Managing Complex Multivariate Relations in the Presence of Incomplete Spatial Data

by

Ryan Matthew Barnett

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

Evaluating the process performance of mining and petroleum operations requires numerical geological models of many related rock properties or variables. Taken together, they provide a characterization of the geologic deposit that forms the basis for engineering design and decision making. Complex multivariate features such as compositional constraints and non-linearity often exist between geological variables and may have a large impact on process performance. This poses a problem for geostatistical modeling, where popular techniques do not capture complex relations. The data could be transformed to be suitable for modeling before using back-transformations to reintroduce the original complexity. Unfortunately, no sequence of available transforms will consistently remove all complex features from a large number of variables. The first contribution of this thesis is a transformation for removing complexity and correlation from data of practical size and dimension. This facilitates independent geostatistical modeling, before the back-transform restores the original relations.

Multivariate transformations may only be applied to data observations that sample all of the variables under consideration. This creates another significant challenge, as it is common for geological data to sample subsets of the variables. Practical solutions will exclude the incomplete observations from transformations, or use basic regression to infer (impute) the missing variables. These approaches usually have consequences, however, in terms of global bias, local accuracy and the reproduction of key properties. The second contribution of this thesis is methodology for the effective imputation and geostatistical modeling of incomplete data. Missing data theory is integrated with geostatistical algorithms to develop imputation methods that are suitable for geological data. Uncertainty of the imputed values is transferred through a modified geostatistical workflow.

These two contributions cumulatively simplify and improve the modeling of potentially complex and unequally sampled geological variables. Their value is demonstrated using real geometallurgical data and associated mine project decision making.

To my parents, Ivan and Nancy,
and to my grandparents, Kenneth and Pauline,
for your love and support at every step.

Acknowledgements

I would like to thank my thesis supervisor, Dr. Clayton Deutsch, for his invaluable instruction, ideas and enthusiasm. The opportunity to learn from Clayton has been one the great fortunes in my life.

I have enjoyed and benefited from working with many students and research associates at the Centre for Computational Geostatistics. I would especially like to thank Jared Deutsch and Dr. John Manchuk, for their help with all of those theory and coding problems along the way.

The financial support of the Centre for Computational Geostatistics member companies, the Natural Sciences and Engineering Research Council of Canada, and Shell Canada was essential to this research.

This would not have been possible without the love and encouragement of my girlfriend Julie, my sisters Amy and Lisa, and the rest of my wonderful family. Nor without my friends, who were an essential source of distraction when I needed it.

Contents

List of Tables

List of Figures

List of Symbols

Symbol	Definition	First Use
K	number of variables	9
i, j	variable indices $(i, j = 1, ..., K)$	9
n	number of data observations	9
α, β	indices for observations or model locations	9
$\mathbf{Z}$	nxK matrix of sampled data	9
$\mathbf{u}$	spatial coordinate vector	9
A	stationary modeling domain	9
Z	random variable in original units	9
F	cumulative distribution function(CDF)	10
$\mathbf{h}$	spatial vector that separates two locations	10
h	distance that $\mathbf{h}$ spans	10
C_{ij}	covariance between the i^{th} and j^{th} variables	10
μ	mean of a random variable	11
σ^2	variance of a random variable	11
$\gamma(\mathbf{h})$	semivariogram value for lag $\mathbf{h}$	11
G	CDF of a standard Gaussian variable	12
Y	transformed random variable	12

$\mathbf{X}$	nxK data matrix before Gaussianization	72
ρ'	Variable loading	73
$\boldsymbol{\theta}$	Kx1 unit vector for projecting multivariate data	74
$\mathbf{p}$	nx1 non-Gaussian projection, $\mathbf{p} = \mathbf{X}\boldsymbol{\theta}$	74
I	Projection index, or non-Gaussianity test statistic	74
M	number of bootstrap distributions	77
$\mathbf{I}$	MxK distribution of I for PPMT stopping criteria	77
N_O	number of ore classifications	144
O_i	ore classification, $O_i, i = 1, ..., N_O$	144

List of Abbreviations

Abbrv.	Definition	First Use
CDF	cumulative distribution function	9
MCS	Monte Carlo simulation	12
SGS	sequential Gaussian simulation	12
LMC	linear model of coregionalization	13
PCA	principal component analysis	15
MAF	minimum/maximum autocorrelation factors	16
SCT	stepwise conditional transformation	16
KDE	kernel density estimation	17
ALR	additive logratio transformation	18
MLR	multiplicative logratio transformation	18
ILR	isometric logratio transformation	18
MLE	maximum likelihood estimation	23
MI	multiple imputation	23
SI	single imputation	23
DE	data exclusion	23
MCAR	missing completely at random	24
MAR	missing at random	25

Chapter 1

Introduction

Background on the problem setting is provided before describing modern challenges with multivariate geostatistical modeling. These challenges motivate the research items that are documented in this thesis, which are outlined before being summarized with a thesis statement.

1.1 Background

Subsurface resource characterization is a critical task for mining, petroleum and environmental projects, as it has an overarching impact on business, technical and operational decision making. A mining context is used in this thesis, although the key concepts are applicable to any subsurface resource. Many measurements may be available for informing ore resource characterization, such as drill core, blast hole and channel samples. These measurements are used to inform the modeling of key variables across the entire deposit. These models of the subsurface may then be used to evaluate process performance. For example, mineral resource and contaminant grades are often modeled to evaluate: i) the global resource for economic feasibility studies, ii) local ore and waste contacts for mine planning, iii) blending strategies, and iv) plant design.

The standard practice for modeling subsurface variables applies a branch of statistics known as geostatistics. Developed in the pioneering work of Matheron (1962), geostatistics is defined as "the study of phenomena that fluctuate in space" (Olea, 1991). Within the current context, geostatistics can be practically described as an approach and set of tools for the numerical modeling of geological variables (Deutsch and Journel, 1998). The true nature of a subsurface deposit will reflect a complex geologic history of physical, biological, and/or chemical processes. Some

understanding of a deposit may be inferred from geologic concepts and numerical measurements, but uncertainty will always exist without exhaustive sampling.

To quantify this uncertainty, geostatistics employs stochastic simulation to generate multiple realizations of the subsurface (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978). Passing each realization of the variables through a transfer function yields a distribution of uncertainty for the process performance response (Figure 1.1).

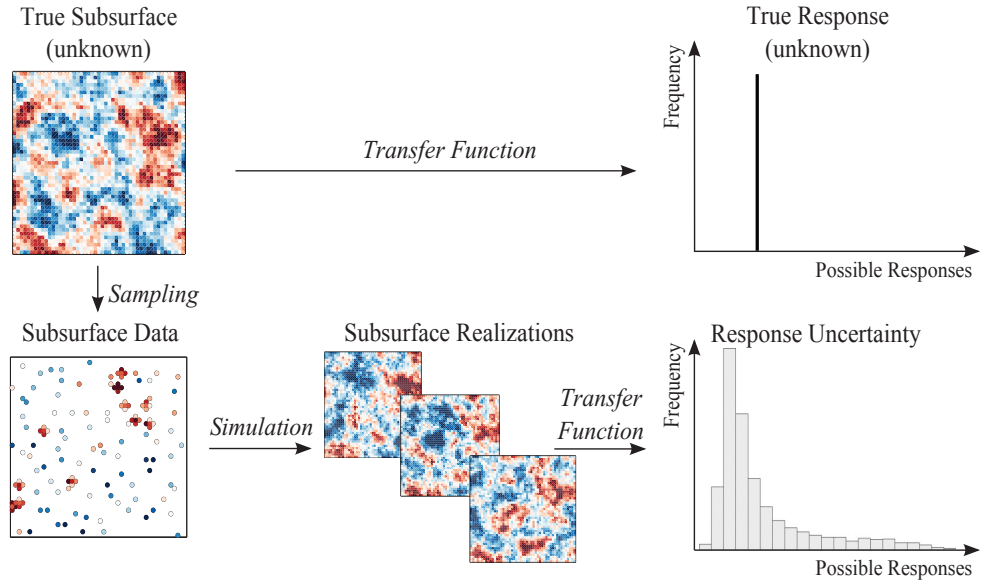


Figure 1.1: Schematic representation of geostatistical simulation and process performance evaluation.

When interpreting Figure 1.1, consider the example of a mineralization and uncertainty of the associated resource. One true realization of the mineralization exists, which an optimal mine plan (transfer function) would exploit to yield a true resource volume (response). As understanding of the mineralization is limited, samples are used with geostatistical inference to generate multiple realizations of the subsurface. Optimal mine plans are generated for each realization (Dimitrakopoulos, 2011; Godoy, 2003), which yield a distribution of uncertainty for the resource.

1.2 Multivariate Modeling

Most mining projects will require the characterization of multiple continuous variables. Relationships between resource and contaminant variables will have a large

impact on process performance forecasting. Consequently, geostatisticians aim to reproduce these relationships in model realizations. Traditional geostatistical modeling techniques will assume that the data are multivariate Gaussian (multiGaussian), so that the multivariate relationships are fully parameterized by the covariance matrix (Chiles and Delfiner, 2012; Isaaks, 1990; Journel and Huijbregts, 1978; Verly, 1983). While this assumption leads to mathematical and computational tractability, geological variables are rarely multiGaussian in nature. Rather than the characteristic elliptical contours of a multiGaussian distribution, geological variables often exhibit complex features such as heteroscedasticity, non-linearity and constraints. Schematic illustrations of these complexities are compared with a Gaussian distribution in Figure 1.2. Though bivariate distributions are shown in this figure, the presented complexities extend to additional dimensions.

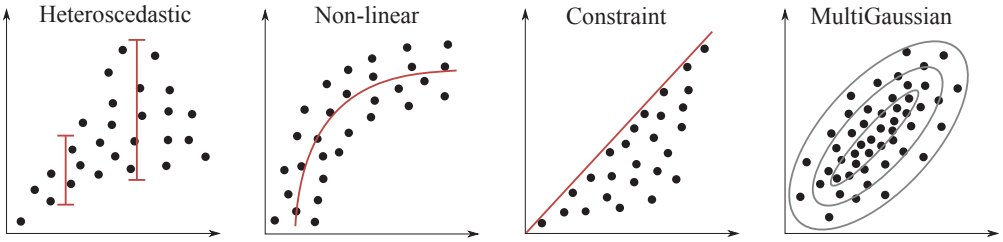


Figure 1.2: Schematic representation of multivariate complexities and a comparative multiGaussian distribution.

Faced with this problem, common practice involves transforming each variable to be univariate Gaussian, before assuming the variables are multiGaussian to apply conventional geostatistical modeling techniques (Leuangthong and Deutsch, 2003). Data may be transformed to be univariate Gaussian using the normal score transformation (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978; Verly, 1983), which is well established in practice and straight forward to execute. MultiGaussian geostatistical model results are returned to the original distributions using the associated back-transformation.

While practical, the assumption that univariate Gaussian variables are multiGaussian is often unrealistic. Figure 1.3 visualizes a cosimulation approach (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978), where two variables are simulated in a manner that considers the data correlation. Two variables are modeled (silica and magnesia), where the relationship between them is very important for the plant processing of nickel laterite ore. Observe from the marginal histograms that a nor-

mal score transform makes the univariate distributions Gaussian, but that bivariate complexities remain (top right). Cosimulation reproduces the targeted normal score correlation, but the resulting bivariate realizations do not possess the complex features of the normal score data (bottom right). Back-transforming to original space exacerbates this problem, as neither the correlation nor the complex features are reproduced (bottom left). This small example illustrates that conventional geostatistical analysis of complex multivariate data may lead to systematic errors.

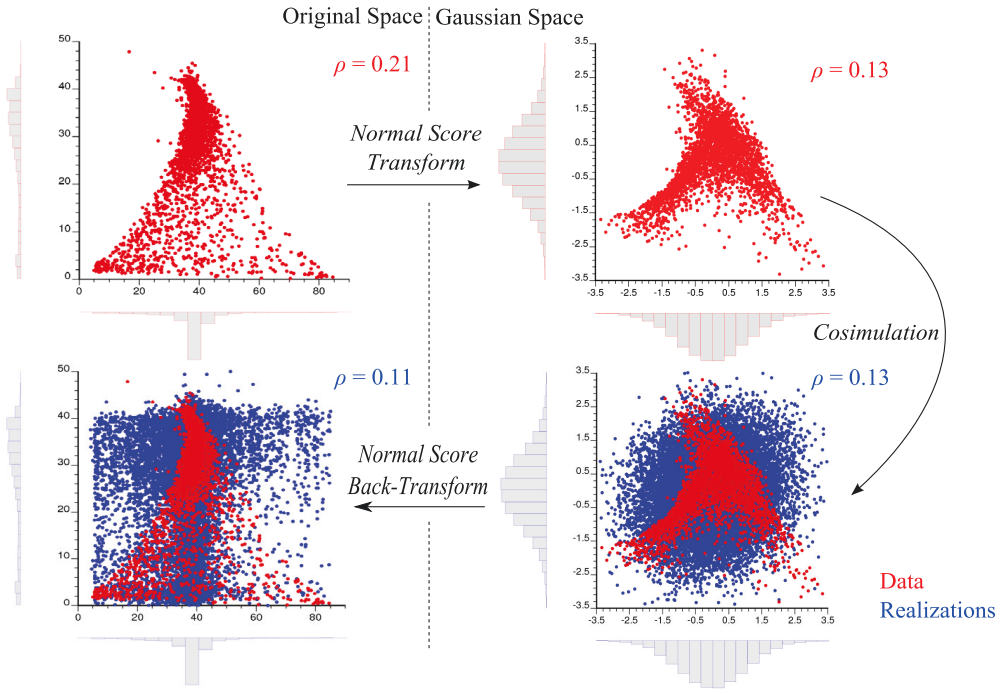


Figure 1.3: Demonstration of geostatistical cosimulation with complex bivariate data, where silica is on the x-axes and magnesia is on the y-axes of each plot.

1.3 Multivariate Transformations

To address this issue, a variety of techniques may be considered for transforming variables to be multiGaussian (Leuangthong, 2003). Conventional modeling methods may then proceed, before using the associated back-transformations to return the original complexity to simulated realizations. Many of these transformations will also decorrelate the variables so that modeling is simplified to independent simulation. Associated back-transformations are then used to return the original correlation to simulated realizations. Figure 1.4 illustrates this concept using the nickel laterite variables.

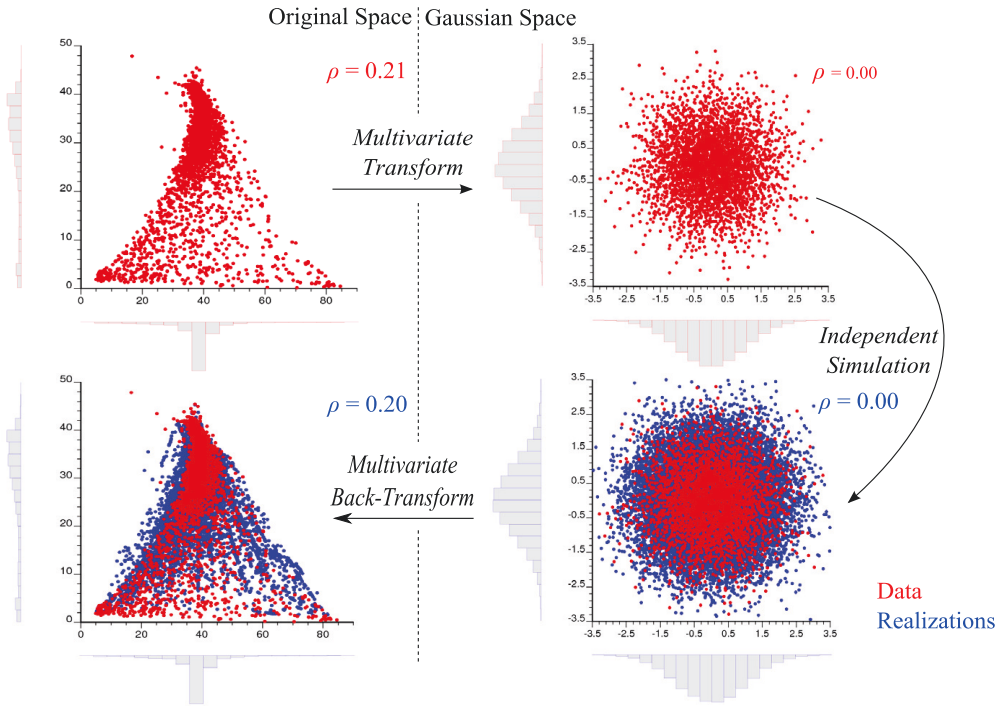


Figure 1.4: Demonstration of multivariate transformation and independent simulation, where SiO_2 is on the x-axes and MgO is on the y-axes.

Observe that the complex data is transformed to be bivariate Gaussian and uncorrelated (top right). As a result, independently simulation yields realizations that match the distribution of the transformed data (bottom right). Back-transformation of the realizations reintroduces the original complexity and correlation (bottom left). Recall that this nickel laterite example is being considered within a process performance evaluation context (Figure 1.1). Given that plant design is highly dependent on the relationship between these two variables, it is expected that the cosimulation approach (Figure 1.3) will lead to inferior evaluation, relative to the multivariate transform approach (Figure 1.4).

1.4 Problem Definition

Although this multivariate transformation concept may appear simple, practical problems surround its application. These problems are described to motivate the primary contributions of this thesis.

1.4.1 Unequal Sampling

The first problem relates to unequal sampling where data observations possess a differing number of measured variables. This commonly occurs with legacy data or when sampling some variables is quite expensive. For example, consider the illustrative schematic in Figure 1.5 where three variables are unequally sampled across six observations. Incomplete or unequal sampled observations are sometimes referred to as heterotopic observations. Multivariate transformations may only be applied to observations that sample all of the variables under consideration (homotopic observations). Heterotopic observations must be excluded from the workflow or have their missing values imputed (inferred). Data exclusion is likely to be problematic for a number of reasons, including the introduction of bias and loss of information (Enders, 2010; Little and Rubin, 2002). Data imputation is the preferred method, though suitable methods for geological variables are not present in the literature. More specifically, available imputation methods do not integrate information from both spatial and colocated data sources.

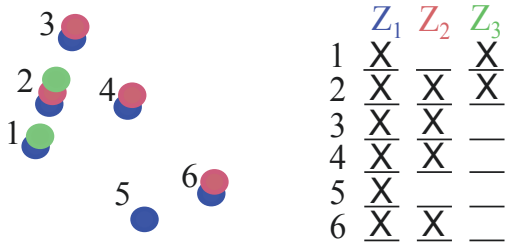


Figure 1.5: Schematic illustration of heterotopic data. A trivariate dataset composed of Z_1, Z_2 and Z_3 variables are heterotopically sampled at six locations (left), leading to a data table where X indicates a sampled value (right).

To address these challenges, a methodology for data imputation in a geostatistical analysis setting has been developed (Barnett and Deutsch, 2015). Imputation theory is integrated with geostatistical tools so that the resultant methods are suitable for geological data. Advancing this methodology into academic and industrial settings should decrease the problematic practice of data exclusion or ad-hoc imputation, while increasing the applicability of multivariate transformations.

1.4.2 Complex Multivariate Data

The second problem relates to functionality gaps that remain in the current set of transformation tools. Consider that the normal score transformation will success-

fully transform any distribution to be univariate Gaussian (disregarding despiking concerns (Deutsch and Journel, 1998)). It is a general technique that is applicable to a wide variety of datasets. Available multivariate transforms do not possess this property, as select transforms will not be successful in the presence of: i) too few observations, ii) too many variables, or iii) complex features.

To address this, a new technique named the projection pursuit multivariate transformation (PPMT) (Barnett et al., 2013) has been developed for transforming potentially complex data to be uncorrelated and multiGaussian. Relative to alternative transforms, the PPMT yields transformed data of improved decorrelation and Gaussianity. More importantly, the PPMT may also be applied to data with more variables and/or less observations. While providing previously unavailable functionality, this new methodology simplifies multivariate geostatistical modeling in practice.

1.5 Thesis Statement and Outline

Improved spatial prediction of geological variables accounting for complex relations and unequal sampling will lead to improved resource management decisions.

Chapter 2 begins with a brief review of fundamental geostatistical theory, which is a necessary precursor to the multivariate techniques that follow. Fundamentals of missing data theory are also reviewed to motivate the methods that are adopted in this research.

Chapter 3 introduces a framework for integrating imputation within geostatistical modeling workflows. Missing data theory is combined with geostatistical algorithms to develop techniques that are optimal for the imputation of geological data. Different techniques may be considered based on properties of the data and priorities of the practitioner. As such, the techniques are demonstrated and compared using two small examples.

Chapter 4 provides a brief overview of transformations that were considered for removing complex features from multivariate data. This early work influences and motivates the PPMT, which was ultimately selected as the preferred multiGaussian transform. Methodology for the PPMT is described in detail in Chapter 5, before demonstrating the technique with a small example.

The PPMT and imputation methodology is applied to a nickel laterite case study

in Chapters 6 and 7, respectively. Modeling results of the new methodologies are compared with that of established techniques. Models are passed through mine project transfer functions to generate response distributions that form the basis for resource management decisions. The value of each contribution is evaluated based on its associated improvement to resource management.

Conclusions and future work are presented in Chapter 8. Methodology that is developed in this thesis has been implemented as Fortran coded programs. These programs are constructed as stand-alone executables that use ASCII parameter and data files. They are used to generate all of the presented results and are available from the author upon request.

Chapter 2

Literature Review

The following chapter provides a literature review of material that is relevant to this thesis. It is divided into three sections: i) conventional multivariate geostatistics, ii) common multivariate transformations and iii) data imputation.

2.1 Conventional Multivariate Geostatistics

Sample and domain definitions are reviewed first, which introduces notation that is used throughout this thesis. Stationarity is described next, as it is a critical precursor to the geostatistical modeling of continuous variables. Univariate modeling fundamentals are then extended to multivariate modeling techniques. These conventional multivariate techniques could be used instead of the multivariate transformations that are introduced in Section 2.2.

2.1.1 Sample, Population, and Stationarity

Multivariate geological data composed of n observations and K continuous variables could be represented as a matrix $\mathbf{Z} : z_{\alpha i}, \alpha = 1, ..., n, i = 1, ..., K$. Assume for now that $\mathbf{Z}$ has no missing values. This geological data is sampled at spatial locations that are represented by the coordinate vectors, $\mathbf{u}_\alpha, \alpha = 1, ..., n$, which is denoted as the set, $\{\mathbf{u}_\alpha, \alpha \in (n)\}$ (Goovaerts, 1997). We are interested in modeling these variables across a model domain, A, that may be discretized into some number of grid locations, $\mathbf{u} \in A$. Since A is a subsurface domain, the specific value of each variable at all locations, $\{z_i(\mathbf{u}), \mathbf{u} \in A, \forall i\}$, is not known.

Consider each variable as a regionalized random variable, $Z_i(\mathbf{u})$, where the potential outcome at each location is represented by its cumulative distribution function (CDF) $F_i(\mathbf{u}; z_i) = Prob\{Z_i(\mathbf{u}) \leq z_i\} \forall z_i$ (Chiles and Delfiner, 2012; Matheron,

1962). The data matrix $\mathbf{Z}$ may be thought of as a sample of specific outcomes from the population A, $\{z_i(\mathbf{u}_\alpha), \alpha \in (n) \subset A, \forall i\}$ (Figure 2.1) (Goovaerts, 1997). Geostatistical modeling will use $\mathbf{Z}$ to infer distributions of potential outcomes at the model grid locations, $\{F_i(\mathbf{u}; z_i), \mathbf{u} \in A, \forall i\}$. To allow for this inference, practitioners must ensure that the data are pooled into stationary subsets. Stationarity implies that parameters observed in $\mathbf{Z}$ may be extended to the modeling domain A. Following basic steps such as declustering, $\mathbf{Z}$ should be representative of A (Chiles and Delfiner, 2012; Matheron, 1962).

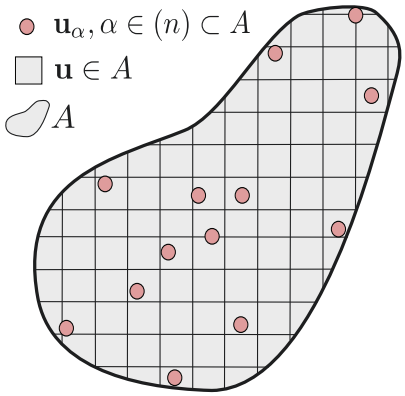


Figure 2.1: Schematic illustration of data locations ($\mathbf{u}_\alpha$), unsampled grid locations ($\mathbf{u}$), and the domain A.

While strict stationarity entails invariance of all high order distributions and related moments, geostatisticians are usually only concerned with first and second order stationarity (Goovaerts, 1997). This relates to stationarity of the first moment, $E\{Z_i(\mathbf{u})\} = \mu_i, \mathbf{u} \in A, \forall i$ and the second moment, $Cov\{Z_i(\mathbf{u}) \cdot Z_j(\mathbf{u}+\mathbf{h})\} = C_{ij}(\mathbf{h}), \mathbf{h}, \mathbf{u} \in A; \forall i, j$. Note that $\mathbf{h}$ is a vector that separates two locations, whereas h will denote the associated distance, $h = |\mathbf{h}|$. As complex multivariate relationships (Figure 1.2) are not characterized by $C_{ij}(\mathbf{h})$ values, practitioners should ensure that those relationships are also stationary. Complex relationships should not be confused with non-stationary relationships, where separate populations are erroneously pooled together. Leuangthong (2003) proposed using techniques such as discriminant analysis (Gnanadesikan and Kettenring, 1999) and cluster analysis (Chatfield and Collins, 1960) to aid in the identification of separate multivariate populations. While any geostatistical modeling workflow and discussion must consider stationarity, the focus of this thesis is the modeling of regionalized variables. The assumption is made in all subsequent sections that a reasonable decision of stationarity has been

made, providing a data matrix $\mathbf{Z}$ that is representative of A.

2.1.2 Univariate Modeling

Recall that the data, $\mathbf{Z}$, is used to infer CDFs of the potential outcomes for each Z_i at the model grid locations, $\mathbf{u} \in A$. Conventional practice would invoke a linear estimation technique known as kriging for geostatistical inference (Chiles and Delfiner, 2012; Matheron, 1962). Kriging calculates the estimate, $z_i^*(\mathbf{u})$, and estimation variance, $\sigma_i^2(\mathbf{u})$, based on the weighted linear combination of nearby spatially correlated data, as well as estimates of the global mean, μ_i, and global variance, σ_i^2.

The notation of $z_i^*(\mathbf{u})$, $\sigma_i^2(\mathbf{u})$, μ_i, and σ_i^2 may cause confusion for some readers, as statistical convention uses 'hat' to distinguish estimated parameters from true underlying parameters. For example, correct statistical presentation would use μ_i for the true underlying global mean, while using $\hat{\mu}_i$ for the estimated global mean. Consistent with geostatistical convention, however, the 'hat' notation is dropped from estimated parameters throughout this thesis. Estimated parameters are presented with far greater frequency, allowing for true underlying parameters to be explicitly defined when they are referred to.

Returning to kriging, the weights for calculating each $z_i^*(\mathbf{u})$ and $\sigma_i^2(\mathbf{u})$ are determined based on the covariance, $C_{ii}(\mathbf{h})$, between $\mathbf{u}$ and the data locations, $\mathbf{u}_\alpha, \alpha = 1, ..., n$. While $C_{ii}(\mathbf{h})$, may be calculated for specific $\mathbf{h}$ vectors from the data, kriging will require that it is represented by a continuous function. Doing so allows for $C_{ii}(\mathbf{h})$ to be determined for all $\mathbf{h}$, which permits calculating $C_{ii}(\mathbf{h})$ between $\mathbf{u}_\alpha, \alpha = 1, ..., n$ and every estimate location, $\mathbf{u} \in A$. This function is referred to as a model of regionalization, which fits a positive definite model to the experimental $C_{ii}(\mathbf{h})$ values (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978). It is worth noting that $C_{ii}(\mathbf{h})$ is frequently referred to as spatial covariance, autocovariance, or spatial continuity. Due to historical practice, spatial variability is frequently calculated according to Equation 2.1.

$$\gamma(\mathbf{h}) = E(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h}))^2/2 \tag{2.1}$$

where $\gamma(\mathbf{h})$ is referred to as a semivariogram (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978; Matheron, 1962). Respecting this convention, examples and case studies in this thesis present spatial continuity by plotting $\gamma(\mathbf{h})$ as a function of distance (e.g., Figure 3.6). Assuming first and second order stationarity, the

semivariogram, $\gamma(\mathbf{h})$, relates to auto-covariance, $C_{ii}(\mathbf{h})$ as:

$$\gamma(\mathbf{h}) = C_{ii}(0) - C_{ii}(\mathbf{h}) \tag{2.2}$$

Although the empirical global CDF of a random variable, $F_i(z_i)$ is often well informed by pooling the available data, the local CDFs, $F_i(\mathbf{u}; z_i)$, are not defined because of few or no replicates of the specific nearby data configuration. This motivates the widespread use of the multiGaussian model within geostatistics. Any local or conditional CDF of a multiGaussian distribution is also Gaussian; it is therefore fully parameterized by the normal equations or simple kriging estimate and variance, $z_i^*(\mathbf{u})$ and $\sigma_i^2(\mathbf{u})$ (Verly, 1983). To facilitate multiGaussian modeling, the normal score transform converts a distribution of any form to be univariate standard Gaussian (Bliss, 1934; Boisvert et al., 2013; Deutsch and Journel, 1998; Verly, 1983). Defining G^{-1} as the inverse of the standard univariate Gaussian CDF, the normal score transform matches probabilities between F_i and G according to Equation 2.3. This transform is schematically illustrated in Figure 2.2.

$$y_{\alpha i} = G^{-1}\left(F_i(z_{\alpha i})\right), \text{ for } \alpha = 1, ..., n, i = 1, ..., K \tag{2.3}$$

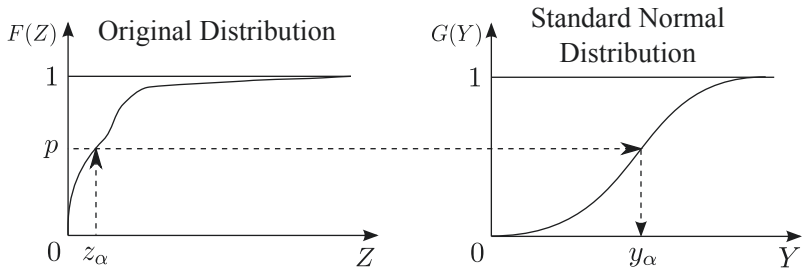


Figure 2.2: Schematic illustration of the normal score transformation.

As the resultant transformed data matrix, $\mathbf{Y}$, is univariate normal, independent modeling of each variable will yield CDFs, $\{G_i(\mathbf{u}), \mathbf{u} \in A\}$, that are fully defined by their mean, $y_i^*(\mathbf{u})$, and variance, $\sigma_i^2(\mathbf{u})$, values from kriging. Monte Carlo simulation (MCS) schemes such as sequential Gaussian simulation (SGS) (Deutsch and Journel, 1998; Isaaks, 1990; Manchuk and Deutsch, 2012; Verly, 1983) may then be used to stochastically sample from these CDFs, generating L realizations of the variables within A, $\{y_{il}(\mathbf{u}), \mathbf{u} \in A, \forall i, l = 1, ..., L\}$, that reproduce the regionalization model of each variable. For convenience, denote the values of a Gaussian realization as $y_{\alpha i}, \alpha = 1, ..., N, i = 1, ..., K$, where N is the number of grid locations,

so that a geostatistical model is cumulatively represented by the locations $\mathbf{u}_\alpha, \alpha = 1, ..., N$. Gaussian realizations are back-transformed to the original distribution according to Equation 2.4.

$$z_{\alpha i} = F_i^{-1}\left(G(y_{\alpha i})\right), \text{ for } \alpha = 1, ..., N, i = 1, ..., K \tag{2.4}$$

Returning to the multivariate context, if the K variables of $\mathbf{Z}$ are dependent, then the CDFs of each location, $F_i(\mathbf{u}), i = 1, ..., K$, should not be determined independently. More specifically, if the covariance matrix, $\mathbf{\Sigma}(\mathbf{h}) : C_{i,j}(\mathbf{h}), i, j = 1, ..., K$, has non-zero off diagonal terms, then a cokriging framework should be applied. This is the focus of the next section. In the presence of complex multivariate relationships that are not characterized by $\mathbf{\Sigma}(\mathbf{h})$, alternative transformations could be used to achieve independence between variables (Section 2.2).

2.1.3 Multivariate Modeling

Kriging considers the auto-covariances, $C_{ii}(\mathbf{h})$; cokriging also considers the cross-covariances, $C_{ij}(\mathbf{h}), \forall i \neq j$. Consider a primary variable to be predicted from $Z_i, i = 1, ..., K$. The CDFs of the primary variable should be inferred in a manner that respects $\mathbf{\Sigma}(\mathbf{h})$ with the remaining $K - 1$ secondary variables. Popular cokriging schemes are differentiated by the coregionalization model that is used for characterizing $\mathbf{\Sigma}(\mathbf{h})$. Coregionalization is the multivariate extension of regionalization, where experimental values of $\mathbf{\Sigma}(\mathbf{h})$ are described as a continuous function for all $\mathbf{h}$. With a coregionalization model defined, cokriging establishes the weight that should be attributed to each observation and variable of $\mathbf{Z}$ when calculating the estimate and estimation variance. These models include the Linear Model of Coregionalization (LMC), the Markov model, and the intrinsic model.

The LMC assumes that $Z_i, i = 1, ..., K$ are the linear combination of a common underlying pool of random variables (Chiles and Delfiner, 2012; Goovaerts, 1994; Journel and Huijbregts, 1978). Practically speaking, the LMC requires the auto and cross-covariance to be fit with a positive semi-definite model. The LMC is utilized for assigning weights to correlated values of $\mathbf{Z}$ in cokriging. More recent modifications and applications of the LMC may be found in Oman and Vakulenko-Lagun (2012) and Mueller and Ferreira (2012), respectively. While semi-automated procedures may be applied for fitting an LMC (Jewbali, 2009; Larrondo et al., 2003; Neufeld and Deutsch, 2004), it is generally considered impractical for massively

multivariate settings. Massively multivariate is a general term that while vague, has very different meaning to different geostatisticians. A number of multivariate methods that are discussed throughout this thesis become difficult or impossible to use beyond approximately three to ten variables. Observing this rough threshold, massively multivariate will refer to $K \geq 5$ in this thesis. Although the lower limit is underwhelming for its name, massively multivariate could refer to a far larger number of variables, such as the $K = 112$ that were modeled in Boisvert et al. (2013).

In settings where the LMC is deemed impractical, the Markov coregionalization model may be considered to avoid fitting the cross-covariances (Almeida and Journel, 1994; Deutsch and Journel, 1998; Xu et al., 1992). The Markov model simplifies the LMC by assuming that colocated secondary data screen the influence of secondary data at any $h > 0$ lag distance. As a result, only the auto-covariance of the primary variable and $\mathbf{\Sigma}(0)$ is required. Since the Markov model only requires colocated information about the secondary variables, it lends its name to colocated cokriging. More recent applications may be found in Babak and Deutsch (2009a) and Boisvert et al. (2013), respectively. While very practical due to simplicity, the cross-covariance assumptions of colocated cokriging often lead to a bias in histogram reproduction (Babak and Deutsch, 2009b; Deutsch and Journel, 1998).

The intrinsic coregionalization model provides a compromise between the convenience of the Markov model and the accurate cross-covariance fitting of the LMC (Babak and Deutsch, 2009b; Wackernagel, 2003). The intrinsic model assumes that the cross-covariance between the i^{th} primary variable and j^{th} secondary variable is given by $C_{ij}(\mathbf{h}) = C_{ii}(\mathbf{h}) \cdot C_{ij}(0)$, that is, the shape of the auto-covariance, scaled to the correct magnitude by the covariance at $h = 0$. This allows for secondary data at $h > 0$ to be considered without requiring a fitted cross-covariance.

Regardless of the coregionalization and cokriging framework that is adopted, no consideration is paid to multivariate relationships that are not characterized by $\mathbf{\Sigma}(\mathbf{h})$. This is a primary motivation for the multivariate transformations that are described in the next section.

2.2 Common Multivariate Transformations

The following section provides an overview of popular multivariate transformations. Some techniques remove the covariance between variables, while others attempt

to also remove complex features. The consideration of realistic multivariate relationships in geostatistical models may lead to issues with histogram reproduction. Another important transform for enforcing histogram reproduction is also reviewed.

2.2.1 Linear Decorrelation Transformations

The first class of transformations are aimed at decorrelating the variables of $\mathbf{Z}$. Considering the difficulty of fitting a coregionalization model (e.g., LMC), these transformations are applied within geostatistics so that the off diagonal terms of $\mathbf{\Sigma}(\mathbf{h})$ may be disregarded for all $\mathbf{h}$. This facilitates independent univariate modeling of each variable (Section 2.1.2), with the associated back-transformation reintroducing the original correlation to the simulated realizations.

Principal component analysis (PCA) is a dimension reduction and decorrelation technique that transforms a correlated multivariate distribution into orthogonal linear combinations of the original variables. This classic technique was developed by Pearson (1901) and Hotelling (1933), before being adapted to geostatistical modeling by Davis and Greenes (1983). More recent applications of PCA in geostatistics include Barnett and Deutsch (2012b) and Boisvert et al. (2013). The first step of PCA performs spectral decomposition of $\mathbf{\Sigma}(0)$, yielding the eigenvector matrix, $\mathbf{V}$ and diagonal eigenvalue matrix, $\mathbf{D}$:

$$\mathbf{\Sigma}(0) = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}} \tag{2.5}$$

The PCA transform is then performed by multiplying $\mathbf{Z}$ with $\mathbf{V}$:

$$\mathbf{Y} = \mathbf{Z}\mathbf{V} \tag{2.6}$$

Equation 2.6 rotates the multivariate data so that the resultant 'principal components' in $\mathbf{Y}$ are uncorrelated. Eigenvalues in $\mathbf{D}$ describe the relative variability that each principal component contributes to the multivariate system, which may allow for components that provide little information to be discarded from subsequent modeling. As a result, PCA may be attractive when a very large number of K variables must be modeled, for example $K > 20$. Consider that the k most important principal components are selected for simulation, where $k \leq K$. Let the k length vectors, $\mathbf{y}_\alpha, \alpha = 1, ..., N$, represent a realization of simulated principal components. The PCA back-transform restores the original dimensionality and correlation to the simulated realization according to:

$$\mathbf{z}_\alpha = \mathbf{y}_\alpha \mathbf{V}^{\mathrm{T}}, \text{ for } \alpha = 1, ..., N \tag{2.7}$$

where $\mathbf{z}_\alpha$ is a K length vector of the simulated values in original units. As discussed, geostatisticians often use PCA to facilitate independent geostatistical simulation of the principal components. While practical, this approach makes two critical assumptions: i) the $h = 0$ multivariate distribution is fully parameterized by $\mathbf{\Sigma}(0)$ and ii) setting $\mathbf{\Sigma}(0) = 0$ will make $\mathbf{\Sigma}(\mathbf{h}) = 0$ for all $\mathbf{h}$. If the first assumption is incorrect (e.g., complex multivariate data), PCA will yield principal components that are uncorrelated but not independent at $h = 0$. Independent simulation and back-transformation of those components will not reproduce those complex features. If the second assumption is incorrect, the principal components will remain correlated at $h > 0$. Independent simulation and back-transformation of spatially correlated components is unlikely to reproduce the original cross-covariance.

Due to issues arising from the second assumption of PCA, minimum/maximum autocorrelation factors (MAF) have become increasingly popular within geostatistics. MAF was first introduced by Switzer and Green (1984) in the field of spatial remote sensing and was popularized in geostatistics by Desbarats and Dimitrakopoulos (2000). Recent geostatistical applications and modifications of MAF may be found in Boucher and Dimitrakopoulos (2012) and Mueller and Ferreira (2012). It is an extension of PCA that performs a two-step spectral decomposition of $\mathbf{\Sigma}(\mathbf{h})$ at $h = 0$ and $h > 0$ lag distances. If $\mathbf{\Sigma}(\mathbf{h})$ is fully described by a two structure LMC for all $\mathbf{h}$, then MAF will remove covariance at all lags. This in turn, should lead to improved cross-covariance reproduction in simulated realizations. Even where the variables are not fully described by a two structure LMC, MAF has still been found to yield better cross-covariance reproduction than PCA (Barnett and Deutsch, 2012b).

2.2.2 Stepwise Conditional Transformation

The stepwise conditional transformation (SCT) attempts to remove complex multivariate features while simultaneously decorrelating the variables at $h = 0$ to form an uncorrelated multiGaussian distribution. This facilitates independent geostatistical modeling of the transformed variables, with the back-transform reintroducing the original correlation and multivariate complexities to simulated realizations. The SCT was introduced by Rosenblatt (1952) and was popularized in geostatistics by

Leuangthong and Deutsch (2003). More recent application and discussion of the SCT may be found in Neufeld et al. (2008) and Pyrcz and Deutsch (2014), respectively.

To apply this technique, the first variable is normal score transformed (Equation 2.3). The second variable is then partitioned according to the probability class of the first variable, before independently normal score transforming the second variable in each discretized bin. The third variable is transformed conditional to probability classes of the first and second variables, and so on. This process is illustrated in Figure 2.3 and is defined as:

$$\begin{aligned} y_{\alpha 1} &= G^{-1}\left(F_1(z_{\alpha 1})\right) \\ y_{\alpha 2} &= G^{-1}\left(F_{2|1}(z_{\alpha 2}|y_{\alpha 1})\right) \\ &\vdots \\ y_{\alpha K} &= G^{-1}\left(F_{K|1,\ldots,K-1}(z_{\alpha K}|y_{\alpha 1}\ldots y_{\alpha K-1})\right), \text{ for } \alpha = 1,\ldots,n \end{aligned} \tag{2.8}$$

If the SCT effectively transforms $\mathbf{Z}$ to an uncorrelated multiGaussian distribution, $\mathbf{Y}$, then independently simulated realizations are expected to possess those properties. The SCT back-transform returns the original complexity and correlation by reversing the forward transformation:

$$\begin{aligned} z_{\alpha K} &= F^{-1}_{K|1,\ldots,K-1}\left(G(y_{\alpha K}|y_{\alpha 1}\ldots y_{\alpha K-1})\right) \\ &\vdots \\ z_{\alpha 2} &= F^{-1}_{2|1}\left(G(y_{\alpha 2}|y_{\alpha 2})\right) \\ &\vdots \\ z_{\alpha 1} &= F^{-1}_{1}\left(G(y_{\alpha 1})\right), \text{ for } \alpha = 1,\ldots,N \end{aligned} \tag{2.9}$$

While possessing many attractive features, the binning nature of SCT makes it heavily restricted to few variables, for example, less than three. The inability to infer a high dimensional non-parametric distribution is sometimes referred to as the curse of dimensionality (Bellman, 1957). To perform a normal score transform, each conditional bin must be populated with a minimum of approximately ten observations (Leuangthong, 2003). Consequently, applying the SCT to $K = 2$ variables requires $n = 10^2$ observations, $K = 3$ requires $n = 10^3$, and so on. It becomes infeasible to execute the SCT with more than two to three variables for most geological datasets.

This binning can also create artifacts in the transformed data and back-transformed simulation results.

Leuangthong (2003) proposed measures to reduce the SCT data requirement, such as a nested application on subsets of the variables, the population of additional data through kernel density estimation (KDE) (Parzen, 1962; Rosenblatt, 1956; Scott, 1992), and overlapping bins. Focusing on the binning artifacts, Manchuk and Deutsch (2011) proposed using a computationally efficient variant of KDE, kernel density networks (Johnston and Kramer, 2011), to execute stepwise in a continuous fashion. While these practical measures may help, using the SCT for greater than three variables is generally not considered.

Similar to PCA, independent geostatistical simulation of the SCT transformed variables assumes that removing complex features and correlation at the $h = 0$ distance removes those properties for all $\mathbf{h}$. Where this assumption is incorrect, systematic issues may arise with the reproduction of univariate, multivariate and spatial properties.

2.2.3 Logratios

Geostatistics is frequently used for modeling natural resources that are geochemical, geophysical or lithological compositions. Considering that all variables of $\mathbf{Z}$ belong to the same composition, basic physical constraints must be respected. First, individual variables must have values greater than or equal to zero, $z_{\alpha i} \geq 0, \forall \alpha, i$. Second, the sum of the variables must equal one, $\sum_{i=1}^{K} z_{\alpha i} = 1, \forall \alpha$. Geostatistical modeling is complicated by the presence of these constraints, as the described linear estimation and stochastic simulation frameworks do not account for their reproduction. Geostatistics is intended for variables that exist in real space, while compositional variables are constrained within simplex space (Aitchison, 1986; Manchuk, 2008).

This motivated the use of logratio transformations where compositional constraints are removed so that modeling may proceed in real space that is not constrained by the simplex. The associated back-transform returns the original constraint to simulated realizations. While many logratio transforms are available, the most commonly applied technique is the additive logratio transform (ALR) (Aitchison, 1986). The ALR transformation is calculated as the logarithm of the ratio between each variable, $z_{\alpha i}$, and a constant divisor variable, $z_{\alpha d}$:

$$y_{\alpha i} = \ln\left(\frac{z_{\alpha i}}{z_{\alpha d}}\right), \text{ for } \alpha = 1, ..., n, i = 1, ..., K, i \neq d \tag{2.10}$$

It is important to note that the transformed data, $\mathbf{Y}$, has its dimension reduced to $K-1$ variables since $z_{\alpha d}$ is removed. Let the $K-1$ length vectors, $\mathbf{y}_\alpha, \alpha = 1, ..., N$, represent a realization of the simulated variables. The logratio back-transform returns the simulated values to simplex space, explicitly reinforcing the compositional constraint according to:

$$z_{\alpha i} = \frac{\exp(y_{\alpha i})}{\sum_{j=1}^{K-1} \exp(y_{\alpha j}) + 1}, \text{ for } \alpha = 1, ..., N, i = 1, ..., K-1 \tag{2.11}$$

The $z_{\alpha d}$ value is restored as the difference between the compositional constraint and $\sum_{i=1}^{K-1} z_{\alpha i}$. Additional logratio transforms are available, including the centered logratio transform (CLR), multiplicative logratio transform (MLR), and isometric logratio transform (ILR). Aitchison (1986) proposed the ALR, CLR, and MLR in his original work on compositional analysis, while the ILR is a more recent development from Egozcue et al. (2003). Selecting from these available transforms requires practitioners to balance simplicity of the technique against required properties of the transformed variables. For example, the ALR is very straight forward to apply, but its transformed variables are correlated. Additional steps must therefore be taken to account for this correlation, whether that involves the subsequent application of a decorrelation transforms to facilitate independent simulation, or the use of cosimulation frameworks. Conversely, the ILR transforms the variables to be orthogonal and constraint free, immediately facilitating independent simulation if the variables are also independent. This comes at the cost, however, of increased complexity in the practical application of the transform relative to the ALR.

The commonality between these transformations is the use of the natural logarithm (from which they derive their name). Consequently, the largest issue for logratios is that they do not permit zero values for any variable to be transformed. Manchuk (2008) outlines many reasons why zero values occur in geological data, as well as the potential solutions that practitioners may use to apply logratios in their presence. When the zeros are considered exactly accurate (termed essential zeros), marginally small numbers are commonly imputed so that logratios may proceed. When the zeros are attributed to rounding errors or measurement detection limits (termed rounded zeros), many methods exist to impute them with informed

non-zero values (Martin-Fernandez et al., 2003). While the imputation of rounded zeros is a recommended practice, there is a legitimate concern with replacing essential zeros so that logratios may be applied. As with several other of the described multivariate techniques, a recent application of logratios may be found in Boisvert et al. (2013).

2.2.4 Histogram Reproduction Transform

The declustered univariate distributions of the conditioning data, $F_i(z_i), i = 1, ..., K$, are usually the most important property for geostatistical model realizations to reproduce. While higher order distributions are certainly important, the univariate histograms will directly impact both the mean and variance of the modeled variables. These two statistics are critical to nearly every transfer function and associated responses, such as resource estimates that are the frequent goal of geostatistical modeling (Figure 1.1).

The multivariate transforms discussed above often cause the simulated realizations to not reproduce the declustered $F_i(z_i), i = 1, ..., K$ for many reasons (Boisvert et al., 2009). Some fluctuation is expected and desired, as this reflects the uncertainty that surrounds a global population. A systematic bias is not desired, however, and could be introduced from many different steps in a modeling workflow. While practitioners should pursue histogram reproduction through improvements to the modeling workflow, a final transformation may be required to ensure reproduction of the original distributions (Deutsch and Journel, 1998; Journel and Xu, 1994).

Working with a single variable for now, let $F^0(z^0)$ represent the global CDF of a simulated realization, $z^0_\alpha, \alpha = 1, ..., N$. Consider that $F^0(z^0)$ is deemed too different from the target CDF, $F(z)$, which is associated with the declustered data. Quantile matching may be applied to transform the simulated values to the targeted CDF:

$$z^t_\alpha = F^{-1}(F^0(z^0_\alpha), \text{ for } \alpha = 1, ..., N \tag{2.12}$$

Observe, however, that Equation 2.12 pays no consideration to exactitude at data locations, $z^t_\alpha \neq z^0_\alpha = z^d$, where z^d is a data value in close spatial proximity to $\mathbf{u}_\alpha$. Simulated values that coincide with data locations should reproduce the sampled values. Further, simulated locations that are near to data locations are relatively certain and should not be heavily influenced by such a global transformation. Conversely, simulated locations that are far from conditioning data are

relatively uncertain and adjustment according to Equation 2.12 is more acceptable. The relative uncertainty of each $\mathbf{u}_\alpha$ may be represented by the kriging variance, σ^2_α, considering the original data only. Consider weighting Equation 2.12 by σ^2_α so that it has a large impact on uncertain locations, while not impacting certain locations. This concept is applied in Equation 2.13, where $\omega_t = \left(\sigma^2_\alpha/\sigma^2\right)^\omega$ and $\omega_o = 1 - \omega_t$ (Journel and Xu, 1994).

$$z^c_\alpha = \omega_t \cdot z^t_\alpha + \omega_0 \cdot z^0_\alpha, \text{ for } \alpha = 1, ..., N \tag{2.13}$$

As can be seen, the final corrected value, z^c_α, is a weighted combination of the original simulated value, z^0_α, and the naive transformed, z^t_α. So long as a reasonable ω is specified by the user, the histogram correction is more heavily enforced in a smoothly increasing nature away from the data. A consequence of this approach, however, is that the target distribution will not be perfectly reproduced. Deutsch (2005a) proposed the iterative application of Equations 2.12 and 2.13 to reduce this discrepancy.

Returning to the multivariate context, consider the independent application of histogram transformations to target the declustered data CDFs, $F_i(z_i), \forall i$. While doing so may enforce reproduction of the univariate properties, multivariate properties of the simulated realizations are not considered and could be distorted.

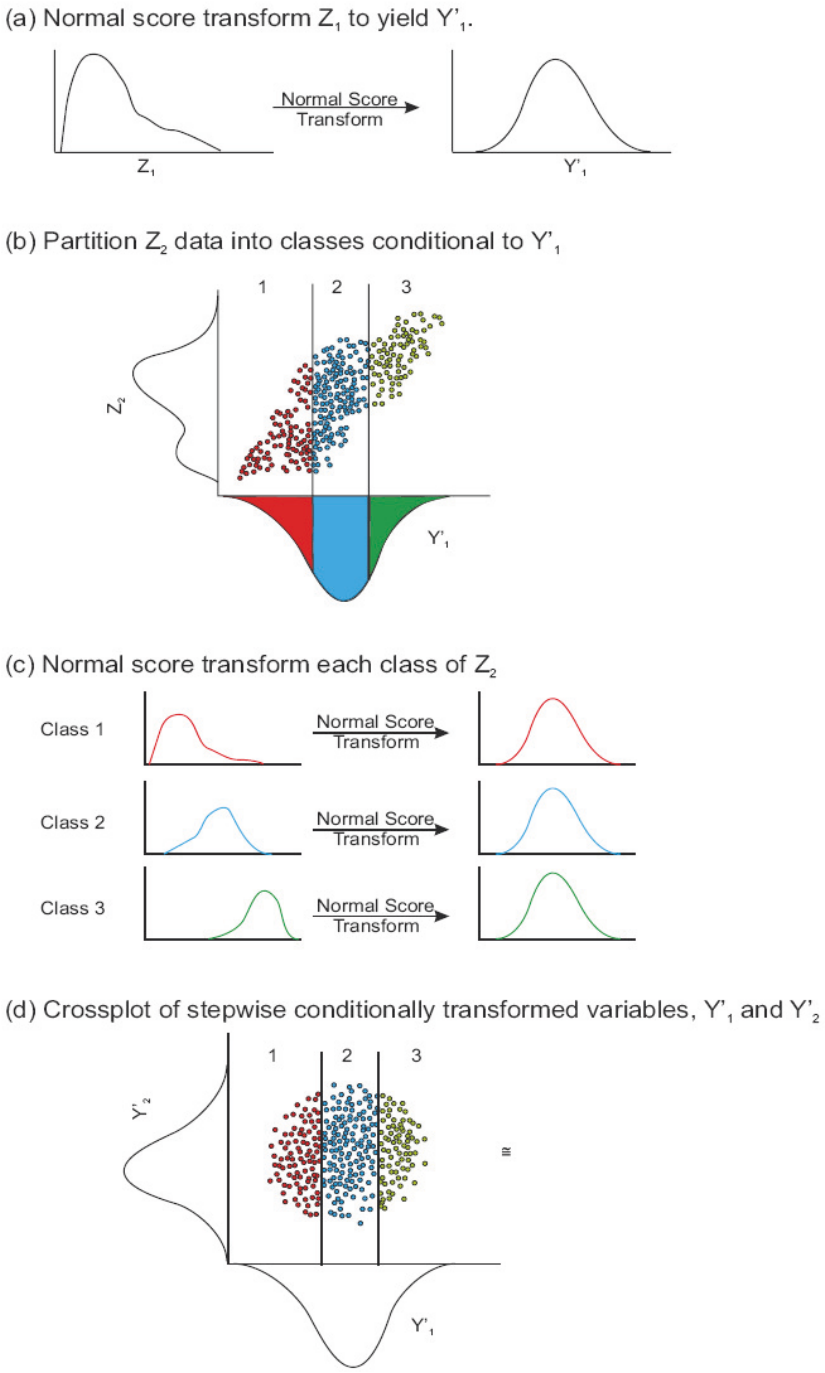


Figure 2.3: Schematic illustration of the stepwise transform for a bivariate case: a) normal score the first variable, b) bin the second variable based on the probability classes of the first variable, c) normal score each bin of the second variable, d) crossplot of the transformed variables are bivariate Gaussianity have approximately zero correlation (Leuangthong, 2003).

2.3 Data Imputation

All of the multivariate transformations that were outlined in the previous section may only be executed on complete data observations that sample every variable under consideration (homotopic data). Unfortunately, geological data usually possess incomplete observations (heterotopic data). When faced with this problem in practical settings, incomplete observations are often excluded so that multivariate transformations and geostatistical modeling may proceed. There are several problems with this approach, including loss of information and data bias. Similarly, 'ad-hoc' imputation procedures may also introduce bias and generally yield suboptimal results.

The following section begins with an overview of techniques that are commonly used for handling missing data in geostatistics and other fields. There are different techniques for different circumstances. Considering missing data mechanisms and other required properties, multiple imputation is selected as the missing data technique that is most suitable for geostatistical analysis. General multiple imputation theory is reviewed, before discussing special considerations for geological data.

2.3.1 Missing Data Techniques

Common missing data analysis techniques are enumerated (Enders, 2010), followed by a discussion of the advocated methods.

i) Listwise deletion: the formal name for excluding heterotopic observations, which is termed data exclusion in this thesis.

ii) Pairwise deletion: reduces the information loss of listwise deletion by only removing observations that do not have the variables under current analysis. Consider calculating each element of a covariance matrix based on subsets that sample each bivariate pair.

iii) Arithmetic mean imputation: replace missing values with the global mean of that variable.

iv) Regression imputation: determine a regression model for a missing variable (response) using the remaining sampled variables as predictors. Each missing value is then estimated based on colocated values.

v) Stochastic regression imputation: the same as regression imputation, but apply stochastic methods to add realistic variability to the regression model assuming normally distributed error.

vi) Hot-deck imputation: randomly replace missing values with data values of other samples that measure similarly according to colocated secondary values.

vii) Similar response pattern imputation: also known as nearest neighbour hot-deck imputation. The random selection is more restricted based on additional factors such as spatial correlation.

viii) Last observation carried forward: the name is derived from its application in time series analysis. The spatial equivalent would be nearest neighbour imputation, where the nearest sampled value is assigned.

ix) Maximum likelihood estimation (MLE): unknown population parameters (e.g. mean and variance) are estimated to maximize the log-likelihood of each observation occurring. In missing data analysis, these parameters are estimated through iterative optimization using various subsets of the data.

x) Multiple imputation (MI): infer a model of the missing values, before stochastically sampling from it to generate multiple realizations of complete data (Figure 2.4). Standard statistical analysis proceeds on the complete data, before combining the results to form an estimate .

Despite their age, MLE (Dempster et al., 1977) and MI (Rubin, 1978) are often referred to as state-of-the-art, whereas the other practical methods are usually referred to as ad-hoc (Enders, 2010; Shafer and Graham, 2002). The ad-hoc techniques may be further subdivided into data exclusion (DE) and single imputation (SI) categories. Although the SI techniques facilitate standard statistical analysis without excluding data, uncertainty of the imputed values cannot be incorporated into subsequent analysis. Further, many of the SI techniques will not reproduce the variability of the data. As discussed, only DE and SI (e.g., regression imputation) are typically used in geostatistical analysis. Methodologists advocate the use of either MLE or MI because they will reflect uncertainty of the imputed values and possess realistic variability. Perhaps even more compelling, MLE and MI are less likely to introduce bias relative to the ad-hoc techniques. Biased analysis relates to reasons why the data are missing, which is described in the next section.

