

https://dspace.org/

# Multi-domain based mineral resource estimation in narrow stratabound tabular deposits

**Ndebele, Mpho**

**2025**

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/49822

Ndebele, Mpho. (2025). Multi-domain based mineral resource estimation in narrow stratabound tabular deposits. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/49822

Downloaded from a DSpace repository. https://dspace.org/

**School of Mining Engineering**



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

# MULTI-DOMAIN BASED MINERAL RESOURCE ESTIMATION IN NARROW STRATABOUND TABULAR DEPOSITS

Mpho Ndebele

**Student Number:** 2533114

**Supervisor:** Prof. Glen T. Nwaila

Date: 28 July 2025

A research report submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science.

Johannesburg, 2025

## DECLARATION

I declare that this report is my own, unaided work. I have read the University Policy on Plagiarism and confirm no plagiarism exists in this report. I also confirm that there is no copying or copyright infringement. I willingly submit to any investigation by the School of Mining Engineering in this regard, and I undertake to abide by the decision of any such investigation.



28 July 2025

Signature of Candidate

Date

## ABSTRACT

The application of geostatistical methods in mineral resource estimation has been at the core of the mining industry for decades. These methods have been extensively studied and their use in quantifying resources and reserves have resulted in many successful techno-economic valuation of mines. The models generated from these methods influence grade control protocols and ore extraction optimisation to realise economic benefits for the mine. Recently, there has been a rise in the integration of machine learning (ML) approaches for mineral resources estimation. This has been influenced by the increase in computational power, the production of large datasets from exploration and mining operations that have proven difficult to handle, and the growth in digitisation and automation applications in the mining and minerals industry. ML methods have shown in many studies that they provide an opportunity to produce high-accuracy models and the ability to decipher hidden patterns within the data that might have been difficult to model using traditional analytical methods. This research aimed to investigate the application of ML algorithms in generating mineral resource estimation domains used in the estimation of in-situ gold grades of the Carbon Leader Reef in the Carletonville Goldfield of the Witwatersrand Basin (South Africa). The ML algorithms used are Random Forest (RF), k Nearest Neighbour (kNN), AdaBoost (AB) and Support Vector Machine (SVM). The total number of primary samples before study selection were 5611 and after narrowing down the study area, they were 2017. The dataset used in this research comprises (a) sample spatial coordinates (X, Y, Z), (b) gold (Au) grade in grams per tonne (g/t), (c) gold accumulation expressed as Au (cm.g/t), (d) sedimentological properties (i.e., vertical thickness of the orebody commonly referred to as channel width (CW) measured in centimetres (cm), percentage conglomerate within the ore zone, pebble size, packing, roundness, sorting, pebble assemblages), and (e) mineralisation descriptions (i.e., carbon content, colour and basal contact). The integration of ML algorithms with traditional geostatistical methods in mineral resource evaluation workflows has shown positive and promising results. The optimal pointwise domain generation was done using traditional a k-means clustering and k-means coupled with signed distance functions and produced two mineral resource estimation domains. The traditional k-means clustering resulted in domain mixing which is unfavourable for resource estimation because it may lead to underestimation or overestimation of grades at these locations and subsequent misclassification of ore and waste. It also may complicate the generation of domain boundaries which is the next step after the pointwise domaining process, the boundary lines may appear obscure and incomprehensible. The signed distance-based k-means clustering model was selected as the best performing model because it resulted in unmixed and clear mineral resource estimation domains which is desirable for grade estimation. These signed distance-based k-means

domains were selected as the base of the domain boundaries. The RF model had the highest performance matrices scores, where both the accuracy and F1 scores were 99.9% while the kNN model had the lowest scores with accuracy and F1 scores were 99.4% each. Despite the RF performing better, the kNN model resulted in a more geologically reasonable domain boundary which is consistent with the geological contacts of the ore body under study. The domain boundary displayed an uneven geometry, typical of Carbon Leader Reef's in variable (e.g. high-to-low energy) depositional environments. These abrupt shifts in energy conditions affect how conglomerates are arranged, which in turn influences grade distributions. Following the definition of estimation domains and subsequent variography, the kNN boundaries were used to guide block ordinary Kriging for grade estimation. Post-estimation metrics showed that both the Kriging Efficiency (KEFF) and the Slope of Regression (SLOR) exceeded 50%, reflecting a favourable sample configuration and confirming the appropriateness of the selected domains and resultant boundaries for resource estimation.

**Keywords:** Domaining; Geostatistics; Gold Deposits; Kriging; Machine Learning; Mineral Resource Estimation

## ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor, Prof. Glen Nwaila. I would like to extend my gratitude to you for providing me with the necessary tools and support to execute this research. Above most thank you for introducing me to Data Science and ML applications in geosciences and mining. I would like to thank Sibanye-Stillwater for allowing me to use their data for this thesis. I believe that fostering development and achieving sustainable mining hinges on collaborative ventures between academia and the industry. I am honoured to be a beneficiary of such. I would like to send my deepest sincere gratitude to the Julian Baring Scholarship Fund for providing all the necessary financial support for my studies and this thesis. Your support has been the driving motivation and encouragement in this journey. I would like to recognise your efforts in empowering young African mining professionals through this academic fund. A big shout out to my family and friends for always rallying behind me, a constant reminder that there is purpose in everything you embark on. Continuous encouragement in nurturing the spirit of hard work and providing unwavering support. You guys are the real MVP.

# TABLE OF CONTENTS

LIST OF FIGURES ..... viii
LIST OF TABLES ..... xvi
ACRONYMS AND ABBREVIATIONS ..... i
1 INTRODUCTION ..... 1
1.1 Research Motivation ..... 2
1.2 Aim ..... 3
1.3 Objectives ..... 3
1.4 Data Context ..... 3
1.5 Research Scope ..... 4
1.6 Research Outline ..... 5
2 LITERATURE REVIEW ..... 7
2.1 Geology of the Witwatersrand Basin ..... 7
2.1.1 Witwatersrand gold fields ..... 8
2.1.2 Carletonville & West Rand gold fields ..... 9
2.1.3 Carletonville Carbon Leader Reef ore ..... 10
2.2 Mineral Resource Estimation ..... 11
2.2.1 Traditional mineral resource estimation workflow ..... 13
2.2.2 Geostatistics ..... 14
2.3 Domaining ..... 15
2.3.1 Types of domain boundaries ..... 15
2.4 Clustering ..... 16
2.4.1 K-Means Clustering ..... 17
2.5 Kriging Estimation ..... 19
2.5.1 Simple and Ordinary Kriging ..... 21
2.5.2 Variogram analysis ..... 22
2.5.3 Quantitative Kriging neighborhood analysis ..... 23
2.5.4 Kriging Performance Analysis ..... 24
2.6 Machine Learning Applications ..... 24
2.6.1 Traditional machine learning workflow ..... 26
2.6.2 Machine learning model selection ..... 27
2.6.3 Machine learning model performance ..... 28
2.6.4 Hybrid machine learning algorithms ..... 28
3 METHODOLOGY ..... 30
3.1 Datasets ..... 30
3.2 Exploratory Data Analysis ..... 30
3.3 Working Environment Setup ..... 30

3.4 Sample Point Domain Analysis ..... 32
3.5 Signed Distance-based Clustering ..... 33
3.6 Post Clustering Metrics ..... 34
3.6.1 Silhouette scores ..... 34
3.6.2 Davies-Bouldin index ..... 35
3.7 Domain Boundary Analysis ..... 35
3.8 Machine Learning Evaluation Metrics ..... 36
3.8.1 K-Fold cross-validation ..... 36
3.8.2 Confusion matrix ..... 36
3.9 Pre Kriging Process ..... 37
3.9.1 Z-Score transformation ..... 37
3.9.2 Normal score transformation ..... 38
3.9.3 Cell declustering ..... 39
3.9.4 Variography ..... 39
3.9.5 Quantitative Kriging neighbourhood analysis ..... 40
3.10 Kriging Block Modelling and Validation ..... 41
3.10.1 Kriging efficiency and slope of regression ..... 42
3.10.2 Declustered sample grades versus block estimation grades ..... 42
4 RESULTS ..... 44
4.1.1 Data preprocessing ..... 44
4.1.2 Study area selection ..... 46
4.2 Exploratory Data Analysis ..... 46
4.2.1 Categorial data conversion ..... 48
4.2.2 Feature analysis ..... 49
4.3 Point Sample Clustering Analysis ..... 54
4.4 Cluster Configuration ..... 54
4.5 Post Clustering Feature Summary Statistics Per Domain ..... 56
4.5.1 Z–Score transformation ..... 56
4.5.2 Box and whisker plot analysis of outlier treated data ..... 58
4.6 Domain Boundary Analysis ..... 63
4.6.1 kNN domain boundary ..... 64
4.6.2 Machine learning internal model parameter importance ..... 65
4.6.3 Machine learning algorithm performance ..... 65
4.7 Cell-based Declustering ..... 69
4.8 Normal Score Transformation ..... 71
4.8.1 Post domaining statistical evaluation ..... 75
4.9 Variography Analysis ..... 79

4.9.1 Fitted variogram models ..... 79
4.9.2 Variogram parameters ..... 82
4.10 Block Discretization ..... 84
4.10.1 Domain 0 discretised blocks ..... 84
4.10.2 Domain 1 discretised blocks ..... 84
4.11 Minimum and maximum number of samples ..... 85
4.11.1 Domain 0 minimum and maximum samples ..... 86
4.11.2 Domain 1 minimum and maximum samples ..... 86
4.12 Kriging Block Models ..... 87
4.13 Post Kriging Analysis ..... 88
5 DISCUSSION ..... 92
5.1 Point Sample Domain Analysis Outcome and Performance Analysis ..... 92
5.2 Machine Learning Domain Boundary Generation Analysis. ..... 93
5.3 Kriging Estimation Optimisation Through Quantitative Kriging Neighbourhood Analysis ..... 94
5.4 Ordinary Kriging Estimation and Block Modelling ..... 95
5.5 Post-Kriging Estimation Performance and Block Model Evaluation ..... 96
5.6 Machine Learning Potential for Clustering and Domain Boundary Generation in Mineral Resources Evaluation and Mineral Resources Management ..... 97
6 CONCLUSIONS ..... 100
7 RECOMMENDATIONS ..... 102
REFERENCES ..... 105
APPENDICES ..... 111

## LIST OF FIGURES

**Figure 1.1:** Location of the study area in Southern African(left) and a simplified geological map of the Witwatersrand Basin and associated goldfields (right). The red polygon corresponds to the location of the study area. Note the younger sequences covering the Witwatersrand Basin have been removed, figure modified from Frimmel & Nwaila, (2020). ..... 4

**Figure 1.2:** An outline indicating the levels of this research and the synergy of different geoscience fields. It includes mining geology (left), showing the onsite sampling type used in data collection, data management and storage and the grade determination methods. Data Science (middle) involves the use of Python as baseline computing language to do clustering and classification of domains. Resource Geology (right) involves the use of traditional geostatistics workflows that includes data pre-processing, quantitative Kriging analysis, ordinary Kriging, estimation and post estimation. ..... 5

**Figure 2.1:** Geological Setting of the Witwatersrand Basin, shown are the structural lineaments of the Colesburg magnetic lineament and Saddleback-Inyoka fault system (SIFS) that squeeze the basin and result in a semi elongated shape with a NE-SW orientation. Also, in shown is the Bushveld igneous complex located NE of the Witwatersrand basin modified from Frimmel & Nwaila, (2020). ..... 8

**Figure 2.2:** Witwatersrand Basin and the various Goldfields which occur in the contact regions between the older West Rand group of the Witwatersrand Supergroup and the younger Central Rand group 1) Welkom goldfield 2) Klerksdorp goldfield 3) Carletonville goldfield 4) West Rand goldfield 5) Central Rand goldfield 6) East Rand goldfield 7) South Rand goldfield 8) Evander goldfield modified from Frimmel & Nwaila, (2020). ..... 9

**Figure 2.3:** Carbon Leader Reef single conglomerate facies showing shorter and less diluted reef units on the (left) and, the multiple banded facies showing higher dilution and cyclic occurrences of individual bands modified from Shango Solutions (2007). ..... 10

**Figure 2.4:** Categorisation of resources and reserves indicating the transitional considerations between the two in internationally recognised reporting standards modified from (CRIRSCO, 2011; SAMREC, 2016). ..... 11

**Figure 2.5:** Mineral resources and reserves demarcation according to drill hole spacing confidence. Smaller drillhole spacing increases confidence from resources to reserves modified form Jowitt and McNulty (2021). ..... 12

**Figure 2.6:** Evolution of mineral resource estimation methods/tools over time showing the older and simper methods on the (lower left) and the recent and more sophisticated methods in the (top right) adapted from Dohm (2021). ..... 13

**Figure 2.7:** Integrated mineral resource estimation workflow which incorporates the 1) data importation of spatial and geological data 2) Data validation where imported data is interrogated and cleaned 3) Automated domaining using spatial data to generate underlying domain points 4) Domain boundary generation using geological knowledge 5) Domain estimation using declustering and Kriging estimation technique modified from Mariz et al., (2016). ..... 14

**Figure 2.8:** Common geological boundaries a) Soft boundary with no distinct break b)Hard boundaries showing hard breaks in between units c) A combination of both soft and hard boundaries which is more complicated to map modified from Kasmaee et al., (2019). ..... 16

**Figure 2.9:** Types of clustering methods 1) Density method that develops clusters based on the concentrations of points 2) Distribution method that uses probability of belonging to determine clusters at used data concentration to determine clusters 3) Centroid methods that determines clusters by the a set number of centroids and closer points to centroids belong to a cluster 4) Hierarchical method where a top down decision tree is used to group data into clusters modified from Geeks for Geeks, (2024). ..... 17

**Figure 2.10:** Steps involved in the K-means clustering process that includes 1) Process initiation 2) Centroid assignment in determining central points for cluster selection 3) Updating the previous step to select optimum K-clusters 4) Repeat the process for a number of iterations to generate optimal cluster numbers and sizes as modified from Sinaga & Yang, (2020). ... 18

**Figure 2.11:** A simplified graphical illustration of cluster generation using k-means cluster with no centroid but proximity of points to one another is used to assign them to a cluster modified from Sinaga and Yang (2020)..... 19

**Figure 2.12:** Semi variogram structure showing the sill, range and the nugget. It also shows the area of correlation and uncorrelated area that are regions beyond continuity adopted from Bohling, (2005)..... 23

**Figure 2.13:** ML categories and their respective algorithms (orange boxes) 1) Supervised learning methods that used labelled data for learning 2) Unsupervised learning methods use unlabelled data and no human interaction (independent methods) 3) Reinforcement learning methods that deal with adopted from Oladipupo, (2010). ..... 26

**Figure 2.14:** A simple ML modelling workflow that shows a summary of the process described above, the identification of problem and model outcome requirements are key in the use of ML models adopted from Andriy, (2019). ..... 27

**Figure 2.15:** ML algorithm selection flow showing the step-by-step process of selecting a model fit for purpose. ..... 28

**Figure 2.16:** Simplified workflow of the combination of super learning and Kriging models for spatial estimation strategy. Illustration, reconstructed from (Erdogan Erten et al., 2022). .... 29

**Figure 3.1:** Simple graphical illustration of methodology steps and modelling environment used at each step. ..... 30

**Figure 3.2:** Simple graphical illustration of Silhouette calculations and the cluster selection criteria (1) Cohesion demonstrated by D1 and D2 (2) Separation demonstrated by D3, D4 and D5 adopted from Saputra et al., (2020). ..... 34

**Figure 4.1:** Outline of the selected study area (dotted black square) which was trimmed from the whole dataset because of the density of data points around the area. ..... 46

**Figure 4.2:** Feature correlation heat map showing the feature correlation from the primary dataset. ..... 50

**Figure 4.3:** Au (g/t) graphical statistics and spatial distribution a) Spatial distribution of Au (g/t) values as shown in the study area, b) Histogram indicating frequency of the occurrence of Au (g/t) values c) Box plot showing the dispersion of outliers from the mean. ..... 51

**Figure 4.4:** CW (cm) graphical statistics and spatial distribution a) Spatial distribution of CW (cm) values as shown in the study area, b) Histogram indicating frequency occurrence of CW (cm) c) Box plot showing a balanced occurrence of high and low CW values. ..... 52

**Figure 4.5:** Conglomerate (%) graphical statistics and spatial distribution a) Spatial distribution of conglomerate (%) values as shown in the study area, b) Histogram indicating frequency occurrence of conglomerate (%) values c) Box plot showing well distributed conglomerate (%). ..... 53

**Figure 4.6**: Pebble size (cm) graphical statistics and spatial distribution a) Spatial distribution of Pebble size (cm) values as shown in the study area, b) Histogram indicating frequency occurrence of Pebble size (cm) values c) Box plot showing the distribution of pebble sizes which are concentrated around 20cm and a few high values. ..... 53

**Figure 4.7:** Cluster analysis a) Silhouette Scores b) Davies-Bouldin Index c) Traditional K-means clustering distribution d) Signed-distance-mds-K-means clustering distribution. ..... 54

**Figure 4.8:** Cluster configuration c) Traditional K-means cluster layout showing cluster mixing. d) Signed-distance-mds-k-means cluster layout showing two distinct clusters. ..... 55

**Figure 4.9:** Histogram and summary statistics for individual domains a) Domain 1 b) Domain 0..... 55

**Figure 4.10:** Feature histogram and summary statistics per domain a) Au (cm.g/t) Domain 0 b) Au (cm.g/t) Domain 1 c) CW (cm) Domain 0 d) CW (cm) Domain 1 e) Au (g/t) Domain f) Au (g/t) Domain 1. ..... 56

**Figure 4.11:** Feature Z–score histograms and the summary statistics per domain a) Au (cm.g/t) Z-score (Domain 0) b) Au (cm.g/t) Z-score (Domain 1) c) CW (cm) Z-score (Domain 0) d) CW (cm) Z-score (Domain 1) e) Au g/t Z-score (Domain 0) f) Au (g/t) Z-score (Domain 1). ..... 57

**Figure 4.12:** Outlier treated data represented as Whisker plots and summary statistics a) Au (cm.g/t) Domain 0 b) Au (cm.g/t) Domain 1 c) CW (cm) Domain 0 d) CW (cm) Domain 1 e) Au (g/t) Domain 1 f) Au (g/t) Domain 1..... 59

**Figure 4.13:** Pairwise plots for all the selected variables of interest in Domain 0. ..... 60

**Figure 4.14:** Pearsons's correlation heatmap of selected variables of interest Au (cm.g/t), Au (g/t) and CW (cm) and their spatial locations (Domain 0). ..... 61

**Figure 4.15:** Pairwise plots for all the selected variables of interest in Domain 1. ..... 62

**Figure 4.16:** Pearson correlation heatmap of selected variables of interest Au (cm.g/t), Au (g/t) and CW (cm) and their spatial locations at Domain 1. ..... 63

**Figure 4.17:** Domain boundaries generated by various ML algorithms a) kNN domain boundary b) SVM domain boundary c) RF domain boundary d) AB domain boundary. ..... 64

**Figure 4.18:** Domain boundary polyline showing demarcation of Domain 0 and Domain 1 features a) Au (g/t) spatial distribution b) Au (cm.g/t) spatial distribution c) CW (cm) spatial distribution. ..... 65

**Figure 4.19:** a) Line graph illustration of average Accuracy and F1 scores across the selected ML algorithms. ..... 68

**Figure 4.20:** ML confusion matrix from the ML algorithms showing predicted vs. true labels a) kNN confusion matrix results b) SVM confusion matrix results c) SVM manually tuned results d) RF confusion matrix results e) AB confusion matrix results. ..... 68

**Figure 4.21:** Cell size selection for Domain 1 Au (cm.g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 69

**Figure 4.22**: Cell size selection for Domain 0 Au (cm.g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 69

**Figure 4.23:** Cell size selection for Domain 1 CW (cm) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 70

**Figure 4.24:** Cell size selection for Domain 0 CW (cm) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 70

**Figure 4.25:** Cell size selection for Domain 1 Au (g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 71

**Figure 4.26:** Cell size selection for Domain 0 Au (g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size. ..... 71

**Figure 4.27**: a) Domain 1 declustered Au (g/t) b) Domain 1 declustered normal score A (g/t) c) Domain 1 declustered Au (g/t) spatial distribution d) Domain 1 normal score Au (g/t) spatial distribution. ..... 72

**Figure 4.28:** Domain 1 declustered CW (cm) b) Domain 1 declustered normal score CW (cm) c) Domain 1 declustered CW (cm) spatial distribution d) Domain 1 normal score CW (cm) spatial distribution. ..... 73

**Figure 4.29:** Domain 1 declustered Au (cm.g/t) b) Domain 1 declustered normal score Au (cm.g/t) c) Domain 1 declustered Au (cm.g/t) spatial distribution d) Domain 1 normal score Au (cm.g/t) spatial distribution. ..... 73

**Figure 4.30:** Domain 0 declustered Au (g/t) b) Domain 0 declustered normal score Au (g/t) c) Domain 0 declustered Au (g/t) spatial distribution d) Domain 0 normal score Au (g/t) spatial distribution. ..... 74

**Figure 4.31:** Domain 0 declustered CW (cm) b) Domain 0 declustered normal score CW (cm) c) Domain 0 declustered CW (cm) spatial distribution d) Domain 0 normal score CW (cm) spatial distribution. ..... 74

**Figure 4.32:** Domain 0 declustered Au (cm.g/t) b) Domain 0 declustered normal score Au (cm.g/t) c) Domain 0 declustered Au (cm.g/t) spatial distribution d) Domain 0 normal score Au (cm.g/t) spatial distribution. ..... 75

**Figure 4.33:** a) Domain 0 Au (cm.g/t) histogram b) Domain 0 Q-Q Au (cm.g/t) plot c) Au (cm.g/t) treated data histogram d) Au (cm.g/t) treated data Q-Q plot. ..... 76

**Figure 4.34:** Domain 0 Au (g/t) histogram b) Domain 0 Q-Q Au (g/t) plot c) Domain Au (g/t) treated data histogram d) Domain Au (g/t) treated data Q-Q plot. ..... 76

**Figure 4.35:** Domain 0 CW (cm) histogram b) Domain 0 Q-Q CW (cm) plot c) Domain 0 CW (cm) treated data histogram d) Domain 0 CW (cm) treated data Q-Q plot. ..... 77

**Figure 4.36:** Domain 1 Au (cm.g/t) histogram b) Domain 1 Q-Q Au (cm.g/t) plot c) Domain 1 Au (cm.g/t) treated data histogram d) Domain 1 Au (cm.g/t) treated data Q-Q plot..... 77

**Figure 4.37:** Domain 1 Au (g/t) histogram b) Domain 1 Q-Q Au (g/t) plot c) Domain 1 Au (g/t) treated data histogram d) Domain 1 Au (g/t) treated data Q-Q plot..... 78

**Figure 4.38:** Domain 1 CW (cm) histogram b) Domain 1 Q-Q CW (cm) plot c) Domain 1 CW (cm) treated data histogram d) Domain 1 CW (cm) treated data Q-Q plot. ..... 78

**Figure 4.39**: Fitted Variogram Model Domain 0 a) Domain 0 Au (cm.g/t) major angle 30 and minor angle 120 b) Domain 0 CW (cm) major angle 15 and minor angle 105 C) Domain 0 Au (g/t) major angle 90 and minor angle 0..... 80

**Figure 4.40:** Fitted variogram model Domain 1 a) Au (cm.g/t) major angle 90 and minor angle 0 b) CW (cm) major angle 75 and minor angle 165 C) Domain 1 Au (g/t) major angle 165 and minor angle 75. ..... 81

**Figure 4.41:** Block discretisation a) Domain 0 Au (cm.g/t) b) Domain 0 Au (g/t) c) Domain 0 CW (cm)..... 84

**Figure 4.42:** Block discretisation results per domain a) Domain 1 Au (cm.g/t) b) Domain 1 Au (g/t) c) Domain 1 CW (cm). ..... 85

**Figure 4.43:** Quantitative Kriging neighbourhood a) Domain 0 Au (cm.g/t) minimum and maximum number of samples b) Domain 0 Au (g/t) minimum and maximum number of samples c) Domain 0 CW (cm) minimum and maximum number of samples. ..... 86

**Figure 4.44:** QKNA number of sample search selection a) Domain 1 Au (cm.g/t) minimum and maximum number of samples b) Domain 1 Au (g/t) minimum and maximum number of samples c) Domain 1 CW (cm) minimum and maximum number of samples. ..... 86

**Figure 4.45:** 2D map showing the spatial distribution of Au (cm.g/t) Kriging results. ..... 87

**Figure 4.46:** 2D map showing the spatial distribution of CW (cm) Kriging results. ..... 87

**Figure 4.47:** 2D map showing the spatial distribution of KEFF results. ..... 88

**Figure 4.48:** Histogram showing the KEFF distribution depicting a bimodal distribution. .... 88

**Figure 4.49:** 2D map showing the spatial distribution of SLOR results. ..... 89

**Figure 4.50:** Histogram showing the SLOR distribution depicting a mixture of values across domains. ..... 89

**Figure 4.51:** X-axis swath plot showing the estimated values compared to actual grades along the E-W direction representing number of samples, samples Au (cm.g/t) and model Au (cm.g/t)..... 90

**Figure 4.52**: Y-axis swath plot showing the estimated values compared to actual grades along the N-S direction representing number of samples, samples Au (cm.g/t) and model Au (cm.g/t). ..... 90

**Figure 4.53:** X-axis swath plot showing the estimated values compared to actual grades along the E-W direction representing number of samples, samples CW (cm) and model values CW (cm). ..... 91

**Figure 4.54:** Y-axis swath plot showing the estimated values compared to actual grades along the N-S direction representing number of samples, samples CW (cm) and model CW (cm). ..... 91

**Figure B1:** Experimental variograms a) Domain 0 Au (cm.g/t) Azimuth 0 b) Domain 0 Au (cm.g/t) Azimuth 15 c) Domain 0 Au (cm.g/t) Azimuth 30 d) Domain 0 Au (cm.g/t) Azimuth 45……113

**Figure B2:** Experimental variograms a) Domain 0 Au (g/t) Azimuth 0 b) Domain 0 Au (g/t) Azimuth 15 c) Domain 0 Au (g/t) Azimuth 30 d) Domain 0 Au (g/t) Azimuth 45. ..... 113

**Figure B3:** Experimental variograms a) Domain 0 CW (cm) Azimuth 0 b) Domain 0 CW (cm) azimuth 15 c) Domain 0 CW (cm) azimuth 30 d) Domain 0 CW (cm) azimuth 45. ..... 114

**Figure B4:** Experimental variograms a) Domain 1 Au (cm.g/t) azimuth 0 b) Domain 1 Au (cm.g/t) azimuth 15 c) Domain 1 Au (cm.g/t) azimuth 30 d) Domain 1 Au (cm.g/t) azimuth 45. ..... 114

**Figure B5**: Experimental variograms a) Domain 1 Au (g/t) azimuth 0 b) Domain 1 Au (g/t) azimuth 15 c) Domain 1 Au (g/t) azimuth 30 d) Domain 1 Au (g/t) azimuth 45. ..... 115

**Figure B6:** Experimental variograms a) Domain 1 CW (cm) azimuth 0 b) Domain 1 CW (cm) azimuth 15 c) Domain 1 CW (cm) azimuth 30 d) Domain 1 CW (cm) azimuth 45. ..... 115

## LIST OF TABLES

**Table 4.1:** Spatial locations (X, Y and Z coordinates) showing the first 5 points. ..... **Error! Bookmark not defined.**

**Table 4.2:** Features from the primary data including grade data and sedimentological data showing the first 5 rows of the data matching the above coordinates.**Error! Bookmark not defined.**

**Table 4.3:** Table showing the first five data columns for grade values Au (g/t) (highlighted in blue lower right corner) values converted by dividing Au (cm.g/t) (highlighted in yellow) and the CW (cm) (highlighted in orange)..... 47

**Table 4.4**: Data types and number of null values per variable. ..... 48

**Table 4.5:** Python extract of the encoded values from the categorical data showing the first 5 columns of the data (represented in 0,1 and 2). ..... 49

**Table 4.6:** Summary statistics for the selected numerical features. ..... 49

**Table 4.7:** Z-Score values (highlighted in orange) generate for selected feature per domain (Domain 0 highlighted in blue and Domain 1 highlighted in yellow). ..... 58

**Table 4.8:** ML algorithm performance (accuracy score and F1 score) and internal best parameters..... **Error! Bookmark not defined.**

**Table 4.9:** ML Accuracy and F1 scores for 10 iterations. ..... 67

**Table 10:** ML Accuracy and F1 average calculations per model..... 67

**Table 4.10:** Variogram parameters per domain for each variable of interest..... 83

## ACRONYMS AND ABBREVIATIONS

| | |
|---|---|
| **AB** | AdaBoost (machine learning algorithm) |
| **Au** | Gold |
| **Cm.g/t** | Centimetre-grams per ton |
| **CW (cm)** | Channel width |
| **KEFF** | Kriging efficiency |
| **kNN** | k-Nearest Neighbour (machine learning algorithm) |
| **MDS** | Multi dimensionality scaling |
| **ML** | Machine learning |
| **MLA** | Machine learning algorithm |

| | |
|---|---|
| **RF** | Random Forest (machine learning algorithm) |
| **SLOR** | Slope of regression |
| **SVM** | Support Vector Machine (machine learning algorithm) |
| **SVM_man** | Manually tuned Support Vector Machine (machine learning algorithm) |

# 1 INTRODUCTION

Mineral resource estimation is the process of qualitatively and quantitatively assessing the mineral content of a geological deposit (Rossi and Deutsch, 2013). The implementation of mineral resource estimation in mining operations forms the basis of mine planning and mine designs used in ore extraction and beneficiation optimisation (Glacken *et al.*, 2001). A proper mineral resource estimation is important because it has an economic impact of any exploration project or mining operation. Mineral resource estimation has been a part of the mining industry for many years and has since evolved with increased computational power and data abundance within the mining space (Dumakor-Dupey and Arya, 2021). Traditional mineral resources estimation methods have been used in the past, even before the use of computers. Most of these methods involved the use of maps and drawings where polygons were drawn by hand to either interpolate or extrapolate close-by data points by some form of geospatial relationships (Dominy and Annels, 2001). Geostatistical methods are used to interpolate values in unsampled areas or zones of poor data knowledge (Patterson, 1959; King, 1982; Stone & Dunn, 1996). Some of the old methods include the sectional method, polygonal method, triangular method, regular & random stratified grid method, inverse distance weighting distance, and contouring methods.

Geostatistical methods involve the application of advanced statistics in the geospatial context (Sabourin, 1983). They operate under the assumption that sample values selected for the estimation process are representative of the larger population they were selected from. Matheron, (1963) introduced the concept of regionalized variable theory of geostatistics and emphasises on the randomness and population representation structures in selecting samples for geostatistical analysis. This fundamental geostatistics principle accounts for spatial correlation and variability of geological data within a deposit. Geostatistical methods are effective when dealing with irregularly spaced data which can be difficult to analyse visually or when using traditional estimation methods. Geostatistical methods were made famous by Danie G. Krige when he developed the “Kriging Estimation Method” to address highly variable mineral grades of the Witwatersrand gold deposits (Gómez-Hernández, 2005). Kriging is an interpolation method for spatial data analysis and spatial prediction of variables over a selected distance and direction (Bostan, 2017). It uses variogram models to measure the spatial dependence of selected pairs of samples at a given distance (Oliver and Webster, 2015). In mineral resources, the variogram is used to measure the extent of continuity and variability of the mineralization which is vital for the quantification of mineral grades (Bargawa and Amri, 2016).

Domaining is an important component of a successful mineral resource estimation process and is based on the principle of stationarity, in which the mean and variance of a population should be constant in space (Armstrong, 1998; Journel & Huijbregts, 1976). Domaining is the process of using geoscientific data such as logging information from drill cores to demarcate the orebody into homogenous and continuous units (Jowitt & McNulty, 2021). This process is fundamental for mineral resource estimation, yet there is still prevalent occurrence of resource downgrades and project devaluations due to poor domain identification (Sterk *et al*., 2019). Poor domaining causes data mixing from different data populations than may not be related. This mixing of data results in the poor understanding the spatial distribution of grades due to data misrepresentation. Emery and Ortiz, (2004) define grade domaining as the partitioning of a mineralised body defined by ore grade intervals. Geodomaining and grade domaining are standard in the industry and the latter is common in gold deposits. These methods are part of many if not all commercial modelling software packages as either explicit or implicit modelling approaches that if compared against each other, generally result in variances of the grades and volumes estimated (Cowan *et al.,* 2002). Structural geometry and the orebody's genetic model are key components in the creation of domains and require a careful approach (Sterk *et al*., 2019).

Due to the increase in computational power over the last decade, there has been an increase in the development and application of more robust and effective computing methods (Karpatne *et al.,* 2018; Lary *et al.,* 2016). Machine learning (ML) as a subset of artificial intelligence (AI) is being applied in a variety of geoscientific projects (Mery and Marcotte, 2022). For mineral resource evaluation industry, ML and AI have been introduced as either replacement of classical estimation techniques (Dumakor-Dupey and Arya, 2021) or integrated with geostatistical methods, thus creating hybrid methods (Dumakor-Dupey and Arya, 2021). ML and AI algorithms can provide insights into hidden data trends that may have been difficult to do using traditional methods and manual. Most ML and AI methods are at the research and development stages, still necessitating further development and refinement in terms of applicability, standardisation and reporting within the context of minerals industry (McCoy and Auret, 2019). Where ML are applied, methods have shown positive and promising results in terms of accuracy, precision, and the ability to handle large and high dimensional datasets (Dumakor-Dupey and Arya, 2021; Zhang *et al*., 2023).

### 1.1 Research Motivation

The use of ML and AI for mineral resource estimation is a relatively new, especially for gold deposits found within the Witwatersrand Basin of South Africa. Primarily, the use of ML for the creation of domains has just recently captured interest in the research field (Cevik *et al*., 2021;

Zhang *et al*., 2023). Poor domaining of orebodies has ramifications, where when poorly executed it may lead to underestimation or overestimation of ore grades (Sinclair and Blackwell, 2006). The Witwatersrand Basin gold deposits are dominated by the nugget effect (Dominy *et al*., 2001), as such, estimating the deposits has proven to be complex and proper domaining is key in addressing this challenge. Overestimation or underestimation of mineral resources has an impact in the business case of exploration and mining operations mainly due to under recovery of gold and this subsequently impacts revenue generated.

### 1.2 Aim

The aim of this research is to apply machine learning algorithms to define mineral resource estimation domains for the Carbon Leader Reef in the Driefontein gold mining complex (Witwatersrand Supergroup, South Africa), which will then be used for in-situ gold grade estimation via ordinary Kriging.

### 1.3 Objectives

The objectives of the research are to:

I. Apply machine learning clustering and classification algorithms using spatial and geological properties to enhance domain selection.
II. Assess ore continuity and directional variation using variograms, then apply ordinary Kriging to estimate grades in unsampled areas while capturing spatial variability.
III. Evaluate the effectiveness of integrating ML algorithms by analysing Kriging outcomes with slope of regression (SLOR) and Kriging efficiency (KEFF).

### 1.4 Data Context

The datasets used in this study were obtained from the Driefontein gold mining complex, owned by Sibanye-Stillwater, one of the world's leading international precious metals mining companies with a diverse asset portfolio. It is located about 70 km west of Johannesburg, near the town of Carletonville (Fig. 1.1), and is part of the Central Rand Group in the Witwatersrand Basin, South Africa. Driefontein is a mature mine that has been operational since 1952. The data used in this study came from channel sampling of the conglomeritic and carbon seams hosted in the Carbon Leader Reef; in this context, “reef” refers to the orebody.

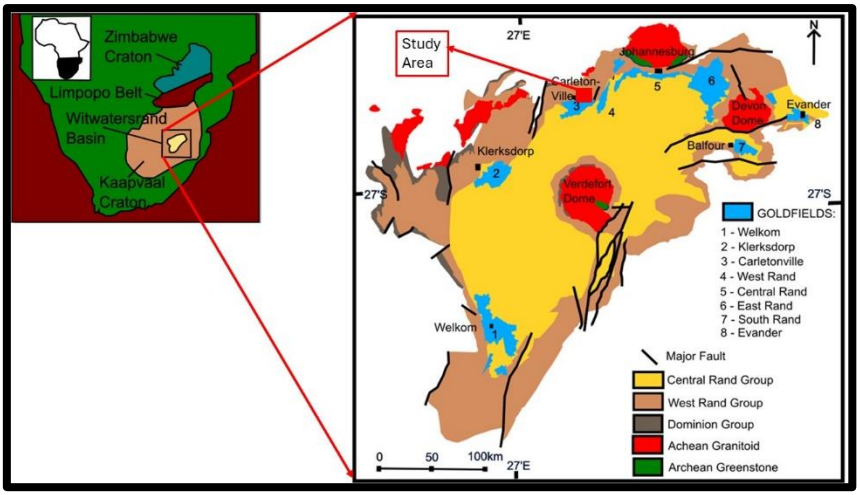


**Figure 1.1:** Location of the study area in Southern African(left) and a simplified geological map of the Witwatersrand Basin and associated goldfields (right). The red polygon corresponds to the location of the study area. Note the younger sequences covering the Witwatersrand Basin have been removed, figure modified from Frimmel & Nwaila, (2020).

## 1.5 Research Scope

This research is a trans-discipline project that incorporates mining geology, mineral resource estimation and modern data analytics. This study explores the use of ML algorithms in the development of mineral resource domains to facilitate the estimation process. The database contains geoscientific information (survey data, grade distribution data and sedimentological data). The database will then be analysed via Python programming language (Jupyter Notebook) and gold grade estimations will be done using the Datamine commercial modelling software. The ML algorithms will be used to build resource domains, in turn used to predict of gold grades in unsampled locations. The performance of these ML algorithms will be assessed using F1 and confusion matrices. Upon completion of the ordinary Kriging process, the SLOR, KEFF and swath plots will be done to draw a conclusion on their applicability of these ML algorithms in generating domains used in the estimation process. Below is the research scope summary indicating interdisciplinary fields to be applied in this research (Fig 1.2).

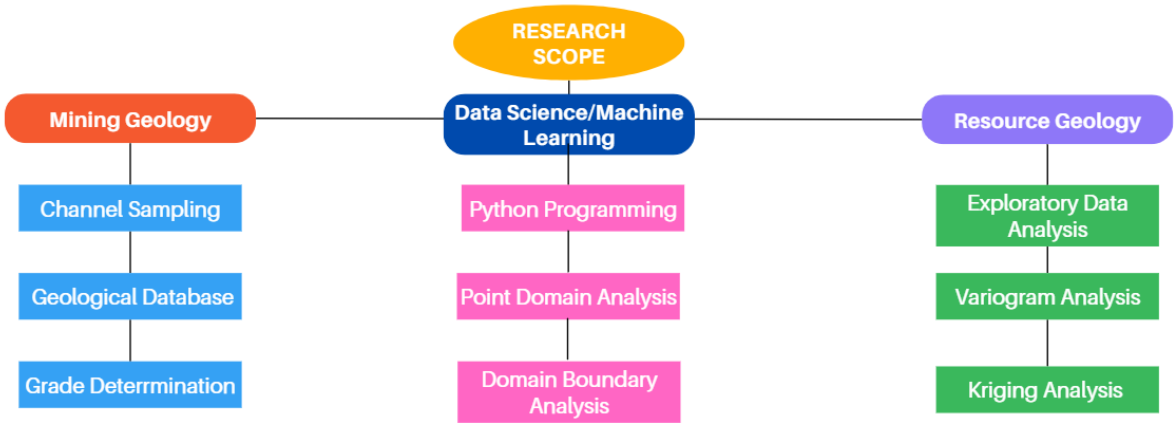


**Figure 1.2:** An outline indicating the levels of this research and the synergy of different geoscience fields. It includes mining geology (left), showing the onsite sampling type used in data collection, data management and storage and the grade determination methods. Data science (middle) involves the use of Python as baseline computing language to do clustering and classification of domains. Resource geology (right) involves the use of traditional geostatistics workflows that includes data pre-processing, quantitative Kriging analysis, ordinary Kriging, estimation and post estimation.

## 1.6 Research Outline

The research is structured as follows:

**Introduction (Chapter 1):** This chapter provides the project overview. It includes the introduction where emphasis is on the project questions and background. Aims and objectives are defined within this chapter to contextualise the need for this research. Finally, the study area and scope of the study are introduced to describe the framework of the research.

**Literature review (Chapter 2):** This chapter encompasses the review of already existing literature on the subject. The study area in the geographical and geological context are also discussed in this chapter. Application analysis of both traditional geostatistical techniques and ML algorithms in geosciences with emphasis on mineral resource estimation is also done in this chapter.

**Methodology (Chapter 3):** This chapter involves the discussion of the proposed methodology and the terms of reference for the research project. This chapter carefully outlines the project data analysis process carried out to yield the desired results.

**Results (Chapter 4):** This is the results presentation chapter, where various visualisation methods are used to display the outcomes of the previous chapter. This will be done in Jupyter notebook using the python programming language.

**Discussion (Chapter 5):** This chapter that involves addressing the key assumption and related aims and objectives of the research project. It also correlates the results outcome of the previous section with some literature to draw up informed conclusions.

**Conclusion (Chapter 6):** This chapter, and it involves the conclusions of the project and reviews the success of addressing the aims and objectives of the study.

**Recommendations (Chapter 7):** This chapter critically examines the inherent limitation of current machine learning methodologies in resource estimation and proposes forward-looking strategies to enhance model robustness, scalability, and domain-specific applicability.

# 2 LITERATURE REVIEW

This chapter focuses on the review of available literature on the regional geological setting of the study area, the Witwatersrand Basin. The various goldfields are discussed, and further emphasis is on the West Rand Goldfield which hosts the Carbon Leader Reef, the orebody under investigation in this study. The mineral resource estimation process is reviewed and various interpolation and estimation techniques used in the exploration and mining industry are reviewed as well. Finally, the ML process and application in geoscientific problems are provided with focal point on mineral resource estimation are provided herein.

## 2.1 Geology of the Witwatersrand Basin

The Witwatersrand Basin is arguably the biggest producer of gold in the world shown in (Fig 2.1) below. The production of gold in the Witwatersrand Basin dates back 120 years before the Anglo-Boer war and continued through to the 1st World War and the 2nd World War (Frimmel and Minter, 2002). Metallogeny and sedimentological processes studies of the Witwatersrand Basin has guided the understanding of the origin and formation of these large gold deposits (Dankert & Hein, 2010). The current mineral reserves are estimated at 40kt and from the geophysical and drilling results it is indicated that the ore occurs in reefs that go as deep as 5-6 km, extensive exploration has indicated that these gold fields extend more than 100 km in length (Cairncross, 2021). Gold orebodies are considered hard to find because of their narrow occurrences. Their occurrence has been localized through drilling and continuous geophysical studies within the goldfields (Phillips, 2011). The Witwatersrand Basin is dated at 3.074-2.78 Ga and is largely made up of sedimentary and sub-volcanic rock (Frimmel *et al*., 2005). It is situated in the Kaapvaal Craton, an ancient portion of thick continental crust within the Southern African continental structure aged between 3.7-2.7 Ga (Robb and Meyer, 1995). The Witwatersrand basin is underlain by granitic basement and widespread greenstone belts hosting much smaller vein-hosted gold deposits which may be remains of a much bigger primary gold source (Heinrich, 2015). This basin is well known for its high-grade tabular reefs that vary in size from a few centimetres which have been targeted for extensive exploration and mining over the years. The mineralized zones of the Witwatersrand basin are made of quartz-pebble-conglomerate hosting both gold and uranium. It is known to have produced roughly a third of the world's gold supply and hosts significant uranium deposits (Frimmel and Minter, 2002). More than 80% of the Witwatersrand gold is concentrated in the <2.90 to >2.78 Ga Central Group and less than 5% of the <2.98 to 2.91 Ga hosted within the West Rand (Frimmel *et al.,* 2009).

