A workflow for defining geological domains using machine learning and geostatistics

Gabriel de Castro Moreira [1]*
Rudi César Comiotto Modena [1]
João Felipe Coimbra Leite Costa [1]
Diego Machado Marques [2]

Abstract

Determining geological domains to be modeled is one of the first steps in the mineral resource evaluation process. Prior knowledge regarding the geology of the deposit is fundamental but, in most cases, not enough for a reasonable definition of these domains. A careful statistical analysis of the available data (e.g. geochemical samples) is also of great importance. In order to avoid mixing different populations of data, samples with similar characteristics should be grouped together. In the context of supervised machine learning, cluster analysis can be especially suited for this matter and there are many different algorithms available in the literature. In this paper, two clustering techniques were investigated: the first is the k-means algorithm, one of the most widely used methods in machine learning, based on the iterative analysis of the statistical distribution, while the other one is based on spatial autocorrelation statistics, which takes into consideration the geographic distribution of samples. The choice of the most appropriate technique, as well as the number of domains can be challenging when performing cluster analysis, and the evaluation of an expert is still necessary, as the results are subjective.

Keywords: Cluster analysis; Geostatistics; Mineral resources; Mining.

1 Introduction

The modeling of spatially correlated variables results from the association between the natural component of the earth sciences and the foundations of mathematics and statistics, in particular the theory of random functions (RF) [1]. Geostatistics [2] involves a set of concepts and techniques aimed at characterizing and modeling spatial phenomena through the analysis of dispersion of the so called 'regionalized variables', and also evaluating the uncertainties of random functions that describe these phenomena on a mathematical perspective.

Defining estimation domains is one of the first tasks when modeling mineral resources, and one of the most important decisions to be made in the entire workflow. A poor definition of these domains can lead to the mixing of populations, which can lead to bad resource estimates, compromising the valuation of grades and tonnages.

A good comprehension of data statistics, along with the geology, provides a better understanding of the deposit, allowing its subdivision into domains for modeling, which is more plausible than dealing with the whole deposit as one individual entity. Multivariate techniques can be valuable when dealing with this problem. In the field of unsupervised machine learning, cluster analysis can be especially suited for this matter, since it separates a set of N samples based on the relationships between the M available variables.

Cluster analysis can play an important role not only in the Earth sciences, but in a variety of topics: social sciences, biology, statistics, data mining, among others. Tan et al. [3] define cluster analysis as a process that groups data based only on their characteristics and relationships. The objective is that objects in each cluster are similar to each other and, at the same time, different from the objects that belong to other clusters. The greater the similarity (or homogeneity) within a group, and the difference between groups, the more efficient is the clustering process. In exploratory analysis, it can be very useful to understand how data can be grouped. For example, to define typologies and estimation domains in a mineral deposit.

Clustering algorithms have been around since the 1960's, when Sokal and Sneath [4] presented the agglomerative hierarchical technique for working in the field of taxonomy, and MacQueen [5] introduced the k-means algorithm.

However, the application of traditional clustering algorithms to geological datasets is quite limited, as these techniques are often used to characterize the relationships based only on statistical parameters, not taking into consideration geological aspects [6].

[1]*Departamento de Engenharia de Minas, Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brasil.*
[2]*Departamento de Geologia, Instituto de Geociências – IGEO, Universidade Federal do Rio Grande do Sul – UFRGS, Porto Alegre, RS, Brasil.*

**Corresponding author: gabrielcm.moreira@gmail.com*

In the last decades, clustering techniques have been applied to spatially-related samples (e.g. Scrucca [7], Romary et al. [8], Fouedjio [9], Martin and Boisvert [10]) with the objective of developing algorithms that take into consideration not only the relationships of samples in the multivariate space, but also in the geographic space, generating clusters that are spatially contiguous and show distinct multivariate properties.

In the context of geology and geostatistics, the geographic contiguity of the clusters and the multivariate delineation are both important factors for the resulting domains to be modeled [11]. According to Martin and Boisvert [10], two criteria should be considered when measuring the "goodness" of the resulting clusters:

i. The contiguity of the domains in the Cartesian (geographic) space;

ii. The separation of populations in the multivariate space.

Although these modern algorithms can divide the data into reasonable groups, both in the geographic and multivariate space, the choice of the optimal number of domains and their validation are still subjective, as already stated by Martin [11]. This article addresses this matter, applying and further discussing some of the methods that can be applied.

Our main objective is to apply two clustering algorithms and discuss the difficulties found when choosing the best configuration of the clusters. As another important contribution, we apply and discuss a method to validate the spatial distribution of the clusters based on correlograms of the indicators. Formal methods for validating the spatial connectivity of the groups are, actually, rarely mentioned in the literature, other than just applying a visual inspection of the results.

2 Methodology

In this study, two clustering algorithms were applied: k-means [5], which is one of the most widely used techniques in machine learning, and the autocorrelation based spatial clustering algorithm, herein mentioned as "acclus", from Scrucca [7], specially designed for dealing with spatial data.

All clustering methods were applied using the web application Jupyter Notebook, with Python 3.6.5 installed via Anaconda; processor Intel® i7-3.20Ghz, with 24.0GB RAM, Windows 10, 64 bit.

For the execution of k-means, the available algorithm on the scikit-learn library [12] was used, with the "k-means ++" option as a centroid initialization parameter [13], which seeks to maximize the separation between the centroids, increasing accuracy and speed.