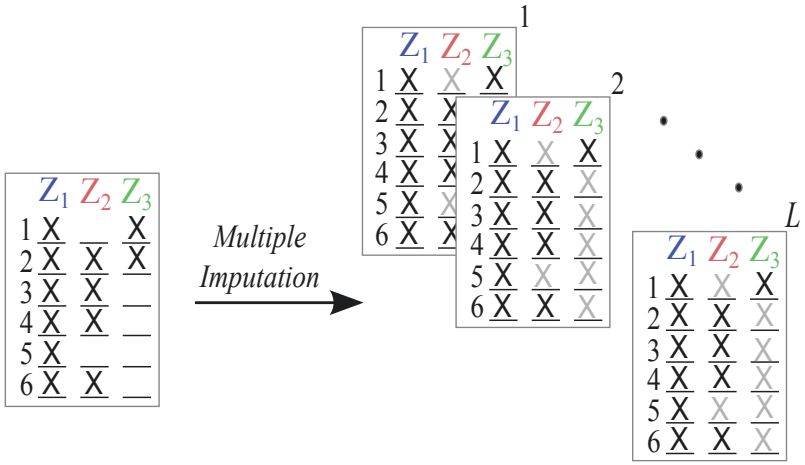


Figure 2.4: Illustration of MI, where an incomplete dataset is used to generate L number of complete data realizations. The sampled values (black) are constant across the realizations, while the imputed values (gray) vary based on the associated uncertainty.

2.3.2 Missing Data Mechanisms

The major advances of missing data analysis (Dempster et al., 1977; Rubin, 1978) coincided with Rubin (1976) proposing a methodology for the mechanisms of missing data (Enders, 2010). Define $\mathbf{M} : m_{\alpha i}, \alpha = 1, ..., n, i = 1, ..., K$ as the missing data indicator matrix, which corresponds with dimensions of the data matrix $\mathbf{Z}$. Each $m_{\alpha i}$ element is a binary indicator, where one indicates a missing value for $z_{\alpha i}$. A major contribution of Rubin (1976) was recognizing that $\mathbf{M}$ may be considered as a random variable and assigned to a distribution. The missing data mechanism is given as the conditional distribution of $\mathbf{M}$ given $\mathbf{Z}$; this may be generalized as $F(\mathbf{M}|\mathbf{Z}, \varphi)$, where φ are parameters that describe the relationship between $\mathbf{Z}$ and $\mathbf{M}$ (Little and Rubin, 2002). While some conceptual understanding of φ may exist, it is difficult to estimate these parameters so that they may be incorporated within statistical analysis. Fortunately, φ may be disregarded depending on the missing data mechanism and the applied analysis technique.

The first missing data mechanism is termed missing completely at random (MCAR) and is represented by Equation 2.14. Observe that the conditional distribution of $\mathbf{M}$ does not depend on the values of $\mathbf{Z}$. MCAR does not imply that the missing data pattern is completely random, but that it is random with respect to the values of $\mathbf{Z}$. MCAR is preferred for analysis since all of the described techniques (including ad-hoc) will not introduce bias when disregarding φ.

$$F(\mathbf{M}|\mathbf{Z},\varphi) = F(\mathbf{M}|\varphi)\forall\mathbf{Z},\varphi \tag{2.14}$$

The second missing data mechanism is termed missing at random (MAR) and is represented by Equation 2.15. Observed values of $\mathbf{Z}$ are specified as $\mathbf{Z}_{obs}$, while the missing values are $\mathbf{Z}_{mis}$. The conditional distribution of $\mathbf{M}$ does not depend on the values of $\mathbf{Z}_{mis}$, but are dependent on the values of $\mathbf{Z}_{obs}$. MAR does not mean that the missing data pattern is random, but that it is random with respect to $\mathbf{Z}_{mis}$. MAR is more challenging than MCAR since ad-hoc techniques will introduce bias unless φ is incorporated. Fortunately, MLE and MI will not introduce bias when disregarding φ.

$$F(\mathbf{M}|\mathbf{Z},\varphi) = F(\mathbf{M}|\mathbf{Z}_{obs},\varphi)\forall\mathbf{Z}_{mis},\varphi \tag{2.15}$$

The third missing data mechanism is termed missing not at random (MNAR), which is given as Equation 2.16. Observe that the conditional distribution of $\mathbf{M}$ depends on the values of $\mathbf{Z}_{mis}$ and $\mathbf{Z}_{obs}$. MNAR is the most challenging mechanism, as none of the reviewed techniques may be used without incorporating φ. Collins et al. (2001) demonstrated that the application of MI and MLE to MNAR data may not introduce substantive bias, but this is dependent on the relationship between $\mathbf{Z}_{mis}$ and $\mathbf{M}$. For MI, Rubin (1996) and Collins et al. (2001) advocate the inclusion of 'auxiliary' variables that help to explain $\mathbf{M}$, regardless of whether they are related to the subsequent analysis. This strategy would effectively shift the data toward a MAR mechanism. If these measures are inadequate, a variety of specialized techniques may considered that directly account for φ (Enders, 2010; Shafer and Graham, 2002).

$$F(\mathbf{M}|\mathbf{Z},\varphi) = F(\mathbf{M}|\mathbf{Z}_{obs},\mathbf{Z}_{mis},\varphi)\forall\varphi \tag{2.16}$$

The only mechanism that may be formally tested is MCAR, which has a large variety of statistical tests available (Chen and Little, 1988; Rubin, 1988). If the data is not MCAR, it falls to practitioner understanding of the variables for determining whether the data is MAR or MNAR.

2.3.3 Multiple Imputation Overview

When choosing a missing data technique for geostatistical analysis, properties that are outlined in the previous sections suggest that MLE or MI should be selected. As

this thesis research will center on MI, the conventional procedure is outlined (Enders, 2010; Huang and Carriere, 2006) before discussing practical considerations. MI may be decomposed into three phases: i) imputation, ii) analysis and iii) pooling. Beginning with the first phase, imputation is performed by constructing a conditional distribution, $F(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$, for the missing values using a prior model and the observed data. The conditional distribution is then stochastically sampled to produce a realization of the missing values, $\mathbf{Z}^l_{mis}$. Iterating this procedure, L number of complete data sets, $\mathbf{Z}_l = (\mathbf{Z}_{obs}, \mathbf{Z}^l_{mis}), l = 1, ..., L$, are generated by combining the observed values with the $\mathbf{Z}^l_{mis}$ missing value realizations. Many techniques could be considered for constructing and sampling from the conditional distributions, including the Gibbs sampler (Geman and Geman, 1984; Little and Rubin, 2002).

Standard statistical (or geostatistical) analysis may then proceed using each $\mathbf{Z}_l$, before combining the results to determine a repeated-imputation inference. Generalize a parameter of interest as θ (e.g., global population mean), where each $\mathbf{Z}_l$ is associated with an estimate, $\widehat{\theta}_l$. The final estimate, $\bar{\theta}$, is determined according to the arithmetic average:

$$\bar{\theta} = {}^1\!/_L \sum_{l=1}^{L} \widehat{\theta}_l \tag{2.17}$$

The associated uncertainty of $\bar{\theta}$ is given by Equation 2.18, which decomposes the total variance, σ^2_T, into the within-imputation variance, σ^2_W, the between-imputation variance, σ^2_B, and the σ^2_B/L term that accounts for a finite number of realizations.

$$\sigma^2_T = \sigma^2_W + \sigma^2_W + \sigma^2_B/L \tag{2.18}$$

Let σ^2_l be the variance that is associated with each $\widehat{\theta}_l$ estimate. The within-imputation variance averages these variances according to Equation 2.19, which represents the variability that would occur if there was no missing data.

$$\sigma^2_W = {}^1\!/_L \sum_{l=1}^{L} \sigma^2_l \tag{2.19}$$

In a complimentary fashion, the between-imputation variance quantifies the variability between data sets according to Equation 2.20, which represents the uncertainty associated with missing values.

$$\sigma_B^2 = {}^1\!/\!_{(L-1)} \sum_{l=1}^{L} (\hat{\theta}_l - \overline{\theta})^2 \tag{2.20}$$

The number of L data sets that is required for MI is approximately three to five according to conventional literature (Little and Rubin, 2002). More recent work, however, has shown that twenty to one hundred data sets are justified in terms of information gained for minimal computational expense (Graham et al., 2007).

2.3.4 Considerations for Geological Data

The performance of MI hinges on the accuracy of the conditional distribution, $F(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$, where the majority of documented techniques assume a multiGaussian form. For example, the Gibbs sampler (Geman and Geman, 1984; Metropolis et al., 1953) is commonly used for generating missing values that converge on the correct covariance, $\mathbf{\Sigma}(0)$ (Little and Rubin, 2002). As complex multivariate features are often present in geological data, such multiGaussian assumptions may lead to inaccurate and unrealistic results. To be successfully applied in geostatistics, $F(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$ should honor any observed complex multivariate features and regionalized correlation, neither of which is captured by $\mathbf{\Sigma}(0)$. This will require that $F(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$ be constructed in a non-parametric fashion and is conditional to both colocated and spatially correlated data.

The MI techniques that are applied to geological data are surprisingly absent from the literature. The nearest work is in the field of compositional data analysis, where imputation is used for the replacement of zero values to facilitate logratio transforms (Martin-Fernandez et al., 2003). While suitable for their purpose, these compositional techniques do not consider $\mathbf{\Sigma}(\mathbf{h})$ for $\mathbf{h} > 0$. Imputation has been applied within other spatial fields, such as agriculture (Lokupitiya et al., 2006), real estate (Knight et al., 1998), traffic flow (Yuebiao and Zhiheng, 2013), and wireless sensor networks (Li and Parker, 2008) and environmental monitoring (Munoz et al., 2010). Techniques for building $F(\mathbf{Z}_{mis}|\mathbf{Z}_{obs})$ in these works include nearest neighbour, universal kriging, KDE, neural networks and regionalized mixed Gaussian models. In all cases, either spatially correlated data or colocated data are considered, but never both sources. Complex multivariate features are not considered for any of the reviewed techniques.

Chapter 3

Imputation of Geological Data

The following chapter presents methodology, examples and discussion relating to the imputation of missing geological data. A summary of this methodology has been published in Barnett and Deutsch (2015) and represents a primary contribution of this thesis. Building on the missing data theory from Section 2.3, multiple imputation (MI) is modified from its conventional form to integrate spatial and potentially complex multivariate data. This improves the imputation of geological data in terms of global bias, local accuracy and reproduction of key properties.

The chapter begins with a discussion on why MI has been selected from the available imputation techniques, before placing its application within the context of geostatistical analysis. After developing the MI simulation framework, a variety of methods are proposed for calculating conditional distributions of missing values within that framework. The methods are compared using two bivariate datasets of varying complexity, providing insight into mechanics of the techniques and the settings where each may be appropriate. The chapter concludes with a discussion on practical considerations and implementation details of the algorithms.

3.1 MI and Geostatistical Analysis

Recall from Section 2.3 that MI and maximum likelihood estimation (MLE) are superior to alternative imputation methods since: i) imputed values match the variability of the data, ii) no bias is introduced with MAR data, and iii) uncertainty of the imputed values is measured and transferred to subsequent analysis. When selecting an imputation method for geostatistical analysis, MI is more immediately suitable since it imputes the missing data values and generates complete data realizations. These data realizations may be used for standard geostatistical analysis,

allowing for seamless integration into popular simulation frameworks. MLE estimates model parameters without imputing data values, making it comparatively difficult to adapt. It should be noted that the two techniques are asymptotically equivalent with a sufficient number of data realizations (Enders, 2010), so the less convenient MLE will not be developed.

MI is now placed within the context of a geostatistical workflow, although additional details are provided in the discussion on practical considerations in Section 3.5. Prior to modeling the subsurface, $l = 1, ..., L$ complete data realizations are generated, where sampled values are constant and imputed values vary in a manner that reflects the uncertainty of each missing value. The value of L is chosen to match the number of geostatistical realizations that is simulated, which usually is one to two hundred realizations in modern workflows (Pyrcz and Deutsch, 2014). Each dataset is passed through the modeling workflow to condition a single geostatistical realization of the subsurface. Following this parallel modeling, results are combined to form a set of simulated realizations that characterize the joint uncertainty of the subsurface. The accuracy and precision of the resultant response distributions from this MI workflow should be superior to that of workflows using data exclusion (DE) or single imputation (SI). The process is schematically illustrated in Figure 3.1.

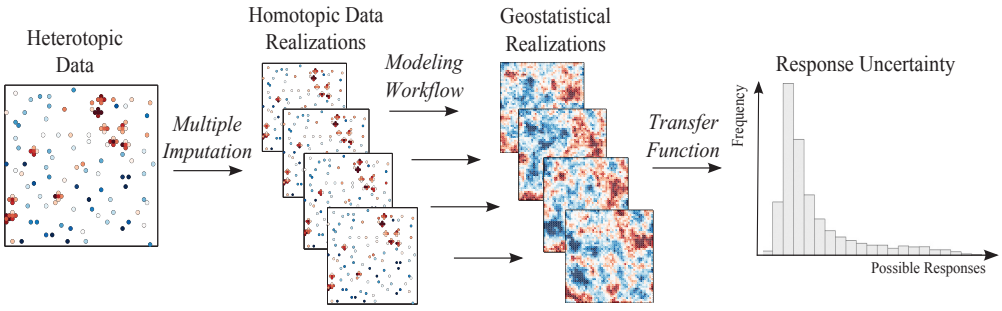


Figure 3.1: Schematic illustration of geostatistical modeling with MI.

The number L is chosen to match the number of geostatistical realizations for several reasons:

i) Diminishing value is gained by using a larger number of realizations than one hundred (Graham et al., 2007).

ii) The concept of creating one high resolution geostatistical realization per dataset follows existent geostatistical methodology. Consider the common practice of

hierarchal modeling, where a matching number of geological facies and continuous properties are stochastically simulated. One continuous property realization is simulated with one facies realization, with the process being repeated L times to explore the joint space of facies and property uncertainty (Pyrcz and Deutsch, 2014). Similarly, the described workflow would simulate one geostatistical realization with one data realization to explore the joint space of subsurface and missing data uncertainty.

iii) Aligning the number of data realizations with the number of geostatistical realizations simplifies workflow scripting and tracking. So long as reasonable automation is used, the user effort should be the same for any number of data realizations.

The motivation for imputation is to facilitate multivariate transformations, which may only be used with homotopic observations. Within the proposed MI workflow, the required multivariate transformations are applied to the L data realizations to generate L transformed data realizations. After using each l^{th} transformed data realization to condition one geostatistical realization, the results are back-transformed to original units. Following this approach, all sampled values of heterotopic data may be used in a multivariate transformation workflow. Note that there is no added practitioner effort when executing multivariate transformations with multiple data realizations, beyond the initial scripting setup as previously described. None of the reviewed transformations require variable input parameters and the computational expense of popular multivariate transformations is not prohibitive.

A potential area of concern is the different modeling parameters for each data realization. For example, a common question is whether semivariograms have to be calculated and modeled for each data realization as input to the simulation? In general, global parameters such as the semivariogram are likely to be quite stable across the data realizations. A primary goal of the imputation methodology that follows is to reproduce a target semivariogram. As such, calculating and modeling the semivariogram of the original data or one transformed dataset is sufficient in most cases.

It is worth noting, however, that the concept of using one imputed data realization to condition one geostatistical realization may be naturally extended to input parameters. Consider settings where sparse sampling leads to significant un-

certainty in parameters such as the global mean and semivariogram. In such cases, one may consider generating L input parameters that span their respective ranges of uncertainty. Using L input parameters, along with L input data realizations, would simultaneously account for parameter and imputation uncertainty in geostatistical modeling.

3.2 Imputation Framework

The imputation framework for generating realizations of missing data values in a geological context will now be documented. The methodology assumes that the data follows either a univariate or multivariate Gaussian distribution, depending on the chosen methodology from the next section. To facilitate the univariate assumption, the first step of the imputation process is to normal score transform the heterotopic data, $\mathbf{Z}$, to the marginally Gaussian heterotopic data, $\mathbf{Y}$. The normal score transform was introduced in Section 2.1.2, where it was represented by Equation 2.3. Note, however, that Equation 2.3 infers that $\mathbf{Z}$ is homotopically sampled since the same n values are transformed for the $i = 1, ..., K$ variables. It is modified for heterotopic data as:

$$y_{\alpha i} = G^{-1}\left(F_i(z_{\alpha i})\right), \alpha = 1, ..., n_i, i = 1, ..., K \tag{3.1}$$

where n_i is the number of sampled values for the i^{th} variable. This simple modification is explicitly defined to emphasize that unlike multivariate transforms, the normal score transform requires no data to be excluded from heterotopic data. It may therefore be used as a preprocessor to the described imputation methodology without loss of information.

Now, working with heterotopic data in standard Gaussian space, MI requires a simulation algorithm for iteratively constructing and sampling from conditional distributions of the missing values, $F(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$. Many techniques could be considered, though the Gibbs sampler (Geman and Geman, 1984; Metropolis et al., 1953) is one of the most popular in missing data literature and practice (Little and Rubin, 2002). The Gibbs sampler is used here to avoid issues that would otherwise be encountered with the curse of dimensionality (as defined in Section 3.3.4). It allows for the joint distribution, $F(Y_1, ..., Y_K)$, to be sampled, while only requiring $F(Y_i|Y_j, j \neq i)$ to be iteratively defined for $i = 1, ..., K$. Illustrating with a bivariate

case, the Gibbs sampler is initialized with a random value, y_2^0, before iteratively drawing from the conditional distributions in Equation 3.2, where the superscript t denotes an iteration of the Gibbs sequence.

$$\begin{aligned} Y_1^t &\sim F\left(y_1 | Y_2^{t-1} = y_2^{t-1}\right) \\ Y_2^t &\sim F\left(y_2 | Y_1^t = y_1^t\right) \end{aligned} \tag{3.2}$$

For imputation, the Gibbs sequence will follow a path across all missing values for each t^{th} iteration. The distribution of each missing value may be conditional to both sampled and previously imputed values. MCS is then used for drawing a sample from the conditional distribution, yielding an iteration of the missing value. Section 3.5 discusses practical implementation considerations for the Gibbs sampler, including potential paths of the sequence and methods for extracting values from the Gibbs sequence (e.g. recording iterations to generate data realizations).

3.3 Imputation Methods

Working within a Gibbs sampler framework in univariate Gaussian space, a method is required for calculating the conditional distribution of each missing value. Four related methods have been developed, which vary based on the sources and assumptions of information that they consider: i) spatial covariance, ii) multivariate covariance, iii) merged spatial and multivariate covariance from techniques (i) and (ii), and iv) merged spatial covariance from technique (i) with non-parametric multivariate information from KDE. Comparing the merged methods, (iii) assumes multivariate Gaussianity, while (iv) captures potentially complex multivariate relationships.

3.3.1 Primary Method

The primary method constructs the conditional distribution of a missing value based on the auto-covariance between it and spatially correlated values. It amounts to the application of simple kriging (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978).

Treat an arbitrary variable that is currently being imputed as the primary variable, Y_p. Similarly, let $\mathbf{u}$ be an arbitrary location where Y_p is currently being imputed. To inform the imputation, $n-1$ sampled and previously imputed primary values, $y_p(\mathbf{u}_\alpha), \alpha = 1, ..., n-1$, are available at spatially correlated locations. Since Gibbs sampling is being used, $n-1$ primary values will always be available for condi-

tioning, though this may include a combination of sampled and previously imputed values.

With the conditioning data assembled, the primary mean, $\bar{y}_p(\mathbf{u})$, and variance, $\sigma_p^2(\mathbf{u})$, of the missing value are calculated according to Equations 3.3 and 3.4, respectively.

$$\bar{y}_p(\mathbf{u}) = \sum_{\alpha=1}^{n-1} \lambda_\alpha \cdot y_p(\mathbf{u}_\alpha) \tag{3.3}$$

$$\sigma_p^2(\mathbf{u}) = 1 - \sum_{\alpha=1}^{n-1} \lambda_\alpha C(\mathbf{u}, \mathbf{u}_\alpha) \tag{3.4}$$

Here, $C(\mathbf{u}, \mathbf{u}_\alpha)$ is the auto-covariance between the missing value at $\mathbf{u}$ and the sampled value at $\mathbf{u}_\alpha$. The weights in Equations 3.3 and 3.4, λ_α, are calculated based on the auto-covariance between $\mathbf{u}$ and $\mathbf{u}_\alpha, \alpha = 1, ..., n$ according to the normal equations:

$$\sum_{\beta=1}^{n-1} \lambda_\beta C(\mathbf{u}_\alpha, \mathbf{u}_\beta) = C(\mathbf{u}, \mathbf{u}_\alpha) \quad \alpha = 1, ..., n-1 \tag{3.5}$$

As any Gaussian distribution is fully defined by its mean and variance, the conditional distribution is now defined by $\bar{y}_p(\mathbf{u})$ and $\sigma_p^2(\mathbf{u})$. MCS may be used for sampling from the resultant distribution, $Y_p^t(\mathbf{u}) \sim F\left(y_p(\mathbf{u})|y_p(\mathbf{u}_\alpha),\ \alpha = 1, ..., n-1\right)$, before proceeding to the next missing value in the Gibbs sequence. Note that $y_p(\mathbf{u}_\alpha),\ \alpha = 1, ..., n$ may contain imputed values from the $t-1$ and t iterations (as seen in Equation 3.2), but this will not be specified going forward.

The implemented Fortran program requires semivariogram models to be provided for the K variables, which defines their respective auto-covariances at all lags. Increasing n can make the system of equations in Equation 3.5 computationally prohibitive. The program allows for a Markov screening assumption (Section 2.1.3) to avoid this issue, where Equations 3.3 to 3.5 are based on a user specified number of nearest observations.

3.3.2 Secondary Method

The secondary method constructs $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs})$ based on the $h = 0$ covariance between missing and colocated variables. It amounts to the application of linear least squares regression (Johnson and Wichern, 1998). While this approach ignores

the spatial covariance of geological data, it is tested for benchmarking since it is the most popular implementation of conventional MI (Shafer and Graham, 2002).

Given that the variable being imputed at location $\mathbf{u}$ has been defined as the primary, Y_p, classify the remaining variables as secondary, $Y_i(\mathbf{u}), i = 1, ..., K-1$. Since Gibbs sampling is being used, $K-1$ secondary variables will always be available at $\mathbf{u}$, though this may include a combination of sampled and previously imputed values.

The covariance between the primary and secondary variables may be used to inform the conditional distribution of the missing value. Calculate the secondary mean, $\bar{y}_s(\mathbf{u})$, and variance, $\sigma_s^2(\mathbf{u})$, of the missing value, according to Equations 3.6 and 3.7, respectively.

$$\bar{y}_s(\mathbf{u}) = \sum_{i=1}^{K-1} \lambda_i \cdot y_i(\mathbf{u}) \tag{3.6}$$

$$\sigma_s^2(\mathbf{u}) = 1 - \sum_{i=1}^{K-1} \lambda_i C_{p,i} \tag{3.7}$$

where $C_{p,i}$ is the covariance between the primary variable and the i^{th} secondary variable. The weights, λ_i, are solved using the normal equations:

$$\sum_{j=1}^{K-1} \lambda_j C_{i,j} = C_{p,i}, \quad i = 1, ..., K-1 \tag{3.8}$$

As with the primary method, the calculated mean and variance fully defines the Gaussian conditional distribution, $Y_p^t(\mathbf{u}) \sim F\left(y_p(\mathbf{u})|y_i(\mathbf{u}),\ i = 1, ..., K-1\right)$. Note that multivariate Gaussianity is assumed by this method, although the normal score transform (Equation 3.1) only ensures univariate Gaussianity. Should complex multivariate features persist through the normal score transform, they will negatively impact imputation results due to the multiGaussian assumption of the method.

3.3.3 Merged Method

The merged method constructs the conditional distribution of each missing value through merging the primary, $F\left(y_p(\mathbf{u})|y_p(\mathbf{u}_\alpha),\ \alpha = 1, ..., n-1\right)$, and secondary, $F\left(y_p(\mathbf{u})|y_i(\mathbf{u}),\ i = 1, ..., K-1\right)$, distributions from the previous sections. This is done using a form of colocated cokriging (Deutsch and Zanon, 2004; Doyen et al., 1996), which uses Equations 3.9 and 3.10 to yield the merged mean, $\bar{y}_m(\mathbf{u})$, and variance, $\sigma_m^2(\mathbf{u})$, respectively.

$$\bar{y}_m(\mathbf{u}) = \frac{\bar{y}_s(\mathbf{u})\sigma_p^2(\mathbf{u}) + \bar{y}_p(\mathbf{u})\sigma_s^2(\mathbf{u})}{\sigma_p^2(\mathbf{u}) - \sigma_p^2(\mathbf{u})\sigma_s^2(\mathbf{u}) + \sigma_s^2(\mathbf{u})} \tag{3.9}$$

$$\sigma_m^2(\mathbf{u}) = \frac{\sigma_s^2(\mathbf{u})\sigma_p^2(\mathbf{u})}{\sigma_p^2(\mathbf{u}) - \sigma_p^2(\mathbf{u})\sigma_s^2(\mathbf{u}) + \sigma_s^2(\mathbf{u})} \tag{3.10}$$

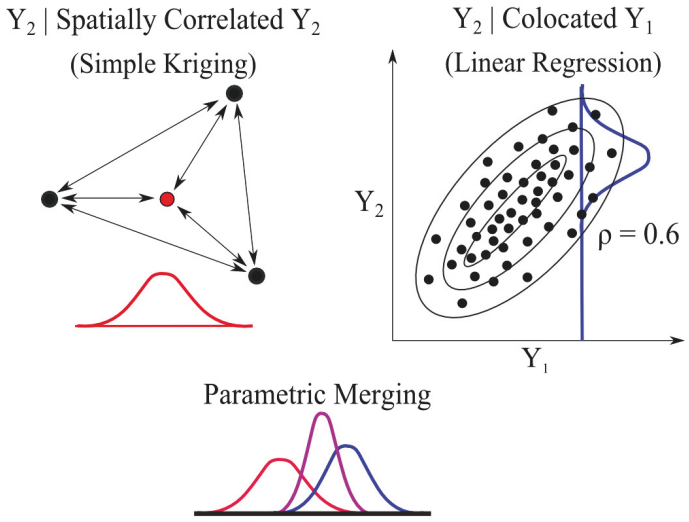


Figure 3.2: Schematic illustration of the merged method.

The process is illustrated in Figure 3.2, where a missing Y_2 value is imputed based on: i) the primary distribution (red curve) of the missing value, which is conditioned by the spatially correlated values (black locations), ii) the secondary distribution (blue curve) of the missing value, which is conditioned by the colocated and correlated Y_1 value, and iii) parametric merging, where the primary and secondary distributions are combined to form the final conditional distribution (purple curve).

The resultant merged distribution,
$Y_p^t(\mathbf{u}) \sim F\left(y_p(\mathbf{u})|y_p(\mathbf{u}_\alpha),\ \alpha = 1, ..., n-1, y_i(\mathbf{u}),\ i = 1, ..., K-1\right)$, integrates both spatial and colocated multivariate information, but contains the strong multiGaussian assumption of the secondary method.

3.3.4 Non-parametric Merged Method

As with its parametric equivalent from the previous section, the non-parametric merged (NPM) method constructs the conditional distribution of each missing value through merging the primary and secondary distributions. To handle complex

multivariate features, however, the secondary distribution is calculated directly from the multivariate data using KDE. The top right panel of Figure 3.3 illustrates this concept, where kernels (grey ellipses) are fitted to the data (black dots) to build a non-parametric distribution of Y_2 conditional to the colocated y_1 value.

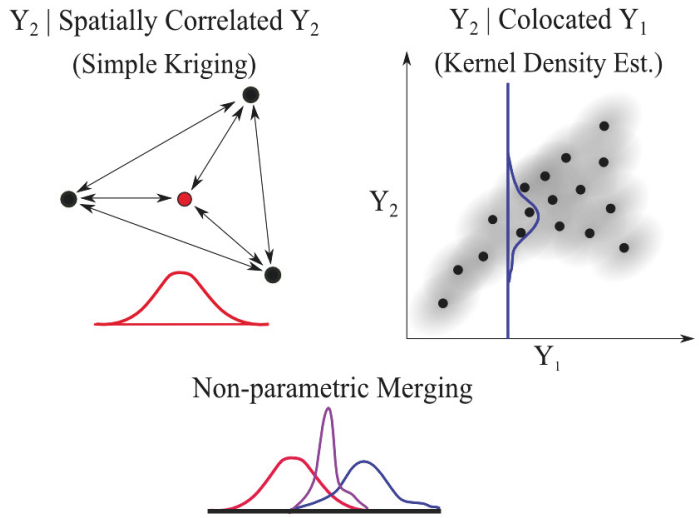


Figure 3.3: Schematic illustration of the NPM method.

A potential concern with this approach is the curse of dimensionality, as calculating the joint KDE for massively multivariate systems is computationally prohibitive. Fortunately, there is an attractive synergy between KDE and the Gibbs sampler. Recall that a Gibbs sequence converges on a joint density through iteratively defining and sampling from conditional distributions. Observing this, KDE estimates are only made along one-dimensional (1-D) vectors in K-variate space.

Consider a set of K length coordinate vectors $\mathbf{y}_\upsilon, \upsilon = 1, ..., d$ as cumulatively representing the discretized locations where $F\left(y_p(\mathbf{u})|y_i(\mathbf{u}),\ i = 1, ..., K-1\right)$ is estimated using KDE. The conditioning variables of each $\mathbf{y}_\upsilon$ are constant according to the colocated values, $y_i(\mathbf{u}),\ i = 1, ..., K-1$, while the variable being imputed $y_p(\mathbf{u})$ varies incrementally across its range. The dashed blue line in Figure 3.3 represents this series of locations, where Y_1 is constant according to the colocated y_1 value and the variable being imputed, Y_2, varies across its range. Multivariate KDE is then performed at each of the d locations using Equation 3.11 (Hong, 2010), where $\mathbf{y}(\mathbf{u}_\alpha), \alpha = 1, ..., n_{hom}$ are the homotopic data values, $\mathbf{H}$ is the symmetric positive definite KxK bandwidth matrix, and $\mathbf{W}$ is a kernel function.

$$f_{KDE}(\mathbf{y}_\upsilon) = \frac{1}{n_{hom}} \sum_{\alpha=1}^{n_{hom}} |\mathbf{H}|^{-1/2} \mathbf{W} \mathbf{H}^{-1/2} (\mathbf{y}_j - \mathbf{y}(\mathbf{u}_\alpha)) \tag{3.11}$$

The black dots in Figure 3.3 represent $\mathbf{y}(\mathbf{u}_\alpha), \alpha = 1, ..., n_{hom}$, while the grey ellipses represent the associated kernels (higher density indicated by darker color). The ellipses illustrate the density that results from each data point, though the KDE calculations are only performed at $\mathbf{y}_\upsilon, \upsilon = 1, ..., d$. The non-parametric secondary distribution is constructed after summing the resultant densities $\sum_{\upsilon=1}^{d} f_{KDE}(\mathbf{y}_\upsilon)$ and insuring that the conditions $f_{KDE}(\mathbf{y}_\upsilon) \geq 0$ and $\sum_{\upsilon=1}^{d} f_{KDE}(\mathbf{y}_\upsilon) = 1$ are met. This discretized distribution is represented by the blue dashed curve in Figure 3.3.

The subjective kernel bandwidth parameter has been studied and is not the focus of this thesis. Nevertheless, the implemented approach is briefly described. The multiGaussian function is used for $\mathbf{W}$ since the normal score transformed data, $\mathbf{Y}$, is marginally Gaussian. The kernel bandwidth parameter is chosen based on iterative visual tuning, which follows the recommended approach for multivariate bandwidth selection from Scott (1992). To streamline this visual tuning, the implemented Fortran program only requires users to input one scalar value, H. This H is used to scale the covariance matrix of the normal score data, $\mathbf{\Sigma}(0)$, yielding the bandwidth matrix $\mathbf{H}$.

The non-parametric secondary distribution may be merged with the primary distribution using a combined probabilities technique that is described in Neufeld and Deutsch (2006). Consider discretizing the primary, secondary and global (standard Gaussian) distributions into a number of values and associated probabilities. If y is a discretized value of interest, its merged probability of occurrence, $P(y|p, s)$, is calculated as:

$$P(y|p, s) = \frac{P(y|p) \cdot P(y|s)}{P(y)} \tag{3.12}$$

where $P(y|p)$ is the probability of y given the primary distribution, $P(y|s)$ is the probability of y given the secondary distribution, and $P(y)$ is the global probability of y. This process is repeated across all d discretizations to yield a non-parametric merged distribution (purple curve in lower panel of Figure 3.3). Neufeld and Deutsch (2006) demonstrate that this method yields an identical result to the parametric approach (Equations 3.9 and 3.10) when applied with Gaussian distributions. Note that the dashed curves in Figure 3.3 symbolize that the input and output of Equation 3.12 is discretized distributions.

3.4 Demonstration

Imputation methods from the previous section will now be compared using two datasets. Variables in the first are approximately bivariate Gaussian, while variables in the second exhibit complex bivariate features. This allows for a demonstration of each method in the presence of these contrasting bivariate forms. Identical imputation workflows are applied to both datasets; jackknife validation is used, where a subset of sampled values are removed so that they may be imputed (Davis, 1987; Efron, 1994; Pyrcz and Deutsch, 2014). The performance of each method is then measured through comparing the true (removed) values with the imputed realizations. The section concludes with a comparison and discussion of the results from both datasets.

Note that the datasets are relatively simple in terms of the number of variables (two), spatial dimension (two), and the missing completely at random (MCAR) missing data mechanism. These simplifications allow for mechanics of the imputation methods to be clearly presented. In a complimenting manner, the case study in Chapter 7 uses real nickel laterite data that is comparatively challenging in terms of the number of variables (four), spatial dimension (three) and missing at random (MAR) missing data mechanism. This provides a realistic geologic setting for testing the described methods and demonstrating their value.

3.4.1 Gaussian Data

The synthetic dataset is composed of two variables, Z_1 and Z_2, which are sampled homotopically at 2,303 locations. Samples are then randomly removed at 1,000 locations (500 for each variable), although this is constrained so that at least one variable must remain for each observation. As a result, 1,803 heterotopic observations and 1,303 homotopic observations remain for informing the imputation. Figure 3.4 displays a map view of the sampled 2-D locations and values, where the missing values (heterotopic samples) are indicated by open spaces in the regular grid of data.

Figures 3.5 to 3.7 provide key statistics of the sampled and missing data, including their univariate distributions (Figure 3.5), spatial variability (Figure 3.6) and bivariate distribution (Figure 3.7). Imputed values are validated against the missing values and statistics from these figures, while the sampled values and statistics are used for construction of the imputation model. More specifically, the normal score

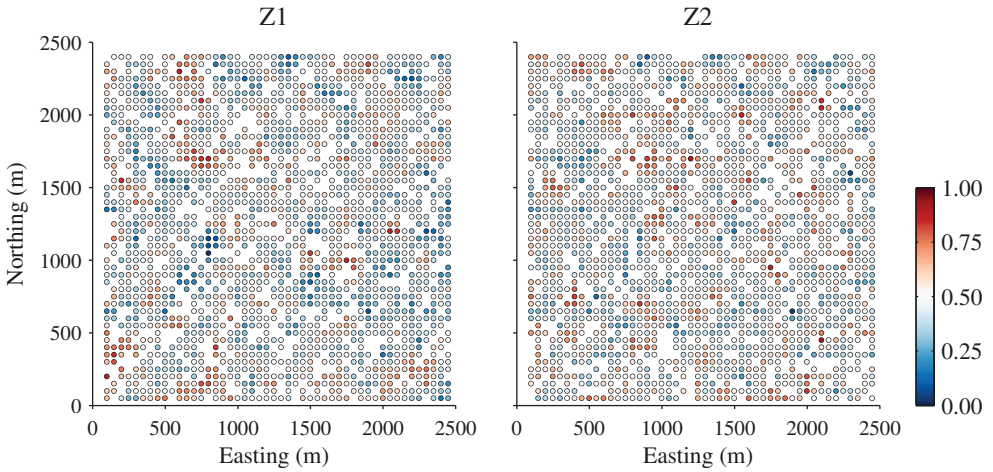


Figure 3.4: Mapview of the 2-D locations, which are colored by the Z_1 and Z_2 values (Gaussian data). Note thatZ_1 and Z_2 range between zero and one to simplify presentation.

sampled data, as well as its associated covariance matrix and semivariogram model are used as input to the imputation.

While the sampled values are generally representative of the missing values, differences do exist in their underlying statistics. These differences arise despite the MCAR mechanism and should be kept in mind when evaluating the imputation results that will follow. The global CDFs in Figure 3.5 are summarized by their mean, μ, and standard deviation, σ. Although the sampled and missing CDFs match in terms of μ, differences exist between their respective σ. Note that the absence of the y-axis label for the Z_2 CDF in Figure 3.5 infers that it matches the y-axis label of the Z_1 CDF. This minimal approach to figure labeling is only necessary for high dimension plots in later sections, although it is applied throughout this thesis for consistency.

The semivariograms in Figure 3.6 (and every other figure) are omnidirectional, where $\gamma(h)$ is calculated at lag distances, h, that are not restricted to specific orientations. To simplify interpretation, each Z_i semivariogram is standardized by the $h = 0$ variance of Z_i, so that $\rho_{ii}(h) = 0$ when $\gamma_{ii}(h) = 1$ (referred to as the sill). Small differences are seen between missing and sampled semivariograms at short scale lag distances, though they are not thought to be consequential or indicative of a methodological issue.

The scatterplots in Figure 3.7 are colored according to the bivariate Gaussian KDE

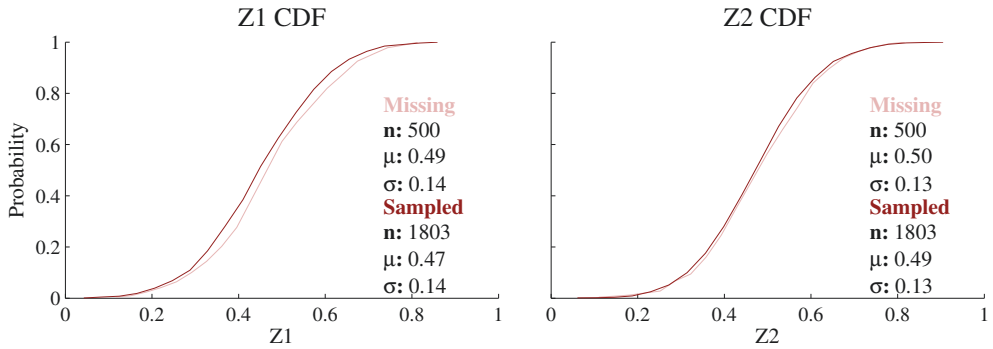


Figure 3.5: CDFs of the missing and sampled values (Gaussian data).

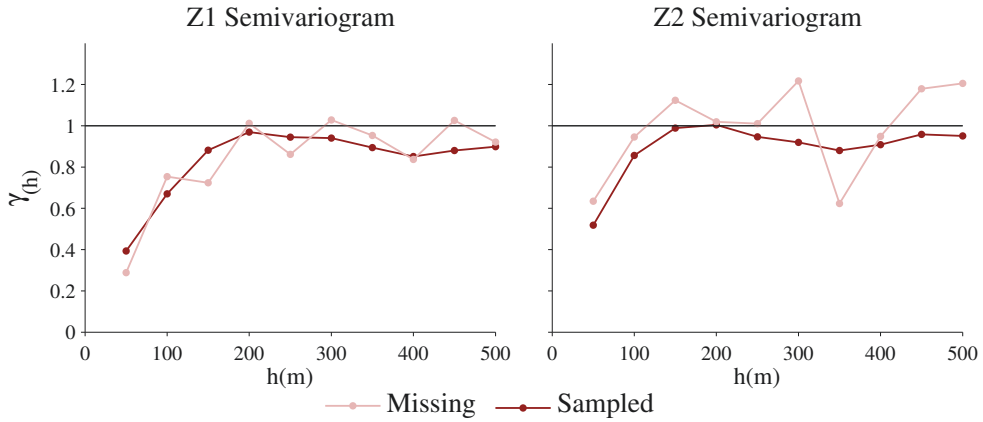


Figure 3.6: Semivariograms of the missing and sampled values (Gaussian data).

of that location. This format is used for visualizing bivariate densities and is referred to as a KDE scatterplot. The single color bar in Figure 3.7 confirms that the same color scaling is applied to both plots in this figure. Again, this minimal approach to figure labeling is only necessary for high dimension plots in later sections, though it is applied throughout this thesis for consistency.

Conditional distributions of missing Z_1 values are displayed for the NPM method in Figure 3.8, although insight into the other imputation methods is also revealed. The grey points are the homotopic data values that are used for KDE. The blue curves represent the non-parametric secondary distributions. The base of each curve is set according to the colocated Z_2 value, so that one can visualize the KDE that results from the nearby homotopic data (imagining the associated Gaussian kernels as in Figure 3.3). The red curves are the primary distributions, all of which exhibit a similar variance due to the semi-regular grid pattern of the data (Figure 3.4). Non-

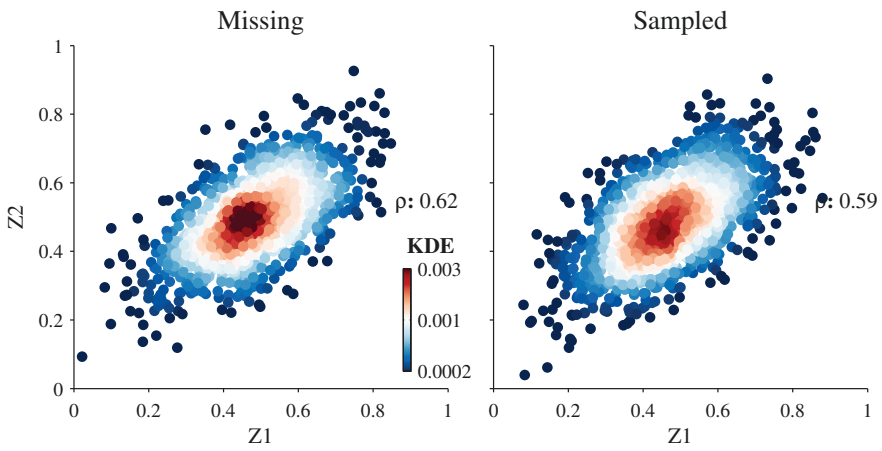


Figure 3.7: KDE scatterplots of the missing and sampled values (Gaussian data).

parametric merging combines the primary and secondary distributions to produce the purple distribution, which would then be drawn using MCS.

Observe that many of the secondary distributions are highly Gaussian, which is expected since Gaussian kernels are describing approximately bivariate Gaussian data. The merging results vary based on the variance and corroboration of the primary and secondary distributions. The locations are somewhat arbitrary, though they are chosen since they display an interesting range of results:

i) Reduction of uncertainty produces a merged distribution that is convex with respect to the primary and secondary (locations 1, 3, 5 and 6).

ii) Reduction of uncertainty produces a non-convex merged distribution (location 2). This behavior is due to the primary and secondary corroborating one another relative to the global.

iii) No reduction of uncertainty as the secondary has a negligible impact on the merged distribution. This occurs since the secondary has non-zero probabilities across the primary and a much larger variance than the primary (location 4).

iv) Non-Gaussian secondary due to poor conditioning for the KDE near the margin of the bivariate distribution (location 3). This non-Gaussianity is reflected in the merged distribution. Given that most of the non-parametric secondary distributions are approximately Gaussian, similar results could be expected for the parametric secondary and merged methods.

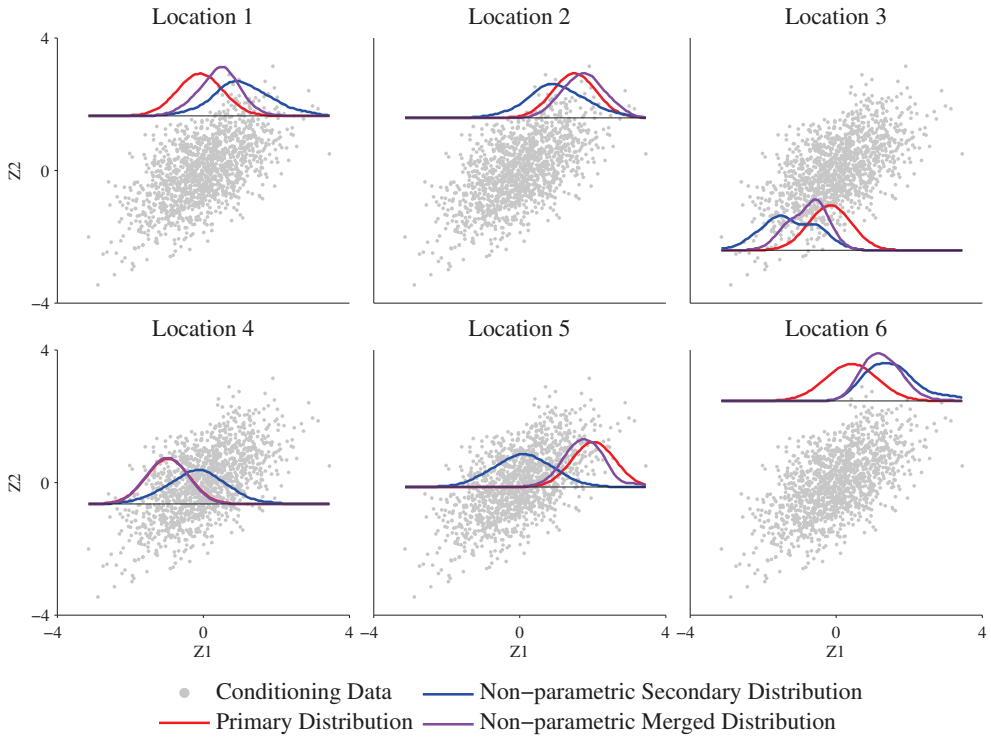


Figure 3.8: Demonstration of the NPM method using Gaussian data.

After using the four methods to impute one hundred data realizations, the results are back-transformed to original units and validated against the true missing values. Bias and accuracy of the techniques are compared in Figure 3.9, which is further summarized in Table 3.1. Each e-type estimate (mean of the one hundred realizations) is plotted against its associated true value to understand the local accuracy and conditional bias. The displayed statistics of root mean square error (RMSE) and Pearson correlation coefficient (ρ) summarize the local accuracy, while the linear regression line summarizes the conditional bias. Global bias is calculated as the error of the displayed global means.

Statistics from Figure 3.9 are standardized in Table 3.1 so that overall performance can be more easily judged. The original statistics are divided by their associated maximum in this table (across all methods, but by variable), so that zero would represent a perfect result and one is the worst result. Z_1 is imputed more accurately than Z_2 in the primary and merged methods since it has greater spatial continuity (Figure 3.6) and resultant predictability.

The best overall result for each measurement is bolded in Table 3.1, highlighting

Table 3.1: Standardized summary statistics that measure the performance of each imputation method in terms of univariate reproduction. The statistics include abs(μ Error), RMSE, and 1 - ρ from Figure 3.9, as well as γ RMSE from Figure 3.10.

Variable	MI Method	μ Error	RMSE	1 - ρ	γ RMSE
Z1	Primary	0.781	0.934	0.818	0.631
	Secondary	1.000	1.000	1.000	1.000
	Merged	0.758	0.855	0.676	0.603
	NPM	0.793	0.854	0.671	0.594
Z2	Primary	0.868	1.000	1.000	0.922
	Secondary	0.371	0.901	0.808	1.000
	Merged	1.000	0.904	0.790	0.777
	NPM	0.962	0.910	0.798	0.809
Average	Primary	0.825	0.967	0.909	0.777
	Secondary	**0.686**	0.950	0.904	1.000
	Merged	0.879	**0.879**	**0.733**	**0.690**
	NPM	0.878	0.882	0.734	0.701

that the merged methods outperform the alternatives for every measure other than global bias. Summary statistics in Figure 3.9 are the same for the merged methods to two decimals places (they differ at three). This is the anticipated result since approximately multiGaussian data is being imputed. Discrepancies between the two techniques occur when the non-parametric secondary distributions are very non-Gaussian (such as location 3 in Figure 3.9), which is relatively infrequent.

Semivariograms of the true values and imputed realizations are displayed in Figure 3.10. To summarize the difference between them, the displayed RMSE statistic is calculated using the semivariogram error at each lag. This RMSE is standardized and averaged in Table 3.1 to simplify the comparison. As expected, the methods that integrate spatial information greatly outperform the secondary method.

Reproduction of the bivariate relationship between the two variables is shown in Figure 3.11. Since the data is approximately bivariate Gaussian, the displayed correlation may be considered as a reasonable statistic for summarizing performance. Given the 0.59 correlation of the sampled data (Figure 3.7), it is unsurprising to see the secondary method yielding a 0.58 correlation. The secondary method is only concerned with converging on the correct covariance and is expected to yield the best results with this Gaussian data. Given that the primary method does not consider $h = 0$ covariance, it is equally unsurprising that it yields the worst result.

The KDE RMSE appearing in this figure is calculated based on the difference

between the KDE of the true missing values and the KDE of each imputation result. In other words, it summarizes the difference between the data and imputation results in Figure 3.11. Observe that the correlation and KDE RMSE statistics corroborate the rank order of the imputation performance; aside from the NPM, which yields the best KDE RMSE result. The KDE RMSE statistic is more useful when evaluating the reproduction of complex bivariate data (beginning in the next section), though it is presented here for consistency.

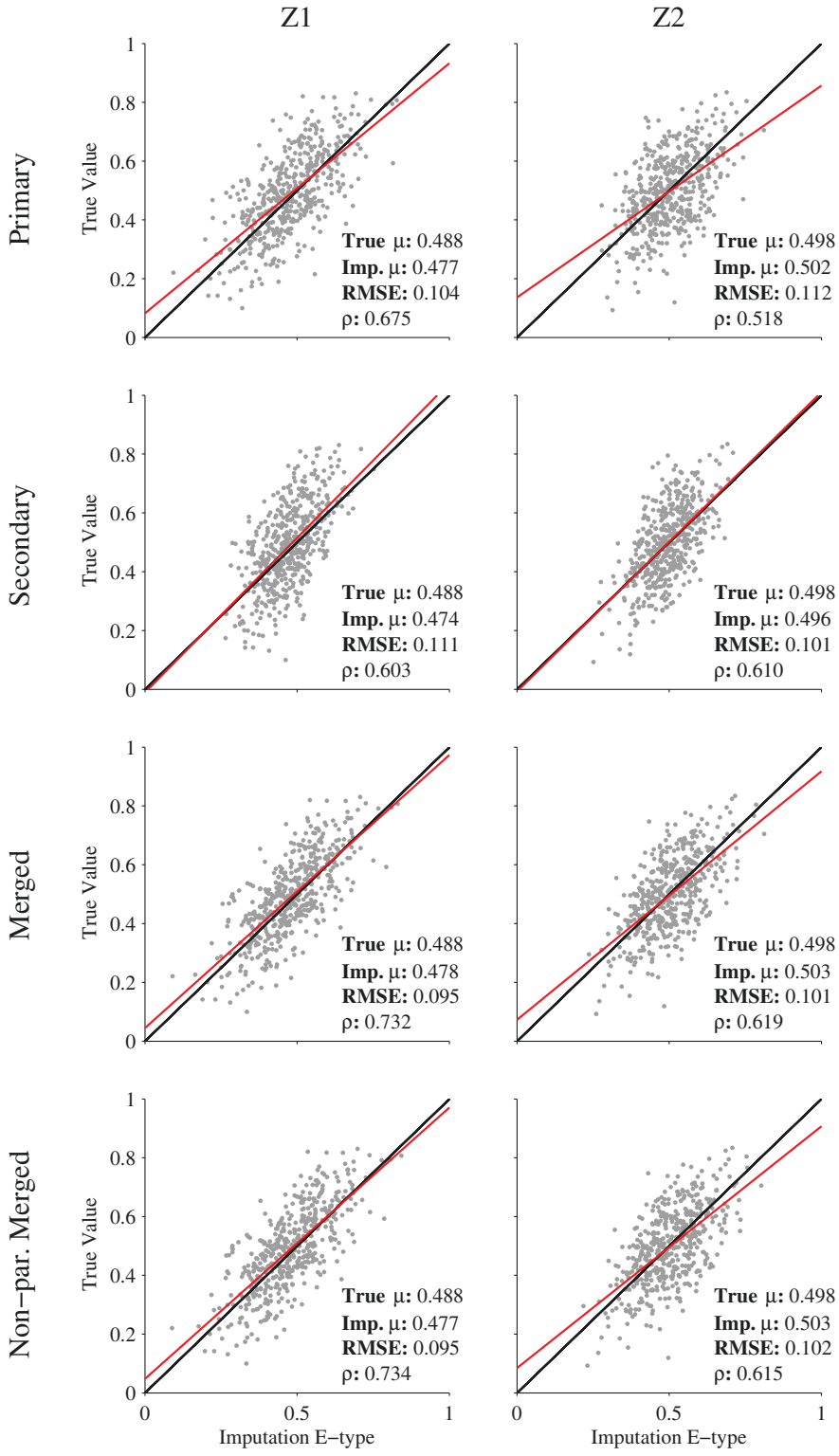


Figure 3.9: Scatterplots and summary statistics that compare the imputed e-type with associated true values (Gaussian data).

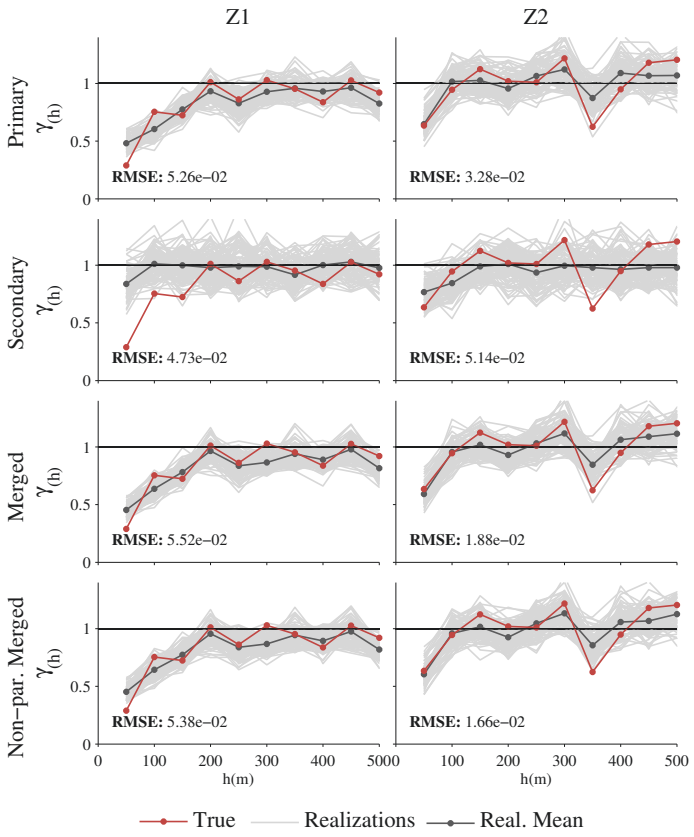


Figure 3.10: Semivariograms of the one hundred imputed realizations for each MI method, with the true semivariograms overlain for comparison (Gaussian data).

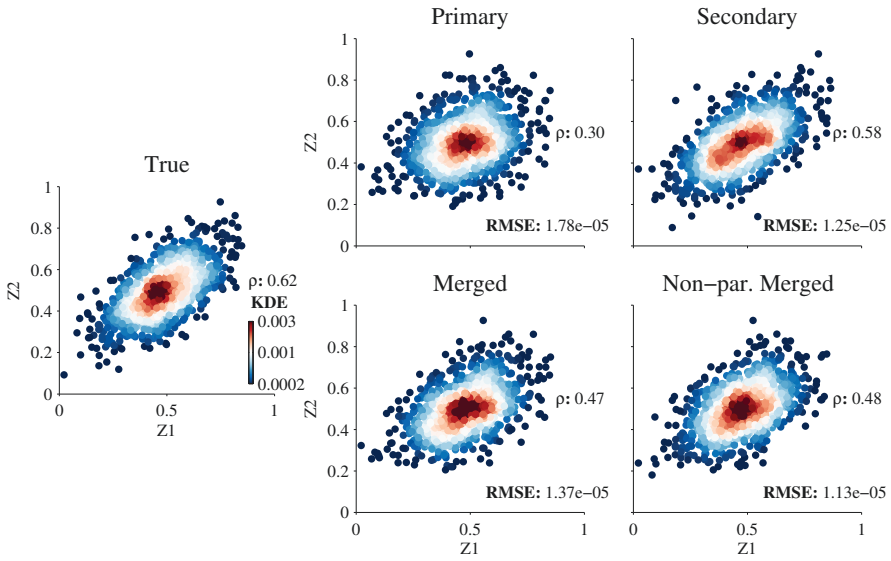


Figure 3.11: KDE scatterplots of the imputed values for each method (one realization), with the true scatter plot displayed for comparison (Gaussian data).

3.4.2 Complex Data

The following considers the same imputation workflow and performance evaluation. While the data being imputed is also identical in terms of the configuration and jackknife removal scheme, the sampled variables exhibit complex bivariate features instead of the bivariate Gaussian features from the previous section. Readers are referred to the previous section for explanations of the figure and table layouts.

Figure 3.12 displays a map view of the sampled 2-D locations and values, while Figures 3.13 to 3.15 provide key statistics of the sampled and missing data. Of particular note, is the heteroscedastic and non-linear features that are present in Figure 3.15. As in the previous section, differences exist in the underlying statistics of the sampled and missing values despite the MCAR missing data mechanism. In particular, the short scale spatial continuity (Figure 3.14) and bivariate distribution (Figure 3.15), both in terms of correlation and the KDE scatterplot.

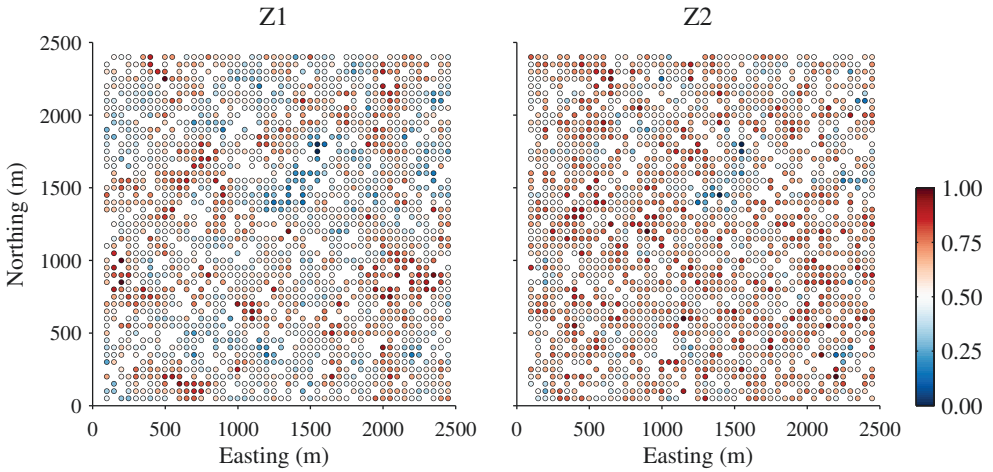


Figure 3.12: Mapview of the 2-D locations and values, where missing locations in the regular grid indicate heterotopic samples (complex data).

