Data Driven Decisions of Stationarity for Improved Numerical Modeling in Geological Environments

by

Ryan Martin

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

Doctor of Philosophy

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Generating representative models of geological domains is critical for decision making and process optimization in natural resource exploitation. Partitioning geological datasets is an important step undertaken early in geostatistical analysis to ensure that subsequent modeling stages consider only related subsets of the domain. The partitioning is hierarchical; the large scale focuses on the geological properties of the domain to ensure the resulting models are geologically representative, and the smaller scale considers the statistical properties of the modeling variables to ensure the assumptions made in geostatistical algorithms are reasonable.

Implicit modeling tools are commonly used to generate geological boundaries, partitioning the domain at the large scale. Commercial implicit modeling tools allow explicit control over all aspects of generated models, abstracting away the actual construction of geological surfaces. However, a manual interpretation of local structural features is required in many geological settings to ensure the boundary models are reasonable. Furthermore, the interpretation must be manually implemented by digitizing polylines or specifying local orientations throughout the domain. An implicit modeling framework is developed to address the inference and interpretation of local features from the dataset. The framework utilizes a well-established strategy for interpolating large datasets, modified for geological modeling by introducing unique local orientations of continuity. An additional methodology is developed to automatically generate local orientations from previous boundary models. The combination of automatic inference of local orientations with the developed boundary modeling algorithm results in a data-guided workflow for capturing curvilinear features in geological domains. The resulting models improve the geological feature reproduction of geological models.

Following the large-scale geological partitioning, explicit decisions about statistically related samples must be made. This decision of stationarity is an important assumption made by geostatistical algorithms used to estimate value at the unsampled locations. Deviations from stationarity imply errors and biases will result in the generated models. There is a consensus in geostatistical literature that improved decisions of stationarity result in improved domains for geostatistical modeling. Increasingly the stationary partitioning of geological datasets is

achieved through clustering, spatial clustering and other machine learning algorithms. Yet, the justification of a set of stationary domains is lacking. Several improvements to stationary domaining are proposed in this work, to aid in the generation and justification of groups of samples used for statistical modeling. Metrics for comparing two different possible stationary configurations are developed that allow the practitioner to choose better domains for geostatistical modeling. Furthermore, two new clustering algorithms are introduced that ease the burden of parameterization, and permit assessment of the uncertainty associated with this component of geostatistical analysis.

This thesis makes several contributions to two related areas of geostatistics that have large consequences on the generated models. Geological boundary models are required to ensure the generated models have the correct geological context. Similarly, decisions of stationarity are required to ensure the dataset is amenable to statistical modeling. The combined contributions of this thesis improve partitioning of geological datasets at all scales for geostatistical modeling.

PREFACE

This thesis is an original work prepared by Ryan Martin. Parts of this thesis have been previously published.

Chapters 3 and 5 are composed by Martin, R., and Boisvert, J. B., *Iterative refinement of implicit boundary models for improved geological feature reproduction*, published by Computers and Geosciences as an original research article.

Chapter 4 is composed by Martin, R., and Boisvert, J. B., *Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics*, published by Computers and Geosciences as an original research article.

For Lisa and Owen.

Acknowledgments

I would like to thank my thesis supervisor Jeff, without whom I would not have completed this program to the fullest extent of my abilities. Jeff has been a constant source of motivation and a positive reinforcement through all stages of this work. The financial, technical and professional support of the Center for Computational Geostatistics (CCG) and member companies is greatly appreciated. Several other professors, researchers and students at the CCG and the University of Alberta have helped me in my journey. Thank you.

I wish to thank the members of my thesis examination committee for insightful comments that have improved the consistency of this thesis.

I would like to thank my wife Lisa for allowing me to be a student for as long as I have been. She provides endless support through all decisions, quests and quandaries in my life. My family and friends in Alberta, Ontario and beyond have also been supportive throughout this journey.

My son Owen showed up late to this party, but provides endless grounding-joy, especially useful after long hours being 'plugged in'. Owen doesn't know it yet, but one day we'll sit around talking about science and programming.

Even if he isn't listening.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Description
A, B	Populations A and B
$ahmax$	Variogram range in the major direction
$ahmin$	Variogram range in the semi-major direction
$avert$	Variogram range in the vertical direction
α, β	Data indexes
ang_1	The first rotation angle from GSLIB
ang_2	The second rotation angle from GSLIB
ang_3	The third rotation angle from GSLIB
r_1	Ratio of the semi-major to major variogram range, $\frac{ahmin}{ahmax}$
r_2	Ratio of the minor to major variogram range, $\frac{avert}{ahmax}$
$G_{i,\alpha}$	Getis local autocorrelation statistic
$I_{i,\alpha}$	Morans local autocorrelation statistic
$C(\mathbf{h})$	Covariance function as a function of distance **h**
C	Uncertainty bandwidth parameter
F	cumulative distribution function (CDF) of the random variable (RV)
C	A $K_1 \times K_2$ co-occurrence matrix, calculated with $\mathbf{B}_1^T\mathbf{B}_2$, counting the number of times a pair of points co-occur in each cluster
B	An $N \times K$ column-wise binary matrix recording membership of each i location to each k cluster
ρ	Correlation
Cov	Covariance
A	Geostatistical domain
x_{ij}	The j^{th} variable value of the i^{th} location

Symbol	Description
$\bar{x}_{kj}$	The j^{th} variable value of the k^{th} cluster center
f	Implicit function
$\gamma(\mathbf{h})$	Variogram at a lag distance **h**
H	Entropy
D	$N \times N$ interpolation matrix
i, j	Location indices
K	The number of clusters, categories, domains for modeling.
k	A single cluster k from the set $\{1, \ldots, K\}$
K_L	The number of clusters for the L^{th} cluster realization
L	Number of realizations
$\boldsymbol{\lambda}$	A vector of weights
u	All locations in domain **A**
$\mathbf{u}_i$	Sample locations in domain **A**
$\mathbf{u}_\square$	Unsampled locations in domain **A**
M	The number of variables sampled in a given dataset
m	Mean
$E\{Z(\mathbf{u})\}$	Expected value of the RV Z
N	Number of samples
G	Standard normal Gaussian CDF
ϕ	Radial kernel
$p_k(\mathbf{u})$	Probability assigned for a given category k to be found at location **u**
$d_{i,j}$	Pairwise distance matrix
$w_{i,j}$	Pairwise weight matrix
R^3	Three dimensional space
R	Rotation matrix calculated considering ang_1, ang_2, ang_3, r_1, and r_2

Symbol	Description
σ	Standard deviation
df	SDF
df^l	Distance function indexed by realization
$\widehat{df}$	Modified SDF considering the C-parameter
SSE	Sum of squared error
$s(\mathbf{u}_i)$	Value of the signed distance function at $\mathbf{u}_i$
wf	The weighting function used in PU interpolation
$Z^*_{sk}(\mathbf{u}_\square)$	Simple kriging (SK) estimate at the unsampled location
$z(\mathbf{u}_i)$	Observed value at $\mathbf{u}_i$
$Z(\mathbf{u})$	Probability distribution about the location $\mathbf{u}$

List of Abbreviations

Abbreviation	Description
ARI	adjusted Rand index
CART	classification and regression trees
CCG	Center for Computational Geostatistics
CDF	cumulative distribution function
CV	cross validation
DS	dual space search
GMM	Gaussian mixture model
IDW	inverse distance weighted
KDE	kernel density estimate
KLD	Kullback-Leibler divergence
LMC	linear model of coregionalization
LNOCV	leave-N-out cross validation
LVA	locally varying anisotropy
MG	multiGaussian
MGK	multiGaussian kriging
MOI	moment of inertia
MPS	multiple point statistics
NE	nugget effect
OK	ordinary kriging
PCA	principal component analysis
pdf	probability density function
PGE	platinum group element
PU	Partition of Unity

Abbreviation	Description
Q-Q	quantile-quantile
RBF	radial basis functions
RF	random forest
RMSE	root mean squared error
RT	rock type
RV	random variable
SA:V	surface area to volume ratio
SDF	signed distance function
SE	spatial entropy
SGS	sequential Gaussian simulation
SIS	sequential indicator simulation
SK	simple kriging
SRF	stationary random function
SSE	sum of squared error
SVD	singular value decomposition
TPG	truncated pluriGaussian simulation
WCSS	within cluster sum of squares

Chapter 1

Introduction

Decisions for resource exploitation require unbiased numerical models that accurately characterize the quantity and quality of the contained materials. The complexity of geological environments and the relatively limited data collected from these domains results in uncertainty that must be incorporated for effective decision making. The final numerical models are the result of a complex series of decisions made by several individuals with the collective goal to maximize profits and minimize environmental impacts. Any model of the subsurface is uncertain; capturing and transferring known uncertainty through numerical modeling is important to optimize downstream planning and decision making based on the most complete information.

1.1 Geostatistics

Rock properties vary spatially and at all scales. Datasets collected to quantify the properties of the subsurface are sparse relative to this variability and provide an incomplete view of the subsurface. Geostatistical techniques are adopted to estimate value and uncertainty where no samples are collected. Two important assumptions are commonly made to facilitate this estimate: 1) the nearby data are related to one another and represent the process being modeled; and 2) the unsampled location is also part of this process. These assumptions simplify the statistical analysis and permit the unbiased characterization of value. Collectively these assumptions comprise stationarity, which is not an intrinsic property of geological domains but instead reflects a series of decisions made by the practitioner to group related samples with one another, model boundaries and boundary relationships, and/or account for deviations from expected behavior. These decisions preclude any geostatistical modeling and thus have large consequences for all downstream decisions including planning, process optimizations, and financial investments that consider the numerical models.

Geostatistics is the practice of statistical inference from samples that are correlated in space

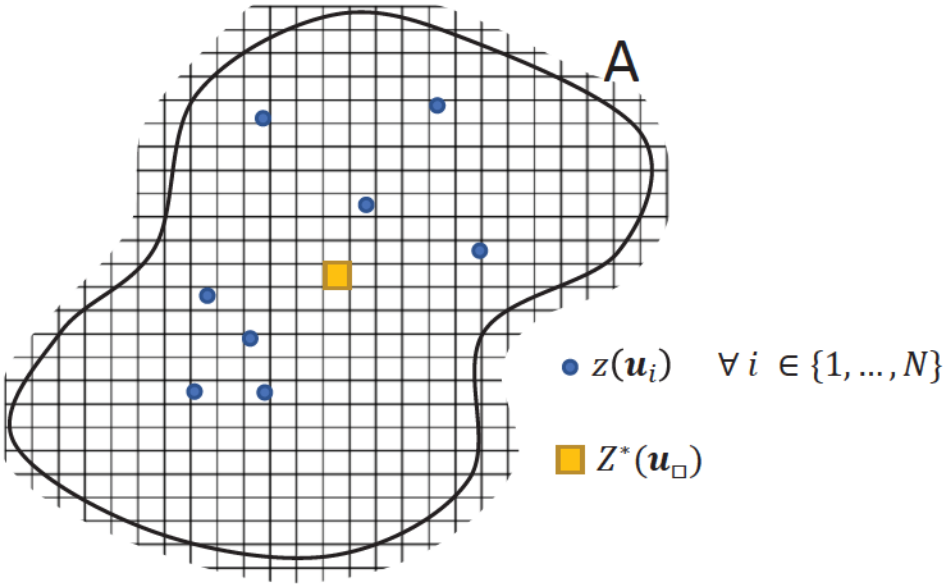


Figure 1.1: Geostatistical domain **A** with sample locations $z(\mathbf{u}_i)$ and unsampled location $\mathbf{u}_\square$. The domain is discretized by a grid suitable for the project engineering parameters.

(Deutsch & Journel, 1998; Matheron, 1963). Usually a geostatistical study is restricted to some domain of interest, **A**, with all locations in the domain denoted **u** (Fig. 1.1). Geostatistics adopts a random variable (RV) formulation such that value at an unsampled location follows a probability distribution $Z(\mathbf{u})$. In this context, the capital Z denotes a distribution of possible values at the $\mathbf{u}_\square$ unsampled location, whereas lowercase $z(\mathbf{u}_i)$ denotes observed values at $\mathbf{u}_i \forall i \in 1, \ldots, N$ sample locations. The probability of a value being observed at a given unsampled location is given by the cumulative distribution function (CDF), $F(z) = Prob\{Z \leq z\}$.

Geological processes result in spatially correlated RVs; this implies that two geographic locations closer to one another are more likely to be related than locations separated by a greater distance. The set of all $Z(\mathbf{u})$ describe a random function $\{Z(\mathbf{u}), \forall\, \mathbf{u} \in \mathbf{A}\}$ with a spatial covariance function $C(\mathbf{h})$ that characterizes the spatial correlation between the random function at different locations. The assumption that the samples $z(\mathbf{u}_i)$ are representative of the underlying random function $Z(\mathbf{u})$ allows inference of the statistical properties of the random function, and in turn, prediction of value and uncertainty at the unsampled locations.

1.2 Estimation at the Unsampled Location

A popular strategy adopted for estimating value at unsampled locations is to consider weighting nearby data:

$$Z^*_{sk}(\mathbf{u}_\square) - m = \sum_{\alpha=1}^{N} \lambda_\alpha (z(\mathbf{u}_\alpha) - m) \tag{1.1}$$

This is referred to as simple kriging (SK) when the weights λ_α minimize the error variance $E\{[Z^*(\mathbf{u}) - z(\mathbf{u})]^2\}$, calculated by solving the relation:

$$C_{\alpha\beta}\lambda_\alpha = C_{\square\alpha} \tag{1.2}$$

for some number of locations $\{\alpha, \beta = 1, ..., N\}$ found in a local search surrounding the unsampled location $_\square$. The associated estimation variance σ^2_{sk} is recovered using the same weights:

$$\sigma^2_{SK} = \sigma^2 - \sum_{\alpha=1}^{N} \lambda_\alpha C_{\square\alpha} \tag{1.3}$$

Under a Gaussian assumption, SK estimates the mean and variance of a Gaussian distribution of uncertainty at the unsampled location, conditional to the observations found in the local search surrounding that location; combined, the Gaussian transform and SK estimator is referred to as multiGaussian kriging (MGK). However, this estimate makes strong assumptions on stationarity, that is, the mean, variance and covariance function of the random function are representative of all locations.

1.3 Stationarity

Stationarity is a key consideration in geostatistical estimation and uncertainty characterization. A stationary random function (SRF) is commonly described in terms of the first and second order moments. A first-order-SRF is one where the mean is invariant with translation through the spatial domain **A**. The mean of the SRF, denoted m, is inferred from the dataset $z(\mathbf{u}_i)$ after ensuring that each sample has a weight reflecting how samples are distributed through the modeling domain. Under first order stationarity, the mean is constant and characteristic of all locations $E\{Z(\mathbf{u})\} = m \ \forall \mathbf{u} \in \mathbf{A}$.

A second-order-SRF is one where the variance is invariant with translation through the spatial domain. The variance of the SRF $Z(\mathbf{u})$, denoted σ^2, is inferred from the sample data $z(\mathbf{u}_i)$ with representative sample weights calculated, as above, to ensure a representative distribution. Under first and second order stationarity, the variance is constant for all locations $E\{Z(\mathbf{u})^2\} - m^2 = \sigma^2 \; \forall \mathbf{u} \in \mathbf{A}$. For geostatistical applications, second-order stationarity more commonly refers to the 2-point statistics of the $z(\mathbf{u}_i)$ sample data, captured by the covariance function $C(\mathbf{h})$. In this context, the orientation and ranges of preferential continuity along each direction are constant with translation, ensuring that the covariance function:

$$C(\mathbf{u}, \mathbf{u}+\mathbf{h}) = E\{z(\mathbf{u})z(\mathbf{u}+\mathbf{h})\} - E\{z(\mathbf{u})\}E\{z(\mathbf{u}+\mathbf{h})\} \quad \forall \mathbf{u}, \mathbf{h} \in \mathbf{A} \tag{1.4}$$

is valid for all **u** and **h** in domain **A**.

The stationary assumption allows inference of the key statistical parameters (m, σ and $C(\mathbf{u}, \mathbf{u}+\mathbf{h})$) from the samples collected to characterize the domain. This, in turn, allows inference of the distributions of uncertainty at the unsampled location through MGK or sequential Gaussian simulation (SGS), or other algorithm. However, geological domains are more commonly deemed non-stationary, resulting from: elemental zoning, gradational contacts, folded structures, structural or hydrothermal overprinting and other features generated from complex and long-lived geological processes. Moreover, the nature of geostatistical problems is increasingly multivariate such that multiple variables must be considered in the generation of distinct sets of samples that meet the above criteria. The omnipresence of non-stationary features and multivariate complexities across all types of geological domains has motivated several practical methodologies to facilitate unbiased prediction in the presence of these features.

1.4 The Decision of Stationarity

Stationarity is an assumption made in the underlying geostatistical algorithms, and not a property of the dataset. In practice a series of decisions are made by the modeler to ensure the assumption of stationarity is reasonable. A decision of stationarity has two main components, where scale and data-support dictate the appropriate tools and techniques. The first component involves defining groups of related samples, which involves partitioning sample dataset within

the context of the total domain (or subdomain), resulting in K modeling groups. Second, the set of related *unsampled* locations are determined by modeling the boundaries between defined sets with a boundary modeling algorithm suitable for the scale of the problem, support of the data and nature of the uncertainty. Boundary modeling maps this choice of K modeling groups to the unsampled locations to ensure that an estimate made at an unsampled location draws from the correct statistical population.

Decisions of stationarity are hierarchical in nature. At the large scale samples may be related through the geological environment, the tectonic setting, or through age relationships. The information supporting delineation at this scale is suitably large-scale in nature; e.g. regional geophysics or mapping datasets. Domains defined by these datasets are deterministic and thus boundaries modeled with deterministic methods considering these geological attributes are often suitable. At a smaller scale, the information collected to characterize the environment consists of direct measurements of continuous variables (e.g., Gold, Bitumen); delineation of domains at this scale requires consideration to the spatial statistical properties of these variables. Domains are generated at different scales to support related activities at each scale. At the very largest regional scale, a subdivision supports further exploration by targeting areas with prospective features dictated by the regional-scale relationships. By contrast, at a smaller scale the delineation of distinct geological bodies, supported by sufficient sample density, may be possible. At the very smallest scale of consideration, subdivision of the dataset facilitates the numerical modeling of the properties contained within the larger-scale geological bodies, within the largest-scale regional framework. Such decisions of stationarity are made at all scales, hierarchically, and based on different supporting information with different final applications. Uncertainty is a component of all decisions and should be quantified and incorporated to numerical models so that decisions based on the models have the most complete information.

1.5 The Importance of Stationarity

Steps from a typical geostatistical study are shown graphically in Figure 1.2. The ultimate goal is to characterize value and uncertainty in the subsurface given the sparse sampling collected to characterize the domain. A generalized set of steps for achieving this goal include: data

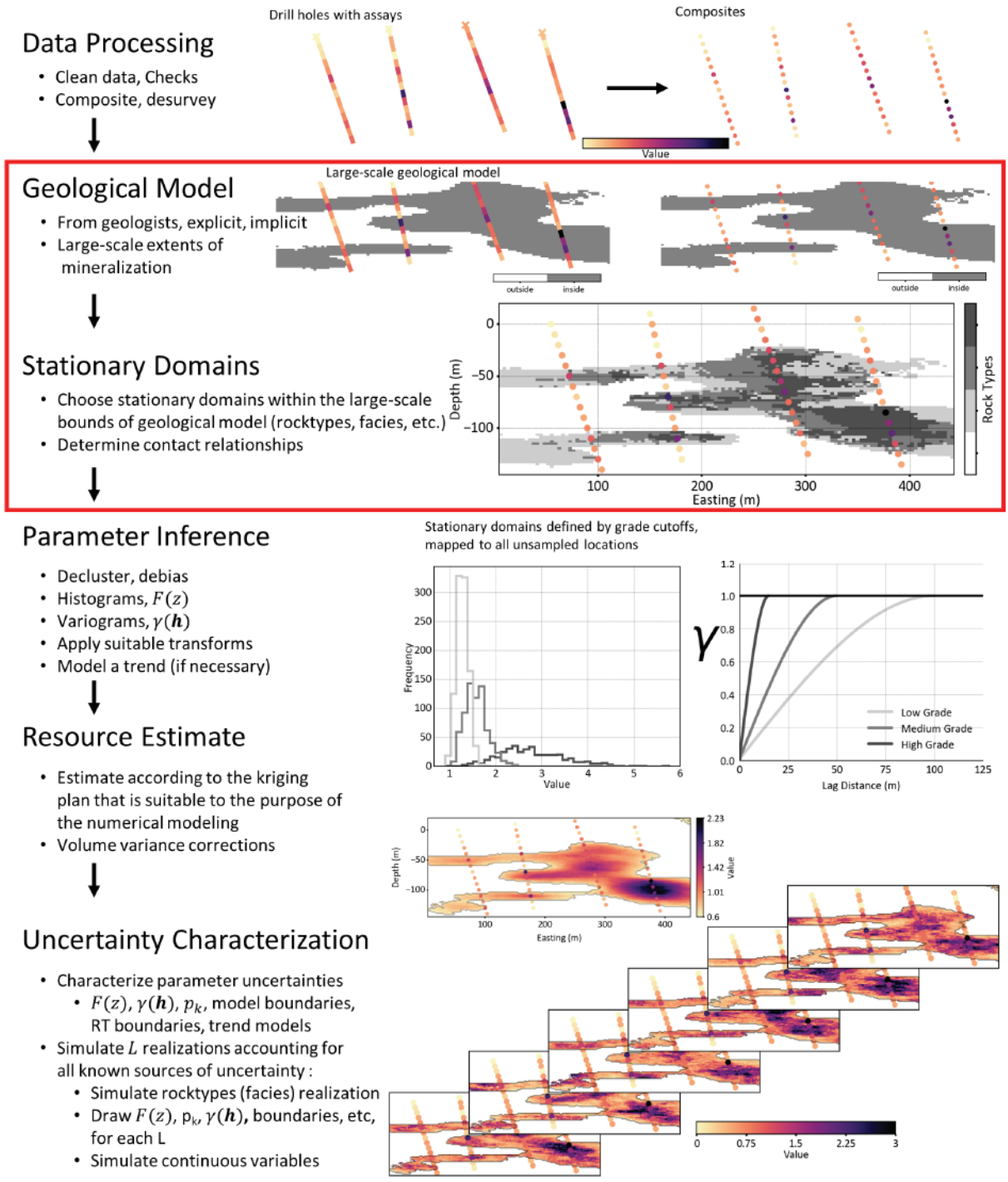


Figure 1.2: Place of this research (highlighted with the red box) within a conceptual geostatistical uncertainty characterization workflow

processing, geological modeling, stationary domaining and geostatistical modeling. Data processing is a required pre-requisite for all statistical modeling to ensure that the samples are representative and free of errors. This component of numerical modeling is essential, but details are beyond the scope of this thesis (see: Abzalov, 2016; Rossi & Deutsch, 2014). Once the data is deemed fit for purpose, the next step is to generate a geological model to impart setting-specific geologic and geometric controls to the numerical models. The geological continuities are commonly generated by experts familiar with the domain, local structure, and mineralizing features, although increasingly computerized tools are used to abstract the details of constructing the model, mainly to increase efficiency. Once the geological domains are deemed reasonable by an expert familiar with the deposit and the geological environment, a set of stationary domains are defined that explicitly pool samples for statistical modeling. The delineation of stationary domains implies that the assumption of stationarity inherent in the geostatistical algorithms that follow is reasonable. Within each stationary domain, traditional geostatistical algorithms estimate value and characterize uncertainty at the unsampled locations. The portion of geostatistical analysis that follows stationary domaining includes: inference of all required parameters, data transformations, trend models, kriging plans, search characteristics, simulation, and post processing. Modeling geological boundaries and the decision of stationarity form an important foundation from which unbiased and representative statistical models of the subsurface are generated.

1.6 Thesis Statement

Incorporating structural features to large-scale geological boundaries and improving the definition of stationary domains results in improved geostatistical models.

1.7 Scope of this Thesis

This thesis focuses on the methodologies and tools used to ensure the assumptions inherent in geostatistical analyses are reasonable. Two areas of domain subdivision are investigated: 1) generating large-scale geologically reasonable boundaries; and 2) improving the definition of stationary domains for numerical modeling. As seen in Figure 1.2, these two activities precede

much of the geostatistical uncertainty characterization workflow and have large effects on all downstream processes, including geostatistical modeling and any decisions that are based on these models.

1.8 Thesis Organization

Chapter 2 provides a literature review detailing the methodologies utilized to generate geological boundaries and stationary domains in this thesis. Chapter 3 develops an iterative locally-anisotropic boundary modeling algorithm. Chapter 4 develops a framework for assessing stationary domains, as well as generating stationary domains with unsupervised learning algorithms. Chapter 5 and 6 present case studies evaluating the methods for implicit modeling and stationary domain definition presented in this thesis. Finally, Chapter 7 provides a summary of the contributions and limitations of the proposed methods, as well as directions for future research.

Chapter 2

Literature Review

This thesis focuses on two main topics: the generation of large-scale boundaries with locally representative features and the methods used to generate stationary domains for geostatistical modeling. The methods are related since each represents a partitioning of the domain based on the properties of the samples. However, the first case partitions the domain based on the geological characteristics and geometric continuities, whereas the second case partitions based on the numerical properties of the samples that characterize the mineralization. This chapter is intended to review the current state of the art with respect to these two related areas of research.

2.1 Geological Boundary Modeling

Boundaries map a decision of related samples to the unsampled locations. Most commonly, the related samples consist of similar geological properties (geological modeling), but related samples may also include decisions of stationarity where the properties of the RV are of primary interest. Boundaries are modeled with the goal of reproducing 'geological features', such as folds, faulting and other complexities. Reproduction of such features ensures that numerical models are geologically reasonable. Boundaries, and the methods to model boundaries, are scale and data-support dependent. Historically, large scale boundaries are modeled using manual interpretation and digitization with computer aided drawing tools. These methods are termed 'explicit' because the manual digitization by an expert explicitly defines the location and extents of geological bodies. Explicit methods allow precise control over the boundary location, extent and the types of features reproduced in the boundary, but they are subjective, unique to each expert, time-intensive, difficult to update or change, and provide no access to uncertainty. Recently, implicit methods are popular for generating geological boundaries since models can be constructed rapidly, are easy to update and provide a tractable workflow that is reproducible by other experts. However, geological and structural interpretations must be manually enforced

by the geomodeler to ensure the implicit model reproduces their interpretation. Without such enforcement, implicit models are not guaranteed to generate 'geological' features; in fact, a common occurrence are geological models that do not look 'geological' (Cowan, 2014).

2.2 Implicit Boundary Modeling

Implicit modeling is a data driven method to extract a surface describing the interface between different geological bodies. Implicit boundary modeling is suitable to model geological features that are relatively large scale and have significant continuity throughout the domain. In this sense the scale of the variability is larger than the scale of the sampling. Data types that can be used to construct implicit models are diverse and include: locations of observed geological properties (rock types, alteration types, etc.); location of contacts; structural measurements; structural interpretations; interpreted constraints (older, younger, etc.); and other information (Fig. 2.1). A categorical dataset is the base for traditional implicit modeling methods, where 1 of K categories is targeted for large-scale boundary modeling. Each category is converted to a volumetric function reflecting the position of each sample relative to a boundary between different units (Fig. 2.3; Wilde & Deutsch, 2012). The implicit model is generated by fitting a function that interpolates the volumetric function at the data locations. The location of the boundary is then said to be known implicitly through the fitted function. A set of contacts and orientation measurements can be used to enhance models by constraining the location and/or orientation of the target boundaries at specific locations (Hillier, Schetselaar, de Kemp, & Perron, 2014; Lajaunie, Courrioux, & Manuel, 1997). If the dataset consists of contacts and orientation measurements, a lower density of point information may be sufficient since the orientation of the surface is dictated by the orientation measurements rather than the arrangement of the point data (Hillier et al., 2014). Although research into implicit geological modeling mainly follows a traditional and strict implicit formulation (e.g., Cowan et al., 2003), today, 'implicit geological modeling' is a more general term describing the process of the data-driven and automatic extraction of geological boundaries. In practice, the methods used to extract the boundaries need not be strictly implicit (Silva, 2015; Wilde & Deutsch, 2012), although there are certain advantages to the strict formulation.

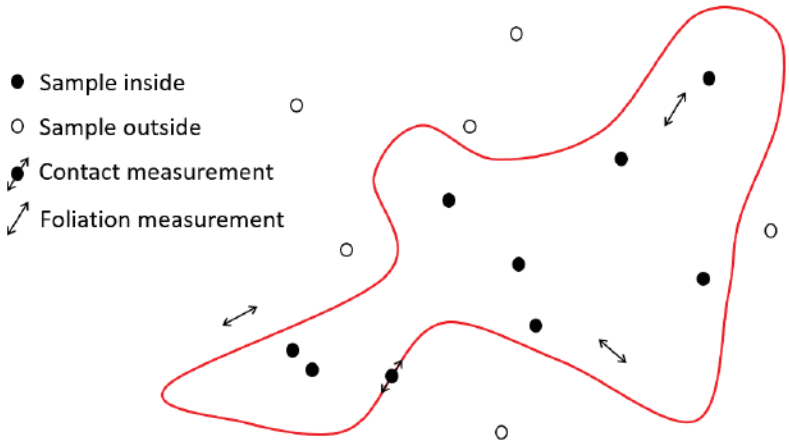


Figure 2.1: Different types of data that can be used in implicit geological boundary reconstruction.

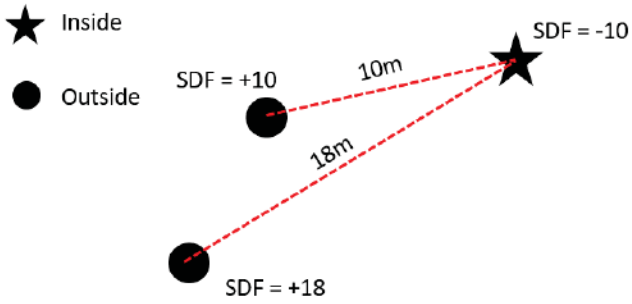


Figure 2.2: Example SDF calculation where the distance to the nearest sample of something different defines the value of the 'signed distance' at each location.

2.2.1 Volumetric Functions

Geological domains are most commonly represented by a signed distance function (SDF) that is calculated from the locations of the categorical data and represents the distance from each sample to a sample of something different, and signed since it is negative if inside and positive if outside (Fig. 2.2; Cowan et al., 2003; Knight, Lane, Ross, Abraham, & Cowan, 2007). The benefit in this representation is that an interpolated value of SDF $= 0$ represents the interface between inside and outside. Thus, to generate a bounding surface of a particular geological category, the first step is to calculate the SDF for the category of interest at all sample locations. Figure 2.3a shows the categorical codes and the corresponding SDF calculated from the point category data. Here, locations where categories are more heterogeneous have SDF values closer to zero, indicating that they are closer to the interface between inside and outside. Other researchers have considered time or stratigraphic position as a function for interpolation (Mal-

let, 2004). The advantage of that representation is that a coordinate system orthogonal to sedimentary deposition implicitly accounts for any locally varying anisotropy that may result from post-depositional modifications like faulting or folding. However, this representation requires specific depositional features that are only found in specific domains. The SDF methodology is indeed more general and more widely used for this reason.

Geostatistical modeling projects rarely involve a single lithology. The SDF is well suited for modeling the interface between two domains (e.g., inside or outside). A modified K category implicit modeling framework is introduced by Silva and Deutsch (2012b) and presented schematically in Figure 2.4. Essentially the domain is represented by K simultaneous SDF modeling problems where each category is modeled independently and the set of K SDF models are post-processed so that each block in the final model is assigned to one of K categories. If the anisotropic parameters governing the continuity of each geological body are identical, weights may be calculated once with Equation 2.5 and the interpolation time required for K categories is significantly reduced. The post-processing of the SDF models chooses the prevailing category at each location as the category with the smallest interpolated SDF value (Fig. 2.4f). This SDF post-processing is suitable for a general problem with K categories and no superposition information (e.g., crosscutting, faulting, etc.). However, if superposition relationships between the different domains are known (e.g., category 1 intrudes category 2) these can be included in the post-processing step by ensuring the cross-cutting category prevails if its interpolated SDF value is negative, even if it is not the most negative SDF value at that location.

2.2.2 Interpolators

A practical consideration in implicit modeling is what interpolator is used for the volumetric function to assign SDF values to the unsampled locations. Many interpolation algorithms are available, each with differing properties and parameterization requirements: inverse distance weighted (IDW) (Hosseini & Deutsch, 2007), kriging (Silva & Deutsch, 2012b; Wilde & Deutsch, 2012), locally varying anisotropy (LVA) kriging (Boisvert, 2013; Lillah & Boisvert, 2013) or radial basis functions (RBF) interpolators (Cowan et al., 2003; Hillier et al., 2014; Knight et al., 2007). The RBF framework is popular in the geological modeling literature (Cowan et al., 2003; Hillier

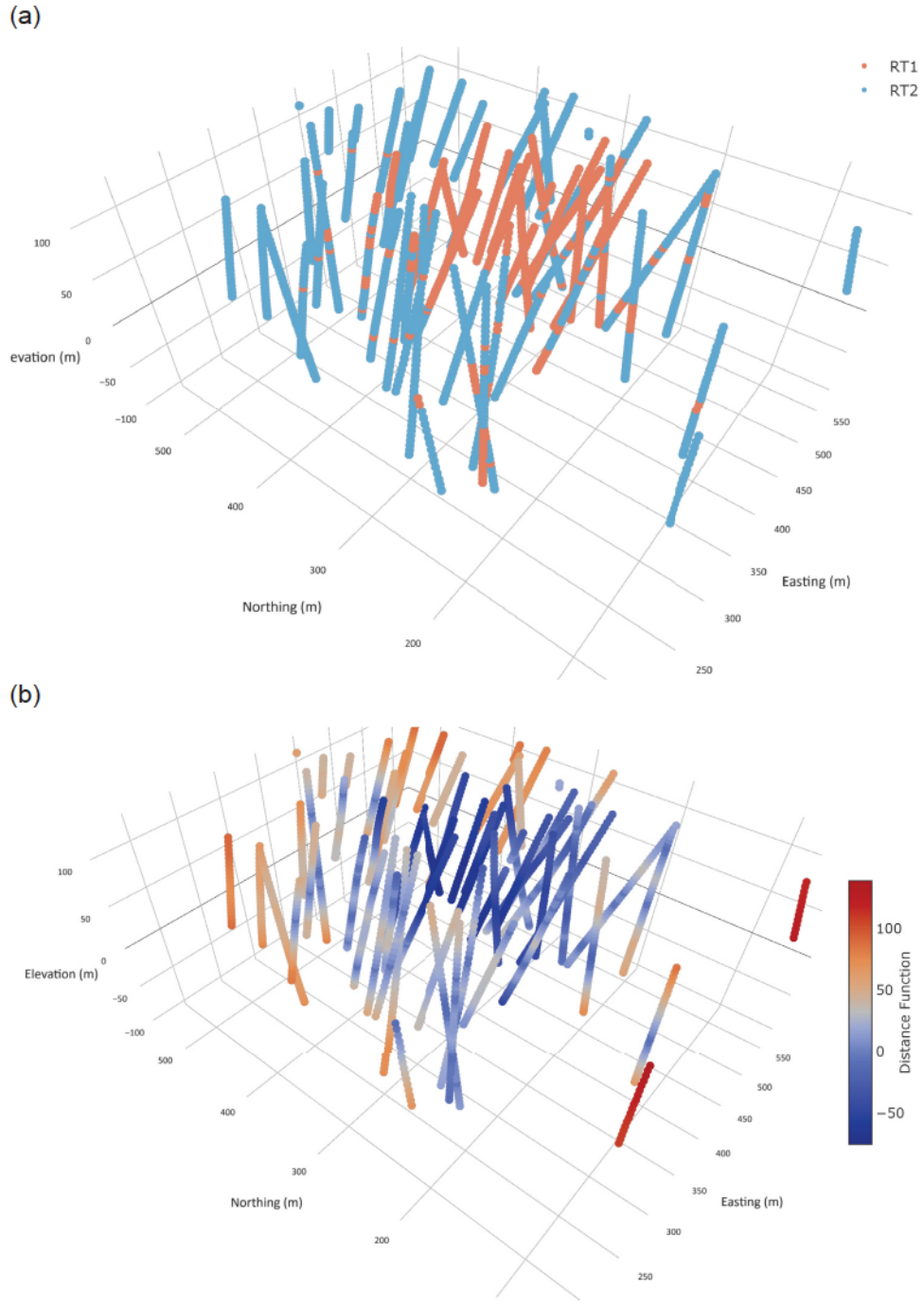


Figure 2.3: Categorical codes along drill holes sampling a geological domain. (a) the categorical codes where 1 is inside and 2 is outside the geological body of interest. (b) The SDF calculated for category 1.

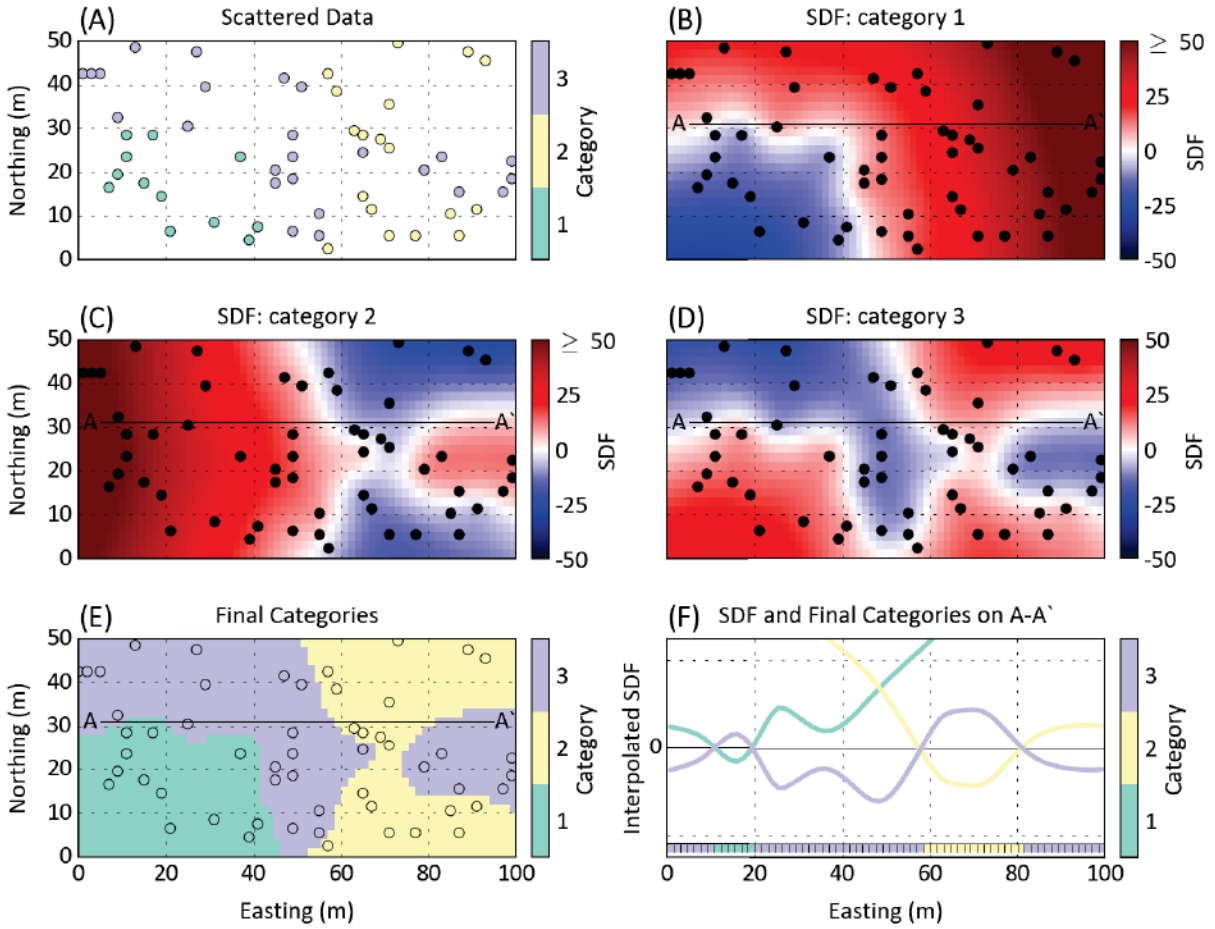


Figure 2.4: K category implicit modeling framework. (a) the set of scattered data for SDF interpolation. (b)-(d) the SDF calculated for indicator 1-3, respectively. (e) the final categories found at each location given the K category configuration found in (a) and the interpolated SDF for each category. (f) a view of the interpolated SDF for each category in this problem along the A-A' transect.

et al., 2014; Knight et al., 2007; Vollgger, Cruden, Aillères, & Cowan, 2015). RBF interpolation is identical to the dual kriging formulation given specific kernel parameterization (Fazio & Roisenberg, 2013; Journel, 1988). The advantages often cited of RBF interpolators include: 1) RBFs can interpolate the SDF without concerns of first-order stationarity; 2) variograms of the SDF are problematic owing to first-order trends and are not required for RBF interpolators; and 3) with respect to geological modeling, RBF interpolation can honor arbitrary shapes instead of depending on a modeled covariance function (Cowan et al., 2003; Hillier et al., 2014; Knight et al., 2007).

Which interpolation algorithm to choose depends on the target properties of the geological boundaries and what information is available for parameterization, e.g., if there is sufficient

sample density for a variogram calculation, or if orientation measurements must be incorporated. IDW interpolation is the simplest to implement, however, kriging is preferred since it minimizes the error variance of the estimate. The main issue with kriging relates to the assumptions on first and second order stationarity, which presents a practical difficulty for SDF interpolation since the SDF is extremely non-stationary (Fig. 2.4b, c, and d), e.g.:

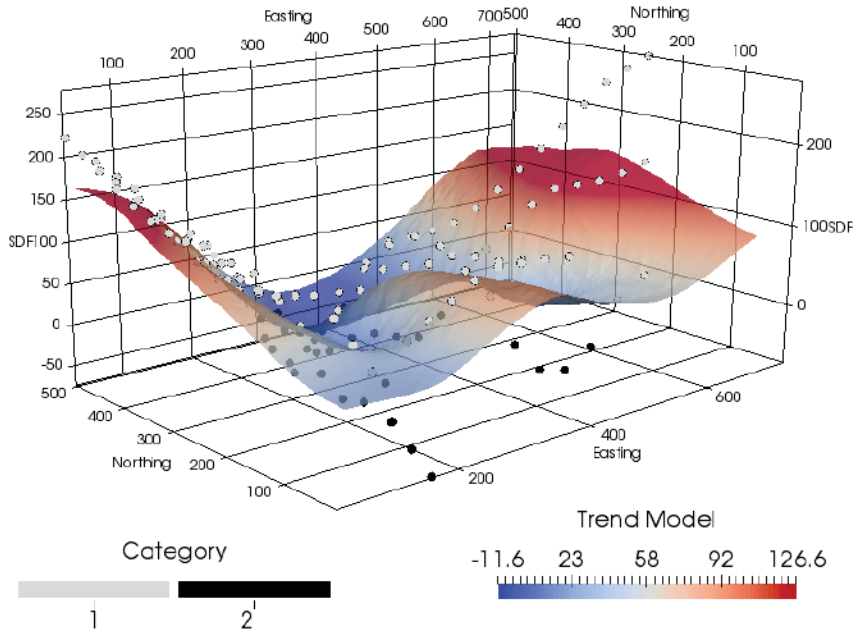
$$E\{z(\mathbf{u})\} \neq m \; \forall \; \mathbf{u} \in \mathbf{A} \tag{2.1}$$

Figure 2.5a shows a 2D categorical modeling problem where the SDF calculated at the sample locations is plotted as the z-value in the oblique rotated view. The surface plotted in the same figure is a trend model calculated using a locally-weighted average of the SDF values. The non-stationary nature of the SDF is evident; it varies from highly negative inside to highly positive outside the domain of interest (category 2 in this case), increasing nearly linearly away from the interface. From the trend model a residual can be calculated from the SDF values; Figure 2.5b plots the residual values on the z-axis. Removing the trend and modeling residuals is a common strategy for geostatistical modeling in the presence of first-order non-stationarity (e.g., Rossi & Deutsch, 2014):

$$z^*(\mathbf{u}) - m(\mathbf{u}) = \sum_{i=1}^{n} \lambda_i (z(\mathbf{u}_i) - m(\mathbf{u}_i)) \tag{2.2}$$

Where the mean m is no longer stationary and instead depends on location $\mathbf{u}$. However, since the trend model is introduced during SDF calculation, residuals either still possess a trend or are planar with no spatial structure (Manchuk & Deutsch, 2015). More recently, the decorrelation of sample data conditional to the trend model is proposed for geostatistical modeling with a trend (Qu & Deutsch, 2017). A simple method to account for this first-order non-stationarity could be to consider ordinary kriging (OK) with a local search. Alternatively, simple kriging with systematic modification of the stationary mean has been proposed as a technique to dilate and erode boundaries calibrated against some known volume (McLennan & Deutsch, 2006). The issue of trends also affects the inference of variograms for the SDF; these issues with kriging estimators have led several authors to prefer RBF interpolators since there are fewer requirements for parameterization and weaker assumptions of first-order stationarity (Cowan et al., 2003; Hillier et al., 2014).

(a) Oblique view of the trend modeled from the SDF dataset



(b) Oblique view of the residuals after removing the trend

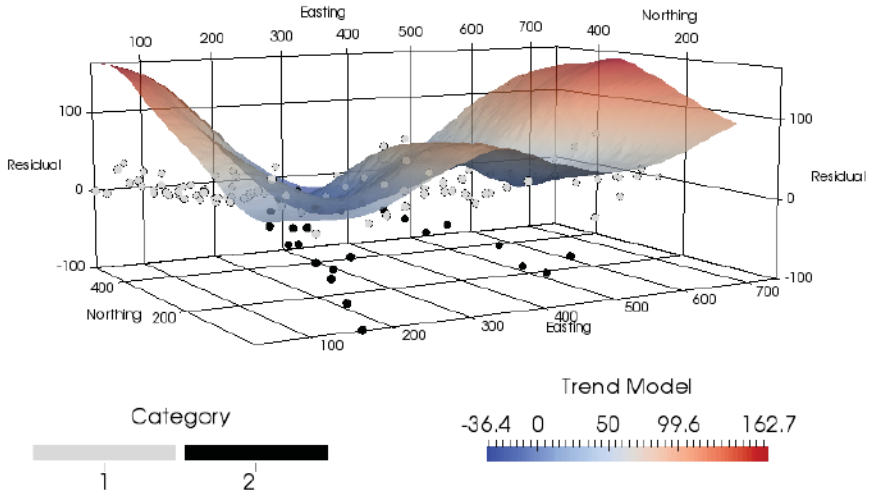


Figure 2.5: Various views of the trend modeled from the SDF, spatial coordinates in meters. (a) Implicit modeling problem with 2 domains, where category 2 is the target of implicit modeling. The z-axis shows the value of the SDF calculated from the locations of the indicator dataset. The surface plotted here is the value of a trend model calculated from the SDF dataset. (b) The z-axis shows the value of the residuals after removing the trend model, points either fall along a plane or still possess a trend after removing the trend from the SDF data.

All discussed interpolators for SDF modeling problems rely on second order stationarity where the spatial relationships between locations are constant throughout the domain. Anisotropy is often present in geological domains but a single set of anisotropic parameters applied globally may not describe all locations in structurally complex domains (Lillah & Boisvert, 2013).

Regardless of chosen interpolation method, the interpolated SDF should be artifact-free to permit further surface generation or extraction of categories on an estimation mesh. Global interpolators ensure an artifact-free map. Search-restricted interpolators like OK or IDW require additional consideration to ensure that artifacts in areas of important thresholds (e.g., near 0) are minimized. A global variant of kriging is generally preferred to search-restricted kriging (Silva & Deutsch, 2012b). In this thesis, the RBF interpolation framework is explored further as it provides interesting opportunities for geological modeling.

2.2.3 Radial Basis Function Interpolator

Consider a set of sample locations in R^3 represented by $\mathbf{u}_i = (x, y, z)$ where the value of the SDF, $s(\mathbf{u}_i)$, is calculated for all $i = 1, \ldots, N$ locations. The RBF interpolator learns a function f that interpolates $s(\mathbf{u}_i)$ at all $\mathbf{u}_i$ sample locations. This interpolant $f(\mathbf{u})$ is a grid-free continuous function defined for all locations $\mathbf{u}$. The value of the SDF at any location $\mathbf{u}$ is given by the weighted linear combination of all data evaluated on a radial kernel:

$$f(\mathbf{u}) = \sum_{i=1}^{N} \lambda_i \phi(|\mathbf{u} - \mathbf{u}_i|) \tag{2.3}$$

where ϕ is a radial kernel chosen for the current problem (Tab. 2.1) and λ_i is a vector of weights obtained by imposing the constraint that the interpolant must equal the data at the data locations:

$$f(\mathbf{u}_i) = s(\mathbf{u}_i) \quad \forall i = 1, \ldots, N \tag{2.4}$$

Which leads to the linear system of equations:

$$\mathbf{D}\boldsymbol{\lambda} = \mathbf{b} \tag{2.5}$$

where the square and symmetric interpolation matrix $\mathbf{D}$ has components $D_{ij} = \phi(|\mathbf{u}_i - \mathbf{u}_j|)$ for $i, j = 1, \ldots, N$, $\mathbf{b}$ is a column vector containing the value of the SDF, $f(\mathbf{u}_i)$, at location $\mathbf{u}_i$

for $i = 1, ..., N$, and $\boldsymbol{\lambda}$ is a column vector of weights determined by solving the linear system of equations. Once the weights are determined, the value of the interpolant can be extracted for any location (Eq. 2.3) and the interface between geological domains is said to be known implicitly as a function of location, $f(\mathbf{u}) = s$.

Parameterization of an RBF interpolator is problem specific and depends on the data configuration, target surface properties, and structural properties of the geological bodies under investigation. The parameters include the kernel type, the support of the kernel (if required), and the anisotropy present in the domain. The kernel defines the spatial relationship between all i, j points evaluated with the Euclidean distance, and in this context is synonymous with the covariance function typically used in kriging estimators. Global anisotropy can be introduced using an anisotropic kernel. An interpolant is commonly evaluated on a mesh discretized for the given problem with the final step to process and extract various levels of interest in the scalar field with algorithms such as marching cubes (Lorensen & Cline, 1987).

2.2.4 Radial Kernels and Parameterization

A set of RBF kernels are shown in Table. 2.1; each kernel has varying properties but all generate positive definite matrices for the linear system of equations (Eq. 2.3; Fasshauer, 2007). Each kernel is suitable for different applications, like hole filling (for sparse data environments) or compact support for dense data environments or large datasets (Carr et al., 2001; Fasshauer, 2007). Determining the best kernel function/parameterization for a given problem is possible with leave-N-out cross validation (LNOCV) (Fasshauer, 2007; Hillier et al., 2014).

The radial kernel establishes the spatial relationship between all points in the domain (Fig. 2.6). Parameterization of the kernel may draw from a variogram modeling workflow. An RBF kernel is fully parameterized with a support parameter and an optional set of anisotropic parameters consisting of rotation angles and anisotropic ratios which are easily extracted from a model variogram, e.g., $sup = ahmax$, ang_1, ang_2, ang_3, $r_1 = \frac{ahmin}{ahmax}$, and $r_2 = \frac{avert}{ahmax}$ (Deutsch & Journel, 1998; Rossi & Deutsch, 2014). The affect of anisotropy on the radial kernels is shown in Figure 2.6c and d. In this simple example, $ahmin$ is reduced to half the size of $ahmax$, resulting in an $r_1 = 0.5$. The effect here is that locations oriented along the $ahmax$ direction are

Table 2.1: RBF kernels and their properties (Fasshauer, 2007). ϵ is the support factor used for each kernel, defining the kernel 'range'

Kernel	Equation	Properties
Gaussian	$\phi(r) = \exp^{-\epsilon^2 r^2}$	Globally supported, positive definite
Spherical	$\phi(r) = 1.5\epsilon r - 0.5(\epsilon r)^3$	Compact support, positive definite
Inverse Multiquadric	$\phi(r) = \frac{1}{\sqrt{1+(\epsilon r)^2}}$	Globally supported, positive definite
Multiquadric	$\phi(r) = \sqrt{1 + (\epsilon r)^2}$	Globally supported, conditionally positive definite
Cubic	$\phi(r) = r^3$	Globally supported, conditionally positive definite, parameter free, hole filling properties
Thin plate spline	$\phi(r) = r^2 \log r$	Globally supported, conditionally positive definite, parameter free, hole filling properties
Wendlands C2	$\phi(r) = (1-r)_+^4(4\epsilon r + 1)$	Compact support, positive definite
Linear	$\phi(r) = r$	Globally supported, conditionally positive definite, parameter free

Note: $\epsilon = 1/sup$

more related to one another than locations in the $ahmin$ direction at the same distance.

If considering a kriging interpolator, a low-nugget Gaussian variogram fit to an experimental variogram calculated from the SDF dataset should be used since the Gaussian structure is continuous at the origin and reflects the properties of a boundary (Manchuk & Deutsch, 2015). Indeed, the Gaussian model is popular for RBF interpolation (Hillier et al., 2014). In other disciplines a support parameter is estimated from the geometric configuration of the sample locations rather than statistical continuity (Fasshauer & Zhang, 2007). Fasshauer (2007) suggest a support distance estimated from the radius of the largest circle or sphere that can be placed between sample locations within the domain. This can be estimated by finding the maximum distance from any one grid cell in the domain to its nearest data (Fasshauer, 2007):

$$sup = \max_{\mathbf{u} \in \mathbf{A}} \left[\min_{\mathbf{u}_j \in \mathbf{u}} (\mathbf{u} - \mathbf{u}_j)^2 \right] \tag{2.6}$$

It is worth noting that for irregularly distributed data in a domain with a rectilinear grid (as is often the case in geostatistical domains), the support parameter may be over-estimated with this method if extrapolation beyond reasonable domain limits is not accounted for. Geostatistical

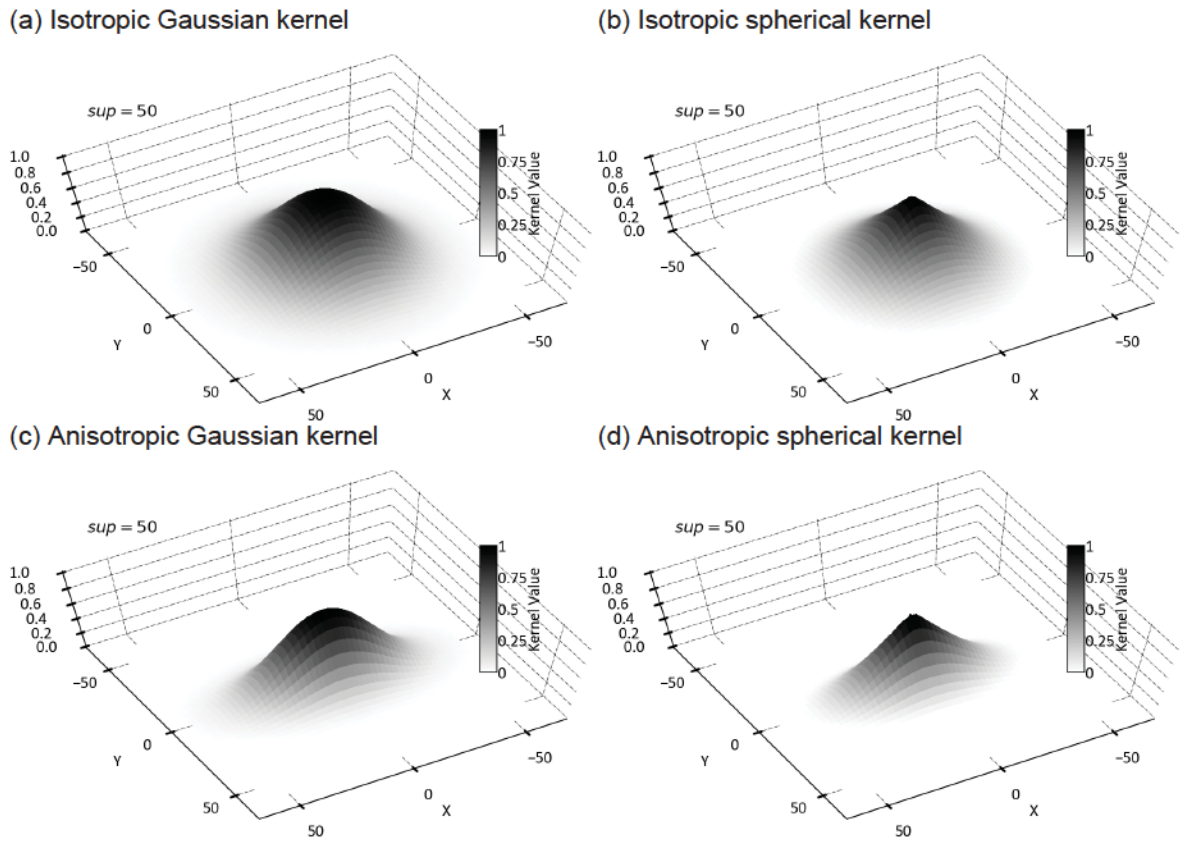


Figure 2.6: Isotropic (a) Gaussian and (b) Spherical radial kernels with a support parameter $\epsilon = 50$. (c) and (d): anisotropic versions of (a) and (b), respectively, where $ang1 = 0$, $ahmax = 50$ and $ahmin = 25$.

domains are also commonly sampled in irregular patterns where sample density increases in areas of interest. In practice this disparity in sample density may result in an estimated support parameter that is not strictly representative of all locations. Regardless, a data spacing analysis is useful in this sense to determine a reasonable support that does not extrapolate excessively from data locations to the unsampled grid locations.

2.2.5 Partitioning

Geostatistical datasets contain 10's to 100's of thousands of data; enough samples that the dense linear system of equations (Eq. 2.5) for a global interpolator is impractical. Methods to overcome this limitation are well researched: a search-restriction following a Markov screening assumption (for a kriging estimator; Deutsch & Journel, 1998); sparse-direct interpolation where the influence of points beyond some distance from one another is negligible (Ohtake, Belyaev, & Seidel, 2006); iterative solving techniques that assume a sparse system of equations and some

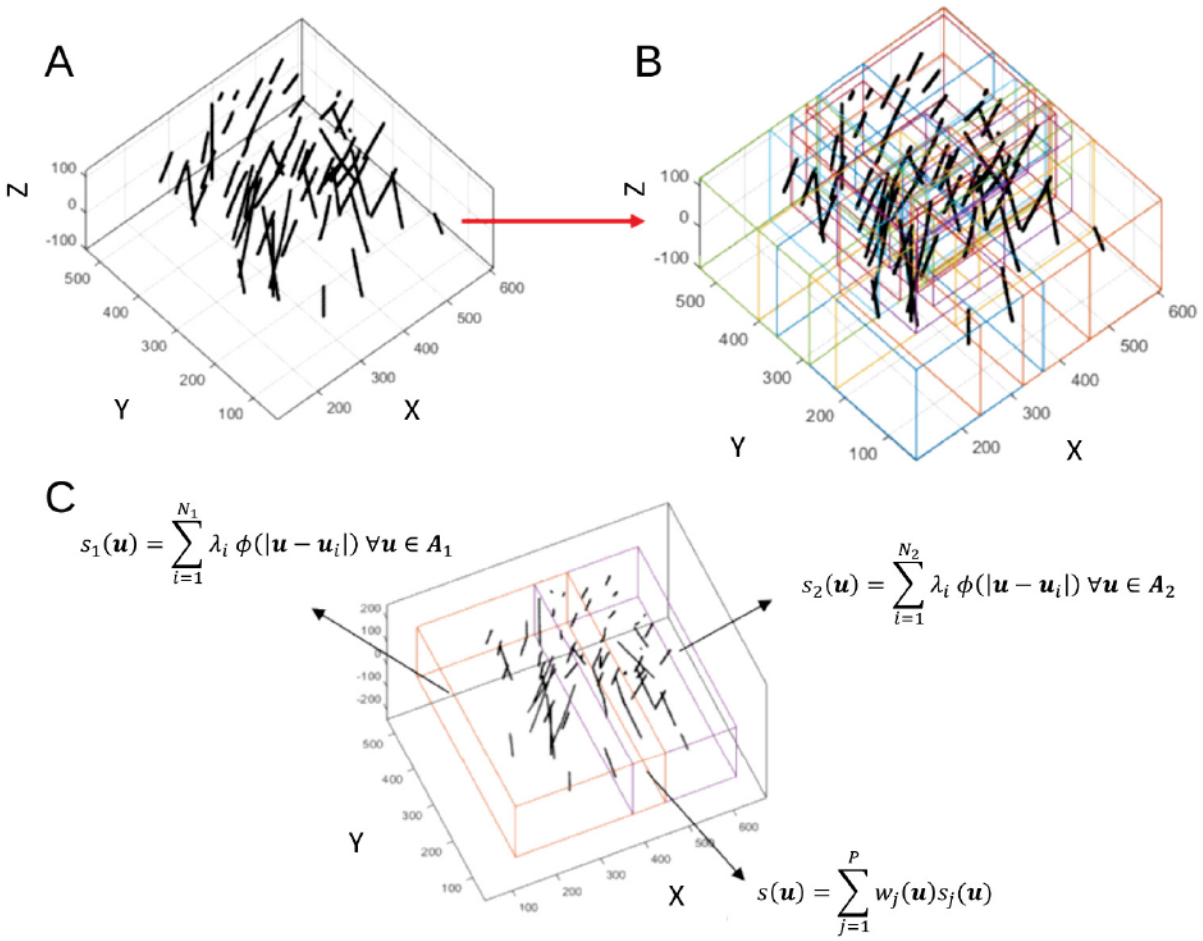


Figure 2.7: Schematic example of a binary partitioning algorithm. (a) the domain with scattered drill holes and mesh, (b) final set of partitions generated from binary decomposition, (c) local interpolants are solved independently in each partition, on overlapping sites the local interpolants $s_j(\mathbf{u})$ are weighted to the global interpolant $s(\mathbf{u})$

level of accuracy (Beatson, Cherrie, & Mouat, 1999; Carr et al., 2001); and domain decomposition techniques which partition the domain into overlapping subsets and solve many small sub-problems independently and often in parallel (Beatson, a Light, & Billings, 2001; Cuomo, Galletti, Giunta, & Starace, 2013; Ohtake, Belyaev, & Seidel, 2003; Pouderoux, Gonzato, Tobor, & Guitton, 2004; Xiaojun, Michael, & Wang, 2005). A partitioning framework has several other beneficial properties when considering geological modeling and is introduced here.