Based on concepts introduced by Ord and Getis [14] for univariate cases, Scrucca [7] applied local autocorrelation statistics to generate spatially interconnected multivariate clusters in a two-step method. First, a new database is generated with local autocorrelation measures for each variable, given the direct relationships with data in the vicinity. Then the traditional k-means algorithm is applied to this new database, so that clusters that have both spatial and statistical coherence are defined. The algorithm is available on GitHub, hosted in the account mentioned in Martin and Boisvert [10].

The techniques were tested on the 2-dimensional isotopic Walker Lake dataset [15] with 470 samples, each containing values for two continuous variables, V and U. Figure 1 shows the spatial distribution of samples, with preferential sampling on high-value areas, especially on its western portion, where it shows a north-south trend of high values.

For clustering, both variables were used and, as the order of magnitude matters, and can influence the results, they were standardized, according to Equation 1. This way they will be on the same basis, with mean equal to zero and standard deviation equal to one. These transformations are typically performed when applying machine learning algorithms.

$$Z = \frac{(X - m)}{s} \tag{1}$$

where Z is the standardized value, X is the original value, m is the mean and s, the standard deviation.

These variable distributions are quite different, as shown in the histograms of Figure 2, with U showing a considerable higher asymmetry than V. The scatter plot in the same figure shows that the variables present positive correlation, although not high.

The Q-Q plot is meant to compare the distributions of different variables by plotting their corresponding quantiles. Figure 3 shows the Q-Q plot of variable V against variable U, evidencing, one more time, the considerable difference in variable distributions.

To evaluate the clustering configurations, and choose the most adequate technique, the following validation methods were applied:

I. Visual inspection of the geographic distribution of samples and scatter plots of the continuous variables, colored by clustering code;

II. The dual space metrics [10], which accounts for the geographic entropy and within clusters sum of squares;

III. Inspection of geographic contiguity of the clusters, using the indicators correlograms;

IV. Contact analysis, which is meant to evaluate the behaviour of a variable having the contacts between different domains as references.

The measure of the geographic distribution of a given configuration of clusters in geographic space, the "spatial

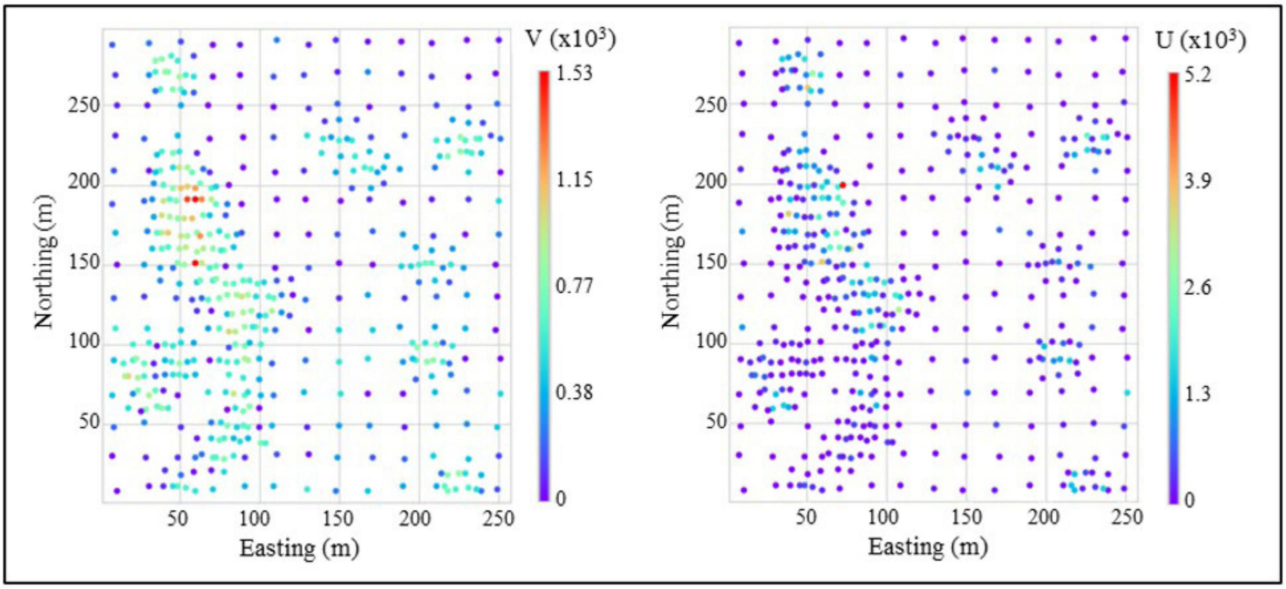

Figure 1. Sample locations and values for each variable present in the Walker Lake dataset.