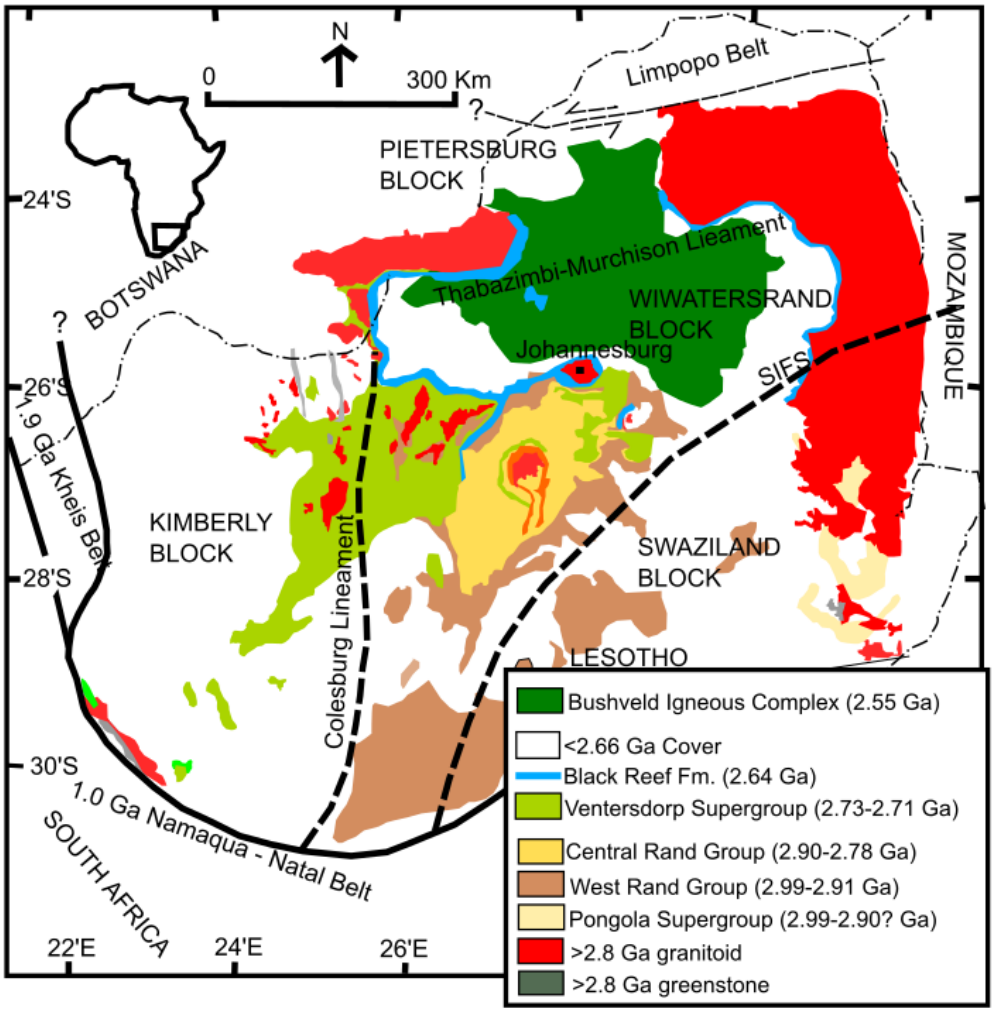


**Figure 2.1:** Geological Setting of the Witwatersrand Basin, shown are the structural lineaments of the Colesburg magnetic lineament and Saddleback-Inyoka fault system (SIFS) that squeeze the basin and result in a semi elongated shape with a NE-SW orientation. Also, in shown is the Bushveld igneous complex located NE of the Witwatersrand basin modified from Frimmel & Nwaila, (2020).

### 2.1.1 Witwatersrand gold fields

The Witwatersrand Basin is well known for its abundant gold occurrences and mining occurs within the predominant eight goldfields. There are some sporadic occurrences of the gold bearing conglomerates across the basin shown in (Fig 2.2) below.

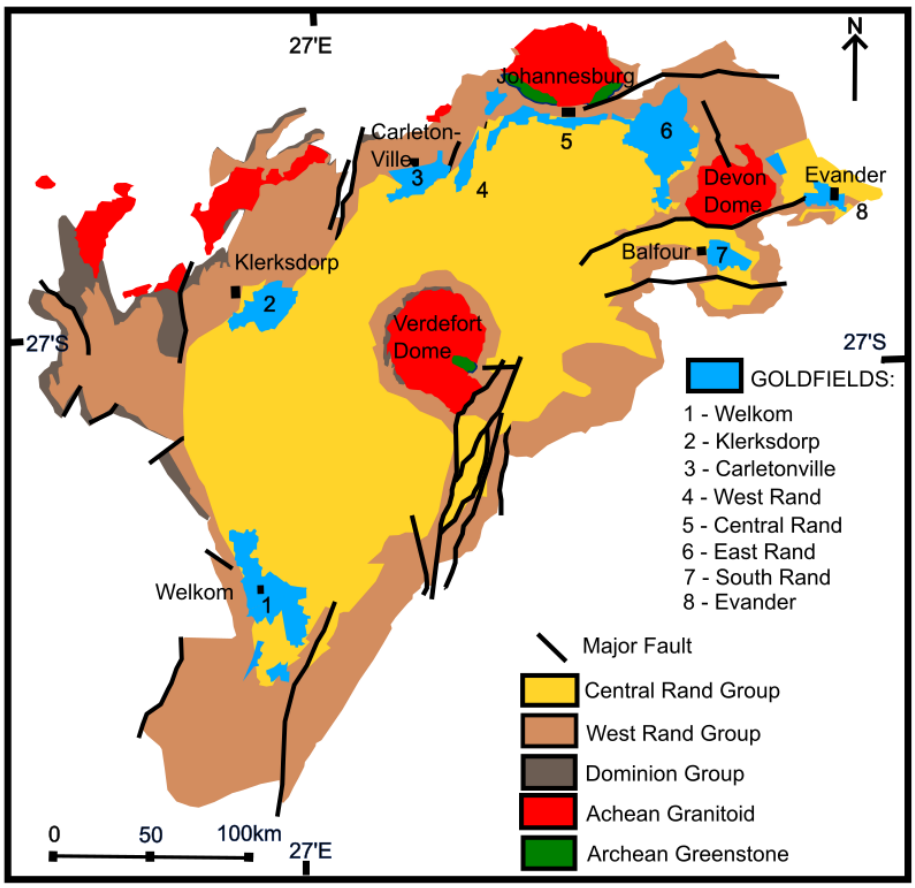


**Figure 2.2:** Witwatersrand Basin and the various Goldfields which occur in the contact regions between the older West Rand group of the Witwatersrand Supergroup and the younger Central Rand group 1) Welkom goldfield 2) Klerksdorp goldfield 3) Carletonville goldfield 4) West Rand goldfield 5) Central Rand goldfield 6) East Rand goldfield 7) South Rand goldfield 8) Evander goldfield modified from Frimmel & Nwaila, (2020).

### 2.1.2 Carletonville & West Rand gold fields

The Carletonville Goldfield is located along the Northen margin of the Witwatersrand Basin shown in (Fig 2.2) above. The West Rand is found South of Carletonville, they are separated by the West Rand Fault (Frimmel & Nwaila, 2020). The Transvaal Supergroup rocks overlain these two goldfields and the pervasive occurrence of interlayering Ventersdorp Supergroup within these Units (Frimmel & Nwaila, 2020). The primary reefs of these Goldfields are the Main and Carbon Leader Reefs and they have the highest mining significance due to the occurrence of high-grade gold deposits (Frimmel et al., 2009; Frimmel & Nwaila, 2020). The Carbon Leader Reef is mined out (Frimmel & Nwaila, 2020) and its average grade was 12.6 g/t in the West Wits Goldfield compared to the average grade of the whole Witwatersrand basin at 8.4 g/t (Tucker *et al*., 2016). The West Rand is relatively small compared to other goldfields and the auriferous and uriniferous reefs have been mined and they overlay the

Dominion Group (Safonov & Prokof'ev, 2006). These conglomerate reefs mostly occur mainly above low angle unconformities (Frimmel & Nwaila, 2020) and significant gold is found where these underlaying units are truncated (Safonov & Prokof'ev, 2006; Dankert & Hein, 2010). Within these goldfields there is the Elsburg, Livingstone and Middelvlei reefs characterised by conglomerates intercalated with quartzite and upward coarsening of units in the Elsburg Reefs which have future potential for mining companies within the area (Frimmel & Nwaila, 2020).

### 2.1.3 Carletonville Carbon Leader Reef ore

This study focuses on the Carbon Leader Reef of the Carletonville goldfields. It is located approximately 100m above the base of the Johannesburg Subgroup and is considered he oldest gold rich unit within the Carletonville goldfield. Majority of the mines within the Carletonville Goldfield occur west of the major Bank Fault. It normally comprises of a carbon rich layer approximately 100mm in thickness. On top is an oligomictic and closely packed pebble unit rich in interstitial carbon matrix (Els, 1988) and is largely mined out within the area (Frimmel & Nwaila, 2020). The Carbon Leader Reef occurs as single bands referred to as single conglomerate facies (SFC) or multi bands hosted referred to as multiple banded facies (MBF) shown on (Fig 2.3) below, they occur within the quarzitic conglomerates giving lateral variations of ore width (Phillips and Law, 2000). It is characterised by a scour surface, evidence of a regional extensive angular unconformity resulting in truncation of underlaying strata towards the northern portion of the goldfield (Buck and Minter, 1985).

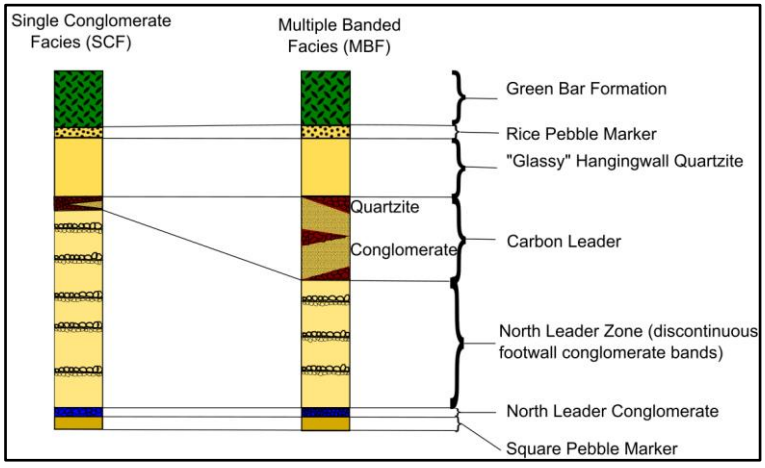


**Figure 2.3:** Carbon Leader Reef single conglomerate facies showing shorter and less diluted reef units on the (left) and, the multiple banded facies showing higher dilution and cyclic occurrences of individual bands modified from Shango Solutions (2007).

## 2.2 Mineral Resource Estimation

The quantification of minerals and metals in the ground is an important aspect of any mining operations as it gives insight on the size and the shape of the orebody in the sub surface needed for mine planning and mine designs (Rossi and Deutsch, 2013). A mineral resource is defined as a naturally occurring material with substantial concentration in the earth's crust that presents economic value (Dominy *et al.,* 2001). A mineral reserve represents a fraction of a mineral resource that can be economically extracted and converted to useful products (Meinert *et al.,* 2016). A mineral resource can be categorised as an inferred, measured and indicated whereas a mineral reserve can be categorised as probable and proved resource depending on the level of confidence and the consideration of the modifying factors (Fig 2.4).

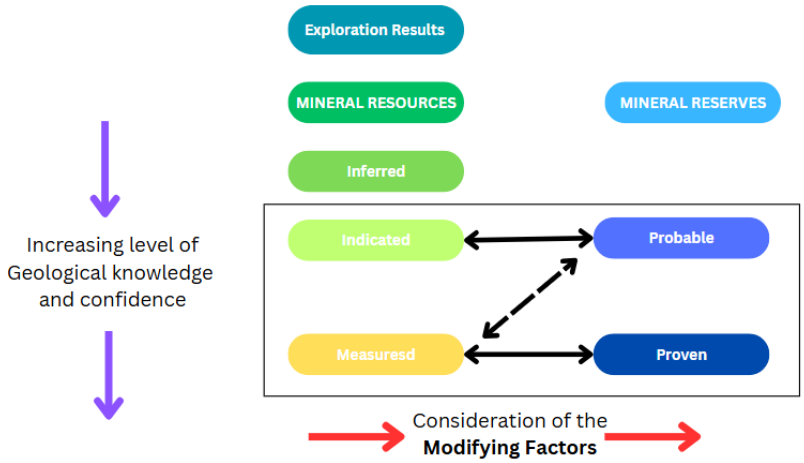


**Figure 2.4:** Categorisation of resources and reserves indicating the transitional considerations between the two in internationally recognised reporting standards modified from (CRIRSCO, 2024; SAMREC, 2016).

The resource confidence is increased by reducing the drill hole spacing to narrow down the resource location (Woodall, 1984). Fig 2.5 below demonstrates the demarcation of resource and reserve based on drilling confidence. Mineral resource estimation generates mineral resource value as it confirms the extent of the mineralization and the quantity to be exploited (Glacken *et al*., 2001). During estimation consideration is put into the spatial variability and the geological uncertainty associated with deposits of various complexities (Rossi & Deutsch, 2013). This is important as it provides proper quantification of the level of confidence in the estimation process.

Mineral resource estimation results are published in line with various reporting guidelines, depending on the jurisdiction of the operation. The internationally recognized reporting standards based on country of practice are; the Joint Ore Reserve Committee (JORC) of

Australia (JORC, 2012), NI-43-101 of Canada (Rendu and Miskelly, 2001), the Committee For Mineral Reserves International Reporting Standards (CRIRSCO, 2024) and the South African Mineral Reporting Codes (SAMREC, 2016). All of these reporting standards share a similar goal for the generation of standard terminology and uniform procedure of the minimum requirements for describing exploration and mining results to the relevant stakeholders; investors, community leaders, regulators, and international agencies (Rendu and Miskelly, 2001).

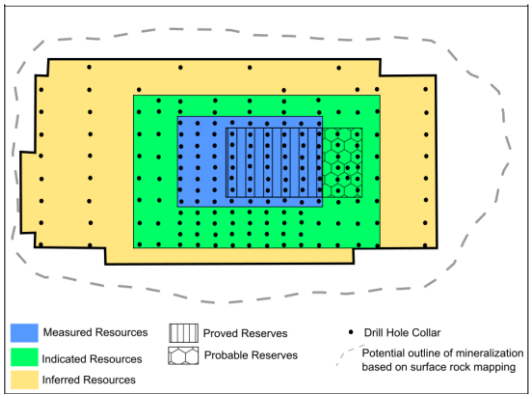


**Figure 2.5:** Mineral resources and reserves demarcation according to drill hole spacing confidence. Smaller drillhole spacing increases confidence from resources to reserves modified form Jowitt and McNulty (2021).

Mineral resource estimation is carried out by implementing various interpolation methods to predict the distribution of metals and minerals underground. The evolution of these mineral resource estimation methods is shown in (Fig 2.6) from (Dohm, 2021). These methods use some underlying mathematical and statistical assumptions to arrive at interpolating values of a variable at unsampled locations using values at sampled locations. Various mineral resource estimation methods are in practice today and are dependent on the complexity of the problem and the availability of data. The simplest mineral resource estimation methods involve the use of geometric functions (Journel & Huijbregts, 1978), such as:

- Polygonal;
- Triangulation; and
- Cross-sectional methods.

There are also geostatistical influenced methods (Rossi and Deutsch, 2013), such as:

- Inverse distance weighted;
- Kriging; and

- Conditional simulation techniques.

Recently, ML algorithms are used for mineral resource estimation (Mahboob *et al.,* 2023), these include the:

- Support vector machine;
- Radial basis function;
- Artificial neural networks
- k-nearest neighbours; and
- Decision trees.

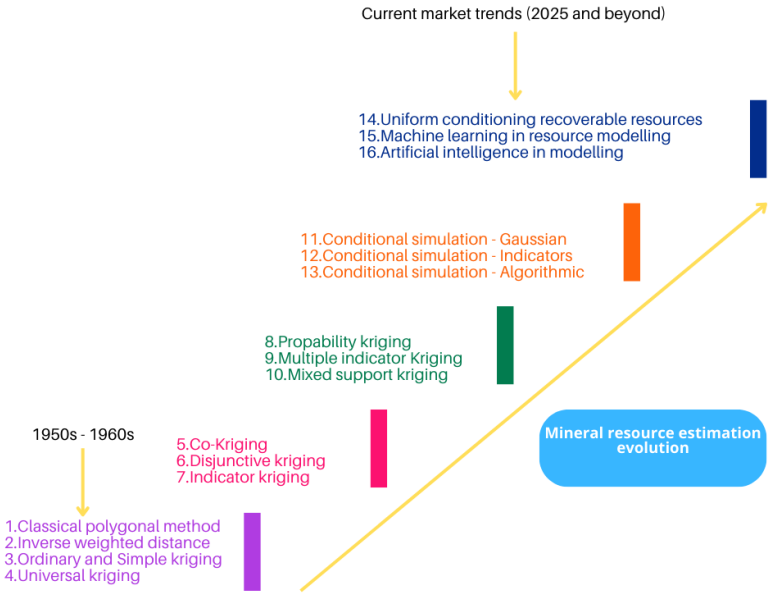


**Figure 2.6:** Evolution of mineral resource estimation methods/tools over time showing the older and simper methods on the (lower left) and the recent and more sophisticated methods in the (top right) adapted from Dohm (2021).

### 2.2.1 Traditional mineral resource estimation workflow

There are various modelling software's in the market that can be used to perform mineral resource estimation and have different components of the methods shown above (Fig 2.6). These software's can either be free as open-source platforms (Santangelo *et al.,* 2016) or commercial software's (Cowan *et al.,* 2002) which are commonly used in exploration and mining projects. Some of the commonly used commercial modelling software's are Leapfrog, Datamine, Micromine and Surpac (Birch, 2014). The modelling software's can combine field data, data interpretation, and user editing functionalities (Alcaraz *et al.*, 2011) that make the visualisation of the resource easier to use for mining purposes. The mineral resource

estimation process involves five vital steps as shown in (Fig 2.7). This workflow depicts a common step-by-step process adopted by most commercial modelling software packages in the mining industry.




**Figure 2.7:** Integrated mineral resource estimation workflow which incorporates the 1) data importation of spatial and geological data 2) Data validation where imported data is interrogated and cleaned 3) Automated domaining using spatial data to generate underlying domain points 4) Domain boundary generation using geological knowledge 5) Domain estimation using declustering and Kriging estimation technique modified from Mariz et al., (2016).

### 2.2.2 Geostatistics

Geostatistics is a branch of statistics that uses spatial and spatiotemporal relationships to predict values at unsampled areas, it also can describe spatial patterns that may be used in various fields such as agriculture, geology and environmental sciences (Matheron, 1963). It uses spatial tools such as semi variograms to determine the spatial pattern, and Kriging to determine the spatial predictions and standard error to quantify uncertainty of the predicted values at unsampled areas (Armstrong, 1998). In the exploration and mining environments geostatistics is used as a tool to predict metal or mineral grades and related volumes that are used to develop block models (Cowan *et al*., 2002). Block modelling involves the use of three dimensional (3D) to represent subsurface geological features. These grades are used for mine

planning purposes for scheduling extraction points at any given time of the life of mine (Sahoo & Pal, 2017).

## 2.3 Domaining

Geological processes are complex in nature and normally result in varying rock properties over a short distance and this may cause complications in mineral resource estimation. The estimation of grades and other rock properties assumes that the selected zone for estimation is continuous and homogenous (Fouedjio *et al.,* 2018). This being said, the estimation of grades should be localized to data points within that particular domain (Sterk *et al.,* 2019). Domaining is the process of dividing an orebody into smaller portions that exhibit consistent properties that aid in the interpolation process (Mohammadi *et al.,* 2022). Despite this understanding it has proven to be difficult for mining professionals to standardize the process of domaining, this presents challenges in mining operations which has negative implications. Poor domaining can result in underestimation or overestimation of grades, which can result in poorly informed mining decisions, thus it is important for mining professionals to take cognizance of this process (Sterk *et al.,* 2019). Good domaining practice relies on the underpinning ore forming processes that are represented by geoscientific information gathered from geological data collection processes (Sterk *et al.,* 2019).

### 2.3.1 Types of domain boundaries

Due to the inherent complexity of most ore deposits the boundaries resulting from these natural phenomena's have influence in the mineral resource estimation. Domain boundaries may be soft, hard or transitional depending on the formational process of the orebody which may be syn-genetic of post genetic. These boundaries influence the approach in mineral resource estimation and proper risk management is required to control their influence in the outcome.

a) Soft boundaries are poorly distinctive transitions from one domain to another as result of a poor confidence level in data interpretation of the geological properties and structures (Fig 2.8a). This is inferred to and can only be conclusive after field and sample verification (Elliot *et al.,* 2001).
b) Hard boundaries are abrupt transitions due to vivid differences in adjacent geological domains shown in (Fig 2.8b) below. These are normally marked by contacts of different geological fabrics caused by changes in mineralogical composition or structure. These can be verified though physical mapping and core logging to establish the location of these boundaries (Ortiz and Emery, 2006).

c) Transitional boundaries are the hardest to deal with as they represent an overlap of both soft and hard boundaries (Fig 2.8c) below. These may result from post-genetic processes such as structural deformation, contact metamorphism and alteration as an overprint on existing domains (Kasmaee et al., 2019).

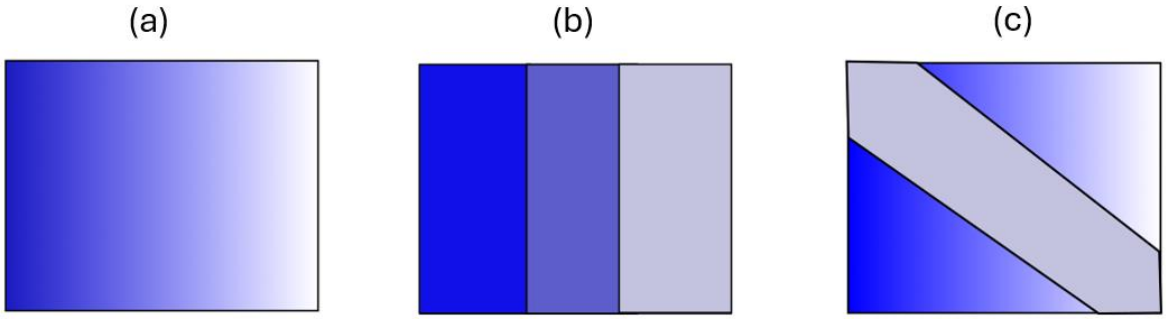


**Figure 2.8:** Common geological boundaries a) Soft boundary with no distinct break b)Hard boundaries showing hard breaks in between units c) A combination of both soft and hard boundaries which is more complicated to map modified from Kasmaee et al., (2019).

## 2.4 Clustering

The process of estimating grades is dependent on the spatial relationships. To establish domains, an unsupervised ML cluster analysis method can be used to define these spatial relationships. There are various clustering techniques which can be used to aggregate points into clusters, the selection of this methods is informed by the data set used, types of clustering techniques are explained in (Fig 2.9). Clustering algorithms populate input data into related groups called clusters, depending on the analysis method chosen (Xu & Wunsch, 2005). A cluster is defined as grouped data points with a relationship to their surrounding data points. The benefit of using clustering methods is that it can decipher hidden relationships amongst data points that may have been hard to recognize from raw data (Xu & Wunsch, 2005). The use of clustering methods in geosciences is common and has proven to be effective in domaining ore bodies by utilising the relationship between spatially related data points which is fundamental in mineral resource estimation. K-means is one of the widely used clustering techniques in data science and is described below (Fig 2.10).

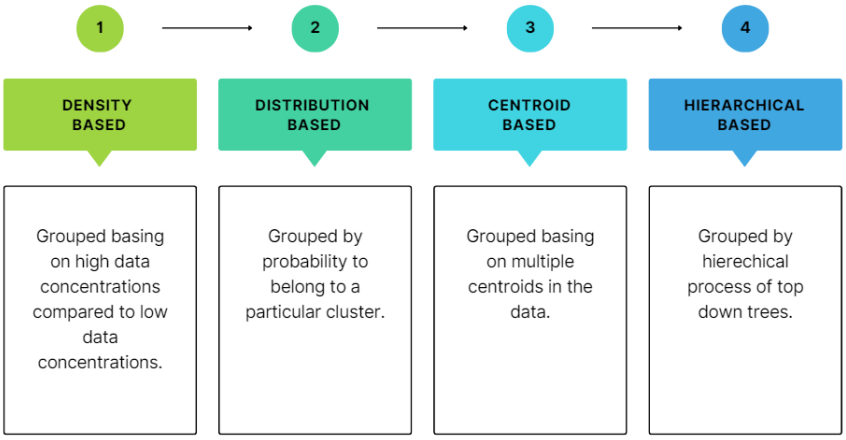


**Figure 2.9:** Types of clustering methods 1) Density method that develops clusters based on the concentrations of points 2) Distribution method that uses probability of belonging to determine clusters at used data concentration to determine clusters 3) Centroid methods that determines clusters by the a set number of centroids and closer points to centroids belong to a cluster 4) Hierarchical method where a top down decision tree is used to group data into clusters modified from Geeks for Geeks, (2024).

### 2.4.1 K-Means Clustering

K-Means clustering is an old simple unsupervised ML algorithm that divides a dataset into 'K' number of clusters (Wu & Wu, 2012). It is the most used clustering technique to generate geodomains in mineral resources estimation. This method is an iterative process that depends on features provided by the data set enabling it to assign every point to a distinct K cluster (Sinaga & Yang, 2020). The process summarised in the (Fig 2.10) below.

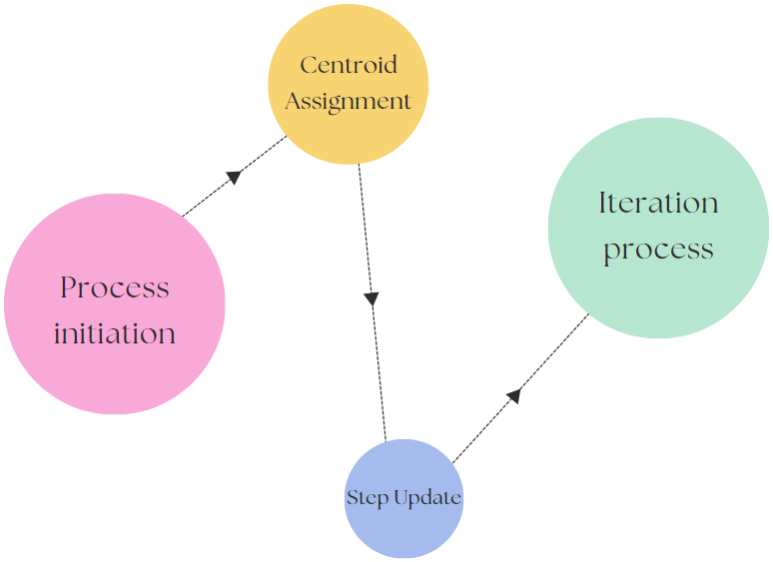


**Figure 2.10:** Steps involved in the K-means clustering process that includes 1) Process initiation 2) Centroid assignment in determining central points for cluster selection 3) Updating the previous step to select optimum K-clusters 4) Repeat the process for a number of iterations to generate optimal cluster numbers and sizes as modified from Sinaga & Yang, (2020).

K-means method optimises the cluster selection by minimising the variance within clusters by the sum of distances between each data points within a specific K cluster (Fränti and Sieranoja, 2019). In mining selection of clusters using traditional methods has proven to be difficult over the years due to their complexities and has implications in many mining tasks. K-means clustering provides an easier and flexible method to handle big data used in mining operations. In mineral resource estimation, K-means clustering method is used as the initial point domaining step to configure the dataset into distinct clusters that are used to determine the geologically continuously homogenous domains (Hong and Oh, 2021). Fig 2.11 below shows a graphical illustration of the outcome of a K-means cluster, resulting in 3 distinctive clusters, with points that are positioned close to each cluster centroid.

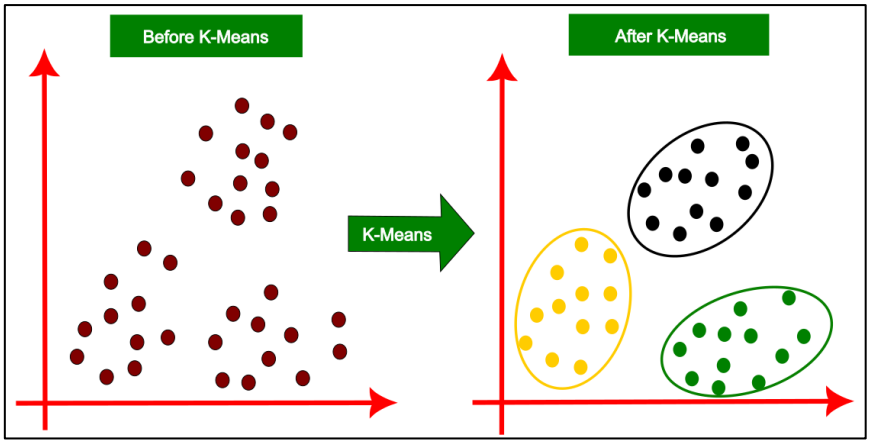


**Figure 2.11:** A simplified graphical illustration of cluster generation using k-means cluster with no centroid but proximity of points to one another is used to assign them to a cluster modified from Sinaga and Yang (2020).

## 2.5 Kriging Estimation

Kriging is a widely used spatial interpolation method for estimating parameters in unsampled locations (Rossi and Deutsch, 2013). It was initially used to estimate highly variable gold grades in the Witwatersrand basin and is named after a world renowned South African mining engineer Danie G. Krige (Van Beers and Kleijnen, 2003). This method has since been applied to other fields beyond mining geology such as hydrogeology, environmental sciences, and remote sensing fields (Van Stein *et al*., 2020). Kriging differs from other interpolation methods such as Inverse weighted distance and Linear regression because it uses the spatial characteristics of points to interpolate unsampled regions (Vann and Guibal, 1998). The output of a Kriging estimate is influenced by:

1. Distance between sampled and unsampled points.
2. Data configuration (Spatial distribution of sampling points).
3. Spatial correlation of data.

Efficiency of Kriging is mainly dependent on the correlation of data within the estimation location. It is a valuable method because it provides a measure of uncertainty which is ideal in mineral resource estimation (Oliver & Webster, 2015). Kriging also considers the spatial anisotropy which captures the direction dependency of spatially correlated data points (Boisvert *et al*., 2009). Kriging weights are assigned based on the spatial correlation deduced from the spatial configuration of the data (Oliver & Webster, 2015). It is considered the **B**est **L**inear **U**nbiassed **E**stimator (BLUE) (Vann & Guibal, 1998).

- **Best:** Kriging aims to minimize the error variance associated with the estimation.

- **Linear:** Estimated values are calculated by a combination of weighted linear equations.
- **Unbiased:** The error mean is zero.

The weights given to points are proportional to the distance they are from the estimated value (Rossi and Deutsch, 2013) shown in eq.1 below. Data clustering is also considered, highly clustered data is given less weight whereas sparsely distributed data is given more weight, and this subsequently affects the influence of each data point in the interpolation process (Van Stein *et al*., 2020). Anisotropy is a critical property in the Kriging method as it is common in interpolation that autocorrelation of data points is stronger in a particular direction (Boisvert *et al*., 2009). Anisotropy is calculated using a spatial variogram model which captures spatial correlation of data points (Rossi and Deutsch, 2013).

Kriging estimates are hypothetically expected not to be lower or higher than the true value used. Equations showing the simplified Kriging calculations:

$$g^* = \sum_{i=1}^{i=n} w_i * g_i \quad (1)$$

$$\sum_{i=1}^{n} w_i = 1 \quad (2)$$

$$R = g^* - G \quad (3)$$

where;

- $w_i$ = weights
- $g^* = estimate$
- $R = estimate\ error$
- $G = true\ value$.

To ensure that optimal weights are achieved, Kriging uses the simultaneous linear equation for selected points on the kriged matrix to ensure all input data are proportionally weighted according to the set distance in the semi variogram (Rossi and Deutsch, 2013). These linear equations are expressed as a matrix notation shown in eq.4 below , where the redundancy is represented by the correlation matrix on the left, and is multiplied by the Kriging weights, themselves regulating the impact of clustering of data within the selected zone of influence, to achieve the minimum error estimator (Coombes, 2008).

$$\begin{bmatrix} C_o & y_{12} & y_{13} & . & y_{1n} & 1 \\ y_{12} & C_o & . & . & y_{2n} & 1 \\ y_{13} & . & C_o & . & y_{3n} & 1 \\ . & . & . & . & . & . \\ y_{1n} & y_{2n} & y_{3n} & . & C_o & 1 \\ 1 & 1 & 1 & . & 1 & 0 \end{bmatrix} \text{ x } \begin{bmatrix} weights_1 \\ weights_2 \\ weights_3 \\ . \\ weights_n \\ 0 \end{bmatrix} = \begin{bmatrix} \overline{y_1} \\ \overline{y_2} \\ \overline{y_3} \\ . \\ \overline{y_n} \\ 1 \end{bmatrix} \quad (4)$$

Where: y is the semivariogram value between sample points.
C is the covariance between sample locations.

### 2.5.1 Simple and Ordinary Kriging

Mineral resource estimation works under the assumption that the data being used are regionalised variables, as such they display some level of randomness (Oliver & Webster, 1990). Any random phenomena can be represented by equation shown in eq.5 below.

$$Z(s) = \mu + \varepsilon(s) \quad (5)$$

Where Z(s) random variable, the $\mu$ is the global mean and the $\varepsilon(s)$ is the residual resulting from the difference between the global mean and the random variable (Daya and Bejari, 2015). In simple Kriging the global mean is known, and it is factored into the estimation calculation by subtracting it from the data and added back into the equation after calculating the residuals. Simple Kriging is easy to compute and functions well with few data (Olea, 1999) shown in eq.6 below.

$$[Z^*_{SK}(u_0) - m] = \sum_{i=1}^{n} \lambda_i . [Z(u_i) - m] \quad (6)$$

Where:

- $Z^*_{SK}(u_0)$: The **estimated value** at the unknown location $u_0$ using **Simple Kriging**.
- $m$: The **known mean** of the random field (assumed constant and known in Simple Kriging).
- $Z(u_i)$: The **observed value** at known location $u_i$.
- $\lambda_i$: The **Kriging weight** assigned to the observation at $u_i$
- $n$: The number of known data points used in the estimation.

In ordinary Kriging the global mean is unknown and is not factored into the estimation calculation. It assumes spatial stationarity of both the mean and variance (Wackernagel, 2003). Ordinary Kriging is computationally intensive compared to simple Kriging and it provides more accuracy as it calculates the local means of the data (Cressie, 1988) shown in eq.7 below. Both simple Kriging and Ordinary Kriging rely on the variogram model to determine

the spatial autocorrelation of points (Daya and Bejari, 2015). Due to its high accuracy ordinary Kriging is mostly used in exploration and mining in case there is sufficient data to model from.

$$Y^*_{OK}(u) = \sum_{i=1}^{n} \lambda_i . Y(u_i) \quad (7)$$

Where:

- $Y^*_{OK}(u)$: The **estimated value** at the unknown location $u$ using **Ordinary Kriging**.
- $Y(u_i)$: The **observed value** at known location $u_i$.
- $\lambda_i$: The **Kriging weight** for the observation at $u_i$, determined based on spatial correlation.
- $n$: The number of known data points used in the estimation.

### 2.5.2 Variogram analysis

A variogram is a geostatistical tool used to capture the spatial variability and spatial dependency within a dataset (Bohling, 2005). It models the spatial structure by using the variance between pairs of sample points as a function of their lag distance (Oliver & Webster, 2015; Bohling, 2005). It is calculated by taking the square difference between pairs of values at an increasing distance and averaging them.

$$\gamma(h) = \frac{1}{2N(h)} \sum_{a=1}^{N(h)} [z(u_a + h) - z(u_a)]^2 \quad (8)$$

Where:

$\gamma(h) = Semivariance\ Function$

$u_a = Vector\ of\ spatial\ corrdinates\ (x, y, z)$

$h = Lag\ distance$

$z = Variable\ of\ Interest$

$z(u_a) = Variable\ of\ interest\ as\ a\ function\ of\ spatial\ location$

$z(u_a + h) = Variable\ of\ interest\ afunction\ of\ spatial\ location\ at\ a\ lag\ distance\ h$

Variogram establishes a locative continuity in which the samples close to each other display greater similarities compared to samples further apart from each other (Oliver & Webster, 1990). When using variograms to analyse continuity of the mineralisation, it is reflected by the growth rate of the variogram over a short distance. The following are variogram components;

1. Sill: This is the value of the semi variance where the variogram graph start to flatten it represents the end of correlation shown in (Fig 2.14).
2. Range: This is the lag distance of the semi variogram where the semi variogram components reaches the sill value (Fig 2.14).
3. Nugget: Represents the short scale noise or randomness related to regionalized variables and is inherent to geological properties. It is illustrated graphically on a variogram plot as a discontinuity from the zero mark of the function (Fig 2.14).

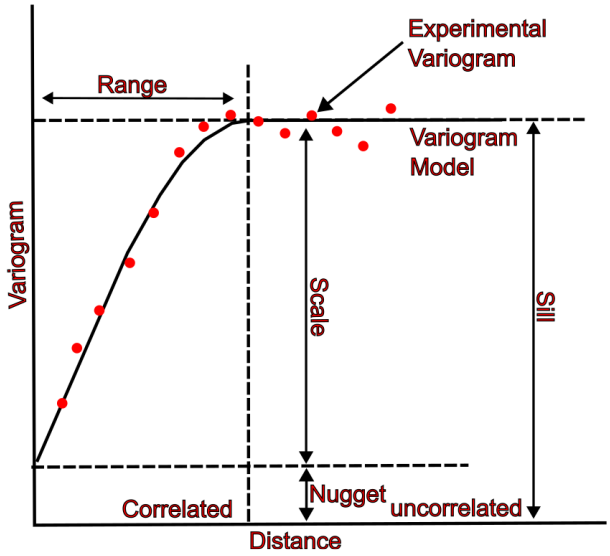


**Figure 2.12:** Semi variogram structure showing the sill, range and the nugget. It also shows the area of correlation and uncorrelated area that are regions beyond continuity adopted from Bohling, (2005).

### 2.5.3 Quantitative Kriging neighborhood analysis

To perform Kriging estimation, firstly deal with the conditional bias or the smoothing effect that is inherent to the Kriging estimation process (Vann *et al*., 2003). This bias may lead to unreliable grade values which impacts mine production. This may be due to overly restricting the search neighbourhood (Vann *et al*., 2003). To address this bias, quantitative Kriging neighbourhood analysis (QKNA) is carried out to determine the optimal Kriging parameters (Talesh Hosseini & Asghari, 2020; Vann *et al*., 2003). The QKNA helps to create the best search volume by taking into cognizance the configurations of either correlated or non-correlated points within a selected block for resource estimation. The following estimation parameters are considered important to run a Kriging estimation and are determined in the QKNA process:

1. Estimation block sizes;
2. Minimum and Maximum number of samples to be considered;
3. Search range and configuration; and
4. Number of discretization points.

Kriging efficiency and Slope of regression are used to determine the best search parameters. These metrics are also used as post Kriging estimation evaluation tools, they are discussed in detail in the methodology section.

### 2.5.4 Kriging Performance Analysis

Model performance analysis is a key component of the Kriging estimation method, it provides insight into the accuracy and the precision of the Kriging estimate. It is important to have a measure of the performance of estimation methods because of their inherent uncertainties (Loquin & Dubois, 2010). This is critical, especially in the mining industry where data is vital for mining decisions. Failure to assess the efficiency of an estimation method can affect the overall performance of the mining operation due to distorted grade-tonnage estimations (Vann and Guibal, 1998). Statistical and mathematical methods are commonly used to qualitatively and quantitatively analyse the performance of various estimation methods and are normally incorporated into commercial modelling software's that is used in mining. The following metrics are commonly used in assessing the performance of Kriging estimations.

Kriging Efficiency

KEFF is a statistical analysis tool that was first introduced by Daniel Krige (Deutsch *et al*., 2014). It measures block estimations and is expressed as Kriging variance normalised by the block variance expressed as a percentage. This method is discussed in detail in the methodology section.

Slope of Regression

SLOR is a measure of conditional bias of the estimation (Deutsch *et al*., 2014). The SLOR is a comparative method where the linear regression of the true data is plotted on the estimated data. This method is discussed in detail in the methodology section.

## 2.6 Machine Learning Applications

Zhou, (2021) defines ML the use and development of computer systems that may learn without following explicit instructions. This is done via the use of algorithms and statistical methods which may draw patterns and inferences from a given dataset. ML have a peculiar learning

process that distinguishes it to traditional from conventional explicit programming method (Jooshaki *et al*., 2021). It provides the ability to learn from data and improve previously studied trends within the data. ML has gained popularity over the past years within the exploration and mining industry (Dumakor-Dupey and Arya, 2021). These tools have the capability of generating predictive models that can be used to generate exploration targets (Sun *et al*., 2020) and assessing geological complexity of deposits (Sun *et al*., 2022). Additionally, in mining operations, ML has been used in the extraction optimisation of mineral resources by studying mineralogical and grade distribution within complex ore deposits (Mahboob *et al*., 2022). In all these applications, ML has shown positive results by revealing insights in datasets that had proven difficult or did not generate value using traditional analytical methods (Karpatne *et al*., 2018; Maniar et al., 2018; Jooshaki et al., 2021). This is mainly due to their computational capabilities as they can handle multitudes of data and can generate output models faster than traditional modelling methods. ML also easily utilises cheap data and most of the programming platforms such as Jupyter and PyCharm are free open access systems. Despite the growing application of ML in geosciences there are still challenges such as standardisation and integration into old systems more especially in mining operations due to flexibility issues. ML is divided into three broad areas (Fig 2.13):

- Supervised learning: Uses labels to assign datapoints and train algorithm to generate predefined outcomes.
- Unsupervised learning: Learns from data without human influence or preset outcomes.
- Reinforcement learning: Environmental based learning in which an agent learns from reward feedback.

ML is a subset of AI which has the capability of empowering smart applications by continuous learning of the set theory as dictated by the data (Anderson, 1983).

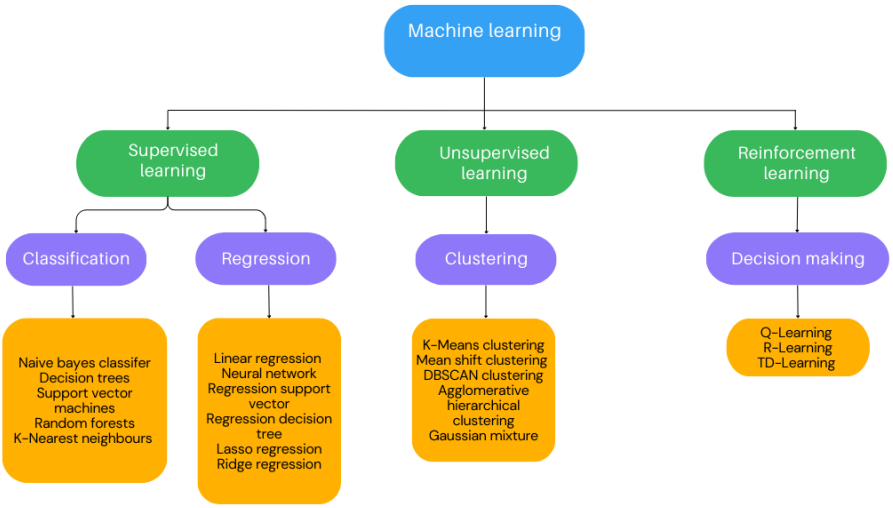


**Figure 2.13:** ML categories and their respective algorithms (orange boxes) 1) Supervised learning methods that used labelled data for learning 2) Unsupervised learning methods use unlabelled data and no human interaction (independent methods) 3) Reinforcement learning methods rely on learning based on cumulative rewards adopted from Ayodele, (2010).

