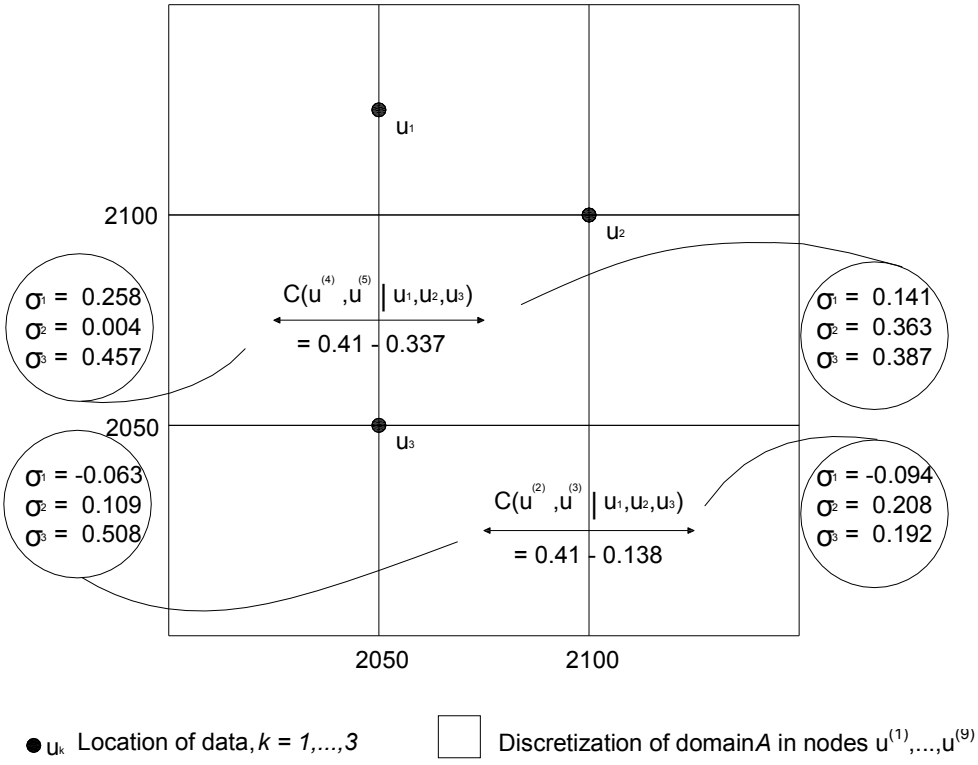


Figure A.3: Graphic of the non stationary covariance in the presence of conditioning data.

The example shows that conditional covariance $C_{45|1,2,3}$ near the data results in 18 % of the covariance model. Otherwise, the conditional covariance $C_{23|1,2,3}$ located a little far from the data results in 66 % of the covariance model (0.41). Those results are verified by simulating 1000 realizations; the $C_{45|1,2,3}$ gives 0.081 and $C_{23|1,2,3}$ gives 0.337. That is, the covariance given n conditioning data increases and becomes close to covariance model as the evaluated nodes are far from the conditioning data.

A.3.2 Verification of Analytical Variance of the Spatial Average

The data and covariance model from the previous example is used to demonstrate the reduction in variance as the domain A increases. The variance of the spatial average of the domain that is discretized by nine nodes with three conditioning data is computed below. The first term corresponds to the covariance average between nodes equal to 0.3204, this value does not depend on the conditioning data; the second term equal to 0.2022 corresponds to the term that accounts the location of the conditioning data. The expected value of the spatial average 0.0964 accounts the values of the input data.

$$
\begin{aligned}
Var\left\{\overline{Z}_A\right\} &= \frac{1}{9^2}\sum_{i=1}^{9}\sum_{j=1}^{9}\left(C_{ij}\right) - \frac{1}{9^2}\sum_{i=1}^{9}\sum_{j=1}^{9}\left(\sum_{k=1}^{3}\sum_{k'=1}^{3}\lambda_k^{(j)}\lambda_{k'}^{(i)}C_{kk'}\right) \\
&= 0.3204 - 0.2022 \\
&= 0.1182 \\
E\left\{\overline{Z}_A\right\} &= \frac{1}{9}\sum_{i=1}^{9}\sum_{k=1}^{3}\lambda_k^{(i)}z_k \\
&= 0.0964
\end{aligned}
$$

These values are shown in Table A., where the parameters of the input covariance model are kept constant as the size of the domain A is increased from 150 meters to 1550 meters. The first term becomes smaller as the size of the domain increases because the covariances between distant nodes are less; the second term becomes smaller because the conditioning data are located farther from the nodes as the size of the domain increases.