The Partition of Unity (PU) domain partitioning algorithm recursively partitions a domain, $\mathbf{A}$, into P overlapping subdomains, $\mathbf{A}_p$, so that $\cup_{p=1}^{P} \mathbf{A}_p = \mathbf{A}$ (Fig. 2.7). The subset of data contained within each subdomain form a local interpolation problem, each of which is relatively small and can be solved efficiently. To reconstruct the global interpolant, a local interpolant is

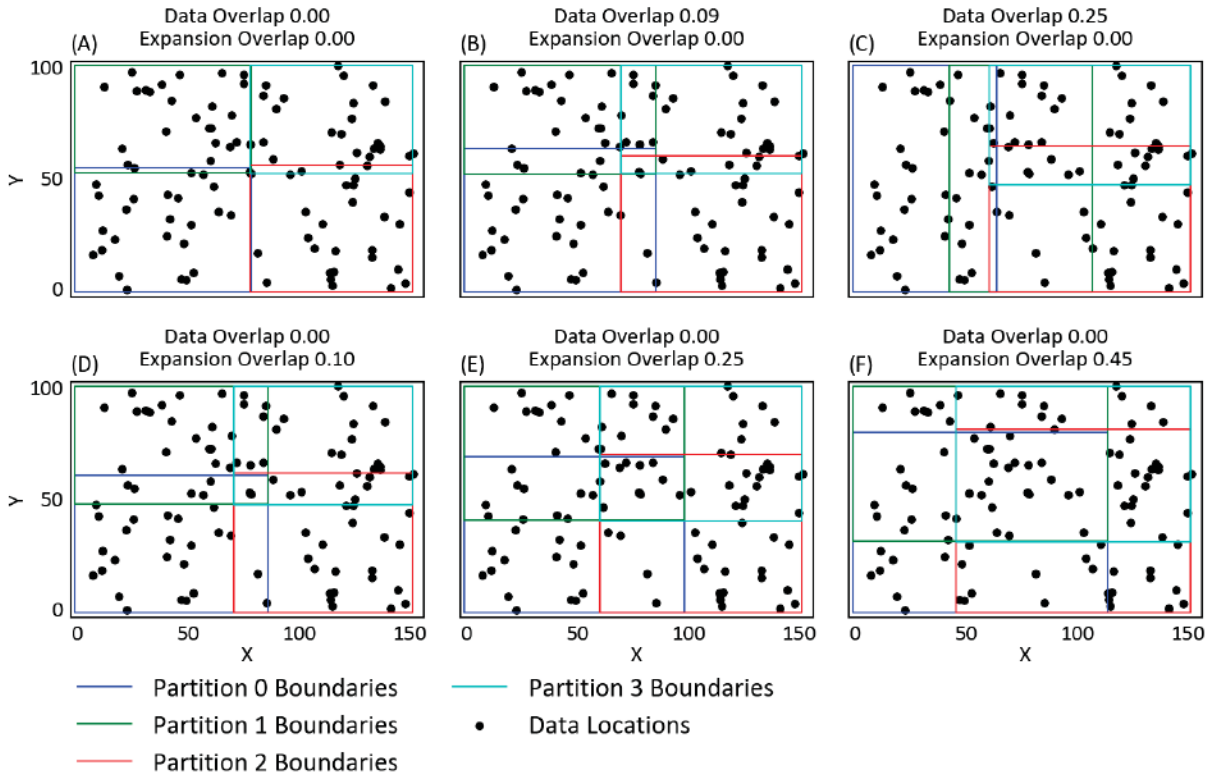


Figure 2.8: Different styles of partitioning overlap to enforce the smooth reconstruction of the global interpolant.

calculated independently in each subdomain, and the global interpolant, $s(\mathbf{u})$, is reconstructed by weighting the interpolants from subdomains on overlapping sites:

$$s(\mathbf{u}) = \sum_{j=1}^{P} w_j(\mathbf{u}) s_j(\mathbf{u}) \tag{2.7}$$

A number of domain partitioning algorithms can be used, including: a regular coarse grid; K-means; bisecting-K-means; binary tree; octree; and others (Klaas & Shephard, 2000; Pouderoux et al., 2004). The simplest partitioning is the regular coarse grid, which can work well for domains of roughly uniform sampling. However, a method that partitions based on the data arrangement is required for geostatistical domains since samples are commonly irregularly distributed. A binary partitioning is considered here since it can rapidly generate partitions with roughly equal data assigned to each, with a suitable amount of overlap, and spanning the associated set of axis-aligned grid locations. Partitions generated by this method also have several valid weighting functions to reconstruct the global interpolant (Pouderoux et al., 2004).

Binary domain partitioning is a recursive partitioning algorithm that considers the total domain as the root node. For each partition a split is considered if the partition contains greater than a user-specific target number of data in each partition. For the current partition, the longest dimension (e.g., either x, y, or z) is calculated from the contained sample data, and the partition is split into 2 overlapping sub-partitions (Fig. 2.7a and c). Overlap between partitions is ensured in two ways: 1) during splitting, a fraction of the data can be taken to be part of both child partitions so overlap is maintained along the boundary (Fig. 2.8b & c); and 2) after partitioning is finalized the partitions can be expanded to guarantee a suitable amount of overlap (Fig. 2.8d, e, and f). In practice, both overlap methods may be required since the data overlap alone can cause issues with recursion for some data configurations and the expansion overlap has potential to expand partitions to encompass too many data (e.g. cyan in Fig. 2.8c versus Fig. 2.8f).

Partitions generated from binary partitioning are axis-aligned rectangles spanning different regions within the domain, and overlapping with potentially several other partitions within the defined coordinate system (Fig. 2.7). Pouderoux et al. (2004) introduce several weighting functions for partitions of this shape:

$$wf = 1 - d \tag{2.8}$$

$$wf = 2d^3 - 3d^2 + 1 \tag{2.9}$$

Where d is the distance from the center of the partition to each cell in the partition, computed as (Pouderoux et al., 2004):

$$d = 1 - \prod_{p \in x,y,z} \frac{4 * (x_p - L_p) * (U_p - x_p)}{(U_p - L_p)^2} \tag{2.10}$$

with x_p the coordinate of the current grid cell, L_p is the lower coordinate of the bounding box, and U_p is the upper coordinate of the bounding box, and the product is taken over each coordinate $p \in x, y, z$. Since the binary partitioning algorithm chooses whether or not to split a partition based on the number of contained data, the final partitions generated from this method are mainly based on the number of data contained within each partition. Thus, there is no guarantee that the data allocated to each partition best reflects all mesh locations allocated to that

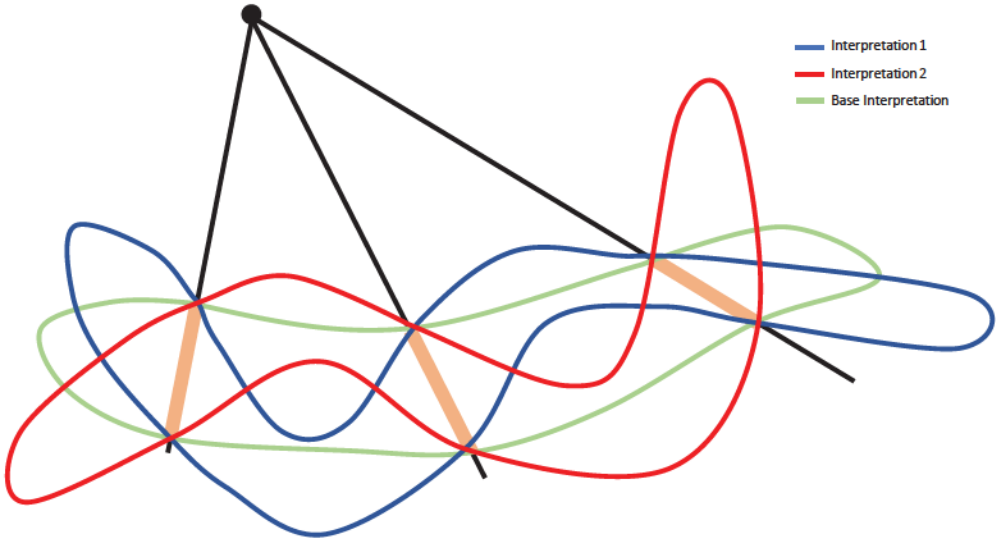
partition. For example, the data may exist only at 1 corner of the partition rather than being distributed uniformly throughout the partition. In this case it is critical that extrapolation of the local interpolant is controlled with some style of clipping.

2.2.6 Uncertainty in Implicit Geological Models

Two general types of uncertainties can be considered in geological modeling projects: 1) volumetric uncertainty; and 2) geometric uncertainty (Fig. 2.9). Volumetric and geometric uncertainties are related; volumetric uncertainty represents the between-sample uncertainty in the volume contained by the bounding surface for a single geometric interpretation. By contrast, geometric uncertainty refers to the uncertainty in the orientation, extent and interconnectedness of the geological bodies; the between drill hole interpretation parameterized by the modeler within the structural setting of the deposit. Both styles of uncertainty are present in all geological modeling projects and are important in the presence of sparse sampling (Lillah & Boisvert, 2013). A critical assumption for most geological modeling projects is that the geological interpretation reasonably accounts for geometric uncertainty, which leaves the more manageable volumetric uncertainty to be quantified. This assumption is reasonable given implicit modeling tools are mostly utilized for the large-scale geological boundaries.

Although geometric uncertainty is assumed to be captured through geological interpretations, this can be a large source of uncertainty when the dataset for implicit modeling consists of point orientation measurements that constrain the orientation of the surface, owing to measurement errors (Lindsay, Aillères, Jessell, de Kemp, & Betts, 2012; Lindsay et al., 2013). Lindsay et al. (2012) propose to simulate sets of orientation input data to generate different boundary models. The suite of boundary models constructed with different simulated inputs allows for the quantification of the probability for each rock type to be found at each location in the domain. In their work the dataset consists solely of point orientation measurements and models are constructed using a potential field co-kriging based interpolation method (Lajaunie et al., 1997; Lindsay et al., 2012). Since there is uncertainty and subjectivity embedded in the measured orientations, their hypothesis is that randomly perturbed sets of input data fed through the modeling algorithm correctly account for these uncertainties. However, for a generalized

(a) Geometric boundary uncertainty



(b) Volumetric boundary uncertainty

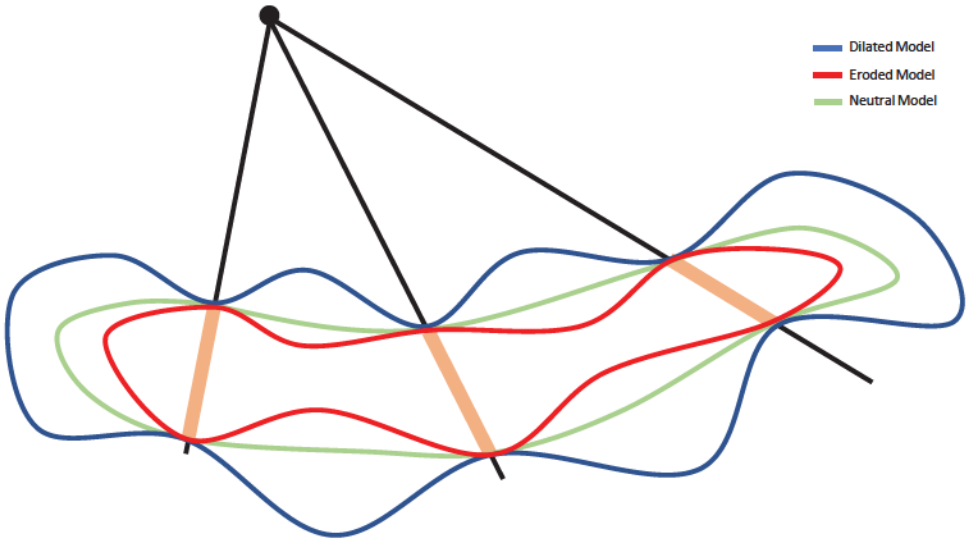


Figure 2.9: Types of boundary uncertainty to consider in implicit modeling projects.

implicit modeling project the main source of information are the locations of indicator data from which volumetric functions are derived. Information on the orientation of different geological boundaries may be secondary. Also, the probabilistic model from which local orientations are simulated is missing from their method (Lindsay et al., 2012). For example, a covariance function should dictate how related orientations are simulated on a particular structure to ensure consistency.

Volumetric uncertainty is important since the volume of the mineralized resource directly affects the tonnage which dictates the total amount of exploitable material found at a given deposit. Lillah and Boisvert (2013) showed that the contribution of boundaries to the total uncertainty is significant in sparsely sampled domains (Fig. 2.10). Thus, in the early stages of a project, or in sparsely sampled areas, the contribution of uncertain boundaries to the total uncertainty will be high relative to the other components affecting the geostatistical model. Several techniques are developed to quantify volumetric uncertainty with implicit modeling (Lillah & Boisvert, 2013; Wilde & Deutsch, 2012). McLennan and Deutsch (2006) proposed to quantify the prior uncertainty in the mean of the SDF with the spatial bootstrap, then generate boundary realizations with SK and a different stationary mean drawn from the uncertain distribution. Alternatively, multiple training-image techniques can be used to combine stochastic and deterministic inputs so that the geometry defined by the implicit model conditions the overall shape and features of stochastic multiple point statistics (MPS) models (Silva & Deutsch, 2012c). This combination of methods can produce both volumetric and geometric uncertainty for a deposit.

For tabular deposits Yamamoto et al. (2015) propose to simulate realizations of vertical drill hole data within some uncertain bandwidth. A similar methodology is proposed by Deutsch (2005) for tabular deposits; the volumetric uncertainty is quantified by rotating the local reference system to the orientation of the tabular deposit and obtaining realizations of the bounding surfaces. These methods can be especially useful if the geometry of the deposit is amenable to unfolding and the bounding surfaces can be extracted and modeled to quantify the distribution of contained volumes.

A separate methodology to quantify boundary uncertainty uses SGS to generate conditional realizations of the SDF directly (Lillah & Boisvert, 2013). Combined with the spatial bootstrap

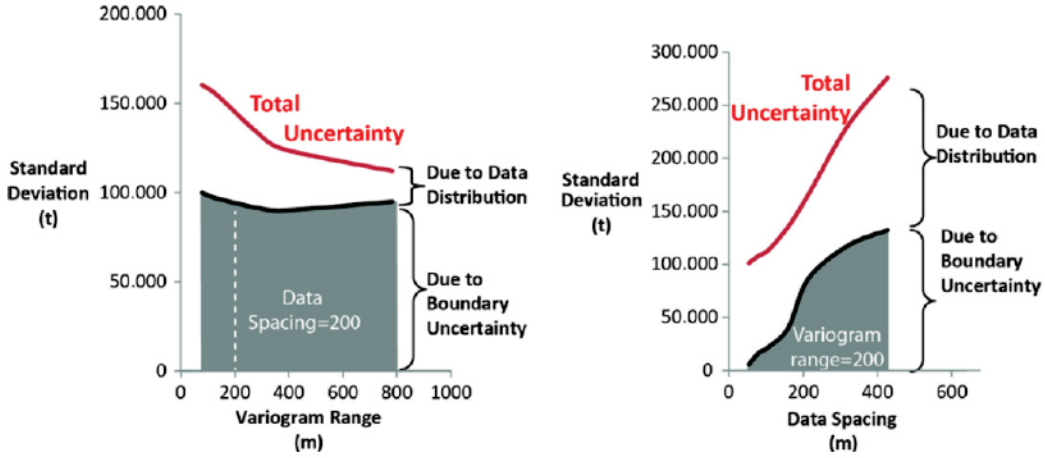


Figure 2.10: Contributions to uncertainty in the total mineralized resource (from: Lillah & Boisvert, 2013).

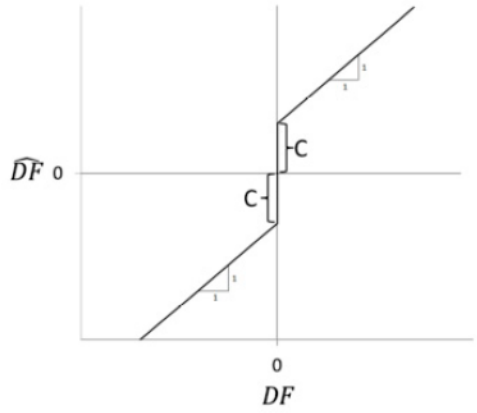


Figure 2.11: The modified distance function considering the single C-parameter (from: Wilde & Deutsch, 2011a)

and other methods for parameter uncertainty this method can provide both a local and global assessment of the volumetric uncertainty. Stochastic simulation of the SDF generally contrasts with the other assumptions inherent in SDF modeling of geological domains (large-scale, variability beyond the data-spacing). The non-stationarity in the SDF also presents practical issues for this method, although in Lillah and Boisvert (2013), an LVA estimation framework accounts for second-order non-stationarity and captures the geological interpretation in the boundary model. The use of SGS to simulate the SDF may generate undesirable features since there is a possibility of short range discontinuities and, over the whole domain, isolated locations inconsistent with the geological interpretation may be simulated.

2.2.7 Uncertainty Bandwidth Parameter

Munroe and Deutsch (2008) proposed a methodology to calibrate a bandwidth of uncertainty in vein-type deposits using a modified SDF. Parameters are added to the SDF calculation to capture the bandwidth of uncertainty and any biases in the volumetric functions. The two parameters, β and C, are added to the SDF calculation as follows (Munroe & Deutsch, 2008):

$$SDF_{mod}(\mathbf{u}_i) = \begin{cases} -(d + \mathrm{C}) * \beta & \text{if } \mathbf{u}_i \text{ is inside} \\ (d + \mathrm{C})/\beta & \text{if } \mathbf{u}_i \text{ is outside} \end{cases} \tag{2.11}$$

where C controls the bandwidth of uncertainty, and β corrects for bias in the calibrated uncertainty bandwidth. The full calibration of both parameters is computationally expensive, which led Wilde and Deutsch (2011a) to simplify the two-parameter uncertainty to a single parameter by omitting β from above (Fig. 2.11). In this case, locations estimated to be less than -C are certain to be inside the model, and locations estimated to be greater than +C are certain to be outside the model (Wilde & Deutsch, 2011a). Any values falling between -C and +C are part of the uncertain bandwidth. Wilde and Deutsch (2011b) generate boundary realizations by unconditionally simulating Gaussian realizations with the same spatial interpolation parameters as the implicit model. Boundary realizations honoring the uncertain bandwidth are then generated by truncating the SDF field between -C and +C. The combination of C-calibration and SDF truncation by unconditional realizations allows assessment of volumetric uncertainty in the boundary model, following the geological interpretation embedded in the implicit model (Manchuk & Deutsch, 2015).

2.3 The Decision of Stationarity

Nearly all geostatistical algorithms make some assumption of stationarity, or explicitly account for non-stationarity, to facilitate the modeling of the spatially correlated SRFs. Common practice is to compile sets of related samples at or near the start of the modeling workflow, which are then considered static for all subsequent geostatistical activities (Fig. 1.2). Stationarity in a geostatistical context is defined in Chapter 1, and states that the samples collected from the

domain are representative of the SRF to be modeled, and the statistics inferred from the sample are representative of all locations in the domain. The geological processes responsible for economic conditions are the result of complex thermal, physical and chemical interactions operating over very long periods of time. Zoning, overprinting, folding and alteration are hallmarks of economic conditions; however, these features imply that several overlapping processes are responsible for the observed conditions. This variability is present at all scales, and thus, geostatistical domains are rarely stationary (Boisvert & Deutsch, 2011; Hadavand & Deutsch, 2017; Qu & Deutsch, 2017).

The motivation for stationary domains follows from the complexity of the geological environment. The assumption is that samples $z(\mathbf{u})$ collected from domain **A** comprise an unknown number K stationary random functions (SRF) $Z(\mathbf{u})$, each with unique first and second order statistical properties. Partitioning the dataset into K mutually exclusive groups to capture the distinct SRFs results in improved numerical models since the estimate at the unsampled location draws from a more representative statistical pool. The stationary delineation operates at the scale of numerical modeling to facilitate statistical inference at the unsampled location.

A number of different methodologies are available to define stationary domains for geostatistical modeling. Common cases are shown in Figure 2.12 and discussed in the following section.

2.3.1 Geologically Defined Stationary Domains

Estimation domains may be naturally defined from the geological logging in cases where specific lithological units have strong control on the spatial distribution of mineralization. For example, in magmatic platinum group element (PGE) deposits it is common for all PGEs to be deposited in specific horizons that are laterally extensive forming a 'reef' of PGE mineralization in the volcanic package (Eckstrand, Roger, & Larry, 2007). In this case there is strong lithological control owing to the fractionation processes, which is readily identified in drill core. Samples identified as being part of the reef would be compiled together to define a single stationary population for numerical modeling. In other geological settings the controls and features identifying different statistical populations may be less prevalent. For example, in a porphyry-style deposit, mineralization is associated with large-scale hydrothermal systems that encompass several rock types,

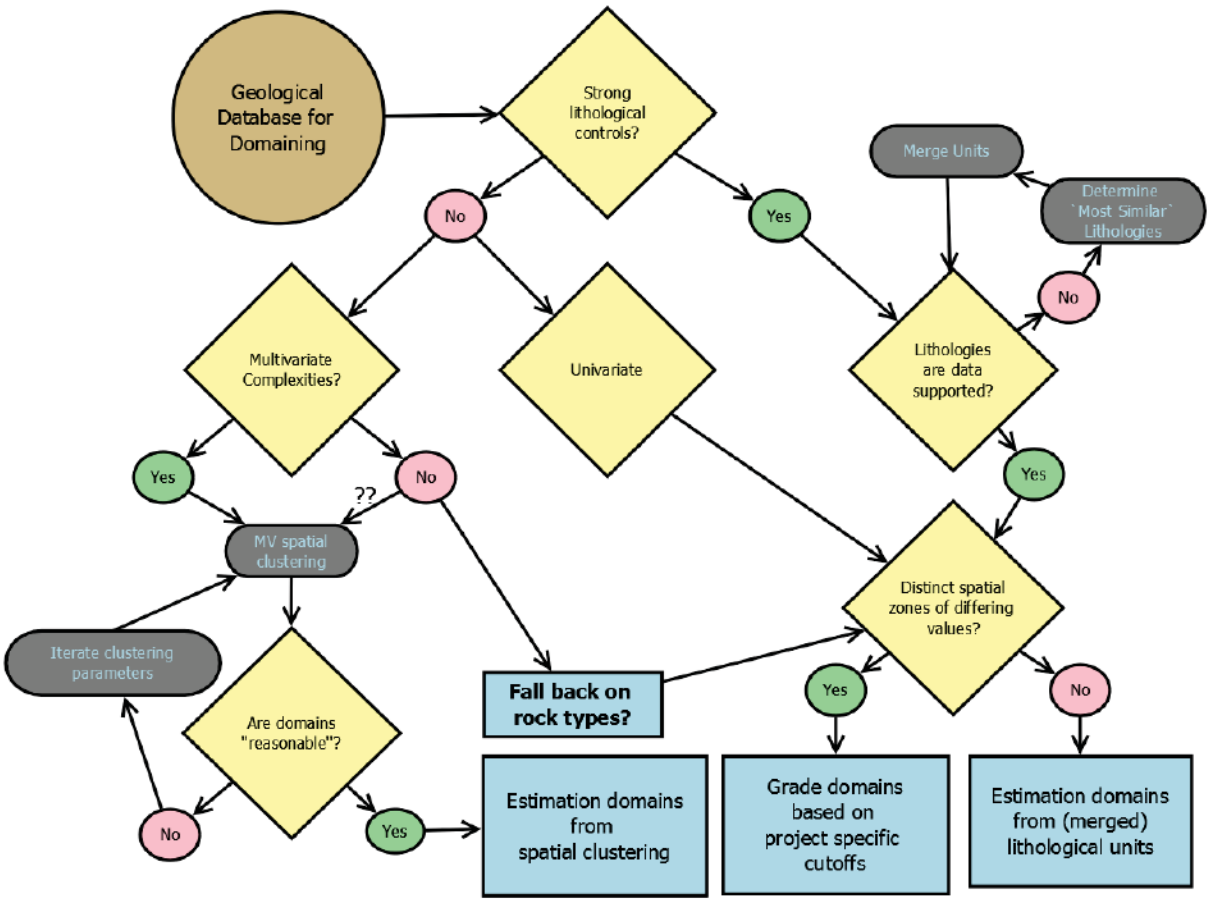


Figure 2.12: A possible workflow to define sets of stationary domains considering the properties of the RV, the spatial configuration, and the other types of geological information available in the core logs.

generating multiple alteration and mineralization styles that overlap and overprint one another (Fig. 2.13). The geological factors may have less consequence for defining stationary domains due to the diffuse and pervasive style of mineralization. However, geostatistical problems are increasingly multivariate with 10's to 100's of variables with complex inter-variable relationships that must be captured by the numerical models (Barnett, 2015). Simultaneously considering the spatial, multivariate and geological characteristics recorded in a set of drill cores is a complex problem.

2.3.2 Merged-Lithology Stationary Domains

A common issue faced in resource estimation is when there are many lithological units, or combinations of factors explaining economic conditions, which are geologically justified, but there are either too few samples to support statistical inference or too many units defined than can

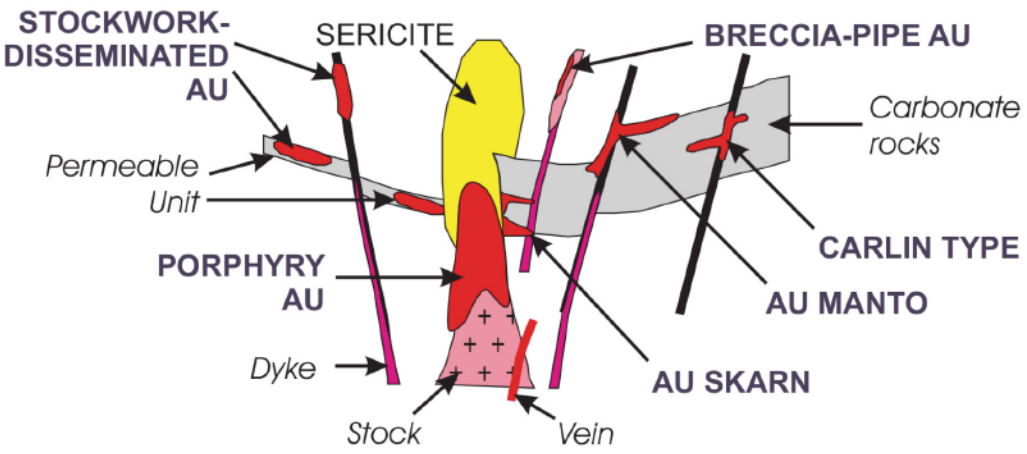


Figure 2.13: Conceptual setting(s) for gold mineralization associated with intrusions (Dubé & Gosselin, 2007)

be practically modeled. In this case the geostatistician must combine a 'large' set of lithological units into 'fewer' (< 7-10) estimation domains that are geologically reasonable, have enough data and capture important mineralizing processes. This workflow must consider the geological, spatial, and statistical properties of each unit to determine reasonable merges and generate a more concise set. Rossi and Deutsch (2014) describe this workflow in detail and note that although the decisions for merging are subjective, they are statistically and geologically supported and ultimately provide the modeler with better knowledge of the database, issues, and the relationships between the domains. This workflow is a significant undertaking since the first and second order statistics, log-quantile-quantile (Q-Q) plots, spatial arrangements and geological properties must be compared to define a metric of 'significant' (dis)similarity to validate each merge. Contact relationships should be investigated following merging to determine the behavior of the RV across boundaries (Fig. 2.14). Better knowledge of the database gained from this workflow allows further justifications for dealing with problem samples, and the insights gained from this analysis improves the parameter inference for the final sets of modeling domains.

2.3.3 Grade-Value Stationary Domains

Grade domaining is a common practice typically applied inside homogeneous lithological units where samples are separated into distinct populations following project specific grade-cutoffs. Cutoffs may correspond to different destinations, or may be based on some combination of

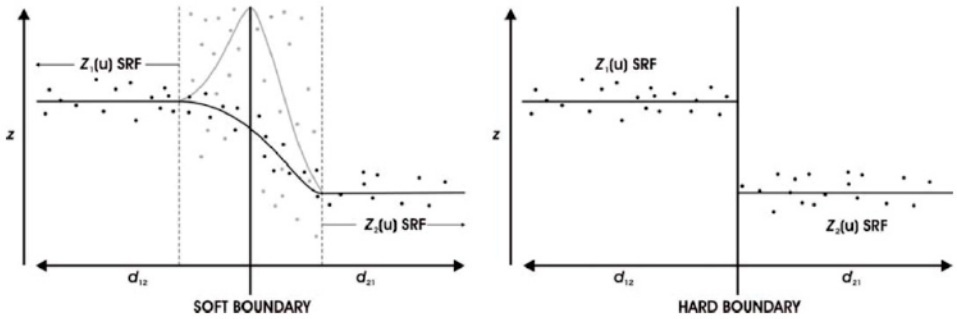


Figure 2.14: Example of contrasting behavior of the RV across contacts. (from: McLennan, 2008)

factors influencing the processing of the materials. This type of domaining could be justified statistically when variables are lognormally distributed (Emery & Ortiz, 2005; Manchuk, Leuangthong, & Deutsch, 2009). Grade domaining is typically undertaken to prevent 'smearing' of high grades into low grade zones and vice versa (Emery & Ortiz, 2005; Leuangthong & Nowak, 2015; Leuangthong & Srivastava, 2012). The delineation of separate populations must be paired with a suitable boundary modeling method which prevents the extrapolation of high grade zones into low grade areas preventing smearing (Dominy & Edgar, 2012; Jewbali, Perry, Allen, & Inglis, 2016; Leuangthong & Nowak, 2015; Stegman, 2001). Grade domains are commonly defined either using the pre-determined cutoffs, or based on inflections in the log-Q-Q plots (Emery & Ortiz, 2005).

Methodologies for multivariate 'grade domains' are only recently addressed with research into spatial clustering. As Figure 2.12 shows, conventional workflows to define stationary domains are poorly suited to complex environments with multiple target variables.

Since grade domains are arbitrary subsets of a spatially correlated RV, samples found in adjacent domains are expected to be correlated (Emery & Ortiz, 2005; McLennan, 2008). Contact analysis and adopting some form of population mixing strategy near the boundaries, or considering the stochastic fluctuations in the location of boundaries is required to properly honor the behavior of the RV at the boundaries between grade domains (Emery & Ortiz, 2005).

2.3.4 Clustering Stationary Domains

Cluster analysis identifies structure in M-dimensional attribute space. An example where M=3 is shown in Figure 2.15. A cluster, by definition, requires a similarity metric, which ensures that

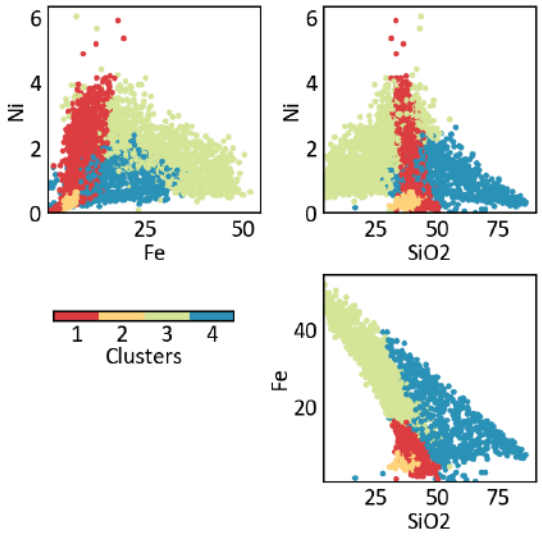


Figure 2.15: Example of hard clustering of a 3-dimensional multivariate space where each sample assigned one of $k \in \{1, ..., K\}$ from a GMM clusterer.

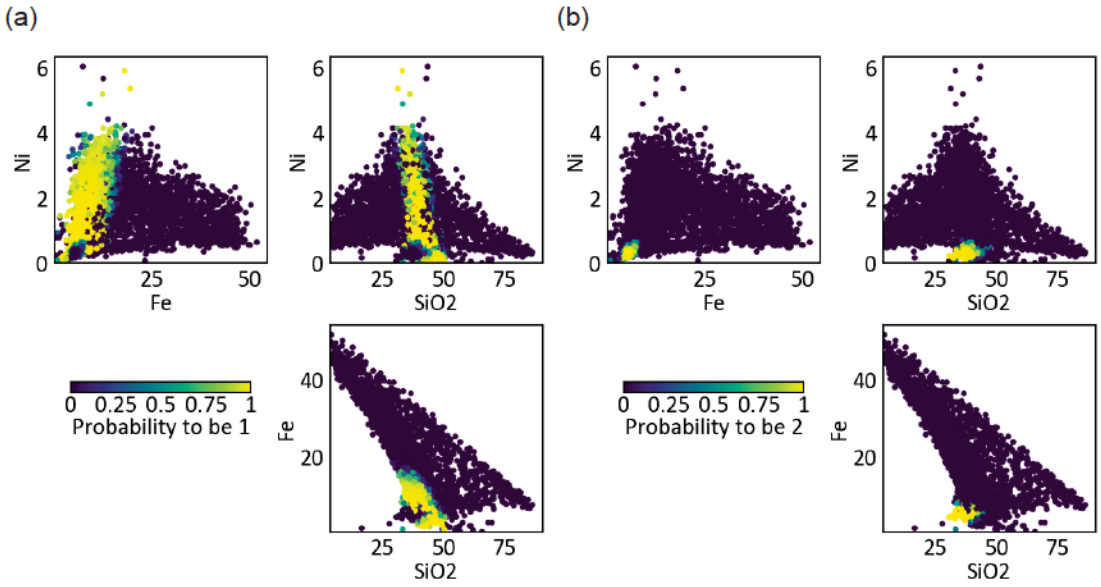


Figure 2.16: Example of soft ('fuzzy') clustering of the 3-dimensional multivariate space in Figure 2.15. Each location assigned a probability $p_k(\mathbf{u})$ to be in each B cluster.

cluster members are more related to each other than they are to members of other clusters. The result of clustering can be discrete or fuzzy. Discreet clusters assign a single category k to each location: $z(\mathbf{u}_i) \in [1, ..., K] \forall \{i = 1, ..., N\}$ (Fig. 2.15). By contrast, fuzzy clustering results in a set of probabilities $p_k(\mathbf{u}_i) \in [0.0, 1.0], \forall \{i = 1, ..., N, k = 1, ..., K\}$ that describe the likelihood that the i^{th} location is part of the k^{th} cluster (Fig. 2.16).

Clustering algorithms are widely applied in reservoir modeling where un-cored wells are

characterized with a set of down-hole geophysical measurements. In such domains, direct observations of the subsurface properties are rare relative to geophysical logs which are available for many un-cored wells. The geophysical responses are interpreted by a trained expert who then delineates the different facies encountered in the subsurface. The classification of geological properties from these logs (called 'electrofacies') is a challenging task; the complexity, time and experience required to process geophysical logs has motivated alternative automated methods to characterize the major units from the well response variables, both to reduce biases and increase efficiency (Mathisen, Heon Lee, & Datta-Gupta, 2001; Perez & Datta-Gupta, 2005; Torghabeh et al., 2014). However, conventional clustering and data-mining algorithms can produce unsatisfactory results since the logs may not clearly delineate distinct and spatially significant facies for geostatistical modeling (Chautru, Chautru, Garner, Srivastava, & Yarus, 2017).

Clustering algorithms follow two main paradigms. Clusters can be represented as a prototype object; for example, represented by a M-dimensional mean or Gaussian kernel (Hastie, Tibshirani, & Friedman, 2009). Alternatively, clusters can be represented in a recursive partition hierarchy where the top partition contains all samples, and the bottom N partitions each contain a single sample. This partition hierarchy is commonly called hierarchical clustering and can either be agglomerative, starting from the bottom and agglomerating upwards, or divisive, starting at the top and dividing partitions downwards (Hastie et al., 2009).

2.3.4.1 K-means

K-means is a popular prototypical clustering algorithm because of its simplicity and speed (Jain, 2010). In K-means, K clusters are represented as the mean of the members of each k cluster. This algorithm operates solely in M dimensional attribute space by minimizing the within cluster variance (Alg. 1; Hastie et al., 2009). K-means results in discrete clustering with a single cluster code k assigned to each location i (Fig. 2.15). A required input to K-means is the number of clusters K. Also, the type of initialization must be chosen. Some practitioners randomly initialize cluster centers for L clustering realizations and choose the single configuration that results in the minimum within-cluster variance over all L. Alternatively the `kmeans++` algorithm initializes

roughly uniformly distributed cluster centers (Arthur & Vassilvitskii, 2007).

The clusters found by K-means are roughly equally dimensioned in multivariate space. This is a good first pass method to generalize and infer initial properties from the dataset, however, this method can produce unsatisfactory results for domains with more complex multivariate relationships, specifically, where features are linearly correlated. For example, consider the relationship between Fe and SiO_2 in Figure 2.16. Linear correlation is clearly an important component of this dataset; a clusterer that accounts for this correlation will benefit the modeling of multivariate properties.

Algorithm 1 K-means Clustering

1: Choose K;
2: Choose the initial cluster centers, randomly or by `kmeans++` (Arthur & Vassilvitskii, 2007);
3: **while** no meaningful change **do**
4: Calculate M dimensional Euclidean distance between each cluster center and each sample;
5: Reassign samples to each clusters based on the minimum calculated distance;
6: Update the cluster center coordinates from the set of points assigned to each cluster;
7: $i(\mathbf{u}_i)$ given by taking the code from the cluster center with the minimum distance to $\mathbf{u}_i$;

2.3.4.2 Gaussian Mixture Model Clustering

Gaussian mixture models (GMM) are another popular prototype clustering method where the clusters are represented by M-dimensional Gaussian kernels. Contrary to the K-means algorithm the GMM clusterer generates a fuzzy clustering by taking the likelihood for each sample to belong to each Gaussian kernel. A final discrete clustering is also available by assigning each sample to the Gaussian kernel that gives the maximum likelihood. The algorithm for GMM clustering is shown in Algorithm 2. A notable difference between this method and K-means is that the Gaussian kernels may be linearly correlated, and thus elongated clusters in M-dimensional attribute space can be found.

The hard and soft clusterings shown in Figure 2.15 and 2.16, respectively, are generated with a 4-component GMM fitted to the multivariate dataset. The choice of how many components is also an input to this clustering method. For this analysis the number is chosen to be 4 since there are 4 rock types defined in the geological logs.

Algorithm 2 GMM Clustering

1: Choose K;
2: Choose the initial cluster centers, randomly or by `kmeans++` (Arthur & Vassilvitskii, 2007);
3: **while** !converged **do**
4: Compute the log-probability for each location to belong to each parameterized Gaussian kernel;
5: Update the kernel weight, mean and covariances given the set of samples that have the maximum likelihood to belong to each kernel;
6: $p_k(\mathbf{u}_i)$ given by computing the probability under each Gaussian kernel (e.g., Fig. 2.16);
7: $i(\mathbf{u}_i)$ given by taking the code from the kernel with the maximum likelihood at $\mathbf{u}_i$;

2.3.4.3 Hierarchical Clustering

Hierarchical clustering contrasts the prototype methods with a recursive partitioning of M dimensional attribute space. The partitioning can either be agglomerative or divisive (Hastie et al., 2009); here the focus is on the agglomerative case. The domain starts with N clusters, one for each sample. Clusters are iteratively merged by finding pairs of clusters that are the most similar based on a chosen similarity metric. Merging continues until there is a single partition left at the topmost level (Alg. 3). Hierarchical clustering requires a metric describing the similarity between two clusters; a thorough summary of such 'population difference' metrics is provided in Chapter 4.

A byproduct of the iterative merging in this style of clustering is a record of the relative distances between merged clusters at each level in the clustering hierarchy. This relative distance can be represented in visual form as a dendrogram (Fig. 2.17, bottom). The visual inspection of the dendrogram permits interpretation of the clustered structure of the dataset, which in turn allows a suitable choice of K. For this reason hierarchical clustering is often considered in initial investigations as a tool for visual inspection of the number of clusters in a dataset.

Algorithm 3 Hierarchical Clustering

1: Compute a pairwise distance matrix $d_{i,j}$ using the chosen distance metric;
2: K = N;
3: **while** K > 1 **do**
4: Calculate the population difference metric for all K*(K-1) cluster combinations;
5: Merge the two clusters that have the minimum population difference;
6: $K = K - 1$;

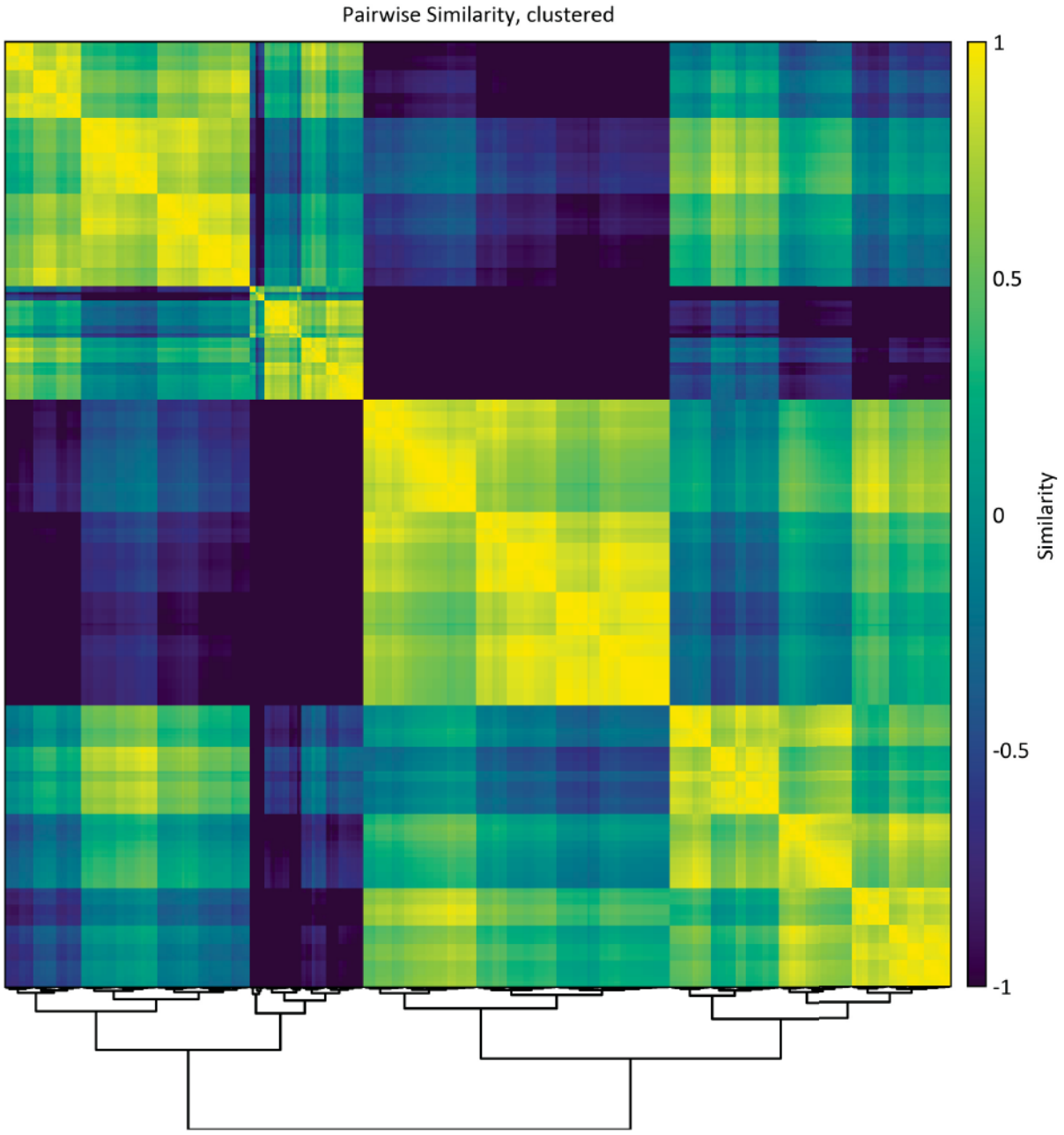


Figure 2.17: Example of hierarchical clustering of a pairwise-similarity matrix for illustration purposes. The bottom axis shows the associated dendrogram which can be visually inspected to determine how many clusters are found in the dataset.

2.3.5 Spatial Clustering for Stationary Domains

Multivariate clustering of spatially correlated data may produce clusters that are overly mixed in the spatial domain, thereby limiting the utility of those clusters for further planning or spatial analysis. This limitation has led several researchers (including this thesis) to explore cluster generation in the presence of spatial correlation.

Spatial clustering describes clustering applied to datasets where samples have some spatial context, e.g., from studies of water chemistry, soil chemistry, epidemiology, ecological studies, geology, and, more recently, natural resource estimation. The goal of clustering geostatistical data is to produce spatially contiguous clusters that have multivariate significance. Thus, the partitioning of M dimensional space is coupled to Cartesian space to ensure that the defined multivariate classes are spatially connected. Two general methodologies have been applied: 1) some form of neighborhood constraint to limit relatedness of distant and uncorrelated samples (Ambroise & Govaert, 1998; Fouedjio, 2016a; Oliver & Webster, 1989; Romary, Ors, Rivoirard, & Deraisme, 2015); and 2) generation of a secondary dataset calculated from the original data with local autocorrelation statistics (Scrucca, 2005).

2.3.5.1 Neighborhood Constraints

Oliver and Webster (1989) modified the pairwise similarity matrix (e.g., Fig. 2.17) using a secondary matrix calculated from the spatial interconnectedness of the sample locations. They justified a variogram model as suitable for generating the secondary matrix, effectively decreasing the relatedness in the pairwise similarity for points separated by large distances. The modified pairwise-similarity matrix is clustered with conventional algorithms to identify multivariate clusters that are spatially contiguous. In their work the spatial compactness of clusters can be tuned by modifying the range and shape of the variogram model. Fouedjio (2016a) propose a similar kernel-based modifier that incorporates direct and cross variogram measures of spatial similarity. Romary et al. (2015) use a spatial neighborhood constraint in hierarchical and spectral clustering algorithms which ensures that only samples in related spatial neighborhoods are paired with one another. Their spatial neighborhood is defined with Delaunay triangulation, modified for 3D to account for the contrasting sample spacing along relative to between drill holes.

The methodology from Ambroise and Govaert (1998) is mainly focused on image segmentation; the local related neighborhood is implied for each pixel by the adjacent pixels. Treatment of the contrasting sample density along versus between drill holes is key for obtaining reasonable spatial clusters with geostatistical datasets.

2.3.5.2 Local Autocorrelation

The second strategy for considering spatial information in clustering algorithms uses local autocorrelation statistics and conventional multivariate clustering algorithms (Scrucca, 2005). The main idea is to calculate a new dataset using local autocorrelation statistics such that the magnitude and sign of the original data is integrated with a measure of the relatedness to nearby locations (Fig. 2.18; Anselin, 1995). Several local autocorrelation statistics are available, two of which are listed in Table 2.2.

To calculate the local autocorrelation, the formulas in Table 2.2 require a spatial weight matrix that specifies spatial relationship between each i, j location. In a simple case this could be a binary weight matrix:

$$w_{i,j} = \begin{cases} 1 & \text{if } i,j \text{ related} \\ 0 & \text{otherwise} \end{cases} \tag{2.12}$$

where *related* requires some spatial neighborhood search, or Delaunay triangulation (as in: Romary et al., 2015). Scrucca (2005) uses a local radius search where samples falling outside some radius from the current location are considered unrelated. Alternatively this weight matrix could contain the inverse distance weights for some number of nearest neighbors around each location.

2.3.6 Validation of Clusters

The validation of a set of clusters is an interesting problem. As clustering is unsupervised, technically there are no 'true' labels for validation. To demonstrate new clustering algorithms some researchers consider clustering a set of multivariate data that has labels. The developed clustering algorithms are used to predict the known labels and the performance of the algorithm against other similar clustering algorithms can be assessed. Validation in this sense is external since

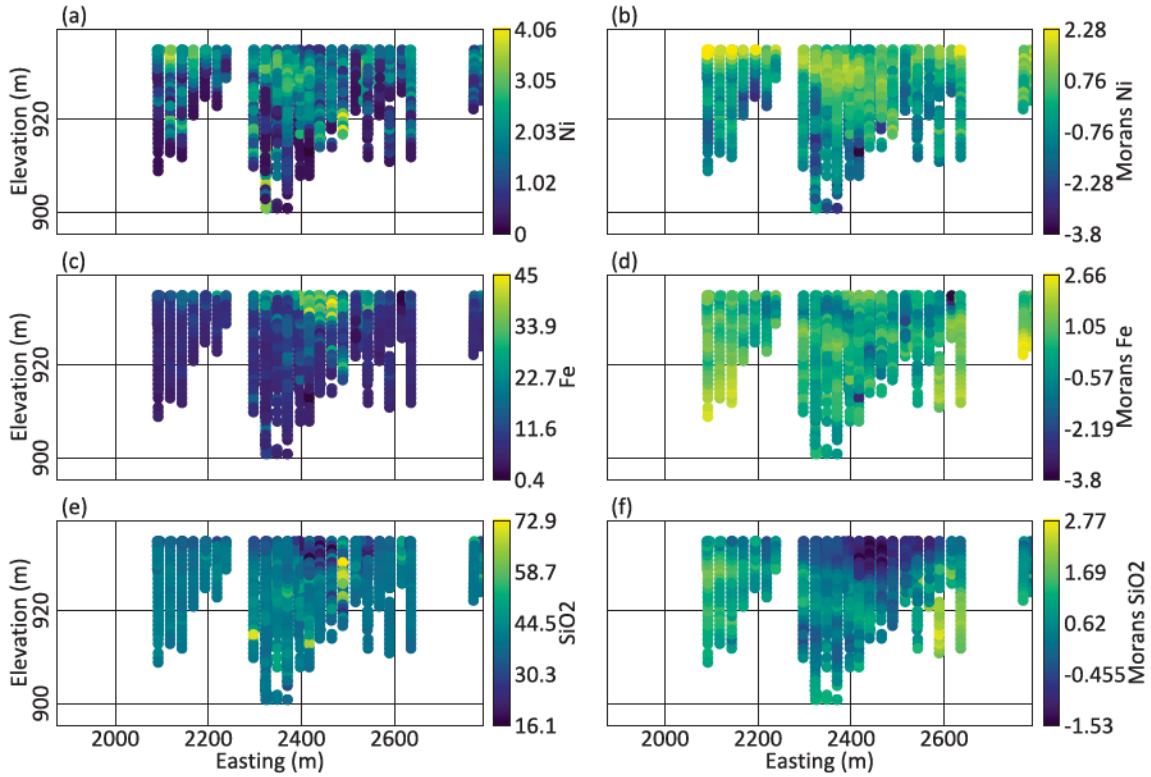


Figure 2.18: Cross sections of the original variable values and the normal scored Moran's local autocorrelation values calculated at the sample locations.

Table 2.2: Local autocorrelation statistics and their properties (Anselin, 1995)

Statistic	Equation	Properties
Morans I	$I_{i,\alpha} = \frac{z_{i,\alpha}\sum_{j=1}^{n} w_{ij} z_{j,\alpha}}{\frac{1}{n}\sum_{j=1}^{n} z_{j,\alpha}^2}$	$\forall i = 1,...,N, \forall \alpha = 1,...,M$. Large positive I_i indicate similar values surrounding the i^{th} location whereas negative I_i indicate dissimilar values surrounding the i^{th} location
Getis G	$G_{i,\alpha} = \frac{\sum_{j=1}^{n} w_{ij} x_j}{\sum_{j=1}^{N} x_j}$	$\forall i = 1,...,N, \forall \alpha = 1,...,M$, Large positive G_i values indicate clusters of high values surrounding the i^{th} location whereas large negative G_i values indicate clusters of low values surrounding the i^{th} location.

w_{ij} is the weight matrix that defines relationship between location i and j
for n samples in a local search, for $i = \{1,...,N\}$ locations, $\alpha = \{1,...,M\}$ variables
$z_{i,\alpha} = (x_{i,\alpha} - \bar{x}_\alpha)$

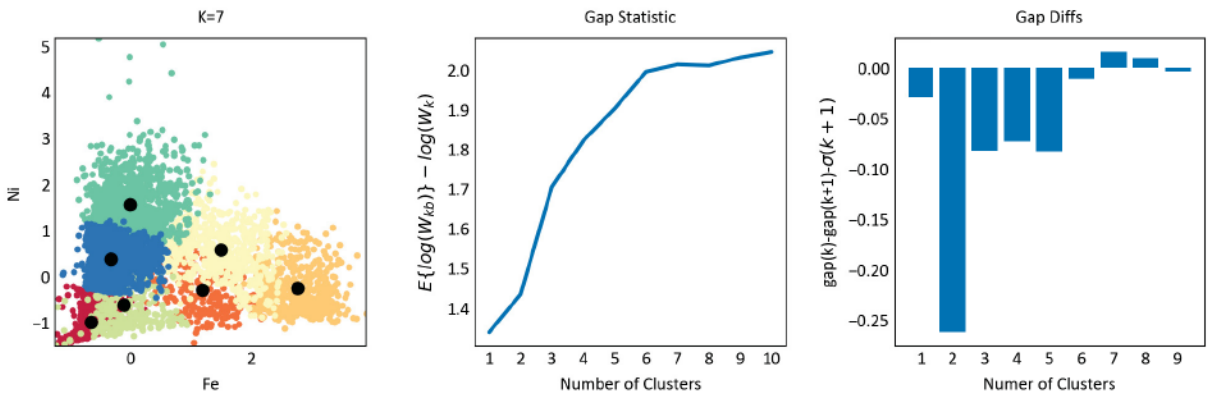


Figure 2.19: Gap statistic calculated for the 3-variate space shown in Figure 2.15.

a set of labels external to the information used to form clusters is used to justify how clustered the configuration is. Alternatively, validation may also take the form of internal metrics where the compactness of the clusters measured with a clustering metric is used to express the relationship between cluster members and prototypes. This is fundamentally similar to determining the K in a given dataset. Determining the K is a subject of much research in the clustering literature; measures such as silhouettes (Rousseeuw, 1987), the gap statistic (Fig. 2.19; Tibshirani, Walther, & Hastie, 2001), or simply estimating the compactness with the inherent clustering metric (Tibshirani & Walther, 2005) can be used to determine K. However, internal metrics are only useful for validation of one clustering over another based on parameterization.

Validation of spatial clusters for stationary decisions is not typically discussed in the literature (Fouedjio, Hill, & Laukamp, 2017; Romary et al., 2015). Since the spatial clustering tools are deployed as an aid to manual orebody domaining, the validation of the result comes from subjective assessment given the geological knowledge of the domain. The generated domains may be statistically supported such that the first and second order stationary characteristics of the domains are improved over other sets of domains, but in general, quantitative validation measures for stationary domains are not developed.

2.4 Ensemble Clustering

Ensemble clustering is a method to extract information from several different clusterings of the same dataset, used to improve results over a single clusterer. Individual clusterings in the

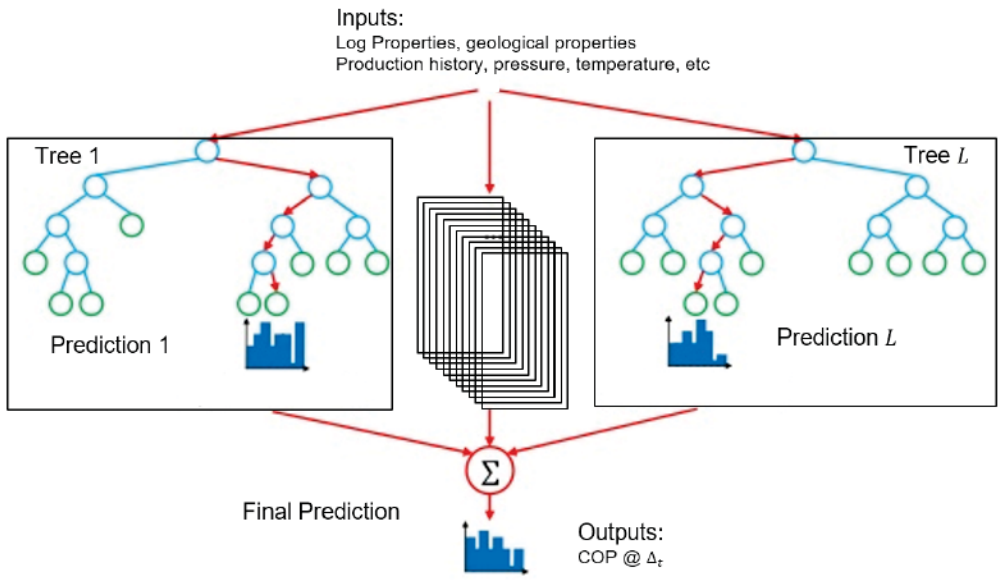


Figure 2.20: Ensemble classification and regression tree predictor (Modified from: KDnuggets, 2018).

ensemble are typically enforced to be diverse, which leads to improved results over a single clusterer as the clusters extracted from the ensemble are less sensitive to parameterization, can fit non-linear features and avoid over-fitting (Topchy, Jain, & Punch, 2005).

A classical example of an ensemble supervised learning technique is the random forest (RF), which comprises L classification and regression tree's (CART), each trained using random-subspace methods to increase the diversity between tree's (Fig. 2.20; Breiman, 2001). Each tree in the ensemble independently makes a prediction of the class label for the i^{th} sample. During training, splits in the tree are chosen to improve the prediction of the response variable. Since each tree is trained against a true response variable, the consensus prediction of the ensemble is the average of all predictions over all trees. The RF can be used in either classification or regression mode. In a classification problem the RF can produce a both fuzzy or discrete classification.

The individual clusterings in an ensemble can be generated by diverse methods, even as simple as random projections to low-dimensional subspaces followed by conventional clustering algorithms applied in that subspace (Rathore, Bezdek, Erfani, Rajasegarar, & Palaniswami, 2018). For the current discussion, consider individual clusterings to be generated using K-means without modifications for spatial features. L training datasets are created using the ran-

dom subspace method and the clustering is generated independently for each dataset. The result of each clustering is stored in an $N \times L$ matrix where each column records the assignment of each sample into K_L clusters for each L realization.

2.4.1 Random-Subspace

The random subspace describes the randomization of the space (variables, set of data, parameters) individual learners in an ensemble are trained on. This technique is commonly used to increase the diversity between learners (Breiman, 2001). For ensemble clustering, each clustering considers a subset of the variables and samples as input. Additionally, the number of clusters or parameters specific to the chosen clustering algorithm can be chosen at random, which ultimately decreases the sensitivity of the outcome to the choice of parameters.

2.4.2 Consensus Functions

A consensus function is required to combine clusterings following the construction of the $N \times L$ clusterings matrix. Strehl and Ghosh (2002) introduce 3 methods by which a consensus clustering can be extracted from a clustering ensemble. The simplest method is to consider an $N \times N$ similarity matrix which is constructed by counting the number of times each sample i is paired with each sample j over all clusterings in the ensemble (Fig. 2.21). In the RF this matrix is referred to as the proximity matrix (Afanador, Smolinska, Tran, & Blanchet, 2016; Breiman, 2001). Finally, clustering this similarity matrix gives clusters that are the consensus of all clusterings in the ensemble. For example, the final clustering could be generated with hierarchical or spectral clustering of the $N \times N$ similarity matrix.

However, this pairwise-similarity-based consensus function scales poorly with the size of the input dataset since the storage and computational requirements are quadratic with N (Rathore et al., 2018; Strehl & Ghosh, 2002). Strehl and Ghosh (2002) develop two other consensus functions that are more complex but scale linearly with N. Romary et al. (2015) alternatively propose a 2-step process where the clustering is obtained on a subset of a large database, followed by supervised classification to predict the cluster labels at the omitted sample locations. Research into alternative ensemble consensus functions is ongoing (Rathore et al., 2018).

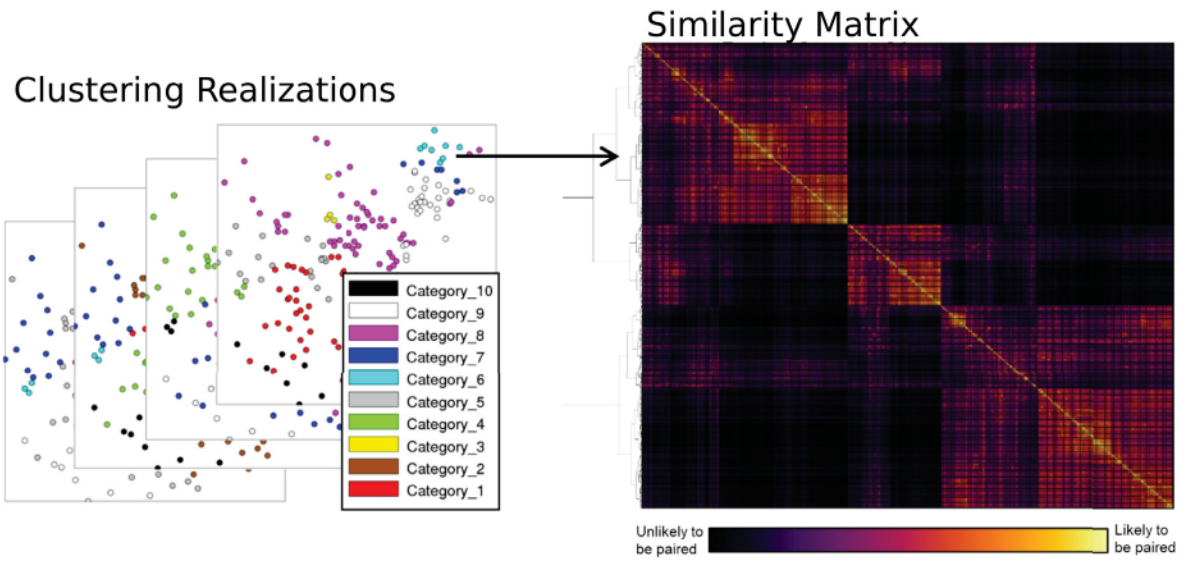


Figure 2.21: Pairwise similarity matrix consensus function. Left: Individual clusterings of a hypothetical dataset. Right: The pairwise similarity matrix, ordered by hierarchical clustering, with a dendrogram on the left axis.

