Multivariate Geostatistical Prediction of Geochemical Measurements for Use in Probabilistic Mineral Prospectivity Modeling

by

Warren Edward Black

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

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Traditional approaches to mineral exploration rely on personal experience, conceptual genetic models, past exploration data, and geological characteristics found in analogous target deposit types to locate and evaluate prospective areas. With the increase in both size and complexity of datasets used in mineral exploration, mineral prospectivity modeling (MPM) provides a means of exploring highly dimensional geological datasets in a meaningful way. When exploring for a specific deposit type, prior knowledge from known mineral deposits within or near the study area and genetic characteristics of the deposit type are used to understand exploration factors that indicate the presence of a mineral deposit (i.e., positive information). A concern is that barren locations (i.e., negative information) are rarely recorded for widespread use by others, yet they are as important as positive locations in training predictive models. It is likely that prospective areas are not being discovered as current methodologies are heuristic in nature and do not consider the full spectrum of the truth.

A proposed novel MPM framework provides a means of passing a stochastic multi-element model and other relevant geological data to a transfer function that calculates the probability that a particular mineral deposit type exists at each location. The use of a multi-element geochemical model allows both positive and negative information to be equally represented while avoiding heuristic searches by not using known mineral occurrences as input. In addition, the multiple realizations of the geochemical model permits uncertainty to be transferred to the final probabilistic values at each location. The principle challenge within the proposed framework is the prediction of the required stochastic geochemical model. It is desired to have a flexible multi-element geochemical model that may be used to perform MPM for many deposit types.

In hopes of providing a straightforward multivariate simulation framework, novel extensions of the decorrelation and direct cosimulation frameworks that operate in the presence of many secondary data are developed; however, they fail to adequately reproduce input multivariate statistics. The introduction of correlation to the once uncorrelated factors during simulation by the conditioning of secondary data renders the decorrelation framework inadequate. While the modification of direct cosimulation is easy to implement, it is hampered by extreme variance inflation and an inability to reproduce the input correlation structure. As the capabilities of the cokriging and hierarchical framework to model highly dimensional problems had not been demonstrated, they were both implemented in an attempt to predict 42 variables. A linear model of coregionalization consisting of 903 direct and cross-variograms was fitted to the data, however, it did not adequately

capture the spatial structure of the input variables. The framework also proved to be very computational expensive. Conversely, the hierarchical framework reasonably reproduces input univariate and collocated multivariate statistics and provides a viable option for simulating multivariate data with many exhaustive secondary data.

The proposed MPM framework is demonstrated in a small example workflow that when passed the geochemical model produced by the hierarchical framework and other relevant geological data, predicted three locations with high probability of deposit discovery. All three locations were in very close proximity (i.e., within 50-2700 meters) to either a showing, drilled prospect, or past-producer, which is a promising sign but requires additional research.

Acknowledgments

Thank you to my supervisor, Dr. Clayton Deutsch, for enabling and supporting this research. Your passion and drive is admirable. Thank you to everyone at the Centre for Computational Geostatistics, specifically Dr. Johnathan Manchuk for your advice and help with my research.

I would like to acknowledge Dr. Maureen Stratton for ensuring that I did not fall between the cracks at an early age. Thank you to Kim Krahn and Charmaine Whitbourne for being instrumental in enabling my success through specialized education while providing confidence and reassurance in my abilities. If it were not for your efforts and passion for education, I would not have the foundation required to complete my post-secondary education.

To my mother, Judith, thank you for providing stability in my life and supporting me through everything.

Finally, to my fiancé, Amélie, you're my rock.

Table of Contents

List of Figures

List of Symbols

Symbol	Description
$^\circ$	degrees as a unit of measurement of angles
$\forall$	for all
A	modeling or geologic domain
Δ	change in value
F	cumulative distribution function of a random function
γ	variogram
Γ	nested variogram structure
$\mathbf{\Gamma}$	vector of nested variogram structures
G^{-1}	standard normal quantile function
$\mathbf{h}$	separation vector
x_{50}	median value of a random function
m	mean of a random function
n	number of observations
ρ	correlation coefficient
σ	standard deviation of a random function
$\mathbf{u}$	coordinate vector
U	overall uncertainty of a probabilistic model

List of Abbreviations

Abbreviation	Description
2-D	two dimensional
3-D	three dimensional
Ag	silver
AOI	area of interest
As	arsnic
Au	gold
Ba	barium
Br	bromine
CCC	collocated Cholesky cosimulation
Cd	cadmium
CD Pb-Zn	clastic-dominated lead-zinc
CDF	cumulative density function
Ce	cerium
Co	cobalt
Cr	chromium
Cs	cesium
Cu	copper
DEM	digital elevation model
EDA	exploratory data analysis
Eu	europium
Exp	exponential variogram structure
F	fluorine
Fe	iron
GB	gigabyte
GHz	gigahertz
GIS	geographic information system
Hf	gafnium
Hg	mercury
ICCK	intrinsic collocated cokriging

Abbreviation	Description
ICP-AES	inductively coupled plasma atomic emission spectroscopy
INA	instrumental neutron activation
km	kilometer
km^2	squared kilometer
La	lanthanum
LMC	linear model of coregionalization
LOI	loss on ignition
LU	lower upper triangular matrix
Lu	lutetium
m	meter
Mag	magnetics
MDS	multidimensional scaling
Mn	manganese
Mo	molybdenum
MPM	mineral prospectivity modeling
MSE	mean squared error
Na	sodium
Ni	nickel
Pb	lead
PC	principle component
PCA	principle component analysis
PDF	probability density function
ppm	parts per million
PPMT	projection pursuit multivariate transform
PVar	primary variable
RAM	random access memory
Rb	rubidium
real	realization distributions
ref	reference distribution
RF	random function
RV	random variable
Sb	antimony

Abbreviation	Description
Sc	scandium
SCK	simple cokriging
SEDEX	sedimentary exhalative
SGS	sequential Gaussian simulation
SGSIM	sequential Gaussian simulation
Sm	samarium
Sn	tin
SoR	slope of regression
Sph	spherical variogram structure
Sphere-R	reverse data sphering
SR	Sphere-R factor
SVar	secondary variable
Ta	tantalum
Tb	terbium
Th	thorium
U	uranium
`USGSIM`	ultimate sequential Gaussian simulation
V	vanadium
`varfit_lmc`	LMC variogram fitting
W	tungsten
Yb	ytterbium
Zn	zinc

Chapter 1

Introduction

1.1 Thesis Setting

Mineral exploration can be described as a two stage process: initial targeting followed by direct detection. Direct detection (e.g., ground geophysics, drilling) is conducted within relatively small prospect areas while initial targeting is completed at a global to district scale (Hronsky & Groves, 2008). The traditional approach to initial targeting uses personal experience, conceptual genetic models, past exploration data, and geological characteristics found in analogous target deposit types to locate and evaluate prospective areas. The generated targets are then represented on a map as highlighted areas of interest. While this method is generally fast, it is limited due to its inherent subjectiveness and bias towards known mineral deposit types (Porwal & Kreuzer, 2010). Compounding with these issues, the geological characteristics commonly sought can vary between adjacent deposits, even if they are similar in style (Kreuzer, Etheridge, Guj, McMahon, & Holden, 2008).

As exploration proceeds within known mineral provinces, mineral deposits with obvious geochemical and/or geophysical signatures are discovered increasing the difficulty of locating new deposits (Groves, 2008). This is illustrated further by a relatively stagnant discovery rate over a period of time, that saw worldwide exploration expenditures rise from 3 billion in 2002 to 30 billion US dollars in 2012. Discovery costs have doubled over this period to the figure of 150 and 180 million US dollars per gold and base metal deposits respectively (Schodde, 2014b). Schodde (2014a) suggests these increases are largely due to the general rise in expenditures and that they are weakly correlated with depth of cover of the deposit. By developing new or advancing mineral exploration techniques and technologies, the increase in discovery costs can be offset (Schodde, 2014b) while potentially increasing the likelihood of new deposit discoveries within known mineral provinces (Hronsky & Groves, 2008).

With the increase in both size and complexity of datasets used in mineral exploration (i.e., geological, geochemical, geophysical, and remote sensing), mineral prospectivity modeling (MPM) provides a means of exploring this data in a meaningful way. Mineral prospectivity models are generated by investigating exhaustive independent data. Factors that are relevant to exploration are subsequently integrated together, producing a prospectivity model specific to a deposit type (Knox-Robinson, 2000). The produced model illustrates the spatial distribution of relative prospectivity and are expressed as ranks, categories, or numerical scores (Lisitsin, Porwal, & Mccuaig, 2014).

Traditionally, mineral exploration is heuristic in nature. Analogies of known mineral deposits

are searched for within known mineral provinces. Prior knowledge, be it known mineral deposit locations or characteristics (i.e., positive information), is applied. Considering this, it is likely that prospective areas are not being discovered as the methodologies are heuristic in nature. Additionally, in a data driven approach, barren negative locations (i.e., negative information) are rarely recorded, yet they are as important as positive locations in training predictive models. As such, the predictive models are derived from censored data.

1.2 Thesis Statement

A proposed novel MPM framework provides a means of passing a stochastic multi-element model and other relevant geological data to a transfer function that calculates the probability that a particular mineral deposit type exists at each location. The use of a multi-element geochemical model allows both positive and negative information to be equally represented while avoiding heuristic searches by not using known mineral occurrences as input. In addition, the multiple realizations of the geochemical model permits uncertainty to be transferred to the final probabilistic values at each location.

To generate the geochemical model, the prediction of regional geochemical measurements that have had many elements measured is required. As a means of improving prediction, the geochemical model may be conditioned to numerous exhaustive geological datasets, increasing the complexity of the geostatistical problem. Stream sediments are commonly sampled regionally; however, it is typical for their geochemical analysis to consist of more than 40 elements when using a technique such as inductively coupled plasma atomic emission spectroscopy (ICP-AES) (ALS Limited, 2009). The transfer function of the proposed MPM framework could be defined prior to simulating the geochemical model enabling the number of elements that require prediction to be known. However, if additional elements are later required, a new geochemical model would need to be generated. The flexibility of the framework is maximized when all available elements in the regional samples are predicted and included in the geochemical model as the transfer function can be altered after its initial construction. In addition, the same geochemical model could then be passed to multiple MPM transfer functions, each designed for different deposit types. With this level of dimensionality it is challenging to ensure that input multivariate statistics are reproduced during spatial prediction. This is the principle challenge of the proposed MPM framework: the multivariate geostatistical prediction of the stochastic multi-element geochemical model in the presence of many secondary data.

While the proposed MPM framework provides the motivation of this thesis, it comprises of only one part of the contributions made by this work. The need for an effective means to generate the required stochastic multi-element geochemical model in the presence of numerous exhaustive secondary data is the primary focus of this thesis. Both novel and existing frameworks are explored

and are proven to not be adequate approaches. One framework is found to be capable of the required spatial prediction and is used to provide the proposed MPM framework with the stochastic geochemical model.

1.3 Thesis Outline

A literature review of relevant concepts and geostatistical techniques to this thesis is found in Chapter 2. Novel frameworks for massively multivariate simulation in the presence of numerous exhaustive secondary data are developed and tested in Chapter 3. Due to the ineffectiveness of the frameworks discussed in Chapter 3, Chapter 4 tests existing multivariate simulation frameworks and discusses the hierarchical framework that is able to model the massively multivariate system. Chapter 5 develops the proposed MPM framework that utilizes a stochastic geochemical model that produces a probabilistic mineral prospectivity model using a binary transfer function. Discussion regarding complexities not appreciably handled in the developed MPM framework, future considerations, and conclusions are discussed in Chapter 6.

The workflows implemented throughout this thesis are not currently available in commercial software and require specialized software and a scripting environment. Computational intensive algorithms are implemented using FORTRAN, the majority of which is sourced from Centre for Computational Geostatistics (2016a) while some are developed for work presented in this thesis. Scripting was completed using the Anaconda Python distribution: Continuum Analytics (2015). A significant contribution to the mixed FORTRAN-Python package `pygeostat` (Centre for Computational Geostatistics, 2016b) was made for work presented in this thesis.

Chapter 2

Literature Review and Background

2.1 Mineral Prospectivity Modeling

2.1.1 Introduction

MPM provides a means of exploring geological data in a meaningful way by integrating different data together producing a model or map of prospectivity, indicating a spatial distribution of relative prospectivity. The various types of geological data available to MPM is discussed in this section in addition to the current MPM frameworks.

2.1.2 Exploration Data

2.1.2.1 Data Types

Data used during mineral exploration can be categorized into three categories: raw, interpreted, or derived. Raw data consists of directly measured data (e.g., geophysics, remote sensing), while interpreted data is extrapolated from a small sample of raw data (e.g., geology maps, structure maps). Derived data is information derived from either raw or interpreted data (e.g., distance to nearest contact, nearest neighbor lithology) (Mccuaig, Porwal, Joly, & Ford, 2013).

2.1.2.2 Stream Sediment Samples

Stream sediment samples are a composite of materials deposited along drainage systems derived from upstream sources that have been subjected to weathering and erosion. Elemental analysis of the samples illustrate the geochemical signature of each sample, that are representations of the background geochemistry upstream and in rare cases, anomalies (Carranza, 2008).

The area of influence associated with stream sediment samples is called a catchment area (Figure 2.1). A catchment area is derived by using a digital elevation model (DEM) to calculate the area that can contribute to the stream at the location sampled (Jones, 2002). In the case when catchment areas overlap due to samples being up or downstream from one another, the catchment area(s) of upstream samples need to be included in the downstream catchment area. Exclusion assumes that the upstream sample(s) have no influence on downstream samples.

2.1.3 Current Mineral Prospectivity Modeling Frameworks

Various methods exist to integrate the factors relevant to exploration. These methods can be categorized as either knowledge driven, data driven, or a hybrid of the two (Knox-Robinson, 2000).

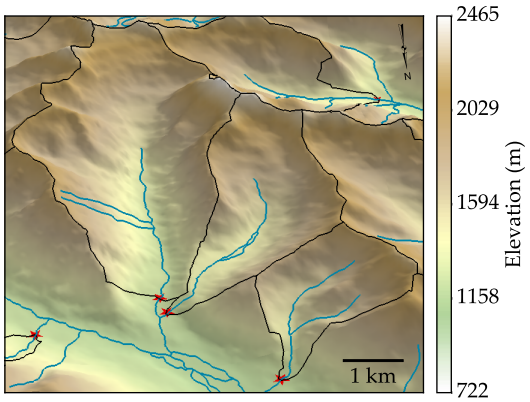

Figure 2.1: An example of delineated catchment areas (black outline) with stream sample locations (red star) and the drainage network (blue lines). The DEM has been shaded using a light source from the south-east at an altitude of 30° to aid in visualizing changing elevation.

The distinction between data driven and knowledge driven mineral prospectivity modeling is defined by the role of statistical relationships derived from the data versus theoretical and subjective considerations. Data driven methods include: weights of evidence (Agterberg, Bonham-Carter, & Wright, 1990; Bonham-Carter, Agterberg, & Wright, 1989; Carranza, 2004; Debba, Carranza, Stein, & van der Meer, 2008; Oh & Lee, 2008; Porwal, González-Álvarez, Markwitz, Mccuaig, & Mamuse, 2010; Raines, 1999; S. Xu, Cui, Yang, & Wang, 1992), logistic regression (Agterberg & Bonham-Carter, 2005; Chung, 1978; Chung & Agterberg, 1980; Harris, Zurcher, Stanley, Marlow, & Pan, 2003; Oh & Lee, 2008; Porwal et al., 2010), likelihood ratio (Chung & Fabbri, 1993; Oh & Lee, 2008, 2010), evidence theory (An & Moon, 1993; Carranza, 2009; Carranza & Sadeghi, 2010; Carranza, van Ruitenbeek, Hecker, van der Meijde, & van der Meer, 2008; Chung & Fabbri, 1993), and artificial neural networks (Brown, Gedeon, Groves, & Barnes, 2000; Harris & Pan, 1999; Harris et al., 2003; Nykänen, 2008; Porwal, Carranza, & Hale, 2003; Singer & Kouda, 1996; Skabar, 2005). All methods use known mineral occurrence locations as the independent dataset to train predictive models which are then used to predict prospectivity. Data driven methodologies can only be applied in mature mineral provinces which contain sufficient training data (Lisitsin et al., 2014). Conversely, knowledge driven methods include: fuzzy logic (Cheng & Agterberg, 1999; de Quadros, Koppe, Strieder, & Costa, 2006; Groves et al., 2010; Knox-Robinson & Wyborn, 1997; Lisitsin et al., 2014; M. Moon & An, 1991; Porwal, Carranza, & Hale, 2006) and belief functions (An, Moon, & Bonham-Carter, 1994; Carranza, Woldai, & Chikambwe, 2005; Chung & Fabbri, 1993; Moon, 1990; Tangestani & Moore, 2002; Wright & Bonham-Carter, 1996). Both use a user defined deposit model to evaluate the independent data and predict prospectivity. Knowledge driven methods are commonly applied in immature mineral provinces where there is insufficient training data (Lisitsin et al., 2014).

2.2 Multivariate Simulation

2.2.1 Introduction

Various methodologies exist that enable the use of multivariate data in spatial prediction. Ideally, multiple variables are jointly modeled, consider all secondary data, and reproduce input statistics such as collocated and spatial correlation structure. As the number of variables increase, implementation challenges arise, including an increase in computational costs and workflow complexity. The following is a summary of methods related to this thesis.

2.2.2 Cokriging Framework

A cokriging framework allows primary variables to be jointly modeled while explicitly considering spatial cross correlation between them. This is done by extending the kriging system of equations to account for conditioning by other variables, in addition to its own conditioning data. When using this framework, spatial direct and cross-covariance functions must be defined by a permissible model of coregionalization; typically, the linear model of coregionalization (LMC) is used (Goovaerts, 1997).

2.2.2.1 Linear Model of Coregionalization (LMC)

Consider K number of coregionalized continuous functions $\{z_k(\mathbf{u}), k = 1, ..., K, \forall \mathbf{u} \in A\}$, for all grid locations $\mathbf{u}$ in the domain A, that are also denoted by the vector $\mathbf{Z}(\mathbf{u})$. The LMC consists of a $[K(K+1)]/2$ set of direct and cross-variograms. The LMC assumes that $\mathbf{Z}(\mathbf{u})$ is constructed by a linear combination of $(L+1)$ underlying independent factors $\{Y^l(\mathbf{u}), l = 0, ..., L, \forall \mathbf{u}\}$. These factors are defined by their variogram models (i.e., nested structures) $\{\Gamma(\mathbf{h})^l, \forall l\}$ such that:

$$\gamma(\mathbf{h})_{kk'} = \sum_{l=0}^{L} b^l_{kk'} \cdot \Gamma(\mathbf{h})^l, \quad \forall k, k'$$

The set of nested variogram structures are also denoted by the vector $\mathbf{\Gamma}(\mathbf{h})$ and remain the same for every direct and cross variogram within the LMC. The index of $l = 0$ is reserved for a nugget effect model and may not be used in all cases; in expressions containing $\forall l$, unless it is otherwise stated, it is assumed that this includes the index for the nugget effect model.

The variance contribution parameters $\{b^l_{kk'}, \forall l, k, k'\}$, that are also denoted by the K by K coregionalization matrix $\mathbf{B}^l$, are adjusted to fit the experimental variograms. For the LMC to licit, the $\mathbf{B}^l$ matrices must be positive definite (Journel & Huijbregts, 1978). A more in depth explanation of the LMC is found in Journel and Huijbregts (1978).

2.2.3 Intrinsic Collocated Cokriging

In situations that contain exhaustive secondary data, alternative methodologies are sometimes implemented. A popular methodology considers a single collocated secondary variable at each location being considered one at a time. It implements a Markov model (Almeida, 1993; W. Xu, Tran, Srivastava, & Journel, 1992) that assumes that the collocated secondary data screens the influence of other distal secondary data. It is an attractive methodology as cross-variograms are not a required input parameters; rather, they come from the variogram model of the variable being predicted that is scaled so the sill is equivalent to the correlation coefficient. However, this methodology is known to overestimate the conditional distributions variance, that when used in an sequential Gaussian simulation (SGS) framework, propagates through the simulation (Babak & Deutsch, 2009b).

To avoid variance inflation within a collocated cokriging framework, Babak and Deutsch (2009b) introduced a new methodology called intrinsic collocated cokriging (ICCK). It implements a cokriging framework using an intrinsic model of coregionalization that considers the collocated secondary data at the location being estimated and from locations that contain conditioning primary data. By accounting for these additional secondary variable observations, the variance of the estimated conditional distributions is correctly calculated.

ICCK can consider many secondary data; however, it is computationally expensive to do so as the system of equations becomes more difficult to solve with an increasing amount of secondary variables. This is handled by utilizing this methodology with a super-secondary variable (Babak & Deutsch, 2009a) described below in Section 2.2.4.

This methodology provides a theoretically correct means of considering many secondary data. By implementing simple cokriging with an intrinsic coregionalization model, correlation between primary and secondary data is reproduced (Babak & Deutsch, 2009a).

2.2.4 Super-Secondary Variables

Due to the increase in the system of equations needing to be solved with ICCK in comparison with collocated cokriging methods that implement a Markov assumption, the number of conditioning secondary variables is an important consideration when implementing the methodology. Babak and Deutsch (2009a) introduces a theoretically valid means to merge any number of exhaustive secondary data into a single super-secondary variable for each of the primary variables. The linear relationships between the primary variable and all secondary variables is defined by a correlation coefficient. A brief summary of these calculations is detailed below.

Consider a system containing K number of primary variables and I number of exhaustive secondary variables. The super-secondary variable for all grid locations $\mathbf{u}$ within the domain A is

calculated from the linear summation of weighted secondary values:

$$y_k^{sup}(\mathbf{u}) = \sum_{i=1}^{I} w_{i,k} \cdot y_i(\mathbf{u}), \quad \forall k, \mathbf{u} \in A$$

The weights are calculated following the procedure:

$$\mathbf{C}_{ss} \cdot \boldsymbol{w} = \mathbf{C}_{sp}$$

The above expression denotes the I by I covariance matrix $\mathbf{C}_{ss}$ of only the secondary variables, the I by K matrix of weights $\boldsymbol{w}$, and the I by K covariance matrix $\mathbf{C}_{sp}$ between the primary variables and all of the secondary variables. The K by 1 vector of correlation coefficients $\boldsymbol{\rho}^{sup}$ between the generated super-secondary variables and the primary variables are calculated following the procedure:

$$\boldsymbol{\rho}^{sup} = diag\left\{\sqrt{\left[\mathbf{C}_{sp}^T \cdot \mathbf{C}_{ss}^{-1} \cdot \mathbf{C}_{sp}\right]}\right\}$$

In the above expression, the diagonal of the calculated K by K matrix is used to populate the $\boldsymbol{\rho}^{sup}$ vector.

The procedure summarized above produces K number of super-secondary variables for all grid locations $\mathbf{u}$ within the domain A, each having a corresponding correlation coefficient $\{\rho_k^{sup}, \forall k\}$.

2.2.5 Hierarchical Framework

A framework that jointly predicts many variables was introduced by Almeida and Journel (1994) that performs as well as a cokriging framework without the use of an LMC. It requires the variables be modeled in an hierarchical fashion rather then simultaneously by using previously modeled variables to condition subsequent variables prediction. The framework assumes a multi-Gaussian system and requires that any secondary data be exhaustive. However, Almeida and Journel (1994) postulates that the approach can perform well in cases that depart from a multi-Gaussian assumption.

This framework may be implemented with SGS using ICCK that allows conditioning to any number of exhaustive secondary data and any previously simulated variables; that is made easer by merging all exhaustive conditioning data into a single super-secondary variable. However, implementation is cumbersome as variables are modeled independently and in the defined hierarchy. Therefore, with K number of variables to model, any number of exhaustive secondary variables, and L number of realizations to simulate, $[(K-1) \cdot L + 1]$ super-secondary variables must be generated. No program exists that allows this workflow to be completed with one call. Instead, the simulation program must be called $(K \cdot L)$ number of times while the super-secondary calculation program must be called $[(K-1) \cdot L + 1]$ number of times.

2.3 Decorrelation Techniques

2.3.1 Introduction

The theory of linear transformations such as principle component analysis (PCA) and reverse data sphering (Sphere-R) are discussed in detail by Barnett and Deutsch (2015). The authors also explain the program `decorrelate` that implements both transformations and the back-transformation program `decorrelate_b`. The following is a summary considered important for this thesis that considers the case where K number of input variables with N number of observations from the grid locations $\mathbf{u}$ within the domain A $\{Z_k(\mathbf{u}_n), k = 1, \ldots, K, n = 1, \ldots, N\}$ that is also donated by the vector $\mathbf{Z}(\mathbf{u_n})$.

2.3.2 Principle Component Analysis

PCA is implemented as a data exploration, dimension reduction, or decorrelation tool. PCA is a linear transformation that rotates the data frame of reference, orthogonalizing the variables. K number of input variables with N number of observations are transformed to principle components (PCs) $\{P_k(\mathbf{u}_n), \forall k, n\}$ that demonstrate a correlation structure such that $\{\rho_{kk'} = 0, \forall k \neq k'\}$.

By determining orthogonal axes that explain the greatest amount of variance within the input data, the linear combination required to achieve the diagonal covariance matrix is found. The linear contribution factors—also known as loadings—of each input variable $Z_k(\mathbf{u}_n)$ resulting in $P_k(\mathbf{u}_n)$ depend on the data correlation structure.

2.3.3 Reverse Data Sphering

Sphere-R is a form of data sphering that modifies PCA with two additional operations. Sphere-R spheres (i.e., standardizes) then rotates the variables $P_k(\mathbf{u}_n)$ back onto the basis of the original variables, generating the variables $\{R_k(\mathbf{u}_n), \forall k, n\}$ that are also denoted by the vector $\mathbf{R}(\mathbf{u}_n)$ and referred to as Sphere-R factors (SRs). Due to these additional steps, the mixing of loadings is minimized. As a result, the correlation between the input and output variables $\rho(Z_k(\mathbf{u}_n), R_k(\mathbf{u}_n))$ is maximized when $k = k$ and still 0 when $k \neq k$.

2.3.4 Application Considerations

Due to the rotation performed by Sphere-R, dimension reduction is not possible as all of the SRs contribute appreciably to the variance of the system. Conversely, each subsequent PC calculated using PCA contributes less to the variance of the system. This feature enables dimension reduction, that when applied ignores a portion of the systems variance. Sphere-R is not suited to data exploration.

Chapter 3

Novel Frameworks for Multivariate Simulation with Many Secondary Data

3.1 Introduction

In a multivariate case consisting of equally sampled continuous variables to be modeled (primary variables), a collocated correlation structure (i.e., at a lag distance of 0) can be calculated to represent the linear relationships between variables. Ideally, these relationships are reproduced during spatial prediction. The cokriging framework presented in Section 2.2.2 reproduces input statistics; however, it requires the use of an LMC. A LMC consists of $[K(K+1)]/2$ number of direct and cross variograms, making it arduous to implement as K increases. It has been shown that $K = 7$ number of variables can be reasonably fit (Jewbali, 2009), which may inform a practicable limit.

There are a limited number of methodologies that jointly model multiple continuous variables without the use of an LMC. In a simulation framework, these include:

1. A hierarchical approach introduced by Almeida and Journel (1994) where variables are consecutively modeled in a user defined order. Once there is more than one previously simulated variable, they are transformed into a single super-secondary variable (Babak & Deutsch, 2009b). Each variable is then conditioned by the calculated super-secondary variable using ICCK, introduced by Babak and Deutsch (2009b).
2. The variables being modeled are decorrelated allowing independent modeling of independent factors that are reconstructed into the variables (Luster, 1985).
3. A block lower upper triangular matrix (LU) simulation with an approximate model of coregionalization (Wang & Deutsch, 2009).

The use of these methods in the presence of exhaustively sampled secondary data (secondary variables) adds additional complexities to the modeling procedure. The hierarchical approach considers all continuous exhaustively sampled secondary data. However, it is only appropriate if the multivariate relationships are fully characterized by the multivariate collocated correlation structure (Rossi & Deutsch, 2014).

This chapter investigates and documents two alternatives utilizing secondary variables: first a decorrelation method and secondly, a novel Gaussian-based approach. Both are found to be ineffective in reproducing the input collocated correlation structure and other important statistics. Additional details regarding multivariate simulation are discussed in Section 2.2.

3.2 Generation of Synthetic Data

3.2.1 Introduction

To evaluate each methodologies ability to simulate multivariate data in the presence of exhaustive secondary data, a series of synthetic case studies is a reasonable test. They require datasets that have a known spatial and collocated correlation structure. One approach for generating synthetic data is detailed in this section that uses a specified LMC to simulate synthetic data.

3.2.2 Simulated Synthetic Data with an LMC

As discussed in Section 2.2.2.1, an LMC is derived from fitting a set of experimental variograms to establish the variogram models and the required $\mathbf{B}^l$ matrices. The $\mathbf{B}^l$ matrices could be derived from $(L+1)$ number of matrices containing the coefficients $\{a_k^l, \forall l, k\}$. These matrices $\mathbf{A}^l$ take the form:

$$\mathbf{A}^l = \begin{bmatrix} a_1^l \\ \vdots \\ a_k^l \end{bmatrix}, \quad \forall l$$

The $\mathbf{B}^l$ matrices can then be calculated from:

$$b_{kk'}^l = a_k^l \cdot a_{k'}^l, \quad \forall l, k, k' \tag{3.1}$$

creating a licit (i.e., positive semi-definite) LMC if the following condition is satisfied:

$$\sum_{l=0}^{L} b_{kk}^l = 1, \quad \forall k \tag{3.2}$$

The correlation matrix $\boldsymbol{\rho}$ of the multivariate system is written as:

$$\boldsymbol{\rho} = \begin{bmatrix} \rho_{11} & \cdots & \rho_{K1} \\ \cdots & \ddots & \vdots \\ \rho_{1K} & \cdots & \rho_{KK} \end{bmatrix}$$

and can be calculated knowing that:

$$\rho_{kk'} = \sum_{l=0}^{L} b_{kk'}^l, \quad \forall k, k' \tag{3.3}$$

Based on the procedure above, the $\mathbf{A}^l$ matrices and the pool of nested variogram structures $\{\Gamma(\mathbf{h})^l, l = 1, \ldots, L\}$ are the only input parameters. They can be generated as deemed appropriate; however there are some limitations. When generating the $\mathbf{A}^l$ matrices, the condition from expression (3.2) must be met. To ensure that the generated correlation matrix of the synthetic data is not singular when using this methodology, there needs to be as many nested structures as there are

variables being generated (i.e., $L = K$). Once the $\{\Gamma(\mathbf{h})^l, l = 1, \ldots, L\}$ variograms are defined, L number of independent random variables $\{Y^l(\mathbf{u}), \forall l, \mathbf{u}\}$ with a mean of 0 and a variance of 1 must be unconditionally simulated. For each $\mathbf{Y}^l$ variable simulated, the corresponding $\{\Gamma(\mathbf{h})^l, l = 1, \ldots, L\}$ is used as the input variogram model. Each synthetic variable can be calculated by:

$$Z_k(\mathbf{u}) = m_k + \sum_{l=1}^{L} a_k^l \cdot y^l(\mathbf{u}), \quad \forall k, \mathbf{u} \tag{3.4}$$

With this procedure, a synthetic dataset is generated with spatial and statistical relationships described by the LMC calculated from the generated $\mathbf{A}^l$ matrices and chosen nested structures $\mathbf{\Gamma}(\mathbf{h})$.

