Independent Factor Simulation for Improved Multivariate Geostatistics

by

Felipe Cabral Pinto

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Mining Engineering

Department of Civil and Environmental Engineering

University of Alberta

Abstract

Multivariate techniques aim to integrate multiple variables and/or data in the same framework to improve uncertainty assessment in high resolution geostatistical models. The necessity to build models that better quantify the uncertainty with limited data that are often collected with different data types has driven the latest developments in multivariate geostatistical modeling. The use of decorrelation techniques facilitates the modeling of equally sampled data, whereas cokriging is more suitable for unequally sampled data and in cases where different data types are considered. The emphasis of this thesis is on the challenges and gaps in current multivariate modeling workflows involving multivariate criteria and geostatistical cosimulation with cokriging. This thesis develops techniques that improve multivariate models of equally and unequally sampled data.

The first contribution of this thesis is on multivariate modeling of equally sampled data. An integrated framework that uses the projection pursuit multivariate transform in the context of estimation and local uncertainty assessment is proposed. Simulation of the independent factors is skipped and the local multivariate distributions are directly back transformed for posterior uncertainty assessement. This framework provides a starting point for modeling more complicated multifactor and extreme value criteria. The applicability of the proposed methodology is shown in the context of exploration geochemistry with geochemical data collected in the Northwest Territories.

The second contribution of the thesis is the development of a methodology that combines the LMC and blind source separation theory for addressing the complexity and limitations of multivariate geostatistical workflows with the LMC and cokriging. The proposed methodology allows for independent simulation of the LMC factors with the most appropriate algorithm, improving variogram reproduction and facilitating model checking. As a consequence, this methodology offers a modern approach to the LMC that increases its applicability in geostatistical multivariate modeling. This methodology is applied in a multivariate modeling of geochemical data.

Because the LMC factors have a single spatial covariance function, the most appropriate Gaussian simulation algorithm may be selected and applied to each factor independently. A third contribution of this thesis is to address the challenges of optimal selection of the simulation algorithm and provide practical recommendations based on different analyses with four common Gaussian simulation algorithms.

Acknowledgements

I would like to thank my supervisor, Dr. Clayton Deutsch, for being an endless source of ideas, for his enthusiasm, for always being in a good mood, and for his invaluable advice in my research. Clayton is an example of dedication and professionalism for me.

This work, and my studies, would not be possible without the support from sponsors of the Centre for Computational Geostatistics. I would also like to thank the support of the Geological Survey of the Northwest Territories, in special Hendrik Falck, for providing the data for my case studies and for his advices and discussions on the local geology. Their financial support is gratefully appreciated.

A special thanks to all my colleagues of the Centre for Computational Geostatistics for sharing their knowledge, for their friendship, and for making these years memorable. And for the good times. I would like to especially thank Dr. Johnathan Manchuk for being such an important piece in the development of the Independent Factor Simulation methodology. Thank you all.

Finally, this special moment of my life would not be possible without the support, encouragement, and love of my family and friends. To Diomedes, Eliana, Daniel and Mariana, and most of all to Cintia, thank you. This thesis is dedicated to you.

Contents

List of Tables

List of Figures

1 Introduction 1

2 Theoretical background 9

3 Multivariate criteria in geostatistical modeling 39

4 Decomposition of multivariate spatial data into latent variables 59

List of Abbreviations

BSS	Blind source separation
CDF	Cumulative distribution function
CCDF	Conditional cumulative distribution function
DFT	Discrete Fourier transform
EXP	Exponential variogram structure
FFT	Fast Fourier transform
GAUS	Gaussian variogram structure
GSC	Geological survey of Canada
GSLIB	Geostatistical software library
ICA	Independent component analysis
ICP-MS	Inductively coupled plasma mass spectrometry
IFS	Independent factor simulation
INAA	Instrumental neutron activation analysis
LMC	Linear model of coregionalization
MA	Moving average
MAF	Minimum/maximum autocorrelation factors
MG	Multi-Gaussian
MV	Multivariate
NPV	Net present value
NS	Normal scores
NTGS	Northwest Territories Geological Survey
PCA	Principal component analysis
PostMG	Post processing of multi-Gaussian kriging
PostPPMT	Post processing of the projection pursuit multivariate transformation
PPMT	Projection pursuit multivariate transform
RF	Random function

RV	Random variable
SCT	Stepwise conditional transform
SEDEX	Sedimentary exhalative deposits
SGS	Sequential Gaussian simulation
Spectral	Spectral simulation
SPH	Spherical variogram structure
SVD	Singular value decomposition
TB	Turning bands
VMS	Volcanogenic massive sulfide
VRD	Variogram range and domain size ratio

Ag	Silver
As	Arsenic
Au	Gold
Ba	Barium
Cu	Copper
Fe	Iron
Pb	Lead
Sc	Scandium
SiO_2	Silica
Ta	Tantalum
Ti	Titanium
Tl	Thallium
V	Vanadium
Zn	Zinc
Zr	Zirconium

Chapter 1

Introduction

This chapter introduces the challenges of multivariate modeling with current techniques that motivate the research documented in this dissertation. Background on the problem setting is provided. The research contributions are reviewed with a thesis statement. The outline of the thesis and a brief summary of each chapter are also provided.

1.1 Problem setting

Evaluation of subsurface resources is a critical task in mining. The complexity of the many geological events that formed a mineral deposit cannot be fully understood or explained by one variable or a single data type. Characterization of mineral deposits is done with the available data that come from different sources and represent different scales. The collected data are used to build numerical representations of a mineral deposit that in the presence of limited data, will always be uncertain. Our understanding of the uncertainty is improved by maximizing the use of all available data.

Geostatistics provides the tools to model spatially correlated data and evaluate recoverable reserves in mineral deposits since the development and formalization of its theory in the 1960s (Matheron, 1962). The continuing development of geostatistics over the 1970s (David, 1977; Journel and Huijbregts, 1978; Matheron, 1971) and 1980s (Isaaks and Srivastava, 1989) expanded its use from interpolation of values at unsampled locations to model the uncertainty about unknown values of the attribute of interest. The release of the geostatistical software library GSLIB (Deutsch and Journel, 1992) in the early 1990s had significant impact in spreading geostatistics among practitioners and the development of new algorithms and techniques in recent years.

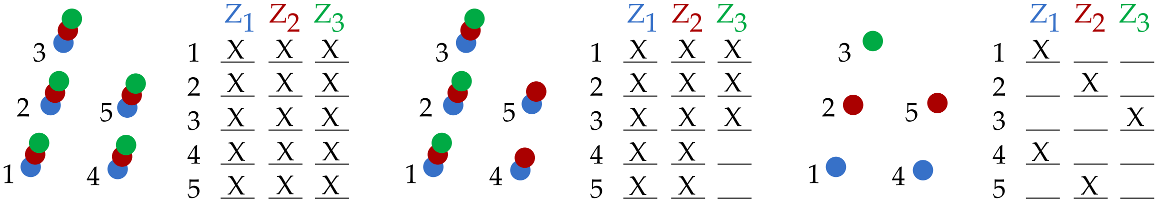
Cokriging is an early multivariate technique in geostatistics (Goovaerts, 1997; Journel and Huijbregts, 1978). Cokriging provides a framework to combine primary and non-exhaustive secondary information for estimation and requires a linear model of coregionalization (LMC) (Isaaks and Srivastava, 1989; Journel and Huijbregts, 1978). The LMC is a tool for modeling the direct and cross covariances of two or more variables. Since its introduction, multivariate geostatistical modeling evolved from the LMC and cokriging for estimation to simpler models of coregionalization and more general use of cosimulation (Wackernagel, 2003). While multivariate modeling allows integration and use of more data, it comes with challenges. Data from different sources are rarely collected at the same locations. Different data types, such as drill core, blast hole, channel samples, and geophysics are measured differently and represent different scales of the data. The integration of all data types for cosimulation with cokriging is challenging. A methodology to facilitate geostatistical workflows with the LMC is currently not available. Another problem in multivariate geostatistical simulation workflows is to choose the appropriate simulation algorithm. The current approach is to choose an algorithm for simulation. The choice of the algorithm is usually left to judgement and experience. However, different algorithms are more efficient at generating realizations for different spatial covariance functions. A methodology that permits independent conditional simulation of LMC factors is currently not available.

While multivariate modeling of unequally sampled data is restricted to the LMC, collocated data is often modelled with decorrelation techniques. Decorrelation techniques such as principal component analysis (PCA) and projection pursuit multivariate transform (PPMT) transform multivariate data with arbitrarily complex behavior to be multivariate Gaussian and uncorrelated (Barnett et al., 2014; Boisvert et al., 2013; Davis and Greenes, 1983; Hotelling, 1933). Simulation and estimation are greatly facilitated since the transformed factors can be considered one at a time. The back transformation restores any complex multivariate behavior. Closure property or hard constraints present in the data (e.g. $Z_2(x) < Z_1(x) \forall\ x$) could be enforced by a ration transform, then PPMT. Simulation of the factors is routinely performed since that provides an assessment of local uncertainty and multilocation uncertainty. There are times when local estimates or local uncertainty measures are the goal of the study. One application is in exploration geology. Assessing the local multivariate joint-probability for a given mineral deposit signature criteria provides insights to narrow exploration targets and highlights the regions where deposits are likely to occur. A computationally efficient methodology that provides accurate multivariate data-value dependent measures of local uncertainty is currently not developed.

The problem of unequally sampled data

The main problem this thesis addresses refers to the many challenges of modeling multivariate data in the presence of unequally sampled data. In geostatistics, heterotopic and homotopic observations refer to unequal and equal sampling respectively, see Figure 1.1. The term unequal refers to incomplete observations of different variables of the same data type, e.g., variable 1 and variable 2 are available at all locations, but variable 3 is missing at some (Figure 1.1b). Unequal sampling may also refer to observations of two or more data types that are available at different locations and never collocated, e.g., reverse circulation and diamond core drilling (Figure 1.1c). Unequal sampling commonly occurs as a result of (1) legacy data, (2) use of exploration data such as delineation drilling and geophysics, (3) different data types being used at different stages of a mineral project to reduce sample errors and improve confidence in the estimates, and (4) different data types collected to represent different scales of variability and/or being used in geometallurgical and geotechnical analyses.

In the presence of missing data, unequal sampling, or different data types, variables are typically cosimulated with some variation of cokriging. Simultaneous modeling of multiple primary variables requires a model of coregionalization. A model of coregionalization like the LMC is required to combine multiple data types measured at different locations and different data support into the same framework,



	Z_1	Z_2	Z_3
1	X	X	X
2	X	X	X
3	X	X	X
4	X	X	X
5	X	X	X

(**a**) Homotopic data with equal sampling

	Z_1	Z_2	Z_3
1	X	X	X
2	X	X	X
3	X	X	X
4	X	X	
5	X	X	

(**b**) Heterotopic data with unequal sampling

	Z_1	Z_2	Z_3
1	X		
2		X	
3			X
4	X		
5		X	

(**c**) Heterotopic data with completely unequal sampling

Figure 1.1: Schematic illustration of heterotopic and homotopic data. A trivariate dataset composed of Z_1, Z_2, and Z_3 are unequally sampled in five locations in (**a**), completely unequally sampled in (**b**), and equally sampled in (**c**).

and in cases where decorrelation techniques do not successfully remove spatial cross covariance. Different variables and data types having sample-specific measurement errors can be used together for estimation with cokriging (Goovaerts, 1997). In the case of completely unequally sampled data (Figure 1.1c), that is, when two or more data types are available and are never collocated, simultaneous modeling of the variables is restricted to the LMC and other models of coregionalization (Liang and Marcotte, 2015; Marcotte, 2012; Paciorek and Schervish, 2006).

Cosimulation with cokriging imposes challenges in geostatistical modeling workflows. Cokriging requires computation of the spatial structure of the direct and cross relationships of all variables. In presence of unequally sampled data, it is necessary to compute $K(K+1)/2$ direct and cross covariances, where K is the number of variables considered. An LMC is assumed and fitted to the computed covariances. The calculation of the experimental covariances takes little computational effort, but fitting an LMC with a large number of variables is challenging. Geological data is often anisotropic, that is, different directions show different spatial continuity. The problem lies in fitting the LMC to the calculated covariances accounting for anisotropy in the data. For example, fitting an LMC to a multivariate data with $k = 5$ variables requires modeling 15 covariances in each anisotropic direction. The covariances models are fitted together in the LMC. The matrix of covariance models must be mathematically valid and the variance of each variables non-negative (Rossi and Deutsch, 2014). Despite the challenge of modeling and LMC with a large number of variables, it remains a useful and mathematically flexible tool.

Simulation algorithms

There are many algorithms for simulation of Gaussian variables including sequential Gaussian simulation, turning bands, moving average methods, random coins, spectral methods, circulant embedding, and matrix methods such as Cholesky factorization and singular value decomposition (Borgman et al., 1984; Chiles and Delfiner, 2012; Dietrich and Newsam, 1997; Emery, 2008; Emery and Lantuejoul, 2006; Goovaerts, 1997; Kyriakidis, 1999; Legchenko et al., 2017; Mantoglou, 1987; Mantoglou and Wilson, 1982; Matheron, 1973; Oliver, 1995; Oliver et al., 2008; Paravarzar et al., 2015; Pardo-Iguzquiza and Chica-Olmo, 1993; Wackernagel, 2003; Yao, 1998b). Each algorithm has a range of spatial covariance functions and grid parameters where they perform with high efficiency and robustness in terms of variogram and histogram reproduction. The current approach in cosimulating with the LMC is to choose an algorithm to simulate all the structures simultaneously. The use of one algorithm may not be optimal. Good variogram and histogram reproduction may not be achieved.

Direct assessement of local multivariate distributions

Another problem this thesis addresses relates to assessing the probability of meeting multivariate criteria. These criteria may involve different rules being applied to many variables at the same time. For example, consider the probability of satisfying the following multivariate rules: variable 1 above a threshold while variable 2 is above a different threshold and variable 3 is below another threshold. One of the applications is in exploration geochemistry. Multivariate criteria are defined based on deposit signatures, for example, Pb/Zn SEDEX deposits usually contains high concentrations of Zn and Pb, while concentrations of other economic elements such as Ag, Au, and Cu vary from low to high (Jebrak and Marcoux, 2008). The assessment of multivariate criteria requires multivariate simulation. Stable assessment of local uncertainty may require hundreds of realizations; a mere one hundred realization would lead to significant noise in the variance or any probability sensitive to the tails of the distribution. Exploration geology deals with a large number of trace elements often collected at the same locations. In the presence of collocated data, Figure 1.1a, decorrelation methods offer a more practical workflow for simulation than cosimulation with cokriging. The application of multivariate modeling in the context of such criteria is developed.

Research contributions and thesis statement

To address the first two challenges, a framework for the spatial modeling of unequal sampling data is proposed in this thesis. This framework combines the LMC and Blind Source Separation (BSS) (Schmidt, 2009) to factorize the multivariate data allowing for missing geological data imputation and direct simulation of the factors. Computation of the original variables from the simulated factors is straightforward. The factors have their own spatial structure and can be modelled independently with the optimal algorithm for each structure. Another convenience is that factors can be analysed and checked independently, as opposed to cosimulation with cokriging. The proposed methodology is referred as to independent factor simulation (IFS).

To address the third challenge, this thesis develops the use of the PPMT (Barnett et al., 2014) in the context of estimation and local uncertainty assessment. Simulation of the independent factors is skipped and the local multivariate distributions are directly back transformed for posterior uncertainty assessement. The proposed methodology is referred as to post-process of PPMT factors (PostPPMT).

The key contributions of this thesis are:

- The development of a framework for addressing the problems of the complexity and limitations of multivariate geostatistical workflows with the LMC and cokriging for a large number of variables.
- A modern approach to the LMC that increases its applicability in geostatistical multivariate modeling.

- The proposed methodology allows for independent simulation of the LMC factors with the most appropriate algorithm, improving variogram reproduction and facilitating model checking.

- The development of a framework for estimation and local uncertainty assessement with multivariate data that provides a starting point for more complicated multifactor and extreme value criteria.

The thesis statement:

> The development of geostatistical modeling with multivariate complex relationships of unequally sampled data modernizes the use of the LMC and leads to improved high resolution geostatistical property models. The practical implementation of probabilistic assessment of uncertainty with multivariate criteria provides ways for multivariate modeling in the context of such criteria and adds practical value and theoretical insight for more complicated multifactor criteria.

At present, multivariate modeling has been a collection of methods and computer programs developed to tackle specific problems in geostatistics. Even though many of these methods have proven their values, many implementation details are missing. In this thesis, the emphasis is on the challenges and gaps in current multivariate geostatistical workflows.

1.2 Thesis outline

Chapter 2 reviews relevant literature and provides the background that motivates the development of this thesis. A summary of the geostatistical theory for multivariate modeling is provided. The focus is on multivariate techniques that use the LMC for cokriging and cosimulation, a brief review of multivariate modelling with decorrelation techniques is also given.

Chapter 3 brings a discussion on multivariate criteria and develops the theory of the PostPPMT methodology. A brief case study on different deposit signatures found in the Northwest Territories, Canada, demonstrates the practical implementation of the methodology.

Chapter 4 starts with a discussion on the limitations of current practices in multivariate modelling with unequal sampling. It introduces a framework for integrating different data types and unequal sampling data within geostatistical modelling workflows. The remainder of the chapter develops the theory of BSS and explains in details how the BSS theory is used for imputation of factor data. When combined with the LMC, BSS provides a framework to impute factor data that reproduce the original data and have the correct spatial structure. This chapter also addresses to the practical aspects, implementation details, and limitations of the BSS theory. A small numerical example demonstrates the step-by-step of the BSS theory.

Chapter 5 brings a discussion on the practical aspects and best practices of selection of simulation algorithm. Because factors are independent they can be simulated independently with the best algorithm for each structure. The process of generating realizations of the factors and computing the original variables is highly parallelizable, not only across realizations, but also across factors and variables. Practical recommendations for algorithm selection are given.

Chapter 6 demonstrates the BSS methodology in a comprehensive case study with geochemical data. The first part of the case study introduces the multivariate data, the fitted LMC model, and the imputation of factor data at sample locations. The second part discusses the simulation of factors and reconstruction of the original variables. Modeling results are compared with that of established techniques.

Chapter 7 summarizes the contributions and results of the developed methodologies. The limitations and proposed future work are highlighted. Each of the methods developed in this thesis are implemented in software. Computationally intensive algorithms are constructed as stand-alone FORTRAN executables in standard GSLIB format and wrapped in Python for easy integration into workflows. These programs are used to generate all results in this thesis and are available from the author upon request. Software description is given in the Appendix.

Chapter 2

Theoretical background

This chapter reviews the geostatistical concepts that are relevant to the development of the proposed methodologies. The independent simulation of the LMC factors calls for a review of the LMC, cokriging, and the normal score transformation. The LMC factors may be simulated with different algorithms; therefore, the most well established Gaussian algorithms in geostatistics are reviewed. The PostPPMT methodology is based on the framework of multi-Gaussian kriging, therefore a review of kriging and decorrelation methods are provided.

2.1 Essentials of Geostatistics

The theory of regionalized variables developed by Georges Matheron (Matheron, 1971) is the foundation of geostatistics. It uses the concepts of random variables (RVs), random functions (RFs), and probability theory to model spatially dependent data. Random functions are an ensemble of spatially related random variables, the inference of the moments of a random function requires stationarity.

Random variables and functions

A random variable is used to quantify outcomes of random processes according to some probability distribution (Goovaerts, 1997). Random variables are either categorical or continuous. A categorical RV has a finite number of outcomes, e.g., facies in a reservoir or rock types in a mineral deposit. Variables with a continuous range of values, such as porosity or mineral grade, are modeled by a continuous RV. Random variables are commonly represented by a capital letter $Z(\mathbf{u})$, while its outcome values are denoted with the corresponding lower-case letter $z(\mathbf{u})$. The location coordinates vector $\mathbf{u}$ is required since the RV model of Z and its probability distribution are location-dependent (Deutsch and Journel, 1992).

The cumulative distribution function (CDF) of $Z(\mathbf{u})$, $F(\mathbf{u};z) = \text{Prob}\,\{Z(\mathbf{u}) \leq z\}$, defines the prior model of uncertainty about the unsampled value $z(\mathbf{u})$. The prior model of uncertainty does not account for the information at neighboring locations $\mathbf{u}^{'}$. In geostatistics this information is usually available from n neighboring non-exhaustive data values $Z(\mathbf{u}_{\alpha}) = z(\mathbf{u}_{\alpha}),\ \alpha = 1,...,n$. The n data values are used to define the conditional distribution function (CCDF) $F(\mathbf{u};z \mid (n)) = \text{Prob}\,\{Z(\mathbf{u}) \leq z \mid (n)\}$ that defines the posterior uncertainty at $z(\mathbf{u})$. This posterior model of uncertainty is the goal of geostatistical modeling and the concept of RF allows such modeling.

A random function, also denoted by $Z(\mathbf{u})$, is a finite set of RVs related to the same attribute z defined within a field of study A, $\{Z(\mathbf{u}),\ \mathbf{u} \in A\}$. A different RF $Y(\mathbf{u})$ is defined to model the uncertainty of the attribute y that may or may not be related to z. In the univariate case, the CDF of the RV $Z(\mathbf{u})$ is used to model the uncertainty about $z(\mathbf{u})$. In the multivariate case, the set of the CDFs of any number K of variables, $k = 1,...,K$, is used to model the joint uncertainty about the K values $z_1(\mathbf{u}),...,z_k(\mathbf{u})$. Inference of the moments of a k-variate CDF requires stationarity.

The decision of stationarity

Inference of any statistics and moments (mean, variance, covariance) of a multivariate CDF requires repetitive sampling at each location $\mathbf{u}$. Exhaustive samples of $Z(\mathbf{u})$ are rarely available in practice and multiple samples at the same location are never available. Stationarity is the decision to pool data together within a domain A to allow for such inference. The decision of stationarity is required to spatially homogenize the attribute under study and allow for statistical inference of $Z(\mathbf{u})$ from samples collected at other locations, $\mathbf{u}_{\alpha} \neq \mathbf{u},\ \alpha = 1,...,n$, within A. Two RVs $Z(\mathbf{u})$ and $Z(\mathbf{u}^{'})$ separated by a vector $\mathbf{h} = \mathbf{u}^{'} - \mathbf{u}$ within a stationary domain A are assumed to have the same multivariate CDF under any translation vector $\mathbf{h}$,

$F(\mathbf{u}_1, ..., \mathbf{u}_n; z_1, ..., z_k) = F(\mathbf{u}_1 + \mathbf{h}, ..., \mathbf{u}_n + \mathbf{h}; z_1, ..., z_k),\ \forall\ \mathbf{h}$ (Deutsch and Journel, 1992; Goovaerts, 1997).

A multivariate RF $\{Z_k(\mathbf{u}),\ k = 1, ..., K\ ;\ \forall\ \mathbf{u} \in A\}$ is said to be stationary of order two if the expected value of each interdependent RF exists and is constant within A, and the covariance and variogram functions exist and depend only on the vector $\mathbf{h}$:

$$m_k = E\left\{Z_k(\mathbf{u})\right\} \qquad k = 1, ..., K \tag{2.1}$$

$$C_{ij}(\mathbf{h}) = E\left\{[Z_i(\mathbf{u}) - m_i] \cdot [Z_j(\mathbf{u}+\mathbf{h}) - m_j]\right\} \qquad \forall\, i, j \in \{1, ..., K\} \tag{2.2}$$

$$\begin{aligned} 2\gamma_{ij}(\mathbf{h}) &= Cov\left\{[Z_i(\mathbf{u}) - Z_i(\mathbf{u}+\mathbf{h})], [Z_j(\mathbf{u}) - Z_j(\mathbf{u}+\mathbf{h})]\right\} \\ &= E\left\{[Z_i(\mathbf{u}) - Z_i(\mathbf{u}+\mathbf{h})] \cdot [Z_j(\mathbf{u}) - Z_j(\mathbf{u}+\mathbf{h})]\right\} \end{aligned} \qquad \forall\, i, j \in \{1, ..., K\} \tag{2.3}$$

When $i = j$ the terms auto or direct are applied, whereas the terms joint or cross are used if $i \neq j$. The semivariogram $\gamma(\mathbf{h})$ has been historically utilized in the place of the variogram $2\gamma(\mathbf{h})$ and is hereafter referred as to as the variogram. Under the assumption of second-order stationarity, the covariance, variogram, and correlogram are related by ($\forall\, i, j \in \{1, ..., K\}$):

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(0) - C_{i,j}(\mathbf{h}) \tag{2.4}$$

$$\rho_{ij}(\mathbf{h}) = \frac{C_{ij}(\mathbf{h})}{\sqrt{C_{ii}(0) \cdot C_{jj}(0)}} \tag{2.5}$$

Some important properties of these moments are defined (Goovaerts, 1997). At $\mid \mathbf{h} \mid = 0$ the correlogram defines the linear correlation between variables. In general the covariance function is not symmetric. In practice such asymmetry is often ignored and the so-called lag effect is rarely modeled. Although a lag effect results in a directional asymmetry, there can be other causes for asymmetries. For example, depositional processes of dunes form asymmetric objects and produce spatially asymmetric grain size distributions, but there is no lag effect of such. Another example, the cross-covariance between a variable and its derivative is not symmetric. It is important to note that asymmetry is not synonymous of lag effect. The variogram and covariance are then considered symmetric in $(\mathbf{h}, -\mathbf{h})$, therefore $C_{ij} = C_{ji}$, and $C_{ij}(\mathbf{h}) = C_{ij}(-\mathbf{h}) = C_{ji}(\mid \mathbf{h} \mid) \quad \forall\, i, j$. In general terms, the correlation between variables tend to zero $C(\mathbf{h}) \rightarrow 0$ and the variogram tends to the a priori stationary variance $\gamma(\mathbf{h}) \rightarrow C(0)$ as the separation distance increases $\mid \mathbf{h} \mid \rightarrow \infty$. The reason geostatisticians prefer the variogram is because it does not require a constant stationary mean and finite variance for the RF $Z(\mathbf{u})$. The variogram only requires that the RF increments $[Z(\mathbf{u}) - Z(\mathbf{u} + \mathbf{h})]$ are stationary of order two (Journel and Huijbregts, 1978).

The decision of stationarity is made prior to calculating relevant statistics in an estimation domain and it is subject to data availability and geological understanding. It is neither a characteristic of the variables being modeled nor a property of the RF, therefore it cannot be checked. Because the decision of pooling data together is one of the first steps of a geostatistical workflow, further steps such as exploratory data analysis and experimental variograms may indicate a lack of homogeneity in the data population. The decision of stationarity may then be reviewed. Variograms are calculated and modeled within stationary domains. They are required for estimation and simulation.

2.2 Spatial variability

Variogram models provide a measure of spatial variability, or continuity, for geostatistical modeling. Conceptual variogram models can be inferred from geology if geological and mineralogical factors are known. Traditionally, direct and cross experimental variograms are calculated from the available data with Equation 2.6 and reconciled with known geology.

$$\gamma_{ij}(\mathbf{h}) = \frac{1}{2 \mid N(\mathbf{h}) \mid} \sum_{\alpha=1}^{N(\mathbf{h})} \left[z_i(\mathbf{u}_\alpha) - z_i(\mathbf{u}_\alpha + \mathbf{h})\right] \left[z_j(\mathbf{u}_\alpha) - z_j(\mathbf{u}_\alpha + \mathbf{h})\right] \qquad \forall i, j \in \{1, ..., K\} \tag{2.6}$$

where $N(\mathbf{h})$ is the number of pairs of data locations a vector $\mathbf{h}$ apart. Experimental variograms are calculated in different directions because geological variability is often anisotropic. Anisotropy is modeled with three main directions orthogonal to each other and their respective ranges of continuity. The major direction is the direction of greatest continuity, the minor direction is the direction of smallest continuity, and the semi direction is the direction orthogonal to the first two directions. Selective sampling often occurs in mining and the calculation of the variogram must account for it. Parameters such as the azimuth and dip, number of lags, lag distance and tolerance can be adjusted to guarantee a reasonable number of data pairs are used to calculate the variogram in the presence of sparse irregular sampling.

Analytical continuous functions are fitted to the experimental points to allow for interpolation of variogram values for any possible lag $\mathbf{h}$ and to smooth out sample fluctuations. These functions also ensure positive-definite condition of the covariance values and non-negativity of the variance of any linear combination of random variables. Such condition guarantees that the kriging system of equations have a solution and such solution is unique. The four most frequently used basic models in geostatistics are the spherical (Equation 2.7), exponential (Equation 2.8), Gaussian (Equation 2.9), and the nugget effect (Equation 2.10). There are other models available in the literature (Gneiting, 1999b; Matern, 1986; Rasmussen and Williams, 2006; Wackernagel, 2003; Webster and Oliver, 2007b) and their uses depend on the area of study, such as geology, environmental, hydrogeology, etc. Most experimental variograms in mining and geology can be fit with these basic functions.

$$Sph(h) = \begin{cases} 1.5h - 0.5h^3 & \text{if } h \leq 1 \\ 1 & \text{otherwise} \end{cases} \tag{2.7}$$

$$Exp(h) = 1 - exp(-3h) \tag{2.8}$$

$$Gaus(h) = 1 - exp(-3h^2) \tag{2.9}$$

$$Nug(h) = \begin{cases} 0 & \text{if } h = 0 \\ 1 & \text{otherwise} \end{cases} \tag{2.10}$$

These models are expressed in their isotropic form and are said to be permissible in three dimensions. The scalar $h =\mid \mathbf{h} \mid$ is the normalized distance calculated from Equation 2.11 to account for the anisotropy and allow inferring the variogram value for any direction and any distance.

$$h = \sqrt{\left(\frac{h_{major}}{a_{major}}\right)^2 + \left(\frac{h_{minor}}{a_{minor}}\right)^2 + \left(\frac{h_{semi}}{a_{semi}}\right)^2} \tag{2.11}$$

where a is the range parameter or the distance the variogram reaches the stationary variance (sill), here set to 1 (standardized models). The nugget effect reaches the sill as soon as $h > 0$. The spherical model reaches the sill at distance a, whereas the exponential and Gaussian models reach their sill asymptotically and therefore have their practical range modeled at 95% of the sill. The three continuous variogram models are illustrated in Figure 2.1. Note the different behavior near the origin of the different models. The Gaussian is highly continuous whereas the spherical and exponential models are linear and steep.

Variograms are modeled using nested structures, that is, the variogram is fit with a sum of valid variogram models. Nested structures allow for different ranges and anisotropy to be modeled together. Each structure can be modeled independent of the other explaining part of the total variance. Note that the variance of all structures must sum up to the stationary variance or to 1 if variograms are standardized.

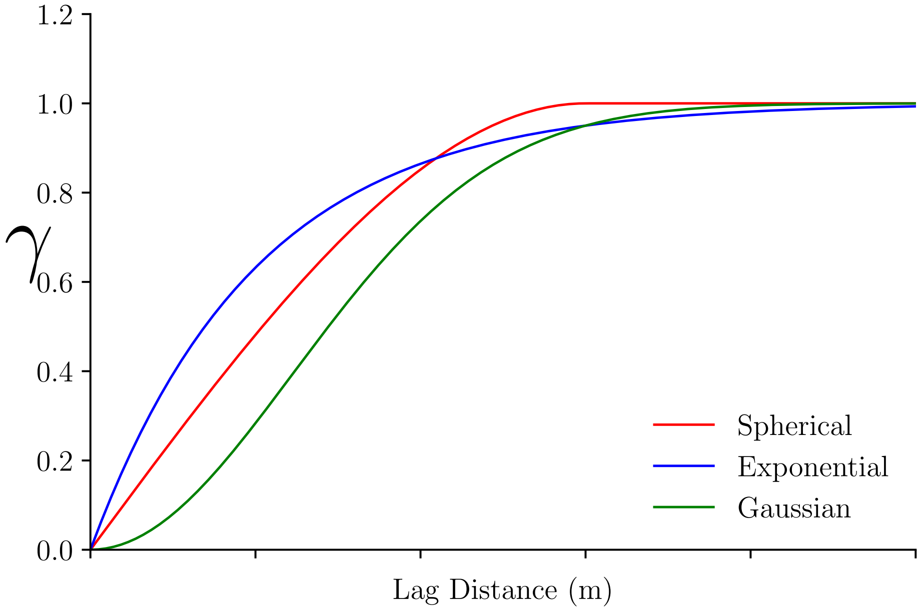


Figure 2.1: The most frequently used variogram models and their shapes. A standardized sill of 1 is used in the illustration.

The variogram tends to zero when the lag distance tends to zero $\gamma(0) = 0$, however an apparent discontinuity at the origin of the variogram may occur. The nugget effect is this discontinuous behavior and is isotropic by definition. It relates to measurement errors and spatial variations at short-scale, usually at distances shorter than the sampling interval (Goovaerts, 1997; Journel and Huijbregts, 1978).

In multivariate cases, the coregionalization of two or more RFs $\Big\{Z_k(\mathbf{u}), k = 1, ..., K \;\; ; \;\; \forall \;\; \mathbf{u} \in A\Big\}$ requires a joint model for the covariance function matrix. This covariance matrix contains the $K(K+1)/2$ direct and cross relationships required for cokriging. The nested structures of the direct and cross covariances cannot be modeled independently from each other. All direct and cross covariances must share the same set of basic structures. Similar to the univariate case, each basic structure must be positive definite, and the contribution or sill of each structure must be positive. To ensure the positive definiteness of the covariance matrix a permissible model such as the LMC must be used to fit the covariances. In the presence of equally sampled data the variograms are most commonly calculated and expression 2.4 is used to convert variogram to covariance values. Unequally sampled data requires the direct calculation of the covariances. The LMC is reviewed thoroughly in the next section due to its importance in the development of the BSS methodology.

Linear Model of Coregionalization

Consider K coregionalized variables at locations $\mathbf{u}$ in a stationary domain A denoted $Z_k(\mathbf{u}), k = 1, ..., K$ $\mathbf{u}$ in A. An LMC is assumed, that is, each variable consists of a sum of independent factors and a stationary mean:

$$Z_k(\mathbf{u}) = m_k + \sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u}) \tag{2.12}$$

Where m_k is the stationary mean of the k^{th} variable, nst is the number of structures or factors, $a_{k,i}$ are the coefficients explaining the contribution of the i^{th} factor to the k^{th} variable and $Y_i(\mathbf{u})$ are independent factors defined by single spatial covariance structure, with the 0^{th} factor representing the nugget effect component. The $nst+1$ factors $Y_i(\mathbf{u})$ have zero mean and unit variance. The $a_{k,i}$ parameters are stationary parameters that are derived in practice from experimental variograms or covariances and a conceptual geological model. The mean and variance of the $Z_k(\mathbf{u})$ variables are given by:

$$\begin{aligned} E\left\{Z_k(\mathbf{u})\right\} &= E\left\{m_k + \sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u})\right\} \\ &= m_k + \sum_{i=0}^{nst} a_{k,i} E\left\{Y_i(\mathbf{u})\right\} = m_k \end{aligned} \qquad k = 1, ..., K \quad \mathbf{u} \text{ in } A \tag{2.13}$$

$$Var\left\{Z_k(\mathbf{u})\right\} = \sum_{i=0}^{nst} a_{k,i}^2 Var\left\{Y_i(\mathbf{u})\right\} = \sum_{i=0}^{nst} a_{k,i}^2 \quad k = 1, ..., K \quad \mathbf{u} \text{ in } A \tag{2.14}$$

In the context of this work, the K variables are standard variables since we could always standardize or normal score transform at the start and reverse the standardization when the models are constructed, that is:

$$m_k = 0 \quad k = 1, ..., K \tag{2.15}$$

$$Var\left\{Z_k\right\} = \sum_{i=0}^{nst} a_{k,i}^2 = 1 \quad k = 1, ..., K \tag{2.16}$$

Given that the factors are independent the direct and cross variograms are given by:

$$\gamma_{k,k'}(\mathbf{h}) = \sum_{i=0}^{nst} a_{k,i} a_{k',i} \Gamma_i(\mathbf{h}) \quad k, k^{'} = 1, ..., K \tag{2.17}$$

This model is widely used in cokriging of unequally sampled data (Chiles and Delfiner, 2012). It is especially useful for multiple data sources that are not sampled at the same location; inference of the cross variograms is done through the cross covariance and the data are combined into best estimates of the variables under consideration (Minnitt and Deutsch, 2014). This model is also widely used as a means to fit direct and cross variograms.

The correlation between Z_k and $Z_{k'}$ at $\mathbf{h} = 0$ is:

$$C_{Z_k, Z_{k'}}(0) = \rho_{Z_k, Z_{k'}}(0) = \sum_{i=1}^{nst} a_{k,i} a_{k',i} \quad k, k^{'} = 1, ..., K \tag{2.18}$$

The direct and cross covariance between variables is:

$$\begin{aligned} Cov\left\{Z_k(\mathbf{u}), Z_{k'}(\mathbf{u}^{'})\right\} &= E\left\{\sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u}) \cdot \sum_{j=1}^{nst} a_{k',j} Y_j(\mathbf{u}^{'})\right\} \\ &= \sum_{i=1}^{nst}\sum_{j=1}^{nst} a_{k,i} a_{k',j} E\left\{Y_i(\mathbf{u}) Y_j(\mathbf{u}^{'})\right\} \\ &= \sum_{i=1}^{nst} a_{k,i} a_{k',i} C_i(\mathbf{u} - \mathbf{u}^{'}) \quad \mathbf{u}, \mathbf{u}^{'} \text{ in } A \end{aligned} \tag{2.19}$$

The cross covariance between the Z and Y values is:

$$
\begin{aligned}
Cov\left\{Z_k(\mathbf{u}), Y_i(\mathbf{u}^{'})\right\} &= E\left\{\sum_{j=0}^{nst} a_{k,j} Y_j(\mathbf{u}) Y_i(\mathbf{u}^{'})\right\} \\
&= a_{k,i} E\left\{Y_i(\mathbf{u}) Y_i(\mathbf{u}^{'})\right\} \\
&= a_{k,i} C_i(\mathbf{u}-\mathbf{u}^{'}) \quad \mathbf{u}, \mathbf{u}^{'} \text{ in } A
\end{aligned}
\tag{2.20}
$$

The Y factors are independent from each other and have their own covariance that is fit from the data:

$$
Cov\left\{Y_i(\mathbf{u}), Y_i(\mathbf{u}^{'})\right\} = C_i(\mathbf{u}-\mathbf{u}^{'}) \quad i = 0, ..., nst \quad \mathbf{u}, \mathbf{u}^{'} \text{ in } A \tag{2.21}
$$

A more typical notation for the LMC found in the literature combines the outer product of the the $a_{k,i}$ parameters into matrices of c coefficients:

$$
\gamma_{k,k^{'}}(\mathbf{h}) = \sum_{i=0}^{nst} c_{k,k^{'}i} \Gamma_i(\mathbf{h}) \quad k, k^{'} = 1, ..., K \tag{2.22}
$$

where each of the $i = 0, ..., nst$ k by k matrices of c coefficients must be positive definite (Goovaerts, 1997; Rossi and Deutsch, 2014). Although both notation styles are interchangeable, fitting algorithms that are used to derive coefficients typically yield the c matrices, rather than the a vectors. The latter is more convenient for the developments in this work.

2.3 Estimation

Kriging and cokriging are briefly reviewed in this section. In the PostPPMT methodology the simple kriging estimate and variance of the independent factors identify the mean and variance of the conditional distribution. Since the factors are Gaussian, the uncertainty at an estimated location is fully determined by these two parameters. Such estimation framework is the basis of multi-Gaussian kriging and sequential simulation. In the BSS methodology, simple cokriging provides a minimum norm solution to Equation 2.12.

Kriging

Kriging is a family of least-squares linear regression algorithms that estimate an unsampled value $z(\mathbf{u})$ from neighboring data values $z(\mathbf{u}_\alpha, \alpha = 1, ..., n)$ (Chiles and Delfiner, 2012; Krige, 1951). The unknown value $z(\mathbf{u})$ and the data values $z(\mathbf{u}_\alpha)$ are realizations of the RVs $Z(\mathbf{u})$ and $Z(\mathbf{u}_\alpha)$. All members of the kriging family aim to minimize the estimation error of the random variable $Z^*(\mathbf{u}) - Z(\mathbf{u})$. This error is also referred as to as kriging or estimation variance and is defined as $\sigma^2_E(\mathbf{u}) = Var\left\{Z^*(\mathbf{u}) - Z(\mathbf{u})\right\}$. Kriging is an exact interpolator that aims to minimizes σ^2 under the constrain of unbiasedness $E\left\{Z^*(\mathbf{u}) - Z(\mathbf{u})\right\} = 0$, therefore it honors data values at their locations $Z^*(\mathbf{u}) = Z(\mathbf{u}_\alpha) \quad \forall\, \mathbf{u} = \mathbf{u}_\alpha,\ \alpha = 1, ..., n$.

In its simplest form, kriging requires a stationary RF model $Z(\mathbf{u})$ with known mean m and covariance $C(\mathbf{h})$ (Deutsch and Journel, 1992). The simple kriging estimator $Z^*_{SK}(\mathbf{u})$ is defined as

$$Z^*_{SK}(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}) \left[Z(\mathbf{u}_\alpha) - m\right] + m \tag{2.23}$$

The number n of data used in the estimation is constrained to the data closest to the location being estimated. In practice, a moving search ellipsoid centered on u is used to limit the data being used in the estimation. This search ellipsoid must account for the anisotropy in the covariance models. The weights $\lambda_\alpha(\mathbf{u})$ assigned to each sample $z(\mathbf{u}_\alpha)$ are a function of the covariance and are determined such as to minimize the kriging unbiasedness constraint. This minimization results in the famous simple kriging system of equations (Equation 2.24), also referred as to the normal equations (Luenberger, 1969). The simple kriging variance (Equation 2.25) is derived from this system of equations.

$$\sum_{\beta=1}^{n} \lambda_\beta(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}_\beta) = C(\mathbf{u}_\alpha - \mathbf{u}) \quad \forall\, \alpha = 1, ..., n \tag{2.24}$$

$$\sigma^2_{SK}(\mathbf{u}) = C(0) - \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}) \tag{2.25}$$

The simple kriging system of equations can be easily represented using matrix notation. Equation 2.24 is written as $C_{SK}\lambda_{SK}(\mathbf{u}) = r_{SK}$, where C_{SK} is the $n \times n$ matrix of covariance between the data, $\lambda_{SK}(\mathbf{u})$ is the vector of simple kriging weights, and r_{SK} is the right hand side vector of covariance between the data and the estimation locations.

The success of kriging is explained by some important properties, such as to account for the geometry of volume being estimated, the distance of the information and configuration of the data, and the structural continuity of the variable. The kriging weights account for the closeness of the data to the location being estimated, redundancy between the data, and the covariance. The smoothness of the kriged estimates can be predicted from the kriging variance, but the kriging variance and weights do not depend on data values. For this reason, the kriging variance does not provide a measure of the uncertainty at an unsampled location.

The traditional SK estimator (Equation 2.23) expresses the estimates as a function of the data values. Such estimator can be defined in its dual form. The dual form of kriging expresses the estimates as a function of the covariance values (Dubrule, 1983; Goovaerts, 1997; Journel, 1989). The dual simple kriging estimator is defined as:

$$Z^*_{SK}(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_{\alpha}^{dual}(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}) + m \tag{2.26}$$

The dual kriging weights $\lambda_{\alpha}^{\text{dual}}(\mathbf{u})$ are derived from the exactitude property of kriging $z^*_{SK}(\mathbf{u}_\alpha) = z(\mathbf{u}_\alpha)$:

$$z^*_{SK}(\mathbf{u}_\alpha) = \sum_{\beta=1}^{n} \lambda_{\beta}^{dual}(\mathbf{u}) C(\mathbf{u}_\alpha - \mathbf{u}_\beta) + m = z(\mathbf{u}_\alpha) \quad \forall\, \alpha = 1, ..., n \tag{2.27}$$

Such dual formalism provides a more efficient way of conditioning realizations generated with Gaussian simulation algorithms (Manchuk and Deutsch, 2017). In the BSS methodology, realizations of the independent LMC factors may be conditioned with dual kriging.

The simple kriging mean and variance can be used to assess the local uncertainty at an unsampled location if used in the multi-Gaussian (MG) approach. A necessary condition is the univariate CDF of a stationary RF $Y(\mathbf{u})$ to be standard normal, i.e., $Y(\mathbf{u})$ follows a Gaussian distribution with a zero mean and unit variance (Deutsch and Journel, 1992). The normal score (NS) transformation (Barnett, 2015; Bliss, 1934; Verly, 1983) is a quantile-by-quantile transformation that converts a distribution to be standard normal:

$$y_k = G^{-1}(F_k(z_k)), \quad \forall\, k = 1, ..., K \tag{2.28}$$

where G^{-1} is the inverse of the standard univariate normal CDF. The resultant distribution is univariate normal and permit independent modeling of the CDFs with simple kriging. In the MG approach, the simple kriged mean $y_k^*(\mathbf{u})$ and variance $\sigma_k^2(\mathbf{u})$ fully define the posterior CCDF $\{G_k(\mathbf{u}), \mathbf{u} \in A\}$ at a location $\mathbf{u}$. The back-transformation to the original distribution is given by the expression:

$$z_k = F_k^{-1}(G(y_k)), \quad \forall\, k = 1, ..., K \tag{2.29}$$

In simple terms, multi-Gaussian kriging is the application of the normal equations to NS transformed variables. Simple kriging is used to estimate the conditional mean and variance at an unsampled location. This conditional distribution follows a non-standard normal distribution with mean and variance equal to the estimated simple kriging mean y_{SK}^* and variance σ_{SK}^2. A number of equally spaced quantiles $p^l, l = 1, ..., L$ of this distribution are defined and back-transformed for post processing and uncertainty assessment:

$$z^l = F^{-1}\left(G\left(\sigma_{SK} G^{-1}(p^l) + y_{SK}^*\right)\right), \quad \forall\, l = 1, ..., L \tag{2.30}$$

The MG approach can be applied to a univariate or multivariate data set. In the multivariate case, cokriging is used to build the local conditional distributions.

Cokriging

Cokriging is a member of the kriging family that uses data from different attributes, also referred as to secondary data, for estimation of the primary variable. There must be a correlation between the variables, otherwise no extra information from the secondary variables are considered and cokriging is reduced to simple kriging. Theoretically, cokriging can be applied to any set of RVs coregionalized with a valid LMC. In practice, cokriging is restricted to a few variables due to the tedious task of fitting large LMCs, the computational requirements to solve large kriging matrices, and challenges of validating the results. Moreover, if the number of data is approximately the same for both primary and secondary variables, the benefits of cokriging may not be worth the additional modeling efforts. Cokriging yields a more significant reduction in the estimation variance when there are many more secondary data than primary data (Rossi and Deutsch, 2014). Cokriging can be useful even with a few secondary data when a physical relationship relates the two variables and this physical relation is taken into account, e.g. a variable and its derivative (Chiles and Delfiner, 2012).