2.5 Uncertainty in Stationary Domains

Two types of uncertainty may be considered when undertaking stationary domaining workflows. First, the allocation of a sample to different stationary groups is inherently uncertain. Two experts undertaking this task for the same dataset are unlikely to code all samples into the same sets of stationary domains. This uncertainty is rarely quantified since there are no strict rules for generating stationary domains, except that they are required, and better stationary domains will improve the quality of the geostatistical models (Rossi & Deutsch, 2014). The second type of stationary domain uncertainty involves the location of boundaries. This was addressed in Section 2.2.6 relating to the large-scale decision of geological stationarity, where uncertainty in implicit geological modeling can be accessed by a calibrated bandwidth of uncertainty (Sec. 2.2.6).

The most widely used method applied to uncertainty in stationary domaining is to consider the allocation of *unsampled* locations to each stationary domain uncertain (Fig. 2.22). In practice this is accomplished using mixing models, co-kriging, or categorical simulation, with mixing relationships or transitional statistics determined globally between stationary domains (Emery & Ortiz, 2005; McLennan, 2008).

The nature of boundaries between different domains is widely studied (McLennan, 2008).

Boundaries can be characterized as either hard or soft. A hard boundary is characterized by abrupt changes in the value of the RV across the boundary. Conversely, a soft boundary is characterized by a gradual change in the RV across the boundary. The nature of the boundaries partly depends on how the boundaries are defined. For example, grade domains defined on arbitrary cutoffs of a spatially correlated RV are expected to have soft boundaries. Without careful attention to the boundary relationships, biases and errors may be introduced to the final numerical models (Emery & Ortiz, 2005).

The nature of the RV near boundaries can be characterized by plotting the expected grade values as a function of distance from the boundary. Using this analysis the contact can either be deemed hard or soft. A strategy for geostatistical modeling near boundaries is required in the presence of soft boundaries. Emery and Ortiz (2005) propose a co-kriging based approach where an estimate made in the transition zone considers samples and statistics from both domains (Fig. 2.14). This requires a linear model of coregionalization (LMC) modeled between domains to capture the inter-domain and between-domain spatial relationships in the transition zone. McLennan (2008) proposes a linear-mixing model that mixes estimates generated in the transition zone independently in each stationary domain. Although this does not consider any non-stationarity in the transition zone, the mixing of estimated values seems reasonable for capturing boundary relationships.

2.6 Validation of Geostatistical Algorithms

Validating numerical models of geological domains is an interesting and important problem and there are several aspects that must be considered. Numerical models can never be fully validated since the exhaustive unknown truth is not accessible (Oreskes, Shrader-Frechette, & Belitz, 1994). The best validation of a numerical model that one can achieve is to compare how predicted values match the measured value found at sample locations. Cross validation (CV) and analysis of prediction errors provides an objective measure of the predictive performance of a statistical model on the collected samples (Fasshauer & Zhang, 2007; Hastie et al., 2009; Rossi & Deutsch, 2014). However, this style of validation does not validate absolutely; there is no measure of goodness against the full truth since this truth is inaccessible. Instead, this

style of validation can be thought of as verification; useful for comparing modeling algorithms or different parameterizations of the same algorithm (Rossi & Deutsch, 2014). The application of CV usually considers a K-fold, where the dataset is partitioned into K disjoint random subsets and the values in the K^{th} partition are considered to be the truth and predicted using the remainder of the dataset (Fig. 2.23; Hastie et al., 2009). This is synonymously referred to as leave N out cross validation or sometimes jackknife. Alternatively if $K = N$, this is simply referred to as cross validation, where each sample location is re-estimated considering the entire dataset. For drill hole datasets in geostatistics, the K-fold must partition based on full drill holes in order to generate representative validation metrics.

Generally the most important validation in geostatistics is obtained by reconciling predicted grades and tonnages with production data. However, as this requires production, it can hardly be used to validate near the beginning of a project. More commonly a set of geostatistical algorithms can initially be validated through K-fold methodologies. Once higher-resolution information is available, either through additional drilling campaigns, delineation drilling, grade control or production, the grades predicted from the initial model must be reconciled with the new information, and model parameters updated to account for any differences.

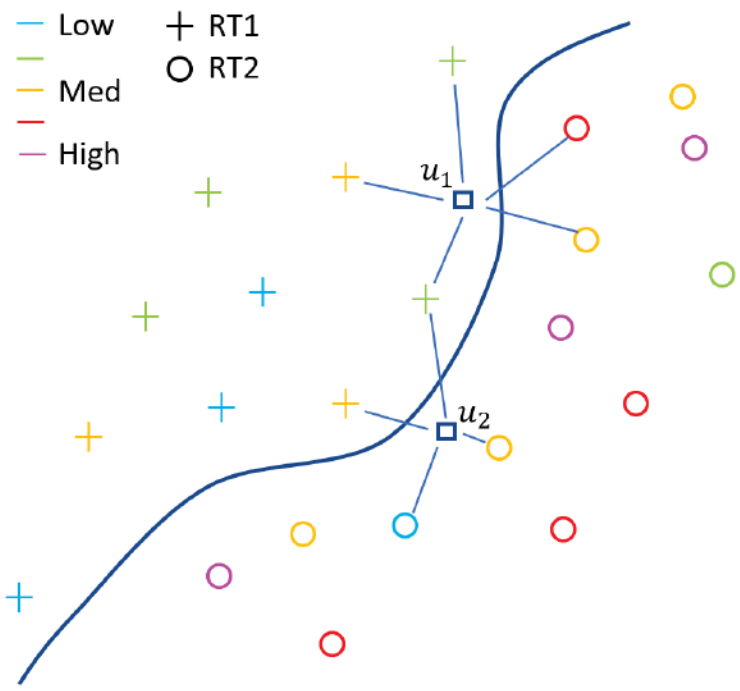


Figure 2.22: Example of mixing stationary populations along the boundary between modeling domains. Estimates at unsampled locations near the boundary consider samples from different stationary populations.

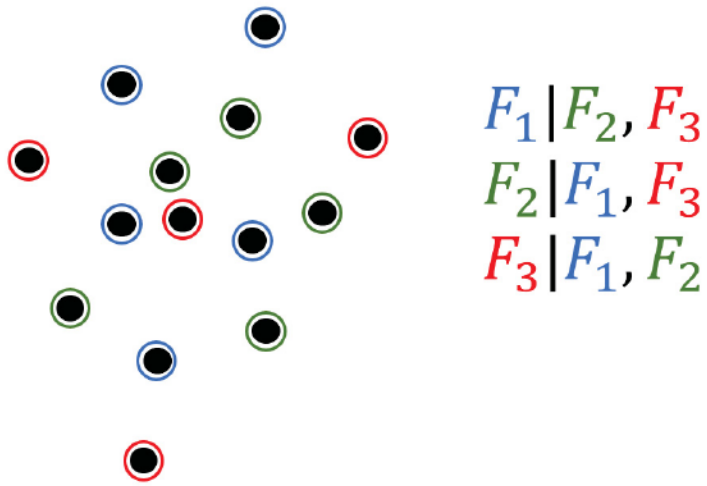


Figure 2.23: K-fold analysis with 3-folds. Validation requires 3 prediction runs where the samples from each fold comprise the test dataset, with the remainder of the data forming the training dataset.

2.7 Review of Main Points

This chapter reviews the current practice in geostatistics for generating boundary models and defining groups of samples for statistical modeling. Specifically:

1. The generation of geological boundaries from large datasets considering RBF interpolators;
2. Methodologies to generate stationary domains considering the information contained in the geological logs;
3. Clustering techniques to define groups of similar samples;
4. Ensemble clustering to reduce sensitivity of results to parameterization;
5. Validation of geostatistical predictions.

2.8 Conclusions

Existing research into geological boundary modeling does not address the need for a generalized interpolation with LVA for dense data environments. There is also a need for tools and methodologies to aid the interpretation of local structural features from the composite datasets. Current practice lacks practical tools and guidelines for the stationary domaining workflow. These two components form an important precursor to geostatistical modeling that should be given care.

Chapter 3

Geological Boundary Modeling with Uncertainty

This chapter develops a methodology to improve geological feature reproduction of implicit models. The technique mainly targets a case of medium to high density sampling relative to the geological variability, where local anisotropic properties can be inferred directly from the samples. This is common in mining, where abundant data are available to construct boundary models and advanced visualization techniques (such as the 'X-Ray Plunge'; Cowan, 2014) are required to correctly interpret and model complex local orientations of continuity in 3D. In these types of domains, implicit models that account for local continuities better capture geological relationships. The proposed methodology results in subjectively and objectively better geological models, and simplified parameter inference for complex domains.

3.1 Motivation

Geologically realistic boundaries are an important precursor to geostatistical models. Boundaries define the location, extents, volume and ultimately tonnage of a mineralized resource. Thus, modeling realistic geological boundaries while simultaneously capturing associated uncertainties is an important component of geostatistical modeling. Geological domains often contain non-linear features that are not adequately described by a single direction of continuity. Second-order non-stationary estimation frameworks can incorporate realistic curvilinear interpretations of subsurface geometries. A RBF based implicit modeling framework is developed in this chapter; using domain decomposition, the framework allows incorporation of locally varying orientations and magnitudes of anisotropy to the generated boundaries. The interpolation framework is paired with a method to automatically infer local structural orientations, which results in a rapid and iterative non-stationary boundary modeling technique that can refine locally

anisotropic geological shapes automatically from the sample data. The method is compatible with existing methodologies for quantifying the volumetric uncertainty.

3.2 Geological Boundary Modeling

Identifying, interpreting and parameterizing geological features to implicit boundary models is not trivial. This component of implicit modeling is traditionally implemented using tools that allow for the explicit control of geological features through interactive 3D-digitization (Fig. 3.1). Interpreting local features from the point dataset forms an important component of this workflow; significant uncertainty is present in this step since interpretations from different individuals are unlikely to be identical and this interpretation has a large and subjective impact on the final model (Lindsay et al., 2012, 2013). Once interpreted, incorporating local features to the boundary models is relatively well explored with second-order non-stationary interpolation techniques, including: partitioning; unfolding; coordinate transformations; or space deformation. Often a preliminary isotropic model is generated for inspection, and to guide interactive refinement of local features and geologically reasonable shapes (Fig. 3.1; Cowan et al., 2003; Leuangthong & Srivastava, 2012). Three-dimensional visualization methodologies can be used to aid identification of local structural orientations by observing grade continuities in different directions with a rotating view, selective opacity and filters (Cowan, 2014). The goal of all methodologies is to identify and digitize local structures such that the SDF interpolation algorithm is able to reproduce the interpreted features.

The idealized geological model has familiar forms and patterns, geological meaning, embedded geological knowledge, consistent geological topology and the correct geological relationships (Hillier et al., 2015). The interpretations between and away from the input data are subject to the discretion of the modeler, the purpose of the model and the algorithms used to characterize the nature of the boundaries. The criteria that characterizes the quality of a geological model is subjective since it depends on the interpretation of the geological modeler, the interpretation of the model evaluator and the purpose of the model. Although several models can be generated, the quality, and thus preference between different models, cannot be assessed quantitatively since there is an unknown true model and any metric of geological quality from

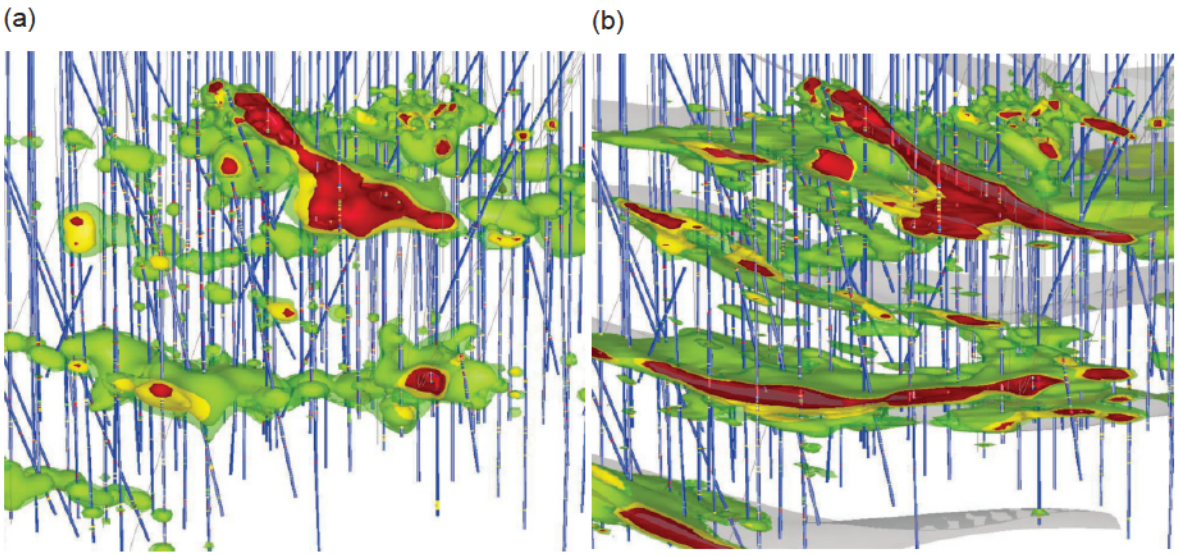


Figure 3.1: (a) Isotropic RBF interpolation showing 'blobby' radial-like shapes throughout the model, (b) the idealized version of (a) with representative local anisotropy (reproduced from: Cowan et al., 2003)

the current expert is implicitly biased. A K-fold cross validation exercise (see: Sec. 2.6) can be used to justify one algorithm or parameterization over another for implicit geological modeling, however, an increase in prediction accuracy does not necessarily ensure the target local features are reproduced.

For the remainder of this chapter, a boundary model that is geologically representative is one that matches the geological interpretation by capturing local features (e.g., Fig. 3.1b). In practice this is achieved by considering LVA, where the LVA is parameterized to represent locally variable geological conditions. Applications of LVA in implicit modeling are not widely explored in the literature, though it is noted that implicit models must be generated with care to respect geological interpretations. Recall the PU RBF interpolation framework presented in Section 2.2.5 where a recursive decomposition of the dataset results in several overlapping partitions with roughly equal numbers of data assigned to each (Fig. 3.3). The remainder of this chapter develops a methodology using the PU interpolation framework to iteratively refine geological features to geological boundary models.

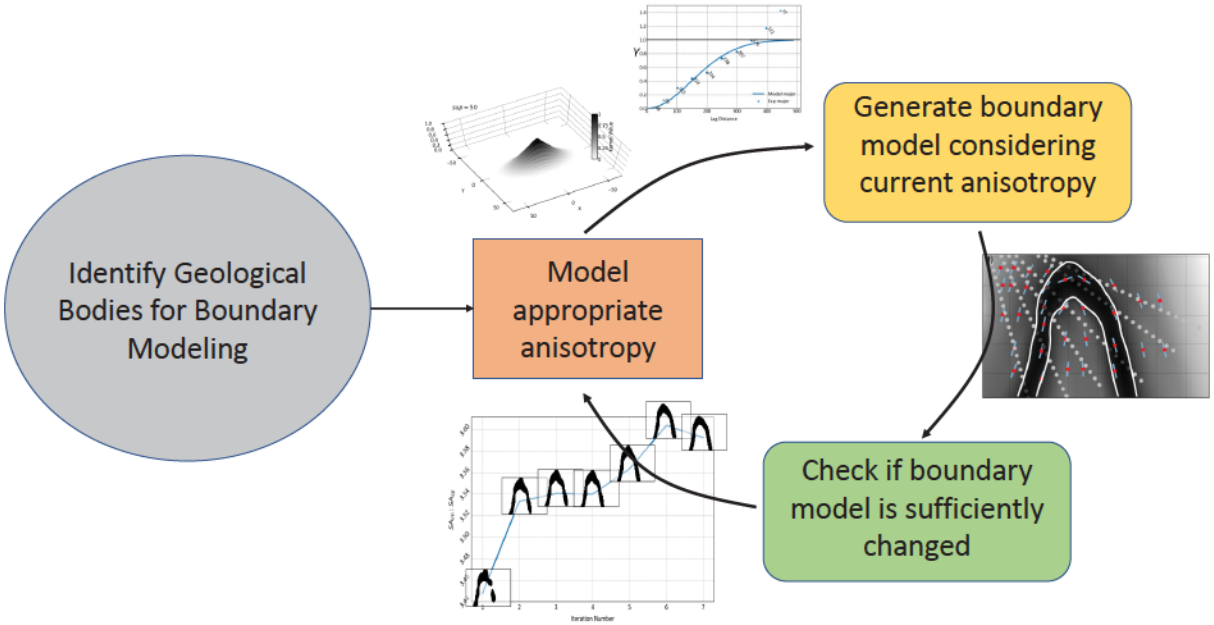


Figure 3.2: Iterative boundary refinement algorithm. Anisotropy is inferred at each stage using appropriate methodologies for the nature of the current information.

3.3 Iterative Refinement of Geological Boundary Models

The proposed algorithm is to iteratively refine geological boundary models by updating local anisotropy parameters generated from previous boundary models. The algorithm has two main components: a method to implicitly model with LVA, and a method to automatically infer local orientations from previous boundary models. The steps of the proposed algorithm are shown in Figure 3.2 where each implicit model considers the anisotropy inferred from the previously generated model.

The proposed iterative refinement algorithm is developed using a test dataset where the target geological structure cannot be adequately modeled using a global set of anisotropic parameters. Figure 3.3a shows the locations of the composite samples that are coded as either inside or outside the domain of interest. The SDF is calculated from the location of the categorical samples (Fig. 3.3b). A set of global anisotropic parameters are extracted by calculating and modeling the experimental variogram of the SDF (Fig. 3.4). An initial boundary model is generated using the PU RBF interpolation algorithm from Section 2.2.5 (Fig. 3.5a). The boundaries between inside and outside are extracted by finding the 0-level contour in the interpolated SDF.

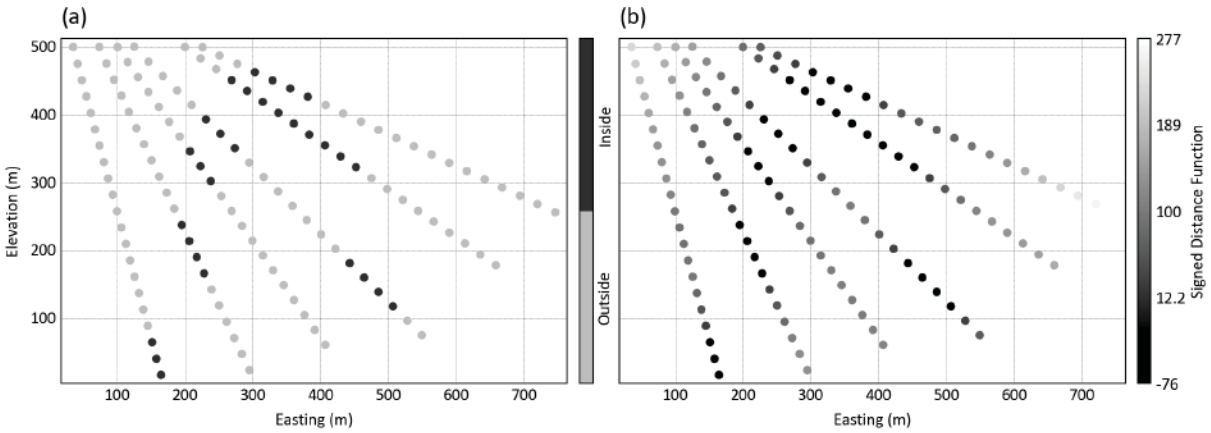


Figure 3.3: Example domain showing the locations of the (a) categorical dataset and (b) the calculated SDF at the data locations.

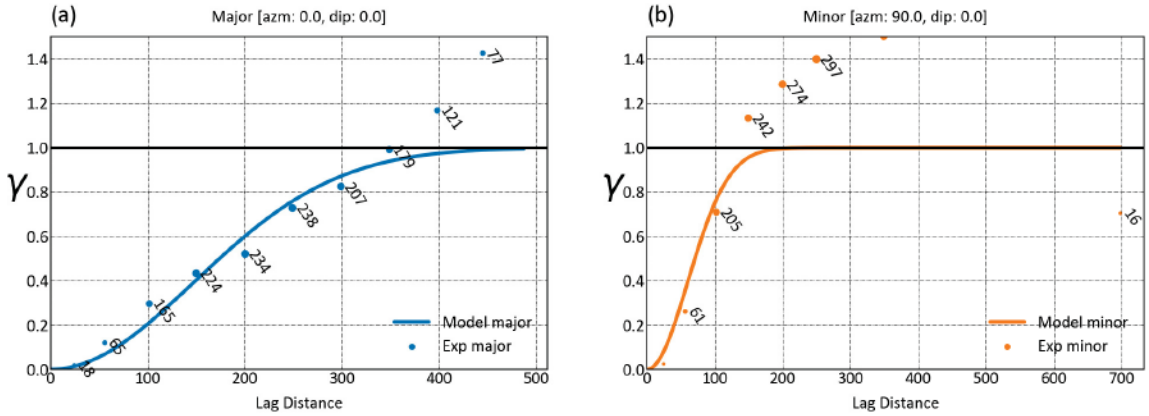


Figure 3.4: Experimental and fitted model variograms calculated for the SDF for indicator 2 in the domain shown in Figure 3.3. Experimental variograms are fitted with a Gaussian model. (a) The major direction, which in this case is vertical. (b) the minor direction, which is horizontal on this cross section.

Although these boundaries generated under an assumption of global anisotropy are reasonable for the left-limb of the folded structure (Fig. 3.5a), the right limb can be improved to better match an interpretation of down-dip continuity.

3.3.1 Local Anisotropy with RBF Interpolators

The PU generates several small overlapping subproblems that can be interpolated independently and combined to the global in a post-processing step. A locally anisotropic kernel in each partition is proposed to incorporate LVA to implicit boundary models. This is formulated

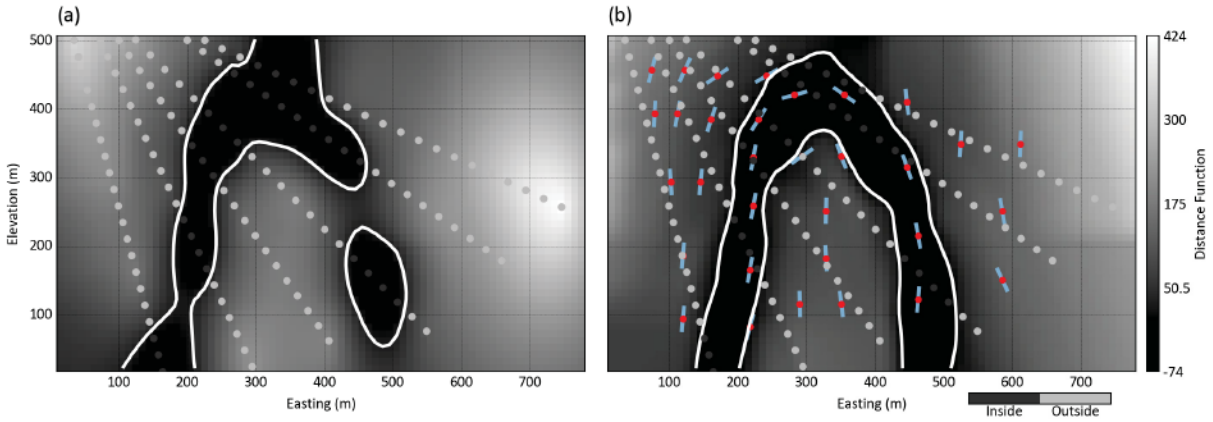


Figure 3.5: (a) Global anisotropy. (b) Interpreted local anisotropy

as a modification to Equation 2.3:

$$f_i(\mathbf{u}) = \sum_{j=1}^{P} \sum_{i=1}^{N} w_{i,j} \lambda_i \phi_j(|\mathbf{u} - \mathbf{u}_i|) \tag{3.1}$$

where for each partition j the kernel ϕ_j is unique and has properties of anisotropy and support that are representative of the local data. As in Figure 2.6c and d, the anisotropic kernel decreases the similarity of points found along non-principle orientations. The result is similar to anisotropy applied globally; disconnected regions become connected and local features stretch to become more representative of the local dataset. For example, a set of local orientations and ranges of anisotropy are generated through manual interpretation, where continuity is expected down-dip on the limbs of the folded structure. The interpreted orientations are shown as blue vectors with the partition centroid in red in Figure 3.5b. The result of PU interpolation with local anisotropy is shown as the white boundary in Figure 3.5b. This boundary has the expected down-dip continuity and improves the geological feature reproduction when compared to the globally anisotropic model.

3.3.2 Inference of LVA Parameters

Inference of LVA parameters is well studied (Fouedjio, 2016b; Lillah & Boisvert, 2015; Machuca-Mory & Deutsch, 2013). Lillah and Boisvert (2015) develop several methodologies to infer the local orientations from both point and gridded data sources. When a point dataset provides insufficient information, Lillah and Boisvert (2015) recommend to generate a neutral gridded

model with global ordinary kriging and a high-nugget isotropic variogram. The goal is to assign values of the sparse point data to the modeling grid to ensure exhaustive information is available for orientation extraction algorithms. Orientations are extracted from gridded datasets by analyzing covariance or gradients calculated in moving windows (Feng & Milanfar, 2002; Hassanpour & Deutsch, 2007; Lillah & Boisvert, 2015).

Feng and Milanfar (2002) develop a multi-level algorithm where block averaging and between-layer weighting extracts orientations from noisy input datasets. Those authors use singular value decomposition (SVD) of an N-long gradient matrix constructed from the gradient values calculated in local windows:

$$G = \begin{bmatrix} \nabla f(1) \\ \nabla f(2) \\ \dots \\ \nabla f(N) \end{bmatrix}, \quad G = U \cdot sv \cdot V^T \tag{3.2}$$

Where G is the gradient matrix for N cells found in each window, sv are the singular values and V is the rotation matrix obtained from SVD. The implementation from Feng and Milanfar (2002) is developed for 2D applications applied to images; the algorithm is modified here for 3D by including calculations in the third coordinate dimension. Thus for 3D domains the G matrix is $N \times 3$, and the SVD of G produces a 3×3 orthogonal rotation matrix V, with the first vector indicating in the direction of the largest change in gradients for each local window. The multi-layer algorithm takes an input gridded dataset and block averages the gridded data to a coarser resolution. Beginning at the coarsest top-most layer, estimated orientations are propagated by weighting the components of the estimated orientation vectors from the current layer with those estimated from the layer above:

$$[x_c, y_c, z_c] = \alpha[x_c, y_c, z_c] + \beta[x_{up}, y_{up}, z_{up}] \tag{3.3}$$

where the c and up indicate the current and upper layer, respectively. Weights for each orientation estimate are calculated as:

$$\alpha = \frac{er_c}{er_c + er_{up}}, \quad \beta = \frac{er_{up}}{er_c + er_{up}} \tag{3.4}$$

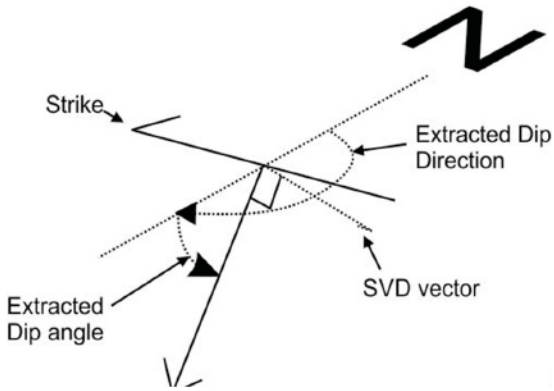


Figure 3.6: Example of the orientation extracted from the gradient-SVD algorithm implemented in 3D

where er are the energies calculated for each layer as:

$$er = \frac{sv(1) - sv(3)}{sv(1) + sv(3)} \tag{3.5}$$

When an orientation is dominant within a local window, the magnitude of the singular values from gradient decomposition will differ significantly, resulting in a high energy calculated for that local window, and thus more weight given to that orientation estimate when considering finer layers. Conversely if the orientation is not dominant, the singular values are more equal in magnitude resulting in a lower energy for that orientation estimate.

The orientation extracted from the gradient algorithm points in the direction of greatest change, and in 2D the conversion from this to the direction of greatest continuity is trivial. However, extra consideration is required to extract orientations in 3D. Figure 3.6 shows the different directions extracted from the 3D-gradient algorithm. The SDF represents an ideal target for this orientation extraction algorithm since the SDF is hyper-continuous as iso-value surfaces parallel with the target bounding surfaces, and the vector extracted from the gradient algorithm is surface-normal to these boundaries (Fig. 3.6). The dip-direction and dip (GSLIB $ang1$, $ang2$) are extracted as:

$$\begin{aligned} dipdir &= \frac{\pi}{2} - arctan2(y, x) \\ dip &= \frac{\pi}{2} - arcsin(z) \end{aligned} \tag{3.6}$$

where x, y and z are the components of the normal vector extracted at each location from the gradient-SVD algorithm, and the extracted $ang1$ and $ang2$ are modified to ensure that the vector is always down-dipping.

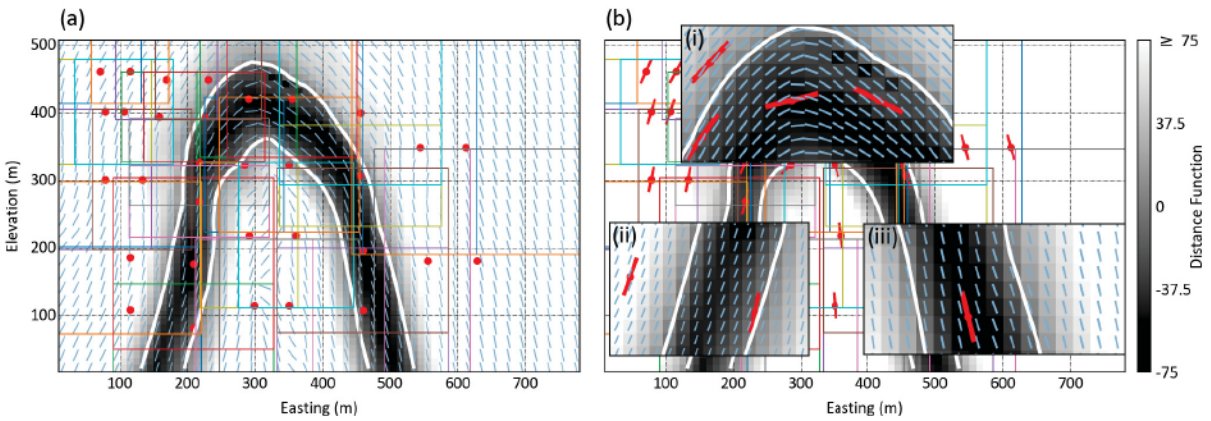


Figure 3.7: Extraction of the partition-representative orientations. (a) Shows the partition centers (red points), boundaries (colored boxes), and the exhaustive local anisotropy extracted using the Feng and Milanfar (2002) algorithm. (b) Shows 3 inset maps where the partition-representative orientation (red) is extracted from the exhaustive orientations (blue).

Figure 3.7 shows an SDF model of the sample domain where the color of the interpolated SDF is binned to between -75 and +75, highlighting the location of the boundary and the locations of important gradients in the SDF (e.g. near the boundary). The set of partition boundaries and centers are shown with the colored boxes and red points, respectively (Fig. 3.7a). The set of local orientations extracted using the gradient-pyramid technique developed above are shown with the blue vectors (Fig. 3.7a). This multilevel gradient-SVD analysis defines a set of anisotropic parameters (strike, dip, plunge, $r_1 = \frac{ahmin}{ahmax}$, $r_1 = \frac{avert}{ahmax}$) for all locations in the domain (Fig. 3.7a). For each partition from the PU decomposition, the problem is now to optimally combine the anisotropic parameters from the partition to represent the most important orientation for that partition.

3.3.3 Extracting a Representative Partition Anisotropy

One useful value generated from the gradient decomposition is the relative energy of each orientation in the gradient window. As shown in Equation 3.5, the energy is derived from the magnitude of the singular values, and gives a measure of the dominance of the orientation estimated within each local window. Feng and Milanfar (2002) use these energies to propagate orientation estimates between layers to ensure that reasonable orientations are extracted in the presence of noise. Here, the goal is to obtain an orientation estimate for each partition considering the set of orientations encompassed within each partition. Using the energy associated

with each orientation, the partition orientation is estimated by weighting orientations estimated in each cell by the associated energy, and ensuring the weights sum to 1:

$$\begin{gathered}\left[x_p, y_p, z_p\right] = \sum_{i=1}^{n} wt_i \left[x_c, y_c, z_c\right]_i \\ \sum_{i=1}^{n} wt_i = 1\end{gathered} \tag{3.7}$$

for all n cells contained in each partition. Since the estimated orientations do not contain a tilt-component (e.g., $ang3 = 0$) the orientation averaging is greatly simplified as the weighted linear combination of the vector components with consideration to the axial nature is suitable (as in: Machuca-Mory, Rees, & Leuangthong, 2015). The resulting partition-orientation represents the most dominant orientation from the estimated energies. Sets of partition orientations estimated from the exhaustive orientations are shown in Figure 3.7b. The inset maps show the representative partition orientation(s) extracted from the exhaustive orientation field. In cases where partitions encompass many predominant sets of orientations, this averaging could result partition orientations that do not reflect the data. In this case, modifying the partitioning overlap in the x, y or z directions to encompass less of the orientation-variability or re-partitioning with fewer data-per-partition is recommended.

3.3.4 Automatic Iterative Refinements

The combination of partition orientation extraction and the PU RBF interpolation algorithm comprises the proposed iterative boundary modeling algorithm. The steps of the iterative algorithm are shown in Figure 3.2: model (local) anisotropy; and interpolate with (local) anisotropy. The first iteration is special in the sense that anisotropy is global and characterized from the composite dataset rather than a previously interpolated SDF model. The first SDF model is generated under stationary conditions using anisotropic parameters estimated from the composite dataset. The iterative algorithm begins at this step by analyzing the local anisotropy using the gradient-SVD algorithm and the energy-weighting to estimate the representative local anisotropy for each partition. The next interpolation step considers this refined local anisotropy, and the process repeats by analyzing local anisotropy and regenerating the boundary model with refined local

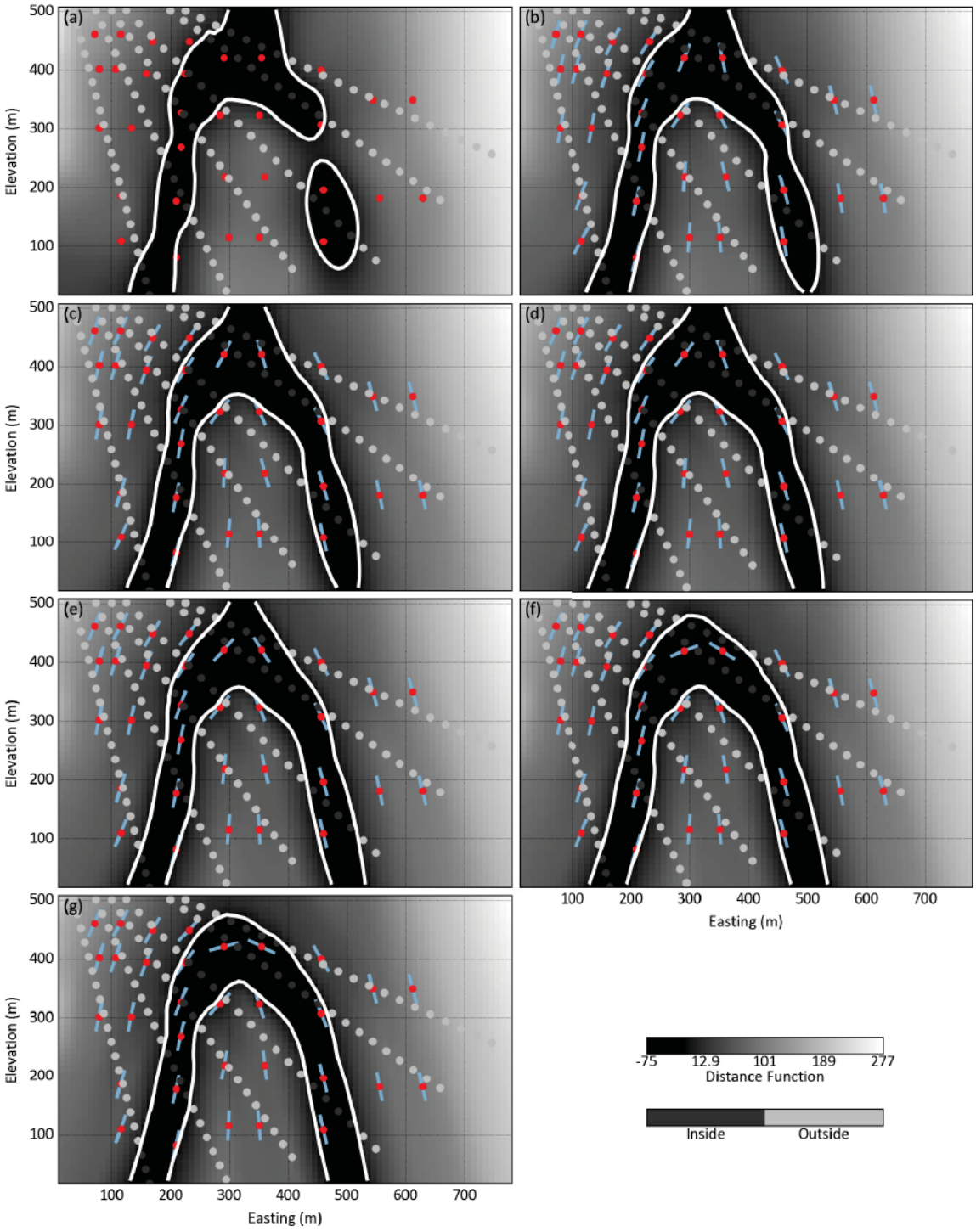


Figure 3.8: Fully automatic iterative refinements of the synthetic test domain

parameters (Fig. 3.2). The individual components of this algorithm are flexible. Any method to estimate local anisotropy could be considered. Similarly, the SDF interpolator is not required to be the PU RBF interpolator; this is simply used here since there are other benefits to considering this formulation for iterative geological modeling.

The final component of the iterative refinements is to choose a stopping criteria. If refinement of a partition has generated no beneficial changes, then further refinement of that partition may be skipped without adversely affecting the boundary model for nodes contained within that partition. Some potential stopping criteria might include: boundary smoothness; volumetric properties; or error on overlapping sites. Here, the implemented metric is to stop refining partitions when there is an insufficient amount of changed cells between iterations. For example, if a cell in a partition changes category between iterations, that cell is considered changed. If fewer than some specified number of cells, say 1%, change between iterations, then this partition could be considered 'refined', and be frozen for all following iterations. In this case, the un-weighted local solution can be stored so that the global interpolant can be quickly recovered with Equation 2.7.

The iterative algorithm is applied to the synthetic test domain and the resulting fully automatic refinements are shown in Figure 3.8. Even a single iteration represents a modest visual improvement over the global anisotropy model (Fig. 3.8b vs a). Further iterations generate improved continuity on the dipping limbs which better matches the interpretation of the folded geological body. By the final iteration the boundary model smoothly reflects the local orientations that were manually inferred in Figure 3.5b.

3.3.5 Parameterization and Sensitivity

Parameters required for this iterative algorithm include the partitioning parameters, the choice of local orientation algorithm, the conditions of the initial boundary interpolation and some form of stopping criteria. The conditions of the initial boundary interpolation are partly subjective. Past experience has shown that representative global anisotropy included in the first boundary model can aid in the iterative refinements to better match the preliminary interpretation. Choosing the local orientation estimation algorithm is also subjective. te Stroet and Snepvangers (2005) use

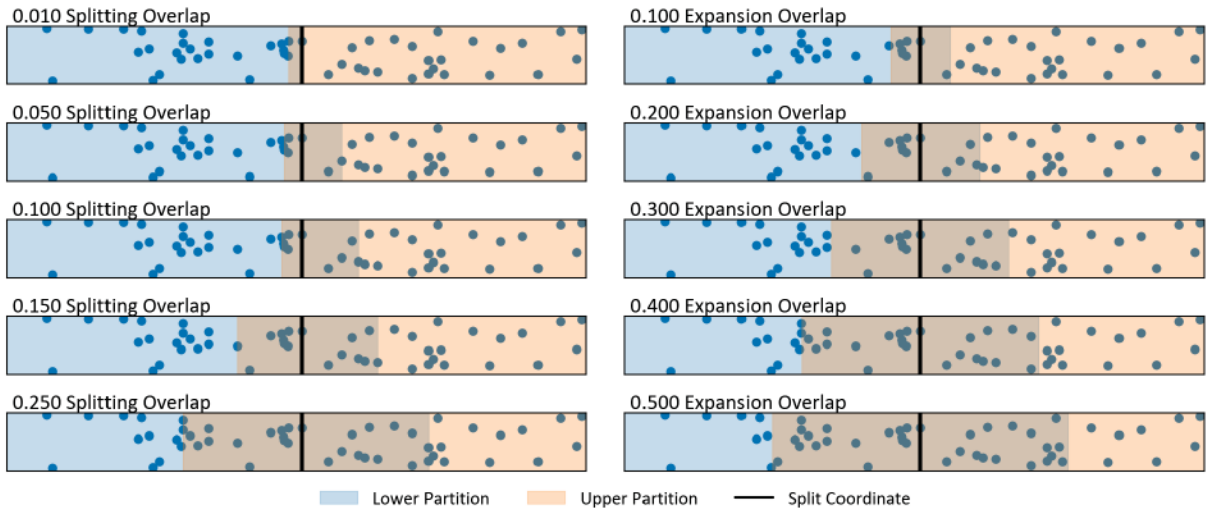


Figure 3.9: Splitting (left) and expansion (right) overlap applied to a 1D domain. Note the asymmetry in overlap between partitions depending on the data configuration in the splitting case.

an iteratively decreasing window size in their local orientation analysis to resolve small-scale details at later stages of their algorithm. A similar strategy could be considered here, though it is worth noting that the PU RBF interpolator used in this work is only able to resolve local features at the scale of partitioning. Thus in sparsely sampled environments, LVA features will be relatively large-scale when compared to densely sampled environments, which naturally follows the support of local information. The most consequential consideration for the PU RBF interpolator is the choice of partitioning parameters, which in turn affect the local anisotropy and iterative refinements. The binary partitioning used in this work requires 3 parameters: the data-per-partition; splitting overlap; and expansion overlap (Sec. 2.2.5).

The data per partition roughly controls the size of each partition. Partition size is inversely proportional to the density of local sampling. Fewer data per partition is more computationally efficient since the local linear system of equations has a smaller N. Fewer data per partition also allows more variable local anisotropy to be considered, since there are more partitions available to specify unique local parameters. However, an anecdotal lower bound, based on several test datasets, appears to be around 6-10 samples for 2D domains and around 15-40 samples in 3D domains, depending on the total number of data and the data configuration.

The different overlap parameters for binary partitioning are discussed in Section 2.2.5 and shown schematically for a 1D case of domain partitioning in Figure 3.9. Splitting overlap is

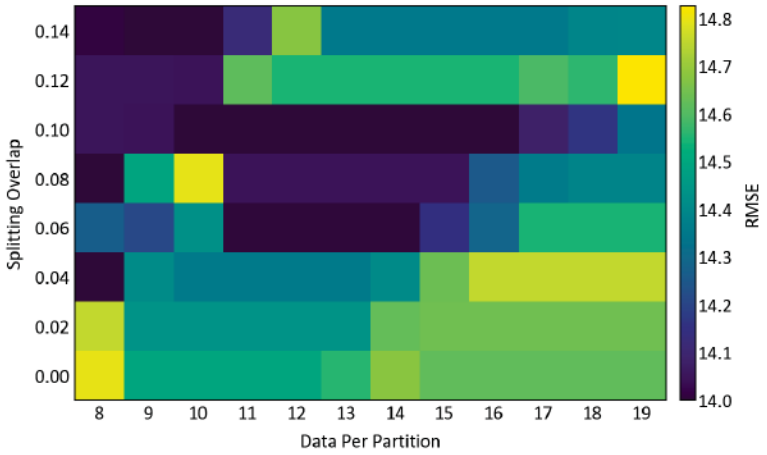
specified as a fraction of the data contained in the parent partition. A large splitting overlap may result in recursion issues for specific data configurations since a split may generate two partitions largely containing the same dataset as the parent. However, specifying a small overlap here may cause artifacts in the final model. Additionally, depending on the data configuration, the overlap on either side of the split may be asymmetric (Fig. 3.9, left). Past experience has shown that taking between $\frac{1}{100}^{th}$ and $\frac{1}{7}^{th}$ of the data found in a partition as common to both child partitions is reasonable. A larger overlap at this stage can potentially generate many redundant partitions with a large amount of overlap with one another. The second overlap parameter expands partition boundaries symmetrically once the partitions are generated (Fig. 3.9). Care must be taken with this parameter since the goal is to ensure a suitable amount of overlap between partitions, simultaneously reducing the number of data contained within each to ensure the efficient solution of the linear system of equations. The recommended method for choosing the required parameters for a given domain is to use a K-fold validation and to choose the minimum possible data per partition and smallest overlap parameters that give a reasonably small and stable root mean squared error (RMSE).

For the synthetic domain the partitioning parameters are determined using a 5-fold validation by permuting the required input parameters and observing the prediction errors associated with each set. Figure 3.10a shows the RMSE obtained by varying the data per partition and the splitting overlap. For a lower data per partition, the interpolation appears to be fairly variable with different splitting overlap. With > 11 data per partition, the error results seem more stable. Thus, a data per partition of 11 and splitting overlap of 0.06 are chosen. Figure 3.10b shows the error associated with different x and y expansion overlap following the partitioning with the chosen data per center and splitting overlap. Here, a final overlap of 0.6 and 0.5 are chosen for the x and y directions, respectively.

3.4 Assessing Geological Boundary Models

Assessment of a geological boundary models at this scale is based on subjective visual inspection and is practitioner dependent. Objective measures such as a K-fold validation and other error metrics are useful to compare models generated with different algorithms or different pa-

(a)



(b)

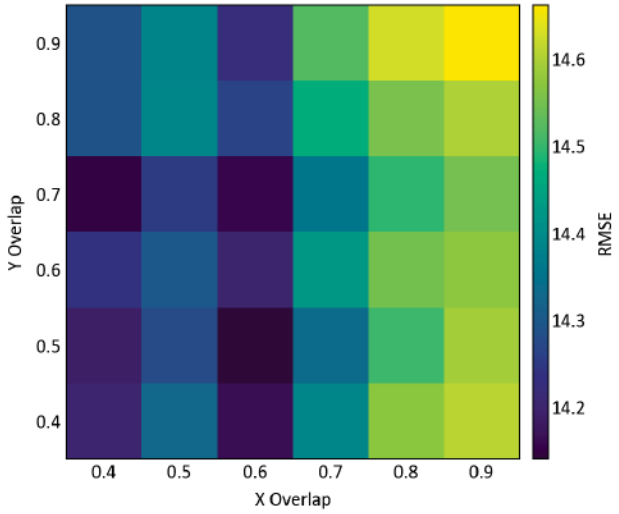


Figure 3.10: K-fold validation RMSE the PU RBF interpolation for different partition parameters. (a) the data per partition and data overlap. (b) the final overlap applied after partitioning.

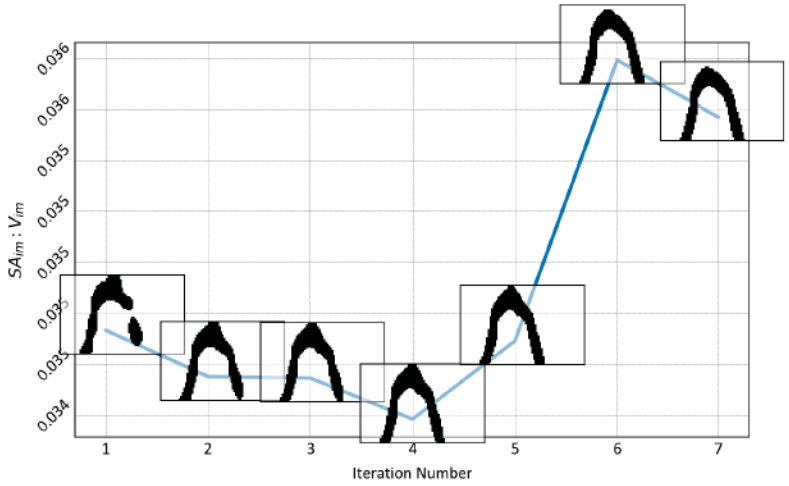
rameterizations of the same algorithm. Generally, a 'good' implicit model is one that accounts for the data, conforms to the conceptual geological model, accounts for uncertainty (perhaps both volumetric and geometric), and has idealized shape properties like smoothness and representative local anisotropy. A 'good' implicit model also lacks features present in subjectively poor implicit models, including: disconnected bodies that could otherwise be connected given the local structural interpretation (Deacon, 2017); and blob- or bubble-like local features that are commonly observed in poorly parameterized implicit models (Fig. 3.1a vs. Fig. 3.1b). The 'blobby' nature of poorly constructed implicit models should be avoided, unless the geological topology warrants such features (Fig. 3.1a). However, these blob-like features are more com-

monly the result of interpolation parameterizations that are not representative of the local data.

The subjectively poor features of an implicit model have distinct shape properties. A sphere, for example, minimizes the surface area to volume ratio (SA:V) for a given volume. By measuring the SA:V of implicit models, models with more blobs or other features approximating spheres (e.g., Fig. 3.1a) will have a lower SA:V metric than models without these features. As demonstrated in this chapter, a reasonable goal for implicit geological modeling is to generate local geometric anisotropies that reflect the underlying geological shapes. Similar shape property-metrics of geological models are proposed by Lindsay et al. (2013). Their metrics (of 'geodiversity') quantify: the top and bottom elevation of each unit; the volume; the surface area of contacts; the surface curvature of contacts; and the geological complexity of a suite of geological models (Lindsay et al., 2013). The goal of their work is to assess the uncertainty in the geological models, identify outlier models that are significantly different from the ensemble, and to come up with models that are most representative of the suite.

The SA:V metrics proposed here, calculated for each iteration of the iterative refinements of the synthetic domain, are shown in Figure 3.11a. One drawback of the SA:V metric is that direct comparison between different models assumes that the volume between models is constant. If this condition is satisfied, this metric essentially quantifies differences in the surface area between two models. However for the synthetic domain presented here, this constraint is not satisfied between iterations; for example, between iteration 1 and 2 the volume of the model increases since the disconnected regions become connected. The result is a higher SA:V metric for a subjectively poorer model in iteration 1 when compared to the model generated in iteration 2. To address this limitation, the ratio of the implicit model surface area to the surface area of a sphere with the same volume is proposed (Fig. 3.11b). In this case, the increase in volume between iterations does not appear to affect the metric as in the case of SA:V and there is a consistent increase in the proposed metric with model development. With K-fold validation and these shape properties, sets of implicit models (or iterations of refined models) can be subjectively and objectively assessed against one another for reproduction of the data, reproduction of the geological interpretation, and a comparison of the shape properties of the model.

(a) Surface area to volume ratio



(b) Implicit surface area to sphere surface area ratio

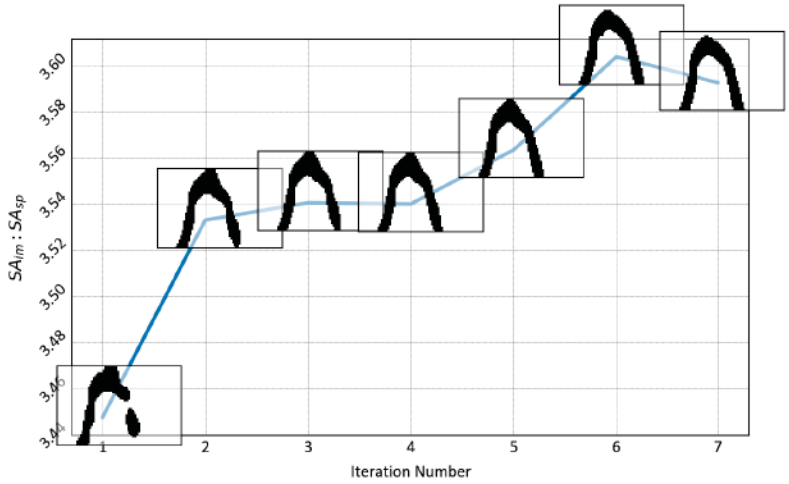


Figure 3.11: (a) Surface to volume ratio for iteratively refined implicit models where the different in model volumes between iterations makes interpretation of the metric difficult. (b) Ratio of implicit surface area to sphere of same volume surface area. Normalizing the surface area of the implicit model by the surface area of a sphere with the same volume as the implicit model seems to generate less dependence on changes in volume between different models.

3.5 Uncertainty Through the C-Parameter Framework

Uncertainty in the location of boundaries is an important consideration for implicit models since these boundaries inform whether or not a location is assigned a value with geostatistical methods. The locally anisotropic RBF interpolator developed in this chapter is compatible with the bandwidth-style uncertainty (Munroe & Deutsch, 2008; Wilde & Deutsch, 2012). Recall that the bandwidth modifies the underlying volumetric function such that two locations in different domains are further away from one another. This bandwidth parameter, C, is a global modifier and must be calibrated for different domains and interpolation conditions. Calibration of C requires a K-fold validation workflow where the prediction errors for different bandwidth parameters are analyzed with a calibration curve, and a representative bandwidth of uncertainty is chosen. The chosen uncertainty bandwidth is deposit and data configuration specific. The C-parameter should be less than the 'effective' data spacing for a particular deposit, and in the absence of the K-fold analysis the data spacing can be used as the C-parameter.

3.6 Review of Main Points

This chapter introduces a practical geological boundary modeling algorithm exploiting domain decomposition to improve geological models where local structural continuity is important. Specifically:

1. A practical geological boundary modeling methodology using PU domain decomposition and independent locally anisotropic parameters;
2. A method for automatically and iteratively inferring local anisotropy from generated boundary models to improve the geological feature reproduction.

3.7 Conclusions

The proposed iterative boundary modeling algorithm is suitable for capturing non-stationary features present at a relatively large-scale. Manual inference of local features at this scale is possible, but is subject to interpretation biases and requires significant time and effort on the

part of the geomodeler. The proposed iterative refinement algorithm provides a better first-pass locally anisotropic model for the expert to consider during geomodeling.

Chapter 4

Stationary Decision Making with Clustering

Cluster analysis has many similarities with a stationary domaining workflow. The goal of clustering is to find natural partitions of a dataset where samples within a partition are more related to each other than they are to samples in other partitions, and 'related' is described in terms of the multivariate properties of the dataset. Finding stationary domains from a geological dataset can be formulated in a similar manner except that the notion of similarity includes spatial, multivariate and geological relationships. However, identifying stationary domains in practice is exceedingly complex and novel techniques reducing subjectivity and biases are needed.

4.1 Introduction

This chapter develops three contributions to improve stationary decision making in geostatistical domains. First, a combined spatial-multivariate metric is proposed to permit objective comparisons between different cluster configurations. Second, a novel random-path spatial clustering algorithm is developed that alleviates parameterization concerns in a spatial-clustering workflow, which in turn allows for straight forward tuning of the spatial contiguity or multivariate delineation of the generated spatial clusters. Finally, an ensemble-clustering based framework is developed to assemble a prior distribution of uncertainty for the decision of stationarity. This prior distribution could represent the parameter uncertainty for the decision of stationarity which can be incorporated in a traditional geostatistical uncertainty characterization workflow.

Note, in the following text the term 'clustering' refers to a set of $K \in \{1, \ldots, K\}$ categories assigned to each sample location following some method that partitions the dataset (e.g., manual or automatic). The following terms are synonymous with 'clustering': stationary domains, partitions, groups, sets, classes, categories, facies, and rock types. All terminology refers to

a set of mutually exclusive categories, each reflecting one category, domain, partition, group, etc., within the K categories, domains, partitions, groups, etc. The only divergence from this notation is with a 'clustering', which comprises K individual clusters.

4.2 Metrics of Stationary Domains

One goal of stationary domaining is to improve the estimate of value at the unsampled location. The estimate is improved in two ways: 1) nearby samples from the same stationary domain are increasingly related, and 2) stationary domains themselves are increasingly contiguous throughout the spatial domain. The estimate at the unsampled location involves many steps including parameter inference, categorical modeling, and choosing a suitable algorithm to infer the distribution of uncertainty conditional to the related nearby data. Thus, validating a set of stationary domains should include a measure of how well the true value has been predicted at an unsampled location. In other words, one reasonable goal of stationary domaining is to minimize prediction errors in a K-fold cross validation study. However, evaluating all N^K possible combinations is impractical not only because of the large combinatorial problem, but also since each combination requires a full geostatistical estimation or simulation workflow to assess the effects.

Consider the porphyry domain show in Figure 4.1 (from: Boisvert, 2010). Figure 4.1a shows two populations that nearly completely overlap with respect to their Cu-Au (hereafter, multivariate) properties, which results in first and second order statistical properties that are similar between domains. By contrast, the spatial configuration in Figure 4.1b shows clearly defined spatial partitions of this domain; presumably the central feature is the porphyry intrusion (RT1) that has intruded the surrounding country rock (RT2). Given the poor delineation of multivariate classes, geostatistical modeling of this dataset should consider additional partitioning of the multivariate space and subsequent boundary analysis in each domain to ensure the assumption of stationarity is reasonable. The high correlation between Cu and Au here suggests a typical 2-variable grade-domain is appropriate. Alternatively, the entire domain could be considered as a single population since the statistical properties of each delineated domain are similar.

A partitioning based on the multivariate properties for this domain could resemble the con-

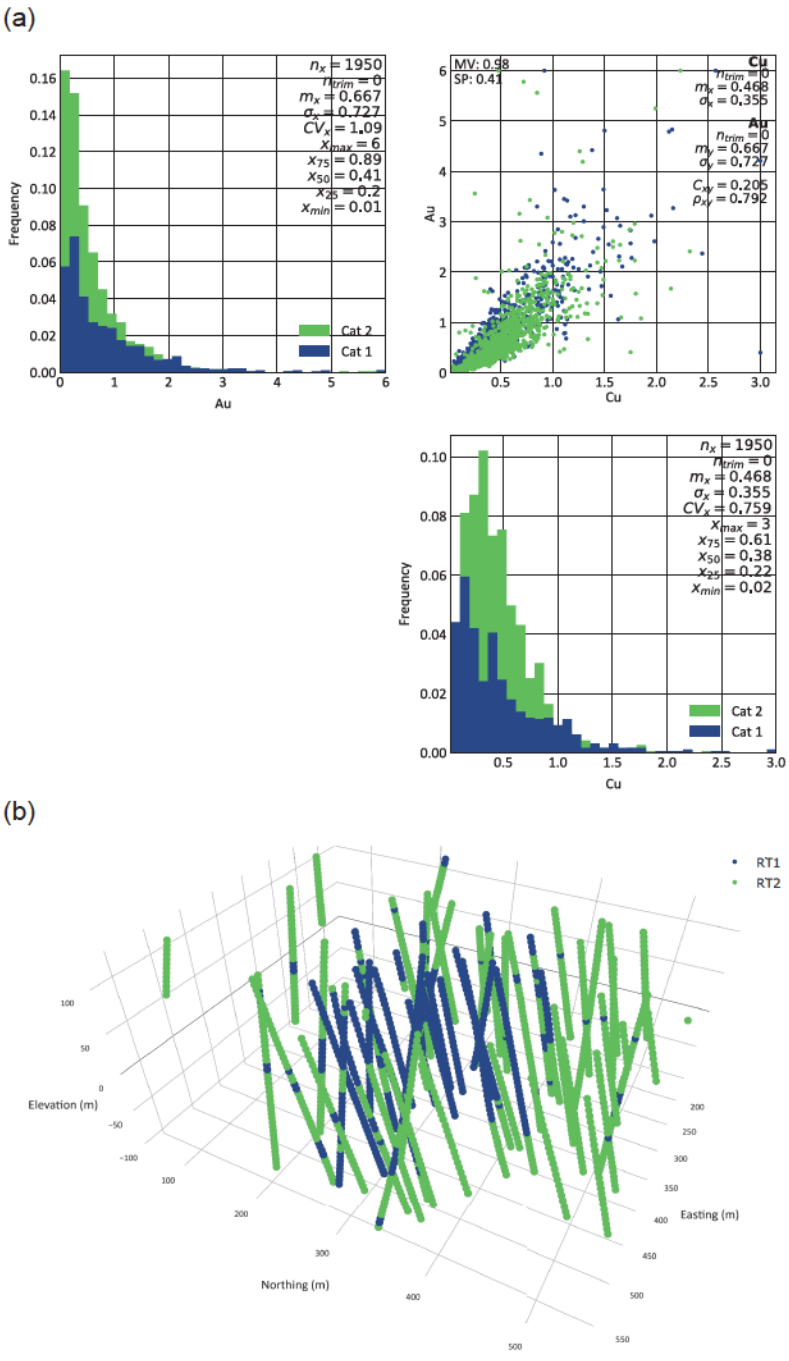


Figure 4.1: (a) Bivariate relationship between Au and Cu colored by rock types defined in the geological log. (b) Oblique projection of the porphyry dataset colored by Rock Types.

