Truncation Trees in Hierarchical Truncated PluriGaussian Simulation

by

Harold Velasquez Sanchez

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

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Geological uncertainty is a major source of risk in resource projects and must be characterized. Extensive research has been undertaken in the development of sophisticated techniques given the economic impact of this uncertainty. The Hierarchical Truncated PluriGaussian (HTPG) simulation is well-known among existing categorical modeling techniques for its ability to portray realistic features. HTPG uses latent Gaussian variables to simulate categories. The rules to map continuous variables to categories, and to introduce juxtaposition constraints are key and known as truncation trees. The number of truncation trees is large, and their structures are very flexible, however, choosing the correct one is a daunting task. This work focuses on developing tools to choose the truncation tree that leads to the optimal model. The process starts with the inference of possible trees from any source of limited data including drillholes, point samples, or images. Then, a tree is chosen based on measures of goodness.

This research has made several contributions. First, it presented tools to enumerate all possible trees in friendly plots given the number of categories. Second, it introduced interval probabilities to quantify the geometrical associations of categories and proposed a dissimilarity matrix based on this concept to summarize the associations and to be used in inference algorithms. Third, it implemented Single Linkage Cluster Analysis (SLCA), and spectral partitioning from graph theory for the inference of trees. Lastly, it showed an optimization framework in a synthetic example that used all trees to recommend transition probabilities as an appropriate measure of goodness. The tools and methodologies were used in a case study where the chosen tree obtained good metrics and respected the geological understanding.

Dedication

To my fiancée Majo

Acknowledgments

I would like to thank my supervisor, Dr. Clayton Deutsch, for his support and expert guidance. Your enthusiasm and endless ideas are inspiring and made this thesis possible.

I would like to acknowledge the Centre for Computational Geostatistics (CCG) affiliates for their financial support. My gratitude is also extended to my colleagues at CCG for their assistance and friendship.

I am indebted to my colleagues in Toquepala and Antamina, their technical and personal advice helped my growth as a professional and encouraged me to pursue a Master's degree. Lastly, I would like to thank my family for always being there for me.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Description
$A_{n \times n}$	Adjacency matrix of $n \times n$ elements
B	Local accuracy parameter
$\mathcal{B}$	Set of categories
$C(h)$	Covariance
C_I	Indicator covariance
$\subseteq$	is a subset of
$E\{\}$	Expected value operator
F_n	Fubini number or ordered Bell number
$\mathbf{h}$	Lag vector
H	Entropy
H_{avg}	Average entropy
$i(u_\alpha, s_k)$	Indicator variable
I	Identity matrix
K	Number of categories
L	Total number of input models
L_G	Laplacian of a graph
nCk	k-combination of a set with n elements
σ	Standard deviation
$\bar{p}$	Expected penalty value
p_k	Probability of category k
δ	Dissimilarity matrix
μ	Mean
u_α	Sample location
$Var\{\}$	Variance operator
ϖ	Bell number
$S(K, p)$	Stirling number
$\lfloor x \rfloor$	Floor function of x
γ_{h}	Variogram
γ_I	Indicator variogram

List of Abbreviations

Abbreviation	Description
2-D	Two-dimensional
3-D	Three-dimensional
DCD	Duluth Complex Database
DHID	Drill Hole ID
ETPE	Embedded Transition Probability Error
ETPM	Embedded Transition Probability Matrix
GSLIB	Geostatistical Software Library
HTPG	Hierarchical Truncated PluriGaussian
IK	Indicator Kriging
MAPS	Maximum a Posteriori Selection
MCC	Matthews Correlation Coefficient
MCS	Monte Carlo Simulation
MDS	Multidimensional Scaling
MSE	Mean Squared Error
MST	Minimum Spanning Tree
OBM	Object-Based Modeling
PE	Prediction Error
RF	Random Function
SGS	Sequential Gaussian Simulation
SGSIM	Sequential Gaussian Simulation GSLIB Program
SIS	Sequential Indicator Simulation
SLCA	Single Linkage Cluster Analysis
TPE	Transition Probability Error
TPGS	Truncated PluriGaussian Simulation
TPM	Transition Probability Matrix

Chapter 1

Introduction

In the resource industry, the characterization of deposits has evolved from deterministic to stochastic models and is essential in mining projects. Sensitive economic evaluation and planning of the extraction methods to forecast resources and reserves rely on input geological and continuous variable models. Extensive geostatistical research has focused on the uncertainty of geologic models as it highly affects engineering designs and economic forecasts. Practitioners must provide an accurate and precise model of the uncertainty of subsurface properties. The quantity of mineral resources is a critical asset and must be clearly reported to investors. Mining projects are capital-intensive, and managing the associated risk has become increasingly important for decision-making. Ideally, the risk would be understood and managed during all stages of an investment.

In resource modeling, categorical variables represent a variety of geological characteristics and are sometimes referred to as lithofacies, (Chiles & Delfiner, 2012); they constitute a major aspect of geological heterogeneity and uncertainty. Lithological geometries are often modeled before modeling of continuous properties, (Journel & Isaaks, 1984) as they subdivide the data to focus the analysis in a partitioned space aiming to portray geological variations and relative geometries with a necessary resolution for engineering purposes, (Journel & Isaaks, 1984; Rossi & Deutsch, 2013).

In cases where numerical models are built without a thorough geological understanding, the validation is limited to the reproduction of input statistics and local data, (Boisvert, 2010; Pyrcz & Deutsch, 2014); the additional geological knowledge permits to check whether the output models show admissible features. In the oil industry, sedimentary and diagenetic processes are the basic mechanisms for the generation of lithofacies. In mining, epigenetic and syngenetic deposits, (Lovering, 1963) are general classifications related to lithological genesis. In both cases the spatial continuity within each domain controls the presence of minerals, outlining the importance of characterizing the uncertainty of categorical variables. Additional geological processes such as tectonics and erosion could be involved making the interpretation and modeling challenging for geologists and geostatisticians. The genetic aspect considered refers to the geometric associations while temporal relations of lithofacies are treated indirectly to some limited extent.

Over the years, diverse techniques have been developed for geostatistical categorical modeling, (Journel & Isaaks, 1984; D. S. F. Silva, 2018). Preference for a particular method depends on the characteristics of the geological settings to be modeled among other factors. In any case, the reproduction of spatial characteristics of the deposit while accounting for the correct uncertainty is desired. Hierarchical Truncated PluriGaussian Simulation (HTPG) is a notable variant of truncated Gaussian techniques that allows the utilization of higher dimensions permitting to model an arbitrary number of categories (D. S. F. Silva, 2018). The hierarchical truncation schema used in HTPG serves as a means to introduce the geological understanding to the model by preserving relevant geometrical associations; however, selecting the correct truncation rule is not clear in most cases, these rules result from vast combinatorics depending on the number of categories. The task becomes puzzling as sparse data is common in the energy and resource industry. The limited amount of available information forces the reliance on geological expertise, while a more data-driven model would be possible with more information. The question lies in selecting a truncation tree in HTPG inferable from interval data such as drillholes or surface sampling, which simultaneously optimize measures of goodness including transition probability errors.

1.1 Problem Motivation

Different techniques for categorical modeling have been developed over the years including Sequential Indicator Simulation (SIS), Object-Based Modeling (OBM), and truncated Gaussian techniques. HTPG is a notable alternative as the hierarchical implementation enables the usage of more complex associations between categories unlike previous variants such as truncated Gaussian and pluri-Gaussian, (Beucher & Renard, 2016; Sadeghi, 2017). In general, the possible geometric relations or associations are summarized in rules of ordering when categories are mapped from the same latent Gaussian variable, and precedence rules whenever the categories or subset of categories belongs to a different Gaussian variable in the hierarchy of the truncation tree. The versatility of HTPG permits to model a variety of geological settings such as non-contact zones, stratigraphic sequences, and geological unconformities. However, due to the newness of the technique, the truncation trees have been incompletely explored. This includes the number and structure of the truncation trees. For a given set of categorical variables, the number of possible truncation rules increases rapidly not only due to the combinatorics of the elements but due to the different structuring within the truncation tree, therefore further work in this area is required.

The geological understanding to perform HTPG may seem trivial when the geological processes involved are fully understood and there is enough information from drilling programs, production, and different sources of validation data; however, the amount of data available in resource projects is insufficient. Geological properties are often sampled very sparsely due to economic reasons and are a small fraction of the volume under study (Chiles & Delfiner, 2012). Geostatisticians integrate different sources of information, e.g. soft and hard data, to generate predictions using a model or data-driven approach. Core samples represent a reliable source for categorical interpretation from the subsurface as well as surface sampling. Core logging and surface mapping are therefore the primary input for categorical analysis. However, there is a lack of practical tools to assist in the detection of geological associations based on cores. The number could be reduced quickly in some cases by visual inspection of the drillholes, nevertheless when the complexity of the patterns or the number of categories increases, more automatic inference methods become useful to reduce the number of plausible truncation trees.

Extensive work has been done in interpreting sequential categorical data from cores, especially in stratigraphic analysis (Carle & Fogg, 1996; Elfeki & Dekking, 2001; Krumbein & Dacey, 1969), and some of the concepts could be extended to tree inference. Recent work has documented tools for the inference of a truncation tree and insights for the selection of a proper tree structure, (J. L. Deutsch & Deutsch, 2013; D. S. F. Silva, 2018). Truncation rules in HTPG are a key aspect that controls the quality of generated models, the selection of a truncation tree depends also on how well the output model performs during validation, therefore, measures of goodness must be recognized and considered during the inference process.

1.2 Thesis Statement and Research Contributions

The selection of a correct truncation tree respects geometric associations of spatial categorical data and benefits from the high dimensionality in the HTPG approach. The research developments are related to the definition of (1) a novel interval-based distance between categories, and (2) a methodology for truncation tree inference and selection. These contributions eased the usage and optimized categorical models in HTPG.

1.2.1 A Measure of Distance

A measure of distance between a pair is a key input for many inference algorithms. It specifies how closely two entities are related (Shepard, 1962). Several other terms are found in literature such as (dis)similarity, closeness, relatedness, and friendliness. Various measures of distance or proximity have been studied in different fields.

In geostatistics, transition probabilities were used to quantify and predict juxtapositional tendencies in stratigraphic sequences (Carle & Carle, 1997; Vistelius, 1949), and to explain fining-upward tendencies and lateral juxtapositions in geological basins. Developments in HTPG used transition probability-based dissimilarities to quantify the relationships between categories (D. S. F. Silva, 2018) to infer a truncation tree, however, this distance depends on stationarity and is sensitive to size/thickness. In general, sedimentary sequences tend to show cyclic behavior but are not stationary; similarly, non-strata-bound mineral deposits are not usually guided by vertical depositions and are strongly non-stationary (Armstrong et al., 2011); a more flexible measure of dissimilarity should be used.

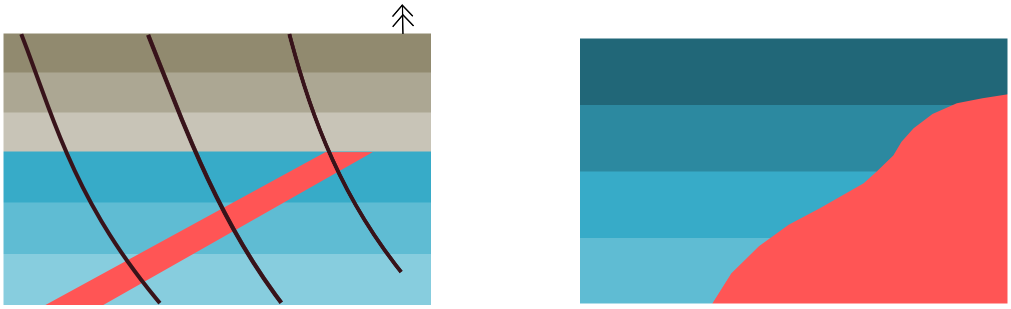
The presented measure of distance is based on interval probabilities and it represents an alternative to interpreting and quantifying distance between categories. In interval probabilities, the variable is the number of intervening intervals between two points in space. Intervals of the same category are not counted, preventing the algorithm from being impacted by repeated patterns along the vector. Moreover, this measure does not assume that both vertical and lateral lithofacies match Walther's Law (Allaby, 2013). It can be applied in more complex geological scenarios. Unlike transition probability-based distances, the number of intervening intervals is a direct measurement of dissimilarity.

1.2.2 Truncation Tree Inference and Selection

The number of possible trees in HTPG is combinatorially large, this makes most optimization approaches unreasonable. Inference algorithms are explored in this thesis to ease the modeling of categories in hierarchical truncated techniques. There is no unique way to infer structures from paired distances, some authors applied Minimum Spanning Tree (MST) on a two-dimensional projection of the distances between categories (D. S. F. Silva, 2018), however, using reduced dimensions may sometimes fail to show hierarchical relations and associations of categories correctly, therefore truncation trees should preserve these hierarchies.

The human ability to order objects in images is biased. Superimposed or cross-cutting objects are interpreted as younger or closer and put at the top of the hierarchy while objects that appear behind or covered are placed at the bottom. Figure 1.1 shows two cases of unconformities, in Figure 1.1a where the vertical sequencing of geological layers governs the temporal aspect and the cross-cutting unit appears younger than the bottom three horizontal layers, however, this naive inference is not always true and setting up automatic algorithms with the premise of cross-cutting units or upper layers being younger may result in incorrect automatic interpretations. For instance, in Figure 1.1b the discordant shape (red) is typically an igneous rock, it could be younger if the intrusion cuts the layers or older if they formed before the sedimentary sequences.

The hierarchies of the categories in the truncation tree may be set differently depending on the geological information. Inference algorithms lack enough information to consider these scenarios, therefore the goal of the tools presented here is to understand the contact relations and associations within an image and to establish hierarchies that would aid with the geological modeling in HTPG. This work introduces the concept of graphs by translating the distances into adjacency values between nodes. A graph contains information on the relationship between categories. When the graph is partitioned properly, subsets of the categories with closer relationships are obtained and a hierarchical structure could be derived. Ideally, the inferred tree would optimize the results of HTPG models, however, additional variables should be considered such as the performance of numerical derivation given a truncation tree, anisotropies of categories, and global proportions among others. Finally, once the set of possible trees is reduced with inference techniques, their structures and ordering of the latent Gaussian variables in the hierarchy can be reconsidered to optimize pre-defined metrics of the output model while preserving a maximum of geological knowledge.



(a)

(b)

Figure 1.1: Two schematic geological cross sections with unconformities. (a) Vertical sequence of layers. The crossing unit (red) cuts the bottom three horizontal layers. Solid black lines represent drillhole traces. (b) Nonconformity. Irregular geometry colored in red represents an igneous rock.

1.3 Thesis Outline

Chapter 2 provides background concepts; the first section presents a summary on categorical variables and the importance of categorical modeling. The second section discusses stationarity and trends within the context of categorical variables. The third section presents the indicator formalism and details on basic concepts of two common geostatistical techniques for categorical modeling.

Chapter 3 introduces general concepts of hierarchical arrangements of data. The first section details hierarchical structures and their relation to truncation trees in HTPG. A simplified notation to arrange categories in trees is explained. Further classification as binary and non-binary structures is discussed. A summary of useful concepts from number theory is reviewed as a means to enumerate and list the truncation trees for a given set of categories. Section two presents a truncation tree graph and describes its components. A partitioning schema blended with simplified notation is presented to calculate the number of trees. Additionally, categorical models generated with HTPG with their respective truncation trees are depicted as examples.

Chapter 4 summarizes current tools for the inference of truncation trees and proposes new algorithms. The first section describes previous measures of distance used in HTPG. The second section presents a novel distance measure based on interval probabilities. The third section explores different tools and algorithms to assist in the detection of truncation trees and develops an explanatory case, these tools include Single Linkage Cluster Analysis (SLCA), Minimum Spanning Tree (MST), and spectral partitioning.

Chapter 5 explores the influence of the truncation tree for HTPG to determine a measure of optimality that validates the inferred trees. The first section evaluates the numerical derivation approach with different tree structures. The second section describes a synthetic workflow used to assess different trees, the results with different measures of goodness considered in the process are reviewed in the next section. The fourth section explains relevant metrics for tree selection in HTPG useful for practical applications.

Chapter 6 presents a case study of categorical modeling with HTPG for the Mesaba Deposit dataset. The framework for tree inference and selection of an optimal tree is applied and the results are discussed.

Chapter 2

Literature Review

This thesis is focused on exploring the truncation trees in Hierarchical Truncated PluriGaussian (HTPG) simulation. Basic literature about categorical modeling and relevant geostatistical techniques are presented.

2.1 Categorical Variables

In geostatistics, categorical variables represent lithofacies or domains of interest for subsequent evaluation of continuous properties. Categorical domains provide a set of alternative scenarios that mimic the actual distributions of domains in space (Chiles & Delfiner, 2012). The probability distribution of categorical variables is defined by the proportion of each category p_k, $k = 1, \ldots, K$. where K is the number of categories, (Rossi & Deutsch, 2013).

2.2 Stationarity and Trend

Stationarity refers to the decision of pooling data together for further geostatistical evaluation. It is not a hypothesis, therefore it cannot be tested (Pyrcz & Deutsch, 2014). The decision of stationarity can be revisited once the geostatistical analysis has started or when more data is available. In categorical variables, the modeling of a trend gains significant importance given that in earth sciences categorical variables are almost always non-stationary. The trend is represented by a spatial model of proportions. In HTPG, the trend controls the local proportions of categories. An over-fit or under-fit trend leads to a wrong assessment of the categorical uncertainty, (Harding & Deutsch, 2021).

2.3 Modeling Categorical Variables

Practitioners often model categorical variables and continuous variables. A simple approach considers signed distance functions with posterior correction for global proportion reproduction, (D. A. Silva & Deutsch, 2013), however, it does not account for uncertainty and small-scale variability. Categorical domains comprise one of the major sources of uncertainty within a geostatistical workflow. Extensive research has been dedicated to improving categorical modeling, (Lajevardi, 2015). Characterizing the uncertainty of geological models is standard practice and several techniques are employed including object-based models (Lantuejoul 2002), SIS (Journel & Alabert, 1990), Multiple Point Statistics (Strebelle, 2002), truncated Gaussian simulation, and pluriGaussian simulation (Armstrong et al., 2011; Matheron et al., 1987).

2.3.1 Indicator Formalism

The indicator formalism allows interpretative information to be translated into binary codes for the application of numerical techniques (Journel & Alabert, 1990). The discrete random function is defined on K mutually exclusive categories $s_k, k = 1, \ldots, K$ within a domain $\mathcal{A}$. In the indicator transformation, Equation (2.1), $i\left(\mathbf{u}_\alpha; s_k\right)$ is the binary indicator value at location $\mathbf{u}_\alpha$ and for category s_k. An additional condition of exhaustivity is required in indicators, any location $\mathbf{u}_\alpha$ belongs to one of the K categories:

$$i\left(\mathbf{u}_\alpha; s_k\right) = \begin{cases} 1, \text{ if } \mathbf{u}_\alpha \in s_k \\ 0, \text{ otherwise} \end{cases}, \quad k = 1, \ldots, K \tag{2.1}$$

Although indicators take values of 0 or 1, the estimates are continuous probabilities, that is, the probability of category s_k at location $\mathbf{u}_\alpha$ is $p_k \in [0, 1]$ and sum up to 1, Equation (2.2). The stationary mean of the binary indicator random function $I(\mathbf{u}; s_k)$ and its stationary variance for category s_k within $\mathcal{A}$ are defined in Equation (2.3) and Equation (2.4).

$$\sum_{k=1}^{K} p_k = 1 \tag{2.2}$$

$$E\{I(\mathbf{u}; s_k)\} = p_k \tag{2.3}$$

$$\text{Var}\{I(\mathbf{u}; s_k)\} = p_k\left(1 - p_k\right) \tag{2.4}$$

Indicators are applied in deterministic and stochastic methods. In estimation mode, indicator kriging (IK) gives the estimated probability. In simulation, SIS allows the evaluation of global and local uncertainty, (Hassanpour, 2007). Stationary indicator covariances and variograms for the category s_k separated by a lag vector $\mathbf{h}$ are calculated following Equation (2.5) or Equation (2.6), where p_k is the stationary proportion of category s_k.

$$C_I(\mathbf{h}; s_k) = E\{I(\mathbf{u}; s_k)I(\mathbf{u} + \mathbf{h}; s_k)\} - p_k^2 \tag{2.5}$$

$$\gamma_I(\mathbf{h}; s_k) = E\left\{[I(\mathbf{u}; s_k) - I(\mathbf{u} + \mathbf{h}; s_k)]^2\right\} \tag{2.6}$$

Assuming stationarity, the indicator variogram and covariance are related by:

$$\gamma_I(\mathbf{h}; s_k) = C_I(0; s_k) - C_I(\mathbf{h}; s_k) \tag{2.7}$$

Some restrictions may arise with covariances and variogram compatibility with particular distributions. For instance, indicator random functions are not compatible with Gaussian variogram models (Armstrong, 1992; Christakos, 1984) even when the condition of positive definiteness is met.

2.3.2 Sequential Indicator Simulation

Sequential Indicator Simulation (SIS) is a common technique to model the uncertainty of either continuous or categorical random variables. Simulation of indicators is based on the indicator kriging formalism, (C. V. Deutsch & Journel, 1998; Journel & Gomez-Hernandez, 1993). Simulated models with SIS reproduce patterns of continuity, values at data location (Suro-Perez & Journel, 1990), and global proportions of categories. Conditional distributions in SIS are estimated non-parametrically to allow the modeling and representation of patterns of non-Gaussian random functions such as categorical or continuous variables with correlated high values (Emery, 2004; Journel & Isaaks, 1984). SIS does not offer explicit geological controls leading to unrealistic transitions between lithofacies. It is usual practice to consider a cleaning algorithm such as MAPS, (C. V. Deutsch, 2005) within an SIS workflow to correct undesired short-scale variations and possible deviations from the global proportion statistics derived from a non-representative sampling. Geological variations and contact relations are better handled in truncated Gaussian techniques, object-based and process-based models.

2.3.3 Hierarchical Truncated pluriGaussian

The truncated Gaussian method was first introduced in Matheron et al. (1987) and naturally extended to pluriGaussian space (Galli, H, G, & B, 1994). The basic idea in truncated Gaussian techniques is to consider one or more Gaussian variables that are truncated to yield categories (Xu, Dowd, Mardia, & Fowell, 2006). The thresholds for the associated Gaussian variables used in the anamorphosis are deduced from the proportions of each facies (Matheron et al., 1987) and are included in the truncation rules.

The hierarchical truncated pluriGaussian method, (D. S. F. Silva, 2018) is a cell-based method for categorical modeling. HTPG can be seen as a generalization of the binary tree structure proposed in Madani and Emery (2015). HTPG is not restricted to the utilization of a determined number of Gaussian variables, therefore it allows the usage of more complex tree structures. These truncation rules define the mapping to the categorical space.

2.3.3.1 Truncation Tree

A key parameter in truncation techniques is the truncation rule, referred to as the truncation tree in this work. These rules have been depicted differently in the past, including matrix-like notations, (Xu et al., 2006) or binary tree-like structures, (Madani & Emery, 2015), however, they showed limitations in describing certain geological settings, especially with a large number of categories and complex contact relations. For instance, the use of 2-D or 3-D drawings to show the truncation rules where the axes represent Gaussian variables fails to generalize with a higher number of Gaussian variables and complex scenarios, Figure 2.1. Moreover, the possible contacts increase exponentially with dimensionality (Xu et al., 2006). Currently, the hierarchical approach eases the utilization of a broad set of truncation rules by using binary and non-binary tree-like structures leaving aside the representation based on orthogonal axes.

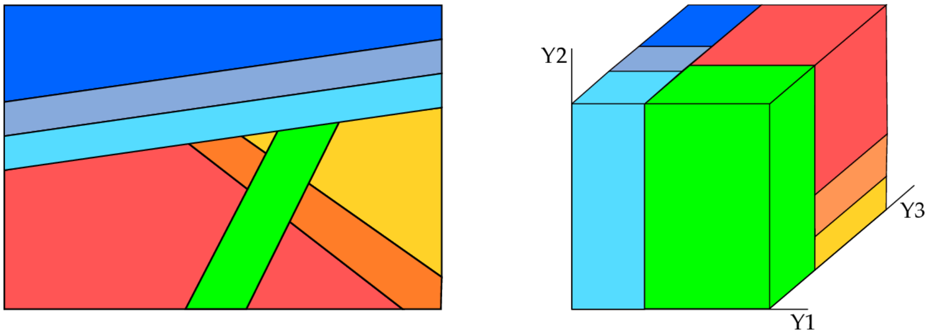


Figure 2.1: (Left) Sketch of two-layered structures separated by an erosional surface. Layers at the bottom are cut by an intrusion (darkest gray). (Right) Truncation rule representation based on orthogonal axes for the Gaussian variables. Y3 controls the transition between the upper layers, and the transition between the bottom layers and the intrusion, however, both cases have different spatial structure. This truncation rule does not represent the geological setting correctly. Taken from (D. S. F. Silva, 2018)

The truncation thresholds are an important aspect of the truncation trees as they control the proportion of categories in the output models. For the set of categories $\mathcal{B} = \{s_k, k = 1 \dots K\}$, the number of thresholds is $K - 1$. Some methods partition the Gaussian space using orthogonal thresholds, (Armstrong et al., 2011; Emery, 2007), and others access to multivariate Gaussian space, (J. L. Deutsch & Deutsch, 2014). In HTPG, the thresholds are calculated utilizing the proportions of categories relevant to each node following Equation (2.8).

$$t_j = \Phi^{-1}\left(\frac{\sum\limits_{k\in\mathcal{B}'_j} p_k}{\sum\limits_{k\in\mathcal{B}'} p_k}\right), \quad j = 1, \dots, K-1 \tag{2.8}$$

where Φ^{-1} is the inverse of the cumulative distribution function of a standard normal distribution. $\mathcal{B}' \subseteq \mathcal{B}$ with relevant categories; and $\mathcal{B}'_j \subseteq \mathcal{B}$ are the relevant categories below the threshold t_j.

2.3.3.2 Non-Stationarity

Stationarity in categorical variables involves having similar proportions over a domain. These categorical proportions specify the thresholds in the truncation rule. In HTPG, the non-stationarity is handled by simulating a stationary Gaussian Random Function and using locally varying thresholds depending on local proportions.

2.3.3.3 Numerical Derivation

Categories are assigned by truncating simulated latent Gaussian variables, therefore the correct spatial structure of the Gaussian variables is important. Continuity models of the latent Gaussian variables are unknown given that the latter is not measured. The variogram models used during the simulation of latent Gaussian variables must output categorical models with reproduced continuity. Zagayevskiy and Deutsch (2015) presented an algorithm to optimize latent variogram values independently at different lags. In HTPG, latent Gaussian variables are independent standard Gaussian, therefore the correlation between two points separated by the lag vector $\mathbf{h}$ is established by Equation (2.9). Correlated pairs $y_{i,j}(0)$ and $y_{i,j}(\mathbf{h})$ is obtained using Equation (2.10), where $z_{\mathrm{A}_{i,j}}$ and $z_{\mathrm{B}_{i,j}}$ are m Monte Carlo Simulations (MCS) from the standard Gaussian distribution and ℓ is the number of latent Gaussian variables in the truncation tree.

$$\rho_i(\mathbf{h}) = 1 - \gamma_i(\mathbf{h}), \quad i = 1, \dots, \ell \tag{2.9}$$

$$\begin{aligned} & y_{i,j}(0) = z_{\mathrm{A}_{i,j}} \\ & y_{i,j}(\mathbf{h}) = \rho_i(\boldsymbol{h}) \times z_{\mathrm{A}_{i,j}} + \sqrt{1 - \rho_i(\boldsymbol{h})^2} \times z_{\mathrm{B}_{i,j}}, \quad i = 1, \dots, \ell; \quad j = 1, \dots, m \end{aligned} \tag{2.10}$$

The method uses an objective function to minimize the overall mismatch between reference models of the indicator variograms, $\gamma_I(\mathbf{h}; s_k)$; and indicator variograms obtained from MCS, $\hat{\gamma}_I(\mathbf{h}; s_k)$ for $k = 1, \dots K$, Equation (2.11). The term ℓ represents the number of latent Gaussian variables in the truncation tree, $\rho_i(\mathbf{h})$ is the correlation from Equation (2.9), p_k is the global proportion of category s_k, and w_k is the relevance given to the reproduction of its indicator variogram.

$$\mathrm{O}\left(\rho_i(\mathbf{h}), \quad i = 1, \dots, \ell\right) = \sum_{k=1}^{K} \frac{w_k}{p_k \left(1 - p_k\right)} \left[\gamma_I(\mathbf{h}; s_k) - \hat{\gamma}_I(\mathbf{h}; s_k)\right]^2 \tag{2.11}$$

The numerical derivation approach is a straightforward and flexible implementation to obtain the continuity in the Gaussian space for different truncation trees and any number of latent Gaussian variables. The numerically derived Gaussian variogram points must be correctly fitted with valid models, otherwise, the final result is compromised. However, depending on the input parameters for numerical derivation such as the reference variogram, categorical proportions, and truncation tree, it may not be possible to obtain a proper fit due to hyper-continuity issues increasing the error between the modeled variogram and the optimized points increases, (D. S. F. Silva, 2018).

Chapter 3

Truncation Trees

Trees in HTPG explain associations between categories. Trees consist of a structure, ordering of the categories within the tree, and the threshold values conform the truncation tree. Extensive research on the methodology for the truncation of latent Gaussian variables in truncated techniques resulted in multiple approaches, (J. L. Deutsch & Deutsch, 2014; Emery, 2007; D. S. F. Silva, 2018; Xu et al., 2006). The ease of communicating the associations between categories should be considered to provide an intuitive tool. The purpose of this work is to more deeply understand truncation trees including the number and structures. Tree-like structures facilitate the visualization and truncation of complex contact relations; however, the number of truncation trees for a set of categories is large and increases logarithmically. The relation between binary and non-binary structures and truncation trees is explored. An intuitive truncation tree plot is presented. Concepts and expressions are provided to list and enumerate trees.

3.1 Hierarchical Structures

The number of latent Gaussian variables to use in HTPG's trees ranges from 1 to $K-1$, where K is the number of categories. The mapping of Gaussian values to categories is governed by thresholds and rules in the tree. Former representations of truncation rules used 2-D or 3-D graphs (Madani & Emery, 2015) where axes represented the latent Gaussian variables; however, it failed in describing complex geological settings, (D. S. F. Silva, 2018). In practice, a higher number of latent Gaussian variables with a complex tree structure is often required and HTPG overcomes previous limitations.

Hierarchical arrangements can be successfully described in trees from graph theory, (Murty & Bondy, 2008). These trees are composed of a parent node sequentially linked to sub-trees or child nodes. In a hierarchical structure, a non-bifurcated node is a leaf. Trees can be classified as binary and non-binary. In a binary tree, a parent node presents zero or two child nodes; in non-binary trees, the parent node presents at least one node with more than two child nodes. Here, this classification is used for the calculations in enumerating and listing possible truncation trees in HTPG. The hierarchical structure used in HTPG depends on the complexity of the associations of categories.

3.1.1 Notation

A simplified notation is used to develop algorithms for enumerating the trees. A way to encode trees and forests (arrangements of a set of trees) uses a sequence of nested parentheses (Knuth, 2013b). For instance, the notation in Figure 3.1 represents the tree structure of the binary tree in Figure 3.2. The numbers inside the parentheses express the cardinality of each leaf node, that is, the digit 2 (starting from the left) indicates the number of categories in that leaf. A notation with a valid hierarchical structure has at least one sequence of ′(   )′. The sequence ′(   )′ represents leaf nodes. There are five leaf nodes in Figure 3.1. Adding up the numbers from the leaves according to the sequence of parentheses decodes the truncation tree structure. The path around the periphery of the tree in Figure 3.2 starting from the leftmost side that puts a ′(′ when a node's left edge is visited and ′)′ for a node's right edge is equivalent to the simplified notation.

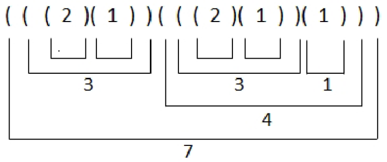


Figure 3.1: Simplified notation of a binary tree.

Hierarchical structures describe the associations but not the ordering of categories within nodes and the hierarchical ordering of the latent Gaussian variables necessary in truncation trees.