Conditional distributions of missing Z_1 values are displayed for the NPM method in Figure 3.16. As with the Gaussian equivalent (Figure 3.8), the locations are chosen since they display an interesting range of results:

i) Reduction of uncertainty produces a merged distribution that is convex with respect to the primary and secondary distributions (locations 1, 4 and 6).

ii) Reduction of uncertainty produces a non-convex merged distribution (locations 3 and 5).

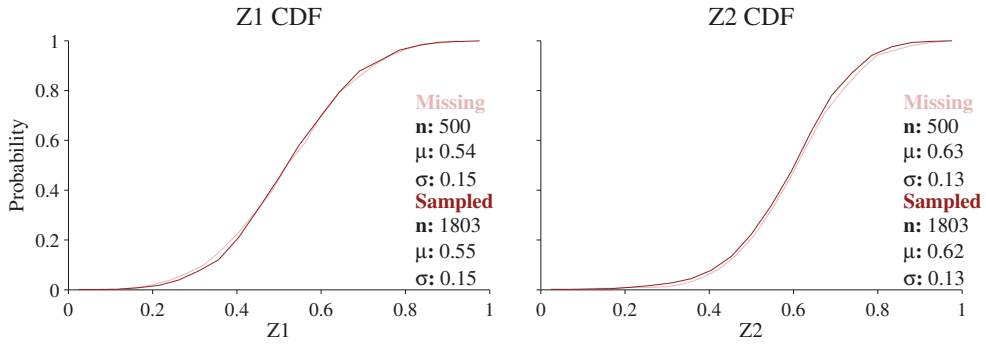


Figure 3.13: CDFs of the missing and sampled values (complex data).

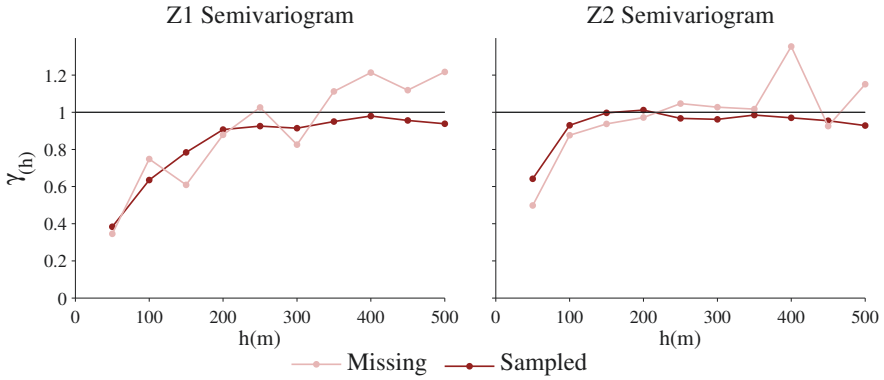


Figure 3.14: Semivariograms of the missing and sampled values (complex data).

iii) No reduction of uncertainty as the secondary distribution has a negligible impact on the merged distribution (location 2).

The secondary distributions in Figure 3.16 are far less Gaussian than the results in Figure 3.8. This is expected given that the underlying bivariate distributions are complex and Gaussian, respectively. This non-Gaussianity is extended to the merged distributions in locations 5 and 6 of Figure 3.16, since the variance of the secondary distributions is sufficiently low relative to the Gaussian primary distributions at these locations.

After using the four methods to impute one hundred realizations, the results are back-transformed to original units and validated against the true missing values. Bias and accuracy of the techniques are compared in Figure 3.17, which is further summarized in Table 3.2. Observe that Z_1 is imputed more accurately than Z_2 for the primary and merged methods since it has greater spatial continuity (Figure 3.14). The best overall result for each measurement is bolded in Table 3.2, highlighting

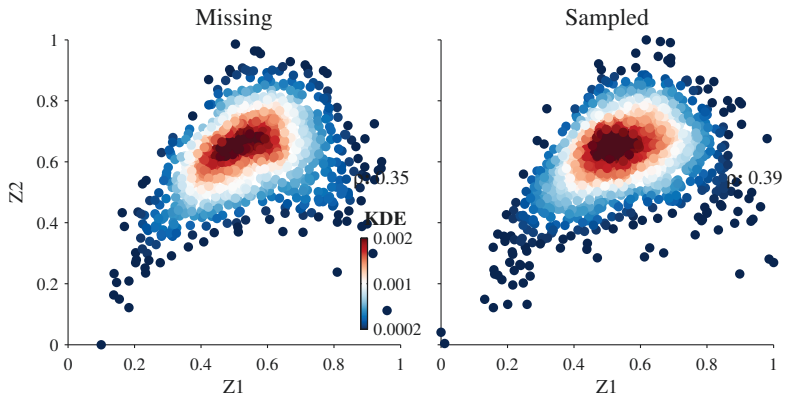


Figure 3.15: KDE scatterplots of the missing and sampled values (complex data).

that the NPM method outperforms the alternatives for every measure. Unlike the Gaussian dataset results (Figure 3.9), the NPM method outperforms its parametric equivalent by a significant margin in terms of global bias and local accuracy. While the obvious advantage of the NPM method is that multivariate complexities should be reproduced, this result demonstrates that improvements in the reproduction of univariate statistics may also be expected.

Table 3.2: Standardized summary statistics that measure the performance of each imputation method in terms of univariate reproduction. The statistics include abs(μ Error), RMSE, and 1 - ρ from Figure 3.17, as well as γ RMSE from Figure 3.18.

Variable	MI Method	μ Error	RMSE	1 - ρ	γ RMSE
Z1	Primary	0.640	0.809	0.530	0.630
	Secondary	1.000	1.000	1.000	1.000
	Merged	0.152	0.781	0.487	0.590
	NPM	0.074	0.745	0.433	0.526
Z2	Primary	0.994	1.000	1.000	0.922
	Secondary	0.672	0.899	0.970	1.000
	Merged	1.000	0.976	0.942	0.894
	NPM	0.840	0.920	0.837	0.842
Average	Primary	0.817	0.904	0.765	0.776
	Secondary	0.836	0.950	0.985	1.000
	Merged	0.576	0.878	0.714	0.742
	NPM	**0.457**	**0.832**	**0.635**	**0.684**

Semivariograms of the true values and imputed realizations are displayed in Figure 3.18. The displayed RMSE statistic is standardized and averaged in Table 3.2. As expected, the methods that integrate spatial information greatly outperform

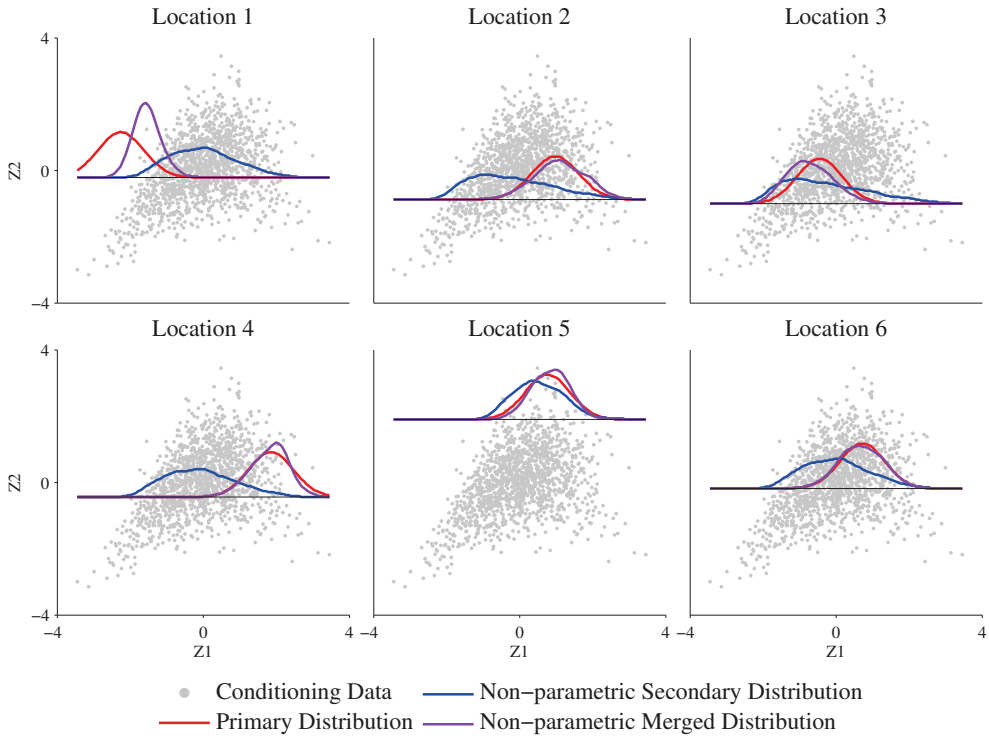


Figure 3.16: Demonstration of the NPM method using complex data.

the secondary method. Discrepancies in the short scale structure are explained by comparing semivariograms of the sampled and missing values (Figure 3.14).

Reproduction of the bivariate distribution is examined in Figure 3.19. As with the Gaussian dataset, the secondary method yields the best reproduction of colocated correlation. Again, this is expected given that the secondary method is only concerned with converging on the correct correlation. Since the data is not bivariate Gaussian, however, the displayed correlation is no longer an appropriate statistic for summarizing performance. The NPM method best reproduces the non-linear and heteroscedastic features of the data according to the KDE scatterplot. This visual observation is supported by the KDE RMSE statistic.

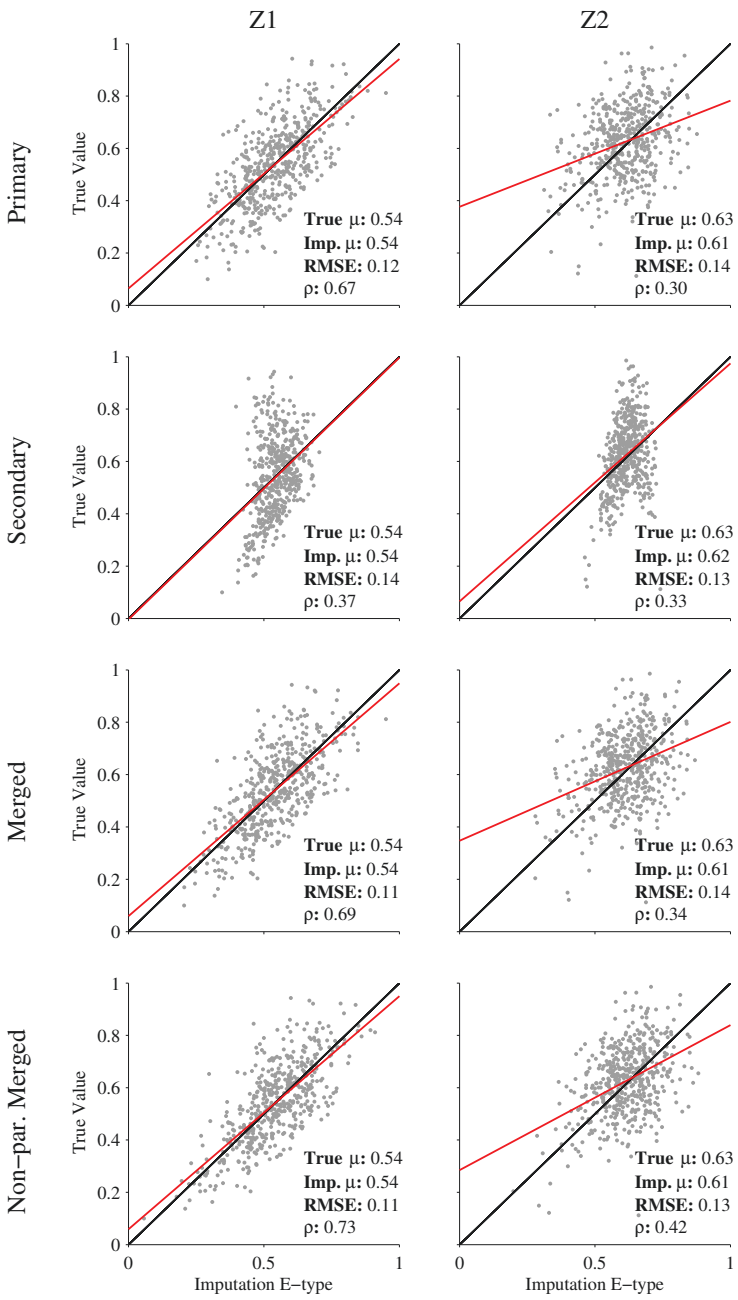


Figure 3.17: Scatterplots and summary statistics that compare the imputed e-type with associated true values (complex data).

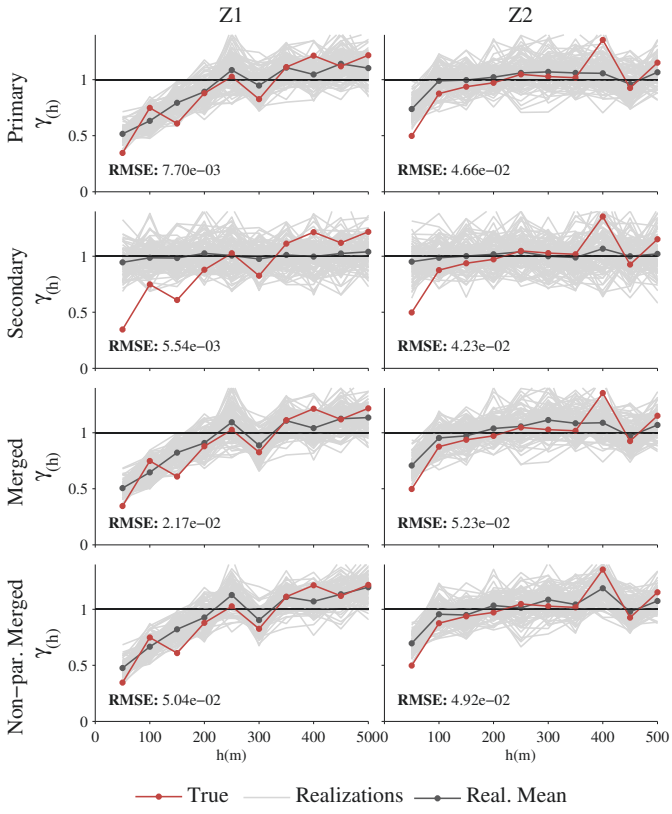


Figure 3.18: Semivariograms of the one hundred imputed realizations for each MI method, with the true semivariograms overlain for comparison (complex data).

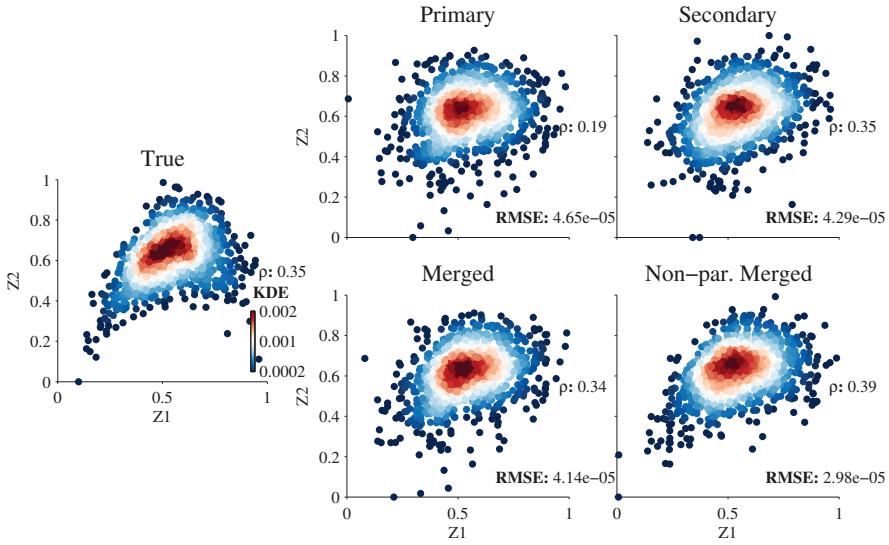


Figure 3.19: KDE scatterplots of the imputed values for each method (one realization), with the true scatter plot displayed for comparison (complex data).

3.4.3 Summary

Synthesizing key findings of the two datasets, the merged methods yield consistently superior results in terms of local accuracy. As evidenced in Figures 3.9 and 3.17 and summarized in Tables 3.1 and 3.2, missing values are imputed nearer to their true values when incorporating information from both spatial and colocated sources. As expected, the merged methods produced very similar results when applied to the Gaussian data, though the NPM method produced significantly better results with the complex data. No consistent benefit was observed for the tested methods in terms of global and conditional bias.

As shown in Figures 3.10 and 3.18 (summarized in Tables 3.1 and 3.2), the methods that incorporate spatial information into the imputation model yielded far better reproduction of spatial variability than the secondary method. Perhaps more surprising, however, is that the merged methods yield better spatial continuity reproduction than the primary method. Integrating the colocated information improves the accuracy of local estimates, which in turn leads to better reproduction of the spatial structure.

The results are more mixed in terms of bivariate reproduction, though this was expected based on critical differences in the data. The secondary method reproduces correlation the best for both datasets, but it yields inferior reproduction of the complex bivariate features (Figures 3.11 and 3.19). The reason for this discrepancy is the bivariate properties of the two datasets. The Gaussian assumptions of the secondary method are correctly applied with the Gaussian data, allowing for it to converge on the correct covariance. In the case of the complex data, however, the covariance fails to capture the bivariate distribution of the data. Consequently, the NPM method outperforms the secondary method and parametric merged methods in non-Gaussian settings.

These examples support the simultaneous inclusion of spatial and colocated information when constructing conditional distributions for MI. Since the NPM method takes longer to execute and has additional complexities (due in both cases to its use of KDE), there is no perceived value in using the method with data that exhibit bivariate Gaussianity. Of course, such simplicity will rarely be encountered with geological data in practice. When applied to complex data, the NPM method produces substantially better results than its parametric equivalent. This conclusion is supported by the case study results in Chapter 7.

3.5 Discussion

The following section discusses practical problems and implementation decisions for the described imputation methods. Note that the described implementations have been used in the Fortran programs that generated the presented results.

3.5.1 Extraction from the Gibbs Sequence

As summarized in Casella and George (1992), several strategies are available for extracting values from a Gibbs sequence to record L realizations of each missing value. A theoretically attractive method executes L independent sequences using L random starting locations. The Gibbs sampler proceeds until convergence is observed for a property of interest considering values of the current iteration across the L sequences. Values of the final iteration for the L sequences are then recorded as the L realizations. This strategy is attractive since it reduces dependence on the starting random value and virtually insures that the recorded values are independent and identically distributed (iid) with respect to each other. It may not be computationally feasible, however, depending on:

i) The number of missing values that the sequence must proceed through.

ii) The number of iterations that are required for convergence.

iii) The chosen method for constructing the conditional distributions. Note that the NPM method is computationally expensive in this regard due to its use of KDE.

Observing these computational challenges, a second option is to execute one long sequence, extracting the values of every r number of iterations until L realizations are recorded (Geyer, 1991). The extracted values are approximately iid for large enough values of r. While this strategy may be considered for the parametric methods that are described in the previous sections, it was found to remain too computationally expensive for the NPM method with typical mining datasets (>10,000 data). The third and most computationally attractive option is to extract consecutive L iterations. Although the realizations are dependent, George and McCulloch (1991) shows that the extractive values will still converge on the underlying joint distribution.

Considering the positives and negatives of each approach, a hybrid of the first and third options has been implemented. User specified b burn-in iterations are executed before extracting the next L iterations. The burn-in iterations reduce dependence on the random starting values, while recording the next L iterations makes the NPM method feasible for large datasets. There is some dependency between realizations using this method; the consequences that this may have on the imputation uncertainty is examined in Chapter 7. A sufficient number of b iterations was found to be critical in terms of increasing the accuracy of the extracted L iterations. Methodology for determining the value of b is a focus of future work (Chapter 8), though it is increases with increasing K and decreases with increasing spatial correlation. Based on jackknife accuracy testing, $b = 2$ is required for the examples in this chapter, while $b = 8$ is required for the case study in Chapter 7.

3.5.2 Path of the Gibbs Sequence

The simulation path is an important consideration for most sequential simulation schemes; the path that the Gibbs sequence takes through the missing values is thought to be no exception. No mention of this practical detail was found in the review of missing data literature; it may be inferred that simulation path is not considered to be of particular import with MI.

A random path is commonly implemented for geostatistical algorithms (Deutsch and Journel, 1998), though there is a concern that this may not be appropriate for imputation. Consider the widely varying uncertainty that could exist for missing values. Some missing values may have almost no conditioning information in terms of colocated secondary values and spatially correlated primary values (sampled rather than previously imputed). A simplified example of this is shown in Figure 1.5, where the conditional distributions of missing values at location 5 are very uncertain relative to the single missing value at location 1. If the simulation path were to start at highly uncertain locations, it is possible that subsequent simulation of the highly certain locations would become 'trapped'. That is to say, the conditioning of previously simulated (but highly uncertain) values could lead to a primary distribution that fails to corroborate the secondary distribution. In turn, this could negatively influence the merged distribution and lead to inferior accuracy of the imputed results. This issue arises due to the Markov model of coregionalization that is implicit to the methods. An imputation scheme that uses cross-covariance, such as the linear

model of coregionalization (LMC) would not suffer from this problem. Coregionalization models that consider cross-covariance have been avoided for practical reasons that were discussed in Section 2.1.3.

This issue is likely overcome with a sufficient number of b iterations for the burn-in sequence. Due to the potential computational expense of each iteration, however, a simulation path is used that avoids it by construction. Prior to executing the Gibbs sampler, the merged method is used to calculate the conditional variance, σ_m^2 (Equations 3.9 and 3.10). Treating σ_m^2 as a proxy of uncertainty, the simulation path will then be ordered according to increasing σ_m^2. This approach is analogous to using simple kriging variance for determining the simulation path of SGS.

Unlike in the Gibbs sampler, only the sampled data is used for determining this uncertainty. The primary and secondary variances that are input to Equations 3.9 and 3.10 are impacted by this change. Variance of the primary distribution (Equations 3.4 and 3.5) is calculated using the number of sampled primary values, $n_s \leq n-1$, rather than $n-1$ sampled and previously imputed values. Variance of the secondary distribution (Equations 3.7 and 3.8) is calculated using the number of colocated sampled variables, $K_s \leq K-1$, rather than $K-1$ sampled and previously imputed variables.

The NPM method is not used for this step since unlike its parametric equivalent, the conditional variance of each missing value cannot be calculated without the $K-1$ secondary variables present at each colocated location. The simulation path will not have separate loops for each variable or location, but varies between variables and locations as each σ_m^2 dictates.

3.5.3 Poorly Informed Bivariate Pairs

The covariance matrix of the normal score data, $\mathbf{\Sigma} : C_{ij}, i, j = 1, ..., K$, plays a central role in the described imputation methods. The merged method uses terms from $\mathbf{\Sigma}$ when calculating the secondary weights in Equation 3.8. The NPM method uses a kernel bandwidth matrix, $\mathbf{H}$, that is scaled by $\mathbf{\Sigma}$. Clearly, it is important that $\mathbf{\Sigma}$ is calculated in a reliable manner. A covariance matrix is traditionally calculated using only homotopic observations to insure that the result is positive-semidefinite. This could pose problems in practice, however, since certain variables may rarely be sampled in colocated locations. In the extreme case, some variables may never be sampled together, meaning that no samples are available for calculating $\mathbf{\Sigma}$.

Despite missing at least one variable, a heterotopic observation may continue to sample the i^{th} and j^{th} variables; it could therefore be considered in the calculation of C_{ij}. In circumstances where only using homotopic observations severely impacts the sample population that is available for calculating $\mathbf{\Sigma}$, it is advocated that all sampled bivariate pairs are used for the calculation of their associated C_{ij} term.

While the resulting heterotopic covariance matrix is not guaranteed to be positive semi-definite, optimal corrections exist for that purpose. A weighted correction proposed by Kumar and Deutsch (2009) is chosen so that some control is maintained on the changes. Consider that each C_{ij} will have varying confidence based on the number of samples that underlies its calculation. The correction is then given by Equation 3.13, where iterative optimization is used to find the positive semi-definite matrix $\Psi : \psi_{ij}, i, j = 1, ..., K$ that minimizes changes to the input covariance matrix. The $\mathbf{\Omega} : \boldsymbol{\omega}_{ij}, i, j = 1, ..., K$ matrix weighs the influence that each component has on the optimization, so that changes may be reduced for C_{ij} that are relatively certain (and vice versa).

$$Minimize : \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \boldsymbol{\omega}_{ij} * (\psi_{ij} - C_{ij})^2 \tag{3.13}$$

The confidence of each C_{ij} does not increase in a linear fashion with respect to the number of sampled data, n_{ij}, used in its calculation. Instead, the covariance confidence relates to the number of independent sampled data, n'_{ij} (Niven and Deutsch, 2008). This value is calculated based on the spatial covariance that exists between the data. Increasing covariance leads to increasing redundancy and decreasing n'_{ij}. The value of n'_{ij} is calculated as:

$$\boldsymbol{\omega}_{i,j} = n'_{ij} = \frac{n_{ij}^2}{\sum_{\alpha=1}^{n_{ij}} \sum_{\beta=1}^{n_{ij}} C_{\alpha\beta}} \tag{3.14}$$

Here, $C_{\alpha\beta}$ is the auto-covariance that exists been the α and β locations where both the i^{th} and j^{th} variables are sampled. As shown, each term of $\mathbf{\Omega}$ in Equation 3.13 is assigned based on the associated number of independent data.

Practically speaking, $C_{\alpha\beta}$ is calculated using the semivariogram model of the i^{th} variable, which will already be required as input to the imputation program. This step therefore requires no additional effort from the user. A binary option specifies whether covariance should be calculated using homotopic or heterotopic data. If

the latter option is selected and the result is not positive semi-definite, then the described correction is applied.

3.5.4 Non-colocated Variables

The previous section presented a methodology for addressing situations where bivariate pairs are poorly sampled. A problem remains, however, if two variables are never colocated, meaning that no information is available for their bivariate pair at $h = 0$. If the merged method is being used, then only the covariance, C_{ij}, is required for informing the relationship between the i^{th} and j^{th} variables. One may consider estimating C_{ij} based on the cross-covariance of the two variables. As detailed by Minnitt and Deutsch (2014), this amounts to extrapolating $C_{ij}(h)$ for $h > 0$ to the $h = 0$ lag distance.

Unfortunately, there is no understood solution for applying the NPM method with non-colocated variables. Only homotopic observations may be considered for conditioning the multivariate KDE (Equation 3.11). While this is an interesting topic of future research, the current work is limited to the merged method when imputing non-colocated variables.

Chapter 4

Exploratory Multivariate Transformations

As detailed in Section 2.2, no technique is available for transforming multivariate data of arbitrary form and dimension to be multiGaussian. Several transformations were developed to address this issue over the course of this thesis, including conditional standardization (CS), the multivariate standard normal transformation (MSNT) and the projection pursuit multivariate transform (PPMT). While the PPMT has emerged as the preferred technique, it is a reflection of challenges that were encountered in development of the former transformations. Further, the PPMT was conceived within the Gaussian mapping transformation framework that was originally proposed for the MSNT.

The following chapter provides a brief overview of the CS and MSNT techniques. Background is provided on the challenges that motivate the PPMT. The PPMT is described and demonstrated in far greater detail in Chapter 5.

4.1 Conditional Standardization

CS (Barnett and Deutsch, 2012b) transforms non-linear and heteroscedastic data to approach linearity and homoscedasticity. While not specifically targeting a multiGaussian distribution, CS yields distributions that may be more suitable for cosimulation frameworks or subsequent linear decorrelation transformations. The original non-linearity and heteroscedasticity is returned to simulated realizations following back-transformation.

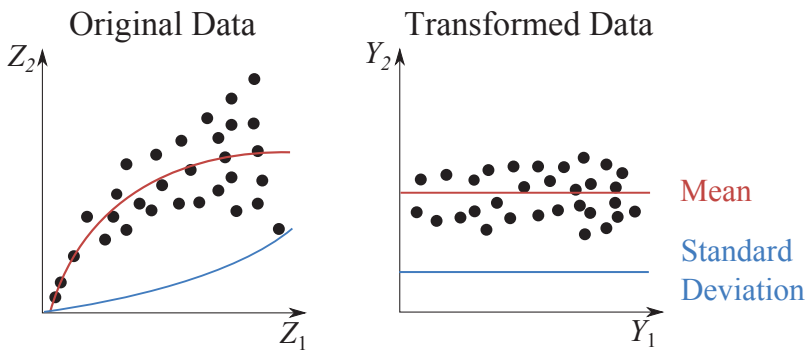


Figure 4.1: Schematic illustration of conditional standardization. Given a non-linear and heteroscedastic bivariate distribution (left), the subtraction of its conditional mean and division of its conditional standard deviation yields a linear and homoscedastic distribution (right).

4.1.1 Forward and Back Transformations

Consider a bivariate distribution consisting of two variables, Z_1 and Z_2, that constitute an n x 2 data matrix, $\mathbf{Z}$. The relationship between these variables may be non-linear and heteroscedastic, such as the schematic distribution in Figure 4.1. Subtracting the Z_2 values by a function that describes the mean of Z_2 conditional to the colocated values of Z_1, will yield a residual distribution that is approximately linear. Similarly, dividing Z_2 by a function that describes the standard deviation of Z_2 conditional to Z_1 will yield a distribution that is approximately homoscedastic.

The bivariate form of CS is given by Equation 4.1, which yields the second column of the transformed data matrix, $\mathbf{Y}$. The first column of $\mathbf{Y}$ is identical to the first column of $\mathbf{Z}$, as the first conditioning variable is not altered by CS. In other words, $z_1(\mathbf{u}_\alpha)$ values could be used interchangeably with the $y_1(\mathbf{u}_\alpha)$ values in Equation 4.1.

$$y_2(\mathbf{u}_\alpha) = \frac{z_2(\mathbf{u}_\alpha) - E\{z_2(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha)\}}{\sqrt{Var\{z_2(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha)\}}}, \alpha = 1, ..., n \tag{4.1}$$

The conditional mean and standard deviation functions may be derived parametrically using forms of regression or non-parametrically through discretizing the distribution according to probability classes of the conditioning variable. This concept may also be extended to higher dimensions, where a variable is transformed conditional to two or more variables. The generalized form of the transformation is given by Equation 4.2, which yields columns 2 to K of the nxK matrix, $\mathbf{Y}$.

$$y_i(\mathbf{u}_\alpha) = \frac{z_i(\mathbf{u}_\alpha) - E\{z_i(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha), ..., y_{i-1}(\mathbf{u}_\alpha)\}}{\sqrt{Var\{z_i(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha), ..., y_{i-1}(\mathbf{u}_\alpha)\}}}, \alpha = 1, ..., n, i = 2, ..., K \quad (4.2)$$

The conditional mean and standard deviation functions are saved, facilitating the back-transformation:

$$z_i(\mathbf{u}_\alpha) = y_i(\mathbf{u}_\alpha) \cdot \sqrt{Var\{z_i(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha), ..., y_{i-1}(\mathbf{u}_\alpha)\}} + E\{z_i(\mathbf{u}_\alpha)|y_1(\mathbf{u}_\alpha), ..., y_{i-1}(\mathbf{u}_\alpha)\}, \text{ for } \alpha = 1, ..., N, i = 2, ..., K \quad (4.3)$$

Equation 4.3 restores the heteroscedasticity and non-linearity to the simulated realizations.

4.1.2 Practical Challenges

The success of this transform is dependent on conditional mean and standard deviation functions that accurately characterize the non-linearity and heteroscedasticity of the distribution. The non-parametric approach is generally expected to produce superior results, since no assumptions of the functional form of the distribution must be made. Parametric application may still be considered, however, in cases where a low number of n observations, or high number of K variables makes the non-parametric approach impractical.

Should the non-parametric approach be chosen, there is no strict rule regarding the number of classes that are required for partitioning the conditioning variable, or the number of observations that are required in each bin for the subsequent calculations of mean and standard deviation. The fewer the classes, the more likely that complex features will remain within the partitioned bins following transformation. Conversely, increasing the number of classes decreases the likelihood of stable conditional statistics. This relationship between the number of observations and the number of bins is analogous to the challenges with SCT that were discussed in Section 2.2.2. As with the SCT, smoothing algorithms where data beyond the class partitions often help, but based on observation it is unlikely that greater than three conditioning variables will be viable. In these cases practitioners may choose between either using a parametric calculation of the conditional functions, or a 'nested' application of the non-parametric approach.

Nested application refers to using only one, two or three conditioning variables to remove complex features from the higher order conditioned variables. Again, this approach was originally proposed for the SCT by Leuangthong and Deutsch (2003). Addressing these selected relationships will often resolve the majority of the complexity between variables that are not transformed conditional to one another. This is not guaranteed, however, and careful decision making must take place regarding the ordering of these variables. Considerations may include:

i) Reproduction of the multivariate relationships between the primary resource variable and secondary variables; this requires that all secondary variables are transformed conditional to the resource variable.

ii) Reproduction of a multivariate relationship between secondary variables that are critical to process performance; one secondary variable must condition the other.

iii) Reproduction of bivariate relationships that exhibit significant complexity; a well behaved transformed distribution will likely require that one variable conditions the other. As discussed by Friedman (1987), complexities in a bivariate distribution are shadows of sharper complexities in higher order dimensions. It follows that bivariate complexities are likely to persist if they are not directly targeted for removal.

The above considerations often lead to difficult decision making, as all of them are unlikely to be satisfied with a nested CS application. Such challenges motivate the parametric approach, which was the original concept when CS was conceived as a method to avoid the curse of dimensionality. The initial parametric approach used least squares regression with linear, squared and cubic functions of the conditioning variables; the specific functions were determined based on the best resulting RMSE of the regression. Ultimately, however, testing found that it was very difficult to fit complex geological data with these mathematical functions. If the conditional functions do not adequately characterize the true relationships, CS will fail to remove the complexity.

4.2 Multivariate Standard Normal Transformation

Observing the shortcomings of transformations such as CS and SCT, the MSNT (Barnett and Deutsch, 2012a; Deutsch, 2011) was proposed for the transformation of complex multivariate data to an uncorrelated multiGaussian distribution. The MSNT was the first method to be proposed under what is now called the Gaussian mapping (GM) transformation framework, which is not sensitive to an increasing number of variables.

4.2.1 Gaussian Mapping

The objective of GM is to map the sampled data, $\mathbf{Z}$, to an uncorrelated multi-Gaussian distribution of matching dimensions and observations, $\mathbf{Y}$. This mapping is recorded as a multivariate transform table, where observations in the original distribution are associated with multiGaussian values. This concept is presented in Figure 4.2 where arrows represent the mapping. Following independent geostatistical modeling of each variable, Gaussian realizations may be back-transformed through interpolating based on the mapped data. This back-transform scheme is shown in the bottom of Figure 4.2, where a single simulated node (square) is interpolated in original space based on its distance to the nearest mapped data (circles) in transformed space.

Observe that GM amounts to the multivariate extension of a normal score transformation. Unlike the SCT and non-parametric CS, no binning or gridding is performed. The transform is defined using only the original data and their multi-Gaussian equivalent, meaning that it can be applied to any arbitrary number of K variables. The specific forward transform that defines a GM and the back-transform that applies it are left purposefully ambiguous for now. Those details are what characterize specific transformations such as the MSNT and PPMT. Properties that are required for a successful GM are discussed first.

The forward transform should map the data to an uncorrelated multiGaussian distribution in a manner that minimizes changes to the relative multivariate configuration of the original distribution. Consider quantifying the relative multivariate configuration using the distance between each observation. It follows that a successful GM should minimize changes to the distances between observations in original and transformed space. When neighbours are mapped adjacent to one another, the

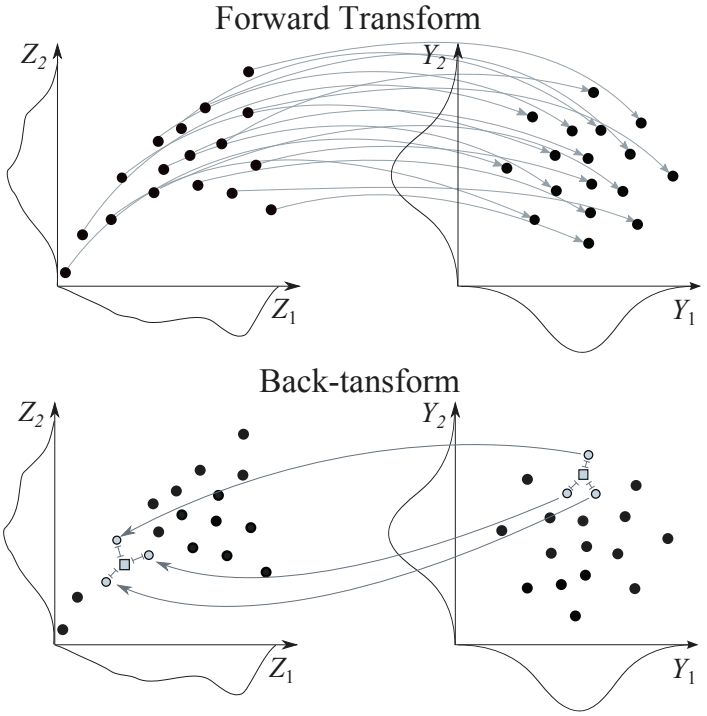


Figure 4.2: Schematic illustration of the forward and back GM transform framework.

back-transform in Figure 4.2 will cause nearby simulated values to be mapped back within their locale. The original variability and declustered joint density of the data should be closely reproduced by realizations in these cases. On the contrary, if the mapping is completely randomized, the back-transform results will converge towards the global mean, dramatically reducing the variability. A poor mapping is illustrated in Figure 4.3, where data that are far apart in original space (grey circles) have been mapped near to each other in transformed space. Observe that this causes the back-transform of a simulated node (square) to be interpolated toward the global mean. If repeated over all simulated nodes, back-transforming with a poor mapping will lead to a severe loss of variability.

Another potential consequence of a poor mapping is a loss of spatial continuity. A systematic increase in the relative multivariate distances will inevitably lead to a decrease in the spatial structure. A random mapping results in semivariograms that are pure nugget effect, $C_{ii}(h) \simeq 0 \ \forall h, i$.

Shifting focus to the GM back-transformation, the chosen interpolation scheme should preserve the configuration between a simulated node and the nearby mapped

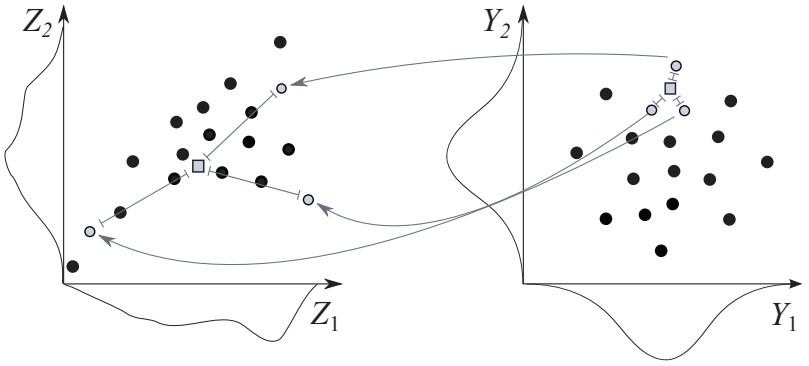


Figure 4.3: Schematic illustration of the GM back-transform based on a poor mapping.

observations. More specifically, the relative distances between a simulated node and the mapped observations in multiGaussian space should be maintained after back-transforming to original space. Assuming an effective forward mapping, this should lead to reproduction of key statistics. It is worth emphasizing that the variability of the original data should be reproduced by back-transformed realizations. A characteristic of many interpolation schemes is smoothed results, which is a concern for this application.

4.2.2 Forward Transformation

The following section will outline the forward MSNT transformation, which may be subdivided into three steps: i) assemble the univariate multivariate Gaussian distributions, ii) initial mapping through dimension reduction, and iii) final mapping through simulated annealing.

Step 1: Assemble the Univariate and Multivariate Gaussian Distributions

As subsequent mapping steps will revolve around the Euclidean distance between observations, the MSNT is very sensitive to drastically different units or outlier values. To address these issues simultaneously, the normal score transform (Section 2.1.2) is applied to transform the K variables of the original data, $\mathbf{Z}$, to the univariate standard Gaussian data, $\mathbf{X}$.

The transformed data, $\mathbf{X}$, is the origin of the subsequent Gaussian mapping. Uncorrelated multiGaussian data, $\mathbf{Y}$, of a matching number of K variables and n observations is now required as the destination of the Gaussian mapping. This multiGaussian data is generated according to Equation 4.4, where G is the standard

Gaussian CDF and p are random probabilities.

$$\mathbf{Y}=G^{-1}\left(\begin{bmatrix} p_{11} & \cdots & p_{K,1} \\ \vdots & p_{\alpha,i} & \vdots \\ p_{n1,} & \cdots & p_{n,K} \end{bmatrix}\right), p_{\alpha,i} \in [0,1] \tag{4.4}$$

MCS could be considered for generating p, but doing so is likely to generate distributions that deviate from the multiGaussian model when n is relatively small. Consider the example in Figure 4.4, where MCS is used to generate data of $K = 2$ and $n = 280$. The plots in this figure are output from the `scatnscores` program (Deutsch and Deutsch, 2011), which plots the correlated multiGaussian probability density contours of 0.25, 0.5 and 0.95 for reference. The program also performs a bivariate standard normal (BVSN) Gaussianity test by comparing the expected and observed density of data that fall in each of these density contours (by quadrant).

While the absence of red text in Figure 4.4 indicates that the MCS sample has passed the BVSN test, the observations are not dispersed in a manner that closely mimics the multiGaussian contours. While expected given the low value of n, this is a cause for concern given the back-transform scheme that is displayed in Figure 4.2. Simulated nodes are interpolated towards the clustered mapped observations in a manner that is not representative of the underlying Gaussian model. Further, the MCS sample is not entirely uncorrelated and will require subsequent decorrelation to be representative of the independently simulated variables.

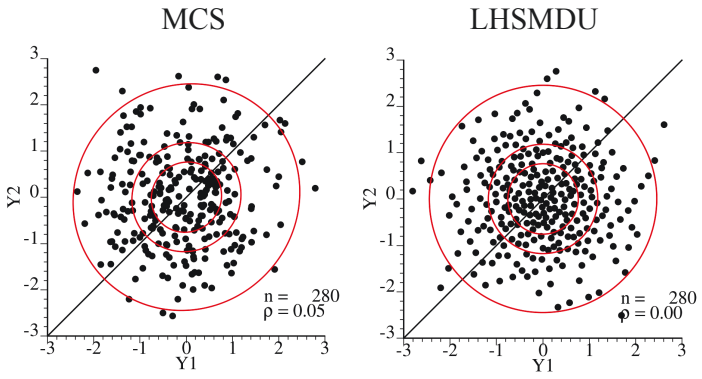


Figure 4.4: `scatnscores` plots and Gaussianity tests of random Gaussian distributions that have been generated using MCS and LHSMDU.

Observing these deficiencies with MCS, latin hypercube sampling with multi-dimensional uniformity (LHSMDU) (Deutsch and Deutsch, 2012b) is applied to generate uncorrelated distributions of p that are uniformly distributed in multi-

variate space. Applying the LHSMDU probabilities with Equation 4.4 yields the results that are displayed in Figure 4.4. While the LHSMDU results continue to exhibit deviations from a perfect bivariate Gaussian cloud, the data are now entirely uncorrelated and evenly distributed.

Step 2: Initial Mapping Through Dimension Reduction

An nx1 vector of indices, $\mathbf{m}$, will now be used to map the univariate Gaussian data, $\mathbf{X}$, to the multiGaussian data, $\mathbf{Y}$. While $\mathbf{m}$ is heuristically optimized in the next step, a reasonable initial mapping will significantly reduce the number of iterations that are required for convergence. This amounts to a dimension reduction problem, as multivariate observations of two distributions must be described by a single measurement that allows for the best possible alignment.

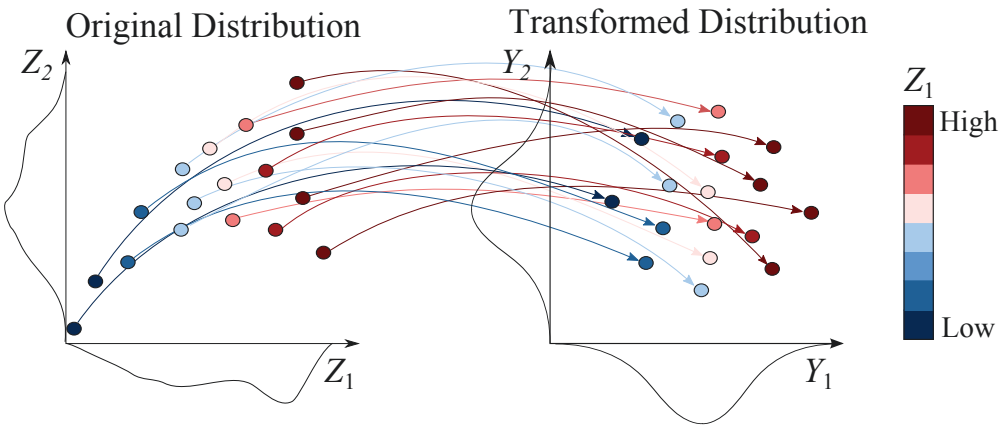


Figure 4.5: Schematic illustration of an initial mapping based on the rank order of Z_1. Also introduces the concept of coloring transformed values by the original values as means for visualizing changes to the multivariate configuration.

One may be influenced by practical priorities for this dimension reduction. Consider basing the initial mapping on the rank order of an original variable of particular interest (e.g., resource variable, the most spatially continuous variable, etc.). Figure 4.5 displays a schematic illustration of an initial mapping based on the rank order of an original variable, Z_1. The arrows in this figure represent $\mathbf{m}$; they are colored by the associated Z_1 value to help visualize changes that are incurred to the multivariate configuration. Observe that while this mapping successfully preserves the rank order of Z_1, large changes are seen in the overall multivariate configuration.

A better option for this dimension reduction is to use PCA (Section 2.2.1), which yields an eigenvector matrix, $\mathbf{V}$, that is sorted according to the variability that each

vector explains. The first row provides a vector rotation of the data that maximizes the variability that one dimension can explain. As a result, multiplying $\mathbf{X}$ and $\mathbf{Y}$ by the first row of $\mathbf{V}$ yields a single measurement that reasonably approximates how closely mapped each multivariate observation should be. Relative to using original variables, this PCA approach to the initial mapping was found to significantly reduce subsequent convergence time.

Step 3: Final Mapping Through Simulated Annealing

Given a mapping index, $\mathbf{m}$, let $d_{x(\alpha,\beta)}$ be the Euclidean distance between the α and β observations in the univariate Gaussian distribution, $\mathbf{X}$, while $d_{y(\alpha,\beta)}$ is the distance between those same observations in the multiGaussian distribution, $\mathbf{Y}$. The final step of the MSNT will minimize changes to $d_{x(\alpha,\beta)}$ according to the objective function:.

$$\min \left[O = \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} \left(d^2_{x(\alpha,\beta)} - d^2_{y(\alpha,\beta)} \right)^2 \right] \tag{4.5}$$

This optimization may not be solved using global linear or convex solvers. Instead, it is accomplished through randomly perturbing $\mathbf{m}$ in a pairwise fashion using a simulated annealing framework (Deutsch, 1992; Metropolis et al., 1953). The decision to square the distances in Equation 4.5 was made after testing a range of powers with multiple datasets. It heavily penalizes observations that are far apart in $\mathbf{X}$ but near in $\mathbf{Y}$ (and vice versa). After a sufficient number of iterations, $\mathbf{m}$ converges on a mapping that minimizes changes to the relative multivariate configuration. Practically speaking, $\mathbf{m}$ is recorded by constructing a table that places the original data observations from $\mathbf{Z}$ with their associated transformed observations from $\mathbf{Y}$. This transformation table facilitates the back-transformation that is described in the next section.

4.2.3 Back-transformation

To simplify presentation of the MSNT back-transformation, consider the original and transformed data matrices as n vectors of size K: $\mathbf{z}_\alpha, \alpha = 1, ..., n$ and $\mathbf{y}_\alpha, \alpha = 1, ..., n$, respectively. Redefine $d_{y(\alpha,\beta)}, \alpha = 1, ..., n$ as the Euclidean distances between a simulated node, $\mathbf{y}_\beta$, and the $\alpha = 1, ..., n$ mapped observations in multiGaussian space. Three $d_{y(\alpha,\beta)}$ distances are represented by the distance symbols within the trans-

formed data in Figure 4.2. As illustrated in this figure, the back-transform aims to preserve the relative distance between the simulated node and its nearest mapped neighbours. This is achieved using Equation 4.6, where original values of the simulated node, $\mathbf{z}_\beta$, are calculated based on the nearest $K+1$ data observations in multiGaussian space. Here, $\mathbf{z}_\alpha$ denotes the original values for the nearest $\mathbf{y}_\alpha$ observation in transformed space. The associated weight, λ_α, is shown to be inversely proportional to $d_{y(\alpha,\beta)}$.

$$\mathbf{z}_\beta = \sum_{\alpha=1}^{K+1} \lambda_\alpha \mathbf{z}_\alpha, \text{ where } \lambda_\alpha \propto \frac{1}{d_{y(\alpha,\beta)}} \text{and } \sum_{\alpha=1}^{K+1} \lambda_\alpha = 1 \text{ for } \beta = 1, ..., N \tag{4.6}$$

Any number of nearest mapped observations could be used for this interpolation, but it is recommended that the number of observations be $K+1$. Using less than this does not adequately constrain the multivariate interpolation, but increasing beyond $K+1$ will begin to converge the back-transformed results towards the global mean. Continuing to draw an analogy with the normal score back-transform, observe that two data observations are used for the linear interpolation of that 1-D transform.

Rather than assigning weights as a linear inverse function of distance (Equation 4.6), consider calculating those weights using more complex systems of equations such as variants of the normal equations. Doing so would incorporate the covariance between the simulated node and nearby mapped observations, as well as the covariance between those observations to determine the optimal weight. Given the small number of observations that are involved in each estimate ($K+1$), however, preliminary testing found that weighting schemes such as ordinary and universal kriging (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978) offered negligible improvements to the results.

4.2.4 Practical Challenges

The concept of the MSNT is novel and conceptually simple. A mapping is established between the original and multiGaussian data of matching K variables and n observations. Following independent Gaussian simulation, the original complexity is restored using a weighted interpolation of the mapped data. The MSNT was conceived as a method that would not be sensitive to increasing K variables. Following the described implementation details, it was found to be an effective method on small test cases.

Unfortunately, however, the MSNT was also found to be very sensitive to increasing n observations. The convergence time for minimizing Equation 4.5 generally increases as a function of n at beyond a quadratic rate. Results from one test case are displayed in Figure 4.6 (Barnett and Deutsch, 2012a), where the execution time becomes prohibitive for $n > 500$.

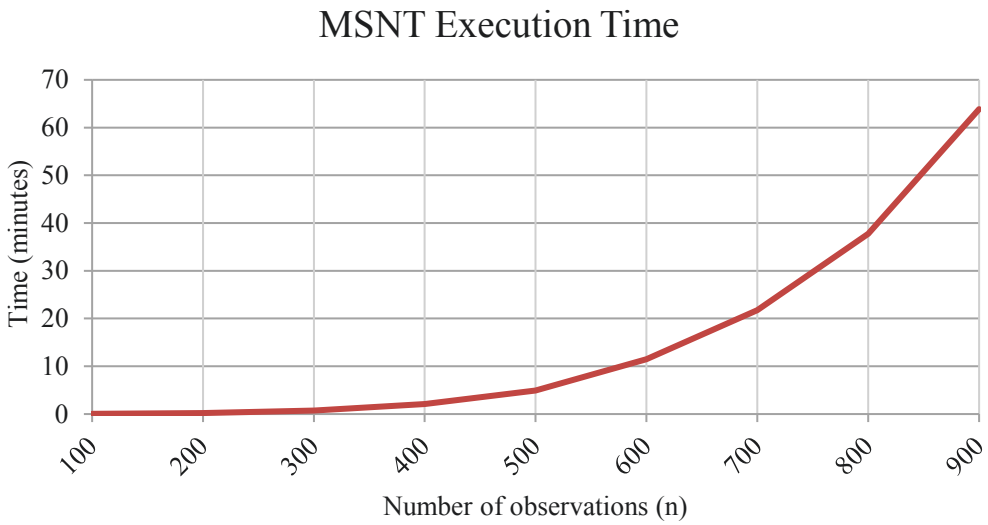


Figure 4.6: MSNT execution time as a function of increasing n observations.

Consequently, the MSNT may not be practical for geological datasets that exceed 500 to 2,000 observations. A range exists that is dependent on practitioner patience and particulars of the multivariate data (the amount of optimization that is required following the initial mapping). As n will frequently exceed 1,000 observations for geological datasets (even after domaining), the MSNT is not applicable to a wide range of practical settings. This problem is exacerbated if using a multiple imputation geostatistical workflow (Section 3.1), since the MSNT will have to be executed for the $l = 1, ..., L$ data realizations.

Chapter 5

Projection Pursuit Multivariate Transformation

The projection pursuit multivariate transformation (PPMT) (Barnett et al., 2014) is the second primary contribution of this thesis. The PPMT applies a modified component of the projection pursuit density estimation (PPDE) algorithm to transform potentially complex and high dimensional geological data to be multiGaussian and uncorrelated. This facilitates independent geostatistical modeling of the variables, before the back-transform restores the original complexity and correlation to simulated realizations. The PPMT is easier to apply than competing transforms and is demonstrated to yield superior modeling results.

Conditional standardization and the multivariate standard normal transformation (MSNT) (Chapter 4) showed conceptual promise in this context since they may be applied to data of any form and dimension. As discussed, however, the quality of their results deteriorate with practical datasets of increasing variables and observations, respectively. Unlike those methods, the PPMT has effectively transformed numerous geological datasets of practical size and dimension.

The chapter begins with an overview of the forward transform, where changes in the conventional PPDE algorithm are highlighted. The PPMT was originally conceived within the Gaussian mapping (GM) framework that was introduced for the MSNT in Section 4.2.1. The PPMT provides a better mapping than the MSNT for large datasets, leading to superior performance with the GM back-transform. Another option for the PPMT back-transform, however, is simply to reverse the forward transformation steps.

The PPMT transformations and geostatistical modeling workflow are demonstrated with a small bivariate example. Alternative modeling workflows are also

applied to provide a benchmark for judging the PPMT results. The chapter concludes with a discussion on practical PPMT considerations and implementation details.

5.1 Forward Transformation

First introduced by Friedman and Tukey (1974), PPDE is a non-parametric technique that is used to determine the PDF of a multivariate distribution. Relative to other density estimation techniques, PPDE performs particularly well with complex and/or high dimensional data (Hwang et al., 1994). As the projection of a multi-Gaussian distribution is also Gaussian, the premise is to detect vectors that yield the most non-Gaussian projections of the data and correct them (Huber, 1985).

The assumption is made that non-Gaussian structures in higher dimensions are exhibited in the lower dimensional projection. Friedman (1987) discusses that projections exhibit a smoothed shadow of what are likely more marked complexities in the higher dimensions. Once the most interesting projection vector has been determined, the multivariate data may be transformed to normalize their projection (termed Gaussianize in the literature). Iterating this search and Gaussianize algorithm, the data is gradually transformed to be multiGaussian. The final step of the PPDE algorithm involves estimating the multivariate density through combining the univariate projections. This final step is not used in the proposed application.

The forward PPMT very closely resembles the PPDE algorithm that has been conceptually described. Major PPMT steps will now be outlined, where deviations from the PPDE algorithm given in Friedman (1987) are highlighted. Though many transformation steps are involved in the PPMT, they are quickly executed by a single Fortran program.

5.1.1 Pre-processing

The first steps of the PPMT will transform the n observation and K variable data matrix, $\mathbf{Z} : z_{\alpha i} : \alpha = 1, ..., n, i = 1, ..., K$, to have properties that are suitable for the subsequent projection pursuit algorithm.

Normal Score Transformation

The first step of PPDE is data sphereing, which is introduced as the second step of the PPMT in the next section. Data sphereing is very sensitive to outliers since

it is based on the $h = 0$ covariance matrix, $\mathbf{\Sigma}_Z$. Friedman discusses that PPDE may benefit from transformations that make the marginal variables more Gaussian. As projection pursuit is mainly concerned with resolving multivariate complexity, this pre-processing would allow for the algorithm to be less influenced by marginal features.

The familiar normal score transformation is used as the first step of the PPMT to simultaneously address these concerns. Each of the K variables are transformed according to Equation 2.3, yielding the univariate Gaussian nxK data matrix, $\mathbf{Y}$. Though multivariate outliers may persist to influence $\mathbf{\Sigma}_Y$, this normal score transform addresses univariate outliers and was found to reduce the required number of projection pursuit iterations.

Data Sphereing

The projection pursuit algorithm will benefit from variables that have an identity covariance matrix, $\mathbf{\Sigma} = \mathbf{I}$, meaning that they are orthogonal with a variance of one. These properties are achieved by data sphereing, which is the second and final pre-processor of the PPMT. The conventional data sphereing calculation is given by Equation 5.1, where $\mathbf{S}^{-1/2}$ is referred to as the sphereing matrix. Spectral decomposition of $\mathbf{\Sigma}_Y$ provides the eigenvector matrix, $\mathbf{V} : v_{i,j}, i, j = 1, ...K$, and the diagonal eigenvalue matrix, $\mathbf{D} : d_{i,j}, i, j = 1, ...K$, according to $\mathbf{\Sigma}_Y = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}}$.

$$\mathbf{X} = (\mathbf{Y} - E\{\mathbf{Y}\})\,\mathbf{S}^{-1/2}, \text{ where } \mathbf{S}^{-1/2} = \mathbf{V}\mathbf{D}^{-1/2} \tag{5.1}$$

Equation 5.1 presents the conventional transformation as it appears in Friedman (1987). Note that the centering portion, $\mathbf{Y} - E\{\mathbf{Y}\}$, is not necessary here since $\mathbf{Y}$ is marginally standard Gaussian (and therefore centered). Observe that the multiplication of $\mathbf{D}^{-1/2}$ is the only difference between data sphereing and PCA (Equation 2.6). The multiplication of $\mathbf{V}$ rotates the variables to an orthogonal axis (PCA), before $\mathbf{D}^{-1/2}$ transforms them to have a variance of one. As with PCA, the descending columns of $\mathbf{X}$ will explain a decreasing amount of variability in the multivariate system. The percentage of variability that is explained by each variable, X_i, is calculated as $d_{i,i}/\mathrm{tr}(\mathbf{D})$.