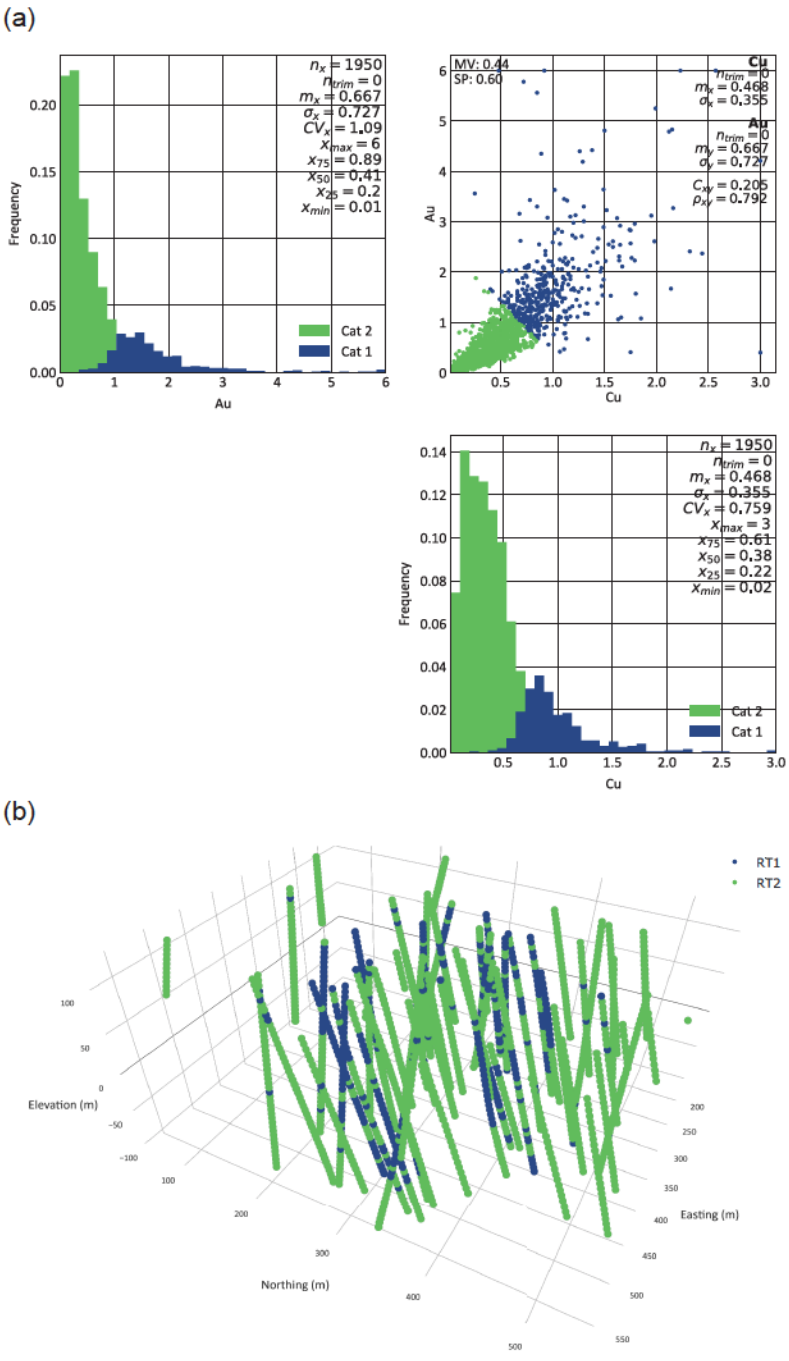


Figure 4.2: (a) Bivariate relationships between Au and Cu colored by multivariate clusters. (b) Oblique projection of the porphyry dataset colored by clusters.

figuration in Figure 4.2. Here the multivariate space is partitioned nearly perfectly between the high and low grade values considering both variables, with minimal mixing of the two populations in the marginal distributions (Fig. 4.2a). However there are less clearly defined spatial classes (Fig. 4.2b). The lack of spatial continuity of the defined classes may adversely affect geostatistical modeling since there may be difficulties obtaining a spatial model of continuity, which will in turn impact how value is assigned to the unsampled location. The relationship between a partitioning of the dataset and prediction error must be established.

Considering the above domain, two criteria can be considered to rank the two configurations for relative 'goodness'. The spatial contiguity of the classes and the multivariate delineation are both important factors for the resulting modeling domains. If the delineated domains are identical in their multivariate and spatial properties, why bother modeling them separately? Similarly, if the delineated domains result in random spatial features, geostatistical modeling will suffer from a lack of continuity in the spatial model of the categorical variables. In practice, both the multivariate and spatial properties should be considered to ensure that a set of domains are suitable for geostatistical modeling. In the following sections, a multivariate-spatial metric is proposed to aid in the interpretation of the quality of a set of modeling domains. The metrics measure the multivariate delineation and spatial contiguity of a set of categories to provide a relative measure of the properties of different stationary delineations made for the same dataset.

4.2.1 Multivariate Metrics

The multivariate metric describes uniqueness of classes with respect to their multivariate properties. Uniqueness in this space is synonymous with compactness and describes the relative delineation of each category given the multivariate properties. For example, categories defined in Figure 4.1a represent a poor delineation of the multivariate space. A measure of the multivariate delineation requires assumptions on the forms of each population. For example, for a set of categories generated using the K-means clustering algorithm, like those found in Figure 4.2a, a suitable measure of separation is the within cluster sum of squares (WCSS), since this measure

is minimized during K-means clustering (Witten & Tibshirani, 2010):

$$WCSS = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in K_k} \sum_{j=1}^{M} (x_{ij} - \bar{x}_{kj})^2 \tag{4.1}$$

For a simple synthetic dataset with $M = 2$ and $K = 3$, a suite of different clusterings are shown in Figure 4.4a. The clustering with the most separation between clusters results in the lowest WCSS. Increased mixing of the populations in this 2-dimensional space leads to higher WCSS. The WCSS is best suited to clusters of equal size and a Mahalanobis-distance-based WCSS could be considered for populations with linearly correlated variables.

A number of other metrics can be considered to measure the relative delineation of multivariate space. In the following discussion consider the populations A and B to each be one of K different clusters of a given dataset, $k_A, k_B \in \{1, \ldots, K\}, k_A \neq k_B$ and $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $B = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ contain n and m samples, respectively. Ward's distance computes the increase in the sum of squared error (SSE) obtained by merging populations A and B (Ward, 1963):

$$D_{ward} = SSE_{AB} - (SSE_A + SSE_B) \tag{4.2}$$

where:

$$SSE_{AB} = \sum_{i=1}^{n+m} |\mathbf{xy}_i - \overline{\mathbf{xy}}|, \quad SSE_A = \sum_{i=1}^{n} |\mathbf{x}_i - \bar{\mathbf{x}}|, \quad SSE_B = \sum_{i=1}^{m} |\mathbf{y}_i - \bar{\mathbf{y}}| \tag{4.3}$$

Alternatively, the energy distance D_e developed by Rizzo and Székely (2016) computes a centered measure of dissimilarity about each population:

$$D_e = \frac{2a - b - c}{2a} \tag{4.4}$$

where:

$$a = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} |\mathbf{x}_i - \mathbf{y}_j|, \quad b = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |\mathbf{x}_i - \mathbf{x}_j|, \quad c = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} |\mathbf{y}_i - \mathbf{y}_j| \tag{4.5}$$

This measure is bounded $[0, 1]$ and equal to zero only if the populations X and Y are identically distributed (Rizzo & Székely, 2016). Both the Ward and energy distances effectively capture differences in the mean of each population, but fail to capture differences in the (co)variance

between populations (Fig. 4.3 a to b vs a to c). The implication is that populations that differ with respect to (co)variance can not be differentiated with these measures. The energy distance is modified here to account for linear correlation using the Mahalanobis distance; the modified Mahalanobis energy distance becomes (Fig. 4.3):

$$D_{em} = \frac{a - b - c}{a} \tag{4.6}$$

where:

$$\begin{aligned} a &= \frac{1}{nm}\left[(\mathbf{x} - \boldsymbol{\mu}_y)\boldsymbol{\Sigma}_y^{-1}(\mathbf{x} - \boldsymbol{\mu}_y) + (\mathbf{y} - \boldsymbol{\mu}_x)\boldsymbol{\Sigma}_x^{-1}(\mathbf{y} - \boldsymbol{\mu}_x)\right], \\ b &= \frac{1}{n^2}\left[(\mathbf{x} - \boldsymbol{\mu}_x)\boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)\right], \\ c &= \frac{1}{m^2}\left[(\mathbf{y} - \boldsymbol{\mu}_y)\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)\right] \end{aligned} \tag{4.7}$$

The Kullback-Leibler divergence (KLD) could also be considered. A non-parametric version is required as the multivariate populations are unlikely to be multiGaussian (e.g., Hershey & Olsen, 2007). Figure 4.3 also shows the scores from a kernel density estimate (KDE)-based KLD calculated between the different populations; the metric shows sensitivity to both mean and covariance differences between populations, but additionally provides sensitivity to non-linear features not captured by linear methods.

4.2.2 Spatial Metrics

A measure of the spatial interconnectedness of the clusters in Cartesian space is required. The WCSS and other multivariate measures are not suitable for Cartesian space since the center of a cluster may not be relevant to how it is distributed and connected throughout the spatial domain. Instead, the entropy calculated in a local window at each location and summed over all locations is proposed. In practice this gives a relative measure of the amount of randomness found in Cartesian space. The total windowed entropy is calculated as:

$$H_{total} = -\sum_{i=1}^{N}\sum_{k=1}^{K} p_k(\mathbf{u}_i) \ln p_k(\mathbf{u}_i) \tag{4.8}$$

where $p_k(\mathbf{u}_i)$ is the probability of finding category k in the local search around the i^{th} location. This is shown graphically in Figure 4.4b; the local entropy is low where a single category is found in a local search and high where several categories are found (e.g., left vs. right circle,

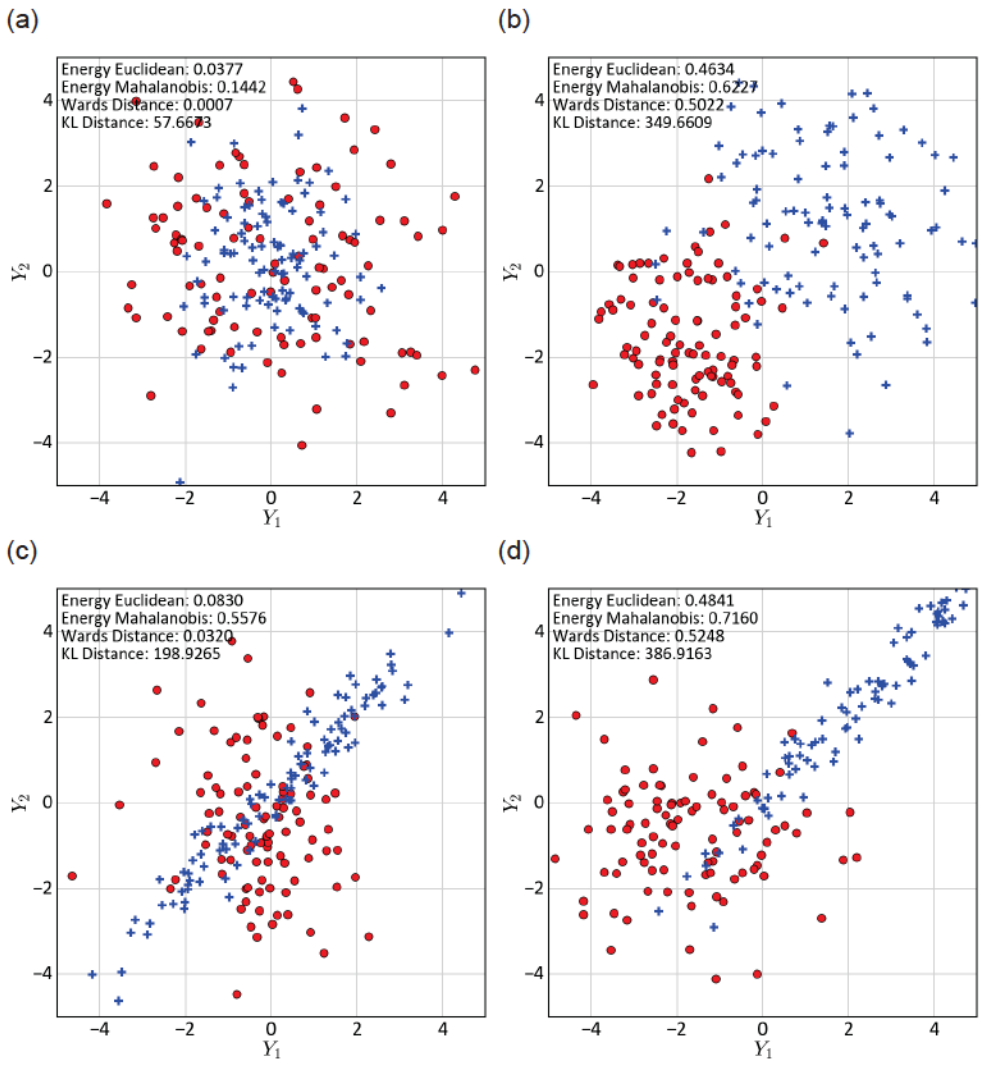


Figure 4.3: The different population difference metrics are calculated for 4 cases between two randomly sampled Gaussian populations generated with different mean and (co)variance. The energy distance, M-energy distance, Wards distance and the KLD calculated between each population is labeled. The scenarios between the red and blue population are as follows: (a) uncorrelated, identical mean, (b) uncorrelated, different mean, (c) linearly correlated (blue) and uncorrelated (red), identical mean, and (d) linearly correlated (blue) and uncorrelated (red), different mean.

respectively, in Fig. 4.4b). The summation over all i locations gives a measure of the connectivity of a particular configuration for a given clustering. In general, a lower total entropy, or higher spatial order, is preferred.

4.2.3 Combined Metrics

The conceptual relationship between the multivariate and spatial metrics is shown in Figure 4.5. Oliver and Webster (1989) noted the trade-off between spatial fidelity of clusters and the multivariate delineation through tuning a kernel bandwidth. Inspection of this metric space is the proposed method for choosing one clustering over another where, for a given level of spatial ordering, a clustering with the lowest WCSS is preferred.

Again, consider the example domain from the beginning of this chapter. Of the two sets of categories, one maximizes the spatial order (Fig. 4.1b), and the other maximizes the multivariate delineation (Fig. 4.2a). In practice some trade-off between spatial and multivariate delineation could be reasonable for a stationary domain. On one hand, additional spatial continuity of the categorical variables is desirable so that the category at the unsampled location can be inferred with minimal uncertainty. Simultaneously, it is desirable to increase the relatedness of samples with respect to their multivariate properties, including their spatial continuity, to ensure the stationary assumption within that domain is reasonable.

4.2.4 Metric Limitations

The main limitation to the proposed metrics comes from comparing configurations with a different number of clusters. Using the domain shown in Figure 4.1 as an example, Figure 4.6a shows the spatial entropy metric for random permutations for each K and the same spatial data configuration. For this experiment the configuration between different numbers of clusters are completely random; the decreasing entropy in Figure 4.6a is entirely due to K. This implies that direct comparison of the spatial entropy between configurations with different K will be problematic. Silva and Deutsch (2012a) also noted that comparison of entropy between configurations is only valid with a consistent K between configurations.

Since clusterings with a different number of clusters do not produce consistent spatial en-

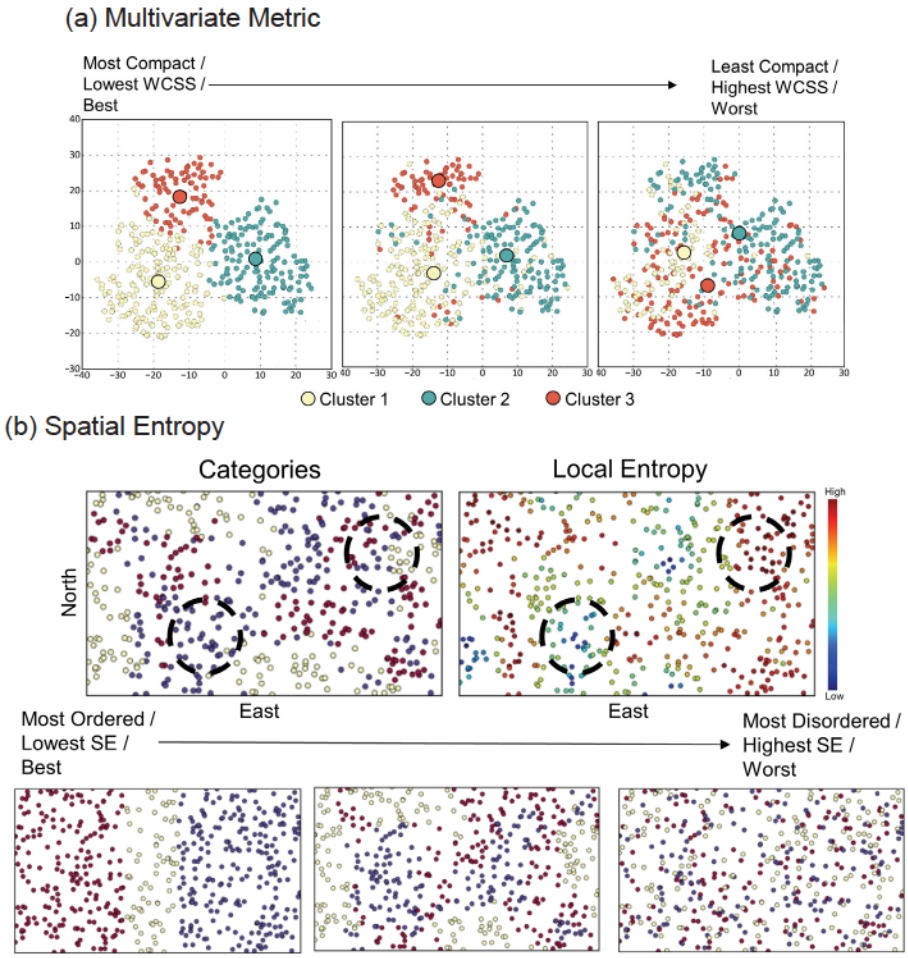


Figure 4.4: Proposed multivariate and spatial metrics to gauge the quality of spatial clusters for the geostatistical workflow. (a) The WCSS shown from best to worst, left to right. (b) Entropy calculated in local windows.

tropy scores they cannot directly be compared. One could argue that the reduced number of categories (and lower scores) from domains with a smaller K reflects a decreased workload for the practitioner and thus represents an improved decision of stationarity. However, a normalized spatial entropy metric could be considered to compare configurations with different K. The entropy calculated for a particular configuration is normalized by the entropy calculated from a random permutation of the cluster labels. For example, if two configurations with a different K are normalized by the entropy of a permutation, the resulting scores can be compared (Fig. 4.6b). The permutation of the cluster codes from each configuration is important as the proportion of each category is constant in the permutation.

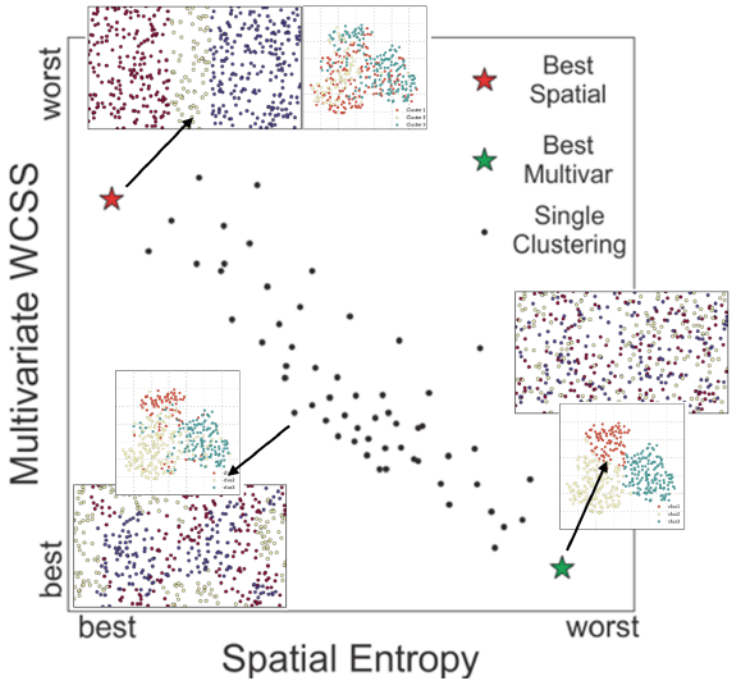


Figure 4.5: Conceptual inverse relationship between the multivariate delineation and the spatial contiguity.

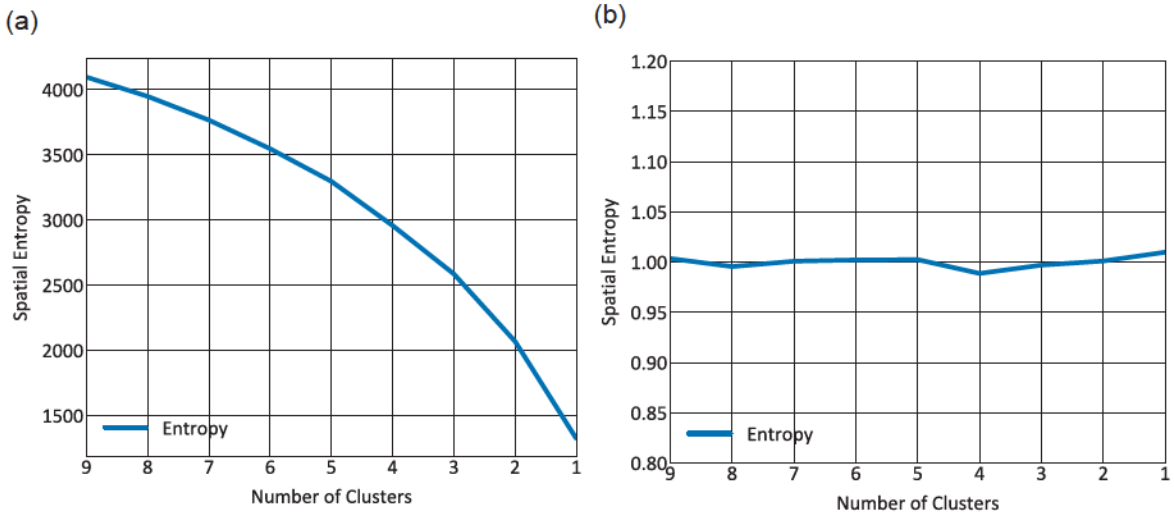


Figure 4.6: (a) Spatial entropy calculated for a constant spatial data configuration. (b) Spatial entropy of the random configuration normalized by the entropy of another random configuration with the same K.

4.3 Dual-Space Ensemble Spatial Clustering

One of the issues with spatial clustering for stationary decisions is that subjective decisions are required to ensure that the generated clusterings are representative and suitable for geostatistical analysis. The clustering metrics introduced in this chapter can be used to guide selection of different decisions of stationarity for a geostatistical analysis. However, parameterizing different algorithms to produce suitable results is still a major challenge for spatial clustering applied to stationary decision making. Furthermore, the results generated between experts and algorithms are different and perhaps incompatible.

A new clustering algorithm is proposed that generates many clustering realizations using a random-path and a 'dual-space' search. The random path and ensemble analysis simplifies the parameter inference required from the user. The implementation uses a K-nearest-neighbor spatial search and the multivariate difference metrics developed above. The full algorithm is given in Algorithm 4. The proposed clustering algorithm has three stages outlined in the following section, and is demonstrated conceptually in Figure 4.7. In the following text consider a set of N geographically located samples where M attributes are recorded at each location.

4.3.1 Stage 1, Primary Spatial Merging

Stage 1 merging generates spatial-first groupings, where samples are merged primarily considering a related spatial neighborhood. For each location along a random path, the neighbors in Cartesian space are identified with a K-nearest-neighbor spatial search that is optionally anisotropic. The Euclidean distance in M-dimensional attribute space is calculated between the current location and those found in the local search. Locations with the smallest M-dimensional distance are merged into mini-clusters. This phase of merging is repeated until every sample belongs to a mini-cluster.

4.3.2 Stage 2, Secondary Multivariate Merging

Stage 2 finalizes a clustering realization by iteratively merging the $K_{mini-clusters}$ generated from stage 1 to the K_{target} number of clusters. At this stage the K_{target} is not required to be the final

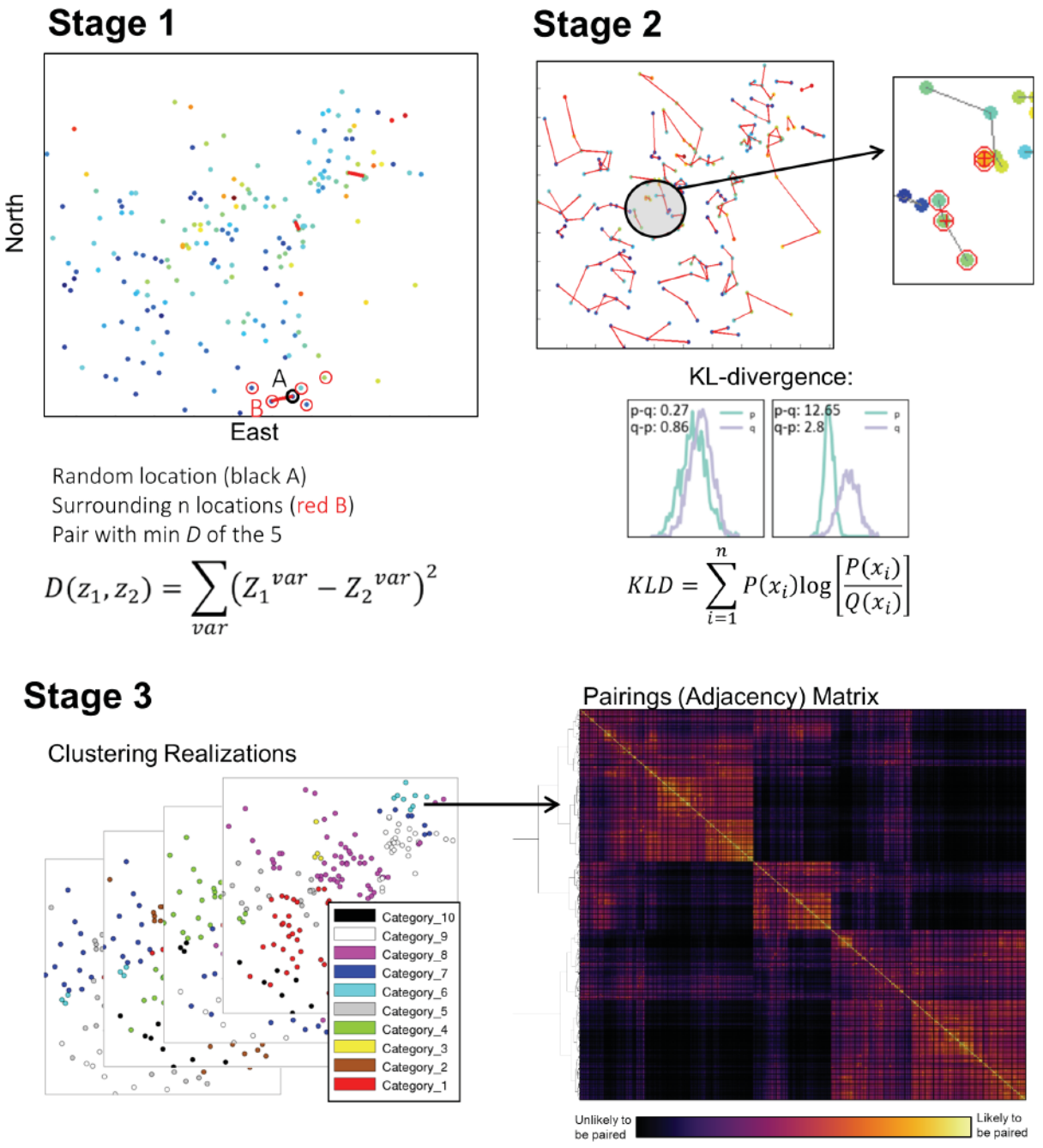


Figure 4.7: Conceptual workflow for random-path iterative-agglomerative spatial clustering. Stage 1: samples merged based on location, Stage 2: mini-clusters merged based on population distances, Stage 3: Calculate the pairings matrix and extract final labels.

Algorithm 4 DS clustering algorithm

$spatial\_tree$ = kdtree built from data locations;
for each L **iteration do**
    Generate a random path through the data locations;

    ▶ Stage 1 Merging
    for each N **data location do**
        Query $spatial\_tree$ for the P closest locations around $\mathbf{x}_i$;
        Compute the M dimensional Euclidean squared distance between the current location and each P neighbor found above;
        Sort P locations based on squared Euclidean distance;
        Merge randomly between 1 and P 'closest' neighbors based on the sorted distances;

    At this stage we have $K_{mini-clusters} >> K_{target}$ clusters that are spatially compact;

    ▶ Stage 2 Merging
    Generate a random path through the mini-clusters;
    $K_{current} = K$;
    for each K **mini clusters do**
        if $K_{current} == K_{target}$ **then**
            Go To stage 3;
        Randomly choose a population difference metric from the available pool;
        Calculate the population difference between the current cluster and all other clusters;
        Merge the two closest mini-clusters;
        $K_{current} = K_{current} - 1$;

▶ Stage 3, Final Class Labels
Build a pairings matrix (adjacency matrix);
Hierarchical clustering of the pairings matrix extracts the dominant configuration for the set of random-path clusterings;

K for the dataset since the ensemble analysis in stage 3 can be used to infer K. Each merge considered in stage 2 uses multivariate distribution difference metrics such as the Ward SSE distance (Székely & Rizzo, 2005; Ward, 1963), the energy distance (Rizzo & Székely, 2016), or the KLD (Hershey & Olsen, 2007) (detailed above).

4.3.3 Stage 3, Final Class Labels

The ensemble of clusterings is recorded in an $N \times L$ matrix where each column contains N labels for each L clustering realization. Ensemble clustering techniques are used to extract the final cluster labels consisting of the consensus of all realizations (Sec. 2.4.2; Strehl & Ghosh, 2002).

The simplest consensus function consists of hierarchical clustering of a sample similarity

matrix (Manita, Khanchel, & Limam, 2012; Strehl & Ghosh, 2002). The $N \times N$ sample similarity matrix counts the number of times each i location is found in the same cluster as each j location, $i, j = 1, ..., N$. The similarity matrix normalized by the number of clusterings L gives the likelihood for each sample to be paired with one another over all clustering realizations. Hierarchical clustering of the similarity matrix with Ward's minimum variance is used to extract the dominant configuration (Strehl & Ghosh, 2002; Ward, 1963).

4.3.4 Synthetic Example

A simple test domain shown in Figure 4.8 is used to illustrate the proposed clustering. The dataset consists of two variables which are approximately normally distributed, with an inter-variable relationship that is not completely described by linear correlation (Fig. 4.8e). Four different clustering algorithms are applied: 1) hierarchical clustering; 2) hierarchical clustering with the coordinates as data; 3) Morans-autocorrelation based spatial clustering (Sec. 2.3.5.2; Scrucca, 2005); and 4) the proposed dual space search (DS) clustering developed above.

The hierarchical clusters generated without considering the coordinates best delineate the multivariate space into 3 clusters (Fig. 4.9a). However, as expected, this generates clusters that have the poorest spatial properties for this domain (Fig. 4.9e). Hierarchical clusters using standardized coordinates are more spatially contiguous at the expense of multivariate delineation (Fig. 4.9b and f). The Moran's clusters have less spatial mixing than the multivariate only clusters (e.g., Fig. 4.9g vs e) and multivariate properties between the first two clusterings. Finally the proposed DS algorithm generates clusters that have similar spatial contiguity compared to the Moran's clustering (e.g., Fig. 4.9h vs g) but with better multivariate delineation (Fig. 4.9d).

To understand how each variable influences each clustering algorithm, a RF classifier is trained to predict the cluster labels using the M input variables and the sample locations (Breiman, 2001; Fouedjio, 2016a; Liaw & Wiener, 2002). Variable importance is calculated from the number of times a variable is used to improve the split results for all branches over all binary decision tree's trained in the RF. The more a variable is used to improve the results of the RF classifier over all decision tree's, the more important that variable is for classification. Predictably, spatial coordinates are the least important variable to predict multivariate-only cluster

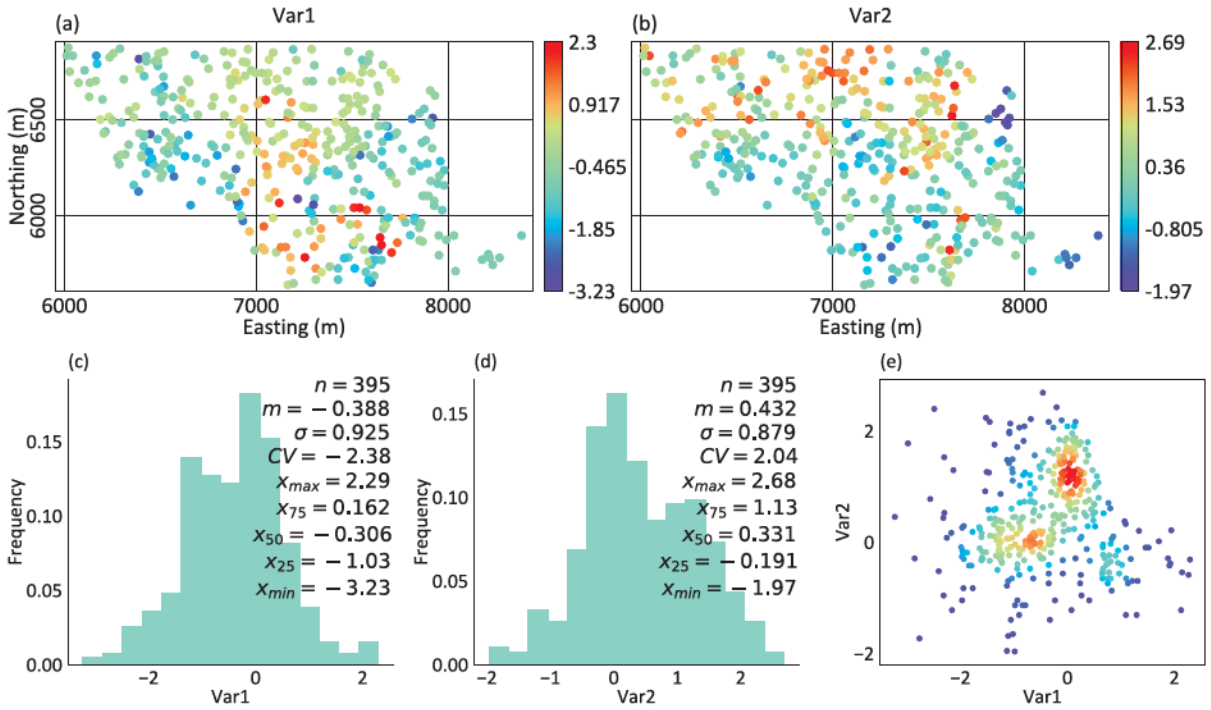


Figure 4.8: (a, b) Synthetic test data locations showing the value of the normal scored variables sampled at each location. (c, d) Univariate histograms and (e) bivariate scatter plot of the two variables in the test dataset.

labels, whereas they are the most important when included as data in the clustering algorithm (Fig. 4.10). Predicting cluster labels from the spatial methods shows increased dependence on coordinates when compared to the multivariate only method (Fig. 4.10).

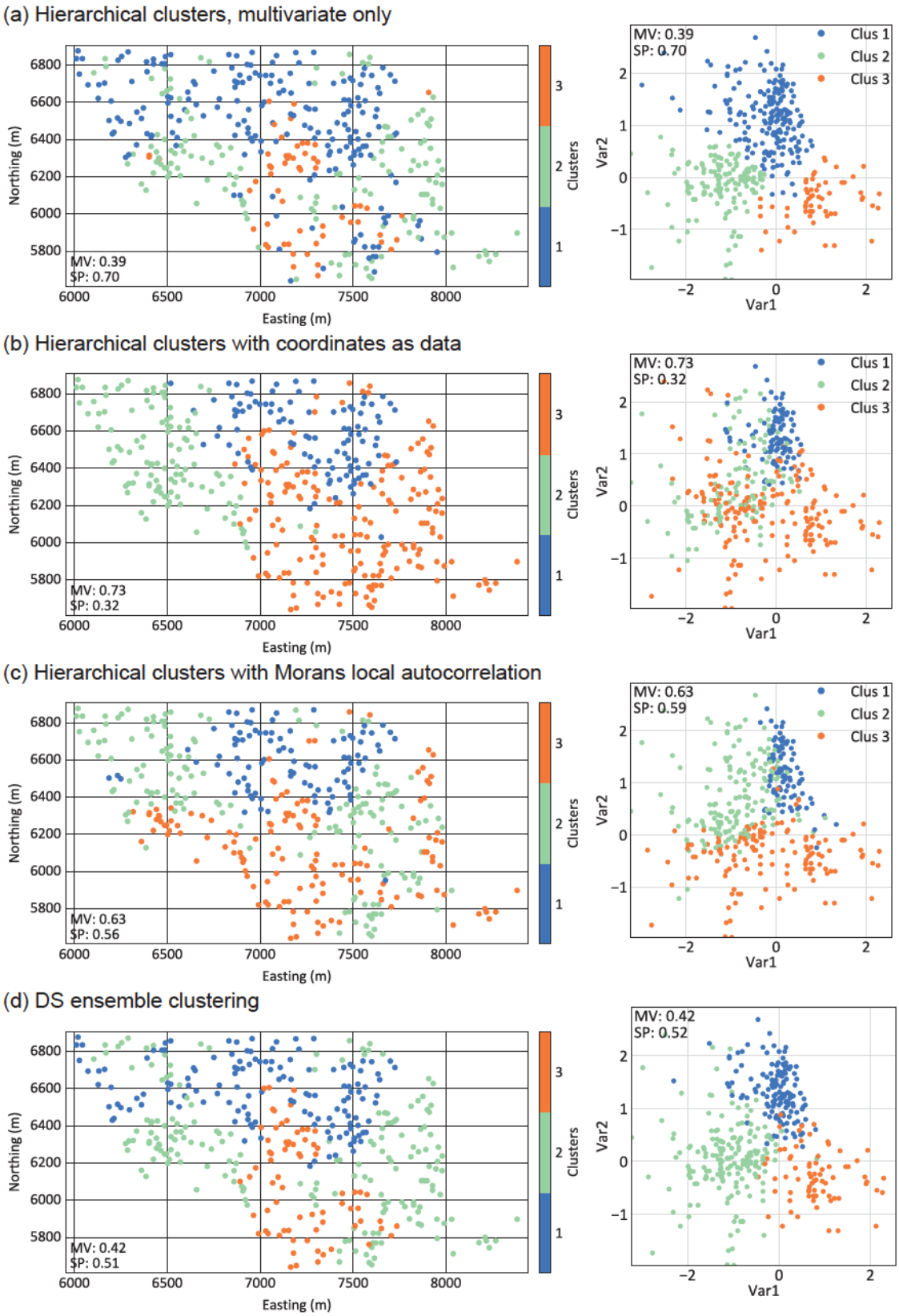


Figure 4.9: Spatial cluster results from (a) hierarchical clustering, (b) hierarchical clusters with standardized coordinates as data, (c) autocorrelation-based spatial clusters, and (d) the proposed random-path clustering algorithm.

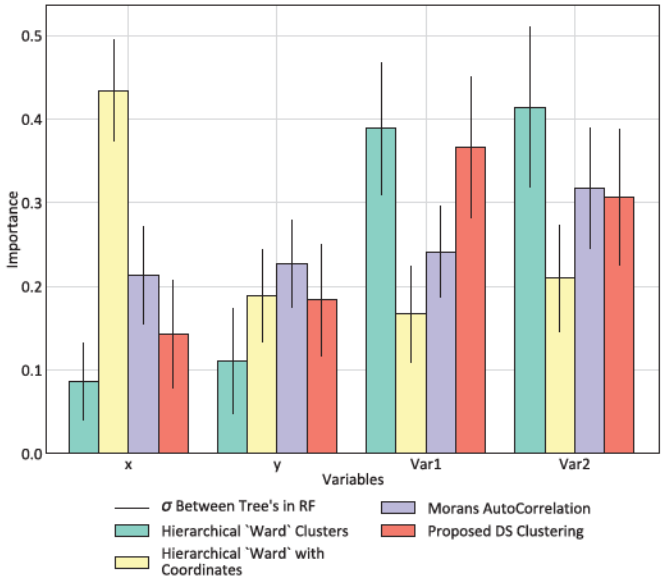


Figure 4.10: Variable importance determined by training a RF classifier to predict the cluster labels for each clustering algorithm, as in (Fouedjio, 2016a)

4.3.5 Addressing Strings of Data

Strings of tightly spaced and highly correlated data are an outstanding issue for spatial clustering of geostatistical datasets (e.g., Fig. 4.1b; Romary et al., 2015). The solution presented here uses an anisotropic spatial search for stage 1 merging to account for the contrasting sampling density based on orientation. The Cu-Au porphyry dataset introduced in this chapter is used to test this solution (Fig. 4.11). An anisotropic search is defined that is isotropic in the horizontal plane (major = minor) with a shortened vertical axis equal to 0.5 * major. A 3×3 rotation matrix **R** is calculated using the $ang1$, $ang2$, $ang3$, r_1, and r_2 conventions (as in: Rossi & Deutsch, 2014) and the anisotropic distance is computed as:

$$D_{aniso} = \mathbf{R} \cdot ((\mathbf{u_1} - \mathbf{u_2})^T)^2 \tag{4.9}$$

The K-nearest-neighbor search is performed in the anisotropic space. The same suite of 4 clustering methods are applied. In this case, the autocorrelation-based clustering and the DS clustering use the anisotropic 3D search.

Clustering results are shown in Figure 4.12. Similar to the synthetic dataset the multivariate delineation is best in the case of the multivariate-only clusters (Fig. 4.12a). The rock types from the geological log generate the most spatially continuous units, however, this comes with the

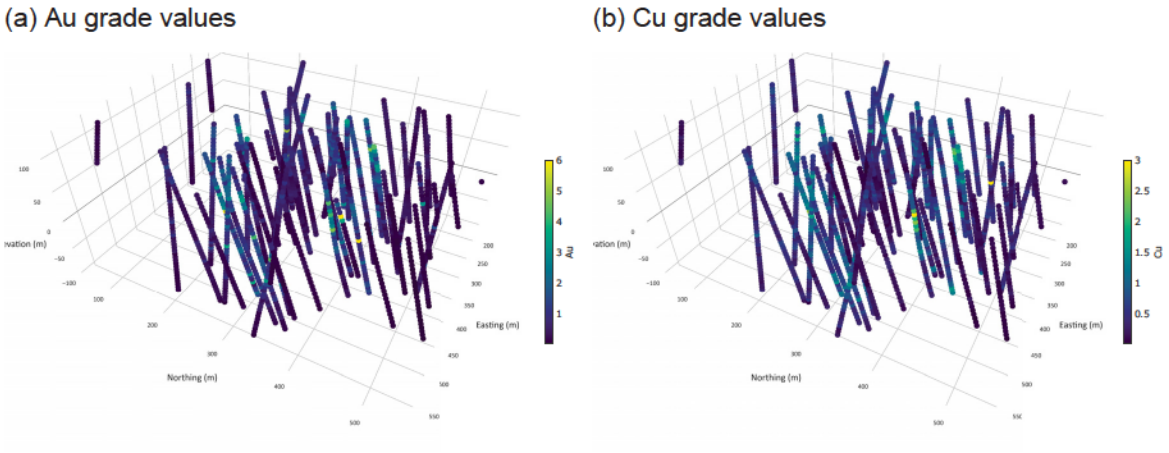
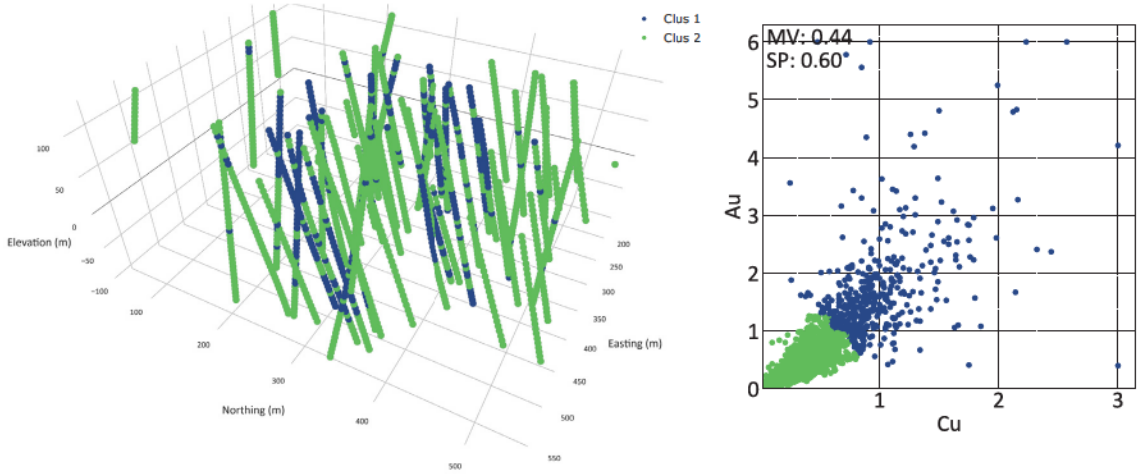


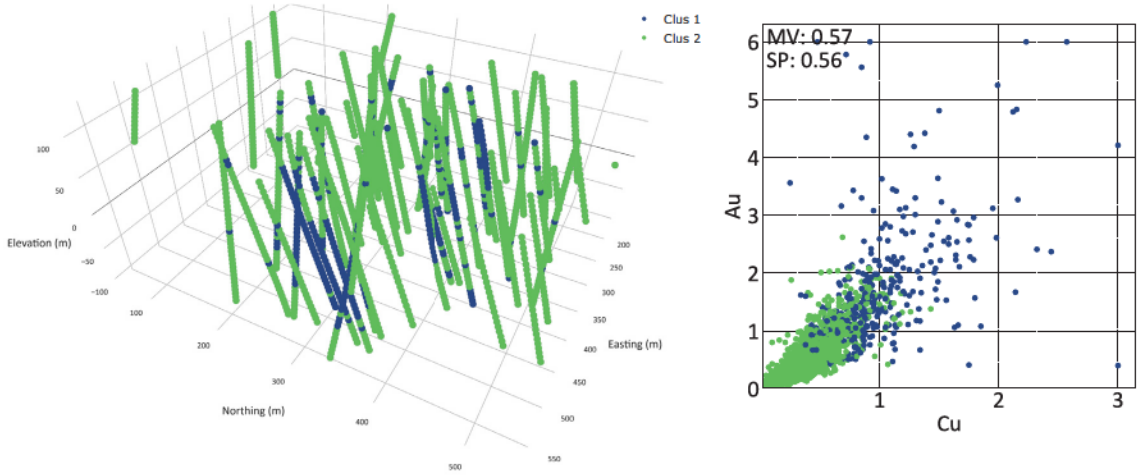
Figure 4.11: Location map colored by (a) Au and (b) Cu grade values at each location.

poorest delineation of multivariate space (Fig. 4.1). It should be noted that the units defined in the geological log are likely defined to capture distinct geological differences (e.g., intrusive phases) rather than the spatial-multivariate populations targeted for geostatistical modeling. Hierarchical clustering with the coordinates as data results in clusters that spatially similar to the multivariate-only clusters but with a higher proportion of cluster 2 (Fig. 4.12b). The result is a significantly more ordered spatial domain (more areas with only 'Clus 2'), but with additional mixing of the populations in multivariate space. The autocorrelation-based spatial clusters are more spatially contiguous than the multivariate-only clusters, however, the multivariate delineation suffers slightly (Fig. 4.12c). DS generates a configuration compatible with the local-autocorrelation clusterer, with similar multivariate delineation and only marginally poorer spatial contiguity (Fig. 4.12d). Depending on the preferences of the modeler, the DS clustering can easily be tuned to generate either better spatial delineation or better multivariate delineation simply by choosing how many samples are merged during stage 1 of the algorithm.

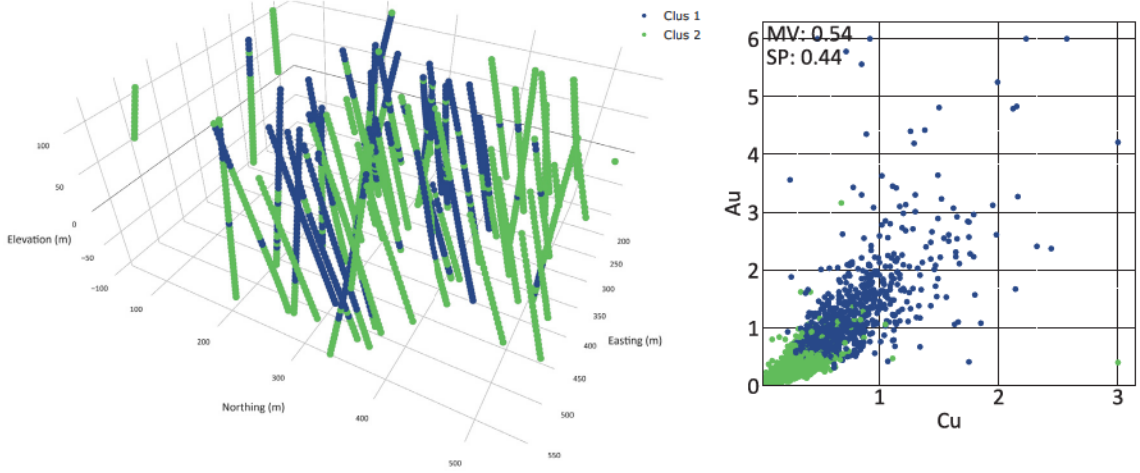
(a) Hierarchical clusters, multivariate only



(b) Hierarchical clusters with coordinates as data



(c) Hierarchical clusters with Morans local autocorrelation



(d) DS ensemble clustering

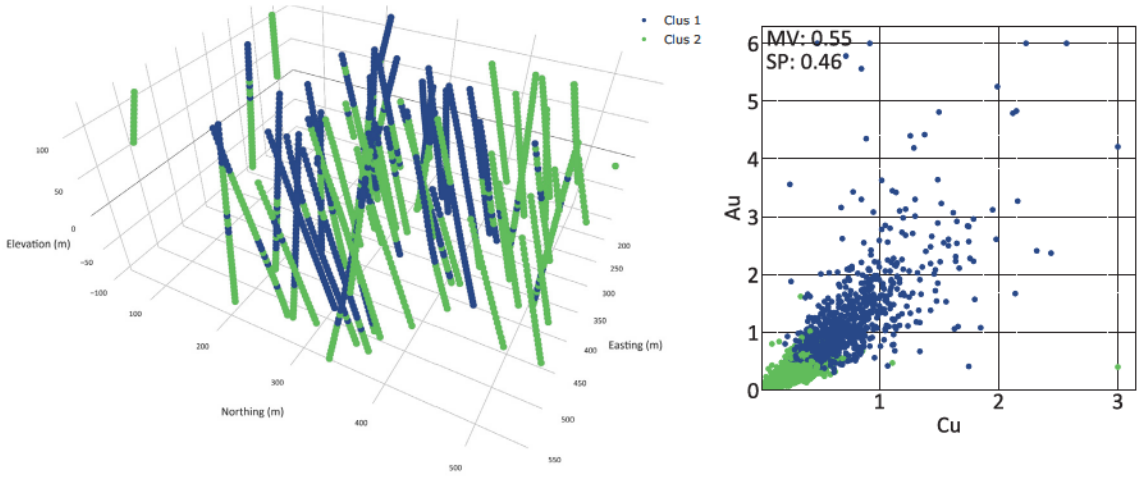


Figure 4.12: Clustering results from (a) hierarchical clustering, (b) hierarchical clusters with standardized coordinates as data, (c) autocorrelation-based spatial clusters, and (d) the proposed random-path clustering algorithm.

4.4 Ensemble Clustering for Uncertain Stationary Decisions

The idea of generating multiple different decisions of stationarity for uncertainty characterization is a novel approach for capturing the uncertainty associated with stationary domaining in geostatistical workflows. The prior uncertainty of input parameters must be expressed for integration into geostatistical modeling. Conventional stationary domaining techniques have previously rendered incorporation of this uncertainty impractical.

A set of consensus functions for traditional ensemble analysis are introduced in Section 2.4.2. The consensus function combines information from all clusterings in an ensemble into a single improved clustering. All consensus functions operate on information shared between clusterers, generally without information about the problem accessible to each individual clusterer (e.g., attributes that define individual clusters). This can be seen as both a strength and a weakness of ensemble clustering, depending on the application (Strehl & Ghosh, 2002). However, most studies note that domain-specific knowledge improves clustering results, since decisions related to parameterization and important subspaces can be tailored to match known characteristics of the domain (Şenbabaoğlu, Michailidis, & Li, 2014; Ghaemi, Sulaiman, Ibrahim, & Mustapha, 2009; Strehl & Ghosh, 2002).

For geostatistical datasets, where one must consider both the multivariate and spatial delineation of the final classes, an enhancement to traditional consensus functions is proposed by using the clustering metrics developed in Section 4.2. The multivariate-spatial metrics are used in a preprocessing step to preferentially select a subset of the clustering ensemble that has improved spatial and multivariate properties. The consensus of the sub-ensemble generates an improved spatial clustering since it incorporates properties of only the best clusterings in the ensemble.

The clustering ensemble is used for three main cases of geostatistical analysis in this work: 1) to improve a single set of clusters for geostatistical estimation; 2) generate a small set of scenarios evaluate the main configurations with estimation; and 3) to generate L equally likely realizations of clusterings that comprise the prior parameter uncertainty for the decision of stationarity, for use in simulation-based uncertainty characterization workflows. The ensemble techniques used in the previous section combined with the clustering metrics comprise the novel geostatistical ensemble consensus function for spatial clustering ensembles.

4.4.1 Improved Stationary Domains and Stationary Domain Realizations

To generate the best set of stationary domains considering the clustering ensemble, typically one would consider the consensus of all clusterings (brown; Fig. 4.13). To improve the single classification, a pre-processing of the ensemble is proposed to prioritize which clusterings form the consensus clustering. This 'geostatistical consensus function' comes from utilizing the metrics of domain goodness combined with conventional ensemble consensus functions.

The proposed geostatistical consensus function is as follows: first, the spatial and multivariate metrics are calculated for all clusterings in a clustering ensemble (black crosses; Fig. 4.13). In this spatial-multivariate metric space, clusterings that have the best multivariate scores for a given spatial score are chosen (blue crosses; Fig. 4.13). This sub-ensemble comprises L clustering realizations that are the 'best' in terms of the spatial and multivariate properties. This sub-ensemble addresses two of three proposed applications of ensemble clustering to geostatistics; either the consensus of the sub-ensemble is taken to generate a single consensus clustering with improved spatial-multivariate properties (pink star; Fig. 4.13), or clusterings from

this sub-ensemble can be used directly in a simulation based workflow to assess the uncertainty associated with assigning samples to different stationary domains (blue crosses; Fig. 4.13).

4.4.2 Stationary Domain Scenarios

The final proposed application of cluster ensembles for geostatistics is to determine the most dominant configurations represented in the ensemble. This requires a pairwise comparison of the similarity of all clusterings in the ensemble. An issue with direct comparison of clusterings for similarity comes from the randomized initialization of each clustering algorithm, which results in different labels assigned to each cluster between clustering realizations. The net effect is incompatible labeling between otherwise identical clusterings (Fig. 4.14). The adjusted Rand index (ARI) (Hubert & Arabie, 1985; Steinley, 2004) is a popular metric used to compare clusterings since it is invariant to permutations, produces scores between $[0, 1]$, and can be used to compare clusterings with different K. Consider two clusterings with K_1 and K_2 clusters, respectively. Calculating the ARI between the clusterings starts by converting the clustering label vectors to binary matrices, $\mathbf{B}_1$ and $\mathbf{B}_2$ (i.e. $\mathbf{B}_1$ is a $N \times K_1$ column-wise binary matrix recording membership of points in each cluster) and then building a $K_1 \times K_2$ matrix $\mathbf{C} = \mathbf{B}_1^T \mathbf{B}_2$. Each entry in the matrix $\mathbf{C}$ records the number of times a pair of points co-occur in each cluster. Once the co-occurrence matrix $\mathbf{C}$ has been calculated the ARI between the clusterings is given by (Hubert & Arabie, 1985):

$$ARI = \frac{\sum_{i,j}\binom{c_{ij}}{2} - \frac{\sum_i \binom{c_i}{2}\sum_j \binom{c_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{c_i}{2} + \sum_j \binom{c_j}{2}\right] - \frac{\sum_i \binom{c_i}{2}\sum_j \binom{c_j}{2}}{\binom{c}{2}}} \tag{4.10}$$

The binomial expansions $\binom{c}{2}$ computes the total number of combinations of each configuration possible between each cluster. To identify the most dominant configurations in the ensemble, the pairwise similarity between clusterings is calculated using the ARI and the result is accumulated in an $L \times L$ cluster similarity matrix. Hierarchical clustering of this clustering similarity matrix results in the dominant cluster configurations from the ensemble, and for each configuration the consensus of the subset is found by building an $N \times N$ pairings matrix using the clusterings from

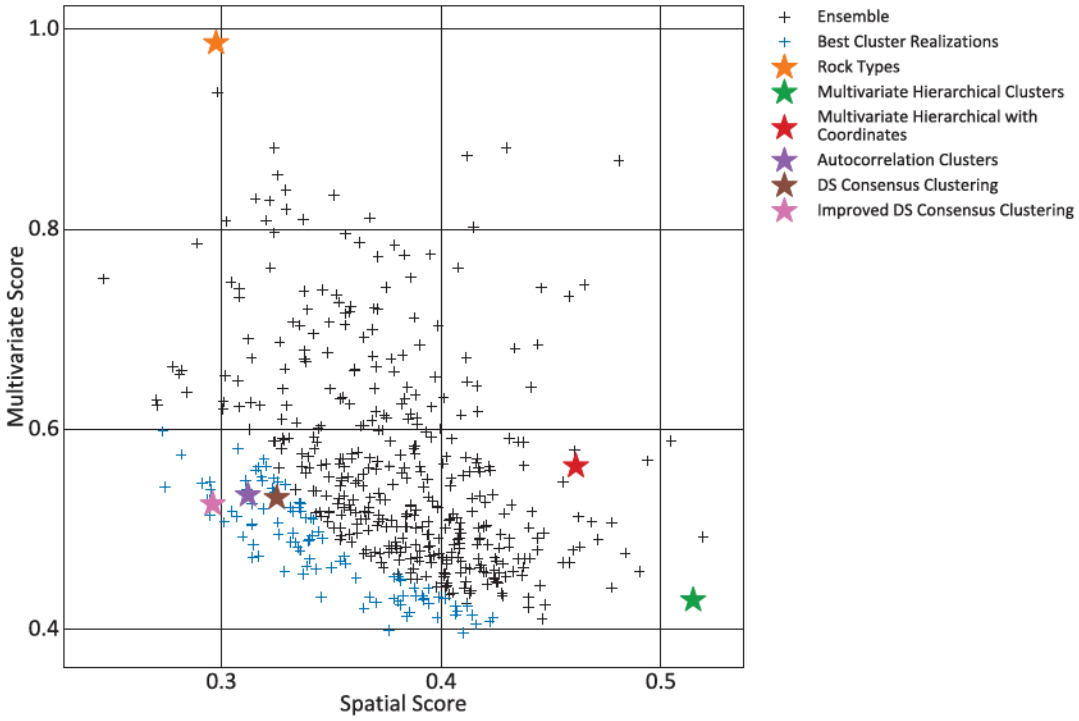


Figure 4.13: Example of the geostatistical consensus function, using the clustering metrics calculated for different stationary domains for the porphyry dataset. Blue crosses represent those that have the best multivariate properties for a given level of spatial contiguity. Selected single realization decisions of stationarity are shown for reference.

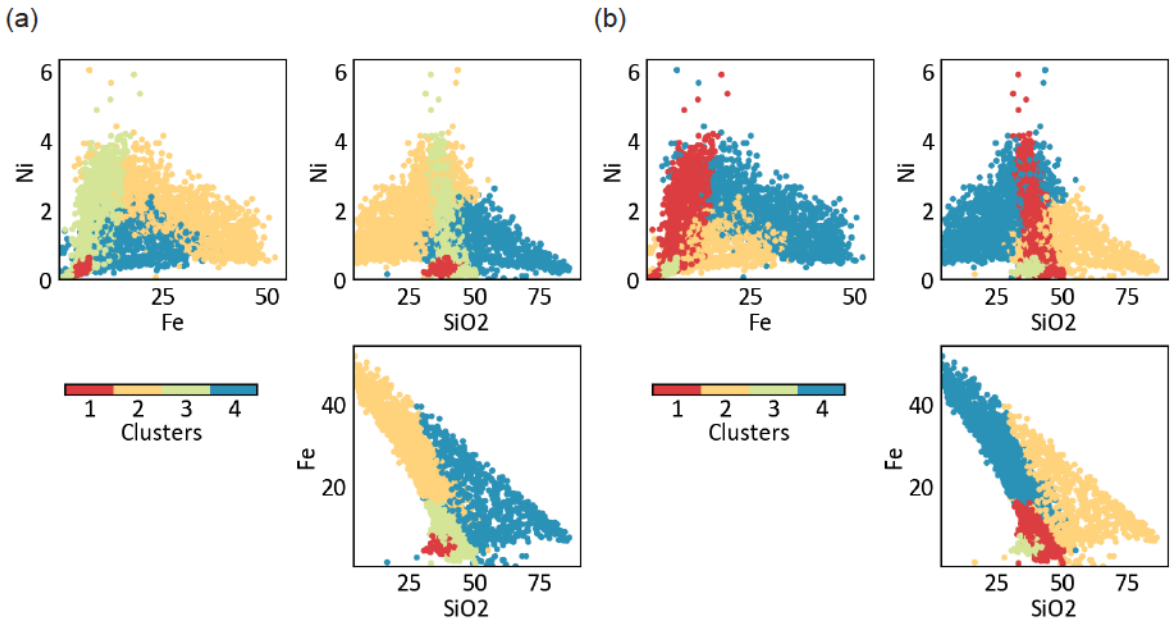


Figure 4.14: The outcome of two different clusterings of the same dataset with the same parameters. The permutation of cluster labels results in incompatible cluster codes between otherwise identical categories.

