Evaluation of PCA with variable selection for cluster typological domains

http://dx.doi.org/10.1590/0370-44672023770071

Silvânia Alves Braga de Castro[1,2,4]
https://orcid.org/0000-0002-1343-660X
André Carlos Silva[3,5]
https://orcid.org/0000-0002-9760-0728

[1]Universidade Federal de Catalão - UFCAT, Catalão – Goiás - Brasil.

[2]Centro Federal de Educação Tecnológica de Minas Gerais - CEFET-MG, Departamento de Mineração, Departamento de Minas e Construção Civil - DMCAX , Araxá - Minas Gerais - Brasil.

[3]Universidade Federal de Catalão – UFCAT, Laboratório de Modelamento e Pesquisa em Processamento Mineral - LaMPPMin, Catalão – Goiás - Brasil.

E-mails: [4]silvaniabraga@cefetmg.br, [5]ancarsil@ufcat.edu.br

Abstract

The modeling of mineral deposits has been improved over the years with the incorporation of mineralogical and metallurgical information obtained from drilling samples that make up the pillars for the construction of resource models. However, sampling data is being made available in large quantities, causing current databases to grow exponentially. The use of machine learning (ML) algorithms has been applied to deal with multidimensional data problems. Principal component analysis (PCA) is a multivariate analysis (MA) technique whose aim is to reduce the dimension of multivariate data. Studies show that results obtained with the reduction of variables were satisfactory in different areas of activity. The purpose of this article is to test variable selection criteria using PCA for geometallurgical data and to check the feasibility of the technique for simplifying variable types and defining typological domains.

Keywords: multivariate analysis, variable selection, geometallurgy.

1. Introduction

Evaluation of process performance within mining operations requires the geostatistical modeling of many related variables. These variables are a combination of grades and other rock properties, which together provide a characterization of the deposit that is necessary for optimizing plant design, blending and stockpile planning (Barnett, *et al*., 2012).

The process of deposit spatial modelling is usually carried out after sampling and identification of homogeneous geological bodies (Braga, 2016). These are known in geostatistics as stationary domains. However, the identification of these domains is not always an easy task because the information available in geoscience databases requires the analysis of a great number of attributes, which may contribute little for the discrimination of the evaluated individual elements, in addition to causing the analysis and data interpretation to become more complex. In general, a small amount of these variables contains the most relevant information, whereas the others do not add much to the interpretation of the results.

Many methods are possible for deciding which variables to reject, but in practice experience and intuition often play a role in the selection (Jolliffe, 1972). According to Milligan (1980), a poor choice of variables for grouping domains leads to inaccurate assignments of observations and the formation of clusters.

The use of machine learning (ML) algorithms is an interesting alternative for multidimensional data problems in deposit modeling. Different approaches have been proposed to select the most discriminating variables, such as Principal Component Analysis, Kernel Principal Component Analysis, Multidimensional Scaling etc. Among the numerous methods, principal component analysis (PCA), linear discriminant analysis (LDA) and the maximum margin criterion (MMC) are the most famous due to their simplicity and effectiveness (Hu, 2021).

The study of variable selection problems dates back more than 50 years (Roy, 1958, Efroymson, 1960; Beale *et al*., 1967 apud Brusco, 2014) and continues to be a relevant topic with recent contributions. In mining, some current work is being done using PCA associated with other machine learning techniques: to reduce geometallurgical variables in neural network models (Mu, 2023); to identify variables that best correlate in geometallurgical studies in a gold mine using self-organizing map clustering (Costa, 2023); in models for predicting water inrush in coal mines using neural networks (DBN) in order to reduce the long training time (Zhang, 2022); in work on the spatial autocorrelation of geotechnical information to reduce the increasing dimensionality of the data set caused by the EDF-Euclidean distance field features (kim *et al.*, 2023); to reduce the dimensionality of hydrochemical data and define the water classification model using clustering techniques at a gold mine in China (Liu *et al.*, 2019); in studies of acid drainage from a galena mine in Japan, identifying the most relevant geochemical components and isotopes for determining water quality (Tomiyama, 2019); to reduce the multidimensionality of compositional and non-compositional data using the scren plots technique in geometallurgical studies in Paracatu (Bhuiyan, 2019); to differentiate lithological variations and hydrothermal changes in multivariate geochemical data and limit the impact of outliers and values below the detection limit, plus hierarchical cluster analysis in mineral exploration work in New Zealand and Australia (Gasley, 2015); in studies to determine the most significant attributes in methane emissions in coal mines, reducing the computational time of predictive regression models and neural networks (Karacan *et al.*, 2011).

