University of Alberta

Accounting for Non-Stationarity Via Hyper-Dimensional Translation of the Domain in Geostatistical Modeling

by

Miguel Angel Cuba Espinoza

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

Fall 2009
Edmonton, Alberta

Examining Committee

Dr. Oy Leuangthong (Supervisor), Civil and Environmental Engineering

Dr. Jozef Szymanski (Chair and Examiner), Civil and Environmental Engineering

Dr. Julian Ortiz (Co-Supervisor), Mining Engineering - University of Chile

Dr. Ivan Mizera (Examiner), Math & Statistical Sciences

To my parents Julian and Isidora,
my brother Julio and my grandmother Feliciana

Abstract

Medium and short term mine planning require models of mineral deposits that account for internal geological structures that permit scheduling of mine production at a weekly and monthly production periods. Modified kriging estimation techniques are used for accounting for such geologic structures. However, in the case of simulation, it is strongly linked to the use of sequential Gaussian simulation which has difficulties in reproducing internal geologic patterns.

This thesis presents: (1) a set of tools to verify the impact of mean and variance trends in a domain; (2) a methodology for identifying highly variable sub-regions within domains; and (3) a simulation methodology that accounts for the internal structures in the domain required by medium and short term planning. Specifically, the simulation approach consists of: (1) moving the domain to a high dimensional space where the features of the internal structures in the domain are more stationary, (2) simulating the realizations via sequential Gaussian simulation, and (3) projecting the results to the initial dimensional space.

Acknowledgements

I would like to thank my supervisor Dr. Oy Leuangthong for her guidance, dedication, support and patience during my studies in the University of Alberta. During the meetings in her office I learned more about this academic world and enjoyed the discussion about the topics presented in this thesis. Thanks also to my co-supervisor Dr. Julian Ortiz for supporting my ideas at the early stages of this research.

I am grateful to the Centre of Computational Geostatistics (CCG) for providing financial support during my studies and its industrial sponsors that keep supporting the research made in the CCG.

Dr. Clayton Deutsch is an inspiration in the CCG, during the group meetings he is a source of motivation and his leadership urges the necessity to keep improving and developing new ideas.

My gratitude also goes to my friends, Yupeng Li, Behrang Koushavand, Tong Wang, Abhay Kumar and especially Enrique Gallardo for their support and willingness to discuss academic and non-academic topics.

Table of Contents

List of Figures

List of Symbols

$\mathbf{u}$	Location vector
$\mathbf{h}$	Separation vector
$\mathbf{l_h}$	Lag distance
$Z(\mathbf{u})$	Continuous random variable at location $\mathbf{u}$
$Y(\mathbf{u})$	Continuous random variable at location $\mathbf{u}$ in normal score scale
$z(\mathbf{u})$	Realization from a random variable $Z(\mathbf{u})$
$z^*(\mathbf{u})$	Kriging estimate at location $\mathbf{u}$
$\gamma(\mathbf{h})$	Semivariogram model for separation vector $\mathbf{h}$
$\hat{\gamma}(\mathbf{h})$	Experimental semivariogram for separation vector $\mathbf{h}$
$f(z)$	Probability density function of Z
$F(z)$	Cumulative density function of Z
λ_i	Weight assigned to sample i
χ_p^2	Chi-square distribution with p degrees of freedom
RF	Random function
SRF	Stationary Random function
IRF	Intrinsic Random function
RV	Random variable
SK	Simple Kriging
OK	Ordinary Kriging
UK	Universal Kriging
SGS	Sequential Gaussian simulation
SMU	Selective Mining Unit
ACD	Alternative conditioning dataset
OCD	Original conditioning dataset
MSE	Mean squared error
uod	Units of distance

1. Introduction

Geostatistics in the mining industry is used for modeling uncertainty of mineral deposits. A mineral deposit model consists of the characterization of many variables; in general, these variables are geologic properties used for economic evaluation, such as metal grades, metal contaminants, rock type, etc. The mineral deposit model is used differently at the different stages of a mine project. In the feasibility stage it is important to quantity the economic potential of the mineral deposit and different mining strategies are proposed. Once the mine project is of economic interest, three model types are built based on the mine planning requirements: (1) long term model, based on 3 to 5 years basis, (2) medium term model, based on 3 months to 1 year basis and (3) short term model, based on 1 week to 1 month basis. Usually, a mineral deposit model is either an estimated model that consists of the estimation of the uncertainty parameters such as mean and variance of the variables of interest at the unsampled locations, or a simulated model that is a set of many realizations that are possible scenarios of the mineral deposit.

For the different types of use of the models, the target in their construction varies. That is, in the feasibility study it is important to get a global view of the mineral deposit and major geologic structures are of interest. Little detail can be added to the model because of the limited information available. Once in the mine project stage as more information is added, the detail of the mineral deposit is required in order to plan the location(s) of the major mining processing and auxiliary facilities during the life of the mine. This is of particular importance because after building a processing plant, deploying waste dumps, etc. it is extremely expensive to relocate them. It is possible that during operations, important mineral bodies of economic interest are found below these major facilities which were not identified before by the model. Not mining them would imply a lost opportunity. At this stage, the mine strategy has already been decided and the model is built at a long term mining scale of support called selective mining unit (SMU) which tries to mimic the mining selection during this 3 to 5 years period.

The other model types are the medium term and the short term models. These two models are required to provide information for monthly and weekly mine planning, respectively. More exploratory information is added as well as production information such as blast-hole data. The two types of models can be compared to past production in order to verify the goodness of the modeling strategy and improving the understanding of the mineral deposit. For the medium term model the scale of support can be similar to or smaller than that used in the long term models. However, for the short term model the scale of support is smaller, and is more evident in highly variable deposits such as skarn type. Notice the degree of accuracy increases as the mine planning time period is reduced. For these types of models, particularly the short term model, the geologic features have to be reproduced more accurately in order to be consistent with the

production rate. If these models are either over-estimated or under-estimated in terms of economic mineral quantity and quality, profit decreases. Therefore, at these stages the models have to be as accurate as possible.

1.1. Problem Setting

This thesis is focused on building medium and short term models. Local reproduction of the geologic features is an important characteristic for such types of models. In practice, the mineral deposit is sub-divided in geologic domains. Each domain tends to share common characteristics in terms of the variables of interest. For example in a case of a skarn deposit, calcium carbonate, exoskarn, intrusive, post mineral, etc. rock types are grouped. Usually, one or two domains contain highly or regular concentrations of the variable of economic interest. A linear estimator is used either for estimation or simulation. It consists of calculating a set of weights for each available data sample in the domain. These weights represent the influence of the dataset for predicting at the unsampled location, the closer the data sample is to the unsampled location the larger is the magnitude of the calculated weight.

In this thesis, reference to *conventional geostatistics* refers to the use of simple kriging for both estimation and simulation. There are also other types of kriging estimators such as ordinary kriging, universal kriging, and others; however, they are used only for estimation. With the use of simple kriging it is intended to minimize the global error variance in the domain. However, the reproduction of the spatial variability features is on average. Particular variability features in some regions tend to be averaged to the rest of the domain. It is because a unique pattern of spatial variability is used, this is the covariance/semivariogram model. One of the consequences is the local estimation variances are a function of the surrounding data configuration rather than of the variability of the data values. Some approaches have been proposed to fix this particular problem. (Pan & Arik, 1993) proposed restricted kriging for controlling the influence of the extreme values during the estimation, (Yamamoto, 2000) proposed an alternative way to calculate the kriging variance based on the sample values, among others. Even when small modifications to the kriging estimator are made, the benefits also involve drawbacks. For instance, the characteristic of minimizing the global error variance is less valid, and, simulation cannot guarantee covariance reproduction.

In this thesis an approach for modeling using conventional geostatistics that accounts for the local features of the domain is presented. Even when the local estimation variances and local means in the proposed approach are a function of the variability of the surrounding data, the characteristics of the kriging estimator are not compromised. Therefore, both estimation and simulation can be carried out without any consistency problems. This is important for modeling medium and short term models, because conventional simulation was used for reproducing *on average* the global features of the domain, making it unsuitable for building medium and short term models.

1.2. Objectives of the Research

The goals of the research presented in this thesis are: (1) propose a set of tools that can be used for testing the stationary conditions in a domain which are a requisite for using conventional geostatistics for modeling and (2) provide a methodology for modeling

mineral geologic domains via estimation and simulation that account for the local geologic features in the domain. These models are intended to be used as an input of medium and short term mine planning processes.

Modeling a non-stationary geologic domain using conventional geostatistics will lead to non-optimal results for many reasons. Mean trends make the range of the semivariogram to be unnecessarily large and anisotropic patterns can be incorrectly inferred, and the sill of the semivariogram of the domain is inflated. In presence of variance trends, the conditional variances are overestimated in some regions and underestimated in others. In presence of highly variable small sub-regions, the range of the semivariogram tends to be reduced, etc. With the use of the proposed tools, the stationary conditions of the domains can be verified, so that the domains can be pre-processed for conventional geostatistics modeling.

After removing mean and variance trends, even when regions with different variability patterns are identified in the domain they may be very difficult to separate them into sub-domains, due to the scale of the sub-structures or lack of information to support the sub-domaining decision. Geologically, the data samples in the sub-regions have different influence than those of the major part of the samples in the domain. Conventional geostatistics tends to average these local features making it unsuitable for medium and short term mine planning. The proposed approach is able to build estimated or simulated models that accounts for the local features in the domain, so that, these models can be used for mine planning at a medium and short term scale.

1.3. Proposed Approach

The mathematical condition that allows the use of conventional geostatistics is called stationarity. In practice, domaining is used for sub-dividing the mineral deposit into more stationary sub-regions so that it becomes suitable for modeling using conventional geostatistics. However, sub-domaining is usually based on geologic knowledge and non-stationary features such as trends and others that are characteristic of a mathematical environment can be omitted during this process. A set of tools are presented for verifying the impact of the non-stationary conditions in the initially proposed geologic domains in order to improve the domain definition of the mineral deposit.

In practice even when the impact of the non-stationary features has been minimized in the domains, conventional geostatistics still models uncertainty in global terms. In this thesis the local features of the domains are accounted for by using extra dimensions as required. That is, the domain is moved from its original dimensional space of usually up to three dimensions (e.g. east, north and elevation) to a higher dimensional space where the spatial variability can be described both locally and globally by a covariance model. Recall this is not the case in practice, where the covariance model describes the spatial variability only in global terms. Once the domain is in this high dimensional space, conventional geostatistics can be applied for modeling uncertainty without any modification to the simple kriging system. Finally, the resulting model in the high dimensional space are projected to the initial space with the benefits that after the projection the local features captured by the dataset are present in the model both in estimation and on each realization of simulation. The extra dimensions are the representation of the local variability that is not captured by the global spatial covariance model.

1.4. Dissertation Outline

Chapter 2 describes the conditions of stationarity required for modeling a domain using conventional geostatistics. The concepts of regionalized variables, stationary random functions, estimation and simulation for modeling continuous variables are summarized and discussed.

Chapter 3 presents a tool for calculating the impact of mean and variance trends in the domain. The experimental semivariogram is used to account for trends both in the mean and in the variance. Different artificial trend cases and two real examples are discussed.

Chapter 4 introduces a methodology for calculating a representative experimental semivariogram that accounts for the spatial continuity of the major part of the domain. Because the experimental semivariogram averages the spatial continuity, small highly variable regions in the domain may tend to affect negatively the experimental semivariograms. These highly variable regions can be identified in the domain and considered for sub-domaining. Three styles of semivariogram model fitting are introduced and the impact of them in modeling is discussed.

Chapter 5 discusses the process of moving the domain to a highly dimensional space where the semivariogram model defines the spatial continuity both for local and global estimation. An algorithm based on cross-validation is proposed.

Chapter 6 proposes a methodology for modeling the domain in the high dimensional space and projecting the results to the initial space. Topics about irregular anisotropic patterns are presented; this is of particular importance because it gives flexibility for calculating conditional distributions.

Chapter 7 contains the conclusions of this thesis and proposes new topics for future work.

2. Summary of Theory Related to Modeling Mineral Deposits

The present chapter is focused on the theoretical background related to conventional geostatistics for simulating mineral deposits. Theoretical aspects of random functions, decision of stationarity, estimation and simulation are presented. This thesis is focused on the problems in the implementation of simulation for building mineral deposit models for medium and short term mine planning purposes, where the decision of stationarity has a big impact in mine production. Topics such as models of coregionalization, indicator approaches for modeling continuous and categorical data, and multiple point geostatistics are not covered because they are considered beyond the scope of this thesis. However, the reader can review the following references (Goovaerts, 1997), (Deutsch & Journel, 1998), (Chilés & Delfiner, 1999), and (Deutsch. 2002) among others.

2.1. Random Variables and Random Functions

A random variable (RV) is the representation of a range of values, discrete or continuous, that may occur at a particular location $\mathbf{u}$ that follows a probability distribution. The discrete variable consists of a number of finite outcomes k. Its distribution consists of proportions of occurrence (2.1). When modeling discrete RVs, indicator variables may be used, they represent the probability of occurrence of a particular outcome in the RV (2.2). Continuous RVs values follow a distribution and are characterized by a probability $f(\mathbf{u};z)$ of an outcome z to happen or a cumulative probability of the outcome is no greater than a certain threshold z_0 (2.3). This thesis is focused on the modeling of continuous RV.

$$p\left(\mathbf{u};z_k\right)=\text{Prob}\left\{Z\left(\mathbf{u}\right)=z_k\right\} \tag{2.1}$$

$$I\left(\mathbf{u};z_k\right)=\begin{cases}1 & \text{if } Z\left(\mathbf{u}\right)=z_k\\ 0 & \text{otherwise}\end{cases} \tag{2.2}$$

$$F\left(\mathbf{u};z_0\right)=\text{Prob}\left\{Z\left(\mathbf{u}\right)\le z_0\right\} \tag{2.3}$$

For simplicity the RV is usually denoted as $Z(\mathbf{u})$ and a realization as $z(\mathbf{u})$. A RF can be seen as a set of RVs, where each RV is located at $\mathbf{u}$. The domain D where $\mathbf{u}$ belongs is infinite, however, for implementation the domain D can be assumed as finite and usually

it is referred to as *field of interest*. As part of nomenclature, when the number of dimensions of the location vector $\mathbf{u}$ is one the RF is usually named as *stochastic process* and when there is more than one dimension involved in $\mathbf{u}$ the RF is named *random field* (Chilés & Delfiner, 1999).

In conventional geostatistics the variable of interest in a domain is assumed to be one realization of a RV. Any variable that is distributed in space is called *regionalized*. A regionalized variable is assumed to be representative of a certain process and the process that generates the regionalized variable is called *regionalization* (Journel & Huijbregts, 1978). In mining applications the regionalized variables are any continuous variable of the mineral deposit that is required to be estimated for mine planning or economic evaluation and the regionalization process refers the geologic events that leaded to the formation of the mineral deposit.

The RV $Z(\mathbf{u})$ is characterized by its cdf (2.3), whereas, the RF is characterized by the multivariate distribution of its RVs (2.4) (Deutsch, 2002). Expression (2.4) is the characterization of the joint distribution for the N values $z(\mathbf{u}_1), \ldots, z(\mathbf{u}_N)$. The multivariate distribution constitutes the spatial law of the RF. However, in practice the analysis involves no more than two locations and their respective moments (Goovaerts, 1997).

$$F\left(\mathbf{u}_1,\ldots,\mathbf{u}_N;z_1,\ldots,z_N\right)=\Pr\left\{Z\left(\mathbf{u}_1\right)\le z_1,\ldots,Z\left(\mathbf{u}_N\right)\le z_N\right\} \tag{2.4}$$

In mining applications only the first two moments of the spatial law are sufficient for solving the problems presented (Journel and Huijbregts, 1978).

First order moment. It is also called expectation and is defined as the mean of the distribution of the RV at location vector $\mathbf{u}$ (2.5); therefore it is a function of the location $\mathbf{u}$ and is assumed to exist.

$$E\left\{Z\left(\mathbf{u}\right)\right\}=m\left(\mathbf{u}\right) \tag{2.5}$$

Second order moments. There are three second order moments that are considered in geostatistics. They are:

1. The variance of the RV (2.6). Similar to the first order moment, it is also a function of the location vector $\mathbf{u}$ and is assumed to exist. It is defined as the second order moment about the expectation of the RV.

$$Var\left\{Z\left(\mathbf{u}\right)\right\}=E\left\{\left[Z\left(\mathbf{u}\right)-m\left(\mathbf{u}\right)\right]^2\right\} \tag{2.6}$$

2. The covariance between two RVs (2.7). Because of the use of one attribute in the RF it is also known as *auto-covariance* and for this case it is a function of two locations $\mathbf{u}_1$ and $\mathbf{u}_2$ of the RF. Notice that the variance is a case of the covariance when $\mathbf{u}_1=\mathbf{u}_2$. Both the covariance and the correlation coefficient are measures of linear dependence, non-linear and non-monotonic relationships are not adequately represented by them (Deutsch, 2002).

$$C\left(Z\left(\mathbf{u}_1\right),Z\left(\mathbf{u}_2\right)\right)=E\left\{\left(Z\left(\mathbf{u}_1\right)-m\left(\mathbf{u}_1\right)\right)\left(Z\left(\mathbf{u}_2\right)-m\left(\mathbf{u}_2\right)\right)\right\} \tag{2.7}$$

3. The variogram between two RVs. It is defined as the variance of the increments between two RVs, for this case it is a function of two locations $\mathbf{u}_1$ and $\mathbf{u}_2$ of the RF (2.8). The nomenclature γ is referred to as the semivariogram. In this thesis semivariogram and variogram are used as two different terms, however, in many documents the term variogram refers to as the semivariogram. It is worth to mention in the Geostatistic Congress in Avignon 1988 it was agreed to use of variogram instead of semivariogram because is less pretentious and more descriptive.

$$2\gamma\left(Z\left(\mathbf{u}_1\right), Z\left(\mathbf{u}_2\right)\right) = Var\left\{Z\left(\mathbf{u}_1\right) - Z\left(\mathbf{u}_2\right)\right\} \tag{2.8}$$

In Figure 2-1 a sketch of a RF is presented. Assuming the expectation of the RVs exists, the local means are drawn as a continuous line along the RF. The variances of the RVs are assumed to be finite; they represent the dispersion of the outcomes on each RV. The two RVs at locations $\mathbf{u}_1$ and $\mathbf{u}_2$ show the shape distributions of the RVs can be different as well as their local means. Assuming reality as one realization of the RF means one outcome of each of the RVs is drawn while preserving the spatial law or in other words preserving the first and second order moments of the RF. The process of generating realizations is addressed later in Section 2.4.

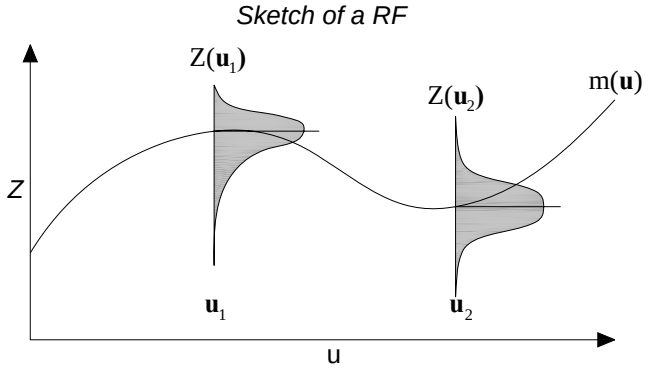

Figure 2-1: Sketch of a RF with to RVs at locations $\mathbf{u}_1$ and $\mathbf{u}_2$, the local means $m(\mathbf{u})$ of the RVs are drawn as a continuous line and the pdfs of the two random variables.

2.2. Decision of Stationarity

In mining, only one outcome at each location $\mathbf{u}$ is available. That is, there is no more than one metal grade at a certain borehole core sample location. Defining the spatial law of the RF under such conditions becomes a very difficult task. The spatial law of the RF has to be inferred with limited data in order to be able to model the deposit using conventional geostatistics approach.

Second order stationarity assumption entails 1) the expectation of all the RVs is constant (2.9), and 2) the covariance of the RF is not a function of locations $\mathbf{u}_1$ and $\mathbf{u}_2$ but of a separation vector $\mathbf{h} = \mathbf{u}_1 - \mathbf{u}_2$ along the domain, that is, the covariance of all the RVs separated by $\mathbf{h}$ is constant (2.10). Once a RF satisfies these two conditions it is said that the RF is stationary of order two (Goovaerts, 1997).

$$E\left\{Z\left(\mathbf{u}\right)\right\} = m\,,\forall \mathbf{u} \in D \tag{2.9}$$

$$C\left(\mathbf{h}\right) = E\left\{Z\left(\mathbf{u}\right)Z\left(\mathbf{u}+\mathbf{h}\right)\right\} - m^2\,,\forall \mathbf{u} \in D \tag{2.10}$$

Stationarity in the covariance also implies 1) the variances of the RVs are constant (2.11) and 2) there is a direct relationship between the covariance and the semivariogram (2.12).

$$Var\left\{Z\left(\mathbf{u}\right)\right\} = C\left(0\right) \tag{2.11}$$

$$\gamma\left(\mathbf{h}\right) = C\left(0\right) - C\left(\mathbf{h}\right) \tag{2.12}$$

A RF is stationary in the strict sense when the spatial law of all the RVs is invariant under translation: $f_{z_1,\ldots,z_N}\left(Z(\mathbf{u}_1),\ldots,Z(\mathbf{u}_N)\right) = f_{z_1,\ldots,z_N}\left(Z(\mathbf{u}_1+\mathbf{h}),\ldots,Z(\mathbf{u}_N+\mathbf{h})\right)$ In conventional geostatistics once any RF is stationary of order two, it can be referred to as a stationary random function SRF. The same nomenclature is used in this thesis.

A RF is said to be intrinsic (IRF) when 1) the expectation of the RVs is constant (2.9) and 2) the variogram is constant for any pair of RVs separated by a separation vector $\mathbf{h}$ (2.13). A SRF implies an IRF, although, the converse is not correct (Journel & Huijbregts, 1978). The IRF is of particular importance in this thesis, especially in the next chapter for dealing with mean and variance trends in the RVs.

$$2\gamma\left(\mathbf{h}\right) = Var\left\{Z\left(\mathbf{u}\right) - Z\left(\mathbf{u}+\mathbf{h}\right)\right\} \tag{2.13}$$

In Figure 2-2 a sketch of SRF is presented. Even when the stationary condition of the covariance is not presented in the picture it is of particular importance to highlight that the local means and variances of the RVs are constant. It can be compared with Figure 2-1 in order to see the big difference between them. In the picture, the confidence limits of the RVs are plotted as constant in order to show the variances of the RVs are also constant. However, this is not a characteristic that is always present in a SRF. In fact, the shape of the distributions of the RVs can be different to each other. Recall that nothing is established about the shape or confidence limits of the RVs distributions of a SRF.

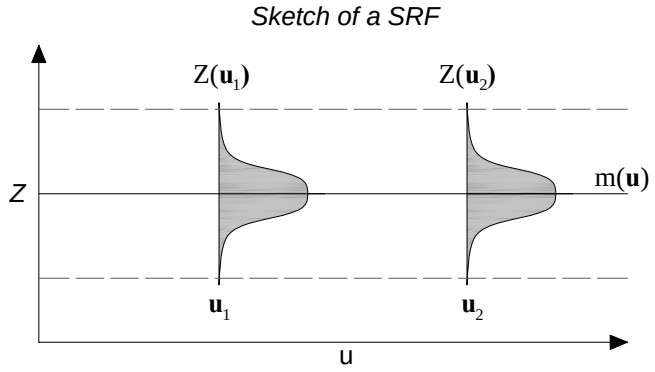

Figure 2-2: Sketch of a SRF with two RVs at locations u_1 and u_2, the constant confidence limits (gray dashed lines) of the RV distributions imply the local variances are constant, although this condition is not always true.

For implementation, the covariance of the SRF is inferred from the available data by fitting the experimental semivariogram (2.14) with a licit function. Notice that the experimental covariance can be calculated instead and fitted also with a licit function. The historical preference in the use of the semivariogram by geostatisticians is because the semivariogram is less demanding than the covariance, that is, only the stationarity of the increments is required (2.13) (Deutsch and Journel, 1998).

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n(\mathbf{h})}\sum_{i=1}^{n(\mathbf{h})}\left[z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h})\right]^2 \tag{2.14}$$

The experimental semivariogram is plotted for different lengths of the separation vector $\mathbf{h}$ (see Figure 2-3). In practice the lengths of the separation vector are usually picked using regular increments called lag distances $\mathbf{l_h}$, that is, $\mathbf{h}_i = i \times \mathbf{l_h}, \forall\, i > 0$. If the existing data is distributed in a sparse configuration the use of angular and distance tolerances are required. The combinations of tolerances act as search volumes for selecting the cloud semivariogram points, $(1/2)\left[z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h}_j)\right]^2$ at $\mathbf{h}_j$ (Deutsch and Journel, 1998). However, it would be more convenient to select the cloud semivariogram points based on clusters in the cloud semivariogram configuration. The clusters need not be aligned in any particular directions unless $\mathbf{u} \in \mathbb{R}^1$, in $\mathbb{R}^2$ the representative clusters define the semivariogram model as a surface and in $\mathbb{R}^3$ as a volume. A semivariogram model is required because the experimental semivariogram only shows the semivariogram at the calculated $\mathbf{h}_j$ vectors, for estimation/simulation the semivariogram has to be known for any $\mathbf{h}$. Finally, the process of fitting a licit semivariogram function to an experimental semivariogram is a subjective task.

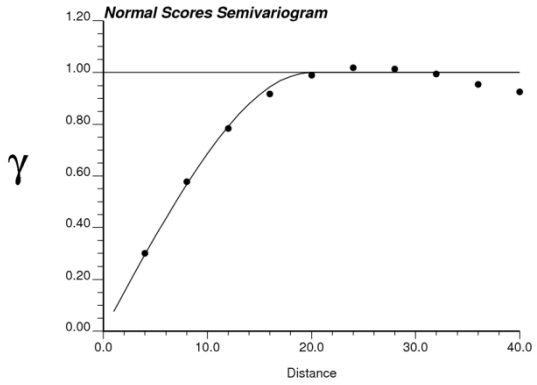

Figure 2-3: Example of an experimental semivariogram (black dots) fitted by a spherical semivariogram model (solid line)

Not just any function or semivariogram model can be used for fitting an experimental semivariogram. This semivariogram model should not allow for negative variances. Recall under the decision of stationarity the covariance can be linked to the semivariogram (2.12). The variance of any linear combination of RVs can be also expressed in terms of the covariances between the locations of such RVs (2.15). In order for the variance to be a non-negative value, the covariance model has to be positive semi-definite (Goovaerts, 1997).

$$Var\left\{\sum_{i=1}^{n}\lambda_i Y\left(\mathbf{u}_i\right)\right\}=\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C\left(\mathbf{u}_i,\mathbf{u}_j\right) \tag{2.15}$$

Expression (2.15) can be rewritten in terms of the semivariogram considering (2.12).

$$Var\left\{\sum_{i=1}^{n}\lambda_i Y\left(\mathbf{u}_i\right)\right\}=C(0)\sum_{i=1}^{n}\lambda_i\sum_{j=1}^{n}\lambda_j-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma\left(\mathbf{u}_i,\mathbf{u}_j\right) \tag{2.16}$$

There are semivariogram models such as the power model which does not have covariance counterpart. For these semivariograms expression (2.16) can still be expressed in terms of the semivariogram by making $\sum_{i=1}^{n}\lambda_i=0$ in order to filter the $C(0)$ term. Then expression (2.16) becomes (2.17). To ensure the non-negativity of the variance of $Y(\mathbf{u})$ the semivariogram has to be negative definite, the condition being $\sum_{i=1}^{n}\lambda_i=0$ (Goovaerts, 1997).

$$Var\left\{\sum_{i=1}^{n}\lambda_i Y\left(\mathbf{u}_i\right)\right\}=-\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\gamma\left(\mathbf{u}_i,\mathbf{u}_j\right) \tag{2.17}$$

A list of common licit semivariogram models are the nugget effect, spherical, exponential and the Gaussian model. It is recommended to use the combinations of them to model more complex structures. From these four models the spherical model is licit up

to $\mathbb{R}^3$; the rest of them have no dimensional limitations. The following semivariogram models are presented with a scaled sill to one.

- Nugget effect.