3.1.2 Binary Structures

Binary tree structures in HTPG follows proper binary tree concept and are used to represent simple contact rules. A proper binary tree is either a single node or a tree where the initial single root node has two child sub-trees that are proper binary trees. A leaf contains up to K categories. Each node in a binary tree shows the number of categories held. Nodes with more than one category represent a latent Gaussian variable. For instance, there are six latent Gaussian variables in Figure 3.2.

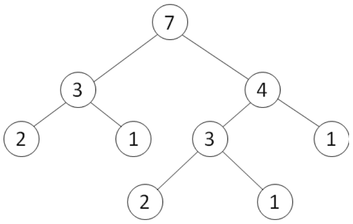


Figure 3.2: A binary tree structure for seven categories. Each parent node is linked to zero or two child nodes.

The number of first level child-node pairs for a parent node with K categories is $\lfloor K \times 0.5 \rfloor$ without considering symmetries. The total number of structures with multiple tree levels requires recursion.

3.1.3 Non-Binary Structures

Binary trees fail to describe complex contact relations as they are limited to one threshold per Gaussian variable in the respective truncation tree. Multiple thresholds in non-leafs enable non-contact rules between categories, Figure 3.3. The ordering of the child nodes or parts in this type of structure is relevant unlike binary trees. The simplified notation encodes also non-binary structures in HTPG. The tree in Figure 3.3 is equivalent to (((1)(3))(1)(2)).

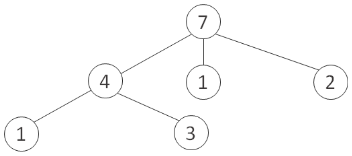


Figure 3.3: A non-binary tree structure for seven categories. A parent node with three child nodes means to use two thresholds in a latent Gaussian variable.

3.1.3.1 Partitions and Complex Tree Structures

Integer partitions or partitions of integer numbers are referred to as partitions to differentiate them from set partitions. A partition of K is a sequence of positive integers, $b_1 \geq b_2 \geq, \cdots, \geq b_i$ named parts that add up K. For instance, partitions of five with three parts are $b_1 = b_2 = 2, b_3 = 1$ and represented by *221*. Seven can be partitioned into two parts *43*, *52*, *61*, or 3 parts *421*, *511*, *331*. In partitions, the order of the parts does not matter, however, they are presented in lexicographic order for convenience. One threshold in a latent Gaussian variable means a two-part partition of the corresponding mapped categories, Figure 3.4. The number of partitions of a positive integer number with fixed p parts is expressed recursively in Equation (3.1). The expression $\begin{vmatrix} K+p \\ p \end{vmatrix}$ represents the number of partitions of K into a maximum of p parts, (Knuth, 2013a). A summary table with the number of partitions for up to ten categories with parts ranging from 1 to 10 is presented in Velasquez and Deutsch (2021).

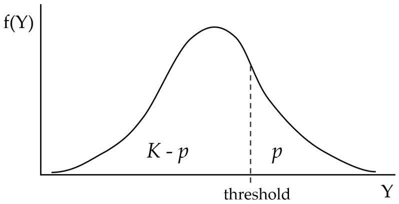


Figure 3.4: Truncation of a latent Gaussian variable by a threshold value.

$$\begin{vmatrix} K \\ p \end{vmatrix} = \begin{vmatrix} K-1 \\ p-1 \end{vmatrix} + \begin{vmatrix} K-p \\ p \end{vmatrix} \tag{3.1}$$

The number of child structures generated from a parent node is defined by the partitions of the number of categories in the node. Recursion is required to count all possible structures from a parent node, however the ordering of the parts of a partition are yet to be specified.

3.1.3.2 Set Partitions and Bell Numbers

The set partitions of a set S are a collection of non-empty blocks or subsets A_i. $A_i \subseteq$ S where, $1 \leq i \leq k$ such that $\cup_{i=1}^{k} A_i$ =S, and $A_i \cap A_j = \emptyset$ for $i \neq j$, (Mansour, 2013). Set partitions of three elements are $\{\{p\}, \{q\}, \{r\}\}, \{\{p, q\}\{r\}\}$, $\{\{p, r\}, \{q\}\}, \{\{q, r\}, \{p\}\}, \{\{p, q, r\}\}$. The order of the blocks and the elements in a block are not considered. Enumerating the number of set partitions does not show a simple closed form. Bell numbers represent the number of set partitions and are denoted as ϖ_K, for instance, $\varpi_3 = 5$, (Belbachir, Djemmada, & Németh, 2021). For efficient computation, the values in the first column of the Peirce's Triangle, (Peirce, 1880) give the Bell numbers and $T_{K,K} = \varpi_K$, e.g. $T_{10,10} = \varpi_{10} = 115,975$, (Sloane, 2022a). Bell numbers can be expressed in terms of second kind Stirling numbers, Equation (3.2). $\varpi_K = \sum_{p=1}^{K} S(K, p)$ where p is the number of parts. In HTPG, when a latent Gaussian variable presents multiple thresholds as in Figure 3.3, the order of blocks controls the contact zones. The number of ordered set partitions of a set S of K elements is known as the ordered Bell number, F_K, (Belbachir et al., 2021). For instance, $F_{10} = 102,247,563$, (Sloane, 2022b). Ordered Bell numbers are helpful to estimate the number of possible structures in HTPG.

$$S(K, p) = S(K-1, p-1) + p \times S(K-1, p), \quad 1 \leq p \leq K \tag{3.2}$$

$$F_K = \sum_{p=0}^{K} p! \times S(K, p) \tag{3.3}$$

The number of truncation structures obtained from K categories is presented in Figure 3.5, the results for six and seven categories are approximations. Most of these structures are binary and many others are symmetric; however, this allows arbitrariness in the specification of truncation structures.

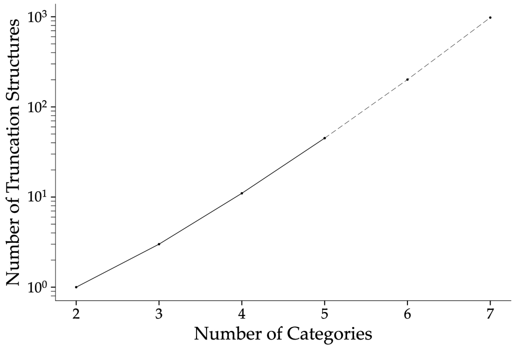


Figure 3.5: Number of truncation structures. Symmetric structures are included. The dashed line indicates approximated values.

3.2 Truncation Trees

Truncation trees are responsible for establishing geometrical relationships between categories. Determining a truncation tree may be one of the most tedious steps in truncated Gaussian techniques (Zagayevskiy & Deutsch, 2015). A truncation tree plot is presented for binary and non-binary structures, Figure 3.6. In the plot, each level represents a latent Gaussian variable. For K categories, there are from 1 to $K-1$ latent Gaussian variables. The number of thresholds is $K-1$. Thresholds are represented by vertical bars and numbered from left to right starting from the first latent Gaussian variable downwards, here the threshold's indexes are omitted. The tree from Figure 3.7 is equivalent to the binary tree in Figure 3.2.

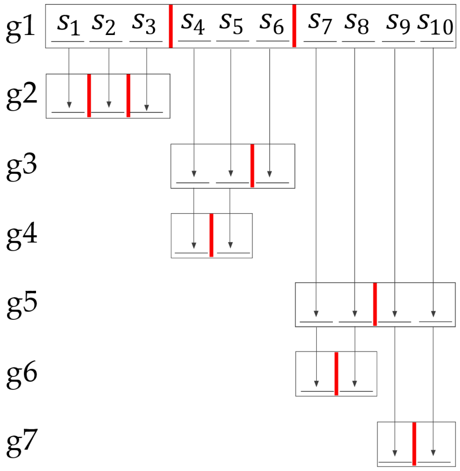


Figure 3.6: Schematic diagram of a truncation tree. Vertical red bars indicate the position of the thresholds. Arrows represent how the categories s_k are hierarchically allocated in Gaussian variables.

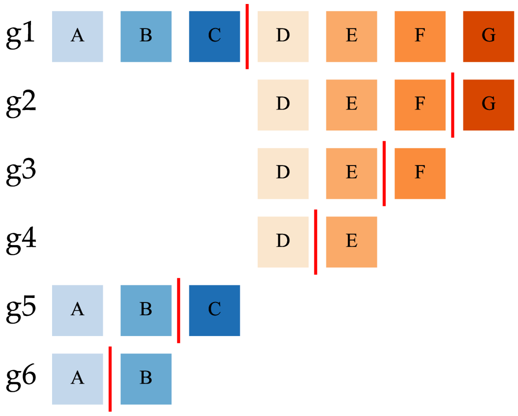


Figure 3.7: A binary truncation tree example.

Using multiple thresholds in non-leaf node further reduce the number of latent Gaussian variables, should be justified by the associations and anisotropies of categories. Figure 3.8 is a truncation tree with a non-binary structure equivalent to the structure in Figure 3.3. Category labels are assigned in the plot as a reference. The Gaussian variable Y_1 is truncated by thresholds one and two and has three parts *412*. Y_2 is truncated by threshold three, the partition is *13*; Y_3 is truncated by threshold four and five and the partition is *111*; Y_4 is truncated by threshold six and the partition is *11*. Reordering the categories changes the tree. The hierarchical truncation of the Gaussian variables of the example in Figure 3.8 is shown in Figure 3.9.

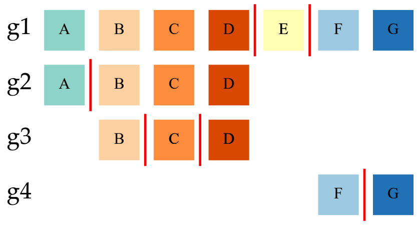


Figure 3.8: A non-binary truncation tree example with seven categories. Category E establishes a non-contact rule between categories A, B, C, D, and F, G.

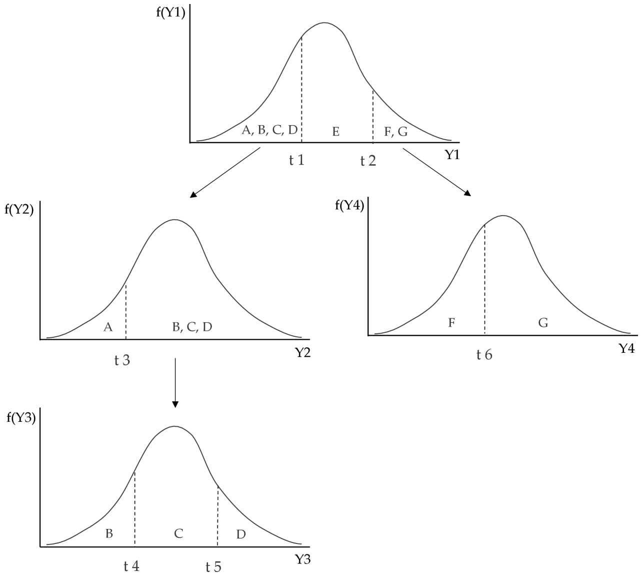


Figure 3.9: Schematic hierarchical truncation of latent Gaussian variables in a non-binary structure. Thresholds 1 and 2 establish a non-contact rule.

Tree structures show the arrangements of categories without specifying the ordering within each block. Allocating the categories into the tree structures is a simple way to obtain truncation trees. Figure 3.10 shows the partitioning of K categories into p_i parts. The number of ways that (A) may be done equals $K!$ in (B). This allows the generation of trees for a specific structure. First, the simplified notation $(((p_1)(p_2))...(p_n))$ of a tree structure is used in (A), then the categories from leaf nodes are placed from left to right in (B). The number of truncation trees is then $R_K \times K!$ where R_K is the number of tree structures for K categories. Figure 3.11 shows the number of truncation trees based on the number of categories. Summaries of truncation structures in 2-D sketches were presented in Armstrong et al. (2011); Emery (2007); however, the position of the categories in the truncation rules was assumed. In simple cases, the ordering is not a concern, nevertheless, when multiple thresholds are required in the latent Gaussian variables, establishing the correct associations of categories for HTPG requires careful analysis.

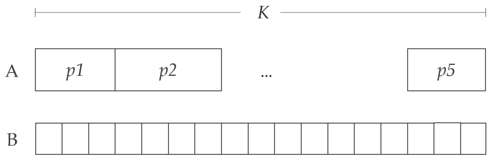


Figure 3.10: Schema of partitioning to calculate the number of trees. K is the number of categories. In A, the parts p_i are positioned. In B, individual categories are positioned.

A general view on the number of truncation trees does not make assumptions on the final ordering as it must be derived from the data. At his stage, symmetrical structures are considered for three reasons (1) it maintains the flexibility in which practitioners choose their truncation tree (2) it is independent of the algorithm for deriving the variograms of the latent Gaussian variables (3) it does not restrict inference algorithms to output specific structures. With more than six categories, the graph in Figure 3.11 shows a projection of the number of trees. The level of recursion explodes in presence of multiple thresholds at non-leaf nodes. With seven categories, the number of possible truncation trees reaches seven orders of magnitude.

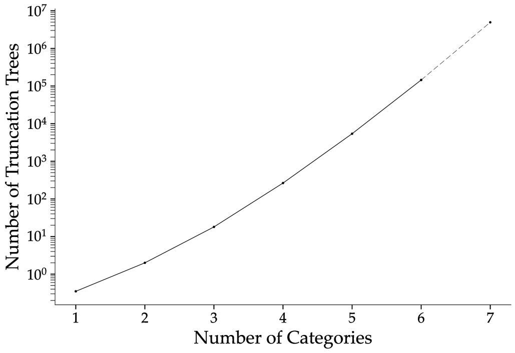


Figure 3.11: Estimated number of truncation trees. Structures and ordering of categories are considered. Solid lines are total values and dashed lines represent approximated values.

3.3 Conclusions

Ordered Bell numbers are used recursively to numerate structures including rules with physical separation. Some categorical HTPG realizations with their truncation tree are shown in Figure 3.12. In non-binary trees, all structures and ordering of categories were considered. A summary of possible tree structures is done by performing the partitions of an initial number of categories and structuring them with nested parentheses notation. In HTPG, characterizing all possible truncation trees allows to evaluate a broad set of associations and patterns between categories. The number of truncation trees with four categories is 264 and 5400 with five categories; however, most are symmetric

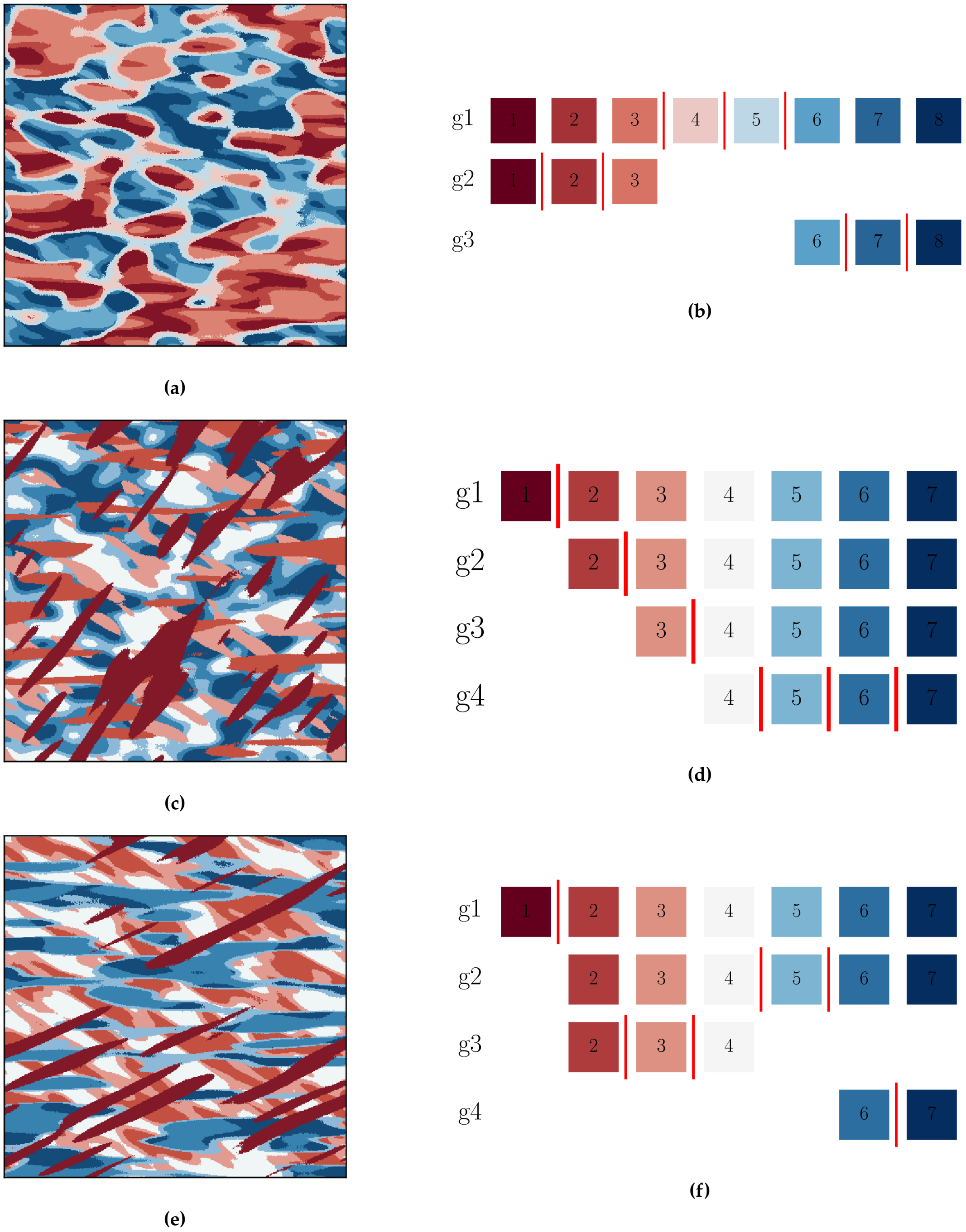


Figure 3.12: Examples of categorical images (left) from HTPG and their respective truncation trees (right). Categories are associated with specific colors.

yielding similar results in HTPG. A first step in describing truncation trees and the number of possible trees has been presented. The next question is how to decrease this number. Finally, HTPG relies strongly on truncation trees, therefore obtaining the optimal or, at least, the geologically reasonable tree is central.

Chapter 4

Truncation Tree Inference

Inferring hierarchical associations using distances between entities or categories has been long studied. Such applications are related to phylogenetic tree inference in biological sciences, (Gascuel, 1997; Makarenkov, 2001; Sattath & Tversky, 1977). Others include graph partitioning techniques for hierarchical image segmentation, (Bourmaud, Mégret, Giremus, & Berthoumieu, 2014). Regardless of the approach, the performance of the inference depends on the measure of dissimilarity. The distance between two categories is expressed as similarity or dissimilarity. Similarities range from zero to one, where one means completely similar. Dissimilarities vary from zero to infinite, where zero is complete similarity.

Truncation trees in control geometric relations in HTPG models. Associations of categories may be obtained from visual inspection of the data, however, practical categorical information in the resource industry is challenging. Concerning HTPG, evaluating possible trees and modeling parameters is time-demanding. The set of trees for K categories should be reduced first. Transition probabilities-based distances have been used to interpret associations of categories and choose a truncation tree, (J. L. Deutsch & Deutsch, 2014; D. S. F. Silva, 2018). This chapter reviews previous work and introduces an interval probabilities and a derived distance matrix. Tools for tree inference and examples and presented.

4.1 Background

Previous work used transition probabilities and Multidimensional Scaling (MDS) in a data-driven approach for the truncation of latent Gaussian variables, (J. L. Deutsch & Deutsch, 2014; Sadeghi & Boisvert, 2012). A transition probability matrix (TPM) uses drillhole data composited at a fixed length $\mathbf{h}$. The elements in a TPM are probabilities of transitioning from category k to k', Equation (4.1). Larger probabilities are related to the closeness between two categories. TPMs are not always symmetric. In practice, TPMs are calculated looking up and looking down a drillhole. Direct interpretation of the TPM is cumbersome when more categories are considered.

$$t_{\mathrm{k,k'}}(\mathbf{h}) = \mathrm{Prob}\left\{ \begin{array}{l} Z(\mathbf{u}) \in \text{ category k} \\ Z(\mathbf{u}+\mathbf{h}) \in \text{ category k}' \end{array} \right\} \tag{4.1}$$

Between-category distances can be obtained from TPM and are input for MDS. to describe n-D data in 2-D or 3-D (Duda, Hart, & Stork, 1973). MDS attaches coordinates to the categories, so the distances between them correspond to experimental dissimilarities, (J. B. Kruskal, 1964). Sometimes these lower-dimension approximations distort high dimensional proximities, (Gower et al., 2016).

An apparent lack of correspondence may be seen between 2-D configurations from MDS and the clustering results due to other dimensions (B. J. B. Kruskal & Wish, 2011).

A recent alternative combines transition probability-related distances, MDS, and MSTs, (Prim, 1957) to infer a truncation tree in a non-automatic fashion (D. S. F. Silva, 2018). It uses an Embedded Transition Probability Matrix (ETPM), (Krumbein & Dacey, 1969) to record state transitions only, then a map built with MDS/MST helps to visualize the association of categories. In this chapter, an alternative measure of dissimilarity based on interval probabilities is developed to analyze the associations of categories.

4.2 Interval Probabilities and a Novel Dissimilarity

Distances between spatial objects were studied in Goldfarb (1985); dissimilarity measures were introduced in J. B. Kruskal (1964); Shepard (1962). They enable the usage of non-metric, indefinite, or non-symmetric distances (Duin & Elżbieta Pe◎kalska, 2012). Dissimilarities allow a straightforward implementation, unlike metric distances that must follow metric space requisites. For instance, the triangle inequality is not obeyed in the touching distance between three objects. In general, distances between objects will be non-metric if the objects under study are not points in a vector space but have a size and shape. Dissimilarities are in general square hollow matrices with pairwise distances between elements, metric or not.

Finding the correct associations of entities in spatial data is a daunting task. Detecting patterns is a human ability learned at an early stage and constantly updated with experience. These observations are primarily triggered by recognizing differences (Edelman, 1999). Even medical doctors struggle to outline algorithms for heart disease detection in an ECG; experts gain consciousness of their recognition process when they are required to outline it such that it can be programmed (Duin & Elżbieta Pe◎kalska, 2012). Similarly, geologists use varied information and methodologies to validate the genesis of complex geological settings. Detecting correct associations of categories is difficult and the data is limited in the resource industry. Yet, practical information related to hierarchical associations of categories can be retrieved from drillholes and sparse sampling. 2-D images are more insightful but less accessible. A novel measure of dissimilarity based on interval probabilities is proposed. This dissimilarity detects geometric associations between categories and can be used for the inference of truncation trees in using limited data.

4.2.1 Intervening Intervals

Rocks are fundamentally described based on lithologic intervals. This is a lithofacies in the core with consistent mineralogy and bounded by different lithofacies. In practice, intervals do not have an inherent genetic connotation as they are sequentially identified as drilling proceeds, however, they represent an alternative to encode geometric relations and allow to establish hierarchies between categories. One way to analyze intervals is to consider the intervening intervals between two points in space. The use of intervening intervals can be extended to images where they are retrieved from vector traces.

Consider the categorical RF $\mathbf{Z}(\mathbf{u})$ in domain $\mathcal{A}$. $\mathbf{Z}(\mathbf{u})$ takes values from a finite set of K categories. Let $\mathrm{k} = z(\mathbf{u}_\alpha)$ and $\mathrm{k}' = z(\mathbf{u}_\beta)$, where $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ are two locations in $\mathcal{A}$, $|\overrightarrow{\mathbf{u}_\alpha\mathbf{u}_\beta}|$ is minimum and k is the closest category to k′. Consider the realizations of $\mathbf{Z}(\mathbf{u})$ along the trace of vector $\overrightarrow{\mathbf{u}_\alpha\mathbf{u}_\beta}$ at certain discretization, then the set $S = \{\mathrm{k}|\mathrm{k} \neq \mathrm{k}'\}$ contains intervening intervals of specific categories. $|S|$ is the number of intervening intervals from k to k′ and is represented by $\delta_{\mathrm{k},\mathrm{k}'}$.

Figure 4.1 shows a referential image of a depositional environment. The number of intervening intervals between Category 03 (red dot) at an arbitrary location and Category 06, k′, is represented by $\delta_{\mathrm{k},\mathrm{k}'}$. The first step is to obtain the closest category k′ and define a vector from the closest k to k′. The intervening intervals relate to categories 04, 05, 02, and 06. The distance between touching categories is one, it is zero if they are the same categories. For instance, $\delta_{5,4} = 1$ and $\delta_{5,5} = 0$. All locations in the image may be visited to obtain a distribution of distances where the variable is the number of intervening intervals.

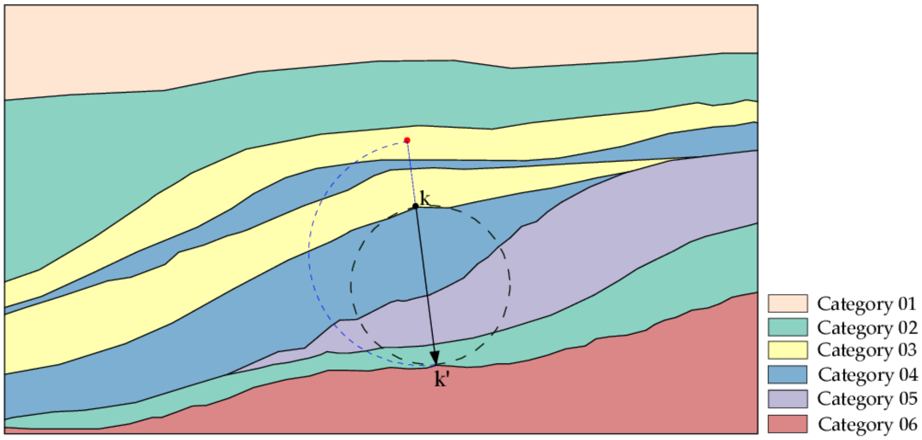


Figure 4.1: Intervening intervals in a referential image of a depositional environment. The red dot is an initial location in the image. The dashed blue semi-circle outlines the minimum distance to k′. The closest k (black dot) to k′ is considered following the definition. The black arrow represents $\overrightarrow{\mathbf{u}_\alpha\mathbf{u}_\beta}$, where $|\overrightarrow{\mathbf{u}_\alpha\mathbf{u}_\beta}|$ is minimum. The number of intervening intervals, $\delta_{k,k'} = 4$.

Figure 4.2 is a sketch of intervening intervals in drillholes. Category 01 is a cross-cutting unit. The composite lengths plotted as reference are not relevant as the definition is based on intervals. In DH01, the number intervening interval between Category 02 and Category 03 is one. In DH02, the number of intervening intervals between Category 03 and Category 05 is two.

Intervening intervals are robust concerning continuity, the intervals are counted only once for a respective category. Intervening intervals are insensitive to boundary extent as long as the proportion of lithofacies contacts is kept. On the contrary, TPMs are affected by the boundary extent and proportions of lithofacies. Unlike TPMs, using intervals is not a good measure for the analysis of

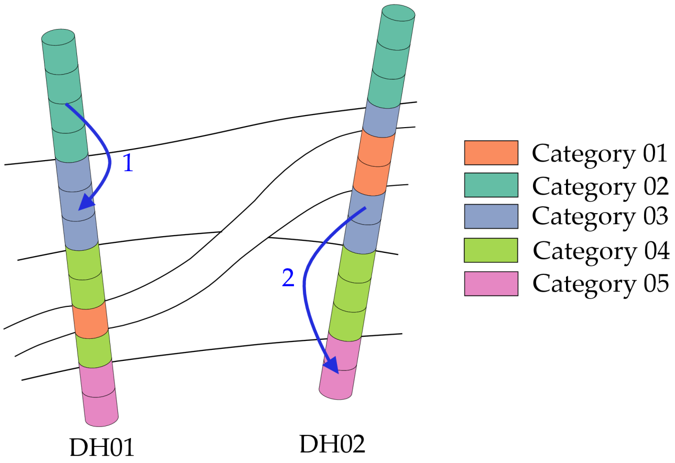


Figure 4.2: Intervening intervals in drillholes. In the background, black lines outline a cross-section. Blue arrows start at the interval of category k and end at the interval of category k′. In DH01, $\delta_{2,3} = 1$. In DH02, $\delta_{3,5} = 2$; the blue arrow starts at the closest Category 03 interval to the Category 05 interval. δ values for intervals of the same categories are zero, for instance, $\delta_{4,4} = 0$.

proportions; however, it is useful to interpret relations between geometries. Intervening intervals are the base to define interval probabilities but they depend more on the contacts.

4.2.2 Interval Probabilities

Interval probability is the probability of the number of intervening categories between a pair of categories k and k′. In interval probabilities, the distances are discrete and take values from 1 to $K-1$. The distributions can be obtained for all pairs k,k′ given K categories, however, the interval probability between the same categorical intervals is always one at zero value.

A novel dissimilarity matrix is proposed using interval probabilities as initial input. This dissimilarity matrix is calculated by taking the expected value of the interval probabilities for all pairs of categories k and k′, Equation (4.2). The resultant square-hollow matrix is non-symmetric, then it is made symmetric before any calculation.

$$\delta = E\left\{\begin{bmatrix} 0 & \delta_{1,2} & \cdots & \delta_{1,K} \\ \delta_{2,1} & 0 & \dots & \delta_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{K,1} & \delta_{K,2} & \cdots & 0 \end{bmatrix}\right\} \tag{4.2}$$

Choosing a measure of distance or dissimilarity in spatial problems is case-based. The proposed dissimilarity matrix was developed to interpret associations between categories for truncation tree inference. Unlike transition probabilities, interval probabilities are insensitive to the domain's size and the proportions of the categories.

4.2.3 Special Tools

Additional tools for tree inference in presence of sparse data are presented as a practical application. The purpose is to apply the proposed interval distance. Figure 4.3 is a schematic representation of sparse surface sampling where the colored dots refer to categories. The colored squares are categories in a NN model built from the samples. The extension limit of the gridded model is represented by the dashed black line and is arbitrary. If the red colored square represents a grid point with categorical value k, and the black colored square is category k', then k' is the closest to k and the number of intervening intervals between k and k' is 1. Similarly, if the green square is now a categorical grid point represented by k', then this k' is the closest to the red grid point, and the intervening intervals between the new k' (green) and k (red) grid points are 2. The intervening intervals between two grid points with the same category are always zero according to the definition. In practice, only grid points at the contacts are considered during the calculation, this ensures that the results are not affected by internal domain sizes and grid model extents. In the graph, the number of intervening intervals between the two most distant red grid points is zero as the algorithm defines the vector targeting the closest category k'.

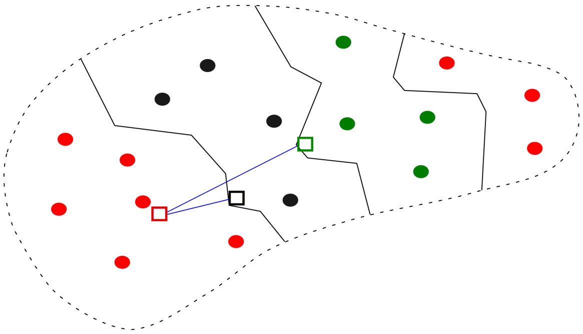


Figure 4.3: Special application of the number of intervening categories in surface sampling. Colored dots represent different categories. Colored squares represent categories in a NN model built from the sample data and the solid black lines are the contacts in the model. The dashed black line is the grid's limit. The solid blue lines represent traces of the vectors defined between two grid points.

The Swiss Jura Rock Type dataset, (Goovaerts, 1997), is used to illustrate the concept. Figure 4.4 shows the categorical data, there are five categories: 1: Argovian, 2: Kimmeridgian, 3: Sequanian, 4: Portlandian, and 5: Quaternary. The interval probabilities are calculated on the NN model using a search radius of 0.2 km and shown in Figure 4.5. The distributions in the diagonal of the plot are not shown as they present a probability of one at a zero distance. The distributions in the graph are similar but not equal as the number of intervening intervals is not symmetric.

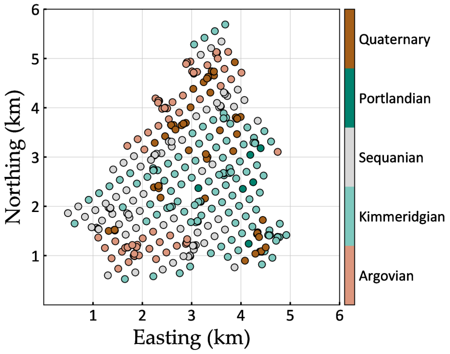


Figure 4.4: Location map of Swiss Jura rock type dataset.