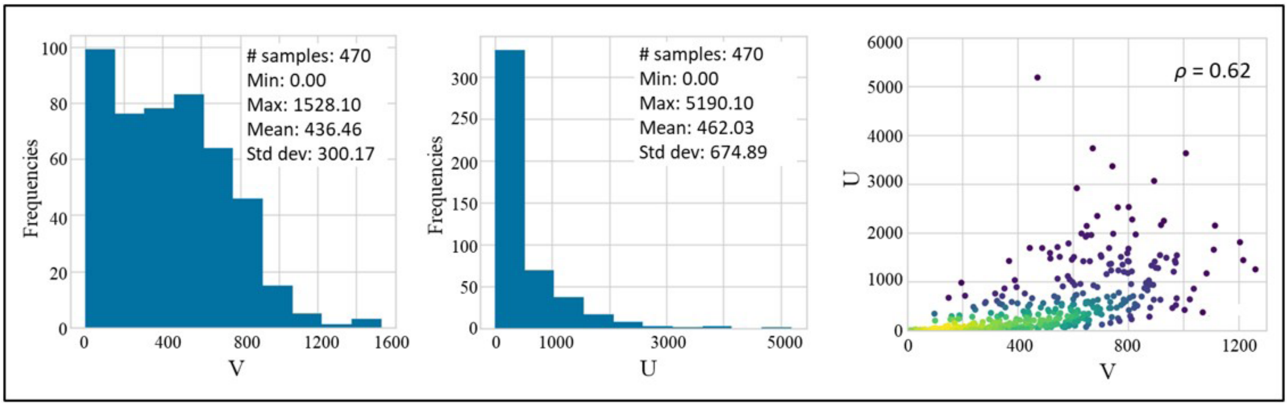

Figure 2. Histograms of the variables V and U and scatter plot V × U, colored by a kernel density estimator.

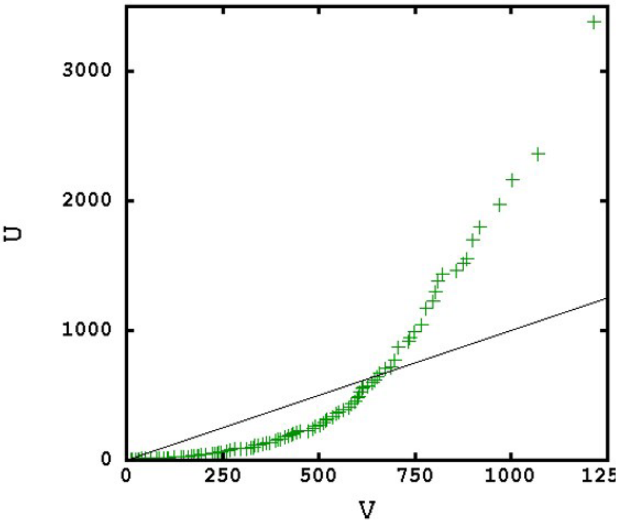

Figure 3. Q-Q plot showing the distributions of variables V and U plotted against each other.

entropy” (H), is calculated using a local window at each location and summing it over all locations, according to Equation 2.

$$H_{total} = -\sum_{i=1}^{N}\sum_{k=1}^{K} p_{i,k} \ln p_{i,k} \qquad (2)$$

where $p_{i,k}$ is the probability of finding category k in the local search around the i^{th} location. Lower values of entropy indicate higher spatial organization, in other words, greater geographic cohesion of clusters.

As for measuring how well each population is separated from one another in the multivariate space, Equation 3 expresses the “within clusters sum of squares” (WCSS):

$$wcss = \sum_{k=1}^{K}\sum_{x_i \hat{o} K_k}\sum_{j=1}^{M}\left(x\mathrm{ij} - \bar{x}\mathrm{kj}\right)^2 \qquad (3)$$

where $(x_{ij} - \bar{x}_{kj})$ represents the distance between a given sample and its respective cluster centroid. Lower WCSS values indicate configurations with more compact groups in the multivariate space, as distances between elements within each group are smaller.

Each metric is calculated independently for a given clustering configuration, but both metrics must be evaluated simultaneously to assess spatial clustering [10], which can be done by plotting spatial entropy versus WCSS.

Although an important aspect in the clustering of spatial data, the formal validation of the geographic distribution of the clusters is not typically discussed in the literature and here we present a method for this verification by measuring the spatial continuity of the clusters (by their indicators correlograms). The semivariogram [2] is the standard instrument for that purpose, as applied by Modena et al. [16]. However, it can be affected by short-distance noise. Thus, in this study, we applied the correlogram [17,18], as it is standardized and more robust than the semivariogram. A binary variable (the indicator) has to be defined, which assumes value 1 for samples within a given domain and 0 for the others. The correlogram of this binary variable was then plotted for various lag separations, which was done with the commercial software Isatis®. Structured and continuous correlograms indicate spatial contiguity of clusters, while fragmented clusters results in noisy correlograms, with high nugget effects, as will be demonstrated in the illustration case that follows.

It was conducted a contact analysis between the defined domains. It displays the mean value of a variable in each domain as a function of the distance of the samples to the contact with another domain.

3 Results and discussions

The *k-means* and acclus algorithms were applied in five different scenarios, each one corresponding to a different number of clusters, from two to six. Figures 4 and 5 show the results for each clustering algorithm: the location maps of the samples colored by cluster code, as well as the scatter plots of V x U with points also colored according to the