Brusco (2014) cites other multivariate methods combined with PCA in studies on multiple linear regression (Gatu & Kontoghiorghes, 2006; Gatu *et al.*, 2007; Hofmann *et al.*, 2007), discriminant analysis (Pacheco *et al.*, 2006; Trendafilov & Jolliffe, 2007; Brusco & Steinley, 2011), cluster analysis (Krzanowski & Hand, 2009), principal component analysis (Jolliffe, 2002; Mori *et al.*, 2007; Brusco *et al.*, 2009[a]; Pacheco *et al.*, 2013) and factor analysis (Kano & Harada, 2000; Hogarty *et al.*, 2004).

It is worth noting that PCA is a technique that does not indicate the number of components and the variables selected (Jiang, 2017). Jolliffe (1972) published a series of comparative analyses using an artificial database. In a revision of this work in 1973, the author used three real databases (of pitprops of Corsican pine, winged aphids and crimes committed in the U.K), but containing a relatively insignificant number of samples when compared to geoscience databases. In the mining studies mentioned above, the selection of variables was based on the expert's knowledge using a single previously defined criterion, usually the most expressive variable in a component obtained by a cut-off value. This cut-off value is based on the percentage contribution of each componente (eigenvector) in order to obtain a share of the total variability. Another widely used criterion for retaining components is Kaiser's (Kaiser, 1958), in which case the value used for the cut-off is the eigenvector, i.e. components with λi (eigenvalue) > 1 represent a sufficient share of the total variation in the data. Matos *et al.*,2019 suggests that different evaluation criteria should always be used in conjunction with others in order to reaffirm the decision although, in most cases, this decision has been made using a single standard.

Ganguli (2022) cites that the current moment is a milestone for the mining industry, because after decades of focusing on data collection, the industry has evolved to where the focus is now on data utilization. However, although this change is taking place, it is common to observe in practice that the information is not all consolidated in the resource model and is absorbed in a piecemeal and/or incomplete way. This happens for various reasons, such as the lack of isotopy in multivariate databases, the non-additivity and non-linearity of some attributes, the lack of studies using more appropriate techniques for using this information, etc. For this reason, variable reduction is essential in geometallurgical studies in order to check which attributes actually have statistical significance for deposit modeling. The advantages of dimensionality reduction techniques applied to a dataset are numerous, such as: (i) decreasing the number of dimensions and data storage space; (ii) Requiring less time for analysis (iii) Irrelevant, noisy and redundant data can be excluded; and (iv) Data quality may well be optimized (v) Helps an algorithm run efficiently and improves accuracy (vi) Allows data visualization (vii) Simplifies classification and also increases performance (Juvonen *et al.*, 2015, Liu *et al.*, 2009);

Al Kandari & Jolliffe (2001) and Al Kandari & JolliffIe (2005), explain some variable selection criteria based on the principal component covariance, as well as cluster analysis algorithms. Grouping analysis algorithms are used for recognizing patterns in multivariate data, thus helping to interpret the information and meaning of geostatistical domains. The combined action of PCA with grouping techniques is interesting because it is possible to achieve more consistent grouping by utilizing a reduced number of variables (Anzanello & Fogliatto, 2011). Cluster analysis can be applied to any data set to determine groupings of samples without a priori knowledge of their spatial or temporal relationships with each other. However, it becomes especially powerful when the grouped data set is multivariate and has already been subjected to PCA, which has reoriented the data cloud so that these first few dimensions summarize a significant proportion of all the variance. There are several approaches to classifying data clusters into supervised and unsupervised areas.

The aim of this study was to obtain geometallurgical domains with a smaller number of attributes in order to solve issues related to the non-isotopic nature of the database, reduce the time and computational demands of updating three-dimensional models and facilitate the interpretation of multivariate data. PCA was the approach chosen for this study because it is a simple and effective method that can be applied widely and works in most cases, as well as being present in most statistical software. Most of the articles and citations visited indicate that PCA is one of the most efficient methodologies for variable selection and dimensionality reduction. However, as already mentioned, PCA does not indicate the number of components and the decision on variable selection is made subjectively. For this reason, the purpose of this article is also to test different techniques for selecting attributes and to see if there are any differences that could have an impact on the clustering result.

The quality of the clusters was measured using the Elbow graph, the David Bowies index and visual validation. The result was satisfactory for the three criteria adopted, since the number of groups obtained was consistent with the geological individualization carried out on the mining fronts.