3.3 PCA and Sphere-R Transformed Multivariate Data

3.3.1 Introduction

PCA can be used to decorrelate the primary variables, enabling independent modeling of each PC (Luster, 1985). The resulting models are then back-transformed into the original units of the input variables, reintroducing the original collocated correlation structure. When using this methodology, a strong assumption is made that the spatial cross-correlation at a lag distances not equal to 0 are also removed during decorrelation.

In the presence of secondary data, the rotation performed during PCA causes the collocated correlation structure between the output PCs and their input primary variables to not resemble the collocated correlation structure observed prior to the rotation. As such, it is possible that strong correlations that once existed between the primary and secondary variables may change. If this change results in the collocated correlation between a PC and a secondary variable to become less than $|0.20|$, the secondary data is not likely to improve to estimation (Cuba, Babak, & Leuangthong, 2009).

An additional decorrelation technique Sphere-R, is introduced by Barnett and Deutsch (2015) for use in geostatistics. Sphere-R consists of two additional steps once PCA is completed: the PCs are first sphered then rotated back onto the original basis of the input variables, minimizing the mixing effect seen with PCA. Due to the reverse rotation, the collocated correlation between the corresponding input and output variables is maximized (Barnett & Deutsch, 2015), as illustrated in Figure 3.1. The resulting SRs maintain a primary-secondary collocated correlation structure similar to the one found with the input variables (Figure 3.2).

The following study explores the use of PCA and Sphere-R by evaluating the reproduction of the collocated correlation structure of multiple small synthetic datasets, each consisting of three exhaustive primary variables and one exhaustive secondary variable. The datasets are generated using their own synthetic LMC, resulting in a multivariate dataset with known spatial and collocated statistical relationships. The primary variables are decorrelated using PCA and Sphere-R

transformations, simulated independently using unconditional SGS with ICCK that considers the exhaustive secondary variable. Once the models are back-transformed to their original units, the collocated correlation structure of the simulated variables from both methodologies are then checked for reproduction. Additional details regarding PCA and Sphere-R are discussed in Section 2.3.

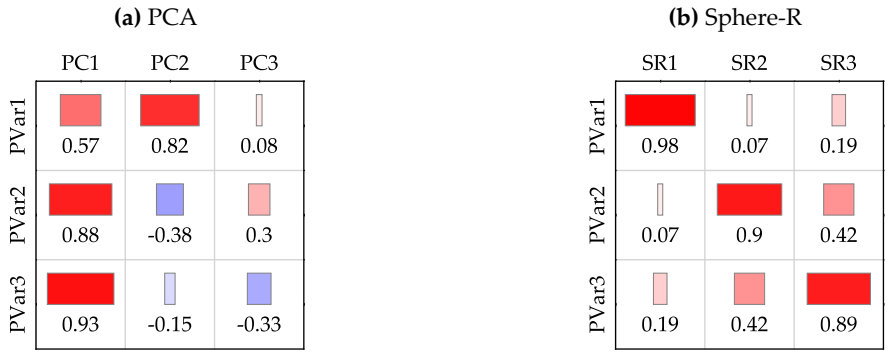

Figure 3.1: Loadings of variables generated from the PCA and Sphere-R transformations.

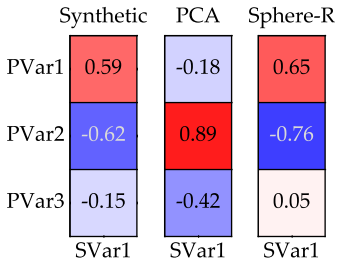

Figure 3.2: The original correlation coefficients between the normal scored primary variables and the normal scored secondary variable and the corresponding correlation coefficients of their PCA and Sphere-R counterparts.

3.3.2 Synthetic Case Study

3.3.2.1 Generation of Synthetic Multivariate Data

To allow predicted models to be checked throughly, synthetic data is generated using a LMC that has known spatial and multivariate relationships. The process outlined in Section 3.2 is used to generate the synthetic data. An example of this process and a description of any necessary parameters is illustrated below.

To test multiple collocated correlation structures, multiple synthetic datasets are generated with four exhaustive variables $K = 4$ using their own synthetic LMC. For this study, it is thought that three primary variables $\{Z_k(\mathbf{u}), k = 1, \ldots, 3, \forall \mathbf{u}\}$ and one secondary variable $\{Z_4(\mathbf{u}), \forall \mathbf{u}\}$ provides adequate dimensionality in addition to the use of 10 unique data sets allows the methodology to be tested sufficiently. A two dimensional (2-D) 200 by 200 cell grid is used for all synthetic datasets along with the same three $L = 3$ nested variogram structures $\{\Gamma(\mathbf{h})^l, L = 1, \ldots, 3\}$ with no nugget

effect Γ^0. Spherical variogram structures are used for the three nested variogram structures with ranges of 16, 32, and 64.

The $\mathbf{A}^l$ coefficients are generated by first populating the $\mathbf{A}^1$ and $\mathbf{A}^3$ matrices with random values between -0.75 and 0.75. The $\mathbf{A}^2$ matrix is then calculated by the calculation:

$$a_k^2 = 1 - a_k^1 - a_k^3, \quad \forall, k$$

The coefficients are permissible if the following condition is satisfied:

$$\sum_{l=1}^{L} (a_k^l \cdot a_k^l) = 1, \quad \forall k$$

If the condition passed, the $\mathbf{A}^l$ matrices is used; otherwise, the procedure is restarted. A positive semi-definite LMC is then calculated from the generated $\mathbf{A}^l$ matrices by completing the procedure outlined in Section 3.2.

The correlation matrix of each LMC generated is then checked so that it meets the following set of conditions:

1. The generated correlation coefficients must have a minimum variance of $|0.20|$.
2. No correlation coefficient can be greater than $|0.75|$.
3. No correlation coefficient can be within $|0.10|$ of another correlation coefficient.
4. No correlation coefficient between primary variables and a secondary variable can be less than $|0.10|$.
5. No primary-secondary correlation coefficient can be within $|0.20|$ of another primary-secondary correlation coefficients.

If these conditions are met, the generated LMC is used to generate a synthetic dataset; otherwise, the LMC is discarded and the process is repeated. Conditions (1), (3), and (5) ensures that the produced correlation matrix does not contain redundant information (e.g., $\rho_{14} = 0.4, \rho_{24} = 0.4$). Condition (2) ensures that no unrealistically informative information is generated (e.g., $\rho_{14} = -0.95$). Condition (4) ensures that no meaningless information is generated (e.g., $\rho_{14} = 0.05$).

Example of Synthetic Data Generation

The following is an example of a set of $\mathbf{A}^l$ matrices and their resulting LMC, calculated contribution matrices $\mathbf{B}^l$, and the synthetic multivariate dataset derived from it. The nested variogram structures outlined in Section 3.3.2.1 are used. The values illustrated below are from one of the synthetic datasets used for this case study.

The $\mathbf{A}^l$ matrices are first generated

$$\mathbf{A}^1 = \begin{bmatrix} -0.47 \\ 0.61 \\ 0.65 \\ -0.70 \end{bmatrix}, \ \mathbf{A}^2 = \begin{bmatrix} 0.84 \\ 0.33 \\ 0.74 \\ 0.47 \end{bmatrix}, \ \mathbf{A}^3 = \begin{bmatrix} 0.28 \\ 0.72 \\ 0.14 \\ -0.54 \end{bmatrix}$$

After which, the symmetric $\mathbf{B}^l$ matrices are calculated using the expression (3.1) with the derived $\mathbf{A}^l$ matrices shown above. As an example, the calculation of $\mathbf{B}^1$ results in:

$$\mathbf{B}_1 = \begin{bmatrix} 0.22 & -0.29 & -0.31 & 0.33 \\ -0.29 & 0.37 & 0.4 & -0.43 \\ -0.31 & 0.4 & 0.43 & -0.46 \\ 0.33 & -0.43 & -0.46 & 0.49 \end{bmatrix}$$

Using the calculated $\mathbf{B}^l$ matrices, the correlation matrix $\boldsymbol{\rho}$ are calculated using the expression (3.3) and is illustrated in Figure 3.3a.

Direct and cross variograms are defined by the generated LMC. The variograms of the factors are required. The covariance contributions for each of the nested variogram structures within $\boldsymbol{\Gamma}(\mathbf{h})$ are contained within the $\mathbf{B}^l$ matrices. In this example case, the direct variograms are as follows:

$$\gamma(\mathbf{h})_{11} = 0.22 \cdot \text{Sph}(\mathbf{h})_{\text{a=16}} + 0.70 \cdot \text{Sph}(\mathbf{h})_{\text{a=32}} + 0.08 \cdot \text{Sph}(\mathbf{h})_{\text{a=64}}$$
$$\gamma(\mathbf{h})_{22} = 0.37 \cdot \text{Sph}(\mathbf{h})_{\text{a=16}} + 0.11 \cdot \text{Sph}(\mathbf{h})_{\text{a=32}} + 0.52 \cdot \text{Sph}(\mathbf{h})_{\text{a=64}}$$
$$\gamma(\mathbf{h})_{33} = 0.43 \cdot \text{Sph}(\mathbf{h})_{\text{a=16}} + 0.55 \cdot \text{Sph}(\mathbf{h})_{\text{a=32}} + 0.02 \cdot \text{Sph}(\mathbf{h})_{\text{a=64}}$$
$$\gamma(\mathbf{h})_{44} = 0.49 \cdot \text{Sph}(\mathbf{h})_{\text{a=16}} + 0.22 \cdot \text{Sph}(\mathbf{h})_{\text{a=32}} + 0.29 \cdot \text{Sph}(\mathbf{h})_{\text{a=64}}$$

Synthetic data are then generated using the expression (3.4) and is summarized in Figure 3.4. The correlation matrix calculated from the synthetic data (Figure 3.3b) shows little change to the LMCs theoretical correlation matrix. For illustration purposes, a heat map of two of the generated

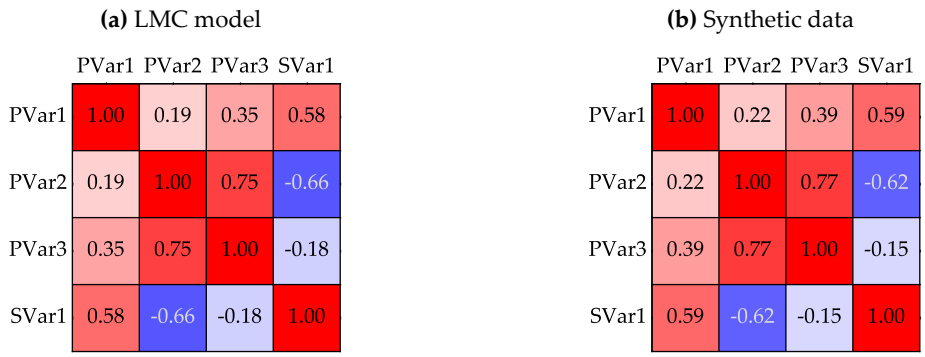

Figure 3.3: Correlation matrices of the input LMC model and the resulting synthetic data.

variables, PVar1 (i.e., $k = 1$) and SVar1 (i.e., $k = 4$), is shown in Figure 3.5. Their calculated experimental variograms are plotting along side their theoretical variogram determined by the LMC are shown in Figure 3.6.

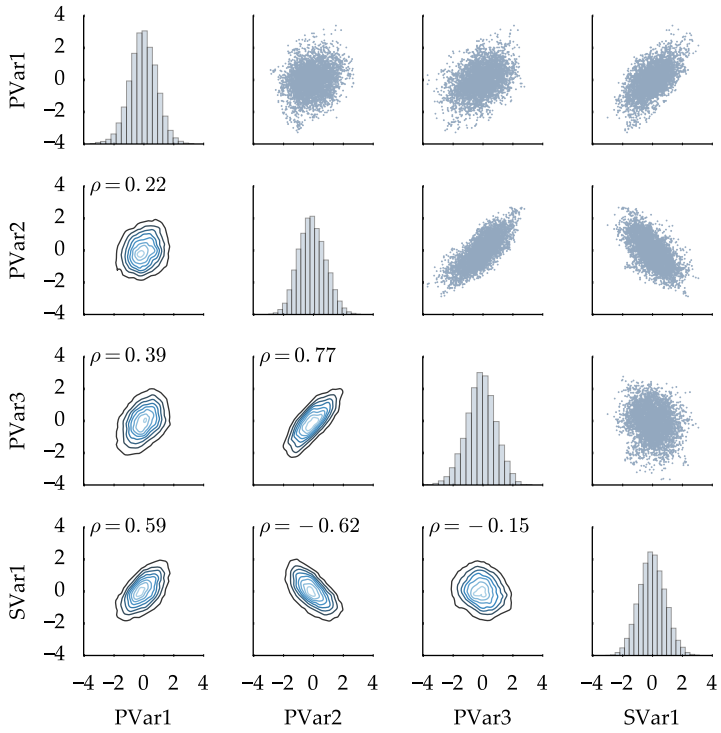

Figure 3.4: Multivariate joint density, histogram, and scatter plot matrix of the synthetic data generated.

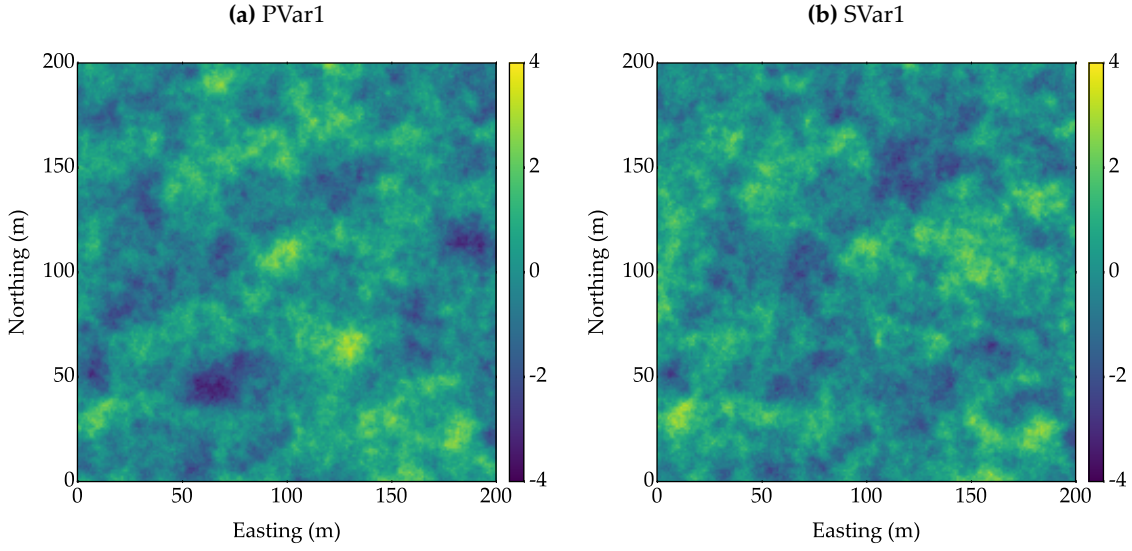

Figure 3.5: Example heatmaps of generated 2-D synthetic data in normal score space.

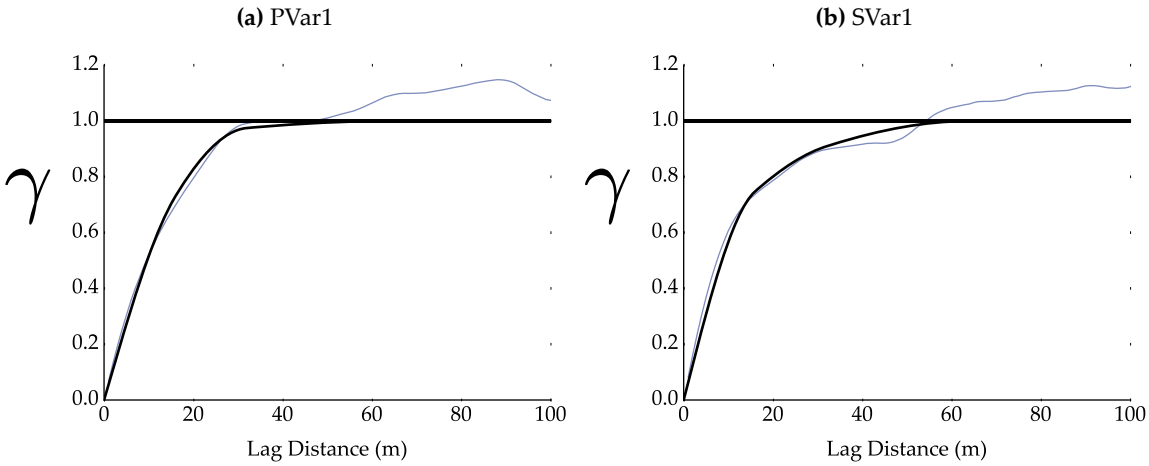

Figure 3.6: Experimental variograms of the synthetic data (blue) at a azimuth of 90° and the theoretical input LCM model variogram (black).

Summary of Synthetic Data Generated

Figure 3.7 illustrates the range of correlation coefficients generated and subsequently tested by the 10 synthetic datasets.

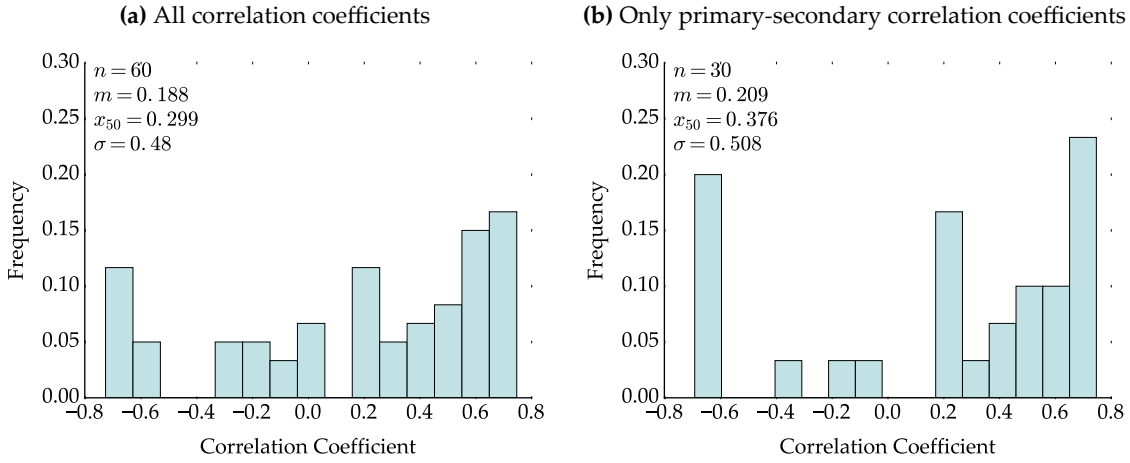

Figure 3.7: Histograms of the correlation coefficients within the 10 generated correlation matrices.

3.3.2.2 Methodology

SGS with ICCK is run unconditionally for each of the 10 synthetic datasets. The variograms for the primary variables are specified by the LMC that generate them. The input correlation matrix is calculated from the decorrelated synthetic exhaustive data once it is normal score transformed. The three primary variables are then simulated. For the purpose of this study, one hundred realizations are used to calculate summaries of local uncertainty. The processes is summarized as follows:

1. Normal score transform the primary and secondary variables
2. Calculate the correlation matrix of the transformed synthetic dataset

3. Decorrelate only the three primary variables using PCA
4. Normal score the generated decorrelated variables
5. Calculate the correlation between the generated PCA variables and the secondary variable, which is still in normal scored space
6. Simulate one hundred realizations of the generated PCs variables using SGS with ICCK that considers the secondary variable and the corresponding correlation coefficient from step 5
7. Complete the following back-transformations in order: normal score back-transform, PCA back-transform, and normal score back-transform
8. Calculate the correlation matrix of the three simulated primary variables in original units with the secondary variable in original units
9. Repeat steps 3 to 8 using Sphere-R instead of PCA to decorrelate the primary variables

3.3.2.3 Results

The PCA methodology produced an average absolute error in primary-primary correlation reproduction of 0.103 (Figure 3.8a) and a mean squared error (MSE) of 0.022 (Figure 3.9a). The Sphere-R methodology has an average absolute error in primary-primary correlation reproduction of 0.304 (Figure 3.8b) and a MSE of 0.192 (Figure 3.9b).

The PCA methodology has an average absolute error in primary-secondary correlation reproduction of 0.037 (Figure 3.10a) and a MSE of 0.002 (Figure 3.11a). The Sphere-R methodology has an average absolute error in primary-primary correlation reproduction of 0.068 (Figure 3.10b) and a MSE of 0.019 (Figure 3.11b).

It is thought by the author that a correlation reproduction error less then $|0.20|$ is reasonable. Figure 3.12 illustrates the percentage of errors greater than this threshold, showing that the PCA methodology does not reproduce primary-primary relationships in 16.7% of the cases while the Sphere-R methodology does not in 56.7% of the cases. The PCA methodology reproduced primary-secondary correlations within this threshold in all cases, while the Sphere-R methodology does not in 6.7% of the cases.

While the simulations are run unconditionally, the variables being modeled are conditioned to secondary data using ICCK. There is an increased amount of variability. While the effect is negligible in regards to histogram reproduction, the secondary data does have a noticeable effect on the realizations variograms. This issue may be mitigated with primary conditioning data.

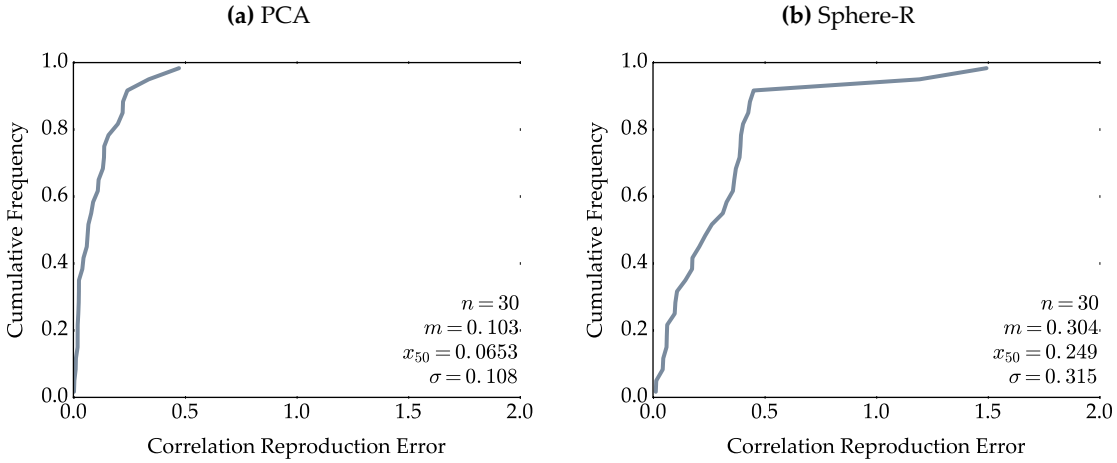

Figure 3.8: Distributions of absolute correlation reproduction errors for primary-primary correlations from the PCA and Sphere-R workflows.

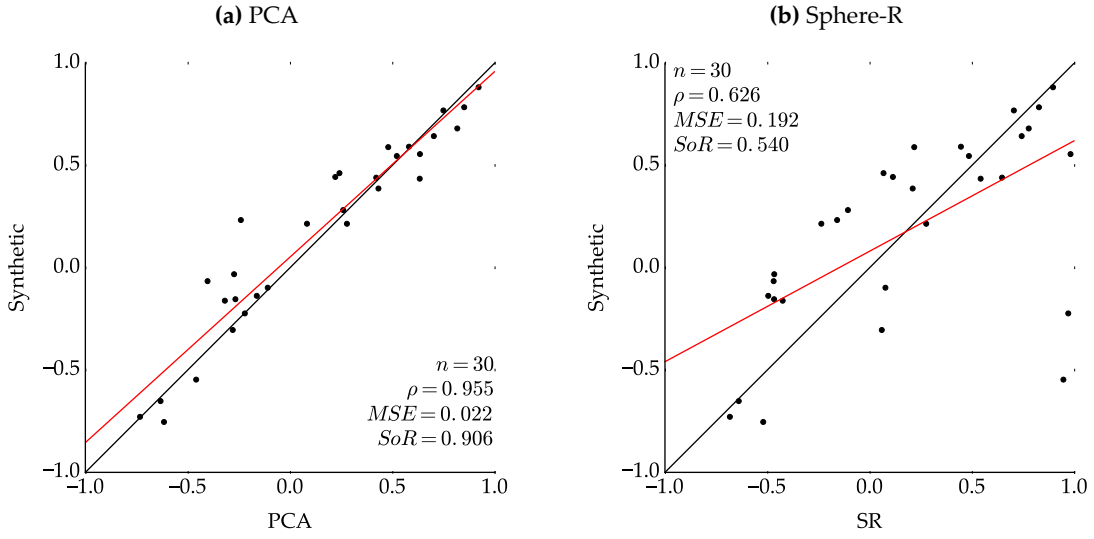

Figure 3.9: Cross validation of the true correlation vs. the produced primary-primary correlations from the PCA (left) and Sphere-R (right) workflows with a fitted regression line (red).

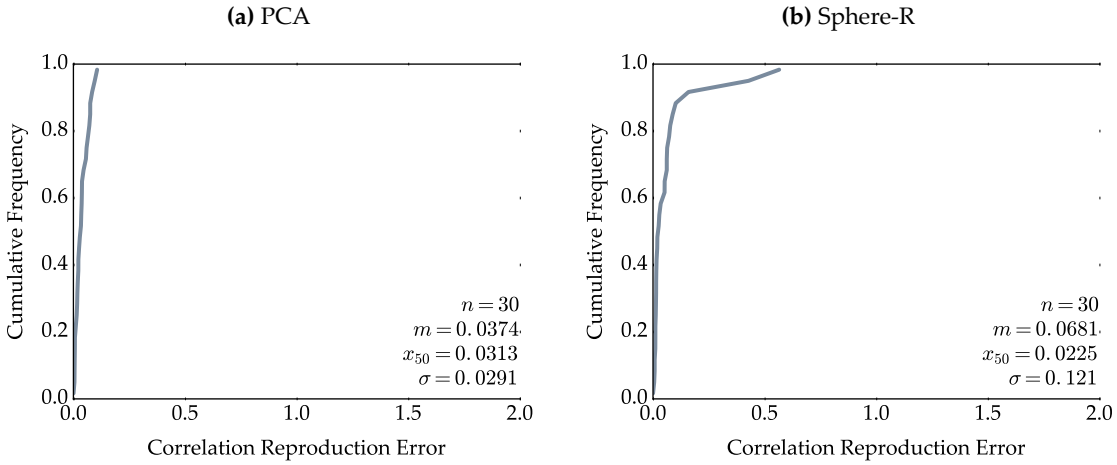

Figure 3.10: Distributions of absolute correlation reproduction errors for primary-secondary correlations from the PCA and Sphere-R workflows.

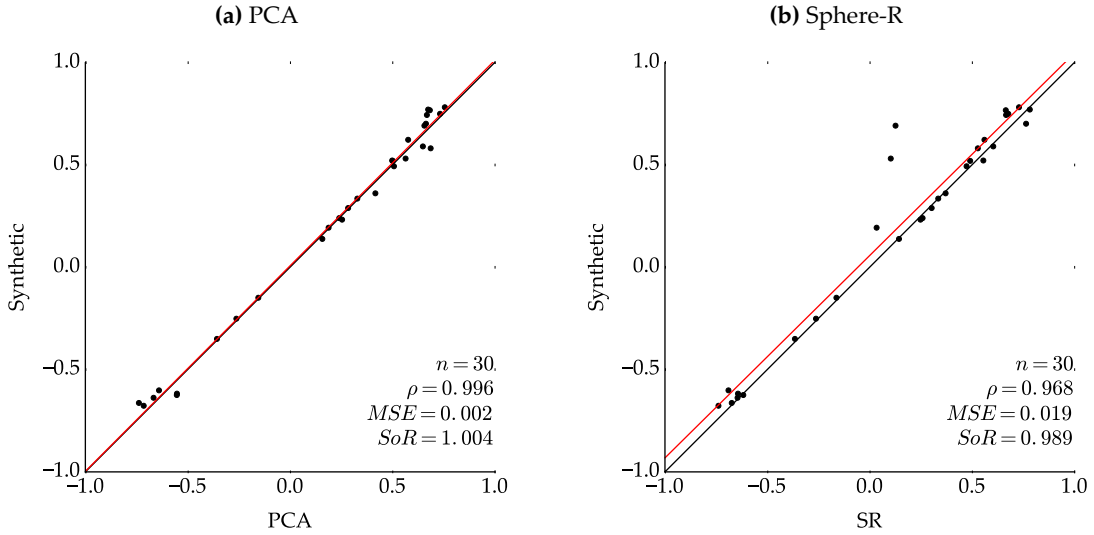

Figure 3.11: Cross validation of the true correlation vs. the produced primary-secondary correlations from the PCA and Sphere-R workflows with a fitted regression line (red).

