Multiple-Indicator Kriging of Gaussian and Non-Gaussian Data

by

Jeremy Dean Vincent

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

Master of Science

in

Mining Engineering

Department of Civil and Environmental Engineering
University of Alberta

Abstract

Multiple-indicator kriging (MIK) manages outlier values through the indicator transform, it generates the distribution of uncertainty non-parametrically through direct estimation of the indicator-probability thresholds, and it readily incorporates secondary, categorical data into the estimate. These features make it attractive for use in the mining industry, especially for mineral deposits containing highly positively skewed data distributions. Order-relations corrections leading to an inconsistent probability distribution present a notable drawback. Furthermore, the use of constant, indicator-class means impacts estimate quality in the upper tail of the distribution, which often comprises significant economic value. The first contribution of this thesis documents the deviations and spatial variability of the estimated probability distribution against a fully consistent, and known probability distribution. Next, the indicator-class-mean component of the MIK estimator is isolated and compared to the known distribution of correct values. This research demonstrates that the indicator-class means vary as a function of the conditioning values. The greatest variability is observed in the lower and upper tails. The second contribution of this thesis is a comparison of MIK and multivariate-Gaussian kriging (MGK) estimates using non-Gaussian data. Motivating the comparison is the assertion that when using Gaussian data, MGK will always generate estimates with lower mean-squared error than MIK. Multiple scenarios ranging from highly non-Gaussian to Gaussian, are generated with the expectation that MIK will outperform MGK once the data were sufficiently non-Gaussian. In the scenarios tested, MGK consistently generates more accurate and precise estimates, demonstrating that MGK can produce robust estimates, even in the presence of highly non-Gaussian data. The place of MIK remains unclear; however, the procedures and standards to assess the relative performance of MIK and other techniques are documented more clearly.

Dedication

To:

P.W. Duckers - My biggest supporter.

T. Henrich - For igniting my curiosity in earth sciences.

Dr. W.S. Board - For always believing in me, even when I doubted myself.

Acknowledgments

I gratefully acknowledge the financial assistance from the Centre of Computation Geostatistics that allowed me to undertake this research. To my supervisor, Dr. Clayton Deutsch, thank you for your guidance and patient support. Your natural ability to motivate and teach is special, and I am grateful to have had the opportunity to learn from you. I would also like to thank my fellow students at the CCG for their friendship, help with questions, and for many fond memories.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Description
$\mathcal{A}$	Domain or region assumed to be stationary
$\forall$	For all
α, β	Sample location indices
$C_{\alpha,\beta}$	Covariance between sample locations, $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$
$C_{\alpha,0}$	Covariance between sample location, $\mathbf{u}_\alpha$, and unsampled location, $\mathbf{u}_0$
$E\{\quad\}$	Expected value
$F_Z(\mathbf{u};z)$	Univariate, cumulative-distribution function of Z-random variable at location, $\mathbf{u}$
$F_Z(\mathbf{u};z\|(n))$	Posterior, cumulative-distribution function of Z-random variable at location, $\mathbf{u}$
γ	Variogram value
$\widehat{\gamma}$	Statistical inference of variogram value
$\mathbf{h}$	Lag vector
$\in$	In, within
$I(\mathbf{u};z_k)$	Indicator transform of the continuous RV Z at location $\mathbf{u}$, at the k^{th} threshold
k	Threshold value, k
λ	Kriging weight
m	Metre
$\overline{m}$	Mean value
m_{ng}	Measure of non-Gaussianity
n	Number of locations
$Prob\{\quad\}$	Probability
P	Number of p-quantiles

Symbol	Description
σ	Standard deviation
$\widehat{\sigma}^2$	Estimation error variance
$\sum$	Sum
$\mathbf{u}$	Location vector in space
Z	Random variable
$z(\mathbf{u})$	Outcome of Z-random variable at location $\mathbf{u}$
$\widehat{z}(\mathbf{u}_0)$	Estimate of z at unsampled location $\mathbf{u}_0$
z_p	p-quantile of Z-random variable

List of Abbreviations

Abbreviation	Description
2-D	two-dimensional
CCDF	conditional, cumulative-distribution function
CCG	Centre for Computational Geostatistics
CDF	cumulative-distribution function
GSLIB	Geostatistical software library
IK	indicator kriging
MG	multivariate-Gaussian
MGK	multivariate-Gaussian kriging
MIK	multiple-indicator kriging
MSE	mean-squared error
OK	ordinary kriging
PDF	probability density function
PLMR	piecewise, linear model of regionalization
RF	random function
RMSE	root-mean-squared error
RV	random variable
SK	simple kriging

Chapter 1

Introduction

The initial and boundary conditions of diverse geological processes that lead to a concentration of minerals in the Earth's crust are never completely understood. The list of circumstances required to yield a mineral deposit include the following: 1) a source fluid bearing the necessary minerals, 2) a network of pathways allowing the fluids to migrate, 3) favourable permeable and impermeable rock types acting respectively as mineralization hosts and barriers to further fluid transport, 4) a chemical mechanism precipitating the minerals from solution, and 5) chemical or physical processes concentrating the minerals. Though unrelated to the aforementioned processes, the roles of uplift and erosion are paramount in providing access for geological sampling and subsequent mineral resource evaluation. If an economic assessment of the resource model demonstrates the deposit to be commercially viable, the deposit can be mined. The desire to study and model the outcomes of these natural processes has led to the development of geostatistics, which combines the fields of applied statistics and geology.

The objective of the mineral resource model is to represent, as best as possible, the outcomes of subsurface phenomena in order to understand technical, environmental, social, and economic impacts of a potential future mine. The mine plan relies on the mineral resource model to economically extract the mineral reserve. The mineral resource model is impacted greatly by sampling, which is often highly selective and focused in zones of higher anticipated value due to high acquisition costs.

Certain mineral deposits have characteristically positively skewed distributions, such as those containing diamonds, gold, uranium, and to a lesser extent, silver and base metals. The metal content represented by the relatively small proportion of samples comprising

the upper tail of the distribution is often significant relative to the samples closer to the mean or median of the distribution. When the coefficient of variation is greater than two, which is not uncommon, outlier values negatively impact estimation quality (Journel, 1983). Logically, accurate delineation and quantification of high-grade zones in these deposits is essential to generate high-quality mine plans and reliable financial forecasts.

Common deterministic estimation techniques used to quantify mineral resources in these types of deposits include, in order of increasing complexity, nearest-neighbour, inverse-distance weighting, and kriging. The nearest-neighbour technique assigns the value of the nearest sample to the unsampled location. Inverse-distance weighting, as the name implies, weights nearby samples according to their proximity to the unsampled location. An exponent, commonly of value two or three, can be assigned to the distance weighting so as to give more weight to nearby samples. Kriging is a geostatistical technique that minimizes estimation error, which is the criterion used to define a ”best” estimator. Kriging considers the spatial relationships and similarities of neighbouring samples to each other and to the unsampled location. While nearest neighbour and inverse-distance weighting are commonly used estimation techniques, they suffer from two primary drawbacks in comparison to kriging: 1) estimation error is high around outlier samples due to inflated estimated grades proximal to the outlier values (Journel, 1983; Rossi & Deutsch, 2014), and 2) they cannot be used to quantify estimation uncertainty. All three techniques require prior treatment of the outlier values through top cutting, also known as grade capping, to mitigate over-estimation at unsampled locations, which paradoxically, can significantly reduce the economic value of the deposit (e.g., Babakhani, 2014; Chiquini & Deutsch, 2017; Costa, 2003; Parker, 1991; Rivoirard, Demange, Freulon, Lécureuil, & Bellot, 2013).

Probabilistic estimation quantifies the distribution of uncertainty at the unsampled location. Simulation techniques such as sequential Gaussian simulation are commonly employed because they are theoretically straightforward and provide a joint distribution of uncertainty between a set of realizations. Furthermore, today’s computer processing power

has greatly reduced simulation run times. Non-simulation based techniques include indicator kriging (IK), commonly implemented as multiple-indicator kriging (MIK), developed by Journel (1983), and multivariate-Gaussian kriging (MGK) developed by Verly (1983). Despite its relatively time-intensive workflow relative to MGK and simulation methods, IK remains a commonly used technique in the mining industry. Three motivating factors for the development of the IK methodology include 1) non-parametric construction of the probability distribution, 2) management of outlier values through the indicator transform, and 3) simple incorporation into the estimate of secondary, categorical data such as the presence or absence of a rock type, geological structure, or some other geological feature (C. Deutsch & Journel, 1998; Goovaerts, 1997). Counterbalancing the positive features of the IK method are several implementation challenges, which form the basis of this research.

The following sections in this chapter review fundamental geostatistical concepts and notation before introduction of the indicator formalism and IK. Estimation accuracy and precision concepts are briefly presented to provide the basis for comparisons of estimate quality. The IK implementation challenges are presented and are used to focus the research goals. This chapter concludes with a thesis statement and an outline of the research in the subsequent chapters.

1.1 Fundamental geostatistical concepts

1.1.1 Random variables and random functions

A key contribution to the geostatistical framework is the underlying random function (RF) model, which permits the inference of statistical properties of the sample distribution (Matheron, 1970). It recognizes the inherent uncertainty of estimation at unsampled locations and provides the mathematical tools, through probabilistic theory, for estimation and assessment of estimation uncertainty (C. Deutsch & Journel, 1998; Goovaerts, 1997; Isaaks & Srivastava, 1989; Journel, 1989). As the field of geostatistics has developed since its incep-

tion under Georges Matheron, mineral resource estimates have evolved from deterministic, single-value outcomes at each unsampled location to more complex probabilistic realizations that assess joint distributions of uncertainty between many variables. Simulation methods are outside the scope of this thesis, but their importance to the field of geostatistics and mineral resource estimation warrants mentioning.

A random variable (RV) can take on a series of values based on a probability distribution, either continuous or discrete. Continuous RVs have a range of values with a natural order. These include elemental concentrations, rock densities, and other rock properties. Discrete RVs have a finite number of outcomes and are often unordered. Geological examples include rock types or the binary presence or absence of a feature such as a fault or alteration style (Goovaerts, 1997). Random variables are denoted by a capital letter, Z, with an outcome of the RV denoted by its lower-case form, z. With earth-science datasets, RVs are often location dependent, thus the notation $Z(\mathbf{u})$ is used to describe the RV at the location-coordinates vector of a finite number of locations, $\mathbf{u}_\alpha,\ \alpha = 1, \dots, n$.

The univariate cumulative-distribution function (CDF) of a continuous RV, $Z(\mathbf{u})$, is defined as the probability that the RV $Z(\mathbf{u})$ is less than or equal to the specified value, z, as: $F_Z(\mathbf{u}; z) = Prob\{Z(\mathbf{u}) \leq z\},\ \forall\, z$. The function must be licit, meaning that it is non-decreasing, $F(\mathbf{u}; z_k) \leq F(\mathbf{u}; z_{k'})\ \forall\ k < k'$, and the cumulative probabilities are bounded within $[0, 1]$. The global univariate CDF does not account for neighbouring information at other locations, $\mathbf{u}'$, and is commonly referred to as the prior uncertainty.

Regionalization of an RV over the finite number of locations in a domain, $\mathcal{A}$, defines an RF: $\{Z(\mathbf{u}_\alpha)\ \forall\ \mathbf{u} \in \mathcal{A}\}$. Consider M-number of data types such as geochemical concentration (grade), mineralogical suites, or other rock properties, where each property is indexed as $m = 1, \dots, M$. The RF-notation defined in Equation 1.1 is unchanged from the RV-notation, except it represents the M-variate CDF (Goovaerts, 1997):

$$F_{Z_1,\dots,Z_M}(\mathbf{u}_1, \dots, \mathbf{u}_n; z_1, \dots, z_M) = Prob\{Z(\mathbf{u}_1) \leq z_1, \dots, Z(\mathbf{u}_n) \leq z_M\} \tag{1.1}$$

The M-variate CDF models the joint uncertainty and spatial variability about the n-number of data locations in the same manner the univariate CDF is used to model uncertainty about the value $z(\mathbf{u}_\alpha)$ (C. Deutsch & Journel, 1998). This permits the definition of the random function probabilistically and facilitates the use of the expected-value operator, $E\{\}$, to express key summary statistics of the random functions (Matheron, 1970). The conditional, cumulative-distribution functions (CCDFs) defining the distributions of uncertainty at a location given the surrounding data values are referred to as the posterior uncertainty (Equation 1.2). Note that unless it is required, the data-type index, m, is dropped henceforth to simplify notation.

$$F_Z(\mathbf{u}; z|(n)) = Prob\{Z(\mathbf{u}) \leq z|(n)\}, \forall\, z \tag{1.2}$$

1.1.2 The decision of stationarity and the variogram

Repetitive sampling is required to make statistical inferences about a distribution and its underlying parameters such as its mean, $\overline{m}$, and variance, σ^2, values. This is impractical in a real-world setting because a sample can only be taken once at each location. No material remains in that precise location to be drawn again. The decision of stationarity is made to navigate this challenge. It permits samples to be grouped and considered for statistical inference as a single "homogeneous" set within the domain limits, $\mathcal{A}$. Note that stationarity needed to infer the characteristics of the RF, but it is not a characteristic of the natural phenomena itself, since it cannot be proven or refuted based on the data (Goovaerts, 1997). Domain limits can be as simple as defining a single rock type containing mineralization, or more complex such as a network of faults acting as fluid conduits. Regardless of the domain complexity, geostatistical modelling requires simplification to infer univariate statistics such as the mean and variance from the histogram, bivariate statistics such as spatial covariance and semi-variogram functions, or other higher-order multivariate statistics.

The challenge of obtaining multiple samples to infer distribution parameters is managed by evaluating all sample pairs, $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$, separated by a lag distance vector, $\mathbf{h}$, where

$\mathbf{h} = \mathbf{u}_\alpha - \mathbf{u}_\beta, \alpha \neq \beta$. The criteria for a second-order stationary random function are 1) each RV has a constant mean, $\overline{m}$, within $\mathcal{A}$ (Equation 1.3), and 2) the covariance function, $C(\mathbf{h})$, between two RVs exists and depends solely on the **h**-lag vector (Equation 1.4).

$$E\{Z(\mathbf{u})\} = \overline{m}_Z \tag{1.3}$$

$$E\{[Z(\mathbf{u}) - \overline{m}_Z][Z(\mathbf{u}+\mathbf{h}) - \overline{m}_Z]\} = C_Z(\mathbf{h}) \tag{1.4}$$

The semi-variogram, $\gamma(\mathbf{h})$, hereafter referred to as the variogram, is commonly used to infer the covariance (Equation 1.5). It is a measure of dissimilarity between two RVs separated by lag vector, **h**. The number of sample pairs separated by, **h**, is denoted by, $N(\mathbf{h})$. The ^-symbol is used to denote inference from a set of samples over a finite number of data.

$$\begin{aligned} \gamma_Z(\mathbf{h}) &= \frac{1}{2} E\{[Z(\mathbf{u}) - Z(\mathbf{u}+\mathbf{h})]^2\} \\ \widehat{\gamma}_Z(\mathbf{h}) &= \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \Big[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})\Big]^2 \end{aligned}, \qquad \forall\ \mathbf{u} \in \mathcal{A} \tag{1.5}$$

where the variogram value, $\gamma(\mathbf{h})$, represents the average of the distance, d, of the data pairs to the $45°$-line in the bivariate h-scatter plot (Figure 1.1). Under second-order stationarity, there is a straightforward relationship between the variogram and the covariance (Equation 1.6).

$$\gamma(\mathbf{h}) = \sigma^2 - C(\mathbf{h}) \tag{1.6}$$

where, σ^2, denotes the stationary variance.

Experimental variogram points are calculated from composited sample data. If the orientations of the local geological structures are not well understood, an omnidirectional variogram is calculated by considering all data pairs at a specified **h**-lag distance. This yields an isotropic variogram. If the geology is understood, then angular constraints may be applied to restrict the data pairs considered in the calculation at each **h**-lag distance.

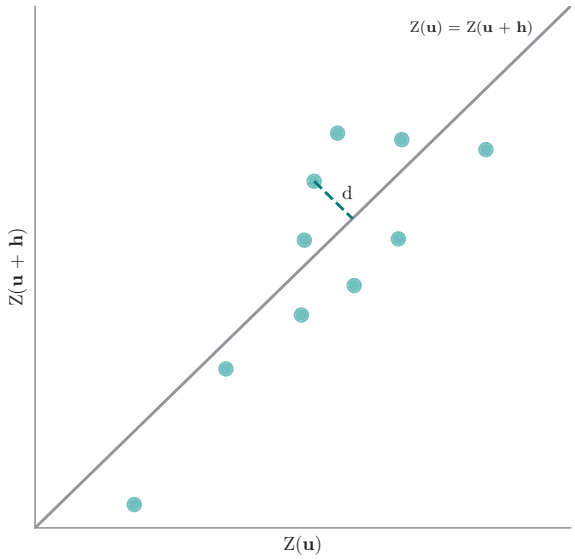


Figure 1.1: Example of bivariate h-scatter plot.

This defines orthogonal directions of maximum and minimum continuity, or in terms of covariance, directions of respective least and greatest dissimilarity. These directions often correspond to the strike and dip of the geological unit or package of units comprising the region or domain, $\mathcal{A}$. A third direction is also defined when the data are in three dimensions. The following terms apply to the directions in decreasing order of spatial continuity: major, minor, and vertical. A common alternative naming convention uses the terms major, semi-major, and minor.

1.1.3 Variogram modelling and the linear model of regionalization

Modelling of the experimental variograms along the major, minor, and vertical axes permits the calculation of the variogram value at any possible distance and direction. Though any analytical, continuous function could be used to model the experimental points, only certain functions that guarantee a positive-definite covariance matrix can be used. The most commonly used functions for modelling geochemical values in mining and petroleum datasets are the spherical and exponential models (Equation 1.7 and Equation 1.8). The

Gaussian model is often used to model surface boundaries or regions where there is significant short-range continuity (Equation 1.9). The nugget effect describes the variogram at lag distance, $\mathbf{h} = 0$ (Equation 1.10). This value is theoretically zero, but due to sampling and analysis errors and mineralogical heterogeneity at ranges smaller than the sampling scale, a discontinuity may be modelled at the variogram origin (Goovaerts, 1997). Other valid analytical functions exist, but they are not used in this work (e.g., Chilés & Delfiner, 2012).

$$Sph(\frac{h}{r}) = \begin{cases} 1.5\frac{h}{r} - 0.5(\frac{h}{r})^3, & \text{if } h \leq r \\ 1, & \text{otherwise} \end{cases} \tag{1.7}$$

$$Exp(h) = 1 - exp\left(\frac{-3h}{r}\right) \tag{1.8}$$

$$Gauss(h) = 1 - exp\left(\frac{-3h^2}{r^2}\right) \tag{1.9}$$

$$Nugg(h) = \begin{cases} 0, & \text{if } h = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1.10}$$

The range parameter, r, defines the distance at which the variogram reaches the stationary variance, also known as the sill. The sill is often standardized to unity to facilitate easier modelling and interpretation of the variogram. The exponential and Gaussian models reach the sill asymptotically, so their effective range values are considered to be at 95-percent of the sill. The models in Equation 1.7 through Equation 1.10 are expressed in their isotropic form, where the scalar, $h = |\mathbf{h}|$. To account for anisotropy, the distance is calculated using the modelled ranges in Equation 1.11.

$$h = \sqrt{\left(\frac{h_{major}}{r_{major}}\right)^2 + \left(\frac{h_{minor}}{r_{minor}}\right)^2 + \left(\frac{h_{vertical}}{r_{vertical}}\right)^2} \tag{1.11}$$

Goovaerts (1997) nicely summarizes the development of the linear model of regionalization that is required to completely describe sample covariance in a region or domain, $\mathcal{A}$. It is often necessary to combine, or nest, two or more variogram functions to adequately fit the calculated experimental variogram points. The linear model of regionalization addresses this need by creating a new random function, $Z(\mathbf{u})$, that is a linear combination of $(L+1)$-independent random functions, $X^l(\mathbf{u})$, each with a mean of zero, that are defined by some positive-definite variogram function, $g^l(\mathbf{h})$ (Equation 1.12).

$$Z(\mathbf{u}) = \sum_{l=0}^{L} c^l X^l(\mathbf{u}) + \overline{m} \tag{1.12}$$

where, c^l, is a scaling factor. Independence of the random functions means their covariance is zero when, $l \neq l'$, which simplifies the inference of the variogram, $\gamma(\mathbf{h})$, to a linear combination of variogram functions, $g^l(\mathbf{h})$ (Equation 1.13):

$$\widehat{\gamma}(\mathbf{h}) = \sum_{l=0}^{L} b^l g^l(\mathbf{h}) \tag{1.13}$$

where, $b^l = (c^l)^2 \geq 0$, is the variance contribution of the $g^l(\mathbf{h})$-variogram model. The nugget effect model is the $(l=0)$-term by convention (Goovaerts, 1997).

1.1.4 Deterministic estimation

Deterministic estimation techniques commonly utilized in mineral resource estimates of metalliferous deposits include nearest neighbour, inverse-distance weighting, and kriging. Further information on the first two techniques can be found in introductory geostatistical texts (e.g., Rossi & Deutsch, 2014).

Kriging is a least-squares, linear regression method that yields an estimate, $\widehat{z}$, at an unsampled location, $\mathbf{u}_0$, using a set of spatially correlated conditioning data (Equation 1.14).

$$\widehat{z}(\mathbf{u}_0) - \overline{m}(\mathbf{u}) = \sum_{\alpha=1}^{n} \lambda_\alpha [z(\mathbf{u}_\alpha) - \overline{m}(\mathbf{u}_\alpha)] \tag{1.14}$$

where, λ_α, is the kriging weight assigned to the sample datum, z_α. The mean term,

$\overline{m}(\mathbf{u})$, is the expected value of the RV, $Z(\mathbf{u})$.

Kriging is an unbiased estimator that minimizes estimation error variance, $\sigma_Z^2(\mathbf{u}) = E\{[\widehat{Z}(\mathbf{u}_0) - Z(\mathbf{u}_0)]^2\}$, which leads to the system of normal equations (Equation 1.15). The system of equations is solved for the kriging weights, which are substituted into the kriging estimator. The notation, $C_{\alpha,\beta}$ and $C_{\alpha,0}$, respectively denote the covariance between two sample locations, $C(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, and the covariance between the sample and the unsampled location, $C(\mathbf{u}_\alpha - \mathbf{u}_0)$. The estimation variance is defined in Equation 1.16.

$$\sum_{\beta=1}^{n} \lambda_\beta C_{\alpha,\beta} = C_{\alpha,0}, \quad \alpha = 1, \ldots, n \tag{1.15}$$

$$\widehat{\sigma}_Z^2(\mathbf{u}_0) = \sigma_Z^2 - 2\sum_{\alpha=1}^{n} \lambda_\alpha C_{\alpha,0} + \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha \lambda_\beta C_{\alpha,\beta} \tag{1.16}$$

There are several variants of kriging that vary by the treatment of the mean term, $\overline{m}$, in Equation 1.14. Simple kriging (SK) assumes strict stationarity, such that the mean is known and constant throughout the domain: $\overline{m}(\mathbf{u}) = \overline{m}$, $\forall\ \mathbf{u} \in \mathcal{A}$ (Equation 1.17). The estimation error, $\widehat{\sigma}_Z^2$, is minimized to yield the optimized kriging weights and the simple-kriging variance, $\widehat{\sigma}_{SK}^2$, (Equation 1.18).

$$\widehat{z}_{SK}(\mathbf{u}_0) = \sum_{\alpha=1}^{n} \lambda_\alpha^{SK} z(\mathbf{u}_\alpha) + \left[1 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK}\right] \cdot \overline{m}_Z(\mathbf{u}_\alpha) \tag{1.17}$$

$$\widehat{\sigma}_{SK}^2(\mathbf{u}_0) = \sigma_Z^2 - \sum_{\alpha=1}^{n} \lambda_\alpha^{SK} C_{\alpha,0} \tag{1.18}$$

In ordinary kriging (OK), the stationarity requirement is relaxed such that the mean, though unknown, is constant only within the local search neighbourhood, $W(\mathbf{u})$, that is centred on $\mathbf{u}$: $\overline{m}(\mathbf{u}') = \overline{m}$, $\forall\ \mathbf{u}' \in W(\mathbf{u})$. An additional constraint is applied such that the kriging weights sum to unity (Equation 1.19). This leads to an updated minimized error variance that considers the constraint (Equation1.20).

$$\widehat{z}_{OK}(\mathbf{u}_0) = \sum_{\alpha=1}^{n} \lambda_\alpha^{OK} z(\mathbf{u}_\alpha), \quad \text{with} \quad \sum_{\alpha=1}^{n} \lambda_\alpha^{OK} = 1 \tag{1.19}$$

$$\widehat{\sigma}^2_{OK}(\mathbf{u}_0) = \sigma^2_z - \sum_{\alpha=1}^{n} \lambda^{OK}_{\alpha} C_{\alpha,0} - \mu_{OK} \tag{1.20}$$

where the LaGrange multiplier, μ, accounts for the constraint forcing the kriging weights to sum to unity. Other kriging variants are not required in this thesis, but they can be referenced in most geostatistical texts (e.g., Chilés & Delfiner, 2012; Goovaerts, 1997; Rossi & Deutsch, 2014).

1.1.5 Probabilistic estimation

Probabilistic estimation is motivated by the need to define the distribution of uncertainty at the estimate location, $z(\mathbf{u}_0)$. Kriging is used as the engine to generate the CCDF (Equation 1.2). Multivariate-Gaussian kriging and IK belong to this family of techniques. With MGK, the $Z(\mathbf{u})$-random variable is transformed from original units to a standard-normal, RV, $Y(\mathbf{u})$, before SK estimation under a multivariate-Gaussian (MG) assumption (e.g., Goovaerts, 1997; Pyrcz & Deutsch, 2018; Verly, 1983) (Equation 1.21).

$$y = G^{-1}_Y(F_Z(z)) \quad \forall\ z \tag{1.21}$$

where G^{-1} is the inverse of the standard-normal, CDF. The normal-score transform is a quantile-to-quantile transform that matches quantiles, p, of the distribution in original units to the same p-quantiles of the distribution in Gaussian units.

It is noteworthy that after a normal-score transform, the relationship between the RVs is not necessarily multivariate Gaussian. This is due to non-linear, constraint, or non-uniform variance (heteroscedastic) relationships (J. Deutsch & Deutsch, 2011). Multivariate normality is difficult to check in practice because it requires locating a sufficient number replicate sets of data with the same spatial configuration. This quickly becomes untenable when there are a large number of unsampled locations. There are multiple statistical tests available for the MG distribution, many of which are well summarized in Gnanadesikan (1997) and Thode Jr. (2002), but they often require significant investigation into the marginal dis-

tributions. For earth-sciences datasets involving many variables, this can be quite time intensive and impractical for the geostatistics practitioner (J. Deutsch & Deutsch, 2011).

