Defining geological units using geochemical data and unsupervised machine learning[1]

Noble E. Potakey (n.potakey@queensu.ca)
Julian M. Ortiz (julian.ortiz@queensu.ca)

Abstract

One of the preliminary and arguably the most crucial step in a mineral resource evaluation campaign is the determination of the geological domains. Conventional geological methods establish domains primarily by grade zoning or spatial clustering techniques. Even though information about the geology is recorded, most model-based domains do not make much of the geological information derived from logging. Domains are best established with the geological information supported by an accurate statistical analysis of the geochemical data and a good understanding of the deposit. The advent of machine learning techniques such as cluster analysis has advanced this course by providing algorithms that can handle large volumes of multivariate data and try to reproduce geological domains. This paper shows the application of a model-based cluster analysis as a machine learning tool to an exploratory drill hole data set from an undisclosed copper porphyry deposit. The K means algorithm, which was applied in this study utilizes the continuous nature of the non-categorical variables to establish domains. The algorithm generated spatial clusters which had some correlation with the alteration unit even though a confusion matrix revealed the flaws of the method in misclassifying most of the geological units. The choice of the most appropriate number of clusters (domains) to be formed, as well as the selection of variables to drive the clustering process can be challenging when performing k means clustering, and the appraisal of an expert is still necessary, as the results are subjective.

1. Introduction

The mining world has not had enough of mineral exploration. Geological mapping, geophysical investigation, sampling of outcrops, logging of drill cores are examples of exploratory data that needs to be analyzed leaving geologists and engineers overwhelmed with large number of variables. Among these data collected are categorical variables about the lithology, alteration and mineralization of an ore body which is largely obtained from core logging and after measuring several physical, chemical, and mineralogical properties of the rock (Bosch et al., 2002). Knowledge about these geological units is important because domains are traditionally established based on them. Domaining in mineral resource evaluation is a big step in mining because it serves as a backbone for subsequent geostatistical estimation and simulation of the ore body. A poor classification of these domains can lead to the mixing of populations, which can result in bad resource estimates, endangering the valuation of grades and tonnages (Emery & Ortiz, 2005). A simulated model of an ore body will be a success or failure depending

[1] Cite as: Noble E. Potakey, Julian M. Ortiz (2022) Defining geological units using geochemical data and unsupervised machine learning, Predictive Geometallurgy and Geostatistics Lab, Queen's University, Annual Report 2022, paper 2022-03, 47-58.

on the accuracy of the domains established. A domain is formed when an ore body is partitioned into groups of similar characteristics. For example, groups of high element concentration or groups of low rock hardness can form a domain.

Domains are best established with the geological information derived from logging supported by an accurate statistical analysis of the geochemical data and a good understanding of the deposit. However, in most model-based domains, attention is not really given to the geological information when characterizing different mineralized zones even though it is proven that ore grades vary in relation to changes in the geological properties such as mineralogy, lithology and alterations (Yasrebi et al., 2013). Despite the non-reliance of geological information to create domains in most mining environments, it remains one of the fundamental information to building a good domain (Sterk et al., 2019). In the context of unsupervised machine learning, cluster analysis emerges as an efficient tool in classifying sample points based on the intrinsic properties of the input variables. K means algorithm clusters data by grouping sample into clusters of equal variances thereby minimizing a phenomenon known as the *inertia* or within-cluster sum-of-squares (Adams, 2018). This results in clusters that contain objects with similar features and at the same time different from objects belonging to a different cluster. Although the algorithm clusters based on statistical parameters, knowledge about the deposit was used to select variables to drive the clustering process based on their significance to the geological units. For instance, most of the variables in our data were attached to the alteration type and hence variables that are trace elements to the porphyry copper deposit were selected for clustering (Mg, Al, Ga, Li, SC, V). This was done to derive a fast, better performing, and easy to understand model.

An important factor in K means clustering is the choice of the optimal number of clusters (Moreira et al., 2021). This paper addresses this issue, applying and further discussing some of the methods that can be applied as well as the difficulties found when choosing the best configuration of the clusters. The model is validated by comparing the clustered data with logged geological units. Furthermore, a confusion matrix is computed to analyze the errors of misclassification. It is an expectation that the clustered geochemical data reflects the geology. Proven methods for verifying the spatial relationship of the clusters are rarely mentioned in the literature, other than just applying a visual examination of the results.

The work has been divided in three sections. First, we show the exploratory data analysis for selected variables and their distribution in space. This included the histograms and probability plots of continuous and categorical variables. Secondly, we performed K means clustering of selected variables to drive the clustering based on their relevance to the alteration unit. And finally, the validation of the clustered model.

2. Exploratory Data Analysis (EDA)

Exploration data analysis forms an essential part of this project. The exploratory data was de-surveyed with a 15m run length compositing while breaking intervals by geology for all drill holes using *Vulcan* software. Data analysis was done using the python programming software. The data contained 1898 drill holes from a porphyry copper deposit with 50 continuous variables made of 50 geochemical elements and three categorical variables (lithology, alteration, and mineralogy). A total of 21,416 composite samples with assay results were created with 20, 19 and 25 different lithologies, alterations and mineralogy respectively. Deep samples beyond the depth of 1200m were removed and variables whose concentrations were unaccounted for were also not considered. The following table shows a summary

statistic of some of the variables in the dataset for brevity. The abbreviations “Lito”, “Alt” and “Minz” mean Lithology, Alteration and Mineralization respectively.

