Enhanced Geologic Modeling of Multiple Categorical Variables

by

Diogo Sousa Figueiredo Silva

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

Doctor of Philosophy

in

Mining Engineering

Department of Civil and Environmental Engineering

University of Alberta

Abstract

Widely spaced data sets from drilling are used in the mining and petroleum industries to model subsurface resources. These data sets have high associated economic and environmental costs. Maximizing the use of information contained in data while minimizing the amount of data required to achieve acceptable understanding of risk and uncertainty is critical in this context. These data sets usually consists of spatially correlated categorical and continuous variables that are modeled using geostatistics, which is a branch of statistics particularly applicable to spatiotemporal variables.

Categorical variables are usually modeled first and utilized to define stationary domains for the modeling of the continuous variables. As a result, the categorical variables are key for the accurate modeling of attributes such as ore concentrations, metallurgical recoveries and structural stability. Each one of these variables are affected in different ways by different combinations of categorical variables. Limitations of existing techniques force the merging of multiple categorical variables into a single variable causing the loss of information and deteriorating the quality of predictions.

Amongst the existing techniques for categorical modeling, truncated pluri-Gaussian simulation (TPGS) is one of the most flexible. The utilization of underlying Gaussian latent variables for the simulation of categories allows for the use of the well established techniques for the simulation of Gaussian random functions (GRFs). The truncation rules utilized to map the continuous variables to the categorical variable allow the introduction of geological constraints. These geological constraints assist the generation of models that are more realistic and accurate. The practical application of TPGS is often limited to the utilization of no more than three Gaussian latent variables. This is mostly attributed to the cur-

rent practice on the definition of truncation rules using truncation masks. This limitation is addressed in this thesis with the development the hierarchical truncated pluri-Gaussian (HTPG). HTPG utilizes a tree structure for the truncation of the Gaussian latent variables facilitating its definition based on geological expertise. The developed methodology allows for the utilization of an arbitrary number of latent variables to model an arbitrary number of categories. As a result, the developed method better explores the potential of the truncated Gaussian method.

The HTPG framework developed in this thesis is extended to the modeling of multiple categorical variables. The extension is achieved by allowing correlation between the latent Gaussian variables defining each categorical variable. This improves the utilization of the information available by preventing the merging of multiple categorical variables into a single set. It is demonstrated that the developed technique leads to significant improvement of the prediction of attributes that depends on the multivariate relationship between the categorical variables.

The research work of this thesis also led to significant contributions in other aspects of the truncated Gaussian methods, such as the numerical derivation of the latent variables variograms and the imputation of the latent variables. Significant contributions are also made on the multiple data imputation for application with multivariate transformations in advanced geostatistical methods.

Preface

This dissertation is an original work by Diogo S. F. Silva and is submitted for degree of Doctor of Philosophy in Mining Engineering at the University of Alberta. All material covered was conducted under the supervision of Professor Clayton V. Deutsch in the Department of Civil and Environmental Engineering.

Part of the research documented in this dissertation led to the publication of two articles. Most of the published material relates to advancements described in chapter 5 and chapter chapter 6. The first article provides an overview of the techniques for the imputation of latent variables for use in truncated Gaussian methods. A framework for improved convergence is also proposed in this article as well as a novel approach based on simulated annealing. The impact of the imputation framework on reserve/resource classification is also investigated. The first article is published as D.S. Silva and C.V. Deutsch, "Multiple imputation framework for data assignment in truncated pluri-Gaussian simulation". *Stochastic Environmental Research and Risk Assessment*, volume 31, issue 9, 2251–2263. The second article outlines the methodology developed for the imputation of missing data for use with multivariate transformations for geostatistical applications. The article is published as "Multivariate data imputation using Gaussian mixture models". *Spatial Statistics*.

To my wife Marina

Acknowledgments

I would like to thank my supervisor Dr. Clayton Deutsch for all his support, guidance and invaluable advice. I would also like to thank all the fellow colleagues from the Centre for Computational Geostatistics (CCG) for their assistance and friendship. Financial support from sponsors of CCG is gratefully appreciated.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Description
AETO	acid east type ore
AWTO	acid west type ore
BETO	basic east type ore
BLUE	best linear unbiased estimation
BU	Bayesian updating
BWTO	basic west type ore
CCK	collocated cokriging
CDF	cumulative density function
DCM	dynamic contact matrix
EM	expectation maximization
GMM	Gaussian mixture model
GRF	Gaussian random function
HG	High Grade
HTPG	hierarchical truncated pluri-Gaussian
ICCK	intrinsic collocated cokriging
IK	indicator kriging
IMC	intrinsic model of coregionalization
KDE	kernel density estimation
KS	Kolmogorov-Smirnov
LG	Low Grade
LMC	linear model of coregionalization
LU	Land Use variable (in Swiss Jura data set)
LVM	locally varying mean
MAF	minimum/maximum auto-correlation factors
MAPE	mean absolute percent error
MAPS	maximum a posteriori selection
MAR	missing at random

Abbreviation	Description
MCS	Monte-Carlo simulation
MDS	multidimensional scaling
MI	multiple imputation
MPS	multiple point statistics
MST	minimum spanning tree
MVHTPG	multivariate hierarchical truncated pluri-Gaussian
NPKDE	non-parametric kernel density estimation
PCA	principal component analysis
PDF	probability density function
PPDE	projection pursuit density estimation
PPMT	projection pursuit multivariate transformation
RF	random function
RMSE	root mean squared error
RSSE	sum of squared error
RT	Rock Type variable
RV	random variable
SCT	stepwise conditional transform
SEDEX	sedimentary exhalative
SGS	sequential Gaussian simulation
SIS	sequential indicator simulation
SK	simple kriging
SPGMM	semi-parametric Gaussian mixture model
TGS	truncated Gaussian simulation
TPGS	truncated pluri-Gaussian simulation

Chapter 1

Introduction

This chapter provides an overview of the problem setting that motivates the research subject of this thesis as well a description of the contributions and the thesis outline. Section 1.1 gives a brief description of the current state of categorical and continuous variable modeling. The motivations and contributions of this thesis are provided in Section 1.2. A thesis statement is given in Section 1.3. The outline of this thesis with a brief summary of each chapter is also provided in Section 1.3.

1.1 Problem Setting

Modeling spatially correlated variables have been focus of geostatistics since its development in the 1960's by Matheron (1962). Even though geostatistics was developed in a mining context, its application has spread to many other fields of research including petroleum, hydrology, forestry, geography, oceanography, agriculture and environmental sciences. Geostatistical methods are particularly applicable to the modeling of spatiotemporal data (Chilès and Delfiner, 1999).

The economic and environmental cost of acquiring samples in the mining and petroleum industries leads to sparse sampling that usually only covers a billionth (Chilès and Delfiner, 1999) of the domain of interest. These scarce data sets are used to model subsurface resources and to support decisions. Maximizing the use of information contained in data while minimizing the amount of data required to achieve acceptable understanding of risk and uncertainty is critical in these circumstances.

Geostatistics uses features and properties observed in the data (e.g. first and second order statistics) to generate equally probable numerical models through stochastic simulation (Goovaerts, 1997). These models reproduce the data at their locations and are built to reproduce observed spatial features, allowing accurate assessment of uncertainty. Most recent developments in geostatisitics are focused on improving the usage of available information and reproduction of observed features.

Geological variables are often divided into two major types: categorical and continuous. Categorical variables are defined as labels or names such as rock types or facies, while continuous variables represent quantifiable attributes such as metal concentration or porosity. The current modeling paradigm is hierarchical with respect to these types. Categorical variables are modeled first followed by the modeling of continuous variables within each category. The continuous variables are assumed to be independent between different categories and each category is often assumed to represent a stationary domain (Pyrcz and Deutsch, 2014; Rossi and Deutsch, 2014).

1.1.1 Modeling of Categorical Variables

The stationary domains defined by categorical variables are the main control on the distribution of continuous variables (Rossi and Deutsch, 2014) and the greatest source of risk (Snowden et al., 2002) in mining projects. There are several methodologies available for the modeling of categorical variables. These techniques can be divided into stochastic and deterministic. The deterministic approaches cannot be used for the characterization of risk and uncertainty as only one scenario is considered. The stochastic methods have different characteristics and their application depends on the modeling goals and geological setting.

Stochastic methods can be further divided into object based and cell based approaches. The object based frameworks are focused on the reproduction of morphological shapes such as meandering channels in fluvial environments. Conditioning the object based models to data observation is difficult and only pertinent when these original shapes are preserved in the subsurface. The original morphology is usually disturbed by geological events and the structures that are left may not resemble the original objects. Cell based methods, on the other hand, are easily conditioned to the observed data. The degree of geological complexity they can represent is variable.

Among cell based methods, multiple point statistics (MPS) is the methodology with the greatest capability of reproducing complex non-linear features. The technique relies on the utilization of reference models often referred to as training images. The resulting models rely heavily on these training images. The main limitation with the use of training images is the difficulty in handling non-stationarity and the limitation of relative simple features embedded in the training image (Pyrcz and Deutsch, 2014). Sequential indicator simulation (SIS) is at the other end of the spectrum of cell based methods. The technique utilizes two

point statistics and provides no means of introducing explicit geological controls. The SIS technique also lacks a fully specified random function that leads to difficulty reproducing reference statistics (Emery, 2004).

TPGS is a cell based method for simulation of categorical variables. The simulation of the categorical variable is undertaken utilizing underlying Gaussian latent variables. These latent variables can be modeled with one of the many well established methods for simulation of GRFs. The mapping between the categorical and continuous variables is realized by truncation rules. These truncation rules allow for the introduction of modeling constraints based on geological contacts. TPGS is a powerful method for the simulation of categorical variables, however, its application have been mostly constrained to no more than two or three Gaussian latent variables. The reasons for this limitation is explored in this thesis and an alternative methodology for the application of TPGS is developed.

1.1.2 Modeling of Continuous Variables

The multivariate modeling of continuous variables in conventional geostatistics is almost always performed under an assumption of multivariate Gaussianity and using the linear model of coregionalization (LMC). Variables are assumed to be multivariate Gaussian after univariate transformation, which is not correct for most geological variables (Barnett, 2015). An incorrect assumption of Gaussianity may lead to poor reproduction of some multivariate relationships. The multivariate relationship between the economic variables and contaminants are important for the prediction of recovery in metallurgical processing.

Considerable advances in multivariate modeling have been made in recent years. Modern workflows based on multivariate transformations such as principal component analysis (PCA) (Davis and Greenes, 1983; Hotelling, 1933), minimum/maximum auto-correlation factors (MAF) (Desbarats and Dimitrakopoulos, 2000; Switzer and Green, 1984), stepwise conditional transform (SCT) (Leuangthong and Deutsch, 2003; Rosenblatt, 1952), projection pursuit multivariate transformation (PPMT) (Barnett et al., 2014; Friedman, 1987), among others have become widely used to overcome the limitations of the classical methods.

1.1.3 Modeling of Multiple Categorical Variables

As with the continuous variables, multiple categorical variables are also available in most cases. The combination of these categorical variables governs many aspects of mining

projects from ore content to metallurgical recovery. Multiple categorical variables, however, did not receive the same attention as the continuous variables and little research on multivariate modeling of categorical variables is available. This problem is addressed in this thesis; a methodology for the multivariate modeling of categorical variable is developed.

1.2 Thesis Statement and Research Contributions

The hierarchical approach for TPGS allows the utilization of higher dimensions and simplifies the geological interpretation of the truncation rule resulting in the enhanced ability to represent complex geological settings. The multivariate modeling of categorical variables mitigates the information loss from category merging and permits better prediction of metallurgical performance.

The key contributions of this thesis are the development of: (1) a HTPG approach that improves the modeling of categorical variables; and (2) a method for the multivariate modeling of categorical variables. Other contributions are the development of method and tools for: (1) continuous variables variogram optimization for truncated Gaussian methods; (2) stable data imputation for HTPG based on Gibbs sampler; and (3) the multivariate data imputation of continuous variables based on Gaussian mixture models (GMMs) for use with multivariate transformation methods.

1.2.1 Hierarchical Approach to Truncated pluri-Gaussian Simulation

The concept of truncating continuous variables to model categorical variables is very flexible. The current application of TPGS for categorical modeling does not explore the full capability of this concept. More often then not, the application of the TPGS is restricted to the utilization of two or three continuous variables, regardless of the number of categories being modeled. There is no theoretical restriction on the number of Gaussian variables that can to be used. In practice, the number of continuous variables can go up to the number of categories being modeled minus one. This provides the degrees of freedom to model any kind of categorical variable.

There are five essential steps for the application of truncated Gaussian methods: (1) definition of a truncation rule; (2) mapping of spatial continuity from categorical to continuous

space; (3) imputation of continuous data subject to categorical observations; (4) simulation of the continuous variables at modeling nodes; and (5) truncation of the simulated models to generate categorical realizations. The first step is critical and has an effect on every other step.

The current methodology for TPGS have been developed around the concept of truncation masks. These masks are graphical representations of the mapping between the continuous and the categorical space (Figures 1.1 and 1.2). The concept of using truncation masks is limiting as the geological interpretation degrades when higher number of categories and/or latent variables are considered. The conventional graphical representation cannot be easily used after three Gaussian variables.

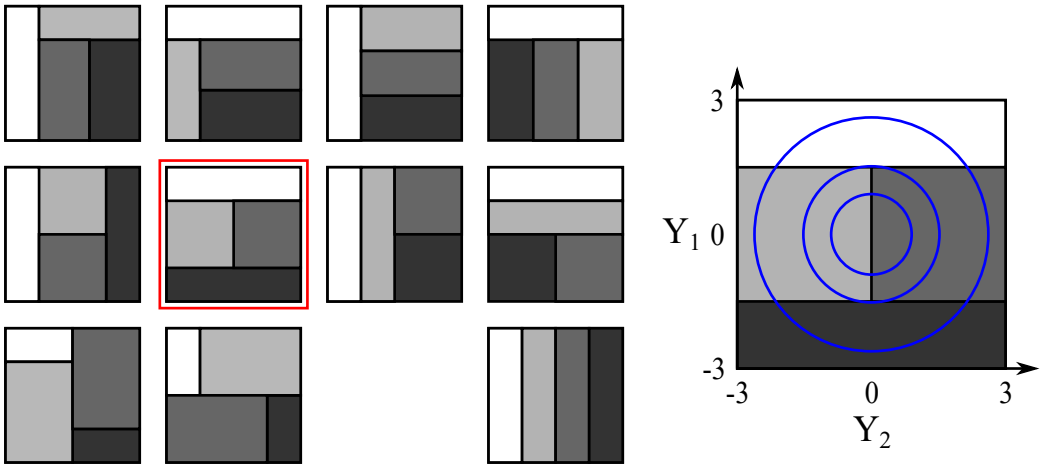


Figure 1.1: Templates for the case of two latent variables and four categories (indicated by differing shades of grey). (modified from Armstrong et al., 2011)

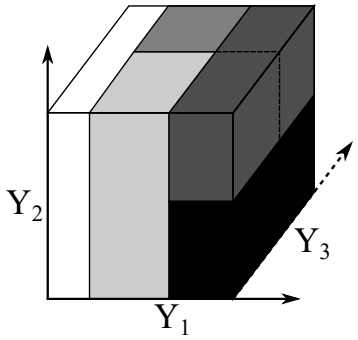


Figure 1.2: TPGS template used to map a categorical variable with 5 categories to a 3D continuous space. (modified from Emery, 2007)

The developed HTPG utilizes a tree structure to define the truncation rules. This facilitates the definition of the mapping between continuous and categorical space making it

possible to efficiently utilize any number of Gaussian variables for the modeling of a categorical variable. Improvements and adaptations to the key steps of the truncated Gaussian framework are developed to ensure the efficient application of the HTPG.

1.2.2 Multivariate Modeling of Categorical Variables

Multiple categorical variables are often available. In mining, for instance, it is common to have several different categorical data such as mineralization zones, alteration zones, and lithology (Figure 1.3). Each of these categorical variables may assume a set of categorical labels. The categorical variables and their combinations are closely related to the mineralogical composition of the rocks, which are linked to metallurgical properties and, therefore, exercise a key role on processing performance (Bowell et al., 2011; Gregory et al., 2013).

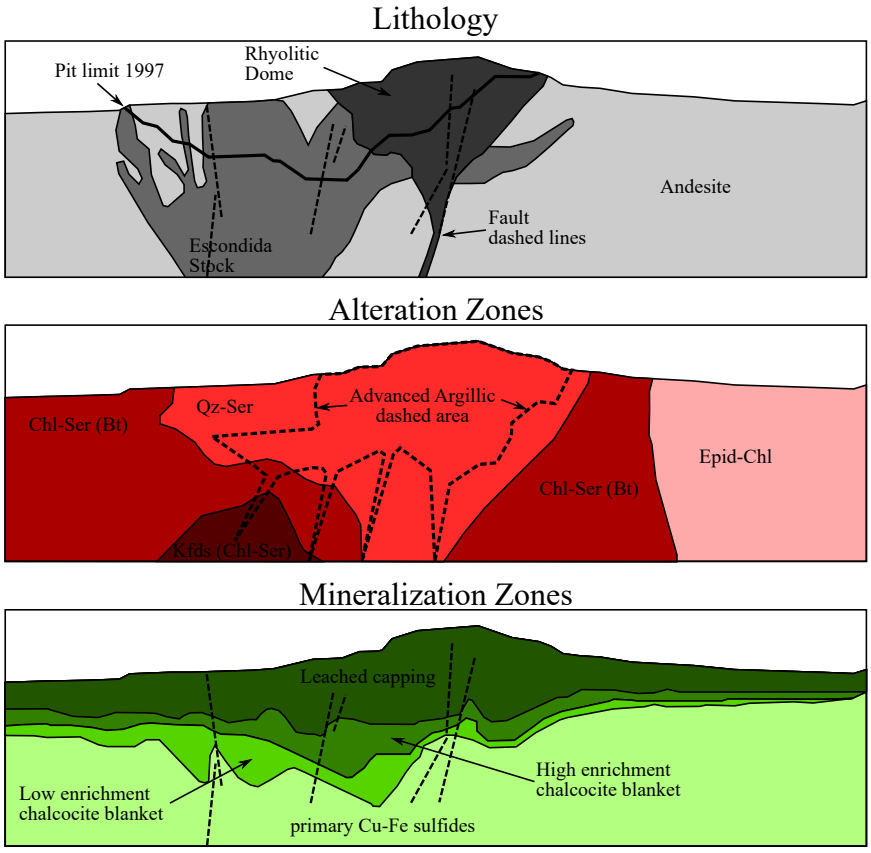


Figure 1.3: Lithology, alteration zones and mineralization zones within a north-south section of La Escondida copper deposit. (modified from Garza et al., 2001)

There has been little research into the modeling of multivariate categorical variables (Emery and Cornejo, 2010). In practice, multiple categorical variables are combined into a

single categorical variable for the purpose of modeling. In some cases, lumping categorical variables is justified in terms of stationary domains, however, this decision is often due to the limitations of existing techniques. The collapse of multiple categorical variables into a single with all possible/observed combinations is likely impractical due to the large number of combinations and difficulty of statistical inference (Rossi and Deutsch, 2014).

Emery and Cornejo (2010) proposed the use of truncated Gaussian simulation (TGS) for the multivariate modeling of categorical variables using LMC to co-simulate the latent variables. The LMC of the latent Gaussian variables is derived iteratively to ensure the reproduction of the spatial structure of the categorical variables. TGS only uses one latent variable per categorical variable. This is too simplistic to reproduce complex categorical ordering and transitions, however, the idea of mapping the categorical variables to a continuous space and using established methodologies for multivariate modeling of continuous variables is promising.

The developed univariate HTPG methodology is extended to perform the multivariate modeling of categorical variables by allowing correlation across the latent variables. This avoids the need for merging categorical variables reducing information loss. The multivariate approach also allows for improved reproduction of multivariate relationships between categorical variables. This ultimately translates into better prediction of resources and reserves including categorical dependent metallurgical performance.

1.2.3 Multiple Data Imputation with Gaussian Mixture Models

Multivariate transformations such as PCA, MAF and PPMT are only possible if all variables are available at all data locations. The SCT method requires a certain ordering in the availability of the variables. Regardless, data imputation is required where some of the variables are not sampled. There have been important advances in the multiple imputation of continuous variables by Barnett and Deutsch (2015). The technique developed by Barnett and Deutsch (2015) relies on the Gibbs sampler for the inference of multivariate distributions. The utilization of a Gibbs sampler is not necessary for this application. This leads to a lack of performance that can be a limiting factor for applications with large data sets. This problem is addressed in this research with the replacement of the Gibbs sampler with a semi-parametric approach using GMMs.

1.3 Thesis Outline

Chapter 2 provides the necessary background for the developments in this thesis. A summary of the geostatistical theory for the modeling of continuous and categorical variables is provided. The Gibbs sampler algorithm is explained followed by the review of the current state of research on data imputation. A review of the main modeling techniques for categorical variables is also provided.

Chapter 3 starts with the mathematical notation for the developed HTPG approach followed by a discussion of the limitations of current practice with the truncated Gaussian methods. The remainder of the chapter is dedicated to the development of the theory and framework for the application of HTPG.

Chapter 4 is dedicated to practical aspects and parameterization of the developed HTPG technique. The chapter starts with the practical considerations for the definition of truncation rules based on qualitative and quantitative factors. Details of the HTPG application in the presence of a categorical trend is also discussed and a framework is proposed. The causes and effects for the hyper-continuous structures observed during the definition of latent variable variograms is also explored in this chapter.

Chapter 5 explores the available implementations of the Gibbs sampler algorithm for the imputation of Gaussian latent variables for the application of the truncated Gaussian methods. A combination of the available techniques is suggested for improved convergence of the algorithm. The impact of multiple versus single data imputation is evaluated with a synthetic example to demonstrate the importance of properly transferring the uncertainty in the unobserved Gaussian latent variables.

Chapter 6 provides a new framework for multiple data imputation of continuous variables for utilization with multivariate transformations such as SCT and PPMT. The framework replaces the utilization of Gibbs sampler for the inference of multivariate conditional distributions by a semi-parametric model based on GMM. The technique is demonstrated for a lateritic nickel data set and compared to common alternatives.

Chapter 7 introduces a new approach for the multivariate modeling of categorical variables based on the HTPG developed in Chapter 3. The chapter starts by highlighting the importance of modeling multiple categories followed by the extension of the HTPG notation to the multivariate case. The framework for the application of the multivariate HTPG

is developed and illustrated with a small 2D data set with two categorical variables. The impact of data correlation on the application of the developed methodology is also explored.

A full scale practical example of the application of both univariate and multivariate HTPG techniques is shown in Chapter 8. The application of the developed methodology is compared with common alternatives. The available data set is split into a training and a test set. The training set is used for modeling and the test set is utilized to validate the results. The effectiveness of each technique are compared in terms of reproduction of data statistics and validation performance.

Concluding remarks are provided in Chapter 9. The chapter starts with a summary of the contributions and results. The limitations of the developed techniques and proposed future work is also outlined in Chapter 9.

Chapter 2

Literature Review and Background

This dissertation involves the modeling of continuous and categorical variables and, therefore, covers a wide range of subjects. The relevant literature is reviewed in the following sections.

2.1 Random Variables, Random Functions and Stationarity

The theory of regionalized variables developed in the 1960's by Georges Matheron (Matheron, 1971) couples probability theory using the concept of random variables (RVs) and random functions (RFs) to spatial problems of modeling subsurface resources. A RV, which is usually denoted by capital letter $Z(\boldsymbol{u})$ where $\boldsymbol{u}$ is a location coordinate vector, is a variable whose values are generated stochastically according to a probabilistic mechanism (Isaaks and Srivastava, 1990) and the RFs are sets of RVs defined within a domain of interest $\{Z(\boldsymbol{u}); \boldsymbol{u} \in \mathcal{A}\}$ and are also denoted as $Z(\boldsymbol{u})$ (Deutsch et al., 1998).

In the multivariate case, a set of n data observations with K variables or properties is represented by $\boldsymbol{z} = \left(\boldsymbol{z}^\top(\boldsymbol{u}_1), \ldots, \boldsymbol{z}^\top(\boldsymbol{u}_n)\right)$ where $\boldsymbol{z}(\boldsymbol{u}_\alpha) \in \mathbb{R}^K$ and is considered to be a realization of the RFs $\boldsymbol{Z}(\boldsymbol{u}_\alpha) = Z_1(\boldsymbol{u}_\alpha), \ldots, Z_K(\boldsymbol{u}_\alpha))$ at the α^{th} data location. A decision of stationarity is required to allow inference. The stationarity decision is often restricted to one-point and two-point statistics (Goovaerts, 1997).

The multivariate cumulative density function (CDF) is deemed invariant under translation ($F_i(z_i) = P\{Z_i(\boldsymbol{u}) \leq z_i\};\ i \in \{1, \ldots, K\};\ \boldsymbol{u} \in \mathcal{A}$); the first order moments are deemed independent of location ($\mathrm{E}\{Z_i(\boldsymbol{u})\} = \mu_i;\ i \in \{1, \ldots, K\};\ \forall \boldsymbol{u} \in \mathcal{A}$); the two-point joint CDFs depend only on separation vector $\boldsymbol{h}$ between two locations ($F_{i,j}(\boldsymbol{h}; z_i, z_j) = P\{Z_i(\boldsymbol{u}) \leq z_i, Z_j(\boldsymbol{u}+\boldsymbol{h}) \leq z_j\};\ \forall i, j \in \{1, \ldots, K\};\ \forall \boldsymbol{u} \in \mathcal{A}$); and the second order moments also depend solely on $\boldsymbol{h}$ ($C_{i,j}(\boldsymbol{h}) = \mathrm{E}\{[Z_i(\boldsymbol{u}) - \mu_i][Z_j(\boldsymbol{u}+\boldsymbol{h}) - \mu_j]\};\ \forall i, j \in \{1, \ldots, K\};\ \forall \boldsymbol{u} \in \mathcal{A}$)

2.2 Spatial Variability and Continuity

The most common measure of spatial variability utilized in geostatistics is the semivariogram ($\gamma(\boldsymbol{u})$). The semivariogram is defined as one half of the variogram. The Equation 2.1 show the variogram definition for a stationary RFs $\boldsymbol{Z}(\boldsymbol{u}) \in \mathbb{R}^K$. As the semivariogram is almost always utilized instead of the variogram, the term variogram is utilized to refer to the semivariogram throughout this dissertation. When $i = j$ in Equation 2.1, the variogram is referred to as direct variogram, whereas when $i \neq j$, the variograms are referred as cross-variograms.

$$\begin{aligned} 2\gamma_{i,j}(\boldsymbol{h}) &= Cov\left\{\left[Z_i(\boldsymbol{u}) - Z_i(\boldsymbol{u}+\boldsymbol{h})\right], \left[Z_j(\boldsymbol{u}) - Z_j(\boldsymbol{u}+\boldsymbol{h})\right]\right\} \\ &= \mathrm{E}\left\{\left[Z_i(\boldsymbol{u}) - Z_i(\boldsymbol{u}+\boldsymbol{h})\right]\left[Z_j(\boldsymbol{u}) - Z_j(\boldsymbol{u}+\boldsymbol{h})\right]\right\} \end{aligned} \qquad \forall i,j \in \{1,\ldots,K\} \quad (2.1)$$

Experimental variogram values are calculated with Equation 2.2. The spatial variability is often anisotropic and the variogram must be inferred for several directions and lag distances. Tolerances are allowed during the calculation to ensure stable estimation of the spatial variability. Three main directions (major, mid and minor) of anisotropy are defined. These directions are orthogonal to each other. The major direction is the direction of greatest continuity, the minor direction is the direction of smallest continuity and the mid directions is locked by the orthogonality requirement and the ellipsoidal nature of the anisotropy modeled in geostatistics.

$$\hat{\gamma}_{i,j}(\boldsymbol{h}) = \frac{1}{2\left|N(\boldsymbol{h})\right|} \sum_{\alpha \in N(\boldsymbol{h})} \left[z_i(\boldsymbol{u}_\alpha) - z_i(\boldsymbol{u}_\alpha+\boldsymbol{h})\right]\left[z_j(\boldsymbol{u}_\alpha) - z_j(\boldsymbol{u}_\alpha+\boldsymbol{h})\right] \qquad \forall i,j \in \{1,\ldots,K\} \quad (2.2)$$

where $N(\boldsymbol{h})$ is the set locations with data pairs that are used for the calculation of the experimental variogram for lag vector $\boldsymbol{h}$.

By further expanding the Equation 2.1 a relationship between the spatial covariance and the variogram can be established (Equation 2.3). This relationship assumes symmetry of the covariance function $\{C_{i,j}(-\boldsymbol{h}) = C_{i,j}(+\boldsymbol{h});\ \forall i,j \in 1,\ldots,K\}$. This relationship is utilized to build the system of equations utilized for estimation as the variograms are often modeled, but the covariance values are the ones required. To ensure that the system of

equations have a solution, the covariance function must be positive definite. To ensure positive definiteness, the experimental variogram points are modeled with specific functions that are known to generate positive definite covariances.

$$\gamma_{i,j}(\boldsymbol{h}) = C_{i,j}(0) - C_{i,j}(\boldsymbol{h}) \qquad \forall i,j \in \{1,\ldots,K\} \tag{2.3}$$

The three most commonly used variogram models are the spherical (Equation 2.4), exponential (Equation 2.5) and the Gaussian variogram (Equation 2.6). Note that h is the scalar normalized distance calculated with the equation of an ellipsoid (Equation 2.7).

$$\text{Sph}(h) = \begin{cases} 1.5h - 0.5h^3, & \text{if } h \leq 1 \\ 1, & \text{otherwise} \end{cases} \tag{2.4}$$

$$\text{Exp}(h) = 1 - \exp(-3h) \tag{2.5}$$

$$\text{Gaus}(h) = 1 - \exp\left(-3h^2\right) \tag{2.6}$$

$$h = \sqrt{\left(\frac{h_{\text{major}}}{a_{\text{major}}}\right)^2 + \left(\frac{h_{\text{minor}}}{a_{\text{minor}}}\right)^2 + \left(\frac{h_{\text{mid}}}{a_{\text{mid}}}\right)^2} \tag{2.7}$$

These models are positive definite in 3D and that also ensures the positive definiteness in lower dimensions. Experimental variograms can be fitted with nested structures composed of multiple variogram types. Nesting structures to fit direct and cross-variograms in the multivariate case must consider a model of coregionalization such as the linear model of coregionalization (LMC) to ensure positive definiteness.

The main features of the variogram are the sill, range and nugget. The sill is equal to the variance of the data used to calculate the variogram. The range is an anisotropic parameter that represents the extension of the spatial continuity. Data locations separated by a distance greater than the variogram range for the given direction are not spatially correlated. The variogram has zero value at the origin, however, experimental points can show relatively high variogram values close to the origin. This is attributed to random measurement errors and small scale variability (Rossi and Deutsch, 2014). This discontinuity at the origin is isotropic and it is called nugget effect.

2.3 Modeling Continuous Variables

The conventional geostatistical methods assume that variables are multivariate Gaussian after univariate transformation and utilize the LMC for modeling. Simplifications of the LMC are commonly used to avoid modeling a full LMC.

In most cases, univariate transformation of geological variables do not lead to multivariate Gaussianity. Complex relationships between the variables remain after the transformation that cannot be matched with conventional methods. Advanced techniques make use of multivariate transformations to improve the Gaussianity of resulting variables and allow for a more efficient use of Gaussian methods.

The latent variables utilized by the truncated Gaussian methods are not observed. The categorical variables are mapped to a multivariate Gaussian space, therefore, conventional LMC and its derivatives are utilized for modeling. Even though the truncated Gaussian approach only makes use of conventional modeling techniques, both conventional and advanced techniques are reviewed in this section. This is because of the secondary contributions of this dissertation to the problem of multiple data data imputation and data transformation using stepwise conditional transform (SCT).

2.3.1 Linear Model of Coregionalization

Modeling a RF requires the assessment of high dimensional conditional and marginal distributions. This is not possible without a flexible parametric model. The first step of the modeling workflows is to map the original variables to the Gaussian space. Gaussian distribution is chosen because it is highly tractable and fully parameterized by mean vector and covariance matrix. The variables are independently transformed to Gaussian using normal scores transformation (Equation 2.8) (Bliss, 1934; Deutsch et al., 1998). This transformation only ensures that marginal distributions are Gaussian, however, the modeling is often carried forward under assumption of multivariate Gaussianity. Note that Φ is utilized throughout this dissertation to denote the Gaussian CDF and Φ^{-1} is used for its inverse. The Gaussian probability density function (PDF) will be denoted by the lower case (ϕ). The mean and covariance parameters are omitted for the standard Gaussian distributions.

$$y_i(\boldsymbol{u}_\alpha) = \Phi^{-1}\left(F_i\left(z_i(\boldsymbol{u}_\alpha)\right)\right), \qquad \alpha \in \{1,\dots,n\} \quad \text{and} \quad i \in \{1,\dots,K\} \tag{2.8}$$

Under multivariate Gaussianity assumption the parameters of any conditional distribution can be calculated by solving the simple cokriging system of equations that are also known as normal equations. This requires the definition of all covariances between all variables ($C_{i,j}(\boldsymbol{h})$;$\forall i,j \in \{1,\dots,K\}$) for any given lag vector $\boldsymbol{h}$. The covariances may be inferred from data for a limited number of lags, however, a valid (positive definite) continuous model must be defined. This allows the calculation of valid covariances for any arrangement of conditioning data to be used for inference at any location $\boldsymbol{u} \in \mathcal{A}$. This continuous model is referred to as coregionalization model in the multivariate case.

The LMC assumes that each RF $Z_i(\boldsymbol{u})$ ($i \in \{1,\dots,K\}$) is a linear combination of underlying standard independent RFs $Y_k^l(\boldsymbol{u})$ (Equation 2.9) with $l \in \{0,\dots,L\}$ and $k \in \{1,\dots,n_l\}$ where $L+1$ is the number of covariance structures and n_l the number of independent factors that shares the same covariance structure $c_l(\boldsymbol{h})$ (Goovaerts, 1997). The resulting coregionalization model is shown in Equation 2.10.

$$Z_i(\boldsymbol{u}) = \sum_{l=0}^{L}\sum_{k=1}^{n_l} a_{i,k}^l Y_k^l(\boldsymbol{u}) + \mu_i \qquad \forall i \in \{1,\dots,K\} \tag{2.9}$$

$$C_{i,j}(\boldsymbol{h}) = \sum_{l=0}^{L}\sum_{k=1}^{n_l} a_{i,k}^l a_{j,k}^l c_l(\boldsymbol{h}) \qquad \forall i,j \in \{1,\dots,K\} \tag{2.10}$$

The LMC entails the definition of $K(K+1)/2$ direct and cross-covariance models. This can become a tedious task with increasing number of RF's to model. Also, the utilization of LMC model for estimation and simulation requires the solution of a system of linear equations that often grows quadratically in terms of number of variables. The computational cost and practicality of the LMC utilization motivates the utilization of simplified alternatives such as the intrinsic model of coregionalization (IMC) (Almeida and Journel, 1994; Babak and Deutsch, 2009; Goovaerts, 1997; Wackernagel, 2003).

2.3.2 Cokriging

Kriging is a least-squares regression utilized in geostatistics for inference of regionalized variables (Goovaerts, 1997) that is named after the work of Krige (1951). Kriging is referred to as cokriging in the multivariate case. The kriging estimator at a location $\boldsymbol{u}_0 \in \mathcal{A}$ for

the RV $Z_k(\boldsymbol{u})$ in a set of multiple RVs $Z_i(\boldsymbol{u})$ ($i \in \{1, \ldots, K\}$), can be written as shown in Equation 2.11. The kriging estimate is also referred to as kriging mean.

$$\hat{Z}_k(\boldsymbol{u}_0) - m_k(\boldsymbol{u}_0) = \sum_{i=1}^{K} \sum_{\alpha \in N_i(\boldsymbol{u}_0)} \lambda_{\alpha,i} \left[Z_i\left(\boldsymbol{u}_\alpha\right) - m_i(\boldsymbol{u}_\alpha) \right] \tag{2.11}$$

where $N_i(\boldsymbol{u}_0)$ is the set of locations where variable i is available and that are utilized for the estimation at location $\boldsymbol{u}_0$.

The weights ($\boldsymbol{\lambda}$) are defined to minimize the error variance (Equation 2.12) under the constraint that the expected error is zero (Equation 2.13). These properties grants kriging the quality of best linear unbiased estimation or BLUE.

$$\sigma_{\mathrm{E}}^2 = \mathrm{Var}\left\{ \hat{Z}_k\left(\boldsymbol{u}\right) - Z_k\left(\boldsymbol{u}\right) \right\} \tag{2.12}$$

$$\mathrm{E}\left\{ \hat{Z}_k\left(\boldsymbol{u}\right) - Z_k\left(\boldsymbol{u}\right) \right\} = 0 \tag{2.13}$$

For a stationary multivariate Gaussian RVs $Y_i(\boldsymbol{u})$ ($i \in \{1, \ldots, K\}$) with zero mean ($\mathrm{E}\{Y_i(\boldsymbol{u})\} = 0 \; \forall i \in \{1, \ldots, K\}$) and covariance function $C_{i,j}(\boldsymbol{h})$ ($i, j \in \{1, \ldots, K\}$), the system of equations that minimizes the error variance (Equation 2.12) can be written in matrix form as shown in Equation 2.14.

$$\begin{bmatrix} \boldsymbol{C}_{1,1} & \boldsymbol{C}_{1,2} & \ldots & \boldsymbol{C}_{1,K} \\ \boldsymbol{C}_{2,1} & \boldsymbol{C}_{1,2} & \ldots & \boldsymbol{C}_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K,1} & \boldsymbol{C}_{K,2} & \ldots & \boldsymbol{C}_{K,K} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{k,1}^{(0)} \\ \boldsymbol{C}_{k,2}^{(0)} \\ \vdots \\ \boldsymbol{C}_{k,K}^{(0)} \end{bmatrix} \tag{2.14}$$

where $\boldsymbol{C}_{i,j}$ ($i, j \in \{1, \ldots, K\}$) is a submatrix with all the covariances between the data locations for the RV's $Y_i(\boldsymbol{u})$ and $Y_j(\boldsymbol{u})$ for the sets $N_i(\boldsymbol{u}_0)$ and $N_j(\boldsymbol{u}_0)$ utilized for the estimation at location $\boldsymbol{u}_0$. The weight vector $\boldsymbol{\lambda}_i$ contains all the weights applied to the variable i at the locations in the set $N_i(\boldsymbol{u}_0)$. The covariance vector $\boldsymbol{C}_{k,i}^{(0)}$ ($k, i \in \{1, \ldots, K\}$) contains all the covariances between the set of data $N_i(\boldsymbol{u}_0)$ and the variable $Y_k(\boldsymbol{u}_0)$.

An estimate of the error variance (Equation 2.12) is also obtained from the solution of the kriging system. The minimized error variance is also known as the kriging variance. The kriging variance for the estimation of variable $Y_k(\boldsymbol{u})$ at location $\boldsymbol{u}_0$ is given by Equation 2.15.

The kriging mean and variance are utilized to parameterize high-dimensional conditional distributions under multiGaussian assumption for the stochastic simulation of Gaussian random functions (GRFs).

$$\hat{\sigma}^2_{E_k}(\boldsymbol{u}_0) = C(0)_{k,k} - \sum_{i=1}^{K} \boldsymbol{\lambda}_i^{\top} \boldsymbol{C}^{(0)}_{k,i} \tag{2.15}$$

The covariance matrix on the left side of the Equation 2.14 must be inverted. The computational time to solve this system raises quickly with the increasing number of variables and data available. Alternatives such as collocated cokriging (CCK) (Almeida and Journel, 1994) and intrinsic intrinsic collocated cokriging (ICCK) (Babak and Deutsch, 2009) are often utilized to decrease the computational burden and to simplify the parameterization task (variogram modeling).

2.3.3 Intrinsic Model of Coregionalization

The LMC model shown in Equation 2.10 is often written in the simplified form shown in Equation 2.16. The the $b^l_{i,j}$ coefficients are defined by Equation 2.17 (Goovaerts, 1997).

$$C_{i,j}(\boldsymbol{h}) = \sum_{l=0}^{L} b^l_{i,j} c_l(\boldsymbol{h}) \qquad \forall i,j \in \{1,\dots,K\} \tag{2.16}$$

$$b^l_{i,j} = \sum_{k=1}^{n_l} a^l_{i,k} a^l_{j,k} \qquad \forall i,j \in \{1,\dots,K\}, \qquad \forall l \in \{1,\dots,L\} \tag{2.17}$$

The IMC is a simplification of the LMC model in which the ratio of correlation between two RFs is independent of the spatial component (Wackernagel, 2003). All coefficients $b^l_{i,j}$ are proportional to each other, that is $b^l_{i,j} = \eta_{i,j} b_l \ \forall i,j \in \{1,\dots,K\}$ and $l \in \{1,\dots,L\}$ (Goovaerts, 1997). As result the, IMC model can be written as shown in Equation 2.18.

$$C_{i,j}(\boldsymbol{h}) = \eta_{i,j} \sum_{l=0}^{L} b^l c_l(\boldsymbol{h}) \qquad \forall i,j \in \{1,\dots,K\} \tag{2.18}$$

Under second-order stationarity, the coefficients $\eta_{i,j}$ are equivalent to the variances $C_{i,i}(0)$ and covariances $C_{i,i}(0)$ ($i \neq j$). The IMC is limited as it entails that all variables share the same basic spatial structure that is scaled based on the zero lag covariance matrix. If the variables are standardized, all direct variograms will be the same.

2.3.4 Intrinsic Collocated Cokriging

In spite of its limitations, the IMC has been extensively used in cases where exhaustive secondary data is available. In these cases, the selection of the data used for estimation is important as the availability of secondary quickly contributes to enlarge the system of equations to be solved (Equation 2.14). Three options for the cokriging neighbourhood are shown in Figure 2.1.

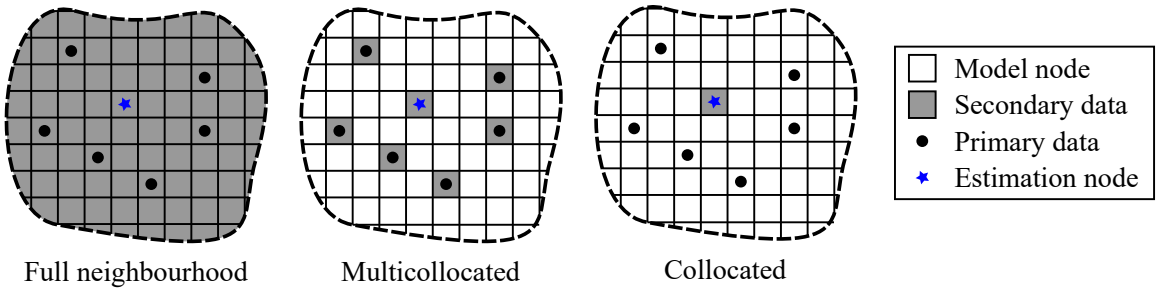


Figure 2.1: Three options of neighbourhoods for cokriging. (modified from Wackernagel, 2003)

The CCK approach is based on an intrinsic model referred to as Markov model (Almeida and Journel, 1994) that uses the collocated neighbourhood (Figure 2.1). The main attraction of CCK is its simplicity and significant gain in computational time, however, the CCK is known to introduce variance inflation. The causes of this variance inflation have been derived by Babak and Deutsch (2009). Babak and Deutsch (2009) also show that the CCK systems are not intrinsic and propose the utilization of the multicollocated neighborhood (Figure 2.1) with an IMC, calling it the ICCK. Manchuk and Deutsch (2016) show that the computational complexity of ICCK can be substantially reduced, to levels comparable to CCK, when it is utilized with exhaustive secondary data.

2.3.5 Sequential Gaussian Simulation

Kriging is used to generate deterministic models with the single best estimated values. It is a moving weighted average that results in smooth models that do not have the correct spatial variability (Deutsch et al., 1998). The correct spatial variability is crucial for reliable prediction of key features such as dilution. Simulation is utilized in order account for the correct spatial variability and to carry uncertainty forward, through transfer functions, to response variables such as dilution, recovery, resources and reserves.

The sequential Gaussian simulation (SGS) framework (Gómez-Hernández and Journel,

1993; Isaaks, 1990) is a Monte-Carlo simulation (MCS) approach utilized to generate realizations of GRFs (Deutsch et al., 1998; Goovaerts, 1997). The SGS algorithm requires the definition of high-dimensional conditional distributions. These distributions cannot be inferred from data due to dimensionality of the problem (Leuangthong et al., 2008). This problem is often referred to as the curse of dimensionality (Bellman, 2003). A practical option is to utilize a parametric model. Under multiGaussian assumption, all the conditional distributions required in the SGS approach can be defined by the kriging mean and variance.

The SGS algorithm for the generation of a numeric model with a realization of the Gaussian RVs proceeds as follows:

(1) define a random path through the model nodes
(2) for each variable from 1 to K:
    (2.1) for each node in the random path:
        (2.1.1) search for conditioning data
        (2.1.2) build and solve kriging system
        (2.1.3) sample the conditional distribution defined by kriging mean and variance
        (2.1.4) assign sampled value to the model node and include the simulated node to the conditioning data ensemble

As the model nodes are populated with simulated values, the number of conditioning data grows. The large number of conditioning data leads to unreasonably large kriging system (Equation 2.14). In practice, only the closest data to the simulation location are used to define the parameters for the conditional distributions. The assumption that the closest data screens out the influence of all other conditioning presumes a Markovian behavior (Gómez-Hernández and Journel, 1993; Manchuk and Deutsch, 2012). To minimize the impact of search restriction on the reproduction of spatial variability, the random path is divided into multiple grids. The grid is populated starting from a coarse grid subset that is refined until the final desired resolution.

2.4 Multivariate Transformations

Geologic variables show complex relationships including non-linearity, compositional constraints and heteroskedasticity which are not removed by univariate normal scores transform (Equation 2.8). This leads to poor reproduction of these complex features in conventional workflows (Barnett, 2015). Multivariate transformations such as SCT and projection pursuit multivariate transformation (PPMT) are developed to ensure the appropriate transformation of original variables to multivariate Gaussian space.

2.4.1 Stepwise Conditional Transform

The SCT is calculated in an ordered fashion where the first variable is independently transformed to normal through normal scores transform, the second variable is transformed based on conditional distributions given the first variable, and the i^{th} variable is transformed based on conditional distributions given the previous $i-1$ variables (Leuangthong and Deutsch, 2003; Rosenblatt, 1952) as in Equation 2.19.

$$\begin{aligned} y_1\left(\boldsymbol{u}_\alpha\right) &= \Phi^{-1}\left(F_1\left(z_1\left(\boldsymbol{u}_\alpha\right)\right)\right) \\ y_2\left(\boldsymbol{u}_\alpha\right) &= \Phi^{-1}\left(F_{2|1}\left(z_2\left(\boldsymbol{u}_\alpha\right) \mid z_1\left(\boldsymbol{u}_\alpha\right)\right)\right) \\ &\vdots \\ y_K\left(\boldsymbol{u}_\alpha\right) &= \Phi^{-1}\left(F_{K|1,\ldots,K-1}\left(z_K\left(\boldsymbol{u}_\alpha\right) \mid z_1\left(\boldsymbol{u}_\alpha\right), \ldots, z_{K-1}\left(\boldsymbol{u}_\alpha\right)\right)\right) \end{aligned} \quad , \qquad \forall \alpha \in \{1, \ldots, n\} \tag{2.19}$$

The resultant transformed variables are multivariate Gaussian and independent at zero lag distance. All the complex relationships between the geological variables are removed with the transformation. There is no guarantee of decorrelation at lags different from zero. The transformed values in Equation 2.19 are used as conditioning data for simulation within the domain of interest. Simulated values ($y_i(\boldsymbol{u}), \forall \boldsymbol{u} \in \mathcal{A}, i \in \{1, \ldots, K\}$) are transformed back to original units using Equation 2.20. The back transformation ensures the reproduction of the complex relationships between variables.

$$
\begin{aligned}
z_K(\boldsymbol{u}) &= F^{-1}_{K|1,\ldots,K-1}\left(\Phi\left(y_K(\boldsymbol{u}) \mid y_1(\boldsymbol{u}),\ldots,y_{K-1}(\boldsymbol{u})\right)\right) \\
&\vdots \\
z_2(\boldsymbol{u}) &= F^{-1}_{2|1}\left(\Phi\left(y_2(\boldsymbol{u}) \mid y_1(\boldsymbol{u})\right)\right) \\
z_1(\boldsymbol{u}) &= F^{-1}_{1}\left(\Phi\left(y_1(\boldsymbol{u})\right)\right)
\end{aligned}
\quad , \qquad \forall \boldsymbol{u} \in \mathcal{A} \tag{2.20}
$$

The methodology suffers from poor variogram reproduction in cases in which it fails to decorrelate variables at lags different from zero. The SCT model relies on the binning of the multivariate space for the calculation of the conditional distributions required for transformation (Leuangthong and Deutsch, 2003). The binning approach makes the SCT very limited with respect to the number of variables that can be modeled.

Alternatives to allow the use of SCT in high dimensions and for sparse data have been proposed to eliminate binning artifacts from discrete partitioning of multivariate space. Leuangthong and Deutsch (2003) proposed the use of kernel density estimation (KDE), Manchuk and Deutsch (2011) proposed kernel density networks that is computationally more efficient than KDE.

2.4.2 Projection Pursuit Multivariate Transform

The PPMT technique developed by Barnett (2015) utilizes a modified component of projection pursuit density estimation (PPDE) developed by Friedman (1987) for multivariate transformation of complex geological data to be multivariate Gaussian. The methodology is applied in steps.

The first step is the univariate normal scores transform (Equation 2.8) that centers all variables and makes them more stable for subsequent steps. The resulting marginally Gaussian data matrix with n transformed data observations is denoted by $\boldsymbol{y} = \left(\boldsymbol{y}^\top(\boldsymbol{u}_1),\ldots,\boldsymbol{y}^\top(\boldsymbol{u}_n)\right)$. A variant of principal component analysis (PCA) sphering is used in PPMT methodology. The sphering step requires the eigen value decomposition of the covariance matrix at lag zero ($\boldsymbol{C}(0) = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$), where $\boldsymbol{V}$ is the matrix of eigenvectors, $\boldsymbol{D}$ is a diagonal matrix with eigenvalues and $\boldsymbol{C}(0)$ is the $K \times K$ covariance matrix at lag zero. The sphered variables are calculated as shown in Equation 2.21 that can be interpreted (reading from the right) as: rotate to the principal components basis, standardize and rotate back to original basis.

$$\boldsymbol{y}_0 = \boldsymbol{V}\boldsymbol{D}^{-1/2}\boldsymbol{V}^\top \boldsymbol{y} \tag{2.21}$$

After the sphering, the PPMT methodology proceeds to the projection pursuit iterations. In each iteration (t), the most non-Gaussian projection ($\boldsymbol{p}_t = \hat{\boldsymbol{\Theta}}_t \boldsymbol{y}_{t-1}$) of the data is found and transformed to be Gaussian. The degree of non-Gaussianity of a projection is defined by the projection index $I(\boldsymbol{\Theta})$, therefore, $\hat{\boldsymbol{\Theta}}$ is defined as $\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\, I(\boldsymbol{\Theta})$. After a number of iterations the data is transformed to be multivariate Gaussian. All the operations on the original data matrix are stored and used later to transform all simulated values to original units restoring the complex relations observed in data.

2.5 Gibbs Sampler

The Gibbs sampler algorithm (Geman and Geman, 1984; Metropolis et al., 1953) is a Markov chain Monte Carlo method commonly used for statistical inference. The technique is utilized to generate samples from distributions for which the direct sampling is difficult. It is particularly useful when marginal distributions can be easily sampled.

Given a target K-dimensional distribution $f(\boldsymbol{Y})$ where $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$, and $\boldsymbol{y} = (y_1, \ldots, y_K)$ is a sample of the RV $\boldsymbol{Y}$. Also, assuming that all the conditional distributions $f(y_i|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_K)\ \forall i \in \{1, \ldots, K\}$ can be readily sampled. The Gibbs sampler algorithm can be utilized to generate a sample of $f(\boldsymbol{Y})$ by the following steps.

(1) initialize the iteration counter: $t = 0$
(2) define initial state by defining a valid arbitrary vector $\boldsymbol{y}^{(0)}$
(3) for each dimension $i = \{1, \ldots, K\}$:
  (3.1) increment iteration counter: $t = t + 1$
  (3.2) set $y_j^{(t)} = y_j^{(t-1)}\ \forall j \neq i$
  (3.3) draw a random sample for $y_i^{(t)}$ from $f(y_i|y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_K)$
  (3.4) terminate the algorithm if $t = t_{\text{MAX}}$

For a sufficient number of iterations t_{MAX} the resulting vector $\boldsymbol{y}^{(t_{\text{MAX}})}$ is a valid sample of the reference joint distribution. The number of iterations needed for the convergence of the Gibbs sampler algorithm is commonly referred to as burn-in iterations. To generate additional samples after the burn-in period, the algorithm can be run from steps 3.1 to 3.3

multiple times recovering the state of $\boldsymbol{y}^{(t)}$ at every n^{th} increment of the iteration counter. This buffer between two consecutive samples is called thinning and is utilized to ensure that the generated samples are independent from each other. The buffer size as well as the number of burn-in iterations relates to the the correlation between the RVs.

One of the applications of Gibbs sampler in geostatistics is for the imputation of missing data (Barnett and Deutsch, 2015; Little and Rubin, 2002). The imputation is important to generate complete (isotopic) data sets from incomplete (heterotopic) ones, allowing their use with multivariate transformations. The Gibbs sampler is also utilized to define the unobserved latent variables for the truncated Gaussian methods (Armstrong et al., 2011; Emery et al., 2014; Galli and Gao, 2001). The imputation of latent variables must respect the mapping between the categorical and continuous space while also matching the categorical data observations.

2.6 Data Imputation for Multivariate Transformations

High dimensional geological data have become standard in the mining and petroleum industry with the increasing availability of multiple measurements per sample. These multivariate data sets present complexities that cannot be reproduced by conventional geostatistical techniques and therefore are more suited to advanced multivariate techniques. Many advanced multivariate geostatistical workflows require multivariate data transformation such as PCA, minimum/maximum auto-correlation factors (MAF), SCT, PPMT, among others. These transformations can only be applied to isotopic observations (PCA, MAF and PPMT) or where there is a particular ordering to the missing data (SCT is performed in ordered manner and the transformation of a variable only requires the previously transformed ones). In practice, however, it is common to find data sets that are not completely informed with all variables at all locations for various reasons. In this context, the heterotopic observations must be discarded or the missing values imputed.

Excluding heterotopic observations would ignore expensive data and often lead to biased models (Little and Rubin, 2002) as missing geological observations are not missing randomly. At times the presence of missing variables is driven by the value of the measured variables. Low values for primary measured variables may not justify the expense of additional measurements. Also, legacy data will have different measurements relative

to more recently collected data. These cases are handled with multiple imputation (MI) (Barnett and Deutsch, 2015; Rubin, 1978) that will allow the use of all available information in subsequent modeling steps without the introduction of bias. The uncertainty from the missing data will be transferred to the final models while maintaining the value from the measured variables.

Barnett and Deutsch (2015) proposed two methodologies for imputation of geological data based on Bayesian updating (BU) (Ren, 2007). The idea is to simulate the missing variables to generate multiple isotopic data sets in order to carry forward the uncertainty from the missing values through the rest of the workflow while reproducing the complexities from the original multivariate data distribution and spatial structure.

The first methodology proposed by Barnett and Deutsch (2015) is called parametric merged method and is referred here as parametric only. This technique is fully parametric and assumes that the data follow a multivariate Gaussian distribution. This assumption makes it difficult to reproduce the complexities of the original data, however, it may perform surprisingly well due to the strong conditioning from nearby and colocated data (Barnett and Deutsch, 2015).

The second methodology is referred as non-parametric and uses univariate KDE and a Gibbs sampler to estimate the conditional distribution given the colocated data used in BU. Barnett and Deutsch (2015) showed that the non-parametric technique improves considerably upon the parametric and works well in reproducing multivariate complexities observed in the data set. The computational expense of the KDE calculations and the need for Gibbs sampler iterations increase considerably with the number of observations and dimensions.

2.7 Modeling Categorical Variables

Agresti (2002) classifies categorical variables in three main groups: (1) nominal, (2) ordinal and (3) interval. Nominal variables are those for which the ordering of categories is irrelevant. Ordinal variables present clear natural ordering, however, the distances between the categories are unknown. Interval variables are ordered and also have defined numerical distance between the categories. Categorical variables can also be binary (dichotomous) or have multiple categories (polychotomous).

Geological domains are most often defined by multiple categories of nominal and ordinal types. Examples of ordinal categorical variables in the geological context are sedimentary sequences formed in dispositional environment and domains defined by intensity levels in a qualitative scale such as degrees of alteration and weathering. Nominal variables in the geological context are diffuse in nature with no clear genetic shape or that have been modified by late stage events such as intrusions and fractures.

In the mining industry, categorical variables are mostly used to represent stationary domains (Pyrcz and Deutsch, 2014; Rossi and Deutsch, 2014) within which continuous variables such as grades are estimated or simulated using geostatistical techniques. There are two principal branches for the modeling of these domains: (1) deterministic and (2) stochastic.

The deterministic approaches include the parametric wireframing process (Bezier et al., 1974), discrete smooth interpolation (Mallet, 2002), interpolation of volume functions (Cowan et al., 2003) and signed distance functions (Hosseini et al., 2009; McLennan, 2007; McLennan and Deutsch, 2006). Deterministic models do not consider uncertainty and do not represent small scale variability (Silva, 2015).

Several stochastic categorical modeling algorithms exists. Each have their own applications depending on the nature of the problem and the modeling goals. The main techniques utilized in conventional geostatistics are reviewed.

2.7.1 Indicator Formalism

The application of indicators in geostatistics is originally proposed for the non-parametric estimation of spatial distributions (Journel, 1983). The approach is particularly useful for modeling variables with long-tailed distributions and high coefficient of variation. The indicator formalism became the basis for the indicator conditional simulation (Deutsch, 2006; Journel and Isaaks, 1984).

The application of indicators for the modeling of continuous variables requires the binning of the data resulting in information loss. This is not the case with categorical variables that inherently have binned distributions. The indicator function utilized to transform the categorical variables to indicators is provided in Equation 2.22.

$$\mathbf{1}_i(x(\boldsymbol{u})) = \mathbf{1}_i(\boldsymbol{u}) = \begin{cases} 1, & \text{if } x(\boldsymbol{u}) = b_i \\ 0, & \text{otherwise} \end{cases}, \qquad \forall i \in \{1, \ldots, B\} \tag{2.22}$$

where $x(\boldsymbol{u})$ is a sample of the categorical RV $X(\boldsymbol{u})$; b_i is the i$^{\text{th}}$ category of a finite set $\mathcal{B}$ of possible categories; and $B = |\mathcal{B}|$.

The indicator variable can be interpreted as the probability of a category at a given location. At hard data locations this probability is either 0 or 1. The estimation of local distributions is performed with kriging. This particular class of kriging is called indicator kriging (IK).

The spatial variability of the indicator variables is modeled through the indicator variogram (Equation 2.23). The indicator variogram is inferred from data similarly to the continuous case (Equation 2.24).

$$\gamma_{\mathbf{1}i,j}(\boldsymbol{h}) = \frac{1}{2}\mathrm{E}\left\{\left[\mathbf{1}_i(\boldsymbol{u}) - \mathbf{1}_i(\boldsymbol{u}+\boldsymbol{h})\right]\left[\mathbf{1}_j(\boldsymbol{u}) - \mathbf{1}_j(\boldsymbol{u}+\boldsymbol{h})\right]\right\}, \qquad i,j \in \{1, \ldots, B\} \tag{2.23}$$

$$\hat{\gamma}_{\mathbf{1}i,j}(\boldsymbol{h}) = \frac{1}{2\,|N(\boldsymbol{h})|}\sum_{\alpha \in N(\boldsymbol{h})}\left[\mathbf{1}_i(\boldsymbol{u}_\alpha) - \mathbf{1}_i(\boldsymbol{u}_\alpha+\boldsymbol{h})\right]\left[\mathbf{1}_j(\boldsymbol{u}_\alpha) - \mathbf{1}_j(\boldsymbol{u}_\alpha+\boldsymbol{h})\right], \quad \forall i,j \in \{1, \ldots, B\} \tag{2.24}$$

where $N(\boldsymbol{h})$ is the set locations with data pairs that are used for the calculation of the experimental variogram at lag vector $\boldsymbol{h}$.

The indicator cross-variograms ($i \neq j$ in Equation 2.24) are often ignored as they show extreme continuity at short lags and cannot be modeled with conventional coregionalization models (Machuca-Mory and Deutsch, 2006).

The experimental indicator variograms are modeled with valid models similarly to the continuous variables. These models are utilized to built the kriging systems for the estimation of the indicator variable at the model nodes. One kriging system per category is solved for each location. The estimated indicator variable represents the probability of the corresponding category. The estimated probabilities are not guaranteed to be non-negative nor to sum to one, as it is required for a closed set of probabilities. Often, non-negative values are reset to zero and the remaining estimates are rescaled by their sum (Deutsch, 2006).

The most common kriging estimators utilized for IK with categorical variables are simple kriging (SK) (Equation 2.25) that assumes stationarity of the global proportions ($p_i(\boldsymbol{u}) = p_i;\ \forall \boldsymbol{u} \in \mathcal{A};\ \forall i \in \{1, \ldots, B\}$) and two versions of non-stationary SK shown in Equations 2.26 and 2.27 (Deutsch, 2006) that utilizes locally varying mean (LVM).

$$\hat{\mathbf{1}}_i^{\text{SK}}(\boldsymbol{u}_0) - p_i = \sum_{\alpha \in N_i(\boldsymbol{u}_0)} \lambda_{\alpha,i} \left[\mathbf{1}_i \left(\boldsymbol{u}_\alpha \right) - p_i \right], \qquad \forall i \in \{1, \ldots, |\mathcal{B}|\} \tag{2.25}$$

where $N(\boldsymbol{u}_0)$ is the set of locations that are utilized for the estimation of location $\boldsymbol{u}_0$.

$$\hat{\mathbf{1}}_i^{\text{LVM}_1}(\boldsymbol{u}_0) - p_i(\boldsymbol{u}_0) = \sum_{\alpha \in N_i(\boldsymbol{u}_0)} \lambda_{\alpha,i} \left[\mathbf{1}_i \left(\boldsymbol{u}_\alpha \right) - p_i(\boldsymbol{u}_\alpha) \right], \qquad \forall i \in \{1, \ldots, B\} \tag{2.26}$$

$$\hat{\mathbf{1}}_i^{\text{LVM}_2}(\boldsymbol{u}_0) - p_i(\boldsymbol{u}_0) = \sum_{\alpha \in N_i(\boldsymbol{u}_0)} \lambda_{\alpha,i} \left[\mathbf{1}_i \left(\boldsymbol{u}_\alpha \right) - p_i(\boldsymbol{u}_0) \right], \qquad \forall i \in \{1, \ldots, B\} \tag{2.27}$$

The difference between the two non-stationary SK is on the utilization of the local proportions. For the LVM_1 (Equation 2.26) estimator, the local proportion on the right side of the equation is taken at each data location, whereas the LVM_2 (Equation 2.27) estimator approximates the local probability at the conditioning data locations by the local probability at the estimation location. The LVM_1 is theoretically correct and the LVM_2 gives more importance to the local hard data (Deutsch, 2006).

2.7.2 Transition Probabilities

Transition probabilities are a measure of spatial variability and are closely related to the indicator variogram (Carle and Fogg, 1996). For a given separation vector $\boldsymbol{h}$, the transition probability between two categories is defined by Equation 2.28 . The transition matrix ($\boldsymbol{T}(\boldsymbol{h})$) is a square matrix of rank B that contains all transition probabilities for the separation vector $\boldsymbol{h}$ (Equation 2.29).

$$\begin{aligned} t_{i,j}(\boldsymbol{h}) =& \mathrm{P}\{\mathbf{1}_j(\boldsymbol{u}+\boldsymbol{h}) = 1 | \mathbf{1}_i(\boldsymbol{u}) = 1\} \\ =& \frac{\mathrm{P}\{\mathbf{1}_j(\boldsymbol{u}+\boldsymbol{h}) = 1 \text{ and } \mathbf{1}_i(\boldsymbol{u}) = 1\}}{\mathrm{P}\{\mathbf{1}_i(\boldsymbol{u}) = 1\}}, \qquad i,j \in \{1, \ldots, B\} \end{aligned} \tag{2.28}$$

$$\boldsymbol{T}(\boldsymbol{h}) = \begin{bmatrix} t_{1,1}(\boldsymbol{h}) & t_{1,2}(\boldsymbol{h}) & \dots & t_{1,B}(\boldsymbol{h}) \\ t_{2,1}(\boldsymbol{h}) & t_{1,2}(\boldsymbol{h}) & \dots & t_{2,B}(\boldsymbol{h}) \\ \vdots & \vdots & \ddots & \vdots \\ t_{B,1}(\boldsymbol{h}) & t_{B,2}(\boldsymbol{h}) & \dots & t_{B,B}(\boldsymbol{h}) \end{bmatrix} \tag{2.29}$$

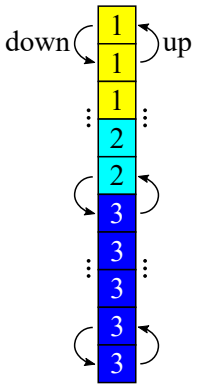
If the bivariate distribution of the indicator variables $\mathbf{1}(\boldsymbol{u})$ and $\mathbf{1}(\boldsymbol{u}+\boldsymbol{h})$ depends only on the separation vector (Equation 2.30) and the categorical proportion is stationary (Equation 2.31), the indicator variogram can be rewritten in terms of transition probabilities (Equation 2.32).

$$\begin{aligned} p_{i,j}\left(\boldsymbol{h}\right) =& \mathrm{P}\left\{\mathbf{1}_i(\boldsymbol{u}+\boldsymbol{h}) = 1 \text{ and } \mathbf{1}_j(\boldsymbol{u}) = 1\right\} \\ =& \mathrm{E}\left\{\mathbf{1}_i\left(\boldsymbol{u}\right)\mathbf{1}_j\left(\boldsymbol{u}+\boldsymbol{h}\right)\right\} \end{aligned}, \qquad i,j \in \{1,\dots,B\} \quad \text{and} \quad \forall \boldsymbol{u} \in \mathcal{A} \tag{2.30}$$

$$p_i = \mathrm{P}\left\{\mathbf{1}_i\left(\boldsymbol{u}\right) = 1\right\} = \mathrm{E}\left\{\mathbf{1}_i\left(\boldsymbol{u}\right)\right\}, \qquad i \in \{1,\dots,B\} \quad \text{and} \quad \forall \boldsymbol{u} \in \mathcal{A} \tag{2.31}$$

$$\gamma_{\mathbf{1}i,j}(\boldsymbol{h}) = p_i\left(t_{i,j}\left(0\right) - \frac{1}{2}\left[t_{i,j}(\boldsymbol{h}) + t_{i,j}(-\boldsymbol{h})\right]\right), \qquad i,j \in \{1,\dots,B\} \tag{2.32}$$

The transition probabilities are often calculated along the drillhole as the data density is higher and the shortest separation lags are well informed. The resulting transition matrices are useful for the definition geological contacts. The transition matrices are often asymmetric ($t_{i,j}(\boldsymbol{h}) \neq t_{i,j}(-\boldsymbol{h})$) and are calculated downwards and upwards along the drilling path (Figure 2.2).



$$\boldsymbol{T}_{\mathrm{d}} = \begin{bmatrix} 0.67 & 0.33 & 0.00 \\ 0.00 & 0.50 & 0.50 \\ 0.00 & 0.00 & 1.00 \end{bmatrix} \qquad \boldsymbol{T}_{\mathrm{u}} = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 \\ 0.00 & 0.20 & 0.80 \end{bmatrix}$$

Figure 2.2: Illustration of the transition probability calculation, upwards and downwards, for a string of categorical data along a drillhole.

For simplicity, the lag vector ($\boldsymbol{h}$) is dropped from the notation of the transition matrix throughout this dissertation. This is justified as the transition matrix is often only calculated for the lag size that corresponds to the composite length and for the upward and downward direction. $\{\boldsymbol{T}_{\mathrm{d}} = (t_{\mathrm{d}i,j}); i,j = 1,\ldots,B\}$ is used for the transition matrix calculated downwards, $\{\boldsymbol{T}_{\mathrm{u}} = (t_{\mathrm{u}i,j}); i,j = 1,\ldots,B\}$ is used for the upward direction and $\boldsymbol{T} = \frac{1}{2}(\boldsymbol{T}_{\mathrm{d}} + \boldsymbol{T}_{\mathrm{u}})$ is used for the combined transition matrix.

2.7.3 Sequential Indicator Simulation

Sequential indicator simulation (SIS) (Deutsch, 2006) is a MCS algorithm that is applied with steps similar to the SGS algorithm. The local conditional distributions are non-parametric and inferred using the IK formalism. The SIS algorithm is summarized below:

(1) define a random path through the model nodes
(2) for each node on the random path:
  (2.1) for each variable from 1 to B:
    (2.1.1) search for conditioning data
    (2.1.2) build and solve kriging system
  (2.2) correct order relations and define the local conditional distribution
  (2.3) sample the conditional distribution
  (2.4) assign sampled value to the model node
  (2.5) calculate the indicator variables at the simulated node and add them to the conditioning data ensemble

SIS is suitable for the simulation of binary categorical variables. Multiple categories are handled through the use of multiple binary indicator transforms. SIS realizations mostly reproduce the proportions of categories and their two-point spatial distribution, however, there is no guarantee of reproducing the ordering between categories specially in a sparsely sampled domain. This makes the technique suitable to nominal categorical variables.