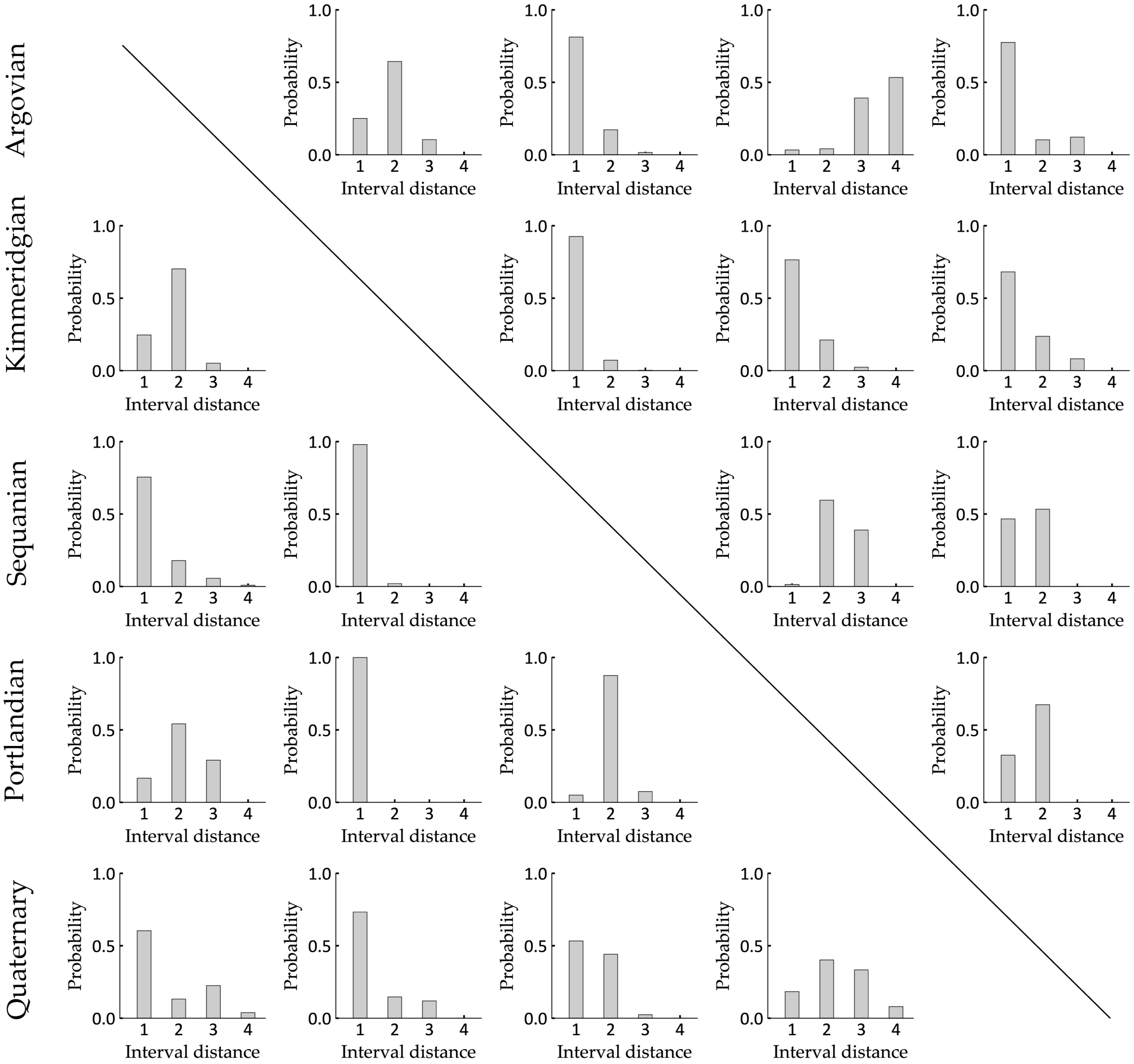


Figure 4.5: Interval probabilities in Jura rock-type dataset. The distributions in the diagonal are not considered as the intervening intervals for equal categories are always zero. The distributions for conjugate pairs are not equal given the non-stationarity in categorical variables and the fact that the interval distance is not symmetric.

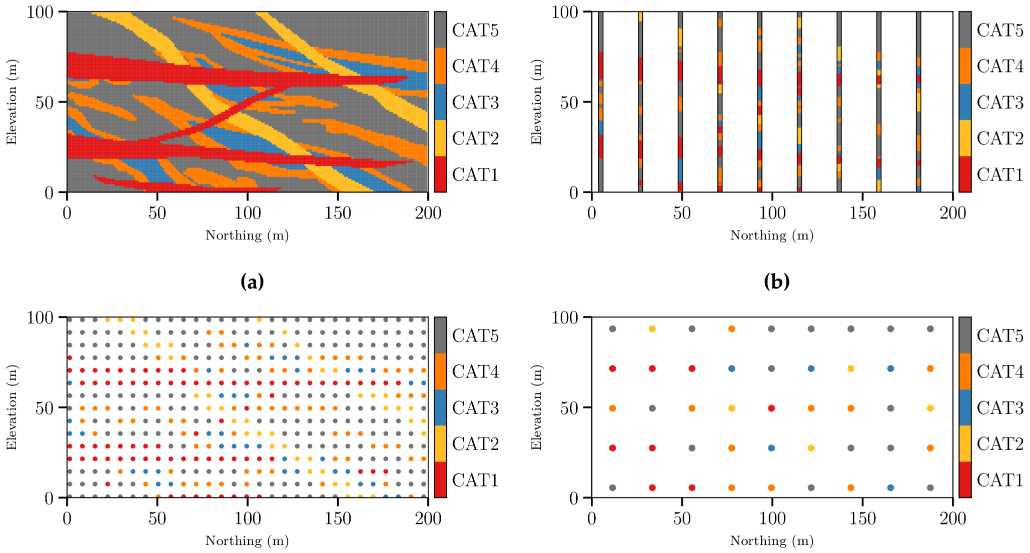
The application for point samples amounts to applying the concept of intervening intervals on images. The interval probabilities for grid models are calculated considering all pairs of grid points or vectors, then the dissimilarity matrix is obtained. The interval probability-based dissimilarity matrix calculated from images differs slightly from the calculation on drillholes. In the case of drillholes, the vectors are fixed by the drilling campaign, therefore calculations to find the closest category k′ follows the drillhole orientation. On the contrary, when the dissimilarity values are calculated from images, all vectors are first evaluated to obtain the closest target category k′. Regardless of these differences, dissimilarity matrices calculated from drillholes and images are similar. Moreover, the significance of the interval probability-based dissimilarity matrix relies on relative dissimilarity values between categories, therefore values of elements in the matrix are less important.

A comparison between the results from the interval-based distance calculated on drillholes and sparse samples at different spacings (Cabral Pinto & Deutsch, 2017; Wilde, 2010) is presented using a synthetic example. The goal is to compare dissimilarity matrices obtained from drillholes and point samples with the calculated from the reference image. Figure 4.6a shows the reference image consisted of 200x100 grid cells. The synthetic drillholes are vertical and evenly spaced. The point samples are taken from a regularly spaced grid aligned to the East and Elevation axes. The dissimilarity matrices of the drillholes (Figure 4.6b) and the sparse samples (Figure 4.6c) are presented in Figure 4.7 as an example. The matrices from the different spacing data configurations are standardized following Equation (4.3), where μ is the mean of the elements δ_{ij} in the matrix and σ is the standard deviation. The matrix norm, (Golub & Van Loan, 1996), of the difference between two standardized matrices is used for the comparison, Equation (4.4).

$$\hat{\delta}_{ij} = \frac{\delta_{ij} - \mu}{\sigma} \tag{4.3}$$

$$Norm = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K}\left|\hat{\delta}_{ij}(image) - \hat{\delta}_{ij}(samples)\right|^2} \tag{4.4}$$

Figure 4.8 shows the results for different spacings ranging from 30m. to 1m. The point samples (black line) show dissimilarities closer to the benchmark as the spacing is decreased. The drillholes (red line) show a steady line, which is explained by visual inspection of the reference. The main anisotropies of categories are sub-horizontal and horizontal, therefore increasing the number of drillholes does not impact the dissimilarity matrix. From the example, different drillhole spacing configurations give similar results compared to using point samples spaced at 7 meters. At spacings close to zero there is still a difference between the the dissimilarities from images and drillholes. The small difference is caused by the additional flexibility during the calculation of interval distances in images as the shortest vectors are searched freely. In practice, the dissimilarities obtained from the NN model of the sparse sampling or the composite data is preferred over the dissimilarities from



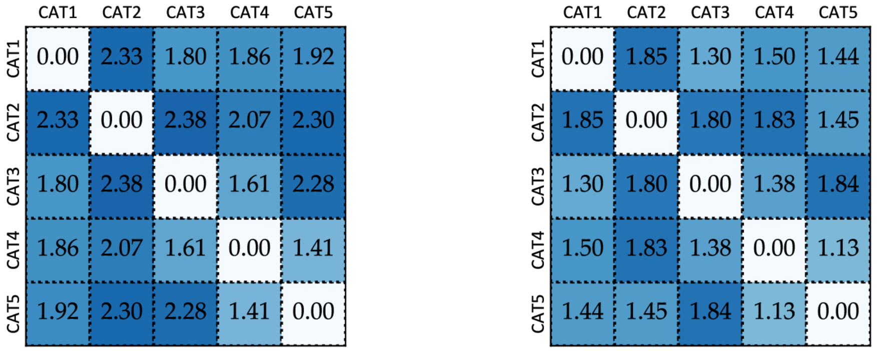
(a)

(b)

(c)

(d)

Figure 4.6: Reference image and samples. (a) Reference image (b) Drillholes evenly spaced at 22m. The bottom graphs are point samples from the reference image with different spacings. (c) Samples with 7m. spacing (2.2% of the exhaustive data), and (d) samples with 22m. spacing, (0.2% of the exhaustive data).



	CAT1	CAT2	CAT3	CAT4	CAT5
CAT1	0.00	2.33	1.80	1.86	1.92
CAT2	2.33	0.00	2.38	2.07	2.30
CAT3	1.80	2.38	0.00	1.61	2.28
CAT4	1.86	2.07	1.61	0.00	1.41
CAT5	1.92	2.30	2.28	1.41	0.00

(a)

	CAT1	CAT2	CAT3	CAT4	CAT5
CAT1	0.00	1.85	1.30	1.50	1.44
CAT2	1.85	0.00	1.80	1.83	1.45
CAT3	1.30	1.80	0.00	1.38	1.84
CAT4	1.50	1.83	1.38	0.00	1.13
CAT5	1.44	1.45	1.84	1.13	0.00

(b)

Figure 4.7: Dissimilarity matrices in example with five categories. (a) Dissimilarity matrix from drillholes (Figure 4.6b) (b) Dissimilarity Matrix from sparse samples (Figure 4.6c).

drillholes for inference purposes.

The applications for interval data and a special case have been reviewed where the final output is a matrix of dissimilarities between categories. The next section focuses on the algorithms to infer categorical associations from a dissimilarity matrix.

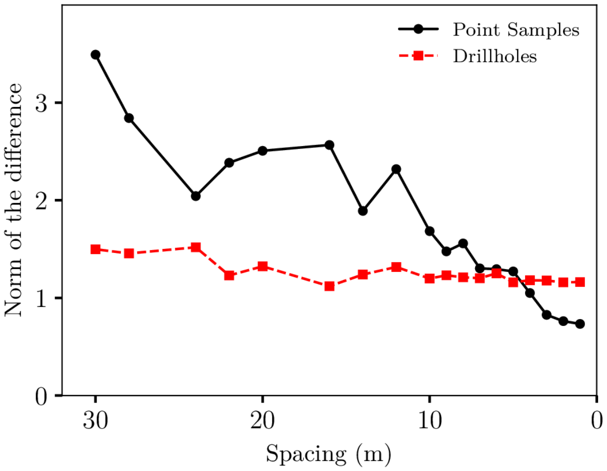


Figure 4.8: Dissimilarity matrices from drillholes and point samples compared to the dissimilarity of a reference image.

4.3 Inference Algorithms

A distance or dissimilarity matrix is the primary input for most tree inference algorithms. Tree-like representations require hierarchies between groups of categories. A different approach considers categorical proportions and spatial continuities to set the ordering of the latent Gaussian variables in the tree, D. S. F. Silva (2018). The focus here is the automatic inference of hierarchical structures and geological associations to provide a reduced set of possible truncation trees. Two techniques are reviewed including hierarchical clustering and spectral graph partitioning.

4.3.1 Hierarchical Clustering

Several clustering techniques output flat descriptions of the data or disjoint clusters, (Duda et al., 1973). Hierarchical methods output non-flat representations and are among the most used unsupervised techniques. They are an appealing option to infer trees. Hierarchical clustering methods use dissimilarity matrices. Sometimes a metric cannot be supplied for multidimensional data, and the n-D points do not have a metric by themselves. However, if dissimilarities can be calculated for the pairs $\delta_{k,k'}$, where $\delta_{k,k'} \geq 0$, with equality holding if and only if $k = k'$, then agglomerative clustering is applicable, (Duda et al., 1973). The agglomerative method, Single Linkage Cluster Analysis (SLCA), is used jointly with Minimum Spanning Tree (MST) to infer trees.

4.3.1.1 MST and Single Linkage Cluster Analysis

Given K points and all their $KC2$ pairwise weights or distances, the MST minimizes the cost, (J. B. Kruskal, 1956; Prim, 1957). Using MST-based SLCA helps when the number of categories increases and high-order information is not optimally described by canonical variates, (Gower et al., 2016). The MST from the dissimilarity matrix δ contains the information for SLCA. The technique uses the minimum distance between agglomerative clusters to generate a dendrogram from the MST, (Gower et al., 2016; Rohlf, 1973). Figure 4.9 shows the MST obtained from the dissimilarity matrix calculated on drillholes (Figure 4.7a). Figure 4.10 shows the dendrogram using SLCA.

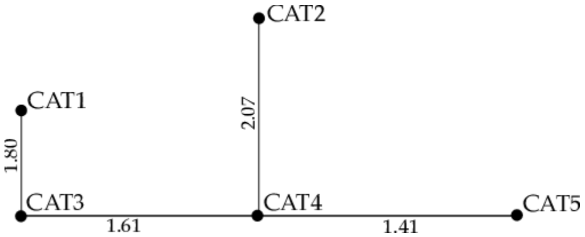


Figure 4.9: Minimum spanning tree calculated from a dissimilarity matrix. The nodes represent categories. The edges are labeled with their weights according to the δ.

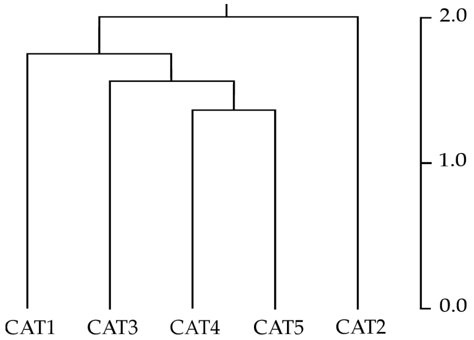


Figure 4.10: Dendrogram from single linkage cluster analysis on a minimum spanning tree with five categories. The positioning of the categories from left to right corresponds to the output of the program `CLUSTER MST`, see Section A.1.3, the documentation is found in Rohlf (1973).

4.3.1.2 Construct Truncation Trees

Dendrograms summarize the output from SLCA and are intuitive to visualize hierarchical associations of categories; however, a truncation tree structure is required. A common alternative to obtain trees from the dendrogram uses thresholds, the categories below a cutting threshold form clusters, and the hierarchical structure is maintained towards the root. To decide the dissimilarity threshold value, the maximum encountered inter-distance is commonly used, however, this approach does not consider any constraint of permissible clustered categories. Figure 4.11 shows a dendrogram where a distance is chosen to cluster all categories below d. For instance, CAT3, CAT4, and CAT5 are grouped below node 2. This cluster and CAT1 are at the same level in the hierarchy. CAT2 and the cluster below 1 conformed by CAT1, CAT3, CAT4, and CAT5 are at the same level. The hierarchical structure is built from d towards the root. The process is repeated for multiple d values to build a set of trees. Using a threshold with the lowest value generates a tree with $K-1$ latent Gaussian variables, and using the maximum threshold d generates a tree with 1 latent Gaussian variable. The interval-based dissimilarity generates monotonic relationships between categories encoding at some level the ordering of categorical clusters in the MST.

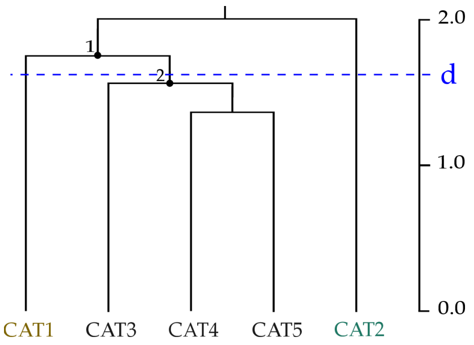
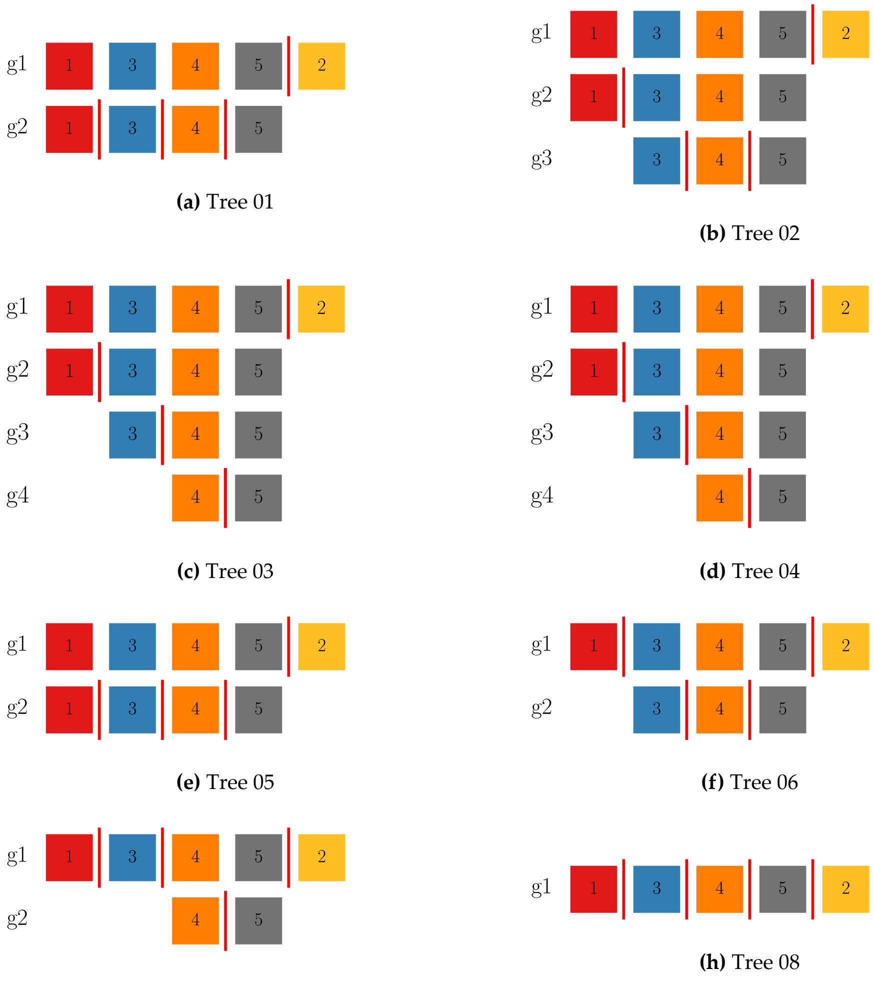


Figure 4.11: Threshold distance d in dendrogram for the generation of a truncation tree.

Figure 4.12 shows automatically generated trees with SLCA and the algorithm to construct trees. According to the categorical image used in this example, the most suitable tree is Tree 02 in Figure 4.12b. Further modifications to the order of the latent Gaussian variables may be applied.



(a) Tree 01

(b) Tree 02

(c) Tree 03

(d) Tree 04

(e) Tree 05

(f) Tree 06

(g) Tree 07

(h) Tree 08

Figure 4.12: Inferred truncation trees with SLCA in example with five categories.

4.3.2 Spectral Graph Partitioning

HTPG potential relies on establishing a suitable tree structure. Graphs may be used to describe geometric and hierarchical relationships of categories in interval data and images, which are encoded as dissimilarities. With a graph built from a dissimilarity, spectral partitioning helps to determine the association of categories by assessing the connectivity of the Fiedler vector. Concepts in graph theory and algebraic methodologies to infer associations are reviewed.

4.3.2.1 Definition of a Graph

A graph G is an ordered pair (V, E), where $V \neq \{\}$ and $E \subseteq \Big\{\{v_1, v_2\} \,|\, (v_1, v_2) \in V \times V\Big\}$, Figure 4.13. V is the set of vertices (nodes) of G. Elements in E are edges of G. If two nodes are connected with more than one edge, the graph is multiedge. Loops are self-connected nodes. Simple graphs have no loops or multi-edges. In undirected graphs, the ordering of the connection is not considered.

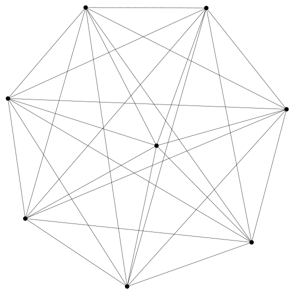


Figure 4.13: Fully connected graph. Nodes are represented by black dots. Edges establish the connectedness between nodes.

The adjacency Matrix $A_{n,n}$ summarizes the connectivity of a graph G, n is the number of nodes. $A_{ij} = 0$ represents unconnected nodes, $A_{ij} = 1$ represents connected nodes. Here the focus is on simple finite weighted undirected graphs where $0 \leq A_{ij} \leq 1$ and $A_{ii} = 0$.

4.3.2.2 Spectral Graph Theory

Spectral graphs associate linear algebra to graphs (Spielman & Teng, 2007). The degree matrix D_{nxn} of G is a diagonal matrix where $D_{ii} = \sum_{j=1}^{n} D_{ij}$. In graph theory, the adjacency and the degree matrix are used to calculate the Laplacian of G denoted by L_G. The Laplacian contains all of the information in the graph.

Typically the L_G works in regular graphs but fails in irregular ones, therefore the normalized Laplacian $\mathscr{L}_G$ is often preferred, (Chung, 1997). Here, the normalized graph Laplacian is referred to simply as Laplacian unless stated otherwise. The first step is to define the normalized adjacency matrix $\mathscr{A}$, Equation (4.5), where A and D are adjacency and degree matrices respectively.

$$\mathscr{A} \equiv D^{-1/2} A D^{-1/2} \tag{4.5}$$

The Laplacian equals the identity matrix subtracted by the normalized adjacency matrix, Equation (4.6). The relationship between the non-normalized Laplacian Matrix L_G and $\mathscr{L}_G$ is shown Equation (4.7).

$$\mathscr{L}_G \equiv I - \mathscr{A} \tag{4.6}$$

$$\mathscr{L}_G = I - \mathscr{A} = D^{-1/2}(D - A)D^{-1/2} = D^{-1/2} L_G D^{-1/2} \tag{4.7}$$

In the normalized matrices $\mathscr{A}$ and $\mathscr{L}_G$, if $\alpha_1 \geq \cdots \geq \alpha_n$ represents the eigenvalues of $\mathscr{A}$; and $\lambda_1 \leq \cdots \lambda_n$, the eigenvalues of $\mathscr{L}_G$. The following relationships hold:

$$1 = \alpha_1 \geq \cdots \geq \alpha_n \geq -1 \tag{4.8}$$

$$0 = \lambda_1 \leq \cdots \leq \lambda_n \leq 2 \tag{4.9}$$

A well-known application of spectral methods is clustering. The eigenvalues and eigenvectors from

L_G or $\mathscr{L}_G$ are central in understanding graphs.

4.3.2.3 Fiedler Vector

The Fiedler vector is the eigenvector corresponding to the second smallest eigenvalue (Fiedler, 1973; Spielman & Teng, 2007). The Fiedler vector of a graph's Laplacian is used to cut the graph and obtain associated classes from a distance matrix. Let v_2 be the eigenvector that corresponds to the second smallest eigenvalue λ_2 of $\mathscr{L}_G$, then λ_2 describes the connectivity of G. In graph partitioning, the quality of the cut is related to smaller values of λ_2. A Fiedler vector is also known as the algebraic connectivity of G; a higher connectivity value relates to a graph with more or stronger edges. For instance, the second-smaller eigenvalue of $\mathscr{L}_G$ in a complete graph of 100 nodes is several orders of magnitude greater than a cycle graph with the same amount of nodes, (Slininger, 2013).

Clusters in G are formed by assigning the $i-th$ element of the Fiedler vector to one of the mutually exclusive clusters C_1 and C_2, with $C_1 \cup C_2 = V$. $C_1 = \{i|v_2(i) > 0\}$ and $C_2 = \{i|v_2(i) < 0\}$.

4.3.2.4 Spectral Partitioning

Spectral partitioning or clustering groups related nodes within G using spectral properties of the Laplacian and the adjacency matrix. The graph's weights are commonly defined by a distance function, (Knyazev, 2018). Spectral partitioning techniques bisect the graph G by analyzing the signs of the components of the Fiedler vector (Fiedler, 1973). It does not make assumptions on the form of the clusters, such as the convexity of the sets, (Von Luxburg, 2007). Some applications result in signed graphs where the weights in the adjacency matrix are either positive or negative, complicating the application of spectral graphs, (Knyazev, 2018). The proposed measure of distance based on intervals (Velasquez & Deutsch, 2022) gives a positively weighted A_G allowing the utilization of the properties of the graph Laplacian related to Fiedler vectors. Additionally, a major reason to consider Fiedler's spectral method is the fact that it does not require coordinates to be assigned.

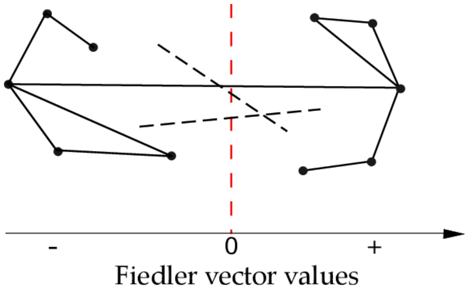


Figure 4.14: Schematic representation of bisecting a graph using the Fiedler vector. The nodes in the graph represent categories. The graph has 10 categories. The bisection is performed at zero value of the Fiedler vector's values (dashed red line). Not all edges are plotted, strong connectivities are represented by solid black lines; weak connectivities are represented by dashed black lines.

A different approach considers multiple eigenvectors simultaneously during the partitioning analysis followed by the application of k-means in the reduced set of the eigenvectors of $\mathscr{L}_G$, (Ng, Jordan, & Weiss, 2001). Here, the classic approach of the Fiedler vector in a positive weighted A_G is used, (Fiedler, 1973).

4.3.2.5 Spectral Partition of a Distance Matrix

The δ contains the expected number of distinct intervals between two points in space, it takes a value of one when two different categories are contiguous. The steps to partition the graph built from a dissimilarity matrix are presented. Adjacency matrices are calculated from δ using the reciprocals of the non-diagonal elements δ_{ij}. The Degree Matrix D is a diagonal matrix obtained from adding the i-th row elements of A in D_{ii}. The Laplacian $\mathscr{L}_G$ is obtained from Equation (4.7). The eigenvalues and eigenvectors of $\mathscr{L}_G$ are calculated. The eigenvector corresponding to the second smallest eigenvalue is of particular interest for partitioning a graph. The categories corresponding to values $v_2(i) > 0$ are assigned to the first cluster C_1 and the categories corresponding to values $v_2(i) < 0$ are assigned to the cluster C_2. Spectral partitioning classifies the categories into sub-groups. The steps are summarized as follows:

1. Let the adjacency matrix $A_{ij} = \delta_{ij}^{-1} \quad \forall i \neq j$, 0 otherwise.
2. With the diagonal matrix $D_{ii} = \sum_{j=1}^{n} x_{ij}$, calculate the $\mathscr{L}_G = D^{-1/2}(D - A)D^{-1/2}$
3. Select the Fiedler vector : $v \mid Lv = \lambda_2 v$
4. Assign elements of Fiedler vector in cluster C_1 and C_2 based on their sign.

The previous steps can be applied recursively or in a subset of the categories depending on each case to obtain clusters. However, the first partitioning is often enough to infer associated categories.

4.3.2.6 Example

A categorical image generated unconditionally with HTPG is presented to demonstrate the application of graph partitioning. Eight category labels from 1 to 8 are positioned in the simplified notation of a tree structure with three latent Gaussian variables, ((1)(4)((1)(1)(1))), the resultant truncation tree is shown in Figure 4.15. Three independent unconditional images of 200x200 cells were generated with SGS. The variogram models of the latent Gaussian variables have one structure. The main ranges for $Y1$ are 30 and 15 units with an azimuth of $90°$, 33 and 8 units with an azimuth of $90°$ for $Y2$, and finally 20 and 8 units with an azimuth of $105°$ for latent Gaussian variable $Y3$. The threshold values in the truncation tree were obtained by using equal global proportions for all categories. One categorical realization is shown in Figure 4.17a. The associations of the categories are clear by visual inspection of the image. Category 1 is not in contact with categories 6,7, and 8. Meanwhile, categories 2,3,4, and 5 are sequenced and establish a non-contact zone between category 1 and the group of categories 6,7, and 8. Categories 6,7 and 8 are also sequenced. Synthetic drillholes were taken from the image, Figure 4.17b to emulate limited data. The methodology for inference of categorical associations is applied to this data.

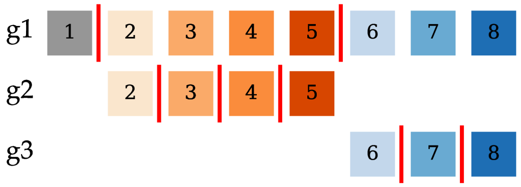


Figure 4.15: Underlying truncation tree.

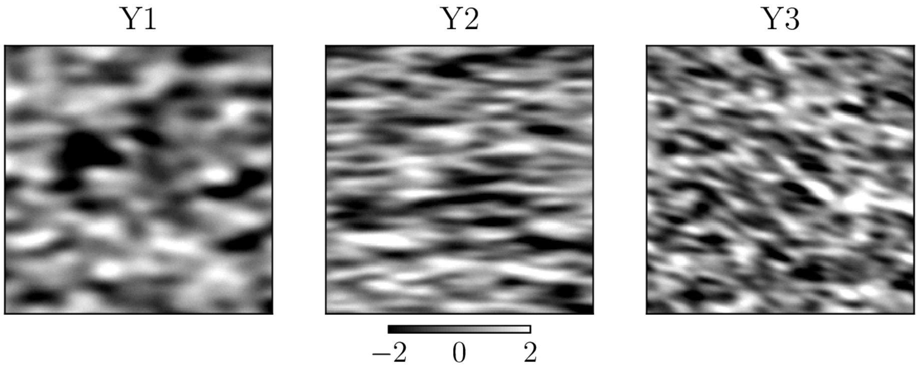
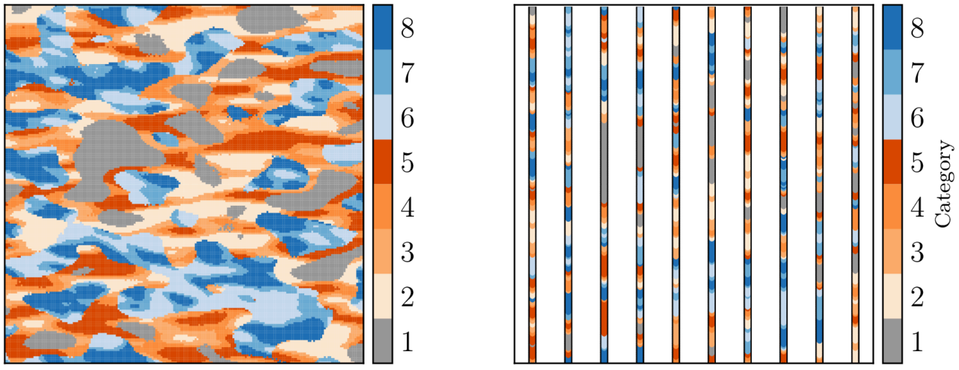


Figure 4.16: A realization of latent Gaussian variables generated with SGSIM. Models of continuity were previously defined

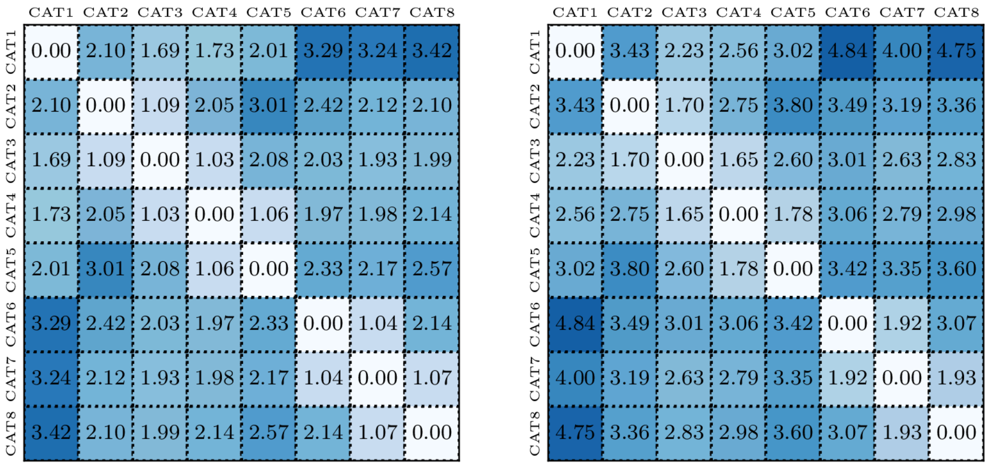


(a) **(b)**

Figure 4.17: (a) Synthetic categorical image (b) Interval data retrieved from the image

First, the dissimilarity matrices are calculated using the drillholes and the image, Figure 4.18 using the programs `INTERVAL` and `INTERVALG`. Section A.1.1 and Section A.1.2 details the documentation of the programs.

