HOLOS
ISSN 1807 - 1600

SELF-ORGANIZING MAPS APPLIED TO DECLUSTERING IN PREFERENTIAL SAMPLING

N. K. AYACHE, A. E. M. SANTOS*, A. E. A. NASCIMENTO, S. A. B. DE CASTRO, D. F. S. DA SILVA
Federal University of Ouro Preto
ORCID ID: 0000-0003-4302-3897*
allan.santos@ufop.edu.br*

Submitted March 27, 2023 - Accepted December 26, 2023

DOI: 10.15628/holos.2023.15200

ABSTRACT

Sampling processes in mineral exploration often result in preferentially sampled areas, with the formation of clustering. Some factors can cause areas to be preferentially sampled, accessibility conditions, attribute values, and the sampling strategy. Clustering impacts statistical inference of area. The objective of the present paper is to propose a new approach to declustering methods using Kohonen network, Self-Organizing Maps (SOM). SOM are a type of artificial neural network used for unsupervised classification. The methodology assigns each sample a weight to calculate the declustered mean. The assignment of weight to each sample in an area is inversely proportional to the densely sampled in area. The declustered mean is given by the sum of the weight multiplication with the attribute value of each sample. Therefore, the logic of assigning weights is similar to Cell Declustering method, but the delimitation of the densified areas is different. SOM identifies areas with non-linear margins, unlike the Cell Declustering method. A case study is presented, using the Walker Lake data set. The present research is not intended to replace classical declustering methods, but rather to present a new approach to a routine problem in reserve evaluation. Although the mathematics of the applied technique is indeed complex, the results can be promising.

KEYWORDS: Self-organizing maps, Kohonen networks, Declustering methods, Preferential sampling.

MAPAS AUTO-ORGANIZÁVEIS APLICADOS AO DESAGRUPAMENTO EM AMOSTRAGEM PREFERENCIAL

RESUMO

Os processos de amostragem na exploração mineral muitas vezes resultam em áreas preferencialmente amostradas, com a formação de agrupamentos, que podem surgir devido a alguns fatores, tais como condições de acessibilidade, valores de atributos e a estratégia de amostragem. Os agrupamentos afetam a inferência estatística da área. O objetivo deste artigo é propor uma nova abordagem para métodos de desagrupamento usando as redes de Kohonen, Self-Organizing Maps (SOM). As SOMs é um tipo de rede neural artificial usada para classificação não supervisionada. A metodologia atribui a cada amostra um peso para calcular a média desagrupada. A atribuição de peso para cada amostra em uma área é inversamente proporcional à área densamente amostrada. A média desagrupada é dada pela soma da multiplicação do peso com o valor do atributo de cada amostra. Portanto, a lógica de atribuição de pesos é semelhante ao método Cell Declustering, porém as SOMs identificam as áreas com margens não lineares, ao contrário do método Cell Declustering. Um estudo de caso é apresentado, usando o conjunto de dados de Walker Lake. A presente pesquisa não pretende substituir os métodos clássicos de desagrupamento, mas sim apresentar uma nova abordagem para um problema rotineiro na avaliação de reservas. Embora a matemática da técnica aplicada seja de fato complexa, os resultados podem ser promissores.

Palavras chave: Mapas auto-organizáveis, Redes de Kohonen, Métodos de desagrupamento, Amostragem preferencial.

HOLOS
ISSN 1807 - 1600

1 INTRODUCTION

Through the multidisciplinary interface among geology, geostatistics, and mineral processing, it is possible to propose the creation of a geometallurgical model (Braga and Costa, 2016). Geometallurgical modeling allows anticipating issues in subsequent stages of mining and ore treatment (Motta, 2014), contributing to better planning, minimizing processing risks, and optimizing production plans in beneficiation plants (Vieira et al., 2015). In this context, the sampling process corresponds to a sequence of systematic operations aimed at representing, through the collection of a small portion called a sample, a specific universe. Therefore, it can be considered that this stage is the key to the success of the mineral exploration phase.Simple or stratified random sampling can cause clusters with a greater densely samples in areas when compared to others areas, for these cases, sampling is said to be preferential. Souza et al. (2001) lists three situations that can lead to preferential sampling in certain areas: accessibility conditions, attribute values, and the sampling strategy.

