Improved Resource Estimates with Multiple Data Types

by

Jinpyo Kim

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

Multiple data types should be used simultaneously to improve resource estimation models. The multivariate relationship between the data is required. One common approach involves using decorrelation transformation techniques to simplify complex relationships, but this method relies on having collocated data. With heterotopic data, these techniques cannot be applied.

A Data Error Model (DEM) is developed as a solution to the challenge of using multiple data types that are not sampled at the same location. This model quantifies relationships between different types of data, even if they are not collocated. The workflow of DEM involves pairing analysis to understand the relationships between variables at different locations. The parameters of the DEM account for errors and biases in different data types. The DEM describes the relationships that emerge in pairing analysis with primary and secondary data types. Applying the DEM to simulated primary data produces collocated secondary data distributions. This allows us to obtain the relationships between two variables.

The thesis proposes a method to improve the accuracy of resource estimation models by facilitating the use of multiple data types. The DEM is used to infer the relationships between different types of data, even if they are not collocated. The relationship inferred from DEM can be expressed as a Gaussian Mixture Model (GMM), which underlies the conditional distribution needed to impute collocated primary or error-free values. Imputation allows for the creation of estimation models conditioning to primary variable and secondary variable data. A case study using data from a Nevada gold mine demonstrates the improved estimates.

Acknowledgments

I want to express my heartfelt gratitude to Dr. Clayton Deutsch for his unwavering support, invaluable advice, and helpful guidance throughout my research journey. Working with Clayton was a true privilege, and his professionalism and enthusiasm have been a huge source of inspiration in my life.

I also want to extend my sincere thanks to my dear friends at CCG for sharing countless enjoyable moments and engaging in motivating academic discussions. I am also grateful to the CCG sponsors for their generous financial support, which was instrumental in the success of my study.

Last but not least, I want to thank my parents for their persistent love and trust in me. Their support has been a constant source of strength and motivation throughout my life.

Table of Contents

List of Figures

List of Symbols

Symbol	Description
a	Relative error parameter of DEM
b	Absolute error parameter of DEM
ϵ	Random number in normal distribution
$C(\quad)$	Covariance
c	Relative bias parameter of DEM
d	Absolute bias parameter of DEM
$f(X\|Y)$	Conditional probability density function of X, given that Y
$E[\quad]$	Expected value
$\mathbf{h}$	Spatial lag vector
i	Index for random variable
j	Index for random variable
k	Index for random variable
nst	Number of structure
V	Variance-covariance matrix
$\ln(\quad)$	Natural logarithm
ρ	Correlation coefficient
σ	Standard deviation
$\mathrm{Var}(\quad)$	Variance
$\gamma(\quad)$	(Semi-) Variogram
x	x-axis index for blocks in the block model
y	y-axis index for blocks in the block model
z	z-axis index for blocks in the block model
Z_{mis}	Missing values of variable Z
Z_{obs}	Observed values of variable Z

List of Abbreviations

Abbreviation	Description
3D	Three Dimension
CDF	Cumulative Distribution Function
DEM	Data Error Model
EM	Expectation-Maximization
`GETSECREAL`	Simulation Data Extraction Program
GMM	Gaussian Mixture Model
GSLIB	Geostatistical Software Library
ICM	Intrinsic Correlation Model
IOCK	Intrinsic Model Ordinary Cokriging
KDE	Kernel Density Estimation
LMC	Linear Model of Coregionalization
MAF	Min/Max Auto-correlation Factors
MG	Multivariate Gaussian
MI	Multiple Imputation
MSE	Mean Squared Error
OCK	Ordinary Cokriging
OK	Ordinary Kriging
`OPTDEM`	Optimal DEM Finder Program
PCA	Principal Component Analysis
RMSE	Root-Mean-Square-Error
RV	Random Variable
SK	Simple Kriging
SOCK	Standardized Ordinary Cokriging Estimator

Chapter 1

Introduction

1.1 Spatial Estimation Modeling

The resources industry relies on the collection and analysis of geological data to predict spatial variations. In this context, variables refer to geological properties that exhibit spatial variability and can be quantitatively measured (Houlding, 1994). Primary variables are those that offer the most accurate and valuable information for geological prediction, obtained through careful collection methods, and treated as error-free. On the other hand, secondary variables are quickly and economically sampled, considered to be less accurate and less reliable. These secondary variables serve as auxiliary data to complement the primary variable and enhance the accuracy of geological estimation models.

Kriging, originally proposed by Krige (1951) and developed by Matheron (1963), is a widely utilized method for estimating at unsampled locations based on a minimum error-variance estimation algorithm. There are different types of kriging. Simple kriging (SK) minimizes the error variance without imposing any constraints on the weights. The mean, inferred from available samples, is considered a known constant for the entire domain. Ordinary kriging (OK) implicitly re-estimates the mean as a constant within each search neighborhood. In kriging practice, OK is often preferred over SK because it estimates a more robust local mean rather than relying on the global mean. This can lead to lower mean squared error (MSE) estimates (J. Deutsch & Deutsch, 2012).

Stochastic simulation, employing conditional cumulative distribution functions and Monte Carlo algorithms, provides a numerical and visual representation of spatial uncertainty. Simulation offers an advantage over kriging by allowing the assessment of uncertainty. Conditional simulation was initially developed to address the smoothing effect observed in maps generated by kriging. This smoothing effect reduces spatial variability, resulting in varying degrees of smoothing across different regions and potentially introducing artificial structures. The estimation map generated by kriging is suitable for illustrating global trends, while conditionally simulated maps are more appropriate for studies focused on local variability patterns (C. V. Deutsch & Journel, 1998).

1.2 Problem Motivation

Multiple data types or variables from different locations, including different sampling vintages, different drilling types and drill hole sizes, and relatively cost-effective chip or channel samples, are used as data sources to create spatial geographic estimation models. Typically, the primary variable is associated with drill holes (exploration data) and the secondary variable is associated with blast

holes (production samples). When multiple variables are measured at the same locations, known as homotopic data, kriging can be employed to estimate each variable individually. Alternatively, decorrelation techniques, such as Principal Component Analysis (Davis & Greenes, 1983), Projection Pursuit Multivariate Transform (Barnett, Manchuk, & Deutsch, 2014), Min/Max Autocorrelation Factors (Desbarats & Dimitrakopoulos, 2000), and Stepwise Conditional Transform (Leuangthong & Deutsch, 2003), can facilitate the simulation of homotopic data. These multivariate geostatistical workflows rely on multivariate transformations to capture the relationships within the data.

However, primary and secondary variables are unequally sampled, resulting in heterotopic data. This issue imposes practical limitations on identifying meaningful multivariate relationships between geological variables. Advancing estimation modeling with multiple data types requires an approach to infer relationships between variables at different locations. Understanding the connections between heterotopic data is a crucial process to establish a comprehensive understanding of their relationships. The inference of relationships between heterotopic data allows for the convenient utilization of data imputation techniques, which, in turn, enhances the accuracy of estimation models. Moreover, the ability to infer relationships facilitates simulation processes that were historically challenging due to the constraints of multivariate transformations when dealing with heterotopic data. This thesis proposes a method for establishing the relationship between heterotopic data and seeks to improve the estimation model of multiple data types using multiple imputation with the inferred relationship.

1.3 Thesis Statement and Research Contribution

This thesis introduces a novel Data Error Model (DEM) workflow that addresses the challenges posed by heterotopic data, thereby enhancing the accuracy of estimation models. The DEM framework enables the inference of relationships between heterotopic variables by simulating hypothetical collocated data, even in cases where variables have not been equally sampled. Through the application of multiple imputation (MI), which utilizes the inferred relationships, the missing data of the error-free primary variable can be imputed, leading to improved accuracy in the estimation model.

MI is a well-established probabilistic approach that quantifies the conditional distribution of missing data based on the observed data. However, obtaining reliable relationships can be challenging in heterotopic scenarios where no data is available at the same location. To overcome this challenge, the DEM approach has been developed to infer relationships between primary and secondary variables by identifying errors and biases within the secondary variables. The DEM workflow involves conducting a pairing analysis of primary and secondary variables to compare correlations and mean differences of pairs, enabling the identification of inferred relationships even in the absence of collocated data. The primary data simulated by a DEM that causes the primary data to have errors and biases of the secondary variable will mimic the secondary data distribution. This facilitates

the identification of inferred relationships between the primary variable and the DEM-applied data. With the conditional distribution from them, MI can be performed even in heterotopic situations.

In conclusion, the mining industry generates a wealth of heterotopic data from multiple variables, presenting significant challenges in analysis and interpretation. The DEM approach offers a valuable tool for inferring relationships between primary and secondary variables in heterotopic situations. By incorporating a DEM into the analysis of heterotopic data, more accurate and reliable estimations at unsampled locations can be made.

1.4 Thesis Outline

Chapter 2: Background Concepts

This chapter provides an overview of the fundamental concepts needed in the thesis. The first section summarizes covariance, correlation coefficient, and variogram, highlighting their significance in geostatistics. The second section explores the Linear Model of Coregionalization (LMC), discussing its application in capturing the relationships between variables. Lastly, the third section introduces the Gaussian Mixture Model (GMM), shedding light on its relevance in analyzing complex spatial distributions.

Chapter 3: Prototype of Alternative Techniques

In this chapter, alternative techniques for building estimation models for heterotopic data are explored. The first section delves into cokriging, emphasizing the utilization of covariance between variables. The second section presents the Intrinsic Correlation Model (ICM) as a simplified version of the LMC. It showcases examples of cokriging using LMC and ICM, followed by a comprehensive comparison of their effectiveness in the third section.

Chapter 4: Data Error Model (DEM)

This chapter introduces the DEM framework and its associated workflows. The first section introduces an approach to the basic error model, which serves as the foundation for the Data Error Model (DEM) formula. The second section provides insight into the development of DEM and presents the underlying mathematical model. The third section demonstrates the impact of varying DEM parameters on the resulting DEM-applied data. Additionally, the fourth section describes pairing analysis, a crucial component of the DEM workflow, and summarizes a comprehensive process for obtaining an optimal DEM. The last section explores factors that influence the accuracy of an inferred DEM.

Chapter 5: Multiple and Mixed Data Type Imputation

This chapter focuses on multiple imputation (MI), a methodology that enhances the accuracy of estimation models and enables uncertainty analysis. It presents the theoretical foundations of

imputation and provides a synthetic MI example incorporating DEM workflows.

Chapter 6: Case Study: Application of DEM Workflow and MI for Multiple Data Types at Rain Mine

In this chapter, a case study is conducted to apply MI using DEM to real data from the Rain Mine data provided by Newmont for CCG training purposes. The DEM and MI are applied to actual data, and the results are discussed and analyzed.

Chapter 2

Background Concepts

This chapter reviews geostatistical techniques for integrating multivariate relationships in space, which is the basis of this thesis. Basic geostatistical concepts are discussed in many books such as Geostatistics: Modeling Spatial Uncertainty by Chiles and Delfiner (2012) or Geostatistical Reservoir Modeling by Pyrcz and Deutsch (2014). Geostatistics Lessons (`http://geostatisticslessons.com`) also provide a wealth of geostatistical knowledge with easy-to-understand examples.

2.1 Covariance, Correlation coefficient, and Variogram

Covariance is a statistical measure that quantifies the relationship between two random variables (RVs) (Rice, 2006). It indicates the extent to which changes in one variable are associated with changes in another variable. The covariance between two variables Z and Y can be calculated using the following formula (Park & Park, 2018):

$$C(Z,Y) = E[(Z - E[Z])(Y - E[Y])] = E[ZY] - E[Z]E[Y] \tag{2.1}$$

In the formula, Z and Y represent two variables, and $E[Z]$ and $E[Y]$ represent the means of Z and Y, respectively. The covariance can be positive, negative, or zero, indicating the direction and strength of the relationship between the variables (Bonamente, 2017). A positive covariance suggests that when one variable increases, the other tends to increase as well, while a negative covariance indicates an inverse relationship, where one variable tends to decrease as the other increases. A covariance of zero suggests no linear relationship between the variables. However, covariance alone does not provide a standardized measure of the strength of the relationship, making it challenging to compare relationships between different variable pairs. For this purpose, the correlation coefficient is commonly used (Rodgers & Nicewander, 1988).

The correlation coefficient is a standardized version of the covariance and measures the strength and direction of the linear relationship between two variables (Taylor, 1990). The most widely used correlation coefficient is Pearson's correlation coefficient, denoted as ρ (rho). It is calculated by dividing the covariance between Z and Y by the product of their standard deviations:

$$\rho_{ZY} = \frac{C(Z,Y)}{\sigma_Z \cdot \sigma_Y} \tag{2.2}$$

Here, $C(Z,Y)$ represents the covariance between Z and Y, and σ_Z and σ_Y represent the standard deviations of Z and Y, respectively. The correlation coefficient ρ ranges from -1 to 1. A value of +1 indicates a perfect direct linear relationship, -1 indicates a perfect inverse linear relationship, and 0 indicates no linear relationship between the variables. It is important to note that the correlation

coefficient measures only linear relationships and may not capture other types of relationships, such as nonlinear or causal relationships (Schober, Boer, & Schwarte, 2018).

Both the covariance and correlation coefficient are useful tools for analyzing relationships between variables, but the correlation coefficient provides a standardized measure that is often easier to interpret and comparable across different variable pairs.

Assuming stationarity, the variogram employed in geostatistics is closely linked to the covariance function, as both serve as measures of statistical correlation strength with respect to distance. They quantify the spatial dependence between two data points (Pyrcz & Deutsch, 2014). Stationarity implies that the statistical properties of the data remain constant across the study area. In such cases, the following equations hold true:

$$E\left[Z(\mathbf{u})\right] = E\left[Z(\mathbf{u}+\mathbf{h})\right] = m,\ \mathrm{Var}\left(Z(\mathbf{u})\right) = \mathrm{Var}\left(Z(\mathbf{u}+\mathbf{h})\right) = \sigma^2,\ \forall\, \mathbf{u}, \mathbf{u}+\mathbf{h} \in A$$

where $\mathbf{u}$ is location vector and $\mathbf{h}$ denoted distance or lag vector.

The (semi-)variogram, denoted by $\gamma(\mathbf{h})$, is computed by taking half of the difference in variance between pairs of data points at varying spatial locations. It measures how the variance of the data changes as a function of distance.

$$2\gamma(\mathbf{h}) = \mathrm{Var}\left(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right) = E\left[\left(Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})\right)^2\right] \tag{2.3}$$

The spatial covariance function, denoted by $C(\mathbf{h})$, represents the covariance between two data points at a given spatial lag vector $\mathbf{h}$.

$$C(\mathbf{h}) = \mathrm{cov}\left(Z(\mathbf{u}), Z(\mathbf{u}+\mathbf{h})\right) = E\left[Z(\mathbf{u})Z(\mathbf{u}+\mathbf{h})\right] - E\left[Z(\mathbf{u})\right] E\left[Z(\mathbf{u}+\mathbf{h})\right] \tag{2.4}$$

The covariance $C(\mathbf{h})$ is zero if the data separated by $\mathbf{h}$ are not linearly correlated. At $\mathbf{h} = 0$, the stationary covariance $C(0)$ equals the stationary variance σ^2.

$$\begin{aligned} C(0) &= E\left[Z(\mathbf{u})Z(\mathbf{u}+0)\right] - E\left[Z(\mathbf{u})\right] E\left[Z(\mathbf{u}+0)\right] \\ &= E\left[Z(\mathbf{u})^2\right] - E\left[Z(\mathbf{u})\right]^2 \\ &= \mathrm{Var}\left(Z(\mathbf{u})\right) = \sigma^2 \end{aligned}$$

The relationship between the variogram and covariance is important in variogram interpretation and in providing covariances to kriging equations. This relation depends on the model decision that the mean and variance are constant and independent of location (Rossi & Deutsch, 2014). The

following equations show the relationship between variogram and covariance.

$$\begin{aligned} 2\gamma(\mathbf{h}) &= \mathrm{Var}\left(Z(\mathbf{u}+\mathbf{h}) - Z(\mathbf{u})\right) \\ &= E\left[Z(\mathbf{u}+\mathbf{h})^2\right] + E\left[Z(\mathbf{u})^2\right] - 2E\left[Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u})\right] \\ &= \mathrm{Var}\left(Z(\mathbf{u}+\mathbf{h})\right) + E\left[Z(\mathbf{u}+\mathbf{h})\right]^2 + \mathrm{Var}\left(Z(\mathbf{u})\right) + E\left[Z(\mathbf{u})\right]^2 \\ &\quad - 2\,\mathrm{cov}\left(Z(\mathbf{u}+\mathbf{h}), Z(\mathbf{u})\right) - 2E\left[Z(\mathbf{u})\right] E\left[Z(\mathbf{u}+\mathbf{h})\right] \\ &= 2C(0) - 2C(\mathbf{h}) \\ \Longrightarrow \gamma(\mathbf{h}) &= C(0) - C(\mathbf{h}) \\ &= C(0)\left(1 - \rho(\mathbf{h})\right) \end{aligned} \tag{2.5}$$

The covariance function evaluates data similarity over distance, while the variogram represents dissimilarity, that is, $\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h})$, as shown in Equation 2.5. Positive correlation is indicated when the semi-variogram is less than the variance, no correlation when they are equal, and negative correlation when the semi-variogram exceeds the variance.

The cross-covariance function $C_{ij}(\mathbf{h})$ of a set of N random functions $Z_i(\mathbf{u})$ is defined in the framework of a joint second order stationarity hypothesis (Wackernagel, 2003).

$$E\left[Z_i(\mathbf{u})\right] = m_i \quad \forall\, \mathbf{u} \in A;\ i = 1, ..., N$$

$$C_{ij}(\mathbf{h}) = E\left[(Z_i(\mathbf{u}) - m_i)(Z_j(\mathbf{u}+\mathbf{h}) - m_j)\right] \quad \forall\, \mathbf{u}, \mathbf{u}+\mathbf{h} \in A;\ i, j = 1, ..., N \tag{2.6}$$

The cross-variogram $\gamma_{ij}(\mathbf{h})$ is defined in the context of a joint intrinsic hypothesis for N random functions. With the same conditions as the cross-covariance equation 2.6, the cross-variogram is expressed as:

$$\gamma_{ij}(\mathbf{h}) = \frac{1}{2}E\left[(Z_i(\mathbf{u}) - Z_i(\mathbf{u}+\mathbf{h}))(Z_j(\mathbf{u}) - Z_j(\mathbf{u}+\mathbf{h}))\right] \tag{2.7}$$

The function representing the relationship between cross-variogram and cross-covariance is expressed as follows, assuming that $C_{ij}(\mathbf{h}) = C_{ji}(\mathbf{h})$

$$\gamma_{ij}(\mathbf{h}) = C_{ij}(0) - C_{ij}(\mathbf{h}) \tag{2.8}$$

Both cross-covariance and cross-variogram depend only on the separation vector $\mathbf{h}$, and the covariance function must be positive definite. Both functions are explained in more detail with the following LMC describing the coregionalization of multivariate.

2.2 Linear Model of Coregionalization (LMC)

The Linear Model of Coregionalization (LMC) is the most commonly used method for analyzing and interpreting the spatial continuity of multiple variables (Goulard & Voltz, 1992). The LMC builds each random function $Z_k(\mathbf{u})$ as a linear combination of independent standard factors $Y_i(\mathbf{u})$

with $i = 0$ corresponding to no spatial structure (C. V. Deutsch, 2021; Journel & Huijbregts, 1976).

$$Z_k(\mathbf{u}) = m_k + \sum_{i=0}^{nst} a_{k,i} Y_i(\mathbf{u}) \quad k = 1, ..., K \tag{2.9}$$

with

- $E\left[Z_k(\mathbf{u})\right] = m_k$
- $E\left[Y_i(\mathbf{u})\right] = 0 \quad \forall\, i$
- $C\left(Y_i(\mathbf{u}), Y_{i'}(\mathbf{u}+\mathbf{h})\right) = c_i(\mathbf{h})$ if $i = i'$, otherwise 0.

The cross-covariance between any two RVs $Z_k(\mathbf{u})$ and $Z_{k'}(\mathbf{u}+\mathbf{h})$ can be expressed as a linear combination of cross-covariances between any two RVs $Y_i(\mathbf{u})$ and $Y_{i'}(\mathbf{u}+\mathbf{h})$. Also, the random functions $Y_i(\mathbf{u})$ are mutually independent, and cross-covariance models of them can be defined LMC as follows(Goovaerts, 1997):

$$\begin{aligned} C_{kk'}(\mathbf{h}) &= \operatorname{cov}\left(Z_k(\mathbf{u}), Z_{k'}(\mathbf{u}+\mathbf{h})\right) \\ &= \sum_{i=0}^{nst} a_{k,i} a_{k',i} c_i(\mathbf{h}) \\ &= \sum_{i=0}^{nst} b_{k,k'} c_i(\mathbf{h}) \quad \forall\, i, k, k' \end{aligned} \tag{2.10}$$

where the sill $b_{k,k'}$ of the basic covariance model $c_i(\mathbf{h})$ is

$$b_{k,k'} = \sum_{i=0}^{nst} a_{k,i} a_{k',i} \tag{2.11}$$

and the variance-covariance matrices $B = [b_{k,k'}]$ must satisfy

$$b_{k,k} \times b_{k',k'} \geq b_{k,k'} \times b_{k',k} \quad \forall\, k, k'$$

By construction, the coefficients $b_{k,k'}$ and $b_{k',k}$ are identical, hence the two cross-covariance models $C_{kk'}(\mathbf{h})$ and $C_{k'k}(\mathbf{h})$ are the same. Furthermore, the $(nst+1)$ coregionalization matrices are all positive semi-definite.

Variograms can also be developed from the LMC using the same process as above. The cross-variogram models $\gamma_{kk'}(\mathbf{h})$ defined by LMC is expressed as:

$$\gamma_{kk'}(\mathbf{h}) = \sum_{i=0}^{nst} b_{kk'} \Gamma_i(\mathbf{h}) \quad \forall\, k, k' \tag{2.12}$$

where each function $\Gamma_i(\mathbf{h})$ is a permissible semi-variogram model and has the relationship $c_i(\mathbf{h}) = 1 - \Gamma_i(\mathbf{h})$ when variograms are standardized. The $(nst+1)$ matrices of the coefficients $b_{kk'}$ corresponding to the sill of the model $\Gamma_i(\mathbf{h})$ are all positive semi-definite.