The analysis continues using the dissimilarity matrix calculated from drillholes, Figure 4.17b. Figure 4.19 is the adjacency matrix obtained from the dissimilarity matrix. The degree matrix Figure 4.20a is required to obtain the Laplacian, Figure 4.20b. The Fiedler vector is selected from $\mathscr{L}_G$.



	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8
CAT1	0.00	2.10	1.69	1.73	2.01	3.29	3.24	3.42
CAT2	2.10	0.00	1.09	2.05	3.01	2.42	2.12	2.10
CAT3	1.69	1.09	0.00	1.03	2.08	2.03	1.93	1.99
CAT4	1.73	2.05	1.03	0.00	1.06	1.97	1.98	2.14
CAT5	2.01	3.01	2.08	1.06	0.00	2.33	2.17	2.57
CAT6	3.29	2.42	2.03	1.97	2.33	0.00	1.04	2.14
CAT7	3.24	2.12	1.93	1.98	2.17	1.04	0.00	1.07
CAT8	3.42	2.10	1.99	2.14	2.57	2.14	1.07	0.00

(a) δ_{Image}

	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8
CAT1	0.00	3.43	2.23	2.56	3.02	4.84	4.00	4.75
CAT2	3.43	0.00	1.70	2.75	3.80	3.49	3.19	3.36
CAT3	2.23	1.70	0.00	1.65	2.60	3.01	2.63	2.83
CAT4	2.56	2.75	1.65	0.00	1.78	3.06	2.79	2.98
CAT5	3.02	3.80	2.60	1.78	0.00	3.42	3.35	3.60
CAT6	4.84	3.49	3.01	3.06	3.42	0.00	1.92	3.07
CAT7	4.00	3.19	2.63	2.79	3.35	1.92	0.00	1.93
CAT8	4.75	3.36	2.83	2.98	3.60	3.07	1.93	0.00

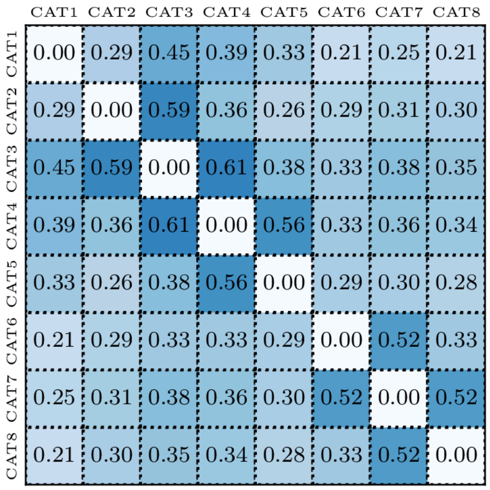
(b) $\delta_{IntervalData}$

Figure 4.18: Dissimilarity matrices in example with eight categories, values are expected dissimilarities. (a) The result from the image (b) The result from drillholes. Respective elements δ_{ij} may differ in value but the relative relationships over all categories prevail shown with the blue shading.

Figure 4.21a shows the eigenvalues in ascending order from left to right. The second smallest eigenvalue is close to 1.0; the vector corresponding to this value is the Fiedler vector. In Figure 4.21b the eigenvector values are divided at a value of zero. The eight categories are subdivided into two sub-graphs or clusters.

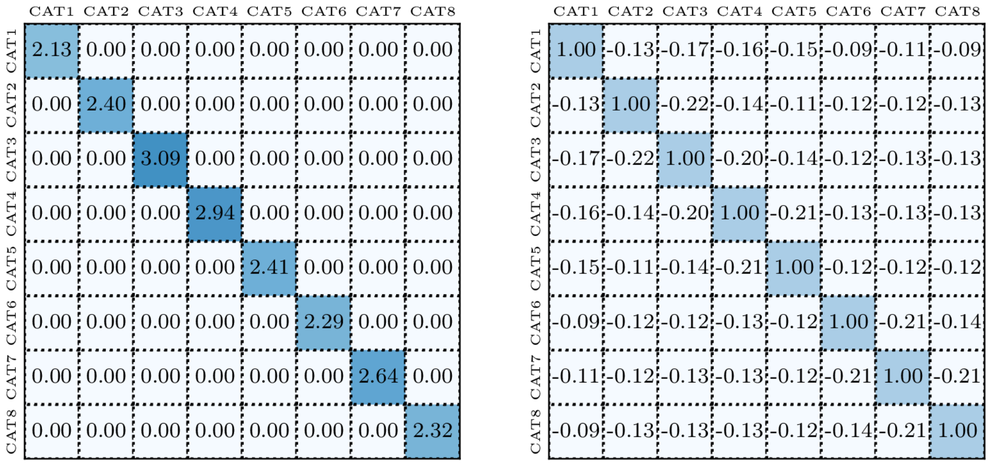
The cluster, C_2, consists of categories 6-7-8 with negative Fiedler vector values, and the cluster, C_1, consists of categories 1-2-3-4-5 with positive Fiedler vector values. From the figure, C_1 may be further divided into two clusters containing categories 1 and 2-3-4-5. Iterative partitioning may be applied to divide each sub-graph; however, the best results were achieved from analyzing the clusters in the first iteration. Note that the categorical labels in the dissimilarity matrices were positioned to match the ordering of the categories observed in the underlying tree from Figure 4.15. Categorical labels in Figure 4.21b follow the same sequence from left to right to ease visualization of the associations; however, this is arbitrary in practice.

With the analysis of the Laplacian, the sequence of the categories in the MST may indicate the actual ordering of the categories, Figure 4.22. The MST in the figure shows the correct ordering of categories 6-7-8 and 2-3-4-5. The presented procedure aims to detect associations of categories automatically as input to obtain possible truncation trees. Figure 4.23 shows possible trees using the inferred associations of categories. Figure 4.23d is an inferred tree that matches the underlying tree. For comparison, the inferred trees following the hierarchical clustering approach are presented in Figure 4.24.



	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8
CAT1	0.00	0.29	0.45	0.39	0.33	0.21	0.25	0.21
CAT2	0.29	0.00	0.59	0.36	0.26	0.29	0.31	0.30
CAT3	0.45	0.59	0.00	0.61	0.38	0.33	0.38	0.35
CAT4	0.39	0.36	0.61	0.00	0.56	0.33	0.36	0.34
CAT5	0.33	0.26	0.38	0.56	0.00	0.29	0.30	0.28
CAT6	0.21	0.29	0.33	0.33	0.29	0.00	0.52	0.33
CAT7	0.25	0.31	0.38	0.36	0.30	0.52	0.00	0.52
CAT8	0.21	0.30	0.35	0.34	0.28	0.33	0.52	0.00

Figure 4.19: Adjacency matrix obtained from inverting the elements in the δ



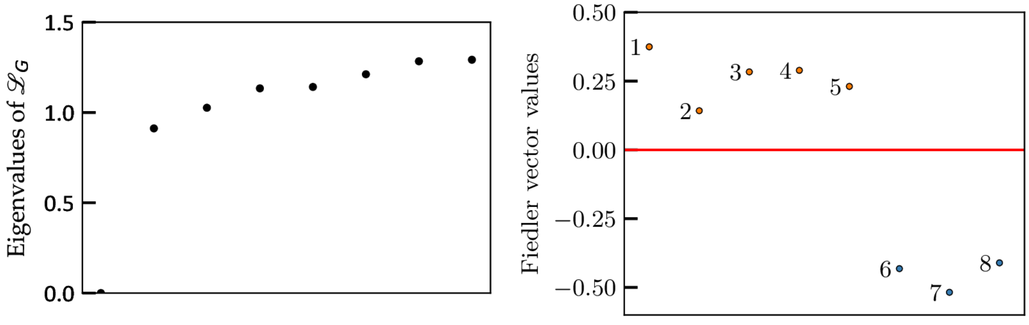
	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8
CAT1	2.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
CAT2	0.00	2.40	0.00	0.00	0.00	0.00	0.00	0.00
CAT3	0.00	0.00	3.09	0.00	0.00	0.00	0.00	0.00
CAT4	0.00	0.00	0.00	2.94	0.00	0.00	0.00	0.00
CAT5	0.00	0.00	0.00	0.00	2.41	0.00	0.00	0.00
CAT6	0.00	0.00	0.00	0.00	0.00	2.29	0.00	0.00
CAT7	0.00	0.00	0.00	0.00	0.00	0.00	2.64	0.00
CAT8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	2.32

(a)

	CAT1	CAT2	CAT3	CAT4	CAT5	CAT6	CAT7	CAT8
CAT1	1.00	-0.13	-0.17	-0.16	-0.15	-0.09	-0.11	-0.09
CAT2	-0.13	1.00	-0.22	-0.14	-0.11	-0.12	-0.12	-0.13
CAT3	-0.17	-0.22	1.00	-0.20	-0.14	-0.12	-0.13	-0.13
CAT4	-0.16	-0.14	-0.20	1.00	-0.21	-0.13	-0.13	-0.13
CAT5	-0.15	-0.11	-0.14	-0.21	1.00	-0.12	-0.12	-0.12
CAT6	-0.09	-0.12	-0.12	-0.13	-0.12	1.00	-0.21	-0.14
CAT7	-0.11	-0.12	-0.13	-0.13	-0.12	-0.21	1.00	-0.21
CAT8	-0.09	-0.13	-0.13	-0.13	-0.12	-0.14	-0.21	1.00

(b)

Figure 4.20: (a) Degree matrix. (b) Laplacian.



(a)

(b)

Figure 4.21: Eigenvalues and Fiedler vector values in graph partitioning. (a) Sorted eigenvalues of the Laplacian of graph G. The noticeable difference between the minimum and the second smallest eigenvalue suggests an adequate partition. (b) Bipartitioning based on Fiedler vector values above and below zero.

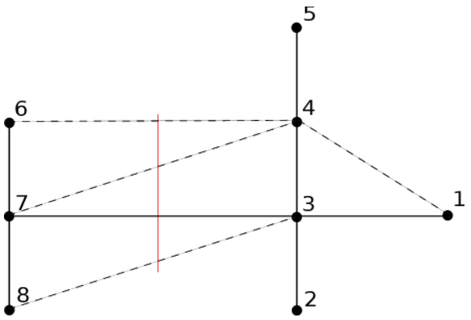
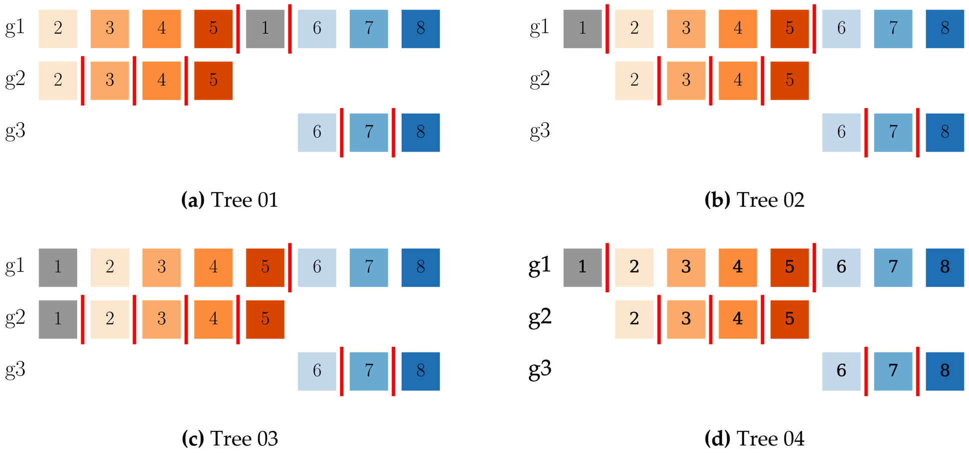


Figure 4.22: Bisected graph with eight categories and an MST. Solid lines are the MST based on the δ. Dashed lines represent moderately connected edges. An adjacency value of 0.5 was set to filter out weak edges. Unconnected nodes in the graph have adjacency values less than 0.5.



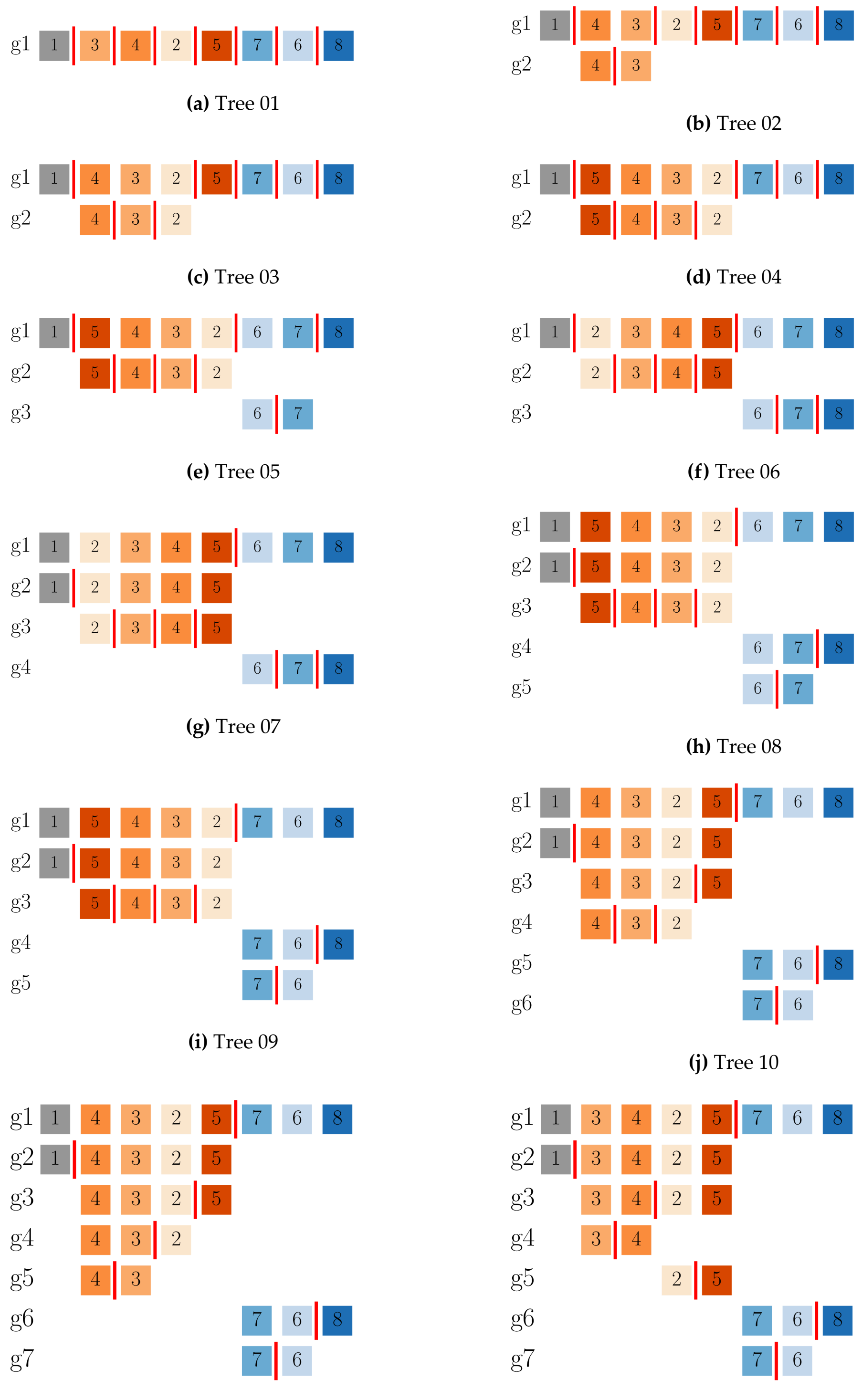
(a) Tree 01

(b) Tree 02

(c) Tree 03

(d) Tree 04

Figure 4.23: Inferred truncation trees with spectral partitioning. Different structures are used for the inferred associations. Symmetric trees may be considered. Tree 02 matches the underlying tree.



(a) Tree 01

(b) Tree 02

(c) Tree 03

(d) Tree 04

(e) Tree 05

(f) Tree 06

(g) Tree 07

(h) Tree 08

(i) Tree 09

(j) Tree 10

(k) Tree 11

(l) Tree 12

Figure 4.24: Inferred truncation trees with SLCA in example with eight categories. Tree 06 matches the underlying tree. Further modifications to the ordering of the latent Gaussian variables may be applied if needed.

4.4 Conclusions

HTPG is flexible in the use of trees, similar associations of categories may be expressed in alternative structures. Choosing a tree for HTPG depends on categorical associations and models of continuity; the inference of associations with limited data was discussed. A novel measure of dissimilarity was proposed to quantify the association of categories in interval data. A special application for surface samples was explained and compared to the results from drillholes at different spacing configurations. As expected, the inference results depend on the quality and quantity of interval data. Table 4.1 shows the advantages and disadvantages of using TPM, ETPM, and intervals for the inference of categorical associations.

Table 4.1: Comments on alternative concepts for establishing a measure of distance between categorical variables.

	Advantages	Disadvantages	Comments
Interval probabilities	Intuitive idea Overall better performance in SLCA and other clustering techniques Independent of the domain size and boundaries	Sensitive to the proportions of contacts between categories	Transition probabilities can be additionally used to assist the ordering Variogram parameters definitions to assist and verify the correct associations The measure can be improved to recognize more complex cutting relations
ETPM	Visual inspection describes the order of categories in simple cases Better performance than traditional TPM	Visual inspection becomes difficult with more categories	
TPM	Visual inspection describes the order of categories in simple cases	Highly dependent on diagonal element values Poor performance in clustering algorithms	Work for cases with a small number of categories and simple geological settings

Algorithms for tree inference were reviewed and applied to stationary and non-stationary cases with appealing results. The methodologies included SLCA and spectral partitioning. Concerning SLCA, the obtained dendrograms demonstrated to be intuitive to understanding the hierarchical structure of the associations; however, some drawbacks need to be considered. In hierarchical clustering, each agglomeration couples two clusters in a major cluster, therefore contact zones are not easily recognized. Using fixed d threshold values to cut the dendrogram and build a tree is rigid and fails to asymmetrically consider clusters.

Spectral graph partitions are practical tools to unravel associations of categories, especially in graphs with large numbers of nodes. The synthetic example in a stationary case gave the correct association of categories. There exist variations of spectral clustering that use multiple eigenvectors coupled with clustering techniques (Ng et al., 2001), those procedures could be integrated into the proposed algorithm. In summary, a single inferred truncation tree is not always achievable with the presented inference tools; there are other parameters to consider in HTPG such as the anisotropy of the categories. Embedding the anisotropies into the tree inference algorithms may help to differentiate cross-cutting units.

Chapter 5

Tree Selection

Inferred truncation trees may be based on technical and modeling decisions as well as geological expertise. Previous work discussed the influence of categorical proportions and continuities improve variogram models, (D. S. F. Silva, 2018), however, the complex associations of categories and the number of latent Gaussian variables in the tree complicate the decision. Measures of optimalities should be considered to obtain the best models. This section uses a simulation-based approach to explore different measures of goodness and assess multiple trees.

5.1 Variograms and Tree Structures

Numerical derivation assists in the determination of the variogram models of the latent Gaussian variables. The method minimizes the mismatch between the reference indicator variograms and the indicator variograms of the categorical models. A better fitting of numerically derived Gaussian variogram values improves modeling results. In some cases, hyper-continuities in the variograms are hard to fit. An evaluation of numerical derivation results with multiple truncation structures is presented. The goal is to quantify the error between the modeled variograms and the optimized variogram of latent Gaussian variables with multiple structures. The example considers five categories with equal isotropic spherical variograms with ranges of 50 units for the reference variograms and equal categorical proportions. The numerical derivation is configured to output 12 variogram values separated at a lag distance of 5 units. The fitted variograms are discretized at the same lag distances as the optimized values. The differences between variogram values are summarized using MSE. This procedure is repeated for all structures.

There is no unique way to index trees. The indexes used for the structures are referential and given by a recursive implementation. Figure 5.1 shows the results. The truncation structures were sorted in ascending MSE values. The error decreases in structures with more latent Gaussian variables. In the left extreme, structure 37, allows the best variogram fitting. The MSE may differ between similar structures. For instance, in the Jura rock-type dataset, the associations based on the geological knowledge require structure 02, where the yellow square represents Quaternary and the adjacent ones are Argovian, Sequanian, Kimmeridgian, and Portlandian, respectively. Structure 01 is symmetric to structure 02, but shows a slightly higher MSE.

In cases where specific categories are less relevant to the model, a structure with a lower MSE could be used. A tree structure should balance a correct association of categories and a low MSE depending on the inferred associations from drillholes and the technical goals. In summary, the number of latent Gaussian variables and the multiple structures affect the numerical approach. Cases with different anisotropies and proportions were not explored.

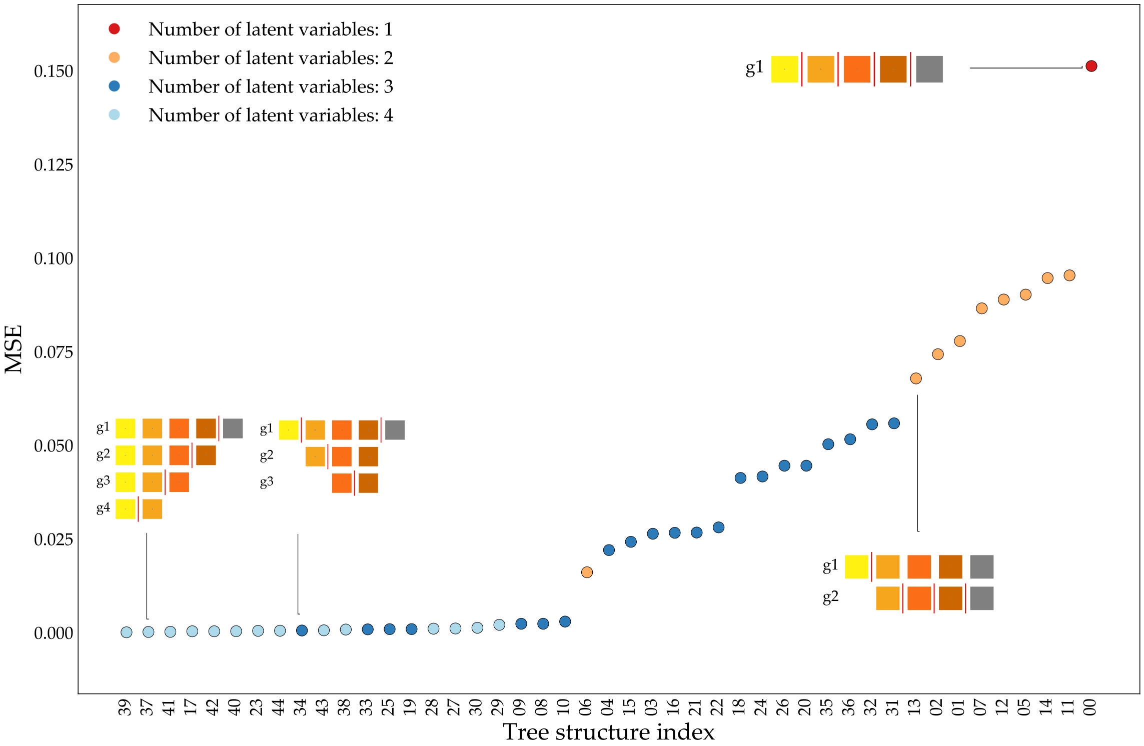


Figure 5.1: MSE between optimized variogram values of latent Gaussian variables and fitted variogram models with multiple truncation tree structures. Example with five categories. The x-axis represents the index of the structures. The y-axis represents the MSE obtained with a specific structure. The MSE values are sorted in ascending order from left to right.

5.2 Tree Optimization Workflow

Trees can be optimized to improve numerical models. Choosing a truncation tree is done from vast possibilities, and the trees are not the only parameter in HTPG. Other modeling parameters such as trends and anisotropies affect the results. The trend should be addressed correctly but is not used for tree inference. The impact that truncation trees have on measures of goodness is explored. Optimizing over all trees is possible but unpractical. Measures of goodness should be established to choose from a set of inferred trees. A simple case with five categories is considered to illustrate the idea. The reference images are generated with a unique reference tree. Trees similar to the reference tree should give best results on measures of goodness. The metrics that lead to the correct tree are used to score the trees in cases without a reference tree.

The example considers an image of 200x100 cells of 1 unit in size and five categories. Figure 5.2 details the work-flow. The tree in Figure 5.3 is used as reference to generate $T = 100$ reference images. The T models are sampled with evenly spaced vertical drillholes and composited at 3 units length. From each sample, $L = 100$ realizations are generated with HTPG, then multiple measures of goodness are calculated on the realizations and summarized. The process is repeated for all trees. The results are averaged across the different samples and trees.

The impact of the truncation trees in the final categorical models is analyzed with different measures of goodness. Two main characteristics are explored (1) the ordering of the categories within the truncation structure, and (2) the number of latent Gaussian variables in the structure. The tendencies in the results are utilized to recognize relevant metrics for tree selection. A scoring system based on individual preferences can also be configured for the tree selection. The expected scores of the T references for nT truncation trees are summarized in a graph as shown at the bottom of Figure 5.2 or using a table. In the graph, the x-axis contains the best-scored trees with a different number of latent Gaussian variables; the y-axis represents the scores. The highest score is the optimal tree. Figure 5.4 shows a list of trees named from Tree 01 to Tree 16. These trees are similar to the reference tree concerning the juxtapositions of categories 3,4, and 5. The reference tree is also on the list as Tree 07. Tree 02 has the juxtapositions and hierarchies of Tree 07, but different structure. These trees are tracked throughout the work-flow to check the performance of the output models with multiple measures of goodness.

5.3 Measures of Model Optimality

Categorical models should comply with traditional probabilistic checks. Reproduction of declustered proportions, experimental indicator variogram, trend model, and others must be correctly addressed. In addition, alternative metrics may be used depending on technical decisions and the goals of the engineering models. The measures of goodness assess different characteristics. There

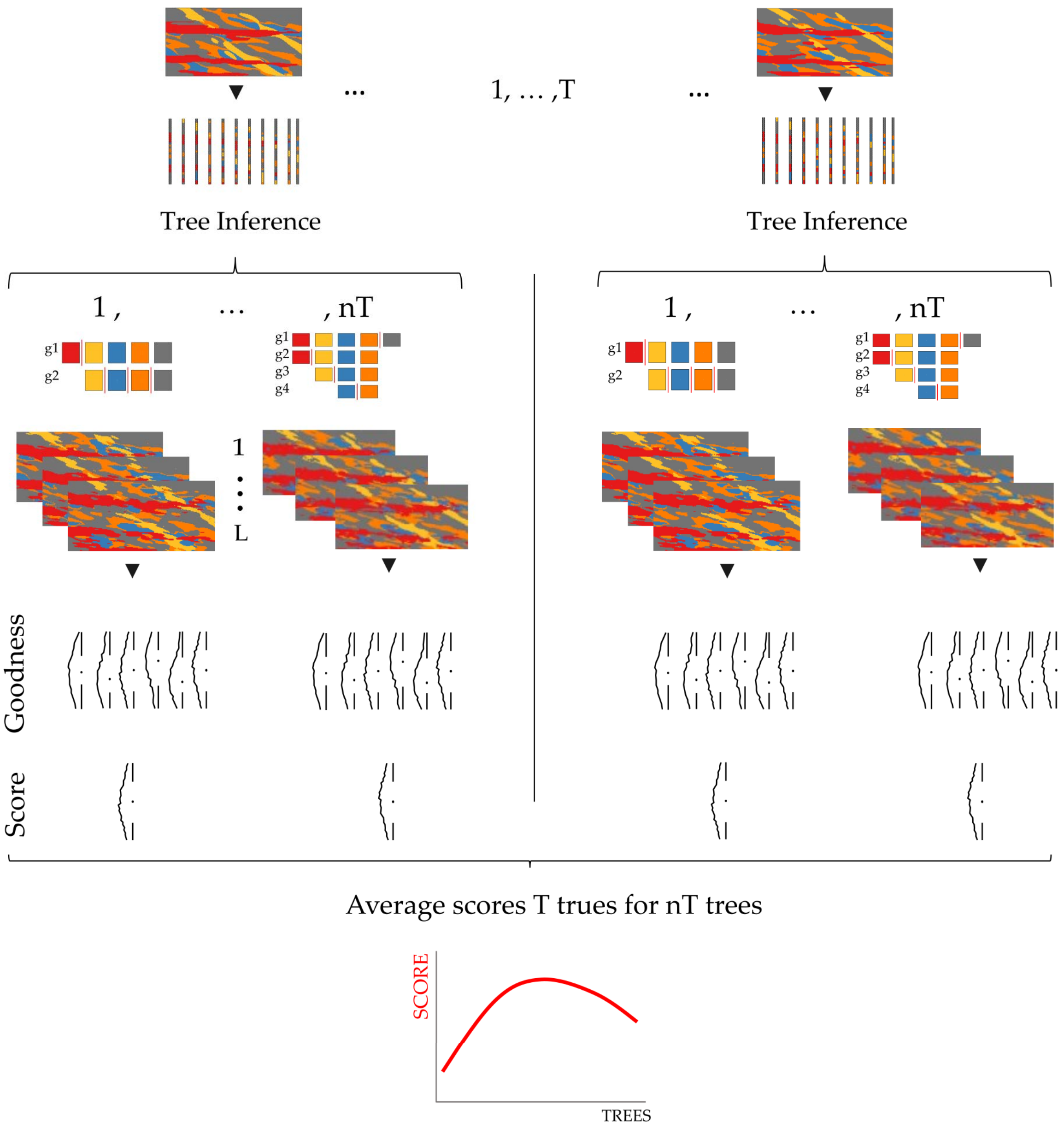


Figure 5.2: Evaluation workflow

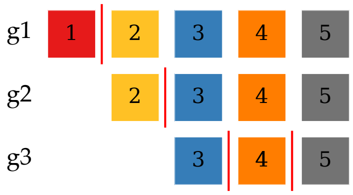
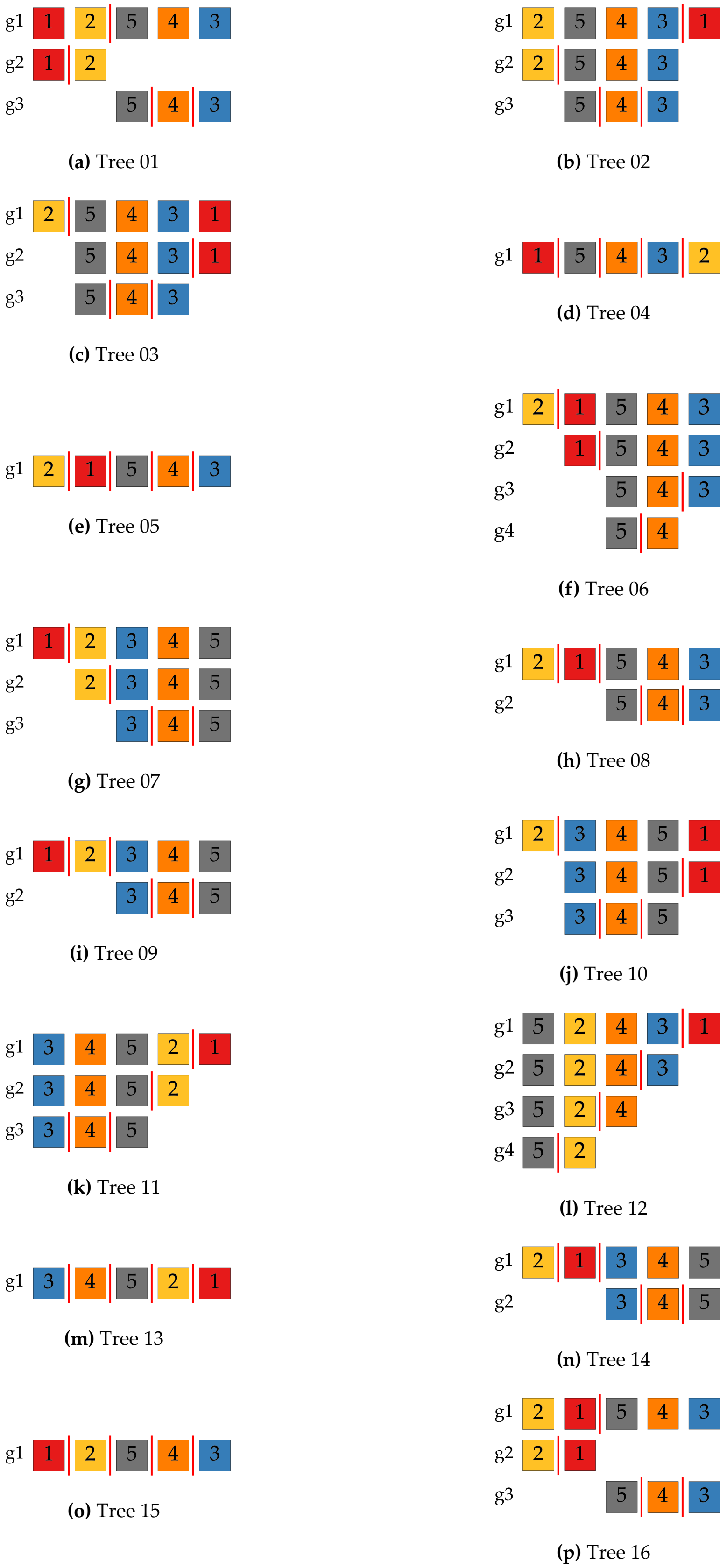


Figure 5.3: Reference truncation tree



(a) Tree 01

(b) Tree 02

(c) Tree 03

(d) Tree 04

(e) Tree 05

(f) Tree 06

(g) Tree 07

(h) Tree 08

(i) Tree 09

(j) Tree 10

(k) Tree 11

(l) Tree 12

(m) Tree 13

(n) Tree 14

(o) Tree 15

(p) Tree 16

Figure 5.4: Reduced set of possible truncation trees. The sixteen trees follow closely the associations from the reference tree. Structures with one to four latent Gaussian variables are considered.

is no clear methodology to compare results from trees in optimization. Trees are not ordinal, interval, or ratio variables. Several modeling parameters are affected including the derived variograms of latent Gaussian variables whenever the tree is changed. The structures condition the permitted associations resulting in completely different models. The measures of goodness in this section summarize model results for all trees.