If a spatial distribution is considered to be bivariate Gaussian, then an MG distribution is assumed (Goovaerts, 1997). It is mathematically tractable and has several useful properties that facilitate its use in geostatistics (Equation 1.22): 1) all lower-order marginal and conditional distributions are Gaussian, 2) all conditional expectations are linear functions of the conditioning data, 3) all conditional variances are independent of the data values (homoscedastic), and 4) if two RVs are uncorrelated, they are also independent (C. Deutsch & Journel, 1998; Goovaerts, 1997).

$$\mathcal{N}(\mathbf{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\left(\sqrt{2\pi}\right)^{d}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right] \tag{1.22}$$

where, d, is the dimensionality of, $\mathbf{y}$. Note that, $\boldsymbol{\mu}$, is a $(d \times 1)$-vector and the covariance matrix, $\boldsymbol{\Sigma}$, is a $(d \times d)$-positive-definite, symmetric matrix. The determinant of $\boldsymbol{\Sigma}$ is denoted as $|\boldsymbol{\Sigma}|$. Simple-kriging estimation under the MG assumption leads to a univariate-Gaussian CCDF at the unsampled location that is fully parameterized by the SK estimate, $\widehat{y}_{SK}$, and the variance, $\widehat{\sigma}^2_{SK}$, values. The probability distribution in original units is calculated through a back-transform of the p-quantiles discretizing the CCDF (Equation 1.23).

$$z_p = F_Z^{-1}(G_Y(\widehat{\sigma}^2_{SK}G_Y^{-1}(p) + \widehat{y}_{SK})), \quad p = 1, \ldots, P \tag{1.23}$$

where the terms, $(\widehat{\sigma}^2_{SK}G_Y^{-1}(p)+\widehat{y}_{SK})$, denote a Gaussian p-quantile of the local CCDF determined from the SK mean and variance values. The term, $G_Y()$ is the corresponding quantile of the global distribution in Gaussian units before the quantile-backtransform, $F_Z^{-1}()$, to original units to yield the z_p-quantile. The quantity of p-quantiles often numbers between one hundred and two hundred. The MGK estimate, $\widehat{z}_{MGK}(\mathbf{u}_0)$, and variance, $\widehat{\sigma}^2_{MGK}(\mathbf{u}_0)$, are inferred through calculation of the mean and variance of the z_p-quantiles.

1.2 Indicator formalism and indicator kriging

The binary indicator transform of the Z-random variable, $I(\mathbf{u}; z)$, can be applied to continuous and discrete variables. The outcomes are either, 1 or 0, depending if they are less than or greater than the threshold, $z_k,\ k = 1, \dots, K$ at location, $\mathbf{u}_\alpha$ (Equation 1.24). Multiple thresholds in increasing order, $z_{k'} > z_k,\ \forall\ k' > k$, can be used to discretize the CDF of the continuous Z-RV. When there is more than one threshold, each data location stores an indicator value corresponding to each threshold. By transforming all values to either one or zero based on their relative position to some threshold z_k, outlier values are eliminated.

$$I(\mathbf{u}; z_k) = Prob\{Z(\mathbf{u}) \leq z_k\} = \begin{cases} 1, & \text{if } Z(\mathbf{u}) \leq z_k \\ 0, & \text{otherwise} \end{cases} \qquad k = 1, \dots, K \tag{1.24}$$

The prior distribution of the binary indicator RV, $I(\mathbf{u}; z_k)$, is the declustered n-number of samples in the domain, which is the stationary mean (Equation 1.25). Since it only has two outcomes, the expected value, $E\{I(\mathbf{u}; z_k)\}$, is easily calculated. The expected variance, $Var\{Z\} = E\{Z^2\} - [E\{Z\}]^2$, reduces to the result in Equation 1.26.

$$\begin{aligned} E\{I(\mathbf{u}; z_k)\} &= 1 \cdot Prob\{Z(\mathbf{u}) \leq z_k\} + 0 \cdot Prob\{Z(\mathbf{u} > z_k\} \\ &= Prob\{Z(\mathbf{u}) \leq z_k\} = F_Z(z_k) = \overline{m}_k \end{aligned} \tag{1.25}$$

$$\begin{aligned} Var\{I(\mathbf{u}; z_k)\} &= Prob\{Z(\mathbf{u}) \leq z_k\} \cdot (1 - Prob\{Z(\mathbf{u}) \leq z_k\}) \\ &= F_Z(z_k) \cdot (1 - F_Z(z_k)) \end{aligned} \tag{1.26}$$

Indicator variograms reflect the probability that the z-values separated by **h**-lag distance are on either side of the threshold, z_k, in a h-scatterplot (Figure 1.2). They do not show the spatial continuity of ones or zeroes by themselves, but the probability of transition between the indicator values as a function of the **h**-lag distance at a threshold, z_k. The indicator variogram value is the number of values falling in the hatched area for each

indicator threshold divided by twice the number of data pairs, $N(\mathbf{h})$ (Equation 1.27).

$$\widehat{\gamma}_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + \mathbf{h}; z_k)]^2, \quad k = 1, \dots, K \tag{1.27}$$

The indicator variograms relate to the same underlying continuous RV (Carvalho & Deutsch, 2017). Figure 1.2 illustrates the transition between two indicator thresholds. The h-scatter plot is contoured by density, with the grey and teal hatched regions representing the data pairs corresponding respectively to the z_k and z_{k+1}-thresholds. Moving from one threshold to the other, some data pairs are lost while others are gained. Many remain shared (cross-hatched region), which means that the experimental, indicator-variogram modelling should not vary significantly between thresholds (Carvalho & Deutsch, 2017).

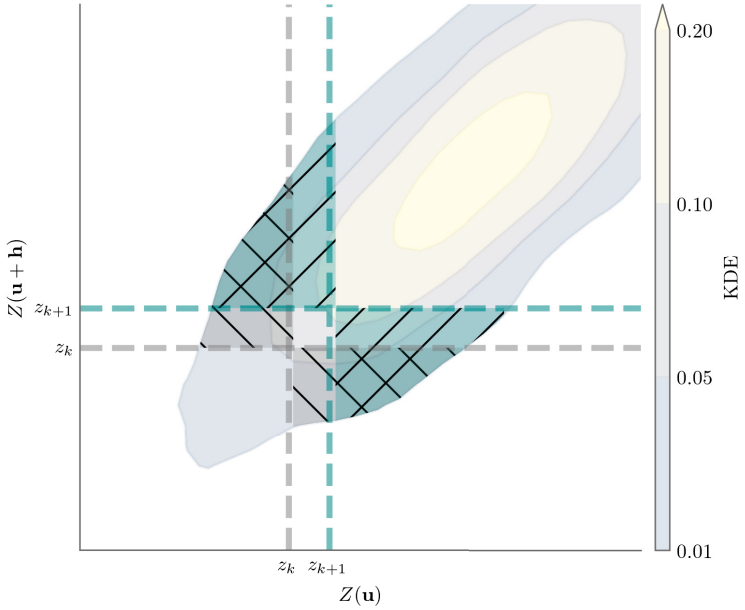


Figure 1.2: Hatched regions denote region of h-scatter plot (contoured by density) contributing to the indicator variogram for indicators, z_k and z_{k+1}.

With bivariate-Gaussian data, indicator variograms exhibit a characteristic symmetrical destructuration away from the median quantile. For example, the $z_{0.1}$ and $z_{0.9}$-quantiles have the same indicator variogram. This is the symmetric property of the Gaussian distribution. It is important to note that the Gaussian distribution does not permit spatial correlation of high or low values (Goovaerts, 1997). Non-bivariate-Gaussian data exhibit

asymmetric destructuration away from the median indicator.

The posterior indicator probability, $\widehat{i}(\mathbf{u}_0; z_k|(n))$, is directly estimated at each indicator threshold by solving the ordinary-indicator-kriging systems of equations at each unsampled location for each threshold, z_k (Equation 1.28). For *K*-thresholds, the CCDF at comprises *K*-points at location $\mathbf{u}_0$. The term MIK is used when more than one indicator threshold is considered.

$$\begin{aligned}\widehat{i}(\mathbf{u}_0; z_k|(n)) &= \widehat{Prob}\{Z(\mathbf{u}_0) \leq z_k|(n)\} \\ &= \sum_{\alpha=1}^{n} \lambda_\alpha(\mathbf{u}; z_k) \cdot i(\mathbf{u}_\alpha; z_k)\end{aligned} \tag{1.28}$$

where the kriging weights, $\lambda_\alpha(\mathbf{u}; z_k)$, are multiplied by the sample indicator value at the k^{th}-threshold, $i(\mathbf{u}; z_k)$. Simple kriging could be implemented. The estimated value, $\widehat{z}_{MIK}(\mathbf{u}_0)$, is the sum of each indicator class mean weighted by the estimated probability of the class (Equation 1.29).

$$\begin{aligned}\widehat{z}_{MIK}(\mathbf{u}_0) = \sum_{k=0}^{K} \left[\widehat{i}(\mathbf{u}_0; z_{k+1}|(n)) - \widehat{i}(\mathbf{u}_0; z_k|(n))\right] \cdot \widehat{m}_k \\ i(\mathbf{u}_0; 0) = 0;\ \ i(\mathbf{u}_0; z_{K+1}) = 1\end{aligned} \tag{1.29}$$

where $i(\mathbf{u}_0; z_{k+1}|(n))$ and $i(\mathbf{u}_0; z_k|(n))$ are the estimated probabilities bounding the k^{th}-indicator class, and $\widehat{m}_k$ is its mean value. Note that $\widehat{m}_k$ is considered stationary. The two constraints ensure the conditional distribution exists for all *z*-values. The volume support of the estimated distribution is the same as the sample data, so a volume-support correction is required in mining applications.

A required post-processing step, referred to as order-relations corrections, involves 1) checking and correcting the conditional probabilities to ensure they are within the range, $[0, 1]$, and 2) the conditional probabilities are monotonically increasing to meet the requirements of a valid CDF. Order-relations errors are a direct result of the varying modelled ranges of the indicator variograms, which means there is departure from the reproduction of the indicator covariance models and constraint intervals (C. Deutsch & Journel, 1998).

Several correction methods are available, but the one employed in the *GSLIB*-software is an average of upward and downward corrections. In the upward direction, the conditional probabilities are checked sequentially from the first threshold to the last, z_1 to z_k, to ensure they meet the order-relations criteria. This check is repeated moving in the opposite direction, z_k to z_1. The upward and downward adjusted probabilities are then averaged to yield the final conditional probability values.

After order-relations corrections, additional post processing is required to interpolate the quantile values between the indicator thresholds, and to extrapolate the lower and upper tails from the bottom and top thresholds, respectively, to satisfy the constraints in Equation 1.29. Several methods exist, including linear and power-model interpolation between the thresholds. Extrapolation of the tails can utilise linear, hyperbolic, and power-model functions. An alternative model that scales the conditional distribution based on the global representative distribution is considered to be best practice (Carvalho & Deutsch, 2017).

1.3 Assessment of estimation quality

Several criteria are used to assess overall estimation quality. Note that the terms *verification* and *validation* are avoided, despite their ubiquitous use in the mining and petroleum industries, because an estimate can never truly be considered valid or verified. This is due to natural systems being open and our incomplete access to them during sampling (Oreskes, Shrader-Frechette, & Belitz, 1994). The estimate can be *confirmed* to be a good predictor of the observed data given the stated hypotheses, but this does not provide certainty as to the future performance of the model given additional data, or a change in the underlying hypotheses.

Confirmation of the quality of the estimate as a predictor is quantified through an assessment of bias, accuracy, and precision. Current best practice utilizes *K*-fold analysis. Synthetic datasets are used in this research, which provide exhaustively known values by which to compare the estimates, and preclude need for *K*-fold analysis. Estimation bias is

the expected difference between the estimated value and known truth value (Equation 1.30). Conditional bias, which represents the slope of regression between the estimate and the truth, is the expected value of the truth given the estimated value (Equation 1.31).

$$Bias = E\{\widehat{Z}(\mathbf{u})\} - E\{Z(\mathbf{u})\} \tag{1.30}$$

$$Conditional\ Bias = E\{Z(\mathbf{u})|\widehat{Z}(\mathbf{u}) = z(\mathbf{u})\} \tag{1.31}$$

Estimation accuracy is assessed by direct comparison of the estimated and known values values in a bivariate plot. When cut-off values are applied, false positives and false negatives capture misclassified estimates. False positives (Type I) occur when the estimated value is greater than the cut-off and the true value is less. False negatives (Type II) occur when the estimated value is lower than the cut-off and the true value is higher. In addition, accuracy can be checked by dividing the distribution of known values into probability intervals of varying widths and then summing the number of estimated values that fall within the probability intervals. Another measure of accuracy, mean-squared error (MSE), is the expected, squared difference between the estimated and known values (Equation 1.32).

$$E\{[\widehat{Z}(\mathbf{u}) - Z(\mathbf{u})]^2\} \tag{1.32}$$

Precision is a measure of the narrowness of a distribution. The mean of the variance of the conditional distributions is compared to assess estimation precision. A lower average variance indicates a more precise estimate. Note that the conditional distributions need to be discretized by the same number of quantiles to avoid overweighting by the tails.

1.4 Problem motivation and thesis outline

The IK methodology faces several challenges that are used to identify research opportunities: 1) it yields an inconsistent probability model that requires order-relations corrections to ensure a licit CCDF, 2) constant, indicator-class means are assumed, and 3) indicator estimation provides an incomplete probability distribution, which requires additional mod-

elling of the distribution tails. This research first seeks to better understand the implementation of MIK. Deviations of the estimated probability distribution from known values are investigated, and the variation of indicator-class means and the factors controlling their behaviour are documented. Next, the estimation performance of MIK and MGK is compared. The motivating research question is based on the assertion that MGK generates a better estimate than MIK in a Gaussian environment. In the presence of non-Gaussian data, when does MIK lead to a better estimate than MGK? This leads to the thesis statement:

> *Theoretical and practical understanding of multiple-indicator kriging leads to improved resource estimates and mining decisions.*

Chapter 2 discusses the research methodology and generation of two-dimensional (2-D), synthetic data used in the subsequent chapters. In Chapter 3 and Chapter 4, the components of the MIK estimator are investigated: the indicator-probability estimates, and the indicator-class means. Chapter 5 compares MIK and MGK estimation performance using several datasets ranging from Gaussian to highly non-Gaussian. The final chapter presents the research conclusions and contributions, limitations of the studies, and future research avenues.

Chapter 2

Study Data

Investigation of the estimated indicator probability model and the indicator class means requires an environment that permits comparison against known values. Real-world datasets are preferred, but the exact probability distributions cannot be known with certainty because of the highly multivariate nature of the geological processes forming the mineral deposits. The chapters that follow in this thesis mostly use synthetic, Gaussian datasets, which through the log-normal transform, can act as realistic analogues to the positively skewed distributions often encountered in precious-metals, uranium, and diamond deposits. Furthermore, use of the MG framework for the investigation simplifies analysis greatly, as the correct, conditional-probability distributions and conditional class mean values are known and easily accessible. Additional benefits include 1) exhaustive, gridded data that do not require declustering, 2) a known covariance function leading to exact indicator variograms and 3) a data distribution that is free of sampling and analysis errors.

The calculation of the correct, conditional-indicator-probability distributions (Chapter 3) and indicator-class-mean values (Chapter 4) using Gaussian datasets allows for a direct comparison to the MIK-estimated values in order to answer the following motivating questions: 1) By how much do they differ? 2) Where do they differ? 3) On which factors do the differences depend?

The motivating research question posed in Section 1.4 is based on the hypothesis that MIK is a better estimator than MGK once the data distribution is sufficiently non-Gaussian. Chapter 5 addresses this question by comparing estimation performance of MIK against MGK in a non-Gaussian environment. The environment is generated through a transform of the reference, Gaussian datasets into non-Gaussian RFs using the novel, piecewise, linear

model of regionalization (PLMR) developed by Pereira and Deutsch (2020a, 2020b).

This chapter is outlined as follows: The first section summarises the software and scripting environment used for this research. The remainder of the chapter describes the processes used to generate the study data used in the following chapters. Comments on the quality of the study data are made to address the robustness of observed results.

2.1 Software environment and executables

All computations are executed in Jupyter Notebooks (Kluyver et al., 2016). Jupyter is a web-based, interactive environment for data science and scientific computing across all programming languages. Python is the computing language used in this research. The following open-source computing packages are imported into Jupyter Notebook environments to undertake the numerical computations and data visualizations:

- Matplotlib, version 2.2.3: data visualization and plotting (Hunter, 2007),
- NumPy, version 1.15.4: numerical computations with data arrays (Harris et al., 2020),
- Pandas, version 1.0.2: data analysis using data structures (dataframes) (McKinney, 2010; Reback et al., 2020),
- Pygeostat, version 0.6.6: Geostatistical software library (GSLIB) executables wrapped in Python code and additional geostatistical functions developed by the CCG (J. Deutsch et al., 2015),
- scikit-learn, version 0.24.2: machine-learning library for Python (Pedregosa et al., 2011),
- SciPy, version 1.1.0: scientific computing including statistical and regression analysis (Virtanen et al., 2020)

Anaconda is the open-source platform used in this research to manage the Python

scripting environment and distribution of open-source computing packages (Anaconda Software Distribution, 2016). Software executables used to generate unconditional realizations, normal-score transforms, and ordinary kriged and indicator kriged estimates are sourced from C. Deutsch and Journel (1998).

2.2 Data generation

2.2.1 Gaussian random functions

All studies in this research utilize unconditionally simulated, 2-D, Gaussian, RFs as reference data. The RFs measure 512 m in the Northing and Easting directions (Figure 2.1). The data points are spaced 1 m apart. Isotropic covariance functions with low nugget values and a range of 64 m are used in Chapter 3 and Chapter 4. In Chapter 5, two RFs with anisotropic covariance functions are needed to generate a non-Gaussian RF (Table 2.1).

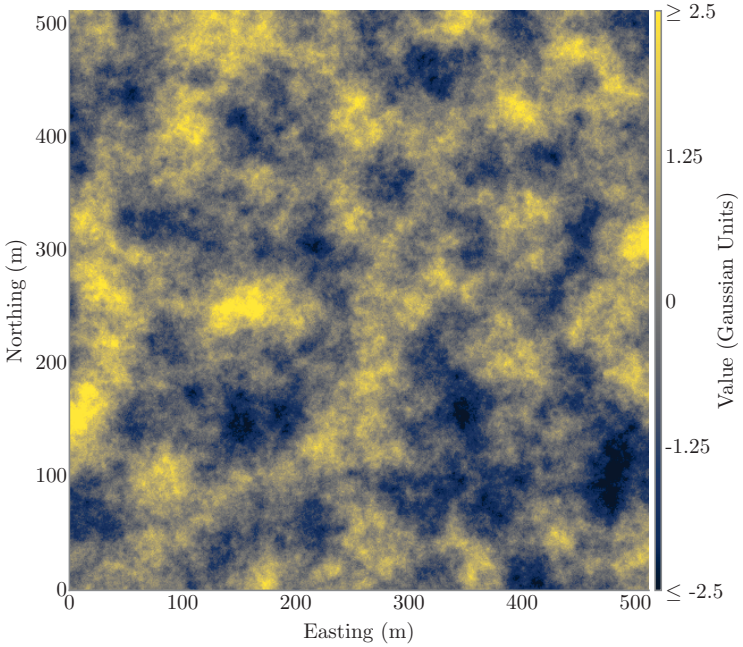


Figure 2.1: Example reference Gaussian RF.

A total of ten reference RFs are used in Chapter 3 and Chapter 5. A single reference RF is used in Chapter 4.

Table 2.1: Spatial structure of the RFs.

Chapter	Random Field Y_i	Factor $X^l(\mathbf{u})$	Structure Type $g(h)$	b^l	Range (m)		Direction (°)	
					r_{major}	r_{minor}	Az.	Dip
3	$i=1$	$l=0$	Nugget	0.00	0	0	0	0
		$l=1$	Spherical	1.00	64	64	0	0
4	$i=1$	$l=0$	Nugget	0.01	0	0	0	0
		$l=1$	Spherical	0.99	64	64	0	0
5	$i=1$	$l=0$	Nugget	0.01	0	0	0	0
		$l=1$	Spherical	0.99	128	32	0	0
	$i=2$	$l=0$	Nugget	0.01	0	0	0	0
		$l=1$	Gaussian	0.99	128	64	90	0

2.2.2 Gaussian variograms and indicator variograms

Indicator covariances are analytically derived through integration of the truncated, bivariate-Gaussian density represented by the hatched region of the bivariate scatterplot in Figure 2.2. This region, truncated at the $y_{0.90}$-indicator-probability threshold, represents the non-zero portion of the non-centred, indicator covariance. The indicator covariance differs from the calculation of the indicator variogram (Figure 1.2). Both calculations yield symmetric densities about the median, $y_{0.50}$-indicator-probability threshold. Note that the bivariate-Gaussian densities are calculated through a power-series expansion because there is no closed-form solution of the integral (Kyriakidis, Deutsch, & Grant, 1999). Calculation of the indicator lag distances and direction requires the Gaussian covariance function and the indicator-probability values as inputs for the GSLIB-program, *bigaus* (Figure 2.3).

All variograms are calculated directly from the gridded reference distributions at a lag distance of $\mathbf{h} = 4$ m. The Gaussian variograms are modelled with two spherical structures to the range of the original covariance function used to generate the reference RFs (Figure 2.3). The correct indicator variograms derived from this covariance function are illustrated in Figure 2.3. Note the symmetric destructuring of the indicator probability classes.

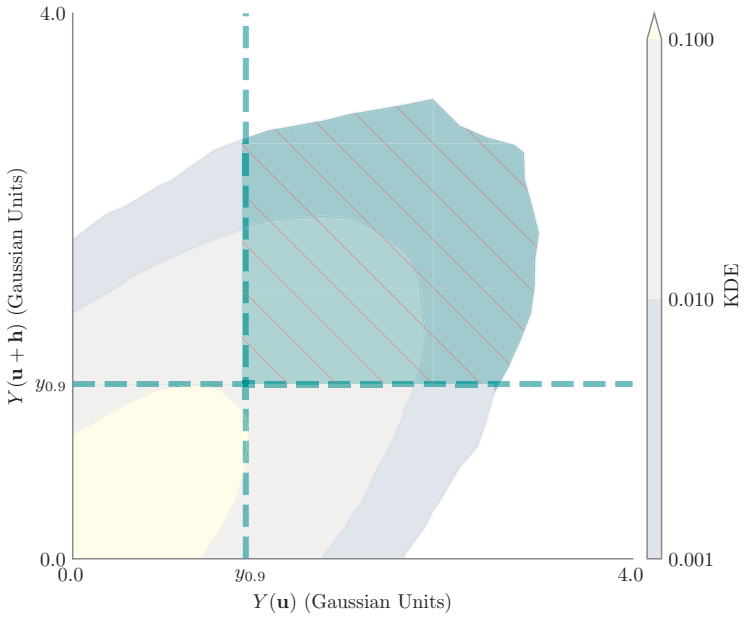


Figure 2.2: Integration region (hatched) of the bivariate-Gaussian density for calculation of the indicator covariance.

2.2.3 Non-Gaussian random functions

In Chapter 5, a novel methodology by Pereira and Deutsch (2020a) to unconditionally simulate a non-Gaussian, regionalized RV using the PLMR is used. It merges two Gaussian RFs through a combination of the indicator formalism (Equation 1.24) and a piecewise, linear mixing rule. Recall that in a purely multivariate-Gaussian framework, the full multivariate distribution is represented by a single covariance function (Journel & Alabert, 1989). This is convenient, but spatial disorder is inflated and leads to a loss of spatial connectivity and correlation in the data tails (Journel & Deutsch, 1993). The PLMR utilizes the indicator formalism to provide detailed structural information of the regionalized variable (RF) across multiple thresholds, while also permitting correlation of the extreme values (Pereira & Deutsch, 2020b). Generation of the non-Gaussian RF is broken down into three steps: 1) independently simulate stationary Gaussian RFs using different arbitrary covariance structures (Table 2.1), 2) define a Gaussian quantile cut-off value about which the RFs undertake a piecewise-linear transform (Equation 2.1), and 3) combine the transformed RF using the mixing rule in (Equation 2.2).

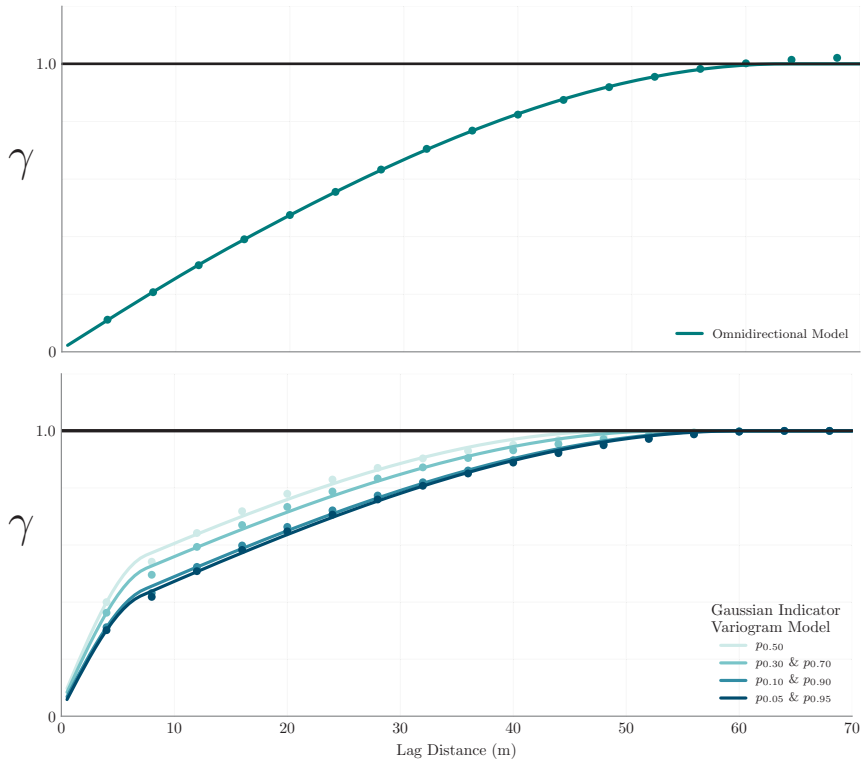


Figure 2.3: Correct indicator variograms (bottom) derived from the Gaussian variogram (top).