2. Materials and methods

The database used comes from an igneous deposit with complex geology formed by the overlapping of various lithologies and with mineral occurrences of phosphate, titanium, rare earth elements and niobium. The chemical database is the most complete with over 40,000 samples, but only 4% of this database is isotopic when considering all the attributes present. For this reason, it is essential to analyze the variables that are statistically important in order to preserve as many samples as possible in the following multivariate analyses. Principal component analysis (PCA) is a technique that allows the reduction of the number of variables to be analyzed, discarding the components that have little variance, studying only those which retain the maximum variation possible present in the data set (Duarte, 1998). This is possible thanks to the transformation of a new set of variables, the principal components (PC), which are not correlated and are ordered in such a way that the first ones retain the biggest part of the variation present in all the other original variables. Thus, PCA problems involve performing a linear data transformation, maximizing transformation variance (Jolliffe, 2002). Consider α as a linear weight and Σ the data covariance matrix. This can be demonstrated for a data set X that *has var*$(\alpha X)=\alpha' \Sigma\alpha$. Using the restriction that the transformed one is independent of $\alpha = 1$, we can stablish PCA as Equation (1):

$$(\sum\alpha\text{-}\lambda I) = 0 \qquad (1)$$

While λ is the Lagrangian, I is a matrix identification. This is the same formula for obtaining eigenvalues (λ) and eigenvectors (α) for a covariance matrix Σ. The eigenvalues λ are ordered according to the covariance percentage. The first eigenvalues provide the maximum percentage of data covariance.

Three approaches for the reduction of variables were tested, namely B2 and B4 methods proposed by Jolliffe (1972), and a variable importance index (VII) which is based on the weight of the variables of linear combinations of PCA presented by Cervo (2015).

In the B2 method, PCA is applied to the normalized data of the correlation n x p matrix, transforming the original set of variables into a new set of variables not correlated among them and arranged in a decreasing order of variance where the first components (eigenvectors) contain the biggest explained variance (eigenvalues) from the original data. For the reduction of variables, the one which represents the biggest coefficient in the main component of smallest variance (eigenvalue) is selected because this attribute is the least important to explain the total variance. This process is repeated for the next variable of the biggest coefficient for the second component of smaller variance and so on and so forth until the discard criterium recommended by Joliffe (1972) is achieved, i.e., the number of discarded variables must be equal to the number of components in which the explained variance (eigenvalue) is inferior to 0.7.

The B4 method follows the same logic as B2, but the procedures are now inverted. The selection process is carried out by retaining the variables with the biggest coefficients for the first components (eigenvectors) whose eigenvalue is bigger than 0.7.

As for the importance of the variable index (IIV), the following steps are followed: PCA is applied to the data. The importance index given by Equation (2) is calculated:

$$IIV_P = \sum_{j=1}^{j} |\alpha_{jp}| . \lambda_j \qquad (2)$$

This index takes into consideration the α_{jp} weight of the variable in each one of the J components (eigenvectors) together with the variance explained by each one of these J components (eigenvalues λ_j). The variables with the biggest indexes were retained for the first components (eigenvectors) in which the eigenvalue is bigger than 0.7.

Once the number of attributes in each of the methods tested hasd been defined, the data was grouped by varying the number of clusters from 2 to 8. The maximum value of k was chosen based on the geological knowledge of the deposit, where 8 main lithologies are recognized. Thus, the number of typological domains is expected to be equal to or less than the 8 lithologies present. The K-means algorithm is one of the most widely used partitioning methods. When using a k-means algorithm, the user must specify the number k of divisions or clusters. Initially, the algorithm randomly chooses k points that will be the centers of the initial clusters (centroids) based on the number of the groups they should be divided into. Each sample point is then assigned to the nearest centroid. The position of each centroid is then updated based on the configuration of the points in each group. This process is repeated until the centroids are no longer modified (Tan *et al.*, 2006). A disadvantage of the method is that the clustering converges to a local minimum. Thus, to find the best cluster, it is necessary to run the algorithm several times with several initial centroids and then choose the best result.

The results of the clusters were then checked. There are various techniques that help assess the quality of clustering, such as the Silhouette coefficient, the Davies Bouldin index, the Elbow method, etc., the latter two of which are presented here. The Elbow method tests the variance of the data in relation to the number of groups, i.e. the ideal k value is related to the smallest sum of squares within the cluster (wcss). Configurations with more compact groups have lower WCSS values because the distance between the elements within each group is smaller. This method is one of the most classic for determining the number of clusters in a data set, as well as being a visual method (Kodinariya & Makwana 2013). The Davies Bouldin index aims to find spherical clusters, with internal compactness and, at the same time, with good separability between the other clusters and is given by the ratio of intra-cluster and extra-cluster dispersion, i.e., it considers the proportion of dispersion within the cluster and the separation between the clusters. To define the index, it is first necessary to define the cluster's dispersion and similarity measure (Ganmawu & Wells, 2007 apud Oliveira *et al.*, 2020). This similarity is calculated as shown below:

$$R_{IJ} = \frac{S_i + S_j}{d_{ij}} \qquad (3)$$

While S_i and S_j are the dispersion measures within clusters C_i and C_j, and d_{ij} is the distance among the centroids of groups C_i and C_j. The Davies-Bouldin (*DB*) index is given by the equation:

$$DB = \frac{1 \sum_{k=1}^{k} R_i}{k} \qquad (4)$$

K: Number of clusters; R_i: is the maximum value R_{ij}.