Consider first the case where a primary variable $\{z_1(\mathbf{u}_{\alpha_1}),\ \alpha_1 = 1, ..., n_1\}$ is used in combination with a single secondary attribute $\{z_2(\mathbf{u}_{\alpha_2}),\ \alpha_2 = 1, ..., n_2\}$. The simple cokriging estimator $Z^*_{SCK}(\mathbf{u})$ of the primary variable z_1 at location $\mathbf{u}$ is defined as

$$Z^*_{SCK}(\mathbf{u}) - m_1 = \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1}(\mathbf{u})[Z_1(\mathbf{u}_{\alpha_1}) - m_1] + \sum_{\alpha_2=1}^{n_2} \lambda_{\alpha_2}(\mathbf{u})[Z_2(\mathbf{u}_{\alpha_2}) - m_2] \quad (2.31)$$

where m_1 and m_2 are the mean of the RVs $Z_1(\mathbf{u})$ and $Z_2(\mathbf{u})$ respectively. This estimator can be extended to any set of $(K-1)$ secondary variables, and the more general simple cokriging estimator for several secondary variables defined as:

$$Z^*_{SCK}(\mathbf{u}) - m_1 = \sum_{\alpha_1=1}^{n_1} \lambda_{\alpha_1}(\mathbf{u})[Z_1(\mathbf{u}_{\alpha_1}) - m_1] + \sum_{k=2}^{K} \sum_{\alpha_i=1}^{n_i} \lambda_{\alpha_i}(\mathbf{u})[Z_k(\mathbf{u}_{\alpha_i}) - m_k] \quad (2.32)$$

The cokriging weights are a function of the direct and cross covariances and similar to simple kriging, are determined such as to minimize the kriging unbiasedness constraint and minimum estimation variance. Such minimization yields the simple cokriging system of $(\sum_{k=1}^{K} n_k)$ equations (Equation 2.33). The simple cokriging variance (Equation 2.34) is derived from the system of equations.

$$\sum_{j=1}^{K}\sum_{\beta_j=1}^{n_j}\lambda_{\beta_j}C_{ij}(\mathbf{u}_{\alpha_i}-\mathbf{u}_{\beta_j}) = C_{k1}(\mathbf{u}_{\alpha_i}-\mathbf{u}) \quad \alpha_i = 1,...,n_i;\ k = 1,...,K \tag{2.33}$$

$$\sigma^2_{SCK}(\mathbf{u}) = C_{11}(0) - \sum_{k=1}^{K}\sum_{\alpha_i=1}^{n_i}\lambda_{\alpha_i}(\mathbf{u})C_{k1}(\mathbf{u}_{\alpha_i}-\mathbf{u}) \tag{2.34}$$

Similar to simple kriging kriging, the system of equations in 2.34 can be represented using matrix notation:

$$\begin{bmatrix} \mathbf{K}_{11} & \cdots & \mathbf{K}_{1K} \\ \vdots & \ddots & \vdots \\ \mathbf{K}_{K1} & \cdots & \mathbf{K}_{KK} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1(\mathbf{u}) \\ \vdots \\ \boldsymbol{\lambda}_K(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} \mathbf{r}_{11} \\ \vdots \\ \mathbf{r}_{K1} \end{bmatrix} \tag{2.35}$$

where $\mathbf{K}_{ij}$ is the $n_i \times n_j$ submatrix of direct and cross covariances $[C_{ij}(\mathbf{u}_{\alpha_i}-\mathbf{u}_{\beta_j})]$, $\boldsymbol{\lambda}_k$ is the vector with the cokriging weights, and $\mathbf{r}_{i1} = [C_{k1}(\mathbf{u}_1-\mathbf{u}),...,C_{k1}(\mathbf{u}_{n_k}-\mathbf{u})]$ is the vector of direct and cross covariances between the data and the estimation location.

The development of the IFS methodology is discussed at length in Chapters 4 and 6 of this thesis. One of the steps in the IFS requires full cokriging of the latent factors at data locations. The factors Y are not directly observed. They are determined from the variables Z. In this aspect, the IFS methodology can be compared to a method known as factorial (co)kriging analysis (Chiles and Delfiner, 2012; Goovaerts, 1992,9; Ma et al., 2014; Matheron, 1982). The development of the IFS methodology provides a direct way to simulate the factors from the observed variables and the fitted LMC that are later used to compute the variables at other locations.

The MG approach reviewed in this section provides a framework for assessing local uncertainty but it does not provide a measure of the joint spatial uncertainty of the variables at several locations taken together. Multi-location spatial uncertainty is assessed with geostatistical simulation.

2.4 Simulation

Kriging produces estimates with less variability than the data, this is also referred as to as the smoothing effect of kriging. For this reason kriging does not reproduce the original data distribution and its spatial variability. Simulation aims at reproducing the input data statistics and the covariance models, preserving the spatial correlation inferred from the sample data and honoring the data values at their locations (Journel and Kyriakidis, 2004). As opposed to kriging, simulation provides a set of $l = 1, ..., L$ realizations of the z-values within a stationary domain $\left\{z^l(\mathbf{u}), \mathbf{u} \in A\right\}$, where l denotes the l^{th} realization. These realizations are equally likely to be drawn, resulting in a distribution of predicted system response values, reflecting the uncertainty (Gotway and Rutherford, 1994). In geostatistical workflows, transfer functions applied to the simulated model are used for risk analysis and decision making.

There are many types of simulation methods and algorithms. The ones more relevant for this work are Gaussian-based approaches such as sequential Gaussian simulation (SGS), turning bands, moving average, and spectral simulation. The process of generating geostatistical conditional realization can be divided into two steps: the generation of unconditional realizations and their posterior conditioning with (dual)kriging. The term unconditional refers to simulated values that follow a standard normal distribution and reproduce the input covariance but do not reproduce the input data values. Posterior conditioning of these realizations will guarantee that the data are reproduced. Of the methods listed above, SGS is the only one that has a conditioning step built-in, and for this reason it is one of the most popular Gaussian algorithms in geostatistics. Another advantage of SGS is that univariate or multivariate simulation are performed with the same framework. Simulation at a location $\mathbf{u}$ requires building the CCDF at that location, which is performed with kriging in the univariate, and cokriging in the multivariate case. The implementation of multivariate simulation in the other algorithms is relatively more complex than in SGS. This complexity is not a concern in the proposed BSS methodology, since the simulation of the LMC factors is a univariate process.

Sequential Gaussian

The sequential simulation paradigm relies on a recursive application of Bayes' law to define the CCDF of the joint distribution of N RVs $\left\{Z(\mathbf{u}_i^{'}),\ i = 1, ..., N\right\}$. The number of simulation locations N is usually very large, e.g., a dense simulation grid discretizing the stationary domain A. Generating realizations of the N RVs conditional to the n original available data values $\{z(\mathbf{u}_\alpha),\ \alpha = 1, ..., n\}$ requires sampling the N-variate CCDF (Goovaerts, 1997):

$$\begin{aligned} F(\mathbf{u}_i^{'}, ..., \mathbf{u}_N^{'}; z_1, ..., z_N \mid (n)) = & F(\mathbf{u}_N^{'}; z_N \mid (n + N - 1)) \\ & \cdot F(\mathbf{u}_{N-1}^{'}; z_{N-1} \mid (n + N - 2)) \cdot\ ... \\ & \cdot F(\mathbf{u}_2^{'}; z_2 \mid (n + 1)) \cdot F(\mathbf{u}_1^{'}; z_1 \mid (n)) \end{aligned} \tag{2.36}$$

Realizations of the RVs are generated in N sequential steps, where each step involves a univariate CCDF defined from decomposition 2.36. The first simulated value $z^{(l)}(\mathbf{u}_1^{'})$ is drawn from the CCDF defined at $\mathbf{u}_1^{'}$ with the n conditioning data. The simulated value $z^{(l)}(\mathbf{u}_1^{'})$ is added to the data set and becomes a conditioning data to simulate the next location $\mathbf{u}_2^{'}$. The CCDF at location $\mathbf{u}_2^{'}$ is then defined on the n original data values and the previously simulated value. These steps repeat until all nodes N are simulated. The simulation at N locations requires the definition of the N univariate CCDFs $F(\mathbf{u}_1^{'}; z \mid (n)), ..., F(\mathbf{u}_N^{'}; z \mid (n + N - 1))$ with increasing level of conditioning. The inference of these distributions is possible under the multivariate Gaussian model.

The SGS algorithm is very simple and straightforward implementation of the sequential paradigm with the multiGaussian RF model (Gomez-Hernandez and Journel, 1993; Isaaks, 1990). Under such assumption, the N univariate CCDFs are assumed Gaussian and fully determined by the simple kriging and variance from the $(n+i-1)$ conditioning data. Because the number of conditioning data increases very quickly as simulation happens, the size of the kriging systems to be solved become prohibitive. Practical implementation of the SGS algorithm limits the number of data for conditioning to data a fixed maximum within a search neighborhood. Good practice consists in using the variogram ranges and anisotropy to define the search. Simulation with SGS proceeds as follows:

- Define a random path for simulation, each node is visited once.
- At the location being simulated, search for the conditioning data and solve the normal equations (simple kriging of the NS data) to determine the conditional mean and variance of the Gaussian CCDF at that location.
- Draw a simulated value from the CCDF and add it to the conditioning data set.
- Move to the next node location in the random path.

Multiple realizations are generated with a different random numbers that are used to define the random path and the sampled quantiles from CCDFs. Simple cokriging is used to define the CCDFs in the second step to simulate correlated variables.

Moving average

Moving average is one of the simplest algorithms to generate unconditional realizations in geostatistics. Simulation of the random function $Y(\mathbf{u})$ with covariance $C_y(\mathbf{h})$ is performed with the convolution product

$$Y(\mathbf{u}) = \int_{-\infty}^{\infty} f(\mathbf{u} - t) X(\mathbf{t}) \, dt \tag{2.37}$$

where $X(\mathbf{t})$ is a second order stationary RF with mean of zero and covariance $C_x(\mathbf{h})$, and $f(\mathbf{u}-t)$ is the weight function applied to each value $X(\mathbf{t})$. Simulation with moving averages simplifies the integral by considering discrete grid points in a simulation grid so that the integral in 2.37 is replaced by the summation over all points within the window. A special case that makes moving average suitable for large applications in geostatistics is when the RF $X(\mathbf{t})$ is a standardized random noise (pure nugget effect) and the weight function has a constant value of 1 within a distance $a/2 =\mid \mathbf{u}-t \mid$ and zero at distances beyond $a/2$ (Luster, 1985). In this case the covariance between $Y(\mathbf{u})$ and $Y(\mathbf{u}+h)$ is equal to the volume of the intersection of two n-dimensional spheres of diameter a centered at $\mathbf{u}$ and $\mathbf{u}+h$. The intersection of these two spheres defines the covariance between pairs of points separated by $\mathbf{h}$. The diameter a of the sphere is the range of the covariance $C_y(\mathbf{h})$.

Simulating a RF with a spherical covariance function is straightforward because the weight function is linear and constant for all data inside the window, but simulating other covariances involve an expensive convolution process. For example, the weight function to simulate a 2D Gaussian type covariance (Equation 2.9) with a range of a is $f(\mathbf{r}) = (4/a^2\pi)^{1/2}exp(-2r^2/a^2)$. Simulation at a location $\mathbf{u}$ requires the calculation of the distance r of every node data to $\mathbf{u}$. This process is computationally expensive for a large number of simulation locations and long covariance ranges. Weight functions and kernels for other covariance functions are found extensively in the literature (Chiles and Delfiner, 2012; Journel, 1974; Oliver, 1995; Oliver et al., 2008).

The simulation of spherical covariance functions are speed up by a clever update of the node data when simulation is performed on a regular grid. When simulation proceeds from a node location $\mathbf{u}$ to the next $\mathbf{u}^{'}$ the contribution of some values are removed and the contribution of other values are added. The algorithm keeps tracking of the nodes getting out of the window (removed) and the nodes getting in the window (added). This implementation, illustrated in Figure 2.2, permits fast simulation of spherical functions in relatively large models (Cabral Pinto and Deutsch, 2017d).

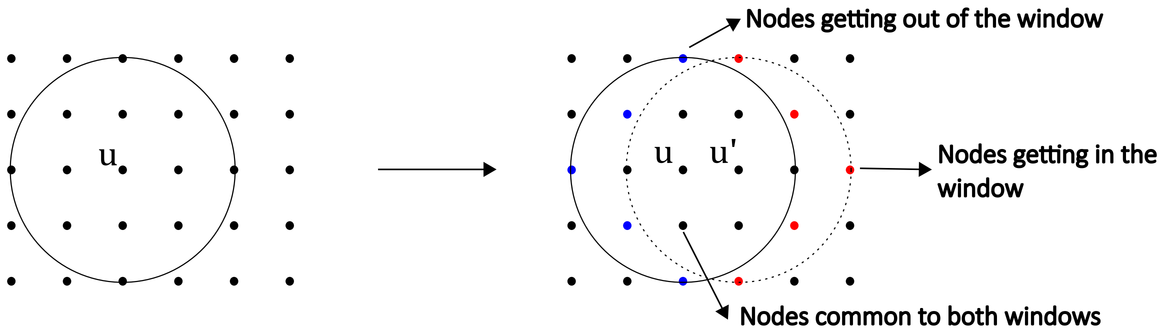

u
u
u'
Nodes getting out of the window
Nodes getting in the window
Nodes common to both windows

Figure 2.2: The process to update the node data indices for moving average simulation is schematically illustrated for two locations in a regular grid.

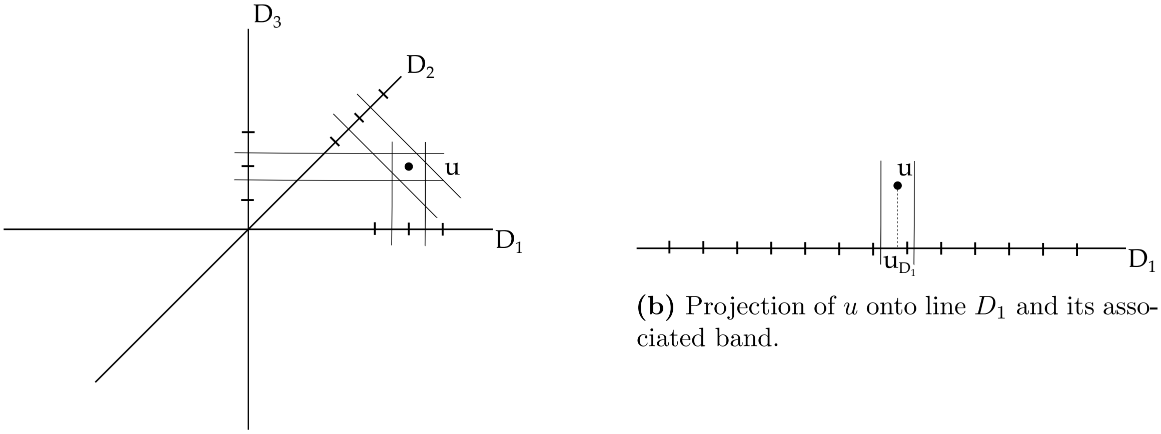
The moving average algorithm as implemented in the GSLIB-like program `MW_SIM` (Cabral Pinto and Deutsch, 2017d) proceeds as follow:

- Project the window onto the coordinate axes and calculate the length of the directional projections.
- Pad the grid based on the projections to avoid artefact in the border of the grid.
- Centre the window at the first cell of the grid and calculate the sum of all values inside the window.
- For each subsequent node until all nodes have been visited:
  - Move the window over the grid and add the values of the node data getting in the window to the previously calculated sum.
  - Subtract the values of the node data getting out the window from the previously calculated sum.
- Average the values considering the number of nodes inside the window.
- Rescale mean and variance as needed.
- Add prior mean as needed.

The moving average algorithm is used to simulate any RF whose covariance can be expressed in terms of the convolution product 2.37. In practice, this method is limited to the simulation of spherical covariance functions in 3D or circular covariance functions in 2D.

Turning bands

The turning bands algorithm was the first 3D simulation algorithm in geostatistics, originally developed to simulate isotropic spherical and exponential covariance functions (Journel, 1974; Matheron, 1973). It was later expanded to other covariance functions (Brooker, 1985; Mantoglou and Wilson, 1982) and more recently updated to simulate multivariate RFs in fast implementations of the algorithm (Emery, 2008; Emery and Lantuejoul, 2006; Marcotte, 2016). Generating realizations with turning bands requires a series of independent 1D realizations on a set of lines. These lines are generated equally distributed in space and radiate from the same point, often set in the grid origin. The simulation node is orthogonally projected onto the lines and is associated to the set of 1D simulated values that fall inside an interval space (or band) of each line. The simulated value for that node in the grid is a function of the sum of these values. The method is schematically illustrated in Figure 2.3.



(b) Projection of u onto line D_1 and its associated band.

(a) Orthogonal projection of a location u being simulated onto the lines.

Figure 2.3: Illustration of the turning bands methods in two dimensions. Each simulation location is projected onto a set of lines. The simulated value at u is a function of the 1D simulated points inside the correspondent bands of all lines. The ticks represent the discretize points along the lines where 1D simulation occurs.

Consider generating realizations $z(\mathbf{u})$ of a RF $Z(\mathbf{u})$ with known covariance $C(\mathbf{h})$.

Consider also a one dimensional RF $Y(u_1)$ on line D_1. This RF is stationary of order two, has zero mean, and covariance $C^1(\mathbf{h})$. A RF $Z_1(\mathbf{u}) = Y(u_{D_1})$ is defined, where u_{D_1} represents the orthogonal projection of the grid location being simulated u onto the line D_1. This RF has zero expectation and a one dimensional stationary covariance $C^1(h_{D_1})$, where h_{D_1} is the projection of vector $\mathbf{h}$ onto D_1. A realization of $z_1(\mathbf{u})$ is generated from the value $y(u_{D_1})$ which lie within a band perpendicular to D_1 (Figure 2.3b) and centered at $y(u_{D_1})$. This process is generalized to N lines $D_1, ..., D_N$ with different directions uniformly distributed on the unit circle (2D realizations) or on the unit sphere (3D realizations). A realization of $Z(\mathbf{u})$ is a linear combination of the RV $Y(u_{D_i})$ simulated on each line D_i:

$$Z(\mathbf{u}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y(u_{D_i}) \tag{2.38}$$

It is important to note that the covariance of the RF $Z(\mathbf{u})$ depends on the implementation of the algorithm and can be expressed in its two dimensional form $C^2(\mathbf{h})$ or three dimensional form $C^3(\mathbf{h})$. As the number of lines becomes very large these covariances tend to the isotropic covariances:

$$C^2(\mathbf{h}) = \frac{2}{\pi \mathbf{h}} \int_0^{\mathbf{h}} \frac{C^1(\mathbf{u})}{\sqrt{1 - (\mathbf{u}/\mathbf{h})^2}} d\mathbf{u} \tag{2.39}$$

$$C^3(\mathbf{h}) = \frac{1}{\mathbf{h}} \int_0^{\mathbf{h}} C^1(\mathbf{u}) d\mathbf{u} \tag{2.40}$$

Because simulation on the lines is performed at discretized points (Figure 2.3), these integrals are solved in their discretized forms. Two or three dimensional simulation in turning bands requires the simulation on the lines to be performed with the right form for the 1D covariance $C^1(\mathbf{h})$. The functions for the 1D covariances are derived in literature (Brooker, 1985; Brooker and Paul, 1982; Chiles and Delfiner, 2012; Mantoglou, 1987). The integral 2.39 is solved using the analytic form of the derived 1D covariances. The solution of equation 2.40 is simpler because the 1D covariance can be reduced to a convolution form. Therefore, different algorithms such as moving average (Journel and Huijbregts, 1978) or spectral simulation (Emery and Lantuejoul, 2006; Mantoglou and Wilson, 1982) can be used to generate realizations on the lines that will reproduce the desired covariance function. For this reason, practical implementation of the turning bands method simulate in 3D space, and 2D realizations are obtained by slicing the model. Anisotropy is achieved by geometric transformation of the grid, and realizations are conditioned with (dual)kriging.

Spectral

Spectral simulation methods generate 3D realizations of a RF with covariance $C(\mathbf{h})$ in the frequency domain (Davis et al., 1981; Emery and Arroyo, 2018; Mejia and Rodriguez-Iturbe, 1974; Yao, 1998a,9). In a particular case with a discrete transformation and limited to a regular grid, the stationary covariance is transformed into the density spectrum with a discrete Fourier transform (DFT). The inverse of DFT converts the simulated values from the frequency domain to the original space domain. The DFT decomposes an original signal waveform into its frequency representation. For a series of N discrete values $f(\tau), \tau = 0, ..., N-1$ the forward one-dimensional DFT of $f(\tau)$ and its inverse are defined respectively as:

$$F(\nu) = DFT(f(\tau)) = \frac{1}{N}\sum_{\tau=0}^{N-1} f(\tau)e^{-i2\pi\tau\nu/N} \quad for \quad \nu = 0, ..., N-1 \tag{2.41}$$

$$f(\tau) = DFT^{-1}(F(\nu)) = \frac{1}{N}\sum_{\nu=0}^{N-1} f(\nu)e^{i2\pi\tau\nu/N} \tag{2.42}$$

where $i = \sqrt{-1}$ is the imaginary unit. There exist other spectral methods that are continuous and can simulate anywhere, not only on regular grids (Chiles and Delfiner, 2012; Emery and Lantuejoul, 2006; Lantuejoul, 2002; Shinozuka, 1971; Shinozuka and Jan, 1972). The fast Fourier transform (FFT) is the most popular algorithm to calculate the DFT of a function (Smith, 1998). The direct computation of the DFT with Equation 2.41 requires N^2 operations. The computation of the

DFT with FFT requires a number $N \times log(N)$ of operations, significantly reducing the computation time. The FFT algorithm is illustrated in Figure 2.4. The first transformation decomposes the N-point signal in two $N/2$ length signal. At this point, the number of operations is of $2 \times (N/2)^2$ and the DFT of the N-point signal is the summation of the DFTs of the two decomposed signals. This process continues until the number of computations is reduced to $N \times log(N)$. Direct computation of Equation 2.41 yields the same result of FFT but it is a much more expensive process for a large number N. For example, the number of operations to compute the DFT of a function with $N = 100,000$ is 10^9, whereas FFT would take only $500,000$ operations.

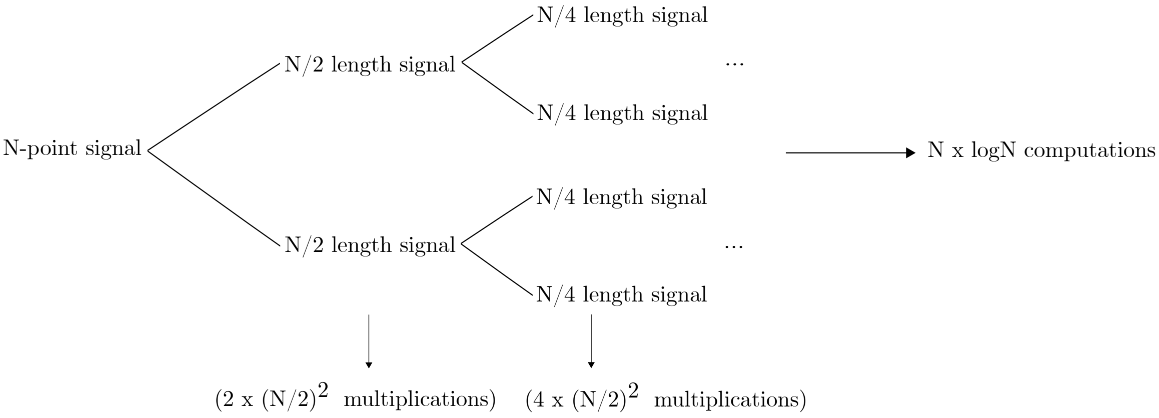


Figure 2.4: The number of computations to calculate the DFT of a signal is greatly decreased with the FFT algorithm.

The FFT of the covariance $C(\mathbf{h})$ of a RF $Z(\mathbf{u})$ is represented in the frequency domain by its density spectrum $s(\omega)$:

$$s(\omega) = FFT(C(\mathbf{h})) = |Z(\omega)|^2 \tag{2.43}$$

where the term $|Z(\omega)|^2$ represents the energy (or amplitude) of the signal. The Fourier coefficient $Z(\omega)$ is calculated as

$$Z(\omega) = |Z(\omega)| \, e^{-i\varphi(\omega)} = \sqrt{s(\omega)} e^{-i\varphi(\omega)} \tag{2.44}$$

with the phase spectrum $\varphi(\omega)$ measured in radians. Realizations of $z(\mathbf{u})$ are generated from its frequency counterparts $s(\omega)$ and $\varphi(\omega)$. Realizations in the spectrum density are generated by drawing phases randomly from a uniform distribution within the interval $[0, 2\pi]$. The inverse FFT of these realization generate realizations of $z(\mathbf{u})$ in the space domain. The spectral simulation algorithm proceeds as follows:

- Define a simulation grid where simulation will occur. Pad the grid with respect to the covariance range to avoid artifacts at the borders.
- Calculate the covariance values from the center of the grid. This avoids artifacts with the periodic nature of the DFT if the stationary covariance function is too continuous.
- Apply FFT on the calculated covariances to yield the density spectrum $s(\omega)$.
- Square root the density spectrum to yield the amplitude spectrum $|Z(\omega)|$.
- Randomly draw phase values $\varphi(\omega)$ from a uniform distribution within $[0, 2\pi]$.
- Calculate the Fourier coefficient $Z(\omega) = |Z(\omega)| \, e^{-i\varphi(\omega)}$.
- Perform the FFT^{-1} on $Z(\omega)$ for a realization of $z(\mathbf{u})$ in the space domain.

Modern implementations of the FFT algorithm extend its application to variogram calculation (Marcotte, 1996) and permit its computation with any grid size (FFTW, 2017). Spectral simulation offers a fast approach to simulate any covariance functions. In practice it is used to simulate very continuous covariance functions, that is, covariances with long range of continuity or with a parabolic behavior near the origin. Short range structures have more high-frequency variation, so $s(\omega)$ tends to zero slowly as ω tends to infinity. The discretization of such a long-tailed density requires large increments of ω in the transformation. This process leads to a loss of local information and increases the cost of simulation (Chatfield, 1980).

2.5 Multivariate transformations

Geological variables often show non-linearity, heteroskedasticity, composition constraints, and other complexities that conventional multivariate geostatistical techniques such as cokriging cannot capture. In univariate modeling, the NS transformation maps the original variables to the Gaussian space. This transformation only ensures that the marginal distributions are normal and does not guarantee multivariate normality. Complex multivariate relationships and constraints may remain after the transformation, however, such transformation is a common first step in most multivariate transformations. Multivariate decorrelation methods aim at removing nonlinearity and other complex relationships and constraints between variables. This section reviews some of these methods.

Stepwise Conditional Transformation (SCT) (Leuangthong, 2003; Rosenblatt, 1952) attempts to remove complex multivariate features by decorrelating the variable in an ordered fashion. The NS transform (Equation 2.28) converts the first variable to normal, the next k^{th} variable is transformed based on the conditional distributions given the previous $k-1$ transformed variables:

$$\left.\begin{aligned} y_1(\mathbf{u}_\alpha) &= G^{-1}\left(F_1(z_1(\mathbf{u}_\alpha))\right) \\ y_2(\mathbf{u}_\alpha) &= G^{-1}\left(F_{2|1}(z_2(\mathbf{u}_\alpha) \mid z_1(\mathbf{u}_\alpha))\right) \\ &\vdots \\ y_K(\mathbf{u}_\alpha) &= G^{-1}\left(F_{K|1,\ldots,K-1}(z_K(\mathbf{u}_\alpha) \mid z_1(\mathbf{u}_\alpha), \ldots, z_{K-1}(\mathbf{u}_\alpha))\right) \end{aligned}\right. \quad , \quad \forall\, \alpha = 1, \ldots, n \tag{2.45}$$

The transformed variables are multivariate Gaussian and independent at their collocated locations. Back transformation is similar to the NS back transformation expression (Equation 2.29) and respect the forward transformations described above. SCT has been used with success in multivariate modeling (Neufeld et al., 2008; Pyrcz and Deutsch, 2014). More recently, the use of kernel density estimation (Leuangthong and Deutsch, 2003), kernel density networks (Manchuk and Deutsch, 2011), and Gaussian mixture models (Silva and Deutsch, 2015) have enhanced the performance of SCT in the presence of sparse data and eliminated binning artifacts.

Minimum and maximum autocorrelation factors (MAF) was first introduced by Switzer (1984) as a method to decorrelate variables while maintaining specific direct and cross spatial continuity. In geostatistics, MAF is an enhancement of principal component analysis (PCA) that uses the direct and cross covariances to improve the PCA transformation of the input variables into uncorrelated and independent factors (Desbarats and Dimitrakopoulos, 2000). The ability to remove covariance at one more lag vector leads to better cross covariance reproduction than PCA, and for this reason MAF has become a popular MV transformation in mining (Barnett, 2015; Boucher and Dimitrakopoulos, 2012).

Projection Pursuit Multivariate Transform (PPMT) (Barnett, 2015; Barnett et al., 2014) is a technique that applies a modified component of the projection pursuit density estimation algorithm (Friedman, 1987) to decorrelate complex and high dimensional data. The first step of PPMT is to apply the NS transformation to all variables, then a variant of PCA (sphering) is used to decompose the eigenvalues of the covariance matrix at lag zero $\mathbf{C}(0) = \mathbf{V}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{V}$ is the matrix of eigenvectors, $\mathbf{D}$ is a diagonal matrix with eigenvalues, and $\mathbf{C}(0)$ is the $K \times K$ covariance matrix at lag zero. The sphering step rotates the data matrix with n normally transformed values $\mathbf{y} = \left(\mathbf{y}^\top(\mathbf{u}_1), ..., \mathbf{y}^\top(\mathbf{u}_n)\right)$ to the principal components basis, standardize and rotate back to original basis, yielding the sphered variables:

$$\mathbf{y}_0 = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top\mathbf{y} \tag{2.46}$$

An iterative process takes place to convert the most non-Gaussian projection of the data to be Gaussian. After a number of iterations, the data is multivariate-Gaussian and variables are uncorrelated. For every iteration the forward transformation is stored, the correspondent back transformation restores the complexities in the original data.

2.6 Matrix transformations

The Cholesky factorization is a method to solve a set of linear equations of form $AX = B$, where $A_{n\times n}$ is a non-singular positive definite square matrix, and $X_{n\times 1}$ and $B_{n\times 1}$ are the solution and right hand side vectors respectively. The Cholesky factorization has many applications in linear algebra, optimization, linear programming, and simulation (Golub and Loan, 2013). A symmetric positive definite matrix A is decomposed into the product of a lower triangular matrix $L_{n\times n}$ with an upper triangular matrix $L^{\top}_{n\times n}$:

$$A = LL^{\top} \tag{2.47}$$

where L is also called the Cholesky factor of A, has positive diagonal elements, and is interpreted as the square root of a positive definite matrix. The diagonal elements l_{kk} and the off-diagonal lower elements $l_{i,k}$ ($\forall\ i > k$) of the matrix L are respectively calculated by Equations 2.48 and 2.49:

$$l_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2} \quad ; \quad k = 1, ..., n \tag{2.48}$$

$$l_{ik} = \frac{1}{l_{kk}} \left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj} \right) \quad ; \quad k = 1, ..., n \quad and \quad \forall\ i > k \tag{2.49}$$

where a_{kk} are the elements of matrix A. This factorization permits an efficient solution to the linear system $AX = B$:

$$\begin{aligned} AX &= B \\ LL^{\top}X &= B \end{aligned} \tag{2.50}$$

defining $L^{\top}X = Y$, the following system of equations can be directly solved with forward and back substitution (Axler, 2015), solving first for Y then for X:

$$\begin{aligned} LY &= B \\ L^{\top}X &= Y \end{aligned} \tag{2.51}$$

The Cholesky factorization is numerically efficient in computational time for small matrices. Computation of large matrices may suffer from sparsity and round-off error propagation leading to suboptimal solutions or infeasible solutions (Dhiflaoui et al., 2003).

The singular value decomposition (SVD) is an orthogonal matrix reduction with applications in many fields such as in data analysis, linear algebra and linear least squares problems, digital processing analysis, blind source separation, and data reduction and compression in computer science (Golub and Reinsch, 1970; Goodfellow et al., 2016; Sadek, 2012; Schmidt, 2009; Strang, 2016; Zhang, 2009). SVD is the factorization of a matrix $A_{m\times n}$ with $m \geq n$ into orthogonal matrices $U_{m\times m}$ and $V_{n\times n}$, and a diagonal matrix $S_{m\times n}$:

$$A = USV^\top \tag{2.52}$$

where $U^\top U = V^\top V = I_n$ and $S = \text{diag}(\sigma_1, ..., \sigma_n)$ with $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$. The diagonal elements of S are called singular values, and the rank r of A is the number of nonzero singular values. The matrices U and V are the left and right singular vectors. The computation of the SVD consists of finding the eigenvectors of $A^\top A$ and $AA^\top$ that make up the columns of V and U respectively. The singular values in S are square roots of eigenvalues from $A^\top A$ or $AA^\top$. The SVD is also represented by sub-matrices partitioned by r:

$$A = USV^\top = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} \tag{2.53}$$

where $S_1 = \text{diag}(\sigma_1, ..., \sigma_r)$, $S_2 = \text{diag}_{(m-r)\times(n-r)}(0, ..., 0)$, $U_1 \in U_{m\times r}$, $U_2 \in U_{m\times(m-r)}$, $V_1 \in V_{n\times r}$, and $V_2 \in V_{n\times(n-r)}$. These matrices provide the range and null spaces for both the column and row spaces of A. U_1 provides an orthonormal basis for the column space of A. V_1 provides an orthonormal basis for the row space o A, and V_2 provides its orthogonal complement $U_1\perp$ or an orthonormal basis for the null space of A. The null space of a matrix A is the matrix $X = \text{null}A$, such that, $AX = 0$. Orthonormal basis are linearly independent vector that are mutually perpendicular and unitary in length (Strang, 2016). The row space of a matrix is the space spanned by the rows of A.

In cases where the matrix A in the linear system of equations $AX = B$ is rank-deficient and the columns of A are not independent, there are infinite solutions to $AX = B$. Usually the solution with the minimum norm $\hat{X}$ is chosen as the solution to the system $A\hat{X} = B$. The solution requires the calculation of the vector $\hat{X} = \min_X \parallel AX - B \parallel_2$. This requires the projections of B and $\hat{X}$ onto the column space of A and the row space of $A^\top$ respectively, leading to the normal equations $(A^\top A)\hat{X} = A^\top B$ (Golub and Loan, 2013). SVD provides a way to solve this problem using the generalized pseudo inverse of a matrix. The pseudo inverse $A^\dagger$ of a matrix $A = USV^\top$ is given by:

$$A^\dagger = VS^\dagger U^\top \tag{2.54}$$

where $S^\dagger$ is calculated by transposing S and inverting all non-zero elements. The solution of $A\hat{X} = B$ is then:

$$\begin{aligned} A\hat{X} &= B \\ \hat{X} &= VS^\dagger U^\top B \end{aligned} \tag{2.55}$$

The computation of SVD is slow and computationally expensive for large matrices and when the matrix size is approximately equal to its rank (Tzeng, 2013). The factorization of sparse matrices may generate suboptimal solutions and lack of convergence (Yang et al., 2014).

The geostatistics concepts and methods reviewed in this chapter are required for the development of the methodologies proposed in this thesis and explained in the next chapters.

Chapter 3

Multivariate criteria in geostatistical modeling

The following chapter presents the theory of the post processing of the projection pursuit multivariate transformation (PostPPMT) methodology. This methodology provides a framework for the direct assessement of local multivariate distributions. A case study with geochemical data demonstrates the application of the methodology for exploration geochemistry.

3.1 Motivation

Multivariate models aim to maximize the use of the information in the data and improve decisions. In recent years, geological modeling has gone from dealing with one economic variable to more complex scenarios with variables of different nature, such as grade, lithology, geomechanical properties, and geometallurgical properties. These multivariate models of uncertainty are built for decisions making. Decisions can be made based on a single threshold value or multiple criteria. An example of a single decision criterion is the cutoff grade concept - a block is mined as ore or waste if the grade is above or below the cutoff grade, see Figure 3.1. Modern decisions, however, involve multiple thresholds and considerations. These criteria may involve different rules being applied to many variables at the same time, see Figure 3.2. For example, the destination of a mined block in a polymetallic mine can be one of many blending and homogenization piles, a stockpile that will feed the mill, the leach pad, or the waste dump. These destinations are chosen based on a series of thresholds applied to multiple grades or ratios of grades, and also depend on other rock properties such as the degree of weathering of the rock, the lithological properties, the presence of contaminants, and so on.

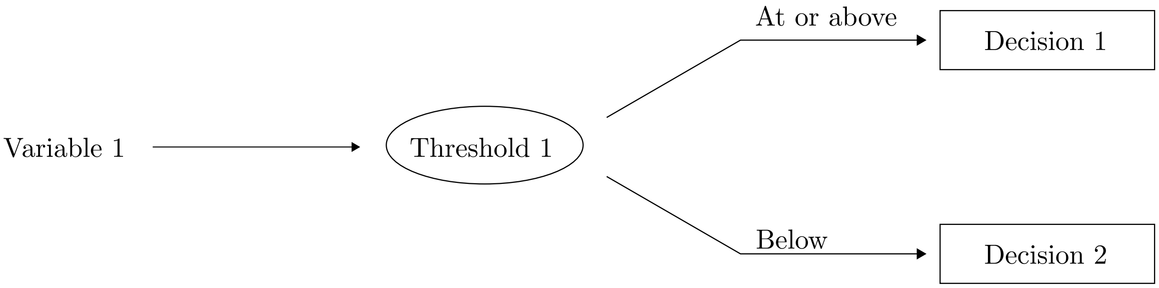


Figure 3.1: Illustrated scheme of a single criterion decision. The criterion is defined based on threshold 1 applied to variable 1. If the value of variable 1 is greater or equal than the threshold, then decision 1 is taken. Otherwise, decision 2 is taken.

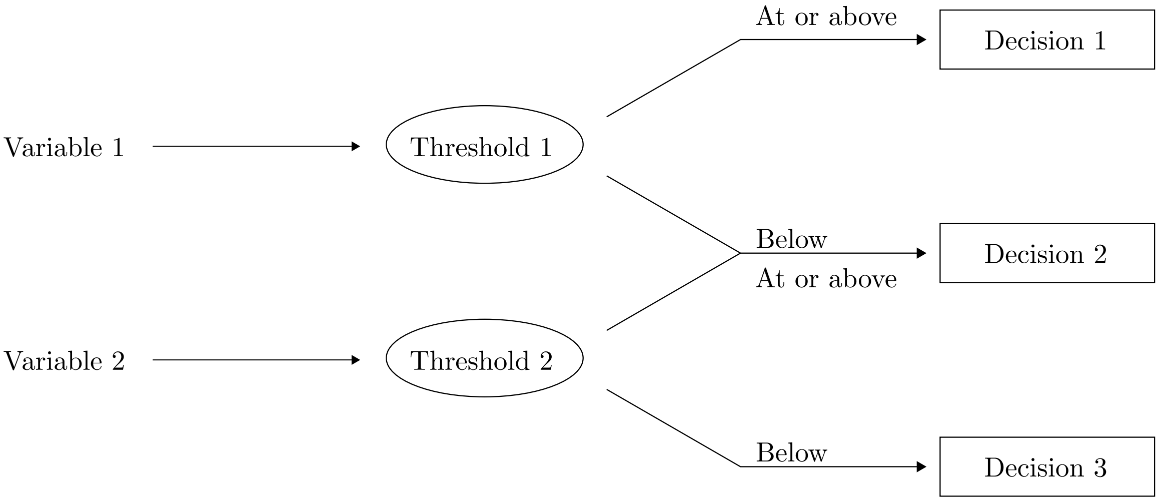


Figure 3.2: Illustrated scheme of multivariate criteria decision. The criteria are defined based on threshold 1 applied to variable 1 and threshold 2 applied to variable 2. Decision 1 depends only on the value of variable 1. Decision 2 is taken if variable 1 is below threshold 1 and variable 2 is greater or equal than threshold 2. Decision 3 depends only on the value of variable 2.

Although geostatistics provides ways for multivariate modeling, the application in the context of such criteria is not clear. The need to improve technical and economic decisions of geostatistical workflows given complex multivariate criteria motivates the development of the PostPPMT methodology.

3.2 Multivariate criteria

Criteria are one element of decision making that spell out or make clear the constraints or targets for a specific decision. Criteria define the desirability of specific outcomes and the negative consequences of other outcomes. Criteria are based on thresholds and objective values using the variables involved. Most decisions in earth sciences involve criteria with multiple aspects; decisions are rarely made based on the value of one variable.

In the field of mineral exploration, multivariate modeling of pathfinder elements is used to model the distribution of such elements in large areas to potentially narrow exploration targets. Complex criteria can be applied to high resolution models of uncertainty to identify anomalous regions for a type of mineral deposit. For example, different types of mineral deposits, such as sedimentary exhalative deposits (SEDEX) or volcanogenic massive sulfide (VMS), have their own signatures. VMS deposits found in northern Canada commonly show high values of Titanium (Ti) and Zirconium (Zr) and low values of Tantalum (Ta) and Scandium (Sc); while SEDEX deposits in the same region usually show very high values of Zinc (Zn), Silver (Ag), and Lead (Pb), and relatively high values of Barium (Ba), Copper (Cu), and Arsenic (As) (Berger, 2015; du Bray, 1995; Fischer et al., 2016; McClenaghan and Peter, 2013; Ootes et al., 2013). New areas for exploration can be defined by translating the deposits signatures into numerical criteria and evaluating the multivariate modeling outcomes with these rules.

Another application is in mineral processing. The multivariate modeling of geometallurgical properties has become as important as grades in many mines. The combination of geological and metallurgical information for production management is necessary for efficient planning, optimal design and maximum metal recovery (Deutsch, 2015b; Deutsch et al., 2016). The old approach of a single cutoff for ore and waste classification has been substituted for more complex decision models based on grades, rock properties, and processing routes (Rendu, 2008). Teck's Red Dog mine complex, destination criteria for mined blocks depend on the grade of Zn, Pb, Ag, Ba, iron (Fe), Silica (SiO_2), and the degree of weathering of the rock. Rocks with high grades of Ba and low grades of Zn and Fe have a processing destination different than those with a high ratio of non-sulphide Pb to total Pb. High Ba/high non-sulphide rocks have another destination. Another destination is for weathered, low-grade, high non-sulphide Pb rocks that do not produce sellable concentrate and are sent to the waste dump. Low to moderately weathered rock with low Cu grades differ from highly weathered ore with low Cu grades, although both produce concentrates, the mined blocks have different destination to different blending piles before floatation. Other complex rules involve Fe as pyrite (high ratios of Pyrite to Zn or Pb), SiO_2, and weathering as deleterious process impacting both Pb and Zn metallurgy (Teck, 2017).

In Newmont's Twin Creeks operation, the resources model and mine planning are reported to be stochastic. The combination of a set of orebody data generated by geostatistical simulation with a production scheduling optimizer increases the net present value (NPV) of the mine complex and provides a basis for assessing different block destination scenarios for decision making (Montiel and Dimitrakopoulos, 2018). Destination criteria involve different types of ore with varying gold grade and cyanide solubility being sent to stockpiles before mill blending in three different facilities. Higher grade oxide ore is processed by conventional milling and cyanide leaching, lower grade material with suitable cyanide solubility is treated on heap leach pads, and refractory ores are fed through grinding mills followed by autoclaves (Ewing, 2016; Kawahata et al., 2016). Assessing the destination criteria for mined blocks can improve mine operations and support technical decisions regarding fleet destination.

Other applications of multivariate modeling with complex criteria are also found in environmental sciences (Lin, 2002; Liu and Koike, 2007; Patil and Rao, 1994; Webster and Oliver, 2007a). The details of these criteria vary from one project to another. The PostPPMT methodology defines how these rules will be transferred and processed in practical geostatistical workflows.

3.3 PostPPMT for estimation and local uncertainty assessment

The PostPPMT methodology is a hybrid of the post-multiGaussian (PostMG) methodology (Deutsch and Journel, 1992; Verly, 1983) and the application of PPMT in the context of estimation and local uncertainty assessment. The step of simulating the independent factors is skipped and the local multivariate distributions are directly back transformed. This is not suitable for situations where multilocation uncertainty is required, but it is a fast and efficient approach for local best estimates and measures of local uncertainty.

The PPMT is widely used for multivariate simulation. Multivariate data observations with no missing values are transformed to be independent and standard normal. Simulation of the transformed factors then proceeds independently or in sequence when secondary data are available. The simulated independent Gaussian variables are back transformed to original units where the multivariate complexity is restored. These realizations could be averaged to obtain a local estimate and other post processing could be considered for measures of local and multilocation uncertainty. Stable assessment of local uncertainty may require hundreds of realizations; a mere one hundred realizations would lead to significant noise in the variance or any probability sensitive to the tails of the distribution. In cases when local estimates or local uncertainty measures are the goal of the study, a direct local back transform with thousands of realizations provides stable results with little computational effort.

PostMG workflow

Considering a single regionalized variable in PPMT is equivalent to the multiGaussian approach long used in geostatistics (Verly, 1983). The core steps to this algorithm could be summarized as:

1. Infer a global representative distribution for the regionalized variable in the

deemed stationary domain by a declustering algorithm. Normal score transform the data and save the transformation table of paired original data and normal score transformed values. The stationary global distribution is standard normal.

2. Infer a variogram model of the normal score data. This stationary variogram model provides all of the covariances needed to parameterize the multivariate Gaussian distribution for the domain.

3. Perform simple kriging to compute a local mean and local variance at each unsampled location. These two moments at each location fully define the local conditional Gaussian distribution; for N locations, the $2 \times N$ numbers define the local parametric distributions of uncertainty.

4. Back transform and post process the local distributions. A large number of quantiles are defined and back transformed using the transformation table from step 1. The quantiles are sometimes regularly spaced, e.g., the 99 percentiles. Many random quantiles could also be sampled to allow the tails of the distribution the correct probability of being sampled. The locations are considered one at a time and local summary statistics are calculated from the back transformed values. It is common to save the local mean, the local variance, the probability to be within a specified tolerance of the mean, and specified local quantiles such as the p_{10}, p_{50} and p_{90}.

This straightforward workflow is computationally efficient and provides estimates and accurate data-value dependent measures of local uncertainty. It is not possible to calculate any multilocation measures of uncertainty; nevertheless, the local measures are useful. The conditional mean values could be used for resource calculation. The conditional variance values could be used to support classification decisions. A high (or low) probability of exceeding a critical threshold would make locations more or less interesting. Finally, the results provide a useful check on simulation-based workflows that are more computationally demanding and prone to error.

PostPPMT methodology

The PPMT transformation has become widely used because of its remarkable ability to transform multivariate data with arbitrarily complex behavior to be multivariate Gaussian and uncorrelated. Simulation and estimation is greatly facilitated since the transformed factors can be considered one at a time. The back transformation restores any complex multivariate behavior. Simulation is routinely performed since that provides an assessment of local uncertainty and multilocation uncertainty. The realizations can be processed in the normal way through resource/reserve calculations or optimization. As with the PostMG workflow, there are situations where stable local estimates and measures of local uncertainty are desired for direct use or for checking of the simulation workflow.

Considering K multiple regionalized variables in PPMT leads to a slightly modified algorithm that could be summarized as:

1. Infer a global representative distribution for all regionalized variables in the deemed stationary domain by a declustering algorithm. There can be no missing values; data imputation would be considered if necessary. PPMT transform the multivariate data and save the transformation. The PPMT transform consists of many steps, but they are all embedded in a single program that greatly simplifies the procedure from a user perspective. The stationary global distribution of each factor is standard normal. The factors are uncorrelated at lag zero.