The transform of the RFs, $Y_1(\mathbf{u})$ and $Y_2(\mathbf{u})$, creates the regionalized variable, $Y_i(\mathbf{u}; a, q)$. It is a function of two parameters, 1) the truncation quantile, q, splits the RFs, and 2) the scalar value, a, factors the RFs linearly above and below the q-truncation value (Equation 2.1):

$$Y(\mathbf{u}; a, q) = \begin{cases} \sqrt{a} Y_i(\mathbf{u}), & \text{if } Y_i(\mathbf{u}) \leq y_q \\ \sqrt{1-a} Y_i(\mathbf{u}), & \text{if } Y_i(\mathbf{u}) > y_q \end{cases}, \quad i = 1, 2;\ a \in (0, 1);\ q \in [0, 1] \tag{2.1}$$

The terms, $\sqrt{a}$ and $\sqrt{1-a}$, sum to one when squared and prevent variance inflation in the $Y_i(\mathbf{u}; a, q)$-RF. The truncation quantile, $q = 0.5$, is considered to the be same value for each standard-normal RF, $Y_i(\mathbf{u})$, and corresponds to a Gaussian value of zero. After the transform, the RFs are combined following the mixing rule defined in Equation 2.2 to

create the new regionalized variable, $Z(\mathbf{u}; a, q)$.

$$Z(\mathbf{u}; a, q) = \begin{cases} \sqrt{a}Y_1(\mathbf{u}) + \sqrt{1-a}Y_2(\mathbf{u}), & \text{if } Y_1(\mathbf{u}) \leq y_q,\ Y_2(\mathbf{u}) \leq y_q \\ \sqrt{a}Y_1(\mathbf{u}) + \sqrt{a}Y_2(\mathbf{u}), & \text{if } Y_1(\mathbf{u}) \leq y_q,\ Y_2(\mathbf{u}) > y_q \\ \sqrt{1-a}Y_1(\mathbf{u}) + \sqrt{1-a}Y_2(\mathbf{u}), & \text{if } Y_1(\mathbf{u}) > y_q,\ Y_2(\mathbf{u}) \leq y_q \\ \sqrt{1-a}Y_1(\mathbf{u}) + \sqrt{a}Y_2(\mathbf{u}), & \text{if } Y_1(\mathbf{u}) > y_q,\ Y_2(\mathbf{u}) > y_q \end{cases}, \ a \in (0,1),\ q \in [0,1] \tag{2.2}$$

The magnitude of the *a*-scalar significantly impacts the transform. When $a = 0.50$, the transformed variables combine linearly and generate a Gaussian scenario, $Z(\mathbf{u}; 0.5, q)$ (Figure 2.4; Step 2 and 3, left side). If $a \neq 0.50$, mixing is non-linear, resulting in a non-Gaussian RF. When the a-parameter is very close to zero, or very close to one, mixing is highly non-linear and generates a highly non-Gaussian end-member (Figure 2.4; Step 2 and 3, right side). In this research, $a = 0.99$ is used to represent the highly non-Gaussian scenario. This mixing rule effectively removes the low-values from Y_1-RF and the high values from the Y_2-RF. It also clearly shows the north-south anisotropy of the high values superimposed over the east-west anisotropy of the low values (Figure 2.4). To adjust the relative "non-Gaussianity" of the $Z(\mathbf{u}; a, q)$-RF, six *a*-parameter values are considered:

$$a = \{0.50, 0.60, 0.70, 0.80, 0.90, 0.99\}$$

The *q*-parameter is kept constant at 0.5.

As a final note, the combination of RFs, Y_1 and Y_2, with principle axes of continuity oriented in perpendicular directions, provides an analogue to real-world geologic environments. For example, there may exist a low-grade dissemination of mineralization broadly aligned to an internal rock fabric that has been cut and re-mineralized by relatively high-grade, structurally controlled mineralization oriented in another direction. The PLMR provides a useful environment to test the performance of MIK and MGK. The degree of non-Gaussianity of the reference $Z(\mathbf{u}; a, q)$-RF can easily be adjusted through the *a*-parameter.

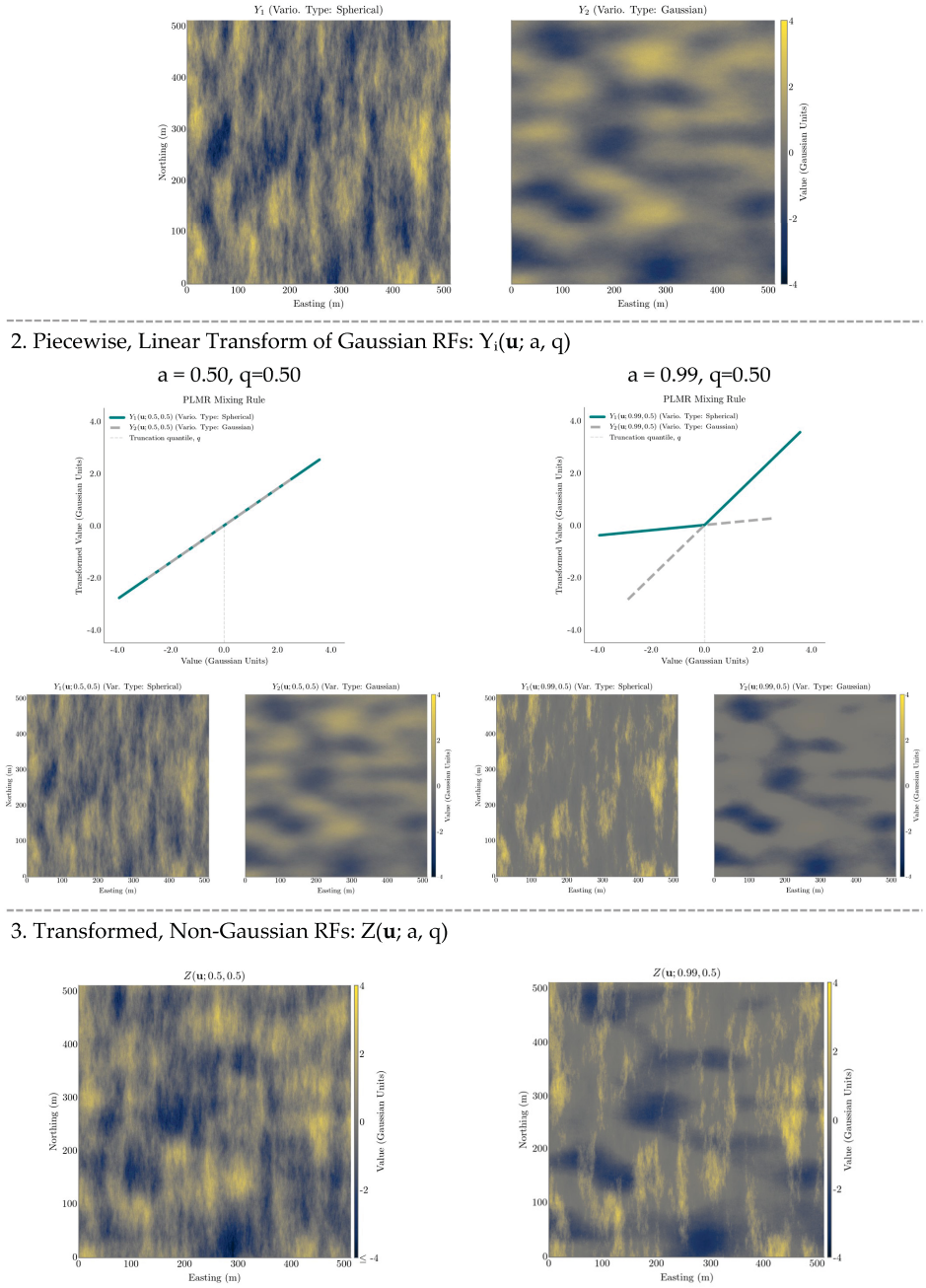


Figure 2.4: Graphical description of PLMR transform; 1. Input Gaussian RFs; 2. Different a-scalar values highlight the effect of the transform above and below, q (vertical, grey, dashed line); 3. The resulting non-Gaussian RF, $Z(\mathbf{u}; a, q)$, from mixing of the Y_i-components.

2.2.4 Estimation and final reference distributions

In Chapter 3 and Chapter 4, multiple, regularly spaced grids are used to sample the Gaussian reference RFs (in metres): 8×8, 16×16, and 32×32. In Chapter 4, an additional 64×64 m sample grid is used. The Gaussian variogram parameters are input for MGK estimation under the MG assumption. The search parameters are aligned with the variograms and consider up to 30 samples within twice the modelled variogram range. The estimates yield exhaustive, correct conditional distributions that are parameterized by the estimated mean, $\widehat{y}_{SK}$, and estimated variance, $\widehat{\sigma}^2_{SK}$.

A log-normal transform is needed in Chapter 4 to understand the impact of the constant, indicator-class means on the estimates. The log-normal distribution serves as an analogue to a positively skewed, precious-metals dataset encountered in the mining industry. Before estimation, the Gaussian sample values are converted to log-normal values that have a target log-normal distribution with mean, $\mu_Z = 1$, and variance, $\sigma^2_Z = 4$ (Equation 2.3). For clarity, standard-normal values are multiplied by beta and then added to alpha before exponentiating to obtain the log-normal value (Equation 2.4).

$$\begin{aligned} \alpha_Y &= \ln(\mu_Z) - \beta^2_Y/2 \\ \beta^2_Y &= \ln(1 + \sigma^2_Z/\mu^2_Z) \end{aligned} \tag{2.3}$$

$$z = \exp\left(y\beta + \alpha\right) \tag{2.4}$$

where α_Y and β^2_Y are the mean and variance, respectively, of Gaussian distribution that yields the log-normal distribution with mean and variance parameters, μ_Z and σ^2_Z.

In Chapter 5, the non-Gaussian RFs are generated following the procedure outlined in Section 2.2.3, and are then sampled by the same regularly spaced sample grids used in Chapter 3. The sample data are then normal-score transformed following Equation 1.21 before variogram calculation and modelling. The normal-score transform is important,

as standard-normal data are required to assume MG before estimation. The search and estimation parameters mentioned previously do not change. For reference, Table 2.2 summarises the details discussed in this section.

Table 2.2: Generation of the reference conditional distributions.

Chapter	Type of Ref. Real.	Data Transform	Sample Grids (m)			
			8×8	16×16	32×32	64×64
3	Gaussian	None	✓	✓	✓	-
4	Gaussian	Log-Normal	✓	✓	✓	✓
5	Non-Gaussian	Normal Score	✓	✓	✓	-

2.3 Representivity of the study data

The conclusions drawn from this research require representative sampling of the bivariate space of uncertainty defined by the conditional mean, y_{SK}, and conditional variance, σ^2_{SK}, output from MGK. The analytical derivation of the valid y_{SK}/σ^2_{SK} boundary is outside the scope of this work, but the limit cases are known for the standard-normal ($\mathcal{N}(0,1)$) distribution: 1) a maximum σ^2_{SK}-value of 1.0 is only reached when the y_{SK}-value is exactly 0.0, and 2) when the σ^2_{SK}-value is 0.0 (at the data locations), the y_{SK}-value can assume any Gaussian value.

The process followed to generate the data in Chapter 4 is used an example to illustrate how the data are assessed for representivity. The four, regularly spaced grid patterns detailed in Table 2.2 are divided into quarters using Easting and Northing values of 256 m. One quarter from each grid is selected and combined with the others to create a new sample grid from which the MGK estimate is generated (Figure 2.5).

After MGK estimation, the y_{SK}/σ^2_{SK}-distribution is assessed (Figure 2.6, left). To improve the range of possible valid y_{SK}/σ^2_{SK} pairs to sample, the symmetric property of the Gaussian distribution is invoked. The y_{SK}/σ^2_{SK} pairs are copied and the y_{SK} value is multiplied by negative one to yield a new set of values that are combined with the original pairs (Figure 2.6, centre). The resulting dataset, comprising nearly $525,000$ data points, is

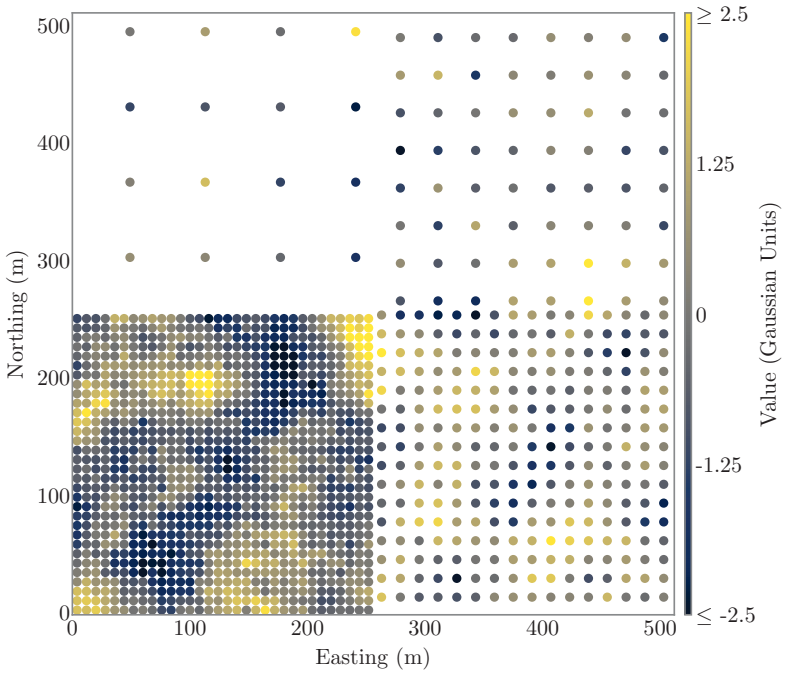


Figure 2.5: Combination of four sample grids used to generate reference MGK estimate in Chapter 4.

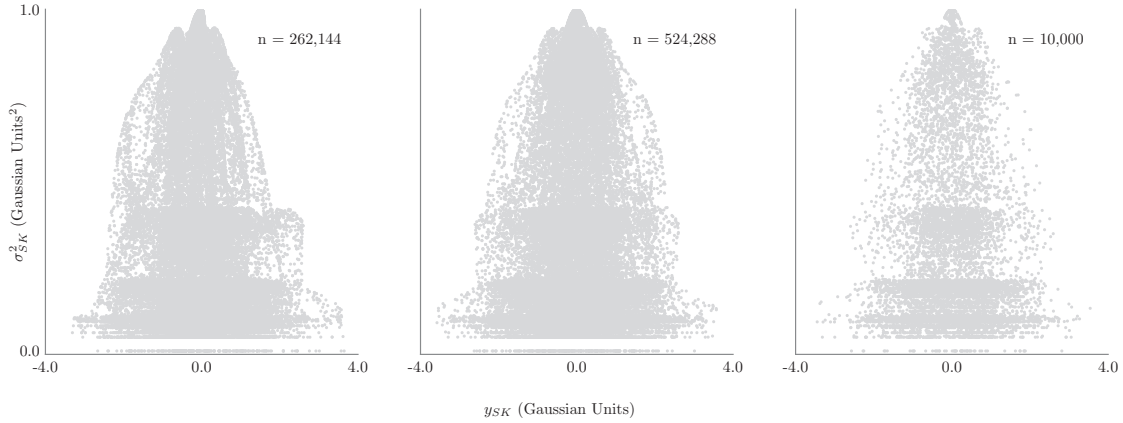


Figure 2.6: Conditional y_{SK}/σ^2_{SK}-pairs after MGK estimation (left); multiplication of y_{SK} by negative one (centre); and data decimation using random sampling (right).

decimated to around $10,000$ data points to reduce computation efforts (Figure 2.6, right). Random sampling is used to decimate the data in order to reserve the relative density of data values across the distribution (Figure 2.7). The data, coloured by kernel density in Figure 2.7 (left), are decimated using random sampling to maintain the correct data density (Figure 2.7, right). Uniform sampling of leads to incorrect results (Figure 2.7, centre). The $10{,}000$ sample pairs drawn are used to analyse the behaviour of the indicator-class means in Chapter 4. Two final steps are completed prior to the data analysis: 1) the data are flagged to identity original estimates ($n = 5{,}024$) versus those where the estimated, conditional-mean value is multiplied by negative one ($n = 4{,}976$), and 2) estimates corresponding to sample locations (i.e., where $\sigma^2_{SK} = 0$) are removed to facilitate computation ($n = 52$). The final dataset comprises $9{,}948$ data values.

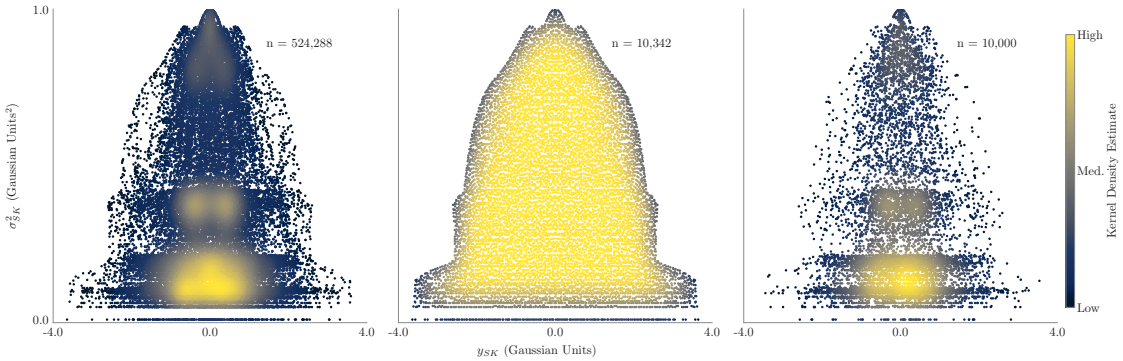


Figure 2.7: Kernel-density colouring of y_{SK}/σ^2_{SK}-estimate pairs before decimation (left); decimation by uniform, random sampling (centre, not used); and decimation using random sampling (right).

In Chapter 3 and Chapter 5, slightly different procedures to yield the final datasets are followed, but the process of multiplying the estimated conditional-mean value by negative one and using several sample-spacing grids is consistent. Data decimation is not undertaken in these chapters. The data used in this research is considered statistically representative and free of bias.

Chapter 3

Indicator Probabilities

Non-parametric construction of the conditional, indicator-probability distribution is an attractive feature of MIK because an explicit mathematical model is not required. Order-relations errors requiring corrections lead to questions about consistency and present a drawback. This chapter utilizes the MG framework to compare the known conditional, indicator-probability distribution output from MGK to the distribution estimated by MIK. The objective is to characterise the differences between the conditional, indicator-probability distributions generated by these estimation techniques. Two comparisons are conducted. The first evaluates the conditional, indicator-probability values, and the second evaluates the indicator-probability-class widths. Results presented in this chapter are generated using seven, unevenly spaced p_k-indicator-probability thresholds that discretize the lower and upper tails more than the centre of the distribution (Equation 3.1). The corresponding Gaussian indicator thresholds, y_k, are presented for reference.

$$\begin{matrix} p_k = \{0.05, 0.10, 0.30, 0.50, 0.70, 0.90, 0.95\} \\ y_k = \{-1.645, -1.282, -0.524, 0, 0.524, 1.282, 1.645\} \end{matrix} \quad , \quad k = 1, \dots, 7 \tag{3.1}$$

Since the Gaussian distribution is symmetric, the discussion focuses on the upper indicators and indicator classes, where $k \geq 4$. Note that log-normal data are not required because the normal-score transform is a quantile-to-quantile transform.

3.1 Comparison of indicator-probability estimates

The correct, conditional, indicator probabilities calculated by MGK, $i(\mathbf{u}_0; y_k|(n))$, are directly compared to those estimated by MIK, $\widehat{i}(\mathbf{u}_0; y_k|(n))$. Checks are completed to ensure all conditional distributions are monotonic and within $[0, 1]$. Figure 3.1 compares the

distribution of indicator probabilities estimated by MGK (grey) and MIK (teal) at each k-threshold. The summary statistics of the two distributions are identical, but deviations are observed near the tails.

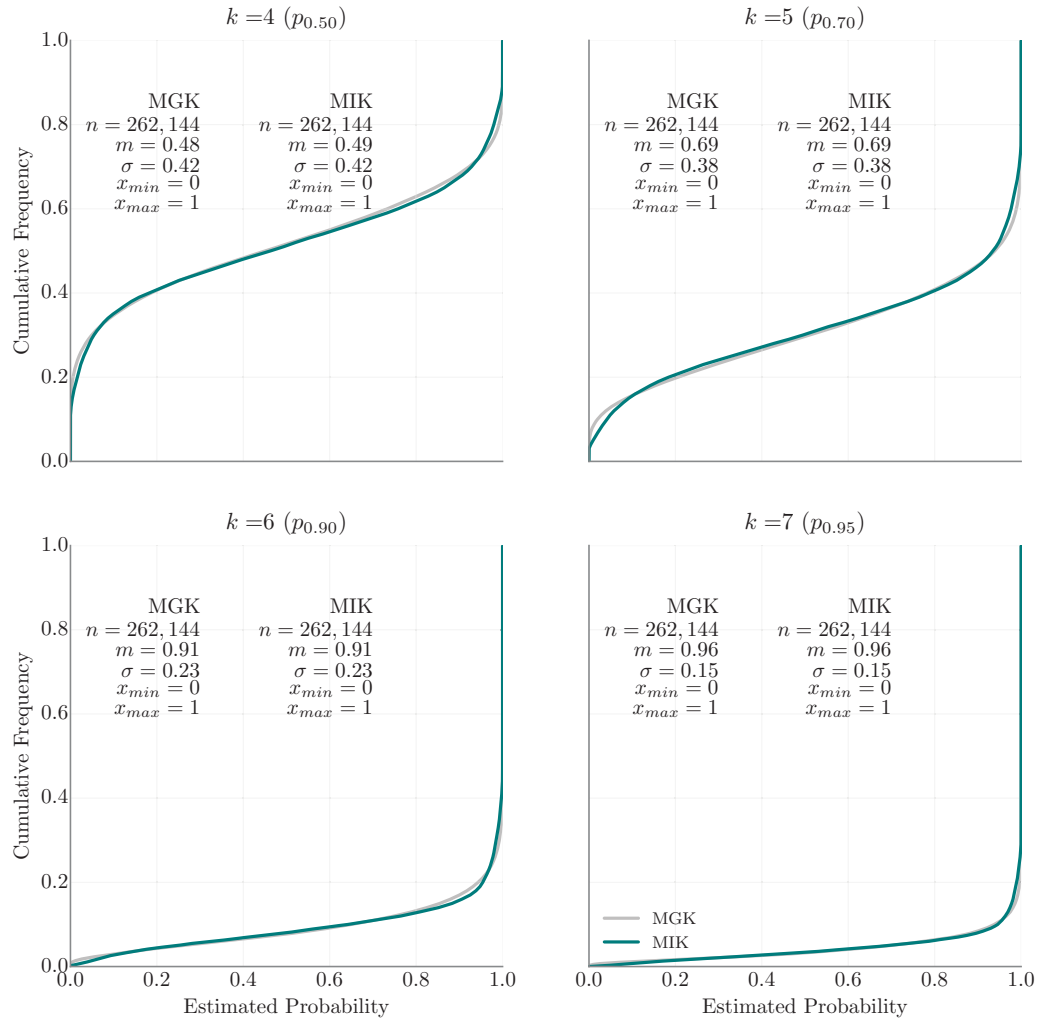


Figure 3.1: CDF of indicator probabilities estimated by MGK (grey) and MIK (teal).

The following figures contrast the differences in the estimated conditional probabilities for the upper indicator thresholds at individual locations. Error calculated at each indicator, denoted as ϵ_{p_k}, is calculated by subtracting the correct, MGK-estimated, indicator probability from the MIK-estimated, indicator probability (Equation 3.2).

$$\epsilon_{p_k} = \widehat{i}(\mathbf{u}_0; y_k|(n))_{MIK} - i(\mathbf{u}_0; y_k|(n))_{MGK}, \quad \in [-1, 1], \ \forall\, k \tag{3.2}$$

Figure 3.2 illustrates the error distribution and confirms the MIK-estimated probabili-

ties are unbiased. The variance of the distributions decreases from the median indicator threshold to the upper indicator threshold (Figure 3.2, upper left to bottom right), which is attributed to the indicator constraints on the data. There are fewer data in the tails of the distribution, so values near the tails are more likely to be zero or one unless the estimate is located in low or high-valued areas. This is also shown in the cross-plots illustrated in Figure 3.3.

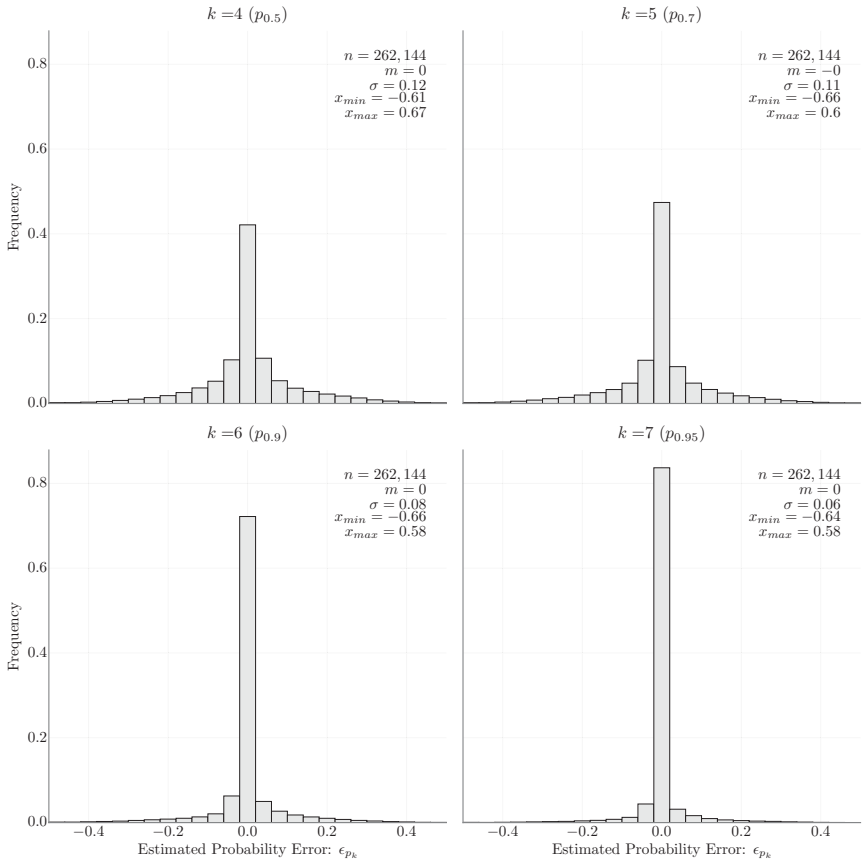


Figure 3.2: Error distribution of MIK-estimated indicator-probability thresholds.

There is strong agreement between the means of the indicator-probability values estimated at each threshold, which is also aligns with the CDFs in Figure 3.1. In addition, the cross-plots illustrate a decrease in the correlation coefficient and an increase in conditional bias towards the upper, indicator-probability threshold (bottom right in Figure 3.3).