3. Results and discussions

The eigenvalues and eigenvectors were obtained from the range of principal components, as shown in Table 1. Note that 12 of the 19 principal components had a variance of less than 0.7 (eigenvalue). This means that 12 attributes can be discarded, leaving seven variables for clustering, a 63% reduction in the number of variables. Table 2 shows the seven variables selected in each of the three methods tested, i.e. methods B2 and B4 using the criterion of excluded and retained attributes, and the IIV method used to calculate the importance index.

The variable $Al_2O_3\_C$ was selected in all the methods tested. Nb_2O_5, Fe_2O_3, SiO_2 and BaO appear in two of the three methods tested (Table 2). Al_2O_3, DIST_MASS, SiO_2_C and TiO_2 are not selected in any of the techniques used and can be discarded. It is worth noting that some of these variables showed high linear correlation values and are therefore redundant. This means that although the three methods have different variables, some of them have similar characteristics and therefore correspond (Figure 1b).

For example, Fe_2O_3 (present in B2 and B4) has a high MMAG correlation (present in IIV), so the high values of one induce the high values of the other and therefore do not need to coexist simultaneously within the same group. The same is true of Nb_2O_5 and $Nb_2O_5\_RF$; P_2O_5, CaO, $P_2O_5\_C$ and CaO_C; MgO and MgO_C. On the other hand, the attributes selected within the same method will be less correlated with each other (Figure 1a).

Table 1 - PCA results.

PC	Eigenvalue	Explained variance proportion	Accumulated explained variance
CP1	5.0429	0.265	0.265
CP2	2.7391	0.144	0.41
CP3	2.1341	0.112	0.522
CP4	1.8765	0.099	0.621
CP5	1.644	0.087	0.707
CP6	1.4885	0.078	0.786
CP7	0.7376	0.039	0.824
CP8	**0.6467**	0.034	0.858
CP9	**0.5396**	0.028	0.887
CP10	**0.4665**	0.025	0.911
CP11	**0.4309**	0.023	0.934
CP12	**0.3438**	0.018	0.952
CP13	**0.2645**	0.014	0.966
CP14	**0.2281**	0.012	0.978
CP15	**0.1608**	0.008	0.987
CP16	**0.1106**	0.006	0.992
CP17	**0.0965**	0.005	0.997
CP18	**0.0371**	0.002	0.999
CP19	**0.012**	0.001	1

Table 2 - Results from the selected variables.

Selected variables		
B2	**B4**	**Importance index**
Nb_2O5	CaO_C	$P_2O_5\_C$
SiO_2 _	Fe_2 O_3	SiO_2 _
Fe_2O_3	BaO	CaO
BaO	MgO	P_2O_5
Fe_2O_3C	LAMA	$Nb_2O_5\_RF$
MgO_C	$Al_2O_3\_C$	MMAG
$Al_2O_3\_C$	Nb_2O_5	Al_2O_3 _C

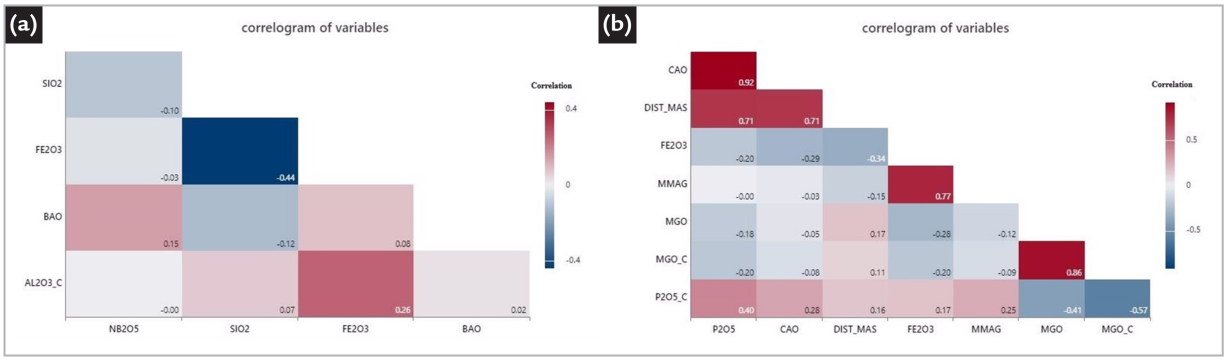

Figure 1 - a) Low correlation of the variables selected in the same method). b) Variables with a high correlation (redundant).

Al_2O_3, Fe_2O_3 and Nb_2O_5 and their correspondents (MMAG and Nb_2O_5_RF) plus SiO_2 and BaO represent 71.4% of all the variables selected in at least two methods. These are the most relevant elements for the grouping and domain definition tests. Clustering was carried out using the k-means method, with the number of clusters (k) varying from 2 to 8. After finalizing the division, the ideal number of domains was chosen, and the quality of the grouping was checked with and without variable reduction. Deciding on the number of clusters is one of the most important considerations in unsupervised clustering algorithms for defining domains.