### 2.6.1 Traditional machine learning workflow.

In the science and engineering fields, it is common practice that a problem must be identified to start any project which will form the initial step of the whole project. Similarly, ML methods require a problem to be identified beforehand which will aid in the selection and design of the methodology of the process because ML algorithms are not a one size fits all (Janiesch *et al.*, 2021). A desired outcome may be determined prior to data collection stages so that proper data collection can be done with the outcome in mind. This is at the discretion of the ML user and in some cases may lead to bias in the modelling process which can be detrimental to the quality of the outcome (Mehrabi *et al.*, 2021). Data is then collected basing on the defined problem and desired outcome, this may be lithological, structural, mineralogy, whole rock geochemistry and grades. Before running a ML model features are to be determined, Features are the key ML models and are defined as any data (numeric or non-numeric), which are used as inputs in the model (Duboue, 2020). They present the information from raw data that is required to predict and decrypt data relationships. Next a model training is done by deploying various ML algorithms to train the majority of the dataset (60-80%) to build the initial response of the prediction model (Nasteski, 2017). Model evaluation follows by testing the model with the remaining data (20-40%). At this stage the success and performance of the ML algorithms are assessed by performance metrics. The commonly used performance matrices are precision, recall and F1 test for classification algorithms (Juba and Le, 2019). For regression algorithms, root mean square error, mean absolute error and root mean square logarithmic

error are used (Botchkarev, 2018). In most cases the best performing algorithm is then selected to form part of the main model that can be used in the future, the ML modelling workflow is shown below (Fig 2.14). Further up the workflow is the model execution, monitoring and maintenance process in which there is continuous improvement and quality checks for any updates on the model modules and dependencies that may affect the use and performance of the model.

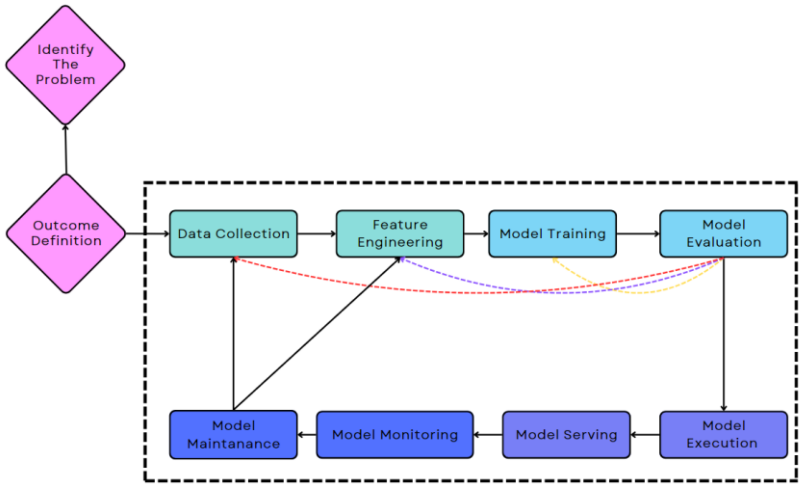


**Figure 2.14:** A simple ML modelling workflow that shows a summary of the process described above, the identification of problem and model outcome requirements are key in the use of ML models adopted from Andriy, (2019).

### 2.6.2 Machine learning model selection

The selection of the right modelling algorithm is not an intuitive process, but a data driven process (Raschka, 2018). Initial understanding of the data is important in the selection of the algorithms, it informs the process. When using classification and regression algorithms the user must have a good understanding of the mathematical and statistical operations that underpins the process of these algorithms to make an informed decision with a desired outcome. The selection is determined by careful consideration of the data size, data quality, data types, feature dimensionality, interpretability and intended results from the modelling (Lee and Shin, 2020). Fig 2.15 demonstrates a step-by-step selection process of a regression algorithm.

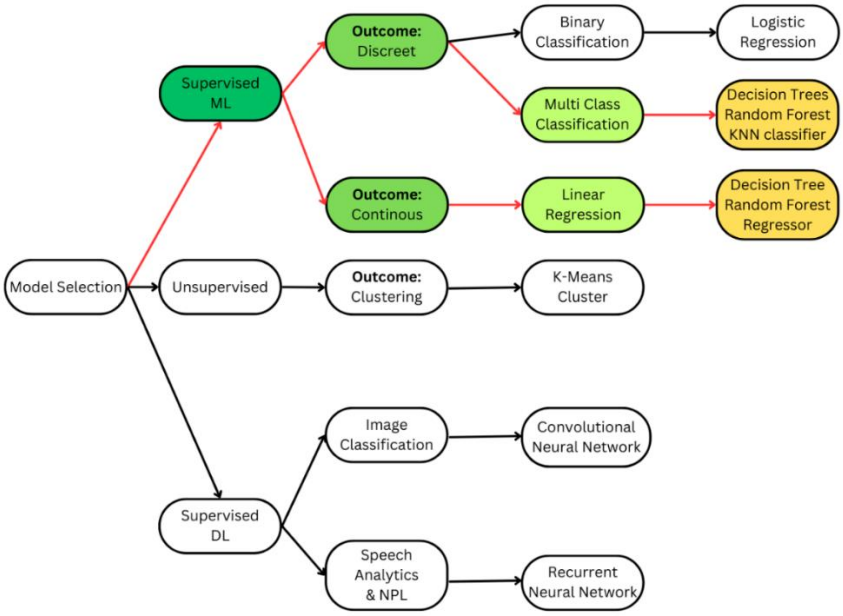


**Figure 2.15:** ML algorithm selection flow showing the step-by-step process of selecting a model fit for purpose.

### 2.6.3 Machine learning model performance

ML models have shown to have potential in optimising mining operations as discussed in the introduction section above. These models require a measure of their performance to ensure that the level of confidence is within desirable limits (Botchkarev, 2018). These methods include the use of validation techniques such as K-Fold validation, accuracy, precision, recall and error value magnitude analysis such as confusion matrix (Santra and Christy, 2012). The evaluation methods used in this research are discussed in depth in the methodology section of this report.

### 2.6.4 Hybrid machine learning algorithms

A combination of ML algorithms and traditional mineral resource estimation methods has recently gained immense interest from industry experts. Fig 2.16 presents a workflow demonstrating the integration of ML in a spatial estimation process (Erdogan Erten *et al.,* 2022). This combination provides robust modelling as it leverages the computational power of ML and the spatial relationship of data points in understanding mineral resources. The use of ML algorithms in mineral resource estimation and evaluation has been reviewed by (Mahboob *et al.* 2023), who have highlighted that even though some of the studies indicate that ML methods outperformed non-ML based algorithms. In some cases, a collaborative process between the two can result in better results.

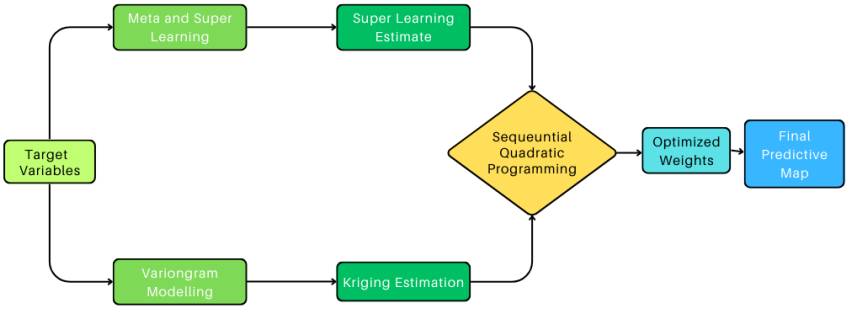


**Figure 2.16:** Simplified workflow of the combination of super learning and Kriging models for spatial estimation strategy. Illustration, reconstructed from (Erdogan Erten et al., 2022).

# 3 METHODOLOGY

This chapter presents the step-by-step description of the methodologies used for this research work. The workflow adopted in this project is illustrated in (Fig 3.1). The methodology is divided into two working environments: (a) Python Jupyter Notebook and (b) Datamine modelling software.

## 3.1 Datasets

The datasets represent channel sampling data from Driefontein gold mining complex in the Carletonville Goldfield of the Witwatersrand Basin, South Africa. The data is tabular and was stored in a Microsoft excel sheet. This data contains spatial coordinates, lithofacies, gold grade and sedimentological parameters.

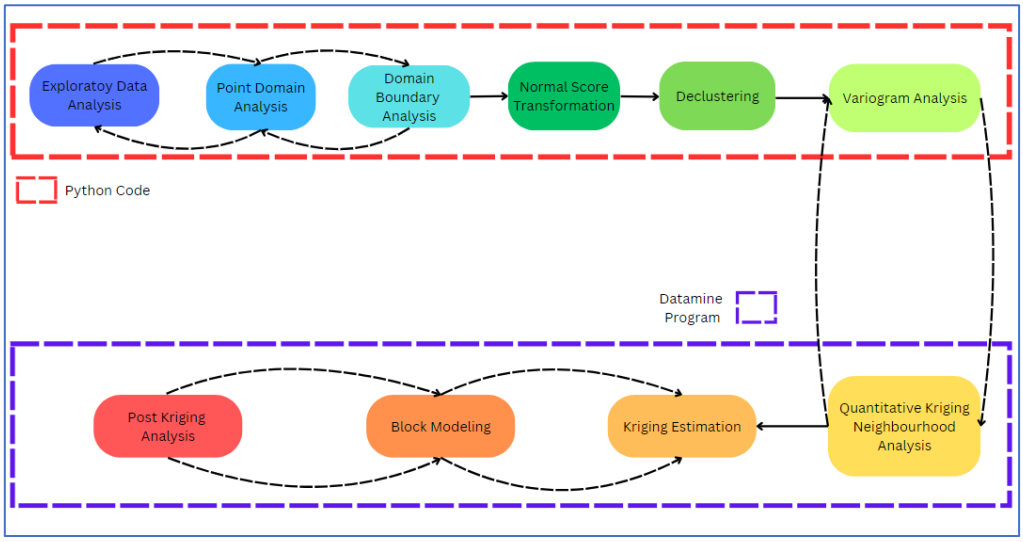


**Figure 3.1:** Simple graphical illustration of methodology steps and modelling environment used at each step.

## 3.2 Exploratory Data Analysis

The data used for this thesis was raw data that has not been pre-processed or altered for any purpose as such the first stage of the method section was to build the working environment and perform exploratory data analysis.

## 3.3 Working Environment Setup

Setting up a working environment on Jupyter Notebook, which will act as the processing platform for this research project. This was done via downloading and installation of the Anaconda software which is the open-source distribution platform for both Python and R

programming languages. The working directory was then setup to enable the initiation of the project. This is then followed by importation of libraries, which can either be inbuilt in python or as an external library defined by a user.

Importing and installing all required libraries and supporting modules such as:

- Import Os – Operating System: Allows to interact with the computers operating system (Windows, MacOs, Linux). This module provides ability to manipulate file directories, extraction of operating system and shell commands.
- Import NumPy as np: Stands for Numerical Python, forms the foundation of computing in python as it enables the use of arrays in different sizes, matrices and execution of mathematical operations.
- Import Pandas as pd: It is fundamentally a Numpy extension which allows the handling and analysis of structured data.
- Import Matplotlib.pyplot as plt: A visualization tool in python that enables the representation of data in static, interactive and animated graphs and plots.
- Import seaborn as sns: This is a statistical visualization tool that is built ontop of Matplolib which is used to present high level statistical graphics.

The primary data (excel) was converted to csv file and imported into the Jupyter platform for initiating the preprocessing of the data. Primary data contains grade values measured in centimetre-grams per ton (cm.g/t), it was then converted to grams per ton (g/t) by channel width (cm) and an extra column was added to the data. This conversion is represented by the equation below:

$$Au\ (gt) = \frac{Au\ (cm.g/t)}{CW} \tag{9}$$

The data summary statistics are then reviewed (using the df.describe function). This is done to check the following statistical properties:

- Count: Total number of entries per variable.
- Mean: Arithmetic mean of the data per variable.
- Std: Standard Deviation of the data per variable.
- Min: Minimum value of the data per variable.
- Max: Maximum value of the data per variable.
- 25 %: 25 percentile of the data per variable.
- 50 %: 50 percentile of the data per variable.

- 75 %: 75 percentile of the data per variable.

The df.info() function is used to check characterization of the data, and this includes the data types (float, object, Boolean, integer or string). This function also checks if there are any missing values in the data.

A Scikit-learn (external library) preprocessing function, OneHotEncoder was imported and ran to convert categorial data into a numerical value. This is done by transforming categorial features into a “dummy” or “one-hot” binary vector with zeros or ones where category position is concerned. This was done on the non-numerical variables (packing, roundness, sorting, mineralisation, carbon, pebble assemblage, colour and basal contacts. The spatial data area was constrained to an area with sufficient data by reconfiguring the coordinates. This resulted in the repositioning of the outline/boundary of the study area, it was done to narrow down the research to data centric areas which may reduce the number of computational calculations in the estimation process. The data features were then displayed on a scatter plot using the matplotlib function (plt.scatter) to display the spatial distribution of the different variables and their measurements.

The data features were displayed on a histogram using the matplotlib function (plt.hist) to display the distribution of the data by a series of bins representing the range and frequency of data falling within that range. The data features were then displayed on a correlation matrix both using the pandas (df.corr) function which is driven by the Pearson correlation coefficient that measures the linear correlation between pairs of variables. The heatmap from the Seaborn function (sns.heatmap) to represent the colour display of the correlation values. The correlation was done in three blocks, one on all the variables, secondly on sedimentary properties only and then gold grades against sedimentary properties. These were done in both correlation integer numbers and then done on percentage correlation using the correlation matrix percentage conversion.

### 3.4 Sample Point Domain Analysis

To build the point domains the Standard K-means clustering technique was used and a new approach “Signed distance -based clustering and combinatorics” were used and multi dimensionality scaling was incorporated to reduce the dimensionality of the data. The Standard K-means clustering is explained in Chapter 2 and the Signed distance-based clustering technique is explained below. The required libraries were imported to facilitate the execution of the K-means clustering and in addition the combination of Multidimensional scaling with K-means clustering method. The data was imported using pandas function df =

pd.read_csv (file name) and ensured that the data format is correct and can be accessed using the pandas functions. Data columns to be used were selected on the feature importance and this data was numeric. The columns included (a) spatial location data (X, Y, X), (b) grade data (Au g/t), (c) sedimentological data such as conglomerate % and pebble size (mm). The data was then scaled using the StandardScaler function of the Scikitlearn.preprocessing which is used to standardize data with different units or scales. This is done by removing the mean and then scaling to a unit variance. The data is then transformed by fitting the scaler to the data using the scaler.fit_transform(data) function. Then selection of the optimal number of clusters is determined using the silhouette score and Davies-Bouldin index.

### 3.5 Signed Distance-based Clustering

This clustering technique was explored in geodomaining by Nwaila *& Frimmel* (2023). This method was applied to Platinum Group Elements deposits of the Bushveld Complex and the gold deposits of the Witwatersrand Basin. It has shown advantages over the commonly used clustering techniques. It is data-driven and eliminates the manual selection and pre-determination of clusters. It solely depends on the underlying structure of the dataset in capturing the spatial components and variation of geological data while preserving the underlying geological systems (Nwaila & Frimmel 2023). This is important in building reliable mineral resources models as it increases the accuracy and precision of the models.

This method considers the distance and directionality of sample points. The signed distance clusters technique has been used in mineral resource evaluation. This new approach provides a combination of the traditional K-means clustering and signed distance functions (Nwaila & Frimmel 2023). When carrying out this method, firstly the features of interest are defined and in this study in this study Au (g/t), CW (cm), and Au (cm.g/t) were selected together with their spatial locations in (X, Y, Z) coordinates. The signed-based distance calculations are then done for each sample point. The method calculates the closeness of these points to the nearest feature boundary. These calculations were done using the fast-matching algorithm where each point is given a signed value to define if it is inside or outside the features (Nwaila & Frimmel 2023). The magnitude is then determined which shows the distance to which the point is nearest to the feature boundary.

After assigning signed distances, the weights were assigned to determine the relative importance of the features. The weighted signed distance values were then subjected to a multi-dimensionality reduction but preserving distances between samples and enhancing the visualisation of relationships and similarities between data points in a low dimensionality space. The Sklearn manifold MDS module was used to achieve this reduction. Finally, the K-

means clustering technique was applied to the dimensionality reduced weighted signed distances values. The final K-means clusters were selected basing on the spatial proximity and similarities of data points in the dataset.

### 3.6 Post Clustering Metrics

Upon completion of the clustering process, a post clustering analysis was carried out to determine the goodness of the clusters generated. The silhouette score and the Davies-Bouldin index are used to analyse these clusters.

#### 3.6.1 Silhouette scores

The silhouette score is a mathematical matrix used to assess the quality of clusters generated by unsupervised clustering algorithms (Rokach & Maimon, 2005). It measures two properties, cohesion and separation shown in Fig 3.2 below, it provides insights into the efficiency and reliability of clustering algorithms (Shutaywi and Kachouie, 2021). Cohesion is the similarity of variables within a singular cluster and the separation is the dissimilarity measure between neighbourhood clusters (Saputra *et al*., 2020). Saputra *et al*., (2020) highlight that the sillhoute scores are a matrix is driven by a distance calculation and in this research a euclidean distance function is used. Fig 3.2 below is an illustration of a clustering process, and the silhouette calculation follows basing on the diagram.

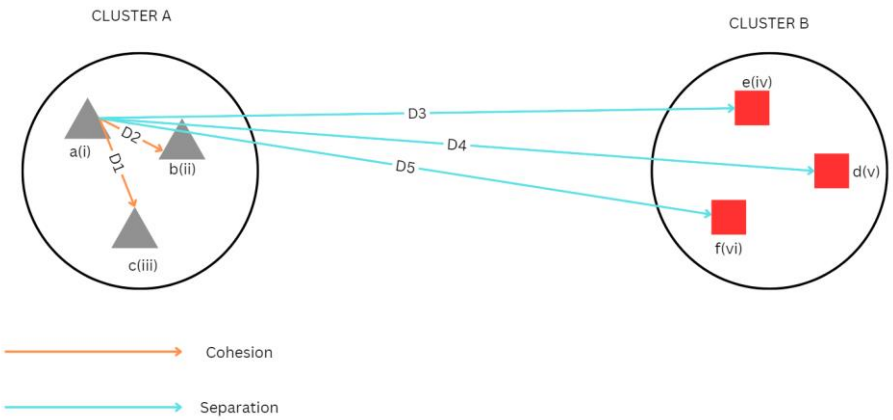


**Figure 3.2:** Simple graphical illustration of Silhouette calculations and the cluster selection criteria (1) Cohesion demonstrated by D1 and D2 (2) Separation demonstrated by D3, D4 and D5 adopted from Saputra et al., (2020).

To calculate the cohesion, separation and the silhouette score the distances of the data points is calculated and the Euclidean distance is used.

$$Euclidean\ distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (10)$$

$$Cohesion = \frac{D1+D2}{2} \quad (11)$$

$$Separation = \frac{D3+D4+D5}{3} \quad (12)$$

$$Silhouette\ coefficient = \frac{Separation - Cohesion}{\max(Cohesion,\ Separation)} \quad (13)$$

### 3.6.2 Davies-Bouldin index

The Davies-Bouldin index is another metric for evaluating clustering models. It measures the similarity of each cluster to its most similar cluster, where similarity is expressed as the ratio of intra-cluster to inter-cluster distances (Petrovic, 2006). In this research, these distances were computed using the Euclidean formula. The standard formulation of DBI is.

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{s_i - s_j}{d_{ij}}\right) \quad (15)$$

Where:

- K is the number of clusters;
- $s_i$ is the average within-cluster distance of cluster i;
- $s_j$ is the average within-cluster distance of cluster j; and
- $d_{ij}$ is the between-cluster distance of cluster i and j.

## 3.7 Domain Boundary Analysis

The generation of domain boundaries was done using the scikit-learn library package, which is a host of many classification and regression ML algorithms. It was used for training, testing and evaluating the model. The dataset was divided into the (a) training (b) testing and (c) cross-validation set. The features (independent variables) and the output labels (dependent variables) are initially defined for all the models. The cross-validation data set makes 20% of the total data and is used in the final model evaluation stage. For this project a single run of multiple classification ML algorithms was used, these algorithms are (kNN, SVM, Decision-trees, RF, AB, Naïve Bayes, Logistic regression, Neural Networks and the Gaussian process). The next step involved defining the dictionary of each classifier and their respective hyperparameters which are model controlling parameters. Upon completion of the training and testing of the data a validation process of leave one out (LOO) is applied which is a 5-fold stratified cross validation method. The predictions were then done by calculating the results

from the cross validation set results, the accuracy and F1 score results from each of the 10 iterations were calculated and the average performance was calculated per model. The best forming models are then saved for future use.

### 3.8 Machine Learning Evaluation Metrics

ML performance is measured using a variety of evaluation matrices which help in the establishment of constructive feedback for the models (Novaković *et al.,* 2017). This is done by the continuous feedback generated from these matrices which helps to improve and fix the model until a desired outcome is achieved (Japkowicz, 2006; Japkowicz and Shah, 2015). These matrices also provide quantitative measures that are used to compare the performance of various models which helps in the selection criteria of the best model to use for future classification (Novaković *et al.,* 2017).

#### 3.8.1 K-Fold cross-validation

The k-Fold method is commonly used in ML to evaluate model performance. This method was used in this project to resample the data and validate the model performance and test the reliability of the model (Rodriguez *et al.*, 2010). Upon completion of the ten iterations, cross-validation was carried out to analyse the performance of the models. The K-Fold validation is done by determining the number of a group of data points as subsets of the train-testing data. These subsets are known as “Folds” and in this project, a total of five folds were chosen. A singular fold is treated as a training set K – 1 times and the testing set 1 time. The average of the cross-validations is used as the model performances.

#### 3.8.2 Confusion matrix

Confusion matrix is used to measure the performance of binary classification models (Japkowicz and Shah, 2015). The performance is visualised in a table to display the comparison of predicted outcomes and the actual data labels (Santra & Christy, 2012). The following class of resultant responses from the model are used to measure the performance of the model:

- True Positives (TP) – Correctly classified/predicted positive labels;
- False Positives (FP) – Incorrectly classified/predicted positive labels;
- True Negatives (TN) – Correctly classified/predicted negative labels; and
- False Negatives (FN) - Incorrectly classified/predicted negative labels.

Accuracy

This captures the proportion of correctly classified/predicted instances out of the total instances shown in eq.17 below (Novaković *et al.,* 2017).

$$Accuracy = \frac{Number\ of\ correctly\ classified\ outcomes}{Total\ number\ of\ occurances} \quad (17)$$

Precision

The ratio of the true positive prediction made by the model to the total number of positive predictions shown in eq.18 (Novaković *et al.,* 2017).

$$Precision = \frac{True\ Positives}{True\ Positives+False\ Positives} \quad (18)$$

Recall

The ratio of correctly predicted positive observations to all the actual positives shown in eq.19 (Novaković *et al.,* 2017).

$$Recall = \frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (19)$$

F1 Score

This is the harmonic mean of the precision and recall. It is calculated to provide a balance between Precision & Recall shown in eq.20 (Novaković *et al.,* 2017).

$$F_1 = 2\ x\ \frac{Precision\ x\ Recall}{Precision\ x\ Recall} \quad (20)$$

### 3.9 Pre Kriging Process

Kriging is a linear-spatial analysis technique that is used in studying the spatial distribution of properties (Rossi and Deutsch, 2013). Before performing Kriging, there are spatial attributes that are determined to achieve an optimal estimation. These include variogram analysis and data transformation to aid in proper estimation (Vann *et al.,* 2003). A quantitative Kriging neighbourhood analysis is also done to determine the block sizes, number of samples to work with, search range, and the total number of discretization points required (Vann *et al.,* 2003).

#### 3.9.1 Z-Score transformation

Mineral resource data normally contains highly variable data, more especially measured grade values and quantitative geochemistry data. On many occasions, the data has outliers amongst them which are defined as measurements of a variable of interest that are above or below

nominal values. This data must be treated in a way that limits or minimises outlier influence in the overall distribution and statistics of the data to be used in estimations. The impacts of including outliers in data analysis have been discussed in many other research fields (Kwak & Kim, 2017; Barnett, 1994) and the methods of identifying and treating outliers have also been extensively studied.

$$Z = \frac{X - \mu}{\sigma} \tag{21}$$

Where:

- Z = Z–Score;
- X = Individual data point;
- μ = Population mean; and
- σ = Standard Deviation.

Treating outliers is a sensitive task because it is data dependent and cutting out outliers or capping data may distort the data and result in misrepresentation of the data. So as much as we want to deal with the impacts of outliers on the overall data, we need to be careful not compromise the quality of the data. To curb this data density analysis is critical and is done on a case-by-case merit. In this project confidence limits are used to determine just how many outliers to deal with. A 98% confidence interval which is equivalent to a 2.33 Z–score value is preferred. This means that all data that plots beyond the 98% interval may be treated as outliers, this helps maintain the integrity of the data. Domain 0 initially had 1477 samples and were reduced to 1386 after transformation so 91 samples were cut out. Domain 1 had 540 samples and reduced to 495, so 45 samples were cut from the data after transformation. The effect of dropped samples is shown in distribution diagrams in the results section.

### 3.9.2 Normal score transformation

Gold grade data is inherently positively skewed and a transformation of the data to a normal distribution is done before the estimation (Nwaila *et al.*, 2020). Normal score transformation was used in this project data. This is done by sorting the data through ranking data from lowest to highest and are assigned an equivalent rank from a standard normal distribution (Yunsel, 2012). These rank values from the normal distribution are then considered as the transformed data.

Kriging is a weight-dependent interpolation method and the distribution of weights is determined by sample locations and distance from the estimation point (Shahbeik *et al.*, 2014). The collection of mining data is informed by the geological significance of the area being

sampled, for example in a mining operation more data is collected within mineralised sections of the mine which introduces sampling bias to the data (Isaaks & Srivastava, 1989).

### 3.9.3 Cell declustering

Weights are distributed based on proximity to the surrounding data within a stationary domain. Generally, data points that are close to each other "clustered" are assigned reduced weights, and those that are far apart get more weights (Goovaerts, 1997). Cell-based declustering minimizes the declustered mean per cell, so a plot showing the various cell sizes and their respective declustered means is generated to choose the optimal cell size with the lowest declustered mean (Isaaks & Srivastava, 1989; Rossi & Deutsch, 2013).

(a) Cell based declustering process.

An original cell network is determined. A grid of equal cell sizes is placed over the domain, the cell size approximately represents sample spacing in a sparsely sampled area of the domain. The total number of cells occupied ($C_{total}$) is calculated and the resultant cells are assigned the same weight. All the occupied cells are given equal weights. In cases where there is only one data point in a cell, it is given as:

$$\frac{1}{C_{total}} \tag{21}$$

In cases where there are multiple data points within a cell then they share the weight assigned to the cell they occupy. Sample weights are calculated as follows:

$$w_i = \frac{\frac{1}{n_x}}{Number\ of\ cells\ within\ data} * n \tag{22}$$

Where:

n=Number of occupied cells;

$w_i$= weight of the individual sample; and

$n_x$= Total samples within the cell which sample *i occupies.*

### 3.9.4 Variography

Once the normality of the data is tested, variography analysis was carried out to understand the spatial properties of the data and generate output data such as nugget, sill, and range. These parameters are required to compute data continuity and correlation of data over a certain distance (Bohling, 2005). The variogram has been the most important tool used in the

Kriging process as it provides insights into the spatial component of exploration and mining data (Oliver and Webster, 2015).

Covariance is given by eq.23 below where the first point is the head and the next point used is the tail Z being the variable of interest at different points (Bohling, 2005).

$$C(h) = \frac{1}{N(h)} \sum_{a=1}^{N(h)} z(u_a) * z(u_a + h) - m_0 * m_{+h} \quad (23)$$

Where:

- $C(h)$ = Covariance function;
- $z(u_a)$ = Head;
- $z(u_a + h)$ = Tail point with lag vector *h;*
- $m_0$ = Mean at head values; and
- $m_{+h}$ = Mean at tail values.

Correlation is given by eq.24 below, where a ratio of the overall covariance and the square of the sum of standard deviations for both tail and head values (Bohling, 2005). $\rho(h) = \frac{C(h)}{\sqrt{\sigma_0 + \sigma_{+h}}}$ (24)

Where:

- $\rho(h)$ = Correlation function;
- $\sigma_0$ = Standard deviation at the head point; and
- $\sigma_{+h}$ = Standard deviation at the tail point with lag vector *h.*

### 3.9.5 Quantitative Kriging neighbourhood analysis

This step involves the calculation of the optimal search volume "neighbourhood". The parameters being considered for this process include the search range, block sizes, the minimum and maximum number of samples, and the total number of discretised points per block (Amwaama, 2018). Kriging neighbourhood analysis minimizes the conditional bias before running a Kriging estimate (Vann *et al.*, 2003).

Search Range

The search range was defined by the output data from variogram models, this includes the search angles, the variogram range (Asghari et al., 2020). Search ranges were defined per

parameter CW (cm), Au (g/t) and Au (cm.g/t). The search range was determined per domain in this research domains were defined as (domains 0 and 1).

Block Discretisation

Block Discretisation is the process of dividing data into the X,Y, and Z dimensions as block sizes which are to be interpolated (Manchuk, 2010). This process is important in the geostatistical workflow as it provides a gridded representation of the variation of rock properties (Sunday & Deutsch, 2020) in this case grades and channel width values. Point data was used in defining data to block average covariances (Sunday & Deutsch, 2020). Block variance and Kriging variance were used to determine the optimal discretised block size and it was calculated per attribute in each domain.

Minimum and Maximum Samples

To run a successful Kriging neighbourhood analysis, it is important to set the dimensionality of the data. In this project, the data is represented in a 2D which affects the vertical and horizontal distribution of data. To determine the optimal minimum and maximum samples, the block size in 2D was set at x=30 m by y=30 m and z=0 m (representing no elevation).

Kriging efficiency scores were plotted against the number of search data for both grade and channel width. The KEFF and the SLOR were plotted and used as guidance to select the optimal minimum and maximum samples.

### 3.10 Kriging Block Modelling and Validation

As explained above, it is important to determine the dimension in which data will be presented (Hao *et al*., 2023) in this case, a 2D grid was designed at a 30 m by 30 m configuration. The block model was generated in Datamine modelling software where the block properties were used to develop a graphical illustration of the grid configuration. Upon completion of the mineral resource estimation and the generation of the block model, model validation is carried out. The model validation is important as it checks the accuracy of the model and ensures that the model is reliable and useful (Rossi and Deutsch, 2013). The performance of the model is critical in making well-informed decisions for mine planning and mine design purposes (Glacken *et al.,* 2001). The number of data points selected impacts model accuracy as fewer data points have a great impact on the overall performance of the mineral resource estimation model (Rossi and Deutsch, 2013).

### 3.10.1 Kriging efficiency and slope of regression

The performance of the model was assessed using KEFF and the SLOR, these are common Kriging evaluation matrices used in the mining industry (Deutsch *et al.*, 2014). They measure the closeness of our estimation and measure the rate of conditional bias.

1. **Kriging Efficiency (KEFF**) **-** Statistical tool/method that measures the effectiveness of the Kriging estimate and is used to analyse the accuracy of the resultant local block estimate. It is described by the eq.25 below (Deutsch *et al.*, 2014):

   $$\text{Kriging Efficiency} = \text{KEFF} = \frac{\sigma_B^2 - \sigma_K^2}{\sigma_B^2} \quad (25)$$

2. **Slope Of Regression (SLOR) -** This statistical tool is considered the conditional bias slope in which it measures the extent/degree of the smoothing effect of both high and low grades within a selected block or domain. Illustrated by the eq.26 below (Deutsch, 2007).

   $$\text{Slope of Regression} = \text{SLOR} = \frac{\sigma_B^2 - \sigma_K^2 + \lambda}{\sigma_B^2 - \sigma_K^2 + 2\lambda} \quad (26)$$

Where:

$\sigma_B^2$ = Block Variance

$\sigma_k^2$ = Kriging Variance

$\lambda$ = LaGrange Multiplier

### 3.10.2 Declustered sample grades versus block estimation grades

Swath plots were used in this research project to compare the declustered sample grades and block grades. These plots can also be used to compare the performance of different estimation methods, which is not within the scope of this project. (Goncalves, 2021) in his thesis demonstrated the comparison between OK, IWD, and radial basis functions methods. The swath plots take an average grade along a certain direction of set intervals. Since this is a 2D project the N-S and E-W directions were considered at a constant 50 m intervals. The swath plots output file contains the number of samples used in each slice/swath represented in a histogram. Absent data grades are also recorded in the sample value records. A pair of line

graphs are generated to compare the model estimation and the sample grades by visually inspecting the correlation between the highs and the lows of the resulting line graphs.

# 4 RESULTS

This chapter of the thesis covers the results from the methodology section, it covers the initial results from the exploratory data analysis process. Presentation of domaining results from the point domain analysis follows then the ML generated boundaries are reviewed. Variogram analysis results are then presented showing the spatial structures generated from the domain data, where spatial properties of the data are deduced for further analysis. The QKNA data is then presented, where optimal Kriging parameters are selected. Declustering and weighting results are then presented, furthermore estimated grades and channel thickness are also presented. These are displayed in a 2D block model, and the resultant performance scores are also outlined in this section.

### 4.1.1 Data preprocessing

The initial raw data consisted of 5610 data entries, and it is made up of the spatial locations of data points, where x=southing coordinates/east-west axis, y=westing coordinates/north-south axis, and z=elevation/height (was not considered for this project) shown in Table 4.1 below. The data also includes the CW (cm) which is the length of sampling zone, the gold grade values in centimetre-grams per ton Au (cm.g/t). It also encompasses the sedimentological descriptions, which are pebble sizes, packing, roundness, and sorting which all describe the maturity and composition of the sedimentary unit of the Carbon Leader Reef. Finally geological descriptions of the mineralisation such as colour, the conglomerate percentage and basal contact that describes the nature of the break between different lithological units show in Table 4.2 below.

**Table 4.1:** Spatial locations (X, Y and Z coordinates) showing the first 5 points.

| X | Y | Z |
|---|---|---|
| 2354.5 | -21088.2 | -2343.2 |
| 2357.1 | -21074.5 | -2334.1 |
| 2363.9 | -21080.9 | -2356.2 |
| 2364.3 | -21074 | -2360.3 |

**Table 4.2:** Features from the primary data including grade data and sedimentological data showing the first 5 rows of the data matching the above coordinates

| **Conglomerate (%)** | **CW (cm)** | **Au (cm.g/t)** | **Pebble Size (mm)** |
|---|---|---|---|
| 22.0 | 11 | 4742 | 20 |
| 60.0 | 12 | 3671 | 20 |
| 77.8 | 21 | 8203 | 22 |
| 76.5 | 13 | 7159 | 30 |
| **Basal Contact** | **Packing** | **Roundness** | **Sorting** |
| Sharp | Well | Sub-rounded | Moderate |
| Sharp | Well | Well rounded | Moderate |
| Sharp | Pebbly | Sub-rounded | Moderate |
| Sharp | Well | Sub-rounded | Moderate |
| **Sulphides Mineralisation** | **Carbon** | **Pebble Assemblage** | **Reef Color** |
| Low | Low | Oligomictic | Dark Grey |
| Low | Low | Oligomictic | Light |
| Low | None | Oligomictic | Dark Grey |
| High | Low | Oligomictic | Dark Grey |

### 4.1.2 Study area selection

An area selection for the study was carried out, this selection was a data density decision because the selected area contains a significant number of data points which is good for estimation purposes. The study area boundary was set on; minimum-X = -21,000, maximum-X = -22,500, minimum-Y = -7,000 and maximum-Y = 9000. The selection of this study area resulted in the reduction of datapoints from 5,610 to 2,017. The study area is displayed in (Fig 4.1) below.

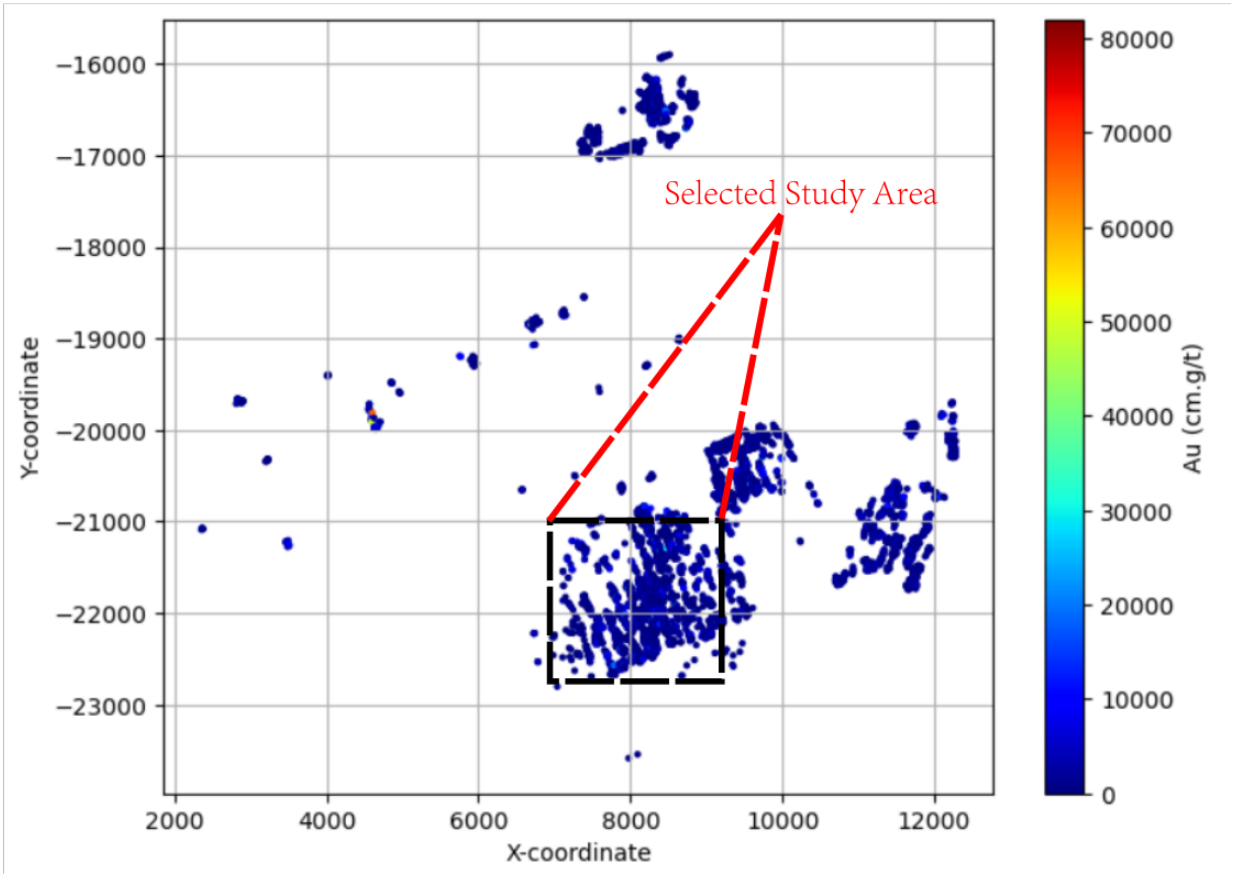


**Figure 4.1:** Outline of the selected study area (dotted black square) which was trimmed from the whole dataset because of the density of data points around the area.

## 4.2 Exploratory Data Analysis

The gold grade received from the mine was in centimetre-grams per ton Au (cm.g/t) so it was converted to grams per tonne Au (g/t). This was done by dividing the Au (cm.g/t) by the CW (cm). Au (g/t) is a feature that represents actual gold grades and has similar distribution as Au (cm.g/t) as one is the function of the other. Table 4.3 represents the resultant column of the grade calculated from the gold accumulation and channel width (highlighted in blue).

**Table 4.3:** Table showing the first five data columns for grade values Au (g/t) (highlighted in blue lower right corner) values converted by dividing Au (cm.g/t) (highlighted in yellow) and the CW (cm) (highlighted in orange).

| **Au (cm.g/t)** | **CW (cm)** | **Pebble Size (mm)** | **Packing** |
|---|---|---|---|
| 673.4240981 | 86.54614 | 20 | Well |
| 508.0313325 | 80.49343 | 20 | Well |
| 310.1647412 | 72.64041 | 17 | Scattered |
| 349.9607355 | 30.69831 | 20 | Scattered |
| | | | |
| **Roundness** | **Sorting** | **Sulphide Mineralisation** | **Carbon** |
| Well rounded | Moderate | High | |
| Sub-rounded | Moderate | High | |
| Well rounded | Poor | High | Low |
| Sub-rounded | Moderate | Low | |
| | | | |
| **Pebble Assemblage** | **Reef Color** | **Basal Contact** | **Au (g/t)** |
| Oligomictic | Dark Grey | Sharp | 7.781098902 |
| Oligomictic | Dark Grey | Sharp | 6.311463415 |
| Oligomictic | Grey | Sharp | 4.269864865 |
| Oligomictic | Grey | Gradational | 11.4 |

The resultant subset of data contained 16 features and is a mixture of both numerical and categorical data. Numerical data included location data (X, Y, and Z), the conglomerate percentages, pebble sizes, channel width, the gold grade Au (g/t) and Au (cm.g/t). In Python numerical data can be represented as an Integer (“int”) for whole numbers, floating point (“float”) for real numbers with decimal points, and categorical data can be represented as an object, string “str” for a sequence of characters (Table 4.4).

**Table 4.4**: Data types and number of null values per variable.

| Before Encoding | | | After Encoding | | |
|---|---|---|---|---|---|
| Column | Non_Null Count | Data type | Column | Non_Null Count | Data type |
| X | 727 | Float64 | X | 727 | Float64 |
| Y | 727 | Float64 | Y | 727 | Float64 |
| Z | 727 | Float64 | Z | 727 | Float64 |
| Conglomerate (%) | 727 | Float64 | Conglomerate (%) | 727 | Float64 |
| CW (cm) | 727 | Float64 | CW (cm) | 727 | Float64 |
| Au (cm.g/t) | 727 | Float64 | Au (cm.g/t) | 727 | Float64 |
| Pebble size (mm) | 727 | int64 | Pebble size (mm) | 727 | int64 |
| Packing | 727 | object | Packing | 727 | int32 |
| Roundness | 727 | object | Roundness | 727 | int32 |
| Sorting | 727 | object | Sorting | 727 | int32 |
| Sulphides Mineralisation | 727 | object | Sulphides Mineralisation | 727 | int32 |
| Carbon | 727 | object | Carbon | 727 | int32 |
| Pebble Assemblage | 727 | object | Pebble Assemblage | 727 | int32 |
| Reef Colour | 727 | object | Reef Colour | 727 | int32 |
| Basal Contact | 727 | object | Basal Contact | 727 | int32 |
| Au (g/t) | 727 | float64 | Au (g/t) | 727 | float64 |

### 4.2.1 Categorial data conversion

To convert non-numerical data to numerical data, a one-hot encoder transformation code was deployed and transformed the categorical data into a set of integer values (Table 4.5).

One-hot encoding is a technique used to convert categorical variables into a numerical format. It works by creating a new binary column for each unique category in the original feature in shown in columns of Table 4.5 below, assigning a value of 0,1 and 2 to indicate the presence of a category. This method prevents models from assuming any ordinal relationship between categories, which can occur with simple label encoding. While effective, one-hot encoding can lead to high-dimensional data when applied to features with many unique values, resulting in sparse matrices that may increase computational complexity as such it should be considered if fit for purpose and domain knowledge applies.

