Uncertainty Modeling of a Vein Deposit

by

Matthew Batty

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

This thesis sets out a workflow to quantify uncertainty of the Arrow Deposit, a tabular, vein type, high-grade, basement-hosted, uranium deposit located on NexGen Energy Limited's 100%-owned Rook I property in northern Saskatchewan. The uncertainty associated with the volume, grade, and density variables of the deposit are the focus of the study, as these variables define the overall metal content of the deposit, the largest input to project economics.

In the study domain, the grade variable is exhaustively sampled at all drill hole locations, but the density variable is missing at approximately 82% of the drill hole locations and biased to high-grade intercepts. Effort was taken to make an interim debiased homotopic representative dataset through a simple imputation process that selected density values from the global deposit-wide dataset. Two uncertainty models were created for the grade and density variables using the representative dataset: a multivariate spatial bootstrap model and a density-imputed model using a Gaussian Mixture Model built on the representative dataset, followed by decorrelation via Projection Pursuit Multivariate Transformation and independent grid Sequential Gaussian Simulation. The multivariate spatial bootstrap model provided an assessment of uncertainty of the histogram parameters while considering the spatial correlation between the variables at data locations. The imputation and simulation process features the transfer of the uncertainty associated with the missing values of the original dataset while accounting for the biased nature of the dataset.

Volume uncertainty was quantified by assessing: the uncertainty of the thickness perpendicular to the plane of continuity via a geometry imputation process, and the uncertainty of the boundary in the plane of continuity via an indicator estimate. The two assessments of domain uncertainty were combined to output volume uncertainty. The volume, grade, and density uncertainty models were combined into a single model; this uncertainty model will be used as the basis to evaluate mine output uncertainty with the mine design as the transfer function.

The metal content of the Arrow Deposit is most sensitive to changes in volume due to the extreme high-grade nature of the deposit, therefore the Boundary model, which has a large range of possible volumes (and range in the determination of ore/not ore), contributes the most to the metal content uncertainty. The Grade/Density model also significantly contributed to the metal content uncertainty, which is credited to the short-range variability of the variables. The thickness perpendicular to the plane of continuity model provided additional uncertainty to the metal content, but to a much lesser degree as the range of possible volumes associated with the model were relatively small.

Dedication

This thesis is dedicated to my supportive family: Alan, Patti, Amber, and Kyle Batty.

Acknowledgement

This thesis would not have been possible without the support of the Centre of Computational Geostatistics (CCG), my work was inspired by and, in part, incorporated the previous work, techniques, and publications developed by the group.

My utmost gratitude is extended to my supervisor, Dr. Jeff Boisvert for his invaluable guidance, patience, and motivation throughout the course of completing this thesis. Jeff continually showed support within and outside the scope of the program; he is wonderful example of modern professionalism and dedication. I would also like to thank my friends at the CCG for sharing their knowledge, camaraderie, and making the experience a fun and memorable one.

My sincere thanks to NexGen Energy Limited for funding my studies, providing employment, logistical support, and access to information. Special thanks to Troy Boisjoli for exposing me to the idea of pursuing my masters and providing the required support to complete it.

Table of Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

In this thesis, multivariate workflows with accompanying implementation details are demonstrated on a case study, the Arrow Deposit. The Arrow Deposit is a high-grade, narrow-vein, basement hosted uranium deposit that has undergone rapid development since discovery in 2014 through aggressive and systematic drilling programs. The Project will be developed through transverse and longitudinal longhole mining methods, pairing primary and secondary stoping techniques with a high degree of equipment mechanization.

The quantification of resource uncertainty at the Arrow Deposit will undoubtedly aid in decision-making as the project progresses through planning through to construction and extraction. This chapter describes the motivation for the thesis, the project, the study domain, and provides an outline to the thesis.

1.1 Motivation

Mining is a capital-intensive endeavour, and the viability of prospective mining operations relies on a sequence of engineering decisions. Understanding the uncertainty associated with the material to be mined can inform the engineering decisions. Geostatistical techniques, if applied correctly, can be used to quantify the uncertainty of the deposit. NexGen is currently approaching major engineering milestones, including detailed design; decisions are necessary to advance the project which can be informed with an uncertainty model.

The metal content at the Arrow Deposit is the largest input to project economics and is the focus of the enclosed uncertainty study. The metal content uncertainty is quantified through the assessment of uncertainties associated with density, grade, and volume. The main goal is to apply appropriate geostatistical techniques to quantify the uncertainty of key variables to provide an accurate and precise representation of the metal content uncertainty at the Arrow Deposit.

There are challenges to quantifying uncertainty for the Arrow Deposit. The highly correlated grade and density variables are heterotopically observed in the study domain dataset, with the grade variable being exhaustively collected at all sample locations, but the density values are missing at 82% of the sample locations and biased to high-grade intercepts. The dataset requires debiasing and the missing data should be accounted for to ensure a representative uncertainty model is created.

Assessing the volume uncertainty of the Arrow Deposit is important for understanding metal content uncertainty, as the metal content is sensitive to changes in volume due to the extreme high-grade nature of the deposit. The volume uncertainty can be assessed by determining the uncertainty of the domain position and thickness perpendicular to the plane of continuity, and the uncertainty of the boundary in the plane of continuity. Appropriate techniques are needed to assess these uncertainties. One challenge is that the vertical nature of the deposit facilitated delineation drilling from either side of the deposit but did not allow for the assessment of true domain thickness due to the drill holes intersecting the mineralization at oblique angles. The inclined drill holes provide known coordinate points defining the hanging wall (HW) and footwall (FW) limits of each domain, but FW and HW position uncertainty exists distal to the known points. The location of hangingwall and footwall position is important to global metal content uncertainty, but also to the positioning of the individual stopes. Optimizing each stope position will undoubtedly improve the economics of the mine.

The objective of the thesis is to develop a multivariate workflow with appropriate geostatistical techniques that can be used to create an uncertainty model for the deposit. The workflow is tested on a study domain to ensure data is reproduced.

1.2 Project Overview

NexGen Energy Limited's (NexGen) 100%-owned Rook I Project (Project) is situated in the Athabasca Basin, the highest-grade uranium district in the world, and hosts the high-grade Arrow Deposit, the largest undeveloped uranium resource in the basin. The Project is currently progressing through advanced engineering and permitting with a proposed new uranium mining and milling operation, including underground and surface facilities to support the extraction and processing of uranium ore from the Arrow Deposit. The Project is located on the southwest edge of the Athabasca Basin, approximately 40 km east of the Saskatchewan-Alberta border, 130 km north of the town of La Loche, and 640 km northwest of the city of Saskatoon.

The following description of the local geology to the Arrow Deposit has been summarized from Hillacre et al. (2020). The Arrow Deposit is hosted in the Paleoproterozoic basement rocks of the Taltson Domain (Figure 1.1). The bedrock geology is comprised of variably silicified and metasomatized intermediate to mafic orthogneisses. Local mafic-rich amphibolite and pyroxenite, ultrabasic and syenitic dykes, and porphyroblastic feldspar- and quartz-rich pegmatites intrude the gneissic granulite facies rocks. The main fabrics and contacts of crystalline basement rocks in the Arrow Deposit area are all steeply dipping, dominantly southeast, with a northeast southwest strike. Basement rocks are unconformably overlain by late Paleoproterozoic to Mesoproterozoic Athabasca Supergroup sandstones of variable thickness, rarely

exceeding 50 m. Devonian and/or Cretaceous sedimentary rocks overlie the Athabasca sandstones, with Quaternary glacial deposits capping the geologic sequence and forming the present-day topography.

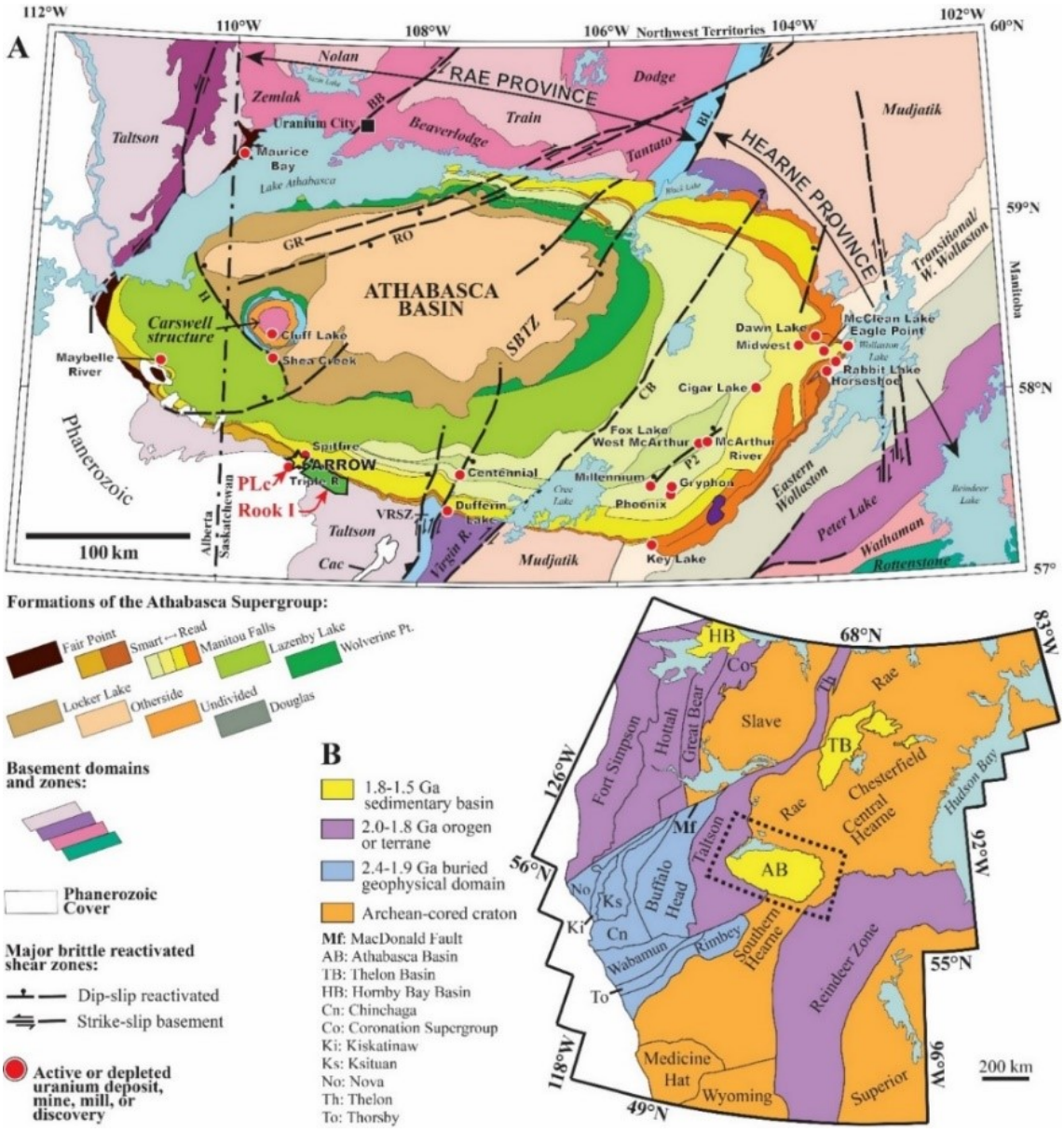

Figure 1.1: (A) Lithostructural domains of northern Saskatchewan and Alberta illustrating the major components of the Churchill Structural Province of the Precambrian Shield and overlying Phanerozoic cover. (B) Cratonic map of western Laurentia showing Figure 5A (dashed box) in context of continent-scale tectonics (Hillacre et al., 2020).

After discovery of the Arrow Deposit in February 2014, mineral resource estimates were completed in 2016, 2017, 2018, and 2021, each supported by successive systematic diamond drill programs. The latest mineral resource estimate was publicly released on February 22nd, 2021, coincident with the release of the Rook I Project Feasibility Study (FS). The deposit is currently interpreted to consist of several high-

grade, near-vertical, uranium veins within at least six reactivated high-strain zones, known as the A0 through A5 Shears (Figure 1.2). The heterogeneous high strain zones hosting the Arrow Deposit evolved through episodic reactivation events creating various small-scale brittle fault linkages oblique to, and connecting, the main fault zones (Hillacre et al., 2020). The main uranium-bearing mineral present at the Arrow Deposit is uraninite, whereas secondary uranium minerals such as coffinite or uranophane may partially or wholly replace uraninite (Hillacre et al., 2020). The mineralized area is 315 m wide with an overall strike of 980 m. Mineralization occurs 100 m below surface and extends to a depth of 950 m. The individual shear zones vary in thickness from 2 m to 60 m. The mineralized shear zones defining the Arrow Deposit comprise part of the Patterson Lake structural corridor, which is host to several other uranium deposits or showings, such as: Fission Uranium's Triple R deposit; NexGen's South Arrow Discovery, Harpoon Discovery, and Bow Occurrence; and Purepoint's Spitfire Deposit (Figure 1.3).

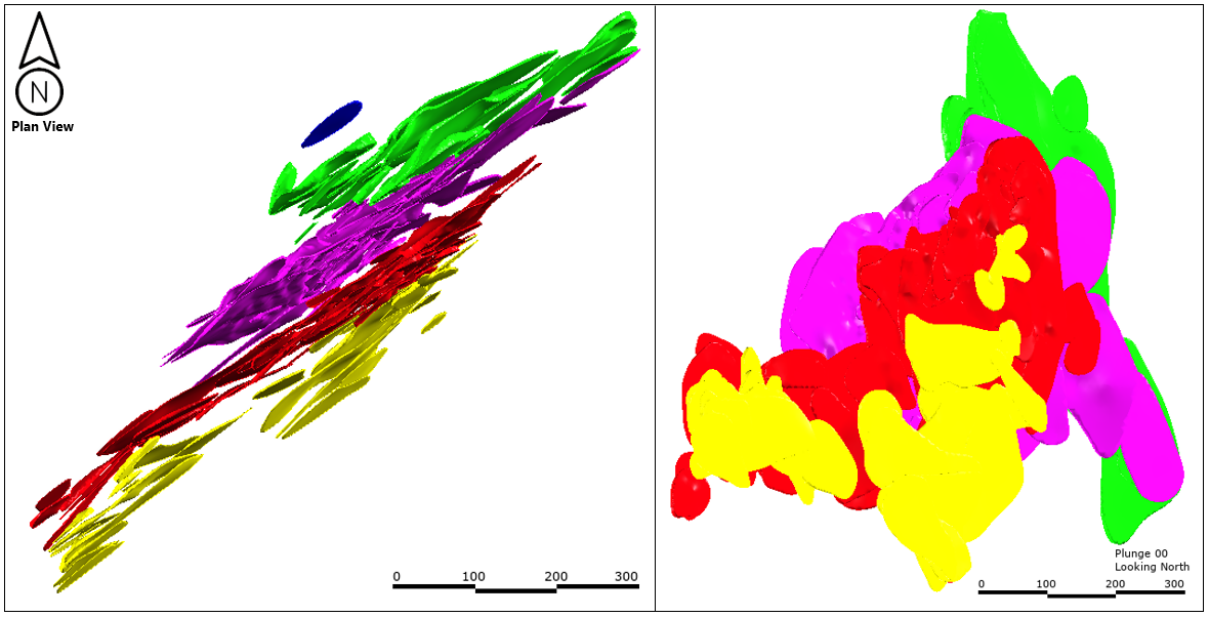

Figure 1.2: Plan view and oblique view of the current mineral resource wireframes. The blue, green, purple, red, and yellow wireframes denote the mineralization in the A0, A1, A2, A3, and A4/A5 Shears, respectively. Scale in metres.

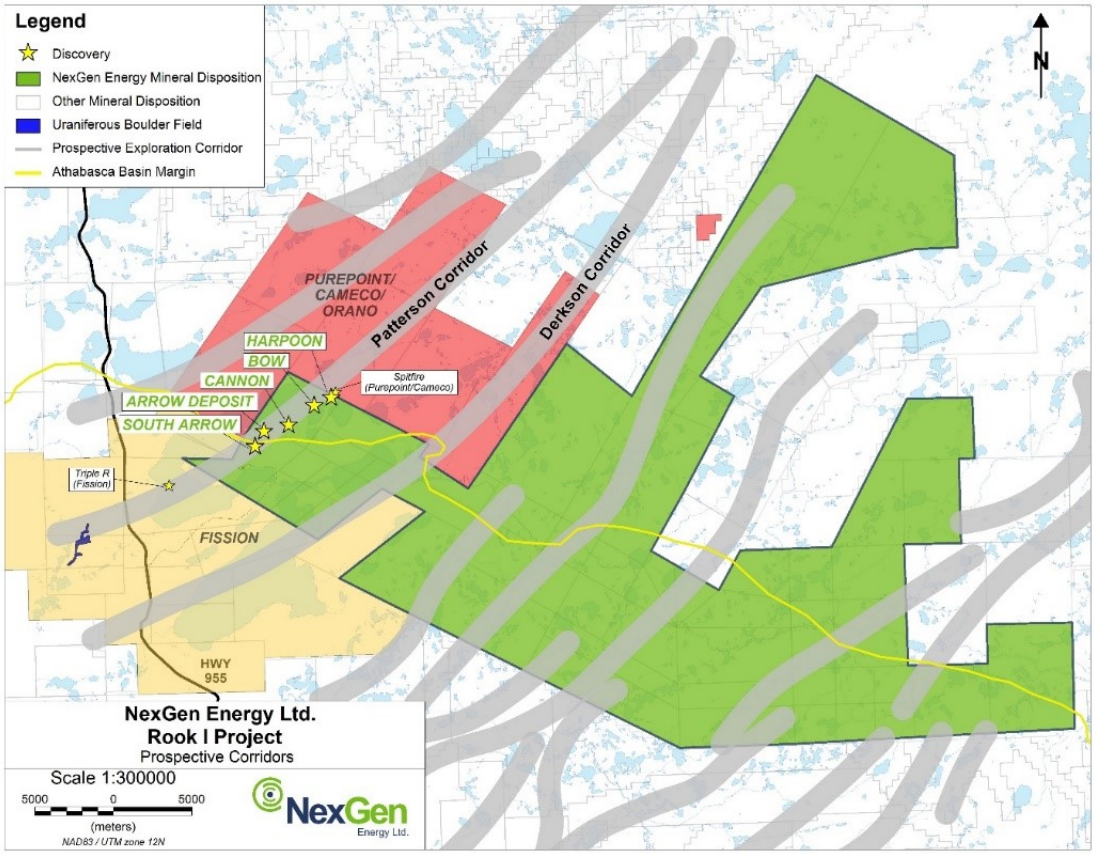

Figure 1.3: Plan view of the southwest Athabasca Bain showing various structural corridors, relevant mineral dispositions, and uranium deposits and/or showings.

The latest mineral resource estimate states that the Arrow Deposit has measured mineral resources of 209.6 million (M) pounds (lbs) of triuranium octoxide (U_3O_8) contained in 2,183 kilotonnes (kt) grading 4.35% U_3O_8, indicated mineral resources of 47.1 M lbs of U_3O_8 contained in 1,572 kt grading 1.36% U_3O_8, and inferred mineral resources of 80.7 M lbs of U_3O_8 contained in 4,399 kt grading 0.83% U_3O_8 (Hatton et al., 2021). The measured and indicated mineral resources are supported by geostatistical drill hole spacing studies. In 2017, Clayton V. Deutsch Consultants Incorporated completed a geostatistical drill hole spacing study, evaluating the uncertainty of thickness, grade, density, and areal limit of the Arrow Deposit to determine a drill hole spacing suitable for indicated mineral resources. Indicated spacing was defined to be 24 to 32 metres ("m") to be within 15% of predicted for a nominal quarterly production volume with a probabilistic tolerance between 75 and 90% (Figure 1.4; Wilde & Deutsch, 2017). A subsequent geostatistical drill hole spacing study on the Arrow Deposit was completed by Resource Modeling Solutions in December 2018 to define measured mineral resource spacing in support of the 2019 drill campaign. The study resulted in a drill hole spacing of 9.00 to 16.75 m to be within 15% of predicted for monthly and quarterly production volumes using a probabilistic tolerance of 90% (Figure

1.5; Barnett & Deutsch, 2018). The measured and indicated mineral resources are limited to the A2 and A3 Shears and were used as the basis for the FS mine design (Figures 1.6).

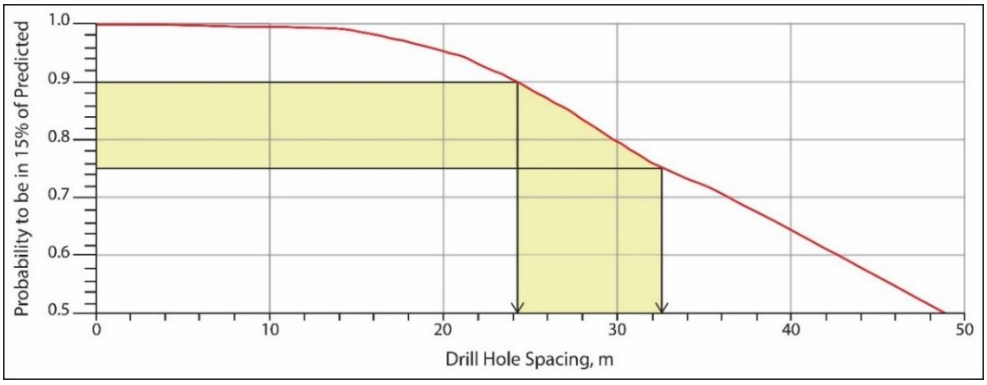

Figure 1.4: Graph showing the drill hole spacing of 24 to 32 m to have a probabilistic tolerance between 75 and 90% to be within 15% of predicted for a nominal quarterly production volume (Wilde & Deutsch, 2017).

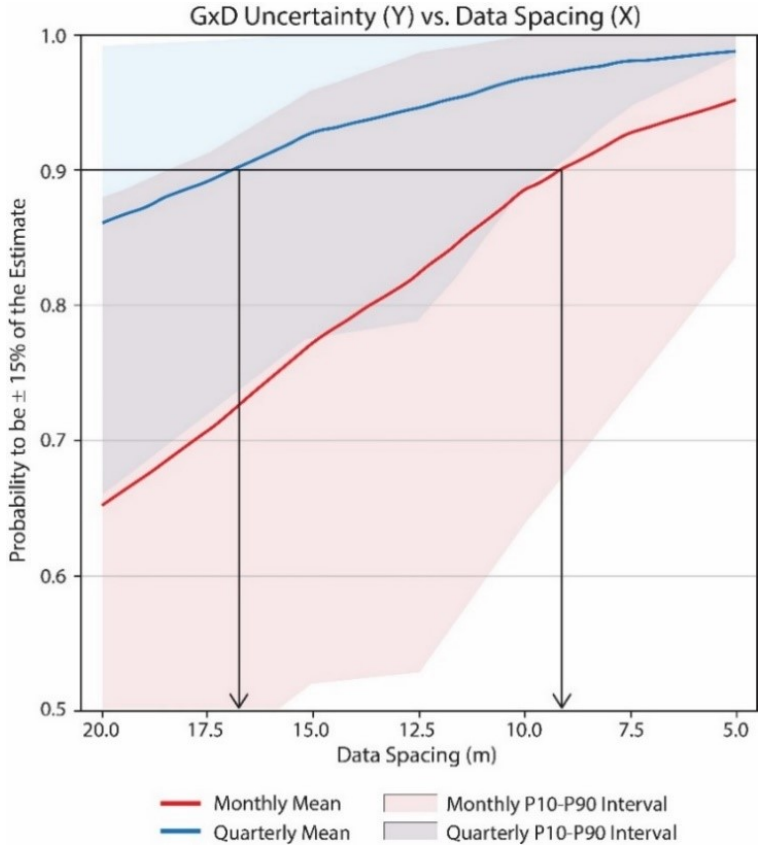

Figure 1.5: Graph showing a drill hole spacing of 9.00 to 16.75 m to be within 15% of predicted for monthly and quarterly production volumes (as per the Pre-Feasibility Study) at a probabilistic tolerance of 90% (Barnett & Deutsch, 2018).

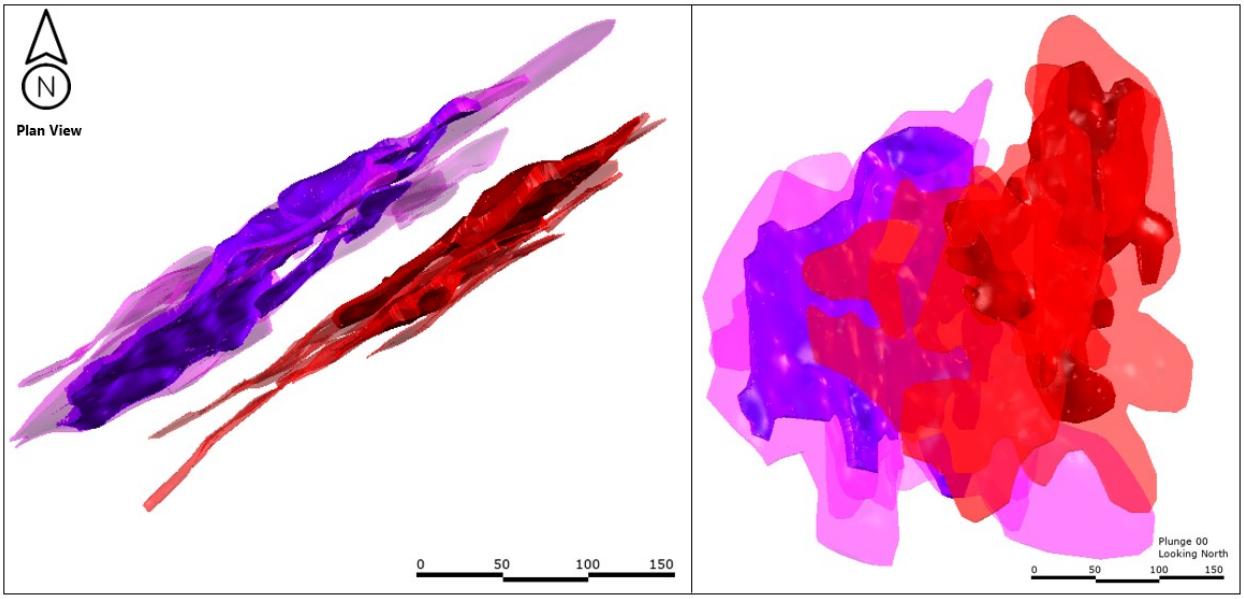

Figure 1.6: Plan view and oblique view (looking north) of the currently defined measured and indicated mineral resources. The purple and red denote the A2 and A3 Shear Zones, the transparent portions of the wireframes signify indicated mineral resources, and the opaque portions demarcate measured mineral resources. Scale in metres.

The mine design and milling process described below are summarized from the FS technical NI 43-101 report (Hatton et al., 2021). The FS proposes two longhole mining methods, transverse and longitudinal stoping, to extract ore from the deposit (Figure 1.7). The stopes are to have a 30 m level spacing and a nominal stope strike length of 15 metres to 30 metres. The mine design predicts a 10.7-year mine life with an average daily throughput of 1,300 tonnes per day and average annual production of 21.7 M lb U_3O_8. The deposit is to be accessed by two shafts, both located in the footwall of the deposit (Figure 1.7). One shaft will be used as a production shaft, and for transportation of personnel and materials into the mine. The production shaft will have divided compartments so that fresh air that encounters ore being skipped to surface will be immediately exhausted within the mine preventing radiation exposure to any personnel. The production shaft will also have a permanent headframe and hoisting house (Figure 1.8). The second shaft will be used as an exhaust ventilation shaft. The ore will be crushed underground and transported to the grinding mill on surface. The ore, after being reduced to a suitable size, will be transported to the leaching circuit to oxidize, and dissolve the uranium that is present in the ore solids. A counter current decantation circuit will wash the uranium-bearing solution from the barren leached residue. The residue is routed towards disposal while the pregnant aqueous solution reports to the solvent extraction circuit for uranium purification. The purified uranium is then precipitated as yellowcake through the addition of hydrogen peroxide. The yellowcake will be dried in a calciner and packaged in 200 litre steel drums. Tailings generated from the mill processes will be used as cemented paste backfill

of stopes, as well as stored in purpose-built Underground Tailing Management Facility (UGTMF) comprised of several dedicated chambers (Figure 1.7).

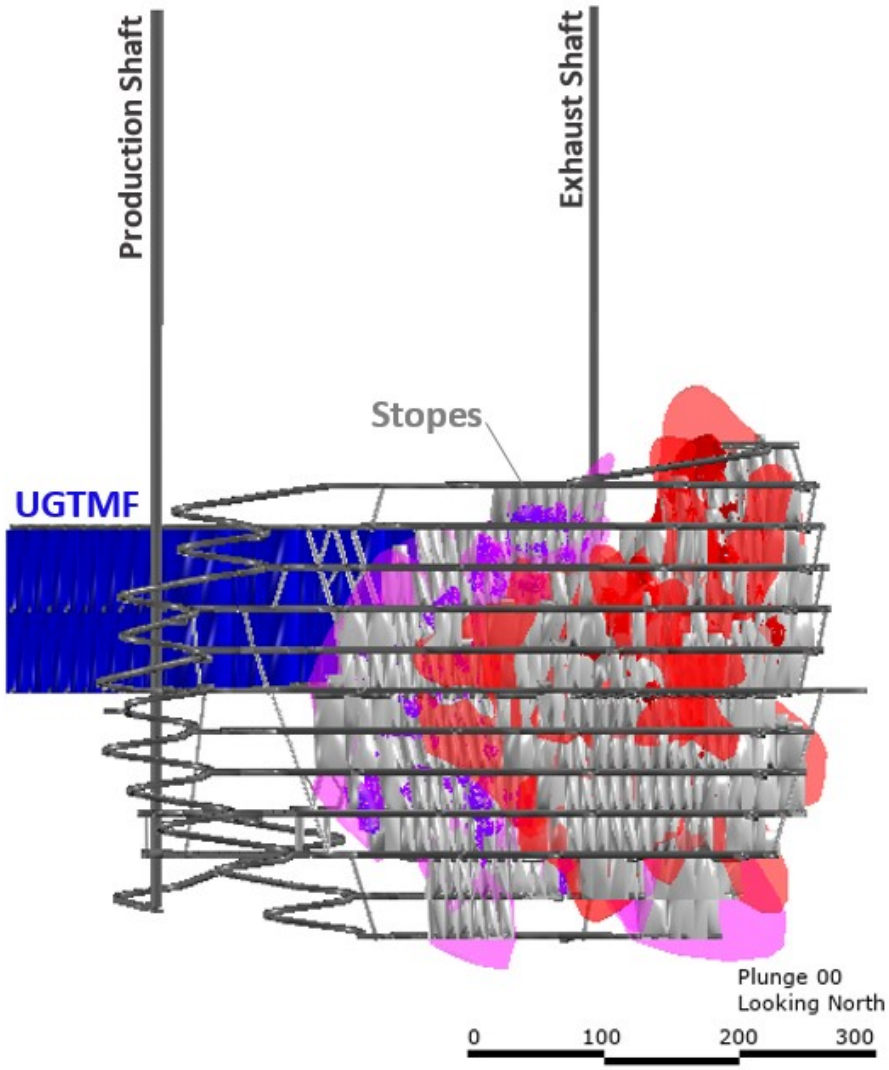

Figure 1.7: Oblique view of the currently defined measured and indicated mineral resources and the FS mine design. The purple and red denote the A2 and A3 Shear Zones, the transparent portions of the wireframes signify indicated mineral resources, and the opaque portions demarcate measured mineral resources. The grey wireframes indicate underground development and stopes, and the blue wireframe represent the UGTMF.

Figure 1.8: Oblique view (looking south) of the proposed surface infrastructure as per the 2021 Feasibility Study (Hatton et al., 2021).

This thesis considers publicly released data available at the initiation of the study, encompassing drill data from 2014 to May 25th, 2018, and mineral resource wireframes released in November 2018, coincident with the release of the Pre-Feasibility Study. The November 2018 mineral resource consists of indicated mineral resources of 256.6 M lbs contained in 2,890 kt grading 4.03% U_3O_8 and inferred mineral resources of 91.7 M lbs contained in 4,840 kt grading 0.86% U_3O_8 (O'Hara et al., 2018). The uranium grade values (% U_3O_8) were converted to gold grade values (grams per tonne) using an undisclosed method to avoid unintentionally disclosing sensitive information.

1.3 Study Domain

A high-grade domain within the A2 Shear of the Arrow Deposit is the subject of this study. The domain is created using a grade intercept limit equal to or greater than one metre with a minimum grade of 5% U_3O_8, although lower grades were incorporated in places to maintain continuity and a reasonable thickness. The drill holes delineating the vertical, northeast striking domain are oriented in two principal directions: 327- and 147-degrees azimuth, dipping between 60 and 80 degrees; however, there is one drill hole drilled vertically through the domain.

The domain is entirely within the defined PFS indicated mineral resources and contains 35 drill holes spaced 0.5 to 25 m apart with a total of 656 composites. The assay grade and density data were composited to 0.5 m lengths within the domain, and composites less than 0.25 m were merged with the

previous composite. Declustering weights for the data set are calculated using the cell declustering method with a 25m (strike) by 25 m (vertical) cell, which is approximately equal to the spacing in the sparsely sampled regions. The data and domain were rotated to a north-south orientation to reduce the grid size, subsequently reducing the computational demand of running the model (Figure 1.9).

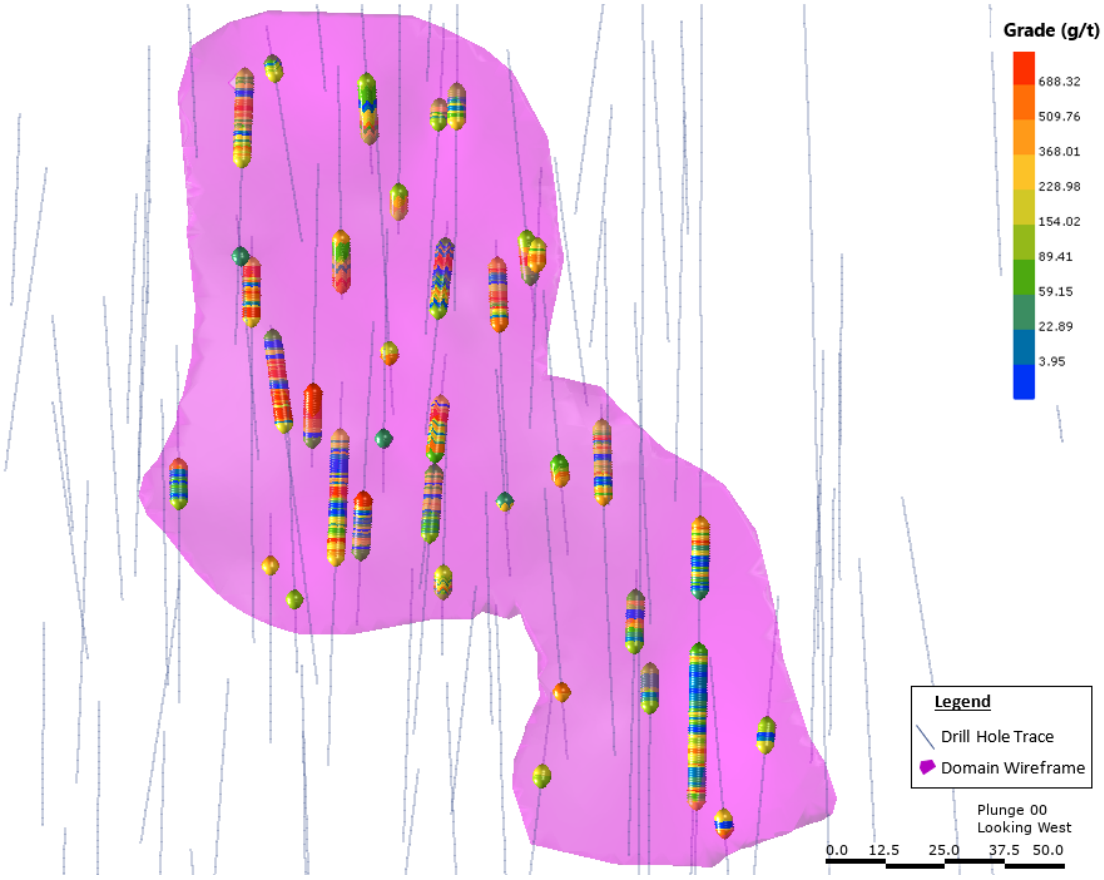

Figure 1.9: Long section (20 m window looking west) of the drill holes constraining the domain, the domain wireframe, and domain composites displaying assay grade results. The composites within the domain are shown as enlarged spheres and the drill traces are lines (exludes assay grades). There are no barren intervals within the domain. Scale in metres.