The Elbow graph shows that after 4 domains, there is an inflection of the curves, i.e. there is no significant change in the variance within the groups. It can therefore be said that the choice of four domains would be appropriate for the case studied for all the methodologies tested. Selection methods B4 and B2 showed greater adherence, but IIV produced better results (Figure 2).

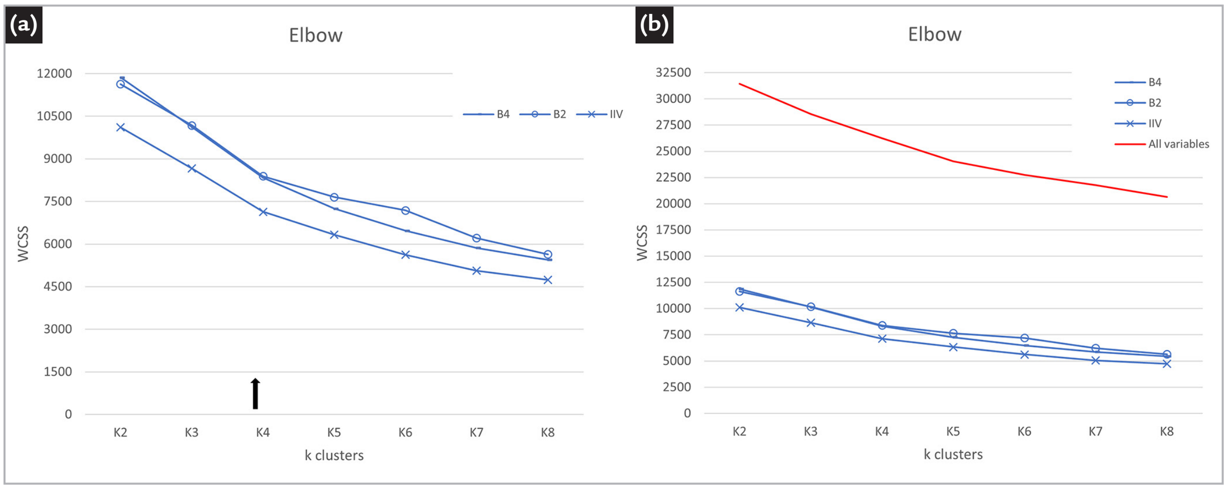

Figure 2 - Elbow graph to check the quality of the clustering, in a) selected variables and in b) comparing the total database.

The Davies Bouldin index is illustrated in Figure 3, which shows the number of clusters. The closer the index is to zero, the better the clustering result. Similar to the Elbow graph, in the Davies index four clusters are indicated in IIV and B2. For B4, a number of clusters equal to 5 has the best index, although with a very small difference for k= 4. This fact reinforces the importance of using more than one technique in order to confirm the appropriate number of clusters, since the results obtained for B2 and B4 in the Davies index are not strongly conclusive. With regard to the adherence of the techniques again, B2 and B4 are more adherent and IIV shows less intra-group variability.

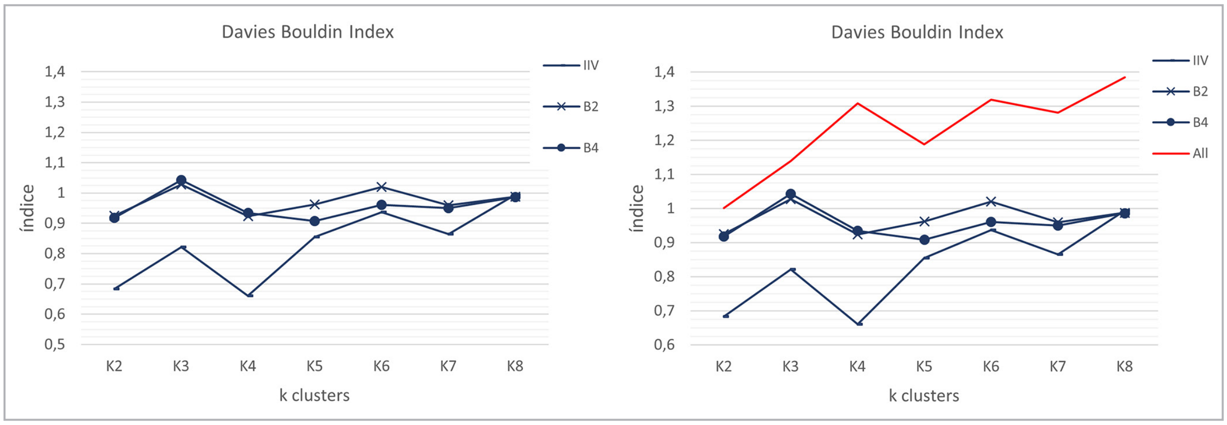

Figure 3 - DB index graph for checking cluster quality.