Table 1 Summary statistics exploratory data. Note that it does not show all the features of the database.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Dhid	21416										
Midx	21416	NaN	NaN	NaN			14971.1				18049.1
Midy	21416	NaN	NaN	NaN			105792.3				107625.6
Midz	21416	NaN	NaN	NaN			1833.9				3110.2
Length	21416	NaN	NaN	NaN			0.04				15
From	21416	NaN	NaN	NaN			0				1197.45
To	21416	NaN	NaN	NaN			0.4				1198.2
Lito	21416	19	50	9603	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Alt	21416	17	51	5621	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Minz	21416	25	70	7712	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Cu_ppm	21416	NaN	NaN	NaN	6112.9	2777.7	46.8	3972	5950	8710	10000
Mo_ppm	21416	NaN	NaN	NaN	92.04652	119.8602	0.45	34.9175	64.7745	113	2890
Mg_ppm	21416	NaN	NaN	NaN	0.442924	0.558493	0.01	0.03	0.094	0.81	3.033
Al_ppm	21416	NaN	NaN	NaN	0.980977	0.661739	0.06	0.46	0.74	1.42	4.519
Ga_ppm	21416	NaN	NaN	NaN	2.591729	2.012614	0.093	1	1.8	3.80525	13.35

Figure 1 below shows the results of the first to third quartiles of all 50 elements to provide us a fair idea of the relationship between the dominant variables and the less dominant ones.

The concentration of copper stands out with extremely high values especially from its third quartile to the maximum. The cumulative probability plot of Cu values shows a fairly log normal distribution with consistently high detection limits of the element as shown in figure 3. The extreme high concentrations of elements could be outliers, measurement or samples errors, or values beyond the detection limit. The negative minimum values of sample 1 (Au) are samples values that were unaccounted for and that was not considered for analysis.

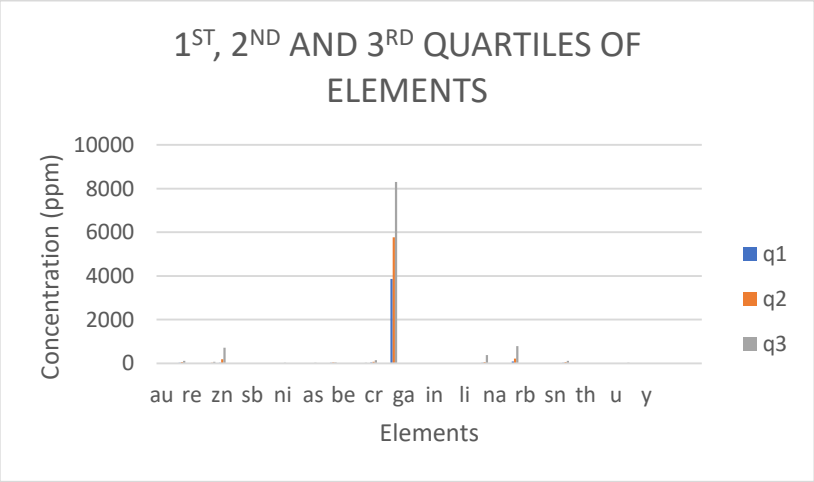

Figure 1 Relatively high Cu concentrations compared other elements.

The following figures represent the box plot of the distribution of copper present in each geological category. This helps us to see the distribution in detail and help identify dominant terrains where our focus should be. Rock code 31, 33 and 50 are the dominant lithologies which host majority of the high grades. Alteration codes 50 and 51 stand out while 50 and 70 are the dominant mineralization codes. Grades of copper are distributed across the features in all three categories and a few outlier values are noted.

