THESIS SUPERVISOR

Dr Nasser Madani

THESIS CO-SUPERVISOR

Dr Mohammad Maleki

Thesis

[MINE 526]

APPLICATION OF SEQUENTIAL INDICATOR SIMULATION TO MODEL NON-STATIONARY GEOLOGICAL DOMAINS COMBINING WITH A MACHINE LEARNING ALGORITHM

Almas Amirzhan

ID: 201781995

17th April 2023

Astana, Kazakhstan

ORIGINALITY STATEMENT

I, Almas Amirzhan, hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at Nazarbayev University or any other educational institution, except where due acknowledgement is made in the thesis.

Any contribution made to the research by others, with whom I have worked at NU or elsewhere is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signed on 2023-04-17

_____

ABSTRACT

Resource estimation is an essential aspect of the development process for any mining project. The geological domains are defined based on data obtained from boreholes, with the goal being to determine the mineral grades in the geological domains. Geostatistics assumes that the joint distribution of geological attribute values is consistent across homogeneous domains and is defined by a stationary covariance function. However, the nature of geological systems often contains uncertainties and variations in structure and behaviour.

Sequential Gaussian and Sequential Indicator Simulation are one of several methods used for simulating continuous and categorical variables in 3D geological modelling. Despite its advantages, this method and other conventional techniques have been criticized for not effectively capturing local mean values, variance, and spatial continuity changes.

The traditional algorithms used in the industry are not suitable for non-stationary geological domains, as they are designed for stationary target simulation variables. This thesis proposes using Multinomial Logistic Regression as an alternative method for simulating the spatial properties of non-stationary geological domains. The technique will be applied to a copper-porphyry deposit that shows clear signs of non-stationarity.

The mineral resource model will be created by weighting the copper grade estimates based on the probability of occurrence of different rock types in various geo-domains. The generated probability maps will be evaluated using various criteria, including visual inspection of realizations, probability maps, replicas of each geo-domain fraction, connectedness metrics, and trend analysis.

ACKNOWLEDGEMENTS

I express my sincerest gratitude to Dr Nasser Madani, my supervisor, for providing me with continuous support throughout my graduate studies and research. His patience, motivation, and extensive knowledge were instrumental in guiding me through this dissertation's research and writing process. I am genuinely grateful to have had him as my supervisor, who served not only as a professor and advisor but also as a mentor and friend.

I also extend my appreciation to the School of Mining and Geosciences at Nazarbayev University for giving me a chance to pursue a master's program and engage in research projects that have offered valuable experiences for my future.

Finally, I want to thank my family for their unwavering support in all my endeavors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 RESEARCH BACKGROUND

Geological modelling is a crucial procedure preceding resource estimation characterized by distinguishing sub-units of the deposit known as the "geological domain". The deterministic approach implies that each geo-domain has homogeneous properties. Deterministic modelling involves estimating the grade distribution the according to geological interpretations, the data from survey boreholes, or by constructing a model based on those data points. By using this model, the geological domains within a deposit are described in a unique manner. This kind of splitting allows us better to describe each geo-domain's traits, especially grade distribution. However, it does not take into account uncertainties resulting from establishing boundaries between geological domains. Stochastic simulations were developed to overcome the limitations of the conventional deterministic approach. By providing a probabilistic description of geological domains, this approach provides a way to quantify uncertainty and enhance geological control for quantitative variables of interest in contrast to deterministic approaches (Dubrule, 1993; Dowd, 1994; Emery and González, 2007a, 2007b). Target variables show homogeneous behaviour inside each geological domain. However, the boundaries of the domains are uncertain (Dubrule, 1993; Dowd, 1994; Emery and González, 2007a, 2007b).

The subject of this thesis is Sequential Indicator Simulation (SIS), the stochastic method widely used in most commercial software, designed to simulate categorical variables (Journel & Alabert, 1988; Alabert, 1987; Journel & Isaaks, 1984; Journel, 1983). The technique was investigated and proposed by François Alabert (Alabert, 1987) and André Journel (Journel & Alabert, 1990). Different authors have constantly modified SIS (Goovaerts, 1994; Deutsch et al., 1992; Goovaerts et al., 1997; Gómez-Hernández & Srivastava, 1990) since it is the most common geostatistical tool for simulation of geological domains. This project will focus on enhancing the current SIS algorithm to model the non-stationary geological domains. It then later will investigate the modelling of ore grade, taking into account the models obtained from the proposed SIS algorithm.

1.2 PROBLEM STATEMENT

The variogram-based stochastic simulation algorithms such as Sequential Indicator Simulation (Journel & Isaaks, 1984), plurigaussian simulations (Dowd et al., 2003) and truncated Gaussian simulation (Galli et al., 1994) were proposed as flawless, innovative techniques. Robust and straightforwardness of stochastic methods were proved by

experimental and practical evidence. Moreover, stochastic simulation techniques can easily incorporate secondary (soft) information to produce more accurate realizations. A substantial disadvantage of variogram-based geostatistical methods is that they rely on stationary assumptions of the deposit.

Stationary assumptions are valid only for homogeneous regionalized variables (Matheron, 1971). There are two recognized types of stationarity assumptions. First-order stationarity expects the constant mean value of a categorical variable within the domain, while second-order stationarity supposes a constant mean and also that the observed covariance between random points leans on the distance between those points. The practice shows that stationary assumptions could be more practical. Categorical variables denoting geo-domains display fluctuating spatial continuity as well as fluctuating mean values. Conventional methods are appropriate when a geological body does not have a precise contour. Another requirement is a detailed variogram analysis of spatial continuity (Mizuno and Deutsch, 2022). Examples of such structures are highly diagenetically altered facies. Therefore conventional Sequential Indicator Simulation technique could be better for modelling the complex heterogeneous geological domains. So distinguished issues are listed below:

- The traditional SIS algorithm improperly reproduces the compactness and spatially adjacent geological properties essential in simulating geological domains.
- Traditional SIS is unsatisfactory in the case of heterogeneous geological domains and features displayed at a large scale.

Generally, SIS is capable of modelling non-stationary complex geo-domains by acknowledging spatially varying attributes like indicator mean value. (Deutsch and Journel, 1992; Ravenne et al., 2002; Beucher et al., 1993). Although obtaining soft data is only sometimes straightforward, implementing this kind of information can be a clue for modelling non-stationary domains. Another problem is that interpretive geo-domains produced by explicit and implicit modelling are deterministic. In this regard, the derived deterministic model must be converted to a probabilistic model. One possible solution is to calculate the local proportions of geological domains at the target block node using the neighbouring data around that location (Madani & Emery, 2015). However, the method is entirely subjective, and the size of the window to include the neighbourhood data needs to be adequately formalised (Madani & Emery, 2017).

Limitations of conventional Sequential Indicator Simulation can be neglected by combining this technique with Multinomial Logistic Regression. The rationale for choosing it is that MLR is powerful enough to produce soft data by using only the conditioning data. In this

way, combining two techniques will allow higher accuracy in modelling non-stationary geological domains. The figure below represents the prime example of non-stationary domains that usually takes place in copper-porphyry deposits, i.e. the investigation is designed to model domains in and calculate grade distribution in such copper-porphyry deposits.

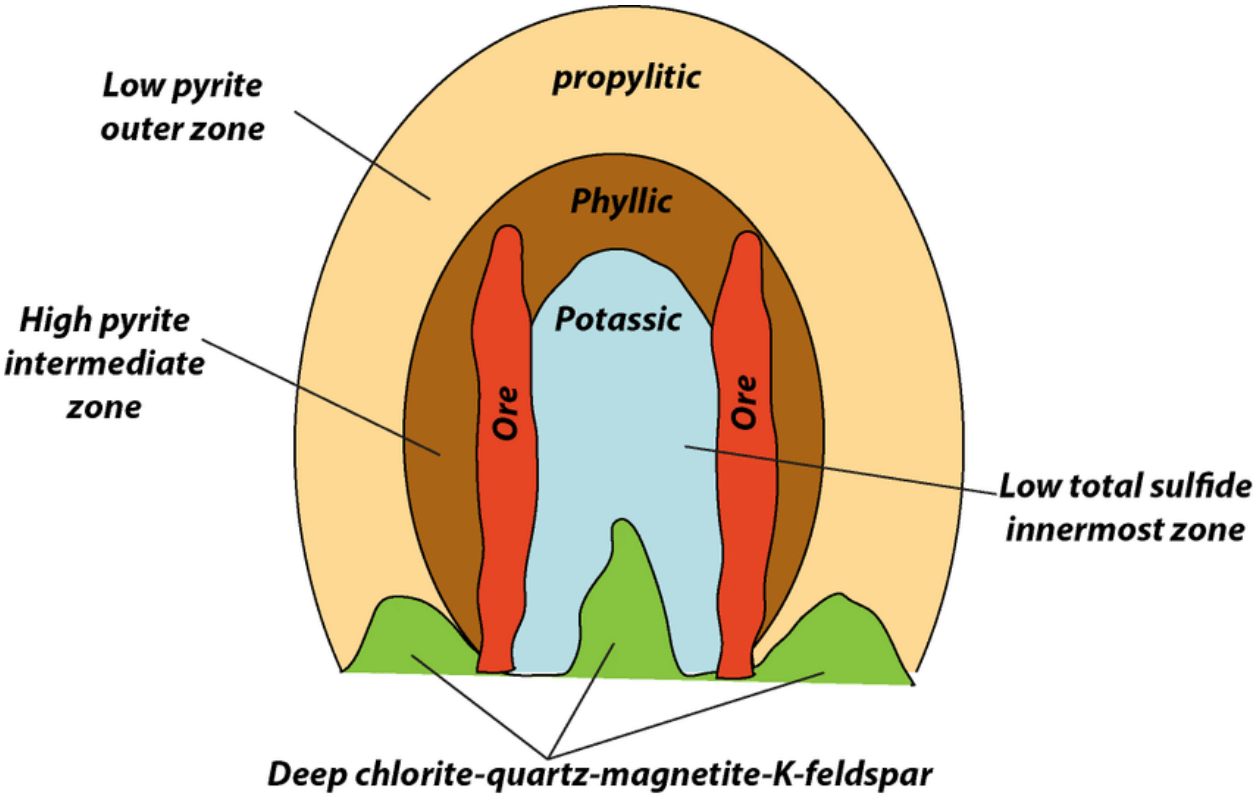


Figure 1. Cross-section of copper-porphyry deposit – prime example of non-stationary geo-domains (Belkacim et al, 2014).

1.3 PROJECT OBJECTIVES

This thesis focuses on using multinomial logistic regression to generate secondary data (soft data) and combining it with a non-stationary sequential indicator simulation to model heterogeneous geological domains at unsampled locations by using the information available at borehole data (hard data). The method proposed in this research needs soft data of geological domains at unsampled locations (target blocks). A multinomial logistic regression algorithm will be used to obtain the local probability (secondary information) of each geological domain at sample points and target grid nodes. In a nutshell, to model the heterogeneous geological domains at the target grid nodes stochastically, a non-stationary sequential indicator simulation paradigm will be used. An actual copper deposit will be used to test the proposed approach. Thus, this thesis aims to accomplish following objectives:

- Develop an algorithm for non-stationary sequential indicator simulation.
- Build the Machine Learning model based on **Multinomial Logistic Regression** and calculate the soft data, required for the latter steps. Check and assess the performance of the model on dataset of synthetic and real existing copper porphyry deposit.
- Evaluate and compare the findings with a traditional sequential indicator simulation.
- Validate the results using different statistical and geostatistical tools.

Besides on being just testes on several case studies, this research aims to produce actual copper grade inside of deposit. Thus, additional specific objective for the case study II is composed as:

- Estimate the copper grade inside of each geo-domains and produce final combined realizations of copper distribution along all three geo-domains by incorporating the simulation results.

1.4 PROJECT SIGNIFICANCE TO THE INDUSTRY

Resource estimation is the central pillar in the mining industry because all further mine development, investment amount, and perspective of the deposit project depend on the model of grades estimated entire the deposit and how reliable this model is. The mining business gains several advantages from the effective fulfilment of this thesis:

1) First of all, the developed algorithm allows an alternative method of combining modern technologies with conventional ones. It can be applied in commercial software programs as an enhanced alternative to the classic technique of SIS.

2) The second reason is that accurately built ore bodies will provide detailed information about the deposit, and minerals will be extracted more effectively than they could be. It can also decrease expenses on excavation at empty or poor-graded areas.

3) Additionally, the automatized machine learning algorithms can potentially helps to get rid of human-made errors and enhance the reliability. Also, ML can increase computational speed and save the time for unnecessary validation, because ML continuously learns at mistakes thereby preventing their repetition.

4) Finally, Multinomial Logistic Regression is a flexible tool as well as other Machine Learning algorithms, meaning that ML model could be modified for narrow task.

2 LITERATURE REVIEW

The ore properties must be thoroughly examined before starting the resource estimation procedure. Primarily these properties are associated with spatial variability and grade continuity. The foremost priority in the grade estimation of a mineral deposit is obviously geological modelling which has to proceed before resource estimation (Sinclair & Blackwell, 2002; Abzalov, 2016; Rossi & Deutsch, 2014). Accurately creating models of ore grades found in a deposit has a significant impact on the long-term planning of a mining operation (Maleki et al., 2021). Geological borehole database itself consist of two main types of variables: categorical or discrete – variable (lithology, mineralization zones, alteration, rock type) and continuous (mineral grade, ore grade). The process of resource estimation usually starts from identifying the target domains from the whole deposit area (Rossi & Deutsch, 2014; Emery & Séguret, 2020). In other words, the deposit is divided in certain zones – sub-domains, and then continuous variables of interest are modelled inside of each of these sub-domains by using of univariate and multivariate geostatistical tools. The main advantage of this method is that prediction over modelled continuous variables is not such complicated, since they are considered to be stationary and homogenous (Sinclair & Blackwell, 2002; Moon et al., 2006; Yunsel & Ersoy, 2011; Haldar, 2013; Rossi & Deutsch, 2014). Sets of non-overlapping geological domains produced as a result of splitting the mineral deposit into smaller parts are assumed to be stationary. Since geological domains are built based on the data from drill holes and examination by mining geologist, they own unique properties and cannot demonstrate totally homogeneous behavior (Dowd, 1986). The idea of assuming grade variability inside each geological domain as homogeneous proved to be unviable in practice because covariance and mean value of categorical variables are not constant throughout the target geological domain. Actually, identifying of estimation domains become very problematic. The challenge occurs because it tooks two stages, first estimation domains have to be distinguished from the borehole data and then modelled at the target points. Different approaches can be implemented for this purpose. The characterization of estimation geo-domains could be made by interpretation of core logging information (Soltani & Hezarkhani, 2011; Adeli & Emery, 2017). Therefore, the geological setting of the deposit plays a key role in determining the categorical variable of interest. To illustrate, when it comes to copper porphyry deposits, the estimation domains can be defined as mineralized zones (whether oxide or sulfide) or types of rock (Madani et al., 2021a). Similarly, lithology can be utilized as an estimation domain in iron deposits (Maleki et al., 2021; Hosseini et al., 2021).

Defining geo domains can also be achieved through grade domaining, which involves delineating regions of similar grade within a deposit (Emery & Ortiz, 2005; Yunsel & Ersoy, 2011; Iliyas & Madani, 2021). This approach allows for a more detailed estimation of the mineral resources within a deposit and can be a useful alternative to using mineralized zones or rock types as estimation domains. While the method of grade domaining is relatively straightforward, it is important to ensure that the resulting domains align with the geological interpretation of the sample points. This requires careful consideration of factors such as lithology, alteration, and structural controls on mineralization. By validating the grade domains against the geological logging data, a more accurate estimation of the mineral resources can be achieved. It is true that these methods, including defining estimation domains based on mineralized zones, rock types, or grade domaining, can be labor-intensive, time-consuming, and subject to subjective interpretation (Fouedjio et al., 2018). In particular, manual interpretation of geological data can be prone to errors and inconsistencies.

Consider the case of two geological domains with unique properties. The clear sign of stationary domains is the presence of “hard boundaries”, meaning that these domains have inimitable properties and ore content. Real cases show that usually, ore bodies have complicated patterns. Mixed zones and transitional zones are everyday things inside the ore body, so the statement of strict grade transition cannot be confirmed in practice. Since this concept does not take place in practical cases, geostatisticians introduced a variogram analysis designed to split the experimental area into two domains and serve as a "soft boundary". It is impossible to define geological units precisely in practice, and geological errors are inevitable, particularly when the geological units are intermingled in a complicated manner. Consequently, the potential uncertainty associated with resources and reserves may need to be assessed appropriately, and a lack of precision or accuracy in grade estimates may result (Stegman, 2001). Furthermore, the estimation cannot be repeated as a result of the subjective interpretation of data and setting the hard boundaries by mining geologists. Uncertainties associated with geological domains must be addressed because it leads to a loss of accuracy of the estimated resources (Stegman, 2001). However, other factors that affect the geo-domains are ignored within the conventional algorithms, and the whole unit is considered stationary. Further conducted investigations also followed this concept and proved efficiency by conducting geostatistical analyses (such as kriging and variogram analysis) and estimating resources inside each geo-domain (Duke and Hanna, 2001; Sarkar et al., 1990). This method allows to better examine the grade concentration inside each geo-domain. However, it fails to account for the uncertainty in the emplacement of the boundaries of the geological units.

To address these challenges, there has been a growing interest in developing automated and semi-automated methods for defining geo domains in mineral deposits. These methods often utilize machine learning algorithms and other computational techniques to process large volumes of geological data and identify patterns and relationships that may not be apparent to human interpreters. A various technique such as hierarchical clustering, K-means, Gaussian mixture and other can be utilized to perform this kind of task. The geo-domains that emerge as a result, exhibit a patchy and disorganized spatial arrangement. Unfortunately, these clusters are not practical for mining operations, as it is crucial to design contiguous, connected, and compact domains that facilitate efficient extraction.

While these approaches are still being refined and validated, they have the potential to greatly improve the efficiency and accuracy of mineral resource estimation. However, it is important to ensure that the results obtained through these methods are validated against existing geological data and expert knowledge to ensure their reliability. To address this problem, various clustering algorithms have been developed that consider the spatial interdependence of the data, enabling the creation of viable geo-domains that can be used for efficient mine planning and exploitation (Oliver & Webster, 1989; Ambroise et al., 1997; Scrucca, 2005; Romary et al., 2012; Romary et al., 2015; Fouedjio, 2016a; Fouedjio, 2016b; Fouedjio, 2017a; Fouedjio, 2017b; Fouedjio et al., 2018; Martin & Boisvert, 2018; D'Urso & Vitale, 2020). The resulting domains do not are compact and spatially connected, but at the same time they produce non-stationary geo-domains, that exhibit significant heterogeneity across the deposit. Therefore, implementation of advanced geostatistical interpretation tools required in order to produce geo-domains with proper configuration. The categorization of methods for this purpose is possible in two ways - deterministic and stochastic approaches. Deterministic methods can only anticipate a solitary geo-domain at untested locations, with no ability to quantify the uncertainty. On the other hand, the stochastic geostatistical techniques hold particular interest in this regard.

The most widely used method among others is Sequential Indicator Simulation (SIS), proposed by Journel and Alabert, 1987; Journel & Alabert, 1990. Nevertheless, SIS is not efficient in case of heterogeneous geo-domains, that exhibit large-scale geological characteristics. Consequently, conventional sequential indicator simulation may not be adequate for reproducing the desired compactness and spatial contiguity of geological features in modelling geo-domains. This is due to the fact that conventional sequential indicator simulation relies on the stationary property of the random function model and only utilizes the variogram as two-point statistics. In situations where geo-domains possess

complex characteristics, one option is to employ secondary information (Deutsch, 2006) to enhance the modelling process.

The present study introduces a novel methodology that integrates the multinomial logistic regression model with non-stationary sequential indicator simulation to generate secondary data for modelling heterogeneous geo-domains. The efficacy of this approach is demonstrated by applying it to a real copper deposit. The methodology involves employing a geo-clustering technique to characterize the geo-domains at sample points, followed by utilizing multinomial logistic regression to produce the local probability (secondary information) of each geo-domain at sample points and target grid nodes. Finally, a non-stationary sequential indicator simulation paradigm is utilized to stochastically model the heterogeneous geo-domains at target grid nodes. Therefore, this thesis works aims following:

- To present the theories of multinomial logistic regression, conventional and non-stationary sequential indicator simulation;
- To validate the proposed algorithm by applying it to a copper porphyry deposit;
- To compare the outcomes with those of conventional sequential indicator simulation and evaluate them using various criteria.

3 METHODOLOGY

3.1 HIERARCHICAL CASCADE SIMULATION

The soft information in the form of probability maps will be implemented in geological control in the resource estimation stage. Since the ultimate goal of this thesis work is to estimate ore grade by incorporating soft information obtained as a result of Logistic Regression prediction over the target points, the total copper grade has to be estimated inside of each geological domain by the probability of occurrence of the corresponding geological domain.

To estimate resources in copper deposits, the initial phase usually involves separately analyzing the variogram of copper grades in each geological domain. Even though the calculation of each sample variogram is based on a portion of the total available data (only those linked to the geological domain being studied), this method enables the capture of structural patterns specific to each lithotype, allowing the modelling of grade continuity in association with the deposit's lithology. An example of this is that the anisotropy may differ among lithotypes.

The next stage involves creating ore grade models pertaining to a specific geological domain. To achieve this, kriging (using both indicator and ordinary methods) can be applied to model the grade within the geological domain being considered, along with the corresponding grade variogram as input. The outcome is a series of grade models, which can then be combined with probability maps of each geological domain. This combination allows for defining the ultimate grade model for each location, as shown in the given expression (Emery, & Gonzalez, 2007):

$$Grade\ Estimate = \sum_{k=1}^{3} Probability\left(k^{th} domain\right)\ x\ Grade\ Estimate(k^{th} domain) \quad (1)$$

This equation describes the methodology called Hierarchical Cascade Technique in the resource estimation paradigm. The total copper grade obtained by this method will likely show better estimation results compared to the one that neglects the influence of geological domains.

The conventional way of assessing resources involves creating a model based on specific rock types and using kriging to estimate grades. This method assumes that the probability of a particular rock type being present is either 1 or 0, meaning there is no uncertainty. An example of this approach is illustrated in the figure, where the most probable rock type is used to define the boundaries. The new proposed methodology implies that grades

associated with a certain rock type at a specific location are adjusted using a probability function that considers the uncertainty of whether that rock type is present or not.

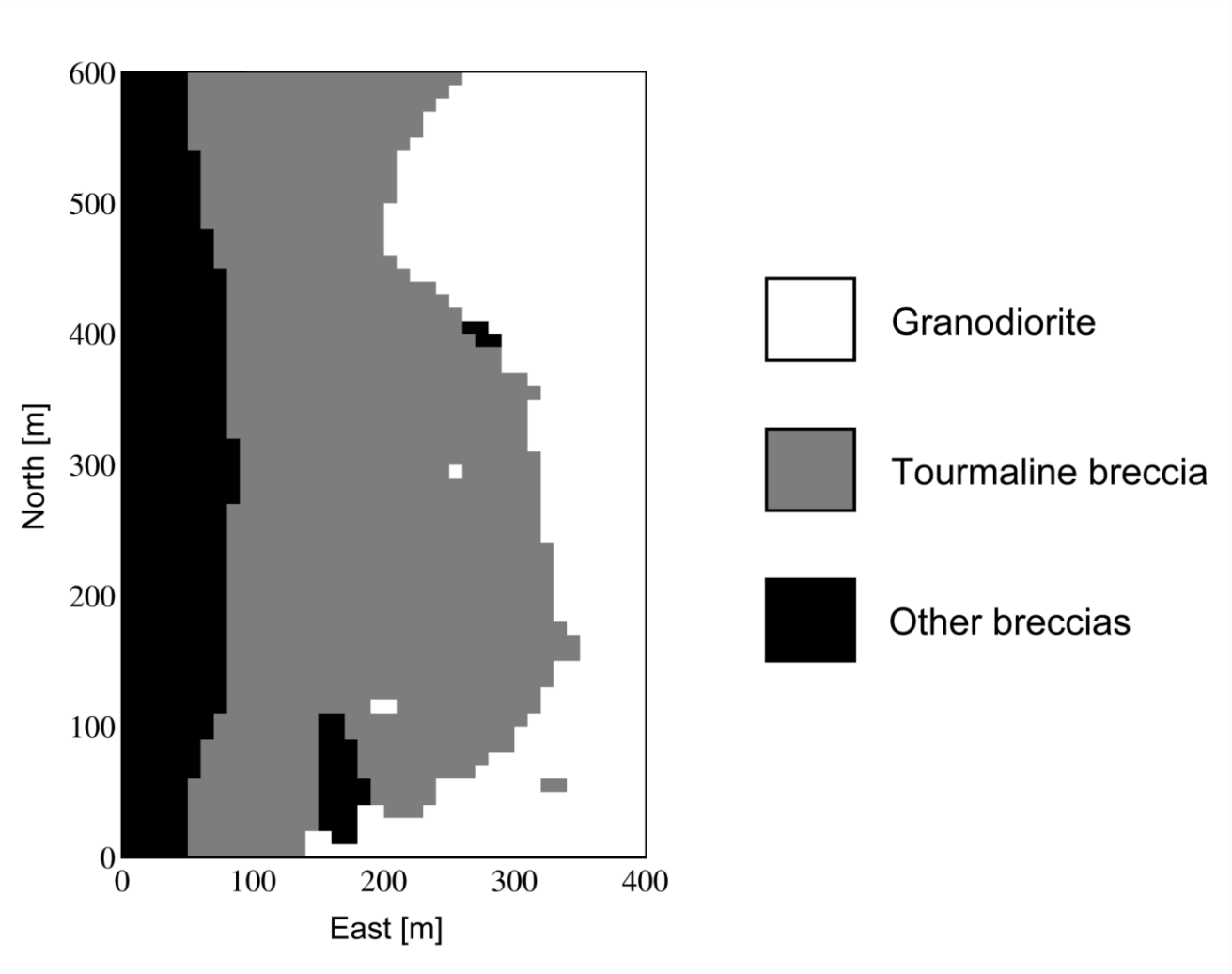


Figure 2. The deterministic lithotype modelling method involves selecting the most likely rock type at each location (Emery & Gonzalez, 2007).