The grade and density variables are strongly and positively correlated with a correlation coefficient of 0.91 (Figure 1.10). The grade variable is sampled at all 656 composite locations, but the density variable is incomplete with only 116 of the composite locations having density values (Figure 1.11). The locations without density values have an average grade of 226.18 grams per tonne (g/t) with a positively skewed grade distribution, whereas the locations with density values have an average grade of 401.75 g/t with a nearly uniform grade distribution (Figure 1.12). The difference in the two sample grade subpopulations

is reflective of preferential collection of density data in high grade samples, representing a bias in the dataset. Therefore, the bias in the heterotopic dataset needs to be addressed to get a representative dataset prior to any uncertainty assessment studies. The bias is likely due to the friable nature of lower-grade samples, which did not allow for reliable density measurements, and, to a lesser degree, the discretion of the logging geologist.

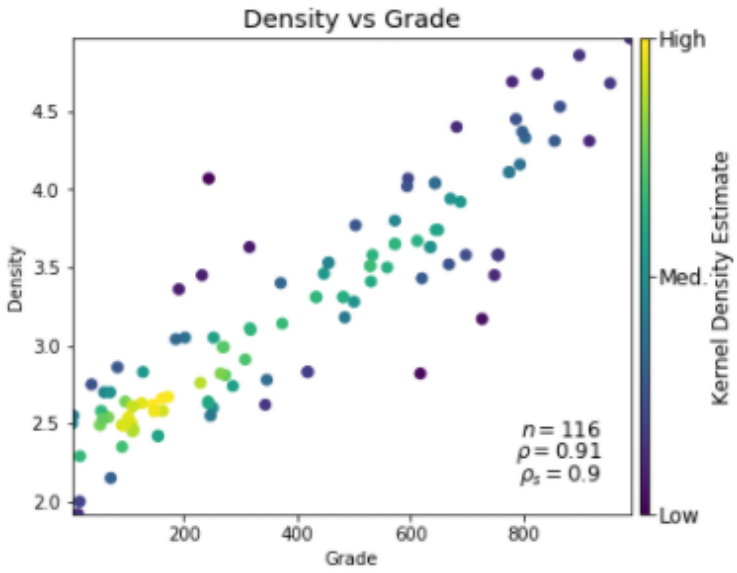

Figure 1.10: Scatter plot of the grade (Au g/t) and density variables of the A2 high-grade domain.

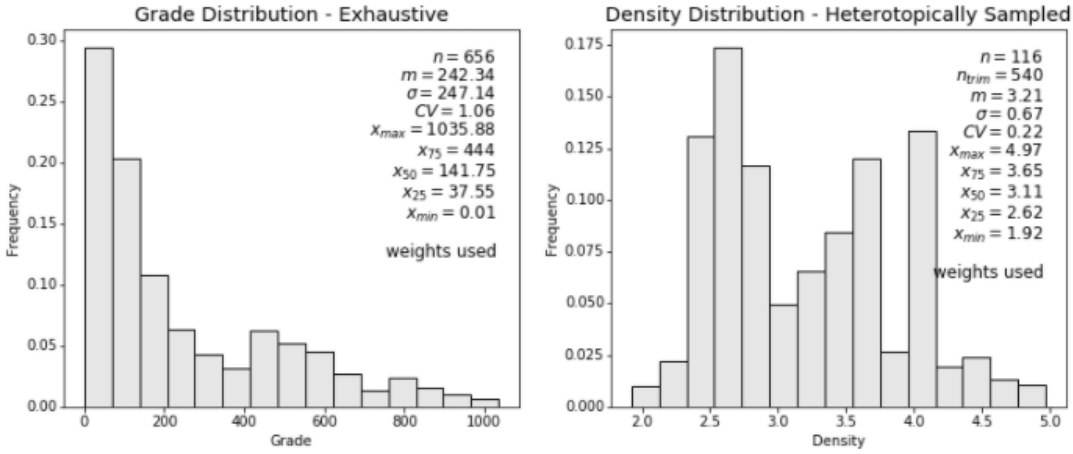

Figure 1.11: Histogram distributions of the grade (Au g/t) and density (S.G.) variables of the A2 high-grade domain. Cell declustering weights were applied to both distributions.

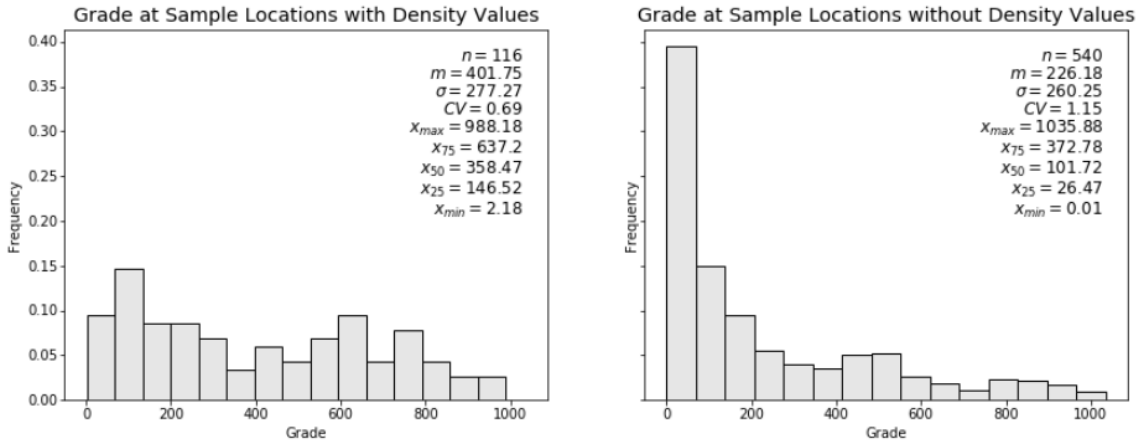

Figure 1.12: Histogram distributions of the grade variable (Au g/t) of the A2 high-grade domain at locations with and without density values.

1.4 Thesis Outline

The thesis sets out a workflow to quantify uncertainty of the Arrow Deposit, testing and using various techniques. Chapter 2 is a background review of the available literature pertinent to the study. Chapter 3 details two uncertainty models of the grade and density variables: a multivariate spatial bootstrap model and a density-imputed model using a Gaussian Mixture Model followed by decorrelation via Projection Pursuit Multivariate Transformation and independent grid Sequential Gaussian Simulation. Chapter 4 is an assessment of thickness perpendicular to the plane of continuity via a geometry imputation process. Chapter 5 describes the process to assess uncertainty in the plane of continuity. The geometry imputation, boundary, grade, and density models were combined into a single model and each individual component was reviewed to evaluate influence on overall uncertainty in Chapter 6. Figure 1.13 is a schematic of the workflow completed in this thesis.

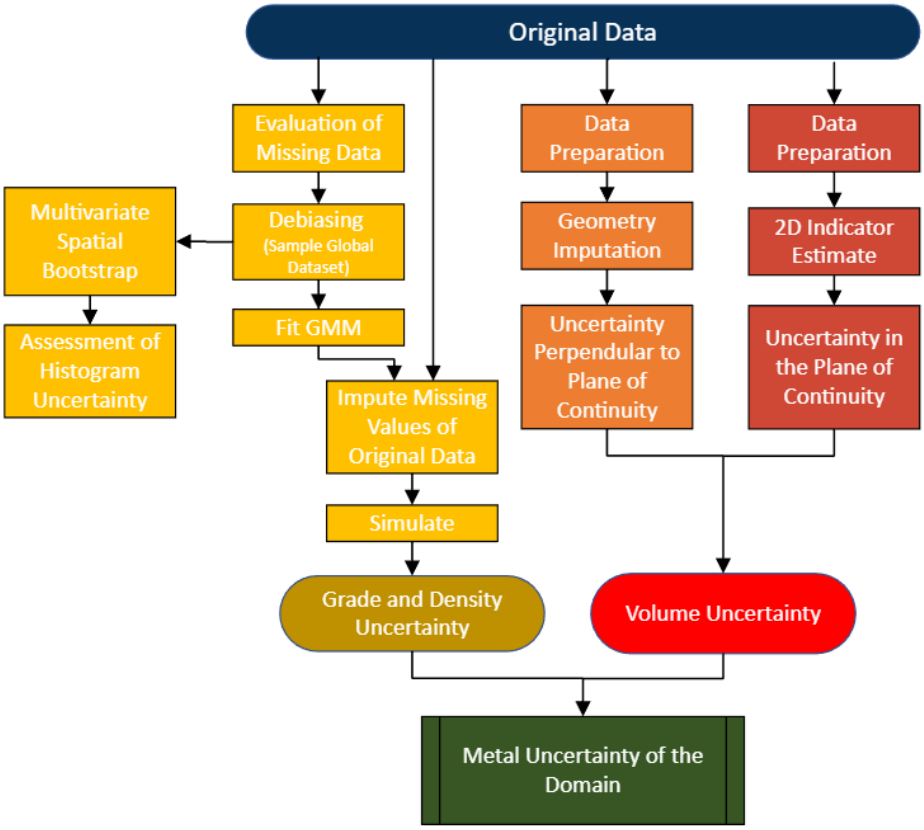

Figure 1.13: Schematic of the thesis workflow

Chapter 2

2. Background Review

Chapter 2 briefly outlines the literature describing the geostatistical techniques and concepts applied to the Arrow Deposit; inclusive to missing data concepts, multivariate data imputation techniques, projection pursuit multivariate transformation, spatial multivariate bootstrap, geometry imputation of vein deposits, and approaches to boundary uncertainty assessment.

The density variable at the Arrow Deposit is not exhaustive for sample locations, therefore understanding and accounting for the mechanism of the missing data will lead to an estimation of the statistical parameters that are more representative of the sampled populations. The missing data is to be imputed in a simulation workflow to transfer the uncertainty associated with the missing values into the model. After imputation, the dataset will be homotopic but may contain multivariate complexities that require a transformation to achieve multivariate Gaussianity, which can be achieved through the projection pursuit multivariate transform. The spatial multivariate bootstrap is used to measure global parameter uncertainty while considering to spatial correlation between the variables at data locations. The metal content of the Arrow Deposit is sensitive to changes in volume due to the extreme high-grade nature of the deposit, therefore understanding the uncertainty of volume and geometry of the domains is important.

2.1 Missing Geological Data

Rubin (1976) proposed that it is possible to relate missing and measured values through three distinct mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Many mining projects unevenly collect data values for various geologic variables resulting in heterotopic datasets that contain missing data; thus, the datasets can be classified as MCAR, MAR, or MNAR.

Consider a random vector of multiple correlated attributes of N geological variables, $\boldsymbol{\Phi}(Z_1, Z_2, \dots, Z_N)$. The MCAR mechanism occurs when the probability of missing value of Z_1 does not depend on the any of the $N-1$ variables observed nor the values of Z_1, the MAR mechanism occurs when the probability of missing value of Z_1 depends on some of the $N-1$ variables observed but not the value Z_1, and the MNAR mechanism occurs when the probability of Z_1 depends on the both some of the $N-1$ variables observed and the value Z_1 itself (Rubin, 1976). Appropriately accounting for the missing data mechanisms has been demonstrated to lead to an estimation of the statistical parameters that are more representative of the sampled populations.

The MCAR mechanism facilitates a complete case analysis where no additional controls are needed to ensure the dataset is representative. However, the MCAR mechanism is not observed in most practical cases, as the incomplete variable is unknowingly biasedly sampled, sometimes due to cost and/or physical restraints. Therefore, identification of the reasons for missing data is critical for a representative predictive model. The MAR mechanism can be accounted for through maximum likelihood estimation and multiple imputation approaches (Rubin, 1988; Schafer and Graham, 2002; Enders, 2010; Silva and Costa, 2019). The MNAR mechanism is more complex and difficult to account for, but nonetheless, the missing data carry important information about the data population that must not be ignored (Silva and Costa, 2019).

Rubin (1987) proposed that the MNAR mechanism be accounted for by using an appropriate approach for missing MAR and using a fixed transformation applied to imputed MNAR values. The fixed transform uses a constant value to compensate for an over or under estimation of the imputed value. Cohen (1988) furthered the studies of Rubin, recommending that the constant value for transform be proportional to one-half of the standard deviation from an imputed scenario, therefore having a direct impact on the mean of the data set (Enders, 2010). More recently, Silva and Costa (2019) account for the MNAR mechanism by inferring the values of an under-sampled variable over the study area through a co-simulation approach and a univariate simulation performed on an artificially complete set. The workflow consists of seven primary steps:

1. Randomly remove 20% of the samples from the incomplete variables in the original data set, $\boldsymbol{\Phi}_o$, to create a test set, $\boldsymbol{\Phi}_t$.
2. Independent spatial continuity modelling for all variables in the original data set, $\boldsymbol{\Phi}_o$.
3. Apply Bayesian Updating (BU) to the test data set, $\boldsymbol{\Phi}_t$, using the spatial continuity model fitted to the original data set, $\boldsymbol{\Phi}_o$. Generate N imputed scenarios for the test set, $\boldsymbol{\Phi}_t$.
4. Calculate the maximum error for each N imputed scenarios for the test set, $\boldsymbol{\Phi}_t$, via a cross-validation process.
5. Apply BU to the original data set, $\boldsymbol{\Phi}_o$, using spatial continuity models fitted to the original data set, generating N imputed scenarios.
6. Apply the fixed transform proposed to every imputed value in each scenario generated.
7. Use the e-type imputed scenarios in the geostatistical analysis.

The maximum imputation error calculated in the cross-validation process noted in step 4 is parameterized by the equation below.

$$Err_{ij} = \frac{z_j^*(\boldsymbol{u}_i) - z(\boldsymbol{u}_i)}{z(\boldsymbol{u}_i)} \quad (2.1)$$

where Err_{ij} is the error for *j* scenario at *i* location, $z_j^*(\boldsymbol{u}_i)$ is the imputed value at location *i* with a missing sample, and $z(\boldsymbol{u}_i)$ is the measured sample at the *i* location.

The fixed transform in step 6 corrects the imputed values to be proportional to the maximum error obtained, as per the equation below.

$$z_j^T(\boldsymbol{u}_i) = z_j^*(\boldsymbol{u}_i) + z_j^*(\boldsymbol{u}_i) \cdot Err_j^{Max} \quad (2.2)$$

Err_j^{Max} is the maximum value of Err_{ij} obtained in scenario *j* and $z_j^T(\boldsymbol{u}_i)$ is the updated value after the fixed transform has been applied to the *j* BU scenario.

The workflow proposed by Silva and Costa accounts for the MNAR mechanism where data imputation is not sufficient.

2.2 Multivariate Data Imputation

Multivariate data transformations are required in many geostatistical workflows, and many of the transformations, such as Projection Pursuit Multivariate Transformation (PPMT), Minimum/Maximum Auto-Correlation Factors (MAF) and Principal Component Analysis (PCA), can only be applied to homotopic data. However, heterotopic data is commonly encountered in practice, and if a multivariate transformation is required in a workflow, data will need to be imputed or discarded to make a homotopic dataset. Discarding heterotopic data observations can lead to biased models, as missing geologic data observations are typically not completely random (Little & Rubin, 2014). Therefore, imputation is generally the chosen method to create a homotopic dataset, and multiple imputation (MI) techniques are preferred as the techniques capture the uncertainty of imputed values (Barnett & Deutsch, 2015b).

Barnett and Deutsch (2015b) proposed two MI techniques for geologic data using Bayesian Updating (BU): a fully parametric method that assumes a multivariate Gaussian distribution and a non-parametric method that uses univariate kernel density estimation (KDE) and Gibbs sampler to calculate likelihood distributions for BU. The assumption of a multivariate Gaussian distribution in the first method makes it difficult to reproduce complexities observed in the data. The non-parametric approach improved upon the first method by consistently reproducing the complexities observed in the original data set. Silva and Deutsch (2015a) continued the MI work in geostatistics and proposed another non-parametric workflow that fits a Gaussian Mixture Model (GMM) to the multivariate data, then uses the GMM to calculate

conditional distributions that defines the likelihood for the Bayesian Updating (BU). The GMM workflow is computationally faster and more accurate than the KDE with Gibbs sampler workflow, as it allows for the quick assessment of any marginal and conditional distribution needed without the use of Gibbs sampler (Silva, 2018). The GMM workflow of Silva (2018) is summarized below.

Consider n observations $z = (z_1^T, \ldots, z_n^T)$ containing K-dimensional homotopic and heterotopic data observations $z_i = (z_{i,1}, \ldots, z_{i,K})$. The GMM imputation process consists of six primary steps.

1. The independent normal score transformation of each variable, resulting in a set of n data observations $y = (y_1^T, \ldots, y_n^T)$. Performing the transformation independently ensures the transformation is not sensitive to heterotopic data observations.

$$y_{i,k} = G^{-1}\left(F\left(z_{i,k}\right)\right), k = 1, \ldots, K \text{ and } i = 1, \ldots, n \tag{2.3}$$

2. The expectation maximization (EM) algorithm is then used to fit a GMM with g components to the transformed data defining the estimated multivariate density function, as per below.

$$\hat{f}(\boldsymbol{y}_j; \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) \tag{2.4}$$

Where $\hat{f}$ is the estimated distribution, ϕ is the multivariate Gaussian probability density function (PDF), and $\boldsymbol{\Psi}$ is the set of unknown parameters $(\pi_1, \ldots, \pi_{g-1}, \mu_1, \ldots, \mu_g, \Sigma_1, \ldots, \Sigma_g)$: weights to each component, the mean values, and the variance-covariance matrices. The EM algorithm maximizes the log likelihood, $\log L(\boldsymbol{\Psi})$, which can be calculated for dataset as:

$$\log L(\boldsymbol{\Psi}) = \sum_{j=1}^{n} log \left\{ \sum_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_j; \boldsymbol{\mu}_i; \boldsymbol{\Sigma}_i) \right\} \tag{2.5}$$

The EM algorithm consists of two steps, which are the expectation (E) and maximization (M) steps that iteratively maximize the log likelihood, $\log L(\boldsymbol{\Psi})$. Every observation is from a component of the mixture, but this information is unknown, hence, the EM algorithm is devised as an incomplete-data problem where the g-dimensional label vectors x_1, ..., x_n indicate whether an observation came from a certain kernel or not. The component $x_{i,j} = (x_j)_i$ is equal to 1 if the j^{th} observation belongs to the i^{th} component of the mixture and 0 otherwise. 1. Note that only a brief description of EM is provided here as it is well established. Interested readers are referred to McLachlan and Krishnan (2008) and McLachlan and Peel (2000) for a more detailed explanation of the algorithm.

3. The prior distribution given the spatial data of the variable to be imputed needs to be defined for each imputation location. After the independent normal score transformation, each variable is assumed to be spatially multi-Gaussian, thus the prior distribution used in BU is Gaussian as

well. Therefore, the prior distribution can be parameterized by the kriging mean and variance, which are obtained by solving the simple kriging system of equations using nearby observations of the same variable being imputed.

Consider the sub-vector y_p for a given variable k to be imputed in data matrix y that contains all n observed and previously simulated values for the same variable. The prior mean, $\bar{\mu}_p$, and variance, $\sigma_p^2(\boldsymbol{u})$, is calculated with the equations below, where $\boldsymbol{u}$ is the vector of spatial coordinates and C is the spatial covariance.

$$\bar{\mu}_p(\boldsymbol{u}) = \sum_{i=1}^{n} \lambda_i y_p(\boldsymbol{u}_i) \tag{2.6}$$

$$\sigma_p^2(\boldsymbol{u}) = 1 - \sum_{i=1}^{n} \lambda_i C(\boldsymbol{u}, \boldsymbol{u}_i) \tag{2.7}$$

$$\sum_{j=1}^{n} \lambda_i C(\boldsymbol{u}_i, \boldsymbol{u}_j) = C(\boldsymbol{u}, \boldsymbol{u}_i) i = 1, \dots, n \tag{2.8}$$

4. A likelihood distribution of the variable to be imputed needs to be obtained for the BU. The likelihood distribution is the marginal distribution of the conditional distribution given the collocated data observations respective to the variable being imputed. The conditional distribution can be extracted from the fitted GMM, where the conditional distribution of all g multivariate Gaussian components is parameterized by their conditional mean and covariance. The conditional mean and covariance are calculated by:

$$\bar{\boldsymbol{\mu}}_m^i = \boldsymbol{\mu}_m^i + \Sigma_{mo}^i \Sigma_{oo}^{i-1} (\boldsymbol{y}_o - \boldsymbol{\mu}_o^i) \quad i = 1, \dots, g \tag{2.9}$$

$$\bar{\Sigma}_{mm}^i = \Sigma_{mm}^i - \Sigma_{mo}^i \Sigma_{oo}^{i-1} \Sigma_{om}^i \qquad i = 1, \dots, g \tag{2.10}$$

Σ_{mo}^i and Σ_{oo}^i are the sub-matrices of Σ_i, representing the covariance between missing and observed variables and the covariance between the observed variables for the i^{th} Gaussian component, g. The sub-vectors $\boldsymbol{\mu}_m^i$ and $\boldsymbol{\mu}_o^i$ of vector $\boldsymbol{\mu}_i$ are the mean of missing and observed variables for the i^{th} component. After parameterizing the multivariate Gaussian components of the conditional distribution, the conditional distribution of the GMM can be defined by:

$$\hat{f}(\boldsymbol{y}_m; \boldsymbol{\Psi}_m) = \sum_{i=1}^{g} \pi'_i \phi(\boldsymbol{y}_m; \bar{\boldsymbol{\mu}}_m^i; \bar{\Sigma}_{mm}^i) \tag{2.11}$$

$$\pi'_i = \frac{\pi_i \phi(\boldsymbol{y}_o; \boldsymbol{\mu}_o^i; \boldsymbol{\Sigma}_{oo}^i)}{\Sigma_{i=1}^{g} \pi_i \phi(\boldsymbol{y}_o; \boldsymbol{\mu}_o^i; \boldsymbol{\Sigma}_{oo}^i)} \tag{2.12}$$

The likelihood distribution is then calculated as the marginal of the conditional GMM distribution relative to the variable being imputed.

5. The prior and likelihood distributions are combined with non-parametric BU to create an updated distribution, as per Neufeld & Deutsch (2006), which is as follows:

$$P(A|B,C) = \frac{P(A|B)\,P(A|C)}{P(A)} \tag{2.13}$$

$P(A|B)$ is the prior distribution, $P(A|C)$ is the likelihood distribution, and $P(A)$ is the global distribution which is N (0,1) from the original independent normal score transformation.

6. Finally, missing data realizations are generated. The updated distribution is sampled with Monte Carlo Simulation to generate imputed realizations.

Hadavand and Deutsch (2020) advocates visualizing marginal distributions to help determine the number of Gaussian components to be used in the GMM. The popular approach to determine the number of Gaussian components is the log likelihood ratio test, but Hadavand and Deutsch (2020) note that the approach is problematic; the test statistic is not bounded for Gaussian mixtures with unequal covariances matrices, and therefore an asymptotic distribution will not always approximate the true null distribution accurately. The null distribution could be generated via resampling, but this is typically computationally expensive.

2.3 Multivariate Data Transformation

Many mining projects require the characterization of multiple continuous variables. Conventional co-simulation workflows assume that datasets are multivariate Gaussian (multi-Gaussian); the assumption allows the user to fully parameterize the multivariate relationships by a covariance matrix (Chiles & Delfiner, 2012; Isaaks, 1990; Journel & Huijbregts, 1978; Verly, 1983). However, multi-Gaussian is rarely observed in geologic data, where complex features such as heteroscedasticity, non-linearity, and constraints are more commonly observed (Barnett & Deutsch, 2015a). The schematic in Figure 2.1 visualizes the typical complex features observed in geologic data.

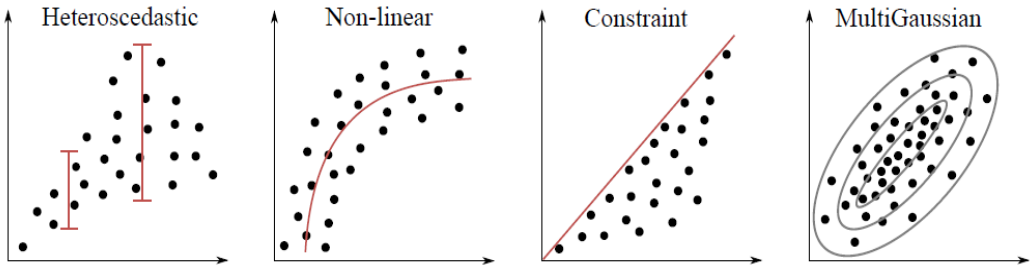

Figure 2.1: Schematic representation of multivariate complexities and a comparative multi-Gaussian distribution (Barnett & Deutsch, 2015a).

Previous authors noted that if the variables were individually transformed to be univariate Gaussian, typically through a normal score transformation, that a multivariate Gaussian assumption was still unrealistic (Barnett & Deutsch, 2015a; Chiles & Delfiner, 2012; Journel & Huijbregts, 1978). Figure 2.2 shows a co-simulation of two variables that considers the data correlation. The original variables, silica (x-axis) and magnesium (y-axis), undergo a normal score transformation producing the univariate Gaussian marginal histograms of the upper right graph. However, the bivariate distribution of the transformed variables is not multi-Gaussian. The co-simulation of the normal score variables reproduces the correlation of the dataset, but not the observed complex features. Backtransforming the simulated values intensifies the issue, where neither the correlation nor complex features are reproduced.

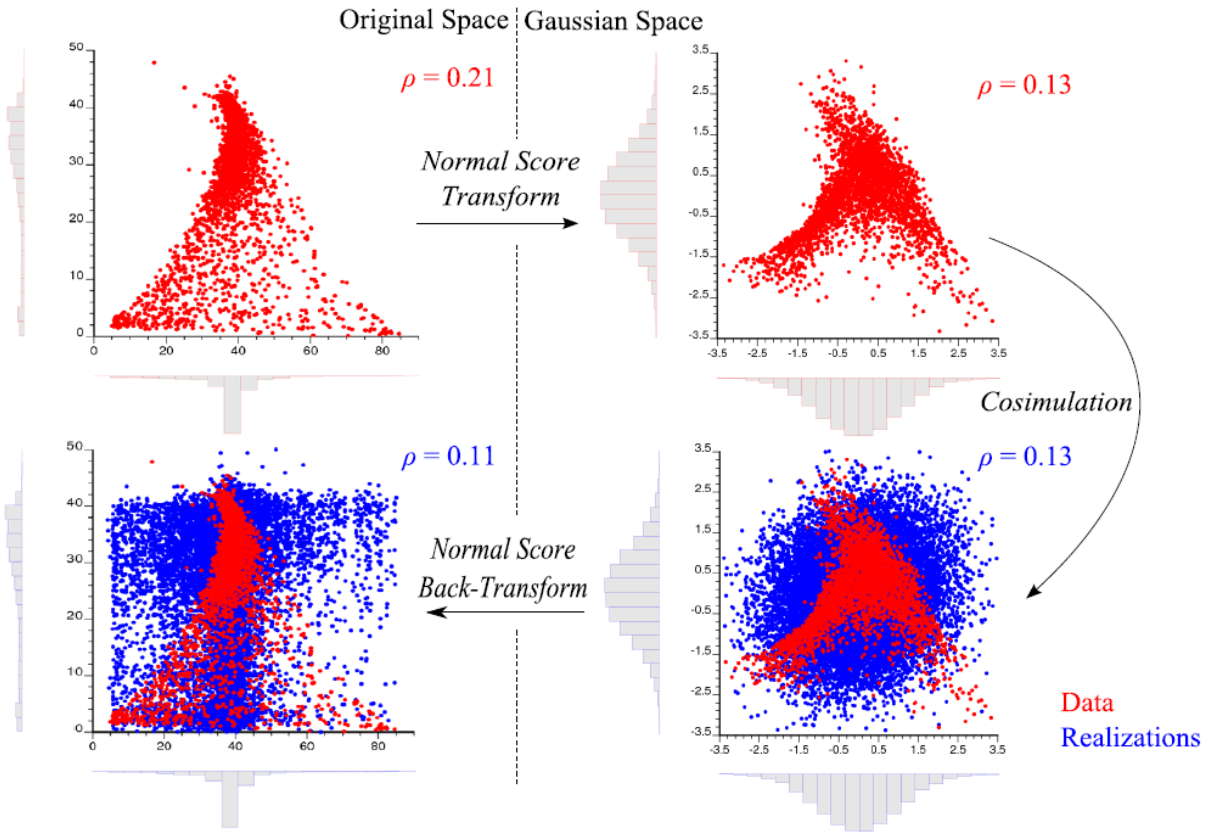

Figure 2.2: Demonstration of geostatistical co-simulation with complex bivariate data (Barnett & Deutsch, 2015a).

In an effort to correct the data replication issues, Leuangthong and Deutsch (2003) adapted a multi-Gaussian transformation to a geostatistical workflow with the stepwise conditional transformation (SCT). However, the SCT was difficult to apply with datasets over two to four variables due to the curse of dimensionality (Bellman, 1957), requiring the implementation of specialized, complicated workflows with difficult decisions (Barnett & Deutsch, 2015a). The difficulties with the SCT motivated developments of new transformations to achieve multi-Gaussianity, such as the multivariate standard normal transformation (MSNT) (Deutsch, 2011) and conditional standardization (Barnett & Deutsch, 2012). Eventually, the Projection Pursuit Multivariate Transformation (PPMT) was developed by Barnett, Manchuk, and Deutsch (2013), which the authors consider to be the current standard for transforming complex data to a multi-Gaussian. The PPMT technique effectively models complex and high dimensional data (Barnett, Manchuk, & Deutsch, 2013).

2.3.1 Overview of Projection Pursuit Multivariate Transformation (PPMT)

The PPMT workflow consists of four primary steps, as described below.

1. Normal score transformation of each variable to make the variables standard univariate Gaussian.
2. Data Sphereing to transform the Gaussian distributions to be uncorrelated with unit variance.
3. The Projection Pursuit algorithm is then used on the multivariate dataset to achieve multi-Gaussianity. The algorithm uses an optimized search that iteratively finds the "most" non-Gaussian projections and transforms the projection to be Gaussian.
4. Backtransforming the data back to original units.

The first two steps are considered pre-processing steps, with the following two steps representing the PPMT workflow. The PPMT transforms data of virtually any form, size, and dimension to an uncorrelated multivariate Gaussian distribution, allowing for independent simulation of the variables (Barnett & Deutsch, 2015a), as visualized in Figure 2.3.

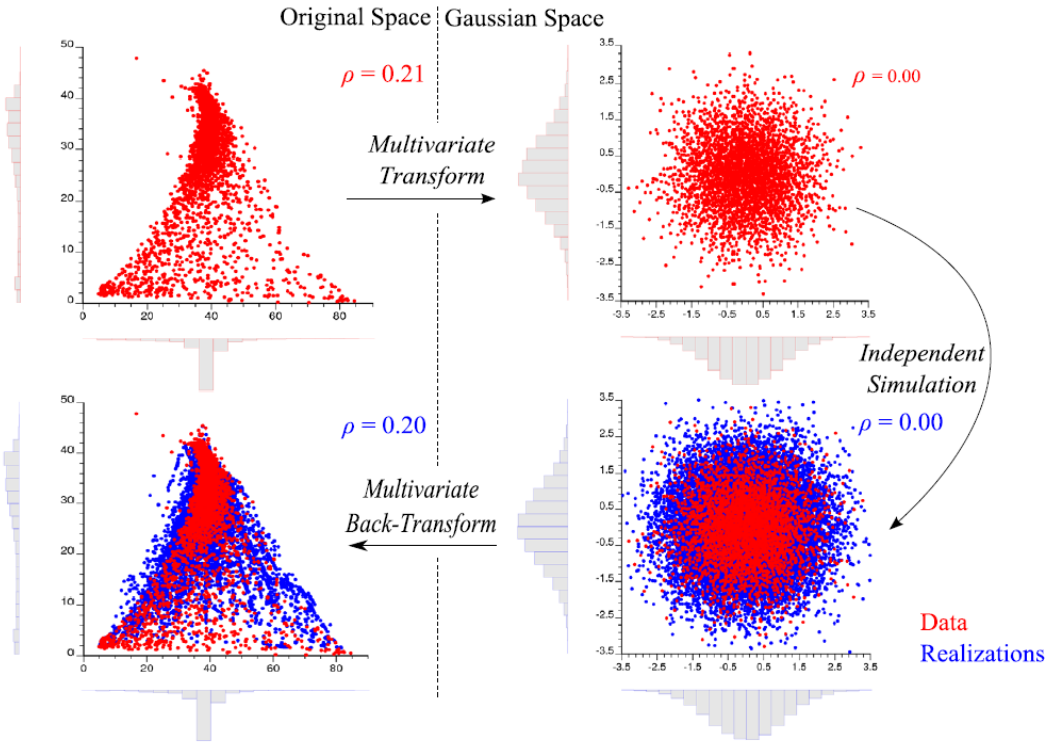

Figure 2.3: Demonstration of multivariate transformation and independent simulation with complex bivariate data (Barnett & Deutsch, 2015a).

2.3.2 Pre-processing of data: Normal Scoring and Data Sphereing

Each of the K variables, in original units, z, are transformed to be Gaussian through a normal score transformation (Bliss, 1934; Boisvert, Rossi, Ehrig, & Deutsch, 2013; Verly, 1983), yielding the univariate Gaussian $n \times K$ data matrix, $\mathbf{Y}$ (Barnett & Deutsch, 2015a). Data sphereing is applied to

transform **Y** to be uncorrelated with unit variance, resulting in **X**, as per the equation below (Fukunaga, 1972; P.J. Huber, 1985).

$$\mathbf{X} = (\mathbf{Y} - \mathbf{E}\{\mathbf{Y}\})\mathbf{S}^{-1/2}, \qquad \text{where } \mathbf{S}^{-1/2} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{V}^{\mathrm{T}} \tag{2.14}$$

$\mathbf{S}^{-1/2}$ is the sphereing matrix, **V** is the eigenvector matrix, and **D** is the diagonal eigenvalue matrix. The eigenvector matrix and diagonal eigenvalue matrix are obtained through the spectral decomposition of the covariance matrix, Σ_Y. In the sphereing matrix:

- the multiplication of **V** rotates the variables to an orthogonal axis (like PCA),
- $\mathbf{D}^{-1/2}$ transforms the variables to have a variance of one, and
- the $\mathbf{V}^T$ component projects the orthogonal variables back to the basis of the original variables, thus maximizing the loading of each original variable.

The data sphereing through spectral decomposition makes the variables uncorrelated with minimal rotation, or minimal mixing in transformed space. There are other forms of data sphereing, such as the dimension reductions sphereing (DRS), but the spectral decomposition method is preferred for PPMT due to the minimal mixing it provides (Barnett & Deutsch, 2015a).

2.3.3 Projection Pursuit Algorithm

The projection pursuit algorithm is performed on the processed **X** dataset and consists of three steps that are iteratively completed: the calculation of the projection index, an optimized search that maximizes the projection index, and Gaussianization along the projection. The projection pursuit algorithm is described below, summarized from Barnett & Deutsch (2015a).