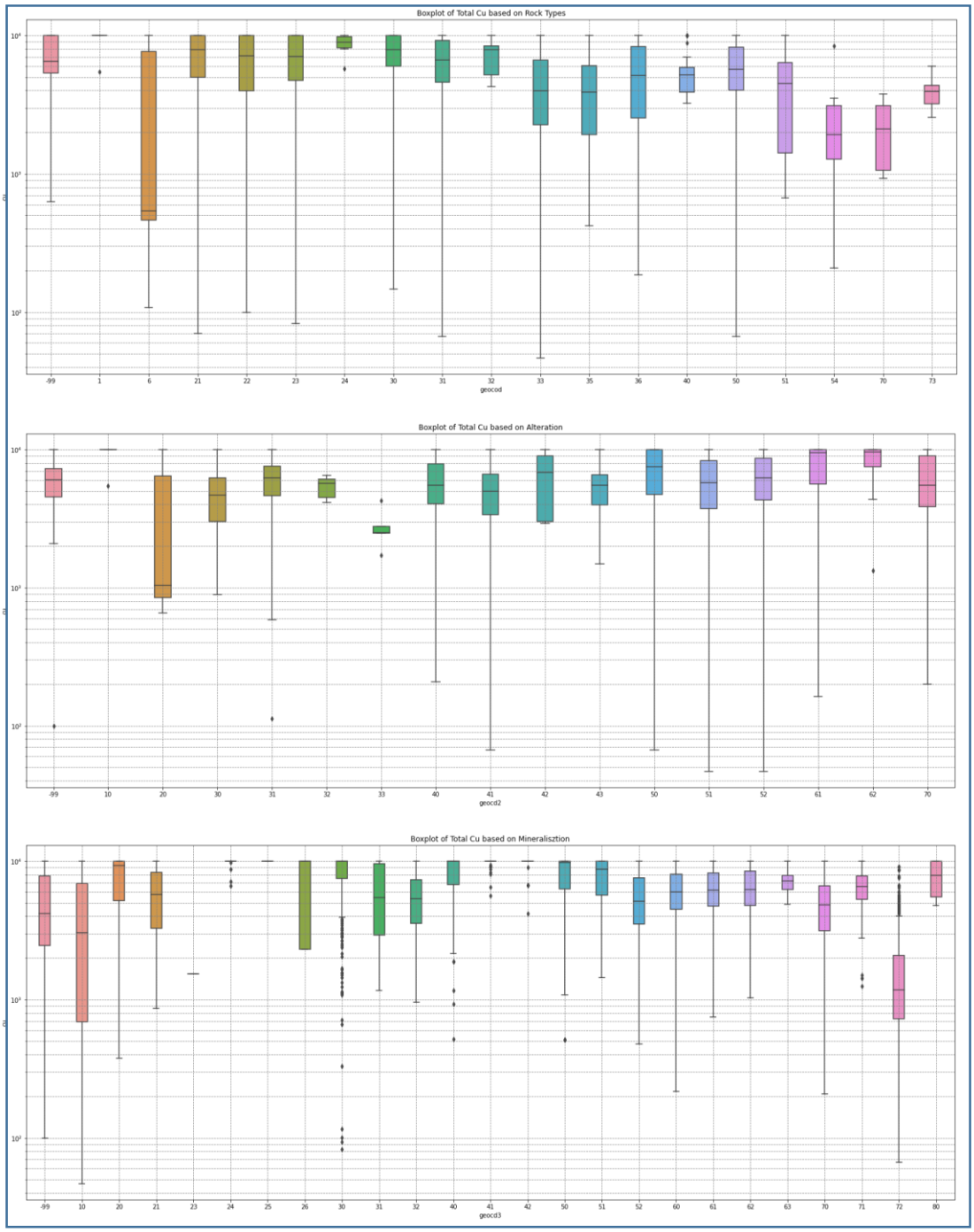

Figure 2 Copper distribution based on the three categories in different features

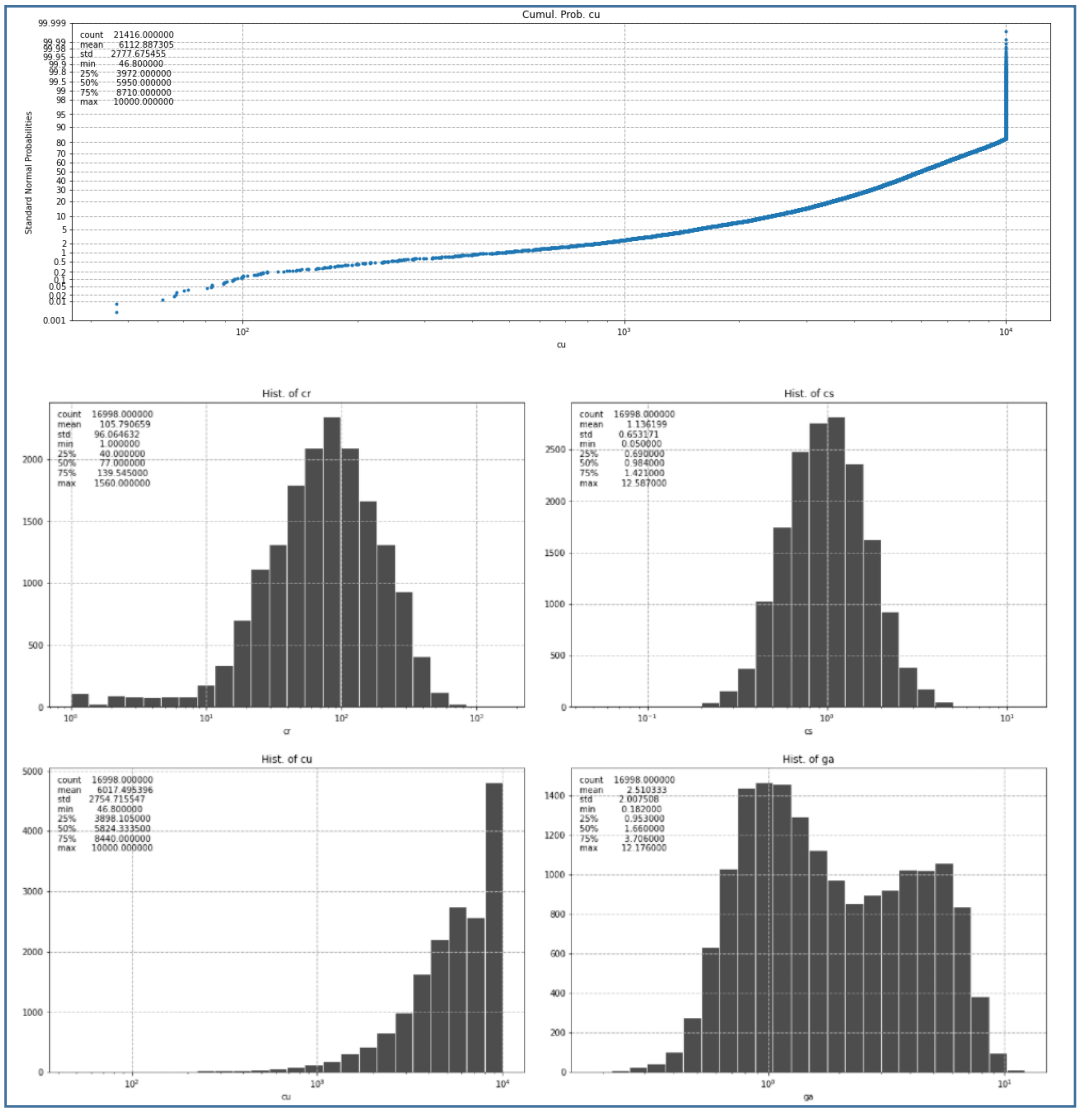

Figure 3 Probability plot of Cu values (top) and Histogram Cr, Cs, Ga and Cu distribution (below)

In the probability plot in figure 4, the variables show a clearer distinction of the populations in the alteration types, a characteristic which is not obvious in the other geological units. This suggests that the alteration is a huge factor when establishing domains and the distribution of most of the elements may vary by alteration.