However, in the proposed methodology, the estimated grades for a particular type of rock at a given location are adjusted based on a probability function that accounts for the uncertainty surrounding the presence or absence of that type of rock. Figure 3 presents a comparison between the map of estimated grades and the map generated by deterministic lithotype modelling.

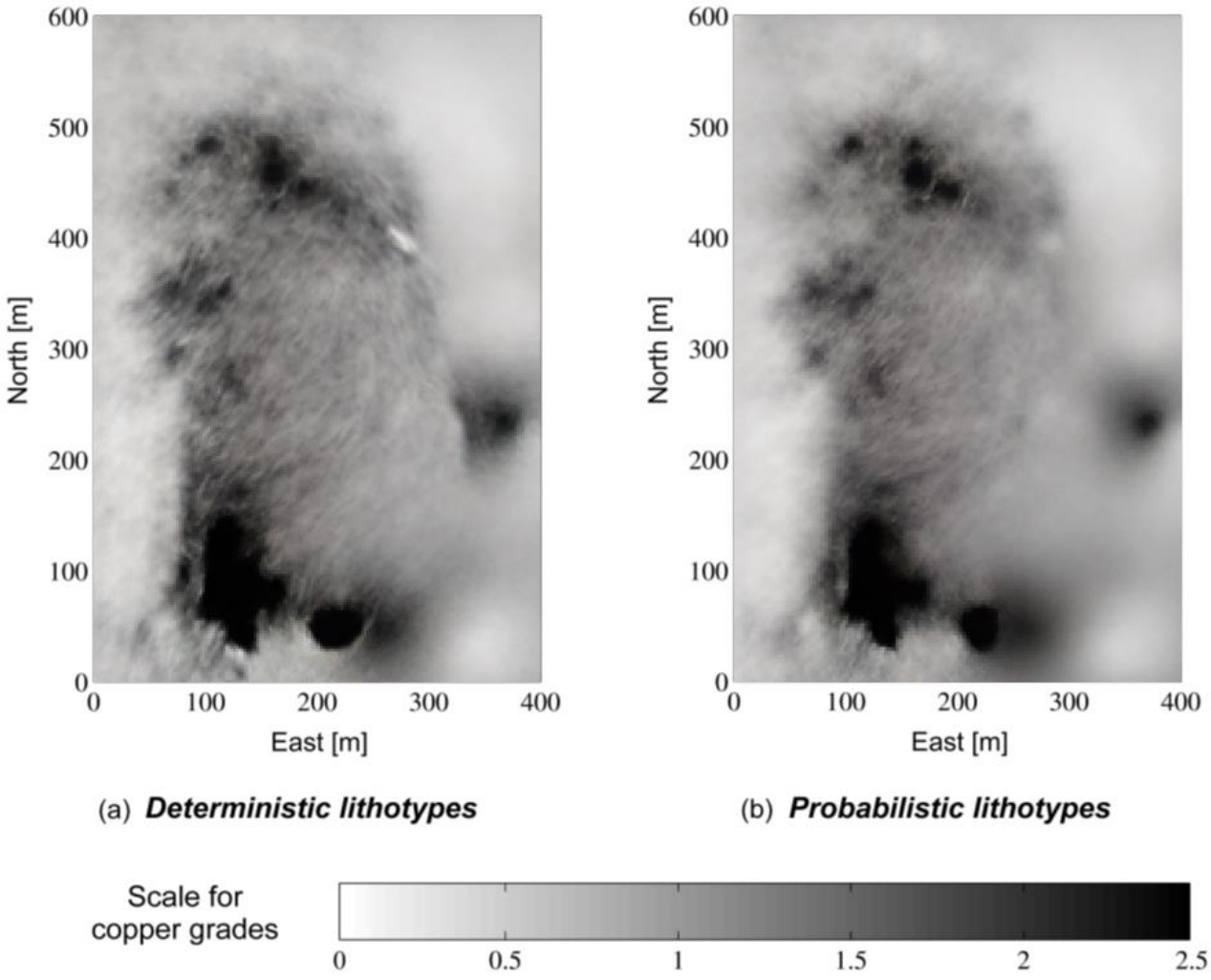


Figure 3. Illustration of the models of copper grade. Obtained with (a), by deterministic and (b), by probabilistic lithotype modelling (Emery & Gonzalez, 2007).

The one limitation of the Probabilistic modelling is producing artificial results because it assumes definite boundaries between domains, creating the appearance of a clear break in copper grade distribution that may not exist or may exist elsewhere. Additionally, deterministic lithotype modelling is heavily dependent on the mining geologist's interpretation and may result in significant variations in the grade model based on boundary contouring. In contrast, probabilistic geo-domain modelling is preferable as it avoids grade discontinuities in locations where the existence of geological boundaries is uncertain. This methodology incorporates two essential aspects of mineral resource evaluation (Emery & Gonzalez, 2007): 1) it considers the uncertainty of the ore body's geology through probability maps instead of a subjective interpretation of geological modelling, and 2) it accounts for the spatial continuity of copper grades within each geological unit using a specific variogram

model for each lithotype, allowing for a fair characterization of the copper grade distribution in space.

3.2 MULTINOMIAL LOGISTIC REGRESSION

Correlation between target and predictor variables can be determined by using statistical methods such as Regression Analysis. Regression Analysis is an advanced technique that can identify the variation of target variables in relation to the specific independent variables, while other predictors remain fixed (Pearson, 1930; Galton, 1984; Allen, 1997). This supervised learning algorithm is designed for forecasting, prediction, and time-series modelling. The graph of regression analysis represents a best-line that goes through all points in the dataset. Spacing between points and best-line shows the strengths of the captured relationship. The general classification of Regression includes the following techniques (Pearson, 1930):

- Linear Regression
- Logistic Regression
- Poisson's Regression

Several Machine Learning algorithms were compared with each other by various criteria as: speed of learning, number of required training samples, training speed, accuracy and flexibility – number of available parameters for tuning. Logistic Regression showed the best performance among other ML algorithms by almost all parameters. Obviously, this algorithm is the most suitable for the classification purposes. Therefore, Logistic Regression is the ultimate tool for providing an accurate probability measurement of each point along the entire area of interest. Figure 4 clearly illustrates the superiority of LR compared with other approaches.

The ambiguity of the problem is that usually the deposit contains several geological domains. However, classical Logistic Regression is limited to handling binary categorical variables, resulting in only two outcomes: zero/one or success/failure. The way to circumvent and make Logistic Regression applicable for modelling non-stationary geo-domains is to use modified version of Logistic Regression, so-called "**Multinomial Logistic Regression**". It is the extension of classical Logistic Regression, a supervised ML algorithm, designed to deal with multiclass classification problems (Long, 1997; Long & Freese, 2006).

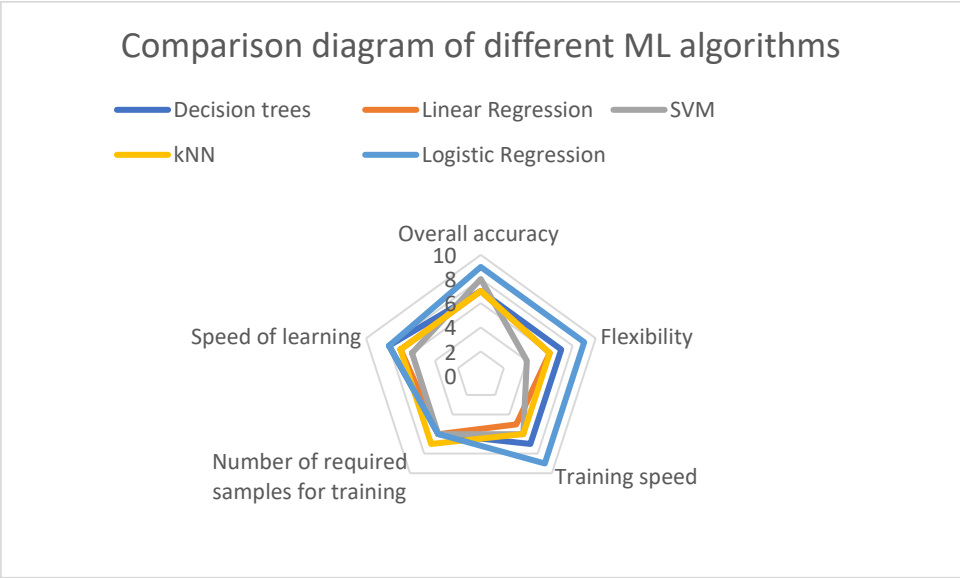


Figure 4. Comparison diagram of different ML algorithms.

In multinomial logistic regression, a categorical variable Y with different possible outcomes N is part of the sample points $\beta = 1 \ldots \tau$ (τ is the total number of observations). To calculate multinomial logistic regression for each geo-, a reference categorical variable N is chosen, and other variables are processed through the prediction process based on initial data.

$$\begin{cases} L_n \dfrac{\mu(Y_\beta = 1)}{\mu(Y_\beta = N)} = \rho_1 + \rho_{11}K_1 + \rho_{12}K_2 + \cdots + \rho_{1_p}K_p \\ L_n = \dfrac{\mu(Y_\beta = 2)}{\mu(Y_\beta = N)} = \rho_2 + \rho_{21}K_1 + \rho_{22}K_2 + \cdots + \rho_{2_p}K_p \end{cases}$$

…

$$L_n = \frac{\mu(Y_\beta = N-1)}{\mu(Y_\beta = N)} = \rho_{N-1} + \rho_{(N-1)1}K_1 + \rho_{(N-1)2}K_2 + \cdots + \rho_{(N-1)_p}K_p \qquad (2)$$

In the formula above $\mu(Y_\beta = 1)$ stands for the probability corresponding to the certain category, $\rho_{1,2,\ldots p}$ is number of independent variables. Using of maximum likelihood technique will allow to solve this equation and determine the regression coefficient - ρ_{N_p}. To obtain regression coefficient associated with the independent variables N – number of category and the ρ independent variables, all equations are solved together. Once the regression coefficients are estimated, they can be used to predict the value of the dependent variable based on the values of the independent variables.

Assuming the nominal dependent model and the condition that the final category has zero coefficients, the probability of being in each category at any given sample point β and the probability of belonging to the N category can be expressed by the formulas below. These

formulas take into account the values of the independent variables and the estimated regression coefficients associated with each category. The probability of falling into a particular category represents the likelihood of observing that specific outcome given the values of the independent variables.

$$\mu(Y_\beta = n) = \frac{e^{\rho^n + \sum_{l=1}^{p} \rho_{nl} K_l}}{1 + \sum_{n=1}^{N-1} e^{\rho_n + \sum_{l=1}^{p} \rho_{nl} K_l}}, n = 1, \dots N-1 \quad (3)$$

The similar formula works for the Nth category:

$$\mu(Y_\beta = N) = \frac{e^{\rho^n + \sum_{l=1}^{p} \rho_{nl} K_l}}{1 + \sum_{n=1}^{N-1} e^{\rho_n + \sum_{l=1}^{p} \rho_{nl} K_l}}, n = 1, \dots N-1 \quad (4)$$

It is also important to estimate the probability of presence inside of each category and inside of each target grid node. Following formula is designed to execute this estimation:

$$\mu(Y^* = n) = \frac{e^{\rho^n + \sum_{l=1}^{p} \rho_{nl} K_l}}{1 + \sum_{n=1}^{N-1} e^{\rho_n + \sum_{l=1}^{p} \rho_{nl} K_l}}, n = 1, \dots N-1 \quad (5)$$

The same principle also for the Nth category:

$$\mu(Y^* = N) = \frac{e^{\rho^n + \sum_{l=1}^{p} \rho_{nl} K_l}}{1 + \sum_{n=1}^{N-1} e^{\rho_n + \sum_{l=1}^{p} \rho_{nl} K_l}}, n = 1, \dots N-1 \quad (6)$$

3.3 SEQUENTIAL INDICATOR SIMULATION

The main tool used in this thesis work is a stochastic simulation method known as Sequential Indicator Simulation or SIS. Sequential Indicator Simulation (SIS) is aimed at providing a statistical framework for modelling geological systems, specifically SIS aimed to model a various category. Generally, SIS is an extension of sequential simulation techniques. The subject of interest in this research is an algorithm proposed by Alabert and Journel (Alabert, 1987; Journel & Alabert, 1990) called “sequential indicator simulation”. In this passage, the term "categories" refers to the geo-clusters or geo-domains that are created through unsupervised clustering techniques and are dependent on spatial factors. These clusters are completely separate from each other at all sampling points if they are identified in a deterministic manner. Once each sampling point is assigned to its corresponding geo-cluster in a deterministic way, the next step involves using conventional sequential indicator simulation to stochastically model the geo-clusters at unsampled locations, or target grid nodes. To do this, the geo-clusters (which act as hard conditioning data) are transformed into a matrix containing N columns of hard indicator data. The decomposition principle of

multivariate spatial distribution is a fundamental principle on which SIS relies. Decomposition creates a sequence of conditional distributions. The arbitrary order helps to avoid the production of artefacts. In this methodology, Indicator Kriging is utilized to estimate conditional distribution during the whole process at each step. A target category is estimated in each grid node using a random number in boundaries of 0 and 1. The SIS procedure continues till all target nodes are simulated. The following figure describes the SIS procedure in detail.

$$Ind(K;n) = \begin{cases} 1, when\ the\ geo-domain\ n\ prevails\ at\ K \\ 0, \qquad\qquad otherwise \end{cases} \quad n = 1, \dots N-1 \tag{7}$$

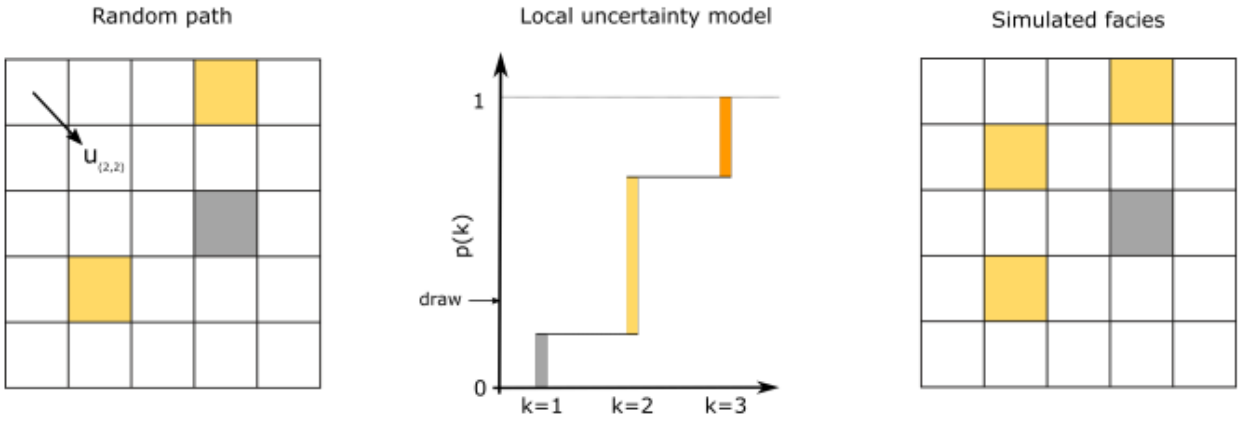


Figure 5. A synopsis of sequential indicator simulation (Mizuno & Deutsch, 2022).

Before starting the sequential simulation, conditional and random cumulative functions have to determine the following:

$$\begin{cases} Prob\,\{Z(x_1) < z_1 \mid (n)\} \\ Prob\,\{Z(x_2) < z_2 \mid (n+1)\} \\ Prob\,\{Z(x_3) < z_3 \mid (n+2)\} \\ \dots \\ Prob\,\{Z(x_N) < z_N \mid (n+N-1)\} \end{cases} \tag{8}$$

The main limitation of implementing sequential indicator methods in confirmed cases is the need for more knowledge of these functions. The new proposal by Journel and Alabert (1989) involved geostatistics in identifying unknown functions in spatial processes. Multi-Gaussian kriging is proposed for sequential Gaussian simulation (SGS) and Indicator Kriging for sequential indicator simulation (SIS). Originally, SIS was designed for the simulation of binary structures. SIS for multi-phase structures can be done in the following way:

1) Choose a random path that occupies all target nodes of the grid
2) Estimate the local probability of a point X by the following formula:

$$F(N) = \sum {}_{\alpha}\lambda_{\alpha} I_k(X_{\alpha}) \tag{9}$$

3) Pick a target grid node x_0. Estimate the probability of belonging the target grid node x_0 to the phases $(X_k, k = 1, \dots K)$:

$$[prob\{x_0 \in X_k\}]^* = [I_k, x_0)]^* \tag{10}$$

Implementing a single structural model (global multi - phase model, for instance) assures that the total probability of $[I_k,\ x_0)]^*$ equal to one. Otherwise, it requires to normalize the sum of probabilities or correct it by another way.

$$\sum\nolimits_{k=1}^{K} [prob\{x_0 \in X_k\}]^* = 1 \tag{11}$$

The estimation of probability in target grid node is given by:

$$[I_k(x_0)]^* = \frac{[I_k(x_0)]^*}{\sum_{k=1}^{K}[I_k(x_0)]^*} \tag{12}$$

4) Create an additional variable $[J_i(x_0)]$ – an aggregate sum of $[I_k(x_0)]^*$:

$$J_i(x_0) = \sum\nolimits_{j=1}^{i} [I_i(x_0)]^* \ where\ i = 1, K \tag{13}$$

A Monte Carlo simulation produces a random number with uniform distribution, which lies between zero and one. Figure 7 demonstrates all aspects of the Monte Carlo simulation. The final simulated value equals:

$$I_{S_i}(x_0) = \begin{cases} 1\ if\ J_{i-1}(x_0) < p \leq J_i(x_0) \\ 0, otherwise \end{cases} \tag{14}$$

5) The next simulating target grid node will use $I_{S_i}(x_0)$ as a conditioning value. The process will continue until all target grid nodes have been reached.

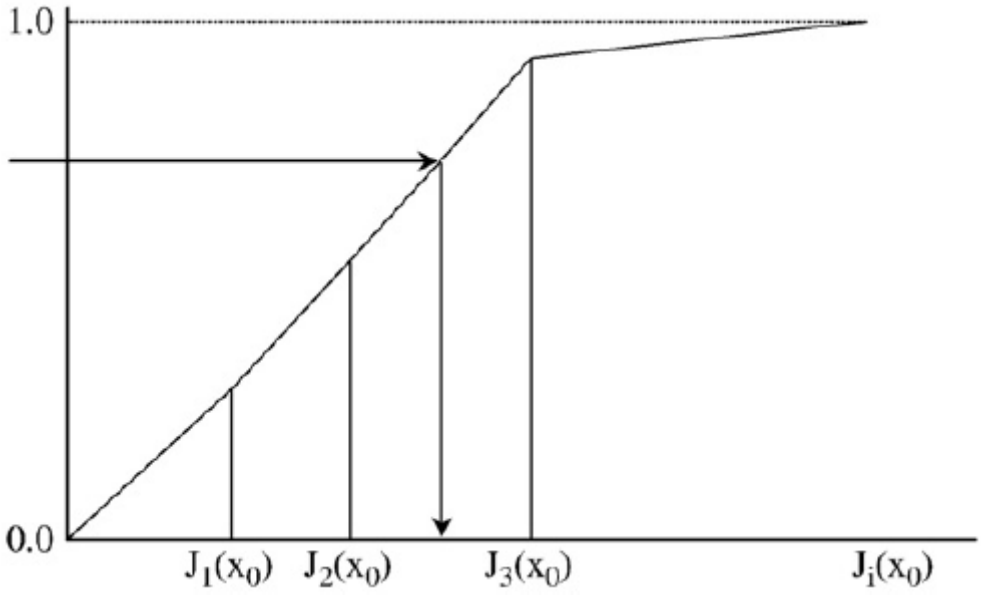


Figure 6. Illustration of Monte Carlo simulation for categorical variables using a pseudo-cumulative histogram. (de Almeida, 2010).

Different authors have mentioned the effectiveness of the SIS technique. The technique is versatile and straightforward. Also, SIS shows itself as a very flexible technique, especially in comparison with other stochastic simulation techniques. Nonetheless, some significant disadvantages have been highlighted, mainly due to the increase in conditioning data, which has a significant impact on $[I_k(x_0)]^*$estimation. The estimation becomes problematic and inaccurate with increasing the amount of conditioning data. The produced outputs will significantly differ from the theoretical continuity model. That is why it is necessary to control the neighboring conditioning samples set chosen during the simulation. To solve this obstacle, Journel (1989) proposed to select *n* neighboring samples of the target x0 randomly. This allows for extending the range of cover in neighboring samples. It is also challenging to reproduce the proportions of each phase, especially those with low proportions, which is one of the primary goals of the simulation.

Obviously, the main limitation initiated this thesis work is that SIS cannot work with non-stationary domains. When dealing with heterogeneous geo-domains, such as geo-clusters with large-scale geological features, conventional sequential indicator simulation may not be the most effective method. The produced realizations appear patchy and unstructured. The simulated categories can be observed all over the deposit (pretending homogeneity), making this method extremely unreliable. This is because conventional SIS struggles to accurately replicate the compact and contiguous geological features that are desired in geo-cluster modelling. The issue arises from the fact that conventional Sequential Indicator Simulation

relies on the stationary nature of random function models and only uses the variogram as a two-point statistic. To address this problem, utilizing secondary information (Deutsch, 2006) could be a viable option when dealing with such complex geo-cluster characteristics. To circumvent this problem, using soft data in the SIS algorithm can be of great help as this information instruct the algorithm to produce the non-stationary geological domains. Required soft data can be obtained from geophysical data and geological interpretation (Deutsch, 2006).