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{if } \mathbf{h} = 0 \\ 1 & \text{if } \mathbf{h} > 0 \end{cases}$$

- Spherical

$$\gamma(\mathbf{h}) = Sph(\mathbf{h}) = \begin{cases} \left[1.5\mathbf{h} - 0.5\mathbf{h}^3\right] & \text{if } \mathbf{h} \leq 0 \\ 1 & \text{if } \mathbf{h} > 0 \end{cases}$$

- Exponential

$$\gamma(\mathbf{h}) = Exp(\mathbf{h}) = 1 - e^{-\mathbf{h}}$$

- Gaussian Model

$$\gamma(\mathbf{h}) = Gau(\mathbf{h}) = 1 - e^{-\mathbf{h}^2}$$

The presented semivariogram models are presented in its isotropic form, in presence of anisotropic preferential directions the separation vector **h** is scaled by a rotation and anisotropic matrices (see Chapter 6).

2.3. Estimation

In estimation the main goal is to calculate the local distributions at unsampled locations in the domain. In conventional geostatistic, a SRF environment is assumed to the access to the spatial law of it, from which estimates can be inferred using the existing data. The existing data is considered as a regionalized variable or part of one realization of the SRF. This assumption limits the set of possible realizations to the ones that honour the existing data at their respective locations. The set of possible realizations is variable at the unsampled locations and such variability is then described by the conditional distributions calculated using a least square optimization method known as kriging.

Kriging is least square regression method and was named after the pioneering work of Danie Krige (1951). It calculates the mean of the estimates at the unsampled location based on a linear estimator form (2.18).

$$z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i \left[z(\mathbf{u}_i) - m(\mathbf{u}_i) \right] \tag{2.18}$$

The vector **u** represents the location to estimate. This location can be anywhere in the domain, even at the locations of the existing data. In that case, the kriging estimate of the mean $z^*(\mathbf{u}_i)$ results in the exact value of the existing data at location $\mathbf{u}_i$,. It is a property that makes kriging an exact estimator. The local means $m(\mathbf{u})$ and $m(\mathbf{u}_i)$ are the

expectations of the RVs of the RF at the estimation and existing sample locations respectively. Recall that under assumption of stationarity all the local means are constant, that is $m(\mathbf{u}) = m(\mathbf{u}_i) = m$. The indices i represent the order of the existing data, $i \in \{1, \dots, n\}$, for the n existing sampled values $z(\mathbf{u}_i)$. The $\lambda_i(\mathbf{u})$ variable acts as weights for each of the existing data. They define the influence of each sample $z(\mathbf{u}_i)$ based on the spatial law of the RF with respect to location $\mathbf{u}$. The λ_i weights are the ones calculated using the kriging approach.

In the kriging approach for simplicity the estimates are based on its residuals, it is, $y(\mathbf{u}) = z(\mathbf{u}) - m$. Then expression (2.18) is rewritten as (2.19).

$$y^*(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i y(\mathbf{u}_i) \tag{2.19}$$

The λ_i weights are calculated by minimizing the variance of the errors in estimation under a condition of unbiasedness (2.20). The variance of the errors in estimation or error variance is defined as the variance of the difference between the residual of the true value at location $\mathbf{u}$ and its respective estimate (2.21). Because the true values $y(\mathbf{u})$ are unknown the optimization problem is solved in terms of expectation.

$$E\left\{y(\mathbf{u}) - y^*(\mathbf{u})\right\} = 0 \tag{2.20}$$

$$\sigma_e^2(\mathbf{u}) = Var\left\{y(\mathbf{u}) - y^*(\mathbf{u})\right\} \tag{2.21}$$

The error variance that accounts for the condition of unbiasedness is then (2.22).

$$\sigma_e^2(\mathbf{u}) = E\left\{\left[y(\mathbf{u}) - y^*(\mathbf{u})\right]^2\right\} \tag{2.22}$$

A discussion about the interpretation of error variance can be found in (Deutsch, 2002). For calculating the λ_i weights that minimize the error variance and account for the unbiasedness condition the partial derivatives of (2.22) are calculated with respect to the λ_i weights.

$$\frac{\partial\left[\sigma_e^2(\mathbf{u})\right]}{\partial\lambda_i} = 0 = 2\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) - 2C(\mathbf{u}, \mathbf{u}_i),\ i = 1, \dots, n$$

$$\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i, \mathbf{u}_j) = C(\mathbf{u}, \mathbf{u}_i),\ i = 1, \dots, n \tag{2.23}$$

Notice that in expression (2.23) there are n variables and n equations. For implementation (2.23) can be expressed in matrix form (2.24).

$$\begin{bmatrix} C(\mathbf{u}_1,\mathbf{u}_2) & \cdots & C(\mathbf{u}_1,\mathbf{u}_n) \\ \vdots & \ddots & \vdots \\ C(\mathbf{u}_n,\mathbf{u}_1) & \cdots & C(\mathbf{u}_n,\mathbf{u}_n) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C(\mathbf{u},\mathbf{u}_1) \\ \vdots \\ C(\mathbf{u},\mathbf{u}_n) \end{bmatrix}$$

$$C_{11}\lambda = C_{10} \tag{2.24}$$

In expression (2.24) C_{11} is the covariance matrix between the existing data, λ is the vector of simple kriging weights and C_{10} the vector of covariances of the existing data and the location to estimate. The λ_i elements are then used in the linear estimator expression (2.19) to calculate the estimated mean of the RV at location **u**. The variance of the RV at location **u** or the local kriging error variance (2.25) is calculated by replacing (2.23) in (2.22).

$$\sigma_e^2(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i \underbrace{\left[\sum_{j=1}^{n} \lambda_j C(\mathbf{u}_i,\mathbf{u}_j) \right]}_{\text{expression (2.23)}} - 2\sum_{i=1}^{n} \lambda_i C(\mathbf{u},\mathbf{u}_i) + C(0)$$

$$= \sum_{i=1}^{n} \lambda_i \left[C(\mathbf{u},\mathbf{u}_i) \right] - 2\sum_{i=1}^{n} \lambda_i C(\mathbf{u},\mathbf{u}_i) + C(0)$$

$$\sigma_e^2(\mathbf{u}) = C(0) - \sum_{i=1}^{n} \lambda_i C(\mathbf{u},\mathbf{u}_i) \tag{2.25}$$

Estimating at the locations of the domain using the kriging approach gives two parameters of the distributions of the realizations; they are the mean (2.19) and variance (2.25) at location **u**. The kriging approach presented is also known as simple kriging (SK). There are other variants of kriging such as: ordinary kriging (OK), universal kriging (UK), kriging with an external drift (KED), etc. However, they are beyond the scope of this thesis because only the simple kriging approach is used in sequential Gaussian simulation (SGS). Interestingly, OK approach is widely used in mining industry because it is not necessary to know the value of the mean of the RVs, OK can be found in many commercial packages used in the mining industry. One important difference between SK and OK is the minimization of the error variance. OK does not give global minimal solution to the error variance. OK is used only for getting estimated models as well as the other kriging variants. The other kriging approaches or kriging flavours can be found in classic geostatistical literature such as Journel & Huijbregts, (1978), Goovaerts, (1997), Deutsch and Journel, (1998), Chilés & Delfiner, (1999), Deutsch, (2002), among others.

2.4. Simulation

Models of estimates give two uncertainty parameters, estimation mean and estimation variance for each location **u** in the domain. However, when proposing a mine plan, uncertainty information about alternative scenarios can be of special importance. For instance, for medium and short term mine planning processes where the time period to make decisions about production is very limited, analyzing different case scenarios is

much more informative than analyzing only estimates. In simulation, realizations of the SRF that honour the existing data at their respective locations are produced.

Drawing one realization from a SRF is equivalent to sampling from the joint multivariate distribution of N RVs variables. Such a large joint distribution can be decomposed into small pieces by using the Bayes' theorem recursively (2.26) (Leuangthong, 2009).

$$\begin{aligned} P\left(A_1,...,A_N\right) &= P\left(A_N \middle| A_1,...,A_{N-1}\right)\cdot P\left(A_1,...,A_{N-1}\right) \\ &= P\left(A_N \middle| A_1,...,A_{N-1}\right)\cdot P\left(A_{N-1} \middle| A_1,...,A_{N-2}\right)\cdot P\left(A_1,...,A_{N-2}\right) \\ &\cdots \\ &= P\left(A_N \middle| A_1,...,A_{N-1}\right)\cdot P\left(A_{N-1} \middle| A_1,...,A_{N-2}\right)\cdots P\left(A_2 \middle| A_1\right)\cdot P\left(A_1\right) \end{aligned} \tag{2.26}$$

After decomposing the N multivariate joint distribution it is possible to proceed to sample sequentially, that is:

- Draw one sample A_1 from the marginal distribution of $P(A_1)$, then
- Draw one sample A_2 from the conditional distribution $P(A_2|A_1 = a_1)$, then
- Draw one sample A_3 from the conditional distribution $P(A_3|A_1 = a_1, A_2 = a_2)$, then
- …
- Draw one sample A_3 from the conditional distribution $P(A_N|A_1 = a_1, A_2 = a_2, \dots, A_{N-1} = a_{N-1})$

For implementation, under assumption of multi-gaussianity of the RF the conditional distributions can be estimated using the SK approach and Monte Carlo simulation. In fact, the multi-gaussian assumption is of special importance because all the conditional distributions are Gaussian and can be fully parameterized by the SK mean and SK variance.

Notice that as sampling proceeds, the number of data samples for estimating the conditional distribution parameters increases. The SK system becomes more cumbersome to solve in practice because it requires the inversion of larger and larger matrices. To overcome this problem, search radius, search strategies, limits in the number of conditioning samples, etc. are implemented at the cost of compromising other features such as the semivariogram model reproduction, fair reproduction of conditional means and variances, etc. It is of special importance to properly set up the sequential simulation algorithm in order to balance cost in time and benefit in quality of the results. One of the most widely used software packages for simulation is SGSIM which is part of GsLib (Deutsch and Journel, 1998), because of the optimization in the programming code this program performs well for relatively large grid definitions of the domain. Implementation algorithms are discussed in Goovaerts, (1997), Deutsch and Journel, (1998), Zanon, (2007), Chilés & Delfiner, (1999) discuss a methodology for conditioning, via SK, unconditional realizations of a SRF. This way unconditional realizations generated by any technique can be conditioned by the SK estimates of mean and variance in order to get conditioned realizations.

Because the attribute $y(\mathbf{u})$ of the SRF is assumed to be a regionalized variable, each of the realizations is considered as one plausible representation of reality. They have to

reproduce the second order moments of the SRF, that is, the semivariogram and consequently the covariance model. In the case of a multi-gaussian RF the global distribution is Gaussian $\sim N(0,1)$. On average, the realizations at the location **u** in the domain have to reproduce the simple kriging estimates of mean and variance.

In the next chapter an approach for identifying non-stationary features in a domain is presented. Variations in the local means and in the local variances are verified and discussed. Other aspects such as stationarity of the increments or intrinsic hypothesis are covered in Chapter 4.

3. Detection of Mean and Variance Trends

In Chapter 2 the conditions of stationarity are discussed. Ideally, a domain should satisfy those conditions in order to model uncertainty if using conventional geostatistics. Unfortunately, in real cases, domains may be influenced by physical conditions that make difficult to assume stationarity in the domain. In this chapter the stationary conditions of constant local mean and constant local variance are verified from the available dataset of the domain. Centered variograms and non-centered variograms are calculated and compared in order to measure and address the consequences of assuming constant mean and variance trends when calculating the spatial continuity in the domain. Trends in the mean and/or variance are a longstanding challenge in mineral resource modeling. Geologic modeling and/or grade domaining are often employed to account for heterogeneities in mineral grades; however, this is only practical up to a certain point since the number of data available for reliable model construction can be severely reduced. In such cases, trends remain present in certain geologic/grade domains and directly impact the modeling methodology; the trend must either be handled explicitly or a suitable approach is adopted to implicitly account for it. This is highly dependent on understanding the source of the trend. Modeling uncertainty in such conditions, using conventional geostatistics leads to wrong results such as inflation of local variances and an incorrect definition of the corresponding semivariogram model among others. In this chapter it is proposed an approach to identify the source of a trend by simply decomposing the semivariogram and quantifying the impact of the local mean and local variance on the semivariogram. This useful assessment can then be accounted for by the choice of modeling methodology. For instance, a trend in the local mean can be mitigated by removing it from the domain while a trend in the variance can be handled by re-scaling the data.

3.1. Sub-Domaining for Mineral Deposits Modeling

In conventional geostatistics, the semivariogram model is used to define the spatial continuity of the domain. This measure is then used to infer metal or mineral grades at unsampled locations via kriging or simulation. A fundamental component of this framework is a decision of stationarity that permits the modeler to extract relevant statistics about the domain using only the available sampled data. Often, an assumption of a SRF, where the mean and covariance are considered constant, is sufficient for inference. In reality, drifts or trends in the domain are common and result in a non-stationary domain. Spatial variations in the local mean are captured by a semivariogram which increases rapidly (Chilés & Delfiner, 1999). However, domains without an apparent drift can still be non-stationary if the local variance is not constant. This

condition is widely seen in practice when two domains are combined; one of them highly variable and the other less variable, both with almost the same local mean.

Consider two different stationary sub-domains with different local variances and same local means. Estimating using conventional geostatistics within the combined domain would lead to wrong results in terms of conditional variances, because the stationary assumption of order two entails the variance of all the RVs in a SRF domain are constant, $C(0) = \sigma^2$, (see Chapter 2). The local uncertainty would tend to be underestimated in the highly variable sub-domain and overestimated in the sub-domain with smaller variability. Similarly, consider two stationary sub-domains with constant global variance and different local means, $m_1(\mathbf{u}) < 0$ and $m_2(\mathbf{u}) > 0$, and a combined global mean, $m(\mathbf{u}) = 0$. For the two sub-domains the estimated local means will tend to the global mean $m(\mathbf{u}) = m$, that is, the local means for the first sub-domain are overestimated and underestimated for the second sub-domain. Further, in a scenario where both the local means and local variances of the RVs are a function of its position vector, the consequences of modeling uncertainty using conventional geostatistics are much more unrealistic.

In practice sub-domaining is carried out without taking into account geologic information which in some cases is so general that it prohibits making any further separation, e.g. exoskarn, when metal grade variability varies for the different types of skarn such as magnetite or garnet among others. That makes some of the resulting domains unsuitable for modeling using conventional geostatistics. Even when the resulting domain is assumed stationary, the estimated uncertainty parameters are inadequate for short and medium term mine planning which require a more accurate prediction or estimation of local uncertainty parameters.

One common natural process that is difficult to model under stationary assumptions is the soft transition of rock types, e.g. mineralized and non-mineralized. The transitions can be hard or soft. A soft transition can be seen as a gradual change in the local mean and/or in the local variance of a RF. On the other hand, a hard transition is an abrupt change usually present when the two sub-domains are independent of each other. A sketch of some possible transitions in a domain that is assumed to be multi-Gaussian is shown in Figure 3-1. In case (1), the different rock types share the same local means and variances, that makes the resulting domain be suitable to be modeled using conventional geostatistics. In case (2), it is shown a case of soft transition between rock types (a) to (c) through (b). Notice that in this case the local mean is constant. In some cases this case is not considered as stationary, only variability in the local means is considered as a unique condition for non-stationarity. In case (3) the local means and variances are assumed as variable with an abrupt change in them. Geologically this case occurs when the two rock types that are assumed to be part of the domain are of independent geologic events e.g. thin post-mineral dikes that cross a mineralized rock type. And case (4) soft transitions of local means and local variances. Case (1) is the only one that can be considered as appropriate enough to be assumed as multi-Gaussian, however, it is the most uncommon in practice. The rest of the cases are simplistic approximations of reality but fairly practical for modeling after re-scaling local mean and variances. Some approaches were presented for dealing with soft boundaries (McLennan, 2007), whereas, for the hard boundary case there is no other solution than sub-domaining because the two sub-domains involved are independent.

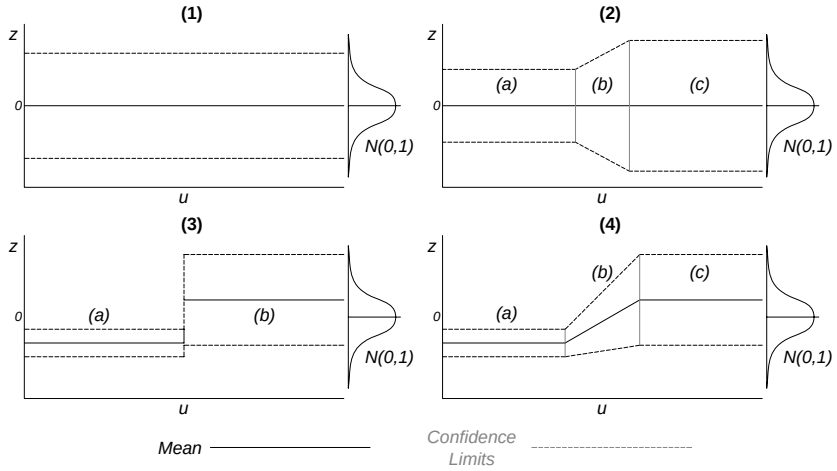

Figure 3-1: schematic 1D Gaussian environments: (1) constant local mean and local variance; (2) two regions of constant local mean but different local variance with a transition zone; (3) two different sub-regions of constant local mean and local variance; (4) two different regions of constant local mean and local variance with a transition zone, the dashed lines represent some confidence limits of a normal distribution.

3.2. Semivariogram in Non-Stationary Environments

The variogram is a second order moment that measures the variability between two locations separated by a vector **h**. It is expressed as the variance of the increment of two random variables (RV) at locations **u** and **u** + **h**, (3.1) (see Chapter 2).

$$2\gamma(\mathbf{u},\mathbf{u}+\mathbf{h}) = Var\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]\right\} \tag{3.1}$$

Experimentally, this is calculated by assuming the variogram of the RF is stationary (see Chapter 2). For this, the calculation is based on the data pairs, $z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})$, found at the two extremes of each particular separation vector, $\mathbf{h}_j$ (3.2). This is often referred to as the method-of-moments approach (Matheron, 1962). However, in practice it is very difficult to find a dataset where data points are separated exactly by **h**. For solving this problem, angular and distance tolerances are used for approximating the experimental semivariogram (Deutsch and Journel, 1998), (Deutsch, 2002). The tolerances define a n-dimensional region for each $\mathbf{h}_j$ for a domain in $\mathbb{R}^n$ in order to capture a representative amount of experimental variogram pairs $0.5[z(\mathbf{u}_i) - \mathrm{z}(\mathbf{u}_i + \mathbf{h})]^2$ from a cloud semivariogram.

$$2\hat{\gamma}(\mathbf{h}) \simeq \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h})\right]^2 \tag{3.2}$$

The calculation of the variogram does not depend on first order stationarity of the RF, in fact, there are some semivariograms that does not have covariance counterparts, such as the power model (Goovaerts, 1997), (Deutsch, 2002), they are the ones so-called unbounded. Historically, the reason why geostatisticians preferred the use of the

semivariogram over covariance is because the former does not require the previous knowledge of the mean of the RVs (Deutsch and Journel, 1998). In the case of an intrinsic random function (IRF), a linear drift is characteristic of it with second order stationary increments (Chilés & Delfiner, 1999). In this case two types of variograms can be obtained, the non-centered (3.3) and the centered (3.4) variograms (Gneiting, Sasvári, & Schlater, 2001). Notice in the absence of a drift, both the centered and non-centered variograms are equal.

$$2\gamma_{NC}(\mathbf{h}) = E\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]^2\right\} \tag{3.3}$$

$$\begin{aligned} 2\gamma_{C}(\mathbf{h}) &= Var\left\{Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right\} \\ &= E\left\{\left[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right]^2\right\} - \left[E\left\{Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right\}\right]^2 \\ &= 2\gamma_{NC}(\mathbf{h}) - \left[E\left\{Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right\}\right]^2 \end{aligned} \tag{3.4}$$

3.3. Impact of non-stationarity in experimental semivariogram calculation

During the calculation of the experimental semivariogram the available dataset is split in two parts. The first part corresponds to the sub-dataset at the head of the separation vector and the second at the tail of the separation vector. When the separation distance is zero, both sub-groups consist of all the conditioning data. Since real data are non-exhaustive, an increase in the separation vector is usually accompanied by a reduction in data pairs to reliably calculate the semivariogram. Ideally, under a decision of second order stationarity, the two sub-datasets at location $\mathbf{u}$ and $\mathbf{u}+\mathbf{h}$ are expected to have the same experimental mean ($m_{\mathbf{u}} \approx m_{\mathbf{u}+\mathbf{h}}$) and variance ($\sigma^2_{\mathbf{u}} \approx \sigma^2_{\mathbf{u}+\mathbf{h}}$). Both the means and variances of the two sub-groups are calculated without using any declustering weights.

The influences of the means and variances of the two sub-groups at the two extremes of the separation vector $\mathbf{h}$ can be obtained by expanding expression (3.2). Notice the experimental semivariogram is used instead of the experimental variogram; this is consistent with industry practice.

$$\begin{aligned} \hat{\gamma}(\mathbf{h}) &= \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[\left(z(\mathbf{u}_i)\right)^2 - 2z(\mathbf{u}_i)z(\mathbf{u}_i+\mathbf{h}) + \left(z(\mathbf{u}_i+\mathbf{h})\right)^2\right] \\ &= \frac{1}{2N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[\left(z(\mathbf{u}_i)\right)^2 + \left(z(\mathbf{u}_i+\mathbf{h})\right)^2\right] - \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_i)z(\mathbf{u}_i+\mathbf{h})\right] \\ &= \frac{1}{2}\left[\sigma^2_{\mathbf{u}} + \sigma^2_{\mathbf{u}+\mathbf{h}}\right] + \frac{1}{2}\left[m_{\mathbf{u}} - m_{\mathbf{u}+\mathbf{h}}\right]^2 - C(\mathbf{h}) \end{aligned} \tag{3.5}$$

Under stationarity, we know that $\gamma(\mathbf{h}) = C(0) - C(\mathbf{h})$; a comparison of this to (3.5) shows that in presence of a drift the global variance of the domain, assuming the available dataset is fully representative of the domain, is modified to $0.5\left(\sigma^2_{\mathbf{u}} + \sigma^2_{\mathbf{u}-\mathbf{h}}\right) + 0.5(m_{\mathbf{u}} - m_{\mathbf{u}+\mathbf{h}})^2$. If the mean is stationary, then the variance of the domain is expressed

as the average of the respective variances of the two sub-groups, that is, $0.5\left(\sigma_{\mathbf{u}}^2 + \sigma_{\mathbf{u-h}}^2\right)$. Notice that the half squared difference of the means of the sub-groups, $0.5(m_{\mathbf{u}} - m_{\mathbf{u+h}})^2$, makes the variance of the dataset larger in presence of a trend in the mean. If the discrete form of the centered semivariogram (3.4) is used instead of the stationary non-centered (3.3) this component is canceled out in (3.5). (Gneiting, Sasvári, & Schlater, 2001) showed that the influence of a trend in the mean in a quadratic form can be filtered using the centered semivariogram.

$$\gamma_{NC}\left(\mathbf{h}\right) = Q\left(\mathbf{h}\right) + \gamma_C\left(\mathbf{h}\right) \tag{3.6}$$

where, $Q(\mathbf{h}) = \sum_{i=1}^{d} a_i h_i^2$, $\mathbf{h} = [h_1, \dots, h_d]^{\mathrm{T}} \in \mathbb{R}^d$ and $a_i \geq 0$ for $i = 1, \dots, d$.

From expression (3.5) the semivariogram can be re-written as a function of the differences of the standard deviations and difference of the means of the sub-groups at locations $\mathbf{u}$ and $\mathbf{u} + \mathbf{h}$ (3.7). The global variance $0.5\left(\sigma_{\mathbf{u}}^2 + \sigma_{\mathbf{u-h}}^2\right)$ is now expressed as $0.5(\sigma_{\mathbf{u}} - \sigma_{\mathbf{u+h}})^2 + \sigma_{\mathbf{u}}\sigma_{\mathbf{u+h}}$. As discussed previously, the only increment to the variance comes from the mean trend component; therefore, the trend in the variance cannot make the experimental semivariogram be greater than the sill.

$$\hat{\gamma}\left(\mathbf{h}\right) = \frac{1}{2}\left[\sigma_{\mathbf{u}} - \sigma_{\mathbf{u+h}}\right]^2 + \frac{1}{2}\left[m_{\mathbf{u}} - m_{\mathbf{u+h}}\right]^2 + \sigma_{\mathbf{u}}\sigma_{\mathbf{u+h}} - C\left(\mathbf{h}\right) \tag{3.7}$$

Assuming multi-gaussianity, expression (3.7) can be re-written so that the correlation coefficient is introduced to the experimental semivariogram expression (3.8).

$$\hat{\gamma}\left(\mathbf{h}\right) = \frac{1}{2}\left[\sigma_{\mathbf{u}} - \sigma_{\mathbf{u+h}}\right]^2 + \frac{1}{2}\left[m_{\mathbf{u}} - m_{\mathbf{u+h}}\right]^2 + \sigma_{\mathbf{u}}\sigma_{\mathbf{u+h}}\left(1 - \rho\left(\mathbf{h}\right)\right) \tag{3.8}$$

From expression (3.7) the experimental semivariogram can be separated into four components: (a) Half of the squared difference of the standard deviations of the two sub-groups, $0.5[\sigma_{\mathbf{u}} - \sigma_{\mathbf{u+h}}]^2$; (b) half of the squared difference of the means of the two sub-groups, $0.5[m_{\mathbf{u}} - m_{\mathbf{u+h}}]^2$; (c) the product of the two standard deviations of the two sub-groups, $\sigma_{\mathbf{u}}\sigma_{\mathbf{u+h}}$; and (d) the covariance between the two sub-groups, $C(\mathbf{h})$. Notice when the means and variances of the two sub-groups are similar (i.e. $\sigma_{\mathbf{u}} \approx \sigma_{\mathbf{u+h}} \approx \sigma$ and $m_{\mathbf{u}} \approx m_{\mathbf{u+h}} \approx m$), the experimental semivariogram expression (3.7) relies only on components (c) and (d) which is close to the stationary form of the semivariogram.

Under stationary conditions for all the **h**-scatter plots of the available dataset, the means and variances of the marginal distributions are expected to be very similar. Systematic differences in them as a function of the separation vector would imply a presence of mean or variance trends in the domain which is captured by the sampled data and from it by the experimental semivariogram (3.7).

The experimental semivariogram is above the variance of the dataset or sill when the correlation between two RV of a SRF is negative (Gringarten & Deutsch, 2001), (Deutsch, 2002). However, from expression (3.7), we can directly quantify the contributions of the trends in the mean of the two sub-groups separated by **h** to the centered experimental semivariogram as well as to capture trends in the variance. Recall that the centered semivariogram is able to filter the effects of mean trends based on differences at **h**. These mean trend contribution can make the experimental

semivariogram be larger than the sill, while still obtaining a positive correlation between the two sub-groups. Expression (3.7) deals with non-stationary RFs which local means and variances exist and vary as a function of the position of their RVs, therefore, the comparison of the semivariograms under the stationary framework is not fair. However, the proposed non-stationary analysis is much more informative in the sense that more information about the nature of the domain can be obtained from the available dataset.

3.4. Case Studies

Three case studies are presented to show the practical applications of the proposed approach. The first case is a synthetic controlled 1D example and the second and third cases are real data applications with mean and variance trends. The first case study presents each trend case isolated from other trend cases. The impact of the presented trends in the experimental semivariogram can be seen in a more practical scenario through the interpretation in the shape of the experimental semivariograms and their respective h-scatter plots.

Case Study 1

Consider a 1D unconditionally simulated dataset of 1000 regularly spaced data points, separated by 1 unit of distance (uod). A spherical semivariogram model of 25 uod is considered (3.9). Three different trend cases are added to this initial dataset (see Figure 3-2):

$$\gamma(\mathbf{h}) = Sph_{25}(\mathbf{h}) \tag{3.9}$$

1. A linear trend in the form $y = ax + b$ with a negative slope a along the entire domain. This case accounts for a large variability in the local means.
2. A symmetrically concave-shaped trend case. This mean trend is presented in order to show the weakness of this approach for accounting for this type of symmetric trends. The limitation of a two point statistics arise as problems for dealing with particular structures.
3. No trends in the mean component, but the variances of the two halves of the domain are re-scaled. The trend in the local variances is usually assumed as not present in a semivariogram analysis.

The experimental semivariograms are calculated using expression (3.7). The semivariogram plots for each case are calculated within a range of one half the size of the domain and the contributions of each component of the experimental semivariogram are color coded for visualization (see Figure 3-3).