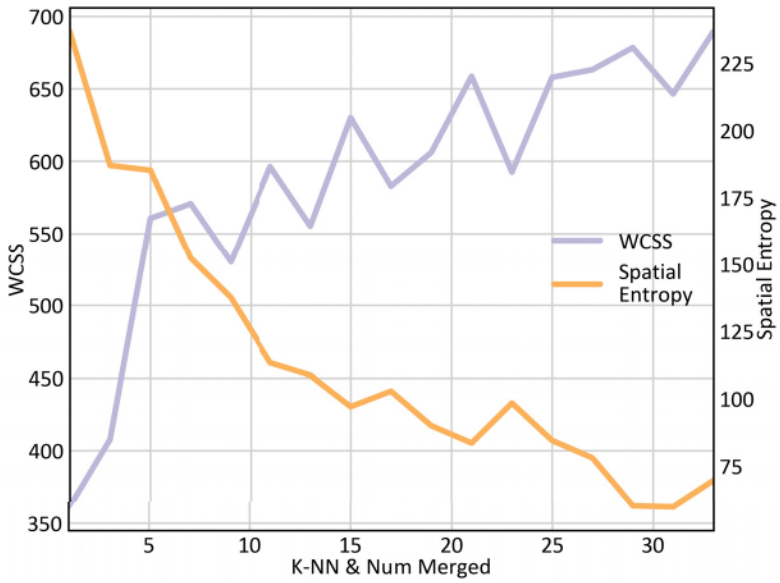


Figure 4.15: WCSS and spatial entropy scores for different number of nearest neighbors and numbers of data merged in the stage 1 search.

each subset, and extracting the consensus clustering from the pairings matrix with hierarchical clustering, as above.

4.5 Parameterization Guidelines

Nearly all clustering algorithms require K as input. The exception is hierarchical clustering, where K can be inferred by inspecting the associated dendrogram. The ensemble clustering methods developed in this chapter benefit from a hierarchical-clustering based consensus function, which allows K to be inferred following ensemble construction. Generally, a lower K is preferable to reduce workflow complexity in the geostatistical modeling that follows.

The DS clustering developed in Section 4.3 requires few input parameters: search parameters (N_{search} and anisotropy), the number of samples merged during stage 1 merging, and a lower proportion threshold to reject clustering realizations.

Search parameters should generally be chosen to match the variability in the underlying variables, with some consideration to differences in sample density along drill holes relative to between drill holes. A general guideline is to mimic the anisotropy and range of the underlying variables. The number of samples found in the spatial search should be similar to the kriging

parameters; 10-25 in 2D and 25-50 in 3D.

The number of samples merged (N_{merged}) in stage 1 can optionally be specified, however in this work it is randomized between 1 and N_{search}. The choice of the number of samples merged during stage 1 merging effectively creates the 'dial' that can be used to tune the multivariate delineation or spatial connectivity of the resulting clusters. Given a constant N_{search}, if N_{merged} during stage 1 is increased, the final result has additional spatial continuity. Conversely, with a small N_{merged}, the output of DS clustering resembles clustering obtained using only multivariate information. The effect of this 'dial' in terms of the metrics developed in Section 4.2 is shown in Figure 4.15. Increasing the number of samples found in the local search in stage 1 causes the DS clusters to have additional spatial connectivity at the expense of multivariate delineation.

The lower threshold rejects clusterings if any one of the categories falls below this proportion. The main idea is to ensure that all clusterings are reasonable and could be used for geostatistical analysis. Generally, with a larger input dataset this threshold can be lower. The key is to ensure that enough samples are available for inference of statistical parameters for modeling.

4.6 Review of Main Points

This chapter introduces clustering metrics and novel clustering algorithms for defining stationary domains for geostatistical analyses. Specifically:

1. The development of several multivariate metrics and a single spatial metric that summarizes differences in stationary domains;
2. The combination of multivariate and spatial properties as an objective measure of stationary domain goodness;
3. Development of a spatial clustering algorithm that reduces the burden of parameterization through the random subspace and ensemble analysis;
4. A method to generate improved spatial clusters considering ensemble analysis and the developed clustering metrics.

4.7 Conclusions

Clustering and spatial clustering tools are effective for partitioning high dimensional spatial datasets in to groups with unique spatial and statistical properties. The clustering metrics proposed in this chapter represent an important contribution for clustering applied to stationary delineation since the result can be assessed based on objective rather than subjective criteria.

Chapter 5

Case Study: Implicit Modeling with Locally Varying Anisotropy

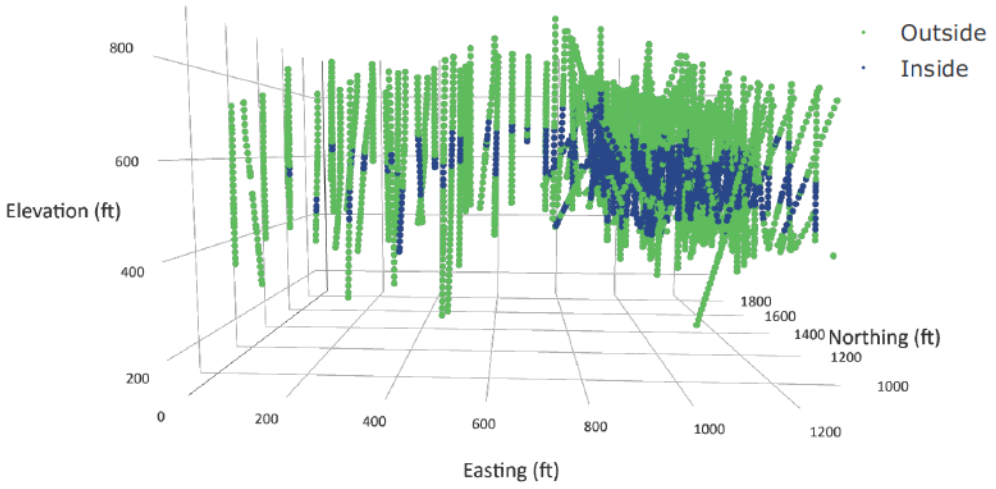
5.1 Porphyry Deposit

A dataset of 4525 samples from 168 drill holes from a Chilean Cu-porphyry deposit is used to demonstrate the implicit geological modeling techniques developed in Chapter 3. The geological log delineates 2 large-scale domains (Fig. 5.1a) that separate the economic supergene-enriched ore from the uneconomic material. The 6 modeling categories are defined to capture smaller scale variations within this large-scale framework (Fig. 5.1b). Categories 3 and 4 comprise the supergene enrichment zone and have the highest average copper grades (Fig. 5.2). This enriched zone is mostly localized to the east where dense drilling is concentrated, however, it is also intersected in the west of the property. Category 1 is mainly unmineralized while 2, 5, and 6 are weakly mineralized. Collectively these units form the unmineralized cap and base of the hydrothermal system. The combined categoriees 3 and 4 comprise the large-scale geological domain that isf modeled in this chapter.

This chapter evaluates different boundary modeling strategies for the interface between the large scale geological domains shown in Figure 5.1a. Implicit modeling is well suited to capturing geological features at this scale since the features are relatively continuous between drill holes. Further, the spatial pattern of the targeted geological domain suggests that local anisotropy of some form may be required to characterize all locations in the domain.

Inspecting the location map in Figure 5.1a shows that an overall horizontal planar structure is the target of boundary modeling. However, from the north-looking view of Figure 5.1a, a gentle folded structure is present, with a central high point and west and east-dipping limbs on the west and east sides, respectively. The portion to the east is thicker than the west, but presumably the two are continuous through the central high feature, and local continuity between the different

(a)



(b)

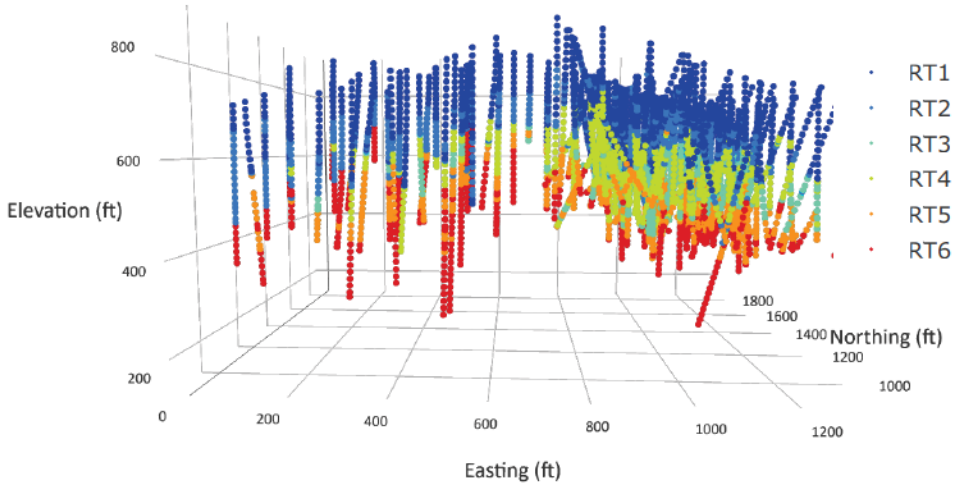


Figure 5.1: Large porphyry location data, looking north. (a) Geological Domains. (b) Categories.

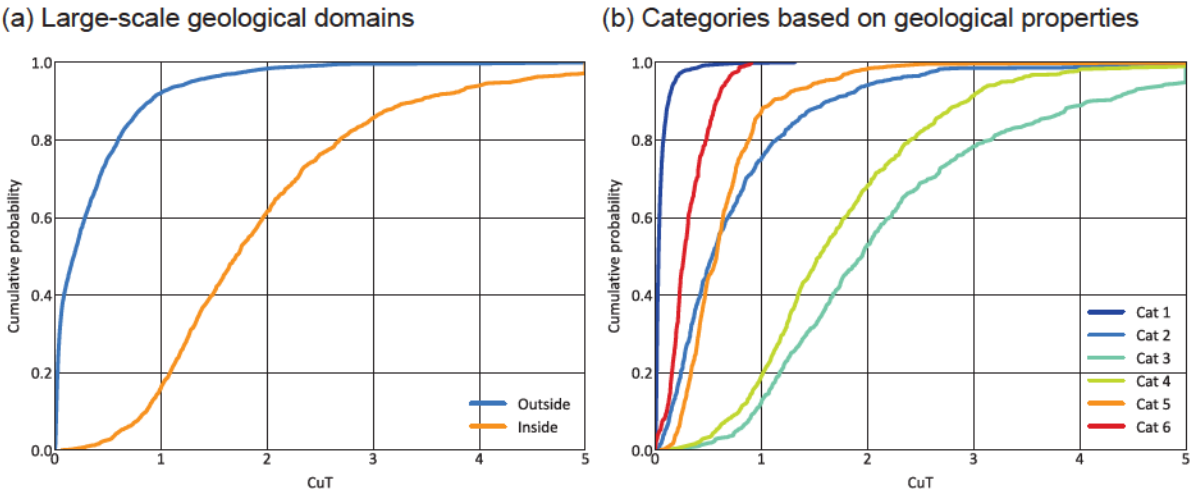


Figure 5.2: Histograms (a) by geological domain, (b) by geological category.

areas should be maintained.

The thick unit on the east side of the domain poses no challenges for implicit modeling. The density of the information in this area imparts strong controls on the recoverable geometry since the dense information controls the local anisotropy, even with isotropic interpolation conditions. However, through the central and west portions of the domain, information is sparser and the local continuity of the geological boundaries may suffer without a representative local interpretation.

All implicit models in this section are generated with a Gaussian kernel in the PU domain decomposition framework developed in Section 2.2.5. Extrapolation from data locations is limited to 200 ft. The parameters of the PU decomposition are constant between all models; 35 data-per-center, 0.01% data overlap during partitioning, and final expansion-overlap of 60% in the x, y and z directions. The models presented in the following sections are built with constant input data, partitioning parameters, and support parameters. The only difference between models is the anisotropy input to the boundary modeling algorithm. Four methods for inference of anisotropy are tested: 1) isotropic; 2) manually inferred global anisotropy; 3) global anisotropy from the variogram model; and 4) iteratively refined local anisotropy (Chap. 3). An isotropic model serves as a baseline for inspection purposes; it is expected that either global or local anisotropy will improve the reproduction of geological features given the interpretation of the geological continuity.

5.2 Globally Isotropic Boundaries

Boundaries generated under isotropic conditions are shown in the north-looking oblique view in Figure 5.3. The reproduction of the dataset is good in the densely sampled area in the east, however, in the central and west-end of the domain the connectivity does not match the interpretation of gently folded geological continuity. The model could also be described as overly rounded, perhaps blob-like, though this is a subjective observation of the shape of the generated boundaries.

5.3 Globally Anisotropic Boundaries

5.3.1 Anisotropy from Manual Inference

The manual inference of anisotropy comes from inspecting the isotropic model and inferring orientations and ranges of continuity that describe the expected continuity of the geological solid. The goal is to address the holes in the sparsely sampled west-end of the deposit. Thus, a horizontal isotropic variogram model without strike or dip rotation is specified; the vertical anisotropic ratio is manually set to $r_2 = \frac{1}{4}$. The resulting boundaries are shown in Figure 5.4.

Boundaries generated with interpreted local anisotropy improve continuity of the geological boundary across the sparse locations by connecting the areas between drill holes in the central portion of the domain (Fig. 5.4b).

(a) Oblique view N

(b) Oblique view SE

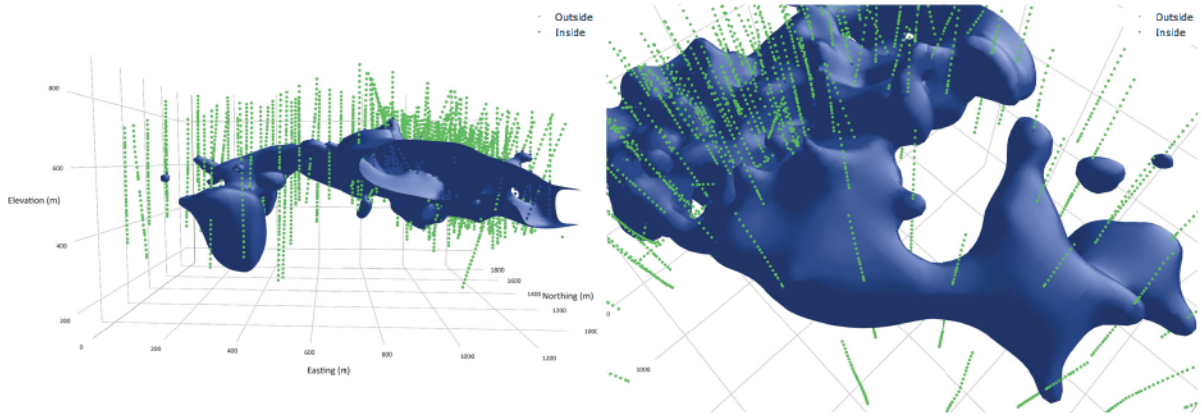


Figure 5.3: Isotropic boundary model.

(a) Oblique view N

(b) Oblique view SE

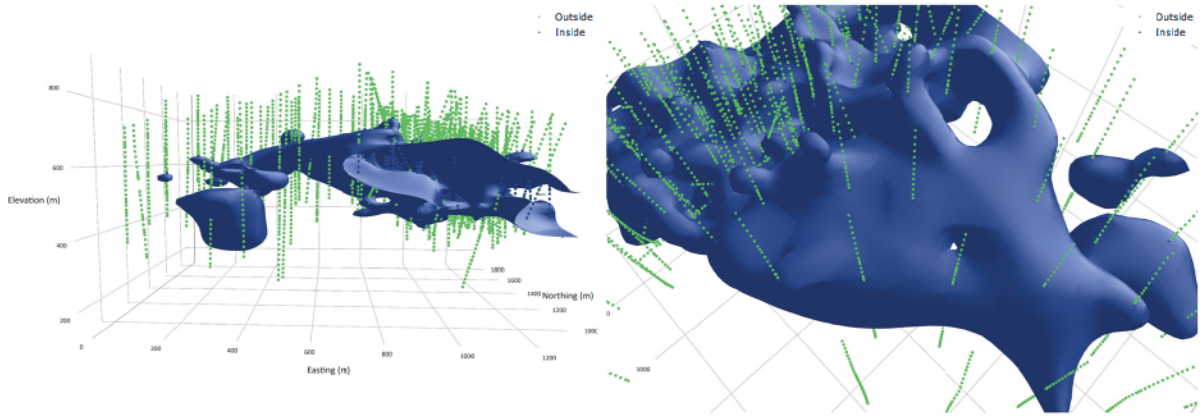


Figure 5.4: Anisotropic boundary model from manual inference.

(a) Oblique view N

(b) Oblique view SE

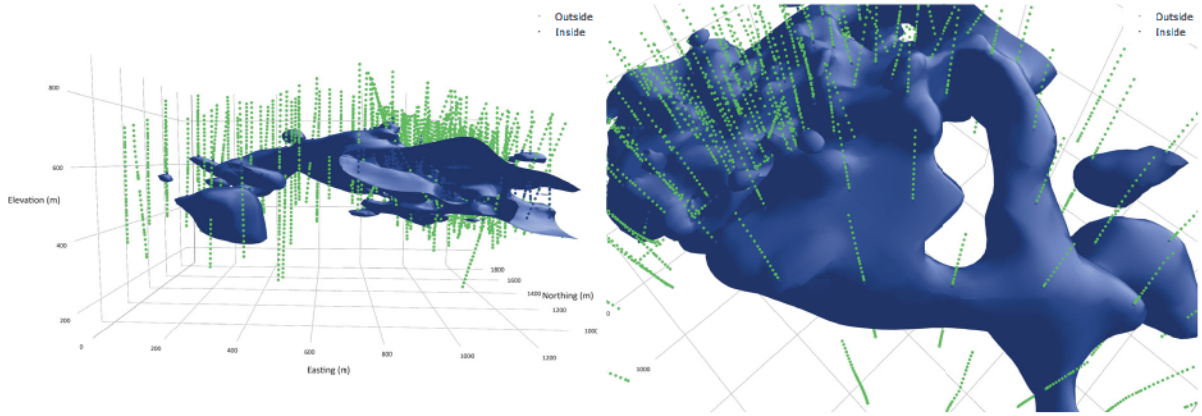


Figure 5.5: Anisotropic boundary model from the variogram model.

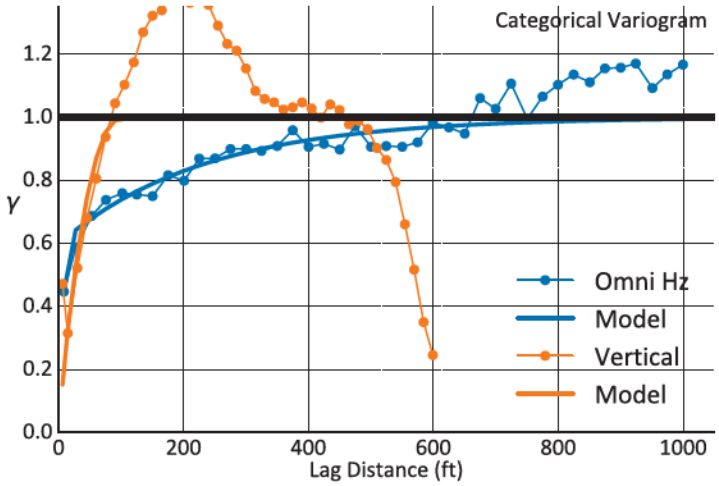


Figure 5.6: Experimental and model indicator variograms for the geological domains.

5.3.2 Anisotropy from Variograms

A set of orientations (ang_1=75, ang_2=-10) for experimental variogram calculation are determined by inspecting the isotropic model in Figure 5.3 and the composite indicator data. An horizontal-omnidirectional indicator variogram is calculated and fit with a model variogram (Fig. 5.6). The range parameters extracted from the fitted variogram models parameterize the RBF kernel with a set of rotation angles and the magnitudes of anisotropy along each direction.

The experimental variograms show two pronounced structures (Fig. 5.6). The first structure accounts for approximately 75% of the variance and shows that there is additional continuity in the vertical direction at a short range. However, the long range structure captures the expected horizontal continuity of the domain, and thus, the set of anisotropic parameters for RBF interpolation are extracted from the second structure. This is a subjective decision according to the interpretation of the continuity and desired properties of the model. Anisotropic ratios $r_1 = 1$ and $r_2 = \frac{a_{vert}}{a_{horz}} = \frac{1}{6}$ are taken from the relative ranges.

Boundary models generated with the variogram-inferred global anisotropy are shown in Figure 5.5. These globally anisotropic models show a minor improvement in continuity in the area of sparse sampling, however, the interpreted global anisotropy does a better job in this area filling out the interpolation.

5.4 Iteratively Refined Local Anisotropy

The iterative refinement algorithm developed in Chapter 3 is applied here to refine local orientations of continuity from the geological boundaries. The iterative algorithm requires a seed model. As shown in Chapter 3, a seed model with the best global anisotropy should be considered.

Figures 5.7 and 5.8 show the N-looking and SE-looking views, respectively, of the seed model (a) and 5 iterative refinements (b) - (f). The local orientations inferred at each stage are shown as vectors in (b) - (f). In each case the vector orientations indicate the local anisotropy that went into building that model, generated from the previous model in the refinements; the final iteration of the algorithm does not generate new orientation vectors. Figures 5.7 and 5.8 focus mainly on the central-west portion of the project since the dense sampling in the east end constrains the shape refinements possible from this technique.

5.5 Geological Model Performance

The goodness of a geological boundary model depends on multiple criteria (Lindsay et al., 2013; Oreskes et al., 1994). Section 3.4 highlights the issues validating implicit geological models citing the deficiency of K-fold validation for capturing the shape properties of the model. In that section, a SA:V metric is proposed to quantify subjectively poor features in implicit models. However, the K-fold validation remains an important criteria to gauge relative predictive performance between different algorithms or parameterizations of the same algorithm operating on the same dataset.

A 5-fold cross-validation and shape-property study is undertaken to assess the performance of each set of anisotropic parameters for reproducing the underlying features. The sample locations colored by fold number are shown in Figure 5.9. The goal of the K-fold analysis is to assess how well the predictive algorithm reproduces the expected domain codes at the removed locations, over all folds. The shape properties of the generated models are assessed by constructing final models considering all samples under each set of anisotropic conditions. Surface properties of each final model are calculated for comparison (as in Sec. 3.4). The different sets of anisotropy include: isotropic (Fig. 5.3); inferred anisotropy (Fig. 5.4); variogram

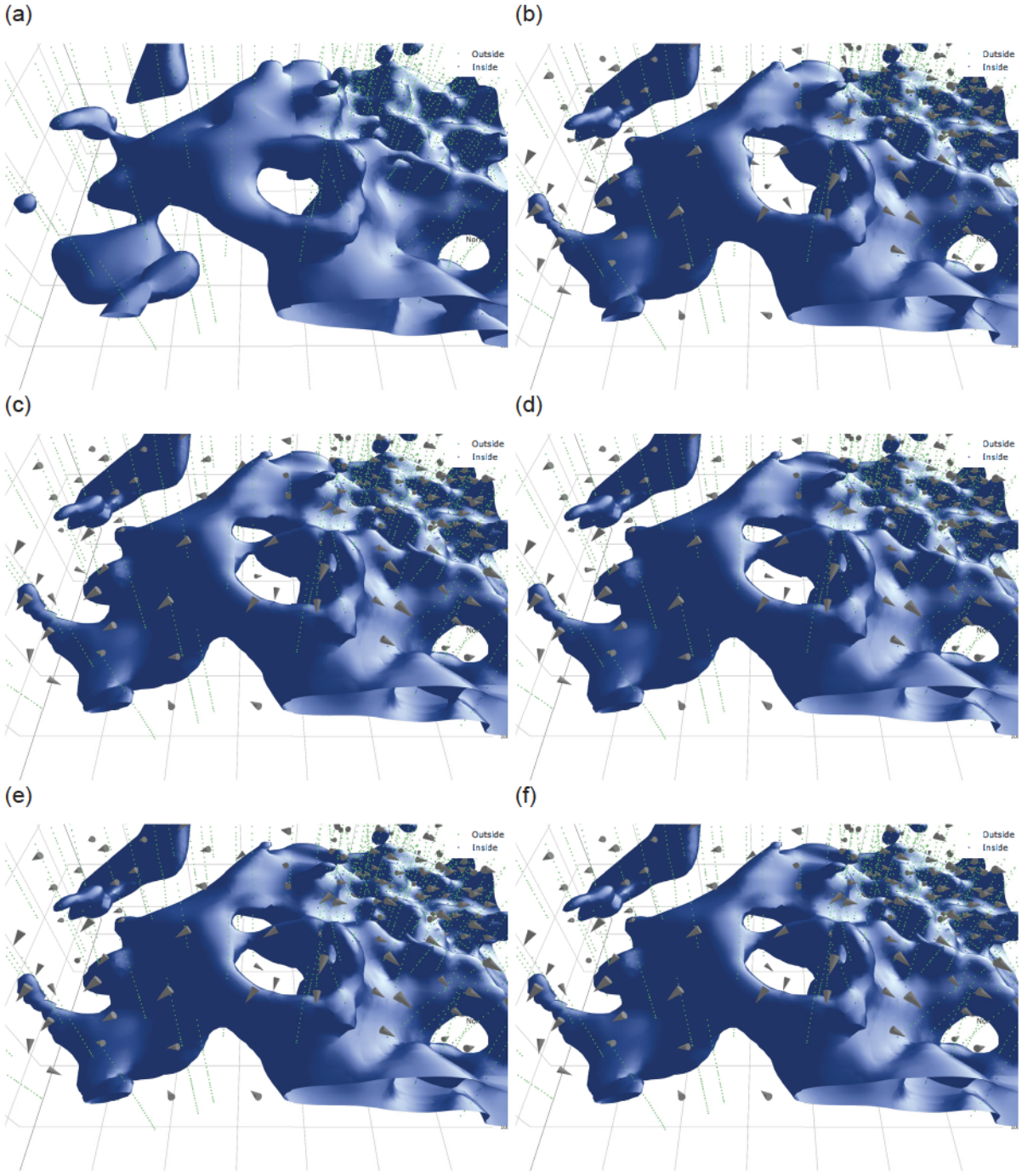


Figure 5.7: Iterative refinements of the boundary model, looking N. (a) Seed model with global anisotropy, (b) to (f) iterations 1 to 5, respectively. Vectors shown in (b) to (f) reflect the orientation of the local anisotropy, sized by the relative magnitude of the anisotropy.

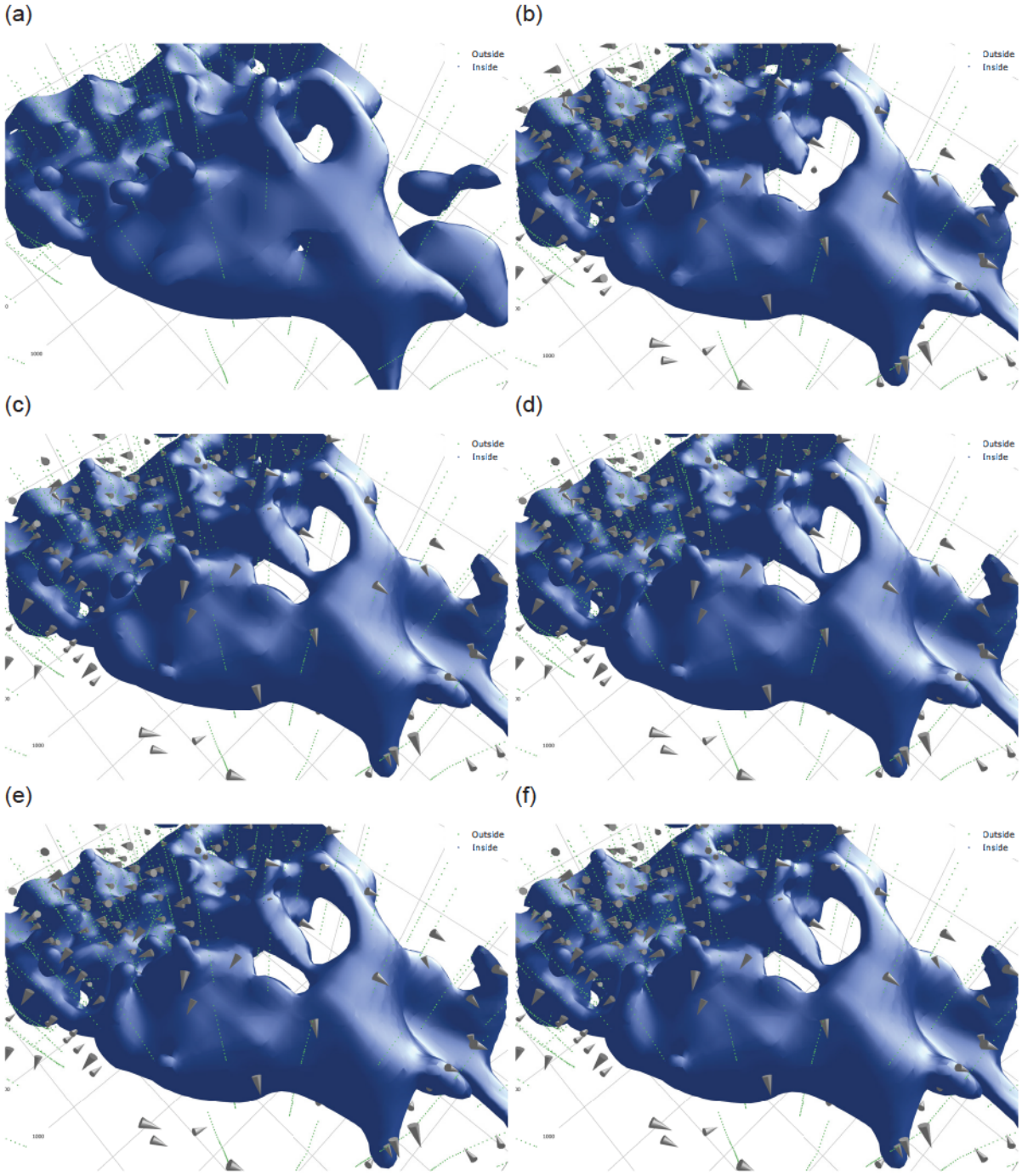


Figure 5.8: Iterative refinements of the boundary model, looking SE. (a) The seed model with global anisotropy, (b - f) Refinement iterations 1 to 5, respectively. Vectors shown in (b - f) represent the orientation of the local anisotropy, sized by the relative magnitude of the anisotropy.

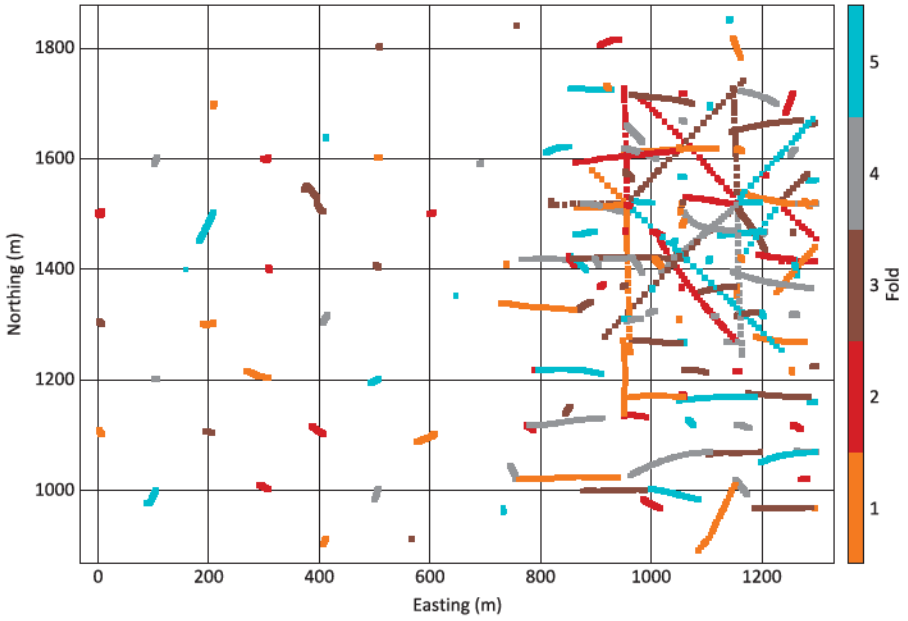


Figure 5.9: Sample locations colored by fold number for K-fold analysis.

anisotropy (Fig. 5.5); and iteratively refined local anisotropy for 5 iterations (Fig. 5.7).

Figure 5.10a shows the percentage of locations where the boundaries generated under each set of anisotropic conditions did not correctly identify the category at the unsampled location. Logically, a lower error is desirable; the best anisotropy is the one that informs boundaries that have the minimum number of misclassified locations. The iteratively refined models generate lower errors than models with inferred or variogram anisotropy. The relatively large difference between the inferred model and the first iterative refinement is interesting as the refinements are seeded with the inferred anisotropy. Interestingly the lowest error is produced with an isotropic interpolator. This is an unexpected result since the subjectively improved properties of the model do not translate into improved cross validation scores.

Figure 5.10b shows the SA:V metric calculated for each implicit model. Here the goal is to minimize the subjectively poor features of bubbly and rounded-looking implicit models. Thus, preference is given to models that maximize the SA:V. Predictably, the isotropic model generates the lowest SA:V score indicating that the model contains blob-like shapes. Global anisotropy improves the models significantly, and the iterative refinements further improve over the global anisotropy. In this case iteration 3, 4, and 5 of local refinements produce the maximum SA:V scores.

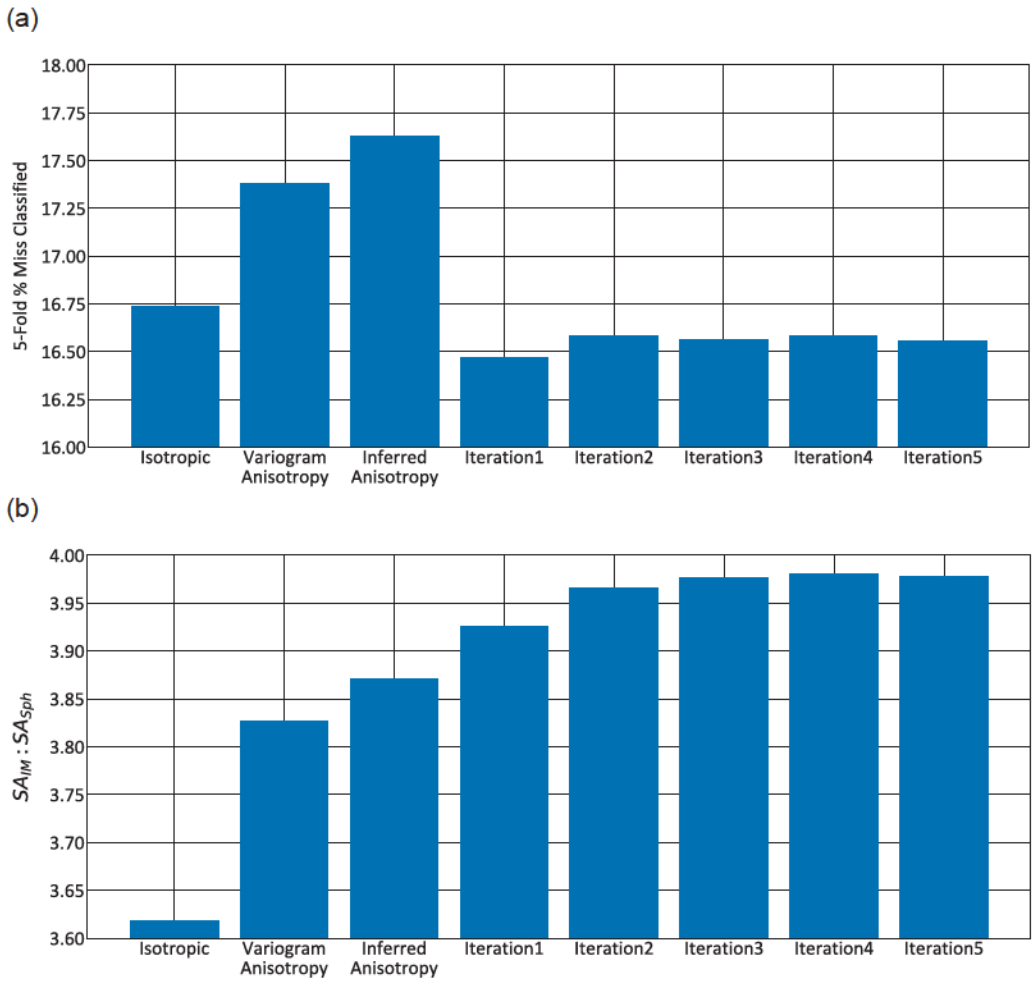


Figure 5.10: Implicit model (a) cross validation and (b) shape properties for the models generated as part of this study. The best model should minimize predictions errors while simultaneously maximizing the $SA_{IM} : SA_{Sph}$.

Ultimately the structural interpretation adopted for implicit modeling is a choice made by the geomodeler based on several criteria, both subjective and objective. This section demonstrates that the iterative refinements proposed for implicit geological models can capture representative local anisotropy that is competitive with global anisotropy in terms if prediction errors and outperforms global anisotropy in terms of shape properties. In conventional workflows the iteratively refined anisotropy could be used to aid the interpretation of local anisotropy for the practitioner - the final choice of global or local anisotropy is left to their discretion.

5.6 Volumetric Uncertainty

The iteratively refined local anisotropy developed in this thesis can be easily incorporated into existing methodologies for quantifying the the volumetric uncertainty in implicit modeling frameworks (Sec. 2.2.7; Wilde & Deutsch, 2012). Twenty-five training-test datasets are generated by randomly assigning 75% and 25% of the drill holes to the training and test set, respectively, for each run. Fifty random C-parameters in the interval [0, 500] are generated. The local anisotropy from iteration 5, above, is used in this case as the representative set of structural anisotropy. The result of error-classification with increasing C-parameter is shown in Figure 5.11.

The guideline to choose a C-parameter from the error in Figure 5.11 is to consider some notion of acceptable error. For example, a large C-parameter tends to dilate boundaries so that a smaller number of locations are incorrectly classified. Manchuk and Deutsch (2015) suggests 2.5% is an acceptable rate for the tabular deposit modeled in that study. There, the C-parameter corresponding to 2.5% error generally represents the larger voids in the sparsely sampled areas. This parameter could also be chosen from expert judgment considering the data spacing and the geometric features. For the current domain the densely sampled regions have an approximate data spacing of 50 ft, whereas the data spacing is > 200 ft in the sparser regions. Therefore, a C-parameter of 250 ft with error of $\approx$4% is chosen (Fig. 5.11). Since this spacing is roughly the spacing of the data in the sparse areas, it is considered reasonable.

A final SDF model is generated by modifying the underlying distance function for boundary uncertainty with the C-parameter, as detailed in Section 2.2.7. This final model also uses the local anisotropy refined above. One-hundred unconditional Gaussian realizations are generated with a Gaussian variogram model and the anisotropy parameters from the case of global anisotropy inferred above. The simulated realizations are first and second-order stationary and thus do not match the local orientations inferred during iterative refinement. SGS with LVA could be considered; local orientations and magnitudes of anisotropy are available at the partition centers, and could be interpolated to the estimation grid (Lillah, 2014). The 100 unconditional realizations are truncated against the final SDF model by transforming the simulated Gaussian values to a uniform distribution between -250 and +250, corresponding to the range of the trained C-parameter on either side of the boundary. For each unconditional Gaussian

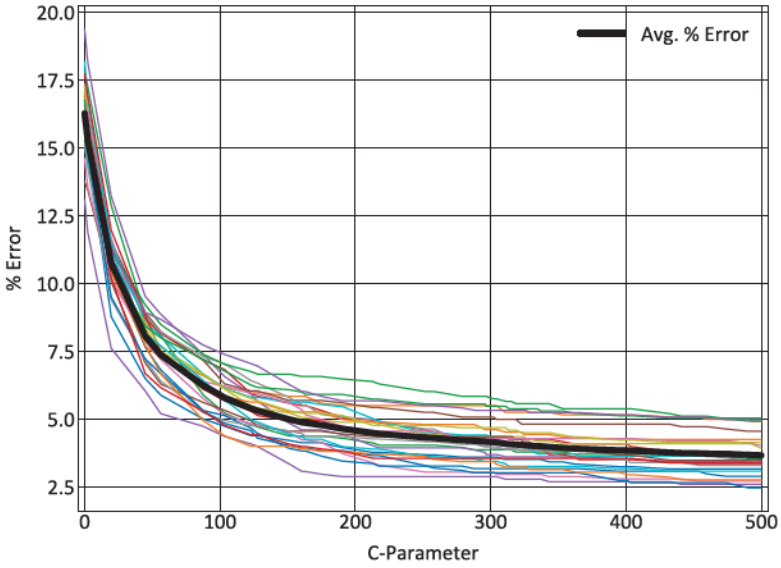


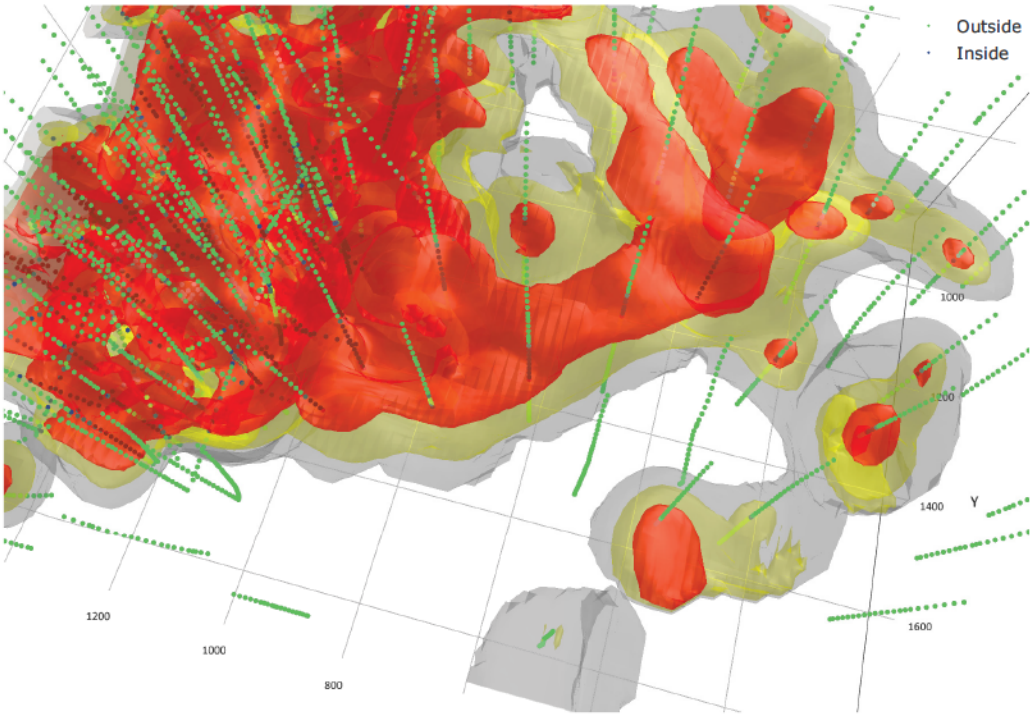
Figure 5.11: Percentage of misclassified locations as a function of C-parameter

realization the final simulated domain code is determined as (Wilde & Deutsch, 2012):

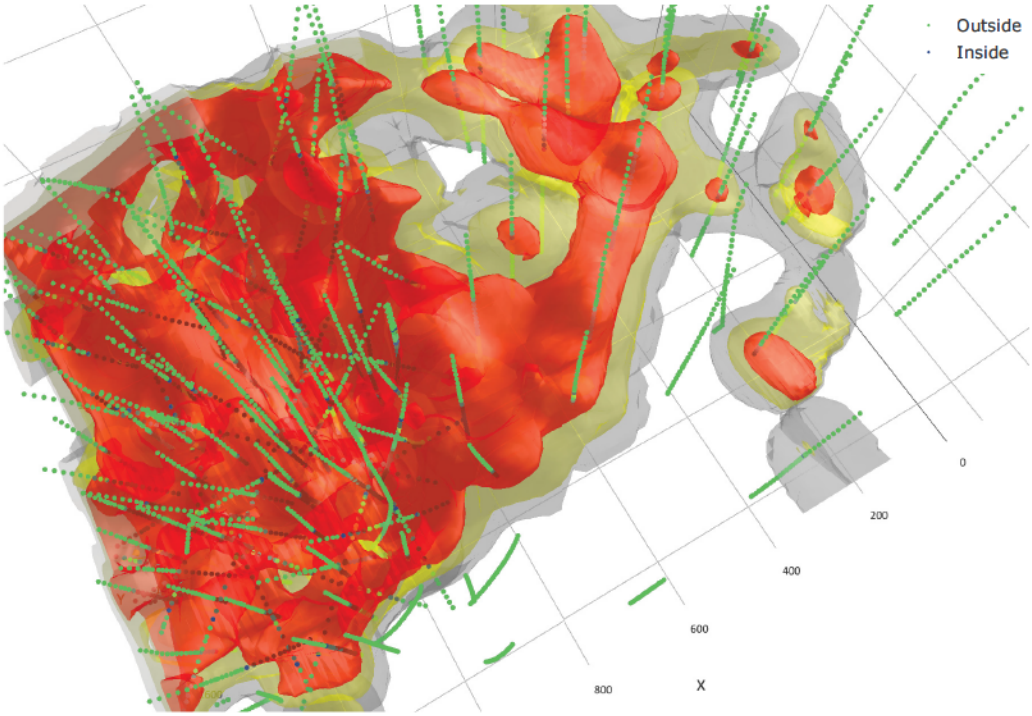
$$i(\mathbf{u}) = \begin{cases} \text{inside} & \text{if } df^l(\mathbf{u}) > \widehat{df}(\mathbf{u}) \\ \text{outside} & \text{if } df^l(\mathbf{u}) < \widehat{df}(\mathbf{u}) \end{cases} \tag{5.1}$$

where df^l is the distance function from unconditional simulation and $\widehat{df}$ is the modified signed distance function accounting for the C-parameter. Truncation with each realization results in 100 boundary realizations generated from the implicit model. Figure 5.12 shows the p_{10}, p_{50} and p_{90} isosurfaces generated by counting how many times each location is found inside the boundary over all realizations. The uncertainty associated with the boundaries is evident in the west portion of the domain where the spread between the red p_{90} and grey p_{10} surface is relatively large. By contrast in the east end, with dense drilling, there is less uncertainty. The distribution of global volumetric uncertainty generated by these realizations is shown in Figure 5.13. The contained volumes range from 59 M to 115 M ft^3.

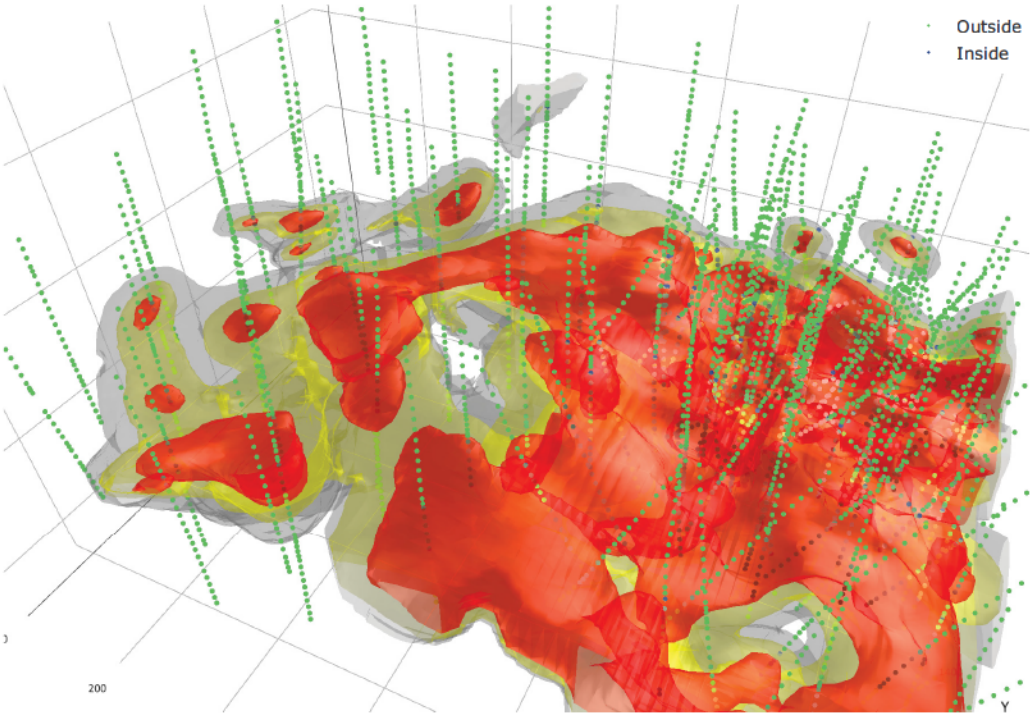
(a) View along 165 degrees dipping 65



(b) View along 210 degrees dipping 65



(c) View along 335 degrees dipping 40



(d) View along 180 degrees dipping 0

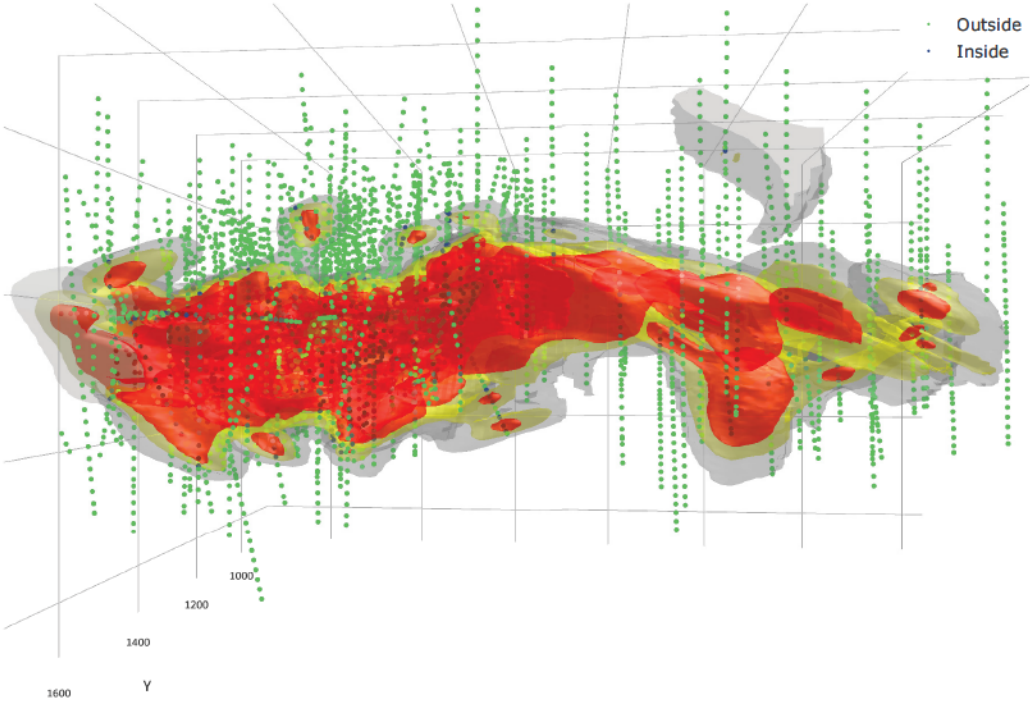


Figure 5.12: Iso-probability shells generated from implicit modeling with uncertainty. The p_{10}, p_{50} and p_{90} surfaces are grey, yellow and red, respectively.

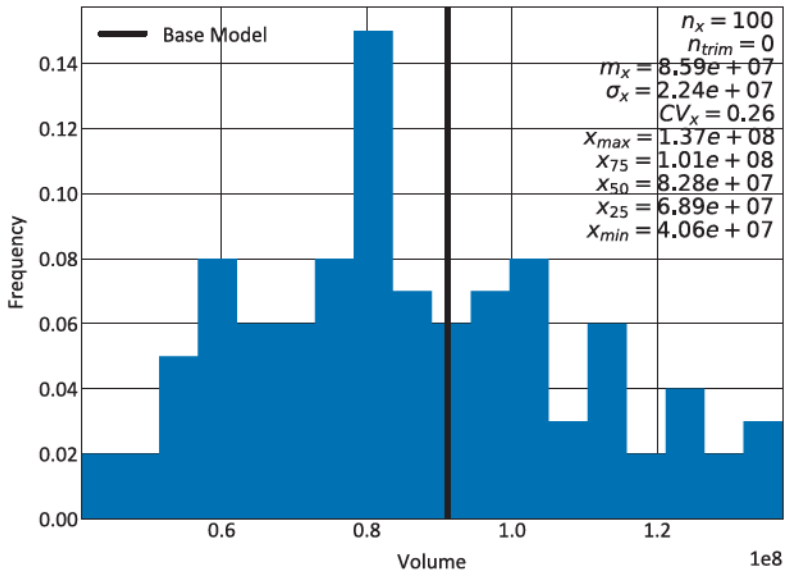


Figure 5.13: Volumetric uncertainty for boundary models generated with local anisotropy.

5.7 Review of Main Points

This chapter demonstrates the implicit modeling and iterative refinement algorithms developed in Chapter 3. Specifically:

1. The PU implicit modeling framework is effective for implicit modeling of geological domains;
2. Local anisotropy refined iteratively from implicit models reasonably captures the expected local features from the targeted domain;
3. Cross validation of implicit models is only one part of implicit model validation; equally important are the shape properties of the generated models, which should minimize bubbles and other subjectively poor features.

5.8 Conclusions

The implicit geological modeling techniques presented in this thesis effectively capture local structures given a composite dataset and targeted geological domains. The inference and execution of local anisotropy to implicit modeling can be time consuming, subjective, and error prone. This chapter demonstrates automatic iterative refinements of local features provides

a reasonable structural interpretation of a domain. Furthermore, the best models in terms of subjective shape properties does not correspond to the best model in terms of prediction errors. The relatively small increase in errors should be carefully considered when implementing local features to implicit models.

Chapter 6

Case Study: The Effect of Improved Stationary Decisions on Geostatistical Predictions

6.1 Introduction

Assessing local and global uncertainty in geostatistical models is critical for decision making and subsequent process optimization. Uncertainty in the input parameters represents an important source that must be quantified and transferred through the geostatistical modeling workflow. Assessing and incorporating known uncertainties to geostatistical models is well studied (Khan & Deutsch, 2016). However, missing from the current state-of-the-art are techniques to quantify the uncertainty associated with partitioning the dataset into stationary domains amenable to statistical modeling. Instead, uncertainies are more commonly characterized in the presence of a single decision of stationarity deemed reasonable at the start of the workflow.

This chapter implements the clustering algorithms and tools developed in Chapter 4 and verifies that the proposed methodologies improve geostatistical models. Four main contributions are developed in Chapter 4: 1) a suite of metrics that aid in the interpretation of different stationary decisions; 2) a novel clustering algorithm that reduces parameterization requirements; 3) a geostatistical consensus function for improved spatial clusters generated from clustering ensembles; and 4) a method to generate realizations of stationary domains for uncertainty characterization. Here, a K-fold validation workflow is undertaken with two datasets to assess what benefit the developed tools provide over other stationary decision making strategies.

The first dataset follows from the geological modeling in Chapter 5. This domain is referred to as porphyry in the following text, and has a single lognormally-distributed modeling variable, Cu total (CuT), from two geologically defined rock types. This dataset can be considered a simplified case of stationary domaining since only a single variable must be considered. The

second oilsands dataset contains 3 variables and 10 facies, with a mix of Gaussian-like and lognormal distributions. However, the multivariate relationships are non-Gaussian.

6.2 Experimental Methodology

The validation methodology employed here consists of a 5-fold cross validation that quantifies prediction error between geostatistical models generated under different decisions of stationarity.

6.2.1 Types of Decisions of Stationarity

For each dataset, a total of 7 decisions of stationarity (hereafter: domains, categories or clusters) are assessed, including: a single domain (no dataset partitioning); geological categories; multivariate clusters; spatial clusters; improved spatial clusters; realizations of spatial clusters; and realizations of random clusters. The single domain ('no domains') tests the assumption that there is only a single stationary population, and subdivision of the dataset is detrimental to the geostatistical analysis. Spatial clusters are generated using the DS clustering algorithm developed in Section 4.3. Improved spatial clusters are generated from the clustering ensemble using the geostatistical consensus function developed in Section 4.4. Finally, the cluster realizations are taken from the sub-ensemble selected to generate the improved spatial clusters. Together, these three methods comprise the novel contributions of Chapter 4 for generating stationary domains and assessing the associated uncertainty.

6.2.2 Validation Methodology

Figure 6.1 left shows a conventional decorrelation-based geostatistical workflow for a constant set of categories. The modification to have different categories for each realization is straight forward (right; Fig. 6.1). Sequential indicator simulation (SIS) and SGS (Deutsch, 2006; Manchuk & Deutsch, 2012) are used for simulating the categorical and continuous variables, respectively. For each geostatistical realization and at each unsampled location, SIS is used to simulate the category given the categories found at nearby sample locations and previously simulated nodes. Within each simulated categorical domain, SGS is used to simulate the continuous variables

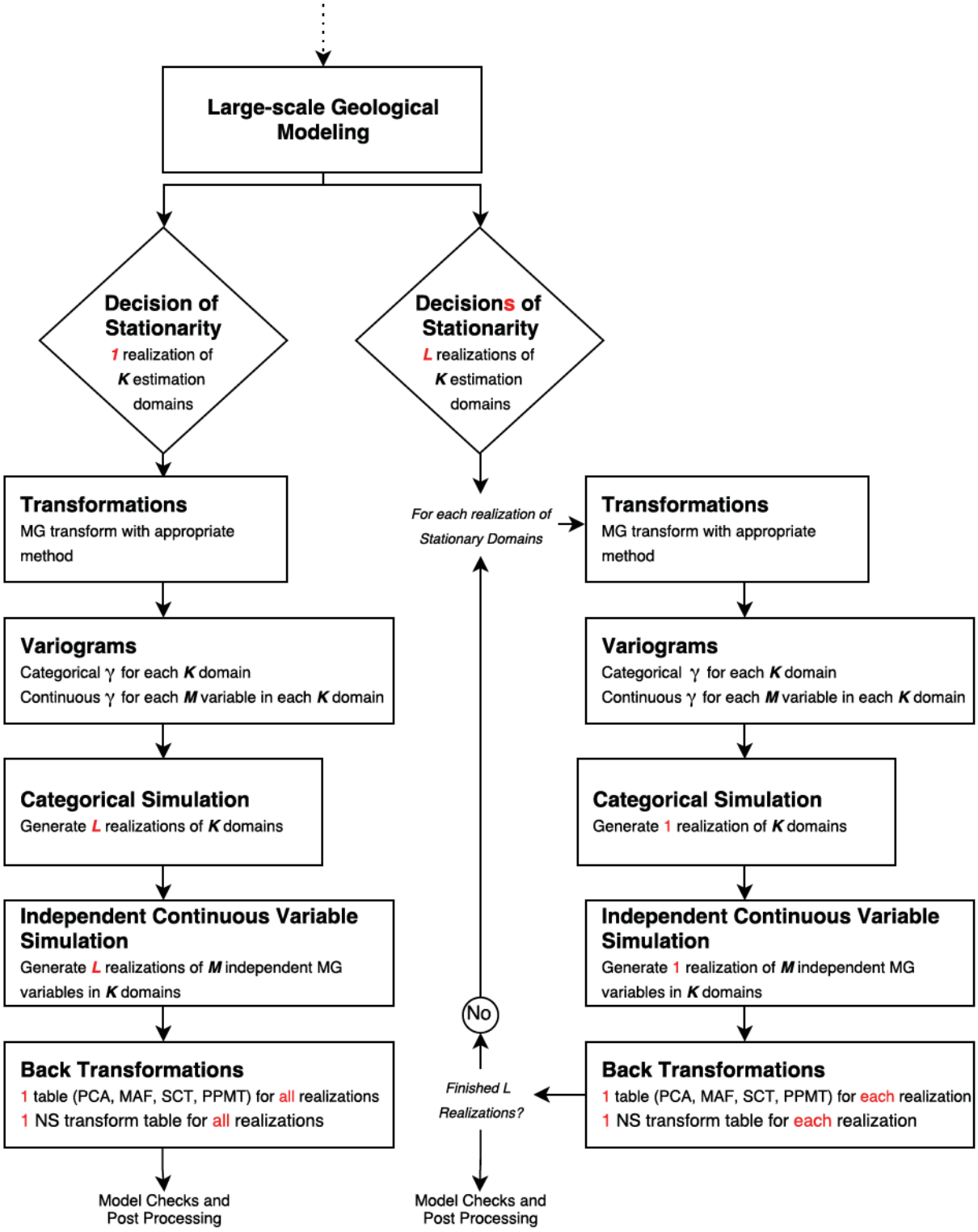


Figure 6.1: Left: Conceptual geostatistical workflow considering a single set of categories. Right: Proposed modified geostatistical workflow for realizations of categories, differences are highlighted in red.

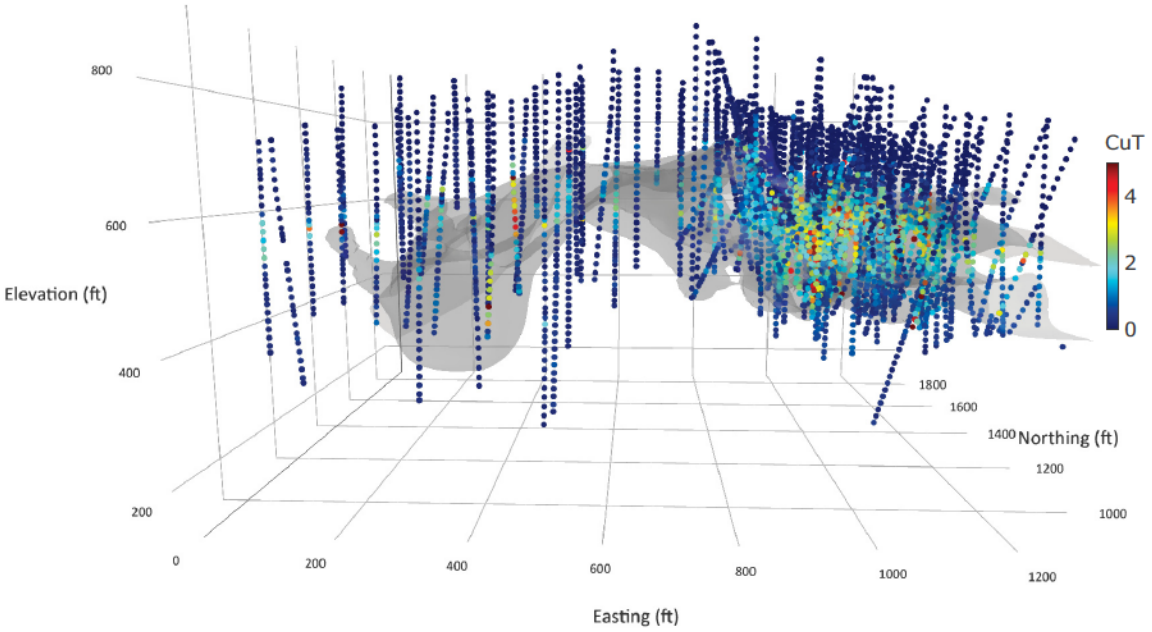


Figure 6.2: Location map of the porphyry deposit colored by CuT.

conditional to the related samples and simulated locations from the same category. For the univariate large-porphyry dataset, a normal-score transformation is performed independently in each category, and independently for each realization of categories in the modified workflow. Similarly, for the oilsands datasets, a linear decorrelation transformation (e.g., sphere-R; Barnett & Deutsch, 2015) is undertaken to permit independent simulation of the Gaussian-transformed variables, and is applied independently for each realization of categories in the modified workflow.