The number of latent Gaussian variables in the example varies from one to four. With one latent Gaussian variable, the multiple juxtapositions are considered. Trees with more latent Gaussian variables are numerous. With four latent Gaussian variables, the contact relations are restricted to the ending leaf node. The plots used for the analysis of the measures of goodness show the metrics grouped by the number of latent Gaussian variables and sorted in ascending order from left to right. The optimal values of the measures of goodness, either the lowest or highest values are inspected to check if the tree used is similar to the reference.

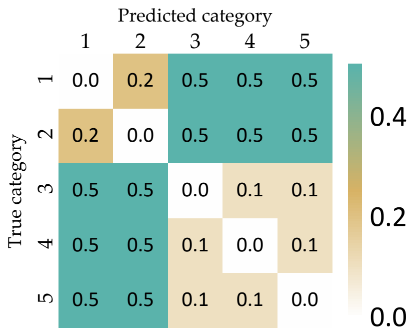
5.3.1 Penalty Matrix

Models are used to solve engineering problems. In a categorical model, some categories are more relevant in terms of economics or related processing issues. In ore-control, the impact of mismatched categories varies depending on the categories. A penalty matrix quantifies the relative importance of the mismatches. One option considers the dissimilarities, δ_{ij}, as penalties. A user-defined option considers the associations in the tree structures for the penalties, for instance, two distant categories in the tree hierarchy, in different latent Gaussian variables, or separated by contact rules in the same latent Gaussian variable have a higher penalty. Conversely, the penalty is less if the mismatch involves contiguous categories in the same latent Gaussian variable.

Figure 5.5 is a penalty matrix for the reference images. The rows represent the true categories and the columns are the predicted categories in the model. The values in the matrix are penalties based on the reference tree. In the example, categories 1 and 2 are considered mineralized units; categories 3,4, and 5 are non-mineralized. Category 1 crosscuts category 2; the mismatch between these categories is more likely in the intersection zone. The penalty between categories 1 and 2 is set with a low value of 0.2 as they are both mineralized. On the contrary, mismatches between either category 1 or 2 and categories 3,4, or 5 receive higher penalties as they cause economic impact. Mismatching categories 3,4 or 5 have the lowest penalty value as they are of no economic interest and belong to the same latent Gaussian variable in the tree.

The penalties are added over the grid points and averaged, Equation (5.1), where L is the number of realizations and N is the number of grid points. The goal is to minimize the expected penalty, $\bar{p}$.

$$\bar{p} = \frac{1}{L \times N} \sum_{l=1}^{L} \sum_{i=1}^{N} penalty_{i,l}(\text{predicted category}, \text{true category}) \tag{5.1}$$



	Predicted category 1	2	3	4	5
True category 1	0.0	0.2	0.5	0.5	0.5
2	0.2	0.0	0.5	0.5	0.5
3	0.5	0.5	0.0	0.1	0.1
4	0.5	0.5	0.1	0.0	0.1
5	0.5	0.5	0.1	0.1	0.0

Figure 5.5: Penalty matrix

Figure 5.6 summarizes the expected penalties for the trees. The x-axis shows the number of latent Gaussian variables. The indexes of the trees are omitted. With one latent Gaussian variable, the best results are related to trees with the sequence of categories 3-4-5 that matches the associations in the reference tree. Trees 04, 05,13, and 15 obtained low penalties compared to other possible trees with one latent Gaussian variable. This analysis is repeated for a different number of latent Gaussian variables. The best results were obtained for Trees 01, 02, and 03 which have tree latent Gaussian variables. Tree 02, which is a symmetric version of the reference tree, gives one of the best three results. Some untracked trees outperform the sixteen trees. Penalty matrices optimize trees based on specific categories and are case-based.

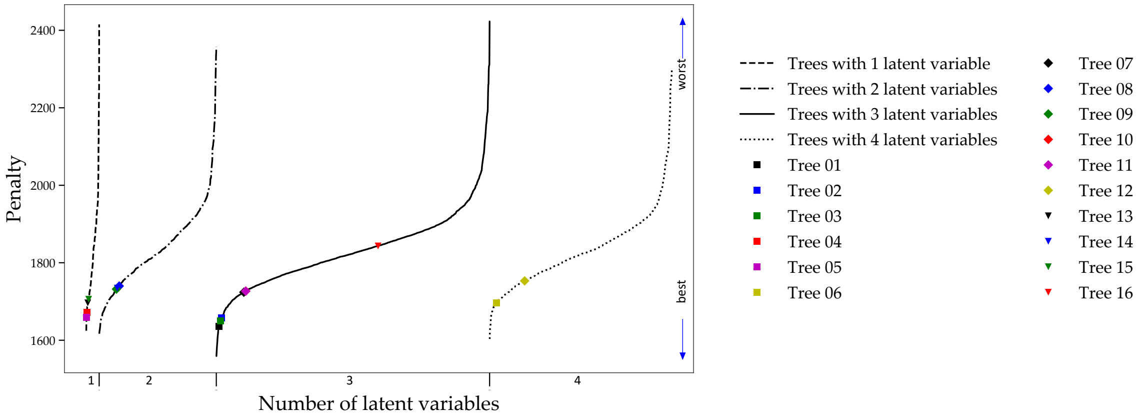


Figure 5.6: Penalty values for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. Y-axis represents penalties for different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows penalty values sorted from best to worst as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

5.3.2 Local Accuracy

Cross-validation is a critical step in geostatistical workflows, it helps in parameter tuning and the comparison of methodologies. The basic idea is to resample or use new data to compare predicted to true values. Checked variables may be continuous or categorical. The quality of the results is usually determined with H and B values. H denotes entropy, it characterizes the amount of information (Shannon, 1948). Entropy is the natural way to represent uncertainty in categorical variables, (J. L. Deutsch & Deutsch, 2012). In categorical variables, the information is represented by the probability of occurrence of categories at a certain location, p_k. In Equation (5.2), K is the number of categories. When there is a certainty of a value, the entropy is zero. The maximum entropy is achieved when all possible values from the set are equally probable. Lower entropies are desired. The average entropy, H_{avg}, Equation (5.3), is a summarizing metric of the local entropy where N is the number of grid cells in the model, and H_i is the entropy at the grid location i.

$$H = -\sum_{k=1}^{K} p_k \ln\left(p_k\right) \tag{5.2}$$

$$H_{avg} = \frac{1}{N}\sum_{i=1}^{N} H_i \tag{5.3}$$

B denotes the difference between the average predicted probability when the true value is 1 and the probability when the true value is zero. In Equation (5.4), K is the number of categories, N is the number of locations, n_{i_k} are the locations where the indicator i_k is 1 or 0, and p_{k,i_k} is the probability of category s_k when i_k is 1 or 0. Higher B values indicate that the presence or absence of a category is predicted correctly, (C. V. Deutsch, 2010; J. L. Deutsch & Deutsch, 2012). In practice, B is a better statistic than H.

$$B = \frac{1}{\sum\limits_{n=1}^{N} n_{i_k=1}} \sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,i_k=1} - \frac{1}{\sum\limits_{n=1}^{N} n_{i_k=0}} \sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,i_k=0} \tag{5.4}$$

In Figure 5.7, the lowest average entropy from the sixteen trees was obtained by Tree 04 with one latent Gaussian variable. With three latent Gaussian variables; Tree 01, 02, and 03 obtained the lowest results which resemble the results obtained with penalty values. The results are less optimal in trees with two and four latent Gaussian variables. Figure 5.8 summarizes the B values for the trees. In this case, the trees with three latent Gaussian variables gave the best results. Tree 02, which has the same associations and similar structure as the reference tree, significantly outperformed the other trees. Tree 07, the reference tree, shows also one of the highest B values. B obtained coherent results.

5.3.3 Variogram MSE

The Variogram Mean Square Error (VMSE) is the MSE between the variogram of categorical realizations and the experimental indicator variograms from drillholes calculated over multiple lags. The relevance of each category is included in weights based on the global proportions; lower-proportion categories have less impact on the total error.

A correct model should reproduce the variogram of the input data. During the HTPG work-

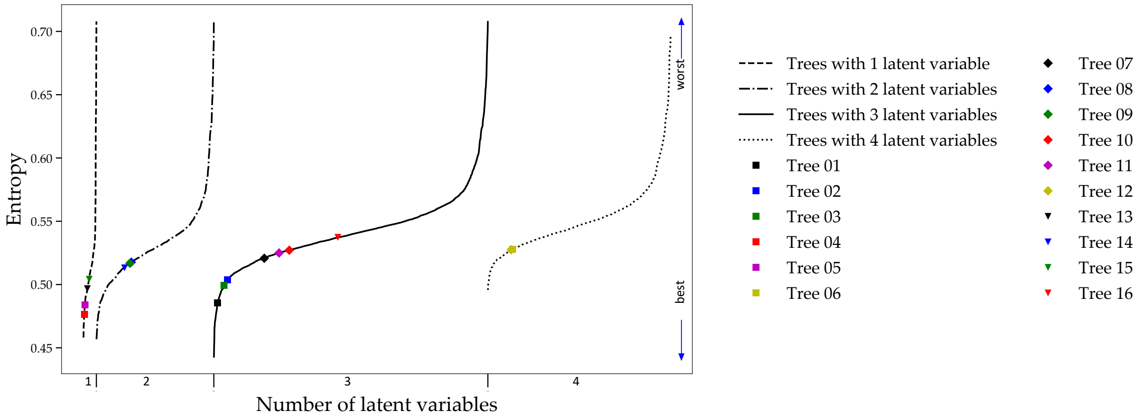


Figure 5.7: Entropy values for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the entropy values for different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows entropy values sorted from best to worst as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

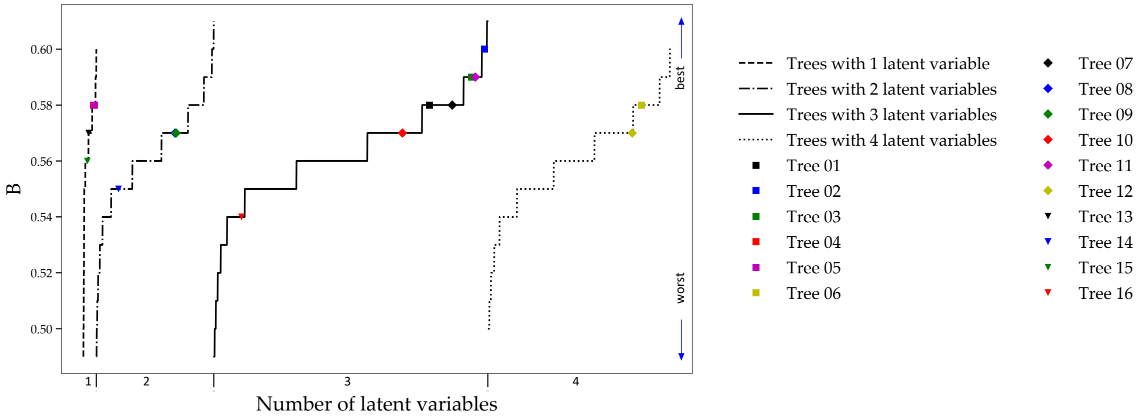


Figure 5.8: B values for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the B values for the different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows B values sorted from worst to best as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

flow, parameter tuning seeks to improve variogram reproduction. As shown previously, the performance of the numerical derivation may be compromised affecting the overall variogram reproduction. The results for different trees are shown in Figure 5.9 to check whether the error is minimized when a tree similar to the reference tree is used. The lowest errors or best results were obtained with Tree 15 with one latent Gaussian variable, and Tree 02 with three latents. The difference in VMSEs between Tree 02 and other trees with three latent variables is significant.

5.3.4 Transition Probabilities

Transition probabilities are used to measure the goodness of categorical models. Former researchers applied transition probabilities for optimization. Sadeghi and Boisvert (2012) considered the per-

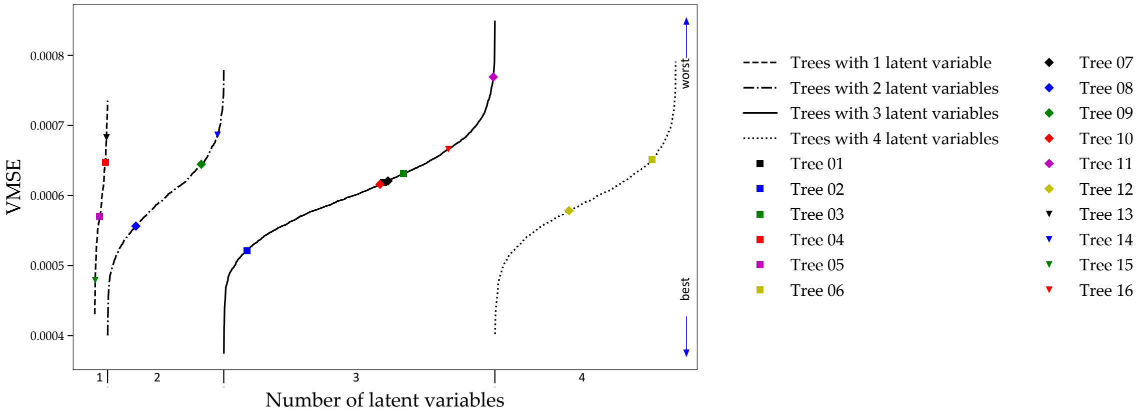


Figure 5.9: VMSE values between experimental indicator variograms and categorical realizations for trees with five categories. The X-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the VMSEs for different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows the VMSEs sorted from best to worst as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

turbation of a random initial truncation mask to minimize the difference between the input and model's TPM, and optimize the thresholds. J. L. Deutsch and Deutsch (2013) used transition probabilities to determine the truncation scheme using MDS. A simple approach is to choose the tree that minimizes the Transition Probability Error (TPE) between the reference and the results, Equation (5.5). Figure 5.10 shows the results with different trees. Unlike previous measures of goodness, results for Trees 01 to 16 are tight and optimal. The lowest errors were obtained in cases with three latent Gaussian variables.

$$TPE = \sqrt{\sum_{i=1}^{K}\sum_{j=1}^{K} \mid \mathrm{TPM}_{ij}^{\mathrm{true}} - \mathrm{TPM}_{ij}^{\mathrm{tree}} \mid} \tag{5.5}$$

ETPM is also considered for the analysis of transition probabilities. ETPMs can be obtained from TPMs by rescaling off-diagonal terms to one and setting diagonal terms to zero. The Embedded Transition Probability Matrix Error (ETPE) is calculated in the same fashion as Equation (5.5) by replacing TPM with ETPM, Figure 5.11 are the results. The difference between the performance of the trees is easily recognizable compared to using TPM errors. The ETPE for trees with three latent Gaussian variables obtained consistently the best results. In this case, Tree 07 gave the minimum ETPE.

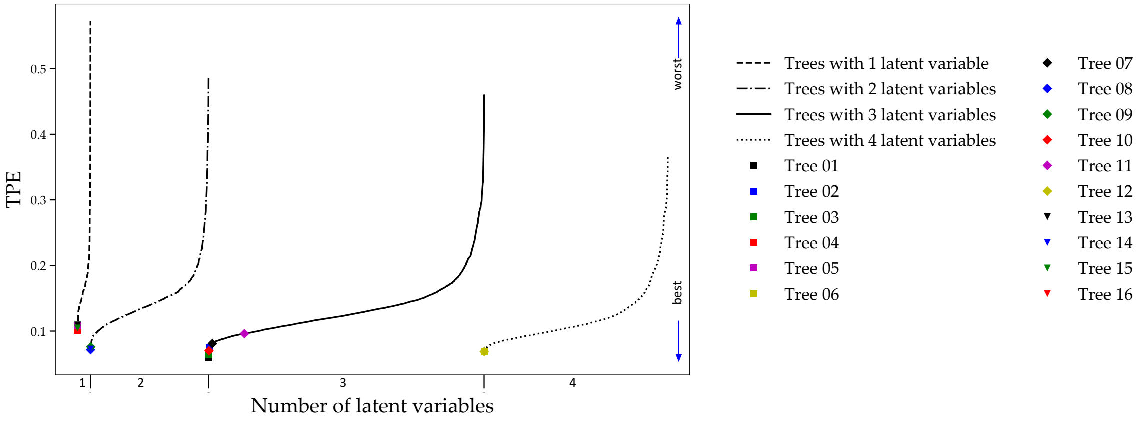


Figure 5.10: TPE values relative to a reference image for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the TPEs for the different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows TPEs sorted from best to worst as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

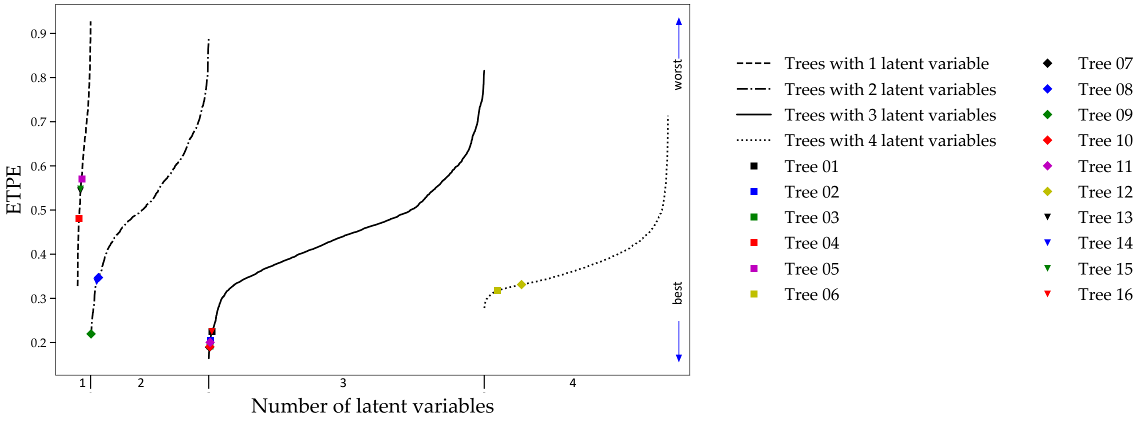


Figure 5.11: ETPE values relative to a reference image for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the ETPEs for the different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows ETPEs sorted from best to worst as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

5.3.5 Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC), (Matthews, 1975), measures the quality of binary classifications following Equation (5.6). It takes values from -1 for total disagreement to +1 for complete correspondence.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{5.6}$$

		Predicted	
		Positive	Negative
True	Positive	tp	fn
	Negative	fp	tn

Figure 5.12: Error Matrix in binary classification.

The MCC in binary cases based on the true and false (+) and (-) from the error matrix, Figure 5.12 is adapted to multiclass cases as denoted in Equation (5.7), t_k is the number of times that a category k truly occurred, e_k is the number of times that k was predicted, c is the number of correct predictions, and s is the number of locations to evaluate. The minimum MCC value in a multiclass case varies from -1 to 0, and the maximum value is 1. Figure 5.13 shows the MCC values obtained for different trees. The best MCC values were obtained for cases with three latent Gaussian variables. Tree 02 gives the best result overall.

$$\text{MCC} = \frac{c \times s - \sum_k^K e_k \times t_k}{\sqrt{\left(s^2 - \sum_k^K e_k^2\right) \times \left(s^2 - \sum_k^K t_k^2\right)}} \tag{5.7}$$

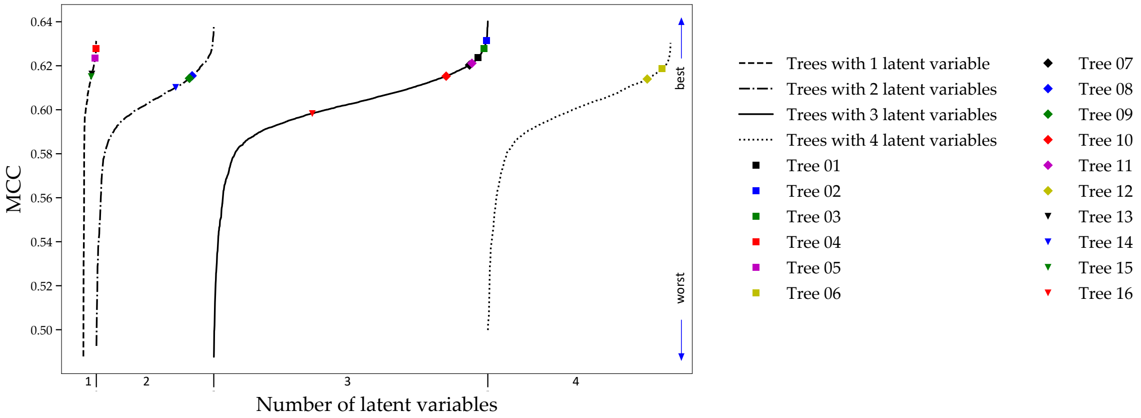


Figure 5.13: Matthews correlation coefficient relative to a reference image for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the MCC values for the different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows MCC values sorted from worst to best as indicated by the blue arrows. Sixteen trees with their respective labels according to Figure 5.4 are also plotted.

Truncation trees are not numerically relatable to each other, therefore a correlation between trees and measures of goodness is not possible. Figure 5.14 shows the coefficient of variations in results of the measures of goodness for all trees.

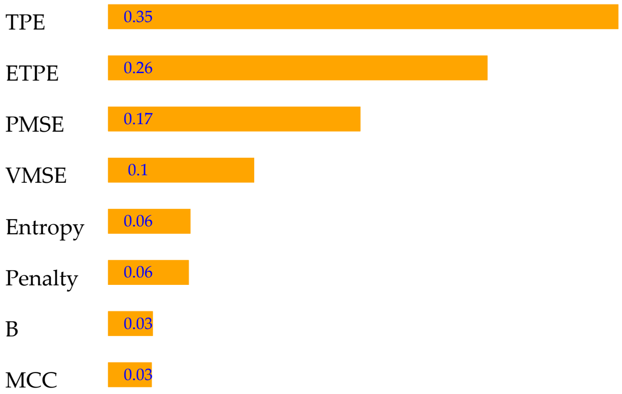
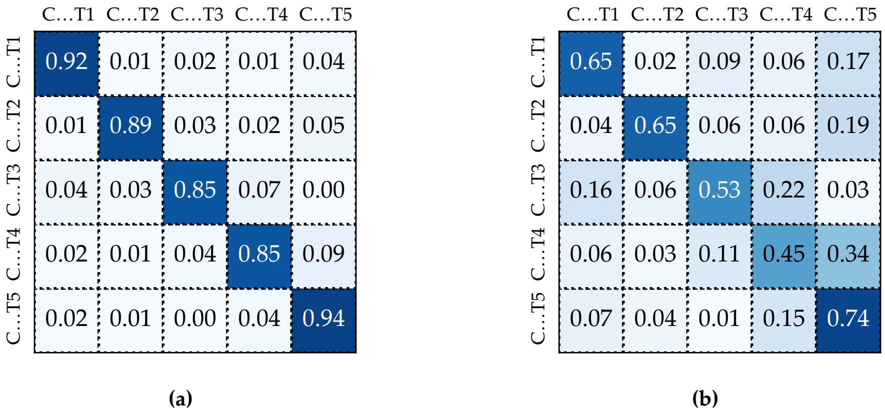


Figure 5.14: Coefficients of variation in the results of measures of goodness for trees

5.4 Measures of Goodness without the Reference Image

Practical applications require the optimal tree to be inferred from limited data. The MSE between the experimental variograms and the variograms from the categorical models is one alternative, but it relies on subjective tolerance parameters during experimental variogram calculation. A penalty matrix requires knowledge of the truncation tree; B and MCC require validation data. More automatic and practical measures of goodness concerning tree optimization should be preferred. TPMs and ETPMs are directly calculated from drillholes, therefore TPEs and ETPEs are strong alternatives.

Figure 5.15 shows the TPM of a reference image and the respective TPM from drillholes. The matrices are significantly different due to the influence of diagonal terms. The norm of the difference between these matrices is 0.78. Figure 5.16 shows the ETPM of a reference image and the respective ETPM from drillholes. The norm of the difference between them is 0.21 and the matrices are similar.



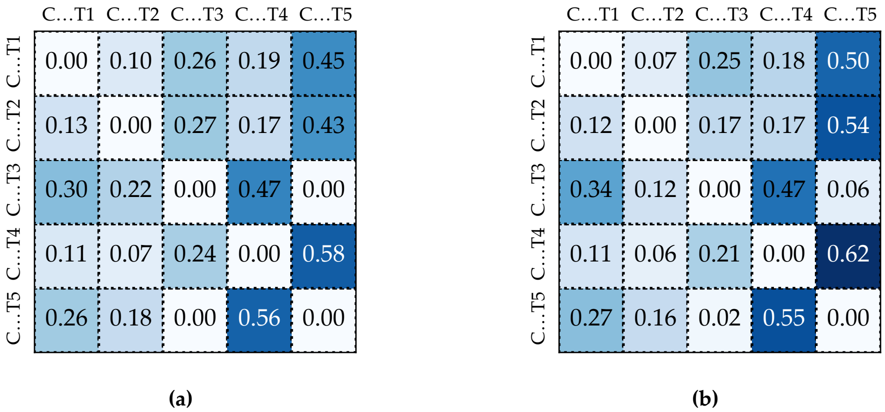
	C...T1	C...T2	C...T3	C...T4	C...T5
C...T1	0.92	0.01	0.02	0.01	0.04
C...T2	0.01	0.89	0.03	0.02	0.05
C...T3	0.04	0.03	0.85	0.07	0.00
C...T4	0.02	0.01	0.04	0.85	0.09
C...T5	0.02	0.01	0.00	0.04	0.94

(a)

	C...T1	C...T2	C...T3	C...T4	C...T5
C...T1	0.65	0.02	0.09	0.06	0.17
C...T2	0.04	0.65	0.06	0.06	0.19
C...T3	0.16	0.06	0.53	0.22	0.03
C...T4	0.06	0.03	0.11	0.45	0.34
C...T5	0.07	0.04	0.01	0.15	0.74

(b)

Figure 5.15: (a) TPM from one reference image (b) TPM from drillholes.



	C...T1	C...T2	C...T3	C...T4	C...T5
C...T1	0.00	0.10	0.26	0.19	0.45
C...T2	0.13	0.00	0.27	0.17	0.43
C...T3	0.30	0.22	0.00	0.47	0.00
C...T4	0.11	0.07	0.24	0.00	0.58
C...T5	0.26	0.18	0.00	0.56	0.00

(a)

	C...T1	C...T2	C...T3	C...T4	C...T5
C...T1	0.00	0.07	0.25	0.18	0.50
C...T2	0.12	0.00	0.17	0.17	0.54
C...T3	0.34	0.12	0.00	0.47	0.06
C...T4	0.11	0.06	0.21	0.00	0.62
C...T5	0.27	0.16	0.02	0.55	0.00

(b)

Figure 5.16: (a) ETPM from one reference image (b) ETPM from drillholes.

ETPE is a robust measure of performance. Figure 5.17 shows ETPE results using drillholes as a benchmark instead of the reference images. The analysis with drillholes led to similar conclusions compared to using ETPE with the images, where Tree 07 achieves gain the minimum ETPE. Overall, the robustness of the results depends on the representativity of the drillholes. ETPMs and share one characteristic, they both perform calculations in the transitions, either transition of state or interval. The idea of using Interval probabilities as a measure of goodness is appealing, however, it was not included in the work-flow due to the high computation time required.

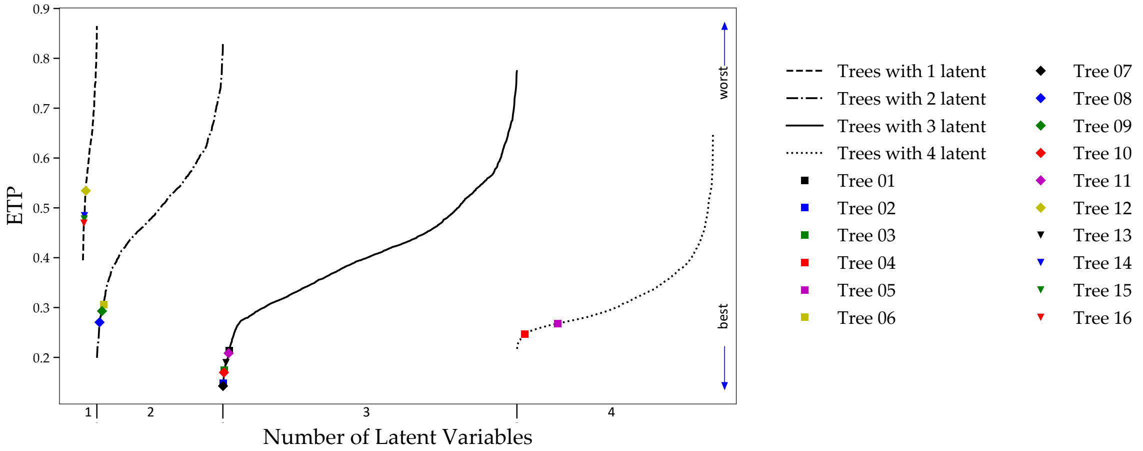


Figure 5.17: ETPE values relative to drillholes for truncation trees with five categories. The x-axis represents the number of latent Gaussian variables in the trees. The y-axis represents the ETPE values for the different trees. Truncation trees are grouped by the number of latent Gaussian variables and are represented by the black lines; each group shows ETPE sorted from best to worst as indicated by the blue arrows.