A(Xd×Yd)	First term	Second term	Analytical Model
150×150	0.3204	0.2022	**0.1182**
350×350	0.0802	0.0158	**0.0644**
550×550	0.0351	0.0026	**0.0325**
750×750	0.0195	0.0008	**0.0187**
950×950	0.0124	0.0003	**0.0121**
1150×1150	0.0086	0.0001	**0.0085**
1350×1350	0.0063	0.0001	**0.0062**
1550×1550	0.0048	0	**0.0048**

Table A.1: Change of the variance of the spatial average with different size of domains.

As expected, the variance becomes smaller as the size of the domain is increased, the variance approaches zero asymptotically, but it is practically zero when the ratio of the domain size and range of correlation is around 10. These analytical values are compared to the numerical model in Figure A.4.

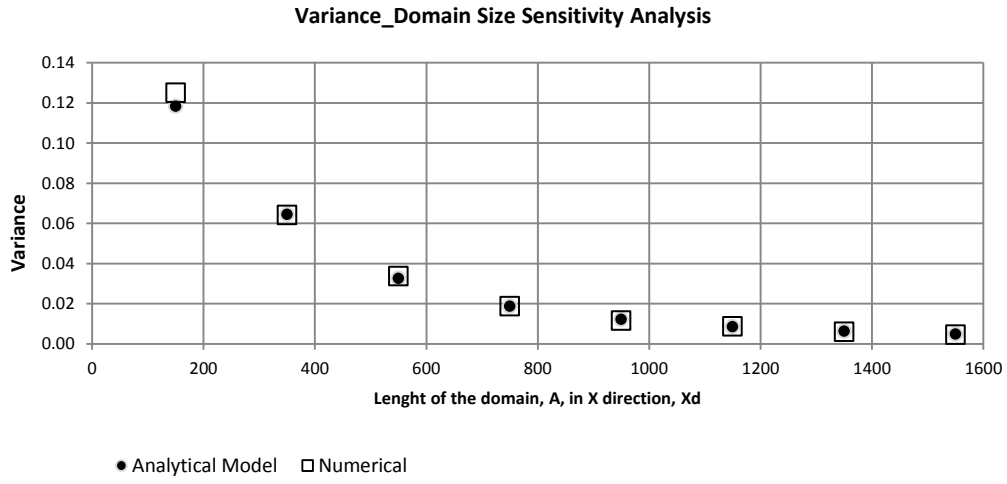


Figure A.4: Synthetic data, non-ergodic variance of spatial average with different domains size.

As expected, both the analytical model and the numerical results show the decrease of the variance of the spatial average as the domain increases; however, slight differences are observed due to the limited number of samples. The ratio of the domain size and the range of correlation in the domain 150 × 150 is 1 and in the domain 350 × 350 is 2.3. Those ratios correspond to significantly non ergodic domains because the ratios are less than 10. The numerical approach shows slight variations of the variance of the spatial average. For instance 200 realizations of the domain 150 × 150 show variance 0.13 and 2000 realizations show variance 0.12.

A second example is used to compare values of variance from 200 simulations (numerical) and analytical model. The "red" data contain 60 values of gold grade from a planar gold vein that are located in an area of 500 meters by 600 meters. These values are transformed to normal Gaussian score and their anisotropic variogram is defined by:

$$\gamma(\mathbf{h}) = 0.2 + 0.8 \cdot sph_{\substack{ah1=250 \\ ah2=150}}(\mathbf{h})$$

The size of nodes is 50 meters × 50 meters, the domain size (*Xd*×*Yd*) 500 meters × 600 meters is increased eight times proportionally until the domain size reach (*Xd*×*Yd*) 2250 meters × 2700 meters. The largest domain is equivalent to 10 times the range of correlation. The previous example used synthetic data that contained 3 sample locations. Meanwhile, the current example shows a real scenario of 60 values. More samples and real scenario evaluate fairly the analytical model against the result of many simulations.