While appropriate for PPDE, this implementation of sphereing will be problematic in the PPMT context for two reasons. First, the GM framework (Section 4.2.1) does not account for transformed variables that explain widely different variability

in original space. More specifically, the interpolation back-transformation (Equation 4.6) assumes that the Euclidean distance in each dimension of transformed space may be considered equally when calculating the weights.

Second, maximizing the multivariate variability that each descending sphere variable explains will increase mixing of the original variables in transformed space. This may not be advantantageous from a geostatistical modeling perspective, as the original variables may have distinct univariate and spatial characteristics that are better kept separate to the maximum possible extent.

More specifically, attempting to load $Y_i, i = 1, ..., K$ onto the first few X_i variables effectively increases their mixing in transformed space. This decreases the likelihood that the distinct characteristics of each Y_i are recovered following geostatistical simulation and back-transformation. Note that a loading, ρ', is closely related to the correlation, ρ, between an original, Y_i, and transformed, X_j, variable according to:

$$\rho'(Y_i, X_j) = v_{i,j} \cdot d_{j,j} = \rho(Y_i, X_j) \cdot \sqrt{Var\{Y_i\}} \tag{5.2}$$

Equation 5.2 shows that a loading is simply the correlation between the original and transformed variables, scaled by the standard deviation of the original variable. Applying Equation 5.1 will maximize the absolute value of $\rho'(Y_i, X_1)$ for $i = 1, ..., K$. Observing these potential challenges, an alternative form of data sphereing (Fukunaga, 1972; Hwang et al., 1994) has been implemented for the PPMT:

$$\mathbf{X} = \mathbf{Y}\mathbf{S}^{-1/2}, \text{ where } \mathbf{S}^{-1/2} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}} \tag{5.3}$$

Aside from the centering portion (dropped for the above noted reasons), the only difference between Equations 5.1 and 5.3 is the additional multiplication of $\mathbf{V}^{\mathrm{T}}$, which projects the orthogonal variables back onto the basis of the original variables. The transformed variables still have the required identity covariance matrix, but the rotation is now performed in a manner that maximizes the absolute value of $\rho'(Y_i, X_j)$ for $i = j$, while minimizing the absolute value of $\rho'(Y_i, X_j)$ for $i \neq j$. Compare Equations 5.1 and 5.3 to the underlying spectral decomposition, $\mathbf{\Sigma}_Y = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}}$. Equation 5.1 drops the $\mathbf{V}^{\mathrm{T}}$ of spectral decomposition and is referred to as dimension reduction sphereing (DRS). Conversely, Equation 5.1 is referred to as spectral decomposition sphereing (SDS).

5.1.2 Projection Pursuit

The projection pursuit algorithm may now proceed on the pre-processed data, $\mathbf{X}$. The algorithm is based on the projection index, which is a statistic that tests for non-Gaussianity. An optimized search finds the projection of $\mathbf{X}$ that maximizes the projection index, meaning that it finds the most non-Gaussian projection. The Gaussianization transform of the multivariate data makes that projection Gaussian. Iterating this search and Gaussianization procedure, $\mathbf{X}$ is transformed to be multiGaussian.

Projection Index

Consider a Kx1 unit length vector, $\boldsymbol{\theta}$, and the associated projection of the data upon it, $\mathbf{p} = \mathbf{X}\boldsymbol{\theta}$. As discussed, any $\boldsymbol{\theta}$ should yield a $\mathbf{p}$ that is univariate Gaussian if $\mathbf{X}$ is multiGaussian. With this in mind, define the projection index, $I(\boldsymbol{\theta})$, as a test statistic that measures univariate non-Gaussianity. For any $\boldsymbol{\theta}$ where the associated $\mathbf{p}$ is perfectly Gaussian, $I(\boldsymbol{\theta})$ is zero.

A key decision when implementing projection pursuit is the projection index that is used to measure the deviation of each projection from the Gaussian distribution. Friedman's $I(\boldsymbol{\theta})$ was used for the initial implementation of the PPMT, though this does not preclude the use of other indexes. It is calculated as:

$$I(\boldsymbol{\theta}) = \sum_{j=1}^{d} \frac{2j+1}{2} E_r^2\{\psi_j(\mathbf{r})\} \tag{5.4}$$

where $\psi_i(\mathbf{r})$ are Legendre polynomials and d is the number of polynomial expansions. The Legendre polynomials are calculated recursively as:

$$\psi_0(\mathbf{r}) = 1,\ \psi_1(\mathbf{r}) = \mathbf{r},\ \text{and}\ \psi_j(\mathbf{r}) = [(2j-1)\,\mathbf{r}\,\psi_{j-1}(\mathbf{r}) - (j-1)\,\psi_{j-2}(\mathbf{r})]/j \quad \text{for } j \geq 2 \tag{5.5}$$

These polynomials are a function of $\mathbf{r}$, which is a transformed version of the projection, $\mathbf{p}$, according to:

$$\mathbf{r} = 2G(\mathbf{p}) - 1,\ \mathbf{r} \in [-1, 1] \tag{5.6}$$

where G is the standard Gaussian CDF. It is important to note that if $\mathbf{p}$ is standard Gaussian distribution, then $I(\boldsymbol{\theta})$ is zero. Friedman chose this form of $I(\boldsymbol{\theta})$ since it

places greater emphasis on the body of the distribution as opposed to the tails. This reflected his belief that important complex structures such as multi-modality and non-linearity will usually occur near the distribution center. Other test statistics were deemed less suitable since they are more sensitive to the tails of a distribution. A practical advantage of this projection index is that its derivative may be calculated, allowing for an gradient based optimization to be used when searching for the $\boldsymbol{\theta}$ that maximizes $I(\boldsymbol{\theta})$. While the equations that are required for the numerical calculation of $I(\boldsymbol{\theta})$ are provided above, interested readers are referred to Friedman (1987) for its full conceptual and theoretical development.

Optimized Projection Search

An optimized search is used to find the $\boldsymbol{\theta}$ that maximizes $I(\boldsymbol{\theta})$. It begins with a coarse grid search along combinations of the principal component axes to minimize the potential of the subsequent gradient based optimization becoming trapped in a local maxima. Once a maximum $I(\boldsymbol{\theta})$ is determined along one of these major directions, $\boldsymbol{\theta}$ is then fine-tuned using steepest ascent optimization. This requires the derivative of Equation 5.4, which is given by Equation 5.7 under the constraint $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta} = 1$.

$$\frac{\partial I(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{d} \mathbf{r} E_r^2\{\psi_j(\mathbf{r})\}\psi_j'(\mathbf{r}) e^{-\mathbf{p}^2/2} (\mathbf{X} - \boldsymbol{\theta}\mathbf{p}) \tag{5.7}$$

Here, $\psi_i'(\mathbf{r})$ is the derivative of the Legendre polynomials, which is calculated as:

$$\psi_1'(\mathbf{r}) = 1, \text{ and } \psi_j'(\mathbf{r}) = \mathbf{r}\psi_{j-1}'(\mathbf{r}) + j\psi_{j-1}(\mathbf{r}), \text{ for } j > 1 \tag{5.8}$$

This search strategy was found to be very fast, completing in under a second for all test cases.

Gaussianization

The final step of each iteration is to transform $\mathbf{X}$ so that its projection $\mathbf{p}$ along $\boldsymbol{\theta}$ is Gaussianized:

$$\tilde{\mathbf{p}} = G^{-1}(F(\mathbf{p})) \tag{5.9}$$

Note that Equation 5.9 amounts to the normal score transformation of the projection, $\mathbf{p}$. The objective, however, is to transform $\mathbf{X}$ so that its projection along $\boldsymbol{\theta}$

is $\tilde{\mathbf{p}}$. Additional steps will therefore be required to achieve the goal that Equation 5.9 represents. Begin by calculating the orthonormal matrix:

$$\mathbf{U} = [\boldsymbol{\theta}, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_{K-1}] \tag{5.10}$$

where the Kx1 unit vectors, $\boldsymbol{\phi}_i$, are calculated using the Gram-Schmidt algorithm (Reed and Simon, 1972). The linear combination of $\mathbf{X}$ and $\mathbf{U}$, results in a transformation where the first column is the projection, $\mathbf{p} = \mathbf{X}\boldsymbol{\theta}$:

$$\mathbf{XU} = [\mathbf{p}, \mathbf{X}\boldsymbol{\phi}_1, \mathbf{X}\boldsymbol{\phi}_2, ..., \mathbf{X}\boldsymbol{\phi}_{K-1}] \tag{5.11}$$

Next, let Θ be a transformation that yields a standard Gaussian projection, $\tilde{\mathbf{p}}$, while leaving the remaining orthogonal directions intact:

$$\Theta(\mathbf{XU}) = [\tilde{\mathbf{p}}, \mathbf{X}\boldsymbol{\phi}_1, \mathbf{X}\boldsymbol{\phi}_2, ..., \mathbf{X}\boldsymbol{\phi}_{K-1}] \tag{5.12}$$

To be clear, Θ amounts to the normal score transformation of the first column of $\mathbf{XU}$. Multiplying this result by $\mathbf{U}^{\mathrm{T}}$ returns $\Theta(\mathbf{XU})$ to the original basis:

$$\tilde{\mathbf{X}} = \Theta(\mathbf{XU})\mathbf{U}^{\mathrm{T}} \tag{5.13}$$

The transformed multivariate data, $\tilde{\mathbf{X}}$, will now yield a Gaussian projection along $\boldsymbol{\theta}$ and therefore have a projection index of $I(\boldsymbol{\theta}) = 0$. The optimized search for the maximum projection index may be repeated on $\tilde{\mathbf{X}}$ to find other complex directions. The multivariate data will eventually approach a standard uncorrelated multiGaussian distribution with iterative application of the search and Gaussianize steps.

Stopping Criteria

Choosing the target value to which the Gaussian test statistic, $I(\boldsymbol{\theta})$, must descend is not straightforward. Indeed, no stopping criteria guidelines were found in the reviewed PPDE literature. Additional variables, K, make the discovery and resolution of complexity in the data more difficult. A smaller number of n observations make the projections less reliable for detecting meaningful multivariate structure. These characteristics are also observed in random samples from a multivariate Gaussian distribution; reducing n and increasing K creates an increasingly non-Gaussian random sample. This was demonstrated using MCS and LHSMDU in Figure 4.4.

Drawing on this parallel, the target test statistic for PPMT stopping is determined by random samples from a multiGaussian distribution. A bootstrapping algorithm is implemented, where MCS is used to randomly sample M distributions of matching K and n from the Gaussian CDF (Equation 4.4). A projection index value, $I(\boldsymbol{\theta})$, is then calculated for all M distributions along K random orthogonal unit vectors. This process yields an MxK distribution of projection indices that is referred to as $\mathbf{I}$; this distribution of random Gaussian projection indices provides a basis for the convergence criterion. A targeted random Gaussian projection index percentile may be specified when executing the PPMT. If this target percentile is not reached after a maximum number of iterations, the achieved percentile is reported.

For example, a user may target the 50^{th} percentile of the $\mathbf{I}$ distribution. The projection pursuit algorithm will iterate until the $I(\boldsymbol{\theta})$ of the transformed data is less than the 50^{th} percentile of $\mathbf{I}$. If this percentile be achieved, then the transformed data is more Gaussian than 50% of the random Gaussian samples.

5.2 Back-Transformation

Two very different options are available for the PPMT back-transformation; they are referred to as GM and reverse projection (RP). The following section presents both back-transforms, before demonstrating and comparing them in the next section.

5.2.1 Gaussian Mapping

As discussed, the PPMT was originally conceived for the GM framework that was introduced with the MSNT in Section 4.2.1. Within this context, the original multivariate data, $\mathbf{Z}$, are recorded with the final transformed data, $\tilde{\mathbf{X}}$, to establish the mapping. The intermediate steps of the PPMT are not required (e.g., normal score, sphereing and projection pursuit); only where each observation starts in original units and finishes in Gaussian units.

Following independent geostatistical simulation in Gaussian space, realizations are back-transformed according to Equation 4.6. Although the PPMT is orders of magnitude faster than the MSNT, it yields similar quality of results for datasets of relatively few observations (say, $n < 500$), in terms of minimizing changes to the original multivariate configuration. For larger datasets, however, the PPMT yields significantly better results, as the MSNT optimization problem becomes too difficult with increasing n. As displayed in Figure 4.6, the MSNT takes an impractical length

of time to converge for $n > 500$.

Beyond the issue of convergence time, the MSNT was found to change the multivariate configuration of the data more than the PPMT for larger test cases. Consequently, the PPMT mapping yields better back-transformation results than the MSNT mapping in terms of reproducing the variability and spatial continuity of the original data. It is for these reasons that the PPMT is recommended for the GM framework over the MSNT.

5.2.2 Reverse Projection

A second option for the PPMT back-transformation is simply to reverse the forward transformation steps. This approach requires that information from each step is recorded so that they may be reversed:

i) Normal score transform: record the original data, $\mathbf{Z}$.

ii) Data sphereing: record the sphereing matrix, $\mathbf{S}^{-1/2}$.

iii) Projection pursuit: record the orthogonal basis, $\mathbf{U}$, and the original projection, $\mathbf{p}$, for each iteration.

Let the N simulated nodes of K independently simulated Gaussian variables be given as the 1 x K vectors, $\tilde{\mathbf{x}}_\alpha, \alpha = 1, ..., N$. The back-transformation begins by applying Equations 5.14 to 5.16 for each projection pursuit iteration (in the reverse of the forward transform order). First, multiply the Gaussian nodes with the recorded orthogonal basis, $\mathbf{U}$:

$$\tilde{\mathbf{x}}_\alpha \mathbf{U} = [\tilde{p}, \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_1, \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_2, ..., \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_{K-1}], \text{ for } \alpha = 1, ..., N \tag{5.14}$$

The first entry of $\tilde{\mathbf{x}}_\alpha \mathbf{U}$ is $\tilde{p} = \tilde{\mathbf{x}}_\alpha \boldsymbol{\theta}$, where $\tilde{p}$ is assumed to lie within the Gaussianized projection of the transformed data, $\tilde{\mathbf{p}} = \tilde{\mathbf{X}}\mathbf{U}$. Next, use the recorded original projection of the data, $\mathbf{p}$, to reconstruct its empirical CDF, $F(\mathbf{p})$. The Gaussianization transformation of Equation 5.12 may then be inverted:

$$\Theta^{-1}(\tilde{\mathbf{x}}_\alpha \mathbf{U}) = [p, \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_1, \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_2, ..., \tilde{\mathbf{x}}_\alpha \boldsymbol{\phi}_{K-1}], \text{ for } \alpha = 1, ..., N \tag{5.15}$$

where Θ^{-1} normal score back-transforms the first entry of $\tilde{\mathbf{x}}_\alpha \mathbf{U}$ as $p = F^{-1}(G(\tilde{p}))$, while leaving the remaining entries unaltered. Now possessing the back-transformed projection value, the simulated nodes are returned to the original basis:

$$\mathbf{x}_\alpha = \Theta^{-1}(\tilde{\mathbf{x}}_\alpha \mathbf{U})\mathbf{U}^{\mathrm{T}}, \text{ for } \alpha = 1, ..., N \tag{5.16}$$

Repeating the above steps for each projection pursuit iteration back-transforms the simulated nodes to the data sphereing space. The sphereing is then inverted using the recorded matrix, $\mathbf{S}^{-1/2}$, to return the simulated nodes to normal score space:

$$\mathbf{y}_\alpha = \mathbf{x}_\alpha \mathbf{S}^{1/2}, \text{ for } \alpha = 1, ..., N \tag{5.17}$$

Finally, the normal score back-transformation is applied to return the simulated nodes to original space:

$$z_{\alpha i} = F_i^{-1}(G(y_{\alpha i})), \text{ for } \alpha = 1, ..., N \text{ and } i = 1, ..., K \tag{5.18}$$

where the empirical CDFs, $F_i(z_i)$ for $i = 1, ...K$, are constructed using the recorded original data, $\mathbf{Z}$.

5.3 Demonstration

Synthetic data is used for demonstrating the PPMT and its related modeling workflow. While this data exhibits bivariate and spatial complexities, it is relatively simple in terms of its spatial configuration (2-D) and the number of variables (two). This allows for a clearer demonstration of the PPMT and additional understanding of its results. In a complimentary manner, the nickel laterite case study in Chapter 6 applies the PPMT to four real and complex geological variables in a 3-D setting.

After performing an inventory of the data, the PPMT is executed to demonstrate and analyze each transformation step from Section 5.1. Properties of the transformed data are then studied in detail, before applying and comparing the back-transformations from Section 5.2. Alternative multivariate transformation and simulation approaches are then used to provide a benchmark for judging the PPMT results.

Please note that most of the figure and table formats in this section (and underlying statistics) are consistent with that of Section 3.4. Readers are referred to Section 3.4 for their explanation, as only formats and statistics that are introduced in this section are explained here.

5.3.1 Data Inventory

A map view of the data values appears in Figure 5.1, where 841 observations are spaced at fifty meter intervals across a regular sampling grid. Univariate and bivariate statistics of the Z_1 and Z_2 variables appear in Figure 5.2, where heteroscedastic and non-linear features are apparent in the KDE scatterplot. All of the properties appearing in Figure 5.2 are considered representative of this domain. Following geostatistical modeling, the simulated realizations will therefore be checked for the reproduction these properties.

Omnidirectional semivariograms and cross-semivariograms are shown in Figure 5.3, where Z_1 is seen to be more far more continuous than Z_2. To simplify interpretation, the semivariograms and cross-semivariogram are standardized by the appropriate term of $\Sigma(h)$ so that $\rho_{ij}(h) = 0$ when $\gamma_{ij}(h) = 1$. Note that the cross-semivariograms are standardized by the As noted, issues are more likely to arise when applying multivariate transformations that mix variables of differing spatial structure. The concern is that the unique spatial character of each variable cannot be recovered following simulation and back-transformation. The goal is to reproduce the spatial continuity that appears in Figure 5.3, which is carefully checked in later sections.

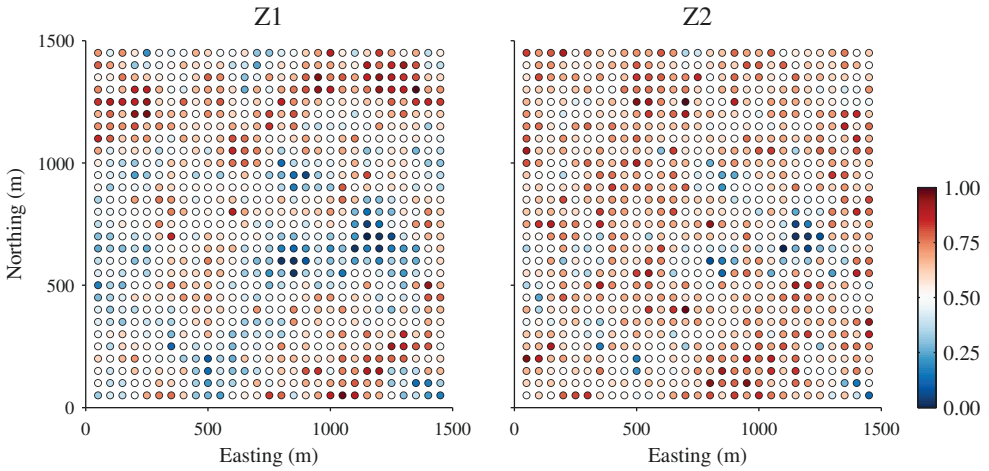


Figure 5.1: Mapview of the 2-D locations and values that are used for demonstrating the PPMT.

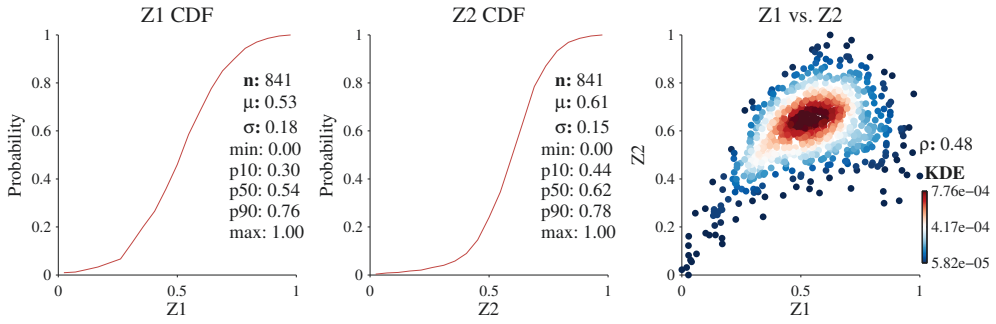


Figure 5.2: CDFs and KDE scatterplot of the variables.

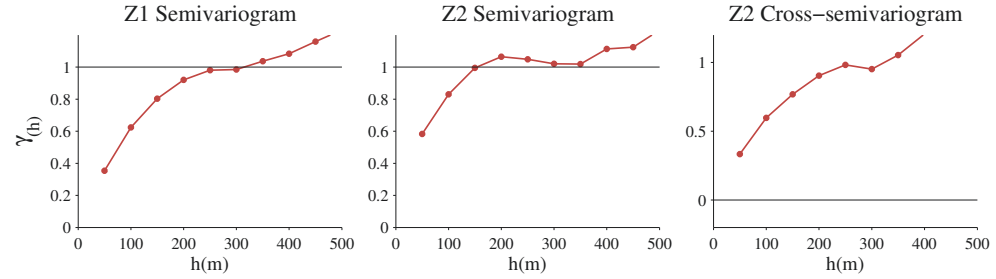


Figure 5.3: Semivariograms and cross-semivariogram of the variables.

5.3.2 Forward Transformation

All of the forward transformation steps from Section 5.1 will now be demonstrated, beginning with the normal score transformation of each variable in Figure 5.4. Note that transformed variables are labeled as Y_1 and Y_2 within subsequent figures (regardless of the specific transformation), whereas Z_1 and Z_2 are reserved for labeling variables in original units. Observe from the CDFs in Figure 5.4 that Y_1 and Y_2 are now standard univariate Gaussian. Unfortunately, the transformed variables continue to exhibit obvious bivariate complexity in the KDE scatterplot. The displayed features and density do not follow the elliptical contours of a Gaussian distribution and would not be characterized by the correlation statistic. Additional measures are required to make these variables bivariate Gaussian, meaning that it is appropriate to consider the PPMT.

Data sphereing is applied next, where Y_1 and Y_2 are rotated using Equation 5.3 to have an identity covariance matrix (Figure 5.5). The variables are now completely uncorrelated with a variance of one. Despite these properties, however, bivariate complexity continues to be evident in the KDE scatterplot.

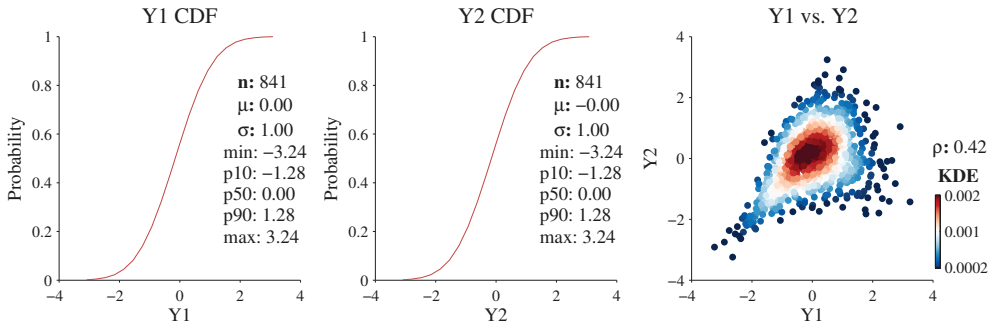


Figure 5.4: CDFs and KDE scatterplot of the normal score transformed variables.

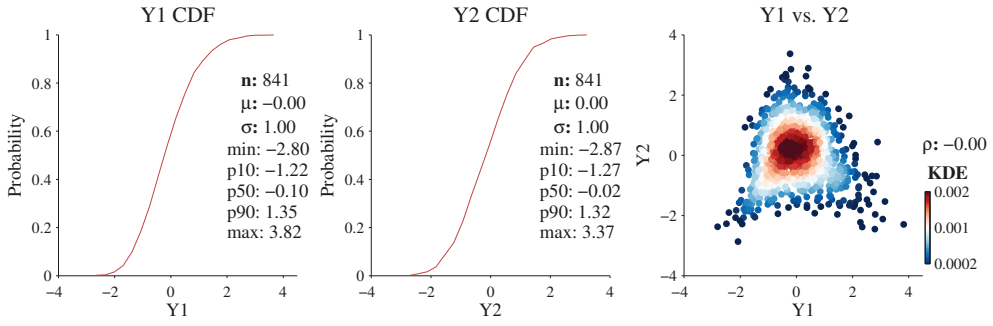


Figure 5.5: CDFs and KDE scatterplot of the sphere variables.

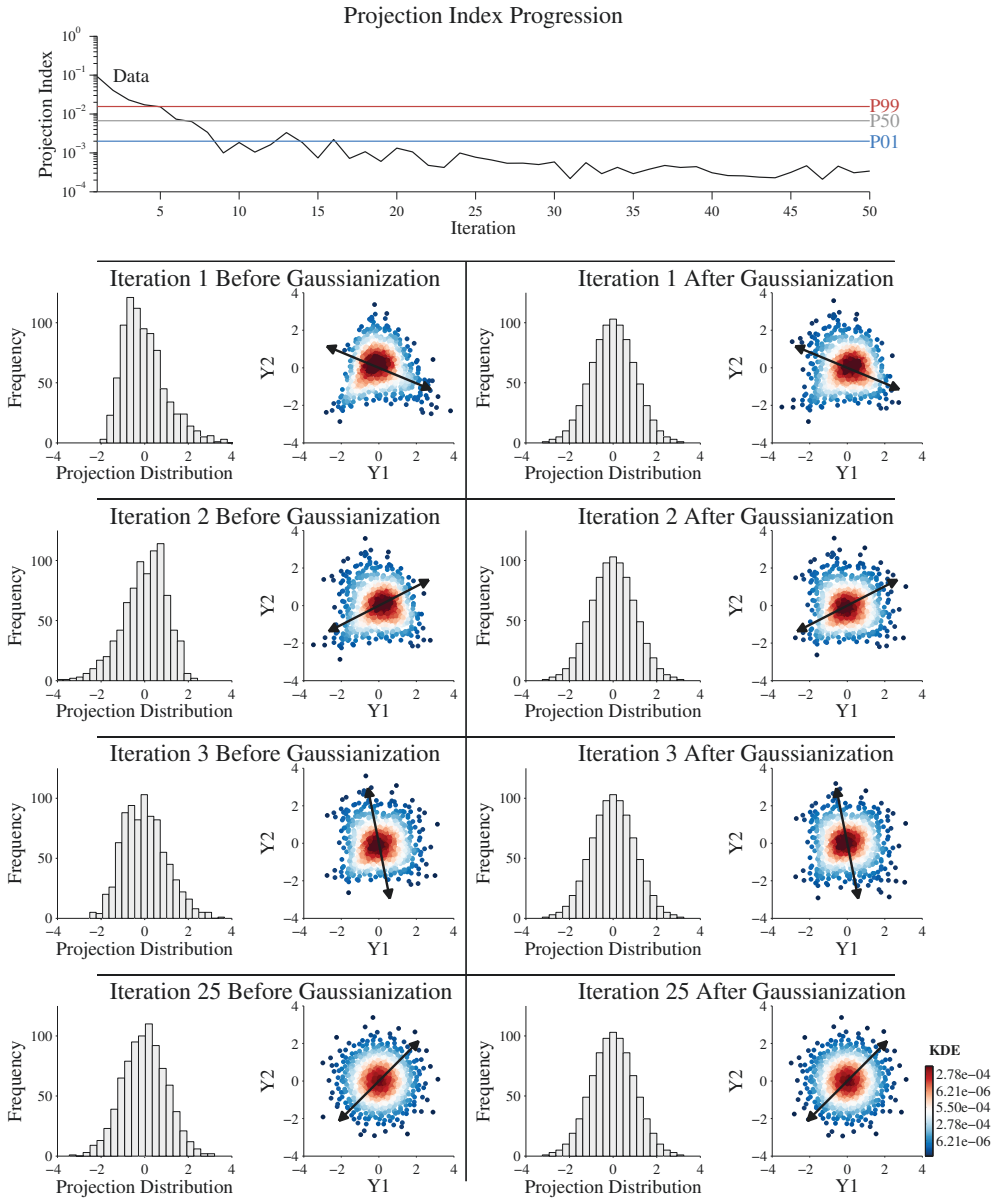


Figure 5.6: Progression of the data through the projection pursuit algorithm.

The projection pursuit algorithm from Section 5.1.2 is demonstrated and summarized in Figure 5.6. The arrows within this figure display the orientation of $\boldsymbol{\theta}$ that was found to maximize $I(\boldsymbol{\theta})$ for each iteration. The left side of the figure displays properties before Gaussianization, including KDE scatterplots of the bivariate data, $\mathbf{X}$, and histograms of the non-Gaussian projection, $\mathbf{p} = \mathbf{X}\boldsymbol{\theta}$. Conversely, the right side of the figure displays properties after Gaussianization, including KDE scatterplots of the bivariate data, $\tilde{\mathbf{X}}$, and histograms of the standard Gaussian projection, $\tilde{\mathbf{p}} = \tilde{\mathbf{X}}\boldsymbol{\theta}$.

Observe that each iteration shifts the bivariate data to make the associated projection univariate Gaussian. Not surprisingly, projections that maximize $I(\boldsymbol{\theta})$ in iterations 1 to 3 are far less Gaussian than that of iteration 25. The KDE scatterplots show the visual progression towards bivariate Gaussianity of the four displayed iterations, while the plot at the top of the figure shows the value of the projection index, $I(\boldsymbol{\theta})$, for the fifty executed iterations. Note the logarithmic y-axis of this plot, meaning that the vast majority of complexity is addressed in the first few iterations.

The 1^{st}, 50^{th} and 99^{th} percentiles of the $\mathbf{I}$ distribution are also displayed in this plot for reference. Recall that $\mathbf{I}$ may be used as stopping criteria for the PPMT, where $I(\boldsymbol{\theta})$ is calculated for M random nxK Gaussian distributions along K random $\boldsymbol{\theta}$ orientations. The 99^{th} percentile indicates that $I(\boldsymbol{\theta})$ is a projection from a distribution that is barely Gaussian, whereas the 1^{st} percentile indicates that $I(\boldsymbol{\theta})$ is a projection from a distribution that is very Gaussian. In this case, the PPMT yields a very Gaussian result after approximately ten iterations, though additional Gaussianity is achieved by additional iterations.

5.3.3 Transformed Properties

Properties of the transformed data following fifty projection pursuit iterations are analyzed in greater detail within Figure 5.7. As expected, the KDE scatterplot appears to mimic the circular density contours of an uncorrelated bivariate Gaussian distribution. Also note that the variables are uncorrelated to the second decimal, with CDF properties that match that of a univariate Gaussian distribution.

The KDE density of Figure 5.7 and projection index of Figure 5.6 suggest a high degree of bivariate Gaussianity in the transformed data. Nevertheless, the bivariate standard normal (BVSN) Gaussianity test of the `scatnscores` program

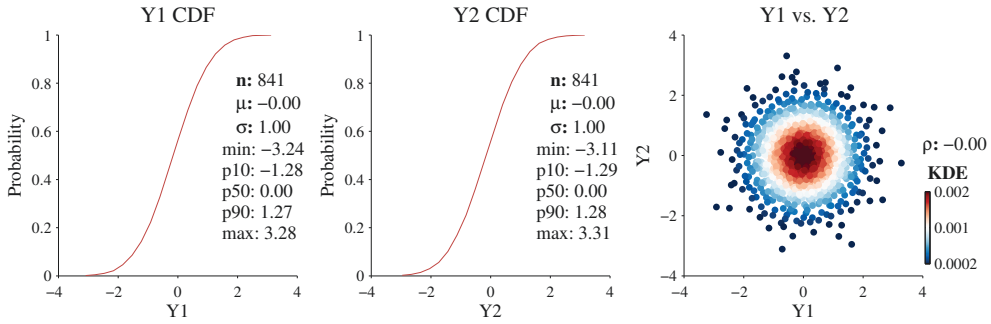


Figure 5.7: CDFs and KDE scatterplot of the PPMT transformed variables.

(Deutsch and Deutsch, 2011) is used as an independent check of this property. Figure 5.8 displays the `scatnscores` plot and associated BVSN test following the normal score transformation, sphere transformation, five iterations of projection pursuit, and fifty iterations of projection pursuit. Refer to the description of Figure 4.4 for an explanation of the `scatnscores` plot format. Red text in the normal score and sphere plots indicate that they have failed the BVSN test.

While the absence of red text in the projection pursuit plots indicate that they have both passed the BVSN test, the fifty iterations result is visually closer to the typical 'Gaussian cloud'. As discussed in Section 4.2.1, the GM back-transformation will benefit from transformed data that is as close as possible to the uncorrelated standard multiGaussian model. The GM workflow assumes that subsequent independent Gaussian simulation will yield geostatistical realizations that closely approximate an uncorrelated multiGaussian model. If deviations exist between the density of simulated realizations and mapped data in transformed space, it is unlikely that the original properties will be reproduced following back-transformation.

Similarly, the RP back-transformation scheme from Section 5.2.2 will generally benefit from transformed data that closely approximate the multiGaussian model. This increases the likelihood that the original data will be representative of the simulated realizations at every step of the RP, leading to simulated realizations that reproduce properties of the original data, $\mathbf{Z}$.

Whereas the sphere data is entirely uncorrelated prior to projection pursuit, minor correlation may be introduced following the Gaussianization transform of each iteration. Correlation of the two variables is plotted for each iteration in Figure 5.9. While the extremely small scale of the y-axis should be noted in this figure (-0.01 to 0.01 correlation), the variables are seen to have less absolute correlation

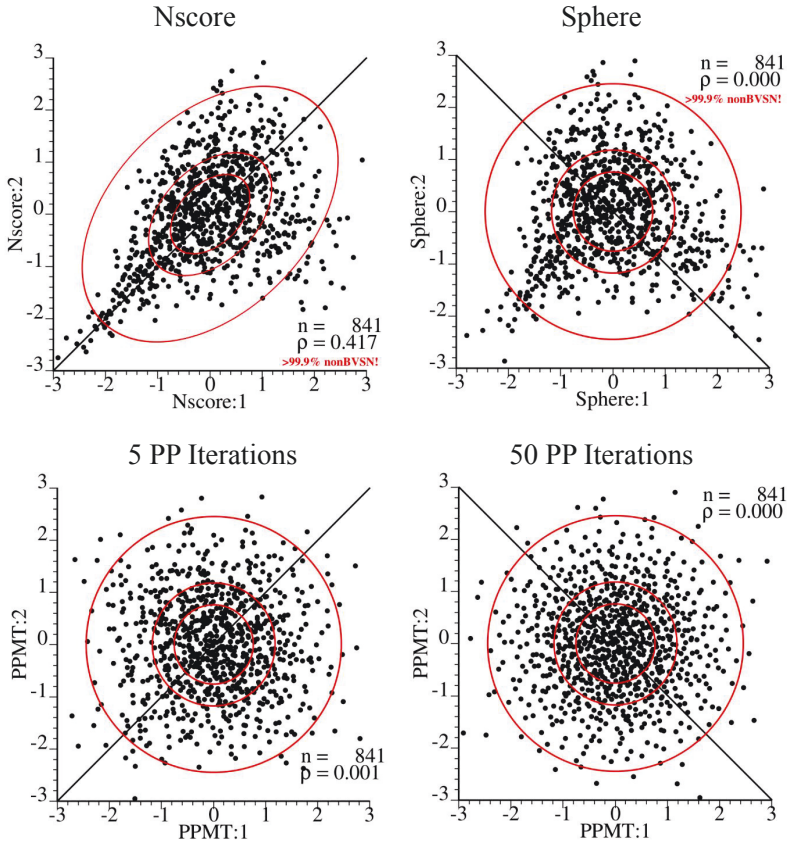


Figure 5.8: `scatnscores` plots and Gaussianity tests following various steps of the PPMT. Red text (>99.9% nonBVSN!) indicates that the test has failed for that plot.

with increasing iterations. Correlation is generally more stable beyond iteration 25, where it has an absolute value of less than 0.001 (with the exception of iterations 42 to 44).

The Gaussianity and correlation results in Figures 5.6 to 5.9 cumulatively establish that there is merit to using a large number of projection pursuit iterations. The question arises, however, as to whether these additional iterations have a negative impact on other properties of the transformed data? Perhaps additional iterations adds significantly to the variable mixing, which should be minimized given the very different spatial continuity of Z_1 and Z_2.

As discussed in Section 5.1.1, the implemented data sphereing algorithm was chosen since it minimizes variable mixing by maximizing the absolute value of the loadings, $\rho'(Y_i, X_j)$ for $i = j$. Consider using the correlation between the original

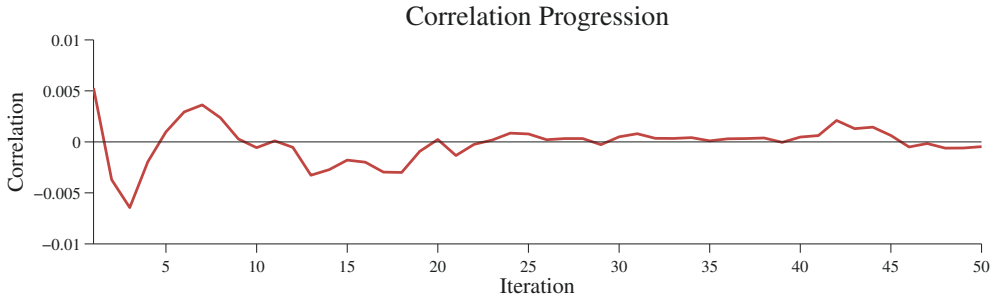


Figure 5.9: Progression of correlation between the two variables across the projection pursuit iterations.

and transformed variables as a metric of mixing, which is simply scaled loadings according to Equation 5.2. Figure 5.10 displays these correlations (they will be referred to as loadings) following data sphereing and fifty projection pursuit iterations. The bars in these plots are scaled according to the absolute value of the displayed loading, while being colored on a gradient of negative loading (blue), zero loading (white) and positive loading (red).

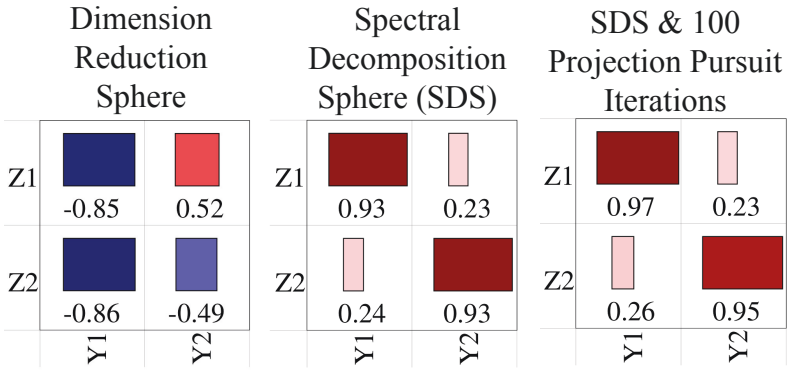


Figure 5.10: Correlation between the original and transformed variables following DRS (left), SDS (middle), and fifty projection pursuit iterations preceded by SDS (right).

Although DRS sphereing (Equation 5.1) is not used by the PPMT, its resultant loadings are shown as a comparative reference in Figure 5.10. As discussed, DRS is similar to PCA in that it maximizes the variability that is explained by the first transformed variable. While the first transformed variable has the largest loadings (as expected), the original variables are heavily mixed across both of the transformed variables. Consequently, the unique continuity of Z_1 and Z_2 is mixed by this transformation and may not be recovered following simulation and back-transformation.

Conversely, the SDS sphereing that is implemented for the PPMT (Equation 5.3) loads the original variables almost entirely on the their transformed equivalent. Though some significant absolute values of $\rho'(Y_i, X_j)$ are seen for $i \neq j$, variable mixing is minimized relative to DRS. Applying fifty iterations of projection pursuit to the SDS data yields the third set of loadings in this plot. Observe that the loadings are very similar following projection pursuit, suggesting that there is little consequence to applying a large number of iterations in terms of additional variable mixing.

A more visual method for inspecting the nature of transformations and their resultant variable mixing is to color the transformed observations according to the original values. Doing so provides insight into the relative shift of the multivariate configuration, as was schematically illustrated in Figure 4.5. The sphere and PPMT transformed data is colored by the original Z_1 and Z_2 values in Figure 5.11.

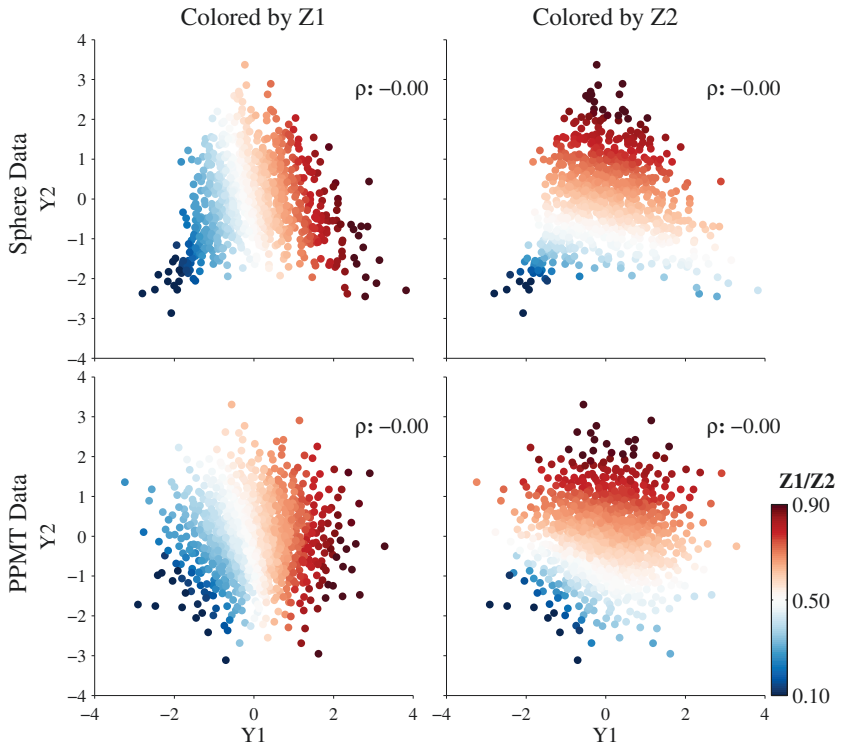


Figure 5.11: Scatterplots of the sphere and PPMT transformed variables, colored by the original values.

As expected, coloring of the sphere data shows that it has been rotated very slightly from the original basis. The PPMT transformed data is colored very similar,

though non-linear changes are seen when compared to the sphere data. This result supports the interpretations that were drawn from the loading plots, suggesting that the PPMT makes relatively 'gentle' changes when transforming the sphere data to be bivariate Gaussian.

Given that the PPMT has transformed the data to be bivariate Gaussian in a manner that minimizes variable mixing, it is expected that the unique spatial continuity of Z_1 and Z_2 will be similar in transformed space. Semivariograms of the original data, normal score data, sphere data, and PPMT transformed data confirms this expectation in Figure 5.12. The cross-semivariograms display that removing correlation at $h = 0$ lag distance has largely removed correlation at $h > 0$ lag distances. Some cross-correlation does remain, however, which may be a concern when considering independent Gaussian simulation of the transformed variables.

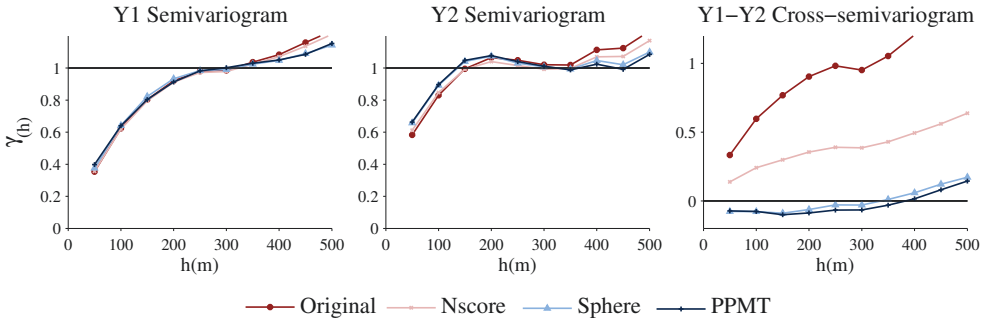


Figure 5.12: Semivariograms and cross-semivariogram of the original and transformed variables.

5.3.4 Back-transformation

With the original data successfully transformed to be uncorrelated and bivariate Gaussian, independent simulation of the variables may now proceed. The simulation will require a model of regionalization so that spatial correlation can be calculated between grid and data locations. To this end, semivariograms of the PPMT transformed data (Figure 5.12) are closely fit using semivariogram models that have zero nugget effect and two spherical nested structures. Using the semivariogram models and transformed data as input, SGSIM is executed to generate one hundred realizations of each variable across a 250x250 node grid. The simulated Gaussian realizations are then returned to original space using both the GM and RP back-transformations.

Maps of three arbitrary realizations are displayed in Figure 5.13, where local conditioning and ergodic fluctuations are evident. The following section will compare the GM and RP back-transformations, though only the RP result is displayed in Figure 5.13 since discrepancies between the techniques are not noteworthy in map view.

Empirical CDFs of the back-transformed realizations are displayed in Figure 5.14, which are summarized with the following statistics:

i) mean μ: the mean of the realization means.

ii) stdev μ: the standard deviation of the realization means.

iii) mean σ: the mean of the realization standard deviations.

iv) stdev σ: the standard deviation of the realization standard deviations.

CDFs of the sampled Z_1 and Z_2 are overlain so that reproduction of the data properties may be evaluated. While the standard deviation of the Z_1 means are higher for RP, both back-transformation methods yield virtually identical CDFs overall. As CDFs of the realizations lie symmetrically about the data CDFs, the PPMT workflow is shown to successfully reproduce the targeted univariate properties.

The KDE scatterplot and correlation of the data is compared with the simulated realizations in Figure 5.15. Differences between the two back-transformations are more evident in this figure. The GM points lie strictly within the concave hull of the data, reflecting the underlying interpolation. Also evident is 'strings' of points along the margins of the distribution, which is another artifact of the interpolation. It occurs when the nearest $K + 1$ mapped observations that are used for the interpolation (Equation 4.6) lie in a relatively sparsely populated region of the original distribution. Though not visually appealing, this artifact it is not thought to be consequential to most applications of geostatistical models (e.g., subsequent transfer functions).

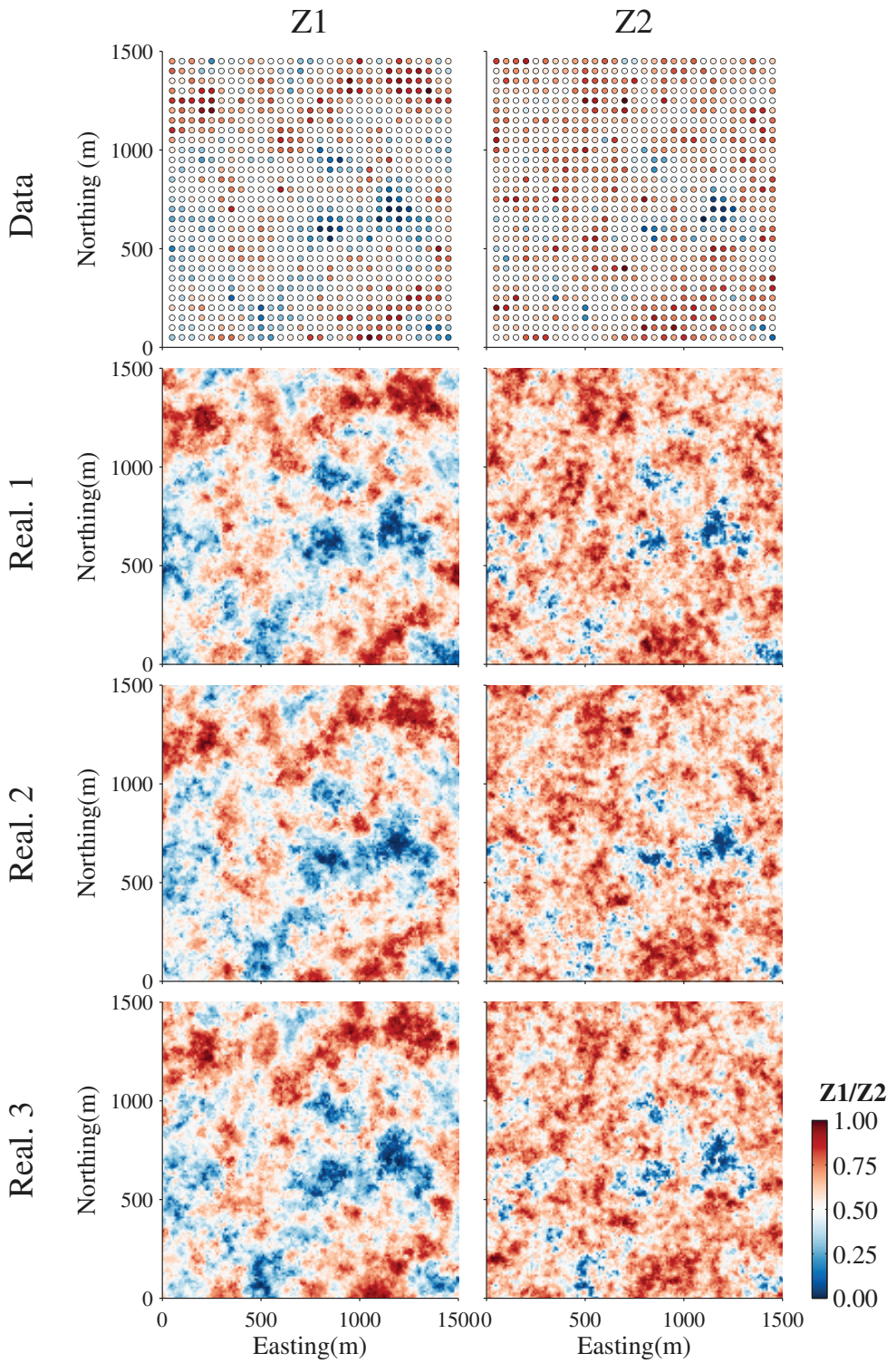


Figure 5.13: Mapview of four arbitrary realizations following the RP back-transform, with the sampled data from Figure 5.1 are shown for comparison.

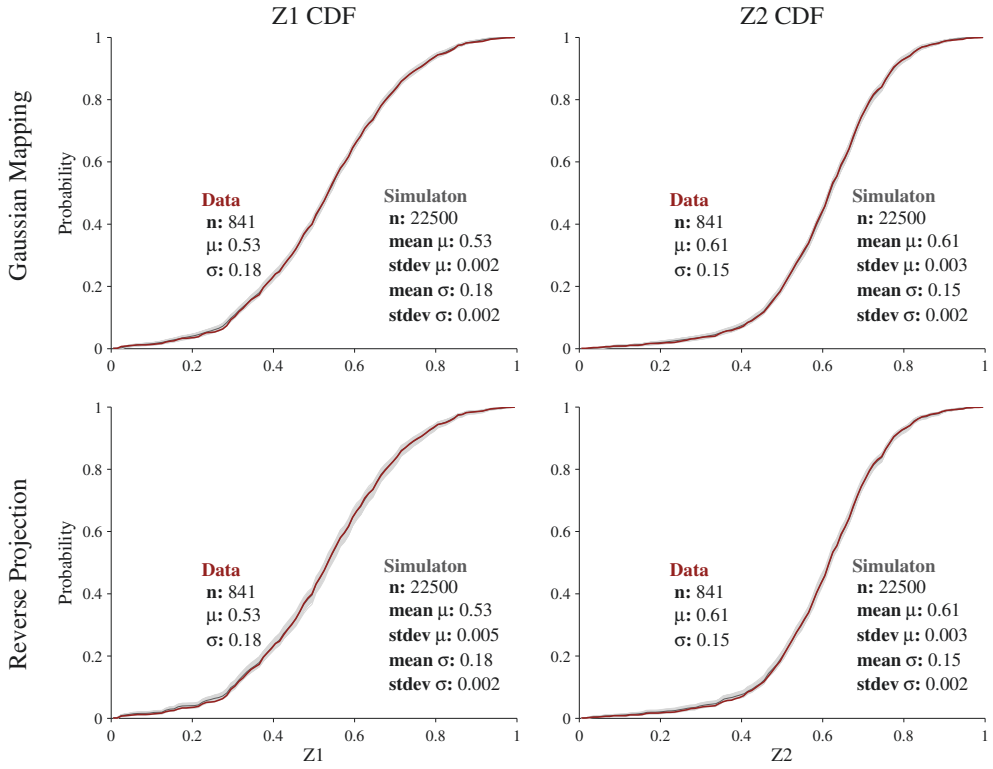


Figure 5.14: CDFs of the simulated realizations following the RP and GM back-transforms. Data CDFs from Figure 5.2 are overlain for comparison.

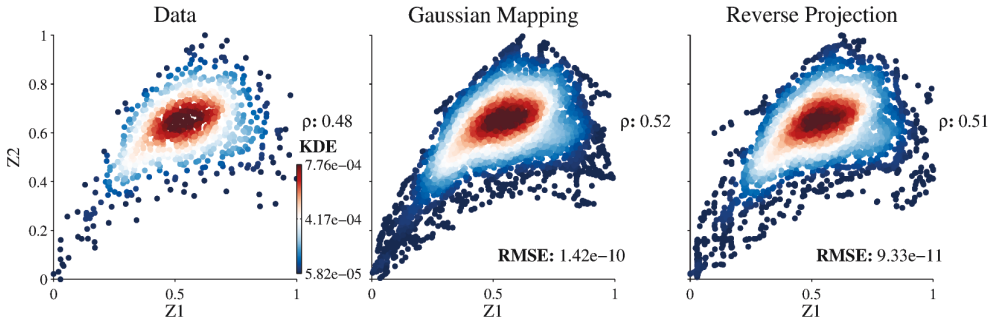


Figure 5.15: KDE scatterplot of simulated realizations following the RP and GM back-transforms. The data scatterplot from Figure 5.2 is shown for comparison.

The RP results do not lie strictly within the concave hull of the data, although the nature of the non-linear boundaries are reproduced. A visual artifact is present along the distribution margins that differs from the described GM artifact. This results from simulated values that fall in the tails of the standard Gaussian projection, beyond the minimum/maximum of the associated data projection. When back-transforming the projection (Equation 5.15), the program constrains these values to the minimum/maximum of the original data projection. As with the visual GM artifact, this is not thought to be consequential for most applications of geostatistical models. One could consider extrapolation schemes, though this may not be straight forward considering the multivariate orientation of the unit vector, $\boldsymbol{\theta}$, that underlies each projection, and the iterative application of Equation 5.15. Initial attempts to implement extrapolation led to simulated points that lie far beyond the multivariate hull of the data in original space.

Aside from the noted discrepancies, Figure 5.15 shows that both back-transforms effectively reproduce the bivariate features of the data. Neither method reproduces the data correlation exactly, although the RP result is slightly better. To summarize the reproduction of bivariate densities, the displayed RMSE statistic is calculated using the difference between the bivariate KDE of the data and that of the realizations. As with correlation, RP is shown to have slightly better reproduction of bivariate density.

Reproduction of the spatial continuity is examined in Figure 5.15. The displayed RMSE statistic in this figure summarizes how well the semivariograms and cross-semivariograms of the data are reproduced by the realizations. It is calculated using the difference between data and the mean of the realizations at the data lag distances.

While it is difficult to visually distinguish the two back-transformation results, the RMSE statistic shows that RP yields better reproduction for both semivariograms, while the GM is a closer match of the cross-semivariogram. The unique continuity of Z_1 and Z_2 has been recovered overall following simulation and back-transformation, although the realizations are too discontinuous at the first lag distance for both variables. More concerning, however, is the overall reproduction of the cross-semivariogram. These concerns are attributed to the cross-correlation that was noted to remain following transformation (Figure 5.12). Independent simulation in the presence of cross-correlation appears to have manifested itself in properties of the back-transformed realizations.

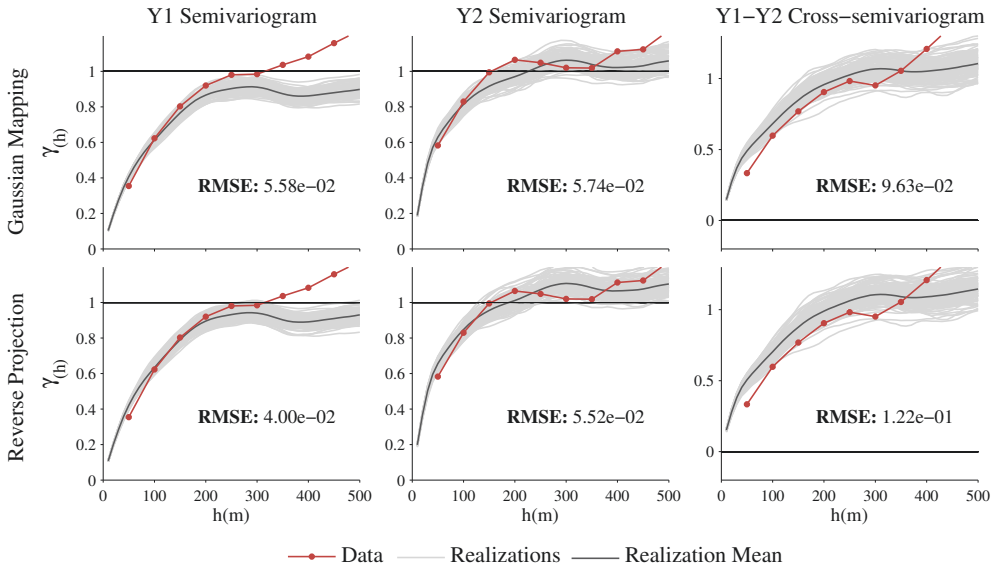


Figure 5.16: Semivariograms and cross-semivariograms of simulated realizations following the RP and GM back-transforms. The data semivariogram and cross-semivariogram from Figure 5.3 overlain for comparison.

5.3.5 Chained MAF Workflow

Given the concern with remnant cross-correlation following the PPMT transformation (Figure 5.12), one may consider the subsequent application of a chained MAF transformation. Doing so will rotate the data to remove correlation at a specified $h > 0$ lag distance, while insuring that the variables remaining uncorrelated at $h = 0$.