5.5 Scoring Trees

The results from the measures of goodness must be reduced to a single value to ease the decision. Min-max scaling was used to standardize the metrics between 0 and 1, where 1 represents the optimal value, then the final score is the summation. Table 5.1 shows the scores for sixteen truncation trees and includes results from SIS. The reference tree, Tree 07, obtained the seventh-highest score, and Tree 02, which is similar to the reference, obtained the second-highest score. Dealing with multiple measures of goodness might be misleading.

Table 5.2 shows only ETPE scores. In this case. The highest-scored is the reference tree followed by Tree 10 that is close to the reference. In practice, the results suggest that ETPEs should be included as measure of goodness to optimize trees. The other measures of goodness may be included with careful interpretation at the time of selecting a tree. Figure 5.18 shows the selected truncation trees in this analysis.

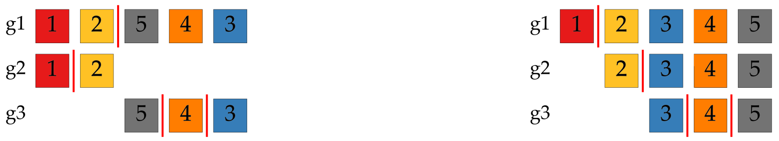
A summary plot for the selection of an optimal tree with all trees mapped may be confusing. Plotting only a few of the best trees with a different number of latent Gaussian variables in the x-axis and the scores in the y-axis is simpler, Figure 5.19.

Table 5.1: Scores of measures of goodness for tree selection in HTPG. Scores $\in [0, 7]$. Results include sixteen truncation trees and SIS.

Tree ID	B	Entropy	VMSE	Penalty	TPE	MCC	PMSE	**Score**
Tree 01	0.58	0.515	0.994	0.836	0.941	0.624	0.992	5.482
Tree 02	0.60	0.496	0.995	0.834	0.926	0.631	0.991	5.474
Tree 03	0.59	0.501	0.994	0.835	0.935	0.628	0.991	5.473
Tree 04	0.58	0.524	0.994	0.833	0.899	0.628	0.993	5.449
Tree 05	0.58	0.516	0.994	0.834	0.892	0.623	0.996	5.435
Tree 06	0.58	0.472	0.993	0.830	0.931	0.619	0.992	5.418
Tree 07	0.58	0.479	0.994	0.828	0.919	0.620	0.992	5.412
Tree 08	0.57	0.482	0.994	0.826	0.928	0.615	0.993	5.409
Tree 09	0.57	0.483	0.994	0.827	0.924	0.614	0.993	5.405
Tree 10	0.57	0.473	0.994	0.827	0.930	0.615	0.993	5.402
Tree 11	0.59	0.475	0.992	0.827	0.904	0.621	0.991	5.401
Tree 12	0.57	0.473	0.994	0.825	0.931	0.614	0.993	5.400
Tree 13	0.57	0.503	0.993	0.830	0.891	0.616	0.996	5.399
Tree 14	0.55	0.487	0.993	0.827	0.927	0.610	0.993	5.387
Tree 15	0.56	0.496	0.995	0.829	0.895	0.615	0.994	5.384
Tree 16	0.54	0.463	0.993	0.816	0.932	0.598	0.991	5.334
SIS	0.67	0.734	0.560	0.853	0.694	0.669	0.937	5.117

Table 5.2: ETPE scores for tree selection in HTPG. Scores $\in [0, 1]$. Results include sixteen truncation trees.

Tree ID	N. Latents	**ETPE Score**
Tree 07	3	0.811
Tree 10	3	0.809
Tree 03	3	0.807
Tree 11	3	0.800
Tree 02	3	0.796
Tree 09	2	0.781
Tree 01	3	0.775
Tree 16	3	0.775
Tree 06	4	0.683
Tree 12	4	0.669
Tree 14	2	0.662
Tree 08	2	0.653
Tree 04	1	0.519
Tree 13	1	0.456
Tree 15	1	0.452
Tree 05	1	0.430



(a) Tree 01

(b) Tree 07

Figure 5.18: Selected trees based on measures of goodness. (a) Using all measures of goodness. (b) Using ETPEs, the selected tree matches the reference.

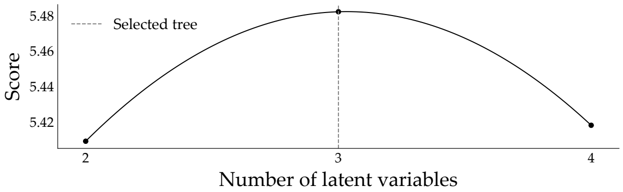


Figure 5.19: Truncation tree selection plot using scores. The x-axis represents the number of latent Gaussian variables in the trees. The dots represent the highest-scored trees depending on the number of latent Gaussian variables.

5.6 Conclusions

Selecting a truncation tree in HTPG is time-demanding. The analysis presented focuses on testing all trees and measures of goodness to identify metrics that correctly optimize the tree. The ordering of the categories showed more impact on the values of the measures of goodness than the number of latent Gaussian variables. The best results were related to trees with the correct structure which is highly constrained to using the correct number of latent Gaussian variables in the tree. This was consistently observed in the case of the penalty matrix, B value, TPE, ETPE, and MCC; however, ETPE was decisive. After identifying practical measures of goodness, they can be applied to a reduced set of inferred trees. The inference tools discussed in Chapter 4 and the results are summarized in the following steps for tree selection: (1) perform HTPG on a reduced set of inferred trees, (2) calculate the scores using measures of goodness such as ETPE to select the optimal tree. Finally, the correct tree optimizes variogram reproduction. Wrong trees give poor variogram reproduction. Transition probabilities and indicator variograms are numerically related, (Carle & Fogg, 1996), therefore minimizing TPM errors would lead to better indicator variogram reproduction.

Chapter 6

Case Study: Categorical Modeling at Mesaba Deposit

The concepts outlined in previous chapters are included in the categorical modeling of the Mesaba deposit, the results from HTPG are also compared to SIS which could be considered standard practice.

6.1 Background

The Mesaba deposit is a Cu-Ni deposit located in the emerging Duluth Mineral District in St. Louis County, Northern Minnesota. It is one of several Cu-Ni sulfide deposits within the Duluth Complex in the trend of existing mines in the Mesabi Iron Range. The mineralization in the Mesaba deposit presents medium to coarse-grained disseminated chalcopyrite, cubanite, pentlandite, and pyrrhotite, (Mayhew, Mean, O'Connor, & Williams, 2009).

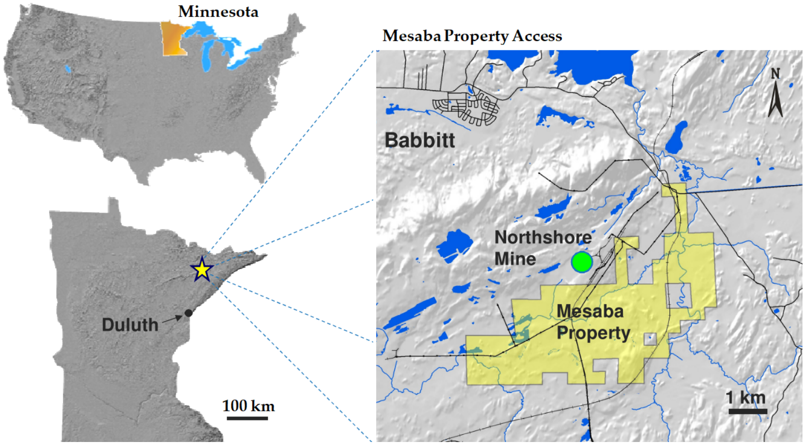


Figure 6.1: Location of Teck's Mesaba property. Taken from Mayhew et al. (2009).

6.2 Data Set

The Duluth Complex Database (DCD) contains drillhole locations, lithological descriptions, copper and nickel assay data, and rock quality gathered in 2,145 exploration drillholes in the Duluth Complex region. There are approximately 1,779,600 feet drilled (Patelke, 2003).

Figure 6.2 encloses the drillholes within the Mesaba deposit region. This region comprises 40,831 interval data with 46 different categorical codes. Grouping and filtering of categories were performed considering the representativity, J. L. Deutsch (2015). The criteria require the categories to

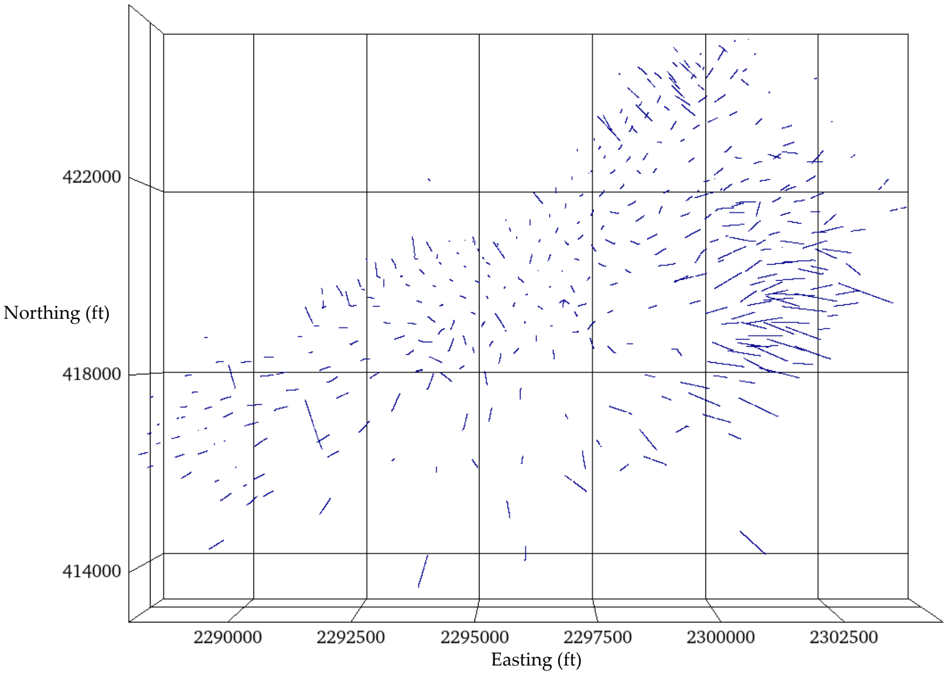


Figure 6.2: Mesaba deposit drillhole data.

be: (1) geologically similar, (2) statistically similar, and (3) spatially similar. Low-proportion categories such as thin layers are grouped with others. Ten categories were determined and geological codes were assigned to the existing categorical labels according to Table 6.1. Figure 6.3 shows the drillholes with the assigned geocodes.

Table 6.1: Geocodes used to replace DCD categorical labels.

Label in DCD	Geocode
3	CAT 01
7	CAT 02
6	CAT 03
5	CAT 04
BTLS	CAT 05
4	CAT 06
1	CAT 07
1S1	CAT 08
VF	CAT 09
BIF	CAT 10

The horizontal dimensions of the grid cells were determined by inspecting the spacing between drillholes, Figure 6.4. The grid cells in the model are aligned to the East and North. Table 6.2 specifies the grid model parameters. The boundary of the model was determined with a data spacing model with a horizontal search radius of 1,000x1,000 ft.

Drillholes are spatially correlated, then a random partitioning of data by sample or by drillhole is not the best option. Assigning the data from one drillhole to different folds is also unreasonably optimistic due to the high proximity at the time of prediction. The drillholes were partitioned into

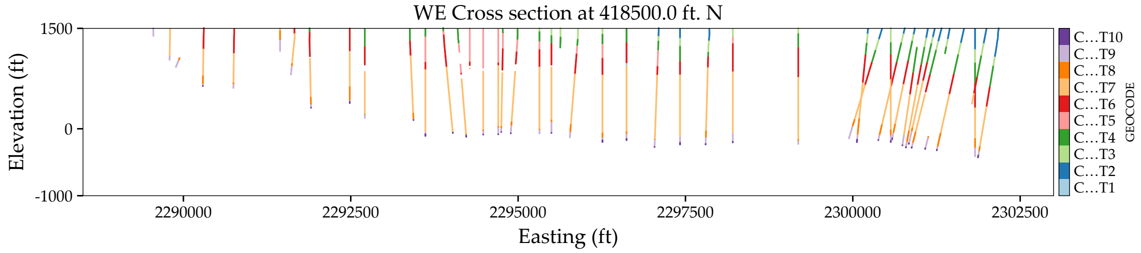


Figure 6.3: Cross-section of Mesaba deposit drillhole dataset. The clipping tolerance is +/- 400 ft.

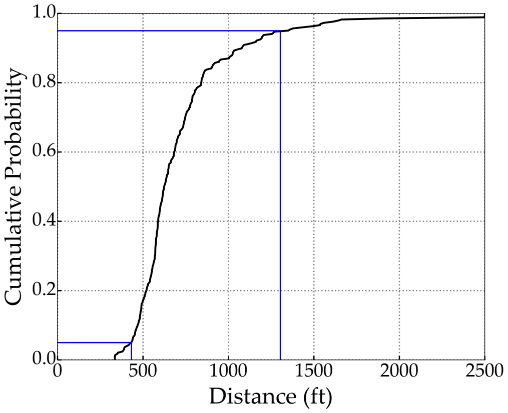


Figure 6.4: Distribution of drillhole spacing in Mesaba dataset. The 5th percentile assists in the selection of initial lags. The 95th percentile assists in determining the offset for model extent.

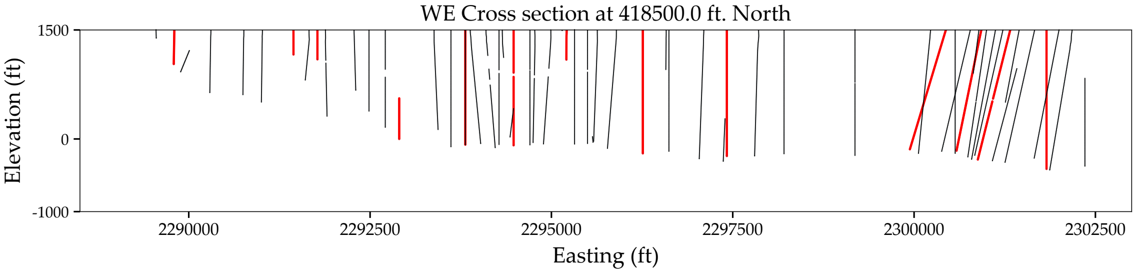
Table 6.2: Grid model definition.

Orientation	Origin (ft)	Number of Cells	Cell Size (ft)
Easting	2287980	72	250.0
Northing	413475	45	250.0
Elevation	-1855	86	40.0

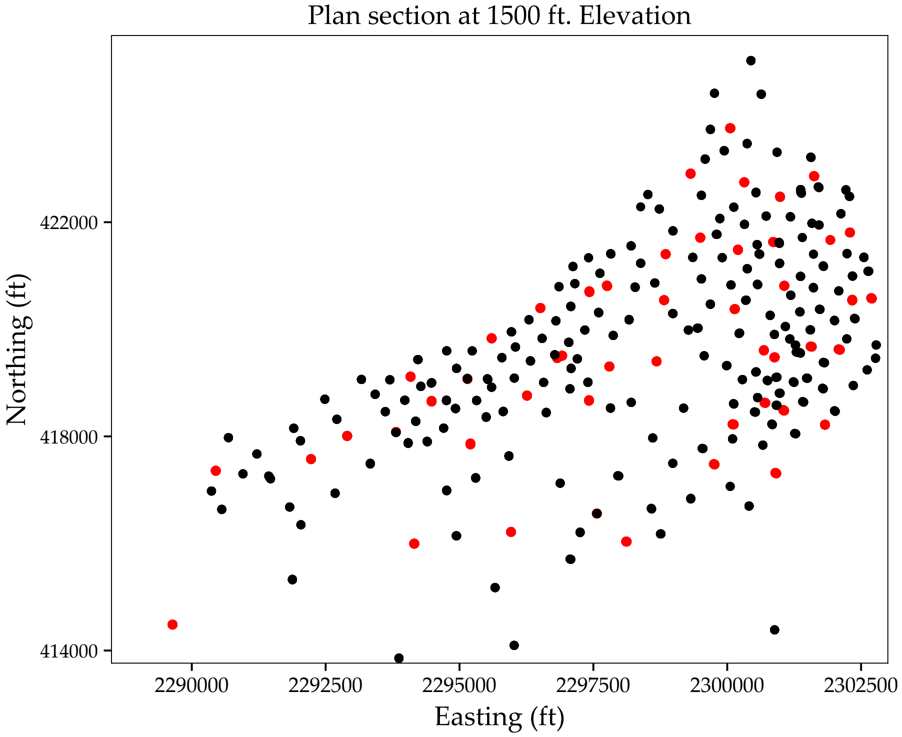
five folds, (C. V. Deutsch, 2018), and modeled with HTPG. Figure 6.5 shows the drillholes from one partitioning, the black lines represent the training data and the red lines represent the validation data. The true values and predictions from different folds are later combined for analysis.

6.2.1 Categorical Proportions

The distribution of the categories impacts directly the resource estimation studies. Any bias in the proportions must be addressed. The studied categories are highly non-stationary as shown in Figure 6.3. The proportion reproduction in models that use local proportions improves if the global proportions from the trends are close to the target global proportions. Trend models also allow to obtain correct transitions of categories in under-sampled areas. The trend models were calculated considering the anisotropy ranges of the Category 09 for the search radius. Subsequently, five Gaussian filter passes were applied as a post-processing step. Figure 6.6 shows the trend models for the Mesaba deposit, the plotted drillhole traces are the projection into the section with a 400 m.



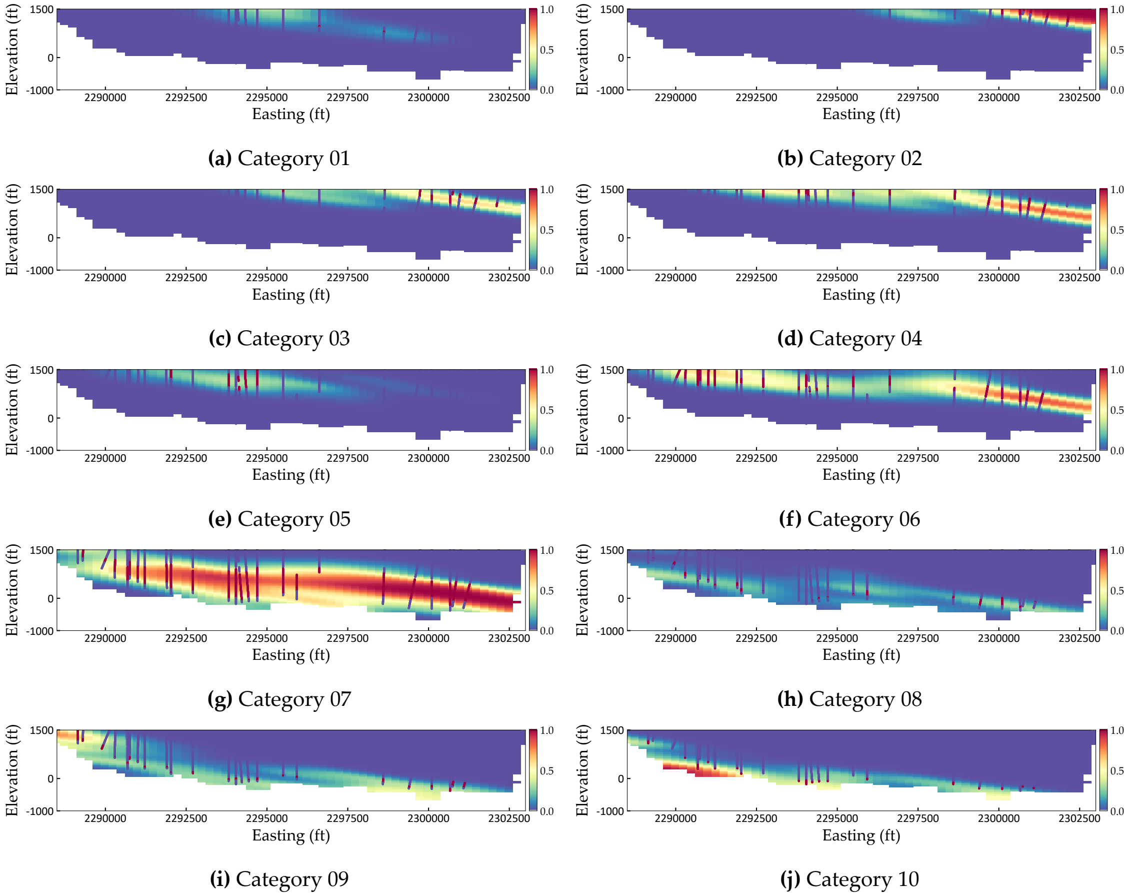
(a)



(b)

Figure 6.5: Cross-section of Mesaba deposit showing one training and validation dataset. The black solid lines represent the training drillhole data and the red ones represent the validation drillhole data. (a) Vertical section with a clipping tolerance of +/- 400 ft. (b) Plan section with a clipping tolerance of +/- 50 ft.

tolerance. Table 6.3 shows the global proportions without declustering weights, the declustered global proportions using an NN algorithm, and the proportions obtained from the trend models. The declustered proportions differ from the proportions without declustering weights for Category 01 and Category 10 values; these categories are found respectively at the top and bottom of the grid model, which is the reason for these overestimated declustered proportions. The categorical proportions obtained from the trend models are similar to the declustered target proportions.



(a) Category 01

(b) Category 02

(c) Category 03

(d) Category 04

(e) Category 05

(f) Category 06

(g) Category 07

(h) Category 08

(i) Category 09

(j) Category 10

Figure 6.6: Local proportions. The vertical section at 418,500.0 ft. North with a clipping tolerance of +/- 400 m. Indicator data from composites is also included with the same color legend.

6.2.2 Variography

The indicator anisotropies are aligned to the geological setting which is a tabular deposit. In tabular deposits, the horizontal directions are aligned to the plane of major continuity. Both vertical and horizontal directions are typically well-defined. The major and intermediate directions of anisotropy are parallel to the depositional layers. To account for the limited continuity in the vertical, tolerances parameters including dip tolerance and vertical bandwidths are restricted, J. L. Deutsch (2015). The experimental indicator variograms are not directly used in the calculation of indicator residuals. The residuals for each category calculated using the local proportions and indicators

Table 6.3: Global categorical proportions in Mesaba data set

Geocode	Global Proportion		
	Clustered	Declustered	Trend Model
Category 01	0.021	0.050	0.057
Category 02	0.026	0.044	0.046
Category 03	0.041	0.058	0.060
Category 04	0.082	0.083	0.087
Category 05	0.028	0.016	0.015
Category 06	0.127	0.123	0.119
Category 07	0.513	0.415	0.390
Category 08	0.057	0.058	0.047
Category 09	0.083	0.104	0.111
Category 10	0.018	0.046	0.065

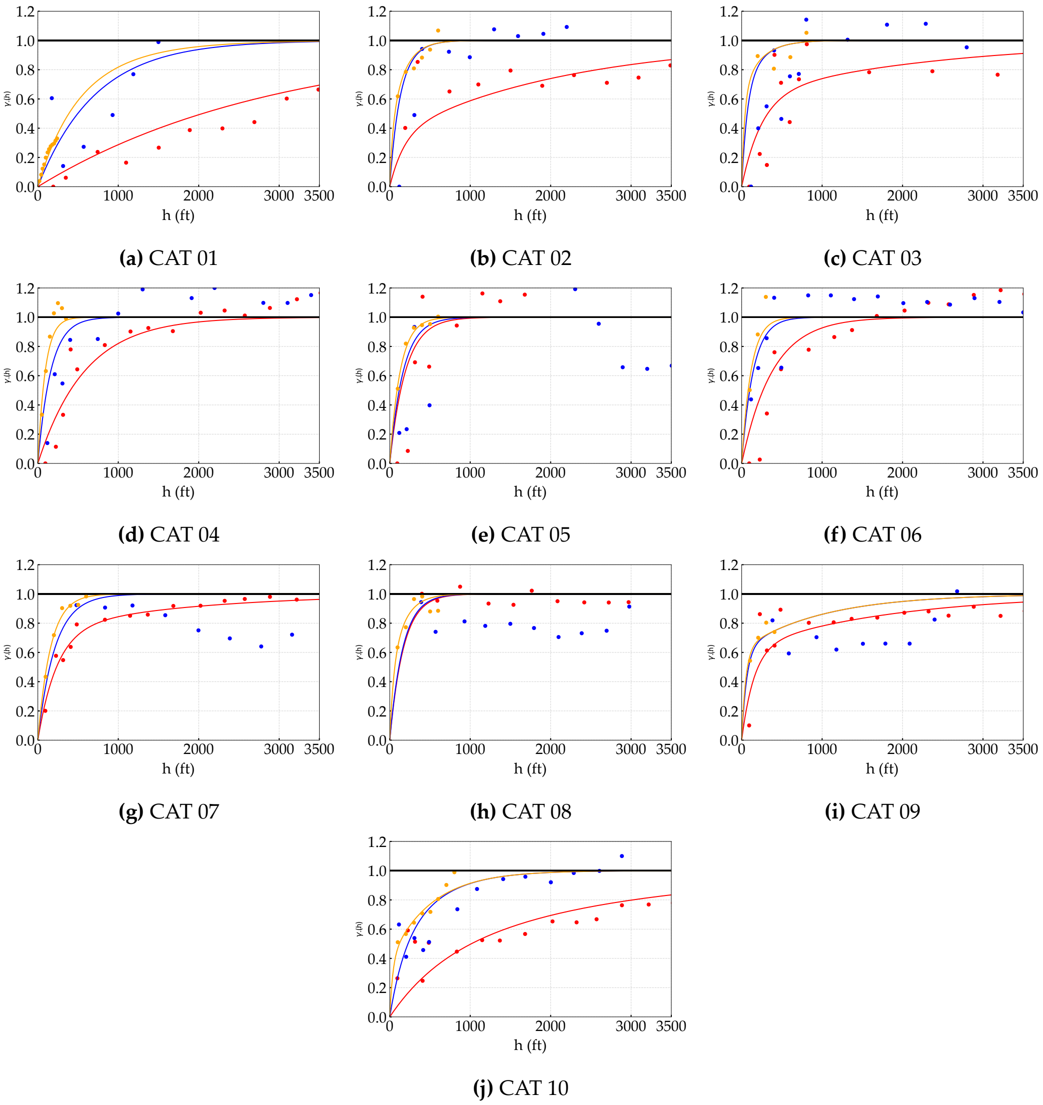
allow to parameterize the spatial structure. Figure 6.7 shows the experimental and modeled variograms of the indicator residuals.

6.3 Truncation Tree Inference

Truncated Gaussian techniques use the truncation tree for the conversion between categories and Gaussian values. The geological associations, the variography, and the proportions of categories define a truncation tree. At first, all drillholes were used for the inference of the tree. Further iterative analyses were also performed on representative drillholes to reduce the influence of local changes in the geological associations. To do so, the drillholes with the most number of different categories and the number of data were isolated. In general, the geological associations must be consistent over the area of study, and minor variations are handled by the trends. Cross-cutting categories also alter the results of the inference process. Filtering out those categories once recognized is a good option. The process of filtering categories and drillholes is iterative until consistent associations, orderings, and crosscutting geometries are obtained.

6.3.1 Transition Probabilities

Transition probabilities assist in understanding geological contacts and for the checking of categorical models. Figure 6.8a shows the TPM for the Mesaba dataset, equal composite lengths were used for the different categories. The analysis of a TPM by visual inspection may be not practical with more categories and complex geological settings. In this dataset, the ETPM is visually more practical than TPM. A sequence of the categories from category 02 to category 10 is shown in Figure 6.8b. However, category 01 is not aligned with the sequence.



(a) CAT 01 **(b)** CAT 02 **(c)** CAT 03 **(d)** CAT 04 **(e)** CAT 05 **(f)** CAT 06 **(g)** CAT 07 **(h)** CAT 08 **(i)** CAT 09 **(j)** CAT 10

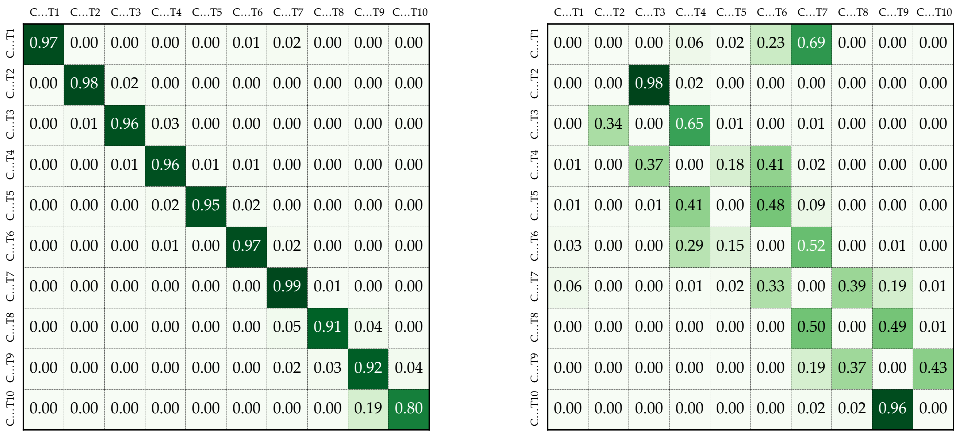
Figure 6.7: Modeled variograms of indicator residuals. The markers are experimental variogram values and the lines are fitted variogram models. The red color is used for the major anisotropy, the blue color is for the minor anisotropy and the orange color is for the anisotropy in the vertical.

6.3.2 Dissimilarity Matrix

Figure 6.9 shows the dissimilarities obtained from the interval data. With a high number of categories, the analysis of any distance matrix is not evident. Shading in the image is used to facilitate the visualization of closer categories based on relative dissimilarities. Categories 7, 8, 9, and 10 are closer to each other. Categories 2, 3, 4, 5, and 6 are more distant from the previously mentioned categories. Category 1 is closer to category 7 than to any other category.

6.3.3 MST and SLCA

The dendrogram built with SLCA summarizes the relative dissimilarities between categories. The associations retrieved from the dendrogram are used, however, the hierarchical structure requires



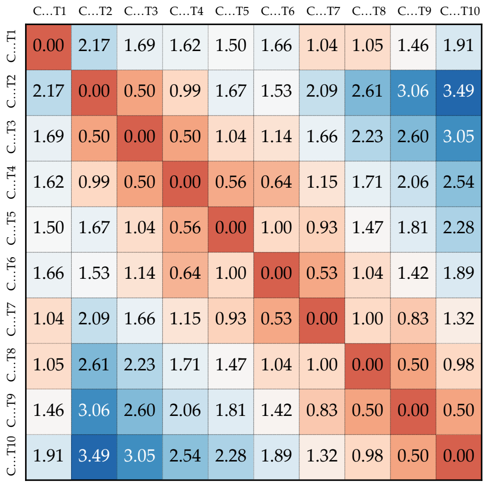
	C...T1	C...T2	C...T3	C...T4	C...T5	C...T6	C...T7	C...T8	C...T9	C...T10
C...T1	0.97	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00
C...T2	0.00	0.98	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C...T3	0.00	0.01	0.96	0.03	0.00	0.00	0.00	0.00	0.00	0.00
C...T4	0.00	0.00	0.01	0.96	0.01	0.01	0.00	0.00	0.00	0.00
C...T5	0.00	0.00	0.00	0.02	0.95	0.02	0.00	0.00	0.00	0.00
C...T6	0.00	0.00	0.00	0.01	0.00	0.97	0.02	0.00	0.00	0.00
C...T7	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00
C...T8	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.91	0.04	0.00
C...T9	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.92	0.04
C...T10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.80

(a)

	C...T1	C...T2	C...T3	C...T4	C...T5	C...T6	C...T7	C...T8	C...T9	C...T10
C...T1	0.00	0.00	0.00	0.06	0.02	0.23	0.69	0.00	0.00	0.00
C...T2	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.00	0.00	0.00
C...T3	0.00	0.34	0.00	0.65	0.01	0.00	0.01	0.00	0.00	0.00
C...T4	0.01	0.00	0.37	0.00	0.18	0.41	0.02	0.00	0.00	0.00
C...T5	0.01	0.00	0.01	0.41	0.00	0.48	0.09	0.00	0.00	0.00
C...T6	0.03	0.00	0.00	0.29	0.15	0.00	0.52	0.00	0.01	0.00
C...T7	0.06	0.00	0.00	0.01	0.02	0.33	0.00	0.39	0.19	0.01
C...T8	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.49	0.01
C...T9	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.37	0.00	0.43
C...T10	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.96	0.00

(b)

Figure 6.8: TPM and ETPM of Mesaba deposit dataset. (a) TPM calculated from composites. (b) ETPM calculated from composites.