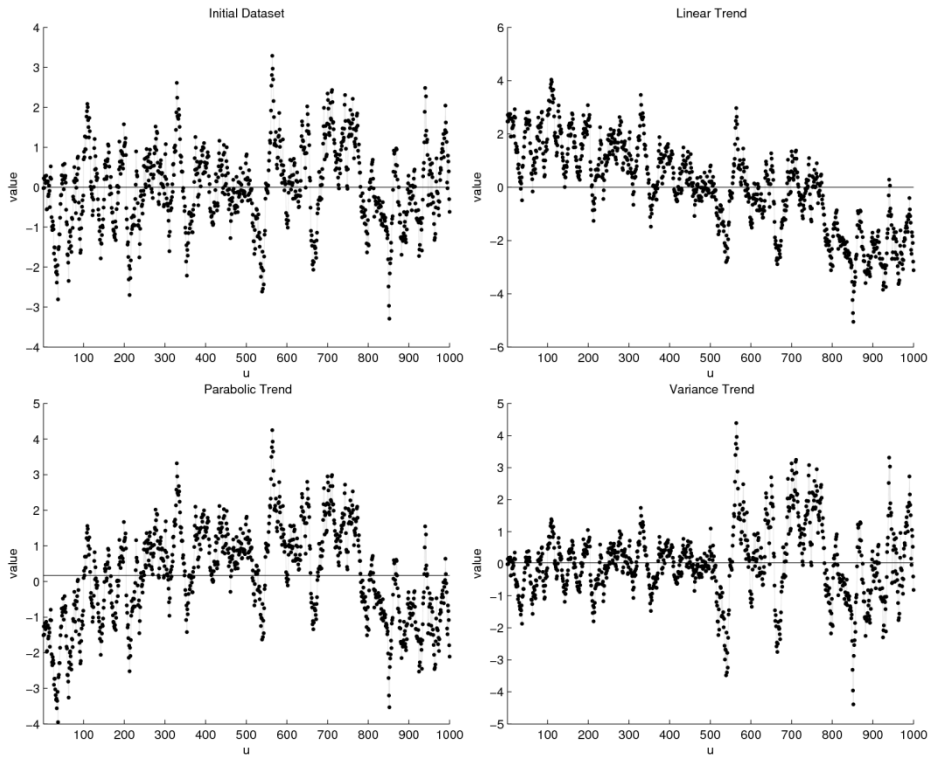

Figure 3-2: Initial dataset (top left) influenced by linear trend (top right), parabolic trend (bottom left) and local variability in variances (bottom right).

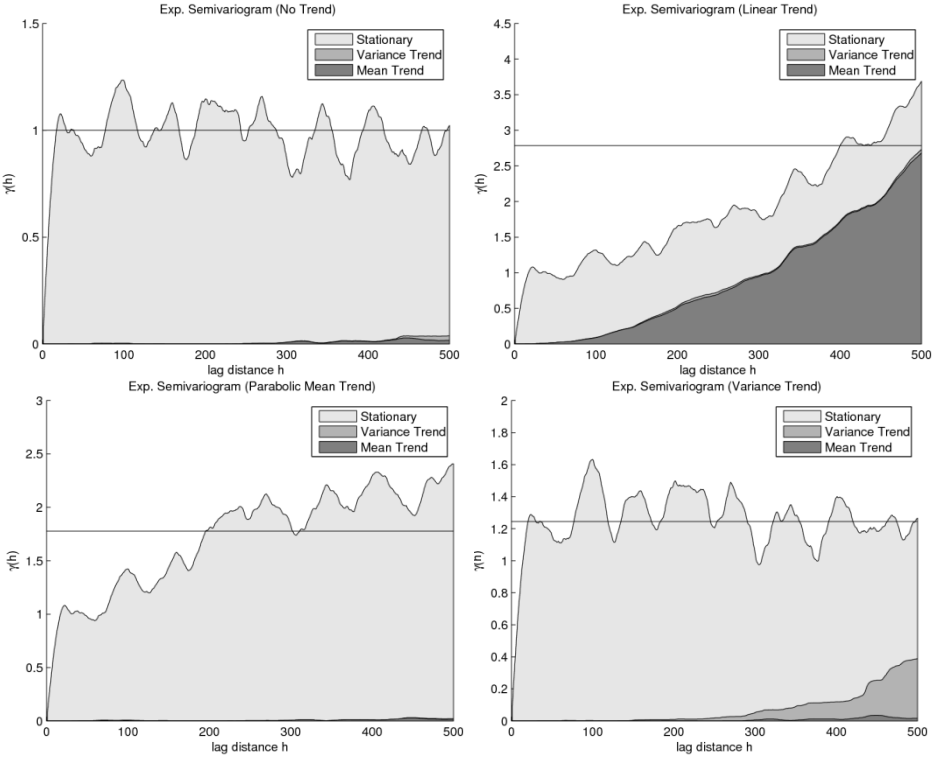

Figure 3-3: experimental variograms for initial dataset (unconditional simulated realization) (top left) and influenced by linear trend (top right), parabolic trend (bottom

left) and locally variable variance (bottom right); in the four plots the blue region represents the variation of the mean trend component, the red region the variation of the variance trend component and the green region the stationary component of the experimental semivariogram.

In the first case (see Figure 3-3 top left) there is no mean or variance trend and the experimental semivariogram reaches the variance of the domain and is relatively constant with some ergodic fluctuations beyond the semivariogram range. The green region represents the stationary part of the experimental semivariogram, $\sigma_{\mathbf{u}}\sigma_{\mathbf{u+h}} - C(\mathbf{h})$, in (3.7) . This is considered as the ideal case of a stationary domain. At each lag separation **h** the **h**-scatter plots the cloud of data pairs are centered along the 45 degree bisector, this implies the local means of the sub-datasets at **u** and $\mathbf{u} + \mathbf{h}$ are similar, as well as the local variances (see Figure 3-4). Some small fluctuations in the correlation coefficients can be seen due to the ergodic fluctuations.

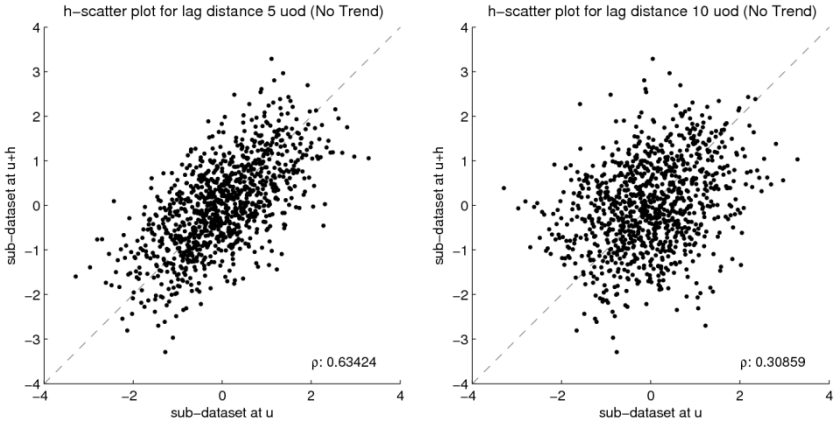

Figure 3-4: h-scatter plots for no mean trend case for lag distances 5 uod (left) and 10 uod (right)

In the second case (see Figure 3-3 top right) the linear mean trend makes the experimental semivariogram increase and at a certain point makes it larger than the variance of the domain. Notice the stationary part remains invariant (green region) and is the mean trend (blue region) the only source of the increment in the experimental semivariogram. When the centered semivariogram is calculated the mean trend contribution is cancelled and only the stationary region remains. The differences in the local means make the data pairs in the **h**-scatter plots are not centered and they are more distant from the 45 degree bisector as the lag distance increases. The experimental semivariogram can be larger than the variance of the domain while the correlation coefficient of the **h**-scatter plot is still positive (see Figure 3-5), however, this is no longer a measure of spatial continuity of a SRF environment.

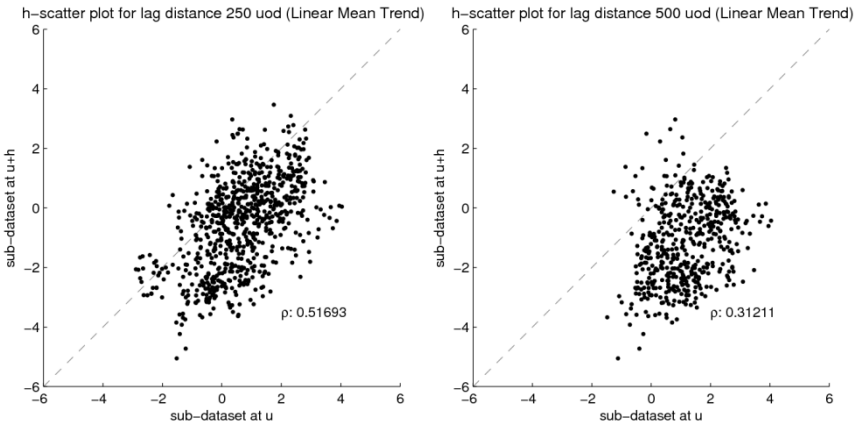

Figure 3-5: h-scatter plots for linear mean trend case for lag distances 250 uod (left) and 500 uod (right)

The third trend scenario (see Figure 3-3 bottom left) results in a very particular case where the contribution of the mean trend component cannot be seen in the experimental plots. Recall that, the mean trend component is captured when differences in the means of the sub-groups appear as the separation distance increases. For this case, this component in the experimental semivariogram is cancelled, $m_{\mathbf{u}} - m_{\mathbf{u+h}}$, therefore, $m_{\mathbf{u}} - m_{\mathbf{u+h}} = 0$, and the trend curve is considered as part of random fluctuation. With this type of trend the semivariograms above the sill have negative correlation coefficients as expected for a SRF (see Figure 3-6).

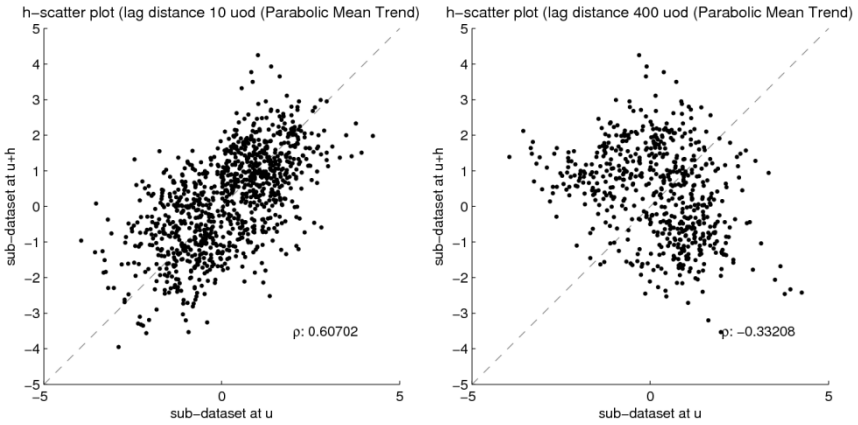

Figure 3-6: h-scatter plots for the parabolic shaped trend case for lag distances 10 uod (left) and 400 uod (right).

In the fourth case the experimental semivariogram does not show any tendency to grow above the sill. This is because the trend in the variance does not contribute to the semivariogram (see Figure 3-3 bottom right). In the **h**-scatter plots this type of trend is present as the difference in the variances of the two sub-groups at the two extremes of the separation vector **h** (see Figure 3-7). Notice the differences in the variances of the two sub-datasets shrink the cloud of data pairs in the horizontal direction because the sub-group at **u** is the one with smaller local variance (see Figure 3-3 bottom right). If the trend variance goes in the opposite direction, the cloud of data pairs will tend to shrink in the opposite direction.

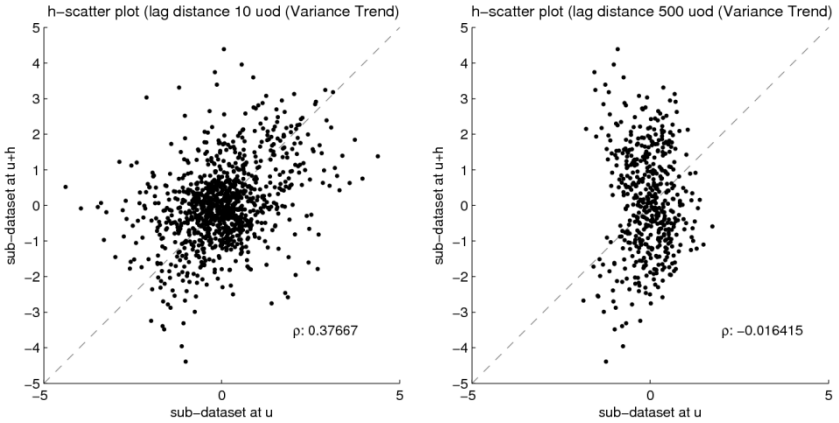

Figure 3-7: h-scatter plots for the parabolic variance trend case for lag distances 10 uod (left) and 500 uod (right)

Case Study 2

The second case consists of a topographic surface where elevations have been transformed to normal score units (see Figure 3-8). In the map, a trend in the mean can be observed in the north direction; there is a low level region in the south west part of the map, thus elevation increases from south to north.

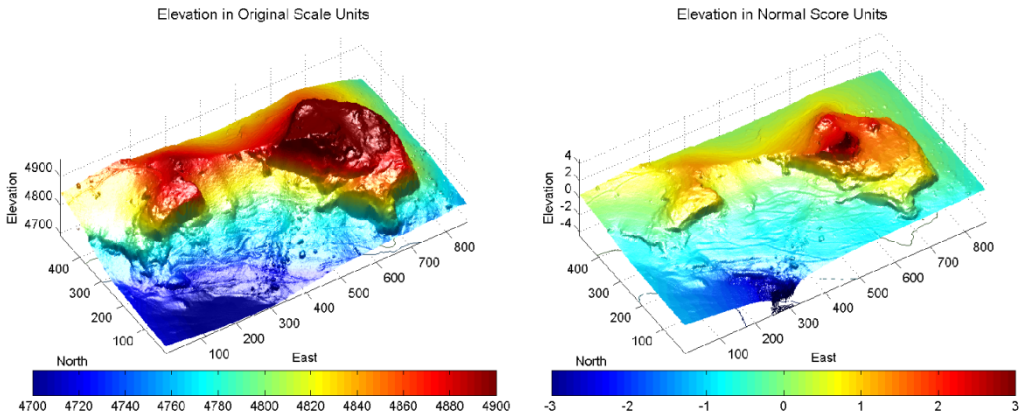

Figure 3-8: Elevations of a topographic map in original units (left) and normal score scale (right)

Two experimental semivariograms are calculated both in the north and west direction (see Figure 3-9). The mean trend in the north direction makes the experimental semivariogram increase above the sill (see Figure 3-9 left) and the increment in the global variance because of the mean trend makes the experimental semivariogram in the west direction not to reach the sill (see Figure 3-9 right). As a consequence, an apparent zonal anisotropy scenario is presented.

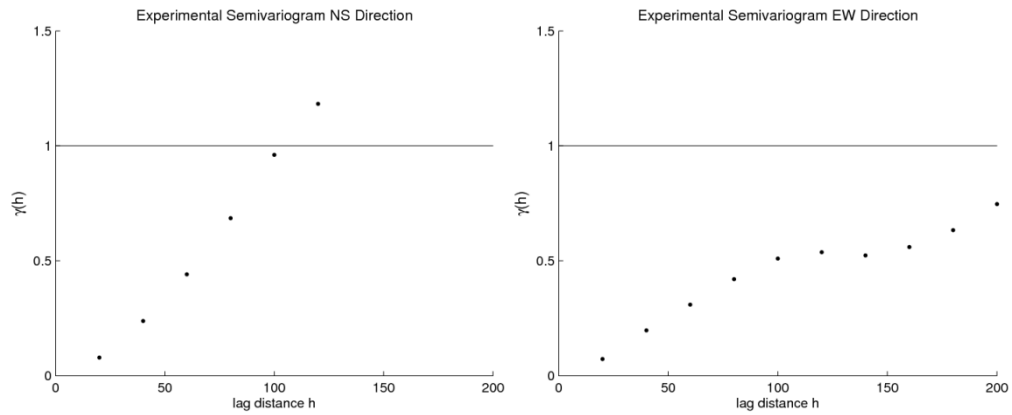

Figure 3-9: Experimental semivariograms for north-south direction (left) and east-west direction (right)

Decomposing the experimental semivariograms in terms of its trend components the mean trend is presented as a blue region in both directions (see Figure 3-10). Also a small variance trend is present in the north direction as a red region, however, this can be considered as negligible because when compared to the mean trend is very small.

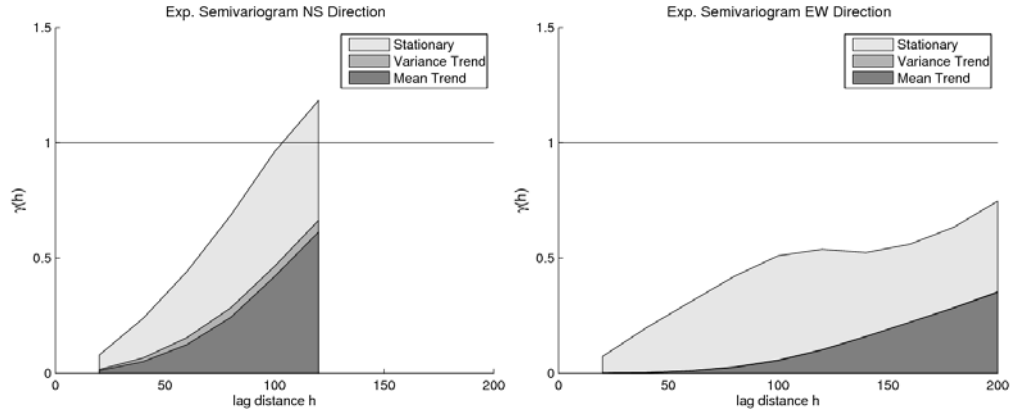

Figure 3-10: Decomposed experimental semivariogram from topography samples in normal score units for north-south (left) and east-west (right) directions

The presence of a mean trend in the topographic map makes it an unsuitable environment for doing conventional geostatistics. If the topographic map is assumed as stationary no reliable semivariogram model can be obtained from fitting the experimental semivariogram. Notice that any fitting of the two experimental semivariograms will result on a semivariogram model with zonal anisotropy while the stationary green region suggest that such anisotropy is not present in the topographic map after removing the trend.

Case Study 3

The second case study considers real data from an east Texas reservoir data set consisting of 62 vertical wells (see Figure 3-11) with 53 samples per well on average (minimum 35 and maximum 66 samples). There are a total of 3303 available data points in the dataset. In the vertical direction the separation distance between the samples is one unit of

distance. Only the vertical direction is considered here as sparsity of wells makes the horizontal direction a poor candidate for reliable variogram inference.

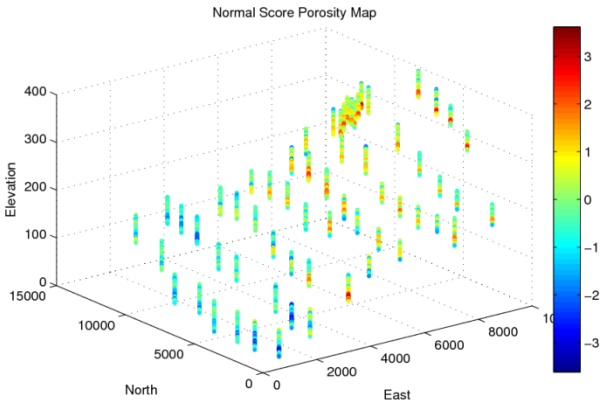

Figure 3-11: View of 62 Wells an East Texas reservoir

A vertical semivariogram plot of porosity, in normal score units, is calculated (see Figure 3-12). The lag interval used is one unit of distance. Traditionally it could be inferred that there is a presence of a mean trend because the experimental semivariogram plot goes above the sill of one beyond 40 distance units. However, the proposed approach reveals that a trend in the mean and variance contribute to the experimental semivariogram.

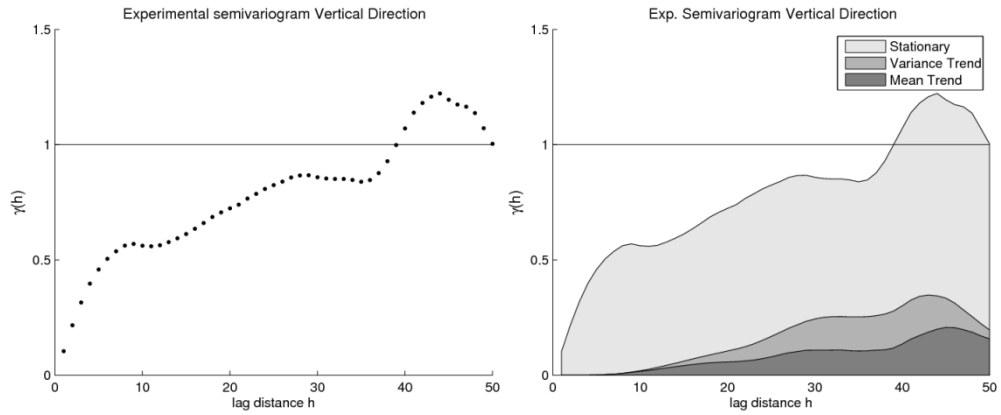

Figure 3-12: Experimental semivariogram (left) and with trend components (right) of vertical direction for porosity in normal score units

3.5. Discussion

It is important to keep in mind the semivariogram is a two point statistic and as such it is limited for capturing complex trend features in the IRF. As a consequence, for determining the influence of the local mean and/or variance on the semivariogram it requires representative sampling of the field. Therefore, in presence of large unsampled regions in the domain or blank spots, this approach would lead to a poor interpretation of the results.

Non-stationary features such as mean and variance trends of the dataset can be captured by analyzing the experimental semivariogram. Differences in the local means

and local variances are identified by comparing the means and variances of the sub-groups of datasets at the two extremes of the separation vector.

The proposed approach for calculating experimental semivariograms can be used to verify the influence of the mean and variance trends in the domain. Decisions of accepting or rejecting a certain domain prior to geostatistical modeling can be made by using the proposed tool.

So far the conditions of constant mean and constant variance of the RF have been addressed. In the next two chapters the so-called “weaker assumption” or intrinsic hypothesis is discussed. This condition is also verified in the dataset for the necessary stationary conditions for using conventional geostatistics when modeling a domain.

4. Experimental Semivariogram Cleaning

A semivariogram is used to characterize the spatial continuity of a variable of interest. In mining, this variable is assumed to be a realization of a SRF, since there is only one value for each sampled location (e.g. metal grade, contaminant concentration, oxidation ratio). In other applications, such as hydrology, many measurements can be taken from the same monitoring location. In order to transfer the spatial continuity information into a geostatistical model, the experimental semivariogram is fitted by a licit semivariogram model (see Chapter 2). However, semivariogram modeling is a subjective task; even when semi-automatic fitting algorithms are used there is still a certain degree of subjectivity. The true semivariogram of a SRF is modeled by comparing and combining a set of available licit semivariogram models. While this technique does not guarantee an exact representation of the semivariogram, it provides a good approximation. Section 4.1 discusses fitting an experimental semivariogram under stationary assumptions.

The stationary assumption is less valid where there is a dense sampling pattern over the domain. If the variable of interest is known at all the locations of the domain, the influence of the geologic structures is determinable. However, geologic processes are non-stationary. Therefore, even when the local means and variances are considered constant along the domain, the intrinsic stationary assumption becomes less valid. This is sufficient to make the domain non-stationary.

The semivariogram model does not, therefore, characterize the spatial continuity of a non-stationary domain in the same way as a stationary domain. The spatial continuity of the domain represented in the semivariogram model, which should be characteristic at all the locations, is now an averaged representation of it. Therefore, the spatial variability characterized by the semivariogram model is misrepresented locally. It is underestimated in some regions and overestimated in others.

Section 4.2 discusses the use of moment-of-inertia to calculate experimental semivariograms. The ideal conditions and characteristics of real data for SGS are discussed in Section 4.3. Section 4.4 describes the impact of non-stationary environments on the semivariogram and explains how to employ a semivariogram model under such conditions. Section 4.5 outlines an iterative approach for dealing with patterns in a domain. It involves removing the influence of outlier data pairs and recalculating the semivariogram model until it describes the spatial continuity of the majority of the dataset. A case study is provided in Section 4.6.

4.1. Fitting Experimental Semivariograms

This section describes the impact of experimental semivariograms on estimation/simulation. In order to avoid any external non-stationary influence, a stationary environment is assumed.

The conventional process of semivariogram modeling consists of fitting an experimental semivariogram using licit semivariogram models (see Chapter 2). In a stationary environment, the available dataset is assumed to be a realization of a SRF. Based on this assumption, the experimental semivariogram calculated from one realization is a fair representation of the spatial continuity of the SRF. The use of a set of licit models makes the semivariogram model an approximation and not necessarily an exact representation of the spatial continuity of the SRF. It is exact when the true semivariogram of the SRF is the same as the model used for fitting. Alternatively, it is still an approximation when the true semivariogram is not present as any combination of the set of licit models used for fitting (see Figure 4-8).

Even under optimal stationary conditions, the process of semivariogram modeling is subjective and dependent on the user's experience and knowledge. To limit the resultant variability in outcome, several semi-automatic algorithms have been developed for use during the modeling process (Larrondo & Neufeld, 2003), (Sinclair & Blackwell, 2004). Even then, the way a user inputs parameters may have a discernable effect on the outcome. Because of this inherent subjectivity, the resulting semivariogram model may be classified as follows (see Figure 4-1):

- *Pessimistic:* when the model is above the experimental semivariogram. As a result, the ensuing conditional variances at the estimated locations are larger than expected. The uncertainty assessed in the estimated domain is inflated; therefore, simulated maps are much more variable.
- *Fair*: when the model closely follows the experimental semivariogram. The estimated conditional variances at unsampled locations are a good approximation of the theoretical variances of the corresponding SRF.
- *Optimistic*: when the model is below the experimental semivariogram. In this case, the spatial continuity is exaggerated; data locations that should not have a significant influence are forced to contribute information. The conditional variances are smaller than they should be and the uncertainty associated with the model is underestimated.

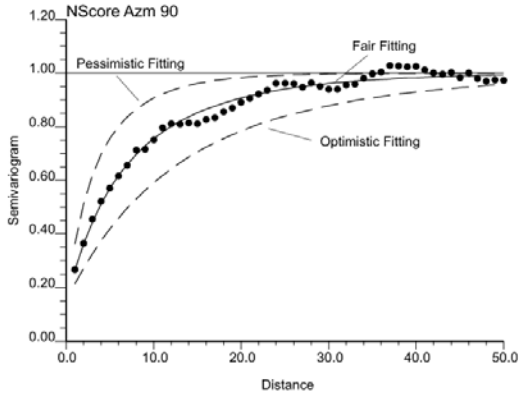

Figure 4-1: Experimental semivariogram (black dots), and three possible cases of semivariogram modeling (black dots)

With a simple configuration of experimental semivariogram points, the semivariogram model should be easy to fit. However, the experimental semivariogram could describe structures that would be difficult to fit using a common set of

semivariogram models (see Figure 4-8). In such a situation, the semivariogram modeling is used to fit a representative portion rather than the whole experimental semivariogram. By doing this, various regions of the experimental semivariogram become: optimistic, pessimistic and fair. The semivariogram model is expected to be globally representative. Since the simulated maps tend to reproduce the semivariogram models, such features of the modeling process are transferred to the simulated model.

4.2. Aspects on the Moment-of-Inertia Experimental Semivariogram Calculation

The experimental semivariogram points can be represented as **h**-scattergrams. In the **h**-scattergram the data values at the two extremes of the separation vector are presented in the form of a scatterplot. The experimental semivariogram can be interpreted as the moment of inertia of the **h**-scattergram about its first bisector. The moment of inertia is the average of the squared orthogonal distances (4.1) of the data pairs to the first bisector (4.2) (Goovaerts, 1997). Under the assumption of multi-gaussianity, the bivariate distribution of the data pairs of any **h**-scattergram is bivariate normal. A multi-gaussian RF is assumed for conventional implementation of SGS.

During the exploratory data analysis stage, the **h**-scattergrams can be used to verify particular features that could correspond to sub-populations. The decision to separate them into small domains is based on the availability of sample points that support such a decision (Goovaerts, 1997). According to (Genton, 1998) this practice is informal and cannot supersede a robust estimation technique of the experimental semivariogram. However, from an engineering perspective, identifying patterns in the dataset improves the knowledge of the natural phenomena under study. The semivariogram is a tool that works well under stationary conditions but not as satisfactorily in real-case scenarios.

$$\begin{aligned} d_i(\mathbf{h}) &= \frac{\sqrt{2}}{2}\left|z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h})\right| \\ &= \cos 45^\circ \left|z(\mathbf{u}_i) - z(\mathbf{u}_i+\mathbf{h})\right| \end{aligned} \tag{4.1}$$

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{N(\mathbf{h})}\sum_{i=1}^{N(\mathbf{h})} d_i^2(\mathbf{h}) \tag{4.2}$$

4.3. Real versus Ideal Environments

In conventional practice of SGS, the attribute of interest is transformed from original units to normal score units and assumed to be multi-gaussian. This transformation is referred to as a normal score transformation and, by construction, will transform any distribution to a univariate standard normal distribution (Deutsch and Journel, 1998). In practice, it is not a requirement for a dataset that is assumed to be sampled from a multi-gaussian process to follow exactly a univariate normal distribution.