Finally, the spatial distribution of the samples was also visualized in two dimensions as shown in figure 5, with preferential sampling on high-value areas, especially on its central portion, where it shows a north-east south-west trend of high values.

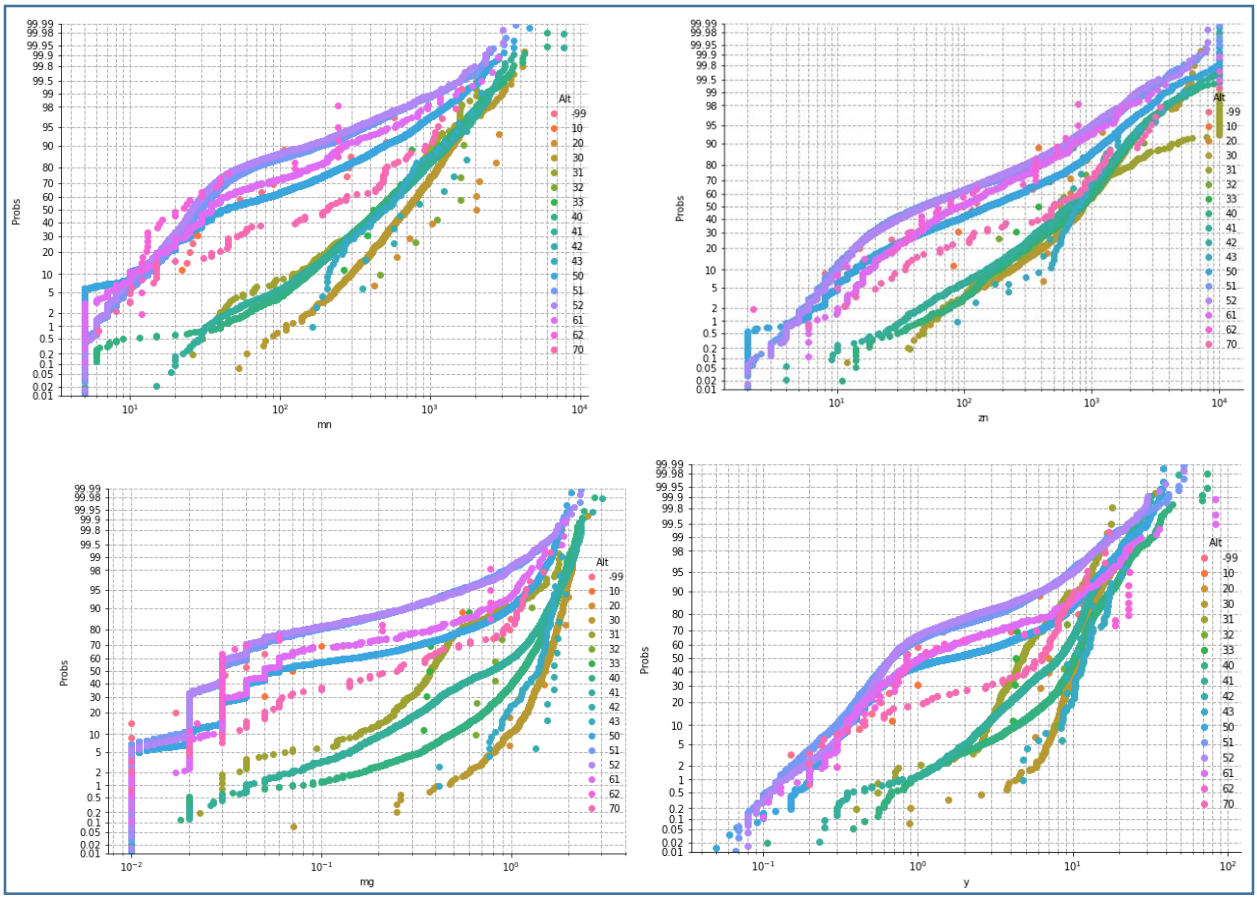

Figure 4 Distribution of Mn, Zn, Mg and Y by alteration. Two distinct populations can be observed

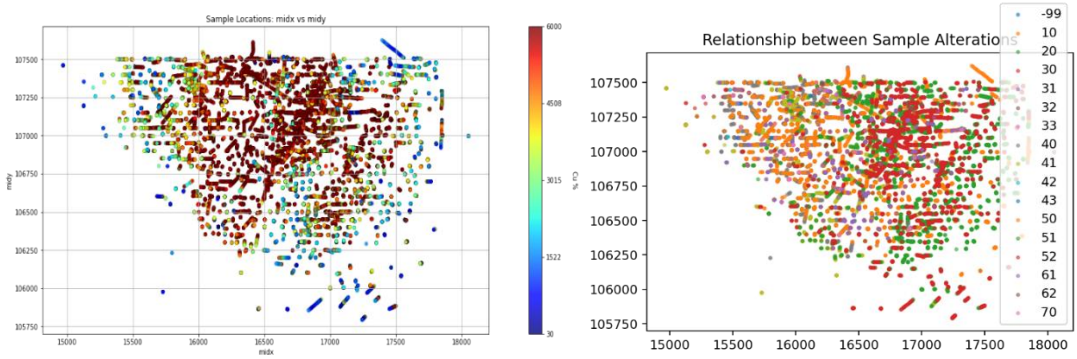

Figure 5 Location map of samples with copper concentration (left) and the various alterations units (right)

From the figures, alteration type 52 seems to be dominant alteration which host most of the high-grade copper values. Alteration type -99 is an unverified alteration and hence not regarded.