In Figure 3.4, the ϵ_{p_k}-error is plotted as a function of the correct, MGK-estimated, indicator probability, $i(\mathbf{u}_0; y_k)_{MGK}$. The MIK estimate appears to overestimate the indicator

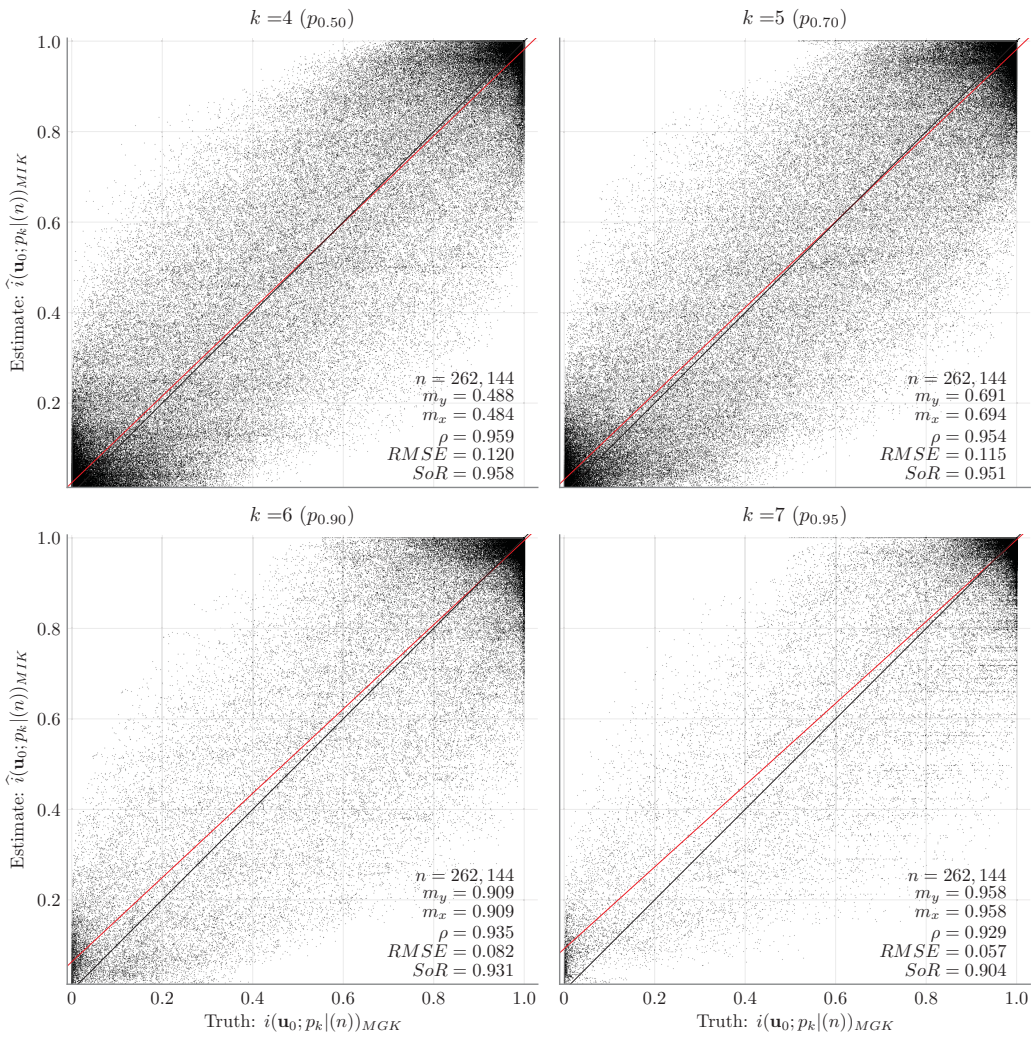


Figure 3.3: Cross-plots comparing the correct MGK-estimated and the MIK-estimated, indicator-probability thresholds.

probability roughly below the $p_{0.25}$-probability threshold, and underestimates the indicator probability above the $p_{0.75}$-probability threshold. The pattern observed in these plots is consistent across the other reference RFs and sample grids, but the inflection points indicating the over/underestimation are not always clear.

The ϵ_{p_k}-error distribution is contoured to assess spatial dependence. The maps in Figure 3.5 and Figure 3.6, illustrate the error contours for the $p_{0.50}$ and the $p_{0.95}$ thresholds, respectively. In Figure 3.5, the positive and negative 0.20-error contours are associated with the transition away from high and low-value regions (refer to inset map). Note that the error contours have a larger spatial extent than the $p_{0.95}$-indicator probability threshold.

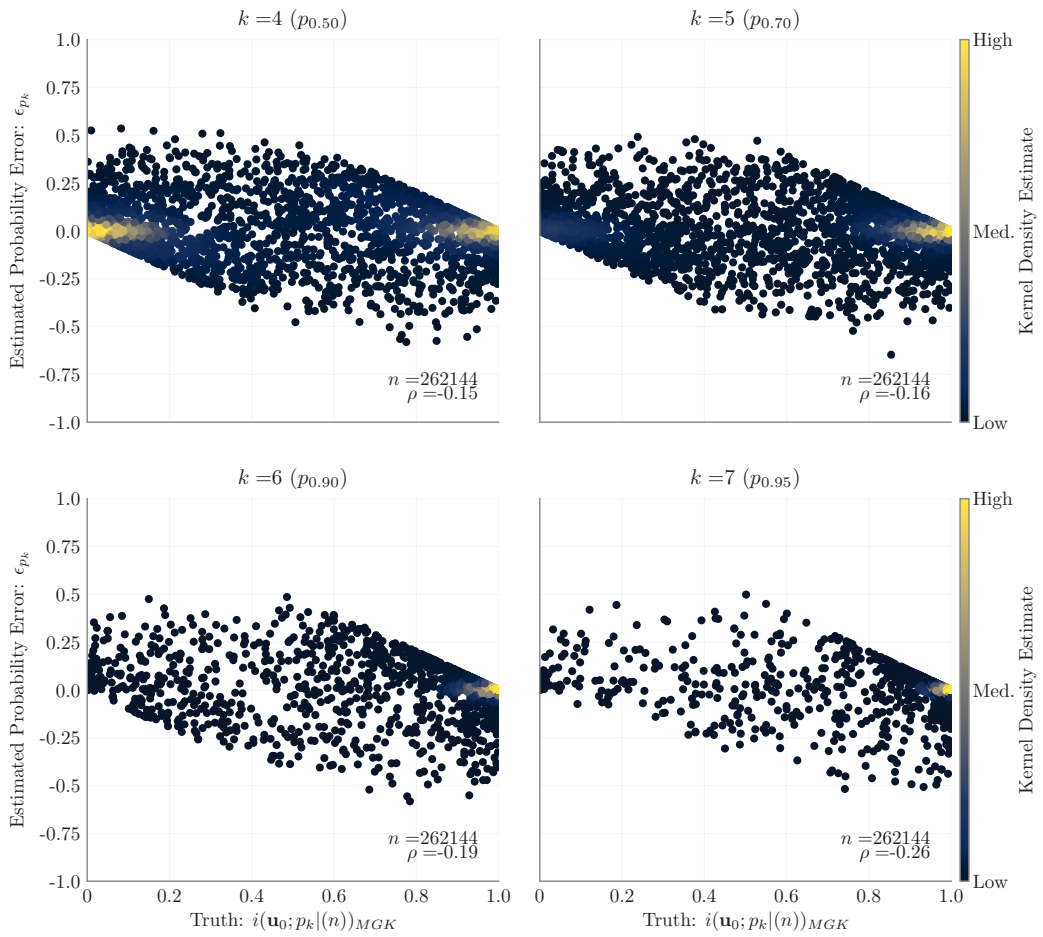


Figure 3.4: Error distribution as a function of the correct, MGK-estimated, indicator-probability thresholds.

In Figure 3.6, the error contours are locally associated with highest values, specifically where the data transition to lower values. It appears that the positive error contours, where $\widehat{i}(\mathbf{u}_0; y_k)_{MIK} > i(\mathbf{u}_0; y_k)_{MGK}$, are spatially associated with higher data values, while the negative contours are associated with lower data values. The spatial association appears weak, so a plot displaying error as a function of distance from the nearest sample would confirm/refute this observation. These observations also apply to the estimates using the larger sample grids, but the spatial extents of the error contours are predictably larger.

The root-mean-squared error (RMSE) and correlation values are averaged across the reference RFs for each indicator-probability threshold to provide representative statistics (Table 3.1). Both the RMSE and correlation decrease as the indicator probability increases.

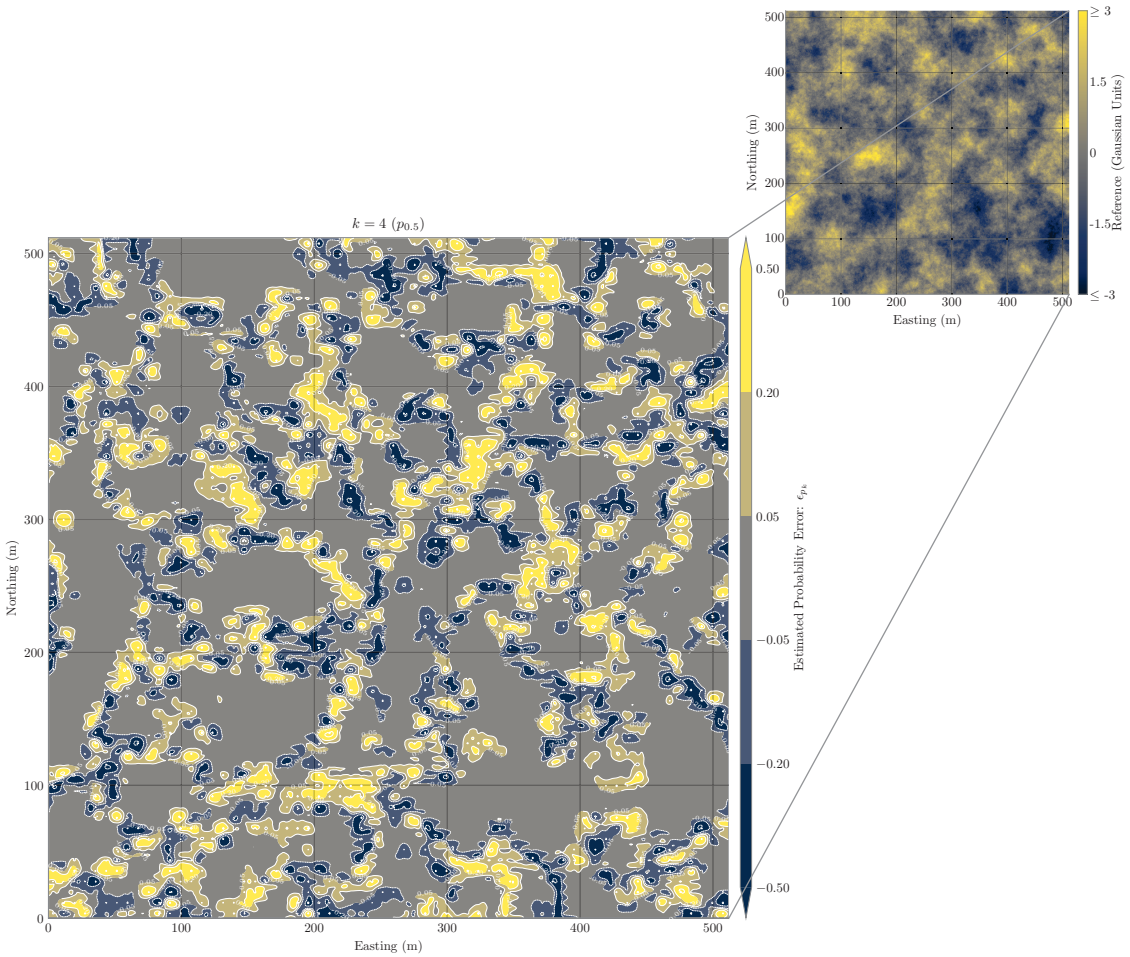


Figure 3.5: Contoured error distribution at the MIK-estimated, $p_{0.50}$-probability threshold. Original data distribution as inset plot for reference.

The apparent decrease in RMSE is misleading due to the constraint on the estimated probability values in the upper indicator-probability thresholds. The correlation coefficient is a better measure of the quality of the indicator estimate. The relative increase in error and the decrease in correlation as the sample spacing increases is expected due to fewer conditioning data.

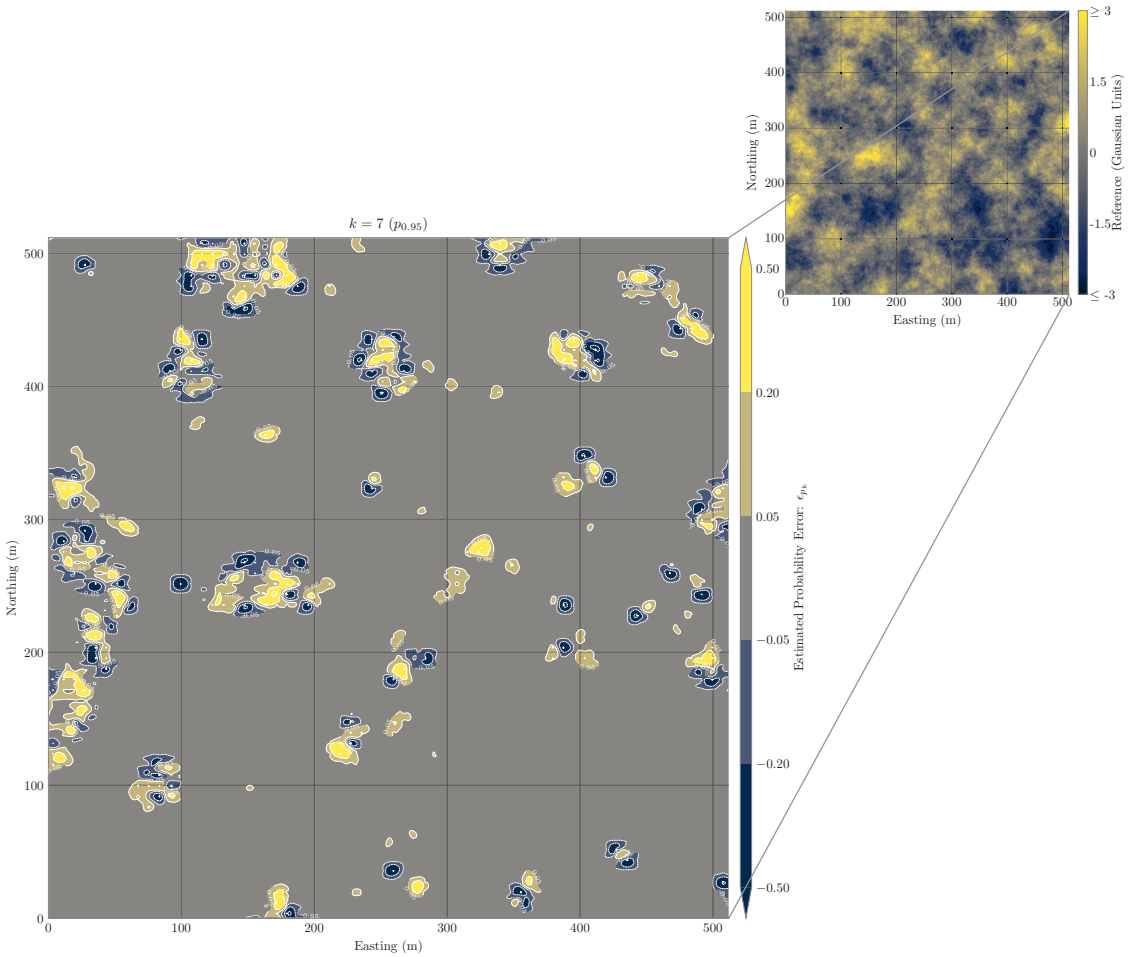


Figure 3.6: Contoured error distribution at the MIK-estimated, $p_{0.95}$-probability threshold. Original data distribution as inset plot for reference.

Table 3.1: Expected RMSE and correlation values between the MIK-estimated indicator probabilities and the known, correct probabilities.

Metric	Sample Grid (m)	MIK Ind. Prob. (p_k)			
		0.50	0.70	0.90	0.95
RMSE	8×8	0.118	0.109	0.078	0.060
	16×16	0.130	0.122	0.089	0.069
	32×32	0.152	0.145	0.102	0.075
ρ	8×8	0.960	0.960	0.946	0.934
	16×16	0.942	0.938	0.910	0.882
	32×32	0.890	0.874	0.814	0.774

3.2 Comparison of indicator-probability-class widths

The indicator-class widths listed in Equation 3.3 are calculated from the p_k-probabilities listed in Equation 3.1:

$$\begin{gathered} p_{cls} = p_{k+1} - p_k = \{0.05, 0.05, 0.20, 0.20, 0.20, 0.20, 0.05, 0.05\} \\ p_0 = 0,\ p_{K+1} = 1 \end{gathered} \tag{3.3}$$

The estimated widths of the conditional, indicator-probability classes, $\widehat{i}(\mathbf{u}_0; y_k, y_{k+1})$, are calculated following Equation 3.4. Calculations are checked to ensure the indicator-class widths sum to unity.

$$\begin{gathered} \widehat{i}(\mathbf{u}_0; y_{cls}) = \widehat{i}(\mathbf{u}_0; y_{k+1}|(n)) - \widehat{i}(\mathbf{u}_0; y_k|(n)) \\ \widehat{i}(\mathbf{u}_0; 0) = 0;\ \widehat{i}(\mathbf{u}_0; y_{K+1}) = 1 \end{gathered} \tag{3.4}$$

where the $\widehat{i}$-terms are the estimated probabilities of adjacent indicator-probability thresholds. Calculation of the indicator-probability-class error, $\epsilon_{p_{cls}}$, follows Equation 3.5.

$$\epsilon_{p_{cls}} = \widehat{i}(\mathbf{u}_0; y_{cls})_{MIK} - i(\mathbf{u}_0; y_{cls})_{MGK}, \quad \in [-1, 1],\ cls = 1, \ldots, K+1 \tag{3.5}$$

Figure 3.7 compares the distribution of indicator probabilities estimated by MGK (grey) and MIK (teal) for the upper classes, $cls \geq 5$. In contrast to the probability estimates in Figure 3.1, the distributions show more variation. Note that the comparison in the upper class ($cls = 8$) appears quite close due to scaling of the ordinate axis. Further investigation (not presented) shows minor local deviations, but they are much less in comparison to the other indicator classes.

The plots in Figure 3.8, Figure 3.9, and Figure 3.10 illustrate similar features observed in the corresponding figures in the previous section, Figure 3.2, Figure 3.3 and Figure 3.4, respectively. The estimated probability-class widths are correct on average, but the variances and RMSE values are higher and the correlation values are lower for all classes except the upper tail ($cls = 8$).

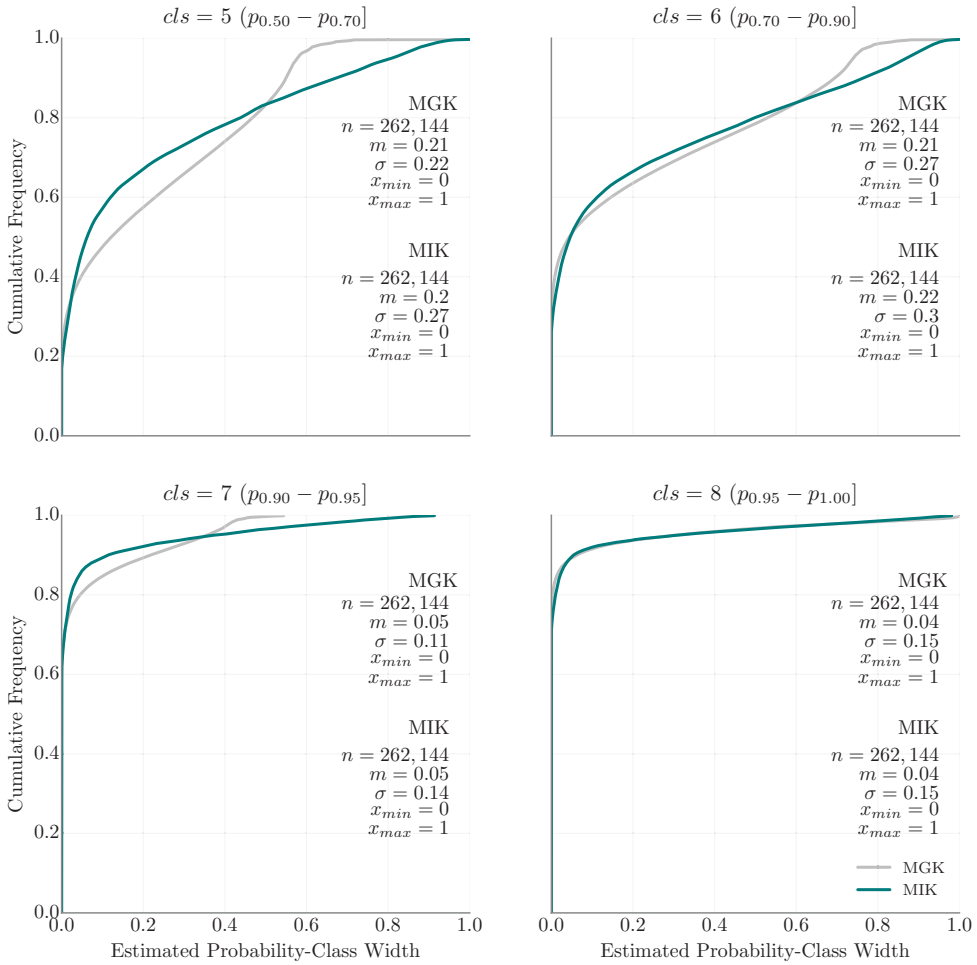


Figure 3.7: CDF of indicator-probability-class widths estimated by MGK (grey) and MIK (teal).

Maps of the error contours of indicator-class-widths for the indicator classes, $cls = 4$ and $cls = 8$, are presented in Figure 3.11 and Figure 3.12, respectively. Both figures show similar features observed in the corresponding maps in Figure 3.5 and Figure 3.6, but the magnitude and spatial extent of the error in the map of the indicator-class widths is much greater. In the upper class, the spatial extents of the contours are unchanged. The magnitudes are inverted due to the symmetry of the Gaussian distribution.

RMSE and correlation values are averaged across the reference RFs for each indicator-probability threshold (Table 3.2). Except for the upper class, the average error increases and the correlation decreases, which is consistent with the observations made in this chapter.

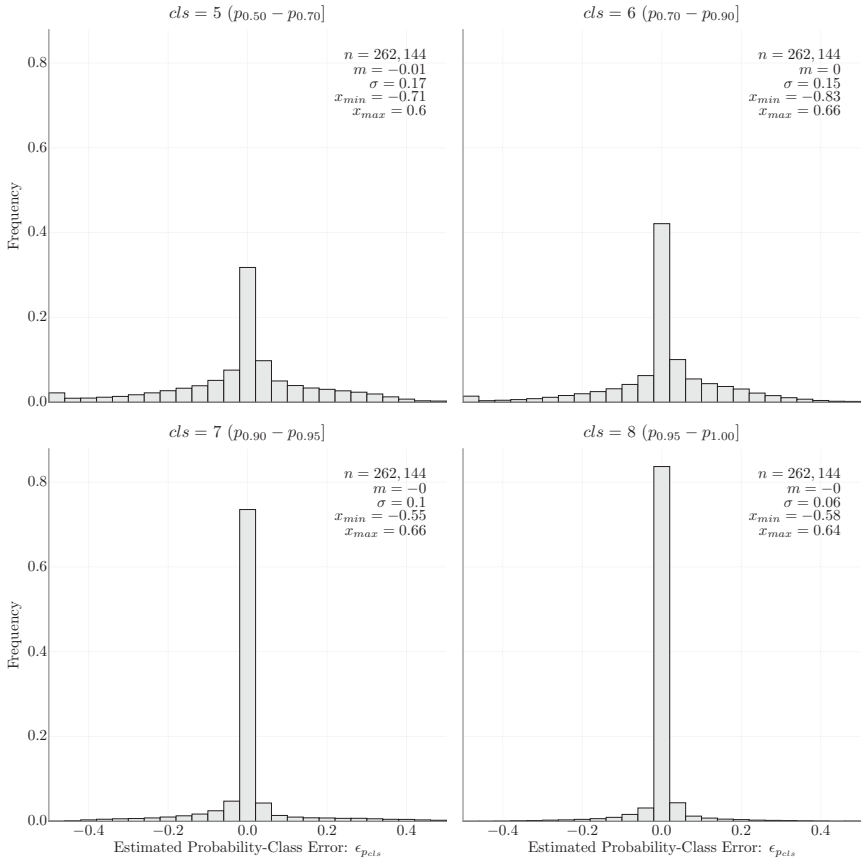


Figure 3.8: Distribution of conditional-indicator probability error between MGK and MIK for median and upper indicator classes.

Table 3.2: Expected RMSE and correlation values between the MIK-estimated indicator-class widths and the known, correct class widths.

Metric	Sample Grid (m)	MIK Ind. Class Width			
		0.20	0.20	0.05	0.05
RMSE	8×8	0.169	0.140	0.096	0.060
	16×16	0.165	0.145	0.090	0.069
	32×32	0.186	0.164	0.090	0.075
ρ	8×8	0.777	0.874	0.750	0.934
	16×16	0.675	0.807	0.667	0.882
	32×32	0.486	0.686	0.548	0.774

3.3 Summary

Non-parametric estimation of the conditional, indicator-probability distribution output by MIK is unbiased in comparison to the correct probability distribution output by MGK. There can be significant local variation in the estimated conditional probability, but positive and negative errors do not show consistent spatial association with high or low values.

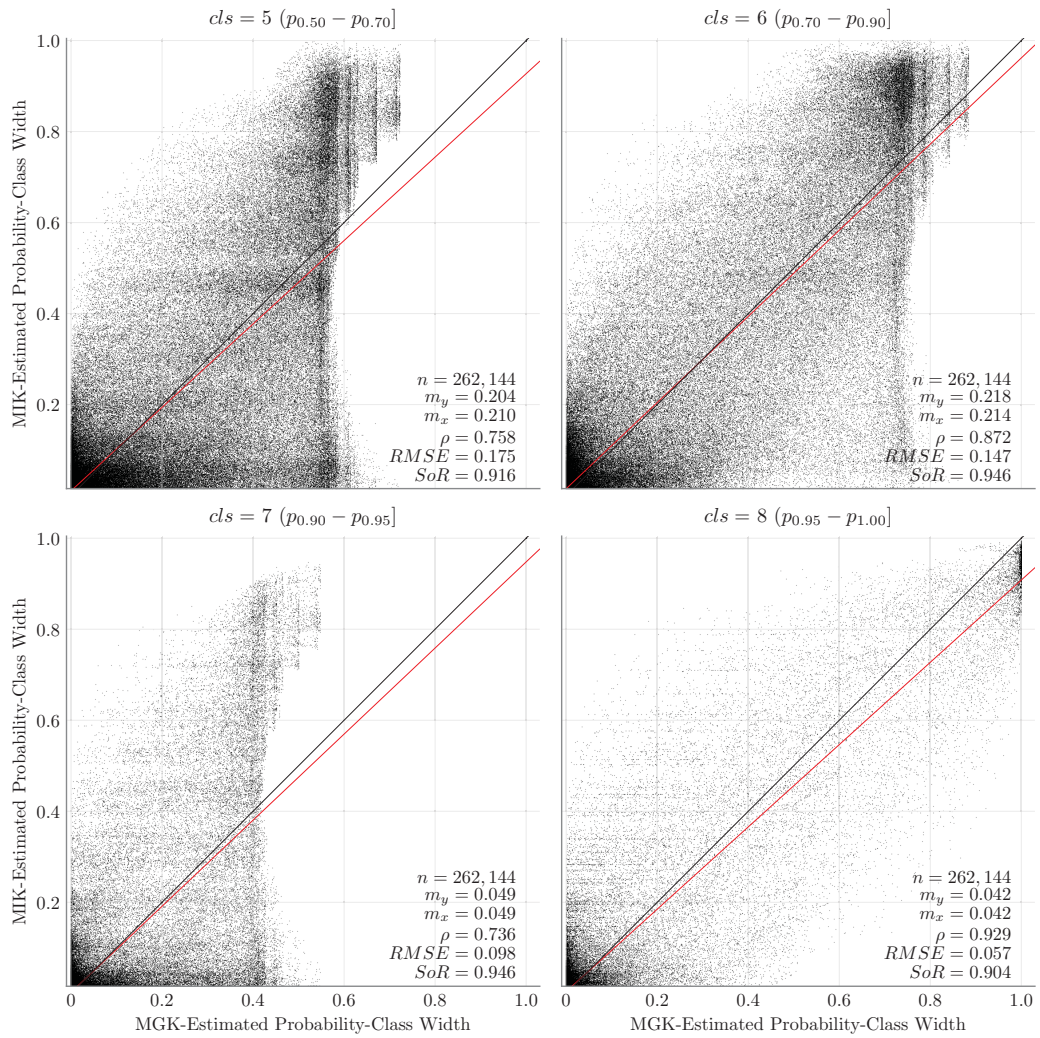


Figure 3.9: Cross-plots comparing conditional-indicator probability between MGK and MIK for median and upper indicator classes.