**Table 4.5:** Python extract of the encoded values from the categorical data showing the first 5 columns of the data (represented in 0,1 and 2).

| Packing | Roundness | Sorting | Sulphides Mineralisation | Carbon | Pebble Assemblage | Reef colour | Basal Contact |
|---|---|---|---|---|---|---|---|
| 2 | 2 | 0 | 0 | 2 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 2 | 0 | 0 | 1 |
| 1 | 2 | 1 | 0 | 1 | 0 | 2 | 1 |
| 1 | 1 | 0 | 1 | 2 | 0 | 2 | 0 |
| 0 | 1 | 1 | 0 | 2 | 0 | 0 | 0 |

### 4.2.2 Feature analysis

A feature analysis was carried out to understand the properties of the 16 features mentioned above. In this section, only quantitative numerical data results are presented because of the ease of plotting and comparing them to one another on a particular scale. Summary statistics data is represented in (Table 4.6) below.

**Table 4.6:** Summary statistics for the selected numerical features.

| | Conglomerate (%) | CW (cm) | Pebble Size (mm) | Au (g/t) | Au (cm.g/t) |
|---|---|---|---|---|---|
| **count** | 2017.00 | 2017.00 | 2017.00 | 2017.00 | 2017.00 |
| **mean** | 51.45 | 105.78 | 16.24 | 26.67 | 2089.73 |
| **std** | 33.99 | 51.70 | 8.16 | 52.98 | 3430.05 |
| **min** | 2.40 | 6.00 | 0.00 | 0.00 | 0.00 |
| **25%** | 21.00 | 67.00 | 11.00 | 4.27 | 385.00 |
| **50%** | 42.70 | 105.00 | 15.00 | 11.47 | 1062.00 |
| **75%** | 90.00 | 143.00 | 20.00 | 26.98 | 2427.00 |
| **max** | 100.00 | 268.00 | 98.00 | 800.49 | 6037.00 |

To understand the relationship between the target variables and the rest of the features, a Pearson correlation matrix is generated to display the strength of the correlation between variables (Fig 4.2).

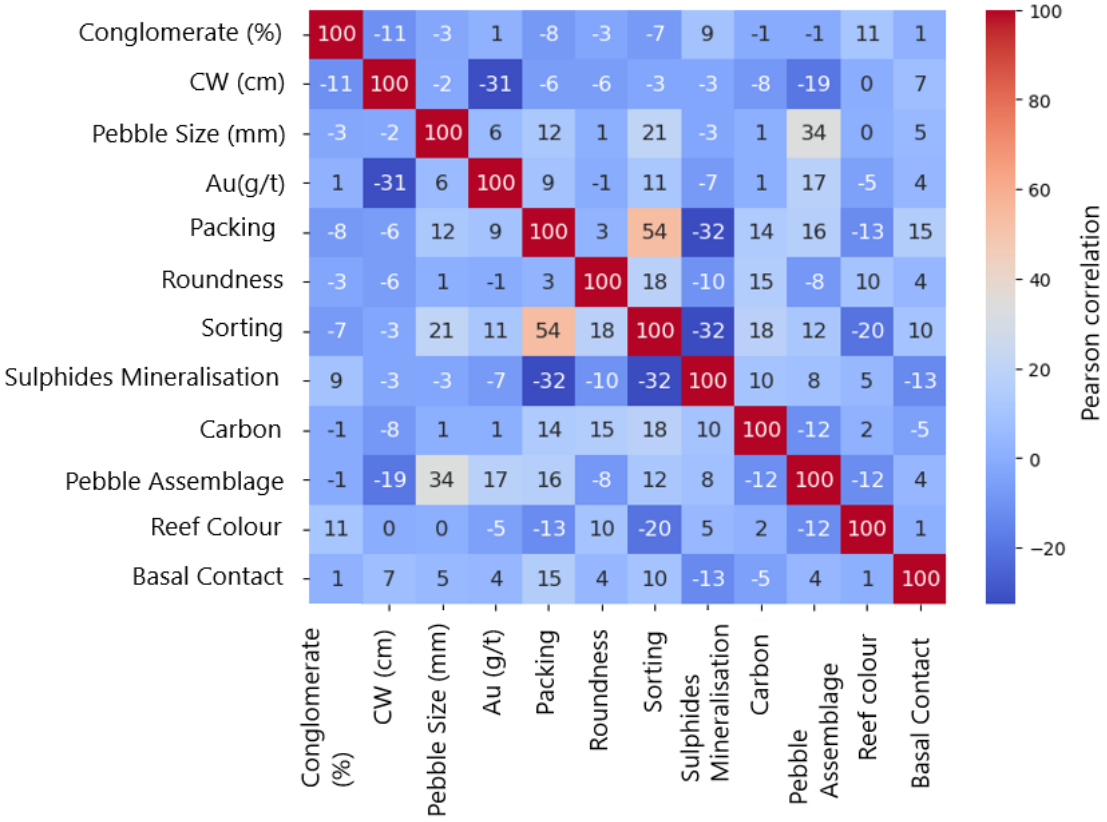


**Figure 4.2:** Feature correlation heat map showing the feature correlation from the primary dataset.

Target Variables statistics

The cumulative frequency distributions and spatial distribution the target variables, Au (cm.g/t), Au (g/t) and CW (cm), shown in (Fig 4.3, Fig 4.4, Fig 4.5, Fig 4.6 and Fig 4.7).

Au (g/t)

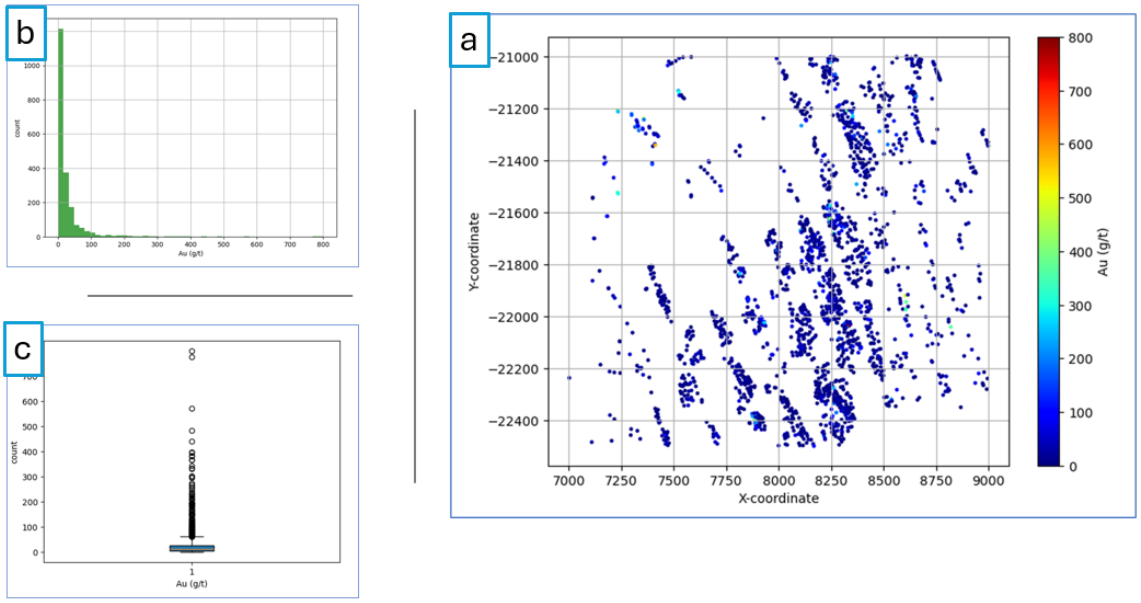


**Figure 4.3:** Au (g/t) graphical statistics and spatial distribution a) Spatial distribution of Au (g/t) values as shown in the study area, b) Histogram indicating frequency of the occurrence of Au (g/t) values c) Box plot showing the dispersion of outliers from the mean.

Au (cm.g/t)

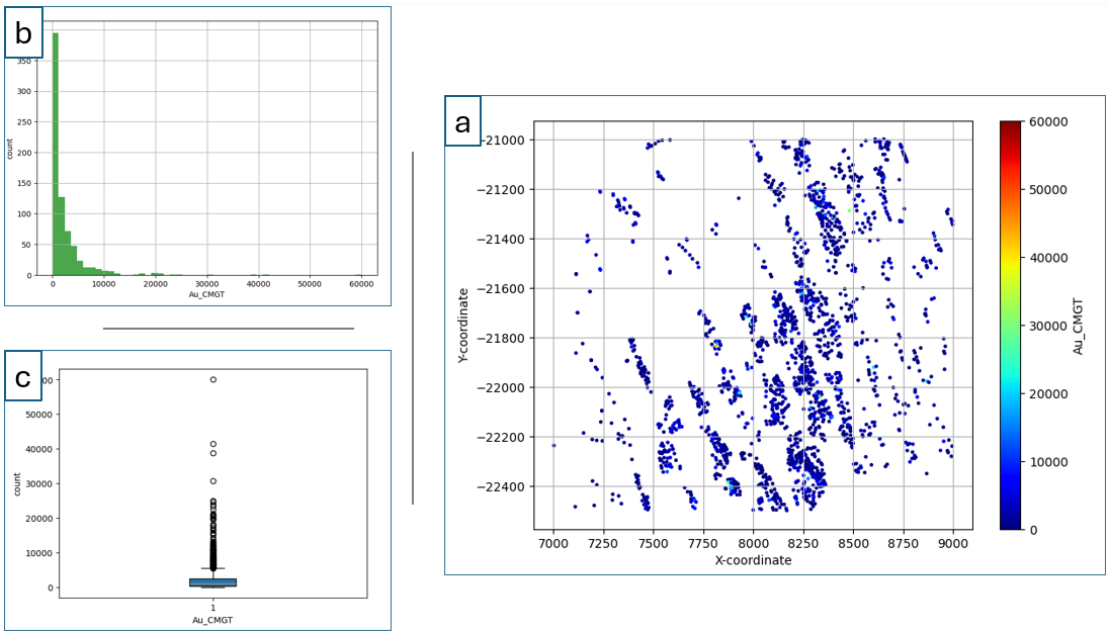


**Figure 4.4**: Au (cm.g/t) graphical statistics and spatial distribution a) Spatial distribution of Au (cm.g/t) values as shown in the study area, b) Histogram indicating frequency occurrence of Au (cm.g/t) values c) Box plot displaying the concentration of data around the mean and some outliers detected beyond that.

Channel width CW (cm)

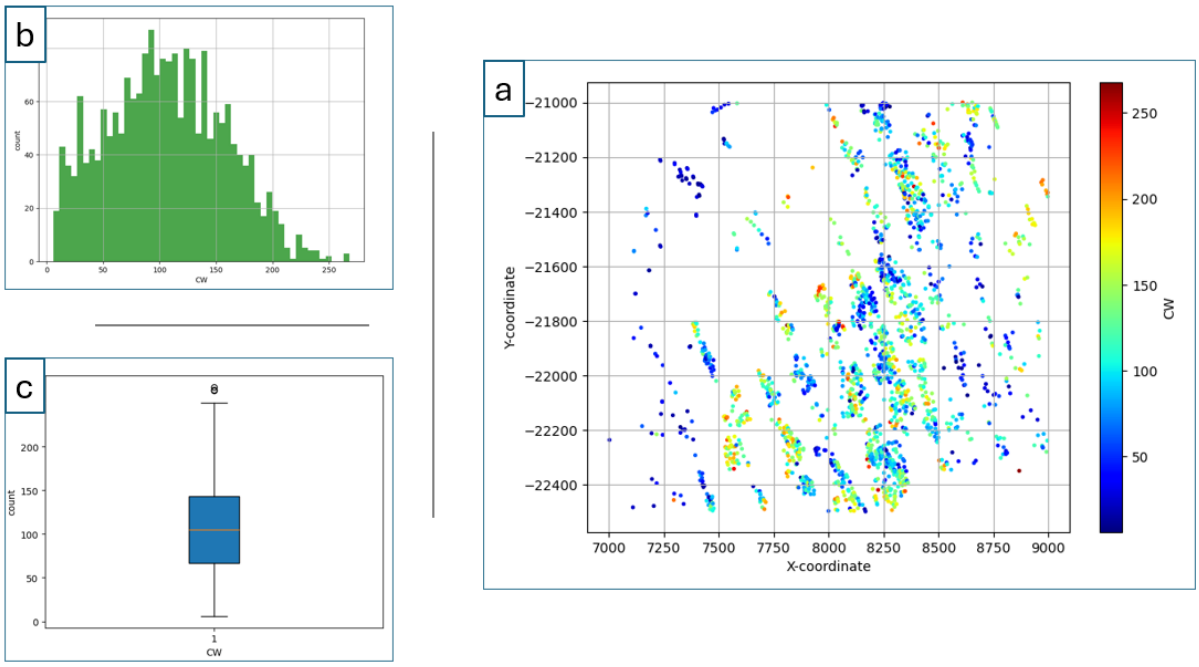


**Figure 4.5:** CW (cm) graphical statistics and spatial distribution a) Spatial distribution of CW (cm) values as shown in the study area, b) Histogram indicating frequency occurrence of CW (cm) c) Box plot showing a balanced occurrence of high and low CW values.

Conglomerate percentage

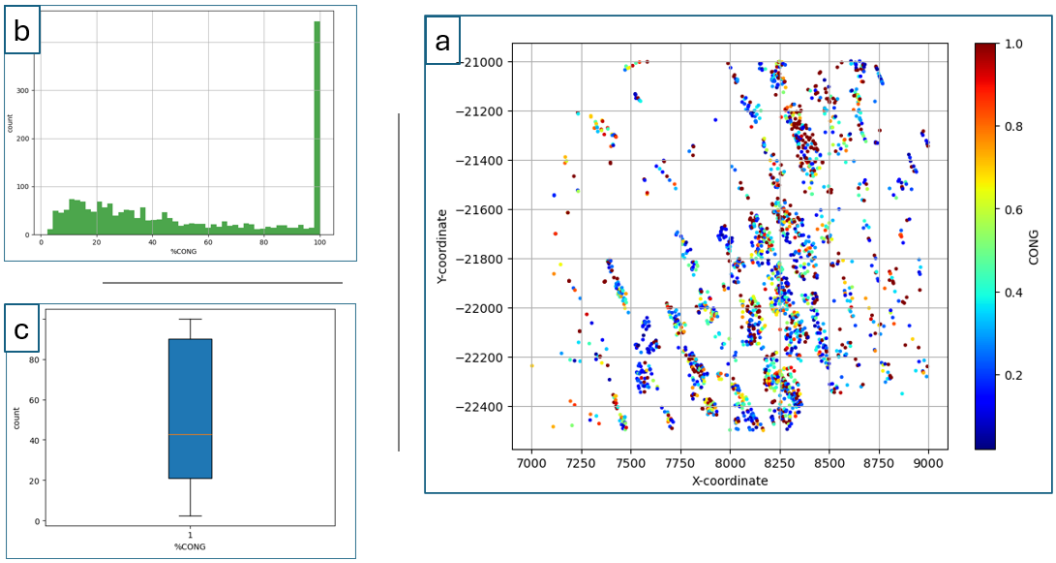


**Figure 4.6:** Conglomerate (%) graphical statistics and spatial distribution a) Spatial distribution of conglomerate (%) values as shown in the study area, b) Histogram indicating frequency occurrence of conglomerate (%) values c) Box plot showing well distributed conglomerate (%).

Pebble size (mm) statistics

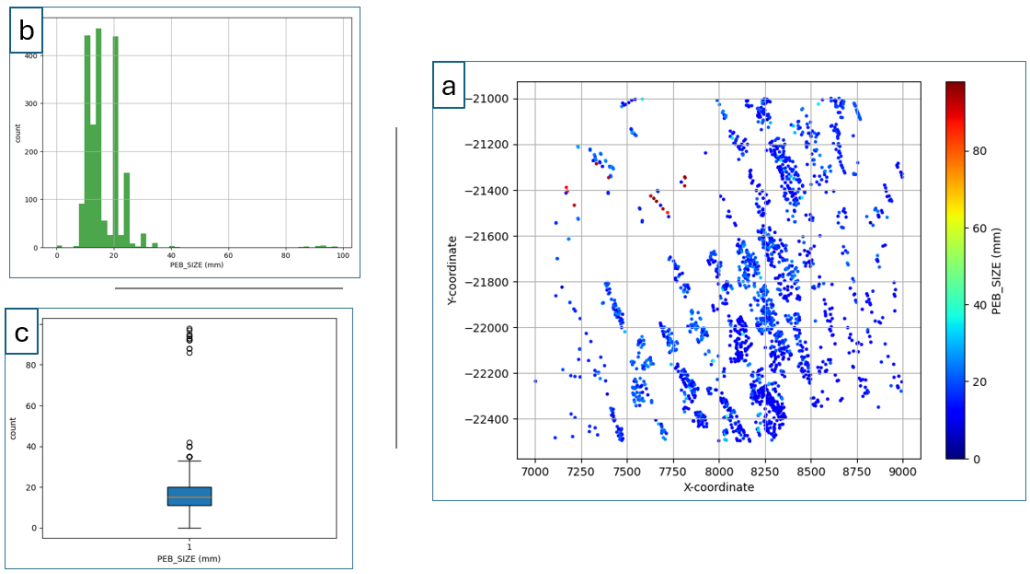


**Figure 4.7**: Pebble size (cm) graphical statistics and spatial distribution a) Spatial distribution of Pebble size (cm) values as shown in the study area, b) Histogram indicating frequency occurrence of Pebble size (cm) values c) Box plot showing the distribution of pebble sizes which are concentrated around 20 cm and a few high values.

## 4.3 Point Sample Clustering Analysis

Traditional K-means and Signed-distance-mds-K-means cluster analysis results are shown in (Fig 4.7). Silhouette Score and the Davies Bouldin Index are used to select optimal number of clusters using the elbow method shown in (Fig 4.8a and Fig 4.8b).

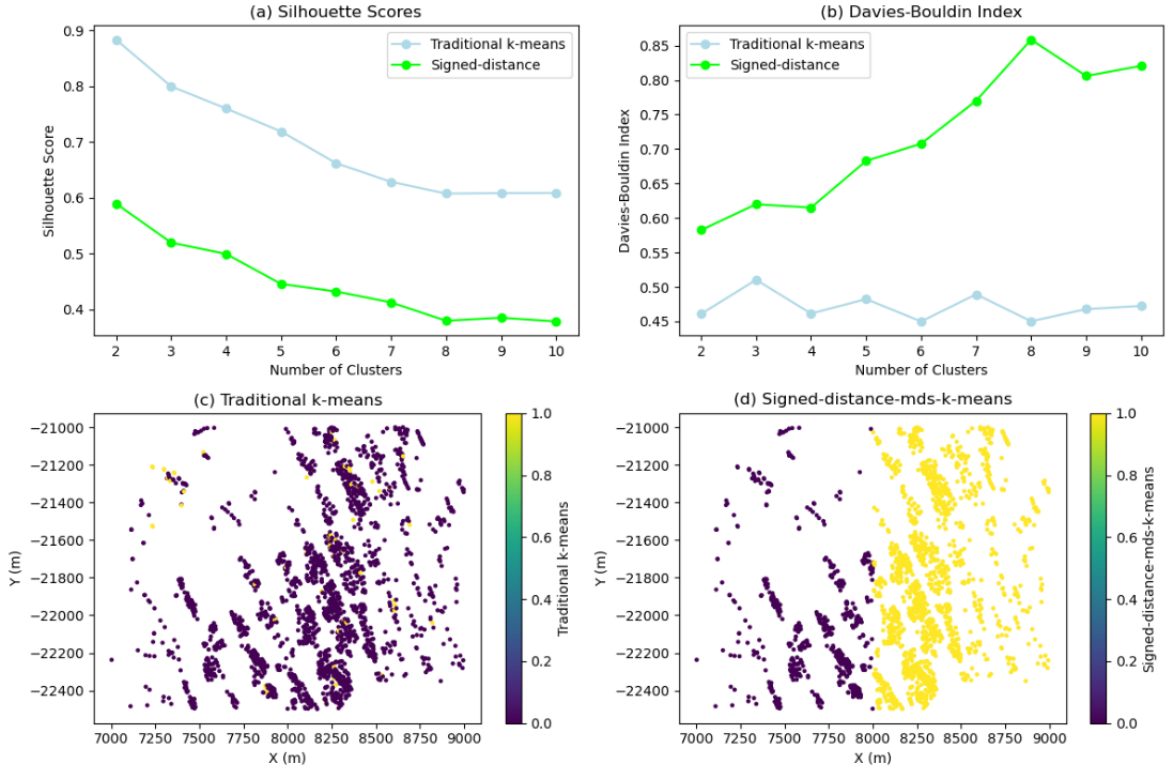


**Figure 4.8:** Cluster analysis a) Silhouette Scores b) Davies-Bouldin Index c) Traditional K-means clustering distribution d) Signed-distance-mds-K-means clustering distribution.

## 4.4 Cluster Configuration

The clusters generated from the above clustering are shown in (Fig 4.9) for Standard K-means clustering and for Signed-distance-mds-k-means in (Fig 4.9d) below. These diagrams show how well the clusters are divided and highlights cluster mixing in the Standard K-means clustering in (Fig 4.9c).

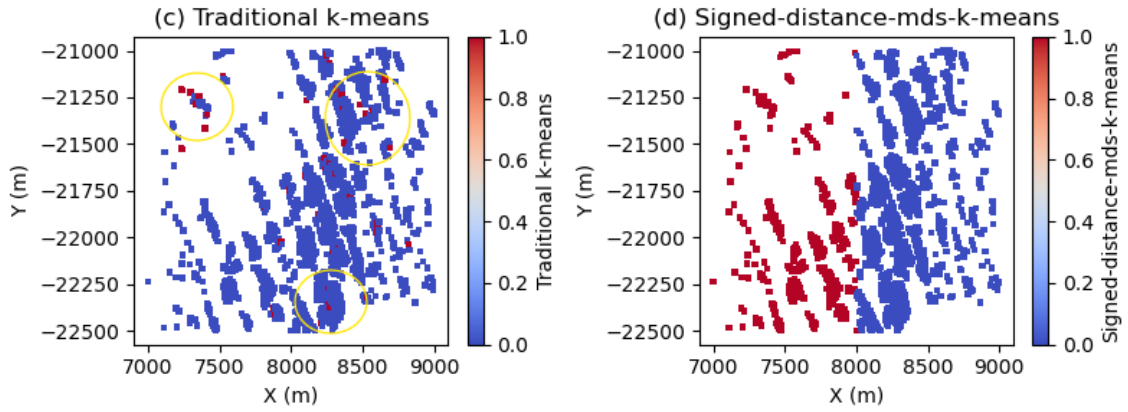


**Figure 4.9:** Cluster configuration c) Traditional K-means cluster layout showing cluster mixing. d) Signed-distance-mds-k-means cluster layout showing two distinct clusters.

Cluster statistics

Summary statistics and histograms from Domain 1 (Fig 4.10a) and Domain 0 (Fig 4.10b) after clustering, showing that Domain 0 has more data points more than Domain 1.

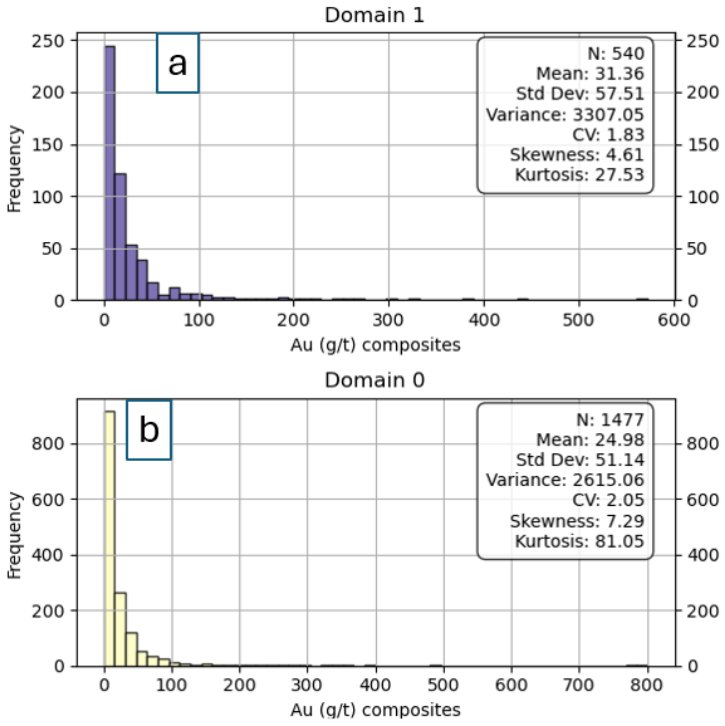


**Figure 4.10:** Histogram and summary statistics for individual domains a) Domain 1 b) Domain 0.

## 4.5 Post Clustering Feature Summary Statistics Per Domain

Summary statistics and histogram representation per domain per target variable before any normalisation or any transformation (Fig 4.11).

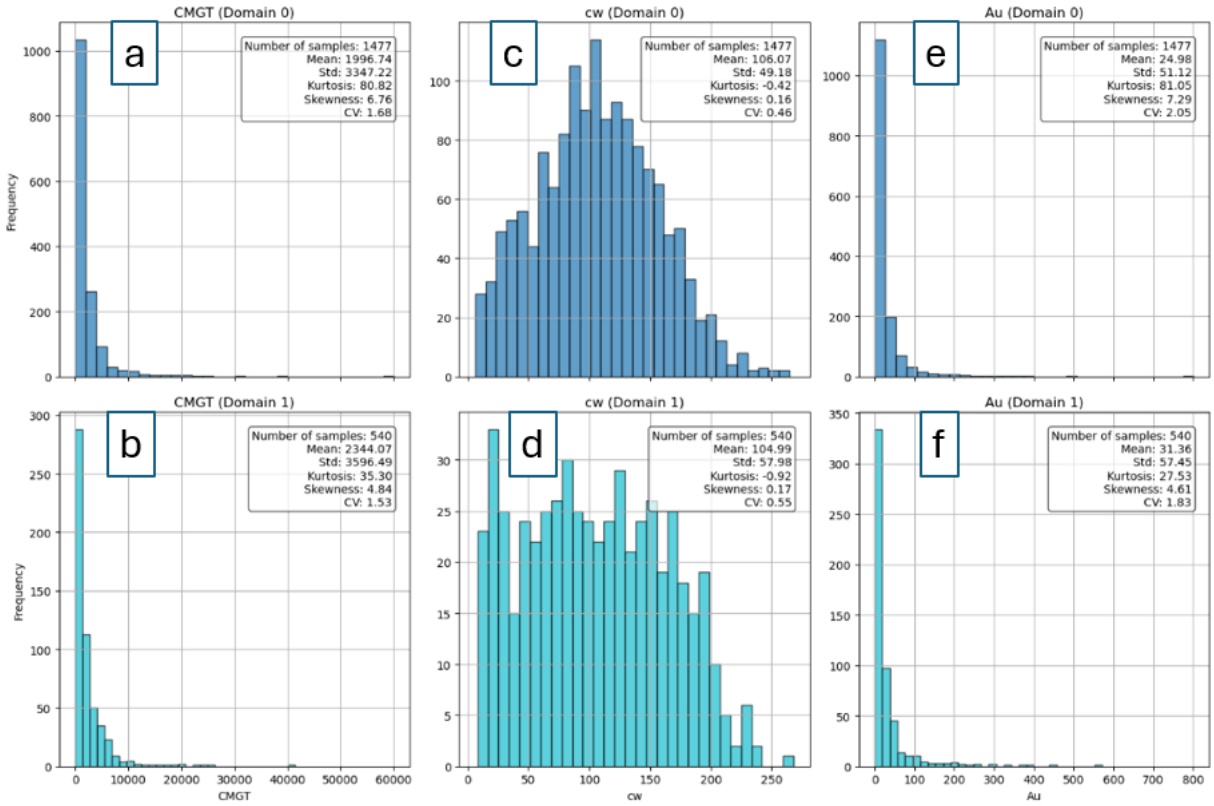


**Figure 4.11:** Feature histogram and summary statistics per domain a) Au (cm.g/t) Domain 0 b) Au (cm.g/t) Domain 1 c) CW (cm) Domain 0 d) CW (cm) Domain 1 e) Au (g/t) Domain f) Au (g/t) Domain 1.

### 4.5.1 Z–Score transformation

To deal with outliers in the feature data, the Z–Score transformation was done. Z–Score is a statistical transformation tool used to convert skewed data to a somewhat normal distribution by converting each point in the data set to the population mean and standard deviation. The Z- scores histograms per domain per target variable are shown in (Fig 4.12) and the resultant Z-score values are displayed in Table 4.7 below.

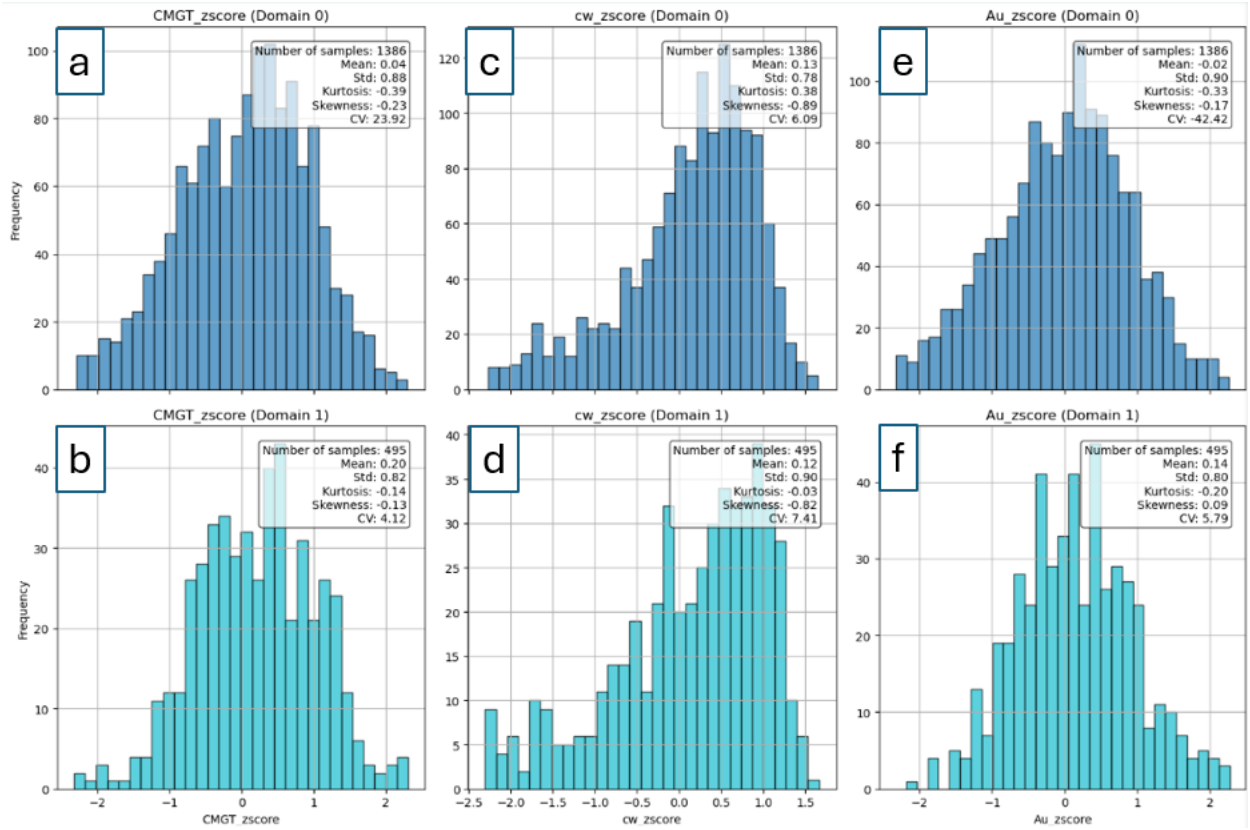


**Figure 4.12:** Feature Z–score histograms and the summary statistics per domain a) Au (cm.g/t) Z-score (Domain 0) b) Au (cm.g/t) Z-score (Domain 1) c) CW (cm) Z-score (Domain 0) d) CW (cm) Z-score (Domain 1) e) Au (g/t) Z-score (Domain 0) f) Au (g/t) Z-score (Domain 1).

**Table 4.7:** Z-Score values (highlighted in orange) generate for selected feature per domain (Domain 0 highlighted in blue and Domain 1 highlighted in yellow).

| X | Y | Z | CW (cm) | Au (cm.g/t) | Au (g/t) |
|---|---|---|---|---|---|
| 7,001.89 | -22,236.80 | -3,136.33 | 21.50 | 567.00 | 26.37 |
| 7,108.91 | -22,483.61 | -3,221.60 | 23.37 | 543.00 | 23.23 |
| 7,112.47 | -21,863.40 | -3,022.10 | 41.69 | 2,045.00 | 49.05 |
| 7,113.28 | -21,545.40 | -2,879.96 | 86.55 | 673.42 | 7.78 |
| 7,113.64 | -21,544.52 | -2,880.36 | 80.50 | 508.03 | 6.31 |
| 8,991.20 | -22,280.25 | -3,426.50 | 132.00 | 133.00 | 1.01 |
| 8,994.00 | -21,336.36 | -3,030.80 | 201.00 | 2,853.00 | 14.19 |
| 8,995.00 | -22,239.31 | -3,408.30 | 107.00 | 1,270.00 | 11.87 |
| 8,997.40 | -21,345.06 | -3,034.60 | 123.00 | 1,731.00 | 14.07 |
| 8,997.60 | -21,330.41 | -3,029.40 | 209.00 | 1,720.00 | 8.23 |
| | | | | | |
| **Domain** | **Au (cm.g/t) Z-score** | **CW (cm) Z-score** | **Au (g/t) Z-score** | | |
| 1 | -0.34 | -2.16 | 0.64 | | |
| 1 | -0.37 | -2.03 | 0.55 | | |
| 1 | 0.56 | -1.16 | 1.05 | | |
| 1 | -0.22 | -0.05 | -0.19 | | |
| 1 | -0.42 | -0.16 | -0.33 | | |
| 0 | -1.35 | 0.59 | -1.57 | | |
| 0 | 0.79 | 1.23 | 0.22 | | |
| 0 | 0.22 | 0.27 | 0.09 | | |
| 0 | 0.44 | 0.48 | 0.21 | | |
| 0 | 0.44 | 1.29 | -0.15 | | |

### 4.5.2 Box and whisker plot analysis of outlier treated data

Outlier treated data per domain for each target variable shown in (Fig 4.13) below. Domain 0 data points reduced from 1477 shown in (Fig 4.11) to 1386 (Fig 4.12) and Domain 1 data points reduced from 540 shown in (Fig 4.11) to 495 (Fig 4.12). The outlier treatment shows a good selection of several points that does not significantly impact the data. Overall, the data points reduced from 2017 to 1881 which is a difference of 136 data points which is just 6.74% of the total.

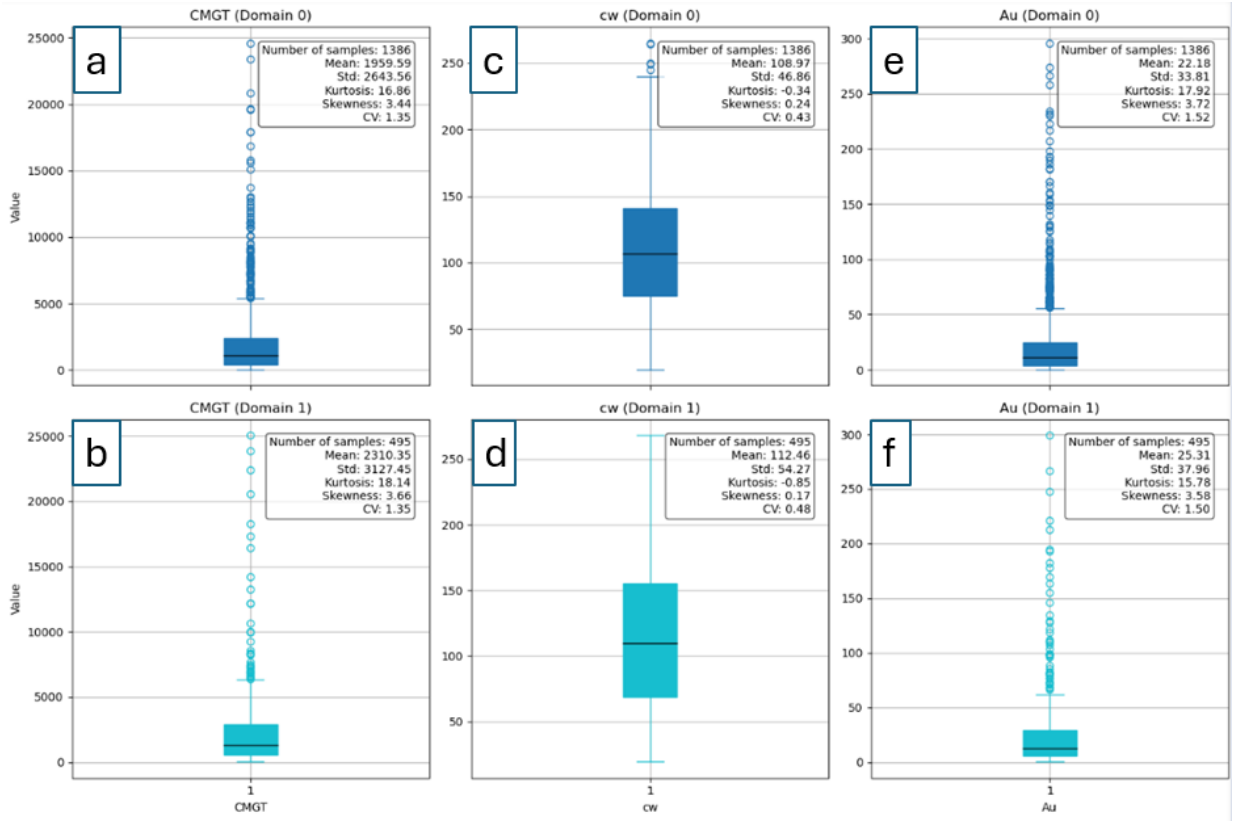


**Figure 4.13:** Outlier treated data represented as Whisker plots and summary statistics a) Au (cm.g/t) Domain 0 b) Au (cm.g/t) Domain 1 c) CW (cm) Domain 0 d) CW (cm) Domain 1 e) Au (g/t) Domain 0 f) Au (g/t) Domain 1.

#### 4.5.2.1 Domain pairwise comparison and correlation plots

The comparative pairwise plots between the three target variables and their geographical locations are shown in (Fig 4.14) for Domain 0 and (Fig 4.16) for Domain 1. A correlation analysis was done per domain for all the target variable for domain 0 is shown in (Fig 4.15) and Domain 1 shown in (Fig 4.17).

Domain 0 Pairwise plots

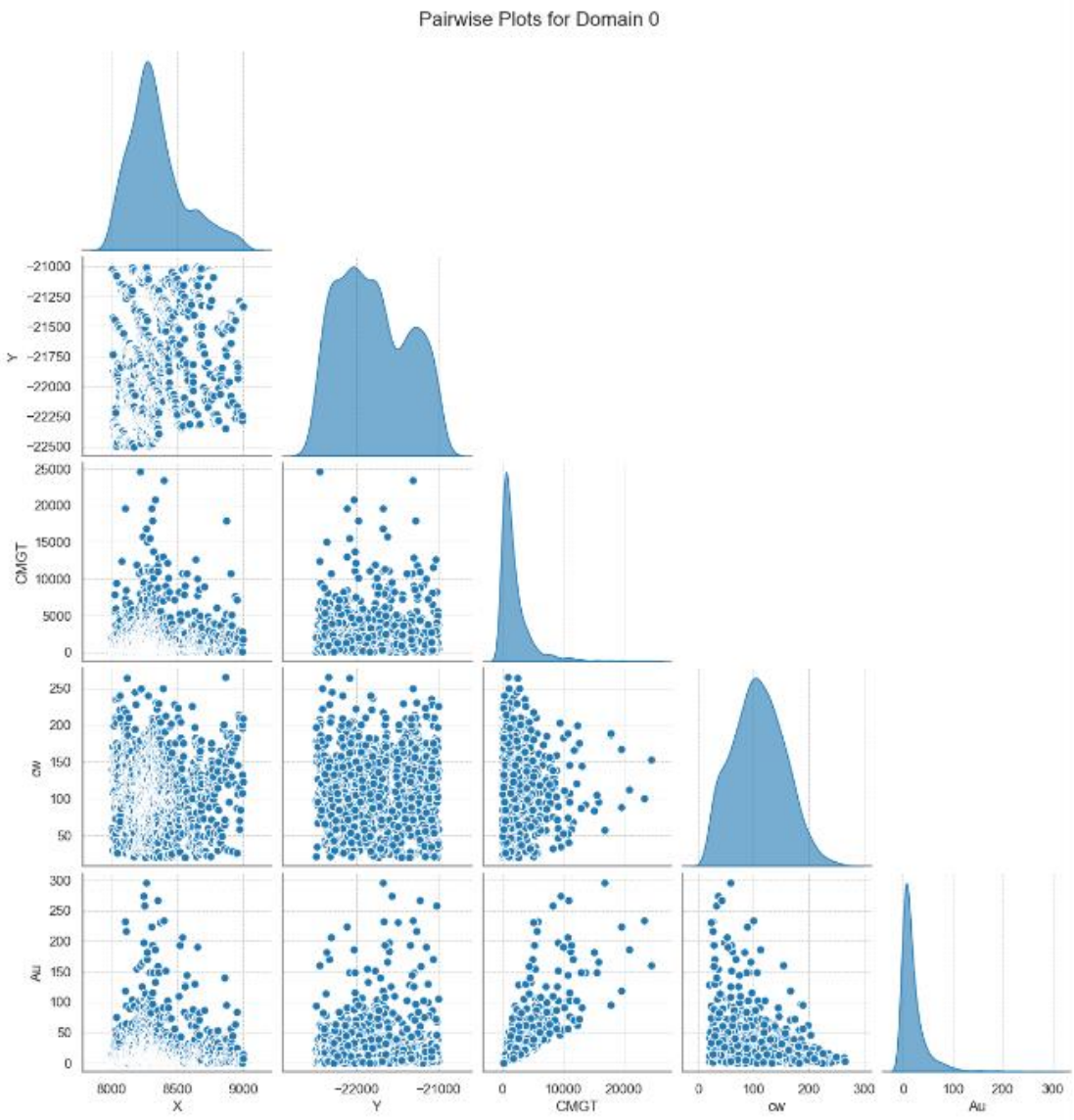


**Figure 4.14:** Pairwise plots for all the selected variables of interest in Domain 0.

The strong positive correlation between the Au (cm.g/t) and Au (g/t) is shown in the correlation matrix below (Fig 4.15). The correlation value is at 81% for the two variables, which is the highest correlation amongst all the variables analysed. The obvious correlation is the X-axis and Y-axis values as they are directly proportional values related to the geographical location of the sample points. These relationships are shown also in the pairwise plot in Fig 4.14.