6.3 Porphyry Dataset

The setting of the porphyry is detailed in Chapter 5. Sample locations, colored by CuT, within the context of the geological domain boundaries are shown in Figure 6.2. The dataset for geostatistical modeling comes from within the geological boundaries and consists of 1299, 10 ft composite assays from 139 drill holes.

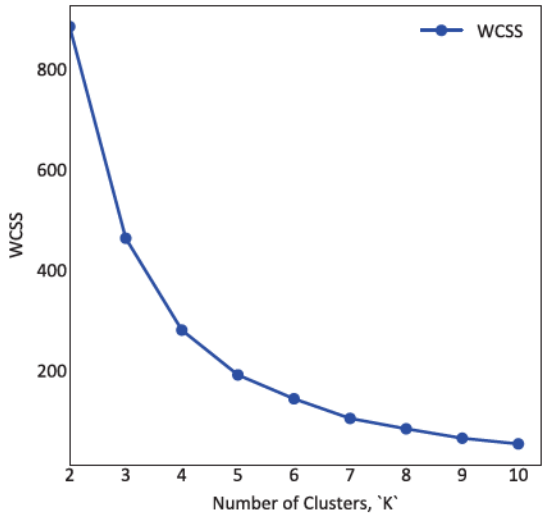


Figure 6.3: Elbow plot for CuT.

6.3.1 Defining Categories

The first step is to choose how many categories to model. An elbow plot measures the compactness of a configuration in multivariate space for different K by calculating the total WCSS of the configuration for each K. A preferred K can be selected by inspecting the plot of K vs WCSS, and looking for inflection points. The elbow plot for this dataset is shown in Figure 6.3 and does not show a clear preferred K. Since there are two geological categories defined in the geological log, $K = 2$ is used for all sets of categories defined in this section.

The univariate properties for all single-realization categories are shown in Figure 6.4, and selected slices through the composite dataset, colored by assigned category, are shown in Figure 6.5. The clustering metrics developed in Chapter 4, calculated for all generated clusterings, are shown in Figure 6.6. For this dataset the WCSS and spatial entropy are selected as the multivariate and spatial metrics, respectively.

6.3.1.1 Geological Categories

Geological categories are defined in the drill logs by site practitioners. The univariate properties of each category are very similar and nearly completely overlap (Fig. 6.4a). However, the spatial delineation of the categories is the best out of all techniques considered here. This combined score is reflected in Figure 6.6 (green) where the geological categories plot with a high

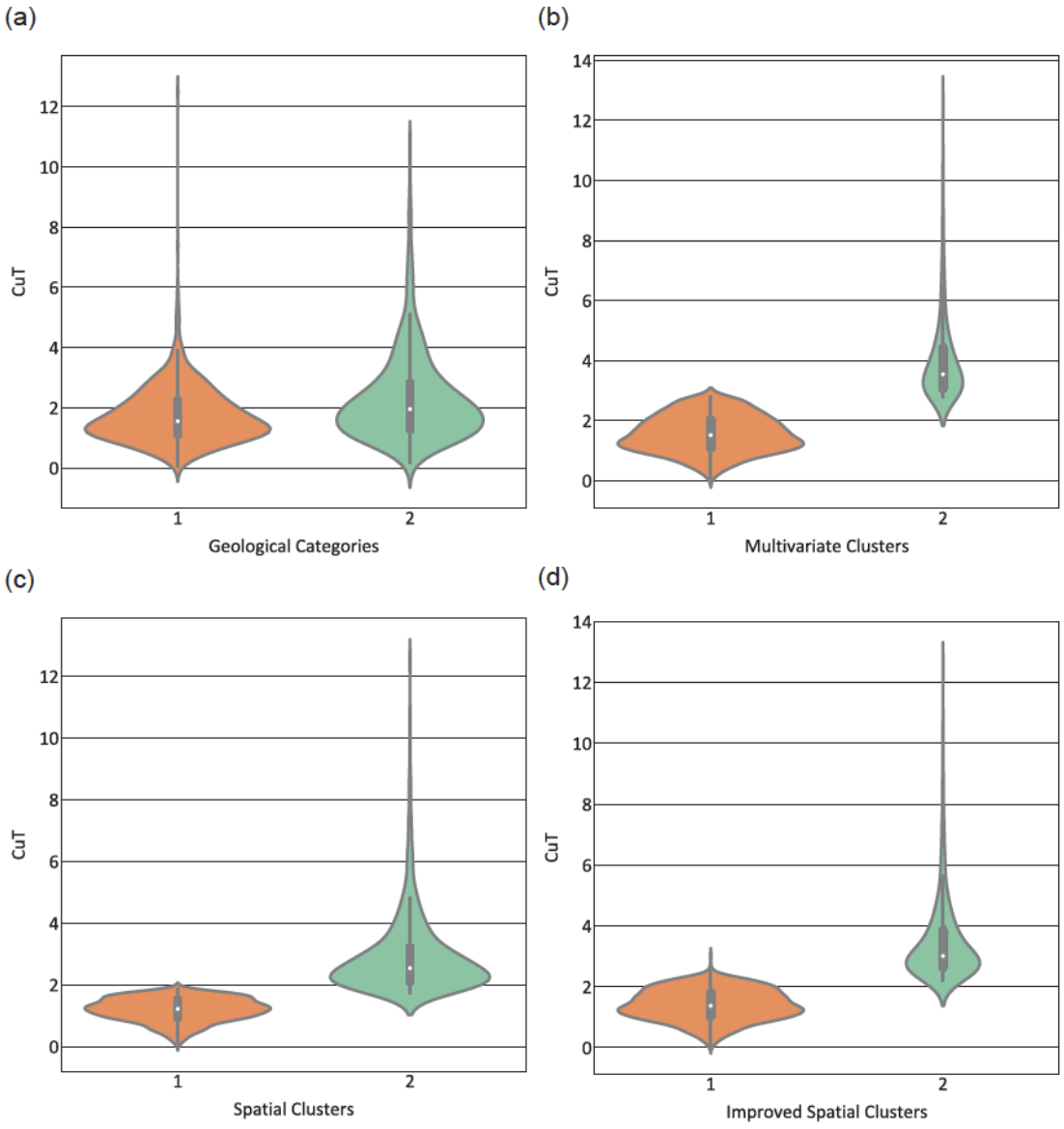


Figure 6.4: KDE (shaded area) and box plots (vertical line) of CuT grades from each defined category from the porphyry dataset. (a) Geological categories, (b) multivariate clusters, (c) spatial clusters, and (d) improved spatial clusters.

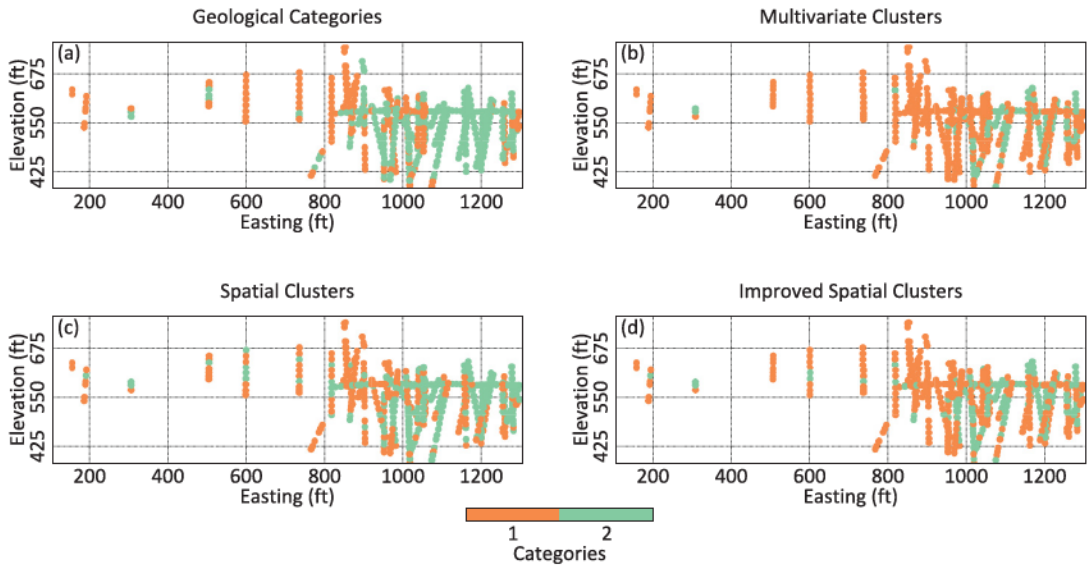


Figure 6.5: E-W slice of the sample locations from the porphyry dataset, colored by category. (a) Geological categories, (b) multivariate clusters, (c) spatial clusters, and (d) improved spatial clusters.

multivariate score and a low spatial score. Recall that, in this metric space, minimizing both metrics is the goal, and a tradeoff between multivariate and spatial scores is expected. The categorical variograms calculated for this set (Fig. 6.7a) support the spatial score, since the spatial continuity is the highest out of all sets. Variograms for CuT within each geological category are shown in Figure 6.8a. The spatial continuity within each category is distinct, further supporting this delineation.

6.3.1.2 Multivariate Clusters

Multivariate clusters are generated with K-means clustering using 100-random initializations. This set differs from the rest since only the continuous variables, and no spatial information, is considered to form the clusters (Fig. 6.4b & 6.5b). Clusters generated from this dataset are not multivariate since only CuT is used; in this sense these categories are identical to grade domains defined on CuT cutoffs. As expected, the resulting categories best delineate the univariate properties of the domain (Fig. 6.4a and red in Fig. 6.6). However, the spatial properties are comparatively poor, especially with respect to the geological domains (Fig. 6.5a vs. b), which is also reflected in the categorical variograms (Fig. 6.7a vs. b). Variograms of CuT within each category show continuity similar to the geological categories, but with slightly shorter ranges

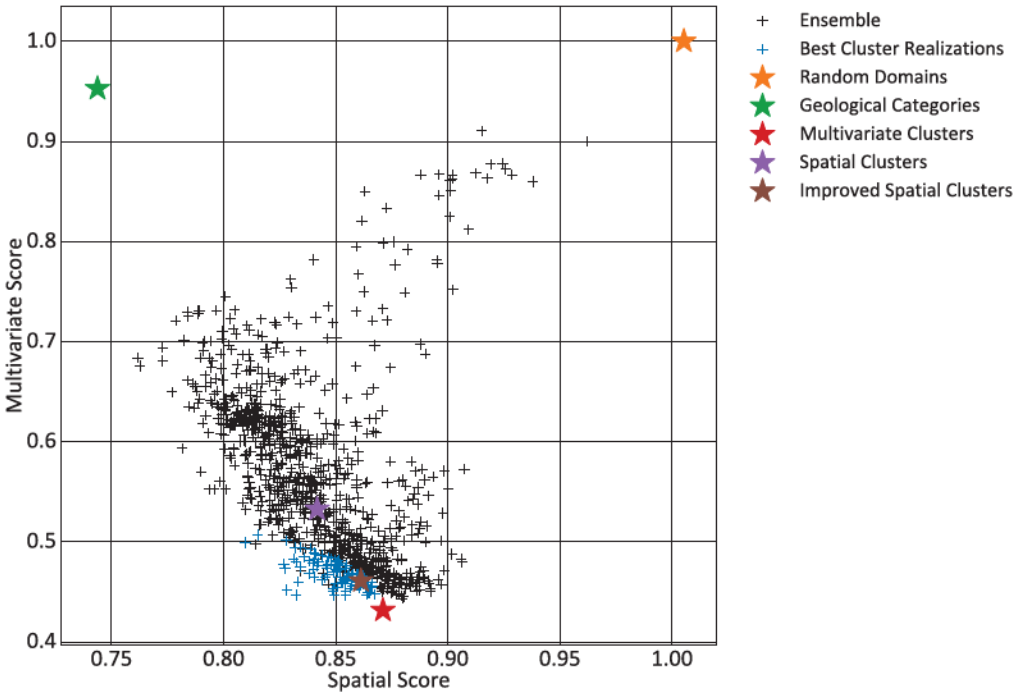


Figure 6.6: Standardized spatial-multivariate metrics calculated for all categories for the porphyry dataset.

(Fig. 6.8b).

6.3.1.3 Spatial Clusters

Spatial clusters are generated using the DS clustering methodology developed in Chapter 4 (Fig. 6.4c & 6.5c). The DS algorithm generates 1000 clusterings, using 35 nearest-neighbors, and merging 3 neighbors per location in the spatial merging stage. Clusterings are rejected from the ensemble if any one of the categories contains less than 25% of the composite database. The search anisotropy is omnidirectional in the horizontal plane with a 2:1 horizontal:vertical anisotropy ratio to preference the local spatial search for samples between, rather than along, drill holes. The spatial and multivariate scores for all clusterings in the ensemble are shown in black in Figure 6.6, and show the expected inverse relationship between multivariate and spatial delineation.

Spatial clusters are generated from the consensus of all clusterings in this ensemble using the pairwise-similarity matrix consensus function; the metric score is shown as the purple star in Figure 6.6. This consensus clustering has properties seemingly representative of the entire

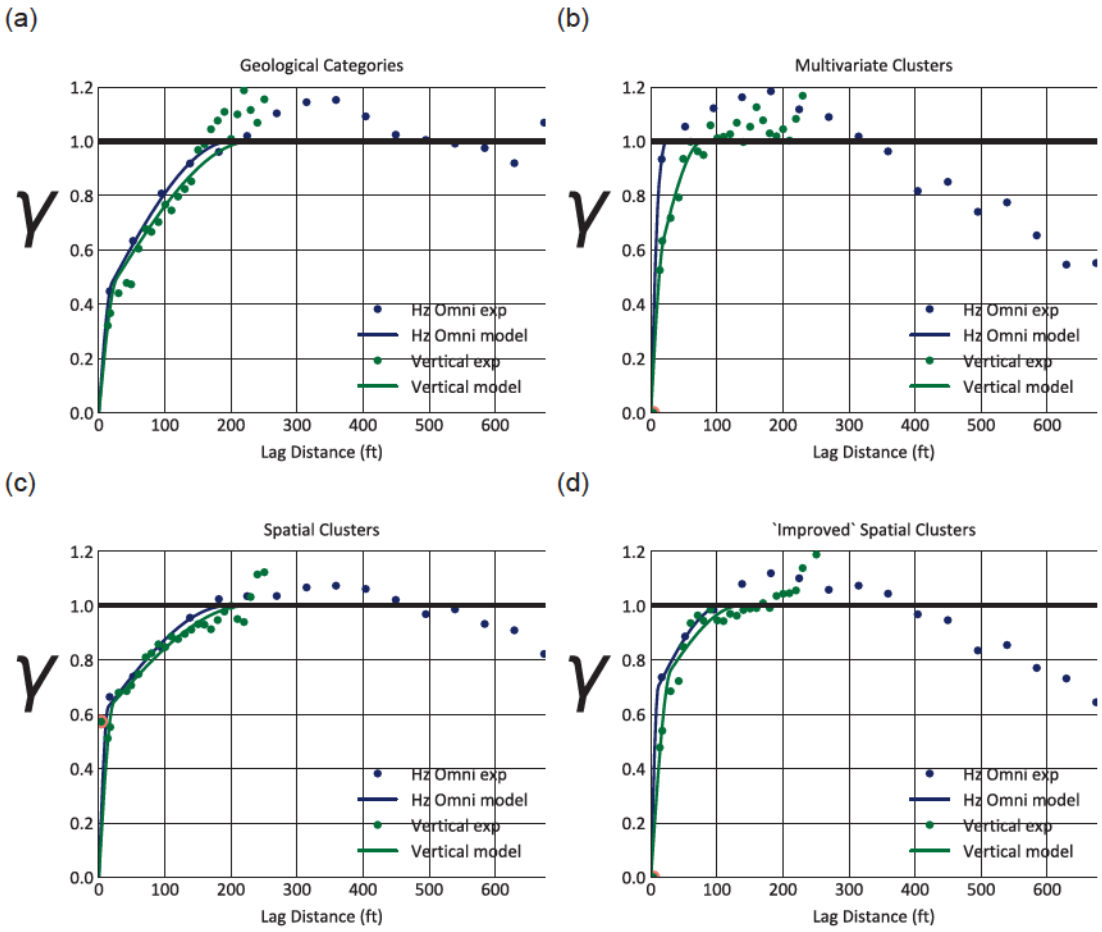


Figure 6.7: Experimental and model categorical variograms for each set of categories. (a) geological categories, (b) multivariate clusters, (c) spatial clusters, and (d) improved spatial clusters.

ensemble. The spatial properties are improved with a corresponding degradation in multivariate properties when compared to the multivariate clusters. In the spatial domain these clusters do appear to be more continuous (Fig. 6.5c), which is also supported by the categorical variograms (Fig. 6.7c vs. b). Variograms of CuT within each cluster are similar to those from the multivariate clusters (Fig. 6.8c).

6.3.1.4 Improved Spatial Clusters

The proposed improvement to spatial clustering chooses a subset of the ensemble that has the best spatial and multivariate properties. These are selected by finding 100 clusterings that have the minimum multivariate score for a given spatial score (blue crosses; Fig. 6.6). Improved spatial clusters are generated from this subset using the same pairwise-similarity matrix consen-

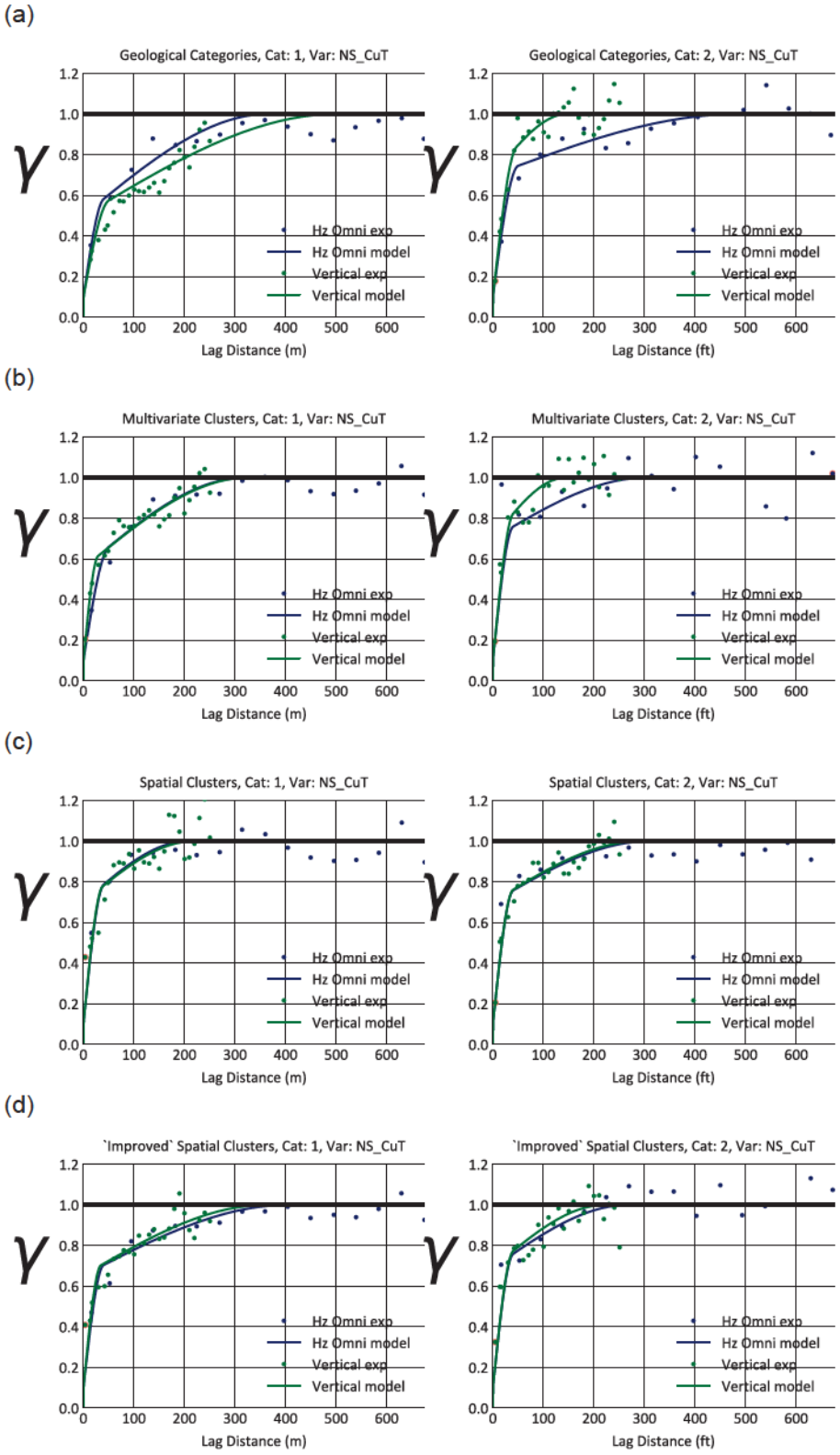


Figure 6.8: Experimental and model variograms of normal scored NS_CuT within each category, for each set of categories.

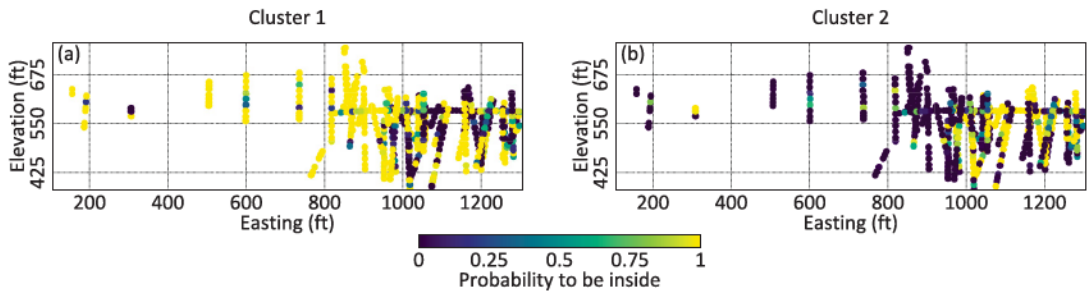


Figure 6.9: E-W slice of the porphyry domain, colored by probability to be part of each category.

sus function, as above (Fig. 6.4d & 6.5d). The resulting categories have improved multivariate properties, with a corresponding decrease in spatial properties when compared to the spatial clusters generated from the total ensemble (brown vs. purple; Fig. 6.6). The categorical variograms also support this decrease in spatial continuity (Fig. 6.7d), whereas variograms of CuT within each category show slightly longer-range continuity when compared to the clusters from the full ensemble (Fig. 6.8d).

6.3.1.5 Cluster Realizations

The suite of improved clusterings selected above each define the categories used for a single realization in the simulation workflow (blue crosses; Fig. 6.6). Each clustering realization is recoded to have the maximum similarity to the improved clusters generated above. The recoded cluster realizations compared against the improved clustering from above gives the likelihood of each location to belong to each category over all realizations (Fig. 6.9). In this case, categorical and continuous variogram models calculated from the improved clustering are used for all realizations, however, automatic fitting of variograms from each category realization could be considered.

The proportion of samples allocated to each category over all realizations is shown in Figure 6.10b. This distribution of uncertain proportions is not typically available in conventional stationary domaining workflows, and instead must be generated through bootstrap analysis and trend modeling (Hadavand & Deutsch, 2017). A set of KDE of CuT contained in each category, colored by category type, are shown in Figure 6.10b. The clustering realizations delineate consistent low and high grade populations in terms of their univariate and spatial properties.

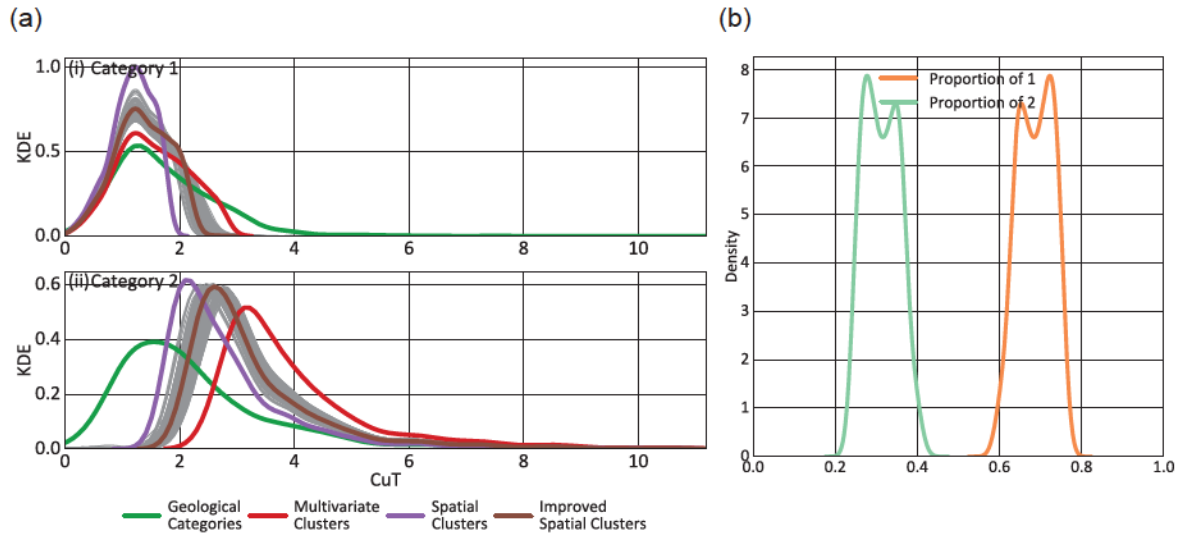


Figure 6.10: (a) KDE for CuT for each category (i) 1 and (ii) 2, for each set of categories. (b) Distribution of proportions of each stationary domain over all realizations.

6.3.1.6 Control Groups

Two control groups are implemented for comparison. The first set considers all samples to be from a single stationary population. This may be a reasonable decision in this domain since the single variable follows a relatively stable lognormal distribution, the controls to mineralization are relatively large-scale and diffuse, and the large-scale geological partitioning mostly separates the low grade from the high grade populations. The second control group considers the stationary domains to be random, and generates a different random set of stationary domains for each realization. According to the developed metrics, considering random categories results in the worst spatial and multivariate delineation out of all clusterings (orange; Fig. 6.6).

6.3.2 K-Fold Results

The geostatistical analysis follows the steps outlined in the left and right of Figure 6.1, for the static and uncertain categories, respectively. The sample locations colored by fold are shown in Figure 6.11. The full SIS and SGS simulation workflow is repeated independently for each fold, using the training set to build geostatistical models, and the test dataset to evaluate how well the true values are predicted at the removed locations.

The performance of each set of categories is assessed by comparing the mean of simulated

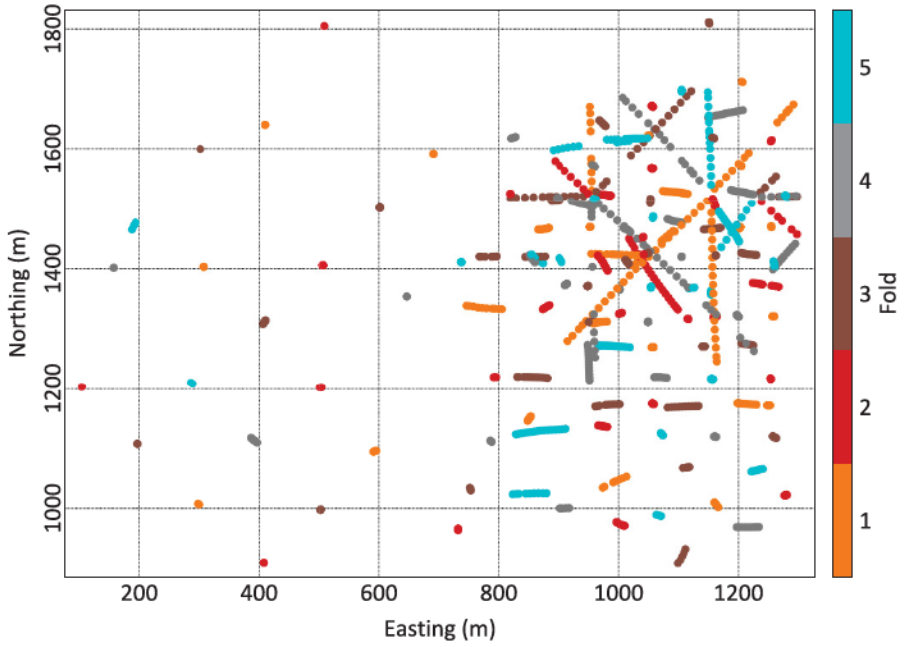


Figure 6.11: Sample locations colored by fold for the large-porphyry dataset.

Table 6.1: CuT K-fold error statistics by category, porphyry dataset.

	Covariance	Correlation	RMSE
Category Realizations	0.33	0.42	1.13
Random Domain	0.33	0.42	1.14
Geological Categories	0.33	0.44	1.11
Multivariate Clusters	0.16	0.37	1.14
Spatial Clusters	0.29	0.42	1.11
Improved Spatial Clusters	0.31	**0.45**	**1.10**
No Domains	**0.37**	**0.45**	1.11

values with the true values at removed sample locations. This E-type estimate is the best estimate given the modeling parameters for each set of categories tested in this workflow (Fig. 6.12 and Tab. 6.1). The goal is to reproduce the true values, minimizing the RMSE and maximizing the correlation and covariance between the estimate and the truth.

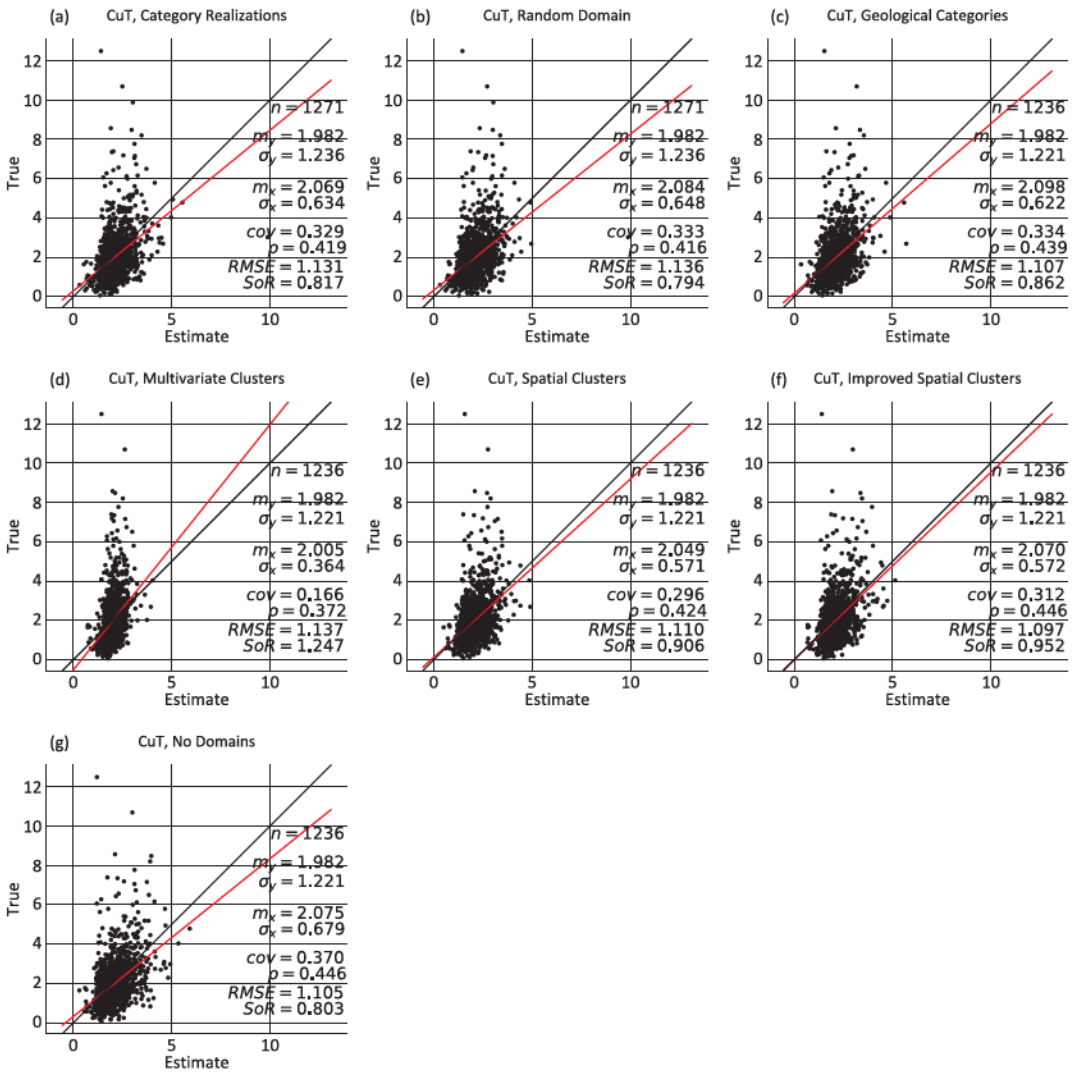


Figure 6.12: Cross plot between the E-type estimate and the true values from the porphyry dataset.

The best set of estimates are obtained from the improved spatial clusters since they give the lowest RMSE and highest correlation (Fig. 6.12f). There is an incremental improvement between multivariate clusters, spatial clusters and improved spatial clusters in terms of correlation and RMSE, confirming that spatial information incorporated in unsupervised learning algorithms for stationary domaining is beneficial for geostatistical modeling.

Another important measure is how the predicted distributions of uncertainty match the data across validated locations; this is assessed using the accuracy plots of Deutsch (1996) (Fig. 6.13). This plot compares the expected and observed number of times a true value falls within a given probability interval for the local distributions of uncertainty. The goal is to gener-

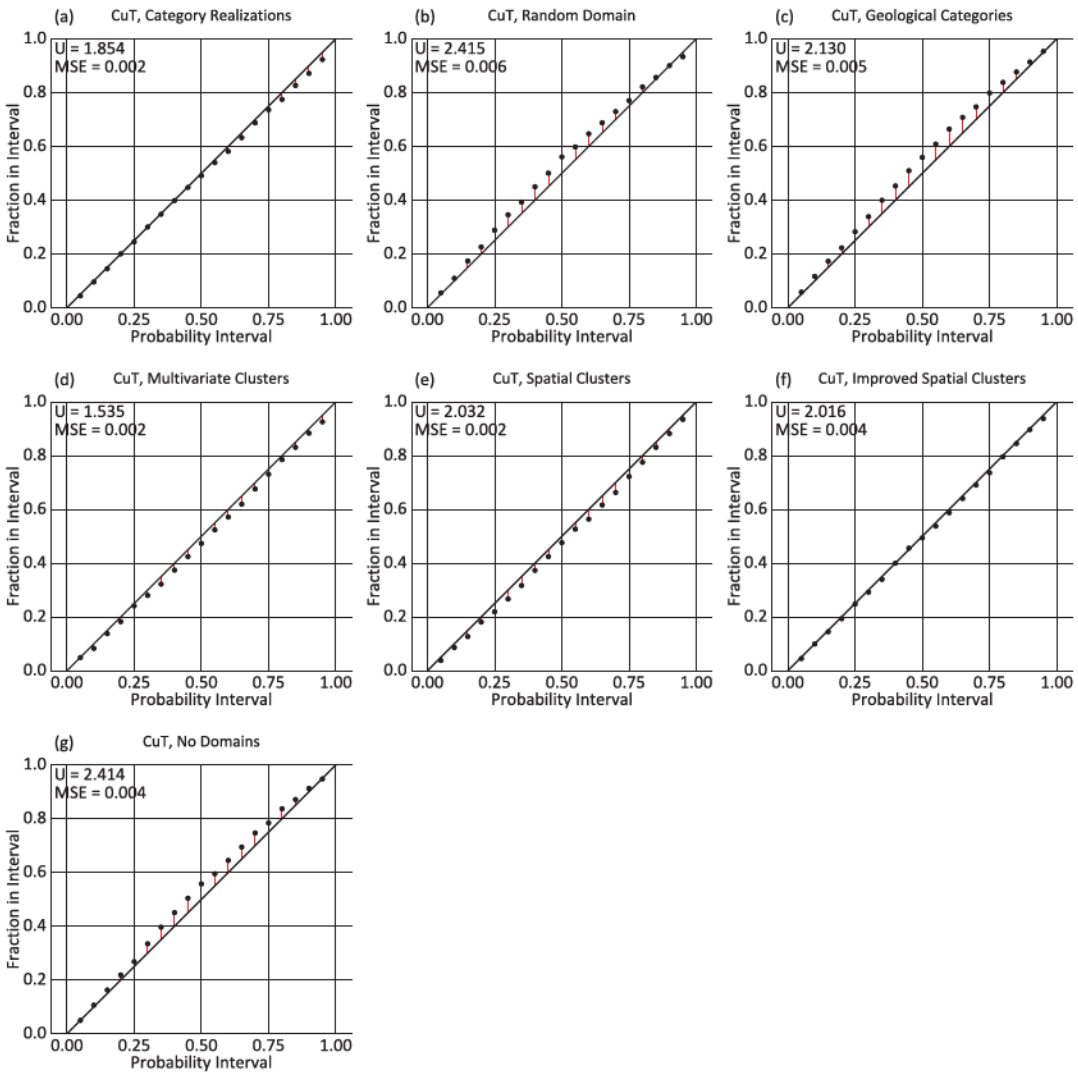


Figure 6.13: Accuracy plots for CuT from the porphyry dataset, for each set of categories.

ate local distributions where the true value always falls within the expected probability interval over all locations in the domain, such a case would result in points plotting along the 1:1 line. Points falling below the 1:1 line indicate that local distributions of uncertainty are too narrow, and points above the 1:1 line indicate the local distributions that are too wide. Either case could be acceptable, but generally, narrower distributions of uncertainty that are equally accurate are preferred (Deutsch, 1996).

For the tested sets, the improved spatial clustering methodology generates the best local distributions of uncertainty (Fig. 6.13a, and f), since points plot closest to the 1:1 line over all probability intervals. Multivariate clusters and regular spatial clusters generate local distribu-

tions that are slightly too narrow (below 1:1), whereas the control groups and the geological categories generate local distributions that are too wide (above 1:1). All methods generate local distributions of uncertainty that could be considered reasonable.

6.3.3 Global Realization Checks

A suite of final realizations are generated with the entire modeling dataset to assess how well realizations generated with different sets of categories reproduce the global statistics.

The first check is to reproduce the representative histogram (Fig. 6.14). All sets of categories result in reasonable histogram reproduction, except for the multivariate clusters which fail to reproduce the medium to high grade values. Here, the control groups and the geological categories generate realizations that have a slightly higher average grade, which could be attributed to uncontrolled extrapolation of high-grade material in the sparsely sampled areas.

Next, the spatial continuity between locations should be reproduced. Variograms calculated on back transformed realizations show reasonable reproduction of the CuT experimental variogram calculated from all data in the dataset (grey vs. red; Fig. 6.15). All methods fail to reproduce the long-range horizontal structure in the global variogram model, though the vertical direction is well reproduced by all methods except the random and multivariate cluster categories. The poor horizontal variogram reproduction could be attributed to the domain clipping by the bounding implicit model.

6.3.4 Discussion

The control groups perform very well in this domain in terms of reproduction of global statistics, and the local error in reproducing the data. However, local distributions overestimate the local uncertainty across the domain. The DS spatial clustering and improved ensemble clustering methodology proposed in Chapter 4 generate categories that are measurably better as stationary domains in terms of local accuracy, representative local distributions and the reproduction of global statistics.

Overall, in terms of reproducing the data, generating representative distributions of uncertainty and reproducing the input statistics, the spatial clustering methodologies match or outper-

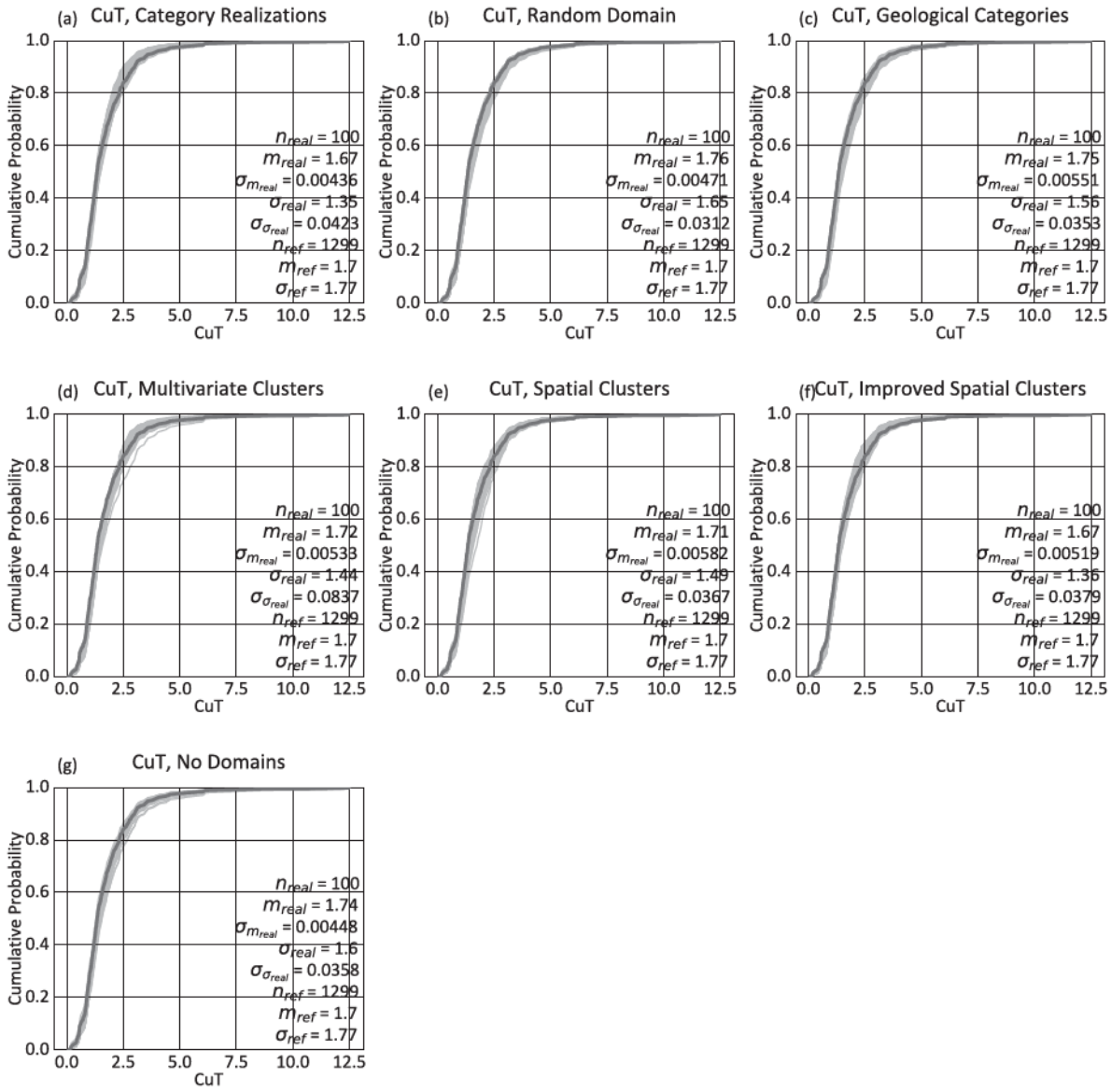


Figure 6.14: Histogram reproduction plots for the porphyry dataset.

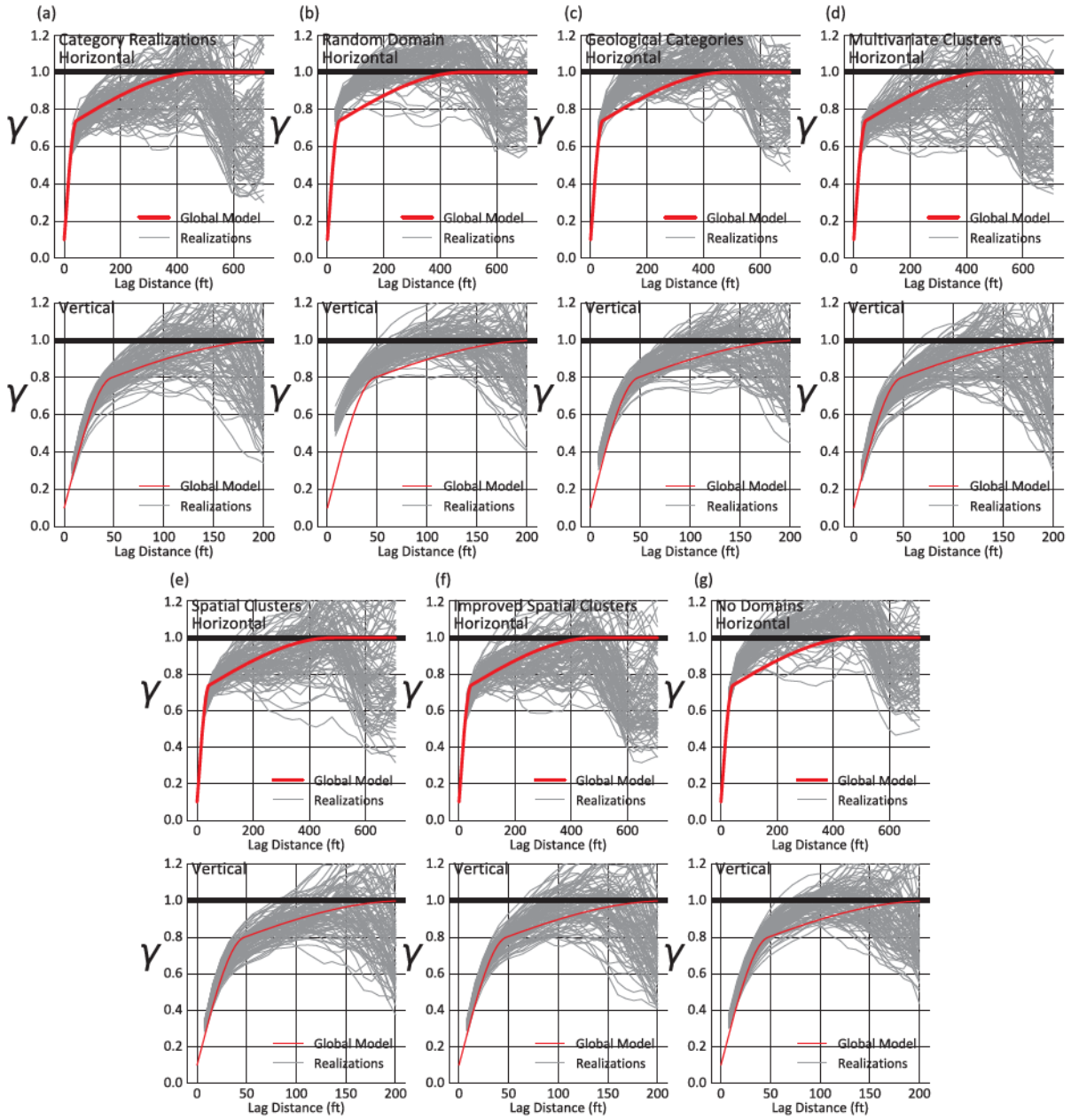


Figure 6.15: Variogram reproduction for realizations generated by each method. Horizontal reproduction shown in the top and vertical shown below for each set of stationary domains (a-g).

form traditional stationary decision making methodologies in this domain. This domain represents a simple case since the spatial properties of only a single variable must be considered and the geological delineation has already imparted some consistency to the delineated populations. Compared to the categories defined in the geological logs, the spatial clusters or the improved spatial clusters represents an improvement in three ways: 1) a reduction in prediction errors; 2) a tractable and reproducible domaining workflow that can easily integrate new information; and 3) assessment of the uncertainty associated with partitioning the domain into stationary sets.

6.4 Oilsands Dataset

The oilsands dataset consists of 3196 3 m composites from 150 wells, sampling 3 variables and 9 unique facies (Fig. 6.16a). Inspection of the normal-scored multivariate distribution in Cartesian and multivariate space suggests that 3 domains will reasonably capture the properties of this domain. However, an elbow plot generated for different K (Fig. 6.17) does not show a conclusive best K, though 3 or 4 could be argued. Vertical zonation is an important feature of this dataset; a laterally continuous bitumen-rich and fines-poor layer forms the center of the domain with a bitumen-poor unit above and a chlorides-rich unit below (Fig. 6.16b). One method to account for this vertical zonation is to partition the domain.

6.4.1 Defining Categories

The same set of 7 methods to generate modeling categories are applied to this dataset. All methods requiring a spatial search use a 50:1 horizontal:vertical anisotropy ratio, which is warranted owing to the long-range horizontal continuity of the underlying variables and the large differences in sample spacing down-well versus between wells. The spatial and multivariate metrics calculated for all generated categories are shown in Figure 6.19. For this dataset the WCSS and spatial entropy are selected as the multivariate and spatial metrics, respectively. KDEs showing the distribution of continuous variables contained within each category, for all sets of categories, are shown in Figure 6.20.

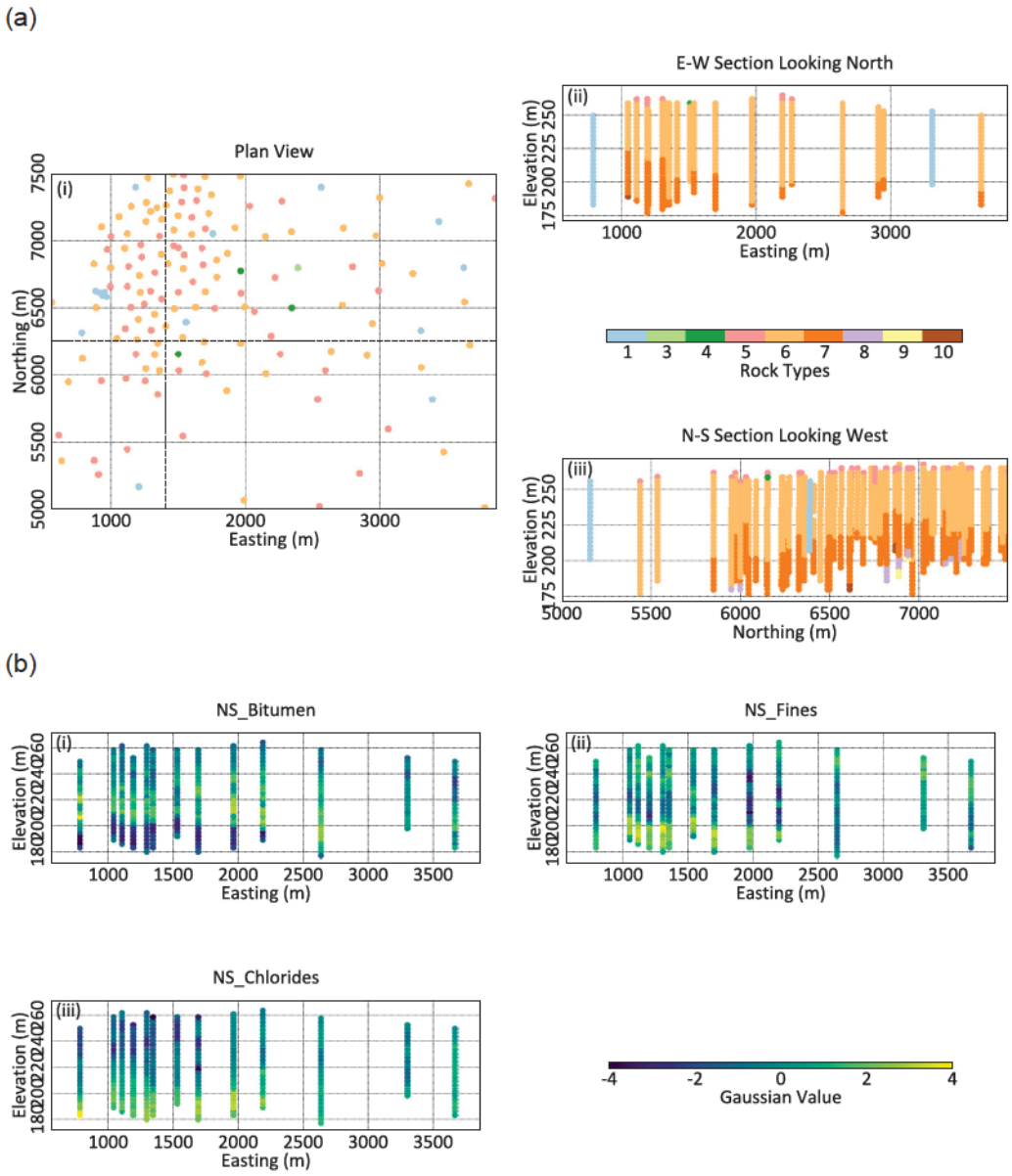


Figure 6.16: (a) Drill hole locations colored by facies. Dotted lines in (ai) show the location of the E-W and N-S slices in (aii) and (aiii), respectively, and (b) Gaussian transformed values of the oilsands dataset.

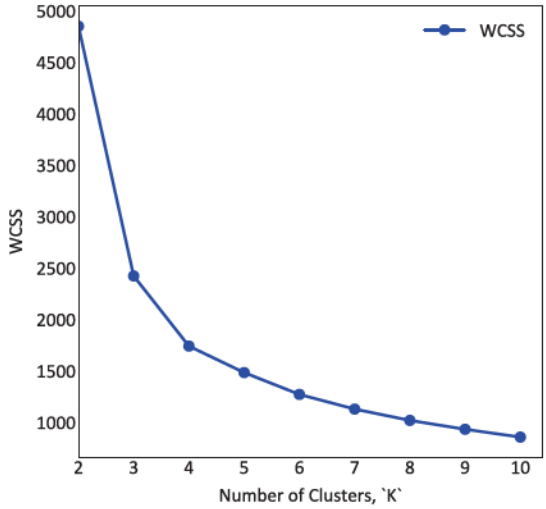


Figure 6.17: Elbow plot for Bitumen, Fines and Chlorides.

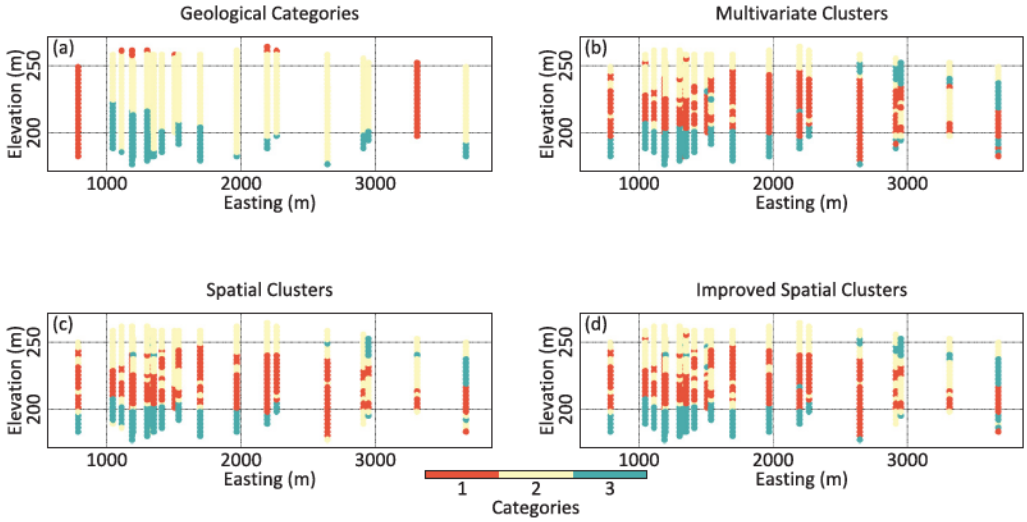


Figure 6.18: Individual sets of categories for the oilsands dataset on the E-W section looking north.

6.4.1.1 Geological Categories

Nine facies are manually merged to 3, by considering the multivariate populations and the large-scale spatial Cartesian relationships. Categories are merged if they form contiguous groups in the Cartesian domain and if the multivariate properties of the two categories are similar (Rossi & Deutsch, 2014). This multivariate similarity could be measured with the cluster metrics proposed in Chapter 4. The resulting merged-geology categories are shown in Figure 6.18a, hereafter referred to as geological categories.

Considering the 3 target variables in Figure 6.16b and 6.20, the delineation of the bitumen-

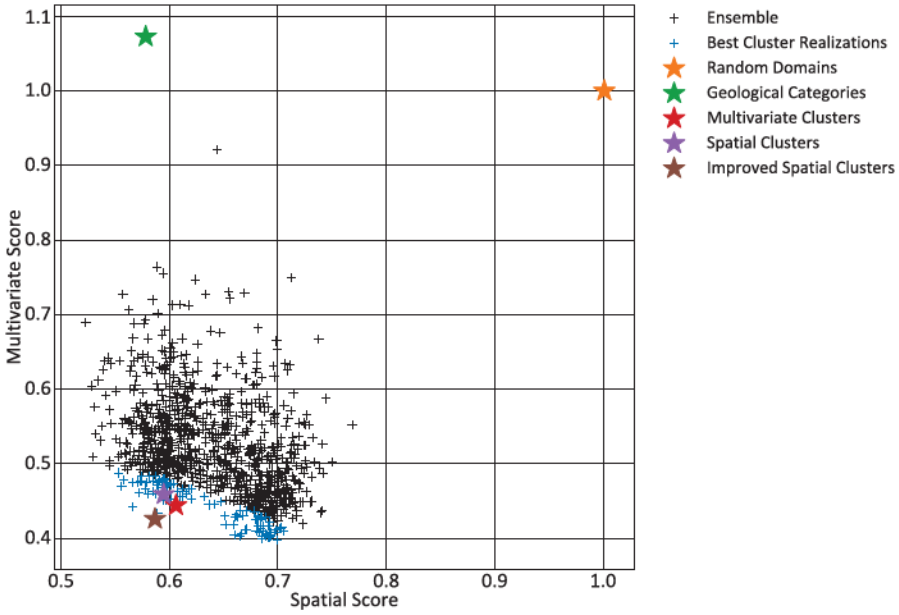


Figure 6.19: Standardized spatial-multivariate metrics calculated for all categories for the oilsands dataset.

rich zone is relatively poor with this method. However, this is a reasonable delineation since there is a clear vertical trend in bitumen content and strong horizontal anisotropy. These categories are the most spatially contiguous out of all considered here, which is reflected by the spatial-multivariate metric value calculated for this set (green; Fig. 6.19).

6.4.1.2 Multivariate Clusters

Categories generated considering only the multivariate properties are shown in Figure 6.18b and red in Figure 6.19 and 6.20. Again, K-means clustering with 100-random initializations is used. The resulting categories have relatively good spatial delineation given that no spatial information is considered in the clustering algorithm. This can be attributed to the high horizontal spatial continuity of the underlying variables and the large-scale zones of consistent variable properties. These categories clearly account for the bitumen-rich fines-poor central layer (red; Fig. 6.20aii & bii).

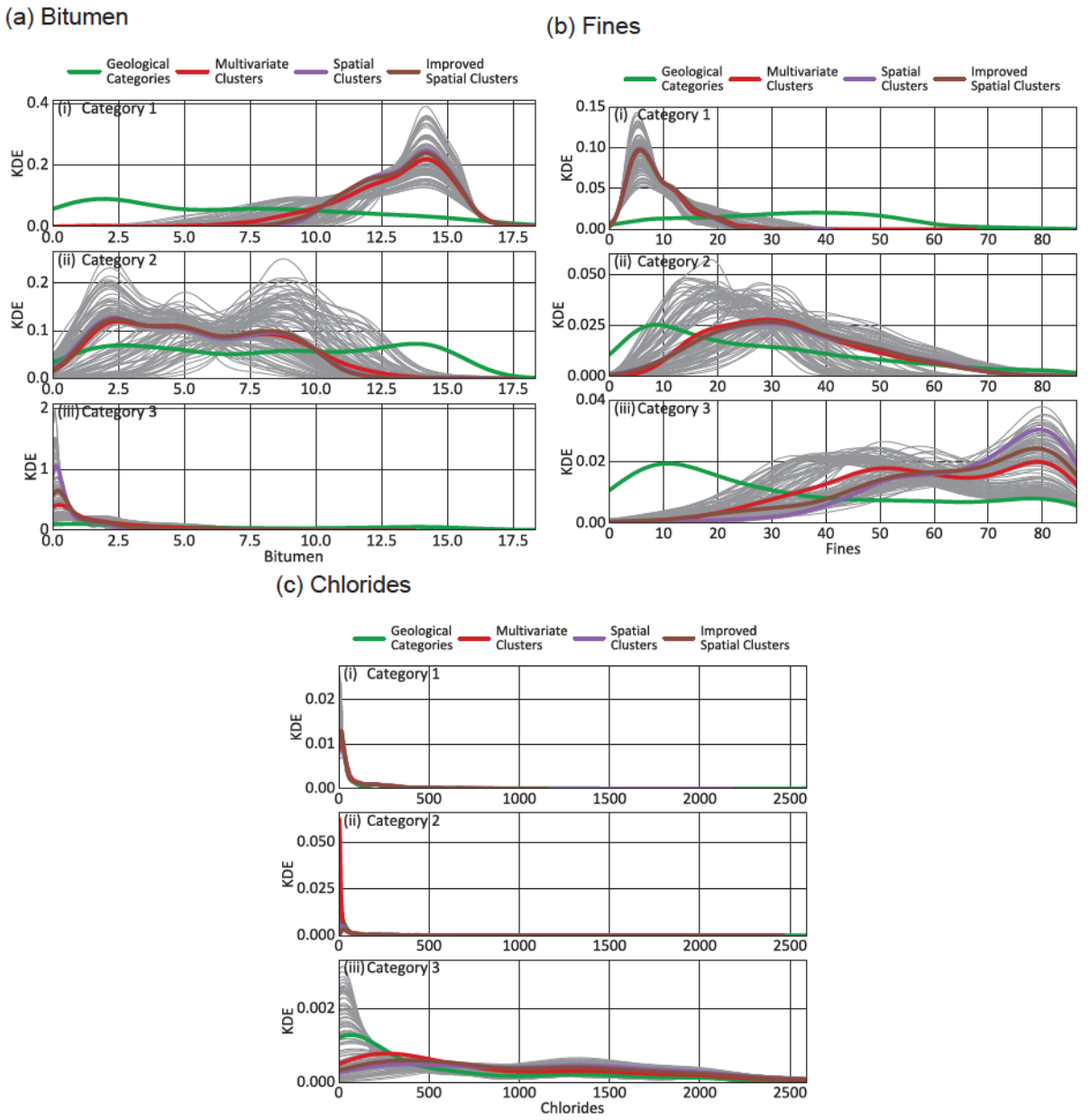


Figure 6.20: KDE for (a) Bitumen, (b) Fines, and (c) Chlorides, for each category (i) 1, (ii) 2, and (iii) 3, for all sets of categories investigated in this study.