2.3 Gaussian Mixture Model (GMM)

Gaussian Mixture Models (GMMs) and the LMC are both used in multivariable or multiple data type modeling. GMMs offer a powerful framework for modeling complex non-Gaussian multivariate relationships among collocated variables (de Souza et al., 2022). This allows complex dependencies to be quantified, making it a useful tool for multivariate modeling when decorrelation techniques are needed, eliminating the need for an LMC approach. This section highlights how GMMs serve as an alternative for capturing multivariate relationships without relying on the assumptions of the LMC method.

A GMM is a probabilistic density function that can be expressed as a weighted sum of Gaussian component densities. Mathematically, a GMM is represented as (Reynolds et al., 2009):

$$p(\mathbf{x} \mid \Psi) = \sum_{i=1}^{M} w_i g\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{2.13}$$

In this equation, $\mathbf{x}$ represents a D-dimensional continuous-valued data vector (i.e., measurement or features), w_i (for $i = 1, ..., M$) are the mixture weights, and $g\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right)$ denotes the Gaussian components characterized by their mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. The parameter set Ψ comprises all the component parameters, including the weights, means, and covariance matrices for each Gaussian component.

GMMs are not constrained by a single parametric form and are not purely data-driven. They offer a flexible and semi-parametric approach for smoothly modeling complex data patterns. Estimating the parameters of GMMs typically involves using the iterative Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Krishnan, 2007).

The implementation of GMMs requires determining the appropriate number of Gaussian components for fitting the data. It is important to strike a balance between over-fitting and under-fitting to ensure that the GMM accurately captures the distribution of the mixed model. Generally, using a smaller number of components is preferred to avoid over-fitting (Gomes, Boisvert, & Deutsch, 2022).

GMMs are particularly well-suited for modeling complex and high-dimensional geostatistic data sets, as they can effectively represent intricate univariate or multivariate distributions (Sarkar, Melnykov, & Zheng, 2020). The GMM can be used for clustering, density estimation, and data generation (Zhang et al., 2021). In this thesis, GMMs produce reasonable conditional distributions for use in multiple imputation purposes for missing data variables. Using GMMs for imputation not only improves the accuracy of the imputed values but also provides computational advantages over alternative methods such as kernel density estimation (D. S. Silva & Deutsch, 2018). Figure 2.1 explains how the GMM looks for multivariate data. GMM shows clusters by showing the relationship between the two variables in a diagram.

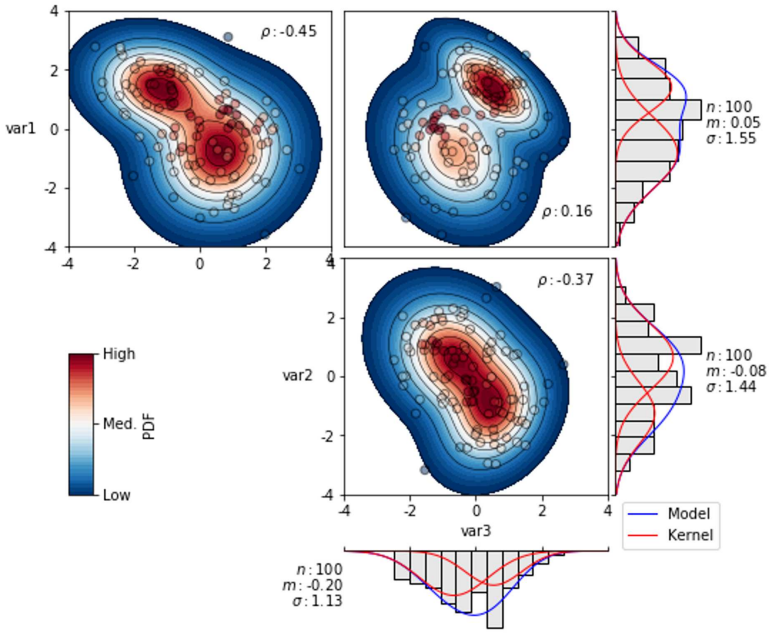


Figure 2.1: Example of a Gaussian mixture model (GMM) representation. It shows a mixture model created for three variables using GMM component value 2. The GMM provides a flexible and smooth model for fitting complex distributions.

Chapter 3

Prototype of Alternative Techniques

This chapter focuses on cokriging, a technique employed to enhance the accuracy of predictive models by incorporating secondary variables that are not located at the same location as the primary variable. The chapter also addresses the challenge associated with conventional cokriging using LMC and introduces the Intrinsic Correlation Model (ICM) as an alternative method. Additionally, a synthetic example is shown comparing the cokriging process using LMC and ICM and the results are discussed.

The utilization of secondary data, such as geological trends, seismic data, and production data, can improve the estimation model for the primary variable. However, the complex nature of the multi-dimensional distributions comprising diverse measurements poses significant challenges to resource estimation. To predict the conditional distribution of uncertainty at unsampled locations, a multivariate distribution between the unsampled location and available sample data within a defined search distance is necessary. Therefore, establishing multivariate relationships is vital in constructing simulation and uncertainty analysis.

Defining these multivariate distributions in a non-parametric manner becomes impractical due to the unique configuration of locations for each unsampled point. As a result, the parametric multivariate Gaussian (MG) distribution is widely adopted (Ortiz & Deutsch, 2022). Geostatistical techniques such as Principal Component Analysis (PCA) (Davis & Greenes, 1983) and Min/Max Autocorrelation Factors (MAF) (Desbarats & Dimitrakopoulos, 2000) are commonly employed for decorrelation transformations, enabling the creation of multivariate Gaussian distributions by transforming collocated data. These methods assume the existence of sufficient collocated data points for each variable to accurately represent their characteristics.

However, in practice, collocated data among multiple data types is not common. Cokriging provides a means to enhance the predictive accuracy of primary data models by incorporating secondary data, even when both data types are not available at the same location.

3.1 Cokriging

Cokriging is a method that utilizes the cross-correlation between a primary variable and a secondary variable to minimize the variance of the estimation error. To simplify the process, all variables are standardized to have a mean of zero and a standard deviation of one (Rossi & Deutsch, 2014). After all computations are completed, the standardized values are reverted to the original units by multiplying by the standard deviation and adding the mean. The Standardized Ordinary Cokrig-

ing Estimator (SOCK) employs standardized variables, ensuring that the mean values of both the primary and secondary variables are equal. SOCK is designed as a complement to ordinary kriging by constraining the sum of all weights to 1, thereby reducing the dependence on the assumption of stationarity (Isaaks & Srivastava, 1989). At a specific location $\mathbf{u}_0$, the SOCK can be expressed by the following equation, assuming the mean values of the two variables, Z and Y, are m_Z and m_Y, respectively (Goovaerts, 1998).

$$\frac{Z_{SOCK}\left(\mathbf{u}_0\right)-m_Z}{\sigma_Z}=\sum_{\alpha=1}^{n_1}\lambda_\alpha\left(\mathbf{u}_0\right)\left[\frac{Z\left(\mathbf{u}_\alpha\right)-m_Z}{\sigma_Z}\right]+\sum_{\beta=1}^{n_2}\lambda'_\beta\left(\mathbf{u}_0\right)\left[\frac{Y\left(\mathbf{u}'_\beta\right)-m_Y}{\sigma_Y}\right] \tag{3.1}$$

The cokriging weights λ_α and λ'_β are constrained as follows.

$$\sum_{\alpha=1}^{n_1}\lambda_\alpha\left(\mathbf{u}_\alpha\right)+\sum_{\beta=1}^{n_2}\lambda'_\beta\left(\mathbf{u}'_\beta\right)=1$$

When analyzing and interpreting multivariate spatial information, it is necessary to model the coregionalization inferred from direct and cross-covariances. The calculation principles for cross-variograms and cross-covariances are as explained in Chapter 2. Let's consider a primary variable $Z(\mathbf{u})$ and a secondary variable $Y(\mathbf{u})$. The direct variogram and cross-variogram can be calculated as follows.

$$2\gamma_Z(\mathbf{h})=E\left[\left(Z(\mathbf{u})-Z(\mathbf{u}+\mathbf{h})\right)^2\right], 2\gamma_Y(\mathbf{h})=E\left[\left(Y(\mathbf{u})-Y(\mathbf{u}+\mathbf{h})\right)^2\right] \tag{3.2}$$

$$2\gamma_{Z,Y}(\mathbf{h})=E\left[\left(Z(\mathbf{u})-Z(\mathbf{u}+\mathbf{h})\right)\left(Y(\mathbf{u})-Y(\mathbf{u}+\mathbf{h})\right)\right] \tag{3.3}$$

In order to calculate the cross-variogram, the variables Z and Y must be located at the same location. However, in practical situations, secondary data are not collocated with the primary data, resulting in unequally sampled data or heterotopic data. Consequently, the cross-variogram cannot be directly computed when multiple data are unequally sampled. In such cases, the cross-variogram can be inferred by leveraging the relationship between covariance and variogram (C. V. Deutsch, 2021). This approach allows for the calculation of cross-covariance without the requirement of collocated data points.

$$C_{Z,Y}(\mathbf{h})=E[Z(\mathbf{u})Y(\mathbf{u})]-E[Z(\mathbf{u})]E[Y(\mathbf{u}+\mathbf{h})] \tag{3.4}$$

$$\gamma_{Z,Y}(\mathbf{h})=C_{Z,Y}(0)-C_{Z,Y}(\mathbf{h}) \tag{3.5}$$

Unfortunately, the collocated correlation, the cross-variogram sill ($C_{Z,Y}(0)$), which represents the cross-covariance of the two variables at lag zero, cannot be directly calculated from unequally sampled data. To estimate $C_{Z,Y}(0)$, an extrapolation method is employed using the experimental cross-covariance. The extrapolation intersects the y-axis, followed by the inclusion of the cross-nugget effect. The nugget effect represents the degree of short-scale variability shared among the variables, and in most cases, the cross-nugget effect is less than or equal to the direct variogram models because of positive definite condition. However, unlike direct variogram models, the cross-variogram's

nugget effect is never used, as the independence of nugget components in random variables means their cross-covariance does not impact the nugget effect of the cross-variogram (Goovaerts, 1997).

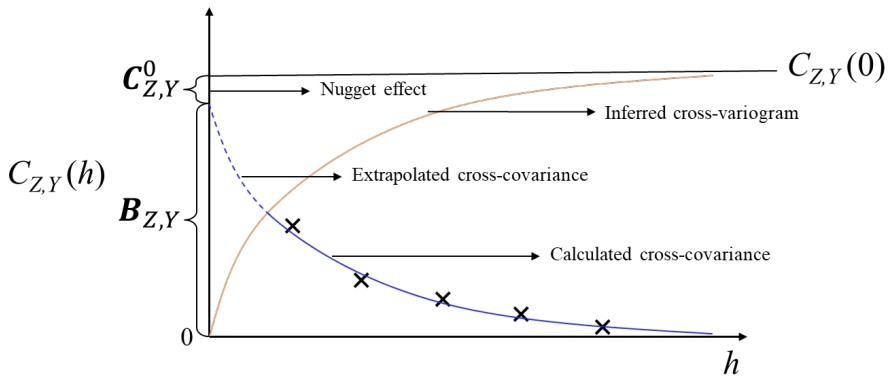


Figure 3.1: Cross-variogram is inferred by extrapolating from the experimental cross-covariance (Wawruch et al., 2002)

3.2 Intrinsic Correlation Model (ICM)

Implementing cokriging through the Linear Model of Coregionalization (LMC) involves calculating the direct and cross-covariances of all variables and fitting them to ensure positive definiteness. However, a major challenge in cokriging lies in the time-consuming and tedious task of fitting all direct and cross-variograms to achieve positive definiteness. Cokriging methods utilizing the LMC are constrained by the requirement that the contribution matrix of the nested structure variogram model must have a positive determinant, as illustrated by Equation 2.12. As the number of variables increases, finding a model that satisfies positive definiteness becomes more challenging.

To address this challenge, the intrinsic model can offer a potential solution. The Intrinsic Correlation Model (ICM) simplifies the process of multivariate covariance modeling by assuming that the spatial covariance function $C(\mathbf{h})$ is the product of a variance-covariance matrix V, representing the relationship between variables, and a spatial correlation function $\rho(\mathbf{h})$ (Wackernagel, 2003).

$$C(\mathbf{h}) = V\rho(\mathbf{h}) \tag{3.6}$$

Note that the spatial correlation function remains the same for all variables. Consequently, all direct and cross-covariance functions can be obtained as scaled versions of the same fundamental spatial correlation function:

$$C_{ij}(\mathbf{h}) = b_{ij}\rho(\mathbf{h}) \quad \forall\, i, j \tag{3.7}$$

Where the coefficient b_{ij} represents the variance when i and j are equal and the covariance when they are different. The concept of intrinsic coregionalization extends this approach by expressing all variograms within the ICM framework as the product of a coregionalization matrix B of coefficients

b_{ij}, satisfying positive definiteness, and a direct variogram $\gamma(\mathbf{h})$ (Chiles & Delfiner, 2012).

$$\Gamma(\mathbf{h}) = B\gamma(\mathbf{h}) \tag{3.8}$$

In the case of two variables, Z and Y, the cross-variogram and the direct variogram under the ICM can be obtained as follows:

$$\begin{aligned} \gamma_{Z,Y}(\mathbf{h}) &= C_{Z,Y}(0)\gamma_{Z,Z}(\mathbf{h}) \\ \gamma_{Y,Y}(\mathbf{h}) &= C_{Y,Y}(0)\gamma_{Z,Z}(\mathbf{h}) \end{aligned} \tag{3.9}$$

In the B matrix, as in the V matrix, the covariance $C_{Z,Y}(0)$ is used to calculate the cross-variogram, and the variance $C_{Y,Y}(0)$ is used to determine the direct-variogram. For heterotopic variables, cross-covariance at the same location, $C_{Z,Y}(0)$, can be obtained by extrapolation as depicted in Figure 3.1.

ICM can be considered as a subset of LMC. However, since ICM ensures proportionality to the main variogram model that satisfies positive definiteness, there is no need for the cumbersome process of model fitting. Additionally, unlike LMC, which requires calculating direct and cross-variograms for all variables, ICM only necessitates the determination of variance, collocated cross-covariance from extrapolation, and a representative variogram. Therefore, cokriging with ICM reduces the time to fit the positive definite requirement of the LMC. This advantage becomes more prominent as the number of variables increases.

3.3 Synthetic Example of Cokriging Using LMC and ICM

The following example demonstrates the practicality and convenience of cokriging using the ICM compared to cokriging based on the LMC. Additionally, ordinary kriging of the primary variable is conducted to assess the impact of incorporating the secondary dataset on estimation accuracy and to compare it with other cokriging results. The example is based on an unconditionally simulated dataset on a regular grid with no outliers.

The reference data consists of a simulated model exhibiting a lognormal distribution with a mean of 1.39 and a standard deviation of 2.62 over a 256m x 256m area with a 1m x 1m resolution. The primary variable, denoted as Z, is sampled at a 20m x 20m square grid spacing within the reference model. These 169 sampled data points represent carefully selected and high-quality data treated as true values. Figures 3.2a and 3.2b depict the sampled locations and the histogram of the primary variable, respectively. On the other hand, the secondary variable, denoted as Y, is sampled from the reference model at a 10m x 10m interval, which is closer than the interval of the primary data. Moreover, the secondary variable is intentionally adjusted for bias and error by modifying its mean and variance. The relative error is generated by multiplying the transformed normal score unit value by the relative error magnitude by a random normal value. The reason for converting to a normal score unit is that the sampling data follows a logarithmic distribution, but the error follows

a Gaussian distribution. Methods for giving errors and biases are provided in detail in later Chapter 4.1. Figure 3.2c and 3.2d display the sampled locations and the histogram of the secondary variable.

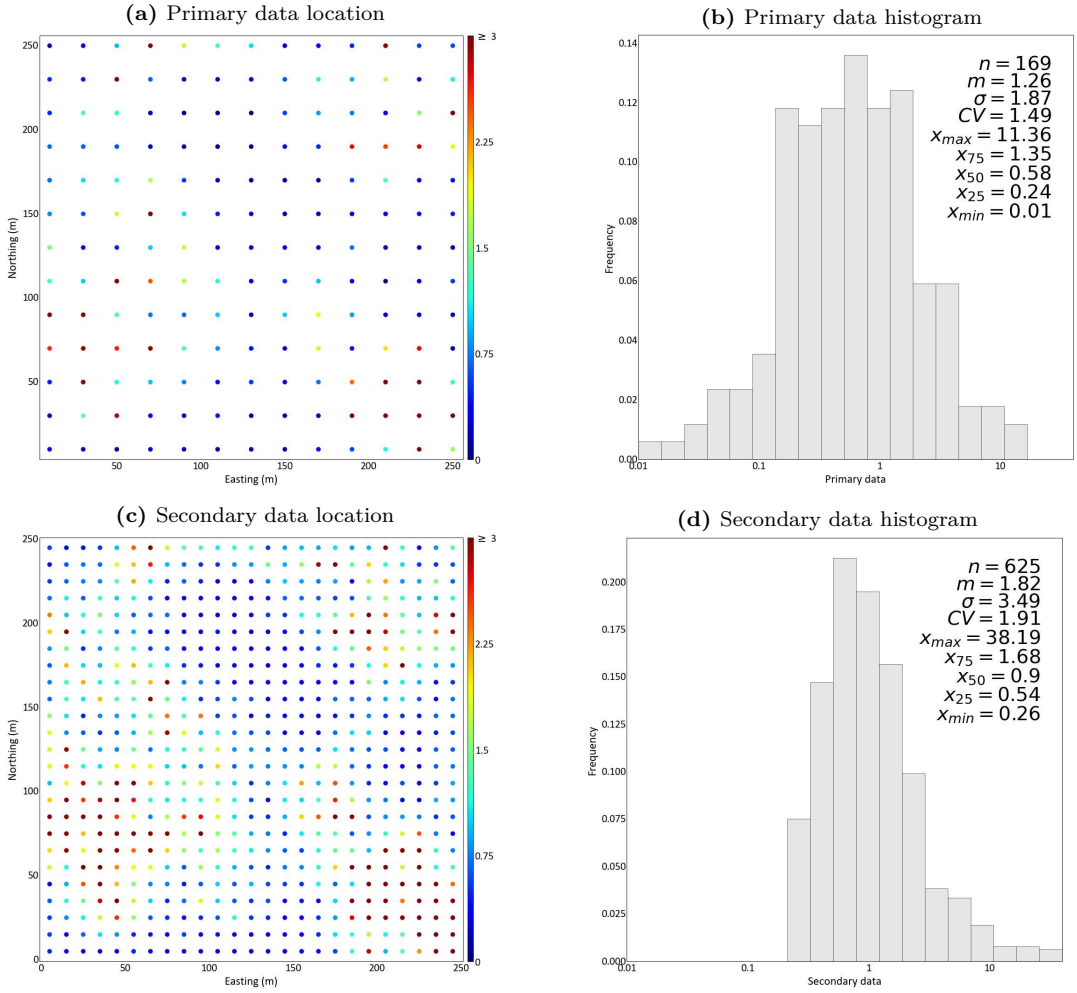


Figure 3.2: Location of primary data spaced 20mX20m (a), and histogram of primary data (b). Secondary data is more densely with spaced 10mX10m (c) and has wider range of value and higher mean (d).

Once the two variables have been established, they undergo standardization before subsequent steps. Direct variogram models for both variables can be obtained through experimental variogram calculations. However, given that the variables Z and Y are not collocated data, this cross-variogram necessitates the calculation of the experimental cross-covariance and subsequent extrapolation to obtain $C_{ZY}(0)$, which represents the sill of the cross-variogram. Figure 3.3 illustrates the extrapolated value of $C_{ZY}(0)$ as 0.73. Subsequently, the cross-variogram is derived by performing $C_{ZY}(0) - C_{ZY}(h)$.

For cokriging, direct variogram models and a cross-variogram model for the two variables are required. The LMC method necessitates iterative fitting to ensure positive definiteness of the covariance matrix and the variogram model. In contrast, ICM simplifies the process by directly multiplying

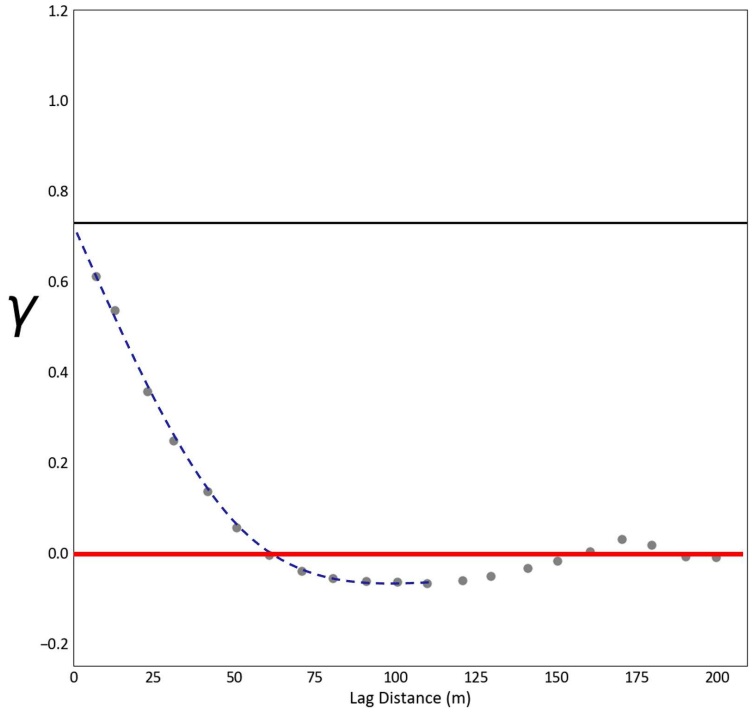


Figure 3.3: The $C_{ZY}(0)$ inferred by extrapolation is 0.73.

the variogram of the primary variable by each sill, and since the value is already standardized, the sill of the direct variogram becomes 1, making it simpler. Therefore, obtaining all cross-variograms becomes straightforward, as long as covariance values at zero lag distance are available. Figure 3.4 shows the direct and cross-variogram models obtained through LMC and ICM for the two variables, Z and Y.

The final step involves creating estimation models using the obtained variogram models. Ordinary kriging (OK) of the primary variable is performed to evaluate whether incorporating the secondary variable through cokriging improves estimation results compared to using only the primary variable. Additionally, cokriging based on LMC and ICM is conducted to compare their respective results. Figure 3.5 displays the results of these three cases, along with the root mean square error (RMSE) values as a comparative measure. The RMSE value between the reference model (true value) and OK is 1.830. Comparing the estimation maps of ordinary cokriging (OCK) using LMC and the intrinsic model ordinary cokriging (ICOK), both models exhibit similar estimation patterns and yield comparable results. The RMSE values compared to the reference model are 1.775 (OCK) and 1.766 (IOCK), with a difference of only 0.09. Also, the correlation coefficient between the two estimation models is very high at 0.986 (Figure 3.6).