The SIS realizations often show inflated short scale variability resulting in noisy categorical models. Also, categories with low proportions are under-sampled and, consequently, the categories with higher proportions are over-sampled resulting in a mismatch in the global proportion statistics. One solution for both issues is the application of the maximum a posteriori selection (MAPS) technique (Deutsch, 1998, 2006). The main drawbacks

of cleaning SIS realizations are the possible removal of real short scale variability and the requirement of parameters arbitrarily defined by the user to ensure the reproduction of conditioning data (Deutsch, 1998).

2.7.4 Truncated Gaussian Simulation

Truncated Gaussian simulation (TGS) proposed by Matheron et al. (1987) is an alternative for the simulation of categorical variables. This technique assumes that categorical variables are generated by the truncation of an underlying latent variable. The latent variable is a realization of a GRF and the truncation rule defines the ordering and proportion of each category. The truncated pluri-Gaussian simulation (TPGS) proposed by Galli et al. (1994) is an extension of the TGS to use an arbitrary number of GRFs as latent variables. This allows the simulation of more complex spatial arrangements of categorical variables including ordered categorical variables such as alteration zones in lateritic nickel or depositional sequences in petroleum reservoirs. The TPGS aims at reproduction of categorical proportions, two-point spatial correlation and transition probabilities.

2.7.4.1 Truncation Mask

The truncation rule, also referred to as lithotype rule, is an important part of the TPGS methodology as it controls the contacts between categories, their transition and proportions. The definition of the truncation rules is usually based on transition probabilities observed in data and conceptual geological model (Mariethoz et al., 2009).

A number of techniques to define the mapping rule have been developed. Armstrong et al. (2011) propose the partition of the multivariate Gaussian space in parallelepipeds (or rectangular areas in bivariate case) through the use of thresholds. A template should be chosen based on the knowledge regarding the geology (Figure 1.1) and the thresholds are calculated in order to match the proportion of each category. This methodology is useful for a bivariate case with few categories. In this situation, the possible templates can be visualized and a decision can be made if the geology is well understood. The methodology becomes more difficult when more categories are considered and virtually impossible for four or more latent variables.

Xu et al. (2006) proposed a methodology for the definition of the truncation rule based on a binary dynamic contact matrix (DCM) that facilitates the application of TPGS to high-

dimensional cases with more complex geological features. The DCM is specified by the user based on geologic knowledge and it is used to partition the multivariate Gaussian space with rectangles and parallelepipeds similarly to what is done by Armstrong et al. (2011).

Allard et al. (2012) introduced the assignation diagram which automatically builds the truncation rule for the truncated biGaussian case allowing for complex contacts. The truncation rule is defined using kernel regression based on auxiliary variables such as seismic attributes and lithofacies information available at drilling locations. Synthetic data may be used in the absence of auxiliary data.

Sadeghi and Boisvert (2012) use simulated annealing to optimize the truncation rule. The objective function is the minimization of the mismatch between the transition probabilities calculated from simulated realizations and the experimental transition probabilities calculated from data. The use of simulated annealing allows for flexible truncation rules which have the potential to reproduce complex features.

Deutsch and Deutsch (2014) use a multidimensional scaling (MDS) methodology to define complex truncation rules with the focus on matching the experimental transition probabilities. The final truncation rule partitions the Gaussian space by nearest distance to B control points, where B is the number of categories. The methodology simplifies the task of defining the truncation rule and is suitable for any number of Gaussian variables up to $B - 1$ and any number of categories.

Astrakova et al. (2015) uses a Bayesian maximum entropy approach coupled to simulated annealing algorithm to optimize the truncation rule of the truncated biGaussian model to match categorical bivariate unit-lag probabilities. The resultant partition of the Gaussian space is also defined by nearest neighbor (Voronoi tessellation) similar to the Deutsch and Deutsch (2014) approach, however, multiple control points for each category are allowed.

Madani and Emery (2015) are the first to propose a hierarchical approach to tackle the problem of simulating many categorical variables using multiple GRFs. The approach proposed in Madani and Emery (2015) is based on the chronological ordering of the geological units and uses one Gaussian variable to separate each category from the remaining resulting in a model with $B - 1$ latent variables (where B is the number of categories). This truncation rule can be seen as a simple binary tree where each node represents a Gaus-

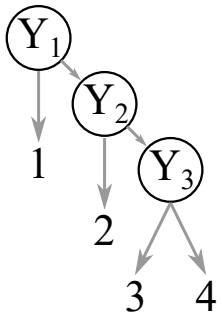


Figure 2.3: Example of linked list with three Gaussian variables Y_1, Y_2 and Y_3 as nodes and four leafs that represents categories.

sian variable and the threshold separates two branches. Each node generates one leaf and one child, with exception of the last node that generates two leafs. This specific case of binary tree can be seen also as a linked list. The root of the tree represents the most recent geological units and the deepest nodes represent the older units.

2.7.4.2 Non-stationarity

The stationarity of a categorical variable is often described in terms of proportions. In truncated pluriGaussian methods, the categorical proportions are defined by the probability density of the region of the continuous space assigned to each category. In order to account for non-stationarity, these regions must change to match the local proportions. These regions are defined by the truncation rule.

In most applications the regions of the truncation rule are defined by thresholds that are orthogonal relative to the continuous variables. This is the case of the 2D and 3D rock-type rules presented in Armstrong et al. (2011) and Emery (2007). For these cases the the thresholds can be locally changed to match the local proportions. Other truncation rules such as the nearest neighbor segmentation of the Gaussian space in Astrakova et al. (2015) and Deutsch and Deutsch (2014) are complex and requires the integration of the multivariate Gaussian space.

2.7.4.3 Mapping Spatial Continuity

The TPGS approach simulates categorical variables indirectly by first simulating Gaussian latent variables and then performing the truncation of these variables to obtain the categorical realizations. One of the objectives of simulation is to reproduce the variogram of observed data. The variogram of the latent variables are unknown as the latent variables

are not observed, however, the variogram of the categorical data can be experimentally calculated from data. The variogram used in the simulation of the latent variables should be defined in order to reproduce the spatial structure of the categorical variables after truncation.

Kyriakidis et al. (1999) uses numerical integration of the bivariate Gaussian distribution to define the variogram of the latent variable from indicator variogram for TPG technique with two categories.

Armstrong et al. (2011) defines the theoretical link between the variogram of Gaussian latent variables and the indicator variogram of the categorical variables for the rectangular (parallelepipeds in higher dimension) partition of the Gaussian space. An iterative approach is used to define the variogram for latent variables that better match the indicator variogram of categories after truncation.

Zagayevskiy and Deutsch (2015) propose a methodology for numerical derivation of latent variables variogram based on MCS. The methodology optimizes a series of lag distances that discretizes the correlation range of the categorical variables independently. The spatial correlation of the Gaussian variables is optimized for each lag distance by simulating thousands of pairs and calculating the average indicator correlation after truncation. The spatial correlation of the latent variables is adjusted until the mismatch is deemed small enough. After defining the optimum correlation for each lag distance for each latent variable, the points are fitted with stable (Chilès and Delfiner, 1999; Zagayevskiy and Deutsch, 2015) variogram models. This is a flexible technique that is suitable to any truncation rule with any number of Gaussian latent variables.

2.7.4.4 Imputation of Latent Variables

The truncated Gaussian techniques assume that the categorical variable is generated by the truncation of underlying latent variables. The latent variables, however, are not observed. In order to condition the models to categorical observations the latent variables must be defined at data locations. The multidimensional truncated Gaussian distribution is complex and cannot be directly sampled, however, marginal conditional distributions can be easily defined. In such situations, the Gibbs sampler algorithm is the best option available (Arroyo et al., 2012; Astrakova et al., 2015; Emery et al., 2014; Galli et al., 1994; Lantuéjoul and Desassis, 2012). The application of the Gibbs sampler depends on the truncation rule

and often suffers from convergence issues.

Another alternative is to assign a valid combination of latent variable that would result in the observed category after applying the truncation rule. The values are often defined by the centroid of each categorical class with respect to the truncation rule (Rossi and Deutsch, 2014). This, however, results in conditioning data that do not have the correct spatial variability.

An important aspect of the data assignment for truncated Gaussian methods that is often overlooked is the uncertainty in the latent variables. For the same combination of categorical observations there are multiple valid combinations of the latent variables that result in the same categorical observation. This uncertainty in the latent variables should be transferred to the final categorical realizations.

2.7.5 Other Modeling Techniques

In addition to the reproduction of proportions and two-point spatial correlation features observed in the data, there is also an interest in the reproduction of pre-conceived morphology. Those morphologic features could include geological structures such as meandering channels, vein systems and other high order continuity. Techniques such as multiple point statistics (MPS) (Guardiano and Srivastava, 1993; Strebelle, 2002) and object based models (Haldorsen et al., 1984; Hassanpour, 2013) are developed to reproduce conceptual geological shapes. These techniques are out of the scope of this dissertation.

2.8 Conclusion

The literature review undertaken in this chapter provides the background required for the further reading of this thesis. Additional literature exist that are not covered here, however, the provided review synthesizes the the most relevant research available that are important to this research subject.

Chapter 3

Hierarchical Truncated pluri-Gaussian

This chapter starts with the mathematical notation for the truncated Gaussian methods, followed by the identification of the limitations with the current application. The theory and procedures for the application of hierarchical truncated pluri-Gaussian (HTPG) are developed in sequence.

3.1 Mathematical Notation and Definitions

The mathematical notation and definitions for the univariate case of HTPG is shown here. This notation is further extended to the multivariate case in Chapter 7.

Consider a categorical random function (RF) $\{X(\boldsymbol{u});\ \forall \boldsymbol{u} \in \mathcal{A}\}$ that can take any value from finite set $\mathcal{B}$ of possible categories. Also, consider a set of latent variables to be represented by the Gaussian random function (GRF) $\{\boldsymbol{Y}(\boldsymbol{u}) = (Y_1(\boldsymbol{u}), \ldots, Y_K(\boldsymbol{u}));\ \forall \boldsymbol{u} \in \mathcal{A}\}$. Finally, consider the truncation rule to be represented by $\mathcal{M}_\theta$ which defines the mapping $\{\mathcal{M}_\theta : \mathbb{R}^K \mapsto \mathcal{B}\}$ such that $\{\mathcal{M}_\theta(\boldsymbol{Y}(\boldsymbol{u})) = X(\boldsymbol{u});\ \forall \boldsymbol{u} \in \mathcal{A}\}$, where θ is the set of parameters that define the truncation rule.

Let the cardinality of the categorical set be defined by $B = |\mathcal{B}|$ and $\{b_i;\ i = 1, \ldots, B\}$ be the categories within the set $\mathcal{B}$. Let $\{\mathcal{C}_i;\ i = 1, \ldots, B\}$ represent the multivariate space where $\{X(\boldsymbol{u}) = b_i \Leftrightarrow \boldsymbol{Y}(\boldsymbol{u}) \in \mathcal{C}_i;\ i = 1, \ldots, B;\ \forall \boldsymbol{u} \in \mathcal{A}\}$.

3.2 Limitation of Current Practice of Truncated Gaussian Methods

The truncated Gaussian methods share four main principles: (1) the truncation rule; (2) the mapping of spatial variability; (3) the assignment of Gaussian data; and (4) the truncation of simulated Gaussian variables. The truncation rule defines the mapping between the categorical and continuous space. The mapping is the most important feature of the

truncated Gaussian methods as it impacts the features observed in the resulting geological models and influences the application of the other principles. Any limitations imposed in the truncation rule will result in limitations on the applicability of the truncated Gaussian method.

The truncated Gaussian simulation (TGS) and truncated pluri-Gaussian simulation (TPGS) were developed with the intention of allowing geological knowledge to be introduced into the modeling workflow by means of the truncation rule. For instance, if a categorical variable representing a simple layered deposit is being modeled and the categories transition in a ordered fashion, a single Gaussian variable is enough to represent the geology (Figure 3.1)

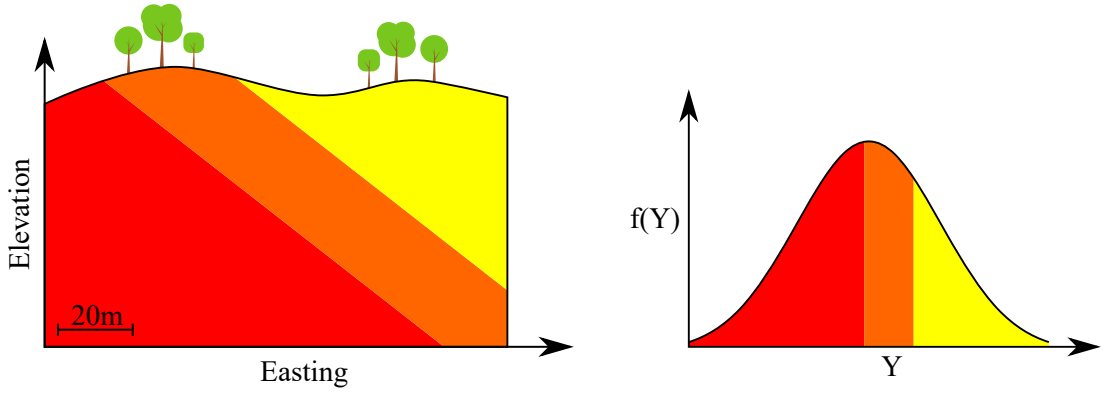


Figure 3.1: Illustrative example of a simple layered structure that can be represented with the truncation of a single Gaussian variable.

Additional Gaussian variables and complex truncation rules are required with increasing geological complexity. For instance, if an intrusion is added to the illustrative example shown in Figure 3.1, an extra Gaussian variable is required to represent the discordant structure (Figure 3.2). In this case, the representation is simple enough and easily visualized with the conventional truncation mask. The first Gaussian variable Y_1 is responsible to separate the intrusion from the layered categories and will control the spatial structure of the intrusion. The second Gaussian variable Y_2 controls the spatial structure of the three layered categories. The spatial structure of the two sets of categories is not mixed because the thresholds are orthogonal in relation to the Gaussian variables.

Most applications of TPGS are restricted to the utilization of two Gaussian variables as it still allows straightforward link between the truncation mask representation and the geological structure. If additional geological complexity is added, more Gaussian variables

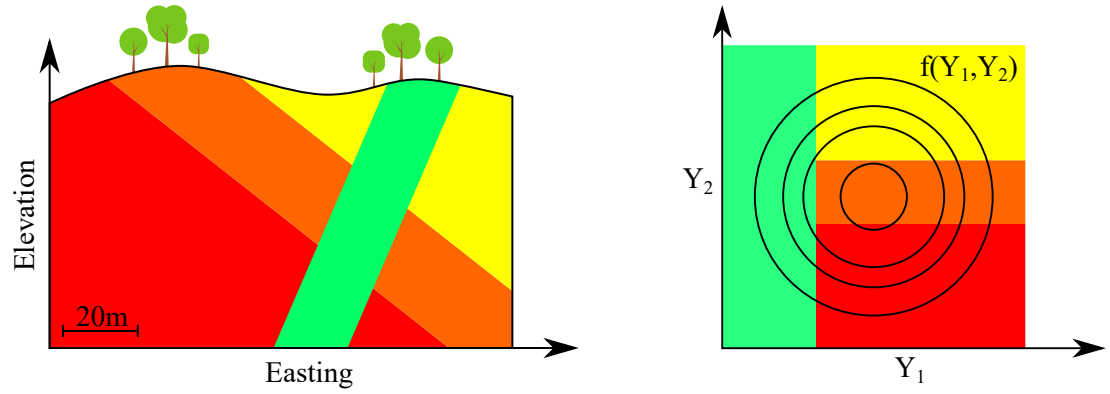


Figure 3.2: Illustrative example of a layered structure cut by an intrusion. The structure can be represented with the truncation of two Gaussian variables.

are required to represent the geological setting. In Figure 3.3, an additional discordant structure is added. An erosional surface is included with additional depositional layers on top of it.

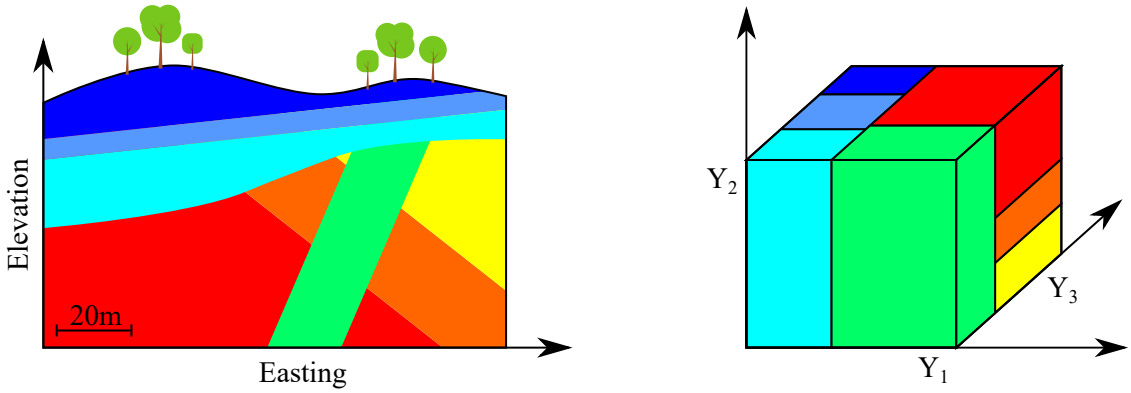


Figure 3.3: Illustrative example of two layered structures separated by a erosional surface. The layers below the erosional surface are cut by a intrusion. A 3D truncation mask that attempts to represent the geological setting is also shown.

At first glance, one would think that one additional Gaussian variable would suffice to represent the geology and that a conventional truncation mask could still be created as shown in Figure 3.3. The Gaussian variable Y_1 controls the transition between the structures above the erosional surface and the structures below ensuring that they do not cut through that boundary. The Gaussian variable Y_2 controls the transition between the layers below the erosional surface. The Gaussian variable Y_3, however, controls the transition between the layers above the erosional surface and the transition between the intrusion and the layers below the erosion. Because these two structures have different spatial structure (direction and shape), the truncation mask shown in Figure 3.3 does not represent the

geological setting that is shown.

Another alternative truncation mask is shown in Figure 3.4. For this example, Y_1 controls the transition between the layers on top of erosional surface. Y_2 controls the transition between the layers below the erosional surface. The Gaussian Y_3 controls the spatial structure of the intrusion. This mask separates each geological unit's spatial variability utilizing an independent Gaussian variable, however, this setting still allows for the intrusion to cut through the layers above the erosional surface. At least four Gaussian variables are required to properly account for all the shapes and boundary constraints shown in this example. An easy visualization and geological interpretation of the conventional truncation mask is not possible for this example.

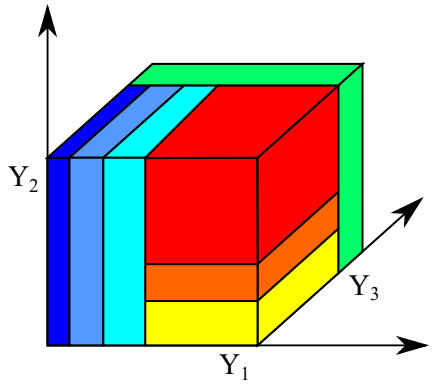


Figure 3.4: Alternative 3D truncation mask for the geological setting shown in Figure 3.3.

There are alternatives to the conventional truncation rules available for utilization with higher dimensions. The multidimensional scaling (MDS) generated masks proposed by Deutsch and Deutsch (2014) maps the multivariate space by Voronoi decomposition. The technique is completely data driven and removes the opportunity to add geological expert knowledge to the truncation rule. In addition, the boundaries of the truncation regions are not completely orthogonal to the variable's axes, resulting in the mixing of spatial structures and making it more difficult to map the spatial structure of the latent variables to the categorical space.

Another approach is proposed by Madani and Emery (2015). The truncation rule is represented by a simple binary tree with a linked list structure where each node is used to segregate one category. The HTPG can be seen as a generalization of this technique where more complex tree structures are allowed depending on the geological structure observed. The technique proposed in Madani and Emery (2015) represents a extreme case in which

the most spatial freedom is given to each category with minimum shared spatial structure.

3.3 Proposed Hierarchical Approach

Most applications of TPGS are restricted to two dimensional truncation rules mostly because the geological interpretation of the conventional truncation rules in higher dimension is difficult. The increased difficulty of mapping the spatial variability of the categorical variable to the continuous space is often used to justify the restriction to the bivariate case (Armstrong et al., 2011). This may be true if one is attempting to analytically resolve the mapping, however, numerical methods can be used to define the mapping for any number of latent variables. In fact, the use of additional dimensions often eases the mapping by providing extra degrees of freedom to the possible spatial structures being mapped.

Methods such as the one proposed by Deutsch and Deutsch (2014) allow for multiple latent variables, however, the definition of the truncation rule is completely data driven with little room for geological understanding and interpretations. The idea of using hierarchical rules with larger number of latent variables to model multiple categories while allowing for geological interpretation is very promising. The binary tree structure proposed by Madani and Emery (2015), however, may not be the best to represent all possible geological structures. The hierarchical approach developed in this dissertation can be seen as a generalization of this technique in which more complex tree structures can be used to describe the contact relationships between geological domains. This allows for more flexibility on what can be represented.

A gridded 2D conceptual model is built to illustrate the HTPG (Figure 3.5). The conceptual model shows the same geological complexity of the illustration in Figure 3.3 that is used to highlight the limitation of the current practice. The template represents the geology resultant of a set of geological events. Categories 5, 6 and 7 were deposited in ordered layers which were tilted from its original horizontal orientation. The category 4 can be interpreted as a dike that came later cutting through categories 5, 6 and 7. The resultant sequence was later eroded and categories 1, 2 and 3 were deposited in layers on top of the erosional surface. In this example the categorical variable $X(\boldsymbol{u})$ can take any category on the set $\mathcal{B} = [1, 2, 3, 4, 5, 6, 7]$ and $B = 7$.

The key contribution of the HTPG approach is the definition of the truncation rule,

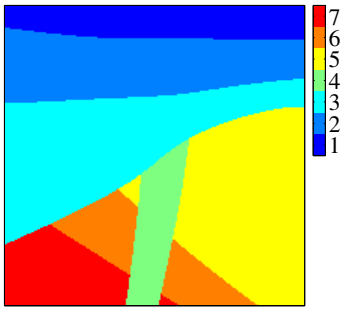


Figure 3.5: 2D conceptual model used to illustrate the HTPG methodology. Categories 5, 6 and 7 were deposited in ordered layers which were tilted from its original horizontal orientation. The category 4 can be interpreted as a dike that cuts through the categories 5, 6 and 7. This sequence was eroded and categories 1, 2 and 3 were deposited in layers on top of the erosional surface.

however, the truncation rule has an impact on all subsequent steps. In fact, the HTPG approach does not only refer to the truncation rule, but to all steps required in its application as defined in this section.

3.3.1 The Hierarchical Truncation Rule

The truncation rule ($\mathcal{M}_\theta$) in HTPG is defined by a decision tree like structure. Every parent (non-leaf) node on the tree structure represents a Gaussian latent variable. The parent nodes are also where the thresholds are applied. The leafs of the tree are the resulting categories. There can only be one leaf per category. This entails that there are only $B-1$ thresholds and that the number of Gaussian variables must be less or equal to the number of thresholds ($K \leq B-1$).

The geological setting shown in Figure 3.5 is complex enough to make the conventional application of truncation masks unworkable, however, the definition of a truncation rule is quite straightforward with the proposed tree structure. The geological setting can be reconstructed with a hierarchical set of truncation rules defined by the tree structure shown in Figure 3.6.

Four Gaussian variables ($K = 4$) are used to represent this geological setting. It was shown that three variables were not enough for this representation. The first Gaussian variable (Y_1) defines the erosional surface that separates the categories 1, 2 and 3 from the categories 4, 5, 6 and 7. The Gaussian variable Y_2 separates the discordant category 4 from the layered categories 5, 6 and 7. The Gaussian variable Y_3 separates the layered categories 1, 2 and 3. Finally the Gaussian variable Y_4 separates the layered categories 5, 6, and 7.

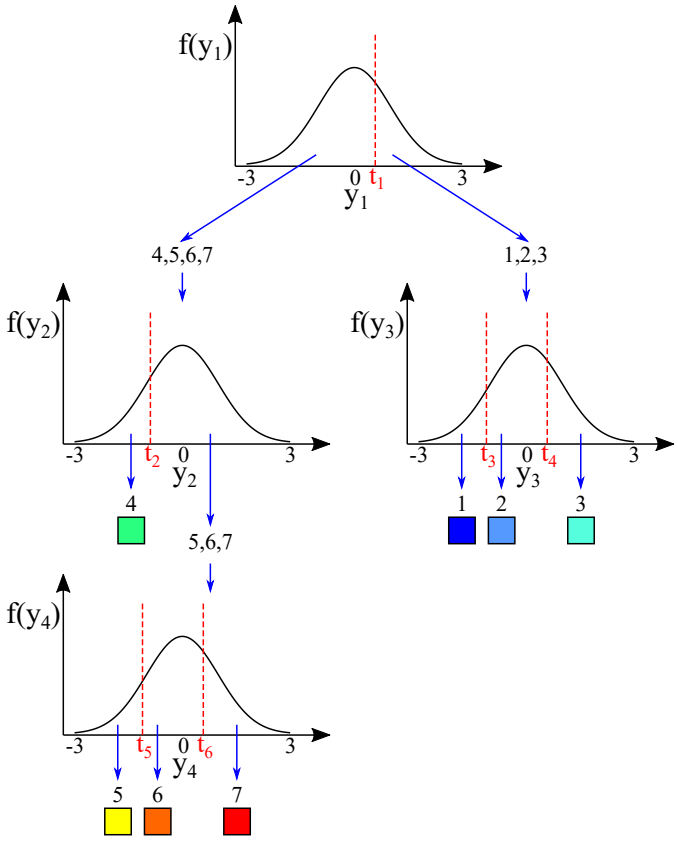


Figure 3.6: Hierarchical set of truncation that describes the geological setting shown in Figure 3.5.

This simple example illustrates how one can define hierarchical truncation rules using multiple Gaussian variables to define the transitions and ordering between different categories. This simple exercise of defining geologically sound truncation rules for this many latent variables and categories would have been much more difficult with conventional practices. The hierarchical procedure can be easily used to separate (1) sets of categories that do not belong together such as sets of rocks types separated by erosional surface, (2) cross cutting (discordant categories) such as intrusive rocks, (3) ordered categories such as sedimentary sequences and (4) background categories.

The hierarchical truncation rule could be built using geological understanding of the domain of interest, however, it is useful to have input information calculated from data to assist the decision. An approach based on MDS, transition probabilities and minimum spanning tree (MST) (Prim, 1957) can be used for quick visualization of categorical transitions and relationship based on transition probabilities calculated from drillholes. More details on the definition of the truncation rule in HTPG is given in Chapter 4.

3.3.2 Thresholds

The truncation thresholds control the proportion of categories on simulated models. The categorical proportion is the probability of observing a certain category at a given location (Equation 3.1).

$$p_i(\boldsymbol{u}) = \mathrm{E}\left\{\mathbf{1}_i(\boldsymbol{u})\right\} \qquad \forall i \in \{1,\ldots,B\} \quad \text{and} \quad \forall \boldsymbol{u} \in \mathcal{A} \tag{3.1}$$

The categorical variable is stationary if the categorical proportions do not change with the location $\{p_i(\boldsymbol{u}) = p_i;\ i = 1,\ldots,B;\ \forall \boldsymbol{u} \in \mathcal{A}\}$. The multivariate Gaussian space is divided into regions $\{\mathcal{C}_i; i = 1,\ldots,B\}$ by the truncation rule. If the latent variables are independent standard Gaussian variables and the categorical proportions are stationary over the domain, the proportion of a category b_i can be calculated by integrating the Gaussian PDF over the respective region (Equation 3.2).

$$p_i = \int_{\mathcal{C}_i} \phi(\boldsymbol{y}) dy \qquad \forall i \in \{1,\ldots,B\} \tag{3.2}$$

The regions $\{\mathcal{C}_i;\ i = 1,\ldots,B\}$ are defined by axis-parallel hyper-rectangles in HTPG. The categorical outcome is only possible if all required inequality conditions are satisfied. For a given category b_i the region delineated by the thresholds is defined by Equation 3.3.

$$\begin{aligned} \mathcal{C}_i = \Big\{ \boldsymbol{y}(\boldsymbol{u}) \in \mathbb{R}^K | & t_{\min}^{(i,1)} \leq y_1(\boldsymbol{u}) \leq t_{\max}^{(i,1)} \wedge t_{\min}^{(i,2)} \leq y_2(\boldsymbol{u}) \leq t_{\max}^{(i,2)} \wedge \\ & \ldots \wedge t_{\min}^{(i,K)} \leq y_K(\boldsymbol{u}) \leq t_{\max}^{(i,K)} \Big\}, \qquad \forall i \in \{1,\ldots,B\} \text{ and } \boldsymbol{u} \in \mathcal{A} \end{aligned} \tag{3.3}$$

where the bounding thresholds $t_{\min}^{(i,j)}$ and $t_{\max}^{(i,j)}$ can take one of the values in $\{-\infty, t_1, \ldots, t_{B-1}, +\infty\}$ while $t_{\min}^{(i,j)} \leq t_{\max}^{(i,j)}$.

Equation 3.2 can be rewritten to that shown in Equation 3.4, using the relationship from Equation 3.3. If a node or Gaussian variable is irrelevant for the definition of a category, the lower and upper thresholds are set to $-\infty$ and $+\infty$, respectively.

$$p_i = \prod_{j=1}^{K} \left[\Phi\left(y_j \leq t_{\max}^{(i,j)}\right) - \Phi\left(y_j \leq t_{\min}^{(i,j)}\right) \right] \qquad \forall i \in \{1,\ldots,B\} \tag{3.4}$$

The tree structure defined in Section 3.3.1 ensures that there is only one possible set

of bounding thresholds $\{t_i;\ i = 1, \ldots, B-1\}$ that defines a closed set of categorical proportions $\left(\sum_{i=1}^{B} p_i = 1.0\right)$. It also ensures that a threshold can be calculated utilizing the proportions of the categories relevant to the node where the threshold is applied (Equation 3.5).

$$t_j = \Phi^{-1}\left(\frac{\sum\limits_{i \in \mathcal{B}_{k,j}} p_i}{\sum\limits_{i \in \mathcal{B}_k} p_i}\right), \qquad \forall j \in \{1, \ldots, B-1\} \tag{3.5}$$

where $\mathcal{B}_k$ is a subset of $\mathcal{B}$ with all the categories that are relevant to the node k where the threshold t_j is applied. $\mathcal{B}_{k,j}$ is a subset of $\mathcal{B}_k$ with all the categories that are defined below the threshold t_j.

3.3.3 Non-stationarity

Categorical variables often are non-stationary. Local proportions are used to account for aspects of non-stationarity. For example, a model that resembles the geology shown in Figure 3.5 using the truncation rule shown in Figure 3.6 cannot be generated without the use of local proportions unless there are enough conditioning data to enforce the observed features. The same truncation rule presented in Figure 3.6 will also generate models such as the one shown in Figure 3.7.

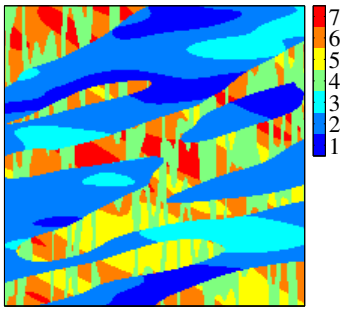


Figure 3.7: 2D model generated without accounting for non-stationarity.

There are two alternatives to account for local proportions with the proposed hierarchical approach. The first is to simulate a stationary GRF and use locally varying thresholds adjusted accordingly with the local proportions. In this case, the non-stationary version of Equation 3.5 is written as in Equation 3.6. The second option is to use fixed thresholds and a non-stationary GRF. In this case, the non-stationary version of Equation 3.5 is written as in Equation 3.7.

$$t_j(\boldsymbol{u}) = \Phi^{-1}\left(\frac{\sum\limits_{i\in\mathcal{B}_{k,j}} p_i(\boldsymbol{u})}{\sum\limits_{i\in\mathcal{B}_k} p_i(\boldsymbol{u})}\right), \qquad \forall j \in \{1,\dots,B-1\} \tag{3.6}$$

$$t_j = \Phi^{-1}\left(\frac{\sum\limits_{i\in\mathcal{B}_{k,j}} p_i(\boldsymbol{u})}{\sum\limits_{i\in\mathcal{B}_k} p_i(\boldsymbol{u})}; \boldsymbol{\mu}(\boldsymbol{u}); \boldsymbol{\Sigma}(\boldsymbol{u})\right), \qquad \forall j \in \{1,\dots,B-1\} \tag{3.7}$$

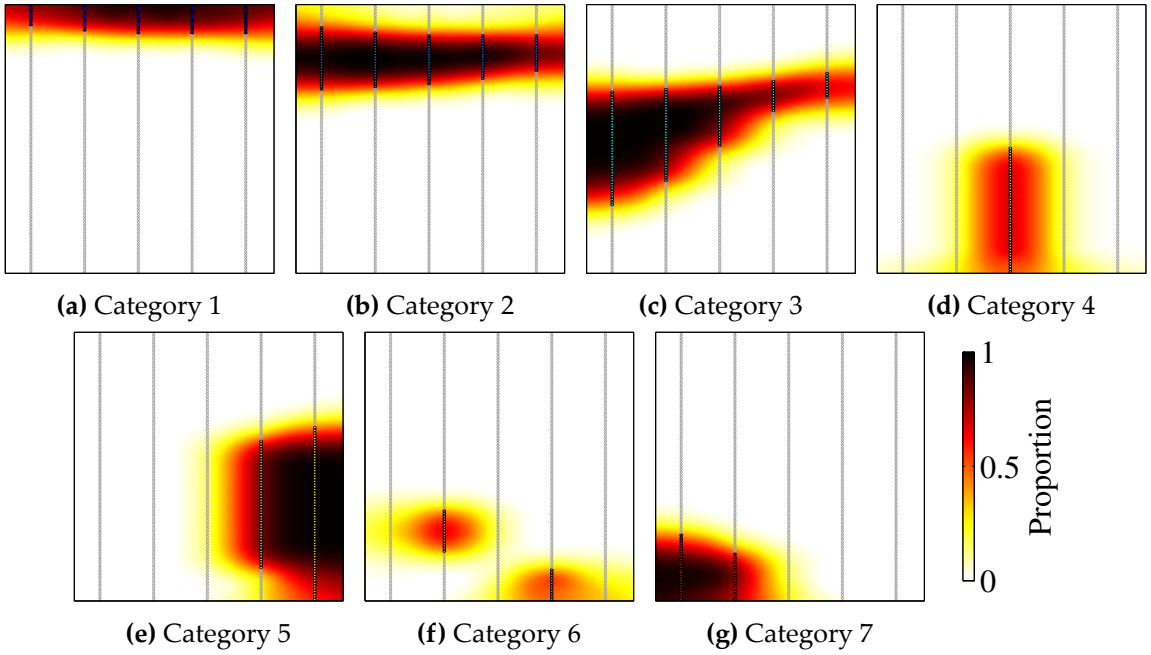
where $\boldsymbol{\mu}(\boldsymbol{u})$ and $\boldsymbol{\Sigma}(\boldsymbol{u})$ are the mean vector and covariance matrix parameterizing the Gaussian distribution. $\boldsymbol{\Sigma}(\boldsymbol{u})$ is a diagonal matrix as the Gaussian variables are independent from each other.

The first option (Equation 3.6) is the simplest and most flexible. It allows for any possible configuration of truncation rules. The second approach (Equation 3.7) is restrictive as it imposes the constraint of using at most two thresholds per Gaussian variable. In this case, only the mean and variance can be used to adjust for the local proportions. There are not enough degrees of freedom to guarantee reproduction of proportions with more than two fixed thresholds per latent variable. It also requires an inversion approach for the definition of the local mean vector and variances that matches the global thresholds for each set of local proportions.

The conceptual model is sampled generating five equally spaced strings of data that emulates drillholes. This data is used to illustrate the definition of the parameters to account for non-stationarity in HTPG. Local proportions are calculated from the samples and shown in Figure 3.8. The local proportions can be used to either adjust the thresholds locally (Figure 3.9) or define locally varying mean and variance (Figure 3.10) for the underlying Gaussian variables.

3.3.4 Mapping Spatial Structure

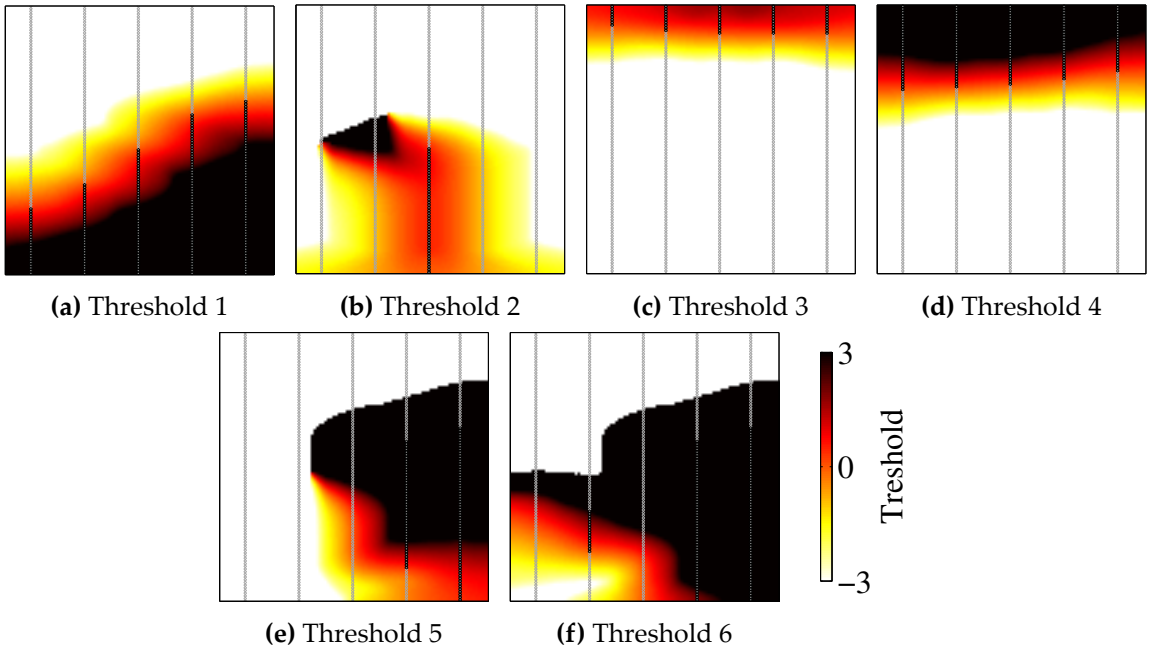
The truncated Gaussian techniques, HTPG included, require the modeling of GRFs to serve as underlying latent variables for the definition of categorical models. This requires the definition of the spatial correlation model for the Gaussian latent variables. In practice, only the indicator variograms of the cateogorical variable can be calculated. The matching spatial structure in continuous space has to be defined to ensure the reproduction of the categorical spatial continuity. The indicator variogram cannot be used directly for the latent variables (Kyriakidis et al., 1999; Matheron, 1989).



(a) Category 1 **(b)** Category 2 **(c)** Category 3 **(d)** Category 4

(e) Category 5 **(f)** Category 6 **(g)** Category 7

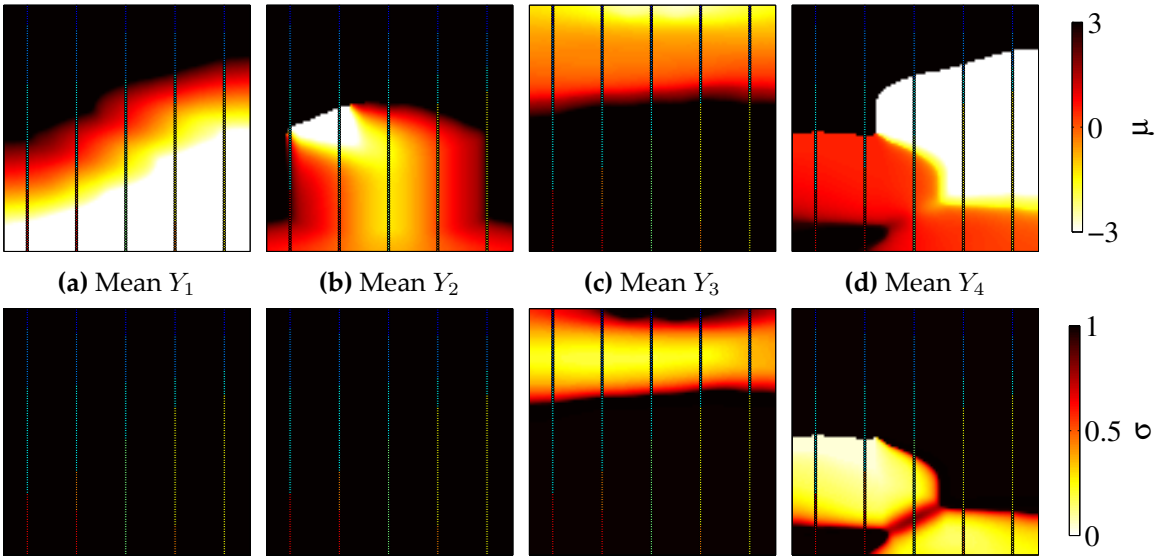
Figure 3.8: Local proportion calculated from sampled data for each category for the 2D example.



(a) Threshold 1 **(b)** Threshold 2 **(c)** Threshold 3 **(d)** Threshold 4

(e) Threshold 5 **(f)** Threshold 6

Figure 3.9: Local threshold adjusted to the local proportion of the categories for the 2D example.



(a) Mean Y_1 **(b)** Mean Y_2 **(c)** Mean Y_3 **(d)** Mean Y_4

(e) Std. Dev. Y_1 **(f)** Std. Dev. Y_2 **(g)** Std. Dev. Y_3 **(h)** Std. Dev. Y_4

Figure 3.10: Local mean and standard deviation for Gaussian variables for the 2D example. Note that the Gaussian variables Y_1 and Y_2 do not require any change in standard deviation. These two variables are truncated by one threshold only. The other two variables (Y_3 and Y_4) are truncated by two thresholds and require varying standard deviation to match the local proportions.

The relationship between the non-centered covariance of the indicator variable and the latent variables is shown in Equation 3.8. A flexible analytical solution for Equation 3.8 for any truncation rule is impossible as there is no closed form solution for these integrals for every possible configuration of regions $\{\mathcal{C}_i; i = 1, \dots, B\}$. A number of numerical solutions have been proposed depending on the application (Armstrong et al., 2011; Kyriakidis et al., 1999; Zagayevskiy and Deutsch, 2015).

$$\begin{aligned} C_{\mathbf{1}i,j}(\boldsymbol{h}) &= \mathrm{E}\left\{\mathbf{1}_i(\boldsymbol{u})\,\mathbf{1}_j(\boldsymbol{u}+\boldsymbol{h})\right\} \\ &= \mathrm{E}\left\{\mathbf{1}_{\boldsymbol{Y}(\boldsymbol{u})\in\mathcal{C}_i}\mathbf{1}_{\boldsymbol{Y}(\boldsymbol{u}+\boldsymbol{h})\in\mathcal{C}_j}\right\} \qquad i,j\in\{1,\dots,B\} \\ &= \int_{\mathcal{C}_i}\int_{\mathcal{C}_j}\phi_h(\boldsymbol{u},\boldsymbol{v})\,d\boldsymbol{u}d\boldsymbol{v} \end{aligned} \tag{3.8}$$

where $\phi_h(\boldsymbol{u},\boldsymbol{v})$ is the $2K$-variate Gaussian density of the vector $(\boldsymbol{y}(\boldsymbol{u}),\boldsymbol{y}(\boldsymbol{u}+\boldsymbol{h}))$, so each integral is over $\mathbb{R}^K$.

The numerical approach proposed by Zagayevskiy and Deutsch (2015) is flexible and can be applied to any type of truncation rule. The technique is adapted to HTPG and further developed to enhance its computational efficiency and practicality. The methodology for

mapping the spatial structure of the categorical data to the continuous space in HTPG is developed in the following paragraphs.

3.3.4.1 Numerical Derivation

The goal of the numerical derivation is to define the variogram model of the Gaussian latent variables that generates realizations of the categorical variable with indicator variograms that are as close to the modeled indicator variograms as possible. This is achieved by an inversion algorithm that utilizes a Monte-Carlo simulation (MCS) framework coupled with a line search optimization. The inversion is illustrated in Figure 3.11. Each lag distance and direction are optimized independently and the resulting points are fitted with valid variogram models. The final state of the numerical derivation is illustrated in Figure 3.12. If the optimized points are fitted well, the mapped points in the categorical space can be seem as the expected variogram reproduction of the HTPG.

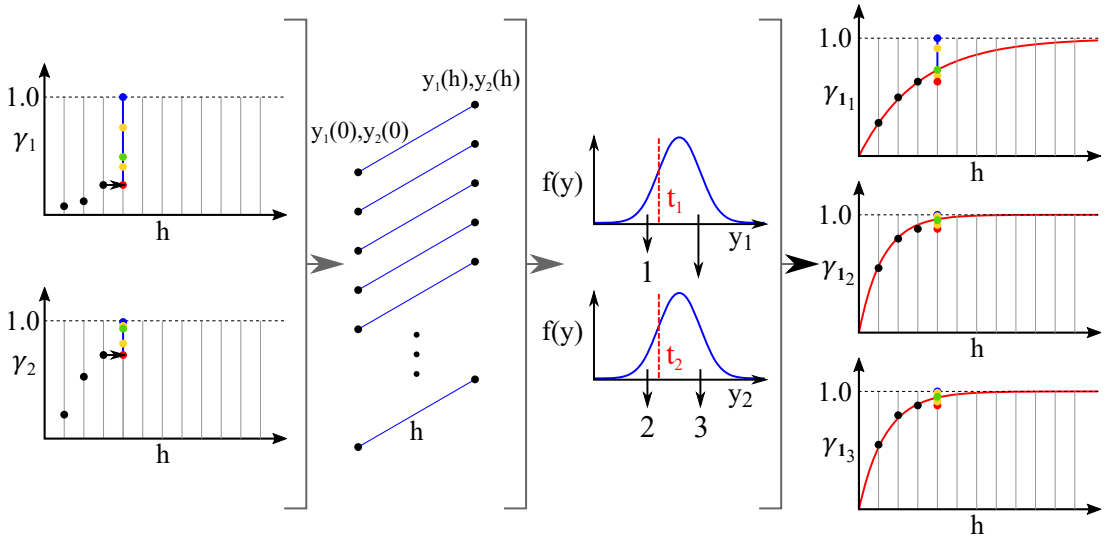


Figure 3.11: Illustration of the MCS based inversion algorithm being applied to the fourth node of the lag discretization. The variograms of the latent variables (left side) are unknown. The first three nodes have been defined before the illustrated state. The lower limit for the line search is the optimized correlation for the previous lag distance. The upper limit is always 1.0. The yellow markers show the iteration points and the green is the final optimum. Each node have its counterpart on the categorical indicator variograms shown on the right side. The red lines are the reference indicator variogram models. The mismatch between the nodes and this line is minimized.

Consider an arbitrary truncation rule represented by M_θ that defines the mapping between the continuous and categorical space $\{\mathcal{M}_\theta : \mathbb{R}^K \mapsto \mathcal{B}\}$. Let $\{\gamma_{\mathbf{1}i}(\boldsymbol{h})\,;\, 1 = 1, \ldots, B\}$ be the modeled direct indicator variograms of the categories in $\mathcal{B}$, and let $\{\gamma_i(\boldsymbol{h})\,;\; 1, \ldots, K\}$ be the variograms of the latent variables. In HTPG, the latent variables are independent

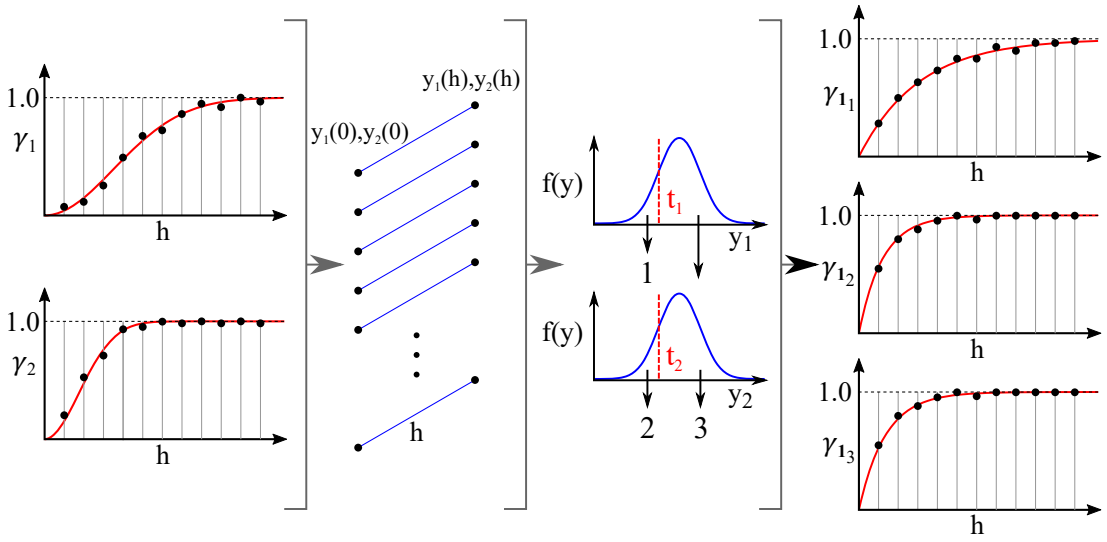


Figure 3.12: Illustration of the final state of the numerical derivation. After all nodes are optimized, the variogram of the latent variables (left side) are fitted with valid variogram models (red lines). The nodes on the right side indicates the expected reproduction of the reference indicator variograms (red lines).

standard Gaussian variables $\left(\boldsymbol{Y}(\boldsymbol{u}) \sim \phi(\boldsymbol{y}; \boldsymbol{0}, \boldsymbol{I})\,;\, \boldsymbol{y} \in \mathbb{R}^K\right)$. The correlation between two points separated by the lag $\boldsymbol{h}$ is calculated by Equation 3.9.

$$\rho_i(\boldsymbol{h}) = 1.0 - \gamma_i(\boldsymbol{h}) \qquad \forall i = 1, \ldots, K \tag{3.9}$$

The inversion approach requires the definition of a link between the correlation $\rho(\boldsymbol{h})$ and the indicator variogram $\gamma_{\boldsymbol{1}}(\boldsymbol{h})$. This is achieved through MCS. Two sets of size $K \times m$ containing m realizations of independent random vectors, $\boldsymbol{z}_{\mathrm{A}} = \left(\boldsymbol{z}_{\mathrm{A}_1}^{\top}, \ldots, \boldsymbol{z}_{\mathrm{A}_m}^{\top}\right)$ and $\boldsymbol{z}_{\mathrm{B}} = \left(\boldsymbol{z}_{\mathrm{B}_1}^{\top}, \ldots, \boldsymbol{z}_{\mathrm{B}_m}^{\top}\right)$, are sampled from the standard Gaussian distribution $(\boldsymbol{z}_{\mathrm{A}}, \boldsymbol{z}_{\mathrm{B}} \sim \phi(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I})$; $\boldsymbol{z} \in \mathbb{R}^K\big)$. These sets are used multiple times throughout the algorithm and can be defined and stored beforehand to save computational time.

The inversion requires the generation of correlated pairs of vectors separated by the lag distance. A set of m realizations of correlated pairs $\boldsymbol{y}(0)$ and $\boldsymbol{y}(\boldsymbol{h})$ are generated using Equation 3.10

$$\begin{aligned} y_{i,j}(0) &= z_{\mathrm{A}_{i,j}} \\ y_{i,j}(\boldsymbol{h}) &= \rho_i(\boldsymbol{h}) \times z_{\mathrm{A}_{i,j}} + \sqrt{1 - \rho_i(\boldsymbol{h})^2} \times z_{\mathrm{B}_{i,j}} \end{aligned} \quad , \quad \forall i \in \{1, \ldots, K\} \quad \text{and} \quad \forall j \in \{1, \ldots, m\} \tag{3.10}$$

The correlated samples are mapped to categorical space using the truncation rule (Equation 3.11). The categorical samples are converted to indicators and the indicator variograms are calculated for each one of the B categories using the m simulated pairs (Equation 3.12).

$$x_j(0) = \mathcal{M}_\theta(\boldsymbol{y}_j(0)) \quad \text{and} \quad x_j(\boldsymbol{h}) = \mathcal{M}_\theta(\boldsymbol{y}_j(\boldsymbol{h})), \qquad \forall j \in \{1,\dots,m\} \tag{3.11}$$

$$\hat{\gamma}_{\mathbf{1}_i}(\boldsymbol{h}) = \frac{1}{2\times m}\sum_{j=1}^{m}[\mathbf{1}_i(z_j(\boldsymbol{u})) - \mathbf{1}_i(z_j(\boldsymbol{u}+\boldsymbol{h}))]^2, \qquad \forall i \in \{1,\dots,B\} \tag{3.12}$$

The objective function (Equation 3.13) is the mismatch between the reference model for the indicator variograms and the indicator variogram resulting from the MCS sampling. The mismatch is measured in terms of the sum of squared errors. The evaluation is performed on standardized variograms to equalize the importance of each category.

$$\mathrm{O}(\rho_i(\boldsymbol{h}); i = 1,\dots,K) = \sum_{j=1}^{B}\frac{w_j}{p_j(1-p_j)}\left[\gamma_{\mathbf{1}_j}(\boldsymbol{h}) - \hat{\gamma}_{\mathbf{1}_j}(\boldsymbol{h})\right]^2 \tag{3.13}$$

where p_j is the global proportion of the j^{th} category and w_j is the weight assigned to the reproduction of the indicator variogram of the j^{th} category.

In practice, the confidence on the reference indicator variogram is not equal across all categories. Some reference models are based on stable experimental variograms estimated from a large number of data pairs. Some reference models are fitted to unstable experimental variograms often defined by few samples. Weights can be assigned to the reproduction of each indicator variogram based on the degree of confidence.

The inversion is performed on the major, mid and minor directions of continuity. For each direction, the range of continuity is discretized into H lag distances and each lag distance is optimized independently. The procedure for the optimization of a given direction discretized into H lag steps is detailed below:

(1) for each lag discretization $\{\boldsymbol{h}_i; i = 1,\dots,H\}$ do:

(1.1) set minimum correlation for each latent variable to:

$$\rho_j^{\min} = \begin{cases} 0, & \text{if } i = 1 \\ \rho_j\left(\boldsymbol{h}_{i-1}\right), & \text{otherwise} \end{cases}, \qquad \forall j \in \{1, \ldots, K\}$$

(1.2) initialize iteration counter: $t = 0$

(1.3) set the correlations to the minimum value:

$$\rho_j^{(t)}\left(\boldsymbol{h}_i\right) = \rho_j^{\min}, \qquad \forall j \in \{1, \ldots, K\}$$

(1.4) initialize the samples of the latent variables $\boldsymbol{y}(0)$ and $\boldsymbol{y}(\boldsymbol{h})$ using Equation 3.10

(1.5) while $t < 10$

(1.5.1) increment the iteration counter $t = t + 1$

(1.5.2) for each latent variable $j = 1, \ldots, K$:

(1.5.2.1) increment counter $t = t + 1$

(1.5.2.2) use the golden-section algorithm (Algorithm 3.1) to define the best value for $\rho_j^{(t)}(\boldsymbol{h}_i)$ between $\rho_j^{\min}$ and 1.0

(1.5.3) stop inversion process if all optimized spatial correlations change by less than 10^{-10} in relation to the previous iteration:

$$\max_{i \leq K} \left\{ \left\| \rho_i^{(t-1)}\left(\boldsymbol{h}\right) - \rho_i^{(t)}\left(\boldsymbol{h}\right) \right\| \right\} \leq 10^{-10}$$

The described procedure results in a set of optimized points informing the major, mid and minor directions of continuity for all latent variables. The stopping criteria for the algorithm is set to 10^{-10}. This is far less than uncertainty in the variograms. The resulting set is converted to variogram values using Equation 3.9 and fitted with a valid variogram model. The set of points are often well behaved and automated variogram fitting tools (Deutsch, 2015; Larrondo et al., 2003) can be used to facilitate this task.

To demonstrate the inversion procedure, two GRFs are unconditionally simulated and truncated to generate a reference categorical model (Figure 3.13). The Gaussian variable Y_1 is created using a Gaussian variogram with ranges of 16 units in the major direction (West-East) and 8 units in the minor direction (South-North). The Gaussian variable Y_2 is created with a isotropic Gaussian variogram with range of 16 units.

The truncation rule utilized for this example is shown in Figure 3.14. Category 1 is defined by the rule $Y_1 \leq 0.0$; the category 2 is defined by the truncation $Y_1 > 0.0$ and $Y_2 \leq 0.0$;

Algorithm 3.1 The golden-section line-search algorithm

procedure LINE-SEARCH(ub: upper bound, lb: lower bound, O($\cdot$): objective function)
  $g_r \leftarrow \frac{\sqrt{5}-1}{2}$ (golden ratio)
  $a \leftarrow$ lb
  $b \leftarrow$ ub
  $c \leftarrow b - g_r \times (b - a)$
  $d \leftarrow a + g_r \times (b - a)$
  $f_c \leftarrow \mathrm{O}(c)$ (Evaluate objective function for parameter c)
  $f_d \leftarrow \mathrm{O}(d)$ (Evaluate objective function for parameter d)
  while $(b - a) > 10^{-10}$ **do**
    if $(f_c < f_d)$ **then**
      $b \leftarrow d$
      $d \leftarrow c$
      $c \leftarrow b - g_r \times (b - a)$
      $f_d \leftarrow f_c$
      $f_c \leftarrow \mathrm{O}(c)$
    else
      $a \leftarrow c$
      $c \leftarrow d$
      $d \leftarrow a + g_r \times (b - a)$
      $f_c \leftarrow f_d$
      $f_d \leftarrow \mathrm{O}(d)$
    end if
  end while
  $o \leftarrow \frac{a+b}{2}$
  $f_o \leftarrow \mathrm{O}(o)$
return (o, f_o) (Return optimum parameter o and objective function value)
end procedure

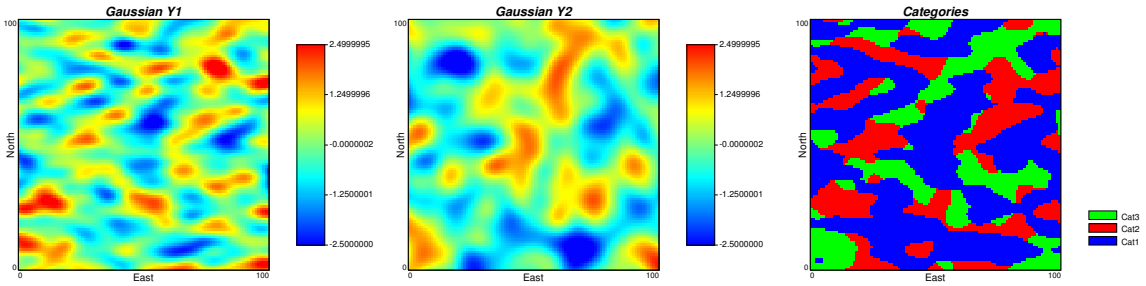


Figure 3.13: Reference model

and the category 3 is defined by Y_1>0.0 and Y_2>0.0. This truncation scheme results in a proportion of 0.5, 0.25 and 0.25 for categories 1, 2 and 3 respectively.

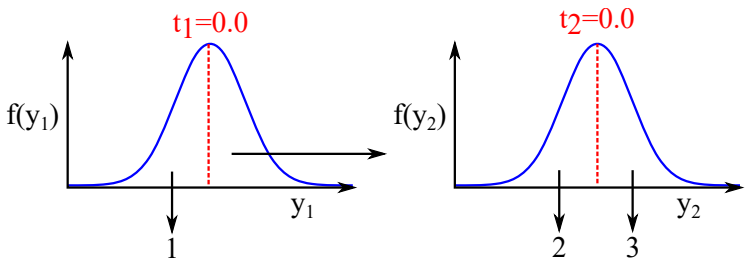


Figure 3.14: Truncation rule for the illustrative example

The experimental indicator variograms of the resultant categorical model are calculated and shown in Figure 3.15. Note that the anisotropy in the variogram of Y_1 has influence in all indicator variograms. After the estimation of the indicator variograms, the inversion procedure is used to infer the variogram of the Gaussian latent variables.

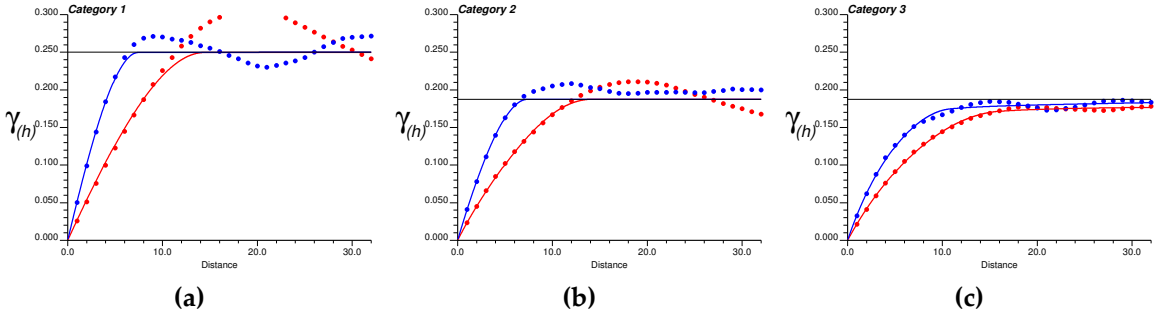


Figure 3.15: Categorical variable indicator variograms for each category

The variograms of the latent variables obtained with the numerical derivation are shown in Figure 3.16. The optimized lags are shown as colored markers (blue for minor direction and red for major direction). The experimental variograms from the original models (Figure 3.13a and 3.13b) are shown as dashed lines in Figure 3.16. Note that the ergodic fluctuations create an apparent anisotropy in the experimental variogram of Y_2. The points defined by the inversion algorithm show good match with the original variogram of Y_1 and also a reasonable match of the original variogram of Y_2. The points are fitted with the Gaussian variogram model (Equation 2.6). The fitted structures are shown as solid lines in Figure 3.16.

The fitted models are used to generate 100 realizations of each variable Y_1 and Y_2. The realizations are truncated to generate realizations of the categorical variable. The first re-

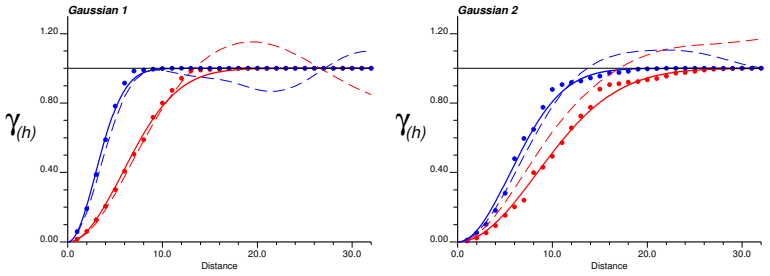


Figure 3.16: Optimized Gaussian variograms

alization of the latent variables as well as the respective categorical model are shown in Figure 3.17.

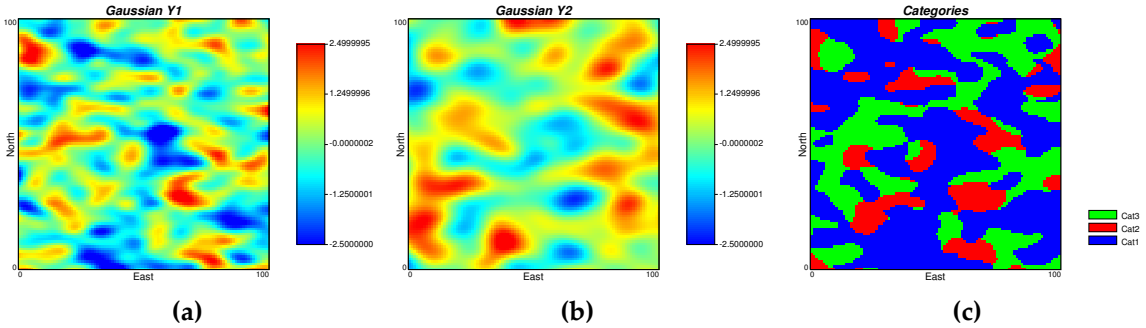


Figure 3.17: One realization of the Gaussian latent variables and resulting categorical variable.

The reproduction of the indicator variograms of the categorical variable is shown in Figure 3.18. The solid lines are the fitted models for the indicator variograms that are used as reference for the optimization. The light gray lines are the variograms of each realization and the dashed lines are the average variogram considering all realizations. Note that the dashed lines have a good match to the fitted models (solid lines).

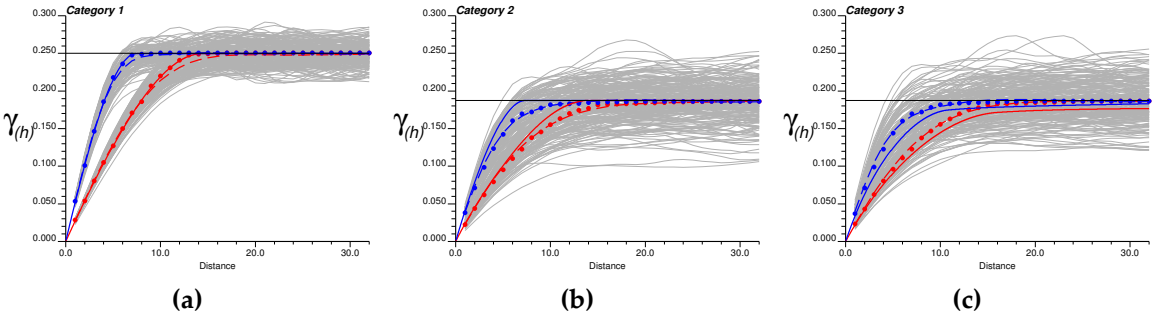


Figure 3.18: Reproduction of indicator variograms of the categorical variable. The solid lines are the reference variogram models. Light gray lines are the variograms of each realization. The dashed lines are the average variograms calculated from all realizations. Markers are the output of the numerical derivation with the optimized points for the indicator variograms.

The indicator variograms resulting from the inversion algorithm are also plotted as colored markers in Figure 3.18. Note that the markers match the dashed lines. The inversion approach provides means of checking the expected variogram reproduction of the HTPG workflow. The mapped points for the indicator variograms can be checked and they can indicate if the adopted procedure, including the chosen truncation rule, is adequate for the problem prior to the simulation of large models.

3.3.5 Imputation of Latent Variables

The Gaussian latent variables used in truncated Gaussian techniques are a model assumption, therefore, these variables are not observed. The truncated Gaussian techniques generate realizations of these latent variables and truncate them to define the categorical models. The only data available are the samples of the categorical variable ($x \sim P\{X = b; b \in \mathcal{B}\}$). The simulated categorical models are expected to match the data at the data locations. If n observations of the categorical variable are available, the condition in Equation 3.14 must be met for all locations.

$$x_i = \mathcal{M}_\theta(\boldsymbol{y}_i), \qquad \forall i \in \{1, \ldots, n\} \tag{3.14}$$

The simplest way of ensuring the reproduction of data is to assign an arbitrary valid number for each location that meets the condition. The center of the truncation interval (Rossi and Deutsch, 2014) is a common choice. Even though this approach enables the data reproduction, the imputed latent variables do not have the correct spatial variability. For instance, if two categorical samples are close in the space and they have different values, there is higher probability that the respective latent variables at that location are close to the boundary. If these samples are assigned to the centroid, artifacts are created near the boundaries.

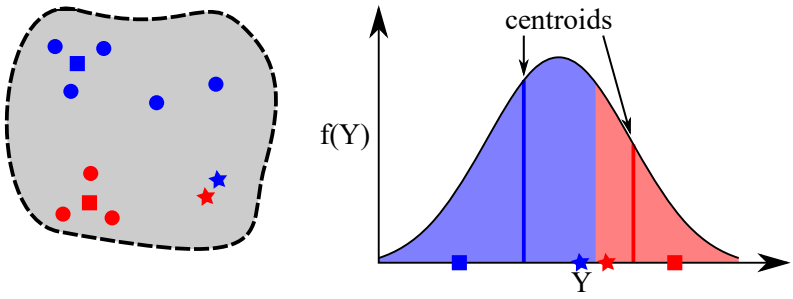


Figure 3.19: Illustrative example of the effect of the spatial configuration of the categorical samples on the underlaying latent variable. The star shaped samples are spatially close to each other, however, they show different categories. In this case the latent variable is expected to be near the threshold if there is spatial correlation between them. The square shaped markers are samples that are within a region surrounded by the same category. In this case, there is a low probability that the latent variable is close to the threshold, in fact, they have a higher chance to be far appart within the standard Gaussian distribution.