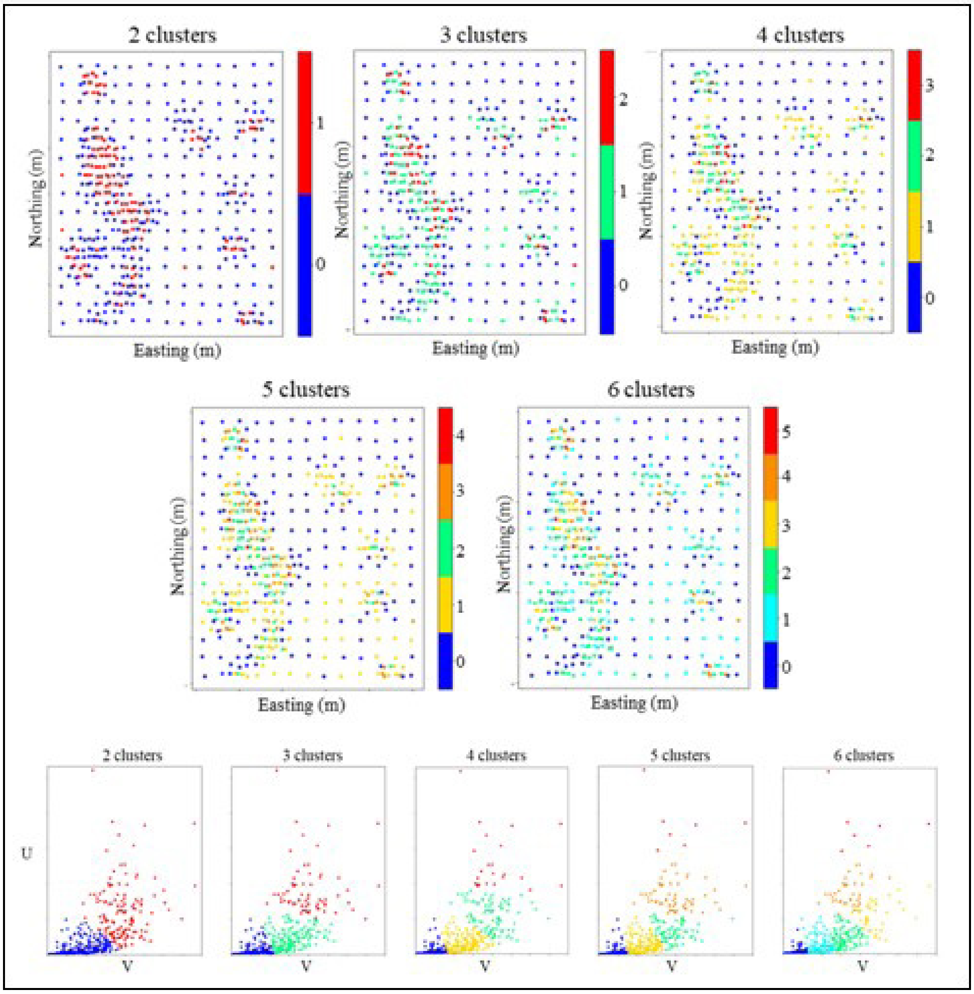

Figure 4. Results of the k-means clustering, each map corresponding to a different clustering configuration (different number of clusters, k). At the lower part, the scatter plots of the continuous variables, V and U, plotted against each other, also colored by cluster code.

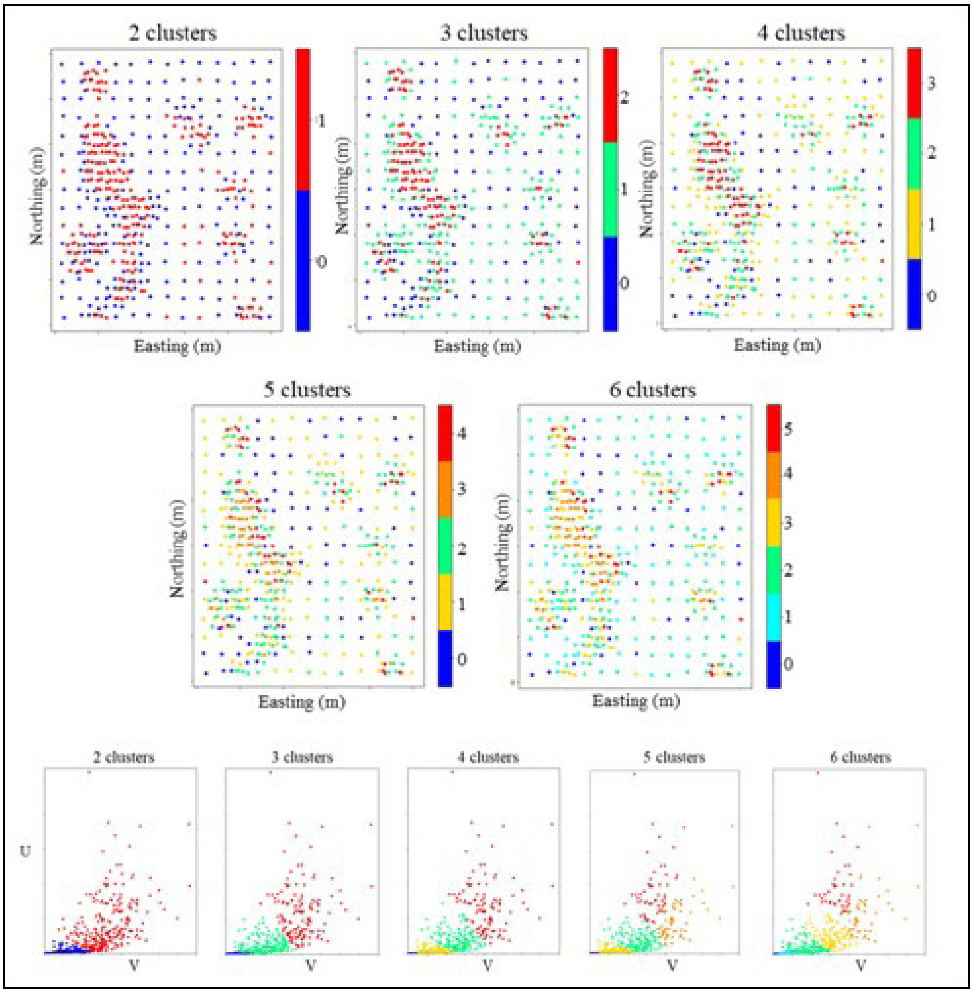

Figure 5. Results of the acclus algorithm, each map corresponding to a different clustering configuration (different number of clusters, k). At the lower part, the scatter plots of the continuous variables, V and U, plotted against each other, also colored by cluster code.

clustering codes. It is noticeable how the acclus provides more geographically contiguous clusters when compared to k-means but, when k = 4 or greater, similarly to k-means, acclus also results in fragmented clusters. Through the scatter plots it becomes apparent that k-means produces more uniformly distributed clusters in the multivariate space.