There is a marked increase in error when evaluating the indicator class widths in comparison to the indicator threshold values. This suggests that the probability model output by MIK is inconsistent. Order-relations corrections are made to ensure the MIK probability distribution is licit, but their impact on its consistency could be investigated further.

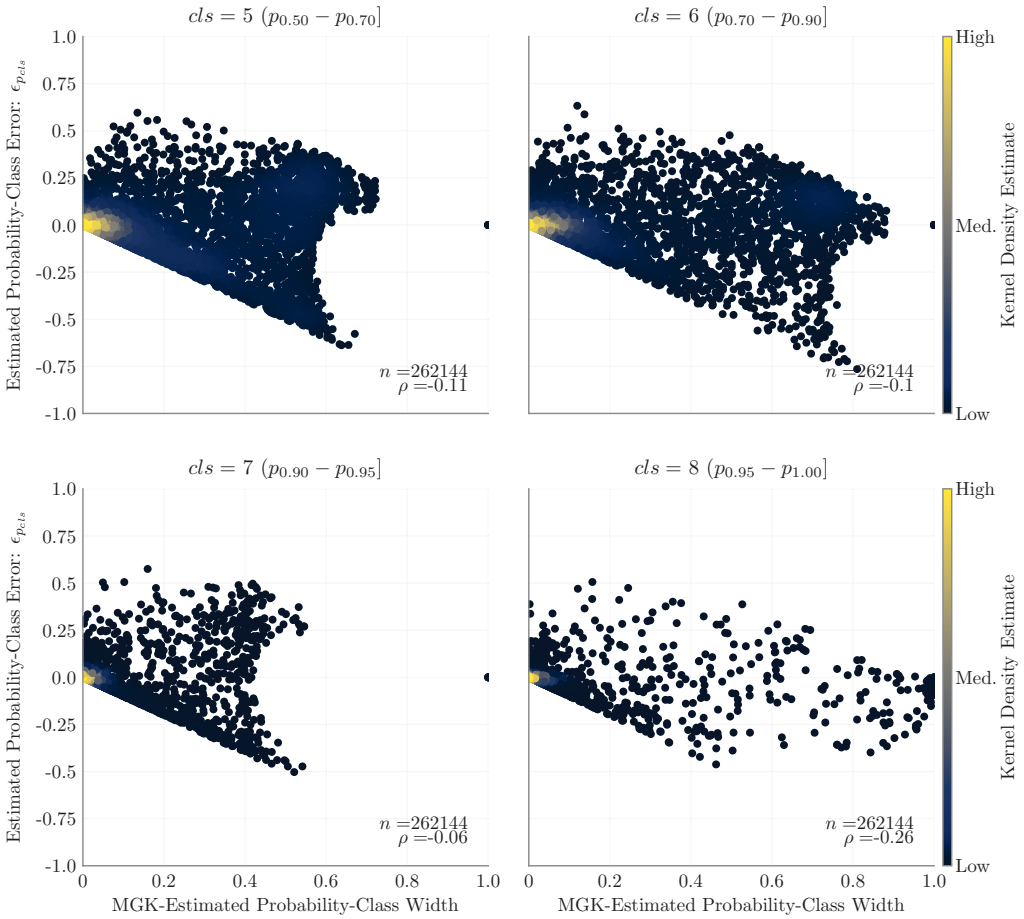


Figure 3.10: Error distribution as a function of the correct, MGK-estimated, indicator-probability-class widths.

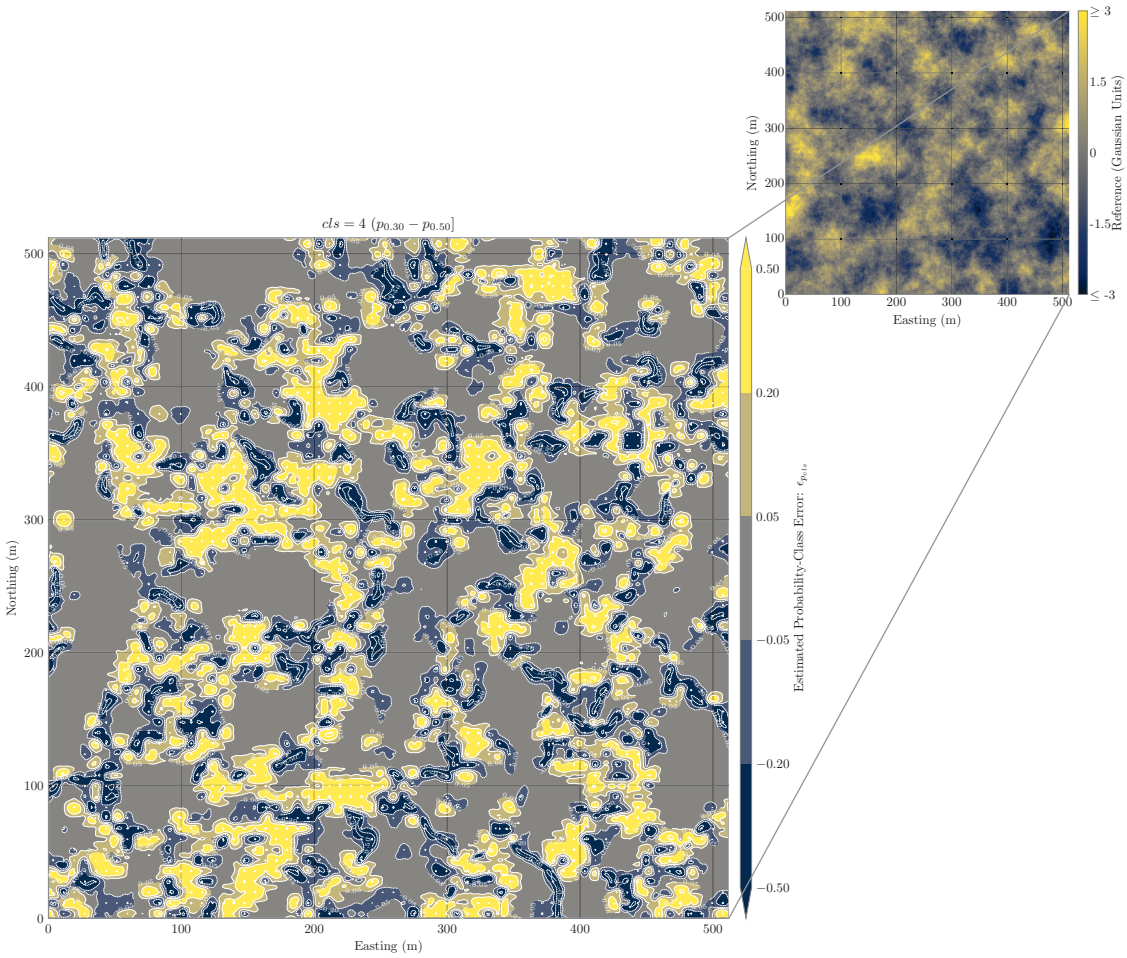


Figure 3.11: Contour map of the indicator-class-width error, $\epsilon_{p_{cls}}$, for the median class ($cls = 4$). Note the increased spatial extent and magnitude of the contours in comparison to Figure 3.5. Original data distribution as inset plot for reference.

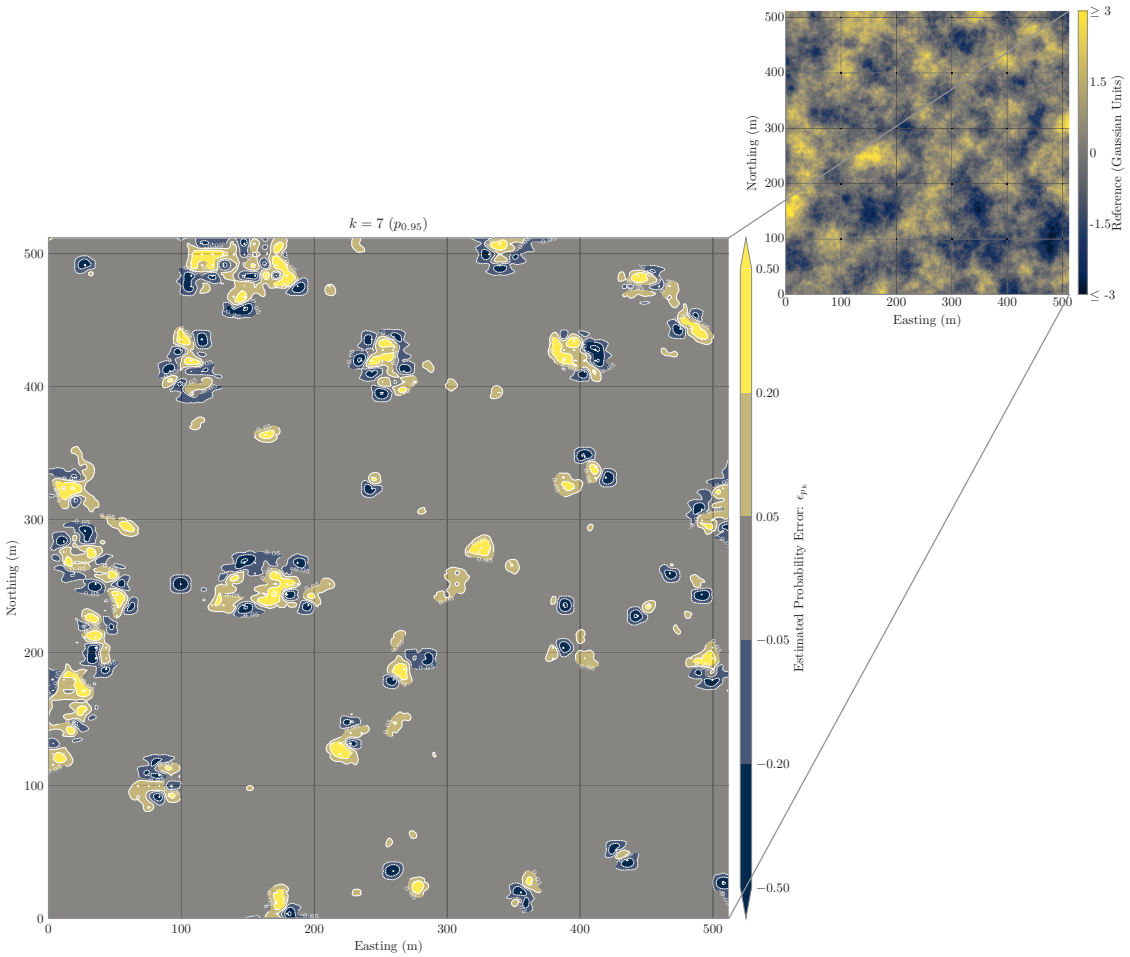


Figure 3.12: Contour map of the indicator-class-width error, $\epsilon_{p_{cls}}$, for the upper class ($cls = 8$). The spatial extent of the contours in comparison to Figure 3.6 has not changed. Original data distribution as inset plot for reference.

Chapter 4

Indicator-Class Means

The MIK estimator assumes stationary indicator-class-mean values, $\widehat{m}_k$. This chapter discusses 1) definition and calculation of $\widehat{m}_k$, 2) $\widehat{m}_k$ behaviour, 3) the factors upon which $\widehat{m}_k$ depends, and 4) how the use of constant $\widehat{m}_k$ impacts the estimate in comparison to the correct indicator-class mean.

This study replicates an MIK estimate by utilizing the properties of the MG distribution to calculate the exact conditional-probability and conditional, indicator-class-mean values, $\widehat{m}_{k_{cond}}$, from the sampled data (Section 2.2.4). This isolates the class-mean component of the MIK-estimator (Equation 1.29). Note that true MIK estimation is not undertaken in this chapter. The practical range of indicator thresholds used in MIK estimation of mining datasets ranges between five and fifteen ($\mathrm{K} = 5, \mathrm{K} = 15$) (Carvalho & Deutsch, 2017). The results mostly reflect the use of $(\mathrm{K} = 10)$-indicator thresholds, but examples using the lower and upper indicator-threshold ranges are also included to illustrate trends where appropriate.

Similar to the presentation of results in Chapter 3, calculations corresponding only to the upper classes of the Gaussian dataset are presented because of the symmetry of the Gaussian distribution. Research results corresponding to the log-normal data are presented only for the upper tail because this class contains a significant proportion of the economic value found in real-world mining datasets.

4.1 Class-mean definition and calculation

Gaussian values are calculated through the inverse of the Gaussian distribution using ten-thousand, evenly spaced p-probability values, where $p_l = \frac{l}{\mathrm{P}+1}$, $l = 1, \ldots, \mathrm{P}$, and $y_l =$

$\mathrm{G}^{-1}(l)$. The corresponding log-normal quantiles, z_l, are calculated following the log-normal transform presented in Equation 2.4.

The $\widehat{m}_k$-values are calculated by averaging the quantiles between the indicator-threshold limits, $(y_{k-1}, y_k]$ (Equation 4.1). The lower-tail class mean is calculated between the interval $(-\infty, y_{k=1}]$ and the upper tail over the interval $(y_{k=K}, \infty)$. The log-normal lower tail class mean is calculated between $(0, z_{k=1})$.

$$\widehat{m}_k = \frac{1}{L} \sum_{y_k}^{y_{k+1}} y_l, \quad y_l \in (-\infty, \infty);\ l = 1, \ldots, L;\ k = 1, \ldots, K \tag{4.1}$$

where L is the number of p_l-quantiles in each indicator class.

Note that the discretization of the Gaussian and log-normal distributions constrains the minimum and maximum quantile values to $[-3.719 \leq y_l \leq 3.719]$ and $[0.004 \leq z_l \leq 50.067]$, respectively. As a check on the effectiveness of discretization in the distribution tails, the upper-tail, class mean is calculated analytically and found to match the value from the discretized distribution to the fourth decimal place. Hereafter, the class mean notation adopted in this thesis is modified to distinguish between three indicator-class-mean values: 1) global (stationary), $\widehat{m}_{k_{glob}}$, 2) conditional, $\widehat{m}_{k_{cond}}$, and 3) expected, $E\{\widehat{m}_{k_{cond}}\}$. These are defined in the following sections.

4.1.1 Global indicator-class mean

The global, indicator-class-mean values are derived from the declustered, sample histogram. They are calculated directly from the known Gaussian and log-normal distributions. Table 4.1 summarizes the $\widehat{m}_{k_{glob}}$ values for Gaussian and log-normal distributions with ten indicator thresholds. The indicator-class intervals and relative locations of the $\widehat{m}_{k_{glob}}$ values are illustrated in Figure 4.1. The inset Gaussian probability density function (PDF), defined by parameters $\alpha_Y = -0.804$ and $\beta_Y^2 = 1.609$, is provided to illustrate the link to the corresponding log-normal distribution with mean, $\mu_Z = 1.0$, and variance, $\sigma_Z^2 = 4.0$. Note that the Gaussian thresholds and global, indicator-class means presented in Table 4.1

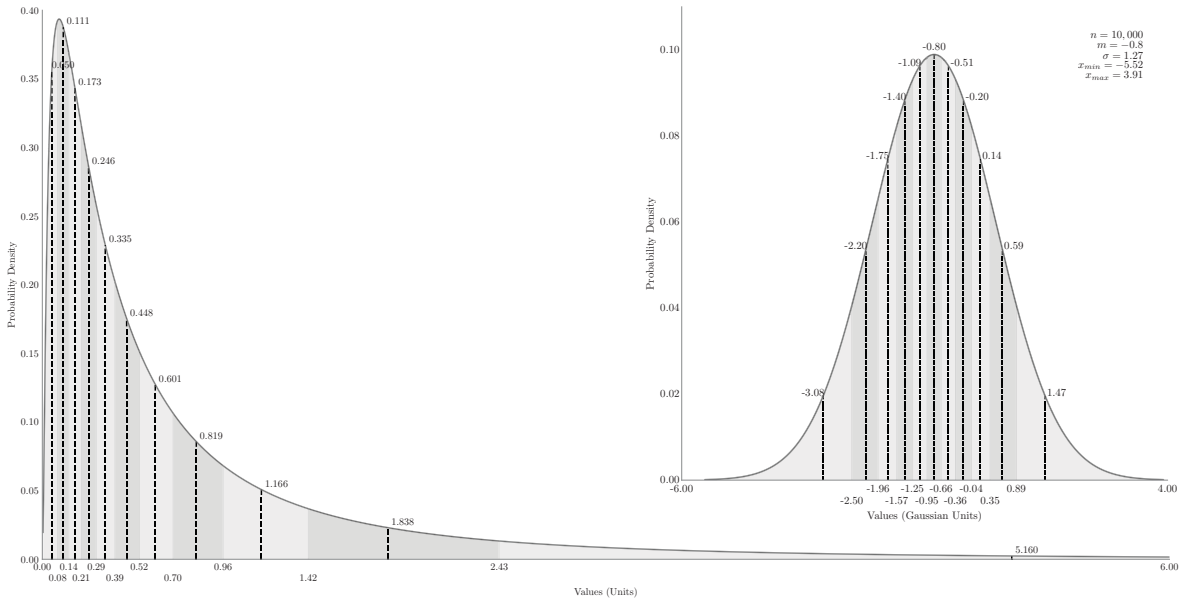


Figure 4.1: Indicator classes (shaded) and $\widehat{m}_{k_{glob}}$ (vertical dashed lines) for $(\mathrm{K} = 10)$-indicator thresholds. The log-normal plot is truncated at 6.0 units for display purposes, but is discretized to an upper-quantile value of 50.067 units.

correspond to the standard-normal Gaussian distribution analyzed in this study, not to the inset plot in Figure 4.1.

Table 4.1: Indicator thresholds $(\mathrm{K} = 10)$ and global, indicator-class means for Gaussian ($\mu_Y = 0$, $\sigma^2_Y = 1$) and log-normal ($\mu_Z = 1$, $\sigma^2_Z = 4$) distributions.

Class Number	Probability	Thresh. Limits (Gauss)	$\widehat{m}_{k_{glob}}$ (Gauss)	Thresh. Limits (Log-Normal)	$\widehat{m}_{k_{glob}}$ (Log-Normal)
1 (Lower Tail)	0.001 - 0.091	$y < -1.335$	-1.798	$z < 0.082$	0.050
2	0.091 - 0.182	$-1.335 < y \leq -0.908$	-1.105	$0.082 < z \leq 0.141$	0.111
3	0.182 - 0.273	$-0.908 < y \leq -0.605$	-0.751	$0.141 < z \leq 0.208$	0.173
4	0.273 - 0.364	$-0.605 < y \leq -0.349$	-0.474	$0.208 < z \leq 0.287$	0.246
5	0.364 - 0.454	$-0.349 < y \leq -0.114$	-0.230	$0.287 < z \leq 0.387$	0.335
6	0.454 - 0.545	$-0.114 < y \leq 0.114$	0.000	$0.387 < z \leq 0.517$	0.448
7	0.545 - 0.636	$0.114 < y \leq 0.349$	0.230	$0.517 < z \leq 0.696$	0.601
8	0.636 - 0.727	$0.349 < y \leq 0.605$	0.474	$0.696 < z \leq 0.963$	0.819
9	0.727 - 0.818	$0.605 < y \leq 0.908$	0.751	$0.963 < z \leq 1.416$	1.166
10	0.818 - 0.909	$0.908 < y \leq 1.335$	1.105	$1.416 < z \leq 2.433$	1.838
11 (Upper Tail)	0.909 - 0.999	$y > 1.335$	1.798	$z > 2.433$	5.160

4.1.2 Conditional indicator-class mean

The $\widehat{y}_{SK}$ and $\widehat{\sigma}_{SK}$ values output from kriging parameterize the conditional distribution at the unsampled location under the MG framework. The spatial arrangement of the conditioning data and their covariance relationships determines the kriged conditional mean

and conditional variance values. This leads to variability in the conditional, indicator-class-mean values.

Figure 4.2 illustrates the calculation of $\widehat{m}_{k_{cond}}$ in the upper tail of the standard normal distribution for three unsampled locations. The global, Gaussian PDF, with $\mu_Y = 0$, $\sigma_Y^2 = 1$, is plotted in black in each figure. The upper class is the dark-grey, shaded region under the curve, with a class limit of 1.335 Gaussian units. There are $(\mathrm{K} = 10)$-thresholds. The conditional distributions are plotted in teal, with the teal, shaded region corresponding to the portion of the conditional distribution situated within the upper class. The conditional, indicator-class mean is denoted by the dashed, teal line. The position of $\widehat{m}_{k_{cond}}$ relative to $\widehat{m}_{k_{glob}}$ depends on the neighbouring conditioning data. The distribution in the left-hand plot is conditioned in an area of dominantly low values, thus only the uppermost part of the conditional distribution overlaps the upper class, and $\widehat{m}_{k_{cond}} < \widehat{m}_{k_{glob}}$. In the middle plot, there is a broad range of conditioning data, so roughly the upper half of the conditional distribution overlaps the upper class, and $\widehat{m}_{k_{cond}} \approx \widehat{m}_{k_{glob}}$. Generally high values condition the distribution in the right plot. The conditional variance is low, thus most of the conditional distribution overlaps the upper class, and $\widehat{m}_{k_{cond}} > \widehat{m}_{k_{glob}}$.

The conditional, indicator-class mean is the average of the quantiles in each class. It can theoretically be calculated for every class because the tails of the Gaussian conditional distributions extend to $(-\infty, \infty)$. In practice this is not done because probability values become exceedingly small. Any indicator classes located sufficiently far into the tails of the conditional distribution may not contain discretized quantiles and are undefined. This explains the varying number of conditional means defining the histograms in Figure 4.3 and Figure 4.4.

The expected, indicator-class mean, $E\{\widehat{m}_{k_{cond}}\}$, is the mean of $\widehat{m}_{k_{cond}}$ across the domain at all unsampled locations. The $E\{\widehat{m}_{k_{cond}}\}$ value is used to assess bias between the stationary $\widehat{m}_{k_{glob}}$ and varying $\widehat{m}_{k_{cond}}$ values.

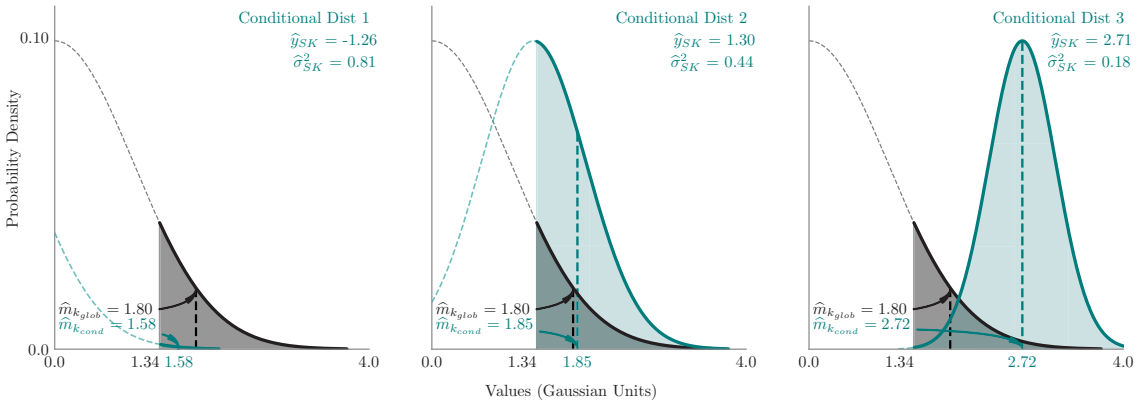


Figure 4.2: Calculation of $\widehat{m}_{k_{cond}}$ from three example conditional distributions (teal) within the upper class (dark-grey shaded region). The $\widehat{m}_{k_{glob}}$ is denoted by the vertical, black dashed lines. The portion of the conditional distribution overlapping the upper class is denoted by a thicker line. The quantiles defining this portion of the conditional distribution are used to calculate $\widehat{m}_{k_{cond}}$ (vertical, teal, dashed lines). The variability of $\widehat{m}_{k_{cond}}$ is noteworthy.

4.2 Statistical variation of class means

The distribution of conditional, indicator-class means for all unsampled locations is presented for the scenario with $(\mathrm{K} = 10)$-thresholds in Figure 4.3. The plotting limits of each histogram correspond to the lower and upper values of each class, aside from the lower and upper limits of the first and last classes. The varying number of data in each class histogram indicates how many conditional distributions have at least one quantile in the class limits. For example, Class 6 (upper-left histogram in Figure 4.3) contains at least one quantile from $8,948$ out of the $9,948$ calculated conditional distributions. There are $1,000$ conditional distributions without any defined quantiles in this class. This means that the tails of the distributions conditioned by either very high or very low data values do not reach into the middle class.

The increasing histogram distribution asymmetry and change in the conditional variance as the class number increases is noteworthy. This causes the global, indicator-class mean (vertical, teal, dashed line in each plot) to diverge from the expected, indicator-class mean (vertical, black, dashed line in each plot). Furthermore, the expected, indicator-class mean is consistently lower than the global value in the upper classes of the Gaussian distribution. The opposite is true in the lower classes (not shown). The divergence between

the expected and global, indicator-class means is consistent across all of the threshold scenarios considered, but the magnitude of the difference in the upper classes decreases with an increasing number of thresholds considered.

The percent difference between the global and expected, indicator-class means (Equation 4.2), is presented in Figure 4.5 for the threshold scenarios: $\{\mathrm{K} = 5, 10, 15\}$. The x-axis denotes the class number (recall there are $\mathrm{K} + 1$ classes).

$$\epsilon_{\widehat{m}_k} = \frac{\overline{m}_{k_{glob}} - E\{\overline{m}_{k_{cond}}\}}{E\{\overline{m}_{k_{cond}}\}} \tag{4.2}$$

The Gaussian data in Figure 4.5a show an diagnostic symmetric profile about the median class in each scenario, with the difference increasing towards the tail classes. The difference between the global and expected, indicator-class mean in the tails of the distributions is significant, nearly five times more than in the penultimate-tail classes. Increasing the number of thresholds improves the discretization of the distribution and results in smaller differences.

Similar trends are seen in Figure 4.5b, though the differences in the upper tail of the log-normal data are significantly greater. The negative differences between the global and expected, indicator-class means observed in the lower tails in Figure 4.5b are interesting, but they are of little practical consequence in the context of economic mineralization.