Valid LMC requires covariance matrices to be semi-positive definite, necessitating efforts to find direct and cross-variogram models that satisfy this condition for cokriging. The intrinsic model simplifies the cokriging process by assuming that all variograms share the same shape scaled to each sill. Although creating a variogram model for LMC requires more time and effort compared to ICM,

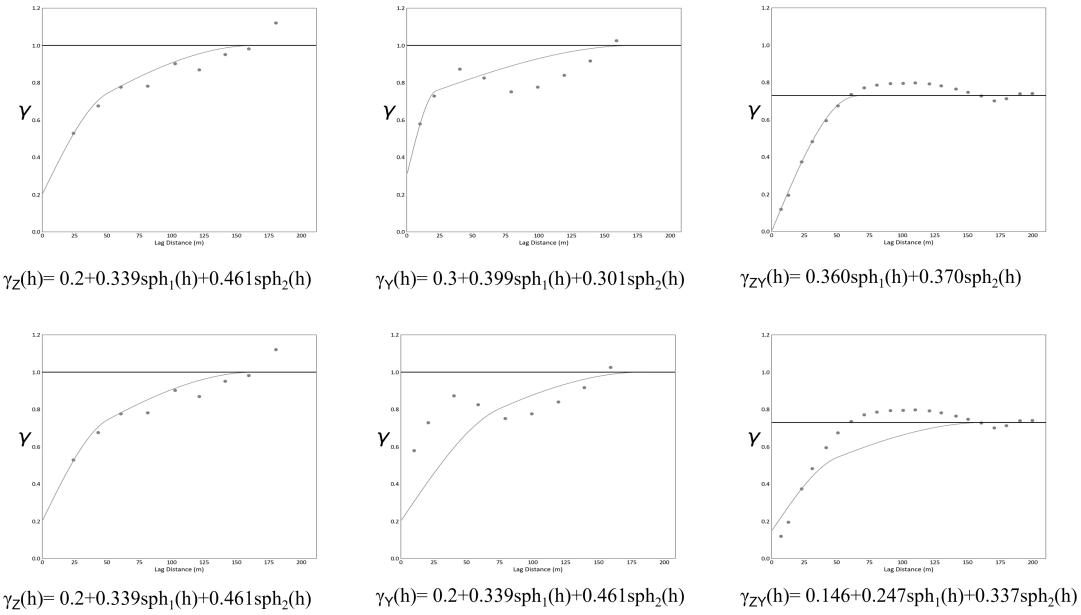


Figure 3.4: The variogram of linear model of coregionalization (top graphs) and the variogram of intrinsic coregionalization model (bottom graphs) about sampling data set.

the estimated results from the two cokriging models demonstrate no significant differences. Hence, ICM can serve as an efficient and reliable alternative to LMC in cokriging involving multiple data types.

It is important to note that in this cokriging example, the results of cokriging with ICM yielded better estimations compared to LMC. However, this does not imply that ICM is universally superior to LMC. The effectiveness of each method depends on the accuracy of the variogram models for the primary and secondary variables, as well as the accuracy of cross-variogram extrapolation at zero lag distance. In general, the accuracy of cokriging with LMC is better because LMC provides a more flexible variogram model that better fits the experimental cross-variograms. ICM is convenient, but less flexible than LMC. Therefore, if the cross-variogram of ICM deviates significantly from the experimental cross-variogram, LMC should be employed, even if it requires more time and effort.

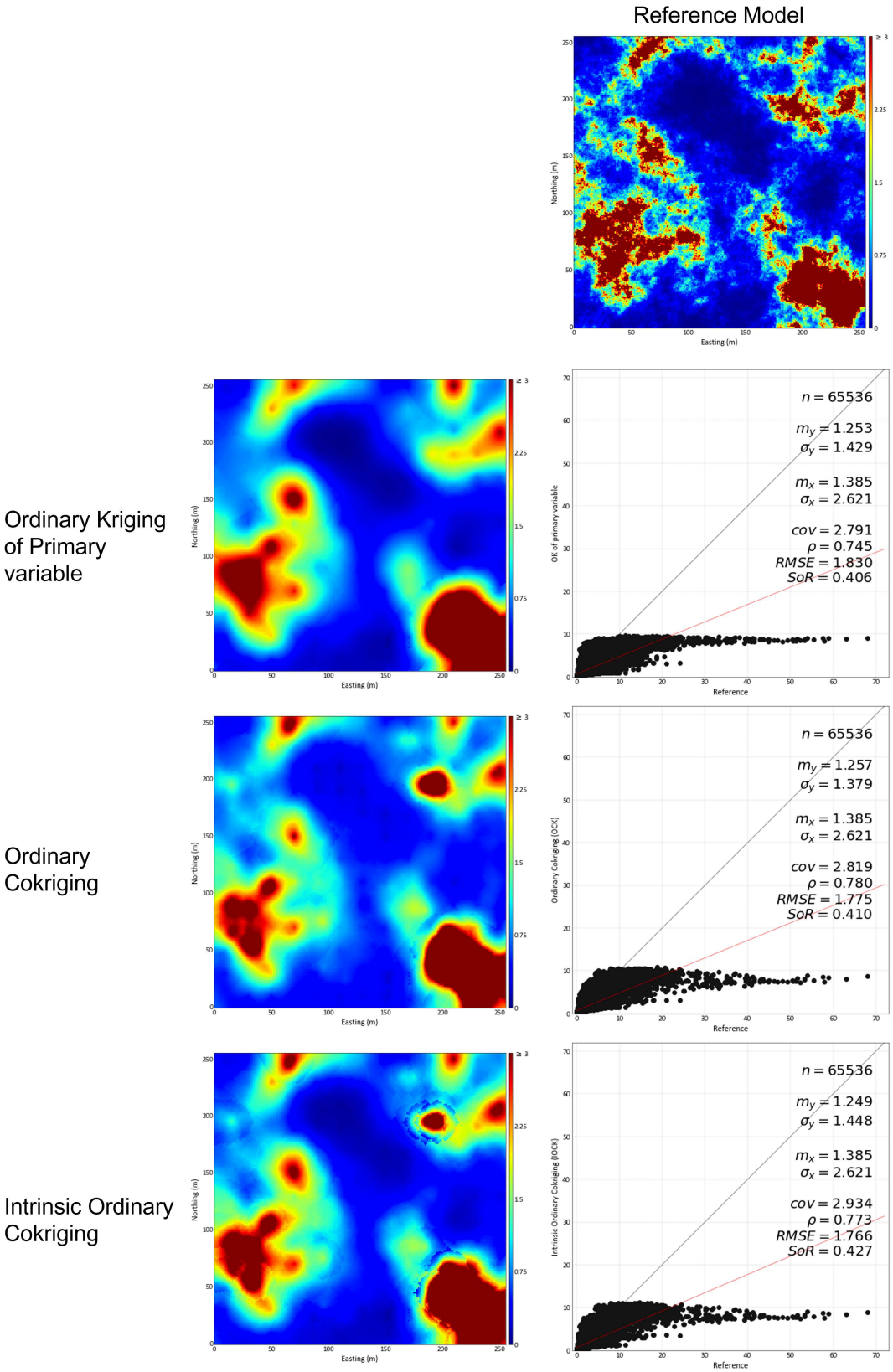


Figure 3.5: Ordinary kriging result with map and comparison with reference model (top), LMC with ordinary cokriging (middle), ICM with ordinary cokriging (bottom).

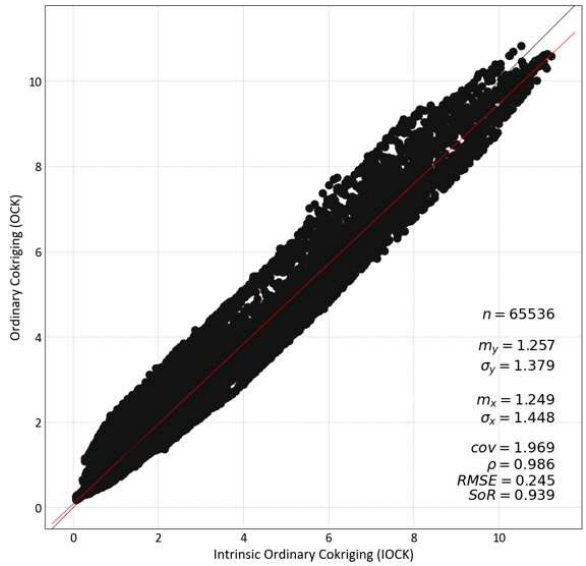


Figure 3.6: The cokriging results of LMC and ICM have high correlation.

Chapter 4

Data Error Model (DEM)

The mining industry generates vast amounts of heterotopic multiple data collected from different sources, vintage, and quality. However, decorrelation transformations aimed at obtaining multivariate relationships are dependent on the availability of collocated data, which inevitably restricts their effectiveness in uncovering relationships within heterotopic data sets.

To address this practical challenge, the Data Error Model (DEM) has been developed as a mathematical model to characterize relationships between primary and secondary variables in the context of heterotopic data. The DEM exhibits flexibility by accommodating errors and biases in the input data. Through an iterative pairing analysis, the DEM effectively tracks the errors and biases associated with a secondary variable, ultimately leading to the identification of suitable relationships.

By employing an optimized DEM, it becomes possible to transform the distribution of primary variables to accurately replicate the distribution of secondary variables and enhance the understanding of complex relationships within spatially heterotopic data. This approach contributes to advancing knowledge in fields that rely on the analysis of multiple data types, such as the mining industry.

4.1 Approach to Error Models

This section is the basis for developing the DEM and explains how to express measured values by dividing them into error or bias and error-free values. In the process of sampling and analyzing measured data, errors are inevitable and cannot be completely eliminated (Gy, 2012). Gaussian-based techniques are well-suited for error modeling, because of efficient calculations in statistical analysis. To apply these techniques effectively, the data must be transformed into normal score units. Consequently, if errors exist in the data, they should also be transformed into normal score units.

The sampled data contains errors in the original units and the data are affected by the sampling error variance (σ_e^2). For a given set of samples generated under consistent conditions and influenced by the same error variance, the measured value at location $\mathbf{u}$ can be expressed as a function of the error magnitude A and the true error-free value $Z_{ef}(\mathbf{u})$.

$$Z(\mathbf{u}) = Z_{ef}(\mathbf{u}) + \varepsilon * A * Z_{ef}(\mathbf{u}) \tag{4.1}$$

Where $Z(\mathbf{u})$ is the observation value with error, $Z_{ef}(\mathbf{u})$ is the true value, ε is a random normal

value following $N(0,1)$, and A is the relative error magnitude.

However, it is not possible to directly convert the relative errors of the original units to normal score units because negative values are physically meaningless. Therefore, in Equation 4.1, a more appropriate representation can be achieved by expressing it in terms of the transformed real value $Y_{ef}(\mathbf{u})$ and the absolute error magnitude B, as demonstrated in Equation 4.2 (Victor M. Silva & Deutsch, 2019).

$$Y(\mathbf{u}) = Y_{ef}(\mathbf{u}) + \varepsilon * B \tag{4.2}$$

Where $Y(\mathbf{u})$ is the observation in normal score units and $Y_{ef}(\mathbf{u})$ is the real value in normal score units.

The DEM is designed to account for both absolute and relative errors, as well as relative and absolute biases. Relative biases (C) and absolute biases (D) can be expressed as:

$$Z(\mathbf{u}) = Z_{ef}(\mathbf{u}) + C * Z_{ef}(\mathbf{u}) \tag{4.3}$$

$$Z(\mathbf{u}) = Z_{ef}(\mathbf{u}) + D \tag{4.4}$$

Where $Z(\mathbf{u})$ is measurement data and $Z_{ef}(\mathbf{u})$ is error-free reference data.

Relative error and absolute error are both experimental errors that can occur in the measurement process. Errors and biases often occur together (Pitard, 2019). Understanding and catching errors and biases can reduce uncertainty in the simulation process.

4.2 Introduction to DEM

The measurement data obtained in various units, such as %metals, gram/ton, ppm, or ppb, often exhibit a distribution that follows either a Poisson or logarithmic distribution (Pitard, 2019). In order to incorporate the errors associated with the sampling data which has a normal distribution, a log transformation of the measured values is necessary.

Given the comprehensive occurrence of errors and biases in sampling, it is crucial to employ a model that can capture both precision and accuracy simultaneously. The DEM provides a framework that accommodates these factors, enabling a more comprehensive analysis of the measurement data. Consider, $Z(\mathbf{u})$ denotes the input variable at location $\mathbf{u}$, and $Y(\mathbf{u})$ represents the DEM output. The DEM is represented by the formula:

$$Y(\mathbf{u}) = e^{\left\{\ln Z(\mathbf{u}) + \varepsilon\left(a \ln Z(\mathbf{u}) + b\right)\right\}} + cZ(\mathbf{u}) + d \tag{4.5}$$

Here, ε represents the standard error that follows a distribution with a mean of 0 and variance of 1. In the DEM, the term 'a' ($a \geq 0$) signifies relative error, 'b' ($b \geq 0$) refers to absolute error, 'c' ($c \in$ Real numbers) presents relative bias, and 'd' ($d \in$ Real numbers) indicates absolute bias. Each parameter can be adjusted to match a data set with specific errors and biases.

4.3 Flexibility of DEM

Figures 4.1 to 4.4 depict the influence of individual parameters on the DEM through scatter plots that compare the input values (x-axis) to the corresponding values obtained after applying the DEM (y-axis). Parameters that are not being adjusted are set to zero to isolate the independent effect of each parameter on the output of the DEM. The input data utilized in these figures follows a log-normal distribution and the output data incorporates error or bias based on the magnitude of the respective parameter. This approach allows for a systematic examination of how variations in parameter values affect the resulting DEM output, providing a clearer understanding of the model's behavior and the impact of each parameter on the transformation process.

The parameters 'a' and 'b' play a role in inputting errors within the DEM. However, 'a' has a more significant effect on the correlation coefficient compared to 'b'. This disparity arises because 'a' can induce substantial errors when the input value is large, even if both 'a' and 'b' are set to the same value. On the other hand, the bias parameters 'c' and 'd' do not have an impact on the correlation coefficient, independently.

The DEM offers flexibility in expressing desired levels of error and bias by adjusting these parameters. Error contributes to the variance of the data distribution, while bias affects the mean. To manipulate the variance and mean of the DEM-applied data distribution, 'a' and 'b' can be used as variance adjustment components, while 'c' and 'd' can be used as mean adjustment components. The relative and absolute error terms, 'a' and 'b', are partially interchangeable. They do capture some slightly different characteristics of error. This interchangeability allows for the fine-tuning of errors to meet specific requirements and achieve the desired data distribution characteristics within the DEM framework.

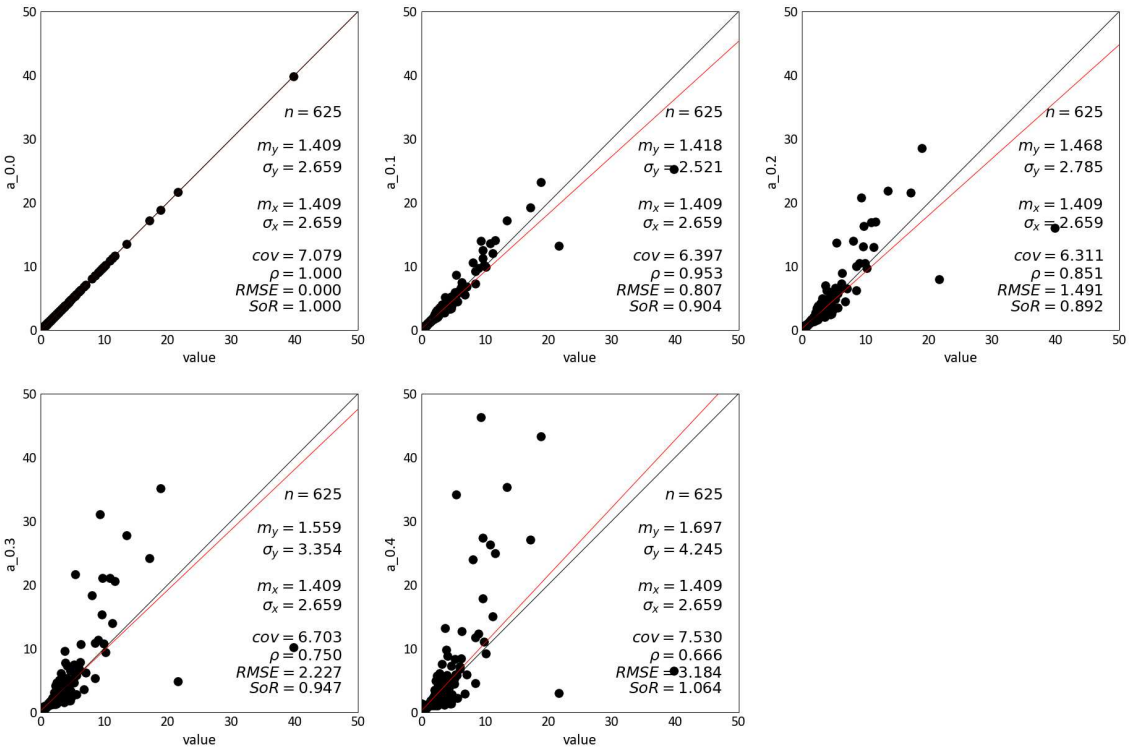


Figure 4.1: 'a' is a relative error parameter. It gives larger errors for larger input values. Also, the higher the 'a' value, the larger the mean of the output values and the smaller the correlation between the input and output values.

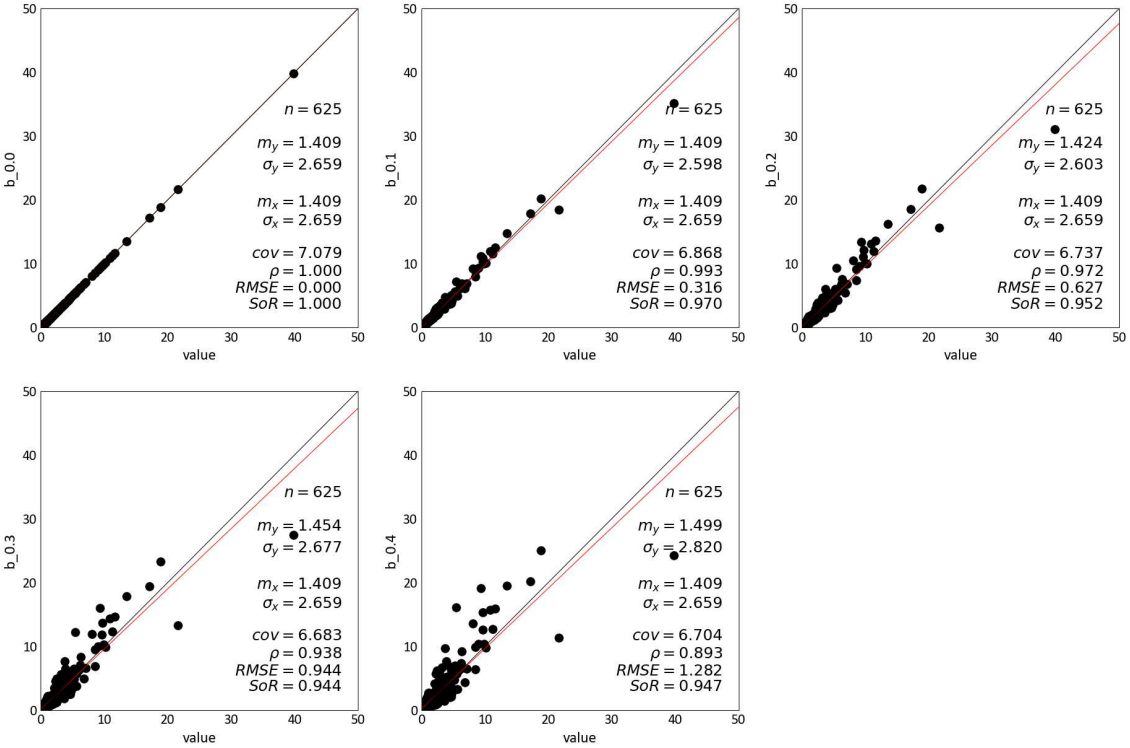


Figure 4.2: 'b' is the absolute error parameter. Adds a fixed error independent of the input value. 'b' has a weaker effect than 'a' on the mean and correlation.

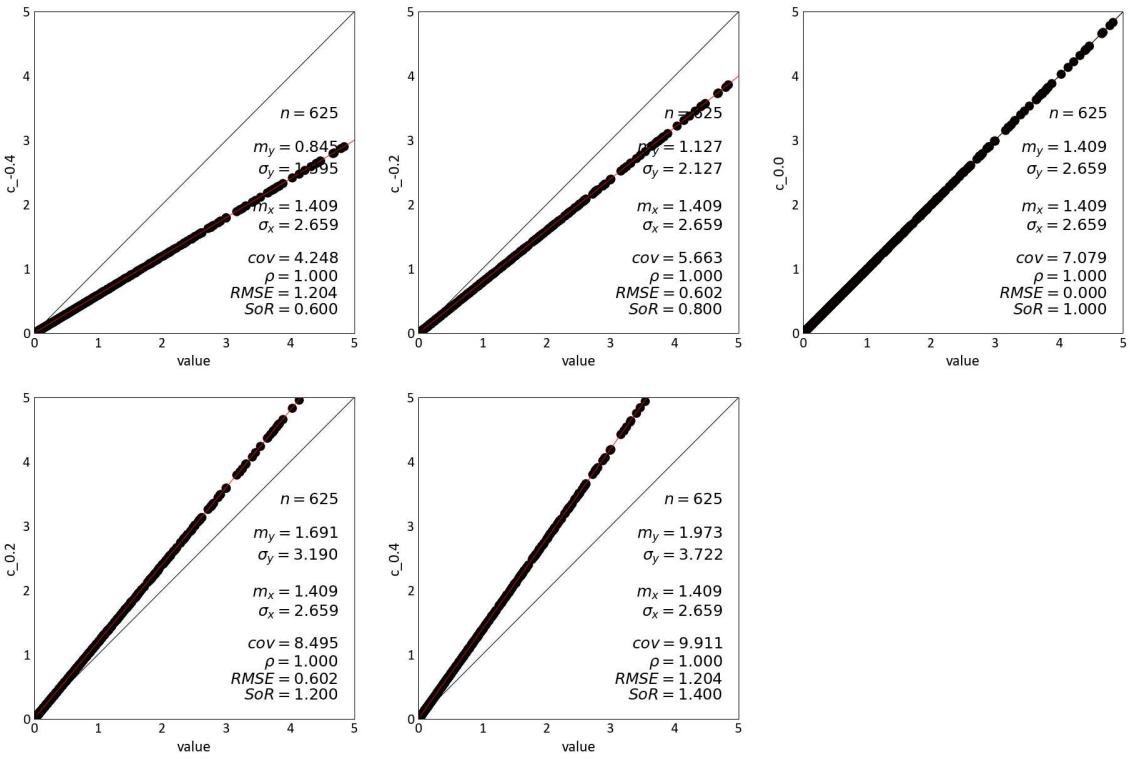


Figure 4.3: 'c' is the relative bias parameter. It is biased to a 'c' multiple of the input value. Negative output values due to a negative 'c' are considered zero.