Consider a $\boldsymbol{K} \times \mathbf{1}$ unit length vector, θ, and the associated projection of the data upon it, $p = \theta^T\mathbf{x}$. If **X** is multi-Gaussian, then any θ should yield a p that is univariate Gaussian. The vector can be examined for non-Gaussianity with the test statistic known as the projection index, $I(\theta)$. If the projection is perfectly Gaussian, the projection index will be zero. The projection index is calculated using Friedman's equation, as proposed by Friedman and Tukey (1974):

$$I(\theta) = \sum_{j=1}^{d} \frac{2j + 1}{2} E_r^2\{\psi_j(\mathbf{r})\} \tag{2.15}$$

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

$$\psi_0(\mathbf{r}) = 1,\ \psi_1(\mathbf{r}) = \boldsymbol{r}, and\ \psi_j(\mathbf{r}) = \left[(2j-1)\boldsymbol{r}\psi_{j-1}(\boldsymbol{r}) - (j-1)\psi_{j-2}(\boldsymbol{r})\right]/j\ for\ j \geq 2 \quad (2.16)$$

The $\boldsymbol{r}$ in the equation is a transformed version of the projection, $\boldsymbol{p}$, corresponding to:

$$\boldsymbol{r} = \mathbf{2}G(\mathbf{p}) - \mathbf{1}, \boldsymbol{r} \in [-\mathbf{1}, \mathbf{1}] \quad (2.17)$$

The derivative of the projection index, $I(\theta)$, may be calculated to allow for a gradient-based optimization while searching for θ that maximizes $I(\theta)$. After finding the vector that produces the maximum $I(\theta)$, the multivariate data is transformed so that the projection is univariate Gaussian, termed Gaussianization:

$$\tilde{\mathbf{p}} = G^{-1}(F(\mathbf{p})) \quad (2.18)$$

Where $\tilde{\mathbf{p}}$ amounts to the normal score transformation of the projection, $\mathbf{p}$. Transforming the pre-processed $\mathbf{X}$ dataset so that its projection along $\boldsymbol{\theta}$ is $\tilde{\mathbf{p}}$ is completed in several steps; the first step is to calculate the orthogonal matrix:

$$\boldsymbol{U} = [\theta, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{K-1}] \quad (2.19)$$

Where the K x 1 unit vectors, $\boldsymbol{\phi}_i$, are calculated using the Gram-Schmidt algorithm (Reed & Simon, 1972). Through the linear combination of the multivariate dataset, $\mathbf{X}$, and the orthogonal matrix, $\mathbf{U}$, results in the first column being the projection, $\mathbf{p} = \mathbf{X}\theta$:

$$\mathbf{XU} = [\mathbf{p}, \boldsymbol{X\phi}_1, \mathbf{X}\boldsymbol{\phi}_2, \ldots, \mathbf{X}\boldsymbol{\phi}_{K-1}] \quad (2.20)$$

Transforming the linear combination, $\mathbf{XU}$, to a standard Gaussian projection in the first column, $\tilde{\mathbf{p}}$, while leaving the orthogonal direction intact, is denoted with Θ.

$$\boldsymbol{\Theta}(\mathbf{XU}) = [\tilde{\mathbf{p}}, \boldsymbol{X\phi}_1, \mathbf{X}\boldsymbol{\phi}_2, \ldots, \mathbf{X}\boldsymbol{\phi}_{K-1}] \quad (2.21)$$

Multiplying by $\mathbf{U^T}$ returns $\boldsymbol{\Theta}(\mathbf{XU})$ to the original basis:

$$\tilde{\mathbf{X}} = \boldsymbol{\Theta}(\mathbf{XU})\mathbf{U^T} \quad (2.22)$$

The transformed multivariate data, $\tilde{\mathbf{X}}$, yields a Gaussian projection along θ, meaning it will have a projection index of $I(\theta) = 0$. The transform to be Gaussian is generalized by the function, ϕ:

$$\mathbf{y} = \phi(\mathbf{x}) \tag{2.23}$$

Iteratively completing the search and performing the gaussian transformation along projections will eventually transform the data to be multivariate Gaussian. The projection pursuit iterations are stopped once a target value of the Gaussian test statistic, $I(\theta)$, is achieved. However, as Barnett, Manchuk, and Deutsch (2013) note, stopping the algorithm in this manner is problematic, typically subjective, and not well defined in literature. If the dataset consists of many variables, discovering complexities in the data with the test statistic is difficult. If the dataset contains a small number of observations, the projections are less reliable for detecting meaningful multivariate structure. Barnett, Manchuk, and Deutsch (2013) used a bootstrapping algorithm to determine the Gaussian test statistic stopping criteria. For the bootstrapping algorithm, Monte Carlo Simulation (MCS) is used to randomly sample M distributions of matching K and n from the Gaussian CDF. A projection index value, $I(\theta)$, is then calculated for all M distributions along K random orthogonal unit vectors to yield a $M \times K$ distribution of projection indices, $\boldsymbol{I}$. The projection indices distribution of random Gaussian projection indices provides a basis for the convergence criterion. A targeted random Gaussian projection index percentile is then specified when executing the PPMT. After the PPMT forward transformation, independent simulation is completed.

2.3.4 Backtransforming

Backtransforming the simulated nodes is completed by inverting the forward transformation steps; therefore, the projection pursuit algorithm is the first step to be reversed, followed by the data sphereing, and then the normal score transformation.

The $\mathbf{1} \times \boldsymbol{K}$ vectors of N simulated nodes of K independently simulated Gaussian variables is denoted $\tilde{\boldsymbol{x}}_\alpha, \alpha = 1, \ldots, N$. First, the Gaussian nodes need to be multiplied with the orthogonal basis, $\mathbf{U}$, used in the projection pursuit iterations:

$$\tilde{\mathbf{x}}_\alpha \mathbf{U} = [\tilde{\boldsymbol{p}}, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_1, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_2, \ldots, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_{K-1}], for\ \alpha = 1, \ldots, N \tag{2.24}$$

Where $\tilde{p}$ is equal to $\tilde{x}_\alpha \theta$, and is assumed to lie within the Gaussianized projection of the transformed data, $\tilde{p} = \tilde{X} U$. The original projection of the data, $\mathbf{p}$, is then used to reconstruct its empirical CDF, F($\mathbf{p}$).

$$\Theta^{-1}(\tilde{\mathbf{x}}_\alpha \mathbf{U}) = [\boldsymbol{p}, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_1, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_2, \ldots, \tilde{\boldsymbol{x}}_\alpha \boldsymbol{\phi}_{K-1}], for\ \alpha = 1, \ldots, N \tag{2.25}$$

The normal score back transformation, Θ^{-1}, back-transforms the first entry of $\tilde{\mathbf{x}}_\alpha \mathbf{U}$ to $p = \boldsymbol{F^{-1}}(\boldsymbol{G}(\tilde{\boldsymbol{p}}))$ while leaving the remaining entries unaffected. The simulated notes, $\tilde{\boldsymbol{x}}_{\boldsymbol{\alpha}}$, are then returned to the original basis:

$$\boldsymbol{x}_{\alpha} = \Theta^{-1}(\tilde{\mathbf{x}}_{\alpha}\mathbf{U})\mathbf{U}^{\boldsymbol{T}}, \qquad for\ \boldsymbol{\alpha} = \mathbf{1},\ldots,N \tag{2.26}$$

The procedure above is repeated for each projection pursuit iteration to back-transform the simulated nodes to sphere space. The sphere space is then returned to normal score space using the sphereing matrix, $\mathbf{S}^{1/2}$:

$$\boldsymbol{y}_{\alpha} = \mathbf{x}_{\alpha}\boldsymbol{S}^{1/2}\ for\ \alpha = 1,\ldots,N \tag{2.27}$$

The simulated nodes then undergo normal score back-transformation to return the nodes back to original units, $z_{\alpha i}$:

$$z_{\alpha i} = F_i^{-1}\big(G(y_{\alpha i})\big)\ for\ \alpha = 1,\ldots,N\ and\ i = 1,\ldots,K \tag{2.28}$$

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

2.3.4 Limitations

In most settings the PPMT surpasses other transformation methods for replicating complex multivariate features, but limitations exist.

PPMT assumes that the transformed data will match the multivariate density of independently simulated Gaussian realizations (Barnett, Manchuk, & Deutsch, 2013). Where this assumption is true, Gaussian mapping should back-transform the simulated realization to reproduce the original density, but where the assumption is untrue, the original density will not be reproduced in the back-transform (Barnett, Manchuk, & Deutsch, 2013). Practitioners are advised by Barnett, Manchuk, and Deutsch (2013) to transform the normal score realizations to target $N(0,1)$ distributions to improve histogram reproduction.

PPMT, like the normal score transformation, requires the extrapolation of tail values. This is difficult in PPMT with the added dimensionality inherent to the transformation, where these tail values cannot be described with a single number. Therefore, the PPMT may not be appropriate in cases where the sampled data does not describe a large portion of the deposit's variability and extreme values are crucial to the related model function (Barnett, Manchuk, & Deutsch, 2013).

2.4 Spatial Multivariate Bootstrap

Considering the uncertainty in global parameters is important in geostatistical workflows because uncertainty in resources can be significantly underestimated if ignored (Khan & Deutsch, 2016). Bootstrapping is a statistical resampling scheme that allows uncertainty in a global input statistical

parameter to be assessed from the data (Rossi & Deutsch, 2014). The multivariate spatial bootstrap is an extension of the bootstrap resampling scheme, where the spatial correlation between the variables at data locations is considered (Rezvandehy & Deutsch, 2016).

The first step of the spatial multivariate bootstrap workflow is the establishment of prior uncertainty. Prior uncertainty is estimated using a resampling algorithm while considering spatial correlation of variables at data locations, essentially equating to unconditional simulation. Interestingly, in the spatial bootstrap method, a dataset with a high degree of spatial correlation will have higher uncertainty than a dataset with a low degree of spatial correlation, as the correlated data are effectively redundant (Khan & Deutsch, 2016). The prior uncertainty is updated through a conditional simulation approach that accounts for domain extents and the conditioning data, resulting in the posterior distribution (Khan & Deutsch, 2016). The posterior distribution represents the uncertainty associated with the measured statistic. The multivariate spatial bootstrap workflow in a geostatistical context was proposed by Babak and Deutsch (2009) and subsequently described by Khan and Deutsch, 2009, then Vincent and Deutsch (2019), these papers serve as the basis for the summary below. Note that the papers focused on histogram uncertainty and did not address uncertainty in a variogram/covariance function. The Lineal Model of Coregionalization (LMC) is a linear function and would not be an appropriate representation of the complex relationships observed in the data and was avoided.

2.4.1 Prior Uncertainty

Babak and Deutsch (2016) proposed a multivariate spatial bootstrap procedure to constitute the prior parameter uncertainty for statistics such as the means, variances, and the multivariate covariance matrix. The multivariate spatial bootstrap procedure uses Monte Carlo simulation (Journel & Bitanov, 2004; Feyen & Caers, 2006) to sample the global statistics to determine the prior uncertainty while considering the spatial correlation of $\boldsymbol{K}$ variables at n data locations.

The multivariate spatial bootstrap resampling procedure for geologic variables to obtain a prior uncertainty consists of six primary steps (Vincent & Deutsch, 2019), as described below.

1. Normal score transformation of the data, $F_{Z_k}(z_k)$ to Gaussian units, $G_{Y_k}(y_k)$ for each variable.
2. Create a Lineal Model of Coregionalization (LMC) by modeling the direct and cross variograms, $\gamma_{Y_k}(\boldsymbol{h})$ and $\gamma_{Y_{kk'}}(\boldsymbol{h})$, between all Gaussian variables, where the subscript of k and k' are the random variables indicating the variables in the cross variogram. There is a total of $K(K+1)/2$ direct and cross variograms. Practitioners can use a simpler intrinsic model of coregionalization if deemed appropriate.

3. A Cholesky LU decomposition of the covariance matrix is performed, $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^T$, where $\mathbf{C}$ is the covariance matrix, $\boldsymbol{L}$ is the lower triangular matrix and $\boldsymbol{L}^T$ is the transpose of the lower triangular matrix. $\boldsymbol{L}$ and $\boldsymbol{L}^T$ are $nK \times nK$ matrices.
4. Unconditional realizations, M, are generated by multiplying the $\boldsymbol{L}$ matrix by $\boldsymbol{w}$, an $n\boldsymbol{K} \times \mathbf{1}$ uncorrelated standard normal deviate vector, for each realization.

$$\boldsymbol{y}^{(m)} = \boldsymbol{L}\boldsymbol{w}^{(m)}, m = 1, \ldots, M \tag{2.29}$$

The result is $\boldsymbol{y}$, a $n\boldsymbol{K} \times \mathbf{1}$ vector of unconditionally simulated Gaussian values with appropriate correlation over M realizations.

5. The values of the realizations, y, are back transformed to original units, z, using the original transformation tables.
6. Generate the prior uncertainty of the parameter of interest (e.g., mean, variance, etc.) for each random variable from the back transformed distributions.

2.4.2 Posterior Uncertainty

The prior uncertainty is transferred to the posterior through a conditional simulation approach that accounts for domain extents and the conditioning data, as per the steps described by Vincent and Deutsch (2019), summarized below.

1. Normal score transform the original data M times using the spatial bootstrap realizations as reference distributions.
2. Use the transformed data to perform M conditional simulations of the multivariate data, then backtransform each realization to original units using the transform tables from the previous step.
3. The backtransformed realizations can be postprocessed to create the posterior uncertainty.

The incorporation of multivariate parameter uncertainty into a geostatistical workflow is thus finalized with the simulation-based workflow described above.

2.5 Geometric Imputation of Vein Deposits

The Arrow Deposit is considered a hydrothermal vein deposit as described by Edwards and Atkinson (1986): Hydrothermal vein deposits are accumulations of economic minerals which form in association with hot aqueous fluids to develop a vein or multiple veins. The estimation domains for vein deposits are typically deterministically defined by explicit modeling, implicit modeling, or surface interpolation (Carvalho, 2018). These traditional methods for the creation of estimation domains are rigid and do not provide uncertainty assessment. However, Carvalho (2018) proposed a framework to capture geometric uncertainty of simple tabular vein domains while considering a local coordinate system, position

uncertainty, and thickness uncertainty. The workflow generated multiple realizations of the vein geometry and boundaries with unstructured grids that fit the shape of the vein (Carvalho, 2018); below is a summary of his framework.

First, Carvalho converted the data from a Cartesian coordinate system to a local coordinate system. The transformation addresses the typical issues associated with the Cartesian coordinate system: it does not typically align with vein geometry and grade anisotropy, and thickness is not always easily calculated due to inclined drill holes (Carvalho, 2018). The transformation used was the one proposed by Ostenberg and Deutsch (2017), which fits a plane to the intercepts points, then defines a new coordinate system that is geologically significant relative to the plane. The plane is fitted via a linear regression modelling technique called Total Least Squares (TLS), which is calculated using drill hole intercept midpoints. TLS plane fitting minimizes the error between the points and the plane by minimizing the sum of squared errors (SSE). Three orthogonal directions were used to define the coordinate axis of the new modified system, referred to as u, v, and w, which correspond to the dip-direction, strike-direction, and perpendicular-to-the-vein-direction (Ostenberg and Deutsch, 2017).

Position and thickness uncertainties, or geometry data uncertainty, are in part due to inclined drill holes intercepting the domain, where true thickness and vein location perpendicular to the intercept-point are not measured. The geometry data could be estimated at the unknown data location, but estimation would produce a deterministic value. Carvalho captured the position and thickness uncertainties with the imputation of geometry data for missing data locations; he deemed missing data locations as intercepts with an angle greater than 34 degrees from perpendicular to the vein in the modified coordinate system (Figure 2.4).

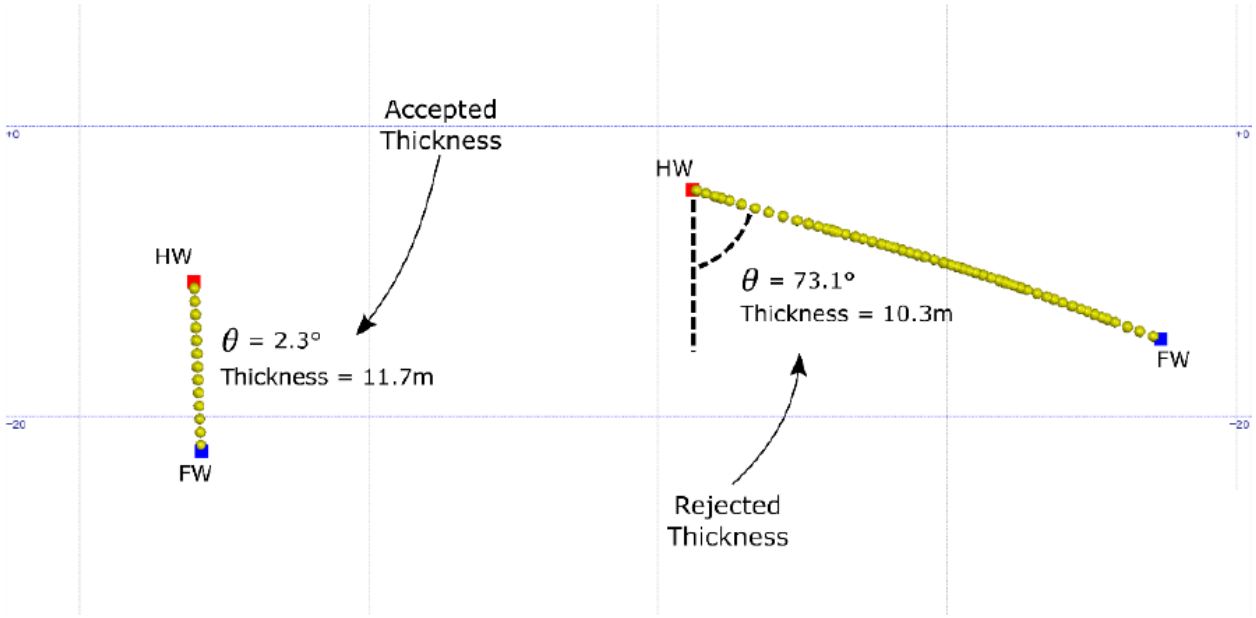

Figure 2.4: Cross section perpendicular to the fitted plane showing Carvalho's example of rejected thickness values for simulation, thus requiring imputation (Carvalho, 2018).

Variograms were calculated and modeled in the modified coordinate system for position and thickness in normal score units. The footwall surface was selected as the base surface, where the hangingwall intercepts were projected to the footwall and the hangingwall position was simulated using Sequential Gaussian Simulation (SGS) based on footwall elevation and thickness values. Subsequently, footwall intercepts are projected to the hangingwall, and the footwall elevation and thickness data are simulated. See Figure 2.5 for a schematic of the simulating procedure. The simulation is carried out at data locations; therefore, Carvalho used an irregular grid.

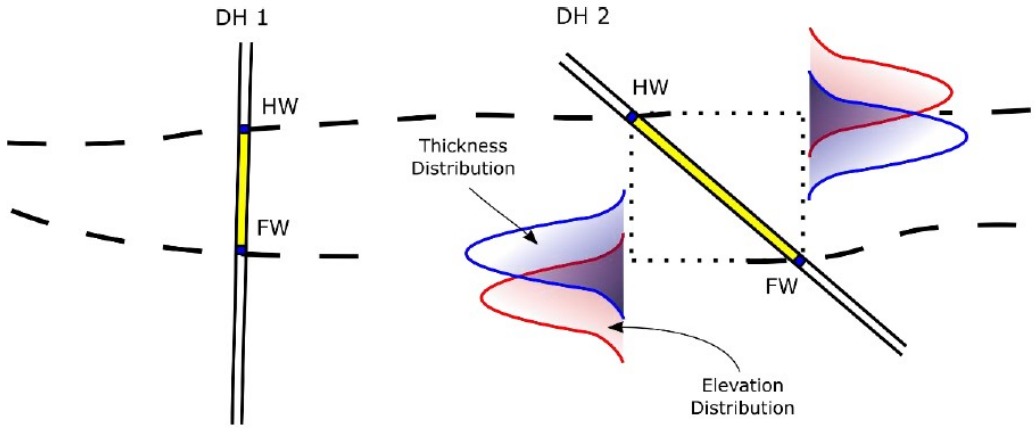

Figure 2.5: Intercepts projection and position simulated (Carvalho, 2018).

The elevation and thickness distributions were merged for the distribution to be randomly sampled. The merged distribution was created by back transforming the elevation values to original units, calculating

thickness values from the elevation values, and then capping the newly formed thickness distribution relative to the global thickness distribution. The newly formed capped thickness distribution is added to simulated thickness distribution, as both distributions are in the same space.

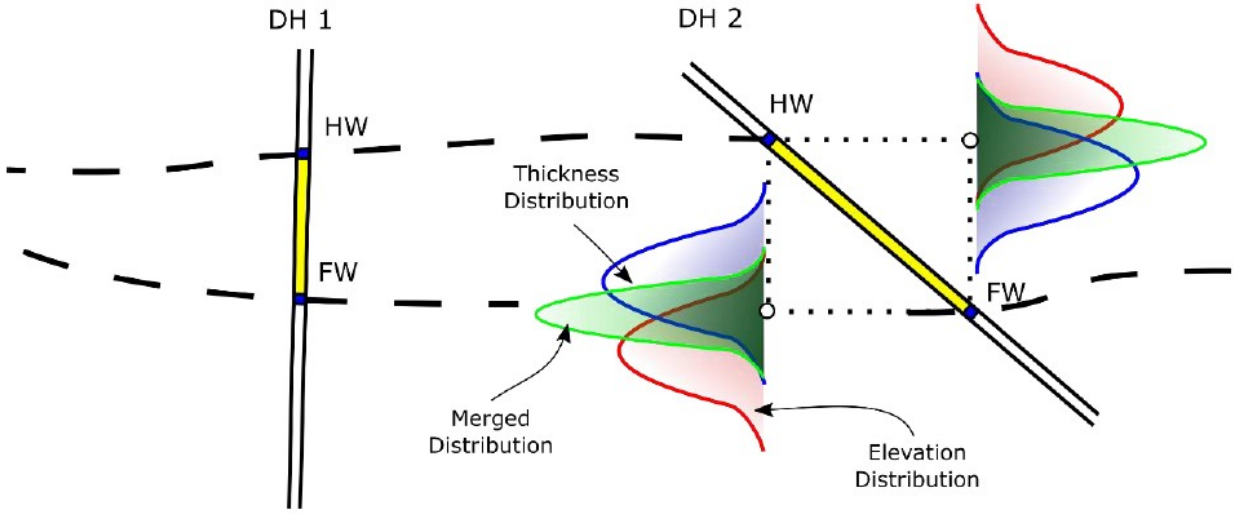

Figure 2.6: Schematic of elevation and thickness distribution merging (Carvalho, 2018).

The distributions were merged using error ellipses with geological data, as per the equation below, which was adapted for two univariate distributions (Rezvandehy & Deutsch, 2014).

$$\bar{X} = C\left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right), \qquad C = \left(\sigma_1^{2\,-1} + \sigma_2^{2\,-1}\right)^{-1} \tag{2.30}$$

The vector correspond to the location of the weighted average is denoted by $\bar{X}$, C is the resulting variance, m_n is the univariate mean, and σ_n^2 is the univariate variance.

A random sample is iteratively selected from the merged distribution to create an imputed distribution and the imputed values are back transformed to original values. Surfaces were created for the footwall and hangingwall with each surface realization using a single and different imputation realization as visualized in Figure 2.7, finalizing the workflow Carvalho proposed to assess geometry uncertainty in tabular vein deposits.

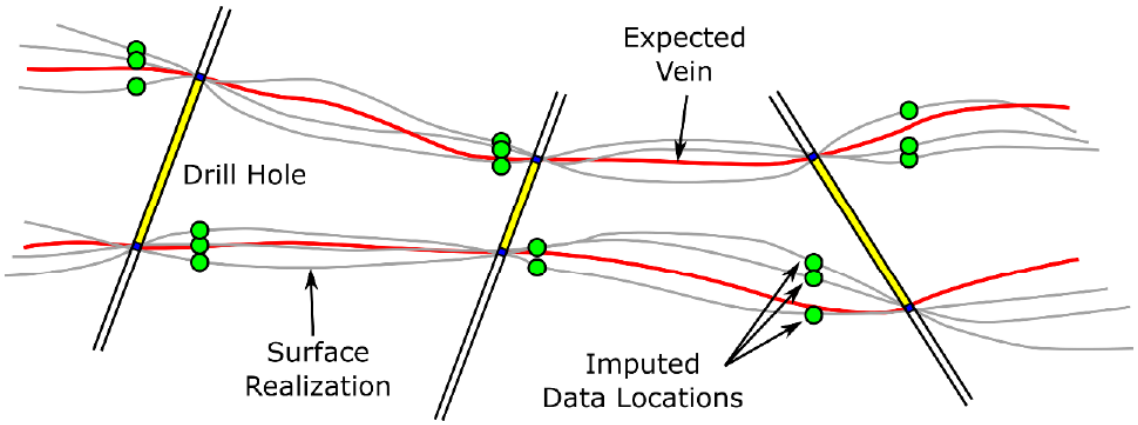

Figure 2.7: Schematic of surface simulation (Carvalho, 2018).

2.6 Signed Distance Functions for Boundary Uncertainty

Defining the areal extent, or the boundary, of estimation domains is important in accurately representing the contained mineral resources. Boundary modeling is traditionally defined through explicit or implicit modeling, resulting in a deterministic model that does not provide uncertainty assessment.

The Signed Distance Function (SDF) with a modification of a constant, known as the C-parameter, has been used to assess boundary uncertainty (Munroe & Deutsch, 2008; Hosseini, 2009; Wilde & Deutsch, 2011). The SDF algorithm is used in implicit modeling as a Euclidean distance measure between the nearest two samples of different indicator values. In SDF, the negative indicator values are 'inside', and the positive indicator values are 'outside'; the indicator formalism for defining the data is:

$$i(\boldsymbol{u}_\alpha) = \begin{cases} -1, if\ \boldsymbol{u}_\alpha \in \text{domain} \\ 1, \quad otherwise \end{cases} for\ \alpha = 1, \dots, n \tag{2.31}$$

Where $i(\boldsymbol{u}_\alpha)$ is the indicator at location $\boldsymbol{u}_\alpha$ and is -1 inside the domain and 1 otherwise calculated for all α conditioning data. The SDF is calculated as:

$$df(\boldsymbol{u}_\alpha) = i(\boldsymbol{u}_\alpha) \cdot min_{\beta = 1,\dots,n} \|\boldsymbol{u}_\alpha - \boldsymbol{u}_\beta\|, for \alpha = 1, \dots, n \text{ and } i(\boldsymbol{u}_\alpha) \neq i(\boldsymbol{u}_\beta) \tag{2.32}$$

The distance function at the data location, $df(\boldsymbol{u}_\alpha)$, is equal to the indicator at that location multiplied by the absolute minimum distance to the nearest sample of an opposing indicator value. Anisotropy can be accounted for by:

$$\|\boldsymbol{u}_\alpha - \boldsymbol{u}_\beta\| = \sqrt{\left(\frac{x_\alpha - x_\beta}{a_x}\right)^2 + \left(\frac{y_\alpha - y_\beta}{a_y}\right)^2 + \left(\frac{z_\alpha - z_\beta}{a_z}\right)^2} \tag{2.33}$$

Where the a values specify the anisotropy, and the x, y, z are coordinate locations of the two samples.

When the SDF is equal to zero, the model is considered to be transitioning. An isosurface is made at the zero horizon, representing the contact between the domain and outside.

The modified-SDF is a modification of the distance function using a C-parameter and β-parameter; the C-parameter is used to create an uncertainty bandwidth and the β-parameter corrects bias to centre the uncertainty bandwidth (Munroe & Deutsch, 2008).

The C-parameter is an additive factor that is added to the data if it is outside the domain and subtracted from the data if inside, increasing the difference between the function values of inside and outside data.

$$DF_{mod} = \begin{cases} df(\boldsymbol{u}_\alpha) - C, \in domain \\ df(\boldsymbol{u}_\alpha) + C, otherwise \end{cases} \tag{2.34}$$

The C-parameter constant is typically calibrated using a jackknife procedure (Hosseini, 2009; Wilde & Deutsch, 2011). The jackknife procedure leaves out a subset of data with subsequent estimation. The estimation is then compared to the left-out data. The procedure is iteratively completed with varying C-values until an acceptable number of misclassifications are achieved between the estimation and the left-out data. Misclassification is a disagreement between the estimation and the left-out data values, for example, a "outside" left-out data location that is estimated to be "inside". The acceptable limit of misclassifications is commonly determined using elbow plots by selecting the distance which little change in misclassifications occurs. The resulting band between - C and + C represents the range of uncertainty where the true boundary exists somewhere in between (Wilde & Deutsch, 2011).

The β-parameter can be used to address a bias within the model by multiplying or dividing the original SDF by a constant:

$$DF_{mod} = \begin{cases} (df(\boldsymbol{u}_\alpha) - C) \cdot \beta, \in domain \\ (df(\boldsymbol{u}_\alpha) + C)/\beta, otherwise \end{cases} \tag{2.35}$$

The β-parameter shifts the uncertainty bandwidth location that is controlled by C (Munroe & Deutsch, 2008). The β-parameter is typically calibrated by several reference models.

The SDF can introduce bias to the model in the presence of asymmetrical, sparse data, which is commonly observed in the mining industry. Mancell (2020) addresses the problem introduced by

asymmetrical, sparse data, by using an indicator estimator to generate a probability field to establish an uncertainty bandwidth.

2.7 Indicator Estimate Approach to Boundary Uncertainty

Mancell (2020) proposed a new implicit modeling technique for boundary modeling that provides a globally unbiased model with a measure of uncertainty. Mancell (2020) generated a field of probabilities through an indicator estimate and applied a threshold that results in an extracted boundary. An uncertainty bandwidth is extracted from the indicator estimate by varying the indicator thresholds to produce eroded and dilated boundaries. The workflow consists of four primary steps, as described below.

1. Creation of a Nearest Neighbour (NN) model to provide an unbiased spatial representation of the conditioning data.
2. Creation of an indicator estimate to generate a field of probabilities of boundary location.
3. Thresholding the indicator estimate to achieve global unbiasedness as derived from the NN model.
4. Thresholding above and below the NN threshold to provide eroded and dilated boundary that gives access to a bandwidth of uncertainty. Mancell determined that values ±0.15 of the NN-model base case are reasonable for bandwidth of uncertainty.

NN volumes are a reasonable representation of global volume, as the model provides an unbiased spatial representation of the conditioning data and is a reasonable interpolator to use in sparse data areas; thus, NN volumes can be used to check an interpreted model's global volume (Rossi & Deutsch, 2014). The NN algorithm designates a geological attribute to a selected node by finding the nearest sample and assigning its value.

The first step in creating the NN model is to define the indicators through indicator formulism:

$$i(\boldsymbol{u}_\alpha) = \begin{cases} 1, if\ \boldsymbol{u}_\alpha \in domain \\ 0, \quad otherwise \end{cases} for\ \alpha = 1, \dots, n \tag{2.36}$$

Where $i(\boldsymbol{u}_\alpha)$ is the indicator at location $\boldsymbol{u}_\alpha$ and is 1 inside the domain and 0 otherwise calculated for all α conditioning data.

The NN-model is created through the Ordinary Kriging of the indicators with a maximum of one datum; using a single informing data point and having the sum of the kriging weights equal one effectively equates to a NN-model. The model is tuned by varying the search distances, selecting a distance that results in a stable model that is geological reasonable. A global volume is obtained by summing all the cells assigned an 'inside' value in the NN-model.

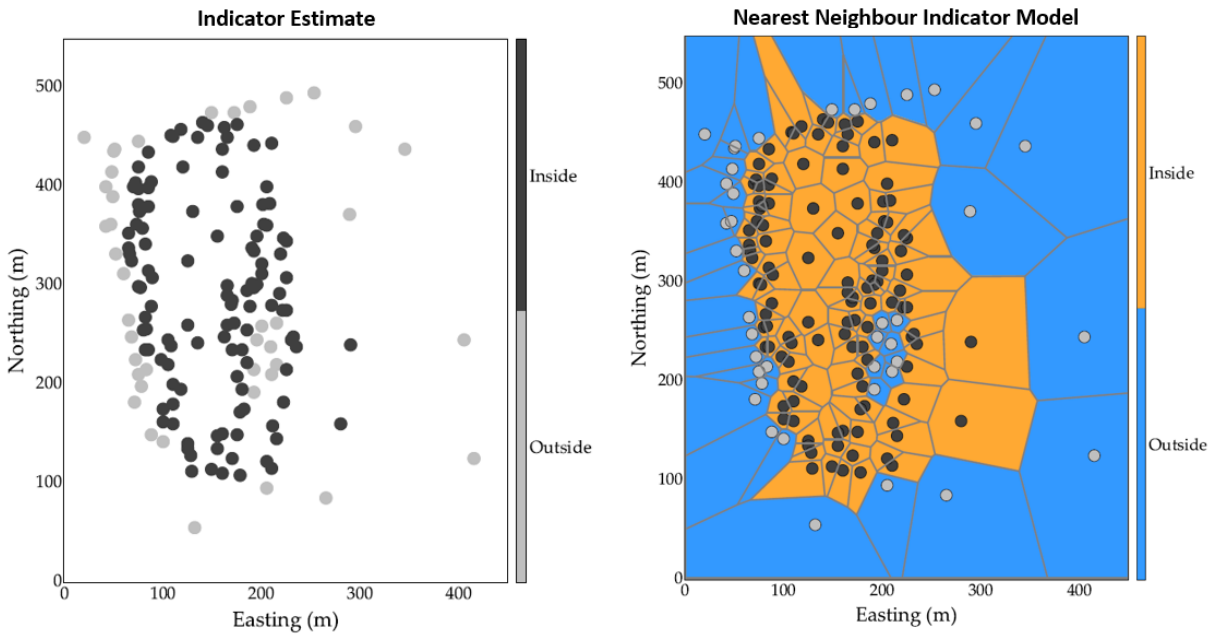

Figure 2.8: Two-dimensional dataset with single domain comprised of inside and outside data (left). Nearest Neighbour model of indicators (right). Mancell (2020).

The indicator estimate can be carried out with various interpolators using the same indicator formalism as the NN-model. Mancell (2020) tested global estimation techniques, such as Global Kriging and Radial Basis Functions (RBF), to create smooth, artifact-free models that considered all the conditioning data (Carvalho, 2018). RBF was selected over kriging in the case-study because the lack of edge effects observed in the RBF estimate, that first-order stationarity was not required, and variograms were not necessary (Martin, 2019). The edge effect is a common problem in boundary estimation where there is uncontrolled extrapolation of estimated domains to the extents of the model. The edge effect is the influence of samples extending into areas of sparse or no data. In Global Ordinary Kriging, the estimates distal to data are relatively constant with values close to the dataset's mean indicator value. Mitigating edge effects in Global Ordinary Kriging is completed by restricting the variogram range or adding artificial controlling samples into the dataset (Maureira, 2015).

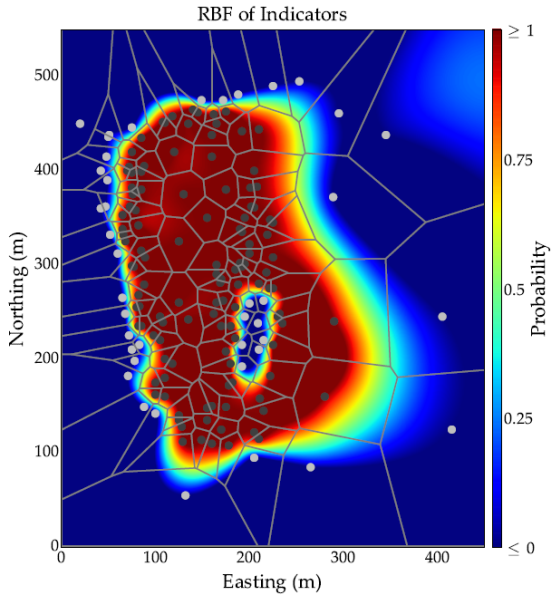

Figure 2.9: RBF estimate of two-dimensional dataset with single domain comprised of inside and outside data (Mancell, 2020).

Mancell (2020) noted that global estimation techniques are problematic with datasets containing N >30,000 due to the computational power required for the dense N x N matrix used in the methods. Ordinary kriging with a restricted search can be used for the large datasets to reduce the computational demand (Deutsch & Journel, 1998). If an RBF is used as the indicator interpolator in these large datasets, sparse iterative or direct solvers can be used to reduce the computation time (Carr et al., 2001; Martin, 2019).

After the NN-model and indicator estimate are finalized, a threshold for the indicator estimate needs to be selected to extract an unbiased boundary. As previously mentioned, NN-models are a reasonable unbiased representation of global volume and can be used to select the indicator estimate threshold value. First, the ratio of inside to outside cells in the NN-model is calculated to provide a NN-thresholding value, th_{NN}, between [0,1] as per the equation below (Mancell, 2020).

$$th_{NN} = 1 - \frac{V_{NN}}{V_{Total}} \tag{2.37}$$

Then, the NN-thresholding value extracts from the indicator estimate cumulative frequency distribution (CDF) the estimate thresholding z-value, z_i^* (Mancell, 2020).

$$z_i^* = F_{i^*}(th_{NN}) \tag{2.38}$$

The z_i^*-value represents the boundary and all cells with a probability above this value are considered 'inside' the domain and all cells with a probability below this value are considered 'outside' the domain. A bandwidth of uncertainty is generated by varying the threshold value by 0.15 above or below the z-value to create eroded and dilated boundaries.