4.3 Estimates with conditional and global indicator-class means

MIK-style estimates are undertaken to quantify the impact of using the global, indicator-class means in comparison to using the correct, conditional, indicator-class means. The conditional, indicator probabilities are calculated directly, so any observed variation is due to the class-mean value used. A direct comparison of the plots in original units (left) and in log-units (right) in Figure 4.6 highlights the significant differences in the upper tail when the conditional, indicator-class mean is used.

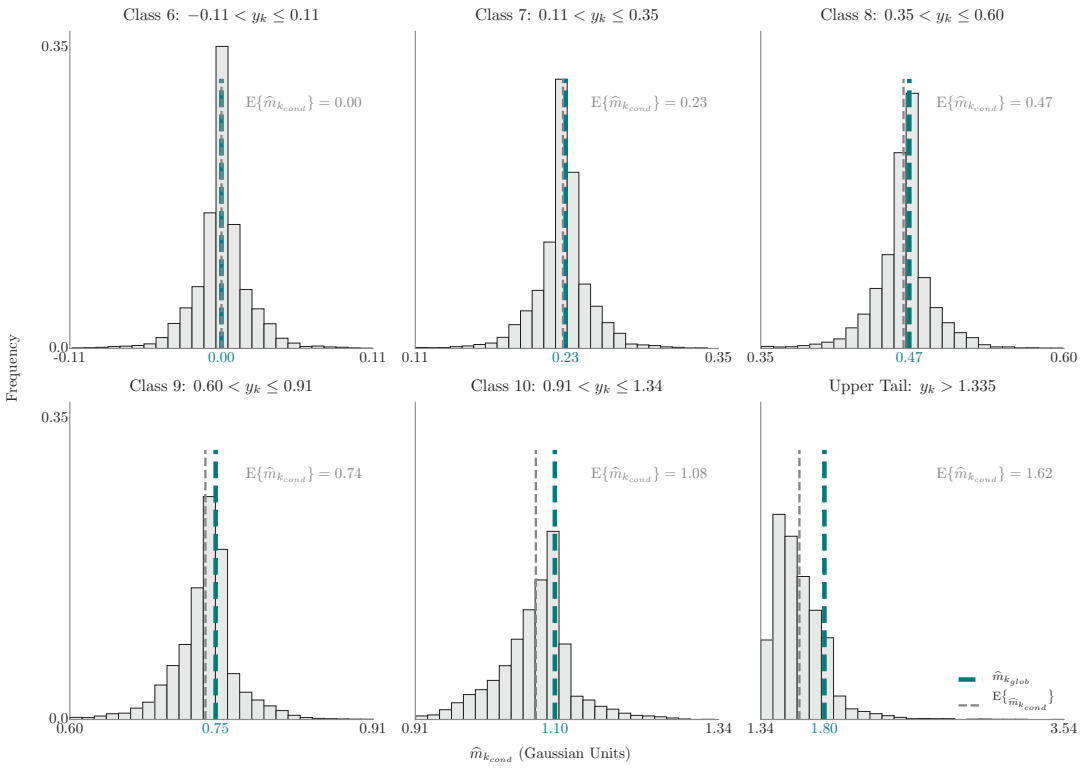


Figure 4.3: Distribution of the Gaussian $\widehat{m}_{k_{cond}}$ (K = 10). Only the upper classes are shown. The vertical, grey, dashed lines represent the mean of the histogram ($E\{\widehat{m}_{k_{cond}}\}$). The vertical, teal, dashed lines represent the global (stationary), indicator-class mean, $\widehat{m}_{k_{glob}}$.

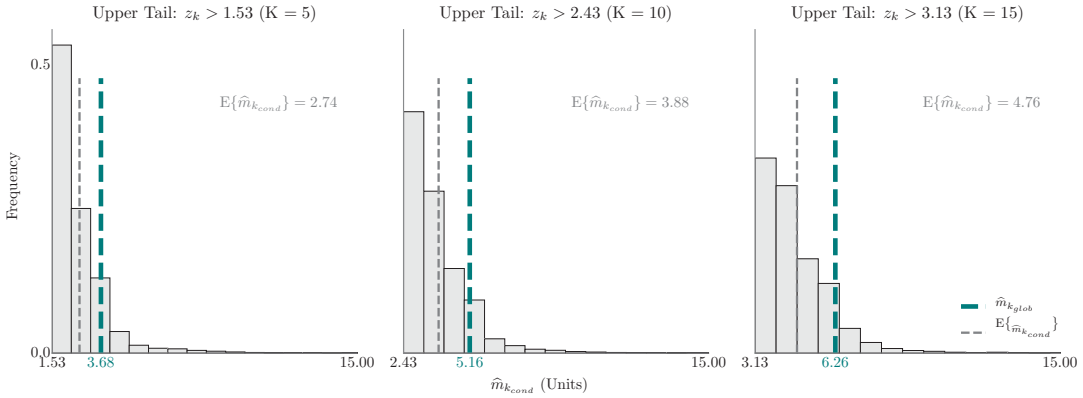
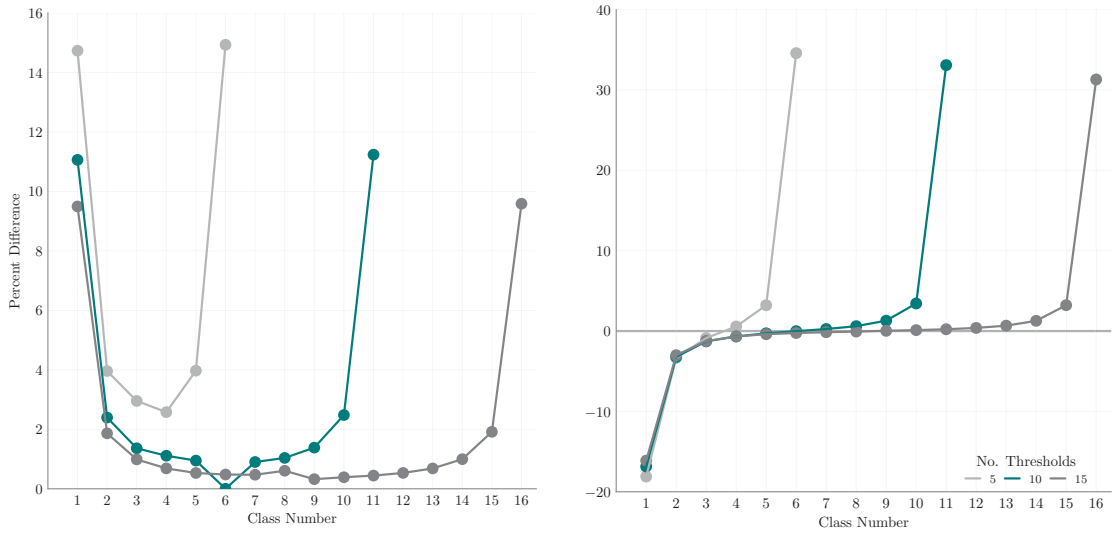


Figure 4.4: Upper-tail, log-normal $\widehat{m}_{k_{cond}}$-distribution (K = 5, 10, 15). The vertical, grey, dashed lines represent the mean of the histogram ($E\{\widehat{m}_{k_{cond}}\}$). The vertical, teal, dashed lines represent the global (stationary), indicator-class mean, $\widehat{m}_{k_{glob}}$. The x-axis is truncated to 15.00 units.

Figure 4.7 maps the difference between estimates using the global and conditional, indicator-class means (defined in Equation 4.2). The log-normal estimates (K = 10) are presented as a set of error contours expressed in percent difference to highlight the relative magnitudes of the positive and negative errors. The conditioning data are overlaid,



(a) Gaussian data.

(b) Log-Normal data.

Figure 4.5: Percent difference between $\widehat{m}_{k_{glob}}$ and $E\{\widehat{m}_{k_{cond}}\}$ for K $= 5$ (light grey), K $= 10$ (teal), K $= 15$ (dark grey). Each threshold scenario has K$+1$ classes, which correspond to the class number values on the x-axis.

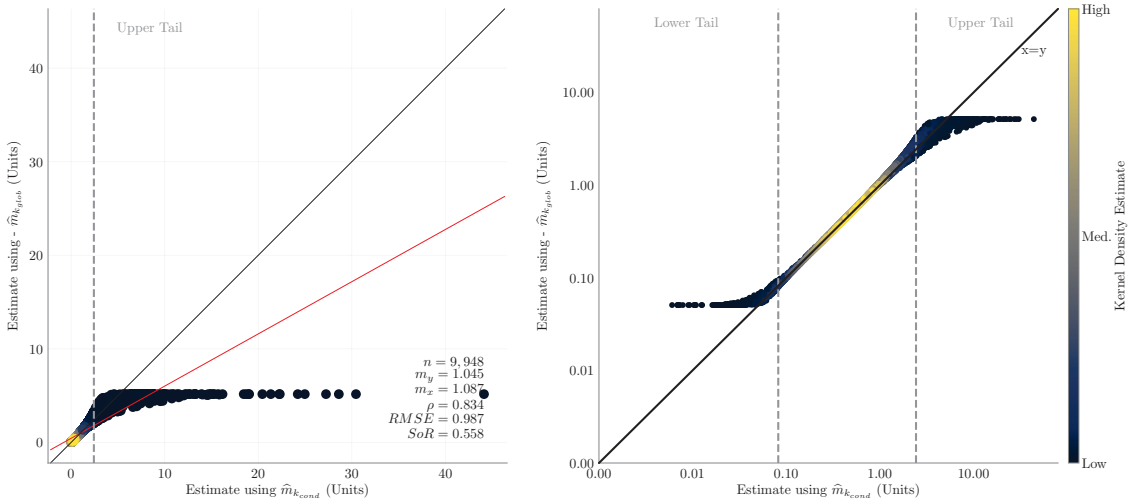


Figure 4.6: Cross-plots comparing estimates using the correct, $\widehat{m}_{k_{cond}}$ and stationary $\widehat{m}_{k_{glob}}$ (log data, K $= 10$). Log-scale plot (right) highlights the estimation error in both the lower and upper tails.

with the lower and upper two indicator classes coloured to highlight their locations relative to the error contours. All conditioning data in the middle classes are coloured grey. The greatest positive and negative errors are located around conditioning data in the lower and upper tails.

In the middle of the plot (Figure 4.7), near coordinates (250 mE, 250 mN), is a region

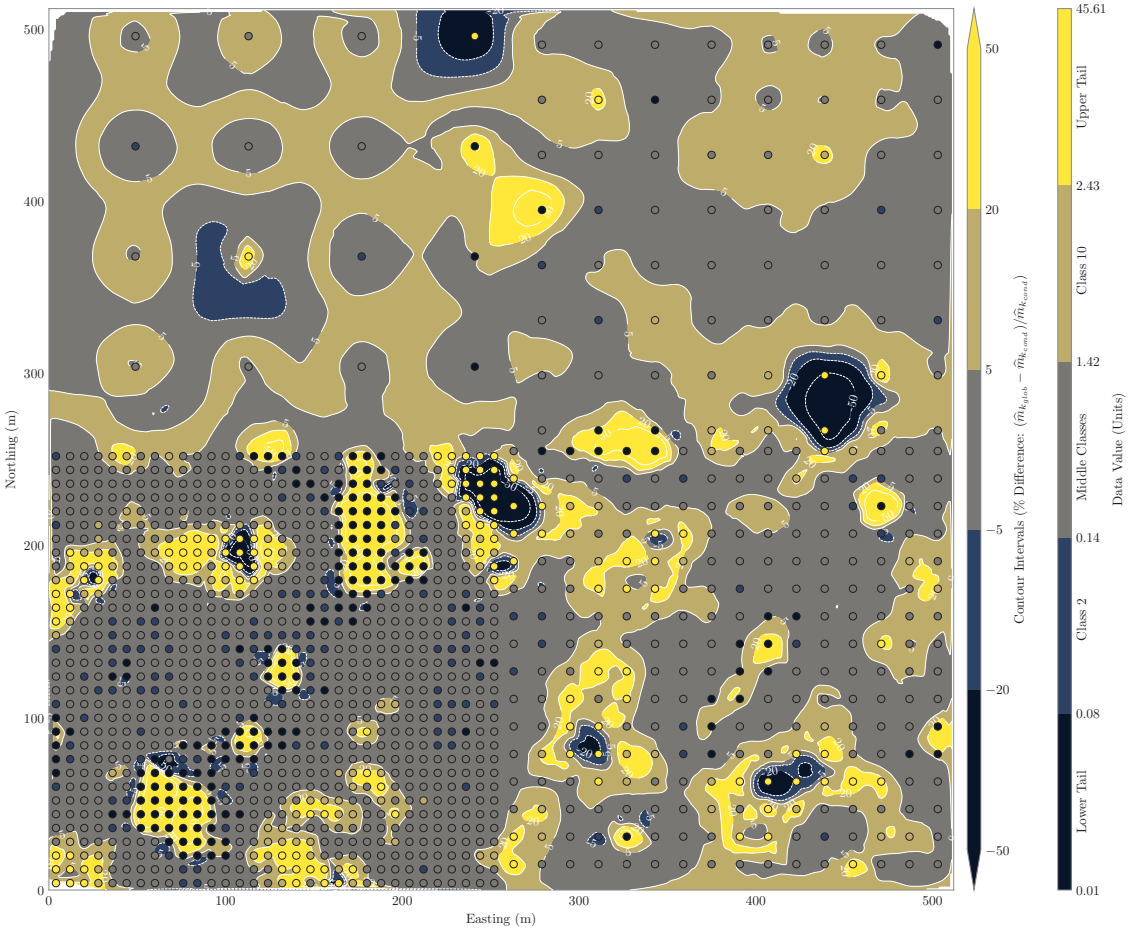


Figure 4.7: Contour map of estimation error between use of the correct, $\widehat{m}_{k_{cond}}$ and the stationary $\widehat{m}_{k_{glob}}$ (log-normal data, K $= 10$). Note the general alignment between the magnitude of the smallest and largest error contours and locations of the lower and upper two indicator classes.

with negative 50 percent error contours (blue fill) overlaid by conditioning data primarily in the upper tail (yellow). Looking closely at the error contours, they transition to a positive error of greater than 20 percent. Two features are explained: 1) the yellow-coloured data within the negative 50 percent contour have conditioning values greater than the global, indicator-class mean of 5.16 units, which results in high conditional, indicator-class means and a strong negative error, and 2) the yellow conditioning data overlaying the positive 10 percent contours immediately to the west and south of the negative 50 percent contour have values less than the global, indicator-class mean in the upper tail, which results in a positive error. The reader is referred to Figure 4.4, middle plot, in conjunction with Figure 4.7 to view the distribution of conditional, indicator-class means relative to the global,

indicator-class means.

Immediately to the west-southwest in Figure 4.7, located near coordinates (180 mE, 200 mN), is a region showing the opposite of the observations described above. The estimation error at this location is greater than 50 percent. It is due to the conditional, indicator-class means being less than the global, indicator-class means. Moreover, several small contours of negative estimation error (-5 percent) surround the region of high positive error and are overlain by conditioning data belonging to the lower two classes. In this instance, the conditional, indicator-class means in the lower tail are greater than the global values. The low magnitude of the conditioning data in the lower tail ($z_{k=0} \leq 0.08$ units), leads to large estimation error in this region, but it has little practical economic implication.

The patterns described for the two locations above are consistent across the study area and are similar across the range of evaluated threshold scenarios and sample grid spacings. With $(\mathrm{K} = 5)$-thresholds, the positive and negative error regions are larger in aerial extent because there is a greater average error between the global and conditional class mean values with fewer thresholds. With $(\mathrm{K} = 15)$-thresholds, the opposite is true.

The error distribution of the log-normal data ($\mathrm{K} = 10$) is asymmetrically centred about zero, with estimates using the global, indicator-class means generally being greater than estimates using the correct, conditional, indicator-class means (Figure 4.8). As expected, the mean-squared error decreases as the number of thresholds increases. The mean of the error is near zero because the magnitude of the negative errors is generally greater even though they are fewer in number.

4.4 Class mean dependence

The spatial arrangement and covariance relationships of the conditioning data dictate the conditional, indicator-class mean. A regression analysis using stochastic gradient boosting is undertaken to better understand the factors upon which the class mean depends

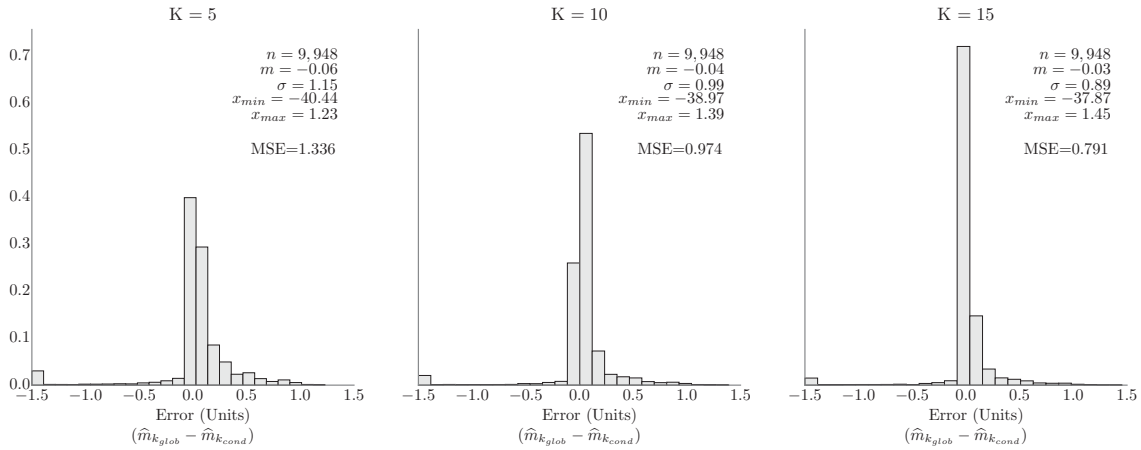


Figure 4.8: Contour map of estimation error between use of the correct, $\widehat{m}_{k_{cond}}$, and stationary, $\widehat{m}_{k_{glob}}$, values (log-normal data, K = 10). Note the general agreement between the magnitude of the error contours and locations of the lower and upper two indicator classes.

(Friedman, 2002; Hastie, Tibshirani, & Friedman, 2009). Stochastic gradient boosting is a supervised, ensemble-machine-learning technique that uses feature vectors to generate a predictions of the response variable, $\widehat{m}_{k_{cond}}$. The feature vectors are $\widehat{y}_{SK}$ and $\widehat{\sigma}_{SK}$. It is an attractive technique because it manages non-linear feature interactions, it is resistant to model overfitting, and it can handle features measured on difference scales. It requires parameter tuning and it is slow to train, but prediction is fast. Figure 4.9 illustrates the fitting and evaluation of the $\widehat{m}_{k_{cond}}$-response surface using all data (left) after training with a test subset of the data (right).

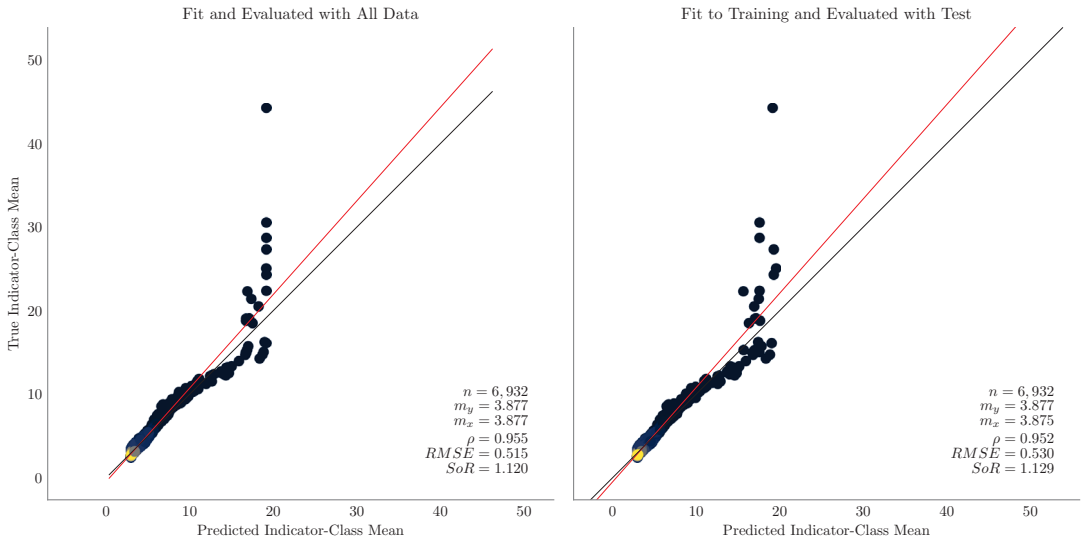


Figure 4.9: Fit and evaluation of a response surface.

The results of the analysis are presented in two ways: 1) feature importance, which expresses the relative importance of one predictor variable in comparison to the others, and 2) partial dependence plots that summarize the influence of a feature averaged over the input domain.

The relative importance of the conditional mean and conditional variance features does not show much variation between the various threshold scenarios considered. The $\widehat{y}_{SK}$-value exerts approximately 60 percent of the influence on the value of the indicator-class mean, with the $\widehat{\sigma}_{SK}$-value comprising the remainder (Table 4.2). The log-normal data reflect the opposite scenario, where the $\widehat{\sigma}_{SK}$-value exerts between $50 - 60$ percent of the influence.

Table 4.2: Relative importance of $\widehat{y}_{SK}$ and $\widehat{\sigma}_{SK}$ features predicting $\widehat{m}_{k_{cond}}$ by indicator class (K $=$ 10).

Feature	Indicator Class										
	L.T.	2	3	4	5	6	7	8	9	10	U.T.
Relative Importance (Gaussian Data)											
$\widehat{y}_{SK}$	0.60	0.56	0.62	0.63	0.64	0.61	0.62	0.64	0.64	0.56	0.61
$\widehat{\sigma}_{SK}$	0.40	0.44	0.38	0.37	0.36	0.39	0.38	0.36	0.36	0.44	0.39
Relative Importance (Log-Normal Data)											
$\widehat{y}_{SK}$	0.58	0.50	0.48	0.46	0.41	0.44	0.44	0.45	0.43	0.39	0.44
$\widehat{\sigma}_{SK}$	0.42	0.50	0.52	0.54	0.59	0.56	0.56	0.55	0.57	0.61	0.56

Note: L.T. = Lower Tail; U.T. = Upper Tail.

Two representative partial dependence plots are presented from the $(\mathrm{K} = 10)$-threshold scenario for the Gaussian and log-normal data sets: Class 6 and the upper tail (Figure 4.10 and Figure 4.11, respectively). The x-axis plots the range of values, with the distribution deciles plotted as small, vertical, black dashes on the axis. The y-axis displays the centred, partial dependence of the $\widehat{m}_{k_{cond}}$-response variable. In the Gaussian example (Figure 4.10), a linear relationship between the response and predictor variables is observed for Class 6. The $\widehat{y}_{SK}$ value results in little fluctuation in the conditional, indicator-class mean (± 0.02 Gaussian units). In the upper tail, there is a non-linear relationship between the predictor variables, with each predictor exerting a ten-times greater influence on the conditional class

mean in comparison to Class 6.

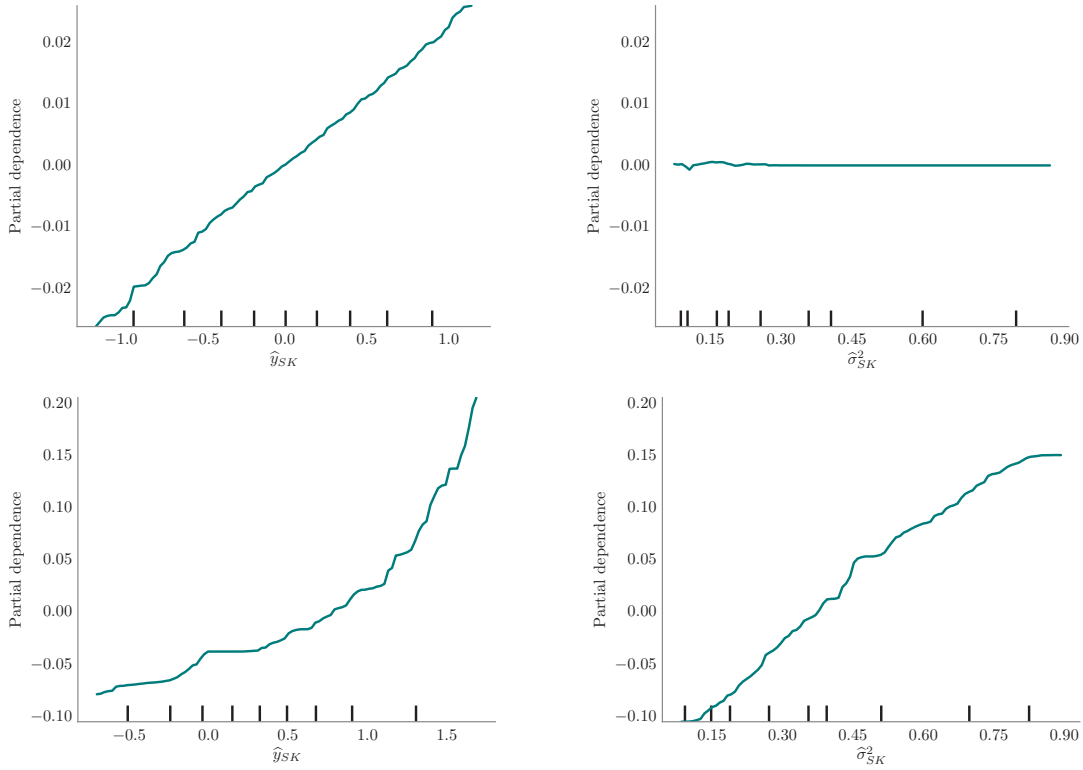


Figure 4.10: Partial dependence relationship of the $\widehat{m}_{k_{cond}}$-response variable to the $\widehat{y}_{SK}$ (left) and $\widehat{\sigma}_{SK}$ (right) for the median (top) and upper-tail (bottom) indicator classes (Gaussian data, $\mathrm{K} = 10$).

In the log-normal example (Figure 4.11), the $\widehat{\sigma}_{SK}$ controls more of the variability of the response variable. The predictor variables have significantly less impact on the response variable in the median indicator class than in the upper tail. Note that the abrupt, vertical change in the upper-right plot ($\widehat{\sigma}_{SK}$, median indicator class) is likely not real.