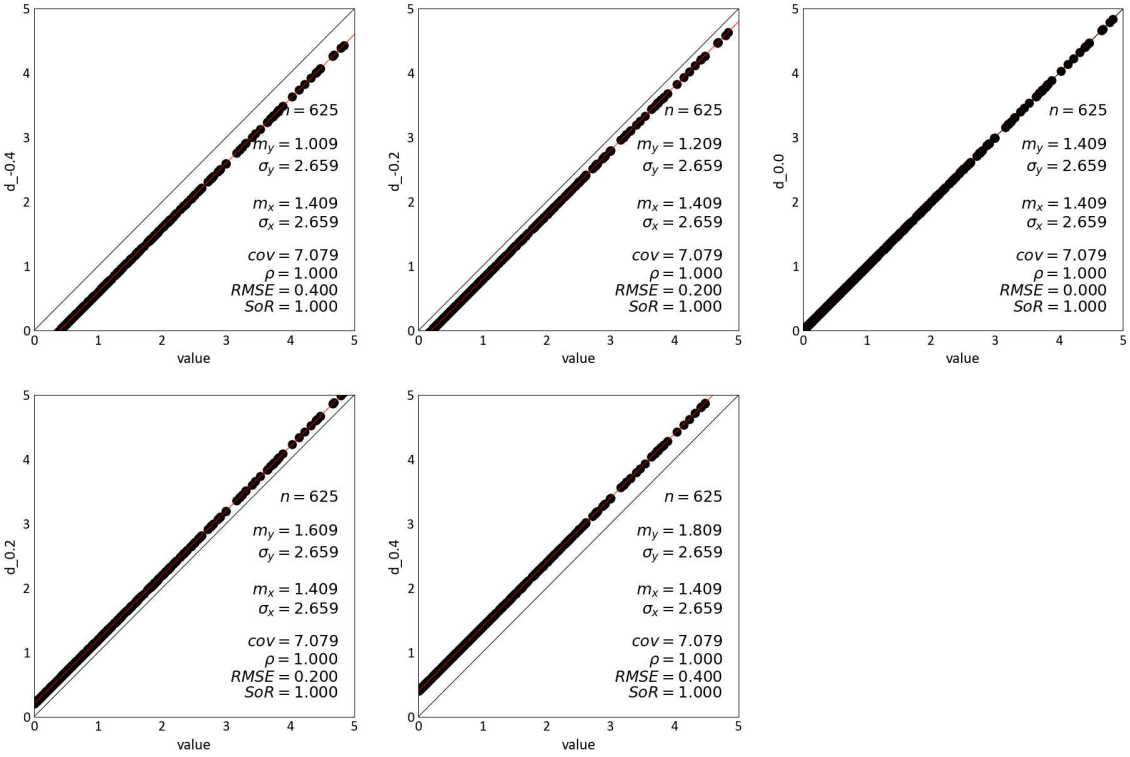


Figure 4.4: 'd' is the absolute bias parameter. All input values are biased with the same 'd' value. Negative output values due to a negative 'd' are considered zero.

4.4 Workflow to Get a Suitable DEM

To obtain a DEM, it is essential to have a primary reference data and the ability to generate reliable simulated primary data at the same locations as the secondary variable by simulation. The simulated primary data is utilized as input data for the DEM. The data applied to the DEM represent inferred secondary data at the same location as the actual secondary data and may contain errors and biases through the DEM.

DEMs make the identification and quantification of errors and biases present in the secondary variable through iterative pairing analysis. Pairing analysis assesses the pairwise correlation and mean difference between the two variables. Pairing analysis enables the detection of spatial relationships between pairs of variables within a defined search radius, even when dealing with heterotopic data.

During the experimental pairing analysis, the inferred secondary data is paired with the primary data, and the results are compared to the true pairing analysis between the true secondary data and primary data. The DEM parameters are adjusted iteratively to ensure that the adjusted experimental pairing analysis by DEM update aligns with the correlation and mean difference observed in the true pairing analysis between the true primary and secondary variables. The suitability of the DEM is evaluated by comparing the pairing analysis results using the updated DEM with the true pairing analysis results. This iterative process continues until the updated analysis fits well with the true pairing analysis, indicating a robust and accurate estimation of the error and bias of the secondary variable. The following list presents steps to find the optimal DEM.

1. Get simulated primary data at secondary data locations and set DEM parameters
2. Do pairing analysis (1) with the primary data and secondary data
3. Apply the simulated primary data to the DEM
4. Do paring analysis (2) with DEM-applied data and the primary data
5. Assess the pairing analysis results (1) and (2) and update DEM parameters to fit (1)
6. Iterate steps 3∼5 until (2) fits well with (1)

If the parameters of DEM change, it affects the result of the pairing analysis. The next figures explain how the DEM-applied data changes in pairing analysis when the DEM parameters are changed. Assuming that there are two variables Z and Y with the same value at the same location ($Z = Y$), Figure 4.5 shows the correlation and mean difference of pairs according to search radius when pairing analysis is performed with these two variables. Naturally, the correlation between pairs of values decreases as the search radius increases, which means that the error increases. A mean difference fixed at 0 means that there is no bias in the means between the pairs. Figure 4.6 shows

the result of pairing analysis between variable Z and DEM-applied data when DEM is applied to variable Y. In order to find out what effect the DEM parameters have, pairing analysis is performed by setting the same search radius to 15m and adjusting each parameter. 'a' and 'b' give errors to the input value, affect the correlation in pairing analysis and show a more sensitive response to 'a' than 'b'. Also, increasing the error parameters increases the mean difference between pairs. 'c' and 'd' do not affect the correlation but shift the mean difference graph in parallel depending on the degree of bias of the DEM-applied data. Figure 4.7 shows how to find a suitable DEM through pairing analysis. The pairing analysis result between DEM-applied data and primary data should have the same correlation and mean difference of pairs that came from true pairing analysis between the true primary and secondary data.

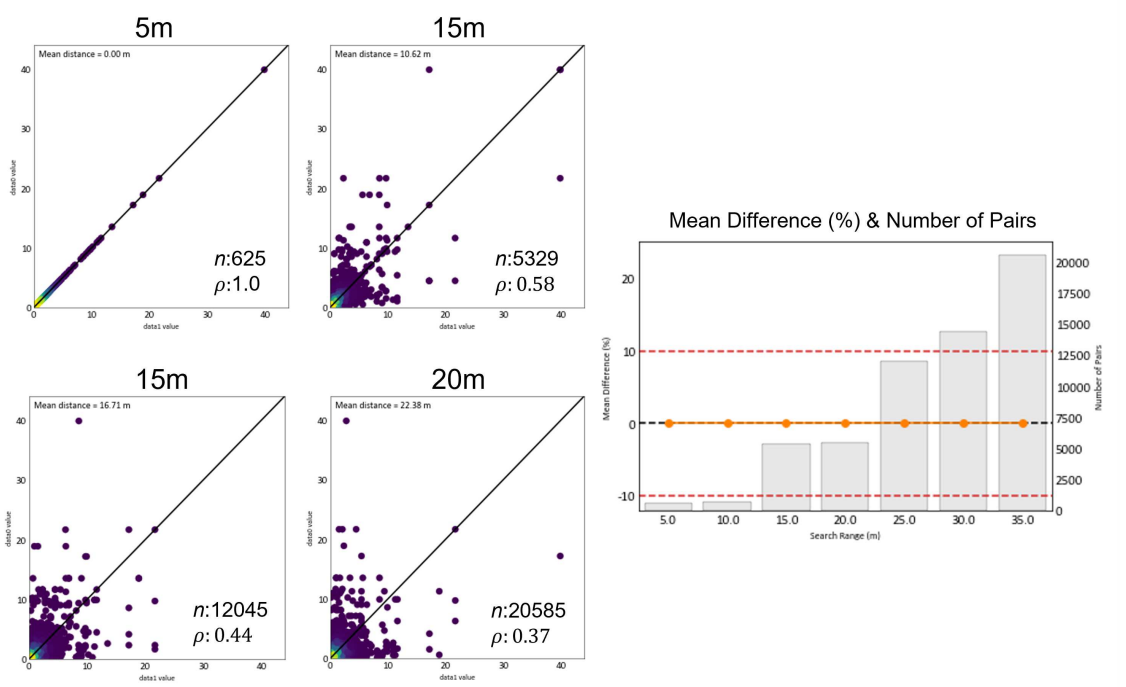


Figure 4.5: When $Z = Y$, scatter plots and correlation coefficient (right), mean difference (left) according to pairing distance.

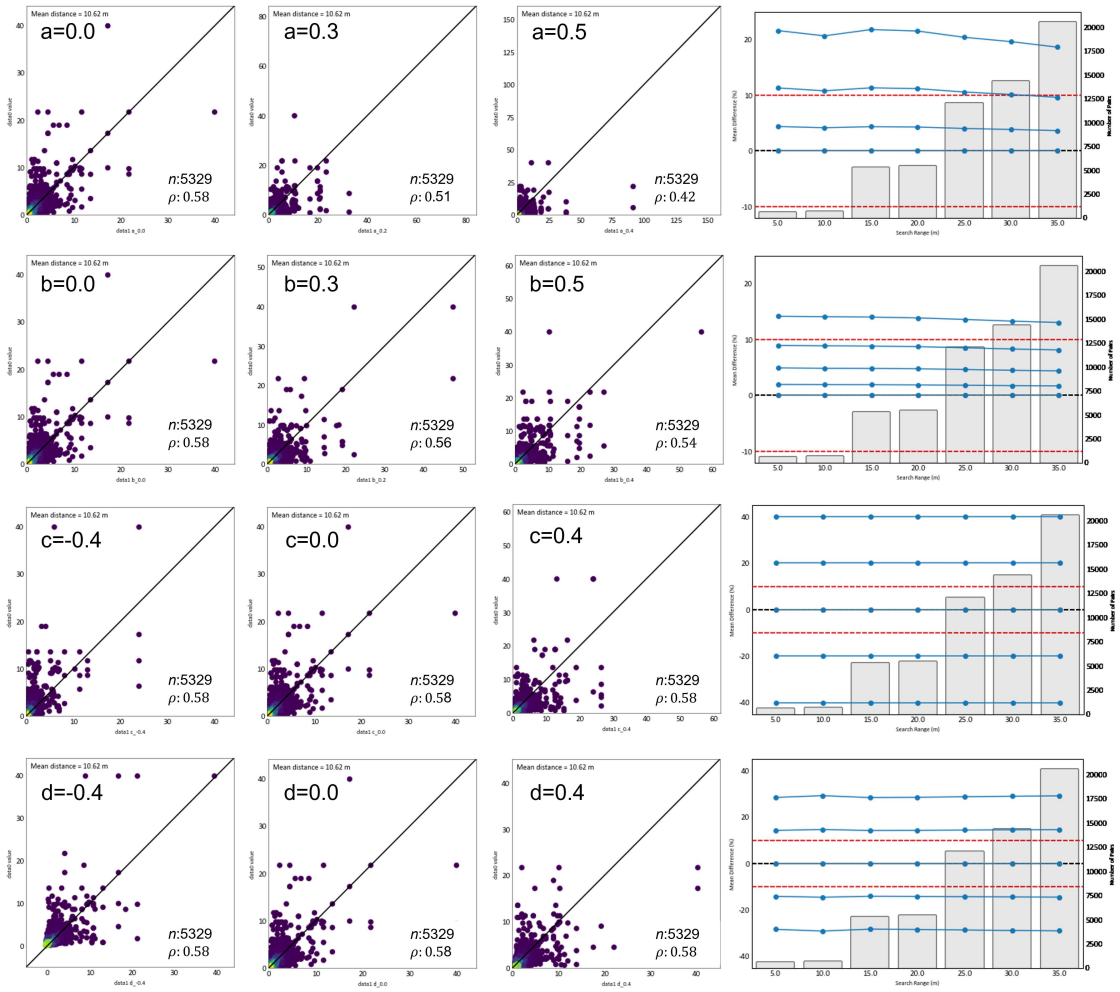


Figure 4.6: Scatter plots and mean difference graphs according to the values of DEM's parameters when the two variables are exactly collocated.

An optimal DEM reproduces the error and bias of the secondary data. This equation can reveal the relationship between the variables by allowing the primary variable to estimate the expected distribution of the secondary data. Through simulation applying the primary data to the DEM, the expected distribution of the secondary data can be obtained if the primary and secondary variables are at the same location. The relationship between two variables can be expressed through a Gaussian mixture model (GMM).

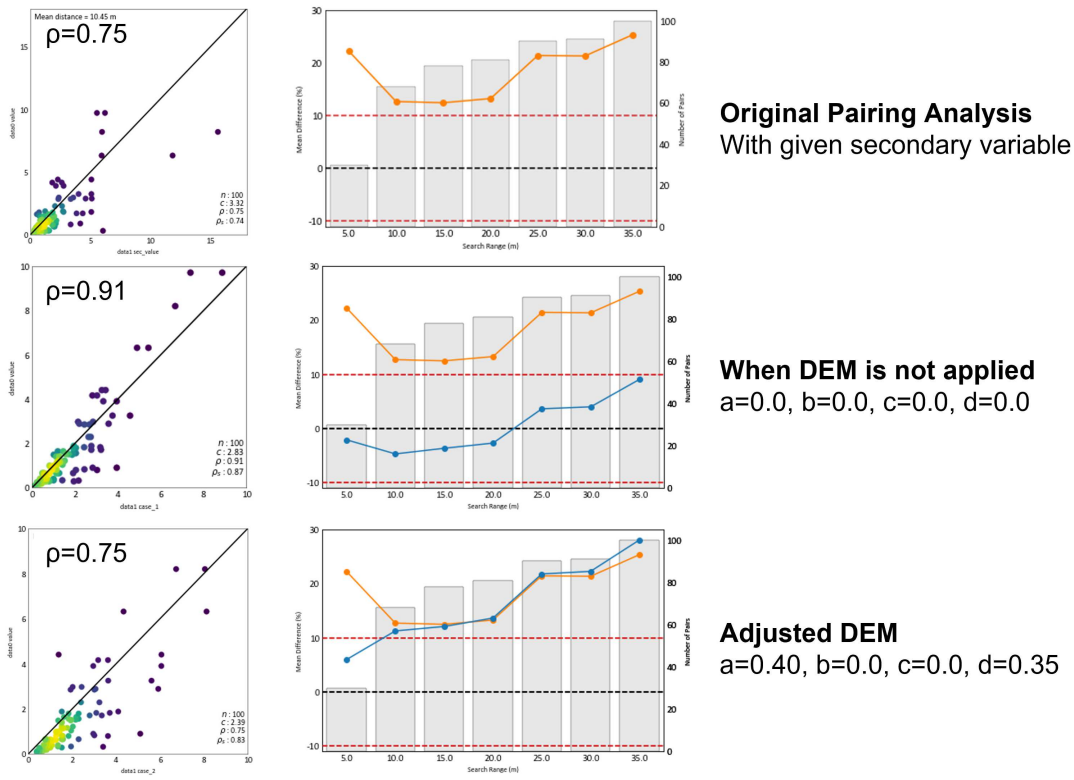


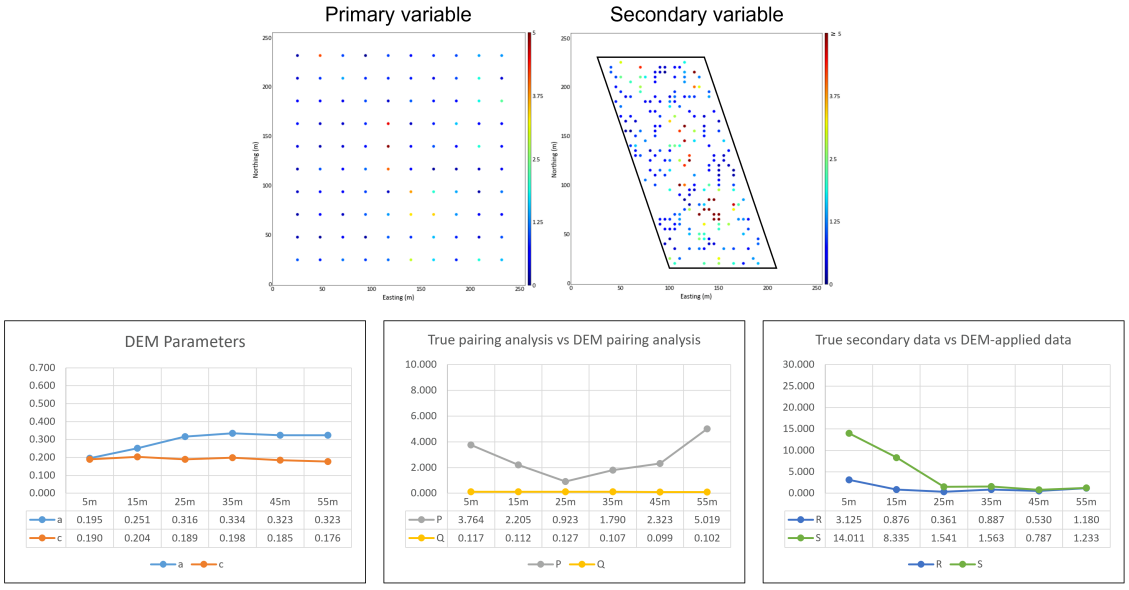
Figure 4.7: A suitable DEM is obtained by iteratively adjusting the DEM parameters until the result of the pairing analysis is close to the mean difference and the correlation coefficient of the true pairing analysis.

4.5 Factors Affecting DEM Accuracy

The accuracy of a DEM can be influenced by various factors, and one critical factor is the search radius used in pairing analysis. The search radius defines the spatial extent within which primary and secondary data are paired. Selecting an appropriate search radius is essential, as it depends on the spatial characteristics of the data, such as the variogram and distance between data of different variables. If the search radius is too small, the DEM may yield noisy and unstable results, failing to represent the entire modeling area adequately. Conversely, a search radius that is too large may lead to the combination of uncorrelated data situated far apart, introducing inaccuracies in the DEM.

To illustrate the impact of the search radius on the DEM, Figure 4.8 presents a practical example. The study employs 100 error-free primary data and 210 secondary data, with no overlapping locations between two variables. The secondary variable is intentionally manipulated to include errors and biases based on a reference DEM where $a = 0.25$ and $c = 0.25$. The graph showing the result is the average value obtained through 3 random samplings of the secondary variable and 50 simulations for each sampling. Through many implementations, the effect of data sampling location and randomness of error is reduced. The results of all experiments in this section are performed in the same way and represent average values. The DEM is obtained for each search radius in increments of 5m.

In this experiment to examine the effect of the search radius, comparing the true pairing analysis results with the DEM-applied pairing analysis results, the smallest difference is shown at the search radius of 25m. The difference in correlation showing a U-shaped parabola explains the importance of setting the search radius that is neither too close nor too far. The difference between the true secondary variable distribution and the DEM-applied data distribution means that the DEM shows an acceptable imitation mean and variance at 25 m or more. However, in the pairing analysis, as the search radius exceeds 25 m, the difference in pairwise correlation increases, which means that the uncertainty of the DEM increases. This observation highlights the importance of carefully selecting an appropriate search radius to achieve the most accurate DEM.



P(%)= correlation coefficient difference between true pairing analysis and experimental pairing analysis

Q(%)= difference of mean difference value between true pairing analysis and experimental pairing analysis

R(%)= mean difference between true secondary variable distribution and DEM-applied data distribution

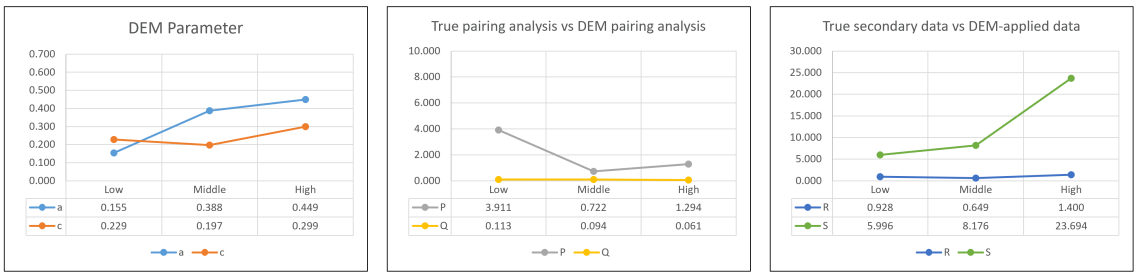
S(%)= standard deviation difference between true secondary variable distribution and DEM-applied data distribution

Figure 4.8: The location map shows the primary data location and the region of interest for the secondary data sampling used in the experiment. The first graph shows the parameters of the DEM obtained for each search radius, the second graph shows the difference between the pairing analysis of the DEM-applied data and the pairing analysis of the actual data, and the third graph shows the difference between the distribution of the DEM-applied data and the actual secondary data distribution. The DEM that shows the smallest difference in pairing analysis and at the same time well described the distribution of the actual secondary variable is obtained at a search radius of 25m.

Another factor that can have an impact on the accuracy of the DEM is the variance of secondary data. When there is higher variance in the secondary data, it can be challenging to identify the model's representative error parameter and the distribution may deviate further from that of the primary data, ultimately reducing the accuracy of the DEM. This is true even when the DEM is applied to collocated simulated primary data. Unfortunately, adjusting the variance of secondary data may not always be possible in real-world scenarios, but it's important to be aware that higher variance can lead to less accurate outcomes when using DEM.

In the provided example, 210 secondary data are randomly sampled from the region of interest, as shown in the location map of Figure 4.8. Three data sets are created with different variances based on three different reference DEMs. The first reference DEM had parameters $a = 0.05$ and $c = 0.25$, the second reference DEM had parameters $a = 0.35$ and $c = 0.25$, and the third reference DEM had parameters $a = 0.55$ and $c = 0.25$. As the 'a' parameter of the primary DEM increased, so did the variance of the secondary data. The pairing analysis is conducted using a fixed search radius of 25 meters.