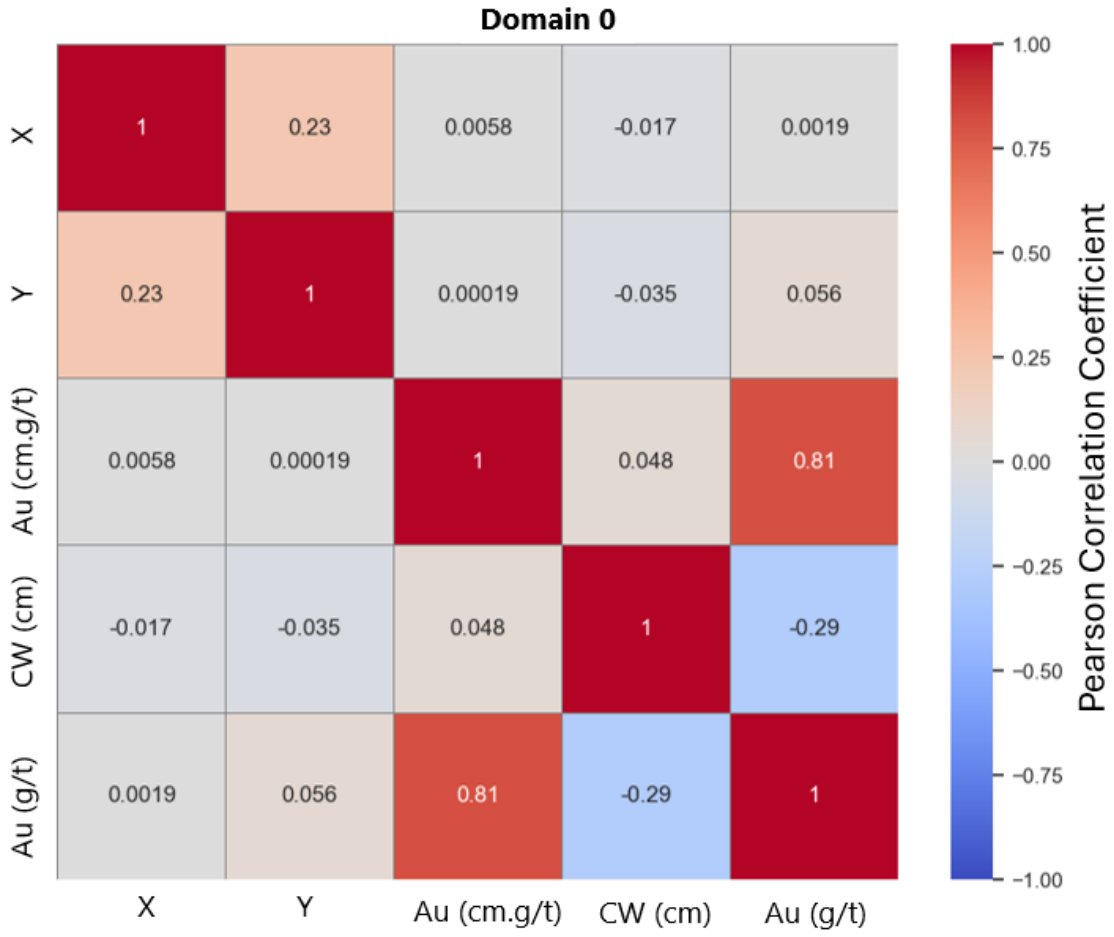


**Figure 4.15:** Pearsons's correlation heatmap of selected variables of interest Au (cm.g/t), Au (g/t) and CW (cm) and their spatial locations (Domain 0).

Domain 1 Pairwise plots

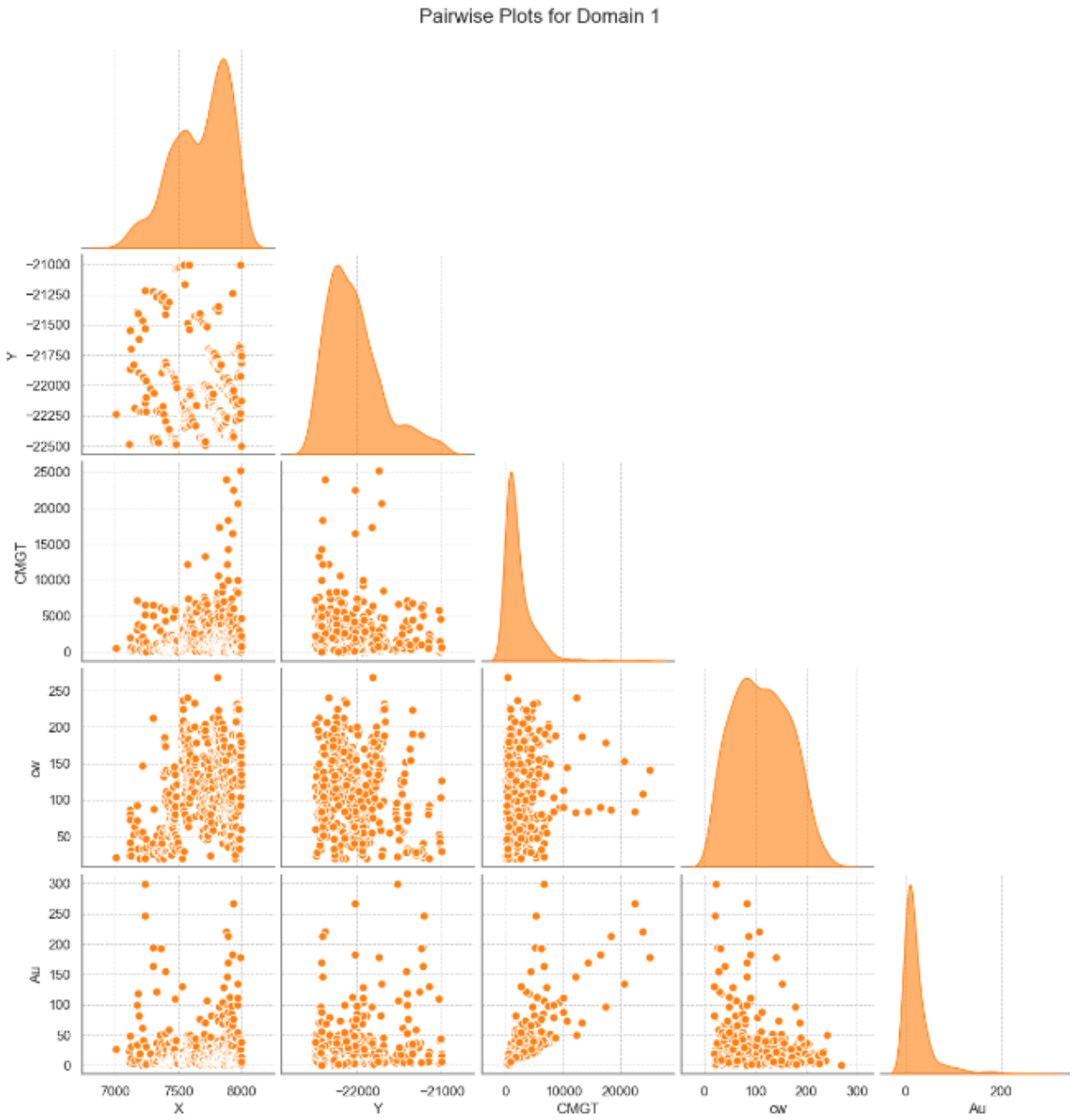


**Figure 4.16:** Pairwise plots for all the selected variables of interest in Domain 1.

Domain 1 correlation heatmap shows that the Au (cm.g/t) and Au (g/t) variables have the highest correlation at 77% (Fig 4.17). There are weak inverse positive and negative correlations between the X-axis, Y-axis and the Au (cm.g/t) these relationships are shown also in the pairwise plot in Fig 4.16. Au (g/t) and CW (cm) variables. Fig 4.17 shows that if the X-axis has a positive correlation with one of the variables of interest, then the Y-axis will have a negative correlation with it, this is also evident vice versa.

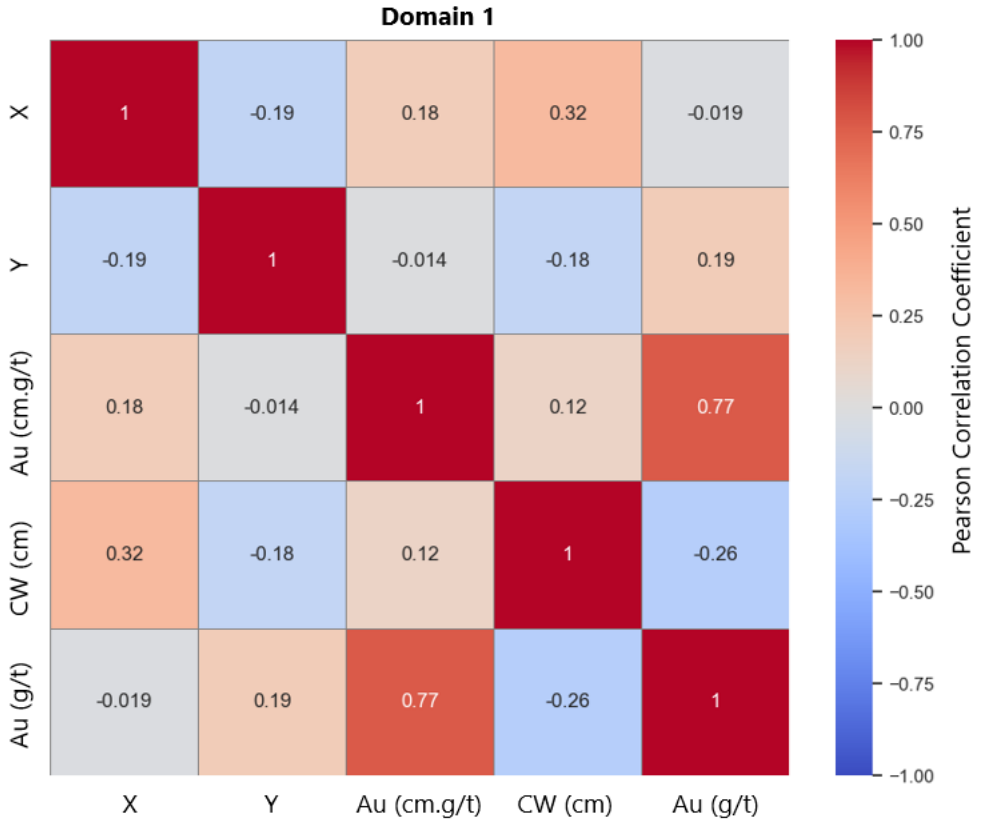


**Figure 4.17:** Pearson correlation heatmap of selected variables of interest Au (cm.g/t), Au (g/t) and CW (cm) and their spatial locations at Domain 1.

## 4.6 Domain Boundary Analysis

The domain boundaries generated from the ML model were displayed graphically to carry out the second step visual assessment of domain boundaries. These boundaries are shown in (Fig 4.18) below, representing domain 1 in yellow and Domain 0 in purple.

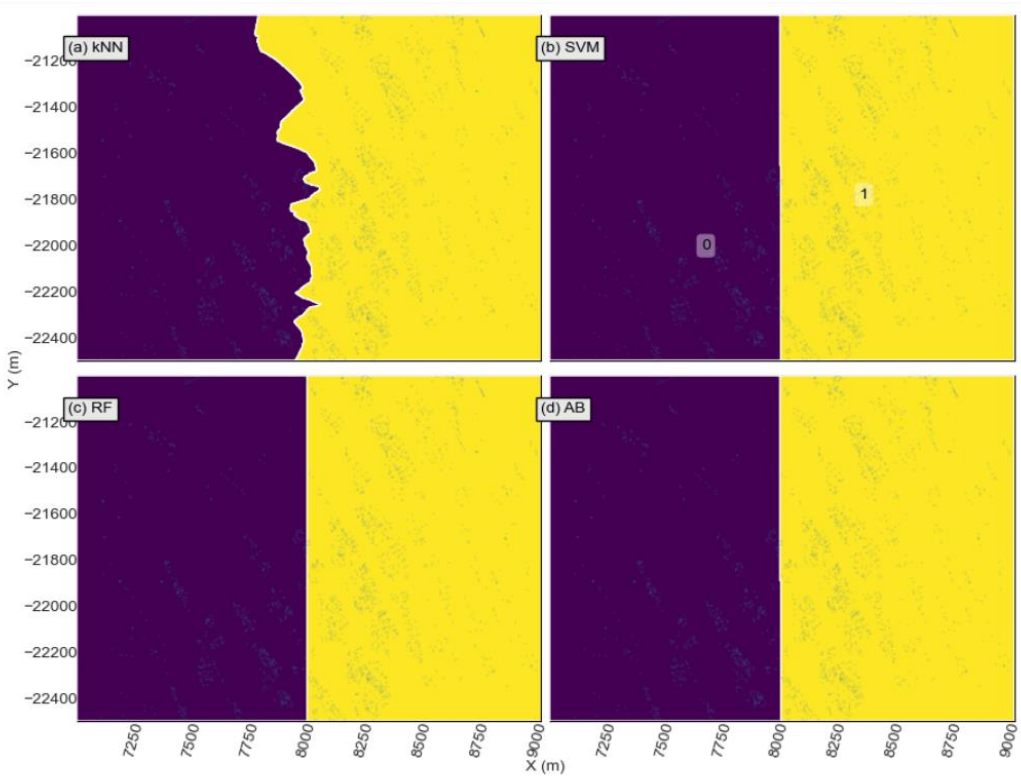


**Figure 4.18:** Domain boundaries generated by various ML algorithms a) kNN domain boundary b) SVM domain boundary c) RF domain boundary d) AB domain boundary.

### 4.6.1 kNN domain boundary

kNN generated the most geologically coherent domain boundary which is consistent with the Carbon Leader Reef deposit. It shows the sharp and undulating contact that is evident in the deposit as a sign of depositional energy transition, this will be explained in the discussion section of this paper. Fig 4.19 shows the spatial distributions of the target features overlain by the domain boundary polygon that will be later used for estimation in Datamine.

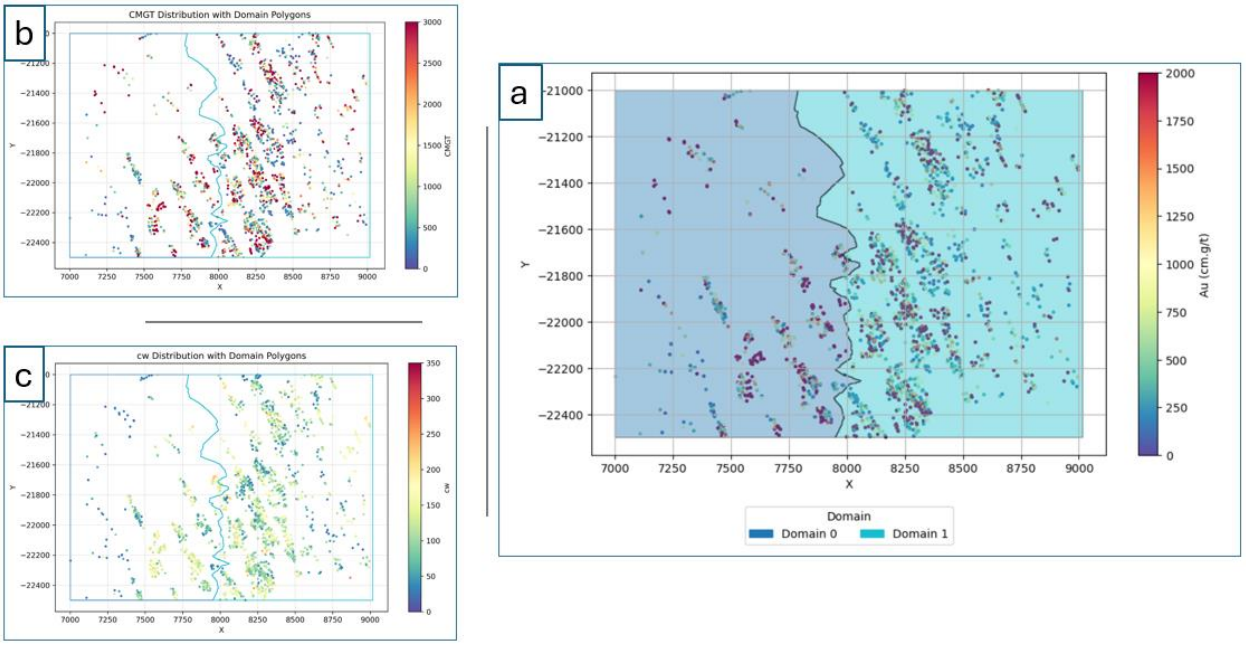


**Figure 4.19:** Domain boundary polyline showing demarcation of Domain 0 and Domain 1 features a) Au (g/t) spatial distribution b) Au (cm.g/t) spatial distribution c) CW (cm) spatial distribution.

### 4.6.2 Machine learning internal model parameter importance

The performance of ML algorithms is controlled by internal parameters; these parameters determine the optimisation of the learning process. They were fine tuned to determine the best parameters to improve for effective predictions and they are shown in Table 4.8 below.

### 4.6.3 Machine learning algorithm performance

To analyse the performance of the various models, the accuracy and F1 score model performance matrices were used to rank the models. This was done after running 10 iterations and a five-fold K-Fold cross-validation.

To examine the performance of the ML algorithms in generating domain boundaries, there were 10 iterations carried out and the results are shown in Table 4.9 below. The average of these iterations per ML algorithm are shown in (Fig 4.20 and Table 4.10).

**Table 4.8:** ML algorithm performance (accuracy score and F1 score) and internal best parameters.

| Algorithm | Call | Accuracy Score | F1 Score | Best Parameters | Model Name |
|---|---|---|---|---|---|
| kNN | KNeighboursCLassifier() | 0.9951 | 0.995 | {'n_neighbors':5} | kNN |
| SVM | SVC() | 0.9951 | 0.995 | {'C','gamma':0.0001,'kernel':linear'} | SVM |
| SVM_Man | SVC() | 1.0000 | 1.000 | {'C':100000,'gamma':0.01'} | SVM_Man |
| RF | RandomForestClassifier() | 1.0000 | 1.000 | {'Criterion':'gini','max_depth':3, 'max_feature… | RF |
| AB | AdaBoostClassifier() | 0.9975 | 0.998 | {'estimator':DecisionTreclassifier(),'estimator…. | AB |
| ANN | MLPClassifier() | 0.9926 | 0.993 | {'activation':'identity','alpha':0.001,'hi... | ANN |
| GP | GaussianProcessClassifier() | 0.9850 | 0.985 | {'kernel':RBF(length_scale=1), 'multi_class'…. | GP |

**Table 4.9:** ML Accuracy and F1 scores for 10 iterations.

| Iteration No | Accuracy - kNN | F1 Score - kNN | Accuracy - SVM | F1 Score - SVM | Accuracy - SVM_Man | F1 Score - SVM_Man | Accuracy - RF | F1 Score - RF | Accuracy - AB | F1 Score - AB |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.994051 | 0.994044 | 0.998513 | 0.998512 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.996034 | 0.996043 |
| 2 | 0.996034 | 0.996029 | 0.998513 | 0.998512 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.998513 | 0.998514 |
| 3 | 0.994051 | 0.994044 | 0.998017 | 0.998017 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.999008 | 0.999009 |
| 4 | 0.994546 | 0.994545 | 0.997025 | 0.997022 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.998008 | 0.999008 |
| 4 | 0.993059 | 0.993059 | 0.998017 | 0.998016 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.998513 | 0.998512 |
| 6 | 0.993555 | 0.993541 | 0.997521 | 0.99752 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.998513 | 0.998514 |
| 7 | 0.993555 | 0.993549 | 0.999008 | 0.999008 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.998017 | 0.999009 |
| 8 | 0.994546 | 0.994542 | 0.999008 | 0.999009 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.996034 | 0.996031 |
| 9 | 0.993555 | 0.993541 | 0.998513 | 0.998512 | 0.999008 | 0.999009 | 0.999504 | 0.999504 | 0.999504 | 0.999504 |
| 10 | 0.995042 | 0.995036 | 0.999504 | 0.999504 | 0.998513 | 0.998514 | 0.999504 | 0.999504 | 0.997025 | 0.997029 |

**Table 10:** ML Accuracy and F1 average calculations per model.

| Method | Accuracy | F1 |
|---|---|---|
| kNN | 0.994199 | 0.994191 |
| SVM | 0.998364 | 0.998363 |
| SVM_Man | 0.99886 | 0.99886 |
| RF | 0.999455 | 0.999455 |
| AB | 0.998017 | 0.998018 |

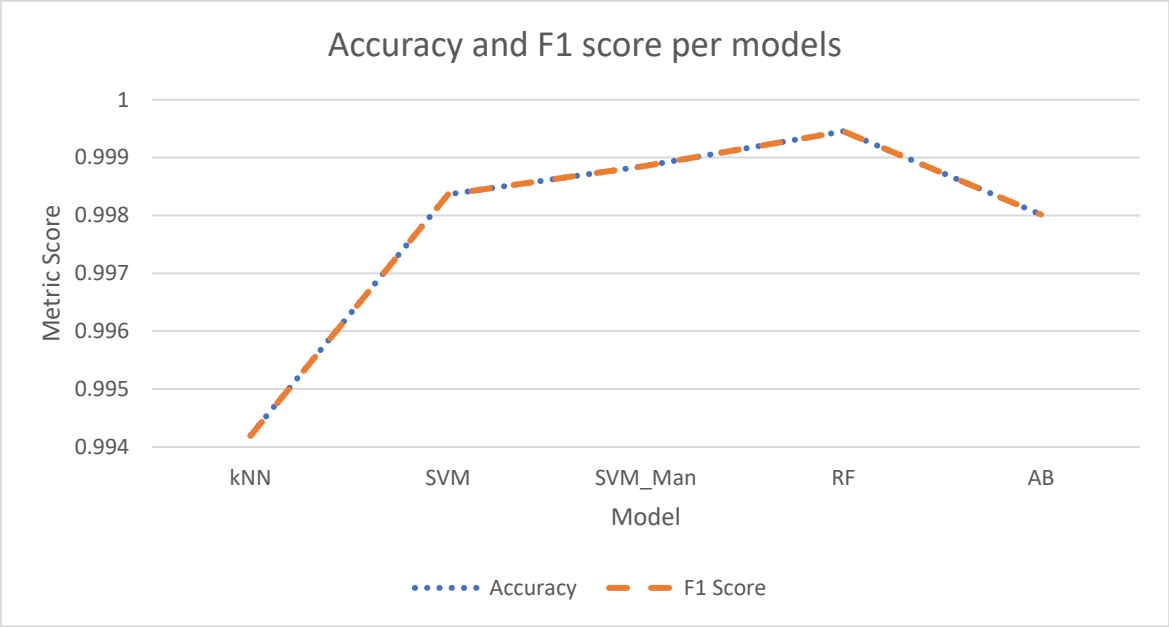


**Figure 4.20:** a) Line graph illustration of average Accuracy and F1 scores across the selected ML algorithms.

Confusion matrix was used as a second performance matrix was used, it displays the false and the true predictions and this is explained in the methodology section above the results are shown in Fig 4.21 below.

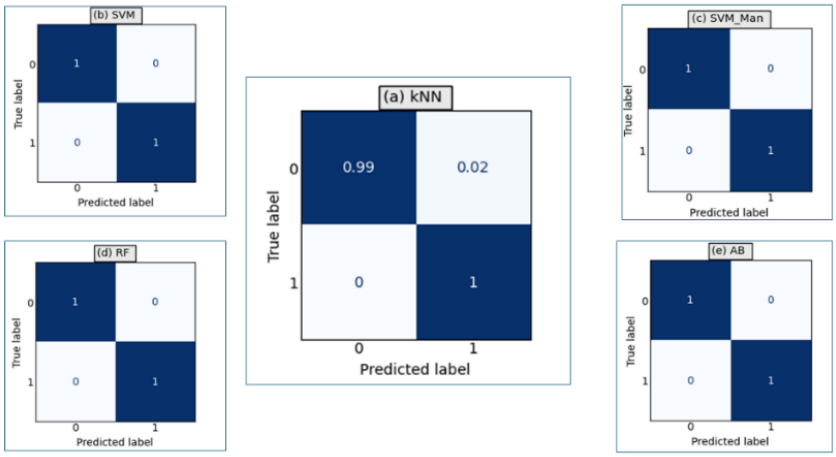


**Figure 4.21***:* ML confusion matrix from the ML algorithms showing predicted vs. true labels a) kNN confusion matrix results b) SVM confusion matrix results c) SVM manually tuned results d) RF confusion matrix results e) AB confusion matrix results.

## 4.7 Cell-based Declustering

Cell based declustering was used to reduce sampling bias by distributing weights according to the sample distribution (highly clustered samples are given less weights while sparsely spaced samples are given more weights), cell size selection is done per domain for each target variable and is shown in (Fig 4.22, Fig 4.23, Fig 4.24, Fig 4.25, Fig 4.26 and Fig 4.27).

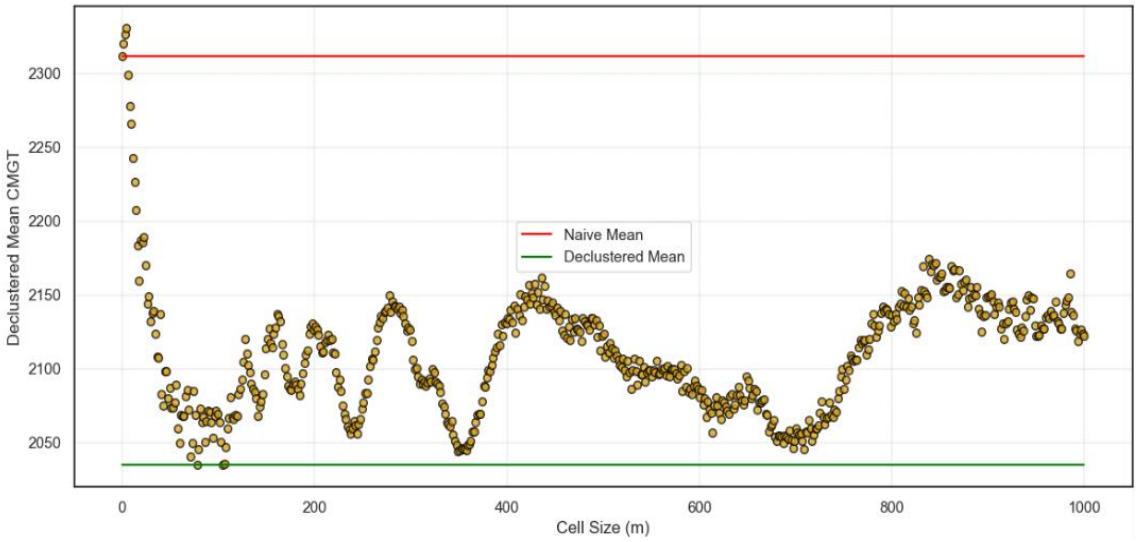


**Figure 4.22:** Cell size selection for Domain 0 Au (cm.g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

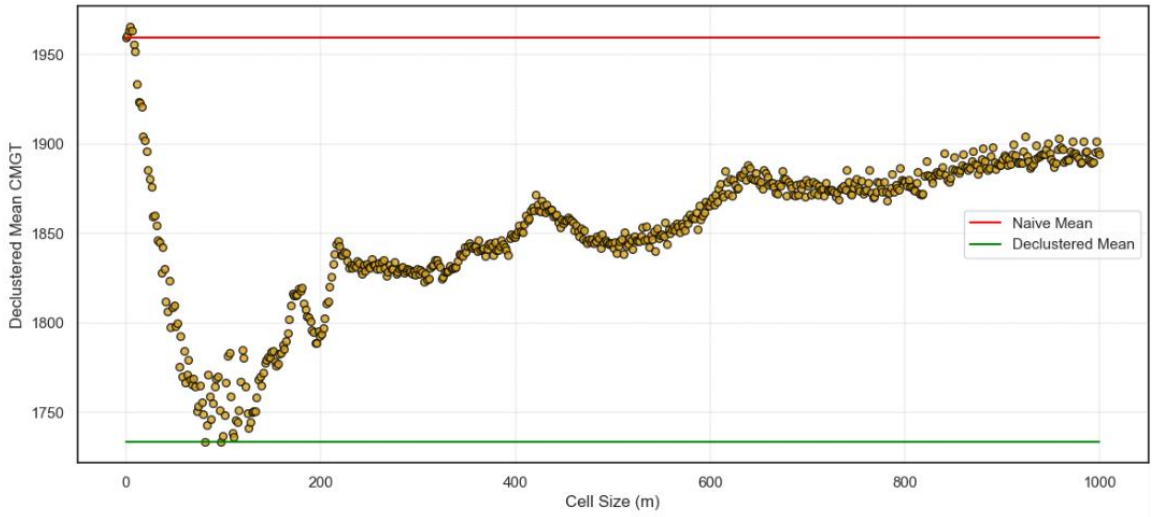


**Figure 4.23***:* Cell size selection for Domain 1 Au (cm.g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

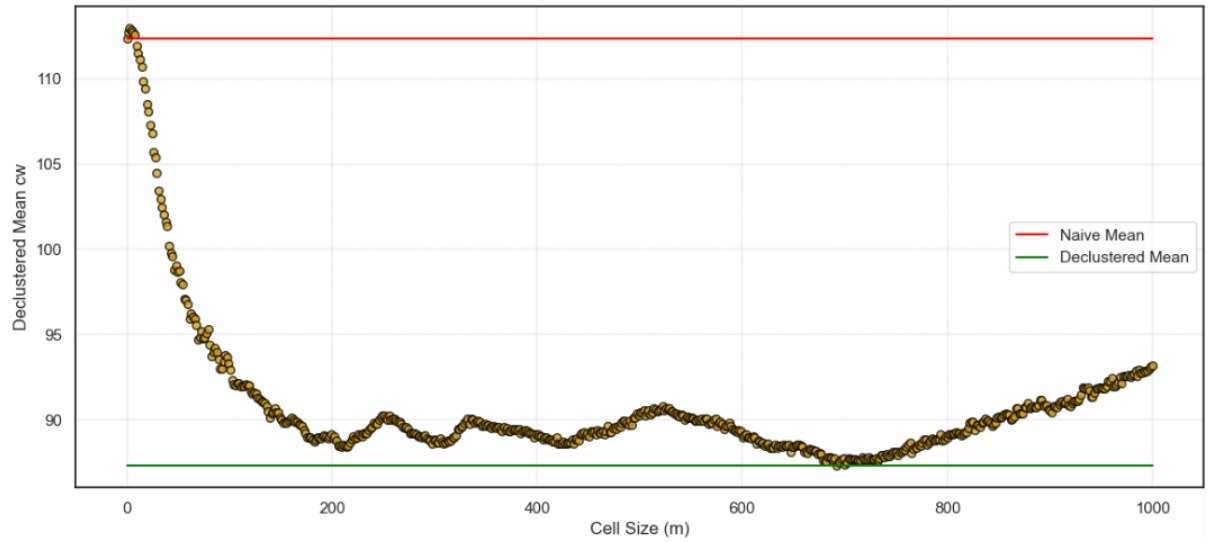


**Figure 4.24:** Cell size selection for Domain 0 CW (cm) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

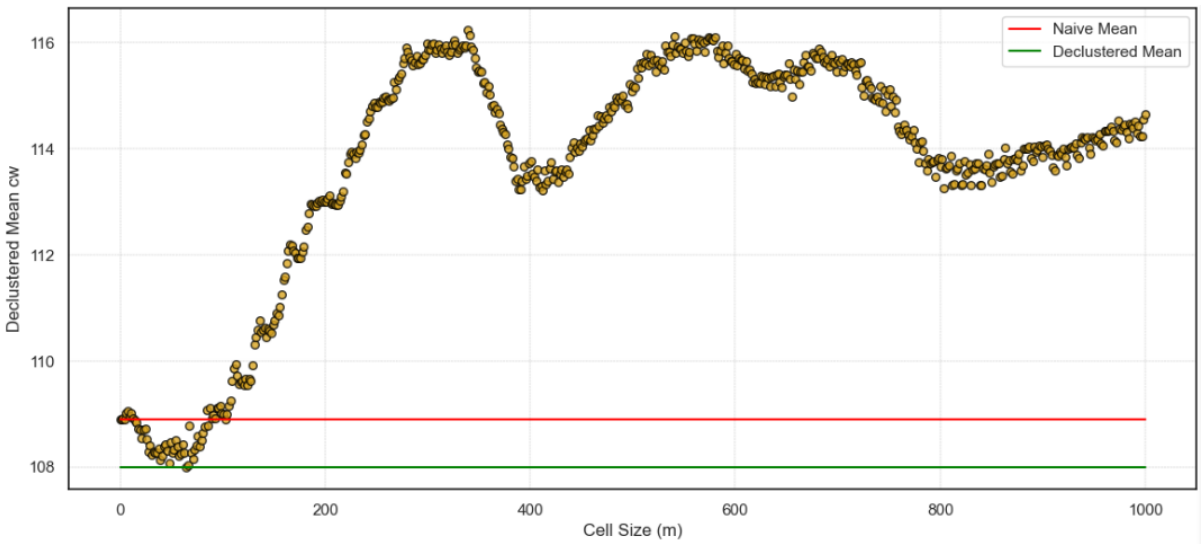


**Figure 4.25:** Cell size selection for Domain 1 CW (cm) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

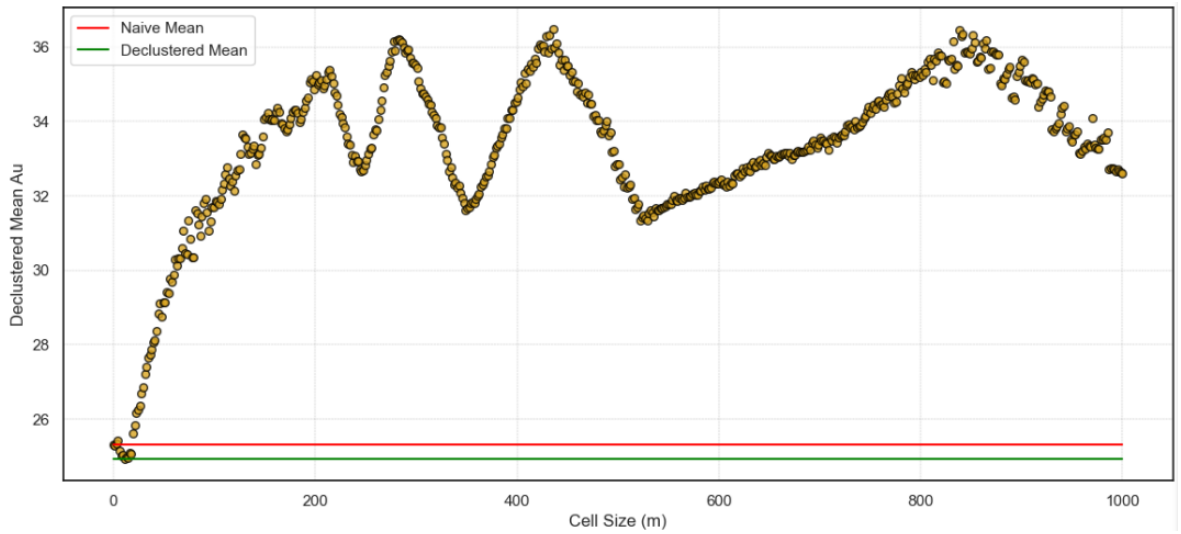


**Figure 4.26:** Cell size selection for Domain 0 Au (g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

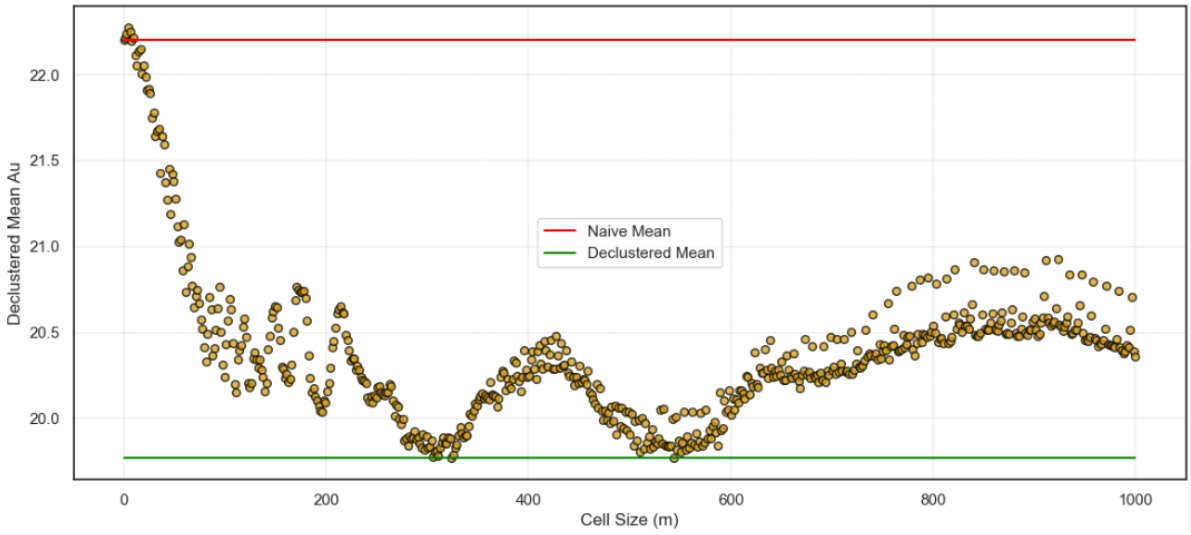


**Figure 4.27:** Cell size selection for Domain 1 Au (g/t) showing the minimum and maximum naïve and declustered mean corresponding to a cell size.

## 4.8 Normal Score Transformation

Normal score transformation was used to transform the declustered data into a gaussian distribution. Gold grades are inherently positively skewed because of their nugget effect and the selective sampling in mining operations influences this distribution. One of the reasons that leads to skewness and heteroscedastic is the proportional effect where there is high variability in the high-grade sampled areas compared to the low-grade sample areas. The normal score method transforms the data by constraining the distribution to a mean of 0 and

a standard deviation of 1 ($\mu$=0 and $\sigma$=1) shown in (Fig 4.28b, Fig 4.29b, Fig 4.30b, Fig 4.31b, Fig 4.32b and Fig 4.33b).

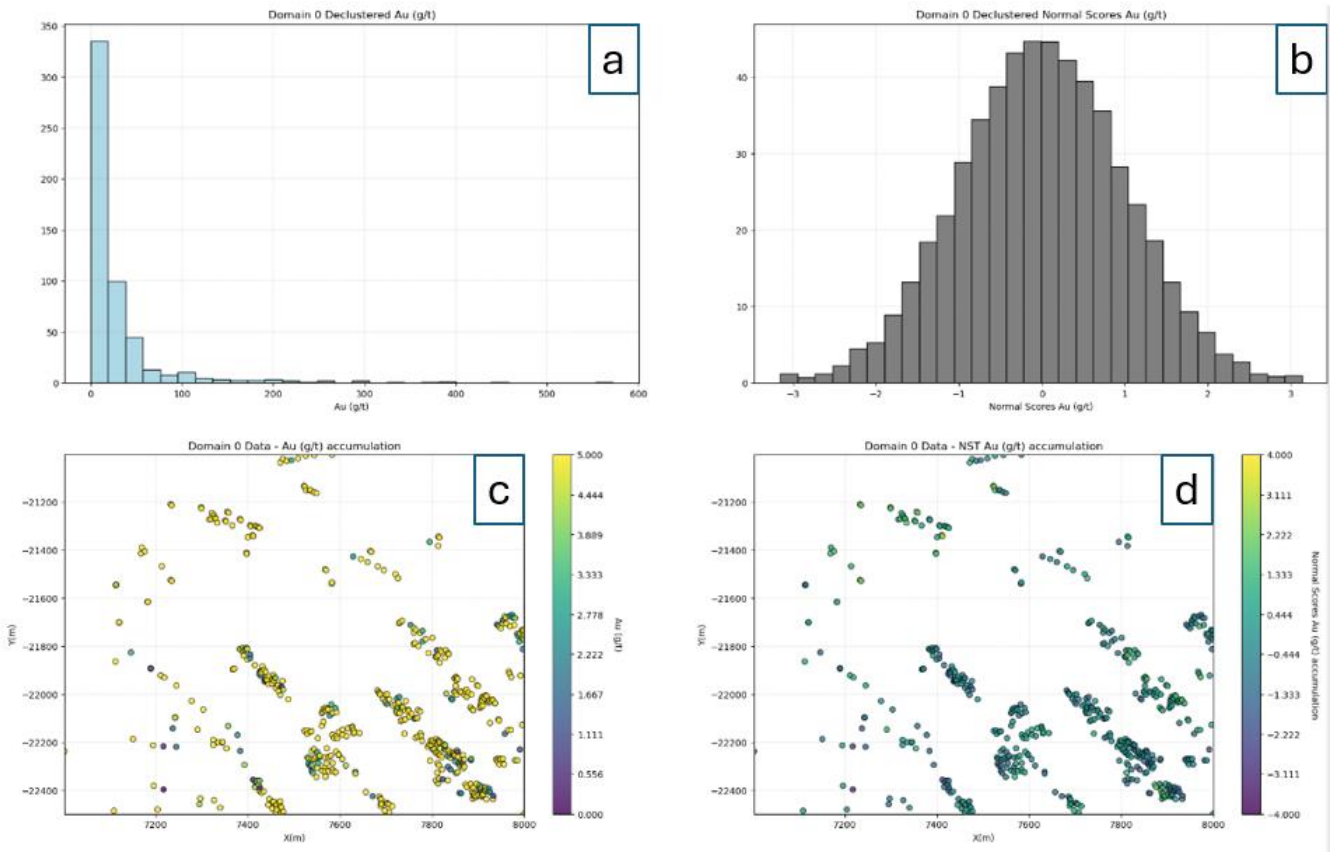


**Figure 4.28***:* Domain 0 declustered Au (g/t) b) Domain 0 declustered normal score Au (g/t) c) Domain 0 declustered Au (g/t) spatial distribution d) Domain 0 normal score Au (g/t) spatial distribution.

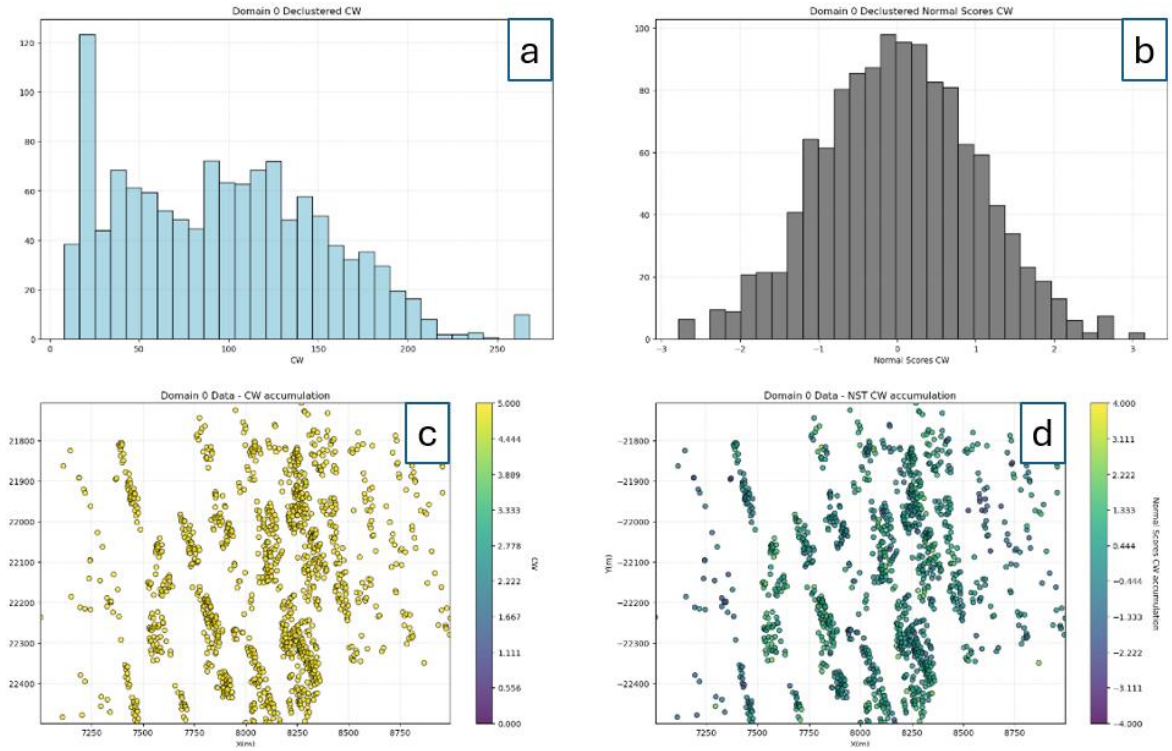


**Figure 4.29:** Domain 0 declustered CW (cm) b) Domain 0 declustered normal score CW (cm) c) Domain 0 declustered CW (cm) spatial distribution d) Domain 0 normal score CW (cm) spatial distribution.