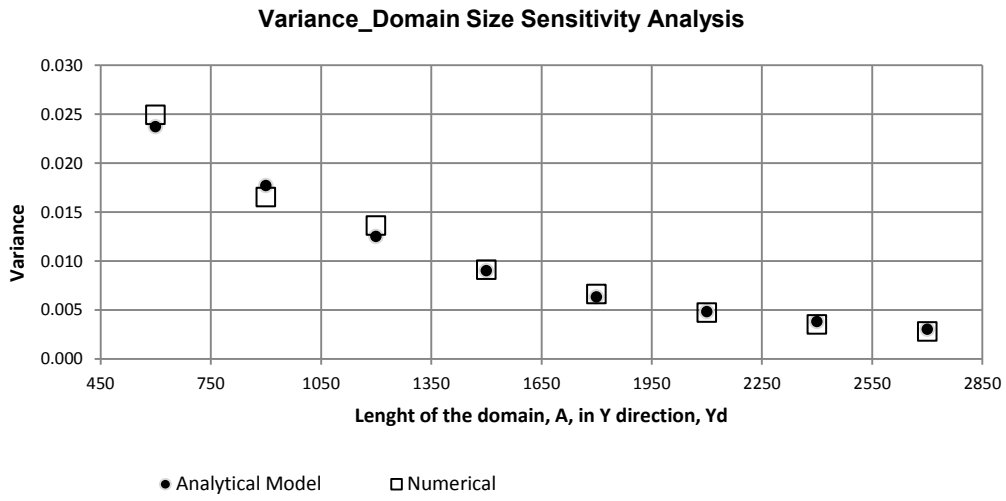


Figure A.5: Variance of the spatial average with different domain sizes for the Red data.

The variance of the analytical model and numerical results are similar, the slight difference is due to the numerical model being sensitive to the random generator of realizations. The decrease of expected fluctuations as the size of the domain increase is reproduced as in the previous example. The examples validate the analytical model.

The concept of ergodicity states that the spatial average of a random stationary function (RF) $Z(\mathbf{u})$ over a domain A converges to the expected value $m=E\{Z(\mathbf{u})\}$ when A tends to infinity. Statistical fluctuations of realizations are reduced as the ratio between domain size and range of correlation increase. The examples show that fluctuations practically reach zero at a ratio of 10.

Appendix B

B.1 Conditional Finite Domain Program (cfdlu.for)

The Conditional Finite Domain(CFD) technique permits quantification of uncertainty in the mean of a univariate distribution. The technique samples from rotated and translated configurations of simulated data. Many configurations are created by random translation and rotation of the data locations relative to the centroid of the original data configuration or limited to some domain. The *order* in CFD is defined as the series of simulation for each of the configurations. The reference distribution is updated for every *order* of simulation. LU simulation is executed for every configuration and the realization is conditioned to the original data. The reference distribution is taken from the previous order. The simulated configurations will give different means that define the uncertainty in the mean. The uncertainty is estabilized with many configurations and orders.

Simulation is done in Gaussian units. The normal scores transformation of the data becomes sensitive to the tail extrapolation options as the order of simulation increases. Reasonable values must be chosen by the user. The parameter file of the program follows the conventions of GSLIB, the name of the data file is asked in the first line. The locations of coordinates *X, Y, Z,* value and declustered weight are required in the next line.

```
data.dat            - input file with data
1  2  3  4  5       - columns for X, Y, Z, variable and weight
```

The trimming limits on the next line removes missing values.

```
-0.1E+04 0.1E+04    - trimming limits LTGT
```

The program will calculate the minimum and maximum tail values as small deviations from the minimum and maximum data values.

```
1                   - itail, permit calculate tails?(1=yes  0=no )
```

The next three parameters are considered only when itail is set equal to 0. Otherwise, the tail options should be specified according the conventions of GSLIB.

```
0  15.0             - data limits(tails)
1   0.0             - lower tail option, parameter
1  15.0             - upper tail option, parameter
```

The number of configurations to be simulated and sampled is set in the next line. The code offers the alternative to run with only translation (0) or translation and rotations (1).

```
100                 - nconf (number of new configuration)
1                   - permit rotation?(1=yes  0=no )
```

Translation is set with respect to the centroid of the configuration data. A window around this location will limit the random translation distance; the dimension of the window could be reference to the median space between data (0), longest distance between samples or apparently domain size (1) or arbitrary value (2). The product of the factor (trfrac) and the reference distance give the size of the window. Sensitivity analysis shows robust CFD uncertainty in the trfrac interval [0.1-0.3] regard to the domain size and trfrac interval [0.2-0.6] regard to the spacing data. The second line is used when the translation is relative to some arbitrary dimension of the window respect to the centroid of the data.

```
0 0.50        - itrans refer(space data=0,domain=1,value=2)& trfrac
0.0           - if itrans = 2 dimension window value
```