To understand the spatial correlation between elements, a correlation matrix was computed which showed poor correlation of copper with other elements in the matrix even though some trace elements of the deposit show some form of correlation with each other. Mg, Al, Ga shows a good correlation.

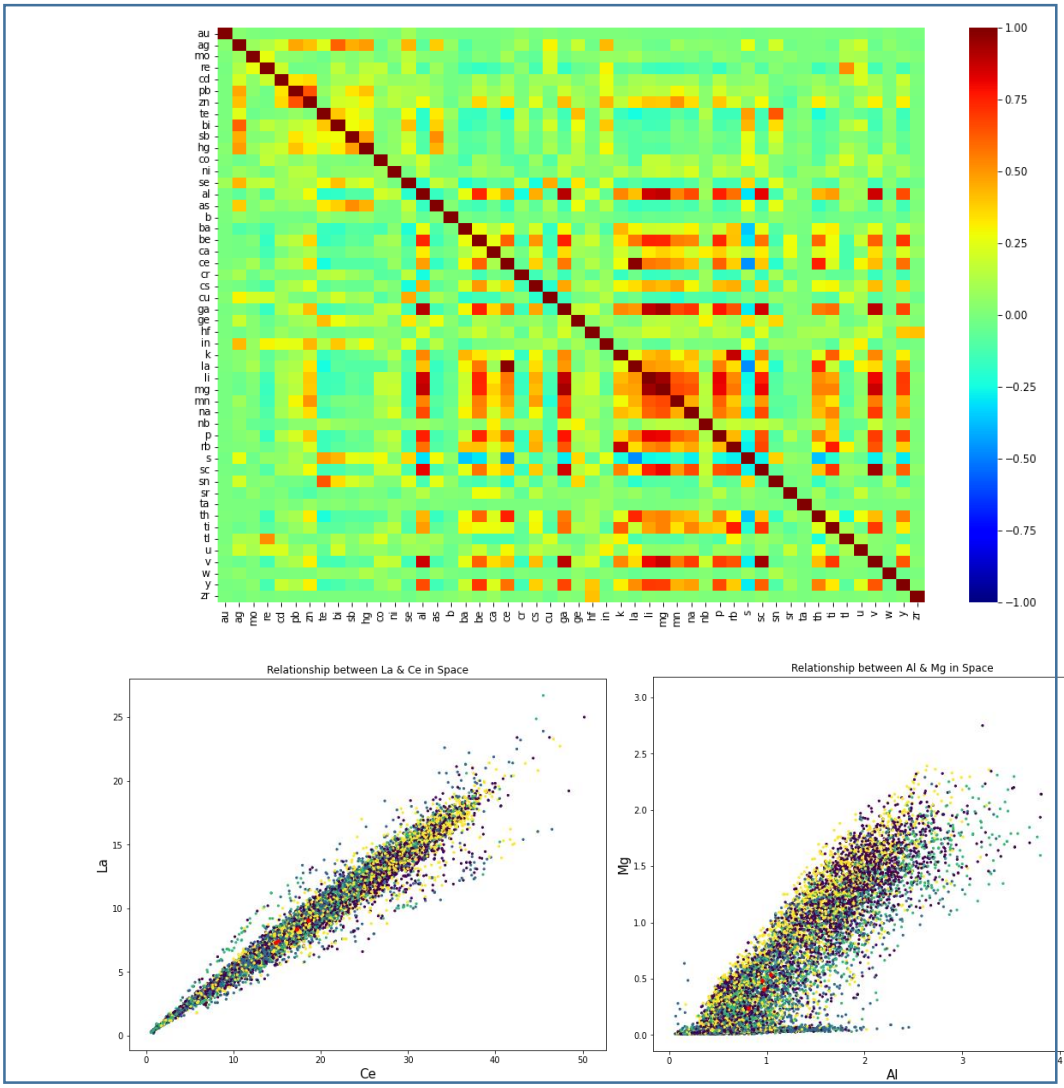

Figure 6 Correlation matrix of 50 continuous variables (top) and scatter plot showing correlation of La and Ce, and Al and Mg respectively. Color codes didn't matter at this point.

In summary, the structure of the data is well understood. Deep samples are removed and variables whose results are unaccounted for are redundant. We see how the samples are distributed in space.

In the next section, unsupervised classification using K means will be conducted. Most of the results will be shown in 3D.

3. Cluster analysis - K Means

The K-Means procedure is one of the most popular machine learning algorithms used in cluster analysis, due to its simplicity, interpretability and application to large amounts of data (Adams, 2018). It is most useful for creating a small number of clusters from many observations. Due to the large number of possible clusters that can be formed, the quality of the output is not guaranteed. The K means algorithm clusters data by separating observations into groups of equal variances, minimizing a phenomenon known as the *inertia* or within-cluster sum-of-squares as shown in the equation below (Davies & Bouldin, 1979).

Clustering was done using the web application Jupyter Notebook, with Python 3.6.5 installed via Anaconda; processor AMD Ryzen 5 3600 6-Core Processor 3.60 GHz, with 16.0GB RAM, Windows 10, 64 bit.