Figure 4.9 illustrates the impact of the variance of the secondary data on DEM accuracy. Although the DEM follows the results of the pairing analysis better when the variance of the secondary variable is High case than when it is Low case, the graph representing the difference in data distribution shows that the ability of the DEM in describing the target distribution diminishes as the variance of the secondary variable increases.



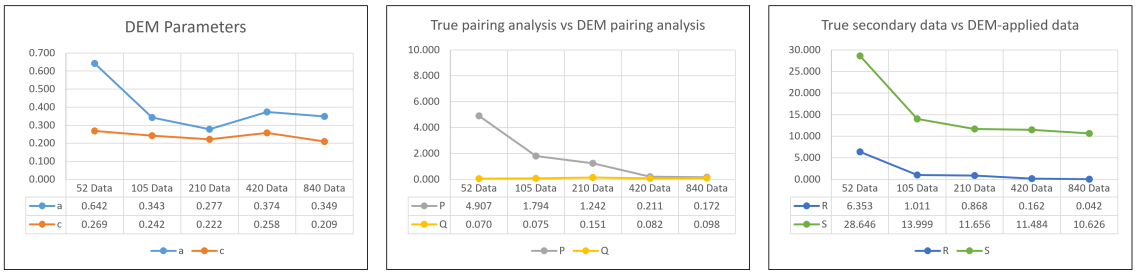
P(%)= correlation coefficient difference between true pairing analysis and experimental pairing analysis
Q(%)= difference of mean difference value between true pairing analysis and experimental pairing analysis
R(%)= mean difference between true secondary variable distribution and DEM-applied data distribution
S(%)= standard deviation difference between true secondary variable distribution and DEM-applied data distribution

Figure 4.9: In the (right) graph comparing the distribution of true secondary data and the distribution of DEM-applied data, it can be seen that the standard deviation difference between the two data sets increases as the variance of the secondary variable increases.

Lastly, the number of secondary data can also affect the accuracy of the DEM. Insufficient data may lead to an unstable DEM. On the other hand, a larger number of data can lead to a more accurate DEM that better predicts the relationship between the variables. Therefore, it is important to carefully consider the number of secondary data used in the analysis and ensure that they are sufficient to obtain reliable results.

To investigate the impact of the number of secondary data on DEM accuracy, 840 secondary data are sampled without overlapping with the primary variable, and errors are given using a reference DEM with parameters $a = 0.25$ and $c = 0.25$. From this 840 sampled data set, 4 additional data sets with 52, 105, 210, and 420 data are generated through random sampling.

Figure 4.10 demonstrates that the more data, the closer the DEM-applied data converges to the distribution of the true secondary data and the pairing analysis results of the true variables. This represents that a larger number of secondary data can enhance the accuracy of the DEM and better describe the relationship between the primary and secondary variables.



P(%)= correlation coefficient difference between true pairing analysis and experimental pairing analysis

Q(%)= difference of mean difference value between true pairing analysis and experimental pairing analysis

R(%)= mean difference between true secondary variable distribution and DEM-applied data distribution

S(%)= standard deviation difference between true secondary variable distribution and DEM-applied data distribution

Figure 4.10: The larger the number of data, the more accurate the DEM. The difference in correlation coefficient between the pairing with the true variable and the pairing of the DEM-applied data and the primary data decrease as the number of secondary data increase (middle graph). This phenomenon is also shown in a graph showing how well the DEM describes the secondary variable (right graph).

In summary, if the search radius is either too close or too far, if the variance of the secondary data is high, or if the number of secondary data points is insufficient, the DEM may fail to effectively capture errors and biases in the secondary data.

To obtain an accurate DEM, it is important to perform the process with different search radii and carefully examine the results to determine the optimal one. Additionally, the accuracy of the DEM can be enhanced by taking steps such as removing outliers. Furthermore, increasing the overlapping area where the primary and secondary data coexist can contribute to more accurate DEM results as the secondary and primary data can make more pairs.

The parameters of the DEM obtained through simulation may not precisely align with the reference DEM provided in the original secondary data. Errors are applied indiscriminately and lack uniqueness. Moreover, when DEM parameters interact, they can influence and obscure each other's effects. For instance, the 'c' parameter in the DEM introduces a relative bias, but when combined with an error parameter, it does not result in a linear bias. This interplay of parameters has significant implications when determining the final DEM parameters.

Chapter 5

Multiple and Mixed Data Type Imputation

Multiple imputation (MI) is a statistical technique that has gained widespread use for handling missing data in various fields including geostatistics. Missing data in geostatistical data sets may arise due to a variety of reasons such as equipment failure, environmental conditions, costs, or sampling errors. MI addresses the missing data problem by generating multiple plausible values for each missing data point based on a statistical model that considers the relationships between variables (Yuan, 2010).

In the context of multivariate, which consists of primary and secondary data obtained from different locations, creating an appropriate imputation model can be challenging because it is difficult to determine the relationship between heterotopic data. However, the use of DEM can help establish the relationship of heterotopic data and facilitate the implementation of MI. Consequently, even if the data is not at the same location, the MI process can be effectively performed using the DEM, so the accuracy of the estimation model using the imputed data can be improved.

5.1 Concept of Multiple Imputation (MI)

Imputation replaces each missing value with acceptable values representing the distribution of possibilities (Barnett & Deutsch, 2013). Multiple imputation (MI) process performs multiple realizations of the imputation process to create completed data sets. There are several steps involved in dealing with missing data in MI. The first step is identifying which variables have missing values. Secondly, it is necessary to explore the relationships between the variables that already have available data. With a better understanding of the data, the next step is to create conditional distributions of the empty data locations to generate multiple plausible values for each missing data point. Finally, these completed data sets can then be used for geostatistical analyses. Multiple imputed data sets reflect uncertainty and can be combined into a single data set representing the MI outcome through averaging.

A key process in the imputation framework is constructing conditional distributions for all locations where data replacement is needed. The observed values of variable Z are labeled as Z_{obs} while the missing values are labeled as Z_{mis}. A conditional probability density function, denoted as $f(Z_{mis}|Z_{obs})$, is generated for the missing values using a prior model and the available observed data. These conditional distributions are utilized in simulations to create multiple realizations of data. This allows for the imputation of missing values (Barnett & Deutsch, 2012). Figure 5.1 demon-

strates how to create conditional distributions at missing data locations. The following workflow is used to make conditional distributions at the locations with missing data.

1. Order locations and variables in decreasing order of information.
2. Establish conditional distribution 1 based on collocated data.
3. Establish conditional distribution 2 based on spatial data.
4. Merge the conditional distribution.
5. Sample merged distribution and continue this loop

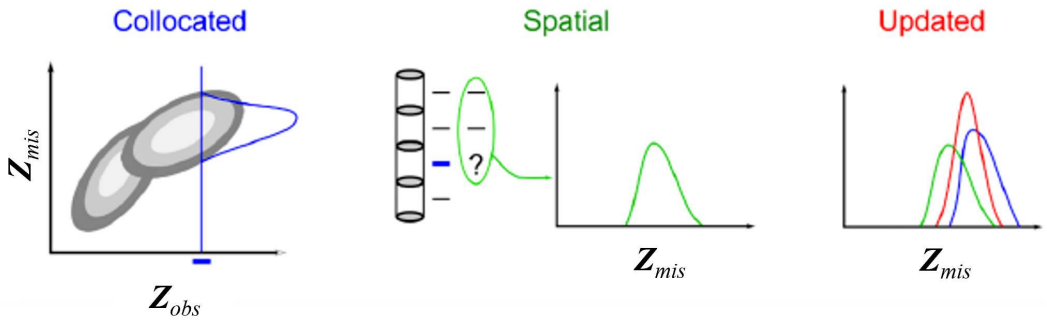


Figure 5.1: Schematic illustration of the process of creating a conditional distribution at missing data locations (Resource_Modeling_Solutions_Ltd, 2022).

Techniques for building $f(Z_{mis}|Z_{obs})$ in these works include nearest neighbor, universal kriging, kernel density estimation, and neural networks (Barnett & Deutsch, 2015). Among them, Kernel density estimation (KDE) (Scott, 2015) with Gibbs sampler (Gelfand & Smith, 1990) which iteratively creates and samples from the missing value distribution is a common method to construct conditional distribution. However, this can be computationally expensive depending on the number of dimensions and data observations (D. S. F. Silva & Deutsch, 2015).

To address these challenges, D. S. Silva and Deutsch (2018) introduced multiple imputation using Gaussian Mixture Models (GMMs) which is based on the expectation maximization (EM) algorithm. The GMM serves as an estimate of the multivariate probability density function and significantly improves computational efficiency. It also allows for quick assessment of any marginal and conditional distributions required for the non-parametric data imputation workflow (D. S. Silva & Deutsch, 2018). A review of GMMs can be found in Chapter 2.3.

MI is one of the powerful methods that can greatly enhance the accuracy and precision of geostatistical estimation models by reducing bias and increasing the sample size. However, it is essential to acknowledge that the quality of the statistical model used for imputation and the extent of missing data influence the accuracy of MI estimates (Rubin, 2004). To obtain reliable results, it

is crucial to ensure accurate variograms and multivariate relationships, making a thorough review of the procedures for obtaining these input parameters necessary.

5.2 Synthetic Example of MI Using DEM

To execute MI, a variogram of the targeted variable for imputation and GMMs that expresses the relationship between all variables are required. However, when the data are heterotopic, imputation can be challenging as it is not possible to obtain a GMM based on collocated data.

A DEM is a tool that can be used to infer the relationship between heterotopic data by detecting errors and biases in the secondary variable relative to the primary variable, which is assumed to be error-free. DEMs become particularly advantageous when dealing with heterotopic data, where developing imputation models can be challenging. By using DEMs to capture relationships between heterotopic data, more accurate imputation models can be created, resulting in improved estimation models that benefit from a larger data set featuring the primary variable.

In this section, the synthetic example shows multiple imputation using DEM when multiple data types have heterotopic data. The reference model used in the experiment has an isotropic variogram of 45m, a mean of 2.05, and a standard deviation of 0.7 in an area of 256m*256m in size. The reference model provides a benchmark to evaluate the accuracy of the final estimation model.

The data sampled from this reference model is divided into a primary variable and a secondary variable. The primary variable is 100 error-free data spread widely in the reference model. The secondary variable collected in specific areas is adjusted to have errors and biases using the parameters $a = 0.20$, $b = 0.05$, $c = 0.10$, and $d = 0.05$ of the reference DEM. Primary and secondary variables do not have data at the same location. Location maps and histograms of the reference model and two variables are shown in Figure 5.2.

To represent the error and bias of secondary data from primary data using DEM, primary data at the location of secondary data are needed. Simulation of the primary variable provides possible values with properties of the primary variable at the location of the secondary variable. Simulations reproduce the original variability observed in the data and allow an assessment of uncertainty (Rossi & Deutsch, 2014). Thus, DEMs made from realizations generated through simulation have the advantage over those made using a kriging model in that they know the probability distributions of the DEM parameters and can allow uncertainty assessments.

Since multiple simulations and transformations are in progress, validation is recommended at each step to check uncertainties and finally obtain more accurate results. The goal of the simulation is to reproduce the input histogram and variogram (Rossi & Deutsch, 2014). Figure 5.3 shows the histogram and variogram reproduction plots of the simulation using the primary variable. If acceptable simulation results are obtained, extract values from the same location as the secondary variable in the simulation model so that the DEM can find the error and bias of the secondary

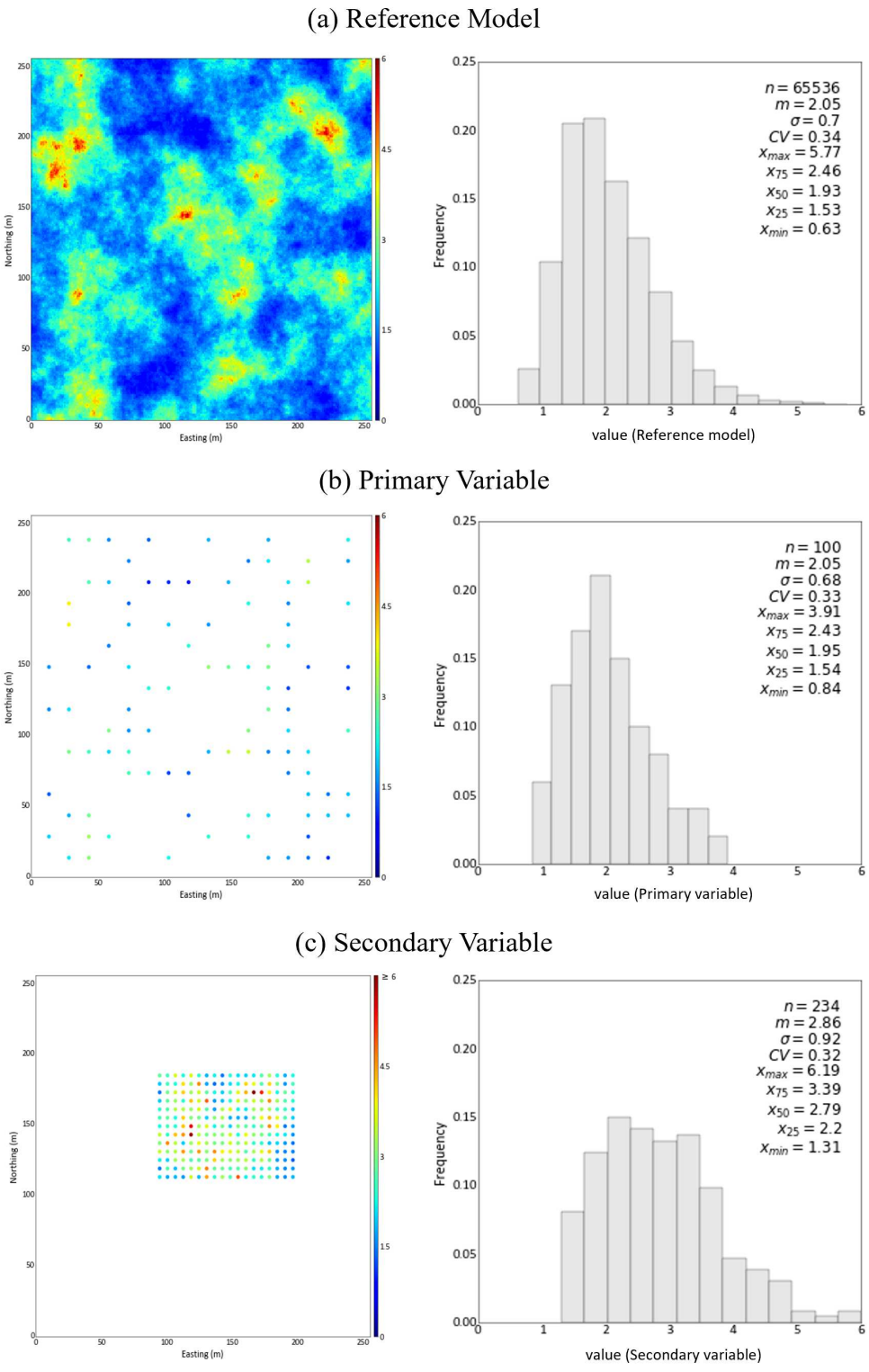


Figure 5.2: The primary variable (b) and secondary variable (c) are sampled from the reference model (a). The secondary variable is adjusted to have errors and biases.

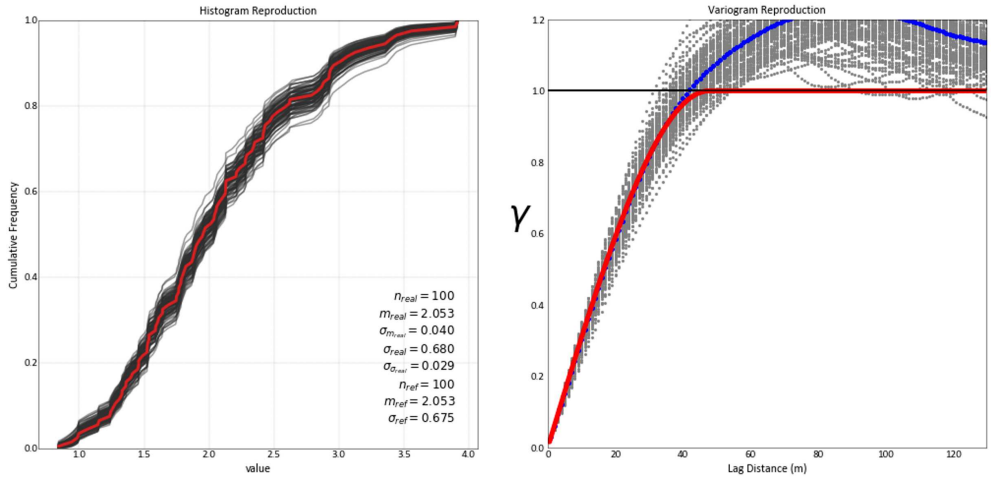


Figure 5.3: Validation is necessary to ensure the accuracy of input data for the DEM process. The histogram and variogram reproductions of simulation results demonstrate that the results closely resemble the original primary data.

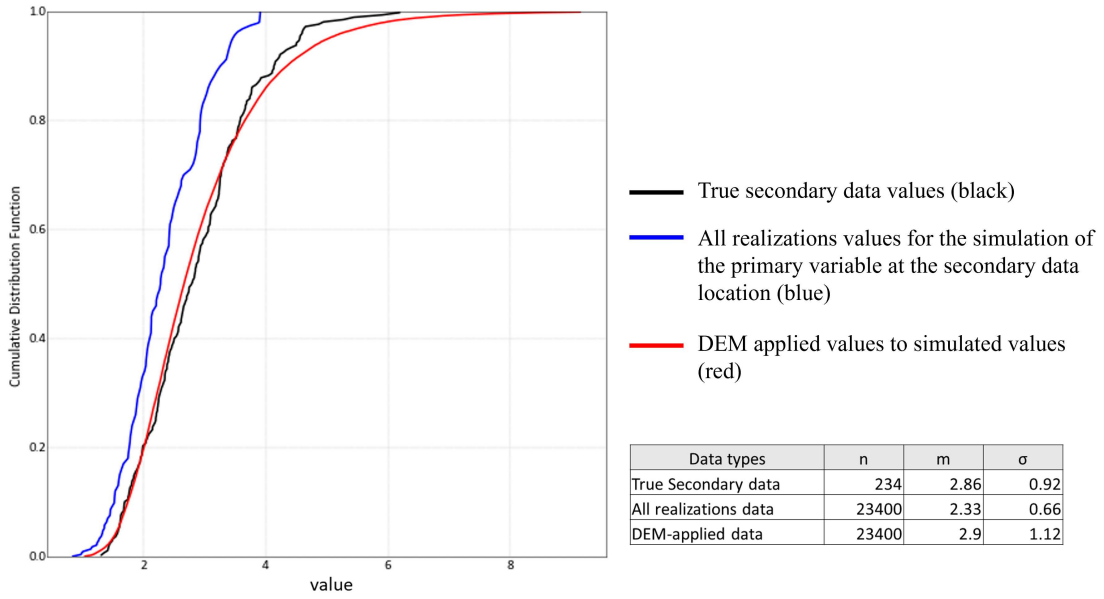
variable in each realization.

The process of identifying suitable DEMs involves conducting pairing analysis and adjusting the parameters of the DEM. The goal is to match the correlation coefficient and mean difference of the pairing analysis results between the primary and DEM-applied data to those obtained from the true secondary and primary data. This matching is done within an appropriate search radius to minimize noise and pair related data. By fine-tuning the DEM parameters to mimic the results of true data pairing analysis, the DEM-applied data can better account for errors in the secondary variable. The average DEM of DEMs obtained from all realizations is then taken as the optimal DEM representative of a given modeling region. This optimal DEM is essentially a mathematical model that transforms the distribution of the primary variable into the distribution of the secondary variable within the region of interest.

In this example, the optimal DEM has $a = 0.23882$, $b = 0.05069$, $c = 0.18526$, and $d = 0.05009$ with a search radius of 5m. The pairing analysis results of the data that underwent DEM-applied data and the primary data exhibit an average difference of 0.35% in correlation coefficient and 0.01% in the mean difference of pairs when compared to the pairing results of the true variables. The difference in pairing analysis results is an indicator of how well the DEM represents the relationship between primary and secondary variables.

In addition, it is important to assess whether the DEM-applied data accurately adheres to the distribution of the secondary variable. Figure 5.4 illustrates the distribution of the simulated data from the primary variable at the location of the secondary data, the data distribution after applying the DEM, and the distribution of the original secondary variable data. The comparison shows that the DEM-applied data aligns well with the distribution of the secondary data after being transformed

from its original distribution, with a mean difference of 1.46% and a standard deviation difference of 22.37%. This shows a better depiction of the secondary variable distribution than DEMs obtained from different search radii, demonstrating a reasonable DEM.



Data types	n	m	σ
True Secondary data	234	2.86	0.92
All realizations data	23400	2.33	0.66
DEM-applied data	23400	2.9	1.12

Figure 5.4: After applying DEM to the simulated data (blue), the result (red) more closely follows the distribution of the secondary variable (black).

Simulation is used to obtain the distribution of the possible data as a secondary variable (output data) at the same location when the DEM (mathematical model) is applied to the primary variable (input data). The simulation method produces a more continuous distribution of outputs from the discrete, sparsely dense primary variable through iterative random sampling of the input variable's distribution and model application processing. This is useful for analyzing uncertain scenarios and providing probabilistic analysis for different situations (Raychaudhuri, 2008). The distribution of inferred secondary variables obtained through simulation by applying the true primary data to the DEM is shown in Figure 5.5. The relationship between the two variables can be inferred as a multivariate GMM using the EM algorithm (Biernacki, Celeux, & Govaert, 2003). Figure 5.6 is the GMM and cross plots obtained based on the simulation results.

As a result, DEMs can overcome the challenge of finding relationships between heterotopic data. The DEM makes it possible to infer the distribution of secondary variables by identifying errors and biases in the secondary variables through simulated primary data and pairing analysis. Through simulation, the inferred relationship that appears if each other exists at the same location can be expressed through GMM, this has been instrumental in enabling multiple imputation (MI) of multiple data types.