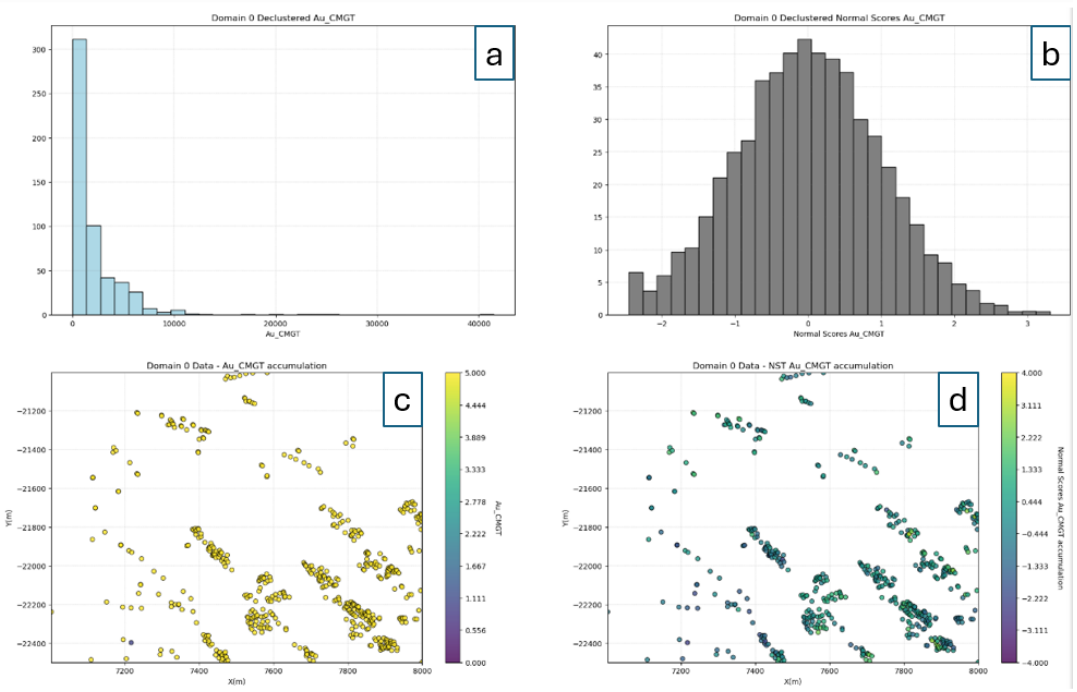


**Figure 4.30:** Domain 0 declustered Au (cm.g/t) b) Domain 0 declustered normal score Au (cm.g/t) c) Domain 0 declustered Au (cm.g/t) spatial distribution d) Domain 0 normal score Au (cm.g/t) spatial distribution.

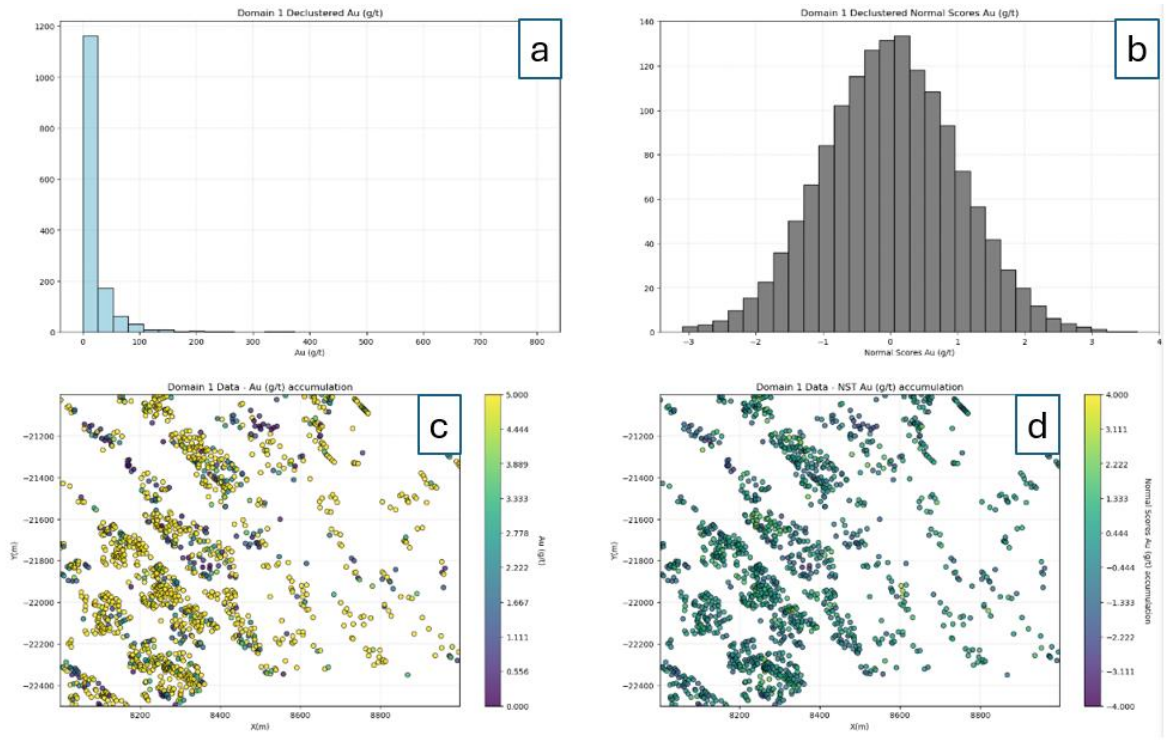


**Figure 4.31:** a) Domain 1 declustered Au (g/t) b) Domain 1 declustered normal score A (g/t) c) Domain 1 declustered Au (g/t) spatial distribution d) Domain 1 normal score Au (g/t) spatial distribution.

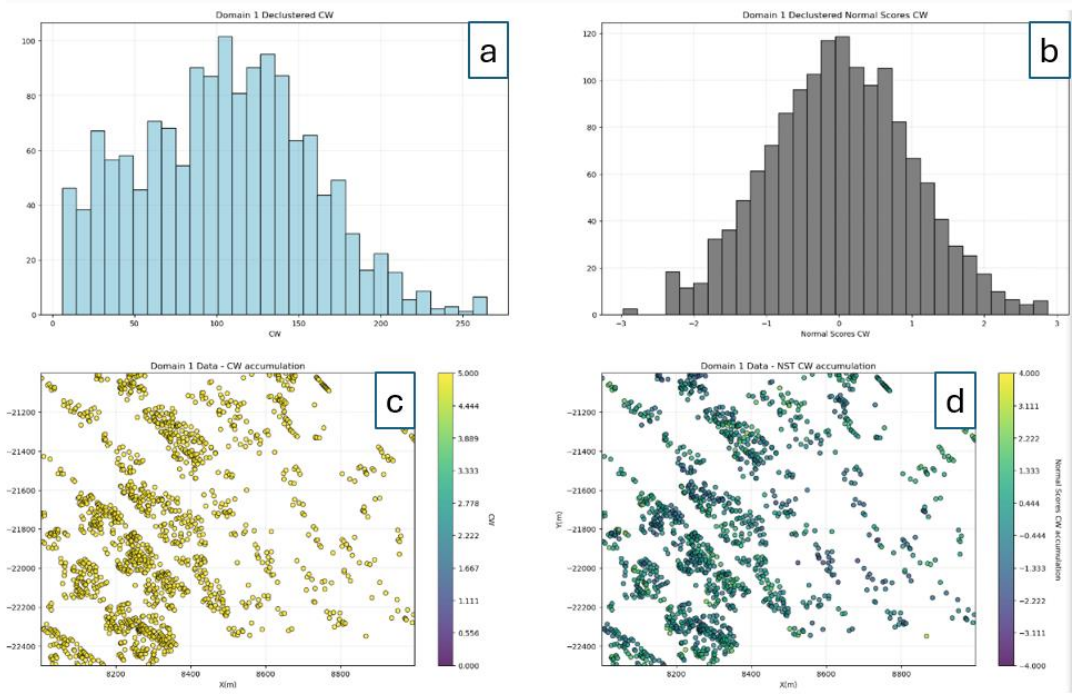


**Figure 4.32: :** Domain 1 declustered CW (cm) b) Domain 1 declustered normal score CW (cm) c) Domain 1 declustered CW (cm) spatial distribution d) Domain 1 normal score CW (cm) spatial distribution.

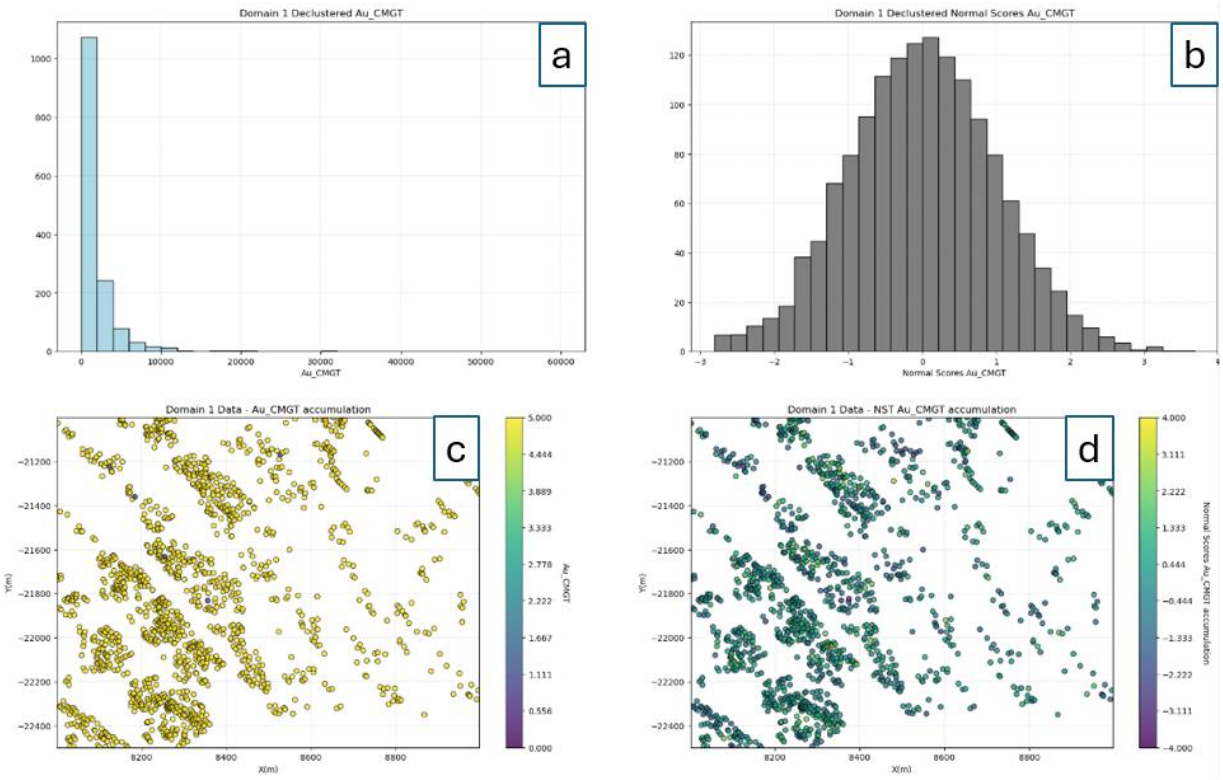


**Figure 4.33:** Domain 1 declustered Au (cm.g/t) b) Domain 1 declustered normal score Au (cm.g/t) c) Domain 1 declustered Au (cm.g/t) spatial distribution d) Domain 1 normal score Au (cm.g/t) spatial distribution.

### 4.8.1 Post domaining statistical evaluation

Upon completion of the Z-score and N-score transformations histograms paired with Q-Q plots are used to compare the data alignment along the 45-degree line. The histogram and the Q-Q plots for the data are presented below in (Fig 4.34b,d, Fig 4.35b,d, Fig 4.36b,d, Fig 4.37b,d, Fig 4.38b,d and Fig 4.39b,d) that shows the data distribution before and after applying the normal score transformation.

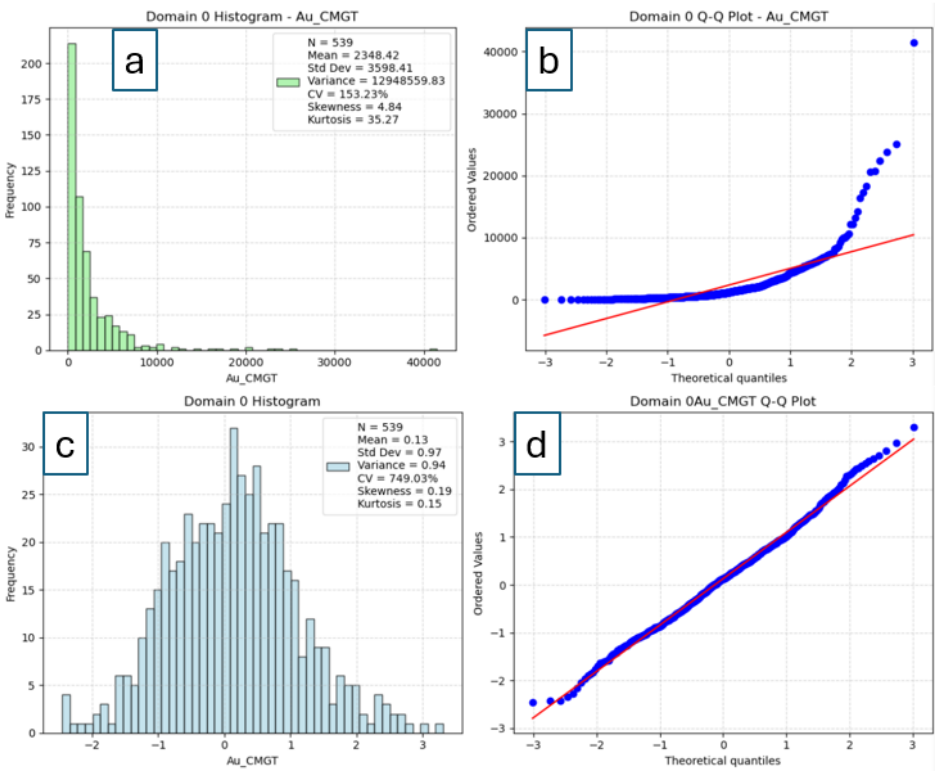


**Figure 4.34:** a) Domain 0 Au (cm.g/t) histogram b) Domain 0 Q-Q Au (cm.g/t) plot c) Au (cm.g/t) treated data histogram d) Au (cm.g/t) treated data Q-Q plot.

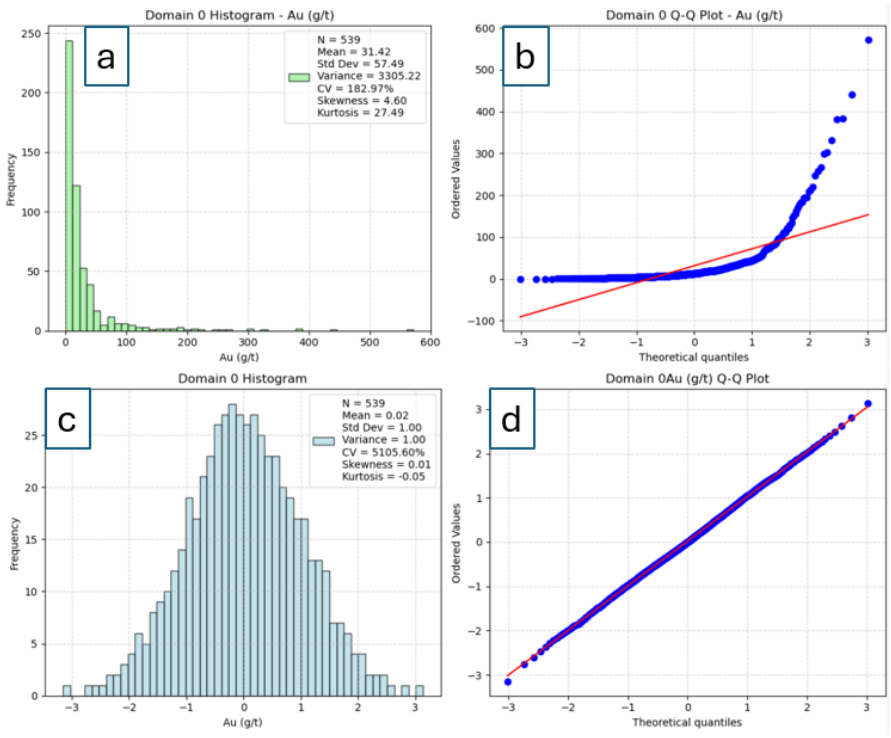


**Figure 4.35:** Domain 0 Au (g/t) histogram b) Domain 0 Q-Q Au (g/t) plot c) Domain Au (g/t) treated data histogram d) Domain Au (g/t) treated data Q-Q plot.

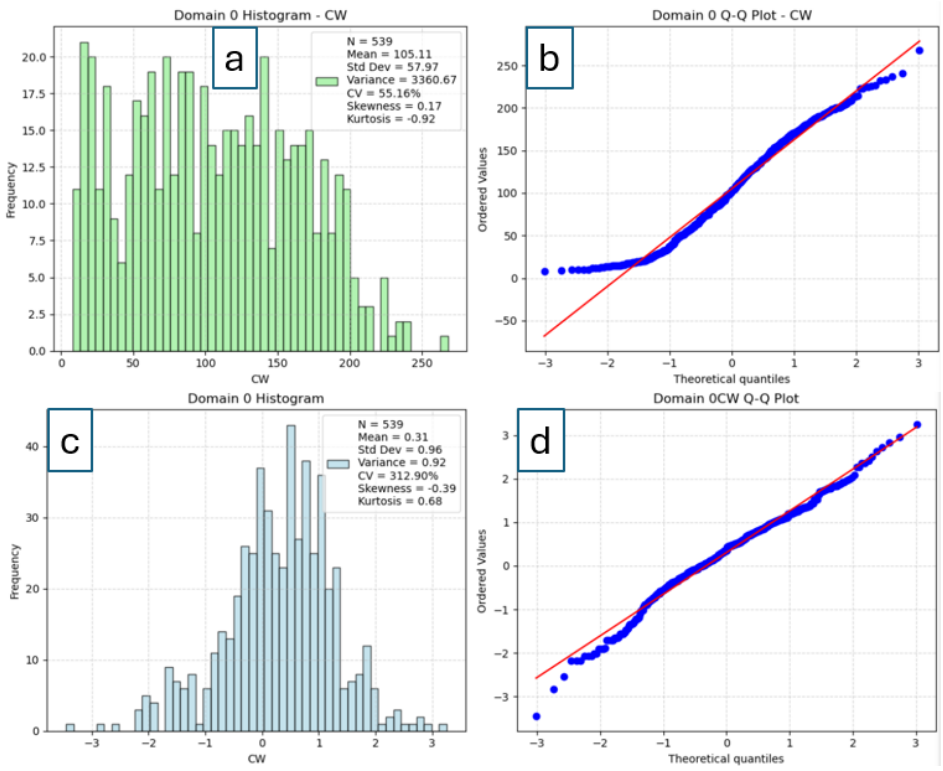


**Figure 4.36:** Domain 0 CW (cm) histogram b) Domain 0 Q-Q CW (cm) plot c) Domain 0 CW (cm) treated data histogram d) Domain 0 CW (cm) treated data Q-Q plot.

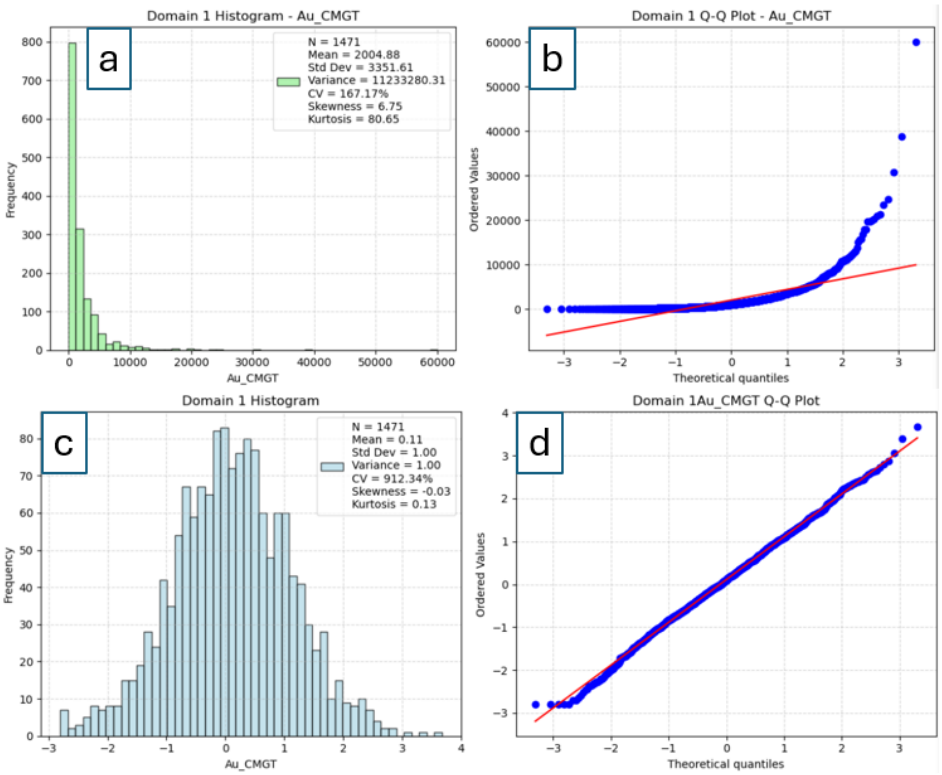


**Figure 4.37:** Domain 1 Au (cm.g/t) histogram b) Domain 1 Q-Q Au (cm.g/t) plot c) Domain 1 Au (cm.g/t) treated data histogram d) Domain 1 Au (cm.g/t) treated data Q-Q plot.

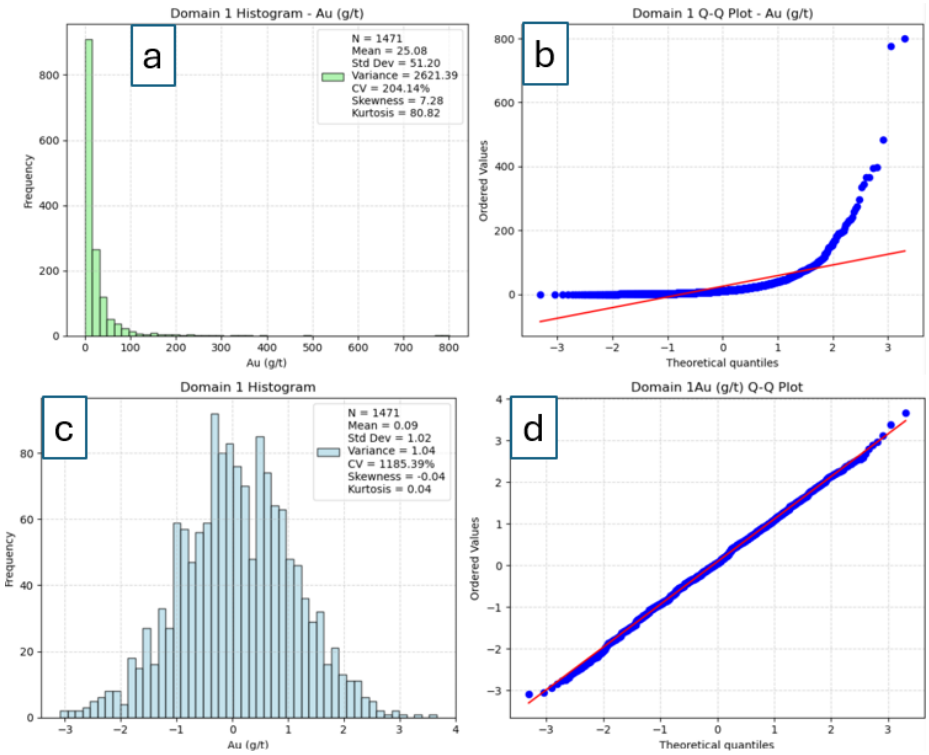


**Figure 4.38:** Domain 1 Au (g/t) histogram b) Domain 1 Q-Q Au (g/t) plot c) Domain 1 Au (g/t) treated data histogram d) Domain 1 Au (g/t) treated data Q-Q plot.

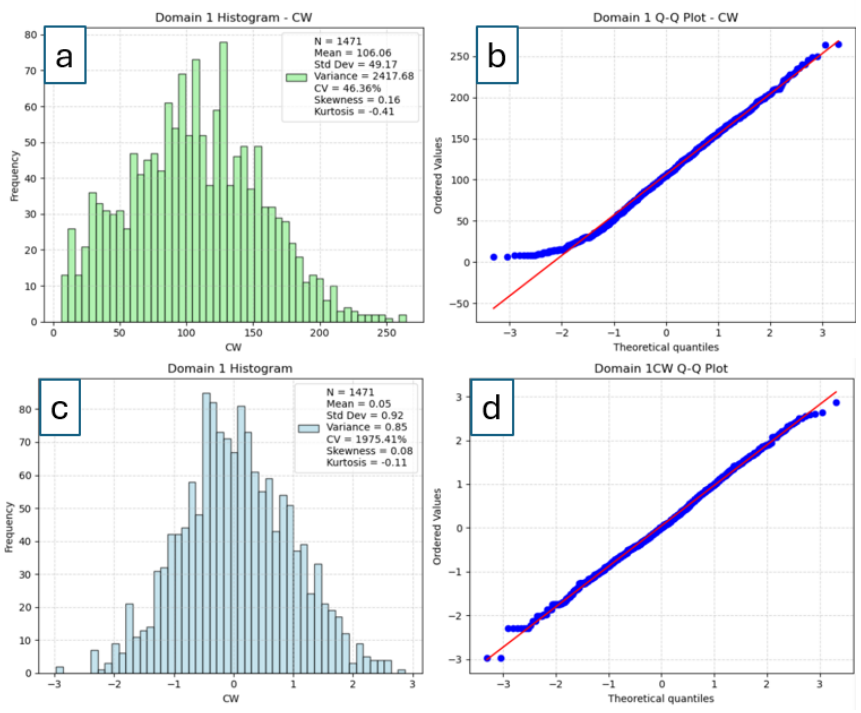


**Figure 4.39:** Domain 1 CW (cm) histogram b) Domain 1 Q-Q CW (cm) plot c) Domain 1 CW (cm) treated data histogram d) Domain 1 CW (cm) treated data Q-Q plot.

## 4.9 Variography Analysis

Variograms were used to model spatial correlations and determine the long and short axis for spatial continuity. Experimental variograms were calculated to determine the variogram parameters, a) sill, b) nugget c) range and d) lag distance.

### 4.9.1 Fitted variogram models

The variogram parameter selection was done using fitted variograms from selected experimental variograms and the major and minor axis were set at the longest and shortest ranges per variable. The variogram model used was set at exponential type characterised by a rapid increase and flattening at the sill which is evident in the fitting variogram models below. These fitted variograms are shown per domain using the three target variables and are shown in (Fig 4.40 and Fig 4.41) below.

(a) Domain 0 fitted variogram models

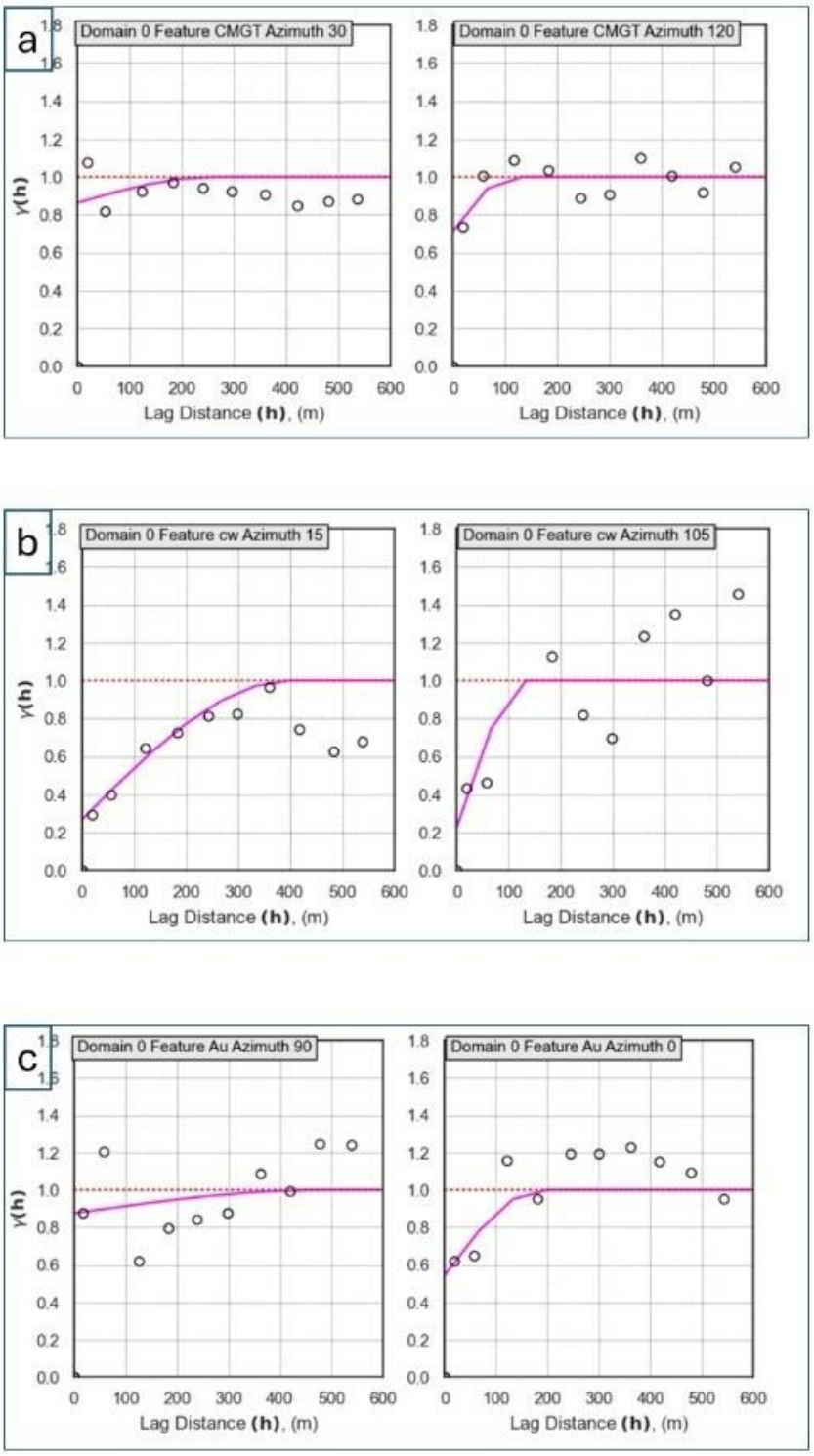


**Figure 4.40***:* Fitted Variogram Model Domain 0 a) Domain 0 Au (cm.g/t) major angle 30 and minor angle 120 b) Domain 0 CW (cm) major angle 15 and minor angle 105 C) Domain 0 Au (g/t) major angle 90 and minor angle 0.

(b) Domain 1 fitted variogram models

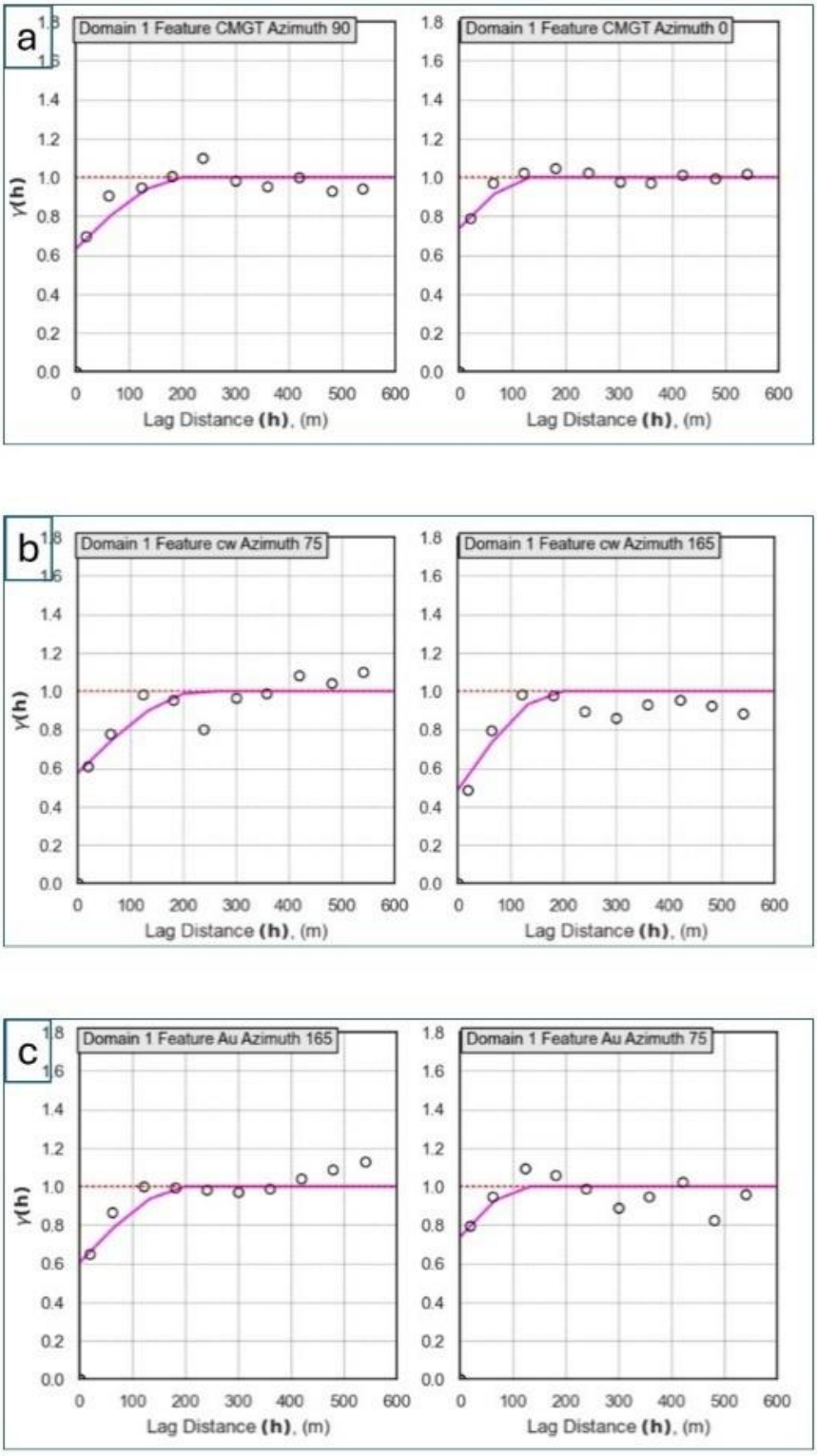


**Figure 4.41:** Fitted variogram model Domain 1 a) Au (cm.g/t) major angle 90 and minor angle 0 b) CW (cm) major angle 75 and minor angle 165 C) Domain 1 Au (g/t) major angle 165 and minor angle 75.

### 4.9.2 Variogram parameters

The variogram parameter to be used in the estimation process are selected from the fitted variograms above and they are show in Table 4.10 below and the parameters are:

- azi1: Azimuth of first structure;
- azi2: Azimuth of second structure;
- cc1: Contribution of first structure;
- cc2: Contribution of second structure;
- hmaj1: Maximum range of first structure;
- hmin1: Minimum range of first structure;
- hmaj2: Maximum range of second structure;
- hmin2: Minimum range of second structure;
- it1: Type of first structure;
- it2: Type of second structure;
- nst: Number of nested structures; and
- nug: Nugget.

**Table 4.11:** Variogram parameters per domain for each variable of interest.

| Parameters | Domain 1 Feature CMGT major | Domain 1 Feature CMGT minor | Domain 1 Feature cw major | Domain 1 Feature cw minor | Domain 1 Feature Au major | Domain 1 Feature Au minor | Domain 0 Feature CMGT major | Domain 0 Feature CMGT minor | Domain 0 Feature cw major | Domain 0 Feature cw minor | Domain 0 Feature Au major | Domain 0 Feature Au minor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| azi1 | 90.0 | 0.0 | 75.0 | 165.0 | 165.0 | 75.0 | 30.0 | 120.0 | 15.0 | 105.0 | 90.0 | 0.0 |
| azi2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| cc1 | 0.37 | 0.26 | 0.4 | 0.51 | 0.39 | 0.26 | 0.1 | 0.29 | 0.73 | 0.78 | 0.12 | 0.46 |
| cc2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| hmaj1 | 205.4 | 0.0 | 234.5 | 204.6 | 0.0 | 0.0 | 258.6 | 0.0 | 398.9 | 0.0 | 475.7 | 0.0 |
| hmaj2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| hmin1 | 135.4 | 0.0 | 194.4 | 0.0 | 0.0 | 0.0 | 111.6 | 0.0 | 136.1 | 0.0 | 184.8 | 0.0 |
| hmin2 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| it1 | 1.0 | 1.0. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| it2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| nug | 0.63 | 0.74 | 0.57 | 0.49 | 0.60 | 0.74 | 0.86 | 0.71 | 0.27 | 0.23 | 0.88 | 0.54 |

### Point Sample Weights

Datapoints were given weights after the declustering task. These weights were distributed according to the data densities and data configuration. Each data point has a weight value for a corresponding feature and weights are unique to that data point.

## 4.10 Block Discretization

To determine the optimal discretisation points per block, the block variance is calculated and plotted against the block discretisation points. The block discretisation block number that corresponds to the graph's inflection point is taken as the optimal number of samples per block.

### 4.10.1 Domain 0 discretised blocks

The discretisation of domain 0 resulted in 5 being selected as the optimal number of points for all the variables of interest. The inflection points for Au (cm.g/t) is 0.87 (Fig 4.42a) corresponds to 5 points per block. Au (g/t) has its inflection point of the block variance curve at 0.88 which corresponds to 5 points as well (Fig 4.42b). CW (cm) has an inflection just above 0.29 sample variance within a block, corresponding to a discretisation of 5 points (Fig 4.42c).

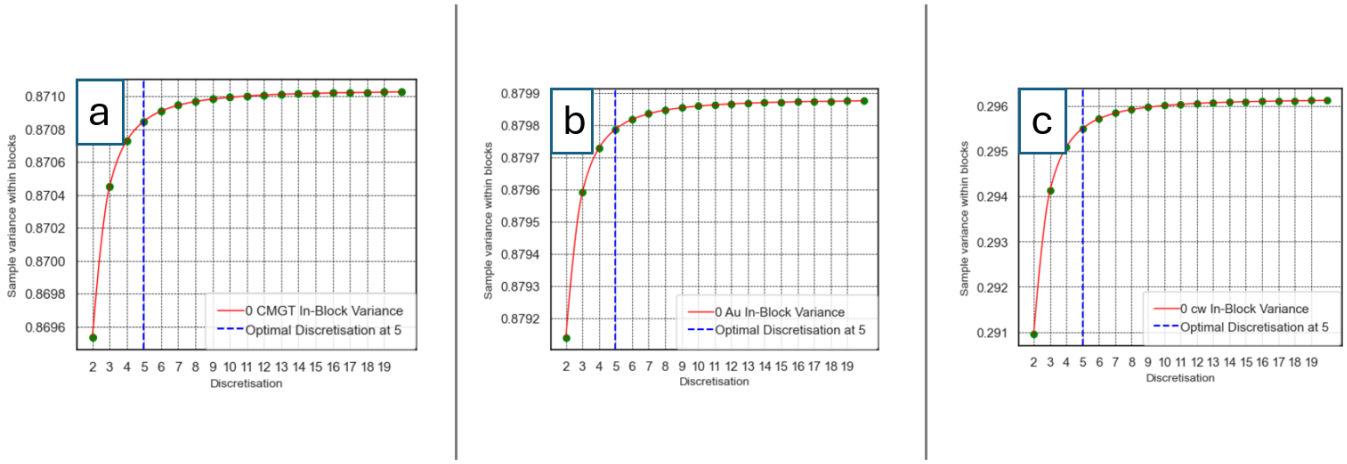


**Figure 4.42:** Block discretisation a) Domain 0 Au (cm.g/t) b) Domain 0 Au (g/t) c) Domain 0 CW (cm).

### 4.10.2 Domain 1 discretised blocks

The discretization of points for domain 1 resulted in 5 being selected as the optimal discretised points for all the variables of interest.). Au (cm.g/t) inflection point is at 0.66 (Fig 4.43a) which corresponds to 5 discretisation points. Au (g/t) has its inflection point of the block variance

curve at 0.63 which corresponds to 5 discretisation points as well (Fig 4.43b). CW (cm) has an inflection point just above 0.60 sample variance within a block which corresponds to a discretisation of 5 points per block (Fig 4.43c).

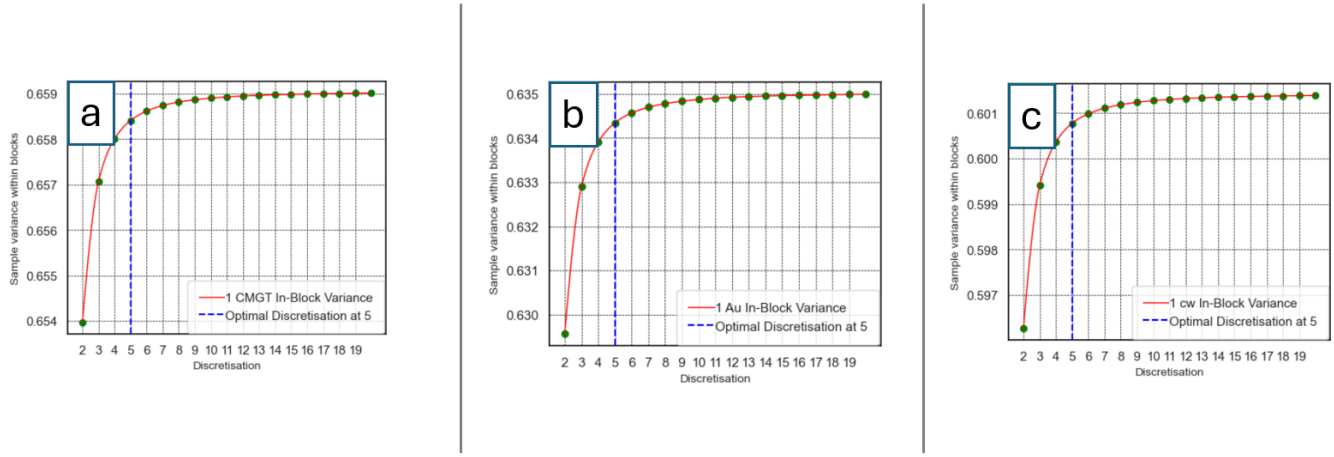


**Figure 4.43:** Block discretisation results per domain a) Domain 1 Au (cm.g/t) b) Domain 1 Au (g/t) c) Domain 1 CW (cm).

## 4.11 Minimum and maximum number of samples

The selected minimum and maximum number of samples for Au (cm.g/t) are 13 and 22 respectively (Fig 4.44a). The minimum and maximum number of samples for Au (g/t) are 12 and 21 respectively (Fig 4.44b). The minimum and maximum number of samples for CW (cm) are 21 and 28 respectively (Fig 4.44c).