A visual inspection was carried out varying the elevation. Figure 4 compares the results of the clusters generated with and without the reduction of variables in a section. There is a better regionalization of the clusters in the simulations that used variable reduction, and a greater dispersion of the samples in the grouping with all the data. The B4 method showed a better separation of the groups. The correlation between the domains generated by k-means using the B4 method and the typological domains in the original data set can be seen in Figure 5. What can be seen is a coherence between the groupings and the predominant typologies, i.e., as expected, ores of the CBMG, NL and FO types tend to group together, as do FL, PI and CBM.

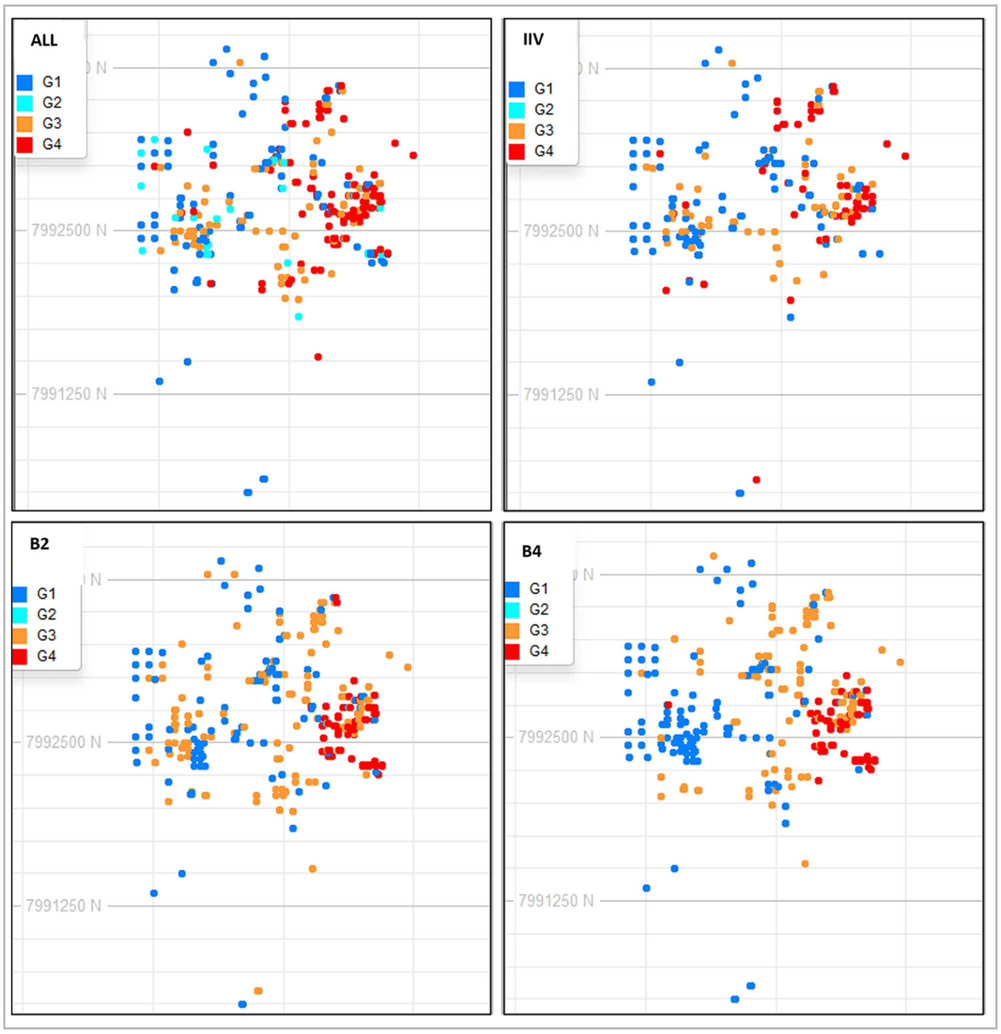

Figure 4 - Clustering result for k=4 clusters. In the upper image (all) using all the variables; in B2, B4 and IIV using the variables selected in each method.