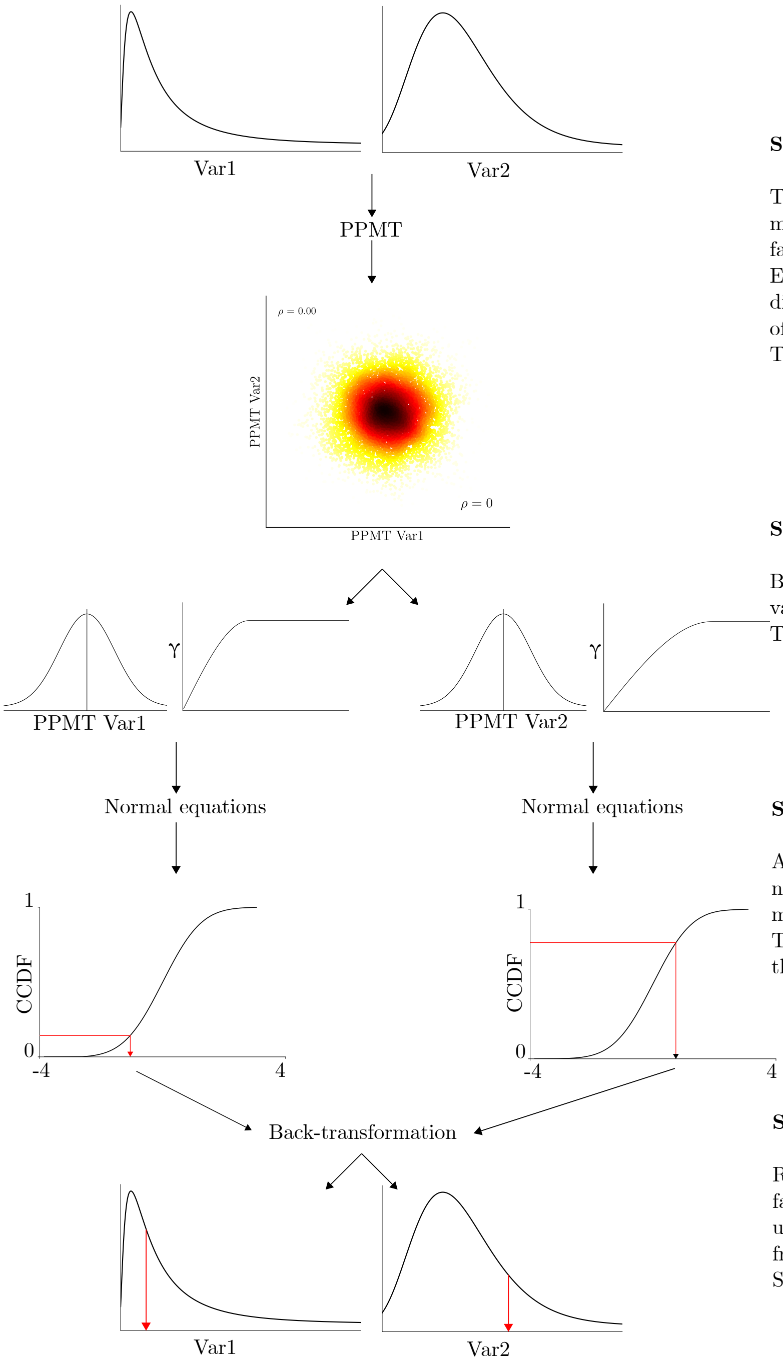
2. Infer a variogram model for each factor. If no MAF-like rotation is considered, then variograms of the normal score transformed variables is recommended. The variograms for the transformed variables are modeled independently.

3. Perform simple kriging (normal equations) to compute a local mean and local variance at each unsampled location for each factor. These moments fully define the local conditional Gaussian distribution; for N locations and K variables, the $2 \times K \times N$ numbers define the local parametric multivariate

distributions of uncertainty. Note that the local multivariate distributions are fully defined by the K univariate non-standard distributions since the local factors are independent.

4. Back transform and post process the local distributions. A large number of quantiles are defined and back transformed using the transformation table from step 1. The quantiles cannot be regularly spaced because of the potentially high dimensionality. Considering 100 regularly spaced quantiles for $K = 10$ variables would lead to 10^{20} values to back transform for each location. This is an excessively large number. An arbitrarily large number of random quantiles is recommended. Low-discrepancy random sequences like the Halton sequence are more uniformly distributed for a same number of quantiles and is an alternative to drawing from a uniform distribution (Halton, 1964). Thousands of randomly chosen back transformed values would provide stable results. The locations are considered one at a time and local summary statistics are calculated from the back transformed quantiles. It is common to save the local means, the local variances, and other measures of uncertainty, see below.

This straightforward workflow, illustrated in Figure 3.3, is computationally efficient and provides estimates and measures of local uncertainty. The results could be used for resources, local multivariate criteria assessment and for checking simulation. Regarding the summary statistics to save, there are many possibilities. The local mean and variance for each variable is always useful for global resource assessment and for checking. The probability of certain variables to be within a tolerance of the mean (e.g., 15%) could be useful in some circumstances. Specific quantiles of certain variables could also provide a measure of uncertainty. The probability of satisfying multivariate rules could be interesting, e.g., variable 1 above a threshold while variable 2 is above a different threshold and variable 3 is below another threshold. Considering the ratios of different variables in the multivariate rules could also be assessed.



STEP 1:

The PPMT transform decorrelates multivariate data into standard normal factors.
Each PPMT factors follows a Gaussian distribution with zero mean and variance of one.
The factors are uncorrelated.

STEP 2:

Because the factors are uncorrelated, the variograms are modeled independenlty. The cross variograms are not needed.

STEP 3:

At each location considered, solve the normal equations to compute a local mean and local variance for each factor. The K-univariate local CCDF fully define the local multivariate distributions.

STEP 4:

Randomly sample the CCDF of each factor. Back-transform the drawn values using the PPMT transformation table from STEP 1.
Save the values for post-processing.

Figure 3.3: Schematic illustration of the PostPPMT methodology.

The postPPMT program implements the PostPPMT methodology in GSLIB format. The run-time to process the local distributions and back transform the large number of quantiles is greatly reduced with the postPPMT FORTRAN program (section A.1). The following case study demonstrates the practical implementation of the methodology in exploration geochemistry with multivariate criteria.

3.4 Application to exploration geochemistry

Publicly available data are used to developed the PostPPMT case study described above. The geochemical data were collected by the Northwest Territories Geological Survey (NTGS) in partnership with the Geological Survey of Canada (GSC) across the Mackenzie Mountains in the Northwest Territories, Canada. This regional geochemical survey is conducted for the evaluation of mineral potential in the area, based on sample collection and analysis protocols developed by the GSC for the National Geochemical Reconnaissance program (Falck et al., 2012).

Data

The Inductively Coupled Plasma Mass Spectrometry (ICP-MS) measurements of stream sediments of Ag, As, Cu, Pb, Thallium (Tl), Vanadium (V), Zn; and the Instrumental Neutron Activation Analysis (INAA) measures of stream sediments of Ba are considered. Duplicated measurements and samples with missing analyses are removed. A subset of the geochemical data containing 8188 samples is considered. These elements are selected as the pathfinders for Ag/Zn/Pb SEDEX deposits in the area. The available geochemical data are plotted on top of the regional geology in Figure 3.4. The distribution of the elements are highly right-skewed due to a large percentage of samples at or below the detection limits and the presence of outliers, see Table 3.1 for a summary of the statistics and Figure 3.5 for the histograms of Ag and Cu.

The analysis of the variable distributions might reveal interesting threshold values that combine with professional judgement to support decisions when formulating

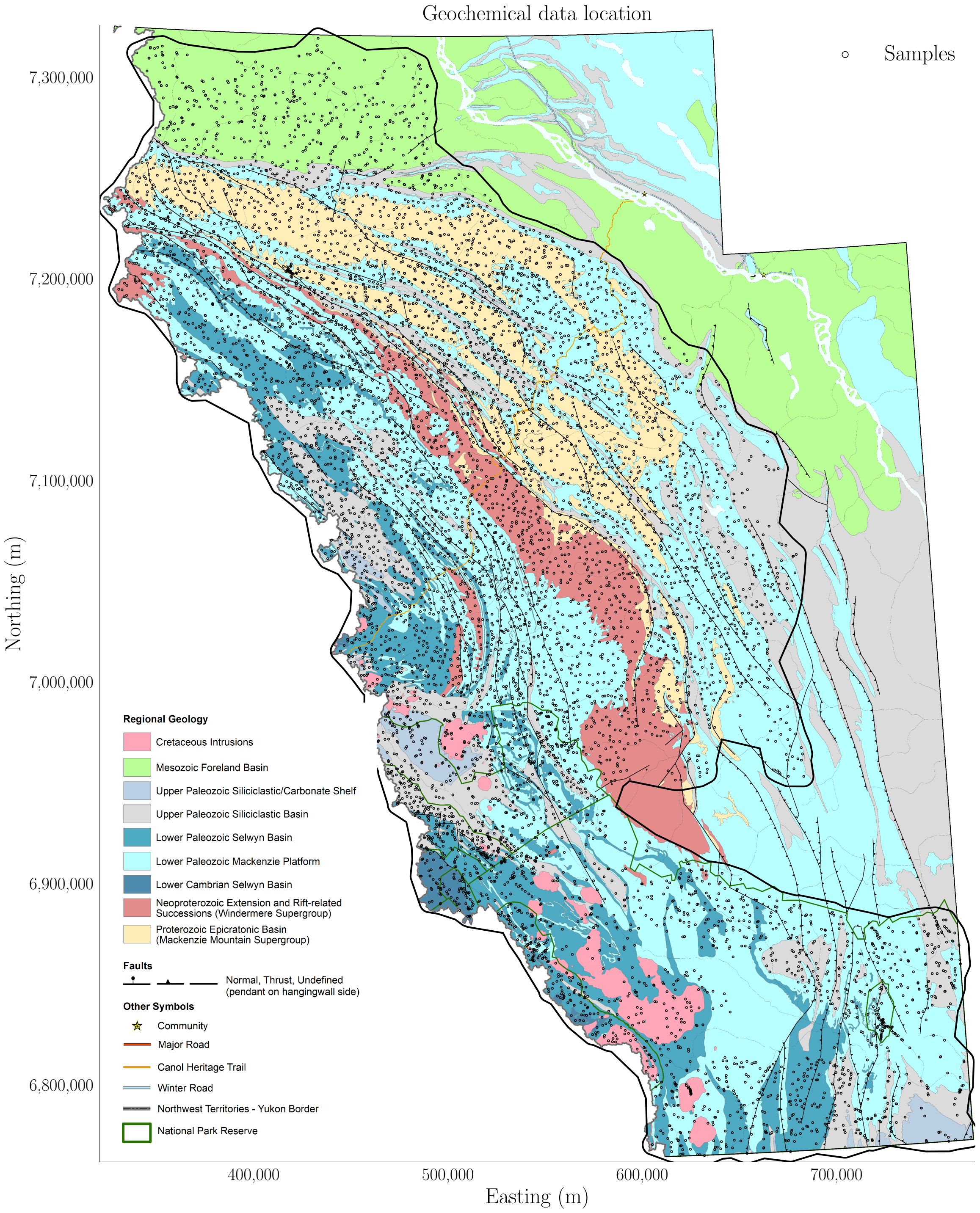

Geochemical data location
○ Samples
Northing (m)
7,300,000
7,200,000
7,100,000
7,000,000
6,900,000
6,800,000
Easting (m)
400,000
500,000
600,000
700,000
Regional Geology
Cretaceous Intrusions
Mesozoic Foreland Basin
Upper Paleozoic Siliciclastic/Carbonate Shelf
Upper Paleozoic Siliciclastic Basin
Lower Paleozoic Selwyn Basin
Lower Paleozoic Mackenzie Platform
Lower Cambrian Selwyn Basin
Neoproterozoic Extension and Rift-related Successions (Windermere Supergroup)
Proterozoic Epicratonic Basin (Mackenzie Mountain Supergroup)
Faults
Normal, Thrust, Undefined (pendant on hangingwall side)
Other Symbols
Community
Major Road
Canol Heritage Trail
Winter Road
Northwest Territories - Yukon Border
National Park Reserve

Figure 3.4: The geochemical samples (black circles) are plotted on top of the geological map of the Mackenzie Mountains. The thick line represents the keyout that delimits the area where estimates are calculated. Geological map provided by the NTGS.

Table 3.1: Summary of statistics for the pathfinder elements. The 25th, 50th (median), 75th, and 90th percentiles are shown along with the mean (μ) and standard deviation (σ).

Element	Minimum	p_{25}	p_{50}	p_{75}	p_{90}	Maximum	μ	σ
Ag(ppb)	0.00	24.00	51.00	121.00	308.000	6732.00	124.01	234.17
As(ppm)	0.00	3.70	6.30	11.10	23.100	1060.80	12.32	30.23
Ba(ppm)	0.00	300.00	590.00	1200.00	3000.00	84000.00	1320.30	2841.69
Cu(ppm)	0.46	7.58	17.09	31.02	57.536	718.36	27.13	38.77
Pb(ppm)	0.50	7.69	11.88	18.49	28.820	2489.07	17.14	41.26
Tl(ppm)	0.00	0.05	0.08	0.16	0.350	6.33	0.16	0.25
V(ppm)	0.00	8.00	16.00	30.00	51.000	2959.00	28.57	69.19
Zn(ppm)	1.50	25.60	65.45	129.60	334.430	10000.00	170.70	428.91

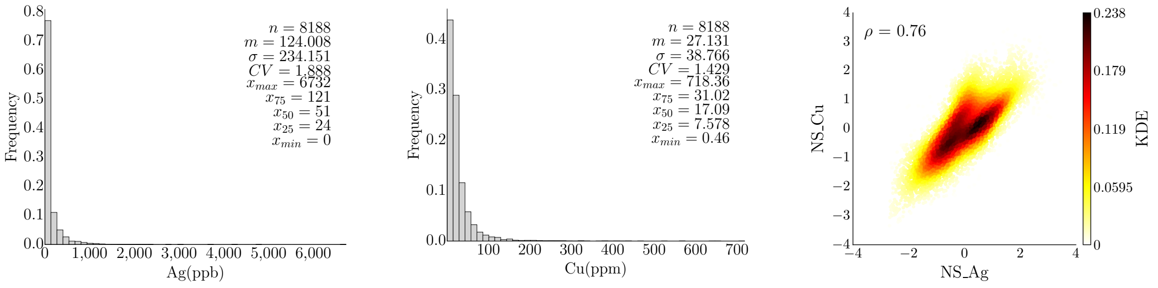
the multivariate criteria rule.

Multivariate criteria rule

The signature for Ag/Zn/Pb SEDEX deposits is defined as high values of specific pathfinder elements. The assessment of the following multivariate criteria is considered for targeting new exploration areas: Ag, Pb, and Zn values above their respective global p_{90} quantiles, and As, Ba, Cu, Tl, and V values above their respective p_{75} quantiles. Locations where the probability of all variables to exceed their thresholds simultaneously are flagged as potential areas for further investigation.

Geostatistical modeling

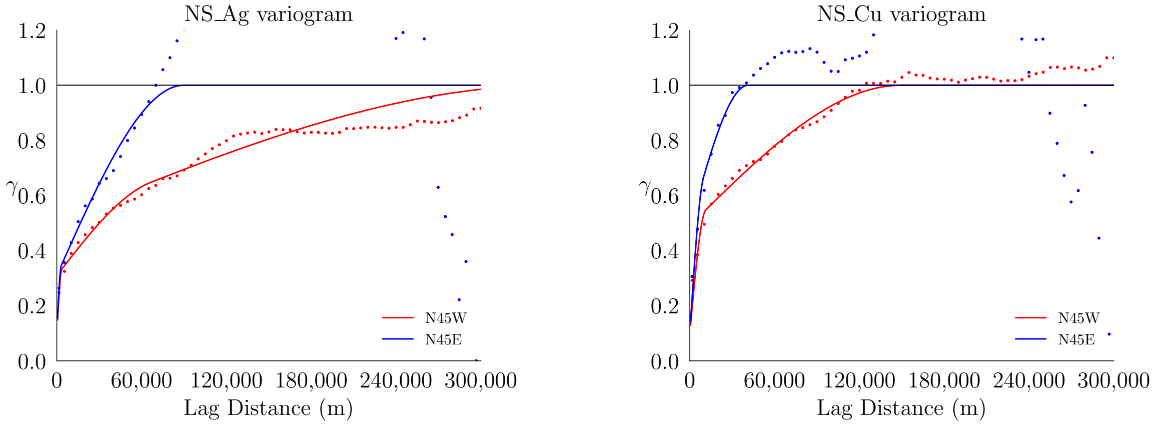
The multivariate NS transformation guarantees that the marginal distributions of the variables are standard normal, but multivariate normality is not always achieved. Figure 3.5c shows the bivariate relationship between the NS of Ag and Cu, note the high correlation between both variables. The NS variograms are an alternative to the PPMT variograms when MAF-like rotation is not required after the PPMT transformation and the correlation between the NS data and the PPMT transformed factors is high. The NS variograms are easier to model and not sensitive to spurious noise when variables are highly correlated, therefore, the NS transform is considered for variogram modeling.



(a) Ag distribution **(b)** Cu distribution **(c)** NS relationship

Figure 3.5: The distribution of Ag (a) and Cu (b) are shown in original units. The relationship of the normal score transforms of both variables is shown in the bivariate plot (c) and colored by the kernel density estimation. The correlation in normal scores is 0.76.

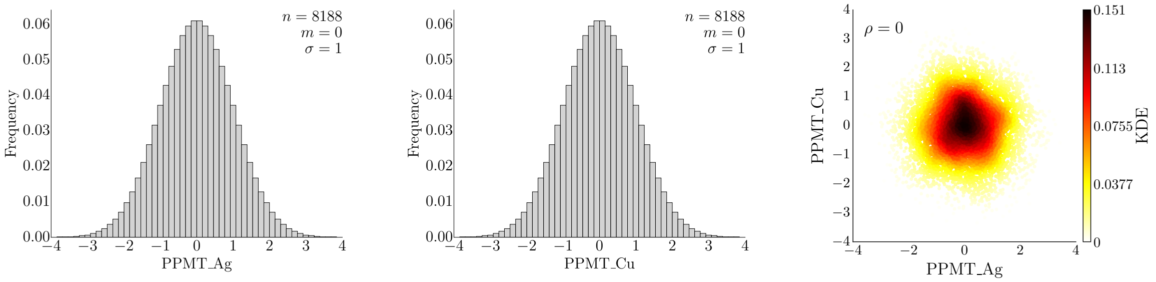
Experimental variograms are calculated at the directions of the apparent anisotropy, with the azimuths of major and minor directions of continuity set respectively to 315 and 45 degrees. The variograms are modeled with a small nugget effect of 10% of the total variance and three spherical structures. The contribution to the sill of each structure depends on the variable. In general, the variogram ranges vary from approximately 40 Km to 300 Km in the minor and major directions. The NS variograms of Ag and Cu are shown in Figure 3.6.



(a) NS Ag **(b)** NS Cu

Figure 3.6: The normal scores variograms of Ag (a) and Cu (b). The dots and lines represent the experimental points and the fitted model respectively. Blue and red colors represent the variograms at the minor and major directions of anisotropy respectively.

The local conditional distributions are defined by the simple kriging and variance of the PPMT factors. PPMT decorrelate variables while ensuring that the marginal and joint distributions are standard normal, see Figure 3.7 for the distribution of the PPMT factors of Ag and Cu and their bivariate relationship.



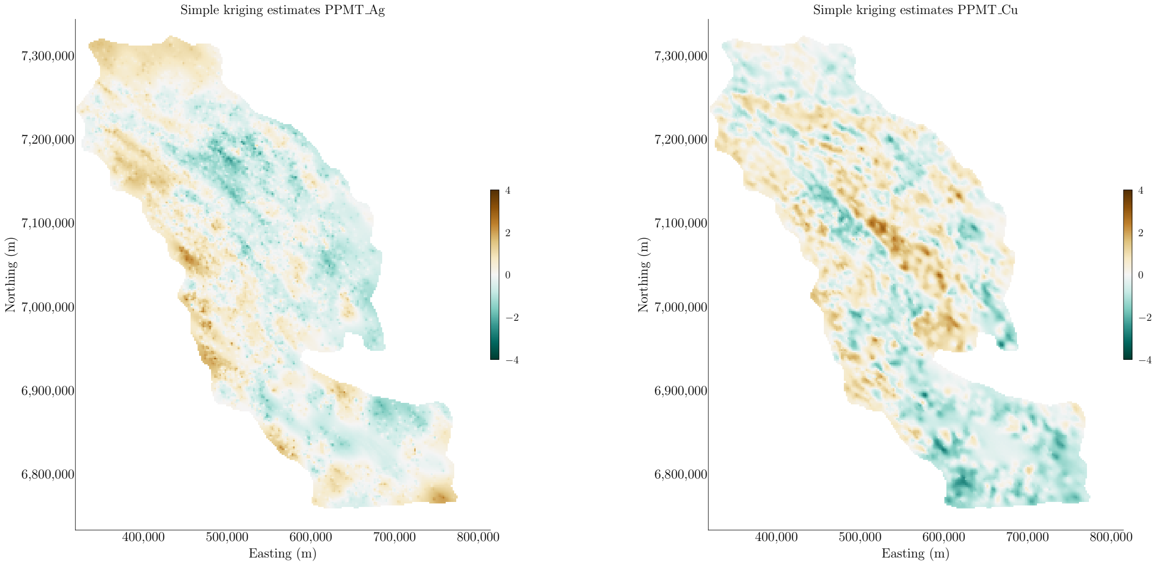
(a) PPMT Ag **(b)** PPMT Cu **(c)** PPMT relationship

Figure 3.7: The distribution of the PPMT factors of Ag (a) and Cu (b). The relationship of the PPMT factors of both variables is shown in the bivariate plot (c) and colored by the kernel density estimation. The factors are uncorrelated.

A grid containing 248 and 305 nodes respectively in the easting and northing directions, with a node spacing at both directions of 2 Km is considered. A keyout is used to ensure that only nodes inside the project area, a total of 29,475 locations, are estimated (Figure 3.4). The PPMT factors are simple kriged on the grid locations using their respective NS variograms and accounting for the anisotropy. The vectors of simple kriging mean and variance of each factor at each location are used to build the local multivariate conditional distributions that are used for local uncertainty assessement. The simple kriging estimates of the PPMT factors of Ag and Cu are shown in Figure 3.8.

Assessing the multivariate criteria

At each location the conditional distribution of each factor is sampled 10,000 times. This large number of quantiles is required to account for the highly skewed distributions of the variables. The quantiles are back transformed and the probability of the individual variable rules and multivariate criteria are considered. The individual rules are defined on a variable by variable basis as defined in subsection 3.4, for example, Ag values above the p_{90} quantile. The joint-probability is calculated counting the number of times that the single rules occur simultaneously at a location. Therefore, the calculation of the joint-probability involves the calculation of single rule probabilities. Figure 3.9 shows the distribution and map of the individ-



(a) Kriging estimates PPMT Ag

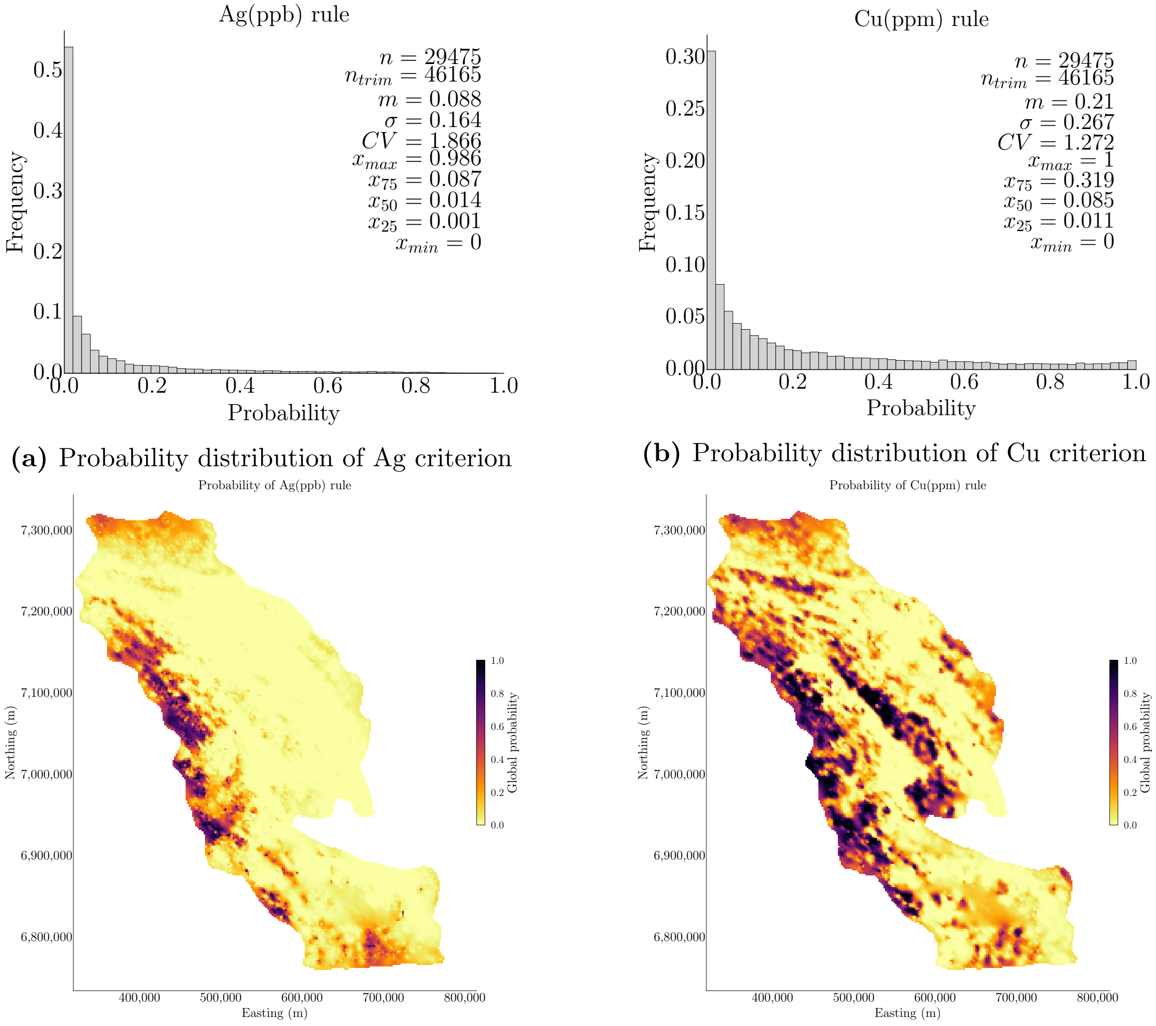
(b) Kriging estimates PPMT Cu

Figure 3.8: The simple kriging estimates of the PPMT factors of Ag (a) and Cu (b) are plotted on the grid.

ual probabilities calculated for Ag and Cu. These results could be used for decision making if decisions are based on a single criterion only.

The multivariate criteria probability is now considered. Locations with a higher probability of the multivariate rule to occur are flagged as new potential exploration areas. In the case of exploration geochemistry, this probability is expected to be low overall. The distribution of the joint-probability in the project area is shown in Figure 3.10. Note the large number of locations where the probability of the multivariate criteria to occur is zero. The expected value over the project area is only 0.7%, with a maximum probability value of 87.2%.

The global joint-probability map is shown in Figure 3.11. Most areas of interest are located in the western region of the Mackenzie Mountains, with peaks of probability concentrated in the southwest region of the map. To improve the visualization of these areas, consider the ratio of local to global probability, that is, the ratio of local values to the mean of 0.7%. Consider also taking the logarithm to the base 10 of the calculated ratio, see Figure 3.12. The value of 0 represent all locations where the local probability is equal to the global average. Positive values represent areas where the local probability is above average, the more positive the value, the higher



(a) Probability distribution of Ag criterion

(b) Probability distribution of Cu criterion

(c) Probability map of Ag criterion

(d) Probability map of Cu criterion

Figure 3.9: The distribution of the calculated probabilities based on the univariate rules of Ag (a) and Cu (b); and the map of the respective calculated probabilities of Ag (c) and Cu (d). The univariate criteria are Ag and Cu above their respective global p_{90} and p_{75} quantiles respectively. Probability values are on a 0 to 1 scale.

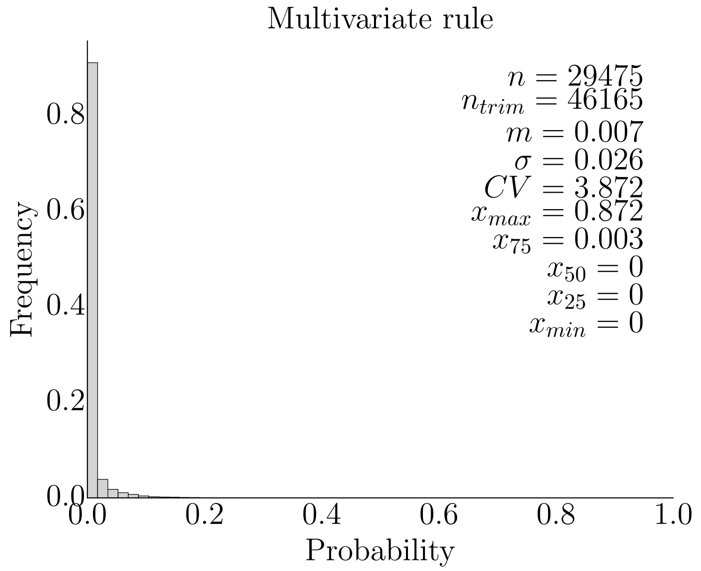


Figure 3.10: The distribution of the multivariate joint-probability. Probability values are on a 0 to 1 scale.

the potential of the area is. In the logarithm scale, a positive value of 2 represents local values 100× greater than the average. Negative values represent areas where the local probability is below the average.

Overall, the western region of the Mackenzie Mountains concentrates most of the potential areas for exploration of Ag/Zn/Pb SEDEX deposits. The analysis of the flagged areas by geologists can support decisions regarding targeting new zones and potentially narrow exploration areas. The pathfinder elements were chosen to illustrate the methodology and would need to be refined for future application. The choice of the elements and the multivariate rule was based on a series of reports, thesis, and articles on the geology of SEDEX deposits and mineral deposits in the Mackenzie Mountains (Berger, 2015; du Bray, 1995; Fischer et al., 2016; Ootes et al., 2013).

3.5 Conclusion

The practical aspects and development of the PostPPMT methodology is discussed in this chapter and demonstrated in a case study with geochemical data. The application of the PPMT transformation for estimation and local uncertainty assessment

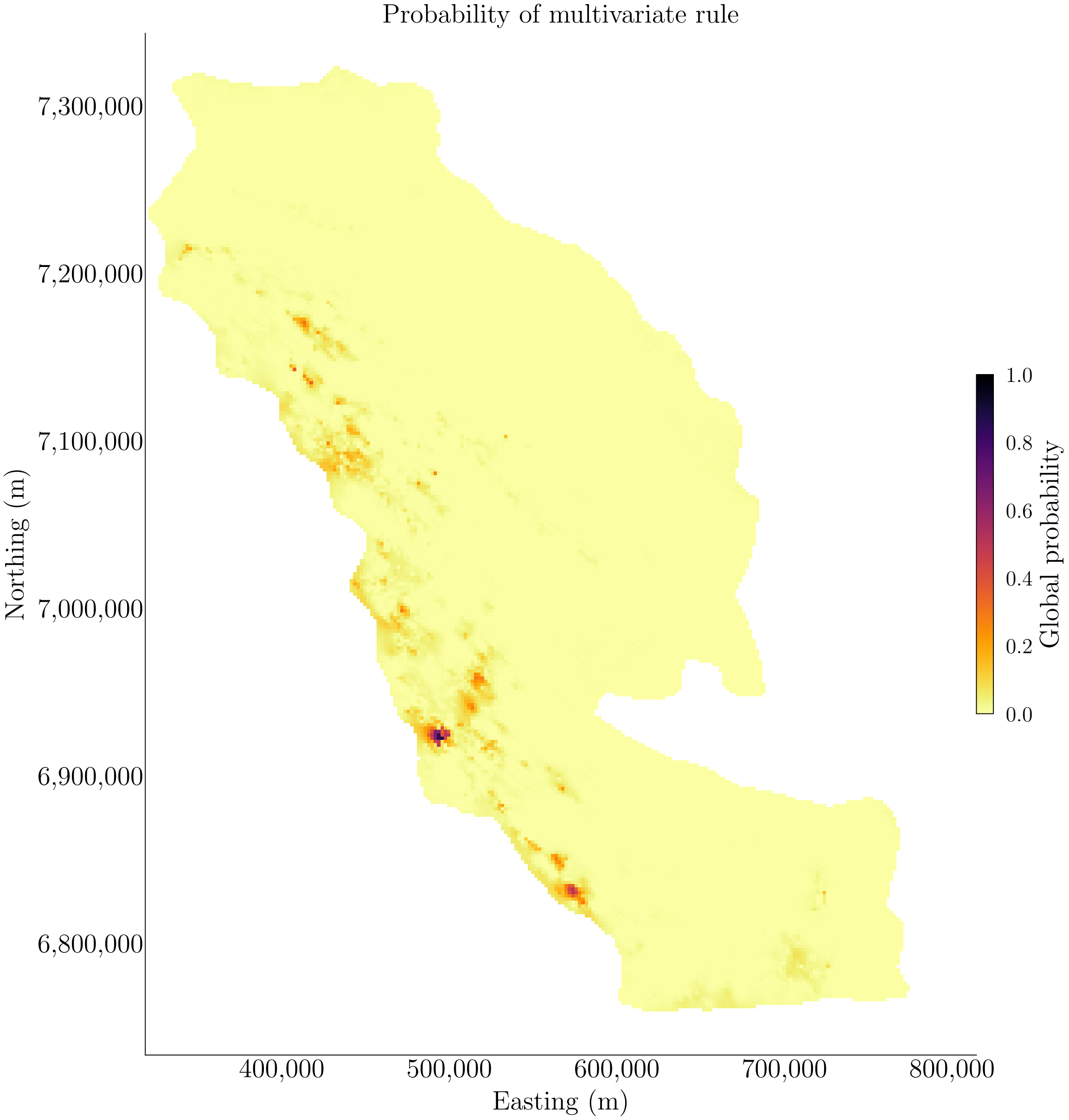


Figure 3.11: Map of the calculated multivariate joint-probability. The global probability is plotted. Probability values are on a 0 to 1 scale.

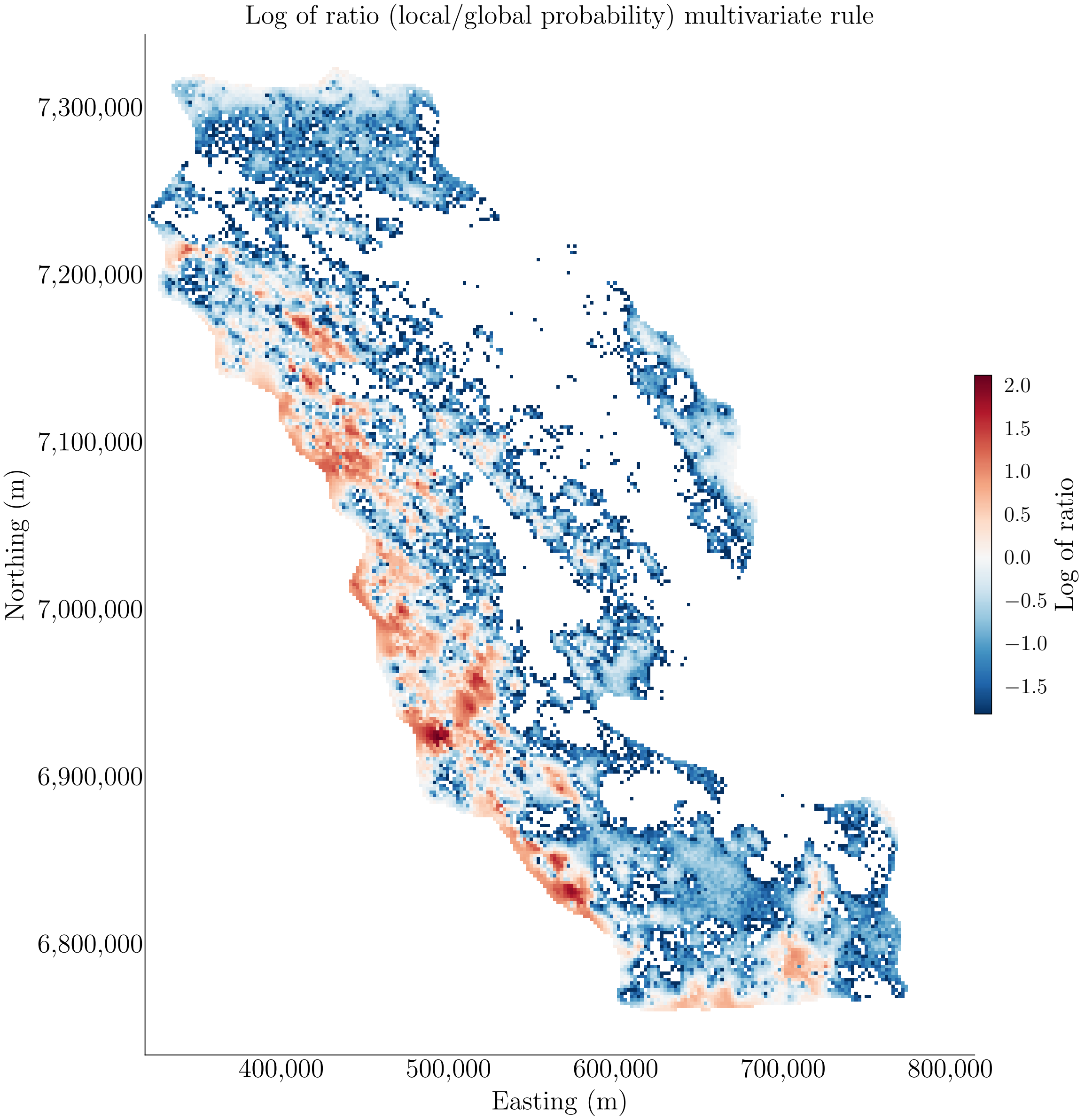


Figure 3.12: Map of the calculated multivariate joint-probability. The logarithm to the base 10 of the ratio of local to global probability is plotted. Negative values are locations where the local probability is below the global average probability. Positive values are all locations where the local probability is greater than the average. Locations with a global probability of 0 are not plotted because the logarithm cannot be calculated.

is straightforward and useful. The PostPPMT methodology provides a solution to this problem and a starting point for more complicated multifactor and extreme value criteria.

Chapter 4

Decomposition of multivariate spatial data into latent variables

Simulating spatial Gaussian realizations is one of the core components of geostatistics and numerous other fields involving uncertainty, risk, and reliability. The use of decorrelation methods and truncated Gaussian algorithms further promotes the use of Gaussian realizations. In multivariate cases, geological variables may represent different scales and exhibit different spatial structures and anisotropy that complicates decorrelation methods. For cases where spatial dependencies cannot be removed using techniques such as the projection pursuit multivariate transform coupled with maximum autocorrelation factors, cokriging is advocated and requires a linear model of coregionalization (LMC). However, blind source separation represents the original multivariate problem as a linear combination of latent source variables, each one having a spatial structure from the LMC. The latent source variables or factors are independent and follow a standard normal distribution facilitating the use of Gaussian simulation algorithms. Moreover, different algorithms may be utilized for each variable given that some are more efficient at generating realizations for different spatial covariance functions. Recovering the original variables afterwards is straightforward. The theory for this decomposition is presented in this chapter. A small numerical example is used to explain the theory. Limitations of the method are discussed.

4.1 Motivation

Geostatistical conditional simulation is used to generate stochastic realizations of the joint spatial variability between variables and it is in the core of modern multivariate modeling (Bailey and Krzanowski, 2012; Chiles and Delfiner, 2012; Wackernagel, 2003). Decorrelation techniques such as the stepwise conditional transformation (Leuangthong and Deutsch, 2003), minimum/maximum autocorrelation factors (MAF) (Desbarats and Dimitrakopoulos, 2000), and projection pursuit multivariate transform (Barnett et al., 2014) are used to decorrelate non-Gaussian relationships between variables for independent simulation. The application of these techniques in geostatistical modeling has grown in the past years with the advancement of algorithms and computer architectures. Such techniques have limitations. A notable one is the requirement for homotopic data, that is, all variables must be available at all sample locations. The requirement for homotopic data is particularly limiting in mining applications since multiple data sources are common, for example, data may originate from diamond drilling, reverse circulation drilling, and blast hole drilling. Another limitation is that decorrelation techniques do not necessarily account for spatial cross covariance. MAF is intended to mitigate cross covariance at a specific lag that, under appropriate circumstances, mitigates cross covariance at all lags; however, it is not always successful. Ignoring residual cross covariance, even if it appears insignificant, can lead to complications with variogram reproduction in geostatistical simulation workflows.

In the presence of missing data, unequal sampling, or different data types, variables are typically cosimulated with some variation of cokriging. This is usually the case when the variables of primary interest are sparsely sampled and one or more densely sampled secondary variables are available that relate to the primary variables being simulated. Markov-type coregionalization models used in collocated cokriging are an alternative to the Linear Model of Coregionalization (LMC) that is required for cokriging (Almeida, 1994; Rivoirard, 2001). However, common implementations of collocated cokriging leads to variance inflation (Babak and Deutsch, 2009). They present a more appropriate technique called intrinsic collocated cokriging that does not lead to variance inflation and expands the range of application of techniques that make use of collocated data. In cases with multiple secondary variables, the problem is simplified by merging them together into a super-secondary variable (Babak and Deutsch, 2008).

Simultaneous modeling of multiple primary variables requires an LMC. Despite the challenge of modeling an LMC with a large number of variables, it remains a useful and mathematically flexible tool. A model of coregionalization like the LMC is required to combine multiple data types measured at different locations and different data support into the same framework, and in cases where decorrelation techniques do not successfully remove spatial cross-covariance. Different variables and data types having sample-specific measurement errors can be used together for estimation with cokriging (Goovaerts, 1997). The classic approach is to fit all direct and cross variograms with the LMC (Chiles and Delfiner, 2012). The LMC is then utilized in cokriging and simulation. The approach proposed here is to decompose or factorize the normal score transform of the original variables into a set of independent normal latent variables using the LMC and blind source separation (BSS) (Schmidt, 2009). This approach is based on the definition of the LMC and therefore accounts for the direct and cross covariance of the original variables. Source variables or factors may then be modeled independently with an appropriate simulation algorithm and used to reconstruct the original variables.

In the case of completely unequally sampled data, that is, when two or more data types are available and are never collocated, simultaneous modeling of the variables is restricted to the LMC. In this context, BSS may be used to generate factors at the locations of all data types facilitating independent simulation. BSS provides a way to simulate factors at all data locations that combined with the LMC creates a framework to compute missing variable values at those locations. The modeling of independent factors also permits practical and easy model checking at each step. The factors can be analysed and checked independently, as opposed to cosimulation with cokriging in which model checking is quite challenging for a large number of variables. Another convenience of BSS is that the process of simulating the factors and computing the variables is highly parallelized, not only across factors but also across variables and structures.

There are many algorithms for simulation of Gaussian variables including sequential Gaussian simulation, turning bands, moving average methods, random coins, spectral methods, circulant embedding, and matrix methods that rely on Cholesky factorization (Borgman et al., 1984; Chiles and Delfiner, 2012; Dietrich and Newsam, 1997; Emery, 2008; Emery and Lantuejoul, 2006; Goovaerts, 1997; Kyriakidis, 1999; Mantoglou, 1987; Mantoglou and Wilson, 1982; Matheron, 1973; Oliver, 1995; Oliver et al., 2008; Paravarzar et al., 2015; Pardo-Iguzquiza and Chica-Olmo, 1993; Wackernagel, 2003; Yao, 1998b). Each algorithm has a range of spatial covariance functions and grid parameters where they perform with high efficiency and robustness in terms of variogram and histogram reproduction. Given that the factors from BSS are independent and have a single spatial covariance function, the most appropriate algorithm may be selected and applied to each factor independently. For example, moving average methods could be applied to factors with short range spherical covariance functions, while spectral methods could be applied to factors having an exponential structure.

Most simulation algorithms are unconditional and the resulting realizations could be conditioned by (co)kriging after unconditional simulation. Once the factors have been simulated and conditioned, the original variables are computed from the definition of the LMC.

4.2 Latent factors

Recall the LMC equation presented in Section 2.2:

$$Z_k(\mathbf{u}) = m_k + \sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u}) \tag{4.1}$$

Where m_k is the stationary mean of the k^{th} variable, nst is the number of structures or factors, $a_{k,i}$ are the coefficients explaining the contribution of the i^{th} factor to the k^{th} variable and $Y_i(\mathbf{u})$ are independent factors defined by single spatial covariance structure, with the 0^{th} factor representing the nugget effect component.

One approach from the early days of geostatistics, is to unconditionally simulate each of the $(nst + 1)$ factors. The factors can be combined with the $a_{k,i}$ coefficients into unconditional realizations of the K variables (Journel and Huijbregts, 1978; Matheron, 1979; Myers, 1982). Then, the realizations can be conditioned by (co)kriging. The simulation of each of the $(nst + 1)$ factors could be done with the best technique for each factor; there is no need to use the same technique for each, then conditioning could be done by global dual kriging for efficient assembly of the final realizations.

Decomposition of data into factors

Consider decomposing the original variables into their underlying factors so that the factors could be kriged and or simulated independently and values reconstructed at the end. The idea is to impute factor data that reproduce the original data and that have the correct spatial structure. If the data are equally sampled, this could be denoted as:

$$Z = AY \tag{4.2}$$

where A is the matrix with the a coefficients, Z is the matrix with the coregionalized variables Z_k, and Y is the matrix with the Y_i independent factors. This development is restricted to cases when there are more factors than there are original variables, that is, $(nst+1) > K$. This is common since a large number of factors would be required to explain the complexity of multivariate spatial data.

The Z and Y data contain equivalent information since they are linked by Equation 4.1. There may be advantages to this decomposition including: (1) the factors are independent, that is, they can be kriged or simulated independently, and (2) unequally sampled data, that is, locations where subsets of the K variables are available could be considered with the $(nst + 1)$ variables carrying the information from each subset.

Inference of the Y values from the available Z data is presented here as an inverse problem; we know the result of the linear combination, but not the factors that went into the combination. The solution is non-unique when $(nst + 1) > K$, which is the case for small K or large nst; the case considered here. The inverse problem is non-unique; there are more y variables to derive than z variables to constrain the results. A simulation framework is developed to create realizations of the factor values and constrain each geostatistical realization with a different data realization. This is similar in principle to the multiple imputation framework (Rubin, 1996). We denote the data:

$$(z_k(\mathbf{u}_j), k = 1, ..., K; j = 1, ..., n) \rightarrow \begin{cases} (y_i^{(1)}(\mathbf{u}_j), i = 0, ..., nst; j = 1, ..., n) & \mathbf{u} \text{ in } D \\ (y_i^{(2)}(\mathbf{u}_j), i = 0, ..., nst; j = 1, ..., n) & \mathbf{u} \text{ in } D \\ \vdots & \\ (y_i^{(L)}(\mathbf{u}_j), i = 0, ..., nst; j = 1, ..., n) & \mathbf{u} \text{ in } D \end{cases} \tag{4.3}$$

where L is the number of realizations. The challenge now is to simulate the underlying latent independent factors at locations where one or more of the original data are observed.