Mineral deposits are the result of the interaction of many different previous geologic events over time. The dataset collected from it captures such geologic features, based on processes that obey physical and chemical laws. Ideally, the variable of interest should be modeled in such a manner that accounts for these laws, which in many cases are

extremely complex to define. The normal score transformation of a dataset sampled from a mineral deposit preserves the geologic features captured by the original scale data values (see Figure 4-2). The geologic structures are represented in the datasets as patterns that are analyzed and interpreted by earth scientists. This condition makes the assumption of multi-gaussianity of the domain inappropriate in the presence of dense sampling over the domain.

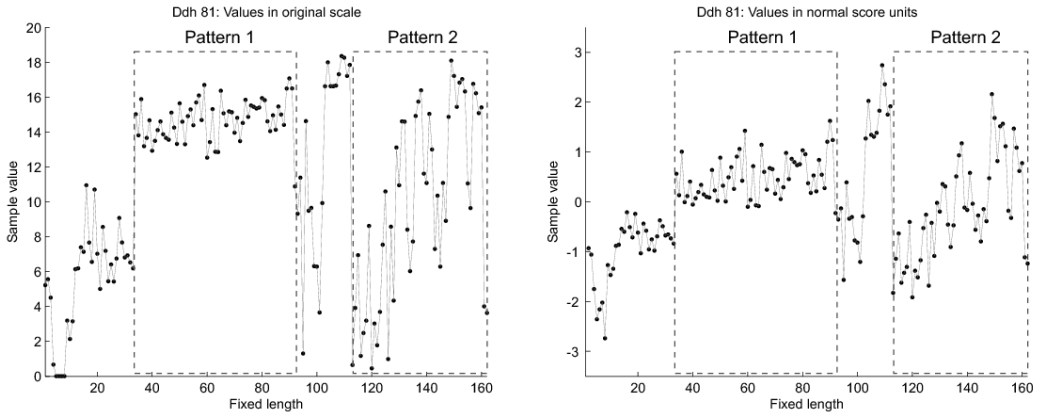

Figure 4-2: Ddh-81sample values for 162 regularly spaced locations in original units (left) and normal score scale (right), two patterns are highlighted in the dataset that are present both in original and normal score scale units.

Conventional geostatistics uses linear estimation to infer a value for variable of interest at an unsampled location based on the same variable or secondary variables. One necessary condition is the assumption of stationarity of the domain. Geologic features cannot be represented by a SRF. Discrepancies in the use of such models arise when estimated/simulated maps are compared against real information, such as geologic mapping of mined faces and mined production. Two sets of 162 data values from an unconditional realization are shown in Figure 4-3. When comparing these two datasets with the normal score values of the real case (Figure 4-2 right) it can be seen the local geologic patterns cannot be reproduced. The unconditional realization was simulated using a semivariogram model fitted of the real dataset.

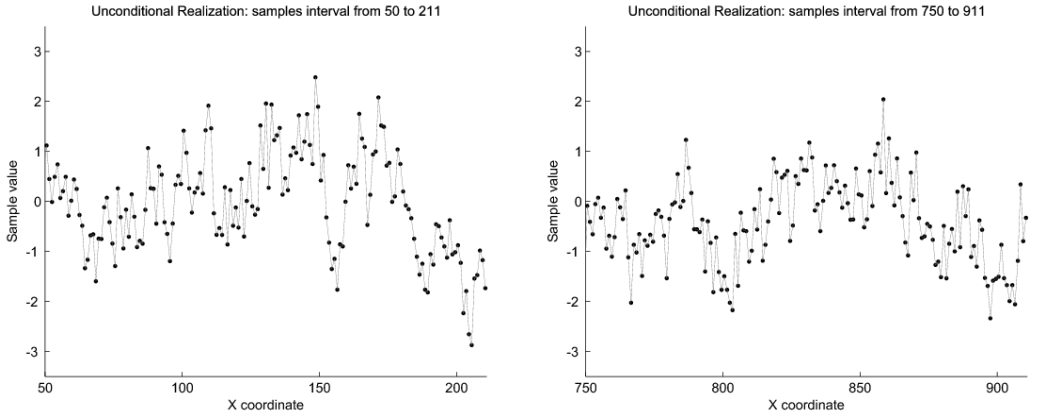

Figure 4-3: Two sub-datasets of 162 data points from an unconditional realization of 1000 data points

In a non-stationary environment (assuming the local mean and local variance are constant), the semivariogram is very sensitive to the structural geologic patterns. If a small portion of the domain presents a different variability than the rest of the domain the semivariogram captures it as a summary of all the structures (Journel & Huijbregts, 1978). Because of the stationary assumption, the information of spatial continuity is generalized to the entire domain.

An exercise is presented that mimics the interaction between two geologic processes in a domain. This is present not because domaining was poorly performed, but spatial structures are inherent to natural processes. Consider one dataset that is the combination of two unconditional realizations, A (80%) and B (20%) with different semivariograms (see Figure 4-4). The semivariogram model of dataset A is $\gamma_A(\mathbf{h}) = Exp_{30}(\mathbf{h})$ and of dataset B is $\gamma_B(\mathbf{h}) = 0.4 + 0.6Exp_{15}(\mathbf{h})$. The resulting combined dataset is non-stationary, despite the local mean and local variance are constant. The absence of the intrinsic hypothesis condition is what makes the dataset non-stationary. As a consequence, the experimental semivariogram is no longer representative of the dataset. Some features of the small region are transferred to the entire domain. The nugget effect of the small sub-dataset B is scaled and assumed present in the entire dataset. A better description of the spatial continuity could be obtained by: (1) separating the domain into two parts, that is, the sub-dataset A and B, (2) calculating and fitting an experimental semivariogram that accounts for the majority of the domain and correct the conditional distributions of the rest of the domain accordingly. The next section proposes an approach to identify minor structures that affect the experimental semivariogram. The cost of working directly in a SRF environment is the generalization of the spatial variability, thereby making it difficult to reproduce local features.

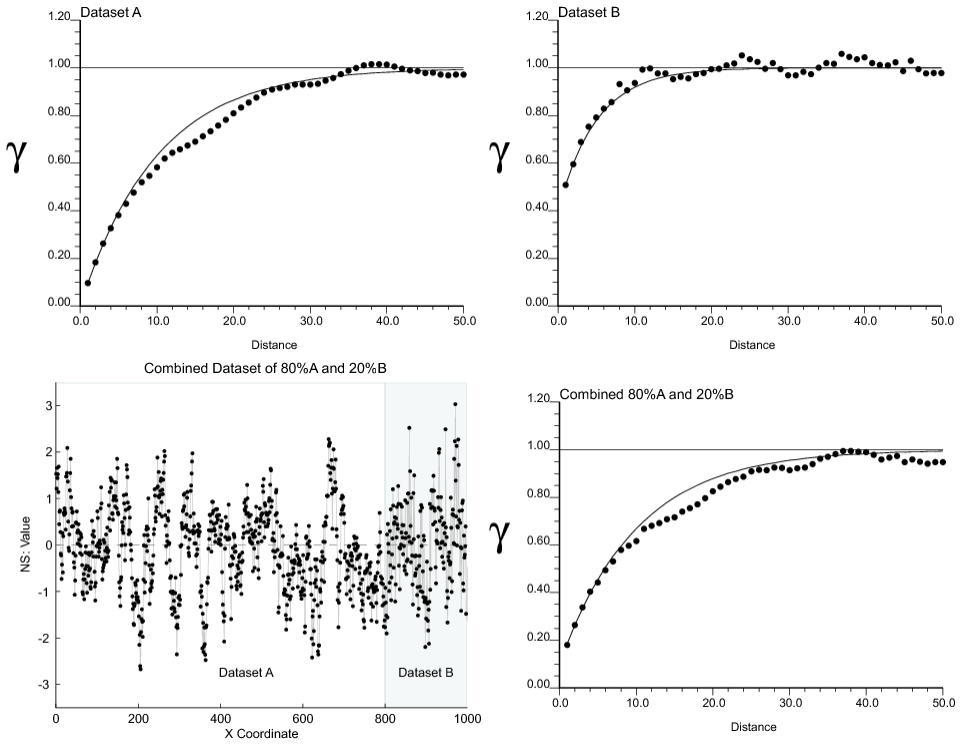

Figure 4-4: Experimental semivariogram of a unconditional realization using an exponential semivariogram model with range 30 uod and C_0=0 (top left), with range 15

uod and C_0=0.4 (top right), combined dataset of 80% of dataset A and 20% of dataset B (bottom left), and its experimental semivariogram (bottom right)

4.4. Experimental Semivariogram Outlier Data Pairs

Outliers are unusual observations that do not appear to belong to a nearby pattern of variability. However, not all the outlier values are wrong numbers; they can be considered as information that could lead to a better understanding of the phenomena being studied (Johnson & Wichern, 2007). The experimental semivariogram of a non-stationary domain presents anomalies when compared to the analytical distributions in a multi-gaussian environment. Such anomalies or outliers can be used to identify patterns in the domain that affect the non-stationary features of the experimental semivariogram.

This approach aims to identify outliers in the experimental semivariogram by defining outlier limits based on the proposed semivariogram model and a probability interval. The outliers are considered to belong to another type of spatial behaviour that affects the experimental semivariogram of the majority of the domain. Once identified, the decision to model them separately or to correct their variability can be made. Two approaches for identifying outliers are discussed in this section.

4.4.1. Control Limit Ellipses

In a multivariate gaussian framework, control limits are used to verify the stability of processes and identify occurrences of special cases of variation that are unlikely part of the process. They make the special variations visible and allow distinguishing them from the common ones in the process (Johnson & Wichern, 2007). Because of dimensionality, control limits are used in the form of charts for multivariate datasets. However, in a two dimensional case a control limit ellipse can be used. The control limit ellipse is defined as a contour of constant density of the bivariate standard normal distribution for a given confidence control value. It can be expressed using the generalized form of the ellipsoid of constant density of *p*-dimensions (4.3) (Johnson & Wichern, 2007).

$$\left(\mathbf{x}-\boldsymbol{\mu}\right)^T \mathbf{C}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right) = \chi_p^2\left(\alpha\right) \tag{4.3}$$

where $\mathbf{x}$ is a vector of observations on different variables, $\boldsymbol{\mu}$ is the vector of means of $\mathbf{x}$, $\mathbf{C}$ is the covariance matrix, p is the number of dimensions or elements in $\mathbf{x}$, and $\chi_p^2(\alpha)$ is the upper (100α)-th percentile of a chi-square distribution with p degrees of freedom. Expression (4.3) gives the contour for the multivariate case that represents $(1-\alpha)100\%$ of the probability.

The experimental semivariogram is analyzed on an **h**-scattergram basis. Analytically the distribution of data pairs on each **h**-scattergram is bivariate normal. Therefore, the control limits are set up to the two dimensional case. Making x_1 be the data values at location $\mathbf{u}$ and x_2 the data values at location $\mathbf{u}+\mathbf{h}$, and for simplicity, assuming both distributions are standard normal. Then the expression of the control limit ellipse is (4.4).

$$\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2} = \chi_2^2\left(\alpha\right) \tag{4.4}$$

However, an analytical model is required that provides the parameters of such bivariate distributions. That information is supplied by the semivariogram model fitted to the experimental semivariogram. The control limit ellipses are used to identify outlier data pair(s) according to the fitted semivariogram model on an **h**-scattergram basis. For a probability of $(1-\alpha)100\%$, the proportion of data pairs that are expected to fall outside each control limit ellipse of the h-scattergrams is $(\alpha)100\%$. In the same way, the proportion of the outlier data pairs for the entire semivariogram model is also $(\alpha)100\%$.

For the case presented, the geometry of the control limit ellipse is a function of the correlation coefficient and the probability confidence limit. The control limit ellipse is rotated so that the major axis is in the direction of the first bisector of the h-scattergram. The length of the major and minor ratios is (4.5), (4.6) (Johnson & Wichern, 2007).

$$rt_{mj} = \chi_2^2(\alpha)\sqrt{1+\rho^2} \tag{4.5}$$

$$rt_{mn} = \chi_2^2(\alpha)\sqrt{1-\rho^2} \tag{4.6}$$

Notice that for the same probability interval, the length of the major ratio is proportional to the correlation coefficient (see Figure 4-5). This is a problem for analyzing the spatial continuity, since data pairs placed close to the first bisector that are considered as good observations for high correlation coefficients would become outliers for small correlation coefficients. It is contradictory to the definition of the semivariogram which is a measure of dissimilarity. The smaller the correlation coefficient, the more tolerant the semivariogram to outlier values is expected to be. In Figure 4-5, consider the data pair P with $x_1 = x_2 = 2.2$. Notice that the data pair P lies within the acceptable region of three out of four control limit ellipses shown in the example. Even when the values of the data pair P are equal some control limit ellipses will identify it as an outlier. An alternative approach is required that identifies outlier limits based on the increments of the data pair values $z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})$ rather than its position on the **h**-scattergram.

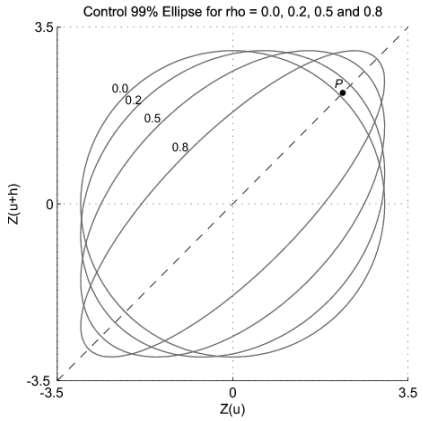

Figure 4-5: Control limit ellipses for 99% of bivariate standard normal distribution for four different values of correlation coefficient; the data point P is placed over the first bisector and is evaluated in the four cases.

4.4.2. Confidence Limits of the Distribution of Experimental Semivariogram Data Pairs

Under the assumption of multi-gaussianity the data pair increments, $z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})$, follow a univariate normal distribution and the square of the increments $[z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})]^2$ a scaled chi-square distribution with one degree of freedom $2\gamma(\mathbf{h})\chi_1^2$ (Cressie & Hawkins, 1980). The previous statement can be re-expressed as $[z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})]^2/2 \sim \gamma(\mathbf{h})\chi_1^2$ or in terms of the orthogonal distance of the data pair to the first bisector $d^2(\mathbf{h}) \sim \gamma(\mathbf{h})\chi_1^2$ (see Figure 4-6). Recall the mean of a chi-square distribution is the number of degrees of freedom, in this case 1, so the expected value of the orthogonal distances is the semivariogram $E\{d^2(\mathbf{h})\} = \gamma(\mathbf{h})$. Similar to the control limit ellipses approach, extreme values of $d^2(\mathbf{h})$ can be discriminated using confidence limits. This way the increments of the data pairs are accounted for directly.

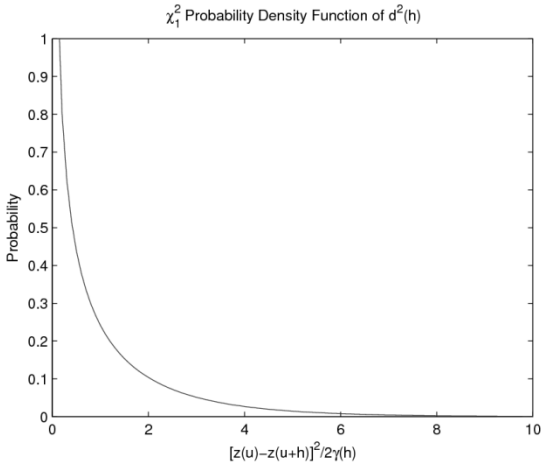

Figure 4-6: χ_1^2 Distribution of the orthogonal distances from data pairs to the first bisector for a separation vector h.

4.5. Experimental Semivariogram Calculation with Confidence Limits

The experimental semivariogram is analyzed in the form of a cloud semivariogram using confidence limits. The semivariogram model defines the ideal conditions of how the conditioning data should be. By comparing the ideal conditions to the dataset that is sampled from the real domain, the patterns that are produced by the geologic features can be identified.

Each data pair consists of two data values separated by a vector. Therefore, there is a reference value of semivariogram model. Then, each data pair can be evaluated independently by its corresponding outlier limit regardless of the tolerances used to calculate the experimental semivariogram. The outlier limit is the maximum squared orthogonal distance for a confidence limit (4.7), where the probability of the chi-square distribution α defines the $(1-\alpha)100\%$ probability of occurrence of the data pairs:

$$d_{MAX}^2(\mathbf{h}) = \chi_1^2(\alpha)\gamma(\mathbf{h}) \tag{4.7}$$

The data pairs that are outside of the limit, that is $d_i^2(\mathbf{h}) > d_{MAX}^2(\mathbf{h})$, are marked as outliers for the given parameters of confidence limits α and semivariogram model $\gamma(\mathbf{h})$. Even when a data pair is marked as an outlier, it does not imply that the corresponding head and tail values are outliers too. Particular patterns of data pairs in the cloud semivariogram provide evidence of the geologic structures present in the available dataset due to the nature of the domain. This makes it possible to assess whether additional sub-domaining is required or find a way to reproduce these local patterns in the geostatistical model. For instance, an abrupt change in metal grades at a short distance could be an indicative of the presence of veins or some other type of small structures in the domain. In mining such structures are of special importance because they can define whether some regions are of economic interest or not. Therefore, they impact directly the economic potential of the mineral deposit.

Once the outlier data pairs are identified, the experimental semivariograms with and without the outlier data pairs should be compared to verify their impact in the semivariogram analysis. If there is a notorious impact in the experimental semivariogram after removing the outlier data pairs, it is recommended to be fitted by a new semivariogram model; otherwise, that would mean the initial proposed semivariogram model is fairly representative of the dataset.

The continuous form of the outlier limits can easily be shown for a 1D dataset in a cloud semivariogram plot. Recall that each data pair is plotted as $0.5[Z(\mathbf{u}_i) - Z(\mathbf{u}_i + \mathbf{h})]^2$, that is $d_i^2(\mathbf{h})$. Since $d_{MAX}^2(\mathbf{h})$ is a continuous function it can be plotted with respect to the semivariogram model for the different confidence limits and for any lag distance (see Figure 4-7). For a multi-gaussian dataset the density of data pairs outside each control limit is uniformly distributed. However, for real dataset particular patterns or clusters of data pairs are present.

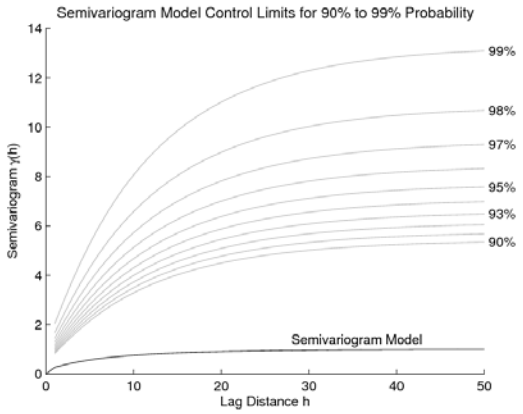

Figure 4-7: Control limits for different probabilities, from 90% to 99% (gray line) with respective semivariogram model (solid line)

The style of semivariogram fitting (see Section 4.1) has an impact on the process of cleaning the experimental semivariogram. In practice, the style of fitting a semivariogram model is subjective. The style of fitting is not applied to the entire experimental semivariogram, but to segments of it. Overall, the experimental semivariogram is recommended to be fitted to the entire experimental semivariogram as close as possible. The consequences of the styles of fitting are:

- *Pessimistic*: The correlation coefficients of the **h**-scattergram are larger than expected, so outliers are more difficult to identify. The theoretical variability is larger than the experimental semivariogram. The limit $d_{MAX}(\mathbf{h})$ tends to be larger.
- *Fair*: The correlation coefficients of the **h**-scattergram try to reproduce the variability of the dataset. This is a good condition to identify outlier data pairs, since it directly compares the analytical to the experimental form of the spatial variability. The limit $d_{MAX}(\mathbf{h})$ is representative.
- *Optimistic:* The correlation coefficients of the **h**-scattergram are smaller. As a consequence, many data pairs are identified as outliers. The limit $d_{MAX}(\mathbf{h})$ tends to be smaller.

4.6. Case Study

Consider the real case presented in the first part of this chapter. The dataset corresponds to a borehole of 162 samples regularly spaced in normal score units (see Figure 4-2 right). Therefore, there is no influence of tolerances or clusters in the experimental semivariogram. The only source of influence comes from the structural patterns in the dataset. The semivariogram model used for the exercise is given in (4.8) and shown in Figure 4-8.

$$\gamma(\mathbf{h}) = 0.15 + 0.35Exp_{12.5}(\mathbf{h}) + 0.50Exp_{35.0}(\mathbf{h}) \tag{4.8}$$

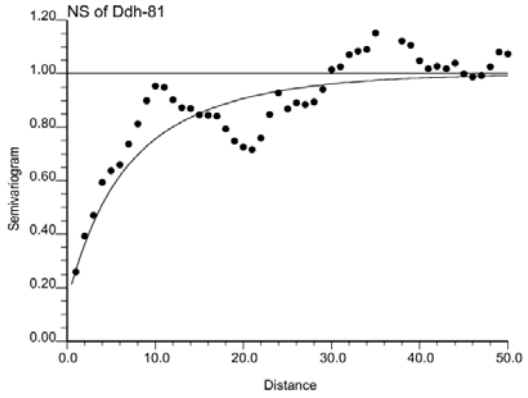

Figure 4-8: Experimental semivariogram (black dots) and semivariogram model (solid line) of Ddh-81 dataset

The selection of the probability limit is based on identifying the smallest number of data pairs as possible that significantly impact the experimental semivariogram. Several control limits were applied to the experimental semivariogram of the dataset Ddh-81 at different ranges in order to identify special patterns in the data pairs (see Figure 4-9). Figure 4-9 (A) considers a range of probabilities from 90% to 99%; based on this range, two patterns can be seen: the first one from lag distances 0 to 18 and the second one from 35 to 38. The range of outliers is reduced to an interval from 97% to 99.5% in Figure 4-9 (B). The first part from the lag distance 0 to 20 becomes more pronounced while the second part recedes. Reducing the range of outliers even more (Figure 4-9 (C)) from 99.5% to 99.9%, the number of data pairs in the first part is still dominant, while the second part virtually disappears. Finally, the limit that identifies the outliers of the first

part is chosen (Figure 4-9 (D)) that corresponds to a 99.7% probability. This removes 0.6% of data pairs (or one data pair) of the experimental semivariogram which is not a significant proportion of the information provided by the dataset.

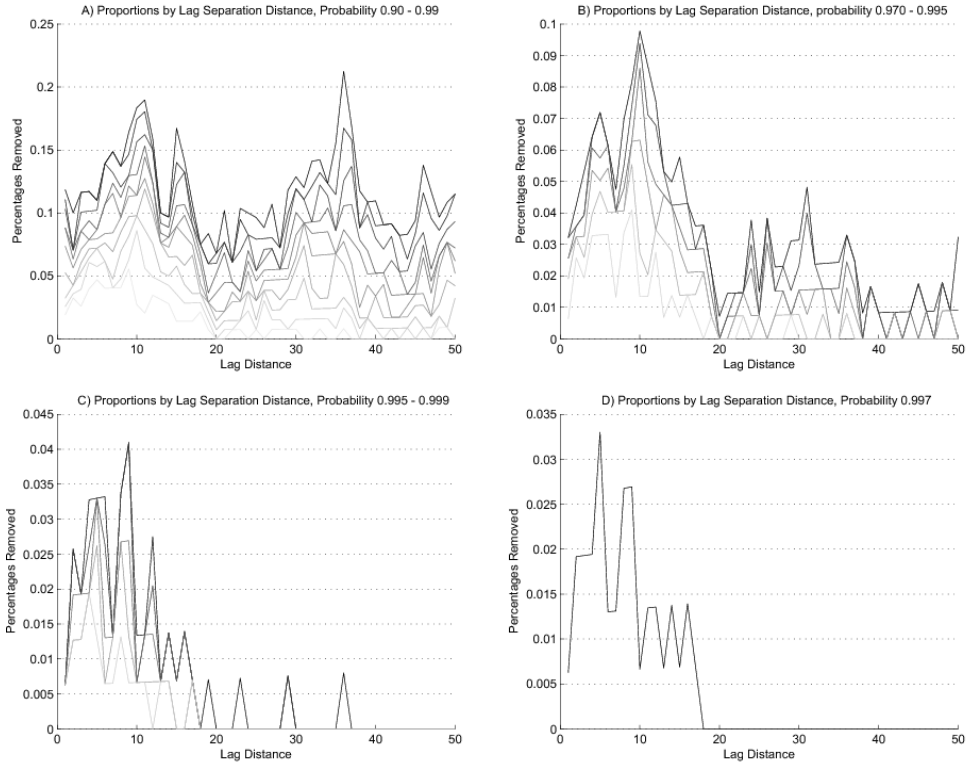

Figure 4-9: Proportions of data pairs identified as outliers for different ranges of control limits for the experimental semivariogram of the dataset Ddh-81 in normal score units.

The data pairs removed from the cloud semivariogram clearly makes a particular pattern (see Figure 4-10 top left). Including the influence of this sub-region in the semivariogram model causes an inflation of the conditional variance to the rest of the domain. On the other hand ignoring such variability means the conditional variances of the small sub-region are unaccounted. Depending on the spatial configuration and size of the sub-region two decisions can be made: (1) split the domain so that two domains can better account for different patterns of variability, or (2) focus attention on the major part of the domain and ignore the variability of the sub-region and correct the pattern of variability locally.

Notice that after removing the 0.6% of the outlier data pairs there are significant changes in the experimental semivariogram. The fixed experimental semivariogram is more continuous with fewer jumps (see Figure 4-10 bottom). Therefore, the fitting of the experimental semivariogram tends to be more straightforward (4.9). This can be seen in the reduction of the mean squared error (MSE) (see Figure 4-10 top right). This, of course, does not change the subjectivity of the fitting process; it merely reduces the trade-offs that the geomodeler must consider while fitting a semivariogram model.

$$\gamma(\mathbf{h}) = 0.12 + 0.37Exp_{20}(\mathbf{h}) + 0.51Exp_{38}(\mathbf{h}) \tag{4.9}$$

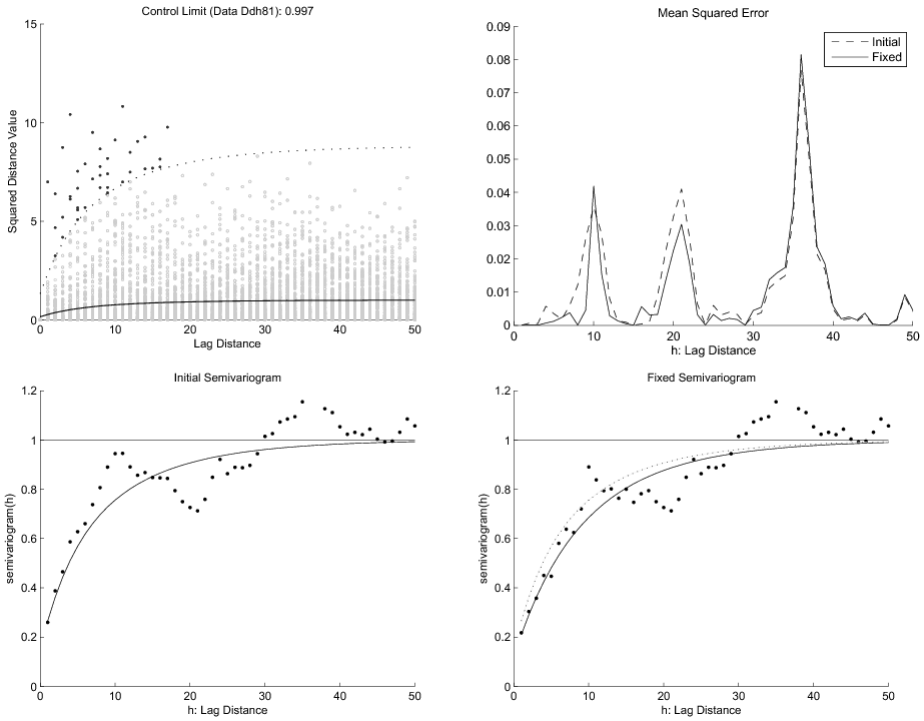

Figure 4-10: Cloud semivariogram with control limit at 99.7% (top left), semivariogram model fitting MSE for initial (dashed line) and fixed (solid line) experimental semivariograms (top right), initial experimental semivariogram (black dots) and semivariogram model (solid line) (bottom left) and fixed experimental semivariogram (black dots) and its respective semivariogram model (solid line), and initial semivariogram model (dotted line) (bottom right)

The second part of the analysis consists of verifying whether the outlier data points are grouped in certain specific parts or are spread around the entire domain. If they are placed over particular regions or follow clear patterns, then this will increase our knowledge of the nature of the domain. A decision of sub-domaining can be considered here. On the other hand, if the data points are dispersed throughout the domain, then this could mean the identified variability is part of the domain and it can be modeled conventionally. For the data set Ddh-81, the data points that produce extra variability are located in a specific sub-region. They can be considered for sub-domaining in order to improve the estimation/simulation of the domain (see Figure 4-11).