Figure 5.17 displays this rotation, where the PPMT and PPMT/MAF (PPMT followed by MAF) transformed data are colored by Z_1 and Z_2. Both the PPMT and PPMT/MAF data are shown to be uncorrelated, though the coloring reveals the orientation of the MAF rotation.

The effect of the MAF rotation is much more obvious when viewing the resultant spatial continuity in Figure 5.17. As can be inferred from this figure, an $h = 150$ lag distance was chosen for the rotation since it displays the largest remnant cross-correlation in the PPMT transformed data. Noting the very small scale of the y-axis, the cross-semivariograms show that decorrelating the variables at $h = 150$ has addressed the majority of cross-correlation at all lags.

The MAF technique orders the rotated variables according to spatial continuity, resulting in large changes to the semivariograms. Note, however, that variable

mixing is not a large concern in this case since the rotation has loaded Z_1 almost entirely on Y_2 (and vice versa). This is evident from the coloring in Figure 5.17 and the semivariograms in Figure 5.19.

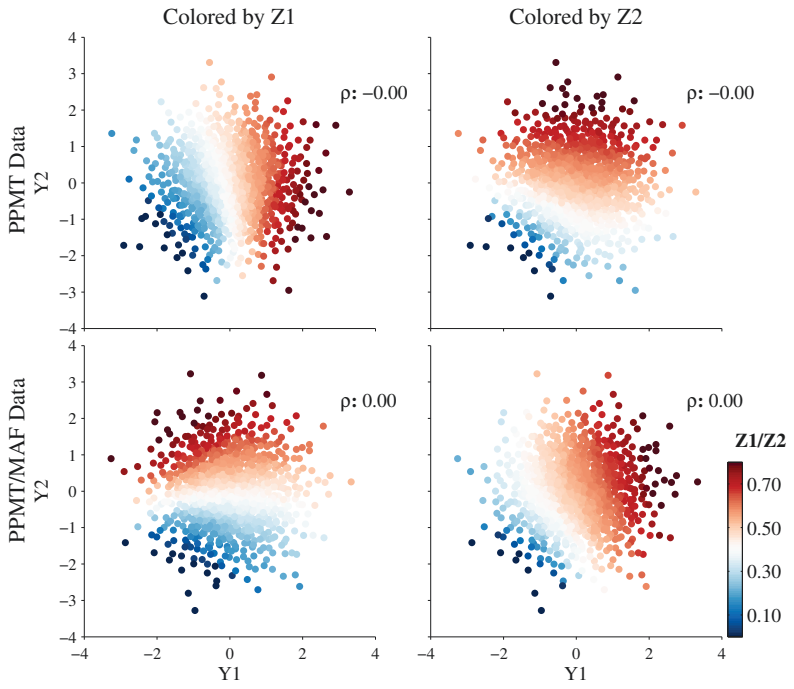


Figure 5.17: Scatterplots of the PPMT and PPMT/MAF transformed variables colored by the original values.

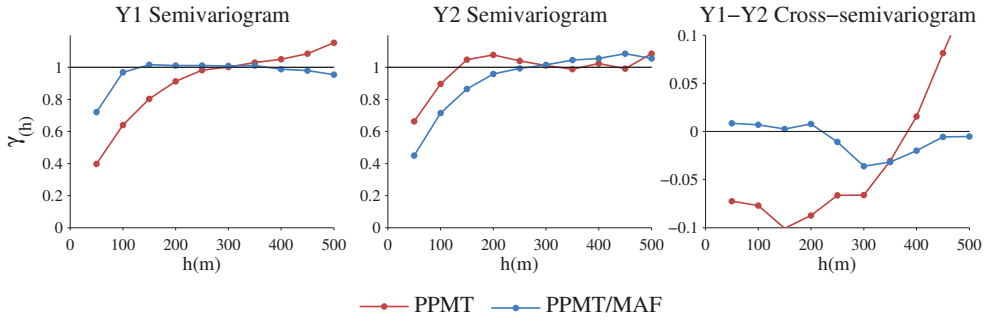


Figure 5.18: Semi-semivariograms and cross-semivariograms of the PPMT and PPMT/MAF transformed variables.

The simulation scheme described in the previous section is applied using the PPMT/MAF transformed data. Simulated realizations are returned to PPMT space using the MAF back-transformation, before using the RP back-transformation to return the realizations to original space. The MAF transformation was found to alter

the multivariate and spatial properties of the simulated realizations. Reproduction of the multivariate density is improved in terms of both correlation and KDE RMSE (Figure 5.19). The scatter is also better constrained to the visual constraints of the data distribution, such as the stray points along the top left margin of the bivariate distribution for the PPMT result.

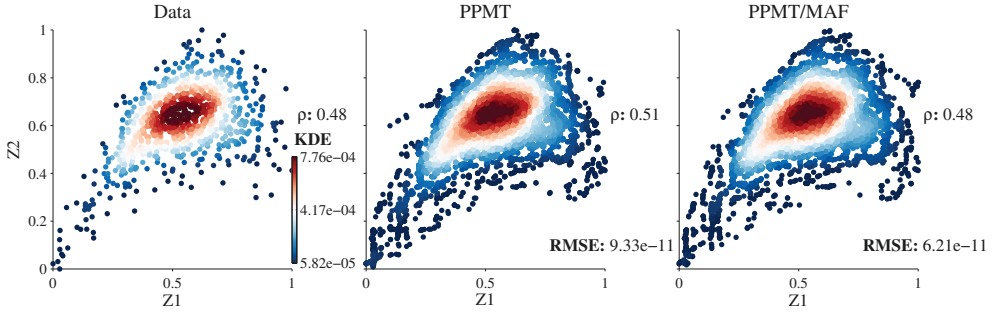


Figure 5.19: KDE scatterplots following the PPMT and PPMT/MAF workflows.

Comparing spatial continuity reproduction in Figure 5.20, the previous concern with the cross-semivariogram has been resolved by the MAF transformation. Perhaps more surprising, however, is that the smaller concern with short scale continuity of the semivariogram has also been improved by the use of MAF.

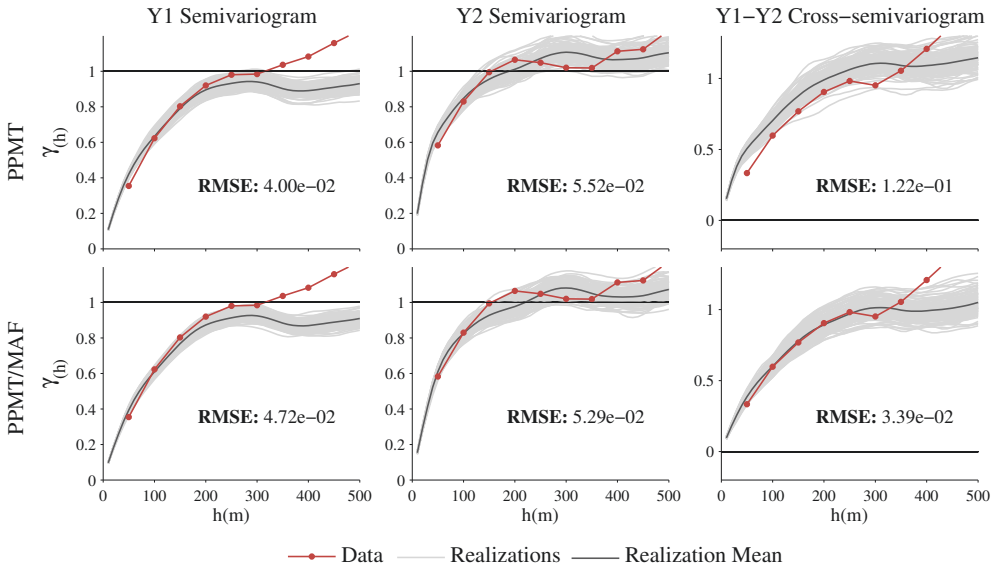


Figure 5.20: Semivariograms and cross-semivariograms of simulated realizations following the PPMT and PPMT/MAF workflows.

5.3.6 Comparative Results

Though the PPMT results appear very promising, it may be difficult to judge their quality without a relative comparison. With this in mind, modeling is repeated using three alternative workflows. These workflows are briefly summarized before comparing key features of the modeling results.

Alternative Approaches

Abbreviated as Nscore/Cosim, the first modeling approach is composed of three primary steps:

i) Normal score transformation of each variable.

ii) Colocated cosimulation using SGSIM with the Markov coregionalization model. More specifically, Z_1 is independently simulated before using its gridded realization as a secondary variable that conditions the cosimulation of Z_2.

iii) Normal score back-transformation of the realizations.

While very practical, this modeling approach usually requires the use of a reduction factor to correct for variance inflation (Babak and Deutsch, 2009a; Deutsch and Journel, 1998). Iterative testing determined that a variance reduction factor of 0.85 corrected for the inflation. Abbreviated as MAF, the second approach is composed of five primary steps:

i) Normal score transformation as a preprocessor that centers the data and removes outliers.

ii) MAF transformation to decorrelate the variables at 0m and 150m lag distances.

iii) Second normal score to transform the variables to be standard univariate Gaussian.

iv) Independent simulation using SGSIM.

v) Required back-transformations of the realizations in the reverse of the forward transform order.

The MAF transformed data is displayed in Figure 5.21, where non-Gaussian bivariate features clearly remain following the rotation. Abbreviated as SCT, the third approach is composed of three primary steps:

i) SCT to transform the variables to an approximately uncorrelated and multi-Gaussian distribution.

ii) Independent simulation using SGSIM.

iii) SCT back-transformation of the realizations.

The SCT transformed data is also displayed in Figure 5.21, where coloring according to Z_1 and Z_2 reveals the mechanics of the transform that were outlined in Section 2.2.2. The visual artifacts that are present along the margins of the distribution is a standard characteristic of the SCT; they result from aligned tail values following the normal score transform of Z_2 in each bin. While not visually appealing, they are not considered to be consequential to the overall multivariate density and SCT back-transformation (Leuangthong and Deutsch, 2003).

All three of the described workflows have enjoyed popular use within geostatistics. While cosimulation and SCT become awkward to apply with an increasing number of variables, this provides an indication of their properties and performance in a bivariate setting where they are better suited.

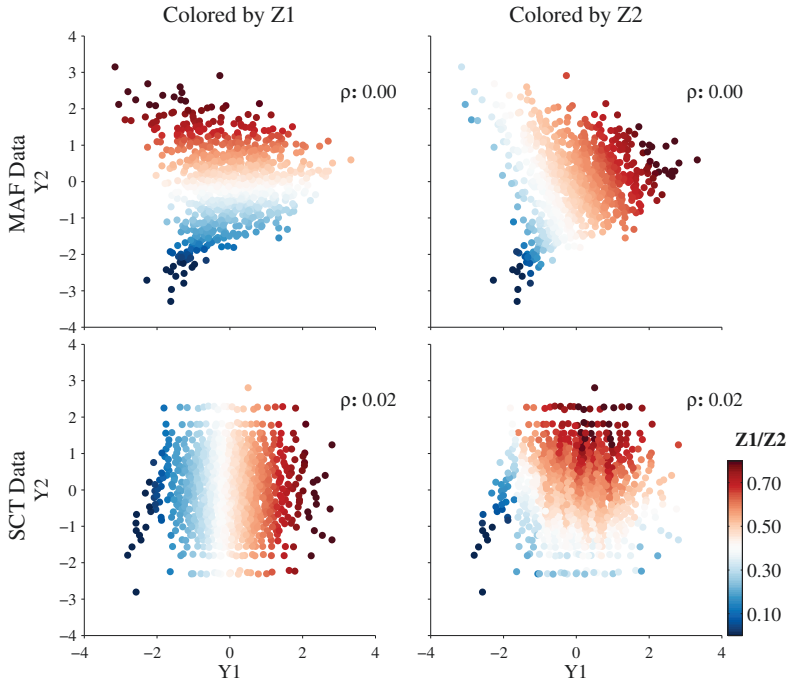


Figure 5.21: Scatterplots of the MAF and SCT transformed variables colored by the original values.

These workflows are compared with the PPMT/MAF workflow from the previous section. All of the underlying modeling details that have been previously applied are held constant so that the outlined features of each workflow are the sole source of discrepancies in their simulation results. This includes semivariogram modeling methodology, SGSIM parameters and grid dimensions.

Results

The bivariate properties of the four workflows are compared in Figure 5.22, where the Nscore/Cosim. and MAF workflows yield the worst results. The complex features of the data are not reproduced since these approaches do not capture and remove them prior to simulation. Consequently, simulated Gaussian realizations do not match the density of the transformed data, leading to notable departures from the density of the original data following back-transformation.

The SCT and PPMT/MAF workflows yield much better results since those complex features are removed prior to simulation and are restored by the back-transformation. Though the binning artifacts of the SCT are distracting, the method achieves good overall reproduction of the bivariate density. Options are available in the latest SCT program (Deutsch, 2005b) for manually setting constraints on the tails of each bin, though this becomes cumbersome with additional variables. The PPMT/MAF workflow yields the best bivariate reproduction according to both visual inspection, KDE RMSE and correlation.

Spatial continuity of each workflow is compared in Figure 5.23. Given the poor reproduction of the $h = 0$ lag correlation in Figure 5.22, it is unsurprising to see that the Nscore/Cosim and MAF workflows yield the worst cross-semivariogram reproduction. PPMT/MAF has the best overall reproduction of the semivariograms and cross-semivariograms according to visual inspection and the RMSE statistic. It is interesting to note, however, that the PPMT/MAF result is the only one that is too discontinuous at the first lag distance for the first semivariogram and the cross-semivariogram (if only slightly).

Summarizing, the PPMT/MAF yields the best overall result for all of the inspected bivariate and spatial properties. Note that univariate properties were not reviewed since all of the workflows performed similar in this regard. Whereas PPMT/-MAF greatly outperformed Nscore/Cosim and MAF, the improvements were more subtle relative to the SCT. It is important to note, however, that the PPMT/MAF

workflow may be applied to increasing K variables without additional practitioner effort. The SCT by comparison, cannot be applied to greater than two to four variables (depending on n observations) without the use of nested workflows. Unlike the PPMT, a nested SCT workflow is unlikely to remove the correlation and complexity between all variables.

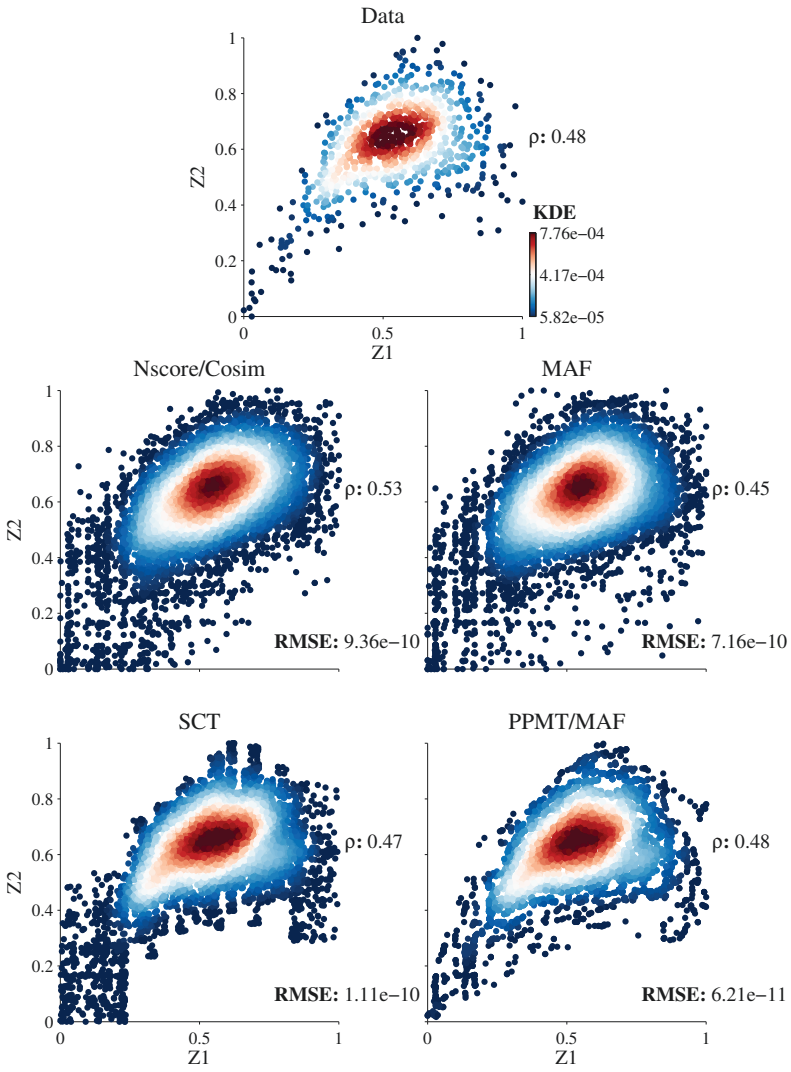


Figure 5.22: KDE scatterplots of the various transformation and simulation workflows.

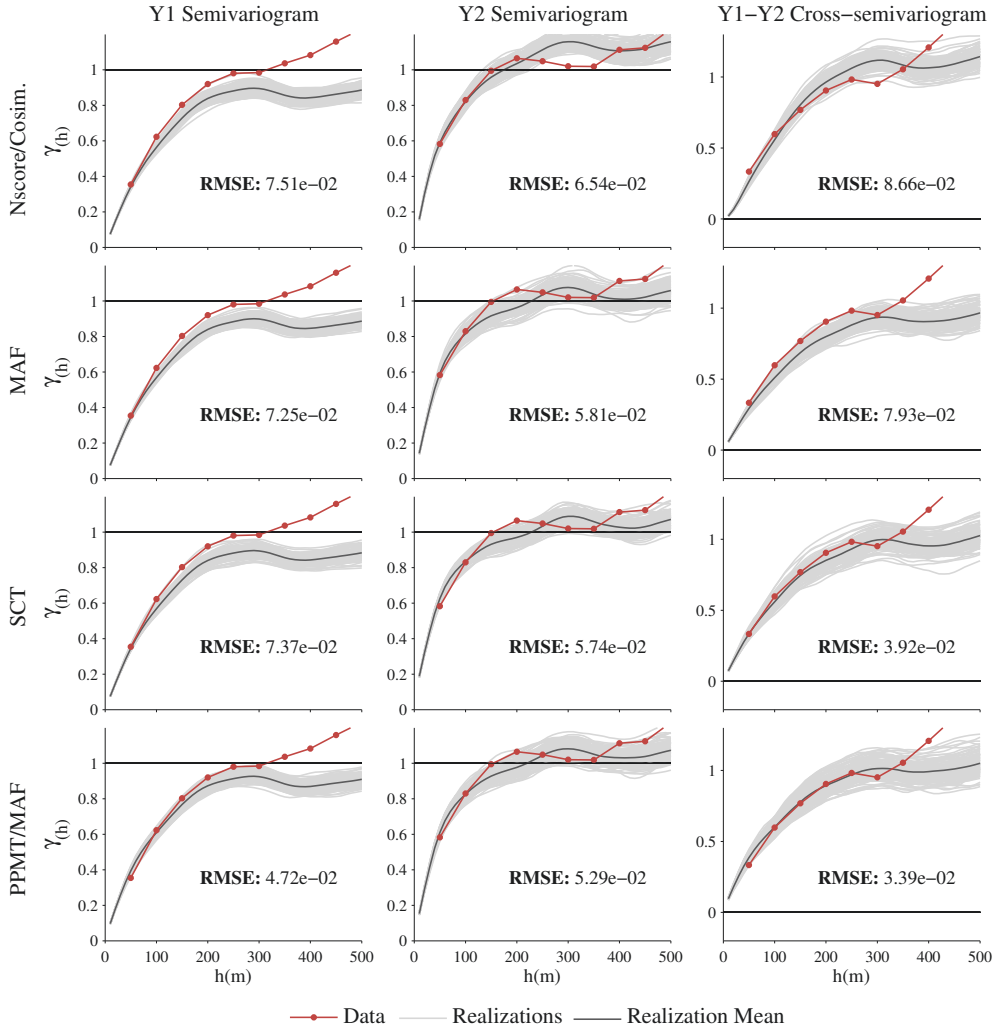


Figure 5.23: Semi-semivariograms and cross-semivariograms of the various transformation and simulation workflows.

5.4 Discussion

The following section will discuss major considerations, limitations, and assumptions of the PPMT.

5.4.1 Back-Transformation Options

As demonstrated in Section 5.3.4, both the GM and RP methods are viable options for the PPMT back-transformation. They generally yield similar results, although the RP method tends to outperform GM in terms of univariate, multivariate and spatial reproduction. Taking a closer look at properties of the GM back-transformation, some other potential disadvantages become apparent:

i) No extrapolation will take place since the interpolation scheme bounds the back-transformed values to the concave hull of the original data.

ii) The intrinsic averaging nature may lead to mean convergence and resultant deviations from the original variability of the data. This is more problematic with increasing K and decreasing n since multivariate space will become poorly informed with sparsely distributed mapped data.

iii) Computational expense, as the execution time increases with greater n since the nearest $K+1$ observations must be found for each simulated node. The distance to each observations is sorted using a 'bubble-sort' (Knuth, 1998), which was found to yield much faster results than other popular sorting algorithms such as Quicksort (Hoare, 1962). Search trees such as a kd-tree (Bentley, 1980) may reduce the execution time further and represent a potential avenue of future work.

Being strictly bounded to the concave hull of the data may actually be a desired feature that encourages the use of GM in some settings. Consider data that have very defined and precise multivariate constraints that are critical to reproduce for subsequent transfer functions. As displayed in Figure 5.15, the RP back-transformation often yields multivariate extrapolation about the concave hull of the data. Consequently, RP is unlikely to reproduce such constraints as precisely as GM. Both back-transformations have been documented since practitioners may prefer either depending on their own set of priorities. The RP method is generally advocated, however, for the listed reasons and demonstrated results.

5.4.2 Standard Gaussian Geostatistical Realizations

Section 5.3.2 stated that it is important for the PPMT to transform the data to be as close to an uncorrelated multiGaussian model as possible. This statement assumes, however, that independent Gaussian simulation will yield an uncorrelated standard multiGaussian distribution. When this is the case, the back-transformed realizations should closely match the univariate and multivariate density of the data.

If the simulated realizations are not uncorrelated standard multiGaussian, however, the back-transformed results may deviate a great deal from the univariate and multivariate properties. For example, consider that long range spatial continuity relative to the domain size commonly leads to geostatistical realizations that have a variance of less than one (in Gaussian space). As the PPMT back-transform is very non-linear, this reduced variance will potentially lead to a large bias of the mean in original space.

Practitioners may consider a histogram correction transformation (Section 2.2.4) in settings where the univariate distributions are not reproduced. Note, however, that histogram corrections may compromise the reproduction of multivariate features. For this reason, the PPMT back-transformation program provides users with the option to correct the univariate distributions prior to back-transforming (transform them to be standard Gaussian). Doing so increases the likelihood of reproducing both the univariate and multivariate features following back-transformation. Note, however, that deviations from a multivariate standard Gaussian distribution may not be corrected by this univariate transformation, allowing issues with univariate reproduction to persist following back-transformation.

5.4.3 Reproduction of Short Scale Continuity

A small but consistent loss of short scale continuity was seen in Section 5.3 for realizations that were simulated with a workflow involving the PPMT. In particular, the semivariograms of both variables were less continuous than that of the data at the first lag distance ($h = 50$). The issue is not the result of the chosen back-transformation method (Figure 5.16), and while a chained MAF transformation helps (Figure 5.20), a small loss of continuity persists. Figure 5.23 shows that although the PPMT leads to superior semivariogram reproduction overall, the alternative modeling workflows do not lead to this loss of short scale continuity.

This issue has been observed when applying the PPMT with other data sets,

including the case study in Chapter 6. It is believed to be the result of forcing dependent variables to be entirely independent at $h = 0$ lag distance, which is usually associated with a loss of short scale continuity in transformed space. Consider the semivariograms in Figure 5.12, where the original and PPMT transformed semivariograms are virtually identical at the majority of lags. The notable exception, however, is the loss of short scale continuity, particularly for Z_2.

A practical solution to this problem is demonstrated in Chapter 6, where semivariogram models of inflated continuity are used as input to the Gaussian simulation. It uses the following workflow:

i) Fit semivariogram models in transformed space, before simulating and back-transforming one realization.

ii) Inspect the semivariogram reproduction and repeat step (i) with semivariogram models of more or less continuity if required.

While not theoretically attractive, this approach has been found to yield reasonable semivariogram reproduction in challenging settings. As discussed in Chapter 8, understanding and correcting this issue is an item of future work.

5.4.4 Chained MAF Workflow

The PPMT and MAF transformations are suited to differing forms of geological data and are not viewed as strictly competing techniques. If multivariate relationships are reasonably linear, MAF is likely to outperform the PPMT since it directly targets spatial correlation. Conversely, the PPMT is advocated for complex multivariate data since it will transform the variables to be truly independent at $h = 0$ lag distance.

Given their complimentary strengths and weaknesses, PPMT and MAF may be effectively used together within chained transformation workflows. As demonstrated in Section 5.3.5, the two techniques have been found to yield consistently superior results when used in combination rather than isolation.

Chapter 6

Nickel Laterite Case Study: Data Transformation

The following chapter uses Anglo American's Barro Alto nickel (Ni) laterite mine to study the performance and value of the projection pursuit multivariate transformation (PPMT) in a complex geological setting. The chapter begins with an overview of the geology, mining, stockpiling, and plant processing of Barro Alto. This background establishes that the multivariate relationships of several Ni laterite variables have a large impact on the extraction process. If geostatistical realizations do not provide realistic multivariate characterization of the deposit, it will have consequences on resource management decisions and subsequent operations.

The value of the PPMT is measured as a function of the improvement in resource management decisions. To facilitate this study, jackknife validation is used for a large subset of the Ni laterite data. Geostatistical modeling then proceeds at the removed sample locations using workflows that do and do not incorporate the PPMT. The resultant realizations are used for technical decisions, which are validated against the correct decision (informed by the removed true value).

Beyond jackknife performance, many properties of the PPMT and its resultant geostatistical models are studied in detail. While it is useful to document the technique in a realistic setting, the dimensionality and geologic complexity will prevent exhaustive presentation and understanding of the results. The synthetic example in Chapter 5 compliments this study, where the reduced dimensionality and controlled setting permits thorough presentation and understanding of the results. The majority of figure and table formats in this chapter, as well as the statistics within them have been introduced in Chapters 3 and 5. Only new formats and statistics will be explained in detail.

6.1 Background

As a critical component in the manufacturing of stainless steel and other non-ferrous alloys, Ni is an important metal to global industry. The metal occurs within two distinct geologic mineralizations. The first is referred to as sulphides, which occur when Ni is precipitated from hydrothermal fluids within ultramafic intrusions. Generally mined using underground methods, sulphides account for $\sim 30\%$ of the world's known Ni, though they yield $\sim 70\%$ of current production (Anglo-American, 2012). The second type is referred to as laterites, which occur when surficial weathering leaches Ni from ultramafic rock. Generally mined using shallow open pit methods, laterites account for $\sim$ 70% of the world's known Ni, though they yield $\sim$ 30% of current production (Anglo-American, 2012). The mining industry has historically favored sulphide deposits since less capital intensive extraction processes are required for Ni recovery. Improving technology has made laterites commercially viable; however, and given their relative abundance, laterites are expected to be more important than sulphides to the future of the Ni industry (Info-Mine, 2012).

Located in the Brazilian state of Goiás, Anglo American's Barro Alto Ni mine exploits a laterite deposit. The mineralization occurs in the Barro Alto mafic-ultramafic complex and is composed of Ni rich saprolite that is overlain by Ni leached laterite. Neufeld et al. (2008) detail that the economic ore body is divided into two geologically distinct zones. West-type-ore (WTO) has relatively high Ni grade, but also high iron (Fe) and silica (SiO_2) to magnesia (MgO) ratio (SMR). As will be explained, high values of Fe and SMR are problematic for the Barro Alto plant process. Conversely, East-type-ore (ETO) has relatively low Ni grades, but also has lower Fe and SMR. As a result, the effective extraction of Ni will require blending of WTO and ETO to regulate Fe and SMR values in feed to the plant.

Barro Alto began production in 2011 and uses a rotary kiln electric furnace (RKEF) process to extract ferronickel (FeNi) (Info-Mine, 2012). While every step that comprises the RKEF process goes beyond the scope of this thesis, the component that may be most impacted by geostatistical modeling is the smelting of ore in electric arc-induction furnaces (Figure 6.1). As described in Neufeld et al. (2008), the Ni, Fe, SiO_2, and MgO content of the furnace feed is critical. First, Ni grade should be high enough to yield sufficient recovery in the final FeNi product (25-30% Ni); Ni feed grade should therefore be maintained above 1.5%. High Fe content can

Figure 6.1: Ni being poured at the Barro Alto plant (Anglo-American, 2011)

lead to poor recoveries due to incomplete separation of the slag (byproduct of smelting); Fe should therefore be held below 18.5%. Finally, high SMR generates excess heat that may damage the furnace lining; SMR should therefore be held below 1.75.

To manage these requirements, mined material is stockpiled, crushed and stacked to homogenize the ore and reduce its variability. Neufeld et al. (2008) present some stockpiling and stacking options that have been considered for Barro Alto. In general, geostatistical modeling is the key input to this process. It is critical for Ni, Fe, SiO_2, MgO and the relationships between them to be realistically characterized. Failure to do so will lead to errors in technical decision making that transfer through plant processing. Material will be allocated to the wrong stockpiles, variability will not be sufficiently reduced by stacking, and furnace feed will not possess the required characteristics. Grade control sampling at each stage may mitigate this issue, but geostatistical modeling remains critical for long term planning.

6.2 Data Inventory and Preparation

The data that has been made available for this case study is a subset of a dated Barro Alto database. It consists of 18,352 observations that homotopically sample Ni, Fe, SiO_2, and MgO. The observations have been sampled using bore holes that are drilled from the surface in vertical and deviated orientations. The spatial coordinates of the observations are present in the form of Easting, Northing, and elevation above sea level. A rocktype variable categorizes the observations as one of four geologically

distinct ore mineralizations. Surface topography was also provided as displayed in Figure 6.2.

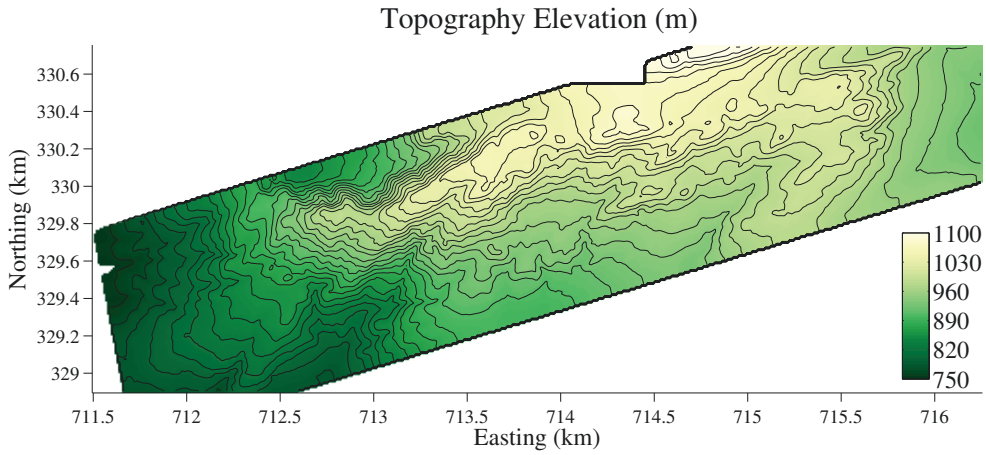


Figure 6.2: Barro Alto surface topography.

6.2.1 Stratigraphic Transformation

Although it is not a perfect relationship, the continuity and anisotropy of a laterite tends to align with its topography, reflecting the underlying surface weathering phenomena that yields the mineralization. If modeling were to proceed using elevation as the vertical coordinate system, the spatial continuity would follow varying orientations of anisotropy across the deposit. This compromises geostatistical modeling workflows that assume stationary continuity of the regionalized variables.

To avoid this issue, the coordinate system of the data is transformed so that the z-axis becomes depth below topography (rather than absolute elevation). Geostatistical modeling proceeds in this flattened space where continuity of the variables is relatively constant, before returning simulated realizations to the original coordinate system as a final step. This approach follows the standard practice of transforming absolute coordinates to a stratigraphically flattened system according to specifics of the geological architecture (Pyrcz and Deutsch, 2014). Figure 6.3 shows various perspectives of the flattened data, which have been transformed relative to the topographic elevation in Figure 6.2.

6.2.2 Jackknife Removal

The data in Figure 6.3 are colored as 'model nodes' or data observations. The model nodes are the 7,339 observations that have been removed for jackknife validation (e.g., test data), leaving the remaining 11,013 observations that will inform geostatistical modeling (e.g., training data). The jackknife removal was performed in a completely random manner so that global properties of the training data are representative of the test data. Given the near proximity of vertical conditioning data, modeling results at these locations are expected to be have greater local accuracy than that of a regular 3-D grid (on average). The comparative geostatistical modeling workflow enjoys this same benefit, however, revealing the accuracy that is attributed to the PPMT rather than specifics of the data conditioning.

One could consider alternative jackknife removal schemes, such as eliminating entire drillholes from the dataset. Such schemes yielded training data that were not representative of the test data, however, adding unnecessary complication to the evaluation of results. Declustering tools are available for determining the spatially representative statistics of a regular domain from irregularly spaced data (Deutsch and Journel, 1998). No such tools are available, however, for declustering data to be representative of a set of isolated locations (removed drillholes).

6.2.3 Univariate, Multivariate and Spatial Properties

Univariate and bivariate properties of the training data are presented in Figure 6.4 using a covariance matrix format, where CDFs appear along the diagonal and scatterplots are placed in the off-diagonal, or upper triangle locations. All of the multivariate complexities that were schematically represented in Figure 1.2 are present, including non-linearity, heteroscedasticity and constraints. Consequently, this multivariate distribution is not expected to be reproduced by geostatistical workflows that fail to remove these complexities prior to the application of Gaussian simulation algorithms.

Strong bivariate dependencies are observed between all of the variables. It is interesting to note, however, that only Fe-SiO_2 and Fe-MgO have strong absolute correlation. The remaining bivariate distributions have absolute correlation of less than 0.16. This is an excellent example of a complex multivariate distribution that is very poorly characterized by correlation coefficients. Bimodality is also present within the scatterplots and CDFs of Figure 6.4. These are primarily explained by the presence of multiple geological rocktypes, which are addressed with stationarity related subsetting in the next section.

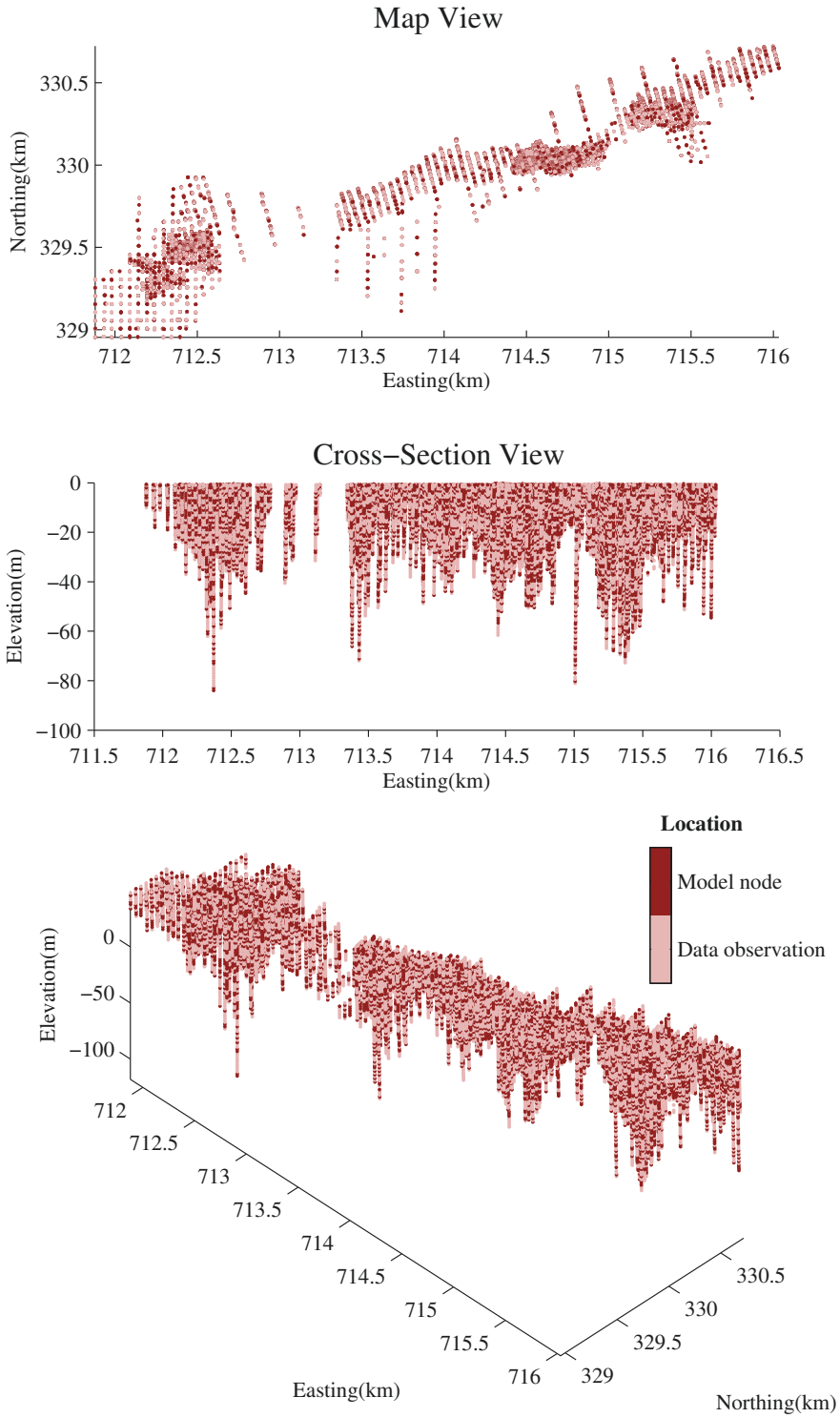


Figure 6.3: Various perspectives of the observation and model node locations.

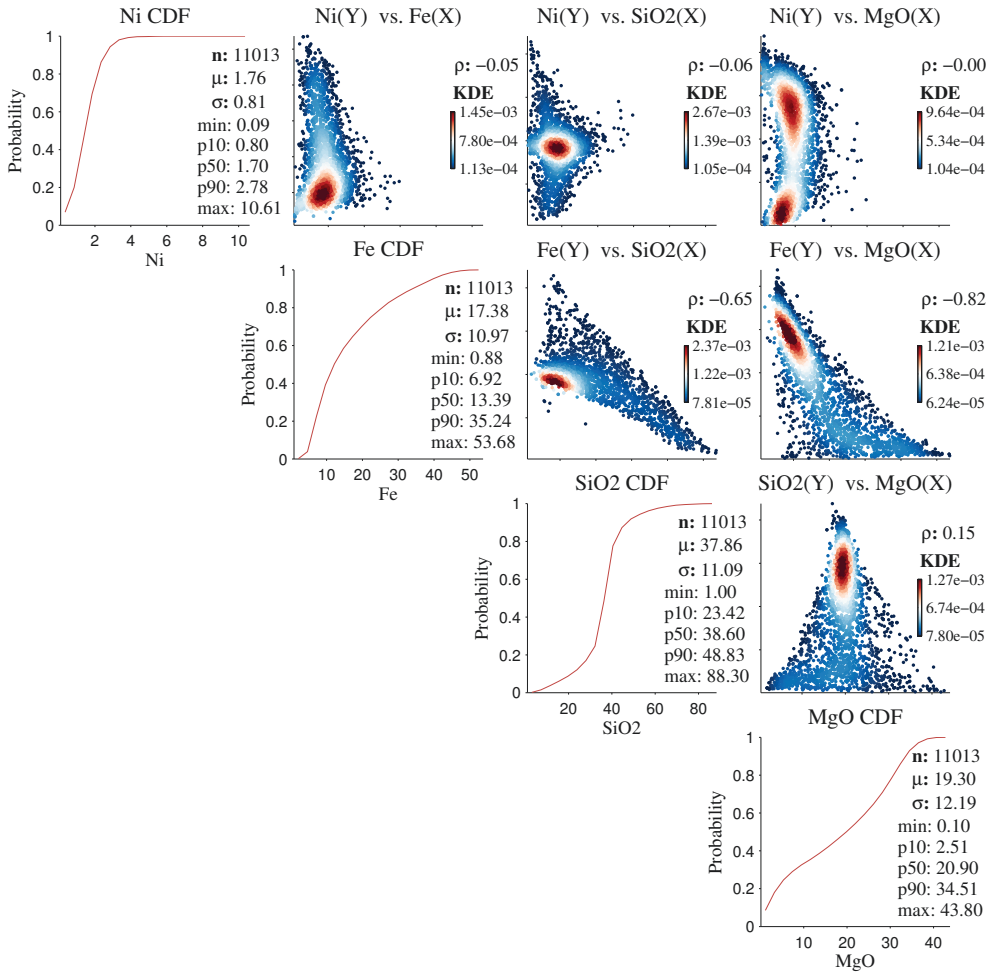


Figure 6.4: CDFs and KDE scatterplots of the data (all rock types).

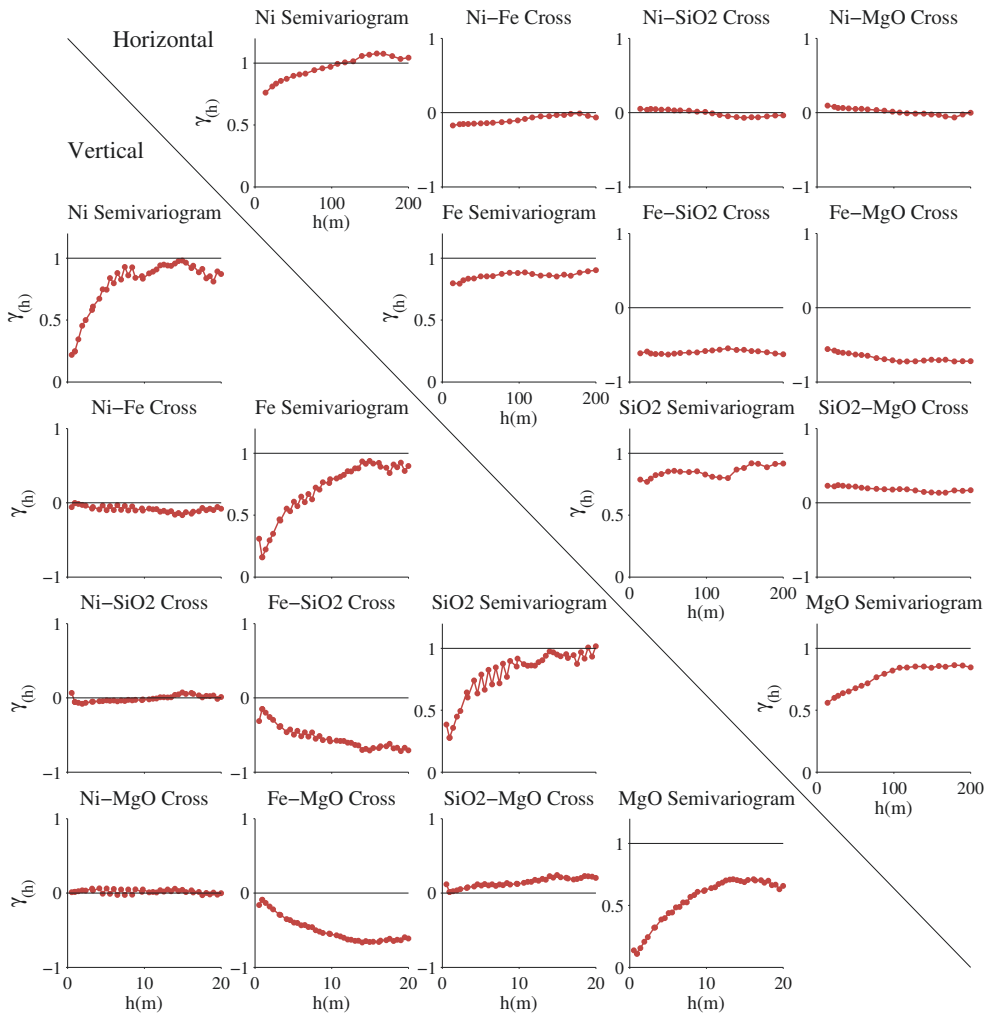


Figure 6.5: Semivariograms and cross-semivariograms of the data in the vertical and horizontal directions (all rocktypes).

Spatial continuity of the variables is presented in Figure 6.5 using a covariance matrix-like format. Due to anisotropy of the deposit, the horizontal direction is more continuous than the vertical and must be presented on a different scale. As such, horizontal semivariograms are placed in the upper triangle displaying lags to $h = 200$m distance; vertical semivariograms are placed in the lower triangle displaying lags to $h = 20$m distance. The 'double-diagonal' locations display the semivariograms, while cross-semivariograms are placed in the off-diagonal locations.

A large range of relative continuity is exhibited by the semivariograms: i)Ni is discontinuous, ii) Fe and SiO_2 are moderately continuous, and iii) MgO is very continuous. These observations apply to both the horizontal and vertical directions. This may be a cause for concern, since the PPMT is transforming variables of widely varying continuity. Depending on the degree of mixing that is required to make them uncorrelated and multiGaussian, it may be difficult to recover the unique spatial structure of each variable.

The nature of the cross-semivariograms are somewhat expected given the correlations at $h = 0$ lag distance in Figure 6.4. Consider that Ni has strong bivariate relationships that are not characterized by the associated correlations. Similarly, although the Ni cross-correlations are near zero at all lags, it is reasonable to expect that complex relationships continue to exist with the other variables at $h > 0$ lag distances. While multivariate relationships at $h > 0$ are only inspected using cross-covariances, this is a subject of future work that is reviewed in Chapter 8.

6.2.4 Rocktype Subsetting

Recall from Section 2.1.1 that prior to geostatistical modeling, the data must be pooled into stationary domains based on geologic factors such as lithology, mineralization, structure or alteration. In this case, the data are subset according to the provided geologic rocktypes. Modeling of the four variables will proceed in parallel for each rocktype, before combining the results.

The rocktypes are numbered as: 1) Basic ETO, 2) Basic WTO, 3) Acid ETO, and 4) Acid WTO. The number of observations and average grades of each rocktype (RT) is provided in Table 6.1. As was described in Section 6.1, the WTO rocktypes have lower Ni grade than their ETO equivalent, but this comes at the expense of higher Fe grades and SMR ratios. Table 6.1 also shows that Ni and MgO grades are higher for the acid rocktypes, while Fe is higher for the basic rocktypes. While

the rocktypes have varying SMR, the acid WTO (RT4) is notably higher than the others.

Table 6.1: Rocktype description, number of observations and the mean grade (%) of each variable.

RT	Description	n	Ni	Fe	SiO_2	MgO
1	Basic ETO	5,968	1.9	12.6	36.7	27.4
2	Basic WTO	4,016	2.0	12.9	38.8	24.9
3	Acid ETO	1,389	1.0	16.4	36.5	22.9
4	Acid WTO	6,979	1.7	24.0	38.4	8.7

Figure 6.6 displays the rocktype classification of each data observation from various orientations. If the map view were divided by a line proceeding roughly southwest to northeast, the majority of ETO rocktypes (dark colors, RTs 1 and 3) would lie to its southeast. Conversely, the majority of WTO rocktypes (light colors, RTs 2 and 4) would to lie to its northwest. Of course, it is from this distinct geographical separation that the ETO and WTO rocktypes derive their name. As discussed in Neufeld and Deutsch (2008), the cross-section view reveals that the WTO rocktype has a thick laterite profile that regularly extends beyond 50m depth. Consistent patterns of spatial variability are more difficult to discern when comparing the basic rocktypes (blue colors, RTs 1 and 2) with the acid rocktypes (red colors, 3 and 4).

CDFs and scatterplots of each rocktype are overlain in Figure 6.7 to summarize the univariate and bivariate populations that are pooled within each rocktype. The bimodality that was noted in Figure 6.4 has largely been resolved, although it continues to exist within MgO for Acid ETO (dark blue, RT 3). Additional subsetting could be considered to address this, although geologists likely chose not to given the relatively low sample population that is available for Acid ETO.

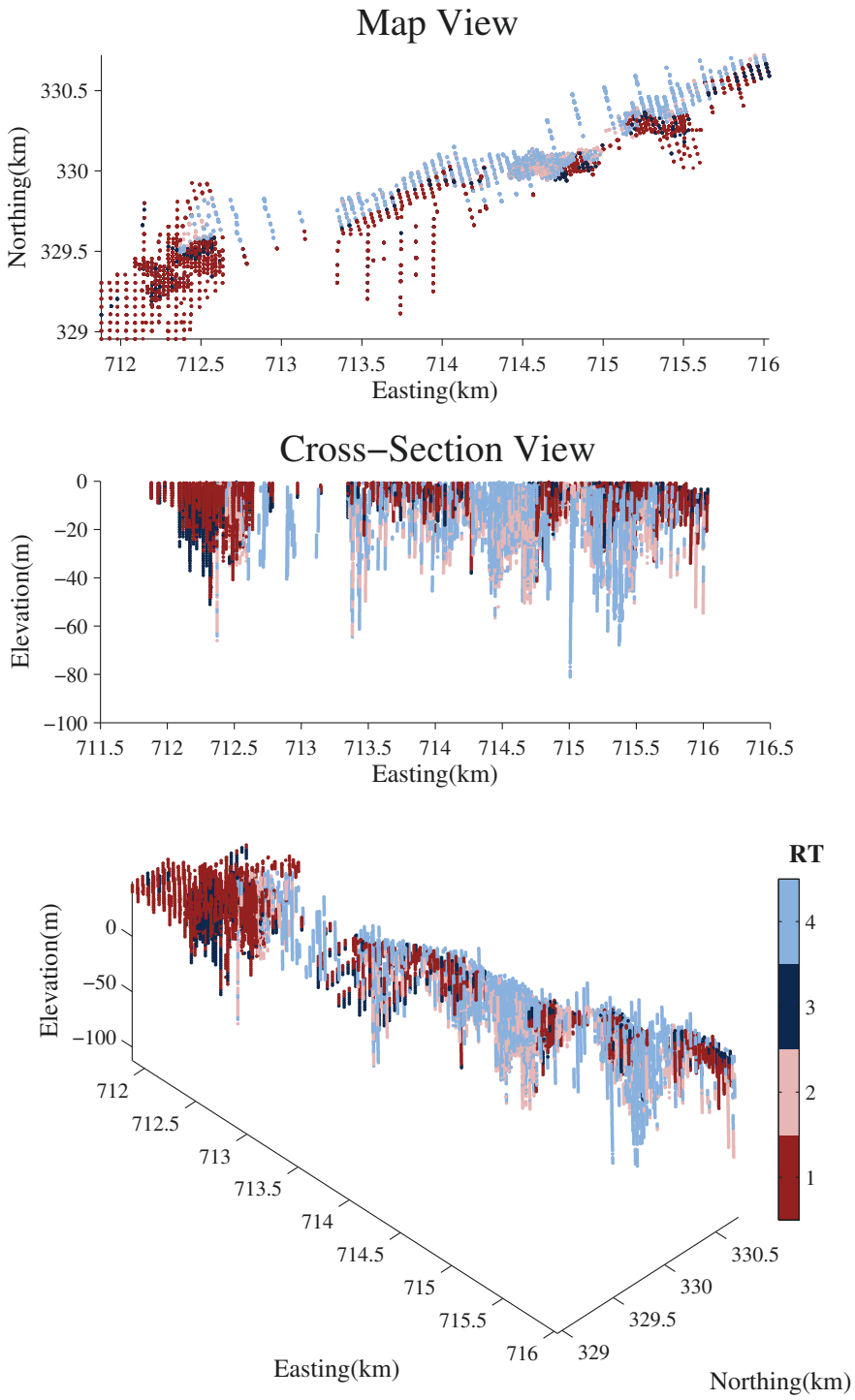


Figure 6.6: Various perspectives of the data locations, colored by rock type.

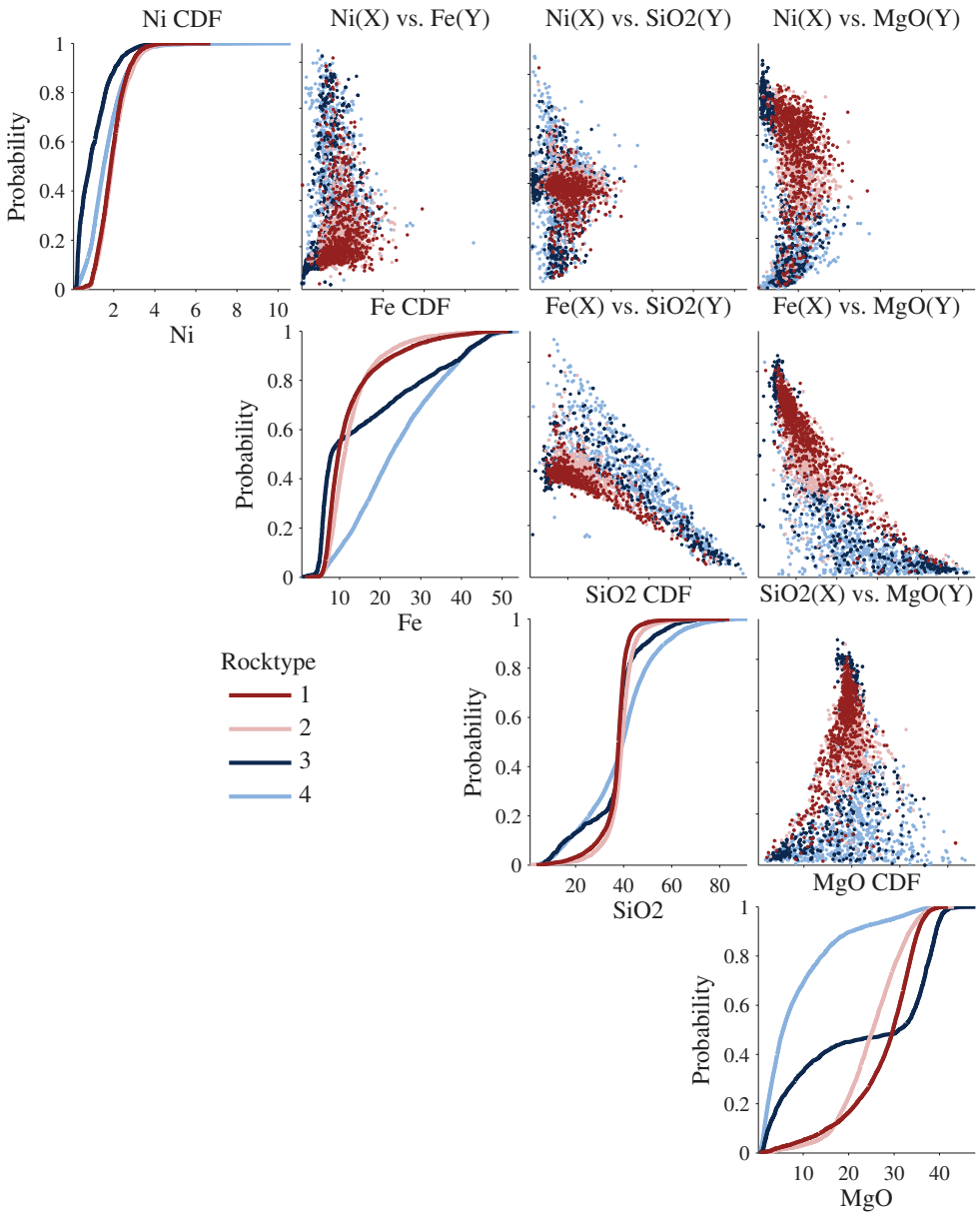


Figure 6.7: CDFs and scatterplots of the data, colored by rock type.

6.3 PPMT Transformation

The following section displays the PPMT transformation of the Ni laterite variables for one rocktype (RT 1). As previously described, multivariate transformation and geostatistical modeling will take place in parallel for each rocktype, before combining the simulated results. Although the different multivariate distributions of the remaining rocktypes leads to differences in their PPMT transformation, the overall nature of the results are similar to RT 1.

6.3.1 Visualization of Each Step

The univariate and bivariate distributions of RT 1 are displayed in Figure 6.8, which show that isolated populations of the full distribution (Figure 6.4) are present within this rocktype.

The PPMT initializes with the normal score transformation, which is applied to remove univariate complexity and outliers prior to subsequent data sphereing. The univariate and bivariate distributions are displayed for the transformed data in Figure 6.8. While CDFs show that the variables have been made univariate standard Gaussian, the KDE scatterplots continue to exhibit a great deal of bivariate complexity. Interestingly, these bivariate relationships appear to be more complex than their original form in Figure 6.8. Since additional steps are clearly required to make this data Gaussian, it is appropriate to consider the PPMT.

The second step of the PPMT, data sphereing, is applied next to decorrelate the normal score data. The univariate and bivariate distributions are displayed for the sphere data in Figure 6.8. Recall that data sphereing yields an identity covariance matrix; in addition to being uncorrelated, all of the variables also have a variance of one. Nevertheless, obvious bivariate complexity remains in the KDE scatterplots.

Finally, one hundred iterations of projection pursuit are used to transform the sphere data to a multiGaussian distribution. The univariate and bivariate distributions are displayed for the PPMT data in Figure 6.8. CDFs and their summary statistics display standard univariate Gaussian values, while the KDE scatterplots mimic the contours of an uncorrelated multiGaussian distribution.