The goal of the MI process in this experiment is to increase the accuracy of the estimation model with more primary data by adding imputation values to empty secondary data locations in

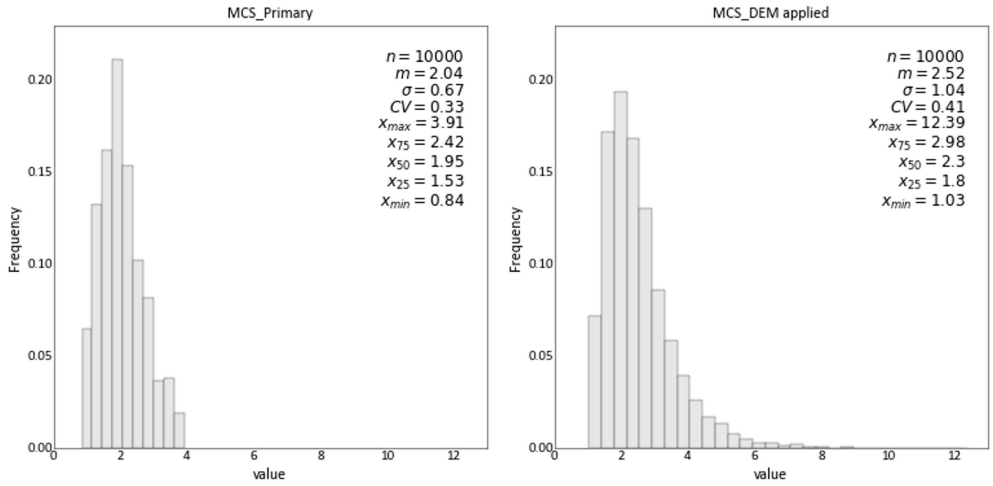


Figure 5.5: The graph shows the distribution of the primary variable and inferred collocated secondary variable obtained using the simulations (10,000 trials) with the optimal DEM as the mathematical model.

the primary data set. Performing MI requires individual models that describe the variability of the variables targeted for imputation, and models that capture the relationships between the variables, such as variogram models and GMM. The variogram model of primary data, which is the object of imputation, can be obtained with experiment variograms. The relationship between primary data and secondary data that do not exist at the same location is solved through DEM.

A total of 100 MI realizations are conducted, and Figure 5.7 depicts the histogram and variogram reproduction plots of the MI outcomes. The histograms of the realizations exhibit a slight positive bias. This is attributed to the MI results following the characteristics of the primary variable, but generating imputed data with numerous high values. The back-transformation of the histogram into the original units further confirms this observation.

The black line in the plot represents the histogram of reference values without error or bias in the region with the secondary variable, and it demonstrates higher values compared to the widespread true primary data (red line). This disparity in values signifies the impact of the imputation process and emphasizes the importance of careful interpretation and consideration of the imputed data.

To assess the impact of MI using DEM on the accuracy of the estimation model, three estimation models are developed using ordinary kriging. The first model uses only primary data for estimation, while the second model utilizes both primary and secondary data. The third model incorporates the average data of homotopic data sets generated through MI, encompassing both the original primary data and the imputed data replaced. This approach checks the performance of the imputed values obtained from MI using the relationships inferred from the DEM through comparative analysis of estimation models.

The estimation models are evaluated and compared with the reference model at both global and local scales. On a global scale (Figure 5.8), the kriging model created using both primary and

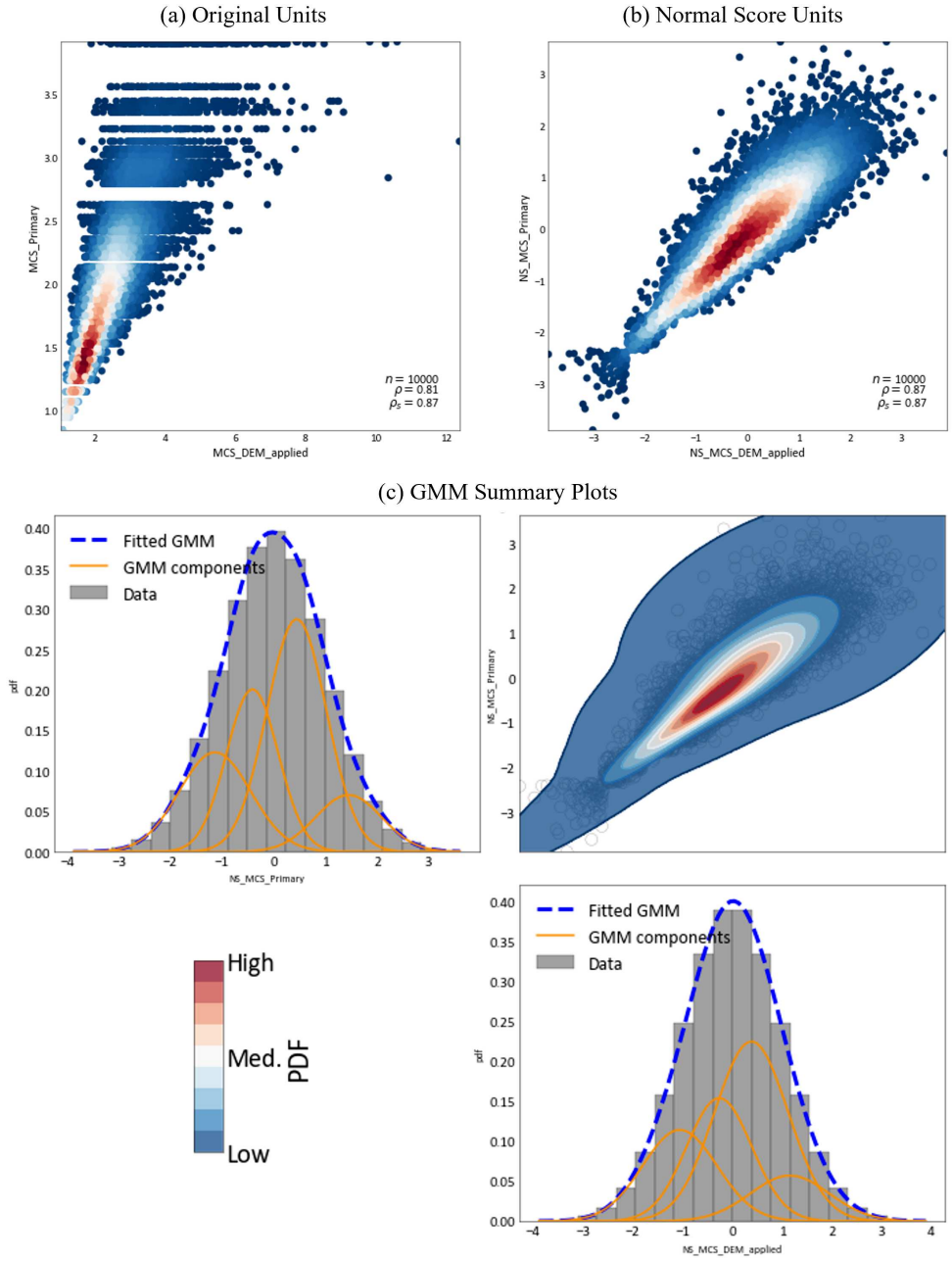


Figure 5.6: The figure shows (a) a cross plot between primary data and DEM applied data after MCS, (b) the relationship between the two variables in normal scores units, and (c) the GMM summary plot between them.

secondary variables exhibits a higher correlation with the reference model than the kriging model created solely with the primary variable. However, the root-mean-square-error (RMSE) is higher due to errors present in the secondary data. The estimation model created using the average of data sets generated by MI shows improved accuracy compared to the previous two models, as it demonstrates high correlation and low RMSE.

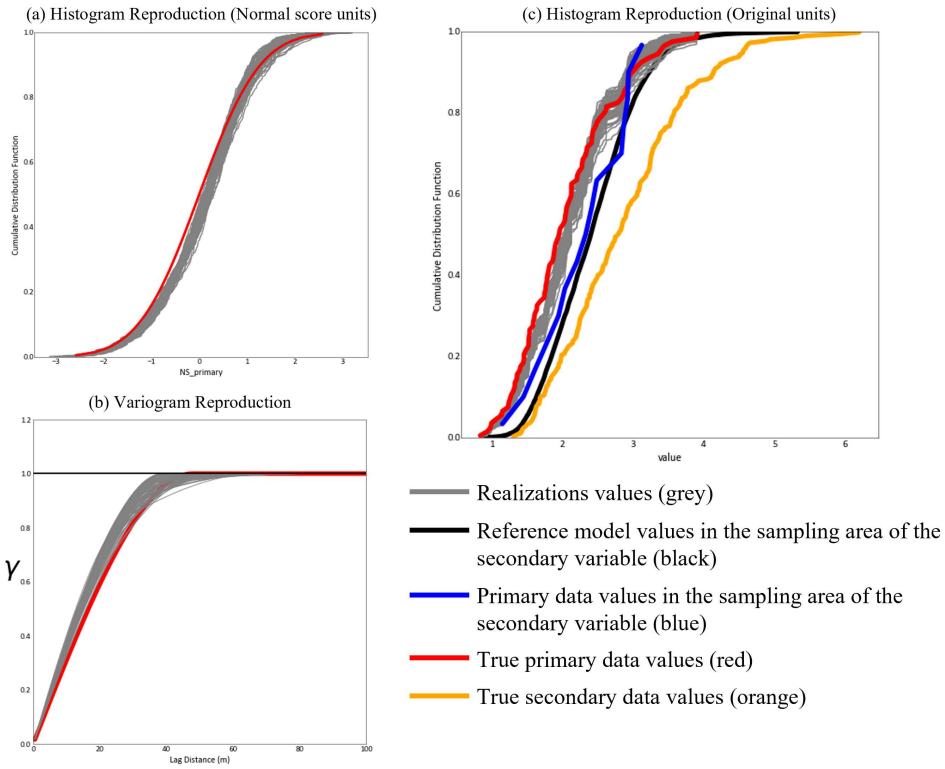


Figure 5.7: After 100 multiple imputations, the histogram and variogram are reproduced to confirm the validity. The histogram plot of the original units demonstrates that the distribution of the imputed results is slightly skewed to the right because the imputed data have higher values than the original primary data.

On the local scale (Figure 5.9), where the secondary variable is sampled, the estimation model using only the original data shows a high RMSE due to the inherent errors in the secondary variable, despite its high correlation. However, upon applying imputation through MI, a significant improvement in accuracy is observed at the local scale. When using the MI, the estimation of the region where the secondary variable exists shows a clear enhancement compared to the global scale estimation.

Indeed, both the global scale and local scale results confirm the validity of the relationship inferred by the DEM within the framework of MI. The successful application of DEM-based imputation methods underscores their capability to yield more reliable estimation model results, particularly in domains where secondary data are present.

Additional experiments are conducted to comprehensively evaluate the effects of the characteristics of secondary variables and the amount of data on DEM and MI. This rigorous approach aims to gain deeper insights into their behavior and performance in various scenarios, ultimately contributing to a more thorough understanding of their impact on data imputation and analysis.

In order to confirm that there is a slight bias in the imputation results depending on the characteristics of the secondary data, the same experiment is repeated by selecting an area with a low value

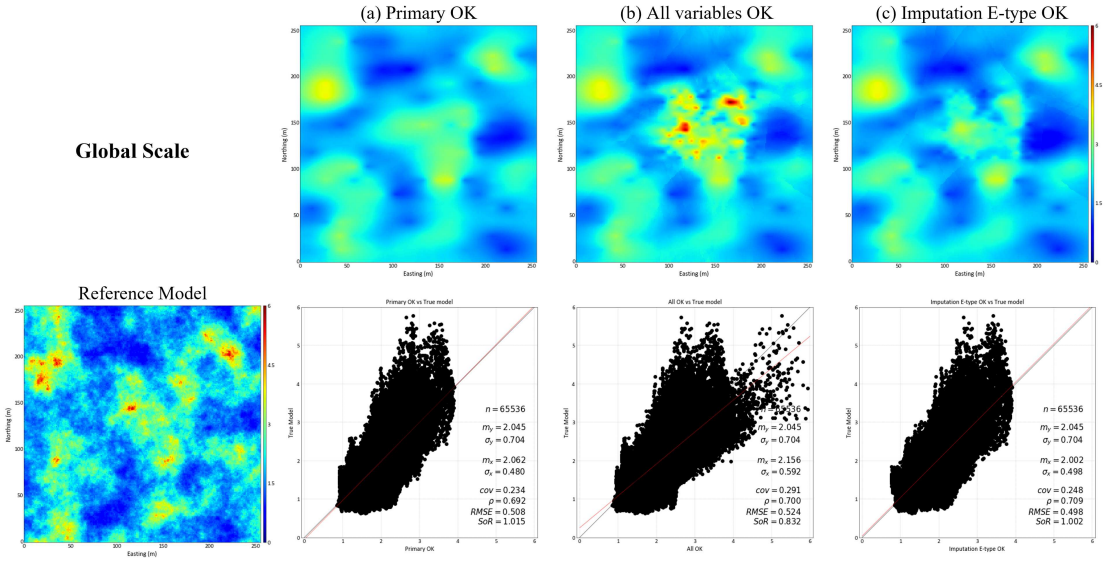


Figure 5.8: (a) $\rho = 0.692$, RMSE $= 0.508$, (b) $\rho = 0.700$, RMSE $= 0.524$, (c) $\rho = 0.709$, RMSE $= 0.498$

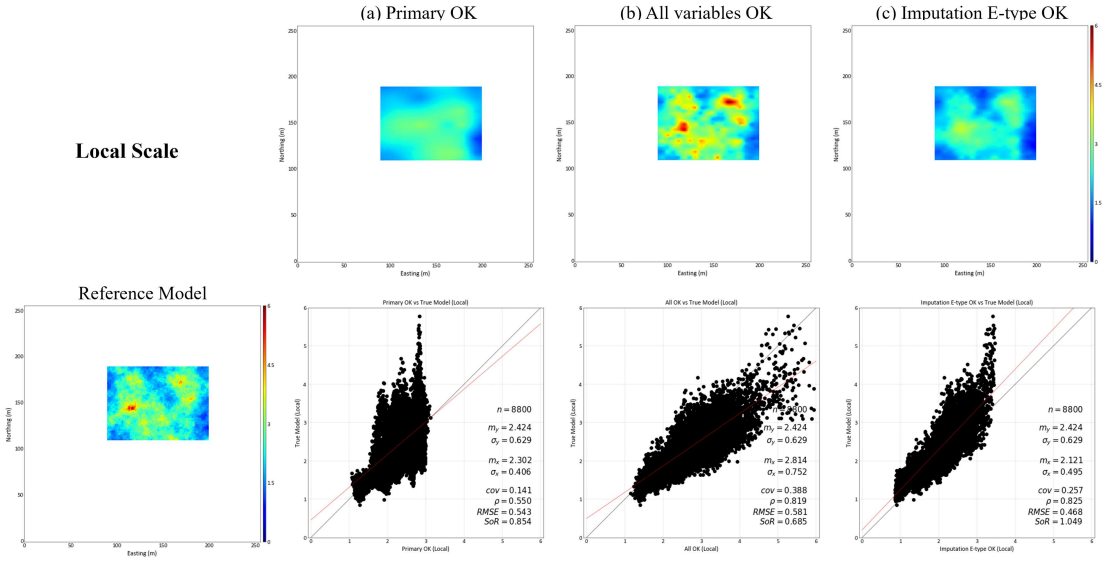


Figure 5.9: (a) $\rho = 0.550$, RMSE $= 0.543$, (b) $\rho = 0.819$, RMSE $= 0.581$, (c) $\rho = 0.821$, RMSE $= 0.468$

in the same reference model as a place to sample the secondary variable. As shown in Figure 5.10, the imputed data sets have many lower values than the original primary variable, and the histogram is skewed to the left. Therefore, it can be said that the MI performs well in these experiments following the characteristics of the primary variable, which is an error-free variable.

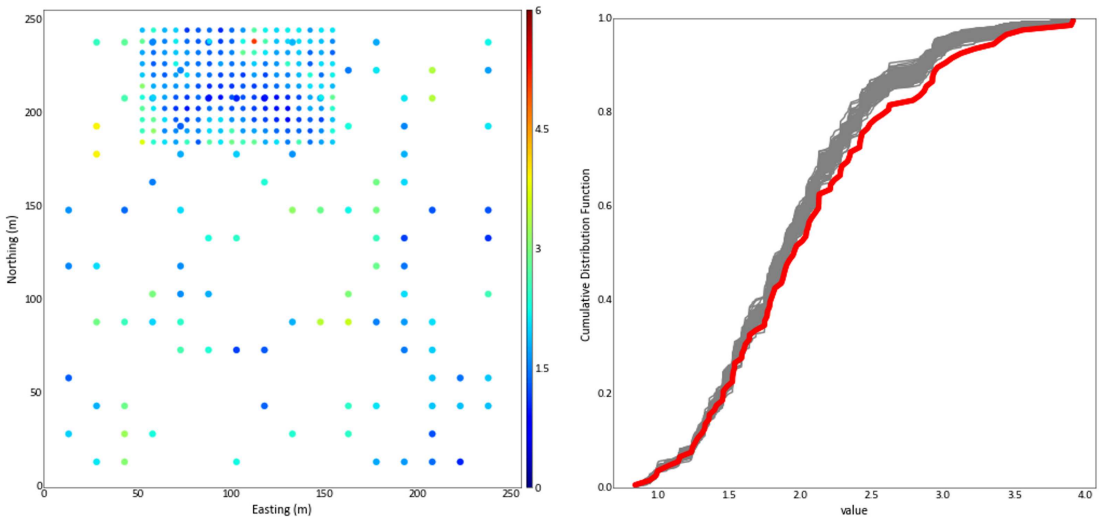


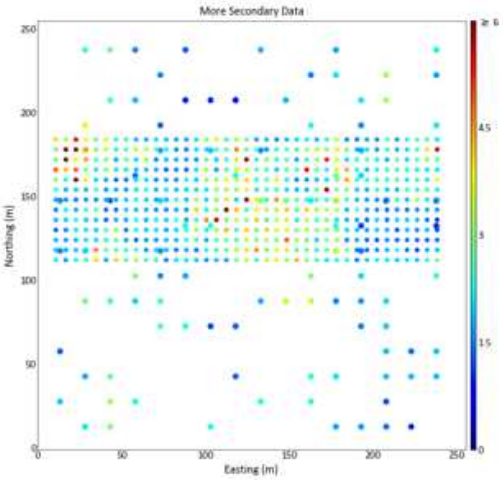
Figure 5.10: When the secondary variable is sampled in the low-value region, the distribution of MI results is skewed to the left of the original distribution of the primary variable.

Another example is conducted to investigate the potential impact of the amount of data on the accuracy of the DEM, MI results validity, and estimation model performance. As in the previous experiment, the data points are not at the same location. However, 507 secondary data are used, more than before.

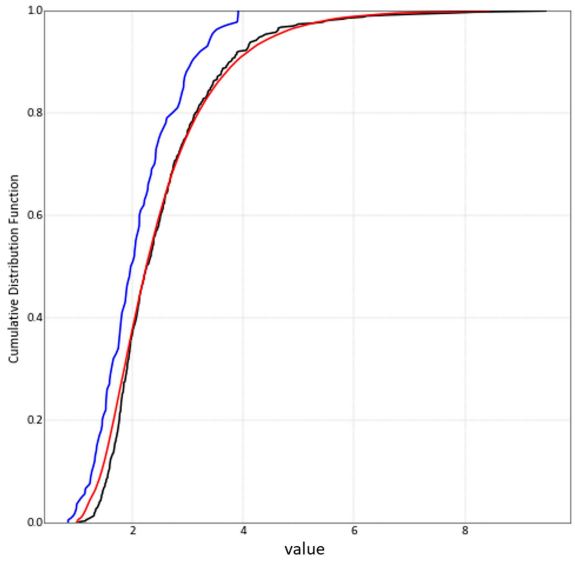
The following Figure 5.11 shows the result when the DEM workflow and MI are performed with more secondary data. Figure 5.11a indicates that the area of overlap between secondary and primary data is wider than before. The distribution of optimal DEM application data is shown in Figure 5.11b shows that the more secondary data, the more accurate the DEM. When DEM is applied to simulated data from primary data and compared to the original secondary data, the average differs by 1.15% and the standard deviation differs by 2.42%. This is an improvement over the main experimental results performed above. The cross plot in Figure 5.11c compares the kriging results using only primary data, the kriging results using primary and secondary data, and the kriging results using imputed data with the reference model in turn. When compared with the reference model, the estimation model using imputed data shows the best results with a correlation coefficient of 0.734 and RMSE of 0.478.

The results indicate that the DEM does a better job of transforming the distribution of the primary variable into that of the secondary variable as the amount of secondary data increases. In addition, the accuracy of the estimation model from MI using DEM improved.

(a) Location map with more data



(b) DEM-applied data distribution



True secondary data values (black)

All realizations values for the simulation of the primary variable at the secondary data location (blue)

DEM applied values to simulated values (red)

Data types	n	m	σ
Original Secondary data	507	2.53	1.04
Simulated data	50700	2.09	0.71
DEM-applied data	50700	2.5	1.06

(c) Cross-validation

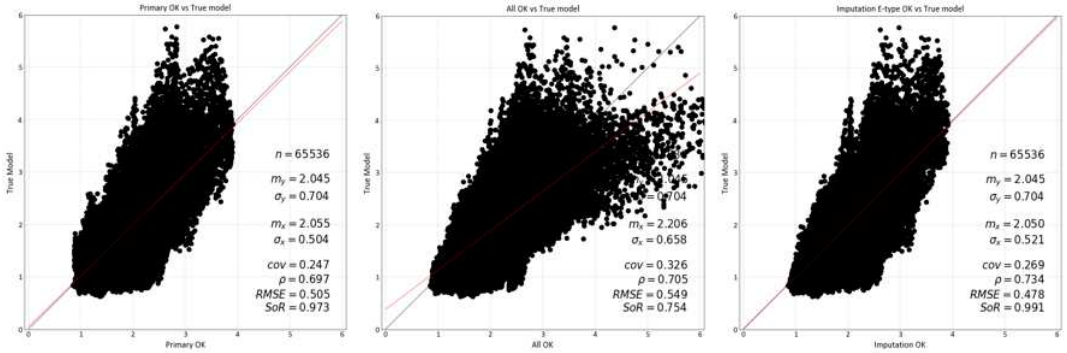


Figure 5.11: The incorporation of more secondary data enhances the ability of the DEM to capture errors and biases present in the secondary variable. This, in turn, significantly contributes to the improvement of the estimation model's accuracy through MI.

Chapter 6

Case Study: Application of DEM Workflow and MI for Multiple Data Types at Rain Mine

The DEM and multiple imputation described in the previous chapters are applied with exploration data and production data from the Rain Mine. Data is provided by the mine owner, Newmont Gold Corporation, for study.