3.4 SEQUENTIAL INDICATOR SIMULATION USING LOCAL MEAN (SIS_LM); SISIM PROGRAM

This thesis work required using of additional enhanced computational program for producing final Sequential Indicator Simulation. This program is written for using by Matlab and was implemented for both simulation cases – without using the soft data and with incorporating the soft data results from the MLR. The program utilized for this thesis work is somehow similar to the BlockSIS program, proposed by Deutsch in 2006. The program requires the variogram parameters of indicators and residuals as input data. This variogram models are mainly built and fitted manually in Isatis.neo. Picture below shows the parameters of the program:

```
                  Parameters for SISIM
                  ********************

START OF PARAMETERS:
data.out                          % file with conditioning data
1 2 3                             %        columns for data coordinates
4                                 %        column(s) for data values
5 6 7                           %        columns for local mean of probability for each category
0.179604262 0.52438936 0.296006378      %        global proportion for each category
grid_2.out                      % file with coordinates of locations targeted for simulation
1 2 3                             %        columns for location coordinates
1 2 3                              %        columns for local mean of probability for each category
1 10 10 10                     %       gridded locations (1=yes, 0=no)? mesh size (0 0 0 if not gridded)
3                                 % number of categories
2 0.00605                              % Category 1:number of nested structures, nugget effect
1 0.10632   20  20  246   0 0 0 1       %        1st structure: it cc a1 a2 a3 angl ang2 ang3 b
1 0.04575   600  328  600 0 0 0 1       %        1st structure: it cc a1 a2 a3 angl ang2 ang3 b
2 0.00918                             % Category 2:number of nested structures, nugget effect
1 0.02562   20  20  246   0 0 0 1      %        1st structure: it cc a1 a2 a3 angl ang2 ang3 b
1 0.14247   600  328  600 0 0 0 1       %        2st structure: it cc a1 a2 a3 angl ang2 ang3 b
2 0.00220                            % Category 3:number of nested structures, nugget effect
1 0.05703   20  20  246   0 0 0 1     %        1st structure: it cc a1 a2 a3 angl ang2 ang3 b
1 0.09494  600  328  600 0 0 0 1     %        2st structure: it cc a1 a2 a3 angl ang2 ang3 b
600 600 600                         % neighborhood for original data: maximum search radii in the rotated system
0 0 0                             %                                 angles for search ellipsoid
0                                 %                                 divide into octants? 1=yes, 0=no
40                                 %                                  number of data per octant (if octant=1) or in total
600 600 600                         % neighborhood for simulated nodes: maximum search radii in the rotated system
0 0 0                             %                                   if scattered locations: angles for search ellipsoid
0                                 %                                                           divide into octants? 1=yes, 0=no
40                                 %                                    number of nodes per octant (if octant=1 and scattered) or in total
1                                 % kriging type: 1=SK, 2=OK
1                                 % SIS type: 1=traditional, 2=Bayesian Updating, 3=non-stationary with local mean probability
100                                 % number of realizations
3                                 % number of refinements (multiple grid simulation) (0=not used)
1                                 % random simulation sequence? (1=yes,0=regular simulation sequence)
9236548                           % seed for random number generation
sisim_TEST_trad.out                        % name of output file
0                                 % create a GSLIB header in the output file? 1=yes, 0=no

Available model types:
        1: spherical
        2: exponential
        3: gamma (parameter b > 0)
        4: stable (parameter b < 2)
        5: cubic
        6: Gaussian
        7: cardinal sine
```

Figure 7. SISIM Program parameters.

The first line contains the information of file with conditioning data. In this study, conditioning data contains spatial parameters of target grid node and category. Lines 2 and 3 stand for the number of geo-clusters and columns with coordinates specification. Next line is responsible for the value of global proportion of each geo-domain (category). This value shows how much certain category weights. The lines below are responsible for location of simulation output, variogram parameters including number of structures and nugget effect, type of kriging and other important parameters. The crucial line here is the line responsible for type of SIS. The program will initiate traditional simulation, if coded by 1 and non-stationary, if coded by 3. Actually the option standing by the number 3 – non-stationary SIS with local mean probability is the proposed simulation method itself.

3.5 PROPOSED SEQUENTIAL INDICATOR SIMULATION

The proposed technique being suggested for the simulation of spatiotemporal geospatial data is known as non-stationary sequential indicator simulation. Geostatistics is the field of study that deals with the modelling of spatial phenomena. Non-stationary sequential indicator simulation is a technique used in geostatistics for simulating geospatial data that exhibit changes in their statistical properties over time and space.

In simpler terms, this technique can model complex geospatial data with varying statistical properties such as trends, changes in variance, or other structures that are difficult to model using traditional geostatistical methods. The simulation process involves the use of various statistical models and techniques, including geostatistical models, time series models, or machine learning methods, to generate realistic geospatial data for research and testing purposes.

The main objective of non-stationary sequential indicator simulation is to create simulations that reflect the dynamic and complex nature of geospatial data in a realistic and meaningful manner. This technique provides a trustworthy method for modelling categorical data with varying properties. In this method, non-stationary simple kriging is combined with residuals derived from the mean probabilities that vary locally (Deutsch, 2006). Non-stationary sequential indicator simulation offers a reliable algorithm to model the categorical data with heterogeneous characteristics. This algorithm uses a non-stationary simple kriging with residuals from the locally varying mean probabilities. It calculates the weights in the same way as stationary simple kriging, with locally varying mean values at every location.

$$i^*_{LM}(u;k) = p_k(u) + \sum_{\alpha=1}^{n} \lambda_\alpha [I(u;z_k) - p_k(u_\alpha)] \tag{15}$$

To calculate the residuals $\mathbf{I}(\mathbf{u};\mathbf{z_k}) - \mathbf{p_k}(\mathbf{u_\alpha})$ in this formula, a regression function is used on the conditioning data points. To obtain these residuals, a regression function must first be fitted to the sample points, which can predict the values at the target location to find $\mathbf{p_k}(\mathbf{u})$. Since the spatial variation of each category depends on the geographic location of the sample points in heterogeneous geo-clusters, a regression function can be created using the coordinates as independent variables and the category as the dependent variable. However, linear regression is not suitable in this scenario due to the dependent variable being represented by integers. To analyze the residuals at the sample points, a variogram analysis should be conducted.

As $\mathbf{p_k(u)}$ and $\mathbf{p_k(u_\alpha)}$ in the equation represent estimated probabilities at the conditioning points and locally varying mean probabilities at the target grid nodes, a flexible regression function that can generate these local probabilities is required in the algorithm. Multinomial Logistic Regression (MLR) is proposed in this study to estimate these probabilities over the sample points and target grids. Once the probability $\mathbf{i^*_{LM}(u;k)}$ is estimated at the target location, the rest of the process is similar to traditional sequential indicator simulation. The proposed approach is referred to as "**SIS-LM**," while the traditional approach is referred to as "**SIS-Trad**."

WORKFLOW

The entire process of modelling non-stationary geological domains described in following steps:

- Firstly, define geo-domain or category in area of interest by using sample data locations.
- Identify main general trends of the categorical variables.
- Apply Multinomial Logistic Regression for each geo-domain. Derive estimated probabilities at the sample points.
- By using obtained probabilities and predicted values by MLR, determine the residuals and then use these residuals to deduce variogram models.
- Determine the locally varying mean probability across the target grid nodes for each geo-cluster to obtain the trend component by utilizing the fitted multinomial logistic regression models.
- Conduct the sequential indicator simulation using conventional and proposed sequential indicator simulation and produce realizations for comparison.

4 RESULTS CASE STUDY 1

4.1 OVERVIEW OF CASE STUDY

The efficacy of the proposed sequential indicator simulation technique utilizing local probability means was evaluated through a thorough analysis of a synthetic dataset. The synthetic map was generated by means of Plurigaussian simulation, as elaborated by Madani (2021), with anisotropy parameters that maximized continuity in the North direction (as illustrated in Figure 8. The resulting reference map exhibited a pronounced heterogeneity, with distinct geo-domains clearly demarcated by color coding. Specifically, geo-domain 1 (represented by the blue color) was observed on the left side of the grid, geo-domain 2 (denoted by the green color) occupied the central region, while geo-domain 3 (indicated by the red color) was predominantly located on the right-hand side.

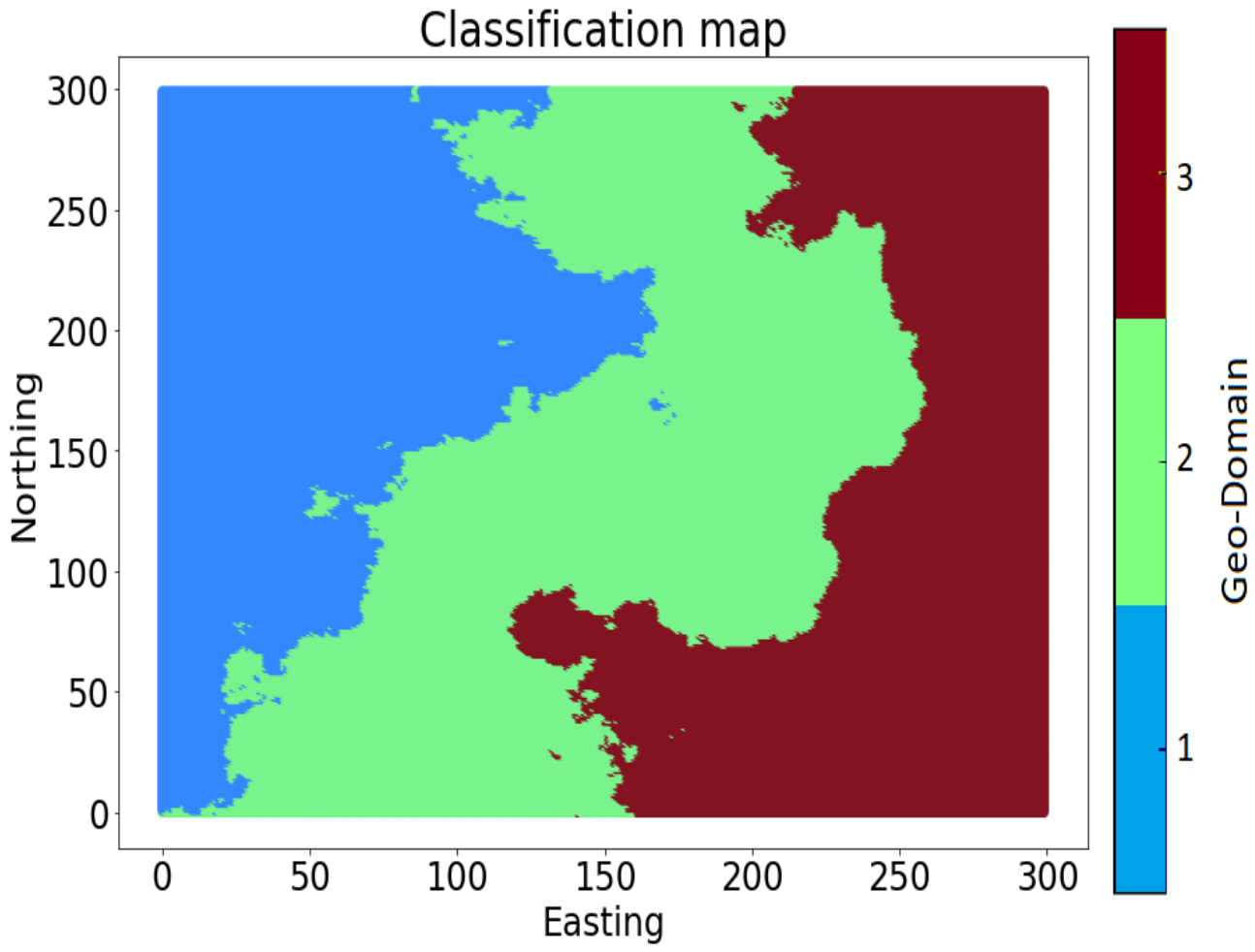


Figure 8. The reference map produced with illustration of main categories: blue – geo-domain 1, green – geo-domain 2, and red – geo-domain 3.

In this particular case study, a random sampling technique was employed to select 50 and 100 sample points from a dataset. These samples were then used as conditioning data for the simulation algorithms. When it comes to classification problems with more than two categories, the Multinomial Logistic Regression technique proves to be a suitable solution. The primary objective of this case study is to apply the Multinomial Logistic Regression technique on the selected data samples to create a predictive model. By utilizing this technique, it is expected that a high degree of accuracy can be achieved in the classification task. The main aims of this case study are:

- Build a ML model base on Multinomial Logistic Regression. Run the algorithm and obtain classification report to understand the performance rate.
- Retrieve probability values for the whole dataset and calculate residuals, which will used further as a soft data for Sequential Indicator Simulation.
- Run conventional SIS and proposed SIS.
- Compare results of both methods, choose the best and make conclusions.

4.2 EDA

A synthetic dataset contains 90,000 values with dimension coordinates and categorical variable responsible for the particular value's geo-domain. Before building a ML model, the proper dataset has to be prepared. 50 and 100 sample points were chosen randomly and divided in proportion of 20% test values and the rest 80% train values. The values of coordinates – X, Y, Z were chosen as a feature variable and geo-domain as a target variable. In simple terms, MLR predicts the target variable – geo-domain by using feature variable – the coordinates of sample points. The ML algorithm was built on the base of 100 and 50 values. Then this model was used to make prediction over the whole dataset consisting of 90,000 values. The procedure was repeated several times using different number random values to evaluate the quality of the ML model. The classification report indicates that the estimated accuracy of the MLR model is 82%. Efforts to improve accuracy through parameter tuning and changing the test/train ratio did not have a significant impact on the overall accuracy in either case.

The next step is the moderation of the predicted dataset. Initial Categorical values were converted into indicators. It is important to note that before performing the SIS-lm, the data must be coded into proper categories (0 or 1), and the residuals for each geo-domain must be calculated:

$$Residuals = Probability - Indicator \quad (16)$$

4.3 SEQUENTIAL INDICATOR SIMULATION

Sequential Indicator simulation for this study was performed in Matlab software. As mentioned before, this study aims to perform SIS twice with the conventional and proposed technique. SISIM program was implemented for both cases.

To provide unbiased results for this study, 100 realizations were generated using both the proposed SIS_LM - with local means and the traditional SIS_Trad. The target block grid dimension was set to 300m x 300m x 1m, forming in total 90,000 nodes, exactly the same as the reference map. As expected, the results from SIS_Trad were unstructured and patchy, both for 50 points and 100 points (Figures 9 and 10). The boundaries of geo-domains does not match with reference map, blocks appears in chaotic manner. Overall performance of SIS_Trad is very poor and unsatisfactorily.

At the same time SIS_LM shows an excellent result almost similar to the reference map. Results of Proposed Methodology devoid of shortcomings of conventional technique. Boundaries between geo-domains are solid, structured and corresponds to the reference map. To illustrate the results, realization number 20 was randomly selected for 50 points and 100 points (Figures 9 and 10).

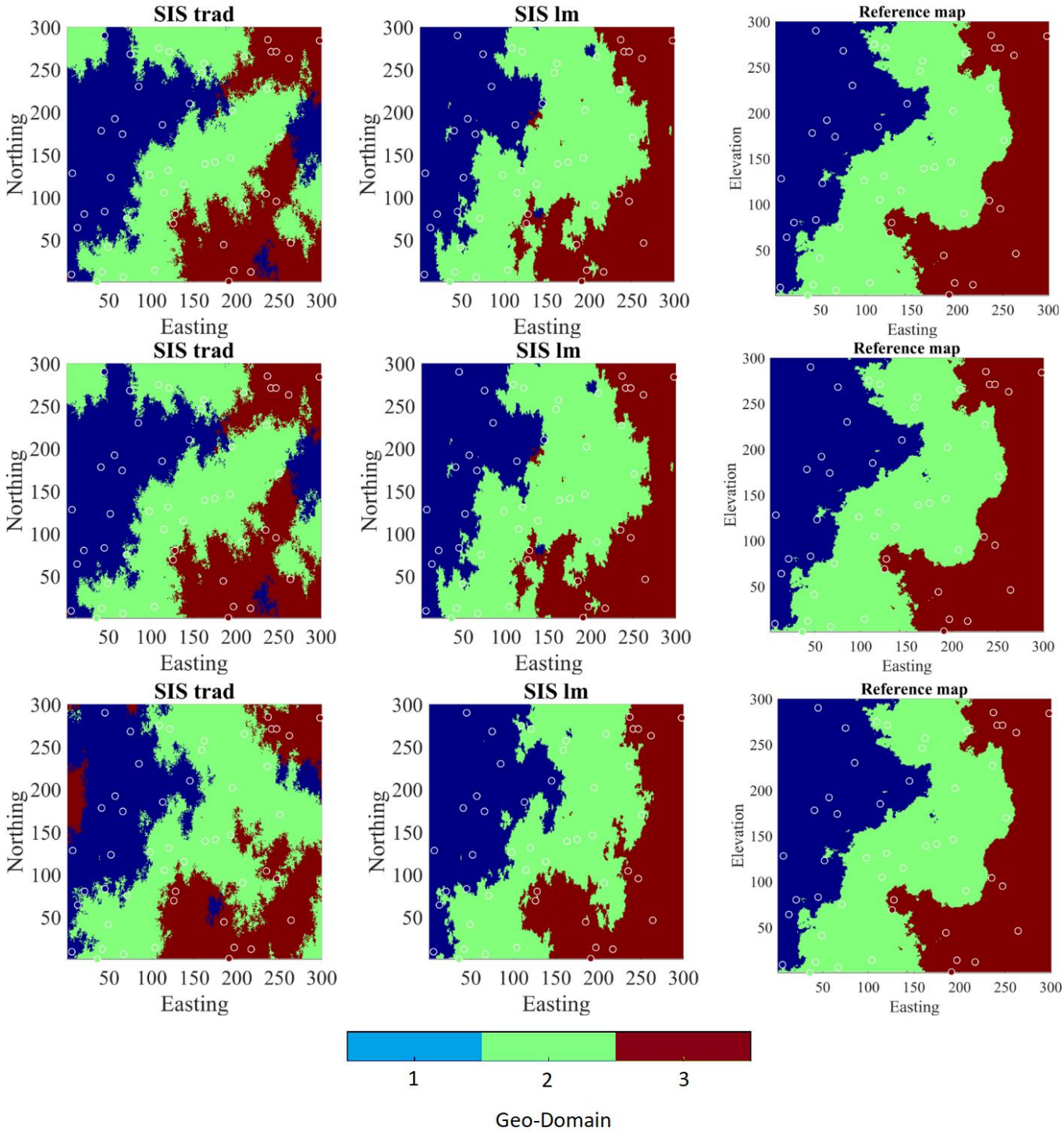


Figure 9. Comparison of realizations obtained by different techniques for 50 points; Left – Traditional SIS, Middle – Proposed SIS, Right – Reference map; blue: geo-domain 1, green: geo-domain 2, and red: geo-domain 3 (Amirzhan & Madani, 2022).

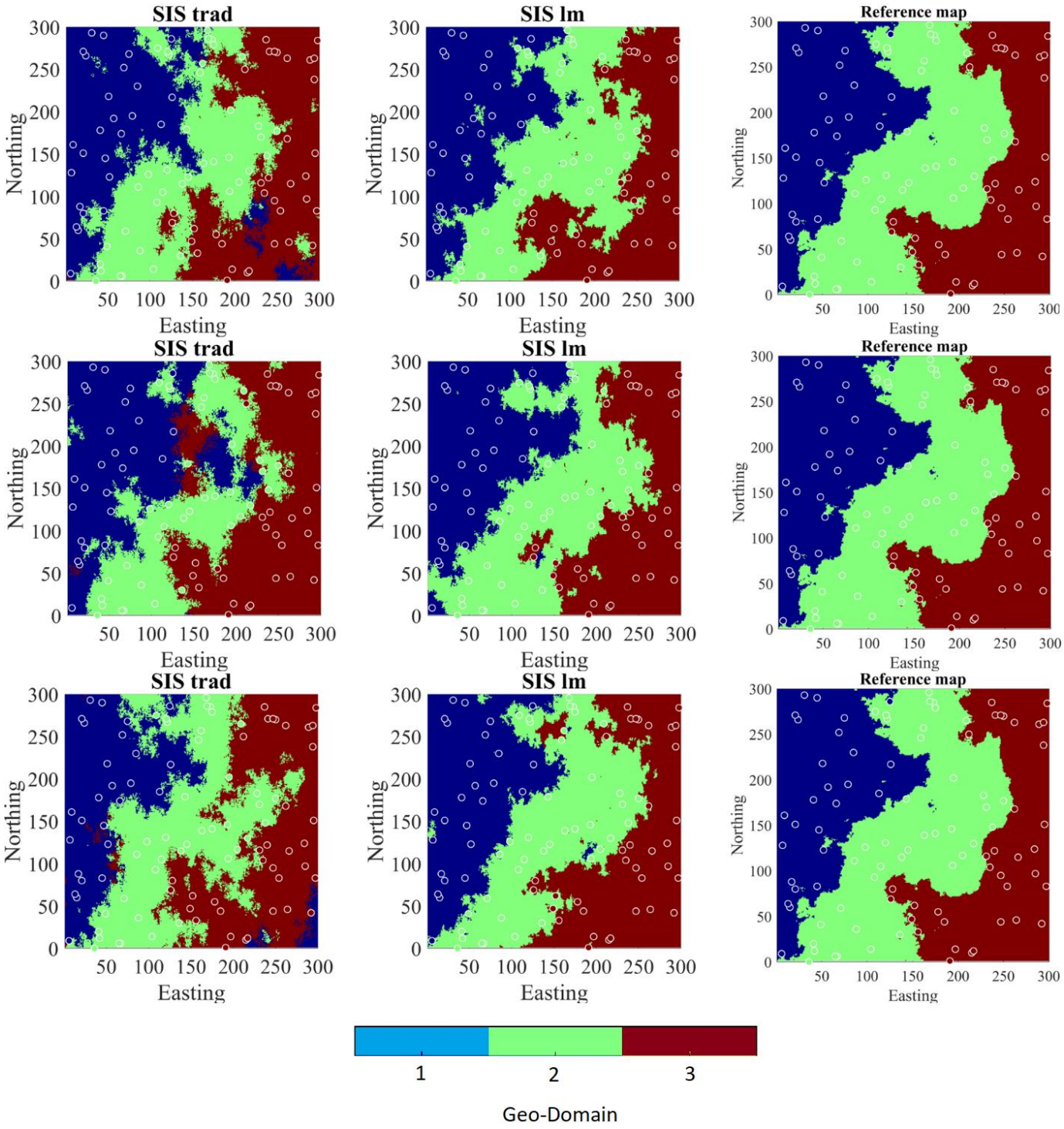


Figure 10. Comparison of realizations obtained by different techniques for 100 points; Left – Traditional SIS, Middle – Proposed SIS, Right – Reference map; blue: geo-domain 1, green: geo-domain 2, and red: geo-domain 3 (Amirzhan & Madani, 2022).

The significance of geological uncertainty in evaluating orebodies is emphasized, as it can impact the boundaries layout (Emery, 2007). This uncertainty can be represented by probabilistically modelling each categorical domain through conditional simulation. Probability maps are generated at a local scale to measure the uncertainty, which is calculated

by determining the frequency of each rock unit's occurrence for each block in 100 conditional realizations. These maps reveal the risk of encountering a mineralized zone that differs from others. The regions with minimal uncertainty are either those with a high probability of a specific rock unit, indicating a low risk of not finding it, or those with a very low probability, indicating a high degree of certainty of not finding it. On the other hand, other regions, depicted in light blue, green, or yellow, are more uncertain. Figures 11 and 12 demonstrates that the proposed approach has produced more robust certainty regarding the presence of categories, especially in areas where the conditioning data may be limited.

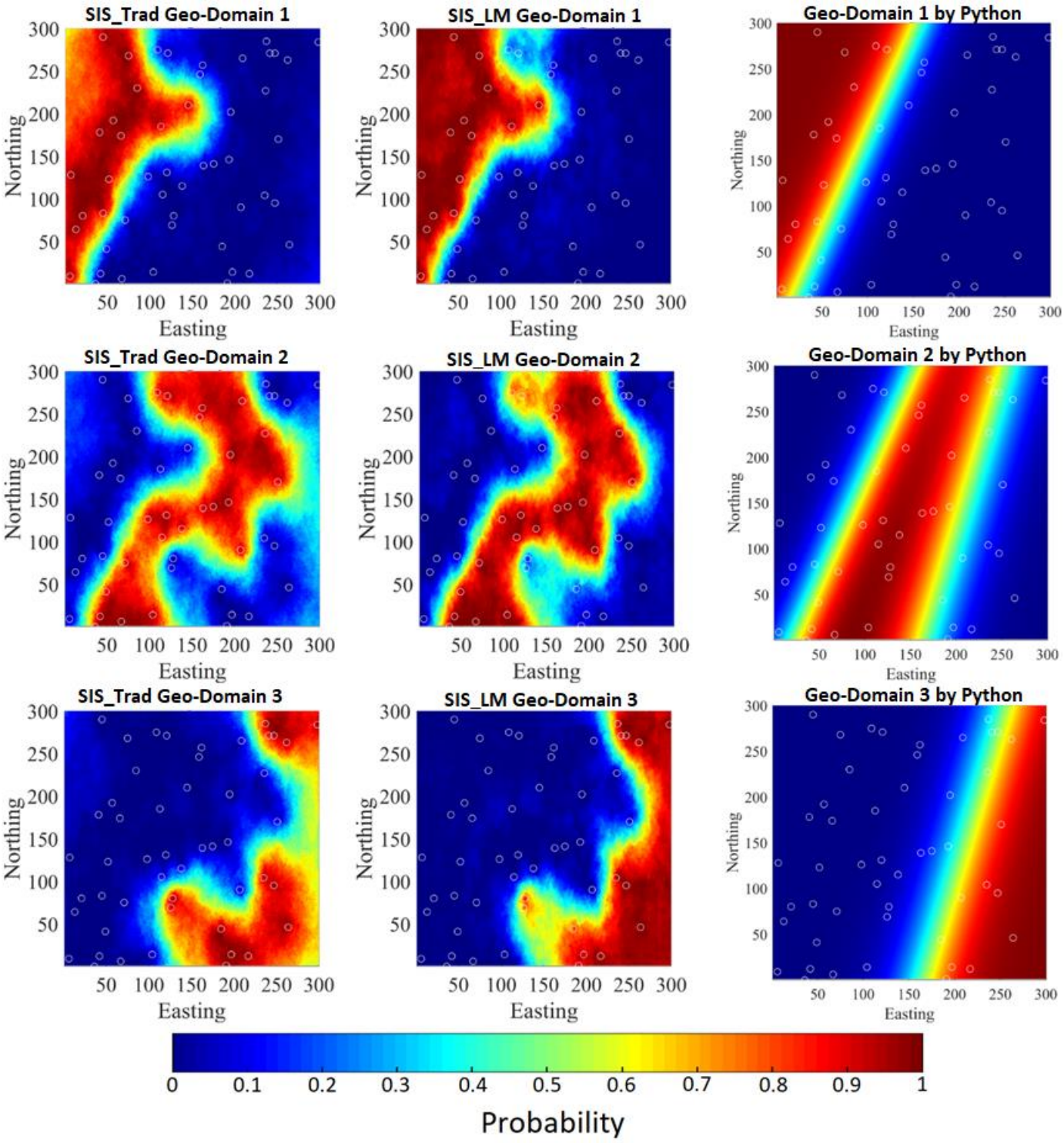


Figure 11. Comparison of probability maps of each geo-domain obtained by different techniques for 50 points: Top – Traditional SIS for geo-clusters 1, 2 and 3; Bottom – Proposed SIS for geo-clusters 1, 2 and 3 (Amirzhan & Madani, 2022).