The definition of the latent variables should account for the spatial structure and also be conditioned to the categorical data observations. Sampling directly from a high-dimensional truncated Gaussian distribution is not feasible, however, sampling from the marginal univariate conditional distributions is simple. The Gibbs sampler method is well suited for the task of indirectly generating samples of a complex multivariate distribution utilizing the univariate conditional distributions. Variations of the Gibbs sampler algorithm for the data imputation in the truncated Gaussian methods have been developed by many researchers (Astrakova et al., 2015; Emery et al., 2014; Galli and Gao, 2001; Geman and Geman, 1984; Lantuéjoul and Desassis, 2012). The convergence of the Gibbs sampler algorithm is a recurring problem with the sampling of correlated variables. The current methodologies for the application of Gibbs sampler are discussed in Chapter 5 and a approach is developed to mitigate the problems with conversion. An alternative to Gibbs sampler based on simulated annealing is also developed in Chapter 5.

Another important factor that is often overlooked is the non-uniqueness of the solution to Equation 3.14. There are multiple realizations of the latent variable that satisfy the same set of categorical data observations. This is illustrated in Figure 3.20. The blue line is the true underlying latent variable that is not observed in practice. The black and white markers are the categorical observations, the only data available. The grey lines are multiple realizations of the latent variable that has the same spatial continuity and that satisfies the observed categorical data. In order to transfer the uncertainty of the unobserved latent variables, a multiple imputation framework should be utilized. To achieve that, each re-

alization of the categorical variable generated with a truncated Gaussian method should utilize a different realization of the latent variable. The impact of different approaches to the imputation of latent variables on resource uncertainty is also discussed in Chapter 5.

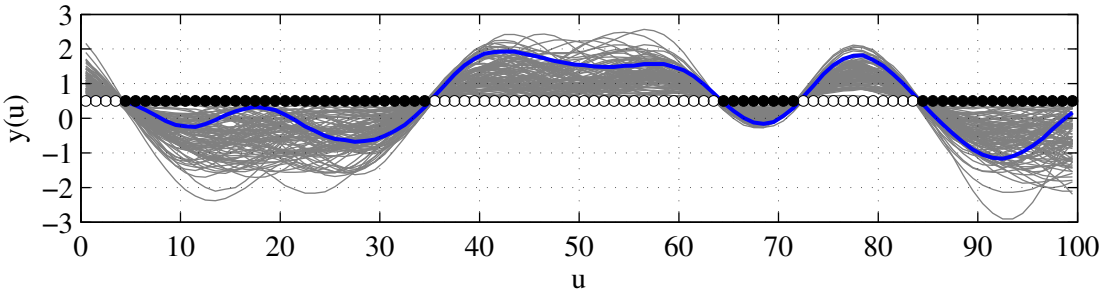


Figure 3.20: Multiple realizations of latent variables that matches the categorical data observation. The black and white markers represent categorical data. The blue line is the true underlying latent variable. The threshold separating black from white is set at $y = 0.5$. The grey lines are a set of 100 realizations of the latent variables that generates the same categorical data observation.

The recommended approach for the imputation of the latent variable in HTPG is the combined Gibbs sampler approach developed in Chapter 5. The methodology improves the convergence and stability of the Gibbs sampler for the sampling of spatially correlated truncated Gaussian vectors. Multiple imputed sets of latent variables, one per realization, should be used to condition the simulation of latent variables and ensure the correct spatial uncertainty in the final categorical models.

3.3.6 Simulation of Latent Variables and Mapping to Categorical Space

After the latent variables are defined at data locations, they can be simulated at all nodes of the modeling grid. Any method for the simulation of Gaussian variables can be used. The most common techniques for the generating of GRFs in geostatistics are the turning bands (Journel, 1974; Matheron, 1973), LU simulation (Alabert, 1987; Davis, 1987), sequential Gaussian simulation (SGS) (Gómez-Hernández and Journel, 1993; Isaaks, 1990), moving average (Black and Freyberg, 1990; Oliver, 1995), and the Fourier integral method (Pardo-Igúzquiza and Chica-Olmo, 1993).

The SGS algorithm is the most popular of all geostatistical simulation techniques (Rossi and Deutsch, 2014). This is due to its simplicity, flexibility and widespread availability in commercial softwares, which makes it the algorithm of choice for most applications of the truncated Gaussian methods. One of the advantages of the truncated Gaussian methods is

that it can readily benefit from any advances in the technology utilized for the generation of GRF's, such as the grid free geostatistical simulation (Zagayevskiy and Deutsch, 2016). As with any other truncated Gaussian technique, HTPG imposes no restriction on the method that is chosen to simulate the Gaussian latent variables.

Once the Gaussian variables are simulated at every grid node, the realizations are mapped back to the categorical space by applying the truncation rule. At this point, if all the preceding steps are properly applied, the categorical realizations should match the categorical data observation, the spatial structure given by the indicator variograms, the categorical proportions and transition probabilities.

3.3.7 Conclusion

A novel HTPG for the application of the truncated Gaussian method is developed. The hierarchical framework allows for the easy incorporation of geological knowledge and understanding to high dimensional cases with many categories and arbitrarily large number of Gaussian latent variables. All the required notation and definitions as well as the properties of the proposed truncation rule are provided.

The technique is suited to the modeling of non-stationary categorical variables through the definition of local parameters for the truncation rule based on local proportions inferred from data. Advances have been made on the numeric derivation of the spatial structure of the Gaussian latent variables. The flexible MCS based inversion technique is robust and can be easily applied to any truncation rule. Important contributions are also made on the imputation of the Gaussian latent variables, however, the topic is discussed with more details in Chapter 5. The multiple data imputation framework should be used in HTPG to ensure the correct characterization of the spatial variability of the categorical variable. Whereas this chapter presents the theoretical grounds of the developed HTPG, the practical considerations on the application and parameterization of the HTPG is the subject of the following Chapter 4.

Chapter 4

Practical Aspects and Parameterization of HTPG

The quality of numerical models is measured in terms of the reproduction of properties observed in the data such as the spatial continuity, global proportions and transition probabilities. The reproduction of morphological characteristics of the geological setting and cross-validation accuracy are also considered.

This chapter is focused on the tools and practical considerations for the definition of the hierarchical truncated pluri-Gaussian (HTPG) parameters to ensure its best performance. The topics of this chapter include: (1) the definition of the truncation rule accounting for transition probabilities, geological expertise and spatial continuity; (2) accounting for locally varying proportions; and (3) the problem of hyper-continuity in the latent variables.

4.1 Defining the Truncation Rule

The truncation rule impacts all steps of the HTPG approach and is the key aspect controlling the quality of the generated models. The main role of the truncation rule is to introduce professional expertise regarding the morphology of the geological setting into the numerical models. The truncation rule controls the possible contacts between categories. Geological expertise should be the main consideration when defining the truncation rule, however, transition probabilities can be used to assist the decisions.

The truncation rule may introduce constraints on the achievable combinations of categorical spatial continuity. As a result, the variograms of the latent variables derived numerically with the inversion approach (Section 3.3.4.1) may show hyper-continuous structures at relatively short ranges. This can often be mitigated by redefining the arrangement of the truncation structure and should be considered.

4.1.1 Geological Expertise

The definition of geological structures involves a mixture of observed factual information and interpretations. Factual information is gathered from core samples, trenches and exposed outcrops, which represents only a infinitesimal portion of the total volume of interest. For this reason, geological interpretation play a significant role on the geological modeling process (Sinclair and Blackwell, 2002).

The geological expertise that goes into the definition of the truncation rule in HTPG is mostly qualitative. Conceptual and genetic models, often based on previous experience and similar deposits, are taken into consideration to assist the interpretation of observed data. The geological sequence and chronology are defined. Fault systems, folding, fractures and veins can extend or disrupt the mineralization depending on their chronological order: before, during, or after, the events that led to the formation of the deposit.

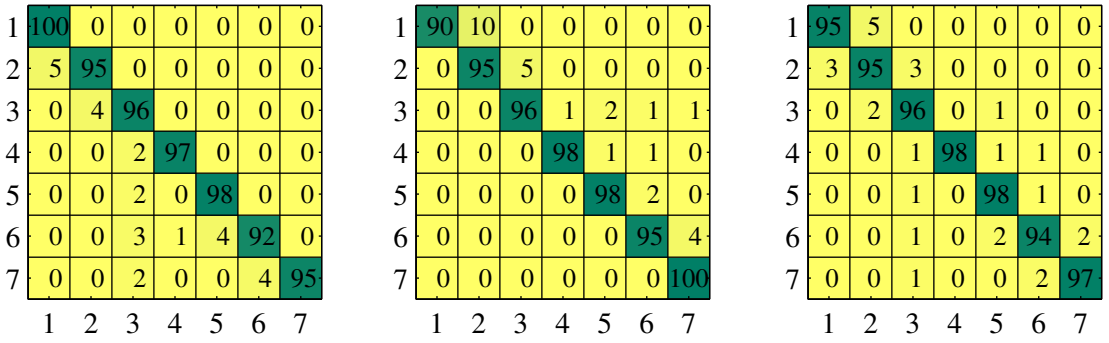
4.1.2 Transition probabilities

Transition probabilities are a useful quantitative measure that have been utilized for the data driven definition of truncation rules (Deutsch and Deutsch, 2014; Sadeghi and Boisvert, 2012). The HTPG methodology is not designed for a blind data driven truncation mask definition. The objective of the HTPG is to facilitate the geological interpretation and allow the modeler to build the truncation tree based on their expertise. This does not impede the practitioner to utilize the information displayed in the transition matrix to assist the interpretation and definition of the mapping rule. The transitions can also be used to check the simulated models.

The diagonal terms of the transition probability represent the transition of a category to itself. The off-diagonal terms are the transition between categories that carries the information regarding the contacts. Depending on the relation between the composite size and the spatial continuity of the categories, the diagonal terms may be much higher than the off-diagonal terms. For visualization purpose, the off-diagonal terms can be rescaled to sum to 1. This facilitates the interpretation of contacts from transition probabilities.

For instance, the transition probabilities for the conceptual model shown in Figure 3.5 are calculated and shown in Figure 4.1. Note that the diagonal terms of the transition probability is much higher than the off-diagonal terms. That makes it difficult to interpret

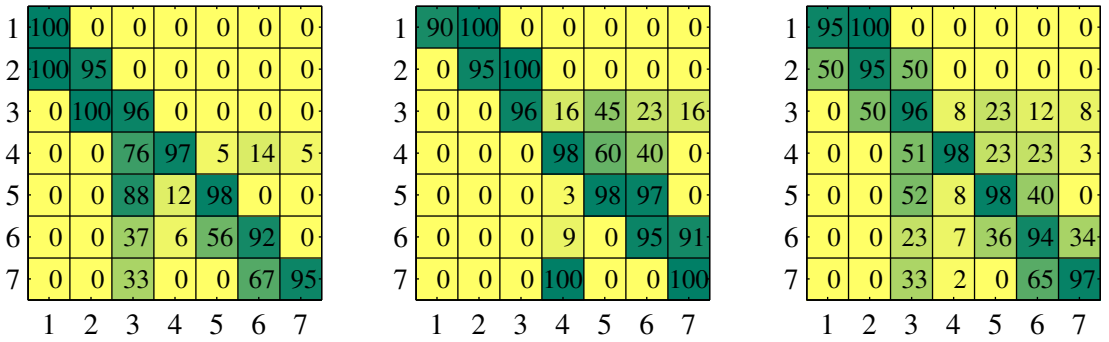
the transitions across different categories.



(a) Up **(b)** Down **(c)** Up and down

Figure 4.1: Transition probability for the conceptual model in Figure 3.5. The displayed values are percentages. In this case the elements in the diagonal are much higher than the off-diagonal elements and the interpretation of the contacts across different categories is compromised.

The transition matrices with rescaled off-diagonal elements are shown in Figure 4.2 for the same case. The rescaled matrices highlights the transition between different categories, facilitating the interpretation. For instance, the first row in Figure 4.2b can be interpreted as: the category 1 transitions to itself 90% of the time and it transitions downwards to category 2 100% of the time, within the remaining 10%. The comparison between the transitions downwards (Figure 4.2a) and upwards (Figure 4.2b) reveals the asymmetry often observed in geological settings.



(a) Up **(b)** Down **(c)** Up and down

Figure 4.2: Transition probability for the conceptual model shown in Figure 3.5, with rescaled off-diagonal elements. The displayed values are percentages. The rescaling highlights the transition across different categories and facilitates the interpretation.

Some of the conclusions that can be drawn from the visualization of the transition probabilities shown in Figure 4.2 are that: (1) category 1 is on top of the entire sequence as it does not transition upwards to any other category; (2) Category 2 is bounded by categories

1 above and 3 below; (3) Category 3 transitions to all categories but category 1 which reveals that is must be located at a discordant boundary.

When the number of categories is relatively high, the interpretation of transition probabilities may become cumbersome. In order to facilitate the interpretation of the transition matrices in such cases, a visualization summary based on multidimensional scaling (MDS) and minimum spanning tree (MST) is proposed. The dimension reduction methodology is similar to the MDS approach utilized by Deutsch and Deutsch (2014) to automatically generate truncations masks, however, the dimension is always reduced to two to facilitate visualization. Also, there is no need to rescale or rotate the coordinates resulting from MDS as the points are not related to the proportions and are not used to truncate the Gaussian distribution. The MST is utilized to improve the visualization of categorical relations as the nearest nodes are connected to each other.

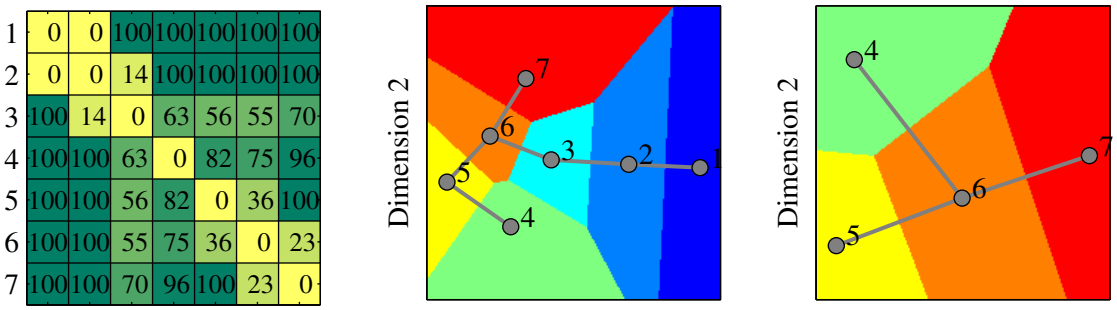
The transition matrix is asymmetric and represents a measure of similarity. The MDS requires the definition of a symmetric dissimilarity matrix $\boldsymbol{S} = (s_{i,j})$. The suggested dissimilarity matrix is defined in Equation 4.1. To achieve symmetry the transition matrix is averaged with its transposed. The terms on the diagonal are set to zeros as a categorical class cannot be dissimilar to itself. The off-diagonal terms are rescaled by the maximum transition and the power of two is applied. The dimension reduction for visualization is always performed to project into two dimensions, therefore, problems with $\boldsymbol{S}$ not being positive definite should not be a concern at such low dimension (Deutsch and Deutsch, 2014).

$$s_{i,j} = \begin{cases} \left(\frac{1-\frac{1}{2}(t_{i,j}+t_{j,i})}{t_{\max}}\right)^2, & \text{if } i \neq j \\ 0, & \text{otherwise} \end{cases}, \qquad t_{\max} = \max_{\substack{1 \leq i,j \leq B \\ i \neq j}} \left\{1 - \frac{1}{2}\left(t_{i,j} + t_{j,i}\right)\right\} \tag{4.1}$$

The dissimilarity matrix calculated from the transition matrix in Figure 4.2c is shown in Figure 4.3a. The respective visual summary of the transition matrix is shown in Figure 4.3b.

From Figure 4.3b it is possible to see how the category 4 is not related to any of the other category. Also it is possible to see how categories 1, 2 and 3 are ordered and how 5, 6 and 7 are also ordered. One can also notice from Figure 4.3b that the sequence formed by 5, 6 and 7 is discordant with that formed by 1, 2 and 3 and that category 3 is the one making contact with the others. This graphical tool together with geological understanding can be utilized

to define the truncation rules. Once a set of rules is defined, i.e. the geologist decides to consider 1, 2, and 3 as an ordered sedimentary unit the procedure can be repeated for the remaining categories (Figure 4.3c).



(a) Dissimilarity matrix **(b)** Visual summary **(c)** Visual summary for 4-7

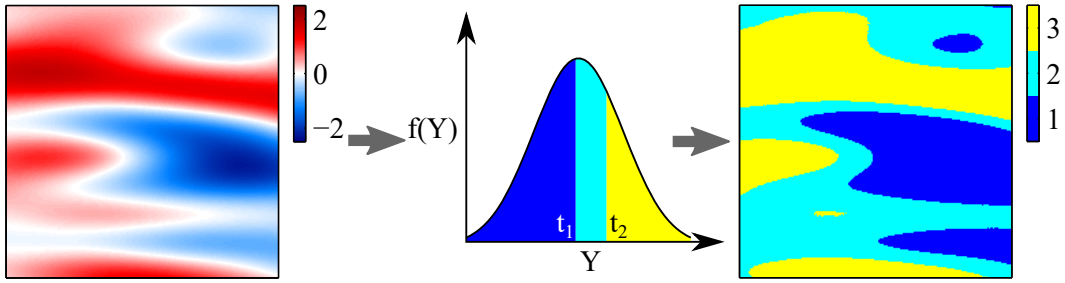
Figure 4.3: Dissimilarity matrix for the conceptual model shown in Figure 3.5 and visual summaries based on MDS and MST.

4.1.3 Asymmetry

The Gaussian random function (GRF) utilized within the HTPG framework are often generated with variogram based Gaussian simulation algorithms suited for stationary variables. The variograms are symmetric functions and the resulting transitions are also symmetric (Equation 2.32). Asymmetry is often observed in transition probabilities calculated in opposing directions (e.g. upwards and downwards). The enforcement of asymmetric transitions in the simulated models is approximately dealt with by locally varying proportions.

The Figure 4.4 shows an example of asymmetry enforcement. The truncation of a stationary GRF with constant thresholds generates symmetric transitions (Figure 4.4a). The sequence in the truncation mask (dark blue-light blue-yellow) can occur in any direction, the only constraint is that light blue occurs between dark blue and yellow. The transition probabilities in Figures 4.4b and 4.4c are symmetric with small deviations caused by ergodic fluctuations.

The same GRF is truncated utilizing locally varying thresholds calculated from the vertical proportion curve (Figure 4.4d). The non-stationary approach allowed the generation of a asymmetric model in which the transition from dark blue to light blue to yellow can only occur upwards whereas the opposite occurs downwards. The resulting asymmetric transitions are shown in Figures 4.4e and 4.4f.



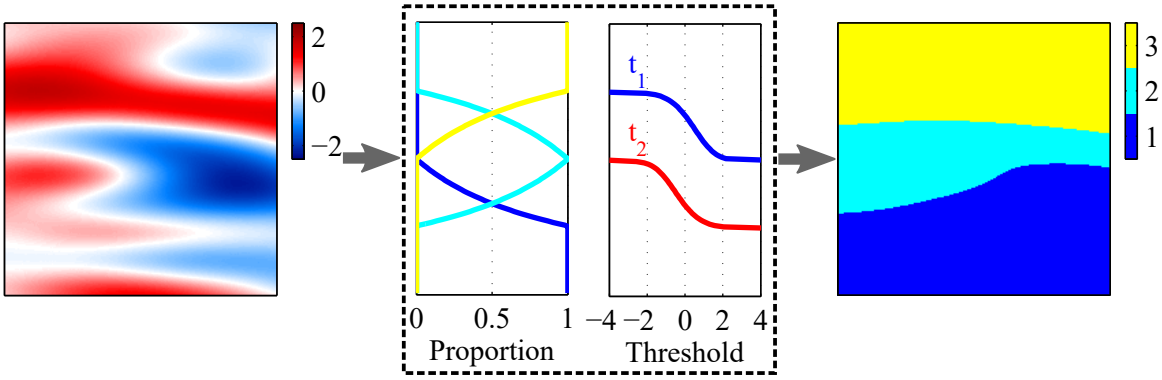
(a) Stationary truncation

	1	2	3
1	97	100	0
2	52	96	48
3	0	100	96

(b) Stationary $\boldsymbol{T}_{\mathrm{u}}$

	1	2	3
1	97	100	0
2	39	95	61
3	0	100	98

(c) Stationary $\boldsymbol{T}_{\mathrm{d}}$



(d) Non-stationary truncation

	1	2	3
1	99	100	0
2	0	98	100
3	0	0	0

(e) Non-stationary $\boldsymbol{T}_{\mathrm{u}}$

	1	2	3
1	0	0	0
2	100	98	0
3	0	100	99

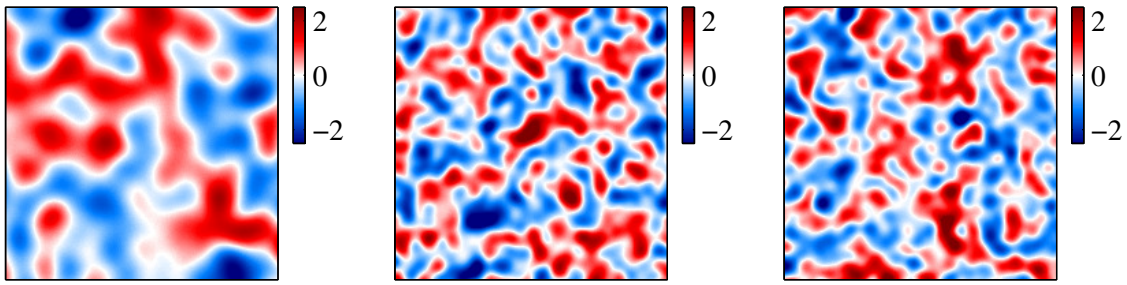
(f) Non-stationary $\boldsymbol{T}_{\mathrm{d}}$

Figure 4.4: Illustrative example of the utilization of non-stationary truncation for asymmetry enforcement. (a) shows the truncation of a stationary GRF with constant global thresholds. (b) and (c) show the resulting transition probabilities for the truncation in (a). The slight asymmetry observed in (b) and (c) are from ergodic fluctuations. (d) shows the truncation of the stationary GRF with locally varying thresholds. (e) and (f) show the resulting transition probabilities for the truncation in (d) and show strong asymmetry enforced by the trend.

4.1.4 Geological Contacts and Non-stationarity

The lack of transition between two categories can be caused by different factors. For instance, the chronological order of a sedimentary sequence may introduce a physical barrier between two geological units in which the transition between one unit to another is not possible without the transition to intermediary units. In this case, there is a sharp well defined barrier between the two geological units and the transition or lack of transition is enforced using the truncation rule. In certain instances, there may be no sharp well defined barrier, and the two categories do not coexist at the same region. In these cases, the lack of transitions could be enforced using locally varying proportions.

A GRF with three variables (Figure 4.5) is utilized to generate categorical variables by applying three different truncation rules (Figure 4.6) to illustrate some of the different geological contacts and how they appear in the transition matrices.

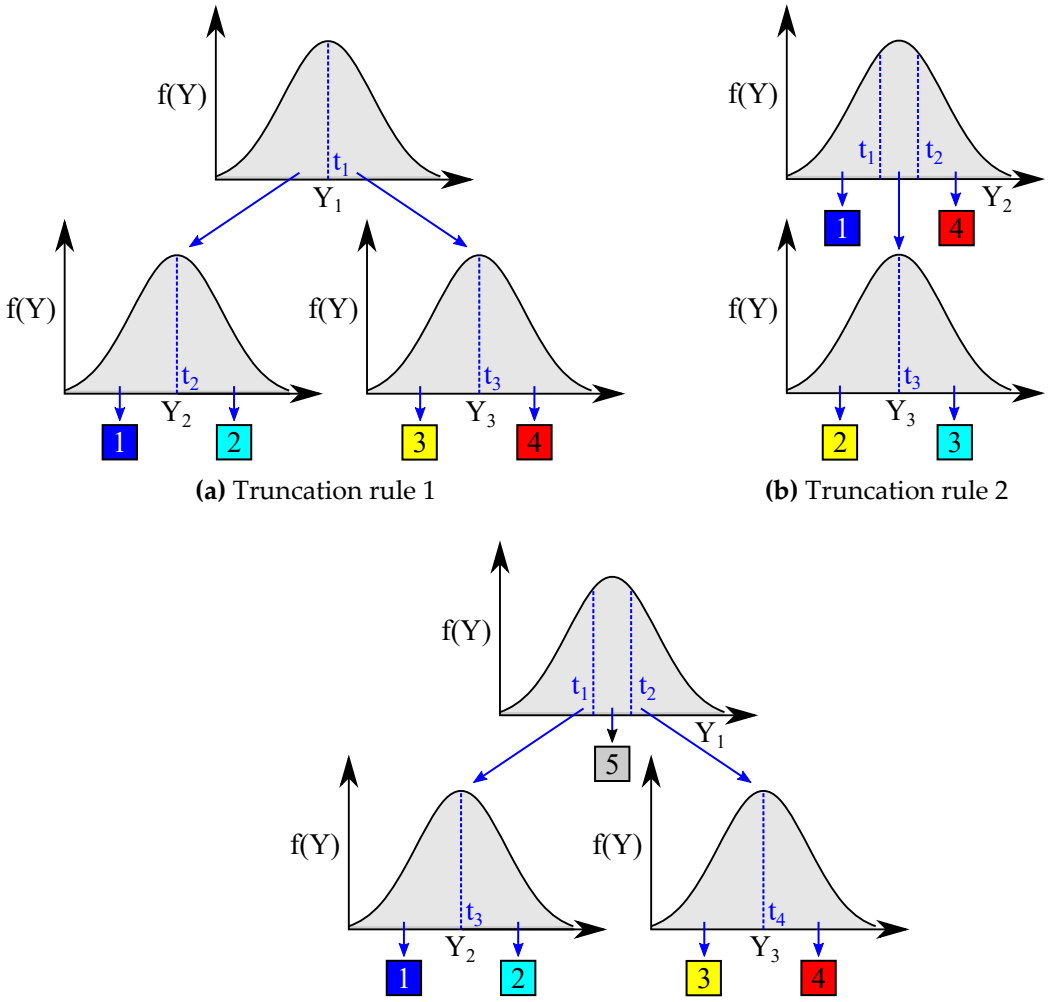


(a) Y_1 **(b)** Y_2 **(c)** Y_3

Figure 4.5: GRF utilized to demonstrate different types of categorical contacts and transitions.

The first example is generated by applying the truncation rule shown in Figure 4.6a with constant thresholds to generate equal proportion (25%) of each category. The truncation results in the categorical model and transitions shown in Figure 4.7. The category pairs 1-2 and 3-4 are separated by Y_1 into two branches that leads to two other nodes where the pairs are defined. The separation of the two pairs as depicted in Figure 4.6a makes the transition between 1-2 and 3-4 more likely, however, it is not enough to remove the possibility of one particular contact. The resulting MDS/MST visualization (Figure 4.7c) alone can be deceiving as it depicts category 1 and 4 far from each other suggesting no contacts, however, that can be quickly checked by looking at the transition probability matrix (Figure 4.7b).

The same truncation rule can be used with the locally varying proportions shown in Figure 4.8. The non-stationary truncation results in the categorical model and transitions

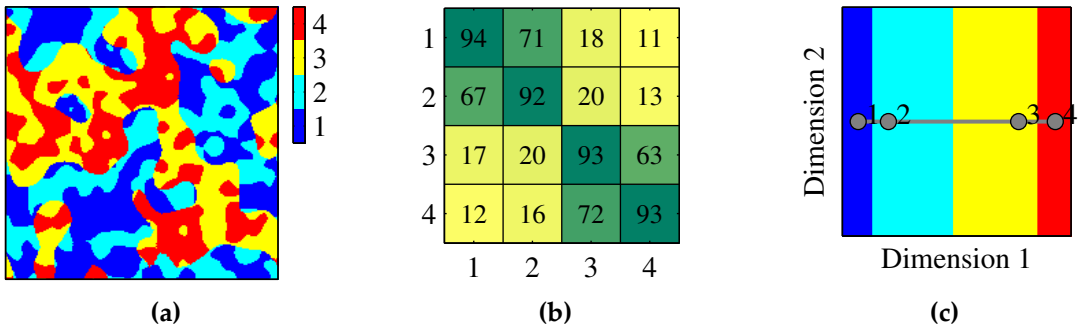


(a) Truncation rule 1

(b) Truncation rule 2

(c) Truncation rule 3

Figure 4.6: Different cases of truncation rule used to demonstrate different types of categorical transitions.



(a) **(b)** **(c)**

Figure 4.7: Illustrative example. (a) stationary categorical model generated by truncating the GRF in Figure 4.5 and applying the truncation rule in Figure 4.6a; (b) resulting transition probabilities with rescaled off-diagonal terms; and (c) visual summary based on MDS and MST.

shown in Figure 4.9. In this case the transition between category 4 and category 1 is constrained by the use of a trend, resulting in the separation of the two categories without a sharp physical barrier.

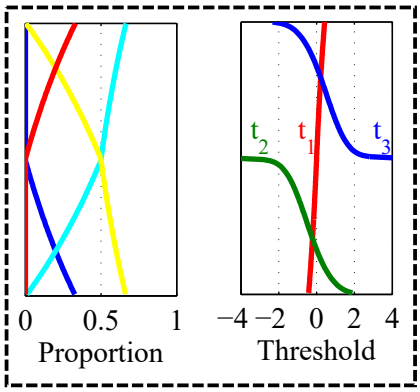


Figure 4.8: Local proportions and respective local thresholds applied to Figure 4.6b.

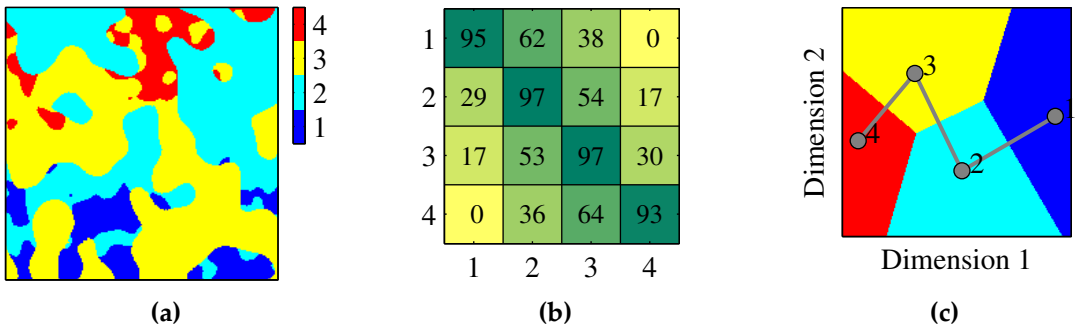


Figure 4.9: Illustrative example. (a) non-stationary categorical model generated by truncating the GRF in Figure 4.5 applying the truncation rule in Figure 4.6a with the local thresholds shown in Figure 4.8; (b) resulting transition probabilities with rescaled off-diagonal terms; and (c) visual summary based on MDS and MST.

The transition matrix 4.9b show the lack of transition between rock type 1 and 4, but it does not necessarily show the nature of the physical separation between the two units. For instance, a similar transition matrix can be achieved using the truncation rule shown in Figure 4.6b without a trend model. Note that this truncation rule is only applied to Y_2 and Y_3. This truncation results in the categorical model and transitions shown in Figure 4.10. The transition probabilities in Figure 4.9b are similar to those in Figure 4.10b even though the nature of the geological contacts are very different.

If categorical pairs 1-2 and 3-4 are disconnected in space, an additional category is required to ensure the physical separation between the two sets. The categorical model and

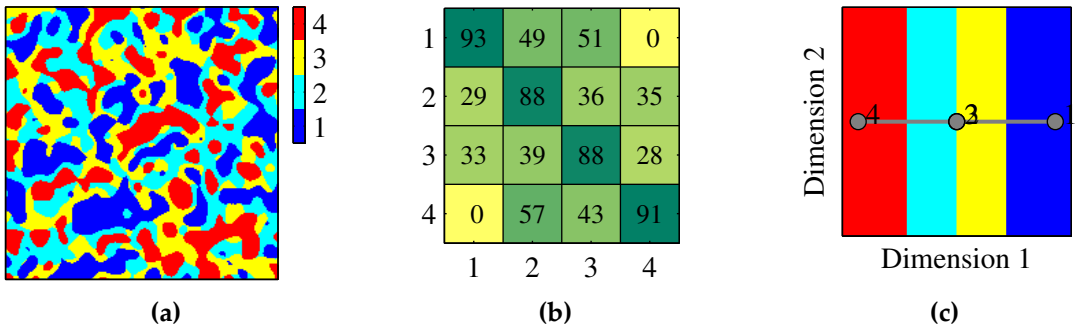


Figure 4.10: Illustrative example. (a) stationary categorical model generated by truncating the GRF in Figure 4.5 applying the truncation rule in Figure 4.6b (note that only Y_1 and Y_2 are used); (b) resulting transition probabilities with rescaled off-diagonal terms; and (c) visual summary based on MDS and MST.

transitions shown in Figure 4.11 are results of the truncation rule shown in Figure 4.6c applied to the GRF. In this case, category 5 removes all the transitions across the two pairs of categories. This is clearly seem in the transition matrix shown in Figure 4.11b and in the MDS/MST visualization in Figure 4.11c.

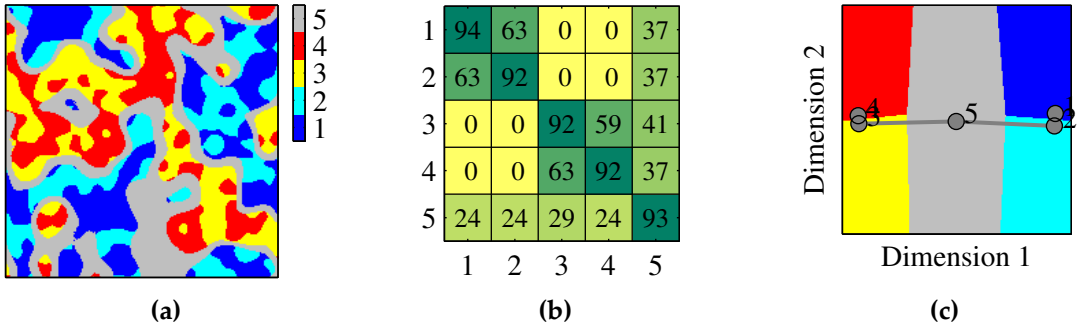


Figure 4.11: Illustrative example. (a) stationary categorical model generated by truncating the GRF in Figure 4.5 applying the truncation rule in Figure 4.6c; (b) resulting transition probabilities with rescaled off-diagonal terms; and (c) visual summary based on MDS and MST.

4.2 Accounting for Locally Varying Proportions

Categorical variables are often non-stationary and the utilization of locally varying proportions is common. The spatial variability of the categorical variable in these cases is a combination of the continuity of the deterministic trend and the stochastic residuals. Indicator variograms calculated directly from categorical data without consideration of the spatial structure of the trend leads to realizations that are more spatially continuous than the un-

derlying variable. In order to achieve appropriate variogram reproduction while modeling categorical variables with trends, it is important to calculate and utilize the variogram of the indicator residuals.

A synthetic 2D example is created to illustrate the required steps for modeling categorical variables in presence of a trend. The model size is set to 100x100 grid cells of 1 unit each. The example is built utilizing a single Gaussian latent variable that is truncated by a single threshold resulting in two categories (0 and 1). A local trend varying with the Y coordinate is created starting from 0.995 proportion of category 1 at lowest Y coordinate to 0.005 proportion of category 1 at highest Y coordinate (Figure 4.12a). Local thresholds are calculated to match the local proportions (Figure 4.12b).

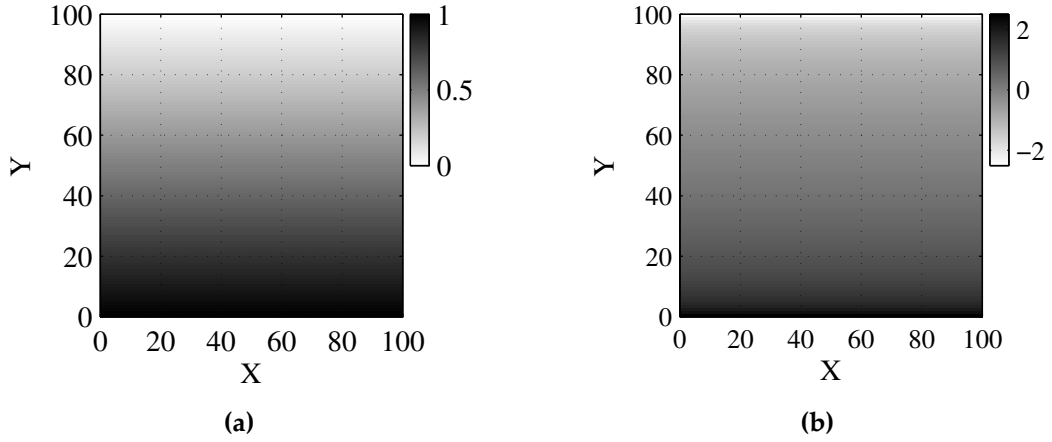
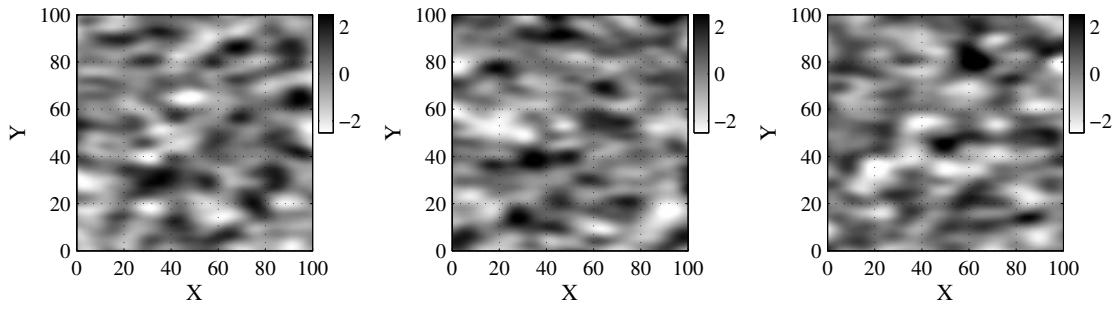


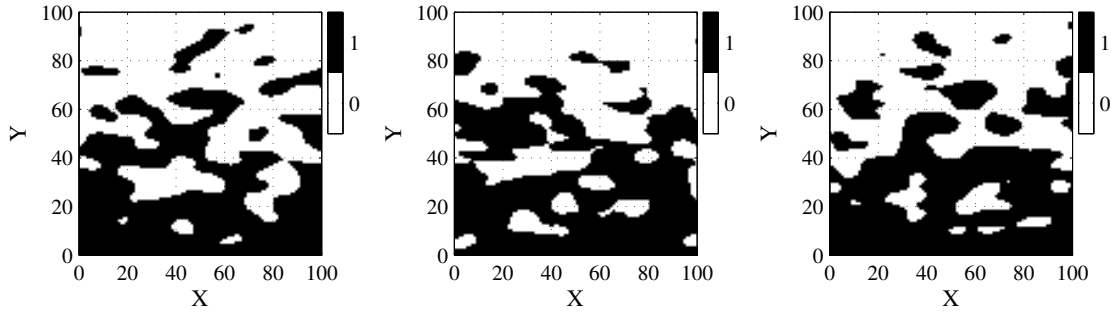
Figure 4.12: Trend model for the 2D synthetic model. (a) Local proportion of category 1. (b) local threshold calculated to match the local proportions.

The variogram of the latent variable has a Gaussian structure (Equation 2.6) with ranges set to 16 and 8 units in X and Y directions respectively. A set of 100 unconditional realizations are generated using sequential Gaussian simulation (SGS). The Gaussian realizations are truncated utilizing the local thresholds (Figure 4.12b) to generate the reference categorical models. Three realizations of the reference models are shown in Figure 4.13.

The variograms of the reference models are shown in Figure 4.14. The light red and light blue are the variograms of each realization and the solid line is the average variogram. The average variogram matches the reference used for the simulation, which is shown as dashed line in Figure 4.14a. In practice, the underlying latent variables are not observed and the categories are the only available data. The true variograms of the latent variables are unknown and the only reference variograms that are observed are the indica-



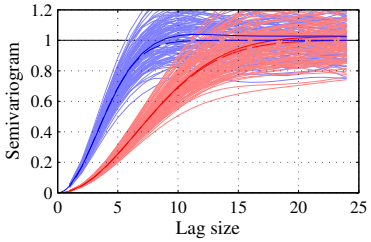
(a) Gaussian realizations



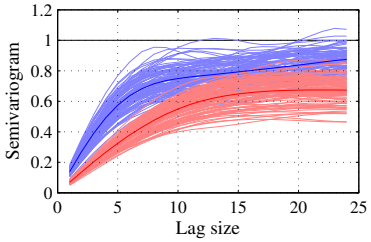
(b) Categorical realizations

Figure 4.13: Three unconditional realizations of the reference Gaussian models (a) and the reference categorical models generated after truncation (b).

tor variograms calculated from the categorical data. In this example, the average indicator variogram shown as solid lines in Figure 4.14b is the equivalent to the data variogram.



(a) Reference Gaussian variogram

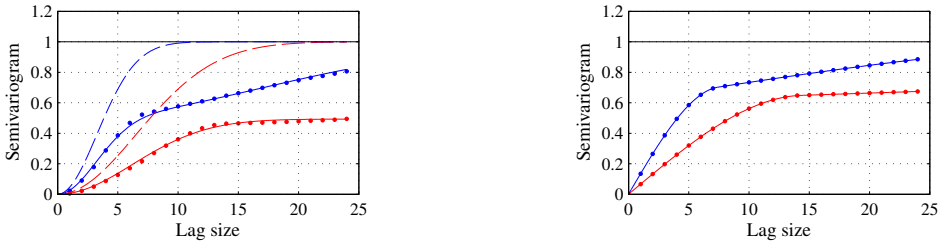


(b) Reference indicator variogram

Figure 4.14: Variogram reproduction for the reference models. Shades of red is utilized for the variograms calculated in X direction and blue is utilized for the variograms in Y direction. Solid dark colored lines are utilized for the average variograms and light colored lines are utilized for the variograms of the realizations. Dark dashed lines in (a) show the reference variogram utilized to generate the Gaussian realizations. The variograms in (b) are standardized.

The indicator variograms are used directly to define the Gaussian variograms. The numerical derivation approach described in Section 3.3.4.1 is utilized for this step. The inversion will consider the desired indicator variogram and truncation parameters and calculate the latent variable variograms that best match the input indicator variogram.

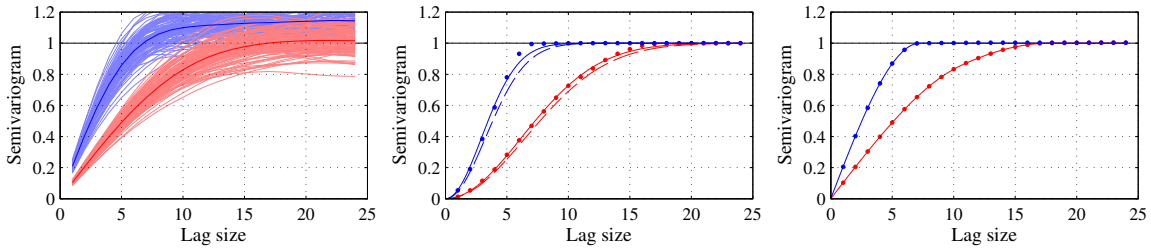
The optimized points are shown as markers in Figure 4.15a and the model fitted to the optimized points is shown as solid lines. The dashed lines in Figure 4.15a are the reference models utilized to build the synthetic example, however, the true variogram of underlying Gaussian variables is unknown in practice. Note that the software matches the reference indicator variogram perfectly (Figure 4.15b), however, by not accounting for the trend during the calculation of the reference indicator variogram, the Gaussian variogram will show greater continuity than it should.



(a) Derived variogram **(b)** Inversion reproduction

Figure 4.15: Results from numerical variogram derivation. Red is utilized for the variograms calculated in X direction and blue is utilized for the variograms in Y direction. Solid dark colored lines are utilized for the fitted models. Markers are the optimized points for the Gaussian variogram in (a) and the expected reproduction for the indicator variogram in (b). The dashed lines in (a) are calculated from the reference Gaussian variogram utilized to build the example.

The second and correct approach is to consider the trend while calculating the reference indicator variogram. One way to consider the trend during the variogram calculation is to calculate and model the variogram of the residuals. The variogram of the residuals calculated for all reference models are shown in Figure 4.16a.



(a) Variogram of residuals **(b)** Derived variogram **(c)** Inversion reproduction

Figure 4.16: Results from numerical variogram derivation utilizing the variogram of the residuals as input. Red is utilized for the variograms calculated in X direction and blue is utilized for the variograms in Y direction. Solid dark colored lines are utilized for the average variogram in (a) and fitted models in (b) and (c). The light colored lines in (a) are the variograms of each realization. Markers are the optimized points for the Gaussian variogram in (b) and the expected reproduction for the indicator variogram in (c). The dashed lines is (b) are calculated from the reference Gaussian variogram utilized to build the example.

The average variogram of the residuals (Figure 4.16a) are standardized, modeled and rescaled to the proper sill accordingly to the declustered proportion of the corresponding category before their utilization to define the latent variable variogram. The variogram is modeled to the sill (solid line in Figure 4.16c), even though the experimental variogram goes slightly above it. The derived points for the latent variable variogram are shown as markers and the fitted model is shown as solid lines in Figure 4.16b. The dashed lines in Figure 4.16b is the reference model utilized to build the synthetic example. Note that the derived variogram (Figure 4.16b) is much closer to the true variogram (Figure 4.14a) than the derived variogram without considering the trend (Figure 4.15a). The inversion approach is able to perfecly match the variogram of the residuals (Figure 4.16c).

Two sets of 100 realizations are generated utilizing the two derived variograms calculated with and without considering the indicator residuals. The same seed number utilized to simulate the initial reference models is also utilized to generate the two sets of realizations to facilitate the visualization of the differences between the two approaches. Three realizations of each case are shown in Figure 4.17. The spatial continuity of the models shown in Figure 4.17a are increased by not considering the trend during the definition of the Gaussian variable variograms. The models generated with the latent variable variogram derived from the variogram of the residuals have virtually the same spatial continuity as the reference models with small differences due to the slightly different variogram (Figure 4.16b).

The variogram reproduction for the two cases are shown in Figure 4.18. As expected, the differences seem in Figure 4.17 are also observed in the variogram reproduction for each case. By defining the variogram of the Gaussian variables from the indicator variogram of the data directly and later utilizing local thresholds to truncate the Gaussian models, the continuity of the trend is introduced two times. This results in models with exaggerated continuity. By defining the Gaussian variogram from the variogram of the indicator residuals, the continuity of the trend is only added during the truncation with local thresholds and the final continuity of the simulated models matches the continuity observed in the raw data.

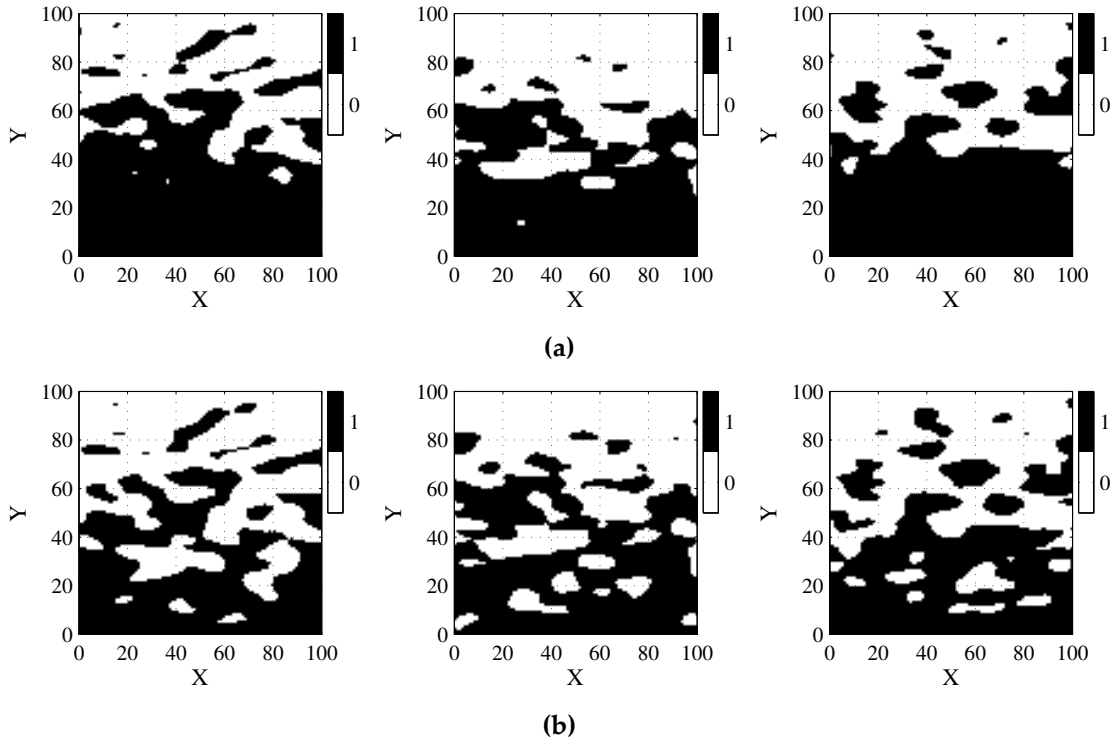


Figure 4.17: Resulting categorical models from the two approaches. The models generated without considering the residuals are shown in (a) and the models generated accounting for the variogram of the residuals are shown in (b).

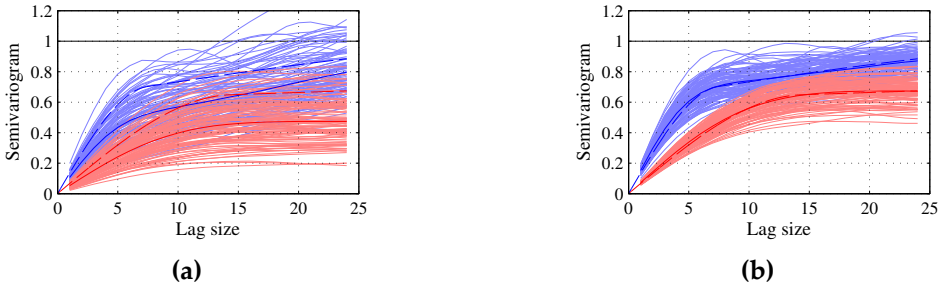


Figure 4.18: Variogram reproduction for the two cases with and without the use of the variogram of the residuals. The variograms in (a) are result of the approach without considering the variogram of the residuals and (b) show the results when the variogram of the residuals is utilized. Shades of red is utilized for the variograms calculated in X direction and blue is utilized for the variograms in Y direction. Solid dark colored lines are utilized for the average variograms and light colored lines are utilized for the variograms of the realizations. Dark dashed lines show the target reference indicator variogram.

4.3 Hyper-continuity and Variogram Reproduction

The application of truncated Gaussian methods, including HTPG, requires the definition of the variograms of the latent variables. Depending on the mapping and categorical proportions, the derived variograms for the latent variables might show extreme continuity at smaller lags and quickly transition to higher variability after a certain distance. This hyper-continuity cannot be modeled with known variogram models. This feature is undesirable and can often be mitigated by rearranging the ordering of the truncation structure in HTPG.

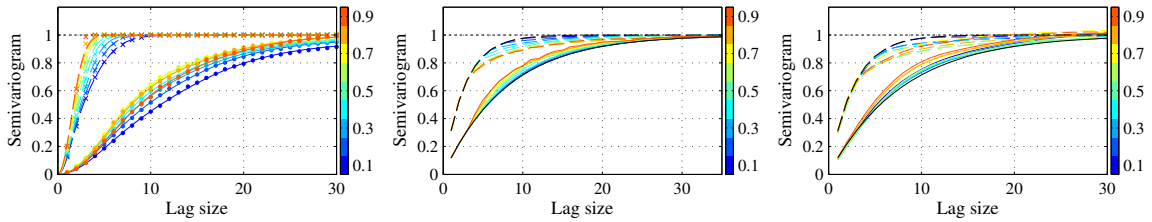
The causes and implications of hyper-continuity are investigated using a truncation structure consisting of two Gaussian latent variables with one truncation threshold applied to each. This configuration results in 3 categories. The first Gaussian variable at the top of the truncation structure is utilized to separate category 1 from categories 2 and 3. The second Gaussian variable is truncated to separate categories 2 from 3. With this configuration, category 1 can have a different variogram structure from category 2 and 3, but 2 and 3 have the same variogram structure as they share an ending node on the truncation tree.

Three main factors are investigated in this section: (1) order in the truncation structure; (2) continuity of the categories; and (3) proportion of each category. To investigate the importance of the categorical proportion, the proportion of the category 1 is changed in 9 steps from 0.1 to 0.9. The remaining proportion is split equally between category 2 and 3 for each case. Three continuity cases are considered. The first case has category 1 as the most continuous with range of 24 units while category 2 and 3 are the least continuous with range of 8 units. The second case equally continuous categories range of 8 units and the third case has category 1 as the least continuous with range of 8 units and categories 2 and 3 being the most continuous with range of 24 units.

For each one of these cases the numerical derivation (Section 3.3.4.1) is run to define the variogram of the latent variables. The resulting expected reproduction of the inversion approach is only correct if the Gaussian model can be reasonably fitted with valid models. If the fitted models are not close enough to the optimized points, the variogram reproduction should be evaluated after generating simulated models. In most cases, fitting the optimized points is straightforward, however, this is not true for hyper-continuous points. In order to demonstrate how much the actual reproduction deviates from the predicted, the optimized

variograms points are fitted with Gaussian variogram structures (Equation 2.6) and used to simulate 100 realizations at grid resolution of 100x100 cells of 1x1 unit each. The average variogram calculated from all realizations is compared with the predicted by the inversion approach for each case.

The results for the case in which the category 1 has the most continuity are shown in Figure 4.19. The derived points for the latent variables variograms are shown in Figure 4.19a. Hyper-continuity does not occur for the case in which category 1 has the most continuity, however, the categorical variogram reproduction deteriorates as the proportion of category 1 increases. The expected variogram reproduction shown in Figure 4.19b is not much different from the actual variogram reproduction shown in Figure 4.19c. This is expected as the optimized points are well matched by the fitted models in Figure 4.19a.

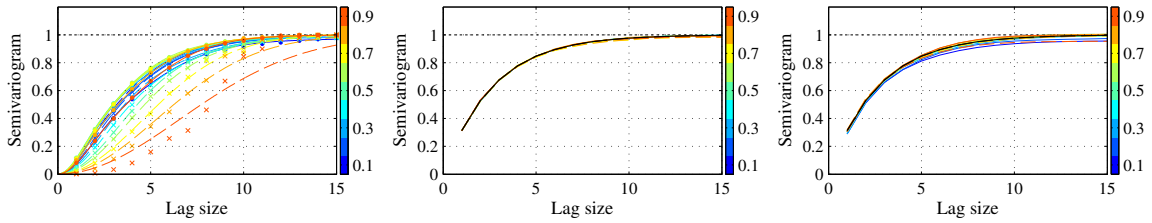


(a) Optimized latent variogram **(b)** Expected reproduction **(c)** Actual reproduction

Figure 4.19: Results for the case in which category 1 is the most continuous. The variograms are colored according with category 1 proportion. The circular markers in (a) are the derived points for the first latent variable and the cross markers are for the second latent variable. The solid lines in (a) are the fitted models for first latent variable and the dashed lines are the fitted models to the second latent variable. (b) shows the expected variogram reproduction for category 1 (solid lines) and categories 2 and 3 (dashed lines). Black lines in (b) are the reference variograms used as input. (c) is similar to (b), however, it is calculated from actual simulated models utilizing the fitted models shown in (a).

The results for the case in which the category 1 has the same continuity as the categories 2 and 3 are shown in Figure 4.20. Hyper-continuity starts to appear in this example, specially for the cases in which category 1 have high proportions (0.8 and 0.9) as shown in Figure 4.20a. The expected variogram reproduction shown in Figure 4.20b is not much different from the actual variogram reproduction shown in Figure 4.20c even though the optimized points are not well matched by the fitted models (Figure 4.20a) for some instances. This is attributed to the fact that all categories have the same structure and the hyper-continuity is not relevant for such high proportion of category 1, which is controlling most of the continuity.

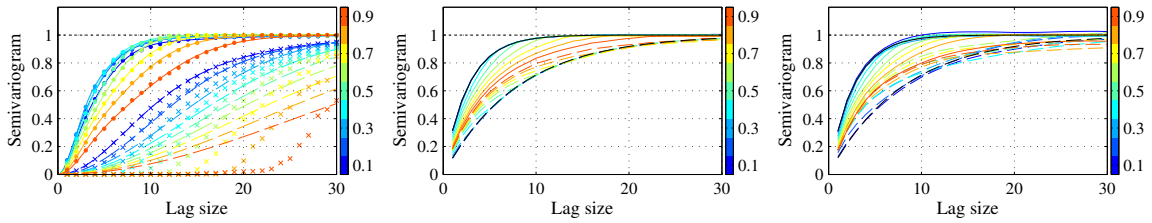
The results for the case in which the category 1 has the least continuity are shown in



(a) Optimized latent variogram **(b)** Expected reproduction **(c)** Actual reproduction

Figure 4.20: Results for the case in which all categories are equally continuous. The variograms are colored according with category 1 proportion. The circular markers in (a) are the derived points for the first latent variable and the cross markers are for the second latent variable. The solid lines in (a) are the fitted models for first latent variable and the dashed lines are the fitted models to the second latent variable. (b) shows the expected variogram reproduction for category 1 (solid lines) and categories 2 and 3 (dashed lines). Black lines in (b) are the reference variograms used as input. (c) is similar to (b), however, it is calculated from actual simulated models utilizing the fitted models shown in (a).

Figure 4.21c. The hyper-continuity is strong in this case. For all the cases in which category 1 have proportion higher than 0.2, the optimized points for the second latent variable show hyper-continuity (Figure 4.21a) and are not well matched by the fitted models. This condition deteriorates the expected variogram reproduction shown in Figure 4.21b. For proportions of category 1 higher than 0.4, it is not possible to reproduce the indicator variograms even for the numerically derived points. The actual variogram reproduction (Figure 4.21c) is worst than the predicted, however, it is reasonably close to the predicted ones considering how poorly matched the optimized points are by the fitted models in Figure 4.21a.



(a) Optimized latent variogram **(b)** Expected reproduction **(c)** Actual reproduction

Figure 4.21: Results for the case in which category 1 is the least continuous. The variograms are colored according with category 1 proportion. The circular markers in (a) are the derived points for the first latent variable and the cross markers are for the second latent variable. The solid lines in (a) are the fitted models for first latent variable and the dashed lines are the fitted models to the second latent variable. (b) shows the expected variogram reproduction for category 1 (solid lines) and categories 2 and 3 (dashed lines). Black lines in (b) are the reference variograms used as input. (c) is similar to (b), however, it is calculated from actual simulated models utilizing the fitted models shown in (a).

The results lead to the conclusion that, whenever possible, the categories with lowest

proportions and highest continuity should be defined earlier on the truncation tree while the least continuous categories with higher proportions should be defined closer to the end nodes of the truncation structure. As the dark blue lines are well behaved in all results (Figures 4.19, 4.20 and 4.21), the proportion criteria should be given priority over continuity if a decision is required. Thankfully, the inversion approach can be quickly run and the resulting variograms assessed before performing simulation at large numeric models.

4.4 Conclusion

The practical aspects and parameterization of the HTPG is discussed in this Chapter. Several factors are taken into consideration to define the truncation rules utilized in HTPG. The hierarchical truncation rules are designed to facilitate the introduction of expert knowledge into the modeling by constraining the geological transitions that are allowed in a given geological setting.

In addition to the professional input, the utilization of quantitative information calculated from data is also discussed. The transition matrix calculated along the drilling path contains information regarding the geological contacts, ordering and symmetry. A graphical approach based on MDS and MST is proposed to facilitate the interpretation. The methodology allows a quick visualization of the relationships between the categories.

The problem of non-stationarity and variogram reproduction while modeling with HTPG is also discussed. It is shown that the indicator variogram calculated from the raw data is not suitable for the definition of the latent variables variogram in presence of a trend. In this situation, the variogram of the indicator residuals must be used instead. This procedure ensures proper variogram reproduction where the utilization of the indicator variogram from the raw data would lead to models with higher continuity than expected.

The causes of hyper-continuous variogram points after latent variogram optimization is also investigated. It is shown that the categorical proportion, spatial continuity and relative position on the truncation structure play major roles in the occurrence of hyper-continuity. It is also shown that categories with lower proportions should be defined as early as possible in the truncation structure as well as categories with the highest continuity. The least continuous categories and categories with highest proportions should be defined closer to the ending nodes of the truncation structure whenever possible.

Chapter 5

Multiple Data Imputation for HTPG

The truncated Gaussian techniques, including hierarchical truncated pluri-Gaussian (HTPG), assume that the categorical variable is the result of the truncation of underlying latent variables. In practice, only the categorical variable is observed. This translates the practical application of HTPG into a missing data problem in which all latent variables are missing. The latent variables are required at data locations in order to condition categorical realizations to the observed categorical data. The imputation of missing latent variables at data locations is often achieved by either assigning valid constant values or spatially simulating latent variables subject to categorical observations. Realizations of latent variables can be used to condition all model realizations. Using a single realization or a constant value to condition all realizations is the same as assuming that the latent variables are known at the data locations and this assumption affects uncertainty near the data locations.

This chapter is focused on the techniques for imputation of latent variables in the truncated Gaussian framework, their impact on uncertainty of simulated categorical models and possible effects on factors affecting decision making. It is shown that the use of a single realization of latent variables leads to underestimation of uncertainty and overestimation of measured resources while the use of constant values for latent variables may lead to considerable over or underestimation of measured resources. The results highlight the importance of multiple data imputation in the context of HTPG.

The techniques for data imputation in the truncated Gaussian framework are mostly based on the Gibbs sampler approach. The convergence of the Gibbs sampler algorithm can be problematic and a framework to mitigate the problems with convergence is proposed in this chapter. An alternative data imputation approach based on simulated annealing is also proposed.

5.1 Introduction

In practice the categorical variable is observed and the underlying latent variables are not; therefore, the assignment of the latent variable values in truncated pluri-Gaussian simulation (TPGS) is a missing data problem (Little and Rubin, 2002). The assignment of Gaussian values to the sample data locations is achieved by either simulating the unknown Gaussian values using the right spatial structure (Arroyo et al., 2012; Astrakova et al., 2015; Emery et al., 2014; Galli et al., 1994; Lantuéjoul and Desassis, 2012) or assigning the centroid of each categorical class with respect to the truncation rule (Rossi and Deutsch, 2014). The latter methodology generates data with exaggerated short range continuity that does not match the variogram of latent variables used for modeling. A single or multiple realizations of the Gaussian latent variables at each data location can be used to condition all model realizations.

Using fixed values for the missing latent variables is the same as assuming that they are known (sampled). This will likely result in incorrect uncertainty assessment as, in reality, the values are unknown. There are multiple realizations for the latent variables that result in the same observed categorical data with the same spatial structure and the same truncation rule. The missing data problem for spatially correlated geological variables is investigated by Barnett and Deutsch (2015) in the context of modeling continuous variables. This chapter is focused on the investigation of the impact of the Gaussian data assignment and multiple data imputation on the uncertainty of categorical models and the practical impact on uncertainty and resource classification affecting decision making. The discussed topics and results presented in this chapter are not only important for the development of the HTPG approach, but are also relevant to the application of all truncated Gaussian techniques.

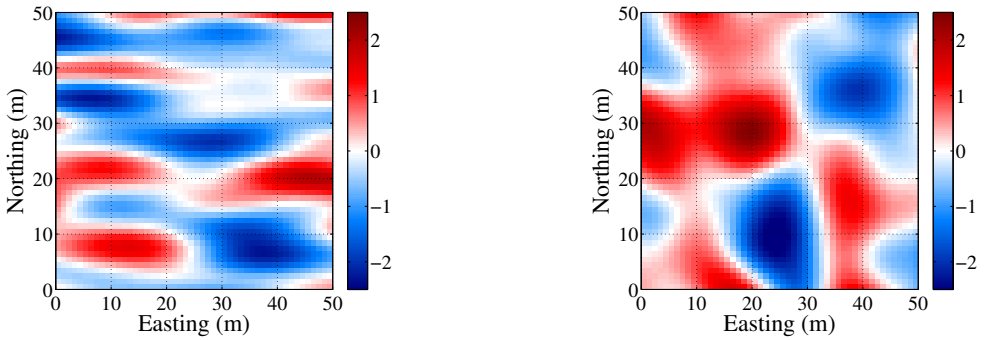
5.2 Simulating latent variables subject to categorical observations

The truncation rule controls the transition probability between categories and their proportions and it also controls the link between the continuity of the latent Gaussian variables and the continuity of the categorical variables. The current practice in the truncated Gaus-

sian methodologies is to define the truncation rule first in order to match the transition probabilities and proportions (Deutsch and Deutsch, 2014) and later define the variogram of the latent variables for the fixed mask to match the categorical variables continuity after truncation.

Simulating spatially correlated variables from a truncated multivariate Gaussian distribution is not trivial, however, the truncated univariate conditional distributions can be easily sampled. The Gibbs sampler algorithm (Geman and Geman, 1984) is the standard choice for simulation in this context as other alternatives such as rejection sampling are not practical (Armstrong et al., 2011). Variations of the Gibbs sampler algorithm for the problem of Gaussian data assignment for use with TPGS have been proposed and are reviewed here. Simulated annealing is a flexible methodology that can be used as an alternative to the Gibbs sampler for simulation of Gaussian random functions (GRFs) (Deutsch and Cockerham, 1994) subject to complex constraints. An algorithm for simulation of spatially correlated Gaussian latent variables subject to categorical data observations using simulated annealing is developed.

A synthetic 2D example is used to illustrate the different simulation methods. Two independent GRFs are generated using unconditional LU simulation (Davis, 1987) to populate a grid of 50x50 cells of size 1x1 meter. The latent variable Y_1 has anisotropic Gaussian variogram with practical ranges of 32 and 8 meters in horizontal and vertical directions respectively while the latent variable Y_2 has isotropic Gaussian variogram with practical range of 16 meters. The resultant simulated fields are shown in Figure 5.1.

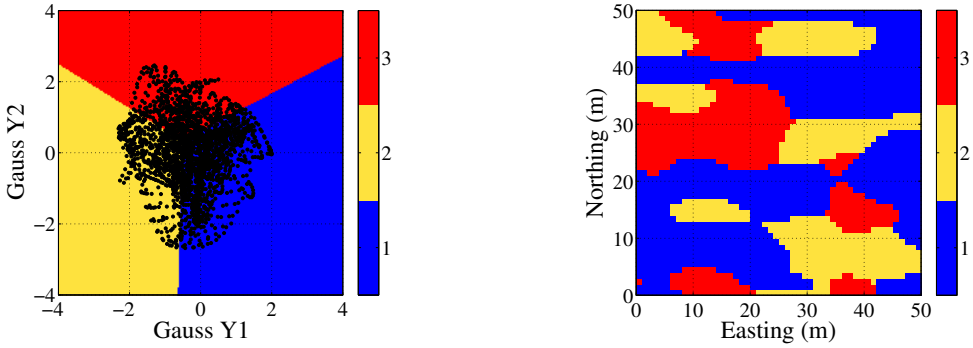


(a) GRF Y_1 **(b)** GRF Y_2

Figure 5.1: Reference GRF's generated for the example

A truncation rule similar to the ones obtained by the multidimensional scaling (MDS) based methodology (Deutsch and Deutsch, 2014) is used to truncate the two random vari-

ables Y_1 and Y_2 into three categorical classes (Figure 5.2). The truncation rule with the simulated bivariate cloud is shown in Figure 5.2a and the resultant categorical model after truncation is shown in Figure 5.2b.