6.1 Background

Rain Mine, located in Elko County, Nevada, is a gold mine located within the Carlin Trend Mining District. The mine employs a combination of surface and underground mining techniques to extract ore comprising cinnabar, calcite, and kaolinite. The waste material primarily consists of barite. The ore body itself has a tabular and irregular shape, measuring a thickness of 106 meters (350 feet). The host rock in this region is shale, formed during the Lower Mississippian epoch approximately 350 million years ago (DiggingsTM, 2023).

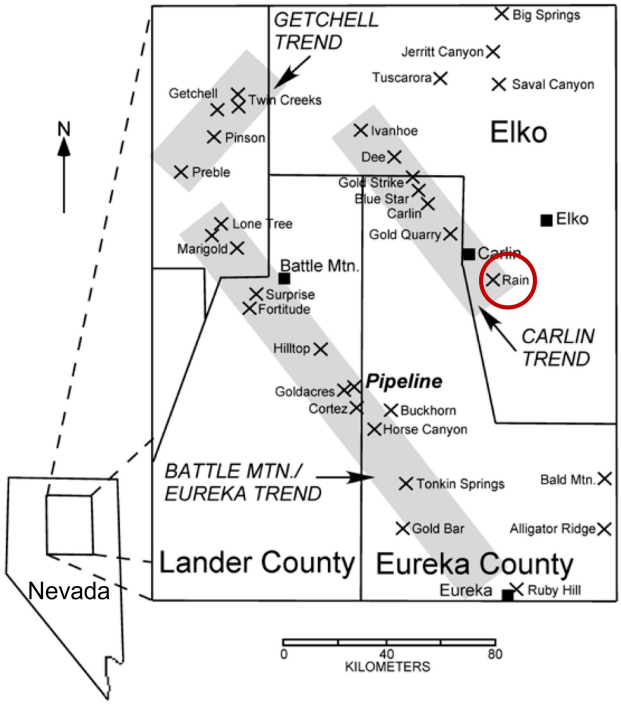


Figure 6.1: Location of Rain Mine in Carlin Trend. Taken from Blamey et al. (2017).

6.2 Data Set

For this experiment, 619 exploration drill holes and 6054 production blast holes are considered in a part of Rain Mine. Figure 6.2 presents a visual representation of the location of drill and blast holes within the area of interest. Exploration data is considered error-free and classified as the primary variable, while production data is assigned as a secondary variable serving as auxiliary data. There is no data at the same location for both primary and secondary variables. The histogram of these two variables is depicted in Figure 6.3. Exploration drills are typically conducted densely in areas where high quality is expected. Therefore, to prevent clustering effects from interfering with the overall interpretation, declustering weights are applied to the primary variable.

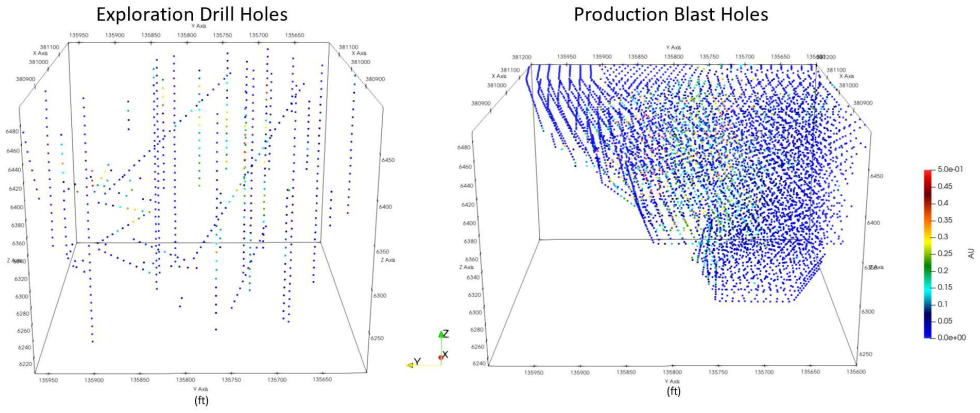


Figure 6.2: Location maps of exploration and production data in Rain Mine.

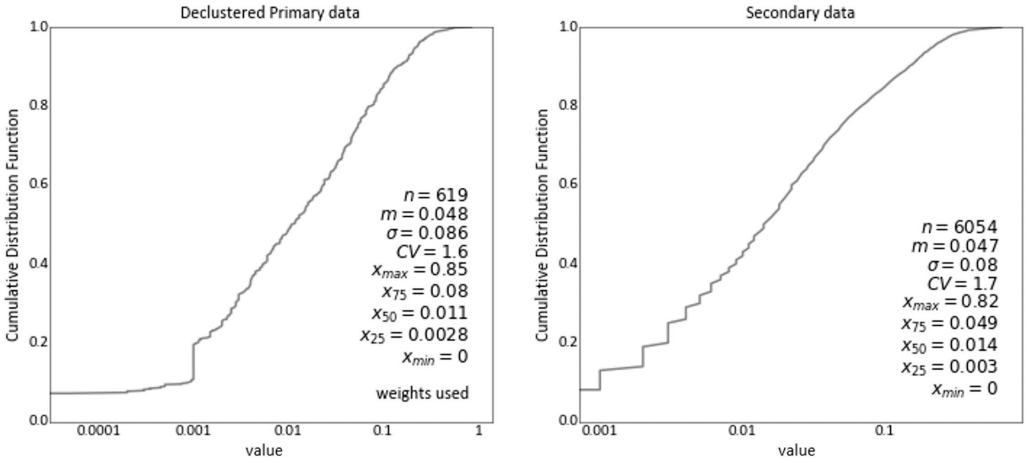
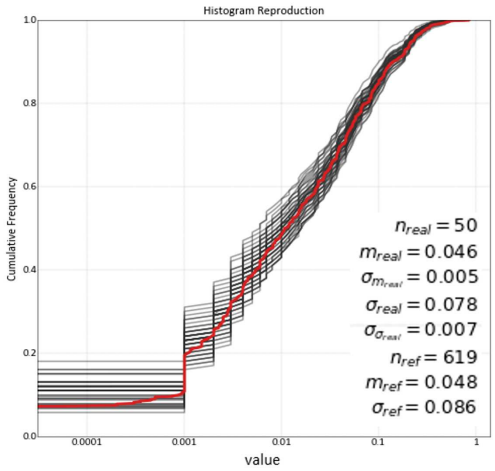


Figure 6.3: Cumulative density function (CDF) of exploration and production data

6.3 Optimal DEM

To facilitate the imputation of heterotopic data, DEM workflows provide relationships between primary and secondary variables. The first step in obtaining a DEM is getting the simulation models of the primary variable. The variogram model is obtained from the experimental variograms and the estimation results are obtained by running 50 simulations. Figure 6.4 shows histogram and variogram reproduction plots to validate this simulation. Since DEMs are created using simulated data, it is important to obtain valid verification. Upon attaining simulation models aligning with primary variable characteristics, data is extracted from secondary variable locations to construct virtual secondary data for DEM application. This process is performed using Simulation Data Extraction Program (`GETSECREAL`). Figure 6.5 illustrates the CDF of simulated primary data at secondary data locations across 50 realizations.

(a) Histogram reproduction of primary data simulation



(b) Variogram reproduction of primary data simulation

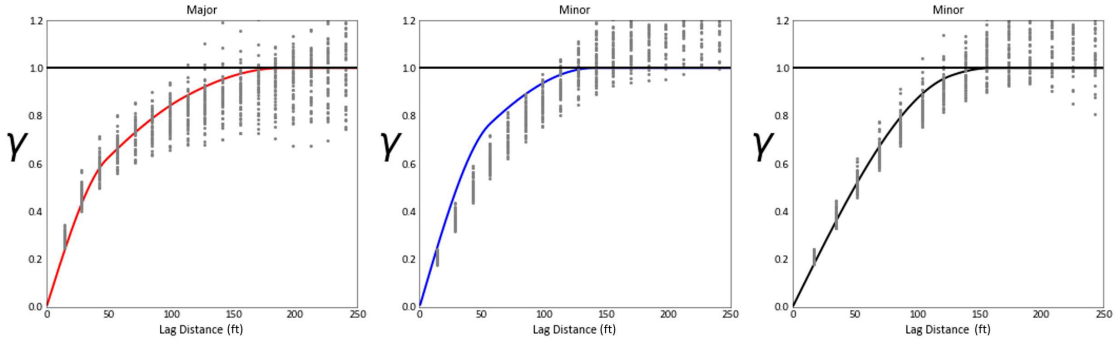


Figure 6.4: (a) shows the histogram reproduction in original units and the CDF of the primary variable. (b) shows the variogram reproduction and the variogram model of the primary variable.

The next step is the pairing analysis between the true primary variable and the true secondary variable, which is the result of the pairing analysis that the relationship of the pairs between the real primary data and the DEM-applied data must follow for suitable DEM describing secondary errors

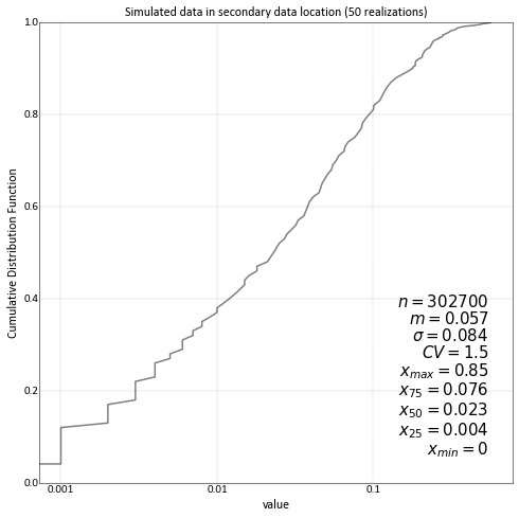
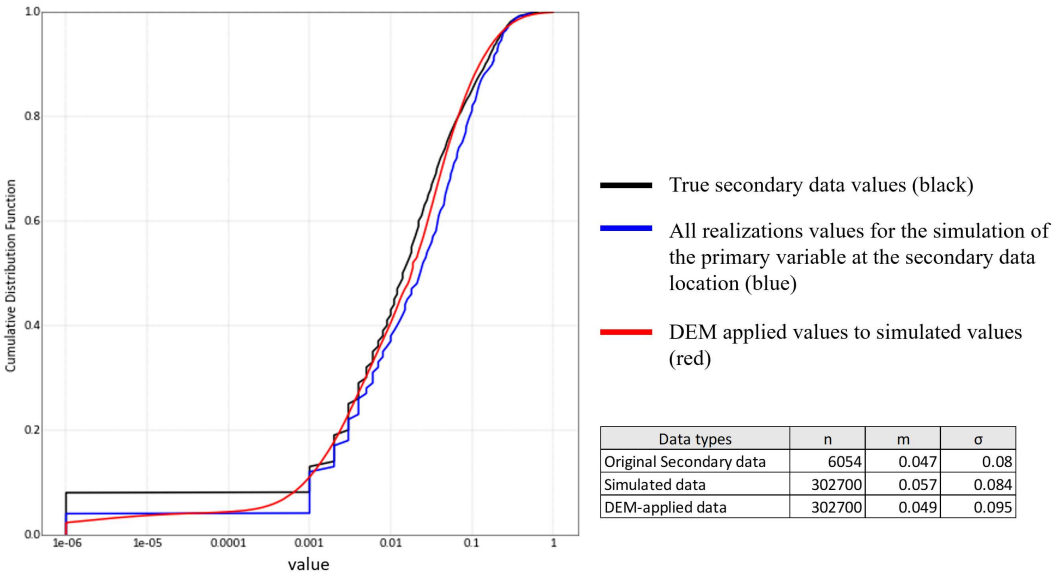


Figure 6.5: This is CDF of the data extracted from secondary data locations from 50 realizations of the primary data simulations. The extracted data becomes the input data for finding a DEM suitable for each realization.

and biases. The DEM parameters are adjusted to ensure that the correlation coefficients and mean differences obtained from the pairing analysis match the true data. Until convergence to the reference result, iterative pairing analysis comparison and DEM parameter update are performed through the Optimal DEM Finder Program (`OPTDEM`). This workflow assumes the average of the DEM parameters for each realization as the optimal DEM representative of the model which describes errors and biases of secondary variable. Utilizing `OPTDEM`, the optimal DEM for the secondary variable is determined with $a = 0.177$, $b = 0.746$, $c = -0.277$, and $d = 0$ within a search region of 30ft. When this optimal DEM is applied to the simulated primary data, the pairing analysis results show a difference of 12.771% in correlation coefficient and 0.202% in mean difference from the criterion result performed with actual data. Figure 6.6 shows the distribution of the simulated data shifted towards the distribution of the real secondary variable after applying the DEM.

The optimized DEM accounts for secondary variable errors and biases relative to the primary variable. Simulation applies the DEM to primary data to produce the expected distribution of secondary data at the same location. Through 10,000 trials of the Monte-Carlo method, data from the primary variable, with declustered weights, is applied to DEM to obtain the expected distribution of the secondary variable. Figure 6.7 illustrates the distribution of primary and inferred secondary data derived from Monte-Carlo simulations. Utilizing this distribution, a GMM is constructed, as shown in Figure 6.8, enabling expression of the inferred relationship between variables required for subsequent MI steps.



Data types	n	m	σ
Original Secondary data	6054	0.047	0.08
Simulated data	302700	0.057	0.084
DEM-applied data	302700	0.049	0.095

Figure 6.6: Applying the optimal DEM to the simulated data for the primary variable moved the data closer to the actual secondary data. For the logarithmic calculation of the DEM, zero-valued data was replaced by an extreme small number 0.000001.

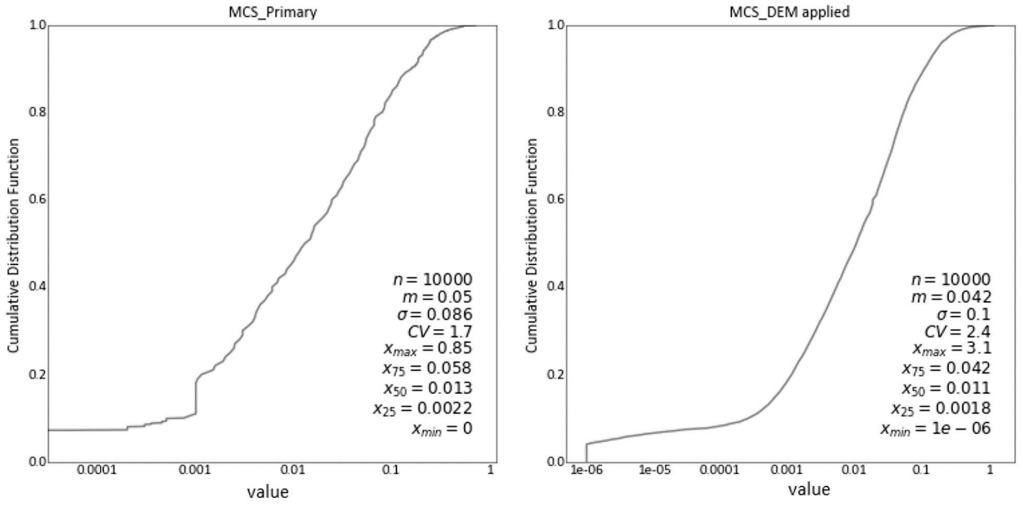


Figure 6.7: The figure on the right shows the CDF of 10,000 random sampling of the primary variable, and the figure on the left shows the CDF of the data where DEM is applied to the sampled data.

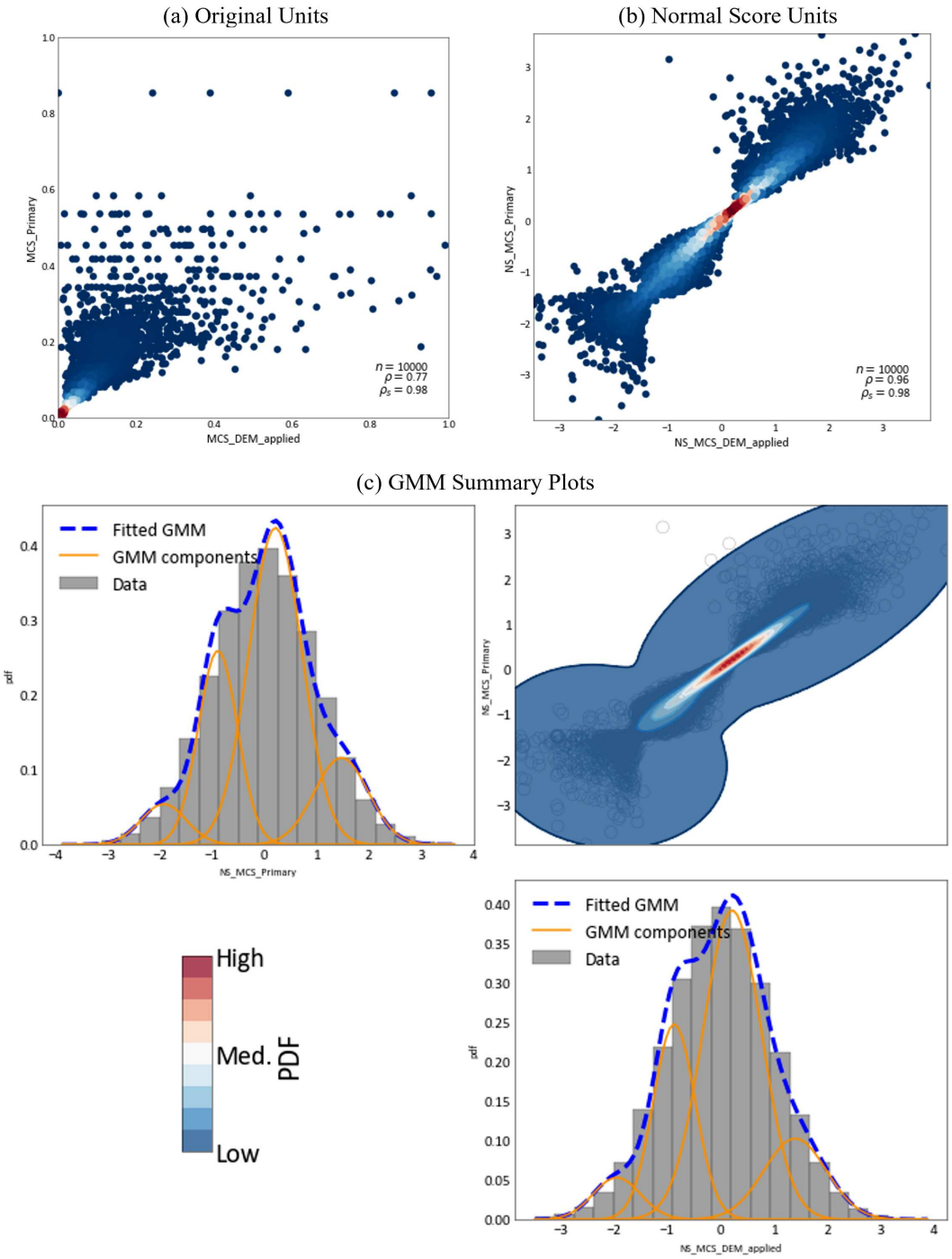
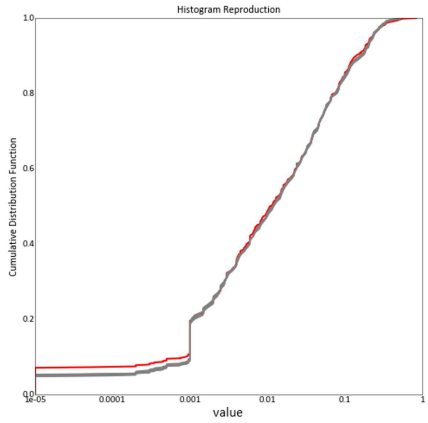


Figure 6.8: The relationship between the two variables can be expressed by creating a GMM through the distributions inferred from DEM and simulation.

6.4 Multiple Imputation

MI targets a primary variable and imputes missing data at the secondary data locations. As a result, more primary data can be obtained through MI. Figure 6.9 shows histogram and variogram reproduction plots for 10 MI realizations. Validation confirms that MI results align with the original primary data. Finally, 3D estimation models using ordinary kriging are constructed, including models based on only primary data, primary and secondary data, and primary data with imputed data. 3D figures and cross-sections of these models are visualized in Figure 6.10. A model using only the primary variable makes it difficult to identify veins. When two variables are used, the veins can be clearly identified, and when imputed data is used, a lower grade is predicted than when secondary data is used.

(a) Histogram reproduction of MI results



(b) Variogram reproduction of MI results

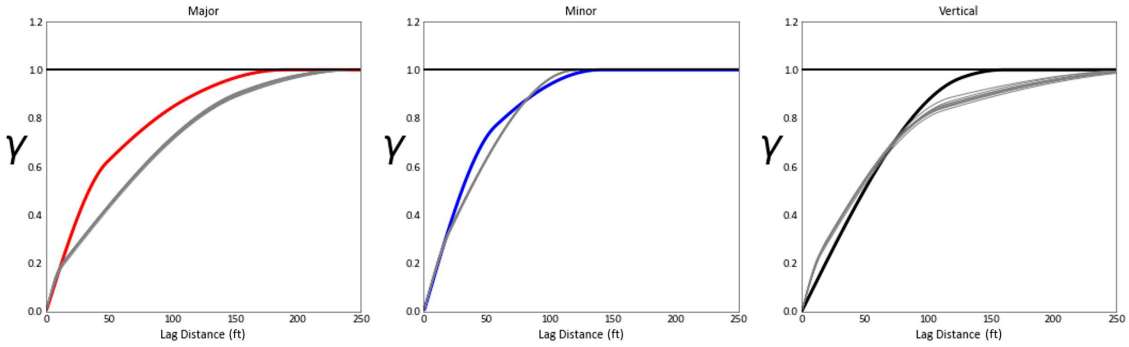
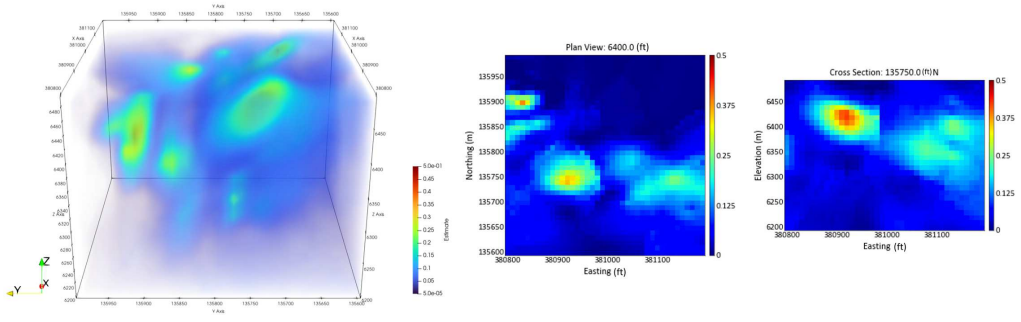
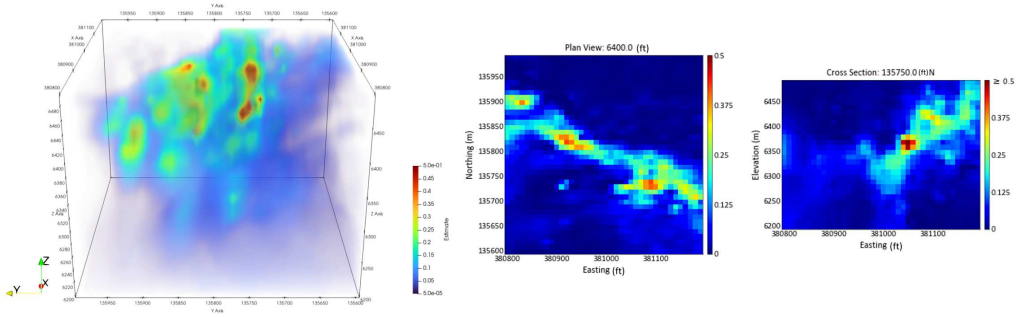


Figure 6.9: There are plots for MI validation. (a) shows the histogram reproduction in original units, and (b) shows the variogram reproduction.