Two important factors that drove this analysis were the number of clusters to choose and the selection of variables to drive the clustering process. The performance of the clustering algorithm depends on the value of K. Therefore, we performed the well-known *elbow analysis* to determine the optimal number of clusters as well as a set of values for *k*. It is also important that the number of values considered should reflect the specific characteristics of the data sets which is the main motivation for performing data clustering (Pham et al., 2005).

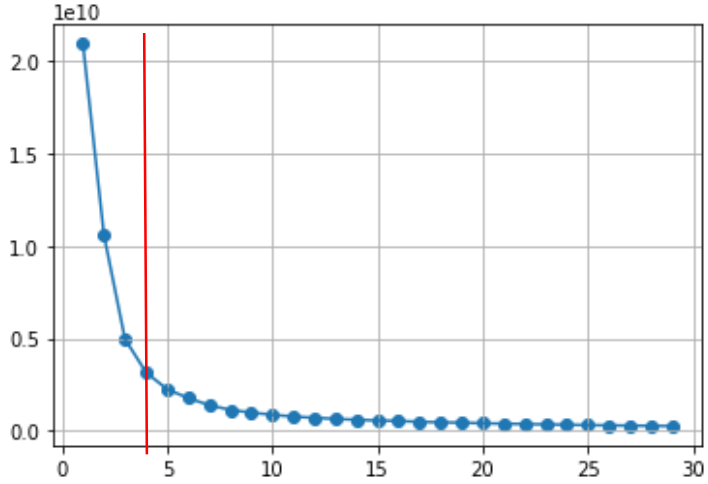

Figure 7 Determination of optimum cluster number using the elbow method. Optimum number was set at 4

The selection of variables from the geochemical data to inform the clustering process forms part of inputting domain knowledge to aid clustering since our objective is to reproduce the geology based on the geochemistry. Six elements associated with the porphyry copper deposit (Al, Ga, Mg, Li, Sc & V) were used. It is important to note that the type of domains formed is a factor of the variables selected (Faraj & Ortiz, 2021). Since most of the elements show significant changes with the alteration, we're expecting our domains to be more consistent with the alteration than the rest of the geology.

The following display shows a three-dimension clustered data of the selected variables with the elbow method optimal number of 4 as well as two other cluster number values (3 and 10).

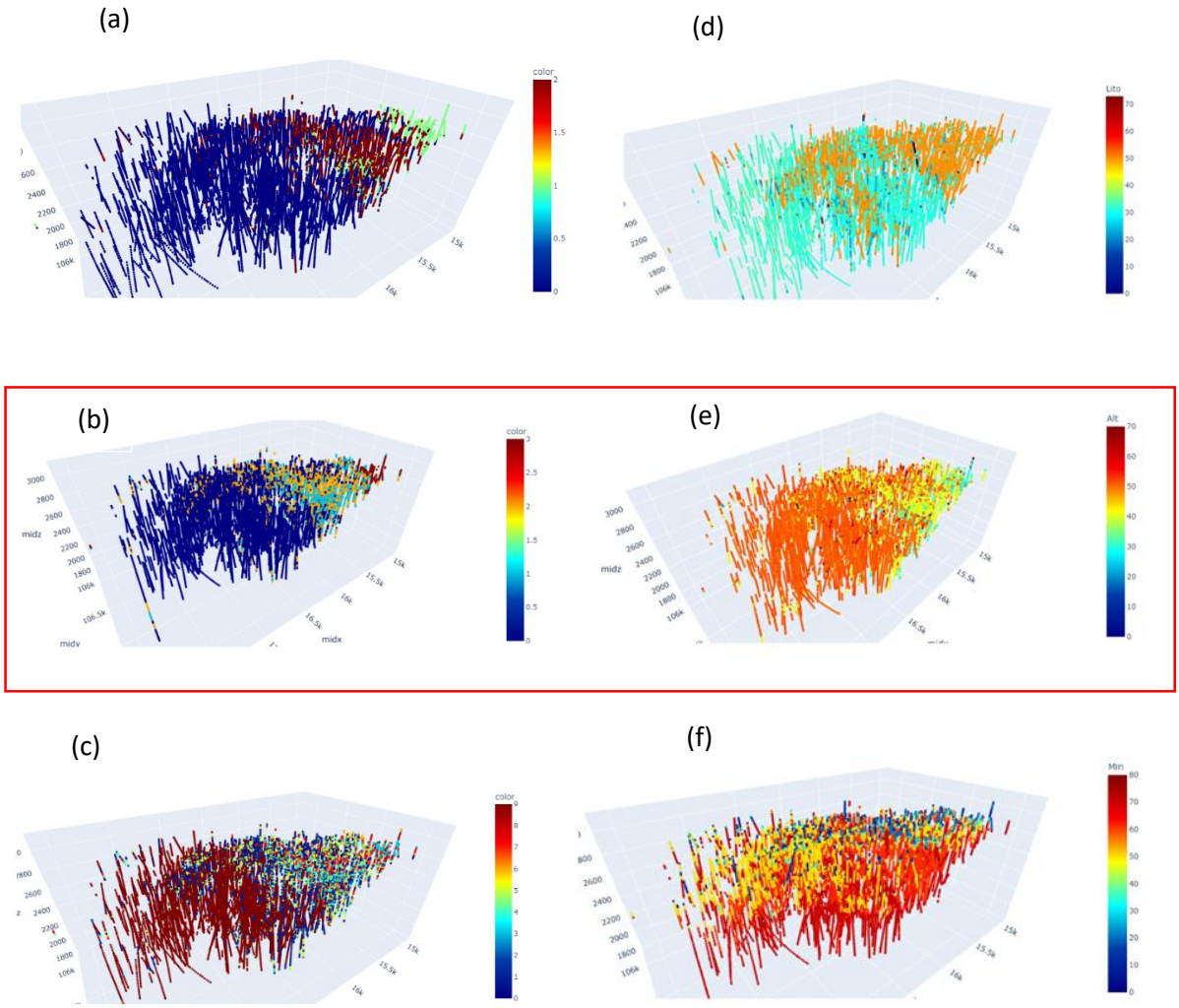

Figure 8 A 3D display of clustered drill hole data showing for cluster numbers 3, 4 and 10 represented as a, b, c respectively and the lithology, alteration and mineralization as d, e, f respectively.