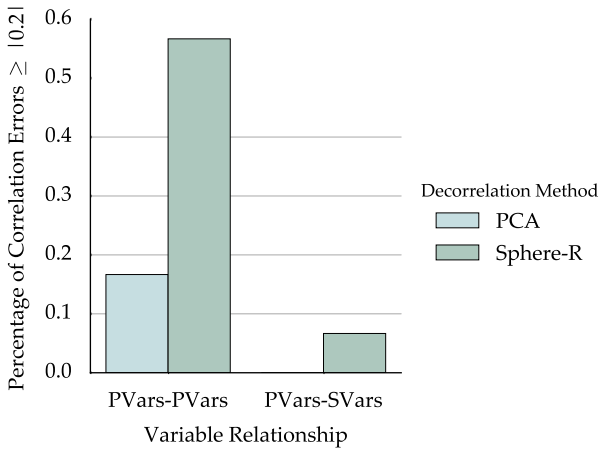

Figure 3.12: Joint bar plot depicting the percentage of correlation reproduction errors above a value of $|0.20|$

3.3.3 Conclusion

One of the benefits thought to exist for PCA and Sphere-R methods was that they would reproduce primary-primary correlations; however, this was not observed in the study. The correlation between the decorrelated variables and the conditioning secondary data introduces correlation between the once decorrelated variables during simulation. After which, the correlation introduced together with the back transformation process fails to properly reproduce original primary-primary correlations. An example of this is illustrated in Figure 3.13; a correlation of -0.49 between SR1 and SR2 is introduced, that when back transformed to original units, becomes -0.24 and not the original 0.22. As both PCA and Sphere-R are linear transformations, the correlation induced by ICCK on the decorrelated variables is maintained during the back-transformation.

While the PCA methodology appears to have performed better than the Sphere-R methodology in regards to producing primary-secondary collocated correlation structure, it is believed this is a product of the different rotations performed by the two methodologies and is not a theoretical advantage. As the PCA transformation does not replicate the original primary-secondary collocated correlation structure, it is believed that this can be advantageous in counteracting the induced correlation during simulation, creating an illusion of better performance. As this unforeseen benefit cannot be controlled, it should not provide confidence in the methodology.

Some cases exhibited erratic behavior, causing correlation coefficients to become nearly 1 (Figure 3.14). This effect is thought to occur when the degrees of freedom is reduced due to redundancy in the collocated correlation structure. This issue occurred in one of the Sphere-R methodology cases, skewing the results.

As the collocated correlation reproduction from both PCA and Sphere-R methods is unpredictable and erratic, this decorrelation framework in the presence of exhaustive secondary data is not robust enough to use with confidence.

(a) Synthetic data in original units

	PVar1	PVar2	PVar3	SVar1
PVar1	1.00	0.22	0.39	0.59
PVar2	0.22	1.00	0.77	-0.62
PVar3	0.39	0.77	1.00	-0.15
SVar1	0.59	-0.62	-0.15	1.00

(b) Normal scored SR

	SR1	SR2	SR3	SVar1
SR1	1.00	0.00	0.00	0.65
SR2	0.00	1.00	0.00	-0.76
SR3	0.00	0.00	1.00	0.05
SVar1	0.65	-0.76	0.05	1.00

(c) Simulated SRs in normal score space

	SR1	SR2	SR3	SVar1
SR1	1.00	-0.49	0.03	0.63
SR2	-0.49	1.00	-0.03	-0.75
SR3	0.03	-0.03	1.00	-0.01
SVar1	0.63	-0.75	-0.01	1.00

(d) Simulated variables in original units

	PVar1	PVar2	PVar3	SVar1
PVar1	1.00	-0.24	0.21	0.60
PVar2	-0.24	1.00	0.71	-0.64
PVar3	0.21	0.71	1.00	-0.16
SVar1	0.60	-0.64	-0.16	1.00

Figure 3.13: Progression of the correlation matrices through the Sphere-R workflow.

(a) Synthetic data in original units

	PVar1	PVar2	PVar3	SVar1
PVar1	1.00	-0.22	0.56	0.69
PVar2	-0.22	1.00	-0.55	0.53
PVar3	0.56	-0.55	1.00	0.19
SVar1	0.69	0.53	0.19	1.00

(b) Simulated variables in original units

	PVar1	PVar2	PVar3	SVar1
PVar1	1.00	0.97	0.98	0.13
PVar2	0.97	1.00	0.95	0.10
PVar3	0.98	0.95	1.00	0.03
SVar1	0.13	0.10	0.03	1.00

Figure 3.14: Correlation matrices in original units of the synthetic data and correlation matrix produced by the Sphere-R method.

3.4 Collocated Cholesky Cosimulation

3.4.1 Introduction

A novel multivariate modeling method that operates within a Gaussian-based simulation framework is introduced and developed. The goal is a straightforward technique aiming to reproduce the collocated correlation structure of multivariate data in the presence of exhaustively sampled secondary data, consider spatial conditioning, and honor measured data.

This methodology modifies the direct cosimulation framework (Verly, 1993) where all primary variables are jointly and independently simulated using SGS with ICCK, that considers the exhaustive secondary data. This framework correlates the deviates that sample the conditional distributes

at each locations aiming to reproduce the input collocated correlation structure. A summary of this framework with the additional procedure is as follows:

1. Normal score transform the primary and secondary variables. A multiGaussian system is now assumed.
2. Calculate the collocated secondary-primary covariance matrix.
3. Decompose the covariance matrix using Cholesky decomposition.
4. For each realization, complete the following:
    a. Define a random path, visiting all grid nodes once.
    b. At each node visited, complete the following:
        i. Independently estimate the conditional mean and variance of each primary variable using a simple kriging framework, that is conditioned to measured data as well as previously simulated values.
        ii. Correlate deviates used to sample the distributions by:
            A. Seed the secondary variables at the current node and calculate their corresponding independent standard normal deviates.
            B. Randomly draw the primary variables independent standard normal deviates.
            C. Multiply the L matrix by the independent standard normal deviates vector, reproducing the seeded secondary data along with unconditional primary values.
        iii. Sample the conditional distributes using the previously correlated deviates.
        iv. Add the simulated values to the set of conditioning data.
        v. Continue to the next unsampled node.

The following section describes how this proposal is implemented, outlines the need and implementation of ad hoc corrections, and presents a synthetic case study. For ease of communication, the proposed method is be referred to as collocated Cholesky cosimulation (CCC). In the end, the method is also not considered a robust and practical methodology.

3.4.2 Methodology

Consider K number of interdependent random functions (RFs) to be modeled $\{Z_{p,k}(\mathbf{u}), k = 1, \ldots, K, \forall \mathbf{u} \in A\}$ for all grid locations $\mathbf{u}$ within the domain A, that is also be denoted by the vector $\mathbf{Z}_p(\mathbf{u})$. The index p identifies the variables being modeled, that are also referred to as primary variables. N number of observations exist for each RF within $\mathbf{Z}_p(\mathbf{u})$ $\{z_{p,k}(\mathbf{u}_n), k = 1, \ldots, K, n = 1, \ldots, N\}$; that is also be denoted by the vector $\mathbf{Z}_p(\mathbf{u}_n)$. If $\mathbf{Z}_p(\mathbf{u}_n)$ is not equally sampled, imputation is be required.

Each random variable (RV) within $\mathbf{Z}_p(\mathbf{u}_n)$ must be normal scored transformed, such that:

$$y_{p,k}(\mathbf{u}_n) = G^{-1}\left(F_{p,k}\left(z_{p,k}(\mathbf{u}_n)\right)\right), \quad k = 1, \ldots, K, n = 1, \ldots, N$$

The results are denoted by the vector $\mathbf{Y}_p(\mathbf{u}_n)$.

In addition, I number of interdependent exhaustively sampled RFs exist $\{x_{s,i}(\mathbf{u}), i = 1, \ldots, I, \forall \mathbf{u} \in A\}$ for all grid locations $\mathbf{u}$ within the domain A, that are also denoted by the vector $\mathbf{X}_s(\mathbf{u})$. The index s identifies the variables used to condition the models of $\mathbf{Z}_p(\mathbf{u})$, that are also referred to as secondary variables. Each RV within $\mathbf{X}_s(\mathbf{u})$ must be normal scored transformed, such that:

$$y_{s,i}(\mathbf{u}) = G^{-1}\left(F_{s,i}\left(x_{s,i}(\mathbf{u})\right)\right), \quad i = 1, \ldots, I, \forall \mathbf{u} \in A$$

The results are also donated by the vector $\mathbf{Y}_s(\mathbf{u})$. The collocated values of $\mathbf{Y}_s(\mathbf{u})$ found at the measured locations of $\mathbf{Y}_p(\mathbf{u}_n)$ $\{y_{s,i}(\mathbf{u}_n), i = 1, \ldots, I, n = 1, \ldots, N\}$ is also donated by the vector $\mathbf{Y}_s(\mathbf{u}_n)$. At this point, the multivariate system is assumed to be multiGaussian.

A $(I + K) \times (I + K)$ covariance matrix $\mathbf{C}$ is computed from the collocated secondary $\mathbf{Y}_s(\mathbf{u}_n)$ and primary $\mathbf{Y}_p(\mathbf{u}_n)$ variables such that:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{ss} & \mathbf{C}_{sp} \\ \mathbf{C}_{ps} & \mathbf{C}_{pp} \end{bmatrix}$$

The above $\mathbf{C}$ matrix contains four sub-matrices:

1. $\mathbf{C}_{ss}$ denotes the $(I \times I)$ covariance matrix of $\mathbf{Y}_s(\mathbf{u}_n)$.
2. $\mathbf{C}_{pp}$ denotes the $(K \times K)$ covariance matrix of $\mathbf{Y}_p(\mathbf{u}_n)$.
3. $\mathbf{C}_{sp} = \mathbf{C}_{ps}^T$ that denotes the $(I \times K)$ cross-covariance matrix between the two vectors $\mathbf{Y}_s(\mathbf{u}_n)$ and $\mathbf{Y}_p(\mathbf{u}_n)$.

If $\mathbf{C}$ is not a positive definite symmetric matrix due to different data used to compute the matricides, it needs to be corrected to become such. The $\mathbf{C}$ matrix then undergoes Cholesky decomposition such that:

$$\mathbf{C} = \mathbf{L}\mathbf{L}^T = \begin{bmatrix} \mathbf{L}_{ss} & 0 \\ \mathbf{L}_{ps} & \mathbf{L}_{pp} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{L}_{ss} & 0 \\ \mathbf{L}_{ps} & \mathbf{L}_{pp} \end{bmatrix}^T \tag{3.5}$$

The produced lower matrix $\mathbf{L}$ is required during the simulation process and consists of 3 submatrices:

1. $\mathbf{L}_{ss}$ denotes the decomposed $(I \times I)$ lower matrix of $\mathbf{Y}_s(\mathbf{u}_n)$.
2. $\mathbf{L}_{pp}$ denotes the portion of the $\mathbf{L}$ matrix that occupies the lower $(K \times K)$ matrix location but is not equivalent to the decomposed $(K \times K)$ lower matrix of $\mathbf{Y}_p(\mathbf{u}_n)$.
3. $\mathbf{L}_{ps}$ denotes the portion of the L matrix that occupies the lower $(K \times I)$ matrix location but is a full valued $(K \times I)$ matrix.

Unconditional simulation of the K number of primary variables at all grid locations is run generating L number of realizations. A joint $\Big(\big(I+K\big)\times 1\Big)$ vector $\mathbf{y}(\mathbf{u})$ is generated, consisting of a $(I\times 1)$ vector $\mathbf{y}_s(\mathbf{u})$of random values $\{y^l_{s,i}(\mathbf{u}), i=1,\ldots,I, l=1,\ldots,L, \forall \mathbf{u}\in A\}$ and a $(K\times 1)$ vector $\mathbf{y}_p(\mathbf{u})$ of random values $\{y^l_{p,k}(\mathbf{u}), k=1,\ldots,K, l=1,\ldots,L, \forall \mathbf{u}\in A\}$ following the procedure:

$$\mathbf{y}(\mathbf{u}) = \mathbf{L}\cdot\boldsymbol{\omega}(\mathbf{u}) \tag{3.6}$$

that can be expanded as:

$$\begin{bmatrix}\mathbf{y}_s(\mathbf{u})\\ \mathbf{y}_p(\mathbf{u})\end{bmatrix} = \begin{bmatrix}\mathbf{L}_{ss} & 0\\ \mathbf{L}_{ps} & \mathbf{L}_{pp}\end{bmatrix}\cdot\begin{bmatrix}\boldsymbol{\omega}_s(\mathbf{u})\\ \boldsymbol{\omega}_p(\mathbf{u})\end{bmatrix} = \begin{bmatrix}\mathbf{L}_{ss} & 0\\ \mathbf{L}_{ps} & \mathbf{L}_{pp}\end{bmatrix}\cdot\begin{bmatrix}\omega^l_{s,1}(\mathbf{u})\\ \vdots\\ \omega^l_{s,i}(\mathbf{u})\\ \omega^l_{p,1}(\mathbf{u})\\ \vdots\\ \omega^l_{p,k}(\mathbf{u})\end{bmatrix} = \begin{bmatrix}y^l_{s,1}(\mathbf{u})\\ \vdots\\ y^l_{s,i}(\mathbf{u})\\ y^l_{p,1}(\mathbf{u})\\ \vdots\\ y^l_{p,k}(\mathbf{u})\end{bmatrix}, \quad \begin{array}{l} i=1,\ldots,I,\\ k=1,\ldots,K,\\ l=1,\ldots,L,\\ \forall \mathbf{u}\in A\end{array}$$

The $\Big(\big(I+K\big)\times 1\Big)$ vector $\boldsymbol{\omega}(\mathbf{u})$ denotes a set of generated independent standard normal deviates consisting of a $(I\times 1)$ vector $\boldsymbol{\omega}_s(\mathbf{u})$ that contains the values $\{\omega_{s,i}(\mathbf{u}), i=1,\ldots,I, \forall \mathbf{u}\in A\}$ and a $(K\times 1)$ vector $\boldsymbol{\omega}_p(\mathbf{u})$ consisting of the values $\{\omega_{p,k}(\mathbf{u}), k=1,\ldots,K, \forall \mathbf{u}\in A\}$.

As $\{\mathbf{y}_s(\mathbf{u}) = \mathbf{Y}_s(\mathbf{u}), \forall \mathbf{u}\in A\}$, the first I values of $\mathbf{y}(\mathbf{u})$ do not need to be simulated as they are known; however they need to be reproduced. This is accomplished by seeding $\mathbf{Y}_s(\mathbf{u})$ into $\mathbf{y}_s(\mathbf{u})$, allowing $\boldsymbol{\omega}_s(\mathbf{u})$ to be calculated, facilitating the reproduction of $\mathbf{Y}_s(\mathbf{u})$. For example:

$$\omega^1_{s,1}(\mathbf{u}) = \frac{Y_{s,1}(\mathbf{u})}{\mathbf{L}_{ss}}, \quad \forall \mathbf{u}\in A$$

The remaining values of $\boldsymbol{\omega}_s(\mathbf{u})$ are calculated recursively. After which, $\boldsymbol{\omega}_s(\mathbf{u})$ is calculated $\boldsymbol{\omega}_p(\mathbf{u})$ with randomly derived independent standard normal deviates. After which, $\mathbf{y}_p(\mathbf{u})$ is calculated, that are unconditional standard Gaussian values that have the same correlation structure of $\mathbf{C}_{pp}$. In addition, the calculated $\mathbf{y}_p(\mathbf{u})$ values have the same cross-correlation structure of $\mathbf{C}_{sp}$ with $\mathbf{Y}_s(\mathbf{u})$.

There is still the task of conditioning the simulation to $\mathbf{Y}_p(\mathbf{u}_n)$ and previously simulated values. Conditioning is accomplished at each location being simulated $\{\mathbf{u}', \forall \mathbf{u}'\in A\}$ by first calculating an estimate through simple kriging independent to the other primary variables for each location in Gaussian space, conditioned to $\mathbf{Y}_p(\mathbf{u}_n)$ and previously simulated values. The calculated simple kriging mean vector $\mathbf{y}^{sk}_p(\mathbf{u}')$ consisting of the values $\{y^{sk,l}_{p,k}(\mathbf{u}'), k=1,\ldots,K, l=1,\ldots,L, \forall \mathbf{u}'\in A\}$ and simple kriging variance vector $\boldsymbol{\sigma}^{sk}_p(\mathbf{u}')$ consisting of the values $\{\sigma^{sk,l}_{p,k}(\mathbf{u}'), k=1,\ldots,K, l=1,\ldots,L, \forall \mathbf{u}'\in A\}$ are used in conjunction with $\mathbf{y}_p(\mathbf{u}')$ to generate a conditioned simulated vector $\mathbf{y}^{sim}_p(\mathbf{u}')$ consisting of the values $\{y^{sim,l}_{p,k}(\mathbf{u}'), k=1,\ldots,K, l=1,\ldots,L, \forall \mathbf{u}'\in A\}$ following the procedure:

$$\mathbf{y}^{sim}_p(\mathbf{u}') = \mathbf{y}_p(\mathbf{u}')\cdot\boldsymbol{\sigma}^{sk}_p(\mathbf{u}') + \mathbf{y}^{sk}_p(\mathbf{u}') \tag{3.7}$$

that can be expanded as:

$$\begin{bmatrix} y_{p,1}^{sim,l}(\mathbf{u}') \\ \vdots \\ y_{p,k}^{sim,l}(\mathbf{u}') \end{bmatrix} = \begin{bmatrix} y_{p,1}^{l}(\mathbf{u}') \\ \vdots \\ y_{p,k}^{l}(\mathbf{u}') \end{bmatrix} \cdot \begin{bmatrix} \sigma_{p,1}^{sk,l}(\mathbf{u}') \\ \vdots \\ \sigma_{p,k}^{sk,l}(\mathbf{u}') \end{bmatrix} + \begin{bmatrix} y_{p,1}^{sk,l}(\mathbf{u}') \\ \vdots \\ y_{p,k}^{sk,l}(\mathbf{u}') \end{bmatrix}, \quad \begin{array}{l} k = 1, \ldots, K, l = 1, \ldots, L, \\ \forall \mathbf{u}' \in A \end{array}$$

The above conditioning step does not simultaneously consider both the collocated multivariate correlation structure and spatial conditioning.

3.4.3 Initial Testing

The proposed CCC methodology is implemented within a modified version of the latest ultimate sequential Gaussian simulation (`USGSIM`) program released by Manchuk and Deutsch (2015). Initial testing revealed issues with collocated correlation reproduction. The following illustrates these findings in a series of case studies used to isolate the source of the problem.

3.4.3.1 Pure Nugget Effect Case Study

To evaluate if CCC can properly reproduce input statistics in a non-spatial case, a test is conducted consisting of three primary variables and two exhaustive secondary variables. The program is run unconditionally with three primary variable input variograms consisting of a pure nugget effect and synthetically generated secondary data. One hundred realizations are simulated using the input correlation matrix depicted in Figure 3.15. The resulting realizations reproduced the input direct variograms and histograms perfectly. The input and output collocated correlation structure is the same to two decimal places.

	PVar1	PVar2	PVar3	SVar1	SVar2
PVar1	1.00	0.40	-0.00	0.37	-0.41
PVar2	0.40	1.00	-0.04	0.61	-0.12
PVar3	-0.00	-0.04	1.00	0.30	-0.35
SVar1	0.37	0.61	0.30	1.00	0.31
SVar2	-0.41	-0.12	-0.35	0.31	1.00

Figure 3.15: Correlation matrix in normal score space used in CCC.

3.4.3.2 Intrinsic Variogram Model Case Study

To evaluate if CCC can properly reproduce input statistics in a simplistic spatial case, a test is conducted consisting of three primary variables and two exhaustive secondary variables. The program

is run unconditionally with the same isotropic semi-variogram for all three primary variables:

$$\gamma(\mathbf{h}) = 0.1 + 0.9 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

Synthetically generated secondary data is used to condition one hundred realizations using the input correlation matrix depicted in Figure 3.15. The correlation matrix for the simulated variables is illustrated in Figure 3.16a and shows considerable difference in primary-primary correlation reproduction from the input correlation structure (Figure 3.16b). Variance inflation is observed for all three primary variables (Figure 3.17). The input variograms are not reproduced due to the observed inflation (Figure 3.17).

(a) Simulated correlation matrix

	PVar1	PVar2	PVar3	SVar1	SVar2
PVar1	1.00	0.78	0.73	0.51	-0.50
PVar2	0.78	1.00	0.64	0.79	-0.12
PVar3	0.73	0.64	1.00	0.46	-0.50
SVar1	0.51	0.79	0.46	1.00	0.31
SVar2	-0.50	-0.12	-0.50	0.31	1.00

(b) Difference between input and output

	PVar1	PVar2	PVar3	SVar1	SVar2
PVar1	0.00	0.38	0.73	0.14	-0.09
PVar2	0.38	0.00	0.68	0.17	0.00
PVar3	0.73	0.68	0.00	0.16	-0.15
SVar1	0.14	0.17	0.16	0.00	-0.00
SVar2	-0.09	0.00	-0.15	-0.00	0.00

Figure 3.16: The resulting correlation matrix of all realizations in normal scored space and its difference to the input correlation matrix in intrinsic case study.

3.4.3.3 Conclusion

Once spatial correlation of the variables being simulated is considered, collocated correlation structure is not reproduced. Variance inflation is an issue as well and in some cases it is extreme. It appears as though an increase in magnitude of a primary-secondary variable correlation coefficient appears to increase the errors observed.

To ensure that no programing errors were made all values used in intermediate calculations—expressions (3.5, 3.6, and 3.7)—were exported and checked. All calculations were being made correctly. A theoretical fix or empirical corrections are needed to fix the output correlation matrix and variance inflation.

3.4.4 Corrections

In an attempt to correct variance inflation and errors in the reproduction of the collocated primary-primary correlation matrix, two empirical fixes are implemented within CCC. The algorithm simultaneously adjusts each of the corrections for each variable being simulated in an iterative tuning fashion. For a single tuning run, 10000 nodes are simulated, checked for errors, and if any exist a

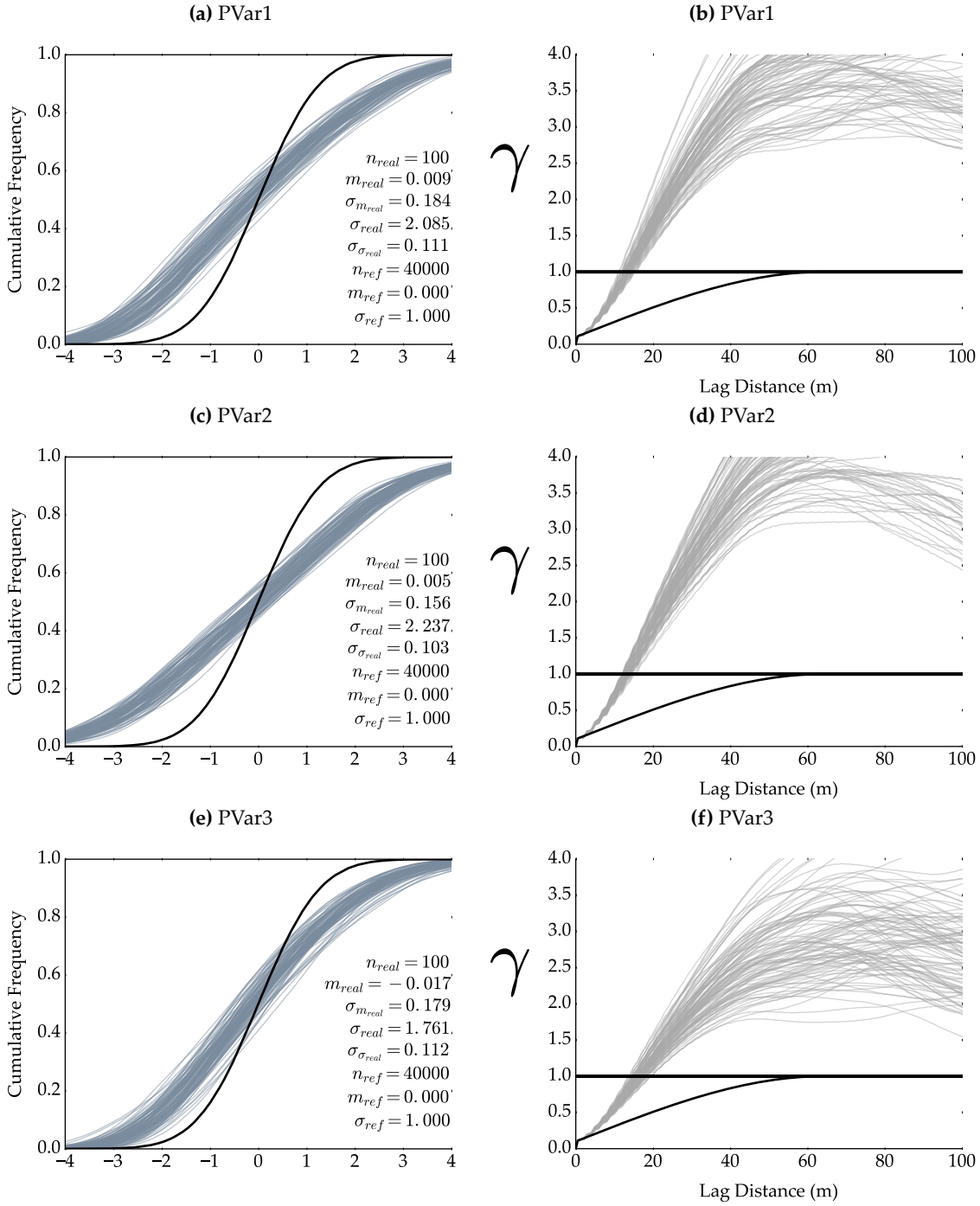

Figure 3.17: Histogram and variogram reproduction of the primary variables in the intrinsic variogram model case study. The variogram reproduction plots contain the input variogram (black) and the variogram for each realization (grey).

fix is attempted. If a set of quit conditions are not met, another tuning run is executed. A summary of this procedure is as follows:

1. Run CCC for 10000 nodes.
2. Calculate the collocated primary-primary correlation matrix and variances of the variables being modeled.
3. Check for errors in the primary-primary correlation matrix.
   a. If required, apply correlation correction (see below); else, keep track of consecutive error free runs.
   b. If five consecutive runs are error free, prevent further changes to the correlation correction.
4. Check for variance inflation.
   a. If required, apply variance correction (see below); else, keep track of consecutive error free runs.
   b. If five consecutive runs are error free or the variance inflation fix reaches a predefined maximum influence, prevent further changes to the variance correction.
5. Run steps 1 to 4 until both the errors are fixed or quit conditions for both fixes is met

To correct variance inflation, a multiplicative factor $\mathbf{f}_p$ consisting of the values $\{f_{p,k}, k = 1, \ldots, K\}$ is implemented that reduces kriging variance for each variable being simulated at each node as follows:

$$\mathbf{y}_p^{sim}(\mathbf{u}') = \mathbf{y}_p(\mathbf{u}') \cdot \boldsymbol{\sigma}_p^{sk}(\mathbf{u}') \cdot \mathbf{f}_p + \mathbf{y}_p^{sk}(\mathbf{u}')$$

$\mathbf{f}_p$ is adjusted iteratively by reducing its value by 0.02 for each tunning run if the variance of the variable is above a value of 1.1. If it reaches a value of 0.04, it is reduced by 0.002 until a value of 0.002 is reached at which point the algorithm does not attempt to fix variance inflation for that variable anymore.

To correct errors in the primary-primary correlation matrix, the $\mathbf{L}_{pp}$ portion of the $\mathbf{L}$ matrix is adjusted, that can be expressed as:

$$\mathbf{L}_{pp} = \begin{bmatrix} L_{I+1,I+1} & 0 & 0 \\ \vdots & \ddots & 0 \\ L_{I+K,I+1} & \ldots & L_{I+K,I+K} \end{bmatrix}$$

During each tuning run, the primary-primary variable correlation matrix of the 10000 simulated nodes is calculated and decomposed using Cholesky decomposition. If the input and output correlation coefficient is different by a value greater than $|0.25|$ a correction is made. Using the input $\mathbf{L}_{pp}^{in}$ matrix used by the last completed tunning run and the calculated $\mathbf{L}_{pp}^{tune}$ matrix from the collocated

correlation matrix calculated from its results, an adjusted $\mathbf{L}_{pp}^{adj}$ matrix is calculated. For each $\rho_{kk'}$ value that requires correction, its corresponding $L_{kk'}^{adj}$ value is calculated as follows:

$$L_{ii'}^{adj} = \begin{cases} L_{ii'}^{tune} - \dfrac{L_{ii'}^{in} - \sum_{j=I+1}^{i-2} L_{i'j} \cdot L_{ij}}{L_{i'i'}^{in}}, & \text{if } i > I+2 \\ L_{ii'}^{tune} - \dfrac{L_{ii'}^{in}}{L_{i'i'}^{in}}, & \text{otherwise} \end{cases}, \quad \begin{array}{l} k = 1, \ldots, K, \\ i = k + I, \\ i' = i - 1 \end{array} \tag{3.8}$$

If no correction is required for a particular $\rho_{kk'}$, its corresponding $L_{kk'}^{adj}$ value is copied from $\mathbf{L}_{pp}^{in}$. The correlation correction is done in an iterative fashion such that the values within each column of the $\mathbf{L}_{pp}^{adj}$ matrix are calculated top down, moving left to right. Values along the diagonal of $\mathbf{L}_{pp}^{adj}$ are calculated once all L^{adj} values in its column are calculated by completing the following procedure:

$$x = \sum_{j=I+1}^{i-1} L_{ij} \cdot L_{ij}$$

$$L_{ii}^{adj} = \begin{cases} \sqrt{1-x}, & \text{if } x < 1 \\ 0.01, & \text{otherwise} \end{cases}, \quad k = 2, \ldots, K, i = k + I$$

Once all corrections are made, the $\mathbf{L}_{pp}^{adj}$ matrix is used for the subsequent tunning run or in the case that all quit conditions have is met, modeling of the variables. By adjusting the $\mathbf{L}_{pp}$ matrix, individual correlation coefficients are adjusted on an as needed basis rather then attempting a global correction.

While testing the implementation of the correlation fix, it was observed that there is a point of diminishing returns. After which, when attempting to fix a correlation coefficient, an increase in the magnitude of its corresponding $\mathbf{L}_{pp}$ value has less influence and exacerbates the variance inflation issue. To prevent this, the value $L_{i'i'}^{in}$ used in the expression (3.8) has a forced floor value of 0.3.

3.4.5 Synthetic Case Study

3.4.5.1 Generation of Synthetic Multivariate Data

To allow generated models during research to be checked, synthetic data is generated using a LMC that has known spatial and multivariate relationships that can be checked for reproduction. The process outlined in Section 3.2 is used to generate the synthetic data. An example of this process and a description of any necessary parameters is illustrated below. The process is very similar to the process used in Section 3.3.2.1; however, and additional exhaustive secondary variable is generated for this case study to increase the complexity of the multivariate relationships.