The ASCCI file of the domain is required, those values should be flagged. The number used to flag the area inside the domain should be specified in the next line.

```
1             - consider domain? iflag(1=yes  0=no )
block.clp     - only if iflag=1,input file with binary values
1             - column of flag, only if iflag=1
```

The code generates an output file of configurations, uncertainty of the mean by order and the means for all configurations and orders. The file cfdlu.out is used to graphic the uncertainty in the mean by every order, this graphic let find the point of convergence of the uncertainty in the mean. A summary table report.out is provided to compare CFD with other techniques that evaluate the uncertainty in the mean of the input distribution.

```
cfdlu_conf.out      - output file with new configuration
cfdlu_sum.out       - output file with uncertainty(conf x order)
cfdlu.out           - output file with uncertainty versus order
report.out          - output file_summary table of uncertainty
```

The number or orders is set in the next line and this value corresponds to the number of simulation of every configuration with the respective previous assembled reference distribution. The option of seed number is similar to traditional GSLIB simulation code.

```
100                 - uncertainty order(number simulation of nconf)
112063              - random number seed
```

The variogram model, number of structure, nugget effect, sill, ranges and angles of anisotropy follow the GSLIB traditional code of simulation or interpolation.

```
1  0.2              - nst, nugget effect
1 0.8 0.0 0.0 0.0   - it,cc,ang1,ang2,ang3
50 50 50            - a_hmax, a_hmin, a_vert
```

The code added routines of conventional bootstrap and spatial bootstrap. Therefore an additional output files without any specification in the parameter file of the code is generated. *Boot_avg.out* - the mean and standard deviation of the realizations using CB technique, *Spatial_bootstrap.out* - the mean and standard deviation of the realizations using SB technique, and *space_data.out* - information of the size domain and median space between data

B.2 Stochastic Trend Program (uncregcorr.for)

The parameter file of the stochastic trend program calculates the uncertainty in the mean of the input distribution. This program is a modification of the *correlate* program (Ren & Deutsch, 2006) and is implemented to develop stochastic trend technique. This code considers linear or quadratic polynomials. The geological model is defined by an equation that contains coefficients, then, the thesis proposes to randomize the regression coefficients to obtain uncertainty in the global mean. The first, second and third lines are the same as available routines of GSLIB programs.

```
s_data.dat          - file with data input
1 2 3 4             - columns X Y Z and variable
-1.0e21   1.0e21    - trimming limits
```

The trend equation defines the mean dependent on the sample location at the domain, the number of drifts or terms that represents the unknown regression coefficients are set in the next line the function of the trend f_l (**u**) is an equation based in an understanding of the trend.

```
1 1 1 0 0 0 0 0 0 - 1=use parameter in:ix,iy,iz,ixx,iyy,izz,ixy,ixz,iyz
```

The next line specifies a file for the coefficients of the trend and their variance that are calculated by using the theory of multiple regression models.

```
regres_coef.out     - file of regression coef output
```

The number of realizations of the coefficients of the trend should be given in the next line and the number of seed is similar to traditional GSLIB simulation code.

```
100                 - simulation number of reg.coef
69069               - random number seed
```

Realizations of the coefficients are correlated and the output file of them is specified.

```
correlated.out      - file of correlated Gaussian output
```

The program gives a stochastic trend, which could be used to evaluate the values at the location of the original data or at nodes of the whole domain.

```
1                   - Result on data location(0) or on the grid(1)
```

Specification of the limit of the simulation

```
40 2.5 5.0          - nx,xmn,xsiz
40 2.5 5.0          - ny,ymn,ysiz
15 11 2             - nz,zmn,zsiz
```

The code could use a three dimensional domain or two dimensional domain and the number used to flag the area inside the domain is specified.

```
1                   - consider domain? iflag(1=yes  0=no )
flagn.clp           - only if iflag=1,input file with binary values
1                   - column of flag, only if iflag=1
```

All the nodes that are calculated with the stochastic trend are reported in the next file and the means for every stochastic trend is written in *unc_sim.out.*

```
regcoef_sim.out     - output, simulation of fit regression coefficient
unc_sim.out         - uncertainty of fit regression coef. Simulation
```

The ASCII file of the domain should have the same dimension as the limit of the simulation, the program works only in the flagged nodes.