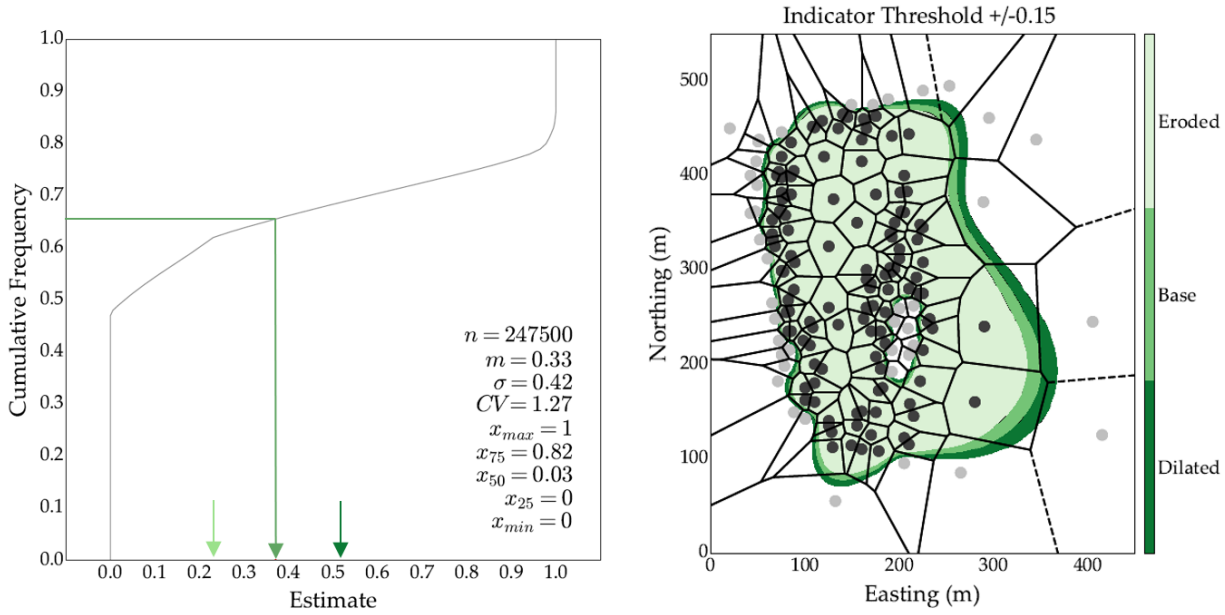

Figure 2.10: Cumulative Distribution Function of RBF indicator estimate with green line and arrow demarcating the threshold to NN model, and the lighter and darker arrows indicating the +/- 0.15 uncertainty thresholds (left; modified from Mancell, 2020). NN volume ratio of 0.66 corresponds to z-value threshold of 0.37. An uncertainty bandwidth was calculated to be between 0.22 and 0.52. Indicator threshold uncertainty model with +/- 0.15 uncertainty thresholds (right; Mancell, 2020).

Mancell (2020) used Probability Threshold Curves (PTCs), which visualize the relationship between threshold values and probability for an indicator estimate, to determine that 0.15 was an appropriate degree to vary the z-value for an uncertainty bandwidth. The first step in constructing PTCs is to generate true scenarios through a simulation process, then subsequently sample the N truths at varying spacings. The indicators were interpolated with the multiple datasets using global kriging and RBF, and the resulting models were thresholded from p100 to p0 in increments of 0.05. Mancell compared the thresholded model volumes to their respective truths to determine the probability of a threshold model to be larger than its truth. An experimental PTC was then constructed by plotting the probabilities against their thresholds. The process was repeated over dozens of scenarios with different structural truth attributes, creating an equal number of PTCs. The PTCs were standardized by setting their midpoints to zero. The standardized PTCs have a near-linear shape and the distance between their midpoints and

where they transition to zero (or one spans a zone of uncertainty thresholds) is predominantly between 0.10 to 0.15.

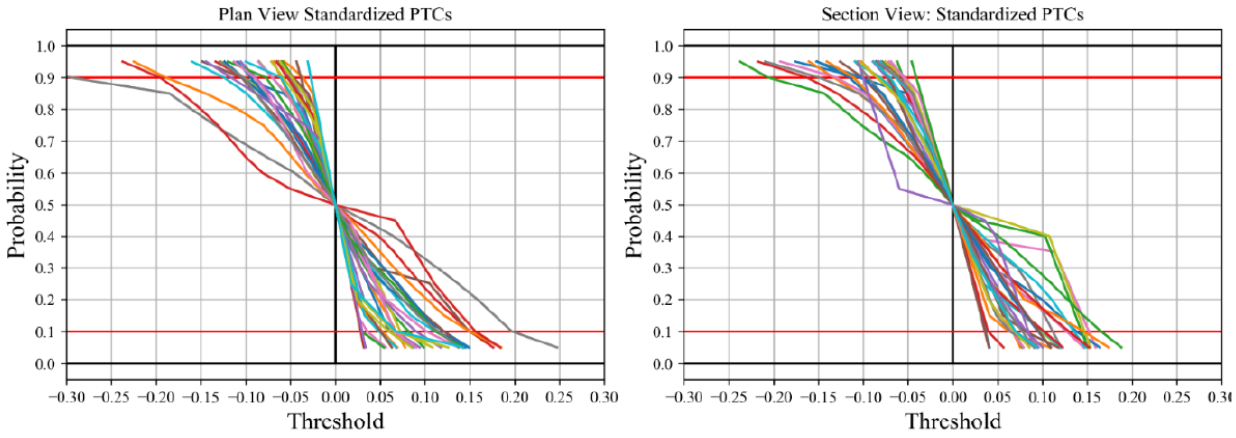

Figure 2.11: Standardized Probability-Threshold Curves for plan and section view with varying drill hole spacing and geology, trimmed to show slight variation at p90 & p10 (Mancell, 2020).

The workflow that Mancell (2020) proposed for boundary uncertainty is straight-forward to implement, accounts for asymmetries in data configuration, and provides a geologically rational output. However, as Mancell (2020) noted, NN modeling is susceptible to edge effects in areas with sparse drilling and is only controlled by adjusting maximum distances, which introduces subjectivity.

Chapter 3

3. Grade and Density Models

Two variables of interest for uncertainty quantification at the Arrow Deposit are the non-linearly, positively, and strongly correlated grade and density variables. The density variable is incomplete relative to the grade variable; the standard within the industry is to impute the density variable to make a homotopic dataset to facilitate a multivariate data transformation, as conventional co-simulation workflows assume that datasets are multivariate Gaussian. Typically, after imputation and transformation, the variables are simulated or co-simulated within a domain to generate realizations on a grid within a coordinate system. The generated realizations are reviewed together then checked.

In the study domain, the grade variable is exhaustive at all sampled drill hole locations, but the density variable is missing at approximately 82% of the sample locations and biased to high-grade intercepts. An interim debiased homotopic representative dataset is created through imputation by selecting density values from the global deposit-wide dataset with equal or near-equal grade values of the study domain's dataset. The purpose of the representative dataset, after a normal score transformation of the variables, is to fit a GMM to it. The GMM was used to impute the missing density values from the original dataset in a simulation workflow. The resulting realizations were decorrelated via Projection Pursuit Multivariate Transformation, variograms were fitted, and independent grid Sequential Gaussian Simulation was completed. The simulated grid was clipped to the boundaries of the study domain and the realizations were back transformed to original units.

Complementing the model described above, the uncertainty in the histograms of the representative dataset was assessed via the multivariate spatial bootstrap method (Khan & Deutsch, 2016). The multivariate spatial bootstrap resamples the dataset while considering the spatial correlation between the variables at data locations to create realizations, which represents prior uncertainty. The prior uncertainty was then updated through a conditional simulation that accounts for domain extents and the conditioning data, resulting in the posterior distribution. In this case, the multivariate spatial bootstrap technique is an imperfect measure of parameter uncertainty as the density variable of the representative dataset was imputed through a simple deterministic imputation process that did not consider spatial correlation and a variety of other factors. However, the output does provide a level of understanding that is helpful in assessing the risk associated with the deposit.

3.1 Workflow Outline

The workflow for the assessment of grade and density uncertainty at the Arrow Deposit consists of four primary steps, listed below and further described in the following sections of this chapter.

1. Define the missing data mechanism of the dataset (MCAR, MAR, or MNAR) and create a representative dataset.

2. Impute the missing density data using a GMM fitted to the representative dataset, where the GMM provides an estimation of likelihood distribution within a Bayesian updating workflow.

3. PPMT of the realizations from the GMM imputation workflow, as they contained multivariate complexities. The PPMT transforms data of virtually any form, size, and dimension to an uncorrelated multivariate Gaussian distribution, allowing for independent simulation of the variables

4. Independent simulation of the variables on a grid within the domain and back transform to original units.

In addition to the workflow described above, the uncertainty in the histograms of the representative dataset was assessed via the multivariate spatial bootstrap method to understand the effect of the conditioning data in its configuration within the study domain.

3.2 Missing Data and Creation of Representative Dataset

In the study domain, the collocated grade and density variables are strongly and positively correlated with a correlation coefficient of 0.91 (Figure 3.1). The grade variable is exhaustive for all 656 composite locations with an average of 242.34 g/t, whereas the density variable is incomplete with only 116 of the composite locations having density values, averaging 3.21 g/cm^3 (Figure 3.2). The density data is incomplete due to the relatively high-cost and labour-intensive nature of collecting density measurements. The locations without density values have an average grade of 226.18 g/t with a positively skewed grade distribution, whereas the locations with density values have an average grade of 401.75 g/t with a nearly uniform grade distribution (Figure 3.3). Reviewing the two subpopulations, it can be concluded that the density data was preferentially collected from high-grade samples, representing a bias in the dataset. The missing density data is considered to be a MNAR mechanism, where the probability of density depends on the grade observed and the density value itself (Rubin, 1976). A debiased representative data set needs to be created to better represent the density values missing at lower grades. The purpose of the representative dataset, after a normal score transformation of the variables, is to fit a GMM to it. Fitting a GMM to the representative dataset provides an estimation of the likelihood

distribution for the imputation workflow, and since it is fitted to the representative dataset, the bias observed in the original dataset is accounted for.

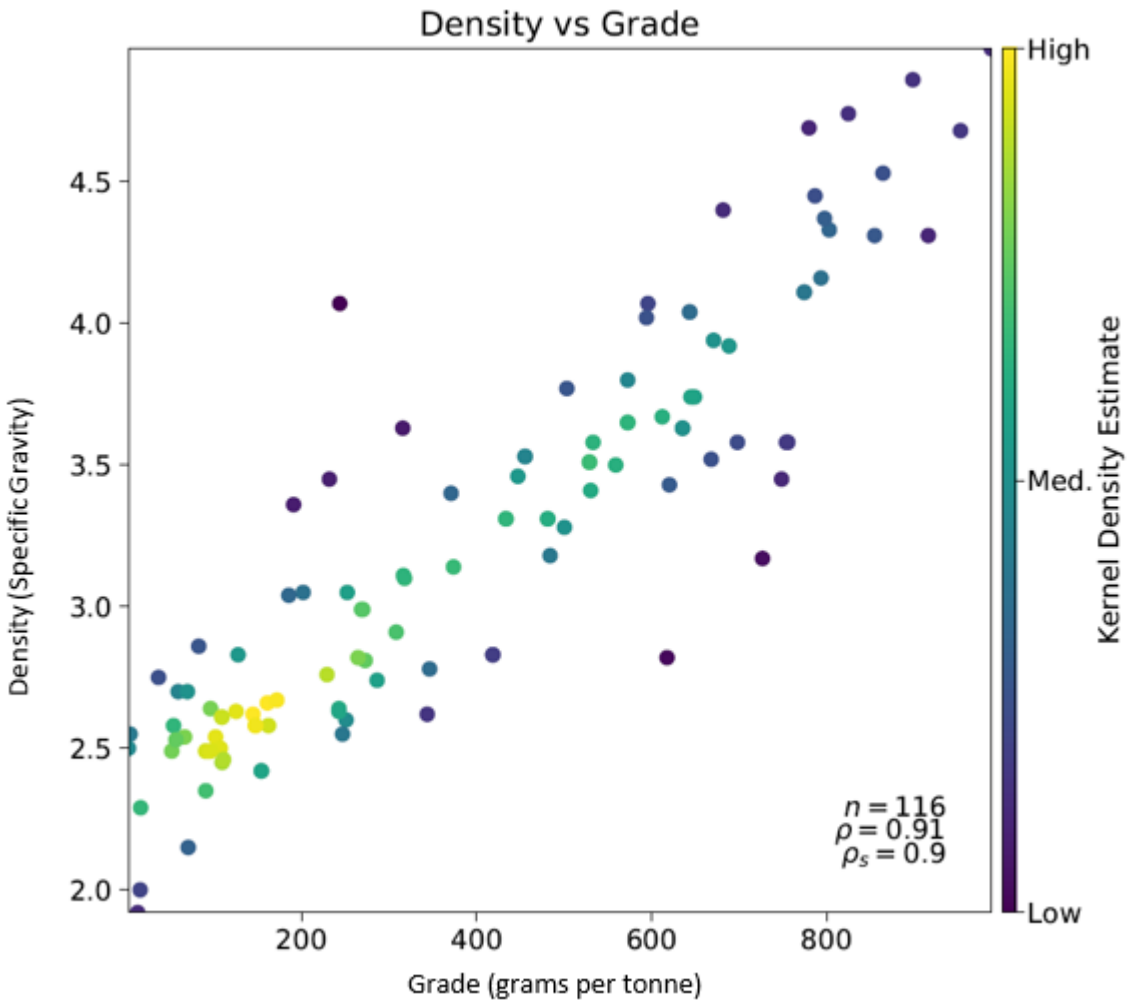

Figure 3.1: Scatter plot of the grade (Au g/t) and density (S.G.) variables of the A2 high-grade domain.

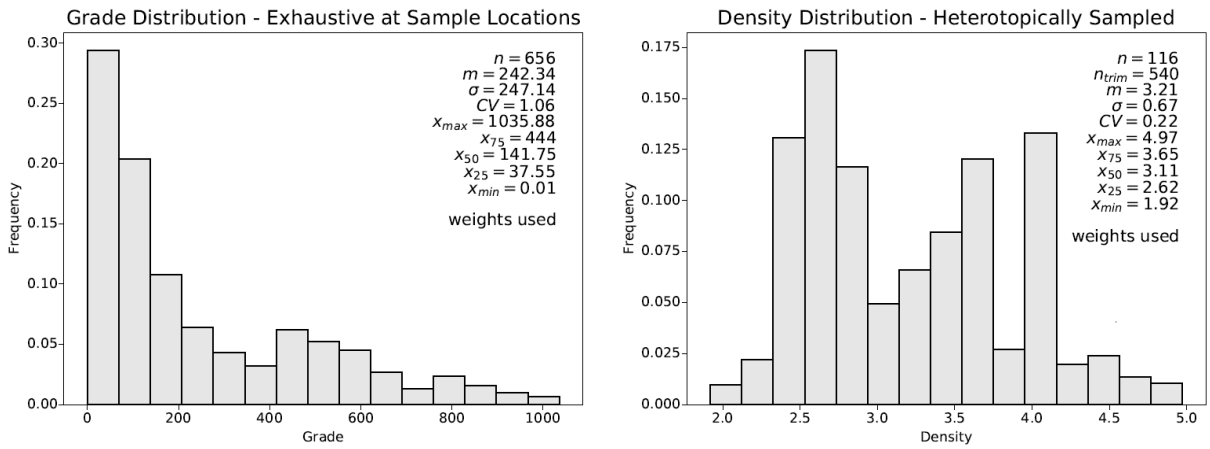

Figure 3.2: Histogram distributions of the grade (Au g/t) and density (S.G.) variables of the A2 high-grade domain. Cell declustering weights were applied to both distributions.

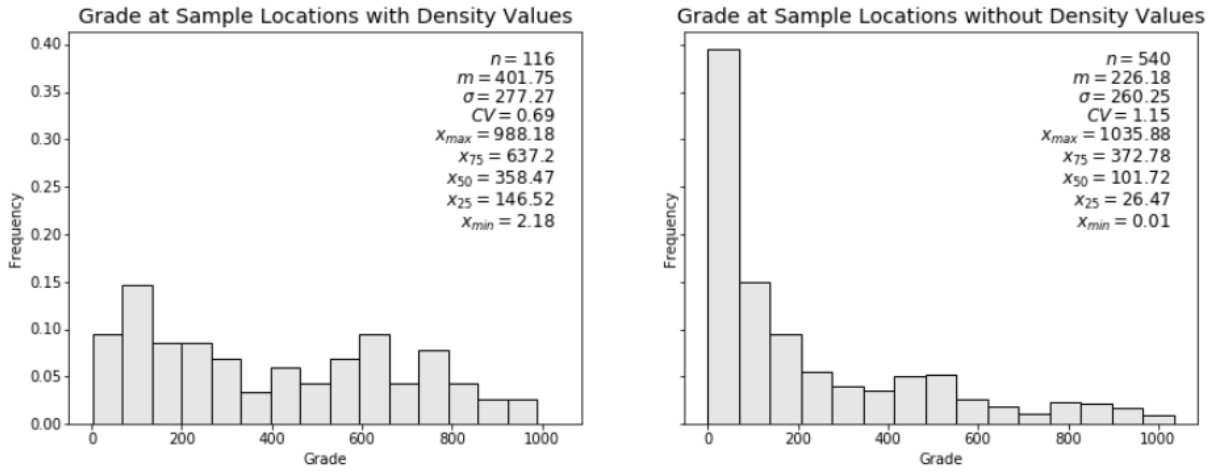

Figure 3.3: Histogram distributions of the grade (Au g/t) variable of the A2 high-grade domain at locations with and without density values.

Silva and Costa (2019) proposed a workflow to account for the MNAR mechanism by inferring the values of the under-sampled variable over the study area through a co-simulation approach and a univariate simulation performed on an artificially complete set. However, this approach does not consider that appropriate data is available in a larger, global deposit data set. At the Arrow Deposit, a global deposit dataset that contains 5850 collocated density (Z_1) and grade (Z_2) samples is available (Figure 3.4). Similar to the study domain data, the grade and density variables of the global deposit dataset are non-linearly, strongly, and positively correlated with a correlation coefficient of 0.82. It was deemed that a simple imputation method could be used to build a reasonable representative dataset by selecting density values from the global dataset rather than complete the complex workflow Silva and Costa proposed.

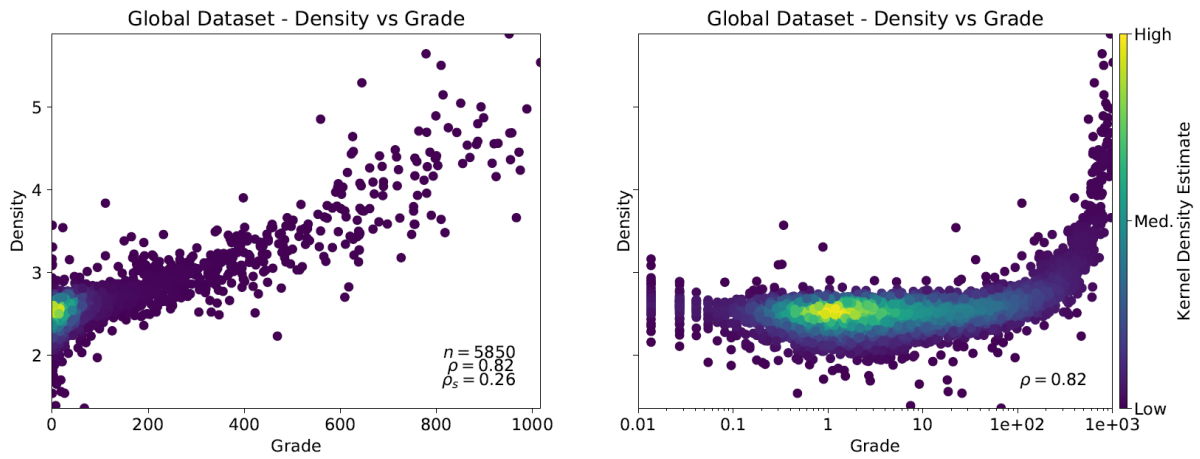

Figure 3.4: Linear and semi-log scatter plots of the grade (Au g/t) and density (S.G.) variables of the global deposit dataset.

The bias in the heterotopic study domain dataset is addressed by replacing the missing density values with suitable density values from the deposit dataset as informed by the strong correlation between grade and density. The grade of the composite with a missing density value is used as a search variable in the collocated deposit dataset. If there is one exact match between the search variable and the grade composite in the deposit dataset, the corresponding deposit density value replaces the missing value in the study domain dataset. If there are more than one matching grade, one of the corresponding density values is randomly assigned to the study domain dataset. If no grades match the search variable, the search variable is expanded by +/- 5% to create an upper and lower bound of grades to find in the deposit dataset, where a subset of related density values is created to randomly select from. The result is a debiased homotopic representative dataset for the domain that is directly built from the site-specific data. The representative dataset contains 116 collocated composites, 287 imputed density values selected from the deposit dataset where the grade values exactly matched, and 253 assigned density values selected from the deposit dataset where the grade values approximately matched.

A polynomial function was fitted to the global dataset and applied to the study domain dataset to check for appropriateness of the representative dataset (Figures 3.5 and 3.6). The polynomial function is as follows:

$$Z_1 = (1.14 * 10^{-6})(Z_2)^2 + (1.32 * 10^{-3})(Z_2) + 2.4739$$

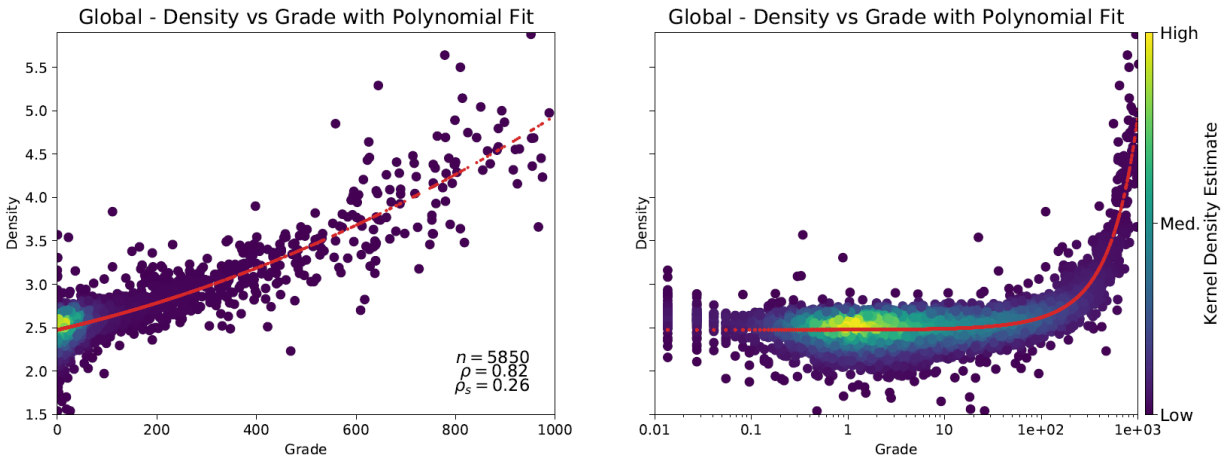

Figure 3.5: Linear and semi-log scatter plots of the grade (Au g/t) and density (S.G.) variables of the global deposit dataset with a fitted polynomial function.

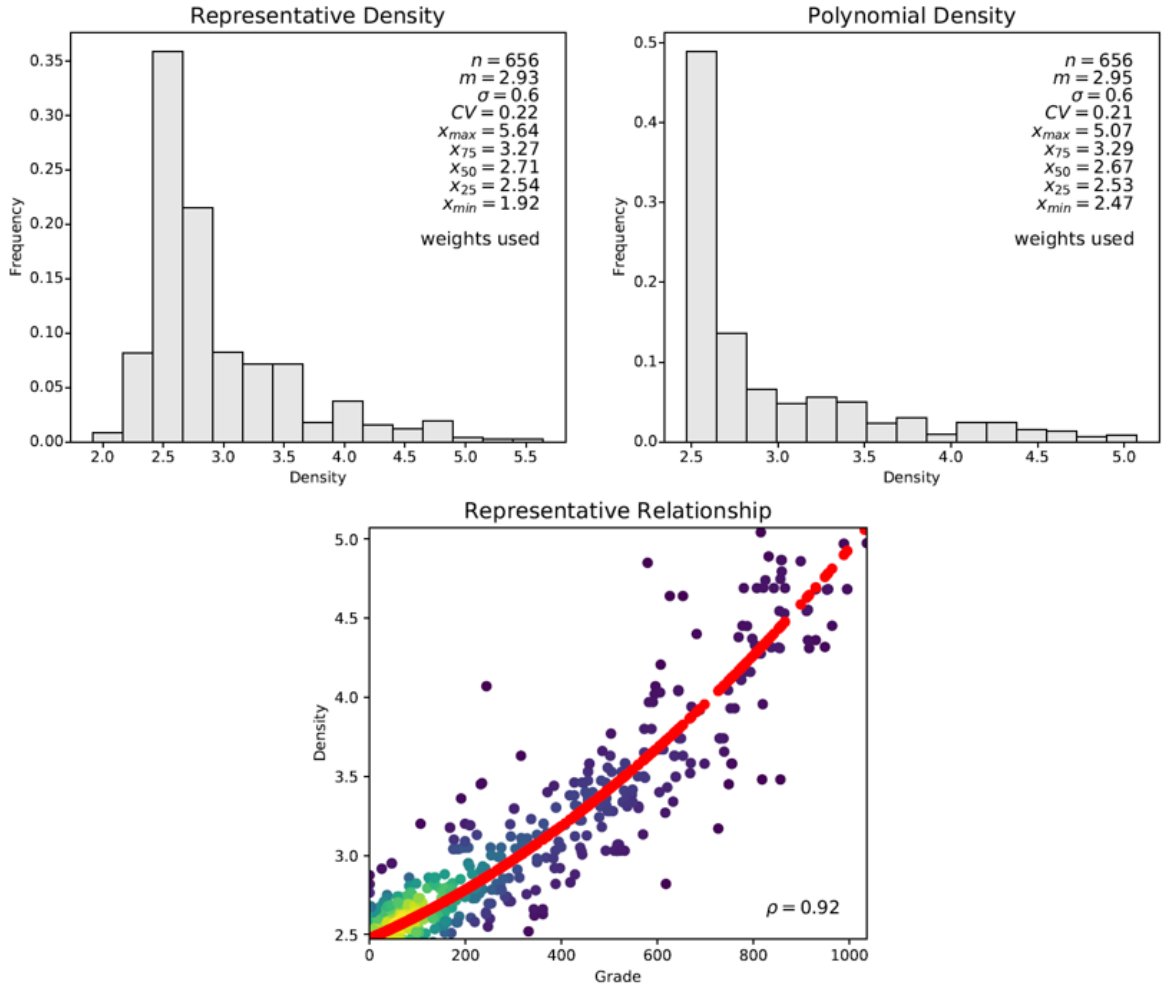

Figure 3.6: Histogram of representative density (S.G.) distribution (upper left), histogram of polynomial imputed density distribution (upper right), and scatterplot of the representative relationship overlain by the polynomial function (lower).

The polynomial imputed density dataset is oversmoothed and underrepresents low density values, but the mean and variance of the dataset closely reproduces the representative dataset, indicating the representative dataset is a reasonable representation of the study domain.

3.3 GMM and Imputation

Silva and Deutsch (2015a) provided a method for multivariate data imputation using GMM for estimation of likelihood distribution within a Bayesian updating workflow, consisting of three primary steps, as per below.

1. Defining the prior distribution at the imputation location: Each variable is assumed to be spatially multi-Gaussian after the normal scores transformation and can be parameterized by the kriging mean and variance through the simple kriging systems of equations.
2. Defining the likelihood distribution from the GMM: The likelihood distribution is defined as the marginal distribution of the conditional distribution given the collocated data observations respective to the variable being imputed.
3. Creating an updated distribution: The information from the prior and likelihood distributions are combined with non-parametric Bayesian updating.

To accommodate the first step, the grade and density variables from the homotopic representative dataset underwent individual normal score transformation to be standard normal.

The grade and density variables are assumed to be spatially multi-Gaussian after the normal scores transformation and can be parameterized by the kriging mean and variance through the simple kriging systems of equations, thus requiring the creation of independent variograms (Figure 3.7). The kriging that produces the prior distribution at each imputation location was completed in the CCG program gmm_impute. The search radii for the kriging were set to 28 m, 10 m, and 5m, for the major (azimuth of 0° and dip of 0°), minor (azimuth of 0° and dip of 90°), and vertical direction (azimuth of 0° and dip of 90°), respectively.

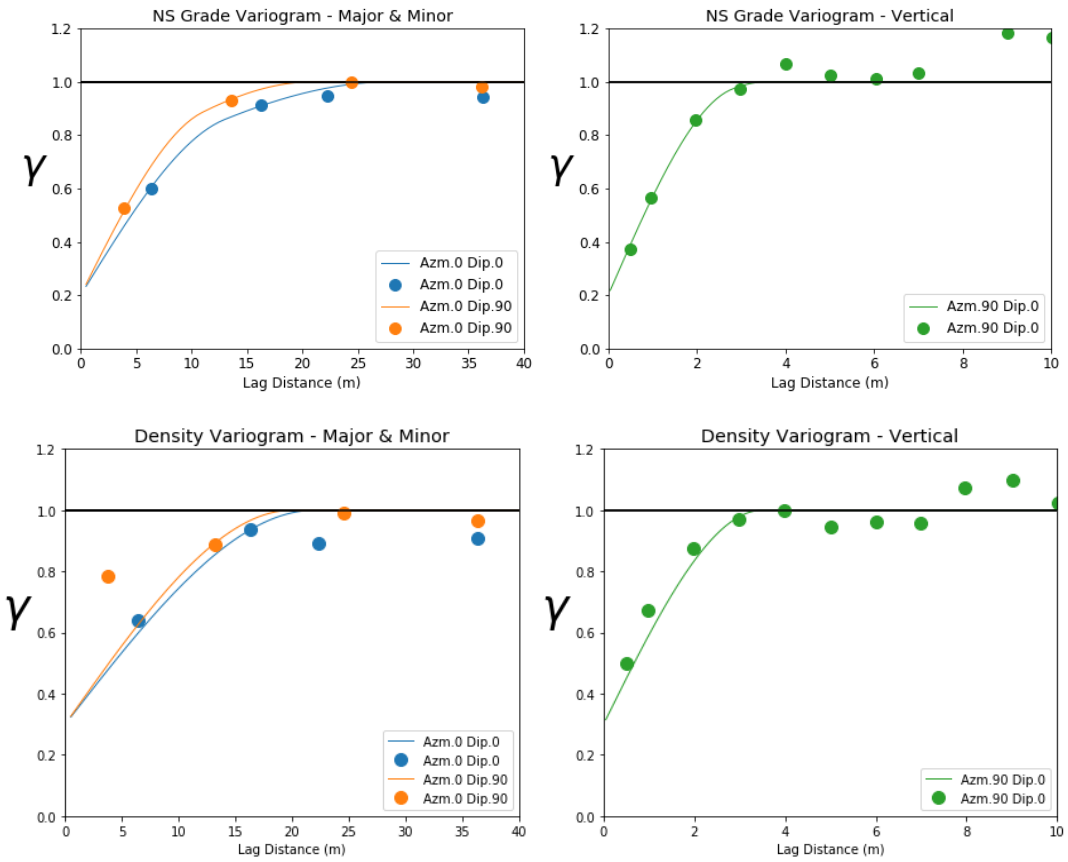

Figure 3.7: Experimental variograms and variogram models of the normal score grade (top) and normal score density (bottom) variables of the representative dataset.

After the prior distribution is created, the likelihood distribution needed to be defined. A bivariate plot was created to assess the relationship between the normal score grade and density variables, where it was apparent that the complex relationship between the two variables could not be captured with a single Gaussian model, a GMM is fit (Figures 3.8 and 3.9). A GMM is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions (i.e., components) with unknown parameters. The GMM fitted to the standard normal representative dataset used four components and an expectation maximization (EM) algorithm, which iteratively maximizes the log likelihood, as described in Chapter 2. The four-component GMM model was selected from sensitivity analysis that compared 2, 3, 4, and 5 components (Figure 3.10). The models that used 2 and 3 components underfitted the data and the model that used 5 components overfitted the data, as apparent in the bivariate and marginal distributions. The underfitted models failed to capture the complexity of the constraints observed in the data, particularly the reduced variance near the densest data cluster (NS Density $\approx$ 0.5 and NS Grade $\approx$ 0.5). The overfitted 5-component model contains a component that describes random noise in the data rather than the relationships between variables, trending from NS

Density ≈ 0.5 and NS Grade ≈ -3.5 to NS Density ≈ -2.5 and NS Grade ≈ 1.5. The selected 4-component GMM allows for the assessment of multi-dimensional probability distributions, and their conditionals and marginals distributions. The GMM represents the likelihood distribution in the imputation workflow.

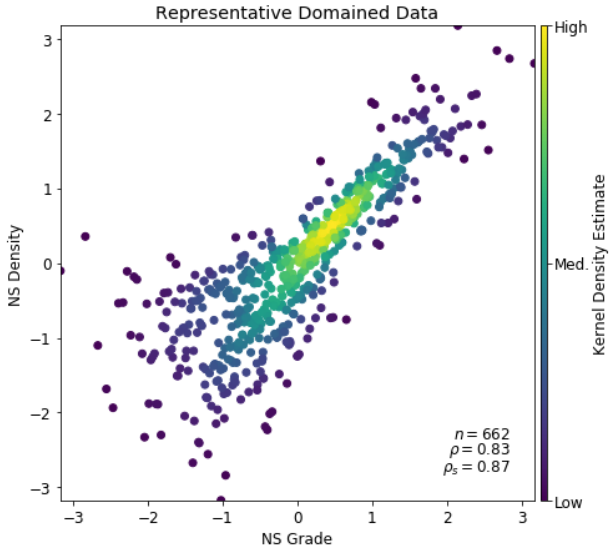

Figure 3.8: Bivariate plot of the representative dataset: normal score density values versus the normal score grade values.

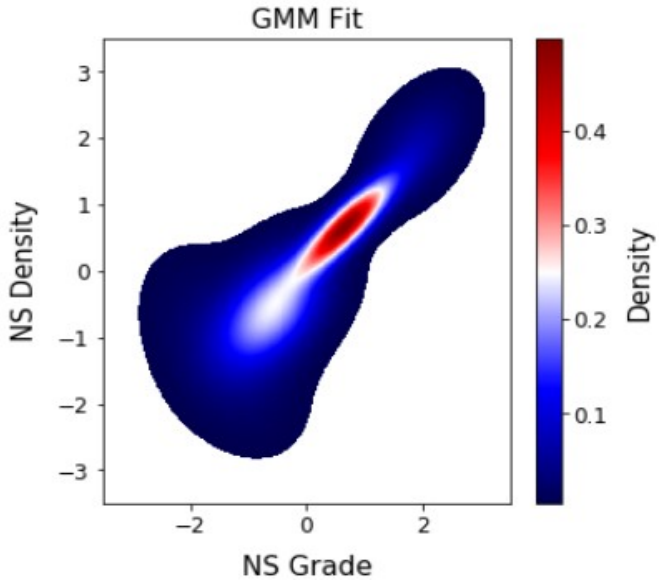

Figure 3.9: The GMM fitted to the homotopic representative dataset.

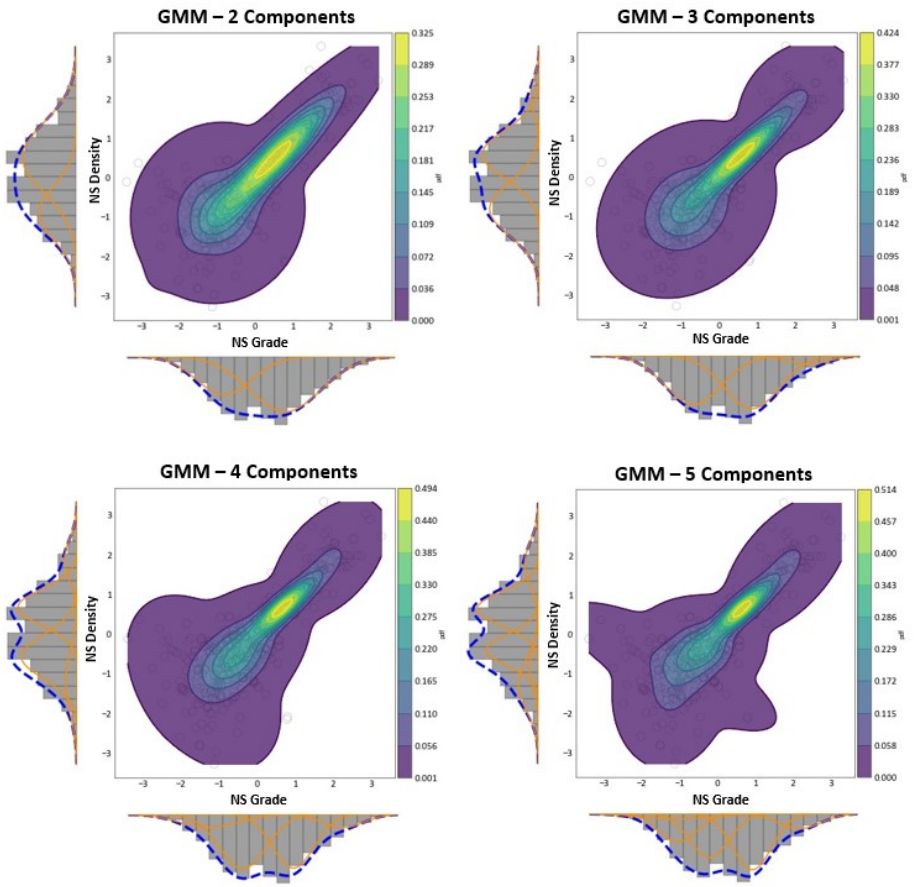

Figure 3.10: Bivariate Plot of the normal score representative data and the fitted GMM with different number of components.