### 4.11.1 Domain 0 minimum and maximum samples

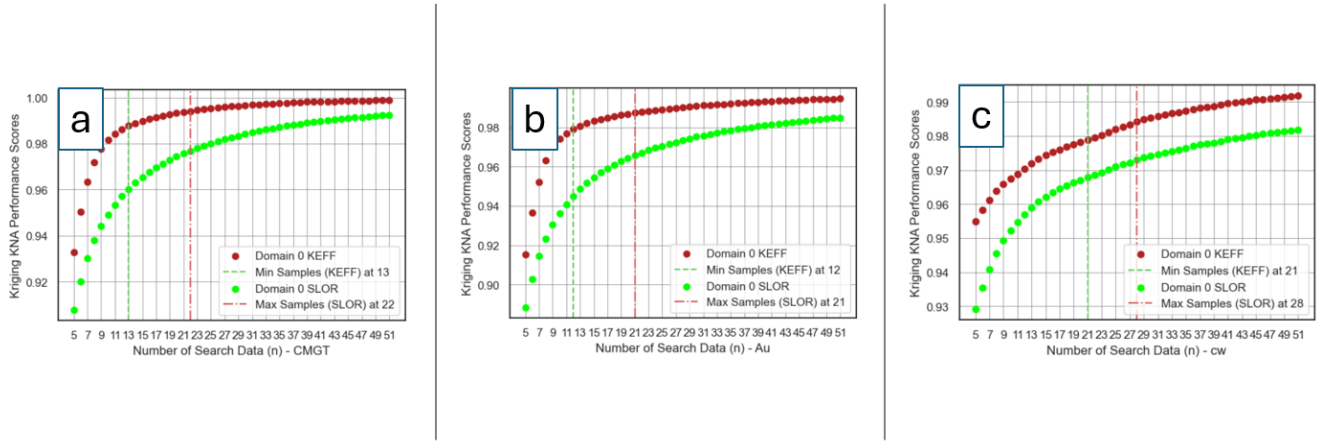


**Figure 4.44:** Quantitative Kriging neighbourhood a) Domain 0 Au (cm.g/t) minimum and maximum number of samples b) Domain 0 Au (g/t) minimum and maximum number of samples c) Domain 0 CW (cm) minimum and maximum number of samples.

### 4.11.2 Domain 1 minimum and maximum samples

The selected minimum and maximum number of samples for Au (cm.g/t) are 15 and 23 respectively (Fig 4.45a). The minimum and maximum number of samples for Au (g/t) are 15 and 23 respectively (Fig 4.45b). The minimum and maximum number of samples for CW (cm) are 15 and 21 respectively (Fig 4.45c).

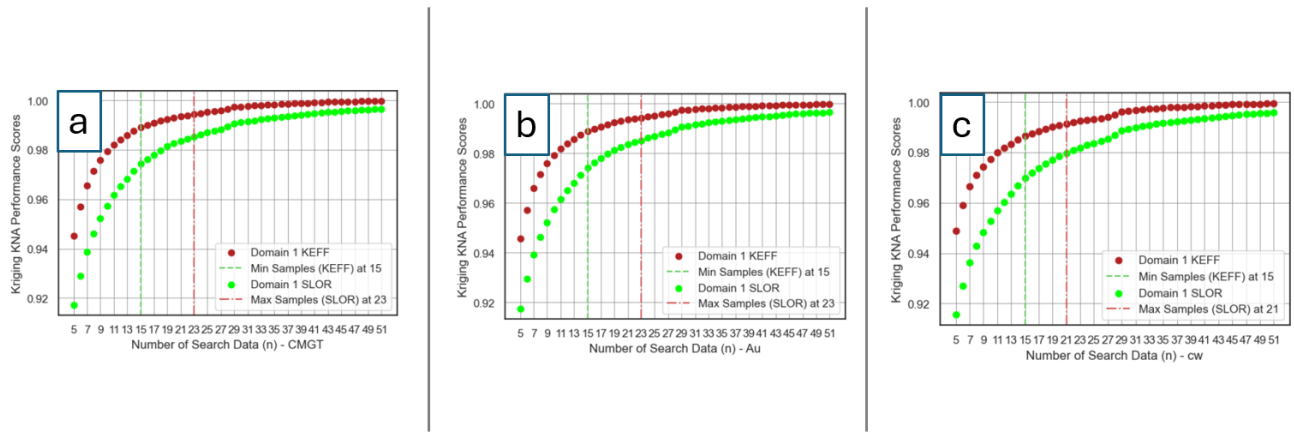


**Figure 4.45:** QKNA number of sample search selection a) Domain 1 Au (cm.g/t) minimum and maximum number of samples b) Domain 1 Au (g/t) minimum and maximum number of samples c) Domain 1 CW (cm) minimum and maximum number of samples.

## 4.12 Kriging Block Models

A 2D Kriging block model was generated in Datamine and using data from the standard Kriging files selection. These files are a) Estimation parameter (Epar) b) Search parameter (Spar) and d) Variogram parameter (Vpar file) e) Empty protom block model. These files contained all the variogram data, estimation model data and the QKNA results and the empty shell model (protom) where the predictions will be done. The 2D models show the spatial distribution of estimated values in unsampled locations and are shown in Fig 4.46 for Au (cm.g/t) and Fig 4.47 for CW (cm).

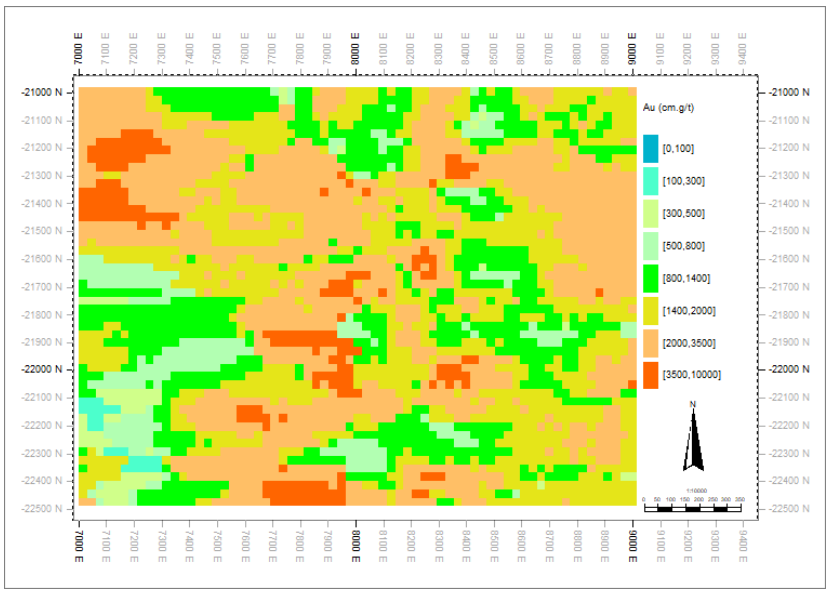


**Figure 4.46:** 2D map showing the spatial distribution of Au (cm.g/t) Kriging results.

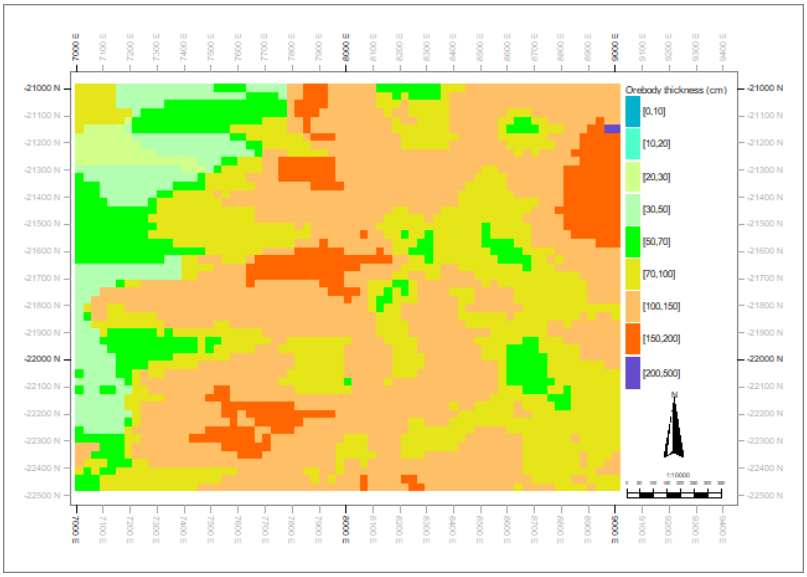


**Figure 4.47:** 2D map showing the spatial distribution of CW (cm) Kriging results.

## 4.13 Post Kriging Analysis

The Kriging estimation performance is validated using the KEFF and the SLOR which are explained in detail in Chapter 2. Fig 4.48 and Fig 4.49 show the spatial distribution of KEFF and SLOR respectively and Fig 4.50. and Fig 4.51 shows the histogram distribution of the performance values.

Kriging Efficiency

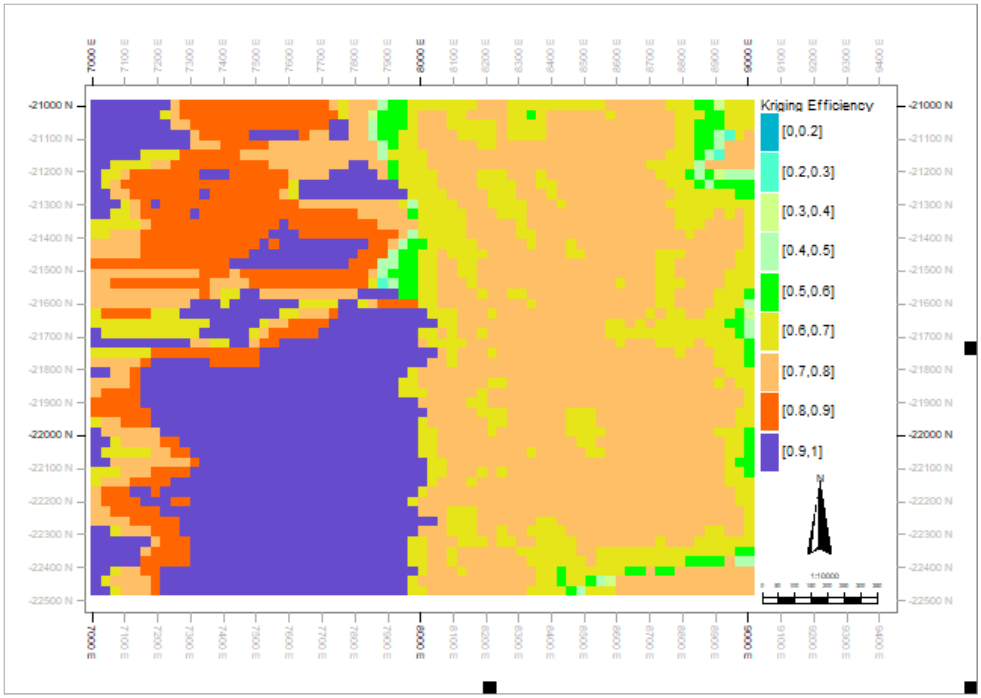


**Figure 4.48:** 2D map showing the spatial distribution of KEFF results.

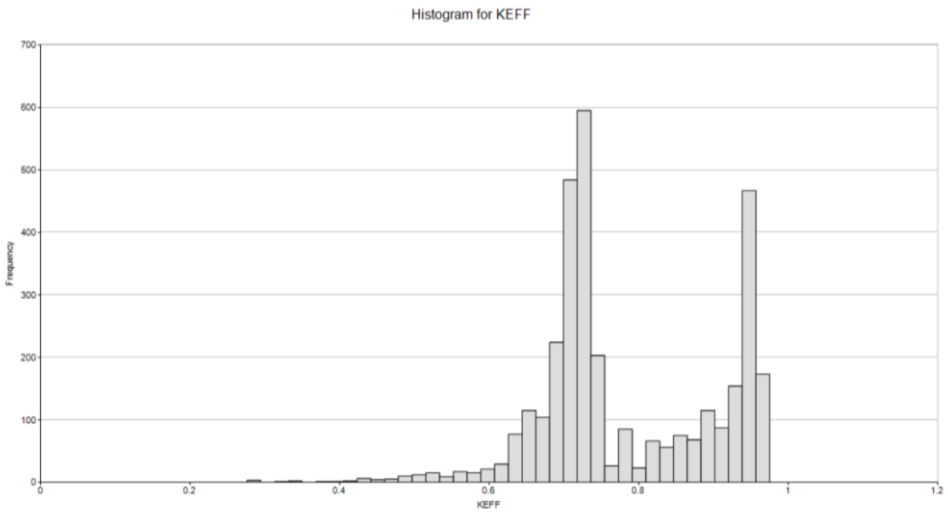


**Figure 4.49:** Histogram showing the KEFF distribution depicting a bimodal distribution.

Slope of Regression

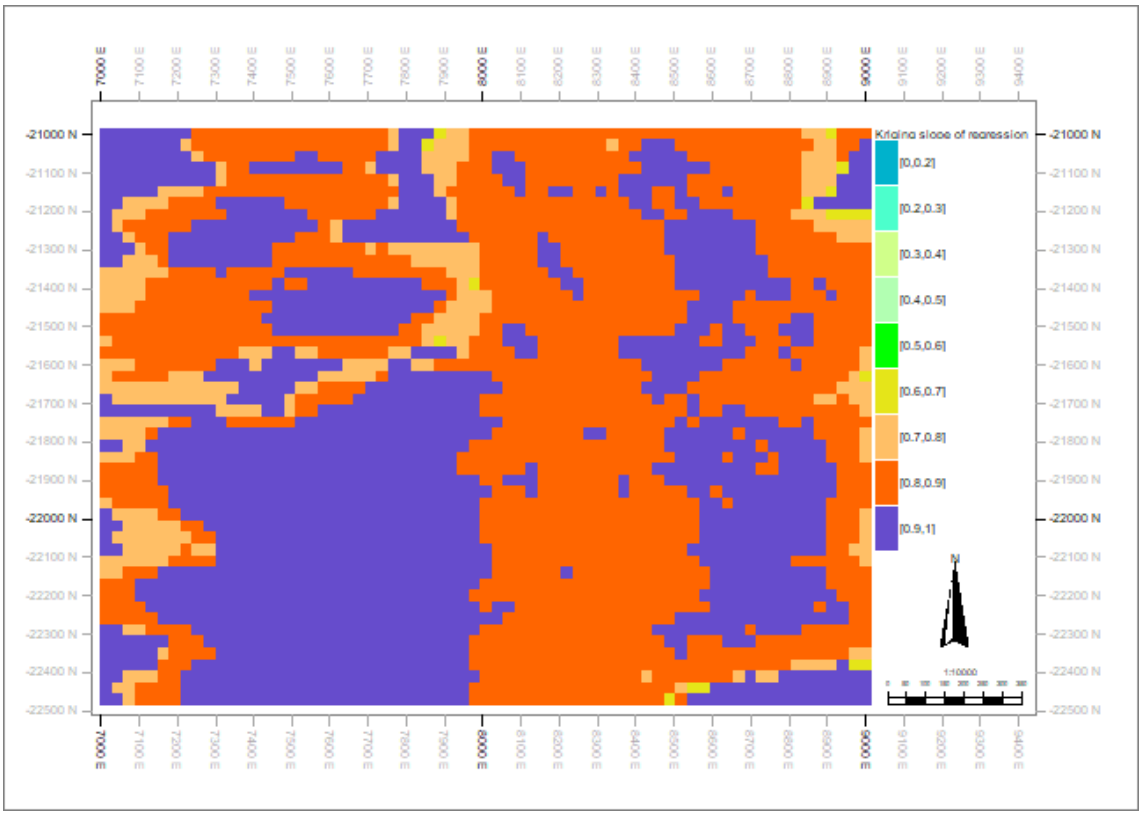


**Figure 4.50:** 2D map showing the spatial distribution of SLOR results.

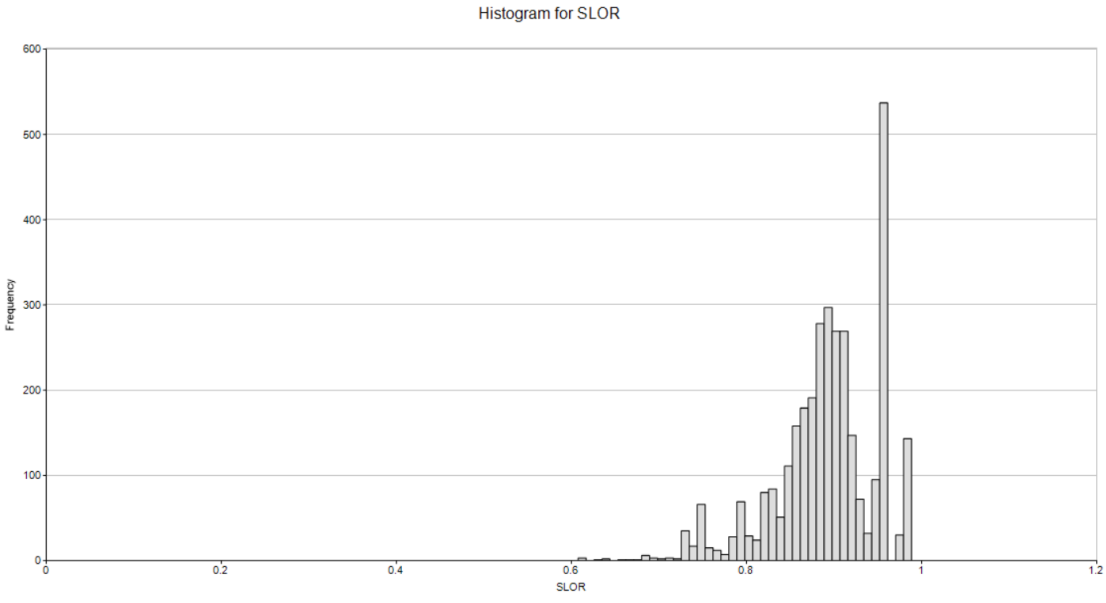


**Figure 4.51:** Histogram showing the SLOR distribution depicting a mixture of values across domains.

Comparative swath plot diagrams between the estimated values and the true values, it also shows the total number of samples used per slice in the X and Y directions. The swath plots indicate the closeness of the prediction to the true values this is shown on Fig 4.52, Fig 4.53 for Au (cm.g/t) and Fig 4.54 and Fig 4.45 CW (cm).

Au (cm.g/t) X-axis swath plot.

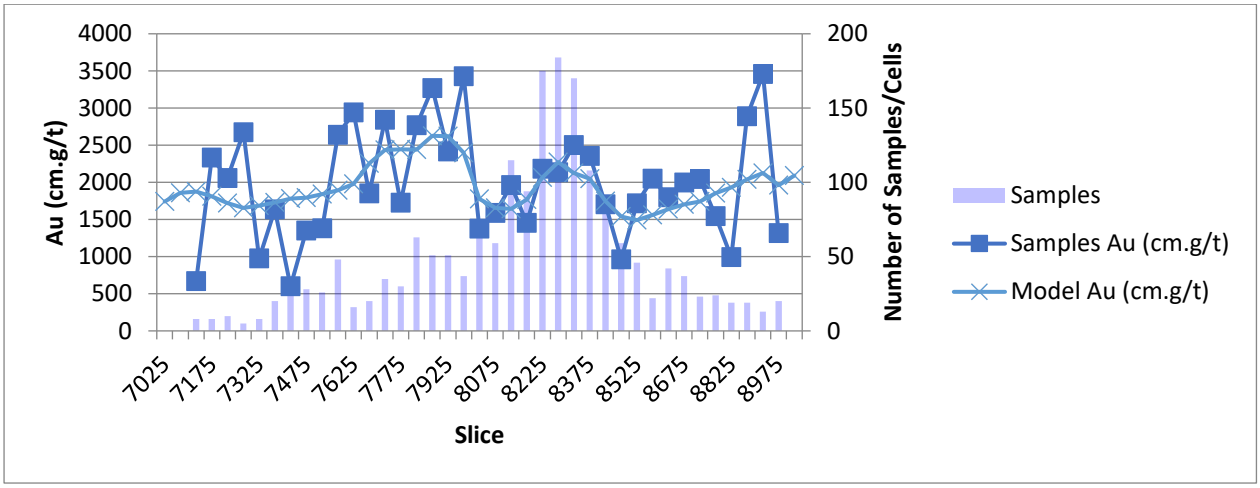


**Figure 4.52:** X-axis swath plot showing the estimated values compared to actual grades along the E-W direction representing number of samples, samples Au (cm.g/t) and model Au (cm.g/t).

Au (cm.g/t) Y-axis swath plot.

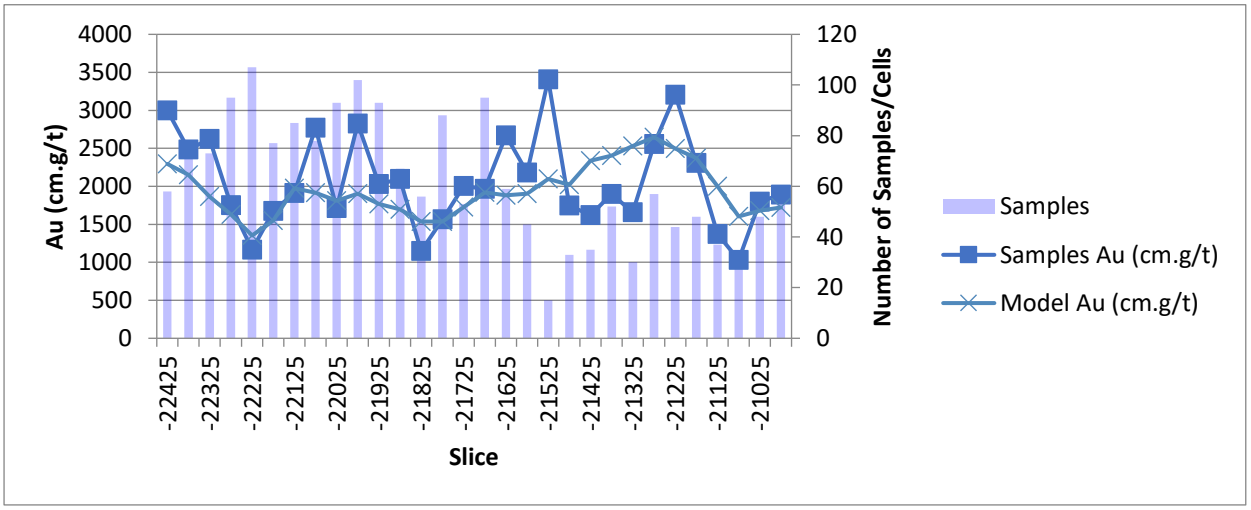


**Figure 4.53**: Y-axis swath plot showing the estimated values compared to actual grades along the N-S direction representing number of samples, samples Au (cm.g/t) and model Au (cm.g/t).

CW (cm) X-axis swath plot.

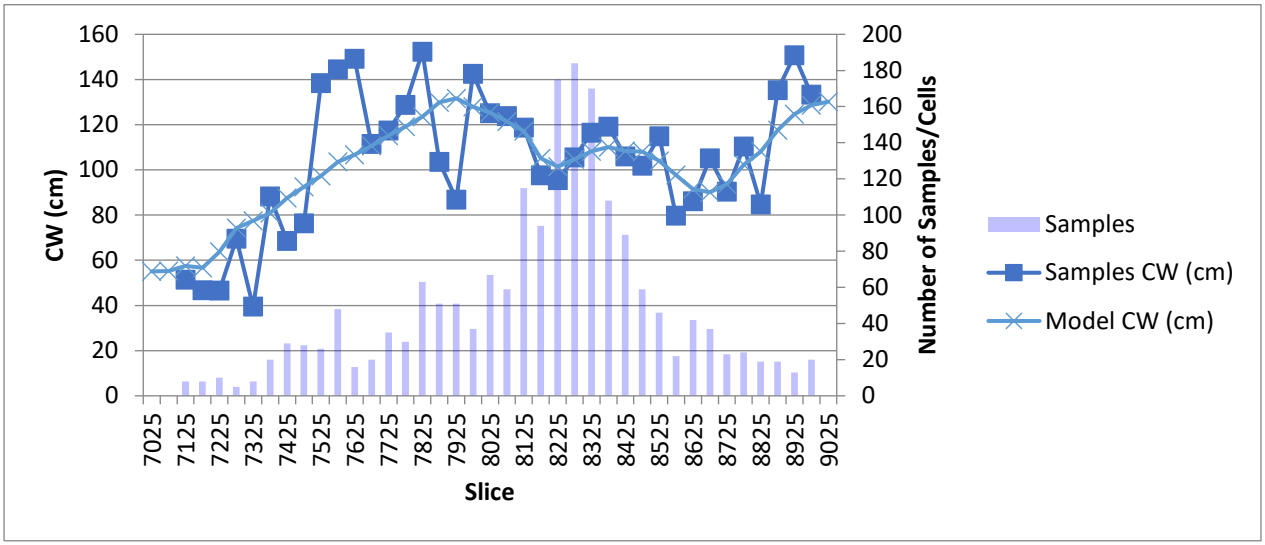


**Figure 4.54:** X-axis swath plot showing the estimated values compared to actual grades along the E-W direction representing number of samples, samples CW (cm) and model values CW (cm).

CW (cm) Y-axis swath plot.

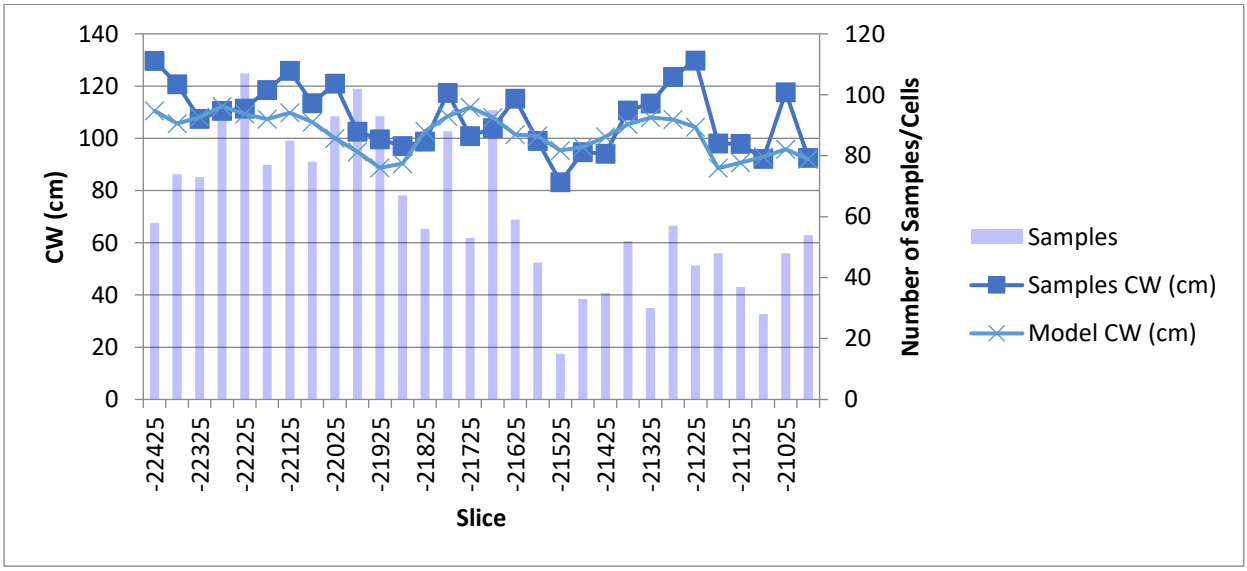


**Figure 4.55:** Y-axis swath plot showing the estimated values compared to actual grades along the N-S direction representing number of samples, samples CW (cm) and model CW (cm).

# 5 DISCUSSION

This chapter of the paper will look at discussing the results from the analysis and give further explanation to the outcome of the domain boundary generation. It will investigate the performance of the two clustering techniques and why signed-distance-based -mds-k-means was selected. Furthermore, the performance of the domain boundary ML algorithms and the reason why kNN algorithm was superior despite the statistical results. The Kriging estimation results, and the post Kriging performance analysis will be discussed and what these mean to the application of this hybrid method in mineral resources estimation of other deposits together with its limitations.

## 5.1 Point Sample Domain Analysis Outcome and Performance Analysis

Point domain analysis was done using the Standard K-means clustering and the signed-distance-mds-k-means clustering techniques. The two methods are described in detail in the methodology section above. The Standard K-means clustering performed better statistically compared to the signed-distance-mds-K-means clustering. It resulted in a higher silhouette score (Fig 4.8a) which shows that the clusters are tight and well separated, it also resulted in a lower Davies-Bouldin score (Fig 4.8b) which indicates that there is lower intra cluster spread and dissimilarity. Even though statistically the Standard K-means clustering performed better when plotting the resultant clusters on a 2D space, it shows cluster mixing (Fig 4.8c) whereas the signed-distance-mds-K-means clustering resulted in a better compact and well represented domains with a clear structural controlled separation that runs N-S. Nwaila et al., (2025) explored an integrated geodata science resource estimation process in the Merensky reef. The traditional k-means clustering resulted in poor cluster separation and the signed-distance-mds-k-means resulted in clearly separated domains and these results are consistent with the results of this research stated above. The incorporation of spatial statistics in ML for geosciences is important because it takes into consideration the spatial relationships of data points, (Ying et al., 2025) demonstrated that a dual supervised learning approach that includes a) spatial component and b) spectral component resulted in superior detection of geochemical anomalies and including the geological knowledge in the ML process improved the interpretability and accuracy of the anomaly identification. Domain data points mixing is a problem in mineral resource estimation that has been studied and highlighted as one of the leading reasons for poor estimation and mineral resource and reserve classifications (Rossi & Deutsch, 2013; Deutsch *et al.,* 2014). The fundamental principle of geostatistical analysis relates to stationarity and continuity, in which domain is homogeneous. Domain sample mixing affects a) plotting of variograms and the selection of variogram parameters b) poor estimation resulting from poor representation if actual values leading to false highs and false lows with

has a materiality impact in mineral resource estimation c) poor mine planning designs with unrealistic recoveries that may impact company cashflow due to poor sales related to the poor recoveries. Atalay, (2025) emphasised the importance of proper domaining before running resource estimations, as the results from the research using XGBoost and Support Vector Regression (SVR) on domained and non-domained data. The results showed that 30% high grade data of the non domained estimation resulted in grade values outside the reasonable range. The signed-distance-mds-k-means resulted domains were selected as the baseline for the domain boundary generation and the subsequent Kriging estimation because of the clear separation. This separation may be attributed to the enhancement of the clustering due to the MDS function that emphasise features and generate clear boundaries. As a result this may increase the intra cluster variance or may, significantly reduce the inter cluster separation which may result in the silhouette and Davies-Bouldin scores looking the way they are in this case. Nwaila and Frimmel, (2023) showed that the signed-distance-mds-k-means resulted in spatially coherent domains and eliminated the bias and influence of generating boundaries. The common practice in developing mineral resource estimation domains in mining involves the use of grade values as indicators for domaining. The cut-off grades are used to determine the boundaries of the domains which is informed by the baseline of the mine. As much as this approach has been applied in many deposits and mining operations it does not take into consideration the underlying geo-spatial relationships and other geological attributes that may provide greater benefit in resource understanding and mining optimisation.

## 5.2 Machine Learning Domain Boundary Generation Analysis.

The use of ML in generating geodomains is promising to provide an alternative to traditional geodomaining methods which have various limitations such as misidentification of domains and long and tedious processes that may result in poor resolution of domains. These methods can also be combined with other methods to generate hybrid models that may optimise the mineral resource estimation process by addressing the shortcomings of both methods in one process. ML for domain delineation have been used in various ore deposit types. (Zhang *et al.*, 2023) explored the use of various classification ML algorithm to generate domain boundaries. The project area is a gold deposit within the same locality as this research and the results show that a combination of qualitative and quantitative analysis is required when working with ML algorithms.

This project's ML approach in resource estimation domaining show that there is a need for assessing model outcomes with a geological understanding as the key deciding factor. It is evident in this research that the post-modelling evaluation statistical results have a different interpretation and implication compared to the visual domain boundaries generated. The RF

algorithm is the best performing with an accuracy score of 0.999 and an F1 score of 0.999 and the KNN was the least performing algorithm with an accuracy score of 0.994 and an F1 score of 0.994 (Table 4.8). Despite the outcome of these performance results, the resultant boundaries show that the kNN algorithm produced a geologically realistic boundary compared to the other algorithms (Fig 4.18a). RF, SVM, manually tuned SVM and the AB algorithms all produced a highly smooth boundary which is not representative of the conditions of the ore-forming processes of the gold-bearing reefs of the Witwatersrand Basin. This would be appropriate for a low-energy depositional setting where the stratigraphic units show a smooth undisturbed contact, and the units are dominated by fine sediments.

The kNN model is the best-performing algorithm qualitatively because it captures the typical geological process associated with the domains of the Carbon Leader Reef. The domain boundaries are normally sharp and distinctive which is easily captured by the logging geologist as shown in the primary data in Table 4.2 under the basal contact description. The transition from a high-energy to a low-energy depositional unit is displayed by the change in pebble sizes, sorting, packing, and carbon content which is also captured in the logging sheet as displayed in Table 4.2. Even though there are notable cross-cutting relationships and structurally induced displacement of geological units evident on site this has little to no impact on the nature of contacts of the reefs. Taking this into cognizance it is important to use geological knowledge and interpretation to analyse ML algorithms application in mineral resource estimation. ML algorithms do not put geological logic into their analysis and are mainly influenced by the nature of the data used. A geologist's input is required to provide high-level quality assurance and quality to ensure that the results from the algorithms make geological sense while putting geological principles and experience into the analytical process.

## 5.3 Kriging Estimation Optimisation Through Quantitative Kriging Neighbourhood Analysis

The determination of optimal parameters for estimation is an important step to ensure accurate estimation of predictions. The selection of the major and minor search ranges influences the size of the ellipse used for running estimation and the subsequent minimum and maximum number of samples required to satisfy the estimation. Domain 1 has longer major and minor search ranges compared to domain 0 as shown in the variogram parameter in Table 4.10. This is because domain 1 has fewer samples (Fig 4.11), and the samples are sparsely located as such more area is required to cover fewer samples and may also be due to the underlying continuity of the sampled area. Domain 1 also has more minimum and maximum number of samples per variable due to the scarcity of the samples within the domain. The number of samples per domain also affects the estimation time per domain, in this case, domain 0 has

1477 points compared to domain 1 with 540 sample points will take longer to estimate. Proper QKNA is required to make a well-informed Kriging estimation and the parameters generated from this process influences the outcome of the values in unsampled areas. Geological knowledge of the ore deposit is important here as it helps in critiquing the outcome in relation to what is already known about the deposit.

## 5.4 Ordinary Kriging Estimation and Block Modelling

Ordinary Kriging estimation was performed and a 2D model for the Au (cm.g/t) and CW (cm) was generated. The model shows the spatial distribution of these two variables in a 30 m by 30 m block dimensions. The selection of block sizes was influenced by the size and the complexity of this mineralisation, and this block size is preferred for most of the Witwatersrand deposits (Nwaila and Frimmel, 2023). The block model represents the geological heterogeneity, grade and volume variability of the mineralisation. The selection of block size affects the accuracy of the estimation model. Selecting large blocks results in the over-smoothing of variables (Khorram *et al.*, 2021). Selecting smaller blocks has a downstream impact as it increases the dilution percentage and it also results in the sudden transition from high to low grades. Jara *et al.,* (2006) further illustrated that the use of smaller block sizes can increase the resolution of the model, but it increases the operational costs due to dense drilling patterns and results in poor productivity. The block model has successfully achieved a good estimation of both variables, and the post-estimation analysis is discussed below.

The model shows distinctive transitions of high to low grades and captures the extent of the variability in space. There is a notable indication of the continuity and preferred orientation of both high and low values within the model. The Au (cm.g/t) higher grades in the NW corner of the study area are consistent with the primary point data (Fig 4.4a) where a cluster of high grades is shown, this coincides with high grades in the same area shown in the 2D model (Fig 4.46). The high-grade values in this area are less dense compared to the lower-grade values in other portions of the study area as such higher weights were allocated to these points compared to the less dense areas when carrying out declustering shown in Table 4.10. These areas should be handled with caution and further interrogation of other relevant data is required to conclusively classify them. Overlapping geological models and sectional drawings may provide insights into the extent of the grade distribution and the constraints of these high-grade values. The CW (cm) values show medium to high estimated values and sporadically occurring low values (Fig 4.47). From this block model, it is evident that there is a direct relationship between the two features, Au (cm.g/t) and the CW (cm). Areas of high grade are related to areas of thicker ore sections this is generally the case in the Witwatersrand basin

because of the abundant occurrence of gold bearing quarzitic pebbles. In many mining operations, targeted mining is carried out in thick reefs to extract higher gold grades.

## 5.5 Post-Kriging Estimation Performance and Block Model Evaluation

The Kriging method aims to reduce estimation variance. This estimation variance indicates how well the estimation is, it quantifies the level of uncertainty associated with the estimation values. One key Kriging performance indicator is the KEFF and is discussed in the methodology section above. It measures the effectiveness and accuracy of the Kriging estimation. KEFF shows the degree of over smoothing for the local block grade, a low KEFF indicates high degree of over smoothing while a high KEFF indicates a low degree of over smoothing (Nwaila et al., 2025). KEFF is a comparative matrix where the estimated block variance is normalized by the true block variance and a high KEFF is a sign that the two are almost equal while a low KEFF indicates that the Kriging variance is higher than the block variance (Deutsch *et al.,* 2014). The KEFF from the block model generated shows that there is low degree of over smoothing as most of the values are above 50% and close to 100% shown in (Fig 4.49). The spatial distribution of the KEFF values is shown in Figure 4.48, where red areas indicate high values and lime-yellow areas represent low values. These low KEFF areas indicate poorly informed areas due to fewer data points used in the Kriging estimation which generally means low levels of confidence and are areas of high uncertainty. This is because domain 0 has sparsely sampled points and are fewer (495 points). High KEFF areas in the other hand are areas of higher confidence and lower uncertainty because of more data used in the estimation. Domain 1 has higher KEFF values because there is better sample density, and more sample points included in the estimation (1386 points). The low KEFF areas are problematic for mining operations because of this uncertainty, they pose a financial risk due to the low levels of confidence. If mine planning and extraction plans are implemented in these zones' resources may be wasted in pursuit of low-grade ore disguised as high grade. So, dealing with these low KEFF requires a strategic approach to collect more data and use geological knowledge of the area to justify further investigation of the area.

The other Kriging performance indicator used is the SLOR, which measures the performance of the Kriging estimation by using regression analysis of the estimated values to the theoretical true values (Rossi & Deutsch, 2013). The SLOR is a measure of the conditional bias from the Kriging estimation. The conditional bias is the deviation of the estimated values from the true values. It is somewhat a natural phenomenon synonymous with Kriging estimation due to its inherent smoothing effect of sparsely distributed data points (McLennan and Deutsch, 2002) and its spatial distributions is shown in (Fig 4.10). The SLOR for the block model shows that most of the value's range between 70% and 100% shown in (Fig 4.51). These values are high

and indicative that the estimated values are closer to the expected values while areas with lower values signal that there is deviation of the estimated values from the expected true values. Low SLOR areas pose a high risk in the mining operation as they may be areas where the grades are either overestimated or underestimated. Just like the KEFF, it has an impact on the business case and the economic baseline of the operation as resources can be wasted mining ore that is indicated as a high grade in the block model while it is low-grade ore. So careful mine planning must go into extracting ore from these zones and consideration should be put into the underlying geological information to guide the selection of mining areas to prevent losses resulting from these areas.

The swath plots for both X-axis and Y-axis slices show that when comparing the block estimated values and the true values there isn't such a significant difference in both directions shown in (Fig 4.52, Fig 4.53. Fig 4.54 and Fig 4.55). The swath plots show that the estimation resulted in a model that can be trusted to represent reality in space, and it meets the permitted confidence margins as recorded by the KEFF and SLOR values. These good estimation performance results show that indeed the pointwise domain generation has succeeded in assigning data points to their rightful populations and the approach can be incorporated into exploration and mining projects.

## 5.6 Machine Learning Potential for Clustering and Domain Boundary Generation in Mineral Resources Evaluation and Mineral Resources Management

The importance of domaining in mineral resource estimation cannot be ignored as it forms the basis of geostatistical estimations. Different domaining approaches have been explored using various methods over time. Even though this is a well-studied area of mineral resource estimation there is still widespread confusion among practicing geostatistitians regarding the appropriate methodology to use (Emery & Ortiz, 2004). This is evident in the different domaining approaches used in various commercial modelling software's. In this era of AI and ML, a new window has been opened to explore other advanced statistical and mathematical methods to use in generating reliable domains for mineral resource estimation (Dumakor-Dupey and Arya, 2021). This research took a leap into this space by exploring the integration of ML methods in the mineral resource estimation process while still upholding the basic principles of stationarity, homogeneity, and continuity of zones in which estimation is to be done. ML methods have proven to provide benefits in handling big datasets and have high computational power that generates results in a short period compared to traditional estimation methods (Zhang *et al.*, 2023; Nwaila *et al.*, 2025; Ying *et al.*, 2025). These methods have the potential to save costs for mining operations as most processes are automated and require less manual work. This can lead to shorter turnaround times and help improve the dynamic

update of geological and resource models. ML methods have recently gained popularity in the mineral resource estimation front. Dumakor-Dupey and Arya, (2021) reviewed ML applications in mineral resource estimation and has presented an increase in the publication of these methods from 1993 to 2020. These numbers are expected to increase exponentially with an increase in understanding of the methods and the establishment of further efficient methods to be used in the automation of some of these methods. Zhang *et al.*, (2023) used a radial basis function motivated ML approach in automating domain boundary delineation using a data set from Driefontein goldmine in the Witwatersrand Basin in South Africa. Their research resulted in spatially coherent geological domains that can be used in mining operations. Even though most of the applications are in research, the mining industry is also embracing this relatively new frontier. The commercial modelling software Maptek introduced their novel approach in domaining named DomainMCF which produces domains or grade models from drilling data in the shortest time compared to the traditional mineral resource estimation methods (Maptek, 2024). Despite all these advancements made in ML, the performance is still dependent on the complexity of the deposit, data density, and the size of the deposit. This research provides an opportunity to use hybrid methods that combine machine ML methods and traditional estimation techniques that have been used in the mining industry for decades. The fusion of these methods creates an inclusive approach that maintains the industry standards while optimising the mineral resources estimation process which may not be realised with individual methods.

**Limitations**

This project was based on a 2D model for a narrow tabular deposit which does not capture the spatial extent of the orebody which may be a challenge as the properties of these deposits may vary significantly in a 3D space. So, exploring a 3D model will be best suited for application in a mining process to realise the full potential of the method. Some of these ML algorithms are super sensitive to the selection of hyperparameters as such their application should be well controlled to prevent poor predictions or overlearning of models. These sensitive algorithms are:

a) SVM – kernel and gamma parameters etc.

b) Neural Networks – number of layers, number of neurons, learning rate and activation function parameters etc.

c) XGBoost – number of trees, learning rate, max depth and subsample ratio etc.

d) kNN – distance metric and number of neighbours.

These models require careful tuning of parameters to attain optimal prediction results. These ML models were trained and tested over Au (cm.g/t), Au (g/t) and CW (cm) which is a small

dataset, so expanding the dataset may yield better predictions and this may include a) structural data which may be less continuous (e.g fold, faults and shear zones), b) geophysical data (e.g magnetic susceptibility, resistance, gravity and seismic data) and c) mineral alteration patters (e.g oxidation , alteration data). All this extra data will be dependent on the type of deposit under investigation so the orebody will forever dictate, and this integration will help improve the classification methods.