4.5 Impact on estimated contained metal

Grade-tonnage curves contrasting estimates of the log-normal data using the conditional and global, indicator class means are presented in Figure 4.12. Three indicator-threshold scenarios are considered: $\{\mathrm{K} = 5, 10, 15\}$. The estimates using the correct, conditional, indicator-class mean (teal lines) predict fewer tonnes and a higher average grade than the estimates using the global, indicator-class mean (grey lines). The plots illustrate the decreasing difference between the estimated tonnes and average grade above a cut-off as the

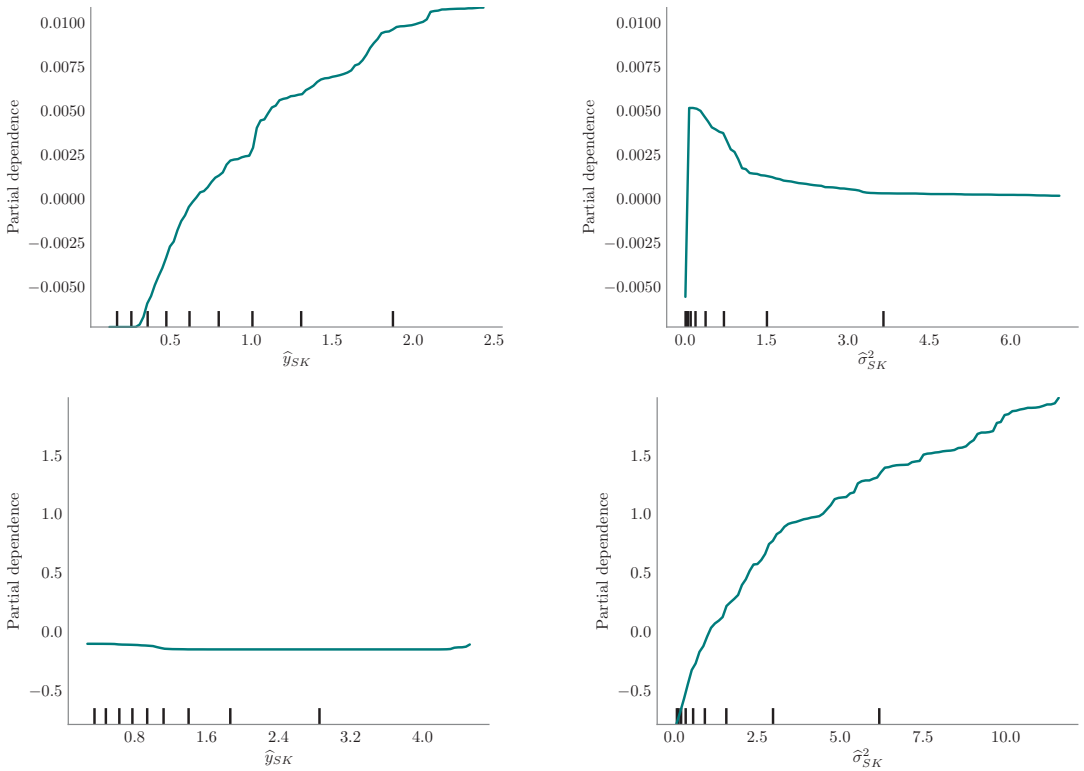


Figure 4.11: Partial dependence relationship of the $\widehat{m}_{k_{cond}}$-response variable to the $\widehat{y}_{SK}$ (left) and $\widehat{\sigma}_{SK}$ (right) for the median (top) and upper-tail (bottom) indicator classes (log-normal data, $\mathrm{K} = 10$).

number of thresholds increases. These results are also expressed in terms of percent difference (Table 4.3). The contained metal is also presented in the last two columns to contrast the ($\mathrm{K} = 5, 15$)-scenarios. The scenario with ($\mathrm{K} = 15$)-thresholds shows the least difference in contained metal content, which is expected given the smaller average error between the global and conditional, indicator-class-mean values.

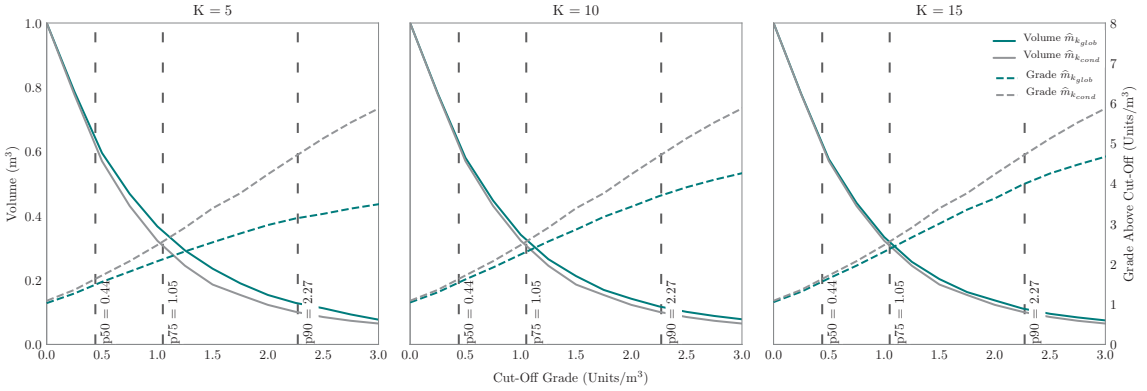


Figure 4.12: Grade-tonnage curves using $\widehat{m}_{k_{glob}}$ (teal) and $\widehat{m}_{k_{cond}}$ (grey) values (log-normal data, $\mathrm{K} = 5, 10, 15$). Vertical, black, dashed lines depict the global $50^{th}, 75^{th}, 90^{th}$-percentile values.

Table 4.3: Comparison of estimates using $\widehat{m}_{k_{glob}}$ and $\widehat{m}_{k_{cond}}$-values (log-normal data, $\mathrm{K} = 10$).

Cut-Off Grade (Units)	Δ Tonnes	Δ Avg. Grade Above Cut-Off	Δ Metal	Δ Metal $\mathrm{K} = 5$	Δ Metal $\mathrm{K} = 15$
0.0	0.0	-3.9	-3.9	-5.3	-3.1
0.5	1.8	-5.8	-4.0	-5.3	-3.2
1.0	5.7	-9.4	-4.2	-4.7	-3.5
1.5	14.2	-15.7	-3.7	-5.1	-3.2
2.0	15.5	-19.1	-6.5	-12.2	-4.5
2.5	18.9	-23.7	-9.3	-16.3	-7.3
3.0	20.4	-27.5	-12.6	-29.6	-8.5

4.6 Weighting of indicator probabilities

The results suggest the global, indicator-class mean in the upper two indicator classes is biased high in comparison to the expected, conditional, indicator-class mean. The magnitude of the bias is a function of the neighbouring conditioning values, with the strongest bias located in the areas of the lowest and highest values. It is important to note that the results comparing estimates using the global and conditional, indicator-class-mean values are not truly representative of a real-world scenario because the experiment is designed to isolate the class-mean component of the MIK estimator. The impact of the conditional, indicator-probability distribution is not yet considered.

Figure 4.13 illustrates the distribution of the Gaussian, conditional, indicator probabilities as a function of the conditional, indicator-class-mean values $(\mathrm{K} = 10)$. In each class, except for the upper tail, there is a roughly symmetric distribution of values about the expected, mean value. The asymmetry in the upper tail causes the majority of the conditional, indicator-class-mean values to be less than the expected value, while the highest conditional-means have a near-zero probability.

When the conditional, indicator-class means are correctly weighted by the conditional, indicator probabilities, they plot much closer to the global (stationary), indicator-class-mean (black, dashed lines in Figure 4.14). In fact, in the upper-tail class the global value appears to be a slightly conservative estimate of the correctly weighted, conditional, indicator-class means. In Figure 4.14, it is approximately seven percent less than the global-mean

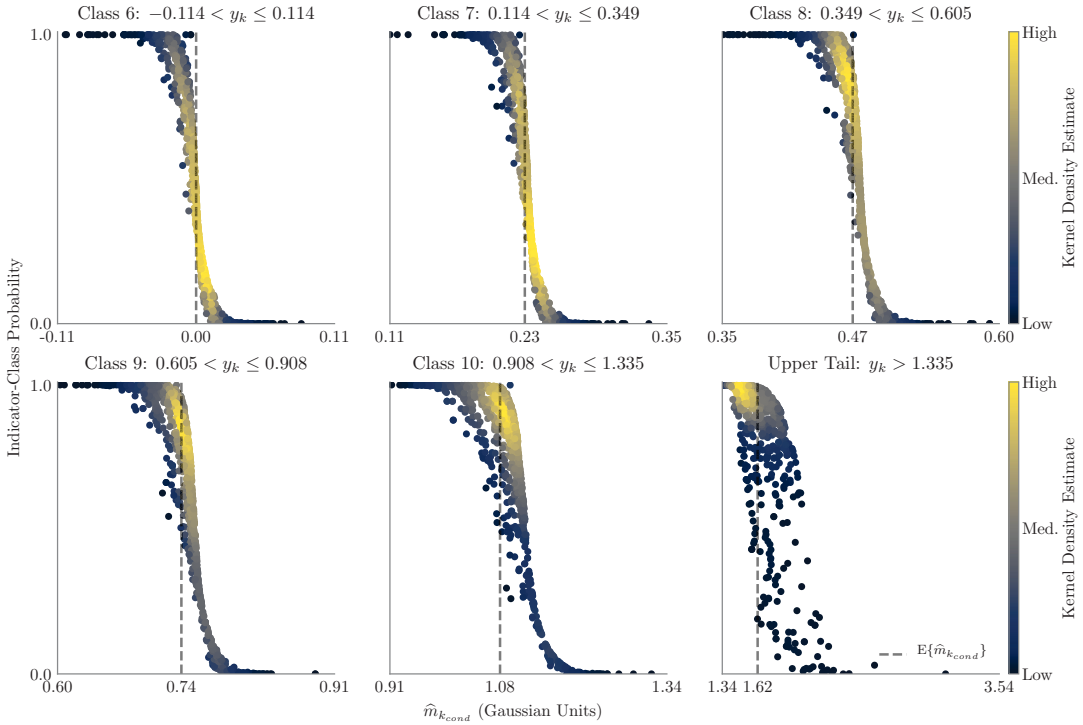


Figure 4.13: Distribution of conditional, indicator probabilities, against $\widehat{m}_{k_{cond}}$ (Gaussian data, K = 10). Vertical, dashed line is $E\{\widehat{m}_{k_{cond}}\}$.

value (K = 10).

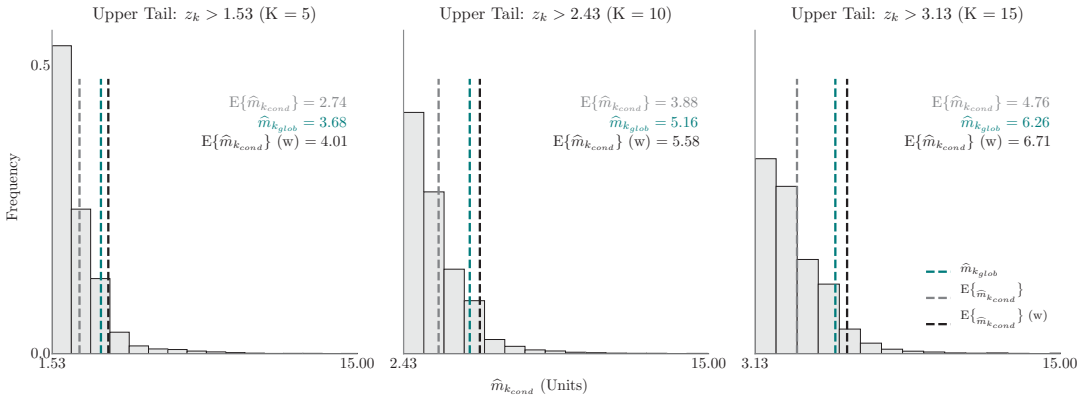


Figure 4.14: Upper-tail, log-normal distribution of $\widehat{m}_{k_{cond}}$ (K = 5, 10, 15), with the weighted, $\widehat{m}_{k_{cond}}$ plotted as the vertical, black, dashed line. The x-axis is truncated to 15.00 units.

4.7 Summary

Evaluation of the indicator-class-mean values as currently implemented in IK-software programs is conducted using both Gaussian and log-normal datasets under a MG frame-

work. A practical range of indicator thresholds is considered: $\{\mathrm{K} = 5, 10, 15\}$. As many real-world mining datasets are positively skewed, the results of the log-normal dataset are the focus of the final remarks. This study demonstrates that the conditional, indicator-class-mean values vary statistically and spatially due to the magnitude of conditioning data and their covariance relationships. The greatest estimation errors are spatially associated with regions of high and low conditioning data.

Under conditions that isolate the class-mean component of the MIK estimator, the conditional, indicator-class means differ the most from the global means in the tails of the distribution. The expected, conditional means, $E\{\widehat{m}_{k_{cond}}\}$, are consistently lower than the global means in the upper classes. As the coefficient of variation increases, the magnitude of the error in the upper tail increases and leads to a larger apparent bias. Though the upper tail is the focus of the analysis because of its disproportionate economic weight, apparent bias is also noteworthy in the lower tail. This could have implications with respect low-grade concentrations such as contaminant elements.

Under the conditions of this study, use of the global, indicator-class mean in the MIK estimator leads to an apparent underestimation of metal content that varies between $3 - 30$ percent depending on the number of thresholds used and the cut-off grade considered. The apparent bias described in this research is not observed in real-world estimates because the conditional-indicator probabilities weight the conditional mean. Probability weighting of the conditional, indicator-class mean is not considered until the end of this chapter, as the objective is to isolate and fully characterise the variability of the global, indicator-class mean. The results demonstrate that assuming the indicator-class mean to be stationary is not robust because a small bias remains even after probability weighting of the correct, conditional, indicator-class-mean values. Furthermore, the global, indicator-class mean is a slightly conservative estimate of the true value in the upper tail.

The stochastic gradient boosting analysis suggests the conditional, indicator-class means depend almost equally on the $\widehat{y}_{SK}$ and $\widehat{\sigma}_{SK}$ values output from MGK. Moreover, the good

fit of the response surface to the predictor variables ($\rho \approx 0.95$) indicates there is potential to generate a good local estimate of the indicator-class mean during MIK-estimation. Last, there is an opportunity to estimate slightly higher grades (and contained metal) in high-grade zones if the correct, conditional-indicator probabilities are used in conjunction with the correct, conditional, indicator-class means.

Chapter 5

Multiple Indicator Kriging of Non-Gaussian Data

Multivariate-Gaussian kriging calculates the correct mean and distribution of uncertainty at the unsampled locations if the regionalized variable is MG. In comparison to an MIK estimate using the same data, the mean-squared error of the MGK estimate will always be lower because its multivariate model is fully consistent and correct. As discussed in Section 1.1, it is difficult to prove a dataset is MG, but a high probability of bivariate Gaussianity can be demonstrated (e.g, J. Deutsch and Deutsch (2011); Mardia (1974); Royston (1992); Shapiro, Wilk, and Chen (1968); Svantesson and Wallace (2003)). Multivariate Gaussianity is then assumed as an extension from bivariate Gaussianity.

Data sets encountered in mining and petroleum environments, however, are commonly not multivariate Gaussian. Since indicator kriging does not rely on an underlying mathematical model, it seems reasonable that MIK should outperform MGK when a data set becomes sufficiently non-multivariate Gaussian. Two prerequisites are needed to test this hypothesis: 1) a controllable environment resembling realistic geological conditions that can generate realizations of varying degrees of non-Gaussianity, and 2) a method to measure the degree of non-Gaussianity.

The PLMR developed by Pereira and Deutsch (2020b) discussed in Section 2.2.3 addresses the first prerequisite. Next, a measure of non-Gaussianity, m_{ng}, is proposed to quantify the non-Gaussianity of a PLMR realization. Concluding this chapter is a comparison of MGK and MIK estimation performance from sampled, highly non-MG and MG realizations. The mean-squared error of the estimates is plotted as a function of the m_{ng} to ascertain if, or at which point MIK outperforms MGK.

5.1 Measure of non-Gaussianity

The degree of non-Gaussianity of the the $Z(\mathbf{u};a,q)$-RF is quantified to compare the MIK and MGK estimates. As a preliminary step, the $Z(\mathbf{u};a,q)$-RF is normal-score transformed to create a new, standardized RF, $X(\mathbf{u};a,q)$, that is required for MGK (Equation 5.1).

$$x = G^{-1}_{X(\mathbf{u};a,q)}(F_{Z(\mathbf{u};a,q)}(z)) \quad \forall\ z \tag{5.1}$$

This research proposes a method that measures the departure of the experimental $X(\mathbf{u};a,q)$-indicator variograms from the theoretically expected indicator variograms of a standard-normal, Gaussian variable. The measure of non-Gaussianity, m_{ng}, calculates the squared difference between the experimental indicator variogram points, γ_{exp}, and the corresponding theoretical indicator variogram points, γ_{theor}. The values are summed over all calculated h-lag distances, $n_{\mathbf{h}}$, and all K-indicators, n_k (Equation 5.2). The factor of 16 accounts for an expected difference of 0.25 between the experimental and theoretical variograms. The calculated value at lag, $\mathbf{h} = 0$, is excluded from the calculation because the variogram value is zero. Variogram points up to the specified ranges of the initial Gaussian Y_1 and Y_2 random fields detailed in Section 2.2.3 (Figure 2.4) are included in the calculation (Table 5.3).

$$m_{ng} = \frac{16}{n_{\mathbf{h}} n_k} \sum_{\mathbf{h}=1}^{n_{\mathbf{h}}} \sum_{k=1}^{n_k} \Big[\gamma_{theor}(\mathbf{h};k) - \gamma_{exp}(\mathbf{h};k)\Big]^2 \tag{5.2}$$

Figure 5.1 shows two indicator variogram h-scatter plots contrasting the difference in point densities in the hatched regions between the $x_{0.05}$ (grey) and the $x_{0.95}$-indicator probability thresholds (teal). The left-side graph represents the data pairs from the linear combination of the piecewise components, where $a = 0.50$. Since the linear combination of Gaussian variables produces a Gaussian distribution, the number of data pairs captured in the lower and upper tails should be equal. Note that the density contours of the data pairs are progressively less elliptical away from the centre of the plot, demonstrating that the distribution is not bivariate-Gaussian. There are approximately twenty percent more

data pairs in the upper tail than in the lower tail.

The right-side plot of Figure 5.1 highlights the strong asymmetry between the lower and upper tails, which is a reflection of the highly non-linear combination of the piecewise components, where $a = 0.99$. There are more than 150-percent more data pairs in the upper tail in comparison to the lower tail. The density contours are not elliptical in shape and reflect the relative dominance of the values above the q-threshold as a function of the a-scaling factor.

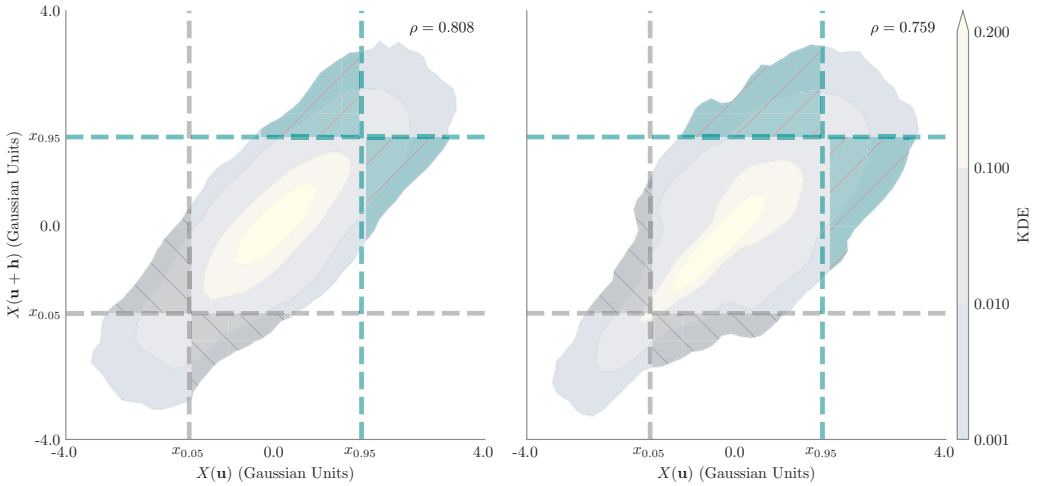


Figure 5.1: Comparison of h-scatter density in the hatched regions for the $z_{0.05}$ (grey) and $z_{0.95}$-indicator probability thresholds (teal) for $X(\mathbf{u}; 0.5, 0.5)$ (Gaussian, left) and a $X(\mathbf{u}; 0.9, 0.5)$ (non-Gaussian, right).

Figure 5.2 illustrates the corresponding indicator variograms of the upper and lower tails calculated at a lag distance, $\mathbf{h} = 4$ m (teal). The theoretical bivariate-Gaussian, indicator variograms (light blue) contrast the difference between the experimental points calculated from the MG scenario ($X(\mathbf{u}; 0.50, 0.50)$) on the left side and the highly non-MG scenario ($X(\mathbf{u}; 0.99, 0.50)$) on the right side. Note that the bivariate-Gaussian, indicator-variogram points are the reference points for comparison to the experimental points. As expected, the difference between the experimental and theoretical points is much greater in the non-MG scenario.

In total, ten realizations of the random fields, Y_1 and Y_2, are generated. They are combined using six different piecewise, linear-transform scenarios to create a total of sixty re-

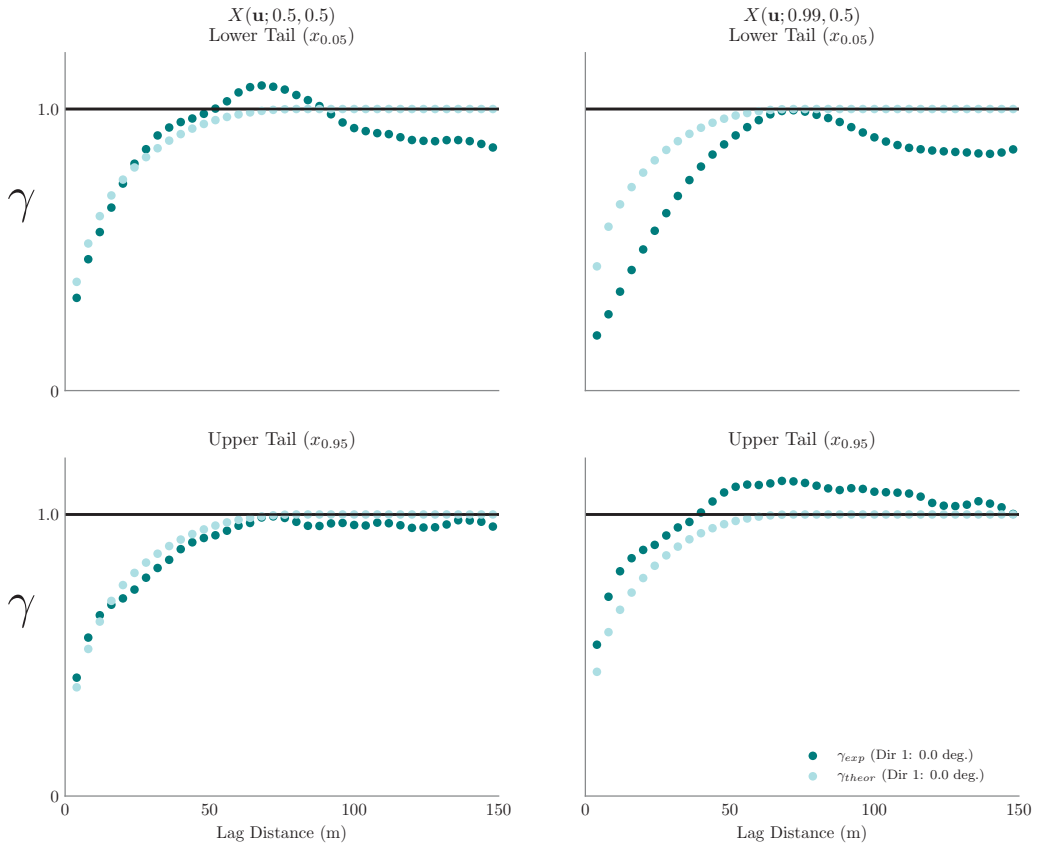


Figure 5.2: Comparison experimental indicator variograms (teal) against the corresponding theoretical bivariate-Gaussian variograms (light blue). Left-side plots are $X(\mathbf{u}; 0.50, 0.50)$ and the right-side plots are $X(\mathbf{u}; 0.99, 0.50)$.

alizations that are normal-score transformed before calculation of the the theoretical and experimental indicator variograms. The expected measure of non-Gaussianity, $E\{m_{ng}\}$, is calculated for the six $X_(\mathbf{u}; a, q)$-scenarios across the ten truth realizations (Table 5.1).

Table 5.1: Expected measure of non-Gaussianity.

$X(\mathbf{u}; a, 0.5)$; where $a =$	0.50	0.60	0.70	0.80	0.90	0.99
$E\{m_{ng}\}$	0.0675	0.0678	0.0756	0.0904	0.1132	0.1531

The expected measure of non-Gaussianity steadily increases as the *a*-parameter in the $X(\mathbf{u}; a, q)$-random variable is increased to be more non-Gaussian. The m_{ng}-value is stable across the different truth realizations.

5.2 Estimation in a non-Gaussian environment

The data comprising the sixty realizations are sampled on regular grid spacings following Table 2.2. The samples are the inputs for the MIK and MGK estimations. Each method uses a maximum of 30 samples. The search orientations are equal to the specified continuity directions of the initial Gaussian Y_1 and Y_2-random fields, and the search distances are equal to twice their specified ranges as listed in Table 2.1. Seven indicator thresholds are chosen to discretize the MIK distribution of uncertainty (Table 5.2).

Table 5.2: MIK cut-off and threshold values.

Indicator quantile (x_p)	$x_{0.05}$	$x_{0.10}$	$x_{0.30}$	$x_{0.50}$	$x_{0.70}$	$x_{0.90}$	$x_{0.95}$
Threshold Value	-1.645	-1.282	-0.524	0.000	0.524	1.282	1.645

A lag distance of $\mathbf{h} = 4$ m is used to calculate the variogram and indicator variograms. They are modelled using three spherical structures and low nugget values. Care is taken during the modelling of the indicator variograms to ensure that ranges change smoothly between each indicator threshold.

Figure 5.3 shows that both MIK (middle) and MGK (right) reproduce the spatial features of the highly non-MG reference realization, $X(\mathbf{u}; 0.99, 0.50)$ (left). The highest and lowest values are not reproduced well by MIK in comparison to MG (Figure 5.4). Also note that the RMSE of the MGK estimate is lower than the MIK estimate. Unexpectedly, the RMSE of all MGK estimates is lower than the MIK estimates, irrespective of the degree of non-Gaussianity of the reference distribution (Table 5.3). The difference in RMSE between the MIK and MGK estimates decreases as the data distribution becomes more non-Gaussian, but a cross-over point where MIK outperforms MGK is not evident. This pattern is consistent for the 16×16 m and 32×32 m-sample grids (Figure 5.5).