The prior and likelihood distributions were combined to create an updated distribution to randomly sample from. For the imputation of missing density values, the measured density values (n = 116) from the normal scored representative data were retained and the "non-measured density values" (n = 540) were deleted. The "non-measured density values" are values that are in the representative dataset but not the original data set (i.e., the density values are part of the interim debiased homotopic representative dataset that used an imputation method from the global deposit wide dataset but are completely missing in the original domain dataset). Then the missing/deleted values were then imputed iteratively from the updated distributions to create a set of homotopic datasets. In total, 200 imputation realizations were created, where 82% of the density values were imputed in each realization. Figure 3.11 is an example of the imputed values of the first realization in normal score space relative to the bivariate relationship of the normal score representative dataset. The normal score realizations were checked for variogram and histogram reproduction (Figures 3.12, and 3.13).

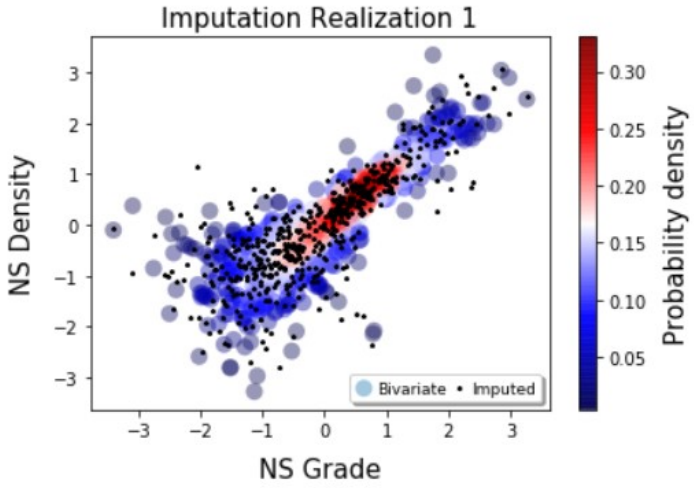

Figure 3.11: Bivariate plot of the imputed normal score density values in realization 1, underlain with data values of representative dataset.

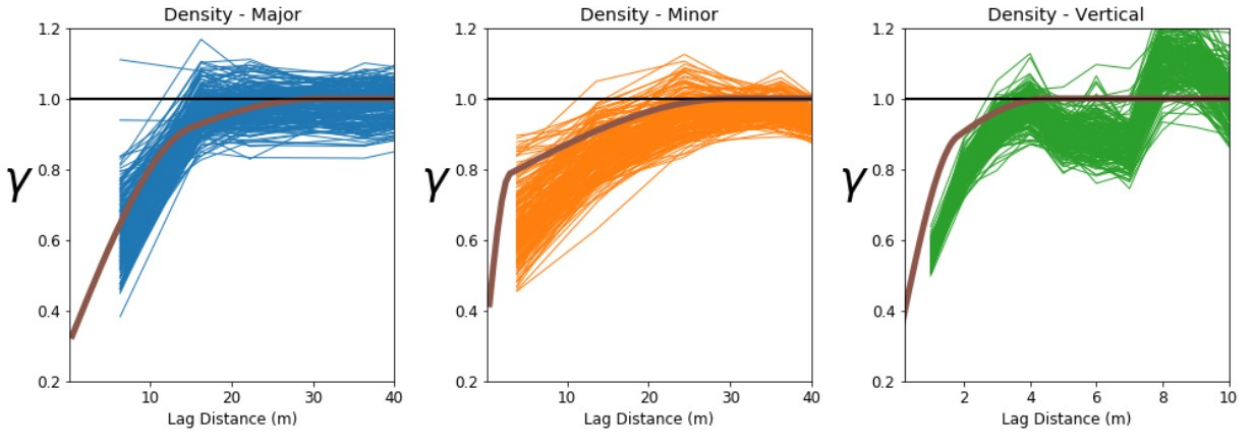

Figure 3.12: Variogram reproduction of imputed density values overlain by representative density variogram.

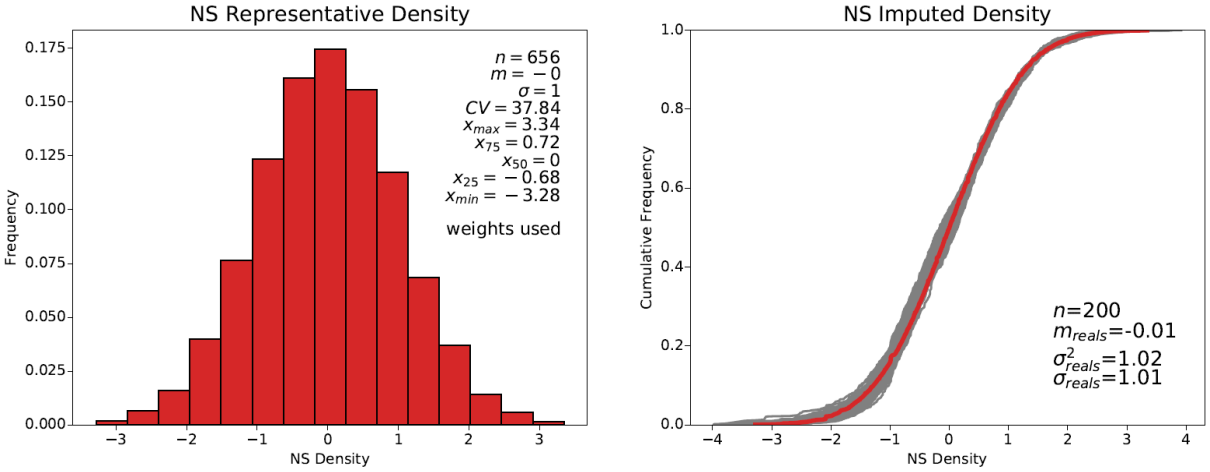

Figure 3.13: Histogram reproduction of imputed density values in normal score space.

In general, the normal score density variograms were well reproduced, but the imputed values did not replicate the short-range variability in the minor or vertical directions. This was determined to be non-consequential due to the extreme short-range variability in the minor direction; a range difference of 1-2 m likely is not biasing the model significantly. The representative density histogram is well replicated by the imputed realizations, but there is a slight bias, which is believed to derive from the GMM model not fully capturing the variability of the representative dataset. The bias is negligible, and the realizations were considered acceptable for further work. The realizations were then backtransformed to original units to check for appropriateness: Figure 3.14 shows the grade and density distributions, as well as a scatter plot of density versus grade of the first two realizations and Figure 3.15 shows the histogram reproduction of the density variable in original units. Note that no grade values were imputed, therefore the grade distribution is the same for every realization. The backtransformed variables preserve the observed relationship noted in the representative database and the realizations replicate the original units reasonably well, but the negligible bias is still present.

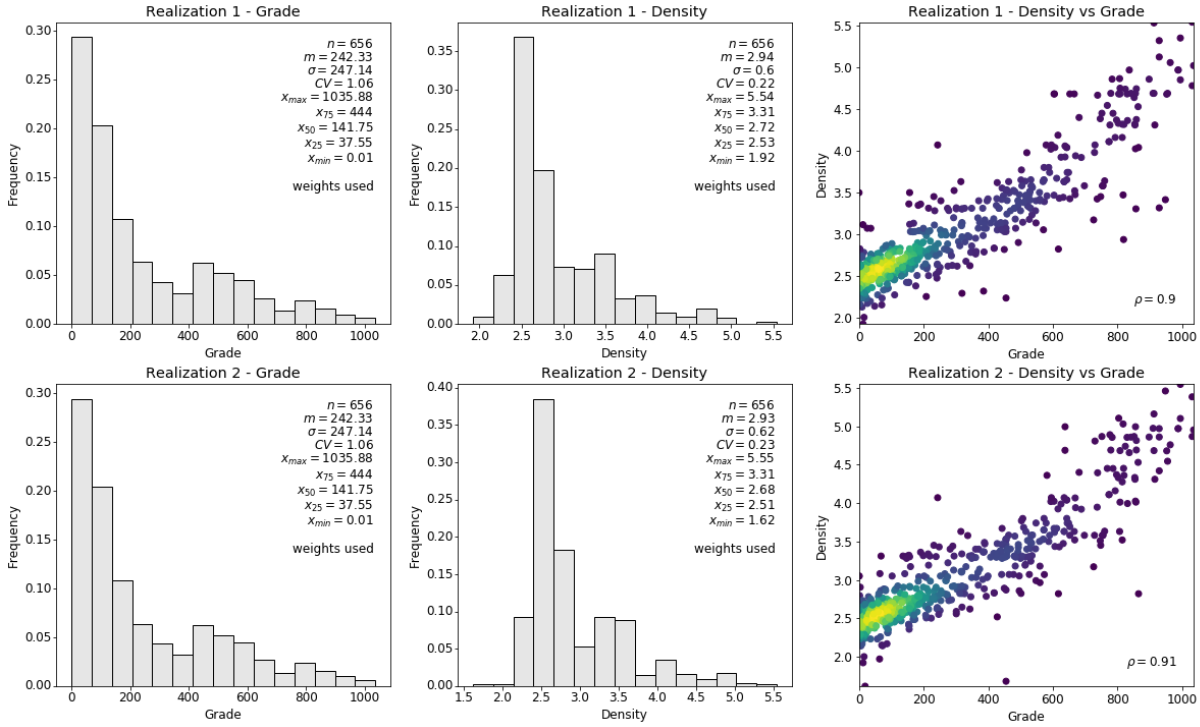

Figure 3.14: Grade (Au g/t) and density (S.G.) histograms, and a scatter plot of density versus grade for the first two imputed realization.

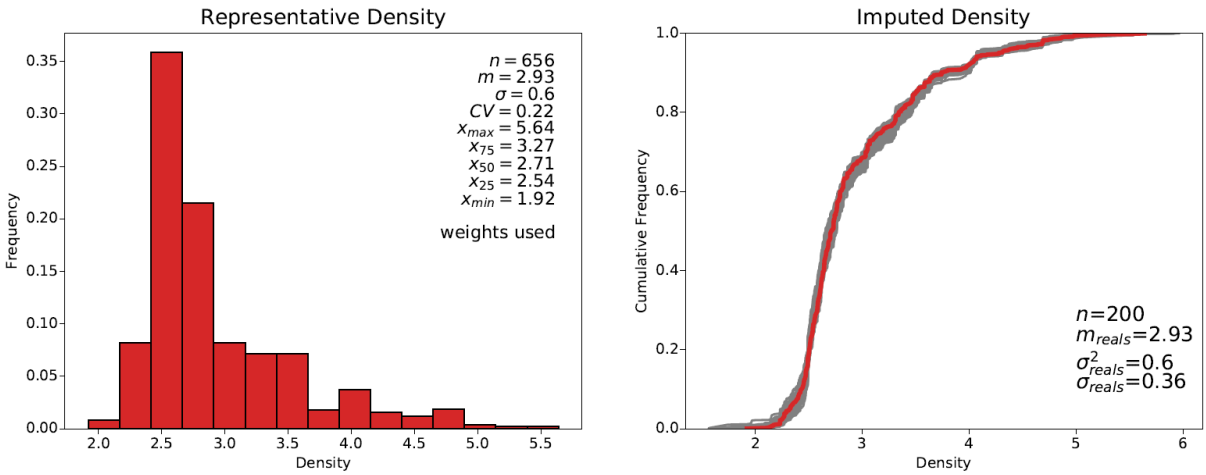

Figure 3.15: Histogram reproduction of imputed density values in original units (S.G.).

After a simple debiasing technique was implemented using site-specific data, the GMM imputation workflow transferred the uncertainty associated with missing values into realizations. Excluding the heterotopic data observation would lead to biases in the model, as the missing density data observations were not missing at random.

3.4 PPMT and Independent Simulation

The output of the GMM imputation workflow contained multivariate complexities that required the use of the non-linear PPMT. The PPMT transforms data of virtually any form, size, and dimension to an uncorrelated multivariate Gaussian distribution, allowing for independent simulation of the variables (Barnett & Deutsch, 2015a). The PPMT consists of two steps after a univariate normal score transformation: data sphereing and the implementation of the projection pursuit algorithm. The data sphereing is applied to transform the Gaussian units to be uncorrelated with unit variance. The projection pursuit algorithm is based on the projection index, a test statistic for non-Gaussianity. An optimized search finds the projection of the sphered data that maximizes the projection index (the most non-Gaussian projection). The Gaussianization transform of the multivariate data makes that projection Gaussian. Iterating this search and Gaussianization procedure, the sphered data is transformed to be multi-Gaussian. The PPMT was completed on each dataset realization to fully decorrelate the variables, Figure 3.16 is the transformed values of the first realization. The targeted Gaussianity percentage for the PPMT was set to 99% with a maximum of 50 iterations.

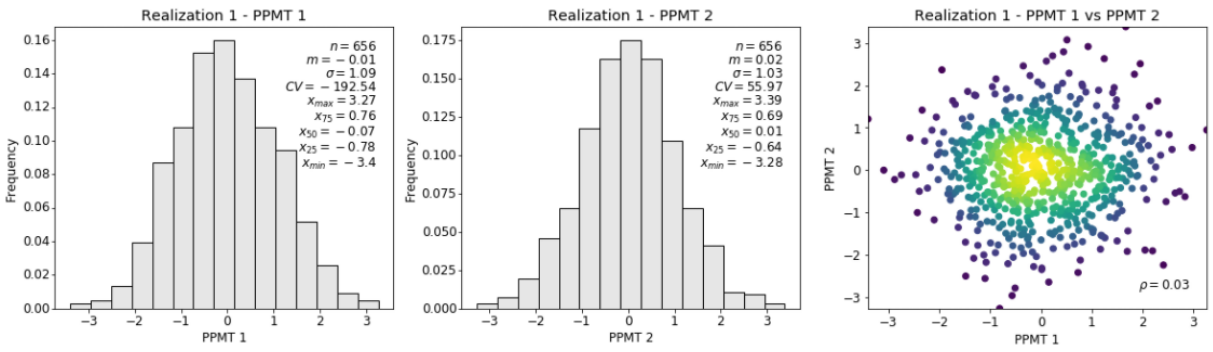

Figure 3.16: Histograms and scatterplot of PPMT values in Realization 1.

Variograms were automatically fitted to the PPMT realizations in the CCG varmodel program. Figures 3.17 and 3.18 are the variogram models of the first two realizations for each variable, now referred to as PPMT 1 and PPMT 2. The variogram structure was nearly lost after the decorrelation of the highly correlated variables.

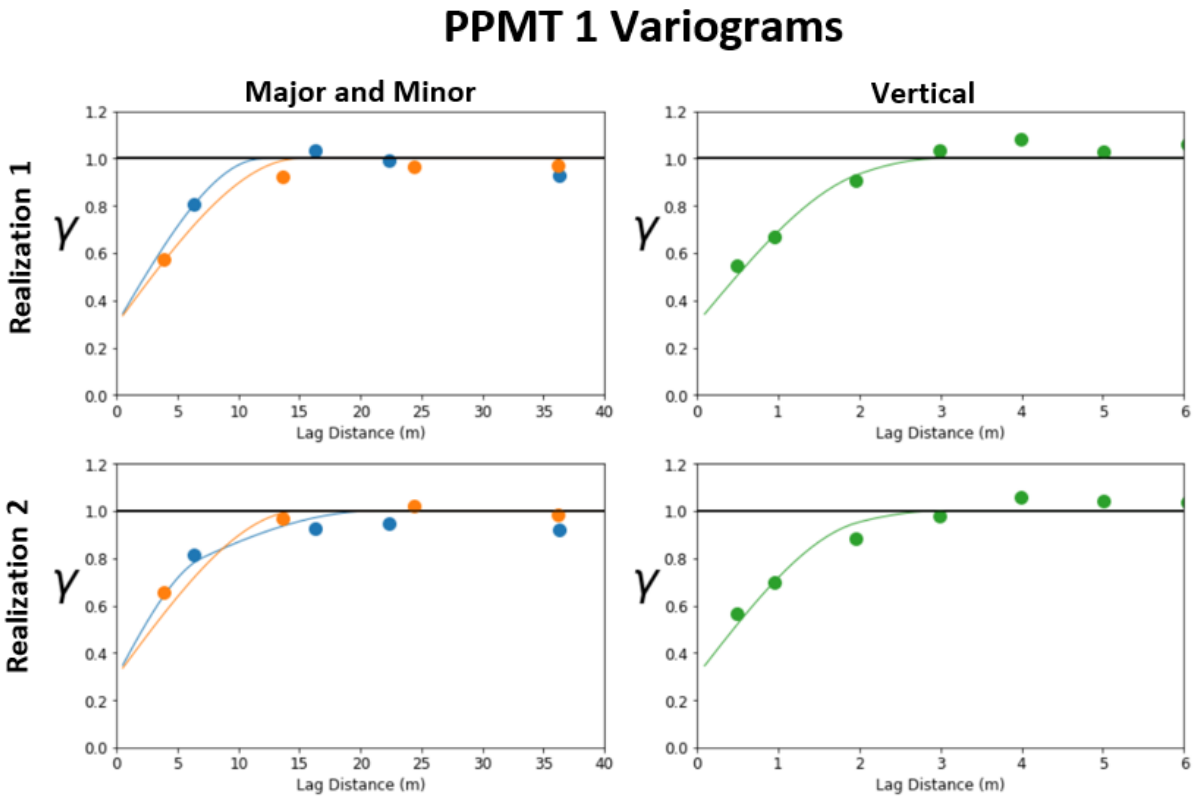

Figure 3.17: Auto-fit variograms to the PPMT 1 variable in realizations 1 and 2. Major, minor and vertical directions shown.

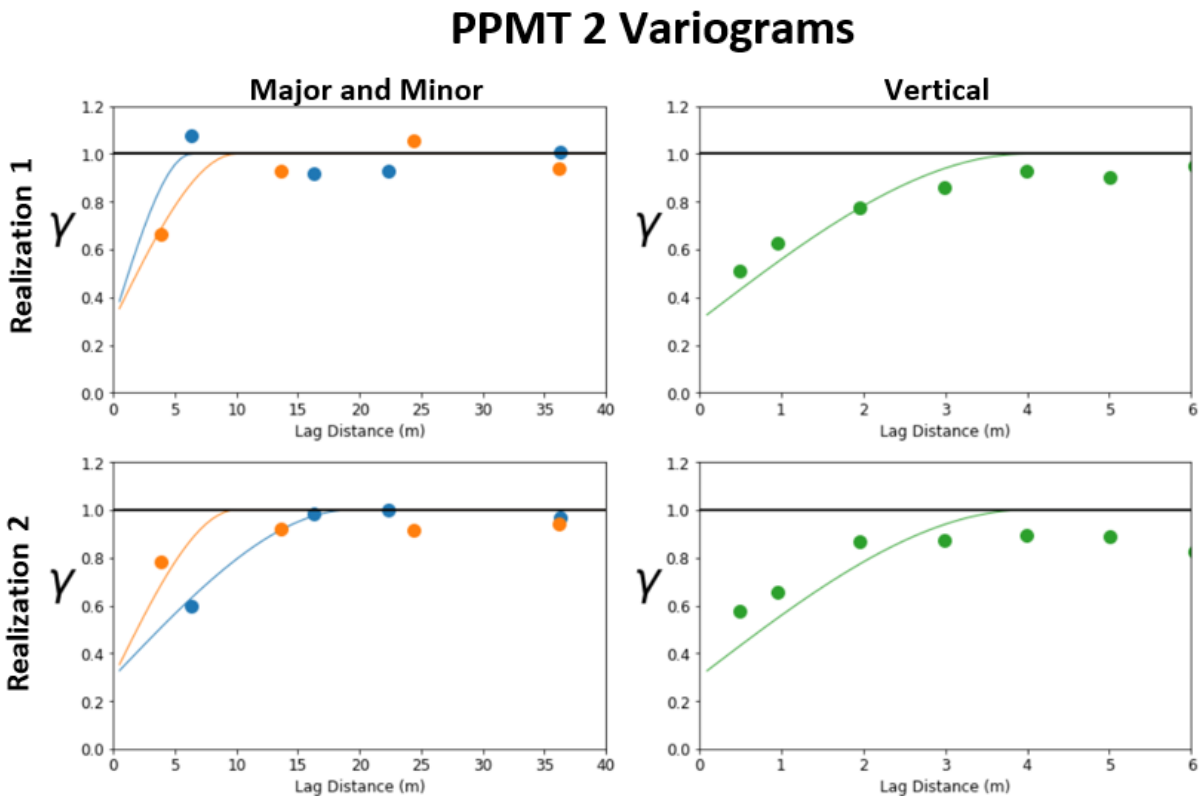

Figure 3.18: Auto-fit variograms to the PPMT 2 variable in realizations 1 and 2. Major, minor and vertical directions shown.

A 68 (x direction) by 320 (y direction) by 400 (z direction) grid of 0.5 m blocks was simulated using the PPMT values and the auto-fitted variograms. Each SGS realization was simulated using a unique decorrelated dataset and a unique PPMT 1 and PPMT 2 variogram fitted to that dataset. The resulting simulated grid was clipped to the domain extents and viewed in Cartesian space for reasonableness. The representative dataset underwent PPMT to check the estimated grid (Figure 3.19). The estimated PPMT values in Cartesian space appear to replicate the informing data reasonably well, where locations of high and low values in the representative dataset are apparent in the simulated grid. The first two realizations were checked to ensure that the grid estimate is decorrelated as per the informing data (Figure 3.20). Variograms were calculated for each realization of the grid simulated model and compared to the average input variogram; it appears that the simulated model reasonably replicates the input variograms for PPMT 1 and PPMT 2, particularly in the shorter distances (Figures 3.21). However, in the minor direction, it appears that the model is slightly more continuous than the input variograms. Histograms were checked for data replication in PPMT space – the simulated model appears to replicate the informing data well, although not perfectly (Figure 3.22). The imperfect replication is likely partially due to the fact that the representative data is not the exact informing data for the model (the imputed realizations were used), and there would be slight differences during the projection pursuit transform.

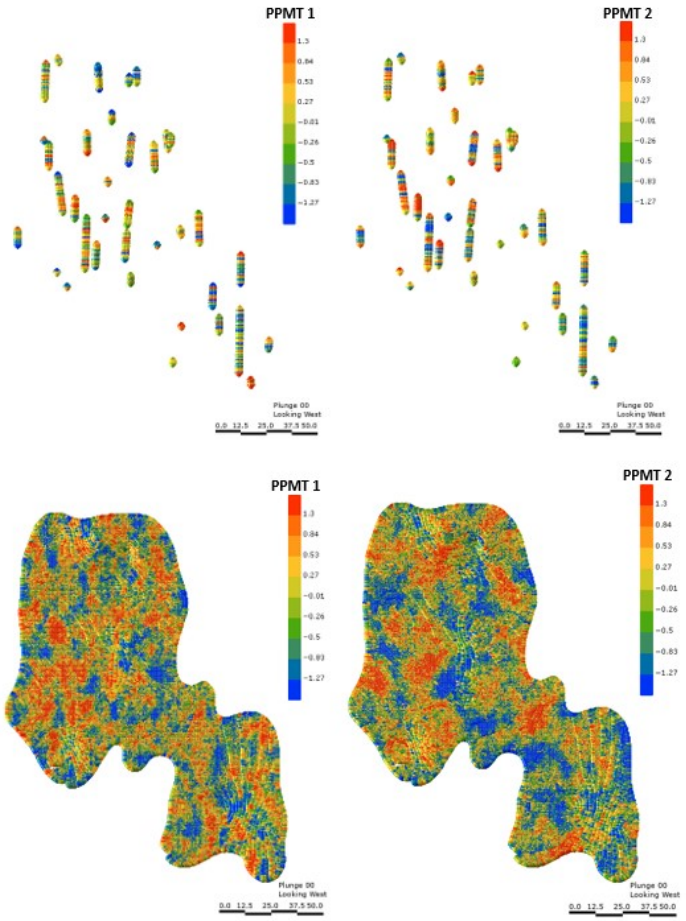

Figure 3.19: Realizations 1/2 of PPMT values used in SGS and the corresponding SGS realizations.

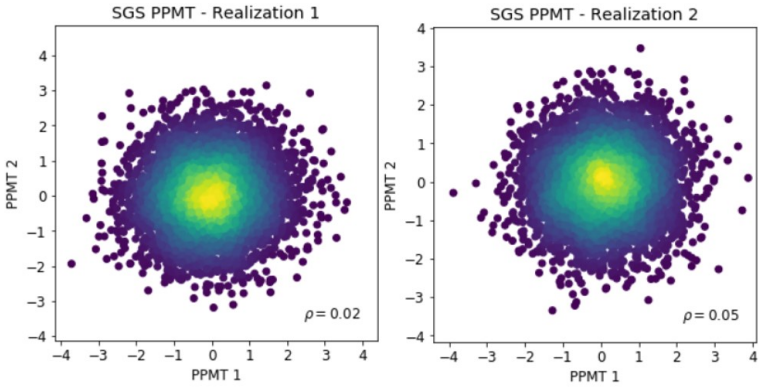

Figure 3.20: Scatterplots of PPMT realizations 1 and 2 from grid SGS.

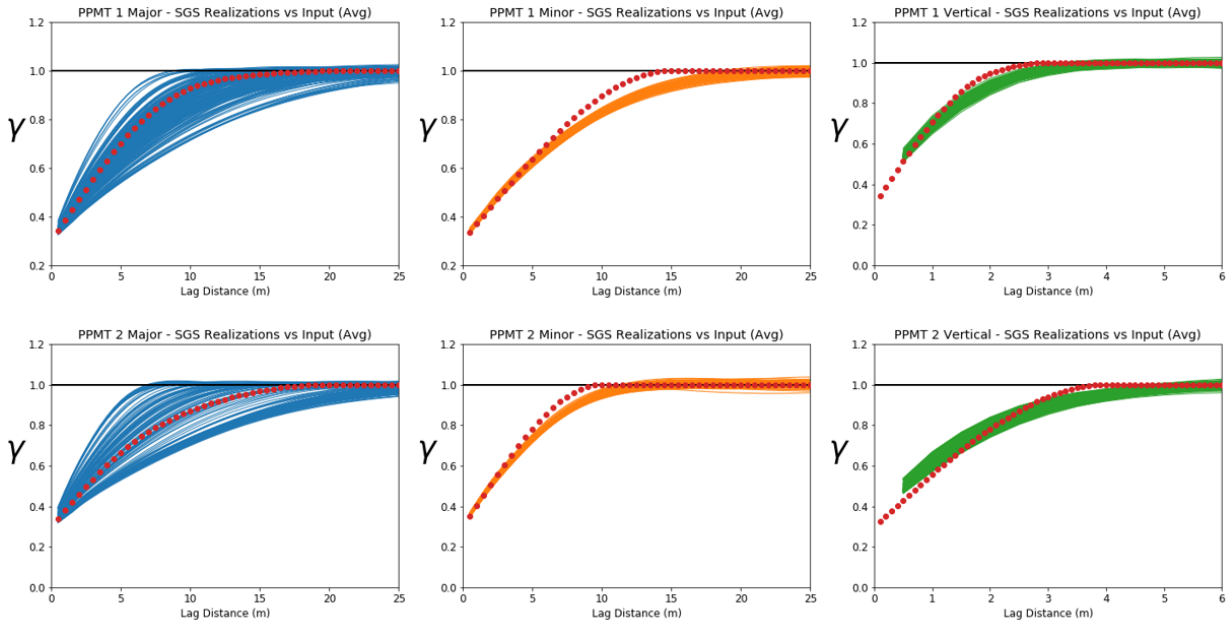

Figure 3.21: Calculated SGS variograms for each realization of PPMT 1 and PPMT 2 overlain by the average input variogram.

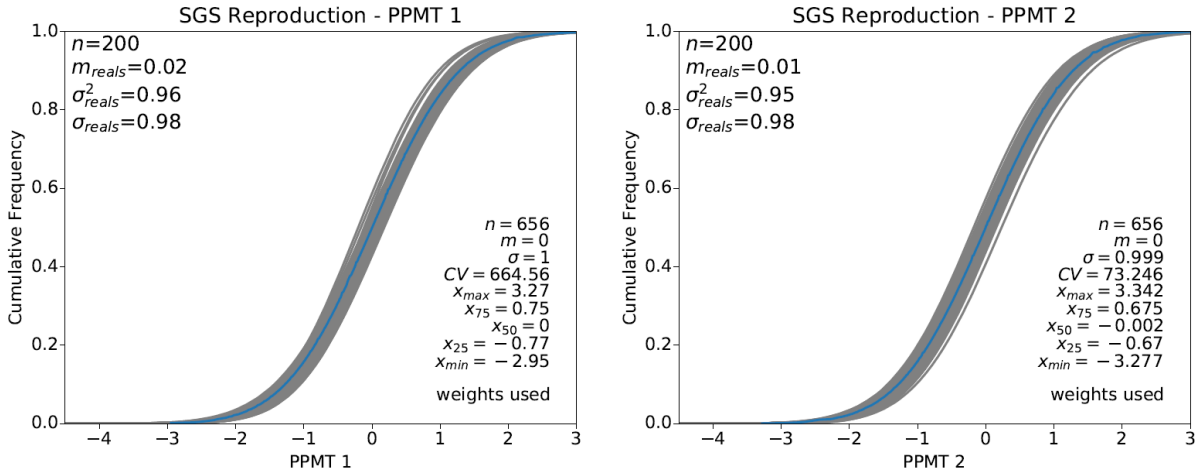

Figure 3.22: Calculated SGS histograms for each realization of PPMT 1 and PPMT 2 overlain by the representative histogram.

The clipped grid was back transformed to original units and the realizations were reviewed in Cartesian space; again, the estimated values appear to replicate the informing data reasonably well (Figures 3.23). The grade and density histograms of all the realizations were compared to the representative dataset – the realizations reasonably replicate the informing data (Figure 3.24). The difference between the representative dataset and the backtransformed model is small, with less than 5 % error between the two, and thus the model is deemed acceptable.

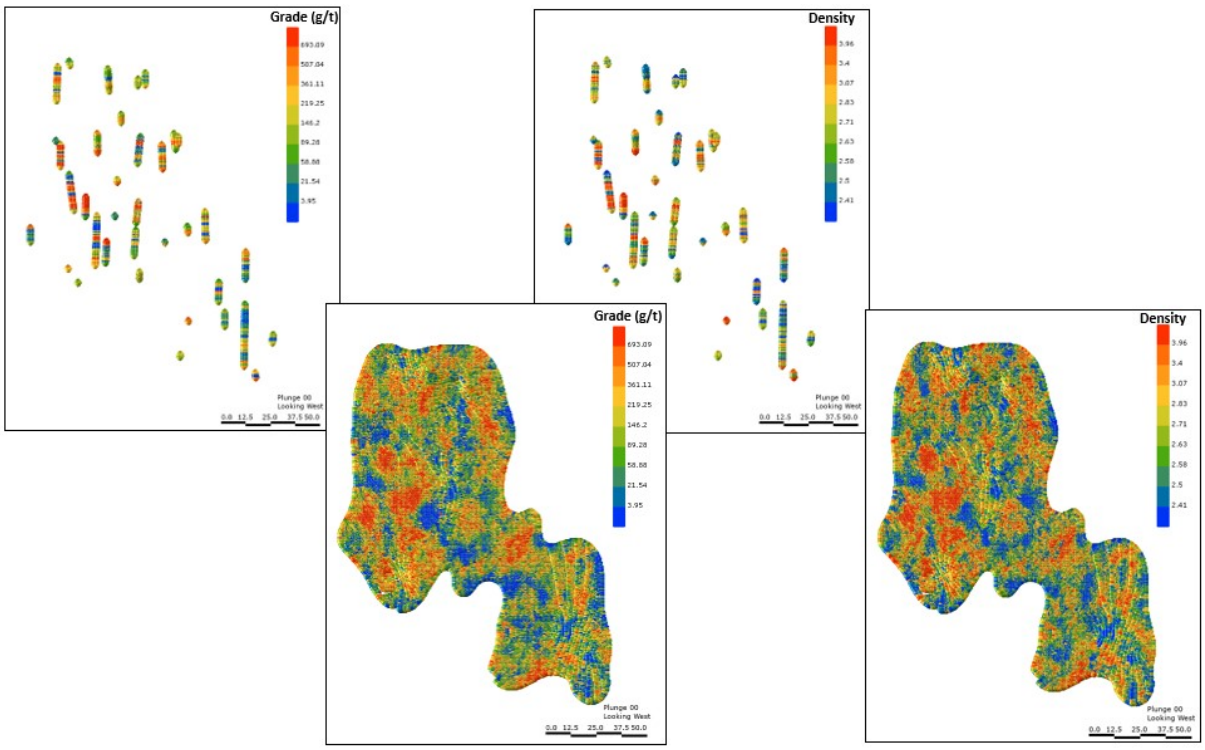

Figure 3.23: The representative dataset (upper) and the back transformed grade and density grid of realization 1 (lower).

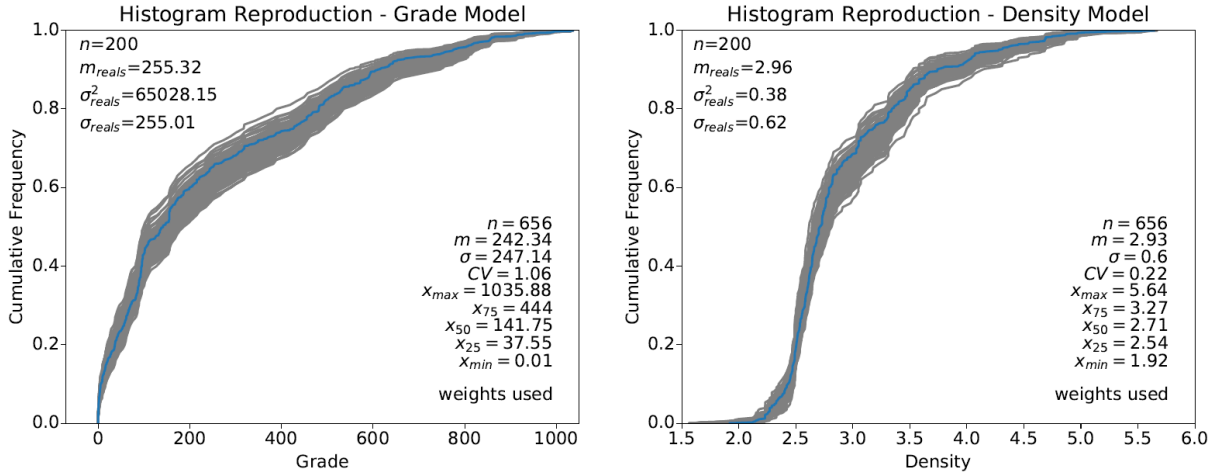

Figure 3.24: Calculated histograms for each realization of the grade and density variables overlain by the representative histograms.

The E-type mean was calculated and reviewed to ensure that it was similar to the representative dataset (Figure 3.25). The average density and grade of each realization was calculated for the study domain; the average grade ranges from 204.22 g/t to 290.23 g/t with a mean of 255.32 g/t, and the average density distribution ranges from 2.83 g/cm^3 to 3.09 g/cm^3 with a mean of 2.96 g/cm^3 (Figure 3.26).

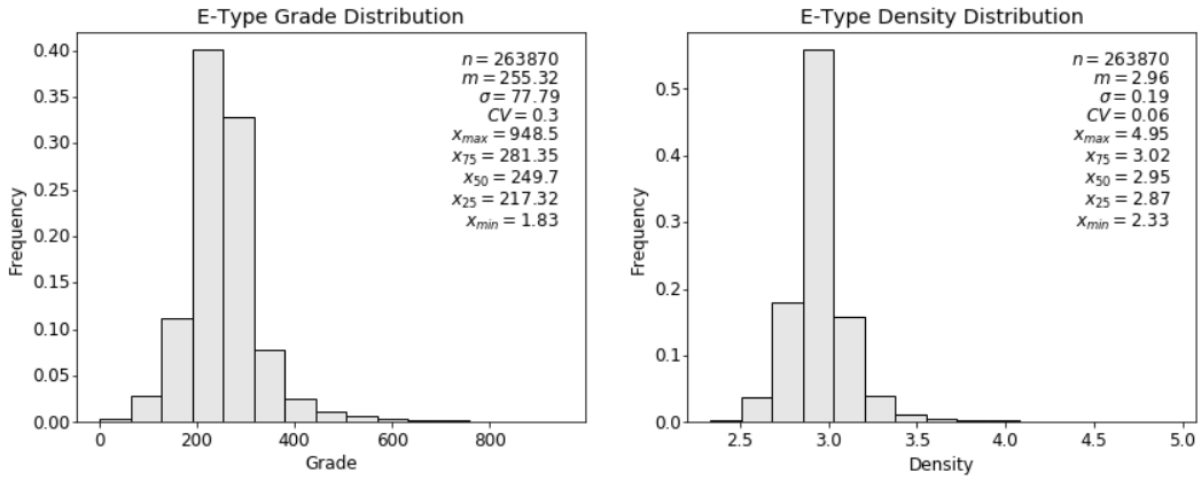

Figure 3.25: Calculated histograms for each realization of the grade (Au g/t) and density (S.G.) variables overlain by the representative histograms.