To test multiple collocated correlation structures, multiple synthetic datasets are generated with five exhaustive variables $K = 5$ using a unique synthetic LMC. For this study, it is thought that three primary variables $\{Z_k(\mathbf{u}), k = 1, \ldots, 3, \forall \mathbf{u}\}$ and two secondary variables $\{Z_k(\mathbf{u}), k = 4, 5, \forall \mathbf{u}\}$ provides adequate dimensionality and the use of 10 unique data sets allows the methodology to be

tested sufficiently. A 2-D 200 by 200 cell grid is used for all synthetic datasets along with the same five $L = 5$ nested variogram structures $\{\Gamma(\mathbf{h})^l, l = 1, \ldots, 5\}$ with no nugget effect Γ^0. Spherical variogram structures are used for the five nested variogram structures with ranges of 16, 28, 40, 52, and 64. Because of the procedure used to generate the synthetic data, there needs to be as many nested structures as there are variables being generated.

Matrices of coefficients $\{a_k^l \forall l, k\}$ are specified to derive synthetic data that honors a LMC. After which, the process outlined in Section 3.2 is used and illustrated below.

The $\mathbf{A}^l$ are generated by populating $L = 5$ number of 5 by 1 matrices with randomly drawn values between -0.75 and 0.75. The $\mathbf{A}^l$ matrices are then corrected so the coefficients met the condition:

$$\sum_{l=1}^{L} b_{kk}^l = 1, \quad \forall k \tag{3.9}$$

For each LMC, the $\mathbf{A}^l$ matrices are corrected by first calculating:

$$\mathbf{a}^{sum} = \sum_{l=1}^{L} {a_k^l}^2 \quad \text{if } k \neq l, \quad \forall k$$

If any values within the $\mathbf{a}^{sum}$ vector is not within the range of $(0, 1)$, the generated $\mathbf{A}^l$ matrices are removed, restarting the process. Otherwise, the value required to satisfy the expression (3.9) is calculated by completing the following calculation:

$$a_k^l = 1 - a_k^{sum} \quad \text{with } l = k, \quad \forall k$$

A positive semi-definite LMC are then calculated from the generated $\mathbf{A}^l$ matrices by completing the procedure outlined in 3.2.

The correlation matrix of each LMC generated is then checked so that it meet a set of conditions, as listed below:

1. The $\mathbf{L}$ matrix calculated by Cholesky decomposition of the correlation matrix must display a minimum value of $|0.20|$ along the diagonal to avoid a singular matrix.
2. There must be a minimum difference of $|0.4|$ between the primary-secondary correlation coefficient for each of the two secondary variables (e.g., $|\rho_{14} - \rho_{15}| > 0.4$).
3. No correlation coefficient can be greater than $|0.75|$.
4. No correlation coefficient between primary variables and a secondary variable can be less than $|0.15|$.
5. The correlation coefficient between the secondary variables can not be greater than $|0.3|$.

If these conditions are met, the randomly generated LMC is used to generate a synthetic dataset; otherwise, the LMC is discarded and the process restarted.

Example of Synthetic Data Generation

The following is an example of a set of derived $\mathbf{A}^l$ matrices and their resulting LMC, calculated contribution matrices $\mathbf{B}^l$, and the synthetic multivariate dataset derived from it. The nested variogram structures outlined in Section 3.4.5.1 are used. The values illustrated below are from one of the synthetic datasets used for this case study.

The $\mathbf{A}^l$ matrices are first generated:

$$\mathbf{A}^1 = \begin{bmatrix} -0.02 \\ 0.35 \\ -0.53 \\ -0.05 \\ -0.35 \end{bmatrix}, \ \mathbf{A}^2 = \begin{bmatrix} -0.05 \\ -0.24 \\ 0.33 \\ 0.03 \\ -0.38 \end{bmatrix}, \ \mathbf{A}^3 = \begin{bmatrix} 0.49 \\ -0.13 \\ 0.54 \\ 0.34 \\ -0.13 \end{bmatrix}, \ \mathbf{A}^4 = \begin{bmatrix} 0.35 \\ 0.49 \\ 0.41 \\ 0.18 \\ -0.37 \end{bmatrix}, \ \mathbf{A}^5 = \begin{bmatrix} 0.02 \\ 0.48 \\ -0.24 \\ 0.70 \\ 0.28 \end{bmatrix}$$

then corrected to:

$$\mathbf{A}^1 = \begin{bmatrix} 0.80 \\ 0.35 \\ -0.53 \\ -0.05 \\ -0.35 \end{bmatrix}, \ \mathbf{A}^2 = \begin{bmatrix} -0.05 \\ 0.62 \\ 0.33 \\ 0.03 \\ -0.38 \end{bmatrix}, \ \mathbf{A}^3 = \begin{bmatrix} 0.49 \\ -0.13 \\ 0.62 \\ 0.34 \\ -0.13 \end{bmatrix}, \ \mathbf{A}^4 = \begin{bmatrix} 0.35 \\ 0.49 \\ 0.41 \\ 0.62 \\ -0.37 \end{bmatrix}, \ \mathbf{A}^5 = \begin{bmatrix} 0.02 \\ 0.48 \\ -0.24 \\ 0.70 \\ 0.76 \end{bmatrix} \tag{3.10}$$

so that the expression (3.9) is satisfied. After which, the symmetrical $\mathbf{B}^l$ matrices are calculated using the expression (3.1) using the derived $\mathbf{A}^l$ matrices (3.10). As an example, the calculation of $\mathbf{B}_1$ results in:

$$\mathbf{B}_1 = \begin{bmatrix} 0.63 & 0.28 & -0.42 & -0.04 & -0.28 \\ 0.28 & 0.12 & -0.19 & -0.02 & -0.12 \\ -0.42 & -0.19 & 0.28 & 0.03 & 0.18 \\ -0.04 & -0.02 & 0.03 & 0.00 & 0.02 \\ -0.28 & -0.12 & 0.18 & 0.02 & 0.12 \end{bmatrix}$$

The correlation matrix $\boldsymbol{\rho}$ calculated using the expression (3.3) is illustrated in Figure 3.18a.

Direct and cross variograms are defined by the generated LMC. As full cokriging is not being used, only the direct variograms are required. The covariance contributions for each of the nested variogram structures within $\boldsymbol{\Gamma}(\mathbf{h})$ as outlined in Section 3.4.5.1 are contained within the $\mathbf{B}^l$ matrices. In this example case, the direct variograms is as follows:

$$\gamma(\mathbf{h})_{11}(\mathbf{h}) = 0.63 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=16} + 0.00 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=28} + 0.24 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=40} + 0.12 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=52} + 0.00 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

$$\gamma(\mathbf{h})_{22}(\mathbf{h}) = 0.12 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=16} + 0.39 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=28} + 0.02 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=40} + 0.24 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=52} + 0.23 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

$$\gamma(\mathbf{h})_{33}(\mathbf{h}) = 0.28 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=16} + 0.11 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=28} + 0.38 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=40} + 0.17 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=52} + 0.06 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

$$\gamma(\mathbf{h})_{44}(\mathbf{h}) = 0.00 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=16} + 0.00 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=28} + 0.12 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=40} + 0.39 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=52} + 0.49 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

$$\gamma(\mathbf{h})_{55}(\mathbf{h}) = 0.12 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=16} + 0.14 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=28} + 0.02 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=40} + 0.14 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=52} + 0.58 \cdot \mathrm{Sph}(\mathbf{h})_{\mathrm{a}=64}$$

(a) LMC model

	PVar1	PVar2	PVar3	SVar1	SVar2
PVar1	1.00	0.36	0.00	0.36	-0.43
PVar2	0.36	1.00	0.02	0.59	-0.15
PVar3	0.00	0.02	1.00	0.34	-0.36
SVar1	0.36	0.59	0.34	1.00	0.27
SVar2	-0.43	-0.15	-0.36	0.27	1.00

(b) Synthetic data

	PVar1	PVar2	PVar3	SVar1	SVar2
PVar1	1.00	0.36	-0.03	0.31	-0.49
PVar2	0.36	1.00	0.01	0.48	-0.35
PVar3	-0.03	0.01	1.00	0.38	-0.38
SVar1	0.31	0.48	0.38	1.00	0.14
SVar2	-0.49	-0.35	-0.38	0.14	1.00

Figure 3.18: Correlation matrices of the input LMC model and the resulting synthetic data.

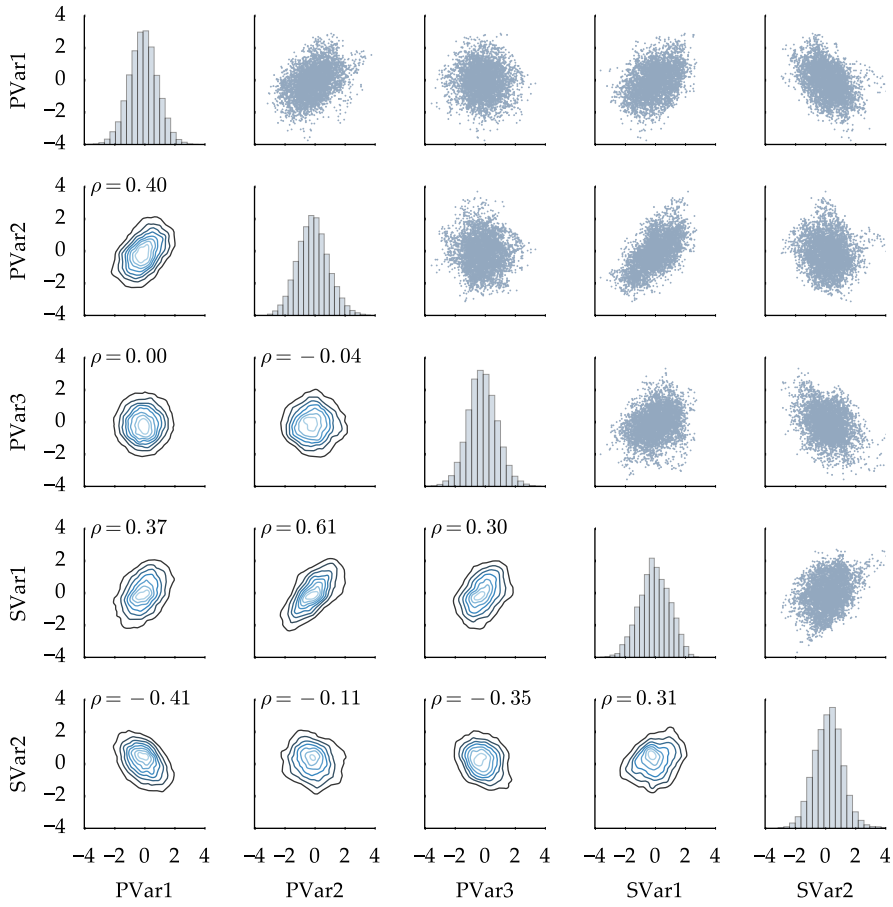

Figure 3.19: Multivariate joint density, histogram, and scatter plot matrix of the synthetic data generated.

Synthetic data is then generated using the expression (3.4) and is summarized in Figure 3.19. The correlation matrix calculated from the synthetic data (Figure 3.18b) showed little change to the LMCs theoretical correlation matrix. For illustration purposes, a heat map of two of the generated variables, PVar1 (i.e., $l = 1$) and PVar2 (i.e., $l = 4$), are shown in Figure 3.20. Their calculated experimental variograms are plotting along side their theoretical variogram determined by the LMC are shown in Figure 3.21.

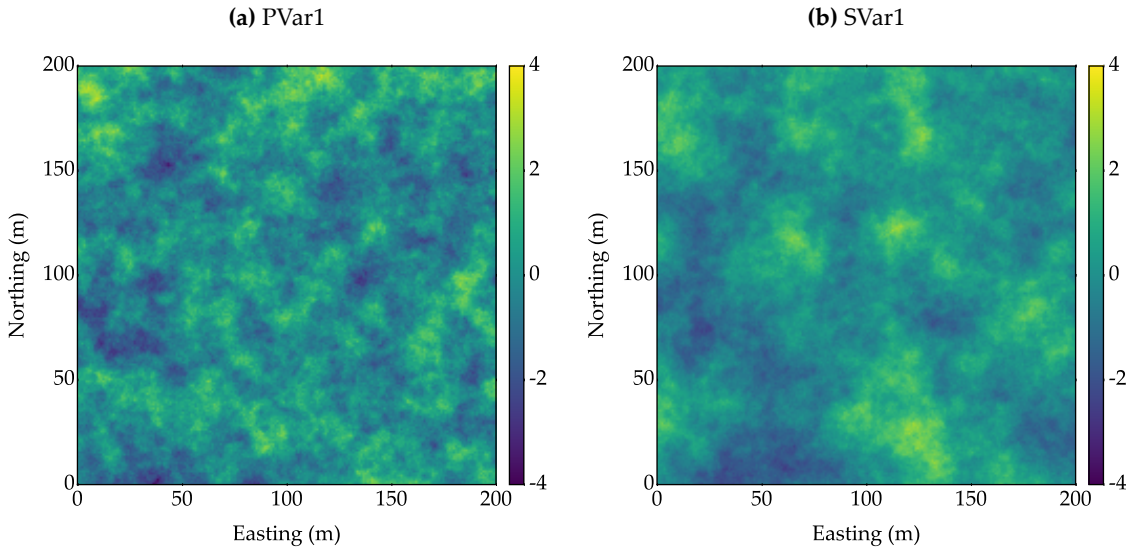

Figure 3.20: Example heat maps of generated 2-D synthetic data in normal scored space.

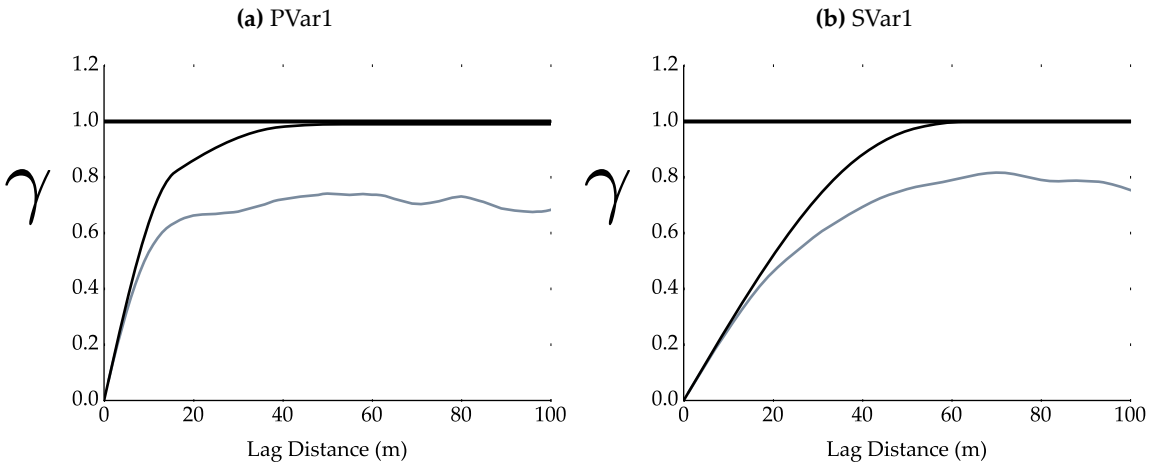

Figure 3.21: Experimental variograms of the synthetic data (blue) at a azimuth of 90° and the theoretical input LCM model variogram (black).

Summary of Synthetic Data Generated

Using the process discussed above, 10 synthetic datasets are generated. Figure 3.22 illustrates the range of correlation coefficients generated and subsequently tested.

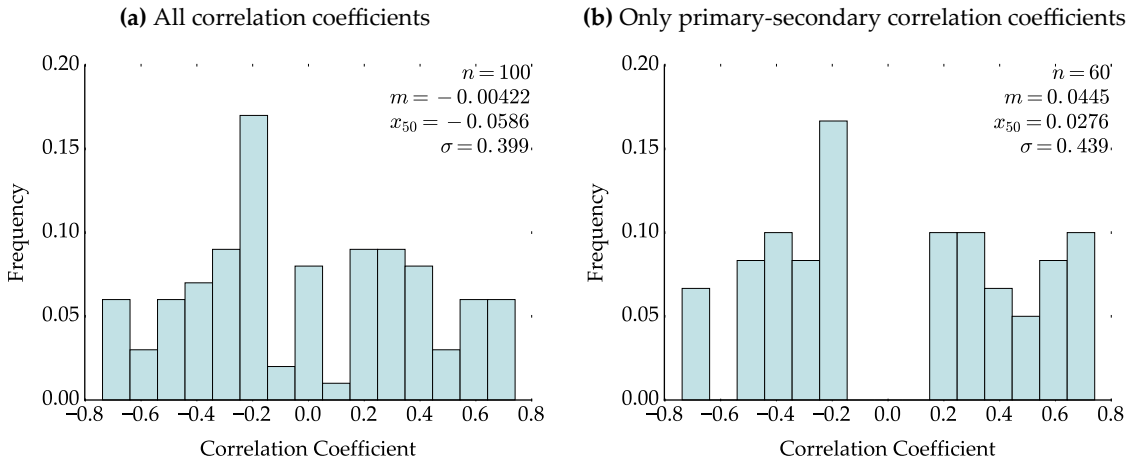

Figure 3.22: Histograms of the correlation coefficients within the 10 generated correlation matrices.

3.4.5.2 Results

CCC does not operate effectively even with both corrections implemented as illustrated by the cross-validation plots in Figure 3.23. Error in primary-primary correlation reproduction appears to increase when the magnitude of the input the correlation coefficient increases.

To summarize histogram and variogram reproduction, reproduction plots of a single synthetic datasets are illustrated in Figure 3.24. Reproduction plots of the other synthetic datasets are not shown or discussed in detail. Overall findings are discussed in Section 3.4.6. In this example case, the histograms of the simulated realizations showed that the method underestimated high values and overestimated low values.

The difference between the input $\mathbf{L}_{pp}$ values and the final values after correction are summarized in Figure 3.25. The variance inflation factors used are summarized in Figure 3.26.

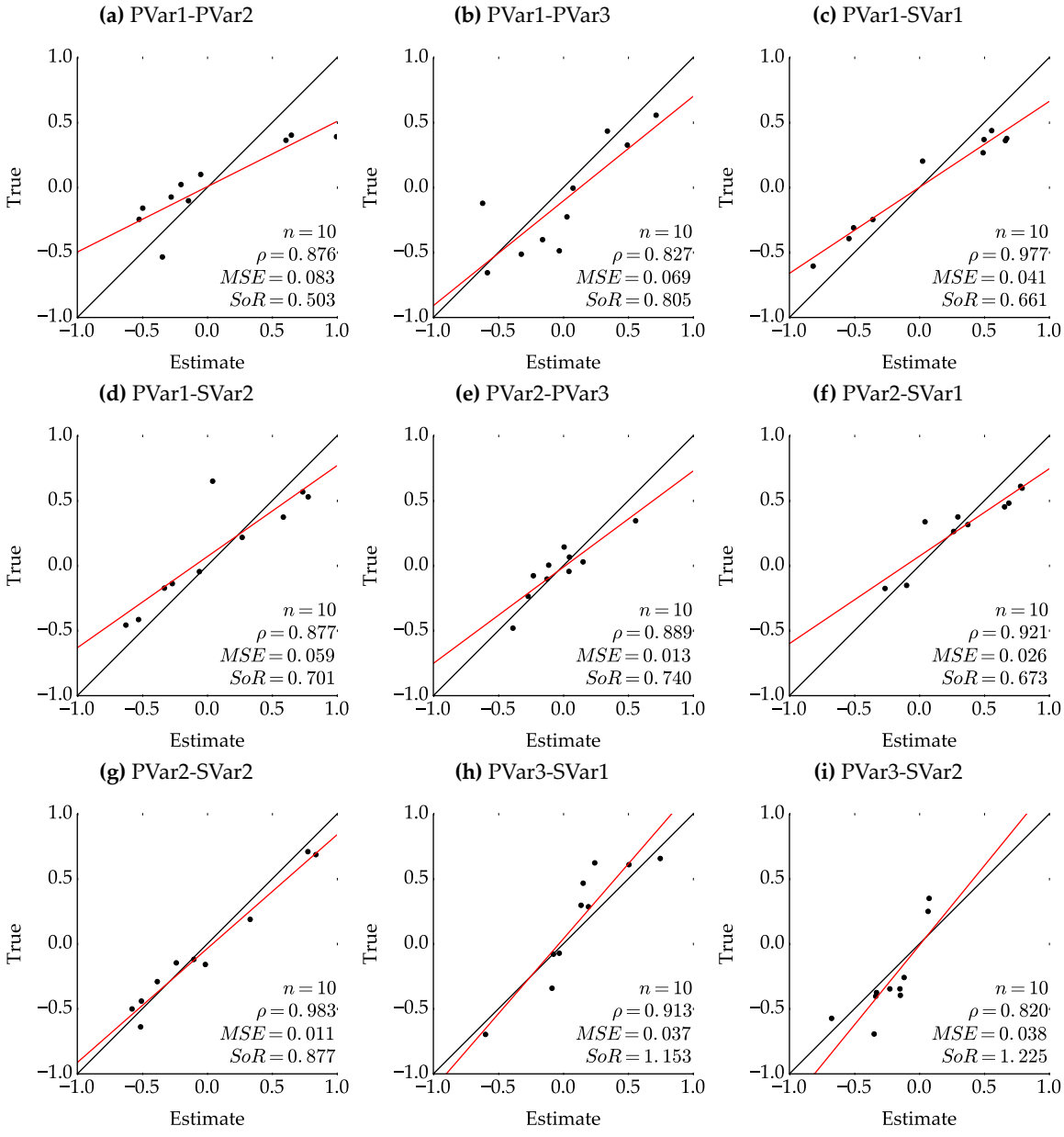

Figure 3.23: Cross validation scatter plots for each correlation coefficient from all 10 synthetic cases.

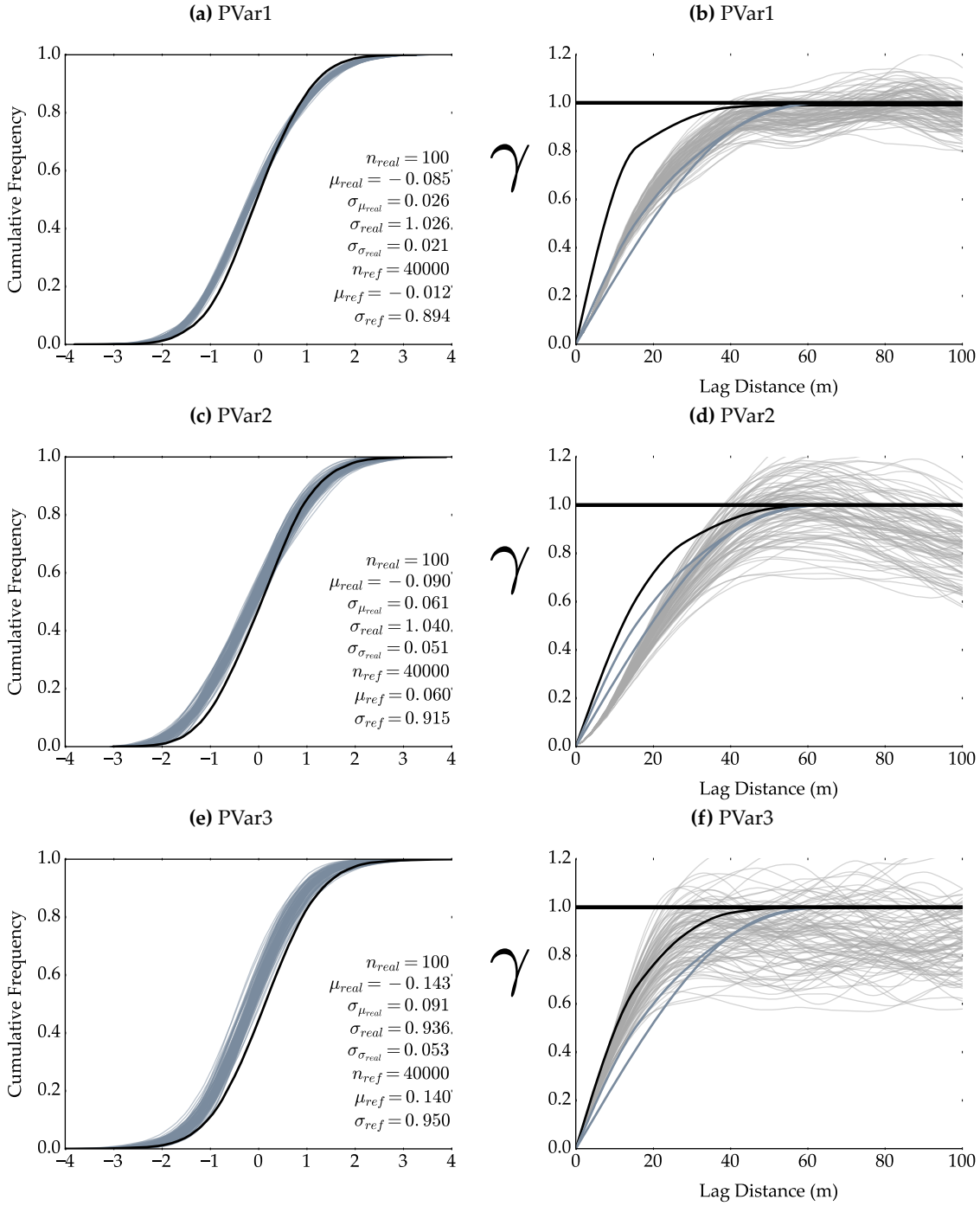

Figure 3.24: Histogram and variogram reproduction of the primary variables simulated from one of the 10 synthetic cases. The variogram reproduction plots contain the input variogram (black), the variogram for each simulated realization (grey), and the variograms from the LMC representing the secondary data (blue).

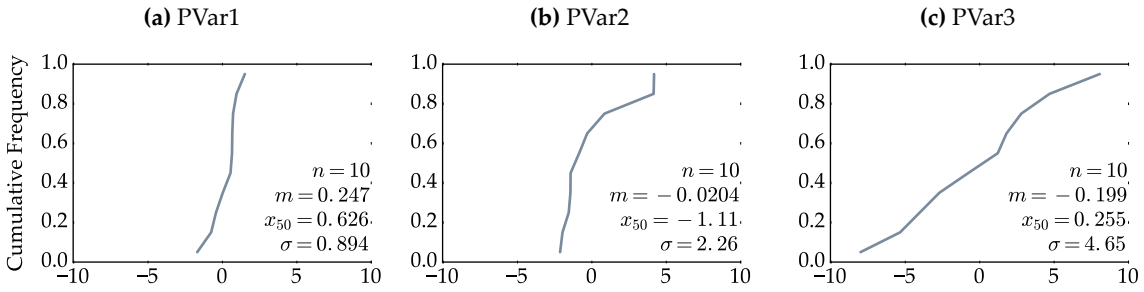

Figure 3.25: Histograms of the difference between the input and fixed lower matrix value controlling each primary-primary correlation coefficient.

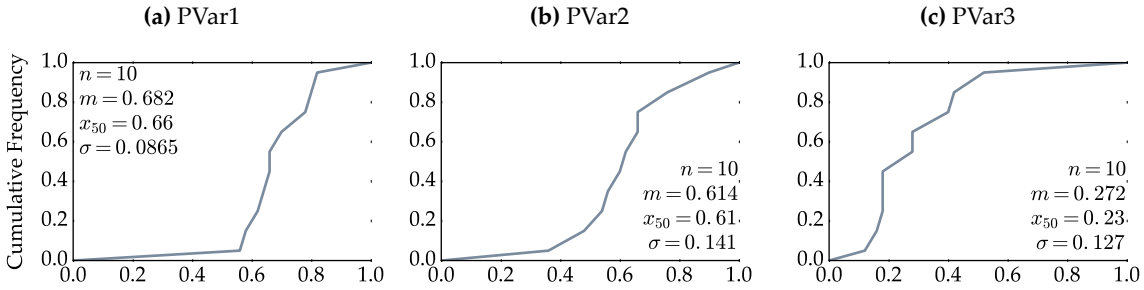

Figure 3.26: Histograms of the variance factors used for each simulated variable in the 10 synthetic cases.

3.4.6 Conclusion

It is thought by the author that a correlation reproduction error less then $|0.20|$ is reasonable. Figure 3.27 illustrates the percentage of errors greater than this threshold observed from the case study discussed in Section 3.4.5. It is clear that CCC could not adequately reproduce the collocated primary-primary correlation structure.

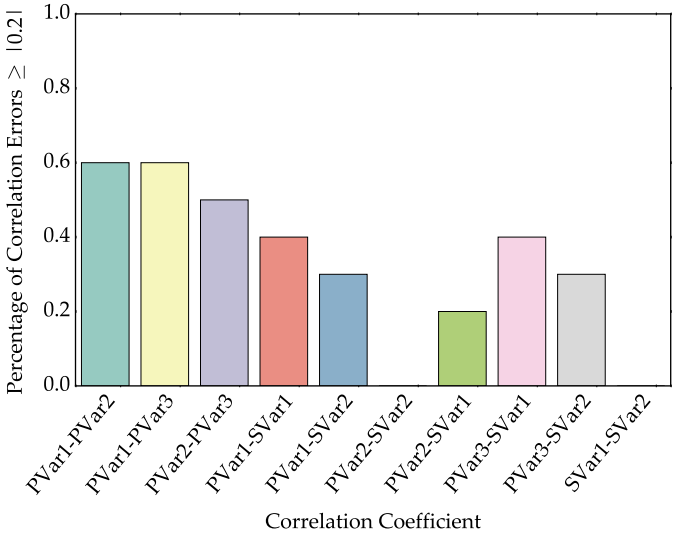

Figure 3.27: Summary bar chart detailing the percentage of times a correlation coefficient was not reproduced within $|0.20|$ of the input value.

Prior to simulation, it is important to ensure the simulation engine effectively reproduces input statistics such as the histogram and variograms. This is demonstrated by the pure nugget effect case study discussed in Section 3.4.3.1. In evaluating the variogram reproduction plots from the case study discussed in Section 3.4.5, it is clear that the secondary data had a strong influence on the spatial structure of the simulated variables. Throughout the 10 cases, it was observed that the simulated realizations of the variables variograms with stronger input primary-secondary correlations would become more similar to the secondary variables variogram and showed a decrease in variability across realizations.