(a) Truncation mask and simulated points **(b)** Truncated categorical model

Figure 5.2: Truncation rule and simulated categorical model

5.2.1 Gibbs sampler algorithm

Consider a set of n observations $\boldsymbol{x} = (x(\boldsymbol{u}_1), \ldots, x(\boldsymbol{u}_n))$ of a categorical random variable $X(\boldsymbol{u})$ with a finite set $\mathcal{B}$ of possible categories, where $\boldsymbol{u}$ is the vector of spatial coordinates. Also, consider the set of K latent variables at all data locations to be represented by $\boldsymbol{y} = \left(\boldsymbol{y}^\top(\boldsymbol{u}_1), \ldots, \boldsymbol{y}^\top(\boldsymbol{u}_n)\right)$ where $\boldsymbol{y}(\boldsymbol{u}_i) \in \mathbb{R}^K$ and is a realization of the Gaussian random variable $\boldsymbol{Y}(\boldsymbol{u}_i) = (Y_1(\boldsymbol{u}_i), \ldots, Y_D(\boldsymbol{u}_i))$ at i^{th} data location. The latent variables $\boldsymbol{Y}(\boldsymbol{u}_i)$ $(i = 1, \ldots, n)$ are assumed to have independent components with zero mean and unit variance $(\boldsymbol{Y}(\boldsymbol{u}_i) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\ \forall i \in \{1, \ldots, n\})$. For notation simplicity the vector of spatial coordinates $\boldsymbol{u}_i$ will be omitted and represented by the subscript i. Finally, consider the truncation rule to be represented by $\mathcal{M}_\theta$ which defines the map $M_\theta : \mathbb{R}^K \mapsto \mathcal{B}$ such that $\mathcal{M}_\theta(\boldsymbol{y}_i) = x_i$ $(i = 1, \ldots, n)$, where θ is the set of parameters that define the truncation rule. The standard Gibbs sampler algorithm proceeds by:

(1) Set iteration counter to zero: $t = 0$

(2) Initialize the set of latent variables $\boldsymbol{y}^{(t)}$ with arbitrary values subject to the mapping constraint: $\mathcal{M}_\theta(\boldsymbol{y}^{(t)}) = \boldsymbol{x}$

(3) Increment iteration counter: $t = t + 1$

(4) Set $\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)}$

(5) Define a random path through data points

(6) Loop for each variable $j = 1, \ldots, K$

(6.1) For each data location in the random path, update its value by a new Gaussian value drawn from the conditional distribution given the surrounding data values.

$$y_{j,i}^{(t)} \sim \mathcal{N}\left(\tilde{\mu}_{j,i}, \tilde{\sigma}_{j,i}\right) \tag{5.1}$$

$$\tilde{\mu}_{j,i} = \sum_{l \in N_i} \lambda_{l,i}^{(j)} \times y_{j,l}^{(t)} \tag{5.2}$$

$$\tilde{\sigma}_{j,i}^2 = 1 - \sum_{l \in N_i} \lambda_{l,i}^{(j)} \times C_{l,i}^{(j)} \tag{5.3}$$

$$\sum_{l \in N_i} \lambda_{l,i}^{(j)} \times C_{l,m}^{(j)} = C_{m,i}^{(j)} \qquad \forall m \in N_i \tag{5.4}$$

where N_i represents the neighboring locations, which can be constant (all neighbors) or it could be restricted by a search around the i^{th} location. The latter results in an approximation of the conditional mean and variance and may affect convergence of the algorithm. $C_{l,m}^{(j)}$ is the spatial covariance between locations l and m for variable j. Note that the Equations 5.2 and 5.3 are derived under an assumption of zero mean and unit variance.

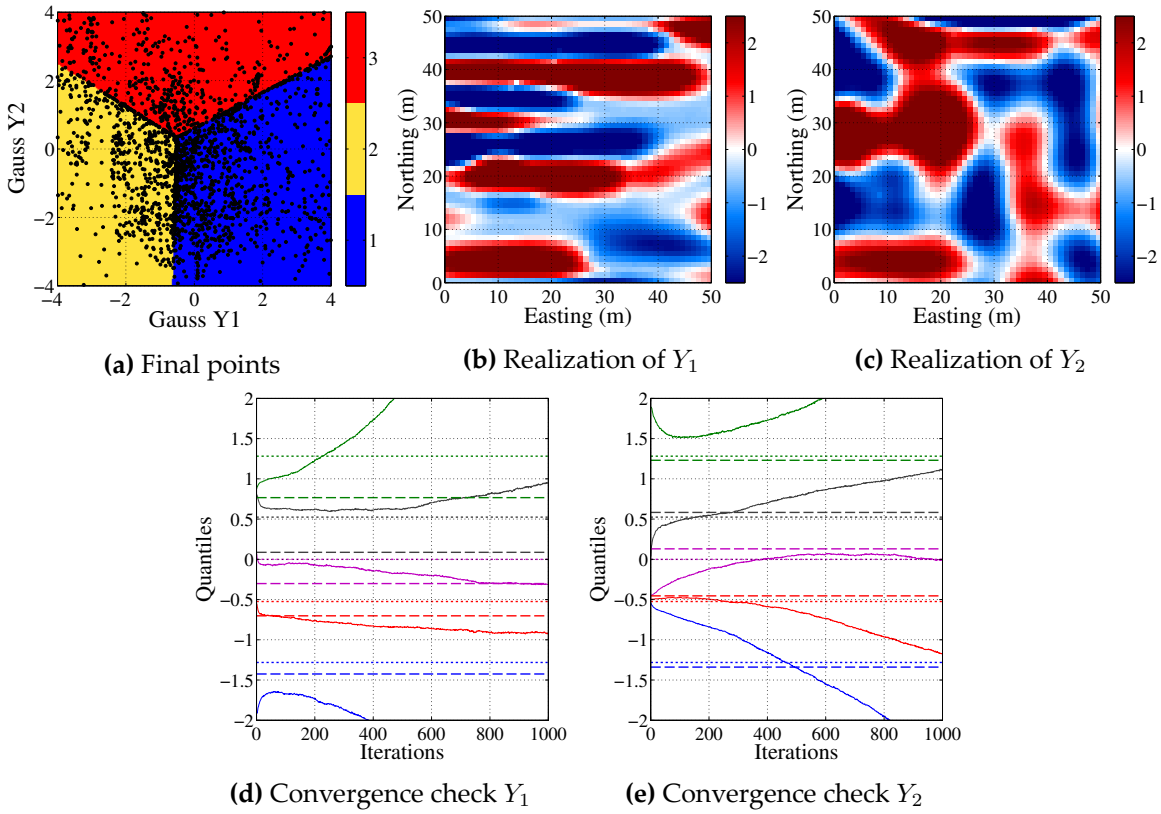
(6.2) Test if the updated value of $y_{j,i}^{(t)}$ satisfies the condition $\mathcal{M}_\theta(\boldsymbol{y}_i^{(t)}) = x_i$. If the mapping is not satisfied, the sampling is repeated until the conditions are met before moving to the next data point in the random path. Note that depending on the map $\mathcal{M}_\theta$ the boundaries for $y_{j,i}^{(t)}$ can be calculated in advance and applied to constrain the distribution in Equation 5.1 to avoid multiple sampling attempts.

(7) Repeat steps from 3 for a maximum number of iterations.

Solving the system of equations (Equation 5.4) considering all neighboring data requires the inverse of a covariance matrix of rank $n - 1$ and quickly becomes impractical with increasing number of data. For practical purposes the size of the covariance matrix must be reduced. An arbitrary number of closest points (N_i) are used to calculate the mean and variance of conditional distributions used in Gibbs iteration. This approximation af-

fects the convergence of the algorithm (Astrakova et al., 2015; Emery et al., 2014; Lantuéjoul and Desassis, 2012).

The Gibbs sampler with restricted neighborhood is run for the 2D example presented in the previous section. The results for 1,000 iterations using nearest 24 neighbors are shown in Figure 5.3. The quantiles of the realization after each iteration are shown in the Figures 5.3d and 5.3e and are compared with the quantiles from $\mathcal{N}(0,1)$ distribution and the reference model (Figure 5.1). It is clear that the values move towards extreme highs and lows with increasing number of iterations.



(a) Final points **(b)** Realization of Y_1 **(c)** Realization of Y_2

(d) Convergence check Y_1 **(e)** Convergence check Y_2

Figure 5.3: Results for 1,000 iterations of the standard Gibbs sampler with restricted neighborhood (nearest 24). It is clear that the realizations do not converge. Items (d) and (e) show the quantiles of the realizations (solid lines), the quantiles from $\mathcal{N}(0,1)$ (dotted lines), and the quantiles from the reference model (dashed lines). The quantiles highlighted correspond to the standard Gaussian percentiles 0.1, 0.3, 0.5, 0.7 and 0.9.

5.2.2 Propagative Gibbs sampler

Galli and Gao (2001) proposed an alternative that does not require the calculation of the inverse of the covariance matrix, therefore, allowing the use of all neighbors even for a large

number of data observations. Lantuéjoul and Desassis (2012) built on the work of Galli and Gao (2001) and introduced the propagative Gibbs sampler. Emery et al. (2014) suggested an algorithm that uses the propagative Gibbs sampler to simulate Gaussian random functions subject to inequality constraints. The latter algorithm is described below:

(1) Set iteration counter to zero: $t = 0$

(2) Initialize the set of latent variables $\boldsymbol{y}^{(t)}$ with arbitrary values subject to the mapping constraint: $\mathcal{M}_\theta(\boldsymbol{y}^{(t)}) = \boldsymbol{x}$

(3) Increment iteration counter: $t = t + 1$

(4) Define a random path through data points

(5) Loop for each variable $j = 1, \ldots, K$

(5.1) Select a pivot i according to the random path and update it according to Equation 5.5.

$$y_{j,i}^{(t)} \sim \mathcal{N}(0, 1) \tag{5.5}$$

(5.2) The new pivot value is used to update all non-pivot locations according to Equation 5.6.

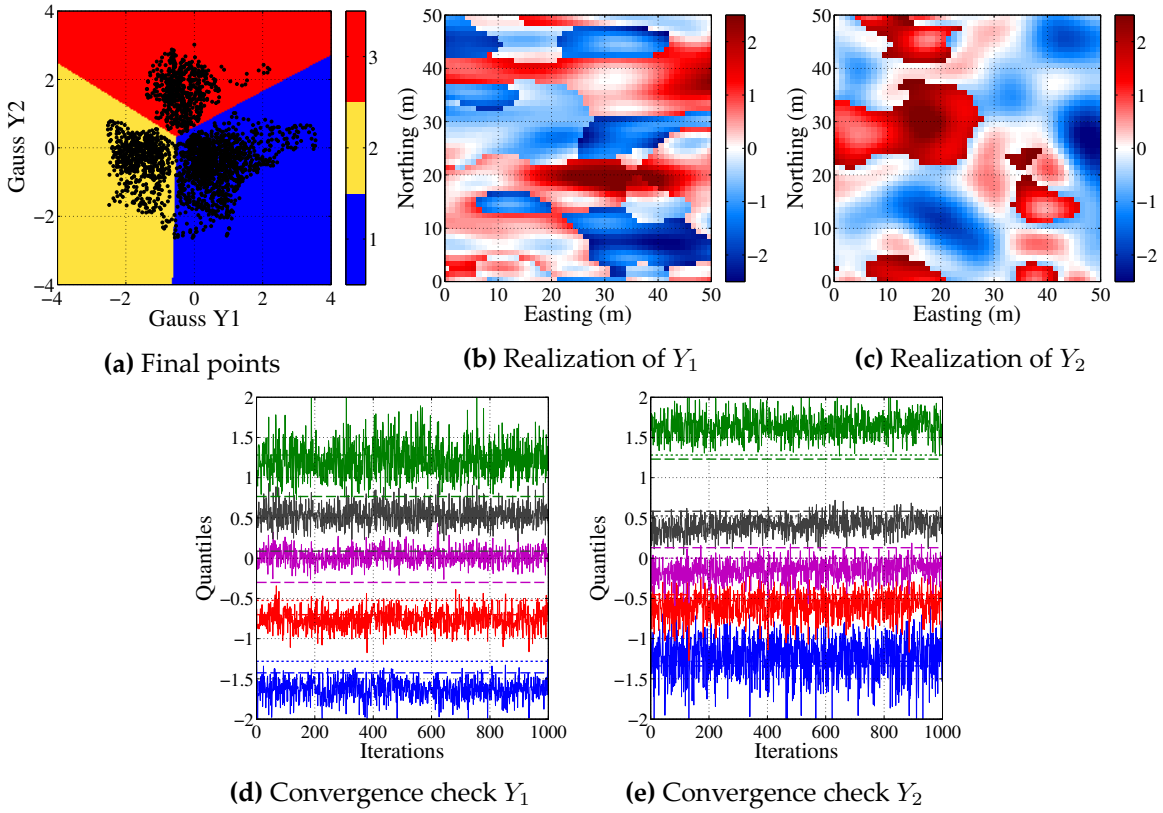
$$y_{j,l}^{(t)} = y_{j,l}^{(t-1)} + \left(-y_{j,i}^{(t-1)} + y_{j,i}^{(t)}\right) C_{i,l}^{(j)} \qquad \forall l \neq i \tag{5.6}$$

(5.3) Test if the updated values satisfy the condition $\mathcal{M}_\theta(\boldsymbol{y}^{(t)}) = \boldsymbol{x}$. If the mapping is not satisfied, the sampling is repeated until the conditions are met before moving to the next data point in the random path. Depending on $\mathcal{M}_\theta$, it may be possible to define boundaries for $y_{j,i}^{(t)}$ in advance and use them to constrain the sampling in Equation 5.5.

(6) Repeat steps from 3 for a maximum number of iterations.

This algorithm is shown to work well for a single latent variable with simple truncation map (Emery et al., 2014). Lantuéjoul and Desassis (2012) found experimentally that initializing the algorithm with zeroes improves convergence. This may not be possible when simulating latent variables subject to complex constraints. In fact, for the 2D example presented here, the algorithm failed to converge after a considerable number of iterations

(Figure 5.4). The algorithm is stable in the sense that it does not diverge to extreme values, however, it also does not converge to the correct covariance of latent variables across different rock types within a reasonable number of iterations.



(a) Final points **(b)** Realization of Y_1 **(c)** Realization of Y_2

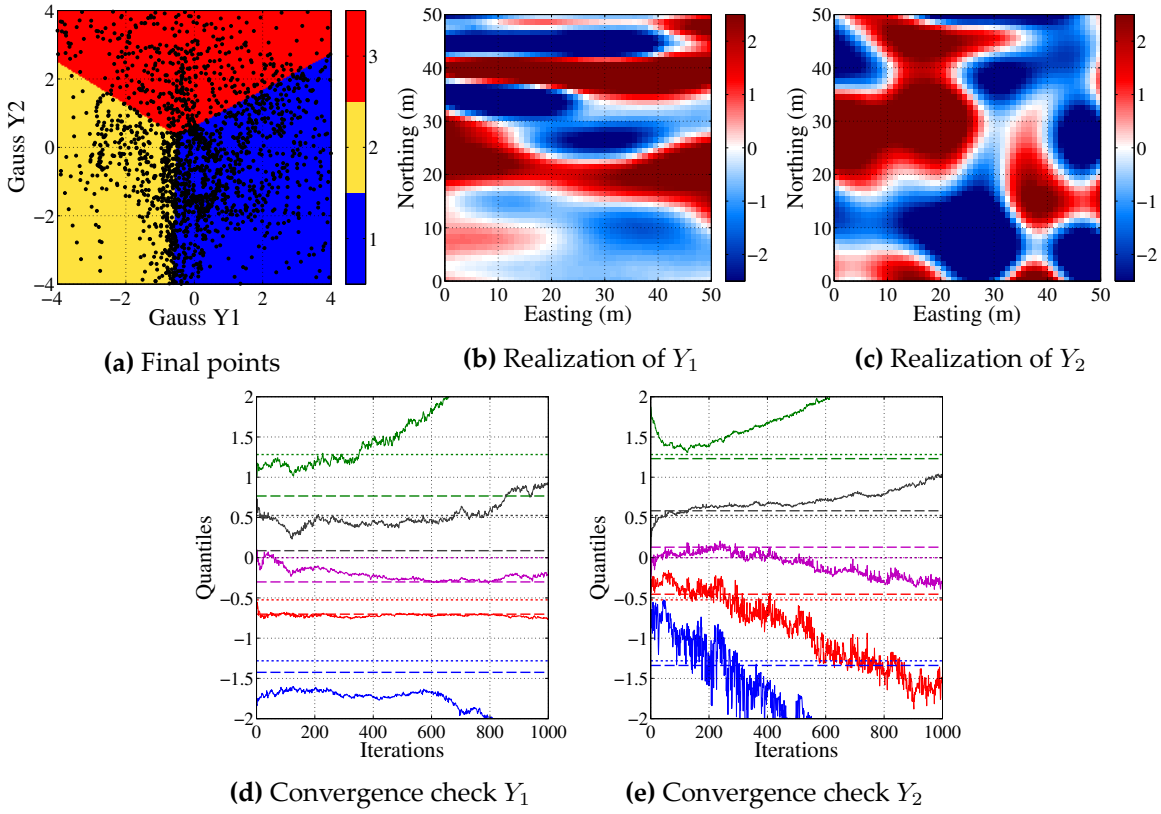
(d) Convergence check Y_1 **(e)** Convergence check Y_2

Figure 5.4: Results for 1,000 iterations of the propagative Gibbs sampler. It is clear that the realizations do not converge to a proper bivariate Gaussian distribution in (a) and show clear hard boundaries between some categories (b and c). Items (d) and (e) show the quantiles of the realizations (solid lines), the quantiles from $\mathcal{N}(0, 1)$ (dotted lines), and the quantiles from the reference model (dashed lines). The quantiles highlighted correspond to the standard Gaussian percentiles 0.1, 0.3, 0.5, 0.7 and 0.9.

5.2.3 Combined Gibbs sampler

The combined Gibbs sampler is proposed by Astrakova et al. (2015) and consists of combining iterations of the standard and propagative Gibbs sampler in order to improve convergence. Astrakova et al. (2015) propose to use several iterations of the standard Gibbs sampler at the start and then proceed with alternating standard and propagative algorithms. The same problems with convergence shown in Figure 5.3 for the standard algorithm is also noticed for alternating the propagative and standard Gibbs sampler (Figure 5.5).

It seems reasonable to use the standard Gibbs sampler at early iterations to improve



(a) Final points **(b)** Realization of Y_1 **(c)** Realization of Y_2

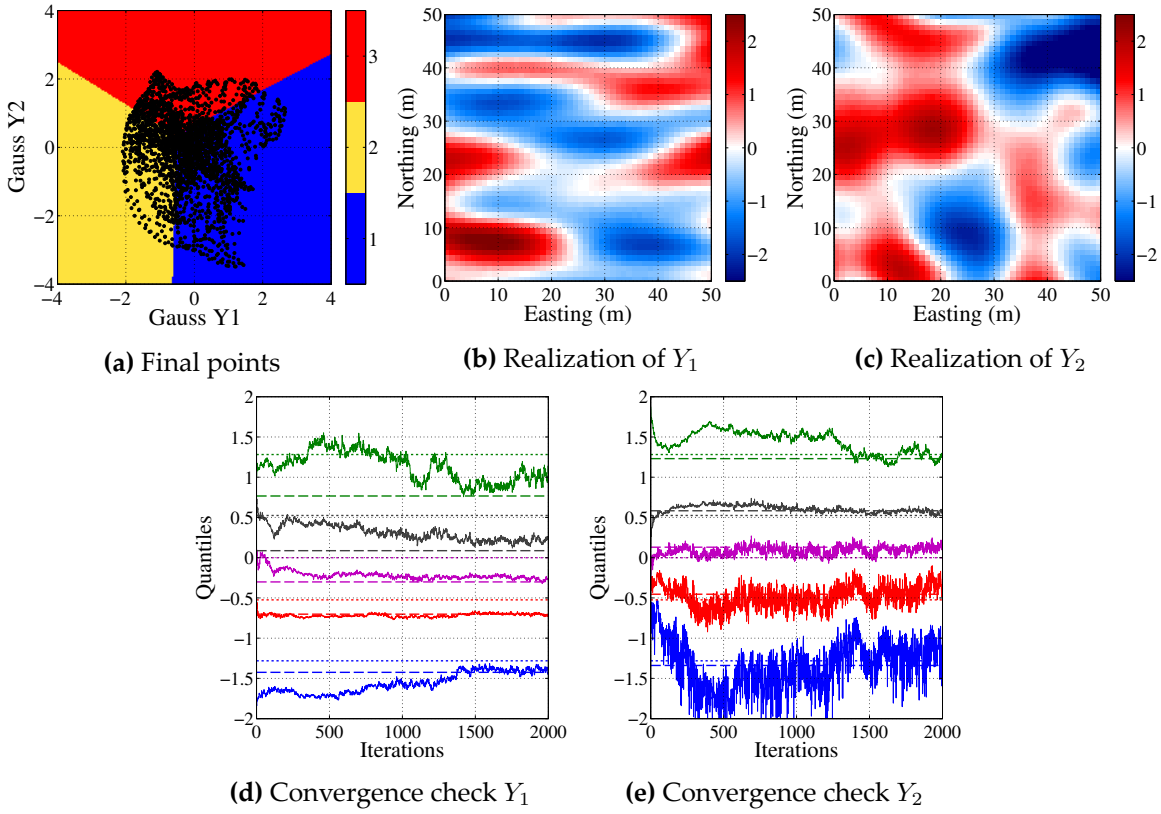
(d) Convergence check Y_1 **(e)** Convergence check Y_2

Figure 5.5: Results for 1,000 iterations of the combined algorithm alternating between the standard and propagative Gibbs sampler. The same problems with convergence is found by alternating the two algorithms. Items (d) and (e) show the quantiles of the realizations (solid lines), the quantiles from $\mathcal{N}(0, 1)$ (dotted lines), and the quantiles from the reference model (dashed lines). The quantiles highlighted correspond to the standard Gaussian percentiles 0.1, 0.3, 0.5, 0.7 and 0.9.

the spatial correlation between locations with different categorical observations, however, continuing with the alternating algorithm does not result in a stable sequence. A second run of the combined Gibbs sampler is attempted, but this time the standard Gibbs sampler iterations are ceased after iteration 400 and only the propagative algorithm is used beyond that iteration. The results are shown in Figure 5.6. It is clear from the Figures 5.6d and 5.6e that this methodology generated a sequence that remains stable even after a large number of iterations (2,000).

5.2.4 Simulated annealing

A methodology similar to the one described by Deutsch and Cockerham (1994) is proposed here for the simulation of spatially correlated latent variables subject to categorical data observations. Modifications to the algorithm are proposed to account for the restrictions



(a) Final points **(b)** Realization of Y_1 **(c)** Realization of Y_2

(d) Convergence check Y_1 **(e)** Convergence check Y_2

Figure 5.6: Results for 2,000 iterations of the combined algorithm alternating between the standard and propagative Gibbs sampler up to 400 iterations and proceeding with propagative Gibbs sampler only. Items (d) and (e) show the quantiles of the realizations (solid lines), the quantiles from $\mathcal{N}(0,1)$ (dotted lines), and the quantiles from the reference model (dashed lines). The quantiles highlighted correspond to the standard Gaussian percentiles 0.1, 0.3, 0.5, 0.7 and 0.9.

given by the truncation rule and to allow its use for simulation of scattered data locations (not gridded). The algorithm is described bellow:

(1) Set iteration counter to zero: $t = 0$

(2) Initialize the set of latent variables $\boldsymbol{y}_i^{(t)} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\ \forall i \in \{1, \ldots, n\}$ using rejection sampling in order to satisfy the map $\mathcal{M}_\theta(\boldsymbol{y}^{(t)}) = \boldsymbol{x}$

(3) Increment iteration counter: $t = t + 1$

(4) Select a random index $i \in \{1, \ldots, n\}$

(5) Select a second random index l such that $l \neq i$ and $x_i = x_l$

(6) Update the latent variables by swapping the location i with location l while maintaining all other locations unchanged.

$$\boldsymbol{y}_i^{(t)} = \boldsymbol{y}_l^{(t-1)} \tag{5.7}$$

$$\boldsymbol{y}_l^{(t)} = \boldsymbol{y}_i^{(t-1)} \tag{5.8}$$

$$\boldsymbol{y}_m^{(t)} = \boldsymbol{y}_m^{(t-1)} \qquad \forall m \neq l \text{ and } m \neq i \tag{5.9}$$

(7) Evaluate the objective function

$$\mathrm{O}\left(\boldsymbol{y}^{(t)}\right) = \sum_{h=1}^{H} \left[\gamma_h - \hat{\gamma}_h\left(\boldsymbol{y}^{(t)}\right)\right]^2 \tag{5.10}$$

where γ_h is the target variogram, $\hat{\gamma}_h$ is the experimental variogram and H is a set of lag vectors.

(8) Draw a random number from a uniform distribution $u \sim \mathcal{U}(0,1)$

(9) Calculate temperature

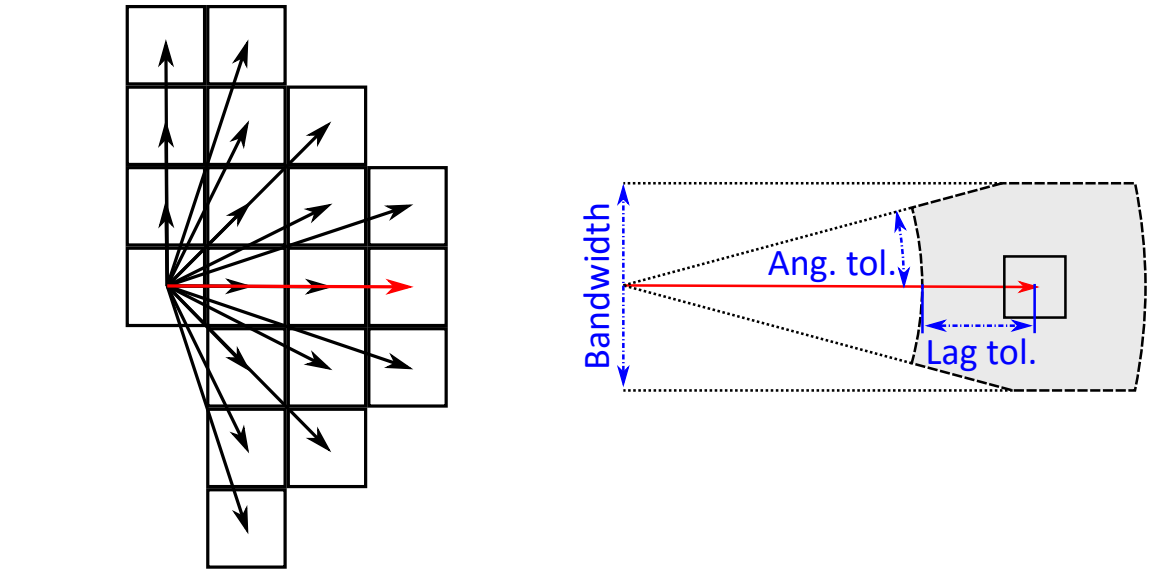
$$t^{(t)} = t^{(0)}\alpha^k \qquad \forall k > 0,\ \alpha \in]0,1[\text{ and } t^{(0)} \in]0,+\infty[\tag{5.11}$$

(10) Accept changes if $u < \exp\{\frac{\mathrm{O}(\boldsymbol{y}^{(t)}) - \mathrm{O}(\boldsymbol{y}^{(t-1)})}{t^{(t)}}\}$, otherwise reject changes and reset values to $\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)}$

(11) Repeat steps from 3 for a maximum number of iterations.

The objective function (Equation 5.10) for the proposed algorithm is the mismatch of spatial structure defined by the variogram model and the spatial structure of the assigned Gaussian values. The spatial structure must be reproduced for all directions and different distance lags. A similar strategy as the one presented in Deutsch and Cockerham (1994) is used in order to ensure the spatial structure. A set of different angles and lags are defined within a certain range (Figure 5.7a) and for each of these lag vectors a search tolerance is allowed in order to make the calculation more stable for scattered data points (Figure 5.7b).

The proposed simulated annealing algorithm is run for 500,000 iterations for the same 2D example. The results are shown in Figure 5.8. Note that the bivariate scatter plot shown in Figure 5.8a is generated at initialization and remains unchanged throughout the iterations. The final realizations shown in the Figures 5.8b and 5.8c have an increased short scale variability, however, this might not be of great concern for the simulation of scattered data points for the purpose of conditioning TPGS realizations. The simulated annealing algorithm represents a viable alternative for cases in which the truncation rule is too complex

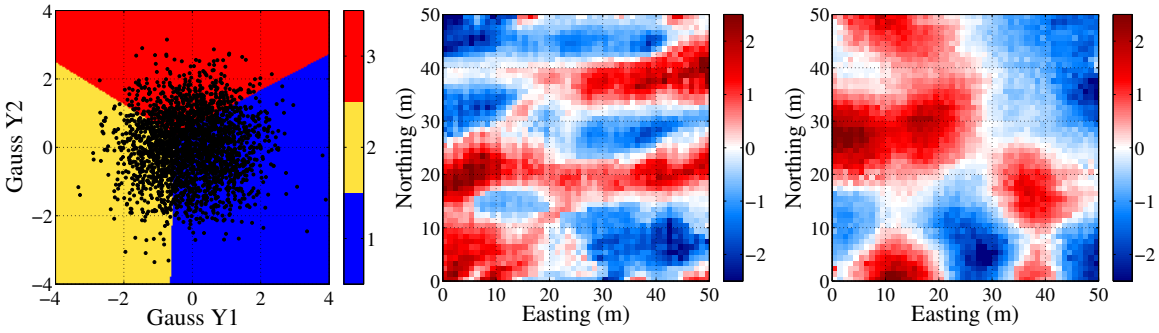


(a) Variogram volume (modified from Deutsch and Cockerham (1994))

(b) Search tolerance

Figure 5.7: Lag vectors and tolerance parameters used for experimental variogram calculation that goes into the objective function within simulated annealing implementation.

for use with Gibbs sampler and for cases in which other important factors are relevant to the simulation and can be easily added to the objective function of the simulated annealing algorithm. Examples of other factors that have been considered with annealing are two-point transition probabilities, seismic data, and multiple-point statistics (Deutsch and Cockerham, 1994).



(a) Final points

(b) Realization of Y_1

(c) Realization of Y_2

Figure 5.8: Results for 500,000 iterations of the simulated annealing algorithm.

An alternative approach based on simulated annealing is also proposed by Emery et al. (2014). The alternative is a modification of the propagative Gibbs sampler in which the constraints in the Gaussian distribution are not directly enforced by the Gibbs sampler

iterations. The simulated annealing targets the minimization of the mismatch between simulated and observed categorical data by penalizing latent variables that are simulated beyond the constraints. The results from this approach may not match the observed data. The simulated annealing approach is CPU intensive and the example in Figure 5.8 takes 20 minutes to run with an Intel® Core™ i7-4790K CPU @ 4.00GHz.

5.3 Impact of multiple data imputation

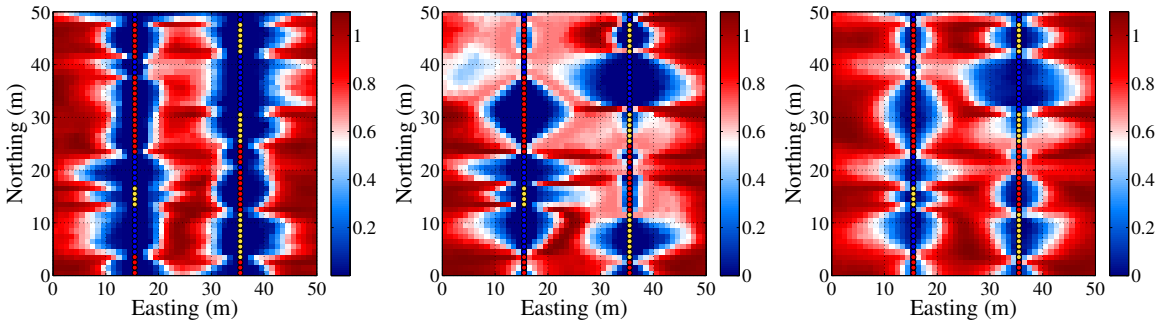
The impact of the different methods for the imputation of missing latent values to categorical data observations in the context of TPGS is investigated in this section. Centroid assignment, single realization and multiple realizations (multiple imputation) are compared here. It is clear that there is a great difference in terms of latent variable assignment, but it is not clear how these strategies affect the uncertainty of the final categorical realizations, which is what matters the most.

The uncertainty in the categorical data can be measured using Shannon's entropy (Christakos, 1990; Shannon, 1948) (Equation 5.12) and used to compare (visualize) the impact of the different data assignment strategies. The Shannon's entropy is adequate for the visualization of local variations of assessed uncertainty resulting from each method. It is also important to investigate the effect of data assignment to relevant practical factors affecting decision making such as resource evaluation and classification. This will provide insight on the global impact of the uncertainty assessment changes.

$$H = -\sum_{i \in C} p_i \log p_i \tag{5.12}$$

Samples of the 2D synthetic model are used to illustrate the effects of the different assignment strategies on the uncertainty of categorical simulations. The Shannon entropy is calculated for 400 TPGS categorical realizations using the three data assignment strategies. The result is shown in Figure 5.9. It is clear that the assessment of uncertainty using simulation with different data assignment methodologies is drastically affected. The uncertainty is changed mostly close to boundaries between categories and in between the drillholes.

Using multiple imputation as the reference, the single realization method consistently underestimates uncertainty. This is an expected result as fixing the data values will translate to less variability between the realizations. The centroid assignment methodology re-



(a) Centroid **(b)** Single **(c)** Multiple

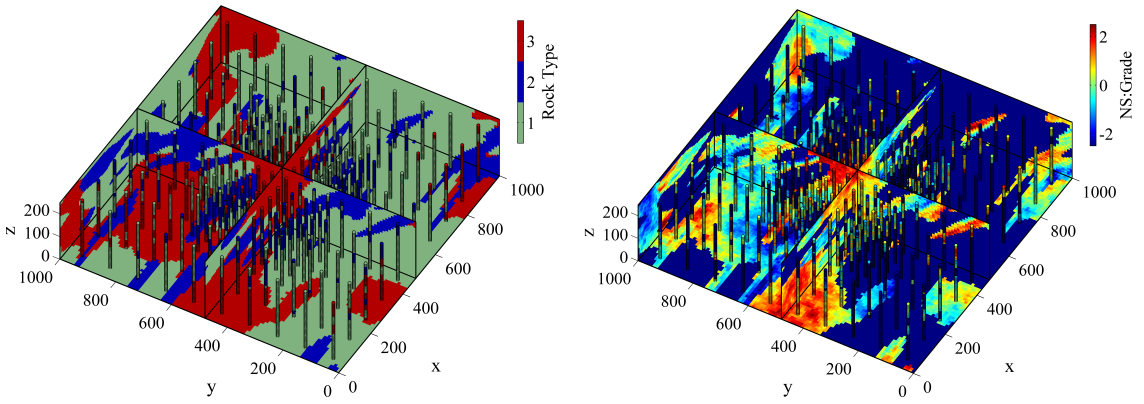
Figure 5.9: Uncertainty of categorical realizations for the 2D example. Colored markers represent the categorical samples from reference model. The gridded model is colored by the Shannon entropy value calculated using 400 categorical realizations from TPGS

sults in less variability close to the data locations specially at boundaries and between the drillholes. At first glance, the increased variability in those areas while using the centroid method is unexpected as multiple imputation leads to varying conditioning data for TPGS and therefore is expected to result in highest uncertainty on categorical realizations when compared to fixed conditioning data. The reason for the larger variability is attributed to the fact that the latent variables at the boundaries between categories are far from what they should be; they are at the centroid of the class and not close to the boundary of another class. This artificially increases the uncertainty in specific areas of the model.

A 3D synthetic model is used to illustrate the impact of using different latent data imputation methodologies on resource evaluation and classification. The reference model is built with the simulation of two latent variables with Gaussian variograms. The first has major direction dipping at 45^o, with azimuth of 90^o and no tilt. The ranges are 300 meters for both major and mid directions and 50 meters for minor direction. The second latent variable has an isotropic structure with range 300 meters. The reference model is built on a grid of 100x100x50 blocks of size 10x10x5 meters. The simulated latent variables are trimmed to generate a reference categorical model shown in Figure 5.10a.

In order to evaluate the resources, a grade variable is simulated per rock type where rock type 1 (Figure 5.10a) is barren rock, rock type 2 and rock type 3 are mineralized with the same average grade but different variability and spatial structure. The grade distribution in rock type 2 follows a lognormal distribution with mean of 0.75% and variance of $0.125\%^2$. The grade distribution in rock type 3 follows a lognormal distribution with mean of 0.75% and variance of $0.0625\%^2$. The grade variable in rock type 2 has a spherical var-

iogram with major direction dipping at 45º angle and with azimuth of 90º and no tilt. It has 300 meter range in major and mid directions, and 25 meter range in minor direction. The grade variable in rock type 3 has isotropic spherical variogram with range of 300 meters. Both categorical and grade models are sampled with a regular drilling pattern with spacing of 100x100 meters closer to the border and 50x50 meters in the central part of the domain. The reference models and samples are shown in Figure 5.10b.

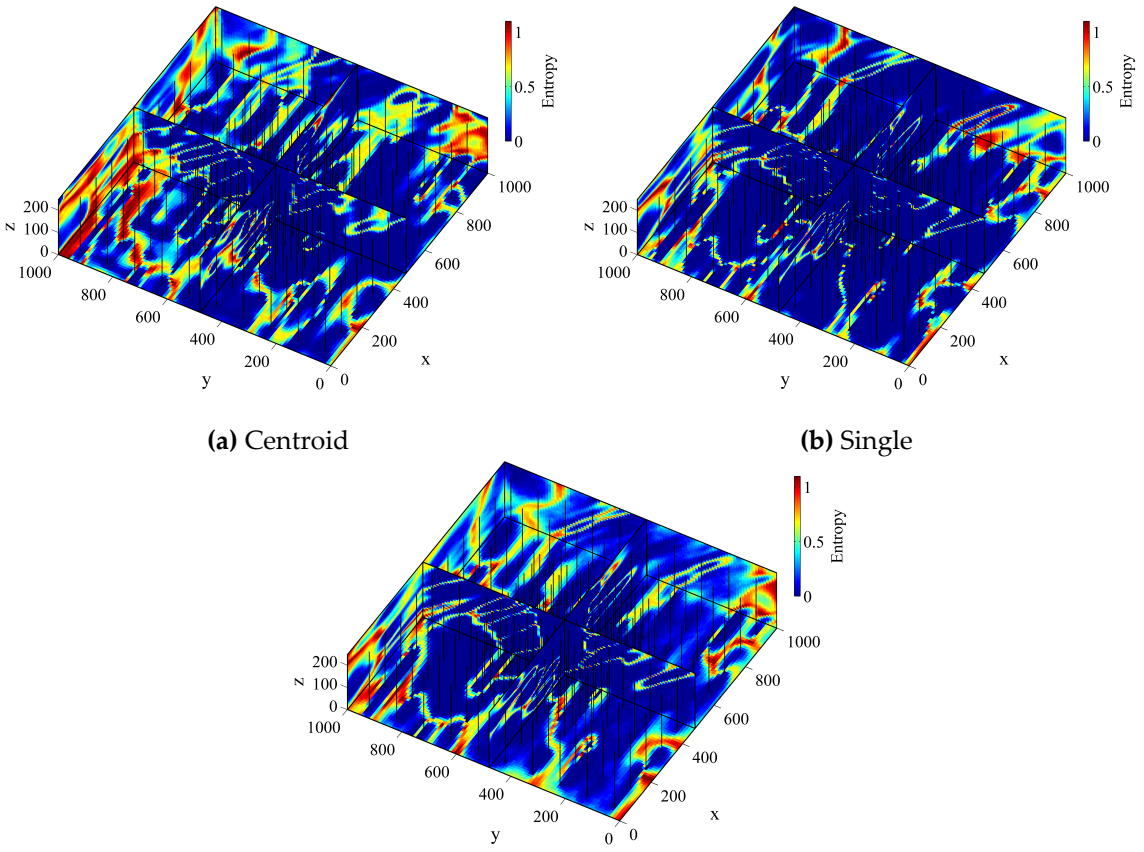


(a) Rock types **(b)** Grade model

Figure 5.10: Simulated 3D model. (a) simulated rock types. Category 1 is barren rock, category 2 and 3 are mineralized rock. (b) reference grade model in Gaussian units. Markers represent sample locations (165 drillholes)

Multiple imputation, single realization and centroid methods for assigning latent variables are used in TPGS workflow for the generation of 400 realizations of rock type for this 3D example. The resultant Shannon's entropy calculated for each methodology is shown in Figure 5.11. Similar results are obtained for the entropy compared to the 2D example. Using the multiple imputation (Figure 5.11c) as reference, the single realization (Figure 5.11b) consistently underestimates uncertainty whereas the centroid method (Figure 5.11a) results in underestimation of uncertainty in densely sampled locations and uncertainty overestimation at more sparsely sampled locations. This is likely the result of the wrong spatial correlation of the latent variables assigned through centroid method that shows hyper continuity at short range within the same category and increased short range variability across different categories due to the sharp transition between them.

The grade variable is also simulated by rock type for each realization. Resource classification was performed using probabilistic criteria following the procedure described by Silva and Boisvert (2014) using a production panel of size 50x50x25 meters. The classifi-



(a) Centroid

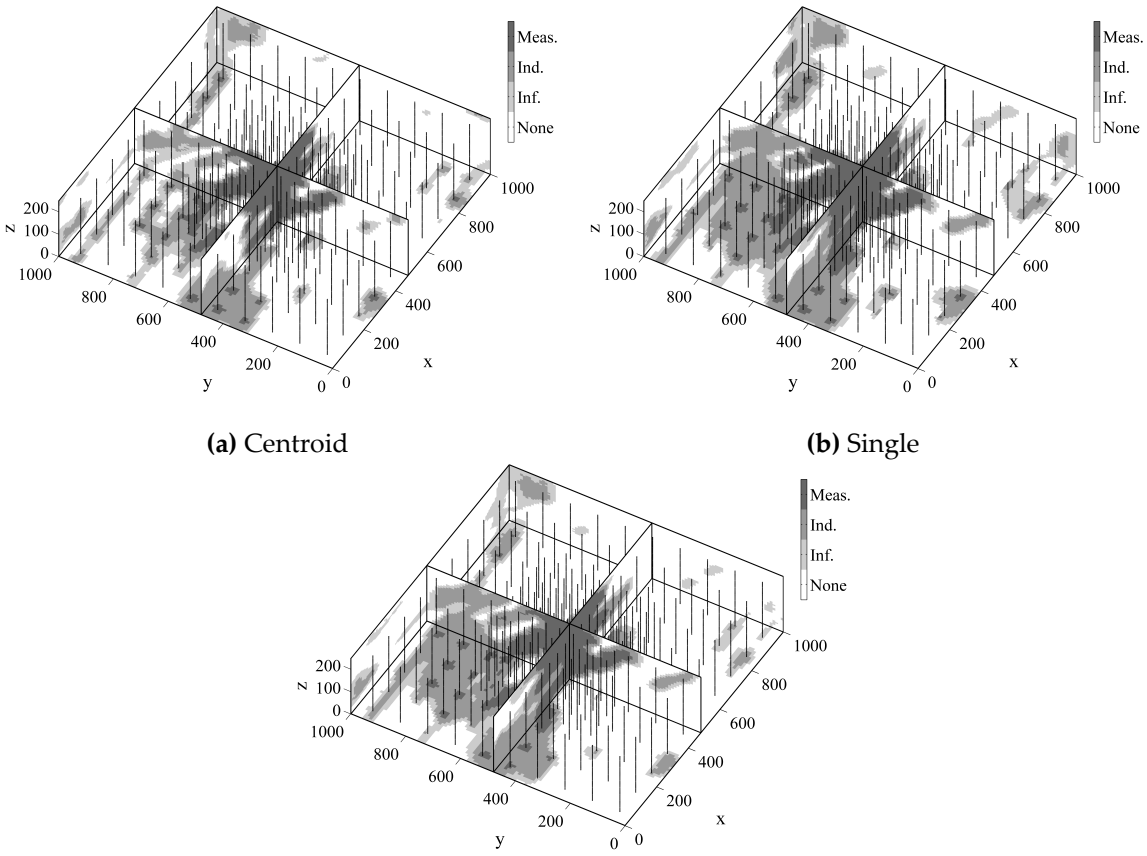
(b) Single

(c) Multiple

Figure 5.11: Uncertainty of categorical realizations for the 3D example. The gridded model is colored by the Shannon entropy values calculated using 400 realizations.

cation of resource is performed according to the following criteria: (1) Measured blocks must have simulated values within ±15% of the mean at least 95% of the time; (2) Indicated blocks must have simulated values within ±30% of the mean at least 95% of the time; and (3) Inferred blocks must be within ±50% of the mean at least 95% of the time.

The classification results shown in Figure 5.12 and summarized at Table 5.1. The results are consistent with the entropy maps. Relative to the multiple imputation methodology, the centroid has higher overall entropy and therefore consistently lower classified resources while single realization methodology has lower entropy and results on consistently higher classified resources. Variations of ±15% on measured plus indicated resources are substantial and have potential to impact economical evaluation of projects and resulting decisions. While the single realization method seems to always underestimate uncertainty, the centroid assignment method may under or overestimate uncertainty depending on the prevalence of higher and lower variability areas.



(a) Centroid

(b) Single

(c) Multiple

Figure 5.12: Results of resource classification for the 3D example.

The multiple imputation methodology is deemed the most accurate in assessing the uncertainty as it transfers the uncertainty of the missing latent variables to the model realizations. In light of the potential deviations on key results affecting decision making depending on the choice of imputation technique, the use of multiple imputation of missing latent variables in TPGS is advised.

Method	Resource Classification (Mt)			
	Measured	Indicated	Inferred	Measured + Indicated
Centroid	535 (0.94)	640 (0.75)	416 (1.14)	1,175 (0.83)
Single	681 (1.20)	928 (1.09)	329 (0.90)	1,609 (1.13)
Multiple	567 (1.00)	851 (1.00)	364 (1.00)	1,418 (1.00)

Table 5.1: Calculated resources for each data assignment methodology. The ratio relative to multiple imputation are shown in brackets. A density of 2.65 g/m^3 is assumed

5.4 Conclusion

The truncated Gaussian techniques are powerful tools for the simulation of spatially correlated categorical variables. They allow for the reproduction of proportions, two-point spatial continuity and transition probabilities and therefore allowing the simulation of categorical variables with complex ordering. The technique requires the use of latent variables which are not sampled and requires imputation.

Three imputation methodologies are compared. The imputation of constant values (centroid), single realization imputation and multiple realization imputation. The first method is simple, fast, and does not require any special algorithm, however, the imputed values will have the wrong spatial structure. The single and multiple realization of latent variables requires the utilization of simulation techniques which are able to reproduce spatial structure and to condition realization(s) to observed categorical variables at data locations.

The available simulation techniques for generating realizations of latent variables at data locations are reviewed. The combination of standard and propagative Gibbs sampler with limited number of standard iterations seems to be adequate for the simulation of latent variables conditioned to categorical observations. A technique based on simulated annealing is also proposed for cases in which Gibbs sampler cannot be applied or additional factors need to be accounted for in the objective function.

Multiple imputation of latent variables is considered to be the most suitable technique for this problem. It accounts for the fact that the latent variables are unknown and transfers the uncertainty from the missing data to the categorical model realizations. It is shown that centroid assignment leads to local under and over estimation of the categorical variable uncertainty as result of fixed data values and wrong spatial structure. It is also shown that centroid assignment could affect resource evaluation and classification which are key factors in decision making. The single realization imputation method leads to consistently lower uncertainty as the data has the right spatial structure, but it is not allowed to fluctuate to other possible states. Single realization leads to consistently higher measured and indicated resources.

The use of multiple imputation in TPGS latent variable assignment is advisable in light of the great differences that may occur from different imputation methodologies and the

potential impact that those differences may have on important decision making factors such as resource evaluation and classification.

Chapter 6

Data Imputation with Gaussian Mixture Models for Continuous Variables

The multiple data imputation is not only required for the truncated Gaussian methods, but it is also required for advanced multivariate geostatistical techniques that relies on multivariate data transformations. The research undertaken in the quest to advance and improve the imputation for the application with hierarchical truncated pluri-Gaussian (HTPG) also resulted in important contributions to multivariate data imputation for multivariate modeling of continuous variables.

Even though the truncated Gaussian methods utilize a single parametric distribution, the categorical conditioning and truncation structure makes it impossible to directly simulate from those distributions requiring the utilization of Gibbs sampler. The simulation of continuous variables for multivariate transformations, on the other hand, do not have the complication of the truncation thresholds. The complexity faced on the imputation of continuous variables for multivariate transformations relates to the shape of the multivariate distribution that is often far from being Gaussian or easily parameterized.

Samples with missing variables are common in geological data sets for many reasons. The missing data must be imputed (inferred) to permit the measured data to be used to their full extent. Imputation methods for geological data should address spatial structure and multivariate complexity. The published techniques that account for these considerations make strong assumptions regarding conditional distributions and are computationally demanding in presence of many data.

A Gaussian mixture model (GMM) fitted to the multivariate data is developed to provide stability in fitting multivariate data and to significantly improve computational efficiency. The approach is demonstrated using a lateritic nickel data set and it is shown to decrease computational time by two orders of magnitude for the example while also consistently improving results in several performance tests.

6.1 Introduction

The developed methodology borrows many aspects of the non-parametric approach proposed by Barnett and Deutsch (2015). The key difference is in the calculation of the conditional distributions given the collocated data for the non-parametric Bayesian updating (BU). These conditional distributions are calculated from a GMM fitted to the multivariate data set. The use of a GMM allows the quick assessment of any marginal and conditional distributions needed for the data imputation workflow improving the speed of the algorithm. The computational expense would be significantly reduced. Moreover, the ability to use unequally sampled locations to inform the fitting of the GMM and the improved robustness against data sparseness leads to more accurate estimation of the multivariate distributions resulting in enhanced performance.

The technique developed here is considered as semi-parametric as opposed to non-parametric as it relies on the mixture of parametric distributions. This methodology is suited to cases where the missingness is independent on the missing values and may be dependent on the observed values. This missing data mechanism is known as missing at random (MAR). The applicability and performance of the methodology is demonstrated for a lateritic nickel data set.

6.2 Background

The Gaussian distribution is extremely tractable and fully parametrized by a mean vector and covariance matrix. Although a single Gaussian model cannot capture the complex relationships that high dimensional geological data often show, a GMM is able to fit the distributions of many multi-dimensional data sets. The most widely used GMM for density estimation is the kernel density estimation (KDE) (Gray and Moore) which has been studied since the 1950's (Fix and Hodges Jr, 1951; Silva and Deutsch, 2015). KDE becomes computationally expensive with increasing number of dimensions and observations as each observation is the center of a Gaussian kernel. One Gaussian kernel per observation is not always required and similar results can be obtained with fewer kernels centered at arbitrary locations that are able to summarize the information of several observations into a single kernel allowing the definition of the multivariate distribution with considerable reduction in the number of kernels/components. Choosing a relatively small number of kernels not

necessarily centered at data locations leads to the GMM formalism.

6.2.1 Expectation Maximization Algorithm

The basic form of the expectation maximization (EM) algorithm for fitting GMMs is well known within the scientific community. This application requires that the EM algorithm be adapted for missing data problems and, therefore, a concise description is presented. Interested readers are referred to McLachlan and Krishnan (2008) and McLachlan and Peel (2000) for a more detailed description of the algorithm. Further details on the missing data (heterotopic samples) problem can be found in Little and Rubin (2002).

Considering a set of n K-dimensional observations $\boldsymbol{y} = \left(\boldsymbol{y}_1^\top, ..., \boldsymbol{y}_n^\top\right)$, the density function estimated by EM algorithm with g Gaussian components is given by Equation 6.1.

$$\hat{f}\left(\boldsymbol{y}_j; \boldsymbol{\Psi}\right) = \sum_{i=1}^{g} \pi_i \phi\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i\right), \qquad \forall j = 1, \ldots, n \tag{6.1}$$

where $\hat{f}$ is the estimated distribution; ϕ is the multivariate Gaussian probability density function (PDF); $\boldsymbol{\Psi}$ is the set of unknown parameters $(\pi_1, ..., \pi_{g-1}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_g)$: weights to each component, the mean values and the variance-covariance matrices. The EM algorithm maximizes the log likelihood $\log L\left(\boldsymbol{\Psi}\right)$ that can be calculated for a given set of observations as:

$$\log L\left(\boldsymbol{\Psi}\right) = \sum_{j=1}^{n} \log \left[\sum_{i=1}^{g} \pi_i \phi\left(\boldsymbol{y}_j; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i\right)\right] \tag{6.2}$$

The algorithm consists of two steps which are the expectation (E) and maximization (M) steps that iteratively maximizes the log likelihood. Each observation is deemed to have originated from a respective component of the mixture, but this information is unknown, thus, the EM algorithm is formulated as an incomplete-data problem where the g-dimensional label vectors $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ indicate whether or not an observation came from a given kernel. The component $x_{i,j} = \left(\boldsymbol{x}_j\right)_i$ is equal to 1 if the j^{th} observation belongs to the i^{th} component of the mixture and 0 otherwise.

6.2.1.1 E-step

An initial set of parameters $\boldsymbol{\Psi}^{(0)}$ is required for the E-step prior to the first iteration of the algorithm. The initialization of the EM algorithm can be done by randomly defining the set of parameters, however, k-means++ algorithm (Arthur and Vassilvitskii, 2007) is often used for initialization to improve stability.

As the label vectors are missing, the expectation step of the $\left(t+1^{th}\right)$ iteration of the algorithm requires the calculation the current conditional expectation of X_{ij} given the observed data $\boldsymbol{y}$ and the current set of parameters $\boldsymbol{\Psi}^{(t)}$ from the previous iteration t. The E-step of the EM algorithm replaces the unknown labels x_{ij} by their conditional expectations $\tau_{ij}^{(t)}$ as follows:

$$\begin{aligned}\tau_i\left(\boldsymbol{y}_j;\boldsymbol{\Psi}^{(t)}\right) &= \mathrm{E}_{\boldsymbol{\Psi}^{(t)}}\left\{X_{i,j} \mid \boldsymbol{y}_j\right\} \\ &= \frac{\pi_i^{(t)}\phi\left(\boldsymbol{y}_j;\boldsymbol{\mu}_i^{(t)};\boldsymbol{\Sigma}_i^{(t)}\right)}{\sum_{k=1}^{g}\pi_k^{(t)}\phi\left(\boldsymbol{y}_j;\boldsymbol{\mu}_k^{(t)};\boldsymbol{\Sigma}_k^{(t)}\right)}, \qquad \forall i=1,\ldots,g \quad \text{and} \quad \forall j=1,\ldots,n \\ &= \tau_{i,j}^{(t)}\end{aligned} \tag{6.3}$$

6.2.1.2 M-step

The M-step consists of the maximization of the log likelihood and can be expressed in closed form for Gaussian components. The solution for the updated mean vector $\boldsymbol{\mu}_i^{(t+1)}$ and covariance matrix $\boldsymbol{\Sigma}_i^{(t+1)}$ are given by Equations 6.4 and 6.5 respectively, while the updated estimate of the component contribution π_i is calculated using Equation 6.6.

$$\boldsymbol{\mu}_i^{(t+1)} = \sum_{j=1}^{n}\tau_{i,j}^{(t)}\boldsymbol{y}_j \Big/ \sum_{j=1}^{n}\tau_{i,j}^{(t)}, \qquad \forall i=1,\ldots,g \tag{6.4}$$

$$\boldsymbol{\Sigma}_i^{(t+1)} = \sum_{j=1}^{n}\tau_{i,j}^{(t)}\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(t+1)}\right)\left(\boldsymbol{y}_j-\boldsymbol{\mu}_i^{(t+1)}\right)^T \Big/ \sum_{j=1}^{n}\tau_{i,j}^{(t)}, \qquad \forall i=1,\ldots,g \tag{6.5}$$

$$\pi_i^{(t+1)} = \sum_{j=1}^{n}\tau_{i,j}^{(t)}/n, \qquad \forall i=1,\ldots,g \tag{6.6}$$

The two steps are repeated until the difference in the likelihood function $L\left(\boldsymbol{\Psi}^{(t+1)}\right)$ and $L\left(\boldsymbol{\Psi}^{(t)}\right)$ is small enough (10^{-10} times the likelihood is used as the criteria here) indicating

convergence of the algorithm to a local optimum. Local optimum are deemed good enough for the proposed application. Inspection for unstable components (e.g. components fitting outliers) is advised.

6.2.2 Expectation Maximization With Missing Data

In geological applications it is common to sample some variables more often than others and the decision on whether to sample all variables is usually related to the value of the more important ones. As the mechanism of missing data is not completely random, disregarding unequally sampled locations will likely lead to bias on the estimated density which would be carried throughout the geostatistical workflow generating biased results.

The methodology described by Little and Rubin (2002) for EM with missing data allows the use of all data available including heterotopic samples. In this version of the EM algorithm, the E-step remains practically the same and the only change is that the multivariate PDF's used in the calculations are marginalized on the observed variables.

Prior to the M-step, the missing variables at the j^{th} observation $\boldsymbol{y}_m^j$ are replaced by their conditional mean $\hat{\boldsymbol{y}}_m^{(i,j)}$ given the set of observed variables $\boldsymbol{y}_o^j$ for each component $(i = 1, ..., g)$ (Equation 6.7).

$$\hat{\boldsymbol{y}}_m^{(i,j)} = \boldsymbol{\mu}_m^{(i,j)} + \boldsymbol{\Sigma}_{mo}^{(i,j)} \boldsymbol{\Sigma}_{oo}^{(i,j)-1} \left(\boldsymbol{y}_o^{(j)} - \boldsymbol{\mu}_o^{(i,j)} \right) \tag{6.7}$$

where $\boldsymbol{\Sigma}_{mo}^{(i,j)}$ and $\boldsymbol{\Sigma}_{oo}^{(i,j)}$ are the sub-matrices of $\boldsymbol{\Sigma}_i$ referring to the covariance between missing and observed variables and covariance between the observed variables for the i^{th} Gaussian component. The vectors $\boldsymbol{\mu}_m^{(i,j)}$ and $\boldsymbol{\mu}_o^{(i,j)}$ are the mean of missing and observed variables for the i^{th} component and are sub-vectors of $\boldsymbol{\mu}_i$.

The calculation of the updated mean $\boldsymbol{\mu}_i^{(t+1)}$ remains the same as in Equation 6.4 and the updated covariance $\boldsymbol{\Sigma}_i^{(t+1)}$ from Equation 6.5 is corrected by adding a matrix $(\boldsymbol{C}_i)$ derived from the sum of all conditional variances (Delalleau et al., 2012; Roberts, 2010) from each observation with missing variables (Equation 6.8). This corrects the smoothing (covariance reduction) effect of the conditional means that are replacing the missing observations for each component.

$$\boldsymbol{C}_i = \sum_{j=1}^{n} \tau_{i,j} \boldsymbol{M}_j^T \left(\boldsymbol{\Sigma}_{mm}^{(i,j)} - \boldsymbol{\Sigma}_{mo}^{(i,j)} \boldsymbol{\Sigma}_{oo}^{(i,j)-1} \boldsymbol{\Sigma}_{om}^{(i,j)} \right) \boldsymbol{M}_j / \sum_{j=1}^{n} \tau_{i,j} \tag{6.8}$$

where $\boldsymbol{M}_j$ is a sub-matrix of an identity matrix that is built by deleting rows corresponding to the missing dimensions for each observation $(j = 1, ..., n)$.

6.3 Methodology

Consider a set of n observations $\boldsymbol{z} = \left(\boldsymbol{z}_1^T, ..., \boldsymbol{z}_n^T\right)$ containing K-dimensional isotopic and heterotopic observations $\boldsymbol{z}_j = (z_{j,1}, ..., z_{j,K})$. The objective is to generate multiple realizations of the missing data that account for the spatial structure and multivariate relationship between the variables. This is achieved by combining two conditional distributions through non-parametric BU and sampling the final combined distribution. The first conditional distribution is calculated using spatial data of the same variable being imputed and the second is calculated using the other observed variables at same location. The steps for the non-parametric data imputation using GMM are as follows:

1. Normal score transformation: Each variable is transformed to be marginally standard Gaussian through normal score transformation (Equation 2.8). This is a univariate transformation done independently for each variable and, therefore, not sensitive to heterotopic observations.

2. Multivariate density estimation: The result of step 1 is a set of n observations $\boldsymbol{y} = \left(\boldsymbol{y}_1^T, ..., \boldsymbol{y}_n^T\right)$. The EM algorithm is then used to fit a GMM with g components to the transformed data defining the estimated multivariate density function as in Equation 6.1

3. Conditional distribution given spatial data: Each variable is assumed to be spatially multivariate Gaussian after the normal scores transformation, therefore, the conditional distribution given spatial data is Gaussian and parameterized by the kriging mean and variance that are calculated by solving the simple kriging system using nearby observations of the same variable being imputed. For a given variable k to be imputed at l^{th} data location, consider the set of neighboring locations $N_l^{(k)}$ containing all observed and previously simulated values (step 6) for the same variable. The conditional mean and variance can be calculated as follows, where $C_{l,j}^{(k)}$ is the spatial covariance between the l^{th} and j^{th} data locations defined for variable k.

$$\bar{\mu}_s^{(l,k)} = \sum_{j \in N_l^{(k)}} \lambda_j^{(l,k)} y_{j,k} \tag{6.9}$$

$$\bar{\sigma}_s^{2\ (l,k)} = 1 - \sum_{j \in N_l^{(k)}} \lambda_j^{(l,k)} C_{l,j}^{(k)} \tag{6.10}$$

$$\sum_{j \in N_l^{(k)}} \lambda_j^{(l,k)} C_{p,j}^{(k)} = C_{l,p}^{(k)} \qquad \forall p \in N_l^{(k)} \tag{6.11}$$

The conditional distribution given the spatial data is defined as:

$$P\left(y_{l,k} \mid Y_{j,k} = y_{j,k},\ \forall j \in N_l^{(k)}\right) = \phi\left(y_{l,k}; \bar{\mu}_s^{(l,k)}; \bar{\sigma}_s^{(l,k)}\right) \tag{6.12}$$

4. Conditional distribution given colocated data: This distribution is defined as the marginal of the conditional GMM given the colocated observed and previously imputed data with respect to the variable being imputed. It can be easily calculated from the GMM fitted in step 2. The conditional GMM requires the calculation of each conditional multivariate Gaussian component. Each one is parameterized by its conditional mean (Equation 6.13) and covariance (Equation 6.14).

$$\bar{\boldsymbol{\mu}}_m^{(i,l)} = \boldsymbol{\mu}_m^{(i,l)} + \boldsymbol{\Sigma}_{mo}^{(i,l)} \boldsymbol{\Sigma}_{oo}^{(i,l)-1} \left(\boldsymbol{y}_o^{(l)} - \boldsymbol{\mu}_o^{(i,l)}\right) \qquad i = 1, ..., g \tag{6.13}$$

$$\bar{\boldsymbol{\Sigma}}_{mm}^{(i,l)} = \boldsymbol{\Sigma}_{mm}^{(i,l)} - \boldsymbol{\Sigma}_{mo}^{(i,l)} \boldsymbol{\Sigma}_{oo}^{(i,l)-1} \boldsymbol{\Sigma}_{om}^{(i,l)} \qquad i = 1, ..., g \tag{6.14}$$

where $\boldsymbol{\Sigma}_{mo}^{(i,l)}$ and $\boldsymbol{\Sigma}_{oo}^{(i,l)}$ are the sub-matrices of $\boldsymbol{\Sigma}_i$ referring to the covariance between missing and observed variables and covariance between the observed variables for the i^{th} Gaussian component at l^{th} data location. The vectors $\boldsymbol{\mu}_m^{(i,l)}$ and $\boldsymbol{\mu}_o^{(i,l)}$ are the mean of missing and observed variables for the i^{th} component at l^{th} data location and are sub-vectors of $\boldsymbol{\mu}_i$ that is the mean of the i^{th} component of the GMM. After calculating the parameters of the conditional distribution of each Gaussian component the conditional distribution of the GMM is defined by:

$$\hat{f}\left(\boldsymbol{y}_m^{(l)}; \boldsymbol{\Psi}_m^{(l)}\right) = \sum_{i=1}^{g} \pi_i' \phi\left(\boldsymbol{y}_m^{(l)}; \bar{\boldsymbol{\mu}}_m^{(i,l)}; \bar{\boldsymbol{\Sigma}}_{mm}^{(i,l)}\right) \tag{6.15}$$

$$\pi_i' = \frac{\pi_i \phi\left(\boldsymbol{y}_o^{(l)}; \boldsymbol{\mu}_o^{(i,l)}; \boldsymbol{\Sigma}_{oo}^{(i,l)}\right)}{\sum_{i=1}^{g} \pi_i \phi\left(\boldsymbol{y}_o^{(l)}; \boldsymbol{\mu}_o^{(i,l)}; \boldsymbol{\Sigma}_{oo}^{(i,l)}\right)} \tag{6.16}$$

The marginal of the conditional GMM respective to the variable k being imputed at l^{th} location is defined as:

$$P\left(y_{l,k} \mid Y_{l,o} = y_{l,o},\ \forall o \in O_l\right) = \sum_{i=1}^{g} \pi_i' \phi\left(y_{l,k}; \bar{\mu}_c^{(i,l,k)}; \bar{\sigma}_c^{(i,l,k)}\right) \tag{6.17}$$

where O_l is the set of observed and previously simulated variables at location l, $\bar{\mu}_c^{(i,l,k)}$ and $\bar{\sigma}_c^{(i,l,k)}$ are the conditional mean and conditional standard deviation of the i^{th} component of the GMM at location l marginalized with respect to the missing variable k.

5. Combined distribution: The two conditional distributions defined in Equations 6.12 and 6.17 are combined with non-parametric BU (Neufeld and Deutsch, 2006). The Equation 6.18 is an approximation as it requires the assumption that the spatial observations of the same variable being imputed are independent from the colocated data from the other observed variables.

$$\begin{aligned} P\left(y_{l,k} \mid Y_{j,k} = y_{j,k},\ \ \forall j \in N_l^{(k)}; Y_{l,o} = y_{l,o},\ \forall o \in O_l\right) = \\ = \frac{\phi\left(y_{l,k}; \bar{\mu}_s^{(l,k)}; \bar{\sigma}_s^{(l,k)}\right) \sum_{i=1}^{g} \pi_i' \phi\left(y_{l,k}; \bar{\mu}_c^{(i,l,k)}; \bar{\sigma}_c^{(i,l,k)}\right)}{\phi\left(y_{l,k}; 0; 1\right)} \end{aligned} \tag{6.18}$$

6. Generate missing data realization: The updated distribution (Equation 6.18) is sampled with Monte-Carlo simulation (MCS) to generate an imputed realization of the missing data.

The steps 1 and 2 are fixed and are required only once. For each variable, the locations where the value for this variable is missing are visited sequentially. At each of these locations steps 3 to 6 are performed. After the imputation the value thus generated is deemed

as observed and used in the imputation of subsequent missing values. Once all missing values are imputed, the realization is stored and the original state of the variables is restored for the generation of the next realization. This is repeated until the desired number of realizations is reached. The realizations may be transformed back to original units after the simulation if required.

Each full-valued isotopic realization of the data is subjected to the geostatistical modeling workflow to create a multivariate realization of all variables at all unsampled locations. As many data realizations are required as geostatistical realizations. This scheme permits the uncertainty due to missing data to be transferred through to the spatial realizations and subsequent resource estimation.

6.4 Application to Lateritic Nickel Data Set

In order to illustrate the proposed methodology a data set with geological variables from a lateritic nickel deposit is used. The data set is composed of measurements of nickel (Ni), which is the most important variable, iron (Fe), silica (SiO_2), and magnesium (MgO). The original data set has 18 352 isotopic observations, but samples were removed to emulate the common scenario in which several less important variables are missing. This synthetic data example is useful because the imputed values can be checked against the known true values. Variables are missing specially where the most important have been already measured to be low and, therefore, not justifying the additional expense of measuring the others. This is the classical situation characterizing MAR mechanism.

The Ni measurements are maintained everywhere while 1/3 of the Fe samples were randomly removed from samples in which Ni values are lower than the median and 1/6 from samples where Ni is above the median resulting in 9178 observed Fe measurements. For MgO and SiO_2 the fraction removed is 1/6 where Ni is lower than the median and 1/10 where Ni is higher than the median resulting in 13 459 observations of these variables. Note that this is the same missing data strategy as used by Barnett and Deutsch (2015) to generate subsets of the same data set resulting in similar but different data set as the selection is random.

The data set is subdivided into four rock types, (1) basic east type ore (BETO), (2) basic west type ore (BWTO), (3) acid east type ore (AETO), and (4) acid west type ore (AWTO). As

with other geostatistical techniques, the imputation methodology proposed here is applied to each population (rock type) independently. The summary of the number of data and proportion missing by rock type are shown in Table 6.1.

Rock type	Number of samples	% missing (Fe)	% missing (SiO_2)	% missing (MgO)
BETO	5968	46.67	25.18	25.17
BWTO	4016	45.69	26.20	26.20
AETO	1389	60.62	30.60	30.60
AWTO	6979	53.19	27.43	27.43

Table 6.1: Number of samples per rock type and percentage missing.

Each variable is transformed to be marginally standard normal. The scatter plots of transformed variables are shown in Figure 6.1 for BETO. Its clear that after the univariate transformation the multivariate distribution remains highly non-linear and heteroscedastic. These complex features of geological data are the main motivation of using multivariate transformation techniques and, therefore, MI techniques such as the one presented here.

The transformed data are fitted with GMMs of 20, 15, 20 and 20 components for BETO, BWTO, AETO and AWTO, respectively. There are a number of techniques to assist in the selection of the number of components used for GMM including split/merge EM algorithms and bootstrap. For this example the number of components is chosen through visual inspection in order to avoid overfitting while capturing complexities of the data. The GMM fitted to the BETO data is shown in Figure 6.2. Note that the standard normal marginal distributions are closely matched by theGMMs. This is possible because heterotopic data are considered during the fitting process using the methodology for fitting GMM's with missing data. The marginal distribution of the isotopic part of the data set is not standard normal due to a higher proportion of missing data at low Ni intervals. Fitting GMMs to heterotopic data is key to avoid bias.