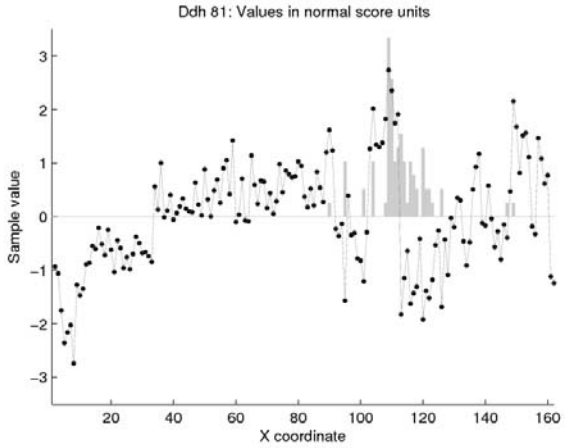

Figure 4-11: Occurrences of data points of outlier data pairs (gray bars) compared with the input dataset Ddh-81 (black dots) in normal score units.

4.7. Discussion

The proposed methodology identifies anomalies in the outlier increments in the domain. However, one of the inputs is the semivariogram model. This makes this methodology to be subjective in the interpretation of the results. Moreover, the anomalies in the distribution of the data pairs has to be interpreted by the user as well as additional semivariogram fitting have to be carried out. Notwithstanding, in a sensitivity analysis this approach gives information about spatial configuration of the samples that produce the outlier increments. Also, the resulting semivariogram model tends to be more continuous due to the influence of the outlier data pairs are removed, this is called cleaned-semivariogram.

In the next chapter the cleaned-semivariogram model is used for analyzing the domain and evaluating the high variable sub-regions. Because of the presence of those sub-regions, the domain is moved to a high dimensional space where spatial variability can be fully defined by the semivariogram model.

5. Conditional Distribution Fitting

In Chapter 4 the semivariogram is used as a tool for identifying the non-stationary features in the domain related to the intrinsic assumption of the RF. Two scenarios remain from the semivariogram analysis: 1) the problematic locations are grouped in sub-regions, so that they can be considered for sub-domaining and 2) the problematic locations are in small groups dispersed throughout the domain such that sub-domaining cannot be performed. This chapter focuses on the latter scenario and proposes an alternative method to account for the influence of these data in the domain so that the local uncertainty calculated using estimation/simulation can be used for a detailed mine plan strategy.

Sequential Gaussian simulation (SGS) is recommended to be used for simulating the variable of interest in a domain due to its simplicity in the implementation and availability in many commercial mining software packages (e.g., Pangeos, Vulcan, MineSight, and Datamine, among others). The implementation of SGS is very similar to simple kriging; both use the same parameters such as the semivariogram model and search specifications. For implementation it is necessary to assume a multi-gaussian environment for SGS since all simulation is performed in Gaussian space. Although SGS generates multiple realizations of the attribute, it can be verified against an SK model since the local average from SGS, taken over many realizations, will tend to SK estimates. The problem with SGS and SK is that it assumes that the conditioning data is part of a SRF, that is, it assumes there are no outlier increments or the semivariogram model adequately defines both local and global uncertainty. For this reason and in these highly variable regions, SGS does not account for the local features of the domain and is not used for simulating medium/short term models. The goal of this chapter is to provide a methodology to prepare the dataset for performing SGS, so that it reproduces the non-stationary features required for mine planning.

In general, only high values are considered problematic because a more pessimistic estimation is preferred over an optimistic one. In some cases, outliers are the result of an erroneous characterization of the geology (e.g., presence of samples that belong to different geologic processes other than their tagged category). In Figure 5-1 a sketch of reality is compared to a tentative geologic interpretation; due to the scale of the interpretation some samples of the high metal grade region are classified as part of the low grade metal domain. Estimating such domains using kriging in a strict manner would smear the high grade value over a considerable region in the low grade domain. As a consequence, the geostatistical model of the low grade domain is neither realistic nor appropriate for proposing a detailed mine plan. In a more general context, the presence of sub-patterns in the domain makes the available dataset behave as a SRF with a spatial continuity cannot be adequately defined by one semivariogram model.

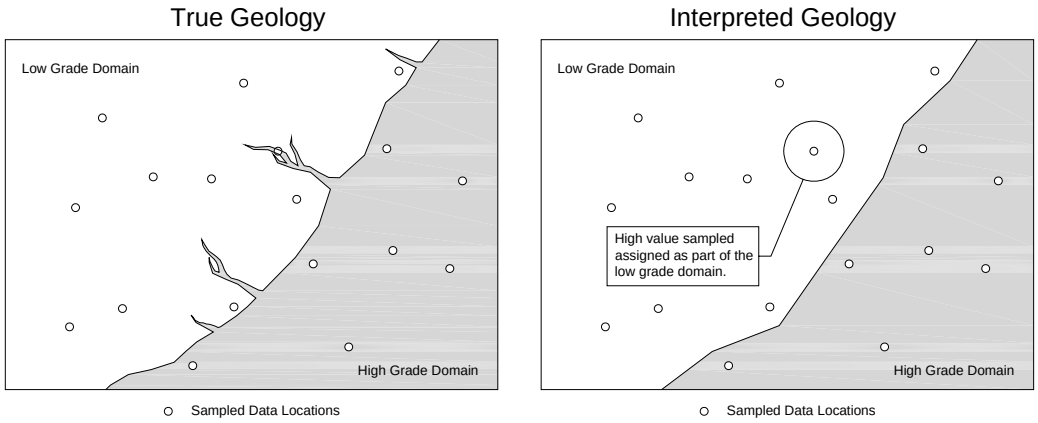

Figure 5-1: Impact of generalization of geology due to the scale of geologic interpretation; reality (left) is not fully characterized when models are built (right).

In this chapter the influences of trends in the mean and in the variance are considered to be absent. However, non-stationarity in the domain occurs when the intrinsic hypothesis of the SRF is not satisfied. Because of this, the estimated model appears locally unrealistic. As shown in chapter 4, internal structures of the domain are present as patterns in the dataset which are not necessarily removed by normal score transformation. The presence of such patterns in the domain makes the decision of stationarity less appropriate.

Proper estimation of the conditional distribution is important for simulation. The farther the true value falls from the confidence limits, the more difficult it is for the simulation to pick a realization that reproduces the true value at such a location and globally reproduces the sub-patterns in the domain. Ideally, a mine plan is based on the analysis of the geologic region to be mined; the impact of the geologic characteristics in the variable of interest is an important input in the process of decision making in mine planning.

Depending on the semivariogram model fitted, the degree of accuracy of the conditional distributions changes (see Chapter 4). A good semivariogram model is one that accounts for the spatial variability of the major part of the domain. The remaining spatial variability that is not properly accounted for by the semivariogram model is under-estimated in some regions and over-estimated in others. In real situations, it is unrealistic to expect that any semivariogram can account for the geologic features of a domain like it may under theoretical conditions. The approach presented in this chapter aims to account for the variability of this minority of the domain, where the spatial variability is under-estimated by tuning the distances between the samples, so that the conditional distributions are consistent with the conditioning information. The distances are modified by adding an extra dimension to the dimensional space of the conditioning dataset. In this way, the influence of the semivariogram model on the estimated parameters of the conditional distributions is approximated until the parameters are consistent with the dataset within some confidence limits. The goodness of the conditional distributions is verified using cross validation and confidence limit parameters.

5.1. Measure of Accuracy

Cross validation is used as a technique to test the quality of the estimated parameters of the conditional distribution with respect to the true values (Chilés & Delfiner, 1999). In this chapter, it is assumed that if the verified conditional distributions using cross validation properly account for their corresponding true values the conditional distributions of the rest of the unsampled locations will also do the same.

For all the data locations, confidence limits are considered as a measure of whether the true value with respect to its estimated conditional distribution is predicted within some tolerance intervals or not (see Figure 5-2). If the true value falls outside the confidence limits, then the proposed semivariogram model is considered to be inadequate and this location is flagged for pre-processing. All the conditional distributions where true values are within the confidence limits are assumed to be consistent with the surrounding information. Theoretically, if 90% probability interval is chosen then 10% of the dataset is expected to fall outside the confidence limits. For this, one basic condition is the dataset is truly part of a SRF realization, recall that a real dataset is non-stationary. Contrary to the theoretical conditions, the proposed approach forces all the data point values to fall within the confidence interval in order to ensure the realizations reproduce the sub-patterns in the domain.

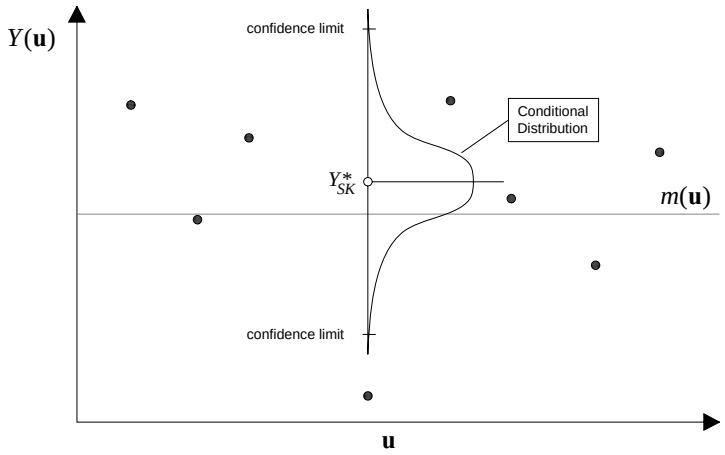

Figure 5-2: Sketch of cross validation where the true value (black dot) falls outside of the confidence limits of the conditional distribution (gray lines) calculated using the rest of the information and the proposed semivariogram model.

The accuracy of the estimates is measured by standardizing to one the distance from the estimated mean to any of the confidence limit values, therefore, if the true value falls outside the confidence limits the standardized distance is greater than one, and the conditional distribution is considered as improperly accounting for the data. Assuming the global univariate distribution of the domain is standard normal, then for the data values that are outside the confidence limits of the standard normal distribution, the standardized accuracy is re-scaled making the true values be the new confidence limits. This is done to ensure that for all values in the dataset there is a conditional distribution that includes the true values within reasonable confidence limits.

5.2. Dimensional Conditional Distribution Fitting

The semivariogram model $\gamma(\mathbf{h})$ is a function of the separation vector $\mathbf{h}$ that fully defines the spatial continuity of a SRF. The kriging estimated parameters of the conditional distribution at the unsampled location are functions of this semivariogram model. The semivariogram model is used to calculate the linear dependence between data values and the unsampled location in the kriging system. The smaller the distance from the conditioning data to the unsampled location, the smaller the conditional variance. Regardless of the values of the conditioning data the surrounding spatial configuration of the unsampled location is what really matters for calculating the conditional variance. The estimated mean tends to be similar in value to the closer surrounding data values.

Since the spatial covariance $C(\mathbf{h})$ is a function of the separation vector $\mathbf{h}$ the estimated parameters of the conditional distribution can be manipulated by modifying the separation distances from the unsampled location to the conditioning data. The influence of the conditioning data for estimating the parameters of the conditional distribution at the unsampled location decreases proportionally as the unsampled location is separated further from the data. The unsampled location becomes gradually more uncertain until the parameters of the conditional distribution are equal to the global distribution parameters, which occur when the influence of the conditioning data is negligible. When there are no samples within the effective range of the semivariogram model the estimated mean will equal the global mean $Y_{SK}^{*} \cong m$ and the estimated variance will equal global variance $\sigma_{SK}^{2} \cong \sigma^{2}$ this is the state of local maximum uncertainty.

A small example is shown in Figure 5-3 to illustrate the sensitivity of the conditional distribution with respect to the position of the unsampled location. Consider a four samples dataset $\sim N(0,1)$ and a location to estimate in 1D (see Figure 5-3-bottom). When the unsampled location is separated gradually from its original position by adding a new dimension the conditional distribution changes as the influence of the conditioning data diminishes (see Figure 5-3-top right). The change is gradual until the conditional distribution equals to the global distribution of the dataset, in this case $\sim N(0,1)$. For the example presented a nested exponential model is used (see Figure 5-3-top left, equation (5.1)). In Figure 5-3-top right notice the SK mean reaches 0 for the effective semivariogram range and the SK variance to the sill value of 1.0 for the effective range. Both parameters approach the global values asymptotically because an exponential model is used.

$$\gamma(\mathbf{h}) = 0.25Exp_{1.5}(\mathbf{h}) + 0.75Exp_{3.0}(\mathbf{h}) \tag{5.1}$$

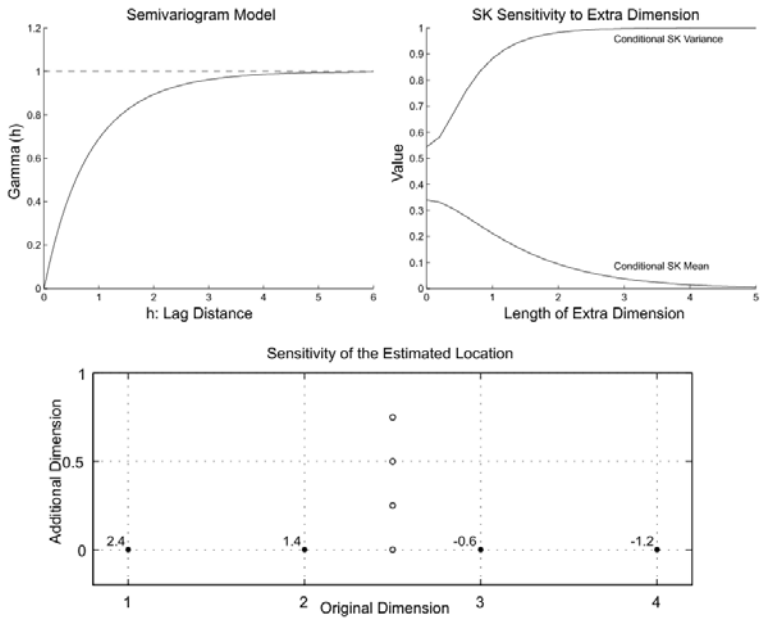

Figure 5-3: Nested exponential semivariogram model (top left), sensitivity of SK variance and SK mean to the inclusion of additional dimension to the original position of the unsampled location (top right) and a sketch of spatial configuration of conditioning data (black dots) and locations of the unsampled location (empty dots) to different lengths of the additional dimension (bottom).

The condition that the separation distances $d_{u-u'}$ between the unsampled location $\mathbf{u}' = [x'_1, x'_2, \dots, x'_n, y'_k]$ and the conditioning data $\mathbf{u}_j = [x_{1j}, x_{2j}, \dots, x_{nj}, 0]$ do not decrease is guaranteed by the additional dimension $y'_k \; \forall \; y'_k \neq 0$. This additional Cartesian component y'_k at the unsampled location is always equal or greater than zero, so it is additive when the separation distance is calculated, $d'_{u_j-u'} = \left[\sum_{i=1}^{n}\left(x_{ij} - x'_i\right)^2 + (y'_k)^2\right]^{1/2}$ therefore $d'_{uj-u'} \geq d_{u-u'}$.

The SK variance is said that it cannot be used as a measure of local variability or accuracy since it is based on a semivariogram model which is a global approximation of the spatial continuity (Journel A. , 1986) or because SK variances are independent of the data values and only provide a comparison of an alternative data configuration (Deutsch and Journel, 1998). The uncertainty assessment of SK fully relies on the assumption that the conditioning data is a realization of a SRF, with two requisites: (1) the distribution of errors is gaussian, and (2) the variance of errors can be predicted (Isaaks & Srivastava, 1989). Let us consider 1000 data points of an unconditional simulation where each cross validation conditional distribution is evaluated using confidence limits with respect to the corresponding true values for different probability confidence intervals. The proportions of samples within the confidence limits are similar to the theoretical expected proportions since the dataset is a realization of a multi-gaussian SRF (see Figure 5-4). Under correct conditions the SK conditional distributions account for local uncertainty properly. These two conditions are (1) the semivariogram is known and (2) the conditional error distribution is calculated with no additional influence of any source of error, (Chilés & Delfiner, 1999). This is not the case of geologic processes where there is no true

semivariogram. In fact, in practice the semivariogram fitting is based on the experience of the person in charge and on the objective of the study (Goovaerts, 1997).

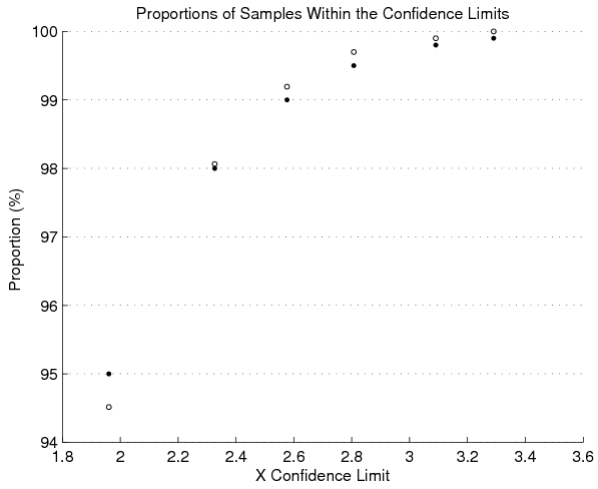

Figure 5-4: Proportions of true values within the confidence intervals (empty dots) compared to the theoretical proportions (black dots) of their respective cross validation conditional distributions. The true values are from an unconditional realization.

Under the assumption of multi-gaussianity the conditional distributions estimated by SK are univariate Gaussian $\sim N(Y_{SK}^{*}, \sigma_{SK}^{2})$. All values are plausible outcomes to occur depending on their probability, even when they are very extreme values. Let us consider a value at a certain location has a probability of 1×10^{-1000} to occur in a conditional distribution calculated using cross validation. Statistically the true value is a valid outcome of the conditional distribution. For a mineral deposit model used for economic decision, such an estimate could have serious consequences if it is considered as a valid result. For the small case presented the true value is smaller than the SK mean and when a realization is drawn at such location due to its very small probability the true value is very unlikely to be simulated. If the true value is unknown there is no way to verify the validity of the estimated conditional distribution. However, problems arise when the true value is known and its cross-validation conditional distribution does not account for it properly. Then when the surrounding locations are estimated, even though the SK means tend to average the big difference in values of the conditioning data, the conditional variances still remain smaller because the semivariogram model does not account for the true value properly.

The proposed approach tunes the conditional distributions to the conditioning data within some confidence limits under the assumption the conditioning dataset is representative of the domain. Using cross validation, once the conditional distributions account properly for their respective true values the resulting models is assumed to account for local and global uncertainty properly. Since the local conditional distributions are tuned to the dataset using additional dimensions the new spatial configuration of the conditioning information is called *proper stationary state* of the dataset. Some geologic features present in the dataset that were not captured by the semivariogram model initially are now explained by the additional dimensions. The additional geological information is added to the dataset in the form of position vectors.

5.3. Cost and Benefit of the Conditional Distribution Fitting

As mentioned before, the main goal of this chapter is to make the conditional distributions account properly for the local uncertainty of the realizations. Modeling conditional distributions is considered an ambitious goal and sometimes unrealistic to achieve without a specified theoretical model (Chilés & Delfiner, 1999). By adding extra dimensions there are many possible solutions to this problem. Many different configurations could give different degrees of fitting of the conditional distributions with respect to their true values. There would be some negative impact on the accuracy of the surrounding data locations of the fixed data location. This problem can be solved considering additional restrictions in the algorithm such as reducing the negative impact on the accuracy of the rest of the samples, reducing the negative impact on the SK means or by trying to use a small number of additional dimensions, etc.

Using small probability intervals could be very restrictive for this approach. The smaller the probability interval the more extra dimensions are necessary to fit the conditional distributions. By making all the data values fall within the confidence limits the sense of probability loses meaning because it does not comply with the theoretical conditions. Theoretically for a 95% confidence interval 5% of the true values is expected to fall outside their respective conditional distributions (see Figure 5-4), while the approach tries to eliminate such proportion of data values. Setting up the confidence interval is a subjective part of this approach. In mining a 95% probability is commonly used (Journel & Huijbregts, 1978). A real example of a Chilean copper mine (Chuquicamata) is presented in (Journel & Huijbregts, 1978) where 96% of the observed errors of mean block grades fall within the 95% interval. On other types of deposits such as skarn type where the grade variability is high and more geologic structures are present such result would be very difficult to obtain because of the complexity of the geologic environment. When the distribution of errors is non-gaussian but continuous and unimodal a confidence interval of $\pm 3\sigma_{SK}^2$ which correspond to 99.73% probability interval is preferable, see discussion in (Chilés & Delfiner, 1999). The distribution of errors is assumed gaussian for this approach.

There is a set of widely used semivariogram models which are present in many mining commercial packages, such as spherical, exponential, gaussian, etc. which are licit models considering dimensional spaces up to $\mathbb{R}^3$. In mining it is very unlikely to deal with data in higher dimensions than $\mathbb{R}^3$. By adding extra dimensions to the conditioning dataset it is highly probable the dimensional space increases to $\mathbb{R}^n$ with $n > 3$ and therefore some of the semivariogram models valid in $\mathbb{R}^3$ would end up not being licit models in such higher dimensions. The problem of using a non licit model is the possibility of get negative conditional variances (see chapter 2). The conditional variance is a linear combination of the covariances $Var\{Y^*(\mathbf{u}_0)\} = \sum_{\alpha=0}^{n}\sum_{\beta=0}^{n}\lambda_\alpha\lambda_\beta C(\mathbf{u}_\alpha - \mathbf{u}_\beta) \geq 0$ and must be non negative (Goovaerts, 1997). To ensure this the covariance must be positive semi-definite and/or the semivariogram negative semi-definite (Goovaerts, 1997). The number of dimensions of the space is important for choosing a semivariogram model, a positive definite covariance function in $\mathbb{R}^m$ is also positive definite in $\mathbb{R}^n$ if $m \geq n$, however it is not necessarily valid for $m < n$ (Chilés & Delfiner, 1999). Two examples are presented in (Armstrong & Jabin, 1981) where it is shown that a semivariogram model would lead to negative conditional variances. The set of covariance models that can be used for the proposed approach are reduced to the ones that are positive-definite in any dimensions. There is a variety of semivariogram models that are proved are licit in any dimensions, some of them are:

- Spherical models based on a sphere of $\mathbb{R}^n$. The spherical semivariogram model without any specification of the dimension n is commonly referred to as the spherical model in $\mathbb{R}^3$, also there are other well known covariance models such as spherical in $\mathbb{R}^2$ is known as the circular model, and in $\mathbb{R}^1$ the triangle model (Chilés & Delfiner, 1999).
- Exponential models are positive definite in $\mathbb{R}^n$. Also Radon transforms of the exponential covariance provide differentiable covariances that are also valid in (Chilés & Delfiner, 1999).
- Gaussian model with a scale parameter greater than zero (Chilés & Delfiner, 1999).
- Generalized Cauchi model (Chilés & Delfiner, 1999).
- K-Bessel model (Chilés & Delfiner, 1999).
- Logarithmic or de Wijsian model (Chilés & Delfiner, 1999).
- Stable model. The exponential and gaussian models belong to this family (Chilés & Delfiner, 1999).
- Matérn model (Rasmussen & Williams, 2008).

5.4. Algorithm for Conditional Distribution Fitting

The only additional parameter different from conventional practice is the definition of a confidence interval. However, the intrinsic hypothesis is not assumed in this approach even when a semivariogram model is proposed. The semivariogram model has to account for the spatial continuity of the major part of the domain, not an average as in conventional practice (see chapter 4). The part of the dataset which spatial variability is under-estimated by the semivariogram model is corrected by the conditional distribution fitting so that the semivariogram model accounts for the entire domain in the fitted higher dimension. However, the region where the spatial variability is over-estimated still remains the same. This is equivalent to a pessimistic fitting of the semivariogram model that makes the conditional distributions be properly accounted but overestimated. And a consequence is the geostatistical model is more uncertain than it should be, see discussion in Chapter 4.

The proposed algorithm can be considered as a prototype of a non-conventional geostatistical modeling which is aimed for the requirements of mining industry. Many different additional strategies of fitting the conditional distributions can be added according to other requirements in the model, such as maximizing the accuracies or maximizing the fitting of the conditional means, etc. The algorithm is presented as a workflow and each of the steps are discussed:

1) **Cross validation and verification of the input parameters.** The goodness of the reproduction of the conditional distributions is verified when compared to their respective true values. The data locations where the true values fall outside the confidence limit parameters are marked for conditional distribution fitting. If the semivariogram model is fitted in a pessimistic manner a very small amount of the conditional distributions will require to be fitted; conversely, when the semivariogram model is fitted in an optimistic manner a very large amount of conditional distributions will require fitting. This is why it is important for the semivariogram model to account for the spatial continuity of the major part of the domain, so that, only an optimal number of locations require fitting. Also, the

selection of the confidence interval influences in the proportion of samples to be fitted, it has to account fairly for the data values of the conditioning data.

2) **Verification of spatial relationship among the marked samples.** It would the case a group of the marked samples are part of secondary populations. The verification is based on a cross validation analysis using only the marked samples. The mutual samples that estimate the parameters of conditional distributions that account for the true values within the given confidence intervals are grouped. Finally for each group of samples and independent samples different dimensions are assigned. It can be interpreted as each identified pattern is assigned to a particular dimension. This is the number of required dimensions.
3) **Tuning the extra dimensions.** The distances on each the additional dimensions are calibrated until the conditional distributions at the marked locations account properly for their respective true values. The calibration process is:
    a. The samples at the marked locations are separated and the linear dependences between them are calculated using cross-validation. Samples that are mutually dependent are grouped and the number of groups becomes the number of extra dimensions to solve the dataset. Each group of samples shares only one extra dimension. This is done for simplicity, otherwise the problem might become intractable to solve.
    b. Small lengths are added to each respective extra dimension at the marked locations. Using cross-validation it is verified if the conditional distribution accounts for the true value within the confidence limits. The same dimension length is added to the non-marked locations where the accuracy of the prediction was affected negatively.
    c. Go to step b and repeat until all the marked locations account for the true values within the specified confidence limits.
    d. Once the marked locations account for the true values the state of the extra dimensions is saved as a solution of the problem. It is worth to mention that there is a negative impact in the some of the surrounding conditional distributions that are minimized in step b.
4) **Save Results.** Store the conditioning dataset including the information of the additional fitted dimensions as the new conditioning dataset.

The goal in the use of extra dimensions is to find sub stationary sub-regions in the domain that are suitable for modeling using conventional techniques. Each sub-region is assigned to a particular extra dimension so that the behaviour of the conditioning data on each dimension is more stable in terms of increments, $|z(\mathbf{u}) - z(\mathbf{u}+\mathbf{h})|$, in the original dimensional space (see Figure 5-5). These sub-regions vanish in the high dimensional space and the whole dataset is considered as stationary and any conventional technique can be used for modeling uncertainty both in local and global terms. In the proposed algorithm step 3-a calculates the number of extra dimensions required for solving the problem according to the confidence interval parameters.

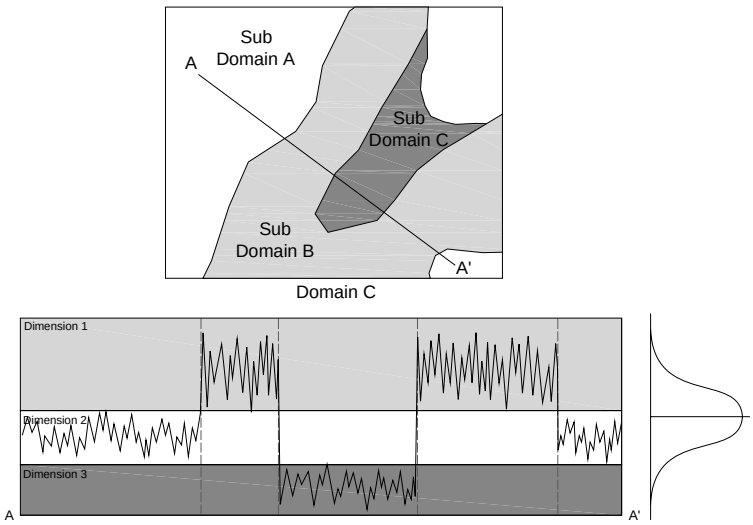

Figure 5-5: Sketch of classification of sub-domains by using extra dimensions

5.5. Case Study

Consider the borehole Ddh-81 dataset in normal score units presented in chapter 4 and the cleaned semivariogram model (4.9) for 99.7% probability as input information for building a geostatistical model. The confidence interval probability parameter chosen for this example is 95%.