	C...T1	C...T2	C...T3	C...T4	C...T5	C...T6	C...T7	C...T8	C...T9	C...T10
C...T1	0.00	2.17	1.69	1.62	1.50	1.66	1.04	1.05	1.46	1.91
C...T2	2.17	0.00	0.50	0.99	1.67	1.53	2.09	2.61	3.06	3.49
C...T3	1.69	0.50	0.00	0.50	1.04	1.14	1.66	2.23	2.60	3.05
C...T4	1.62	0.99	0.50	0.00	0.56	0.64	1.15	1.71	2.06	2.54
C...T5	1.50	1.67	1.04	0.56	0.00	1.00	0.93	1.47	1.81	2.28
C...T6	1.66	1.53	1.14	0.64	1.00	0.00	0.53	1.04	1.42	1.89
C...T7	1.04	2.09	1.66	1.15	0.93	0.53	0.00	1.00	0.83	1.32
C...T8	1.05	2.61	2.23	1.71	1.47	1.04	1.00	0.00	0.50	0.98
C...T9	1.46	3.06	2.60	2.06	1.81	1.42	0.83	0.50	0.00	0.50
C...T10	1.91	3.49	3.05	2.54	2.28	1.89	1.32	0.98	0.50	0.00

Figure 6.9: Dissimilarity matrix of Mesaba deposit.

careful analysis before considering it in trees. In Figure 6.10, three main associations of categories are present: 6-7, 2-3-4-5, and 8-9-10. Category 1 is distant from the others, this suggests that category 01 might cross-cut some of the clusters but not necessarily all.

6.3.4 Spectral Partitioning

Spectral partitioning of a graph summarizes the closeness between the categories. Figure 6.11a shows the values of the eigenvalues of Fiedler's vector. Figure 6.11b shows the results from spectral partitioning. The ten categories are separated into two groups. The first cluster considers categories 2, 3, 4, 5, and 6; the second cluster considers categories 1, 7, 8, 9, and 10. From the dissimilarity matrix, category 1 is closer to category 7 than category 6. The inspection of the drillholes validates these results, category 1 is inside or cross-cut category 7; in some drillholes, it cross-cuts category 6.

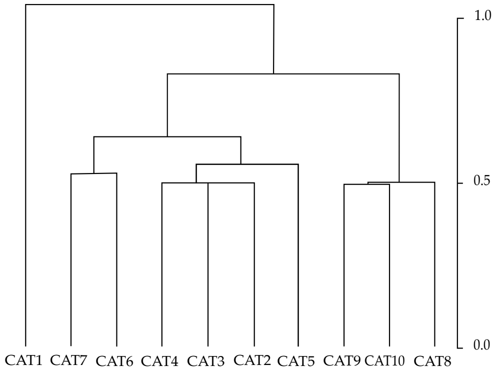


Figure 6.10: Dendrogram built from the dissimilarity matrix using SLCA.

Spectral partitioning is more robust than the SLCA for detecting associations of categories, however, the dendrogram from SLCA is valuable for first calculations. Usually, the first iteration in spectral partitioning gives a more correct idea of categorical associations than subsequent iterations.

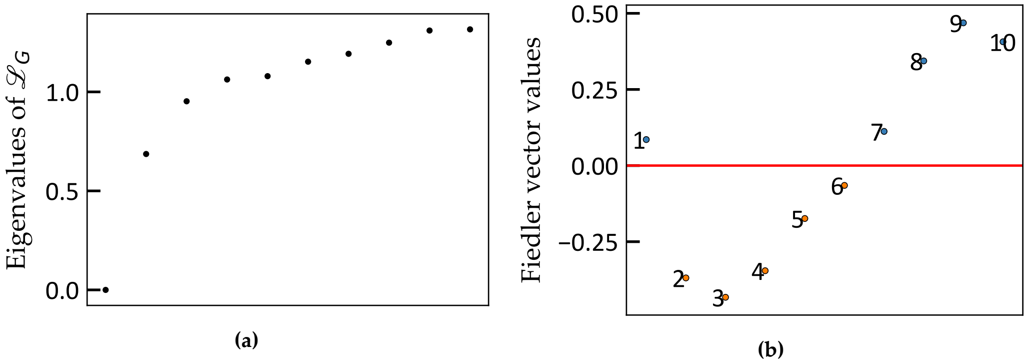
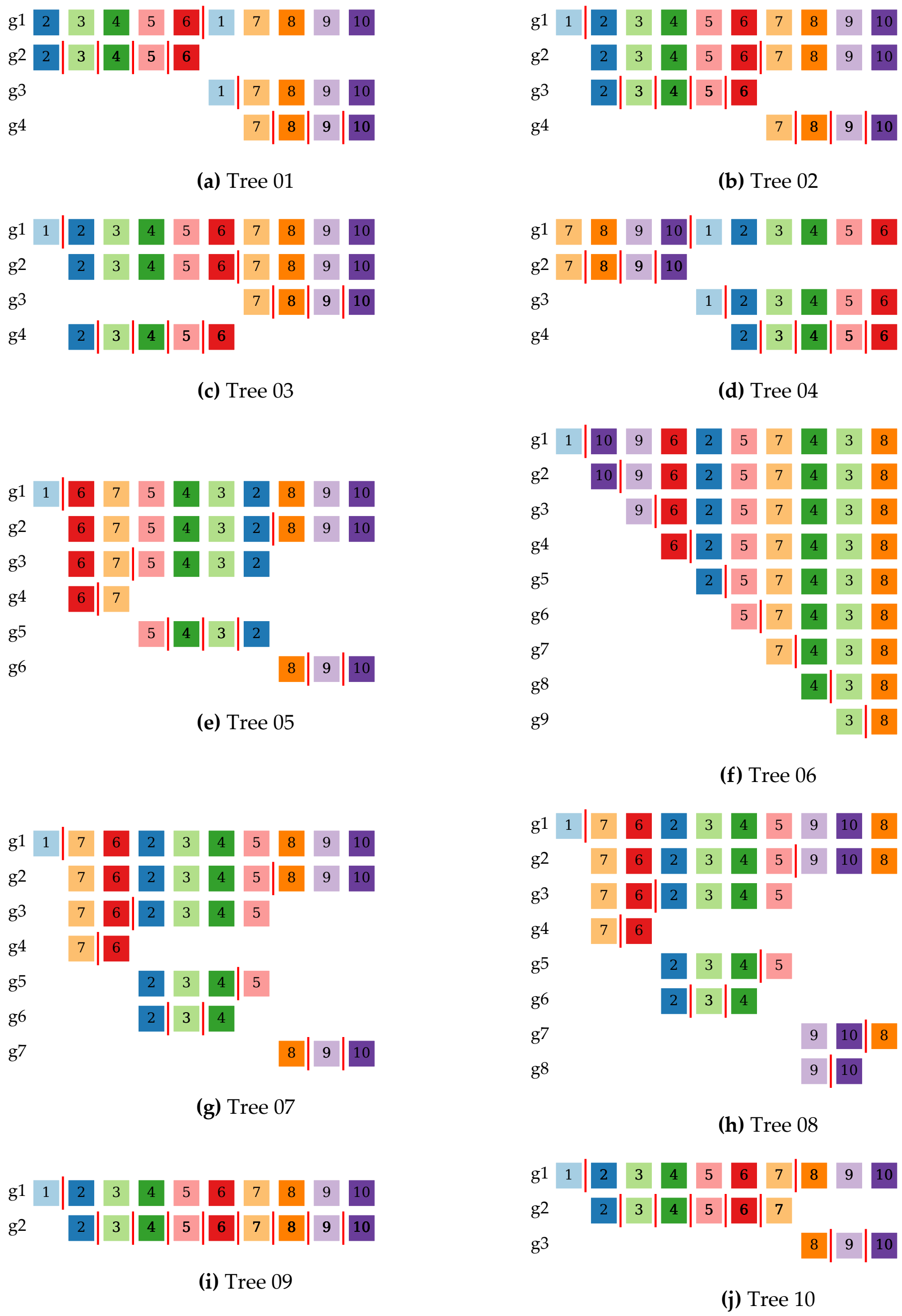


Figure 6.11: Eigenvalues and Fiedler vector values of rock types in Mesaba deposit. (a) Sorted eigenvalues of the Laplacian of Graph G. The noticeable difference between the minimum and the second smallest eigenvalue suggest an adequate partition. (b) Bipartitioning based on Fiedler vector values above and below zero.

The ordering of the categories is obtained by selecting the most representative drillholes and repeating the process. The ordering of categories was retrieved from the analysis of the TPM. Figure 6.12 shows a list of possible trees for the Mesaba deposit. These trees should be scored to obtain the optimal tree.



(a) Tree 01

(b) Tree 02

(c) Tree 03

(d) Tree 04

(e) Tree 05

(f) Tree 06

(g) Tree 07

(h) Tree 08

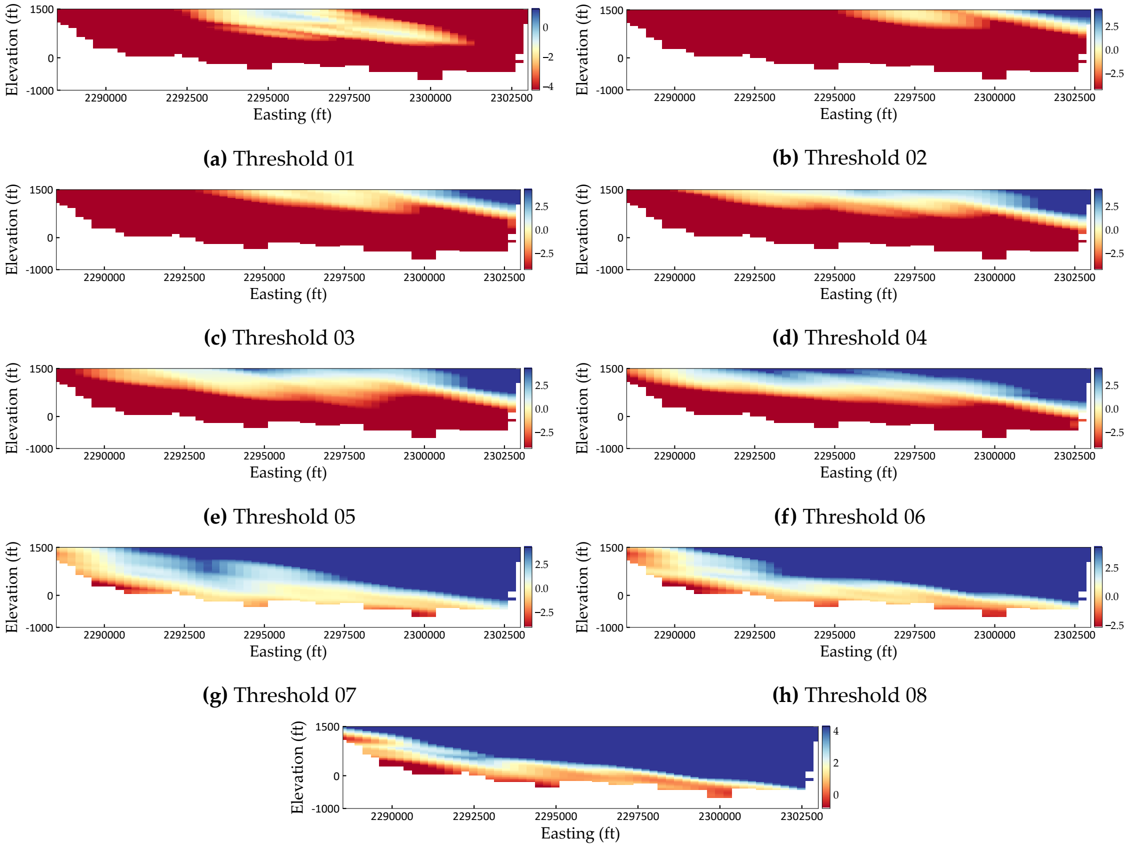
(i) Tree 09

(j) Tree 10

Figure 6.12: Inferred truncation trees with SLCA in example with ten categories. The associations of categories in Trees 01, 02 03, and 04 were inferred from spectral partitioning. The associations and order of categories in Trees 05, 07, and 08 were inferred with SLCA. Tree 06 was generated considering proportions and variogram ranges, and presents the maximum number of latent Gaussian variables. The ordering and associations in Trees 09 and 10 were inferred with TPM and SLCA.

6.4 HTPG Parameters

The parameters for the HTPG workflow are the truncation tree for the rock type categorical variable, the locally varying thresholds, and the model of continuity of the latent Gaussian variables. The truncation tree is arguably the most important parameter in truncated Gaussian methods. It contains information on the geological associations and also considers the spatial structure of categories and proportions. The global thresholds contained in the truncation tree are calculated based on global categorical proportions and are required for the numerical derivation of the variograms of underlying Gaussian variables. The set of inferred trees is used in this example. The local proportions are used to update the thresholds in the truncation tree, it accounts for the non-stationarity. Figure 6.13 shows the nine local threshold maps for ten categories.



(a) Threshold 01

(b) Threshold 02

(c) Threshold 03

(d) Threshold 04

(e) Threshold 05

(f) Threshold 06

(g) Threshold 07

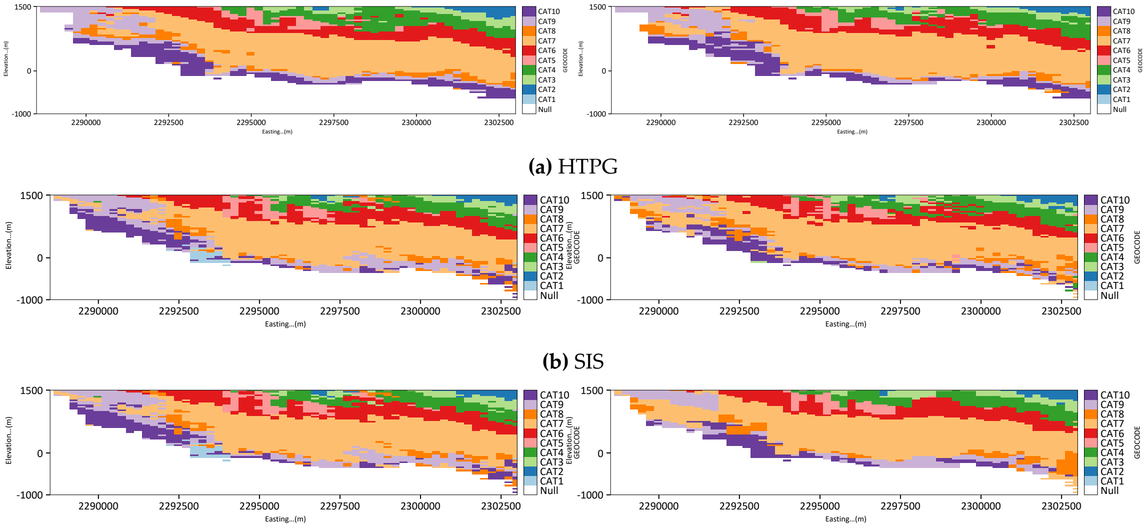
(h) Threshold 08

(i) Threshold 09

Figure 6.13: Local thresholds.

6.5 Results

Ten sets of 100 realizations were generated for each inferred truncation tree. Additionally, two more sets of 100 realizations were included for SIS and SIS+MAPS. The results from SIS+MAPS are referred to as MAPS. Figure 6.14 shows two realizations for the different techniques. At the top, the realizations were generated with HTPG using Tree 08. The realizations in the middle of the image were obtained with SIS. The realizations at the bottom were obtained with MAPS. The HTPG realizations show large structures. The realizations obtained with SIS show less structure and higher variability of the categories, this noise is geologically unrealistic and commonly mitigated with MAPS. The realizations after applying MAPS are smooth and more similar to HTPG results. The correct spatial continuity from the different techniques is checked with variogram reproduction.



(a) HTPG

(b) SIS

(c) MAPS

Figure 6.14: Realizations of the rock type variable generated with HTPG, SIS, and MAPS. The HTPG realizations were generated with Tree 09. Cross section at 418,500 ft. North.

6.5.1 Variogram Reproduction

The spatial continuity in the realizations was compared to the spatial continuity of the data, Figure 6.15. The major, intermediate and vertical variograms of the realizations were compared to the experimental indicator variograms. The light red lines are variograms of the realizations in the major direction. The light blue lines are variograms of the realizations in the mid-direction. The light yellow lines are variograms of the realizations in the vertical direction. The solid black lines represent the respective average variograms. The connected red markers represent the experimental variogram of the data.

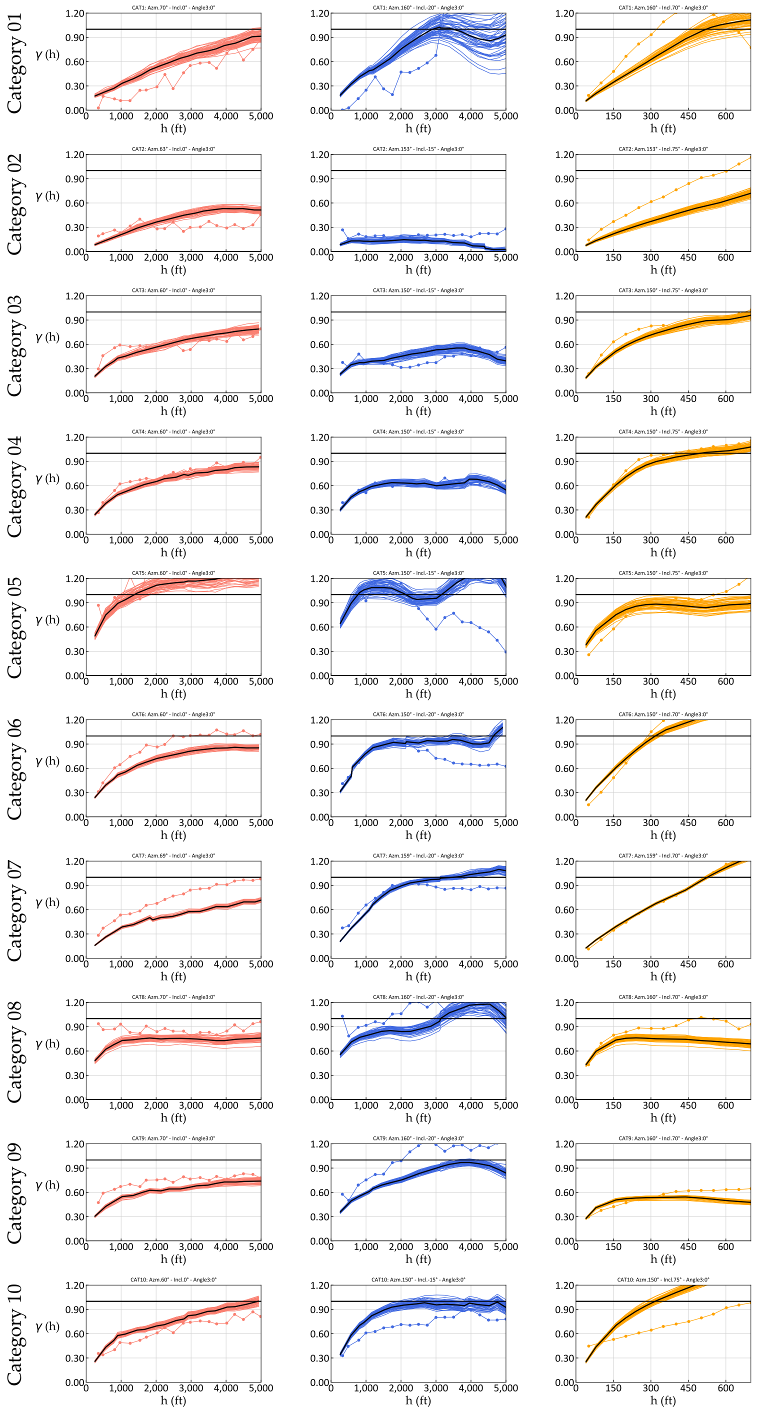


Figure 6.15: Indicator variogram reproduction with HTPG. The light red lines are variograms of the realizations in the major direction. The light blue lines are variograms of the realizations in the mid-direction. The light yellow lines are variograms of the realizations in the vertical direction. The solid black lines are the respective average variogram of the realizations. The connected markers are the data's experimental variograms in the respective anisotropy directions.

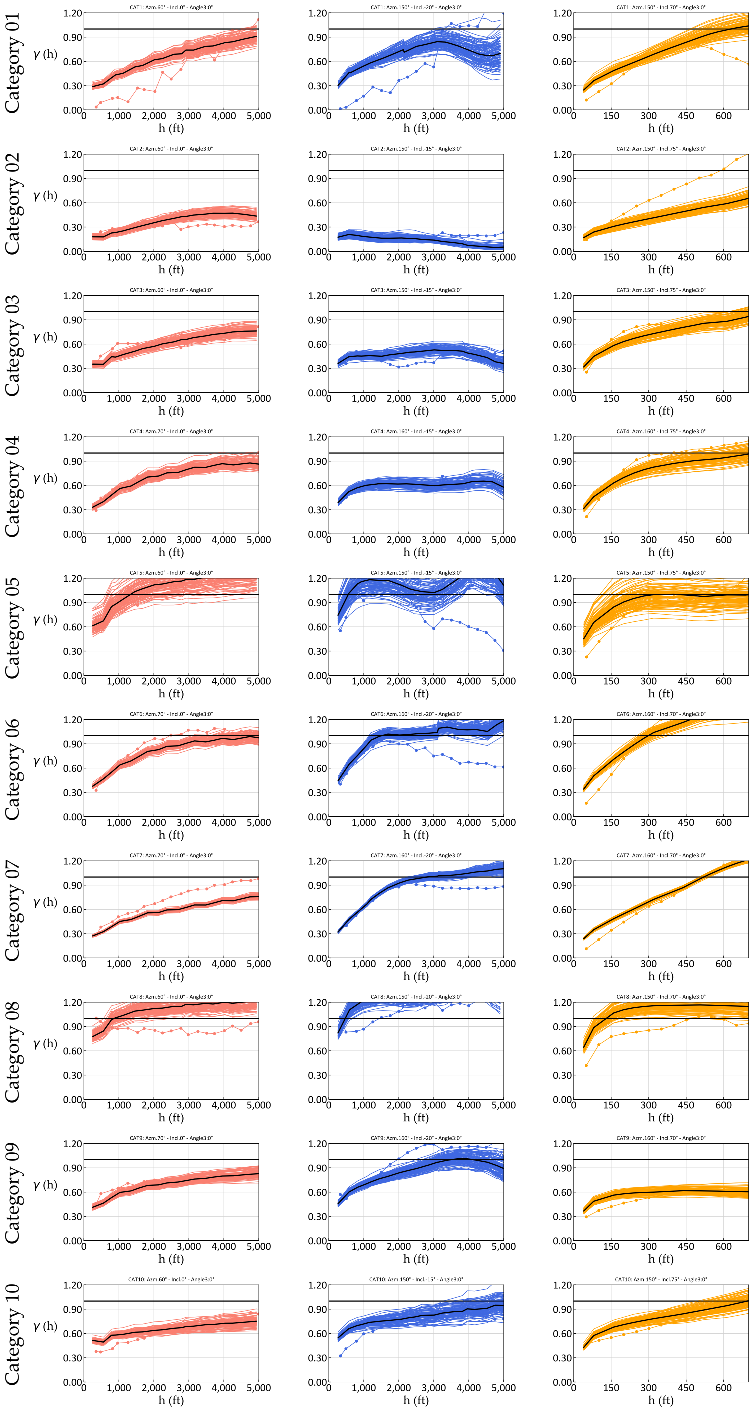


Figure 6.16: Indicator variogram reproduction with SIS. The light red lines are variograms of the realizations in the major direction. The light blue lines are variograms of the realizations in the mid-direction. The light yellow lines are variograms of the realizations in the vertical direction. The solid black lines are the respective average variogram of the realizations. The connected markers are the data's experimental variograms in the respective anisotropy directions.

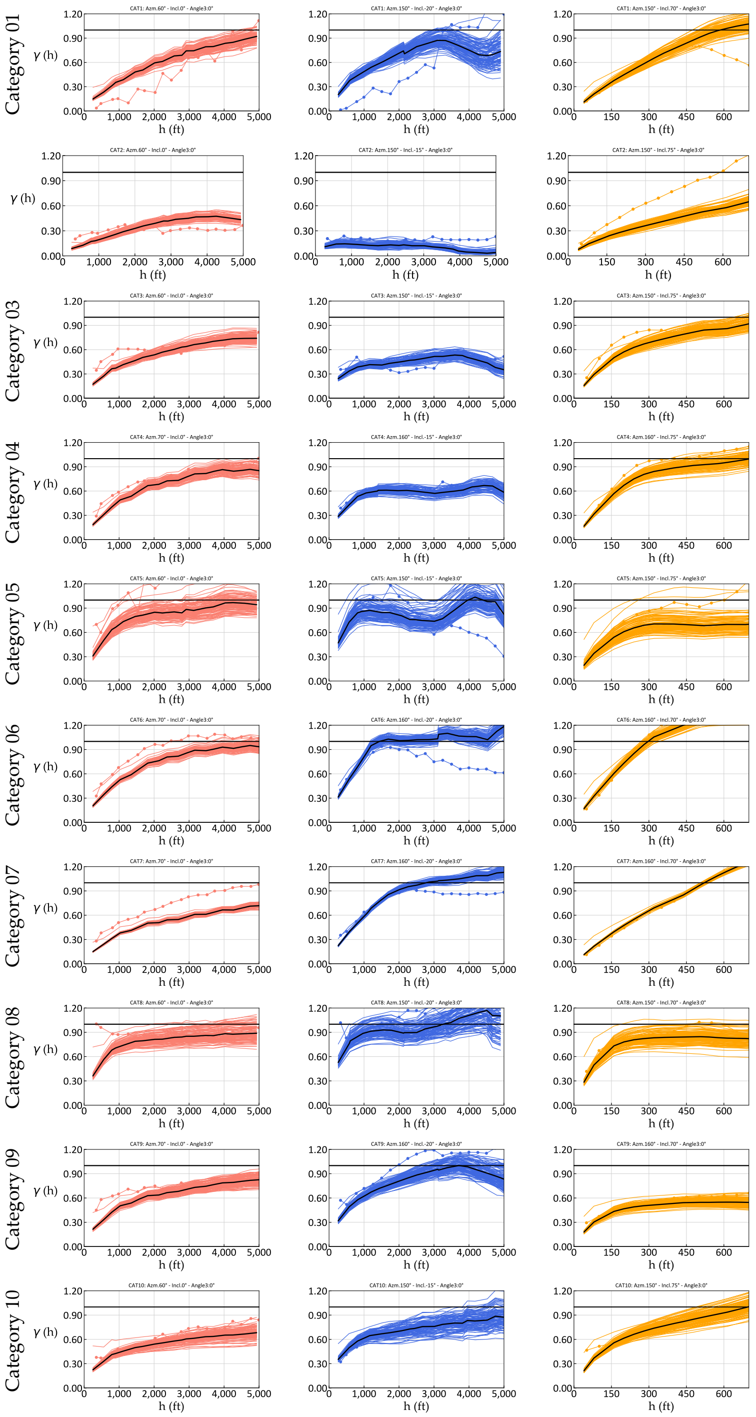


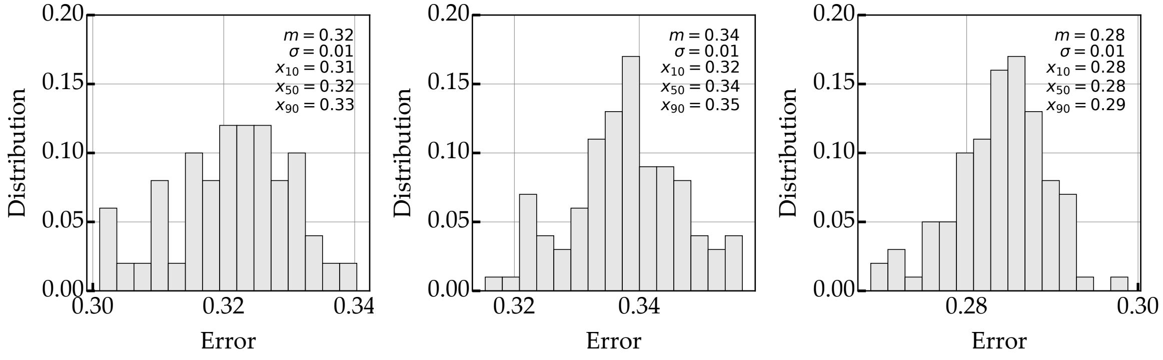
Figure 6.17: Indicator variogram reproduction with MAPS. The light red lines are variograms of the realizations in the major direction. The light blue lines are variograms of the realizations in the mid-direction. The light yellow lines are variograms of the realizations in the vertical direction. The solid black lines are the respective average variogram of the realizations. The connected markers are the data's experimental variograms in the respective anisotropy directions.

6.5.2 Validation

The validation data from the folds are used to check the performance of the modeling. HTPG presented a more consistent and realistic characterization of local uncertainty when compared to MAPS. Special care should be taken when choosing the level of cleaning in MAPS to avoid increased continuity and unreliable characterization of the uncertainty.

6.5.2.1 Prediction Error

Prediction Error (PE) calculates the percentage of error when comparing true categorical values from the validation set and the values from realizations. The distribution of this error is plotted in Figure 6.18. The mean in PE using HTPG is lower than SIS which suggests betters results. MAPS has better PE results than HTPG.

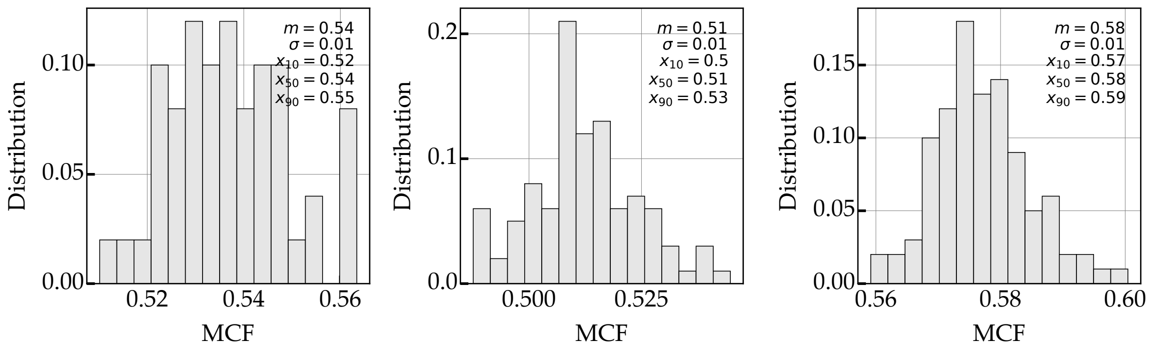


(a) HTPG **(b)** SIS **(c)** MAPS

Figure 6.18: Prediction error for rock types in Mesaba deposit

6.5.2.2 Matthews Correlation Factor

The locations of the validation set were used to retrieve the values from the categorical realizations, each true value is then compared to the closest grid point in the models. The distributions of the MCFs are plotted in Figure 6.19 for HTPG, SIS, and MAPS. Higher MCFs indicate a better correspondence of the predicted categories. The mean MCF obtained with HTPG is better than the corresponding result with SIS. The mean MCF obtained with MAPS is higher than SIS, however, this metric does not verify a correct assessment of the uncertainty.