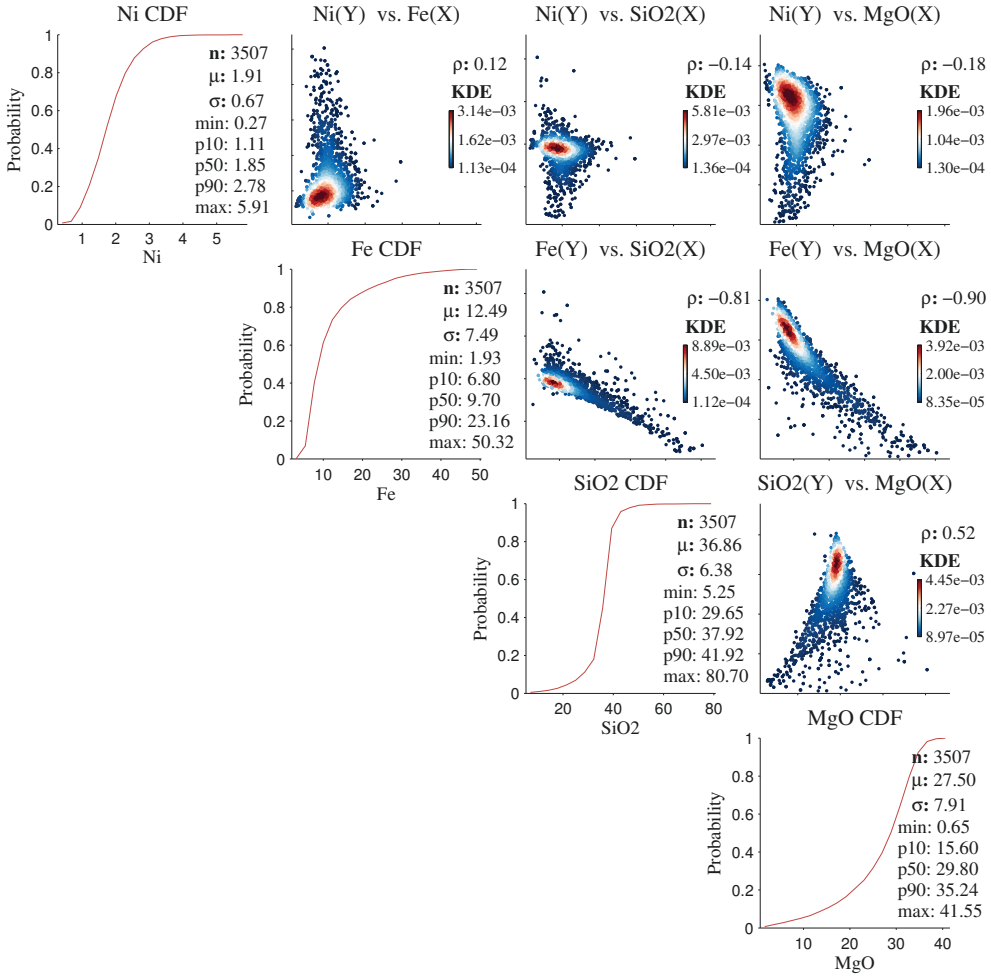


Figure 6.8: CDFs and scatterplots of the original data (RT 1).

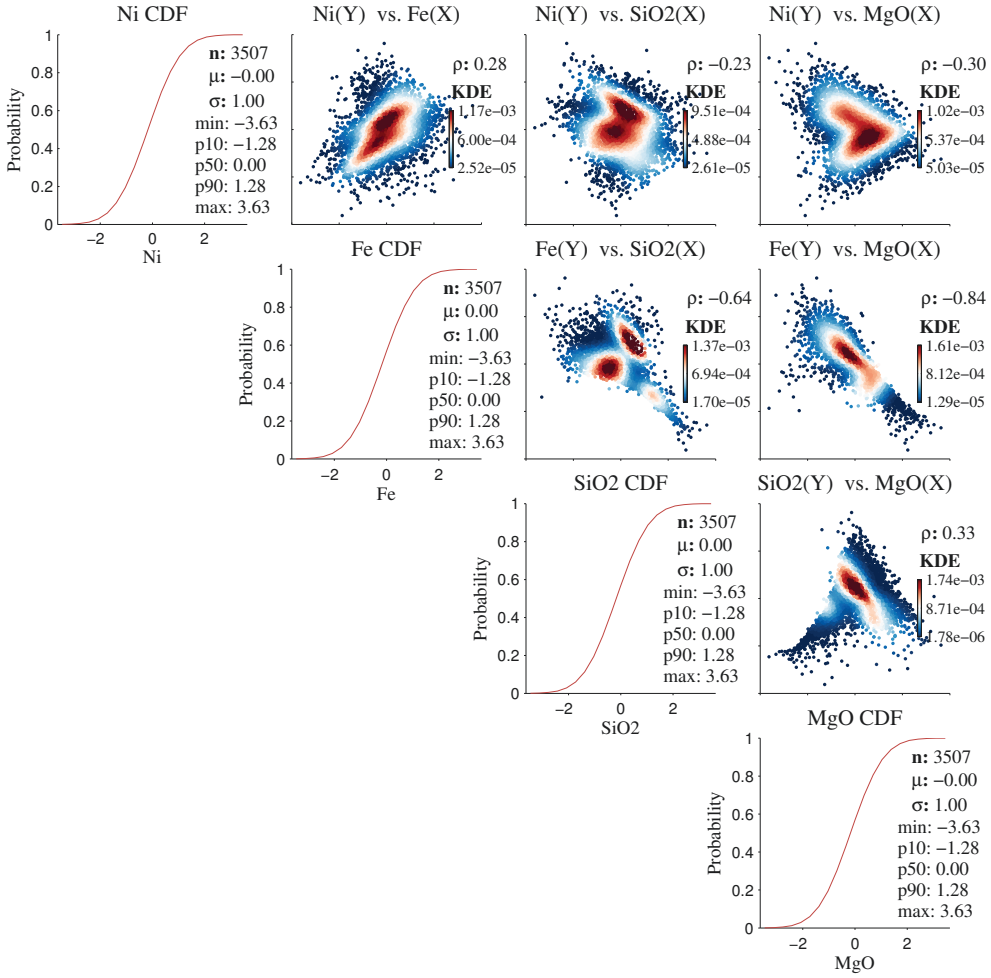


Figure 6.9: CDFs and scatterplots of the normal score data.

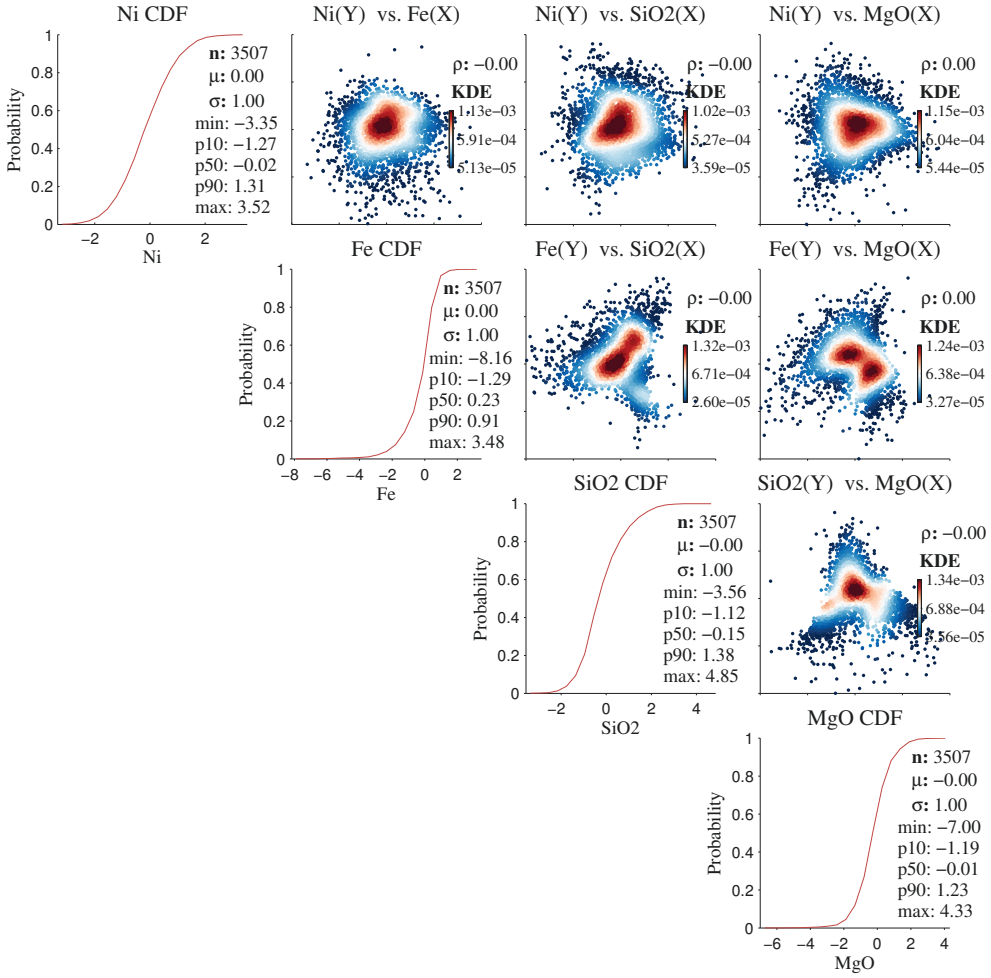


Figure 6.10: CDFs and scatterplots of the sphere data.

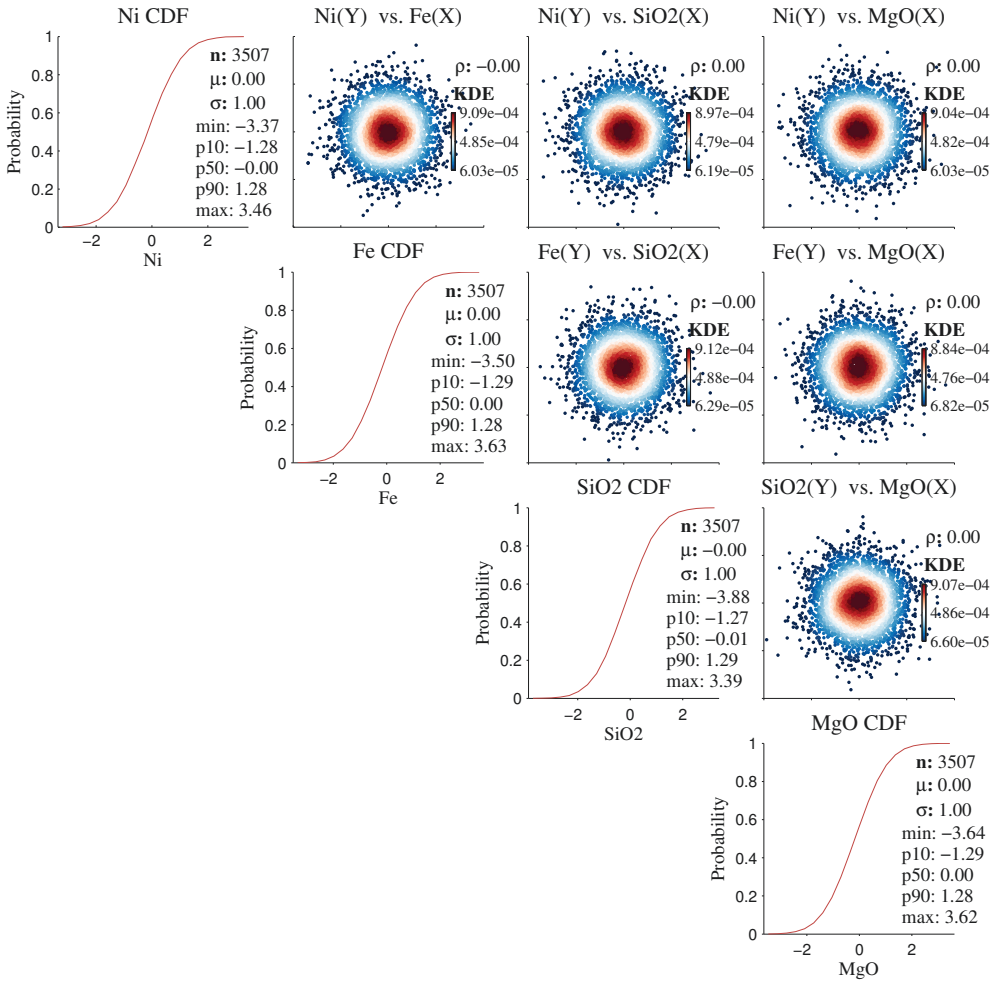


Figure 6.11: CDFs and scatterplots of the PPMT data.

6.3.2 Gaussianity and Decorrelation

While KDE scatterplots in Figure 6.11 suggest that the transformed data is very multiGaussian, this property will now be quantified. The maximum projection index, $I(\boldsymbol{\theta})$, of each projection pursuit iteration is displayed in Figure 6.12. Recall that $I(\boldsymbol{\theta})$ is a test statistic that is large when the projection, $\mathbf{p} = \mathbf{X}\boldsymbol{\theta}$, is non-Gaussian, and zero when $\mathbf{p}$ is standard Gaussian. The overlain percentiles are drawn from the distribution of projection indices, $\mathbf{I}$, which is calculated as the basis for PPMT stopping criteria. The distribution is generated through calculating I within random Gaussian distributions of matching K variables and n observations.

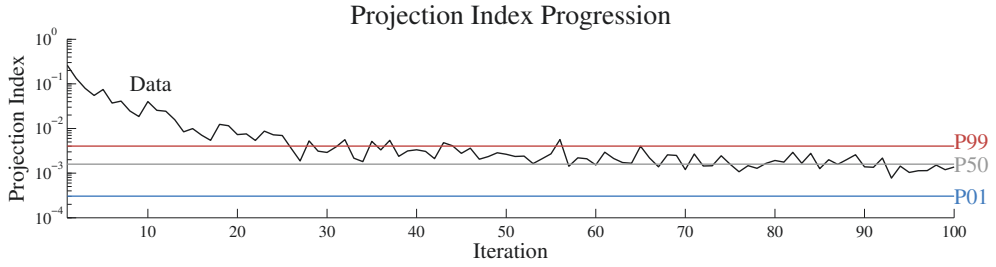
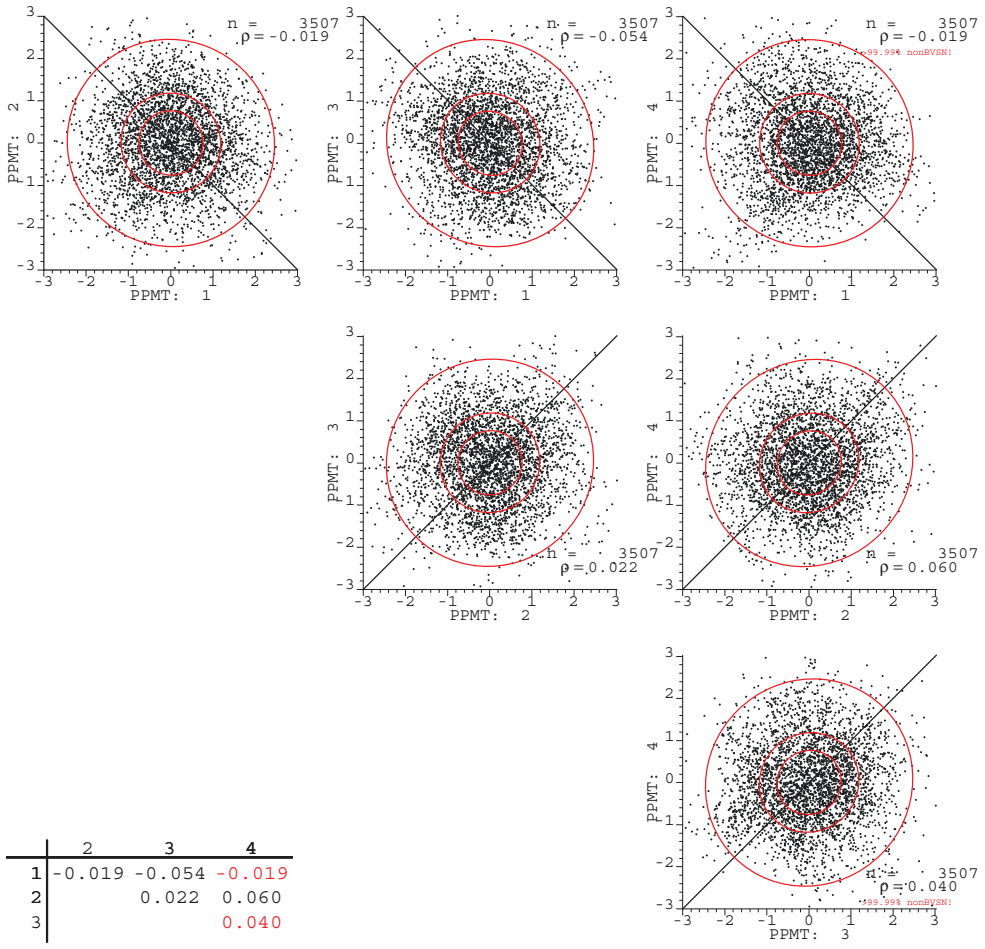


Figure 6.12: Projection index across one hundred iterations, with percentiles of $\mathbf{I}$ overlain for comparison.

Noting the logarithmic y-axis of Figure 6.12, the descending maximum projection index indicates that the majority of complexity is resolved within the first twenty five to thirty iterations. According to $\mathbf{I}$, however, the distribution is barely multi-Gaussian after this many iterations. While complexity is resolved at a much slower rate for iterations 30 to 100, improvements are nevertheless made to the Gaussianity of the distribution. After one hundred iterations, the 50^{th} percentile of $\mathbf{I}$ indicates that the data matches the Gaussianity of a typical random Gaussian distribution.

The `scatnscores` BVSN test (Section 4.2.2) is used once again as an independent check of the PPMT Gaussianity. Figures 6.13 and 6.14 display the `scatnscores` plots of the data following ten and twenty projection pursuit iteration, respectively. For $K > 2$ variables, `scatnscores` outputs a correlation table in the bottom left of its plot. These correlations are colored red if they fail the BVSN. Four of the six pairs pass the BVSN after ten iterations, while all of the distributions pass after twenty iterations. This result corroborates the stopping criteria of the PPMT.

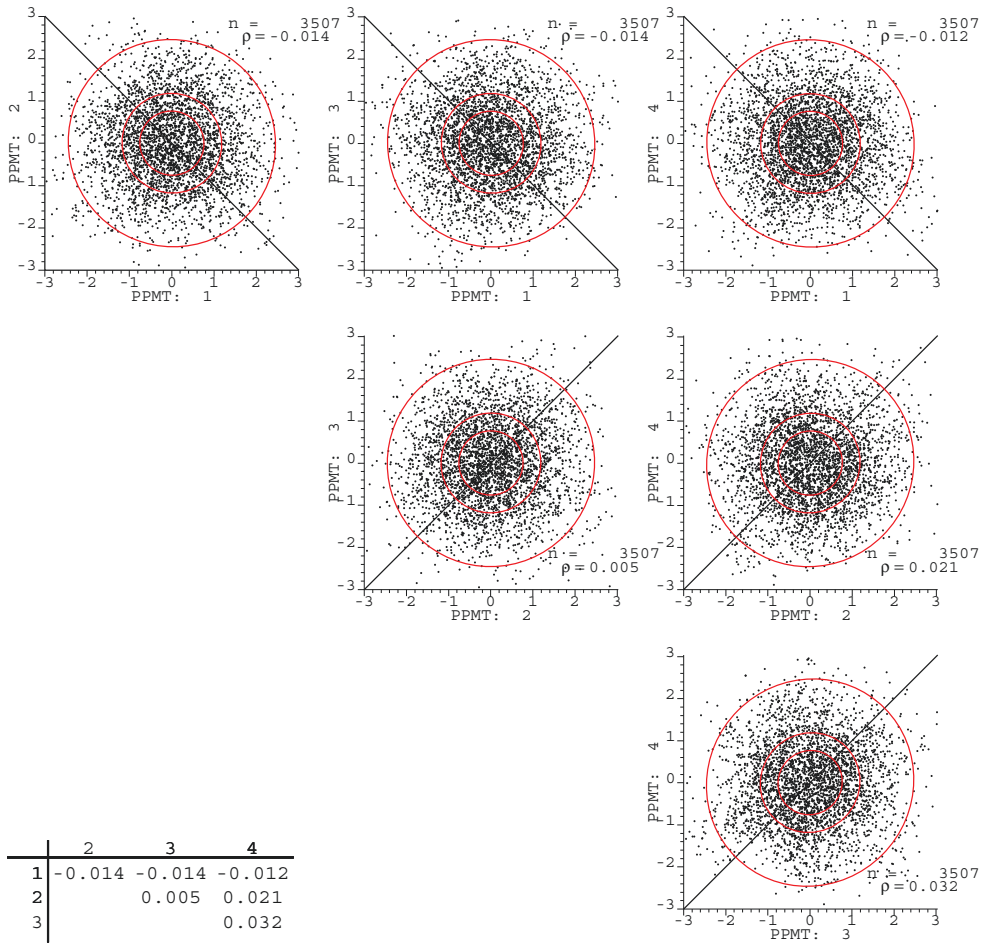
The progression of decorrelation across the projection pursuit iterations is displayed in Figure 6.15. While bearing the small scale of the y-axis in mind (-.14



	2	3	4
1	-0.019	-0.054	-0.019
2		0.022	0.060
3			0.040

Figure 6.13: `scatnscores` plot and Gaussianity test of the 10^{th} projection pursuit iteration .

to 0.15), the variables are not entirely decorrelated in this case until the $\sim 60^{th}$ iteration. Summarizing, the PPMT transforms these complex variables to an uncorrelated multiGaussian distribution, although many projection pursuit iterations are required to so. Some concern may arise in terms of the changes that these iterations could incur to the original multivariate configuration and spatial continuity; this is examined in the next section.



	2	3	4
1	-0.014	-0.014	-0.012
2		0.005	0.021
3			0.032

Figure 6.14: `scatnscores` plot and Gaussianity test of the 20^{th} projection pursuit iteration .

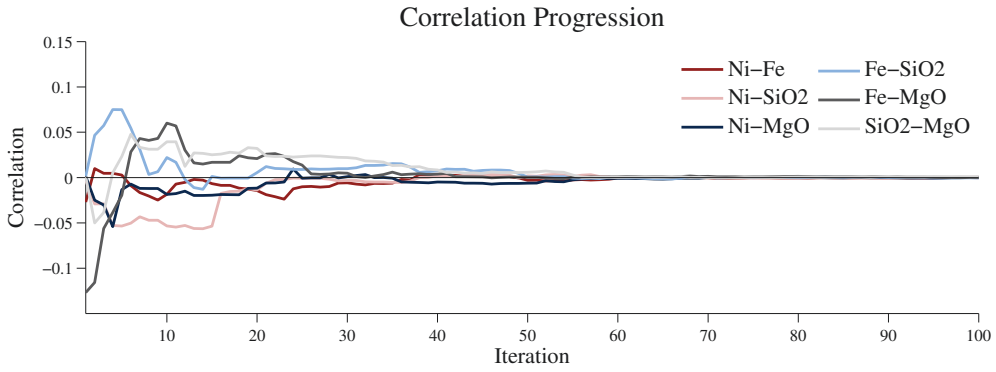


Figure 6.15: Correlation of each pair across one hundred projection pursuit iterations.

6.3.3 Mixing and Spatial Structure

The previous section established that the PPMT can effectively transform the Ni laterite data to an uncorrelated multiGaussian distribution. The nature of the underlying transformation and its effect on the spatial structure of the data will now be examined. Plots in Figure 6.16 display how the original variables have been loaded onto the transformed variables.

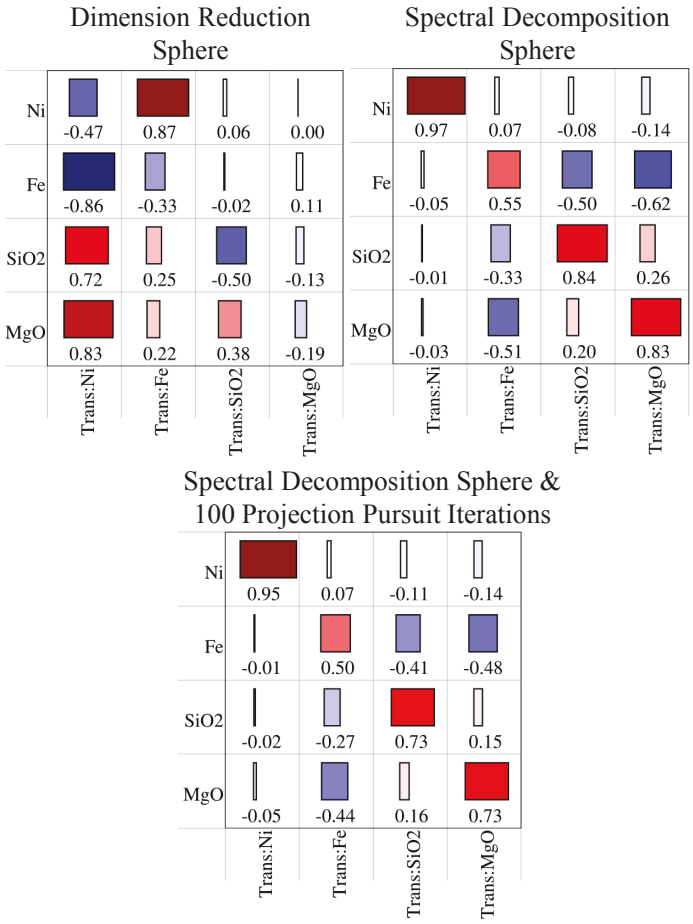


Figure 6.16: Loading plot of the transformed variables following DRS (top left), SDS (top right), and fifty projection pursuit iterations preceded by SDS (bottom).

As in Section 5.3.3, the dimension reduction sphereing (DRS) result is provided as a relative comparison to the implemented spectral decomposition sphereing (SDS) result. DRS maximizes the variability that each descending transformed variable explains, leading to large absolute loadings for the first transformed variable in Figure 6.16. Consequently, the original variables are heavily mixed within the first

transformed variable, potentially complicating the recovery of their distinct spatial continuity following simulation and back-transformation.

This is the motivation for the SDS sphereing that is applied by the PPMT, which primarily loads each variable onto the associated transformed variable. While this effectively minimizes mixing of the variables, Figure 6.16 shows that the decorrelation of this multivariate system requires the mixing of Fe, SiO_2 and MgO in transformed space. As with the bivariate example in Section 5.3.3, the subsequent PPMT iterations are seen to shift the loadings, though the overall structure of the sphereing step is largely preserved.

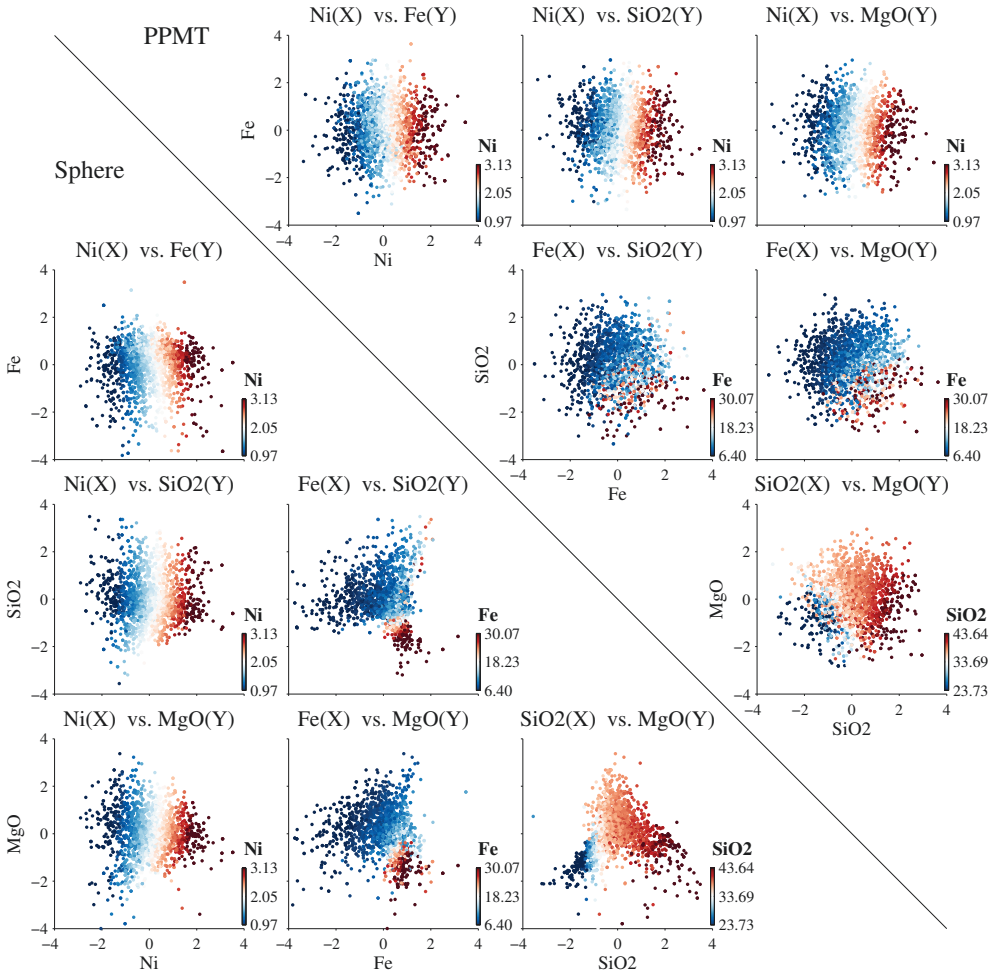


Figure 6.17: Scatterplots of the sphere and PPMT data, colored by their original x-axis value.

A more visual method for inspecting the nature of transformations and their resultant variable mixing is to color the transformed observations according to the original values. Doing so provides insight into the relative shift of the multivariate configuration, as was schematically illustrated in Figure 4.5. The sphere and PPMT transformed data are colored by the original values in Figure 6.17. The coloring in this figure corroborates the loadings in Figure 6.16. For example, Ni is mixed the least according to its loadings; plots that are colored by Ni show little rotation and a smooth gradient of color. Conversely, Fe is mixed the most according to its loadings; plots that are colored by Fe show the largest rotation and a relative mixing of colors. The overall nature of the PPMT coloring clearly reflects the sphereing step. Ni continues to exhibit a smooth gradient of color, whereas the mixing that was noted for Fe appears to be exacerbated by subsequent projection pursuit.

Given the above observations, it is expected that the spatial continuity of Fe is altered the most in transformed space, whereas Ni is altered the least. Semivariograms of the original, normal score, sphere, and PPMT data confirm this expectation in Figure 6.18. Fe continuity decreases dramatically in the horizontal and vertical direction, whereas Ni continuity increases only slightly in the horizontal direction. It is also interesting to note that the SiO_2 continuity is largely preserved, despite the large correlation and cross-correlation that previously existed for Fe-SiO_2 . This suggests that forcing two strongly correlated regionalized random variables to be independent at $h = 0$ will lead to destructuring of at least one variable in transformed space.

The cross-semivariograms display that removing correlation at $h = 0$ lag distance has largely removed correlation at $h > 0$ lag distances. Some cross-correlation does remain, however, which motivates a min./max. auto-correlation factors (MAF) transformation that is used for further spatial decorrelation in the applied workflow.

Figure 6.18 also shows that spatial continuity is significantly altered by projection pursuit in some cases, according to differences that exist between sphere and PPMT semivariograms. To understand how these changes progress, Figure 6.19 plots the semivariograms of each projection pursuit iteration. With the exception of horizontal MgO, the large continuity changes are incurred in the first twenty five projection pursuit iterations. This is not surprising given that the largest changes to the multivariate system occur in the those iterations according to the projection index (Figure 6.12). Given that most changes occur in the early states of projec-

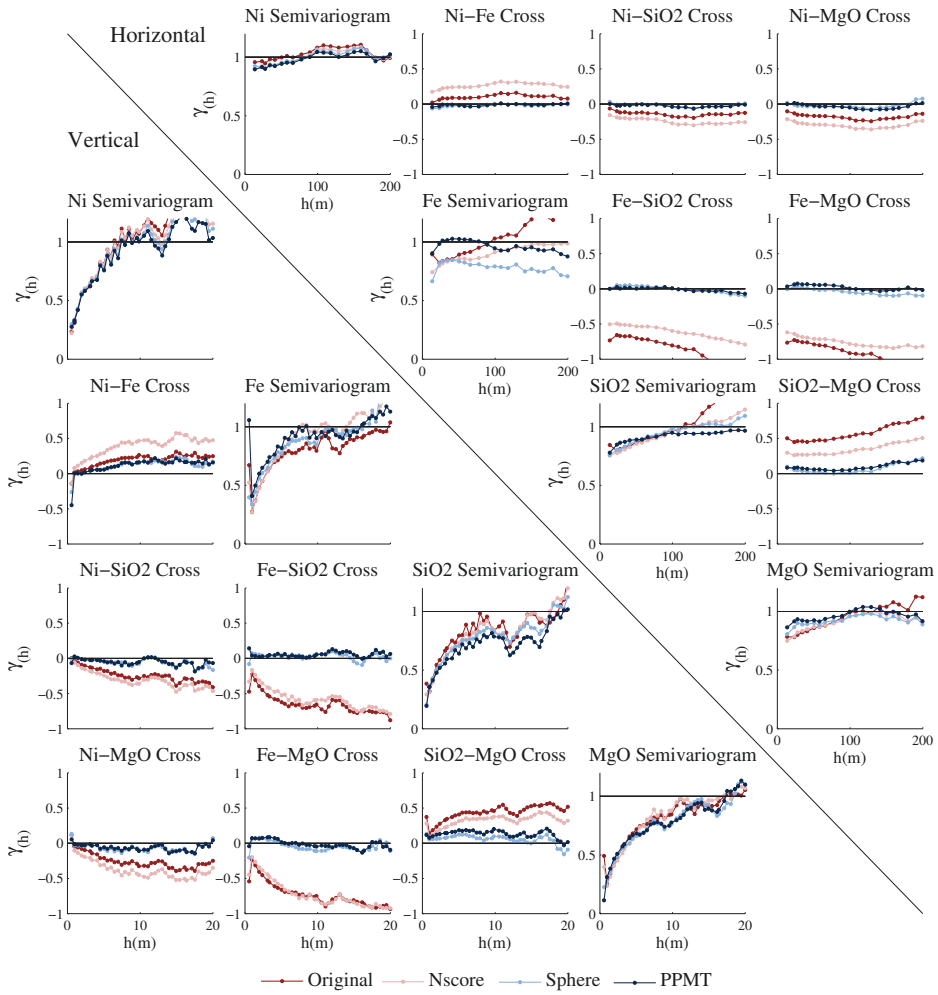


Figure 6.18: Semivariograms and cross-semivariograms for the original and transformed data .

tion pursuit, this result supports the previous conclusion that a large number of iterations is not overly consequential to spatial continuity.

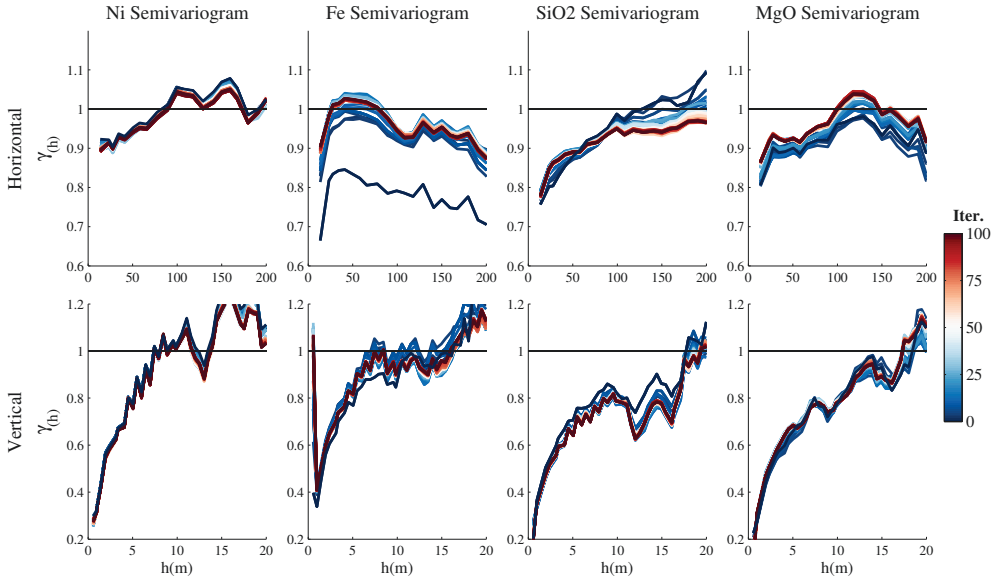


Figure 6.19: Semivariograms and cross-semivariograms following each of the one hundred projection pursuit iterations.

6.4 Simulation Results

Following the PPMT transformation from the previous section, the data have been made uncorrelated multiGaussian at $h = 0$. Observing the remnant cross-correlation at $h > 0$m distances (Figure 6.18), MAF is applied next to decorrelate the variables at $h = 10$m. Simulation then proceeds at the jackknife validation locations that were displayed in Figure 6.3. Using semivariogram models and the transformed data as input, SGSIM is executed to independently simulate one hundred realizations of each variable. The realizations are returned to original space using the MAF and PPMT reverse projection (RP) back-transformations. The RP method is used rather than Gaussian mapping (GM) for the back-transformation for the reasons that were developed and discussed throughout Chapter 5.

After combining rocktypes, properties of the realizations are validated against the removed jackknife data. A geostatistical workflow that does not incorporate the PPMT is used to provide a relative benchmark of the results. This includes a comparison of the resource management decisions that each modeling workflow provides, which is used as a metric of value in this case study.

6.4.1 Initial Results

Reproduction of the global univariate distributions is shown in Figure 6.20. Overall reproduction is confirmed, showing that the multivariate transformation workflow has successfully restored the univariate distribution of each variable. The largest mean bias (though small) is observed for the Fe and SiO_2variables, which may relate to the relatively high mixing of those variables in the transformation.

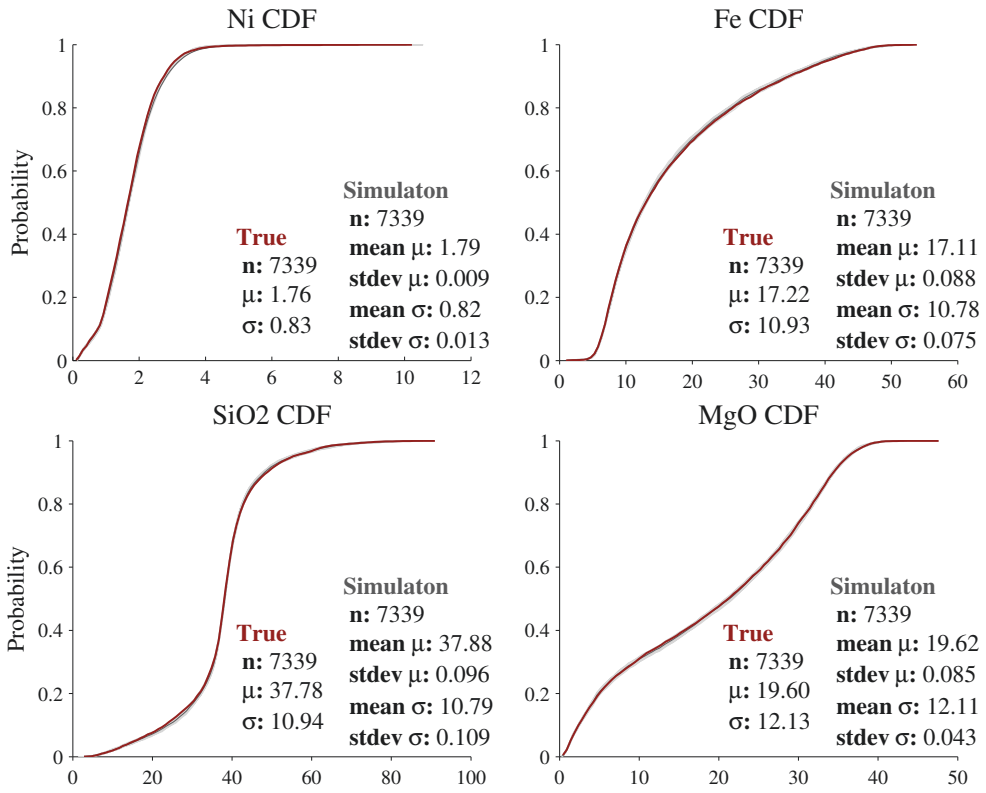


Figure 6.20: CDFs of the simulated realizations with the true CDFs overlain for comparison.

Local accuracy of the simulated realizations is examined in Figure 6.21, where the e-type mean of each location is compared with the true value. The spatial continuity of each variable is reflected in its local accuracy, as high continuity generally improves prediction (e.g., MgO). Everything else being equal, effectively characterizing the multivariate distribution of the variables should lead to increased accuracy, which is demonstrated in the subsequent comparisons.

Reproduction of the bivariate distributions is examined in Figure 6.22, where KDE scatterplots of the jackknife data (upper triangle) are compared with a sim-

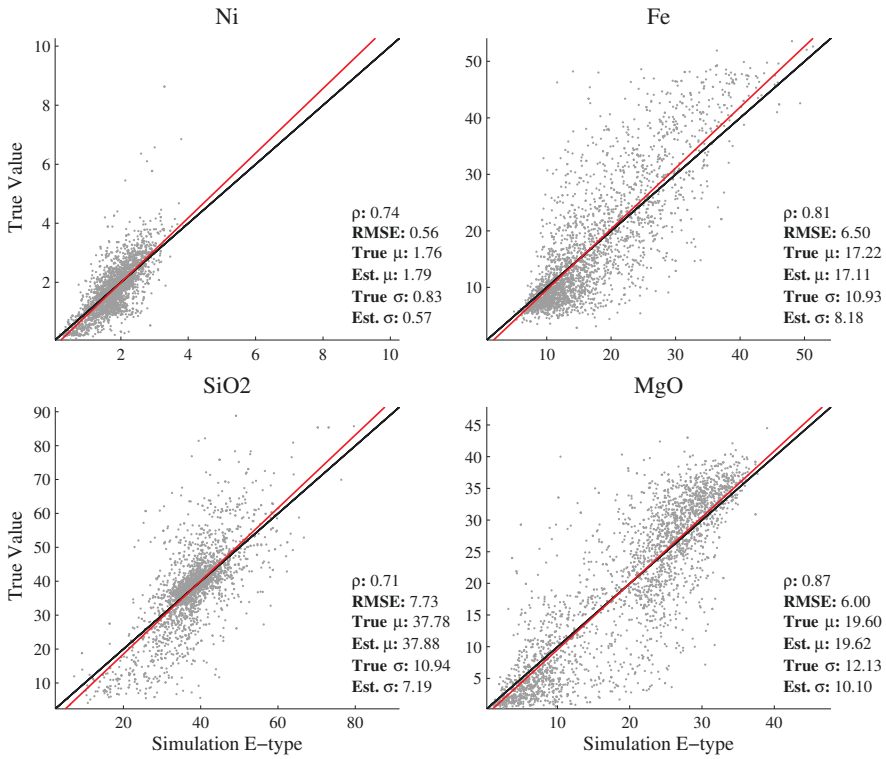


Figure 6.21: Scatterplots and summary statistics that compare the simulated e-type with associated true values.

ulated realization (lower triangle). Excellent reproduction of the complex features is seen overall. One small concern is the infrequent occurrence of simulated points in space that appears to lie beyond the concave hull of the data. This is most evident in the Fe-SiO_2, Fe-Mgo and SiO_2-MgO plots. This is a potential issue of the RP back-transformation documented earlier. The GM back-transformation could be considered as an alternative if strict adherence to multivariate constraints is a priority.

Spatial continuity is examined in Figure 6.23, where semivariograms and cross-semivariograms of the simulated realizations are overlain with that of the jackknife data. Reproduction is excellent in the horizontal direction, as the unique continuity of each variable has been recovered despite the mixing concerns that were noted in the previous section. A loss of short scale structure is present in the vertical direction. Following spatial destructuring of the transformation (Figure 6.18), the back-transformation does not restore the full original vertical continuity to the simulated realizations.

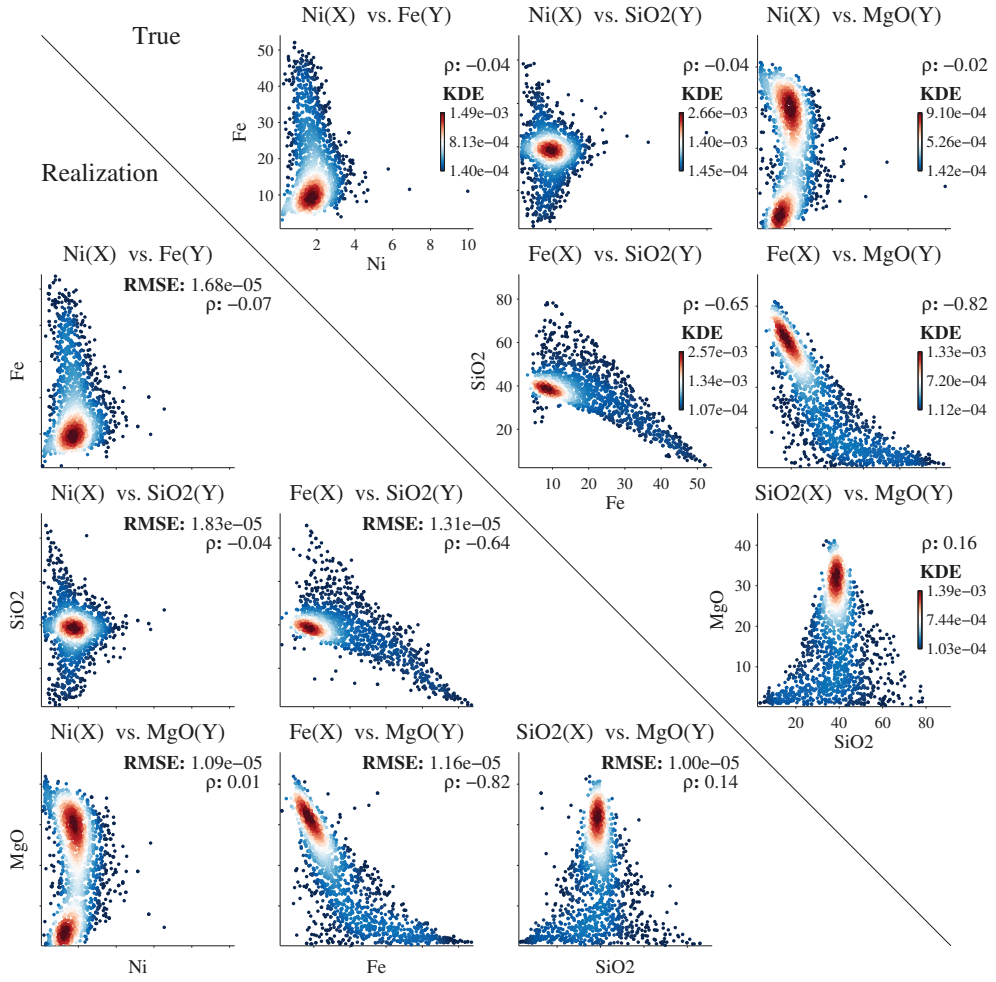


Figure 6.22: KDE scatterplots of a realization, with the true values shown for comparison.

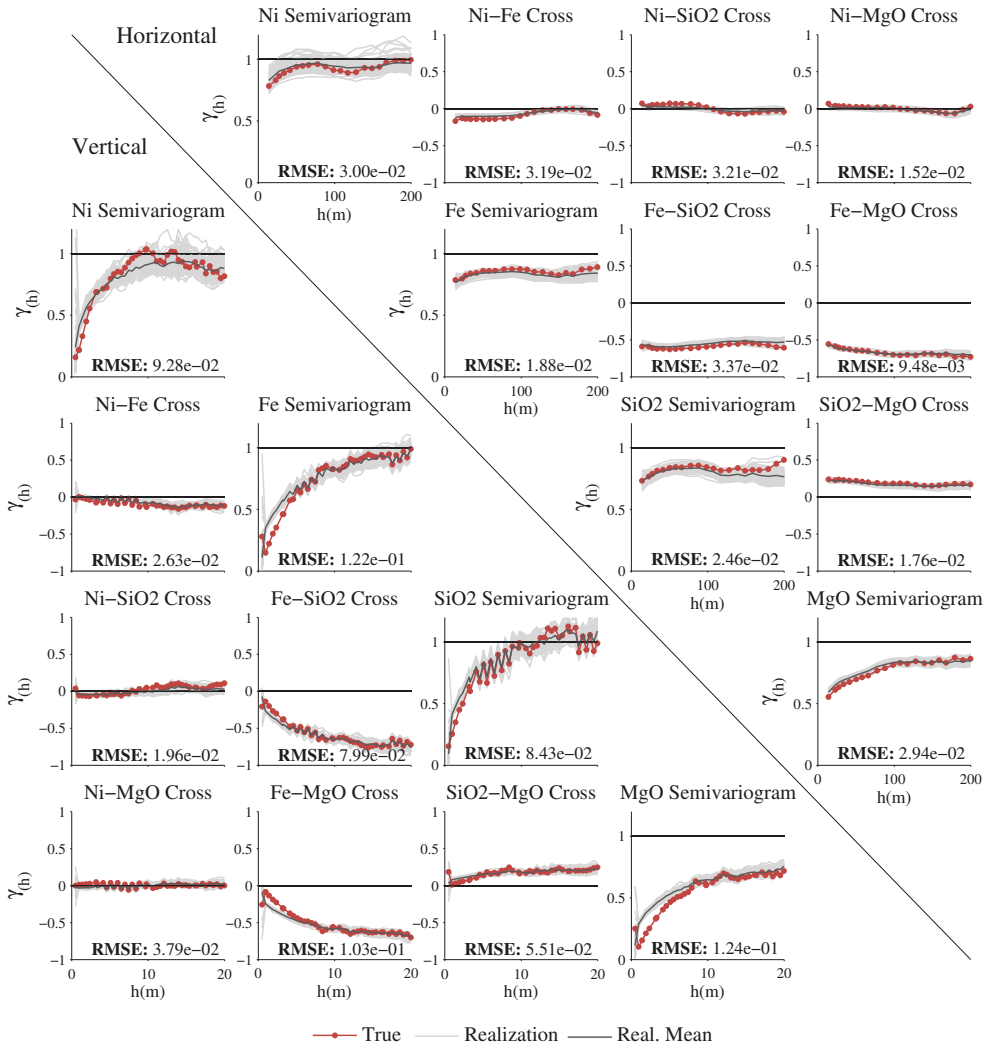


Figure 6.23: Semivariograms and cross-semivariograms of the true values and simulated realizations.

6.4.2 Practical Measures for Semivariogram Reproduction

Simulated results in the previous section displayed excellent reproduction of the univariate, bivariate and horizontal continuity properties. A concern, however, is the vertical continuity. As discussed in Chapter 8, this problem is a primary focus of future work that is not resolved by this thesis.

In this case, simulated realizations in the more poorly conditioned and aggressively destructured horizontal direction are very well reproduced. In a counter-intuitive manner, however, simulated realizations in the well conditioned and gently destructured vertical direction are not well reproduced. The problem is believed to relate to forcing related variables to be independent at $h = 0$ lag, which has a larger impact on the continuity at short scale lag distances. Despite subsetting, non-stationary vertical trends persist within the rocktypes; this degrades vertical continuity in a manner that is unrelated to the PPMT.

To overcome this issue, one could consider the very practical measure of increasing the continuity of the semivariogram models in transformed space. Figure 6.24 displays a slice through the vertical component of the semivariogram models that are used as input to SGSIM. The models from the previous section are constructed with 0.01 nugget effect and three spherical nested structures. In general, they are fit very closely to the vertical variability of the transformed data, though they err on the side of too-continuous when the exact shape cannot be fit. Also displayed in this figure are semivariogram models of inflated continuity, which increase the ranges of the first, second and third nested structures by factors of four, two and one, respectively. The closely fit semivariogram models and resultant geostatistical realizations are referred to as Varg0, while the inflated continuity semivariogram models and resultant geostatistical realizations are referred to as Varg1.

Figure 6.25 displays semivariogram reproduction where the Varg1 models have been used as input to simulation. Relative to the Varg0 results (Figure 6.23), the vertical semivariogram and cross-semivariogram reproduction is improved a great deal. Despite the extremely exaggerated continuity in transformed space, the back-transformation has yielded realizations where only Ni is too continuous (overall) in the vertical direction. Further, at the first vertical lag distance, every variable other than SiO_2 remains too discontinuous.

The improved vertical reproduction comes at the expense of horizontal repro-

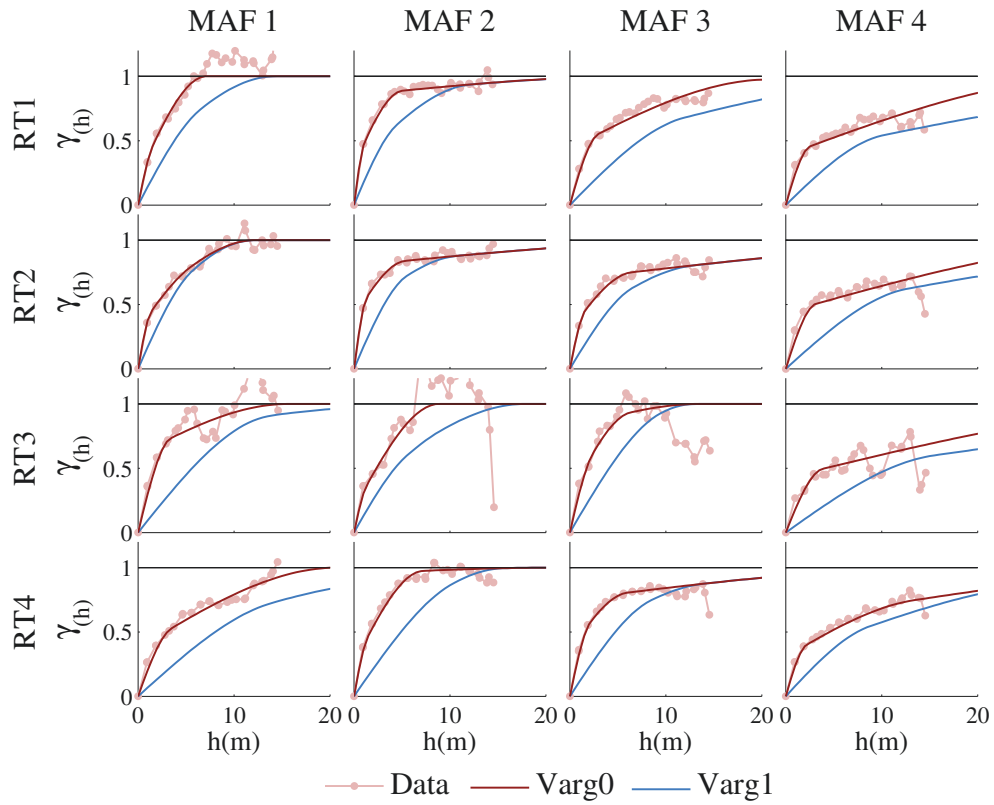


Figure 6.24: Vertical semivariograms of the transformed data, with the original (Varg0) and inflated continuity (Varg1) models.

duction, as the increased continuity in the vertical direction has led to Fe and SiO_2 semivariograms that are slightly too continuous in the horizontal direction (Figure 6.25). The best result will be judged based on the priorities of the setting and practitioner; here, the best result is judged based on local accuracy. Comparing their local accuracy results in Figures 6.21 and 6.26, the increased continuity of Varg1 is seen to improve local accuracy relative to Varg0 for all modeled variables.

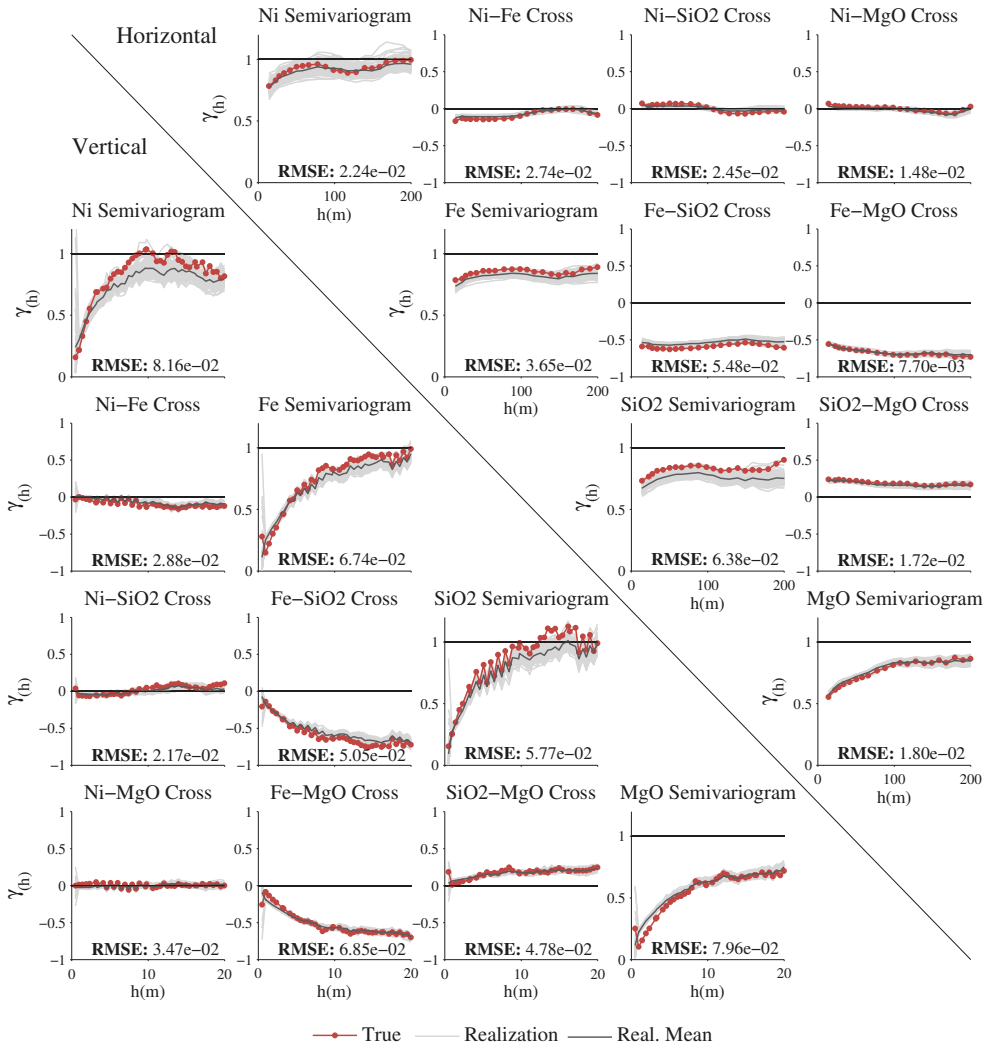


Figure 6.25: Semivariograms and cross-semivariograms of the true values and simulated realizations (Varg1 workflow).