The algorithm solves the dataset using three additional dimensions at twenty data point locations. From the three additional extra dimensions the first one consists of eleven data locations, the second one of five data locations and the third one of four data locations. The presence of outlier data points in the semivariogram cleaning exercise in chapter 4 showed that some data pairs are not accounted for by the semivariogram model (4.9) due to their large increments in the data pairs. The same locations are also identified in this approach. The initial accuracy of the conditional distributions using the conventional approach show some locations which true values fall outside their confidence limits (see Figure 5-6-top left). Even when the algorithm approaches the conditional distributions to the accuracy targets, there might be some locations that cannot solved completely. However, they will tend to be close to the fitting conditions and are accepted as solutions using small tolerances in the approximation. That is not the case in this example, the accuracy of all the locations are solved successfully (see Figure 5-6-top right). The improvement in the calculation of the conditional distributions is evident, however the accuracy of a small proportion of locations is slightly negatively affected at very few locations but they are still within the fitting target limits (see Figure 5-6-bottom).

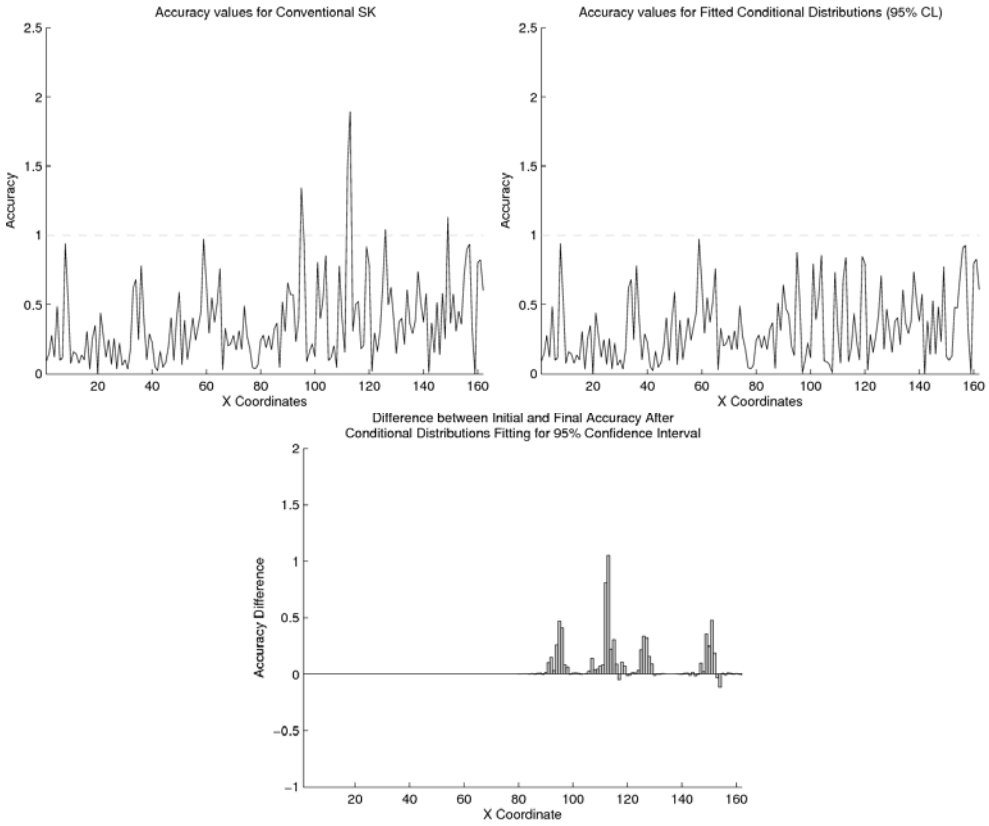

Figure 5-6: Initial status of the accuracy of the conditional distributions calculated using the conventional approach (top left), conditional distribution fitting (top –right) and comparison of them (initial - fitted) (bottom)

The fitting of the conditional distributions have an inevitable impact in the estimation of the SK means. For the locations where the conditional distributions are fitted there are negligible improvements in the prediction of the SK means. Because, what the approach does is to make them locally more uncertain. However, the influence of fitted data values (outliers) on the surrounding data locations is reduced. This makes the surrounding samples to be influenced mostly by the non-outlier data values when their conditional distributions are calculated. The improvement of the SK means can be seen when the conventional approach is compared to the conditional distribution fitting approach (see Figure 5-7). The correlation coefficients of the SK means compared to their true values show a small improvement, that is, 0.8253 for the conventional approach and 0.8671 for the proposed approach. Graphically the SK means of the proposed approach (right side) show a better local fitting than in the conventional case (see Figure 5-7-left side).

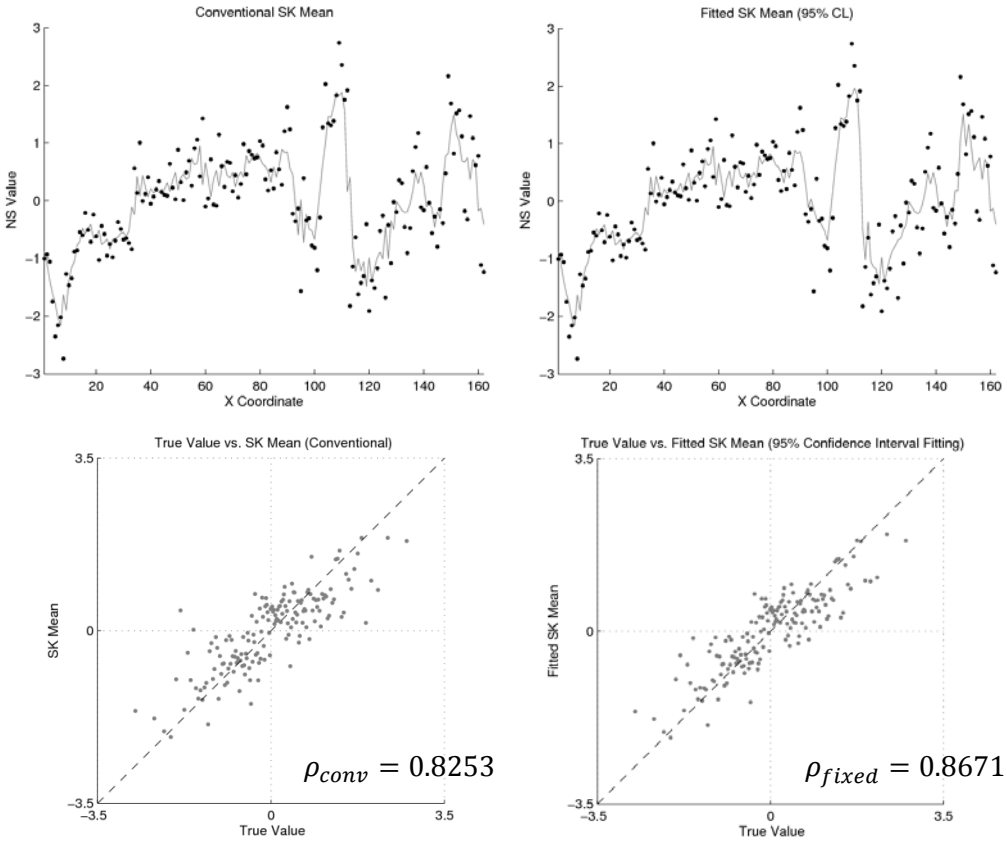

Figure 5-7: Cross validation SK means calculated using conventional approach (top – left), conditional distributions fitting (top right), and scatter plots of cross validation SK mean versus true values using conventional approach (bottom left), conditional distributions fitting (bottom right)

As well as the SK means the conditional variances are also affected. At the fitted locations the conditional variances tend to increase until the true values fit within the confidence limits of the conditional distribution (see Figure 5-8). The conditional variances can be considered as data dependent in the original dimensional space, since the tuning of the conditional distribution is based on the occurrence of the true values within their respective conditional distributions. While they still remain stationary and configuration dependent in the fitted higher dimensional space.

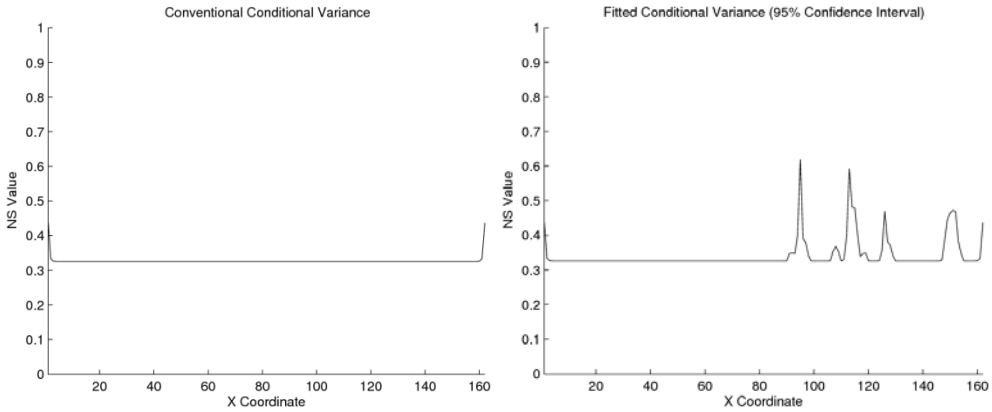

Figure 5-8: Conditional variances of conventional approach (left) and conditional distribution fitting (right)

The number of additional dimensions and locations that require conditional distribution fitting tend to decrease as the confidence limits of the conditional distributions increases, that is because fitting to larger probability intervals is less demanding for the proposed algorithm (see Figure 5-9). For the probability intervals of 97% and 99% the number of required dimension are three and two respectively and the outlier sample locations fifteen and twelve.

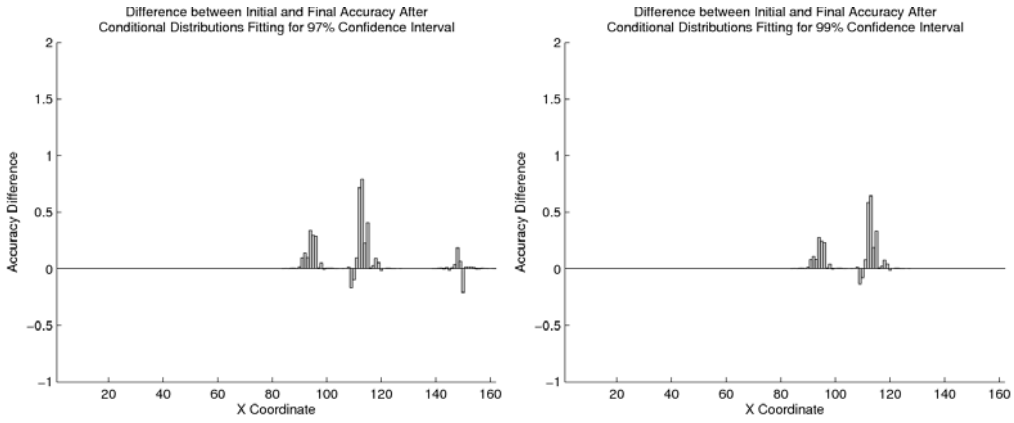

Figure 5-9: Improvements in the accuracy of conditional distributions for 97% confidence interval (left) and 99% confidence interval (right)

5.6. Discussion

The extra dimensions capture the information of patterns in the domain that are unaccounted for by the proposed semivariogram model. Patterns in the domain are present due to the different geologic processes that deposit the concentrations of metal grades or any other element of interest in the domain. Modeling the domain without taking into account such patterns may result in a non-realistic representation of the domain. Consider the same domain as in Figure 5-5 which consists of three stationary sub-domains A, B and C (see Figure 5-10-top). The resulting domain becomes non-stationary even when the local mean and variance are assumed to be constant. This can be seen in a cross section of the values of the domain. Notice that in Figure 5-10-bottom, for

calculating the conditional distributions in the regions of sub-domain A the SK weights of the data points of the sub-domains B and C should be less relevant than of the sub-domain A. In the proposed approach, the degree of relevance in the estimation of the conditional distributions at any location of the domain is tuned by the extra dimensions. When estimating at any location of the sub-domain A the samples of sub-domain B and C become less relevant because the extra dimensions tend to move the samples from domain B and C away the domain A or in another words make other sub-domains samples much more different. The effect of the additional dimensions for fitting the conditional distributions along the domain can be considered as an enhanced form of anisotropy since the directions of preferential continuity are defined more precisely and are shaped by the existing data, that is, data dependent.

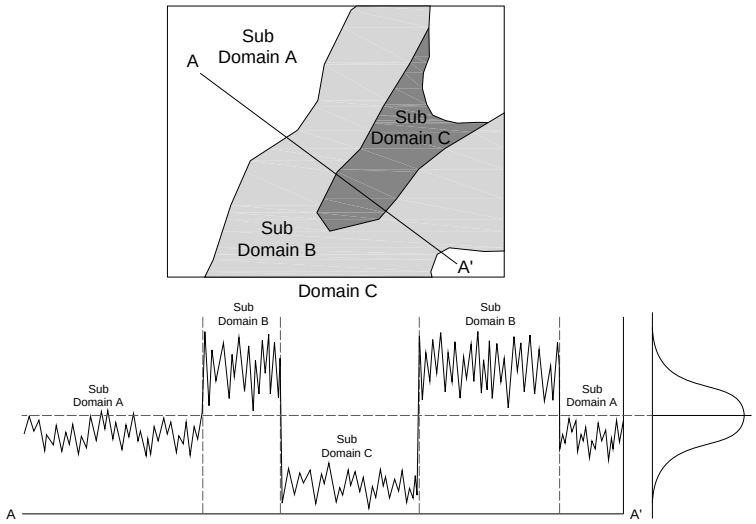

Figure 5-10: Sketch of combination o three stationary sub-domains A, B and C into a bigger one that mimic a geologic process (top), section of the resulting non-stationary domain which shows the patterns in data values (bottom)

In this proposed approach, cross validation is used in the process of fitting the conditional distributions. For each data location the best condition to analyze the estimation of the parameters of its conditional distribution is by using as much information as possible, because for modeling the unsampled locations in the domain the entire dataset is used. The larger the dataset, the better the understanding of the domain. Other testing techniques such as jackknife are not considered because the analysis has to be done locally, that is, sample by sample. Jackknife tends to be more global and that is not the purpose of this approach.

The over-estimation of the spatial variability in some regions of the domain cannot be identified by using the experimental semivariogram and the semivariogram model only. The solution of these sub-regions would require the reduction of the distances between the existing data locations and perhaps the definition of a new more continuous semivariogram model. There is no easy way to find out the conditions when the conditional distributions have to be narrowed. Finally, for identifying the conditional distributions which are over-estimated the analysis has to be made in groups of data locations rather than individual samples.

In this chapter the high-dimensional configuration of the conditional dataset is assumed to behave more stationary in the sense that the conditional distributions

calculated via cross-validation account properly for the true data. This condition is extrapolated to the rest of the domain, that is, the estimated conditional distributions of the domain at the unsampled locations in the same dimensional space also accounts for reality as the conditional data does, both local and global. In the next chapter, a methodology for simulating in the domain accounting for the high-dimensional space of the conditional dataset is presented. The features of the non-regular shape anisotropic patterns are also discussed.

6. Sequential Gaussian Simulation of High Dimensional Stationary Data

In Chapter 4 the influence in the experimental semivariogram of highly variable sub-regions in the domain was identified and removed. The resulting cleaned experimental semivariogram is fitted to get a semivariogram model and this represents the spatial continuity of the rest of the domain without the influence of the highly variable sub-regions. Therefore, the variability of the highly variable sub-regions is not accounted for by the cleaned semivariogram model. Consequently, this semivariogram may tend to be more continuous in presence of highly variable sub-regions. The conditional distributions of the data locations in these sub-regions are not properly represented by the conventional semivariogram model, and even less so by the cleaned semivariogram model.
Chapter 5 proposed a method to fix the conditional distributions at these locations by separating them from the rest of the conditioning data, and then adding extra dimensions to the initial Euclidean space in order to inject more uncertainty to the conditional distributions. The resulting spatial configuration is a high dimensional version of the initial conditional data or alternative conditioning data (ACD), that behaves more stationary than in the initial space or of the original conditioning data (OCD) in that the conditional distributions properly account for local uncertainty.

In this chapter, this new, high dimensional dataset is used to simulate geostatistical realizations of the domain. Transferring the highly dimensional information to the domain and an alternative representation of anisotropy that does not rely on the conventional elliptic pattern are discussed. After getting a high dimensional state of the conditioning dataset and transferring that information to the domain, virtually any conventional approach can be applied to reproduce the non-stationary features related to the intrinsic assumption of the RF.

6.1. Proposed spatial analysis and approximation of additional dimensions

To show and compare the proposed approach, a real dataset is used. Specifically, the Jura dataset (Goovaerts, 1997) consists of 100 data points placed over a fairly regular grid configuration (see Figure 6-1). This particular data configuration helps to get a better experimental semivariogram without using large tolerances and also to highlight the differences between the conventional and the proposed dimensional modeling approaches. The element analyzed is Cobalt (Co). The data values are transformed to normal scores for sequential Gaussian simulation and it is assumed there is no presence of trends in either the mean or the variance.

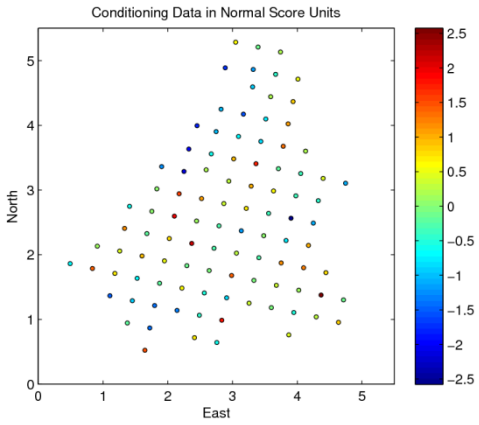

Figure 6-1: Configuration of the Jura dataset of Co variable in normal score units

The direction of major continuity is calculated at 45 degrees azimuth based on contour lines plotted from a semivariogram map (see Figure 6-2). The proposed semivariogram model consists of two anisotropic structures, both can be fit as exponential structures (see Figure 6-3). The first structure is isotropic with range equal to 0.6 distance units and the second one is anisotropic with the major axis equal to 5.0 units and minor axis is equal to 0.6 units (6.1).

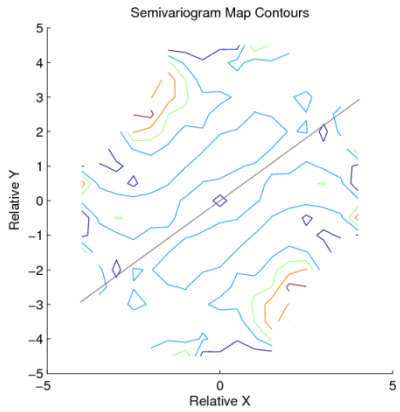

Figure 6-2: Semivariogram map of NS Co

$$\gamma(\mathbf{h}) = 0.6Exp_{\substack{mj=0.6 \\ mn=0.6}}(\mathbf{h}) + 0.4Exp_{\substack{mj=5.0 \\ mn=0.6}}(\mathbf{h}) \qquad (6.1)$$

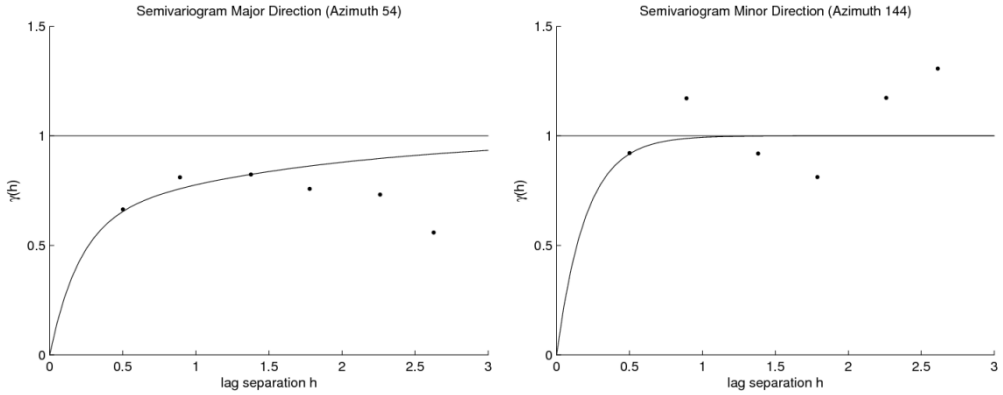

Figure 6-3: Semivariogram model (solid lines) and experimental semivariograms (black dots) of major (left) and minor direction (right)

The experimental semivariogram of the OCD is cleaned using a 99% probability cut-off. For the 2D case, both the control limit and the semivariogram model are surfaces. The control limit splits the cloud semivariogram in two parts. The first part contains the valid increments of the experimental semivariogram and the second part the outlier increments. The latter represents the impact of the variability of the high variable sub-regions in the domain. The cleaned experimental semivariogram is calculated using the valid increments in order to eliminate the influence of the highly variable sub-regions if they exist (see Figure 6-4).

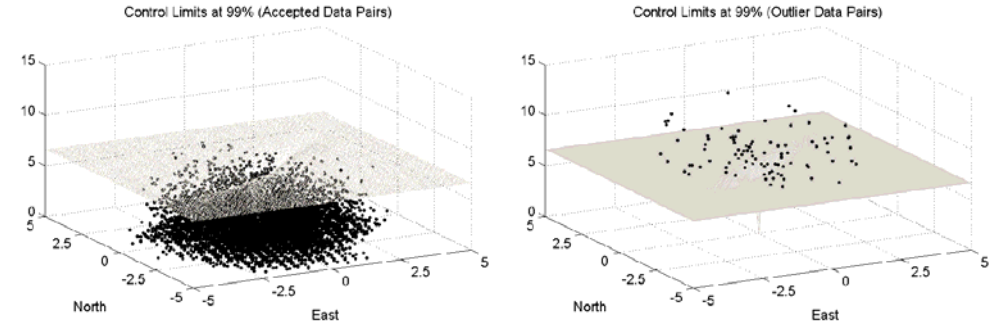

Figure 6-4: Cloud semivariogram split in two parts by a control limit at 99% probability, valid increments (left) and outlier increments (right)

In presence of highly variable regions in the domain some significant changes in the cloud semivariogram may result. Those changes might lead to differences in the main anisotropic orientations, the use of different semivariogram model for fitting the cleaned experimental semivariogram and larger ranges of the new semivariogram model. For this dataset, the differences in the experimental semivariogram values are too minor to support the use of a different semivariogram model (see Figure 6-5) or a different anisotropic configuration (see Figure 6-6).

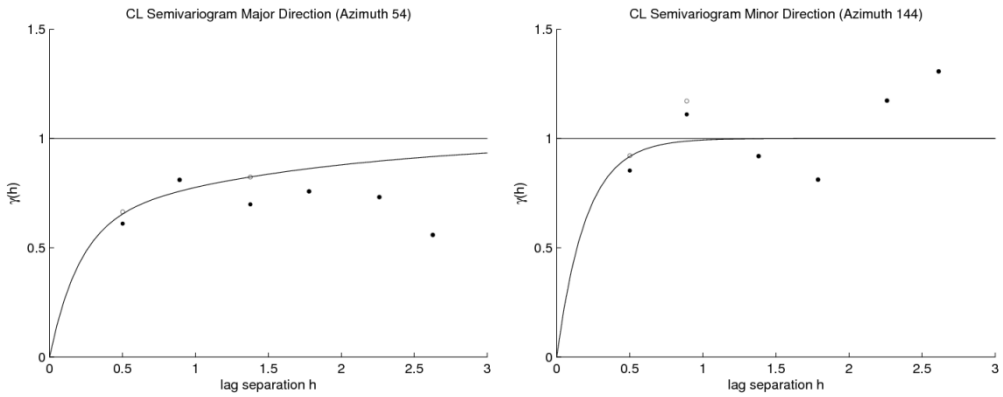

Figure 6-5: Semivariogram model (solid lines) and Experimental semivariograms (black dots) of major (left) and minor direction (right) after experimental semivariogram cleaning at 99% probability cut-off (black dots) and original experimental semivariogram points (empty gray dots)

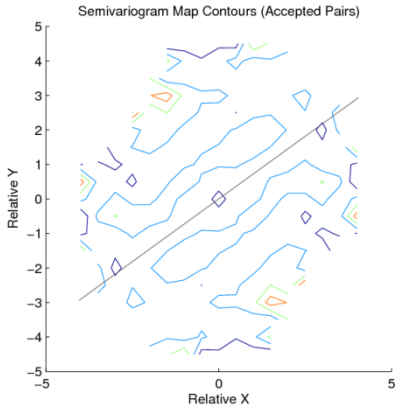

Figure 6-6: Semivariogram map after cleaning increment outliers at 99% probability

The locations of the data grouped in the outlier data pairs are plotted in Figure 6-7. The size of the black dots in the map represents the proportions of the number of times each sample location participates in the outlier increments, the data locations with large occurrences can be considered as problematic locations during modeling because they make large increments occur in the experimental semivariogram. This information can be used to identify sub-patterns in the domain that may cause problems when modeling the variable of interest. The sample locations with small occurrences are not considered as indicators of potential sub-domains. In Figure 6-7 there is a group of problematic data locations on the west side that may be separated into a sub-domain, however, the frequency of those locations are relatively small compared to the two more problematic locations in the east and south-east region. For the present case study no sub-domaining is considered.

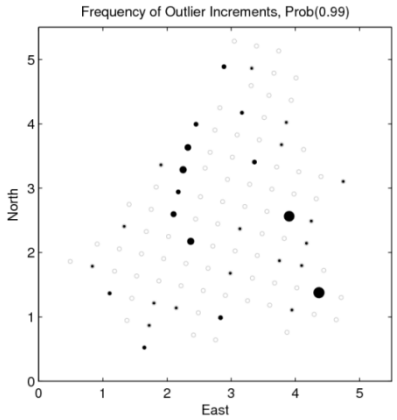

Figure 6-7: Occurrences of data locations that make outlier data pairs for a cut-off probability of 99%

6.2. Transferring ACD high-dimensional information into the domain

In order to be able to use the ACD in estimation/simulation, the domain of the geologic deposit has to be in the same dimensional space as the ACD. The process of moving the domain from $\mathbb{R}^n$ to $\mathbb{R}^m, \forall\, m > n$ has to be bijective. This is important for two reasons: 1) after estimating and/or simulating in $\mathbb{R}^m$ the model of the domain is analysed and studied in the OCD space, that is, $\mathbb{R}^n$; and 2) there has to be a unique correspondence between $\mathbb{R}^m$ and $\mathbb{R}^n$ because the OCD has to be exactly reproduced in the domain. The latter means the ACD projected in $\mathbb{R}^n$ has to be the OCD and no other in order to ensure the exact reproduction of the conditioning data in the domain.

The process of moving the domain from $\mathbb{R}^n$ to $\mathbb{R}^m$ consist of moving the nodes that represent the domain, each node has to have the same dimensions as the ACD. However, the values of the extra dimensions at the node locations are unknown and the process of transferring the extra dimension of the ACD can be subject to many interpretations. The extra dimensions can be seen as summarizing extra physical information about the domain (see Chapter 5) and would be more convenient if they are modeled as such. This results in an additional estimation/simulation problem to get the extra dimensions at the nodes of the domain.

When modeling a mineral deposit, an outlier increment is the result of an abrupt change in the metal grade values of the data pair that is uncommon in the rest of the domain. If the problematic locations cannot be separated into another sub-domain, the outlier increments may occur because of the presence of a minor geologic structure within the domain, such as veins or faults that are difficult to separate into sub-domains because of their small scale (see Figure 6-8). For example, in a Cu skarn deposit, such an abrupt transition might be due to the presence of a boundary between chalcopyrite ($CuFeS_2$) and bornite (Cu_5FeS_4) minerals. When the deposit is separated in domains, both are classified as exoskarn.