For verifying spatial connectivity, as well as multivariate cohesion, simultaneously, the dual space metrics of Martin and Boisvert [10] were applied to the results of both algorithms, to each scenario. Results can be observed in Figure 6A. Note that the WCSS is consistently lower for k-means, which is expected, as this clustering algorithm finds the best configuration for clusters in the multivariate space only. Conversely, the spatial entropy is lower for the acclus, as it accounts for the spatial distribution of the clusters.

There is an inflection point for WCSS when k = 3 for both algorithms, especially for acclus, meaning that there is no significant change in the within cluster variance from that point on.

For a given configuration to be considered acceptable, low values of both WCSS and spatial entropy are desirable. However, these two metrics are inversely proportional, as exposed in Figure 6B, that is, greater cohesion in the multivariate space implies in geographical fragmentation of clusters.

Thus, considering the dual space metrics, for a given level of multivariate ordering (WCSS), a clustering with the lowest spatial entropy is preferred; therefore, the acclus is a better choice over k-means. Furthermore, the inflection point when k = 3 suggests that this could be an appropriate choice.

Also for validating the choice of k we used the measurement of spatial continuity of the clusters through the omnidirectional correlograms of the indicators of the acclus codes, which are presented in Figures 7 and 8, for k = 2 and k = 3, respectively.

Both correlograms in Figure 7, related to domains 0 and 1 respectively, show "good" structure, attesting an acceptable geographic continuity of the clusters. It is noticeable that those correlograms are identical, as they are complementary, that is, the clusters are equivalent.

Figure 8 shows the correlograms of the indicators for domains 0, 1 and 2 in the three-domain scenario and, as can be noted, the correlograms of domains 0 and 3 are not satisfactory, given that they present considerably high nugget effects. Thus, although this is the suggested

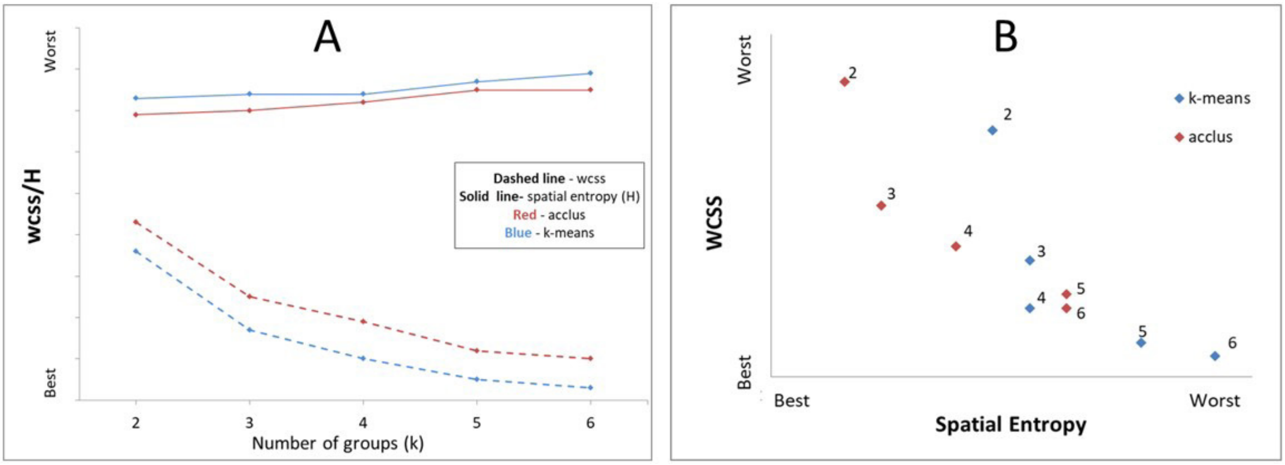

Figure 6. (A) The variations of the scores of the dual space metrics: within cluster sum of squares (WCSS) and spatial entropy (H), as the number of clusters (k) changes; (B) The same metrics plotted against each other for each clustering configuration; the numbers indicate the corresponding k.