The imputation methodology proposed here is applied to generate 100 realizations of the data for each rock type. The parametric and non-parametric techniques proposed by Barnett and Deutsch (2015) are also applied to this data set for comparison purposes. The non-parametric technique proposed by Barnett and Deutsch (2015) is referred to as non-parametric kernel density estimation (NPKDE) while the technique proposed in this paper is referred to as semi-parametric Gaussian mixture model (SPGMM) highlighting the main difference between the two that is the method of estimating conditional distributions from

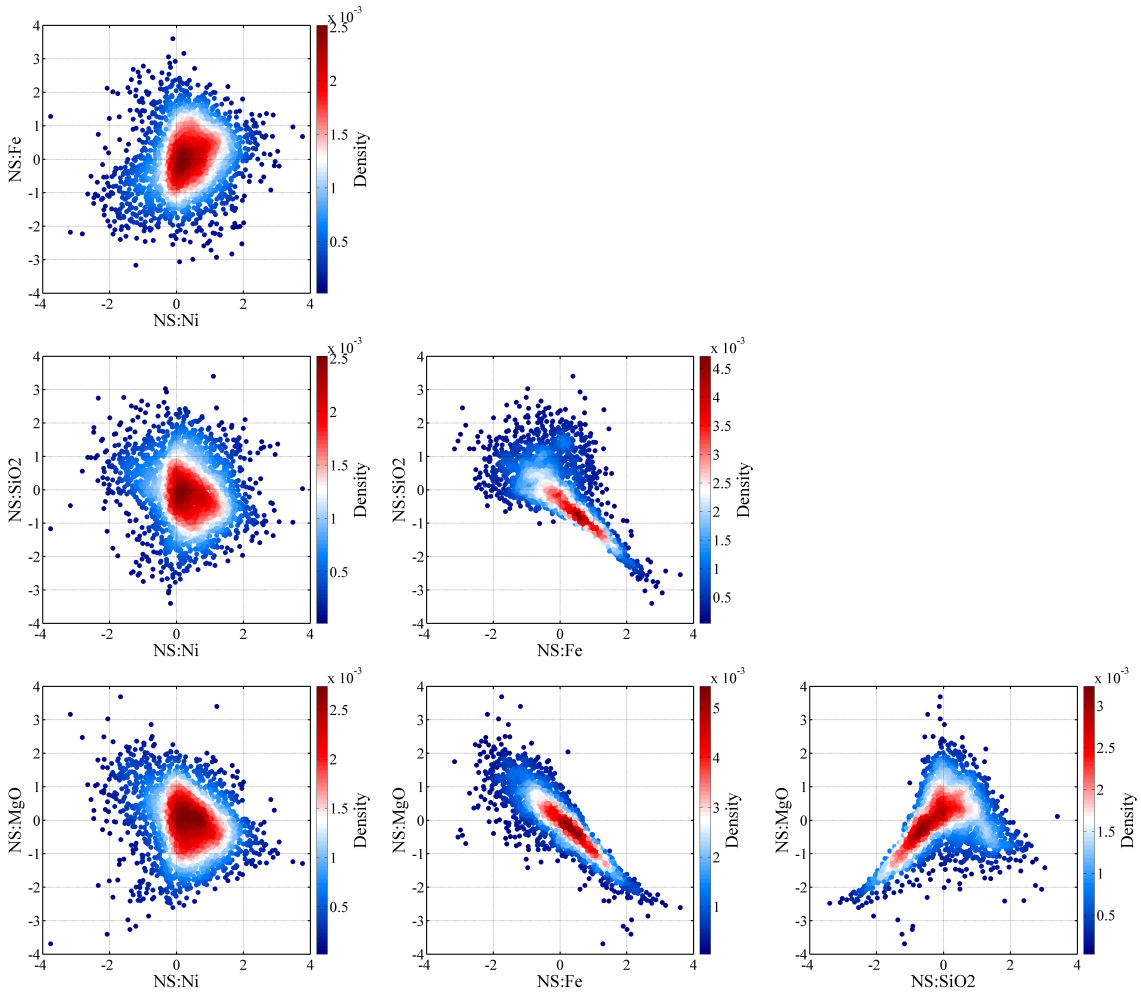


Figure 6.1: Bivariate scatter plots of normal scores transformed data for BETO rock type coloured by the bivariate KDE. Only isotopic observations are used for plotting.

collocated data.

One motivation for development of this methodology is to reduce computational time required for imputation while keeping the advantages of the non-parametric technique. The computational time required to run each method to this data set is shown in Table 6.2. These results were obtained with an Intel® Core™ i7-4790K CPU @ 4.00GHz. The time is improved considerably. The SPGMM for this data set took from 3 to nearly 5 times longer than the parametric technique while the NPKDE took from 80 to 700 times longer.

In order to check if the proposed technique is able to keep the advantages of the non-parametric technique, the known true values are used to cross-validate the results from each technique based on several univariate and bivariate performance measures. The univariate performance measures are cumulative density function (CDF) reproduction that

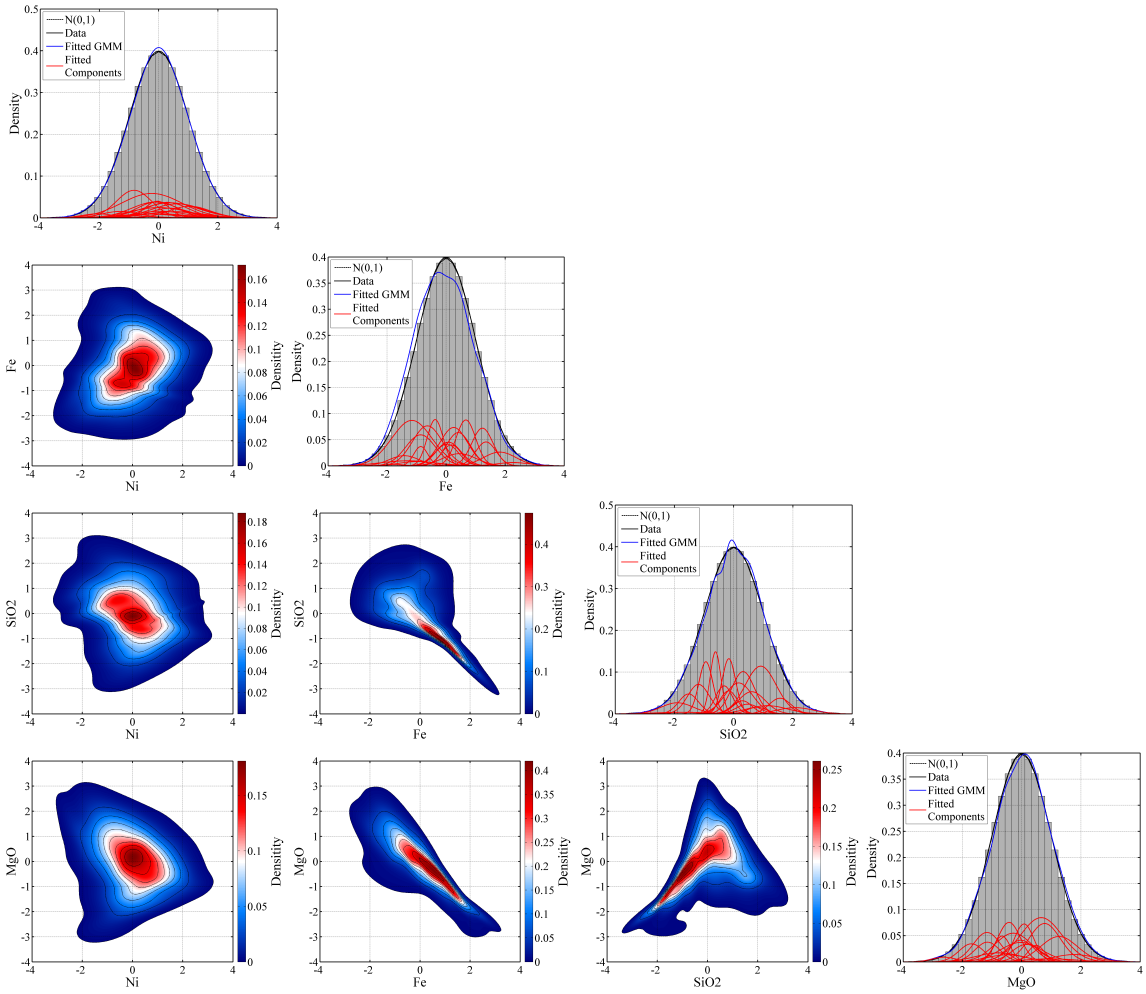


Figure 6.2: Bivariate and univariate marginals of the GMM fitted to all observations in BETO rock type. As heterotopic observations are allowed during the fitting, the marginals that are known to be standard Gaussian are reasonably matched, avoiding introduction of bias. The data is fitted with 20 multivariate Gaussian components.

Rock type	Number missing	Time (s)		
		Parametric	NPKDE	SPGMM
BETO	5789	17.11 (1)	8390.64 (490.4)	55.43 (3.2)
BWTO	3939	9.78 (1)	3848.86 (393.5)	29.96 (3.1)
AETO	1692	4.97 (1)	401.86 (80.9)	16.64 (3.4)
AWTO	7540	15.31 (1)	10968.60 (716.4)	70.97 (4.6)

Table 6.2: Time in seconds per technique and per rock type. The time relative to the fastest is shown in brackets.

includes mean and variance reproduction, correlation between true and imputed values, root mean squared error (RMSE) between true and imputed values and RMSE between the true variogram (vertical and horizontal) and the variogram of the imputed data. The bivariate performance measures considered in this article are the reproduction of the correlation and RMSE on the bivariate KDE between true and imputed data.

Box plots of the univariate measures calculated using all realizations are shown in Figure 6.3. The relative results are summarized in Table 6.3 to facilitate the comparison between techniques. The best average results are highlighted. The performance results are surprisingly good considering that the initial intent was to improve time while keeping the performance. On average, the proposed SPGMM yielded the best results in all univariate performance checks (Table 6.3) even where the parametric method outperformed the NPKDE (σ^2 error). The SPGMM is outperformed only for the reproduction of SiO_2 variance.

Measure	Method	Fe	SiO_2	MgO	Avg.
μ Error	Parametric	1.00	0.55	0.99	0.84
	NPKDE	0.28	1.00	1.00	0.76
	SPGMM	0.14	0.45	0.67	**0.42**
σ^2 Error	Parametric	0.30	0.83	0.80	0.64
	NPKDE	1.00	1.00	1.00	1.00
	SPGMM	0.15	0.91	0.46	**0.51**
$1-\rho$	Parametric	1.00	1.00	1.00	1.00
	NPKDE	0.63	0.69	0.73	0.68
	SPGMM	0.55	0.62	0.65	**0.60**
RMSE	Parametric	1.00	1.00	1.00	1.00
	NPKDE	0.81	0.86	0.87	0.84
	SPGMM	0.74	0.78	0.81	**0.78**
γ Error	Parametric	1.00	1.00	1.00	1.00
	NPKDE	0.71	0.75	0.80	0.75
	SPGMM	0.57	0.61	0.77	**0.65**

Table 6.3: Summary of univariate performance measures. The best average result is highlighted. All results are stated relative to the worst case.

Univariate statistics are the easiest to reproduce and, not surprisingly, all techniques result in appropriate reproduction of the CDFs as shown in Figure 6.4. It is visually clear that the semi and the non-parametric techniques improve upon the parametric. It is also clear that the SPGMM technique improves the reproduction of the distribution tails in this

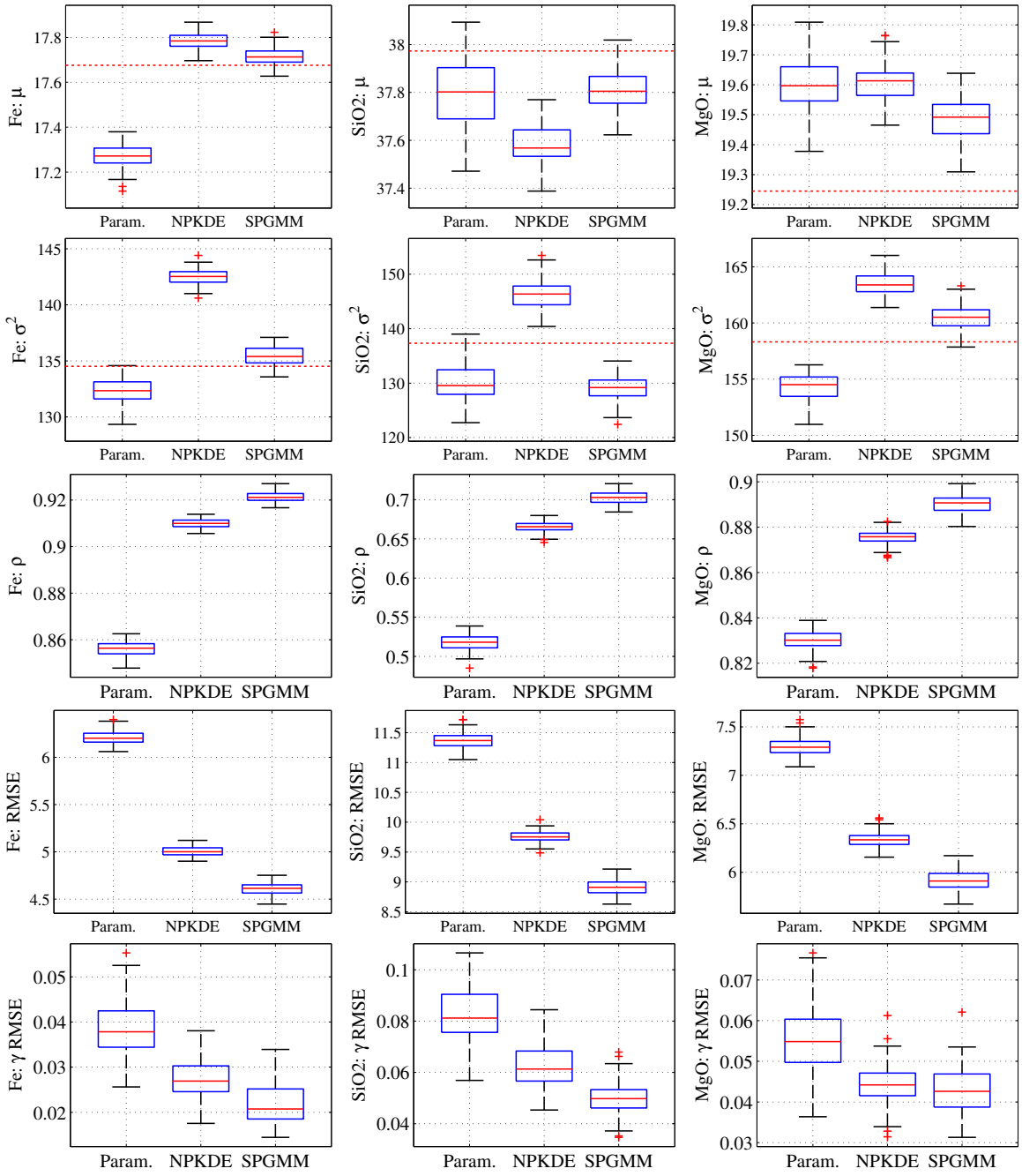


Figure 6.3: Box plot summary of univariate measures for all realizations. Dashed red lines represents known true values when applicable.

example.

The scatter plots of true values vs e-type estimated values are shown in Figure 6.5. Again, all techniques perform well in the missing data estimation task mostly due to the highly conditioned environment. While conditional bias is not evident for the parametric and the SPGMM it seems slightly more pronounced for the NPKDE technique. This was

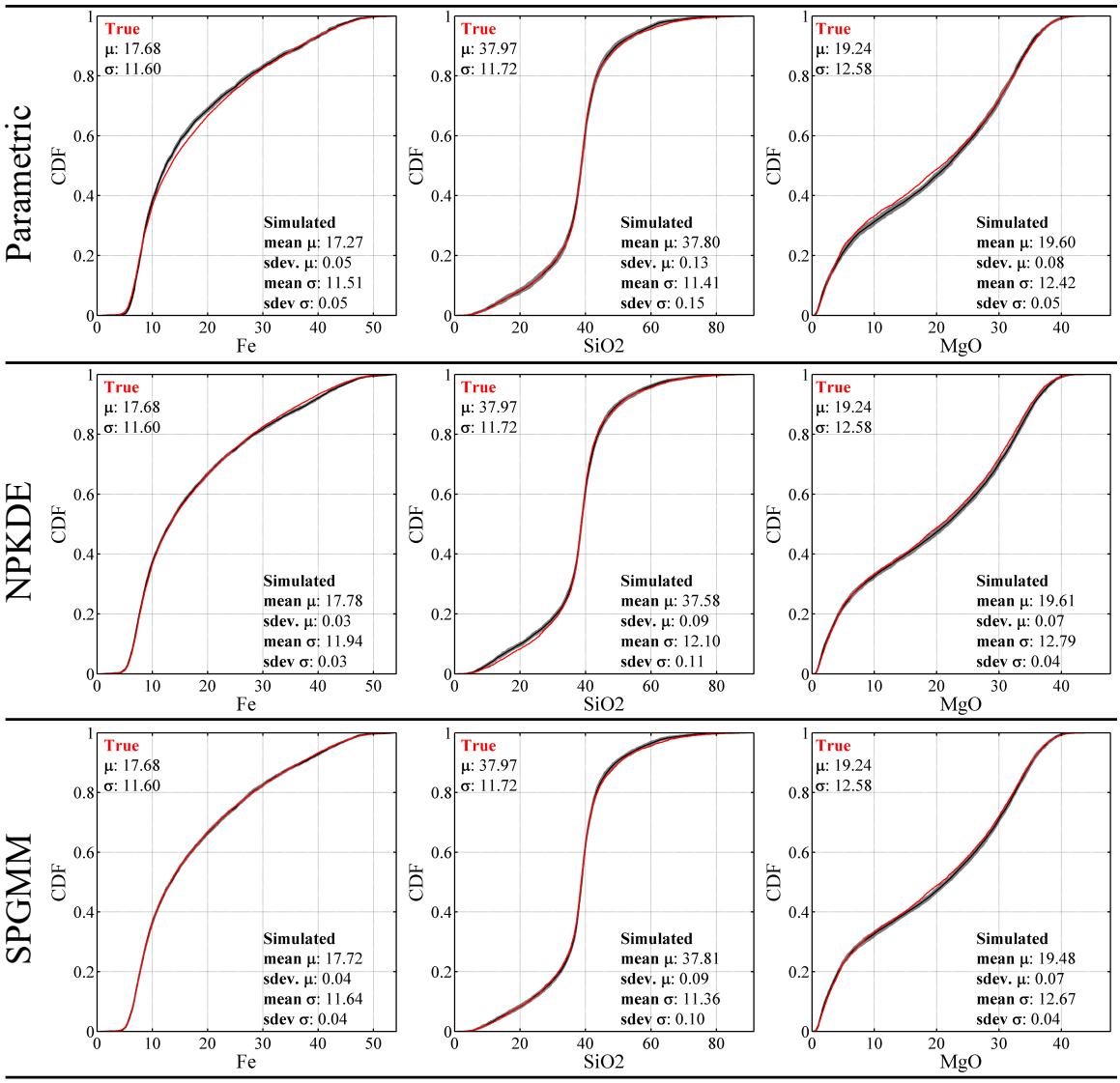


Figure 6.4: Reproduction of univariate CDF. Grey lines are the CDF of each realization, the black line is the average of all realizations and red line is the CDF of the original values.

not observed in the work by Barnett and Deutsch (2015) and could be related to modeling decisions and parameters or perhaps due to the different random selection of missing data. This highlights an advantage of using GMM to derive the multivariate conditional distributions rather than using a built-in Gibbs sampler. The fitted GMMs can be visualized by the user with plots like the one shown in Figure 6.2 allowing the user to evaluate the adequacy of the fitted model and change fitting parameters if required to potentially improve results.

The reproduction of direct horizontal and vertical variogram are shown in Figs. 6.6 and 6.7. The overall variogram reproduction for the semi and non-parametric techniques are better compared to the parametric approach. The SPGMM shows considerable improvement es-

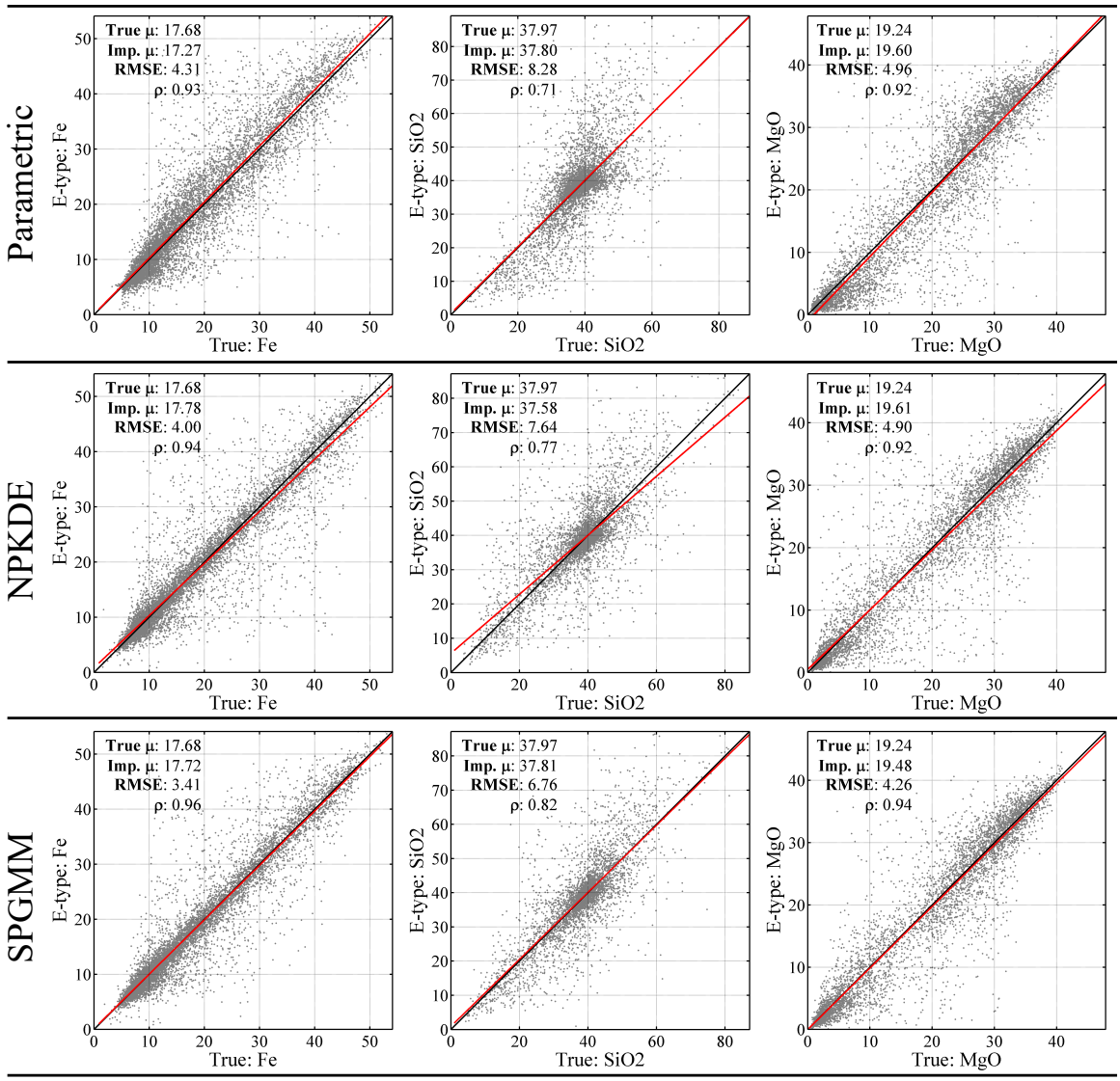


Figure 6.5: Scatter plot of true vs e-type estimate of the missing values. The black line is the 45 degree line and the red is the linear regression.

pecially regarding the horizontal variogram.

Box plots of the bivariate measures calculated using all data realizations are shown in Figure 6.8. The relative results are summarized in Table 6.4. The best average results are highlighted. Again, the average performance results are better for the proposed SPGMM technique for both measurements ρ error and KDE RMSE.

The bivariate distributions are shown in Figure 6.9. The benefits of semi and non-parametric approaches are more evident when dealing with the reproduction of the multivariate complexities. The superiority of the semi and non-parametric distribution is quite evident upon visual inspection of the scatter plots shown in Figure 6.9. Note that distribu-

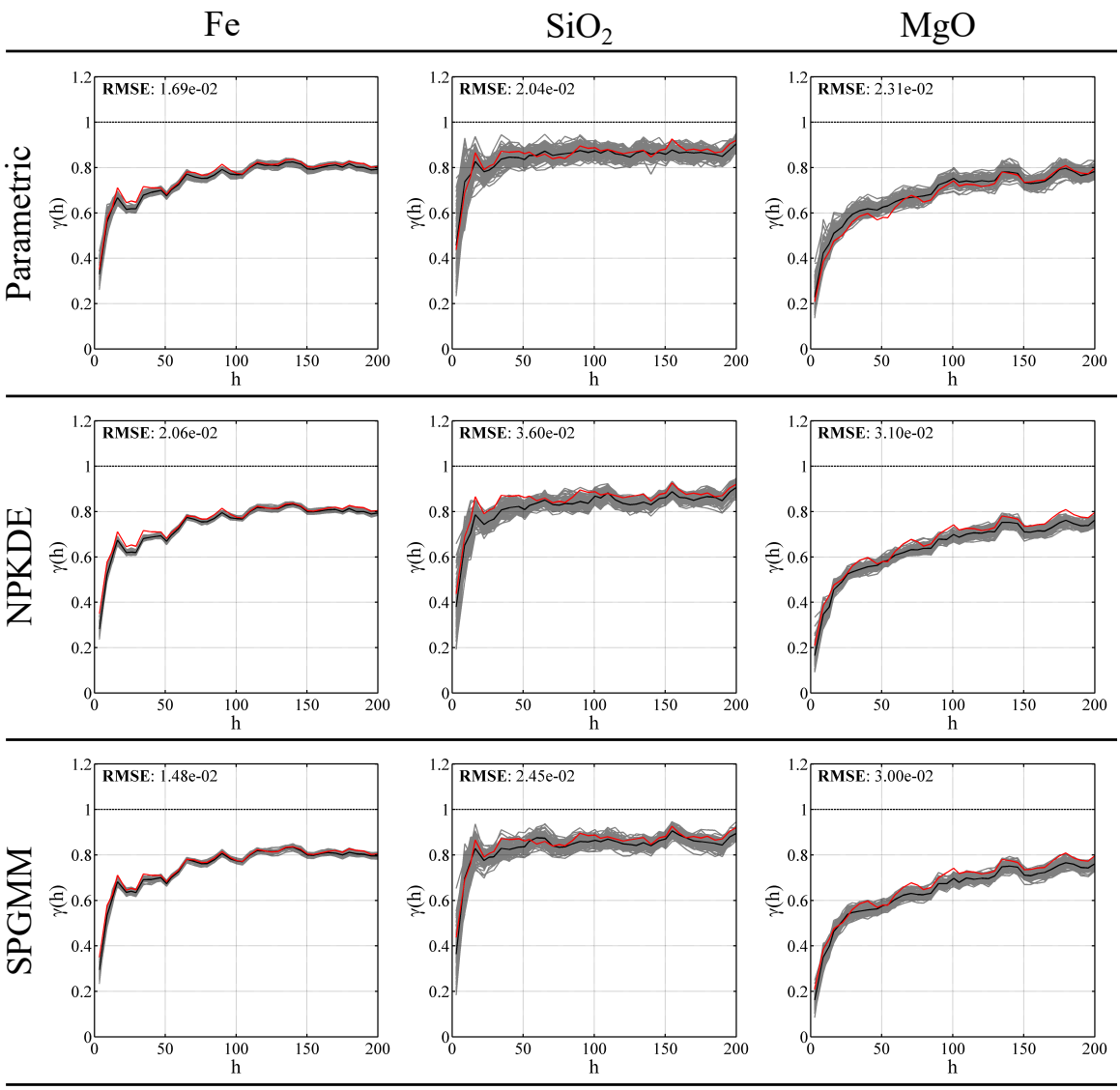


Figure 6.6: Horizontal variogram reproduction for each technique. Grey lines are the variogram of each realization, the black line is the average of all realizations and red line is the variogram of the original values.

Measure	Method	Fe-SiO_2	Fe-MgO	SiO_2-MgO	Avg.
ρ Error	Parametric	1.00	1.00	0.70	0.90
	NPKDE	0.23	0.03	1.00	0.42
	SPGMM	0.41	0.07	0.48	**0.32**
KDE RMSE	Parametric	1.00	1.00	1.00	1.00
	NPKDE	0.90	0.91	0.90	0.91
	SPGMM	0.87	0.85	0.91	**0.88**

Table 6.4: Summary of bivariate performance measures. Best average result is highlighted. All results are stated relative to the worst case.

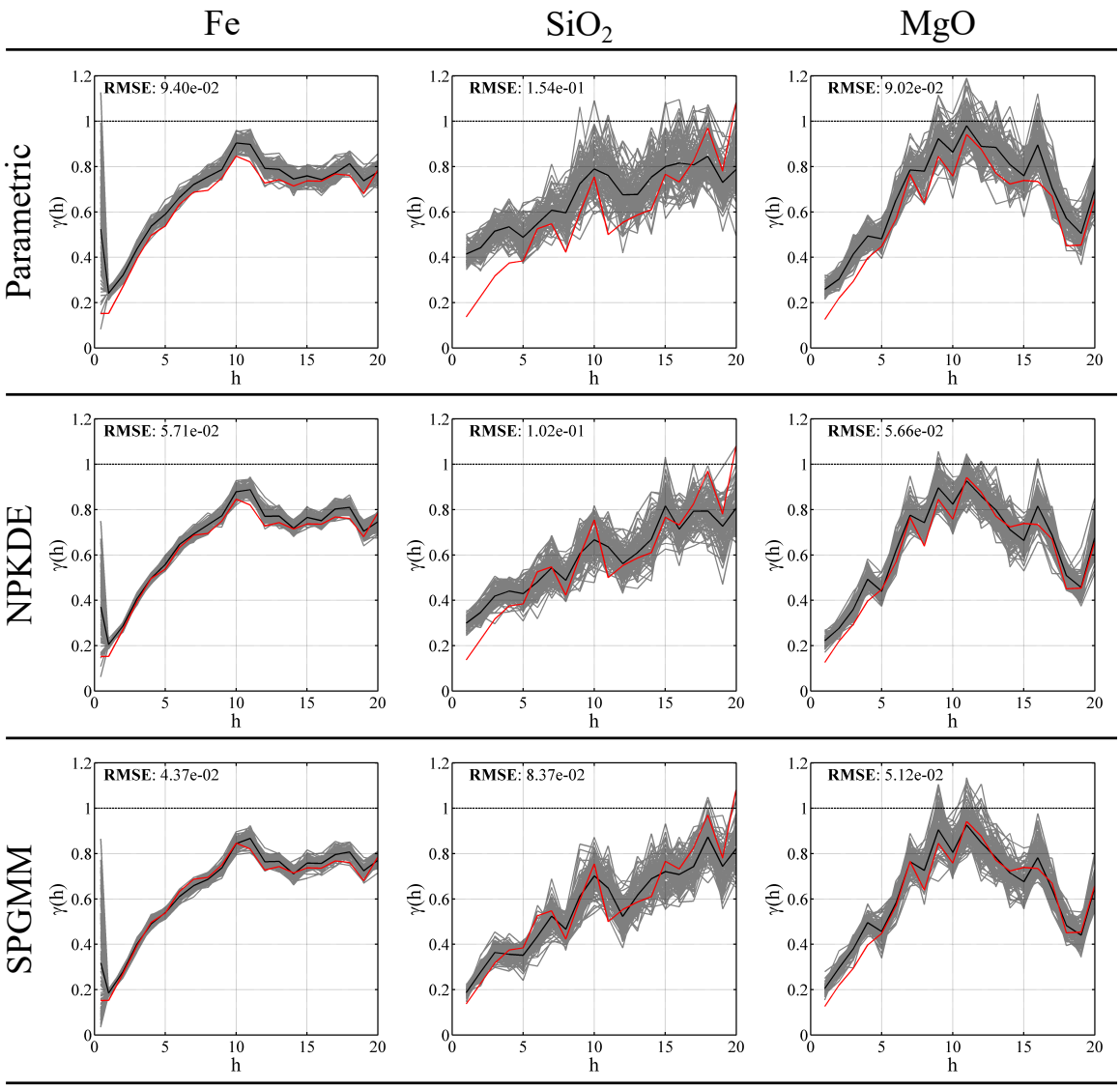


Figure 6.7: Vertical variogram reproduction for each technique. Grey lines are the variogram of each realization, the black line is the average of all realizations and red line is the variogram of the original values.

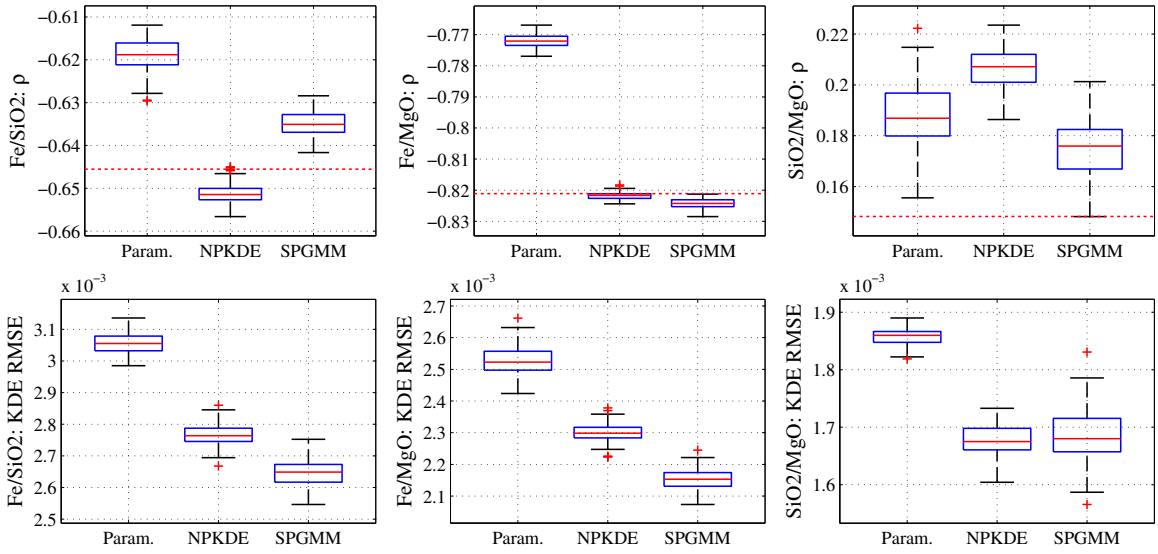


Figure 6.8: Box plot summary of bivariate measures for all realizations. Dashed red lines represents known trues value when applicable.

tions with Ni are not accounted in the analysis as they are relatively well reproduced by all techniques due to the fact that no Ni value is missing.

In light of the significant improvement of the SPGMM results over the NPKDE, further investigation of the differences between the two is carried out. The consistent improvement in performance can be related to the quality of the estimated conditional distributions. The current methodology for NPKDE only uses the isotopic observations for the calculation of univariate conditionals used in the Gibbs sampler while SPGMM uses a GMM fitted to all observations including locations where some variables are missing. In addition to a more well informed multivariate distribution, this feature reduces the bias caused by data MAR in which the missing data mechanism is dependent on the values of the observed variables but independent of the missing values (Barnett and Deutsch, 2015; Little and Rubin, 2002).

The differences between the conditional distributions from NPKDE and SPGMM are quantified through three measures. The first is the well known Kolmogorov-Smirnov (KS) d-distance which is defined by Equation 6.19. The KS d-distance can be very high for highly conditioned distributions (low variance) with no significant difference in the actual Gaussian values, therefore, the Equation 6.20 is used to measure the distance between the distributions in Gaussian units. The Equation 6.20 can be sensitive to the tails of the distributions and, for this reason, the difference between the median values is also calculated and shown separately.

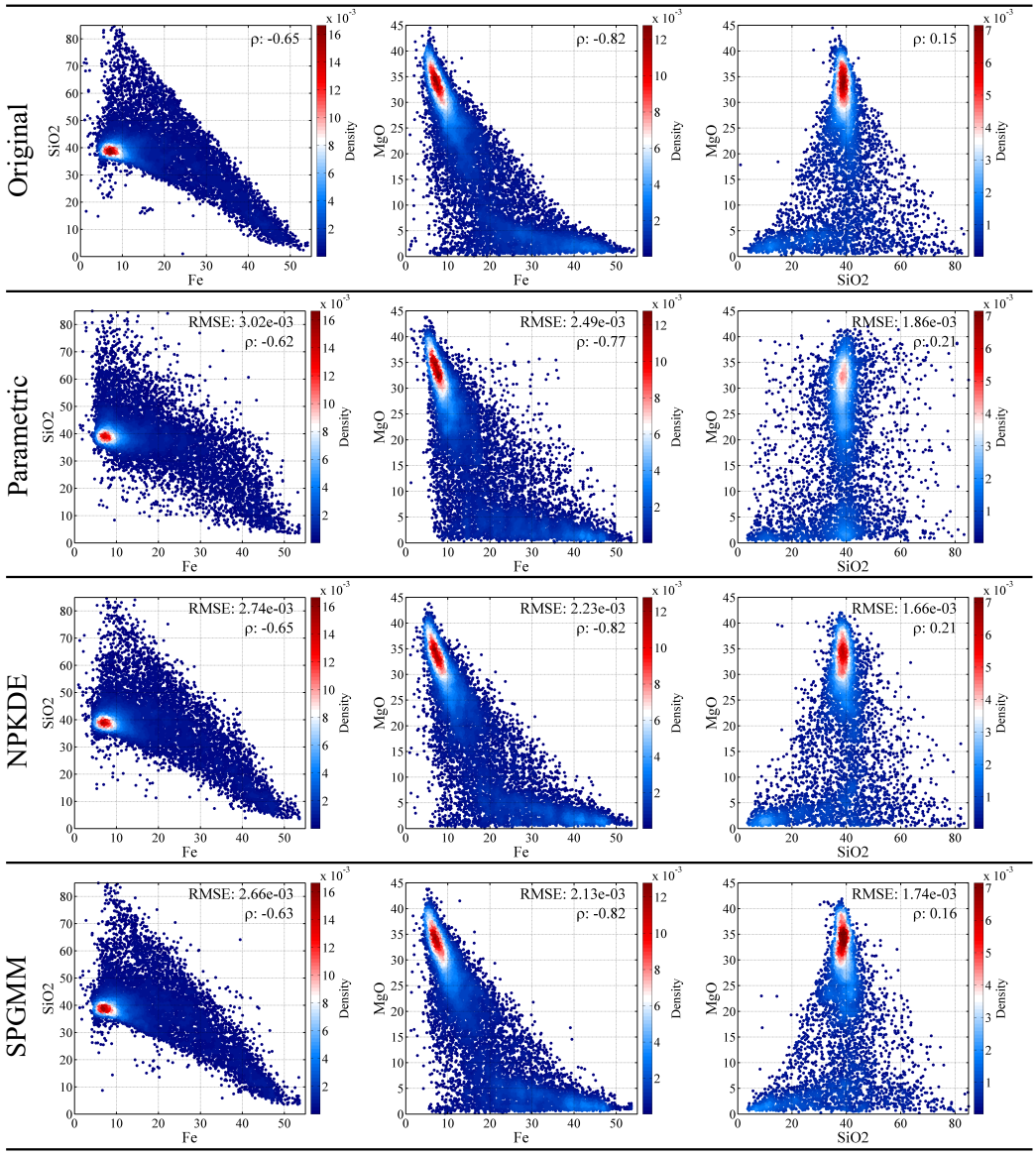


Figure 6.9: Bivariate scatter plots coloured by KDE. Orginal distributions are shown on top for comparison.

$$\mathrm{D}_{\mathrm{KS}j,k} = \max_{y} \left| F_{\mathrm{KDE}j,k}\left(y\right) - F_{\mathrm{GMM}j,k}\left(y\right) \right| \tag{6.19}$$

where $\mathrm{D}_{\mathrm{KS}j,k}$ is the KS d-distance for the missing variable k at location j, $F_{\mathrm{KDE}j,k}(y)$ is the conditional CDF of variable k at location j estimated using KDE and $F_{\mathrm{GMM}j,k}\left(y\right)$ is the conditional CDF of variable k at location j estimated using GMM.

$$\mathrm{D}_{\mathrm{Q}j,k} = \max_{p} \left| F^{-1}_{\mathrm{KDE}j,k}\left(p\right) - F^{-1}_{\mathrm{GMM}j,k}\left(p\right) \right| \tag{6.20}$$

where $\mathrm{D}_{\mathrm{Q}j,k}$ is the quantile distance for the missing variable k at location j, $F^{-1}_{\mathrm{KDE}j,k}(y)$ is the inverse of the conditional CDF of the variable k at location j estimated using KDE and $F^{-1}_{\mathrm{GMM}j,k}(y)$ is the inverse of the conditional CDF of the variable k at location j estimated using GMM.

The conditional distributions are calculated for every missing variable in the BETO rock type and the three measurements are calculated for each pair of CDFs coming from each technique. The histograms of each measurement are shown in Figure 6.10. Both histograms for KS d-distance (Figure 6.10a) and quantile distance (Figure 6.10b) show substantial differences between the distributions from each technique for this well informed example. Figure 6.10c also shows substantial differences between the median values, but it also shows that NPKDE is unbiased in relation to SPGMM in a global sense as the mean is close to zero. The data imputation problem is highly conditioned by collocated data and therefore such differences in estimated distributions are significant.

6.5 Conclusion

A methodology for multivariate data imputation using a GMM for estimation of the multivariate distribution within Bayesian updating workflow that improves upon existing ones is developed. The most important improvement is related to the computational time, which is reduced by two orders of magnitude compared to the existing technique that uses KDE and Gibbs sampler for the estimation of multivariate conditional distributions. This improvement would make the approach practical versus waiting days.

The technique shows consistent performance improvements in both univariate and bivariate tests. The differences between the multivariate conditional distributions estimated using univariate KDE coupled with Gibbs sampler and the distributions estimated using

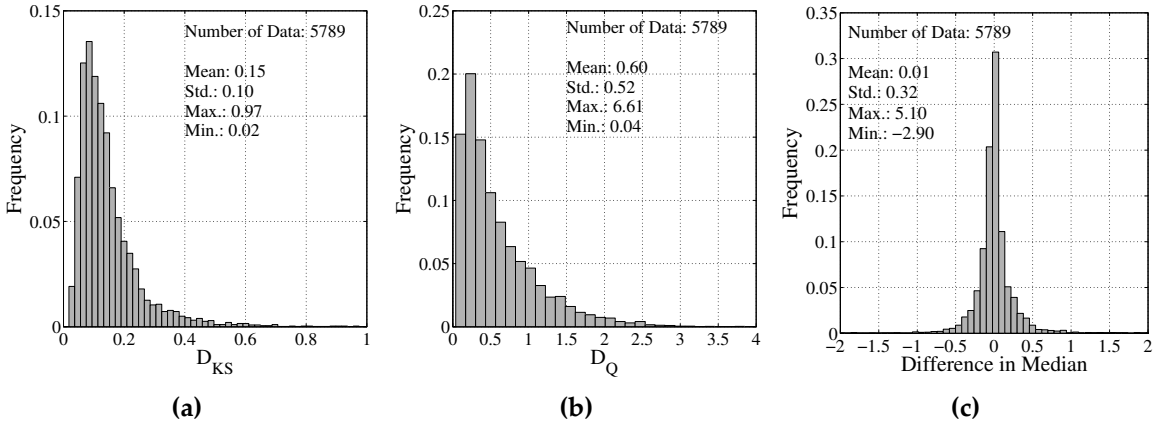


Figure 6.10: Measures of difference between conditional distributions from NPKDE and SPGMM. The histogram of KS d-distance between the conditional from the two sources for each missing variable is shown in (a); a similar measure to d-distance, but for the maximum distance between quantiles is shown in (b); and the histogram of the differences on the median of each distribution is shown in (c).

GMM are substantial. The improved performance is attributed to superior estimation of multivariate conditional distributions given colocated data. The multivariate density estimated using GMM accounting for all data including the unequally sampled locations reduces local bias and improves the robustness against sparse data sets.

Even though the results were satisfactory, there is room for improvement. The number of components for the GMM in this study was chosen by visual inspection, however, there are a variety of techniques that can be used as alternative such as split/merge expectation maximization algorithm and bootstrap techniques. Other possible improvements could include accounting for the spatial correlation across different variables and investigate the impact of declustering weights in case of spatial preferential sampling. The potential performance gain should be enough to justify the increased complexity in these cases.

Chapter 7

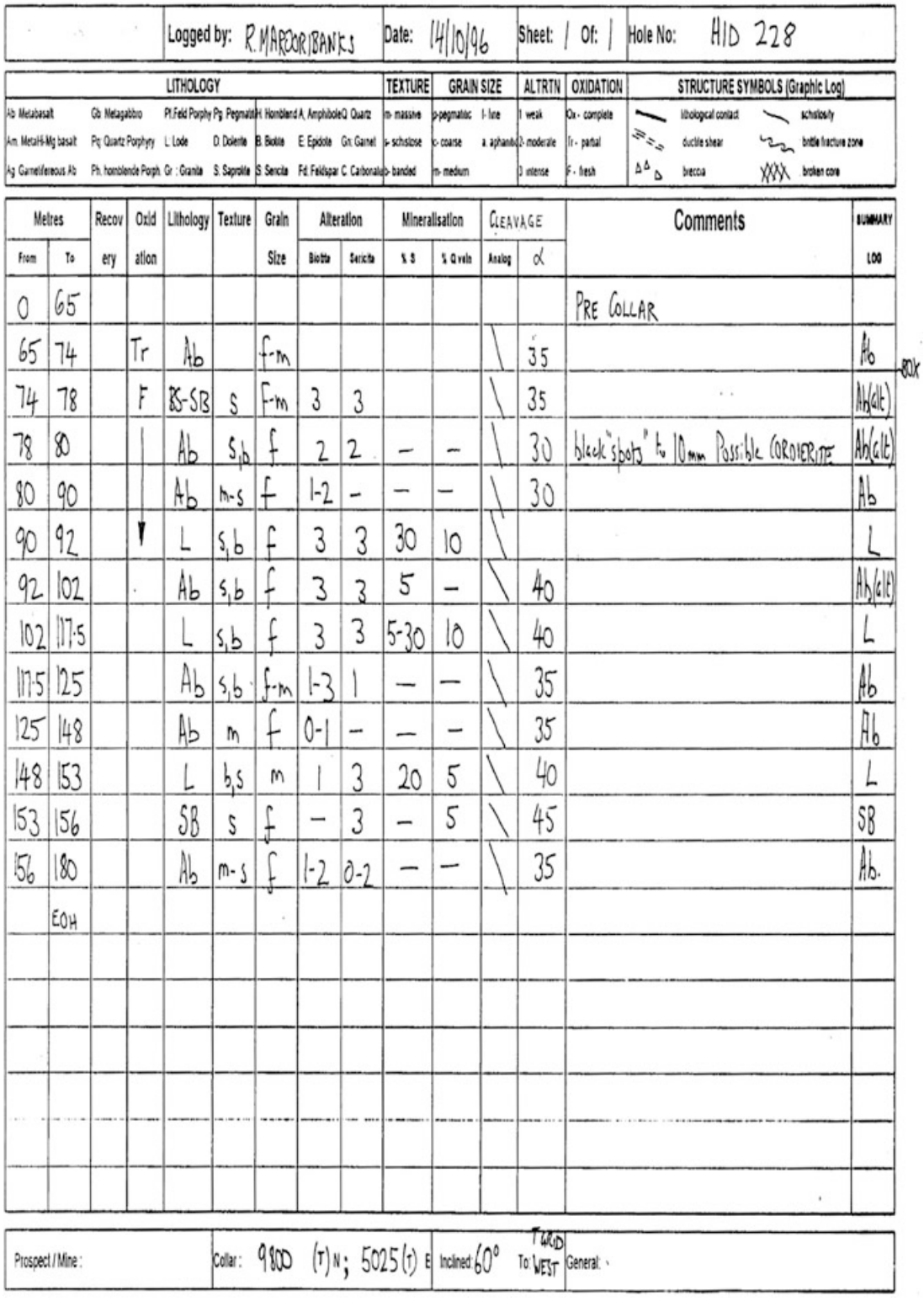
Multivariate Categorical Modeling with HTPG

Multiple categorical variables are often available for geostatistical modeling. Each categorical variable has a number of possible categorical outcomes. The current approach for numerical modeling of categorical variables is to either combine the categorical variables or to model them independently. The collapse of multiple categorical variables into a single variable with all possible/observed combinations is impractical due to the large number of combinations. The independent modeling of each categorical variable will fail to reproduce the joint categorical proportions. A methodology for the multivariate modeling of categorical variables utilizing the hierarchical truncated pluri-Gaussian (HTPG) approach is developed in this chapter and illustrated with the Swiss Jura data set.

7.1 Importance of Multiple Categorical Models

Multiple spatially distributed categorical variables are common in many areas of geoscience and yet there has been little research on the multivariate modeling of categorical variables (Emery and Cornejo, 2010). In mining, for instance, it is common to have categorical variables such as lithology, alteration, mineralization, oxidation, and structural domains (Bye, 2011; Rossi and Deutsch, 2014; Tonder et al., 2010). Each one of these categorical variables are defined by several mutually exclusive categories. An example of drillhole logging spread sheet with multiple categorical variables defined for each core section is shown in Figure 7.1.

Categorical variables have been mostly utilized for the definition of stationary domains for simulation and/or estimation of grade variables. The categories that are identified as the main controls of grade variability often belong to different categorical variables and they must be combined to form these stationary domains. The relevant categories to grade estimation/simulation are case specific. For instance, in a gold deposit, the main factors affecting metal content may be defined by the oxidation, alteration and structural controls



Logged by: R. MARJORIBANKS | Date: 14/10/96 | Sheet: 1 Of: 1 | Hole No: HID 228

LITHOLOGY: Ab Metabasalt; Gb Metagabbro; Pf Feld Porphy; Pg Pegmatite; H Hornblend; A Amphibole; Q Quartz; Am Meta Hi-Mg basalt; Pq Quartz Porphyry; L Lode; D Dolerite; B Biotite; E Epidote; Gt Garnet; Ag Garnetiferous Ab; Ph hornblende Porph; Gr Granite; S Saprolite; S Sericite; Fd Feldspar; C Carbonate

TEXTURE: m massive; s schistose; b banded

GRAIN SIZE: p pegmatoic; c coarse; m medium; f fine; a aphanitic

ALTRTN: 1 weak; 2 moderate; 3 intense

OXIDATION: Ox complete; Pr partial; F fresh

STRUCTURE SYMBOLS (Graphic Log): lithological contact; schistosity; ductile shear; brittle fracture zone; breccia; broken core

Metres From	Metres To	Recovery	Oxidation	Lithology	Texture	Grain Size	Alteration Biotite	Alteration Sericite	Mineralisation % S	Mineralisation % Q vein	CLEAVAGE Analog	CLEAVAGE α	Comments	SUMMARY LOG
0	65												PRE COLLAR	
65	74		Tr	Ab		f-m					\	35		Ab
74	78		F	SS-SB	s	f-m	3	3			\	35		Ab(alt)
78	80		|	Ab	s,b	f	2	2	–	–	\	30	black "spots" to 10mm Possible CORDIERITE	Ab(alt)
80	90		|	Ab	m-s	f	1-2	–	–	–	\	30		Ab
90	92		▼	L	s,b	f	3	3	30	10	\			L
92	102		.	Ab	s,b	f	3	3	5	–	\	40		Ab(alt)
102	117.5			L	s,b	f	3	3	5-30	10	\	40		L
117.5	125			Ab	s,b	f-m	1-3	1	–	–	\	35		Ab
125	148			Ab	m	f	0-1	–	–	–	\	35		Ab
148	153			L	b,s	m	1	3	20	5	\	40		L
153	156			SB	s	f	–	3	–	5	\	45		SB
156	180			Ab	m-s	f	1-2	0-2	–	–	\	35		Ab.
	EOH													

Prospect / Mine: | Collar: 9800 (T) N; 5025 (T) E | Inclined 60° To TRUE WEST | General:

Figure 7.1: Drillhole logging spread sheet with an example of multiple categorical variables that are interpreted for each core section. This log sheet includes oxidation, lithology, texture, recovery, grain size, alteration, and mineralization. This figure is obtained from Marjoribanks (2010).

whereas in diamondiferous kimberlite, the lithology will represent the main control, Rossi and Deutsch (2014).

In addition to the definition of stationary domains for the modeling of grades, the categorical variables are also relevant for the definition of other domain types such as geometallurgical (Acar, 2016; Angove and Acar, 2016; Deutsch et al., 2016; Hunt and Berry, 2017) and structural domains (Hunt and Berry, 2017; Rossi and Deutsch, 2014). Geometallurgical domains are often defined based on the categories, across the multiple categorical variables, that have the most impact on metallurgical processing. This includes recovery, chemical reagent and energy consumption, mill throughput, environmental impact (acid drainage), among others. The geometallurgical variables are as important as the grade variables to the economic success of a mining project (Scheffel et al., 2016), especially for low grade projects (Deutsch et al., 2016). The structural domains are modeled and utilized to inform mine/slope stability and blasting design (Hunt and Berry, 2017).

The different stationary domains defined for each particular application utilize different combinations of categories within the categorical variables (Bye, 2011; Hunt and Berry, 2017). The two options for the modeling of these categorical variables with univariate workflows are to define a single categorical variable by combining the categories of each set or to model each set independently. If the first option is chosen, the combinatorial nature of the problem would lead to a high number of merged categories. The definition of many categories often result in lack of data to infer the required parameters for modeling within each domain (Rossi and Deutsch, 2014) and lumping is required to lower the number of categories to a manageable level. Information is inevitably lost during the lumping process. Moreover, if stationary domains are required for multiple objectives such as resource and metallurgical variables estimation, one or both objectives would have to be compromised to build a model with a single set of categories.

The second alternative with the current univariate workflows is to model each categorical variable independently and perform the combination and lumping of the categories after modeling to define the stationary domains for different applications. In this case, the inference problem for the parameters required for the categorical modeling is solved as the data do not require to be subset based on the multiple possible combinations. The drawback with this approach is that it does not consider any multivariate relationship that may exist across the different categorical variables.

A lithology unit, for instance, may be more or less susceptible to alteration depending on its mineral composition and the stability of the minerals. The degree of alteration may also be affected by other factors such as texture and fractures. In some instances, the lithology can have strong control on the mineralization in some regions of the deposit with low alteration and be completely overprinted by alteration in other regions where it no longer represents a strong control (Rossi and Deutsch, 2014). The categorical variables are often not completely independent from each other and their spatial overlap has significant effects on the continuous attributes.

Emery and Cornejo (2010) proposed a methodology based on the truncated Gaussian simulation (TGS) for the multivariate modeling of categorical variables using a linear model of coregionalization (LMC) to co-simulate the latent variables. The LMC of the latent Gaussian variables is derived iteratively to ensure the reproduction of the spatial structure of the categorical variables. The TGS only uses one latent variable per categorical variable. This is too simplistic to reproduce complex categorical ordering and transitions, however, the idea of mapping the categorical variables to a continuous space and using established methodologies for multivariate modeling of continuous variables is promising.

The HTPG technique developed for the univariate modeling of categorical variables simplified the modeling of complex geological variables with the utilization of underlying Gaussian latent variables and laid the groundwork for the creation of a multivariate approach. A novel technique based on HTPG for the multivariate modeling of categorical variables is developed in this chapter. The technique is focused on the reproduction of the joint multivariate relationship across the different categorical variables and the improvement of the prediction of any response variables that are sensitive to this relationship. This is achieved by introducing correlation across the Gaussian latent variables utilized for the modeling of each categorical variable.

7.2 Mathematical Notation and Definitions

The mathematical notation outlined in Section 3.1 is further extended here to the multivariate case of HTPG. Consider M random functions (RFs) $\{X_i(\boldsymbol{u});\ \forall \boldsymbol{u} \in \mathcal{A};\ i = 1, \ldots, M\}$, each one with a corresponding finite set $\{\mathcal{B}_i;\ i = 1, \ldots, M\}$ of possible categorical output. Let the cardinality of each categorical set be defined by $\{B_i = |\mathcal{B}|_i ; i = 1, \ldots, M\}$.

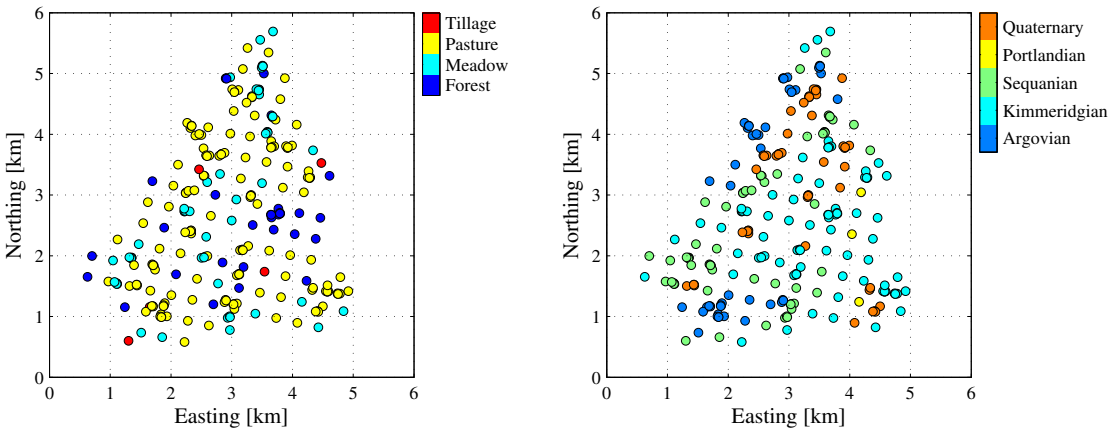
Also, consider M sets of latent variables to be represented by the Gaussian random functions (GRFs) $\{\boldsymbol{Y}_i(\boldsymbol{u}) = (Y_{i,1}(\boldsymbol{u}), \ldots, Y_{i,K_i}(\boldsymbol{u}));\ \forall \boldsymbol{u} \in \mathcal{A};\ i = 1, \ldots, M\}$. Finally, consider the truncation rules to be represented by $\{\mathcal{M}_{\theta_i};\ i = 1, \ldots, M\}$ that defines the mapping $\{\mathcal{M}_{\theta_i} : \mathbb{R}^{K_i} \mapsto \mathcal{B}_i;\ i = 1, \ldots, M\}$ such that $\{\mathcal{M}_{\theta_i}(\boldsymbol{Y}_i(\boldsymbol{u})) = X_i(\boldsymbol{u});\ \forall \boldsymbol{u} \in \mathcal{A};\ i = 1, \ldots, M\}$, where θ_i is the set of parameters that define the truncation rule for the i^{th} categorical variable. Note that for the conventional truncated Gaussian methodologies, only a single categorical variable is considered ($M = 1$).

In some instances, it is more convenient to use a compact notation. The categorical variables are defined by $\{\boldsymbol{X}(\boldsymbol{u}) = (X_i(\boldsymbol{u}), \ldots, X_M(\boldsymbol{u})); \forall \boldsymbol{u} \in \mathcal{A}\}$. The sets of Gaussian latent variables are considered together in a D-dimensional GRF $\{\boldsymbol{Y}(\boldsymbol{u}) = (\boldsymbol{Y}_i(\boldsymbol{u}), \ldots, \boldsymbol{Y}_M(\boldsymbol{u})); \forall \boldsymbol{u} \in \mathcal{A}\}$ where $D = \sum_{i=1}^{M} K_i$. The notation for the truncation rule becomes $M_{\boldsymbol{\theta}}$ that defines the mapping $\{\mathcal{M}_{\boldsymbol{\theta}} : \mathbb{R}^D \mapsto \prod_{i=1}^{M} \mathcal{B}_i\}$ such that $\{\mathcal{M}_{\boldsymbol{\theta}}(\boldsymbol{Y}(\boldsymbol{u})) = \boldsymbol{X}(\boldsymbol{u});\ \forall \boldsymbol{u} \in \mathcal{A}\}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ is the set of parameters for all truncation rules.

7.3 Methodology for the Modeling of Multiple Categorical Variables

In order to illustrate the developed methodology for the multivariate modeling of categorical variables, the well known Swiss Jura data set is used (Goovaerts, 1997). The data set comprises of 360 samples with two categorical variables: Land Use (LU) and Rock Type (RT). As well as seven continuous variables: Cd, Co, Cr, Cu, Ni, Pb, and Zn. The data is often split into two sets, one for prediction with 259 samples and one for validation with 100 samples. Only the categorical variables are required for the demonstration of the developed technique.

A map with the location of the samples is shown in Figure 7.2. The LU variable has four categories: (1) Forest; (2) Meadow; (3) Pasture; and (4) Tillage. The RT variable has five categories: (1) Argovian; (2) Kimmeridgian; (3) Squanian; (4) Portlandian; and (5) Quaternary. The area is covered partially with evenly spaced samples with some areas that are densely sampled with additional clusters of data.



(a) Land Use **(b)** Rock Type

Figure 7.2: Location map of the 259 samples of the prediction subset of the Swiss Jura data

7.3.1 Relationship Between Categorical Variables

Complex relationships amongst continuous geological attributes are often observed in their joint distribution. These complex features should be properly addressed and reproduced by the modeling workflow (Barnett and Deutsch, 2015; Barnett et al., 2014). Similar complexity is also observed in the joint distribution of geological categorical variables, however, its detection may not be as apparent as with the continuous case due to its discrete nature.

Geological knowledge of the processes involved and interactions between the different categorical variables is always a valuable criteria to determine the existence of interdependence between them. The complex relationships can be observed on the categorical joint distribution. If a set of categorical variables are independent from each other, the joint probability density function (PDF) is defined by the product of the marginal probabilities (Equation 7.1).

$$P\left(X_1 = x_1, \ldots, X_M = x_M\right) = \prod_{i=1}^{M} P\left(X_i = x_i\right) \tag{7.1}$$

The developed multivariate approach to HTPG is heavily focused on the reproduction of the multivariate joint PDF. It is important, in this case, to define a representative experimental distribution from data that will serve as reference for the modeling workflow. In geoscience, it is a common practice to oversample few areas of interest while sparsely sampling marginal areas. Declustering techniques are utilized to mitigate the impact of preferential sampling and assist in the definition of a representative distribution. The rep-

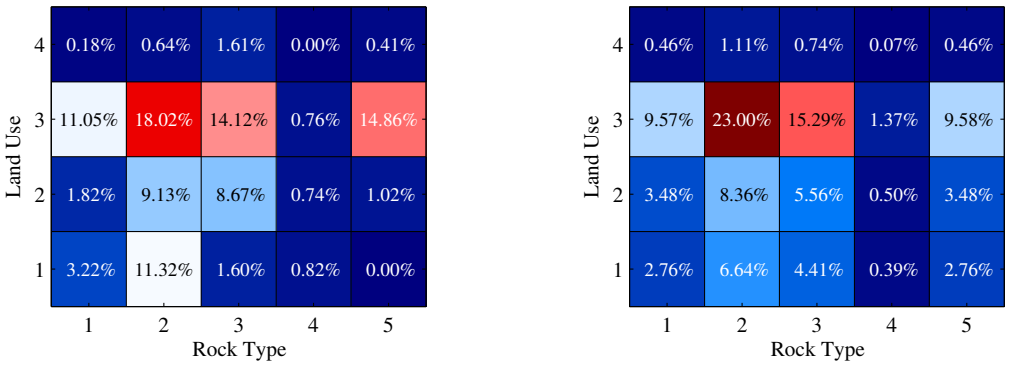
resentative experimental joint PDF calculated from data is compared to the theoretical independent case (Equation 7.1). The departure from the independent PDF is an indication of the existence of complex relationships that must be addressed by the modeling workflow.

Cell declustering (Deutsch et al., 1998) is utilized with the Swiss Jura data set to define the weights for the calculation of the declustered global proportion of each category of both LU and RT variables. The global proportions are shown in Table 7.1. Both LU and RT are unevenly distributed with proportion ranging from 2% and 3% for Tillage LU and Portlandian RT to 59% for Pasture LU and 40% for Kimmeridgian RT.

Table 7.1: Global proportions for Swiss Jura data set

Variable	Category	Code	Global Proportion (%)	
			Clustered	Declustered
Land Use	Forest	1	12.74	16.96
	Meadow	2	21.62	21.37
	Pasture	3	63.71	58.81
	Tillage	4	1.93	2.85
Rock Type	Argovian	1	20.46	16.27
	Kimmeridgian	2	32.82	39.11
	Sequanian	3	24.32	26.00
	Portlandian	4	1.16	2.32
	Quaternary	5	21.24	16.30

The reference declustered bivariate distribution for LU and RT variables (Figure 7.3a) is calculated and compared with the theoretical distribution considering independent variables (Figure 7.3b). The reference distribution calculated from data is fairly different from the distribution expected if the variables were independent.



(a) Reference **(b)** Theoretical for independent

Figure 7.3: Joint distribution of categorical variables LU and RT

The likelihood of finding the most common category of each variable, Pasture LU and Sequanian RT, at the same location is reduced by 5%. The probability of occurrence of Pasture LU in the Quarternary RT is increased by 5%. Also, no occurrence of Forest LU in the Quaternary RT is observed in the data resulting in 0% while the occurrence for the independent case is of nearly 3%. The difference between the two distributions is one of the motivations to consider the multivariate workflow.

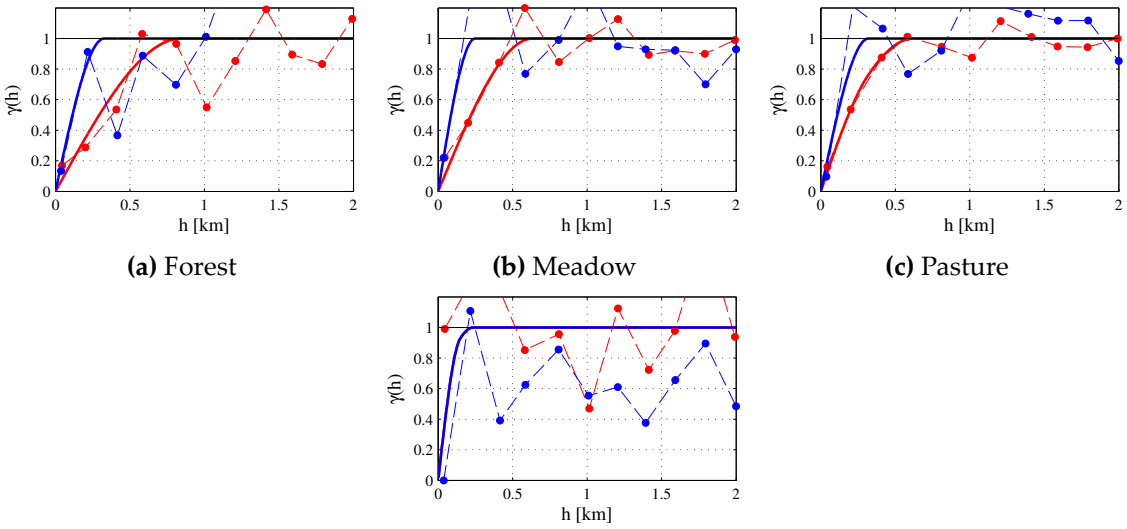
7.3.2 Truncation Rule and Spatial Continuity

The workflow for the multivariate HTPG shares many aspects with its univariate counterpart. The initial steps taken for the the modeling remains unchanged. The definition of truncation rule and the mapping of the spatial continuity from the continuous to categorical space is undertaken as if the categorical variables were to be modeled independently. Geological knowledge, spatial structure, categorical proportion and transition probabilities are utilized to determine the truncation rule. Once the truncation rule is defined, the respective variograms for the Gaussian latent variables are defined utilizing the numerical derivation developed in Section 3.3.4.1 for the univariate HTPG.

For the Jura data set, the spatial continuity is used as the main factor for the definition of the truncation rule. The indicator variograms are calculated for each category of the two variables. The variograms are calculated for the major direction at an azimuth of 67.5^o and minor direction at an azimuth of 157.5^o. The indicator variograms for the LU categories are shown in Figure 7.4. The Tillage LU only has 4 samples (Figure 7.2) scattered over the domain, therefore the range of the variogram was considered isotropic and smaller than the data spacing (Figure 7.4d).

The indicator variogram for the RT categories are shown in Figure 7.5. The variograms of the RT categories are better defined compared to the variograms of the LU categories. The anisotropy is very strong for the Argovian and Sequanian RTs. Similarly to the Tillage LU, the Portlandian RT also only has few scattered samples, therefore, its range is also set to a value smaller than the data spacing.