4.3 Blind source separation

Blind source separation (BSS) methods are used to process mixed sensor observations and infer the most probable source estimates, without or with limited information about the source signal. Independent component analysis (ICA), PCA, and SVD are the most used techniques for BSS to reveal the hidden factors in the observed signals. BSS methods have applications in image processing, medical imaging, music, wavelets and signal processing, speech recognition, telecommunications, and machine learning (Comon and Jutten, 2010; Naik and Wang, 2014; Yu et al., 2014).

BSS is often explained with the cocktail party problem (Choi and Cichocki, 1997; Handel, 1989). Consider a party with people talking simultaneously. There might also be some background noise such as music or sound coming from outside the room. Consider the analysis of the sound recorded by microphones installed in the room. The problem is to separate the mixture of sound signals with no a-priori knowledge of the sources. This kind of problem is referred as to as blind source separation. BSS aims at separate the signals based on their mixture only, without accessing the signal themselves, hence the term blind. In most applications of BSS, the signals are assumed to be stationary and have zero-mean, the sources are statistically independent, and the number of sensors (observations) exceeds or equal the number of sources (Kofidis, 2016).

The paradigm of BSS is that some unknown matrix of sources Z is mixed by some linear matrix of constants A, also referred as to the mixing coefficient matrix. The sources are projected from the original source space to an observation space X in which the signals are obtained:

$$X = AZ \tag{4.4}$$

where X is the mixed signal matrix. Both source signal Z and how the source signal are mixed are unknown. The goal of BSS is to calculate a demixing matrix $W \approx A^{-1}$ such that

$$Y = WX = \hat{Z} \tag{4.5}$$

is a good estimation and approximation of the real source signal Z. To achieve this, the observations in X are transposed into an estimated source space in which the estimates of the sources, Y are projected. Such projection highlights different patterns in the data along different projection axes. This allows filtering solutions of Y that corresponds to noise signals or unwanted solutions. This projection or transformation can be done with orthogonalization. ICA and SVD are used to perform BSS and estimate W, since both methods assume linear independence between the sources and yield statistically independent estimated Y (Clifford, 2008). SVD attempts to find an independent set of vectors onto which data is transformed, that is, the principal components of a multi-dimensional signal. SVD separates the signal into a subspace of signal and another of noise. Maximum independence between these subspaces is achieved by requiring them to be orthogonal. The projected data onto each vector are the independent sources. Once these vectors are discovered, they are used to calculate the inverse W^{-1} of the demixing matrix and reproject the data back into the observation space.

Problem illustration

The following example is used to illustrate the problem and early motivation to the development of the proposed methodology. At a location $\mathbf{u}$ consider the following LMC:

$$\begin{cases} Z_1 = a_{1,1}Y_1 + a_{1,2}Y_2 \\ Z_2 = a_{2,1}Y_1 + a_{2,2}Y_2 \end{cases} \tag{4.6}$$

The Z variables are standard normal with correlation equal to $\rho_{1,2} = a_{1,1}a_{2,1} + a_{1,2}a_{2,2}$ (Equation 2.18). This model is very constrained since Equation 4.6 imposes $a_{1,1}^2 + a_{1,2}^2 = 1$ and $a_{2,1}^2 + a_{2,2}^2 = 1$ as both Y and Z have variance of 1. Equation 4.6 can be reorganized such as the factors are isolated from the variables:

$$Y_1 = \frac{Z_1 - \frac{a_{1,2}}{a_{2,2}}Z_2}{a_{1,1} - \frac{a_{1,2}}{a_{2,2}}a_{2,1}} \tag{4.7}$$

$$Y_2 = \frac{Z_2 - \frac{a_{2,1}}{a_{1,1}}Z_1}{a_{2,2} - \frac{a_{2,1}}{a_{1,1}}a_{1,2}} \tag{4.8}$$

The equations above are used to compute the variables from the factors or the factors from the variables. They define the linear relationship between the bivariate distributions shown in Figure 4.1.

In the context of unequally sampled data, one of the variables is missing and a unique observation of Z_1 or Z_2 is available. The available variable defines the conditional distribution of the other. For example, for an observed value $Z_1 = z_1$ the conditional distribution $Z_2|Z_1 = z_1$ has a mean equal to $\mu_{Z_2} = \rho_{1,2} \times z_1$ and variance $\sigma_{Z_2}^2 = 1 - \rho_{1,2}^2$. This conditional distribution in the Z space defines the space of uncertainty of Z_2. It also defines the space of uncertainty of Y since Y

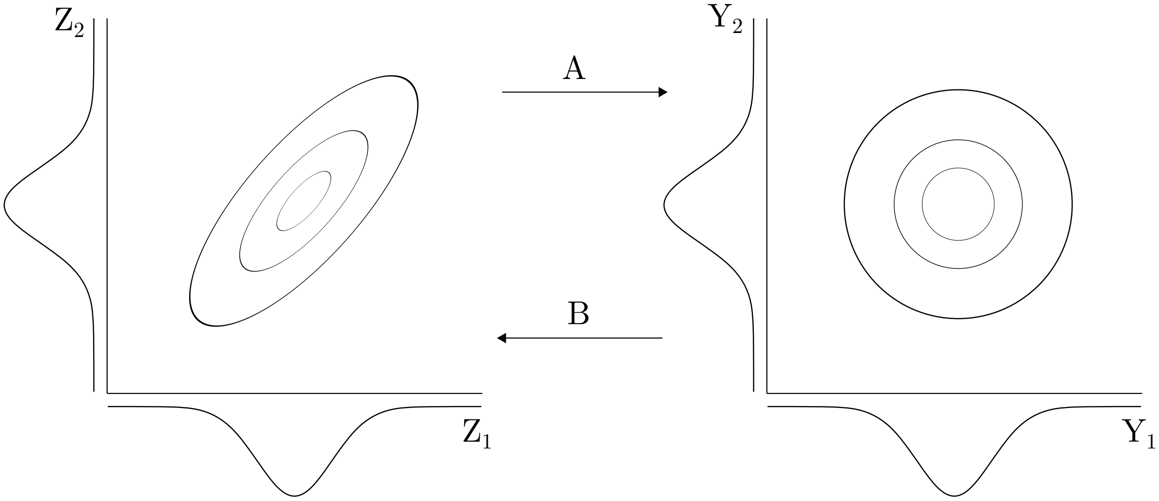
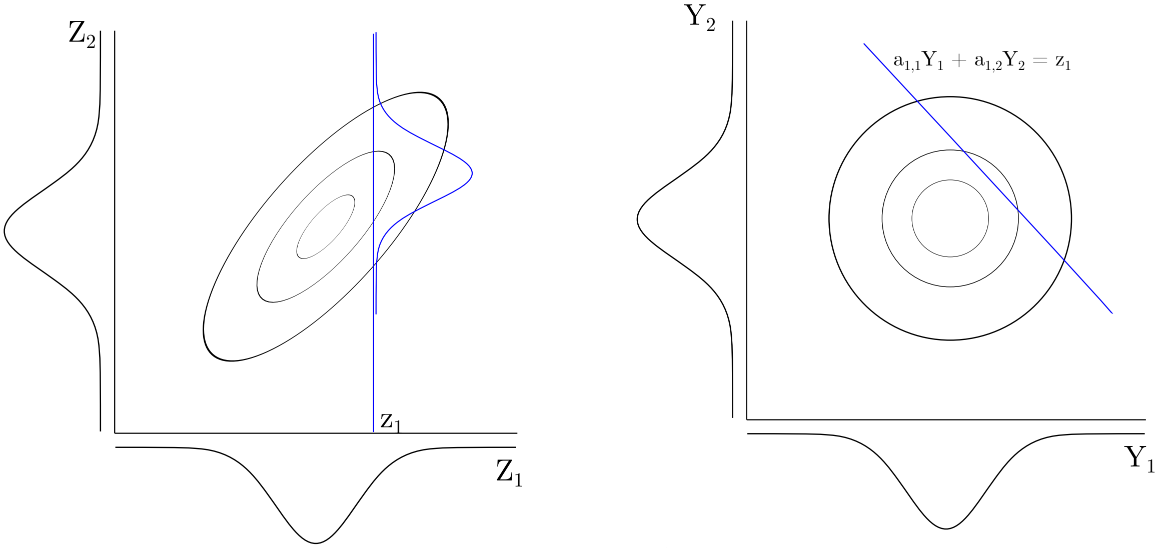


Figure 4.1: The bivariate relationships between the Z variables (left) and Y factors (right) are illustrated. They are linked by paths A (Equations 4.7 and 4.8) and B (Equations 4.6).

is constrained to Z by the relationship in Equation 4.6. For the observed value $Z_1 = z_1$, this space of uncertainty is a plane whose intersection with the XY plane is a line with equation $a_{1,1}Y_1 + a_{1,2}Y_2 = z_1$, see Figure 4.2.



(a) Space of uncertainty in Z space

(b) Corresponding space of uncertainty in Y space

Figure 4.2: The conditional distribution $Z_2|Z_1 = z_1$ (a) defines the space of uncertainty of Y (b) since the variables are linked by Equation 4.6. The space of uncertainty in Y is defined by the line $a_{1,1}Y_1 + a_{1,2}Y_2 = z_1$ (b).

We are interested in sampling either conditional distribution $Y_1|Z_1 = z_1$ or $Y_2|Z_1 = z_1$ and calculate the other. These conditional distributions in Y are defined by available Z and are linearly constrained by the LMC. For convenience, the conditional distributions in Y must be standard normal and independent. BSS provides a solution to this problem by reducing the space of uncertainty to a subspace where all vectors are orthogonal. It constrains the sampling space of the multivariate Gaussian distribution to a region that can be sampled to generate realizations of the Y factors such that the Z are reproduced. In this example with two variables, this space is illustrated in Figure 4.3.

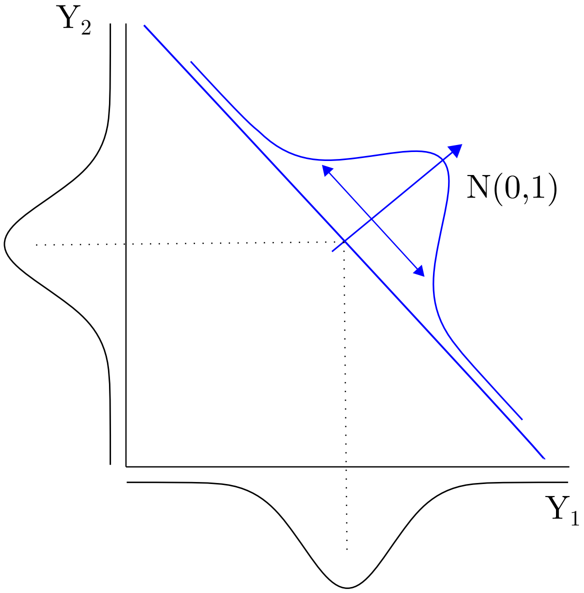


Figure 4.3: BSS is used to calculate the space of uncertainty shown in Figure 4.2b.

Given heterotopic samples of Z and a valid LMC, BSS is used in the proposed methodology to constrain the sampling space of the multivariate Gaussian distributions such that the Y factors are imputed at locations where at least one variable Z is measured. The Y values must be valid solutions of Equation 4.1, must be standard normal, and have the same spatial structure from the LMC. The theory of the proposed methodology is discussed with more details in the next section.

4.4 Theory

The process of simulating y values is equivalent to generating random samples from a multivariate Gaussian distribution subject to linear equality constraints. This is an essential component in matrix factorization methods used in BSS (Schmidt, 2009). The following theory was adapted from Schmidt (2009) to the proposed independent factor simulation (IFS) methodology. Assuming that $rank(A) > K$ and using matrix notation, the problem consists of solving the following underdetermined system of equations:

$$Z = AY$$

Consider generating random samples $y \in D$ from a multivariate Gaussian density $p(y)$ subject to the linear equality:

$$p(y) \propto \begin{cases} \mathcal{N}(y \mid \mu_y, \Sigma_y) & \text{if } z = Ay \\ 0 & otherwise \end{cases} \tag{4.9}$$

where μ_y is the mean and Σ_y the covariance matrix of the y variables. The equality constraints restrict the distribution of $p(y)$ to an affine subspace of the multivariate Gaussian space. This distribution can be mapped onto this subspace by computing an orthonormal basis T and its orthogonal complement $T_\perp$ for the constraints using singular value decomposition (SVD) (Golub and Reinsch, 1970):

$$A = USV^\top = US \begin{bmatrix} T \\ T_\perp \end{bmatrix} \tag{4.10}$$

Given a minimum norm solution, y_0, to $Z = AY$, which is obtained using cokriging, a random variable x that is orthogonal to the residuals $y - y_0$ is defined:

$$x = T_\perp(y - y_0) \tag{4.11}$$

The minimum norm solution is equivalent to the linear combination closest to the true unknown value measured with a least squared norm (minimum mean squared error). Since the variables have a known covariance function, the solution is equivalent to cokriging that minimizes the estimation variance and hence minimizes the mean squared error. The dual form of cokriging is applicable when the number of samples multiplied by the number of variables is not excessive. The cokriging solution (Equations 4.12) requires the calculation of the covariance matrix Σ_z between z values, and the cross covariance matrix Σ_{yz} between y and z values. These covariances are calculated from the Equations 2.19 and 2.20 of the LMC.

$$\begin{cases} \Sigma_z = A^T \Sigma_y A \\ \Sigma_{yz} = \Sigma_y A^T \\ y_0 = \Sigma_{yz} \Sigma_z^{-1} z \end{cases} \tag{4.12}$$

The conditional distribution of x given the equality constraint is Gaussian with the following mean and covariance matrix

$$p(x \mid z = Ay) \propto \mathcal{N}(x \mid \mu_x, \Sigma_x) \tag{4.13}$$

$$\begin{cases} \mu_x = \Lambda(\mu_y - y_0) \\ \Sigma_x = \Lambda \Sigma_y T_\perp^\top \end{cases} \tag{4.14}$$

where $\Lambda = T_\perp(I - \Sigma_y T^\top (T \Sigma_y T^\top)^{-1} T)$. Given the Cholesky decomposition $LL^T = \Sigma_x$, a vector, r, of uncorrelated Gaussian random variables is defined:

$$r = L^{-1}(x - \mu_x) \tag{4.15}$$

Sampling from this distribution is used to generate realizations of the original variable y by substituting them into Equation 4.16 (from equations 4.11 and 4.15):

$$y = T_\perp^\top (L^T r + \mu_x) + y_0 \tag{4.16}$$

The resulting realizations of the same y variable are equally likely to be drawn from the constrained multivariate Gaussian distribution and for this reason realizations of the same factor are not independent. However, the different Y values may be independent, there is no cross-correlation between the values of the factors. These values follow a standard normal distribution, have the correct spatial structure, and satisfy $Z = AY$.

Summary The methodology described above can be summarized by the following steps. These steps will be illustrated by a small example in the next section.

1. Fit an LMC to the normal scores of z, which yields A.

2. Use the LMC to calculate Σ_y matrix.

3. Use Σ_y and A matrices to calculate the Σ_z and Σ_{yz} and solve the cokriging equation for y_0, Equations 4.12.

4. Decompose the matrix A with SVD to compute T and $T_\perp$.

5. Calculate Λ and Σ_x as in Equations 4.14.

6. Decompose Σ_x with Cholesky decomposition and calculate L.

7. Generate a vector with random standard Gaussian values r and calculate y with Equation 4.16.

Once the Y factors are imputed at data locations, their simulation at grid locations will call for a series of univariate conditional simulations, that is, conditional to the Y at data locations, observed or imputed.

4.5 Small example and implementation details

In this small example, the theory and equations from the previous section are used to illustrate the simulation of the LMC factors and then reconstruct the variables at the data locations.

Problem setup

Consider a realization of two standard normal random variables Z_1 and Z_2 at locations $\mathbf{u}_1$ and $\mathbf{u}_2$ with an arbitrary distance of $|\mathbf{u}_1 - \mathbf{u}_2| = 1$ unit:

$$\begin{bmatrix} z_1(\mathbf{u}_1) & z_1(\mathbf{u}_2) \\ z_2(\mathbf{u}_1) & z_2(\mathbf{u}_2) \end{bmatrix} = \begin{bmatrix} 0.146 & -1.207 \\ -0.264 & 1.155 \end{bmatrix}$$

A LMC is assumed with no nugget effect and four nested isotropic spherical structures $i = 1, ..., 4$ with respective ranges of 2, 4, 7, and 10 units:

$$\begin{cases} \gamma_{1,1}(\mathbf{h}) = 0.711\gamma_1(\mathbf{h}) + 0.254\gamma_2(\mathbf{h}) + 0.028\gamma_3(\mathbf{h}) + 0.007\gamma_4(\mathbf{h}) \\ \gamma_{1,2}(\mathbf{h}) = 0.293\gamma_1(\mathbf{h}) + 0.175\gamma_2(\mathbf{h}) + 0.146\gamma_3(\mathbf{h}) + 0.007\gamma_4(\mathbf{h}) \\ \gamma_{2,2}(\mathbf{h}) = 0.120\gamma_1(\mathbf{h}) + 0.120\gamma_2(\mathbf{h}) + 0.752\gamma_3(\mathbf{h}) + 0.008\gamma_4(\mathbf{h}) \end{cases}$$

This is a valid LMC model that respects the constraints in Equations 2.16 and 2.17. The A matrix is given by:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & a_{2,2} & a_{2,3} & a_{2,4} \end{bmatrix} = \begin{bmatrix} 0.843 & 0.504 & 0.168 & 0.084 \\ 0.347 & 0.347 & 0.867 & 0.087 \end{bmatrix}$$

Let us reshape matrix A to a new matrix A with dimensions equal to $(nvar \times nloc)$ rows and $(nfac \times nloc)$ columns, where $nvar$ is the number of variables, $nloc$ the number of locations, and $nfac$ the number of factors. The system of equations $Z = AY$ for this problem is then given in matrix form by Equation 4.17:

$$\begin{bmatrix} z_1(\mathbf{u}_1) \\ z_2(\mathbf{u}_1) \\ z_1(\mathbf{u}_2) \\ z_2(\mathbf{u}_2) \end{bmatrix} = \begin{bmatrix} 0.843 & 0.504 & 0.168 & 0.084 & 0 & 0 & 0 & 0 \\ 0.347 & 0.347 & 0.867 & 0.087 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.843 & 0.504 & 0.168 & 0.084 \\ 0 & 0 & 0 & 0 & 0.347 & 0.347 & 0.867 & 0.087 \end{bmatrix} \begin{bmatrix} y_1(\mathbf{u}_1) \\ y_2(\mathbf{u}_1) \\ y_3(\mathbf{u}_1) \\ y_4(\mathbf{u}_1) \\ y_1(\mathbf{u}_2) \\ y_2(\mathbf{u}_2) \\ y_3(\mathbf{u}_2) \\ y_4(\mathbf{u}_2) \end{bmatrix} \tag{4.17}$$

This system of equations is underdetermined, with two z data and four y values at each location. An infinite number of solutions exist for this system of equations.

Covariances and cokriging

The required covariance matrices for cokriging are calculated as in the system of equations 4.12:

$$\Sigma_y = \begin{bmatrix} C_1 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 \\ 0 & 0 & C_3 & 0 \\ 0 & 0 & 0 & C_4 \end{bmatrix} = \begin{bmatrix} 1.000 & 0.312 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.312 & 1.000 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.000 & 0.633 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.633 & 1.000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.000 & 0.787 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.787 & 1.000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.000 & 0.851 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.851 & 1.000 \end{bmatrix}$$

$$\Sigma_{yz} = \Sigma_y A^T = \begin{bmatrix} 0.843 & 0.263 & 0.347 & 0.108 \\ 0.263 & 0.843 & 0.108 & 0.347 \\ 0.504 & 0.319 & 0.347 & 0.220 \\ 0.319 & 0.504 & 0.220 & 0.347 \\ 0.168 & 0.132 & 0.867 & 0.682 \\ 0.132 & 0.168 & 0.682 & 0.867 \\ 0.084 & 0.071 & 0.087 & 0.074 \\ 0.071 & 0.084 & 0.074 & 0.087 \end{bmatrix}$$

$$\Sigma_z = A\Sigma_y A^T = \begin{bmatrix} 1.000 & 0.411 & 0.620 & 0.323 \\ 0.411 & 1.000 & 0.323 & 0.620 \\ 0.620 & 0.323 & 1.000 & 0.712 \\ 0.323 & 0.620 & 0.712 & 1.000 \end{bmatrix}$$

The matrix with the cokriging solution $y_0 = \Sigma_{yz}\Sigma_z^{-1}z$ is calculated and reshaped to match the size of the original A matrix:

$$y_0 = \begin{bmatrix} y_{0,1}(\mathbf{u}_1) & y_{0,2}(\mathbf{u}_1) & y_{0,3}(\mathbf{u}_1) & y_{0,4}(\mathbf{u}_1) \\ y_{0,1}(\mathbf{u}_2) & y_{0,2}(\mathbf{u}_2) & y_{0,3}(\mathbf{u}_2) & y_{0,4}(\mathbf{u}_2) \end{bmatrix} = \begin{bmatrix} 0.452 & -0.343 & -0.343 & -0.053 \\ -1.705 & -0.251 & 2.113 & 0.020 \end{bmatrix}$$

The cokriging estimates $z^* = Ay_0^\top = z$ are verified to ensure numerical consistency:

$$\begin{bmatrix} z_1^*(\mathbf{u}_1) & z_1^*(\mathbf{u}_2) \\ z_2^*(\mathbf{u}_1) & z_2^*(\mathbf{u}_2) \end{bmatrix} = \begin{bmatrix} 0.843 & 0.504 & 0.168 & 0.084 \\ 0.347 & 0.347 & 0.867 & 0.087 \end{bmatrix} \begin{bmatrix} 0.452 & -1.705 \\ -0.343 & -0.251 \\ -0.343 & 2.113 \\ -0.053 & 0.020 \end{bmatrix}$$
$$= \begin{bmatrix} 0.146 & -1.207 \\ -0.264 & 1.155 \end{bmatrix}$$

Note that the cokriging variance is not used anywhere in the BSS theory, therefore the dual form of cokriging is used to calculate the minimum norm solution.

Singular value decomposition

The null spaces of the big matrix A are calculated with SVD and defined based on the rank rk of A (Equation 4.10). The first $rk = 4$ rows of V form the orthonormal basis T, whereas the remaining rows form the orthogonal complement $T_\perp$:

$$\begin{aligned}
A &= USV \\
&= US \begin{bmatrix} T \\ T_\perp \end{bmatrix} \\
&= US \begin{bmatrix}
0 & 0.661 & 0 & 0.473 & 0 & 0.575 & 0 & 0.095 \\
-0.661 & 0 & -0.473 & 0 & -0.575 & 0 & -0.095 & 0 \\
0.569 & 0 & 0.180 & 0 & -0.802 & 0 & -0.003 & 0 \\
0 & -0.569 & 0 & -0.180 & 0 & 0.802 & 0 & 0.003 \\
0.486 & 0 & -0.860 & 0 & 0.152 & 0 & -0.027 & 0 \\
0 & 0.486 & 0 & -0.860 & 0 & 0.152 & 0 & -0.027 \\
-0.048 & 0 & -0.068 & 0 & -0.054 & 0 & 0.995 & 0 \\
0 & -0.048 & 0 & -0.068 & 0 & -0.054 & 0 & 0.995
\end{bmatrix}
\end{aligned}$$

The matrices Σ_y, T, and $T_\perp$ are used in the calculation of Λ, μ_x, and Σ_x, as in Equation 4.14:

$$\Lambda = \begin{bmatrix}
0.430 & 0.144 & -0.882 & 0.067 & 0.210 & -0.078 & -0.028 & 0.007 \\
0.144 & 0.430 & 0.067 & -0.882 & -0.078 & 0.210 & 0.007 & -0.028 \\
-0.036 & -0.030 & -0.060 & -0.019 & -0.044 & -0.011 & 0.997 & -0.003 \\
-0.030 & -0.036 & -0.019 & -0.060 & -0.011 & -0.044 & -0.003 & 0.997
\end{bmatrix}$$

$$\Sigma_x = \begin{bmatrix}
0.976 & 0.571 & 0.004 & -0.001 \\
0.571 & 0.976 & -0.001 & 0.004 \\
0.004 & -0.001 & 0.999 & 0.849 \\
-0.001 & 0.004 & 0.849 & 0.999
\end{bmatrix}$$

Since the mean of the y is zero, the mean of x is $\mu_x = \Lambda y_0^\top = 0$.

Simulation of the factors

Two realizations of the Y factors are generated with Equation 4.16. The lower triangular matrix L is computed from the Cholesky decomposition of Σ_x:

$$L = \begin{bmatrix} 0.988 & 0 & 0 & 0 \\ 0.578 & 0.801 & 0 & 0 \\ 0.004 & -0.005 & 0.999 & 0 \\ -0.001 & 0.006 & 0.849 & 0.527 \end{bmatrix}$$

Let $r^{(1)}$ and $r^{(2)}$ be two different realizations of a vector with random samples generated from a standard Gaussian distribution:

$$r^{(1)} = \begin{bmatrix} -0.355 \\ 1.909 \\ 0.593 \\ -1.076 \end{bmatrix} \quad ; \quad r^{(2)} = \begin{bmatrix} -0.260 \\ 0.129 \\ -0.698 \\ 1.482 \end{bmatrix}$$

Equation 4.16 is evaluated for $r^{(1)}$ and $r^{(2)}$ to generate two realizations $y^{(1)}$ and $y^{(2)}$ of the Y factors:

$$\begin{cases} y^1 = \begin{bmatrix} 0.254 & -0.081 & -0.427 & 0.536 & -1.058 & -1.386 & 2.317 & -0.066 \end{bmatrix} \\ \\ y^2 = \begin{bmatrix} 0.361 & -0.075 & -0.344 & -0.742 & -1.737 & -0.224 & 2.096 & 0.210 \end{bmatrix} \end{cases}$$

Substituting these two vectors into Equation 4.17 yields two realizations, $z^{(1)}$ and $z^{(2)}$, where $z^{(1)} = z^{(2)} = z$. The theory presented in the previous section is demonstrated in this small example. The computed $z^{(l)}$ values at the data locations are exactly the same of the input z values.

4.6 Limitations

One of the limitations of the BSS theory is regarding the nugget effect. The nugget effect is explained by (1) artificial error caused by sampling error, and an (2) intrinsic short scale variability caused by geologic factors. In the theory of sampling, the nugget effect is also referred as to the variance of strictly random fluctuations (Pitard, 2019). It includes the variance of the estimation errors, the variance of the intrinsic heterogeneity of the material, and it is always positive (Minnitt and Esbensen, 2017). The variance of the measurements is expected to be zero, that is, $V(0) = 0$. In fact, the variogram tends to zero when $\mathbf{h}$ tends to zero. In the theory of sampling, however, $V(0)$ represents the errors from the sampling, preparation, subsampling, and analytical errors. This variance is not zero and affects the shape of the variogram. The most effective way to estimate the nugget effect is by extrapolating the variogram values calculated between $\mathbf{h} = 0$ and the next lag(s). In geostatistics, this extrapolation is often done on the experimental variogram calculated in the most informative direction, e.g., the down-hole variogram (Deutsch, 2015a).

In the old view of the LMC there is no need to split the nugget effect (Chiles and Delfiner, 2012; Goovaerts, 1997). It is directly taken into account in the calculation of the covariances used for cokriging. In the new approach to the LMC introduced by the BSS methodology, each factor is isolated and simulated independently. Therefore, proper decomposition of the nugget effect requires the identification and isolation of all of its sources. In the proposed methodology, it is unrealistic to have one completely shared nugget. It is also unrealistic to have a different nugget effect for each variable. The most realistic scenario would consider the many possible combinations of the nugget effect, that is, the nugget that is shared across all variables, the nugget shared in subsets of the variables, and the independent truly variable specific nugget effect. Consider the variable gold measured from reverse circulation (RC) and diamond drilling (DDH). There are three different nugget effect to isolate, two drilling specific and a large shared nugget effect. Different nugget factors may also be considered for different mineralogy. For example, the copper minerals chalcopyrite ($CuFeS_2$) and chalcocite (Cu_2S) have a shared nugget effect between copper, and an independent nugget effect for each mineral.

The many combinations for the nugget effect for K variables taken r at a time is calculated as:

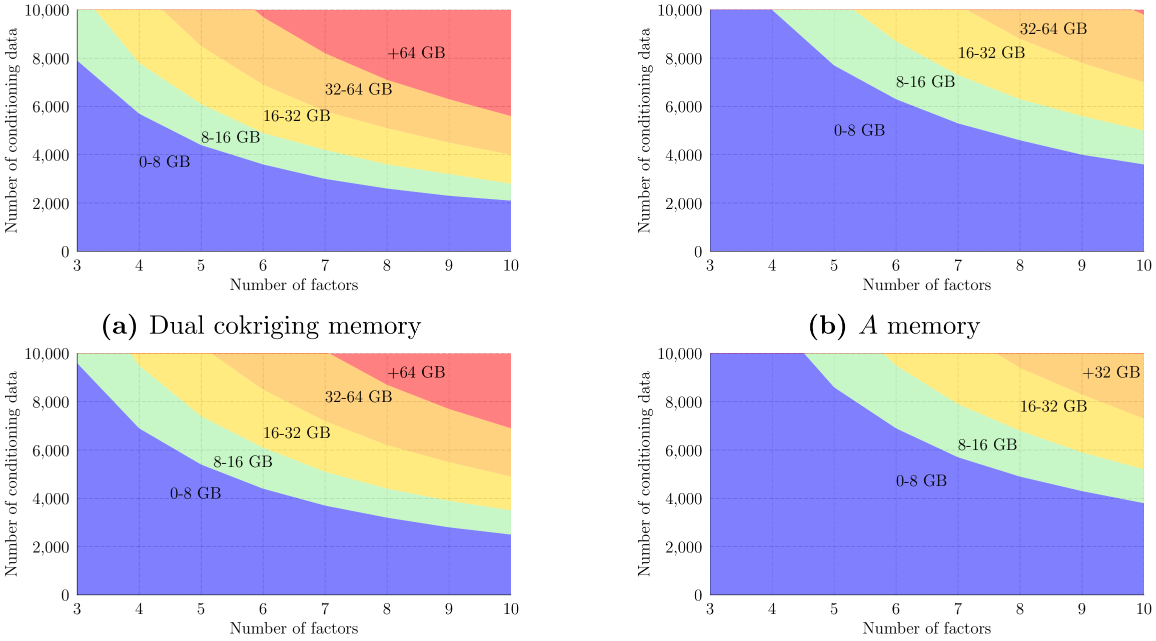
$$kCr = \frac{k!}{(k-r)!r!}$$

For any K number of variables the combined number of independent and shared nugget effect is calculated by $2^k - 1$. This is a large number for only a few variables. For example, for two different data types, and three variables, there would be a total of $2^6 - 1 = 63$ different nugget effect factors to be considered in the BSS decomposition. The nugget components could never be isolated from data itself and additional information and professional judgement must be used. In practice, the nugget effect is still isolated and simulated independently as a single unique factor.

Another limitation is related with the size of the model. Memory allocation for dual cokriging, SVD, and Cholesky decomposition becomes expensive as the number of variables, factors and locations grow. The methodology of BSS is implemented in the GSLIB programs `LMC_IMP` and `LMC_COMP` (There will be a link here to the Appendix). These programs, written in FORTRAN, are used to assess the memory required for each major operation in the methodology. The required memory is calculated based on the maximum memory that must be allocated inside each subroutine. The LAPACK library for FORTRAN is used for SVD, Cholesky decomposition, and to solve linear system of equations (Anderson et al., 1999). The computation of the memory considers that an $n \times m$ matrix with double-precision floats requires $n * m * 8/10^9$ gigabytes (GB) of memory RAM to be allocated.

Figure 4.4 shows the memory required for each operation for a different number of factors and conditioning data. Memory-wise the most expensive step of the implementation is dual cokriging, followed by SVD. Note that cokriging requires the calculation of the covariance matrices Σ_y, Σ_z, and the cross covariance Σ_{yz}

between y and z variables. Memory for SVD also depends on the dimensions of the big A matrix, that is, after reshaping as in the Equation 4.17 of the example. The memory required to allocate A is shown in Figure 4.4b. Another expensive operation is the Cholesky decomposition of the matrix Σ_x. The Σ_x and Σ_z matrices have the same dimensions, that is, $(nvar \times nloc) \times (nvar \times nloc)$.



(a) Dual cokriging memory

(b) A memory

(c) SVD memory

(d) Σ_x memory

Figure 4.4: Expected memory in gigabytes required to allocate all matrices for the operations given different number of data locations and factors.

SVD is the slowest operation in the methodology. While the memory required to perform SVD on a model with 10 factors and 2,000 conditioning data is less than 8 Gb, the expected run-time is more than 30 minutes, see Figure 4.5.

Despite the expensive computation of dual cokriging and SVD for large models, these operations are performed once. Practical implementation of the BSS methodology should perform cokriging and SVD upfront and store all matrices requires for simulation. Multiple realizations of the factors are generated with one single program call, as oppose to multiple parallel calls of the program.

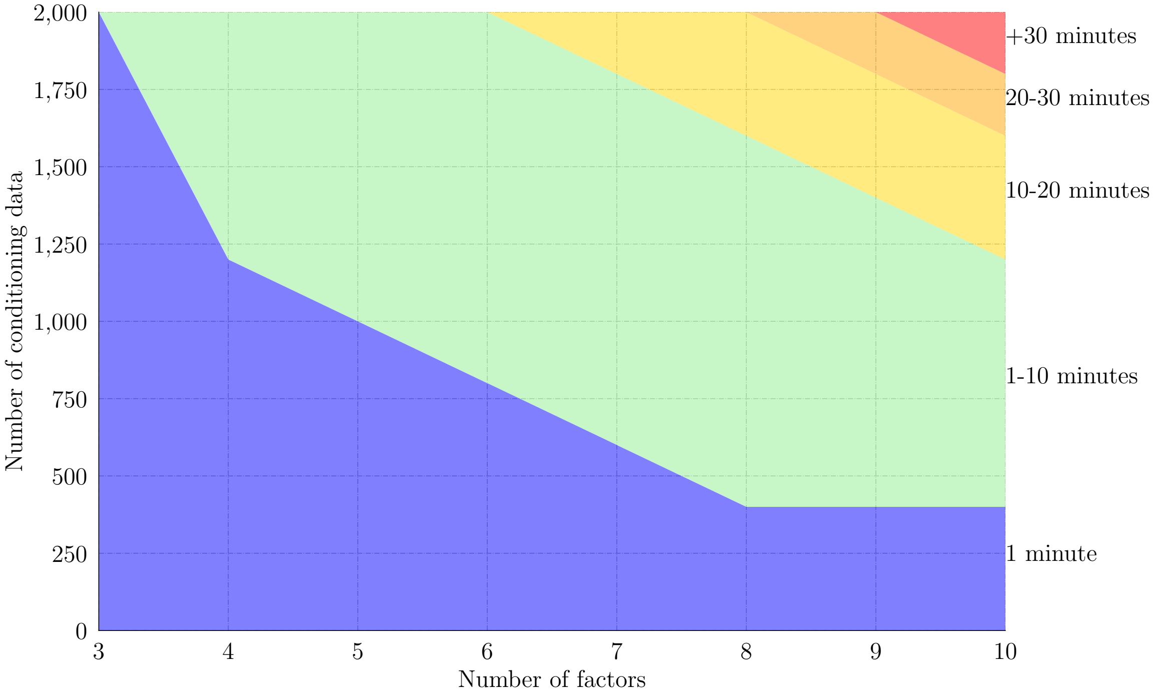


Figure 4.5: Expected run-time of the SVD operation given different number of data locations and factors.

4.7 Conclusion

An approach to simulate independent standard factors of the LMC is presented. This approach uses the theory of BSS to sample a constrained multivariate Gaussian distribution. The methodology relies on the SVD of the coefficients explaining the variance of each structure of the LMC to calculate the null spaces that are needed in further steps of the methodology. To satisfy the linear constraints, a vector with the minimum norm solution has to be calculated at the data locations that is obtained with dual cokriging. All covariance matrices required to solve the system of equations of dual cokriging are calculated from the LMC. A transformed variable is then calculated using the null spaces and the covariance matrix of the independent factors. Decorrelation of these factors is achieved with Cholesky decomposition of the latter covariance matrix. A vector of simulated LMC factors is then generated conditioned to the minimum norm solution, the null space, and the decomposed matrix. Original variables are recovered directed from the LMC at the sample locations or in a grid. The factors can be simulated independently. The variables can be computed at any locations where all factors have been simulated.

Chapter 5

Best practices of selection of simulation algorithm

The IFS methodology discussed in the previous chapter provides a framework for imputation and simulation of independent factors that are used to reconstruct the original variables. In a simulation context it offers an alternative to cosimulation with cokriging. The current approach is to choose an algorithm to simulate all variables. In the IFS methodology, the variables are computed from the simulated factors. The factors are independent, facilitating the use of Gaussian simulation algorithms. A different algorithm may be used to simulate each factor. The choice of the algorithm must account for the unique spatial covariance function of the factors, that is, the range of correlation, the structure type, and the anisotropy. This chapter discusses the practical aspects and best practices of selection of four common simulation algorithms in geostatistics: moving average (MA), sequential Gaussian simulation (SGS), turning bands (TB), and spectral simulation (SS). The non-conditional simulated mean, variance, and variogram are checked and compared against the theoretical expectations. The heavy mathematical foundation of these methods is put aside for a more hands-on analysis (Cabral Pinto and Deutsch, 2017a).

5.1 Motivation

The choice of the simulation algorithm depends on many factors, such as, the practical implementation and software availability, the target variogram function, and the modeler expertise and experience. Given software availability and assuming that all best practices are implemented in each algorithm, the choice of the simulation algorithm reduces to one goal: choosing the algorithm that best reproduces the input variogram.

The different nature and implementation of these algorithms yield to different reproduction of first and second order statistics. For example, the practice of SGS has shown that short range structures with a high contribution generates realizations with high average variance. Also regarding SGS, another observation from practice is that long range structures are simulated with less variability and the average simulated variogram is more continuous than the reference one. In the MA method, the reproduction of first and second order statistics depends on the number of data inside the window, for a reasonable large number of data the variogram is well reproduced. There are many spectral methods. The one reviewed in this thesis refers to spectral method with DFT and is referred as to Spectral simulation (SS) throughout the chapter. Other methods include the FFT Moving Average (FFT-MA) based approach (Le Ravalec et al., 2000) that is exact on a grid, and continuous spectral methods that are exact and not limited to a grid, see section (?). The SS method has shown good reproduction of long range structures, but poor reproduction of short range structures and covariance functions that are not smooth at the origin. Practice has also shown that the simulated variograms with SS can suffer from the periodic nature of DFT, which may generate artifacts in the borders of the grid. DFT and continuous spectral method have more difficulty with variogram models having a linear behavior at the origin (e.g. spherical and exponential). Continuous approach has no problem simulating smooth covariances at the origin for all ranges. Such methods can also simulate covariances with linear behavior provided the sampling at high frequencies is well done. In TB the number of lines used for simulation impacts the variogram reproduction. The number of lines depends on the dimension of the model and variogram structure, a few lines and the artifact banding is perceptible in the realizations. Overall, significant statistical fluctuations are expected in any Gaussian algorithms as the variogram range increases with respect to the domain size. Optimal selection of the parameters of each algorithm is ideal, however, some parameters may be more important than others to ensure a good variogram reproduction. In this chapter the following are discussed:

- The problem of discretization:
  - The number of data inside the window in MA.
  - Searching strategy and the number of data for kriging in SGS.
  - The number of lines in TB.
- Ergodicity
- The variance of the simulated mean of each method and the theoretical error.

A better reproduction of the input statistics is achieved with optimal selection of the simulation algorithm. In addition to choosing the algorithm, the understanding of the practical implementations and user-input parameters of each method is important to improve geostatistical workflows involving uncertainty assessment with simulation.

5.2 Combining spatial structures

In the IFS methodology, the factors are extracted from the LMC and used to reconstruct the original variables. The reproduction of the LMC over all realizations is then directly affected by the simulated factors. Simulated factors that do not honor their spatial structures may depart from the multivariate Gaussian assumptions and potentially introduce less or more variability to the computed variables, which directly affects the simulated LMC. In the IFS methodology, each realization of the simulated factors are combined to generate a final realization of the input LMC through the computation of the variables. The concept of simulate different structures of the variogram independently is not new in geostatistics (Goovaerts, 1997; Journel, 1974). Each structure of the variogram is simulated independently, the generated maps are scaled to their variogram contributions and combined together to produce unconditional realizations that honors the input model.

To illustrate, a few unconditional realizations of an isotropic spherical variogram with two structures and no nugget effect are generated on a grid. The variance contribution of the two structures are 80% and 20% of the total standardized variance, with the respective ranges of 100 m and 400 m. The grid contains 100 nodes in the easting and northing directions, with a node spacing at both directions of 10 m. A total of 50 realizations are generated with SGS and SS. The simulated histogram and map of the first realization are shown along with the variogram reproduction of each method in Figure 5.1.

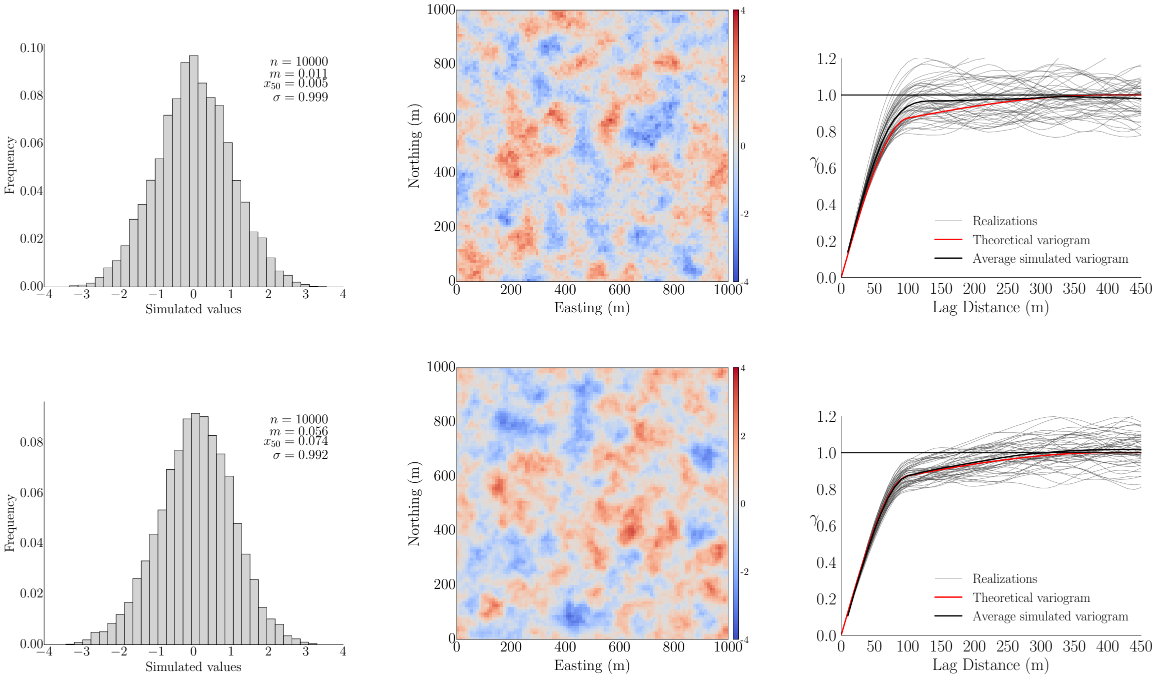


Figure 5.1: The distribution and map of the first realization of a simulated random variable and the respective variogram reproduction are shown for SGS (top) and SS (bottom). The grey lines represent the simulated variograms, the black line represents the average variogram of the realizations, and the red line is the input variogram model.

Note that both simulated histograms shown in Figure 5.1 are standard normal, and the simulated maps appear similar; however, the SS map appears smoother than SGS. The SS reproduces the input variogram more accurately. The biggest difference in the simulated variograms between the two methods is seen in the second structure. The average simulated variogram with SGS shows more discrepancy from the input model. Consider simulating the two structures of the variogram independently. The first and second structures are simulated respectively with SGS and SS. Realizations of the input variogram model are generated by adding the SGS realizations to the correspondent SS realizations. This is represented by:

$$\begin{aligned}\gamma(h) &= 0.8 \times Sph_{a=100} + 0.2 \times Sph_{a=400} \\ \gamma(h) &= 0.8 \times \gamma(h)_{SGS} + 0.2 \times \gamma(h)_{Spectral}\end{aligned} \tag{5.1}$$

The final model is a combination of the realizations generated independently with SGS and SS. The results are shown in Figure 5.2. Note that the combined model reproduces better the input variogram when compared to the SGS variograms shown in Figure 5.1. The high variability seen in the second structure of the SGS model is replaced by the better behaved simulated variograms of the SS method.

This example illustrates the potential gains when simulating structures independently, which permits mixing different algorithms and optimal setting of simulation parameters. The next sections discuss important practical implementations in each algorithm. Recommendations for algorithm selection are given in the end.

5.3 The problem of discretization

This section covers the problem of the number of data inside the window in MA, the number of data used for kriging and search strategy in SGS, and the number of lines for simulation in TB.