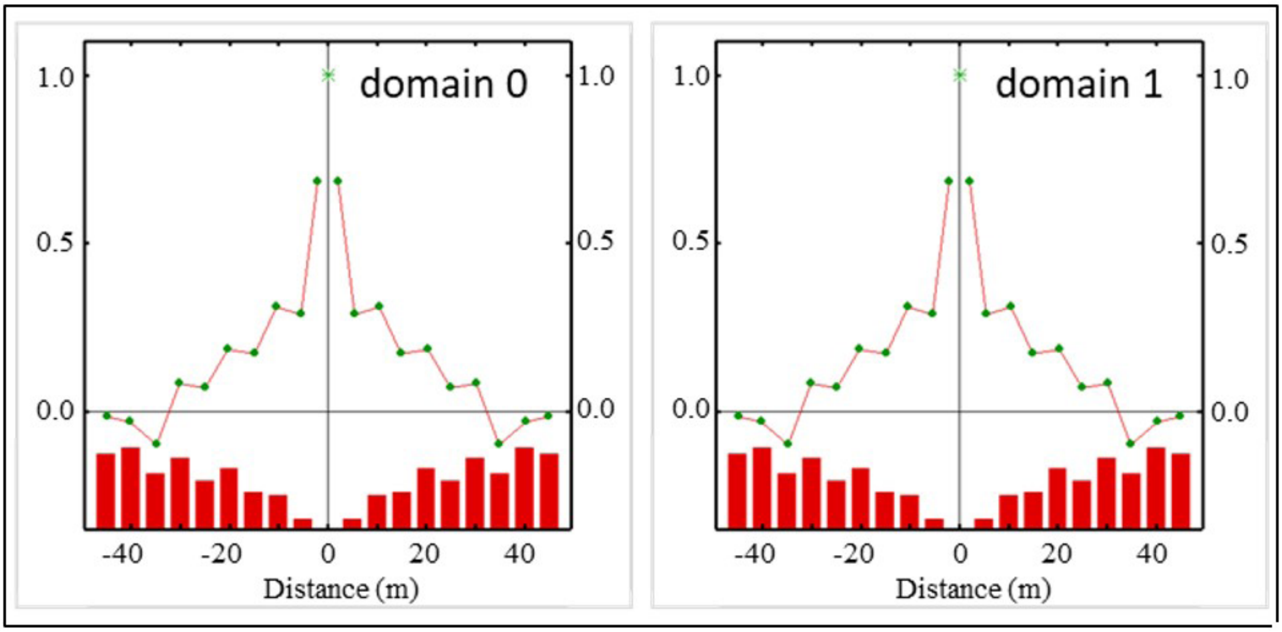

Figure 7. Omnidirectional correlograms of domains 0 and 1, in the two domains scenario.

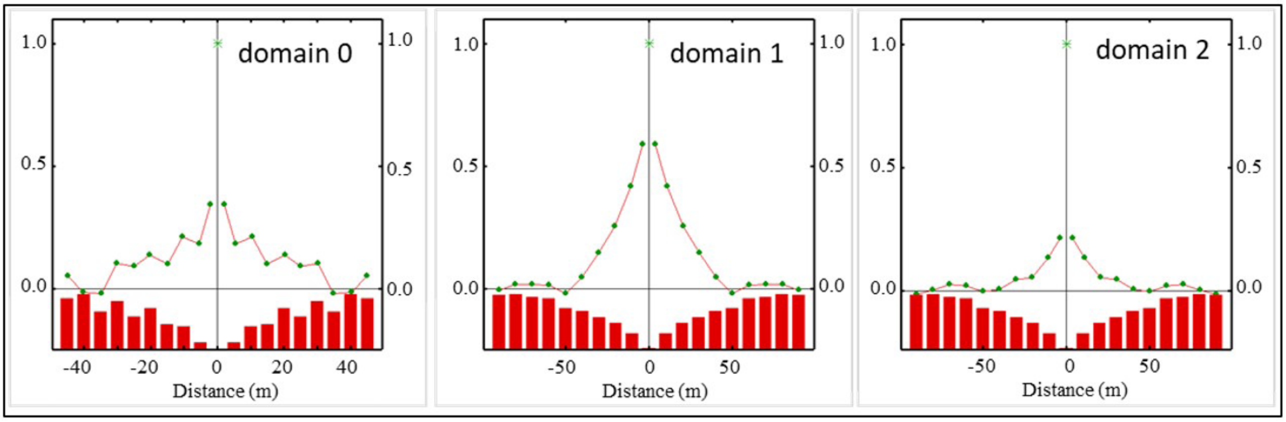

Figure 8. Omnidirectional correlograms of domains 0, 1 and 2, in the three domains scenario.

configuration as the most appropriate by statistical analysis, data segmentation into three domains does not present suitable spatial connectivity, but rather geographically fragmented clusters, as seen in Figure 9, which shows the location maps for the two- and three-domain scenarios, including the contours of these domains, manually drawn. Note that, in the three-domain scenario, there are many samples that belong to a particular cluster located within another cluster, unlike the other case, with very few samples in this condition. Therefore, the two-domain scenario seems to be the most appropriate in this case. The boxplots from Figure 10 show the statistical distributions in each domain.

Lastly, the contact analysis between the domains in the two-domain scenario was performed and, as it can be observed in Figure 11, there is a considerable difference of the mean values from samples separated by a certain distance (represented by the red lines) from the contact, for both variables. Note that there are samples almost up to 100m from the contact in domain 1 and above 110m for domain 0, which indicates that there is considerable geographic connectivity within each of these domains.