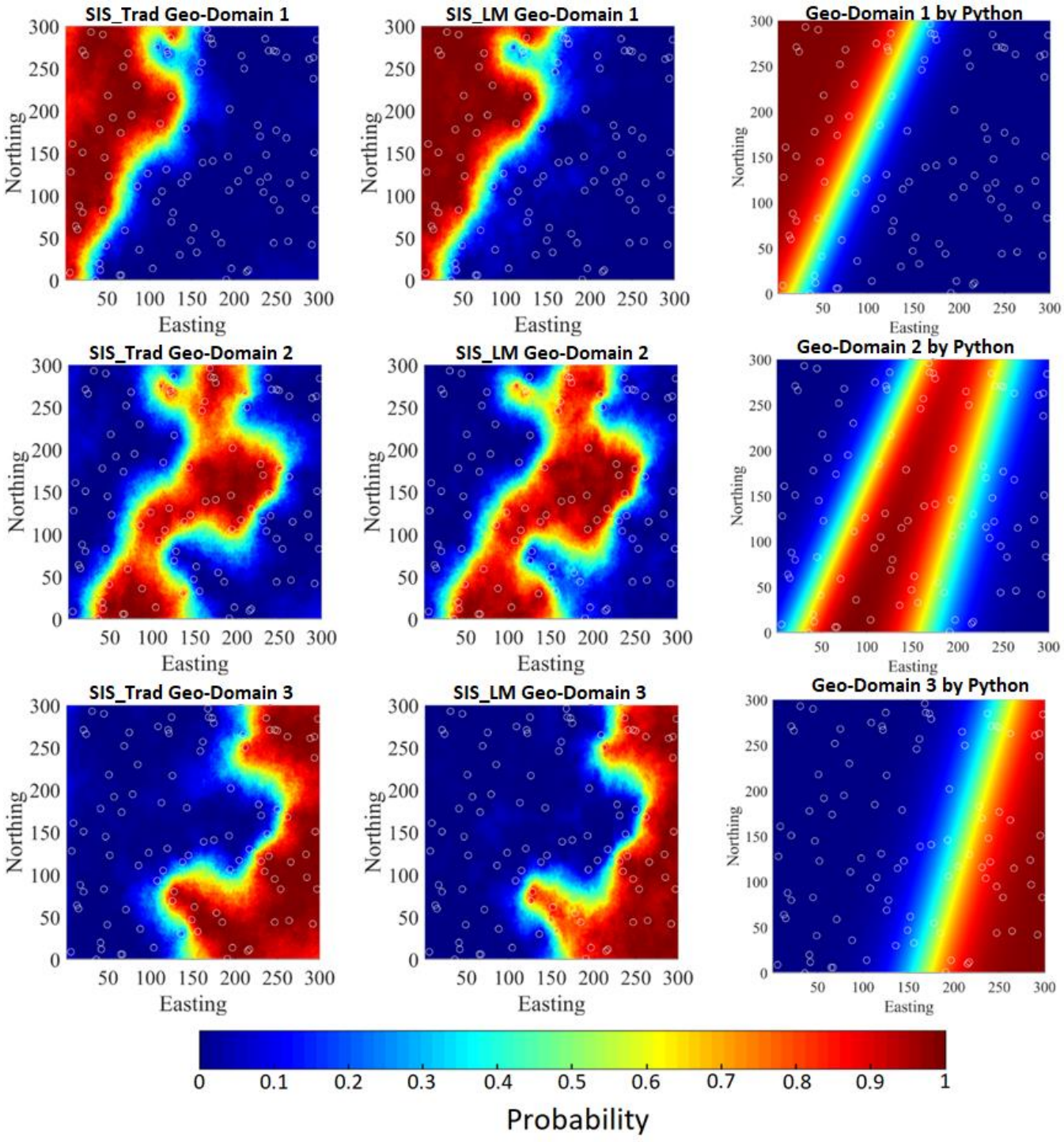


Figure 12. Comparison of probability maps of each geo-domain obtained by different techniques for 100 points; Top – Traditional SIS for geo-clusters 1, 2 and 3; Bottom – Proposed SIS for geo-clusters 1, 2 and 3 (Amirzhan & Madani, 2022).

To validate the results of the simulations, it is necessary to calculate the frequency of each geo-domain in the simulated data. This measure of global uncertainty provides a valuable way to compare the properties of the simulated data with the experimental data. Additionally, it shows how well each geo-domain is represented in the simulations. Table 1 presents the average global proportions of each geo-domain for the 100 realizations produced by the proposed approaches. Although there is a slight difference in the global proportions of each geo-domain between the two methods, both methods produce the global proportions accurately.

Table 1 - comparison of global proportions reproduced by Traditional SIS and Proposed SIS with an original proportion (Amirzhan & Madani, 2022).

	Geo-Domain 1	Geo-Domain 2	Geo-Domain 3
Original proportion (Reference map)	0.296	0.401	0.302
SIS_LM (50 points)	0.318	0.411	0.270
SIS_Trad (50 points)	0.297	0.422	0.279
SIS_LM (100 points)	0.324	0.373	0.301
SIS_Trad (100 points)	0.303	0.342	0.354

Another approach to assess uncertainty is to determine the relative error (RE) between the estimated proportions and the actual proportions. Table 2 compares the RE of traditional SIS and SIS with local mean for 50 and 100 points. It can be observed that the proposed methodology produced Relative Error significantly lower than conventional technique. Overall results prove effectiveness of proposed methodology.

Table 2 - comparison of relative errors evaluated by Traditional SIS and Proposed SIS with original proportion (Amirzhan & Madani, 2022).

	Geo-Domain 1	Geo-Domain 2	Geo-Domain 3	Sum of errors
SIS_LM (50 points)	**0.005**	**0.025**	**-0.106**	**-0.006**
SIS_Trad (50 points)	0.006	0.052	-0.076	-0.016

SIS_LM (100 points)	**0.096**	**-0.06**	**-0.004**	**0.024**
SIS_Trad (100 points)	0.024	-0.147	0.171	0.048

Another validation method used for this case study is comparison of Global Proportions for each method. Global Proportions of each geo-domains were calculated over the realizations for each case separately. Purpose of this manipulations is checking the reproducibility of Global Proportions over SIS_Trad and SIS_LM. In a nutshell, the Global Proportion histogram of each geo-domains were plotted according to values of these Global Proportions, calculated over all three geo-domains. After that original (reference) Global Proportion was superimposed on mean Global Proportion of each approach – SIS_Trad and SIS_LM.

$$Global\ Proportion\ of\ geo-domain = \frac{nuber\ of\ simulated\ blocks}{total\ number\ of\ blocks} \quad (17)$$

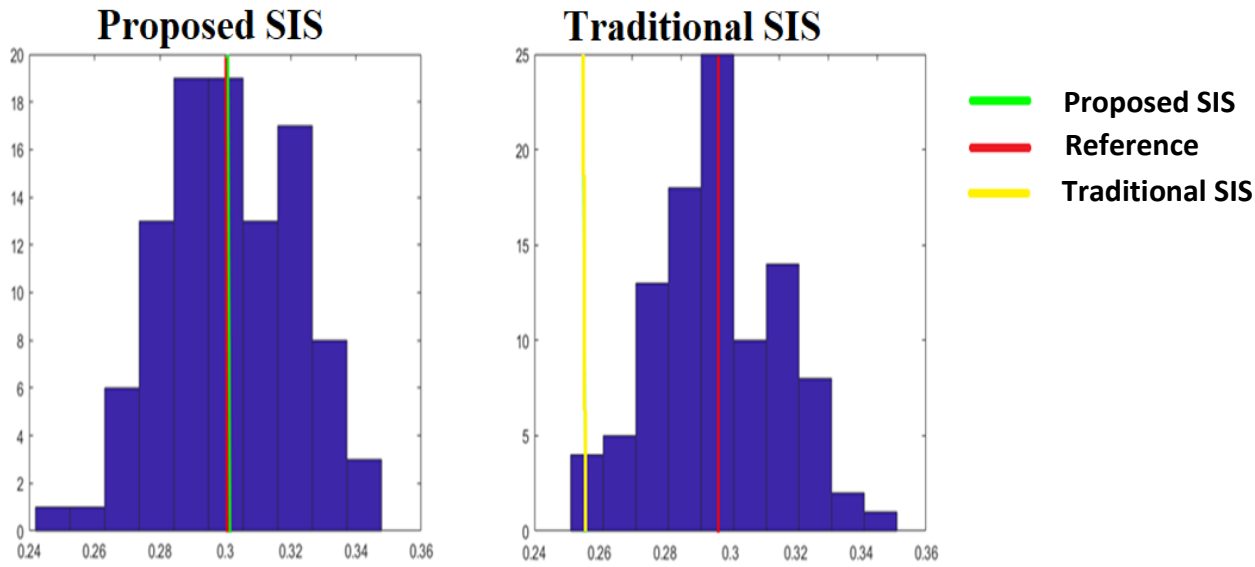


Figure 13. Comparison of global proportions of reference map, conventional SIS and proposed SIS (output from Matlab).

As it can be seen from the histograms, proposed SIS has the value of Global Proportion almost the same reference's. Traditional SIS failed in reproduction of Global Proportions, showing a huge deviation.

Summarizing all of the above, Proposed SIS showed the way better results than Traditional SIS. Realizations visually showed the complete incapacity of SIS_Trad in case of modelling

of non-stationary geo-domains. Two steps of validation again demonstrated the superiority of Proposed SIS under the Traditional SIS.

5 CASE STUDY 2

5.1 GEOLOGICAL SETTING

A real dataset copper-porphyry deposit was used as a source of information for executing the proposed algorithm. Detailed information of location, name and other geological parameters cannot be disclosed due to the privacy policy. The idea is to model the clustering variable (as it is identified as the estimation geo-domain in this deposit) and then model the copper grade inside of each geo-domain. In fact, this is the ultimate goal. However, before starting the modelling process, firstly need to do an exploratory data analysis for the target variable, copper.

5.2 EDA

The dataset pertains to a copper-porphyry drilling campaign with 67 boreholes arranged in a semi-regular pattern. The dataset of the case study shows a complete isotropic sampling pattern. This means that all variables of interest within the study area are identified and located at the exact sample coordinates. The primary continuous variables reported from borehole assaying were Cu (ppm), Mo (ppm) and four categorical variables (Clustering, Alteration, Rock Type, and Mineralization zones). To maintain confidentiality, the continuous variables were scaled, and local coordinates were mapped. The deposit is characterized by two categorical variables that consist of eleven mineralization zones, seventeen alteration types, and ten rock types.

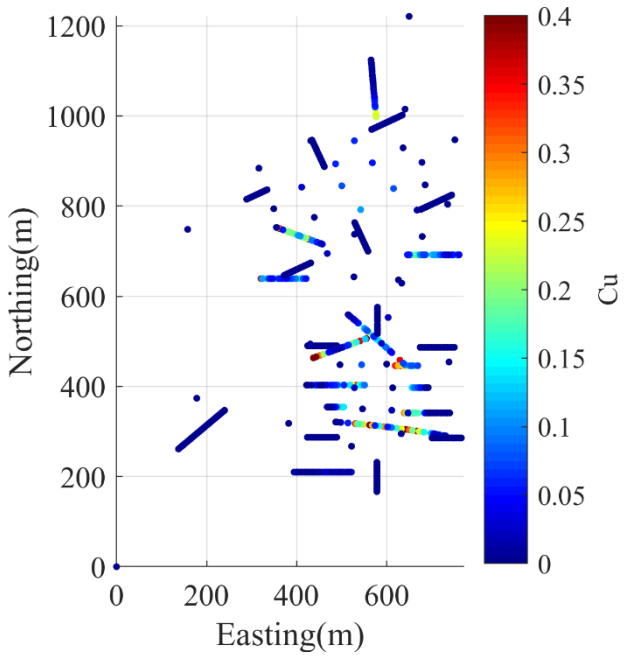


Figure 14. Visualization of sample points (boreholes) in planar view with Cu concentration.

The general overview of the dataset shows that it consists of three main geo-domains with high, medium and low coper concentrations. The mineralization zones recognized through core logging include UNK, hypogene (HYP), CLS, supergene hypogene (SUP-HYP), oxidized (OXI), supergene (SUP), leached-hypogene (LEA-HYP), oxidized-supergene(OXI-SUP), leached (LEA), leached-oxidized (LEA-OXI), leached supergene (LEA-SUP); The seventeen alteration types are phyllic (PHY), pyritic (PYR), propylitic (PRP), CAL, BLE, limestone (LIM), siliciclastic (SLC), NON, UNA, CLS, argillic (ARG), potassic (POT), HYD, sericitic (SER), and UNK, while the rock types include DAC, ALL, TON, LTT, tuffs (TUF), diorite (DIO-D), andesite (ANS), quartz diorite (QDI), CLS, diorite (DIO).

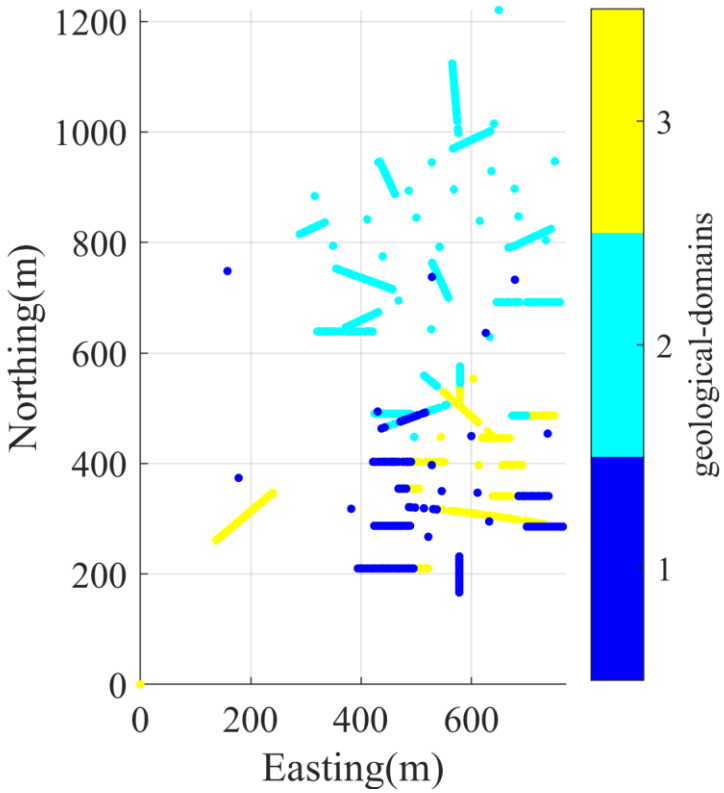


Figure 15. Visualization of geo-domains in planar view.

One of the initial steps in exploratory data analysis is to decrease the number of categories in Alteration, Rock Type, and Mineralization zones. This procedure is required since there is a numerous alterations and rock types. The aim is to decrease this number as much as possible and obtain 3-4 target categories, removing outliers and merge them. This procedure allows to calculate the associations, that will show the strength of connection between continuous and categorical variables. The boxplot is a best tool for performing this task since it visually shows the median value that helps to combine different alterations, rock types and mineralization zones. Whole procedure was done by using Isatis.neo.

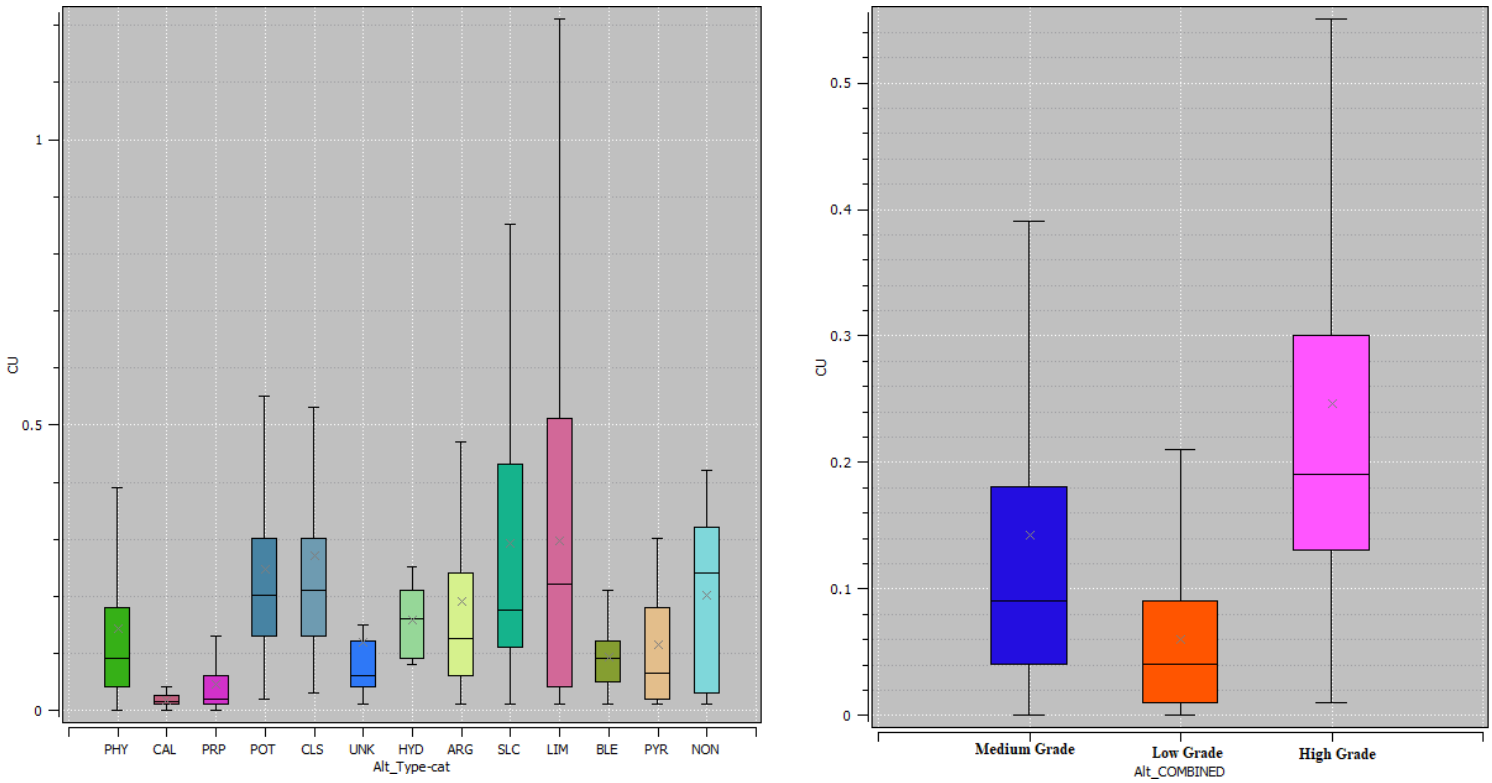


Figure 16. Initial (left) and newly formed Alterations (right).

The Figure 16 shows the variety of Alterations and final combined Alterations. The grouping of alterations was based on the principle of combining two alterations with a similar median value. Thus, 3 alterations were extracted from the variety of alterations for the convenience of further calculation of Associations.

Table 3 - The content of each newly formed Alterations.

Medium Grade	v1 == 'PHY' or v1 == 'PYR'
Low grade	v1 == 'PRP' or v1 == 'CAL' or v1 == 'BLE'
High Grade	v1 == 'LIM' or v1 == 'SLC' or v1 == 'POT' or v1 == 'NON' or v1 == 'UNA' or v1 == '_' or v1 == 'CLS' or v1 == 'ARG' or v1 == 'HYD' or v1 == 'SER' or v1 == 'UNK'

The same procedure was implemented for Rock Type and Mineralization Zones. The boxplots are plotted between Cu grade – continuous variable and Rock Type/Mineralization Zones – categorical variables. Boxplot shows the copper concentration inside of each target categorical variables and copper grade in combined Rock Type/Mineralization Zones

respectively. In a nutshell, different rock types and mineralization zones were combined according to their median values. Figures below stands for the explanation of procedure done.

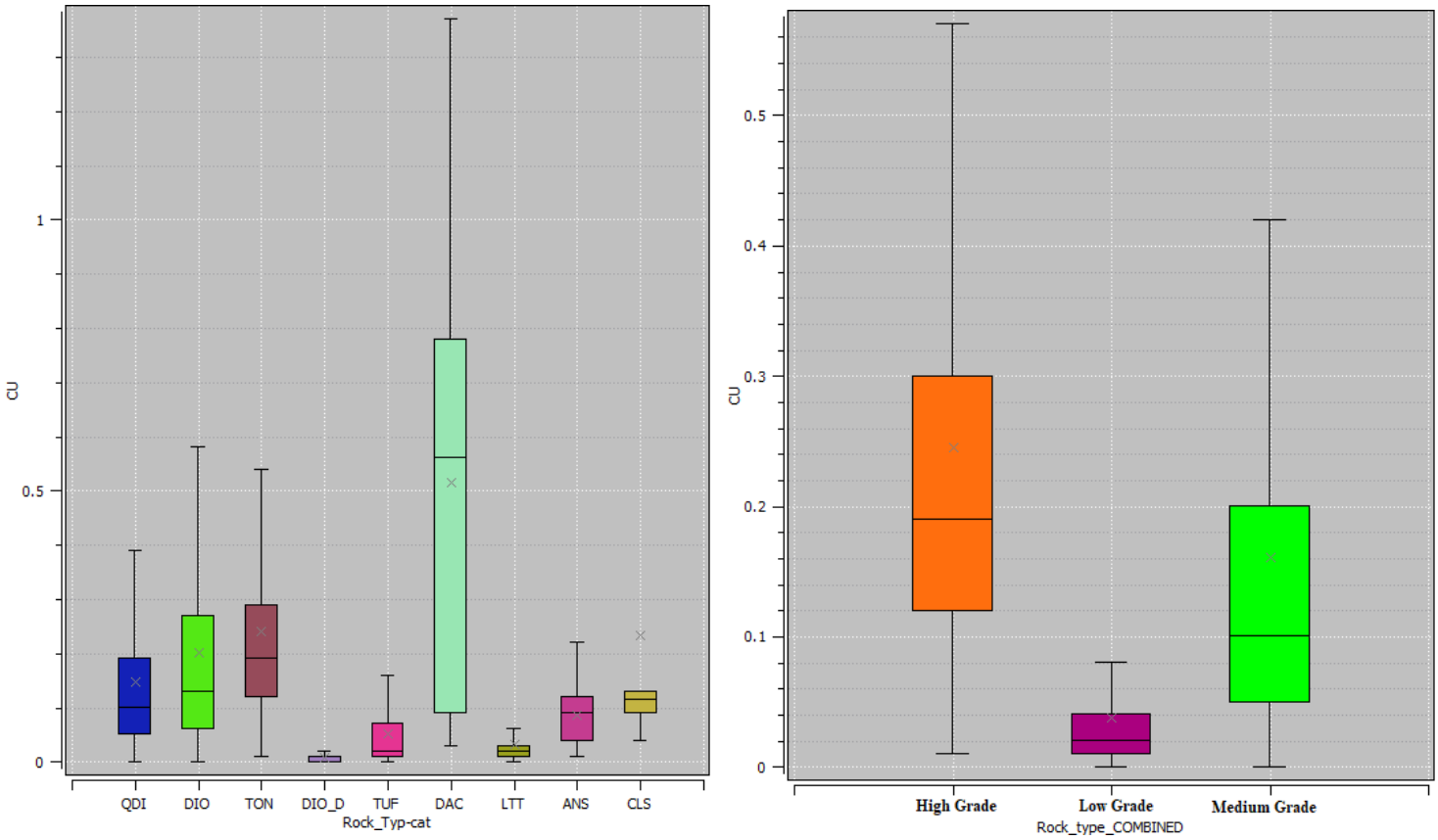


Figure 17. Initial (left) and combined Rock Type (right).