It can be observed that all the three clustered data follow the alteration pattern better than either lithology or mineralization. Having established this fact, clustering with four clusters seems the better of the other cluster numbers in reproducing the alteration features. This validates the selection of the optimum number (4) by the elbow method. A cluster number of 10 was used because of its proximity to the total number of alteration features but it could not reproduce the alteration due to the larger volumes of the predominant features of the unit which overshadowed the fewer units.

Majority of the cluster labelled '0' (in blue colour) falls within alteration code 50 and 51. Clusters labelled 1 and 2 fall in alteration code 40 and 41 while the final cluster lies in the alteration 30. The K means algorithm assigns a cluster to a geological feature based on the highest number of clustered elements found in each feature. For example, if majority of the members in a particular cluster belongs to a particular lithology, K means predicts the cluster as that lithology.

4. Validation

To further validate the predicted clustered domains, a confusion matrix was computed. The confusion matrix shows the ways in which the algorithm is confused when it makes predictions, and highlights the errors made by the classifier. Since there are three geological categories present, this section validates the clustered data by these categories.

The major lithological groups present in the data were represented by codes 31 and 50 as seen in figure 9. K means prediction for the major groups attained an average accuracy of about 80% when compared to the logged lithology. The spatial distribution of the logged lithology showed that about 95% of the lithology with code 31 were rightly identified by the clustering algorithm. Rock code 50 which is the most abundant lithology was also rightly predicted. However, about 30% of lithology code 50 was misclassified as 30 by the algorithm as well as a few blocks of lithology code 30 was misclassified as 50. Due to the large number of lithologies present, misclassifications of the fewer groups are expected which leads to the decline in accuracy as it becomes trickier predicting delicate differences in units that are not largely represented or have significant similarities with the major groups.

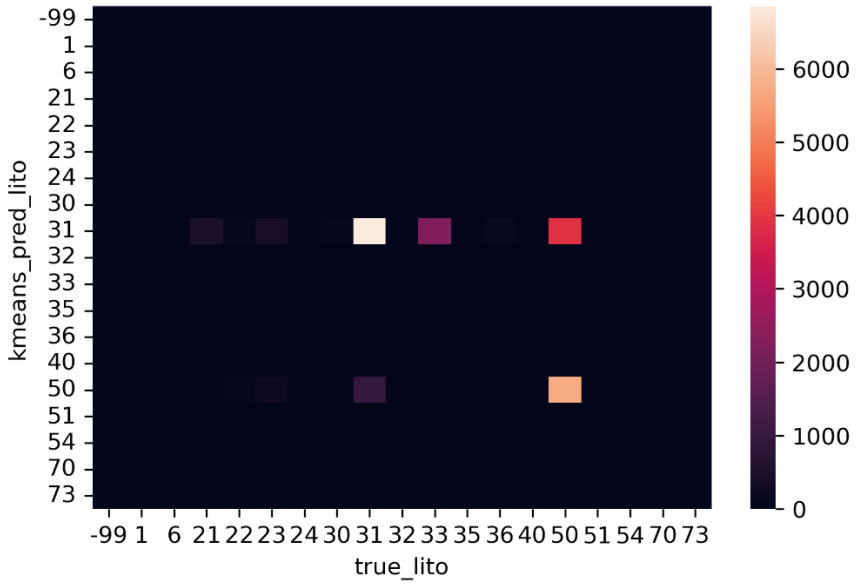

Figure 9 Confusion matrix of K means predicted units versus logged lithology

The logged alteration units contained five major units (40, 41, 50, 51, 52) with code 51 being the most abundant although the difference is not large. The algorithm predicted accurately for alteration codes 51 and 40. However, alterations that are close to code 51 are predicted as 51 as shown in figure 10. This may be due to close similarities in alterations which was evident in the grades. As captured in the probability plot in figure 5, the algorithm identifies two distinct groups of alterations. Similarly, misclassifications can be due to the insignificant representation of other alterations or due to their similarities.

The mineralogy exhibited a monopoly of predicted mineralization zones. The algorithm assigned every mineralogical zone to code 70. The algorithm performed poorly by not being able to predict the various mineralogical units present in the data. This shown in figure 11.