It is clear that the constant class mean is impacting the accuracy of the MIK estimate in the lower and upper tails, but the cross-validation plots in Figure 5.4 do not provide a good assessment of the relative accuracy of the estimated probability distributions. A

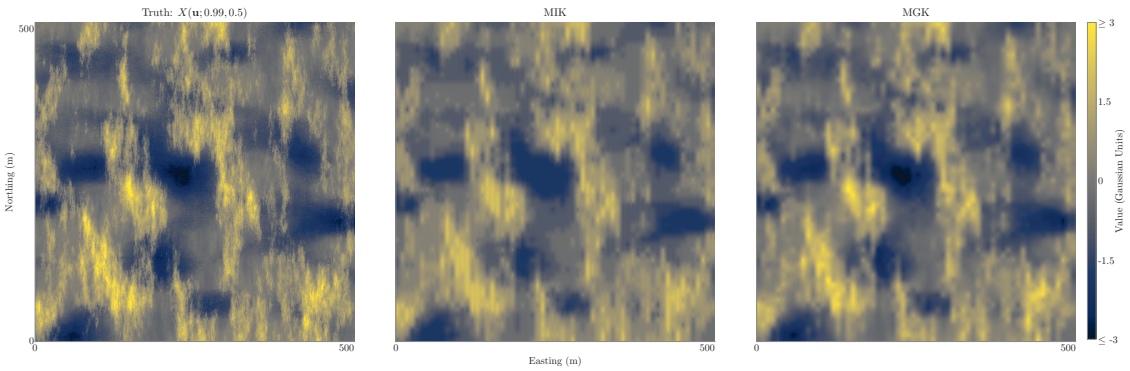


Figure 5.3: Location maps comparing the reference realization (left) to the MIK (middle) and MGK estimates (right).

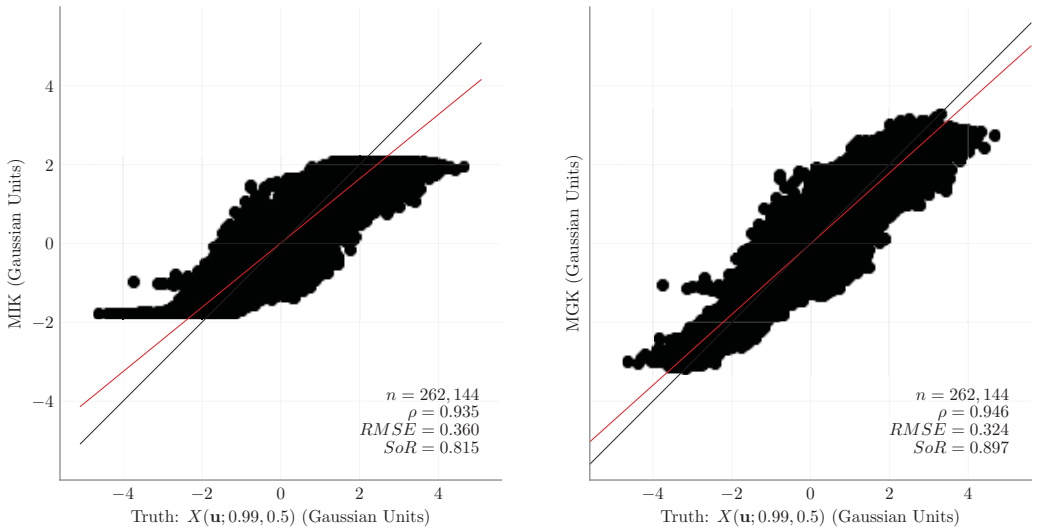


Figure 5.4: Cross validation plots comparing MIK (left) and MGK (right) estimates against reference values.

Table 5.3: Expected RMSE of MIK and MGK estimates in comparison to reference realizations.

Sample Grid (m)	Estimate	$X(\mathbf{u}; a, 0.5)$					
		$a = 0.50$	0.60	0.70	0.80	0.90	0.99
8×8	MIK	0.308	0.308	0.309	0.314	0.321	0.339
	MGK	0.270	0.271	0.273	0.277	0.285	0.305
	MIK - MGK	0.039	0.038	0.036	0.037	0.036	0.034
16×16	MIK	0.412	0.414	0.414	0.419	0.426	0.445
	MGK	0.370	0.371	0.374	0.379	0.387	0.410
	MIK - MGK	0.043	0.043	0.040	0.040	0.039	0.035
32×32	MIK	0.592	0.592	0.596	0.596	0.601	0.617
	MGK	0.545	0.546	0.549	0.554	0.563	0.584
	MIK - MGK	0.047	0.046	0.047	0.042	0.038	0.033

comparison of the estimated MIK and MGK probabilities at each indicator threshold shows that in the upper tail, MGK consistently correctly estimates the value to belong to the upper-

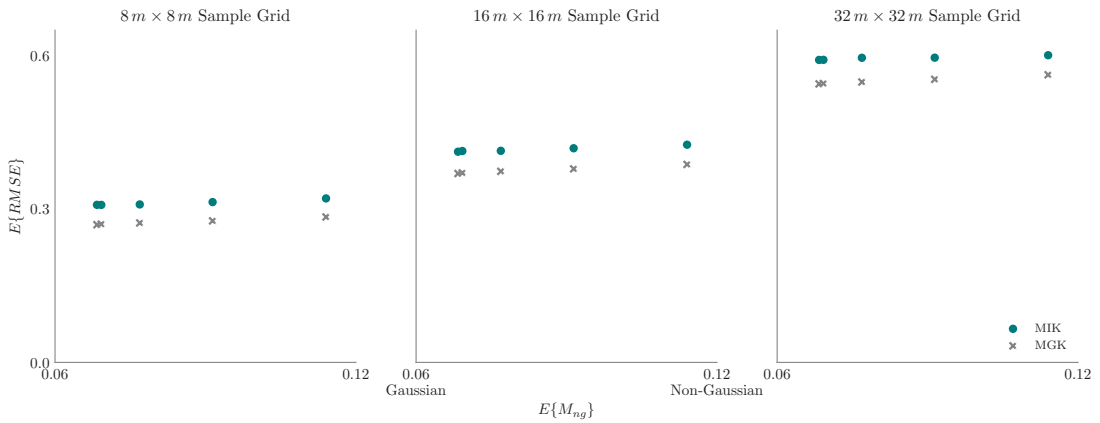


Figure 5.5: Comparison of expected RMSE as a function of expected measure of non-Gaussianity.

tail class more frequently than MIK. This is independent of the degree of non-Gaussianity of the distribution.

Estimation accuracy and precision of the MIK and MGK estimates of the six $X(\mathbf{u}; a, q)$-scenarios are compared in Figure 5.6. The MIK estimates occupy the first two rows of the plot and the MGK occupy the last two rows. The $X(\mathbf{u}; a, q)$-scenarios become more Gaussian moving from left to right. The accuracy plots show that both estimation methods are accurate (plot on or above the $45°$-line) and are in agreement with the RMSE values calculated in Table 5.3. The data points at each probability interval reflect poorer precision of MIK in comparison to MGK, as they plot further away from the $45°$-line. Furthermore, the MGK estimates noticeably increase in precision across the probability intervals as the $X(\mathbf{u}; a, q)$-realizations become more Gaussian. There is minor improvement in the precision of the MIK estimates.

Figure 5.7 quantifies the precision of the estimates, by calculating the mean of the variance of the conditional distributions (in squared, Gaussian units). One-hundred quantiles are used to discretize the MIK conditional distribution. The layout of the plot similar to Figure 5.6. The MIK estimates decrease from a mean conditional variance of 0.27 to 0.22. Note that the scale of the x-axis changes between the MIK and MGK plots in the figure. It is significantly higher than the MGK values, which decrease from 0.11 to 0.08. Note that the mean of the MIK conditional variances could be skewed upward due to the discretiza-

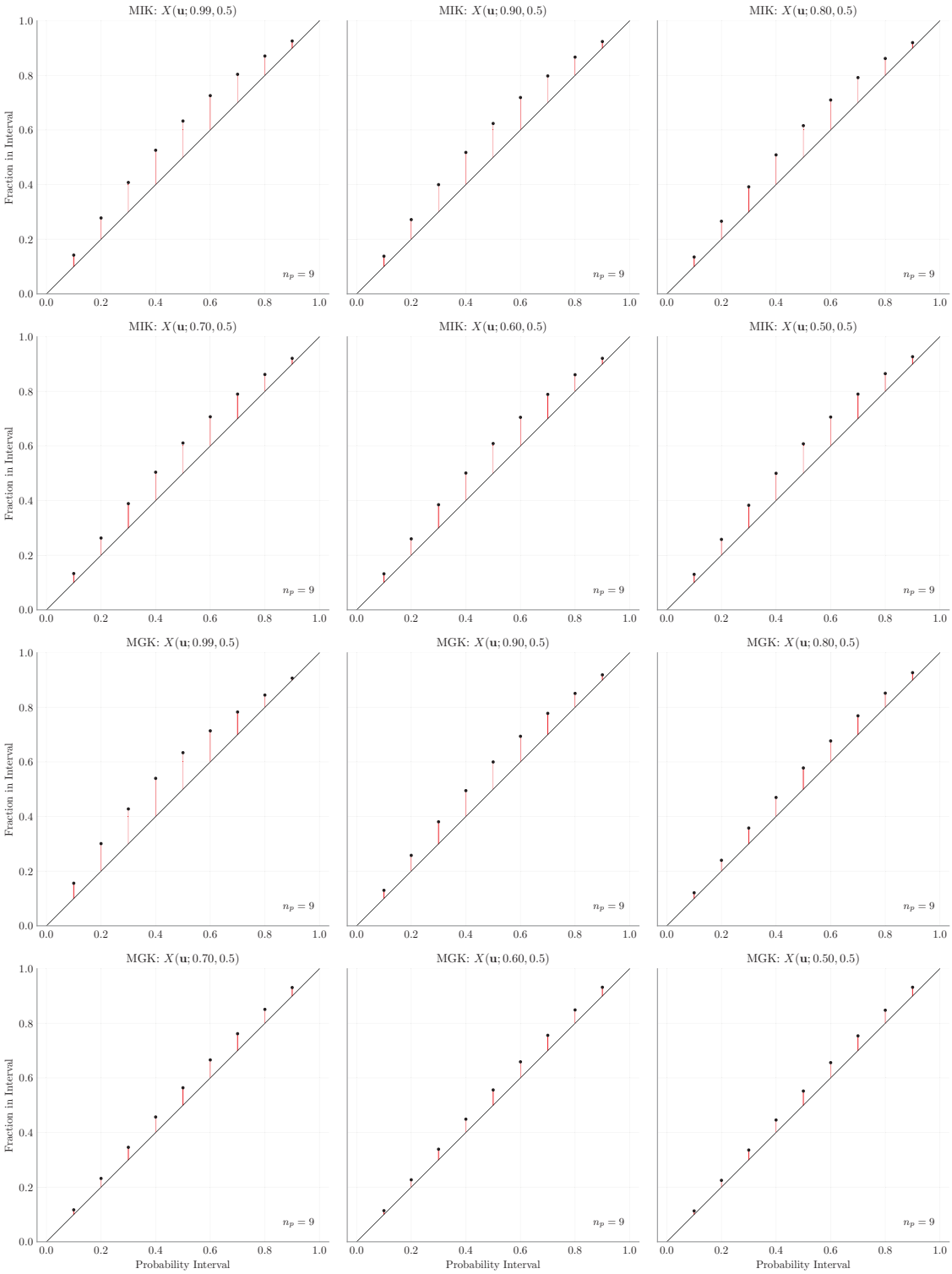


Figure 5.6: Accuracy of MIK (top two rows) and MGK (bottom two rows) for the six $X(\mathbf{u}; a, q)$-scenarios. 8×8 sample grid. Measure of non-Gaussianity decreases from left to right.

tion of the conditional distribution, so the interquartile range is also compared to discount any potential effects of the distribution tails. In all cases, the upper quartile of the MIK

distribution is greater than the maximum value of the MGK distribution.

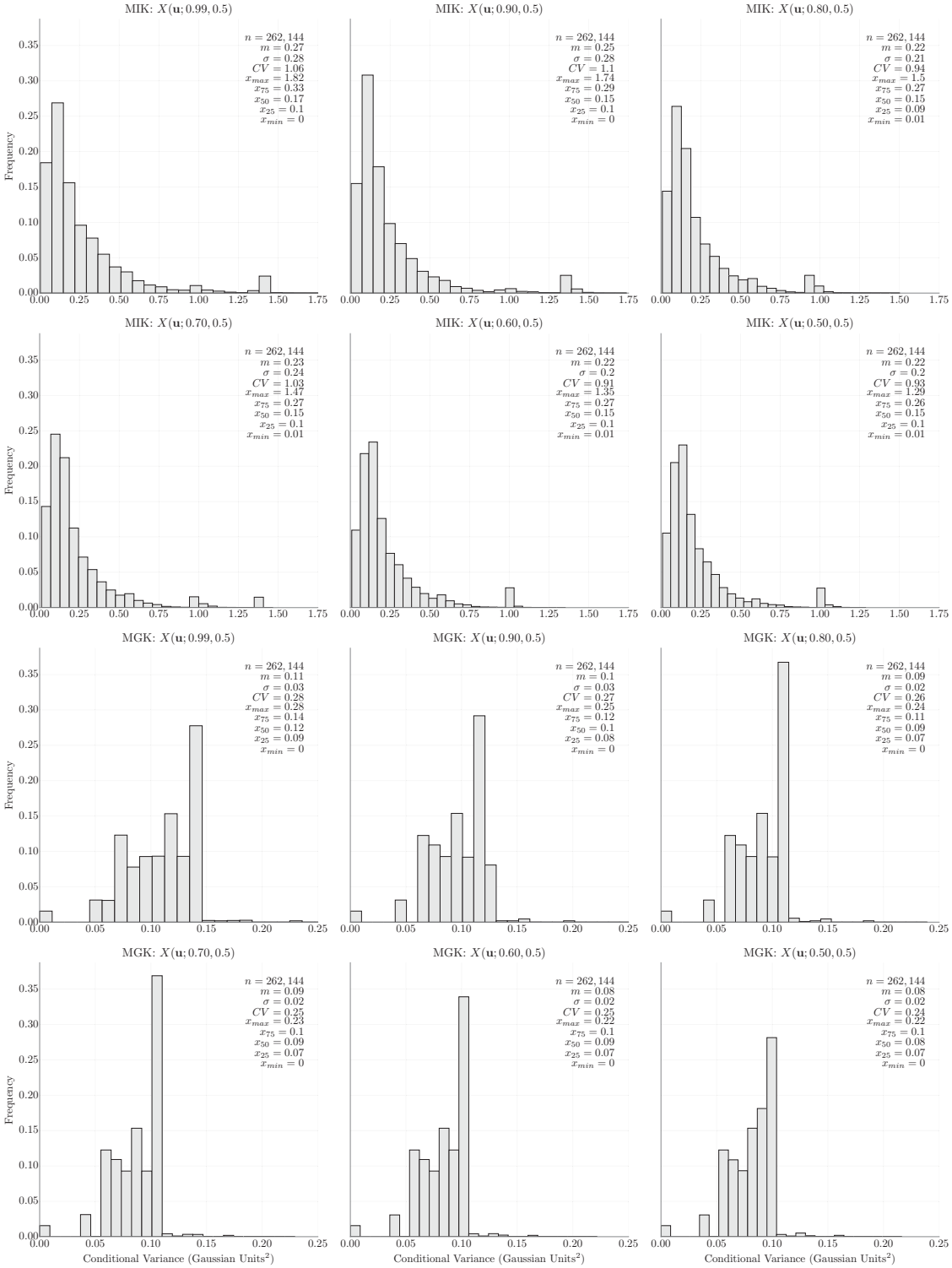


Figure 5.7: Conditional-variance distributions of MIK (top two rows) and MGK (bottom two rows) for the six $X(\mathbf{u}; a, q)$-scenarios. 8×8 sample grid. Measure of non-Gaussianity decreases from left to right. Note change of scale in x-axis between MIK and MGK estimates.

5.3 Summary

Unconditional realizations are combined using the novel PLMR transform by Pereira and Deutsch (2020b) to generate a set of realizations ranging from highly non-MG to MG that act as synthetic analogues to real-world geological environments. A method is presented to quantify the degree of non-Gaussianity of the realizations. The MIK and MGK estimates generate accurate predictions for each of the realizations, but the distribution of uncertainty predicted by MIK is consistently greater than the MGK estimate.

Comparison of estimation error shows that MGK is consistently lower than MIK across all of the reference realizations, even in the highly non-MG scenario. Initially, it was expected that there would be a cross-over point where MIK would outperform MGK once the distribution became sufficiently non-Gaussian. This warrants additional investigation. The results show the multivariate-Gaussian probability distribution is more consistent than the distribution generated non-parametrically by the indicator model. Values are more often correctly estimated to be in the upper tail by MGK. The indicator-class mean negatively impacts estimation quality in the upper tail and does not reproduce the highest values.

Chapter 6

Conclusions and Future Work

Multiple-indicator kriging is commonly used in the mining industry. It is attractive to geostatistics practitioners because the indicator transform manages outlier values and the conditional distribution is non-parametric. This research utilizes an MG environment to study the behaviour and dependencies of the component parts of the MIK estimator in isolation: 1) the conditional, indicator-probability distribution, and 2) the indicator-class means. Synthetic non-Gaussian datasets provide analogues of geological conditions and are used to compare the estimation performance of MIK and MGK. The bivariate conditional-mean / conditional-variance distribution is used to assess the representivity of the results. Multiple reference realizations sampled using several grid spacings provide representative reference values for the studies in this research.

6.1 Research contributions

In Chapter 3, estimation of the indicator-probability thresholds is shown to be unbiased, but estimation accuracy varies locally over small distances and does not appear to exhibit spatial dependence. When considering estimation of the indicator-class probability intervals, the estimate remains unbiased, but the error of the interval probabilities relative to the absolute CDF probabilities appears larger in both magnitude and spatial extent. The research demonstrates that MIK generates unbiased, but erroneous values.

In Chapter 4, Gaussian and log-normally distributed datasets demonstrate that the indicator-class means vary as a function of the conditioning data and their spatial covariance relationships. Comparison of the global (stationary), indicator-class means to the correct values reveals the greatest variability in the lower and upper-tail classes. In the

upper-tail class, the highest values in the distribution cause a consistently high bias in the global, indicator class mean in comparison to the expected, correct value. The opposite is true in the lower-tail class. The number of indicator classes and the characteristics of the tail distributions explain the magnitude of the observed bias.

Estimation using the global, indicator-class means reveals that the greatest estimation error occurs in regions with the lowest and highest conditioning data. Changes in the magnitude of the error are also greatest in these regions. Furthermore, the positive and negative estimation errors exhibit spatial dependence on the global, indicator-class-mean value. It is also noteworthy that estimates using the global, indicator-class mean are biased low when the indicator-probability distribution is *not* considered. The apparent low bias disappears during normal MIK estimation because the indicator probability distribution weights the indicator-class means. An additional research contribution shows that the correct, indicator-class-mean value can be predicted by the conditional-mean and conditional-variance values. This finding represents an opportunity to improve MIK estimates by varying the mean locally.

In Chapter 5, the PLMR transform is used to control the degree of non-Gaussianity of the data distribution. A method is presented to quantify the degree of non-Gaussianity. Estimates using samples of the non-Gaussian reference data consistently show that MIK has higher estimation error than MGK. The uncertainty distribution determined by MIK is also consistently wider and less precise than MGK. These results are unexpected, but they demonstrate that MGK is a robust estimation technique across a variety of non-Gaussian scenarios. The place of MIK remains unclear; however, the procedures and standards to assess the relative performance of MIK and other techniques are documented more clearly.

6.2 Research limitations

Outlier management is not addressed in this research. Indicator kriging is presented as a technique that manages outlier data, but there is no prior treatment of extreme values in the

synthetic data sets or assessment of their affect on the upper-tail class mean. This research shows that the estimated contained metal is sensitive to the number of indicator thresholds, and most likely, to the extreme values in the distribution tail. An improved framework would consider these in tandem. Furthermore, the assessment of the significance of the indicator-class mean values is limited because it is case specific.

The findings show that MGK outperforms MIK in all tested non-Gaussian environments. It is important to note that it is not possible to test all non-Gaussian settings, so a general conclusion cannot be reached. The synthetic non-Gaussian data, while necessary to control the degree of non-Gaussianity, limits the findings. A carefully controlled study using real-world mining data would be beneficial. Though synthetic data are required to adequately control the experiments, it cannot replace real-world data.

Multiple-indicator kriging yields an estimate at the scale of the data. In mining applications, a volume-support correction is required to understand the grade-tonnage distribution at the scale of mining, commonly referred to as the selective mining unit (SMU). Localized indicator kriging (LIK) developed by Hardtke, Allen, and Douglas (2011) is commonly used to predict the distribution of grades inside the SMU since the local grade distribution inside the SMU is not known after the change of support. This research does not consider a change of support or localization of grades.

6.3 Future work

Avenues for future research could focus on topics of indicator-threshold selection, indicator-variogram modelling, the estimation algorithm, construction of the CCDF, and post processing. No formal methodology is followed during indicator selection. Common practice relies on "rules of thumb" and the experience of the practitioner. The CDF is generally modelled at the distribution deciles, with changes made to account for features in the CDF, cut-off grade, and additional discretization of high grade bins. Research opportunities include optimization and placement of thresholds based on estimation performance criteria

such as MSE, selecting thresholds based on the CCDF, and development of validation measures for threshold selection. A formal methodology, except for some "rules of thumb", also does not exist for calculation and modelling of indicator variograms. There is an opportunity to develop a technique to validate the indicator discretization of the bivariate distribution in order to reduce order-relations errors.

Ordinary kriging is almost always used in MIK because it handles non-stationarity better than simple kriging. Comparison of MIK using OK and SK with a locally varying mean would be interesting. Also, the IK estimation algorithm could be modified to consider all data as possible thresholds by projecting them on the unsampled locations to construct the CCDFs. Distant data would be given a weight of zero and close data would be considered to a variable extent depending on the indicator variogram.

The CCDF construction could be improved by modifying the scaling process to better represent local conditions. Post processing of the MIK estimate could be improved when applications require estimated values at a different volume support (e.g., a mining volume). Current options available in software include affine and log-normal corrections, but these are considered inadequate by current practice standards. Implementation of a global-change-of-support option could be investigated.

References

Anaconda Software Distribution. (2016). *Conda.* Anaconda.org. Retrieved from https://www.anaconda.com

Babakhani, M. (2014). *Geostatistical Modeling in presence of extreme values* (Unpublished master's thesis). University of Alberta, Edmonton, AB, Canada.

Carvalho, D., & Deutsch, C. (2017). In J. Deutsch (Ed.), *An overview of multiple indicator kriging.* Geostatistics Lessons. Retrieved from http://www.geostatisticslessons.com/lessons/mikoverview

Chilés, J., & Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty* (2nd ed.). Hoboken, New Jersey: Wiley.

Chiquini, A., & Deutsch, C. (2017). In J. Deutsch (Ed.), *A simulation approach to calibrate outlier capping.* Geostatistics Lessons. Retrieved from http://www.geostatisticslessons.com/lessons/simulationcapping

Costa, J. (2003). Reducing the impact of outliers in ore reserves estimation. *Mathematical Geology*, *35*(3), 323–345.

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

Deutsch, J., & Deutsch, C. (2011). Plotting and checking the bivariate distributions of multiple gaussian data. *Computers & Geoscience*, *37*, 1677–1684.

Deutsch, J., Deutsch, M., Martin, R., Black, W., Acorn, T., Barnett, R., … Samson, M. (2015). *Pygeostat version 0.6.6.* Centre for Computational Geostatistics, University of Alberta.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. Retrieved from https://www.sciencedirect.com/science/article/pii/S0167947301000652 doi: https://doi.org/10.1016/S0167-9473(01)00065-2

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations* (Second ed.). John Wiley & Sons, Inc., New York. Retrieved from https://doi-org.login.ezproxy.library.ualberta.ca/10.1002/9781118032671 (A Wiley-Interscience Publication) doi: 10.1002/9781118032671

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press, Inc.

Hardtke, W., Allen, L., & Douglas, I. (2011). Localised indicator kriging. *Publication Series - Australasian Institute of Mining and Metallurgy*, *11*, 141 - 147.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). *Array programming with NumPy* (Vol. 585) (No. 7825). Springer Science and Business Media LLC. Retrieved from https://doi.org/10.1038/s41586-020-2649-2 doi: 10.1038/s41586-020-2649-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer, New York. Retrieved from http://www-stat.stanford.edu/~tibs/ElemStatLearn/

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.5281/zenodo.1343133

Isaaks, I., & Srivastava, R. (1989). *An Introduction to Applied Geostatistics*. Oxford University Press, Inc.

Journel, A. (1983). Nonparametric estimation of spatial distributions. *Mathematical Geology*, *15*(3), 445-468.

Journel, A. (1989). *Fundamentals of geostatistics in five lessons* (Vol. 8). Washington, DC: American Geophysical Union.

Journel, A., & Alabert, F. (1989). *Non-gaussian data expansion in the earth sciences* (Vol. 1).

Journel, A., & Deutsch, C. (1993). *Entropy and spatial disorder* (Vol. 25).

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational

workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (p. 87 - 90).

Kyriakidis, P., Deutsch, C., & Grant, M. (1999). Calculation of the normal scores variogram used for truncated gaussian lithofacies simulation: theory and FORTRAN code. *Computers & Geosciences*, *25*, 161–169.

Mardia, K. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *36*(2), 115–128.

Matheron, G. (1970). Random functions and their application in geology. In D. Merriam (Ed.), *Geostatistics, a colloquium.* (pp. 79–87). Plenum Press; (Computer Applications in the Earth Sciences).

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 56 - 61). doi: 10.25080/Majora-92bf1922-00a

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, 641-646.

Parker, H. (1991). Statistical treatment of outlier data in epithermal gold deposit reserve estimation. *Mathematical Geology*, *23*(2), 175–199.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pereira, F., & Deutsch, C. (2020a). The bivariate and multivariate property of the PLMR. In C. Deutsch (Ed.), *CCG annual report 22* (pp. 126-1 – 126-11). Centre for Computational Geostatistics.

Pereira, F., & Deutsch, C. (2020b). A new non-Gaussian simulation framework: the PLMR. In C. Deutsch (Ed.), *CCG annual report 22* (pp. 125-1 – 125-10). Centre for Computational Geostatistics.

Pyrcz, M., & Deutsch, C. (2018).

In J. Deutsch (Ed.), *Transforming data to a gaussian distribution.* Geostatistics Lessons. Retrieved from `Retrievedfromhttp://www.geostatisticslessons.com/lessons/normalscore`

Reback, J., McKinney, W., jbrockmendel, den Bossche, J. V., Augspurger, T., Cloud, P., … Mehyar, M. (2020). *pandas-dev/pandas: Pandas 1.0.2.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.3708035` doi: 10.5281/zenodo.3708035

Rivoirard, J., Demange, C., Freulon, X., Lécureuil, A., & Bellot, N. (2013). A top-cut model for deposits with heavy-tailed grade distribution. *Mathematical Geosciences*, *45*(8), 967–982.

Rossi, M., & Deutsch, C. (2014). *Mineral Resource Estimation*. Springer.

Royston, J. (1992). Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, *2*, 117–119.

Shapiro, S., Wilk, M., & Chen, H. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, *63*(324), 1343-1372.

Svantesson, T., & Wallace, J. (2003). Tests for assessing multivariate normality and the covariance structure of MIMO data. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).* (Vol. 4, p. IV-656).

Thode Jr., H. (2002). *Testing for Normality*. Marcel Dekker, New York, NY. doi: https://doi-org.login.ezproxy.library.ualberta.ca/10.1201/9780203910894

Verly, G. (1983). The multigaussian approach and its applications to the estimation of local reserves. *Journal of the International Association for Mathematical Geology*, *15*(2), 259–286.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., … SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2