Table 4 - The content of each newly formed Rock Type.

1st	v1 == 'DAC' or v1 == 'ALL' or v1 == 'TON'
2nd	v1 == 'LTT' or v1 == 'TUF' or v1 == 'DIO_D'
3rd	v1 == 'ANS' or v1=='QDI' or v1 == 'CLS' or v1 == 'DIO'

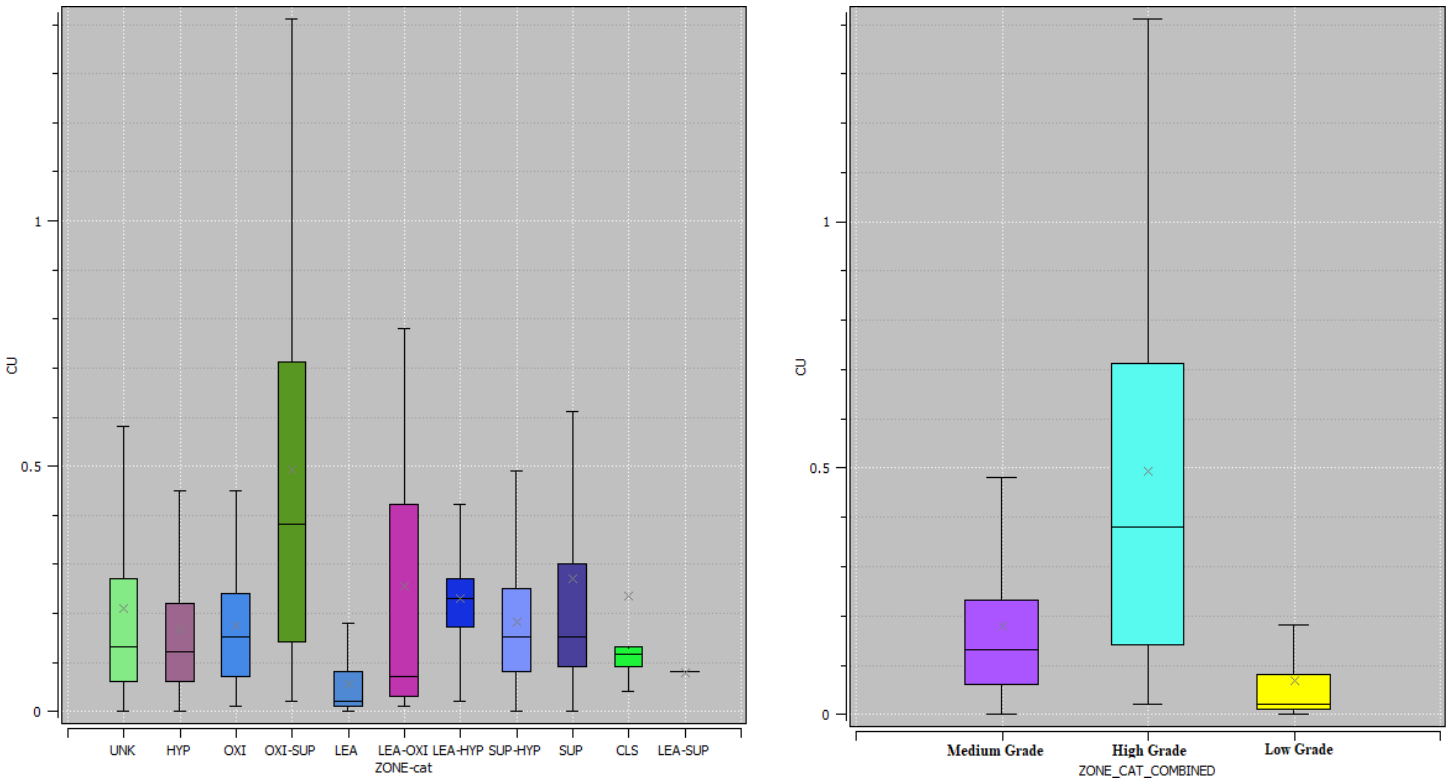


Figure 18. Initial (left) and combined Mineralization Zones (right).

Table 5 - The content of each newly formed Mineralization Zones.

1st	v1 == 'UNK' or v1 == 'HYP' or v1 == 'CLS' or v1 == 'SUP-HYP' or v1 == 'OXI' or v1 == 'SUP' or v1 == 'LEA-HYP'
2nd	v1 == 'OXI-SUP'
3rd	v1 == 'LEA' or v1 == 'LEA-OXI' or v1 == 'LEA-SUP'

Results shows that geo-clusters are in very strong associations with alteration. This means that the obtained domains (clustering) are less in alignment with rock type and zones.

The next step is to identify the association between:

1- Clustering & Rock Type

2- Clustering & Alteration

3- Clustering & Zones

4- Cu & Clustering

5- Cu & Rock Type

6- Cu & Alteration

7- Cu & Zones

Multivariate statistics tools were used due to the presence of several continuous and categorical variables. In resource estimation and priori data analysis, it is of interest to find the relationship between: 1- Continuous variables: e.g., between Fe & Mo 2- Categorical variables: e.g., between Alteration & Mineralization 3- Continuous and categorical variables: e.g., between Cu & Alteration. Clustering is the estimation domain of this deposit, then the goal is to find whether or not there are any associations between Clustering and other variables.

To examine the correlation between continuous variables, the correlation coefficient was utilized. However, this method is not practical for categorical data. Instead, Cramer's V coefficient is recommended as an alternative (Cramér, 2016).

$$V = \sqrt{\frac{\chi^2}{n(q-1)}} \tag{18}$$

In this equation χ^2 stands for the Chi-squared test statistic, derived from the contingency table, q is the minimal number of rows and columns in this table and n is the total samples amount.

This coefficient is a measure of the dependency or association between discrete variables and ranges from 0 (poor association) to 1 (perfect association). The absence of association is defined by $0 < V < 0.05$, then weak association range lies between $0.05 < V < 0.10$, the range of moderate association starts from 0.10 and lasts till 0.15. The rest two range is for strong association between $0.15 < V < 0.25$ and more than 0.25 till the 1 mean that here is very strong associations. In order to compute the coefficient, the Chi-squared test statistic derived from the contingency table, the total number of sample locations, and the number of rows and columns within the table must be utilized. In situations where the objective is to determine the level of association between categorical and continuous variables, the continuous data can be transformed into categorical data by utilizing quartile thresholds. This permits the application of Cramer's V coefficient to assess the potential strength of association between the converted continuous variable and other categorical variables. Table 6 furnishes data pertaining to the degree of interrelationships and associations.

Table 6 - Level of relationship between continuous-continuous variables (upper diagonal: Pearson linear correlation, and lower diagonal: Spearman non-linear correlation); categorical-

categorical variables (Cramer's V coefficient); and continuous-categorical variables (Cramer's V coefficient)

	Cu	Mo	Zones	Alteration	Rock type	Clustering
Cu	1	0.1097	VS	VS	VS	VS
Mo	0.2948	1	M	VS	VS	VS
Zones	VS	M	--	M	VS	W
Alteration	VS	VS	M	--	VS	VS
Rock type	VS	VS	VS	VS	--	S

W - Weak association; M - Moderate association; S - Strong association; VS - very strong association.

As can be seen, the correlation between Molybdenum and Copper is negligible (0.2948). Therefore, the presence of Molybdenum will be ignored and all further calculations will be made over the Copper. Mineralization Zones are associated very strongly rock type, and associated moderate with alteration. Alteration is in very strong associations with rock type and clustering. It also can be noticed that Copper is associated very strongly with zones, clustering, rock types and alteration while Molybdenum has moderate association with mineralization zones. In line with a common approach to modelling copper deposits, rock type appears to be a significant factor to consider when identifying estimation domains for modelling the continuous variables within the deposit. However, this method can overlook the impact of mineralization zones and alteration on the definition of the target estimation domains. To address this issue, machine learning algorithms can be employed to determine the domains by incorporating multiple variables. In this study, a clustering-based machine learning approach is used to identify the estimation domains, which takes into account not only rock types but also Cu, Mo, mineralization, alteration, and clustering. Determining ore grades is crucial because they directly affect the mining plan for this deposit.

5.3 GEOSTATISTICAL MODELLING OF GEO-CLUSTERS

The previous section's statistical analysis confirmed that the spatial variability of the resulting geo-clusters does not follow a stationary assumption. This leads to the use of geostatistical simulation methods that are designed for modelling such complex non-stationary geo-domains. The non-stationary sequential indicator simulation method proposed in this study

uses the residuals from a locally varying mean probabilities to stochastically model the entire geo-clusters of this copper deposit. The same block model previously identified for nearest neighborhood prediction is used, which consists of rectangular blocks with a mesh size of $10m \times 10m \times 10m$, resulting in a total of 709,800 blocks for the entire deposit.

After identifying the geo-clusters at sample points (boreholes), the next step is to fit multinomial logistic regression model over these points for each geo-domain and obtain the coefficients. The prediction will be made over the target variable - geo-domain by using feature variable - geographical coordinates (X, Y, Z). The aim is to create multinomial logistic regression models to predict the geo-clusters as a function of geographical coordinates. The procedure entirely repeats steps from case study I. Randomly chosen 50 and 100 points will be implemented for building ML model and then this model will make prediction over the whole dataset. Test/Train ratio also remains the same 20% and 80% respectively. The ScikitLearn library was selected as the tool to perform Multinomial Logistic Regression.

```
print(classification_report(y_pred, Y))
```

	precision	recall	f1-score	support
1	0.62	0.67	0.65	2286
2	0.95	0.95	0.95	7278
3	0.88	0.84	0.86	4233
accuracy			0.87	13797
macro avg	0.82	0.82	0.82	13797
weighted avg	0.87	0.87	0.87	13797

```
[45] clf.score(X, Y)
```

```
0.8711314053779807
```

Figure 19. Classification report of MLR model produced by Python, indicqating major properties.

Like in previous case, the fitted model resulted a very high accuracy rate – 87%. Relying on the results of classification report, it can be concluded that calculations are reliable. Next step is to estimate the probability of each geo-cluster at sample points. The probability of each sample point was measured over all three domains, in order to increase the accuracy and ensure that final resulting probability is correctly filtered out of other two values. Since the probability usually varies from 0 to 1, increasing the likelihood of event occurrence with numerical value, the MLR model choose value that is closest to 1.

```
points.head(50)
```

	X	Y	Z	Category	Ind1	Ind2	Ind3	predictions	1_proba	2_proba	3_proba	proba_final
0	615.0	839.0	2478.79	2	0	1	0	2	0.003185	0.993034	0.003781	0.993034
1	615.0	839.0	2477.79	2	0	1	0	2	0.003170	0.993053	0.003777	0.993053
2	615.0	839.0	2476.79	2	0	1	0	2	0.003154	0.993072	0.003774	0.993072
3	615.0	839.0	2475.79	2	0	1	0	2	0.003138	0.993092	0.003770	0.993092
4	615.0	839.0	2474.79	2	0	1	0	2	0.003123	0.993111	0.003767	0.993111
5	615.0	839.0	2473.79	2	0	1	0	2	0.003107	0.993130	0.003763	0.993130

Figure 20. Prediction made by MLR over the random 50 points.

The block model (grid) was created by using Isatis.neo software and exported in csv file, which contains only a dimensional values (X, Y, Z coordinates) without a categorical variable, indicating geo-domain. The Machine Learning model was implemented on entire block, i.e., prediction was made over the 709,800 sample points. Soft information was derived by the same principle as for borehole values, the probability of each grid point was estimated and the values **closest to 1** was chosen as a total probability value.

```
grid
```

	X	Y	Z	predictions	1_proba	2_proba	3_proba	predict	proba_final
0	157.87	210	1910.73	1	9.868499e-01	0.006815	6.335452e-03	1	0.986850
1	167.87	210	1910.73	1	9.853078e-01	0.007055	7.637384e-03	1	0.985308
2	177.87	210	1910.73	1	9.834943e-01	0.007301	9.204300e-03	1	0.983494
3	187.87	210	1910.73	1	9.813568e-01	0.007554	1.108899e-02	1	0.981357
4	197.87	210	1910.73	1	9.788332e-01	0.007813	1.335427e-02	1	0.978833
...	...	...	...	...	...	...	...	...	...
709795	887.87	1500	2600.73	2	8.193039e-09	1.000000	3.145108e-07	2	1.000000
709796	897.87	1500	2600.73	2	7.901739e-09	1.000000	3.662347e-07	2	1.000000
709797	907.87	1500	2600.73	2	7.620796e-09	1.000000	4.264650e-07	2	1.000000
709798	917.87	1500	2600.73	2	7.349842e-09	0.999999	4.966006e-07	2	0.999999
709799	927.87	1500	2600.73	2	7.088522e-09	0.999999	5.782706e-07	2	0.999999

709800 rows × 9 columns

Figure 21. Prediction made over the entire block model by MLR.

Using the estimated probabilities at sample points, one can calculate the residuals:

$$Ind(K_\beta; n) - \mu(Y_\beta = n) \tag{19}$$

It can be done by subtracting the geo-cluster indicators $Ind(K_\beta; n)$ from the estimated probabilities - $\mu(Y_\beta = n)$. To apply the proposed algorithm, the input (hard conditioning data) should be the sought residuals, thus requiring a variogram analysis. The anisotropy of each

geo-cluster in the region was quantified, and two directions of anisotropy in the horizontal and vertical directions were identified.

The initial step in resource estimation is to conduct a variogram analysis of the copper grades for each lithotype. Even though the sample variogram calculation only employs a fraction of the available data (those pertaining to the lithotype being examined), this approach allows one to capture the appropriate structural patterns for each lithotype and model the grade continuity based on the deposit's lithology. For example, it is evident that anisotropy varies across lithotypes. Spherical variogram models were fitted to the experimental variograms of the residuals and indicators, taking into account proper nugget effect and maximum and minimum continuities along the vertical and horizontal directions, respectively.

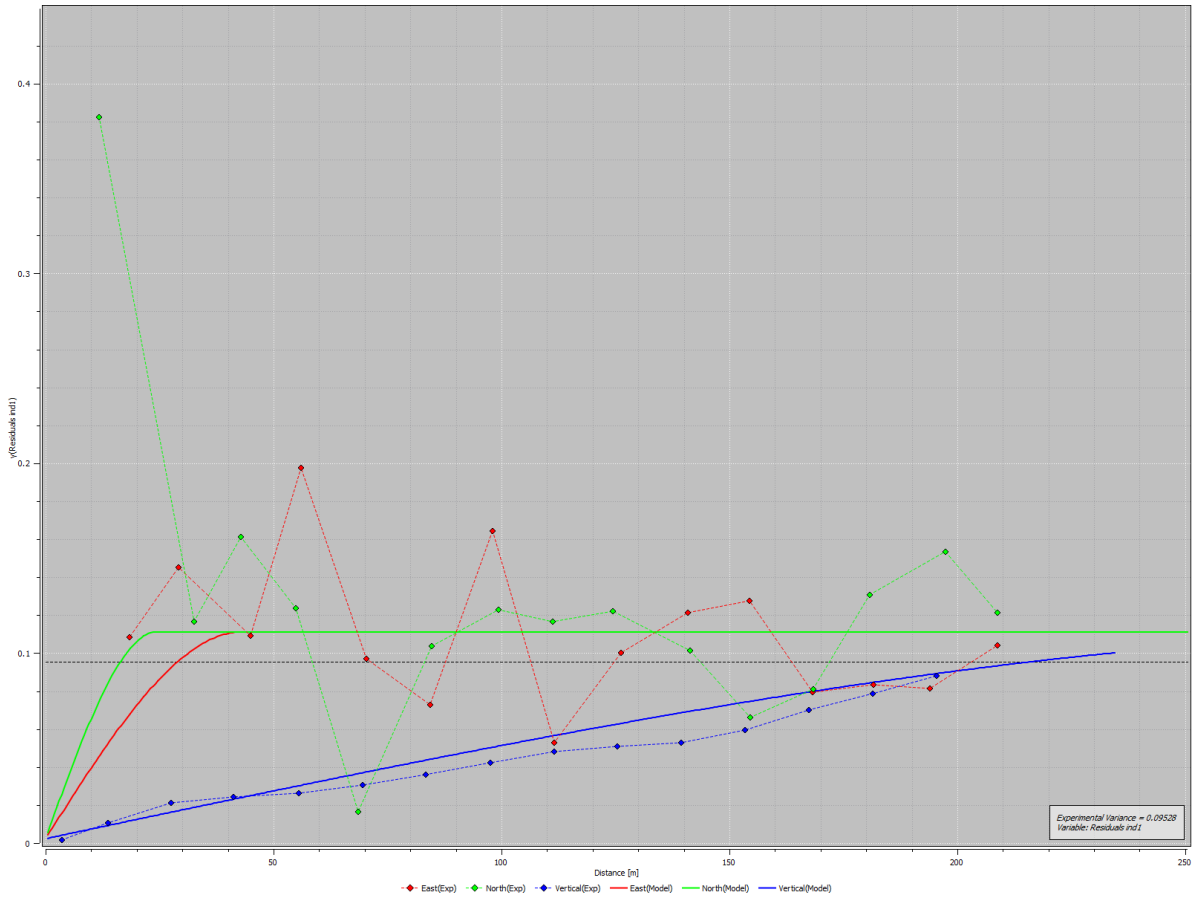


Figure 22. The experimental variogram of residuals at Category 1

$$\gamma Res{-}1 = 0.11Sph(43m, 24m, 322m).$$

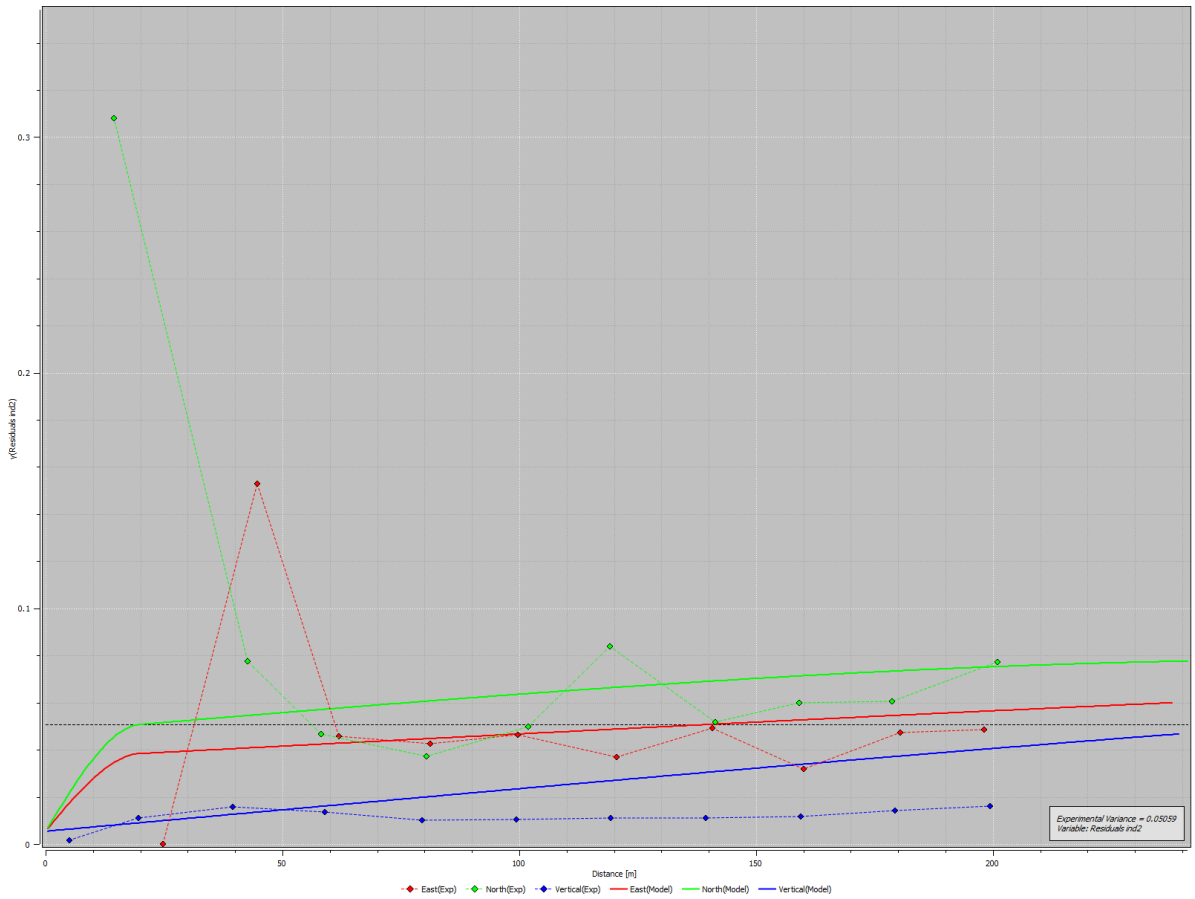


Figure 23. The experimental variogram of residuals at Category 2

$\gamma Res{-}2 = 0.01nugget + 0.03Sph(20m, 272m, 599m) + 0.04Sph(598m, 20m, 598m)$.