The corrections implemented in CCC showed diminishing returns as it became increasingly difficult to correct the $\mathbf{L}_{pp}$ matrix as the number of correlation coefficients fixed increased. Subsequent corrections had less flexibility and needed a more drastic correction to account for previous corrections (Figure 3.25). The compounding nature of the correlation fix caused variance inflation to become exacerbated as the diagonal values along the $\mathbf{L}_{pp}$ matrix became smaller with additional corrections. This issue is reflected in Figure 3.26 as the variance correction factors required by each subsequent variable became smaller (i.e., more correction).

The magnitude of the corrections needed by CCC are disconcerting and bring the validity of the methodology into question. It is clear that when spatial correlation is considered, regardless of the complexity, variance inflation and collocated correlation reproduction become an issue. It is not clear as to why these issues exist; however, it has been postulated that these issues stem from the fact that the collocated variance of the spatial conditioning data is not considered (J.G. Manchuk, personal communication, January 11, 2016). This idea has not been investigated to date.

3.5 Conclusion

The decorrelation framework does not work well when considering exhaustive secondary data. Inducing correlation between the simulated factors renders the back transformations of both PCA and Sphere-R unable to reproduce the original correlation structure.

CCC required excessive ad hoc corrections that could not fully control variance inflation nor correct erroneous correlation coefficients. Future work could modify the methodology, removing the ad hoc corrections and ensuring that the collocated secondary data of conditioning data is also considered when calculating the conditional variance.

Unfortunately, neither the decorrelation method nor CCC proved to be an adequate methodology for multivariate simulation with many secondary data. Existing nontrivial frameworks will need to be explored.

Chapter 4

Implementation of Multivariate Simulation with Many Secondary Data

4.1 Introduction

Effective reproduction of collocated correlation structure and other statistics when simulating multivariate data with many secondary data requires frameworks not tested in Chapter 3 to be explored. Unlike the synthetic case studies described in Sections 3.3.2 and 3.4.5, a massively multivariate dataset suitable for MPM is used to test alternatives.

This chapter investigates and documents two frameworks: first the cokriging framework that requires an LMC and secondly the hierarchical framework. A discussion on the dataset being tested is followed by two case studies. The cokriging framework is found to be ineffective as the LMC is too restrictive in a massively multivariate setting. The hierarchical framework is found to adequately reproduce input statistics while characterizing uncertainty.

4.2 Data Source and Processing

4.2.1 Introduction

As discussed in Section 1.2, the geochemical signature of stream sediment samples are thought to represent the full spectrum of the truth, minimizing the use of censored data and heuristic searches utilized by current MPM frameworks. Additionally, by considering all available exhaustive secondary data during the modeling processes, the resulting model improves even if they are poorly correlated (Cuba et al., 2009). Therefore, it is desired to test a stream sediment geochemical dataset with exhaustive geological data.

The following section describes the datasets retrieved, processing performed on them, and an exploratory data analysis (EDA). Data commonly used for mineral exploration and MPM is described in Section 2.1.2.

4.2.2 Data Source

Due to the extremely extensive geological data available in the Yukon, Canada, it is an excellent location to source data from. Stream sediment samples (Héon, 2003) and bedrock geology data (Yukon Geological Survey, 2016b) is collected from Geomatics Yukon (2014). DEM data (Canada Centre for Mapping and Earth Observation, 2014) is collected from GeoGratis (Natural Resources

Canada, 2014a). Aeromagnetic data (Geophysical Data Centre, 2014) data is collected from the Geoscience Data Repository for Geophysical Data (Natural Resources Canada, 2014b).

As the datasets covers most of the Yukon, an area of interest (AOI) is selected using two criteria: (1) the area must contain a sufficient number of stream sediment samples of the same vintage so that the type of geochemical analysis performed on all of the samples are the same, (2) the exhaustive secondary data—geology, geophysics and DEM—must cover the entire area. A 100 by 100 kilometer (km) AOI is selected that satisfies these conditions and is fully contained within the Yukon (Figure 4.1).

Figure 4.1: Location of the selected AOI. Base map data provided by ESRI (2012).

4.2.3 Primary Variables

Stream sediment samples from the AOI are extracted and their geochemical signatures are used as primary variables. The geochemical analysis performed on the samples is extensive and in some cases, varied between samples. Of the geochemical variables available, 42 are found to be analysed similarly and are deemed appropriate to use for the purpose of this study. After reduction from the original 1040, 813 stream sediment samples remain within the AOI and are roughly evenly distributed (Figure 4.2). The selected 42 geochemical variables are referred to as the primary variables.

4.2.3.1 Calculating Catchment Areas

As discussed in Section 2.1.2.2, stream sediment samples are a representation of their corresponding catchment area. To calculate the stream sediments catchment area, a collection of tools are used. They include ArcGIS Desktop 10.1 (ESRI, 2012), the Optimized Pit Removal tool (Jackson, 2013), and geographic information system (GIS) functionality developed for this thesis in the mixed

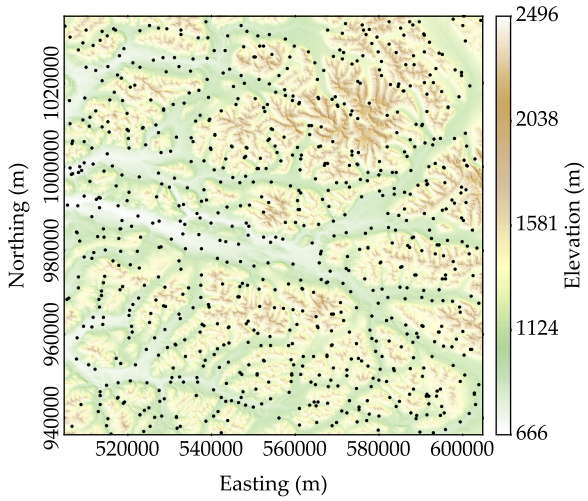

Figure 4.2: Map illustrating the distribution of the stream sediment samples found within the AOI overlying a DEM.

FORTRAN-Python package `pygeostat` (Centre for Computational Geostatistics, 2016b). A more in depth explanation of calculating catchment areas using a DEM is found in Jones (2002). Below is a summary of the workflow used for this thesis:

1. Upscale the DEM to match the grid definition.
2. Remove pits from the the upscaled DEM, burn drainage network into it, and remove pits again from the burned DEM.
3. Calculate a flow accumulation grid and define the cutoff to calculate a drainage network.
4. Snap stream sediment samples to nearest drainage network derived from the previous step, fixing the locations of the pour points of the catchment areas. Visually validate the snapped pour points.
5. Calculate a flow direction grid.
6. Calculate catchment areas using the flow direction grid and the snapped pour points.
7. Visually inspect all catchment areas and move pour points that require adjusting.
8. Repeat steps 6 and 7 until the calculated catchment areas properly represent their contributing area.
9. Buffer final catchment areas, increasing their size equal to a single cell.
10. Perform an intersection with the relocated pour points and the buffered catchment areas. Pairs of upstream-downstream catchment areas can now be identified if the intersected pour points overlap two catchment areas.
11. Determine the hierarchy of the catchment pairs based on their elevation.
12. Build the stream sediment sample networks and determine each stream samples predecessors.

13. Merge the catchment areas to include their predecessors.

The calculated catchment areas then indicate the cells that contribute to each stream sediments sample location or catchment pour point (Figure 4.3). If a network of samples exist, the downstream samples catchment area also includes the upstream samples catchment area. Spatial declustering is performed with a cell size of 5 km. All 42 primary variables are normal score transformed using the declustering weights to ensure that all variables are univariate Gaussian; further, they are all assumed to be multi-Gaussian.

As a means of distinguishing between the background geochemistry and possible anomaly signatures in the measured stream sediment samples geochemical signature, various studies have attempted to calculate the background geochemistry signature so that it may be removed leaving the residual signature (Arne & Bluemel, 2011; Bonham-Carter & Goodfellow, 1986; Carranza, 2010; Kramar, 1995; Mackie, Arne, & Brown, 2015; Rantitsch, 2000; Rose, Dahlberg, & Keith, 1970). It is thought that with proper decisions of stationarity and by constructing MPM transfer functions , that concerns of anomalies being masked is minimized. As such, the case studies within this chapter uses the measured stream sediment samples geochemical signature.

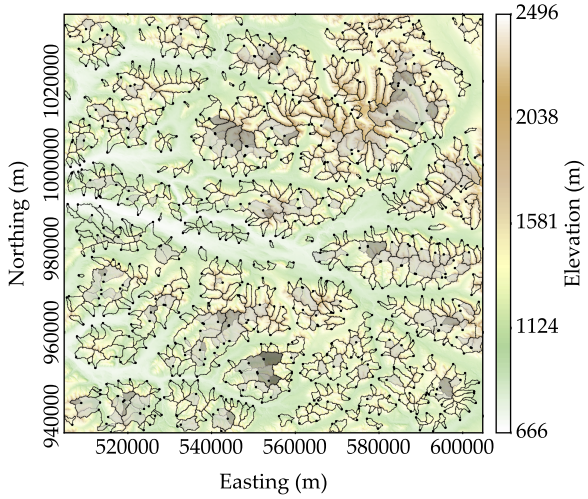

Figure 4.3: Map illustrating the distribution of the calculated catchment areas overlying a DEM. Each network of connected stream sediment samples are grouped together as indicated by the darker outline. Within the network, catchment areas are darker as you move upstream.

4.2.4 Exhaustive Secondary Variables

The DEM and aeromagnteic data—that includes total magnetic and first vertical derivative surveys—are used as exhaustive secondary variables. The bedrock geology data retrieved is also used to derive an additional three exhaustive secondary variables. Distance grids to the nearest lithology contact, fold, and fault are calculated.

Based on the size of the AOI selected, a 4 million cell 2-D grid with 2000 cells in each direc-

tion and a resolution of 50 by 50 meter (m) is defined. The exhaustive secondary data is rescaled to the defined grid. The DEM is upscaled while the aeromagnetic data is downscaled; both processes utilizing bilinear interpolation. The 3 geological feature distance grids are calculated at the appropriate scale and do not require adjustment. The normal score transformation is performed on all exhaustive secondary variables to ensure they are on the same basis (Figure 4.4). These 6 processed exhaustive secondary datasets are used to condition the geostatistical models of the primary variables.

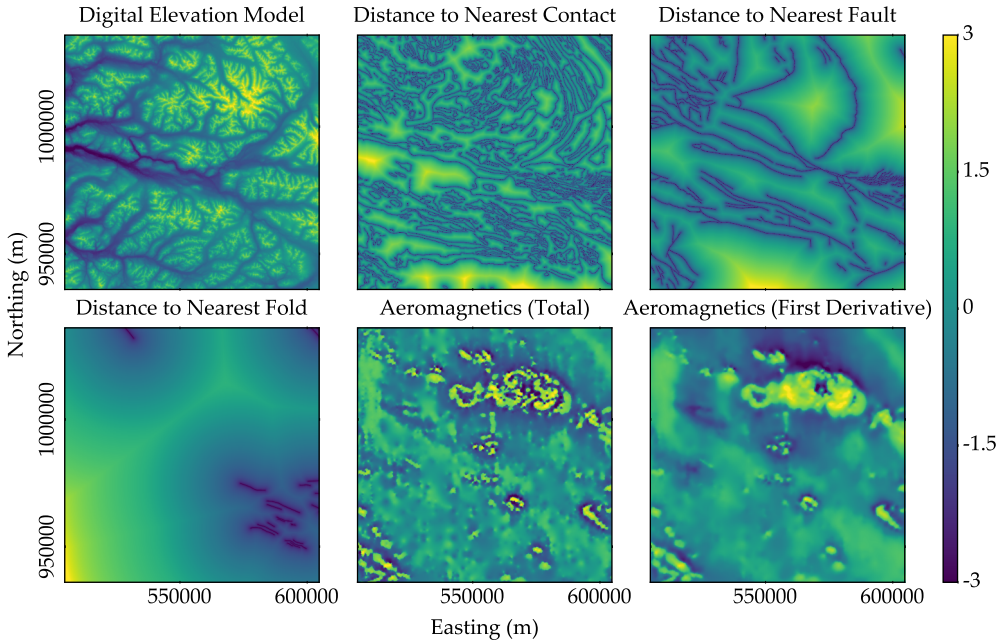

Figure 4.4: Heat maps of the 6 exhaustive secondary variables in normal score space

4.3 Exploration Data Analysis

4.3.1 Introduction

To better understand the multivariate relationships of the dataset, an EDA is performed. This considers if the primary variables are a true representation of their catchment area, what point location should be used for the spatial prediction, and the multivariate relationships between the variables.

4.3.2 Catchment Scale and Sample Location

In an ideal case, stream sediment samples are a perfect representation of their catchment areas background geochemistry and when present, anomalies. This would require that the entire surface of the catchment area to be evenly weathered and equally represented by the stream sediment sample. This is not the case and the reality is extremely complex. Understanding the true nature of

erosion within each individual catchment area, the transportation of eroded sediment, and its final deposition is not possible. Therefore, assumptions must be made on how the samples are utilized.

Most geostatistical techniques require conditioning data to represent a point support, the best location of the stream sediment samples must be investigated. It is thought that if the geochemical signature represents the area, the centroid is more informative, as opposed to the actual sample location or pour point. To evaluate how informative a specific location within a catchment area is, the correlation between the primary variables and the secondary variables is checked in addition to the relationship the catchments size has with the primary variables. The absolute primary-secondary variable correlation coefficients of the centroid are subtracted from the corresponding values from the pour point. The distribution of this difference is illustrated in Figure 4.5. If the average value of the calculated differences is negative, that indicates that the pour point location is more informative. Conversely, if the mean is positive, that indicates that the centroid is more informative and that the geochemical signature is a closer representation of the area. However, the distribution illustrated in Figure 4.5 has a mean of 0.0108 (nearly zero). This does not provide guidance on the best location to assign to the stream samples.

The bivariate relationships between the primary variables and the size of the catchment area they represent are illustrated in Figure 4.6. Of the primary variables, only loss on ignition (LOI) displays a moderate linear relationship with the size of the catchment. Sibbick (1994) attributes this finding to an increased amount of stream energy in smaller catchments causing moss mats sediments to be transported, increasing the organic content of the stream sediment sample. If stream sediment samples are representations of areas, it is expected that the primary variables from smaller catchments would display more variability. Box plots of three primary variables are illustrated in Figure 4.7 where the size of the catchment areas has been divided into three size ranges with the same number of catchments. The smaller third of catchment areas shows a minimal increase in variance compared to the larger catchment areas.

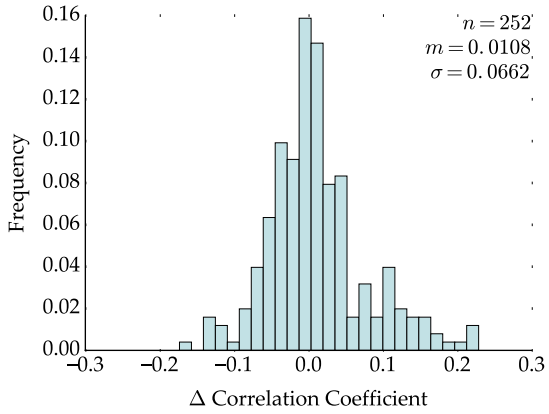

Figure 4.5: Histogram of the difference between the absolute primary-secondary variable correlation coefficients at the centroid and pour point of all catchment areas.

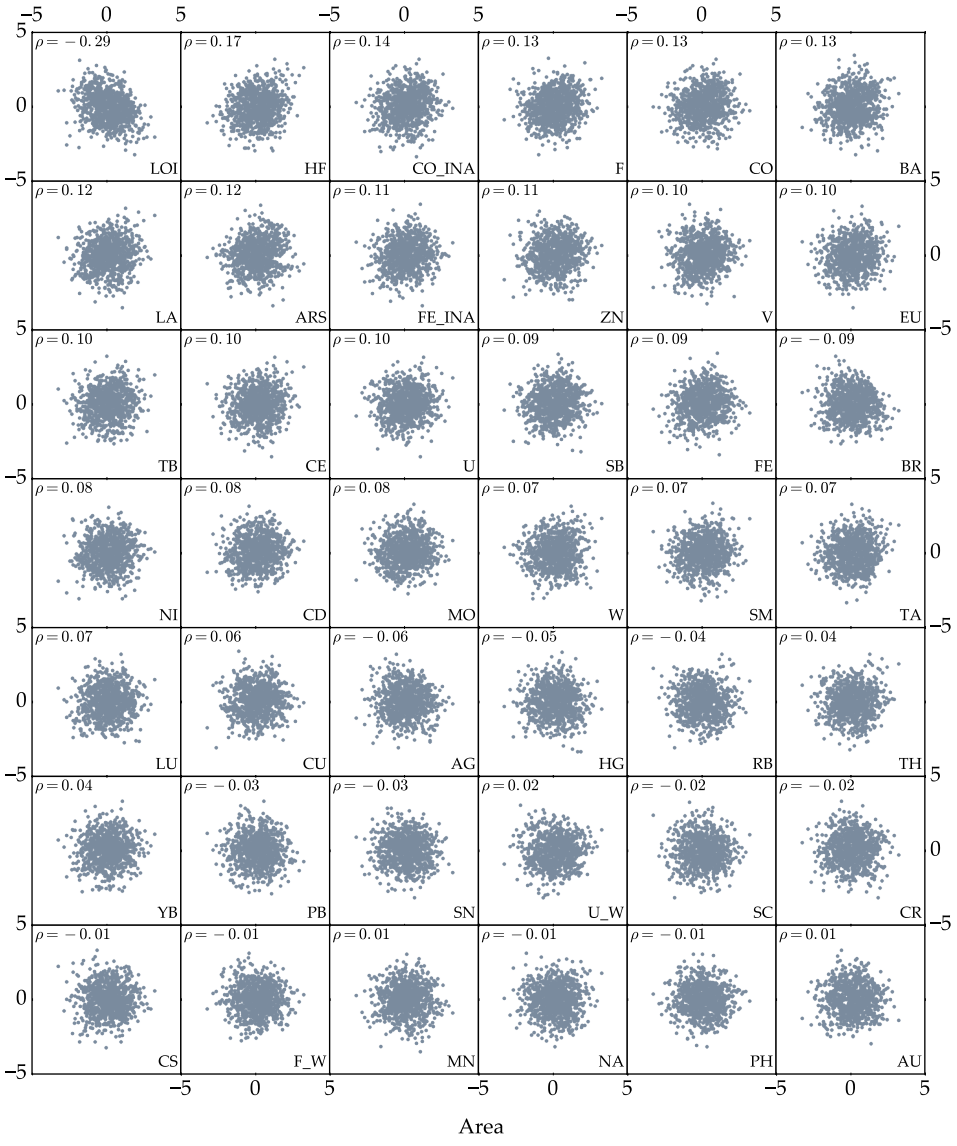

Figure 4.6: Bivariate scatter plots of all primary variables and catchment area size in normal score space. The primary variables indicated within each subplot are plotted along their respective y-axis.

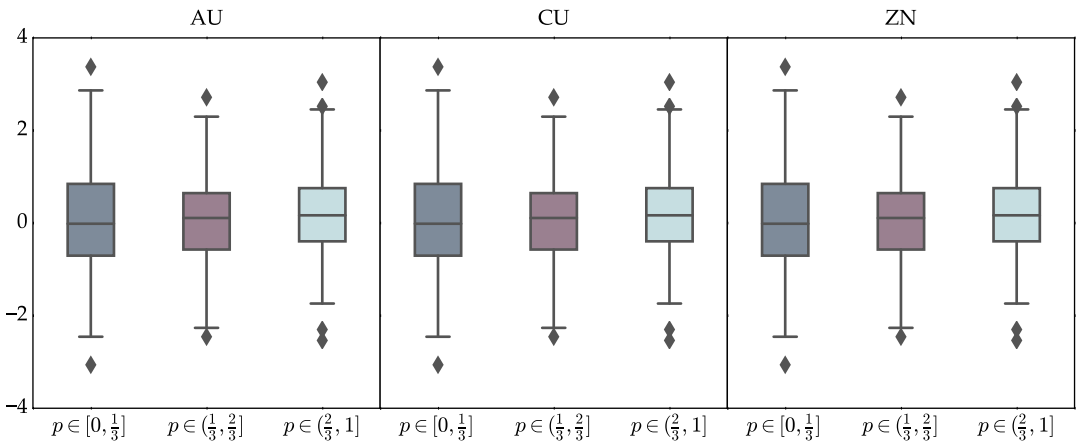

Figure 4.7: Box plots illustrating the variance of three primary variables that are categorized based on the size of their catchment area.

These findings may indicate that the geochemical signature of stream sediment samples are not a representation of the entire catchment area. They may represent a smaller area within the catchment area. This issue may warrant further research; however, for the purpose of the case studies conducted in this chapter, the centroid of the calculated catchment areas is used as the location of the corresponding stream sediment sample.

4.3.3 Multivariate Relationships

To better understand the multivariate relationships between the primary variables and their relationships with the exhaustive secondary variables, a series of statistical exploration techniques are utilized. Specific attention is paid to the collocated correlation structure between primary variables as it is a desired to reproduce these statistics in the case studies discussed in this chapter. The possibility of dimension reduction is also explored.

All 861 correlation coefficients are illustrated in Figure 4.8 and are clustered using the hierarchical Ward methodology (Ward, 1963). There are distinct clusters of variables that display strong linear relationships; however, the linear relationships between variables within clusters and those outside of it varies indicating that the variables within the larger clusters are not redundant. The clustered correlation matrix does not illustrate enough redundancy between primary variables to reduce the dimension the a more manageable number of variables (see Section 3.1).

As a means of visualizing the similarity of the primary variables, multidimensional scaling (MDS) is performed and the first three coordinates are illustrated in Figure 4.9. Similar to the findings from the above correlation matrix, there are distinct clusters of variables; however, there is not enough redundancy displayed where the primary variables may have their dimensionality reduced to a more manageable number of variables (i.e., $K \leq 7$).

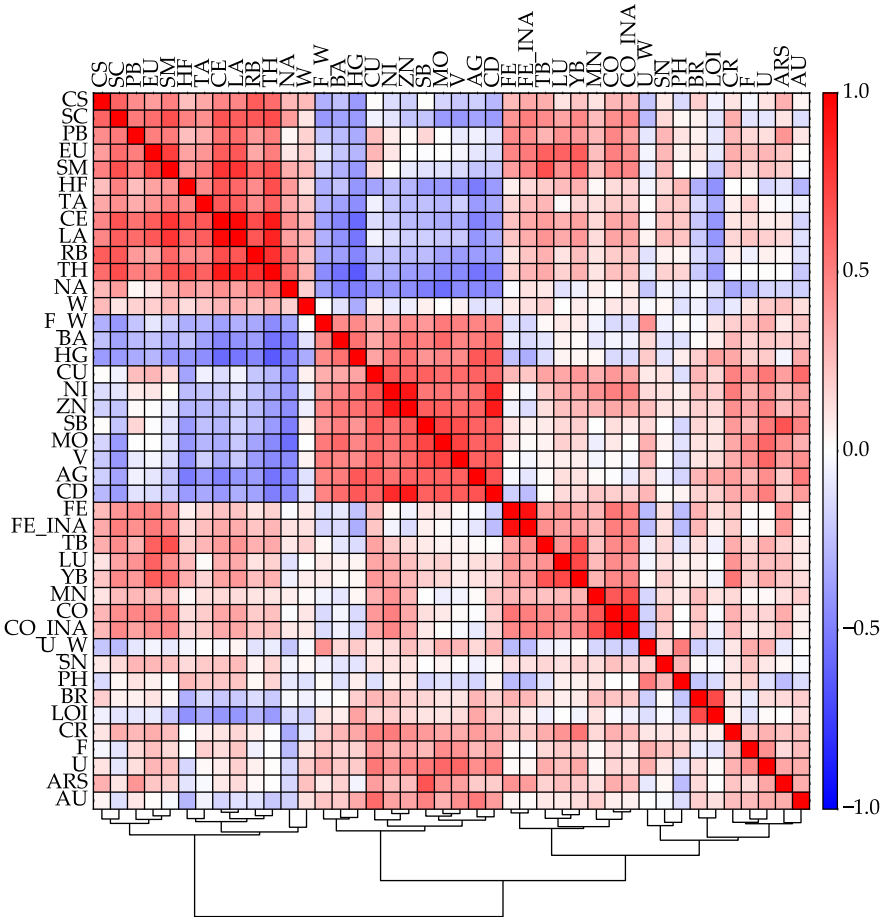

Figure 4.8: Correlation matrix of all the primary variables clustered using the hierarchical Ward method (Ward, 1963).

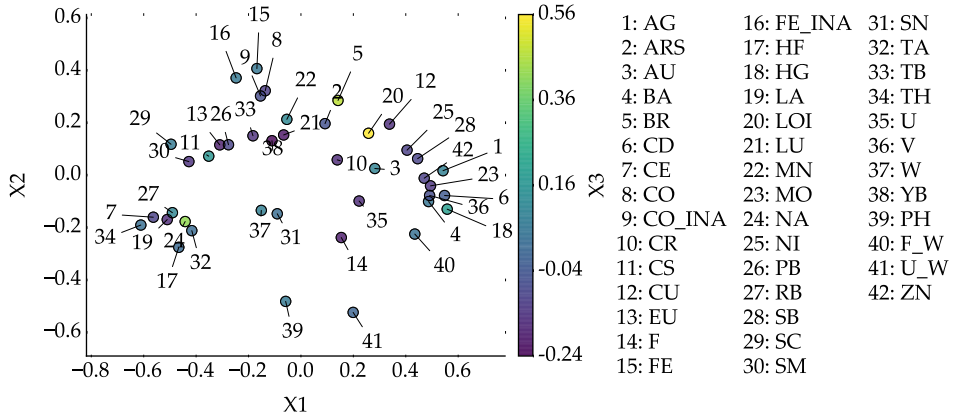

Figure 4.9: MDS plot of each primary variables first 2 MDS coordinates that are colored base on their third MDS coordinate.

As discussed in Section 2.2.3, ICCK utilizes exhaustive secondary data to inform spatial prediction by using the linear relationships between the primary variable being predicted and the collocated secondary data. Those relationships are summarized in Figure 4.10. The only exhaustive secondary variables that displays significant correlation with the primary variables is the DEM. This may be due to weathering of bedrock at higher elevations. While the overall correlation coefficients are not significant for the other geological secondary variables, they do display some moderately significant linear relationships with the primary variables. However, as stated by Cuba et al. (2009), spatial prediction improves when using exhaustive secondary data even if the secondary data are weakly correlated with the primary variables.

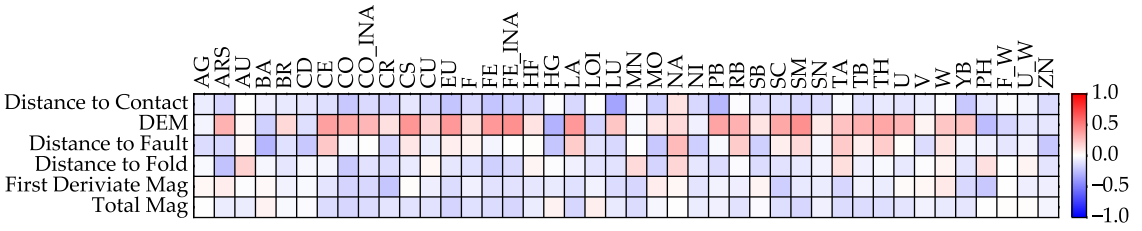

Figure 4.10: Correlation matrix between the primary variables and the collocated exhaustive secondary data at the centroid of each catchment area.

4.4 Cokriging Framework

4.4.1 Introduction

While a cokriging framework is capable of jointly modeling all primary variables, requires a LMC. As discussed in Section 2.2.2, a LMC parameterizes the spatial direct and cross-correlation of each primary variable and primary variable pair. The LMC requires $[K(K+1)]/2$ number of direct and cross variograms, making it arduous to implement as K increases. While it has been shown that $K = 7$ number of variables can be reasonably fit (Jewbali, 2009), that is vastly different than the 42 primary variables in the dataset being used here. With $K = 42$ variables, 903 direct and cross variograms would need to be simultaneously fit with the LMC. Due to the limited number of alternatives to adequately model a massively multivariate dataset, it is interesting to explore the use of a cokriging framework.

The following section describes the implementation of a cokriging framework. Specifically, the building of a 42 variable LMC and implementation challenges encountered are described.

4.4.2 Case Study

4.4.2.1 Methodology

Using the normal score transforms of the primary variables, experimental variograms are calculated, and fit to a LMC. To simplify the procedure, all variograms are omnidirectional. For each of the primary variables being modeled, the exhaustive secondary variables are merged into a single super-secondary variable using the process outlined in Section 2.2.4. SGS with simple cokriging (SCK) and ICCK is run conditional to the stream sediment samples to model the $K = 42$ primary variables.

4.4.2.2 Building the LMC

To model the spatial correlation of the primary variables, 903 experimental omnidirectional direct and cross variograms are calculated (Figure 4.11). Fitting a large number of variograms to an LMC is difficult as the ability to adjust the variance contribution parameters $\{b^l_{kk'}, \forall l, k, k'\}$ of the LMC is limited by the condition that the final K by K matrices $\mathbf{B}^l$ must be positive definite.

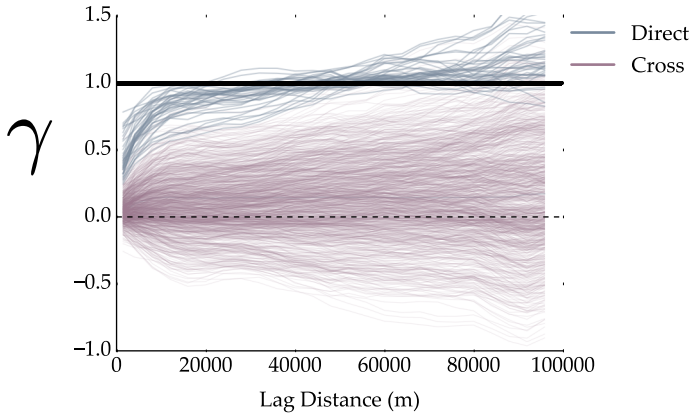

Figure 4.11: The 903 omnidirectional experimental direct and cross variograms requiring fitting with an LMC

By testing various combinations of variogram parameters (i.e., the number, type, and ranges of each variogram structure), a fit is found. Two variogram structures $L = 2$ are used: the first variogram structure is a spherical model with a range of 20000 m, the second variogram structure is an exponential model with a range of 45000 m. This variogram model can be expressed as:

$$\gamma_{kk'}(\mathbf{h}) = b^0_{kk'} + b^1_{kk'} \cdot \text{Sph}(\mathbf{h})_{a=20000} + b^2_{kk'} \cdot \text{Exp}(\mathbf{h})_{a=45000}, \quad \forall k, k'$$

To fit the experimental omnidirectional direct and cross variograms to a LMC, the LMC variogram fitting (`varfit_lmc`) program developed by Larrondo, Neufeld, and Deutsch (2003) is used. The resulting direct variogram models from the LMC are illustrated in Figure 4.12. As displaying the 861 cross variograms is excessive, a subset is displayed in Figure 4.13. Only cross variograms

from the LMC with an absolute cross-correlation greater than $|0.65|$ are illustrated as they are the most informative.