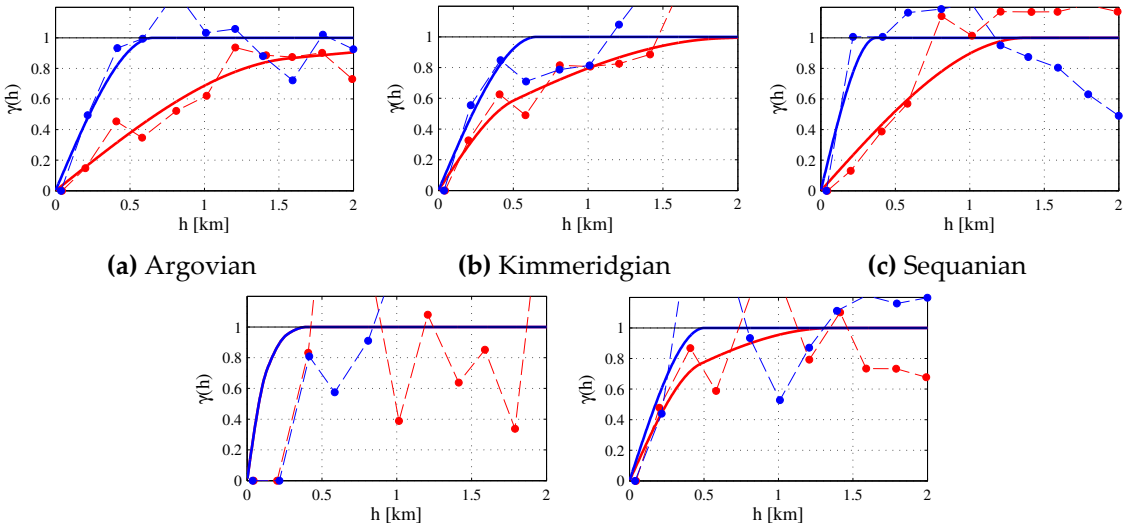
The hierarchical truncation chosen for this example utilizes the most number of Gaussian variables possible for each categorical variable. The LU variable is modeled with 3 Gaussian variables and RT with 4. This choice avoids unwanted restrictions and gives more flexibility to the spatial configuration of the categories as there are not enough data



(a) Forest **(b)** Meadow **(c)** Pasture

(d) Tillage

Figure 7.4: Experimental indicator variograms for LU categories. The red variogram represents azimuth of 67.5^o and blue is used for the azimuth of 157.5^o. Markers represent the experimental variograms and the solid lines are the fitted model. The variograms shown are standardized.

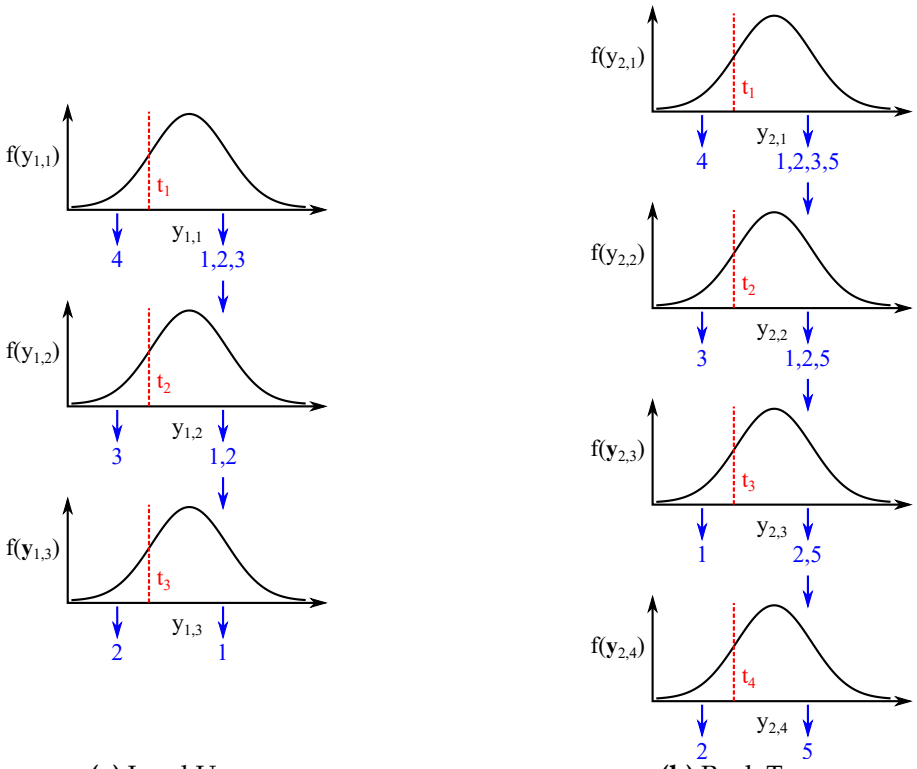


(a) Argovian **(b)** Kimmeridgian **(c)** Sequanian

(d) Portlandian **(e)** Quaternary

Figure 7.5: Experimental indicator variograms for RT categories. The red variogram represents azimuth of 67.5^o and blue is used for the azimuth of 157.5^o. Markers represent the experimental variograms and the solid lines are the fitted model. The variograms shown are standardized.

to support a more complex structure for this particular example. The truncation rule for both LU and RT variables are shown in Figure 7.6. The most different variograms should appear higher in the truncation tree while the most similar variograms should share the final Gaussian variable. For this reason, the Tillage LU and the Portlandian RT are at the top of the truncation structures while Forest and Meadow share the last Gaussian variable for LU and Kimmeridgian and Quaternary share the last Gaussian variable for RT.



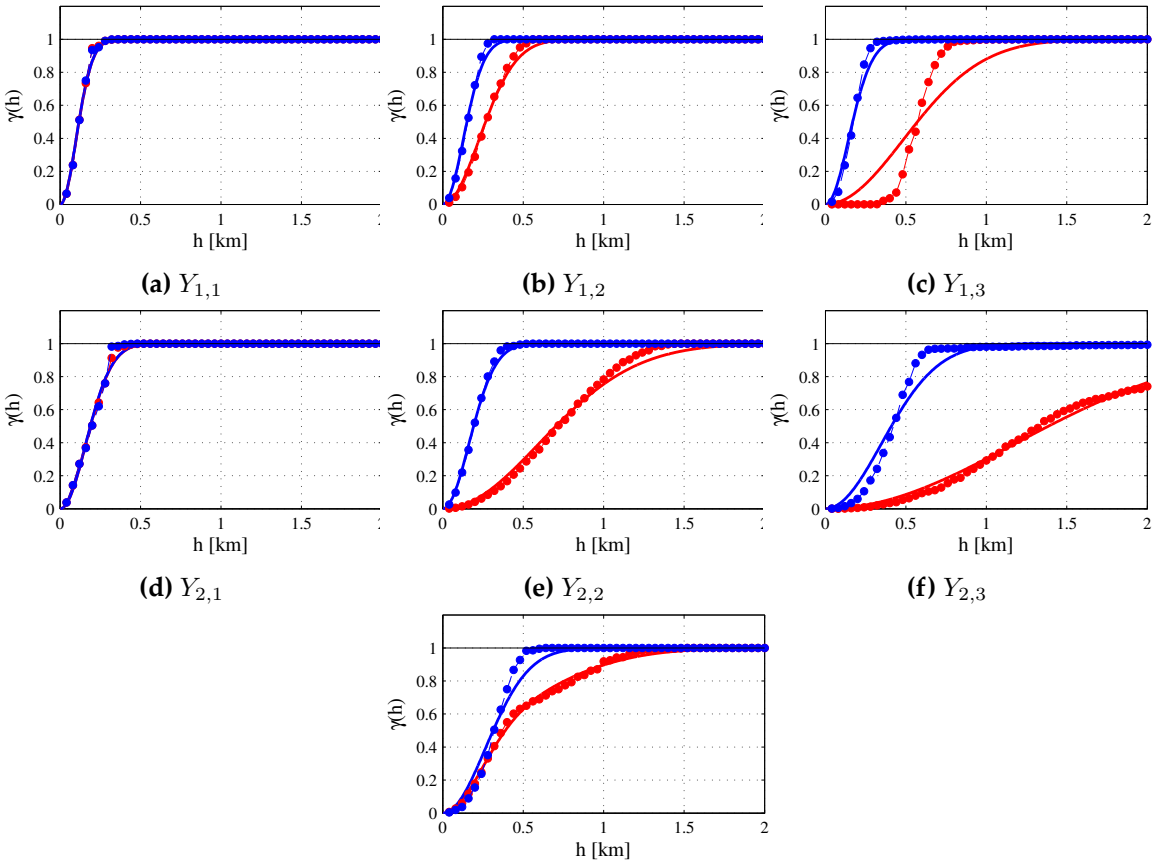
(a) Land Use **(b)** Rock Type

Figure 7.6: Hierarchical truncation scheme for LU and RT variables

The thresholds are defined in the same way as with the univariate HTPG. This is possible because the Gaussian variables are only correlated across different categorical variables and remain independent from each other within the same categorical variable. The covariances between the Gaussian latent variables representing the same categorical variable are zero (Equation 7.2)

$$\text{Cov}\left(Y_{i,j}\left(\boldsymbol{u}\right), Y_{i,k}\left(\boldsymbol{u}+\boldsymbol{h}\right)\right) = 0, \quad i = 1,\ldots,M; \quad j,k = 1,\ldots,K_i; \quad j \neq k \quad \text{and} \quad \forall \boldsymbol{u} \in \mathcal{A} \tag{7.2}$$

The truncation rules and indicator variograms are utilized to define the variogram of the Gaussian latent variables. The numerical approach developed for the univariate HTPG is utilized for the definition of the variogram of the latent variables independently for each categorical variable. The cross-variograms are not required as the latent variables are modeled with an intrinsic model of coregionalization (IMC) approximation. The optimized variograms and the fitted models are shown in Figure 7.7. Most of the optimized values are well fitted by the selected models. The variables $Y_{1,3}$ and $Y_{2,3}$, however, show hyper-continuity which is difficult to fit with valid numeric models.



(a) $Y_{1,1}$ **(b)** $Y_{1,2}$ **(c)** $Y_{1,3}$

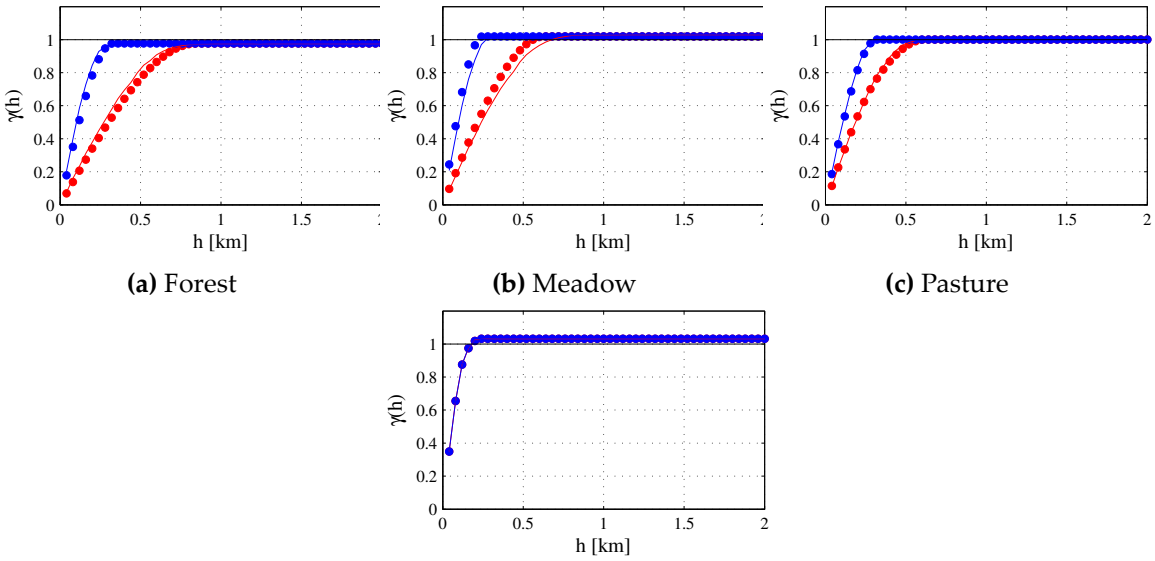
(d) $Y_{2,1}$ **(e)** $Y_{2,2}$ **(f)** $Y_{2,3}$

(g) $Y_{2,4}$

Figure 7.7: Result from variogram calculation. Markers and dashed lines represent the optimized lags and the solid lines represent the fitted models. Red color is utilized for the major direction at azimuth of 67.5^o and blue for the minor direction at 157.5^o.

The quality of the optimization can be evaluated by examining the expected indicator variogram reproduction that is generated by the optimization software. The variogram reproduction for the LU categories is shown in Figure 7.8. Forest and Meadow LUs share the same Gaussian at the last node. This leads to similar indicator variogram for both

categories. This is the reason why the categories with similar experimental variograms are chosen to share the terminating node of the truncation tree. The variogram reproduction for the RT categories is shown in Figure 7.9. Note that Kimmeridgian and Quarternary RTs also share the same terminating node and will have the similar expected indicator variogram reproduction.



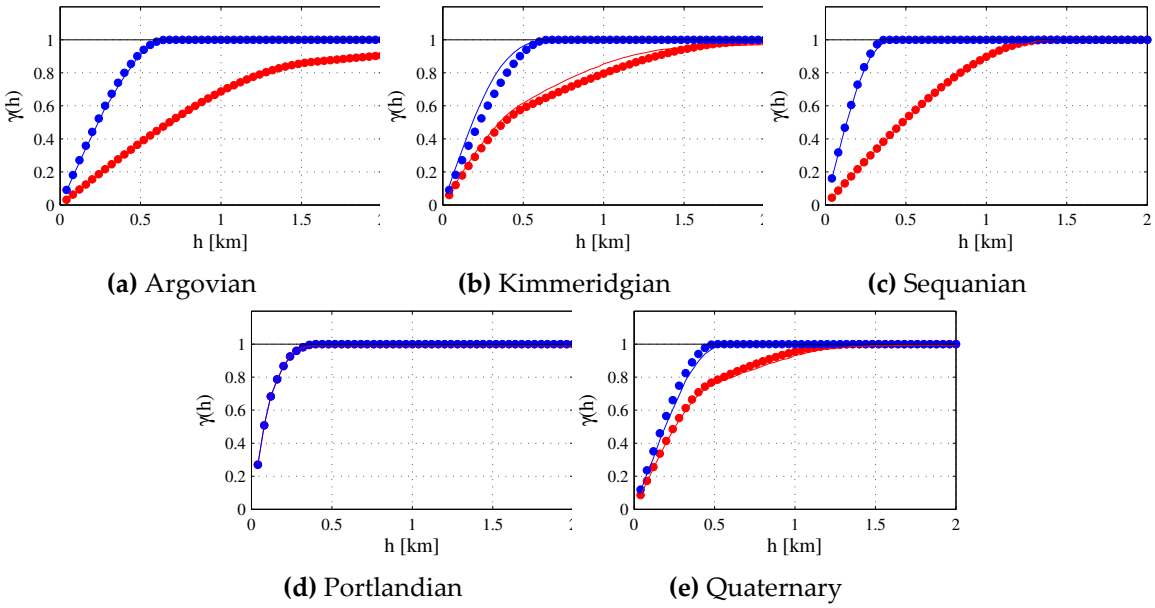
(a) Forest **(b)** Meadow **(c)** Pasture

(d) Tillage

Figure 7.8: Expected indicator variogram reproduction for LU categories. The red variogram represents azimuth of 67.5^o and blue is used for the azimuth of 157.5^o. Markers represent the experimental variograms and the solid lines are the fitted model. The variograms are standardized.

7.3.3 Defining the Correlation Structure

The truncation rules and the mapping of spatial correlation is unchanged from the univariate approach. The multivariate relationships observed on the categorical joint PDF are reproduced by introducing cross-correlation between the latent variables utilized to model different categorical sets. This cross-correlation affects the imputation of the Gaussian latent variables at data locations and the simulation of these variables at the modeling grid nodes. If all cross-correlations between all latent variables utilized in the modeling are set to zero, the resulting joint PDF is expected to reproduce the PDF of the independent case (Figure 7.3b). If the experimental PDF calculated from data show a departure from the independent case, as observed for the Jura data set (Figure 7.3a), the correlation between the latent variables is utilized improve the match with the reference PDF. Conditioning data



(a) Argovian **(b)** Kimmeridgian **(c)** Sequanian

(d) Portlandian **(e)** Quaternary

Figure 7.9: Expected indicator variogram reproduction for RT categories. The red variogram represents azimuth of 67.5^o and blue is used for the azimuth of 157.5^o. Markers represent the experimental variograms and the solid lines are the fitted model. The variograms shown are standardized.

would likely improve reproduction of the joint PDF.

For the Jura data set, for instance, the probability of having Pasture LU at the same location as a Quaternary RT is increased by nearly 5% from the independent case. The categorical code for Pasture LU is 3 and the code for Quaternary RT is 5 (Table 7.1). These categorical codes are defined by the variables $Y_{1,2}$ and $Y_{2,4}$ where these categories are leafs of the truncation tree structure (Figure 7.6). The Pasture LU occurs when $Y_{1,2}$ is low (below the threshold) and the Quaternary RT occurs when $Y_{2,4}$ is high (above the threshold). This suggests that a negative correlation between these two variables would improve the reproduction of the joint PDF across multiple realizations of simulated models. This assumption, however, only considers the local impact of that specific correlation. The definition of a valid correlation matrix that improves the expected reproduction of the joint PDF is a complex optimization problem.

A localized random search optimization algorithm (Spall, 2005) is developed to define the optimal correlation matrix. The randomized search algorithms are highly customizable and are able to perform an extensive exploration of the solution space. If applied well, this family of algorithms are able to provide reasonable solutions at a reasonable computational cost (Silva et al., 2018) to various complex problems for which an analytical treatment is not straightforward.

As requirement for the application of the multivariate HTPG, all Gaussian latent variables must be standard normal and uncorrelated within the same categorical set (Equation 7.3). This means that some of the covariances are set to zero and one and remain unchanged throughout the optimization. The covariance matrix with all collocated covariances between the all Gaussian latent variables will have the structure shown in Equation 7.4, as result of Equation 7.3.

$$\boldsymbol{Y}_i(\boldsymbol{u}) \sim \phi(\boldsymbol{y}_i; \boldsymbol{0}, \boldsymbol{I}), \qquad \boldsymbol{y}_i \in \mathbb{R}^{K_i} \quad \text{and} \quad i \in \{1, \ldots, M\} \tag{7.3}$$

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I}_{K_1} & \boldsymbol{C}_{1,2} & \ldots & \boldsymbol{C}_{1,M} \\ \boldsymbol{C}_{2,1} & \boldsymbol{I}_{K_2} & \ldots & \boldsymbol{C}_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{M,1} & \boldsymbol{C}_{M,2} & \ldots & \boldsymbol{I}_{K_M} \end{bmatrix} \tag{7.4}$$

where $\boldsymbol{C}$ is a $D \times D$ matrix containing all the covariances between all latent variables; $\{I_{K_i}; 1 = 1, \ldots, M\}$ are identity matrices of order $\{K_i; 1 = 1, \ldots, M\}$; and $\{\boldsymbol{C}_{i,j}; i, j = 1, \ldots, M; i \neq j\}$ are sub-matrices of $\boldsymbol{C}$ containing the covariances to be optimized.

The covariance matrix $\boldsymbol{C}$ must be symmetric. This entails that $\left\{\boldsymbol{C}_{i,j} = \boldsymbol{C}_{j,i}^{\top}; \forall i, j = 1, \ldots, M\right\}$. As a result, the number n_{c} of covariance terms being optimized is defined by Equation 7.5. These covariance values can be mapped into an one dimensional vector with linear indexing ($\boldsymbol{c} = (c_i, \ldots, c_{n_{\mathrm{c}}})$) to facilitate the notation. Each one of these variables are mapped into two locations in $\boldsymbol{C}$, one at the upper triangle and one at the lower triangle (symmetry), outside of the fixed identity sub-matrices.

$$n_{\mathrm{c}} = \sum_{i=1}^{M-1} K_i \times \left(\sum_{j=i}^{M} K_j \right) \tag{7.5}$$

The joint categorical PDF can be defined by a $B_1 \times B_2 \times \ldots \times B_M$ probability matrix $\boldsymbol{P}$. Let $\boldsymbol{p} = \left(p_i, \ldots, p_{n_{\mathrm{p}}}\right) = \mathrm{vec}(\boldsymbol{P})$ be the vector with all reference probabilities calculated from data, where $n_{\mathrm{p}} = \prod_{i=1}^{M} B_i$.

A Monte-Carlo simulation (MCS) approach is utilized for the evaluation of the objective function, similarly to the procedure for the numerical derivation of the variograms of latent variables. A random set of Gaussian deviates are simulated and correlated utilizing the respective state of the covariance matrix at each iteration of the algorithm. The

simulated latent values are truncated in accordance with the truncation rules to define categorical realizations. The joint distribution of the simulated categorical set is evaluated and the mismatch between the PDF of simulated values and the reference PDF is utilized to calculate the objective function.

A set of size $D \times m$ containing m realizations of independent random vectors $\boldsymbol{z} = \left(\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_m^\top\right)$ are sampled from the standard Gaussian distribution $\left(\boldsymbol{Z} \sim \phi\left(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}\right); \boldsymbol{z} \in \mathbb{R}^D\right)$. This set is utilized multiple times throughout the optimization and can be defined and stored beforehand to save computational time. A set of correlated vectors $\boldsymbol{y} = \left(\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_m^\top\right)$ of size $D \times m$ is generated from the independent set $\boldsymbol{z}$, utilizing the lower triangle matrix resulting from the Cholesky decomposition of the covariance matrix (Equation 7.6).

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{z} \qquad \text{where} \qquad \boldsymbol{C} = \boldsymbol{L}^\top \boldsymbol{L} \tag{7.6}$$

The set of simulated categorical variables of size $M \times m$ is generated by applying the truncation rule to each simulated sample of the latent variables (Equation 7.7). A probability matrix $\hat{\boldsymbol{P}}\left(\boldsymbol{c}\right)$ can be calculated from the categorical set $\boldsymbol{x} = \left(\boldsymbol{x}_i^\top, \ldots, \boldsymbol{x}_m^\top\right)$ simulated using $\boldsymbol{c}$. The probability matrix is used to define $\hat{\boldsymbol{p}}\left(\boldsymbol{c}\right) = \left(\hat{p}_i\left(\boldsymbol{c}\right), \ldots, \hat{p}_{n_\mathrm{p}}\left(\boldsymbol{c}\right)\right) = \mathrm{vec}\left(\hat{\boldsymbol{P}}\left(\boldsymbol{c}\right)\right)$. The objective function is then defined by the Equation 7.8 as function of the covariance terms being optimized.

$$\boldsymbol{x}_i = M_{\boldsymbol{\theta}}\left(\boldsymbol{y}_i\right), \qquad i = 1, \ldots, m \tag{7.7}$$

$$\mathrm{O}\left(\boldsymbol{c}\right) = \sum_{i=1}^{n_\mathrm{p}} [p_i - \hat{p}_i\left(\boldsymbol{c}\right)]^2, \qquad i = 1, \ldots, m \tag{7.8}$$

The optimization algorithm requires the initialization of the covariances. The initialization can be performed in two ways. One is the random initialization that sets the initial covariances to small random deviates uniformly distributed between -0.1 and 0.1. Small numbers are utilized to ensure that the initial covariance matrix at iteration $t = 0$ is a valid positive semi-definite matrix. The second option is to independently simulate the Gaussian deviates at data locations subject to the truncation rule and categorical observations and calculate the resulting covariance matrix. A large number of data realizations are utilized to ensure stable initial covariance values. The second option often generates better initial

correlation values. Even though the random deviates are generated without correlation, the categorical data observation together with the truncation structure often introduces enough structure to allow faster convergence.

The localized random search algorithm proposed for the optimization of the objective function in Equation 7.8 is outlined below:

(1) Set iteration counter to zero: $t = 0$

(2) Initialize the set of covariances: $\boldsymbol{c}^{(0)} = \left(c_1^{(0)}, \ldots, c_{n_c}^{(0)}\right)$

(3) Evaluate the initial objective function: $\mathrm{O}\left(\boldsymbol{c}^{(0)}\right)$

(4) While $t < t_{\max}$

(4.1) Increment iteration counter: $t = t + 1$

(4.2) Set $\boldsymbol{c}^{(t)} = \boldsymbol{c}^{(t-1)}$

(4.3) Generate a random deviate $u \sim \mathcal{U}(0, 1)$

(4.4) Randomly select a index i between 1 and n_c

(4.5) Update the covariance element: $c_i^{(t)} = c_i^{(t-1)} \times \left(1 + \frac{t \times (u - 0.5)}{t_{\max}}\right)$

(4.6) If $\boldsymbol{C}^{(t)}$ is not positive definite, reset the covariance value to $c_i^{(t)} = c_i^{(t-1)}$ and go to step 4.3

(4.7) Evaluate objective function $\mathrm{O}\left(\boldsymbol{c}^{(t)}\right)$

(4.8) Keep the change if the objective is improved or reset covariance to previous state otherwise:

$$c_i^{(t)} = \begin{cases} c_i^{(t)}, & \text{if } \mathrm{O}\left(\boldsymbol{c}^{(t)}\right) \leq \mathrm{O}\left(\boldsymbol{c}^{(t-1)}\right) \\ c_i^{(t-1)}, & \text{otherwise} \end{cases}$$

(5) For $i = 1, \ldots, n_c$ do:

(5.1) Increment counter: $t = t + 1$

(5.2) Set selected covariance to zero: $c_i^{(t)} = 0$

(5.3) If $\boldsymbol{C}$ is not positive definite, reset the covariance value to $c_i^{(t)} = c_i^{(t-1)}$ and cycle the loop to next index i

(5.4) Evaluate objective function $\mathrm{O}\left(\boldsymbol{c}^{(t)}\right)$

(5.5) Keep the change if the objective is improved or reset covariance to previous state otherwise:

$$c_i^{(t)} = \begin{cases} 0, & \text{if } \mathrm{O}\left(\boldsymbol{c}^{(t)}\right) \leq \mathrm{O}\left(\boldsymbol{c}^{(t-1)}\right) \\ c_i^{(t-1)}, & \text{otherwise} \end{cases}$$

The first set of iterations (step 4) is used to define a reasonable solution for the covariance matrix, however, the nature of the iterations may lead to unnecessary covariances adding unwanted complexity to the problem. In order to eliminate any non-zero covariance values that are not contributing to the solution a second set of iterations over all the covariances is utilized, starting in step 5. This second round of iterations will set the covariances to zero wherever the change does not improve the solution.

This is a greedy algorithm and perhaps not the most efficient, however, it is able to find reasonable solutions within a reasonable amount of time. The algorithm is stochastic in nature and the results from multiple runs are non-unique. The utilization of multiple random restarts is recommended to avoid unreasonable solutions from local minima. The global minimum is unlikely to be found, however, the algorithm is capable of finding solutions within very small deviations from the reference joint PDF.

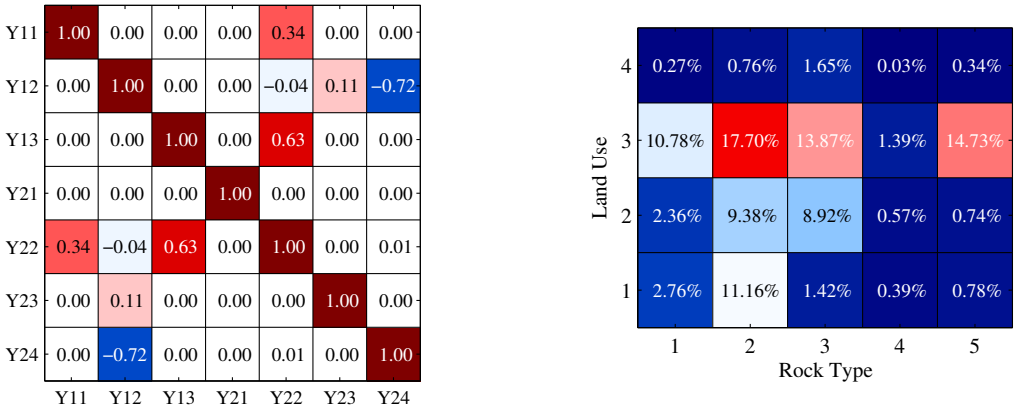
The algorithm is applied to the Swiss Jura data set example. The algorithm is initialized using the covariance matrix calculated from 100 realizations of independently simulated latent variables at each categorical data observation. The initial correlation matrix is shown in Figure 7.10

	Y11	Y12	Y13	Y21	Y22	Y23	Y24
Y11	1.00	−0.00	0.00	−0.00	0.03	−0.02	−0.00
Y12	−0.00	1.00	0.02	−0.02	−0.02	0.05	−0.14
Y13	0.00	0.02	1.00	−0.00	0.08	0.00	−0.03
Y21	−0.00	−0.02	−0.00	1.00	0.00	−0.00	0.01
Y22	0.03	−0.02	0.08	0.00	1.00	−0.01	−0.01
Y23	−0.02	0.05	0.00	−0.00	−0.01	1.00	−0.01
Y24	−0.00	−0.14	−0.03	0.01	−0.01	−0.01	1.00

Figure 7.10: Initial correlation matrix utilized to initialize the optimization algorithm for the Swiss Jura data set.

The optimization is run utilizing the declustered distribution (Figure 7.3a) as reference. The software is allowed to run for 10,000 iterations ($t_{\max}$), 10,000 samples (m) and 120 random restarts to ensure a solution that is close to the reference. The error measure utilized for the optimization is the root of the sum of squared error (RSSE) between each bin of the categorical joint distribution of the simulated realizations and the reference distribution. The best solution found has a RSSE of 1.5%. The correlation resultant from optimization

and the respective expected joint distribution are shown in Figure 7.11. The obtained distribution (Figure 7.11b) is close to the reference (Figure 7.3a). The Gaussian variables representing the same categorical variables remained uncorrelated. This is required to ensure the reproduction of categorical spatial correlation and proportions.



(a) Correlation **(b)** Optimized Distribution

Figure 7.11: Results from the correlation matrix optimization for the Jura data set.

7.3.4 Gibbs Sampler Algorithm for Multivariate HTPG

The latent variables must be jointly simulated to allow the reproduction of the covariance matrix and consequently the reproduction of the joint categorical PDF. The joint simulation introduces further complexity to the imputation of latent variables at data locations. The Gibbs sampler algorithm must be adapted to accommodate and enforce the correlation between the latent variables. The utilization of an IMC with a merged secondary attribute (Babak and Deutsch, 2008) is proposed for the adaptation of the standard Gibbs sampler algorithm (Section 5.2.1).

Consider a set of n observations $\boldsymbol{x} = \left(\boldsymbol{x}^\top(\boldsymbol{u}_1), \ldots, \boldsymbol{x}^\top(\boldsymbol{u}_n)\right)$ of a M-dimensional categorical random variable $\boldsymbol{X}(\boldsymbol{u})$, where $\boldsymbol{u}$ is the vector of spatial coordinates. Also, consider the set of D latent variables at all data locations to be represented by $\boldsymbol{y} = \left(\boldsymbol{y}^\top(\boldsymbol{u}_1), \ldots, \boldsymbol{y}^\top(\boldsymbol{u}_n)\right)$ where $\boldsymbol{y}(\boldsymbol{u}_i) \in \mathbb{R}^D$ and is a realization of the Gaussian random variable $\boldsymbol{Y}(\boldsymbol{u}_i) = (Y_1(\boldsymbol{u}_i), \ldots, Y_D(\boldsymbol{u}_i))$ at i^{th} data location. The latent variables $\boldsymbol{Y}(\boldsymbol{u}_i)$ $(i = 1, \ldots, n)$ are centered Gaussian variables with covariance matrix $\boldsymbol{C}$. For notation simplicity the vector of spatial coordinates $\boldsymbol{u}_i$ will be omitted and represented by the subscript i. Finally, consider the truncation rule to be represented by $\mathcal{M}_{\boldsymbol{\theta}}$ such that $\mathcal{M}_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = \boldsymbol{x}_i$ $(i = 1, \ldots, n)$. The proposed Gibbs

sampler algorithm for the multivariate HTPG proceeds by:

(1) Set iteration counter to zero: $t = 0$

(2) Initialize the set of latent variables $\boldsymbol{y}^{(t)}$ with arbitrary values subject to the mapping constraint: $\mathcal{M}_{\boldsymbol{\theta}}(\boldsymbol{y}^{(t)}) = \boldsymbol{x}$

(3) Increment iteration counter: $t = t + 1$

(4) Set $\boldsymbol{y}^{(t)} = \boldsymbol{y}^{(t-1)}$

(5) Define a random path through data points

(6) For each location $i \in \{1, \ldots, m\}$ in the random path do:

(6.1) Loop for each variable $j = 1, \ldots, D$

(6.1.1) Merge all variables $\{k = 1, \ldots, D; k \neq j\}$ into a secondary attribute at each neighboring location within the set N_i:

$$y_{\mathrm{s}_{j,l}}^{(t)} = \frac{\sum_{\substack{k=1 \\ k \neq j}}^{D} \omega_k^{(j)} y_{k,l}^{(t)}}{c_{\mathrm{ps}_j}}, \qquad \forall l \in N_i \tag{7.9}$$

$$c_{\mathrm{ps}_j} = \sum_{\substack{k=1 \\ k \neq j}}^{D} \omega_k^{(j)} c_{j,k} \tag{7.10}$$

$$\sum_{\substack{k=1 \\ k \neq j}}^{D} \omega_k^{(j)} c_{k,k'} = c_{j,k'}, \qquad k' = 1, \ldots, D \quad \text{and} \quad k' \neq j \tag{7.11}$$

where $y_{\mathrm{s}_{j,l}}^{(t)}$ is the merged secondary attribute; $c_{j,k}$ is the covariance between the j^{th} and k^{th} variable from the optimized matrix of covariances (Equation 7.4); c_{ps_j} is the new covariance between the merged secondary attribute and the primary when j is the variable being updated (deemed primary).

(6.1.2) Update the selected variable's value by a new Gaussian value drawn from the conditional distribution given the surrounding data and merged attributes:

$$y_{j,i}^{(t)} \sim \phi\left(y; \bar{\mu}_{j,i}, \bar{\sigma}_{j,i}\right) \tag{7.12}$$

$$\bar{\mu}_{j,i} = \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{p}_{l,i}} \times y^{(t)}_{\mathrm{p}_{j,l}} + \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{s}_{l,i}} \times y^{(t)}_{\mathrm{s}_{j,l}} + \lambda^{(j)}_{\mathrm{s}_{i,i}} \times y^{(t)}_{\mathrm{s}_{j,i}} \tag{7.13}$$

$$\bar{\sigma}^2_{j,i} = 1 - \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{p}_{l,i}} \times C^{(j)}_{\mathrm{pp}_{l,i}} + \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{s}_{l,i}} \times C^{(j)}_{\mathrm{ps}_{l,i}} + \lambda^{(j)}_{\mathrm{s}_{i,i}} \times C^{(j)}_{\mathrm{ps}_{i,i}} \tag{7.14}$$

$$\begin{aligned}
\sum_{l \in N_i} \lambda^{(j)}_{\mathrm{p}_{l,i}} \times C^{(j)}_{\mathrm{pp}_{l,m}} + \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{s}_{l,i}} \times C^{(j)}_{\mathrm{ps}_{l,m}} + \lambda^{(j)}_{\mathrm{s}_{i,i}} \times C^{(j)}_{\mathrm{ps}_{i,i}} &= C^{(j)}_{\mathrm{pp}_{m,i}}, \quad \forall m \in N_i \\
\sum_{l \in N_i} \lambda^{(j)}_{\mathrm{p}_{l,i}} \times C^{(j)}_{\mathrm{ps}_{l,m}} + \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{s}_{l,i}} \times C^{(j)}_{\mathrm{ss}_{l,m}} + \lambda^{(j)}_{\mathrm{s}_{i,i}} \times C^{(j)}_{\mathrm{ss}_{i,i}} &= C^{(j)}_{\mathrm{ps}_{m,i}}, \quad \forall m \in N_i \\
\sum_{l \in N_i} \lambda^{(j)}_{\mathrm{p}_{l,i}} \times C^{(j)}_{\mathrm{ps}_{l,i}} + \sum_{l \in N_i} \lambda^{(j)}_{\mathrm{s}_{l,i}} \times C^{(j)}_{\mathrm{ss}_{l,i}} + \lambda^{(j)}_{\mathrm{s}_{i,i}} \times C^{(j)}_{\mathrm{ss}_{i,i}} &= C^{(j)}_{\mathrm{ps}_i}
\end{aligned} \tag{7.15}$$

where $y^{(t)}_{\mathrm{p}_{j,l}} = y^{(t)}_{j,l}$ is the original variable being updated and deemed to be the primary attribute; $\bar{\mu}_{i,j}$ and $\bar{\sigma}_{i,j}$ are the mean and variance parameters used for the updating of the j^{th} variable at the i^{th} data location; $C^{(j)}_{\mathrm{pp}_{l,i}} = C^{(j)}_{l,m}$ is the original spatial covariance between the locations l and m for variable j (primary); $C^{(j)}_{\mathrm{ps}_{l,m}} = c_{\mathrm{ps}_j} \times C^{(j)}_{l,m}$ is the intrinsic spatial covariance between the primary variable at location l and merged attribute at location m when the variable being updated is j; $C^{(j)}_{\mathrm{ss}_{l,m}}$ is the covariance between the merged secondary attributes at locations l and m. This covariance is the same as $C^{(j)}_{\mathrm{pp}_{l,m}}$ under IMC.

(6.1.3) Test if the updated value of $y^{(t)}_{j,i}$ satisfies the condition $\mathcal{M}_{\boldsymbol{\theta}}\left(\boldsymbol{y}^{(t)}_i\right) = \boldsymbol{x}_i$. If the mapping is not satisfied, the sampling is repeated until the conditions are met before moving to the next variable. Note that depending on the map $\mathcal{M}_{\boldsymbol{\theta}}$, the boundaries for $y^{(t)}_{j,i}$ can be calculated in advance and applied to constrain the distribution in Equation 7.12 to avoid multiple sampling attempts.

(7) Repeat steps from 3 for a maximum number of iterations.

Only the direct variograms of each Gaussian latent variable are required with the intrinsic assumption. As a result, there is no additional work for the definition of variograms

when compared to the univariate HTPG. The utilization of a merged secondary attribute reduces the computational cost by lowering the dimensionality of the covariance matrices that requires inversion. As with the univariate case of the standard Gibbs sampler, the utilization of a restricted neighborhood (N_i) leads to approximate parameters for the conditional distribution in Equation 7.12 and can potentially result in convergence issues. The propagative Gibbs sampler could be extended for the multivariate case, however, the requirement to store all covariances between all data locations and all variables would render it not feasible for practical applications.

The results of the proposed algorithm must be checked for convergence. The convergence is only a concern for the data imputation step and it is not a problem with the simulation at the modeling grid. In cases where the imputation with the proposed approach is not stable, the independent Gibbs sampler from univariate case could be utilized for the imputation while the correlation is considered during the modeling at the grid nodes. The impact of correlation on the imputation of multiple Gaussian latent variables is investigated in Section 7.4.

7.3.5 Simulation of Latent Variables and Mapping to Categorical Space

After the imputation of the latent variables at the data locations, the imputed data are used to condition the simulation at the modeling grid. The same IMC approach with merged secondary attribute utilized in the proposed multivariate Gibbs sampler is also utilized for the simulation at the modeling nodes. The simulation is performed sequentially for each latent variable. During the simulation of a given latent variable, all the previous variables already simulated are merged into a secondary attribute and utilized with the IMC.

The ordering of the latent variable is important in this case as the IMC is an approximation and does not necessarily represent the true nature of the spatial structure. The latent variables must be ordered accordingly with the most important categorical variables coming first. The IMC assumption has no effect on the modeling of the latent variables of the first categorical variable, as the latent variables within one categorical variable are all independent from each other. The effects of the IMC approximation, if any, are only observed for the latent variables of the second and subsequent categorical variables. Once all latent variables are simulated at all grid nodes, the truncation rule is applied for the definition of the categorical realizations.

7.3.6 Results for Jura Data Set

The data imputation is run with 10,000 iterations of the standard Gibbs sampler approach to generate 100 realizations of data. The full neighborhood search is used to ensure stability of the algorithm. The imputed data is utilized as conditioning data for the subsequent modeling in order to transfer the uncertainty of the unsampled latent variables. Two sets of imputed data are generated. One set uses the optimized correlation matrix with intrinsic model and supersecondary variables and the other set has independent latent variables. These two sets are compared.

The sequential Gaussian simulation (SGS) algorithm is used to simulate the latent variables at grid locations. The 2D grid have 173x205 blocks of 25x25 m. Five sets of realizations are generated. These sets are utilized to compare the multivariate modeling approach with the independent approach. These sets are: (1) correlated conditioning data with correlated SGS; (2) independent data with correlated SGS; (3) independent data with independent SGS; (4) unconditional correlated SGS; (5) unconditional correlated SGS (6). The methodology proposed here is better represented by set 1, however, the utilization of the Gibbs sampler with a large number of data may be unpractical and the user might only be able to generate independent latent variables (set 2). The sets 1 and 2 are compared to the current approach (set 3) in which both data imputation and simulation are performed with independent latent variables. As the conditioning is very strong due to dense data. The effects on unconditional realizations are also checked (sets 4 and 5).

The joint probability reproduction of each realization of each one of the described sets are checked. The results are summarized in box-plots and shown in Figure 7.12. The proposed framework resulted in the least amount of error (7.7%) in average, however, it shows a large variability in the error magnitude. The average error for the case with independent data and correlated simulation is slightly higher than the proposed methodology at 8.7%. The conventional independent modeling resulted in the highest error amongst the conditional realizations sitting at 9.5% that represents a 23.5% increase in error compared to the proposed framework. The error is pushed down due to the effect of conditioning data that helps reproducing the joint proportion. The error difference between the proposed multivariate approach and the conventional independent approach is better visualized with the unconditional realizations. The error of the unconditional independent simulation is 60% higher then the amount for the unconditional multivariate simulation. The mean er-

ror for the unconditional multivariate approach is lower than for conditional independent approach.

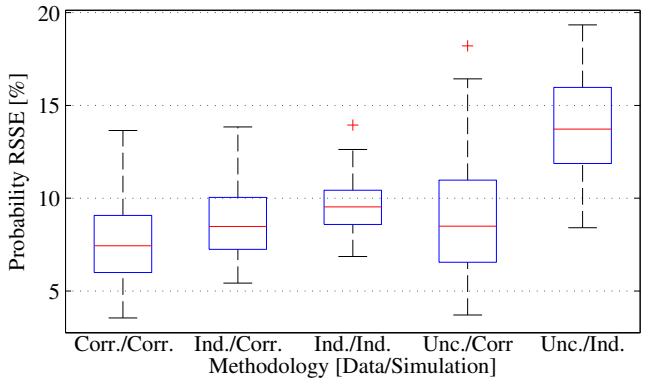


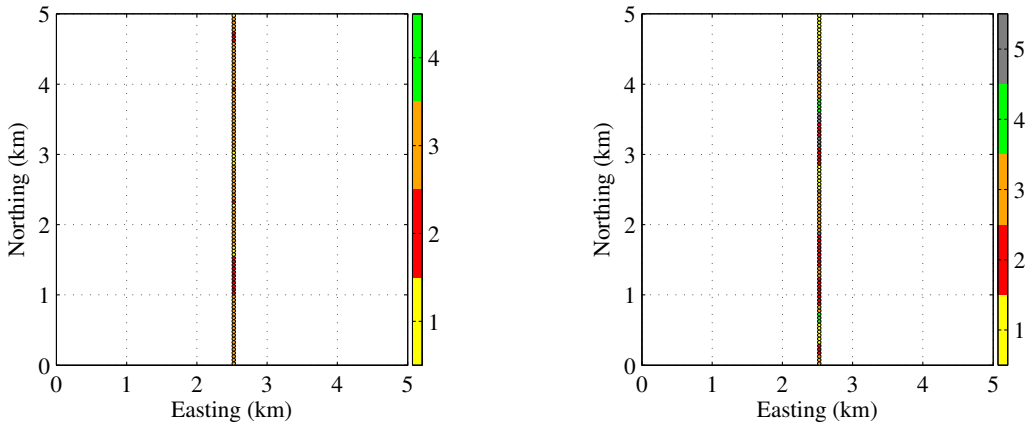
Figure 7.12: Summary of RSSE results for each testing set.

7.4 Effects of Data Correlation on Multivariate Categorical Modeling

A synthetic example is created to further investigate the effects of data correlation on the proposed multivariate HTPG. A grid with 100x100 cells of 50x50m each is defined for discretization of an area of 5x5 km in easting and northing directions. The the categorical variables and modeling parameters are defined based on the Swiss Jura data set. Unconditional simulation is utilized to generate 1,000 reference models. A high number of realizations is used to ensure smooth results. The reference models are sampled with a string of data containing all data for a fixed central location at easting direction (Figure 7.13). This creates a string of data oriented from south to north.

Three cases are built for comparison. The first case utilizes the standard Gibbs sampler considering the cross-correlation for data imputation and also utilizes the cross-correlation for the simulation at grid nodes. The second case utilizes the independent Gibbs sampler to perform the data imputation, however, the simulation at the grid nodes is performed using the correlation between the Gaussian latent variables. The third case utilizes independent modeling for the imputed data and for the simulation over the grid. All three cases are compared against the reference models.

For each of the three cases 1,000 realizations were generated. The joint proportion reproduction is checked for each location in easting direction by evaluating the error for all



(a) Variable 1 **(b)** Variable 2

Figure 7.13: Sample locations with categorical observations for variable 1 and 2 for realization 1 of the reference models.

realizations and all locations in northing direction. This results in the error as function of the easting coordinate. The RSSE of each bin of the joint distribution is utilized as error measure. The results for all cases including the reference is shown in Figure 7.14 and the error relative to the reference model is shown in Figure 7.15.

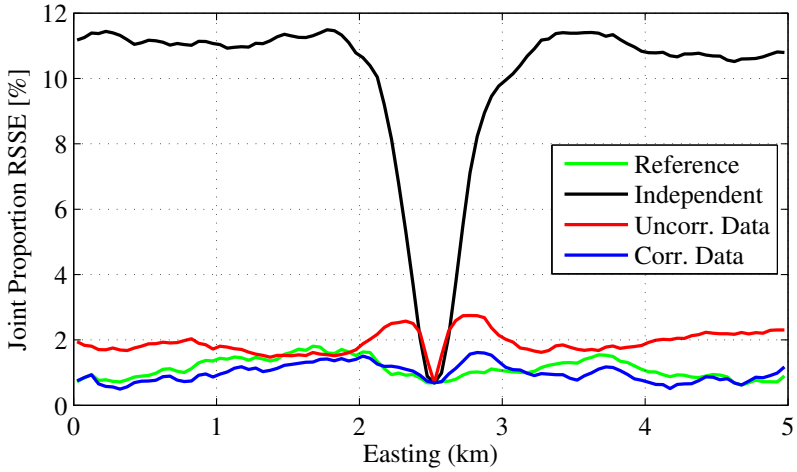


Figure 7.14: Sample locations with categorical observations for variable 1 and 2 for realization 1 of the reference models.

All cases have have the same error as the reference models at data location as the imputation is constrained to match the categorical data observations. This error is due to ergodic fluctuations in the categorical joint distribution. As the distance from the data increases, the error is different for each case. The first case with correlated imputation has the same error level as the reference models. This indicates that the standard Gibbs sampler converged

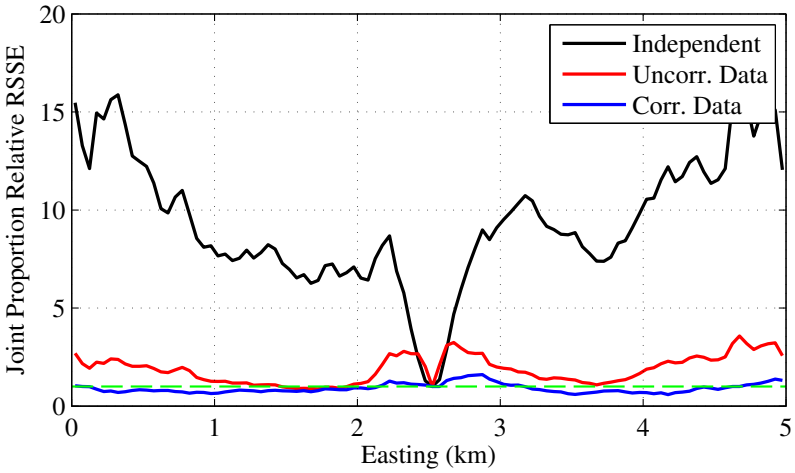


Figure 7.15: Sample locations with categorical observations for variable 1 and 2 for realization 1 of the reference models.

to the appropriate correlation between the variables (Figure 7.11). The error levels for the second and third cases are higher than the reference as the imputation does not consider correlation. The error for the second case is comparable to the third case up to a short distance from data and it starts to decline as the distance to the data increases. This happens as the effect of conditioning data gets less strong. In the third case, the error only increases until reaching the error expected for completely independent categorical variables around 11%. For this example the error from completely independent simulation reaches levels up to 15 times higher than the expected error from ergodic fluctuations (Figure 7.15). Not considering correlation during the imputation, but utilizing it for the simulation at grid locations, leads to errors up to 3 times higher than the reference models.

These results suggest that the joint modeling of categorical variables should be performed even if it is not possible to introduce the correlation during the imputation step. The joint probability of independent categorical variables can be calculated from the univariate proportions and compared to the declustered joint probability calculated from data. The multivariate modeling of categorical variables should be considered if these two distributions are different.

7.5 Conclusion

A methodology for multivariate categorical modeling using the HTPG framework is developed. The technique is aimed at the reproduction of the joint categorical proportion.

The Swiss Jura data set is used as an illustrative example to show that the methodology improves the reproduction of the joint distribution when compared to the conventional independent modeling. The reproduction is controlled by the correlation between the underlying latent variables that are utilized for the modeling of each categorical variable. The correlation is considered during the simulation through the use of IMC and merged secondary attributes.

The imputation of Gaussian latent variables with correlation when dealing with large data sets is not always possible due to computational limitations. The effect of the data correlation on the multivariate HTPG is investigated. The correlation between Gaussian latent variables is important for the reproduction of the joint categorical distribution. In cases that the correlation cannot be introduced in the imputation step due to computational constraints, it is recommended that the simulation at grid nodes are still performed with correlation as it reduces the error on the reproduction of joint distribution. The proposed methodology is demonstrated for a full scale practical application in Chapter 8.

Chapter 8

Case Study: Categorical Modeling at Red Dog Mine

The hierarchical truncated pluri-Gaussian (HTPG) approach for categorical modeling developed in this dissertation along with its extension to the multivariate case are demonstrated in this chapter for a data set from the Red Dog Mine. The univariate and multivariate HTPG approaches are applied and compared with the sequential indicator simulation (SIS) technique. The results are compared through a test data set left out of the modeling workflow and the advantages and disadvantages of each method are highlighted. The impact of considering or not a multivariate workflow for the modeling of multiple categorical variables is assessed by comparing the predicted metallurgical recovery against the real recovery for the test data set.

8.1 Background

The Red Dog Mine is located in the Western Brooks Range in northern Alaska and is one of the world's largest producers of zinc concentrate. The deposits in the Red Dog district are described as shale-hosted massive sulfide Zn-Pb-Ag deposits formed by sedimentary exhalative (SEDEX) and replacement processes (Kelley and Jennings, 2004; Moore et al., 1986). The deposits in Red Dog have a strong structural control as a result of thrust faults that formed a stacking structure. Within the Red Dog thrust plate, the mineralization is separated by thrust sheets (plates). The deposits in Red Dog district are characterized by veins and breccias at the lower portions followed by massive sulfide, silica rock or silicified barite and barite with high to low sulphide content at the upper portions (Krolak et al., 2017).

The metallurgy in Red Dog mine is complex and sensitive to the multiple ore types. The high relative iron content as pyrite has a negative impact on metallurgical recovery and silica is the most important variable affecting the metallurgical throughput (Krolak et al., 2017). Weathering is also an important factor affecting the overall metallurgy at Red

Dog deposits (Krolak et al., 2017).

8.2 Available Data

The data set available for this case study consists of a total of 1134 drillholes and 41,343 samples composited to a length of 12.5 ft and distributed at a nominal 100 ft spacing. The location of the available drillholes are shown in Figure 8.1. The parameters of the block model utilized for the modeling in this chapter is shown in Table 8.1.

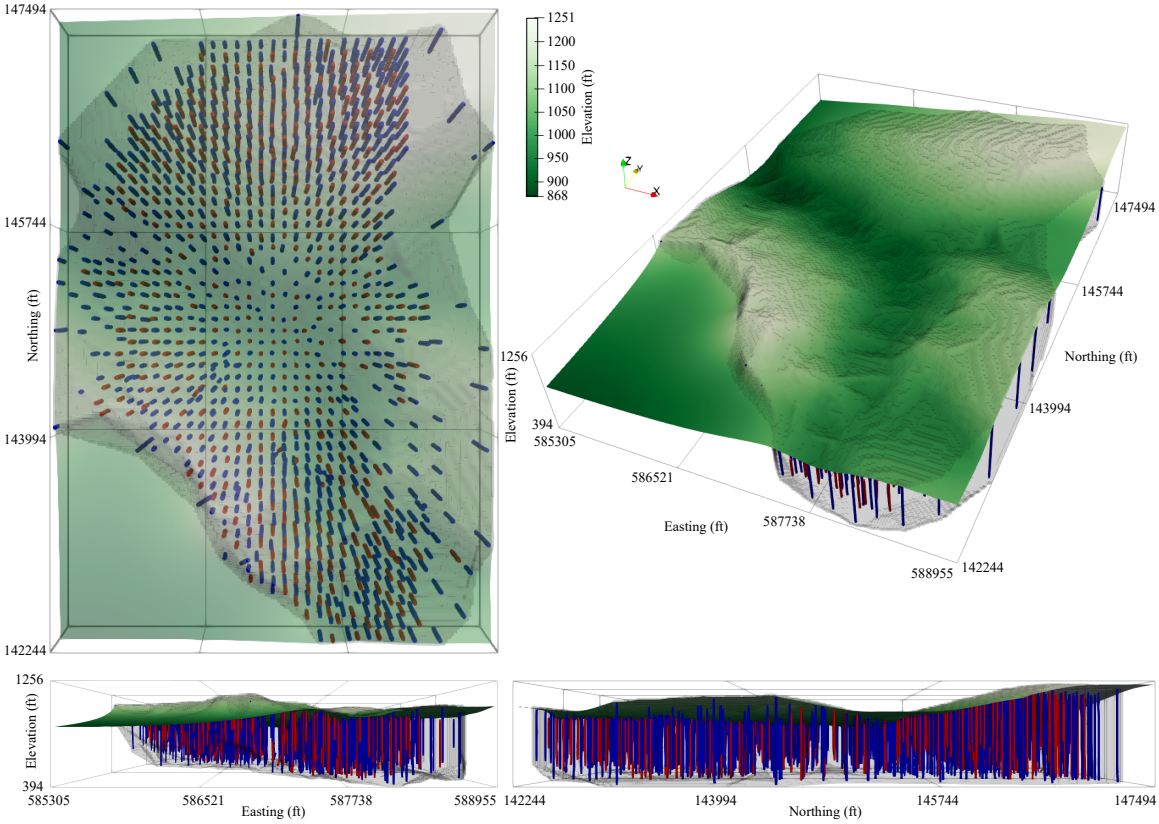


Figure 8.1: Views of the location of the 1134 drillholes available for modeling at Red Dog mine. There are 712 blue drillholes and 422 red drillholes. The blue drillholes are used for modeling and the red ones are used for performance check.

Table 8.1: Parameters of the block model utilized for the modeling of the area where the data is available.

Direction	Origin	Number of Blocks	Block Size (ft)
Northing (ft)	585,317	146	25.0
Easting (ft)	142,256	210	25.0
Elevation (ft)	400	69	12.5

The categorical variables available for modeling are the Plates and Rock Type (RT). The Plates variable has five categories: (1) Upper; (2) Middle; (3) Lower; (4) Sub-lower; and (5) Other. The RT variable also has five categories: (1) Barite; (2) Low Grade (LG); (3) High Grade (HG); (4) Vein; and (5) Host. The data set is divided into two sets of 712 drillholes with 26,074 samples and 422 drillholes with 15,269 samples. The larger set is used as train data and is shown as blue drillholes in Figure 8.1 and the smaller set is used as test data and is shown as red drillholes. The separation of the two sets is performed in such way that the resulting sets have nearly evenly spaced drillholes and the test set appear as infill drilling in relation to the training set. All the modeling in this chapter is undertaken with the training set and the quality of the models and performance of the methodologies are evaluated with the test set.

8.2.1 Recovery Parameters

The mineral composition, texture and rock/ore type are cited as important factors for the metallurgical recovery and throughput in Red Dog Mine (Krolak et al., 2017), however, quantitative factors relating the metallurgical recovery to the categorical variables are not found in the literature. The goal of this chapter is to demonstrate the impact of the developed methodologies for categorical modeling on key response variables such as metallurgical recovery. Leuangthong (2003) utilizes a metallurgical recovery prediction model based on the Zn, Fe, and Ba content for the Red Dog Mine. A metallurgical recovery model based on the categorical variables is built by applying the same rules to the available composites and defining the average recovery for each combination of the categorical variables. The rule utilized in (Leuangthong, 2003) proceeds as follows:

(1) if the sample belongs to the Vein RT, the Zn recovery is fixed at 89%
(2) if Ba is greater or equal than 7%, the recovery is given by: $27.182 \times \ln(\text{Zn}) - 3.4834$; to a maximum of 85%
(3) if Ba is less than 7% and Fe is less than 15.5%, the recovery is given by: $89.4 - 0.7 \times \text{Fe}$
(4) if Ba is less than 7% and Fe is greater or equal than 15.5%, the recovery is given by: $(-0.4205 \times \text{Fe} + 90.196) - (55 - (-0.531 \times \text{Fe} + 60.036)) \times 1.6$

Two additional rules are added for this case study: (1) recovery is 0 for the Host RT as it is outside of mineralization and (2) minimum recovery is set to 0% as the rule above

resulted in negative recoveries for some samples in the data set. The resulting recovery model for the combination of the categorical variables is shown in Table 8.2.

Table 8.2: Recovery as function of the combination of categorical variables RT and Plates. Missing values are combinations that are not observed in the data.

Plates	Rock Type				
	Barite	LG	HG	Vein	Host
Upper	20.24	60.57	-	-	-
Median	18.84	71.02	83.25	89.00	-
Lower	37.32	79.08	84.18	89.00	-
Sub-lower	31.03	76.50	-	89.00	-
Other	-	71.01	79.46	-	0.00

8.2.2 Categorical Proportions

Categorical variables are modeled to define stationary domains for the modeling of continuous variables. The distribution of categories have a direct contribution to resource estimation. Bias in the categorical proportions can potentially lead to under or overestimation of resources with serious economic impact. The definition of global proportion for categories is an important aspect of the modeling. In a multivariate modeling workflow such as the multivariate HTPG developed in this dissertation, the definition of a joint categorical distribution is also required.

The training set utilized for modeling is fairly regularly distributed in space, but there are still some areas that are more or less densely sampled. To account for preferential sampling, declustering weights are calculated for each sample. Cell declustering is a robust technique that only requires the definition of the cell size parameter. Cell size is often defined by largest reasonable drillhole spacing within the area of interest (Rossi and Deutsch, 2014). To assist the selection of cell size the distribution of drill hole spacing is calculated (Figure 8.2) and the 95^{th} percentile is highlighted as reference for the cell size selection. The cell size of 200x200x25 ft is selected for this data set.

There is a clear trend from bottom to top in the RT categories (Figure 8.3). The Vein RT is more likely to appear at lower elevations (Figure 8.3d) whereas the Barite is observed more often in the upper portions (Figure 8.3a). The Plates categorical variable is also highly non-stationary (Figure 8.4) with the Upper and Middle Plates located mostly on the South region (Figures 8.4a and 8.4b) and Lower and Sub-lower located towards the North region

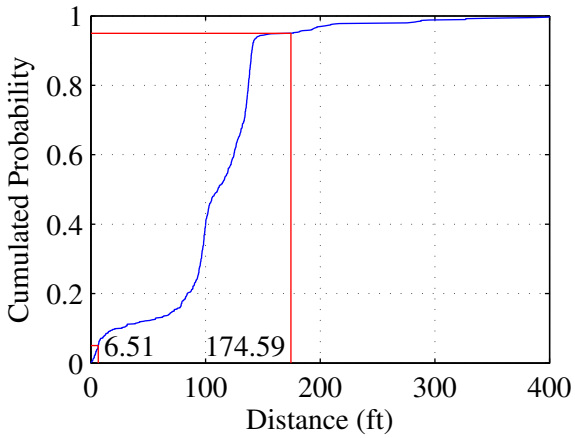
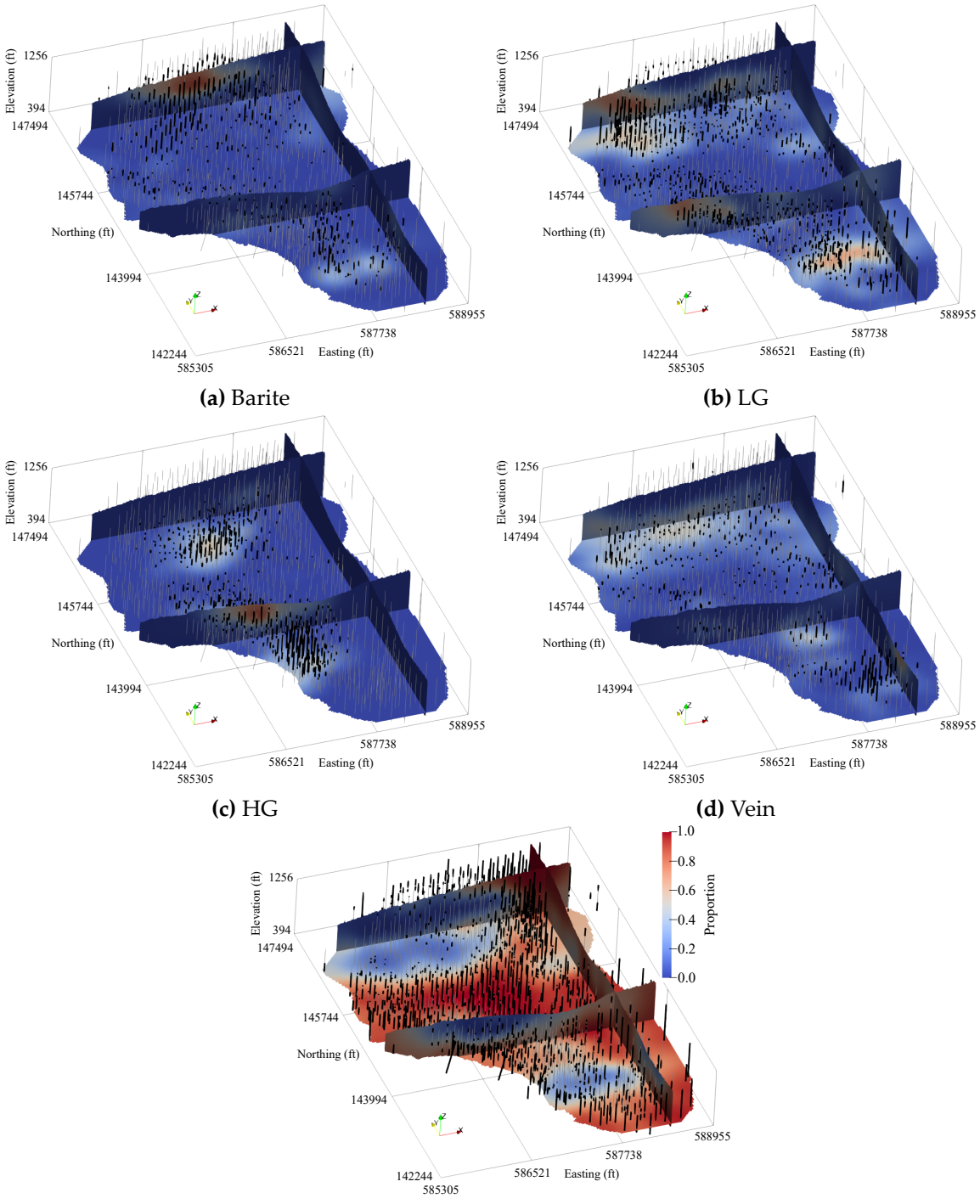


Figure 8.2: Distribution of drillhole spacing for the train data set. The 95^{th} percentile is highlighted to assist the definition of cell size for cell declustering and the 5^{th} percentile is highlighted to assist the definition of the smallest lag sizes for horizontal variogram calculation.

(Figures 8.4c and 8.4d) of the Red Dog area. The Host category on RT variable is similar to the Other category on the Plates variable. The main difference is that the weathered LG and HG is included in the Other category.

The global proportions calculated from the trend models need to be checked against the reference distribution calculated from the data to avoid bias in the simulated models. The categorical global proportions calculated with and without declustering weights as well as the proportions calculated from the trend model are shown in Table 8.3. The declustered proportions do not differ significantly from the proportions calculated without the declustering weights. This is due to the fairly regular drillhole spacing in the area of interest. A small change in the proportion, however, may lead to significant bias on the resource estimation. The models generated with the use of local proportions are more likely to reproduce the distribution inferred from data if the global proportions calculated from the trend model matches the target global proportions. The proportion calculated from the trend model generated for this case study is close to the declustered proportions (Table 8.3).

The joint categorical probability density function (PDF) is calculated for the training data set (Figure 8.5a) and compared with the theoretical distribution for independent categories (Figure 8.5b). There is a clear deviation between the joint PDF of the data and the one for independent variables. A multivariate modeling approach is necessary to enforce the relationship between the two categorical variables observed in the data.

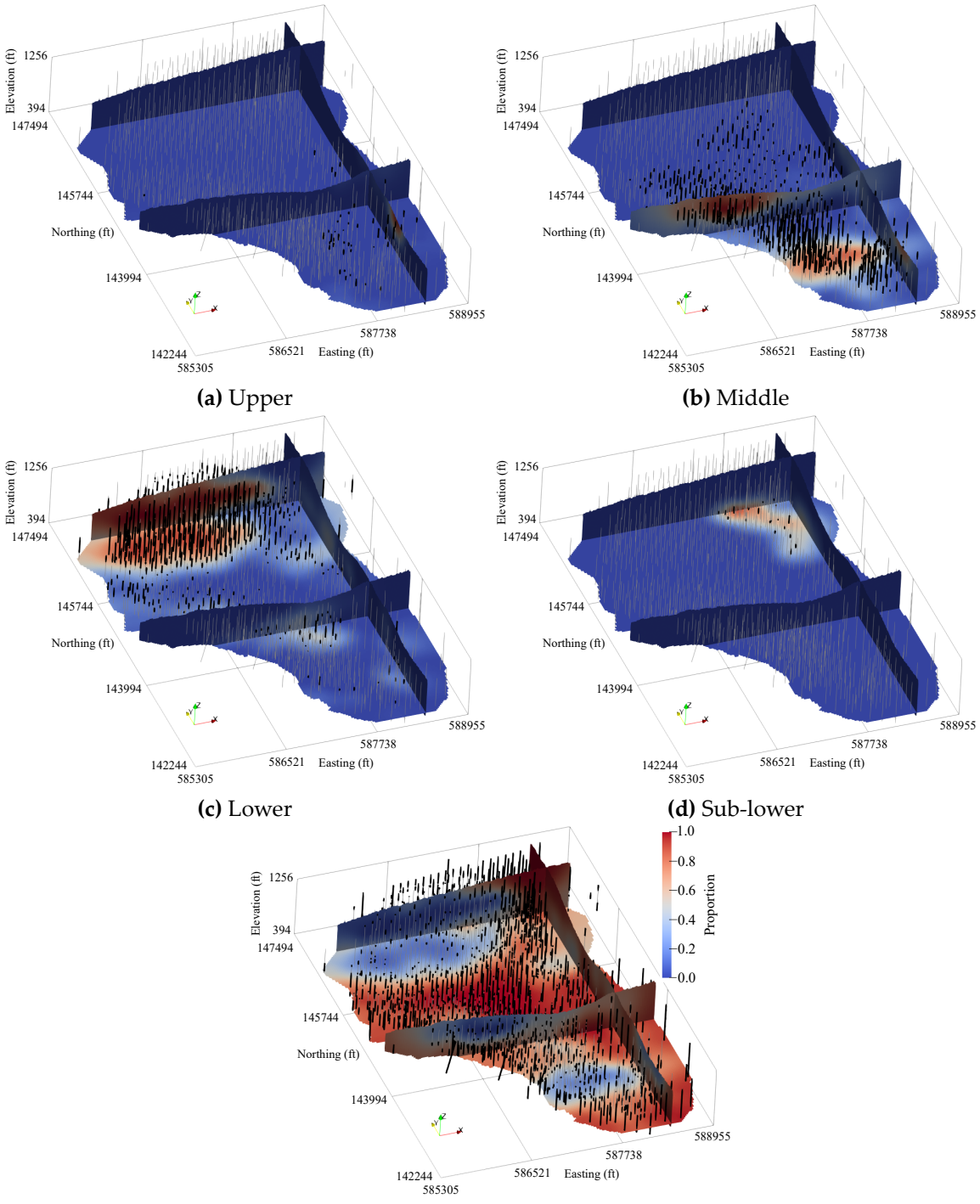


(a) Barite **(b)** LG

(c) HG **(d)** Vein

(e) Host

Figure 8.3: Local proportion for the RT categorical variable.