6.4.1.3 Spatial Clusters

Spatial clusters are generated using the DS clustering algorithm, with 1000 clustering realizations, 55 spatial neighbors with 3 locations merged during during the spatial search phase. An additional constraint given to the clustering algorithm is to reject clusterings from the ensemble if any one of the categories contains less than 10% of the composite database, to ensure that a suitable number of samples are found within each delineated domain for geostatistical modeling. The ensemble of clusterings (black; Fig. 6.19) is processed with the pairwise-similarity consensus function; the result is shown in Figure 6.18c and purple in Figure 6.19 and 6.20. Similar to the multivariate clusters, these also account for the central bitumen-rich fines-poor layer. However, these clusters have slightly better spatial and poorer multivariate scores when compared to the multivariate clusters in the cluster metric space (purple vs. red; Fig. 6.19). Spatial patterns between the two sets of categories are similar (Fig. 6.18b vs. c).

6.4.1.4 Improved Spatial Clusters

A sub-ensemble is selected from the total set of clusterings generated in the previous section by finding clusterings with the best multivariate properties for different levels of spatial continuity (blue; Fig. 6.19). The same pairwise-similarity consensus function generates the improved spatial clusters from this sub-ensemble (Fig. 6.18d; brown in Fig. 6.19 & 6.20). The resulting clusters are improved in terms of spatial and multivariate properties with respect to the multivariate clusters and the spatial clusters generated from the full ensemble.

6.4.1.5 Cluster Realizations

The clusterings selected to form the improved spatial clusters are used as realizations of categories (blue; Fig. 6.19). After ensuring the cluster codes between realizations are compatible with the improved spatial clusters from above, the probability for each location to be in each category can be assessed over all realizations (Fig. 6.21). The spatial configuration of all clusters realizations is very similar to the improved spatial clusters, but with uncertainty mainly along the boundaries between the large-scale consistently-categorized zones. Furthermore, the continuous variables contained within each clustering realization are similar to those from the con-

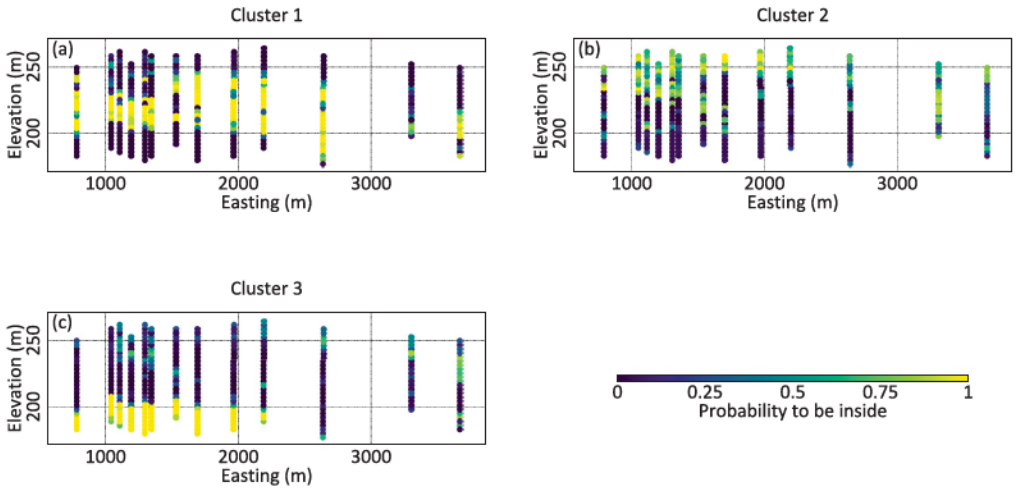


Figure 6.21: E-W section, looking north colored by the probability for each location to be part of each category.

sensus clustering (e.g., grey vs. brown; Fig. 6.20), though several of the cluster realizations delineate multivariate populations that are better delineated than the other methods.

6.4.1.6 Control Groups

Two control sets are tested; the first considers the total domain to be a single stationary population. In general, in a domain with trends, a single stationary population without proper attention to the trend component should produce poor results in sparse areas. However, given the strong stationary anisotropy in this domain, at any given unsampled location, a local search for conditioning data will ensure that related samples are found generate the estimate at that location. The second control group of randomized categories essentially tests three superimposed global populations, which should perform similar to the single population control group.

6.4.2 K-Fold Results

The geostatistical analysis follows the same steps outlined in the left and right of Figure 6.1 for the static and uncertain categories, respectively. The drillhole locations ordered by fold are shown in Figure 6.22. Scatter plots between the E-type estimate and the true values are shown in Figure 6.23, B.6 and B.7, and error metrics tabulated in Table 6.2, 6.4 and 6.3 for bitumen, fines and chlorides, respectively.

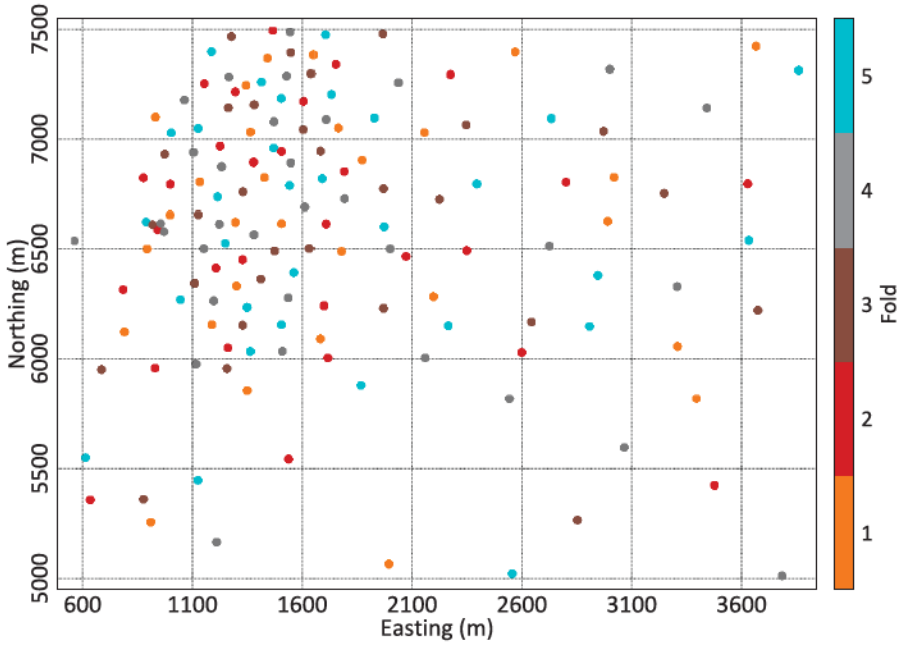


Figure 6.22: Sample locations colored by fold for the oilsands dataset

Table 6.2: Bitumen K-fold error statistics by category, oilsands dataset.

	Covariance	Correlation	RMSE
Category Realizations	16.28	**0.78**	3.16
Random Domain	14.57	0.77	3.26
Geological Categories	15.22	0.77	3.20
Multivariate Clusters	16.23	**0.78**	**3.15**
Spatial Clusters	16.27	**0.78**	3.18
Improved Spatial Clusters	**16.63**	**0.78**	**3.15**
No Domains	15.22	0.77	3.23

Table 6.3: Fines K-fold error statistics by category, oilsands dataset.

	Covariance	Correlation	RMSE
Category Realizations	253	**0.70**	16.01
Random Domain	235	0.68	16.38
Geological Categories	246	0.69	16.15
Multivariate Clusters	248	0.69	16.17
Spatial Clusters	253	**0.70**	16.13
Improved Spatial Clusters	**264**	**0.70**	**15.97**
No Domains	250	0.68	16.57

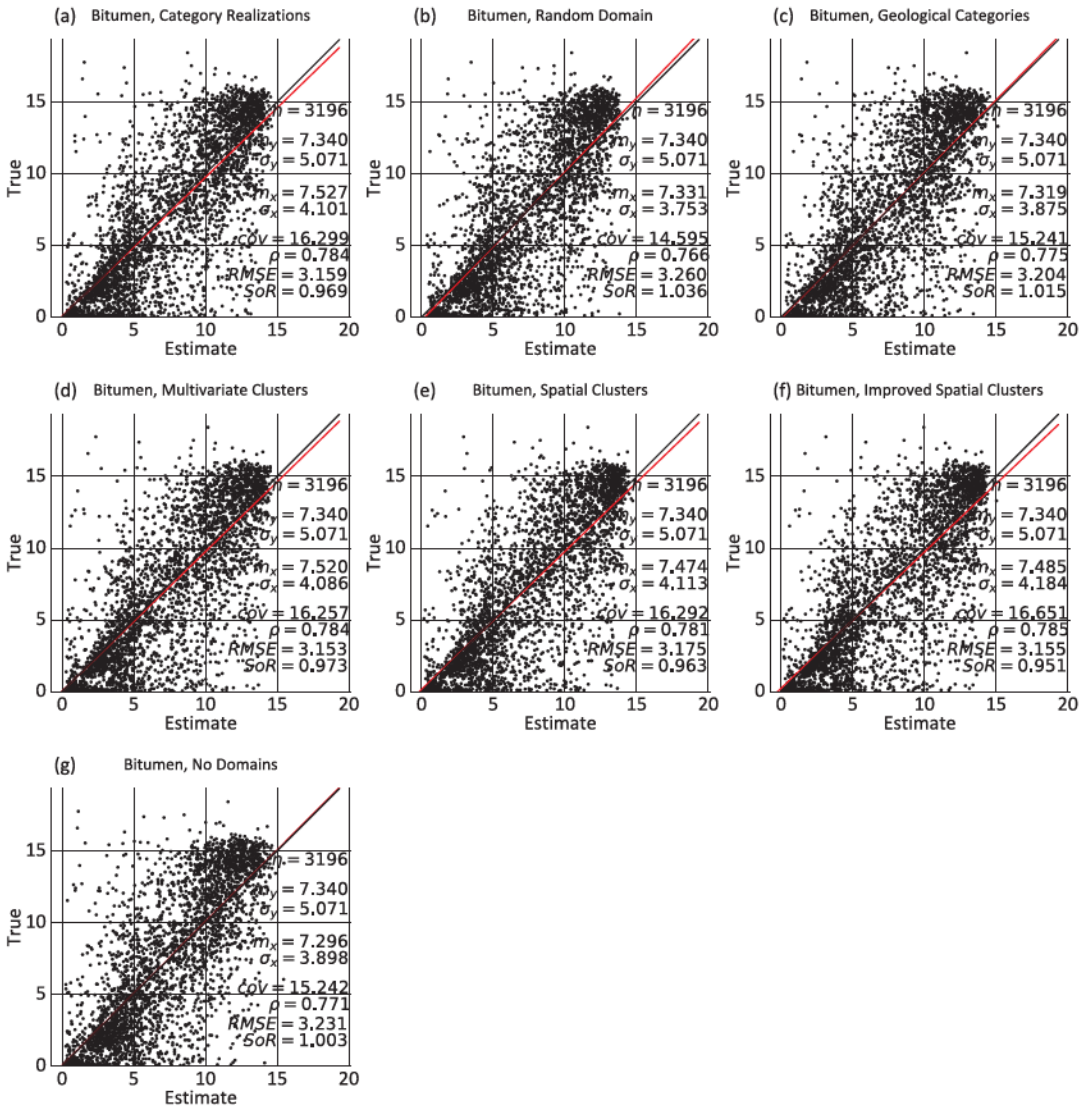


Figure 6.23: Cross plot between the E-type estimate and the true values from the oilsands dataset.

Table 6.4: Chlorides K-fold error statistics by category, oilsands dataset.

	Covariance	Correlation	RMSE
Category Realizations	129725	0.80	272
Random Domain	125968	0.80	274
Geological Categories	128111	**0.81**	**268**
Multivariate Clusters	131339	**0.81**	270
Spatial Clusters	131340	0.80	272
Improved Spatial Clusters	**131815**	**0.81**	269
No Domains	131718	0.80	273

For this dataset all methods produce global error metrics that are similar to one another, with a very small preference to the spatial clustering methodologies for stationary domains. The control groups consistently perform the worst with respect to covariance and RMSE, likely a consequence of the vertical zonation that is not accounted for in the adopted geostatistical modeling workflow. However, it should be noted that with respect to global prediction error assessed over 100 realizations, all methods generate E-type estimated values that are consistent with one another.

The error statistics generated by comparing the E-type estimate to the truth are corroborated by accuracy plots (Fig. 6.24, B.8 & B.9). Here the control groups generate local distributions of uncertainty that slightly over-predict local uncertainty for bitumen (Fig. 6.24b & g) and under-predict local uncertainty for fines (Fig. B.8b & g). Multivariate clustering and the category realizations generate local distributions that are both accurate and precise for bitumen, as points are fairly close to the 1:1 relationship. Interestingly, for bitumen and fines the best combined prediction of local uncertainty is obtained with the proposed category realizations. However, with respect to chlorides, the best prediction of local uncertainty is obtained by considering no domains (Fig. B.9g). Ultimately the practitioners stance on risk determines which prediction of local uncertainty is most suitable.

6.4.3 Global Realization Checks

A suite of final geostatistical realizations are generated using the entire modeling dataset; histogram reproduction plots are shown in Figure 6.25, B.10, and B.11 and variogram reproduction plots in Figure 6.26, B.12, and B.13 for bitumen, fines and chlorides, respectively.

All sets of categories generate realizations that slightly under predict quantiles of the reference bitumen distribution. Realizations from the control groups are slightly poorer than those from the other groups. The uncertain categories (Fig. 6.25a) are the only set that encompass the reference distribution and, relative to the improved spatial clusters, category realizations have captured greater uncertainty (Fig. 6.25a vs. 6.25f).

Variograms calculated from bitumen realizations show reasonable reproduction of the global variogram calculated considering all sample locations (Fig. 6.26). The vertical direction is well

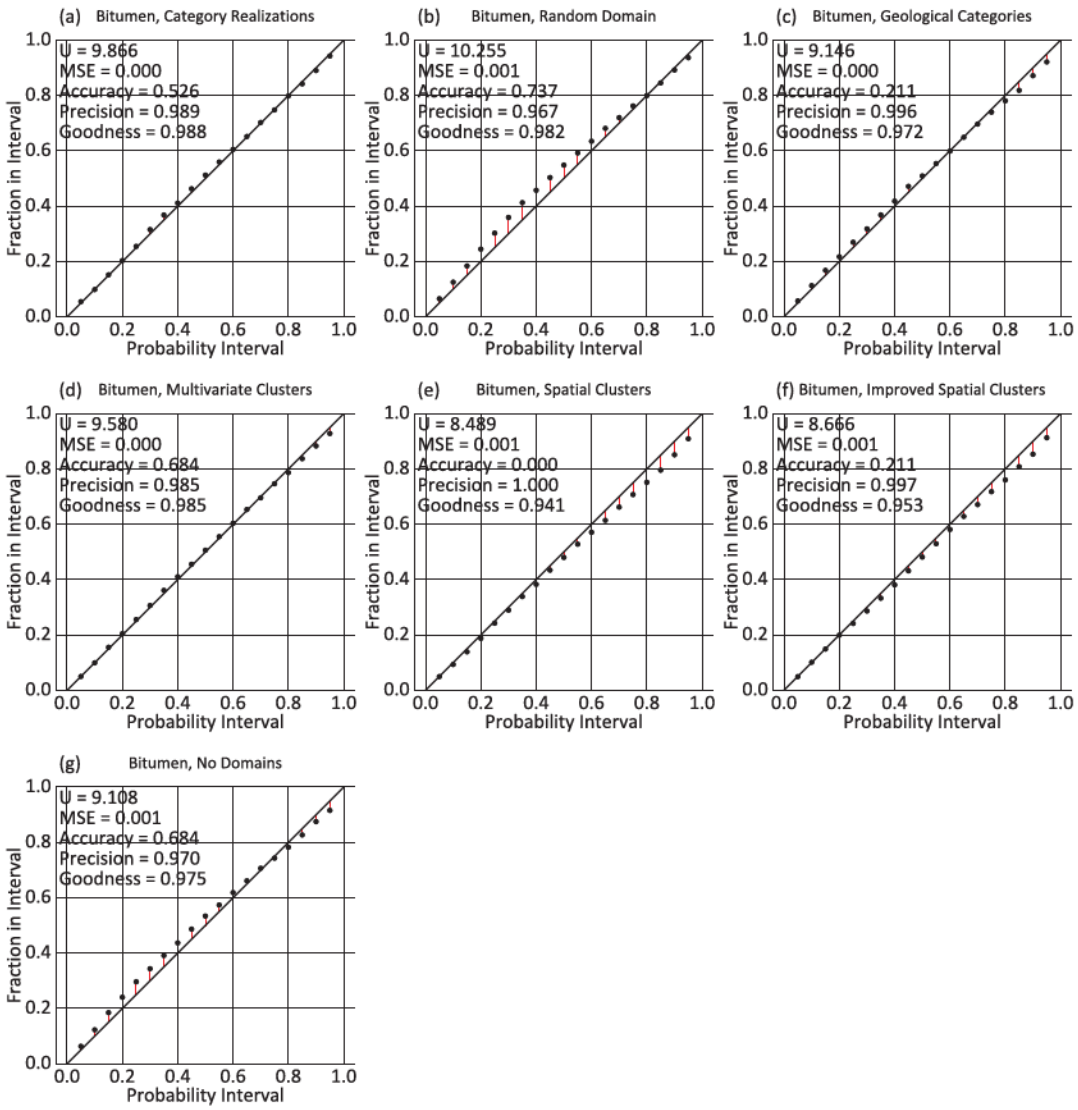


Figure 6.24: Accuracy plots generated from the K-fold validation analysis for bitumen.

reproduced by all methods, whereas reproduction of the short-range horizontal continuity is problematic for all methods.

6.4.4 Discussion

The oilsands dataset displays results that are consistent with the porphyry dataset. Interestingly, even with clear non-stationarity found in this domain, the high horizontal continuity combined with representative search parameters results in estimates that are largely un-affected by this non-stationarity; predictions generated considering no domains perform only slightly worse than the best methods for stationary domaining. However, over all methods there is a slight

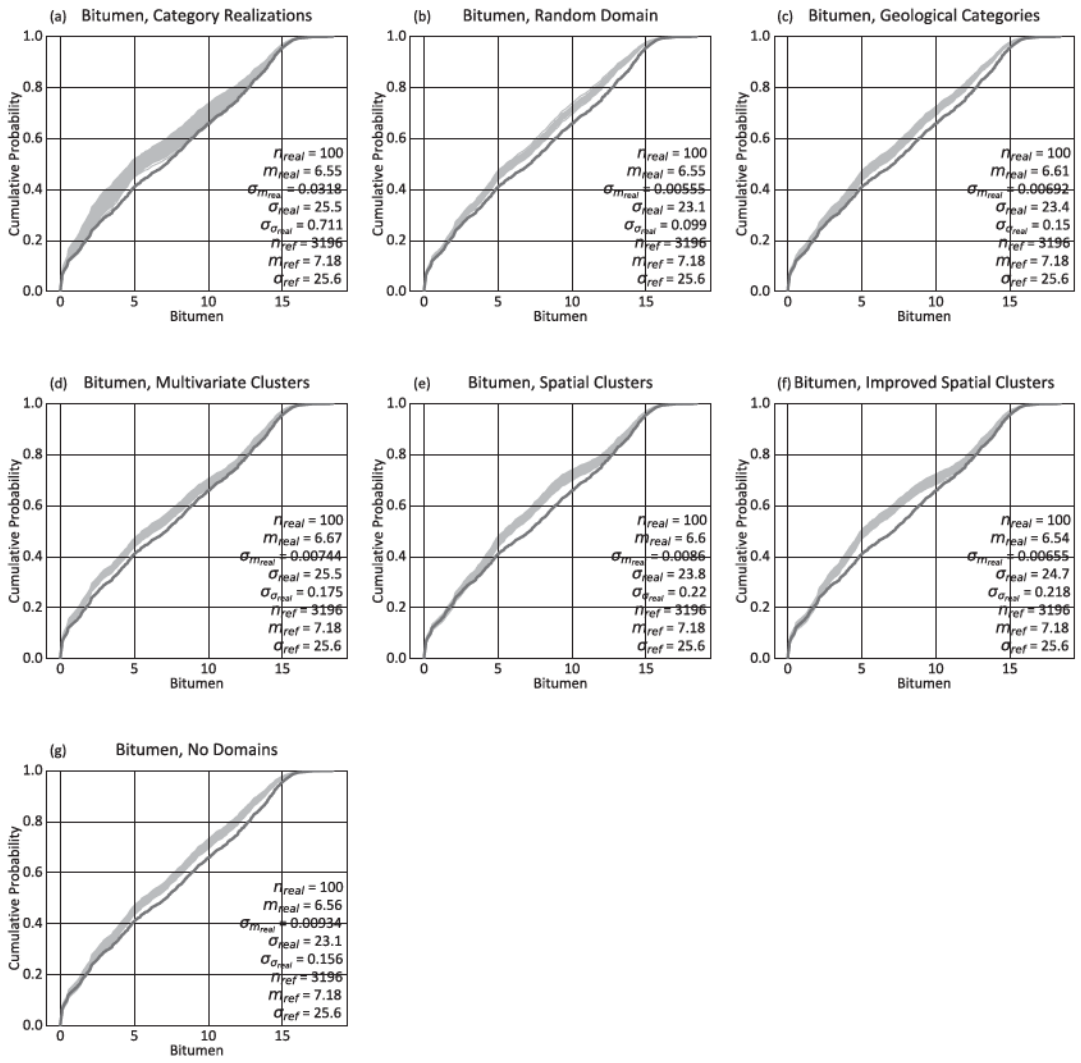


Figure 6.25: Histogram reproduction plots for bitumen from the oilsands dataset.

preference to the spatial clustering methodologies in terms of maximizing the covariance and correlation and minimizing the RMSE between the estimate and the truth. Spatial clustering and cluster realization stationary decisions result in the best prediction of local uncertainty. Considering the final suite of realizations, the additional variability between realizations captured by the clustering-realization stationary domains suggests that the uncertainty associated with stationary domaining has been incorporated without adversely affecting the other qualities of the generated realizations.

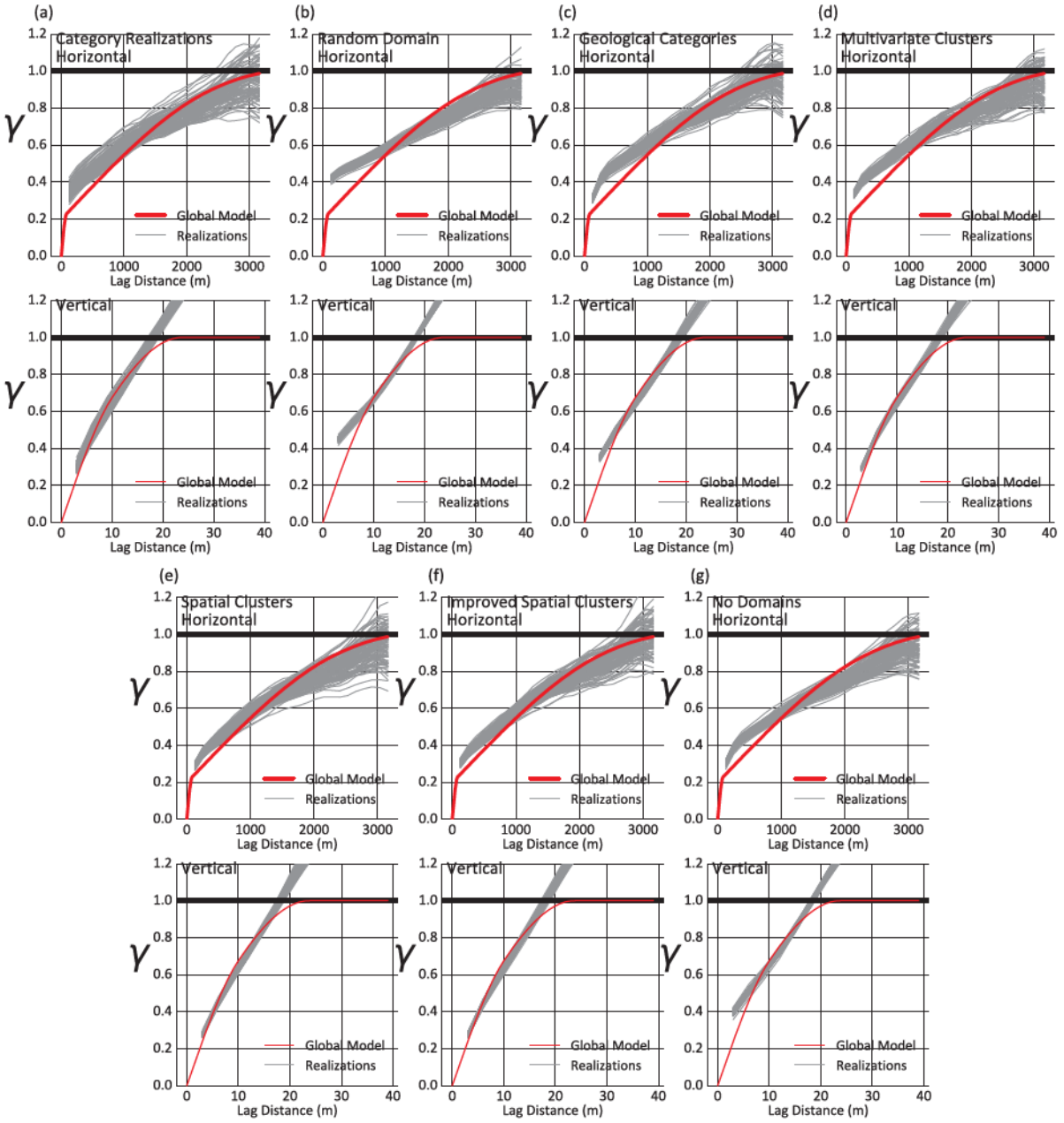


Figure 6.26: Variogram reproduction for realizations of bitumen generated by each method. Horizontal reproduction shown in the top and vertical shown below for each set of stationary domains (a-g).

6.5 Review of Main Points

This chapter demonstrates the consequence of different stationary domaining methodologies, followed through conventional geostatistical analyses. Specific contributions of this chapter include:

1. Demonstration of the insight provided by the cluster metrics developed in Chapter 4;
2. Confirmation that spatial clusters represent a better set of categories for geostatistical analysis over traditional multivariate-only clusters;
3. The DS clustering algorithm and geostatistical ensemble analysis in Chapter 4 are effective for generating stationary domains within a large-scale geological context;
4. The uncertainty associated with defining categories can be incorporated into existing uncertainty characterization workflows with minimal scripting efforts;

6.6 Conclusions

Decisions made during delineation of stationary domains impact the predictive performance of geostatistical models. The proposed cluster-metrics provide a useful summary of how stationary domains will perform in subsequent geostatistical analysis, thereby facilitating a reproducible stationary domaining workflow without an arduous geostatistical validation workflow. Spatial clusters and spatial cluster realizations as stationary domains are indeed beneficial to subsequent geostatistical modeling in the form of increased local and global accuracy. However, given the marginal differences between control groups and the developed methods for stationary delineation, further research into when stationary domaining is appropriate for a given domain is warranted.

Chapter 7

Conclusions

This dissertation develops contributions to implicit geological modeling and stationary decision making that improve decisions made early early in geostatistical analysis. The developed methodologies and workflows are practical and help capture complex features in geostatistical datasets.

7.1 Contributions to Geological Boundary Modeling

Geological boundaries are an important component of numerical models as they control domain extents and inform the setting of mineralized features.

Choosing the algorithm by which a volumetric function is interpolated in implicit modeling is critical since it affects the speed of the interpolation, the types of anisotropy that can be included, and the information required to parameterize the interpolation framework. Global interpolators, like dual kriging or RBF, are advantageous since weights are calculated once and the interpolators result in an artifact-free map that is ideal for extracting the geological surface. However, global methods suffer from CPU and RAM limitations for large datasets.

The practitioner must also choose whether or not to model with LVA. This decision depends on the structural relationships for the given modeling project. This decision may require domain knowledge or sufficient data to identify LVA as a necessary component for capturing the geological relationships. Once the decision has been made to implicitly model with LVA, the next decision is what kind of LVA to implement. A number of different techniques are available ranging from manual domain partitioning, stratigraphic transformations, unfolding, distance-weighted local statistics, automated domain partitioning, or the full LVA implementation using shortest-paths.

Three main contributions are made to implicit geological modeling in this thesis: 1) a boundary modeling algorithm for large datasets that can incorporate local anisotropy; 2) an iterative

local anisotropy refinement algorithm combining automatic local orientation extraction with the developed boundary modeling algorithm; and 3) a method to quantify shape properties of subjectively poor implicit models.

7.1.1 Implicit Geological Modeling with Local Anisotropy

Previous works studying implicit modeling with LVA in RBF interpolation frameworks mainly target a sparse data environment, where local structural orientations integrated into the interpolation improves the generated surfaces where no point information is present (Hillier et al., 2014). Outside of RBF interpolators, several strategies for interpolation considering LVA are available. However, the unique properties of the underlying SDF restricts which interpolators are suitable for SDF interpolation. The implicit modeling framework developed in Chapter 3 specifically targets a dense data environment, where the point dataset comprises the bulk of the information for boundary reconstruction and global interpolation is problematic from a computational standpoint. Such domains are common in earth sciences, where a wealth of point information is available. Partitioning of the domain into overlapping subdomains has several benefits for implicit geological modeling. The characteristics of local surfaces can be improved since parameters can be locally representative. The decomposition of large problems into small overlapping subproblems intrinsically speeds up boundary interpolation since the large $N \times N$ system of equations is avoided. Interpolation speed can be increased further by parallelizing partitions across several CPU threads. Moreover, since partitions are independent, if a particular area receives more information, the model update step only requires re-interpolation of affected partitions to account for the new data. This greatly reduces the computational complexity of implicit modeling of geological domains and allows rapid iteration between different interpretations or model updates in the presence of new data.

7.1.2 Iterative Refinement of Geological Features

Identifying and modeling local features in geological domains can be challenging (Lillah & Boisvert, 2015). The geomodeler must consider the structural environment and the composite point dataset in three dimensions to determine reasonable structures represented in the

dataset that can be captured by the geological models. This interpretation of structural continuity is largely achieved through inspection of the composite database with a 3D viewer (Cowan, 2014) and manually iterating through different structural interpretations to generate reasonable models. An algorithm to iteratively and automatically infer local orientations from the data is developed by combining local orientation extraction and boundary modeling with LVA. A similar algorithm is proposed by te Stroet and Snepvangers (2005), however, in that study the refinement of local anisotropy requires a local search such that different kriging parameters are used at each location. The proposed algorithm in this thesis starts by inferring local orientations from a boundary model generated considering the best global anisotropy determined for the target boundaries. Each refinement is generated under the local anisotropic conditions inferred from the previous step. Given the structural features are present at a scale suitable for implicit modeling, the result is a reasonable reconstruction of local structural features directly from the data. The resulting local properties and models can be used to guide further interpretation of the dataset.

7.1.3 Shape Properties of Geological Boundary Models

Subjectively poor implicit models contain bubble-like features where the local conditions of the boundary model do not reflect the underlying features of the dataset. A metric is proposed to quantify these features such that different models generated under different anisotropic conditions can be compared.

7.2 Contributions to the Decision of Stationarity

The decision of stationarity is a critical decision made early in the geostatistical workflow and influences all subsequent modeling steps. Best practice and conventional wisdom states that a better decision of stationarity improves geostatistical analysis, yet, justification of different possible decisions of stationarity is commonly subjective, prone to errors and may bias the geostatistical modeling that follows.

This dissertation develops several contributions that improve the generation and assessment of stationary domains, including: 1) combined spatial and multivariate metrics that express

the targeted properties of stationary domains for interpretation; 2) novel clustering algorithms that reduce parameterization requirements; and 3) improvements to conventional ensemble analysis specifically for geostatistical datasets. Moreover, this dissertation develops a novel method to assess the uncertainty associated with the decision of stationarity through ensemble clustering and a practical modification to conventional geostatistical workflows.

7.2.1 Spatial and Multivariate Metrics

Assessment of different stationary domains for geostatistical analysis is a difficult problem, typically requiring subjective decisions made with statistical justification (Rossi & Deutsch, 2014). Justification of a clustering algorithm itself, without considering a spatial component, is a similarly difficult problem (Tibshirani & Walther, 2005; Tibshirani et al., 2001). Most studies agree that domain-specific knowledge improves the results of clustering since decisions can be tailored to the unique properties of the domain. The combined spatial-multivariate metrics developed in this thesis provide a useful geostatistics-specific criteria for comparing different sets of categories defined for the same dataset. By considering the relative properties between different configurations, the choice of which stationary delineation to consider for modeling can be improved.

7.2.2 Spatial Clustering Algorithms

Spatial clustering algorithms developed in the literature are well suited to forming spatial-multivariate clusters in spatial domains. However, the results of those algorithms are subject to the parameterization determined by the modeling expert. Although the domain specific knowledge is considered an asset in this work, the outcome is based on the subjective criteria of the expert and does not account for uncertainty.

The random-path ensemble spatial clusterer developed in Chapter 4 generates improved spatial clusters that require fewer input parameters when compared to other spatial clustering methodologies. The consensus of an ensemble of clusterings is shown to generate reasonable clusters for further geostatistical analysis in spatial domains. The random-path and 2-stage merging is an appropriate strategy for ensemble generation in spatial domains.

7.2.3 Geostatistical Ensemble Clustering

The use of the spatial-multivariate metric space as a pre-processor for ensemble analysis is a novel strategy for imparting domain specific knowledge into consensus clustering in spatial domains. The result is a clustering that is improved in terms of the spatial and multivariate properties.

7.2.4 Investigation of the Effect of Stationary Domaining on Prediction Errors

Finally, this dissertation develops two case studies that demonstrate the effects of different stationary decisions on the outcome of geostatistical uncertainty characterization. The two domains are very different in terms of the stationary domaining requirements. In the porphyry case, the large-scale geological partitioning of the domain accounts for some of the non-stationary features. Subsequent partitioning with traditional methods resulted in marginal improvements. In this case, only the improved spatial clusters matched the performance of the single stationary population.

The second oilsands dataset reflects conditions more typical of geological domains, where gradational features may not be captured by a single population. However, in this domain, the strong anisotropy and stationary nature of the spatial variability resulted in only marginal differences between methods that partitioned the dataset into stationary domains compared to considering a single global population.

7.3 Limitations and Future Work

There are several areas for improvement in the developed implicit modeling and stationary domaining methodologies.

7.3.1 Implicit Geological Modeling

The algorithm chosen to partition the domain in Chapter 3 is important as it controls the time required for boundary reconstruction and the potential for artefacts at the boundaries of partitions. The axis-aligned partitioning algorithm chosen for implementation in this thesis may not be ideal

for all datasets and all domains. Furthermore, the partitioning is solely based on the number of data contained within each partition and not the properties of the underlying variables. For example, the partition size should generally be related to the local variability in the underlying variable; greater variability warrants additional partitioning to better-capture the local features.

The partitioning framework is not immune to other issues of statistical inference. Extrapolation is an outstanding issue for all estimation/interpolation methods. Extrapolation in the partitioning framework is likely to cause noticeable partition-boundary artefacts in the interpolated SDF. This is not usually an issue for implicit models since the 0-level is of primary interest, and is usually well informed by sample locations. However, the issue is exacerbated in the PU-LVA framework since adjacent partitions with different anisotropic properties may display inconsistencies along partition boundaries.

For the iterative local refinement algorithm the criteria by which models are considered finished can be improved. Stopping iterations when there is no further change is an acceptable first pass, however, iterating model orientations based on cross validation scores could be considered.

The algorithm used to infer local anisotropy to the boundary models could be simplified; a moment of inertia (MOI) (Hassanpour & Deutsch, 2007) considering the local composite information or the interpolated SDF values could be considered in place of the gradient-SVD method used in this thesis.

7.3.2 Limitations of the Proposed Stationary Domaining Methodologies

First and foremost, decisions of stationarity must be made within some larger geological context. However, knowing when geological controls have been reasonably accounted for in the delineation of distinct stationary domains is a challenging task. Ideally the clustering algorithm would consider the geological partitioning in the generation of clusters for modeling. In this sense a supervised clustering algorithm should be the target of future works.

The spatial clustering implementations developed for stationary domaining in this thesis mainly target the unsupervised classification of stationary domains from the spatial-multivariate dataset. However, future works should integrate logged rock properties from the geological logs

to improve the classification extracted from the spatial-multivariate dataset.

Although the developed metrics account for spatial and multivariate properties of the dataset, a third metric that measures the 2-point spatial continuity of the samples contained within each delineated domain could be considered. This is suggested to account for contrasting relationships observed during verification of the cluster metrics for informing goodness of the clusters as stationary domains for geostatistical analysis. The developed metrics provide an important start for unbiased assessment of stationary delineation, but additional investigation into the factors explaining 'goodness' of a stationary delineation is warranted.

7.4 Final Words

Recall the statement of this thesis: *Incorporating structural features to large-scale geological boundaries and improving the definition of stationary domains results in improved geostatistical models.*

The incorporation of local anisotropy to geological boundary models improves the local feature reproduction and provides reasonable geological boundaries inside which resources can be estimated. Inferring local anisotropy from dense composite datasets through bootstrap of previous boundary models results in a reasonable first-pass local structural interpretation of the features in geological domains.

Improving the delineation of stationary populations by considering spatial-clustering and spatial-multivariate metrics results in geostatistical models that better reproduce the true values at unsampled locations. This translates into improved geostatistical models for further decision making. Automating the delineation of stationary domains with unsupervised learning algorithms and the developed clustering metrics is reasonable in the context of large-scale geological features.

References

Abzalov, M. (2016). *Applied Mining Geology* (Vol. 12). Cham: Springer International Publishing. doi: 10.1007/978-3-319-39264-6

Afanador, N. L., Smolinska, A., Tran, T. N., & Blanchet, L. (2016). Unsupervised random forest: A tutorial with case studies. *Journal of Chemometrics*, *30*(5), 232-241. doi: 10.1002/cem.2790

Ambroise, C., & Govaert, G. (1998). Convergence of an EM-type algorithm for spatial clustering. *Pattern Recognition Letters*, *19*(10), 919-927.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, *27*(2), 93-115. doi: 10.1111/j.1538-4632.1995.tb00338.x

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, *8*, 1027-1035. doi: 10.1145/1283383.1283494

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

Barnett, R. M., & Deutsch, C. V. (2015). *Linear Rotations : Options for Decorrelation and Analysis* (CCG Annual Report 17). Edmonton, AB: University of Alberta.

Beatson, R. K., a Light, W., & Billings, S. (2001). Fast solution of the radial basis function interpolation equations: Domain decomposition methods. *SIAM Journal on Scientific Computing*, *22*(5), 1717-1740.

Beatson, R. K., Cherrie, J. B., & Mouat, C. T. (1999). Fast Fitting of radial basis functions: Methods based on preconditioned GMRES iteration. *Advances in Computational Mathematics*, *11*, 253-270. doi: 10.1023/A:1018932227617

Boisvert, J. B. (2010). *Geostatistics with Locally Varying Anisotropy* (PhD). University of Alberta.

Boisvert, J. B. (2013). *Automatic Geological Modeling with Iterative Refinement* (CCG Annual Report 15). Edmonton, AB: University of Alberta.

Boisvert, J. B., & Deutsch, C. V. (2011). Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. *Computers and Geosciences*, *37*(4), 495-510. doi: 10.1016/j.cageo.2010.03.021

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. doi: 10.1023/A:1010933404324

Carr, J. C., Beatson, R. K., Cherrie, J. B., Mitchell, T. J., Fright, W. R., McCallum, B. C., & Evans, T. R. (2001). Reconstruction and Representation of 3D Objects with Radial Basis Functions. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 67-76. doi: 10.1145/383259.383266

Chautru, J.-M., Chautru, E., Garner, D., Srivastava, R. M., & Yarus, J. (2017). Using Spatial Constraints in Clustering for Electrofacies Calculation. *Geostatistics Valencia 2016 Quantitative Geology and Geostatistics*, *19*. doi: 10.1007/978-3-319-46819-8_31

Cowan, J. (2014). 'X-ray Plunge Projection'— Understanding Structural Geology from Grade Data 'X-ray Plunge Projection'— Understanding Structural Geology from Grade Data. *AusIMM Monograph 30: Mineral Resource and Ore Reserve Estimation — The AusIMM Guide to Good Practice*, *2*, p. 207-220.

Cowan, J., Beatson, R. K., Ross, H., Fright, W. R., McLennan, T., Evans, T. R., … Titley, M. (2003). Practical implicit geological modelling. In *5th International Mining Geology Conference* (p. 89-99).

Cuomo, S., Galletti, A., Giunta, G., & Starace, A. (2013). Surface reconstruction from scattered point via RBF interpolation on GPU. In *Computer Science and Information Systems (FedCSIS)* (p. 433-440).

Deacon, J. (2017). *Explicit and Implicit Geological Modeling Methods on Resource Definition and Resource Utilization - Sishen Iorn Ore Deposit Case Study* (MSc). University of Witwatersrand.

Deutsch, C. V. (1996). Direct Assessment of Local Accuracy and Precision. In *GEOSTATISTICS WOLLONGONG 96 - PROCEEDINGS OF THE FIFTH INTERNATIONAL GEOSTATISTICS CONGRESS*. Wollongong, Australia.

Deutsch, C. V. (2005). *A Short Note on Prediction of Uncertainty in Tonnage and Grade of Vein*

Type Deposits (CCG Annual Report 07). Edmonton, AB: University of Alberta.

Deutsch, C. V. (2006, December). A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. *Computers & Geosciences*, *32*(10), 1669-1681. doi: 10.1016/j.cageo.2006.03.005

Deutsch, C. V., & Journel, A. G. (1998). *GSLIB: Geostatistical software library and user's guide* (2nd ed.). New York: Oxford University Press.

Dominy, S. C., & Edgar, W. B. (2012). Approaches to reporting grade uncertainty in high nugget gold veins. *Applied Earth Science*, *121*(1), 29-42. doi: 10.1179/1743275812Y .0000000013

Dubé, B., & Gosselin, P. (2007). Greenstone-hosted quartz-carbonate vein deposits. *Mineral Deposits of Canada: A synthesis of major deposit-types, district metallogeny, the evolution of geological provinces, and exploration methods: Geological Association of Canada, Mineral Deposits Division, Special Publication*, *5*, 49-73.

Eckstrand, O., Roger, H., & Larry, J. (2007). Magmatic nickel-copper-platinum group element deposits. *Mineral deposits of Canada: A synthesis of major deposit types, district metallogeny, the evolution of geological provinces, and exploration methods: Geological association of Canada, mineral deposits division, special publication*, *5*, 205-222.

Emery, X., & Ortiz, J. M. (2005). Estimation of mineral resources using grade domains. *The Journal of the South African Institute of Mining and Metallurgy*, *105*, 10.

Şenbabaoğlu, Y., Michailidis, G., & Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. *Scientific reports*, *4*, 6207. doi: 10.1038/srep06207

Fasshauer, G. E. (2007). *Meshfree Approximation Methods with MATLAB*. Singapore: World Scientific.

Fasshauer, G. E., & Zhang, J. G. (2007). On choosing "optimal" shape parameters for RBF approximation. *Numerical Algorithms*, *45*(1-4), 345-368. doi: 10.1007/s11075-007-9072 -8

Fazio, V. S., & Roisenberg, M. (2013). Spatial Interpolation: An Analytical Comparison Between Kriging and RBF Networks. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC '13* (Vol. 2, p. 2-7). Coimbra, Portugal: ACM Press. doi:

10.1145/2480362.2480364

Feng, X., & Milanfar, P. (2002). Multiscale principal components analysis for image local orientation estimation. *Proc. Thirty-Sixth Asilomar Conference on Signals, Systems and Computers*, *1*, 478-482. doi: 10.1109/ACSSC.2002.1197228

Fouedjio, F. (2016a). A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*. doi: 10.1016/j.spasta.2016.07.003

Fouedjio, F. (2016b). Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, 1-20. doi: 10.1007/s00477-016-1274-y

Fouedjio, F., Hill, E. J., & Laukamp, C. (2017). Geostatistical clustering as an aid for ore body domaining: Case study at the Rocklea Dome channel iron ore deposit, Western Australia. *Applied Earth Science*. doi: 10.1080/03717453.2017.1415114

Ghaemi, R., Sulaiman, N., Ibrahim, H., & Mustapha, N. (2009). A Survey: Clustering Ensemble Techniques. *International Journal of Computer and Information Engineering*, *3*(2).

Hadavand, M., & Deutsch, C. V. (2017). Facies proportion uncertainty in presence of a trend. *Journal of Petroleum Science and Engineering*, *153*, 59-69. doi: 10.1016/j.petrol.2017.03.036

Hassanpour, R. M., & Deutsch, C. V. (2007). *Determination of Locally Varying Directions through Mass Moment of Inertia Tensor* (CCG Annual Report 09). Edmonton, AB: University of Alberta.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Hershey, J. R., & Olsen, P. A. (2007). Approximating the Kullback Leibler divergence between Gaussian mixture models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Vol. 4, p. IV-317-IV-320). IEEE. doi: 10.1109/ICASSP.2007.366913

Hillier, M. J., Schetselaar, E., Courrioux, G., Caumon, G., Wellmann, F., Lindsay, M., … Jessell, M. (2015). *Achieving Geologically Reasonable 3D Models.*

Hillier, M. J., Schetselaar, E. M., de Kemp, E. A., & Perron, G. (2014). Three-Dimensional Mod-

elling of Geological Surfaces Using Generalized Interpolation with Radial Basis Functions. *Mathematical Geosciences*, 931-953. doi: 10.1007/s11004-014-9540-3

Hosseini, A. H., & Deutsch, C. V. (2007). *A Distance Function Based Algorithm to Quantify Uncertainty in Areal Limits* (CCG Annual Report 09). Edmonton, AB: University of Alberta.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193-218. doi: 10.1007/BF01908075

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651-666. doi: 10.1016/j.patrec.2009.09.011

Jewbali, A., Perry, B., Allen, L., & Inglis, R. (2016). Implications of Algorithm and Parameter Choice: Impacts of Geological Uncertainty Simulation Methods on Project Decision Making. *Quantitative Geology and Geostatistics*, *19*, 225-243. doi: 10.1007/978-3-319-46819-8_15

Journel, A. G. (1988). *Fundamentals of Geostatistics in Five Lessons* (Tech. Rep.).

Khan, K. D., & Deutsch, C. V. (2016). Practical Incorporation of Multivariate Parameter Uncertainty in Geostatistical Resource Modeling. *Natural Resources Research*, *25*(1), 51-70. doi: 10.1007/s11053-015-9267-y

Klaas, O., & Shephard, M. S. (2000). Automatic generation of octree-based three-dimensional discretizations for Partition of Unity methods. *Computational Mechanics*, *25*(2-3), 296-304. doi: 10.1007/s004660050478

Knight, R. H., Lane, R. G., Ross, H. J., Abraham, a. P. G., & Cowan, J. (2007). Implicit Ore Delineation. In *Proceedings of Exploration 07: Fifth Decennial International Conference on Mineral Exploration* (p. 1165-1169).

Lajaunie, C., Courrioux, G., & Manuel, L. (1997). Foliation fields and 3D cartography in geology: Principles of a method based on potential interpolation. *Mathematical Geology*, *29*(4), 571-584. doi: 10.1007/BF02775087

Leuangthong, O., & Nowak, M. (2015). Dealing with high-grade data in resource estimation. *The Journal of The Southern African Institute of Mining and Metallurgy*, *115*, 27-36.

Leuangthong, O., & Srivastava, R. M. (2012). On the Use of MultiGaussian Kriging for Grade Domaining in Mineral Resource Characterization. In *Ninth International Geostatistics*

Congress. Oslo.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R news*, *2*(December), 18-22. doi: 10.1177/154405910408300516

Lillah, M. (2014). *Inference of Locally Varying Anisotropy Fields* (MSc). University of Alberta.

Lillah, M., & Boisvert, J. B. (2013). Stochastic Distance Based Geological Boundary Modeling with Curvilinear Features. *Math Geoscience*, *45*, 651-665.

Lillah, M., & Boisvert, J. B. (2015). Inference of locally varying anisotropy fields from diverse data sources. *Computers & Geosciences*, *82*, 170-182. doi: 10.1016/j.cageo.2015.05.015

Lindsay, M. D., Aillères, L., Jessell, M. W., de Kemp, E. A., & Betts, P. G. (2012). Locating and quantifying geological uncertainty in three-dimensional models: Analysis of the Gippsland Basin, southeastern Australia. *Tectonophysics*, *546-547*, 10-27. doi: 10.1016/j.tecto.2012.04.007

Lindsay, M. D., Jessell, M. W., Ailleres, L., Perrouty, S., De Kemp, E., & Betts, P. G. (2013). Geodiversity: Exploration of 3D geological model space. *Tectonophysics*, *594*, 27-37. doi: 10.1016/j.tecto.2013.03.013

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, *21*(4), 163-169. doi: 10.1145/37402.37422

Machuca-Mory, D. F., & Deutsch, C. V. (2013). Non-stationary Geostatistical Modeling Based on Distance Weighted Statistics and Distributions. *Mathematical Geosciences*, *45*(1), 31-48. doi: 10.1007/s11004-012-9428-z

Machuca-Mory, D. F., Rees, H., & Leuangthong, O. (2015). Grade Modelling with Local Anisotropy Angles: A Practical Point of View. In *APCOM 2015*.

Mallet, J. L. (2004). Space-time mathematical framework for sedimentary geology. *Mathematical Geology*, *36*(1), 1-32. doi: 10.1023/B:MATG.0000016228.75495.7c

Manchuk, J. G., & Deutsch, C. V. (2012). A flexible sequential Gaussian simulation program: USGSIM. *Computers & Geosciences*, *41*, 208-216. doi: 10.1016/J.CAGEO.2011.08.013

Manchuk, J. G., & Deutsch, C. V. (2015). *Geometric Modeling of Irregular Tabular Deposits* (CCG Annual Report 17). Edmonton, AB: University of Alberta.

Manchuk, J. G., Leuangthong, O., & Deutsch, C. V. (2009). The Proportional Effect. *Mathematical Geosciences*, *41*(7), 799-816. doi: 10.1007/s11004-008-9195-z

Manita, G., Khanchel, R., & Limam, M. (2012). Consensus Functions for Clustering Ensembles. *Applied Artificial Intelligence*, *26*(6), 598-614. doi: 10.1080/08839514.2012.687668

Martin, R., Manchuk, J. G., & Boisvert, J. B. (2015). *Automatic LVA field generation in 3D* (CCG Annual Report 17). Edmonton, AB: University of Alberta.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, *58*(8), 1246. doi: 10.2113/gsecongeo.58.8.1246

Mathisen, T., Heon Lee, S., & Datta-Gupta, A. (2001). Improved Permeability Estimates in Carbonate Reservoirs Using Electrofacies Characterization: A Case Study of the North Robertson Unit, West Texas. In *SPE Permian Basin Oil and Gas Recovery Conference*. Society of Petroleum Engineers.

McLennan, J. A. (2008). *The Decision of Stationarity* (PhD). University of Alberta.

McLennan, J. A., & Deutsch, C. V. (2006). *Implicit Boundary Modeling (BOUNDSIM)* (CCG Annual Report 08). Edmonton, AB: University of Alberta.

Munroe, M. J., & Deutsch, C. V. (2008). *A Methodology for Modeling Vein-Type Deposit Tonnage Uncertainty* (CCG Annual Report 10). Edmonton, AB: University of Alberta.

Ohtake, Y., Belyaev, A., & Seidel, H. P. (2003). A multi-scale approach to 3D scattered data interpolation with compactly supported basis functions. In *Shape Modeling International, 2003* (p. 153-161). doi: 10.1109/SMI.2003.1199611

Ohtake, Y., Belyaev, A., & Seidel, H. P. (2006). Sparse surface reconstruction with adaptive partition of unity and radial basis functions. *Graphical Models*, *68*(1), 15-24. doi: 10.1016/j.gmod.2005.08.001

Oliver, M. A., & Webster, R. (1989). A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, *21*(1), 15-35. doi: 10.1007/BF00897238

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, *263*.

Perez, H. H., & Datta-Gupta, A. (2005). The Role of Electrofacies, Lithofacies, and Hydraulic Flow Units in Permeability Prediction From Well Logs: A Comparative Analysis Using

Classification Trees. *SPE Resevoir Evaluation & Engineering*.

Pouderoux, J., Gonzato, J.-C., Tobor, I., & Guitton, P. (2004). Adaptive hierarchical RBF interpolation for creating smooth digital elevation models. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems - GIS '04* (p. 232). doi: 10.1145/1032222.1032256

Qu, J., & Deutsch, C. V. (2017). Geostatistical Simulation with a Trend Using Gaussian Mixture Models. *Natural Resources Research*, 1-17. doi: 10.1007/s11053-017-9354-3

Rathore, P., Bezdek, J. C., Erfani, S. M., Rajasegarar, S., & Palaniswami, M. (2018). Ensemble Fuzzy Clustering Using Cumulative Aggregation on Random Projections. *IEEE Transactions on Fuzzy Systems*, *26*(3), 1510-1524. doi: 10.1109/TFUZZ.2017.2729501

Rizzo, M. L., & Székely, G. J. (2016). Energy Distance. *WIREs Comput Stat*, *8*(1), 27-38. doi: 10.1002/wics.1375

Romary, T., Ors, F., Rivoirard, J., & Deraisme, J. (2015). Unsupervised classification of multivariate geostatistical data: Two algorithms. *Computers & Geosciences*, *85*, 96-103. doi: 10.1016/j.cageo.2015.05.019

Rossi, M. E., & Deutsch, C. V. (2014). *Mineral Resource Estimation*. Springer Science. doi: 10.1007/978-1-4020-5717-5

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53-65.

Scrucca, L. (2005). *Clustering multivariate spatial data based on local measures of spatial autocorrelation* (Tech. Rep.). Perugia, Italy: Universita a degli Studi di Perugia.

Silva, D. (2015). *Enhanced Geologic Modeling with Data-Driven Training Images for Improved Resources and Recoverable Reserves* (PhD). University of Alberta, Edmonton.

Silva, D., & Deutsch, C. V. (2012a). *An Entropy-Based Measure of Continuity for Weighting Multiple Training Images* (CCG Annual Report 14). Edmonton, AB: University of Alberta.

Silva, D., & Deutsch, C. V. (2012b). *Modeling Multiple Rock Types with Distance Functions : Methodology and Software* (CCG Annual Report 14). Edmonton, AB: University of Alberta.

Silva, D., & Deutsch, C. V. (2012c). *Multiple Point Statistics with Multiple Training Images* (CCG

Annual Report 14). Edmonton, AB: University of Alberta.

Stegman, C. L. (2001). How Domain Envelopes Impact on the Resource Estimate — Case Studies from the Cobar Gold Field, NSW, Australia. In A. C. Edwards (Ed.), *Mineral Resource and Ore Reserve Estimation – The AusIMM Guide to Good Practice* (p. 221-236). Melbourne.

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological methods*, *9*(3), 386-396. doi: 10.1037/1082-989X.9.3.386

Strehl, A., & Ghosh, J. (2002). Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. *Journal of Machine Learning Research*, *3*, 583-617.

Székely, G. J., & Rizzo, M. L. (2005). Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, *22*, 151-183.

te Stroet, C., & Snepvangers, J. (2005). Mapping Curvilinear Structures with Local Anisotropy Kriging. *Mathematical Geology*, *37*(6).

Tibshirani, R., & Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511-528. doi: 10.1198/106186005X59243

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411-423. doi: 10.1111/1467-9868.00293

Topchy, A., Jain, A., & Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1866-1881. doi: 10.1109/TPAMI.2005.237

Torghabeh, A., Rezaee, R., Moussavi-Harami, R., Pradhan, B., Kamali, M., & Kadkhodaie-Ilkhchi, A. (2014). Electrofacies in gas shale from well log data via cluster analysis: A case study of the Perth Basin, Western Australia. *Open Geosciences*, *6*(3). doi: 10.2478/s13533-012-0177-9

Vasylchuk, Y. V., & Deutsch, C. V. (2015). *A Short Note on Optimal Kriging Grid Size relative to Data Spacing* (CCG Annual Report 17). Edmonton, AB: University of Alberta.

Vollgger, S. A., Cruden, A. R., Aillères, L., & Cowan, J. (2015). Regional dome evolution and its control on ore-grade distribution: Insights from 3D implicit modelling of the Navachab gold

deposit, Namibia. *Ore Geology Reviews*, *69*, 268-284. doi: 10.1016/j.oregeorev.2015.02.020

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function*. *Journal of the American Statistical Association1*, *58*(301), 236-244. doi: dx.doi.org/10.1080/01621459.1963.10500845

Wilde, B. J., & Deutsch, C. V. (2011a). *A New Way to Calibrate Distance Function Uncertainty* (CCG Annual Report 13). Edmonton, AB: University of Alberta.

Wilde, B. J., & Deutsch, C. V. (2011b). *Simulating Boundary Realizations* (CCG Annual Report 13). Edmonton, AB: University of Alberta.

Wilde, B. J., & Deutsch, C. V. (2012). Kriging and Simulation in Presence of Stationary Domains: Developments in B. In *Quantitative Geology and Geostatistics 17* (p. 12). Oslo.

Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, *105*(490), 713-726. doi: 10.1198/jasa.2010.tm09415

Xiaojun, W., Michael, Y., & Wang, Q. X. (2005). Implicit fitting and smoothing using radial basis functions with partition of unity. In *Proceedings - Ninth International Conference on Computer Aided Design and Computer Graphics, CAD/CG 2005* (Vol. 2005, p. 139-148). doi: 10.1109/CAD-CG.2005.50

Yamamoto, J. K., Kikuda, A. T., Koike, K., Campanha, G. A. d. C., Leite, C. B. B., Endlen, A., & Lopes, S. D. (2015). Determination of volumetric uncertainty for geological bodies detected by boreholes. *Measurement: Journal of the International Measurement Confederation*, *66*, 45-53. doi: 10.1016/j.measurement.2015.01.023

Appendix A

Software

Software implementations are an integral component of this research. The implementations range from self-contained Fortran binaries to python modules.

A.1 Boundary Modeling

The PU RBF implicit modeling techniques described in Chapter 3 and 5 are realized as standalone Fortran binaries following the GSLIB-style parameter-file implementation. Three programs are developed: `rbfdfmod`, `rbfiterref` and `rbfuncert`, for generalized interpolation, iterative refinement, and bandwidth of uncertainty calibration, respectively. The examples demonstrated in Chapter 3 and the case study presented in Chapter 5 utilize these tools for all implicit modeling steps.

These programs are suitable for small to medium sized projects. Commercial RBF-based implicit modelers permit boundary reconstructions for > 1 M data points using optimized sparse iterative solvers (Cowan et al., 2003). In this case geological boundaries are constructed one category at a time following the geological rules of superposition. The programs presented here use the direct-dense LU-decomposition based solver from LAPACK, and the K-category SDF modeling algorithm from Silva and Deutsch (2012b). For the 'direct' solver method (equivalent to global kriging), this is possible for up to 30 K data sites (tested on a machine with 16 GB of memory), but will require significant CPU and RAM resources even for > 15 K input data. However, the PU framework is implemented to decompose the domain so that many small (and efficient) subproblems are solved independently and recombined to the global solution. In the PU framework the resource requirements for very large datasets are modest, however, optimal overlap and partition parameters are still required for best results considering smoothness and model run times. For large projects with significant null sites, a keyout can be valuable for omitting locations that are not required to be modeled. Run times scale linearly with the number

of blocks in the evaluation grid. The block size in a grid is an important consideration; the grid should adequately resolve the smallest geological feature and be relevant to the mining process, but should also be reasonable from a CPU-time point of view (Boisvert, 2013). Further discussion on this is provided at the end of Chapter 5 (or in: Vasylchuk & Deutsch, 2015).