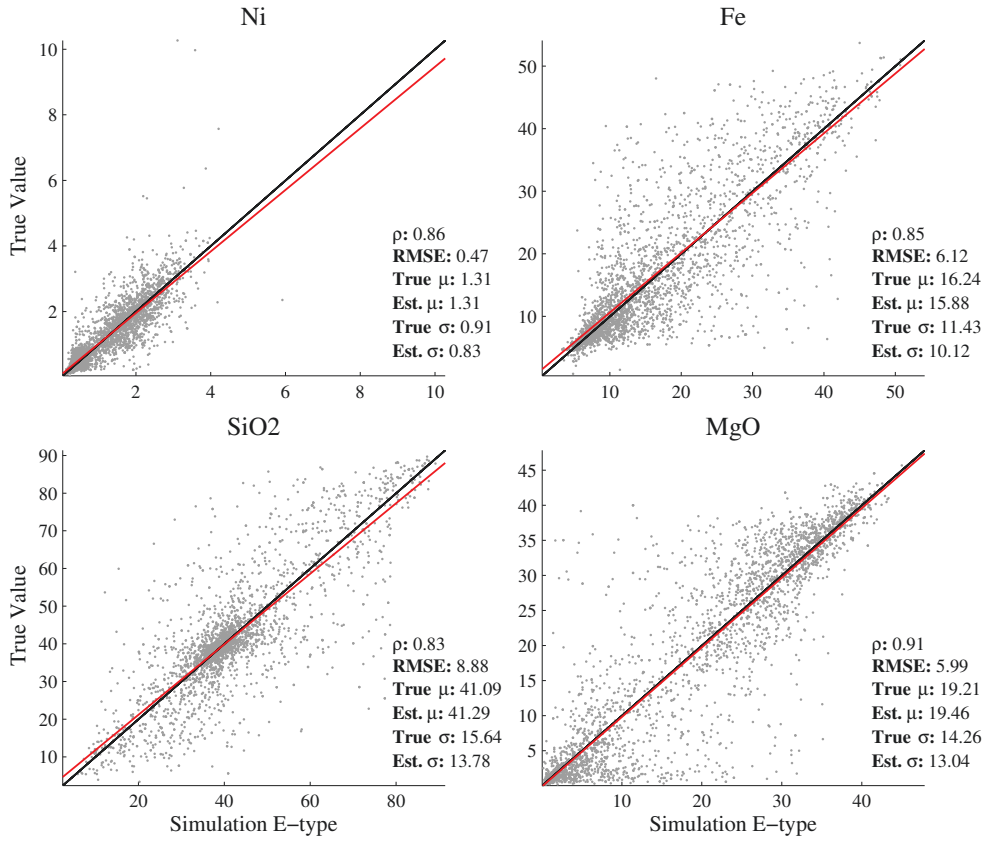


Figure 6.26: Scatterplots and summary statistics that compare the simulated e-type with associated true values (Varg1 workflow).

6.4.3 MAF Comparison

To provide a relative benchmark for judging the PPMT results, modeling of the jackknife locations will now proceed with the following 'MAF' workflow:

i) Normal score transformation as a preprocessor that centers the data and removes outliers.

ii) MAF transformation to decorrelate the variables at 0m and 10m lag distances.

iii) Second normal score to transform the variables to be standard univariate Gaussian.

iv) Independent simulation using SGSIM.

v) Back-transformation of the realizations in the reverse order.

Results from this MAF workflow are compared with the PPMT/MAF workflows that appeared in the previous sections, where the Varg0 and Varg1 semivariogram model approaches are denoted as PPMT/MAF0 and PPMT/MAF1, respectively. The MAF workflow is identical to the PPMT/MAF workflows, except that a normal score transform is used for step (i) rather than the PPMT. The semivariogram modeling methodology follows the PPMT/MAF0 workflow methodology, where the models are fit to the transformed data rather than being overly continuous.

Alternative workflows could be considered such as the cosimulation and stepwise conditional transformation (SCT) approaches that were compared to the PPMT in Section 5.3.6. The cosimulation approach is not considered since it underperformed MAF in that bivariate setting and is not expected to offer significant advantage in this one. The SCT outperformed MAF in the bivariate setting, though it underperformed the PPMT. The data requirements of the SCT and dimensionality of this setting mean that nested workflows must be implemented. The n observations that are available for some rocktypes (Table 6.1) only permit two conditioning variables, increasing the likelihood that multivariate complexity will remain following the SCT transformation.

Univariate results are compared in Table 6.2, summarizing the performance of each workflow in terms of the global mean error (μ Error), global standard deviation error (σ Error), one minus the cross-validation correlation ($1 - \rho$), RMSE of the cross-validation (RMSE), and RMSE of the semivariogram (γ RMSE). To simplify comparison, these statistics have been standardized through dividing by their

maximum (by variable). Following this standardization, one represents the worst result for each variable, while zero would represent a perfect result. The underlying raw values of these statistics may be found and visualized for the PPMT/MAF0 workflow in Figures 6.20(μ and σ), 6.21 (ρ and RMSE) and 6.23 (γ RMSE).

Table 6.2: Standardized univariate performance statistics for the MAF and PPMT/MAF workflows.

Measure	Method	Ni	Fe	SiO2	MgO	Avg.
μ Error	PPMT/MAF0	0.97	0.71	0.80	0.23	**0.68**
	PPMT/MAF1	1.00	1.00	1.00	0.41	0.85
	MAF	0.96	0.43	0.76	1.00	0.79
σ Error	PPMT/MAF0	0.44	0.58	0.34	0.11	**0.37**
	PPMT/MAF1	1.00	1.00	1.00	0.37	0.84
	MAF	0.70	0.24	0.40	1.00	0.59
$1-\rho$	PPMT/MAF0	1.00	1.00	1.00	1.00	1.00
	PPMT/MAF1	0.95	0.94	1.00	0.93	**0.95**
	MAF	0.97	0.94	0.98	0.92	0.96
RMSE	PPMT/MAF0	1.00	1.00	1.00	1.00	1.00
	PPMT/MAF1	0.97	0.97	1.00	0.96	**0.98**
	MAF	0.99	0.97	0.99	0.96	0.98
γ RMSE	PPMT/MAF0	1.00	0.97	0.75	0.93	0.91
	PPMT/MAF1	0.85	0.72	0.84	0.60	**0.75**
	MAF	0.87	1.00	1.00	1.00	0.97

To further summarize the comparison, each statistic is averaged across the variables (Avg. column) before bolding the best result. Workflows that incorporate the PPMT produce the best results for each statistic. PPMT/MAF0 yields the best reproduction of the global statistics (μ Error and σ Error), while PPMT/MAF1 yields the best local accuracy ($1-\rho$ and RMSE) and semivariogram reproduction (γ RMSE).

As spatial variability was the one area of concern for the PPMT results, it may be surprising to see that it outperforms MAF according to the γ RMSE. Semivariogram reproduction of the MAF workflow is displayed in Figure 6.27, where an issue with vertical continuity is present that is similar in nature to the PPMT/MAF0 result. This supports the claim that this issue relates to stationarity problems that arise from vertical trends, as well as a general problem with forcing dependent variables to be independent.

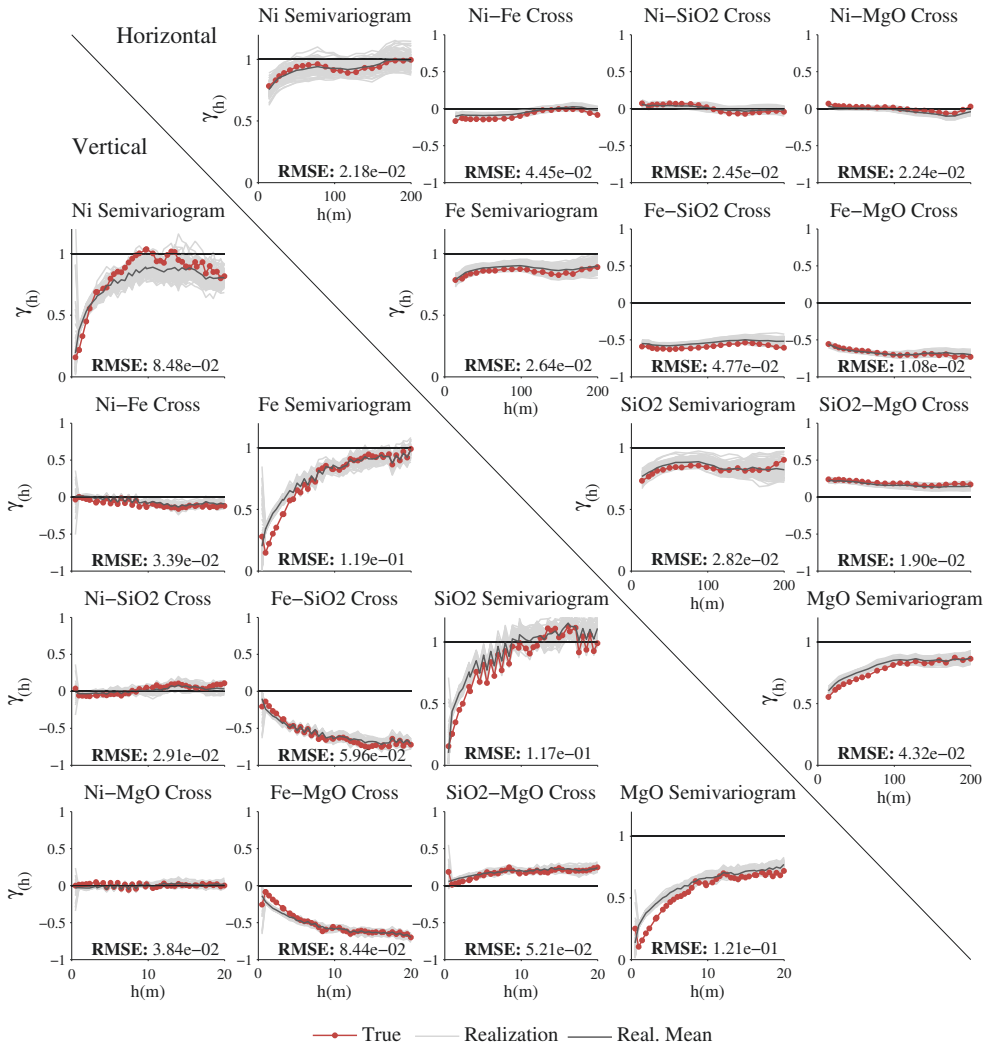


Figure 6.27: Semivariograms and cross-semivariograms of the true values and simulated realizations (MAF workflow).

Bivariate results are compared in Table 6.3, presenting the performance of each method in terms of correlation error (ρ Error), RMSE of bivariate KDE (KDE RMSE) and RMSE of the cross-semivariogram (γ RMSE). Following this standardization, one represents the worst result for each bivariate pair, while zero would represent a perfect result. Note that the pairs are abbreviated by the first letter of each variable in this table due to space constraints.

The raw values of these statistics may be found and visualized for the PPMT/-MAF0 workflow in Figures 6.22 (ρ and KDE RMSE) and 6.23 (γ). To further summarize the comparison, each statistic is averaged across the bivariate pairs (Avg. column) before bolding the best result. As with the univariate performance, the PPMT workflows yield the best reproduction of these bivariate statistics.

Table 6.3: Standardized bivariate performance statistics for the MAF and PPMT/-MAF workflows.

Measure	Method	N-F	N-S	N-M	F-S	F-M	S-M	Avg.
ρ Error	PPMT/MAF0	1.00	0.18	1.00	0.25	0.20	0.68	0.61
	PPMT/MAF1	0.62	0.45	0.80	0.26	0.02	0.44	**0.53**
	MAF	0.03	1.00	0.81	1.00	1.00	1.00	0.71
KDE RMSE	PPMT/MAF0	0.54	0.37	0.39	0.32	0.61	0.32	**0.41**
	PPMT/MAF1	0.50	0.44	0.43	0.39	0.73	0.41	0.44
	MAF	1.00	1.00	1.00	1.00	1.00	1.00	1.00
γ RMSE	PPMT/MAF0	0.74	0.97	0.88	1.00	1.00	1.00	0.90
	PPMT/MAF1	0.72	0.86	0.82	0.93	0.68	0.89	**0.83**
	MAF	1.00	1.00	1.00	0.94	0.85	0.98	0.99

The KDE RMSE statistic in Table 6.3 indicates that the PPMT workflows significantly outperform the MAF workflow in terms of reproducing the complex bivariate density. As displayed in Figure 6.28, however, it should be noted that the MAF workflow does a reasonable job of reproducing complex features that are present in the data. Although it is visually worse than the PPMT/MAF reproduction (Figure 6.22), this result emphasizes the importance of stationarity decisions. Rocktype subsetting has isolated distinct multivariate populations of the data. These populations account for a substantial portion of the multivariate complexity in the overall system. Following geostatistical modeling of the properties within each rocktype, this complexity is restored by the recombination of rocktypes. In other words, the PPMT improves reproduction of the complex features that are present within each rocktype. Substantial complexity also exists between rocktypes, however, allowing for the MAF workflow to exhibit reasonable reproduction of the overall complexity.

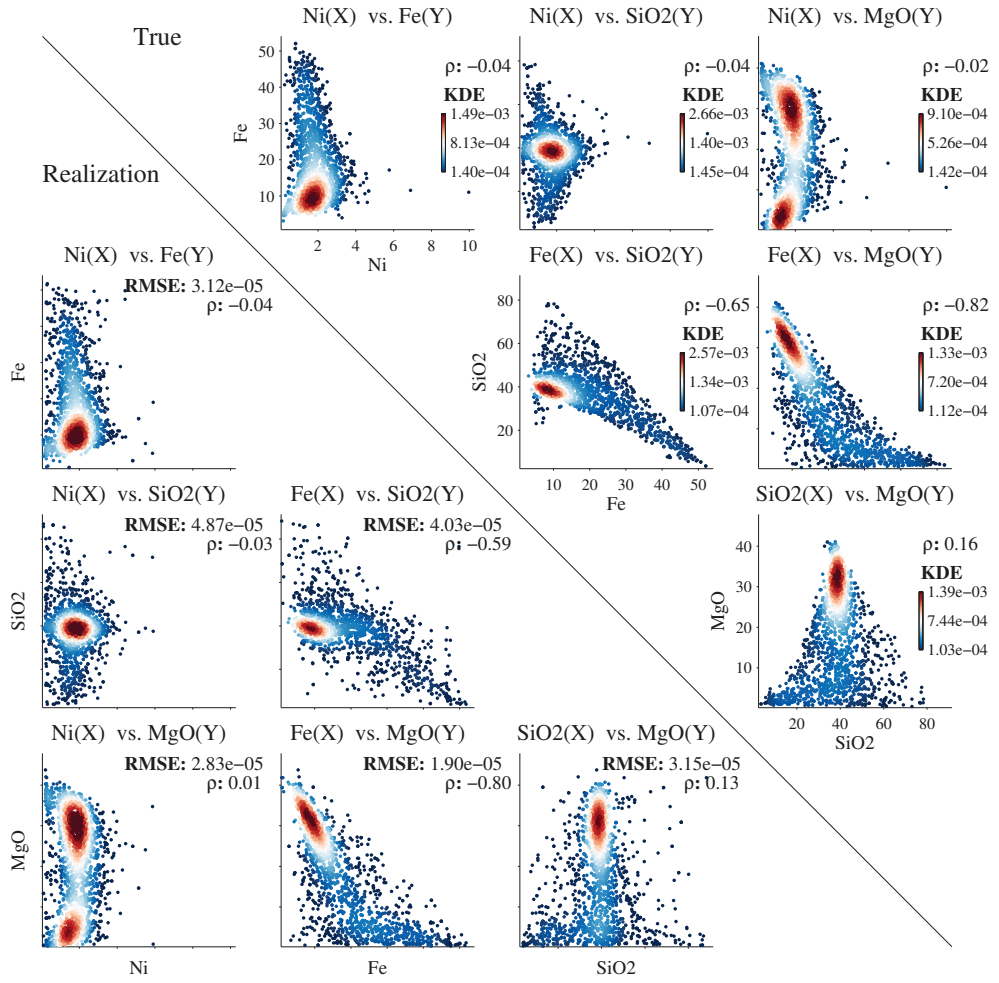


Figure 6.28: KDE scatterplots of a realization, with the true values shown for comparison (MAF workflow).

6.5 Impact on Process Performance

The previous section establishes that incorporating the PPMT in geostatistical modeling improves characterization of the univariate, bivariate, and spatial properties of the Ni laterite deposit. The value that this improved characterization offers in terms of technical decision making and process performance will now be measured.

Unfortunately, information such as the Barro Alto plant recovery functions and blending methodology are not available to this thesis. As a result, value is measured in a conceptual manner following methodology that is available in Neufeld et al. (2008) to be as realistic as possible. Section 6.1 explains that mined ore is stockpiled and stacked to reduce variability and maintain key properties. Feed to

the furnace should have an average Ni grade above 1.5%, Fe grade below 18.5%, and an SMR ratio below 1.75. Observing the critical thresholds of the furnace feed, consider categorizing each simulated location, $\mathbf{u}_\alpha$, as one of $N_O = 8$ classifications, $O_i(\mathbf{u}_\alpha), i = 1, ..., N_O$, according to Table 6.4. This classification is used for directing the material to appropriate stockpiles, and is tracked for directing subsequent stacking to achieve the correct furnace feed blend.

Table 6.4: Table of ore types that are used for blending and stockpiling.

O_i	Group	SMR	Description
1	**Low Fe Ore**	High	Ni $\geq$ 1.5%, SMR $\geq$ 1.75, Fe $<$ 18.5%
2		Low	Ni $\geq$ 1.5%, SMR $<$ 1.75, Fe $<$ 18%
3	**High Fe Ore**	High	Ni $\geq$ 1.5%, SMR $\geq$ 1.75, Fe $\geq$ 18.5%
4		Low	Ni $\geq$1.5%, SMR $<$ 1.75, Fe $\geq$ 18%
5	**Low Fe Waste**	High	Ni $<$ 1.5%, SMR $\geq$ 1.75, Fe $<$ 18.5%
6		Low	Ni $<$ 1.5%, SMR $<$ 1.75, Fe $<$ 18%
7	**High Fe Waste**	High	Ni $<$ 1.5%, SMR $\geq$ 1.75, Fe $\geq$ 18.5%
8		Low	Ni $<$ 1.5%, SMR $<$ 1.75, Fe $\geq$ 18%

Varying economic cost is associated with the misclassification of each ore type. Table 6.5 presents the economic cost of misclassification, C_{ij}, resulting from a classification of $O_i(\mathbf{u})$ when the true ore type is $O_j(\mathbf{u})$. Each C_{ij} is calculated as the summation of:

i) SMR misclassification (0.4), where SMR is predicted to be ≥ 1.75 but the true value is < 1.75 (or vice versa). SMR is assigned a relatively high cost since it has a critical impact on the plant process. Furnace feed that is well above the SMR target will damage the furnace lining. Conversely, furnace feed that is well below the SMR target effectively wastes low SMR material that could have been used more effectively for blending high SMR material in the future.

ii) Fe misclassification (0.2), where Fe is predicted to be $\geq 18.5\%$ but the true value is $< 18.5\%$ (or vice versa). Fe is assigned a lower cost than SMR misclassification since it has less severe consequences overall. Furnace feed that is well above the Fe target will lead to poor recoveries. Conversely, furnace feed that is well below the Fe target effectively wastes low Fe material that could have been used more effectively for blending high Fe material in the future.

iii) Ni misclassification (0.2), where Ni was predicted to be $\geq 1.5\%$ but the true

value is $< 1.5\%$ (or vice versa). Ni is assigned the same cost as Fe misclassification since they both relate to economic recovery.

Table 6.5: Cost associated with misclassification, which differs based on the predicted and true ore type.

			Predicted Type							
			Low Fe Ore		High Fe Ore		Low Fe Waste		High Fe Waste	
		SMR	High	Low	High	Low	High	Low	High	Low
True Type	Low Fe Ore	High	0	0.4	0.2	0.6	0.4	0.8	0.6	1
		Low	0.4	0	0.6	0.2	0.8	0.4	1	0.6
	High Fe Ore	High	0.2	0.6	0	0.4	0.6	1	0.4	0.8
		Low	0.6	0.2	0.4	0	1	0.6	0.8	0.4
	Low Fe Waste	High	0.4	0.8	0.6	1	0	0.4	0.2	0.6
		Low	0.8	0.4	1	0.6	0.4	0	0.6	0.2
	High Fe Waste	High	0.6	1	0.4	0.8	0.2	0.6	0	0
		Low	1	0.6	0.8	0.4	0.6	0.2	0	0

As demonstrated for strategic mine planning in Dimitrakopoulos (2011) and conceptualized in Figure 1.1, basing resource management decisions on multiple geostatistical realizations allows for the mitigation of risk that is associated with geologic uncertainty. With this in mind, the final ore type classification, $O_i(\mathbf{u}_\alpha)$, minimizes the total economic loss that is incurred based on the $L = 100$ realizations of ore types:

$$\underset{i\in(1,N_O)}{\arg\min}\left[\sum_{l=1}^{L}\sum_{j=1}^{N_O} C_{ij} \bullet \Delta_l(\mathbf{u}_\alpha; O_j)\right] \tag{6.1}$$

where $\Delta_l(\mathbf{u}_\alpha; O_j)$ is the binary indicator:

$$\Delta_l(\mathbf{u}_\alpha; O_j) = \begin{cases} 1, \text{ if ore type } O_j \text{ is present at } \mathbf{u}_\alpha \text{ for the } l^{th} \text{ realization} \\ 0, \text{ if not} \end{cases} \tag{6.2}$$

Using this approach, ore is classified in a manner that integrates uncertainty to mitigate risk. Each classification, $O_i(\mathbf{u}_\alpha)$, may then be evaluated using the true ore type, $O_j(\mathbf{u}_\alpha)$, to determine the economic cost that occurs if $\mathbf{u}_\alpha$ is misclassified. The performance of each workflow is summarized based on the total economic loss:

$$\text{Loss} = \sum_{\alpha=1}^{N}\sum_{i=1}^{N_O}\sum_{j=1}^{N_O} c_{ij} \bullet \widehat{\Delta}(\mathbf{u}_\alpha; O_i) \bullet \Delta(\mathbf{u}_\alpha; O_i) \tag{6.3}$$

where $\widehat{\Delta}(\mathbf{u}_\alpha; O_i)$ is the binary indicator of the predicted ore type (Equation 6.4) and $\Delta(\mathbf{u}_\alpha; O_j)$ is the binary indicator of the true ore type (Equation 6.5).

$$\widehat{\Delta}(\mathbf{u}_\alpha; O_i) = \begin{cases} 1, \text{ if ore type } O_i \text{ is predicted at } \mathbf{u}_\alpha \\ 0, \text{ if not} \end{cases} \tag{6.4}$$

$$\Delta(\mathbf{u}_\alpha; O_j) = \begin{cases} 1, \text{ if ore type } O_j \text{ is truly present at } \mathbf{u}_\alpha \\ 0, \text{ if not} \end{cases} \tag{6.5}$$

This economic loss metric is standardized in Table 6.6 so that one represents the worst result (largest economic loss) and zero represents a perfect result (no economic loss). Consider that this metric is expected to reduce with increasing accuracy and precision for the distribution of ore type uncertainty. Given that PPMT/MAF1 workflow provided the best multivariate and spatial characterization of the Ni laterite deposit, it is unsurprising to see that it yields the lowest economic loss.

Table 6.6: Standardized process performance loss, as well as the a and b statistics from Figure 6.29

Method	**Loss**	a	$1-b$
PPMT/MAF0	1.000	1.000	0.999
PPMT/MAF1	**0.974**	**0.716**	**0.936**
MAF	0.999	0.716	1.000

Since the metric rewards an accurate and precise distribution of uncertainty, these properties are evaluated in a more conventional manner. Figure 6.29 displays the overall accuracy of the ore type uncertainty, where the predicted probability of each ore type occurring is plotted against the true fraction of its occurrence (Deutsch and Deutsch, 2012a). Specifics of the plot construction are described using an arbitrary probability interval (0.5) and ore type (O_1):

i) The simulated locations that have a 0.45 to 0.55 probability of O_1 are determined, before calculating the true fraction of occurrence for O_1 across those locations.

ii) Repeat step (i) for the remaining $N_O = 8$ ore types.

iii) Average the true fraction of occurrence across the N_O ore types and plot against the 0.5 probability interval.

The coloring in Figure 6.29 summarizes how frequently the various certainty intervals are observed. The fraction of simulated locations that fall into each probability interval (np/nt) underly the coloring, where the scale is set based on the

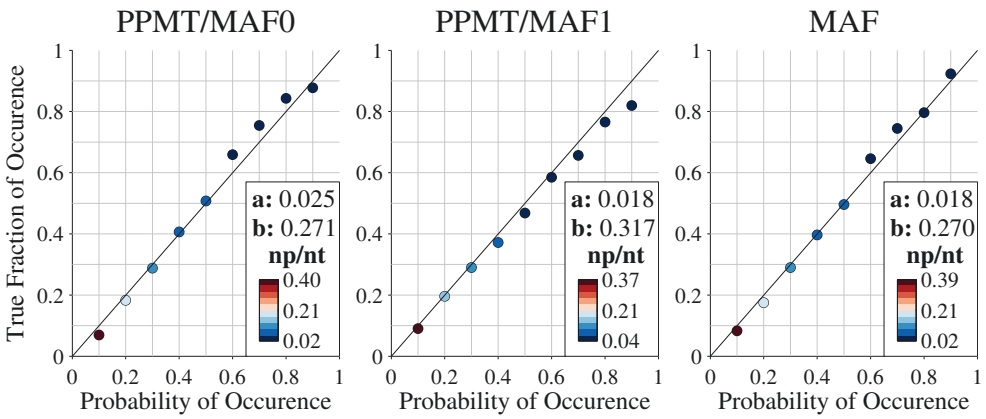


Figure 6.29: Accuracy plots of the ore type distributions resulting from each workflow.

minimum and maximum of this fraction. Using the Varg0 results as an example, the average probability of occurrence is 0.1 at 40% of the locations, whereas the average probability of occurrence is 0.9 at only 2% of the locations.

The displayed a and b statistics in Figure 6.29 summarize the accuracy and precision of each workflow's uncertainty. The a statistic is calculated as the average difference between the predicted probability of occurrence and true frequency of occurrence. This averaging is weighted by the np/nt that underlies each probability interval. The a statistic can be visualized as the average difference between the points and the forty five degree line in Figure 6.29, weighted by the associated np/nt color. As described in Deutsch and Deutsch (2012a), the b statistic is calculated as the average difference between the predicted probability of a category when it actually occurs and the predicted probability when is does not actually occur. The a and b statistics are standardized in Table 6.6 in the usual manner to aid in comparison. Observe that the PPMT/MAF1 workflow yields the best overall result, in terms of the conceptual process performance (Loss), the uncertainty accuracy (a) and uncertainty precision (b).

6.6 Summary

The Barro Alto Ni laterite data was chosen for this case study since it presents a large challenge for the PPMT and multivariate geostatistical modeling in general. The variables have widely varying spatial continuity, while exhibiting very complex multivariate features between them.

Relative to the bivariate example in Chapter 5, additional iterations (> 50) were required to transform the variables to an uncorrelated multiGaussian distribution. This is attributed to the additional dimensions and complexity of the Ni laterite variables. The multivariate configuration and spatial continuity of the variables was largely preserved by the transformation. The exception, however, was the mixing and spatial destructuring of Fe. A strong dependence exists between Fe and SiO_2, meaning that one of those two variables must become destructured to make them independent.

After applying a subsequent MAF transformation to decorrelate the variables at $h > 0$, the resultant data is appropriate for independent Gaussian simulation. Simulated realizations are back-transformed to original space, where they were found to effectively characterize the Ni laterite deposit. A conventional MAF workflow that does not incorporate the PPMT was used as a comparative benchmark of the results. This comparison established that utilizing the PPMT leads to improved reproduction of univariate, multivariate and spatial properties. A conceptual loss function that is based on priorities of the Barro Alto project was used to measure the value of this improved characterization in terms of improved decision making.

The one concern with the presented results is the loss of continuity that was observed in the vertical semivariograms of simulated realizations. This problem was not unique to the PPMT, as the MAF workflow also failed to reproduce the vertical semivariograms. The issue is primarily attributed to non-stationarity vertical trends that exist despite stationarity related rocktype subsetting.

Chapter 7

Nickel Laterite Case Study: Data Imputation

The following chapter uses Anglo American's Barro Alto Ni laterite mine to study the performance and value of data imputation in a complex geologic setting. This is a direct extension of Chapter 6; all of the previously described information such the Barro Alto background and data inventory continues to apply and will not be repeated.

Further, the jackknife locations of the Ni laterite dataset will continue to be modeled using the described PPMT/MAF1 geostatistical workflow. The previous chapter established the accuracy that this workflow is best. That data will now be decimated using a missing at random (MAR) mechanism, resulting in heterotopic and homotopic observations. Consequently, there are three geostatistical modeling options:

i) Use data exclusion (DE), where heterotopic observations are excluded from modeling to facilitate the multivariate transformations.

ii) Use single imputation (SI), where missing values are imputed with a single value. The resultant homotopic data is used for modeling so that sampled values are not excluded.

iii) Use multiple imputation (MI), where missing values are imputed with multiple realizations that sample their uncertainty. The resultant homotopic data realizations are used for modeling so that sampled values are not excluded and uncertainty of the imputed values is incorporated.

Geostatistical modeling is performed at the jackknife locations to measure the improvement that MI (option iii) offers over DE and SI (options i and ii) in terms

of local accuracy and resource characterization. Option (iii) is repeated using the four imputation methods from Chapter 3 to compare each method's performance with real Ni laterite data. This also provides insight into how improved imputation results translate to improved geostatistical modeling results. The conceptual loss metric from the previous chapter is used to measure the relative impact that the missing data schemes (three options) and imputation results (four methods) have on resource management decisions of the Barro Alto project.

The majority of figure and table formats in this chapter, as well as the statistics within them have been introduced in Chapters 3, 5 and 6. Only new formats and statistics will be explained in detail.

7.1 Missing Data Mechanism

The Ni laterite data set is composed of 11,013 homotopic observations, which is decimated with a MAR mechanism to facilitate jackknife validation of the imputation process. Consider that it is very common for mines to spend additional money on sampling and lab testing of observations in higher grade areas. With this in mind, the data removal scheme is described as:

i) Ni is the resource and therefore is not missing from any observations.

ii) Fe is missing in $\sim 2/3$ of the observations from low grade regions (Ni less than the median of its distribution), while only missing in $\sim 1/3$ of the observations from high grade regions (Ni greater than or equal to the median of its distribution).

iii) MgO and SiO_2 are missing in $\sim 1/3$ of the observations from low grade regions, while only missing in $\sim 1/5$ of the observations from high grade regions. They are always missing together as SMR motivates their sampling; both or none would be missing.

Consequently, the resultant heterotopic data will preferentially sample Fe, SiO_2 and MgO from high grade Ni regions. While no Ni values are missing, the variable is included in the imputation model for several reasons: i) to prevent a missing not at random (MNAR) mechanism that would lead to bias results, since the missingness relates to Ni grade, ii) to insure that the imputed variables reproduce their relationship with Ni, and iii) to increase the imputation accuracy through additional condi-

tioning information. The number of missing values, N_{mis} for each variable is shown in Table 7.1, which leads to 5,098 homotopic observations and 5,915 heterotopic observations. Figure 7.1 displays the locations of the heterotopic and homotopic observations, which are shown to be relatively dispersed across the dataset from a spatial perspective.

Table 7.1: The number of missing values (N_{mis}) and their associated fraction of the total 11,013 observations (N_{mis}/N_{tot}).

Variable	**High Grade**		**Low Grade**		**Total**	
	Nmis	***Nmis/Ntot***	***Nmis***	***Nmis/Ntot***	***Nmis***	***Nmis/Ntot***
Ni	0	0.00	0	0.00	0	0.00
Fe	1,694	0.15	2,814	0.26	4,508	0.41
MgO	1,075	0.10	1,712	0.16	2,787	0.25
SiO2	1,075	0.10	1,712	0.16	2,787	0.25

CDFs in Figure 7.2 reveal that the MAR mechanism has yielded differences between the sampled and missing distributions. The differences are relatively small, however, meaning that DE or SI is not expected to add significant bias to the dataset and subsequent geostatistical modeling from a univariate perspective.

Figure 7.3 displays the bivariate relationships of the missing and sampled values in the upper and lower triangles, respectively. As with the CDFs, significant bias is not present in the bivariate distributions in spite of the MAR mechanism. As a result, DE or SI is also not expected to add significant bias from a bivariate perspective.

The complex features that are present in Figure 7.3 are only expected to be reproduced by the non-parametric merged (NPM) imputation method. Note, however, that all of the rocktypes are present in this figure. As with geostatistical modeling in the previous section, the data is imputed by rocktype. Rocktype subsetting and recombination has been shown to remove and reintroduce a significant proportion of the Ni laterite complexity. This permitted the MAF workflow to reasonably reproduce complex bivariate features (Figure 6.28); it may also allow for the imputation methods that make multiGaussian assumptions to reasonably reproduce the bivariate distributions in Figure 7.3.

Figure 7.4 displays spatial continuity of the missing and sampled values in the upper and lower triangles, respectively. The 'double-diagonal' locations display the semivariograms, while cross-semivariograms appear in the off-diagonal locations.

Unlike the univariate and bivariate distributions, the missing and sampled data have very different spatial continuity. In particular, the semivariograms are less continuous for the sampled data than they are for the missing data in both the horizontal and vertical directions. Consequently, DE or SI is expected to bias the continuity of the data and subsequent geostatistical models.

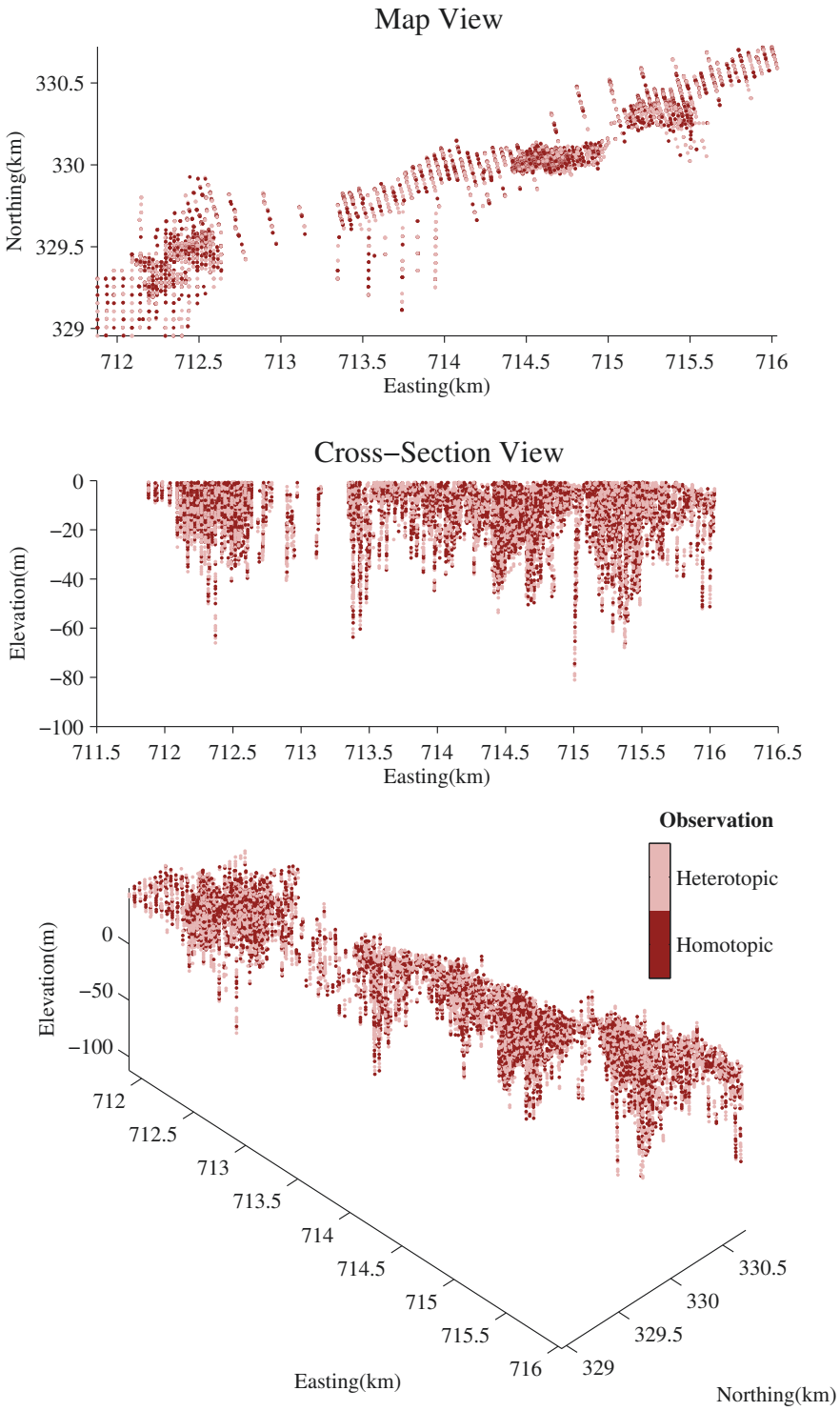


Figure 7.1: Locations of heterotopic and homotopic data observations.

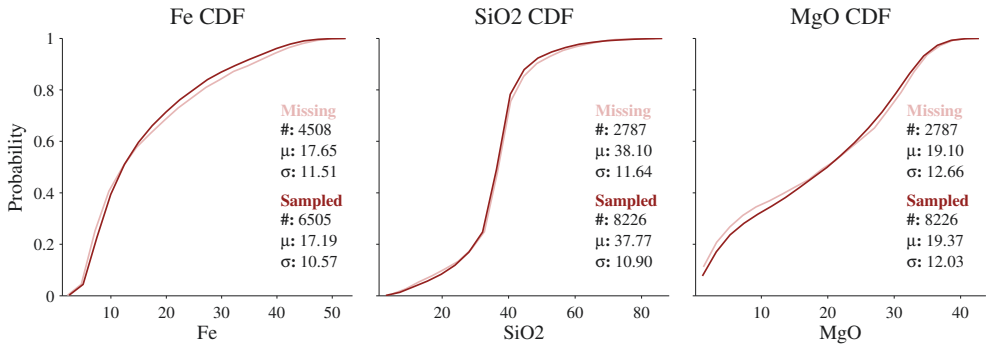


Figure 7.2: CDFs of the missing and sampled data.

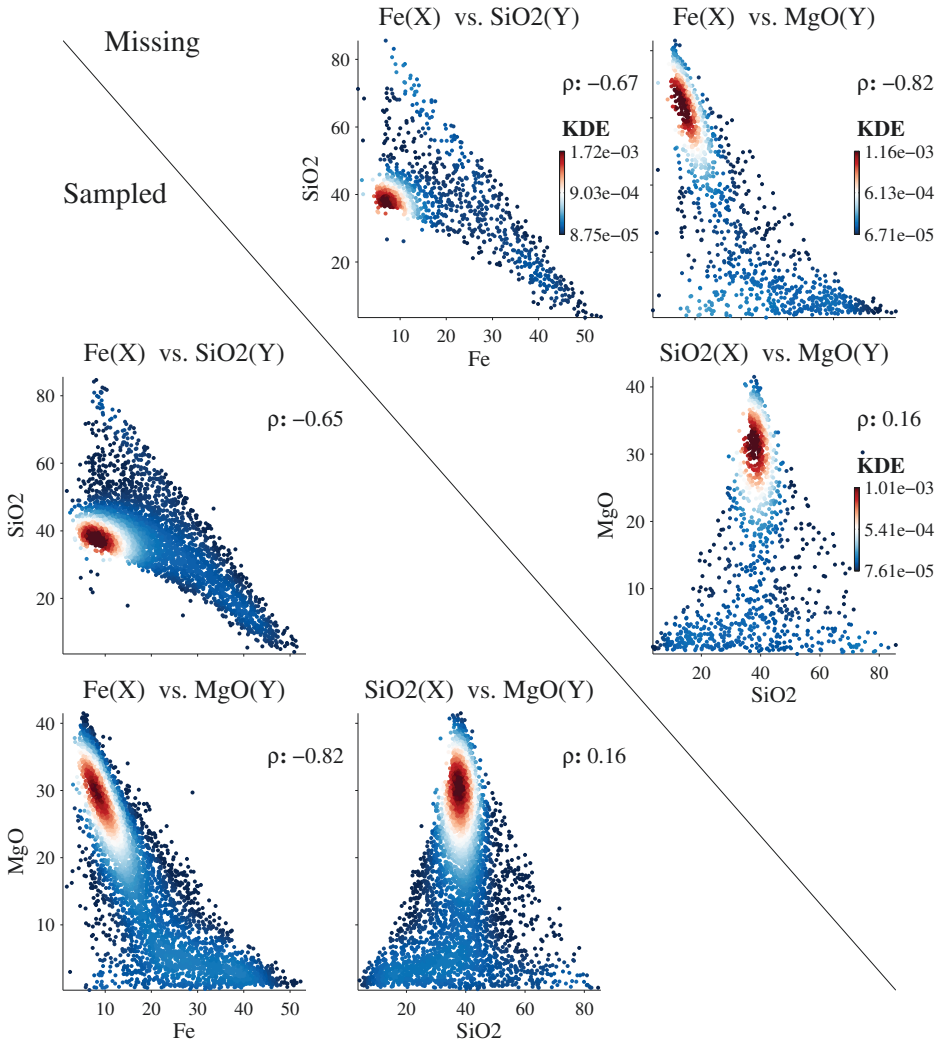


Figure 7.3: KDE scatterplots of the missing and sampled data.

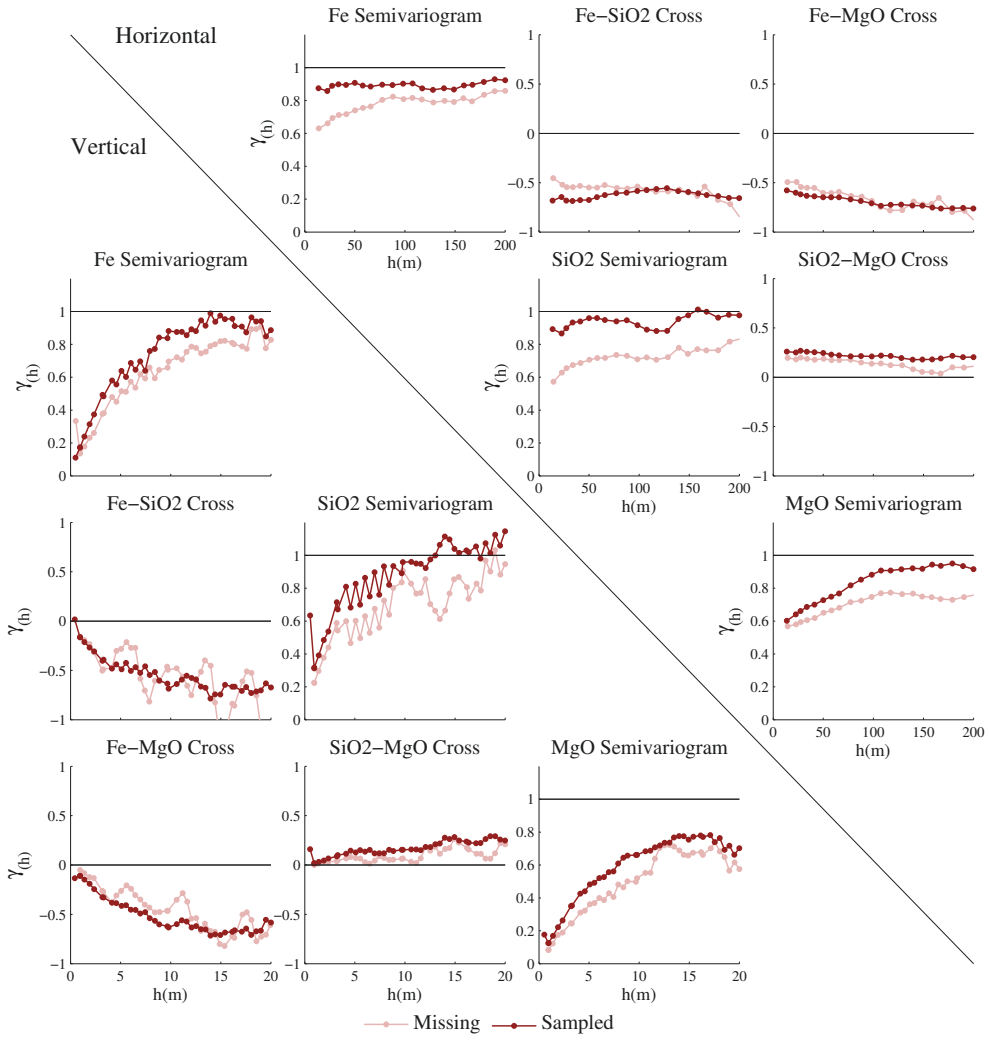


Figure 7.4: Semivariograms and cross-semivariograms of the missing and sampled data.

7.2 Imputation Results

After subsetting the heterotopic data by rocktype, the four MI methods from Section 3.3 are used to impute one hundred realizations of the missing values. After recombining the rocktypes, the performance of each imputation method may be judged by comparing the realizations against the true removed values.

7.2.1 Univariate Reproduction

Reproduction of the global univariate distributions is shown in Figure 7.5. Overall reproduction of the missing CDFs is excellent, showing that values can be imputed by MI without introducing global bias. CDF reproduction is summarized based on the displayed mean and standard deviation errors in Figure 7.5 to aid in the comparison of each imputation method. These statistics appear in Table 7.2, where they are labeled as μ and σ Error, respectively. As in earlier chapters, the statistics are standardized so that one represents the worst result for each variable; zero would represent a perfect result. To further summarize the comparison, each statistic is averaged across the variables (Avg. column) before bolding the best result. Keeping in mind that each method yielded excellent CDF reproduction, the merged and NPM methods produce the best overall result in terms of μ and σ, respectively. The primary method yields the worst result according to these statistics, as is visually apparent in Figure 7.5. This indicates that incorporating colocated information of the heterotopic observations reduces global bias in the imputed results.

Local accuracy is examined in Figure 7.6, where the e-type mean of the imputed realizations is plotted against the associated true value. Performance is summarized by the correlation $(1-\rho)$ and RMSE statistics in Figure 7.6, which are standardized in Table 7.2. According to these statistics, a consistent gradient of improvement is observed when proceeding across the imputation methods. Only considering spatially correlated information (primary) yields worse accuracy than only considering colocated information (secondary), though methods incorporating both sources of information perform best (merged and NPM). Looking at the scatter characteristics in Figure 7.6, observe that the NPM method yields the most normally distributed error. This reflects that the NPM does not assume that the data is multiGaussian. Compare this to the complex distribution of error for the secondary method. The secondary method revolves around a multiGaussian assumption, leading to more

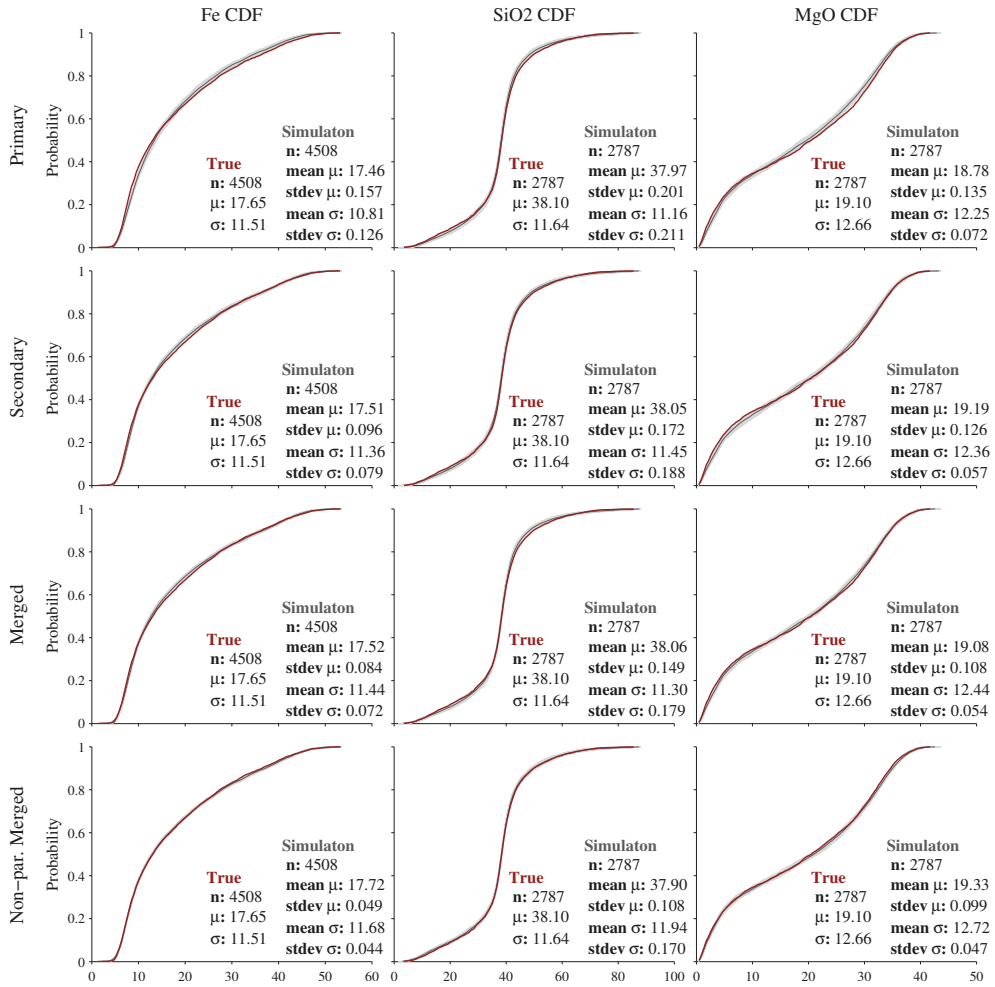


Figure 7.5: CDFs of the imputed realizations with the true CDFs overlain for comparison.

frequent occurrences of 'very bad' estimates than the other methods.

Reproduction of horizontal and vertical continuity is examined in Figures 7.7 and 7.8, respectively. Semivariograms of the imputed realizations are overlain with that of the removed true values. The RMSE statistic in these figures are averaged and standardized in Table 7.2.

Recall from Figure 7.4 that large differences exist between the continuity of the missing and sampled data. It should be noted that the imputation methods that incorporate spatial information had semivariogram models that were closely fit to semivariograms of the sampled data. Although this case study is aware of the missing properties, this would not occur in reality and was not incorporated

Table 7.2: Standardized univariate performance statistics for each imputation method.

Measure	Method	Fe	SiO_2	MgO	Avg.
μ Error	Primary	1.00	0.66	1.00	0.89
	Secondary	0.75	0.26	0.28	0.43
	Merged	0.70	0.20	0.06	**0.32**
	NPM	0.36	1.00	0.75	0.70
σ Error	Primary	1.00	1.00	1.00	1.00
	Secondary	0.22	0.40	0.71	0.44
	Merged	0.10	0.71	0.53	0.45
	NPM	0.23	0.62	0.16	**0.34**
$1 - \rho$	Primary	1.00	0.83	0.86	0.90
	Secondary	0.51	1.00	1.00	0.84
	Merged	0.31	0.60	0.59	0.50
	NPM	0.26	0.48	0.50	**0.41**
RMSE	Primary	1.00	0.93	0.94	0.96
	Secondary	0.74	1.00	1.00	0.91
	Merged	0.59	0.82	0.78	0.73
	NPM	0.54	0.76	0.72	**0.67**
γ RMSE	Primary	0.81	0.45	0.62	0.63
	Secondary	1.00	1.00	1.00	1.00
	Merged	0.27	0.41	0.47	0.38
	NPM	0.19	0.26	0.31	**0.25**

into the imputation. Despite using these semivariogram models, Figures 7.7 and 7.8 demonstrate that the differing continuity of the missing values is generally well reproduced. As expected, the secondary method yields the worst reproduction of continuity since it does not incorporate spatial information.

Figure 7.6 demonstrated that the NPM method yields the greatest local accuracy. It is this increased accuracy that is attributed to its superior semivariogram reproduction, as values that are nearer to their truth will indirectly improve the spatial variability. Interestingly, the one concern with the semivariogram reproduction is at short scale vertical lags. This is reminiscent of the issue that was seen with geostatistical realizations in the previous section.

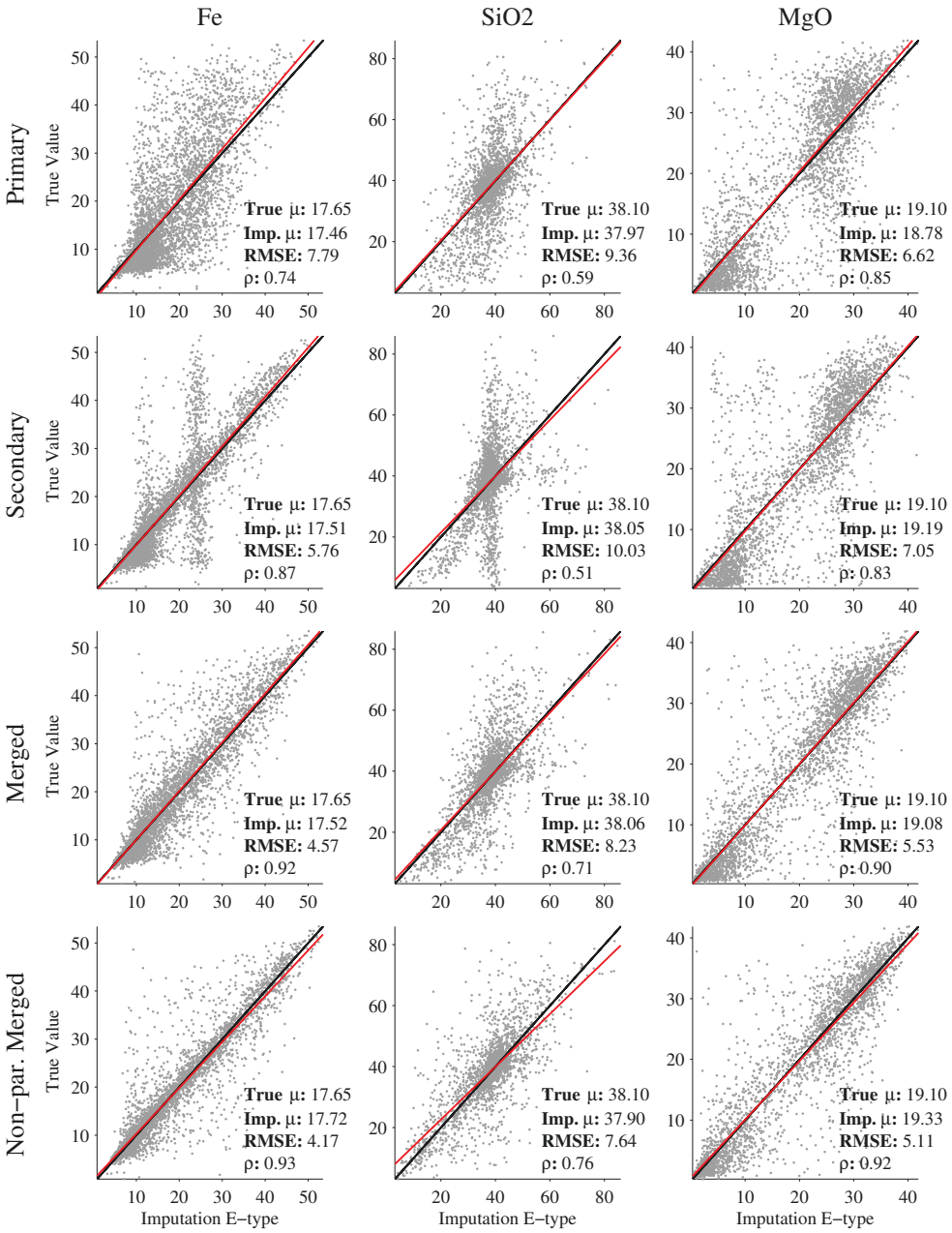


Figure 7.6: Scatterplots and summary statistics that compare the imputed e-type with associated true values.

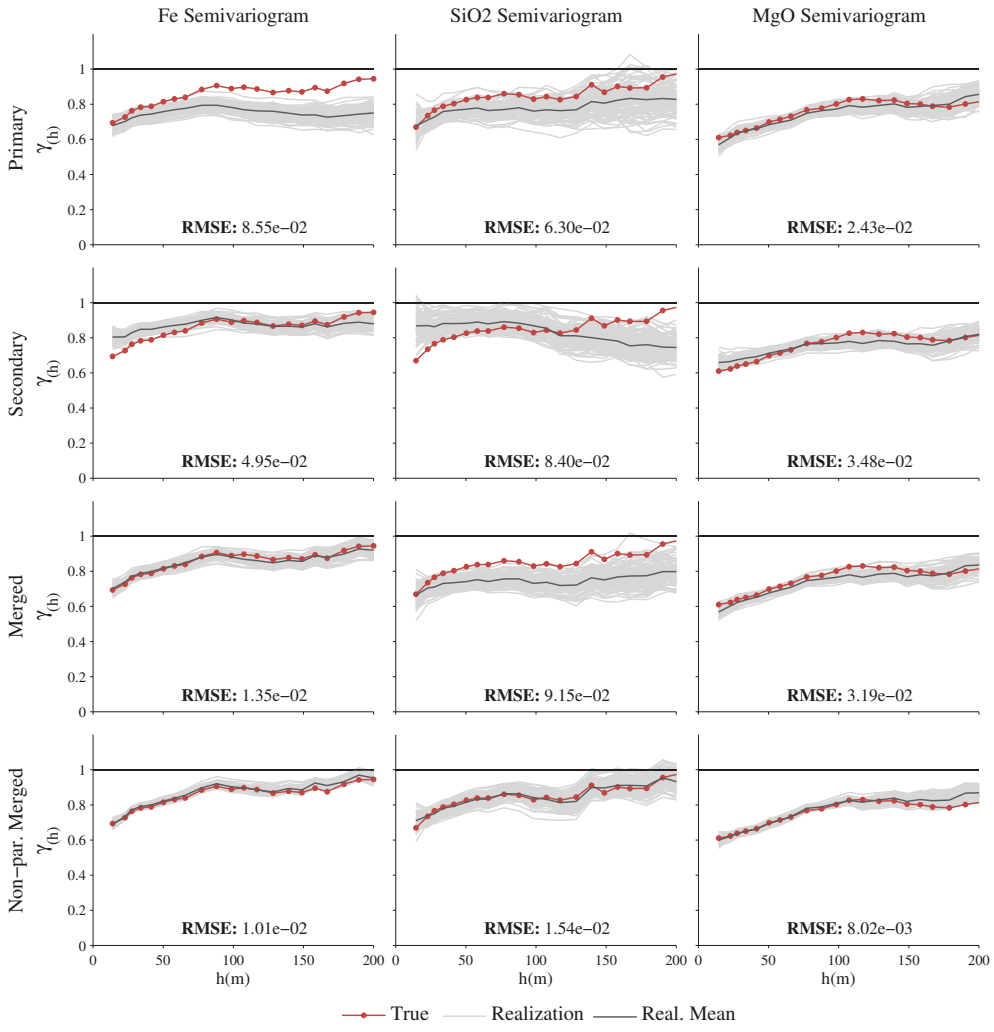


Figure 7.7: Horizontal semivariograms of the true values and imputed realizations.