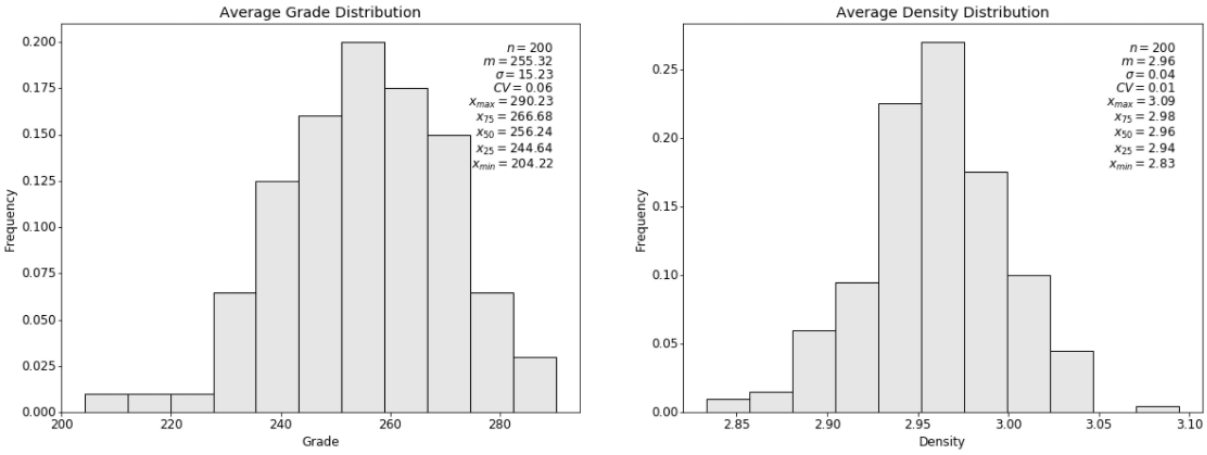

Figure 3.26: Average grade (Au g/t) and density (S.G.) distributions within study domain.

3.5 Multivariate Spatial Bootstrap

Bootstrapping is a statistical resampling scheme that allows uncertainty in an input statistical parameter to be assessed from the data (Rossi & Deutsch, 2014). The multivariate spatial bootstrap is a resampling scheme that considers the spatial correlation between the variables at data locations (Rezvandehy & Deutsch, 2016). The technique is used to create realizations representing the prior uncertainty, and the realizations are used as reference distributions for the conditional simulation, an uncertainty updating process to produce the posterior uncertainty (Vincent & Deutsch, 2019). The application of the process to the Arrow Deposit is described below.

The density and grade variables of the representative dataset are transformed to Gaussian units through a quantile transform (Figures 3.27 and 3.28). A Linear Model of Coregionalization (LMC) is created by modeling the direct variograms of the gaussian density and grade variables and the corresponding cross

variogram, while maintaining a positive definite covariance matrix (Figures 3.29, 3.30, and 3.31). The direct and cross variograms have a nugget effect of 0.2 and two spherical structures. The ranges and directions for the structures in all three variograms are identical, but the contributions to the structures differ for each variogram (Table 3.1). The sill of the cross variogram is equal to the correlation between grade and density, $\rho = 0.82$.

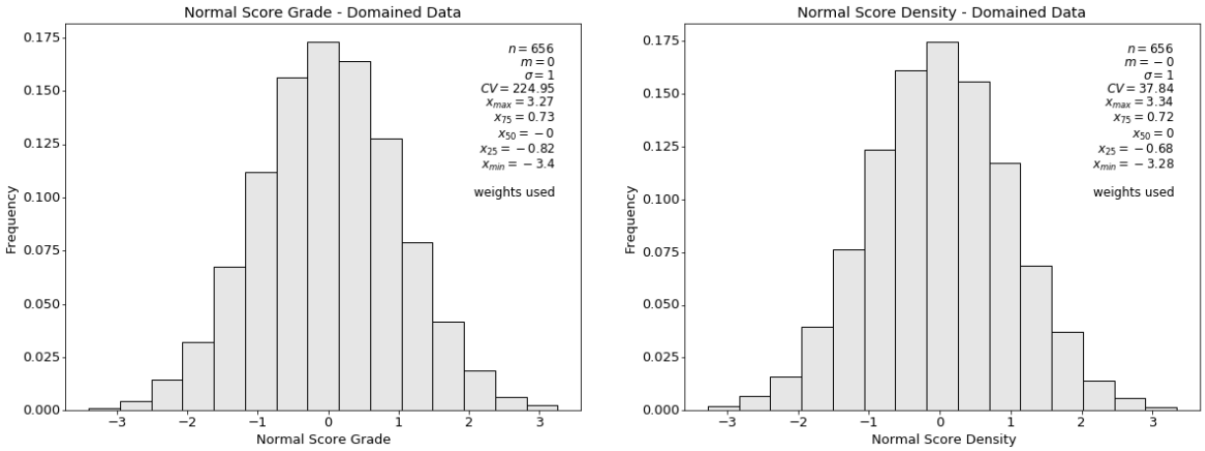

Figure 3.27: Histograms of the grade (Au g/t) and density (S.G.) variables in Gaussian units.

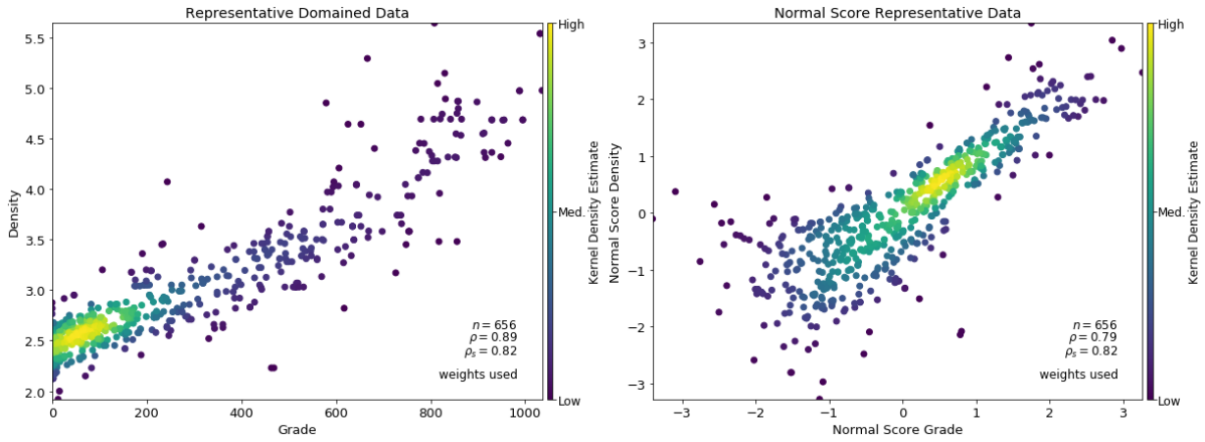

Figure 3.28: Scatterplots of the density (S.G.)and grade (Au g/t) variables in original and Gaussian units.

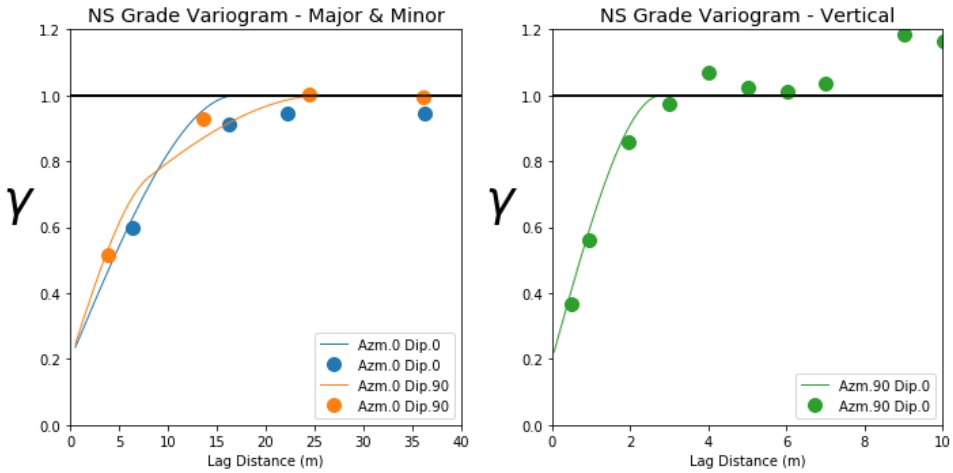

Figure 3.29: Modeled direct variogram of the normal score grade variable

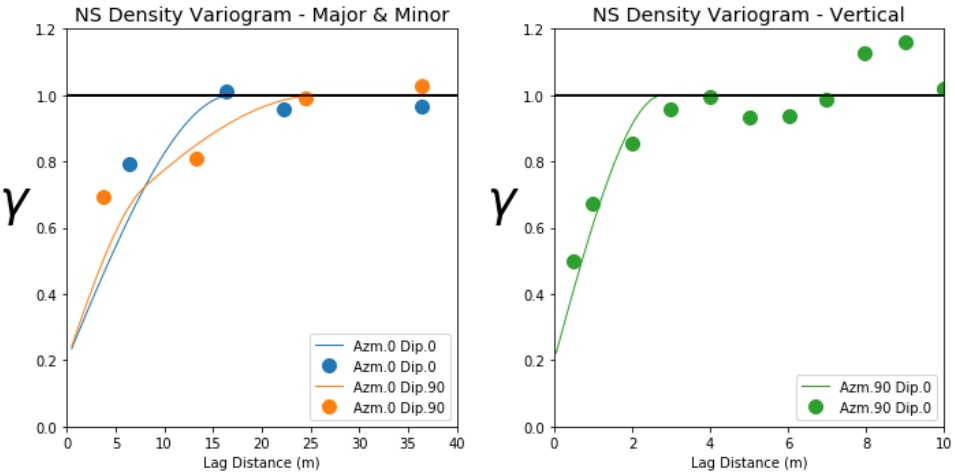

Figure 3.30: Modeled direct variogram of the normal score density variable

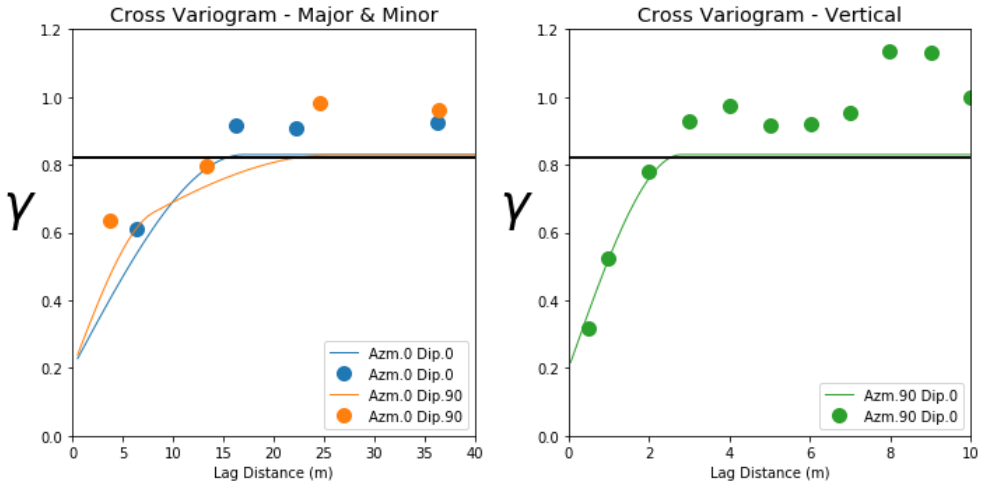

Figure 3.31: Modeled cross variogram of the normal score variables.

Structure	Contribution (NS Grade)	Contribution (NS Density)	Contribution (Cross)	Major Range	Minor Range	Vertical Range
C_0	0.20	0.20	0.20	-	-	-
C_1	0.30	0.55	0.40	13.0	5.0	2.8
C_2	0.50	0.25	0.22	31.0	21.0	2.8

Table 3-1: Direct and cross variogram structure contributions and ranges.

A Cholesky LU decomposition of the covariance matrix is performed, $C = LL^T$, where C is the covariance matrix, L is the lower triangular matrix and L^T is the transpose of the lower triangular matrix (Rezvandehy & Deutsch, 2016). Two hundred unconditional realizations with 656 collocated samples are produced by multiplying the L matrix by w, an uncorrelated standard normal deviate vector, for each realization (Rezvandehy & Deutsch, 2016). The realizations are back transformed to original units using the original transformation tables and the distributions of uncertainty were generated for grade and density (Figures 3.32 & 3.33). Spatial correlation directly affects prior uncertainty, increasing spatial correlation leads to greater uncertainty because the data are more redundant (Khan, Deutsch, & Deutsch, 2014).

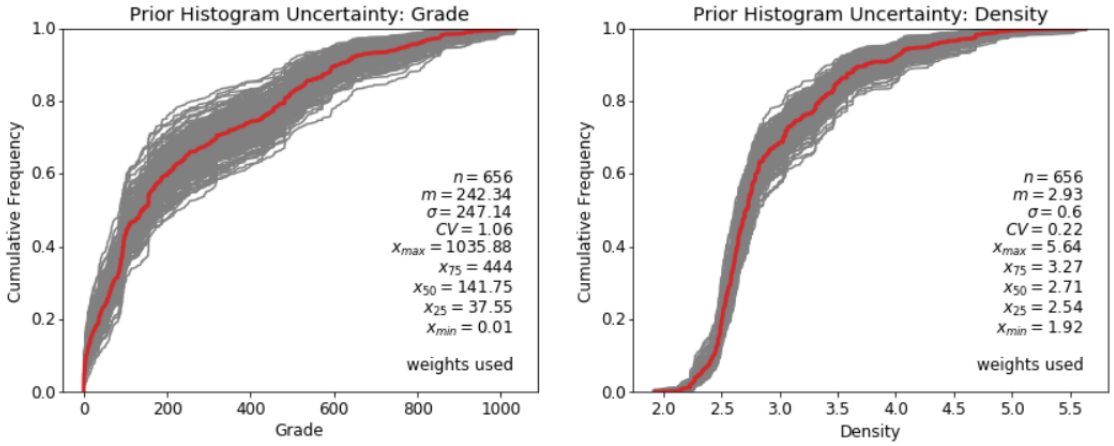

Figure 3.32: Prior histogram uncertainty of grade (Au g/t) and density (S.G.)

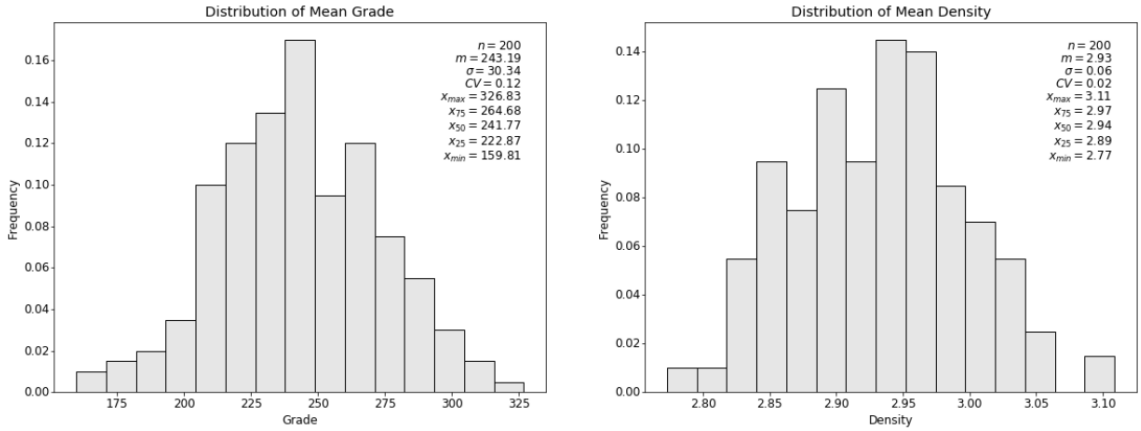

Figure 3.33: Distribution of mean grade (Au g/t) and density (S.G.) of prior distribution.

The realizations of the multivariate spatial bootstrap are used as reference distributions during the normal score transform of the representative data (Figure 3.34).

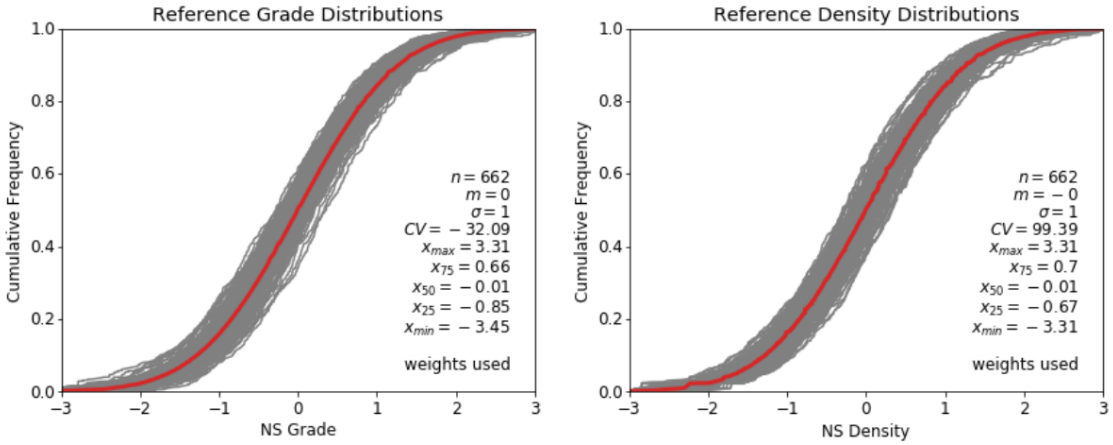

Figure 3.34: Normal score transformation of grade (Au g/t) and density (S.G.) variables using prior distributions as reference distributions.

The density variable is simulated independently at the grid locations using the previously modeled density variogram, and the grade variable is simulated using the grade variogram model and the simulated density realizations as a secondary data during co-simulation. The simulated grid is clipped to the domain dimensions, then the realizations were back transformed to the original units using the spatial bootstrap reference distributions (Figure 3.35). The back transformed grid was checked for histogram reproduction (Figure 3.36).

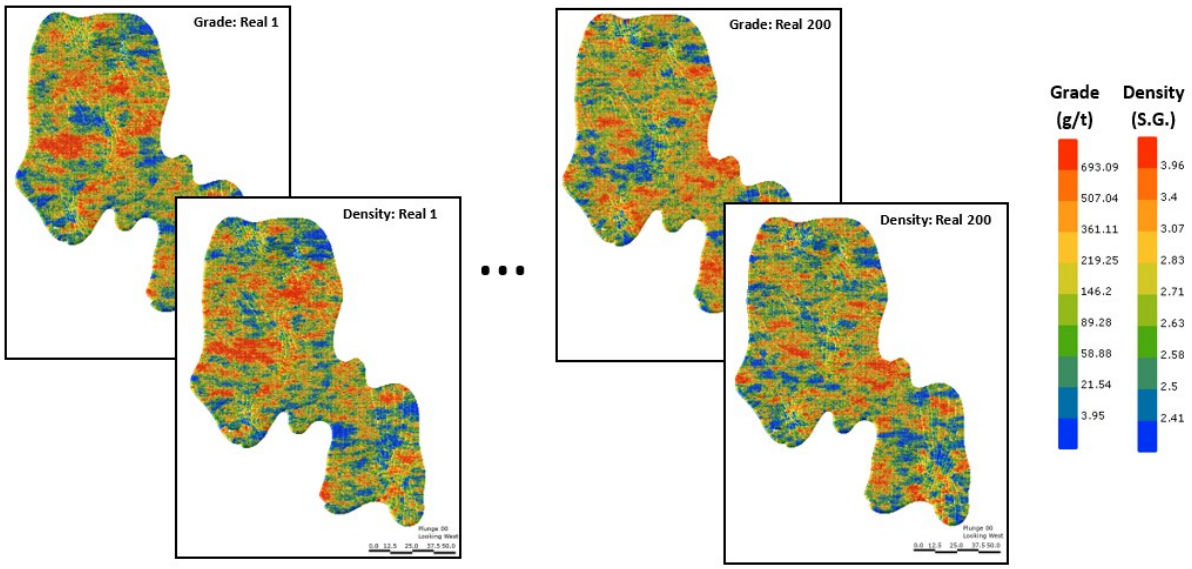

Figure 3.35: Realizations of the grade (Au g/t) and density (S.G.) variables using prior distributions as reference.

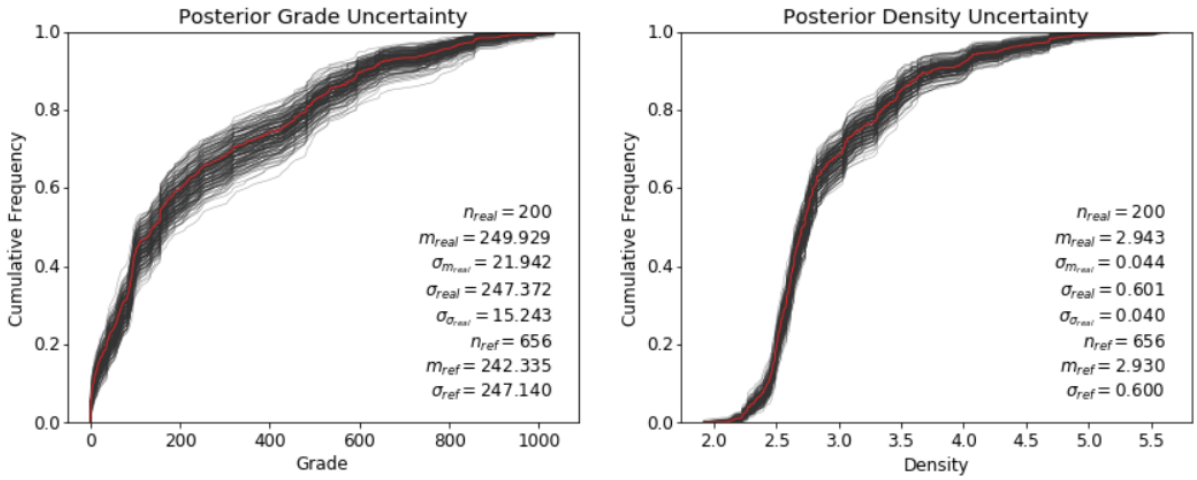

Figure 3.36: Histogram reproduction of grade (Au g/t) and density (S.G.) variable in original units.

Notably, the posterior uncertainty is only slightly reduced from the prior uncertainty (Figure 3.37); therefore, the conditioning data in its configuration in the study domain has little effect on the global uncertainty of the grade and density variables.

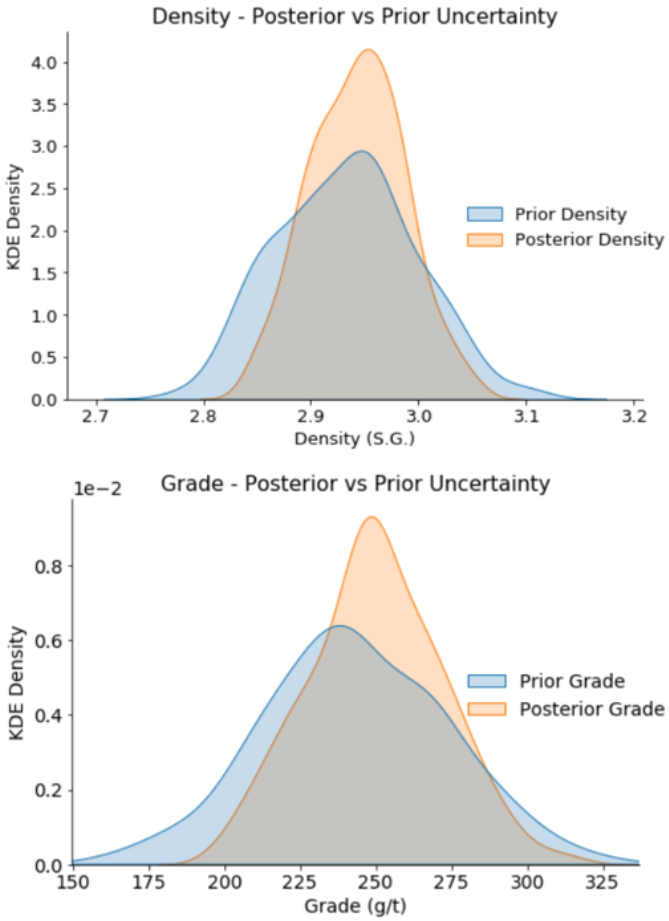

Figure 3.37: Comparison of prior and posterior mean uncertainty of density (upper) and grade (lower).

Chapter 4

4. Geometric Imputation of Domain

Carvalho (2018) developed a technique for assessing domain geometry uncertainty in tabular vein deposits. The workflow considers a local coordinate system, position uncertainty (i.e., vein location perpendicular to the intercept-point) and thickness uncertainty. In the workflow, drill holes with shallow angle intersections are imputed to assess position and thickness uncertainty. Position and thickness uncertainties, or geometry data uncertainty, are in part due to inclined drill holes intercepting the domain, where true thickness and vein location perpendicular to the intercept-point are not measured. The workflow generated multiple realizations of the vein geometry and boundaries with unstructured grids that fit the shape of the vein (Carvalho, 2018). Prior to development of geometry uncertainty workflows, estimation domains for vein deposits were usually deterministically defined by explicit modeling, implicit modeling, or surface interpolation, which do not provide uncertainty assessment (Carvalho, 2018).

At the Arrow Deposit, all the informing drill holes intersect the vertical deposit obliquely, thus requiring imputation for uncertainty analysis. Note, that unlike the workflow Carvalho proposed, only thicknesses were considered in the evaluation as that was the primary interest in the deposit. Position uncertainty was not considered in this thesis. The uncertainty of the domain thickness perpendicular to the plane of continuity is defined as “FW/HW” (footwall/hangingwall) uncertainty.

4.1 Data preparation

Nearly all of the drill holes delineating the vertical, northeast striking study domain are oriented in two main directions: 327- and 147-degrees azimuth, dipping between 60 and 80 degrees; however, there is one drill hole drilled vertically through the domain. The data and domain were rotated to a north-south orientation so that coordiniate directions are more geologically significant and to reduce the grid size (Figures 4.1 & 4.2). The domain contains 35 drill holes spaced 0.5 to 25 m apart and the deterministically explicitly modeled wireframe representing the domain ranges in thicknesses between 1 and 7m with a maximum vertical extent of 175 m along a strike of 145 m (Figure 4.3). The one drill hole drilled vertically through the domain was not considered in the imputation workflow described below.

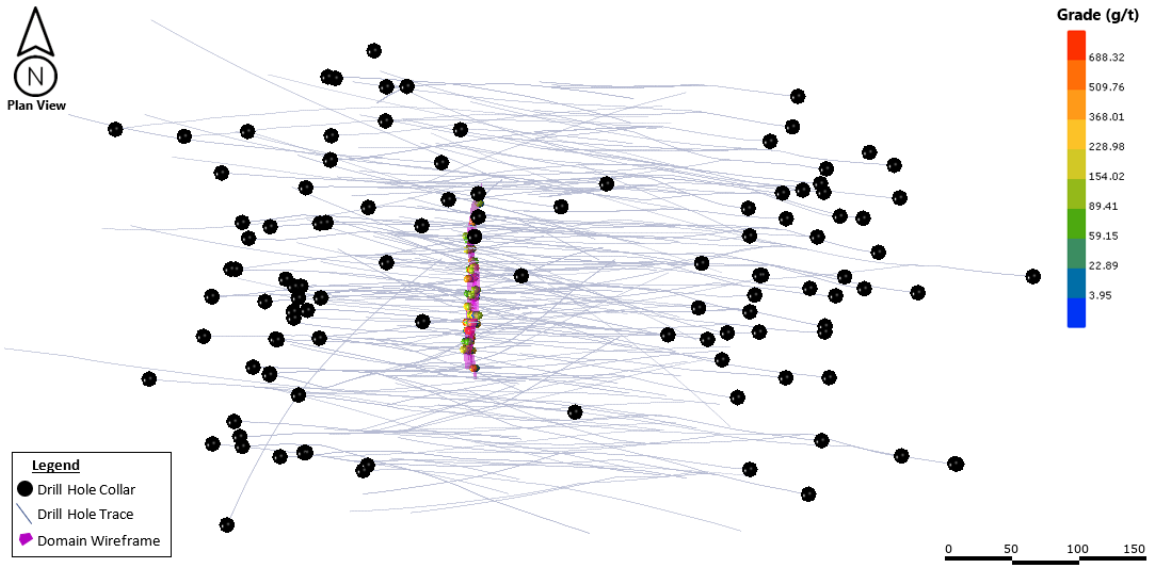

Figure 4.1: Plan view of the drill holes constraining the domain, the domain wireframe, and domain composites displaying assay grade (Au g/t) results. Scale in metres.

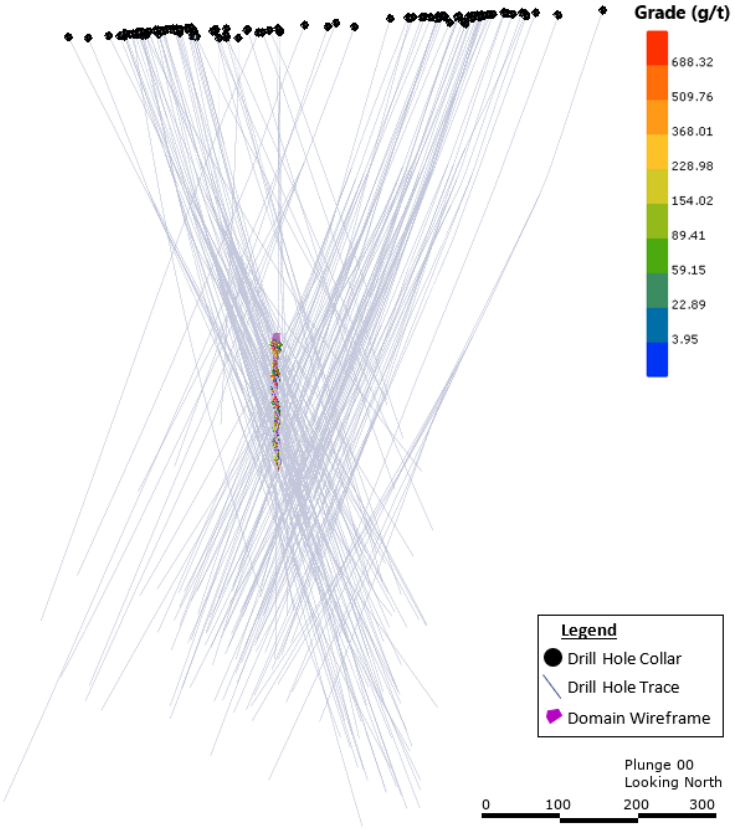

Figure 4.2: Oblique view (looking north) of the drill holes constraining the domain, the domain wireframe, and domain composites displaying assay grade (Au g/t) results. Scale in metres.

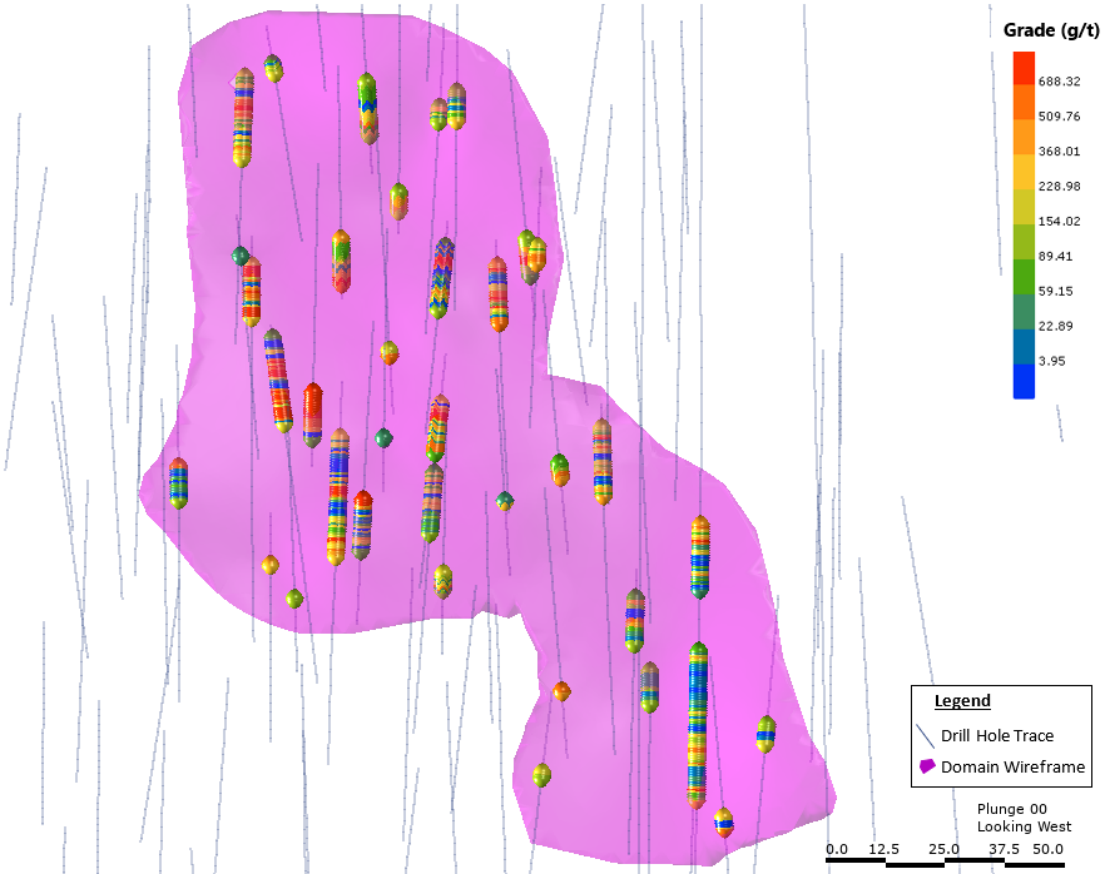

Figure 4.3: Long section (20 m window looking west) of the drill holes constraining the domain, the deterministic domain wireframe (pink vein created in Leapfrog), and domain composites displaying assay grade (Au g/t) results. Scale in metres.

The inclined drill holes intersect the domain mineralization at an oblique angle; thus, the length of the mineralized interval along the drill hole is not a true representation of the thickness of the domain. In lieu of true thickness measurements, a database of interim thickness is created by calculating the distance between the intersection of the drill hole with the wireframe and the wireframe location opposite to the intersection (Figure 4.4). The assessed thickness is assigned to the wireframe intersection coordinate. The declustered average thickness is 2.2 m with values ranging between 0.36 m and 6.29 m (Figure 4.4). As expected, the lower values are located at the peripheral edges of the wireframe and at minor bends in the wireframe. The composite grades and lengths have no appreciable relationship with one another (Figure 4.4). The use of the wireframe and drill hole intersection coordinates to create thickness values - although not a direct observation from raw data - imparts geologic knowledge of the deposit onto the workflow. Each assessed thickness value will be replaced through the simulation workflow and the uncertainty will be assessed.