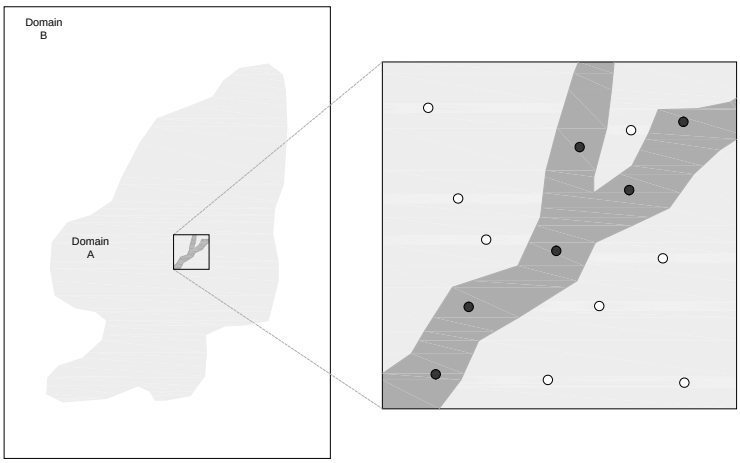

Figure 6-8: Sketch of a sub-structure (right) identified using outlier increments present in a large domain (left), samples with extra dimension (black dots) show the presence of an anomaly in Domain A when compared to the rest of the samples (empty dots)

Consider two data points with values in normal score units at locations **u1** and **u2** that make an outlier data pair (see Figure 6-9 top left). The outlier data pair implies that for this particular increment the averaged semivariogram model $\gamma(\mathbf{d})$ is not representative locally and the same is also assumed to be valid for the locations in between. The node locations of the domain have to be projected in the same higher dimensional space as the ACD prior to estimation/simulation. However, the only known locations in the high-dimensional space are the ones that correspond to the ACD (see Figure 6-9 bottom left). They are taken as reference for the transformation.

The technique proposed for transformation takes each of the extra dimensions of the ACD, one at a time, and consider them as additional variables to predict. That is, keeping the original dimensional space of the OCD make the dimensions the new variables to predict at the grid node locations. After all the dimensions are modeled in the domain as additional variables, it is assumed the domain is in the higher dimensional space. The transition of each extra dimension in the domain between sample locations in the original space can be modeled considering many different scenarios. For example, they can be modeled considering a linear transition, parabolic, convex, etc., also using geological interpretations of the extra dimension based on the fact they represent physical properties of the domain (see Figure 6-9 case1-4). Also, geostatistics can be used to model them. Therefore, the uncertainty associated to the extra dimensions can be introduced to the model. However, this might give unrealistic results because of the assumption of stationarity. The extra dimensions are the non-stationary part of the domain, and are only located in specific regions of the domain. Each extra dimension is not present throughout the domain because they would then become part of the random variability and can be captured by the nugget effect in the conventional experimental semivariogram.

Assigning predefined shapes of transitions may introduce bias due to the subjectivity of the decision. The most simplistic way to transfer the extra dimensions is the linear transition which in 2D and 3D original spaces is triangulation (see Figure 6-9 top middle). The advantage of triangulation is simplicity in terms of parameters. This is suitable when knowledge of the physical characteristics of the extra dimensions in the domain is lacking. Many techniques are available for implementing triangulation such as Delaunay triangulation algorithm which can be used in 2D and 3D with no problems. However, it may still be necessary to define boundaries of influence for the triangulation

algorithm, which control the influence of the extra dimensions in the domain. These boundaries can be generated using nearest neighbour techniques.

The previous approach only gives one scenario of the dimensions. However, modeling the uncertainty associated to the extra dimensions would also be required. The lack of knowledge of the physical properties may lead to an unsuitable modeling approach of the dimensions. As mentioned before conventional geostatistics is inappropriate for modeling the extra dimensions because the extra dimensions cannot be assumed as part of a SRF. Any modeling technique that accounts for physical parameters would be more appropriate. Finally, the decision between the simplistic and geologically more realistic approach involves a high price when the complexity of the domain is modeled. However, a mineral deposit is of such complexity and it becomes more important when locally consistent models are required.

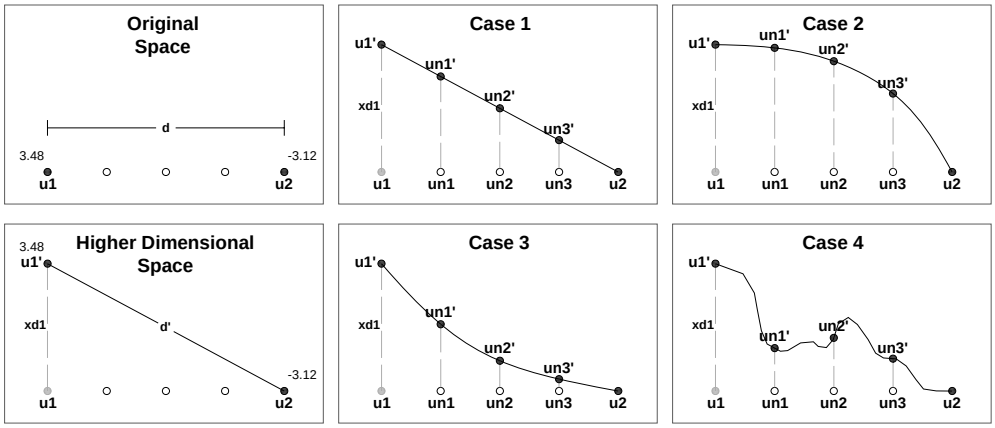

Figure 6-9: Sketches of cases of transferring extra dimension into grid nodes locations (empty dots), five transition cases are presented: linear (top middle), convex (top right), linear (bottom left), concave (bottom middle) and irregular (bottom right)

For the case study the extra dimensions are calculated considering a 95% probability confidence interval of the conditional distributions. One solution of the ACD dataset is found using five extra dimensions, and a triangulation technique using influence boundaries is considered for modeling the dimensions in the domain. The boundary is used to limit the influence of the extra dimensions (see Figure 6-10).

The new location vectors of the domain grid nodes are the combination of the initial location vectors and extra dimensions modeled in the five maps. Recall that the location vector of the ACD is the combination of the initial location vector and the extra dimensions. However, only one of the extra dimensions has a value greater than zero. In chapter 5 this condition was set up for simplicity. Since the new location vectors are the combination of the initial location vectors and the additional models of extra dimensions, the new location vectors of the nodes may end up having more than one extra dimensional component with values greater than zero. This is still valid and there is no problem associated to the consistency of the dimensional space.

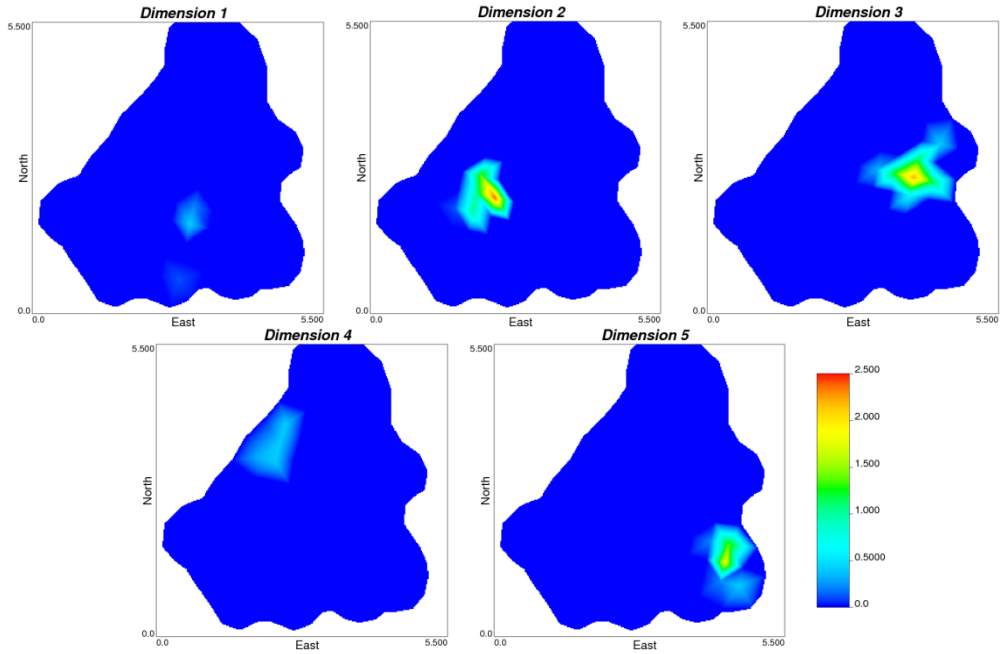

Figure 6-10: ACD extra dimensions transferred to domain node locations using a triangulation algorithm

After estimating/simulating in the high dimensional space the results have to be returned to the initial space to analyze the results. To do this, the estimated/simulated nodes of the domain are projected into the initial space by simply making all the extra dimensions equal to zero. There are no two node locations in the higher dimensional space that, after the projection in the initial space, fall in the same position because the components of the initial dimension remain the same. No matter the combination of the extra dimensions, the initial components will remain the same and that guarantees the bijective feature of the dimensional transformation of the domain. Consider a surface in a Euclidean $\mathbb{R}^3$ space, where X and Y axes correspond to east and north respectively and Z to elevation. If the surface is sampled over a regular grid for projecting in a XY $\mathbb{R}^2$ plane, Z simply becomes zero and there is no problem in the projection because there are no two similar combinations of X and Y in $\mathbb{R}^3$. If the projection is made over the plane XZ or YZ the projection will result in non-unique projections because there are similar values of X and Y for the projected planes respectively. In the case of the domains in hyper-dimensional space the same logic is applied, as long as there is a unique vector of the initial space, the rest of the components will not make the projection problematic.

6.3. Anisotropy in Original and Alternative Dimensional Spaces

The Euclidean space is considered as an isotropic environment because the measure of the distances is not preferential in any direction. On the other hand, an anisotropic environment can be seen as a deformation of a Euclidean space in some particular directions. The distance between two points in this case becomes a function of these anisotropic directions. A simple form of anisotropy is the ellipsoidal pattern. In $\mathbb{R}^3$ the preferential directions are defined by the three axes of the ellipsoid. They are usually referred to as major, minor and vertical axes. This form of regular shape anisotropy is suitable to be used in conventional geostatistics because of the assumption second order

stationarity of the IRF $\gamma(\mathbf{h}) = \gamma(-\mathbf{h})$. Assuming an ellipsoidal anisotropic environment the separation vector **h** can be scaled by a square matrix **A** (Isaaks & Srivastava, 1989):

$$\mathbf{h}_a = \mathbf{A} \times \mathbf{h} \tag{6.2}$$

The form of the anisotropic scaling matrix for any number of dimensions is (6.3):

$$\mathbf{A} = \left[a_{i,j} \middle| \begin{matrix} a_{i,j} = t_i, i = j \\ a_{i,j} = 0 \text{ elsewhere} \end{matrix} \right], \forall i, j \in \{1, ..., n\} \tag{6.3}$$

The diagonal elements of the anisotropy matrix are the scaling factors for each dimension of the separation vector **h**. The t_i factors in the anisotropy matrix **A** scale the elements of the separation vector **h** linearly, each anisotropic component of the separation vector $\mathbf{h}_a$ is $h_{ai} = t_i h_i$. In the case of conventional estimation/simulation the problem is finding the scaling factors of the elements of the location vectors **u** that make the spatial continuity of the new scaled space be defined by the proposed semivariogram in its isotropic form for each of the nested structures. The anisotropic deformation of the distances is inversely proportional to the semivariogram anisotropic scheme. In the direction of major continuity the anisotropic distance has to be reduced in order to increase the covariance value and conversely in the direction of minor continuity. That is, $t_i = 1/q_i$, where q_i is the length of the anisotropic semivariogram ellipsoid axis that corresponds to dimension i.

An anisotropic scheme defined in $\mathbb{R}^n$ does not affect the extra elements of a separation vector **h** in $\mathbb{R}^m, \forall\, m > n$. The anisotropic matrix can be expanded to a higher dimension m by simply equalling the size of matrix **A** to the new total number of dimensions and adding one in the diagonal extra elements. That makes the new extra components remain the same $h_{ai} = 1h_i$.

The anisotropic matrix **A** deforms the original space in the directions of the elements of the **h** vector. However, the anisotropic ellipsoid can also involve rotation in space. The vector **h** can be rotated by multiplying it by a rotation matrix **R** (6.4). The form of the rotation matrix for any number of dimensions is (6.5); it rotates the plane of dimensions a, b (Aguilera & Pérez-Aguila, 2004). Notice the rotation directions are based on the order of the dimensions a, b. That is, θ is positive or counter-clockwise for plane a, b and negative or clockwise for plane b, a. The rotation matrix **R** can be the resulting combination of different rotated planes. This is done by rotating vector **h** by each of the different planar rotation matrices.

$$\mathbf{h}_r = \mathbf{R} \times \mathbf{h} \tag{6.4}$$

$$\mathbf{R}_{a,b} = \left[r_{i,j} \middle| \begin{matrix} r_{a,a} = \cos\theta \\ r_{b,b} = \cos\theta \\ r_{a,b} = -\sin\theta \\ r_{b,a} = \sin\theta \\ r_{j,j} = 1,\ j \neq a,\ j \neq b \\ r_{i,j} = 0 \text{ elsewhere} \end{matrix} \right] \tag{6.5}$$

The rotation matrix **A** rotates vector **h** according to the specific rotation angles. However, the system is meant to be rotated rather than the vector. For rotating the system the vector **h** is multiplied by the transpose of the rotation matrix (6.6). Finally, the spatial configuration of the anisotropic ellipsoid is integrated by multiplying the transpose of the rotation matrix **R** and the anisotropic matrix **A** (6.7). Recall that **RA** affects the vector, whereas $(\mathbf{RA})^T$ affects the system. This is consistent with the anisotropic effect in the separation vector **h** in GsLib (Deutsch and Journel, 1998). There is no limit in the number of dimensions for accounting for anisotropy and it is easy to write programming code in the matrix form.

$$\mathbf{h}_s = \mathbf{S} \times \mathbf{h} = \mathbf{R}^T \times \mathbf{h} \tag{6.6}$$

$$\mathbf{h}_{as} = \left(\mathbf{RA}\right)^T \mathbf{h} \tag{6.7}$$

The following is a 2D example of accounting for anisotropy for a separation vector **h**:

$$\begin{aligned}
\mathbf{h}_{as} &= \left(\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} t_x & 0 \\ 0 & t_y \end{bmatrix} \right)^T \begin{bmatrix} h_x \\ h_y \end{bmatrix} \\
&= \left(\begin{bmatrix} t_x \cos\theta & -t_y \sin\theta \\ t_x \sin\theta & t_y \cos\theta \end{bmatrix} \right)^T \begin{bmatrix} h_x \\ h_y \end{bmatrix} \\
&= \begin{bmatrix} t_x \cos\theta & t_x \sin\theta \\ -t_y \sin\theta & t_y \cos\theta \end{bmatrix} \begin{bmatrix} h_x \\ h_y \end{bmatrix}
\end{aligned}$$

Notice that after rotating and deforming the system the new elements of the separation vector are a function of the scaling factors that correspond to each dimension $h_{ax} = t_x\left(h_x\cos\theta + h_y\sin\theta\right)$ and $h_{ay} = t_y\left(-h_x\sin\theta + h_y\cos\theta\right)$.

In practice, because of the different geologic events that formed the mineral deposit, the ideal scenario of linear estimator is one which accounts for locally different and of irregular-shape anisotropic patterns. Performing conventional estimation/simulation in such a complex environment would be virtually intractable. Even when a mineralized region has been exposed to a folding event it does not necessarily mean the anisotropic distances must be measured along the fold in order to get a better estimate. In such a situation, some approaches may consider unfolding the dimensional space of the domain. However, this may suggests that the unfolded domain is fairly stationary for estimation/simulation. A stationary environment in a high dimensional space when projected in the initial dimensional space can account for these irregular anisotropic patterns. The extra dimensions provide the necessary customized deformations in the initial dimensional space and at the same time the estimated conditional distributions account properly for both global and local uncertainty. Let us consider nine data point locations placed over a regular grid (see Figure 6-11). An isotropic pattern is compared to three anisotropy cases that are the result of adding one extra dimension to some of the nine sample point locations (see Figure 6-11). The values of the maps represent the distances measured from the center of the map to the rest of the locations. For the first case the distances are measured in the 2D plane directly. The second case corresponds to the sampling of data #6 which requires an extra dimension of length 5 units (see Figure

6-12-top right). For the third case, the previous configuration is kept and to sample data #1, we must assign an extra dimension of 2.5 units (see Figure 6-12-bottom left). Finally, for the fourth case, the configuration of case 2 is kept and sampling data #7 requires an extra dimension of 1.75 (see Figure 6-12-bottom right). For the three anisotropic cases only one extra dimension is used to modify the initial 2D space and a triangulation scheme is used to transfer the influence of the extra dimensions in the domain. Notice, projecting from a 3D surface to a 2D plane is sufficient for getting irregular anisotropic patterns in 2D. Recall that the extra dimensions account for the variability that cannot be represented in the initial space.

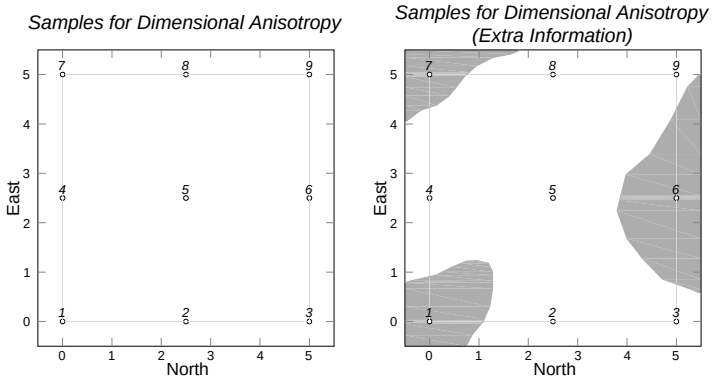

Figure 6-11: Initial configuration of nine data point locations in 2D for showing the effect of irregular shaped anisotropic patterns

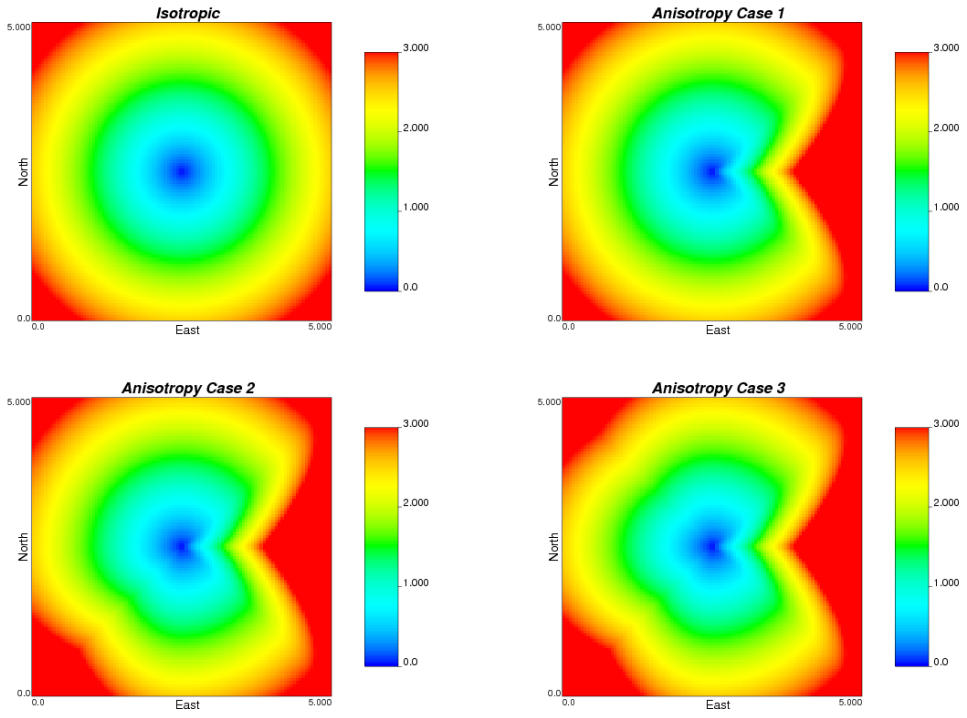

Figure 6-12: Comparison of an isotropic pattern (top left) against three cases of irregular shape anisotropic patterns with one sample with an extra dimension (top right), two

samples with extra dimension (bottom left) and three samples with extra dimension (bottom right)

The irregular anisotropic patterns account for the locations in the domain where the local uncertainty is under-estimated. The irregular anisotropic patterns are more flexible than conventional regular ones. Even when they can vary locally there is still the problem related to the symmetry of the anisotropic pattern. Samples that are located in one side of the location to estimate have to have exactly the same influence as the samples on the opposite side. In the conventional approach the influence of the surrounding conditioning data cannot be customized even in presence of geologic information that supports such decision. The additional information that might support irregularities in the symmetric influence of conditioning data can be extra geologic information. Recall that considering the variable to model is a regionalized variable and that any additional information is disregarded from the modeling process unless a co-regionalization environment is considered. However, co-regionalization considers dependence between primary and secondary variables, the dependence of the variable of interest might not be direct with each of the secondary variables one-by-one but a relationship of the primary variable with a combination of the rest of the variables. In Figure 6-11 left, for estimating/simulating at location #5 the conventional approach considers the spatial continuity is only a function of the variable of interest, whereas the domain usually has additional features that control the continuity variable of interest.

6.4. Comparison between Conventional and Proposed Approach.

The conventional approach shares the same anisotropic pattern all along the domain as a result of assuming second order stationarity. As a result, the model may be unrealistic because the estimates capture a global uncertainty, whereas the local features remain averaged. However, both the conventional and the proposed approach rely on the assumption of multi-gaussianity and reproduces features, such as semivariogram and global distribution (see Figure 6-13). For comparing the experimental semivariogram reproduction between the conventional and the dimensional approach, in the case of the latter only part of the domain share the same dimensional locations, that is, the regions where the extra dimensions are zero. Recall that the location vectors of the nodes affected by the extra dimensions (see Figure 6-10) are not the same as in the original space XY. Therefore, the experimental semivariograms calculated in the plane XY do not consider such node locations. From the total of 75625 nodes in the original dimensional space without considering any boundaries 58331 are not affected by the dimensional approach. The experimental semivariograms of the realizations in the dimensional approach (see Figure 6-13 top-left) are supported by the 58331 node locations, whereas, the experimental semivariograms of the conventional approach by 75625 nodes, this is one of the reasons why the experimental semivariograms reproduction of the dimensional approach looks more variable. In the case of the verification of the global distribution reproduction (see Figure 6-13 bottom) all the 75625 nodes are considered because this verification is not dimensional dependent.

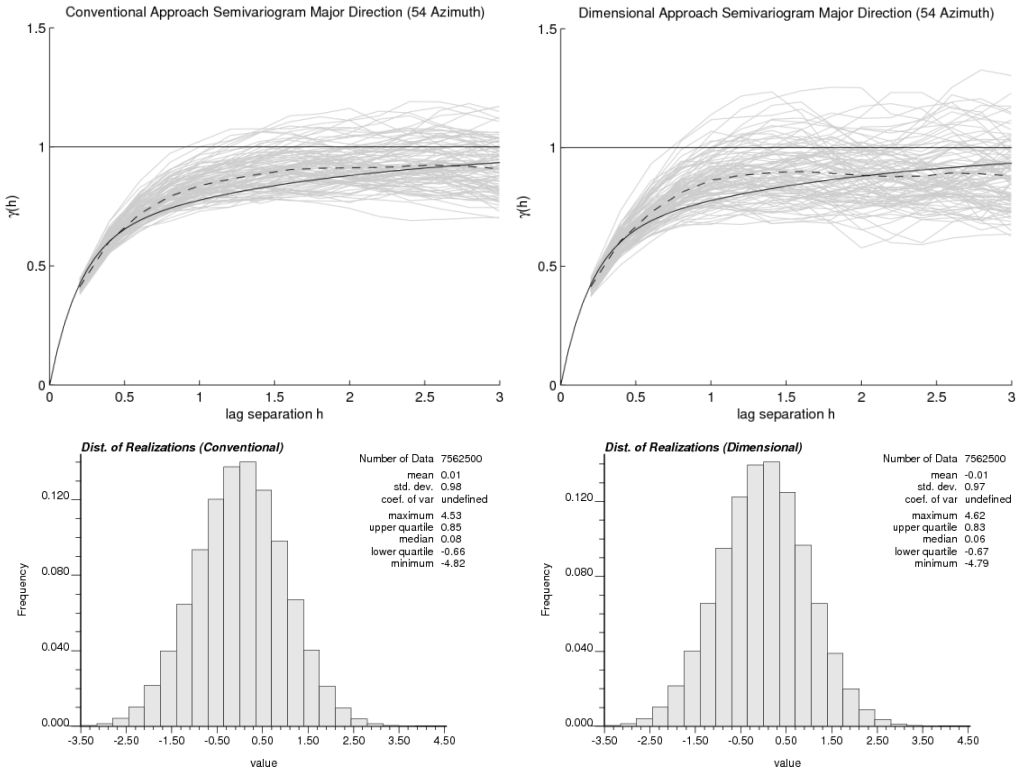

Figure 6-13: Experimental semivariogram and global distribution reproduction of 100 realizations for conventional approach (left side) and dimensional approach (right side)

The simulated nodes in the dimensional approach because of the deformation of the initial space reproduce non-stationary features of the semivariogram when projected in the initial dimension (see Figure 6-14-right). The conventional approach shows a constant anisotropic pattern along the domain (see Figure 6-14-left). In both cases the standard normal distribution is reproduced. In the conventional approach the semivariogram is reproduced in the initial space, however, as mentioned initially the semivariogram is not fully representative of the domain in the initial space and there is no reason for this. The approach relies on the domain being stationary in a hyper space; in that space the semivariogram is reproduced and the anisotropic patterns are constant.

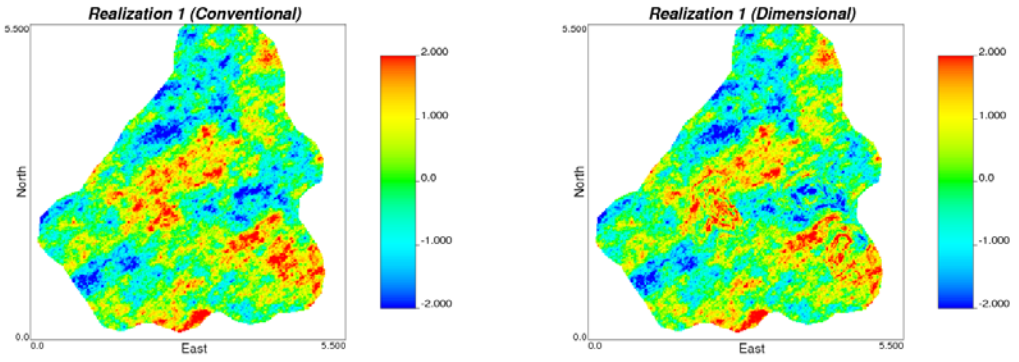

Figure 6-14: First realization using conventional simulation (left) and dimensional proposed approach (right)

After averaging many realizations the conditional estimated means for the dimensional approach also reproduce the non-stationary features as the realization maps and in the conventional approach the anisotropic pattern is constant (see Figure 6-15). Notice in the dimensional approach the data locations that were identified in the semivariogram analysis stage as problematic locations are not spread in the domain. The problematic data locations are extreme values either positive or negative. They become problematic when the samples that surround them are so different that the increments cannot be accounted by the semivariogram model, therefore the conditional distribution is under-estimated.

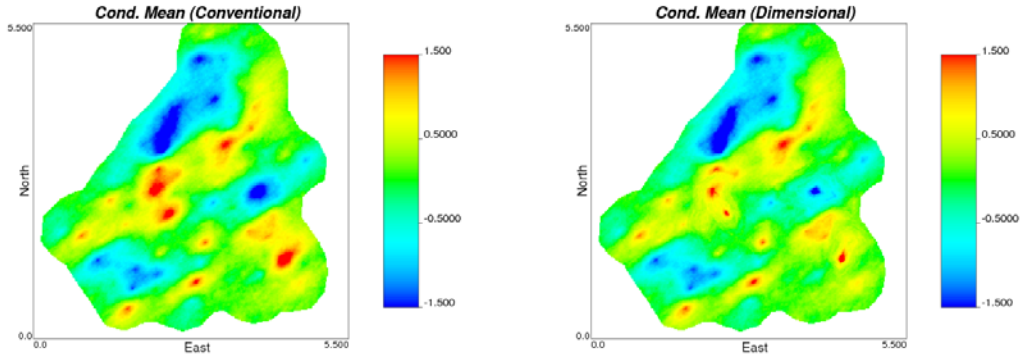

Figure 6-15: E-type map of 100 realizations in normal score units of conventional approach (left) and dimensional proposed approach (right)

The difference of the conditional means between the conventional and the dimensional approach for this exercise shows there is a preferential trend to constrain the higher values rather than lower values (see Figure 6-16). There is no condition that makes the mean of the distributions is centered to zero or the distribution is unbiased. It is a function of the occurrence of the problematic locations. The map of the differences shows that most of the regions constrain the high values and only in one region the influence of a low value sample is limited (see Figure 6-15 left).