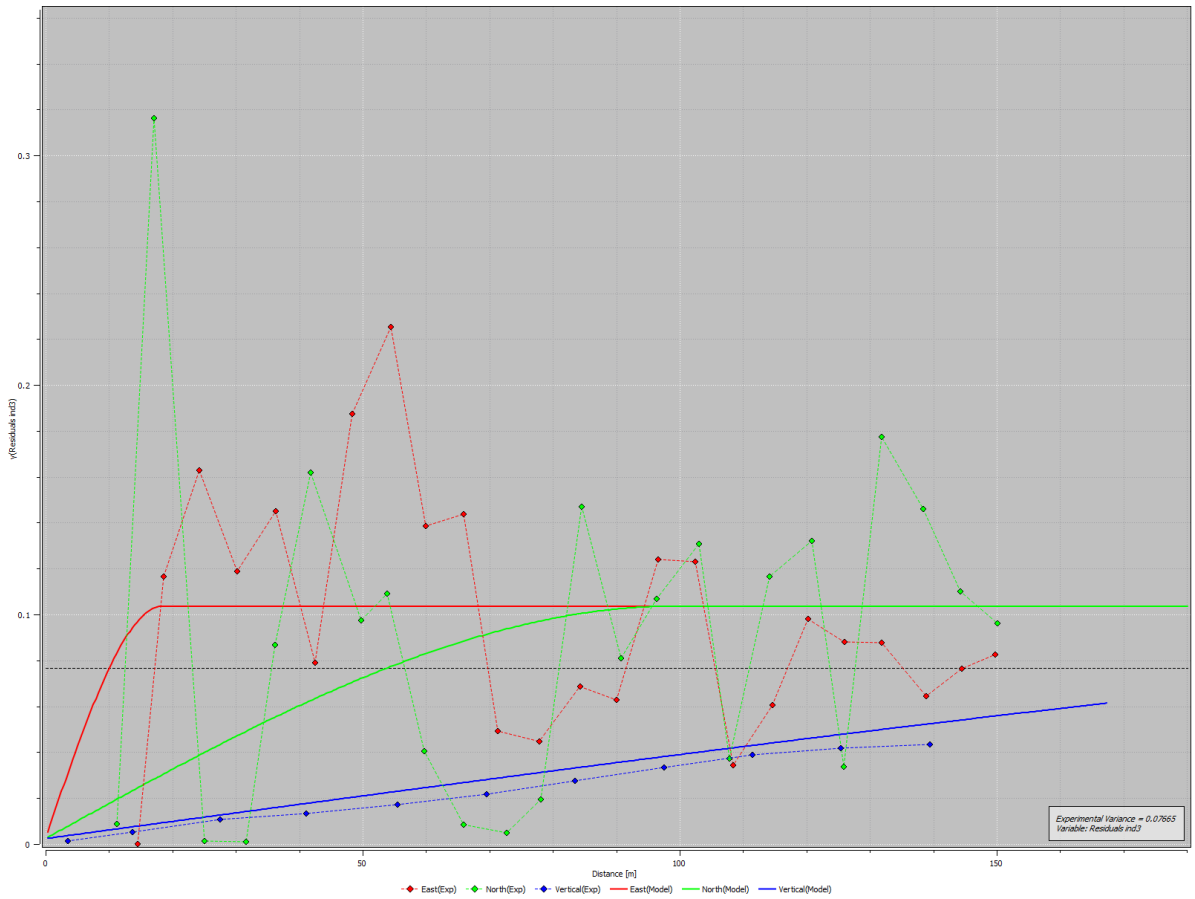


Figure 24. The experimental variogram of residuals at Category 3

$$\gamma Res{-}3 = 0.01 Sph(19m, 99m, 407m).$$

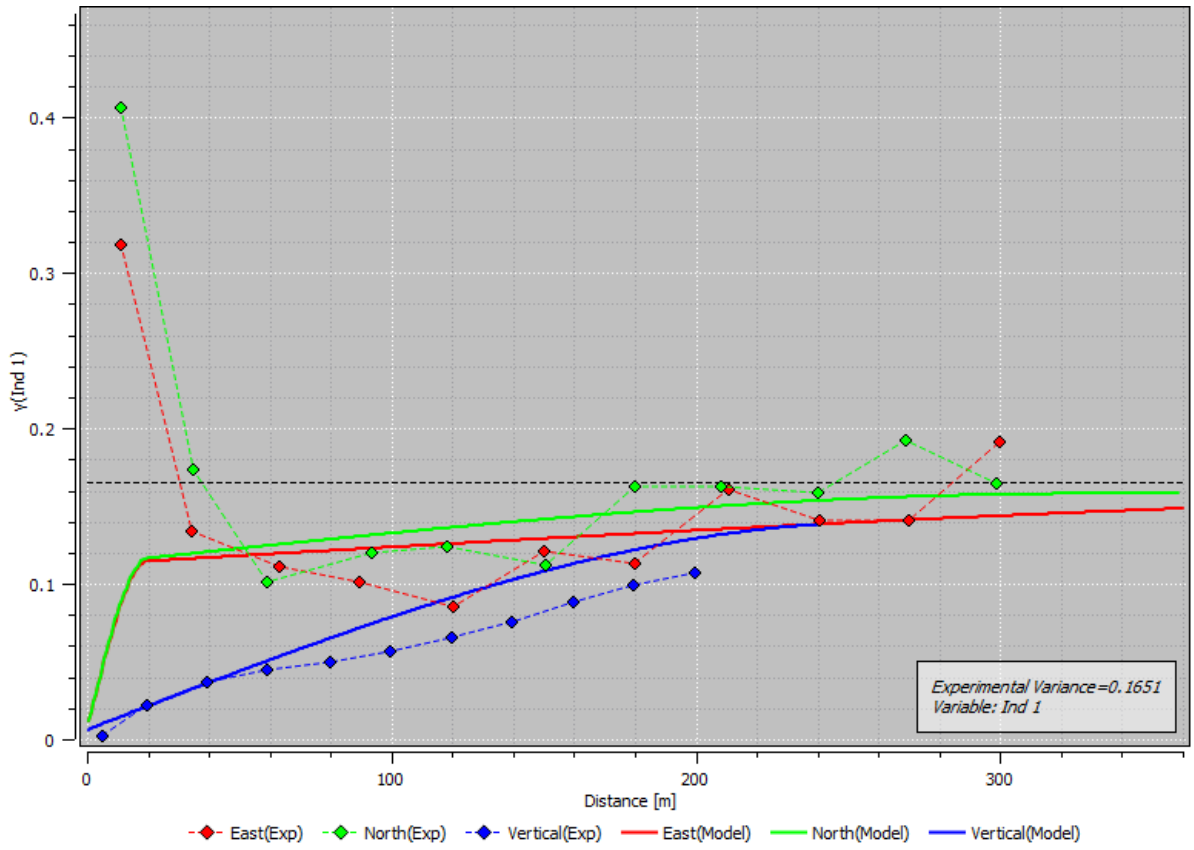


Figure 25. The experimental variogram of indicators at Category 1

$\gamma\text{Ind}-1 = 0.1084Sph + 0.07642(20m, 523m, 600m) + 0.10077Sph(600m, 20m, 318.9m).$

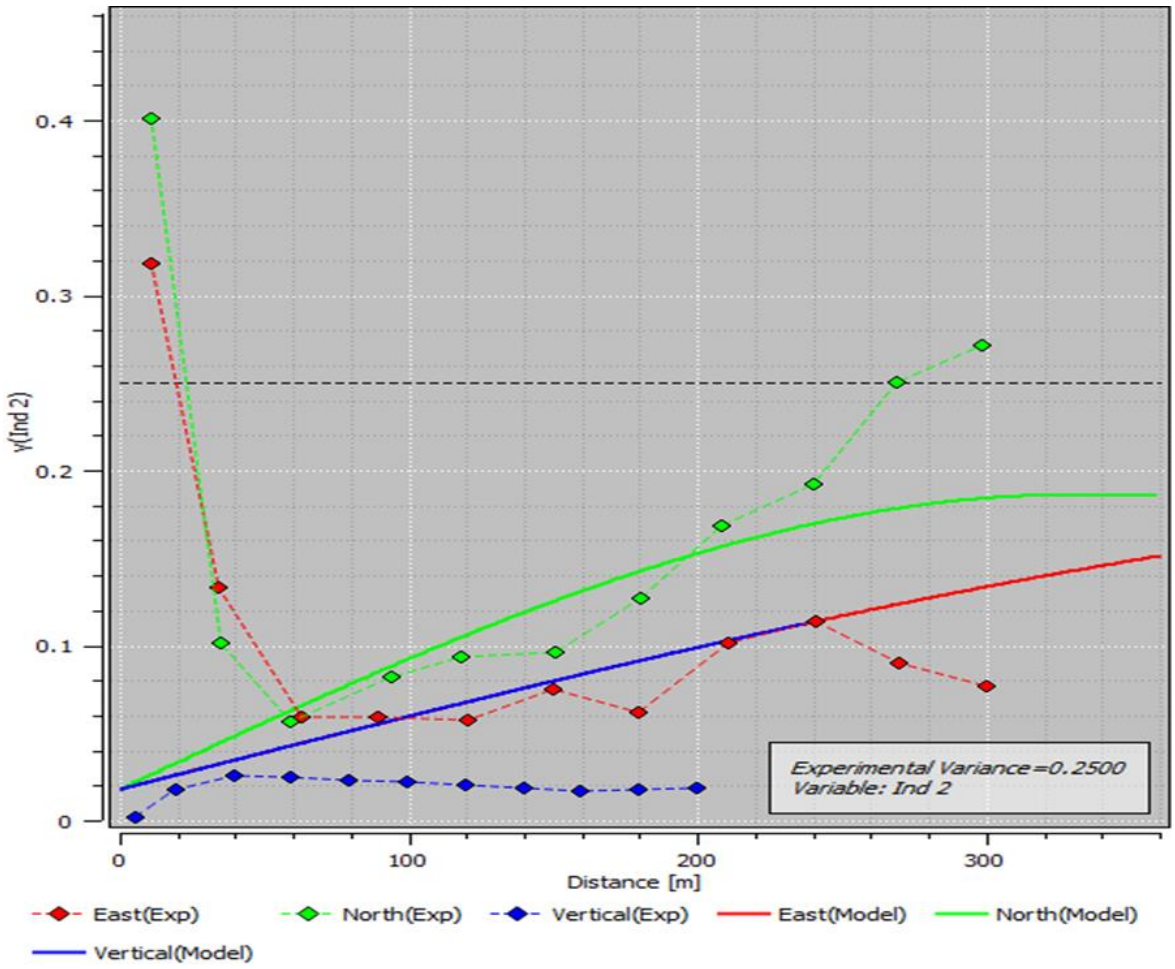


Figure 26. The experimental variogram of indicators at Category 2

$\gamma \text{Ind}-2 = 0.01775\ Sph + 0.16832 Sph(600m, 328.5m, 600m)$.

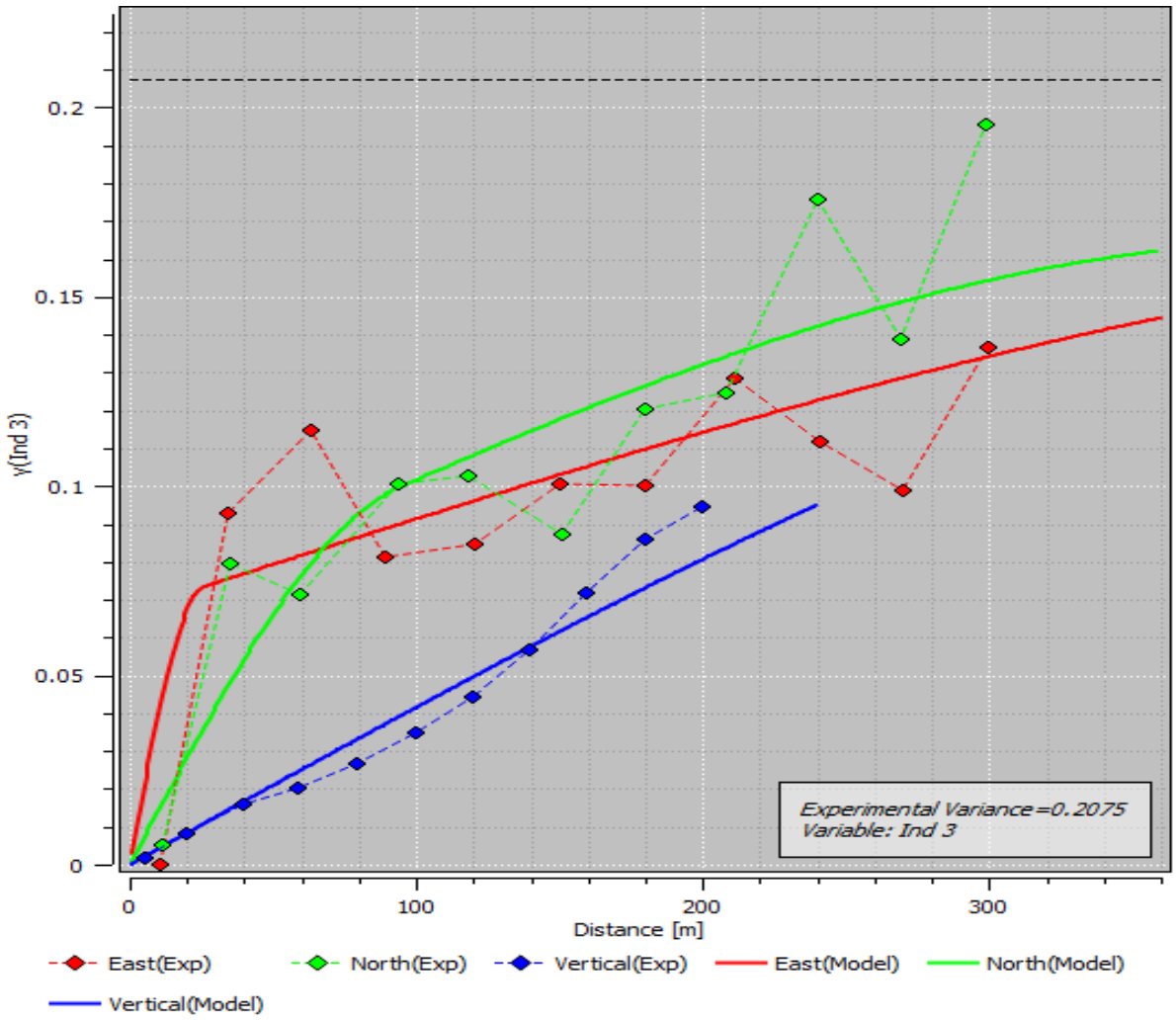


Figure 27. The experimental variogram of indicators at Category 3

$\gamma\text{Ind}-3 = 0.06709Sph(25.92m, 94.19m, 575.2m) + 0.09768Sph(600m, 416.5m, 600m)$.

The variograms of residuals and indicators show finite sill, implying a stationary hypothesis for these variables. The next step involves computing the locally varying mean probabilities or trend component $\mu(Y^* = n), n = 1, \ldots 3,$ which are obtained using the fitted multinomial regression function, with the geographical coordinates of the target grid nodes Y^*as independent variables. The resulting maps are displayed in Figure 30. The estimated probable areas of each geo-cluster are consistent with their spatial distribution over the borehole dataset, as illustrated in Figure 28. Geo-clusters 1, 2, and 3 are highly likely to be found in the lower west, lower east, and upper central parts of the deposit, respectively. This information can serve as a secondary component for simulation using the proposed approach, by adding the estimation conditional probabilities of residuals to obtain the final estimated geo-clusters.

These probabilities are then used along with the residuals, which are calculated over the sample points, and the derived variogram model for each residual as inputs into the proposed sequential indicator simulation algorithm.

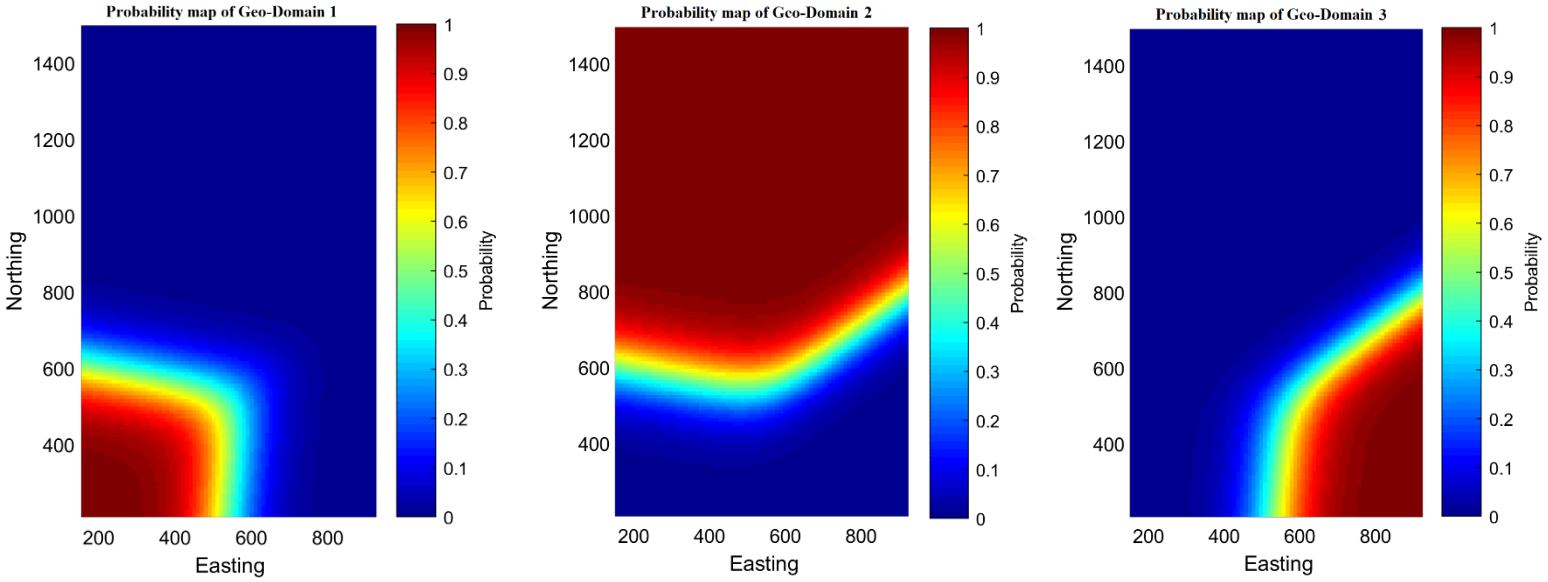
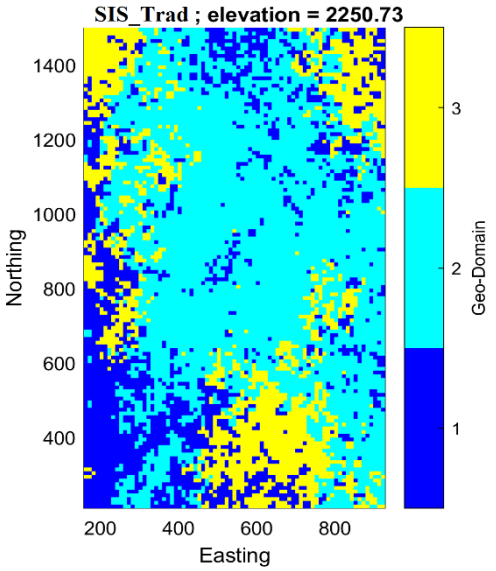
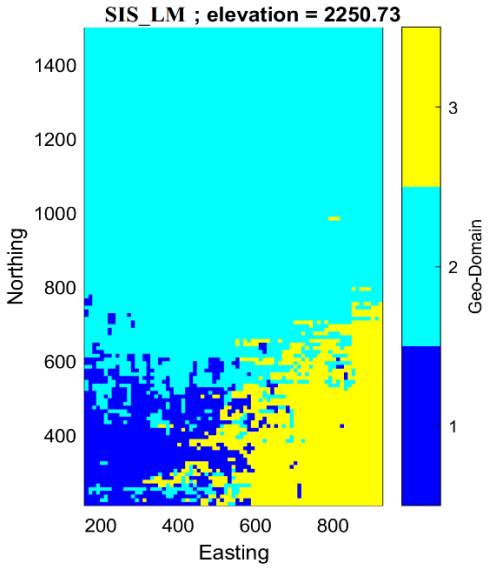
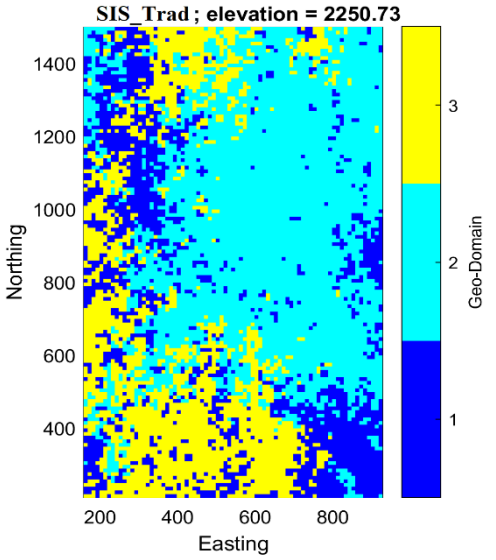


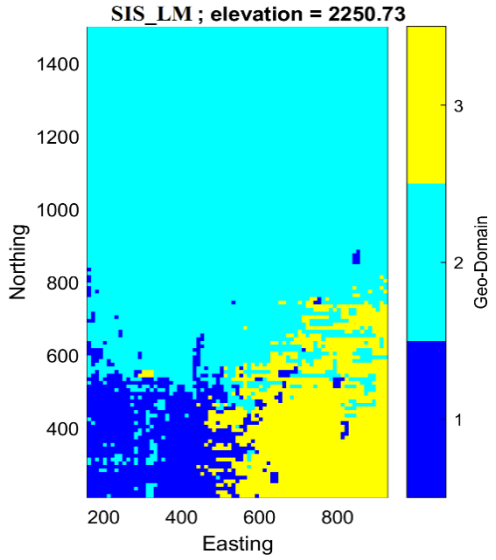
Figure 28. The probability maps for all three Geo-domains calculated by Python. Left for Geo-Domain 1, Middle for Geo-Domain 2, Right for Geo-Domain 3.

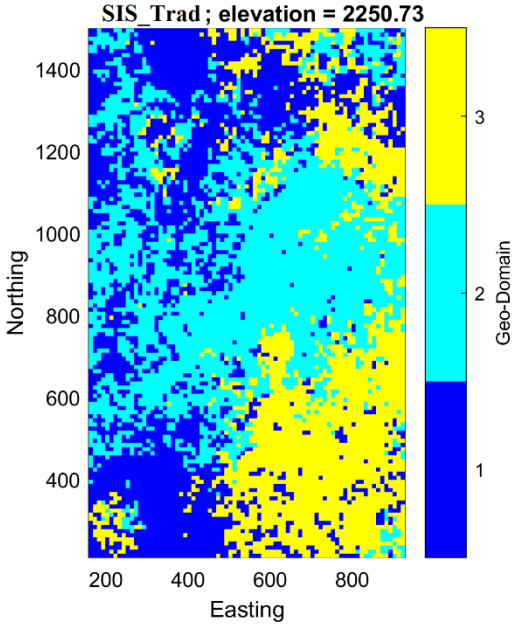
The next step is incorporating soft data and run traditional and proposed SIS approaches. The comparison involves two cases, namely SIS-LM - proposed Sequential Indicator Simulation using non-stationary simple kriging integrated with soft data from Multinomial Logistic Regression, and SIS-Trad - conventional Sequential Indicator Simulation, which does not use secondary data. The comparison is done over the same block model that was discussed earlier, using an identical moving neighborhood for both cases. In order to retrieve unbiased and accurate results, 100 realizations were produced by both methods. The results of 100 realizations for each method are shown in Fig. 29, where it can be observed that the proposed approach (SIS-LM) is better at reproducing the non-stationary characteristics of the geo-domains compared to the traditional SIS_Trad, which produced completely chaotic realization. All


SIS_Trad ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3


SIS_LM ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3


SIS_Trad ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3


SIS_LM ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3


SIS_Trad ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3

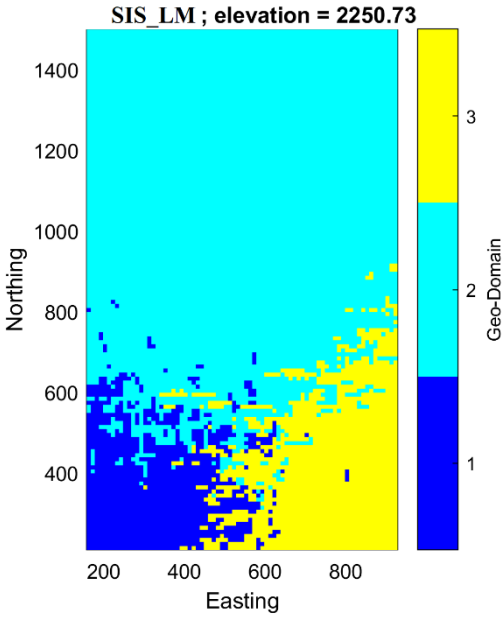

SIS_LM ; elevation = 2250.73
Northing
Easting
Geo-Domain
1400
1200
1000
800
600
400
200
400
600
800
1
2
3

Figure 29. Comparison of realizations obtained by Traditional and Proposed SIS at the same elevations; Left – Traditional SIS; Right – Proposed SIS; blue: geo-domain 1, light blue: geo-domain 2, and yellow: geo-domain 3.

The probability maps can be used to estimate the uncertainty of the geo-domains on a node-by-node basis. These maps can be generated by calculating the proportion of each geo-domains over the 100 simulations. The areas that have low uncertainty are represented by the color red, indicating that the risk of not locating the geo-domain is minimal, while areas with low probability are represented by a light blue color, signifying that there is a high level of certainty that the geo-domain is not present in these areas. The areas represented by colors such as green or yellow indicate more uncertainty. The results obtained from the SIS-LM method are more reasonable and provide a better agreement with the spatial distribution of the geo-domains, which is similar to the conceptual model.