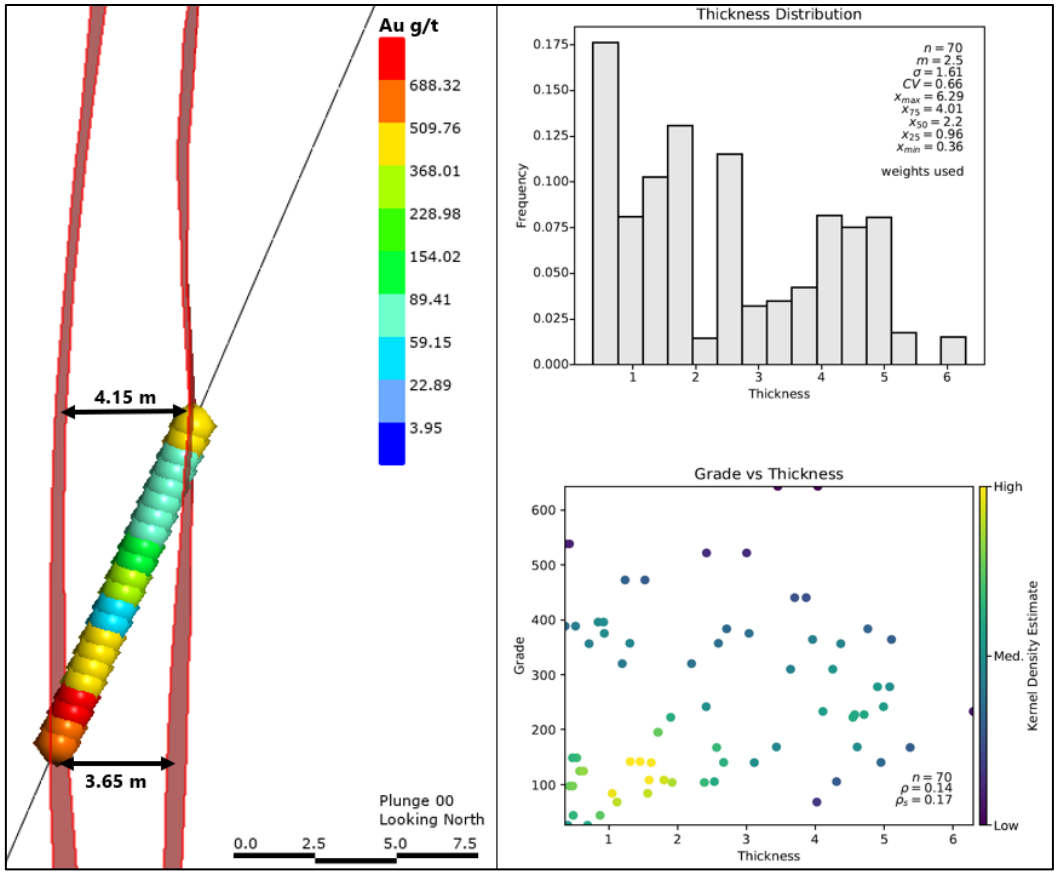

Figure 4.4: Cross section of the deterministic wireframe, a drill hole with the associated internal composites, and calculated thickness at intersection points (left). Scale in metres. Histogram of declustered assessed thicknesses (upper right) and scatterplot of the grade (Au g/t) versus thickness (lower right).

4.2 Sequential Gaussian Simulation

The thickness variable underwent a normal score transformation and a variogram model is fitted to the normal score data (Figure 4.5). SGS is performed on a 0.5 m (strike) by 0.5 m (width) by 0.5 m (vertical) grid for 200 realizations and back transformed to original units. The locations where the drill hole intersected the wireframe were populated with the thickness realizations of the closest block. The realizations reproduced the input histogram and variogram reasonably well (Figures 4.6 and 4.7).

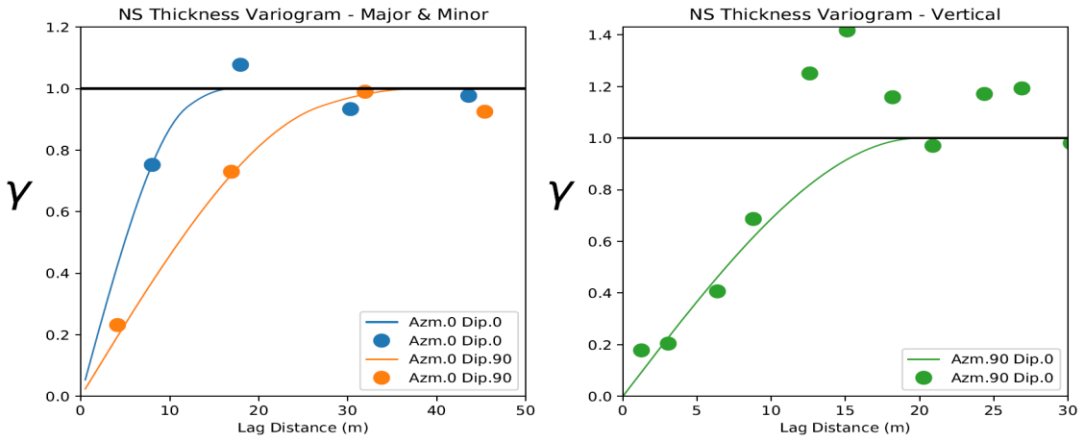

Figure 4.5: Variogram of assessed normal score thickness variable.

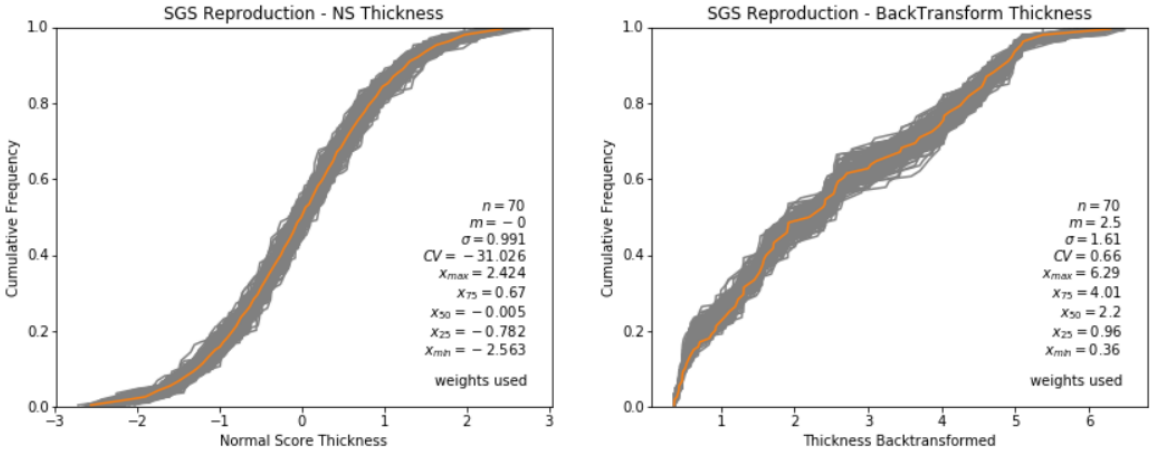

Figure 4.6: Histogram reproduction of the thickness variable in normal and original units. The grey lines represent the realizations and the orange line represents the input data.

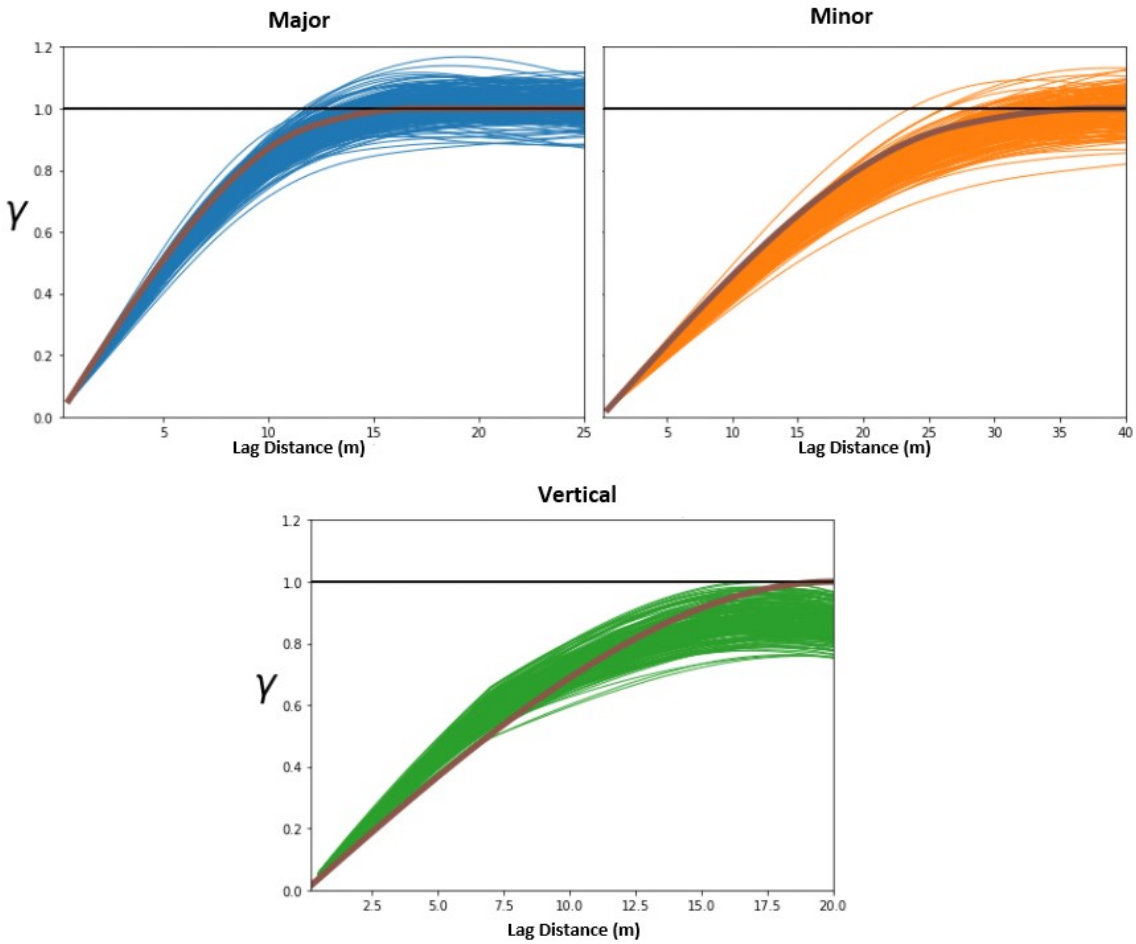

Figure 4.7: Variogram reproduction of the normal score thickness variable in the major (blue), minor (orange), and vertical (green) directions. The bolded brown line is the input variogram.

4.3 Wireframe Creation and Thickness Uncertainty

The drill hole intercept points are fixed in Cartesian space and have the associated imputed thickness values. The north-south orientation of the rotated domain allowed for the simple conversion of the thickness values to coordinate points, where the thickness value was either added or subtracted to the X coordinate of the intercept. On average, the calculated coordinate position distribution across from each intercept point are centered on the originally assessed thickness measurement (Figure 4.8). The data was divided into HW and FW subsets, both containing coordinate distributions and fixed intercept locations.

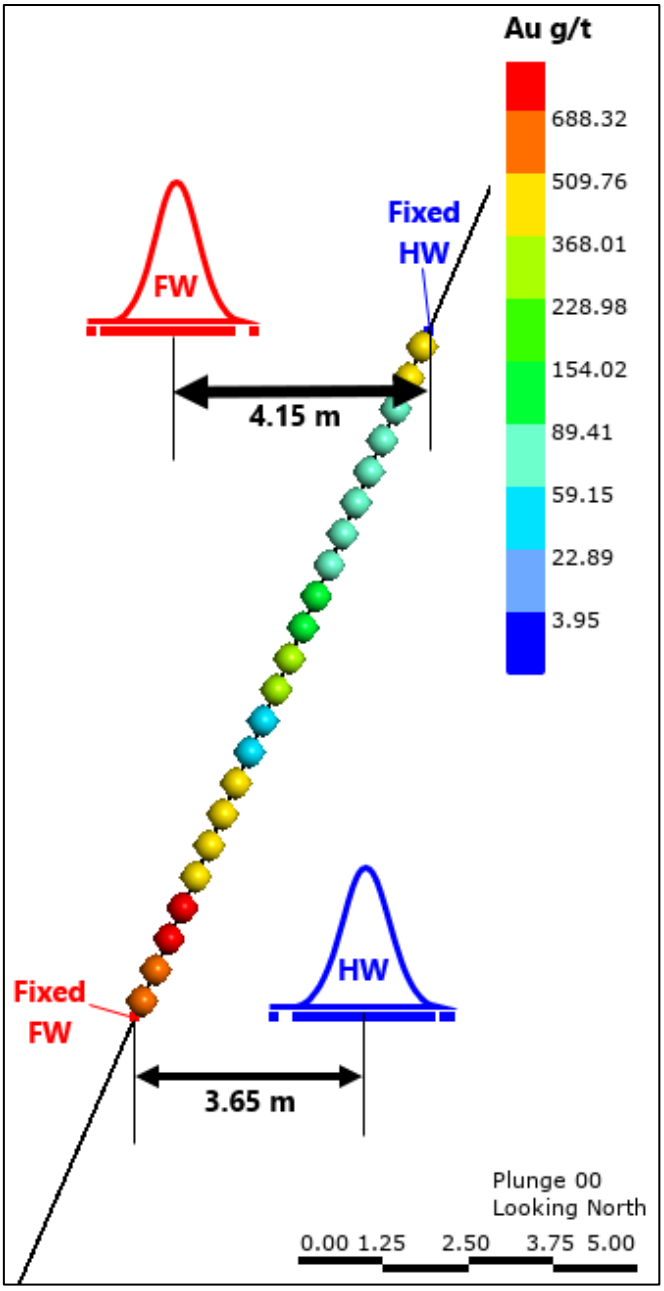

Figure 4.8: Cross section schematic of the coordinate distribution opposite to the fixed drill hole intercept point with composite grades (Au g/t) and the originally assessed thickness distances. Scale in metres.

The FW and HW surface wireframes were created from the data using the LeapFrog Geo version 5.0.1 program from SeeQuent Limited. The program uses a radial basis function to interpolate a surface from the given points. The interpolant, which snapped to each individual point, used a linear model with total sill of 300 and base range of 50 m with no nugget effect. Two hundred wireframe realizations were created for both the FW and HW datasets (Figure 4.9).

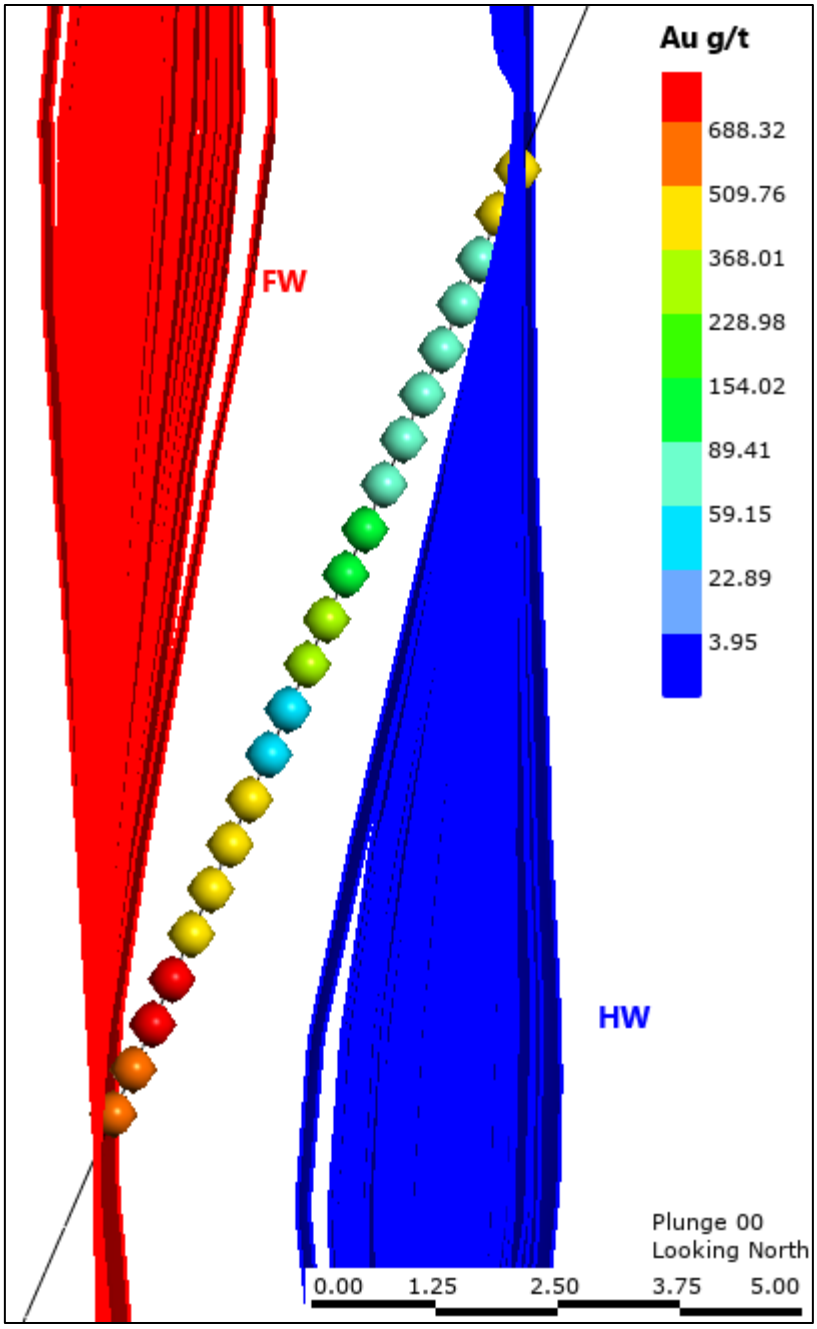

Figure 4.9: Cross section of the generated HW and FW wireframe realizations with composite grades. Scale in metres.

Each realization consists of a different imputed value perpendicular to the intercept point of the opposite dataset and the intercept point of the dataset being evaluated (i.e., the fixed data point). This accounts for uncertainty in the boundary but is fixed at data locations. Automated implicit modelling was necessary to create the 400 surfaces (200 for FW and 200 for HW) being evaluated, as manual creation of the surfaces would be prohibitively time consuming.

Chapter 5

5. Boundary Uncertainty Assessment

Assessing volume uncertainty of the Arrow Deposit is important for understanding metal content uncertainty, as the metal content is sensitive to changes in volume due to the extreme high-grade nature of the deposit. The volume uncertainty can be assessed by determining the uncertainty of the domain perpendicular to the plane of continuity and in the plane of continuity. In Chapter 4, the assessment of domain position and thickness uncertainty quantified uncertainty perpendicular to the plane of continuity; this chapter describes the process to assess uncertainty in the plane of continuity, termed the "Boundary" uncertainty.

Workflows using SDFs were not used, as they can introduce bias to the model in the presence of asymmetrical, sparse data, which is evident at the Arrow Deposit. Rather, the workflow Mancell (2020) proposed was used, as it addresses the problem through the use of an indicator estimator to generate a probability field to establish an uncertainty bandwidth. Since the uncertainty perpendicular to the plane of continuity was already assessed in Chapter 4, the boundary uncertainty assessment was limited to the plane of continuity, thus the data set was considered in two dimensions.

5.1 Workflow Outline

The workflow for the assessment of boundary uncertainty at the Arrow Deposit is derived from the boundary uncertainty work completed by Mancell (2020). The primary five steps implemented in this chapter are listed below – steps two to five are a summary of Mancell (2020).

1. Reduce the data to relevant extents and create a representative dataset. Note that the uncertainty study only considers the boundary model in the plane of continuity (i.e., two dimensions), as the uncertainty perpendicular to the plane of continuity is quantified through different means.

2. Creation of a Nearest Neighbour (NN) model to provide an unbiased spatial representation of the conditioning data.

3. Creation of an indicator estimate to generate a field of probabilities of boundary location. Note that two estimator types were tested, Global Ordinary Kriging and RBF, and the one that is most geologically representative and reasonably replicates the data was selected.

4. Thresholding the indicator estimate to achieve global unbiasedness as derived from the NN model.

5.Thresholding above and below the NN threshold to provide eroded and dilated boundary that gives access to a bandwidth of uncertainty.

5.2 Data Preparation

The drill data was clipped to relevant extents around the domain to inform the estimation; the resulting data encompasses a volume that is approximately 14 m wide (perpendicular to the plane of continuity), 195 m along strike, and 242 m along the dip direction (Figure 5.1). The dataset is further reduced by removing "outside" indicators from drill holes that contain "inside" indicators, as these "outside" indicators are not relevant when the boundary is only being considered in the plane of continuity. In total, the number of "inside" indicators remained the same at 656 composites, but the number of "outside" indicators was reduced to 3,296 composites from over 259,000 composites.

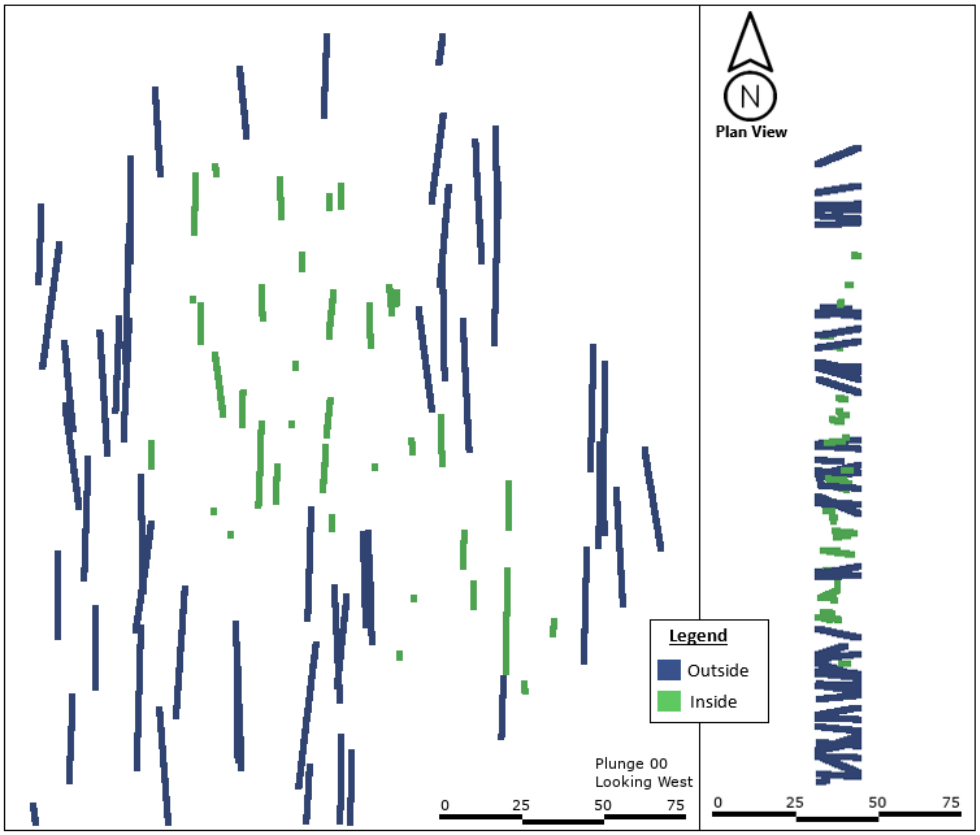

Figure 5.1: Long section and plan view of the reduced drill hole dataset. Scale in metres.

5.3 Nearest Neighbour Model

Nearest Neighbour (NN) volumes are a reasonable representation of global volume and can be used to check an interpreted model's global volume (Rossi & Deutsch, 2014). The first step in creating the NN model is to define the indicators through indicator formulism.

$$i(\boldsymbol{u}_\alpha) = \begin{cases} 1, if\ u_\alpha\ \in domain \\ 0, otherwise \end{cases} for\ \alpha\ =\ 1,\ldots,n$$

Where $i(\boldsymbol{u}_\alpha)$ is the indicator at location $\boldsymbol{u}_\alpha$ and is 1 inside the domain and 0 otherwise calculated for all α conditioning data.

The NN model is created through the Ordinary Kriging of the indicators with a maximum of one datum; using a single informing data point and having the sum of the kriging weights equal one effectively equates to a NN model. The NN model was created on a grid of 0.5 m by 0.5 m cells extending 410 cells in the y direction and 478 cells in the z direction, for a total of 195,980 cells. The NN model used a search of 80 m (z direction) by 80 m (y direction), resulting in 49,132 cells with an indicator of 1, equating to 12,283 metres squared (Figure 5.2). For reference, the deterministic study domain has a two-dimensional area of 14,076 metres squared.

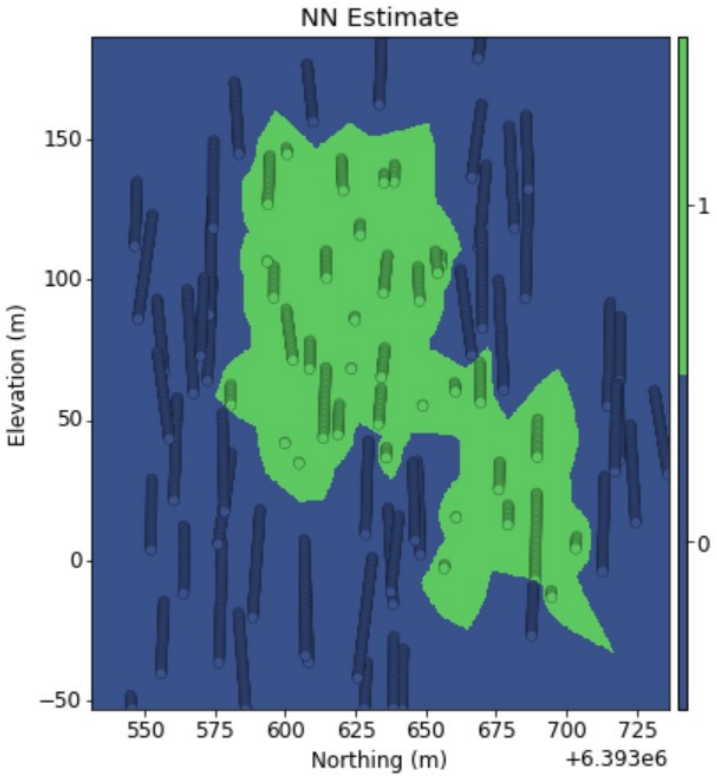

Figure 5.2: Long section of the reduced drill hole dataset and NN model. Scale in metres.

5.4 Indicator Estimation – Global Simple Kriging

Interpolation of the indicator values maps a field of probabilities from which a boundary interface can be extracted (Mancell, 2020). Global Simple Kriging was used to interpolate indicator values on the grid; Global Simple Kriging uses all available conditioning data and produces a smooth, artifact free model (Carvalho, 2018). A spherical variogram was fitted to the omnidirectional experimental variogram to

inform the estimate (Figure 5.3). Overall, no conventional variogram model fit the data well, so focus was on fitting the model to the short-range variability. Note that the Gaussian model fit the data better than the spherical variogram model, but resulted in an unrealistic, unstable model. The resulting interpolated field of probabilities of the indicator values from the Global Simple Kriging is visualized in Figure 5.4.

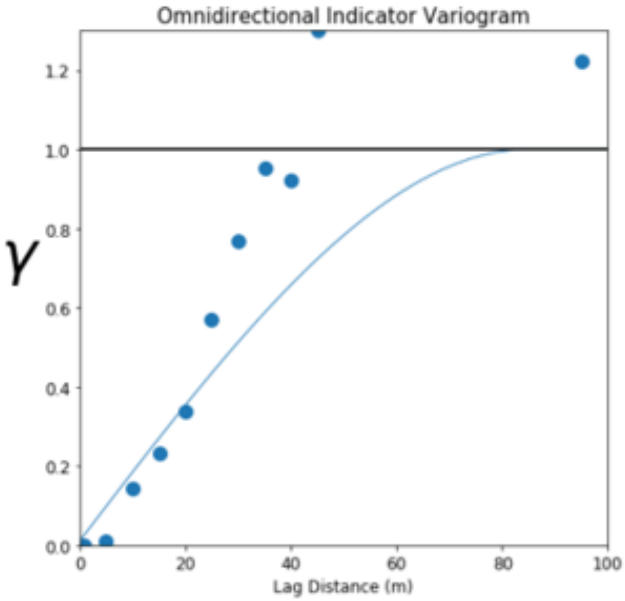

Figure 5.3: Omnidirectional experimental and spherical model variogram.

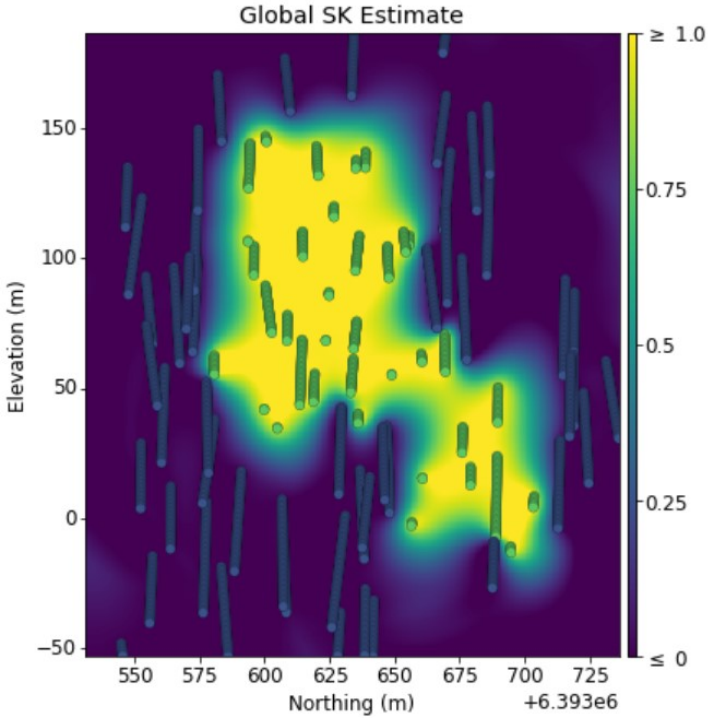

Figure 5.4: Global Simple Kriging of Indicators.

5.5 Indicator Estimation – RBF

Mancell recommends RBF over kriging because the lack of edge effects observed in the RBF estimate, that first-order stationarity was not required, and variograms were not necessary (Martin, 2019). The edge effect is a common problem in boundary estimation where there is uncontrolled extrapolation of estimated domains to the extents of the model. The edge effect is the influence of samples extending into areas of sparse or no data. In Global Ordinary Kriging, the estimates distal to data are relatively constant with values close to the dataset's mean indicator value. Mitigating edge effects in Global Ordinary Kriging is completed by restricting the variogram range or adding artificial controlling samples into the dataset (Maureira, 2015).

In RBF estimation, a function is fitted to the conditioning data, and the function requires a kernel type to be selected. In the study domain, the Gaussian kernel was selected because it replicates short-scale continuity well, resulting in smooth models. The Gaussian kernel is defined as follows:

$$\phi(r) = exp^{-\epsilon^2 r^2}$$

Where $\phi(r)$ is the Gaussian kernel as a function of radius and ϵ is the support parameter. A reasonable support parameter can be estimated from the geometrical configuration of the data. This is accomplished by finding the radius of the largest circle that can be placed inside the data configuration with all locations considered. Using the provided data, the support parameter was calculated to be 56.82; the equation and the output of the equation are below.

$$support = \underset{\boldsymbol{u} \in \boldsymbol{A}}{Max}\left[\underset{\boldsymbol{u_\alpha} \in \boldsymbol{A}}{Min} |\boldsymbol{u} - \boldsymbol{u}_\alpha| \right]$$

Viewing the estimated model, minor edge effects are observed at depth in the northern portion of the model (Figure 5.5); however, the model is left unchanged as it will not be an issue during thresholding for an uncertainty bandwidth.

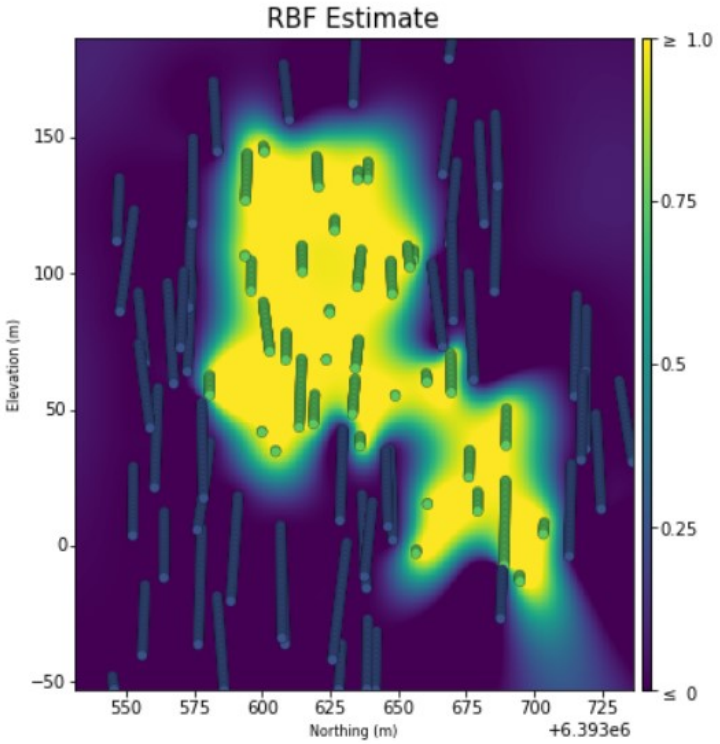

Figure 5.5: RBF of Indicators.

5.6 Thresholding and Uncertainty Assessment

As previously mentioned, NN models are a reasonable unbiased representation of global volume or area. The ratio of inside to outside cells in the NN model provides a NN-thresholding value, th_{NN}, between [0,1] as per the equation below (Mancell, 2020).

$$th_{NN} = 1 - \frac{A_{NN}}{A_{Total}}$$

The NN-thresholding value extracts from the indicator estimate cumulative frequency distribution (CDF) the estimate thresholding z-value, z_i^* (Figure 5.6; Mancell, 2020)

$$z_i^* = F_{i^*}(th_{NN})$$

In the study area, th_{NN} is equal to 0.749 and z_i^* is equal to 0.456 for the Global SK estimate and 0.462 for the RBF estimate. Figure 5.5 shows the cumulative distribution function (CDF) of the indicator estimates with the NN threshold and corresponding z_i^* value (yellow arrow). A bandwidth of uncertainty is generated by varying the z_i^* value by ±15% to create eroded and dilated boundaries – the eroded and dilated limits were chosen at the discretion of the author based on geological knowledge of the deposit

but are in line with the recommendations of Mancell, where the end members of boundary uncertainty are the extents that appears to be geologically reasonably (Figure 5.7).

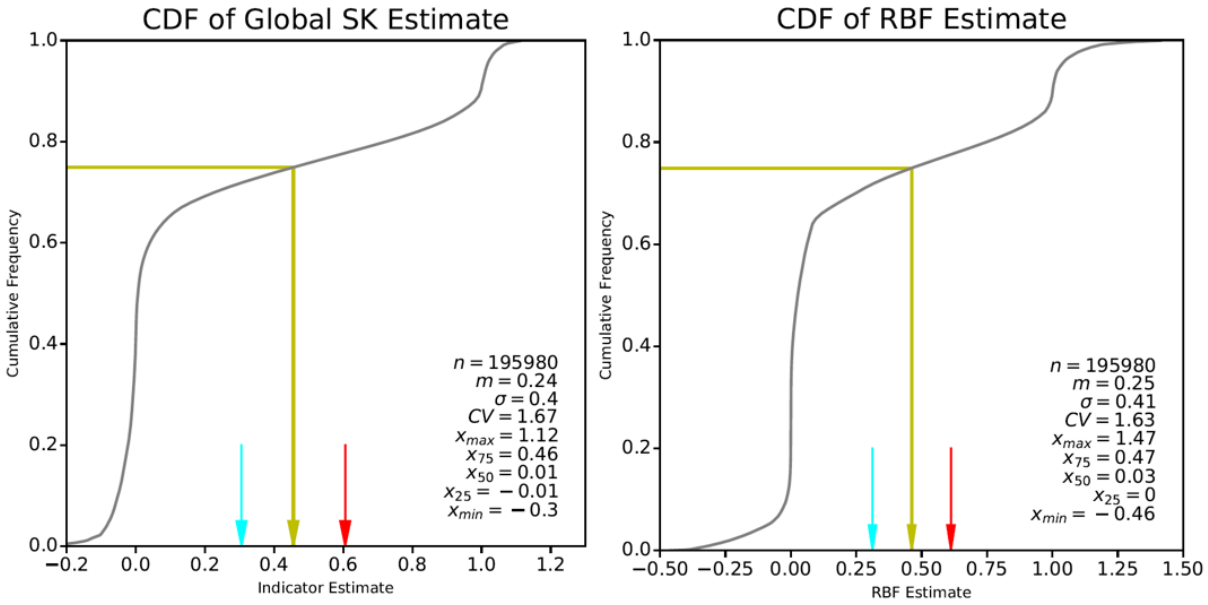

Figure 5.6: Cumulative Distribution Function of indicator estimate (Global SK Estimate and RBF Estimate), showing NN threshold, the corresponding z_i^* value (yellow arrow), and the bandwidth uncertainty end members (blue and red arrows).