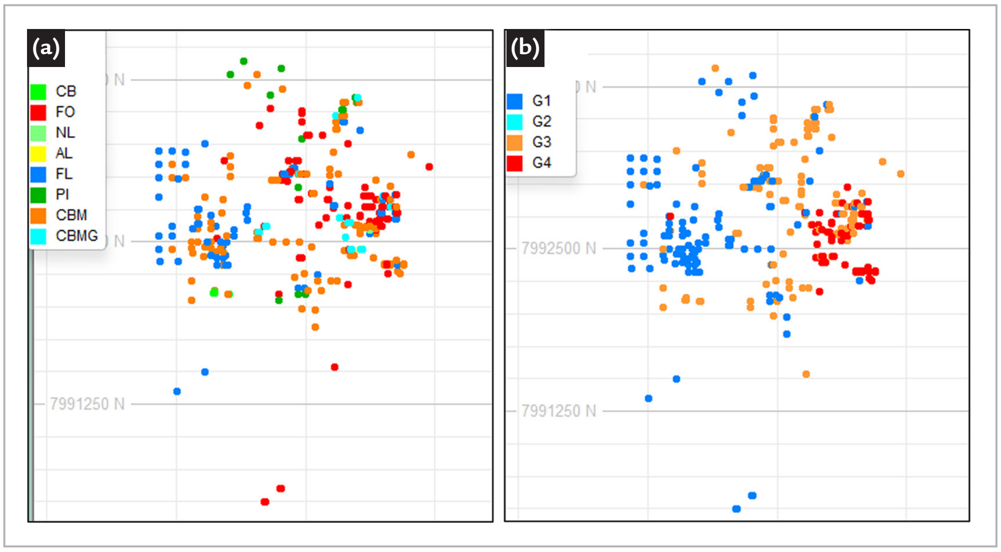

Figure 5 - On the right the samples according to typology (geological description), on the left the result of grouping technique B4.

4. Conclusions

The following variables Nb_2O_5, Fe_2O_3, BaO and SiO_2 are important in explaining the variability of the data, being present in 2 of the 3 methods selected. Al_2O_3C was selected in all three techniques tested. The metrics used to check the quality of the clusters show that the methodologies used are consistent. The least satisfactory result in terms of the compactness of the groups occurred when the database with all the variables was used. This was expected, since irrelevant attributes can distort the quality of the clusters formed. Thus, we can eliminate these redundant features, as well as those that are difficult to measure, which reduces the time and cost of the experiments. It is important to note that the metrics obtained by these indices do not dispense with visual inspection and expert evaluation in order to check for coherence between the clustered domains. The clusters obtained with dimensionality reduction through PCA were generally consistent with the typological domains and their characteristics, although the distribution of the groups appears visually different in the three techniques, with B2 and B4 being more similar and IIV more discrepant. This indicates that it is necessary to be careful regarding the selection criteria, since these can impact the quality of the grouping, making it important to use more than one technique, which does not always happen.

In future projects, it is recommended that other variable selection methodologies be tested, taking into account the non-linearity characteristics that some metallurgical variables may present, certifying the attributes that best define the typological domains. Mining companies can benefit from the results of this research when making decisions, especially with regard to the analysis of variables and their order of importance, since some of these are not so relevant for modeling and controlling ore quality.

Another important consideration is that only one of the five explanatory variables used in the cluster algorithm is part of the metallurgical database, reducing the number of non-additive variables and facilitating other analyses involving geostatistics and domain modeling.

Acknowledgements

The authors thank CEFET-MG, UFCAT for their support during the research.

References

AL KANDARI, N. M.; JOLLIFFE, I. T. Variable selection and interpretation of covariance principal components. ***Communications in Statistics Simulation and Computation,*** 30, p. 339-354, 2001.

AL KANDARI, N. M.; JOLLIFFE, I. T. Variable selection and interpretation of correlation principal component. ***Environmetrics,*** 16, p. 659-672, 2005. Version online 28 Apr. 2005 in Wiley InterScience. Available in: www.interscience.wiley.com

ANZANELLO, M. J.; FOGLIATTO, F. F. Selecting the best clustering variables for grouping mass-customized products involving Workers learning. ***International Journal of Production Economics P***, 2011.

BARNETT, R. M.; DEUTSCH, C. V. Practical implementation of non-linear transforms for modeling geometallurgical variables. *In*: ABRAHAMSEN, P.; HAUGE, R.; KOLBJORNSEN, O. (org.) Geoestatística. ***Geologia Quantitativa e Geoestatística***, v. 17, 2012. Oslo: Springer, Dordrecht.

BHUIYAN, M.; ESMAEILI, K.; ORDÓÑEZ-CALDERÓN, J. C. Application of data analytics techniques to establish geometallurgical relationships to bond work index at the Paracatu mine, Minas Gerais, Brazil. ***Minerals***, v. 9, n. 5, 302, 2019.

BRAGA, S. A.; COSTA, J. F. C. L Krigagem dos indicadores aplicada a modelagem das tipologias de minério fosfatados da Mina F4. ***Holos***, 1, 394-403, 2016. DOI: https://doi.org/10.15628/holos.2016.3870

BRUSCO, M. J. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. ***Computational Statistics & Data Analysis***, 77, p. 38-53, 2014.

CERVO, V. L.; ANZANELLO, M. J. Seleção de variáveis para clusterização de bateladas produtivas através de ACP e remapeamento kernel. ***Production,*** v. 25, n. 4, p. 826-8332015, 2015.