(a) Upper **(b)** Middle

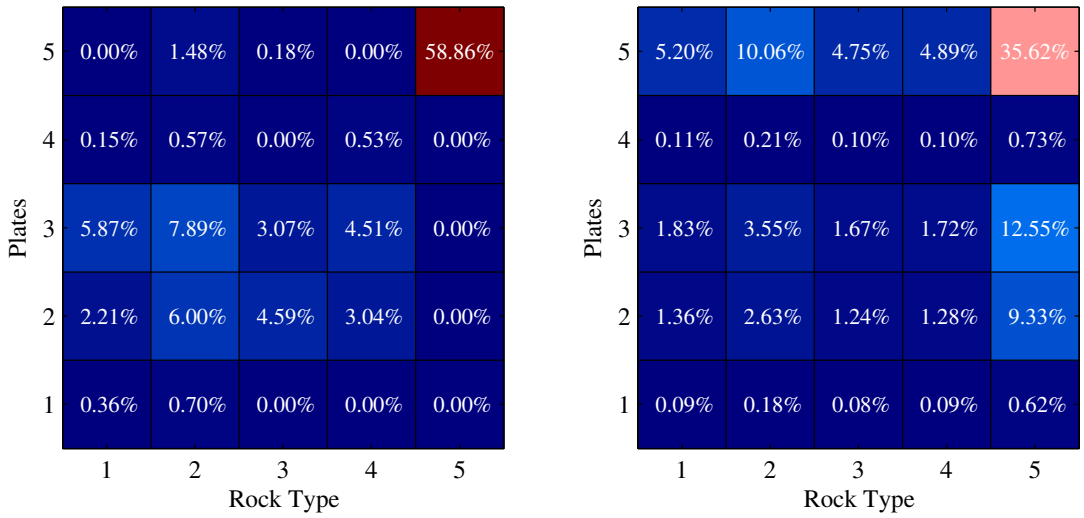
(c) Lower **(d)** Sub-lower

(e) Other

Figure 8.4: Local proportion for the Plates categorical variable.

Table 8.3: Global proportions for Swiss Jura data set

Variable	Category	Code	Global Proportion (%)		
			Clustered	Declustered	Trend Model
Plates	Upper	1	0.95	1.06	0.94
	Median	2	20.00	15.84	15.71
	Lower	3	23.97	21.33	21.78
	Sub-lower	4	1.21	1.25	01.36
	Other	5	53.86	60.52	60.22
Rock Type	Barite	1	9.75	8.59	7.75
	LG	2	18.30	16.63	16.83
	HG	3	10.77	7.85	7.48
	Vein	4	8.86	8.08	8.63
	Host	5	52.32	58.86	59.31

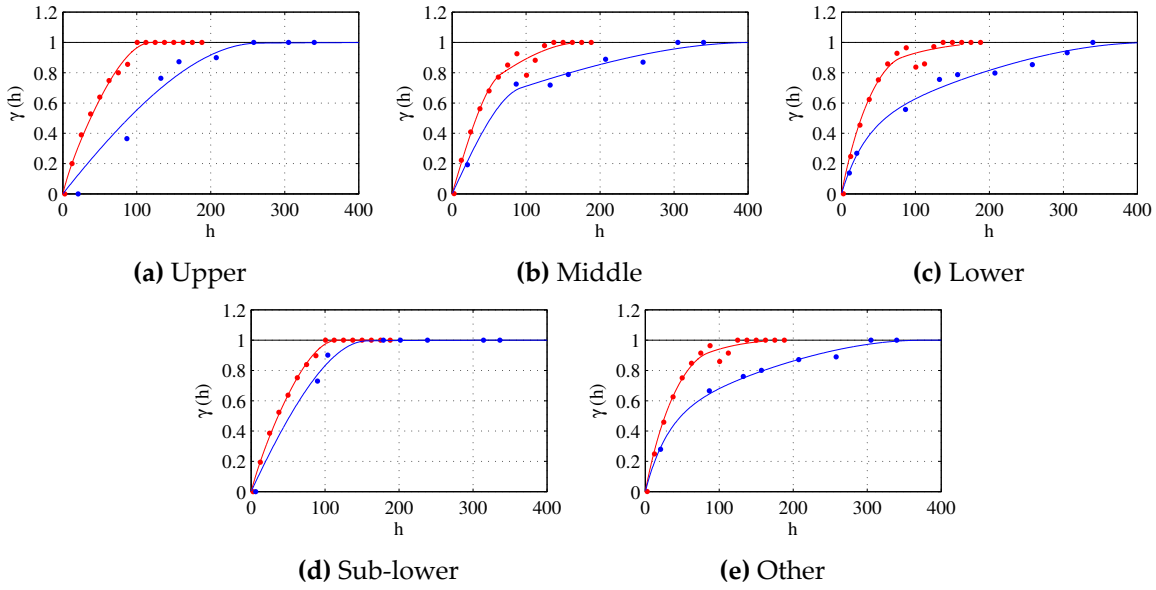


(a) Data PDF **(b)** Independent PDF

Figure 8.5: Multivariate categorical PDF calculated from data and the theoretical joint PDF for independent categories.

8.2.3 Variography

The parameterization of the spatial structure is performed by calculating the indicators for each categorical variable and the respective residuals considering the trend models. The only strong observed anisotropy is from horizontal to vertical for all Plates and RT indicators. The major and mid directions are isotropic and parallel to the horizontal plane whereas the minor direction is aligned with the vertical direction. The experimental variograms of the indicator residuals for the Plates variable are shown in Figure 8.6 and the variograms for the RT categories are shown in Figure 8.7e.



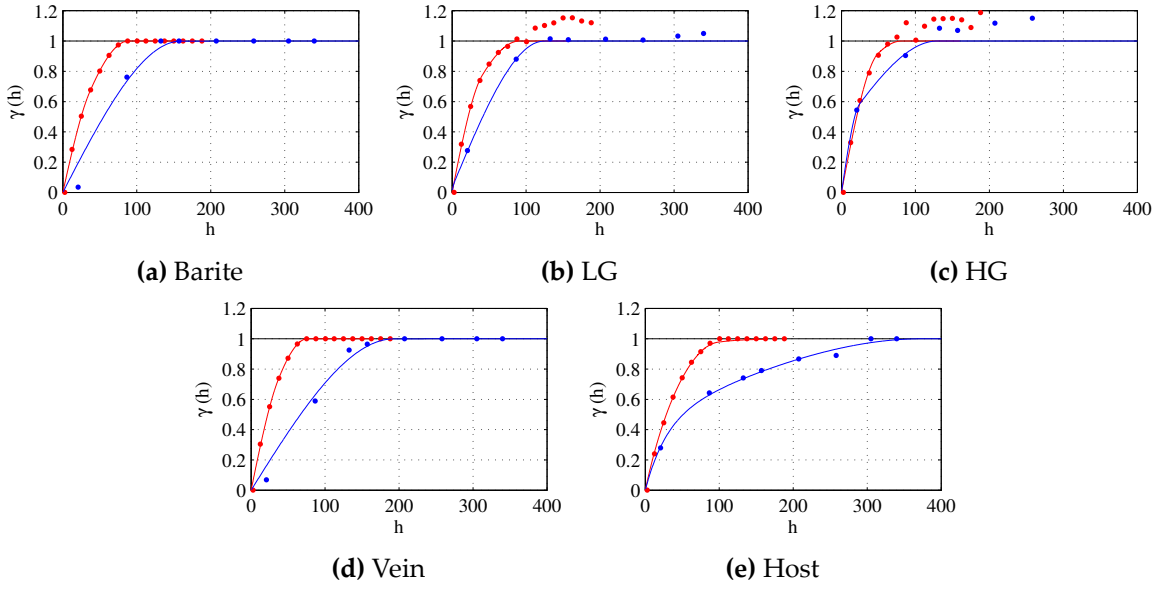
(a) Upper **(b)** Middle **(c)** Lower

(d) Sub-lower **(e)** Other

Figure 8.6: Modeled experimental variograms of the indicator residuals for the Plates variable. The red color is used for the vertical direction and the blue color is used for the horizontal direction. The markers are the experimental points and the lines are the fitted models.

8.3 HTPG Parameters

The parameters for the HTPG are defined in this section. The required parameters are the structure of the truncation tree for each categorical variable, the truncation thresholds and the spatial structure of the latent variables. The multivariate HTPG also requires the definition of the correlation structure between the Gaussian latent variables across all categorical variables.



(a) Barite **(b)** LG **(c)** HG

(d) Vein **(e)** Host

Figure 8.7: Modeled experimental variograms of the indicator residuals for the RT variable. The red color is used for the vertical direction and the blue color is used for the horizontal direction. The markers are the experimental points and the lines are the fitted models.

8.3.1 Truncation Rule

The definition of the mapping between the categorical and the continuous space is the most important step in the application of truncated Gaussian methods. Geological contacts, spatial structure and proportions are the main factors considered for the definition of the truncation rule.

A cross-section in the South-North direction located at 570840 ft easting coordinate is shown in Figure 8.8a together with the transition probability (Figure 8.8b) and multi-dimensional scaling (MDS)/minimum spanning tree (MST) visualization (Figure 8.8c) for the Plates categorical variable. Most of the times the Upper, Middle, Lower and Sub-lower plates are separated. Each one of these categories transitions to the Other category. In few instances, transition between consecutive plates is observed, such as Upper to Middle, Middle to Lower, and Lower to Sub-Lower. The categories in the Plates variable represent large features concentrated in certain areas of the domain. Given the spatial separation of the categories, the trend model can be used to enforce the transitions. The hierarchical truncation tree chosen for the Plates variable is a linked list shown in Figure 8.9. The linked lists have the most number of Gaussian variables possible in a truncation structure. The truncation ordering is based on the spatial continuity with the most continuous variables

at the top of the truncation structure and the least continuous at the bottom.

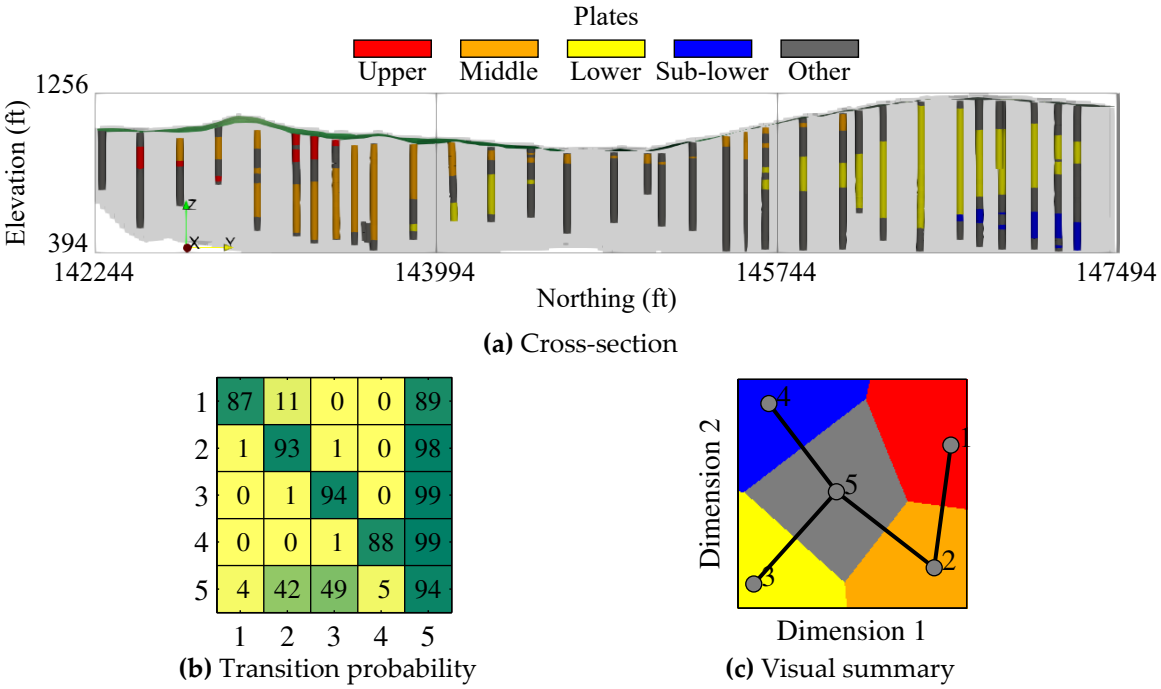


Figure 8.8: Cross-section, dissimilarity matrix and visual summaries based on MDS and MST for the Plates categorical variable.

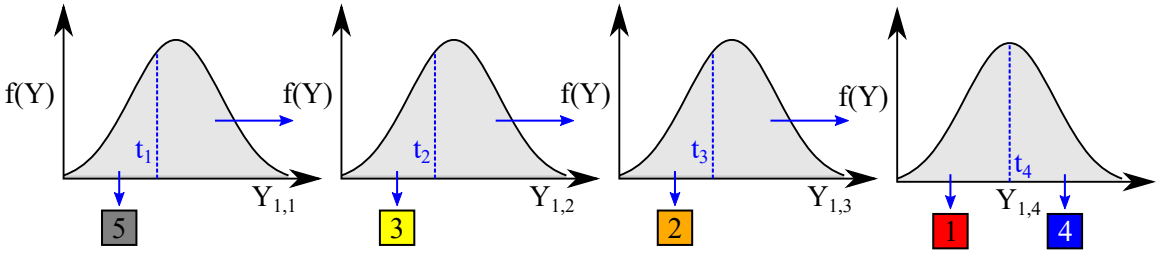


Figure 8.9: Truncation tree for the Plates variable.

The procedure utilized for the Plates variable is repeated for the RT variable. The same cross-section is shown in Figure 8.10a for the RT variable and the transition matrix and MDS/MST visualization are shown in Figures 8.10b and 8.10c. There are transitions between all categories to a certain extension. The cross-section show that the Barite is often at the top followed by LG and HG categories before transitioning to the Vein category. The transition from Barite to Vein is rare and it happens in the absence of the LG and HG categories. The LG and HG appear together most of the time and the MDS/MST visualization show a close connection between the two categories. In light of these observations, the truncation structure shown in Figure 8.11 is chosen for the RT variable.

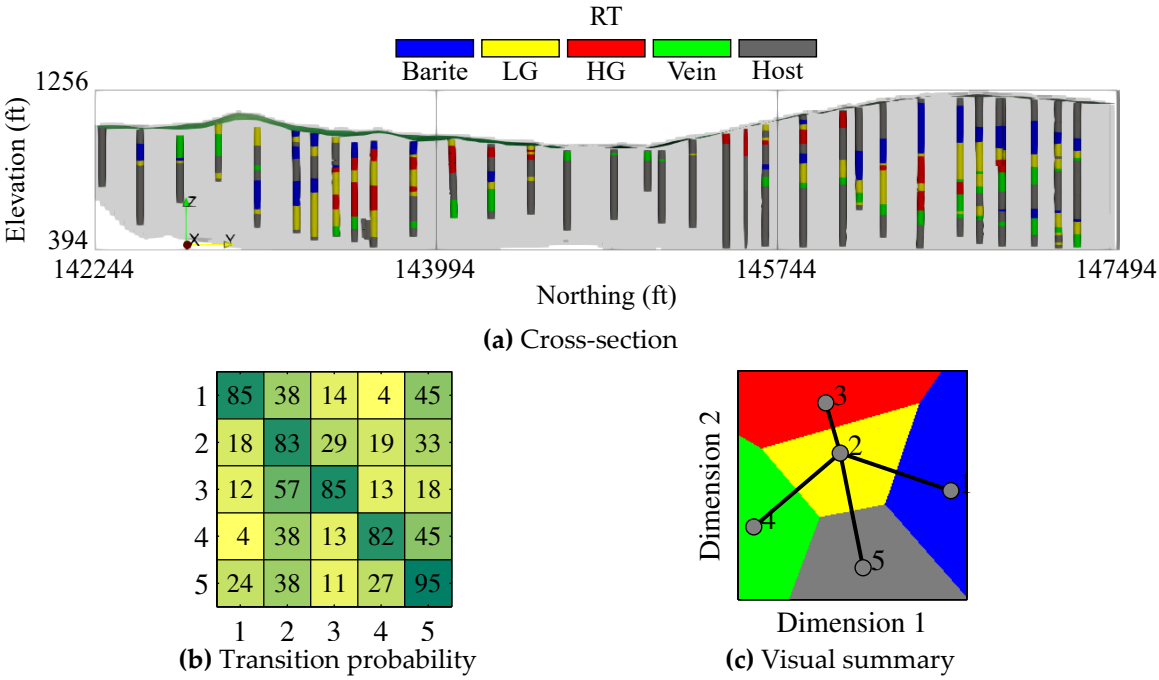


Figure 8.10: Cross-section, dissimilarity matrix and visual summaries based on MDS and MST for the RT categorical variable.

The truncation structure defined for the RT variable consists of three Gaussian variables with the first defining the Host category (the most continuous), the second Gaussian variable is utilized to separate Barite from Vein by placing the LG and HG in between, and the third Gaussian variable defines the separation between LG from the HG. In the stationary case, this configuration would result in no transition between the Barite and Vein categories, however, in the non-stationary case the transition between the Barite and Vein can happen where the local proportion of LG and HG are low.

8.3.2 Thresholds

Once the the truncation rules are defined, the thresholds are calculated based on the categorical proportions. A global set of thresholds must be defined using the global proportions. These are used in the numerical derivation of the variograms of the latent variables. The global thresholds for both Plates and RT variables are shown in Table 8.4.

Local proportions are being considered for the modeling of the categorical variables. In this case, the matching set of local thresholds must be defined. The local thresholds for the Plates variable are shown in Figure 8.12 and the local thresholds calculated for the RT variable are shown in Figure 8.13.

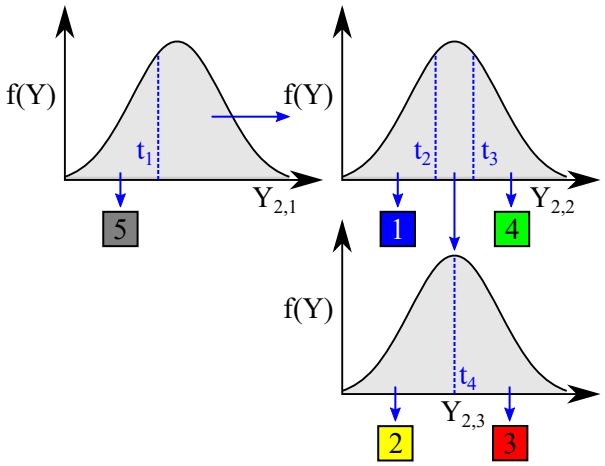


Figure 8.11: Truncation tree for the RT categorical variable.

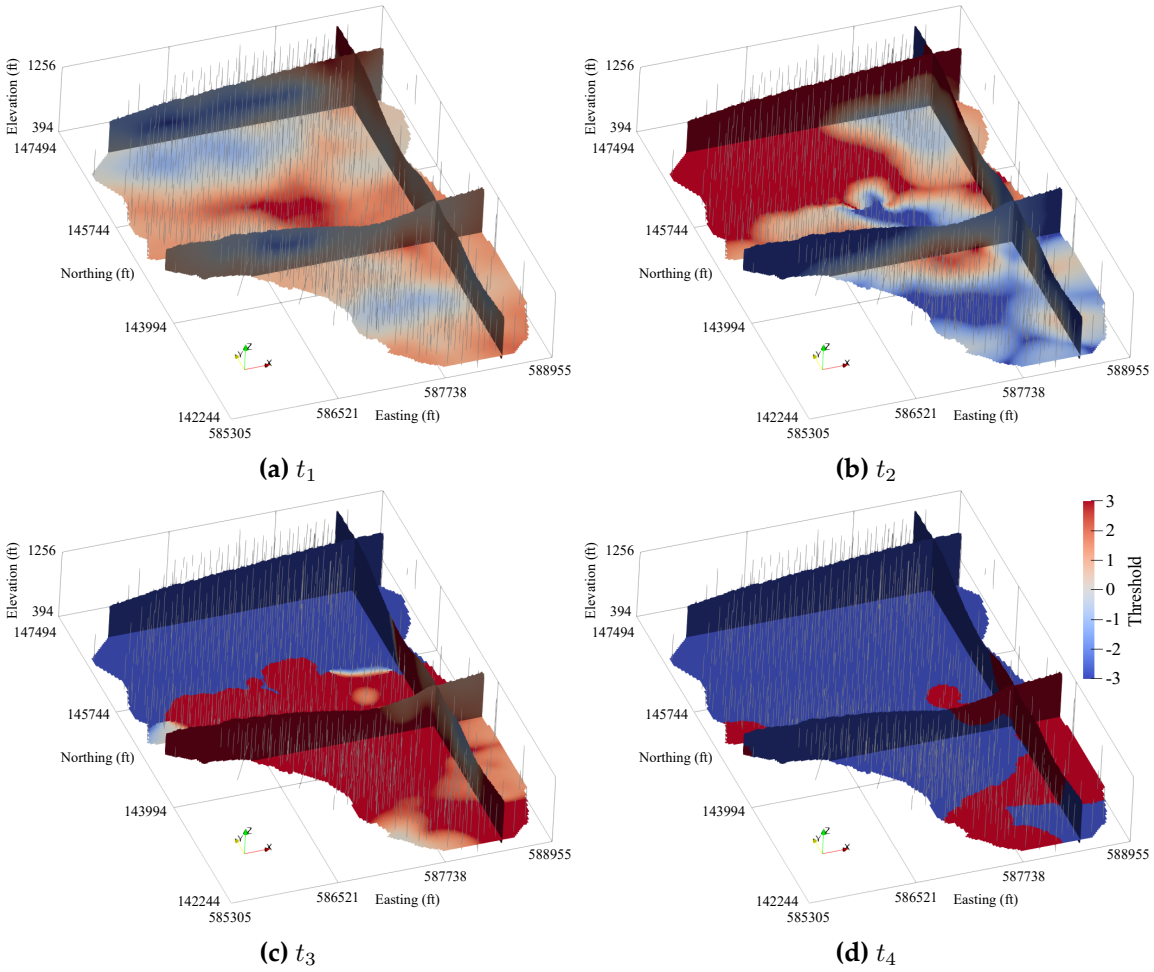
Table 8.4: Global thresholds for the Red Dog case study.

Variable	Global Thresholds			
	t_1	t_2	t_3	t_4
Plates	0.2589	0.1193	1.1395	-0.2313
RT	0.2354	-0.8764	0.7992	0.5022

8.3.3 Numerical Variogram Derivation

With the truncation rule established and the global thresholds calculated, the numerical derivation of the Gaussian variables variograms is undertaken utilizing the variogram of the indicator residuals as reference. The numerical derivation is performed for the horizontal and vertical directions of continuity utilizing 100,000 samples for the Monte-Carlo simulation (MCS). The horizontal direction is discretized into 20 steps of 20 ft and the vertical direction is discretized into 20 steps of 10 ft.

The results of the numerical derivation for the Gaussian variables utilized for the Plates variable are shown in Figure 8.14. The variograms in Figures 8.14a to 8.14d are the resulting variograms for the Gaussian variables. The variograms for Gaussian $Y_{1,1}$ and $Y_{1,2}$ are well behaved whereas the variograms for $Y_{1,3}$ and $Y_{1,4}$ show hyper-continuous structure. The hyper-continuity in this case is difficult to avoid given the very different categorical proportions for this variable. The Upper and Sub-lower categories show small proportions compared with the other variables. Even in the presence of unfitted hyper-continuous structure, the reproduction of categorical variograms may be just slightly affected. This can only be determined when checking the simulated models.



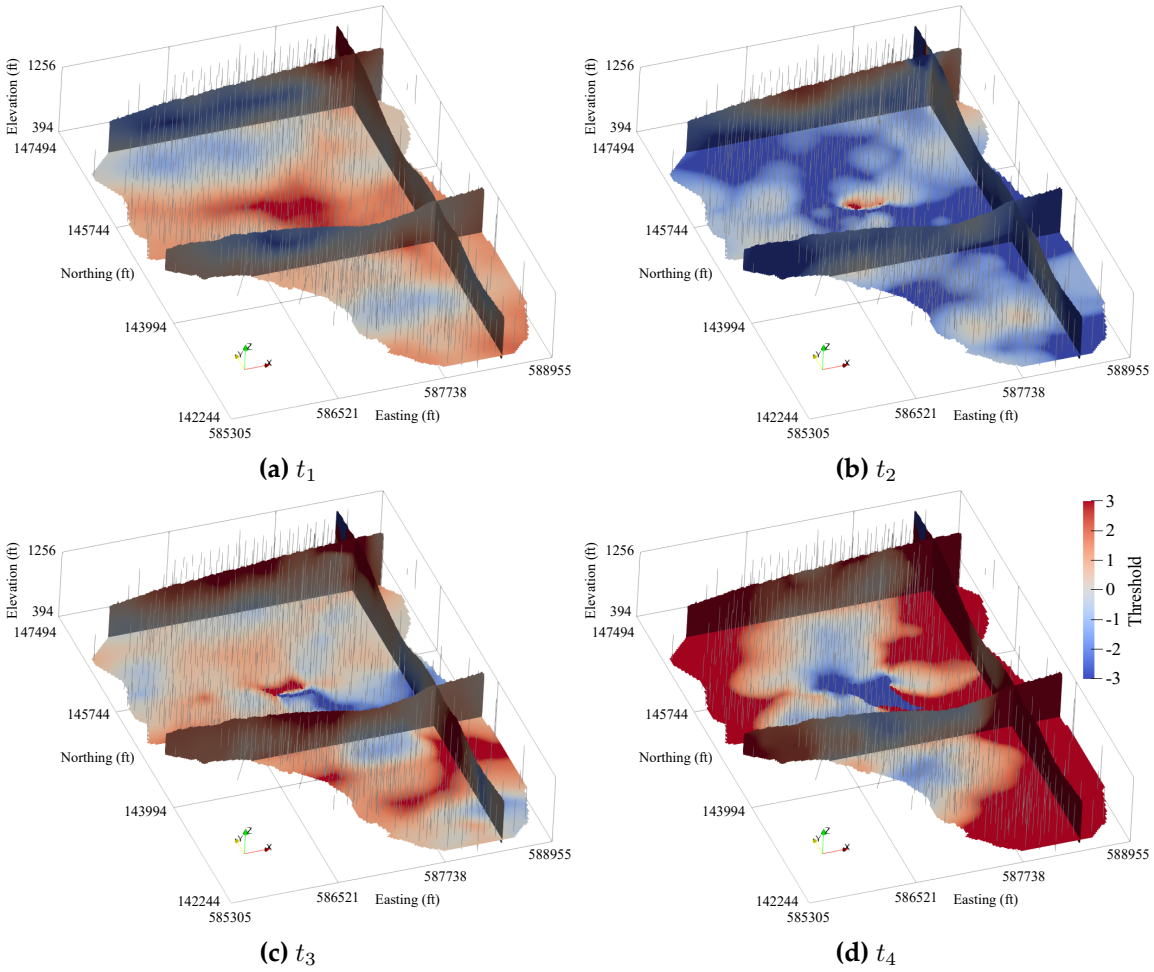
(a) t_1 **(b)** t_2

(c) t_3 **(d)** t_4

Figure 8.12: Local thresholds calculated for the Plates variable.

The Upper and Sub-lower categories share the last node of the truncation tree and their variograms are not well matched in the horizontal direction (Figures 8.14e and 8.14h). This is not a major concern as these categories did not have stable horizontal experimental variograms. The well informed vertical direction is well matched. The other categories were matched by the numerical derivation. Again, the unfitted hyper-continuous structures may affect final variogram reproduction.

The results of the numerical variogram derivation for the RT variable are shown in Figure 8.15. The variogram of the Gaussian variables are better behaved with a slight presence of hyper-continuity observed for $Y_{2,2}$ (Figure 8.15b). The variogram models are well matched for all categories in RT variable (Figures 8.15d to 8.15h).



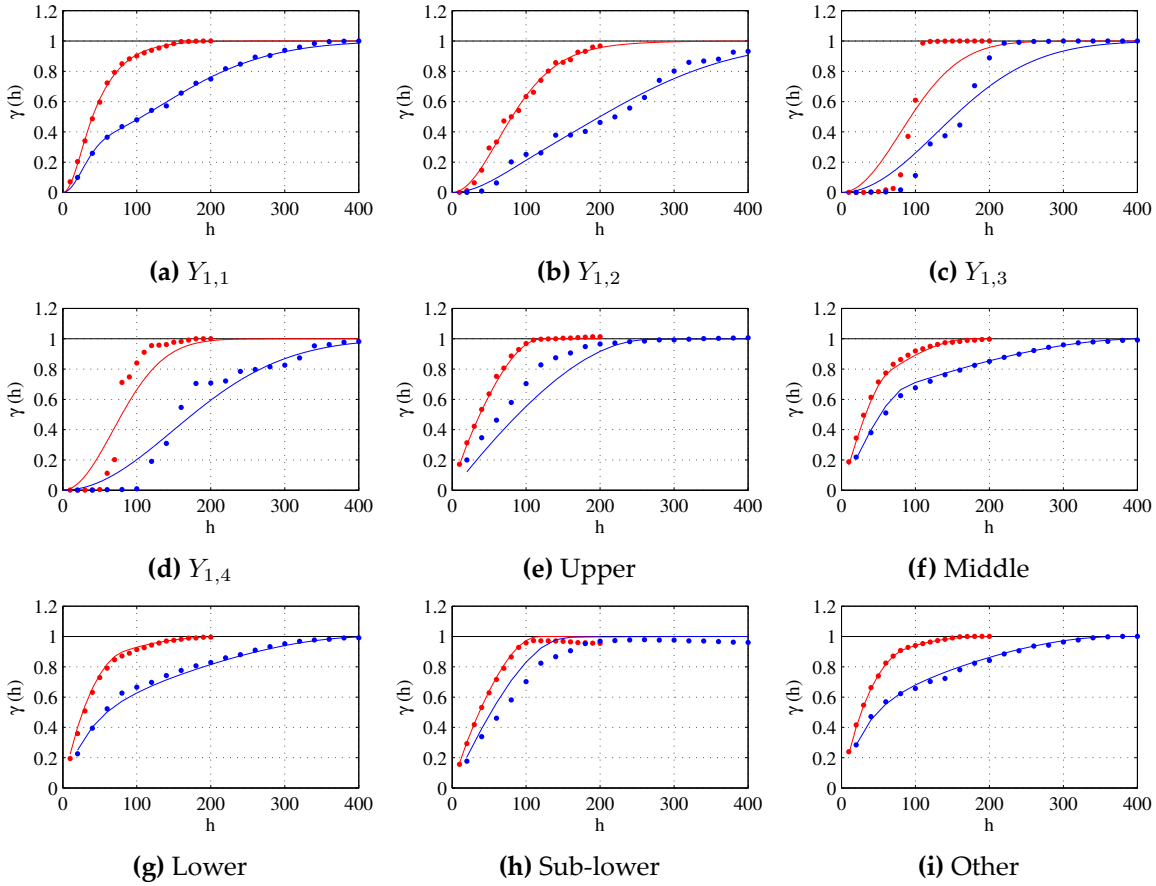
(a) t_1 **(b)** t_2

(c) t_3 **(d)** t_4

Figure 8.13: Local thresholds calculated for the RT variable.

8.3.4 Correlation for Multivariate HTPG

The algorithm for the optimization of the correlation matrix is run to match the declustered categorical multivariate distribution (Figure 8.5a). The software is allowed to run for 5,000 iterations with 20,000 samples and 120 random restarts. The best solution found has a sum of squared error (RSSE) of 2.18%. The correlation matrix resulting from the optimization is shown in Figure 8.16a and the respective expected joint PDF is shown in Figure 8.16b. The obtained distribution is close to the reference distribution compared to the distribution considering independent variables (Figure 8.5b). The Gaussian variable $Y_{1,1}$ defines the the Other category for the Plates variable and the $Y_{2,1}$ defines the Host category for the RT variable. These categories are very similar and overlap each other in space for the most part. As a result the optimized correlation between these two variables is 0.9985, which rounds to 1.00 in Figure 8.16a. The high correlation can be a limiting factor in the application of



(a) $Y_{1,1}$ **(b)** $Y_{1,2}$ **(c)** $Y_{1,3}$

(d) $Y_{1,4}$ **(e)** Upper **(f)** Middle

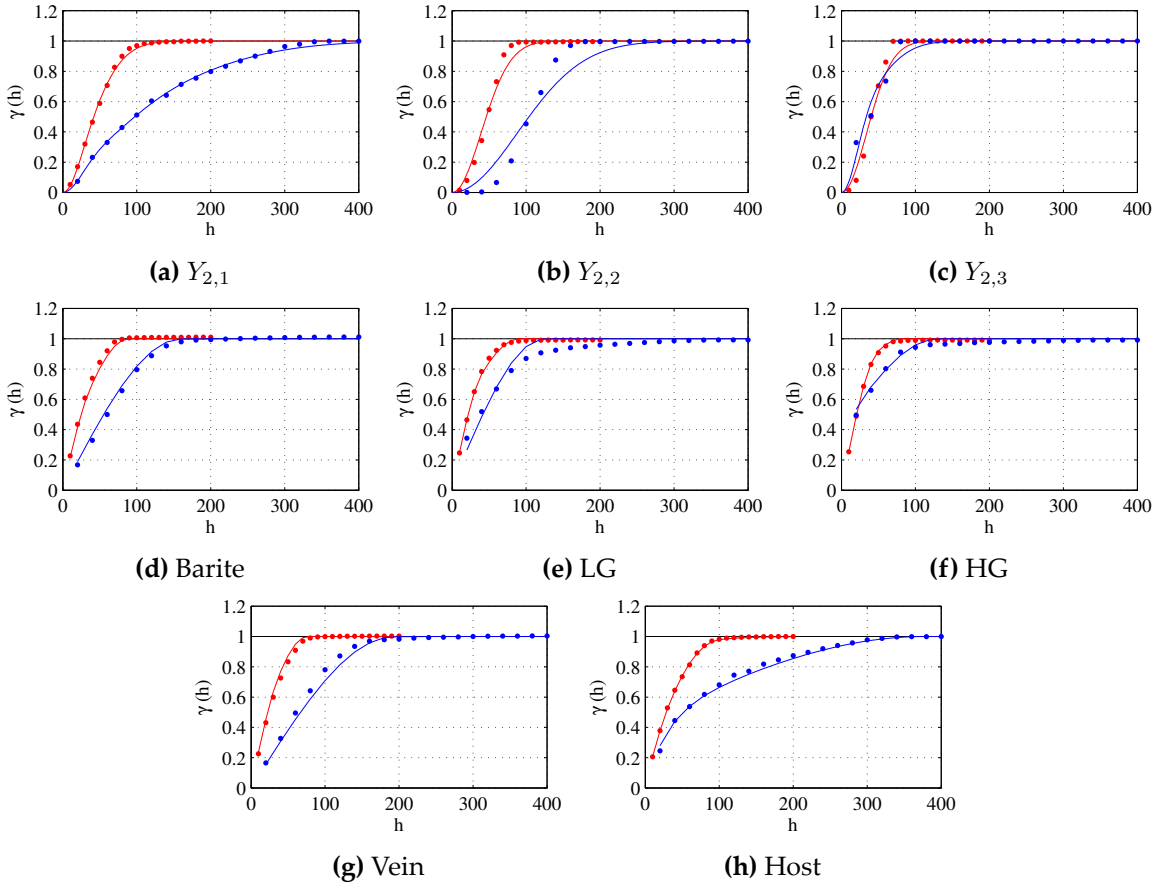
(g) Lower **(h)** Sub-lower **(i)** Other

Figure 8.14: Results from the numerical derivation of the Gaussian variables variograms for the Plates categorical variable. The red color is used for the vertical direction and the blue color is used for the horizontal direction. The markers are the results of numerical derivation and the lines are the reference models (categories) and fitted models (Gaussian variables).

the multivariate Gibbs sampler as highly correlated variables leads to slow convergence of the algorithm.

8.3.5 Imputation of Gaussian Variables

Two sets of imputed Gaussian variables are generated for the Red Dog case study. One is generated with independent Gaussian latent variables and the other is generated considering the optimized correlation matrix shown in Figure 8.16a. The independent data imputation is utilized with the univariate case of the HTPG and the correlated is utilized for the multivariate HTPG.

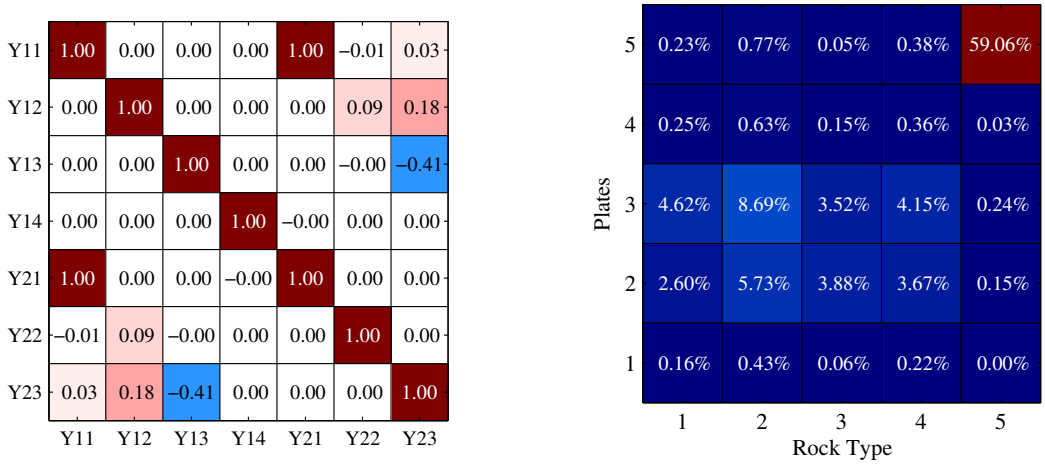


(a) $Y_{2,1}$ **(b)** $Y_{2,2}$ **(c)** $Y_{2,3}$

(d) Barite **(e)** LG **(f)** HG

(g) Vein **(h)** Host

Figure 8.15: Results from the numerical derivation of the Gaussian variables variograms for the RT categorical variable. The red color is used for the vertical direction and the blue color is used for the horizontal direction. The markers are the results of numerical derivation and the lines are the reference models (categories) and fitted models (Gaussian variables).



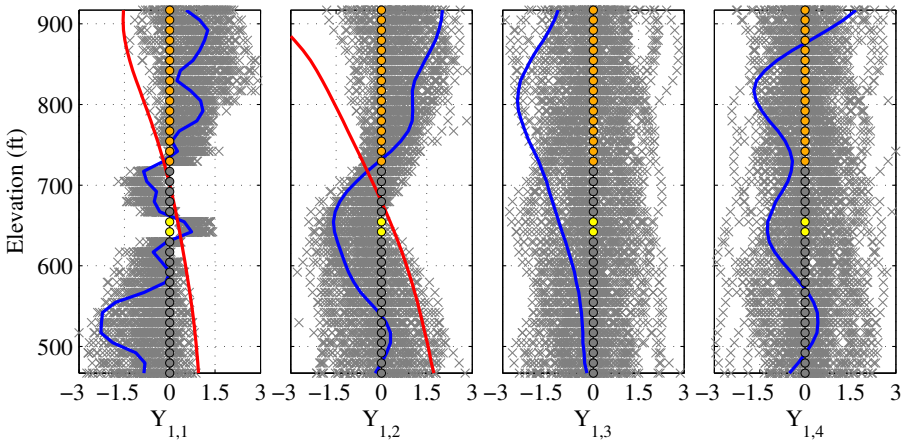
(a) Correlation **(b)** Optimized Distribution

Figure 8.16: Results from the correlation matrix optimization for the train data set.

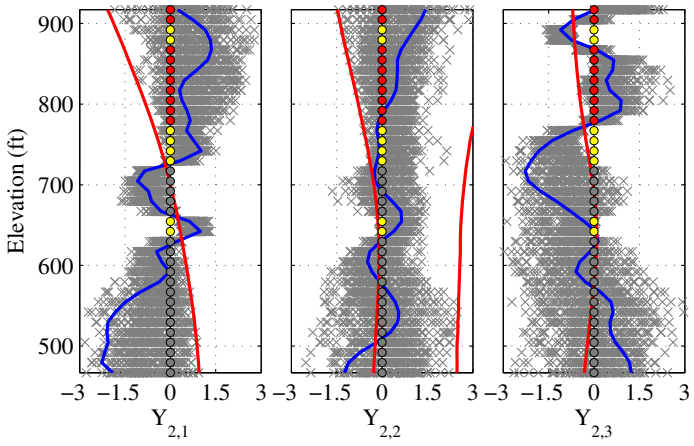
8.3.5.1 Independent Gibbs Sampler

The combined Gibbs sampler approach proposed in Section 5.2.3 is utilized for the independent data imputation in this case study. At each iteration, the Gibbs sampler algorithm loops through all data locations. As a result, the more data points in the data set the less burn-in iterations are required. Only eight burn-in iterations were utilized for the data imputation with the training data set due to the large number of composites available.

The imputation results for drillhole 1267 are shown in Figure 8.17. The blue lines highlight one data realization, the red lines represent the thresholds applied to each Gaussian variable and the gray markers show all 100 realizations for each data point.



(a) Plates



(b) RT

Figure 8.17: Results from independent data imputation utilizing the combined Gibbs sampler algorithm for drillhole 1267. The marker color coding in (a) is: orange for Middle; gray for Other and yellow for Lower Plates. The marker color coding in (b) is: red for HG; yellow for LG and gray for Host RT. The red lines represent the thresholds, the blue lines highlight a single data realization and the gray markers are all 100 data realizations in each data location.

The Middle, Lower and Other categories are observed in this drillhole for the Plates variable (Figure 8.17a). The Middle plate (orange markers) must be above the threshold for the first two Gaussian variables ($Y_{1,1}$ and $Y_{1,2}$) and below the threshold for the third ($Y_{1,3}$). The threshold for $Y_{1,3}$ is not visible in Figure 8.17a as it is above the shown interval. The variable $Y_{1,4}$ is irrelevant for the Middle plate (see Figure 8.9). The Lower plate (yellow markers) must be above the threshold for $Y_{1,1}$ and below the threshold for $Y_{1,2}$. The Other category (gray markers) must be below the threshold for $Y_{1,1}$ and all other Gaussian variables are irrelevant for this category.

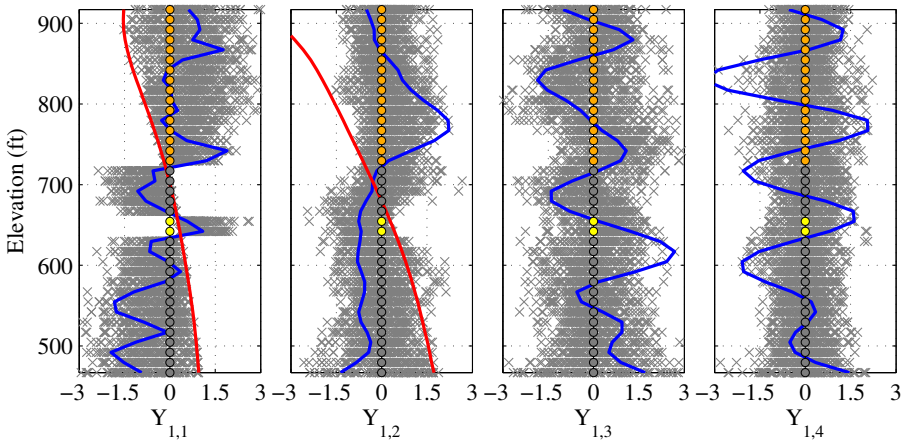
There are three main points to observe when visually checking the results from the Gibbs sampler: (1) check if the simulated values matches the truncation rule; (2) check for extreme highs and lows that may indicate instability in the algorithm; and (3) check for reasonable fluctuations where conditioning is not strong (e.g. variable $Y_{1,4}$ in Figure 8.17a), this may indicate whether or not more iterations are required.

The simulated Gaussian variables for the RT variable are shown in Figure 8.17b. The LG and HG categories (yellow and red markers) must be above the threshold for the first variable ($Y_{2,1}$), between the lower and upper thresholds for the second variable $Y_{2,2}$ and the third variable ($Y_{2,3}$) is the one that separates the two with HG above the threshold and LG below. This is observed in Figure 8.17b. Similar results of the independent Gibbs sampler are observed for the other drillholes in the training data set and are deemed adequate for use with the univariate HTPG workflow.

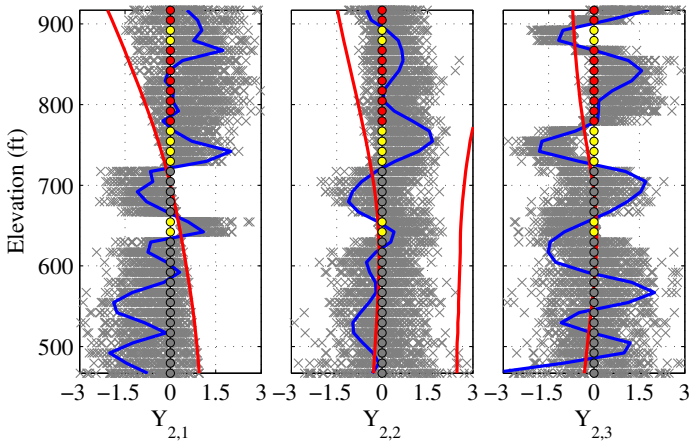
8.3.5.2 Multivariate Gibbs Sampler

The algorithm for the multivariate Gibbs sampler developed in Section 7.3.4 is applied for the simulation of the missing latent variables for the application of the multivariate HTPG. The results for drillhole 1267 are shown in Figure 8.18. The correlation between variable $Y_{1,1}$ and $Y_{2,1}$ is almost 1.0. This high correlation between the two variables makes the convergence of the Gibbs sampler algorithm slow and it is impractical to run the algorithm long enough for convergence. The values observed in Figure 8.18 for these two variables are almost identical to the initial values assigned to data locations and do not show the expected smoothness of the Gaussian variogram structure as observed in Figure 8.17. Also, the variables $Y_{1,3}$ and $Y_{2,3}$ have a correlation of -0.41 and different spatial structures (Figures 8.14c and 8.15c). The intrinsic model of coregionalization (IMC) approximation utilized with the

multivariate Gibbs sampler causes the mixing of the spatial structure when the true structure is not intrinsic. Despite the poor convergence observed for the variables $Y_{1,1}$ and $Y_{2,1}$ and the mixing of spatial structure from IMC approximation, the imputation matches the data and show reasonable fluctuation. The multiple stochastic imputation will transfer the uncertainty on unobserved latent variables to the final models unlike the alternative fixed class center approach.



(a) Plates

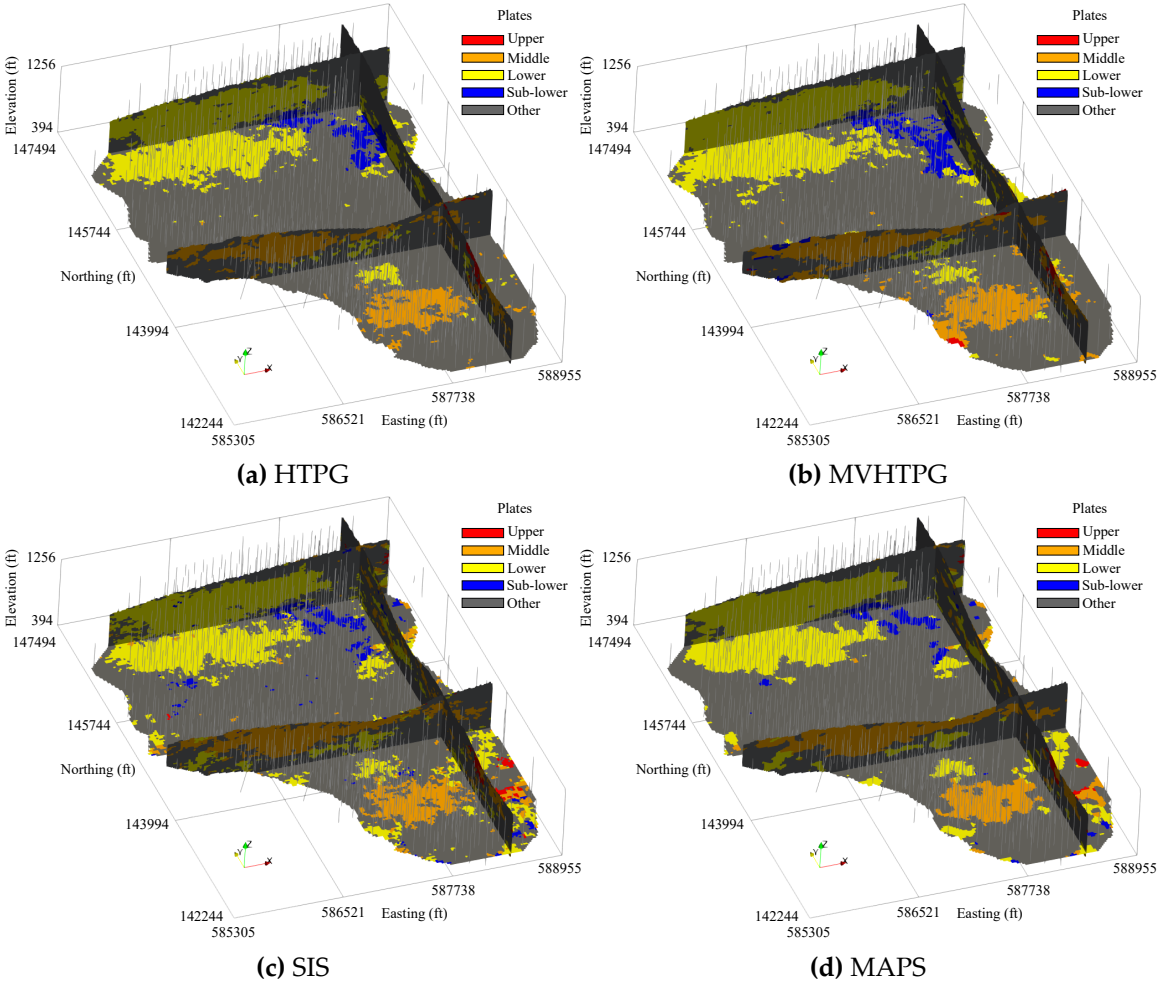


(b) RT

Figure 8.18: Results from correlated data imputation utilizing the standard Gibbs sampler algorithm with IMC for drillhole 1267. The marker color coding in (a) is: orange for Middle; gray for Other and yellow for Lower Plates. The marker color coding in (b) is: red for HG; yellow for LG and gray for Host RT. The red lines represent the thresholds, the blue lines highlight a single data realization and the gray markers are all 100 data realizations in each data location.

8.4 Results

Four sets of 100 realizations of each categorical variable are generated for this case study. The first set is generated using the univariate HTPG, the second set is generated with the multivariate HTPG and will be referred to as MVHTPG, the third set is generated with the SIS approach and the fourth set is generated with SIS and cleaned with maximum a posteriori selection (MAPS) post processing. This is a common practice with the application of SIS and it is used to remove the noise artifacts from SIS and improve the reproduction of categorical proportions (Deutsch, 1998, 2006). The fourth set will be referred to as MAPS. One realization of each set is shown in Figure 8.19.



(a) HTPG **(b)** MVHTPG

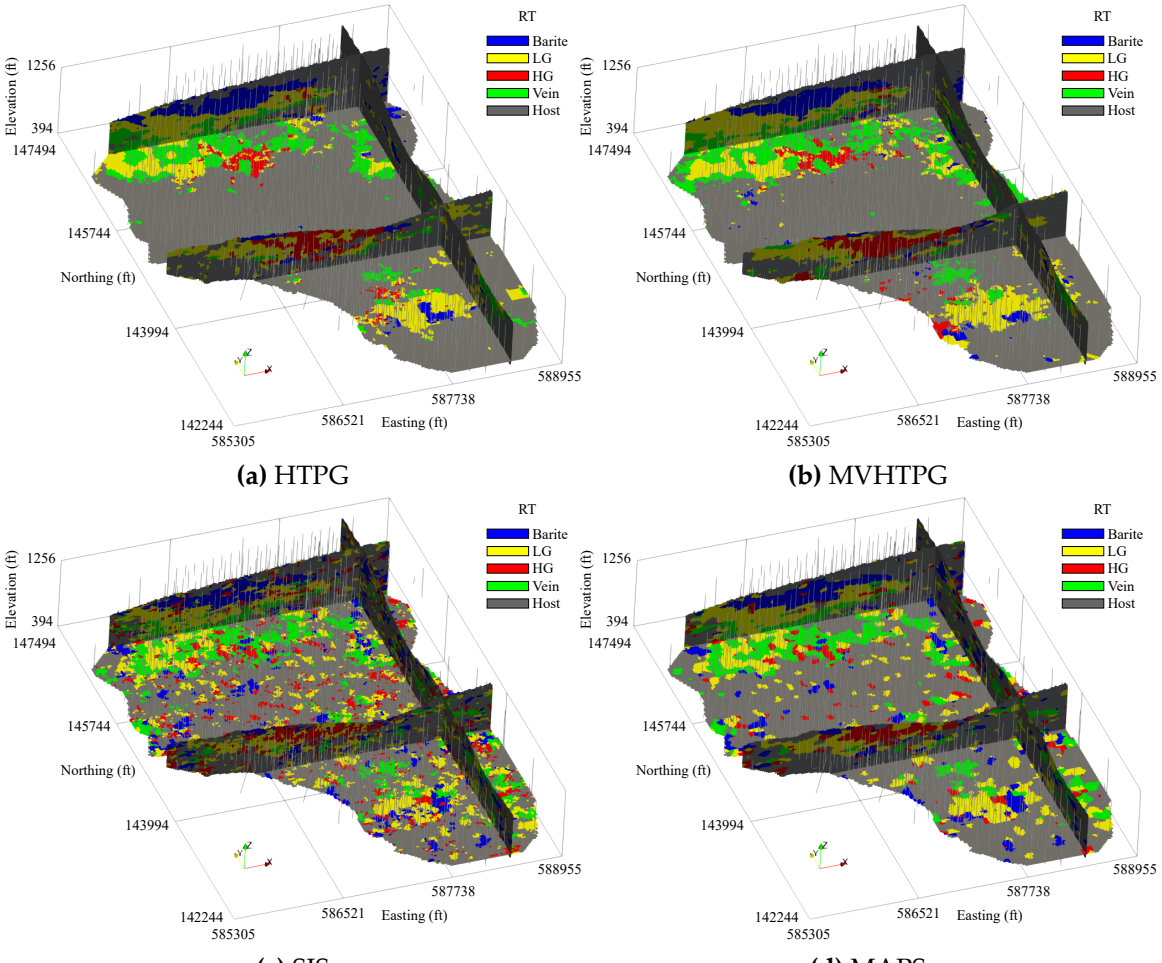
(c) SIS **(d)** MAPS

Figure 8.19: One realization of the Plates variable generated with HTPG, MVHTPG, SIS and MAPS.

The Plates indicators have large structures and the realizations are similar for the HTPG and MVHTPG approaches (Figures 8.19a and 8.19b). The realization for SIS (Figure 8.19c) is

noisiest of all realizations shown and the MAPS realization (Figure 8.19d) is the smoothest. The correctness of the spatial variability is evaluated by checking the variogram reproduction for each methodology.

One realization of the RT variable generated with each technique is shown in Figure 8.20. Again, both HTPG and MVHTPG (Figures 8.20a and 8.20b) seem very similar in terms of spatial variability. Also for the RT variable, SIS generated the noisiest realization and MAPS generated the smoothest. There is a visible difference between SIS and HTPG based models in terms of spatial distribution of the categories.



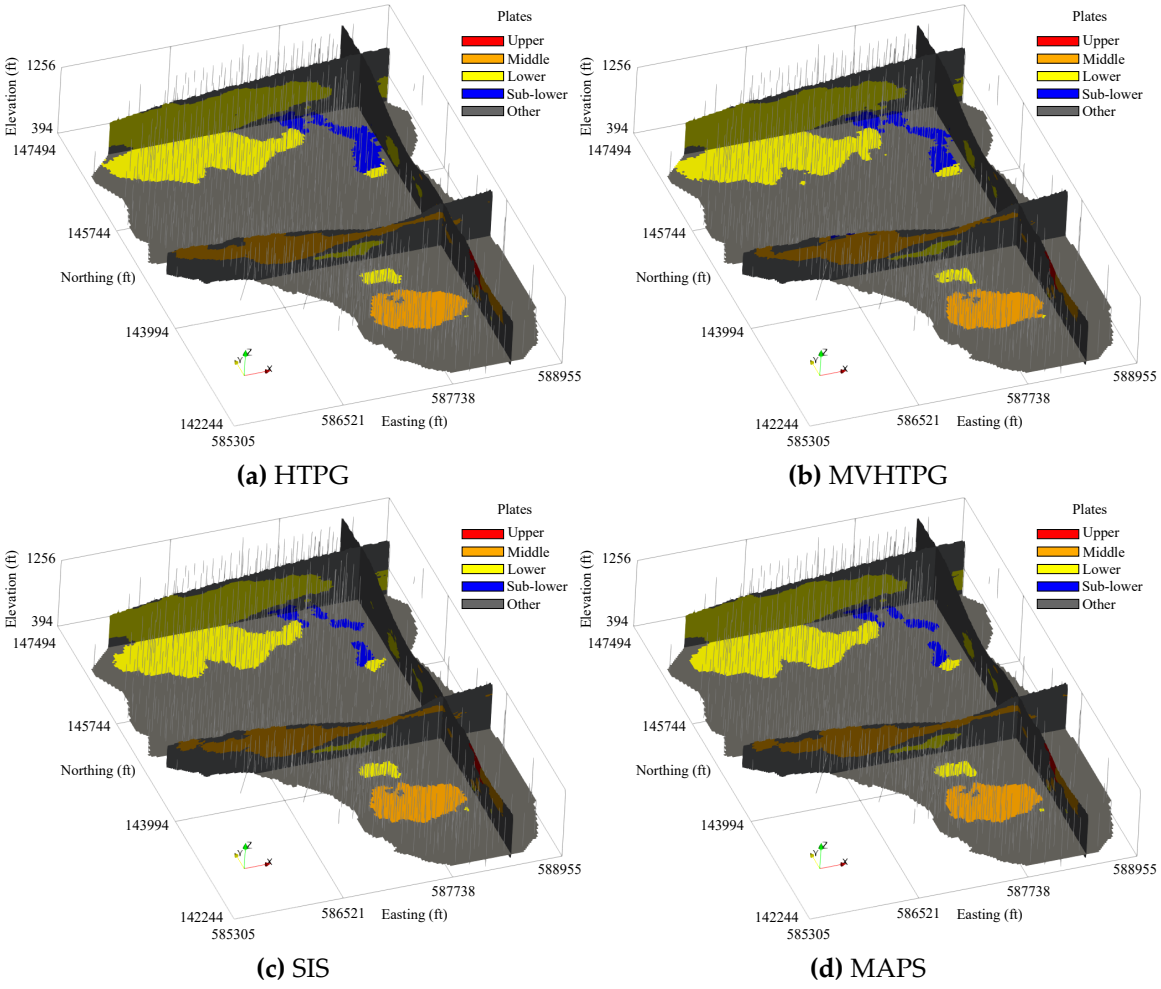
(a) HTPG **(b)** MVHTPG

(c) SIS **(d)** MAPS

Figure 8.20: One realization of the RT variable generated with HTPG, MVHTPG, SIS and MAPS.

The most likely categories over all 100 realizations generated with each technique are shown in Figure 8.21 for the Plates variable. The models for all techniques are similar. This is not a surprise as they use the same trend model. The only noticeable difference is for the the Sub-lower category that shows much higher proportions in the HTPG and MVHTPG

models (Figures 8.21a and 8.21b) when compared to the models generated with SIS and MAPS (Figures 8.21c and 8.21d).



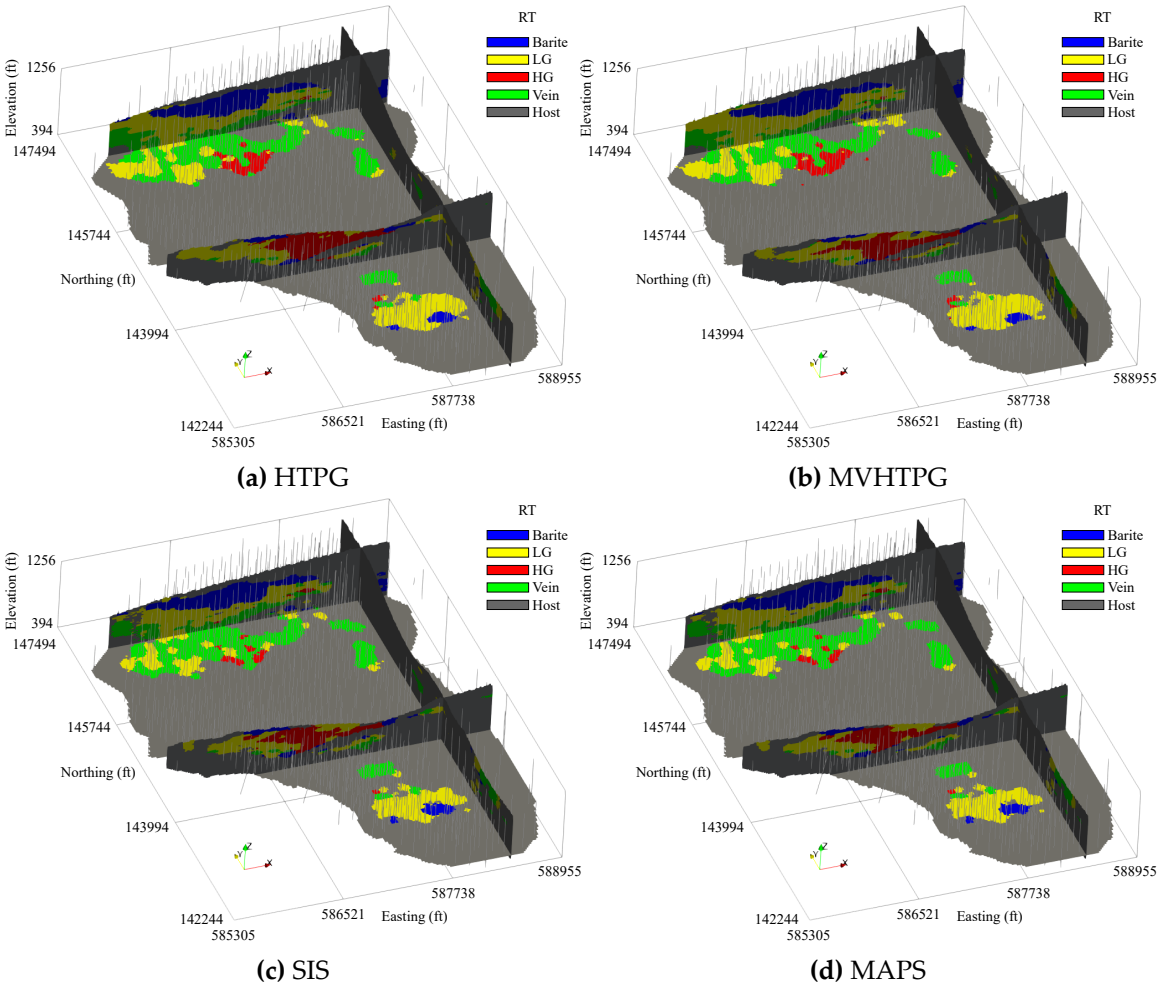
(a) HTPG **(b)** MVHTPG

(c) SIS **(d)** MAPS

Figure 8.21: The most likely category across all realizations for the Plates variable generated with HTPG, MVHTPG, SIS and MAPS.

The most likely categories over all 100 realizations generated with each technique are shown in Figure 8.22 for the RT variable. Again, the models for all techniques are similar. The west-east middle cross-section in Figure 8.22b for MVHTPG show different transition structure between the Barite to LG to HG categories that is not observed in the other models.

The visualization of single realizations offers only a hint of the heterogeneity of the models. Shannon's entropy is used to summarize the uncertainty. The entropy is a direct measure of categorical disorder and it is highest where the uncertainty is highest and lowest where the uncertainty is lowest. The theoretical maximum entropy value is observed when all categories have equal probability. Both Plates and RT variables have five categories and



(a) HTPG **(b)** MVHTPG

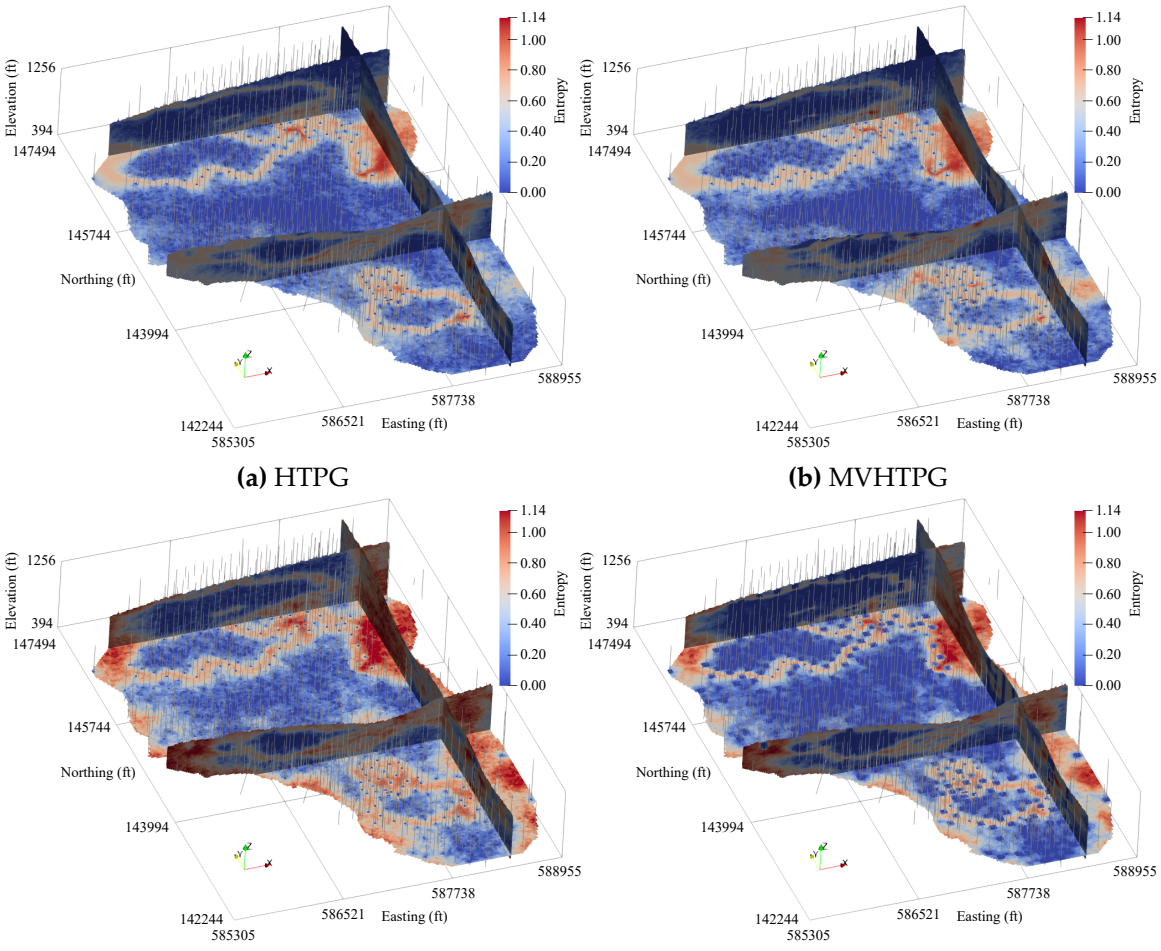
(c) SIS **(d)** MAPS

Figure 8.22: The most likely category across all realizations for the RT variable generated with HTPG, MVHTPG, SIS and MAPS.

the maximum possible entropy for both variables is 1.6. The Shannon entropy is calculated for each grid node based probability of each category calculated over all realizations.

The entropy models for the Plates variable resulting from the application of each modeling technique are shown in Figure 8.23. The entropy is not expected to be different for the Plates variable between the MVHTPG (Figure 8.23b) and HTPG (Figure 8.23a), however, the entropy for the MVHTPG is slightly lower inside the large features away from the categorical boundaries and slightly higher close to the boundaries. This could be a symptom of the less than ideal convergence of the Gibbs sampler algorithm, but, the differences are minor compared to the entropy for SIS (Figure 8.23c) and MAPS (Figure 8.23d).