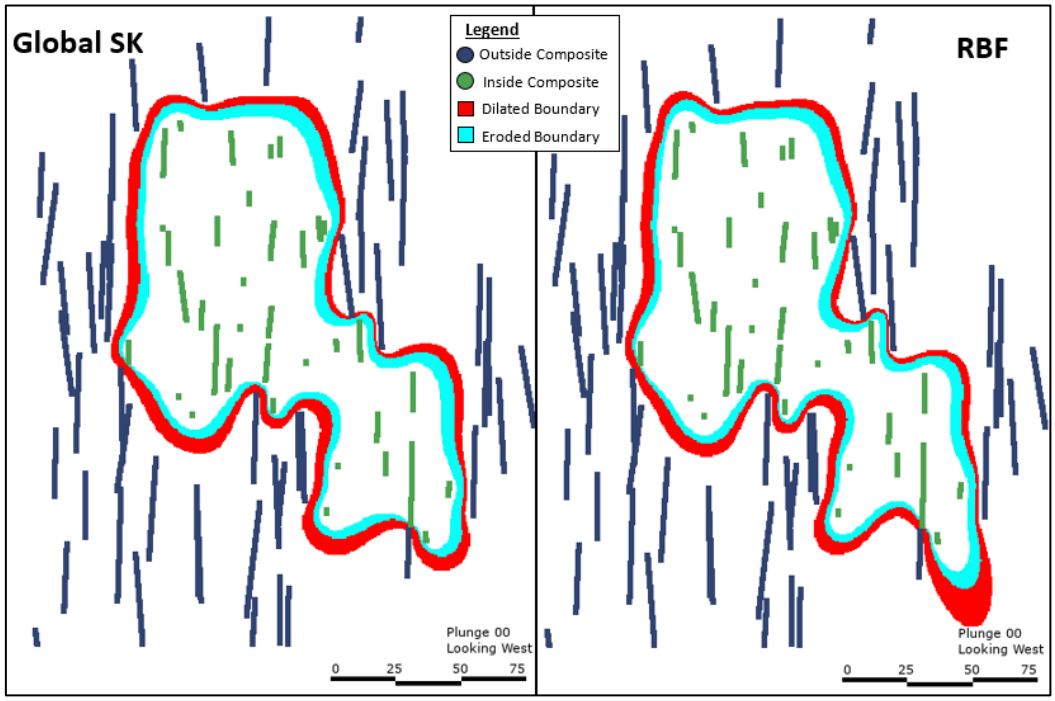

Figure 5.7: The bandwidth uncertainty of the boundary (left: Global SK, right: RBF) and indicator composite dataset. Note that in areas where the boundary appears to overlap the informing data, this is an artifact in the visualizations, where the point data is enlarged for viewing purposes.

5.7 Conclusions

The resulting boundaries corresponding with the z_i^* value and associated boundary uncertainty appears to be geologically reasonable. However, both the RBF and Global SK estimates have values below 0 (25.1% of RBF estimated cells and 41.7% of Global SK estimated cells) and above 1 (10.2% of RBF estimated cells and 9.2% of Global SK estimated cells), which is attributed to the screening effect during estimation. The estimated cells with values outside the values of the informing dataset was considered to be of negligible importance in the evaluation of the boundary, as the estimates with the extreme values are in areas distal to the boundary location, typically in between drill hole intersections with the same indicator classification, as visualized in Figure 5.8.

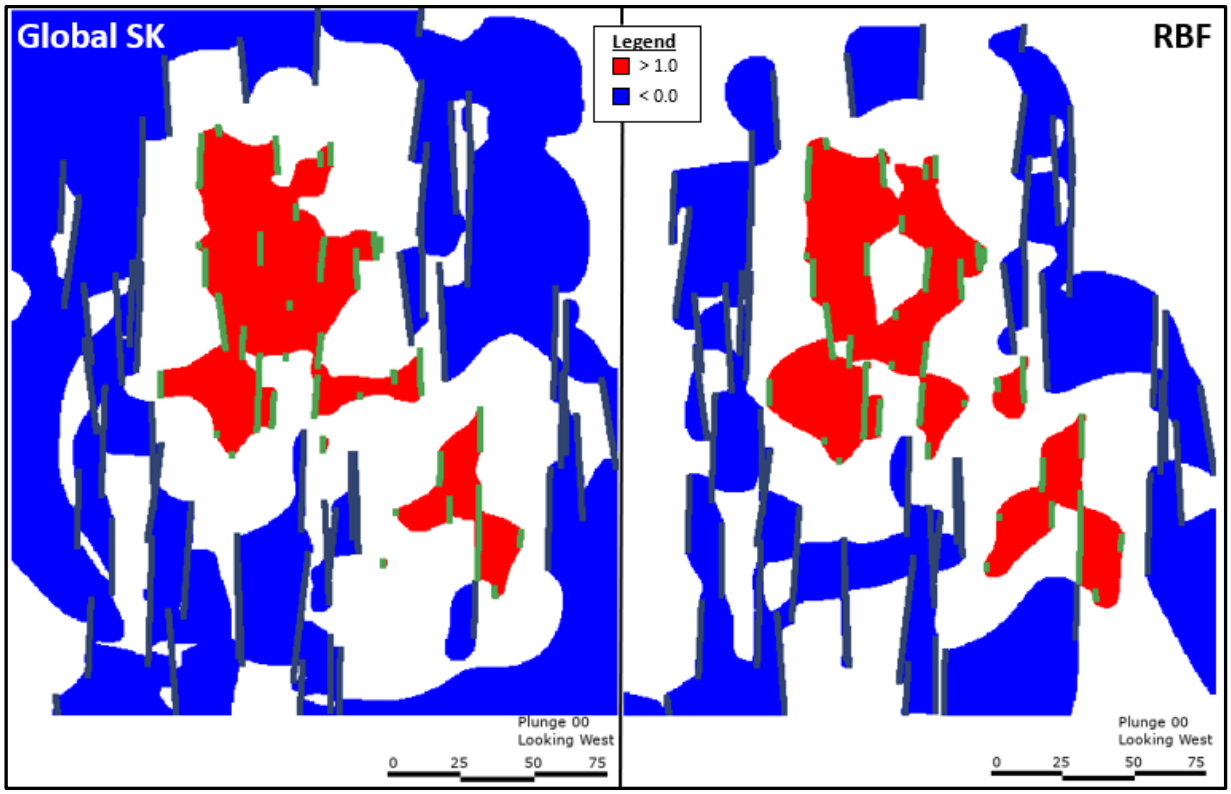

Figure 5.8: The Global SK (left) and RBF (right) estimated cells with extreme values

The RBF estimate was selected for the uncertainty analysis as it appeared to honour the sparsely sampled areas better than the Global SK estimate, which is likely a function of a poorly fitting variogram used in the Global SK estimate.

The workflow that Mancell (2020) proposed for boundary uncertainty is straight-forward to implement, accounts for asymmetries in data configuration, and provides a geologically rational output. However, as Mancell noted, NN modeling is susceptible to edge effects in areas with sparse drilling and is only controlled by adjusting maximum distances, which introduces subjectivity. The endmembers of the

uncertainty bandwidth are chosen rather arbitrarily by the user, as well as the clipping limits of the data, again introducing subjectivity. Still, as one of the cannons in the field of geostatistics is to ensure that the resulting model is representative for the related geology, which is believed to have been achieved; where the unbiased boundary closely matches the deterministic geologic interpretation.

Chapter 6

6. Uncertainty Review of the Arrow Deposit

The geometry imputation, boundary, grade, and density models were combined into a single model and each individual component was reviewed to evaluate influence on overall uncertainty. For the purposes of comparison, the uncertainty associated with the boundary in the plane of continuity is termed "Boundary", the uncertainty of the domain thickness perpendicular to the plane of continuity is labelled "FW/HW" (footwall/hangingwall), the grade and density model is appropriately named "Grade/Density", and the combined model considering all sources of uncertainty is denoted "Complete". The uncertainty of each component is isolated for evaluation by using the E-type mean of the other components. In total, the domain contains 800 kilo (K) troy ounces of gold contained in 97 K tonnes at a grade of 259 grams per tonne gold, as per the E-type mean of the uncertainty assessments. No classification scheme or criteria was applied to the model.

Unsurprisingly, the Grade/Density model is the largest input to specific gravity and grade uncertainties, with very little influence from the FW/HW and Boundary models (Figures 6.1 and 6.2). The limited uncertainty associated with FW/HW and Boundary models is from each realization occupying slightly more or less blocks, introducing minor fluctuations in the average of the domain.

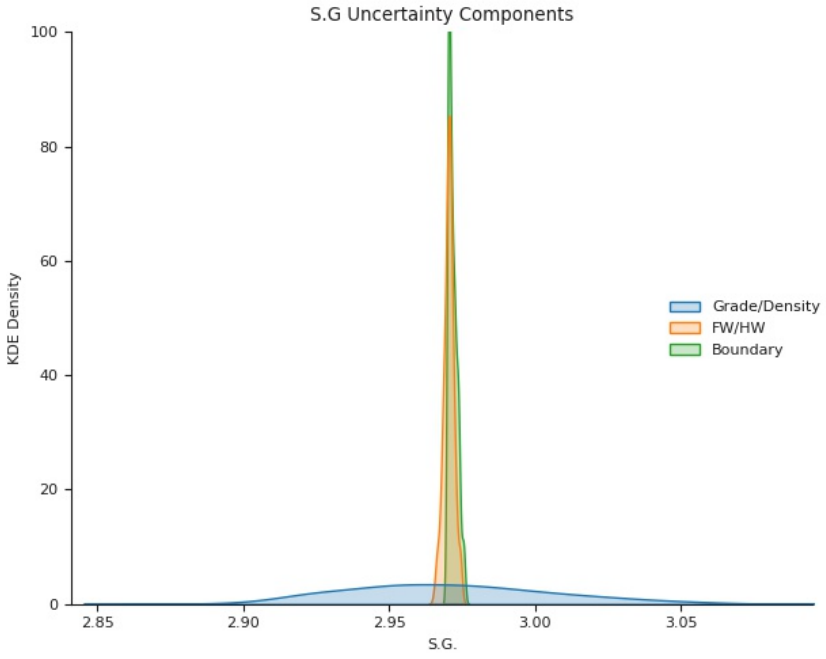

Figure 6.1: Kernel Density Estimates from histogram of specific gravity uncertainty pertaining to the "Grade/Density", "FW/HW", and "Boundary" models.

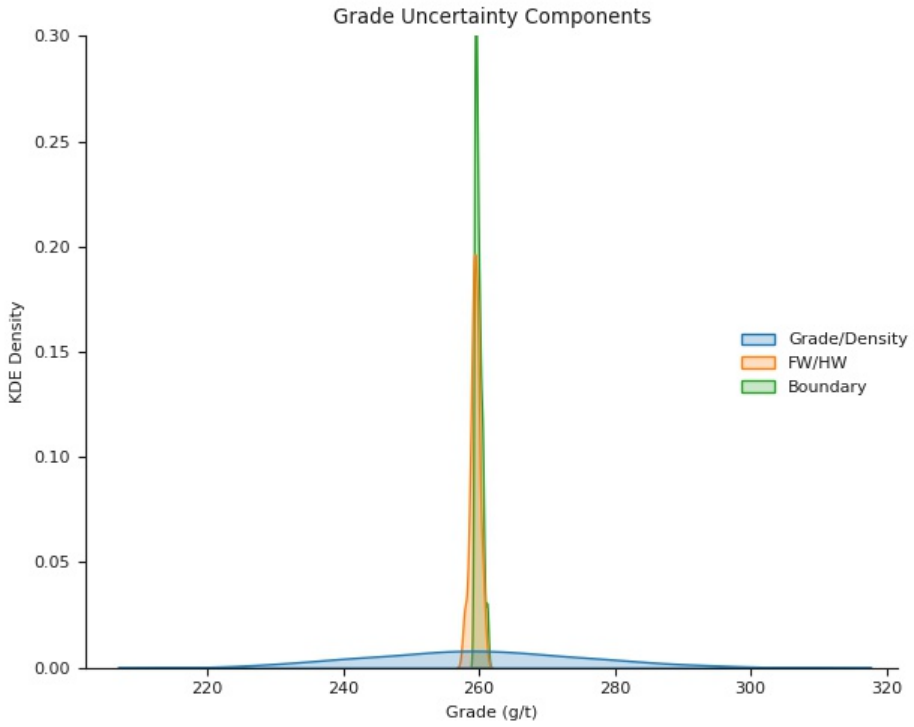

Figure 6.2: Kernel Density Estimates from histogram of specific gravity uncertainty pertaining to the "Grade/Density", "FW/HW", and "Boundary" models.

The Boundary model contributes the most to tonnage uncertainty; the FW/HW and Grade/Density models contribute about the same degree of uncertainty, with the "FW/HW" contributing slightly more (Figure 6.3). The Boundary model has the largest influence on tonnage because the model has the largest range in volume realizations. The Boundary model defines a bandwidth for uncertainty assessment by varying the z_i^* value by ±15%. The selected percentage to vary the z_i^* value is in line with the recommendations of Mancell, and appears to be geologically reasonably, but the model is sensitive to percentage selected to create the bandwidth; thus, requires further research.

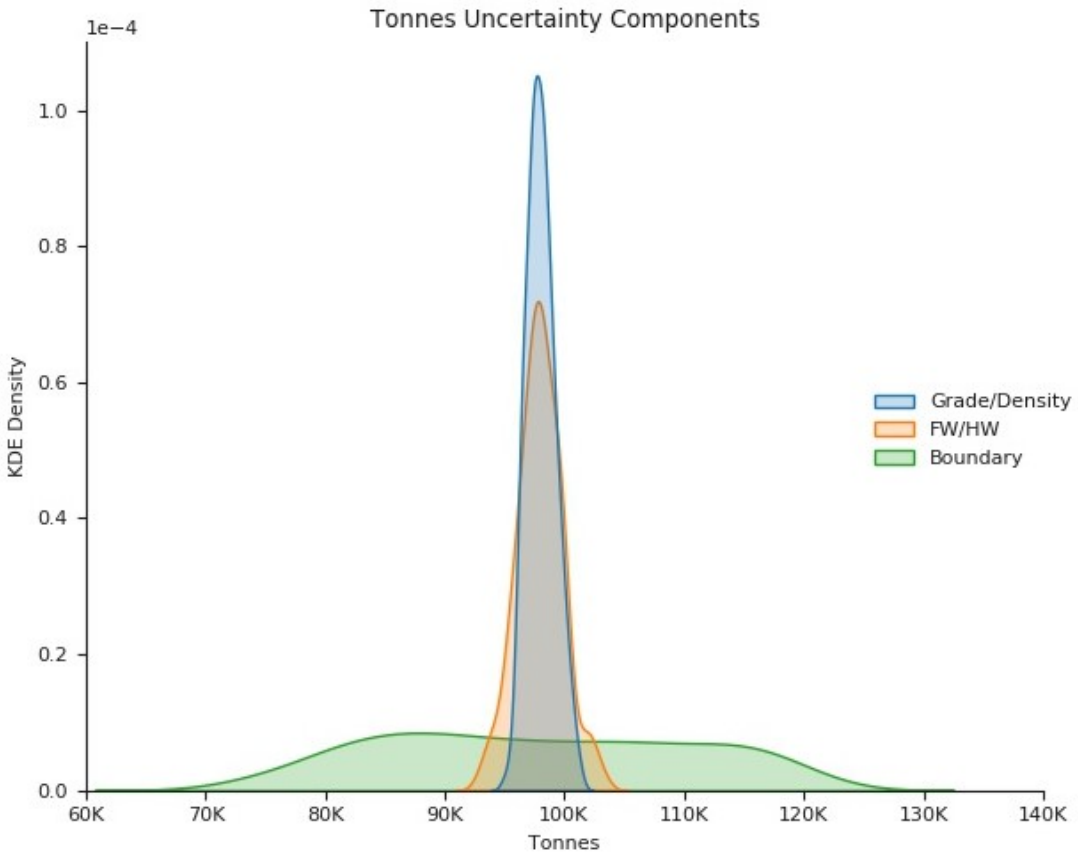

Figure 6.3: Kernel Density Estimates from histogram of tonnes uncertainty pertaining to the "Grade/Density", "FW/HW", and "Boundary" models.

The Boundary model contributes the most to the metal content (gold troy ounces) uncertainty, followed by the Grade/Density model with minor influence from the FW/HW model (Figure 6.4). The impact of the Boundary model to the metal content is attributed to the associated tonnage uncertainty within the defined volume and the high-grade nature of the deposit, whereas the effect of the Grade/Density model is attributed to the short-range variability of the variables. Furthermore, the Boundary model is poorly sampled, where the sampling equates to drill hole spacing, meaning more uncertainty than other well sampled variables. The FW/HW model appears to reflect that thickness is more continuous than the sampling density.

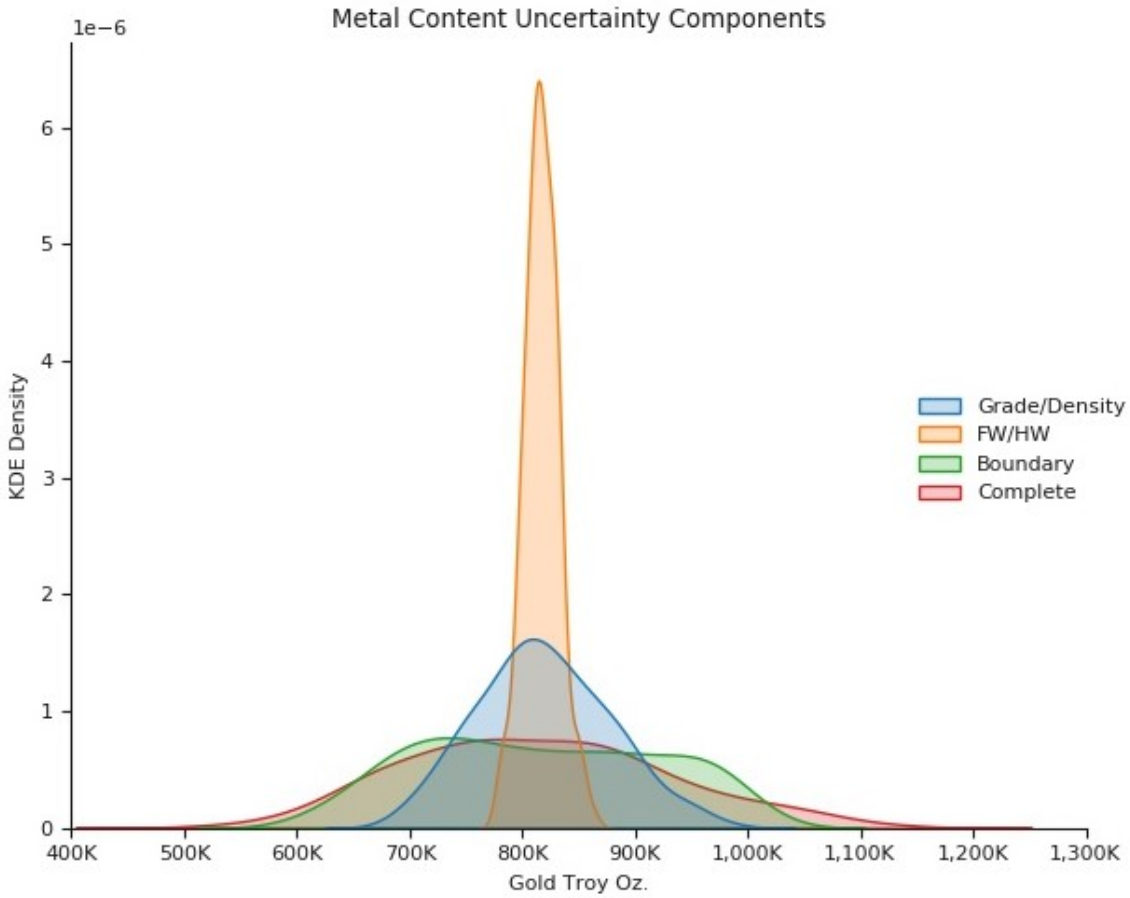

Figure 6.4: Kernel Density Estimates from histogram of tonnes uncertainty pertaining to the "Grade/Density", "FW/HW", and "Boundary" models, as well as the combined model, noted as "Complete".

The uncertainty associated with the volume, grade, and density variables of the deposit are the focus of the study, as these variables define the overall metal content of the deposit, the largest input to project economics. The metal content of the Arrow Deposit is sensitive to changes in volume due to the extreme high-grade nature of the deposit, therefore the Boundary model, which has a large range of possible volumes (and range in the determination of ore/not ore), contributes the most to the metal content uncertainty; demonstrating the importance of applying the boundary technique with great care to ensure geologic reasonableness is achieved in a narrow vein, high-grade, tabular deposit. The Grade/Density model also significantly contributed to the metal content uncertainty, which is credited to the short-range variability of the variables. The FW/HW model provided additional uncertainty to the metal content, but to a much lesser degree as the range of possible volumes associated with the model were relatively small.

Chapter 7

7. Conclusion

This thesis details a workflow to quantify uncertainty of the Arrow Deposit, using a single, representative high-grade domain for the study. The grade variable is exhaustively sampled at all drill hole locations within the study domain, but the density variable is missing at approximately 82% of the drill hole locations and biased to high-grade intercepts. The density variable was de-biased through a simple imputation process that selected density values from the global deposit-wide dataset. A multivariate spatial bootstrap model was created from the debiased dataset, as well as a density-imputed model using a GMM followed by decorrelation via PPMT and independent SGS. The multivariate spatial bootstrap model demonstrated that the conditioning data in its configuration in the study domain has little effect on the global uncertainty of the grade and density variables. The GMM imputed model captures the uncertainty associated with the missing density data, thus is considered a more complete model (and there was little histogram parameter uncertainty) and was used in the final model. Appropriate checks were completed on the two models.

Volume uncertainty was quantified by assessing the uncertainty of the thickness perpendicular to the plane of continuity via a geometry imputation process and the uncertainty of the boundary in the plane of continuity via an indicator estimate.

The grade/density model was combined with the volume models and reviewed for contribution to the metal content uncertainty. The boundary model, which has a large range of possible volumes (and range in the determination of ore/not ore), contributes the most to the metal content uncertainty.

The domain contains 800 kilo (K) troy ounces of gold contained in 97 K tonnes at a grade of 259 grams per tonne gold, as per the E-type mean of the uncertainty assessments. No classification scheme or criteria was applied to the model, further work is necessary to do so, including cost and revenue analysis.

7.1 Limitations

The original dataset was debiased by sampling the deposit-wide dataset assuming that any sample location with both a grade and density measurement in the deposit are representative of a location within the study domain. This assumption is supported by the author's lithological, structural, and alteration knowledge of the deposit, and is thought to be pragmatically reasonable, but has not been tested with detailed academic examination; further research is recommended.

Uncertainty perpendicular to the plane of continuity was assessed through a thickness imputation process, which only contributed a minor amount to global metal content uncertainty. Carvalho (2018) developed a technique for assessing domain geometry uncertainty that used both position and thickness uncertainty. Position uncertainty was not the focus of this thesis and adding it to the workflow can reasonably assumed to have little effect on global metal content uncertainty. However future analysis of the Arrow Deposit may require the addition of position uncertainty and should be considered in future workflows.

In the boundary model, both the RBF and Global SK indicator estimates have values below 0 and above 1 due to the screening effect during estimation. These cells are of negligible importance in the evaluation of the boundary, as the cells are distal to the boundary location to be assessed. Further research should be completed on the estimation types for the boundary model outside of the methods proposed by Mancell (2020) to reduce the number of extreme value cells.

The uncertainty of the boundary model was assessed by randomly sampling the uncertainty bandwidth produced from the estimate, producing an irregular distribution. Further research is needed on the sampling of the bandwidth to produce a more justifiable outcome.

In addition to the visual and reproduction of data checks performed on the model, cross-validation or k-fold validation are to be completed in future research for an assessment of the estimate relative to the truth.

The workflow can be optimized to account for flattening of the cartesian data (i.e., unfolding) and include the outcomes of the spatial bootstrap for both the multivariate (i.e., the grade and density model) and univariate (i.e., thickness) cases. The unfolding accounts for local anisotropy observed in the domain through a coordinate transformation whereas integrating the spatial bootstrap into the workflow, rather than as a side study, incorporates the uncertainty in global parameters in the final uncertainty. Figure 7.1 is a schematic of the optimized workflow.

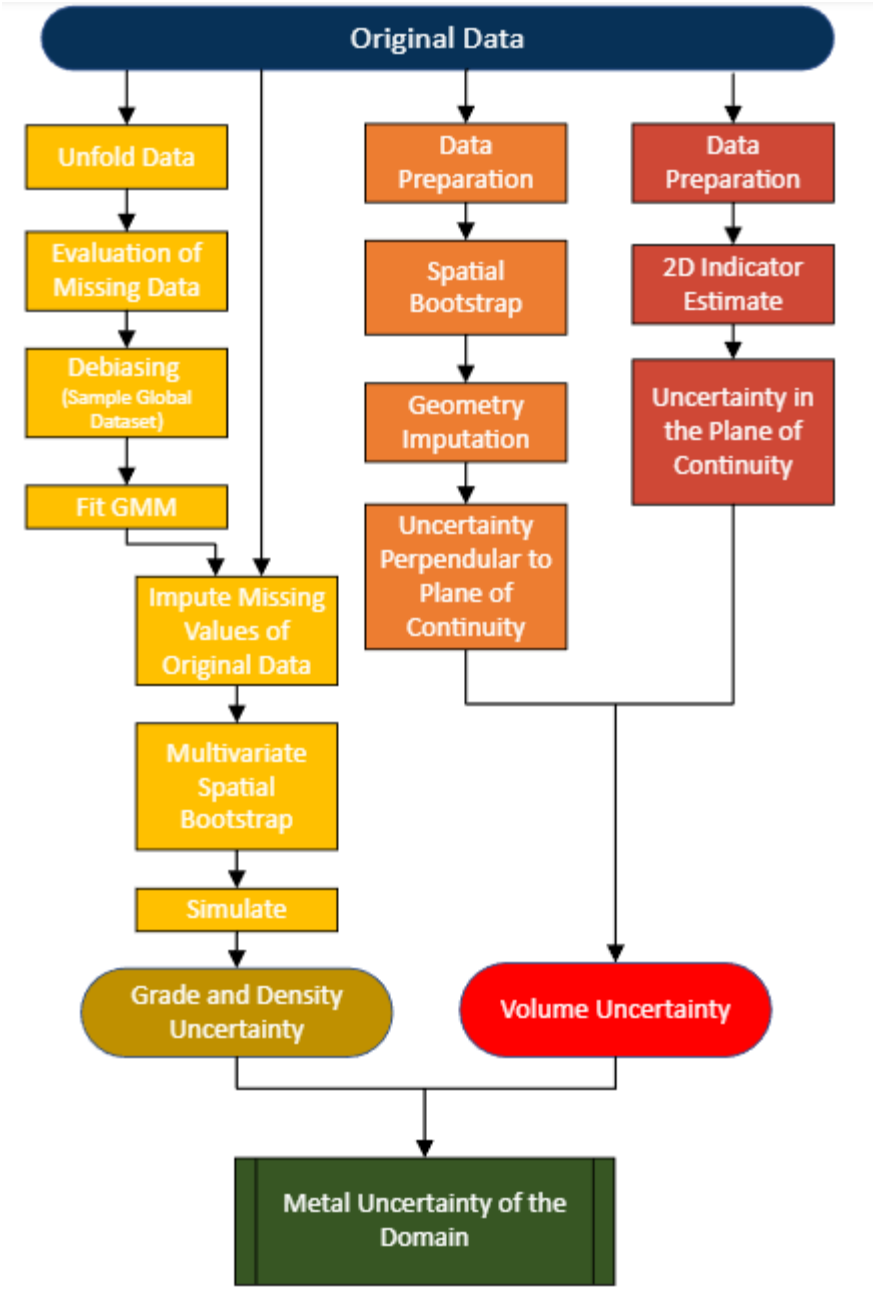

Figure 7.1: a schematic of the proposed workflow.

7.2 Contributions

The workflow within this thesis intends to provide a pragmatic approach to uncertainty quantification for a tabular vein type deposit with heterogeneously sampled data; most of the workflows can be completed with commercial software, which is likely useful to industry practitioners. The workflow, in part or in full and with modifications can be used for Mineral Resource reporting or used as the basis for classification criteria (i.e., reviewed against monthly or quarterly production panels for classification

based on uncertainty tolerances). The workflow can be generalized to be used on unfolded tabular veins or stratigraphic units.

The workflow for the assessment of grade and density uncertainty demonstrates how to define the missing data mechanism (MCAR, MAR, or MNAR) and a simple way to create a representative dataset for later GMM fitting and imputation, thus de-biasing the data and maintaining uncertainty associated with the missing data. The fitting of the GMM is described as well as the testing recommended to select the appropriate number of components. The output of the GMM imputation workflow contained multivariate complexities that required the use of the non-linear PPMT. The PPMT, independent simulation on a grid, and back transformation is described with all the appropriate checks. The entirety of the workflow, which includes the linking and combining of multiple geostatistical techniques, can be used as a case study in the mining industry.

The boundary and thickness uncertainties are interpreted to be independent from one another – the workflow describes how to decouple the data, prepare the data, and perform the various techniques to output the associated uncertainty, then combine the models to output the final volume uncertainty model.

References

Babak, O., & Deutsch, C. V. (2009). Accounting for parameter uncertainty in reservoir uncertainty assessment: the conditional finite-domain approach. Natural Resources Research, 18(1), 7–17.

Barnett, R., & Deutsch, C. (2015a). Guide to Multivariate Modeling with the PPMT. CCG Guidebook Series Vol. 20, Centre for Computational Geostatistics, University of Alberta.

Barnett, R., & Deutsch, C. (2015b). Multivariate imputation of unequally sampled geological variables. Mathematical Geosciences, 1–27.

Barnett, R., & Deutsch, C., (2018). Production Scale Uncertainty vs. Data Spacing. Report prepared by Resource Modeling Solutions for NexGen Energy Ltd.

Barnett, R., Manchuk, J., & Deutsch, C. (2013). Advances in the projection pursuit multivariate transform. CCG Paper 2013-106, Centre for Computational Geostatistics, University of Alberta.

Bellman, R. E. (1957). Dynamic programming. New Jersey: Princeton University Press.

Bliss, C. (1934). The method of probits. Science, 79, 39-39.

Boisvert, J. B., Rossi, M. E., Ehrig, K., & Deutsch, C. V. (2013). Geometallugical modeling at Olympic Dam Mine, South Australia. Mathematical Geosciences, 45, 901-925.

Carr, J. C., Beatson, R. K., Cherrie, J. B., Mitchell, T. J., Fright, W. R., McCallum, B. C., & Evans, T. R. (2001). Reconstruction and representation of 3d objects with radial basis functions. In Proceedings of the 28th annual conference on computer graphics and interactive techniques (p. 67–76). New York, NY, USA: Association for Computing Machinery.

Carvalho, D., (2018). Probabilistic Resource Modeling of Vein Deposits. (Unpublished master's thesis), University of Alberta, Edmonton, Canada.

Chiles, J. P., & Delfiner, P. (2012). Modeling spatial uncertainty, 2nd edition. Wiley, New York.

Cohen, J., (1988), Statistical power analysis for behavioral sciences. New York, NY. Lawrence Erlbaum Associates, Publishers.

Deutsch, C. (2011). Multivariate standard normal transformation. CCG Annual Report 13, Paper 101, Centre for Computational Geostatistics, University of Alberta.

Edwards, R., & Atkinson, K. (1986). Ore deposit geology and its influence on mineral exploration. - 466 pp, London, New York (Chapman and Hall).

Enders, C. K. (2010). Applied missing data analysis. New York, NY. The Guilford Press.

Feyen, L., & Caers, J. (2006). Quantifying geological uncertainty for flow and transport modeling in multi-modal heterogeneous formations. Advances in Water Resources, 29(6), 912–929.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, C-23, 881-890.

Fukunaga, K. (1972). Introduction to statistical pattern recognition. Academic Press

Hatton, M., O'Hara, P., Mathisen, M., & Walker, D., 2021, Arrow deposit, Rook I Property, NI 43-101 Technical report on Feasibility Study Arrow deposit: technical report prepared by Stantec, Wood, RPA, and Golder for NexGen Energy Ltd., effective date February 22, 2021.

Hillacre, S., Ansdell, K., & McEwan, B., 2020, Structural analysis and paragenesis of the Arrow uranium deposit, western Athabasca Basin, Saskatchewan, Canada: Implications for the development of the Patterson Lake corridor: Economic Geology.

Hosseini, A. (2009). Probabilistic modeling of natural attenuation of petroleum hydrocarbons (Unpublished doctoral dissertation). University of Alberta.

Huber, P. (1981). Projection pursuit, research report pjh-6 (Tech. Rep.). Massachusetts: Harvard University, Dept. of Statistics.

Isaaks, E. H. (1990). The application of monte carlo methods to the analysis of spatially correlated data (Unpublished doctoral dissertation). Stanford University

Jefferson, C.W., Thomas, D.J., Gandhi, S.S., Ramaekers, P., Delaney, G., Brisbin, D., Cutts, C., Portella, P., & Olson, R.A., 2007, Unconformity-associated uranium deposits of the Athabasca Basin, Saskatchewan and Alberta: EXTECH-IV: Geology of uranium EXploration TECHnology of the Proterozoic Athabasca Basin, Saskatchewan and Alberta, Jefferson, C.W. and Delaney, G. (eds.), Geological Survey of Canada Bulletin 588, Saskatchewan Geological Society Special Publication 18, Mineral Deposits Division (GAC) Special Publication 4, p. 23-68.

Journel, A. G., & Bitanov, A. (2004). Uncertainty in N/G ratio in early reservoir development. Journal of Petroleum Science and Engineering, 44(1), 415–430.

Journel, A. G., & Huijbregts, C. J. (1978). Mining geostatistics. London: Academic Press

Khan, D., & Deutsch, C. (2016). Practical incorporation of multivariate parameter uncertainty in geostatistical resource modeling. Natural Resources Research. Vol. 25.

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

Little, R. J., & Rubin, D. B. (2014). Statistical analysis with missing data. John Wiley & Sons.

Mancell, S.A., (2020), Boundary Modeling with Uncertainty, (Unpublished master's thesis), University of Alberta, Edmonton, Canada.

Martin, R. (2019). Data driven decisions of stationarity for improved numerical modeling in geological environments (Unpublished doctoral dissertation). University of Alberta.

Maureira, D. (2015). Enhanced geologic modeling with data-driven training images for improved resources and recoverable reserves (Unpublished doctoral dissertation). University of Alberta.

McLachlan, G. & Peel, D. (2000). Finite Mixture Models. John Wiley & Sons, Inc.

McLachlan, G. J. & Krishnan, T. (2008). The EM Algorithm and Extensions. John Wiley & Sons, Inc., 2 edition.

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

Neufeld, R., & Deutsch, C. (2006). Data integration with non-parametric bayesian updating. CCG Paper 2006-105, Centre for Computational Geostatistics, University of Alberta.

O'Hara, P., Cox, J., Robson, D., & Mathisen, M. (2018). Arrow Deposit, Rook I Project, Saskatchewan, NI 43-101 Technical Report on Pre-Feasibility Study. NI 43-101 Report prepared by Wood Canada and Roscoe Postle Associates for NexGen Energy Ltd.

Ostenberg, B. C., & Deutsch, C. (2017). Transforming Vein Deposit Coordinates for Improved Resource Estimation. CCG Annual Report 19, Paper 312

Reed, M., & Simon, B. (1972). Functional analysis (Vol. 1). New York: Academic Press.

Rezvandehy, M. (2016). Geostatistical reservoir modelling with parameter uncertainty in presence of limited well data. Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada.

Rezvandehy, M., & Deutsch, C. (2014). Combining Multivariate Gaussian Distributions from Different Sources. CCG Annual Report 16, Paper 127

Rezvandehy, M. & Deutsch, C. (2016). Geostatistical Reservoir Modeling with Parameter Uncertainty in Presence of Limited Well Data. Unpublished doctoral dissertation, University of Alberta, Edmonton, Canada.

Rossi, M., & Deutsch, C. (2014). Mineral resource estimation. Springer Dordrecht Heidelberg New York London.

Rubin, D. B. (1976) Inference and missing data. Biometrika, Volume 63, Issue 3, p 581-592.

Rubin, D. B. (1987) Multiple imputation for nonresponse in surveys. New York, NY. John Wiley and Sons

Schafer. J. L. & Graham, J. W. (2002) Missing data: our view of the state of the art. Psychological Methods, 7, 147-177.

Silva, C., & Costa, J.F. (2019). Missing not at random geological data. CCG Paper 2019-125, Centre for Computational Geostatistics, University of Alberta.

Silva, D. (2018). Enhanced geologic modeling of multiple categorical variables (Unpublished doctoral dissertation). University of Alberta.

Silva, D., & Deutsch, C. (2015a). Multivariate Data Imputation using Gaussian Mixture Models. CCG Paper 2015-104, Centre for Computational Geostatistics, University of Alberta.

Silva, D., & Deutsch, C. (2015b). Program for fitting Gaussian mixture models based on EM algorithm and geostatistical applications. CCG Annual Report 17, Centre for Computational Geostatistics, University of Alberta.

Verly, G. (1983). The multi-Gaussian approach and its applications to the estimation of local reserves. Mathematical Geology, 15, 249-286.

Vincent, J. & Deutsch, C. (2019). The Multivariate Spatial Bootstrap. University of Alberta, Edmonton, Canada. Retrieved from http://geostatisticslessons.com/pdfs/spatialbootstrap.pdf.

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

Wilde, B., & Deutsch, C. (2017). Geostatistical Support for Uncertainty Quantification and Drill Spacing Optimization at the Arrow Deposit. Report prepared by Clayton V. Deutsch Consultants Inc. for NexGen Energy Ltd.