COSTA, F. R.; CARNEIRO, C. D. C.; ULSEN, C. Self-organizing maps analysis of chemical-mineralogical gold ore characterization in support of geometallurgy. ***Mining,*** 3, 230-240, 2023. Available in: https://doi.org/10.3390/ mining3020014

DUARTE, J. B. ***Introdução à análise de componentes principais.*** São Paulo, 1998. Available in : https://files.cercomp.ufg.br/weby/up/396/o/ACP.pdf. Access in: oct. 2022.

GANGULI, R.; DESSUREAULT, S.; ROGERS, P. Introduction to the special issue: advances in computational intelligence applications in the mining industry. ***Minerals,*** 2022. DOI: https://doi.org/10.3390/ min12010067

GAZLEY, M. F.; COLLINS, K. S.; ROBERSTON, J.; HINES, B. R.; FISHER, L.A.; McFARLANE, A. Application of principal component analysis and cluster analysis to mineral exploration and mine geology. CONFERÊNCIA ANUAL DA FILIAL AusIMM, Nova Zelândia, 2015.

HU, H.; FENG, D.-Z.; CHEN, Q.-Y. ***A Novel dimensionality reduction method***: similarity order preserving discriminant analysis. Signal Process Elsevier, 2021.

JIANG, Z.; LI, J. High dimensional structural reliability with dimension reduction. ***Structural Safety***, 69, p. 35-46, 2021.

JOLLIFFE, I. T. Discarding variables in a principal component analysis - I: artificial data. ***Journal of the Royal Statistical Society,*** n. 2, p.160-173, 1972.

JOLLIFFE, I.T. Discarding variables in a principal component analysis -II: real data. ***Applied Statistics***, v. 22, p. 21-31, 1973.

JOLLIFFE, I.T. Mathematical and statistical properties of population principal components. *In*: ***Principal component analysis.*** New York, NY: Springer, 2002. (Springer Series in Statistics).

JUVONEN, A.; SIPOLA, T.; HÄMÄLÄINEN, T. Online anomaly detection using dimensionality reduction techniques for HTTP log analysis, ***Computer Networks,*** v. 91, p. 46-56, 2015. DOI: 10.1016/j.comnet.2015.07.019

KAISER, H. F. The Varimax criterion for analytic rotation in factor analysis. ***Psychometrika,*** 23, 187-200, 1958. DOI: https://doi.org/10.1007/BF02289233

KARACAN, C. O.; RUIZ, F. A.; COTE, M.; PHIPPS, S. Coal mine methane: a review of capture and utilization practices with benefits to mining safety and greenhouse gas reduction. ***International Journal of Coal Geology,*** p. 121-156, 2011.

KIM, H.-J.; MAWUNTU, K. B. A.; PARK, T.-W.; KIM, H.-S.; PARK, J.-Y.; JEONG, Y.-S. Spatial autocorrelation incorporated machine learning model for geotechnical subsurface modeling. ***Applied. Sciences,*** 2023. DOI: https://doi.org/10.3390/app13074497

LIU, L.; ZSU, M. T. ***Encyclopedia of database systems.*** Springer Publishing Company, 2009.

LIU, G.; MA, F.; LIU, G.; ZHAO, H.; GUO, J.; CAO, J. Application of multivariate statistical analysis to identify water sources in A coastal gold mine, Shandong, China. ***Sustainability,*** v. 11, p. 3345. DOI: https://doi.org/10.3390/su11123345

MATOS, D. A. S.; RODRIGUES, E. C. ***Análise fatorial.*** Brasília: Enap, 2019. (Coleção Metodologia de Pesquisa).

MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. ***Psychometrika,*** v. 45, n. 3, p. 325-342, 1980.

MU, Y.; SALAS, J. C. Data-driven synthesis of a geometallurgical model for a copper deposit. ***Processes,*** 2023, 11, 1775. DOI: https://doi.org/10.3390/pr11061775

OLIVEIRA, F. R. *et al.* Clusterização de clientes: um modelo utilizando variáveis categóricas e numéricas. CONGRESSO BRASILEIRO DE ENGENHARIA DE PRODUÇÃO, 2020.

TAN, P.; MICHAEL, S.; KUMAR, V. Cluster analysis: basic concepts and algorithms. *In*: ***Introduction to Data Mining.*** New York: Pearson Education, 2006.

TOMIYAMA, S.; TOSHIFUMI, I.; TABELIN, C. B.; TANGIROON, P.; HIROYUKI, L. Acid mine drainage sources and hydrogeochemistry at the Yatani mine, Yamagata, Japan: a geochemical and isotopic study. ***Journal of Contaminant Hydrology,*** 2019.

ZHANG, Y.; TANG, S.; SHI, K. Risk assessment of coal mine water inrush based on PCA-DBN. ***Nature,*** 2022. https://doi.org/10.1038/s41598-022-05473-8

Received: 30 June 2023 - Accepted: 24 October 2023.