Window size and grid discretization for moving average

The grid discretization in MA defines the number of data n falling inside the window. For a fixed window size, the variance of the simulated values decreases when n increases. The error in simulation of the spherical structure is assessed for 2D and 3D grids. The software `MW_SIM` (Cabral Pinto and Deutsch, 2017d), see sections 2.4 and , is used. To assess the error in simulation as a function of n, a simulation exercise is performed. A set of thirteen different numbers n ranging from a minimum of 10 to a maximum of 100,000 is defined. A total of 500 realizations are generated

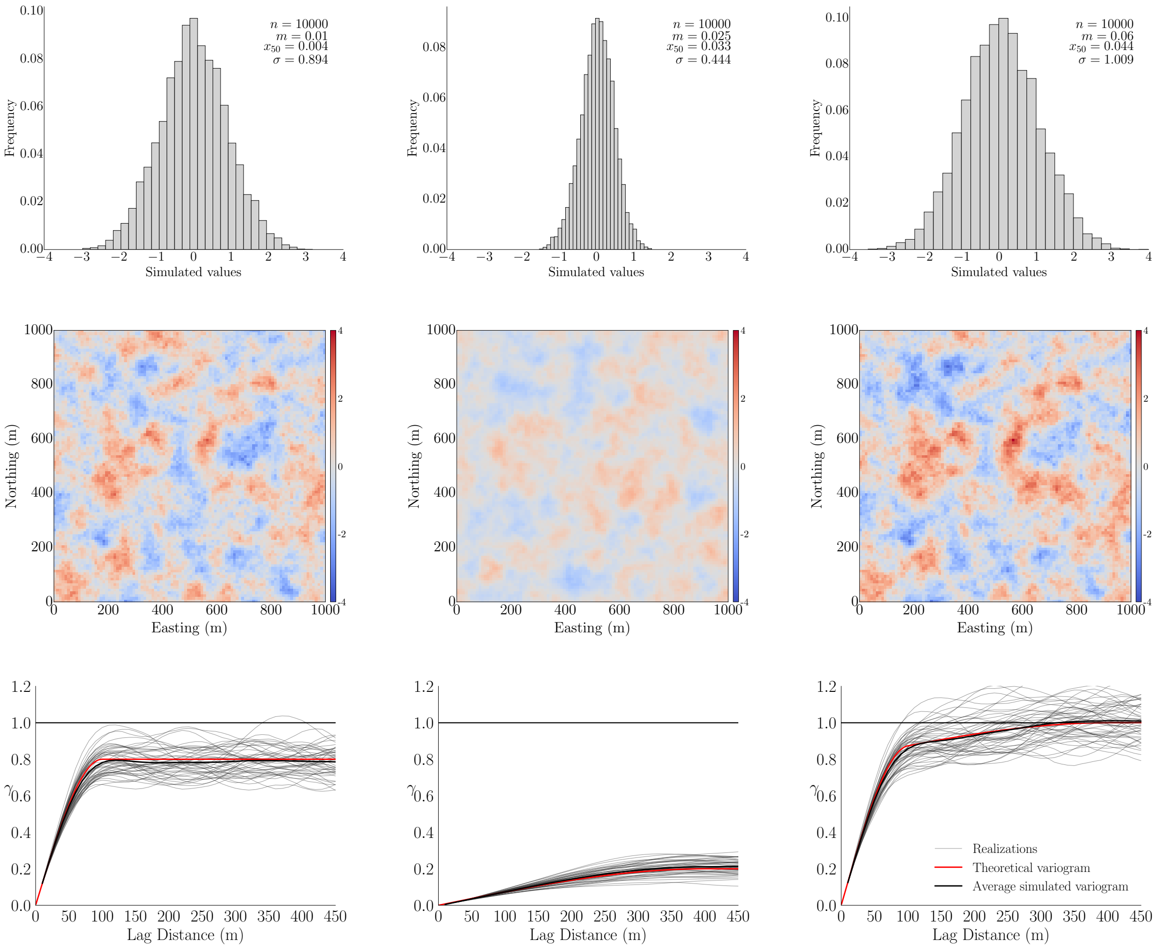
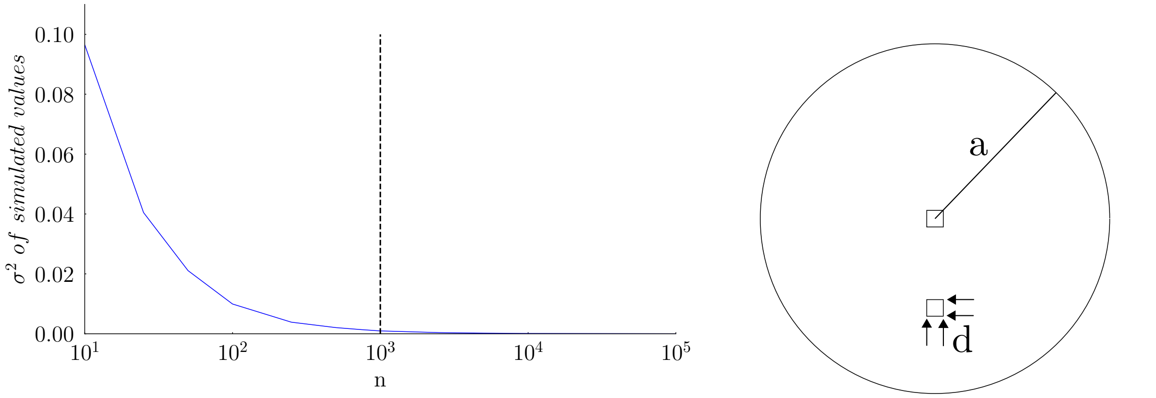


Figure 5.2: The distribution and map of the first realization, and the variogram reproduction generated with SGS for the first structure of the input variogram model are shown in the left column. The same results generated with SS for the second structure of the variogram are shown in the middle column. The combined models are shown in the column to the right. The grey lines represent the simulated variograms, the black line represents the average variogram of the realizations, and the red line is the input variogram model.

for each n. The n numbers are randomly drawn from a standard normal distribution. The average of these numbers represents the simulated value for each realization. The variance of the 500 simulated values is calculated and plotted against n in Figure 5.3a.



(a) Error in simulation for different n values

(b) Schematic illustration of d and a

Figure 5.3: The effect of discretization d in MA. The variance of the simulated values is a function of the number n of data inside a window with radius a.

The discretization d in MA, shown in Figure 5.3b, is defined as a function of the window radius a and the number of data n inside the window. The error in the simulated values approaches zero as n tends to infinity. In practice, as shown in Figure 5.3a, there are no significant changes in the error for $n > 1000$. This number is used in the Equations 5.2 and 5.3 to calculate the discretization d as a function of the variogram range in 2D and 3D models respectively.

$$\begin{aligned} n &= \frac{\pi a^2}{d^2} \\ d &= a \times \sqrt{\frac{\pi}{1000}} \\ d &\approx 0.05 \times a \\ d &\approx 0.025 \times \gamma_{range} \end{aligned} \tag{5.2}$$

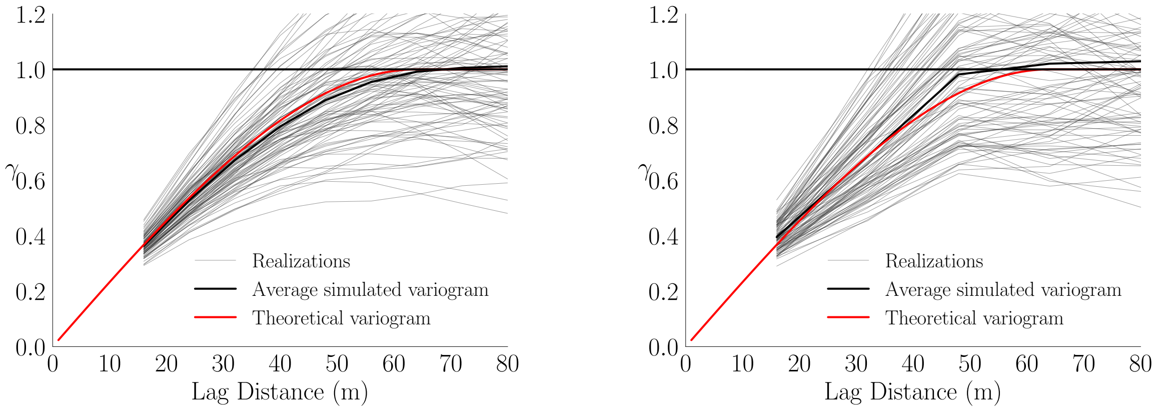
$$
\begin{aligned}
n &= \frac{4}{3}\frac{\pi a^3}{d^3} \\
d &= a \times \sqrt[3]{\frac{4}{3}\frac{\pi}{1000}} \\
d &\approx 0.16 \times a \\
d &\approx 0.08 \times \gamma_{range}
\end{aligned}
\tag{5.3}
$$

The grid discretization values of at least 2.5% and 8% of the variogram range for 2D and 3D models respectively, are values calculated based on an isotropic window size for a reference value $n = 1000$. For the same variogram range the discretization required in 2D models is approximately 3x finer than the required in 3D. For an acceptable higher error in simulation, a coarser discretization can be calculated from Figure 5.3a and Equations 5.2 and 5.3. The discretization must accommodate the anisotropy, and a different grid discretization in the anisotropic directions must be set accordingly.

Considerations for sequential Gaussian simulation

The grid discretization is an important parameter in MA. The error in simulation is directly related to the number of data inside the window. Simulated values from SGS are less sensitive to the grid discretization and good variogram reproduction is still achieved in coarse grids. For instance, consider a 2D grid with node spacing of 16 m in both directions. A total of 50 realizations are generated with MA and SGS. The target variogram is isotropic with one spherical structure with a range of 64 m. Note that the grid discretization is 10x coarser than the recommended discretization for MA, calculated with Equation 5.2. The variogram reproduction of both methods are shown in Figure 5.4.

Note the difference in the average simulated variogram, the SGS variogram shows a better reproduction for any lags values whereas the MA variogram is above the reference for range values greater than 35 m. The MA variogram reaches the sill at a lag distance of approximately 50 m and the variogram values beyond this range are consistently higher than the input variogram. This illustrates the high



(a) SGS

(b) MA

Figure 5.4: Variogram reproduction of realizations on a grid with coarse discretization generated with SGS (a) and MA (b). The target variogram range is four times larger than the grid node spacing.

variance in MA simulation for a few n inside the window. In SGS, the number of data n used in kriging to conditioning the local distributions is a critical parameter. Using all data in SGS is in practice prohibited (Emery, 2004; Emery and Pelaez, 2011; Safikhani et al., 2016). As simulation proceeds, previously simulated nodes are added to the original input data matrix and become new conditioning data to simulate the next location. A few interactions and the high computational requirements to solve the kriging equations invalidate the method. Practical implementations such as a moving search window and multiple-grid search (multigrid) provide a workaround to the problem but limit the number of data used as conditioning. This limitation reduces the performance of SGS and affects variogram reproduction. The optimal selection of the number of data inside the search is considered, and the balance between good variogram reproduction and computational performance is sought.

SGS is a covariance based algorithm, the covariance matrix used in the kriging equations is a function of the n data found inside the search. The simulated values are more normal distributed for large n. To illustrate, unconditional realizations of an isotropic variogram with no nugget and one spherical structure with a range of correlation of 100 m are generated on a 3D grid. The grid extends 1000 m in easting, northing, and elevation directions. The node spacing is 10 m in each direction. The variogram range is then 1/10 of the domain size. The number of data n is set variable, ranging from $n = 10$ to $n = 50$. For each n a total of 100 realizations are generated. In the presence of conditioning data, n is the total number of data used for conditioning, that is, hard input data and previously simulated data. In unconditional simulation cases, there is no conditioning data and n coincides with the number of previously simulated data and the total number of data used for kriging. The variance of the realizations are calculated and plotted against n in Figure 5.5.

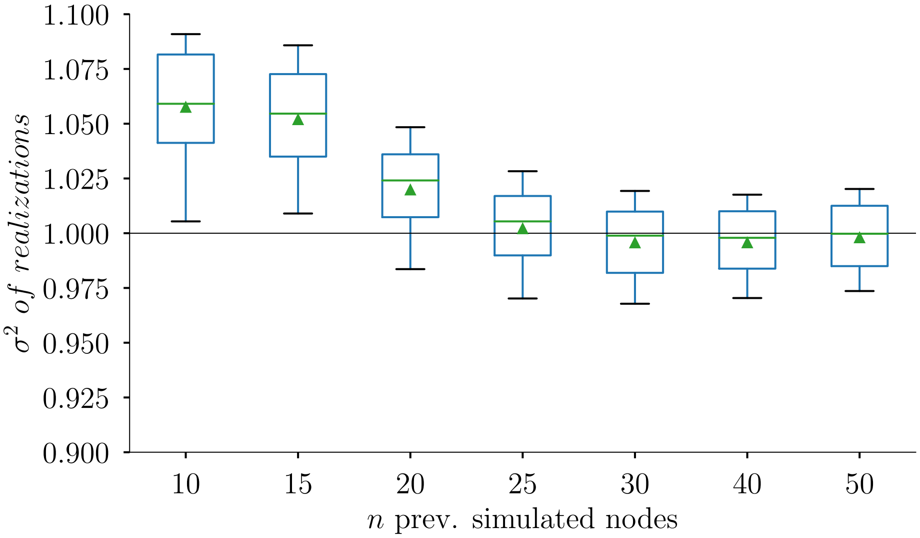
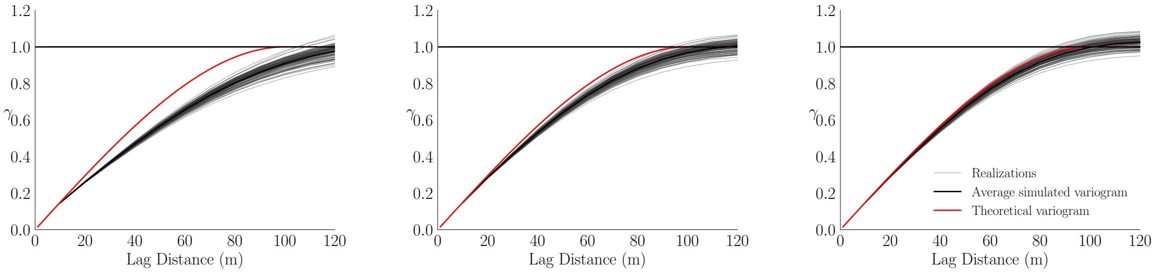


Figure 5.5: Box plots of the variance of the realizations generated with different number of previously simulated nodes. The whiskers represent 10% and 90% of the distribution, the green line in each box represents the median, and the mean is represented by the green triangle.

Note that the mean of the distributions approaches to one and that the dispersion of the distributions (see the whiskers) decreases as n increases. There is no significant change in the distributions for $n > 30$. The number of data inside the search and how the search is performed is critical for SGS. To illustrate the impact of n in the variogram reproduction, the variograms from the same analysis are calculated for $n = 10$, $n = 30$, and $n = 50$ and plotted in Figure 5.6. The simulated variograms approach the reference model as n increases, that is, more data is required to reproduce the input spatial variability.

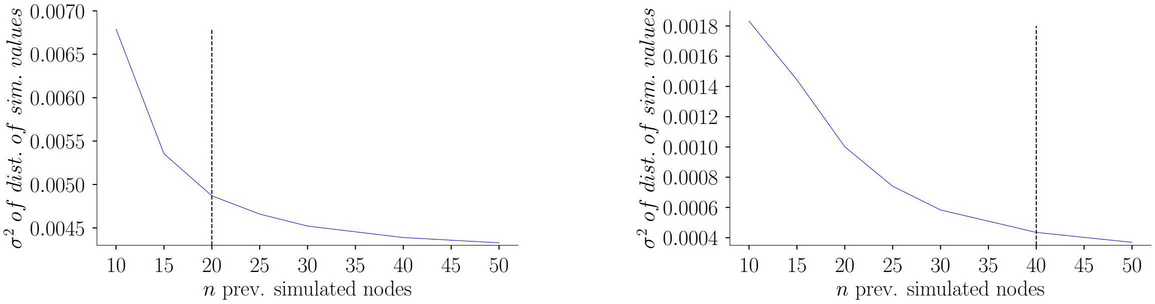
A similar analysis is performed to n in 2D and 3D grid models. The input



(a) $n = 10$ **(b)** $n = 30$ **(c)** $n = 50$

Figure 5.6: The variogram reproduction with SGS for a different number of data n inside the search. The simulated values from the example of Figure 5.5 are used.

variogram, number of realization, and grid specs are the same of the previous example. The mean of the simulated values is calculated for each realization, then the variance of the mean values is calculated and used as the measure of the error in simulation for each n. The distribution of the simulated values for each realization is expected to be standard normal. Departure from normality is expected for small n values, and less error is expected when a large number of data is used. The results are shown separately for the 2D and 3D models in Figure 5.7. The analysis of the plots indicates a diminishing changing rate in the measured error for $n > 20$ and $n > 40$ values in the 2D and 3D models respectively. These parameter values can be used as a reference when SGS is considered and fine tunned after checking the simulation results.

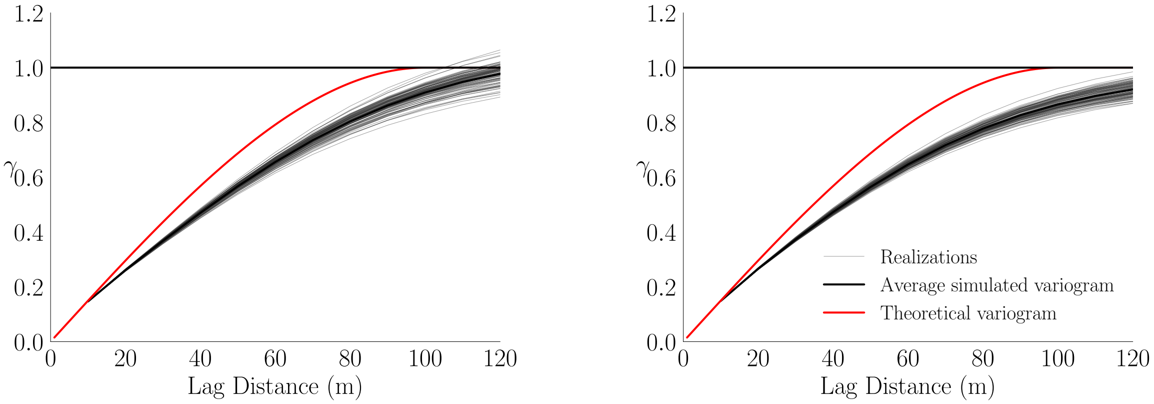


(a) 2D **(b)** 3D

Figure 5.7: The variability in the simulated mean with SGS as a function of n for 2D (a) and 3D (b) models. The highlighted $n > 20$ (2D) and $n > 40$ (3D) values are used as a reference for a diminishing changing rate in the measured error.

The multigrid random paths is another search strategy found in most implementations of SGS (Deutsch and Journel, 1997; Gomez-Hernandez and Journel, 1993;

Isaaks, 1990; Manchuk and Deutsch, 2012; Tran, 1994). It is designed to improve the reproduction of long range structures by allowing data that are further away from the simulation location, but still inside the search range, to be used for kriging. These data points are found using a coarse grid searching strategy. By refining the coarse grid to a finer grid the multigrid does not compromise the reproduction of short range structures, thus, there is no harm in using it. In most implementations, the multigrid is a parameter that can be turned on and off and does not required tunning. The improvements in the variogram reproduction are perceptible, although not as significant as the ones seen when changing the number of data in the search window. Consider the case $n = 10$ in the example used to generate Figure 5.6. The variogram reproduction with $n = 30$ and $n = 50$ are significantly better than using $n = 10$. The multigrid search is used during simulation. A comparison is made with the variograms of realizations generated without multigrid, see Figure 5.8. The average and simulated variograms with multigrid are slightly closer to the input reference variogram. This is visually more evident when taking the range of 120 m as a reference point.



(a) With multigrid

(b) No multigrid

Figure 5.8: The variogram reproduction with SGS with (a) and without (b) multigrid search strategy. The simulated values for $n = 10$ in the example generated for Figure 5.6 are used.

The number of data to search and the search strategy in SGS are important input parameter for SGS. A balance between the number of data to search and computational performance must be found. The analyses made on the simulation error and variogram reproduction for different n data indicate that a minimum of $n = 20$ and $n = 40$ provide a good starting point for 2D and 3D models respectively. The use of a multigrid search is suggested since it improves reproduction of the long structure range of the variogram without sacrificing short range structures, and does not require any user-input adjustments.

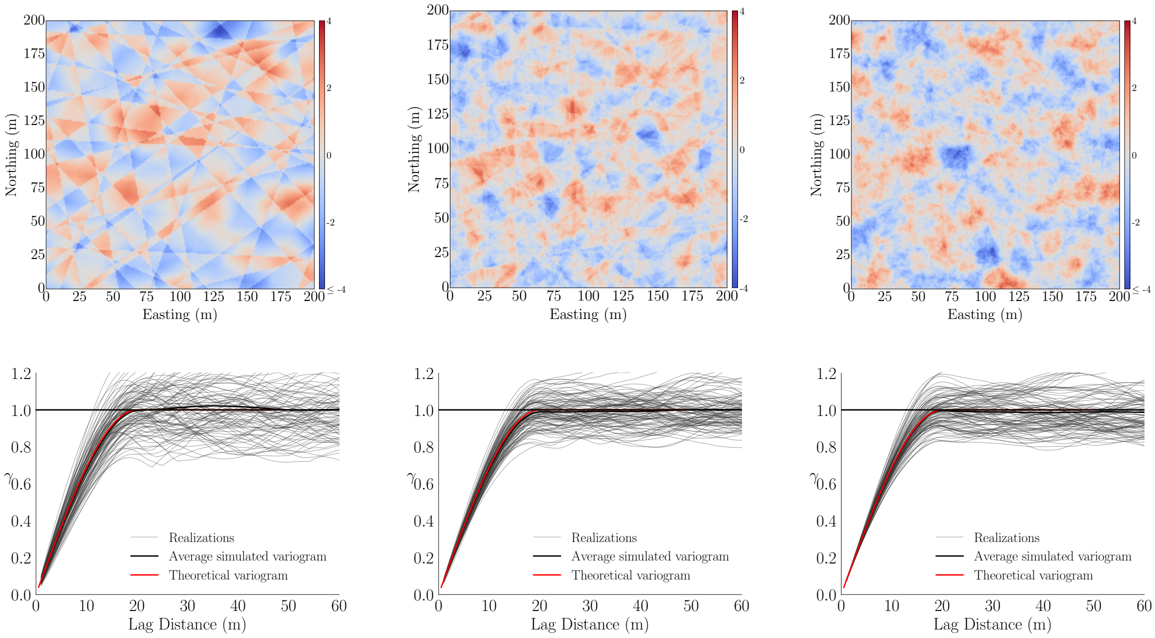
Number of lines in turning bands

Practical implementations of the TB algorithm found in the literature use different approaches to generate line directions for simulation. To cite a few, there is the icosahedron approximation (Deutsch and Journel, 1992; Journel, 1974), algorithms to generate lines that are uniformly or equidistributed over the sphere (Brooker, 1985; Chiles, 1977; Hunger et al., 2015; Tompson et al., 1989), or lines that follow directions from a van der Corput sequence (Corput, 1935; Emery and Lantuejoul, 2006; Lantuejoul, 1994). It is demonstrated that the input covariance is honored in these methods (Chiles and Delfiner, 2012) and that TB lines generated with the van der Corput sequence improve ergodic properties (Emery and Lantuejoul, 2006; Freulon and de Fouquet, 1991; Lantuejoul, 1994). The current implementation for the analyses in this chapter uses the van der Corput sequence (Cabral Pinto and Deutsch, 2018; Mol, 2018).

The number of lines N in TB depends on several factors, such as, the distribution of the lines, the spatial structure type, and the algorithm to simulate the one dimensional random fields onto the lines. There is no magic number that works for all implementations. There is, however, a range of values for N that practitioners can use as a reference when using TB in geostatistical workflows. The recommended number of lines for 2D and 3D models varies in the literature. Values ranging from $N = 4$ to $N = 180$ are recommended for 2D models (Chiles, 1977; Gneiting, 1999a; Mantoglou and Wilson, 1982; Tompson et al., 1989), whereas for 3D models these values range from $N = 15$ (Journel and Huijbregts, 1978) to hundreds of lines $N > 100$ (Emery and Lantuejoul, 2006; Lantuejoul, 2002; Tompson et al., 1989).

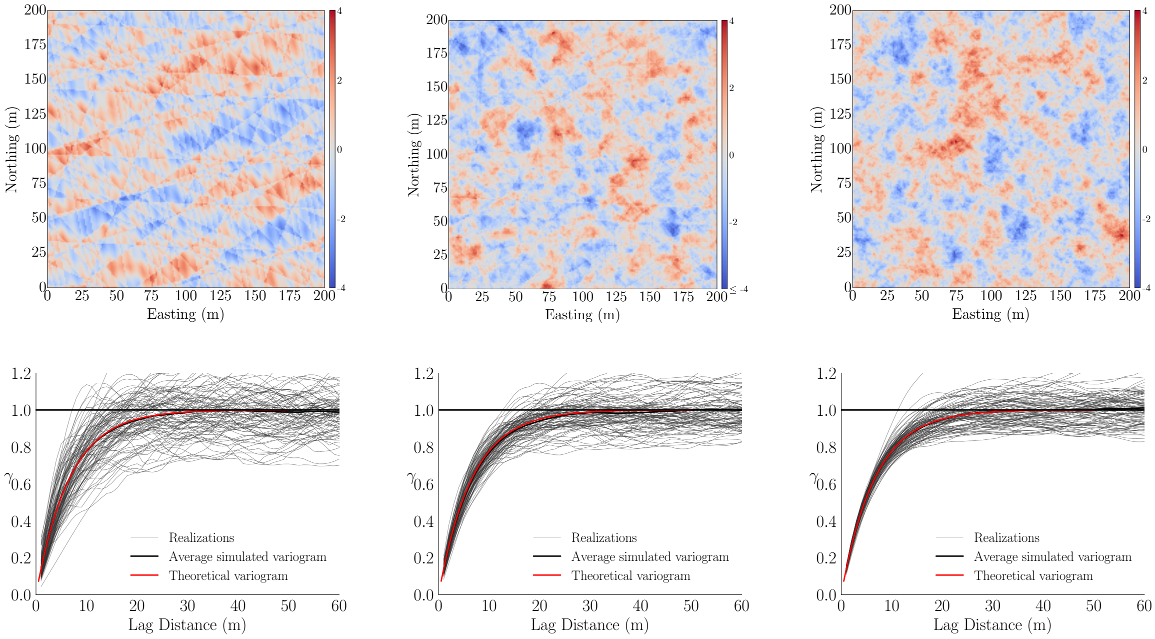
In practice, the number of lines to use in TB can be choose by generating a few unconditional realizations with the desired input variogram and visually checking the results. Simulated maps with artifact banding indicates that more lines should be used. It is also recommended to check the variogram reproduction, since visual appreciation of the banding is subjective. To illustrate, consider generating unconditional realizations on a grid with a different number of lines and variogram types. The simulated map of a realization and the variogram reproduction for each N are checked. Three variograms with one isotropic structure and no nugget effect are considered: a spherical type with range of 20 m, and a exponential and Gaussian variograms with effective range of 20 m. The Gaussian variogram is fit with a small nugget effect for computation stability. A total of 100 realizations on a grid with 200 node cells in easting and northing directions are generated. The cell spacing is 1 m in each direction. The results are shown in Figures 5.9 to 5.11. Note the artifact banding seen in the spherical and exponential models for $N = 10$. These covariance functions are simulated with the partition method (Lantuejoul, 1994) and require more lines to reproduce the input function. The spectral method is utilized to simulate the Gaussian covariance and fewer lines are required in simulation since the simulation is not performed into intervals of the line. The banding effect becomes less perceptible when N increases. The inspection of the simulated variograms for all models shows that a larger N improves variogram reproduction.

Practical implementations of TB found in the literature consider different approaches to generate the lines, different algorithms for simulation on the lines, and the discretization used for this simulation. Practical recommendations for choosing the number of lines to use in TB is based on the visual analysis of the simulated maps and the variogram reproduction. Departures from normality and the input variogram model are more unlikely to happen for a large number of lines, regardless of the implementation. The process of generating the lines is not computationally expensive and can be performed in parallel. The analysis made in this section indicates that $N > 100$ are suitable for generating realizations that reproduce the input variogram. This number can be reconsidered to improve variogram reproduction.



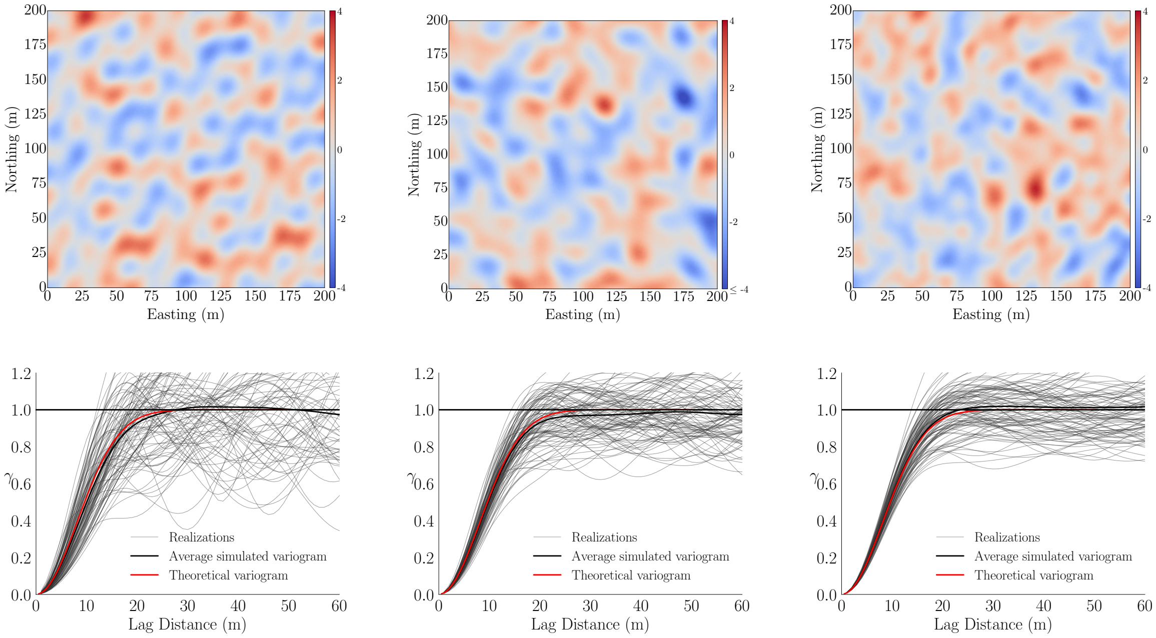
(a) $N = 10$ **(b)** $N = 50$ **(c)** $N = 250$

Figure 5.9: 2D realization maps of an isotropic spherical variogram with range of correlation of 20 m generated with a different number N of lines.



(a) $N = 10$ **(b)** $N = 50$ **(c)** $N = 250$

Figure 5.10: 2D realization maps of an isotropic exponential variogram with effective range of correlation of 20 m generated with a different number N of lines.



(a) $N = 10$ **(b)** $N = 50$ **(c)** $N = 250$

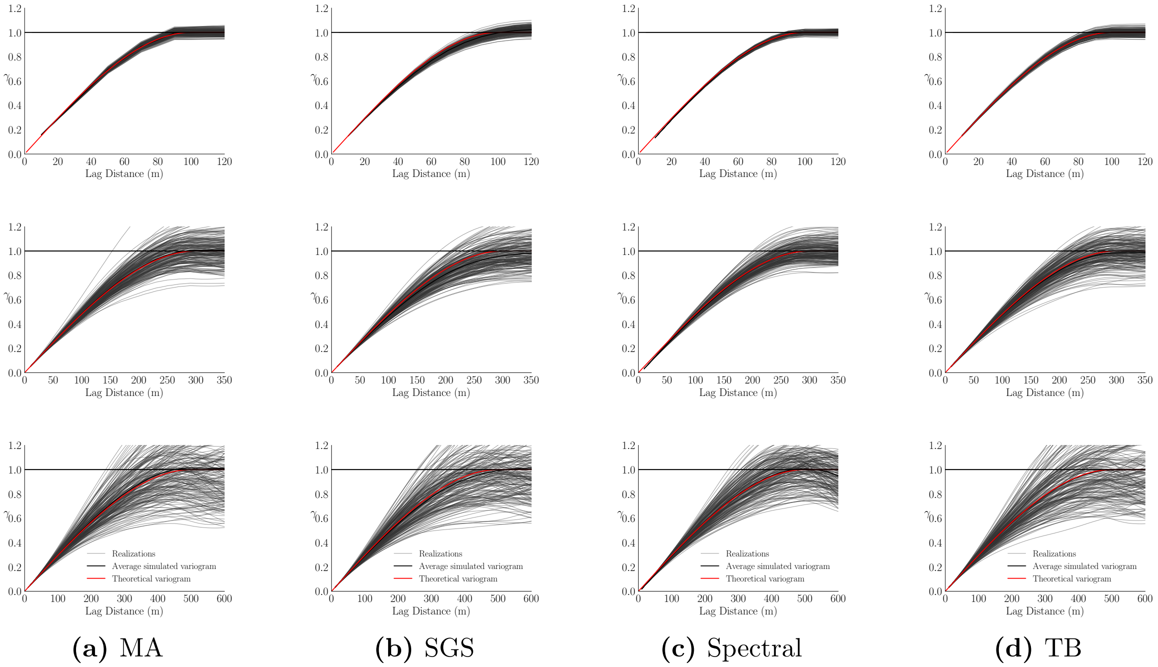
Figure 5.11: 2D realization maps of an isotropic Gaussian variogram with effective range of correlation of 20 m generated with a different number N of lines.

5.4 Ergodic fluctuations of different simulation algorithms

In geostatistical simulation, the discrepancy between the simulated values over a set of realizations and the corresponding model parameters is referred as to ergodic fluctuations (Deutsch and Journel, 1997). Ergodicity allows inference of the statistical parameters of stationary random functions from realizations statistics. Simulated values tend towards normality as the size of the model increases as regarding the range of correlation. Fluctuations in the simulated statistics are expected in any Gaussian simulation algorithm. The ratio between the variogram range and the domain size (VRD) dictates the degree of discrepancy between the simulated values and the reference model parameters.

To illustrate the expected degree of ergodic fluctuations in different algorithms, consider a 3D grid extending 1000 m in all three directions with an equal grid discretization of 10 m. Five isotropic spherical variograms with ranges of 100, 200, 300, 400, and 500 meters are considered. These values correspond to VRD ratios ranging from 10% to 50%. A total of 200 realizations are generated. The search window in SGS matches the variogram range with the maximum number of data for kriging set to 48. A multigrid search is used. For TB, 250 lines are used for simulation. Consider the variogram reproduction plots shown in Figure 5.12. Note that the dispersion of the simulated variograms around the reference model depends on the algorithm and the VRD ratio. Note also, that despite the high dispersion of the simulated variograms as the range increases, the input variogram is on average reproduced. SS provides good results also at short distances but with less variogram fluctuations than with the other methods. The lower fluctuations is explained by the implementation (section (**?**)) using $\sqrt{s(\omega)}$ instead of a Gaussian variable with variance proportional to $s(\omega)$ as describe in Chiles and Delfiner (2012). This is seen in the simulated variograms of SS for small lag distances. Long range structures, including spherical and exponential ones, are still recommended to be simulated with Spectral, since a fine discretization of the spectrum is still achieved at long range. In fact, the Spectral method shows the best variogram reproduction for VRD=0.3 and VRD=0.5. The average variogram generated with SGS is more continuous than the reference for any VRD ratios. The increase in the search allows for more data to be used in the kriging equations which leads to a better reproduction of the variogram for VRD=0.5. The simulated variograms with TB are similar to SGS, however, TB shows a better average variogram for any VRD.

The mean (μ) and variance (σ^2) of the simulated values are calculated for each realization and their distributions are shown as box plots in Figures 5.13 and 5.14. These two figures show the ergodic fluctuations in the first and second order statistics as a function of the variogram range and domain size. Note that the despite an increasing dispersion of the simulated mean for large VRD values, the average and median of the distributions fluctuate around zero. The analysis of the dispersion of the variance shows that as the variogram range increases compared to the domain size the realizations have a lower average variance. As shown in Figure 5.12, the reference variograms are still reproduced over many realizations. It is recommended



(a) MA **(b)** SGS **(c)** Spectral **(d)** TB

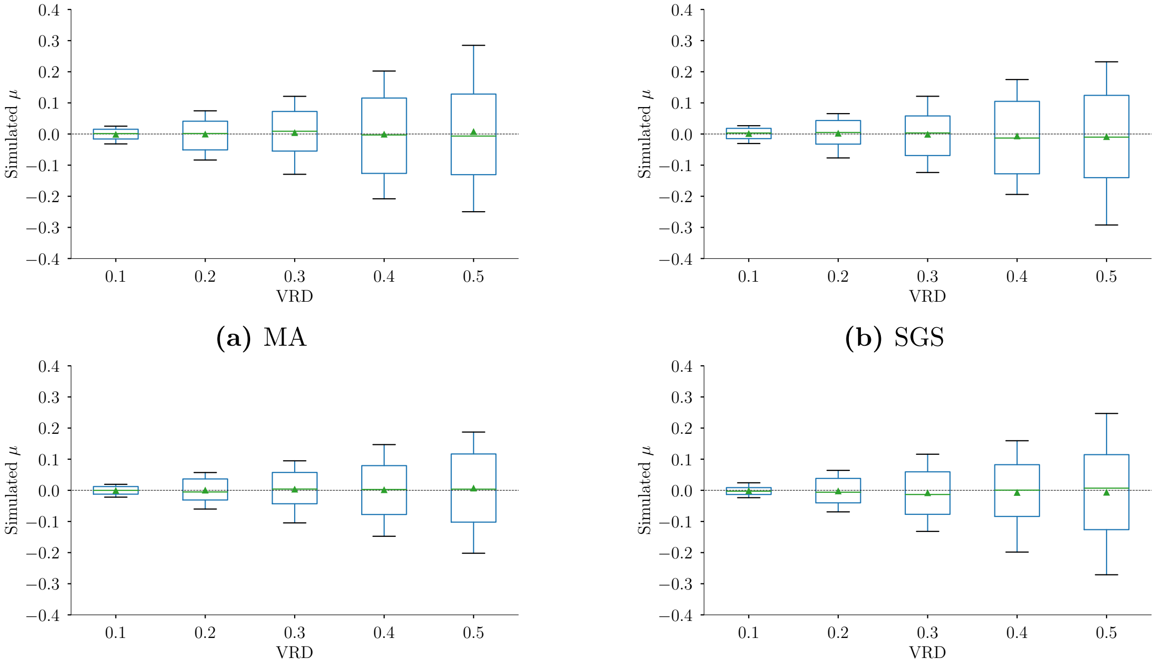
Figure 5.12: The reproduction of the variogram of each method for different VRD ratios. A VRD=0.1 is shown on the top row, VRD=0.3 on the middle row, and VRD=0.5 on the bottom row. The methods are plotted column-wise.

to increase the number of realizations when simulating structures with long ranges of correlation.

The visual analysis of the variograms and the analyses on the distributions of the simulated mean and variance are recommended when different algorithms demonstrate similar performance. For example, the simulated variograms with MA and SGS for VRD=0.5 appear similar. The distribution of the simulated mean in both methods also look similar, however, there is more dispersion in the simulated variance with SGS, which can be seen comparing Figures 5.14a and 5.14b. The visual inspection of the variograms and simulated distributions are hands-on analyses to compare different algorithms.

5.5 Recommendations for algorithm selection

The analyses made in the previous section illustrate that all four algorithms perform well when compared against the expected error and input statistical param-



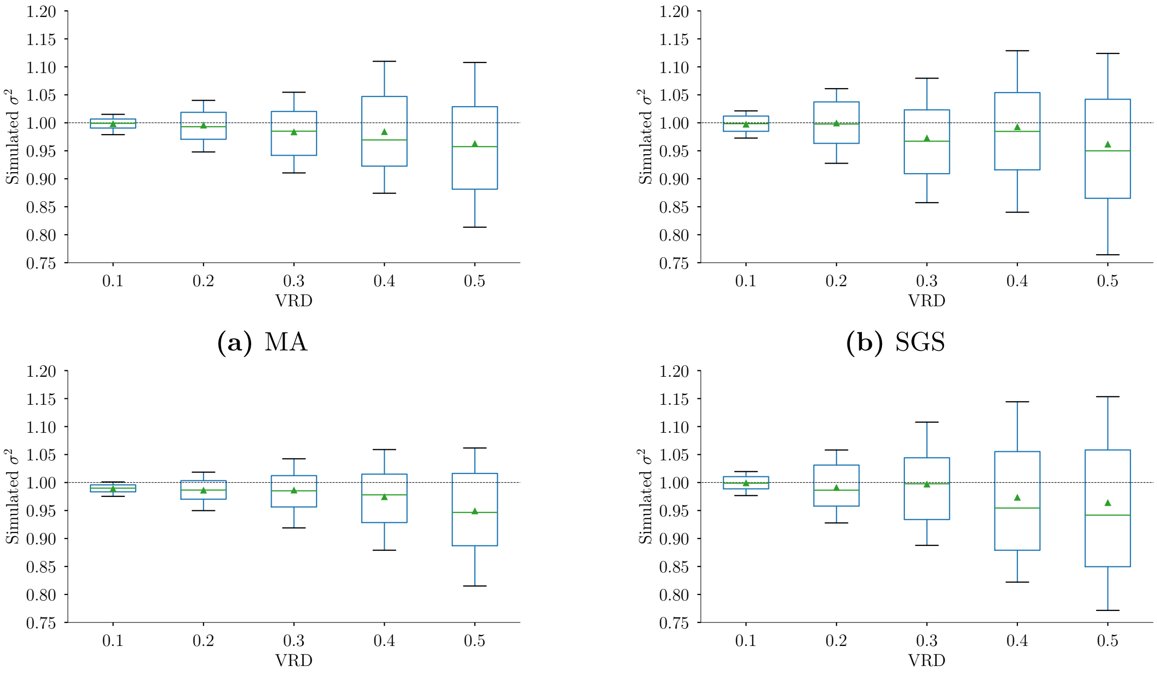
(a) MA

(b) SGS

(c) Spectral

(d) TB

Figure 5.13: The dispersion of the simulated mean over all realizations for different algorithms and VRD ratios. The whiskers represent 10% and 90% of the distribution, the green line is the median, and the mean is represented by the green triangle.



(a) MA

(b) SGS

(c) Spectral

(d) TB

Figure 5.14: The dispersion of the simulated variance over all realizations for different algorithms and VRD ratios. The whiskers represent 10% and 90% of the distribution, the green line is the median, and the mean is represented by the green triangle.

eters. There are however, different implementation aspects in each method that when used correctly improve performance for specific input variogram functions and provide a better modeling tool at a given scenario. The understanding of these practical implementations when selecting the algorithm for geostatistical simulation workflows directly impacts the simulation results. All these algorithms are often available in software and different implementations of the same algorithm exist. A comprehensive understanding of what is available and implemented can also save time on further post-processing and tunning. Based on the analyses made and review of the literature, the following are recommended:

- Different structures of the variogram can be simulated independently and combined together to generate a final realization that reproduces the input model. The realizations generated for each structure must be scaled accordingly to their sill contributions.

- MA offers a fast approach to simulate spherical structures. Improved variogram reproduction is achieved for a number $n > 1000$ of data inside the window. This number represents a grid discretization of 2.5% and 8% of the variogram range for 2D and 3D models respectively. The discretization must change to accommodate anisotropy.

- SGS is less sensitive to the grid discretization and more dependent on the number of data in the search window and other search strategies. The multigrid search does not harm simulation and helps improving long range structures without sacrificing short range variability. The recommended number n of data to search, accounting for original input data and previously simulated nodes, is $n > 20$ in 2D and $n > 40$ in 3D models. Using more data improves the reproduction of the variogram but affect computational performance since more data is used to solve the kriging equations.

- The recommended number of lines N in TB simulation depends on how the lines are generated. For most applications $N > 250$ is a good stating point.

Fewer lines may be required to simulate continuous covariance functions near the origin if spectral methods are implemented to simulate such structures onto the lines.

- Long range structures are better reproduce than short range structures with Spectral simulation with discrete spectral decomposition. Algorithms based on continuous spectral decomposition can be used to achieve good overall reproduction of short and long range structures and less continuous covariance functions at the origin.
- Spectral or TB implementations that use continuous spectral decomposition functions or DFT are recommended for simulating continuous structures with short or long range of correlation.
- SGS is recommended to simulate short range structures that are less continuous near the origin.
- Ergodic fluctuations in the simulation statistics is expected for model with a relatively larger range of correlation with relation to the simulation field size.

5.6 Conclusion

This chapters discusses important implementation details of moving average, sequential Gaussian simulation, spectral simulation, and turning bands. These are common Gaussian simulation algorithms used in geostatistics and different implementations exist in the literature. Each algorithm performs differently when simulating different spatial structures. The choice of the algorithm depends on the implementation and the nature of the algorithm. Recommendations for selecting the best algorithm are given based on the analyses performed on the simulated mean, variance, and variogram reproduction of each method with different grid discretization and the variogram range to the domain size ratio.

Chapter 6

Case study: multivariate modeling of geochemical data with IFS methodology

The following chapter demonstrates the application of the Independent Factor Simulation (IFS) methodology for multivariate modeling. The proposed approach is illustrated through a case study with geochemical sample data. The steps of the methodology discussed in Chapter 4 are highlighted and comments on the results are given. The results of the IFS method are compared to the conventional cosimulation of the variables with cokriging.

6.1 Motivation

In this case study, two geochemical variables are modelled with the IFS methodology. To demonstrate all practical aspects of the proposed method, a subset of the data set is selected that contains collocated and non-collocated sample locations. These locations are chosen randomly. This case study demonstrates that:

- The IFS methodology can be used to impute missing factors at the data locations.
- The imputed factors at the data locations follow a standard normal distribution and are independent.
- At the collocated data locations, the computed variables are exact for any realizations of the imputed factors.
- At the non-collocate data locations, the missing geological variables can be computed from the imputed factors.

- The factors have a unique spatial structure fitted by the LMC.
- The factors are simulated independently and their spatial structures are reproduced over many realizations.
- The original variables can be computed from the simulated factors.
- The LMC is reproduced over many realizations.
- The distribution of the original variables is reproduced over many realizations.

The two variables are modeled with a conventional approach and results are compared to the proposed methodology.

6.2 Data

Publicly available data collected by the Northwest Territories Geological Survey (NTGS) in partnership with the Geological Survey of Canada (GSC) across the Mackenzie Mountains in the Northwest Territories, Canada, are used to illustrate the IFS methodology (Figure 6.1). The geochemical data were collected in several regional surveys for the National Geochemical Reconnaissance program and published by many authors in a series of open file reports (Day et al., 2009,1,0; du Bray, 1995; Falck and Day, 2008; Falck et al., 2014,1,1; Fischer et al., 2016; McCurdy et al., 2009a,0,0; Ozyer, 2010,1).

A subset of the geochemical data containing 2,000 Inductively Coupled Plasma Mass Spectrometry (ICP-MS) measures of stream sediments of Copper (Cu), and the Instrumental Neutron Activation Analysis (INAA) measures of stream sediments of Lanthanum (La) is considered. The geochemical data were collected over an area of approximately 438 Km easting and 546 Km northing, with an average sampling spacing of 6.65 Km. Homotopic measurements of Cu and La are available at 1,500 locations. At the remaining 500 locations, Cu and La are unequally sampled, with a total of 250 measurements of each variable at unique locations. The sample locations are shown in Figure 6.2. A larger subset of the data containing

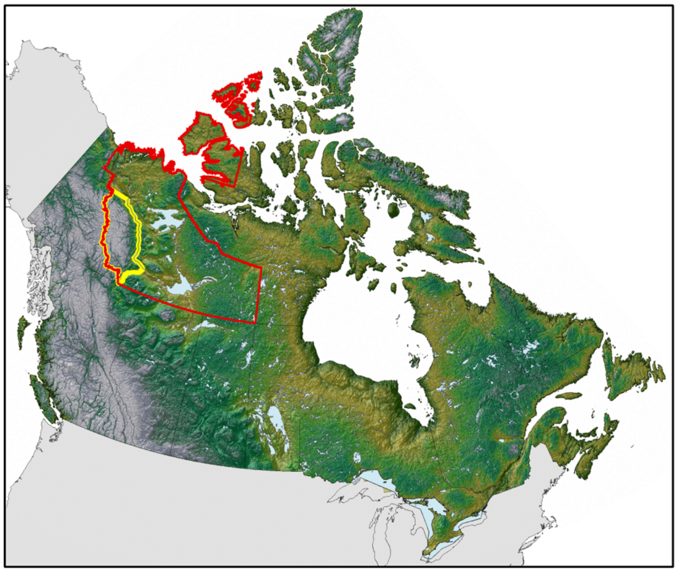


Figure 6.1: Project location: the Mackenzie Mountains (in yellow) in the Northwest Territories (red line). Courtesy of the NTGS.

4,000 measures of Cu and La are left out of the analysis for validation. The accuracy plot of the computed variables over all realizations, and the scatterplot of the smooth e-type mean of each location is compared with the true value of the validation samples.