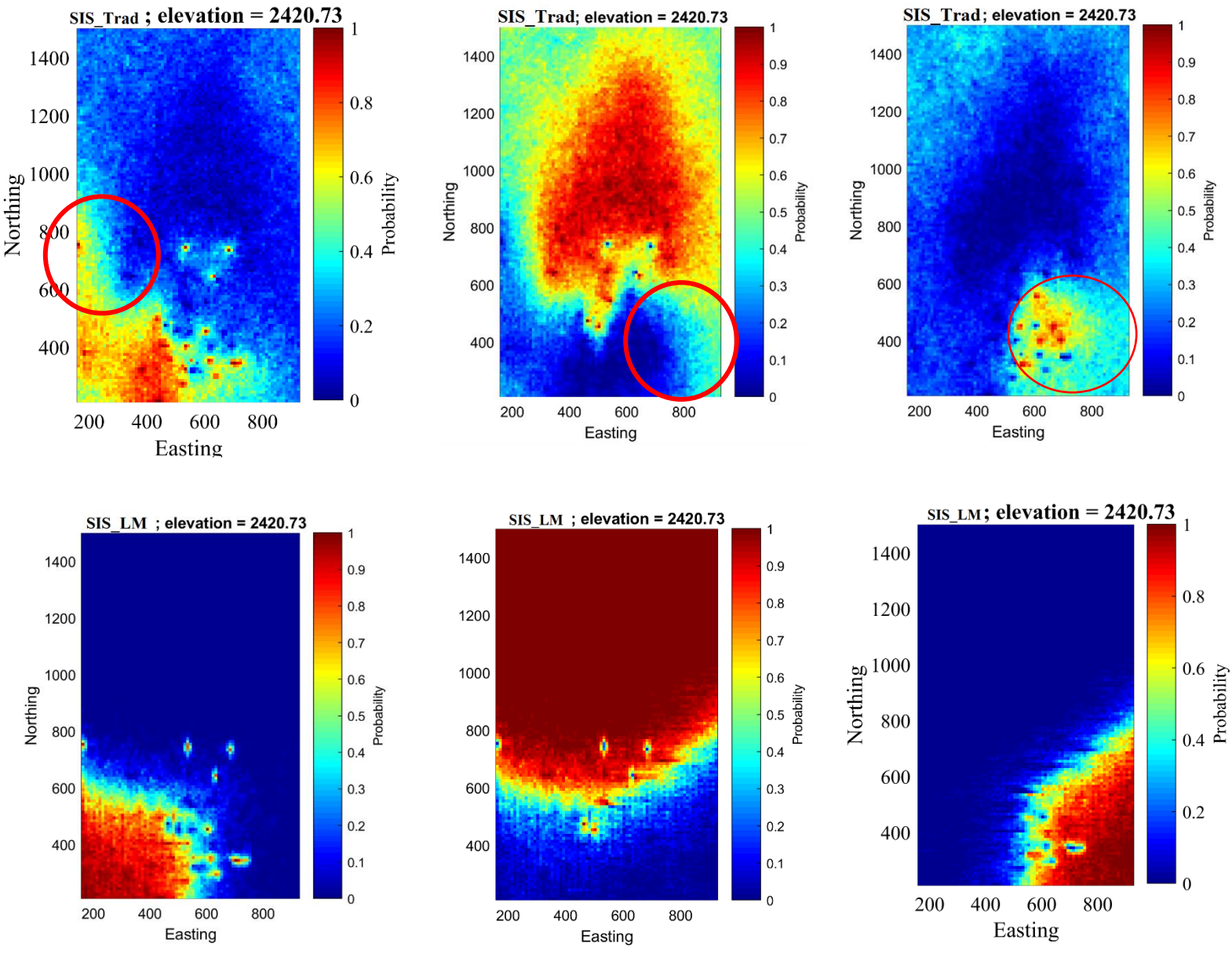


Figure 30. Probability maps obtained with 100 realizations for Proposed SIS and Traditional SIS. Top three realizations were produced by SIS_Trad. Bottom three realizations were produced by SIS_LM.

The probability maps obtained of Traditional SIS shows a weak saturation in all three geo-domains. Furthermore, the borders of geo-domains are not clear, they do not match with their actual locations. The maps show unreliable results, especially probability occurs at points that does not corresponds to certain geo-domain. In contrary, probability maps obtained by Proposed SIS shows completely opposite results with strongly saturated zones in all three geo-domains. According to obtained results, it could be concluded that Proposed SIS produced proper and reliable results in comparison with Traditional SIS.

5.4 STATISTICAL VALIDATION

The next step involves verifying the accuracy of the original trends for each geo-domain. To do this, the trend is calculated for each realization obtained using each SIS method, and then their averages are plotted against the coordinates. An example of this can be seen in Fig. 31, 32, 33, which displays the trends of each geo-domain along the northing, easting, and elevation.

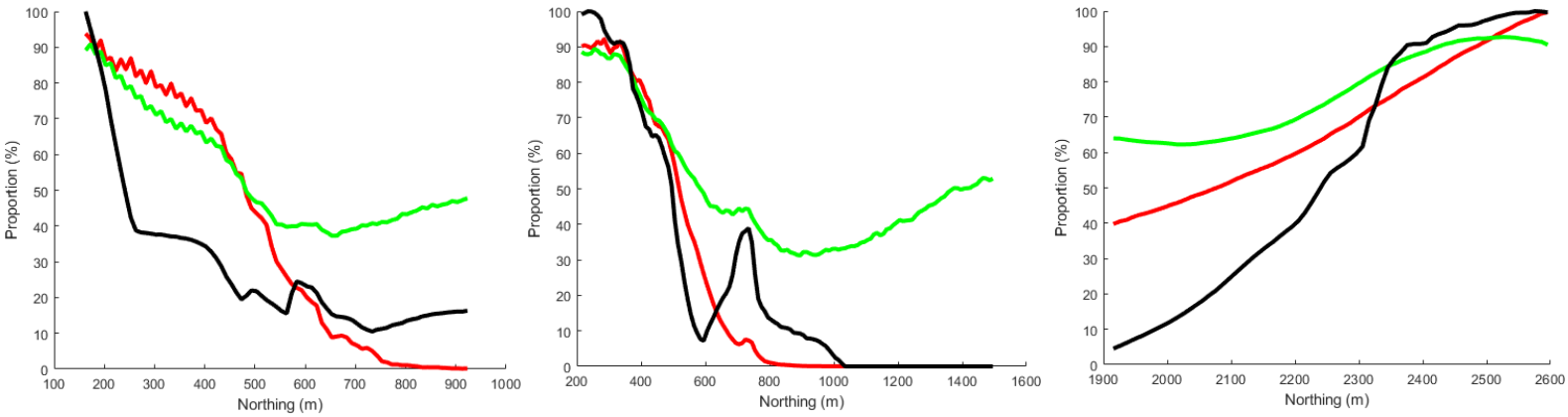


Figure 31. Trend analysis reproduction along easting over the simulation results for geo-domain 1. Black line: original trend; Red line: average of trends over 100 realizations obtained with Proposed SIS; and Green line: average of trends over 100 realizations obtained with Traditional SIS.

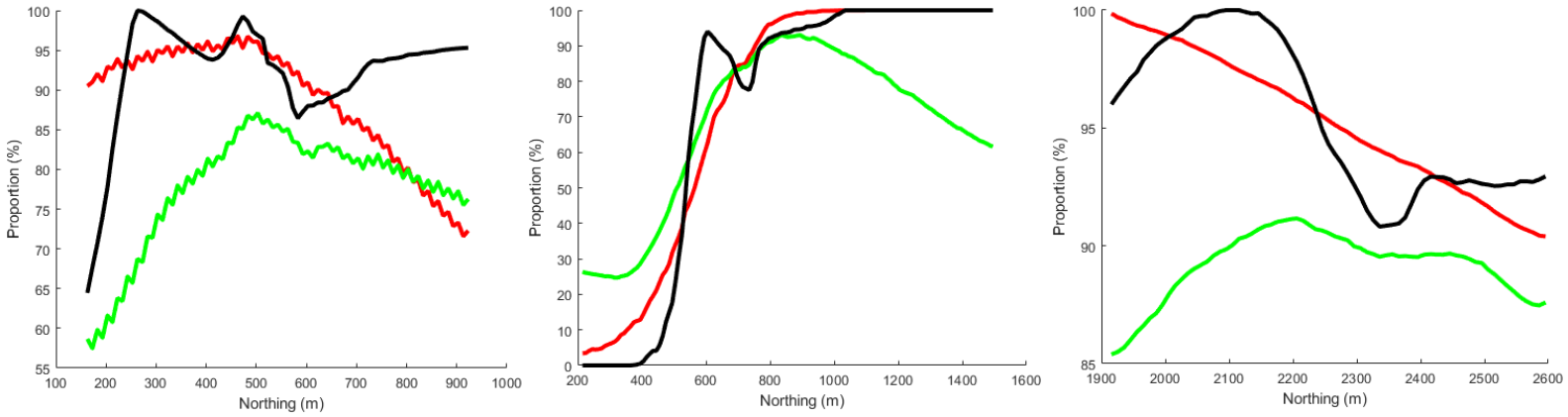


Figure 32. Trend analysis reproduction along easting over the simulation results for geo-domain 2. Black line: original trend; Red line: average of trends over 100 realizations obtained with Proposed SIS; and Green line: average of trends over 100 realizations obtained with Traditional SIS.

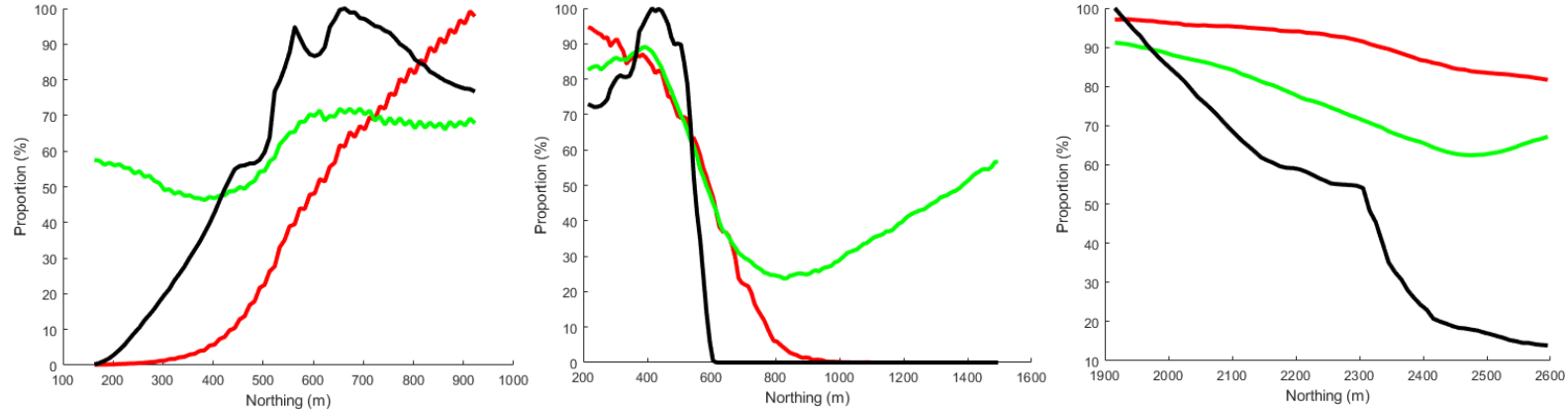


Figure 33. Trend analysis reproduction along easting over the simulation results for geo-domain 3. Black line: original trend; Red line: average of trends over 100 realizations obtained with Proposed SIS; and Green line: average of trends over 100 realizations obtained with Traditional SIS.

The trend reproduced by the Proposed SIS (SIS_lm) method is more consistent with the original trend when compared to the Traditional SIS (SIS_trad) method. In most cases, the Proposed SIS method outperforms the Traditional SIS method, indicating that the stronger the trend component, the more likely it is that the trend along the coordinate can be accurately reproduced. The reason

for the superior performance of the Proposed SIS method is due to its ability to incorporate the trend component into the simulation algorithm, which is informed by Multinomial Logistic Regression.

5.5 COPPER GRADE MODELLING

The next stage involves estimating the copper grade inside of each geo-domain. To obtain these models, simple kriging is utilized with the input of the associated grade variogram and the samples that belong to the respective geo-domain. Then estimation results were merged with simulation outputs, in order to define the ultimate copper grade inside of each geo-domain.

The copper grade modelling uses preliminary kriging estimation to estimate the ore grade at each target node, build the block model and the combine this data with Traditional and Proposed Sequential Indicator Simulation results. Estimation was done by using SGeMS software. Both types of kriging – ordinary and simple were implemented for this study. Simple kriging produced trustworthy estimation maps, while estimation maps by ordinary kriging produced significant number of artifacts and shows low accuracy. Also, Simple Kriging is more reliable because final estimation maps does not affect by smoothing effect (overestimation of high grades and underestimation of low grades) as in case of Ordinary Kriging. Therefore, outputs of ordinary kriging will be neglected and no longer used for the further steps. The kriging results are illustrated on Figure 34.

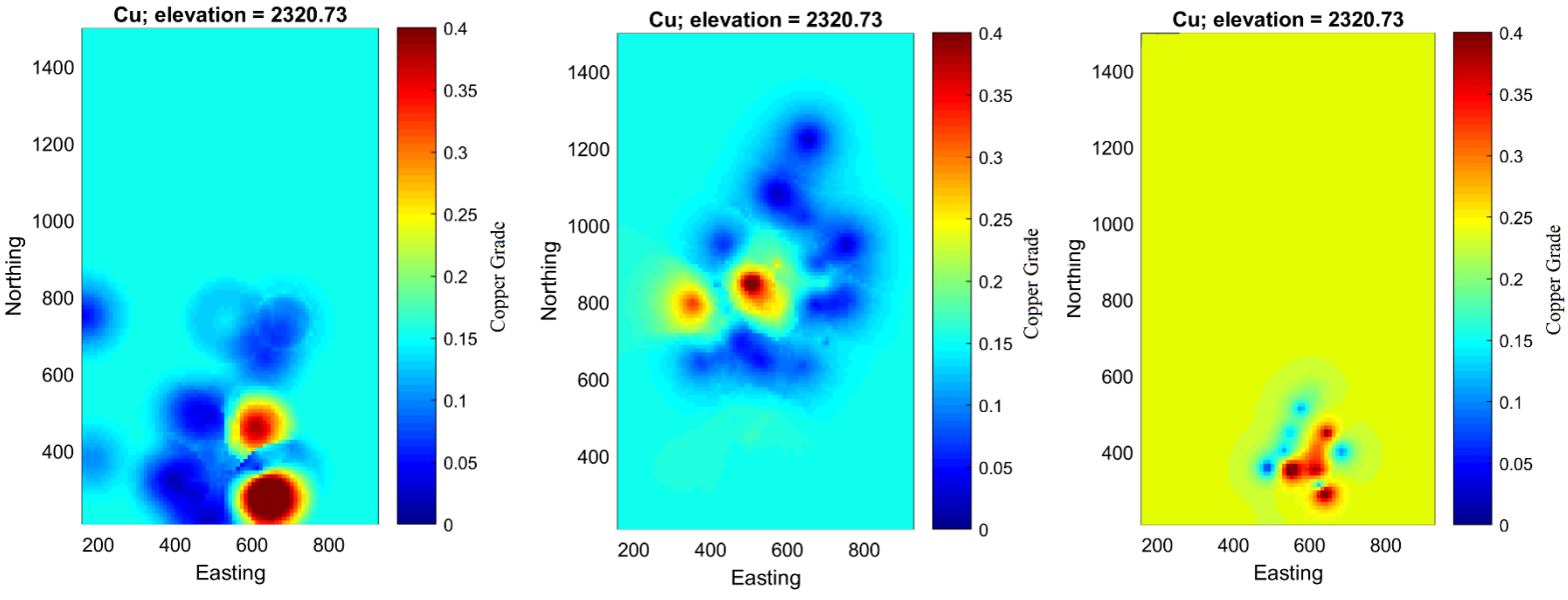


Figure 34. Modelled Copper grade produced by Simple Kriging. Left is for geo-domain 1, Middle is for geo-domain 2, Right is for geo-domain 3.

The estimation results then were transferred in Matlab, where they were combined with simulation outputs of Traditional and Proposed SIS. This combination is necessary to modelling copper grade simultaneously in all three geo-domains. This is achieved using the following formula (Emery & Gonzalez, 2007):

$$Grade\ Estimate = \sum_{k=1}^{3} Probability\,(k^{th} domain)\ x\ Grade\ Estimate(k^{th} domain) \qquad (20)$$

For the sake of unbiased comparison, final copper grade maps at the same elevations were selected from each of the approaches. Since the target grid consists of 70 level, the medium layers as #35, #45 and #55 were. Results are demonstrated on the Figures 35 and 36.

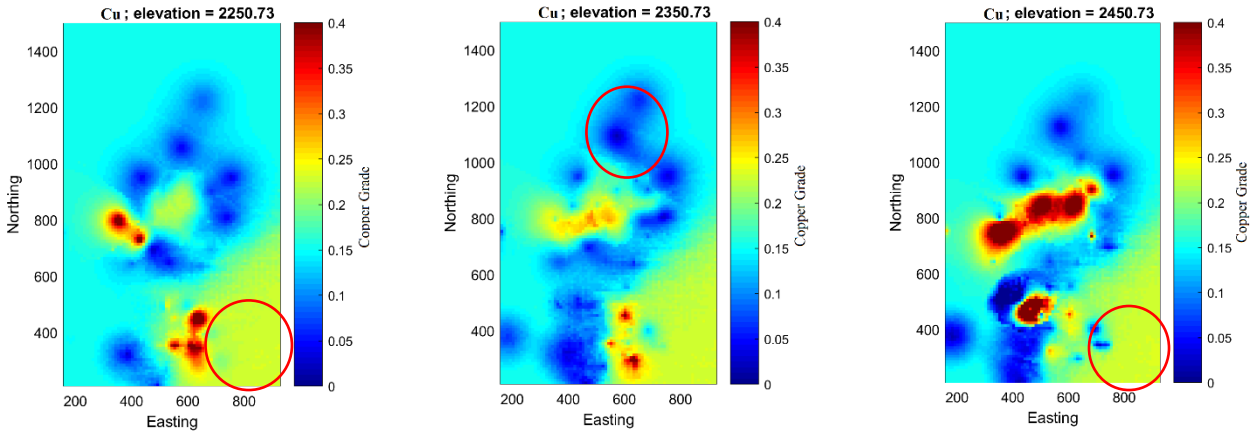


Figure 35. Final Copper Grade Produced by Simple Kriging by combining with Proposed SIS at different elevations. Left for elevation #35, Middle for elevation #50, Right for elevation #55.

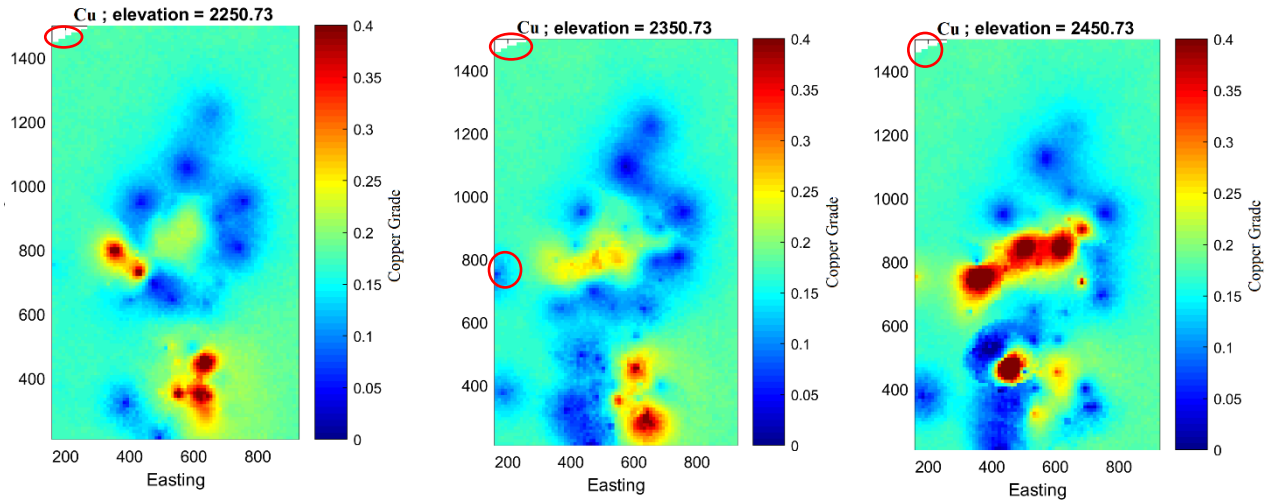


Figure 36. Final Copper Grade Produced by Simple Kriging by combining with Traditional SIS at different elevations. Left for elevation #35, Middle for elevation #50, Right for elevation #55.

Total copper grade maps produced by integrating Proposed SIS performed significantly better than maps provided by Traditional SIS, as expected. Mainly Traditional SIS failed in reproducing of geo-domain #3. Moreover, realizations made by Traditional SIS has a missing upper-left corner. The Traditional SIS realizations also shows underestimation of low-grade zones and overestimation of high-grade zones which are consequences of smoothing effect. On the contrary, Proposed SIS proved to be less susceptible for this limitation of SIS_Trad. Also, poor quality and graininess are direct outcomes of smoothing effect. The maps of final copper grade again prove that Proposed SIS is more robust technique especially in cases of non-stationary domains as this copper-porphyry deposit.

6 DISCUSSIONS

Sequential Indicator Simulation is a very straightforward, rapid and robust technique, however this study shows that Traditional SIS used in most of commercial software is unable to reproduce non-stationary geological domains. Results of Proposed SIS are better by all criteria. Proposed SIS properly shows compact geo-domain, especially it is obvious in direct comparison with results of Traditional SIS. Moreover, realization obtained by Proposed SIS matches with original distribution of geo-domains calculated over the boreholes.

Reliability of the resource estimation is the key factor in any mine project. This study does not require identifying geo-domain of interest, since dataset with distinguished geo-domains was provided. Therefore, the problem which can be met is properly point out boundaries of each geo-domain. This procedure is very sensitive for subjective interpretation of geologists, thus the data provided cannot be always claimed as trustworthy. In particular, manual interpretation of geological data can be prone to errors and inconsistencies. Entailed consequences could decrease of accuracy in resource assessment and could be disastrous for further resource estimation and mine planning stages. The way to minimize errors related to this procedure is use geostatistical hierarchical clustering technique.

Difficulties remain with the order relation problem and neighborhood due to the use of sequential indicator simulation as the base of proposed algorithm (Emery, 2004; Deutsch, 2006). The proposed approach has potential for further improvement, particularly since the simulation process can be slow due to the abundance of hard data. One possible solution to this issue is to use parallel computing. Additionally, the proposed method can be tested on other datasets, coming from not only copper-porphyry deposits or by modelling several minerals simultaneously.

Despite of proved efficiency of Proposed SIS, the realizations may exhibit patchiness along the borders of adjacent geo-domains or contain small spots of another geo-domain. One potential solution to overcome this issue is to employ an image cleaning algorithm based on a maximum a posteriori selection (Deutsch, 1998). This study provided pure results without implementation of cleaning algorithms.

7 CONCLUSION

In the field of geological modelling, creating accurate representations of non-stationary geological domains is critical for making informed engineering decisions. Traditional methods, such as sequential indicator simulation (SIS), have limitations in accurately capturing the complex spatial patterns and trends found in geological domains. However, recent developments in machine learning and statistical modelling have opened up new possibilities for improved geological modelling techniques. One such method is the use of sequential indicator simulation with local mean probabilities and residuals calculated from multinomial logistic regression. This approach allows for the modelling of secondary information and guidance in the modelling of non-stationary trends in geological domains. Compared to traditional SIS, this method results in more accurate representations of each geological domain, as demonstrated by improved probability maps and realizations with reduced error. This study proposes a novel technique for modelling various types of geological domains resulting from spatially-dependent clustering machine learning algorithms. The proposed method employs multinomial logistic regression to model secondary information and guide the modelling of non-stationary sequential indicator simulation trends. The resulting approach outperforms traditional SIS in terms of visual representation of geo-domains in resulting maps, reproduction of geo-domain proportions, and reproduction of indicator variogram, connectivity measures, and trend component. The proposed method can model any geo-domain, including geo-domains with trend components, and is designed to produce compact and contiguous domains. However, some minor patchiness or tiny spots of geo-domains at the borders of adjacent geo-domains may exist. This issue can be resolved using image cleaning based on maximum a posteriori selection. It is important to note that this approach utilizes sequential indicator simulation, and thus problems related to order relation and neighborhood still exist. Future research can explore non-stationary approaches using plurigaussian simulation or multiple-point statistics. Overall, this proposed approach provides a potential avenue for better support for engineering decisions in geological modelling by improving the accuracy and reliability of geological domain representations.

8 REFERENCE LIST

Belkacim, S., Ikenne, M., Souhassou, M., Elbasbas, A., & Toummite, A. (2014). The Cu-Mo±Au mineralizations associated to the High-K calc-alkaline granitoids from Tifnoute valley (Siroua massif, anti-atlas, Morocco): An arc-Type porphyry in the late neoproterozoic series. *J. Environ. Earth Sci*, *4*, 90-106.

Alabert, F. (1987). The practice of fast conditional simulations through the LU decomposition of the covariance matrix. *Mathematical Geology*, *19*(5), 369-386.

Beucher, H., Galli, A., Loc'h, G. L., Ravenne, C., & Heresim Group. (1993). Including a regional trend in reservoir modelling using the truncated Gaussian method. In *Geostatistics Tróia'92* (pp. 555-566). Springer, Dordrecht.