# 6 CONCLUSIONS

In this research, machine learning algorithms were used to develop mineral resource estimation domains of a narrow tabular orebody (i.e. Carbon Leader Reef) of the Witwatersrand gold deposits. The proposed method is a hybrid method that combines ML algorithms and ordinary Kriging. The development of mineral resource estimation domains in this project was done by splitting samples into discrete domains using point sample classification algorithms. These algorithms are the Standard K-means clustering and the enhanced signed-distance-mds-K-means clustering method. Both methods resulted in two domains each when using the silhouette scores and the Davies-Bouldin index statistical analysis. The Standard K-means clustering showed better intra cluster separation, even though this is the case domain knowledge is important in interpreting ML performances. The signed-distance-mds-K-means did not have good clustering results statistically but visually the cluster configuration showed well separated and unmixed domains compared to the Standard K-means clustering method as such it was the domains were selected for estimation. This mixing of representative spatial points across domains that they don't belong has the potential of negatively impacting the grade estimation as it may lead to over estimation or underestimation of grade values in the unsampled locations. These poor estimates lead to misclassification of mineral resources and mineral reserves that affect the overall economical assessment of the deposit.

The point-based clustering above is the first step of the two staged approach to building domains by developing spatial clusters that will inform the domain demarcations. The second stage is the delineation of the boundary lines separating adjacent domains using further classification algorithms which underpins the primary objective of this research. Various algorithms were explored for this purpose, and they included: (a) SVM, (b) kNN (c) RF, (d) AB and (e) SVM manually tuned. RF had the best performance statistics in its F1 and accuracy scores that are at 0.999 each while kNN had the lowest scores at 0.994 each. Even though RF algorithm resulted in better statistics compared to the kNN algorithm, the kNN algorithm produced a domain boundary that was consistent with the reef contacts of this Carbon Leader Reef deposit that is characterised by uneven boundaries showing a transition from high energy depositional environment to a low energy depositional environment. This reef is characterised by a sharp and planar contact to the quartzite footwall and at times the carbon layer is located on a competent quartzite base that results in this hard boundary transition. Furthermore, it can be indicative of unconformity in the depositional process of the conglomerates that generates erosional features that can display the hard boundaries in this deposit.

This underpins many concerns of using purely data driven methods in geoscientific problems, as such these attests to the importance of including geological understanding of different ore forming process in selecting algorithms to use. This also shows that ML algorithms are not a one size fits all and careful consideration needs to go into the selection criteria and process development to fit the particular purpose.

The kNN domain polygon from the boundary generation was used for ordinary Kriging estimation and this included the domain points generated by the signed-distance-mds-K-means clustering method, the post estimation analysis show that the KEFF and the SLOR both measured above 50% which is acceptable in the mining industry. KEFF and SLOR above 50% indicate that there is minimal conditional bias and minimal smoothing due to weak data structure. Furthermore, the swath plots show good grades estimations when compared with the true grade values as such this gives high confidence in creating reliable block models and determining cut off grades for mine planning and designing phases.

Finally, kNN algorithm has improved the Kriging estimation and has potential to be replicated in other deposits and can enhance other spatial driven predictive models. It should be noted that to confidently run any estimation model there must be quantification of accuracy and precision of the predictive model. As such statistical methods such as KEFF, SLOR and swath plots should be used to measure the performance of the model to ensure quality results.

# 7 RECOMMENDATIONS

The use of ML workflows in mineral resource estimation domains development in this project has shown to offer flexible and robust options to traditional methods. It can be a key improvement in mining operations where there is continuous drilling and multiple datasets being produced in a short period of time. The following recommendations are derived from the outcome of this research:

- Point sample domain clustering step both the K-means and the signed-based-distance-K-means clustering took a very long time (45 minutes to 1hr 30minutes) to run and may require a computer with a higher computational capability to handle larger datasets and more complex deposits. The time taken to run an algorithm is also predominantly determined by the number of iterations ran for the predictions. Careful consideration must be given to the selection of model parameters and the process of hyperparameter tuning that greatly affects processing time. Zhao et al., (2024) highlighted that the use of ML and AI in geosciences is faced with a challenge of computational demands to run models optimally, this is reality in areas generating high volume data for processing.

- The combination of K-means clustering with geologically derived features can significantly improve the point domain classification and enhance the domain boundary generation using the kNN algorithm. Khorram et al., (2025) showed that the use of both qualitative and quantitative data in domain delineation using Self Organizing Maps model that combined geological layers a) minerilsation zone b) lithologies c) alterations, Geochem layers including d) Cu, Mo, Ag and Au grades and e) sample spatial locations resulted in a robust classification of the domains.

- The mixture of sample points across domains should not be accepted and the domain boundaries should be representative of the kind of geological processes influencing the morphology and grade distribution of the deposit being studied. Ying et al., (2025) shows the effect of poorly domained samples in estimations that resulted in 30% high grade wrong predictions.

- It is important to test the consistency and accuracy of these domains by some traditional statistics and geostatistical calculations such as KEFF and SLOR to test the outcome of estimations which is affected by the domain selections process. Furthermore, domains should be carefully selected taking into cognisance the

underlaying geoscientific properties such as stationarity which are key in geostatistical analysis. Nwaila et al., (2025) emphasised the need for resource model performance assessment to determined how well the estimated values match the true values and be cautious not to have unreasonably large blooks with overly smoothed data that maybe detrimental for short term mine planning.

# 8 REFERENCES


Alcaraz, S., Lane, R., Spragg, K., Milicich, S., Sepulveda, F., Bignall, G., 2011. 3d geological modelling using new Leapfrog Geothermal software, in: Proceedings of the 36th Workshop on Geothermal Reservoir Engineering, Stanford University, Stanford, CA, USA.

Amwaama, M.M.N., 2018. Comparison of Kriging neighbourhood analysis (KNA) results from two different software packages: A case study.

Anderson, J.R., 1983. Machine learning: An artificial intelligence approach. Morgan Kaufmann, 95 First Street , Los Altos, California 94022.

Armstrong, M., 1998. Basic linear geostatistics. Springer Science & Business Media.

Asghari, O., Safikhani, M., Talesh Hosseini, S., 2020. Determining the optimum search range for 2D and 3D mapping based on Kriging through quantitative analysis. Boll. Geofis. Teor. Ed Appl. 61.

Atalay, F., 2025. Effect of Domaining in Mineral Resource Estimation with Machine Learning. MDPI 15, 330.

Ayodele, T. O. (2010). Types of machine learning algorithms. *New advances in machine learning*, *3*(19-48), 5-1.

Bargawa, W.S., Amri, N.A., 2016. Mineral resources estimation based on block modeling, in: AIP Conference Proceedings. AIP Publishing.

Barnett, V., 1994. Outliers in statistical data. John Wiley Sons Google Sch. 2, 705–708.

Birch, C., 2014. New systems for geological modelling-black box or best practice? J. South. Afr. Inst. Min. Metall. 114, 993–1000.

Bohling, G., 2005. Introduction to geostatistics and variogram analysis. Kans. Geol. Surv. 1, 1–20.

Boisvert, J.B., Manchuk, J.G., Deutsch, C.V., 2009. Kriging in the presence of locally varying anisotropy using non-Euclidean distances. Math. Geosci. 41, 585–601.

Bostan, P., 2017. Basic Kriging methods in geostatistics. Yuzuncu Il Univ. J. Agric. Sci. 27, 10–20.

Botchkarev, A., 2018. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. ArXiv Prepr. ArXiv180903006.

Buck, S., Minter, W., 1985. Placer formation by fluvial degradation of an alluvial fan sequence: the Proterozoic Carbon Leader placer, Witwatersrand Supergroup, South Africa. J. Geol. Soc. 142, 757–764.

Burkov, A. (2019). *The hundred-page machine learning book* (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.

Cairncross, B., 2021. The Witwatersrand Goldfield, South Africa. Taylor Francis 96, 296–351.

Cevik, I.S., Leuangthong, O., Cate, A., Ortiz, J.M., 2021. On the use of machine learning for mineral resource classification. Min. Metall. Explor. 38, 2055–2073.

Coombes, J., 2008. The art and science of resource estimation: a practical guide for geologists and engineers. Coombes Capability, Perth Australia.

Goncalves, Rui. (2021). *Practical Implementation Of Radial Basis Functions In Sparse Datasets in Resource Estimation.* [Master of Science in Mining Engineering, Witwatersrand]. https://wiredspace.wits.ac.za/server/api/core/bitstreams/777ad42f-d966-4e2e-b5f8-7ed0c8438774/content

Cowan, E.J., Beatson, R.K., Fright, W.R., McLennan, T.J., Mitchell, T.J., 2002. Rapid geological modelling, in: Applied Structural Geology for Mineral Exploration and Mining, International Symposium. pp. 23–25.

Cressie, N., 1988. Spatial prediction and ordinary Kriging. Math. Geol. 20, 405–421.

CRIRSCO (2014) 'Committe for mineral reserves international reporting standards.' CRIRSCO. Available at: https://crirsco.com/wp-content/uploads/2024/04/5_standard_definitions.pdf (Accessed: 12 April 2025).

CRIRSCO. (2024, June). *INTERNATIONAL REPORTING TEMPLATE FOR THE PUBLIC REPORTING OF EXPLORATION TARGETS, EXPLORATION RESULTS, MINERAL RESOURCES AND MINERAL RESERVES*.CRIRSCO.https://crirsco.com/wp content/uploads/woocommerce_uploads/2024/06/CRIRSCO_International_Reporting_Template_June2024_Update_Approved_for_Release_20240627-dl8515.pdf

Dankert, B.T., Hein, K.A., 2010. Evaluating the structural character and tectonic history of the Witwatersrand Basin. Precambrian Res. 177, 1–22.

Daya, A.A., Bejari, H., 2015. A comparative study between simple Kriging and ordinary Kriging for estimating and modeling the Cu concentration in Chehlkureh deposit, SE Iran. Arab. J. Geosci. 8, 6003–6020.

Deutsch, C.V., 2007. The slope of regression for Kriging estimators. Cent. Comput Geostat 9, 311.

Deutsch, J.L., Szymanski, J., Deutsch, C.V., 2014. Checks and measures of performance for Kriging estimates. J. South. Afr. Inst. Min. Metall. 114, 223–223.

Dohm, C., 2021. MINN 7007 Statistical evaluation of mineral resources lecture notes.

Dominy, S.C., Annels, A.E., 2001. Evaluation of gold deposits—Part 1: review of mineral resource estimation methodology applied to fault-and fracture-related systems. Appl. Earth Sci. 110, 145–166.

Dominy, S.C., Stephenson, P.R. and Annels, A.E. (2001) 'Classification and Reporting of Mineral Resources for High-Nugget Effect Gold Vein Deposits', *Exploration and Mining Geology*, 10(3), pp. 215–233. Available at: https://doi.org/10.2113/0100215.

Duboue, P., 2020. The art of feature engineering: essentials for machine learning. Cambridge University Press.

Dumakor-Dupey, N.K., Arya, S., 2021b. Machine Learning—A Review of Applications in Mineral Resource Estimation. Energies 14, 4079. https://doi.org/10.3390/en14144079

Elliot, S.M., Snowden, D.V., Bywater, A., Standing, C.A., Ryba, A., 2001. Reconciliation of the McKinnons gold deposit, Cobar. New South Wales AC Edw. Ed Miner. Resour. Ore Reserve Estim. AusIMM Guide Good Pract. Australas. Inst. Min. Metall.

Els, B., 1988. Textural and mineralogical composition of some fluvial Witwatersrand rocks of the West Wits Line Goldfield. South Afr. J. Geol. 91, 543–549.

Emery, X., Ortiz, J.M., 2004. Defining geological units by grade domaining. Technical report, Universidad de Chile.

Erdogan Erten, G., Yavuz, M., Deutsch, C.V., 2022. Combination of Machine Learning and Kriging for Spatial Estimation of Geological Attributes. Nat. Resour. Res. 31, 191–213. https://doi.org/10.1007/s11053-021-10003-w

Khorram, F., O. Asghari, H. Memarian, A. Hoseein Morshedy, X. M. Emery, 2021. Adaptive multi-size block modeling for mineral resources and ore reserves evaluation. Shahrood Univ. Technol. 12, 339--350.

Fouedjio, F., Hill, E.J., Laukamp, C., 2018. Geostatistical clustering as an aid for ore body domaining: case study at the Rocklea Dome channel iron ore deposit, Western Australia. Appl. Earth Sci. 127, 15–29.

Fränti, P., Sieranoja, S., 2019. How much can k-means be improved by using better initialization and repeats? Pattern Recognit. 93, 95–112. https://doi.org/10.1016/j.patcog.2019.04.014

Frimmel, H.E., Groves, D.I., Kirk, J., Ruiz, J., Chesley, J., Minter, W.E.L., 2005. The formation and preservation of the Witwatersrand goldfields, the world's largest gold province.

Frimmel, H.E., Minter, W.E.L., 2002. Recent developments concerning the geological history and genesis of the Witwatersrand gold deposits, South Africa.

Frimmel, H.E., Nwaila, G.T., 2020. Geologic evidence of syngenetic gold in the Witwatersrand Goldfields, South Africa.

Frimmel, H.E., Zeh, A., Lehrmann, B., Hallbauer, D., Frank, W., 2009. Geochemical and geochronological constraints on the nature of the immediate basement next to the Mesoarchaean auriferous Witwatersrand Basin, South Africa. J. Petrol. 50, 2187–2220.

Geeks for Geeks. (2024, March 20). *Clustering in Machine Learning* [Online Post]. https://www.geeksforgeeks.org/clustering-in-machine-learning/
Glacken, I.M., Snowden, D.V., Edwards, A.C., 2001a. Mineral resource estimation. Miner. Resour. Ore Reserve Estim.- Aus Inst Min. Metall. Guide Good Pract. 189–198.
Gómez-Hernández, J.J., 2005. Geostatistics, in: Hydrogeophysics. Springer, pp. 59–83.
Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Oxford University Press, USA.
Goncalves, Rui., 2021. PRACTICAL IMPLEMENTATION OF RADIAL BASIS FUNCTIONS IN SPARSE DATASETS IN RESOURCE ESTIMATION (Master of Science in Mining Engineering). Witwatersrand, Johannesburg.
Hao, D.V., Nguyen, D.C., Klityński, W., Zygo, W., Nowak, J., 2023. 3D block modelling of the Sin Quyen IOCG deposit, North Vietnam. Geol. Geophys. Environ. 49.
Heinrich, C.A., 2015. Witwatersrand gold deposits formed by volcanic rain, anoxic rivers and Archaean life. Nat. Geosci. 8, 206–209.
Hong, J., Oh, S., 2021. Model selection for mineral resource assessment considering geological and grade uncertainties: Application of multiple-point geostatistics and a cluster analysis to an iron deposit. Nat. Resour. Res. 30, 2047–2065.
Isaaks, E. H., & Srivastava, R. M. (1989). Applied geostatistics.
Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electron. Mark. 31, 685–695.
Japkowicz, N., 2006. Why question machine learning evaluation methods, in: AAAI Workshop on Evaluation Methods for Machine Learning. University of Ottawa.
Japkowicz, N., Shah, M., 2015. Performance evaluation in machine learning. Mach. Learn. Radiat. Oncol. Theory Appl. 41–56.
Jara, RM., Couble, A., Emery, X., Magri, EJ., Ortiz, JM., 2006. Block size selection and its impact on open-pit design and mine planning. South. Afr. Inst. Min. Metall. 106, 205--211.
Jooshaki, M., Nad, A., Michaux, S., 2021. A systematic review on the application of machine learning in exploiting mineralogical data in mining and mineral industry. Minerals 11, 816.
JORC, 2012. Australasian Code for Reporting of Exploration Results, Mineral Resources and Ore Reserves.
Journel, A.G., Huijbregts, C.J., 1976. Mining geostatistics. Blackburn Press.
Jowitt, S.M., McNulty, B.A., 2021. Geology and Mining: Mineral Resources and Reserves: Their Estimation, Use, and Abuse. SEG Discov. 27–36. https://doi.org/10.5382/Geo-and-Mining-11
Juba, B., Le, H.S., 2019. Precision-recall versus accuracy and the role of large data sets, in: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4039–4048.
Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., Kumar, V., 2018. Machine learning for the geosciences: Challenges and opportunities. IEEE Trans. Knowl. Data Eng. 31, 1544–1554.
Kasmaee, S., Raspa, G., de Fouquet, C., Tinti, F., Bonduà, S., Bruno, R., 2019. Geostatistical estimation of multi-domain deposits with transitional boundaries: A sensitivity study for the Sechahun iron mine. Minerals 9, 115.
Khorram, F., Rayes, M., Pais, G., 2025. Domain Delineation Using Geological Data, Variogram Analysis, and Clustering Algorithms. J. Geosci. Environ. Prot. 13, 31–47.
King, H.F., 1982. Guide to the understanding of ore reserve estimation. Australas Inst Min Met. ProcAustralia.
Kwak, S.K., Kim, J.H., 2017. Statistical data preparation: management of missing values and outliers. kja 70, 407–411. https://doi.org/10.4097/kjae.2017.70.4.407
Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10.
Lee, I., Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Bus. Horiz. 63, 157–170.
Loquin, K., Dubois, D., 2010. Kriging and epistemic uncertainty: a critical discussion. Methods Handl. Imperfect Spat. Inf. 269–305.

Mahboob, M.A., Celik, T., Genc, B., 2023. Review of machine learning-based Mineral Resource estimation. J. South. Afr. Inst. Min. Metall. 122, 1–10. https://doi.org/10.17159/2411-9717/1250/2022
Manchuk, J.G., 2010. Geostatistical modeling of unstructured grids for flow simulation.
Maniar, H., Ryali, S., Kulkarni, M.S., Abubakar, A., 2018. Machine-learning methods in geoscience, in: SEG International Exposition and Annual Meeting. SEG, p. SEG-2018-2997218.
Maptek, 2024. Machine learning for domain modelling. Mach. Learn. Domain Model. URL https://www.maptek.com/products/domainmcf/ (accessed 7.2.24).
Mariz, C., Wagner, L., Pilger, G., Whitehouse, P., Bessin, C., & Laurent, D. (n.d.). *Geostatistical Clustering, Potential Modeling and traditional geostatistics: A new coherent workflow*.
Matheron, G., 1963. Principles of geostatistics. Econ. Geol. 58, 1246–1266.
McCoy, J.T., Auret, L., 2019. Machine learning applications in minerals processing: A review. Miner. Eng. 132, 95–109. https://doi.org/10.1016/j.mineng.2018.12.004
McLennan J, Deutsch C, 2002. Conditional bias of geostatistical simulation for estimation of recoverable reserves. CIM Proc. Vanc.
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A., 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv. CSUR 54, 1–35.
Meinert, L.D., Robinson Jr, G.R., Nassar, N.T., 2016. Mineral resources: Reserves, peak production and the future. Resources 5, 14.
Mery, N., Marcotte, D., 2022. Quantifying mineral resources and their uncertainty using two existing machine learning methods. Math. Geosci. 54, 363–387.
Mohammadi, H., Hosseini, S.T., Asghari, O., Harouni, P.A., 2022. Ore body domaining by clustering of multiple-point data events; a case study from the Dalli porphyry copper-gold deposit, central Iran. Ore Energy Resour. Geol. 10, 100018.
Nasteski, V., 2017. An overview of the supervised machine learning methods. Horiz. B 4, 56.
Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., & Tomović, M. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics & Computer Science*, *7*(1), 39.
Nwaila, G.T., Rose, D.H., Frimmel, H.E., Ghorbani, Y., 2025. An Integrated Geodata Science Workflow for Resource Estimation: A Case Study from the Merensky Reef, Bushveld Comple. Nat. Resour. Res. 1–29.
Nwaila, Glen T. and Frimmel, Hartwig E., A New Approach to Geodomaining: Signed Distance-Based Clustering and Combinatorics. Available at SSRN: https://ssrn.com/abstract=4475019 or http://dx.doi.org/10.2139/ssrn.4475019
Nwaila, G.T., Zhang, S.E., Frimmel, H.E., Manzi, M.S., Dohm, C., Durrheim, R.J., Burnett, M., Tolmay, L., 2020. Local and target exploration of conglomerate-hosted gold deposits using machine learning algorithms: a case study of the Witwatersrand gold ores, South Africa. Nat. Resour. Res. 29, 135–159.
Olea, R.A., 1999. Simple Kriging. Geostat. Eng. Earth Sci. 7–30.
Oliver, M.A., Webster, R., 2015. Basic steps in geostatistics: the variogram and Kriging. Springer.
Oliver, M.A., Webster, R., 1990. Kriging: a method of interpolation for geographical information systems. Int. J. Geogr. Inf. Syst. 4, 313–332.
Ortiz, J.M., Emery, X., 2006. Geostatistical estimation of mineral resources with soft geological boundaries: a comparative study. J.-SOUTH Afr. Inst. Min. Metall. 106, 577.
Patterson, J.A., 1959. Estimating ore reserves follows logical steps. Eng. Min. J. 160, 111–115.
Petrovic, S. (2006, October). A comparison between the silhouette index and the davies-bouldin index in labelling ids clusters. In *Proceedings of the 11th Nordic workshop of secure IT systems* (Vol. 2006, pp. 53-64). Citeseer.
Phillips, G.N., Law, J.D., 2000. Witwatersrand gold fields: geology, genesis, and exploration.
Phillips, G. N., & Powell, R. (2011). Origin of Witwatersrand gold: a metamorphic devolatilisation–hydrothermal replacement model. *Applied Earth Science*, *120*(3), 112-129.Raschka, S., 2018.

Model evaluation, model selection, and algorithm selection in machine learning. ArXiv Prepr. ArXiv181112808.Rendu, J.M., Miskelly, N., 2001. Mineral resources and mineral reserves: Progress on international definitions and reporting standards. Min. Technol. 110, 133–138.
Phillips, G.N, 2011. Gold exploration success. Appl. Earth Sci. 120, 7–20.

Robb, L.J., Meyer, F.M., 1995. The Witwatersrand Basin, South Africa: geological framework and mineralization processes. Ore Geol. Rev. 10, 67–94.
Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Trans. Pattern Anal. Mach. Intell. 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187
Rokach, L., Maimon, O., 2005. Clustering methods. Data Min. Knowl. Discov. Handb. 321–352.
Rossi, Mario E., Deutsch, C.V., 2013. Mineral resource estimation. Springer Science & Business Media.
Sabourin, R., 1983. Geostatistics as a tool to define various categories of resources. J. Int. Assoc. Math. Geol. 15, 131–143.
Safonov, Y.G., Prokof'ev, V.Y., 2006. Gold-bearing reefs of the Witwatersrand Basin: a model of synsedimentation hydrothermal formation. Geol. Ore Depos. 48, 415–447.
Sahoo, M.M., Pal, B.K., 2017. Geological modelling of a deposit and application using Surpac. J. Mines Met. Fuels 65, 417–422.
SAMREC CODE (2016) 'The South African Code For Reporting Of Epxloration Results, Mineral Resources and Mineral Reserves.' SAMREC. Available at: https://www.samcode.co.za/news/194-download-latest-samrec-code (Accessed: 12 April 2025).
Santangelo, M., Marchesini, I., Mirabella, F., Bucci, F., Cardinali, M., Guzzetti, F., 2016. From photo-geology to 3D geological models. An open source approach, in: Proceedings of the 4th Open Source Geospatial Research and Education Symposium (OGRS2016). Perugia. p. 14.
Santra, A.K., Christy, C.J., 2012. Genetic algorithm and confusion matrix for document clustering. Int. J. Comput. Sci. Issues IJCSI 9, 322.
Saputra, D.M., Saputra, D., Oswari, L.D., 2020. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method, in: Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019). Atlantis Press, pp. 341–346.
Shahbeik, S., Afzal, P., Moarefvand, P., Qumarsy, M., 2014. Comparison between ordinary Kriging (OK) and inverse distance weighted (IDW) based on estimation error. Case study: Dardevey iron ore deposit, NE Iran. Arab. J. Geosci. 7, 3693–3704.
Shango Solutions. (2007). *Identification of the Driefontein Multiple-band Carbon Leader Domains* (Technical Report No. S078/06-1; p. 29). Sibanye Stillwater.
Shutaywi, M., Kachouie, N.N., 2021. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy 23, 759.
Sinaga, K.P., Yang, M.-S., 2020. Unsupervised K-means clustering algorithm. IEEE Access 8, 80716–80727.
Sinclair, A.J., Blackwell, G.H., 2006. Applied mineral inventory estimation. Cambridge University Press.
Sterk, R., de Jong, K., Partington, G., Kerkvliet, S., van de Ven, M., 2019. Domaining in Mineral Resource Estimation: A Stock-Take of 2019 Com-mon Practice.
Stone, J.G., Dunn, P.G., 1996. Ore reserve estimates in the real world. Society of Economic Geologists Littleton.
Sun, G., Zeng, Q., Zhou, J., 2022. Machine learning coupled with mineral geochemistry reveals the origin of ore deposits. Ore Geol. Rev. 142, 104753.
Sun, T., Li, H., Wu, K., Chen, F., Zhu, Z., Hu, Z., 2020. Data-driven predictive modelling of mineral prospectivity using machine learning and deep learning methods: A case study from southern Jiangxi Province, China. Minerals 10, 102.

Sunday, C.A., Deutsch, C.V., 2020. Choosing the Discretization Level for Block Property Estimation. Geostat. Lessons.
Talesh Hosseini, S., Asghari, O., 2020. Design Optimization of Estimation Process Using the Quantitative Kriging Neighbourhood Analysis in the Gol-E-Gohar No. 1 Mine. J. Miner. Resour. Eng. 5, 1–18.
Tucker, R.F., Viljoen, R.P., Viljoen, M.J., 2016. A Review of the Witwatersrand Basin-The World's greatest goldfield. Episodes J. Int. Geosci. 39, 104–133.
Van Beers, W.C., Kleijnen, J.P., 2003. Kriging for interpolation in random simulation. J. Oper. Res. Soc. 54, 255–262.
Van Stein, B., Wang, H., Kowalczyk, W., Emmerich, M., Bäck, T., 2020. Cluster-based Kriging approximation algorithms for complexity reduction. Appl. Intell. 50, 778–791.
Vann, J., Guibal, D., 1998. Beyond Ordinary Kriging–An overview of non-linear estimation, in: Proceedings of a One Day Symposium: Beyond Ordinary Kriging. p. 32.
Vann, J., Jackson, S., Bertoli, O., 2003. Quantitative Kriging neighbourhood analysis for the mining geologist-a description of the method with worked case examples, in: 5th International Mining Geology Conference. Bendigo, Australia. Melbourne: Australian Inst Mining & Metallurgy, pp. 215–223.
Wackernagel, H., 2003. Ordinary Kriging. Multivar. Geostat. Introd. Appl. 79–88.
Woodall, R., 1984. Success in mineral exploration: confidence in science and ore deposit models. Geosci. Can.
Wu, J., Wu, J., 2012. Cluster analysis and K-means clustering: an introduction. Adv. K-Means Clust. Data Min. Think. 1–16.
Xu, R., Wunsch, D., 2005. Survey of clustering algorithms. IEEE Trans. Neural Netw. 16, 645–678.
Ying, X., Renguang, Z., Yang, B., 2025. Geological knowledge-guided dual-branch deep learning model for identification of geochemical anomalies related to mineralization. J. Geophys. Res. Mach. Learn. Comput. 2.
Yunsel, T.Y., 2012. Risk quantification in grade variability of gold deposits using sequential Gaussian simulation. J. Cent. South Univ. 19, 3244–3255.
Zhang, S.E., Nwaila, G.T., Bourdeau, J.E., Ghorbani, Y., Carranza, E.J.M., 2023. Machine Learning-Based Delineation of Geodomain Boundaries: A Proof-of-Concept Study Using Data from the Witwatersrand Goldfields. Nat. Resour. Res. 32, 879–900.
Zhao, T., Wang, S., Ouyang, C., Chen, M., Liu, C., Zhang, Jin, Yu, L., Wang, Fei, Xie, Y., Li, J., Wang, Fang, Grunwald, S., Wong, B.M., Zhang, F., Qian, Z., Xu, Y., Yu, C., Han, W., Sun, T., Shao, Z., Qian, T., Chen, Z., Zeng, J., Zhang, Huai, Letu, H., Zhang, B., Wang, Li, Luo, L., Shi, C., Su, H., Zhang, Hongsheng, Yin, S., Huang, N., Zhao, W., Li, N., Zheng, C., Zhou, Y., Huang, C., Feng, D., Xu, Q., Wu, Y., Hong, D., Wang, Z., Lin, Y., Zhang, T., Kumar, P., Plaza, A., Chanussot, J., Zhang, Jiabao, Shi, J., Wang, Lizhe, 2024. Artificial intelligence for geoscience: Progress, challenges, and perspectives. The Innovation 5, 100691. https://doi.org/10.1016/j.xinn.2024.100691
Zhou, Z.-H., 2021. Machine learning. Springer Singapore.

# 9 APPENDICES

## APPENDIX A: CATEGORIAL DATA STATISTICS

Packing

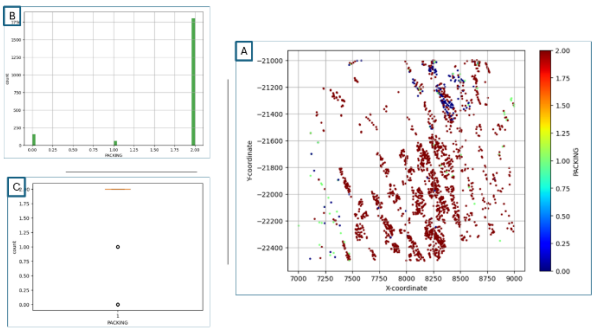


**Figure A1:** Packing graphical statistics and spatial distribution a) Spatial distribution of Packing values as shown in the study area, b) Histogram indicating frequency occurrence of Packing values.

Roundness

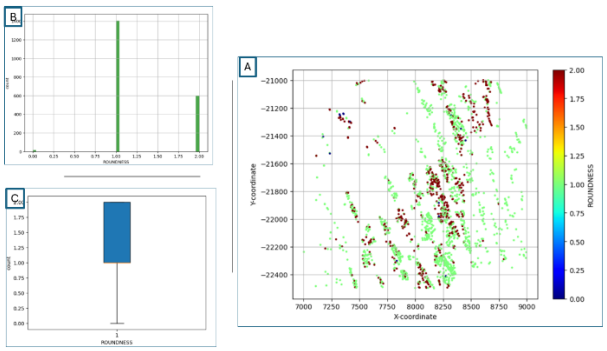


**Figure A2:** Roundness graphical statistics and spatial distribution a) Spatial distribution of Roundness values as shown in the study area, b) Histogram indicating frequency occurrence of Roundness values.

Sorting

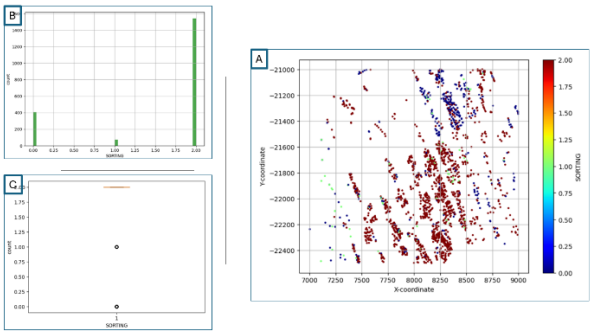


**Figure A3:** Sorting graphical statistics and spatial distribution a) Spatial distribution of Sorting values as shown in the study area, b) Histogram indicating frequency occurrence of Sorting values.

## APPENDIX B: EXPERIMENTAL VARIOGRAMS

Domain 0 Au (cm.g/t)

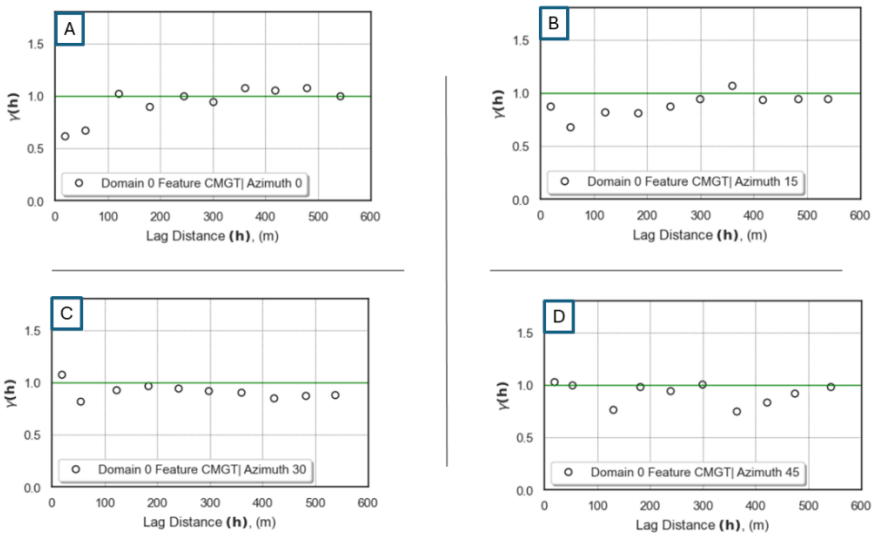


**Figure B1:** Experimental variograms a) Domain 0 Au (cm.g/t) Azimuth 0 b) Domain 0 Au (cm.g/t) Azimuth 15 c) Domain 0 Au (cm.g/t) Azimuth 30 d) Domain 0 Au (cm.g/t) Azimuth 45.

Domain 0 Au (g/t)

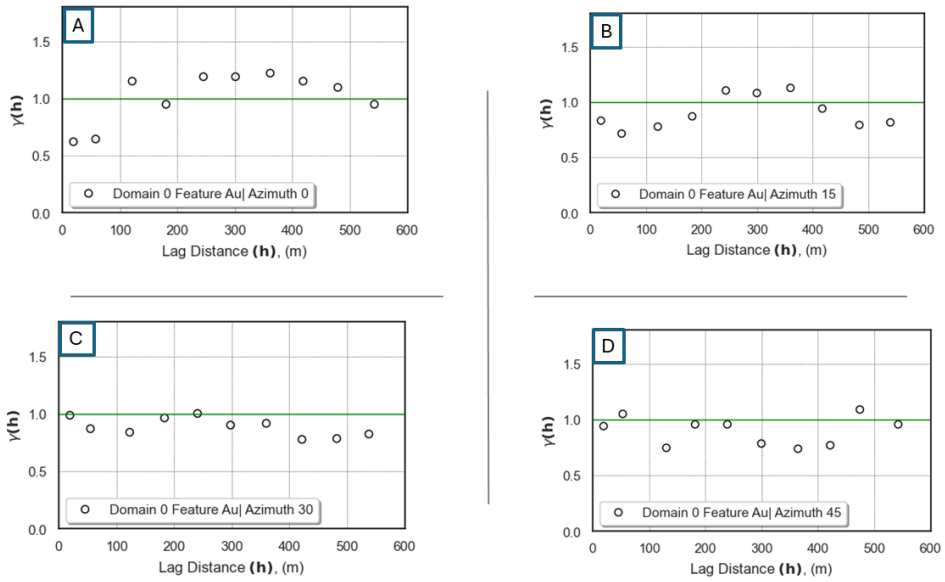


**Figure B2:** Experimental variograms a) Domain 0 Au (g/t) Azimuth 0 b) Domain 0 Au (g/t) Azimuth 15 c) Domain 0 Au (g/t) Azimuth 30 d) Domain 0 Au (g/t) Azimuth 45.

Domain 0 CW (cm)

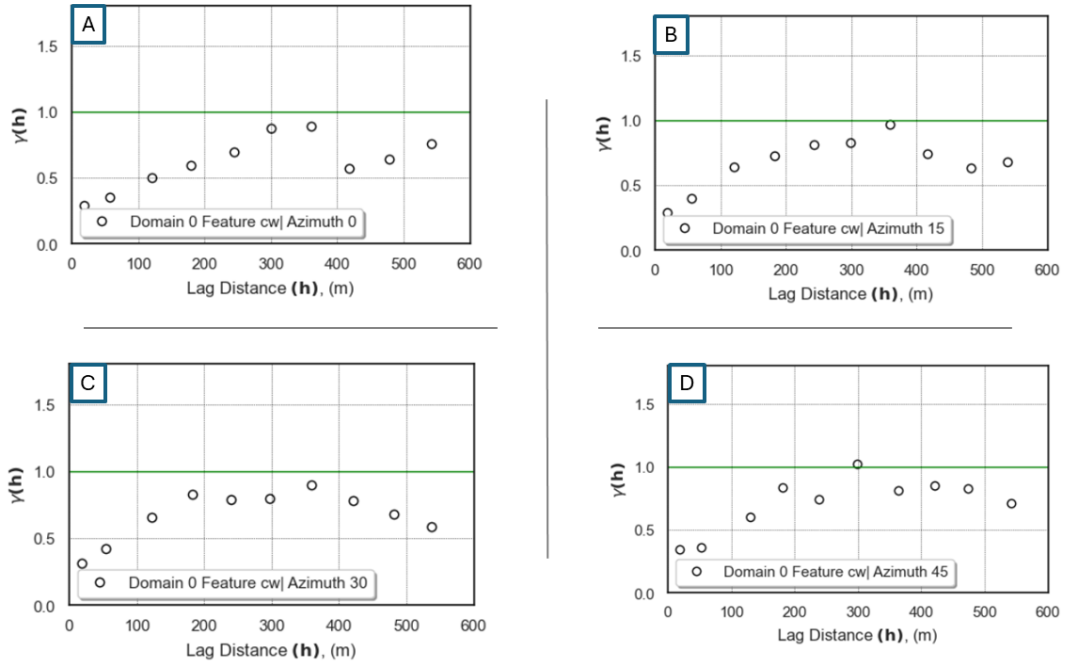


**Figure B3:** Experimental variograms a) Domain 0 CW (cm) Azimuth 0 b) Domain 0 CW (cm) azimuth 15 c) Domain 0 CW (cm) azimuth 30 d) Domain 0 CW (cm) azimuth 45.

Domain 1 Au (cm.g/t)

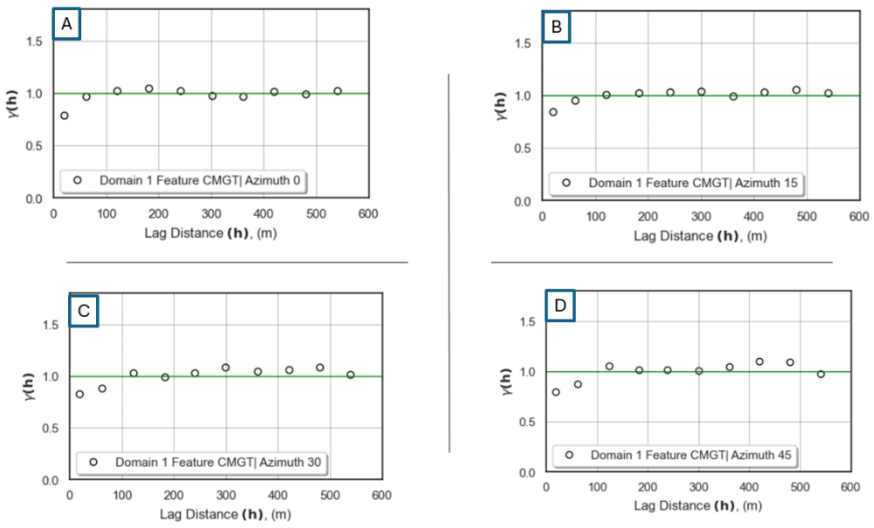


**Figure B4:** Experimental variograms a) Domain 1 Au (cm.g/t) azimuth 0 b) Domain 1 Au (cm.g/t) azimuth 15 c) Domain 1 Au (cm.g/t) azimuth 30 d) Domain 1 Au (cm.g/t) azimuth 45.

Domain 1 Au (g/t)

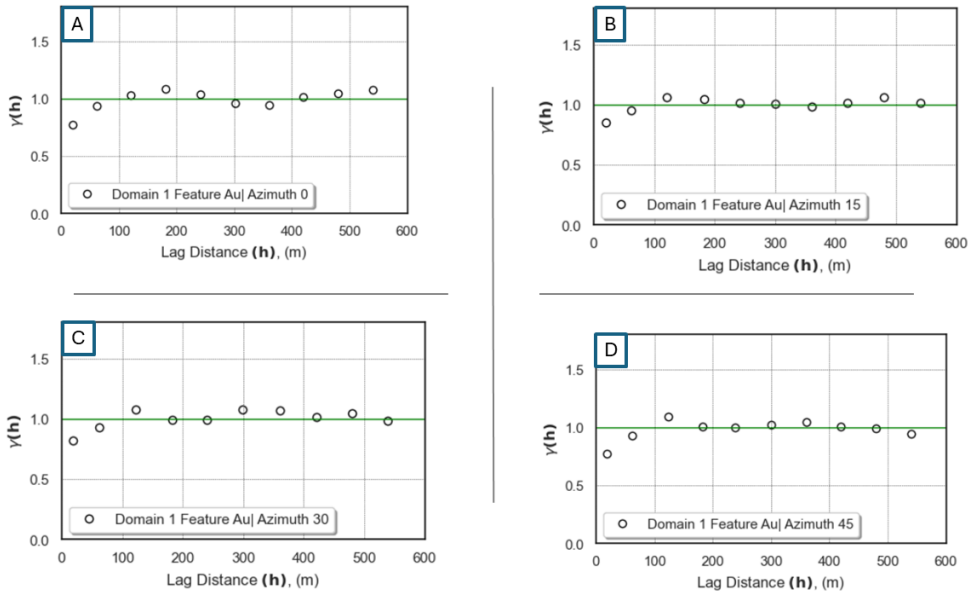


**Figure B5***:* Experimental variograms a) Domain 1 Au (g/t) azimuth 0 b) Domain 1 Au (g/t) azimuth 15 c) Domain 1 Au (g/t) azimuth 30 d) Domain 1 Au (g/t) azimuth 45.

Domain 1 CW (cm)

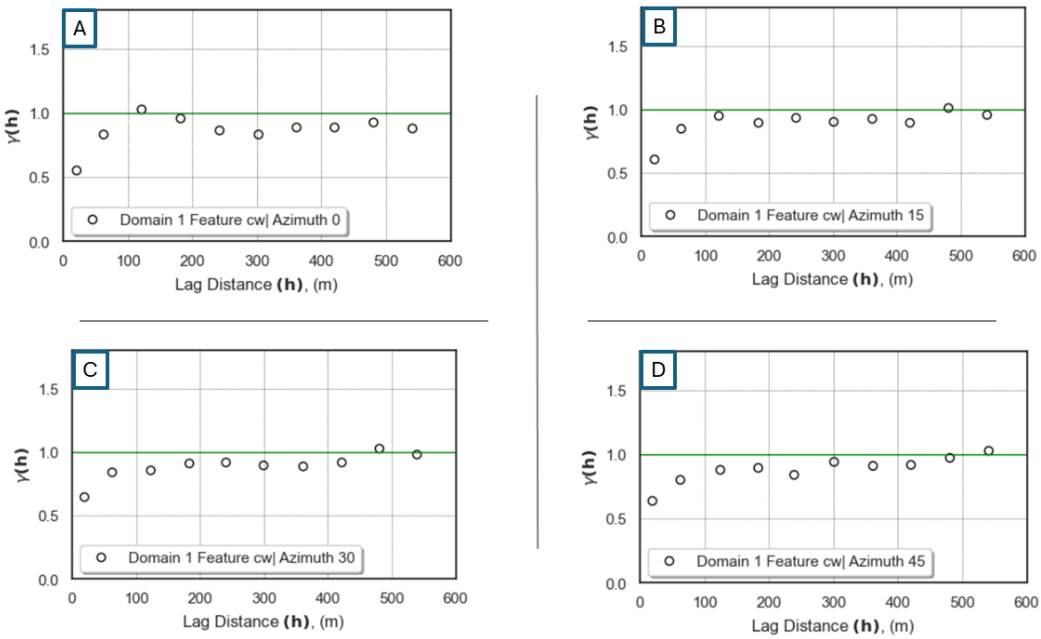


**Figure B6:** Experimental variograms a) Domain 1 CW (cm) azimuth 0 b) Domain 1 CW (cm) azimuth 15 c) Domain 1 CW (cm) azimuth 30 d) Domain 1 CW (cm) azimuth 45.