The SIS models resulted in the highest entropy. This is a expected feature in SIS models. They often show increased small scale variability that results in artificial increase in the



(a) HTPG **(b)** MVHTPG

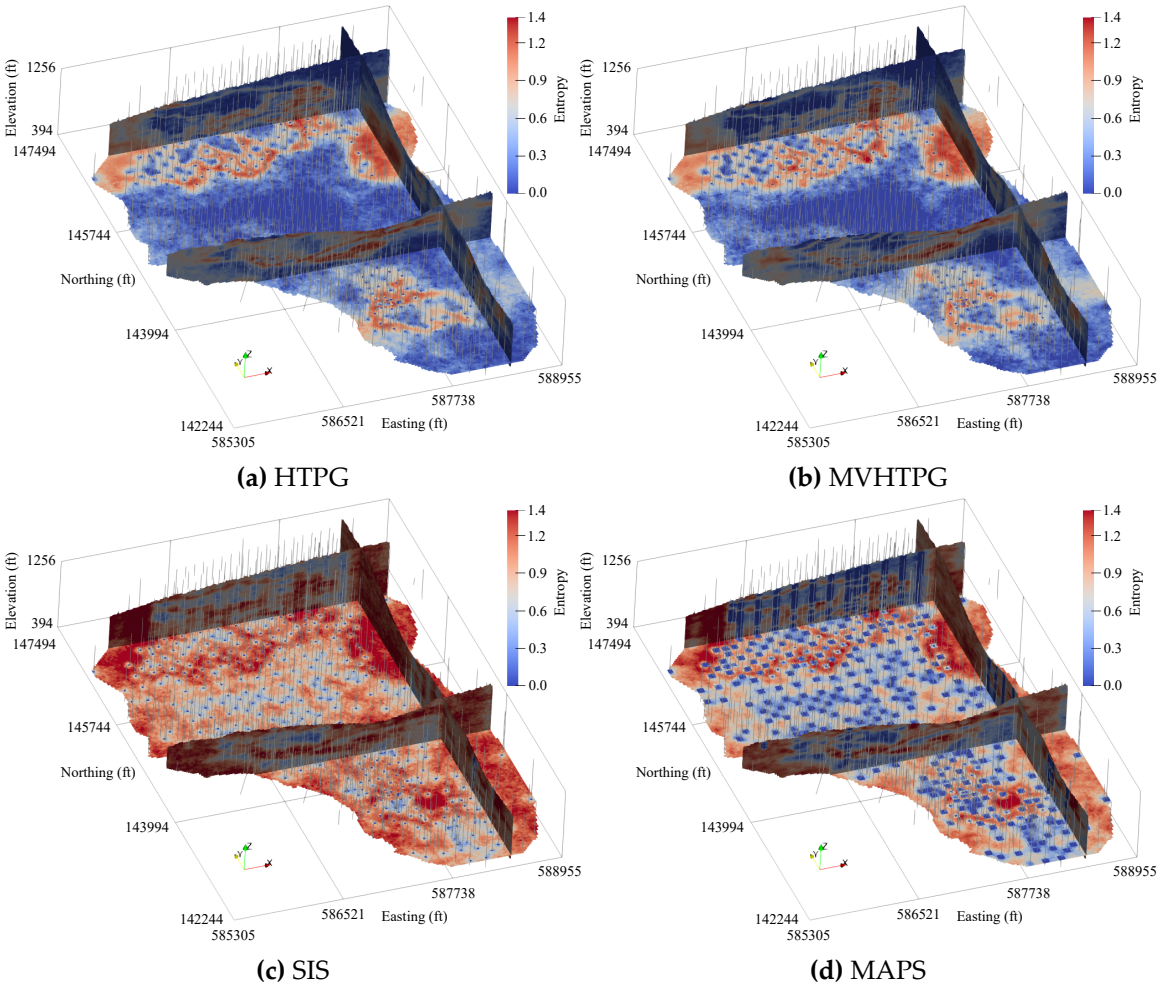
(c) SIS **(d)** MAPS

Figure 8.23: Categorical uncertainty represented by the Shannon entropy calculated over all realizations for the Plates variable generated with HTPG, MVHTPG, SIS and MAPS.

categorical disorder. The entropy is clearly decreased for the cleaned SIS models (MAPS), however, there is an evident short scale contrast close to the drillhole locations which could be artificially introduced by the weighting function utilized with cleaning process. An evident difference in entropy is also observed at the domain boundaries between the HTPG and SIS based models.

Shannon's entropy is also calculated for the realizations of the RT variable generated with each modeling technique. The resulting entropy models are shown in Figure 8.24. The RT variable has categories with shorter spatial continuity than the categories of the Plates variables. This results in an overall increase in entropy values across the models at relatively shorter distance from the conditioning data.

The entropy for the RT variable is similar for the HTPG and MVHTPG models (Fig-



(a) HTPG **(b)** MVHTPG

(c) SIS **(d)** MAPS

Figure 8.24: Categorical uncertainty represented by the Shannon entropy calculated over all realizations for the RT variable generated with HTPG, MVHTPG, SIS and MAPS.

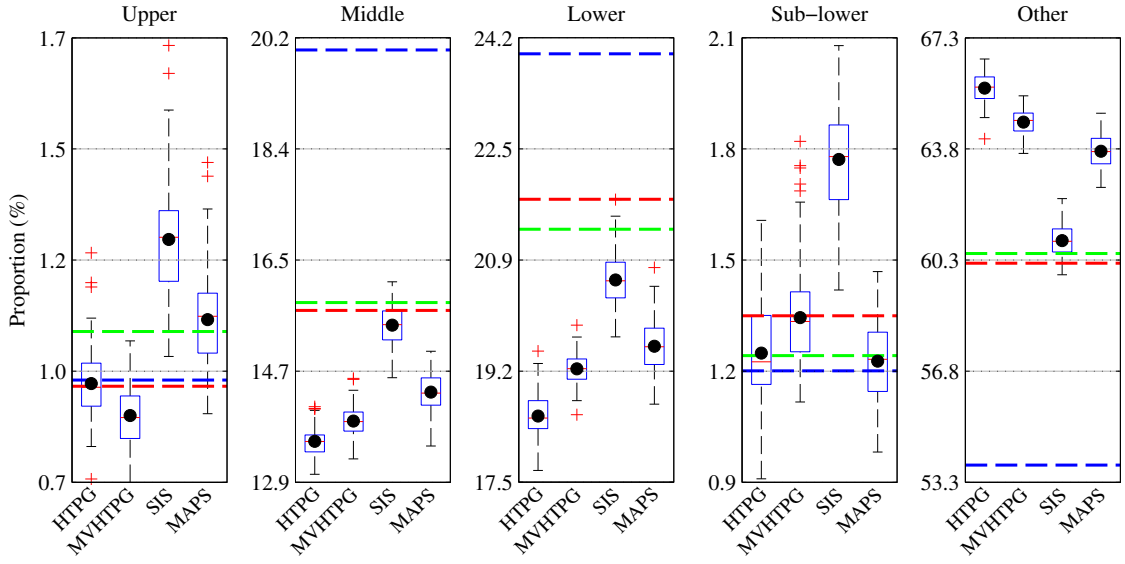
ures 8.24a and 8.24b). The slight differences observed are expected as the RT models are conditioned to the previously simulated models for the Plates variable and are slightly more contained by the enforced multivariate structure. The contrast between the entropy of the HTPG based techniques and the SIS based techniques is clear. Both SIS and MAPS have considerably higher entropy (Figures 8.24c and 8.24d). One of the reasons for the discrepancy is that the Gaussian latent variables utilized with the HTPG techniques carries information regarding the distance to the boundary. The latent variables are constrained to crossing the thresholds depending on the categorical contacts and the spatial structure enforced by the Gibbs sampler determines how quickly the Gaussian values can fluctuate back to values close to the thresholds. These constraints are transfered to the categorical realizations resulting in very different local uncertainty compared to SIS based models.

8.4.1 Reproduction of Global Proportions

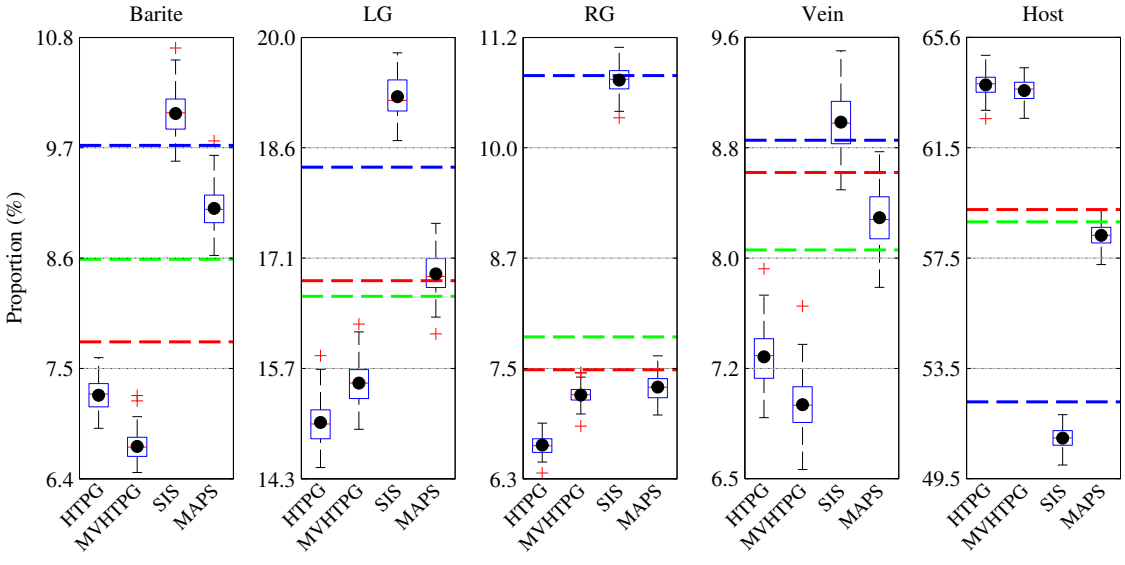
Categorical variables are often utilized as stationary domains for the modeling of continuous variables such as metal concentrations (Rossi and Deutsch, 2014). The proportion of categorical variables have a direct impact on the resources and must be checked carefully. Box-plots are utilized to summarize the global proportions calculated for each category of each categorical variable for each one of the modeling techniques being compared. The results are shown in Figure 8.25. The target global proportions calculated with the weights from cell declustering are shown as green dashed lines in Figure 8.25 and the global proportions calculated from the trend model are shown as red dashed lines. These are the target proportions that are expected to be reproduced. The global proportions calculated without declustering weights are also shown as blue dashed lines in Figure 8.25.

The most noteworthy feature of the global proportion reproduction in this case study is with respect to the SIS modeling techniques. It is well documented in the literature that SIS often underestimates lower proportion categories in favor of the categories with higher proportions due mostly to the order relations corrections that take place within its application (Deutsch, 2006), The MAPS correction is not only utilized to remove noise from the SIS realizations, but also to improve the reproduction of categorical proportions (Deutsch, 1998). In this example the opposite behavior is observed. The lower proportion categories such as Upper and Sub-lower Plates as well as Barite, HG and Vein RT have the highest proportions in SIS simulated models and are lowered when MAPS is used.

It is not possible to state that one particular technique performed better in terms of reproducing the global proportions by visually inspecting the box-plot summary. The mean absolute percent error (MAPE) is calculated for each case using the global proportions from the trend model (red dashed lines) in Figure 8.25 as reference. The MAPE is chosen as it is a relative error measure that equalizes the importance of reproducing the low proportion categories and high proportion categories. The results are shown in Table 8.5 and the best results are highlighted. The techniques performed similarly with MVHTPG showing the best reproduction for the Plates variable and MAPS resulted in the best overall reproduction for the RT variable.



(a) Plates



(b) RT

Figure 8.25: Reproduction of global proportions for the categorical variables utilizing the different modeling techniques. The red dashed line is the global proportion calculated from the trend model, the green dashed line is the declustered global proportion and the blue dashed line is the clustered global proportion.

Table 8.5: Reproduction of categorical proportions for Plates and RT variables utilizing different techniques. The error is measured in terms of the MAPE. The best (lowest error) is highlighted.

Variable	Category	Technique			
		HTPG	MVHTPG	SIS	MAPS
Plates	Upper	6.4	8.7	35.5	16.7
	Middle	13.7	11.6	**2.1**	8.6
	Lower	15.0	11.7	**5.6**	10.1
	Sub-lower	10.4	**8.0**	30.4	9.6
	Other	9.2	7.4	**1.2**	5.9
	Mean	10.9	**9.5**	15.0	10.2
RT	Barite	**7.0**	13.7	30.0	17.5
	LG	10.9	7.9	14.1	**1.4**
	HG	11.3	3.8	43.3	**2.8**
	Vein	15.2	19.1	4.2	**3.9**
	Host	7.6	7.3	14.0	**1.6**
	Mean	10.4	10.4	21.1	**5.4**

8.4.2 Variogram Reproduction

As observed in Figures 8.19 and 8.20, the SIS approach resulted in the noisiest models and MAPS resulted in the smoothest models. The HTPG and MVHTPG generated models with intermediate spatial variability. To state that one is better than the other, the spatial continuity of the realizations must be compared with the spatial continuity of the data. The horizontal and vertical variograms of each realization generated by each technique are calculated and compared to the reference.

The variogram reproduction for the Plates variable is shown in Figure 8.26. The red color is utilized for the vertical direction and the blue color is utilized for the horizontal direction. The light colored lines are the variograms of the realizations and the solid lines are the average variogram over all realizations. The markers connected by dashed lines are the experimental indicator variograms calculated from data. Note that a variogram model for the indicators is not available as it was not necessary for the modeling with a trend. The noisy features observed of the SIS models are clearly shown as inflated short range variability in the variogram reproduction. The HTPG, MVHTPG and MAPS have similar variogram reproduction with MAPS showing an overall better horizontal variogram reproduction while the HTPG and MVHTPG techniques show better vertical variogram reproduction for the Plates variable.

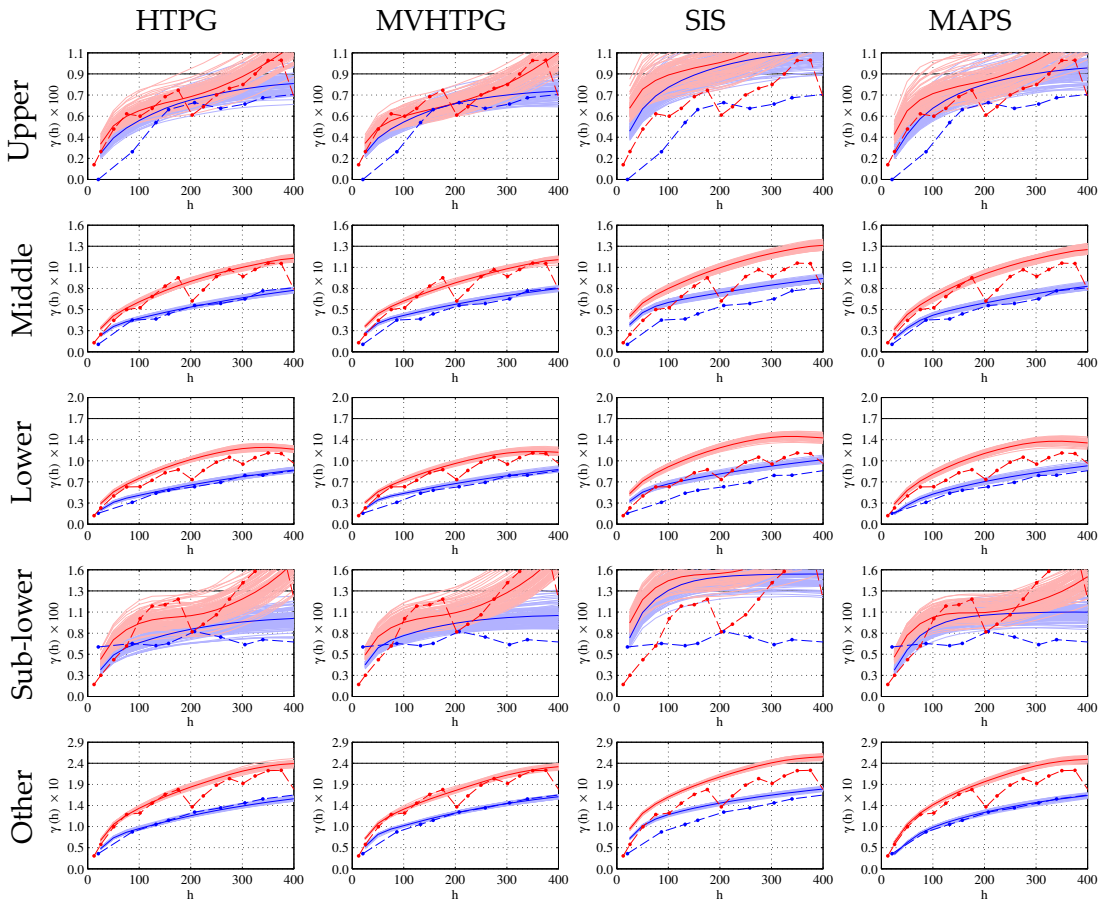


Figure 8.26: Variogram reproduction for the categories of the Plates variable. The red color is utilized for the vertical direction and the blue color is utilized for the horizontal direction. The light colored lines are the variograms of the realizations and the solid lines are the average variogram over all realizations. The markers connected by dashed lines are the experimental indicator variograms.

The variogram reproduction for the RT variable is shown in Figure 8.27. Again, the same behavior is observed for the models generated by the SIS technique. MAPS in this case was not sufficient to remove the inflated variability. The HTPG and MVHTPG have similar variogram reproduction and have clearly superior performance when compared with the SIS and MAPS techniques.

8.4.3 Transition Probabilities

Transition probabilities are a quantitative measure of the geological structure. The truncated Gaussian methods have the ability to explicitly control some features of the transitions between categories, whereas, the SIS method does not have the same capability. It is noted, however, that the local trend is utilized for the control of transitions for the Plates

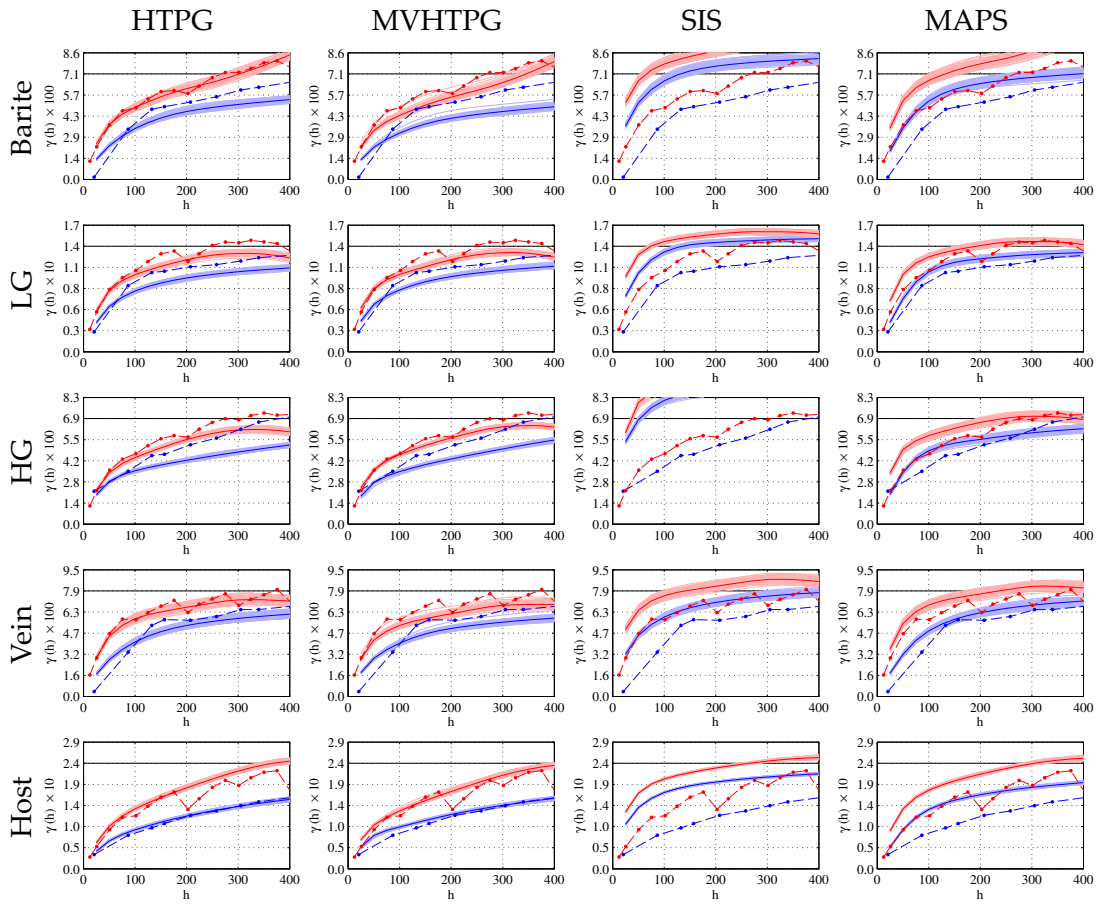


Figure 8.27: Variogram reproduction for the categories of the RT variable. The red color is utilized for the vertical direction and the blue color is utilized for the horizontal direction. The light colored lines are the variograms of the realizations and the solid lines are the average variogram over all realizations. The markers connected by dashed lines are the experimental indicator variograms.

variables and no explicit geological control is utilized in the truncation rule defined for the Plates variable. In light of these observations the reproduction of transition probabilities of HTPG and SIS based methods for the Plates variable is expected to be somewhat similar.

The RT variable, on the other hand, has the transition enforced by both trend model and truncation structure. The Gaussian variable $Y_{2,2}$ in Figure 8.11, for instance, is utilized to decrease the likelihood of transition between Barite and Vein categories. The HTPG and MVHTPG are expected to have better transition reproduction for the RT variables when compared with SIS and MAPS due to the combination of the truncation rule and trend model for the enforcement of transitions.

The average scaled transition probabilities calculated over all realizations for each modeling technique and for each variable are shown in Figure 8.28. The reproduction of the transition probabilities for the Plates variable is similar across all modeling techniques. This is

expected due to the aforementioned reasons. The separation between the Upper and Middle plates from the Lower and Sub-lower is slightly better reproduced by the HTPG and MVHTPG methods. The transition between the Other plate to the Upper plate and Sub-lower plate is inflated for all models compared to the reference (Figure 8.8b) and, as result, there are less than expected transitions to Middle and Lower plates. This can potentially be improved by a more restrictive truncation rule for the HTPG techniques perhaps explicitly separating the Upper and Sub-lower plates from the Middle and Lower by placing the Other category in between. This would be similar to what is done for the RT variable.

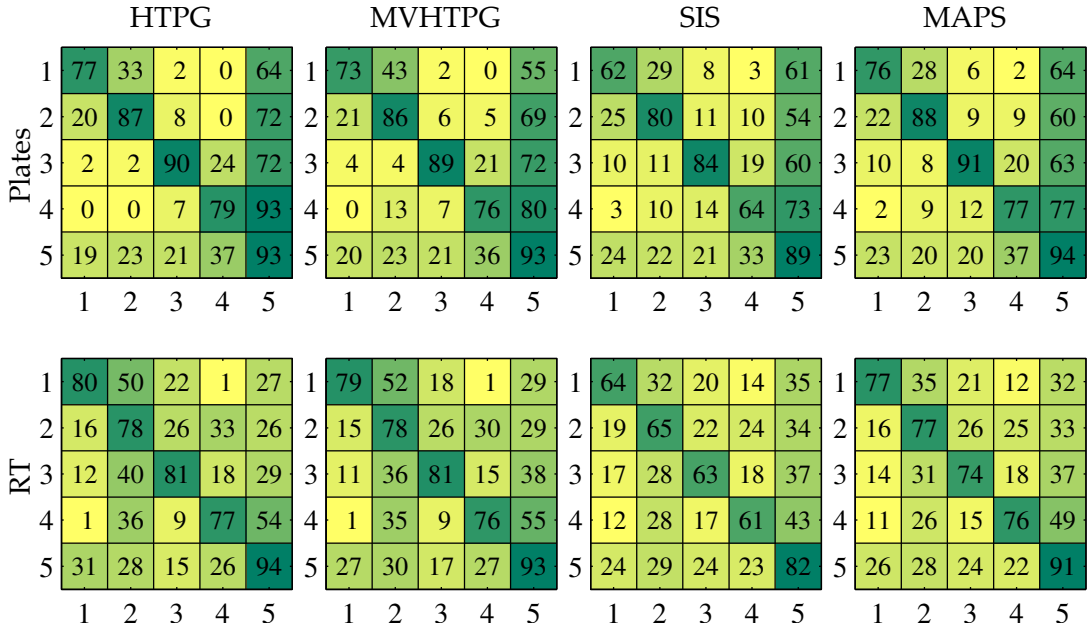


Figure 8.28: Average scaled transition probabilities calculated over all realizations for each modeling technique for Plates and RT variables.

The reproduction of the transition probabilities for the RT variable is slightly better for the HTPG and MVHTPG approaches. The restriction on the transition between the Barite and Vein categories can be clearly observed by comparing the matrices for SIS and MAPS versus the matrices for HTPG and MVHTPG in Figure 8.28. The root mean squared error (RMSE) is calculated for each realization and the distribution for each modeling technique is summarized in Figure 8.29. There is no theoretical reason for the lower performance of MVHTPG compared to the HTPG, specially for the Plates variable. The only difference between the two is in the conditioning data. The differences can be attributed to the performance of the Gibbs sampler utilized in each application.

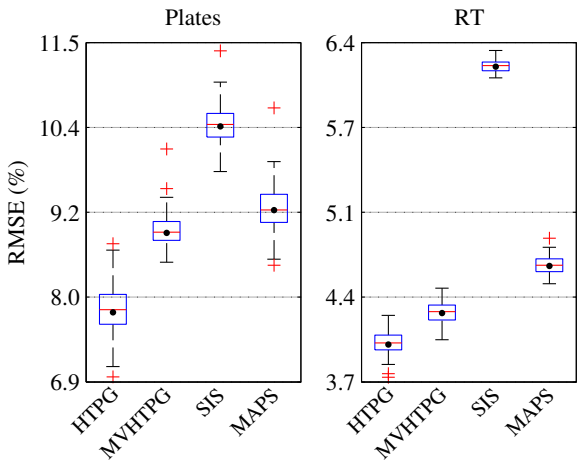


Figure 8.29: Distribution of error on the reproduction of the transition probabilities for each modeling technique.

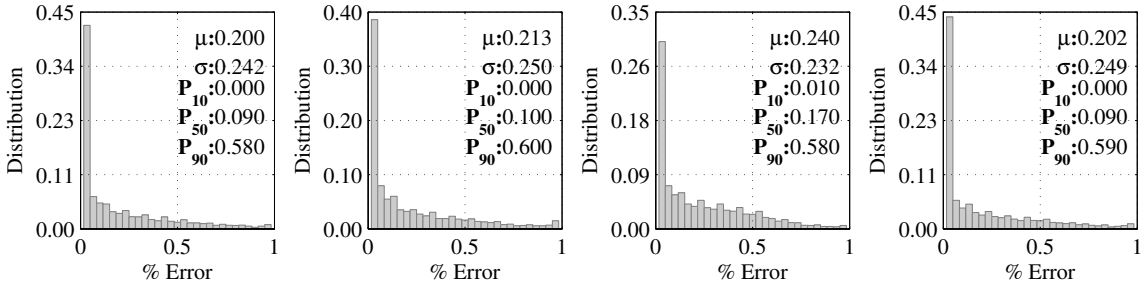
8.4.4 Validation

HTPG and MVHTPG are shown to have very different characterization of the local uncertainty if compared to SIS and MAPS techniques. HTPG and MVHTPG show an overall decrease in local uncertainty especially for the RT variable. A lower uncertainty does not necessarily mean better models. This is only the case if the models are also accurate. The test data set left out of the modeling workflow is utilized to evaluate the accuracy of the models.

8.4.4.1 Prediction Error

The most obvious validation to perform is to check the reproduction of the categorical observations at the test data locations. For each data location the true category is compared with the simulated value at the closest grid node and the mismatch are counted. The total count is divided by the number of realization and a percent error is calculated. The percent error calculated for each data location is utilized to generate an error distribution. This distribution is shown in Figure 8.30 for the Plates variable.

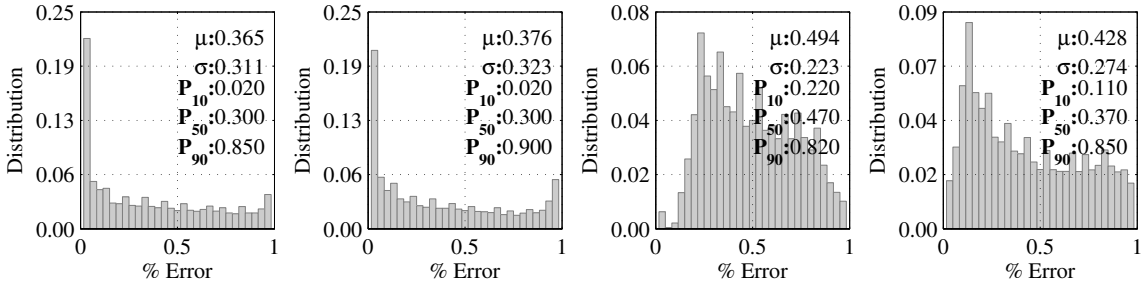
All techniques perform similarly with relation to reproducing the categories of the Plates variable. This is not surprising as even with the data left out the structures of this variable are large if compared with the drilling density. HTPG resulted in the lowest error followed closely by MAPS. SIS resulted in the highest mismatch between the true and simulated categories.



(a) HTPG **(b)** MVHTPG **(c)** SIS **(d)** MAPS

Figure 8.30: Prediction error for the Plates variable for each modeling technique.

The error distribution for the RT variable is shown in Figure 8.31. The RT variable has much smaller features relative to the drillhole spacing that are more difficult to model accurately. The differences in performance between HTPG and MVHTPG compared to SIS and MAPS are visible. The developed methodologies show better reproduction of the categories of the RT variable.



(a) HTPG **(b)** MVHTPG **(c)** SIS **(d)** MAPS

Figure 8.31: Prediction error for the RT variable for each modeling technique.

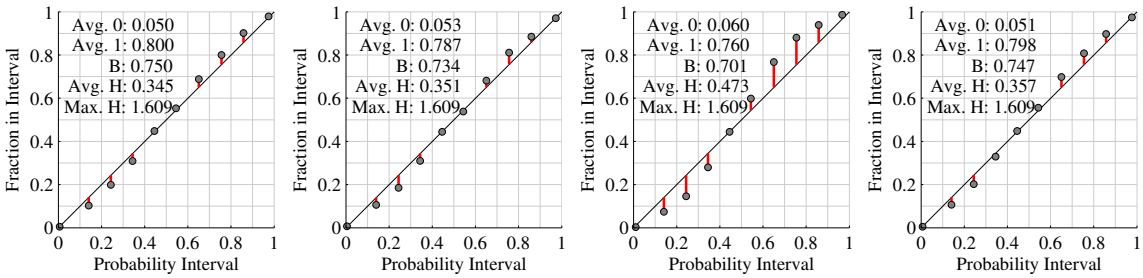
8.4.4.2 Probabilistic Accuracy

The probabilistic accuracy is evaluated by discretizing the probability interval in several bins, calculating the predicted frequency in each one of these bins and then comparing the predicted frequency with the actual fraction of each category in each one of these bins (Deutsch and Deutsch, 2012). The predicted and actual fraction in each interval are utilized to generated an accuracy plot where being close to the 45 degree line indicates that the modeling technique is accurate. Other summary statistics such as entropy and B values are also shown (Deutsch and Deutsch, 2012). The B value is the average predicted probability when the true value is 1 and when the true value is 0. The highest the B value the better

reproduction of the categorical indicator. The B value has similar meaning to the mean prediction error shown in the previous section.

The expected features in an accuracy plot are low entropy coupled with high B value and little deviation from 45 degree line. This means that the modeling technique is both accurate and precise. Low entropy and low B value indicates bias and is usually coupled with points consistently above or below the 45 degree line. A high entropy with low B value indicates a models that are inaccurate and not precise and are usually characterized by large spread from the 45 degree line.

The accuracy plots for the Plates variable are shown in Figure 8.32. The B values show the same result as observed for the prediction error. The accuracy is similar for the HTPG, MVHTPG and MAPS whereas the SIS technique show features of a precise but inaccurate models for this variable.



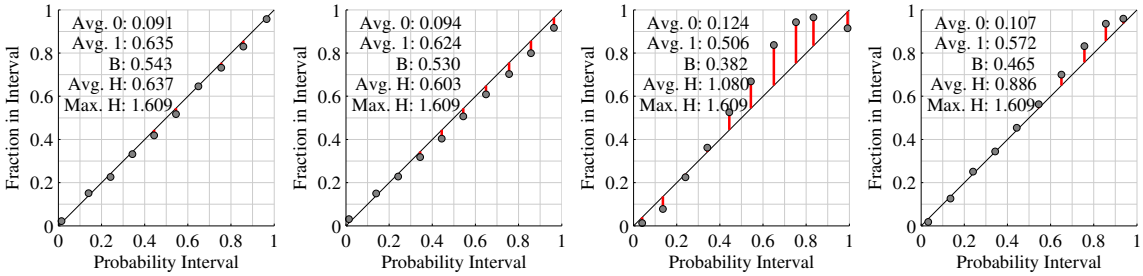
(a) HTPG **(b)** MVHTPG **(c)** SIS **(d)** MAPS

Figure 8.32: Accuracy plot for Plates variable.

The accuracy plots for the RT variable are shown in Figure 8.33. HTPG shows the desired features with the most accurate and precise models amongst all the modeling techniques. The MVHTPG is also precise but it shows slight bias which means that it is not as accurate as the HTPG models. The SIS and MAPS are significantly less precise than HTPG and MVHTPG. The SIS models are also less accurate compared to all other models as it shows significantly higher uncertainty as indicated by the entropy values.

8.4.4.3 Metallurgical Recovery

All the performance evaluations undertaken so far are checking univariate features. None of the measures have a direct dependence on the multivariate relationships between the categorical variables. To evaluate the impact of the developed multivariate workflow represented by MVHTPG against the univariate alternatives HTPG, SIS and MAPS, the met-



(a) HTPG **(b)** MVHTPG **(c)** SIS **(d)** MAPS

Figure 8.33: Accuracy plot for RT variable.

allurgical recoveries shown in Table 8.2 are applied to the composites of the test data. The true overall metallurgical recovery is the mean value considering all the composites and it is calculated to be 31.42%.

The simulated values at the closest nodes to the test data set are utilized to calculate the overall metallurgical recovery for each realization for each modeling technique. The 100 values for the simulated overall recovery are utilized to generate a histogram for comparison with the true value for each technique. The results are shown in Figure 8.34.

As the recovery is highly dependent on the combination of categories, it is not a surprise that the only methodology that is able to accurately predict the overall metallurgical recovery is the only multivariate approach considered. The MVHTPG results are far superior to all the others with RMSE of 1.36% whereas the next best result is given by the HTPG technique with a RMSE of 32.78%. All the univariate approaches show highly biased results. The MVHTPG also show the least uncertainty which makes it the most accurate and the most precise modeling technique.

The improved performance on the reproduction of the metallurgical recovery is attributed to the enforcement of the joint categorical PDF by the MVHTPG workflow. The average joint PDF calculated over all realizations for each of the modeling techniques are shown in Figure 8.35. As expected, the best reproduction is achieved with the MVHTPG with a RMSE of 4.8%. The joint PDF is enforced to some degree for all methodologies. This is attributed to the trend model and conditioning data. Amongst the univariate techniques the HTPG resulted in the lowest RMSE and the SIS resulted in the highest.

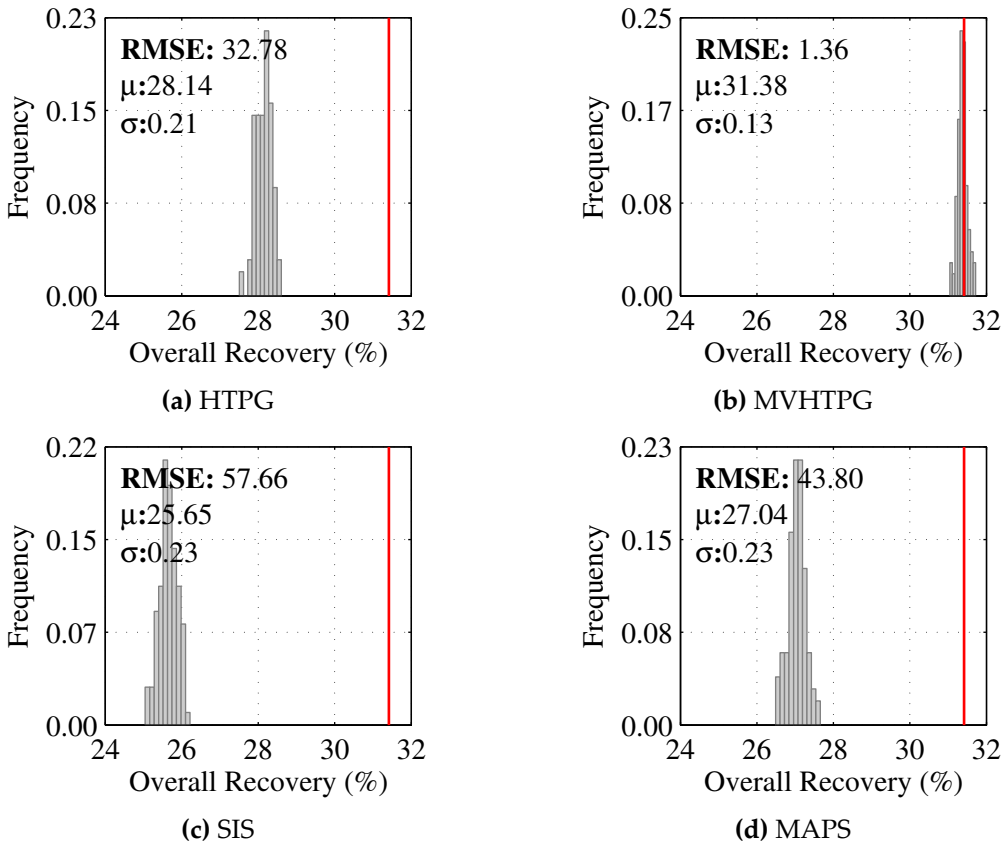
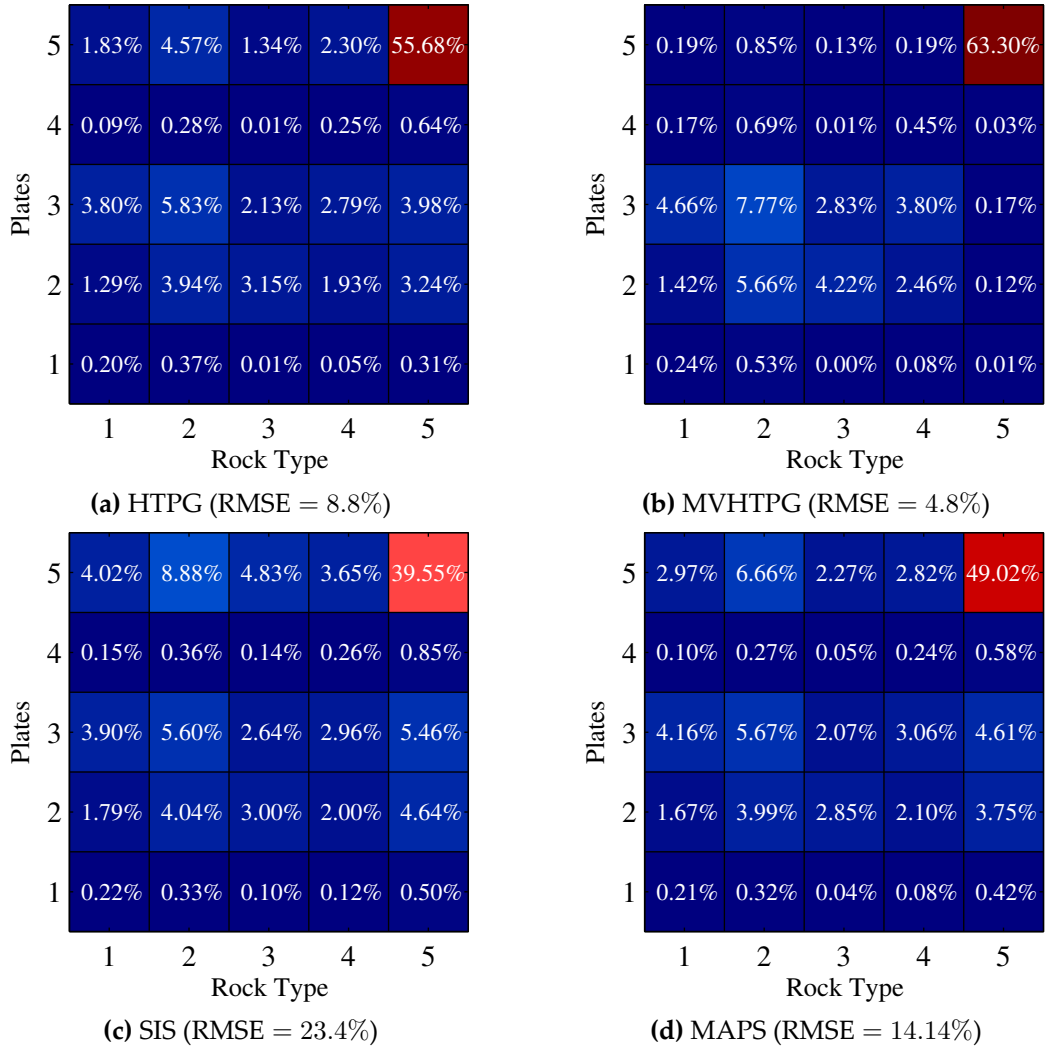


Figure 8.34: Histogram of the overall recovery calculated from the simulated categorical variables for each modeling technique. The true overall recovery is shown as a vertical red line at 31.42%.

8.5 Conclusion

The two techniques for categorical modeling developed in this dissertation are demonstrated for a data set from the Red Dog Mine. The univariate and multivariate versions of the HTPG are compared against the SIS modeling technique. The comparison also accounts for the cleaned SIS models (MAPS) as it is a common practice to post process the models generated by SIS to remove noise and improve reproduction of categorical proportions.

The SIS technique underperformed in all model checking measurements. The overall results show good accuracy in the reproduction of the univariate parameters for the HTPG, MVHTPG and MAPS, however, HTPG and MVHTPG were more precise as they consistently showed lower uncertainty. The categories of the Plates variable are spatially continuous and the model validation with the test data set was reasonable with all techniques. The HTPG and MVHTPG showed clear performance improvement for the less



(a) HTPG (RMSE $= 8.8\%$) **(b)** MVHTPG (RMSE $= 4.8\%$)

(c) SIS (RMSE $= 23.4\%$) **(d)** MAPS (RMSE $= 14.14\%$)

Figure 8.35: Average joint categorical PDF calculated over all realizations for each modeling technique and respective RMSE calculated using the declustered distribution calculated from data as reference.

spatially continuous categories from the RT variable with a significantly better prediction of the categories from the test data set.

The univariate HTPG technique outperformed the MVHTPG for the univariate measures. This is mainly attributed to the less than ideal convergence the multivariate Gibbs sampler utilized with the MVHTPG. Even with the convergence difficulties, MVHTPG outperformed the SIS based models in most univariate measurements.

As expected, MVHTPG showed the best reproduction of the joint PDF. The impact of the multivariate workflow is also shown with regard to the reproduction of the overall metallurgical recovery of the test data set. MVHTPG was the only technique to achieve

unbiased predictions and was also the technique with the lowest uncertainty. This makes MVHTPG the most precise and accurate amongst all tested techniques. Being precise and accurate in the prediction of key attributes affecting the economical feasibility of a mining project is key for mitigating the risk of such projects.

Chapter 9

Concluding Remarks

Multiple techniques are available for modeling categorical variables. The goal is to reproduce the key features observed in data and achieve the the most precise and accurate predictions to minimize risk and maximize the value of the data acquired. This dissertation develops a new framework for the application of truncated Gaussian methods that allows for greater flexibility and facilitates the use of geological interpretation in a straightforward manner.

The development of this new framework led to the adaptation and improvement of existing techniques that are required for the application of any truncated Gaussian method. This includes the mapping of the spatial structure from categorical to continuous variables and the implementation of the Gibbs sampler algorithm for the imputation of the Gaussian latent variables. The work undertaken for multiple data imputation also generated important contributions for multivariate modeling of continuous variables.

The foundations developed for the univariate hierarchical truncated pluri-Gaussian (HTPG) created an opportunity to explore the multivariate modeling of categorical variables. This is an area of little research and limited available options despite the fact that the multiple categorical variables are the key controls for grade distribution and other important variables such as metallurgical recovery. A summary of the key contributions of this thesis as well as the main limitations and future work are outlined in this chapter.

9.1 Summary of Contributions

The main contributions of this thesis are in the field of truncated Gaussian techniques for the modeling of categorical variables. There are also important contributions to multiple data imputation applied to the modeling of continuous variables.

9.1.1 Hierarchical Truncated pluri-Gaussian

The HTPG modeling framework developed in this thesis addresses some of the limitations of the existing applications of truncated Gaussian methods. Truncated Gaussian methods were created to facilitate the introduction of transition constraints based on knowledge of the geological setting being modeled. The truncated Gaussian simulation (TGS) method applied with a single Gaussian variable allowed the modeling of simple geological settings with clear categorical ordering. Even in a simple geological setting, the control on transitions is not possible with other techniques such as sequential indicator simulation (SIS). The potential of the truncated Gaussian techniques was extended to the utilization of multiple Gaussian variables with the truncated pluri-Gaussian simulation (TPGS). The utilization of multiple Gaussian variable allowed the construction of more complex mappings between the categorical variables and continuous latent Gaussian variables, however, the application of TPGS was mostly restricted to the utilization of two Gaussian variables. This is attributed to the increased difficulty of building geologically sound truncation rules and also to the increased complexity in the required steps such as the mapping of the spatial continuity between the categorical and continuous space.

The limitations on the utilization of multiple Gaussian variables with TPGS framework is overcome with the introduction of the hierarchical approach. The HTPG technique developed in this thesis facilitates the definition of truncation rules for complex geological settings without sacrificing the ability of devising a sound geological interpretation. The development of the new truncation framework led to the development and adaptation all the required steps of the application of truncated Gaussian methods and, as result, contributions are made on the numerical derivation of the latent variable variograms and latent variable imputation.

A detailed description of the practical considerations for the application and parameterization of the HTPG is provided that includes: (1) the tools for the definition of the truncation rule based on a transition matrix; (2) modeling in a non-stationary setting; and (3) identification and mitigation of hyper-continuity for improved variogram reproduction.

The technique is demonstrated for a data set from the Red Dog Mine and shown to perform well compared with alternative approaches such as the SIS based models. The utilization of truncation rules leads to more constrained models that show greater accuracy with lower uncertainty. This is crucial for risk management and maximization of infor-

mation usage. The multivariate HTPG makes the most use of the data when it comes to the prediction of attributes that are dependent on the complex relations resulting from the different combinations of multiple categorical variables.

9.1.1.1 Numerical Derivation of Latent Variables Variogram

The tree based truncation rule scheme utilized by the HTPG approach facilitates the utilization of a large number of latent variables for the modeling of an arbitrary number of categories. Even though the framework is designed to facilitate the geological interpretation, there is an increased complexity in other steps of the modeling workflow. One of these steps is the derivation of the variograms for the Gaussian latent variables. This step is mentioned by Armstrong et al. (2011) as a deterring factor for the utilization of more than two Gaussian variables with the TPGS method.

The framework developed for the HTPG utilizes a numerical derivation technique improved upon the method proposed by Zagayevskiy and Deutsch (2015). The technique is shown to work well and it is only limited by the degrees of freedom of the truncation itself. This is often the restricting factor for the reproduction of categorical variograms with reasonable spatial structure within the Gaussian space.

Hyper-continuity is a common issue with the derivation of Gaussian variograms. This is not a limitation of the numerical derivation procedure itself. This is a limitation of the physically possible combination of spatial structure for the categorical variable given the chosen truncation rule. The causes and effects of hyper-continuity are investigated in this thesis and mitigating actions are suggested. In many cases, the hyper-continuity cannot be addressed, however, in some instances the closest possible fitting is good enough in practice. For instance, hyper-continuous structures are observed for the case study in Chapter 8 and it did not prevent the HTPG methods from outperforming SIS at variogram reproduction.

9.1.1.2 Multiple Data Imputation of Latent Variables

The Gibbs sampler is utilized for the imputation of the Gaussian latent variables subject to the categorical data observations. A detailed overview of the available Gibbs sampler algorithms is presented and suggestions are made to improve the stability and convergence

of such algorithms. The suggested combination of the standard Gibbs sampler and propagative Gibbs sampler is shown to improve stability and convergence.

An alternative based on simulated annealing is proposed. The technique can be used in some instances where additional constraints and objectives are present or when convergence with the conventional Gibbs sampler approach cannot be achieved.

9.1.2 Multiple Data Imputation with Gaussian Mixture Models

The latent variables in the truncated Gaussian methods are nearly always a model assumption and are not observed in practice. This makes the application of truncated Gaussian techniques a missing data problem. Research in the field of multiple data imputation has led to the improvement of the Gibbs sampler approach for imputation of latent variables. Research has also led to the development of an alternative to Gibbs sampler approach for the application with the multivariate modeling of continuous variables in the context of multivariate transformations.

The truncated multivariate Gaussian distributions are complex and the sharp boundaries cannot be easily parameterized with any parametric of semi-parametric model. This makes the Gibbs sampler the best available practical alternative for the imputation of latent variables given that the convergence of the algorithm can be achieved within reasonable computational time.

The joint distributions observed for the continuous geological variables, on the other hand, can be well represented by semi-parametric models such as with the fitting of a Gaussian mixture model (GMM). The utilization of a semi-parametric model removes the requirement for the utilization of the Gibbs sampler. A technique for multiple data imputation using a GMM is developed in Chapter 6 and it is shown to not only increase the computational performance by orders of magnitude for large data sets, but it also improves the overall accuracy and precision of the imputed data. This is demonstrated for a lateritic nickel data set in Chapter 6 and it represents an important contribution to the multivariate geostatistical modeling of continuous variables.

9.1.3 Multivariate Categorical Modeling

Categorical variables controls the modeling of many important aspects of a mining project such as ore concentrations, metallurgical recovery and structural stability. Each one of

these features are affected by a different combination of these categorical variables. The current paradigm is to combine these categorical variables into a single model with a limited number of categories. The combined categories are utilized as modeling domains for the continuous attributes. This practice leads to the underutilization of important and expensive information. Each aspect of interest depends on a different combination of categorical variables. The univariate methods can only work with one particular combination and a trade off must be considered when merging categorical information. This leads to inevitable deterioration of the model prediction potential.

There is very limited research in the field of multivariate geostatistical modeling of categorical variables. The development of the univariate HTPG and the tools for its application created the groundwork for the development of a multivariate categorical modeling technique. The extension of the HTPG to the multivariate case is done by introducing correlation between the Gaussian latent variables that are independently defined for the modeling of each categorical variable. The correlation is calculated by targeting the reproduction of the categorical joint probability density function (PDF) calculated experimentally from the data.

The developed multivariate HTPG technique shows similar performance for the reproduction of the univariate parameters such as global proportions, variogram models, prediction error and probabilistic accuracy. The greatest advantage of the multivariate approach is observed for the features that are highly dependent on the combination of the categorical variables such as the metallurgical recovery. It is shown for the Red Dog Mine case study that the multivariate categorical modeling developed in this thesis is able to precisely and accurately predict the overall metallurgical recovery of the test data set, left out of the modeling workflow.

The multivariate HTPG technique developed in this thesis represents an important contribution to the geostatistics. This is perhaps the most flexible approach available at this time as it has no theoretical restriction on the number of Gaussian latent variables and categorical variables that can be modeled together. The modeling framework is setup in such way that any combination that can be modeled independently with HTPG can be easily extended to the multivariate case by introducing the correlation across the latent variables representing different categorical variables.

9.2 Limitations and Future Work

Despite the developments made in this dissertation there are several limitations to the application of the developed techniques. Important limitations of the univariate and multivariate HTPG are described as well as the limitations with the multiple data imputation of continuous variables with GMM.

9.2.1 Imputation of Latent Variables

The main limitation of the application of the HTPG technique is the convergence rate of the Gibbs sampler approach. The combination of the propagative Gibbs sampler and the standard approach improved the convergence, however, it can still pose a problem when the latent variables show long range spatial correlation. The highly correlated variables have a much slower convergence rate and in some instances the number of burn-in iterations required to ensure convergence is impractical.

The propagative Gibbs sampler requires the computation of the covariance values between all data available and at each burn-in iteration several loops through all the data is required to define the probabilistic boundaries for sampling. Also, the change in one data location must be propagated through all data locations. As a result the propagative Gibbs sampler requires a large amount of memory and computational time that grows exponentially with increasing number of data. It may be impractical to run the Gibbs sampler algorithm until convergence for large data sets with long range correlation structures.

For the multivariate HTPG approach, there are even greater limitations. The number of latent variables being modeled are usually larger than for univariate HTPG. The multivariate modeling of three categorical variables, for instance, can potentially have three times more latent variables than what would be considered in the univariate case. In addition to the added complexity by number of variables alone, there is also the added complexity from the introduction of correlation between the latent variables. A propagative Gibbs sampler approach could still be outlined in theory, however, its practical application would be seriously limited due to memory and computational time requirements. For this reason, only the standard Gibbs sampler approach has been adapted for the multivariate case.

The standard Gibbs sampler for the multivariate HTPG is subject to the same limitations regarding the convergence as for the univariate case. The practitioner must be aware of this

limitation and check the imputed data. In some instances, when full convergence cannot be achieved the imputed values can still be used as long as the values do not diverge to extreme highs and lows. The utilization of values that are not fully converged will lead to a loss in the overall performance of the technique. The results are expected to be somewhere between the correct uncertainty characterization and the characterization utilizing the fixed class centroid alternative investigated in Chapter 5. The multivariate Gibbs sampler also utilizes a IMC approximation. This can lead to the mixing of spatial structure between latent variables if two correlated variables have different spatial structure.

9.2.2 Multivariate HTPG

The current approach utilized in the multivariate HTPG is only focused on the reproduction of the joint PDF of the categorical variables utilizing collocated correlation and the IMC approximation. The information contained in the joint PDF is equivalent to the non-zero elements of the zero lag transition probability matrix. The transition probabilities between the categories of the same set are zero (mutually exclusive) and the non-zero terms are transitions between categories in different sets. The transition probabilities can be defined at several lag sizes and may contain important information regarding higher order relations between the categorical variables. This is currently not accounted for by the developed technique.

9.2.3 GMM Based Multiple Data Imputation

The developed methodology for multiple data imputation for the multivariate modeling of continuous variables utilizes a fitted semi-parametric model to define the conditional distributions given the collocated data. The spatial correlation across the different variables is disregarded in this process and the only spatial correlation accounted for is for the primary variables being imputed. Disregarding the spatial contribution of the non-collocated secondary variables can lead to variance inflation similar to what is observed with collocated cokriging (Babak and Deutsch, 2009).

The current implementation has no means of automatically defining the number of Gaussian kernels to be utilized for the fitting of the GMM. This can potentially lead to over/under-fitting of the data multivariate PDF and result in unreasonable imputed values. Also, the current implementation has no means of accounting for preferential sampling.

The fitted GMM can be potentially biased. This is not major issue for data imputation as the realizations of data are highly conditioned by the nearby available data.

9.2.4 Future Work

Future work could be focused on the solution of the major limiting factors for the application of the developed methodology. Additional work is necessary on the development of a stable implementation of the Gibbs sampler algorithm for utilization with the univariate and multivariate HTPG. The development of practical alternatives to the Gibbs sampler is also a possibility. The simulated annealing alternative presented in this dissertation can be improved to allow for its practical application with both univariate and multivariate HTPG.

The impact of higher order relationships between categorical variables at multiple lag distances should also be focus of future research. The possible benefits of considering a full linear model of coregionalization (LMC) over the currently proposed intrinsic model of coregionalization (IMC) approximation can also be evaluated in future work.

References

Acar, S. (2016). Process development metallurgical studies for gold cyanidation process. *Minerals & Metallurgical Processing*, 33(4).

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2 edition.

Alabert, F. (1987). The practice of fast conditional simulations through the lu decomposition of the covariance matrix. *Mathematical Geology*, 19(5):369–386.

Allard, D., D'Or, D., Biver, P., and Froidevaux, R. (2012). Non-parametric diagrams for pluri-gaussian simulations of lithologies. In *9th international geostatistical congress, Oslo, Norway*, volume 1115.

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

Angove, J. and Acar, S. (2016). Metallurgical test work: Gold processing options, physical ore properties, and cyanide management. In Adams, M. D., editor, *Gold Ore Processing (Second Edition)*, pages 131 – 140. Elsevier, second edition edition.

Armstrong, M., Galli, A., Beucher, H., Loc'h, G., Renard, D., Doligez, B., Eschard, R., and Geffroy, F. (2011). *Plurigaussian simulations in geosciences*. Springer-Verlag Berlin Heidelberg, 2 edition.

Arroyo, D., Emery, X., and Peláez, M. (2012). An enhanced gibbs sampler algorithm for non-conditional simulation of gaussian random vectors. *Computers & Geosciences*, 46(Supplement C):138 – 148.

Arthur, D. and Vassilvitskii, S. (2007). K-means++: the advantages of careful seeding. In *In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*.

Astrakova, A., Oliver, D. S., and Lantuéjoul, C. (2015). Truncation map estimation based on bivariate probabilities and validation for the truncated plurigaussian model. *arXiv preprint arXiv:1508.01090*.

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). An intrinsic model of coregionalization that solves variance inflation in collocated cokriging. *Computers & Geosciences*, 35(3):603 – 614.

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

Barnett, R. M. and Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. *Mathematical Geosciences*, 47(7):791–817.

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

Bellman, R. (2003). *Dynamic Programming (Dover Books on Computer Science)*. Dover Publications.

Bezier, P. et al. (1974). Mathematical and practical possibilities of unisurf. *Computer Aided Geometric Design*, 1(1).

Black, T. C. and Freyberg, D. L. (1990). Simulation of one-dimensional correlated fields using a matrix-factorization moving average approach. *Mathematical Geology*, 22(1):39–62.

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

Bowell, R., Grogan, J., Hutton-Ashkenny, M., Brough, C., Penman, K., and Sapsford, D. (2011). Geometallurgy of uranium deposits. *Minerals Engineering*, 24(12):1305 – 1313. Special Issue : Process Mineralogy.

Bye, A. (2011). Case studies demonstrating value from geometallurgy initiatives. In *GeoMet 2011-1st AusIMM International Geometallurgy Conference 2011*, pages 9–30. AusIMM: Australasian Institute of Mining and Metallurgy.

Carle, S. F. and Fogg, G. E. (1996). Transition probability-based indicator geostatistics. *Mathematical Geology*, 28(4):453–476.

Chilès, J. and Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty*. Wiley series in probability and statistics. John Wiley & Sons.

Christakos, G. (1990). A bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22(7):763–777.

Cowan, E., Beatson, R., Ross, H., Fright, W., McLennan, T., Evans, T., Carr, J., Lane, R., Bright, D., Gillman, A., et al. (2003). Practical implicit geological modelling. In *Fifth International Mining Geology Conference*, pages 89–99, Bendigo, Victoria. Australasian Institute of Mining and Metallurgy.

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

Davis, M. W. (1987). Production of conditional simulations via the lu triangular decomposition of the covariance matrix. *Mathematical Geology*, 19(2):91–98.

Delalleau, O., Courville, A., and Bengio, Y. (2012). Efficient em training of gaussian mixtures with missing data. *arXiv preprint arXiv:1209.0521*.

Desbarats, A. 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. (1998). Cleaning categorical variable (lithofacies) realizations with maximum a-posteriori selection. *Computers & Geosciences*, 24(6):551 – 562.

Deutsch, C. V. (2006). A sequential indicator simulation program for categorical variables with point and block data: Blocksis. *Computers & Geosciences*, 32(10):1669 – 1681.

Deutsch, C. V. and Cockerham, P. W. (1994). Practical considerations in the application of simulated annealing to stochastic simulation. *Mathematical Geology*, 26(1):67–82.

Deutsch, C. V., Journel, A. G., et al. (1998). *Geostatistical software library and user's guide*. Oxford University Press, 2 edition.

Deutsch, J. L. (2015). Variogram program refresh. Technical report, Centre for Computational Geostatistics, University of Alberta.

Deutsch, J. L. and Deutsch, C. V. (2012). Accuracy plots for categorical variables. Technical report, Centre for Computational Geostatistics, University of Alberta.

Deutsch, J. L. and Deutsch, C. V. (2014). A multidimensional scaling approach to enforce reproduction of transition probabilities in truncated plurigaussian simulation. *Stochastic Environmental Research and Risk Assessment*, 28(3):707–716.

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.

Emery, X. (2004). Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, 18(6):414–424.

Emery, X. (2007). Simulation of geological domains using the plurigaussian model: New developments and computer programs. *Computers & Geosciences*, 33(9):1189 – 1201.

Emery, X., Arroyo, D., and Peláez, M. (2014). Simulating large gaussian random vectors subject to inequality constraints by gibbs sampling. *Mathematical Geosciences*, 46(3):265–283.

Emery, X. and Cornejo, J. (2010). Truncated gaussian simulation of discrete-valued, ordinal coregionalized variables. *Computers & Geosciences*, 36(10):1325 – 1338.

Fix, E. and Hodges Jr, J. L. (1951). Discriminatory analysis-nonparametric discrimination: consistency properties. Report number 4, USAF School of Aviation Medicine.

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

Galli, A., Beucher, H., Le Loc'h, G., Doligez, B., et al. (1994). The pros and cons of the truncated gaussian method. In *Armstrong M., Dowd P.A. (eds) Geostatistical Simulations*, volume 7, pages 217–233. Springer, Dordrecht.

Galli, A. and Gao, H. (2001). Rate of convergence of the gibbs sampler in the gaussian case. *Mathematical Geology*, 33(6):653–677.

Garza, R. A. P., Titley, S. R., and Pimentel B., F. (2001). Geology of the escondida porphyry copper deposit, antofagasta region, chile. *Economic Geology*, 96(2):307–324.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741.

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

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

Gray, A. G. and Moore, A. W. *Nonparametric Density Estimation: Toward Computational Tractability*, pages 203–211.

Gregory, M. J., Lang, J. R., Gilbert, S., and Hoal, K. O. (2013). Geometallurgy of the pebble porphyry copper-gold-molybdenum deposit, alaska: Implications for gold distribution and paragenesis. *Economic Geology*, 108(3):463–482.

Guardiano, F. B. and Srivastava, R. M. (1993). Multivariate geostatistics: beyond bivariate moments. In *Soares A. (eds) Geostatistics Troia'92*, volume 5 of *Quantitative Geology and Geostatistics*, pages 133–144. Springer, Dordrecht.

Haldorsen, H. H., Lake, L. W., et al. (1984). A new approach to shale management in field-scale models. *Society of Petroleum Engineers Journal*, 24(04):447–457.

Hassanpour, R. M. (2013). *Grid-free Facies Modelling of Inclined Heterolithic Strata in the McMurray Formation*. PhD thesis, University of Alberta, Edmonton, Alberta.

Hosseini, A. H. et al. (2009). *Probabilistic modeling of natural attenuation of petroleum hydrocarbons*. PhD thesis, University of Alberta, Edmonton.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417–441.

Hunt, J. A. and Berry, R. F. (2017). Geological contributions to geometallurgy: A review. *Geoscience Canada*, 44(3):103–118.

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. (1990). *An introduction to applied geostatistics*, volume 1. Oxford University Press.

Journel, A. G. (1974). Geostatistics for conditional simulation of ore bodies. *Economic Geology*, 69(5):673.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15(3):445–468.

Journel, A. G. and Isaaks, E. H. (1984). Conditional indicator simulation: Application to a saskatchewan uranium deposit. *Journal of the International Association for Mathematical Geology*, 16(7):685–718.

Kelley, K. D. and Jennings, S. (2004). A special issue devoted to barite and zn-pb-ag deposits in the red dog district, western brooks range, northern alaska. *Economic Geology*, 99(7):1267–1280.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of Chemical, Metallurgical, and Mining Society of South Africa*.

Krolak, T., Palmer, K., Lacouture, B., and Paley, N. (2017). Ni 43-101 technical report, red dog mine, alaska, usa. Technical report, Teck Alaska Incorporated.

Kyriakidis, P. C., Deutsch, C. V., and Grant, M. L. (1999). Calculation of the normal scores variogram used for truncated gaussian lithofacies simulation: theory and fortran code. *Computers & Geosciences*, 25(2):161 – 169.

Lantuéjoul, C. and Desassis, N. (2012). Simulation of a gaussian random vector: A propagative version of the gibbs sampler. In *The 9th International Geostatistics Congress*, Oslo, Norway.

Larrondo, P. F., Neufeld, C. T., and Deutsch, C. (2003). Varfit: A program for semi-automatic variogram modelling. Technical report, Centre for Computational Geostatistics, University of Alberta.

Leuangthong, O. (2003). *Stepwise Conditional Transformation for Multivariate Geostatistical*

Simulation. PhD thesis, University of Alberta, Edmonton, Alberta.

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

Leuangthong, O., Khan, K. D., and Deutsch, C. V. (2008). *Solved Problems in Geostatistics*. Wiley-Interscience.

Little, R. J. and Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons, 2 edition.

Machuca-Mory, D. and Deutsch, C. (2006). Deriving indicator direct and cross variograms from a normal scores variogram model. Ccg paper 2006-404, Centre for Computational Geostatistics, University of Alberta.

Madani, N. and Emery, X. (2015). Simulation of geo-domains accounting for chronology and contact relationships: application to the río blanco copper deposit. *Stochastic Environmental Research and Risk Assessment*, 29(8):2173–2191.

Mallet, J.-L. L. (2002). *Geomodeling*. Applied Geostatistics. Oxford University Press, Inc., New York, NY, USA, 1 edition.

Manchuk, J. and Deutsch, C. (2011). A program for data transformations and kernel density estimation. Ccg paper 2011-116, Centre for Computational Geostatistics, University of Alberta.

Manchuk, J. and Deutsch, C. (2016). Revisiting the intrinsic cokriging model for integrating secondary. Ccg paper 2016-128, Centre for Computational Geostatistics, University of Alberta.

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

Mariethoz, G., Renard, P., Cornaton, F., and Jaquet, O. (2009). Truncated plurigaussian simulations to characterize aquifer heterogeneity. *Ground Water*, 47(1):13–24.

Marjoribanks, R. (2010). *Diamond Drilling*, pages 99–136. Springer Berlin Heidelberg, Berlin, Heidelberg.

Matheron, G. (1962). *Traité de géostatistique appliquée*, volume 1 of *Memoires du Bureau de Recherches Géologiques et Miniéres, No. 14*. Editions Technip, Paris.

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

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in*

Applied Probability, 5(3):439–468.

Matheron, G. (1989). The internal consistency of models in geostatistics. In Armstrong, M., editor, *Geostatistics*, pages 21–38, Dordrecht. Springer Netherlands.

Matheron, G., Beucher, H., De Fouquet, C., Galli, A., Guerillot, D., Ravenne, C., et al. (1987). Conditional simulation of the geometry of fluvio-deltaic reservoirs. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, Inc.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, Inc., 2 edition.

McLennan, J. A. (2007). *The decision of stationarity*. PhD thesis, University of Alberta, Edmonton, Alberta.

McLennan, J. A. and Deutsch, C. (2006). Implicit boundary modeling (boundsim). Technical report, Centre for Computational Geostatistics, University of Alberta.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Moore, D. W., Young, L. E., Modene, J. S., and Plahuta, J. T. (1986). Geologic setting and genesis of the red dog zinc-lead-silver deposit, western brooks range, alaska. *Economic Geology*, 81(7):1696.

Neufeld, C. and Deutsch, C. (2006). Data integration with non-parametric bayesian updating. Ccg paper 2006-105, Centre for Computational Geostatistics, University of Alberta. Available at http://www.ccgalberta.com/ccgresources/ report08/2006-105-non_parametric_bu.pdf.

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

Pardo-Igúzquiza, 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.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36(6):1389–1401.

Pyrcz, M. J. and Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford University Press, 2 edition.

Ren, W. (2007). Large scale modeling by bayesian updating techniques. Tech-

nical report, Centre for Computational Geostatistics, University of Alberta. Available at http://www.ccgalberta.com/ccgresources/report09/2007-129_bayesian_updating.pdf.

Roberts, W. J. J. (2010). Application of a gaussian, missing-data model to product recommendation. *IEEE Signal Processing Letters*, 17(5):509–512.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472.

Rossi, M. E. and Deutsch, C. V. (2014). *Mineral resource estimation*. Springer Netherlands.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, volume 1, pages 20–34. American Statistical Association.

Sadeghi, S. and Boisvert, J. (2012). Optimizing thresholds in truncated pluri-gaussian simulation. Ccg paper 2012-129, Centre for Computational Geostatistics, University of Alberta.

Scheffel, R., Guzman, A., and Dreier, J. (2016). Development metallurgy guidelines for copper heap leach. *Minerals & Metallurgical Processing*, 33(4).

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Silva, D. and Boisvert, J. B. (2014). Mineral resource classification: a comparison of new and existing techniques. *Journal of the Southern African Institute of Mining and Metallurgy*, 114:265 – 273.

Silva, D. and Deutsch, C. (2015). Transformation for multivariate modeling using gaussian mixtures with exhaustive secondary data. CCG paper 2015-105, Centre for Computational Geostatistics, University of Alberta.

Silva, D. A. (2015). *Enhanced Geologic Modeling with Data-Driven Training Images for Improved Resources and Recoverable Reserves*. PhD thesis, University of Alberta, Edmonton, Alberta.

Silva, D. S., Jewbali, A., Boisvert, J. B., and Deutsch, C. V. (2018). Drillhole placement subject to constraints for improved resource classification. *CIM Journal*, 9(1):21–32.

Sinclair, A. J. and Blackwell, G. H. (2002). *Applied Mineral Inventory Estimation*. Cambridge University Press.

Snowden, D. V., Glacken, I., and Noppé, M. A. (2002). Dealing with demands of technical variability and uncertainty along the mine value chain. In *Value Tracking Symposium*, volume 69, pages 93–100, Brisbane, Australia. Australasian Institute of Mining and Metallurgy.

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons.

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. *Mathematical Geology*, 34(1):1–21.

Switzer, P. and Green, A. A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Technical report no. 6, Department of Statistics, Stanford University.

Tonder, E. V., Deglon, D., and Napier-Munn, T. (2010). The effect of ore blends on the mineral processing of platinum ores. *Minerals Engineering*, 23(8):621 – 626.

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

Xu, C., Dowd, P. A., Mardia, K. V., and Fowell, R. J. (2006). A flexible true plurigaussian code for spatial facies simulations. *Computers & Geosciences*, 32(10):1629 – 1645.

Zagayevskiy, Y. and Deutsch, C. (2015). Numerical derivation of gaussian variogram models for truncated pluri-gaussian simulation. Technical report, Centre for Computational Geostatistics, University of Alberta.

Zagayevskiy, Y. and Deutsch, C. V. (2016). Multivariate geostatistical grid-free simulation of natural phenomena. *Mathematical Geosciences*, 48(8):891–920.