4.4.2.3 Implementation Challenges

The proposed framework is implemented in the latest version of the program `USGSIM` released by Manchuk and Deutsch (2015). The size of the system of equations increases with an increasing number of variables that require modeling. If the number of conditional data available is limited to 10, a matrix with a maximum size of 486 by 486 is multiplied by a matrix with a maximum size of 486 by 42. With this configuration, it takes 0.0368 seconds per estimate. With four million cells requiring simulation, it takes approximately 41 hours to simulate a single realization (J.G. Manchuk, personal communication, March 19, 2016). Due to the computational cost, this framework is abandoned.

4.4.3 Conclusion

It is evident that the variograms illustrated in Figures 4.12 and 4.13 are not adequately fit. This is due to the inability of the fitting algorithm of `varfit_lmc` to meaningfully alter the $\mathbf{B}^l$ matrices from their starting position due to the condition that they must remain positive definite. In Figure 4.12, the direct variogram models for mercury (Hg) appears reasonably fit by the LMC; however, more often than not, this is not the case. If a variables input spatial correlation structure is similar to what the LMC starts at during the fitting procedure, then it is likely to be fit well. Of the cross variograms displayed in Figure 4.13, this issue is also apparent.

It is demonstrated by Guo and Deutsch (2002) that for a 2-D problem, kriging variance converges and the kriging weights become effectively 0 when 40 conditioning data are used. If the simulation is limited to less than 40 conditioning data the accuracy of the resulting models will be affected. A maximum number of 10 conditioning data is considered. This is significantly less then the recommended maximum of 40. By increasing the number of conditioning data to 40, the procedure becomes even more expensive.

The poorly fit direct and cross variograms is concerning on its own; however, when combined with the computational cost, it becomes clear that the cokriging framework for simulating massively multivariate datasets is not practicable.

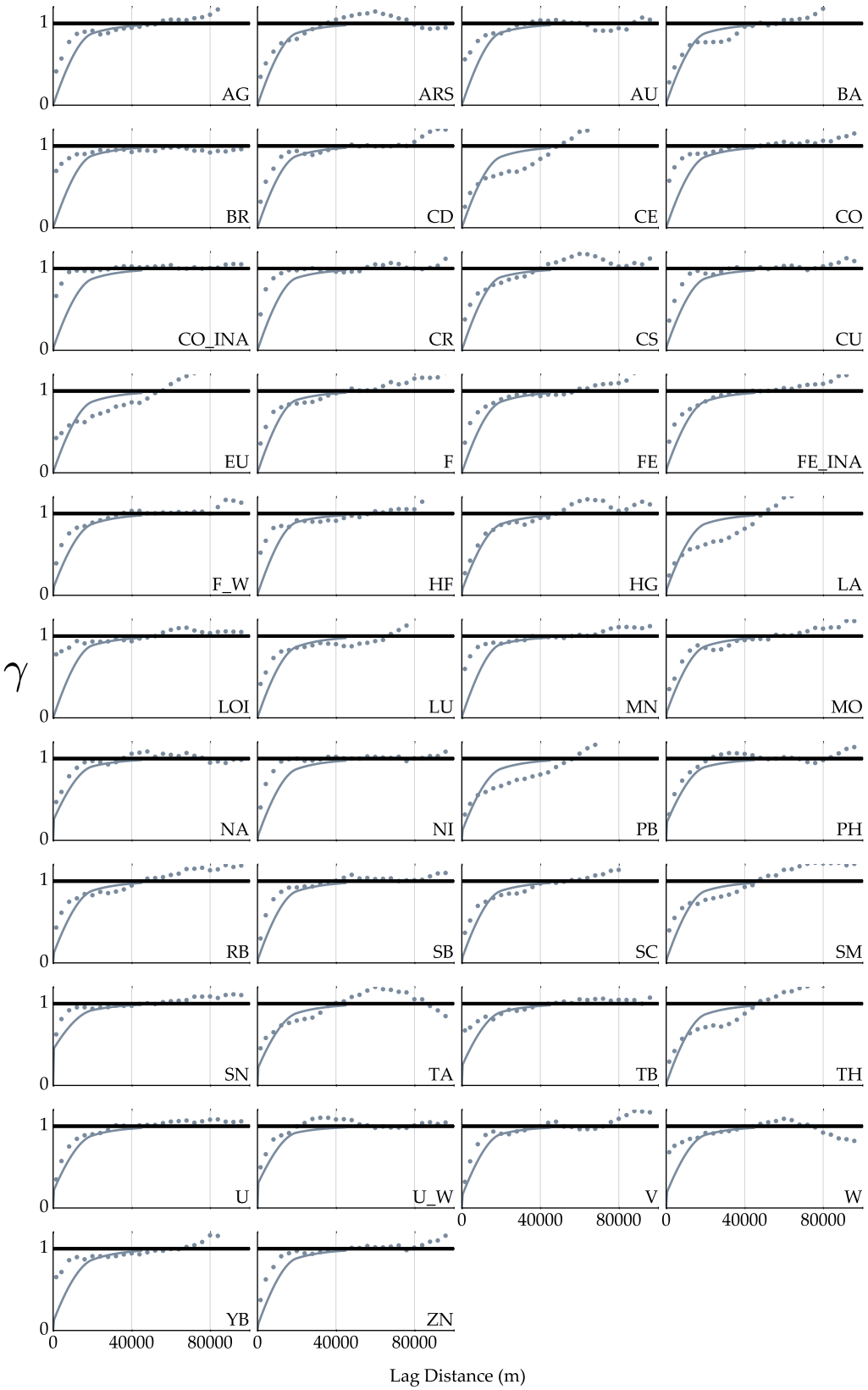

Figure 4.12: Direct variogram models from the LMC for all 42 primary variables in normal score space.

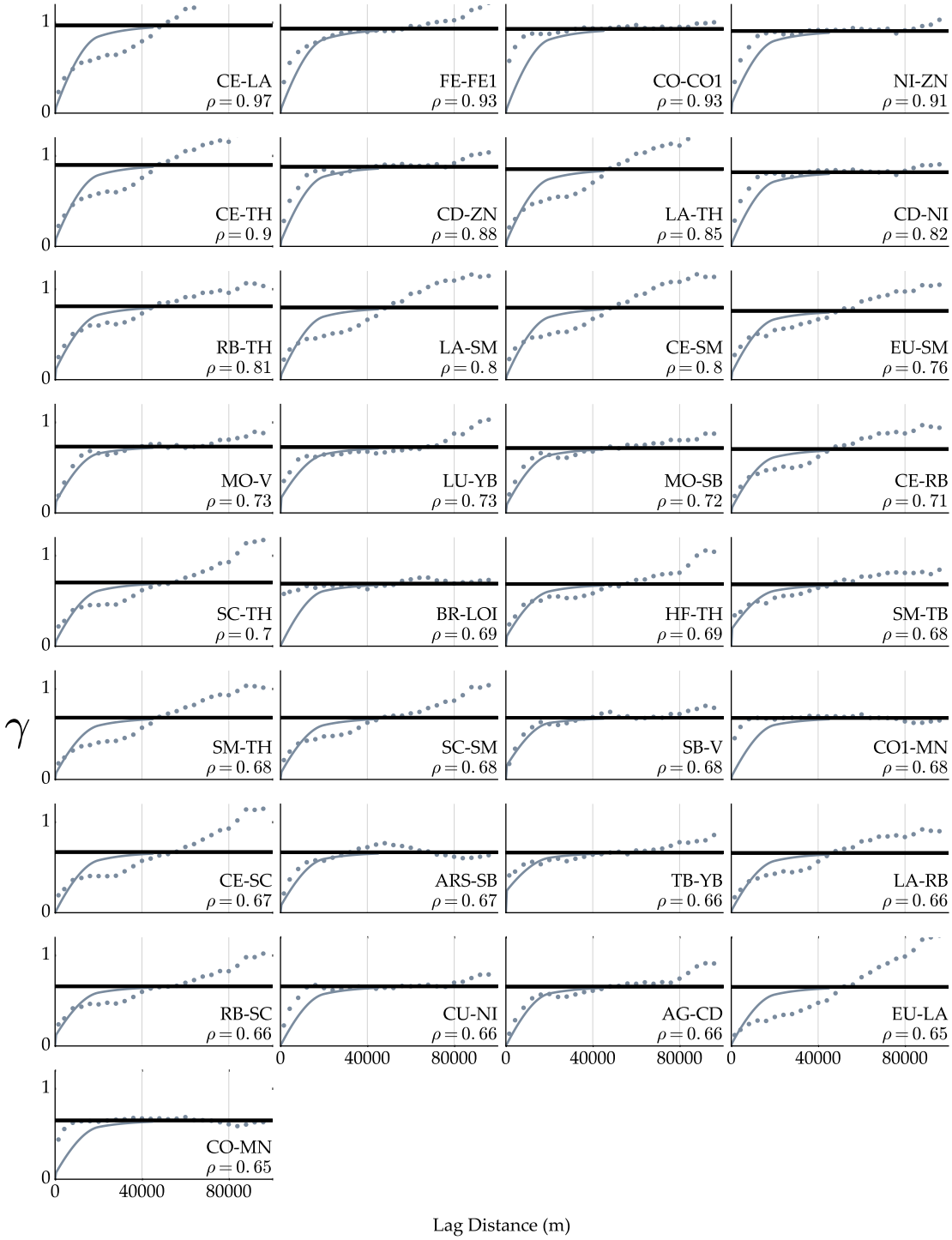

Figure 4.13: Cross variogram models from the LMC that have an absolute cross-correlation greater than 0.65 in normal score space.

4.5 Hierarchical Framework

4.5.1 Introduction

As discussed in Section 2.2.5, the hierarchical framework reproduces input multivariate relationships. Previously modeled variables are used to condition subsequently modeled variables, allowing all multivariate relationships to be honored.

The following section describes the implementation of this framework in a case study that uses jackknife cross-validation as a means to evaluate the predicted models accuracy and precision. The determination of the variable hierarchy and the results are described.

4.5.2 Case Study

4.5.2.1 Methodology

SGS with ICCK is run conditional to the stream sediment samples and secondary data to model the $K = 42$ primary variables using a hierarchical framework. The direct variograms used for each of the the primary variables are modeled and a hierarchy is specified. Conditioning the model to the sequentially generated super-secondary variables with ICCK considers collocated primary-primary and primary-secondary correlation. One hundred realizations is deemed to provide a reasonable assessment of uncertainty. Jackknife cross-validation is implemented as a means of model checking; 20% of the stream sediment samples are randomly removed, leaving the remaining 80% of the samples to condition the spatial prediction. The process is summarized as follows:

1. Normal score transform the primary and secondary variables
2. Model the primary variables variograms
3. Determine a hierarchy
4. Remove 20% of the primary data and reserve for model checking
5. For each realization, complete the following:
    a. For each primary variable in the order specified by the hierarchy, complete the following:
        i. Merge any previously simulated primary variable(s) with the exhaustive secondary variables into a single super-secondary variable
        ii. Simulate a single realization
        iii. Add the realization to the pool of previously simulated primary variables
6. Back-transform realizations to original units

4.5.2.2 Implementation

Variography

The hierarchical workflow requires direct variograms for each of the primary variables. Of the 42 primary variable variograms 15 are modeled with a omnidirectional variogram while the remaining 27 are modeled using a major and minor direction of continuity (Figure 4.14).

Determining Hierarchy

Two criteria are used to determine the order variables are modeled: (1) each variables correlation with its corresponding super-secondary variable built from the six exhaustive secondary variables and (2) the continuity of the primary variables variogram.

The correlation between a primary variable and its corresponding super-secondary variable calculated from the exhaustive secondary data is calculated. The primary variables are then sorted based on their absolute correlation coefficient in decreasing order. This ensures that the primary variables that are most informed (i.e., high correlation with super-secondary variable) are modeled first. It is thought that this method of sorting improves the estimates of subsequent primary variables that initially are not well informed (i.e., low correlation with super-secondary variable).

The continuity of each variables variogram model is evaluated and used to create a second rank value in decreasing order. To do this, the variogram values ranging from $(0.0, sill]$ are discretized into three zones. The three zones range from $(0.00, 0.20]$, $(0.20, 0.50]$, and $(0.50, 1.0]$ (Figure 4.15). Within each zone, the $\mathbf{h}$ values for each 0.01 variogram value interval is calculated and summed together. This processes is completed for each of the zones, generating a rank for each zone. The three zones ranks are then summed and used to rank the variables in increasing order. The lower the summed rank, the more continuous the variogram model is for the purpose of this ranking scheme. To place a higher emphasis on the variogram models short range, two of the three zones cover the first half of the variogram value range. This ensures that the primary variables that are most continuous are modeled first. It is thought that this method of sorting improves the estimates of subsequent primary variables that initially are not as informed (i.e., less long range conditioning).

The correlation and continuity rankings are used to determine the hierarchy by setting the correlation rank as the primary determinant and allowing the continuity rank to shift the correlation rankings by ± 5 rankings. For example, if a variable has a correlation rank of 7, the highest final rank it can occupy is 2 and the lowest rank it can occupy is 12.

It is thought that this procedure results in a more accurate model by delaying prediction of variables that are not as informed or continuous as other primary variables. With additional conditioning data (i.e., previously simulated primary variables), the correlation between the variable being modeled and its corresponding super-secondary variable may increase, improving spatial prediction. This idea is not specifically tested in this thesis.

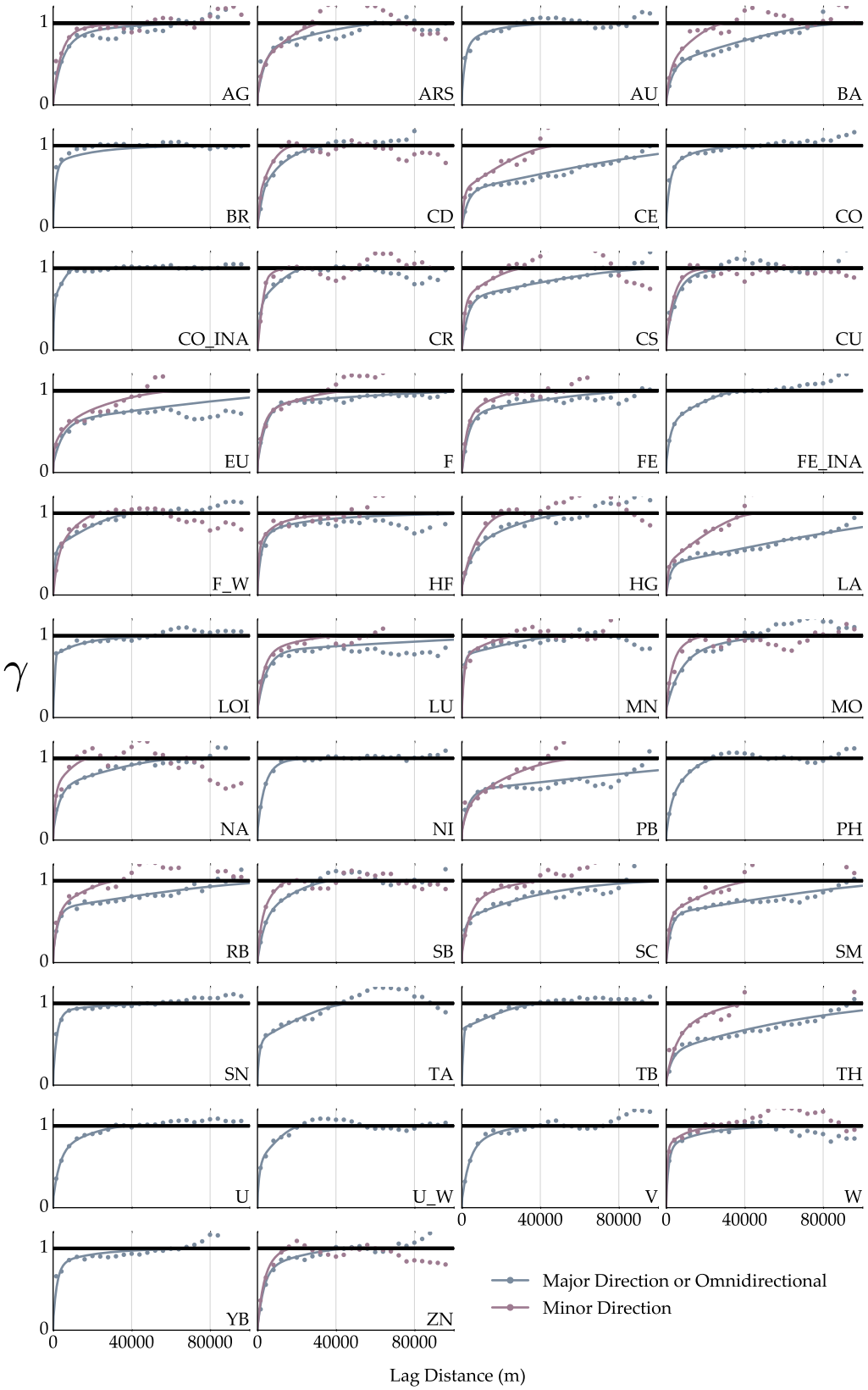

Figure 4.14: Modeled variograms for all 42 primary variables.

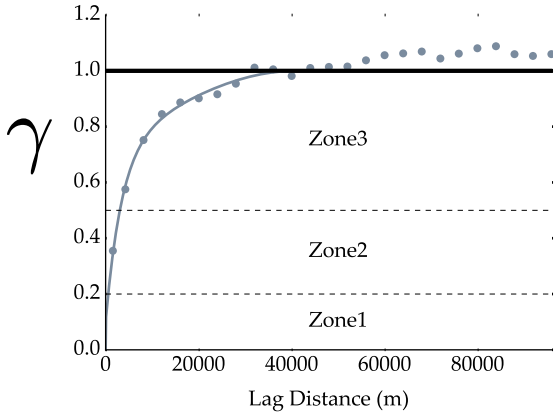

Figure 4.15: Zones used to calculate the continuity rankings when determining the hierarchy of the primary variables

Workflow Implementation

The hierarchical simulation workflow is scripted within a Python environment (Continuum Analytics, 2015), utilizing specialized geostatistical software; specifically `USGSIM` (Manchuk & Deutsch, 2015) and the geostatistical Python package `pygeostat` (Centre for Computational Geostatistics, 2016b). The script is run in parallel using five cores; each computing a single realization at a time. For example, a single core completes the following procedure:

1. Create super-secondary variable combining all exhaustive secondary data.
2. Simulate a single realization of the first primary variable in the specified hierarchy.
3. For the remaining primary variables, complete the following in the sequence specified by the hierarchy:
   a. Combine all previously simulated primary variable for the current realization including the exhaustive secondary variables.
   b. Simulate a single realization of the primary variable currently being modeled.

To complete the process using an Intel i7-4790 3.60 gigahertz (GHz) processor and 16.0 gigabyte (GB) of random access memory (RAM), the simulation takes approximately 2 days and 17 hours to complete.

4.5.2.3 Results

To summarize histogram and variogram reproduction, two plots are selected to illustrate the best and worst reproduction (Figure 4.16). Reproduction plots of the other specific variables are not shown or discussed in detail. The histograms of the simulated realizations for manganese (Mn) in normal score space shows that the input histogram is reproduced very well. Conversely, the variable reproduction of cobalt (Co) is the worst of the 42 variables modeled and shows that all realizations are underestimated. Consistent with the two variables shown here, all of the variables

modeled display very little variation between realizations. Additional summary statistics for all variables are illustrated in Figure 4.17 and shows very little variation.

Variogram reproduction varied as illustrated in Figure 4.18 that shows the best and worst variogram reproduction. All of the variables variogram reproduction appear reasonable.

There are 861 primary-primary variable correlation coefficients, each reproduced in its own way. As an example, three are illustrated in Figure 4.19. To better understand the overall performance, the coefficients for all of the realizations are averaged then subtracted by its corresponding input correlation coefficient producing an error measure (Figure 4.20). All of the 861 average coefficients are within ± 0.15 of their corresponding input value.

The results of the Jackknife for each of the primary variables modeled in normal score space are illustrated in Figure 4.21 and ordered based on their MSE. The accuracy and precision of each variable appears reasonable. In the worst cases, 10% of the true values do not fall within the specified probability interval. This may indicate that the variograms of the variables that displaced this error may be too continuous.

As a means of visualizing multiple realizations, the cell-by-cell average (i.e., E-type average) and variance (i.e., E-type variance) of the realizations is calculated and displayed in Figures 4.22 and 4.23 respectively. The influence of the exhaustive secondary data is clear in the E-type estimates as features from the secondary data is visible. The E-type variance illustrate the reproduction of the conditioning primary data and the influence the multivariate relationships have on reducing the variance of lead (Pb) (Figure 4.23a) relative to zinc (Zn) (Figure 4.23b).

4.5.3 Conclusion

Overall, histogram reproduction was reasonably achieved as shown by the histograms of Figure 4.17. The mean of all realizations in normal score space for all primary variables modeled were within $\pm 6\%$ of their input mean. Additionally, all of the modeled primary variables displayed minimal variance between each realizations distribution (Figure 4.17b).

It is thought by the author that a correlation reproduction error less then $|0.20|$ is reasonable. As the average correlation coefficient from each primary variables realizations are within this range, in addition to that the variation between realizations is minimal, it can be said that the hierarchical simulation framework also reproduced collocated multivariate relationships.

Variogram reproduction was not achieved for all variables. It is thought that better variogram reproduction can be achieved with proper decisions of stationarity.

The Jackknife study also showed positive results, with most variables displaying acceptable accuracy and precision. Fair accuracy results shown by some variables may indicate their variograms are too continuous. Adjustments to the variogram model may improve model accuracy in these cases.

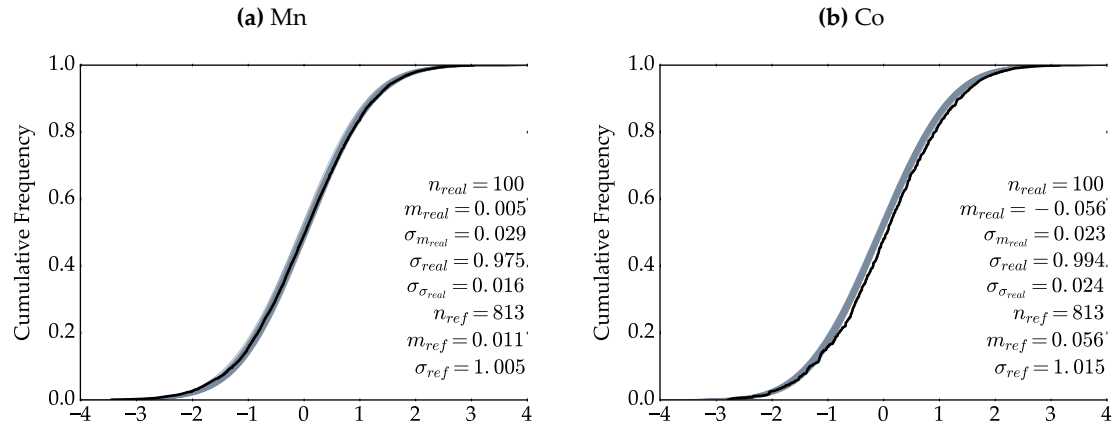

Figure 4.16: Examples of the variables with the best (Mn) and worst (Co) histogram reproduction in normal score space.

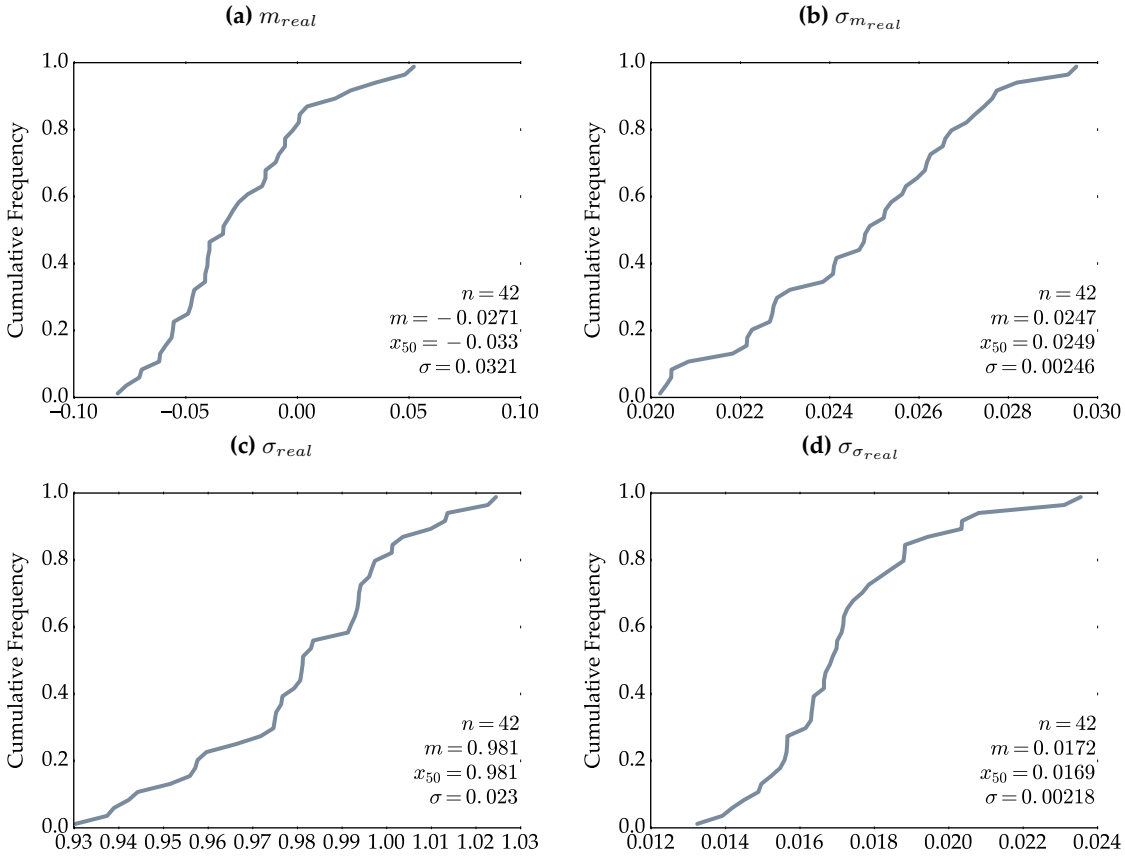

Figure 4.17: Histograms of histogram reproduction summary statistics from each of the 42 primary variables realizations in normal score space.

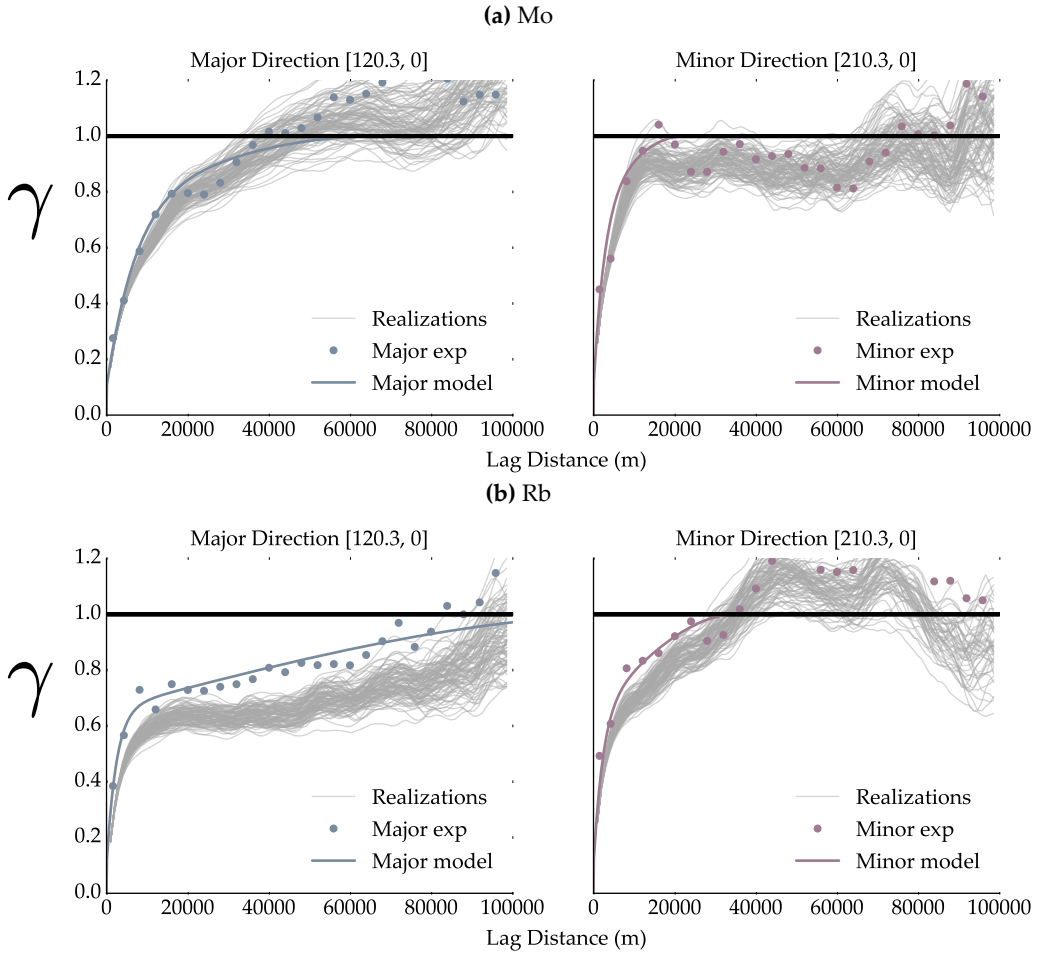

Figure 4.18: Example of the variables with the best (Mo) and worst (Rb) variogram reproduction in normal score space.