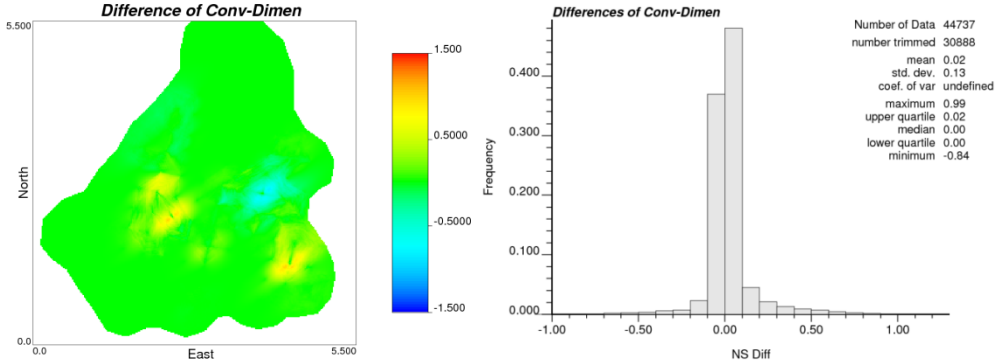

Figure 6-16: Map (left) and distribution (right) of difference between conventional and dimensional means.

The conditional variances are data configuration dependent; they are almost similar for regular configuration patterns of the conditioning data. In the example presented the conditioning data is placed fairly over a regular grid, the conditional variances are also in

a regular pattern because of the data configuration. In the dimensional approach the conditional variances are larger around the problematic locations (see Figure 6-17)

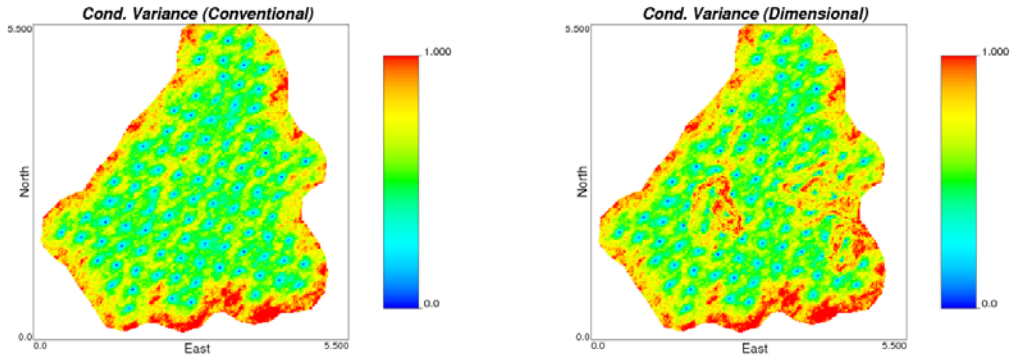

Figure 6-17: Local conditional variances map of 100 realizations in normal score units of conventional approach (left) and dimensional approach (right)

6.5. Discussion

The proposed approach accounts one second order non-stationary property of the domain, more specifically, the intrinsic hypothesis. Only the regions in the domain where the conditional variability are underestimated are targeted by this approach, which leaves the regions where the variability is over estimated invariant. The reason for this is the extra dimensions where they exist make the data locations more dissimilar than they are in the conventional approach, therefore, increasing the degree of local uncertainty. For reducing the variability there should exist a combination extra dimensions that when they are added makes the separation distances closer than in the original dimensional space, that is not possible using real numbers, however, the use of complex number can be considered which will end up making the problem almost intractable to solve. Moreover, for identifying locations where the local uncertainty is under-estimated it is only necessary to evaluate each data location using cross-validation technique. Whereas, for estimating locations where the local variability is over estimated not one location but many have to be evaluated in order to verify the same condition happens in all of them. From an engineering perspective, the resulting model can be seen as a pessimistic model where the local variability is overestimated.

The non-stationary features are solved by using a stationary methodology. Recall that there is nothing special in the estimation/simulation technique. In the higher dimensional space the domain is assumed and solved as stationary, that is, using a semivariogram/covariance model and a global symmetric anisotropic pattern. The non-stationary features are highlighted only after the higher dimensional domain is projected onto the original dimensional space. It is worth to mention, the assumption of regionalized variable tries to diminish the impact of additional physical information. Consider in a mineral deposit two data locations are remarkably different because some other geological variable or variables have also suffered also an abrupt change compared to the rest of the domain. In the proposed dimensional approach the extra dimensions try to account that ‘reason’ of outlier variability in terms of dimensions. That is, two locations are highly variable in the original dimensional space but not as much as variable in the higher dimensional space, the abrupt change in their extra dimensions tries to explain the high variability in the original space. The use of extra dimensions is also

another scenario to be modeled as the variable of interest is, there is uncertainty associated to the extra-dimensions. The proposed approach uses a simplistic approach when considers only one scenario of many. However, it is shown that the use of them helps to improve the prediction in terms of performance of the conditional distributions.

7. Conclusions

The use of conventional geostatistics in simulation is mostly aimed at long term models where the reproduction of global features is more important than of local features. Long term models are used for planning on a yearly basis. Therefore, compensating for the under-estimation of some regions with the over-estimation of others does not have a big impact in resource estimation over such a long period of time. The global assumption that the spatial continuity of the domain can be defined by a single semivariogram model and its corresponding anisotropic pattern may work properly.

Modeling uncertainty for planning on a smaller period of time than large term mine plans requires making modifications to the kriging system so that it accounts for local patterns. Some approaches that constrain the influence of extreme values were proposed (Arik, 1992), (Pan & Arik, 1993), the influence of extreme values is modified so that they are not spread to the rest of the values that are close to the mean of the global distribution. The resulting model is a pessimistic version of the conventional model because the extreme values do not have the same effect of contributing information as the rest of the samples. These techniques are a modification of ordinary kriging and some of them are implemented in many of the mining software packages e.g. MineSight®, Vulcan®, etc. Because these models are conservative it is safer but not optimal to make mine production plans on a weekly or monthly basis. However, the restriction made to the extreme values may have a geologic support in practice; for instance, the high extreme values are usually linked to high concentrations of minerals in the form of veins, mineralized faults or small anomalies that are unusual in the domain. Only estimated models can be built using these types of approaches.

The stationary conditions of the domain can be verified through the semivariogram. Trends in the mean and variance can be verified by expanding the experimental semivariogram expression. However, because the semivariogram is a two point statistic the available dataset has to be fairly representative of the domain. In presence of large unsampled regions in the domain the results may be misinterpreted. The condition of the intrinsic hypothesis of the semivariogram can be also verified by using the experimental semivariogram and the semivariogram model. It is necessary to assume multi-gaussianity if it is required to build a simulated model. Moreover, the semivariogram model is assumed to be a reasonable fit of the experimental semivariogram, neither optimistic nor pessimistic as discussed in Chapter 4. By using the semivariogram model, the analytical form of the distributions of the data pairs for each lag separation distance can be inferred. These analytical distributions are compared to the experimental distributions in order to find preferential patterns in the data pairs that are evidence of sub-regions with different spatial variability patterns.

In practice it is very difficult to find a valid situation where the stationarity of the semivariogram is valid, unless the domain was previously generated using an unconditional simulation. Real domains contain sub-regions with different spatial variability that is the result of different geologic processes. The sub-regions where spatial variability is under-estimated by the proposed semivariogram model can be identified in the domain. However, the decision to separate them into sub-domains lies with the geomodeler, because those sub-regions may be placed in specific areas or in the form of small internal sub-structures that are difficult to separate due to scale.

When the internal highly variable structures in the domain cannot be separated they can be modeled with the use of extra dimensions. These extra dimensions correct the distances of the samples in the internal structures to the rest of the domain, so that, in the high dimensional space they are more dissimilar. This correction is based on the consistency of the covariance model. In the original dimensional space the sub-set of samples that belong to the internal structures inflate the experimental semivariogram because the increments with respect to the rest of the samples are noticeably larger than the increments of the samples that belong to the majority of the domain. Moreover, when the samples of the internal structures are used for estimation they tend to give erroneous estimates. In practice, the goodness of the estimates can be verified by using cross validation and confidence limits. The use of the samples of the internal structures in estimation may make the prediction of the true value with respect to its conditional distribution inaccurate; on the other hand, if the influence of those samples is reduced the accuracy increases. The influence is controlled by the extra dimensions.

The proposed approach is based on the correction of the highly variable sub-regions that are not accounted by the conventional model. Recall the semivariogram model represents the average of the spatial variability of the domain. As an average there are sub-regions where spatial variability is underestimated as well as others that are overestimated. In the sub-regions where spatial variability is underestimated the dimensional approach corrects such variability. However, the sub-regions where spatial variability is overestimated are not modified. This makes the resulting model a sort of pessimistic model in terms of spatial variability. Reality may be expected to be less variable than the final model. This may be considered a good scenario from the economic analysis perspective, because the range of overestimated spatial variability gives a tolerance when the model is analyzed.

The extra dimensions can be calculated in many different ways. In this thesis the proposed algorithm minimizes the negative impact in the accuracy of the conditional distributions when calculating the extra dimensions of the samples in the internal structures and at the same time to use the smallest number of extra dimensions as possible. Other targets can be implemented in the algorithm such as maximizing accuracy without taking care of the number of extra dimensions. There is no exact solution for getting the extra dimensions. However, the goal is to fix the influence of the sample locations that belong to internal structures in the domain. Even after getting a set of extra dimensions that improves the accuracy of the estimation in a cross-validation basis, transferring such information to the rest of the domain for modeling uncertainty ends up in additional modeling problems. In this thesis a unique scenario is considered for simplicity.

In the proposed approach there are few problems for modeling in higher dimensional spaces and then projecting the results to the original dimensional space. The main item to take into account is to select a covariance model that is positive definite in such high dimensional space. In the high dimensional space the modeling is carried out under the

conventional approach, however, after projecting the results to the original space the regular shape anisotropic patterns are transformed to an irregular shape which may be geologically more realistic.

7.1. Contributions

A set of tools based on the experimental semivariogram that are used for verifying the impact of mean and variance trends in the domain are proposed in this thesis (see Chapter 3). The assumption of multi-gaussianity is not a requisite. They can be used for evaluating the sub-domaining process and verifying whether the techniques for trend removal modeling account completely for the effect of trends in the domain. These set of tools can be considered as complementary in the data analysis stage prior to modeling the mineral deposit. However, it requires the available dataset to be fairly sampled over the domain.

The stationary assumption of the semivariogram can be verified by using the methodology proposed in Chapter 4. The proposed approach requires setting up confidence limits and a proposed semivariogram model. This makes the approach a subjective interpretation of the results. Two examples are presented in Chapter 4 and Chapter 6 respectively where highly variable sub-regions are identified in the domain. Basically, this approach identifies groups of data samples that may be problematic during the modeling of the domain.

Highly variable regions in a domain make the stationary assumption of the semivariogram not valid for the domain. In practice such condition is usually omitted. The proposed approach in Chapter 5 moves the dataset from its initial dimensional space to a high dimensional space where the semivariogram accounts for all the sampled locations in the domain. However, this approach requires the use of confidence limits which requires a subjective interpretation of the results.

A novel approach for building a simulated model based on the pseudo-stationary high-dimensional dataset is proposed in Chapter 6. The realizations are simulated in the same dimensional space of the input dataset and then they are projected to the initial dimensional space. Once the projection is carried out non-regular anisotropic patterns can be found. Local structural features captured by the dataset are reproduced in the domain without compromising the validity of the multi-gaussian assumption. For transferring the domain to the dimensional space of the conditioning dataset a triangulation approach is used for simplicity. Alternative techniques that account for uncertainty of the extra dimensions are recommended to be used instead.

7.2. Future Work

By adding one extra dimension to a sample location makes the influence of it in the modeling process decrease because the separation distance is incremented. At the sample location in a cross-validation analysis the conditional distribution is more uncertain. This way, the locations where the local uncertainty is underestimated by the conventional approach is fixed. On the other hand, in order to reduce the uncertainty of the conditional distribution, the distances after adding the extra dimension should be reduced. This happens when the extra dimension is a complex number. The use of complex number in the dimensional approach can be analyzed in order to be able to expand the spatial continuity of the covariance model. Aspects of positive definite conditions under such environments can be studied in order to ensure the consistency of the approach. It is

worth to mention that modifying the initial dimensions may end up being impossible to project the results into the original space.

The proposed algorithm uses only one extra dimension per location of the conditioning data, this is done for simplicity. The use of more than one extra element in the position vector can be studied in order to make more flexible to deal with the negative impact in the accuracy of the surrounding samples. Solving the problem by using more than one extra element may increase significantly the complexity of the proposed algorithm.

In this thesis the transference of the extra dimensions to the domain has been carried out using a triangulation algorithm and considering only one scenario. Uncertainty associated to the transference of the extra dimensions at the unsampled locations can be studied. However, it is not recommended any geostatistical method because the extra dimensions are located in the domain following particular structures, that is, the variability of the extra dimensions only happen at specific regions in the domain. The use of a geostatistical approach may transfer the localized variability of the elements all over the domain because of the assumption of stationarity that is required. It is recommended to study a object based methodology in order to give more geologic realism to the identified geologic structures in the domain.

8. Bibliography

Aguilera, A., & Pérez-Aguila, R. (2004). General n-Dimensional Rotations. *WSCG'2004.* Plzen: UNION Agency - Science Press.

Arik, A. (1992). Outlier Restricted Kriging; A New Kriging Algorithm for Handling of Outlier High Grade Data in Ore Reserve Estimation. *APCOM'92*, (pp. 181-187). Tucson.

Armstrong, M., & Jabin, R. (1981). Variogram Models Must Be Positive Definite. *Mathematical Geology* , 455-459.

Chilés, J. P., & Delfiner, P. (1999). *Geostatistics, Modeling Spatial Uncertainty.* New York: Wiley-Interscience Publication.

Cressie, N., & Hawkins, D. (1980). Robust Estimation of the Variogram I. *Mathematical Geology* , 115-125.

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

Deutsch, C. V., & Journel, A. (1998). *GSLIB Geostatistical Software Library and User's Guide.* New York: Oxford Press.

Genton, M. (1998). Highly Robust Variogram Estimation. *Mathematical Geology* , 213-221.

Gneiting, T., Sasvári, Z., & Schlater, M. (2001). Analogies and Correspondences Between Variograms and Covariance Functions. *Advances in Applied Probability* , 617-630.

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

Gringarten, E., & Deutsch, C. V. (2001). Teacherès aide - Variogram interpretationand modeling. *Mathematical Geology* , 507-534.

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

Johnson, R., & Wichern, D. (2007). *Applied Multivariate Statistical Analysis.* New Jersey: Pearson Prentice Hall.

Journel, A. (1986). Models and Tools for Earth Sciences. *Mathematical Geology* , 119-140.

Journel, A., & Huijbregts, C. J. (1978). *Mining Geostatistics.* New York, USA: The Blackburn Press.

Larrondo, P., & Neufeld, C. (2003). *VARFIT: A Program for Semi-Automatic Variogram Modelling.* Edmonton: Centre of Computational Geostatistics.

Leuangthong, O. (2009, January). MIN E 613, Advanced Geostatistics I. *Non-Parametric and Multi-Variate Geostatistics* . Edmonton, Alberta, Canada: University of Alberta, Department of Civil and Environmental Engineering, School of Mining and Petroleoum.

Matheron, G. (1962). *Traité de Géostatistiqué Appliquée.* Paris: Technip.

McLennan, J. A. (2007). *The Decision of Stationarity.* Edmonton: University of Alberta.

Pan, G. P., & Arik, A. (1993). Restricted Kriging for Mixture of Grade Models. *Mathematical Geology* , 713-736.

Rasmussen, C., & Williams, C. (2008). *Gaussian Processes for Machine Learning.* The MIT Press.

Sinclair, A., & Blackwell, G. (2004). *Applied Mineral Inventory Estimation.* Cambridge University Press.

Yamamoto, J. K. (2000). An Alternative Measure of the Reliability of Ordinary Kriging Estimates. *Mathematical Geology* , 489-509.

Zanon, S. (2007). *Advanced Aspects of Sequential Gaussian Simulation.* Edmonton: University of Alberta.

Appendix A

Program Details

Individual programs were written for Chapters 3, 4, 5 and 6. Visual Basic .net was used for writing programs in GsLib format for Chapters 3 and 5. The algorithm proposed in Chapter 4 was written in MATLAB and the GsLib program SGSim was modified for simulating the results in Chapter 6.

A.1. Program for trend detection via experimental semivariogram calculation

The program **ExpVarM1** calculates directional experimental semivariograms in specified directions using expression (3.7). The output file is an ASCII file, formatted according to the output format option in its parameter file (see Figure A-1). The output file contains information of:

- Lag distance
- Mean of the sub-dataset at the head of the separation vector
- Mean of the sub-dataset at the tail of the separation vector
- Standard deviation of the sub-dataset at the head of the separation vector
- Standard deviation of the sub-dataset at the tail of the separation vector
- Covariance between the sub-datasets at the head and tail of the separation vector
- Experimental semivariogram value

```
                  Experimental Semivariogram
                  **************************

[INPUT-OUTPUT]
<outopt>                                                    *Output option
<inpfile>                                                   *Input Datafile
<outfile>                                                   *Output Datafile
<rptfile>t                                                  *Report File

[RUN-OPTIONS]
<colx> <coly> <colz>                                        *Col Ids for x,y,z coordinates
<colv>                                                      *Col Ids for variables
<ndir>                                                      *Number of directions
<azmt> <hzag> <hzbw> <dipa> <vtag> <vtbw> <lagh> <lagt> <nlag> *Direction i
```

Figure A-1: Parameters of **ExpVarM1**

The parameter file consists of:

- **outopt:** output file formatting option. (1) GsLib format where each of the calculated fields correspond to a column, one extra column is added which corresponds to the experimental semivariogram number, (2) table format where the first row contains headers and each of the calculated fields are written in columns, one extra column is added to identify the experimental semivariogram number.
- **inpfile:** input data file in GsLib format
- **outfile:** output data file according to **outopt** option
- **rptfile:** program execution report file
- **colx, coly, colz:** column ids in the input data file for *x*, *y* and *z* coordinates
- **colv:** column id for the variable to calculate the experimental semivariogram
- **ndir:** number of directions to calculate the experimental semivariogram
- **azmt, hzag, hzbw, dipa, vtag, vtbw, lagh, lagt, nlag:** azimuth, horizontal angular tolerance, horizontal bandwidth, dip angle, vertical angular tolerance, vertical bandwidth, lag separation distance, lag tolerance, and number of lags to calculate. It is the same format to define the directional semivariograms in GsLib format.

A.2. MATLAB scripts for experimental semivariogram cleaning

The experimental semivariogram cleaning algorithm was written in MATLAB code. The code was extracted from a larger script file in three parts (see Figure A-2, Figure A-3 Figure A-4). The code uses the statistical toolbox of MATLAB to calculate the confidence limits of a chi-square distribution. However, it is not difficult to rewrite the code in a standalone GsLib format program, an extra module that calculates the probabilities and corresponding values from an analytical function of a chi-square distribution has to be added. The part of the code presented is for solving a 2D dataset. In Chapter 4 a 1D example is presented, it was solved using a simplified version of the 2D code. There is no problem to adapt the code for a 3D dataset.

The first part of the code (see Figure A-2) consists of reading a GsLib format input file and setting up the upper confidence limit for a given probability (line 9).

```
1   %--------------------------------------------------------------------------
2   %Cloud Semivariogram Cleaning
3   %By Miguel A. Cuba Espinoza
4   %University of Alberta
5   %--------------------------------------------------------------------------
6
7   %Control limit probability
8 - prob=0.99;
9 - c2x=chi2inv(prob,1);
10
11  %Load Data
12 - load('ns_co.out');
13 - data=ns_co;
14 - clear ns_co;
15
16  %Specify Fields
17 - xcoord=data(1:length(data),1);
18 - ycoord=data(1:length(data),2);
19 - value=data(1:length(data),12);
20 - clear data;
```

Figure A-2: First part of MATLAB experimental semivariogram cleaning script

In the second part (see Figure A-3) the parameters of the proposed semivariogram model are configured (lines 42 - 45). For optimizing the code, the maximum number of data pairs is calculated for initializing the data pair arrays (lines 48 - 58), after that the necessary arrays are initialized (lines 62 - 70).

```
41      %Semivariogram parameters
42 -    range=[0.6 5.0;0.6 0.6];    %anisotropic ranges
43 -    contr=[0.6 0.4];            %contributions
44 -    rot=[cos(Azimuth) -1*sin(Azimuth);...
45           sin(Azimuth) cos(Azimuth)];  %rotation matrix
46
47      %Calculate Number of Pairs
48 -    NumPairs=0;
49 -    for i=1:length(value)
50 -        for j=1:length(value)
51 -            if i ~= j
52 -                dist=sqrt(((xcoord(i)-xcoord(j))^2)+((ycoord(i)-ycoord(j))^2));
53 -                if dist<=MaxDist
54 -                    NumPairs=NumPairs+1;
55 -                end
56 -            end
57 -        end
58 -    end
59
60      %Calculate Cloud Semivariogram and Control Limit
61      %-Semivariogram
62 -    SemiVarg=zeros(NumPairs,1);
63 -    RefX=zeros(NumPairs,1);
64 -    RefY=zeros(NumPairs,1);
65 -    SepDst=zeros(NumPairs,1);
66      %-Control Limit
67 -    C_Limit=zeros(NumPairs,1);
68 -    C_Flag=zeros(NumPairs,1);
69 -    head_idx=zeros(NumPairs,1);
70 -    tail_idx=zeros(NumPairs,1);
```

Figure A-3: Second part of MATLAB experimental semivariogram cleaning script

In the third part of the code (see Figure A-4), the values of the semivariogram data pairs are calculated and stored (line 79), as well as the corresponding separation vector (lines 81 - 83). Each data pair is compared to the control limit surface calculated from the semivariogram model and the defined upper confidence limits (lines 98 - 103). The semivariogram data pairs that fall outside the control limit are flagged as 1 and the data pairs that do not as 0.

```
  NumPairs=0;
for i=1:length(value)
    for j=1:length(value)
        if i ~= j
            dist=sqrt(((xcoord(i)-xcoord(j))^2)+((ycoord(i)-ycoord(j))^2));
            if dist<=MaxDist   %Proceed to Calculate
                NumPairs=NumPairs+1;
                SemiVarg(NumPairs)=0.5*((value(i)-value(j))^2);

                %i index is the base index
                RefX(NumPairs)=xcoord(j)-xcoord(i);
                RefY(NumPairs)=ycoord(j)-ycoord(i);
                SepDst(NumPairs)=abs(dist);
                head_idx(NumPairs)=i;
                tail_idx(NumPairs)=j;
                %control limit parameters
                u=[RefX(NumPairs) RefY(NumPairs)];
                semivrg_model=0;
                for k=1:length(contr)
                    ani=[1/1 0;0 1/(range(2,k)/range(1,k))];
                    rotsys=(rot*ani)';
                    u_relat=rotsys*u';
                    dist_sv=sqrt((u_relat(1)^2)+(u_relat(2)^2));
                    exp_semivrg=[contr(k) range(1,k)];
                    semivrg_model=semivrg_model+...
                        fnc_expmodel(0,dist_sv,exp_semivrg);
                end
                C_Limit(NumPairs)=semivrg_model*c2x;
                if SemiVarg(NumPairs)<C_Limit(NumPairs)
                    C_Flag(NumPairs)=0; %data pair is within the range
                else
                    C_Flag(NumPairs)=1; %data pair is an outlier
                end
            end
        end
    end
end

  idx_in=find(C_Flag==0);     %index of accepted data pairs
  idx_out=find(C_Flag==1);    %index of outlier data pairs
```

Figure A-4: Third part of MATLAB experimental semivariogram cleaning script

A.3. Program for conditional distribution fitting

The program for calculating the extra dimensions of a dataset in normal score units that behave more stationary is **Covariance_FittingM1**. The output file of the program consists of the initial data file plus the extra dimensions in GsLib format. In the program the extra dimension are approximated using intervals, they are specified in the parameter file (see Figure A-5).

```
                  Conditional Distribution Fitting
                  ********************************

[INPUT-OUTPUT]
<inpfile>                                          *Input Datafile in normal score scale
<outfile>                                          *Output Datafile
<rptfile>                                          *Report File

[RUN-OPTIONS]
<colx> <coly> <colz> <colv>                        *Col Ids for x,y,z coordinates, NS variable
<pint> <dinc>                                      *Prob. interval and distance increment
<nvar> <neff>                                      *nst, nugget effect
<vrtp> <incr> <ang1> <ang2> <ang3>                 *it,cc,ang1,ang2,ang3
              <rng1> <rng2> <rng3>                 *a_hmax, a_hmin, a_vert
```

Figure A-5: Parameters of **Covariance_FittingM1**

The parameter file is explained below:

- **inpfile:** input data file in GsLib format
- **outfile:** output data file according to **outopt** option
- **rptfile:** program execution report file
- **colx, coly, colz:** column ids in the input data file for *x*, *y* and *z* coordinates
- **colv:** column id for the variable in normal score units to calculate the experimental semivariogram
- **pint:** is the confidence interval in decimal notation
- **dinc:** extra dimensions distance increment. The extra dimension increases in these intervals.
- **nvar, neff:** are the number of semivariogram structures and nugget effect of the semivariogram model
- **vrtp, incr, ang1, ang2, ang3:** are semivariogram type, contribution and angular definitions of traditional GsLib programs
- **rng1, rng2, rng3:** ranges of major, minor and vertical direction

A.4. Program for sequential Gaussian simulation using high dimensional data

The GsLib program SGSim has been modified in order to read the extra dimensional maps and the extra dimensions of the conditioning data. Four extra columns were added to the parameter file (see Figure A-6). The output file consists of a gridded configuration of nodes that represent the domain.

The four extra lines in the parameter file (lines 7 - 10) are explained below:

- **nxtr:** number of extra dimensions in the conditioning dataset calculated by **Covariance_FittingM1**
- **xtf(***i***):** *i*-th extra dimension column id in the input data file
- **xtrfile:** gridded map of the domain with the extra dimensions, there should be one column for each extra dimension
- **xtd**(*i*)**:** *i*-th extra dimension column id in the map with extra dimensions of the domain

```
                  Parameters for SGSIM_MD
                  ***********************

START OF PARAMETERS:
../01-data/covariance_fittingm1.out      -file with data
1  2  0  8  0  0                -  columns for X,Y,Z,vr,wt,sec.var
<nxtr>                           -  num. of extra dimensions
<xtf1> <xtf2> ... <xtfn>         -  column ids of extra dimensions in data file
<xtrfile>                        -file with dimension 1
<xtd1> <xtd2> ... <xtdn>         -column ids with extra dimensions
-1.0e21    1.0e21                -  trimming limits
0                                -transform the data (0=no, 1=yes)
sgsim.trn                        -  file for output trans table
0                                -  consider ref. dist (0=no, 1=yes)
histsmth.out                     -  file with ref. dist distribution
1  2                             -  columns for vr and wt
0.0    15.0                      -  zmin,zmax(tail extrapolation)
1       0.0                      -  lower tail option, parameter
1      15.0                      -  upper tail option, parameter
1                                -debugging level: 0,1,2,3
sgsim.dbg                        -file for debugging output
../03-output/d_sgsim.out         -file for simulation output
100                              -number of realizations to generate
275   0.01   0.02                -nx,xmn,xsiz
275   0.01   0.02                -ny,ymn,ysiz
1     0.00   0.50                -nz,zmn,zsiz
69069                            -random number seed
0     8                          -min and max original data for sim
12                               -number of simulated nodes to use
1                                -assign data to nodes (0=no, 1=yes)
0     3                          -multiple grid search (0=no, 1=yes),num
0                                -maximum data per octant (0=not used)
6.00  6.00  6.00                 -maximum search radii (hmax,hmin,vert)
 0.0   0.0   0.0                 -angles for search ellipsoid
51    51    11                  -size of covariance lookup table
0     0.60   1.0                 -ktype: 0=SK,1=OK,2=LVM,3=EXDR,4=COLC
../data/ydata.dat                -  file with LVM, EXDR, or COLC variable
4                                -  column for secondary variable
2   0.00                                        -nst, nugget effect
2   0.60 54.0   0.0   0.0                       -it,cc,ang1,ang2,ang3
           0.6   0.6   0.6                       -a_hmax, a_hmin, a_vert
2   0.40 54.0   0.0   0.0                       -it,cc,ang1,ang2,ang3
           5.0   0.6   5.0                       -a_hmax, a_hmin, a_vert
```

Figure A-6: Parameters of **SGSIM_MD**