Cell declustering is considered to calculate a representative distribution of the variables that accounts for the different sampling distances in the area. The cell size is chosen to be the average spacing in areas of sparse sampling and the same cell size is used for both variables. The declustered mean of Cu and La are 6.17% and 1.04% lower than their respectively clustered mean. The naive and declustered distributions of both variables are shown in Figure 6.3.

A requirement of the IFS methodology is that variables are standardized. The original units of Cu and La are transformed to a standard Gaussian distribution with a normal scores (NS) transformation. The distribution of the transformed variables are shown in Figure 6.4. This transformation ensures a zero mean and unit variance consistent with the BSS theory and LMC constraints in Equations 2.12 to 2.16. Moreover, this transformation diminishes the impact of common features in geochemical data such as despikes and outliers on the variograms. The sample

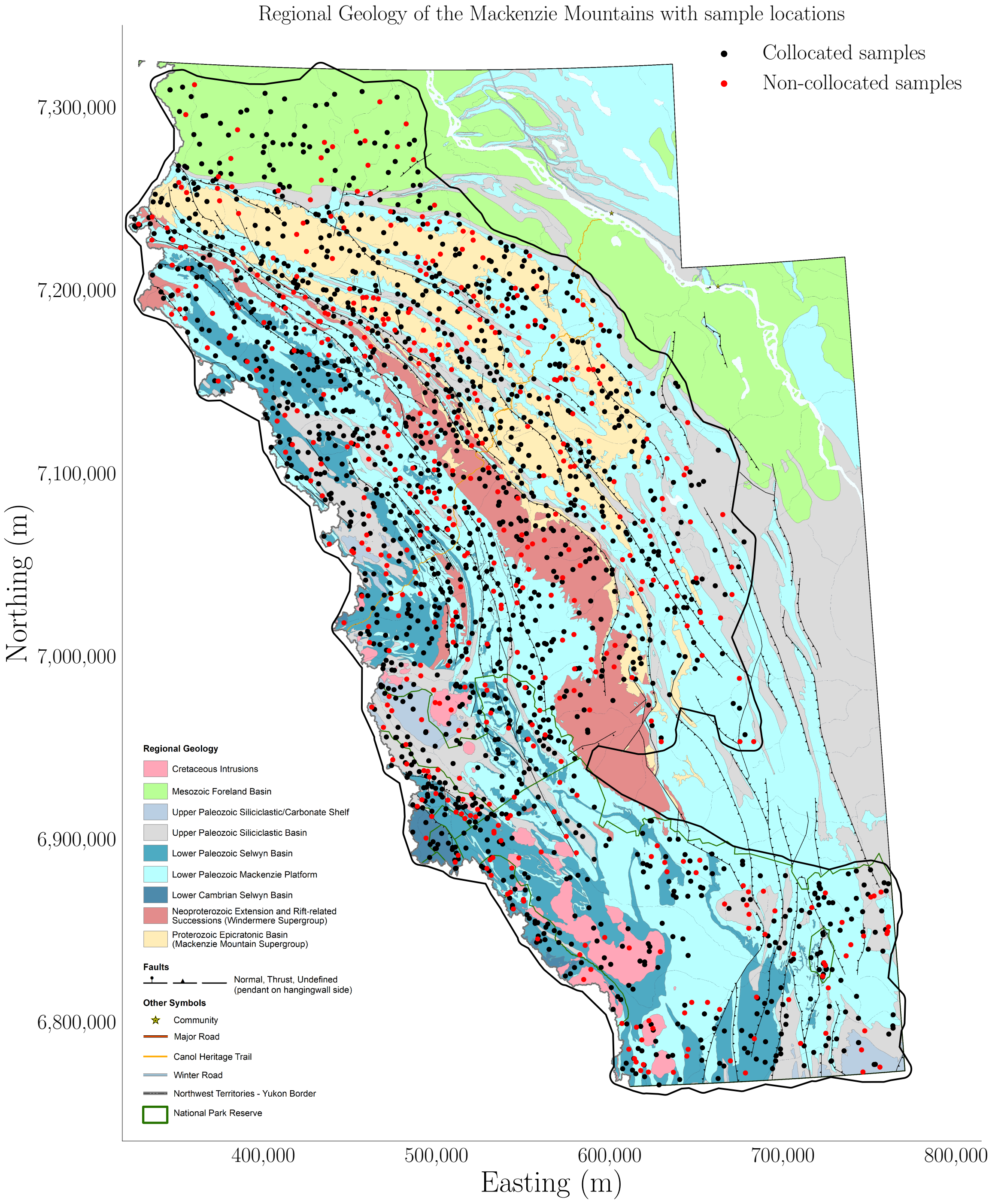
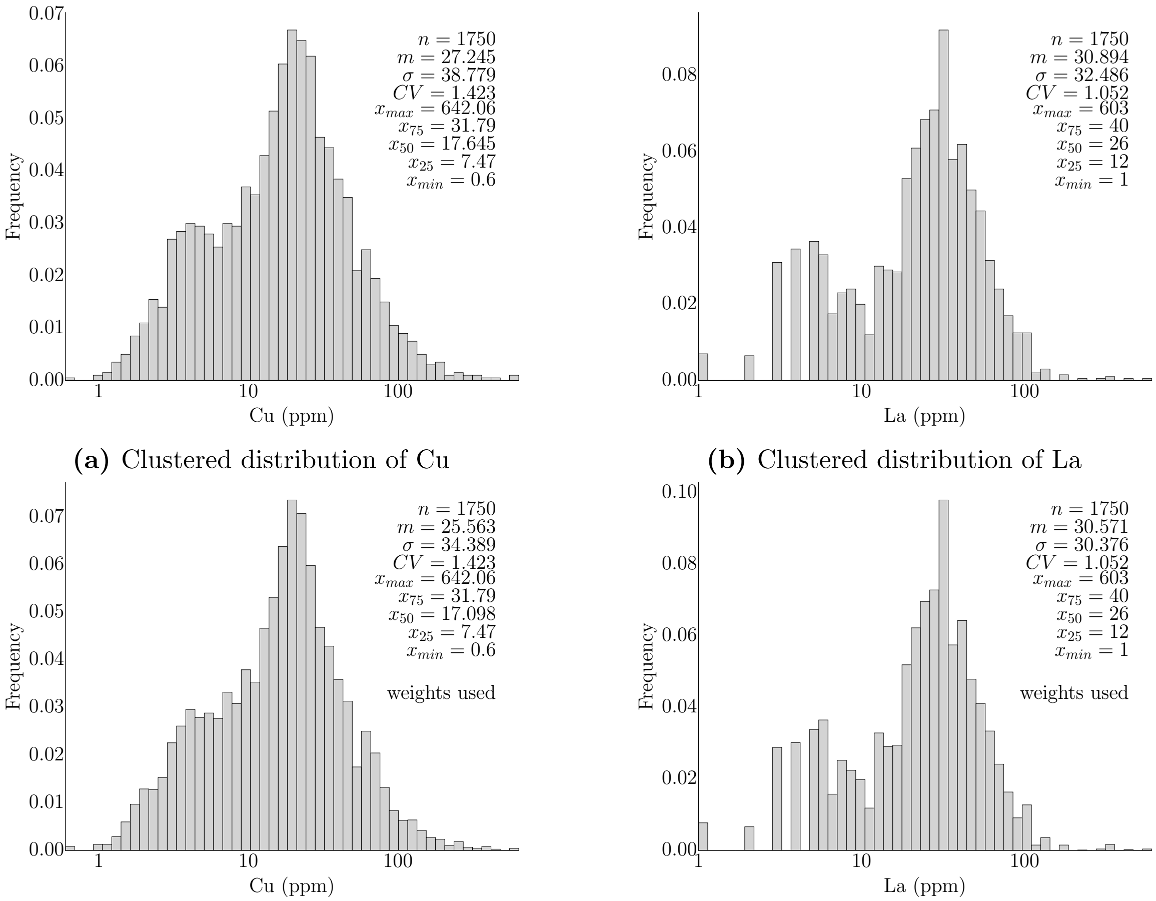

Regional Geology of the Mackenzie Mountains with sample locations
● Collocated samples
● Non-collocated samples
Regional Geology
Cretaceous Intrusions
Mesozoic Foreland Basin
Upper Paleozoic Siliciclastic/Carbonate Shelf
Upper Paleozoic Siliciclastic Basin
Lower Paleozoic Selwyn Basin
Lower Paleozoic Mackenzie Platform
Lower Cambrian Selwyn Basin
Neoproterozoic Extension and Rift-related Successions (Windermere Supergroup)
Proterozoic Epicratonic Basin (Mackenzie Mountain Supergroup)
Faults
Normal, Thrust, Undefined (pendant on hangingwall side)
Other Symbols
Community
Major Road
Canol Heritage Trail
Winter Road
Northwest Territories - Yukon Border
National Park Reserve
Northing (m)
7,300,000
7,200,000
7,100,000
7,000,000
6,900,000
6,800,000
Easting (m)
400,000
500,000
600,000
700,000
800,000

Figure 6.2: The geochemical samples (circles) are plotted on top of the geological map of the Mackenzie Mountains. The black circles represent all 1,500 locations where Cu and La measures are available. The red circles represent the 500 locations where only one variable is sampled, that is, 250 unique measures of Cu or La. The thick line represents the keyout that delimits the area where estimates are calculated. Geological map provided by the NTGS.



(a) Clustered distribution of Cu

(b) Clustered distribution of La

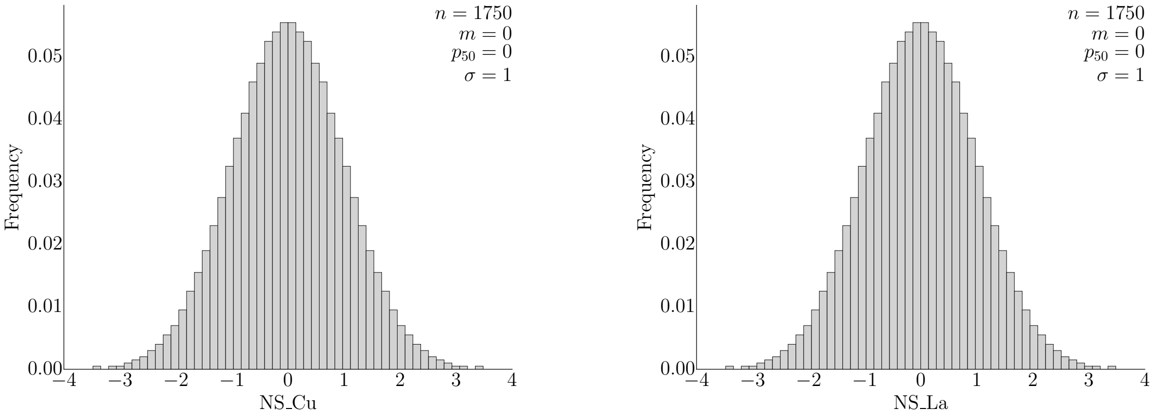
(c) Declusterd distribution of Cu

(d) Declusterd distribution of La

Figure 6.3: The clustered (a,b) and declustered (c,d) distributions of Cu and La are shown in logarithmic scale.

locations of each variable in original units and normal scores are plotted in Figure 6.5.

The spatial distribution of the variables follow the regional structural trend seen in the geological map of Figure 6.2. High measurements of both variables are located in two regions of the Mackenzie Mountains, (1) on all the extension of the Selwyn Basin in the southwest portion of the map and (2) on the centre part of the mountains on the Windermere Supergroup. Average to high measures are located to the northeast, in parts of the Mackenzie Mountains Supergroup and Siliciclastic Basin extensions.



(a) Histogram of NS Cu

(b) Histogram of NS La

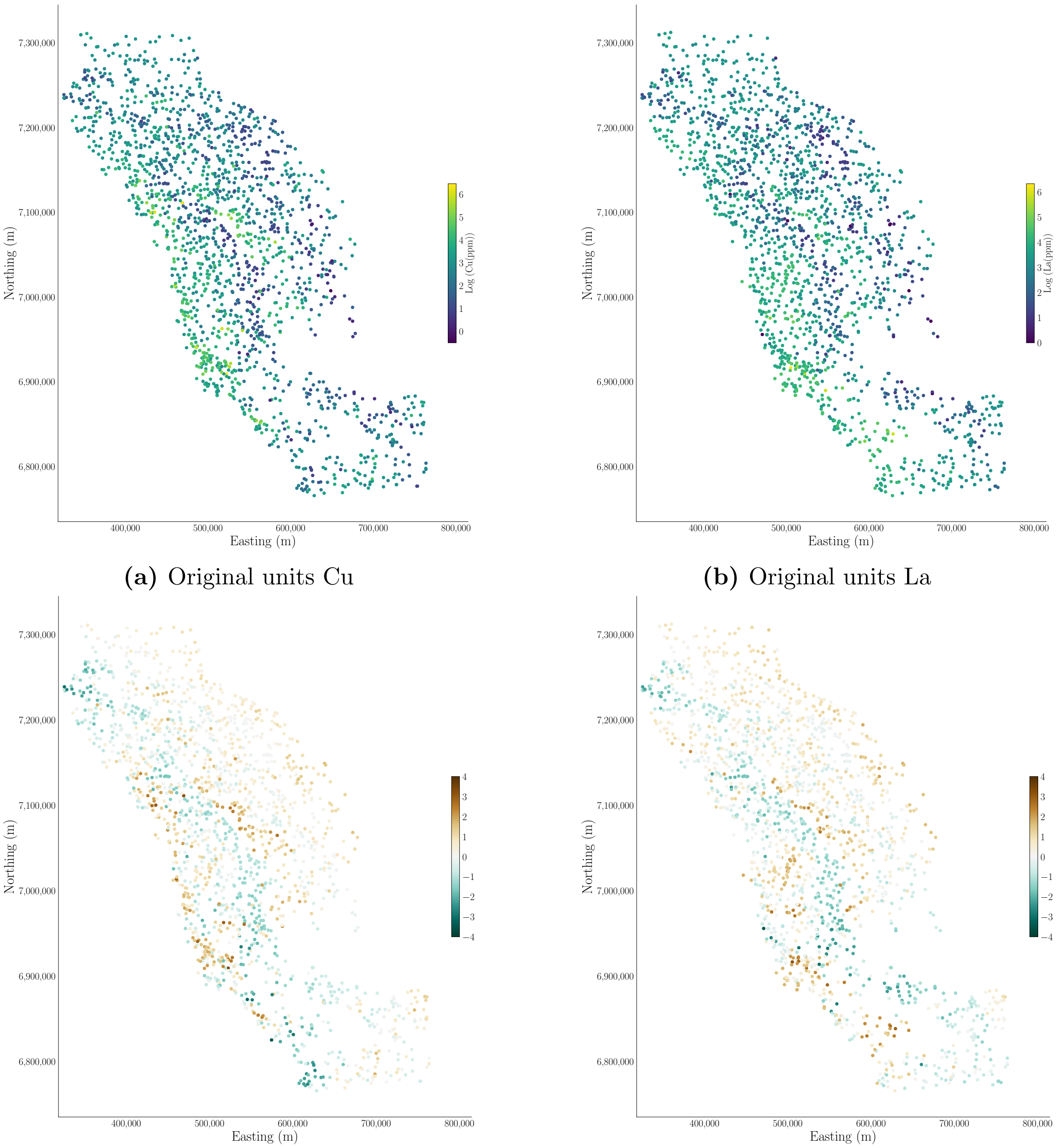
Figure 6.4: The distribution of the normal scores of Cu and La.

6.3 Linear model of coregionalization

The spatial distribution shows zones of low and high values with the highest continuity in the northwest to southeast direction. In the presence of collocated data the variograms are often calculated, and the correlation coefficient is calculated directly from the data. In the presence of unequally sampled data, covariances are calculated and the correlation can be estimated from the cross-covariance.

Experimental covariances are calculated at the directions of apparent anisotropy, with the azimuths of major and minor directions of continuity set respectively to 30 degrees west and 60 degrees east. An LMC is fitted with four structures and no nugget effect (factors Y_1, Y_2, Y_3, and Y_4). The first and last structures receive contributions from both variables, whereas the second and third structures are fitted based on NS La. The correlation between NS Cu and NS La is estimated from the cross-covariance and fitted to the value of $C_{12}(0) = \rho_{12} = 0.515$. The fitted LMC is shown in Table 6.1 and Figure 6.6. The variograms are calculated using the relationship given in Equation 2.4, and are shown in Figure 6.7.

The fitted LMC yields the matrix A that is used to calculate all the covariances required for cokriging, as given in Equations 4.12.



(a) Original units Cu

(b) Original units La

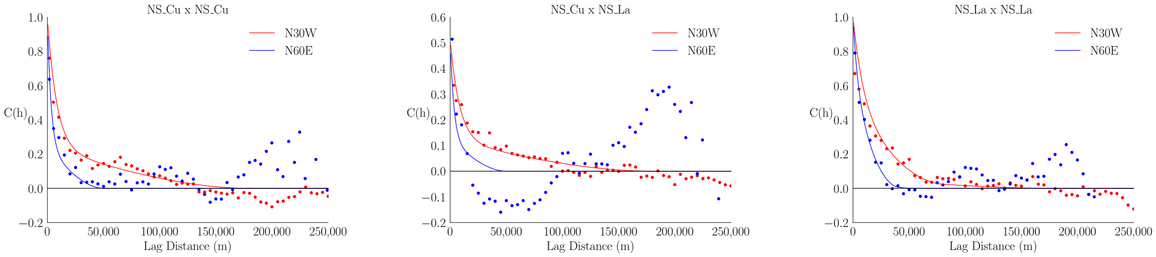
(c) NS Cu

(d) NS La

Figure 6.5: Sample locations of Cu and La colored by logarithm of their original units (a) and (b), and their normal scores (c) and (d).

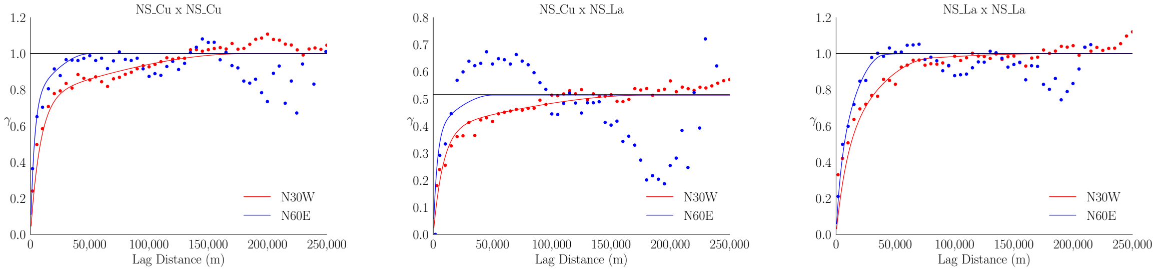
Table 6.1: Fitted LMC parameters.

Factor (i)	Structure type	$a^2_{1,i}$	$a^2_{2,i}$	Range major (Km)	Range minor (Km)
i=1 (Y_1)	Exponential	0.766	0.203	25	10
i=2 (Y_2)	Exponential	0.000	0.390	35	25
i=3 (Y_3)	Spherical	0.000	0.345	80	40
i=4 (Y_4)	Spherical	0.234	0.062	180	50



(a) Direct covariance NS Cu **(b)** Cross-covariance **(c)** Direct covariance NS La

Figure 6.6: The fitted LMC of the normal scores of Cu and La. The dots and continuous lines represent the experimental covariances and the fitted model respectively. The major direction of continuity is shown in red, whereas the minor direction is shown in blue.



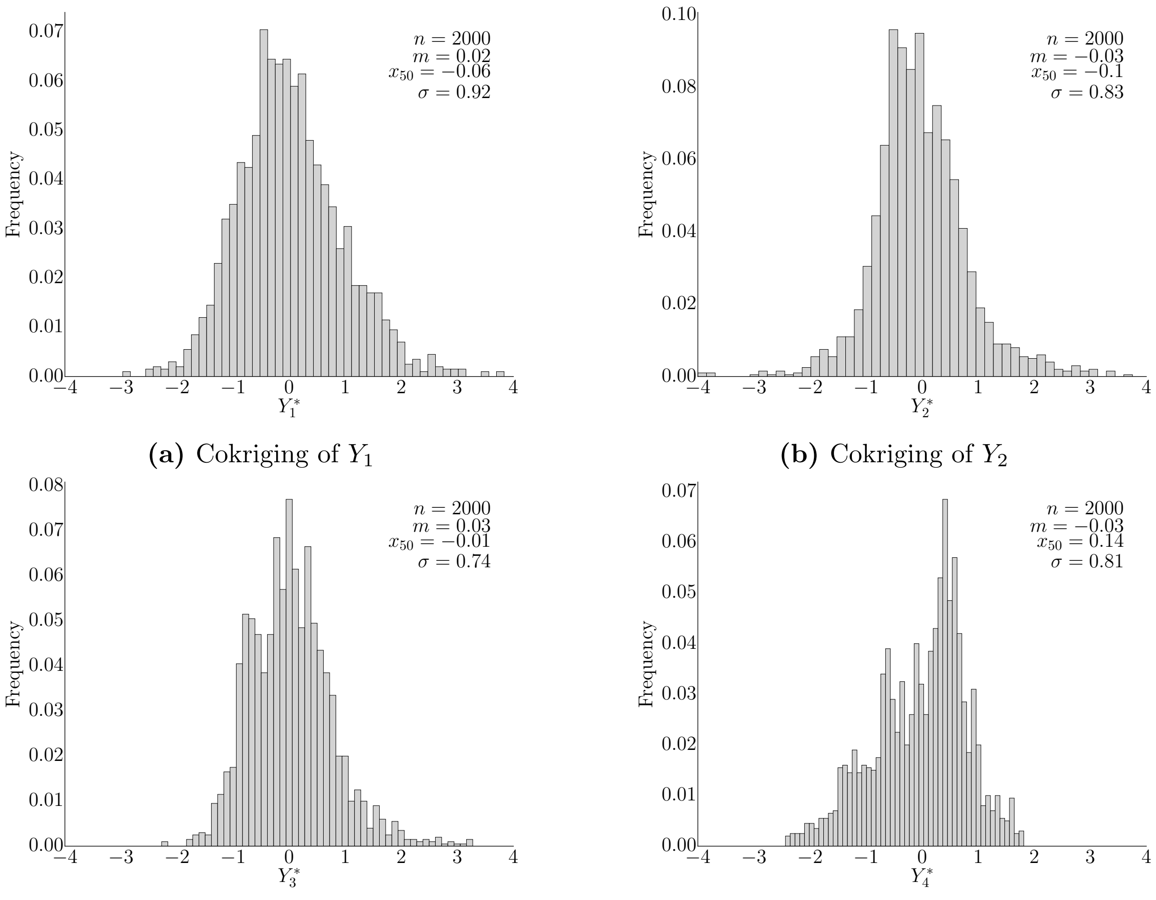
(a) Direct variogram NS Cu **(b)** Cross-variogram **(c)** Direct variogram NS La

Figure 6.7: The fitted LMC of the normal scores of Cu and La. The dots and continuous lines represent the experimental variograms and the fitted model respectively. The major direction of continuity is shown in red, whereas the minor direction is shown in blue.

6.4 Minimum norm solution

The minimum norm solution $y_0 = \Sigma_{yz}\Sigma_z^{-1}z$ to the system of equations $Z = AY$ (Equation 4.11) is obtained by cokriging the variables and factors. Cokriging is performed at the data locations and has all the kriging features introduced in Section 2.3, such as a smoothing effect dictated by the covariances and data configuration. Such effect is seen in the distribution of the cokriged factors shown in Figure 6.8 and on a map in Figure 6.9. Note that the continuity seen on the map are the ones fitted in the LMC.

The LMC and cokriging sections cover the steps 1 to 3 of the methodology, summarized in Section 4.4. These steps involve the calculation of the covariance matrices for dual cokriging Σ_z, Σ_y, Σ_{yz} from matrix A. The minimum norm solution y_0 is required in the next steps of the methodology (steps 4 to 7) that are used to impute the factors at data locations.



(a) Cokriging of Y_1

(b) Cokriging of Y_2

(c) Cokriging of Y_3

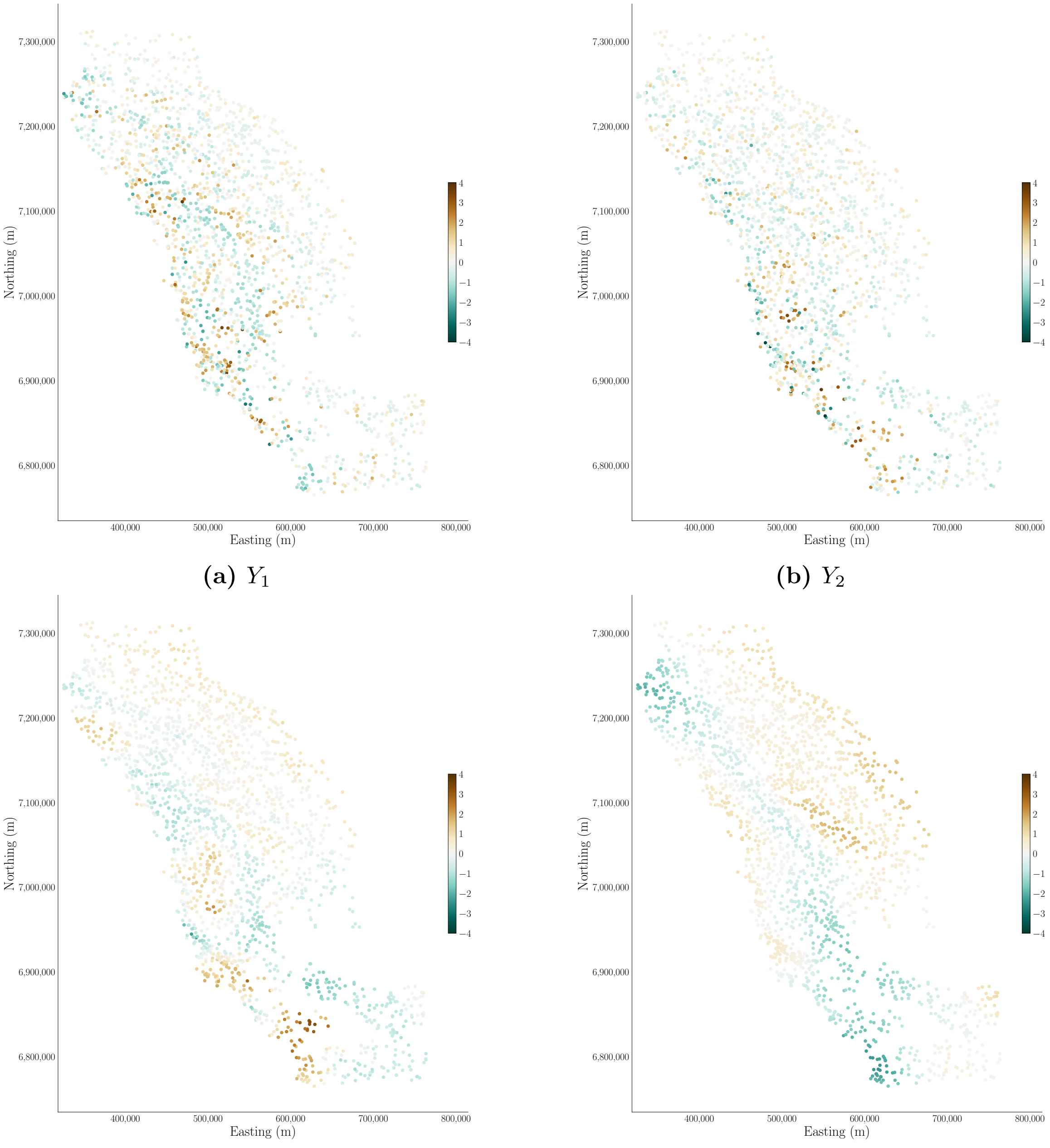
(d) Cokriging of Y_4

Figure 6.8: Distribution of the cokriged factors at the data locations.

6.5 Imputation of the factors at the data locations and checking

The orthonormal basis T and its orthogonal complement $T_\perp$ matrices provide a null solution space to the multivariate distributions, that is, an initial solution to the problem that ensures the factors are independent. Both matrices are calculated from the SVD of the A matrix. At this stage, all matrices required for the calculation of Λ and Σ_x (Equation 4.14) are already calculated. The Cholesky decomposition of Σ_x is used in combination with y_0 in 4.16 to generate a total of 200 realizations of the Y factors at the data locations.

To demonstrate that the factors are standard normal, the average mean and standard deviation of each factor over all realizations is calculated and shown in



(a) Y_1

(b) Y_2

(c) Y_3

(d) Y_4

Figure 6.9: Cokriging of each factor Y at the data locations.

Table 6.2. The average correlation between the imputed factors is shown in Table 6.3. Note that there is no correlation between the imputed factors.

Table 6.2: Average mean (μ) and standard deviation (σ) of the imputed factors at the data locations.

	Y_1	Y_2	Y_3	Y_4
Average μ	0.008	-0.030	0.032	-0.015
Average σ	0.982	1.029	0.978	1.015

Table 6.3: Average correlation between the imputed factors at the data locations.

	Y_1	Y_2	Y_3	Y_4
Y_1	1.000	-0.022	-0.005	0.024
Y_2	-0.022	1.000	-0.010	0.011
Y_3	-0.005	-0.010	1.000	0.039
Y_4	0.024	0.011	0.039	1.000

The distribution of the imputed factors and their cross-correlation are shown in Figure 6.10 for realization #50. This realization is randomly chosen for illustration purpose only. The imputed factors are plotted on a map and shown in Figure 6.11. Note that the continuity of the factors are as fitted in the LMC, but the smoothing effect of cokriging (Figure 6.9) is not seen. The imputed factors are standard normal with a mean of zero and variance of one, and have the correct spatial structure fitted in the LMC.

Realizations of NS Cu and NS La at data locations are generated by the matrix operation $Z = AY$ with the different realizations of the y values. For each realization of the Ys a realization of the Zs is computed. At the collocated data locations the computed values are exact, that is, for any realization of the Ys the computed Zs values are equal to the sample values. This is demonstrated in Figure 6.12 in which the reference sample values are plotted against the computed values.

At the non-collocated data locations, the different realizations of the imputed Ys are used to compute realizations of the missing variable. One realization of the computed variables is shown in the histograms and scatterplot of Figure 6.13.

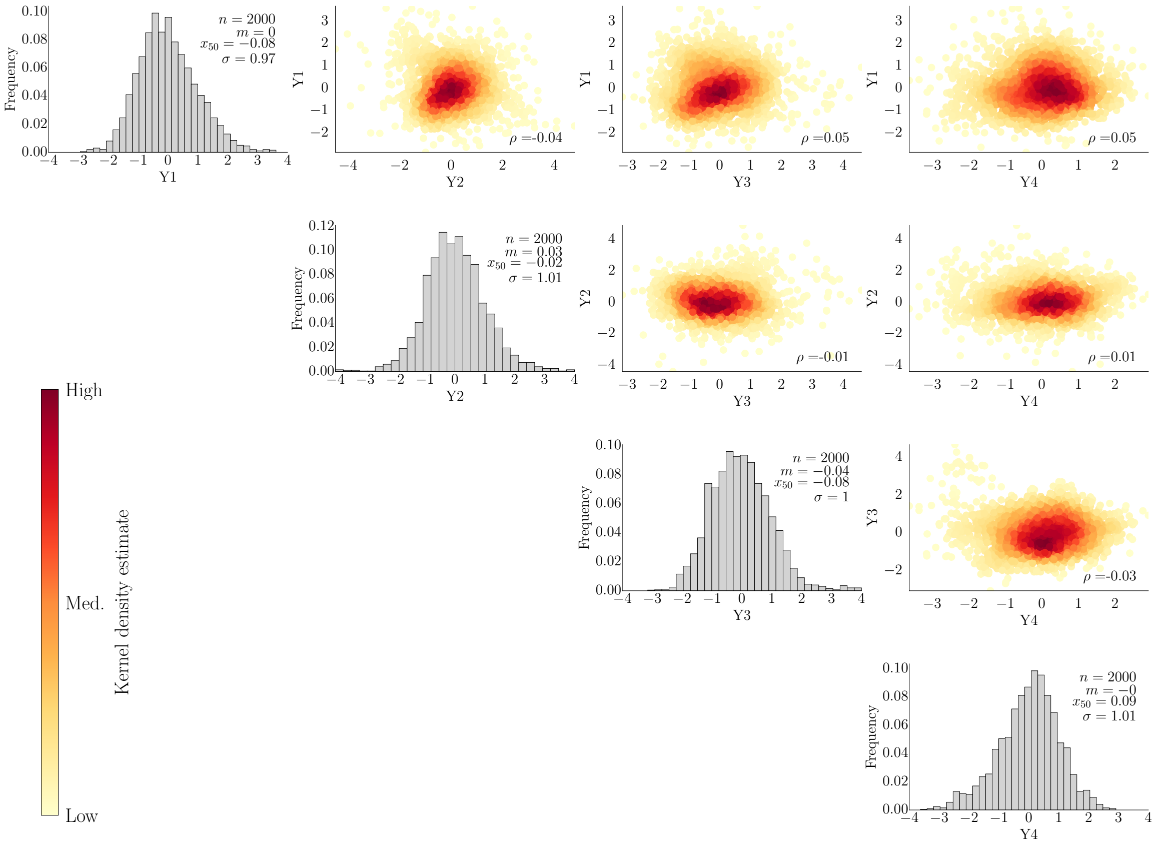
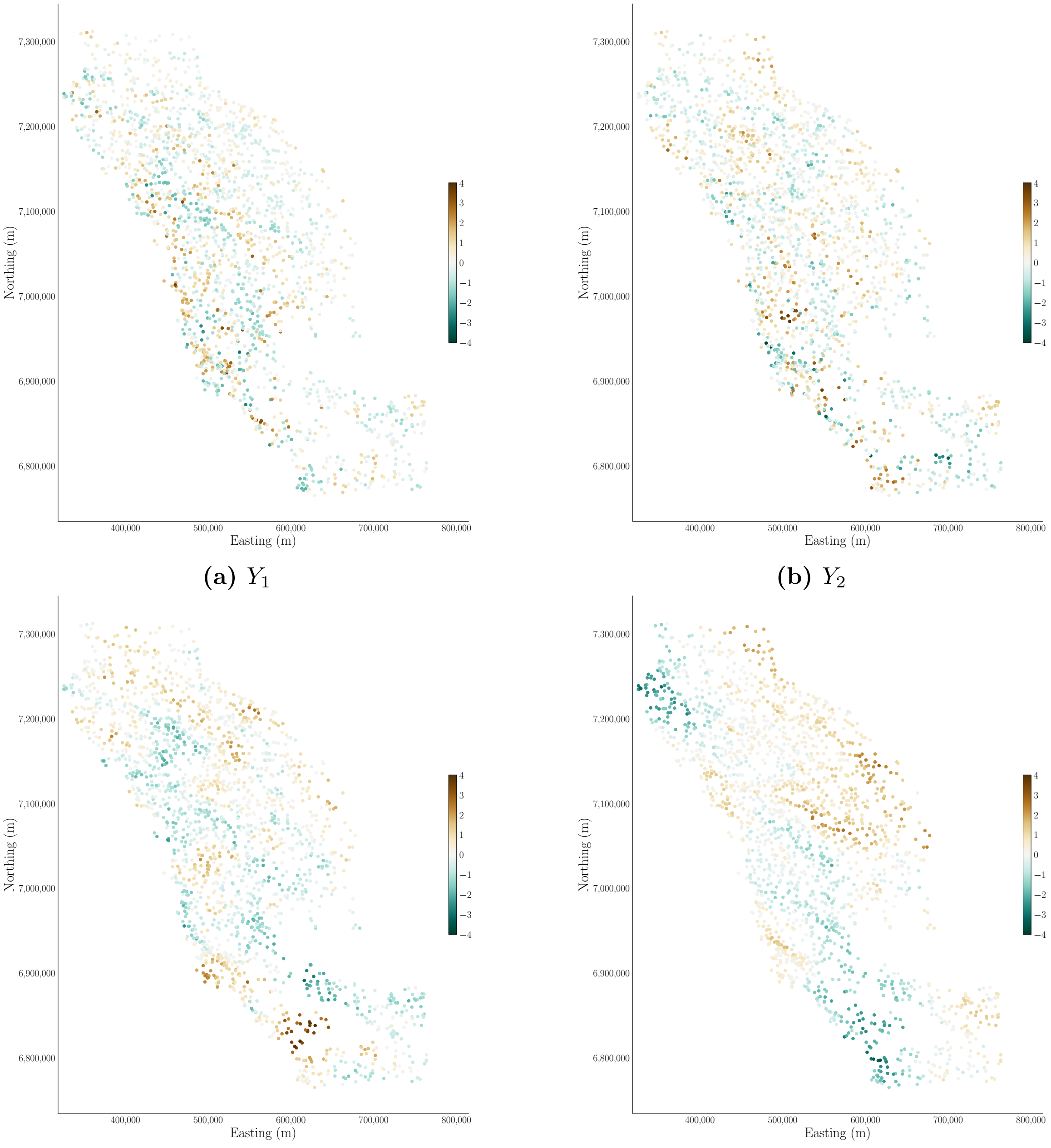


Figure 6.10: The distribution and cross correlation of the imputed factors in realization #50. The kernel density estimator is plotted.

The direct and cross variograms of the imputed factors were calculated for all realizations and are plotted in Figures 6.14 and 6.15 for the major and minor directions of continuity. Note that the ergodicity is more relevant in the minor direction and in the structures with longer ranges. The correlation seen in the cross variograms for a few realizations is not a reason for concern since on average the imputed variograms show no significant departure from zero. Any departure from the theoretical variograms is explained by the limited number of sample locations. A good variogram reproduction is expected when simulating the factors on a grid (next section).

An important practical aspect of the imputation workflow concerns the drawing of uncorrelated Gaussian random variables used in the simulation of the y values, see vector r in Equation 4.16. All values from that distribution are equally probable to be drawn. It was noted during the case study that extreme values yield simulated



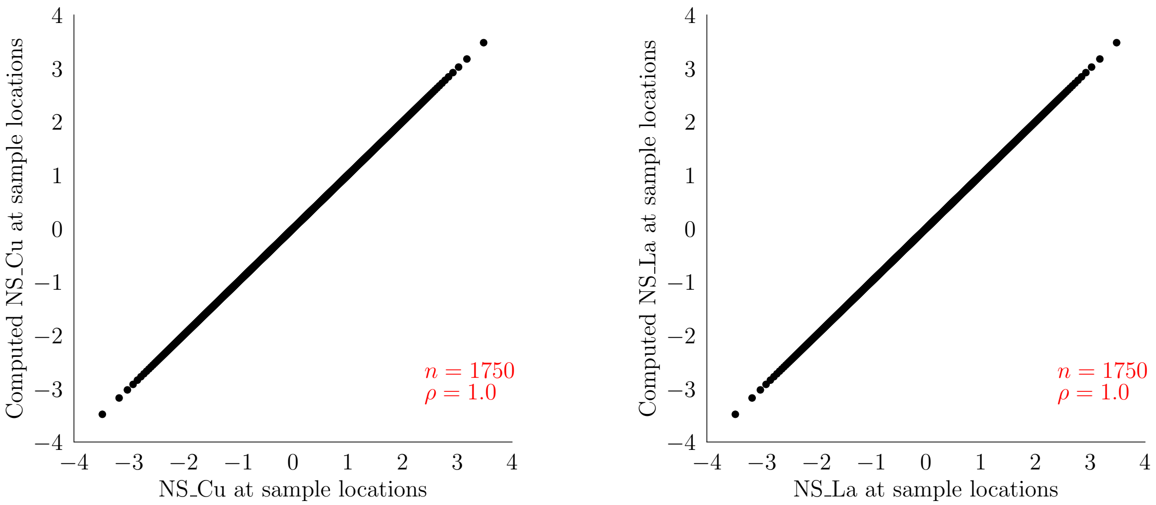
(a) Y_1

(b) Y_2

(c) Y_3

(d) Y_4

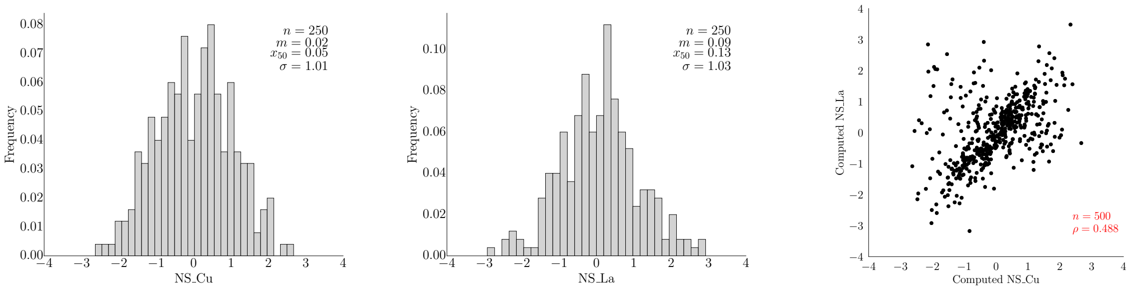
Figure 6.11: Realization #50 of each imputed factor Y at the data locations.



(a) Reference versus computed values of NS Cu

(b) Reference versus computed values of NS La

Figure 6.12: Scatterplot of the original data values and the respective computed values from any realization of the imputed Y factors at the collocated data locations.



(a) NS Cu

(b) NS La

(c) Simulated correlation

Figure 6.13: Distribution and correlation of the computed variables at the non-collocated data locations for realization #50.

factors that departure significantly from the multivariate normal with considerable cross-correlation. A temporary solution is to discard any set of y values generated from extreme values in r. In the case study a total of 11 realizations were discarded and regenerated, that is, approximately 1 every 18 realizations.

The imputed factors can be used to compute missing variables at the data locations. Another aspect of the IFS methodology is that the imputed factors can be then simulated on any locations, allowing for the computation of the variables at unsampled locations. Simulated factors over a large number of locations are expected to show an improved variogram reproduction when compared to the imputed

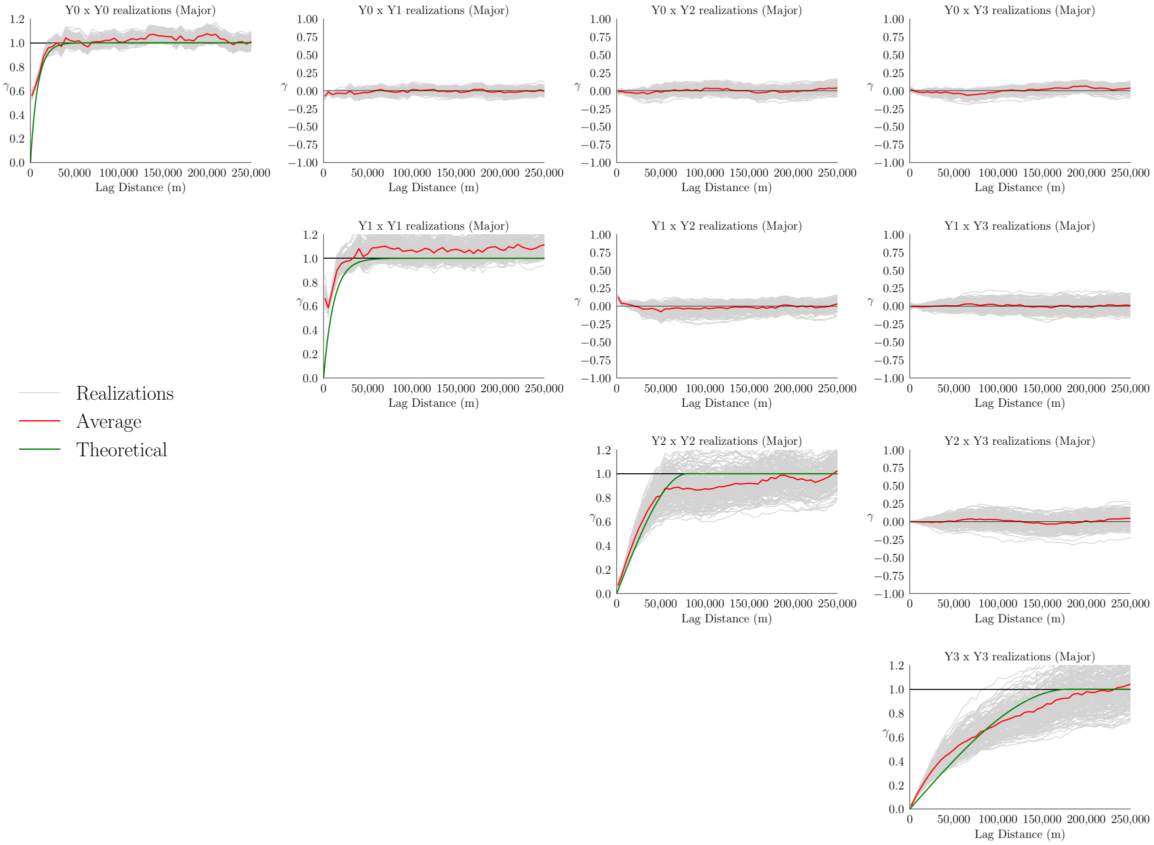


Figure 6.14: Variogram reproduction of the imputed Y factors at the data locations at the major direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the theoretical variogram of that factor as fitted in the LMC.

factors, as per the central limit theorem.