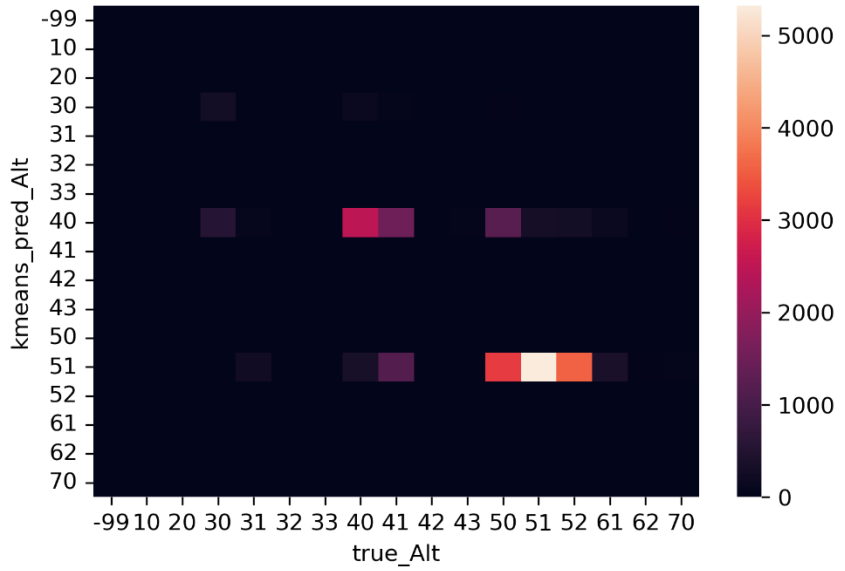

Figure 10 Confusion matrix of predicted alteration zones versus logged alteration

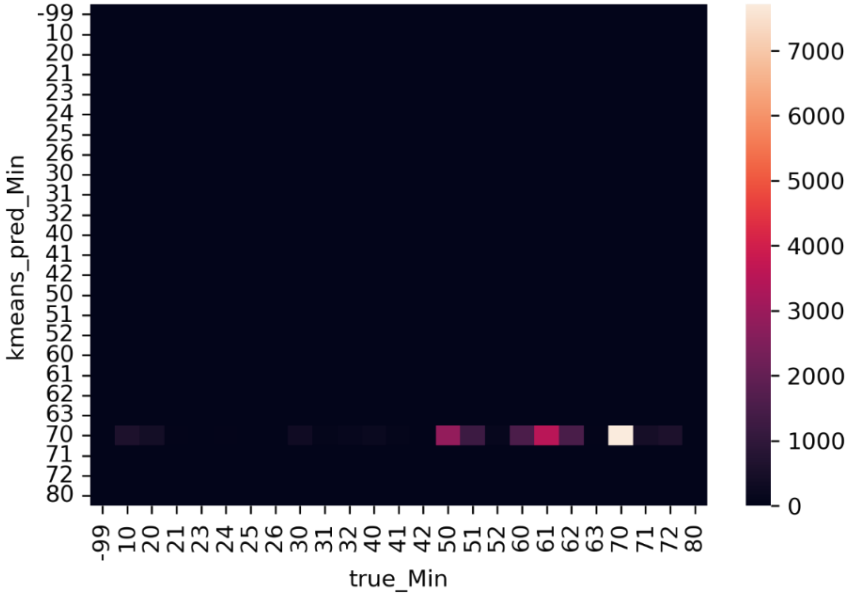

Figure 11 Confusion matrix of predicted mineralization zones versus actual mineralization

5. Conclusion

The results from this study show that although very effective and is one of the most used algorithms in machine learning, the sole application of the classical k-means is quite handicapped in geological modelling, despite the spatial contiguity it exhibits. Outliers present in the data were not properly captured which accounts for some of the misclassification.

According to the model, clusters delineated by the algorithm show some form of consistency with the logged alteration unit but forms an insignificant correlation with the rest of the categorical variables. The confusion matrix revealed the major flaws of the algorithm in its inability to classify units that are similar

or close to each other. Even though the clustered model showed a correlation with the alteration, the confusion matrix exposed its weakness in the number of misclassified units. The clustering resulted in a mixed-up population with the major units overlapping each other. The probability plot of the alteration unit indicates the presence of two major populations.

A more adequate approach is needed to account also for the geographic distribution of samples, which is done by some modern clustering techniques, such as the local autocorrelation-based clustering algorithm. The selection of variables that do not properly represent the difference in the geological units would lead to a poor discrimination by algorithm, reducing the accuracy of the model.

6. Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2017-04200 and RGPAS-2017-507956.

7. References

Adams, R. P. (2018). *K-Means Clustering and Related Algorithms*. Princeton University, 18 p.

Bosch, M., Zamora, M., & Utama, W. (2002). Lithology discrimination from physical rock properties. *Geophysics*, *67*(2), 573–581. https://doi.org/10.1190/1.1468618

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Emery, X., & Ortiz, J.M. (2005) Estimation of mineral resources using grade domains: critical analysis and a suggested methodology, Journal of the South African Institute of Mining and Metallurgy, 105(4): 247-255.

Faraj, F., & Ortiz, J. M. (2021). A Simple Unsupervised Classification Workflow for Defining Geological Domains Using Multivariate Data. *Mining, Metallurgy & Exploration*, *38*(3), 1609–1623. https://doi.org/10.1007/s42461-021-00428-5

Moreira, G. de C., Modena, R. C. C., Costa, J. F. C. L., & Marques, D. M. (2021). A workflow for defining geological domains using machine learning and geostatistics. *Tecnologia Em Metalurgia, Materiais e Mineração*, *18*, e2472. https://doi.org/10.4322/2176-1523.20212472

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of *K* in *K* -means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, *219*(1), 103–119. https://doi.org/10.1243/095440605X8298

Sterk, R., de Jong, K., Partington, G., Kerkvliet, S., & van de Ven, M. (2019). *Domaining in Mineral Resource Estimation: A Stock-Take of 2019 Common Practice*. Mining Geology 2019 Conference, Perth 25-26 November, 2019, 13 p.

Yasrebi, A. B., Afzal, P., Wetherelt, A., Foster, P., & Esfahanipour, R. (2013). Correlation between geology and concentration-volume fractal models: Significance for Cu and Mo mineralized zones separation in the Kahang porphyry deposit (Central Iran). *Geologica Carpathica*, *64*(2), 153–163. https://doi.org/10.2478/geoca-2013-0011