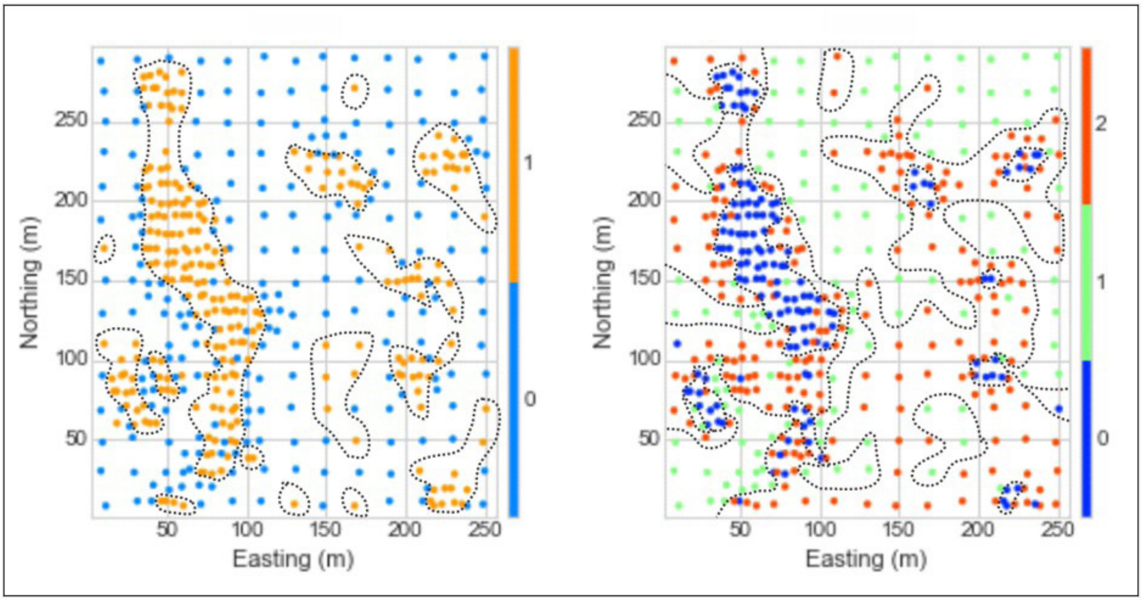

Figure 9. Location maps of the configurations resulting from the acclus algorithm for the two (left) and three (right) domains scenarios, including domain contours, manually drawn. Note that clustering into three domains results in scattered groups.

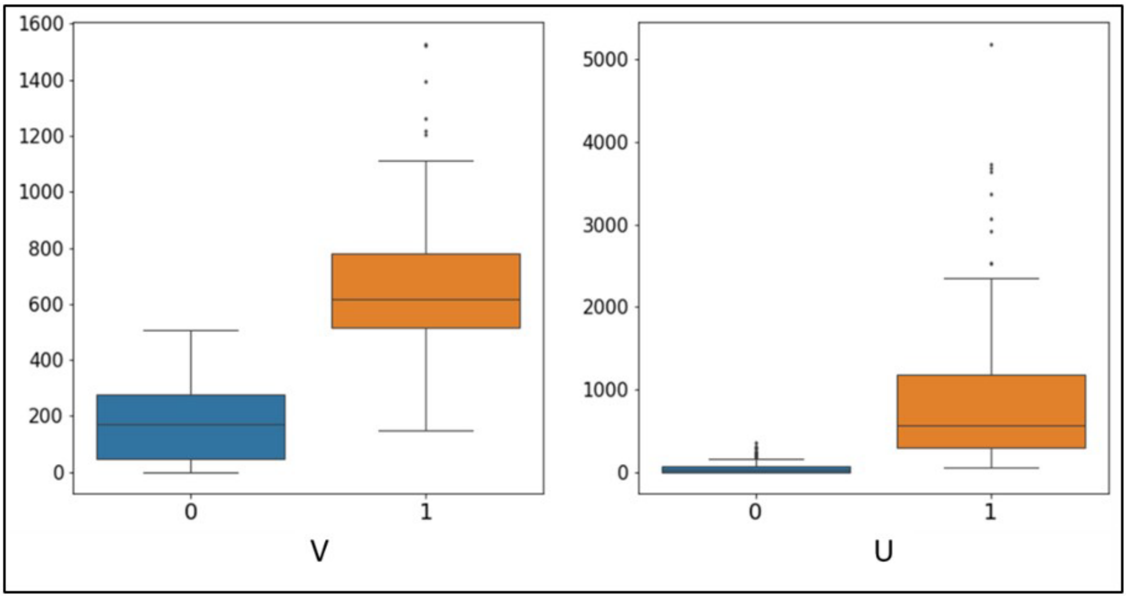

Figure 10. Boxplots showing the statistical distributions of V and U in each domain.