A.1.1 `rbfdfmod`

The parameter file for this program is shown below. Line 4-6 define the input data file: the columns for sample locations, the column for the indicator data to be modeled and the null values in the datafile. Line 8 and 9 define a keyout file, the column of the data in the keyout and the value at each location to keep. The keyout must be the same dimensions as the modeling grid, defined on lines 10-12. On line 13, the solver type can be chosen to either be 'direct' (1) or 'POU' (2). The following parameters on the same line define the bandwidth parameter to consider for each category when calculating the SDF. **Note:** on this line, a bandwidth parameter must be present for each category that is defined on line 7. Five different kernels can be used for interpolation; Gaussian, cubic, Wendland, linear and spherical. Details of each kernel are tabulated in Table 2.1. Line 14 also includes a nugget effect (NE) that is implemented for decreasing kernels only. This parameter should be used with caution; in the current implementation an exponent of 1.8 in instead of 2 is used to alleviate matrix instability with the Gaussian kernel. Introduction of even a small (0.05) NE to the interpolation problems has a large effect on the interpolation, which tends to result in un-honored data points.

```
1                          rbfdfmod
2                          ********
3   START OF PARAMETERS:
4   input.dat               -file with input dataset
5   1  2  3  4              - column for x, y, z, indicator data
6   -999                    - trimming value
7   3 2 3 5                 -ncats, category indicator
8   keyout.dat              -file with the keyout model
9   1  1                    - column for keyout variable, value to keep
10  50   0.5  1             -nx,xmn,xsz - interpolation grid definitions
```

```
11   50    0.5   1              -ny,ymn,ysz
12   50    0.5   1              -nz,zmn,zsz
13   1   0   0   0              -RBFinterp type (1:direct, 2:POU)-NOTE 1, tra..
14   1   0.001                  - RBF kernel (NOTE 2), Nugget Effect
15   0   350   350              - Support (0:manual,1:auto) - NOTE 2, support..
16   0                          -Use different anisotropy for each category? ..
17   0   0   0                  - Anisotropic orientations (str,dip,plunge) /..
18   1   1                      - Anisotropic ratios (minor/major, vertical/m..
19   1                          -Ouput 1:indicators, 2:signed distacnce funct..
20   outfl.out                  -file for interpolated output
21   *** BPT parameter section: ***
22   95   0.06   1              -Partition paramters: dpc, d overlap, weight ..
23    0.5 0.5 0.5               - final partition overlap in x, y, z directio..
24   0                          -read BPT parameters from file(1=Yes,0=No)
25   BPT_bounds.out             -file with partition boundaries
26   BPT_aniso_cat2_ref4.out    -file(s) with anisotropy for each category
```

The support is an important parameter for this implementation. For kriging, the range of continuity is modeled from the variogram along different orientations. The 'support' used in the RBF framework is synonymous with the range of the variogram, and often this support for RBF problems can be parameterized from the variogram range. Figure A.2 and A.3 show the affects of the support parameter for a boundary interpolation problem with the Gaussian and cubic kernels, respectively. The estimated support parameter for this domain modeled from the indicator variogram is roughly 20 m (Fig. A.1). As can be seen with the cubic kernel, the support parameter does not affect the interpolation. Conversely for the Gaussian kernel, with a very small support relative to the data configuration, the influence of data on the interpolated scalar field is clearly localized, and away from data the interpolant is zero. The idealized support parameter (~20 m) from the variogram range generates a smooth interpretation of the delineated categories, whereas increasing the support beyond the ideal value has some affect, but still largely produces consistent results.

The support parameter can also be estimated from the configuration of the data locations by

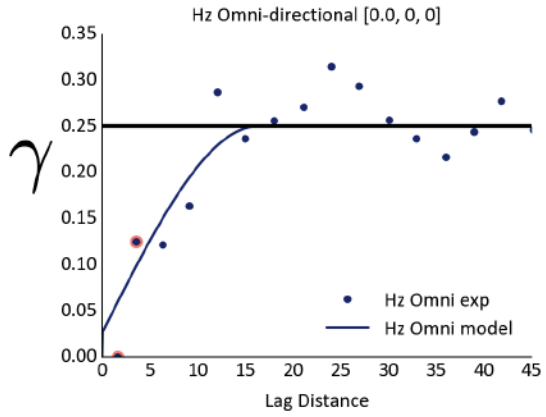


Figure A.1: Categorical omnidirectional variogram for the simple test domain shown in Figure A.2 and A.3

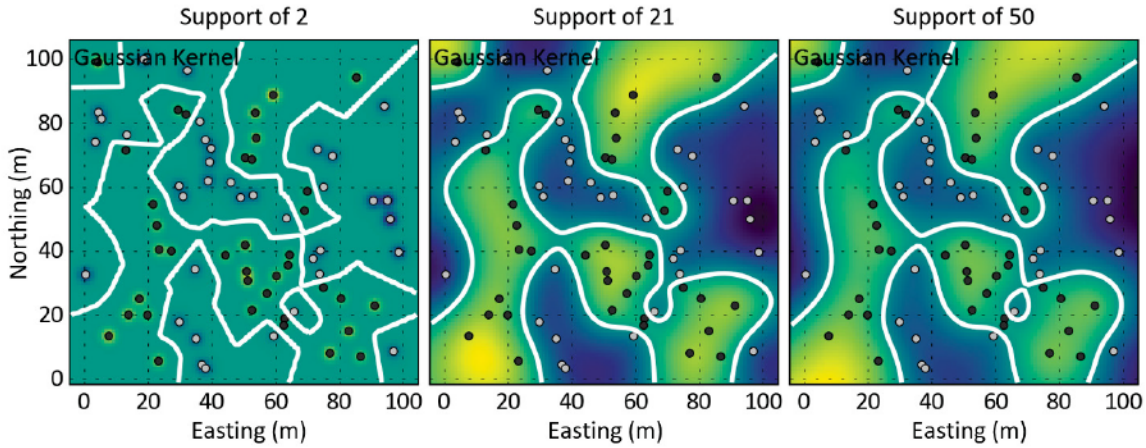


Figure A.2: Influence of different support parameters on the boundary location for the decreasing Gaussian kernel

finding the maximum distance from any 1 grid cell to the closest data location (Fasshauer, 2007). The support parameter estimated by this method for this simple dataset is 34 m. Line 15 controls the support parameters. Following this the anisotropy is parameterized with ratios obtained from $r1 = ahmin/ahmax$ and $r2 = avert/ahmax$. The extrapolation distance provides the ability to control how far from a data location will be interpolated in the domain; this will also affect the estimation of the support parameter for the domain. This parameter is designed to limit extrapolation, and to reduce run times in the absence of a keyout for oddly configured data arrangements relative to the axis aligned grid.

As discussed above the kernel, anisotropic parameters and support distance define the required properties of the RBF interpolation problem. The parameter on line 16 controls how

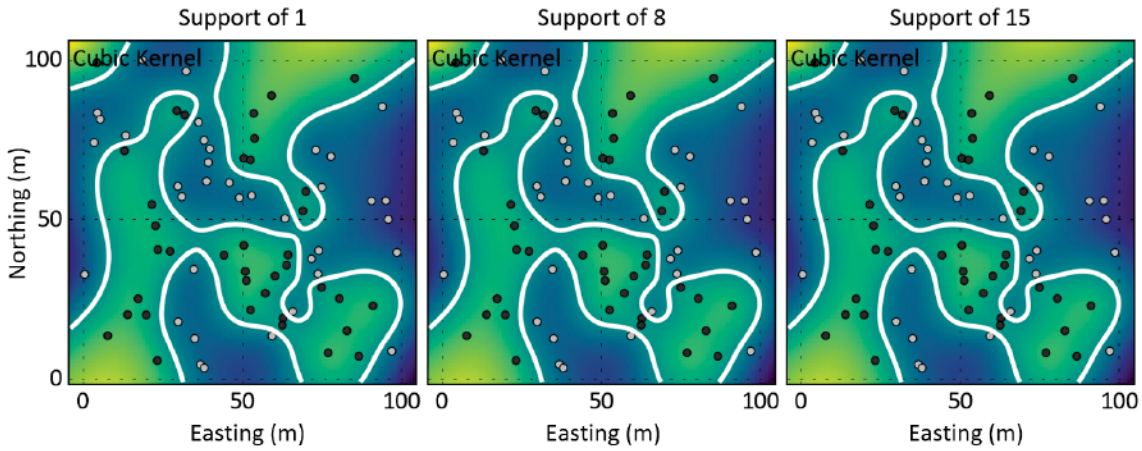


Figure A.3: Influence of different support parameters on the boundary location for the increasing cubic kernel

(a) Anisotropy File Format

```
BPT Aniso Cat  2
6   40
partNo
strike
dip
plunge
r1
r2
  1   243.50908        0.0000000        0.000000
  2   238.20627        0.0000000        0.000000
```

(b) Boundaries File Format

```
BPT Boundaries
8   40   2
partNo
minx
miny
minz
maxx
maxy
maxz
support
  1   30.215000        12.037000        0.0000000
  2   30.215000        59.296625        0.0000000
```

Figure A.4: Format of the anisotropy and boundary files for defining a set of partitions from the scattered data points

many sets of anisotropic parameters are to be used in the K-category interpolation problem. If line 16 is 1, then a set of anisotropic parameters are required for each category that is defined on line 7. Line 19 defines the output format, which can be either a categorical model, or the interpolated SDF for each category. Setting this option to 2 is useful when considering the SDF with bandwidth parameters, so that different isosurfaces viewed or used to calculate the volumes corresponding to different levels in the SDF. At this point it is worthy to note that the K-category interpolation can be completed completely within `rbfdfmod`, however, if many categories with abundant data and large grids are being interpolated, consider scripting `rbfdfmod` in parallel by exporting the SDF for each category, and post-processing K-SDF models to generate the required indicator grid.

Lines 22-26 define all required parameters for partitioning, including the ability to read in the boundaries and anisotropy output from the `rbfiterref` program. Line 22 defines the data-per-center, which guides the decision on whether or not a partition is split, the 'data overlap',

which is the overlap applied to the partitions *during* recursive splitting and the weighting function type calculated on overlapping sites. The data-per-center parameter is self explanatory - partitions are split until the number of data contained within a partition is less than this number. The data-overlap parameter controls the % of data allowed to overlap between partitions during the splitting stage of the algorithm. In general the data overlap parameter should be small, < 0.15, or recursive problems may arise for some data configurations, which will be indicated in the output. The final overlap parameters defined on line 23 control the post-splitting overlap applied by expanding partitions along the diagonal according to the percentage in each direction. This can be used to ensure there is sufficient overlap between partitions if no data-overlap is used; generally some mix of data overlap and post-partitioning overlap should be used. Additional post-partitioning overlap may be required in the case of extreme differences in anisotropy between adjacent partitions.

Line 24-26 define the use of partitioning parameters output from `rbfiterref`; set line 24 to 1 to read in and use these files, or 0 to generate a new set of partitions according to the partitioning parameters on the lines above. The file defined on line 25 contains the partition number, the lower and upper corner of the partition, and the estimated support distances. The file on line 26 contains the anisotropic properties for each of the partitions. **Note:** an anisotropy file must be defined for each of the categories defined on line 7. The format of the boundaries and anisotropy files are shown in Figure A.4. The second line of the anisotropy file contains the number of variables and the number of partitions found in the domain. Each entry in this file contains the strike, dip, plunge, r1 and r2 corresponding to the numbered partition. The second line in the boundaries file contains the number of variables, the number of partitions and the weighting function to be used for interpolation, which can be either 1 or 2. Each entry in this file defines the extent of a partition based on the lower and upper corners of the bounding box. The final column defines the local support parameter to consider in each partition.

A.1.2 `rbfiterref`

This program is used to generate the iteratively refined local anisotropy for a boundary model considering the partitioning parameters, the `iterref` parameters and the `imorient` parameters.

The current program uses % changed as a stopping criteria, but in the future a jackknife or cross validation metric could be considered to evaluate if a partition is 'refined'. In the future a randomized optimization of the anisotropy angles based on cross validation could also be considered. Several of the key parameters used in this program are identical to those described in `rbfdfmod` above and are not repeated.

Line 13-14 define the partitioning parameters and line 15 defines the parameters for the iterative refinement. The target number of refinements can be defined, however according to the % difference stopping criteria defined on the same line, there may be fewer actual refinements than the defined number. On the same line, the max r1 and max r2 parameters define maximum anisotropy that is allowed, which can be used to permit or prevent excessive anisotropic magnitudes from being inferred to each partition. Line 16 defines the number of layers, the window size, and the layer for the final orientations that are extracted with the `imorient` subroutines. For more detailed information on these parameters, see Martin, Manchuk, and Boisvert (2015). It should be noted that in the `imorient` subroutines the anisotropic magnitudes are calculated from the singular values in SVD which reflects relative dominance of a particular orientation in a window surrounding each location. Although these ranges of anisotropy are related to the dominance of orientations in the boundary model, they are not calculated from the point data. Line 17 allows definition of the type of model to refine orientations from (0: SDF model, 1: indicator model), and the parameters for IDW interpolation for orientations between layers. In general the SDF carries more orientation information than the indicator model, especially away from boundaries. Most commonly the default IDW parameters for `imorient` are sufficient.

The final parameters of importance are on line 24 and 26. The files on these two lines define the project prefix and output extension for all files from this program. For example, on line 24 if '`Data/BPT.out`' is defined, the program will export a series of files in the '`Data/`' folder with the prefix '`BPT`', i.e. '`Data/BPT_bounds.out`' and '`Data/BPT_diffs.out`'. **Note:** in this case the folder '`Data/`' must exist. Similarly, on line 26, a file '`Data/outfl.out`' will generate a series of files '`Data/outfl<ref>.out`' where ref will take values between 1 and the number of refinements that are completed during the program.

```
                          rbfiterref
                          **********
START OF PARAMETERS:
input.dat                 -file with categorical data
1  2  3  4                - column for x,y,z, indicator data
-999  1                   - trimming value, NCPU for parallel solve
3 2 3 5                   -ncats, category indicator
keyout.dat                -file with the keyout for the geological model
1  1                      - column for keyout variable, value to keep
50   0.5  1               -nx,xmn,xsz - definitions for interpolation grid
50   0.5  1               -ny,ymn,ysz
50   0.5  1               -nz,zmn,zsz
95  0.06  1               -Partition paramters: dpc, data overlap, weight ..
  0.35 0.35 0.35           - final partition overlap in x, y, z directions
5   6     0.1   0.1       -num refinements, %diff stopping, maxr1, maxr2 -..
3   10    1               -Imorient parameters: nlayers, window, final lay..
  0  2.  25                - orients from SDF model (0) or ind model (1), i..
1  0.000                  -RBF kernel (NOTE 2.5), nugget effect
0                         -Starting Anisotropy, different for each categor..
0  0  0                   - Anisotropic orientations (str,dip,plunge) / Ca..
1  1                      - Anisotropic ratios (minor/major, vertical/majo..
1  350  350               - Support (0:manual,1:auto), support distance, e..
1                         -writeout binary partition LVA parameters (1=Yes..
BPT.out                   - BPTflnames (becomes `BPT` project prefix) - NO..
1                         -Ouput 1:indicators, 2:signed distance function
outfl.out                 -file for interpolated output - NOTE 5
```

A.1.3 `rbfuncert`

This program is used to train the C-parameter from Wilde and Deutsch (2012) using the PU RBF interpolation implemented in this work. All steps of the workflow are contained in this program to generate the required summary table that can help make the decision as to the

correct bandwidth of uncertainty to consider.

Several of the input parameters to this program are identical to the above `rbfdfmod` and are not repeated. Notably, for 3D datasets the drill hole ID column is required (line 5) to perform jackknife analysis. Line 11 defines the number of random datasets, the number of C-parameters, the random seed and the number of CPU's to use. All random jackknife datasets are generated inside this program. Line 12 defines the maximum C-parameter(s) that will be considered in generating the random C-parameters for each category. A first approximation for this parameter may include the data spacing.

Similar to `rbfdfmod`, this program can read in a set of partitioning files generated with `rbfiterref` to define the partitioning parameters and local anisotropy that is refined from that program.

```
                        rbfuncert
                        *********
START OF PARAMETERS:
input.dat                  -file with input indicator dataset
1  2  3  4  5              - column for dhid(0 if 2D), x,y,z, indicator..
-999                       - trimming value
2 2 3                      -ncategories, categories
50   0.5  1                -nx,xmn,xsz - interpolation grid definitions
50   0.5  1                -ny,ymn,ysz
50   0.5  1                -nz,zmn,zsz
15   25   69069  4         -num random datasets, num C parameters, rsee..
200.  500.                 -max c parameter, for each categorical varia..
0.25                       -proportion DHS/samples to remove
1                          -RBFinterp type, (1:direct, 2:PU) - NOTE 1
1   0.001                  -RBF kernel (NOTE 1.5), nugget effect
0   350   350              - Support (0:manual,1:auto) - NOTE 2, suppor..
1                          -different anisotropic parameters for each c..
0  0  0                    - Anisotropic orientations (str,dip,plunge)
1  1                       - Anisotropic ratios (minor/major, vertical/..
0  0  0                    - Anisotropic orientations (str,dip,plunge)
1  1                       - Anisotropic ratios (minor/major, vertical/..
cparamsummary.out          -file for summary output - NOTE 4
*** BPT parameter section: ***
95  0.06  1                -Partition paramters: dpc, d overlap, weight..
  0.35 0.35 0.35           - final partition overlap in x, y, z directi..
0                          -read BPT parameters from file(1=Yes,0=No)
BPT_bounds.out             -file with partition boundaries
BPT_aniso_cat2_ref4.out    -file(s) with anisotropy for each category
BPT_aniso_cat3_ref4.out    -file(s) with anisotropy for each category
```

A.2 Spatial Clustering Python Package

This software is distributed as a python library using a `.whl` file, which contains all source code, a list of dependencies that are resolved on installation, and a set of 'entry points' that expose functionality in the package to GSLIB-style command-line callable programs (see below). If utilizing this package with an existing python installation, packages upon which this package depends (e.g., numpy) may require updates. In the case of unresolvable differences between this package and others installed to the target Python environment, a unique Python environment for this project, generated with `conda create -n <envname> && activate <envname>`, is recommended.

All requirements for installation are contained in the `spatialcluster<...>.whl`. The package can be installed to the current python distribution by opening a command prompt in the same folder as the `.whl` file and issuing the following command: `pip install [.\]sp` and pressing 'tab' to complete the sentence (NOTE: the '.\' preceding the `.whl` file may be required if using 'powershell' on windows). Ensure that the machine either already has all required dependencies (this requirement is met with the Anaconda Python distribution), or has web access so that dependencies can be automatically resolved. If working with environments, remember to activate the target environment prior to installation and when running any scripts or notebooks depending on this package.

Functionality provided in this package can be accessed in two ways. First, all classes and functions can be called from Python providing all functionality developed in this package to be accessed from a Python-based workflow; this is the most flexible use case since plotting and other data analysis tools can be freely mixed with the developed software. Second, the pip-installation of this package to the local Python distribution generates several GSLIB-style command-line 'hooks' (Note: these executables are not portable) that wrap the main functionality of each clustering algorithm. Following pip-installation, the executables `spatialclusterex`, `acclus`, `acens`, and `dssens` are available on the path. For example, 'which acclus' issued to the bash shell should point to an acclus.exe in the `%PYTHONDIR%/scripts` directory.

The parameter files for the programs contained in this package are shown below and outlined in more detail in this section. Detailed usage, both the GSLIB-style executable and the

recommended Python-only method, are provided as a set of Jupyter notebooks in the examples folder. These examples are automatically installed alongside the main package in the `spatialcluster/examples` folder. These Jupyter notebooks also serve as the documentation and the guide to usage. A simple 2D test dataset is provided to illustrate the functionality embedded in this package. Note: the executable `spatialclusterex` will launch a jupyter notebook in this folder so that examples can be viewed and utilized.

A.2.1 `acclus`

```
1         ACCluster
2         ----------
3
4     START OF DATA:
5     datafile.dat    - file with the input dataset
6     1  2  0         - columns for dh, x, y, z data
7     3  4  5  6      - columns for variables (implicit nvar)
8     outfile.out     - file for clustering output
9     1               - append output to input file? (0=No, 1=Yes)
10
11    START OF AC:    # autocorrelation settings
12    25              - number of nearest neighbors
13    0  0  0         - search anisotropy/ ang1, ang2, ang3
14    500 500 500     - range1, range2, range3
15    kmeans          - cluster method: one of `kmeans`, `gmm`, `hi..
16    1               - autocorrelation metrics (0=None, 1=Morans, ..
17    5               - number of clusters
```

The first block in the `ACCLuster` parameter file is the `DATA` block which specifies the filename, x, y, z columns, the input variable columns, the output file and whether or not the results of clustering should be appended to the output file. These lines are self explanatory. The `AC` block defines the parameters for autocorrelation-based clustering. The number of nearest neighbors on line 12 is what is considered in the spatial weighting, more neighbors has more influence

from distant points. The search is a K-nearest neighbors search. The search anisotropy defined on lines 13 and 14 should be chosen to account for differences between versus along drill holes (in 3D). The clustering method on line 15 is a string, one of 'kmeans', 'gmm' or 'hier', reflecting the final clustering method that will be used on the generated autocorrelation dataset. The autocorrelation metric is one of 0, 1, 2 for None, morans and getis, respectively. If 'None', the program simply performs the chosen clustering on the multivariate dataset. Finally, line 17 specifies the number of clusters for output.

A.2.2 acens

```
1       ACEnsemble
2       ----------
3
4   START OF DATA:
5   datafile.dat        - file with the input dataset
6   0  1  2  0          - columns for dh, x, y, z data
7   3  4  5  6          - columns for variables (implicit nvar)
8   outfile.out         - file for clustering output
9   1    1              - save all reals? recode clusters? (0=No, 1=Y..
10  1                   - append output to input file? (0=No, 1=Yes)
11
12  START OF AC:        # autocorrelation settings
13  25                  - number of nearest neighbors
14  0  0  0             - search anisotropy/ ang1, ang2, ang3
15  500 500 500         - range1, range2, range3
16  kmeans              - cluster method: one of `kmeans`, `gmm`, `hi..
17  1  2                - autocorrelation metrics (0=None, 1=Morans, ..
18  0.5 0.5             - autocorrelation props
19
20  START OF ENS:       # ensemble settings
21  523151              - random seed
22  100                 - num ensemble
```

```
-1                  - min number of variables, -1 to use all
4  6                - final nclus, target nclus
0.0 0.15            - min and max proportion to remove
0.001 0.999         - min and max proportion in found clustering
spec                - final consensus method, `spec`(tral) or `hi..
```

The DATA and AC blocks for ACEnsemble are similar to ACCluster except that multiple autocorrelation metrics can be chosen to be randomly used in the ensemble. The ENS block specifies the settings of the ensemble of clusterings. The number of clusterings in the ensemble is chosen up front. Randomization with the random subspace is controlled by specifying a number of samples to use and the min and max proportion of samples to remove in each clustering. Furthermore, it is possible to control how much or little of a given cluster is permissible in a clustering, for example, to ensure that clusters are equally sized or may be very different in size. Finally, the final consensus method defaults to spectral clustering for efficiency reasons. If 'hier' is chosen, then hierarchical clustering with Wards linkage is applied.

A.2.3 dssens

```
        DSSEnsemble
        -----------

   START OF DATA:
   datafile.dat         - file with the input dataset
   1  2  0              - columns for x, y, z data
   3  4  5  6           - columns for variables (implicit nvar)
   outfile.out          - file for clustering output
   1    1               - save all reals? recode clusters? (0=No, 1=Y..
   1                    - append output to input file? (0=No, 1=Yes)
   1                    - number of parallel processes

   START OF DSS:        # dual-space-search settings
   25                   - number of nearest neighbors
```

```
15      10                      - number of neighbors merged at spatial search
16      0  0  0                 - search anisotropy/ ang1, ang2, ang3
17      500 500 500             - range1, range2, range3
18
19      START OF ENS:           # ensemble settings
20      512671                  - random seed
21      100                     - number of clusterings to generate
22      4  6                    - final nclus, target nclus
23      0.001 0.999             - min and max proportion in found clustering
24      spec                    - final consensus method, `spec(tral)` or `hi..
```

The DS clusterer has similar input parameters to the other two algorithms. Of note, in the DS section the number of neighbors and number merged during the spatial search are the two parameters that have the greatest impact on the spatial contiguity of the final clusters. Anisotropy should again be chosen to ensure that disparity in sample density with orientation is accounted for.

Appendix B

Chapter 6 Supporting Figures

B.1 Porphyry Case Study

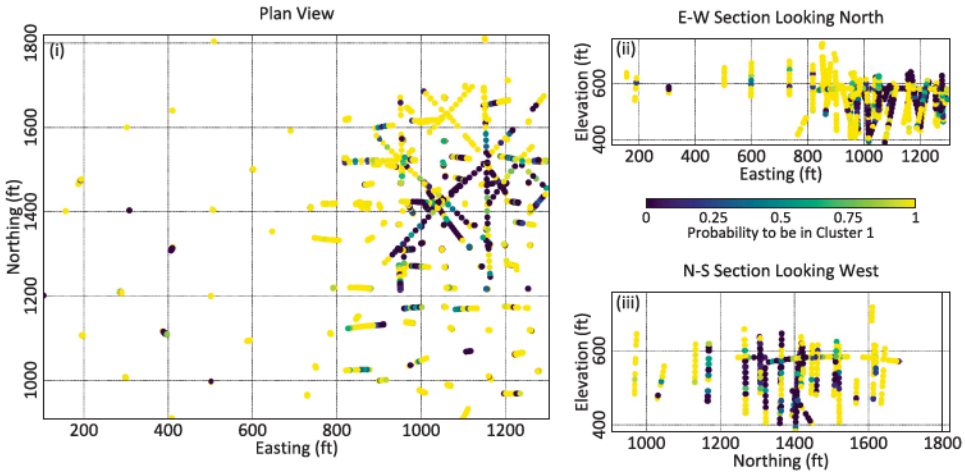


Figure B.1: Probability to be cluster 1 in the porphyry dataset.

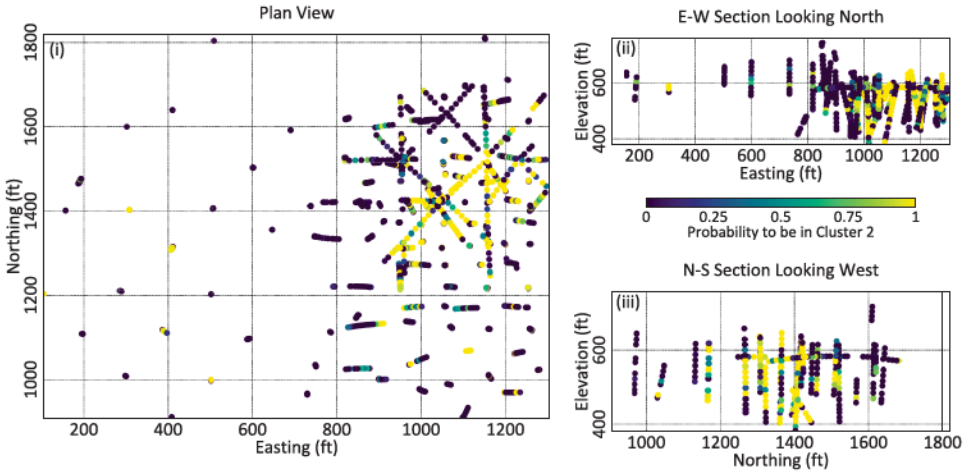


Figure B.2: Probability to be cluster 2 in the porphyry dataset.

B.2 Oilsands Case Study

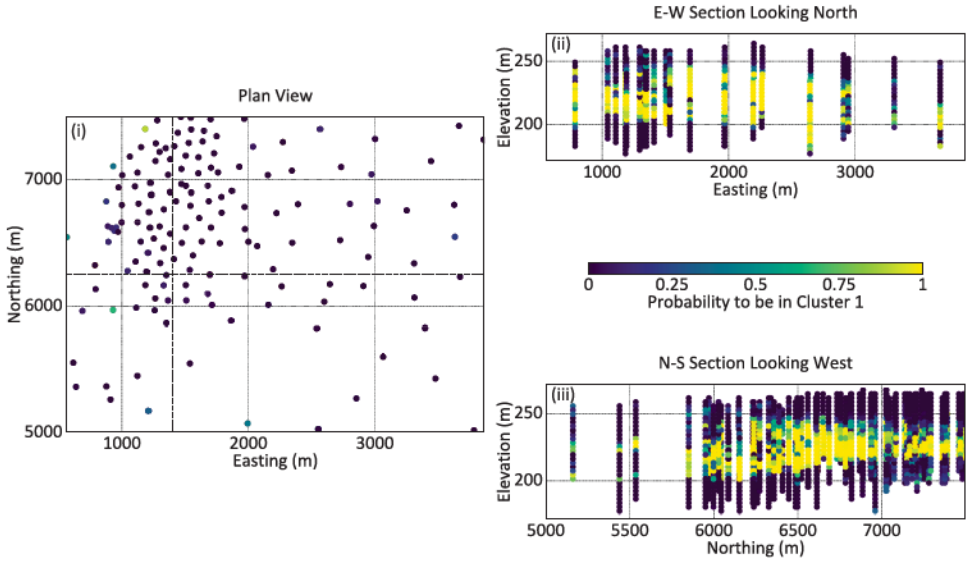


Figure B.3: Probability to be cluster 1 in the oilsands dataset.

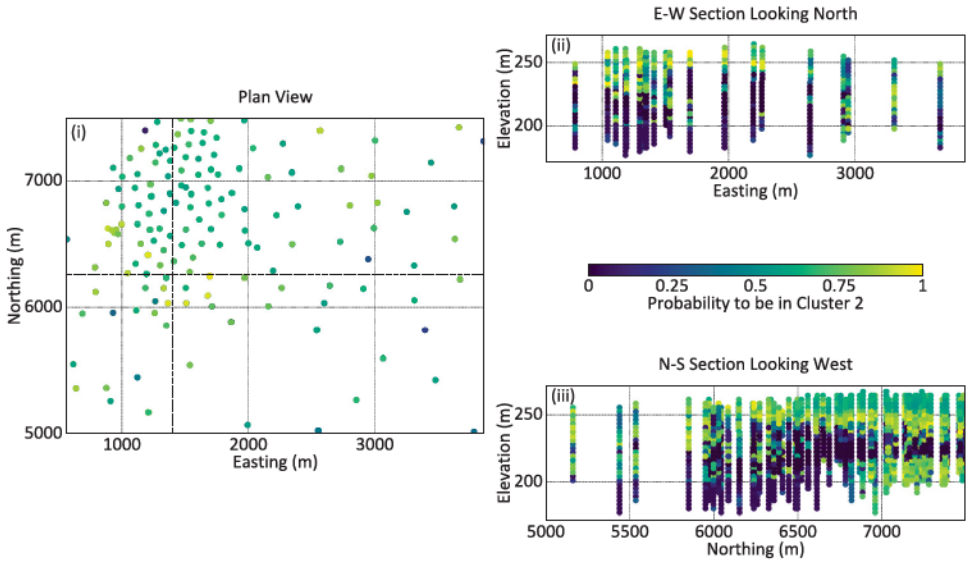


Figure B.4: Probability to be cluster 2 in the oilsands dataset.

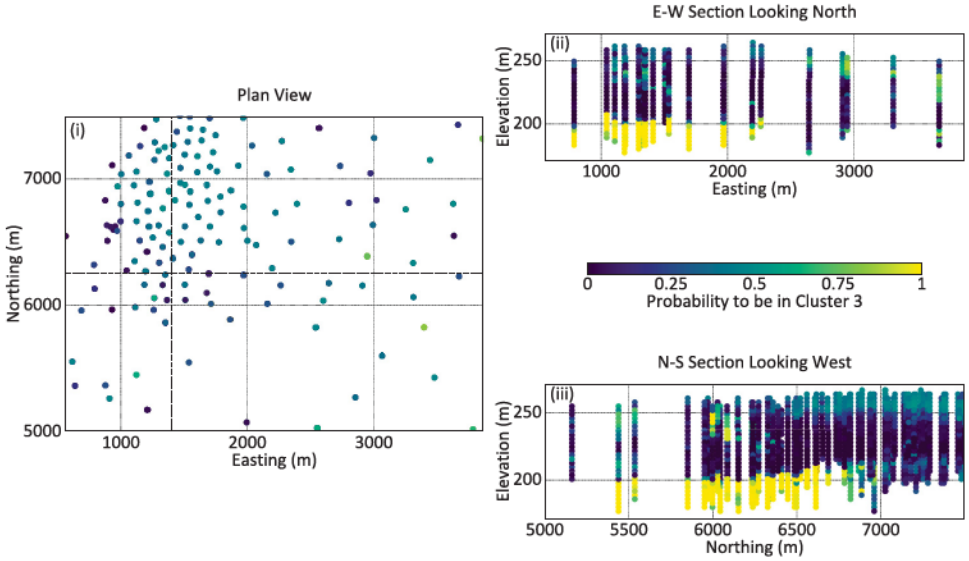


Figure B.5: Probability to be cluster 3 in the oilsands dataset.

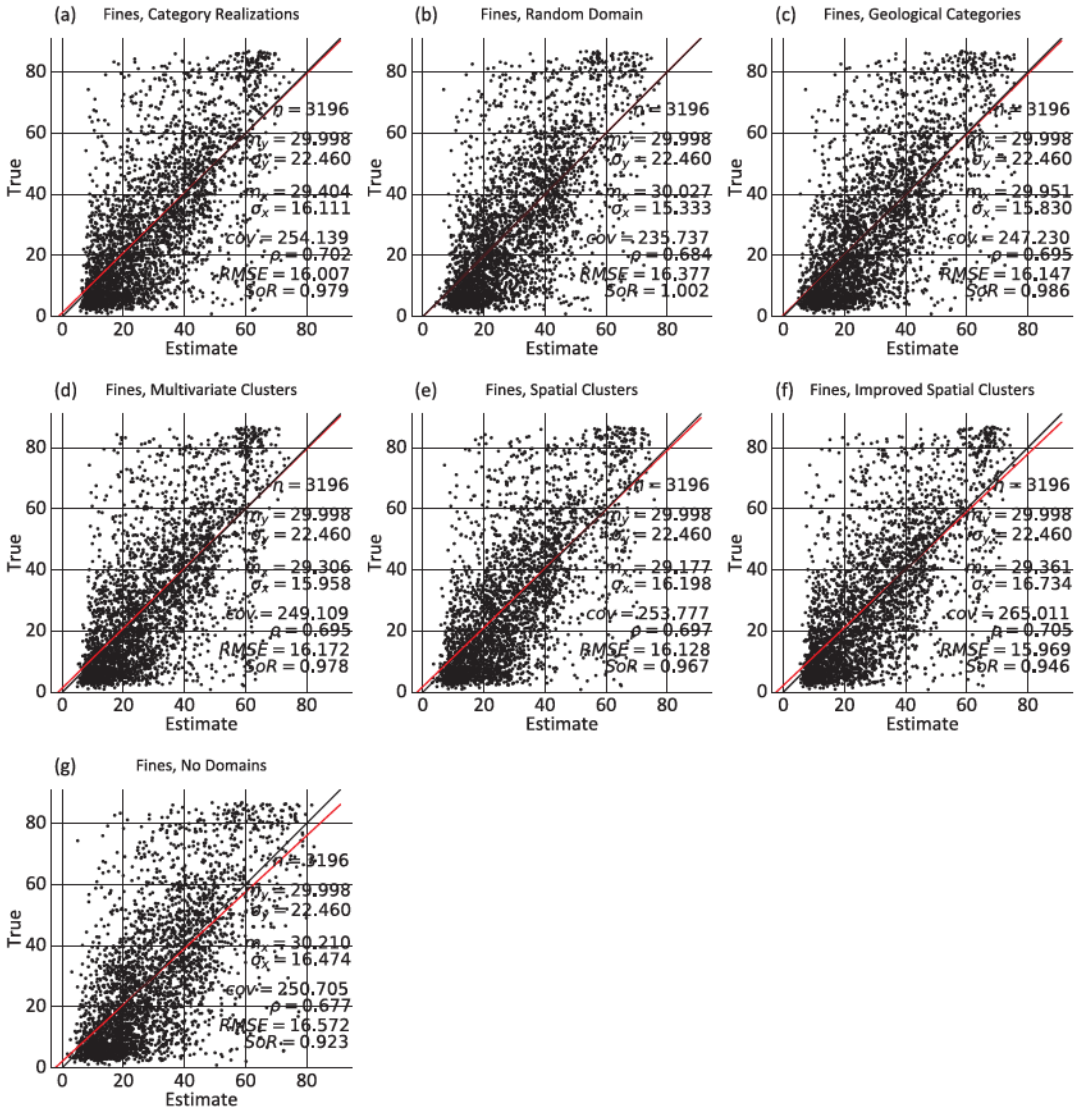


Figure B.6: Cross plot between the E-type estimate and the true values for fines from the oilsands dataset.

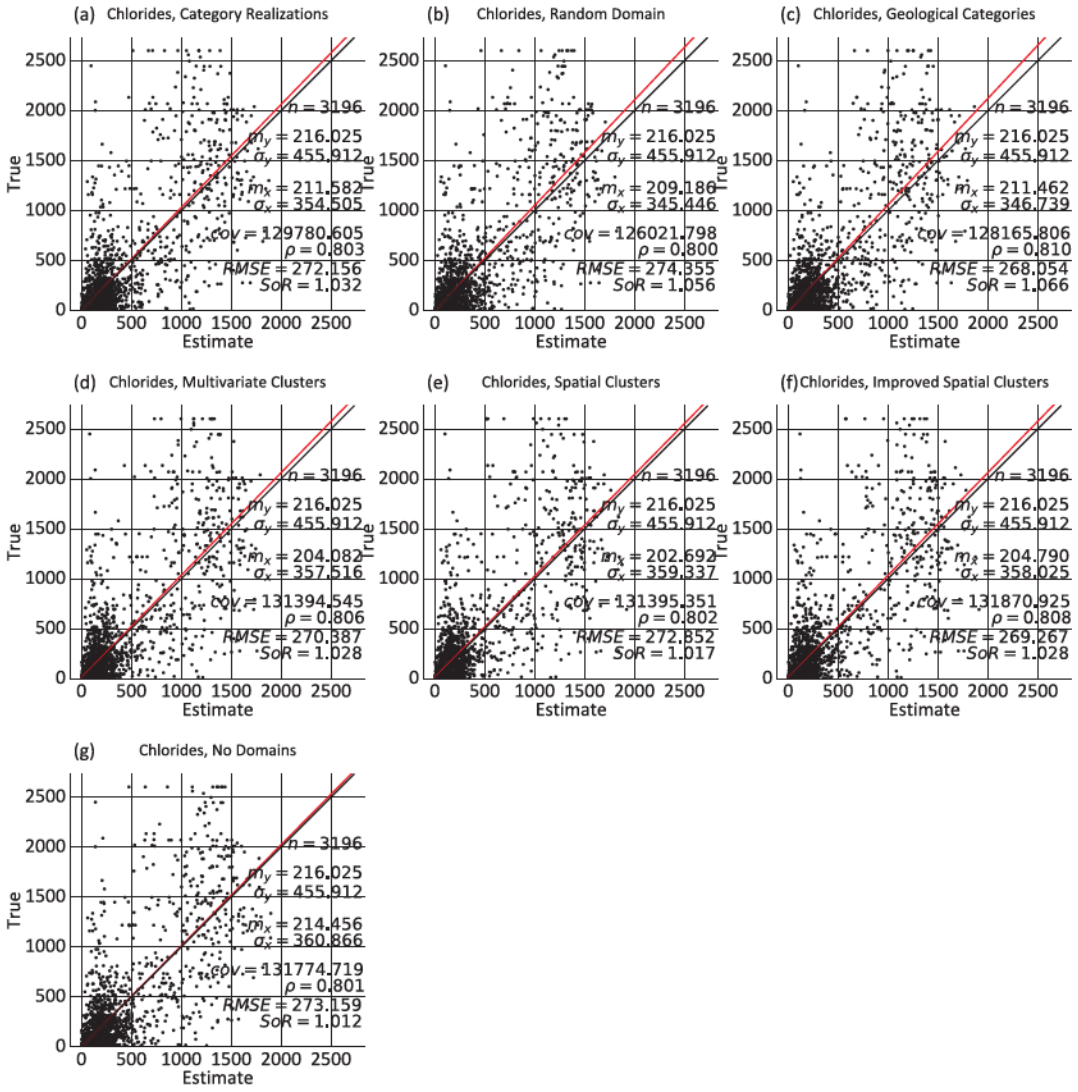


Figure B.7: Cross plot between the E-type estimate and the true values for chlorides from the oilsands dataset.

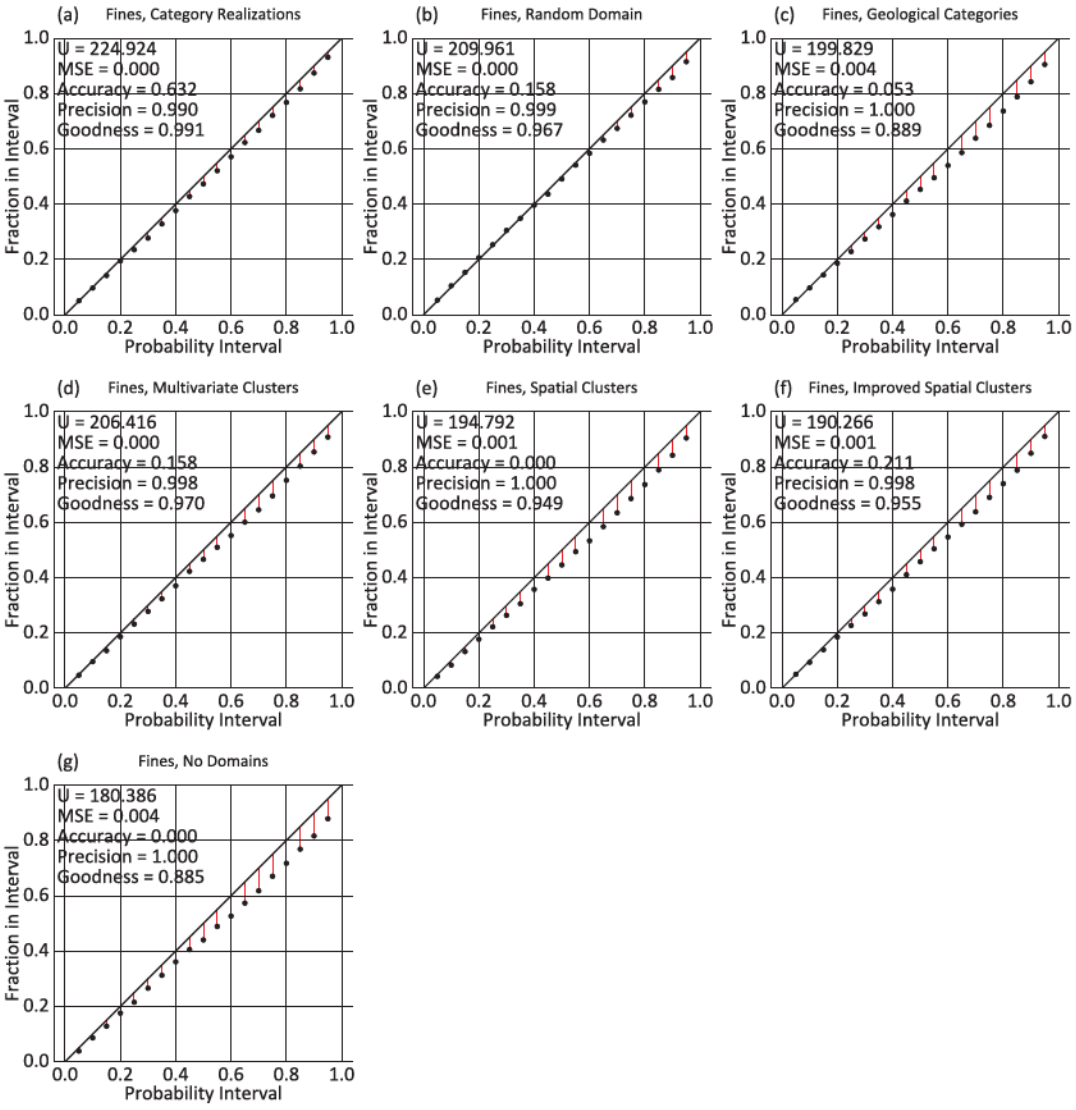


Figure B.8: Accuracy plots generated from the K-fold validation analysis for fines.

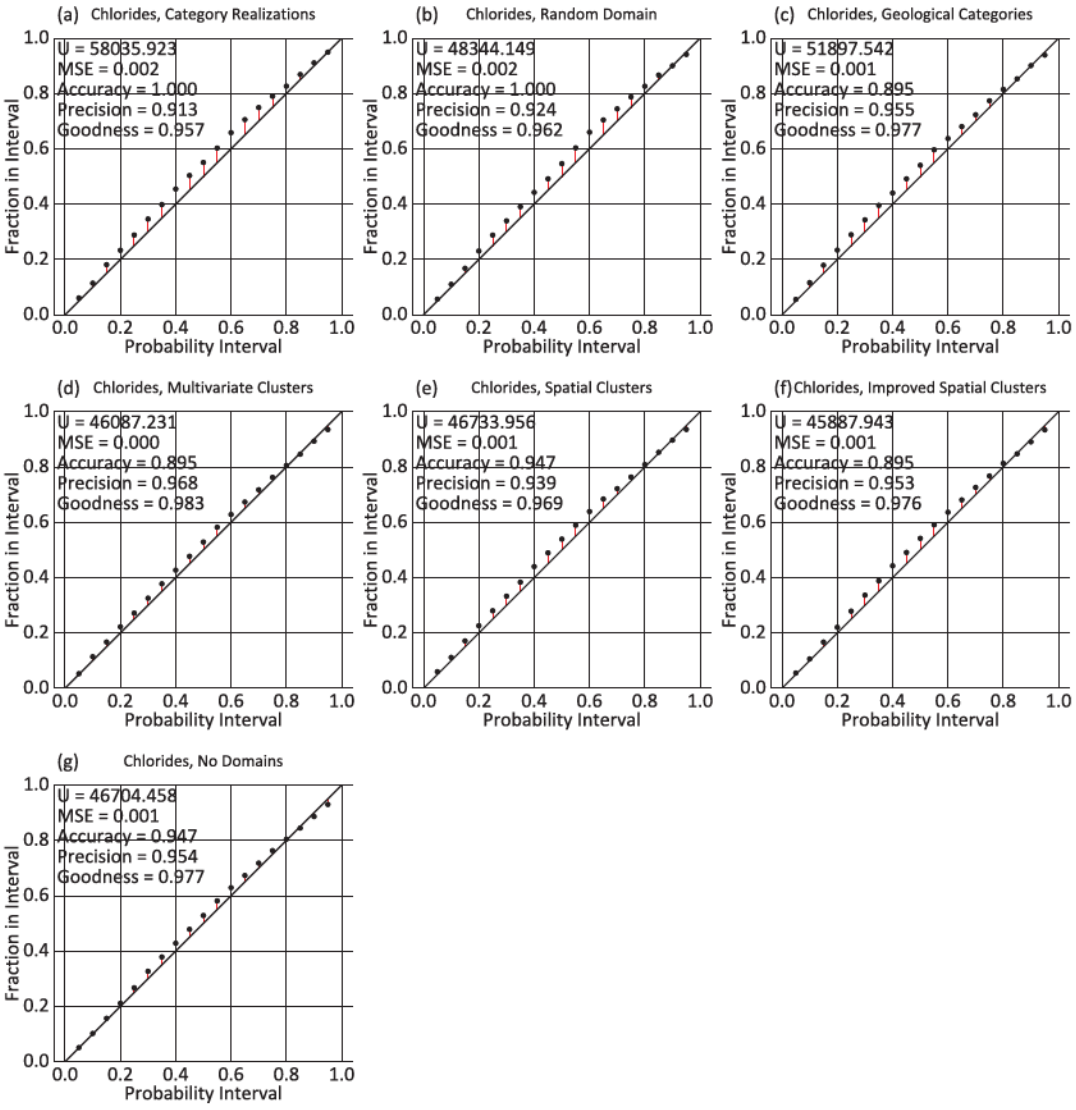


Figure B.9: Accuracy plots generated from the K-fold validation analysis for chlorides.

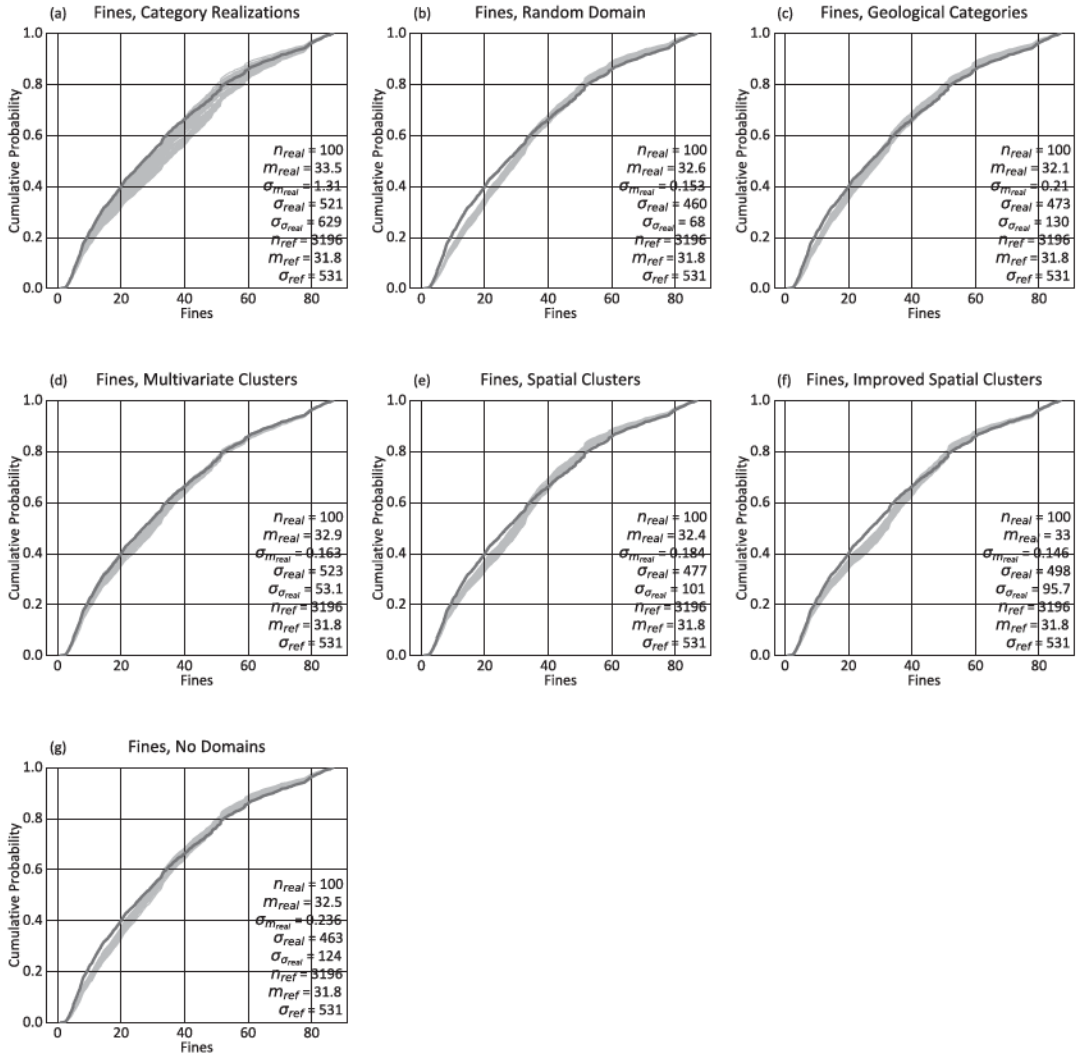


Figure B.10: Histogram reproduction plots for fines from the oilsands dataset.

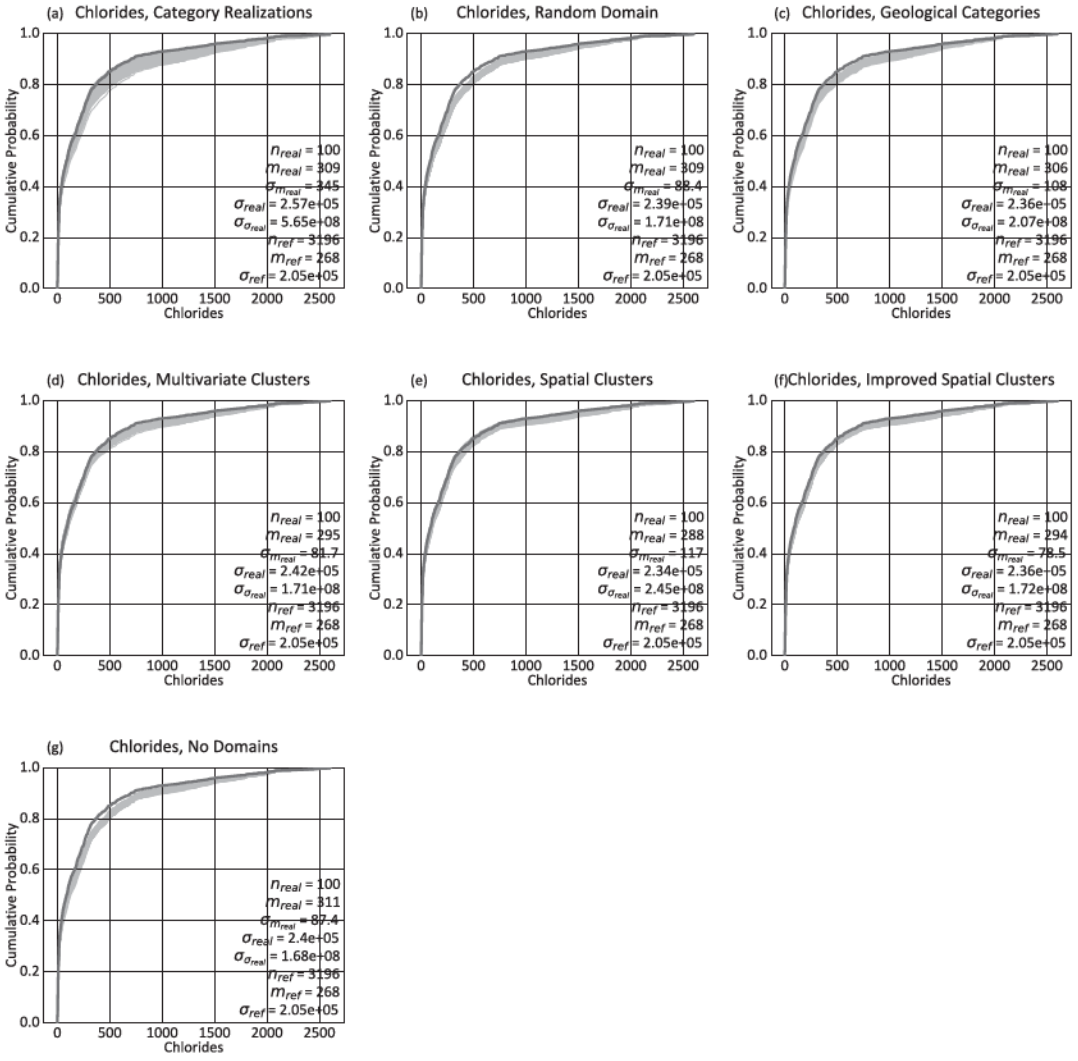


Figure B.11: Histogram reproduction plots for chlorides from the oilsands dataset.

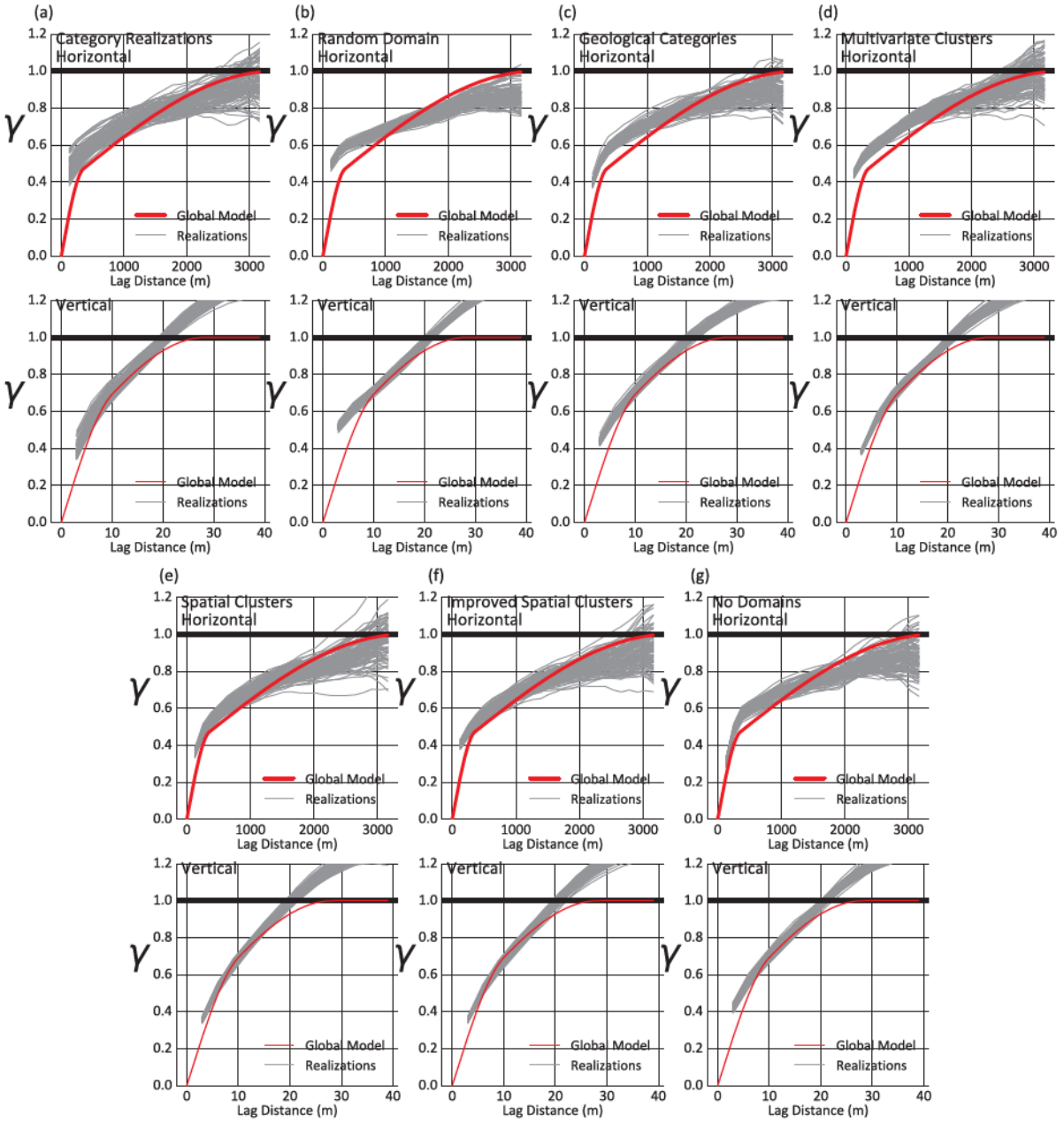


Figure B.12: Variogram reproduction for realizations of fines generated by each method. Horizontal reproduction shown in the top and vertical shown below for each set of stationary domains (a-g).

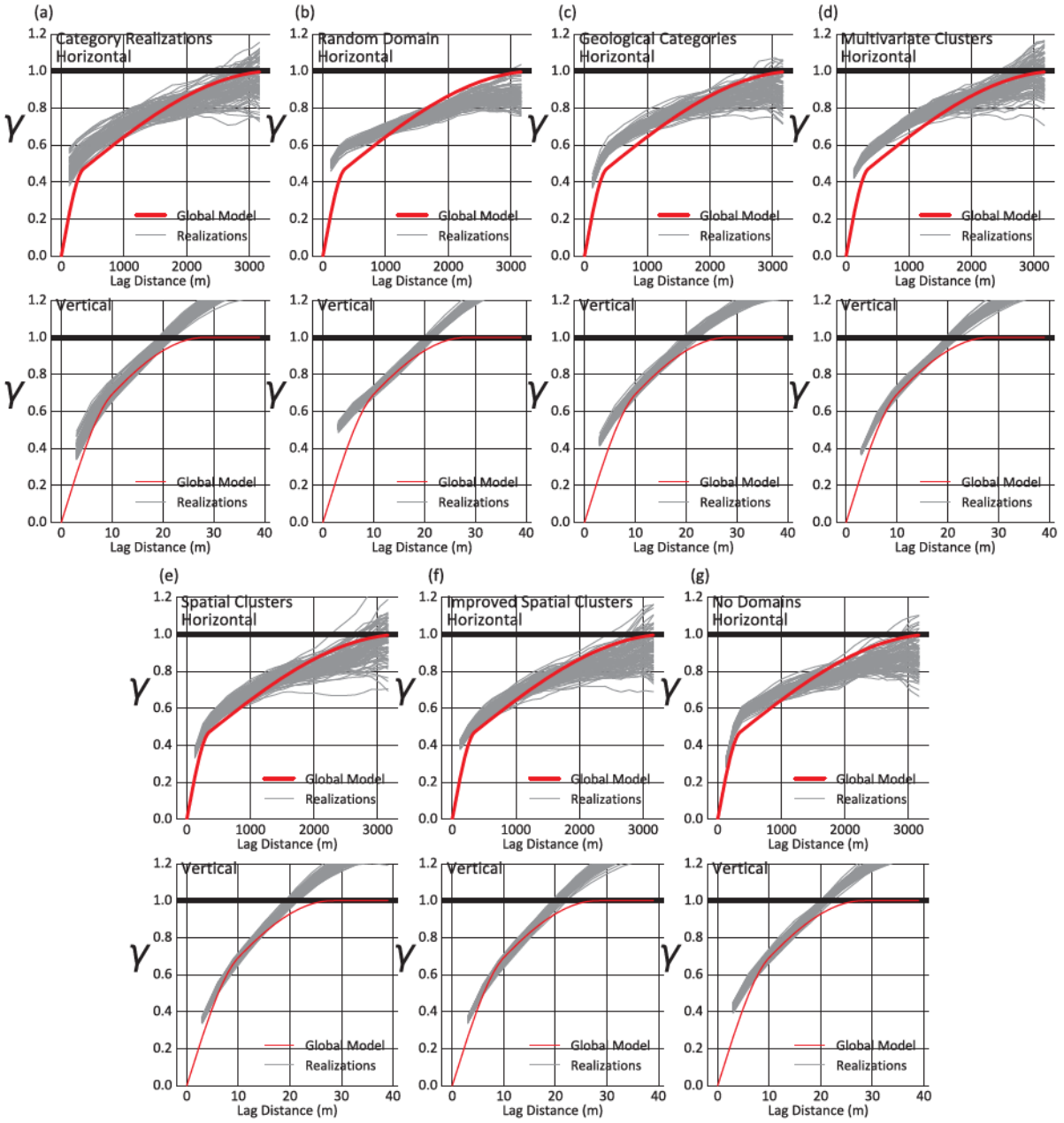


Figure B.13: Variogram reproduction for realizations of chlorides generated by each method. Horizontal reproduction shown in the top and vertical shown below for each set of stationary domains (a-g).