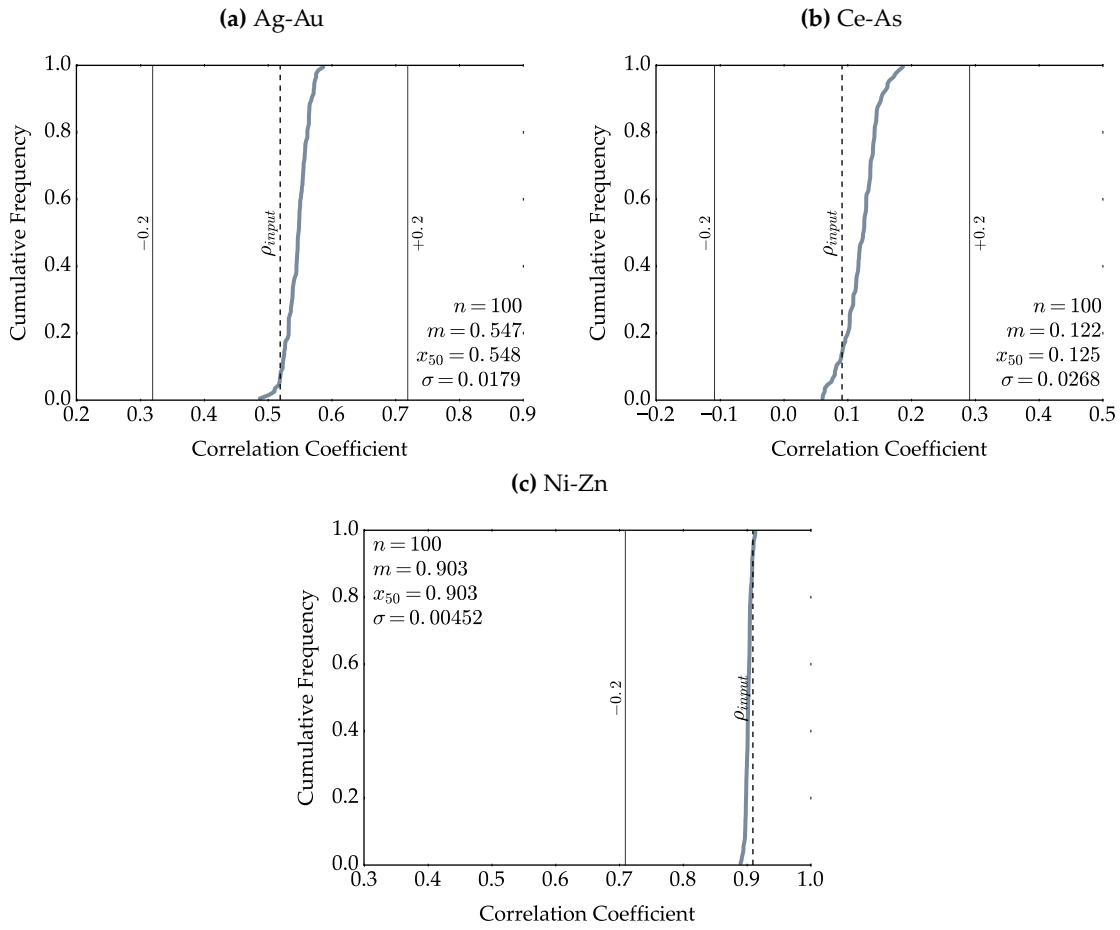

Figure 4.19: Histograms of two primary-primary variable correlation coefficients calculated from 100 simulated realizations with its input value indicated.

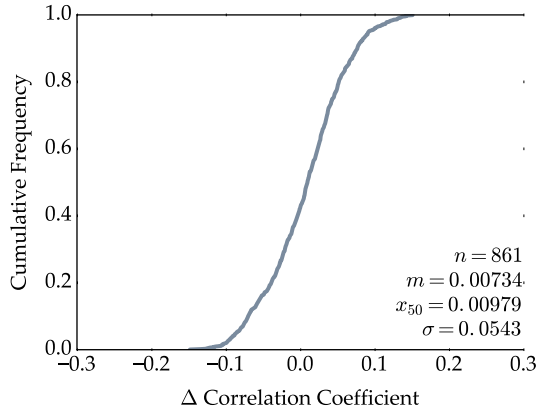

Figure 4.20: Histogram of correlation coefficient errors for all primary-primary variable pairs between the average value across all realizations and the input correlation matrix.

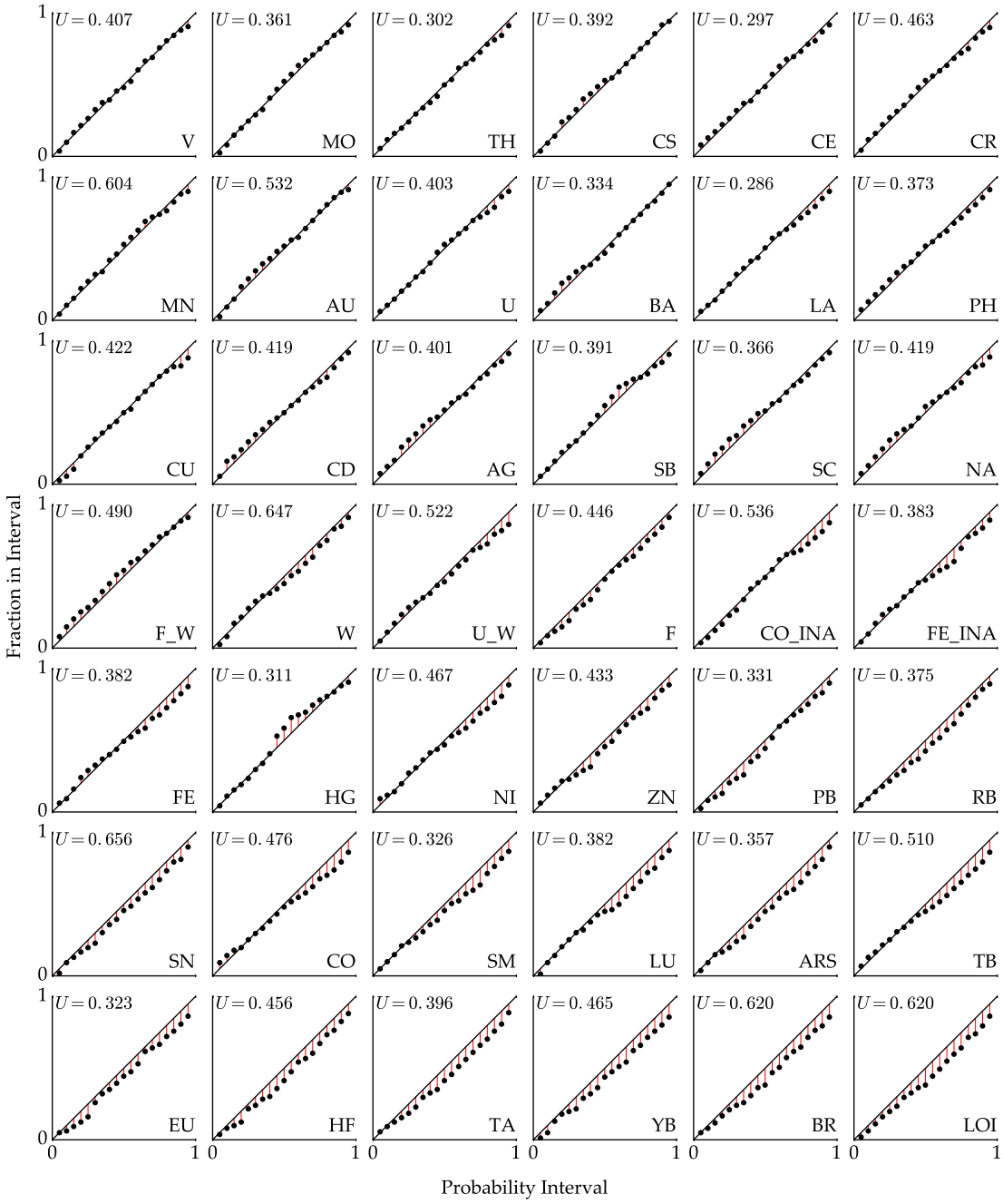

Figure 4.21: Accuracy plots depicting results from the Jackknife cross-validation study.

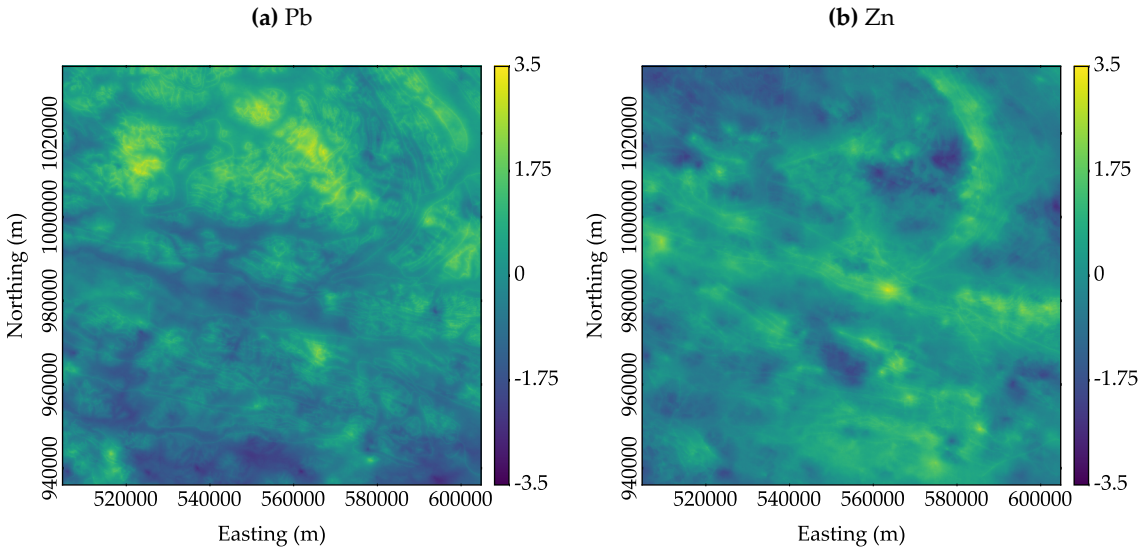

Figure 4.22: Heatmaps of the cell-by-cell average of the realizations (i.e., E-type estimate) in normal score space

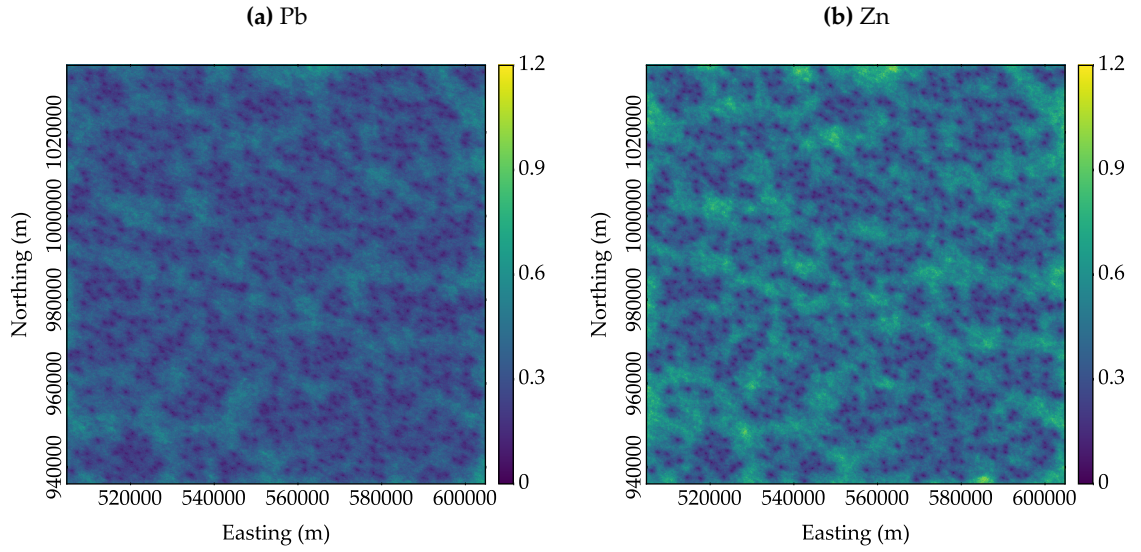

Figure 4.23: Heatmaps of the cell-by-cell variance of the realizations (i.e., E-type variance) in normal score space

4.6 Conclusion

Many issues exist when attempting to use a cokriging framework with a massively multivariate system. Due to the constraining nature of fitting an LMC to many variables, it is unable to adequately fit direct and cross variograms as the fitting algorithm can move very little from its initial starting position. Additionally, the computation cost of solving the large system of equations required by the framework with an appropriate amount of conditioning data is expensive. These two factors render the framework impractical.

The hierarchical framework provides a viable option for simulating multivariate data with many

secondary data. In the case study, collocated correlation structure and input histograms were adequately reproduced. Variogram reproduction was not as close; however, it is thought that with proper care and attention to decisions of stationarity, this concern can be minimized, if not eliminated. The results from the Jackknife study showed that uncertainty was predicted in a precise and accurate manner. The workflow is cumbersome to implement and debugging is a challenge due to the number of program calls. With scripting, the task is practical and repeatable.

Chapter 5

Probabilistic Mineral Prospectivity Modeling

5.1 Introduction

As discussed in Sections 1.2 and 2.1.3, current MPM frameworks integrate factors that are determined—either statistically or based on expert knowledge—to be relevant in mineral exploration in order to produce a model illustrating the spatial distribution of relative prospectivity. This process is complicated due to the massively multivariate nature of geochemical and remotely sensed data, most of which have unequal support. The key idea is to search for analogies of known mineral deposits. While the known deposits illustrate positive information, it is censored in that negative information is not used or is typically not as extensive or accurate as the positive data set.

A novel MPM framework is proposed that provides a means of passing a stochastic multi-element model and other relevant geological data to a transfer function that generates a model illustrating the probability that a mineral deposit exists at each location. By using a multi-element geochemical model, both positive and negative information is equally represented. Bias searches are avoided by not using the location of known mineral deposits to discover exploration factors that indicate the presence of a mineral deposit. In addition, by having multiple realizations of the geochemical model, uncertainty can be be transferred to the final prospectivity modeling that is represented by a single probabilistic value at each location.

This chapter describes the proposed framework, the type of data that can be passed to the MPM transfer function, and describes a small example workflow demonstrating how the proposed framework may be implemented. The transfer function defined in the example workflow is passed the stochastic multi-element geochemical model generated using the hierarchical framework in Section 4.5 in addition to bedrock geology data (Yukon Geological Survey, 2016b) to provide illustrative results.

5.2 Proposed Framework

5.2.1 Introduction

The proposed framework passes a stochastic geochemical model and exhaustive secondary data to a MPM transfer function. For each realization within the geochemical model, a binary response is calculated that indicates the presence or absence of a deposit. The average of the binary reposes

across all realizations at each location produces a probability that a mineral deposit exists at that location.

By having the full spectrum of possibilities represented by the stochastic geochemical model, concerns relating to heuristic searches and censored data are minimized, while having multiple realizations allows uncertainty to be passed to the final predictive model. Discussion regarding the required data and the function methodology is provided.

5.2.2 Input Data

The proposed framework requires the generation of a stochastic multi-element geochemical model using a simulation framework (Figure 5.1a). This may be done using the methodology discussed in Section 4.5, where stream sediment samples are conditioned by exhaustive secondary data. Rather then premeditatedly determine what elements to predict and include in the geochemical model, it is thought that modeling all possible elements provides the flexibility to change the MPM transfer function after its initial construction. In addition, the same geochemical model could then be passed to multiple MPM transfer functions each designed for different deposit types. Additional geological features are represented by exhaustive secondary data (Figure 5.1b).

(a) Stochastic geochemistry model **(b)** Exhaustive secondary data

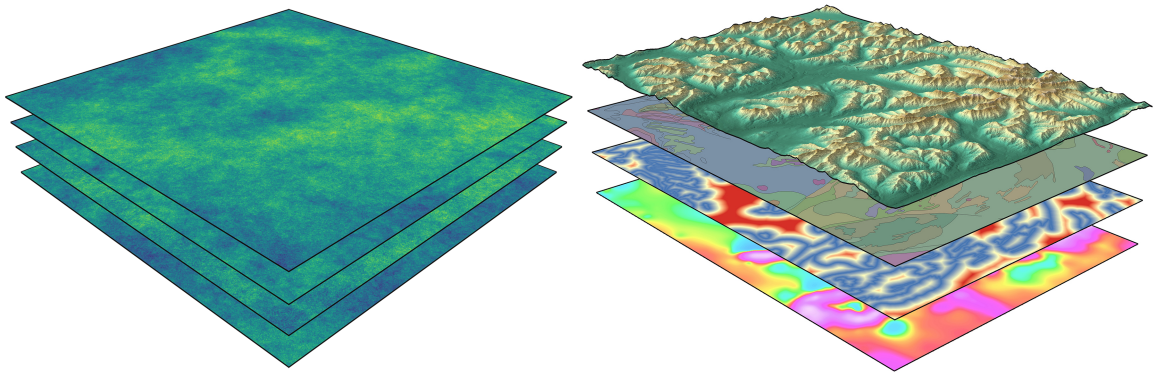

Figure 5.1: Schematic illustration of the datasets used in the proposed MPM framework.

5.2.3 Transfer Function Methodology

Consider K number of interdependent RFs that are stochastically modeled $\{z_k^l(\mathbf{u}), k = 1, \ldots, K, l = 1, \ldots, L, \forall \mathbf{u} \in A\}$ for L number of realizations at all grid locations $\mathbf{u}$ within the domain A, that is also denoted by the vector $\mathbf{Z}(\mathbf{u})$. In addition, I number of interdependent exhaustively sampled RFs exist $\{x_i(\mathbf{u}), i = 1, \ldots, I, \forall \mathbf{u} \in A\}$ that are also denoted by the vector $\mathbf{X}(\mathbf{u})$.

The transfer function consists of R number of binary functions $\{F_r, r = 1, \ldots, R\}$ defined by the user that is also donated by the vector $\mathbf{F}$. The binary functions $\mathbf{F}$ are constructed in a way that when $\mathbf{Z}(\mathbf{u})$ and $\mathbf{X}(\mathbf{u})$ is passed to it, R number of binary responses are generated $\{i_r^l(\mathbf{u}), \forall l, r, \mathbf{u}\}$

using the expression:

$$i_r^l(\mathbf{u}) = F_r(z^l(\mathbf{u}) \text{ or } x(\mathbf{u})), \quad \forall l, \mathbf{u} \in A$$

For each location and each realization, a binary indicator of any mineral deposit $\{d^l(\mathbf{u}), \forall l, \mathbf{u} \in A\}$ is calculated using the expression:

$$d^l(\mathbf{u}) = \begin{cases} 1, & \text{if } i_r^l(\mathbf{u}) = 1 \quad \forall r \\ 0, & \text{otherwise} \end{cases}, \quad \forall l, \mathbf{u} \in A \tag{5.1}$$

that are also denoted by the vector $\mathbf{D}(\mathbf{u})$. The rule set that constructs the transfer function $\mathbf{F}$ indicates the presence or absence of specific deposit types. Exploration factors for the deposit type under investigation are determined and their relationships to the geochemical model variables or exhaustive secondary data are identified and evaluated using the transfer function.

The final probabilistic value at each location is then calculated by averaging $\mathbf{D}(\mathbf{u})$ at each location:

$$p(\mathbf{u}) = \frac{1}{L} \sum_{l=1}^{L} d^l(\mathbf{u}), \quad \forall \mathbf{u} \in A$$

For illustration purposes, Figure 5.2 displays a histogram of $\mathbf{D}(\mathbf{u})$ at a single location with $L = 100$ realizations. The average of those responses $p(\mathbf{u})$ is the probability of a mineral deposit at that location. Ideally, uncertainty in the probabilistic predictions across all locations is low, as illustrated in Figure 5.3a. In this case, there are minimal locations displaying a moderate probability of a mineral deposit, rather each location displays a low or high probability. Conversely, higher uncertainty in the final model, as illustrated in Figure 5.3b, is present when the frequency of locations displaying moderate probability of containing a mineral deposit is increased.

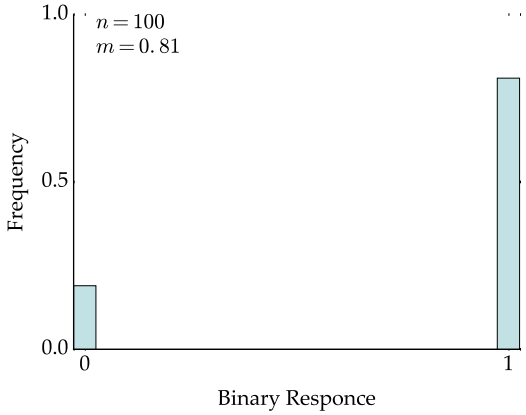

Figure 5.2: Example of the possible outcome $\mathbf{D}(\mathbf{u})$ from the transfer function at a single location across all realizations

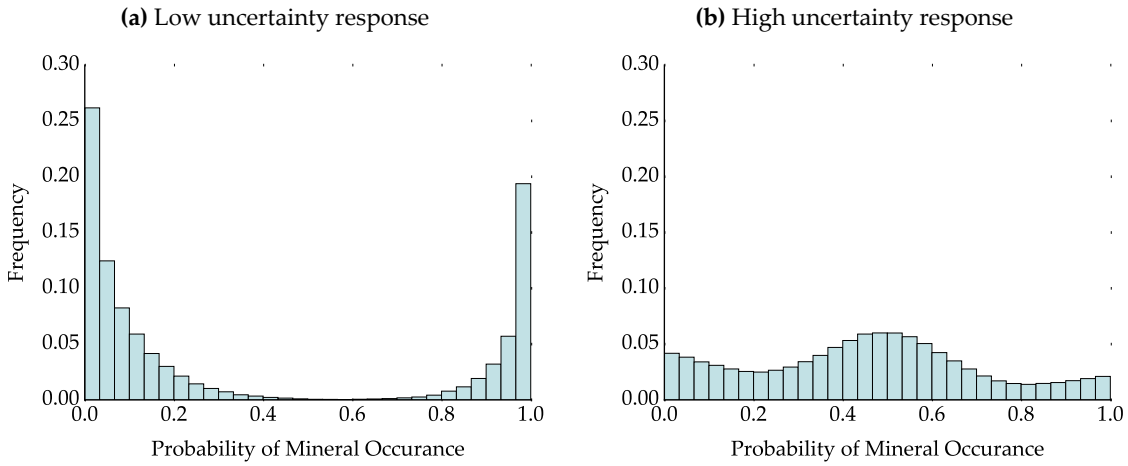

Figure 5.3: Example of the possible histogram of the final prospectivity model values $\{p(\mathbf{u}), \forall \mathbf{u} \in A\}$.

5.3 Clastic-Dominated Pb-Zn Deposit Model Illustration

5.3.1 Introduction

As a means of a demonstrating the proposed MPM framework, a small example workflow is illustrated. The transfer function is designed to predict the probability of clastic-dominated lead-zinc (CD Pb-Zn) deposit discovery at each location within the AOI selected in Section 4.2.2 (Figure 4.1). A brief introduction to CD Pb-Zn deposits is provided, exploration factors and their geochemical proxies are discussed, a rule set for a transfer function is proposed, and results are discussed.

5.3.2 Clastic-Dominated Lead-Zinc Deposits

CD Pb-Zn deposits are a subtype of sediment-hosted Pb-Zn deposits that includes deposits traditionally refereed to as sedimentary exhalative (SEDEX) deposits. They are important sources of Pb and Zn with some deposits containing economical quantities of silver (Ag), copper (Cu), and/or gold (Au). CD Pb-Zn deposits are precipitated from basinal brines within passive margin, continental rift, and sag basin tectonic settings that are dominated by sedimentary sequences. Host rocks include shale, sandstone, siltstone, or mixed clastic rocks (Leach et al., 2010). Deposition occurs in syngenetic to early diagenetic environments or during early burial diagenesis meaning mineralization occurs at the same time as deposition or close to. The tectonic setting is the main controls on mineralization and deposit characteristics (Leach et al., 2005). The primary ore minerals are typically sphalerite and galena with secondary iron sulfides. Gangue mineralogy typically consists of barite and carbonates (Leach et al., 2010).

5.3.3 Transfer Function

Based on the brief deposit description provided in Section 5.3.2, three exploration factors must be present for a CD Pb-Zn deposit to exist: (1) an anomalous Pb geochemical signature, (2) an anomalous Zn geochemical signature, and (3) shale, sandstone, siltstone, or mixed clastic host rock. Proxies for these exploration factors are found within the geochemical model for both Pb and Zn that was simulated in Section 4.5 and the bedrock geology data (Yukon Geological Survey, 2016b) sourced from Geomatics Yukon (2014).

For simplicity, bedrock geology is reclassified into a binary system: clastic or not clastic (Figure 5.4). To determine anomalous geochemical signatures for Pb and Zn, probability plots are used to identify outliers (Figure 5.5). It is determined that Pb anomalies display concentrations ≥ 35 parts per million (ppm) and Zn anomalies display concentrations ≥ 2500 ppm.

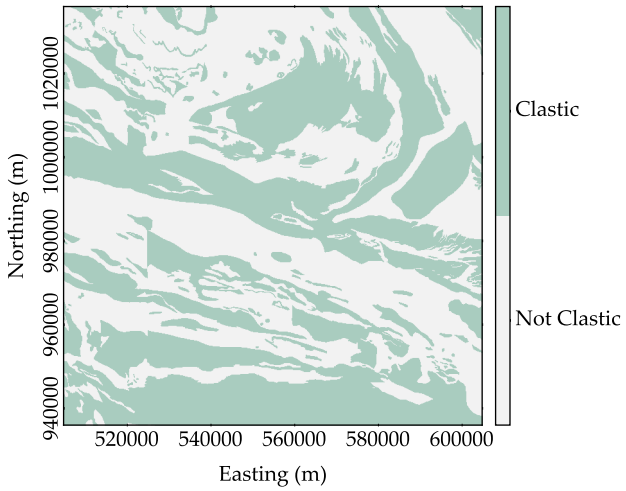

Figure 5.4: Lithology reclassified into a binary system within the AOI.

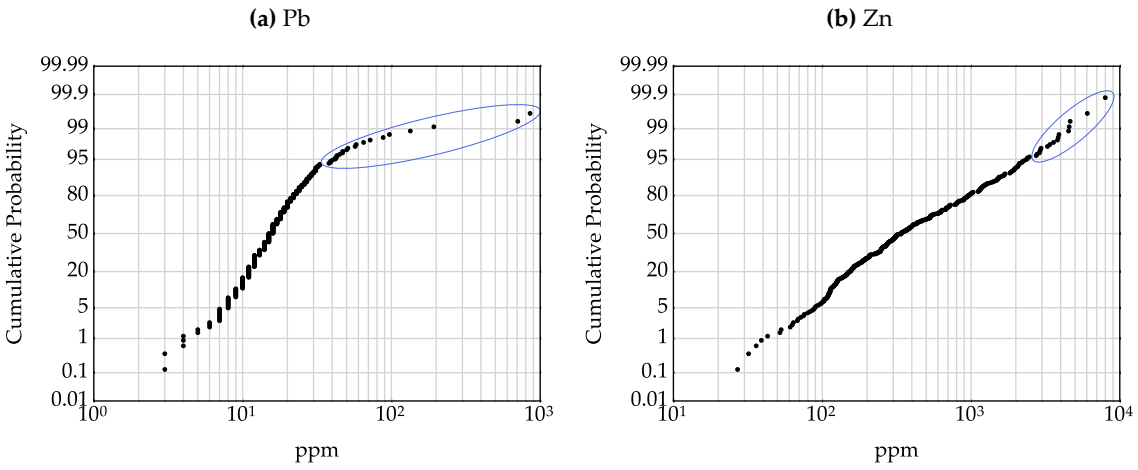

Figure 5.5: Probability plots of the two geochemical variables in original units with the identified range of anomalous signatures.

To determine if a CD Pb-Zn deposit exists at a location the following binary rules must be satisfied: (1) the Pb signature must $\geq$ 35 ppm, (2) Zn signature must be $\geq$ 2500 ppm, and (3) the lithology must be clastic. In practice, these rules would be fine tuned for each geological province and considering local geological conditions.

5.3.4 Results

The stochastic geochemical model for Pb and Zn simulated in Section 4.5 and the binary lithology data is passed to the transfer function generating a predictive model (Figure 5.6). Very little uncertainty is observed as 99.939% of the four million locations display 0% probability of CD Pb-Zn deposit discovery. Of the 2425 locations that displayed some chance of discovery, only 27 locations showed probability $\geq 3\%$ (Figure 5.7).

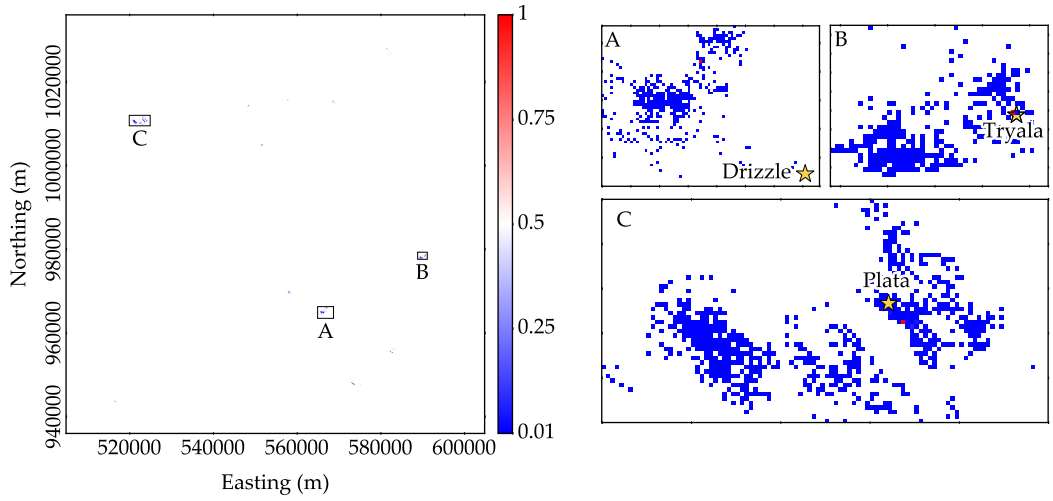

Figure 5.6: Map illustrating the final probabilistic model of CD Pb-Zn deposit discovery. Mineral occurrences (Yukon Geological Survey, 2016a) are displayed near high potential areas.