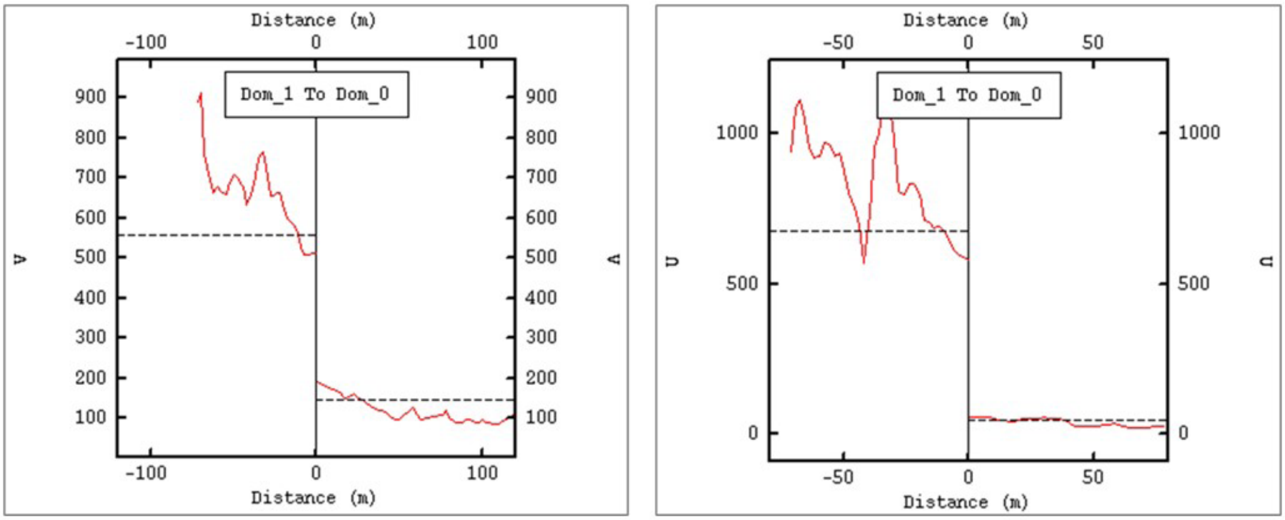

Figure 11. Contact analysis between domains 0 and 1 for variables V (left) and U (right). The red lines show the variation of the mean values from samples separated by a certain distance from the contact, which is represented by the black vertical line at the center of each plot. The dashed horizontal lines represent the overall mean values of each domain.

4 Conclusions

Although very effective and one of the most used algorithms in machine learning, the application of the traditional k-means is quite limited in geological modeling, as it only considers data distribution in the multivariate space, which can produce geographically fragmented clusters. A more adequate approach is 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 applied in this study. Results of the case study described herein corroborate with this statement.

Defining the number of groups and validating the results are not trivial, and should be done with great care, so that there is no mixing of statistical populations and that there are not too many clusters, which can result in geographically fragmented domains. This would naturally lead to unnecessary complications in the subsequent steps for the resource modeling workflow, such as contour modeling, estimation and simulation.

The illustration case is a satisfactory example for the application of the suggested methodology, which can be used in virtually any context, including complex mineral deposits with many variables. A full geological model can be built after defining the contours around the designated clusters (as illustrated in Figure 9). Variographic modeling of the variables to be estimated follows, and then, the grade-estimation, usually done by ordinary kriging (other interpolation methods can be applied, as well as geostatistical simulation). This full-modeling workflow is intended to be addressed in a future work.

The applied metrics, along with the proposed method of using indicator correlograms for validating the spatial distribution of clusters are good techniques when working with cluster analysis for resource modeling. Yet, expert knowledge and evaluation are still necessary on these rather subjective tasks, which still require parameterization and validation. We believe that, with the advances on machine learning algorithms and their applications on mineral resource modeling, these interventions will be less important with time.

Acknowledgements

The authors would like to thank the Mineral Exploration and Mining Planning Laboratory (LPM) at the Federal University of Rio Grande do Sul (UFRGS), for providing the necessary conditions for the development of this work. Luiz Englert Foundation (FLE), the Coordination for the Improvement of Higher Education Personnel (Capes) and the National Council for Scientific and Technological Development (CNPq) are acknowledged for their financial support.

References

1 Soares A. Geoestatística para as ciências da terra e do ambiente. Lisboa: Instituto Superior Técnico; 2000.

2 Matheron G. Principles of geostatistics. Economic Geology and the Bulletin of the Society of Economic Geologists. 1963;58:1246-1266.

3 Tan P-N, Steinbach M, Kumar V. Introduction do data mining. Boston: Pearson Education; 2005.

4 Sokal RR, Sneath PHA. Principles of numerical taxonomy. Journal of Mammalogy. 1965;46:111-112.

5 MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probabability; 1967; Berkeley. Berkeley: University of California Press; 1967. p. 281-296.

6 Rossi ME, Deutsch CV. Mineral resource estimation. Dordrecht: Springer Science & Business Media; 2014.

7 Scrucca L. Clustering multivariate spatial data based on local measures of spatial autocorrelation. Quaderni del Dipartimento di Economia. Finanza e Statica. 2005;20:1-25.

8 Romary T, Ors F, Rivoirard J, Deraisme J. Unsupervised classification of multivariate geostatistical data: two algorithms. Computers & Geosciences. 2015;85:96-103.

9 Fouedjio F. A hierarchical clustering method for multivariate geostatistical data. Spatial Statistics. 2016;18:333-351.

10 Martin R, Boisvert J. Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics. Computers & Geosciences. 2018;120:82-96.

11 Martin R. Data driven decisions of stationarity for improved numerical modeling in geological environments [thesis]. Edmonton: University of Alberta; 2019.

12 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825-2830.

13 Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. In: Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms; 2007 January 7-9; New Orleans, Louisiana, USA. New Orleans: ACM-SIAM; 2007. p. 1027-1035.

14 Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis. 1995;27:286-306.

15 Isaaks EH, Srivastava RM. An introduction to applied geostatistics. New York: Oxford University Press; 1989.

16 Modena RCC, Moreira GC, Marques DM, Costa JFCL. Avaliação de técnicas de agrupamento para definição de domínios estacionários com o auxílio de geoestatística. In: Associação Brasileira de Metalurgia, Materiais e Mineração. Proceedings of the 20th Mining Symposium - International; 2019 October 1-3; São Paulo, Brazil. São Paulo: ABM; 2019. p. 91-100.

17 Srivastava RM, Parker HM. Robust measures of spatial continuity. In: Proceedings of the Third International Geostatistics Congress; 1988 September 5-9; Avignon, France. Avignon: Springer; 1988. p. 295-308.

18 Moreira GC. Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais [thesis]. Porto Alegre: Universidade Federal do Rio Grande do Sul; 2020.

Received: 20 Aug 2020
Accepted: 21 Nov. 2020