6.6 Simulation of the factors on a grid and checking

In this section, the results of the simulation of the factors on a grid are discussed. A grid containing 248 and 305 nodes respectively in the easting and northing directions, with a node spacing in both directions of 2 Km is considered. A keyout is used to ensure that only nodes inside the project area are simulated. The realizations of the Y factors at the data locations are used as the input data in the simulation of the factors in the grid. Sequential Gaussian Simulation is used to conditionally simulate

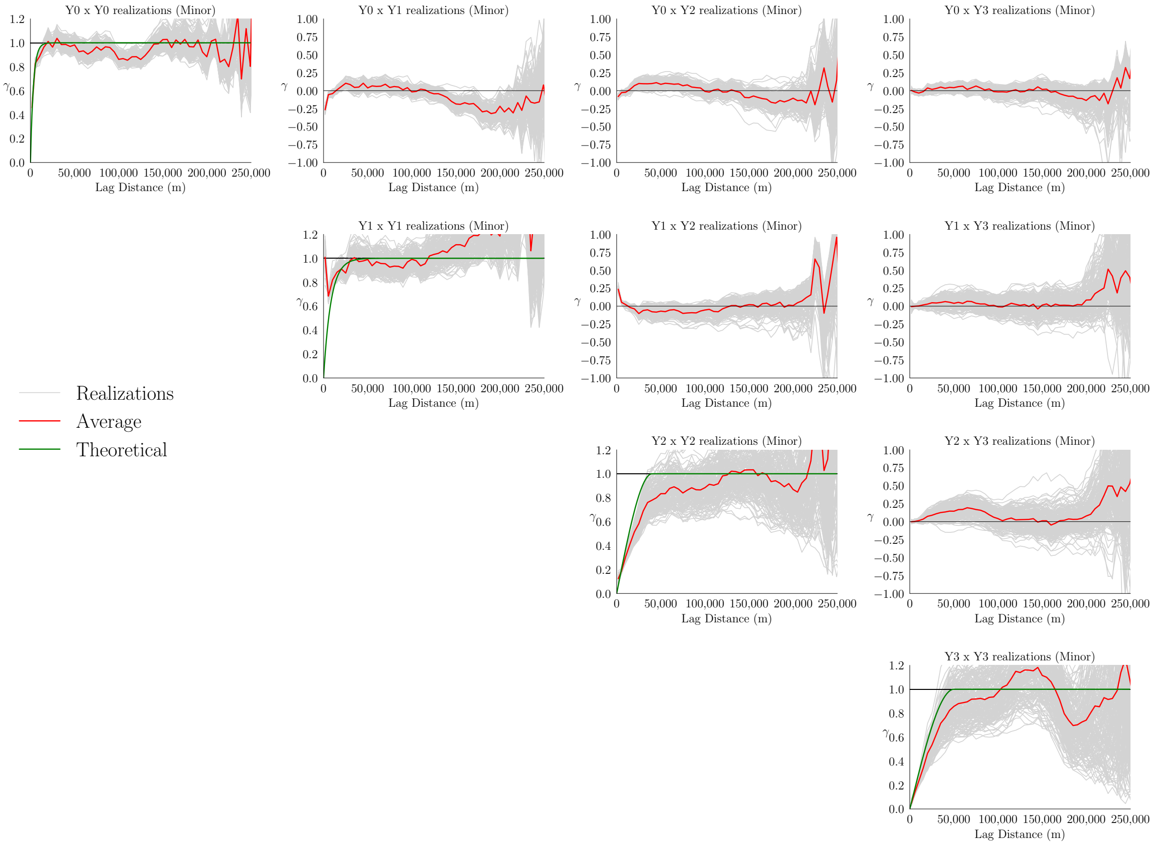
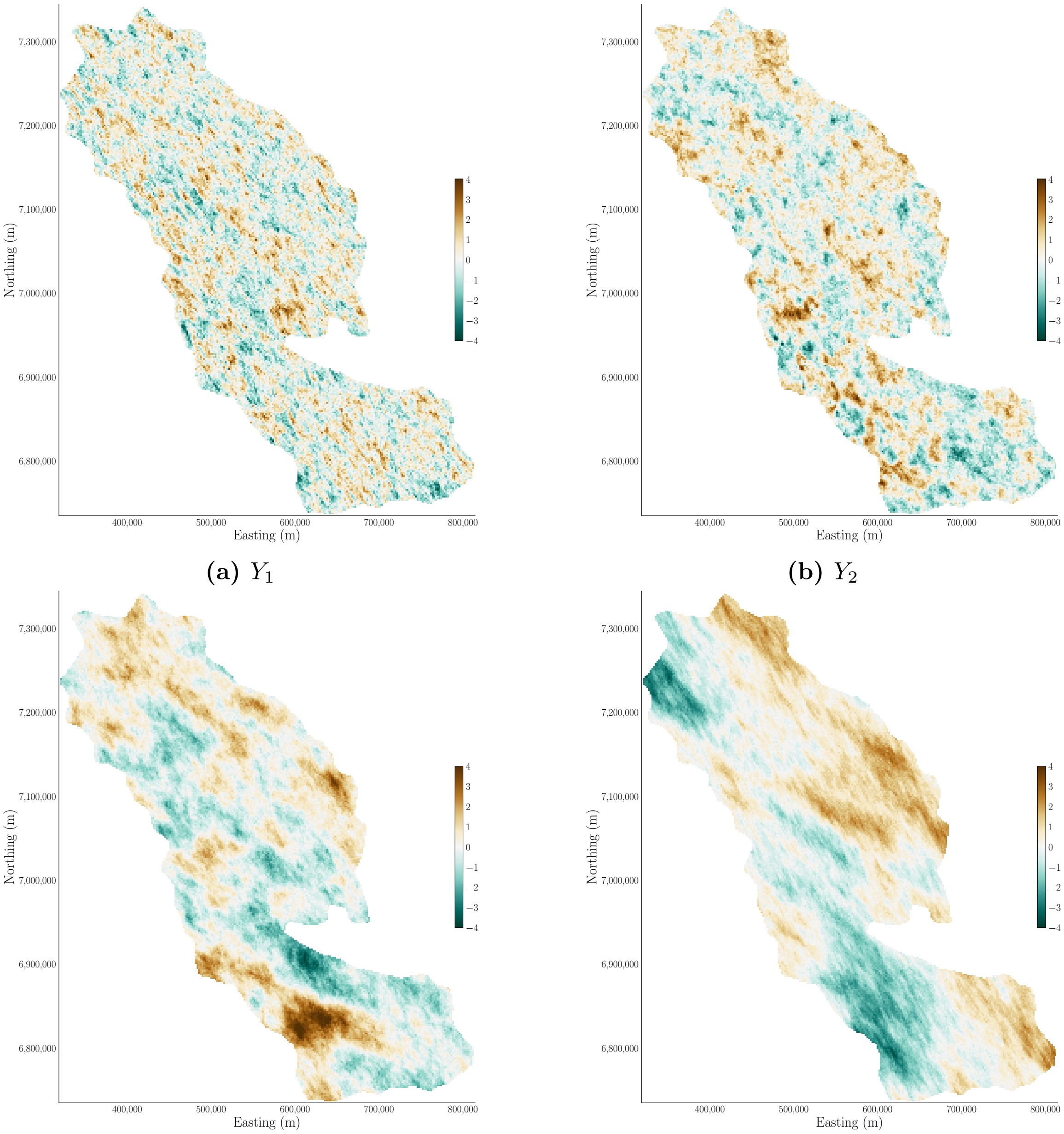


Figure 6.15: Variogram reproduction of the imputed Y factors at the data locations at the minor direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the theoretical variogram of that factor as fitted in the LMC.

Y_1 and Y_2 factors since it shows good reproduction of the exponential structure with a short range (Manchuk and Deutsch, 2015). Moving average is used to simulate Y_3 factor because of its fast implementation with a spherical structure and good reproduction when the variogram range is relatively larger than the node spacing (Cabral Pinto and Deutsch, 2017d). The last factor shows a large range in both directions of anisotropy and for this reason it is simulated with spectral simulation, a method more suitable for trend-like structures (Cabral Pinto and Deutsch, 2017b). The realizations generated by moving average and spectral method are conditioned by parallel global kriging (Manchuk and Deutsch, 2017). Realization #50 of the simulated factors on the grid is shown in Figure 6.16.

Reproduction of the direct and cross variograms is also checked in both directions of continuity and shown in Figures 6.17 and 6.18. There is good reproduction of the



(a) Y_1

(b) Y_2

(c) Y_3

(d) Y_4

Figure 6.16: The conditional realization #50 of each factor Y on the grid locations.

direct variograms in both directions, and that on average, there is no spatial cross-correlation between the factors as expected. Note that compared to variograms of the imputed factors (Figures 6.14 and 6.15) most of the correlation seen in the cross variograms is mitigated and that on average the direct simulated variograms are very close to the reference models.

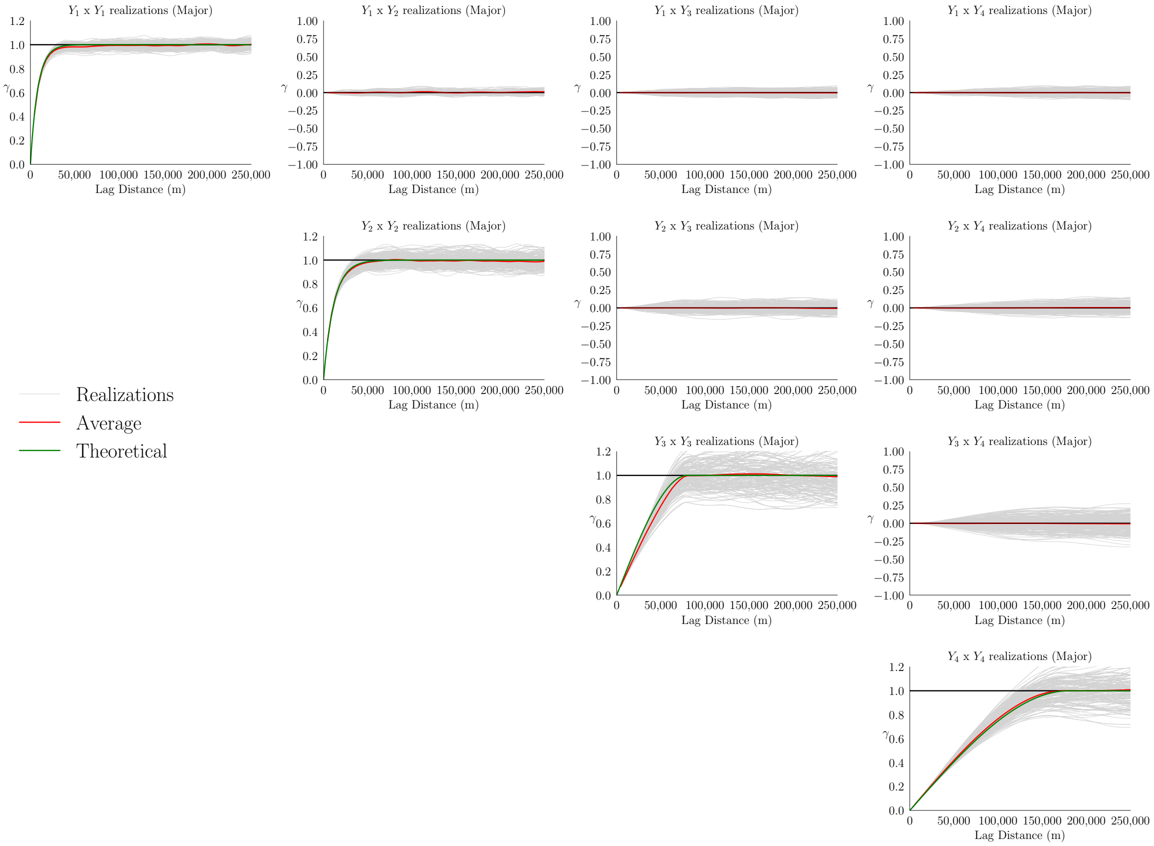


Figure 6.17: Variogram reproduction of the simulated Y factors on the grid at the major direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the theoretical variogram of that factor as fitted in the LMC.

The simulated factors are used to compute realizations of the variables on the grid, one realization is shown in Figure 6.19. The simulated histograms are standard Gaussian with little fluctuation around the reference mean. The correlation coefficient is also reproduced, with a mean of 0.517 over all realizations, as shown in Figure 6.20.

The reproduction of the LMC for the computed variables is also checked and plotted for both directions of continuity in Figures 6.21 and 6.22. There is good

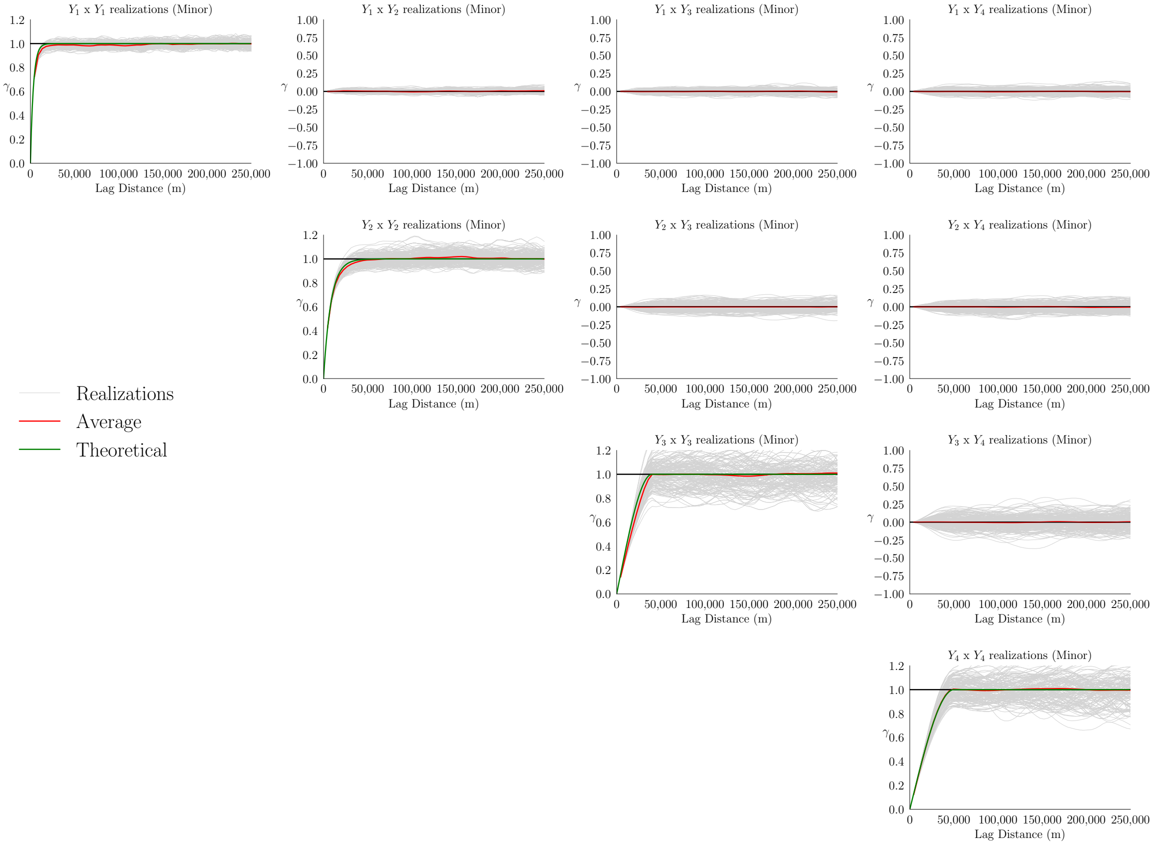
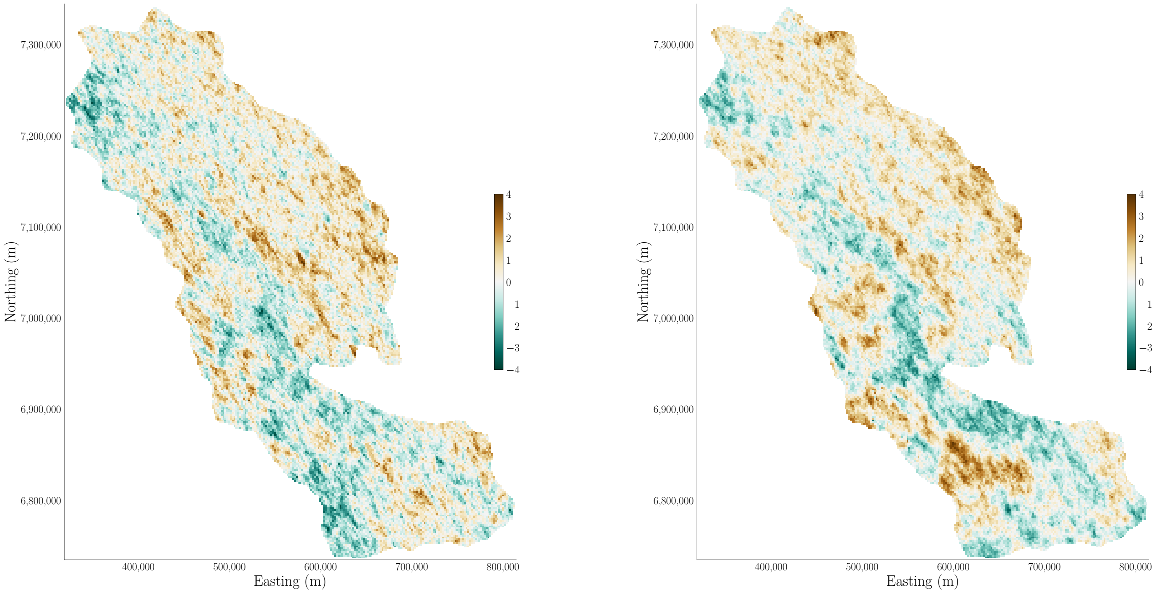


Figure 6.18: Variogram reproduction of the simulated Y factors on the grid at the minor direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the theoretical variogram of that factor as fitted in the LMC.

reproduction of the variograms in both directions.

The simulated variable values are back-transformed to their original units and the reproduction of the histogram is checked and plotted in Figure 6.23. There is good reproduction of the mean for both variables, the simulated mean of Cu and La are 0.13% and 1.56% higher than their respective reference declustered mean.

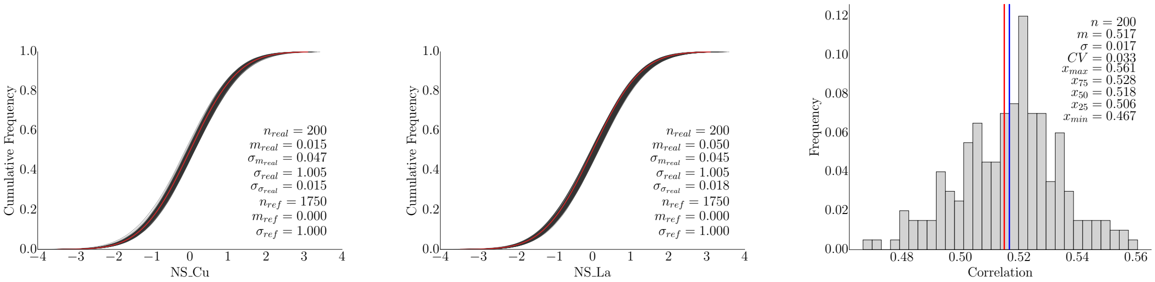
The probabilistic accuracy is evaluated for all realizations by discretizing the probability interval in nine bins, equally spaced in probability intervals of 0.1. The calculated predicted frequency in each bin is compared with the actual fraction of the data in each one of these bins. This analysis is performed on the original units of the variables at the validation locations. The realizations are sampled at the left out data locations and compared against the true values of the data. The predicted and actual fraction in each interval are plotted in Figure 6.24. Points falling on the 45



(a) NS_Cu

(b) NS_La

Figure 6.19: The conditional realization #50 of each variable on the grid.



(a) NS_Cu

(b) NS_La

(c) Correlation

Figure 6.20: Reproduction of histogram and correlation of the normal scores of the variables over all realizations. The red line in the histogram reproduction plots represent the reference distribution, and the blue line in the correlation plot represent the mean. The reference correlation is 0.515.

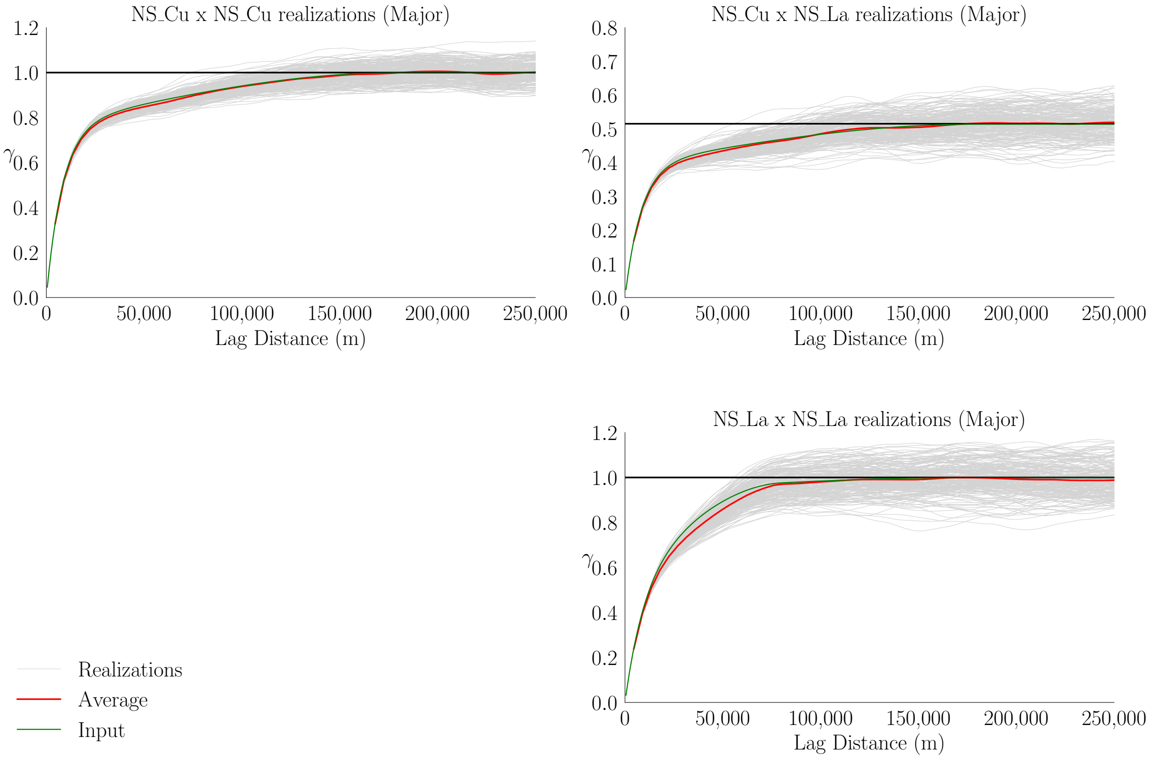


Figure 6.21: LMC reproduction at major direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the input variogram from the LMC.

degrees line represent the predictions are accurate and precise. Points falling above the 45 degrees line indicate that the local uncertainty may be too high, whereas points falling below the line indicate a low variance in the estimates. The analysis of Figure 6.24 indicates good accuracy and precision for both variables.

The local accuracy of the simulated variables is examined. The simulated e-type of all realizations of the original units of Cu and La are compared against the validation data. The realizations are sampled at the left out data locations and averaged over all realizations to compute the smooth e-type means of the variables. The e-type means are then compared against the true values and results are plotted in the scatterplots of Figure 6.25. For both variables, the root mean square error (RMSE) is small, the slope of the regression line (SoR) is slightly smaller than one and may indicate conditional bias. This bias is acceptable at this moment given the complex nature of the geochemical data and the smoothing nature of the (e-type) estimates. Overall there is good correlation between the e-type means the

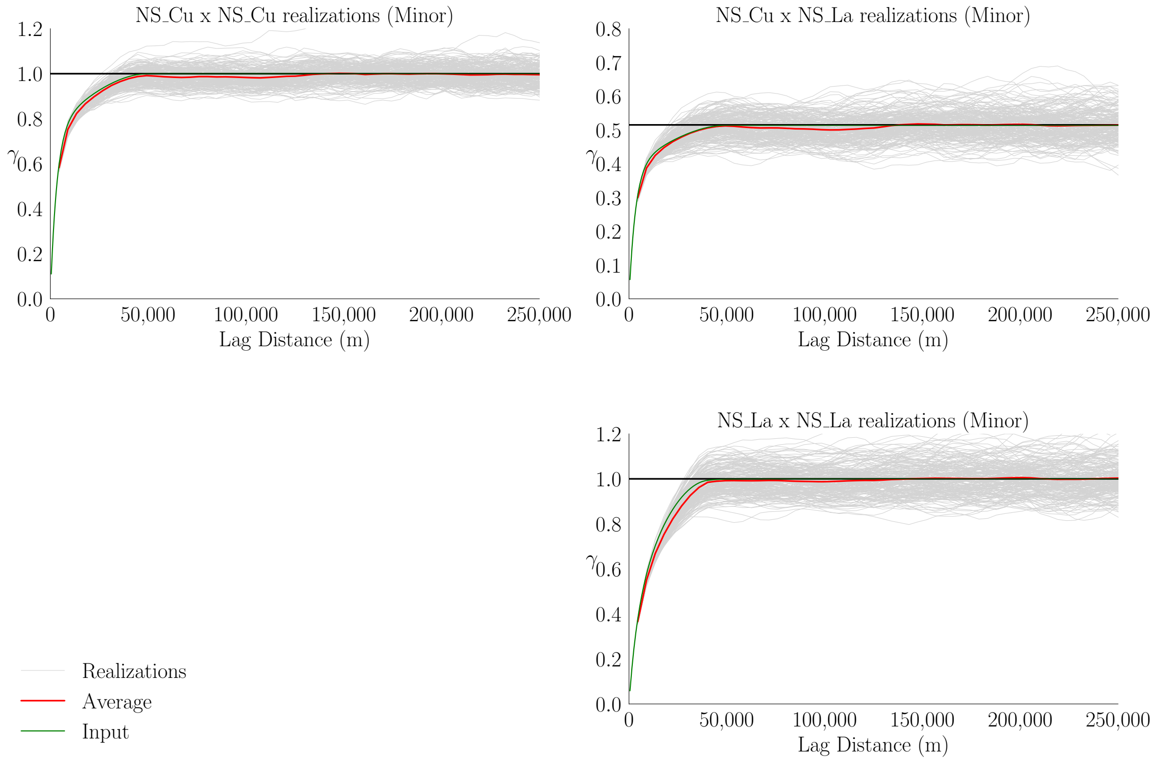
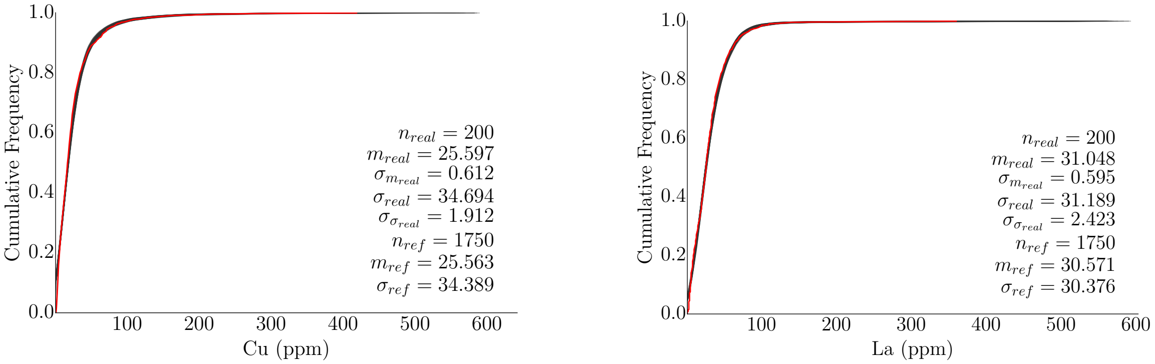


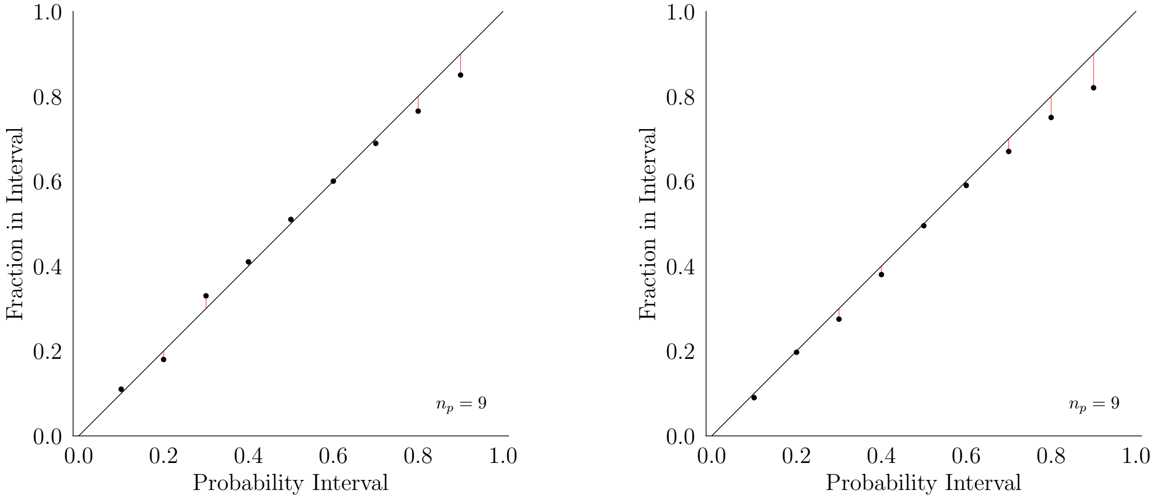
Figure 6.22: LMC reproduction at minor direction of continuity. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the input variogram from the LMC.



(a) Histogram reproduction of Cu

(b) Histogram reproduction of La

Figure 6.23: Reproduction of histogram of the original unit variables over all realizations. The cumulative distribution function is shown. The red line represents the reference distribution.

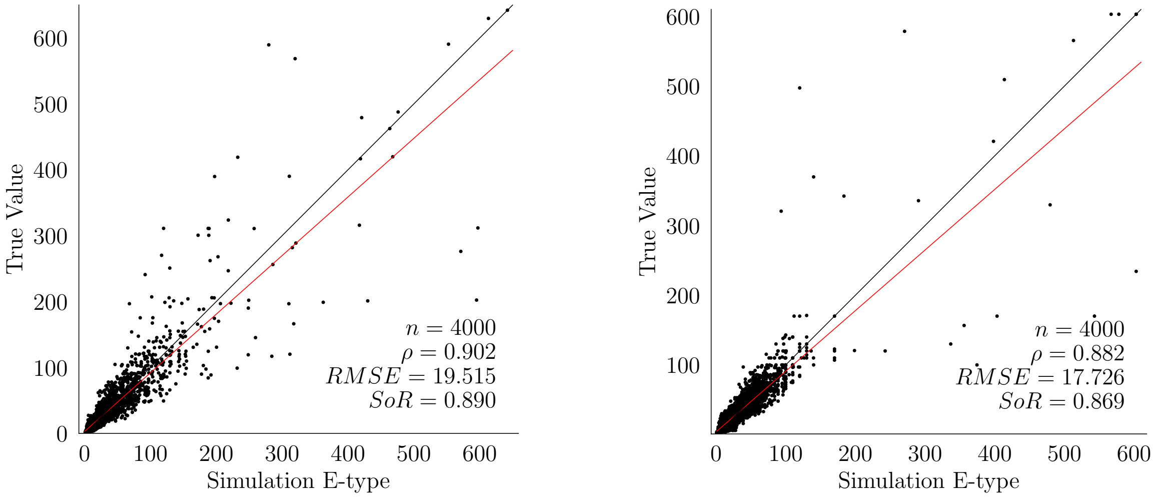


(a) Accuracy plot of Cu

(b) Accuracy plot of La

Figure 6.24: Accuracy plots of the computed distributions of Cu (a) and La (b) over all realizations at the validation locations.

validation data.



(a) Scatterplot of Cu

(b) Scatterplot of La

Figure 6.25: Scatterplots and summary statistics that compare the computed e-type with associated true values for Cu (a) and La (b) at the validation locations. The correlation, root mean square error (RMSE) and the slope of the regression (SoR) (red line) are shown.

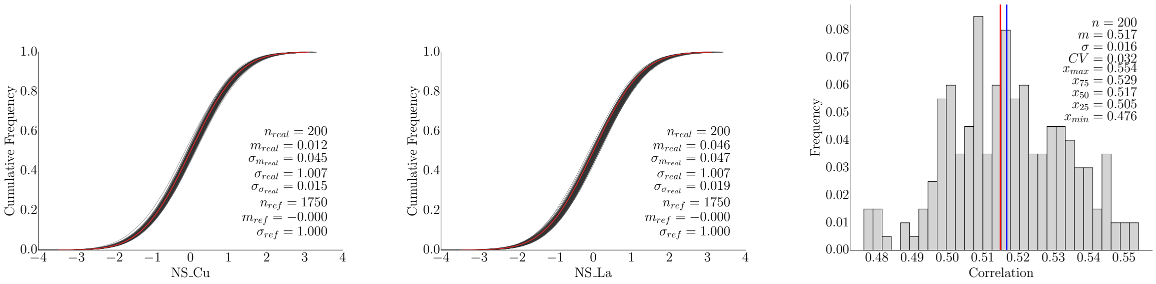
In the IFS methodology, the simulation of Cu and La requires fitting an LMC to the NS transform of the variables. The matrix with the LMC coefficients is then used in the decomposition of the variables into independent factors. The factors are imputed at the data locations and can be used to compute missing geological variables. The factors can be simulated at any locations and the original variables computed afterwards. In the IFS methodology the variables are not simulated, they are rather computed from the imputed or simulated factors. The conventional approach is to directly simulate the variables with cokriging. The next section compares the result of IFS and cosimulation with cokriging methods in simulating Cu and La.

6.7 Conventional cosimulation with cokriging

Cosimulation with cokriging is the conventional method to simulate unequally sampled variables. SGS is often chosen as the simulation algorithm since conditioning is done on the fly and all variables are simulated simultaneously. Local distributions are inferred from cokriging and different realizations are generated by sampling these distributions. Because the same algorithm is used in simulation, the reproduction of different structures of the LMC may be affected and departures from the input reference model may occur. In the IFS methodology, the LMC structures are simulated independently with the corresponding best algorithm for that structure. Therefore, a better reproduction of the LMC is achieved.

To compare the IFS method to the conventional approach, SGS is used to cosimulate Cu and La at the same grid locations used in the IFS case study. For consistency, the same declustered distributions and LMC model are used. A large number of data is used in simulation and a total of 200 realizations are generated. In normal scores, the simulated variables and correlation from both methods are very similar, as shown in Figure 6.26. The same average simulated correlation of 0.517 is seen in the IFS and in the cosimulation methods.

The reproduction of the variograms are shown in Figures 6.27 and 6.28 for both directions of continuity. Variograms are analyzed for their reproduction for distances less or equal the variogram range. A visual inspection shows a better reproduction of the direct variogram of Cu and the cross variogram in the IFS methodology for both directions of continuity. The direct variogram of La is better



(a) NS Cu **(b)** NS La **(c)** Correlation

Figure 6.26: Reproduction of histogram and correlation of the normal scores of the variables over all realizations with the conventional approach. The red line in the histogram reproduction plots represent the reference distribution, and the blue line in the correlation plot represent the mean. The reference correlation is 0.515. To be compared to Figure 6.20.

reproduced with the conventional approach. The error in the simulated variogram is numerically calculated from the difference between the average simulated variogram and the reference model for a set of lag distances. The average variogram error in both directions is calculated and shown in Table 6.4. The calculated error supports the results from the visual analysis of the variograms. Despite both methods showing a good overall variogram reproduction, the IFS methodology overtakes the conventional approach in four of the six variograms.

Table 6.4: Average error of the average simulated variograms and the LMC model. Text in bold represent the model with the average error closer to zero.

	Major			Minor		
	NS Cu x NS Cu	NS Cu x NS La	NS La x NS La	NS Cu x NS Cu	NS Cu x NS La	NS La x NS La
IFS	**-0.005**	**-0.002**	-0.012	**-0.013**	**-0.004**	-0.023
Conventional	-0.016	0.02	**-0.001**	-0.021	0.005	**-0.019**

Similar to the IFS method, an overall good reproduction of the histograms of the original units is achieved, as shown in Figure 6.29. The simulated mean of Cu and La are respectively 1.27% and 2.90% higher than their reference declustered means. These values are approximately 9.76 and 1.85 times greater than the simulated means of Cu and La with the IFS methodology.

In this cases study, the results from the conventional approach are similar to those generated with the IFS method. The IFS methodology shows a better reproduction of the variograms of Cu and the cross variogram between Cu and La. The

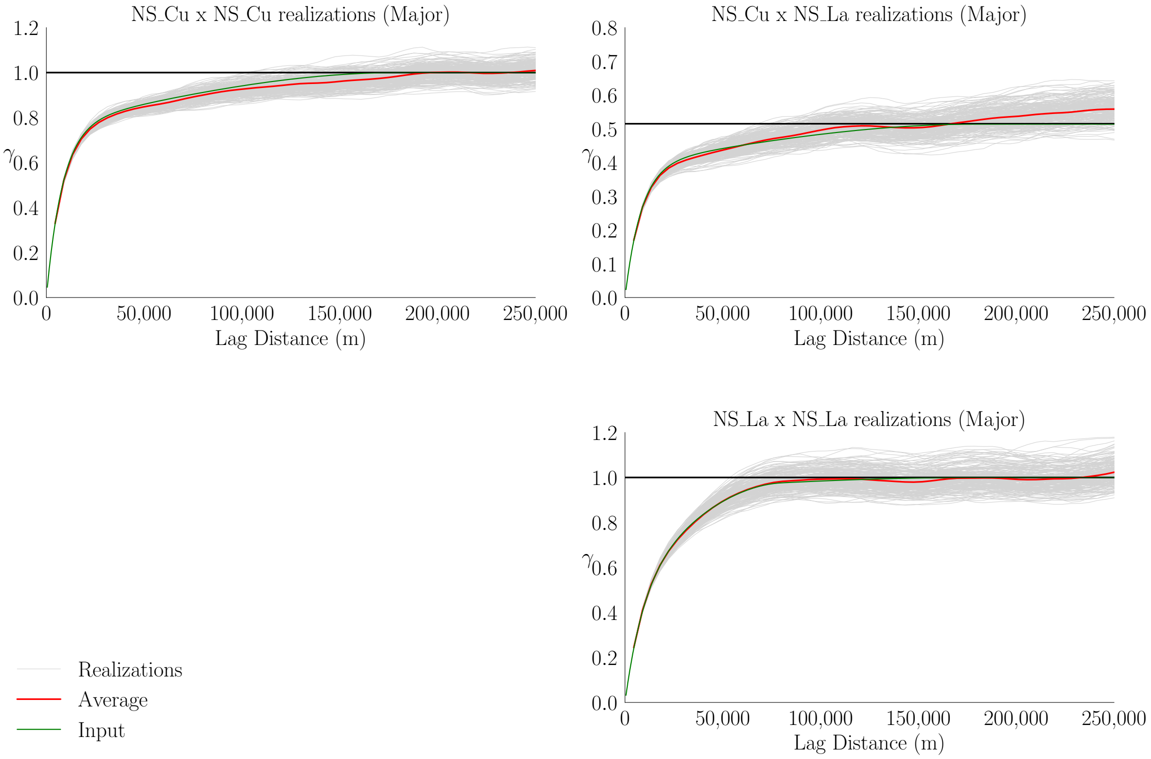


Figure 6.27: LMC reproduction at major direction of continuity of the conventional approach. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the input variogram from the LMC. To be compared to Figure 6.21.

conventional approach reproduces better the variogram of La. A better reproduction of the histograms is achieved with the IFS methodology.

6.8 Conclusion

This case study demonstrates in practice the use of the IFS methodology in multivariate modeling. An LMC is fitted to the normal scores transform of two unequally sampled geochemical variables. The LMC is used to decompose the variables into a set of independent normally distributed factors that are imputed at the data locations. At locations where both variables are available the computed variables from any of the 200 realizations of the imputed factors are exactly the same of the existing variables. At locations where a variable is missing the imputed factors are used to compute different realizations of the missing variable. Once the factors are im-

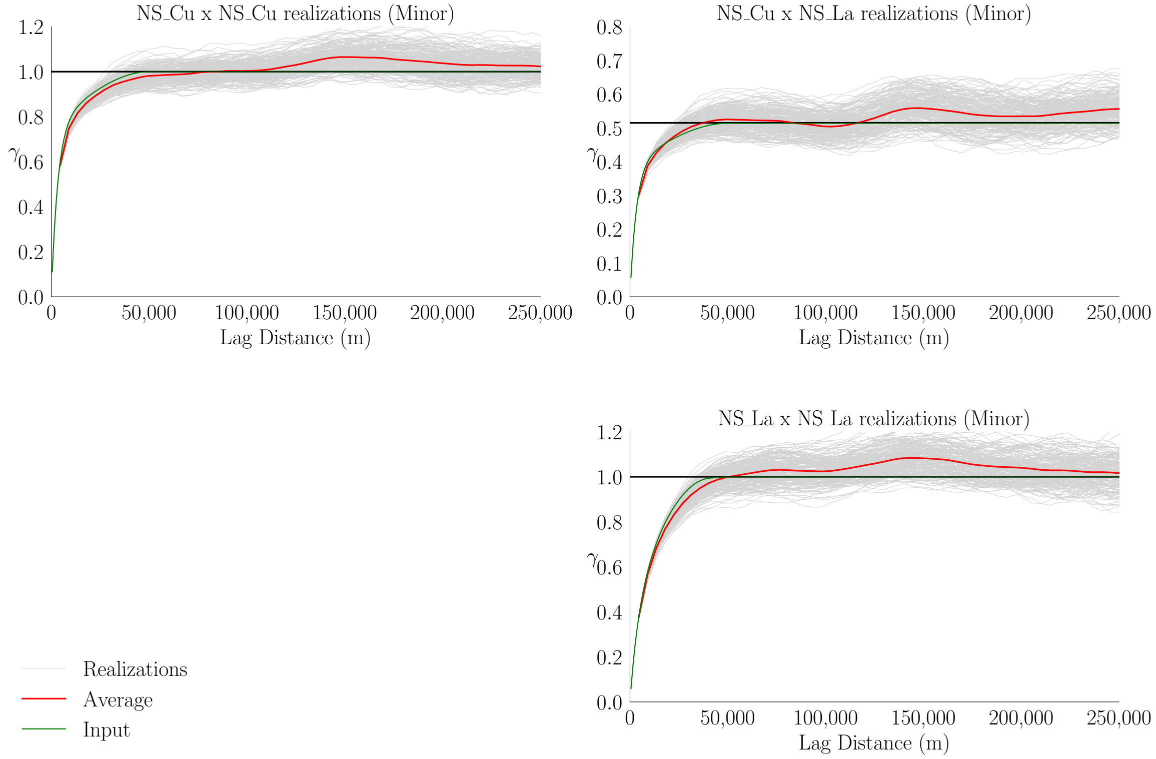
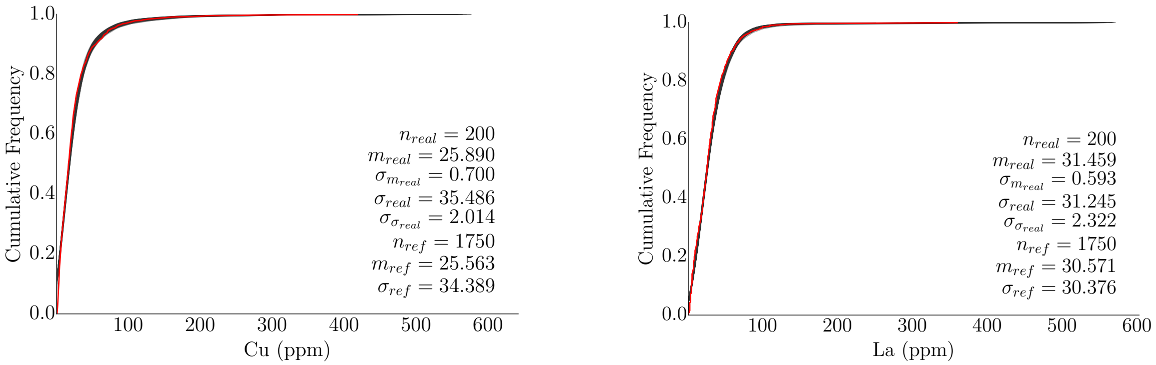


Figure 6.28: LMC reproduction at major direction of continuity of the conventional approach. The grey lines represent the variograms of the realizations, the red line represents the average variogram of the realizations, and the green line is the input variogram from the LMC. To be compared to Figure 6.22.



(a) Histogram reproduction of Cu

(b) Histogram reproduction of La

Figure 6.29: Reproduction of histogram of the original unit variables over all realizations. The cumulative distribution function is shown. The red line represents the reference distribution. To be compared to Figure 6.23.

puted they are simulated on a grid, each factor is simulated with the best algorithm for the respective spatial structure. SGS, moving average, and spectral simulation are used as simulation algorithm. The simulated factors are independent, follow a standard normal distribution, and reproduce the input spatial structure. The original variables are computed from the simulated factors and show good histogram and variogram reproduction. When compared to cosimulation with cokriging, the IFS methodology demonstrates a better capacity of reproducing input modeling parameters such as the histograms and the LMC.

Chapter 7

Conclusions

This final chapter reviews the contributions, discusses the limitations of the methodologies, and propose future work for the developments in this thesis. This thesis makes primary contributions to the field of multivariate modeling of equally sampled data with the PostPPMT methodology, and unequally sampled data with the IFS methodology. These two methodologies integrate and expand the use of well established and recently developed geostatistical tools such as the multiGaussian approach and PPMT, modernize the use of the LMC, and benefit from the flexibility of cokriging in integrating different multiple data types. These contributions were motivated by a series of studies and reports carried out on a large data set of geochemical data collected in the Northwest Territories in Canada. The findings and developments of these case studies are presented in Chapters 3 and 6.

7.1 Probabilistic assessement of multivariate criteria

One of the gaps in current multivariate modeling techniques is the lack of a computationally efficient methodology that provides accurate multivariate data-value dependent measures of local uncertainty. The development of such methodology provides a solution when local estimates or local uncertainty measures are the goal of the study. The PostPPMT methodology, developed in Chapter 3, proposes a framework for assessing the local multivariate joint-probability for a given multivariate criteria to occur. Multivariate criteria are rules being applied to many variables at the same time. Some applications of these criteria in multivariate geostatistical modeling are in exploration geology, mine planning and operations. The PostPPMT methodology provides a way for multivariate modeling in the context of such criteria.

The PostPPMT methodology is a hybrid of the post-multiGaussian approach and the PPMT transform to decorrelate the multivariate data. Considering a single independent variable in PPMT is equivalent to the multiGaussian approach long used in geostatistics. The idea is to use the PPMT transform to remove the complex relationship between the variables and then model each factor independently. The variograms models are inferred and simple kriging is performed to compute a local mean and local variance at each unsampled location for each factor. This allows optimal selection of the kriging parameters such as number of data used for kriging and search strategies for each factor. The kriged mean and variance fully define the local conditional Gaussian distribution. The back transformation and post processing of the local distributions are considered for resources, local multivariate criteria assessment, and for checking simulation. A large number of random quantiles is recommended to provide stable results. The process of sampling the local distributions and back transform the quantiles with the PPMT transformation table is highly parallelizable. The calculation of joint-probabilities is fast and efficient even for a large number of realizations.

The proposed workflow is subject to the standard geostatistical limitations, such as the requirement for stationarity and usage of the variogram as a measure of two-point statistics. The assumptions of multivariate Gaussianity are still valid, the normal scores transformation does not guarantee such assumption, thus the PPMT transform is used. Within these limitations it is possible to check and verify models with conformance to modeling assumptions. When the PostPPMT methodology is considered, the many steps of the workflow can be checked to ensure the methodology is correctly being used. Consider the following practical recommendations. Declustering is recommended to calculate a representative multivariate distribution that accounts for different sampling density zones. The PPMT factors must be checked for normality, since they are expected to be standard normal. Checking the correlation of the PPMT factors is also recommended. PPMT guarantees decorrelation at lag zero but correlation at other lags may exist. The calculation of experimental cross-variograms can be used to investigate cross-correlation at non-zero lag distances. Practice has shown that a correlation of 20% is significant to introduce potential bias in the model. In such cases, MAF is recommended to decorrelate the factors at other lags with considerable correlation. Due to all transformations that PPMT imposes to the multivariate data, the PPMT variograms can be difficult to model. The normal scores variogram is recommended when the correlation between the normal scores and the correspondent PPMT factor is high and a reliable PPMT variogram cannot be achieved. An important step of the methodology is the sampling algorithm and number of realizations. A random sampling is recommended over a regularly spaced approach because of the potentially high dimensionality. The sampling algorithm must also be checked to guarantee that quantiles are drawn uniformly over many realizations. The number of realizations must be large to avoid significant noise in the variance or any probability sensitive to the tails of the distributions. Thousands of realizations are recommended to provide stable results.

The postPPMT is not suitable for situations where multilocation uncertainty is required. The probability of meeting multivariate criteria is calculated from the local distributions built from the kriged factors. The kriged factors are calculated based on the multivariate data inside the search defined by the variograms of the factors. The probabilities calculated in a location are not used to assess the probability of the next location, thus the calculation of any multilocation measures of uncertainty is not possible. Geostatistical simulation is recommended in such cases. The methodology is limited to cases when all variables are equally sampled because the PPMT transform requires homotopic data.