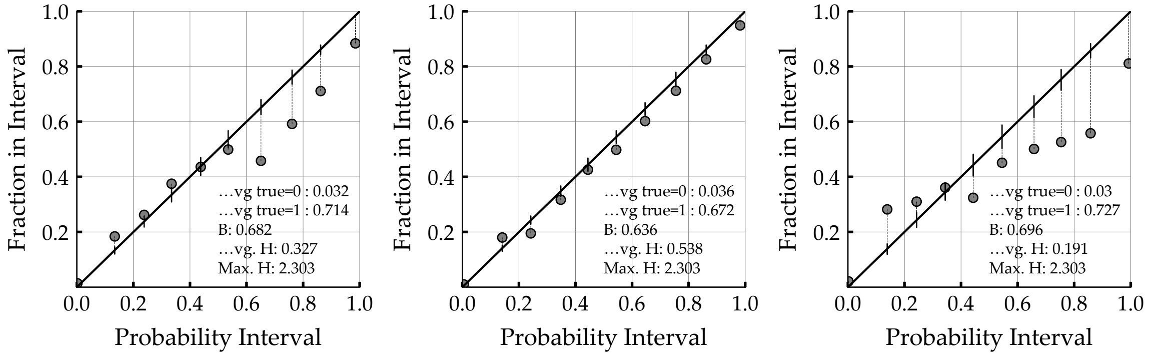


(a) HTPG **(b)** SIS **(c)** MAPS

Figure 6.19: Matthews correlation factor for rock types in Mesaba deposit

6.5.2.3 Probabilistic Accuracy

Final models must be accurate and precise. Accuracy plots should present points close to the 45° line in addition to low entropy and high B values. High H and low B relates to inaccurate and imprecise models. The accuracy plots for the rock type variable are presented in Figure 6.20. The HTPG model shows overall the best features with a higher B value, lower H than SIS, points closer to the 45° line, and is more balanced compared to SIS. MAPS results obtained the highest B values, however, the points show a high departure from the 45° line; in addition, the entropy obtained with MAPS is underestimated compared to SIS results, which suggests a poor characterization of the uncertainty.



(a) HTPG **(b)** SIS **(c)** MAPS

Figure 6.20: Accuracy plot for rock types in Mesaba deposit

6.5.3 Scoring

A summary of the results with the inferred trees is presented in Table 6.4. The table includes scores from six main measures of goodness that are min-max scaled from zero to one, where one is the best result, therefore the maximum attainable overall score is six. Results from SIS and MAPS are also included in the table. SIS results obtained the lowest score. Tree 06 with nine latent Gaussian variables was automatically generated using the proportions and variogram ranges criteria, (D. S. F. Silva, 2018), it shows one of the worst results from the set of inferred trees along with Tree 01 and 10.

Trees 02, 04, and 03 showed reasonable results and occupy the positions from the eighth to the sixth best score; these trees consider categories 06 and 07 in different latent Gaussian variables as could be inferred from the bisection of Fiedler vector values, Figure 6.11b. MAPS obtained the third-best score, however, the result from the VMSE is low which indicates that the reproduction of the indicator variograms is not optimal. Trees 07, 09, and 08 obtained the best scores occupying the fourth, second, and first positions. Trees 07 and 08 have a hierarchical structure closer to the dendrogram obtained with SLCA. It is noted that Tree 08 and Tree 09 are very different trees in relation to the structure, and they obtained the two best overall scores. However, only Tree 09 matches with both the TPM and inference process using the interval probability-based dissimilarity matrix. Tree 09 presents a realistic and simple geological understanding, its low VMSE score may be related to the effectiveness of numerical derivation. For the selection of a truncation tree, the presented measuring factors and simplicity should be considered. The correct choice is Tree 09 with a total score of 3.72, it presents a low number of latent Gaussian variables and aligns with the geological understanding from the section's views of drillholes. From the previous chapter, it was also shown that the ETPE as a measure of goodness performed satisfactorily for truncation tree inference; in the case study, Tree 09 obtained the highest ETPE score.

Table 6.4: Measures of goodness for tree selection in HTPG. Scores $\in [0, 6]$.

Position	Tree ID	N. Lat.	B	PE	VMSE	PMSE	TPE	ETPE	**Score**
01	Tree 08	8	0.39	0.26	0.83	1.00	1.00	0.44	3.92
02	Tree 09	2	0.77	0.53	0.00	0.58	0.84	1.00	3.72
03	MAPS	-	1.00	1.00	0.17	0.53	0.81	0.13	3.64
04	Tree 07	7	0.10	0.06	1.00	0.76	0.63	0.57	3.12
05	Tree 05	6	0.24	0.16	0.23	0.73	0.77	0.68	2.81
06	Tree 03	4	0.08	0.06	0.87	0.78	0.17	0.82	2.77
07	Tree 04	4	0.21	0.14	0.72	0.70	0.13	0.85	2.74
08	Tree 02	4	0.29	0.19	0.58	0.79	0.00	0.83	2.67
09	Tree 10	3	0.00	0.00	0.76	0.84	0.17	0.70	2.47
10	Tree 06	9	0.10	0.07	0.55	0.66	0.78	0.28	2.43
11	Tree 01	4	0.26	0.17	0.50	0.00	0.05	0.78	1.76
12	SIS	-	0.03	0.32	0.71	0.21	0.43	0.00	1.70

6.6 Discussions

A comprehensive methodology for tree inference in the Mesaba deposit has been presented. The geological setting from this deposit is not complex, however, it is a good demonstrative case to test the inference algorithms. The interval-based dissimilarity is useful in varied geological settings, nevertheless, ETPM and TPM are practical in depositional environments. HTPG was used to generate the categorical realizations. A framework for selecting from a set of possible trees based on measures of goodness has been included in the decision of the final tree. Additionally, HTPG's results have been compared to SIS and post-processed models with MAPS. The SIS technique obtained in general the worst results for the different measures of goodness and checking measurements. HTPG results improved the results compared to SIS and are slightly better than the results after using MAPS.

Chapter 7

Conclusions

This chapter starts by reviewing the research motivation. A summary of the main contributions is included followed by some limitations encountered during the development of this thesis. Future work is suggested.

7.1 Review of the Motivation

Categorical modeling is essential in geostatistical workflows and can largely impact subsequent steps during resource evaluation. Several geostatistical techniques are available for the modeling of categorical variables, however, this work is focused on HTPG due to its versatility to impose geological constraints utilizing a truncation tree rule. The challenge of choosing the truncation tree is the main motivation of this research. Attention is centered on the truncation trees and choosing the tree for the best possible categorical models.

7.2 Summary of Contributions

During the development of this work, the number and the structure of the truncation trees are investigated to develop simple ways to communicate the categorical associations of the truncation tree. A second contribution is the development of a novel measure of dissimilarity applicable to drillholes that encodes geometrical distances between categories and can later be used in inference algorithms. A third contribution is the development of tools for truncation tree inference based on the novel measure of dissimilarity. The fourth contribution is the assessment of measures of optimalities in a simulation-based approach to assist in the selection of the best truncation tree in HTPG from a reduced set.

7.2.1 Number and Structure of Truncation Trees

Practical tools are developed and made available to the geostatistical community. These tools allow the calculation of the number of trees based on the number of categories, the generation of a list containing all truncation trees in a parentheses notation, the conversion from the parenthesis-based notation to a dictionary-based notation of truncation trees, and the plotting of trees in a matrix-like format. Concerning the number of truncation trees, the possible combinatorics escalates logarithmically which means that assessing all possible trees is not a practical solution to choosing a tree, however, the possible trees are easily reducible by adding restrictions in the structures and the possible categorical associations based on geological knowledge or information inferred from drillholes.

7.2.2 Interval Probability-based Dissimilarity

Reducing or inferring trees based on limited data contained in drillholes is the next step after exploring the number of trees. A novel measure of dissimilarity is developed to quantify and encode the associations or geometrical relations between categories. This measure of dissimilarity relies on the use of the interval data such as drillholes which includes from-to columns with categorical data represented by rock types and alterations. The conventional approach is to consider constant-length composite data. The proposed dissimilarity is tested in several geological settings and proves to be consistent in interpreting the geological associations. This dissimilarity is calculated for all categorical pairs and summarized in a matrix. This matrix is the basic input for any inference algorithm. GSLIB-like programs are developed to obtain interval-based dissimilarity matrices from drillholes and gridded models.

This dissimilarity is not symmetric and needs to be made symmetric before any further calculation. One advantage of this dissimilarity is that it is insensitive to the domain size. Although the obvious application is on drillhole data, this dissimilarity can be calculated on sparse surface 2-D sampling by creating an interpolated model of the categories. The methodology for surface samples led to similar dissimilarity matrices to the one obtained from drillholes when the number of samples was high, however, it was clear that the drillholes are more reliable and consistent information than point 2-D samples.

7.2.3 Tree Inference Algorithms

A dissimilarity or distance matrix is the input for the tree inference algorithms. The inference algorithms use the interval probability-based dissimilarity matrix as input. Two techniques are considered, the first technique relied on the application of SLCA on the MST calculated from the matrix. This approach outputs a dendrogram that summarizes the hierarchical associations of categories. The advantage of using dendrograms is the direct visualization of the hierarchical associations of categories. However, since the ordering of an independent cluster of categories depends on the algorithm, the specification of contact relations and the ordering of categories in the tree are not definitive.

The second approach uses spectral partitioning, a graph theory technique, to divide the categories into two subclusters. In this approach, the inverse of the dissimilarities is used as elements of the adjacency matrix to obtain a weighted graph. The algorithm divides the categories with positive and negative Fiedler vector values. The inspection of the bisection plot summarizes the associations of categories. The main drawback is that spectral partitioning tends to divide the categories into clusters with a similar number of categories, therefore if the geological setting contains a set of sequenced layers and only one cross-cutting unit, the latter will not be adequately separated. Other approaches consider the use of multiple eigenvectors to perform a K-means clustering that may improve the results.

7.2.4 Assessment of Measures of Optimalities for Tree Selection

The algorithms for truncation tree inference aim to reduce the set of trees before testing possible trees. A synthetic simulation-based approach for the analysis of measures of optimalities with different trees is presented. The approach applies min-max scaling to the results of specific metrics to obtain an overall score. Several measures of optimality were presented, scaled from 0 to 1, and given equal relevance in the final scores of the trees. The results show that transition probability and embedded transition probability errors are minimized when the tree is similar to the reference tree. Other measures of goodness are not as conclusive. The measures of optimality to consider for the scoring is case-based, however, in practice, special consideration should be given to TPM and ETPM for the selection of a tree.

7.3 Limitations and Future Work

The automatic generation of trees was calculated for up to five categories. The results for a higher number of categories were approximations due to the level of recursion required and the combinatorics of categories. Concerning the dissimilarity distance, several experiments could be explored such as an anisotropy-based weighting or adding additional constraints for geological settings with several cross-cutting units to enhance the ability of the dissimilarity to decode the categorical associations. The developed dissimilarities were based on the expected values of the interval probabilities found over each categorical pair. From interval probabilities, the distribution of the intervening intervals for each pair of categories was not thoroughly discussed, however, it contains relevant information about the confidence of the results and could be used as a metric to evaluate models. The programs `INTERVAL` and `INTERVALG` could be modified to output this information. The calculation of the interval probabilities and corresponding dissimilarity matrix from images using the program `INTERVALG` is $O(nK(K-1))$, where n is the number of grid points and K is the number of categories. This program takes considerable time in practical cases with large grids and several categories. Improvements to the program could be made to ease its usage including the introduction of anisotropies, and a search window to reduce the computation time. Concerning the automatic tree inference process using SLCA, the ordering of the categories in the inferred trees is not always the best as it depends on how the algorithm that calculates the dendrogram orders the clusters of categories. The spectral partitioning algorithm is not stable if applied recursively, this technique tends to output clusters with a similar number of categories. Variants of spectral partitioning consider more eigenvectors and K-means to obtain the clusters. Those variants could be explored to obtain more flexible partitions. Concerning the variogram derivation, the only parameter explored was the tree structures. The results showed that structures with more latent Gaussian variables gave a better fit. The influence of the anisotropies and proportions in complex tree structures is an area for further research.

References

Allaby, M. (2013). *A dictionary of geology and earth sciences* (3rd ed.). Oxford University Press.

Armstrong, M. (1992). Positive definiteness is not enough. *Mathematical Geology*, *24*(1), 135–143. doi: 10.1007/BF00890092

Armstrong, M., Galli, A. G., Le Loch, G., Renard, D., Doligez, B., Eschard, R., & Geffroy, F. (2011). *Plurigaussian simulations in geosciences*. Springer Science \& Business Media. doi: 10.1007/978-3-662-12718-6

Belbachir, H., Djemmada, Y., & Németh, L. (2021). The deranged Bell numbers. *arXiv preprint arXiv:2102.00139*, 1–11. Retrieved from `http://arxiv.org/abs/2102.00139`

Beucher, H., & Renard, D. (2016). Truncated gaussian and derived methods. *Comptes Rendus Geoscience*, *348*(7), 510–519.

Boisvert, J. (2010). *Geostatistics with Locally Variable Anisotropy* (CCG Thesis). University of Alberta, Edmonton, AB.

Bourmaud, G., Mégret, R., Giremus, A., & Berthoumieu, Y. (2014). Global motion estimation from relative measurements using iterated extended kalman filter on matrix lie groups. In *2014 ieee international conference on image processing (icip)* (pp. 3362–3366).

Cabral Pinto, F. A., & Deutsch, C. V. (2017). *Calculation of High Resolution Data Spacing Models* (In J.L. Deutsch ed.; Geostatistics Lessons). Retrieved from `http://geostatisticslessons.com/lessons/dataspacing`

Carle, S. F., & Carle, S. F. (1997). *Integration of Geologic Interpretation into Geostatistical Simulation* (Tech. Rep.). Livermore, California: Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States).

Carle, S. F., & Fogg, G. E. (1996). Transition probability-Based indicator geostatistics. *Mathematical Geology*, *28*(4), 453–476. doi: 10.1007/bf02083656

Chiles, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons.

Christakos, G. (1984). On the Problem of Permissible Covariance and Variogram Models. *Water Resources Research*, *20*(2), 251–265. doi: 10.1029/WR020i002p00251

Chung, F. R. (1997). *Spectral graph theory* (Vol. 92). American Mathematical Soc.

Deutsch, C. V. (2005). *A sequential indicator simulation program for categorical variables with point and block data: BlockSIS* (CCG Report 07). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Deutsch, C. V. (2010). *Display of Cross Validation / Jackknife Results* (CCG Report 12). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Deutsch, C. V. (2018). *Partitioning drill hole data into K folds* (CCG Annual Report 20). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

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

Deutsch, J. L. (2015). *Multivariate Spatial Modeling of Metallurgical Rock Properties* ((CCG Thesis)). University of Alberta, Edmonton, AB.

Deutsch, J. L., & Deutsch, C. V. (2012). *Accuracy Plots for Categorical Variables* (CCG Report 14). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Deutsch, J. L., & Deutsch, C. V. (2013). *Advances in Truncated Plurigaussian Simulation for Reproduction of Transition* (CCG Report 15). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Deutsch, J. L., & 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. doi: 10.1007/s00477-013-0783-1

Duda, R. O., Hart, P. E., & Stork, D. G. (1973). *Pattern classification and scene analysis* (2nd ed., Vol. 3). New York: Wiley.

Duin, R. P., & Elżbieta Pe@kalska. (2012). The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, *33*(7), 826–832. doi: 10.1016/j.patrec.2011.04.019

Edelman, S. (1999). *Representation and Recognition in Vision*. MIT press.

Elfeki, A., & Dekking, M. (2001). A Markov Chain Model for Subsurface Characterization: theory and applications. *Mathematical Geology*, *33*(5), 569–589.

Emery, X. (2004). Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, *18*(6), 414–424. doi: 10.1007/s00477-004-0213-5

Emery, X. (2007). Simulation of geological domains using the plurigaussian model: New developments and computer programs. *Computers and Geosciences*, *33*(9), 1189–1201. doi: 10.1016/j.cageo.2007.01.006

Fiedler, M. (1973). Algebraic Connectivity of Graphs. *Czechoslovak Mathematical Journal*, *23*(2), 298–305.

Galli, A., H, B., G, L. L., & B, D. (1994). The pros and the cons of the truncated gaussian method. In *Geostatistical simulations* (pp. 217–233). Springer.

Gascuel, O. (1997). *Concerning the NJ algorithm and its unweighted version, UNJ* (Tech. Rep.). Montpellier: Deoartenebt d'Informatique Fondamentale.

Goldfarb, L. (1985). A new approach to pattern recognition. *Progress in Pattern Recognition*, *2*, 241–402.

Golub, G., & Van Loan, C. F. (1996). *Matrix Computations - Johns Hopkins studies in mathematical sciences* (3rd ed.). JHU Press.

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

Gower, A. J. C., Ross, G. J. S., Journal, S., Statistical, R., Series, S., & Statistics, C. A. (2016). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *18*(1), 54–64.

Harding, B., & Deutsch, C. (2021). *Trend Modeling and Modeling with a Trend.* (In J.L. Deutsch ed.; Geostatistics Lessons). Retrieved from `http://www.geostatisticslessons.com/pdfs/trendmodeling.pdf`

Hassanpour, R. (2007). *Tools for Multivariate Modeling of Permeability Tensors and Geometric Parameters for Unstructured Grids* (CCG Thesis). University of Alberta, Edmonton, AB.

Journel, A. G., & Alabert, F. G. (1990). New method for reservoir Mapping. *Journal de petroleum technology*, 212–218.

Journel, A. G., & Gomez-Hernandez, J. J. (1993). Stochastic imaging of the Wilmington clastic sequence. *SPE Formation Evaluation*, *8*(1), 33–40. doi: 10.2118/19857-pa

Journel, A. G., & 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. doi: 10.1007/BF01033030

Knuth, D. E. (2013a). *The art of computer programming: Generating all partitions*. Addison-Wesley Publishing Company.

Knuth, D. E. (2013b). *The art of computer programming: Generating all trees - history of combinatorial generation* [Vol.4 Fascicle 4]. Addison-Wesley Publishing Company.

Knyazev, A. (2018). On spectral partitioning of signed graphs. In (pp. 11–22). doi: 10.1137/1.9781611975215.2

Krumbein, W. C., & Dacey, M. F. (1969). Markov chains and embedded Markov chains in geology. *Journal of the International Association for Mathematical Geology*, *1*(1), 79–96. doi: 10.1007/BF02047072

Kruskal, B. J. B., & Wish, M. (2011). Dimensionality in: multidimensional scaling. *SAGE Research Methods*, 49–60.

Kruskal, J. B. (1956). On the shortest spanning subtree of a araph and the traveling salesman problem. *Proceedings of the American Mathematical society*, *7*(1), 48–50.

Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness of Fit to non-metric Hypothesis. *Psychonometrika*, *29*(1), 1–27.

Lajevardi, S. (2015). *Improved Probabilistic Representation of Facies Through Developments in Geostatistical Practice* (CCG Thesis). University of Alberta, Edmonton, AB.

Lovering, T. S. (1963). Epigenetic, diplogenetic, syngenetic and lithogene deposits. *Economic Geology*, *58*(3), 315–331.

Madani, N., & 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*(8), 2173–2191. doi: 10.1007/s00477-014-0997-x

Makarenkov, V. (2001). T-REX: Reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, *17*(7), 664–668. doi: 10.1093/bioinformatics/17.7.664

Mansour, T. (2013). *Combinatorics of set partitions*. CRC Press Boca Raton.

Matheron, G., Beucher, H., de Fouquet, C., Galli, A., Guerillot, D., & Ravenne, C. (1987). Conditional simulation of the geometry of fluvio-deltaic reservoirs. In *Spe annual technical conference and exhibition.*

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451.

Mayhew, K., Mean, R., O'Connor, L., & Williams, T. (2009). Nickel and cobalt recovery from mesaba concentrate. In (pp. 25–27).

Murty, U. S. R., & Bondy, A. (2008). *Graduate texts in mathematics: Graph Theory)* (Vol. 244). Springer London, Limited.

Ng, A., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th international conference on neural information processing systems: Natural and synthetic* (Vol. 14, pp. 849–856).

Patelke, R. (2003). *Exploration Drill Hole Lithology, Geologic Unit, Copper-Nickel Assay, and Location Database for the Keweenawan Duluth Complex, Northeastern Minnesota* (UMN Digital Conservancy). Duluth, MN: University of Minnesota Duluth. Retrieved from https://hdl.handle.net/11299/187133

Peirce, C. S. (1880). On the algebra of logic. *American Journal of Mathematics*, *3*(1), 15–57.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell System Technical Journal*, *36*(6), 1389–1401. doi: 10.1002/j.1538-7305.1957.tb01515.x

Pyrcz, M., & Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling* (2nd ed.). Oxford University Press.

Rohlf, F. J. (1973). Algorithm 76. Hierarchical Clustering Using the Minimum Spanning Tree. *Computer Journal*, *16*(1), 93–95.

Rossi, M. E., & Deutsch, C. (2013). *Mineral Resource Estimation*. Springer Science & Business Meda.

Sadeghi, S. (2017). *Geostatistical Categorical Variable Modeling using Optimization Techniques with Truncated Plurigaussian Simulation* (CCG Thesis). University of Alberta, Edmonton, AB.

Sadeghi, S., & Boisvert, J. B. (2012). *Optimizing Thresholds in Truncated Pluri-Gaussian Simulation* (CCG Annual Report 14). Edmonton, AB: University of Alberta. Retrieved from http://www.ccgalberta.com

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345. doi: 10.1007/BF02293654

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*(3), 219–246.

Silva, D. A., & Deutsch, C. V. (2013). *Correcting distance function models to correct proportions* (CCG Annual Report 15). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Silva, D. S. F. (2018). *Enhanced Geologic Modeling of Multiple Categorical Variables* (CCG Thesis). University of Alberta, Edmonton, AB.

Slininger, B. (2013). *Fiedler's Theory of Spectral Graph Partitioning* (Tech. Rep.). California, CA: University of California.

Sloane, N. (2022a). *The Online Encyclopedia of Integer Sequences* (OEIS Foundation Inc.) Retrieved from `https://oeis.org/A000110`

Sloane, N. (2022b). *The Online Encyclopedia of Integer Sequences* (OEIS Foundation Inc.) Retrieved from `https://oeis.org/A000670`

Spielman, D. A., & Teng, S. H. (2007). Spectral partitioning works: planar graphs and finite element meshes. *Linear algebra and its applications*, 284–305.

Strebelle, S. (2002). Conditional simulation of complex geological structures Using multiple point statistics. *Mathematical Geology*, *34*(1), 1–21.

Suro-Perez, V., & Journel, A. (1990). Stochastic Simulation of lithofacies: an improved sequential indicator approach. In *Proceedings of 2nd european conference on the mathematics of oil recovery (ecmor).* Paris.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Velasquez, H. G., & Deutsch, C. V. (2021). *The Number and Structure of Truncation Trees in Hierarchical Truncated PluriGaussian* (CCG Annual Report 23). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Velasquez, H. G., & Deutsch, C. V. (2022). *Alternative measure of distance for Truncation Tree Inference in HTPG* (CCG Annual Report 24). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Vistelius, A. (1949). On the question of the mechanism of formation of strata. *Dokl. akad. naug. sssr*(65), 191–194.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, *17*(4), 395–416.

Wilde, B. J. (2010). *Data spacing and uncertainty* (CCG Thesis). University of Alberta, Edmonton, AB.

Xu, C., Dowd, P. A., Mardia, K. V., & Fowell, R. J. (2006). A flexible true plurigaussian code for spatial facies simulations. *Computers and Geosciences*, *32*(10), 1629–1645. doi: 10.1016/j.cageo.2006.03.002

Zagayevskiy, Y. V., & Deutsch, C. V. (2015). *Numerical Derivation of Gaussian Variogram Models for Truncated Pluri-Gaussian Simulation* (CCG Annual Report 17). Edmonton, AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Appendix A

Appendices

A.1 Software

Two GSLIB-like programs are developed to calculate interval probability-based dissimilarity matrices as the first step. The program INTERVAL is applied to GSLIB-formatted composited drillhole data, and the program INTERVALG is applied to gridded 2-D and 3-D images. The second step is to use the dissimilarity matrix to infer truncation trees. The program CLUSTER MST is developed. This software performs SLCA on an MST built from dissimilarities. The output of the program is a dendrogram and encoded information of the MST. The third step is implemented in a Python class, (Van Rossum & Drake Jr, 1995). It uses the output of step two to infer truncation trees. This Python class performs the following steps: (1) the dendrogram is cut at different thresholds obtaining clusters from the branches which are checked with the MST for permissible sequences of categories, if some sequences are not possible in the MST, the clusters are discarded, (2) the validated clusters are used to build the hierarchical structure in the dendrogram and obtain a tree, and (3) the trees are translated to a compatible pygeostat dictionary notation and plotted.

A.1.1 INTERVAL

The INTERVAL is implemented as a standalone program, which follows the geostatistical software library conventions (GSLIB). From **Line 1** to **Line 3** the content can be omitted. **Line 4** specifies the start of the parameter file. It is important to keep the word START at the beginning of the line. **Line 5** defines the input data file, it must be a GeoEAS formatted file containing composites and correspondent DHID. DHIDs must be integers. Composites in each drillhole must be already sorted, that is, each drillhole represents a string of sequenced categories in space. **Line 6** specifies the DHIDs column and categories' column. **Line 7** specifies the number of categories considered. If the number of categories used in Line 7 is less than the actual categories in the file, the omitted categories will not be included in the calculation, affecting the matrix. However, sometimes it may be useful to interpret dissimilarities for a subset of the total categories. **Line 8** specifies the categorical codes in the input file. There must be the same number of categorical codes in the parameter file as the number of categories used in Line 7. It also specifies the order of elements in the output matrix.

Line 9 specifies the output file name for the dissimilarity matrix. In the output file, the first three rows are informative. The actual dissimilarity matrix is written from row four to the end of the file. **Line 10** is a user option to obtain the dissimilarity matrix in its raw asymmetric form (option 0), or as a symmetric matrix (option 1). The default option is 1. **Line 11** is the output file name for the interval probabilities. **Line 12** is a user option to calculate the interval probabilities downhole (option 0) or in both directions (option 1).

```
 1            Parameters for INTERVAL
 2            ***********************
 3
 4       START OF PARAMETERS:
 5       reservoir_data.dat      -file with data
 6       1   7                   -columns for dhid, category
 7       5                       -number of categories
 8       1 2 3 4 5               -category codes (int>0)
 9       distance.out            -file for dissimilarity output
10       1                       -  0: non-symmetric 1: symmetric
11       interval.out            -file for interval prob. output
12       1                       -  1: up-down, 0:downhole
```

A.1.2 INTERVALG

INTERVALG is a standalone program that follows the geostatistical software library conventions (GSLIB). It implements interval probability-based dissimilarities for gridded 2-D and 3-D models. This program is useful when categorical images are available. **Line 4** identifies the start of the parameter file. **Line 5** specifies the name of the grid image. **Line 6** must contain the number of category codes to use followed by the respective category codes in the same line. Category codes must be integer positive numbers. **Line 7** to **9** is for the grid definition. **Line 10** defines the level of discretization considered along the trace between two points k and k′. The default is 1 in the three axes, that is, the vector kk′ will be discretized by jumping over a contiguous grid cell that lies within the vector footprint. **Line 12** is a flag option to generate a debugging file if option 1 is selected, any other value will not generate a debug file. **Line 13** specifies the output file containing the interval probabilities. **Line 14** specifies the name for the output file containing the dissimilarity matrix. In the output file the first row shows general information, the second row has the number of categories, and the third row contains the category codes in the same order as stated in Line 6. The dissimilarity matrix is contained from the fourth row to the end of the output file. In real-sized models, the time of computation escalates rapidly and depends on the number of cells in the grid model. **Line 15** to **Line 17** is a note at the end of the parameter file and can be deleted. It indicates the approximate computation time in seconds for a given number of cells, e.g, for a model bigger than 1,000,000 cells, a discretization of 5, 5, 5 would be reasonable.

```
 1             Parameters for INTERVALG
 2             ************************
 3
 4      START OF MAIN:
 5      model.dat                   -file with primary data
 6      4 1 2 3 4                   -number of categories, categories
 7      10   0.5    1               -nx,xmn,xsiz
 8      10   0.5    1               -ny,ymn,ysiz
 9      10   0.5    1               -nz,zmn,zsiz
10      1  1  1                     -Discretize along ix,iy,iz
11      1                           -debug file 0=None,1=basic
12      dbg.out                     -file for debugging output
13      interval.out                -file for interval prob. output
14      distance.out                -file for dissimilarity output
15      Note:
16      ----
17      Estimated Comp. Time (s): 0.0039*ncells - 30
```

A.1.3 CLUSTER MST

The CLUSTER MST is a standalone program. It outputs a dendrogram using the distance matrix from INTERVAL or INTERVALG. A user-defined symmetric dissimilarity matrix can be used if it follows the format of the two mentioned programs. **Line 4** identifies the start of the parameter file. **Line 5** specifies the name of the input file. This file is the output dissimilarity matrix from INTERVAL or INTERVALG. **Line 6** specifies the number of categories in the input matrix. **Line 7** is for the categorical codes, they must be in the same order as in the dissimilarity matrix. The categorical codes are assigned to the matrix elements respecting this order. **Line 8** is the name for the output MST file. This file is a GSLIB formatted file containing three columns that define an MST, the from-node, the to-node, and its respective weight or distance. **Line 9** and **10** specify the name of the output dendrogram file and PostScript plot file. Both, the MST and the dendrogram files are used to infer trees.

```
1              Parameters for CLUSTER_MST
2              **************************
3
4  START OF PARAMETERS:
```

```
distance.dat                    -file with data
7                               -number of categories
1 2 3 4 5 6 7                   -category labels
mst.out                         -output file
dendrogram.out                  -output dendrogram
dendrogram.eps                  -output postScript file
```

A.2 Example of Truncation Structures

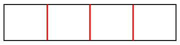
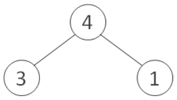
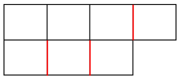
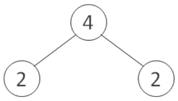
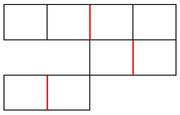
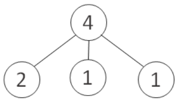
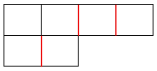
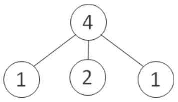
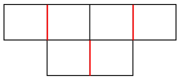
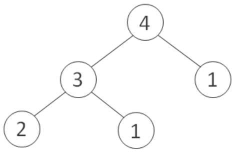
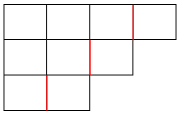
Structure	Parentheses Notation	e.g. $\{a, b, c, d\}$	Truncation Tree Structure
	(4)	(abcd)	
	((3) (1))	((abc)(d))	
	((2) (2))	((ab)(cd))	
	((2) (1) (1))	((ab)(c)(d))	
	((1) (2) (1))	((a)(bc)(d))	
	(((2) (1)) (1))	(((ab)(c))(d))	

Table A.1: Example of truncation tree structures with four categories. There are six different structures for $K = 4$, symmetric structures are not included. The first column shows the linked-node representation of trees. The second column shows the equivalent parenthesis notation of truncation structures, the integers specify the cardinality of the node. The third column shows examples of trees with labels. The fourth column shows the truncation tree structure as a matrix plot where the red bars are the position of thresholds.

A.3 Example of MSE in Numerical Derivation

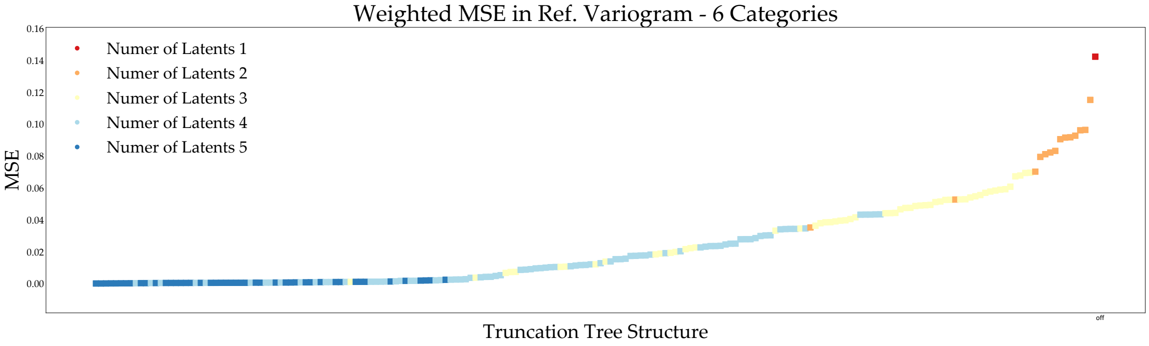


Figure A.1: Weigthed MSE between optimized points and reference variogram. Example with six categories and all truncation structures.