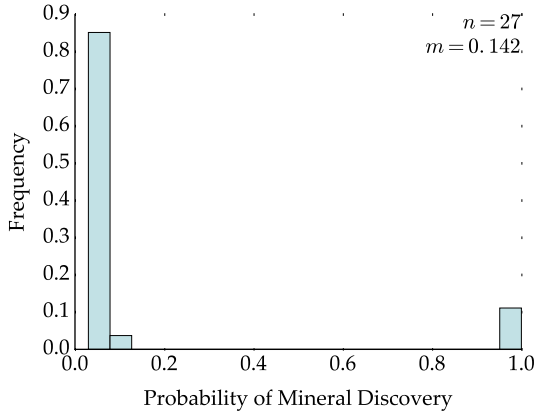

Figure 5.7: Histogram of the locations probability of CD Pb-Zn deposit discovery that displayed probabilities $\geq 3\%$.

As a means of validation, current mineral occurrences were sourced from Yukon Geological Survey (2016a) and plotted against the three locations that displayed 100% probability of CD Pb-Zn deposit discovery (Figure 5.6). All of these locations occur in very close proximity or near a mineral occurrence. The location displayed in Figure 5.6a lies approximately 2.7 km to the shale-hosted nickel (Ni)-Zn-barium (Ba) showing Drizzle. The location displayed in Figure 5.6b nearly overlies the drilled Tryala prospect, a sediment-hosted Ba-Zn deposit. The location displayed in Figure 5.6c lies directly adjacent to past producer Plata, a Au-Pb-Ag vein deposit.

5.3.5 Conclusion

In this small example workflow, the transfer function may be too restrictive in that nearly no locations displayed an intermediate probability of deposit discovery. The 3 locations that displayed high probability of deposit discovery were in very close proximity to either a showing, drilled prospect, or past-producer, which is a promising sign. However, these occurrences are not of the desired deposit type.

While only two elements from the geochemical model that contains 42 predicted elements were used, in practice more complex geochemical rules would be defined utilizing more elements. As all elements measured in the stream sediment samples were predicted, the flexibility of the geochemical model is maximized allowing the transfer function to be modified as desired. In cases where the elements predicted and included in the geochemical model were tailored to a predefined transfer function, if additional elements are later required, a new geochemical model would need to be predicted.

5.4 Conclusion

The CD Pb-Zn deposit MPM example workflow shows promise for the proposed framework, however, it requires thorough investigation and evaluation. It is hoped that with a stochastic geochemical model calculated to suit the needs of a carefully constructed transfer function, that MPM can be completed in an unbiased manor considering the full spectrum of the truth, while passing uncertainty through the workflow.

Chapter 6

Conclusions

6.1 Introduction

Exploratory research presented above has focused on understanding current MPM frameworks and their shortcomings. Their dependence on heuristic methodologies and censored data that only explain factors that indicate the presence of mineral deposit and not their absence is limiting. The stochastic multi-element geochemical model provides both positive and negative information and allows uncertainty to be transferred to a final prospectivity model. Multiple realizations of the geochemical model and additional exhaustive secondary data can then be passed to a MPM transfer function that calculates the probability that a mineral deposit exists at each location. This proposed framework alleviates the concerns of heuristic searches and the use of censored data.

The desire to develop a novel and improved MPM framework motivates this thesis. Two novel multivariate simulation frameworks that consider many exhaustive secondary data are formulated and tested in hopes of developing a straightforward technique; however, neither proved to be adequate. The ability of the cokriging and hierarchical frameworks to model 42 variables in the presence of 6 exhaustive secondary variables is evaluated. The cokriging framework is found to be computationally expensive and limited by the restrictive nature of the LMC when a large number of variables are considered. The hierarchical framework reasonably reproduces input univariate and multivariate statistics and produced the stochastic geochemical model that was passed to a contrived MPM transfer function that shows promise; however, the concept requires additional research. A summary of the contributions made by this thesis and their conclusions are described below with a discussion regarding future research.

6.2 Research Contributions

6.2.1 Novel Frameworks for Multivariate Simulation with Many Secondary Data

To generate a stochastic multi-element geochemical model, a framework that can reproduce the input univariate and multivariate statistics in the presence of many exhaustive secondary data is required. The cokriging framework can achieve this; however, it is limited due the need of a LMC. There may be a practicable limit of about seven variables (Jewbali, 2009) motivating two novel and straightforward multivariate simulation frameworks were presented and tested in Chapter 3.

The decorrelation framework was implemented using both PCA and Sphere-R as methods to transform the variables requiring modeling into uncorrelated factors. Once correlation is removed,

the factors can be independently simulated and the original multivariate relationships is reintroduced during back-transformation. However, by conditioning the uncorrelated factors to exhaustive secondary data, correlation between the simulated factors was introduced. This rendered the back transformations of both PCA and Sphere-R unable to reproduce the input multivariate relationships.

A novel framework based on the direct cosimulation framework (Verly, 1993) was proposed that would jointly simulate multiple variables while considering many exhaustive secondary data. It operates by independently simulating the variables that require modeling and correlating the deviates that sample the conditional distributions calculated at each location. Secondary data is considered by seeding the known values into the calculation of the deviates. Implementation is very straightforward; however, variance inflation was an issue and in some cases extreme. Ad-hoc empirical corrections were implemented that could not fully control the issue. In addition, multivariate relationships were not reasonably reproduced.

6.2.2 Implementation of Multivariate Simulation with Many Secondary Data

The cokriging and hierarchical frameworks were tested in Chapter 4 by implementing them in case studies and evaluating their ability to model 42 primary variables in the presence of six exhaustive secondary variables.

A LMC was fit to the variables and the cokriging framework successfully ran. However, the LMC poorly fit most of the 903 direct and cross-variograms as the algorithm was unable to meaningfully alter the initial covariance contributions of the variograms. The computational cost of the framework is very expensive, as it would have taken 41 hours to simulate a single realization using only 10 conditioning data (J.G. Manchuk, personal communication, January 11, 2016), which is not practical.

The hierarchical framework was implemented using SGS with ICCK conditional to the stream sediment samples and exhaustive secondary data. The framework proved to be a viable option for simulating multivariate data with many secondary data. It was able to reasonably reproduce the input univariate and multivariate statistics in the presence of many exhaustive secondary data and is practical to implement with scripting.

6.2.3 Probabilistic Mineral Prospectivity Modeling

Once the appropriate framework for simulating multivariate data with many secondary data had been identified, a novel MPM framework was proposed and illustrated in Chapter 5. The proposed framework requires a stochastic multi-element geochemical model and other relevant exhaustive geological data to a transfer function that calculates the probability that a mineral deposit exists at each location.

Contrary to existing MPM frameworks that have heuristic and biased searches, the proposed MPM framework allows the full spectrum of the truth to be considered. By using a multi-element geochemical model, both positive and negative information is considered. Biased searches are avoided by not using the location of known mineral deposits to discover exploration factors that indicate the presence of a mineral deposit. In addition, by having multiple realizations of the geochemical model, uncertainty can be be transferred to the final prospectivity modeling that is represented by a single probabilistic value at each location.

As an illustration, a rule set that contains binary functions is defined that evaluates a input stochastic geochemical model and exhaustive secondary data in hopes of discovering CD Pb-Zn deposits. Three exploration factors are considered by the transfer function and produce a single probabilistic prospectivity model. The contrived transfer function is far too restrictive as very few locations displayed probabilities above 0%. The locations that displayed 100% probability of deposit discovery were in very close proximity to either a showing, drilled prospect, or past-producer. This is a promising sign, however, none were of the desired deposit type.

6.3 Future Research

Much of the work completed in this thesis focused on discovering the appropriate multivariate simulation framework to calculate the geochemical model. This focus of efforts motivates two further areas of research: (1) how stream sediment samples should be utilized in a geostatistical workflow, and (2) the construction of a MPM transfer function that transforms input exploration factors into measures of frequency rather than binary responses.

The complexity of the natural system that deposits the stream sediments is immense and difficult to unravel. The catchment area from which the sampled stream sediments are sourced can be calculated, however, the measured geochemical signature is not a perfect representation of that area. Slope, vegetation, mobility of elements, differential weathering of bedrock, and natural homogenization of sediments are all factors that control the measured geochemical signature. In the case studies completed in Chapter 4, the stream sediment samples geochemical signature were considered at a point support. As stream sediment samples are a representation of an area, this is a simplification. A small exploratory analysis was conducted to better understand the multi-scaled nature of catchment areas and what location within the catchment area is best suited to represent the stream sediment samples as a point of support. The use of these complex measurements in a geostatistical workflow is a topic that warrants future research. Developing methodologies to determine the optimal location of the point of support or adapting geostatistical algorithms to consider an area of support could possibly improve the prediction of the geochemical signature of stream sediment samples.

The MPM transfer function utilized in Chapter 5 is limited due to it consisting of only binary

functions to evaluate the input exploration factors and the limited range of geochemical signatures considered. Developing functions that generate a frequency response would be interesting. The final probabilistic prospectivity model would consider the frequency responses combined into a single probabilistic value.

References

Agterberg, F. P., & Bonham-Carter, G. F. (2005). Measuring the Performance of Mineral-Potential Maps. *Natural Resources Research*, *14*(1), 1–17. doi: 10.1007/s11053-005-4674-0

Agterberg, F. P., Bonham-Carter, G. F., & Wright, D. F. (1990). Statistical Pattern Integration for Mineral Exploration. In *Computer applications in resource estimation* (pp. 1–21). Elsevier. doi: 10.1016/B978-0-08-037245-7.50006-8

Almeida, A. S. (1993). *Joint Simulation of Multiple Variables with a Markov-Type Coregionalization Model* (PhD Thesis). Stanford University, Stanford, CA.

Almeida, A. S., & Journel, A. G. (1994). Joint simulation of multiple variables with a Markov-type coregionalization model. *Mathematical Geology*, *26*(5), 565–588. doi: 10.1007/BF02089242

ALS Limited. (2009). *ME-MS41 Aqua Regia ICP AES ICPMS Multielement Method.* Retrieved 2016-04-27, from `http://www.alsglobal.com/Our-Services/Minerals/Geochemistry`

An, P., & Moon, W. M. (1993). An evidential reasoning structure for integrating geophysical, geological and remote sensing data. In *Proceedings of igarss '93 - ieee international geoscience and remote sensing symposium* (pp. 1359–1361). IEEE. doi: 10.1109/IGARSS.1993.322084

An, P., Moon, W. M., & Bonham-Carter, G. F. (1994). Uncertainty management in integration of exploration data using the belief function. *Nonrenewable Resources*, *3*(1), 60–71. doi: 10.1007/BF02261716

Arne, D. C., & Bluemel, E. B. (2011). *Catchment Analysis and Interpretation of Stream Sediment Data from QUEST South, British Columbia* (Report 2011-5). Vancouver BC: Geoscience BC. Retrieved from `http://www.geosciencebc.com/s/2011-05.asp`

Babak, O., & Deutsch, C. V. (2009a). Collocated Cokriging Based on Merged Secondary Attributes. *Mathematical Geosciences*, *41*(8), 921–926. doi: 10.1007/s11004-008-9192-2

Babak, O., & Deutsch, C. V. (2009b). An intrinsic model of coregionalization that solves variance inflation in collocated cokriging. *Computers & Geosciences*, *35*(3), 603–614. doi: 10.1016/j.cageo.2008.02.025

Barnett, R. M., & Deutsch, C. V. (2015). *Linear Rotations : Options for Decorrelation and Analysis* (CCG Annual Report 17). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Bonham-Carter, G. F., Agterberg, F. P., & Wright, D. F. (1989). Weights of evidence modelling: a new approach to mapping mineral potential. In F. P. Agterberg & B.-C. C. F (Eds.), *Statistical applications in the erath sciences* (pp. 171–183). Geological Survey of Canada. doi: 10.4095/128125

Bonham-Carter, G. F., & Goodfellow, W. (1986, mar). Background corrections to stream geochemical data using digitized drainage and geological maps: application to selwyn basin,

yukon and northwest territories. *Journal of Geochemical Exploration*, *25*(1-2), 139–155. doi: 10.1016/0375-6742(86)90011-7

Brown, W. M., Gedeon, T. D., Groves, D. I., & Barnes, R. G. (2000). Artificial neural networks: A new method for mineral prospectivity mapping. *Australian Journal of Earth Sciences*, *47*(4), 757–770. doi: 10.1046/j.1440-0952.2000.00807.x

Canada Centre for Mapping and Earth Observation. (2014). *Canadian Digital Elevation Data.* Ottawa, ON: Natural Resources Canada. Retrieved from `http://ftp2.cits.rncan.gc.ca/pub/geobase/official/cded`

Carranza, E. J. M. (2004). Weights of Evidence Modeling of Mineral Potential: A Case Study Using Small Number of Prospects, Abra, Philippines. *Natural Resources Research*, *13*(3), 173–187. doi: 10.1023/B:NARR.0000046919.87758.f5

Carranza, E. J. M. (2008). *Geochemical anomaly and mineral prospectivity mapping in GIS*. Saint Louis, MO: Elsevier. Retrieved from `www.sciencedirect.com/science/handbooks/18742734/11`

Carranza, E. J. M. (2009). Controls on mineral deposit occurrence inferred from analysis of their spatial pattern and spatial association with geological features. *Ore Geology Reviews*, *35*(3-4), 383–400. doi: 10.1016/j.oregeorev.2009.01.001

Carranza, E. J. M. (2010). Catchment basin modelling of stream sediment anomalies revisited: incorporation of EDA and fractal analysis. *Geochemistry: Exploration, Environment, Analysis*, *10*(4), 365–381. doi: 10.1144/1467-7873/09-224

Carranza, E. J. M., & Sadeghi, M. (2010). Predictive mapping of prospectivity and quantitative estimation of undiscovered VMS deposits in Skellefte district (Sweden). *Ore Geology Reviews*, *38*(3), 219–241. doi: 10.1016/j.oregeorev.2010.02.003

Carranza, E. J. M., van Ruitenbeek, F. J. A., Hecker, C., van der Meijde, M., & van der Meer, F. D. (2008). Knowledge-guided data-driven evidential belief modeling of mineral prospectivity in Cabo de Gata, SE Spain. *International Journal of Applied Earth Observation and Geoinformation*, *10*(3), 374–387. doi: 10.1016/j.jag.2008.02.008

Carranza, E. J. M., Woldai, T., & Chikambwe, E. M. (2005). Application of Data-Driven Evidential Belief Functions to Prospectivity Mapping for Aquamarine-Bearing Pegmatites, Lundazi District, Zambia. *Natural Resources Research*, *14*(1), 47–63. doi: 10.1007/s11053-005-4678-9

Centre for Computational Geostatistics. (2016a). *Centre for Computational Geostatisics Software.* Edmonton AB: Centre for Computational Geostatistics. Retrieved from `http://www.ccgalberta.com`

Centre for Computational Geostatistics. (2016b). *pygeostat.* Edmonton AB: Centre for Computational Geostatistics. Retrieved from `http://www.ccgalberta.com`

Cheng, Q., & Agterberg, F. P. (1999). Fuzzy Weights of Evidence Method and Its Application in Mineral Potential Mapping. *Natural Resources Research*, *8*(1), 27–35. doi: 10.1023/A:1021677510649

Chung, C.-J. F. (1978). *Computer program for the logistic model to estimate the probability of occurrence of*

discrete events (Paper 78-11). Ottawa, ON: Geological Survey of Canada. doi: 10.4095/103392

Chung, C.-J. F., & Agterberg, F. P. (1980). Regression models for estimating mineral resources from geological map data. *Journal of the International Association for Mathematical Geology*, *12*(5), 473–488. doi: 10.1007/BF01028881

Chung, C.-J. F., & Fabbri, A. G. (1993). The representation of geoscience information for data integration. *Nonrenewable Resources*, *2*(2), 122–139. doi: 10.1007/BF02272809

Continuum Analytics. (2015). *Anaconda Software Distribution.* Retrieved from `https://continuum.io`

Cuba, M., Babak, O., & Leuangthong, O. (2009). *On the Selection of Secondary Variables for Cokriging and Cosimulation* (CCG Annual Report 11). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Debba, P., Carranza, E. J. M., Stein, A., & van der Meer, F. D. (2008). Deriving Optimal Exploration Target Zones on Mineral Prospectivity Maps. *Mathematical Geosciences*, *41*(4), 421–446. doi: 10.1007/s11004-008-9181-5

de Quadros, T. F. P., Koppe, J. C., Strieder, A. J., & Costa, J. F. C. L. (2006). Mineral-Potential Mapping: A Comparison of Weights-of-Evidence and Fuzzy Methods. *Natural Resources Research*, *15*(1), 49–65. doi: 10.1007/s11053-006-9010-9

ESRI. (2012). *ArcGIS Desktop 10.1.* Redlands CA: Environmental Systems Research Institute.

Geomatics Yukon. (2014). *Yukon geographic data.* Whitehorse YK. Retrieved from `http://www.geomaticsyukon.ca`

Geophysical Data Centre. (2014). *Canadian Aeromagnetic Data Base.* Ottawa, ON: Natural Resources Canada. Retrieved from `http://gdr.agg.nrcan.gc.ca/gdrdap/dap/search-eng.php`

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York, NY: Oxford University Press.

Groves, D. I. (2008). Conceptual mineral exploration. *Australian Journal of Earth Sciences*, *55*(1), 1–2. doi: 10.1080/08120090701673310

Groves, D. I., Goldfarb, R. J., Knox-Robinson, C. M., Ojala, J., Gardoll, S., Yun, G. Y., & Holyland, P. (2010). Late-kinematic timing of orogenic gold deposits and significance for computer-based exploration techniques with emphasis on the Yilgarn Block, Western Australia. *Ore Geology Reviews*, *17*(1-2), 1–38. doi: 10.1016/S0169-1368(00)00002-0

Guo, H., & Deutsch, C. V. (2002). *Choosing an Adequate Number of Conditioning Data for Kriging* (CCG Annual Report 4). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Harris, D., & Pan, G. (1999). Mineral Favorability Mapping: A Comparison of Artificial Neural Networks, Logistic Regression, and Discriminant Analysis. *Natural Resources Research*, *8*(2), 93–109. doi: 10.1023/A:1021886501912

Harris, D., Zurcher, L., Stanley, M., Marlow, J., & Pan, G. (2003). A Comparative Analysis of

Favorability Mappings by Weights of Evidence, Probabilistic Neural Networks, Discriminant Analysis, and Logistic Regression. *Natural Resources Research*, *12*(4), 241–255. doi: 10.1023/B:NARR.0000007804.27450.e8

Héon, D. (2003). *Yukon Regional Geochemical Database 2003 - Stream Sediment Analyses.* Whitehorse, YK: Yukon Geological Survey. Retrieved from `http://www.geomaticsyukon.ca`

Hronsky, J. M. A., & Groves, D. I. (2008). Science of targeting: definition, strategies, targeting and performance measurement. *Australian Journal of Earth Sciences*, *55*(1), 3–12. doi: 10.1080/08120090701581356

Jackson, S. (2013). *Optimized Pit Removal.* Austin TX: Center for Research in Water Resources. Retrieved from `http://tools.crwr.utexas.edu/`

Jewbali, A. (2009). *Finding the Nearest Positive Definite Matrix for Input to Semi- automatic Variogram Fitting (varfit _ lmc)* (CCG Annual Report 11). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Jones, R. (2002). Algorithms for using a DEM for mapping catchment areas of stream sediment samples. *Computers & Geosciences*, *28*(9), 1051–1060. doi: 10.1016/S0098-3004(02)00022-5

Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics*. New York, NY: Acedemic Press Inc.

Knox-Robinson, C. M. (2000). Vectorial fuzzy logic: A novel technique for enhanced mineral prospectivity mapping, with reference to the orogenic gold mineralisation potential of the Kalgoorlie Terrane, Western Australia. *Australian Journal of Earth Sciences*, *47*(5), 929–941. doi: 10.1046/j.1440-0952.2000.00816.x

Knox-Robinson, C. M., & Wyborn, L. A. I. (1997). Towards a holistic exploration strategy: Using Geographic Information Systems as a tool to enhance exploration. *Australian Journal of Earth Sciences*, *44*(4), 453–463. doi: 10.1080/08120099708728326

Kramar, U. (1995, dec). Application of limited fuzzy clusters to anomaly recognition in complex geological environments. *Journal of Geochemical Exploration*, *55*(1-3), 81–92. doi: 10.1016/0375-6742(95)00028-3

Kreuzer, O. P., Etheridge, M. A., Guj, P., McMahon, M. E., & Holden, D. J. (2008). Linking Mineral Deposit Models to Quantitative Risk Analysis and Decision-Making in Exploration. *Economic Geology*, *103*(4), 829–850. doi: 10.2113/gsecongeo.103.4.829

Larrondo, P. F., Neufeld, C. T., & Deutsch, C. V. (2003). *VARFIT: A Program for Semi-Automatic Variogram Modelling* (CCG Annual Report 5 No. Lmc). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Leach, D. L., Bradley, D. C., Huston, D., Pisarevsky, S. A., Taylor, R. D., & Gardoll, S. J. (2010). Sediment-Hosted Lead-Zinc Deposits in Earth History. *Economic Geology*, *105*(3), 593–625. doi: 10.2113/gsecongeo.105.3.593

Leach, D. L., Sangster, D. F., Kelley, K. D., Large, R. R., Garven, G., Allen, C. R., … Walters, S. (2005). Sediment-hosted lead-zinc deposits; a global perspective. In J. W. Hedenquist, J. F. H. Thomp-

son, R. J. Goldfarb, & J. P. Richards (Eds.), *Economic geology; one hundredth anniversary volume, 1905-2005* (pp. 561–607). Ottawa ON: Society of Economic Geologists.

Lisitsin, V. A., Porwal, A. K., & Mccuaig, T. C. (2014). Probabilistic Fuzzy Logic Modeling: Quantifying Uncertainty of Mineral Prospectivity Models Using Monte Carlo Simulations. *Mathematical Geosciences*, *46*(6), 747–769. doi: 10.1007/s11004-014-9534-1

Luster, G. R. (1985). *Raw Materials for Portland Cement: Applications of Conditional Simulation of Coregionalization* (PhD Thesis). Stanford University, Stanford, CA.

M. Moon, W., & An, P. (1991, may). Integration of mineral exploration data using fuzzy set theory. In *Canadian jounral of exploration* (Vol. 27, pp. 1–11). doi: 10.3997/2214-4609.201410970

Mackie, R. A., Arne, D. C., & Brown, O. (2015). *Enhanced interpretation of regional geochemical stream sediment data from Yukon: catchment basin analysis and weighted sums modeling* (Open File 2015-10). Whitehorse YK: Yukon Geological Survey. Retrieved from `http://www.geology.gov.yk.ca/`

Manchuk, J. G., & Deutsch, C. V. (2015). *Latest SGSIM Program* (CCG Annual Report 17). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Mccuaig, T. C., Porwal, A. K., Joly, A., & Ford, A. (2013). Managing Uncertainty in Exploration Targeting. In *Et 2013 masters course* (p. 37). Perth: Center for Exploration Targeting (CET). Retrieved from `www.cet.edu.au/research-outcomes/presentations/managing-uncertainty-in-exploration-targeting`

Moon, W. M. (1990). Integration Of Geophysical And Geological Data Using Evidential Belief Function. *IEEE Transactions on Geoscience and Remote Sensing*, *28*(4), 711–720. doi: 10.1109/TGRS.1990.572988

Natural Resources Canada. (2014a). *GeoGratis.* Retrieved from `http://ftp2.cits.rncan.gc.ca/pub/geobase/official/cded`

Natural Resources Canada. (2014b). *Geoscience Data Repository for Geophysical Data.* Retrieved from `http://gdr.agg.nrcan.gc.ca/gdrdap/dap/search-eng.php`

Nykänen, V. (2008). Radial Basis Functional Link Nets Used as a Prospectivity Mapping Tool for Orogenic Gold Deposits Within the Central Lapland Greenstone Belt, Northern Fennoscandian Shield. *Natural Resources Research*, *17*(1), 29–48. doi: 10.1007/s11053-008-9062-0

Oh, H.-J., & Lee, S. (2008). Regional Probabilistic and Statistical Mineral Potential Mapping of Gold–Silver Deposits Using GIS in the Gangreung Area, Korea. *Resource Geology*, *58*(2), 171–187. doi: 10.1111/j.1751-3928.2008.00050.x

Oh, H.-J., & Lee, S. (2010). Application of Artificial Neural Network for Gold–Silver Deposits Potential Mapping: A Case Study of Korea. *Natural Resources Research*, *19*(2), 103–124. doi: 10.1007/s11053-010-9112-2

Porwal, A. K., Carranza, E. J. M., & Hale, M. (2003). Artificial Neural Networks for Mineral-Potential Mapping: A Case Study from Aravalli Province, Western India. *Natural Resources Research*,

12(3), 155–171. doi: 10.1023/A:1025171803637

Porwal, A. K., Carranza, E. J. M., & Hale, M. (2006). A Hybrid Fuzzy Weights-of-Evidence Model for Mineral Potential Mapping. *Natural Resources Research*, *15*(1), 1–14. doi: 10.1007/s11053-006-9012-7

Porwal, A. K., González-Álvarez, I., Markwitz, V., Mccuaig, T. C., & Mamuse, A. (2010). Weights-of-evidence and logistic regression modeling of magmatic nickel sulfide prospectivity in the Yilgarn Craton, Western Australia. *Ore Geology Reviews*, *38*(3), 184–196. doi: 10.1016/j.oregeorev.2010.04.002

Porwal, A. K., & Kreuzer, O. P. (2010). Introduction to the Special Issue: Mineral prospectivity analysis and quantitative resource estimation. *Ore Geology Reviews*, *38*(3), 121–127. doi: 10.1016/j.oregeorev.2010.06.002

Raines, G. L. (1999). Evaluation of Weights of Evidence to Predict Epithermal-Gold Deposits in the Great Basin of the Western United States. *Natural Resources Research*, *8*(4), 257–276. doi: 10.1023/A:1021602316101

Rantitsch, G. (2000, oct). Application of fuzzy clusters to quantify lithological background concentrations in stream-sediment geochemistry. *Journal of Geochemical Exploration*, *71*(1), 73–82. doi: 10.1016/S0375-6742(00)00143-6

Rose, A. W., Dahlberg, E. C., & Keith, M. L. (1970, apr). A multiple regression technique for adjusting background values in stream sediment geochemistry. *Economic Geology*, *65*(2), 156–165. doi: 10.2113/gsecongeo.65.2.156

Rossi, M. E., & Deutsch, C. V. (2014). *Mineral Resource Estimation* (CCG Annual Report 16). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com` doi: 10.1007/978-1-4020-5717-5

Schodde, R. C. (2014a). The Global Shift to Undercover Exploration - How fast? How effective? In *Society of economic geologists (seg) 2014 annual conference* (p. 47). Keystone CO: MinEx Consulting. Retrieved from `www.minexconsulting.com/publications/sep2014b.html`

Schodde, R. C. (2014b). Uncovering exploration trends and the future: Where's exploration going? In *International mining and resources conference (imarc)* (p. 49). Melbourne: MinEx Consulting. Retrieved from `www.minexconsulting.com/publications/sep2014.html`

Sibbick, S. J. (1994). *Preliminary Report on the Application of Catchment Basin Analysis to Regional Geochemical Survey Data, Northern Vancouver Island (NTS 92L/03, 04, 05 and 06)* (Paper 1994-1). Vancouver BC: Ministry of Energy Mines and Petroleum Resources. Retrieved from `http://www.empr.gov.bc.ca/Mining/Geoscience/Geochemistry/CatchmentBasins/Pages/CatchmentExample.aspx`

Singer, D. A., & Kouda, R. (1996). Application of a feedforward neural network in the search for Kuroko deposits in the Hokuroku district, Japan. *Mathematical Geology*, *28*(8), 1017–1023. doi: 10.1007/BF02068587

Skabar, A. A. (2005). Mapping Mineralization Probabilities using Multilayer Perceptrons. *Natural Resources Research*, *14*(2), 109–123. doi: 10.1007/s11053-005-6955-z

Tangestani, M. H., & Moore, F. (2002). The use of Dempster–Shafer model and GIS in integration of geoscientific data for porphyry copper potential mapping, north of Shahr-e-Babak, Iran. *International Journal of Applied Earth Observation and Geoinformation*, *4*(1), 65–74. doi: 10.1016/S0303-2434(02)00008-9

Verly, G. W. (1993). Sequential Gaussian Cosimulation: A Simulation Method Integrating Several Types of Information. In A. Soares (Ed.), *Geostatistics troia '92,* (pp. 543–554). Dordrecht NL: Kluwer Academic Publisher. doi: 10.1007/978-94-011-1739-5_42

Wang, T., & Deutsch, C. V. (2009). *Application of Block LU Simulation with an Approximate Model of Coregionalization* (Vol. 2009; CCG Annual Report 11). Edmonton AB: University of Alberta. Retrieved from `http://www.ccgalberta.com`

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. doi: 10.1080/01621459.1963.10500845

Wright, D. F., & Bonham-Carter, G. F. (1996). VHMS favourability mapping with GIS-based integration models, Chisel Lake-Anderson Lake area. In G. F. Bonham-Carter, A. G. Galley, & G. E. M. Hall (Eds.), *Extech i: a multidisciplinary approach to massive sulphide research in the rusty lake-snow lake greenstone belts, manitoba* (Bulletin 4 ed., p. 402). Ottawa, ON: Natural Resources Canada. Retrieved from `http://geogratis.gc.ca`

Xu, S., Cui, Z., Yang, X., & Wang, G. (1992). A preliminary application of weights of evidence in gold exploration in Xiong-er Mountain Region, He-Nan Province. *Mathematical Geology*, *24*(6), 663–674. doi: 10.1007/BF00894232

Xu, W., Tran, T., Srivastava, R., & Journel, A. G. (1992). Integrating Seismic Data in Reservoir Modeling: The Collocated Cokriging Alternative. In *Spe annual technical conference and exhibition.* Society of Petroleum Engineers. doi: 10.2118/24742-MS

Yukon Geological Survey. (2016a). *MINFILE.* Whitehorse YK. Retrieved from `http://www.geology.gov.yk.ca`

Yukon Geological Survey. (2016b). *Yukon Digital Bedrock Geology.* Whitehorse YK. Retrieved from `http://www.geomaticsyukon.ca`