Some ideas are suggested for future research to improve the methodology. Future work could consider the implementation of PPMT with exhaustive secondary data. Although the primary variables can be decorrelated, they remain dependent through the secondary data. A hierarchical decorrelation approach with stepwise conditional transformation is proposed to address this problem (Manchuk et al., 2019) and could be implemented into the PostPPMT framework. Although the

PPMT transform is used, further work could explore other decorrelation techniques such as PCA, MAF, sphering, stepwise conditional transform, and Gaussian mixture models. The advantage of using PPMT is that the forward and backward transformations are stored together in the same transformation table and are easily accessed in the algorithm. In the presence of missing data, multivariate data imputation techniques can be considered (Barnett and Deutsch, 2015; Silva and Deutsch, 2018). Additional work may consider the LMC and cokriging as an alternative to geostatistical imputation. Cokriging may be used to define the local conditional distributions. Another interesting area to explore is the use of indicator kriging where the indicator would take the value one when the multivariate criteria is met at a data point and zero otherwise. A cross-validation study can be used to compare both approaches.

7.2 Independent factor simulation methodology

The main contribution of this thesis is in the modeling of multivariate data in the presence of unequally sampled data. In the presence of missing data, unequal sampling, or different data types, variables are cosimulated with the LMC and cokriging. The LMC is a model of coregionalization that combines multiple data types at different locations and different data support into the same framework for estimation with cokriging. Fitting a LMC to a large number of variables is challenging. The problem lies in fitting the LMC to the experimental covariances accounting for anisotropy in the data while respecting the LMC constraints. Fitting an LMC may require the use of a large number of nested structures with different covariance shapes and variance contributions. The current approach is to simulate the variables with the fitted LMC and cokriging. The simulation is often performed with a chosen algorithm. Given the complexity of the LMC, using the same algorithm to simulate all structures of the LMC is not optimal. The development of the IFS methodology discussed in Chapters 4 and 6 proposes a solution to this problem.

The IFS methodology uses the LMC and BSS to decompose and factorize the normal scores transform of the original variables into a set of independent normal latent variables. The factors are then modeled independently with an appropriate algorithm and used to reconstruct the original variables. In the context of completely unequally sampled data, the methodology provides a way to impute factors at the data locations of all data types facilitating further independent simulation. Given that the factors are independent and have a single spatial covariance function, the most appropriate algorithm may be selected and applied to each factor independently. The modeling of independent factors permits practical and easy model checking at each step. As a consequence of modeling independent factors, another contribution of this thesis is the discussion on common Gaussian algorithms and practical recommendations of optimal algorithm selection. This is discussed in Chapter 5. Each algorithm has a range of spatial covariance functions and grid parameters where they perform with high efficiency in terms of variogram and histogram reproduction. Using the optimal algorithm leads to a better variogram reproduction for each factor and better histogram reproduction of each variable, as shown in the case study with geochemical data.

There are limitations to the IFS methodology, and some ideas for future work are given to address them. The IFS methodology can be applied to equally and unequally sampled data. In the presence of homotopic data, decorrelation techniques may provide a more efficient and simpler workflow than IFS. However, in cases where decorrelation techniques do not successfully remove spatial cross-correlation the IFS methodology may be considered. In the presence of unequally sampled data, simulation is restricted to the LMC and cokriging. In such cases the IFS methodology provides an alternative to cosimulation with cokriging that simplifies modeling check and improve reproduction of the input modeling parameters. Similar to the PostPPMT, the IFS methodology is subject to the standard geostatistical limitations, such as the requirement for stationarity. Practice has shown that the methodology is sensitive to second order non-stationarity. Proper trending modeling must be considered prior LMC modeling. The assumptions of multivariate Gaussianity are still valid. Such assumptions are difficult to check in the presence of different data types and heterotopic data. The normal scores transformation is a requirement of the methodology and does not guarantee multivariate normality.

Another limitation is that the IFS is restricted to cases when there are more factors than there are original variables. The number of structures of the LMC must be always greater than the number of variables. Most software used in the mining industry limit the number of structures of the LMC to a few, usually less than five. Practical implementation of the IFS methodology must consider an increasing number of factors. The singular value decomposition poses another limitation to the methodology. SVD provides a way to find the two norm solutions of a matrix A. The algorithm may converge to an approximate solution of A. SVD stands out from other matrix decomposition techniques because it provides the best approximation to A. However, this comes at a high cost, SVD is considered an expensive decomposition (Chan, 1986; Watkins, 2002). The run-time and memory required to perform SVD is demonstrated in Chapter 4 for a different number of data and factors. The current implementation limits the number of data and factors to less than 10,000 and 10 respectively. Further work may consider alternatives to SVD, such as independent component analysis, nonnegative matrix factorization, sparse component analysis, eigenvalue decomposition, Cholesky factorization, and uncorrelated component analysis (Cant et al., 2015; Chang et al., 2006, 1999; Comon, 1994). A tradeoff between the calculation of low-rank approximations to a matrix and other solutions for rank-deficient matrices must be found. Only SVD is implemented in the current methodology.

Finally, future work would improve how the nugget effect is utilized in the methodology. Proper decomposition of the nugget effect requires the identification and isolation of all possible combinations of the nugget effect. Each component would then be considered an isolate factor and modelled separately. The current implementation considers one completely shared nugget between the variables. Future development in the methodology would also consider the integration of parameter uncertainty, including histogram, correlation, and variogram uncertainty. No parameter uncertainty is considered in the current approach.

7.3 Software

A number of programs are developed for this thesis. They are developed for research and proof of concept purposes, but permit the application of the methodologies with large data sets and high resolution models. Major software developed as part of this thesis are discussed with details in the Appendix.

The `postPPMT` program implements the PostPPMT methodology in GSLIB format. The run-time to process the local distributions and back transform the large number of quantiles is greatly reduced with the `postPPMT` FORTRAN program (see section A.1). The program does not perform the PPMT transform but takes the transformation table as an input.

The IFS methodology is implemented in the GSLIB programs `LMC_IMP` and `LMC_COMP`, see section A.3. Both programs are compiled with INTEL compiler and multithread libraries. GSLIB subroutines are used for most of the calculations and data processing. The subroutines for matrix operations and SVD are imported from the LAPACK library (Anderson et al., 1999). The subroutines are collected and distributed with the original FORTRAN files to facilitate compilation. The `LMC_IMP` program implements all steps of the IFS methodology for factor extraction and imputation. The `LMC_COMP` program computes the Z variable values given the simulated Y factors and the LMC coefficients matrix.

A fast implementation of the moving average algorithm for spherical variograms is developed, see section A.2.

7.4 Final comments

Recall the thesis statement proposed in chapter 1: *The development of geostatistical modeling with multivariate complex relationships of unequally sampled data modernizes the use of the LMC and leads to improved high resolution geostatistical property models. The practical implementation of probabilistic assessment of uncertainty with multivariate criteria provides ways for multivariate modeling in the context of such criteria and adds* *practical value and theoretical insight for more complicated multifactor criteria.*

In this thesis, a methodology that uses the LMC to model complex relationships of unequally sampled data was developed. The LMC and BSS are combined together to create a framework that allows for independent factor imputation and simulation. These factors are decomposed from the multivariate data and linked together by the definition of the LMC. The current approach is to fit the LMC and use the LMC for cokriging. The IFS methodology proposes a modern use to the LMC. In this methodology, the variables are not simulated, but rather computed from the simulated factors. The independency of the factors allow for optimal selection of the simulation algorithm which improves the reproduction of the geostatistical property models.

The application of geostatistical modeling in the context of multivariate criteria is unclear. The implementation of the PostPPMT provides ways for modeling of complex multivariate rules in the context of estimation and local uncertainty assessment. The results could be used for resource calculation, to support classification decisions, provide a useful check on simulation-based workflows, provide a measure of uncertainty, and assessment of the probability of satisfying multivariate rules. These add practical value and theoretical insight in complex multivariate models.

References

Almeida, Alberto S.and Journel, A. G. (1994). Joint simulation of multiple variables with a markov-type coregionalization model. Mathematical Geology, 26(5):565–588.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). LAPACK Users' Guide. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition.

Axler, S. (2015). Linear Algebra Done Right. Springer International Publishing, 3 edition.

Babak, O. and Deutsch, C. V. (2008). Collocated cokriging based on merged secondary attributes. Mathematical Geosciences, 41(8):921.

Babak, O. and Deutsch, C. V. (2009). Improved spatial modeling by merging multiple secondary data for intrinsic collocated cokriging. Journal of Petroleum Science and Engineering, 69(1):93 – 99.

Bailey, T. C. and Krzanowski, W. J. (2012). An overview of approaches to the analysis and modelling of multivariate geostatistical data. Mathematical Geosciences, 44(4):381–393.

Barnett, R. M. (2015). Managing Complex Multivariate Relations in the Presence of Incomplete Spatial Data. PhD thesis, University of Alberta, Edmonton, Canada.

Barnett, R. M. and Deutsch, C. V. (2014). A compressed binary format for large geostatistical models. Centre for Computational Geostatistics annual report 16, Paper 413.

Barnett, R. M. and Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. Mathematical Geosciences, 47:791–817.

Barnett, R. M., Manchuk, J. G., and Deutsch, C. V. (2014). Projection pursuit multivariate transform. Mathematical Geosciences, 46(3):337–359.

Berger, A. J. (2015). Stratigraphy, geochemistry, isotopic signatures and VMS potential of late archean volcanic rocks from the southern Slave craton, Northwest Territories. Master's thesis, Carleton University, Ontario.

Bliss, C. I. (1934). The method of probits. Science, 79(2037):38–39.

Boisvert, J. B., Rossi, M. E., Ehrig, K., and Deutsch, C. V. (2013). Geometallurgical modeling at Olympic Dam mine, South Australia. Mathematical Geosciences, 45(8):901–925.

Borgman, L., Taheri, M., and Hagan, R. (1984). Three-dimensional, frequency-domain simulations of geological variables. In Verly, G., David, M., Journel, A. G., and Marechal, A., editors, Geostatistics for Natural Resources Characterization: Part 1, pages 517–541. Springer Netherlands, Dordrecht.

Boucher, A. and Dimitrakopoulos, R. (2012). Multivariate block-support simulation of the yandi iron ore deposit, western australia. Mathematical Geosciences, 44(4):449–468.

Brooker, P. I. (1985). Two-dimensional simulation by turning bands. Journal of the International Association for Mathematical Geology, 17(1):81–90.

Brooker, P. I. and Paul, C. J. (1982). Numerical simulation of a two dimensional orebody. In Proceedings of the fifth biennial conference of simulation society of Australia, pages 133–136.

Cabral Pinto, F. A., Barnett, R. M., and Deutsch, C. V. (2018). Postppmt for estimation and local uncertainty assessment. Centre for Computational Geostatistics annual report 20, Paper 126.

Cabral Pinto, F. A. and Deutsch, C. V. (2017a). Comparison of multivariate spatial distributions from different simulation algorithms. Centre for Computational Geostatistics annual report 19, Paper 112.

Cabral Pinto, F. A. and Deutsch, C. V. (2017b). Development of fft-based simulation. Centre for Computational Geostatistics annual report 19, Paper 113.

Cabral Pinto, F. A. and Deutsch, C. V. (2017c). Processing airborne geophysics for subsurface prediction. Centre for Computational Geostatistics annual report 19, Paper 123.

Cabral Pinto, F. A. and Deutsch, C. V. (2017d). Unconditional simulation of spherical variogram with moving window average. Centre for Computational Geostatistics annual report 19, Paper 403.

Cabral Pinto, F. A. and Deutsch, C. V. (2018). Turning bands simulation program - short note. Centre for Computational Geostatistics annual report 20, Paper 403.

Cabral Pinto, F. A. and Deutsch, C. V. (2019). Factor extraction and other software tools. Centre for Computational Geostatistics annual report 21, Paper 404.

Cabral Pinto, F. A., Manchuk, J., and Deutsch, C. V. (2019). Decomposition of data to underlying factors and simulation. Centre for Computational Geostatistics annual report 21, Paper 404.

Cant , R., Pel ez, M. J., and Urbano, A. M. (2015). Full rank cholesky factorization for rank deficient matrices. Applied Mathematics Letters, 40:17 – 22.

Chan, T. F. (1986). Alternative to the SVD: Rank revealing QR-factorizations. In Speiser, J. M., editor, Advanced Algorithms and Architectures for Signal Processing I, volume 0696, pages 31 – 38. International Society for Optics and Photonics, SPIE.

Chang, C., Fung, P. C. W., and Hung, Y. S. (2006). On a sparse component analysis approach to blind source separation. In Rosca, J., Erdogmus, D., Principe, J. C., and Haykin, S., editors, Independent Component Analysis and Blind Signal Separation, pages 765–772, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chang, C., Yau, S. F., Kwok, P., Chan, F. H., and Lam, F. (1999). Uncorrelated component analysis for blind source separation. Circuits, Systems and Signal Processing, 18(3):225–239.

Chatfield, C. (1980). The analysis of time series : an introduction. London : Chapman and Hall ; New York, N.Y. : Chapman and Hall in association with Methuen, 2nd ed edition.

Chiles, J.-P. (1977). Geostatistique des phenomenes non stationnaires (dans le plan). PhD thesis, University de Nancy I, Nancy, France.

Chiles, J.-P. and Delfiner, P. (2012). Geostatistics: Modeling Spatial Uncertainty. 2nd Ed. John Wiley & Sons.

Choi, S. and Cichocki, A. (1997). Adaptive blind separation of speech signals: Cocktail party problem. In International Conference on Speech Processing, pages 617–622.

Clifford, G. D. (2008). Blind source separation: Principal and independent component analysis. Chapter 15 for HST-582J/6.555J/16.456J Biomedical Signal and Image Processing 2005-2018.

Comon, P. (1994). Independent component analysis, a new concept? Signal processing, 36(3):287–314.

Comon, P. and Jutten, C. (2010). Handbook of Blind Source Separation: Independent component analysis and applications. Academic press.

Corput, J. (1935). Verteilungsfunktionen i, nederl. Akad. Wetensch. Proc. Ser. B, 38(38):813–821.

Cuba, M. A. and Silva, D. S. F. (2013). Spectralsim: A program for unconditional spectral simulation. Centre for Computational Geostatistics annual report 15, Paper 410.

David, M. (1977). Geostatistical ore reserve estimation. Elsevier, Amsterdan.

Davis, B., Hagan, R., and Borgman, L. (1981). A program for the finite fourier transform simulation of realizations from a one-dimensional random function with known covariance. Computers & Geosciences, 7(2):199 – 206.

Davis, B. M. and Greenes, K. A. (1983). Estimation using spatially distributed multivariate data: an example with coal quality. Journal of the International Association for Mathematical Geology, 15(2):287–300.

Day, S., Falck, H., Friske, P., Pronk, A., McCurdy, M., McNeil, R., Adcock, S., and Grenier, A. (2009). Regional stream sediment and water geochemical data, Mount Eduni area, northern Mackenzie Mountains, NT (NTS 106a and part of 106b). Geological survey of Canada, open file 6312/Northwest Territories geoscience office, NWT open report 2009-004. digital files., Geological Survey of Canada.

Day, S., Falck, H., McCurdy, M., and McNeil, R. (2012). Regional stream sediment and water geochemical data, Cranswick river area, Northwest Territories (parts of nts 106f and g). Geological survey of Canada, open file 6721/Northwest Territories geoscience office, NWT open report 2010-010. digital files., Geological Survey of Canada.

Day, S., Lariviere, J., Friske, P., Gochnauer, K., MacFarlane, K., McCurdy, M., and McNeil, R. (2005). National geochemical reconnaissance (NGR): Regional stream sediment and water geochemical data, Macmillan Pass - Sekwi mountain, Northwest Territories. GSC open report 2008-013. digital files., Geological Survey of Canada.

Desbarats, A. J. and Dimitrakopoulos, R. (2000). Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Mathematical Geology, 32(8):919–942.

Deutsch, C. V. and Journel, A. G. (1992). GSLIB: geostatistical software library and user's guide. Oxford University Press, New York.

Deutsch, C. V. and Journel, A. G. (1997). GSLIB: geostatistical software library and user's guide, 2nd Ed. Oxford University Press, New York.

Deutsch, J. L. (2015a). Experimental variogram tolerance parameters. In Deutsch, J. L., editor, Geostatistics Lessons. Retrieved from http://www.geostatisticslessons.com/lessons/variogramparameters.

Deutsch, J. L. (2015b). Multivariate Spatial Modeling of Metallurgical Rock Properties. PhD thesis, University of Alberta, Edmonton, Alberta.

Deutsch, J. L., Palmer, K., Deutsch, C. V., Szymanski, J., and Etsell, T. H. (2016). Spatial modeling of geometallurgical properties: Techniques and a case study. Natural Resources Research, 25(2):161–181.

Dhiflaoui, M., Funke, S., Kwappik, C., Mehlhorn, K., Seel, M., Schomer, E., Schulte, R., and Weber, D. (2003). Certifying and repairing solutions to large LPs how good are LP-solvers? In Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '03, page 255–256, USA. Society for Industrial and Applied Mathematics.

Dietrich, C. R. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. SIAM Journal on Scientific Computing, 18(4):1088–1107.

du Bray, E. A. (1995). Preliminary compilation of descriptive geoenvironmental mineral deposit models. Open-file report 95-831, U.S. Geological Survey.

Dubrule, O. (1983). Two methods with different objectives: Splines and kriging. Journal of the International Association for Mathematical Geology, 15(2):245–257.

Emery, X. (2004). Testing the correctness of the sequential algorithm for simulating Gaussian random fields. Stochastic Environmental Research and Risk Assessment, 18.

Emery, X. (2008). A turning bands program for conditional co-simulation of cross-correlated Gaussian random fields. Computers & Geosciences, 34(12):1850 – 1862.

Emery, X. and Arroyo, D. (2018). On a continuous spectral algorithm for simulating non-stationary Gaussian random fields. Stochastic environmental research and risk assessment, 32(4):905–919.

Emery, X. and Lantuejoul, C. (2006). Tbsim: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method. Computers & Geosciences, 32(10):1615 – 1628.

Emery, X. and Pelaez, M. (2011). Assessing the accuracy of sequential Gaussian simulation and cosimulation. Computational Geosciences, 15:673–689.

Ewing, I. (2016). Simulated deposits, real profits: stochastic mine planning trial at Newmont's Twin Creeks mine shows value generation potential. CIM Magazine, August 31, 2016.

Falck, H. and Day, S. (2008). Preliminary regional stream sediment and water geochemical data, backbone ranges area, west-central Northwest Territories. NWT open file 2008-013. digital files., Northwest Territories Geological Survey.

Falck, H., Day, S., Pierce, K., and Cairns, S. (2014). Geochemical, mineralogical and indicator mineral data for stream silt sediment, heavy mineral concentrates and waters, Cranswick river area northewest territories, (part of NTS 106f). NWT open file 2014-012. digital files., Northwest Territories Geological Survey.

Falck, H., Day, S., Pierce, K., Cairns, S., and Watson, D. (2015). Geochemical, mineralogical and indicator mineral data for stream silt sediment, heavy mineral concentrates and waters, Flat river area, Northwest Territories, (part of NTS 95e, 105h and 105i). NWT open file 2015-002. digital files., Northwest Territories Geological Survey.

Falck, H., Day, S., Pierce, K., Rentmeister, K., Ozyer, C., and Watson, D. (2012). A compilation of heavy mineral concentrates: results from stream sediment samples collected 2007-2010, Mackenzie Mountains, NWT. NWT open report 2012-001, Northwest Territories Geoscience Office.

FFTW (2017). Fastest fourier transform in the west. `http:////www.fftw.org/`. Accessed: 2017-08-10.

Fischer, B., Martel, E., and Falck, H. (2016). Geology of the Mactung tungsten skarn and area – review and 2016 field observations. NWT open file 2018-02, Northwest Territories Geological Survey.

Freulon, X. and de Fouquet, C. (1991). Remarques sur la pratique des bandes tournantes a trois dimensions. In Armstrong, M. and Dowd, P. A., editors, Cahiers de geostatistique, Fascicule 1, pages 101–117, Fontainebleau. Centre de Geostatistique, Ecole des Mines de Paris.

Friedman, J. H. (1987). Exploratory projection pursuit. Journal of the American statistical association, 82(397):249–266.

Gneiting, T. (1999a). The correlation bias for two-dimensional simulations by turning bands. Mathematical geology, 31(2):195–211.

Gneiting, T. (1999b). Isotropic correlation functions on d-dimensional balls. Advances in Applied Probability, 31:625–631.

Golub, G. H. and Loan, C. F. V. (2013). Matrix computations, volume 4. Johns Hopkins University Press.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. Numerische Mathematik, 14(5):403–420.

Gomez-Hernandez, J. J. and Journel, A. G. (1993). Joint Sequential Simulation of MultiGaussian Fields, pages 85–94. Springer Netherlands, Dordrecht.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. MIT press.

Goovaerts, P. (1992). Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. Journal of Soil Science, 43(4):597–619.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press.

Gotway, C. A. and Rutherford, B. M. (1994). Stochastic simultation for imaging spatial uncertainty: Comparison and evaluation of available algorithms. In Armstrong, M. and Dowd, P. A., editors, Geostatistical Simulations, pages 1–21, Dordrecht. Springer Netherlands.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. Commun. ACM, 7(12):701–702.

Handel, S. (1989). Listening: An Introduction to the Perception of Auditory Events. MIT Press.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of educational psychology, 24(6):417.

Hunger, L., Cosenza, B., Kimeswenger, S., and Fahringer, T. (2015). Spectral turning bands for efficient Gaussian random fields generation on GPUs and accelerators. Concurrency and Computation: Practice and Experience, 27.

Isaaks, E. H. (1990). The application of Monte Carlo methods to the analysis of spatially correlated data. PhD thesis, Stanford University.

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

Jebrak, M. and Marcoux, E. (2008). Geologie des ressources minerales. Quebec: Ministere des ressources naturelles et de la faune.

Journel, A. G. (1974). Geostatistics for Conditional Simulation of Ore Bodies.

Economic Geology, 69(5):673–687.

Journel, A. G. (1989). Fundamentals of geostatistics in five lessons. Short course in Geology. American Geophysical Union, 8.

Journel, A. G. and Huijbregts, C. J. (1978). Mining Geostatistics. Blackburn Press, New York.

Journel, A. G. and Kyriakidis, P. C. (2004). Evaluation of mineral reserves: a simulation approach. Oxford University Press.

Kawahata, K., Schumacher, P., and Criss, K. (2016). Large-scale mine production scheduling optimisation with mill blending constraints at Newmont's Twin Creeks operation. Mining Technology, 125(4):249–253.

Kofidis, E. (2016). Blind source separation: Fundamentals and recent advances (a tutorial overview presented at sbrt-2001).

Krige, D. G. (1951). A statisticial approach to some mine valuations and allied problems at the witwatersrand. Master's thesis, University of Witwatersrand.

Kyriakidis, P. C. (1999). Stochastic modeling of spatial temporal distributions. PhD thesis, Stanford University, Stanford, United States.

Lantuejoul, C. (1994). Non conditional simulation of stationary isotropic multi-Gaussian random functions. In Armstrong, M. and Dowd, P. A., editors, Geostatistical Simulations, pages 147–177, Dordrecht. Springer Netherlands.

Lantuejoul, C. (2002). Geostatistical simulation: models and algorithms. Springer, Berlin, Heidelberg.

Le Ravalec, M., Noetinger, B., and Hu, L. (2000). The fft moving average (fft-ma) generator: An efficient numerical method for generating and conditioning gaussian simulations. Mathematical Geology, 32:701–723.

Legchenko, A., Comte, J.-C., Ofterdinger, U., Vouillamoz, J.-M., Lawson, F. M. A., and Walsh, J. (2017). Joint use of singular value decomposition and Monte-Carlo simulation for estimating uncertainty in surface nmr inversion. Journal of Applied Geophysics, 144:28 – 36.

Leuangthong, O. (2003). Stepwise Conditional Transformation for Multivariate Geostatistical Simulation. PhD thesis, University of Alberta.

Leuangthong, O. and Deutsch, C. V. (2003). Stepwise conditional transformation for simulation of multiple variables. Mathematical Geology, 35(2):155–173.

Liang, M. and Marcotte, D. (2015). A class of non-stationary covariance functions with compact support. Stochastic Environmental Research and Risk Assessment.

Lin, Y.-P. (2002). Multivariate geostatistical methods to identify and map spatial variations of soil heavy metals. Environmental Geology, 42(1):1–10.

Liu, C. and Koike, K. (2007). Extending multivariate space-time geostatistics for environmental data analysis. Mathematical Geology, 39(3):289–305.

Luenberger, D. G. (1969). Gaussian Processes for Machine Learning. Wiley, New York.

Luster, G. R. (1985). Raw materials for porland cement: applications of conditional simulation of coregionalization. PhD thesis, Stanford University.

Ma, Z., Royer, J.-J., Wang, H., Y, W., and Zhang, T. (2014). Factorial kriging for multiscale modelling. Journal- South African Institute of Mining and Metallurgy, 114:651–657.

Manchuk, J. G. and Deutsch, C. V. (2011). A program for data transformations and kernel density estimation. Centre for Computational Geostatistics annual report 13, Paper 116.

Manchuk, J. G. and Deutsch, C. V. (2012). A flexible sequential Gaussian simulation program: USGSIM. Computers & Geosciences, 41:208 – 216.

Manchuk, J. G. and Deutsch, C. V. (2015). Latest sgsim program. Centre for Computational Geostatistics annual report 17, Paper 401.

Manchuk, J. G. and Deutsch, C. V. (2017). Global kriging and conditioning realizations in parallel. Centre for Computational Geostatistics annual report 19, Paper 114.

Manchuk, J. G., Qu, J., and Deutsch, C. V. (2019). Simulation of decorrelated factors in presence of secondary data. Spatial Statistics, 33:100385.

Mantoglou, A. (1987). Digital simulation of multivariate two and three-dimensional stochastic processes with a spectral turning bands method. Mathematical

Geology, 19(2):129–149.

Mantoglou, A. and Wilson, J. L. (1982). The turning bands method for simulation of random fields using line generation by a spectral method. Water Resources Research, 18(5):1379–1394.

Marcotte, D. (1996). Fast variogram computation with FFT. Computers & Geosciences, 22(10):1175 – 1186.

Marcotte, D. (2012). Revisiting the Linear Model of Coregionalization, volume 17, pages 67–78.

Marcotte, D. (2016). Spatial turning bands simulation of anisotropic non-linear models of coregionalization with symmetric cross-covariances. Computers & Geosciences, 89:232 – 238.

Matern, B. (1986). Spatial Variation. Springer-Verlag, New York.

Matheron, G. (1962). Traite de geostatistique appliquee, volume 1 of Memoires du Bureau de Recherches Geologiques et Mini res, No. 14. Editions Technip, Paris.

Matheron, G. (1971). The theory of regionalized variables and its applications, volume 5. Ecole national superieure des mines.

Matheron, G. (1973). The intrinsic random functions and their applications. Advances in Applied Probability, 5(3):439–468.

Matheron, G. (1979). Recherche de simplification dans un probleme de cokrigeage,(rapport n-628, cg, Ecole des Mines de Paris).

Matheron, G. (1982). Pour une analyse krigeante des donnees. research note n-732 centre de geostatistique fontainebleau.

McClenaghan, M. B. and Peter, J. M. (2013). Till geochemical signatures of volcanogenic massive sulphide deposits in glaciated terrain: a summary of canadian examples. Open file 7354, Geological Survey of Canada.

McCurdy, M., Day, S., Friske, P., McNeil, R., and Hornbrook, E. (2009a). Regional stream sediment and water geochemical data, Frances Lake area, southeastern Yukon (NTS 105h). GSC open file 6043. digital files., Geological Survey of Canada.

McCurdy, M., Friske, P., McNeil, R., Day, S., and Goodfellow, W. (2009b). Regional stream sediment and water geochemical data, eastern Yukon and western Northwest Territories (NTS 105i). GSC open file 6271. digital files., Geological Survey of Canada.

McCurdy, M., McNeil, R., Friske, P., Day, S., and Wilson, R. (2007). Stream sediment geochemistry in the proposed extension to the nahanni park reserve. In Mineral and energy resource assessment of the Greater Nahanni Ecosystem under consideration for the expansion of the Nahanni National Park Reserve, Northwest Territories, pages 75–98. GSC Geological Survey of Canada, Open File 5344. Digital files.

Mejia, J. M. and Rodriguez-Iturbe, I. (1974). On the synthesis of random field sampling from the spectrum: An application to the generation of hydrologic spatial processes. Water Resources Research, 10(4):705–711.

Minnitt, R. and Deutsch, C. (2014). Cokriging for optimal mineral resource estimates in mining operations. Journal of the Southern African Institute of Mining and Metallurgy, 114:189 – 189.

Minnitt, R. and Esbensen, K. (2017). Pierre Gy's development of the theory of sampling: a retrospective summary with a didactic tutorial on quantitative sampling of one-dimensional lots. TOS forum, page 7.

Mol, M. (2018). Van der corput sequence in python. `https://rosettacode.org/wiki/Van_der_Corput_sequence#Python`. Accessed: 2018-05-23.

Montiel, L. and Dimitrakopoulos, R. (2018). Simultaneous stochastic optimization of production scheduling at Twin Creeks mining complex, Nevada. Mining Engineering, 70:48–56.

Myers, D. E. (1982). Matrix formulation of co-kriging. Journal of the International Association for Mathematical Geology, 14(3):249–257.

Naik, G. and Wang, W. (2014). Blind Source Separation: Advances in Theory, Algorithms and Applications. Springer-Verlag Berlin Heidelberg.

Neufeld, C., Deutsch, C. V., and Lyall, G. (2008). Simulation of grade control, stockpiling and stacking for compliance testing of blending strategies. In 8th

International Geostaitiscs Congress. Santiago, Chile.

Oliver, D. S. (1995). Moving averages for Gaussian simulation in two and three dimensions. Mathematical Geology, 27(8):939–960.

Oliver, D. S., Reynolds, A. C., and Liu, N. (2008). Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press.

Ootes, L., Gleeson, S., Turner, E., Rasmussen, K., Gordey, S., Falck, H., Martel, E., and Pierce, K. (2013). Metallogenic evolution of the Mackenzie and eastern Selwyn mountains of Canada's northern cordillera, Northwest Territories: A compilation and review. Geoscience Canada, 40(1).

Ozyer, C. (2010). Sude Niline tueyeta (Ramparts river and wetlands) candidate protected area phase ii non-renewable resource assessment – minerals, Northwest Territories, Canada. NWT open file 2010-07. digital files., Northwest Territories Geological Survey.

Ozyer, C. (2012). Shuhtagotine nene candidate protected area phase ii non-renewable resource assessment – minerals, Northwest Territories, Canada. NWT open file 2012-01. digital files., Northwest Territories Geological Survey.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. Environmetrics, 17(5):483–506.

Paravarzar, S., Emery, X., and Madani, N. (2015). Comparing sequential Gaussian and turning bands algorithms for cosimulating grades in multi-element deposits. Comptes Rendus Geoscience, 347(2):84 – 93.

Pardo-Iguzquiza, E. and Chica-Olmo, M. (1993). The fourier integral method: An efficient spectral method for simulation of random fields. Mathematical Geology, 25(2):177–217.

Patil, G. and Rao, C. (1994). Multivariate Environmental Statistics, Volume 6. North Holland.

Pitard, F. (2019). Theory of Sampling and Sampling Practice, Third Edition. New York: Chapman and Hall/CRC.

Pyrcz, M. J. and Deutsch, C. V. (2014). Geostatistical Reservoir Modeling, volume 2. Oxford university press.

Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. The MIT Press.

Rendu, J.-M. (2008). An introduction to cut-off grade estimation. SME: Society for Mining, Metallurgy and Exploration.

Rivoirard, J. (2001). Which models for collocated cokriging? Mathematical Geology, 33(2):117–131.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. The annals of mathematical statistics, pages 470–472.

Rossi, M. and Deutsch, C. V. (2014). Mineral Resource Estimation. Springer Netherlands.

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434):473–489.

Sadek, R. A. (2012). Svd based image processing applications: State of the art, contributions and research challenges. International Journal of Advanced Computer Science and Applications, 3(7).

Safikhani, M., Asghari, O., and Emery, X. (2016). Assessing the accuracy of sequential Gaussian simulation through statistical testing. Stochastic Environmental Research and Risk Assessment, 31.

Schmidt, M. (2009). Linearly constrained bayesian matrix factorization for blind source separation. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, Advances in Neural Information Processing Systems 22, pages 1624–1632. Curran Associates, Inc.

Shinozuka, M. (1971). Simulation of multivariate and multidimensional random processes. The Journal of the Acoustical Society of America, 49(1B):357–368.

Shinozuka, M. and Jan, C. (1972). Digital simulation of random processes and its applications. Journal of Sound and Vibration, 25(1):111 – 128.

Silva, D. and Deutsch, C. (2015). Transformation for multivariate modeling using Gaussian mixtures with exhaustive secondary data.

Silva, D. S. and Deutsch, C. V. (2018). Multivariate data imputation using Gaussian mixture models. Spatial Statistics, 27:74 – 90.

Smith, S. W. (1998). The Scientist and Engineer's Guide to Digital Signal Processing.

Strang, G. (2016). Introduction to Linear Algebra, Fifth Edition. Wellesley-Cambridge Press.

Switzer, P. (1984). Min/max autocorrelation factors for multivariate spatial imagery. 16.

Teck (2017). Ni 43-101 technical report, Red Dog mine, Alaska, USA. report prepared for Teck Resources Limited, 2017. https://www.miningdataonline.com/reports/Red%20Dog%20Mine_TR12312016.pdf [Accessed: 2018-07].

Tompson, A. F. B., Ababou, R., and Gelhar, L. W. (1989). Implementation of the three-dimensional turning bands random field generator. Water Resources Research, 25(10):2227–2243.

Tran, T. T. (1994). Improving variogram reproduction on dense simulation grids. Computers & Geosciences, 20(7):1161 – 1168.

Tzeng, J. (2013). Split-and-combine singular value decomposition for large-scale matrix. Journal of Applied Mathematics, 2013.

Verly, G. (1983). The multiGaussian approach and its applications to the estimation of local reserves. Journal of the International Association for Mathematical Geology, 15(2):259–286.

Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer-Verlag Berlin Heidelberg, 3 edition.

Watkins, D. S. (2002). Fundamentals of Matrix Computations, Second Edition. John Wiley & Sons, Inc.

Webster, R. and Oliver, M. A. (2007a). Geostatistics for Environmental Scientists, 2nd Edition. John Wiley and Sons, Ltd.

Webster, R. and Oliver, M. A. (2007b). Spatial Variation. John Wiley & Sons, Ltd.

Yang, D., Ma, Z., and Buja, A. (2014). A sparse singular value decomposition method for high-dimensional data. Journal of Computational and Graphical

Statistics, 23.

Yao, T. (1998a). Automatic covariance modeling and conditional spectral simulation with Fast Fourier Transform. PhD thesis, Stanford University.

Yao, T. (1998b). Conditional spectral simulation with phase identification. Mathematical Geology, 30(3):285–308.

Yao, T. (1998c). specsim: A fortran-77 program for conditional spectral simulation in 3d. Computers & Geosciences, 24(10):911 – 921.

Yu, X., Hu, D., and Xu, J. (2014). Blind source separation: theory and applications. John Wiley & Sons.

Zhang, M. (2009). Blind source separation using generalized singular value decomposition. In 2009 First International Conference on Information Science and Engineering, pages 509–511. IEEE.

Appendix A

Software

The software developed for the methodologies discussed in this thesis are presented in this appendix. Some of the software use pre-compiled subroutines and modules written in FORTRAN for the original GSLIB codes (Deutsch and Journel, 1992). Some programs, such as the PostPPMT, make use of subroutines recently developed for research purposes in other GSLIB programs, such as PPMT (Barnett et al., 2014). Software for simulation are available within GSLIB and published under Manchuk and Deutsch (2015) (SGS), Cuba and Silva (2013)(Spectral), and Deutsch and Journel (1992) (TB). New versions of existing GSLIB programs are developed to accommodate modern computer practices and research purposes, such as SpectralSim (Cabral Pinto and Deutsch, 2017b) and TB3D (Cabral Pinto and Deutsch, 2018). The parameter files of new software developed and used in this thesis are presented.

A.1 PostPPMT

The PostPPMT program (Cabral Pinto et al., 2018) provides the essential functionality described in Chapter 3. The parameter file is given below.

```
    Parameters for PostPPMT
    ***********************
START OF PARAMETERS:
PostPPMT.out      -file for output
ppmt.trn          -file with input transformation table
kt3dn_1.gsb       -file with kriged NS mean and variance (Var 1)
1  2              -  columns with NS mean and variance
kt3dn_2.dat       -file with kriged NS mean and variance (Var 2)
1  2              -  columns with NS mean and variance
kt3dn_2.dat       -file with kriged NS mean and variance (Var 3)
1  2              -  columns with NS mean and variance
-10   1.0e21      -  trimming limits
```

```
13 1000 69069         -sampling realizations and seed
14 PROBABILITY TO BE INSIDE A TOLERANCE OF THE MEAN
15 3                  -# of inputs to check
16 1 0.15             -  var, probability
17 1 0.25             -  var, probability
18 2 0.15             -  var, probability
19 QUANTILES
20 5                  -# of inputs to check
21 1 0.5              -  var, percentile
22 1 0.1              -  var, percentile
23 1 0.9              -  var, percentile
24 2 0.5              -  var, percentile
25 3 0.5              -  var, percentile
26 SINGLE MULTIVARIATE RULE
27 3                  -# of single var. conditions
28 1 1 0.9            -  var, rule, threshold
29 2 0 1.6            -  var, rule, threshold
30 3 0 1.3            -  var, rule, threshold
31 RATIO MULTIVARIATE RULE
32 2                  -# of ratio conditions
33 1 2 1 1.1          -  var 1, var 2, rule, threshold
34 3 2 0 0.8          -  var 1, var 2, rule, threshold
```

The output file (line 4) is a standard GSLIB-format file. The transformation table input is from the PPMT program (line 5). Note that there is no standard approach to save the transformation – legacy transformation files or files from other software will not likely work. This program uses the transformation table format compiled in PPMT. The transformation table contains the number of variables. The user must provide the files and columns with the kriged mean and variance for each factor, lines 6 to 11. The input files are given in compressed gsb format (Barnett and Deutsch, 2014) or the standard GSLIB-format file, the file extension specifies the format. The back transformation will proceed for all entries in the input files – until the end of file is reached in one or more input files. The trimming limits are set in line 12. Trimming limits are applied across variables. The number of samples to draw from the conditional distributions and the random number seed are set in line 13. A 1000 should be enough in most cases. A larger number could be use to increase discretization of the tails of the distribution. The random number seed permits reproducibility of the results.

The calculation in the four blocks starting with PROB, QUAN, SING and RATI (lines 14, 19, 26, and 31) are optional. Calculations are skipped if the numbers in lines 15, 20, 27, and 32 are set to zero. The PROB block calculates the probability to be within a specified tolerance of the mean for specified variables. The variable numbers and probabilities are set in lines 16 to 18. The QUAN block calculates the quantiles for given variables, lines 21 to 25. A joint multivariate rule is defined by multiple univariate criteria applied to any independent set of variables or/and their ratios. The probability of all criteria being jointly satisfied is reported to the output file. Multiple runs could be made for multiple rules or modifications could be made to consider multiple sets of rules. Single multivariate rules are defined in lines 28 to 30, ratio rules are defined in lines 33 and 34.

The `PostPPMT` program always outputs the conditional mean and standard deviation (default). The number and order of the variables are taken from the transformation table. One file with kriged mean and variance is required for each variable in the `PPMT` transformation table. There is no need to specify the file size, the size is read from the input files, however, all files with the kriged mean and variance are expected to be of the same size. The multivariate rules are set as 0 = below threshold and 1 = above threshold. Ratios are calculated as Var_1/Var_2. The programs accepts GSB functionality for any input file, but the output file is in standard Geo-EAS and ASCII format.

A.2 `MW_SIM`

The moving window simulation `MW_SIM` program (Cabral Pinto and Deutsch, 2017d) implements the clever update of the indices (section 2.4) for a fast simulation approach of spherical covariance structures. The code is simple but it offers a fast and robust implementation of the moving window algorithm. The program was developed and tested during studies done on Airborne survey (Cabral Pinto and Deutsch, 2017c).

A classical moving windows algorithm works by visiting each node of the grid, centring the window at that node, searching for all other nodes inside the window, and averaging all node data values (including the node at the location being simulated). The average step per si is already inefficient, because it requires two operations, the summation of all values and the division by the number of data in the window. This process repeats for every node in the cell grid. Another issue is in the border of the grid cell, that is, the window centred at nodes at the border of the grid cell contains less data than nodes located in the inner part of the grid. This usually introduces artifacts because different locations of the grid are simulated with different number of data. All these setbacks are handled in MW_SIM. The clever update of the indices saves memory and speed up the code by reducing the number of data searched and by performing one single average in the end. The parameter file is given below.

```
 1         Parameters for MW SIM
 2         *********************
 3  START OF PARAMETERS:
 4  mw_sim.out             - output file
 5  128   0.5   1.0        - grid: nx,xmn,xsiz
 6  128   0.5   1.0        -       ny,ymn,ysiz
 7  1     0.5   1.0        -       nz,zmn,zsiz
 8  69069                  - Random number seed
 9  0.0                    - Prior mean
10  1.0                    - Contribution factor
11  16  16  1              - Window radius (major, semi, minor)
12  0  0   0               - Anisotropy angles (major, semi, minor)
13  NOTES:
14  Use .gsb for binary output
```

The output file, line 4, is written in standard Geo-EAS format with no data compression (ASCII format) or in GSB binary format. The simulation grid definition is defined in lines 5 to 7. A subroutine calculates and pads the grid based on the window radius and anisotropy angles. The random number seed is set up in line 8, it is used to populate the padded grid with random standard Gaussian values. The prior mean and contribution factors are given in lines 9 and 10. The window radius (not the diameter) and the anisotropy angles are defined in lines 11 and 12. The window radius in any direction must be greater than zero and set to 1 if a dimension is not considered.

A.3 IFS programs

The IFS methodology is implemented in the GSLIB programs LMC_IMP and LMC_COMP (Cabral Pinto and Deutsch, 2019; Cabral Pinto et al., 2019). Both programs are compiled with INTEL compiler and multithread libraries. GSLIB subroutines are used for most of the calculations and data processing. The subroutines for matrix operations and SVD are imported from the LAPACK library (Anderson et al., 1999). The subroutines are collected and distributed with the original FORTRAN files to facilitate compilation.

LMC_IMP

The LMC_IMP program implements all steps of the IFS methodology for factor extraction and imputation. The data inputs are the normal scores of the variables, the data locations, and the LMC parameters. The program outputs the minimum norm solution (dual cokriging) and the imputed Y factors. The program checks if the LMC factors are adding up to one. The program writes in the terminal the LMC a and a^2 coefficients in addition to the correlation matrix between the variables calculated from the LMC coefficients. The parameter file is given below.

```
    Parameters for LMC_IMP
    **********************
START OF PARAMETERS:
nscore.out               -file with data
1  2  0                  -   columns for X, Y, Z
-10    10                -   trimming limits
2                        -   number of variables
3 4                      -   columns for variables
69069                    -random number seed
1                        -number of realizations to generate
lmc_ys.out               -output file with independent factors
min_norm.out             -output file with minimum norm solution
4                        -number of LMC factors
0.875 0.000 0.000 0.484      -LMC "a" coeff. matrix:1-1, 1-2,...
0.451 0.624 0.587 0.249      -                      2-1, 2-2,...
2 -45.0 0.0 0.0 250 100 1.0   -LMC parameters matrix (*): str. 1
2 -45.0 0.0 0.0 350 250 1.0   -                           str. 2
1 -45.0 0.0 0.0 800 400 1.0   -                           str. 3
```

```
19 1 -45.0 0.0 0.0 1800 500 1.0  -                            str. 4
20
21 * LMC parameters matrix format:
22 structure type, ang1, ang2, ang3, a\_hmax, a\_hmin, a\_hvert
```

The input file with the NS transforms of the variables is given in line 4. The trimming limits, in line 6, are applied to all locations with no variable measurements. An array with trimmed locations is stored for later use. Locations with at least one variable available are kept. The number of variables and columns in the input file are set in lines 7 and 8. The seed number and number of realizations are set in lines 9 and 10. The output file with the imputed Y factors is set in line 11. The output file contains one column for factor, with the realizations written row-wise. To preserve the input file size, the output file will contain the same number of rows of the input file. Trimmed locations are written in the file as missing -999. The output file with dual cokriging estimates is set in line 12. The cokriging solution does not depend on the seed number neither on the number of realizations. There is an unique solution for the same data configuration and LMC coefficients. The number of LMC factors is set in line 13. The LMC coefficients matrix (lines 14 and 15) depend on the number of variables and factors. The matrix follows the standard notation of the theory: $a_{k,i}$ are the coefficients explaining the contribution of the i^{th} factor to the k^{th} variable. The LMC parameters are defined in lines 16 to 19; for each factor the user must enter the structure type, and the three angles and ranges defining the anisotropy.

LMC_COMP

The LMC_COMP program computes the Z variable values given the simulated Y factors and the LMC coefficients matrix. The program accepts data locations for imputed factors or gridded data for simulated factors. In both cases the number of data per realization must be given and set by the variables nx, ny, and nz. If not gridded data, then nx must be set to the number of locations in the data, and $ny = nz = 1$. These variables are used to allocate the array that is used to read-in the data. The LMC coefficients matrix used in LMC_COMP and LMC_IMP must be the same. The parameter file is given below.

```
1        Parameters for LMC_COMP
2        ***********************
3  START OF PARAMETERS:
4  computed_z.out           -output file with computed variables
5  2                        -number of variables
6  4                        -number of LMC factors
7  0.875 0.000 0.000 0.484  -LMC "a" coeff. matrix:1-1, 1-2,...
8  0.451 0.624 0.587 0.249  -                      2-1, 2-2,...
9  -10    10              -trimming limits
10 1                      -number of realizations
11 248 305 0              -nx, ny, nz
12 4                      -number of files with simulated Y factors
13 usgsim_y1.out          - file #1
14 1 1                    -   number of factors and columns
15 usgsim_y2.out          - file #2
16 1 1                    -   number of factors and columns
17 mwsim_y3.out           - file #3
18 1 1                    -   number of factors and columns
19 spectralsim_y4.out     - file #4
20 1 1                    -   number of factors and columns
```

The output file with the computed Z variables is set in line 4. The number of variables and LMC factors are defined in lines 5 and 6. The LMC coefficients matrix (lines 7 and 8) depends on the number of variables and factors. Lines 5 to 8 must be set to match the same parameters defined in the LMC_IMP program. The trimming limits in line 10 are used when reading the files with the simulated factors (lines 13 to 20). The variables are only computed at locations where all factors are available. Trimmed locations are written out as missing -999 values. The number of realizations is set in line 10. The model size, either for data locations or gridded data is set in line 11. The LMC factors can be simulated with different algorithms and read in the program from different files. The number of files with simulated factors is set in line 14. The number of factors and columns must be defined for each file (lines 13 to 20). The program expects the total number of factors in all files to match the number of factors set in line 6.