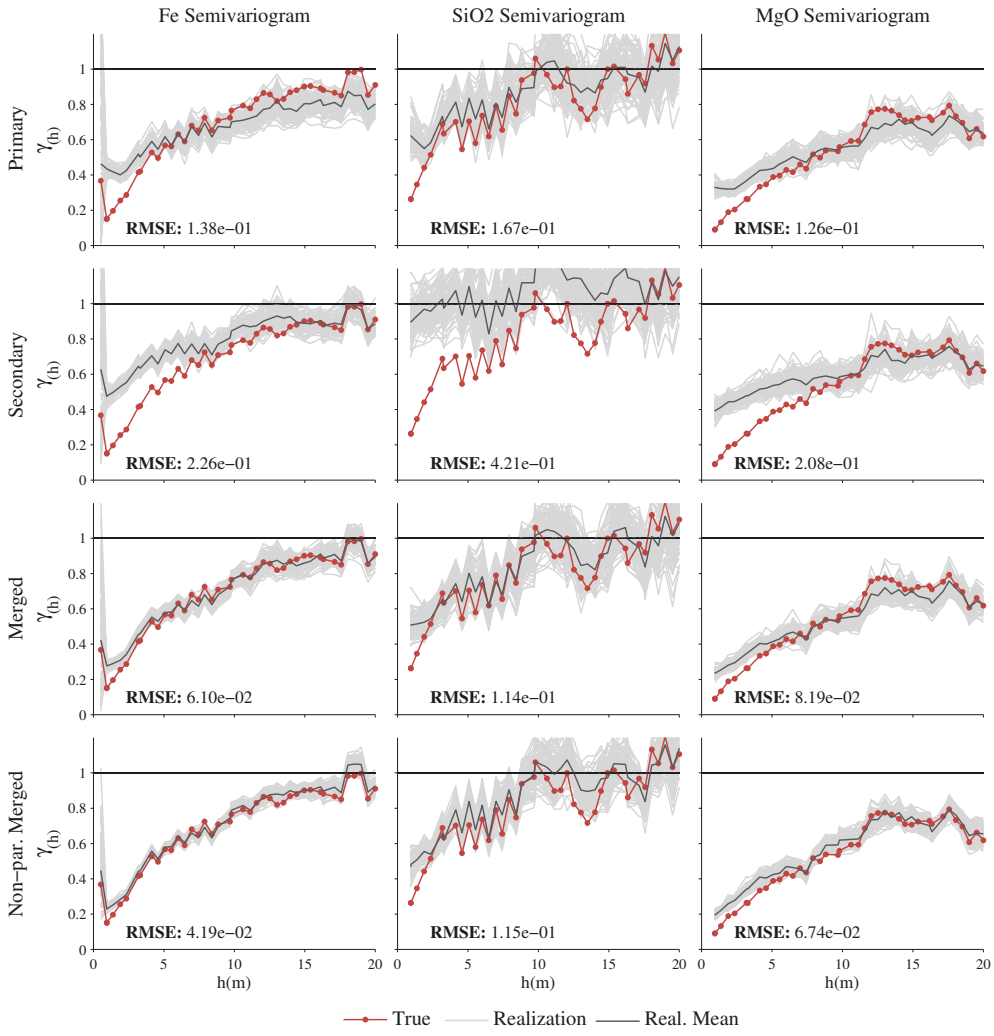


Figure 7.8: Vertical semivariograms of the true values and imputed realizations.

7.2.2 Multivariate Reproduction

The reproduction of bivariate densities is examined in Figure 7.9. The performance of each method is summarized using standardized KDE RMSE and correlation error (ρ Error) in Table 7.3.

Table 7.3: Standardized multivariate performance statistics for each imputation method.

Measure	Method	Fe-SiO_2	Fe-MgO	SiO_2-MgO	Avg.
ρ Error	Primary	1.00	1.00	1.00	1.00
	Secondary	0.32	0.34	0.64	0.43
	Merged	0.22	0.24	0.16	0.21
	NPM	0.00	0.02	0.12	**0.05**
KDE RMSE	Primary	1.00	1.00	1.00	1.00
	Secondary	0.78	0.46	0.96	0.73
	Merged	0.65	0.41	0.80	0.62
	NPM	0.30	0.25	0.28	**0.28**
γ RMSE	Primary	1.00	1.00	1.00	1.00
	Secondary	0.42	0.65	0.87	0.65
	Merged	0.42	0.44	0.48	0.45
	NPM	0.19	0.29	0.32	**0.27**

All statistics in Table 7.3 are standardized so that one represents the worst result for that bivariate pair; zero would represent a perfect result. According to these metrics and visual inspection, the NPM method yields the best result. The improved local accuracy and spatial variability of the NPM method is attributed to its correct integration of multivariate information. As noted, however, imputation by rocktype subdivision has allowed for some bivariate complexity to be reproduced by the methods that make Gaussian assumptions.

Reproduction of horizontal and vertical cross-correlation is examined in Figures 7.7 and 7.8, respectively. Cross-semivariograms of the imputed realizations are overlain with that of the removed true values. The RMSE statistic in these figures are averaged and standardized in Table 7.3. As with the other properties of interest, the NPM method is seen to yield the best cross-correlation.

Summarizing, the NPM imputation method yields the best reproduction of the missing univariate, multivariate and spatial features, while also yielding the best local accuracy. This method incorporates colocated secondary information in a manner that accounts for multivariate complexity. The imputed values better reproduce the multivariate properties of the missing Ni laterite data, which also leads to improved local accuracy and spatial continuity reproduction.

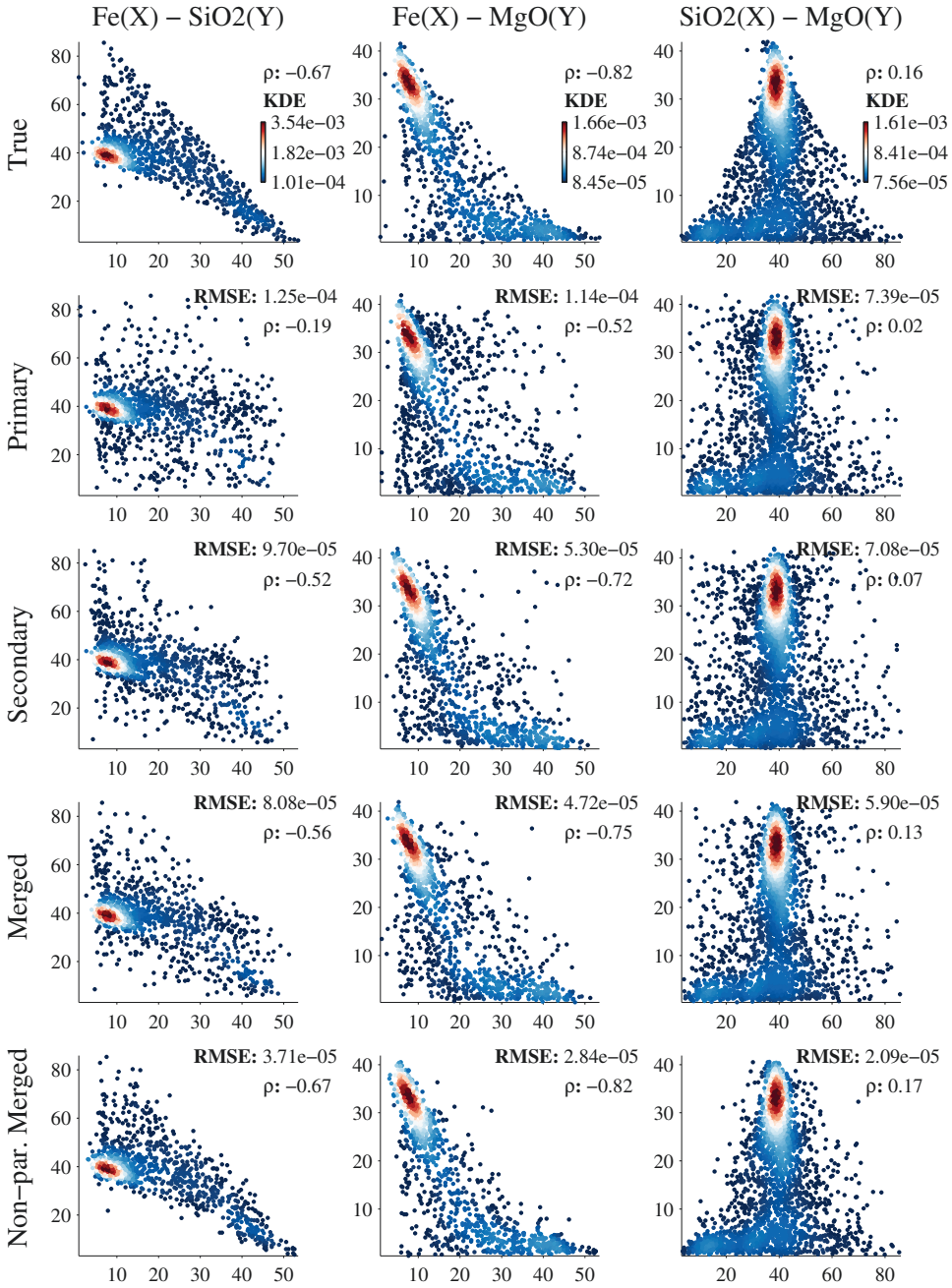


Figure 7.9: KDE scatterplots of an imputed realization, with the true values shown for comparison.

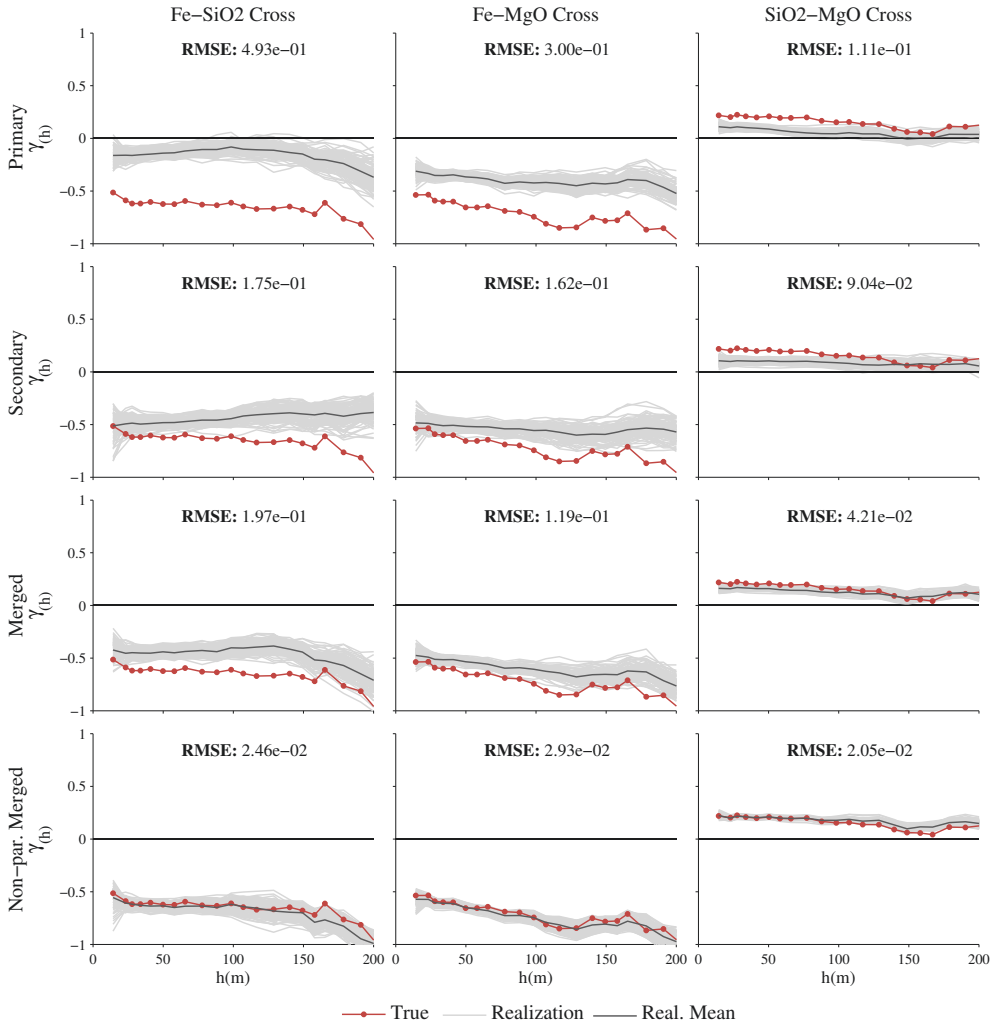


Figure 7.10: Horizontal cross-semivariograms of the true values and imputed realizations.

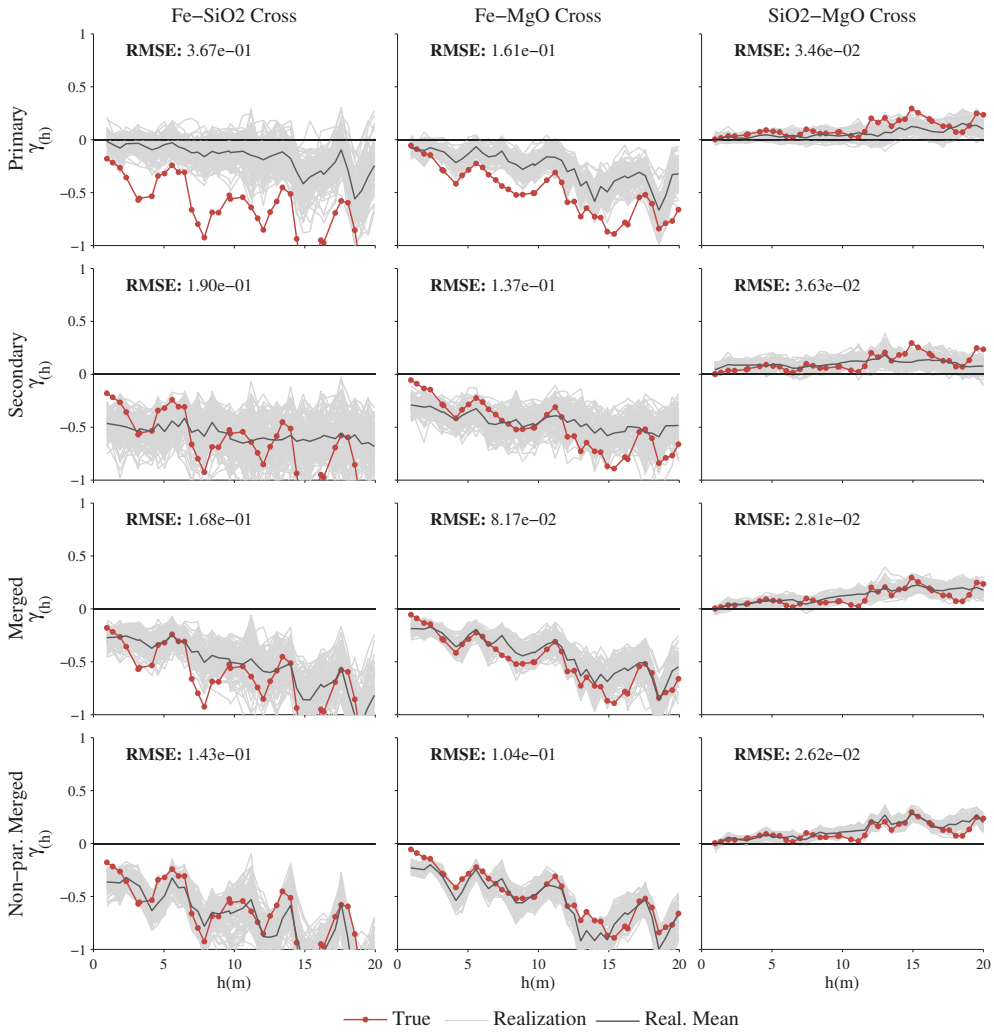


Figure 7.11: Vertical cross-semivariograms of the true values and imputed realizations.

7.3 Impact on Geostatistical Modeling

The previous section establishes that missing Ni laterite data is imputed most effectively by the NPM method. The focus now shifts to geostatistical modeling with imputed homotopic data realizations. A comparison with the more primitive DE and SI methods will also be shown. Performance can be compared using the true jackknife values that have been left out from the beginning. The PPMT/MAF1 modeling workflow from Section 6.4 is applied with a variety of input data. The workflows and modeling results are labeled according their input:

i) No Missing: the missing values were never removed, meaning that true values are used in place of the imputed values. This matches the workflow and results that were presented in Section 6.4 under the PPMT/MAF1 label. This modeling result is expected to outperform the remaining workflows, where the missing values are excluded or imputed. This workflow is included in the summary tables as a relative comparison.

ii) No Impute: the heterotopic data observations are removed, so that only homotopic observations are used for geostatistical modeling. This workflow represents geostatistical modeling with DE.

iii) Merged Mean: a single homotopic dataset is created by imputing each missing value with the e-type of one hundred realizations (merged method). This workflow represents geostatistical modeling with SI.

iv) Primary, Secondary, Merged, NPM: one hundred imputed realizations are used for geostatistical modeling according to the MI framework from Section 3.1. Each workflow is labeled according to the imputation method that was used to generate the applied data realizations. These workflows represent geostatistical modeling with MI when compared to workflows (i), (ii) and (iii). When compared to each other, they demonstrate how improved imputation quality translates to improved geostatistical modeling results.

Steps of the PPMT/MAF1 workflow are held constant for the above described input data. It is used to predict the jackknife locations that are presented and described in Section 6.2.2. In the case of the MI workflows, the PPMT and MAF transformations are repeated on the $L = 100$ data realizations to yield L transformed

data realizations. Modeling is then performed using the L data realizations for simulation conditioning, before using the L recorded transformation tables to return the associated realizations to original space.

Table 7.4 presents the univariate performance of the geostatistical models that result from each input data. Refer to the description of Table 6.2 for an explanation of the format and statistics. As discussed, the No Missing workflow uses true values in place of the missing values. It is expected to yield the best overall result; in cases where it does, both it and the next best workflow are bolded in the Avg. column. Similarly, the No Missing result is excluded when describing the best result, though it is present in the tables for benchmarking the heterotopic data results.

As expected, the NPM MI workflow yields the best univariate and spatial reproduction according to almost every statistic. Conversely, the Merged Mean workflow yields the worst result in terms of global standard deviation and semivariogram reproduction. This is the anticipated consequence of imputing missing values with a single estimate. Variability has been unrealistically reduced in the data, which is translated to the geostatistical realizations. The Merged Mean workflow underperforms the Merged workflow for every statistic in Table 7.4. Geostatistical modeling with SI yields worse univariate reproduction than MI. The No Impute workflow yields the worst local accuracy according to the $1 - \rho$ and RMSE statistics. This is the expected consequence of DE, where valuable local information that is contained in heterotopic observations is excluded from geostatistical modeling.

Table 7.5 presents the multivariate performance of geostatistical modeling with the described input data. Refer to the description of Table 6.3 for an explanation of the format and statistics. The NPM workflow yields the best reproduction of the bivariate densities and cross-semivariograms according to the KDE RMSE and γ RMSE, respectively. The Merged workflow yields the best reproduction of the bivariate correlation, although it is interesting to note that both it and the NPM workflow yield better reproduction than the No Missing workflow (if only narrowly). The SI and DE workflows cumulatively produce the worst multivariate results. The No Impute workflow (DE) yields the worst bivariate density reproduction and cross-semivariogram reproduction, while the Merged Mean workflow (SI) has the worst correlation reproduction.

Finally, Table 7.6 presents the conceptual economic loss that results from each workflow. As discussed in Section 6.5, this economic loss is based on the misclas-

sification of ore types. The ore types are classified using the $L = 100$ realizations to minimize the expected economic loss. The a and $1 - b$ statistics in Table 7.6 display the accuracy and precision of the ore type distributions that underly the economic loss statistic. Figure 7.12 displays the raw value of these statistics in their associated uncertainty accuracy plots.

The NPM workflow yields the best economic loss, which is expected due to its superior characterization of the Ni laterite deposit. Interestingly, this characterization does not lead to the best a and $1 - b$ statistics, although it yields the best overall result for both statistics. The Secondary workflow yields the most accurate distribution of ore type uncertainty, but it also yields the second worst economic loss due to its lack of precision.

The Merged Mean workflow yields the most precise distribution of ore type uncertainty, which may be confusing given that it underperformed the NPM workflow according to every statistic in Tables 7.4 and 7.5. As the Merged Mean method leads to smoother geostatistical realizations, the results are shifted away from realistic variability and towards a deterministic estimation. While unrealistic, these deterministic estimates are less likely to have 'very wrong' multivariate values, leading to improved precision according to the b statistic. Given that the Merged Mean workflow poorly characterized the Ni laterite deposit (Tables 7.4 and 7.5), however, this does not support the application of the SI approach in geostatistical modeling. Further, the smoothed realizations lead to the least accurate distribution of uncertainty for the ore types according to the a statistic. In turn, this leads to worse economic loss than the NPM and Merged MI workflows since risk is not accurately characterized and integrated into resource management decision making.

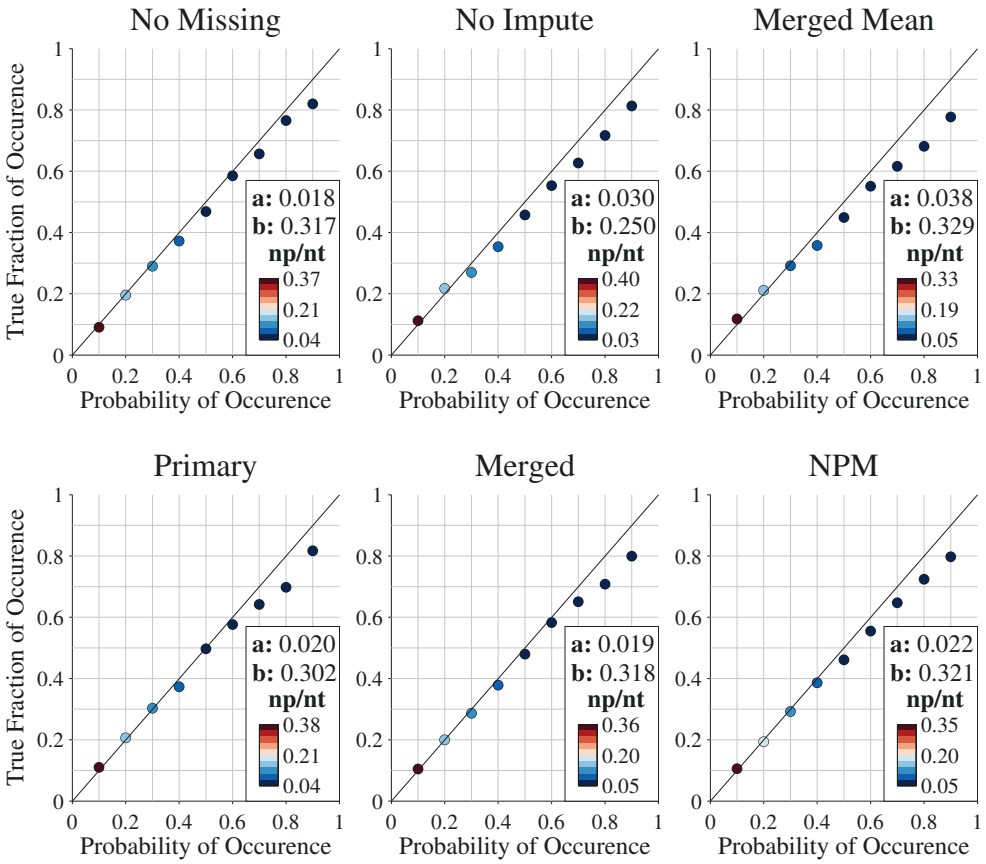


Figure 7.12: Accuracy plots of the ore type distributions resulting from each workflow.

Table 7.4: Standardized univariate performance statistics for geostatistical modeling with various input data.

Measure	Method	Ni	Fe	SiO_2	MgO	Avg.
	No Missing	0.20	0.64	1.00	0.16	0.50
	No Impute	1.00	0.72	0.35	0.45	0.63
	Merged Mean	0.15	0.97	0.76	0.67	0.64
μ Error	Primary	0.20	0.91	0.82	0.47	0.60
	Secondary	0.22	1.00	0.25	0.60	**0.52**
	Merged	0.21	0.97	0.49	0.73	0.60
	NPM	0.20	0.73	0.72	1.00	0.66
	No Missing	0.68	0.41	0.37	0.13	**0.40**
	No Impute	0.56	1.00	0.60	1.00	0.79
	Merged Mean	0.90	1.00	1.00	0.91	0.95
σ Error	Primary	0.92	0.78	0.54	0.37	0.65
	Secondary	1.00	0.51	0.63	0.41	0.64
	Merged	0.88	0.34	0.51	0.26	0.50
	NPM	0.94	0.27	0.31	0.09	**0.40**
	No Missing	0.65	0.66	0.68	0.70	**0.67**
	No Impute	1.00	1.00	1.00	1.00	1.00
	Merged Mean	0.66	0.70	0.77	0.73	0.71
$1 - \rho$	Primary	0.65	0.82	0.78	0.75	0.75
	Secondary	0.65	0.70	0.80	0.76	0.73
	Merged	0.64	0.67	0.75	0.71	0.69
	NPM	0.64	0.66	0.73	0.72	**0.69**
	No Missing	0.82	0.83	0.86	0.85	**0.84**
	No Impute	1.00	1.00	1.00	1.00	1.00
	Merged Mean	0.83	0.86	0.90	0.87	0.86
RMSE	Primary	0.82	0.92	0.90	0.88	0.88
	Secondary	0.82	0.86	0.91	0.88	0.87
	Merged	0.82	0.84	0.89	0.85	0.85
	NPM	0.82	0.83	0.89	0.86	**0.85**
	No Missing	0.60	0.55	0.37	0.89	**0.60**
	No Impute	0.71	0.92	0.71	0.96	0.83
	Merged Mean	0.87	1.00	1.00	1.00	0.97
γ RMSE	Primary	0.92	0.73	0.56	0.74	0.74
	Secondary	1.00	0.62	0.48	0.85	0.74
	Merged	0.86	0.55	0.52	0.74	0.67
	NPM	0.88	0.54	0.36	0.77	**0.64**

Table 7.5: Standardized multivariate performance statistics for geostatistical modeling with various input data.

Measure	Method	N-F	N-S	N-M	F-S	F-M	S-M	Avg.
ρ Error	No Missing	0.54	0.35	0.63	0.09	0.01	0.48	0.40
	No Impute	1.00	0.14	0.88	0.08	0.11	0.33	0.53
	Merged Mean	0.94	0.43	1.00	0.06	0.10	1.00	0.61
	Primary	0.80	0.04	0.42	1.00	1.00	0.22	0.56
	Secondary	0.39	1.00	0.55	0.19	0.14	0.28	0.53
	Merged	0.19	0.32	0.57	0.18	0.11	0.87	**0.32**
	NPM	0.52	0.55	0.43	0.05	0.05	0.44	0.39
KDE RMSE	No Missing	0.52	0.51	0.49	0.36	0.36	0.52	**0.47**
	No Impute	1.00	1.00	1.00	0.49	0.52	0.78	0.87
	Merged Mean	0.73	0.83	0.66	0.79	0.56	1.00	0.75
	Primary	0.89	0.82	0.78	1.00	1.00	0.77	0.87
	Secondary	0.80	1.00	0.80	0.62	0.38	0.86	0.80
	Merged	0.65	0.80	0.73	0.54	0.38	0.77	0.68
	NPM	0.69	0.72	0.67	0.45	0.39	0.58	**0.63**
γ RMSE	No Missing	0.66	0.76	0.58	0.29	0.39	0.89	**0.57**
	No Impute	1.00	0.86	1.00	0.46	0.50	0.93	0.83
	Merged Mean	0.80	0.84	0.60	0.47	0.54	0.84	0.68
	Primary	0.56	1.00	0.47	1.00	1.00	1.00	0.76
	Secondary	0.74	0.94	0.56	0.37	0.47	0.88	0.65
	Merged	0.67	0.89	0.56	0.38	0.46	0.79	0.63
	NPM	0.64	0.87	0.65	0.31	0.42	0.81	**0.62**

Table 7.6: Raw and standardized loss function values for each workflow.

Method	Loss	a	$1-b$
No Missing	0.845	**0.479**	0.911
No Impute	1.000	0.792	1.000
Merged Mean	0.827	1.000	**0.894**
Primary	0.834	0.522	0.930
Secondary	0.843	**0.501**	0.922
Merged	0.825	0.511	0.909
NPM	**0.821**	0.588	0.905

7.4 Summary

The NPM method yields the best imputation of missing Ni laterite data. The resultant data realizations outperformed the other options in terms of reproducing the univariate, multivariate and spatial features of the data, while also yielding superior local accuracy. The merged method generally yielded the second best results. All of the described methods have been implemented in Fortran code that has not been optimized for speed. Nevertheless, one hundred realizations of the presented missing Ni laterite data were imputed using the parametric and NPM methods in 6.68 minutes and 208 minutes, respectively. While significant, the execution time of the NPM method is unlikely to prohibit its use in practice since it must only be performed once. So long as multivariate complexities exist in the imputed data, the extra time is likely to be justified by improved results.

Following imputation of the missing values, the data realizations from each method were applied in identical geostatistical modeling workflows that utilize MI, DE (heterotopic data excluded) and SI (each missing value imputed with an estimate). In general, the DE and SI workflows yielded the worst results. The reduced variability of the SI data translated to poor reproduction of univariate and spatial variability in geostatistical realizations. The excluded information of the DE data led to poor local accuracy. The MI workflows yielded the best results, where improved imputation results (e.g., NPM method) translated to the anticipated relative improvement in geostatistical realizations.

Chapter 8

Conclusions

The following chapter begins with a brief review of the problems that motivated this research. The primary and secondary contributions of this thesis are then summarized. Inevitable limitations remain for these contributions. The limitations are listed along with items for future work. The thesis statement is then revisited before making final concluding remarks.

8.1 Review of the Motivation

Almost every resource modeling setting requires the spatial prediction of K geological variables within the subsurface. Geostatistical modeling tools are the most popular method of prediction, where L realizations of the subsurface are simulated to characterize the associated uncertainty. Conventional geostatistical cosimulation algorithms assume that the K variables follow a multiGaussian distribution. In settings where this is not the case, the original multivariate distribution will not be reproduced by simulated realizations. This unrealistic characterization of the geological variables will negatively impact most resource management decisions that geostatistical models are used for. Moreover, those cosimulation algorithms become difficult to apply in massively multivariate settings ($K > 5$).

The above issues motivated the adaptation of multivariate transformations to geostatistical modeling. Linear decorrelation techniques such as principal component analysis (PCA) and min./max. correlation factors (MAF) are used to remove correlation from the data prior to simulation. The decorrelated variables are independently simulated, before returning correlation to the realizations with the associated back-transform. This simplifies geostatistical modeling significantly and is not sensitive to increasing K. Applying linear decorrelation techniques to complex

multivariate data fails to remove the complexity and dependence. Systematic errors result from simulating complex and dependent variables under an independent multi-Gaussian assumption. This motivates multiGaussian transforms such the stepwise conditional transformation (SCT), which transform variables to an approximately uncorrelated and multiGaussian distribution prior to modeling. While effective in low dimensional settings, the SCT suffers in massively multivariate settings due to binning and limited number of data for the transform of high dimensional variables.

Another restriction on every multivariate transformation technique, is that they may only be applied to homotopic equally sampled observations. Geostatisticians have typically used one of two options to address this issue. The first option, data exclusion (DE), excludes heterotopic observations from multivariate transformations and subsequent geostatistical modeling. DE leads to loss of information, as sampled values of the heterotopic observations are not available for local conditioning and global statistics. The second option, single imputation (SI), imputes the missing values based on regression so that all sampled values may be used for the multivariate transformation and subsequent modeling. SI leads to unrealistic smoothing in the resultant data, which compromises numerous properties of resultant geostatistical models. Both options also introduce bias if the values are not missing completely at random (MCAR).

8.2 Summary of Contributions

The issues described in the previous section motivate the two primary contributions of this thesis: i) methods for the multivariate imputation and geostatistical modeling of heterotopic geological data, and ii) the projection pursuit multivariate transformation (PPMT) for the transformation and modeling of complex geological data. Secondary contributions of this thesis include increased understanding and documentation of alternative multivariate transforms, as well as software for data imputation and transformation.

8.2.1 Multivariate Imputation of Geological Data

The imputation of missing values is a well-established practice in many scientific fields where heterotopic data are encountered. Rather than SI, approaches such as multiple imputation (MI) and maximum likelihood estimation (MLE) are favored since they allow for: i) uncertainty of the imputed values to be passed through the

subsequent analysis, ii) imputed values to possess realistic features of the data, and iii) bias to be avoided when the missing data mechanism is not MCAR.

Consider a typical geostatistical simulation workflow: i) a multivariate transformation removes complexities from homotopic data, ii) the transformed data conditions L geostatistical realizations, and iii) the L realizations are back-transformed to original space using the recorded transform table. In the presence of heterotopic data, the above workflow is modified by this thesis to naturally integrate MI: i) generate L realizations of homotopic data; sampled values are constant across the realizations while imputed values vary according to their uncertainty, ii) multivariate transformations remove complexities from the L homotopic data realizations, iii) the l^{th} transformed data realization conditions the l^{th} geostatistical realization for $l = 1, ..., L$, and iv) the l^{th} realization is back-transformed to original space using the l^{th} recorded transform table. Following this approach, the L realizations characterize the joint uncertainty of the subsurface and missing data. Using heterotopic Ni laterite data, this MI workflow is demonstrated to significantly improve geostatistical characterization of the subsurface relative to equivalent SI and DE workflows. This translated to improved resource management decisions according to a conceptual economic loss statistic.

MI algorithms use iterative simulation techniques such as the Gibbs sampler to generate realizations of the missing data values. Conventional algorithms typically assume that the data are multiGaussian and converge on the correct covariance between colocated values. As a result, these algorithms fail to reproduce the complex multivariate distributions of geological data. Further, they fail to reproduce the spatial variability of geological data data since only the colocated information is considered in the imputation model. Spatial information is incorporated in MI algorithms that have been developed in spatial fields such as environmental monitoring and remote sensing. All reviewed algorithms, however, fail to consider both spatial and colocated information. This prevents accurate reproduction of both colocated and spatial properties of geological data, while also limiting precision since less information is incorporated.

Observing these issues with available MI algorithms, this thesis developed the merged and non-parametric merged (NPM) methods for the imputation of geological data. Working within a Gibbs sampler framework, these techniques incorporate both spatial and colocated information. The merged method assumes the data are

multiGaussian and converges on the correct covariance between colocated values. The NPM method uses KDE to calculate the conditional distribution directly from the colocated values. In doing so, the NPM method imputes missing values in a manner that effectively reproduces complex multivariate distributions. Relative to MI methods that only consider spatial or colocated information, the merged and NPM methods were demonstrated to significantly improve the realism and accuracy of imputed values. The merged method may be considered in settings where the data are reasonably multiGaussian, as it is far faster to execute. The NPM method should be considered in settings that exhibit complex multivariate features. The NPM method yielded significantly better imputation results with the Ni laterite data, which translated to significantly better geostatistical realizations and resource management decision making.

8.2.2 Exploratory Multivariate Transformations

The goal at the outset of this research, was to develop a technique that would transform complex geological data of any K variables and n observations to an uncorrelated multiGaussian distribution. The first attempted transformation was conditional standardization (CS), which removes non-linearity and heteroscedasticity from a multivariate distribution through fitting conditional functions of the means and standard deviations. Subtracting the conditional mean and dividing by the conditional standard deviation yields transformed distributions that are linear and homoscedastic. Simulated realizations may be back-transformed by adding and multiplying the recorded conditional means and standard deviations, respectively. Though conceptually attractive for its simplicity, the effectiveness of CS hinges on the accuracy of its conditional functions. It yielded the best results with a non-parametric approach, where the multivariate distribution is binned by the conditioning variables before calculating the requisite statistics. Unfortunately, however, this binning means that CS suffers from the same dimensionality restrictions as the SCT technique; it is not feasible for K greater than three to four. For this reason, CS was originally conceived using parametric conditional functions that may be calculated for any K dimension. Unfortunately, increasing K makes fitting complex data with mathematical functions increasingly difficult. Initial testing found that the non-linearity and heteroscedasticity could not be adequately characterized. As a result, multivariate complexity persisted in the transformed data.

The second attempted transformation was the multivariate standard normal transformation (MSNT), which mapped the original complex data to a multiGaussian distribution of matching K and n. The MSNT poses this transformation as an optimization problem. Observations in the original distribution are mapped to observations in the multiGaussian distribution in a manner that minimizes changes to the relative multivariate configuration. The relative configuration is defined based on the distances between observations in original space. With that in mind, the MSNT uses simulated annealing to determine the mapping that minimizes changes to these distances. After converging, the mapping is recorded so that simulated realizations may be back-transformed to original space. The nearest $K+1$ neighbours are determined for each simulated node in multiGaussian space. The node is then back-transformed by interpolating its location in original space using the original multivariate values of the $K+1$ neighbours. The MSNT was a direct response to the SCT and CS techniques; it was designed to be appropriate for increasing K variables. While successful in this regard, the MSNT was ultimately abandoned since it is sensitive instead to increasing n observations. The time that is required for it to converge on a suitable mapping increases at beyond a quadratic rate with increasing n, making it infeasible for $n > 1000$.

8.2.3 Projection Pursuit Multivariate Transformation

The basic idea of the MSNT is conceptually attractive. Determine the mapping between original multivariate observations and transformed multiGaussian observations that minimizes changes to the multivariate configuration. Following simulation in multiGaussian space, back-transform the realizations through a multivariate interpolation that is based on the recorded mapping. Referred to as Gaussian mapping (GM), this framework amounts the multivariate extension of a normal score back-transformation. The optimization of the MSNT failed to provide a good mapping for increasing n. This motivated development of the PPMT for the GM framework, which provides a mapping that is not sensitive to increasing n.

A component of the projection pursuit density estimation (PPDE) algorithm transforms high dimensional data to an uncorrelated multiGaussian distribution. After orthogonalizing the data with data sphereing, PPDE uses an optimized search to find non-Gaussian univariate projections of the multivariate data. The multivariate data are then transformed to make that projection Gaussian. Iterating this search

and transformation procedure, PPDE transforms complex multivariate data to be multiGaussian. Given the geostatistical modeling context of the PPMT, changes to the conventional PPDE algorithm (Friedman, 1987) include:

i) A normal score transformation of the K variables to make the data univariate Gaussian. Subsequent data sphereing and projection pursuit steps benefit from this preprocessing, as univariate outliers and complexities are removed.

ii) A modified data sphereing approach that projects the orthogonal data back onto the original basis. Originally applied to PPDE by Hwang et al. (1994), this approach avoids the dimension reduction and variable mixing of conventional data sphereing. Dimension reduction would cause stress to the GM back-transform, since each dimension in transformed space accounts for widely different variability in original space.

iii) A stopping criteria for the projection pursuit iterations, which is based on the Gaussianity of bootstrap random Gaussian distributions of matching n and K.

Following the final projection pursuit iteration, the multiGaussian observations are recorded with their original values to facilitate the described GM back-transform. Reverse projection (RP) is a second back-transformation option, however, where each step of the forward transformation is reversed with the simulated multiGaussian realizations. An advantage of GM is that it explicitly reproduces multivariate constraints of the original distribution. The RP usually leads to multivariate extrapolation beyond the original distribution, which may or may not be realistic depending on specifics of the variables. In general, the RP is favored since it usually yields better overall reproduction of the univariate and multivariate densities, while also being faster to execute with increasing n. GM must search for the nearest $K+1$ neighbours to each simulated node; as such, execution time increases as a function of n. The RP back-transform is not sensitive to increasing n, though its execution time does increase linearly with increasing projection pursuit iterations.

The PPMT is only concerned with removing covariance between the variables at $h = 0$ lag distance. It therefore benefits from a subsequent MAF transformation to remove cross-covariance at $h > 0$ lag distance. The PPMT/MAF geostatistical workflow outperforms popular alternatives such as cosimulation, SCT and MAF in terms of multivariate and spatial characterization. This is demonstrated using a

controlled synthetic example, as well as a real Ni laterite case study. In turn, this characterization is shown to improve resource management decisions according to conceptual economic loss.

8.2.4 Software

A large amount of software was implemented throughout this research. Not only for testing new methodologies, but also for conventional methods that were required for comparison and benchmarking. This software forms a secondary but practical contribution of this thesis; it is divided into the following categories:

i) Imputation software, including programs for the primary (`impute_pri`), secondary (`impute_sec`), and merged/NPM (`impute`) imputation methods.

ii) Conventional transformation software, including programs for PCA (`pca` and `pca_b`), MAF (`pca` and `pca_b`), SCT (`stepwise` and `stepwise_b`), logratios (`logratio` and `logratio_b`), and histogram corrections (`histcorrect`).

iii) New transformation software, including programs for CS (`constd` and `constd_b`), MSNT (`msnt` and `msnt_b`), and PPMT (`ppmt` and `ppmt_b_gm` and `ppmt_b_rp`).

Each program is coded in Fortran and follows a GSLIB-style convention (Deutsch and Journel, 1998), meaning that they require ASCII parameter and data files as input. The software is available from the author upon request.

8.3 Limitations and Future Work

Limitations remain for the described contributions, which motivates future work in the areas of multivariate imputation and transformation.

8.3.1 Multivariate Imputation

Although the NPM method yields the best imputation results in all real test cases, many implementation details could be improved for the technique. First, the coding should be reviewed to determine bottlenecks in the algorithm. The current execution time of the NPM method does not prevent its use, as it usually finishes within a few hours. That said, faster execution would further encourage its use and allow for iterative parameter experimentation. Potential improvements include: i) dividing

independent Gibbs sampler sequences across multiple processors, which is advantageous in terms of both execution speed and accuracy of uncertainty (see below), and ii) computationally efficient kernel density networks (KDN) rather than conventional KDE, since KDN does not require the fitting of kernels to every observation.

The current speed challenges motivate the chosen method for extracting realizations from the Gibbs sampler. Following b burn-in iterations, consecutive L realizations are extracted from L iterations of a single Gibbs sequence. The b iterations mitigate dependency of the extracted values on the random starting location, which improves resultant accuracy. The extraction of consecutive iterations allows for a reduction of execution time. Increasing execution time should be weighed against the shortcoming of this practical extraction method. Dependency exists between the extracted realizations, which may artificially increase the certainty of imputed values. Users could be provided with flexibility to execute multiple Gibbs sequences and/or periodic extraction to mitigate this issue when execution time allows. Although limited by the number of processors, a number of independent Gibbs sequences could be executed without incurring additional execution time.

KDE lies at the heart of the NPM method, where it is used to build conditional distributions of a missing value based on colocated samples. The imputation program is implemented to streamline KDE parameterization; the user specifies a single H parameter, which is scaled by the data correlation to populate the bandwidth matrix $\mathbf{H}$. This thesis relied on the iterative tuning of H to match visual expectation. Future work will focus on automated procedures for fitting $\mathbf{H}$ to further streamline this process for users and potentially improve the results. The previously described KDN research precedes this work in priority since it will likely impact the chosen bandwidth tuning methodology.

Attention will also be paid to additional complexities that are encountered in geological data, such as building on the work of Martin-Fernandez et al. (2003) and (Tjelmeland and Lund, 2003) for the reproduction of compositional sum constraints (Section 2.2.3).

8.3.2 Multivariate Transformations

Although the PPMT is currently advocated over the CS and MSNT techniques, this does not preclude the applicability of the latter transformations if large advances are made with key implementation details. Both are attractive in terms of their

conceptual simplicity. To be viable, however, the CS will require further research into fitting parametric functions of the conditional mean and standard deviation functions to high dimension distributions. The current CS algorithm does not adequately characterize non-linearity and heteroscedasticity with increasing K, so that those features persist following the transformation. The MSNT will require further research into the optimization that is used for mapping complex multivariate data to a multiGaussian distribution. The current MSNT algorithm takes too long to converge for increasing n.

Immediate research, however, will focus on advances to the PPMT. The primary concern with the PPMT workflow, is a loss of spatial continuity (if only slight) that is usually observed in geostatistical realizations at short scale $h > 0$ lag distances. This is attributed to forcing dependent variables to be entirely independent at $h = 0$, which leads to spatial destructuring of at least one variable in transformed space. Following simulation and back-transformation, the original continuity is not entirely recovered. One ad-hoc solution is to inflate the continuity of semivariogram models that are used as input to the Gaussian simulation algorithm. A less ad-hoc solution, however, may be the use of dimension reduction data sphereing in the PPMT pre-processing. It makes logical sense that the spatially destructured transformed variables are likely to contribute less information to the multivariate system. As a result, attributing less of the original variability to these destructured variables may mitigate their impact on the back-transformed continuity. Although the GM back-transform may be complicated by this dimension reduction, the RP back-transform should remain suitable.

This thesis advocates the use of PPMT and MAF in combination; the former makes the variables independent at $h = 0$ while the latter decorrelates the variables at $h > 0$. A future modification of the PPMT could integrate spatial decorrelation and remove the need for a subsequent MAF transformation. Cross-correlation could be calculated periodically within the projection pursuit iterations at various h lag distances. Data sphereing could then be used to orthogonalize the data at each lag distance, where the sphereing matrix is calculated based on the spectral decomposition of the cross-covariance at each h. This approach would simplify the workflow since only one transformation is required; it may also yield superior removal of cross-correlation since multiple lag distances would be iteratively decorrelated.

Another interesting avenue would be to consider multivariate mixture models to

fit complex multivariate distributions. If this were possible, then imputation and multivariate transformation would be facilitated.

8.4 Final Remarks

Recall the thesis statement: *Improved spatial prediction of geological variables accounting for complex relations and unequal sampling will lead to improved resource management decisions.*

To address unequal sampling, this thesis contributes methodology for the imputation and geostatistical modeling of heterotopic geological data. This methodology is demonstrated to improve the accuracy and realism of imputed values relative to conventional MI and other ad-hoc missing data schemes. This translates to improved accuracy and realism of subsequent geostatistical models, which leads to improved resource management decisions.

To address complex relations, this thesis contributes PPMT methodology for the transformation and geostatistical modeling of complex geological data. This methodology outperforms alternatives such as the SCT in terms of decorrelation and multivariate Gaussianity of the transformed data. This translates to improved accuracy and realism of subsequent geostatistical models, which leads to improved resource management decisions.

Bibliography

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data: Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

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

Anglo-American (2011). Barro Alto plant picture, accessed April 2, 2015:[https://www. flickr. com/photos/angloamerican/8047243340].

Anglo-American (2012). Barro Alto fact sheet, accessed April 2, 2015:[http://www. angloamerican. com/ /media/files/a/anglo-american-plc/media/angloamerican_fs_barro%20alto_2012_print. pdf].

Babak, O. and Deutsch, C. V. (2009a). Collocated cokriging based on merged secondary attributes. *Mathematical Geosciences*, 41:921–926.

Babak, O. and Deutsch, C. V. (2009b). An intrinsic model of corregionalization that solves variance inflation in collocated cokriging. *Computers and Geosciences*, 35:603–614.

Barnett, R. M. and Deutsch, C. V. (2012a). Multivariate standard normal transformation: Advances and case studies, paper 101. In *CCG Annual Report 14*, University of Alberta, Edmonton.

Barnett, R. M. and Deutsch, C. V. (2012b). Practical implementation of non-linear transforms for modeling geometallurgical variables. In Abrahamsen, P., Hauge, R., and Kolbjornsen, O., editors, *Geostatistics Oslo 2012*, pages 409–422. Springer, Netherlands.

Barnett, R. M. and Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. *Mathematical Geosciences*, pages 1–27 [in press].

Barnett, R. M., Manchuk, J. G., and Deutsch, C. V. (2013). Advances in the projection pursuit multivariate transform, paper 102. In *CCG Annual Report 15*, University of Alberta, Edmonton.

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

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, New Jersey.

Bentley, J. L. (1980). Multidimensional divide and conquer. *Communications of the ACM*, 23:214–229.

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

Boisvert, J. B., Rossi, M. E., and Deutsch, C. V. (2009). Multivariate geostatistical simulation of proportions and nonadditive geometallurgical variables, paper 303. In *CCG Annual Report 11*, University of Alberta, Edmonton.

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

Boucher, A. and Dimitrakopoulos, R. (2012). Multivariate block-support simulation of the Yandi iron ore deposit, Western Australia. *Mathematical Geosciences*, 44:449–468.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46:167–174.

Chatfield, C. and Collins, A. (1960). *Introduction to Multivariate Analysis*. Chapman and Hall, London.

Chen, H. Y. and Little, R. (1988). A test of missing completely at random for generalized estimating equations with missing data. *Biometrika*, 86:1–13.

Chiles, J. P. and Delfiner, P. (2012). *Modeling Spatial Uncertainty, 2nd edn.* Wiley, New York.

Collins, L. M., Shafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psychological Methods*, 6:330–351.

Davis, B. M. (1987). Uses and abuses of cross-validation in geostatistics. *Mathematical Geology*, 19:241–248.

Davis, B. M. and Greenes, K. A. (1983). Estimating using spatiallly distributed multivariate data: an example with with coal quality. *Mathematical Geology*, 15:287–300.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Desbarats, A. J. and Dimitrakopoulos, R. (2000). Geostatistical simulation of regionalized poresize distributions using min/max autocorrelations factors. *Mathematical Geology*, 32:919–942.

Deutsch, C. V. (1992). *Annealing Techniques Applied to Reservoir Modeling and the Integratioin of Geological and Engineering (Well Test) Data*. PhD thesis, Stanford University.

Deutsch, C. V. (2005a). A new TRANS program for histogram and trend reproduction, paper 306. In *CCG Annual Report 7*, University of Alberta, Edmonton.

Deutsch, C. V. (2005b). Order relations correction and tail extrapolation for stepwise conditional transformation, paper 109. In *CCG Annual Report 7*, University of Alberta, Edmonton.

Deutsch, C. V. (2011). Multivariate standard normal transformation, paper 101. In *CCG Annual Report 13*, University of Alberta, Edmonton.

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

Deutsch, C. V. and Zanon, S. (2004). Direct prediction of reservoir performance with Bayesian updating under a multivariate Gaussian model.

Deutsch, J. L. and Deutsch, C. V. (2011). Plotting and checking the bivariate distributions of multiple Gaussian data. *Computers and Geosciences*, 37:1677–1684.

Deutsch, J. L. and Deutsch, C. V. (2012a). Accuracy plots for categorical variables, paper 404. In *CCG Annual Report 14*, University of Alberta, Edmonton.

Deutsch, J. L. and Deutsch, C. V. (2012b). Latin hypercumbe sampling with multidimensional uniformity. *Journal Statistical Planning and Inference*, 142:763–772.

Dimitrakopoulos, R. (2011). Strategic mine planning under uncertainty. *Journal of Mining Science*, 47:138–150.

Doyen, P. M., Boer, L. D., and Pilley, W. R. (1996). Seismic porosity mapping in the ekofisk field using a new form of collocated cokriging. In *Society of Petroleum Engineers, SPE 36498 presented at the SPE Annual Conference and Exhibition*.

Efron, B. (1994). The jackknife, the bootstrap, and other resampling plans. *Society for Industrial and Applied Math*, 26:197–204.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35:279–300.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Press, New York.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23:881–890.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition.* Academic Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

George, E. I. and McCulloch, R. E. (1991). Variable selection via Gibbs sampling. *Technical Report, University of Chicago, Graduate School of Business*.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Seattle, Washington.

Gnanadesikan, R. and Kettenring, J. (1999). Discriminant analysis and clustering: Panel on discriminant analysis, classification, and clustering. *Statistical Science*, 4:34–69.

Godoy, M. (2003). *The Effective Management of Geological Risk in Long term Production Scheduling of Open Pit Mines*. PhD thesis, University of Queensland.

Goovaerts, P. (1994). On a controversial method for modeling a coregionalization. *Mathematical Geology*, 26:197–204.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8:206–213.

Hoare, C. A. R. (1962). Quicksort. *The Computer Journal*, 5:10–16.

Hong, S. (2010). *Multivariate Analysis of Diverse Data for Improved Geostatistical Reservoir Modeling*. PhD thesis, University of Alberta.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441.

Huang, R. and Carriere, K. C. (2006). Comparison of methods for incomplete repeated measures data anlaysis in small samples. *Journal of Statistical Planning and Inference*, 136:235–247.

Huber, P. J. (1985). Projection pursuit. *Annals of Statistics*, 13:435–475.

Hwang, J., Lay, S., and Lippman, A. (1994). Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42:2795–2810.

Info-Mine (2012). Anglo's new nickel, accessed April 2, 2015:[http://www. infomine. com/library/publications/docs/internationalmining/moore2012h. pdf].

Isaaks, E. H. (1990). *The Application of Monte Carlo Methods to the Analysis of Spatially Correlated Data*. PhD thesis, Stanford University.

Jewbali, A. (2009). Finding the nearest postive defininte matrix for input to semi-automatric variogram fitting, paper 402. In *CCG Annual Report 11*, University of Alberta, Edmonton.

Johnson, R. J. and Wichern, D. W. (1998). *Applied Multivariate Statistical Analysis, 4th edn.* Prentice Hall, New Jersey.

Johnston, L. P. and Kramer, M. A. (2011). Probability density estimation using elliptic basis functions. *American Institute of Chemical Engineers Journal*, 40:1639–1649.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*. Academic Press, London.

Journel, A. G. and Xu, W. (1994). Posterior identification of histograms conditional to local data. *Mathematical Geology*, 26:323–359.

Knight, J. R., Sirmons, C. F., Gelfand, A. E., and Ghosh, S. K. (1998). Analyzing real estate data problems using the Gibbs sampler. *Real Estate Economics*, 26:469–492.

Knuth, D. (1998). *The Art of Computer Programming: Volume 3: Sorting and Searching, 2nd edn.* Addison-Wesley Professional.

Kumar, A. and Deutsch, C. V. (2009). Optimal correction of indefinite correlation matrices. *CCG Annual Report 11*, Paper 401.

Larrondo, P. F., Neufeld, C., and Deutsch, C. V. (2003). Varfit: A program for semi-automatric variogram modeling, paper 122. In *CCG Annual Report 5*, University of Alberta, Edmonton.

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

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

Li, Y. Y. and Parker, L. E. (2008). A spatial-temporal imputation technique for classification with missing data in a wireless sensor network.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd edn.* Joh Wiley & Sons Inc., New Jersey.

Lokupitiya, R. S., Erandathie, L., and Paustian, K. (2006). Comparison of missing value imputation methods for crop yield data. *Environmetrics*, 17:339–349.

Manchuk, J. G. (2008). *CCG Guidebook Series Vol. 7: Guide to Geostatistics with Compositional Data*. University of Alberta, Edmonton, Canada.

Manchuk, J. G. and Deutsch, C. V. (2011). A program for data transformations and kernel density estimation, paper 116. In *CCG Report 13*, University of Alberta, Edmonton.

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

Martin-Fernandez, J. A., Barcelo-Vidal, C., and Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35:253–278.

Matheron, G. (1962). *Traité de Géostatistique Appliquée, vol. 1.* Editions Technip, Paris.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

Minnitt, R. C. A. and Deutsch, C. V. (2014). Cokriging for optimal recoverable reserve estimates in mining operations. *SAIMM*, 114:189–203.

Mueller, U. A. and Ferreira, J. (2012). The U-WEDGE transformation method for multivariate geostatistical simulation. *Mathematical Geosciences*, 44:427–448.

Munoz, B., Lesser, V. M., and Smith, R. A. (2010). Applying multiple impuation with geostatistical models to account for item nonresponse in environmental data. *Journal of Modern Applied Statistical Methods*, 9:276–286.

Neufeld, C. and Deutsch, C. V. (2004). Developments in semiautomatic variogram fitting, paper 404. In *CCG Annual Report 6*, University of Alberta, Edmonton.

Neufeld, C. and Deutsch, C. V. (2006). Data integration with non-parametric Bayesian updating, paper 105. In *CCG Annual Report 8*, University of Alberta,Edmonton.

Neufeld, C. and Deutsch, C. V. (2008). Sequential approach to covariance correction for p-field simulation. *CCG annual Report 10*, 121.

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

Niven, E. B. and Deutsch, C. V. (2008). Application of the sampling distribution of the correlation coefficient. *CCG Annual Report 10*, Paper 117.

Olea, R. (1991). *Geostatistical Library and Multilingual Dictionary.* Oxford University Press, New York.

Oman, S. D. and Vakulenko-Lagun, B. (2012). A remark on the use of a weight matrix in the linear model of coregionalization. *Mathematical Geosciences*, 44:505–512.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statitistics*, 33:1065–1076.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2*, 11:559–572.

Pyrcz, M. J. and Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling, 2nd edn.* Oxford University Press.

Reed, M. and Simon, B. (1972). *Functional Analysis*, volume 1 of *Methods of Modern Mathematical Physics*. Academic Press, New York.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statitistics*, 23:470–472.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rubin, D. B. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pages 30–34.

Rubin, D. B. (1988). A test of missing completely at random for multivaraite data with missing values. *Journal of the American Statistical Association*, 83:1198–1202.

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

Scott, D. W. (1992). *Multivariate Density Estimation: Thoery, Practice, and Visualization*. John Wiley and Sons, Inc., New York.

Shafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177.

Switzer, P. and Green, A. A. (1984). Min/max autocorrelation factors for multivariate spatial imaging. Technical report, Stanford University.

Tjelmeland, H. and Lund, K. V. (2003). Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, 30:87–100.

Verly, G. (1983). The multiGaussian approach and its applications to the estimation of local reserves. *Mathematical Geology*, 15:249–286.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications, 3rd edn.* Springer, New York, NY.

Xu, W., Tran, T. T., Srivastava, R. M., and Journel, A. G. (1992). Integrating seismic data in reservoir modeling: the collocated cokriging alternative. In *Socienty of Petroleum Engineers Annual Technical Conference and Exhibition*, pages 833–842, Washington, D. C.

Yuebiao, L. and Zhiheng, L. (2013). Efficient missing data imputing for traffic flow by considering temporal and spatial dependence. *Transportation Research Part C: Emerging Technologies*, 34:108–120.