(a) Estimation model from primary data



(b) Estimation model from primary data and secondary data



(c) Estimation model from primary data and imputed data

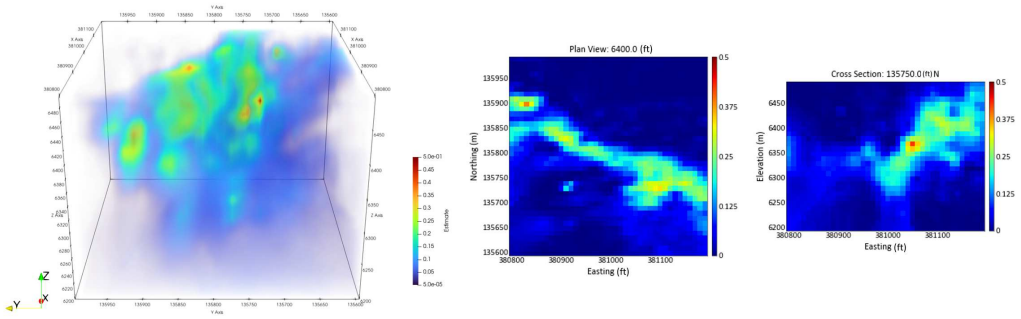
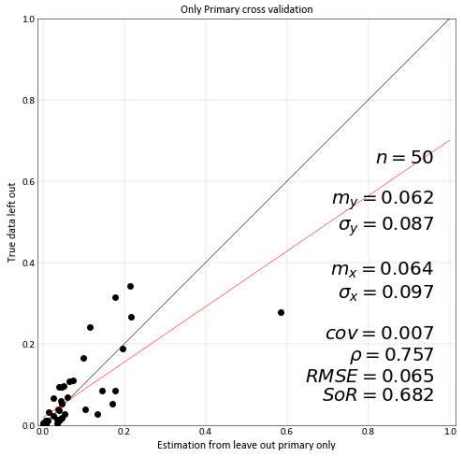


Figure 6.10: There are 3D estimation models for three cases. Ordinary kriging was used for estimation.

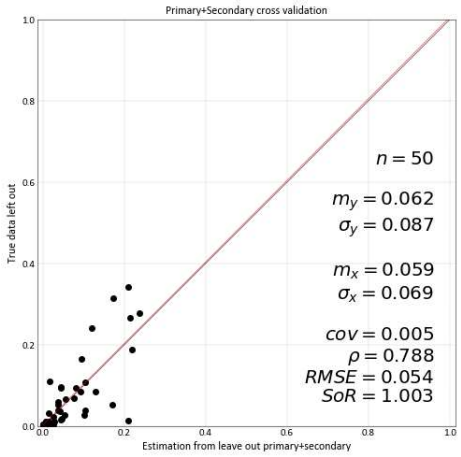
6.5 Cross-Validation

Imputation within the primary variable adds data with error-free properties to the location of the secondary data. Consequently, a more accurate estimation model can be created, substantiated by cross-validation. After removing 50 data points from the primary data, the estimation model is reconstructed and the data generated from the removed data locations are compared. Figure 6.11 displays cross-validation plots for the three cases of estimation models, which are OK based on primary data, primary and secondary data, and primary with imputed data. The estimation model with imputed data shows the highest correlation and lowest RMSE compared to other models. This indicates that the relationship between the heterotopic data inferred by the DEM worked well in the MI and consequently increased the accuracy of the estimation model.

(a) Cross-validation of the model using primary data



(b) Cross-validation of the model using primary data and secondary data



(c) Cross-validation of the model using primary data and imputed data

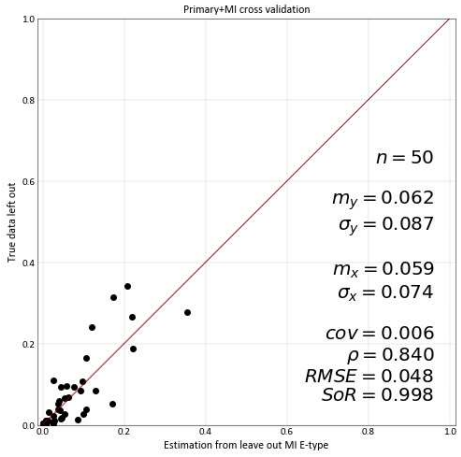


Figure 6.11: The cross-validation results for the three cases. The estimation model based on primary data and MI results have the highest correlation and lowest RMSE than those using primary data alone or primary and secondary data.

Chapter 7

Conclusion

This chapter reviews the motivation and summarizes the contributions to the advancement of estimation models involving multiple data types. Additionally, future research directions are described, setting the stage for future investigations in this area.

7.1 Review of the Motivation

Integrating multiple data types is presented as a way to increase the accuracy of estimation models. However, understanding the relationships between heterotopic data remains challenging. While several geostatistical methodologies utilize collocated data to enable multivariate modeling, the focus of this study is to consider heterotopic data and utilize novel models to improve modeling. The main motivation is to create a model that proficiently infers relationships between different data types to facilitate multiple imputation (MI) as a way to improve estimation models.

7.2 Summary of DEM and Contribution

A data error model (DEM) quantifies relative and absolute errors along with relative and absolute biases. DEM-based data captures errors and biases in secondary variables by iterative pairing analysis and DEM parameter updating. Pairing analysis allows interpretation of pairwise relationships such as correlations and mean differences. This analysis establishes connections between data that exist at different locations, and pairing analysis between the primary and secondary variables presents the relationships of DEM-applied data and the primary data that must follow.

The DEM workflow describes the relationship between primary and secondary variables. Inferred relationships from the DEM can be expressed in the form of a Gaussian mixture model (GMM). The efficacy of DEM is demonstrated in multiple imputation of heterotopic data.

By providing relationships between heterotopic data, DEMs facilitate imputations that require relationships between multiple variables that have heterotopic data. MI for the primary variable makes it possible to create estimation models with more data by replacing secondary data locations with data that has the characteristics of the primary variable. The examples provided show that as a result of conducting MI using the inferred relationship obtained by DEM, the estimation model made using imputed data is more accurate than the estimation model made using only the primary variable or the primary and secondary variables. In the case study using Nevada gold mine data, the estimation model created with primary data and the MI result using DEM provided a more accurate model than the estimation model made with only primary data or primary data and secondary data.

In conclusion, DEM facilitates MI of various data types and the results of MI improve estimation models.

7.3 Future Work

To improve the reliability of inferences drawn from DEMs, it is important to consider the variations in parameter values and properties across different geographical subsets. One way to achieve this is by partitioning the region into subsets and creating localized DEMs for each. These DEMs can then be subjected to sensitivity assessments to gain a comprehensive understanding of their dynamic behavior.

It may also be valuable to explore non-stationary conditions, as they can reveal spatial trends in various aspects. By reflecting on potential non-stationary phenomena, the DEM can better understand the complexity of geological data.

Expanding the DEM to include scenarios involving three or more data types can be helpful. It would also be interesting to explore using different scales, distribution shapes, and variable types within the DEM framework. Research on DEMs that incorporate more variables with more diverse properties will enable the inference of more complex multivariate relationships.

DEM infers the relationship between error and bias of the primary and secondary variables based on the primary variable, assuming that there is no error in the primary variable. However, there may be errors in the primary data. Future research efforts to account for primary variable errors will contribute to creating more accurate estimate models.

To ensure the integrity of primary data, it is important to account for procedural knowledge and sampling protocols. Also, by integrating laboratory and sampling processes, more accurate errors and biases of data can be known.

The presence of extreme values within a data set can have a significant impact on the accuracy of the DEM. In addition, a GMM created after applying DEM to sample values with a log distribution and converting them to normal score units tends to have a twisted appearance where small values exist. Analyzing and understanding these phenomena will more accurately describe DEMs and suggest more effective work frames.

Finally, the usability of the DEM can be improved by simplifying the MI technique. Efforts on MI that circumvent complex procedures while yielding accurate estimates instead of providing multiple simulated values using conditional distributions represent an interesting way to improve the efficiency of creating an estimation model and evaluation.

References

Barnett, R. M., & Deutsch, C. V. (2012). Missing data replacement in a complex multivariate context. *CCG Annual Report 14*, 113.

Barnett, R. M., & Deutsch, C. V. (2013). Imputation of geologic data. *CCG Annual Report 15*, 102.

Barnett, R. M., & Deutsch, C. V. (2015). Multivariate imputation of unequally sampled geological variables. *Mathematical Geosciences*, *47*(7), 791–817.

Barnett, R. M., Manchuk, J. G., & Deutsch, C. V. (2014). Projection pursuit multivariate transform. *Mathematical Geosciences*, *46*(3), 337–359.

Biernacki, C., Celeux, G., & Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, *41*(3-4), 561–575.

Blamey, N. J., Campbell, A. R., & Heizler, M. T. (2017). The hydrothermal fluid evolution of vein sets at the pipeline gold mine, nevada. *Minerals*, *7*(6), 100.

Bonamente, M. (2017). *Statistics and analysis of scientific data*. Springer.

Chiles, J.-P., & Delfiner, P. (2012). *Geostatistics: modeling spatial uncertainty* (Vol. 713). John Wiley & Sons.

Davis, B. M., & Greenes, K. A. (1983). Estimation using spatially distributed multivariate data: an example with coal quality. *Journal of the International Association for Mathematical Geology*, *15*(2), 287–300.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, *39*(1), 1–22.

Desbarats, A., & Dimitrakopoulos, R. (2000). Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. *Mathematical Geology*, *32*(8), 919–942.

de Souza et al., G. F. M. (2022). Chapter 5 - engineering systems fault detection methods. In G. F. M. de Souza, A. Caminada Netto, A. H. de Andrade Melani, M. A. de Carvalho Michalski, & R. F. da Silva (Eds.), *Reliability analysis and asset management of engineering systems* (p. 119-164). Elsevier. Retrieved from `https://www.sciencedirect.com/science/article/pii/B9780128235218000037` doi: https://doi.org/10.1016/B978-0-12-823521-8.00003-7

Deutsch, C. V. (2021). Citation in applied geostatistics.

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

Deutsch, J., & Deutsch, C. (2012). Kriging, stationarity and optimal estimation: Measures and suggestions. *Centre for Computational Geostatistics*, *14*, 306.

DiggingsTM. (2023). *Rain gold mine near carlin, nevada.* Retrieved 2023-08-09, from `http://thediggings.com/mines/usgs10310534`

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Gomes, C. G., Boisvert, J., & Deutsch, C. (2022). Gaussian mixture models. *Geostatistics Lessons*. Retrieved from `http://www.geostatisticslessons.com/lessons/gmm`

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

Goovaerts, P. (1998). Ordinary cokriging revisited. *Mathematical Geology*, *30*(1), 21–42.

Goulard, M., & Voltz, M. (1992). Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Mathematical Geology*, *24*, 269–286.

Gy, P. (2012). *Sampling of particulate materials theory and practice*. Elsevier.

Houlding, S. W. (1994). *The geological characterization process*. Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-642-79012-6_2` doi: 10.1007/978-3-642-79012-6_2

Isaaks, E. H., & Srivastava, M. R. (1989). *Applied geostatistics* (No. 551.72 ISA).

Journel, A. G., & Huijbregts, C. J. (1976). Mining geostatistics.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, *52*(6), 119–139.

Leuangthong, O., & Deutsch, C. V. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology*, *35*(2), 155–173.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, *58*(8), 1246–1266.

McLachlan, G. J., & Krishnan, T. (2007). *The em algorithm and extensions*. John Wiley & Sons.

Ortiz, R. B., & Deutsch, C. V. (2022). Multivariate gaussian distribution. *Geostatistics Lessons*. Retrieved from `http://www.geostatisticslessons.com/lessons/multigaussian`

Park, K. I., & Park, M. (2018). *Fundamentals of probability and stochastic processes with applications to communications*. Springer.

Pitard, F. F. (2019). *Theory of sampling and sampling practice*. Chapman and Hall/CRC.

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical reservoir modeling*. Oxford University Press, USA.

Raychaudhuri, S. (2008). Introduction to monte carlo simulation. In *2008 winter simulation conference* (pp. 91–100).

Resource_Modeling_Solutions_Ltd. (2022). *To impute or not to impute.* Retrieved 2022-02-01, from `https://resourcemodelingsolutions.com/to-impute-or-not-to-impute`

Reynolds, D. A., et al. (2009). Gaussian mixture models. *Encyclopedia of biometrics*, *741*(659-663).

Rice, J. A. (2006). *Mathematical statistics and data analysis*. Cengage Learning.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient.

American statistician, 59–66.

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

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Sarkar, S., Melnykov, V., & Zheng, R. (2020). Gaussian mixture modeling and model-based clustering under measurement inconsistency. *Advances in Data Analysis and Classification*, *14*, 379–413.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763–1768.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silva, D. S., & Deutsch, C. V. (2018). Multivariate data imputation using gaussian mixture models. *Spatial statistics*, *27*, 74–90.

Silva, D. S. F., & Deutsch, C. V. (2015). Program for fitting gaussian mixture models based on em algorithm and geostatistical applications. *CCG Annual Report 17*, 407.

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, *6*(1), 35–39.

Victor M. Silva, J. F. C. C., & Deutsch, C. V. (2019). A short note on the relationship between relative original units error and absolute normal score error. *CCG Annual Report 21*.

Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Wawruch, T., Deutsch, C., & McLennan, J. (2002). Geostatistical analysis of multiple data types that are not available at the same locations. *Centre For Computational Geostatistics*, *4*.

Yuan, Y. C. (2010). Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, *49*(1-11), 12.

Zhang, Y., Li, M., Wang, S., Dai, S., Luo, L., Zhu, E., … Zhou, H. (2021). Gaussian mixture model clustering with incomplete data. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *17*(1s), 1–14.

Appendix A

Appendices

Two geostatistical software library (GSLIB)-based programs have been developed to automatically obtain optimal DEMs. The first program `GETSECREAL` is applied to extract simulated primary values that exist at the location of secondary variables in each realization. The second program, `OPTDEM`, uses the results from `GETSECREAL` to provide a DEM for each realization. Both programs are implemented as standalone programs and follow the GSLIB conventions. This appendix provides the parameter files and code for both programs.

A.1 `GETSECREAL` Parameter

```
1                Parameters for GETSECREAL
2                 **********************
3
4   START OF PARAMETERS:
5   secondary.dat              -file with data
6   1  2  3  4                 -   columns for X,Y,Z,val
7   -1.0    1.0e21             -   trimming limits
8   backtr_sgsim.out           -file with realizations to extract
9   100                        -   number of realizations
10  256  0.0  1.0              -nx, xmn, xsiz
11  256  0.0  1.0              -ny, ymn, ysiz
12    1  0.0  0.5              -nz, zmn, zsiz
13  getsecreal.out             -file for output
```

From Line 1 to 3 of the parameter file can be ignored. Line 4 specifies the start of the parameter file, and it's crucial to start with the word "START" at the beginning of the line. Line 5 defines the input data file including the locations to extract from each realization. It would be a secondary data file in the DEM workflow. Line 6 specifies the XYZ coordinate column and values column for the input data. Line 7 sets the range of values in the input file and excludes data with values outside the range from the output values. Line 8 is a sequential Gaussian simulation GSLIB program output file. Line 9 indicates the number of realizations and Line 10 to 12 defines the simulation grid. Line 13 specifies the name of the output file that contains the XYZ coordinate and values of each realization.

A.2 OPTDEM Parameter

```
1                Parameters for OPTDEM
2                ********************
3
4   START OF PARAMETERS:
5   primary.dat                 -file with primary data
6   1 2 3 4                     -   columns for X,Y,Z,val
7   -1.0     1.0e21             -   trimming limits
8   secondary.dat               -file with secondary data
9   1 2 3 4                     -   columns for X,Y,Z,val
10  -1.0     1.0e21             -   trimming limits
11  getsecreal.out              -file with realizations at secondary
        data locations
12  1 2 3 4                     -   columns for X,Y,Z,val
13  100                         -number of realizations
14  optdem.out                  -file for output
15  3                           -number of radii for pairing
16  10.0   20.0   30.0          -   radii for pairing
17  1                           -Choose the number of radii to apply to
        the DEM
18  68516                       -random number seed
19  0   1.0                     -range for "a" parameter
20  0   0.1                     -range for "b" parameter
21  0   1.0                     -range for "c" parameter
22  0   0.1                     -range for "d" parameter
```

Line 5 identifies the primary data file. Line 6 sets the XYZ coordinate column and the value column of the primary data file. Lines 8 to 10 is for secondary data file and has the same meaning as above. Line 11 identifies the data file extracted from the simulation, which is the output file of GETSECREAL. Line 12 is for coordinate and value columns of GETSECREAL output file. Line 13 indicates the number of realizations. Line 14 should specify the name of the output file containing the parameters of the DEM, the differences in correlation coefficients, and the differences in mean differences from the pairing analysis results. The difference value should be calculated as the difference between the pairing analysis result of the primary variable and the secondary variable, and the pairing analysis result of the primary variable and DEM-applied data. Line 15 is the number of search radii specified. Perform pairing analysis of primary and secondary variables according to

the search radius entered in Line 16. Line 17 determines which radius to calculate DEM among the search radius presented in Line 16. Line 18 is the random number for generating random numbers in the code. Line 19 to 22 is for the range of DEM parameters.

A.3 DEM Workflow Chart

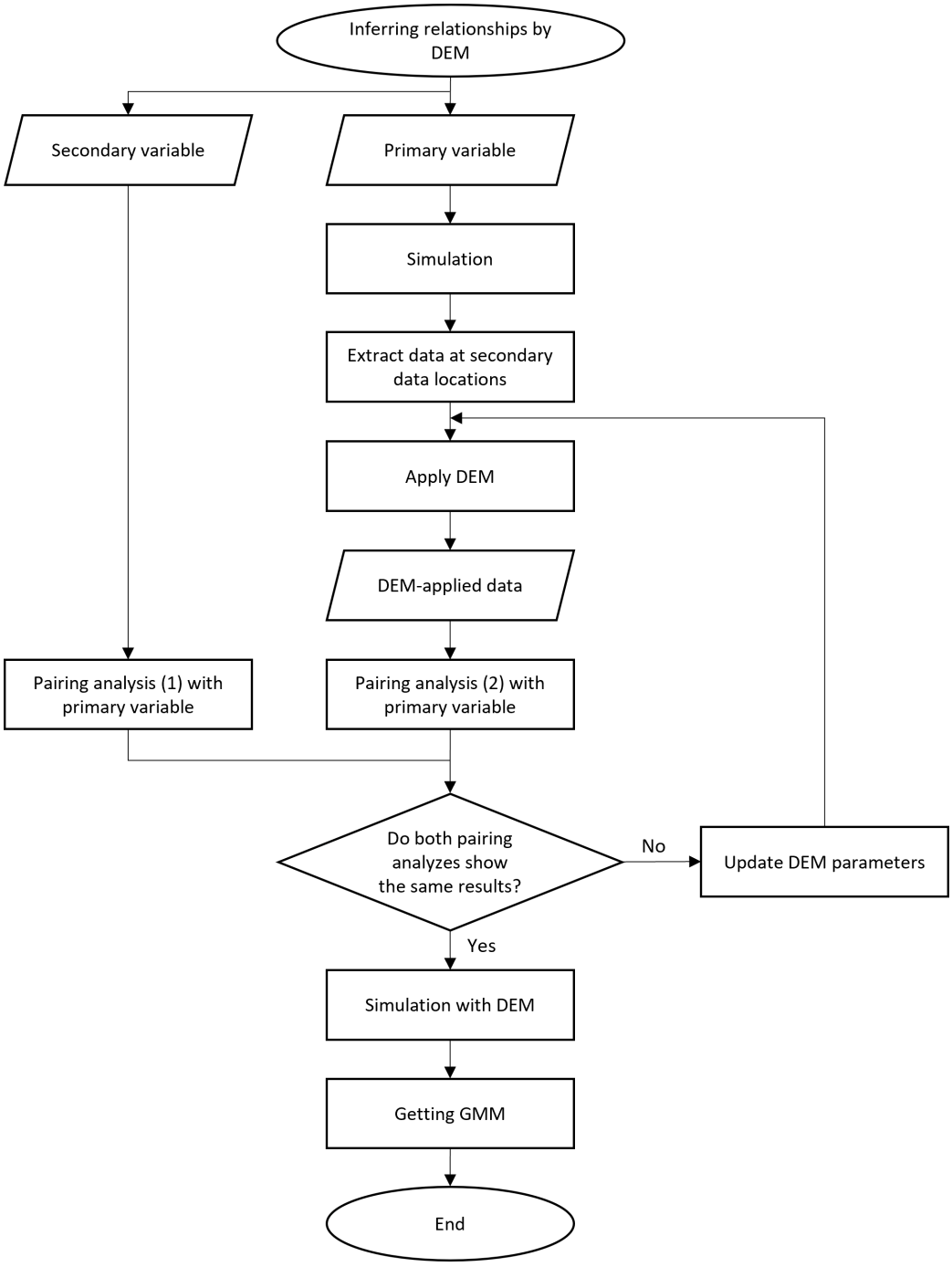


Figure A.1: DEM Workflow Chart

Data extraction at secondary data locations can be done via `GETSECREAL`. From pairing analysis using actual primary and secondary data to iterative DEM parameter updates, `OPTDEM` can be used.