Chilès JP, Delfiner P, 2012. Geostatistics: Modeling Spatial Uncertainty, 2th edition. Wiley, New York.

Cox, D. R., & Miller, H. D. (2017). *The theory of stochastic processes*. Routledge.

Cressie, N., and Chan, N. H. (1989). Spatial modeling of regional variables. Journal of the American Statistical Association, 84(406), 393-401.

de Almeida, J. A. (2010). Stochastic simulation methods for characterization of lithoclasses in carbonate reservoirs. *Earth-Science Reviews*, *101*(3-4), 250-270.

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., & Timonin, V. (1998). Neural network residual kriging application for climatic data. *Journal of Geographic Information and Decision Analysis*, *2*(2), 215-232.

Deutsch, C. 1992. AnneaJing Techniques Applied to Reservoir Modeling and the Integration of Geological and Engineering (Weil Test) Data. PhD thesis, Stanford University, California.

Deutsch, C. V. (2006). A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. Computers and Geosciences, 32(10), 1669-1681.

Deutsch, C. V., and Journel, A. G. (1992). GSLIB: geostatistical library and user's guide.

Deutsch, C.V. and Journel, A.G., (1997). GSLIB Geostatistical Software Library and User's Guide, Oxford University Press, New York, second edition. 369 pages

Dowd, P. (1986). Geometrical and geological controls in geostatistical estimation and orebody modelling.

Dowd, P. A. (1991). A review of recent developments in geostatistics. *Computers and Geosciences*, *17*(10), 1481-1500.

Dowd, P. A. (1994). Risk assessment in reserve estimation and open-pit planning. *Transactions of the Institution of Mining and Metallurgy-Section A-Mining Industry*, *103*, A148.

Dowd, P.A., Pardo-Igúzquiza, E., Xu, C., 2003. Plurigau: a computer program for simulating spatial facies using the truncated plurigaussian method. Computers & Geosciences 29 (2), 123–141.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Advances in geophysics*, *61*, 1-55.

Dubrule, O. (1993). Introducing more geology in stochastic reservoir modelling. In *Geostatistics Tróia'92* (pp. 351-369). Springer, Dordrecht.

Duke, J. H., & Hanna, P. J. (2001). Geological interpretation for resource modelling and estimation. *Mineral resource and ore reserve estimation—The AusIMM guide to good practice*, 147-156.

Emery, X. (2004). Properties and limitations of sequential indicator simulation. *Stochastic Environmental Research and Risk Assessment*, *18*(6), 414-424.

Emery, X. (2007). Simulation of geological domains using the plurigaussian model: new developments and computer programs. Computers and geosciences, 33(9), 1189-1201.

Emery, X., González, K., 2007a. Incorporating the uncertainty in geological boundaries into mineral resources evaluation. J. Geol. Soc. India 69 (1), 29–38.

Emery, X., González, K.E., 2007b. Probabilistic modelling of lithological domains and its application to resources evaluation. J. South. Afr. Inst. Min. Metall. 107 (12), 803–809.

Galli A, Beucher H, Le Loc'h G, Doligez B (1994) The pros and cons of the truncated Gaussian method. In: Armstrong M, Dowd PA (eds) Geostatistical simulations. Kluwer, Dordrecht, pp 217–233

Galli, A., Armstrong, M., and Jehl, B. (1999, March). Comparing three methods for evaluating oil projects: option pricing, decision trees, and Monte Carlo simulations. In *SPE hydrocarbon economics and evaluation symposium*. OnePetro.

Gómez-Hernández, J. J., and Srivastava, R. M. (1990). ISIM3D: An ANSI-C three-dimensional multiple indicator conditional simulation program. Computers and Geosciences, 16(4), 395-440.

Goovaerts, P. (1994). Comparative performance of indicator algorithms for modeling conditional probability distribution functions. Mathematical Geology, 26(3), 389-411.

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

Goovaerts, P., Webster, R., & Dubois, J. P. (1997). Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and ecological Statistics*, *4*(1), 49-64.

Journel, A. B., and Alabert, F. G. (1989). Focusing on spatial connectivity of extreme-valued attributes: Stochastic indicator models of reservoir heterogeneities. *AAPG Bull.;(United States)*, *73*(CONF-890404-).

Journel, A. G. (1974). Geostatistics for conditional simulation of ore bodies. *Economic Geology*, *69*(5), 673-687.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology, 15(3), 445-468.

Journel, A. G., and Alabert, F. G. (1990). New method for reservoir mapping. *Journal of Petroleum technology*, *42*(02), 212-218.

Journel, A. G., and Isaaks, E. H. (1984). Conditional indicator simulation: application to a Saskatchewan uranium deposit. Journal of the International Association for Mathematical Geology, 16(7), 685-718.

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

Journel, A.G. and Alabert, F. 1988. Focusing on Spatial Connectivity of Extreme-valued Attributes: Stochastic Indicator Models of Reservoir Heterogeneities, SPE 18,324.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London.

Kanevski, M. (2009). *Machine learning for spatial environmental data: theory, applications, and software*. EPFL press.

Kanevski, M., & Demyanov, V. (2015). Statistical learning in geoscience modelling: novel algorithms and challenging case studies. *Computers and Geosciences*, *85*, 1-2.

Kanevski, M., Kanevski, M. F., & Maignan, M. (2004). *Analysis and modelling of spatial environmental data* (Vol. 6501). EPFL press.

Kim, K. H., Chiu, J. M., Pujol, J., Chen, K. C., Huang, B. S., Yeh, Y. H., & Shen, P. (2005). Three-dimensional VP and VS structural models associated with the active subduction and collision tectonics in the Taiwan region. *Geophysical Journal International*, *162*(1), 204-220.

Kupfersberger, H., Deutsch, C. V., and Journel, A. G. (1998). Deriving constraints on small-scale variograms due to variograms of large-scale data. Mathematical geology, 30(7), 837-852.

Madani N. (2021) Plurigaussian Simulations. In: Daya Sagar B., Cheng Q., McKinley J., Agterberg F. (eds) Encyclopedia of Mathematical Geosciences. Encyclopedia of Earth Sciences Series. Springer, Cham. https://doi.org/10.1007/978-3-030-26050-7_251-1

Madani, N., & Emery, X. (2015). Simulation of geo-domains accounting for chronology and contact relationships: application to the Río Blanco copper deposit. *Stochastic environmental research and risk assessment*, *29*(8), 2173-2191.

Madani, N., & Emery, X. (2017). Plurigaussian modeling of geological domains based on the truncation of non-stationary Gaussian random fields. *Stochastic environmental research and risk assessment*, *31*(4), 893-913.

Madani, N., Maleki, M., and Soltani-Mohammadi, S. (2022). Geostatistical Modeling of Heterogeneous Geo-clusters in a Copper Deposit Integrated with Multinomial Logistic Regression: an Exercise on Resource Estimation. *Ore Geology Reviews*, 105132.

Matern, B. (1960), Spatial Variation, Meddelanden fran Statens Skogsforskningsinstitut, 495. Second ed. (1986), Lecture Notes in Statistics 36, New York: Springe

Matheron, G. (1967). Kriging or polynomial interpolation procedures. *CIMM Transactions*, *70*(1), 240-244.

Matheron, G. (1971). The theory of regionalised variables and its applications. *Les Cahiers du Centre de Morphologie Mathématique*, *5*, 212.

Mizuno, T. A., and Deutsch, C. V. Sequential Indicator Simulation (SIS).

Myers, D. E. (1989). To be or not to be... stationary? That is the question. *Mathematical Geology*, *21*(3), 347-362.

Ortiz, JM & Emery, X. (2006). Geostatistical estimation of mineral resources with soft geological boundaries: a comparative study. *Journal of the Southern African Institute of Mining and Metallurgy*, *106*(8), 577-584.

Pawlowsky, V., Olea, R. A., and Davis, J. C. (1993). Boundary assessment under uncertainty: a case study. *Mathematical Geology*, *25*(2), 125-144.

Pearson, E. S. (1930). A further development of tests for normality. *Biometrika*, 239-249.

Pearson, K. (1895). VII. Note on regression and inheritance in the case of two parents. *proceedings of the royal society of London*, *58*(347-352), 240-242.

Ravenne, C., Galli, A., Doligez, B., Beucher, H., & Eschard, R. (2002). Quantification of facies relationships via proportion curves. In *Geostatistics Rio 2000* (pp. 19-39). Springer, Dordrecht.

Remy, É., Ruet, P., and Thieffry, D. (2008). Graphic requirements for multistability and attractive cycles in a Boolean dynamical framework. Advances in Applied Mathematics, 41(3), 335-350.

Samson, M., & Deutsch, C. V. (2022). A hybrid estimation technique using elliptical radial basis neural networks and cokriging. *Mathematical Geosciences*, *54*(3), 573-591.

Sarkar, A., Ramesh, R., Bhattacharya, S. K., & Rajagopalan, G. (1990). Oxygen isotope evidence for a stronger winter monsoon current during the last glaciation. *Nature*, *343*(6258), 549-551.

Soares, A. (1992). Geostatistical estimation of multi-phase structures. *Mathematical geology*, *24*(2), 149-160.

Sojdehee, M., Rasa, I., Nezafati, N., Abedini, M. V., Madani, N., and Zeinedini, E. (2015). Probabilistic modeling of mineralized zones in Daralu copper deposit (SE Iran) using sequential indicator simulation. *Arabian Journal of Geosciences*, *8*(10), 8449-8459.

Stegman, C. L. (2001). Cobar deposits: still defying classification!. *SEG Discovery*, (44), 1-26.

Tabachnick, B. G., Fidell, L. S., and Ullman, J. B. (2007). Using multivariate statistics (Vol. 5, pp. 481-498). Boston, MA: pearson.

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., & Emery, W. J. (2009). Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(7), 2218-2232.

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2, No. 3, p. 4). Cambridge, MA: MIT press.

9 APPENDICE

A1. PYTHON CODE FOR BUILDING MULTINOMIAL LOGISTIC REGRESSION IN CASE STUDY I

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
#50 points
points = pd.read_excel('file.xlsx')
X = points[['X', 'Y', 'Z']]
Y = points['Category']
rs = 42
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2,
 shuffle = True, random_state = rs)
clf = LogisticRegression(random_state = rs, multi_class = 'multinomial').f
it(X_train, y_train)
y_pred = clf.predict(X)
print(classification_report(y_pred, Y))
X_vis = X.copy()
X_vis['predictions'] = y_pred
X_vis['Categories'] = Y
X_vis.head()
y_vis = np.array(X_vis['predictions'])
points_pred = clf.predict(points[['X', 'Y', 'Z']])
points['predictions'] = points_pred
first_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 0]
second_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 1]
third_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 2]
points['1_proba'], points['2_proba'], points['3_proba'] = first_proba, sec
ond_proba, third_proba
arr_proba = []
probas = clf.predict_proba(points[['X', 'Y', 'Z']])
cnt = 0
for i in points['predictions']:
    arr_proba.append(probas[cnt, i - 1])
    cnt+= 1
```

```
points['proba_final'] = arr_proba
points.head()
#90 k
data = pd.read_excel('50_with_proba.xlsx')
data.head()
data_pred = clf.predict(data[['X', 'Y', 'Z']])
data['predictions'] = data_pred
first_proba = clf.predict_proba(data[['X', 'Y', 'Z']])[:, 0]
second_proba = clf.predict_proba(data[['X', 'Y', 'Z']])[:, 1]
third_proba = clf.predict_proba(data[['X', 'Y', 'Z']])[:, 2]
data['1_proba'], data['2_proba'], data['3_proba'] = first_proba, second_pr
oba, third_proba
arr_proba = []
probas = clf.predict_proba(data[['X', 'Y', 'Z']])
cnt = 0
for i in data['predictions']:
    arr_proba.append(probas[cnt, i - 1])
    cnt+= 1
data['proba_final'] = arr_proba
data
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred_test = model.predict(X_test)
y_predict = model.predict(data[['X','Y', 'Z']])
data['predict'] = y_predict
y_vis_2 = np.array(points['proba_final'])
from matplotlib.colors import ListedColormap
cmap_bold_2 = ListedColormap(['none'])
fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,10))

sc = plt.scatter(data[data.predict==1].X, data[data.predict==1].Y, c = dat
a[data.predict==3].proba_final, cmap='jet')
plt.scatter(points['X'], points['Y'], edgecolors='w', facecolor='none')

plt.title('Category 1', size=20)
plt.xlabel('Easting', size=20)
plt.ylabel('Northing', size=20)
```

```
cbar = plt.colorbar(sc, )
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1.0])

# cbar.ax.set_ylim(bottom=0)

plt.rcParams['axes.facecolor']='#03045c'


#plt.fill(150, 150, "b")
ax = plt.gca()
ax.set_xlim([0, 300])
ax.set_ylim([0, 300])
plt.xticks(range(0,350,50))
fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,9))

sc = plt.scatter(data[data.predict==2].X, data[data.predict==2].Y, c = dat
a[data.predict==2].proba_final, cmap='jet')
plt.scatter(points['X'], points['Y'], edgecolors='w', facecolor='none')

plt.title('Category 2', size=20)
plt.xlabel('Easting', size=20)
plt.ylabel('Northing', size=20)

cbar = plt.colorbar(sc, )
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1])
cbar.ax.locator_params(nbins=6)
cbar.ax.set_ylim(bottom=0)

plt.rcParams['axes.facecolor']='#03045c'


plt.fill(150, 150, "b")
ax = plt.gca()
ax.set_xlim([0, 300])
ax.set_ylim([0, 300])
plt.xticks(range(0,350,50))
fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,10))
```

```
sc = plt.scatter(data[data.predict==3].X, data[data.predict==3].Y, c = dat
a[data.predict==3].proba_final, cmap='jet')
plt.scatter(points['X'], points['Y'], edgecolors='w', facecolor='none')

plt.title('Category 3', size=20)
plt.xlabel('Easting', size=20)
plt.ylabel('Northing', size=20)

cbar = plt.colorbar(sc, )
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1.0])

# cbar.ax.set_ylim(bottom=0)

plt.rcParams['axes.facecolor']='#03045c'


#plt.fill(150, 150, "b")
ax = plt.gca()
ax.set_xlim([0, 300])
ax.set_ylim([0, 300])
plt.xticks(range(0,350,50))
```

A2. MATLAB CODE FOR SEQUENTIAL INDICATOR SIMULATION

```
%clear

%z=load('all_predicted_.txt'); % Final predicted map

%z=tb_r;

%data=load('randomly_selected.txt'); % randomly selected dataset

i = 1; % the number o fcolumn in z

nx = 78; ny = 130;nz=70; level = 35; % number of blocks in x, y, and z,
level: it is the corresponidng plane, in 2D, level is 1.

nreal=100;

figure(1);

 %real =  mean(cu(:,1:100)==3,2);

 real =  calc_trad(:,1);

 % real =  sisim_TEST_trad(:,15);

real = reshape(real,nx*ny,nz);

      set(gcf,'DefaultAxesFontName','Times','DefaultAxesFontSize',14)

      pcolor(reshape(real(:,level),nx,ny)');

      axis('image');

      shading('flat');

      xlabel('Easting');

      ylabel('Elevation');

      title('SIS__LM Hierarchical SK lvl 60')

    colormap('jet')

 %  hold on; plot(borehole(:,1)-min(borehole(:,1)),borehole(:,2)-
min(borehole(:,2)),'wo')

   caxis([0.1 0.40])

  return

      figure(2);
```

```
  %real = mean(sis_trad(:,1:100)==3,2);
 real = calc_lm(:,7);
 % real = sisim_TEST(:,1);
real = reshape(real,nx*ny,nz);
     set(gcf,'DefaultAxesFontName','Times','DefaultAxesFontSize',14)
     pcolor(reshape(real(:,level),nx,ny)');
     axis('image');
     shading('flat');
     xlabel('Easting');
     ylabel('Northing');
     title('SIS__LM Hierarchical OK lvl 60')
   colormap('jet')
 hold on;
% hold on; plot(data_50(:,1),data_50(:,2),'wo')
 caxis([0.1 0.40])
  return
   figure(3);
% real = mean(sisim_TEST(:,1:100)==4,1);
 real = grid_2(:,6);
 % real = sisim_TEST(:,1);
real = reshape(real,nx*ny,nz);
     set(gcf,'DefaultAxesFontName','Times','DefaultAxesFontSize',14)
     pcolor(reshape(real(:,level),nx,ny)');
     axis('image');
     shading('flat');
     xlabel('Easting');
     ylabel('Northing');
     title('Category 2')
```

```
   colormap('jet')
 hold on;
% hold on; plot(data_50(:,1),data_50(:,2),'wo')
% caxis([0 1])
```

A3. SISIM PROGRAM SOR SIS_LM AND SIS_TRAD

```
                  Parameters for SISIM

                  ********************


START OF PARAMETERS:

data.out                          % file with conditioning data

1  2  3                           %        columns for data coordinates

4                                 %        column(s) for data values

5  6  7                          %        columns for local mean of
probability for each category

0.179604262 0.52438936 0.296006378      %        global proportion for
each category

grid_2.out                        % file with coordinates of locations
targeted for simulation

1  2  3                           %        columns for location coordinates

1  2  3                            %        columns for local mean of
probability for each category

1  10  10  10                    %        gridded locations (1=yes, 0=no)?
mesh size (0  0  0 if not gridded)

3                                 % number of categories

2  0.00605                              % Category 1:number of nested
structures, nugget effect

1  0.10632   20   20   246    0  0  0  1       %        1st structure: it cc a1
a2  a3  ang1  ang2  ang3  b

1  0.04575   600   328   600  0  0  0  1       %        1st structure: it cc a1
a2  a3  ang1  ang2  ang3  b

2  0.00918                           % Category 2:number of nested
structures, nugget effect

1  0.02562   20   20   246    0  0  0  1      %        1st structure: it cc a1 a2
a3  ang1  ang2  ang3  b
```

```
1 0.14247   600  328  600 0 0 0 1       %        2st structure: it cc a1
a2 a3 ang1 ang2 ang3 b

2 0.00220                            % Category 3:number of nested
structures, nugget effect

1 0.05703   20  20  246   0 0 0 1     %        1st structure: it cc a1 a2
a3 ang1 ang2 ang3 b

1 0.09494  600  328  600 0 0 0 1     %        2st structure: it cc a1 a2
a3 ang1 ang2 ang3 b

600 600 600                          % neighborhood for original data:
maximum search radii in the rotated system

0 0 0                                %                                  angles
for search ellipsoid

0                                    %                                  divide
into octants? 1=yes, 0=no

40                                    %                                  number
of data per octant (if octant=1) or in total

600 600 600                           % neighborhood for simulated nodes:
maximum search radii in the rotated system

0 0 0                                %                                   if
scattered locations: angles for search ellipsoid

0                                    %
divide into octants? 1=yes, 0=no

40                                    %
number of nodes per octant (if octant=1 and scattered) or in total

1                                    % kriging type: 1=SK, 2=OK

1                                    % SIS type: 1=traditional, 2=Bayesian
Updating, 3=non-stationary with local mean probability

100                                    % number of realizations

3                                    % number of refinements (multiple grid
simulation) (0=not used)

1                                    % random simulation sequence?
(1=yes,0=regular simulation sequence)

9236548                              % seed for random number generation
```

```
sisim_TEST_trad.out                         % name of output file

0                                  % create a GSLIB header in the output
file? 1=yes, 0=no


Available model types:

        1: spherical

        2: exponential

        3: gamma (parameter b > 0)

        4: stable (parameter b < 2)

        5: cubic

        6: Gaussian

        7: cardinal sine

        8: J-Bessel (parameter b > 0.5)

        9: K-Bessel (parameter b > 0)

       10: generalized Cauchy (parameter b > 0)

       11: exponential sine
```

A4. PYTHON CODE FOR BUILDING MULTINOMIAL LOGISTIC REGRESSION IN CASE STUDY II

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
#50 points
points = pd.read_excel('data.xlsx')
X = points[['X', 'Y', 'Z']]
Y = points['Category']
rs = 42
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size = 0.2,
 shuffle = True, random_state = rs)
clf = LogisticRegression(random_state = rs, multi_class = 'multinomial').f
it(X_train, y_train)
y_pred = clf.predict(X)
print(classification_report(y_pred, Y))
X_vis = X.copy()
X_vis['predictions'] = y_pred
X_vis['Categories'] = Y
X_vis.head()
y_vis = np.array(X_vis['predictions'])
points_pred = clf.predict(points[['X', 'Y', 'Z']])
points['predictions'] = points_pred
first_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 0]
second_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 1]
third_proba = clf.predict_proba(points[['X', 'Y', 'Z']])[:, 2]
points['1_proba'], points['2_proba'], points['3_proba'] = first_proba, sec
ond_proba, third_proba
arr_proba = []
probas = clf.predict_proba(points[['X', 'Y', 'Z']])
cnt = 0
for i in points['predictions']:
    arr_proba.append(probas[cnt, i - 1])
    cnt+= 1
```

```
points['proba_final'] = arr_proba
points.head()
#90 k
grid = pd.read_excel('grid.xlsx')
grid.head()
grid_pred = clf.predict(grid[['X', 'Y', 'Z']])
grid['predictions'] = grid_pred
first_proba = clf.predict_proba(grid[['X', 'Y', 'Z']])[:, 0]
second_proba = clf.predict_proba(grid[['X', 'Y', 'Z']])[:, 1]
third_proba = clf.predict_proba(grid[['X', 'Y', 'Z']])[:, 2]
grid['1_proba'], grid['2_proba'], grid['3_proba'] = first_proba, second_pr
oba, third_proba
arr_proba = []
probas = clf.predict_proba(grid[['X', 'Y', 'Z']])
cnt = 0
for i in grid['predictions']:
    arr_proba.append(probas[cnt, i - 1])
    cnt+= 1
grid['proba_final'] = arr_proba
grid
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred_test = model.predict(X_test)
y_predict = model.predict(grid[['X','Y', 'Z']])
grid['predict'] = y_predict
y_vis_2 = np.array(grid['proba_final'])
from matplotlib.colors import ListedColormap
cmap_bold_2 = ListedColormap(['none'])
fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,10))
plt.scatter(grid[grid.predict==1].X, grid[grid.predict==1].Y, c = grid[gri
d.predict==1].proba_final, cmap='jet')
#plt.scatter(points['X'], points['Y'], c = y_vis_2, cmap = 'jet', edgecolo
rs='w' )
plt.title('Category 1', size=20)
plt.xlabel('X', size=10)
plt.ylabel('Y', size=10)
cbar = plt.colorbar()
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1])
```

```
cbar.ax.set_ylim(bottom=0)
#cbar.ax.set_ylim(top=1)
plt.rcParams['axes.facecolor']='#03045c'
#plt.fill(150, 150, "b")
ax = plt.gca()
#ax.set_xlim([0, 300])
#ax.set_ylim([0, 300])
#plt.xticks(range(0,350,50))

fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,10))
plt.scatter(grid[grid.predict==2].X, grid[grid.predict==2].Y, c = grid[gri
d.predict==2].proba_final, cmap='jet')
#plt.scatter(points['X'], points['Y'], c = y_vis_2, cmap = 'jet', edgecolo
rs='w' )
plt.title('Category 2', size=20)
plt.xlabel('X', size=10)
plt.ylabel('Y', size=10)
cbar = plt.colorbar()
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1])
cbar.ax.set_ylim(bottom=0)
#cbar.ax.set_ylim(top=1)
plt.rcParams['axes.facecolor']='#03045c'
#plt.fill(150, 150, "b")
ax = plt.gca()
#ax.set_xlim([0, 300])
#ax.set_ylim([0, 300])
#plt.xticks(range(0,350,50))

fig = plt.figure(facecolor='blue')
plt.figure(figsize = (15,10))
plt.scatter(grid[grid.predict==3].X, grid[grid.predict==3].Y, c = grid[gri
d.predict==3].proba_final, cmap='jet')
#plt.scatter(points['X'], points['Y'], c = y_vis_2, cmap = 'jet', edgecolo
rs='w' )
plt.title('Category 3', size=20)
plt.xlabel('X', size=10)
plt.ylabel('Y', size=10)
cbar = plt.colorbar()
cbar.ax.set_yticklabels([0,0.2,0.4,0.6,0.8,1])
```

```
cbar.ax.set_ylim(bottom=0)
#cbar.ax.set_ylim(top=1)
plt.rcParams['axes.facecolor']='#03045c'
#plt.fill(150, 150, "b")
ax = plt.gca()
#ax.set_xlim([0, 300])
#ax.set_ylim([0, 300])
#plt.xticks(range(0,350,50))
```