Souza et al. (2001) also explain that when the database does not include a sufficient amount of information to guarantee reliability for the inference, it is necessary to perform the declustering of the data. This procedure consists of assigning weights to the data, to attenuate or moderate the influence of sparse data. Consequently, data from densely sampled areas may be given less weight than data from sparsely sampled areas.

In current engineering practice, there are declustering methods that are applicable to any sample data set, the polygonal method (Isaaks and Srivastava, 1989) and cell declustering method (Journel, 1983; Deutsch, 1989). In methods a weighted linear combination is applied of all available sample values to estimate the declustered mean. These methods correct the weight for clustered samples.

The polygonal method, assigns a polygon of influence to each sample. The areas of these polygons are then used as the declustering weights. The cell declustering method, uses the moving window concept to calculate how many samples fall within particular regions or cells. The declustering weight assigned to a sample is inversely proportional to the number of other samples that fall within the same cell (Isaaks and Srivastava, 1989).

This paper presents an evaluation of the declustering using Self-Organizing Maps (SOM), with a case study applied to the Walker Laker data set (Isaaks and Srivastava, 1989). The SOM can be defined as a neural network of unsupervised learning, where the network seeks to group the input data based on their similarities forming classes. In this way, from the results obtained, initial evidence of applicability can be measured as an alternative tool to the classical methods to be used in the disaggregation of sample data.

The major relevance of the proposed methodology is identifies areas with non-linear margins. Thus, the methodology is like the Polygon Method in the neighborhood construction, and similar as Cell Declustering method in declustering weights construction. The SOM works as methodology between the two classical methods declustering. This is believed to be the main justification for choosing SOM as a declustering method.

HOLOS
ISSN 1807 - 1600

2 MATERIALS AND METHODS

2.1 Dataset

The database used as a case study was the Walker Lake dataset, according to Isaaks and Srivastava (1989), 470 sampled sites. The walker lake database is located in the northeast district of Nevada, it corresponds to a saline, lacustrine deposit. The variable V was used as an attribute of interest. The preferential clustering zones are concentrated in areas with a high value for variable V (see Figure 1). The methodology was developed in R language (R Core Team, 2016).

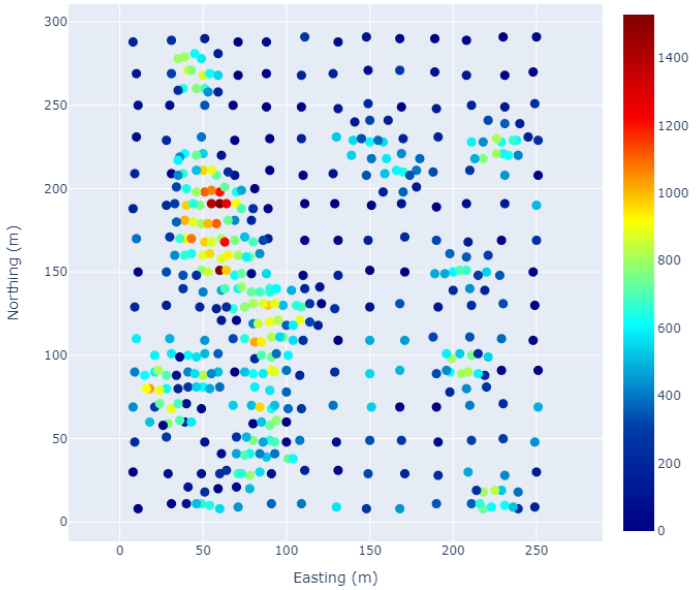


Figure 1: Map of variable V.

2.2 Similarity metric

The Manhattan distance was used, whose metric is such that the distance between two points is the sum of the absolute differences of their coordinates as shown in Equation 1, where d (*i, j*) is the Manhattan distance between samples *i* and *j*.

$$d(i,j) = \sqrt{\left(\left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right|\right)}$$

(1)

According to the data from the distance matrix, a selection of the 5 smallest distances obtained for each sample from the database was made (SOM_A, SOM_B, SOM_C, SOM_D and SOM_E). Once selected, it was possible to generate a new database with 5 variables that represented, in ascending order, the 5 smallest distances for each sample. These new variables were used in the Kohonen Networks (Kohonen, 1981a; Kohonen, 1981b; Kohonen, 1981c). The header of database used is presented in Table 1.

Table 1: Database header.

Sample	SOM_A	SOM_B	SOM_C	SOM_D	SOM_E
1	0.954	0.955	0.909	0.671	0.725
2	0.839	0.894	0.680	0.666	0.641

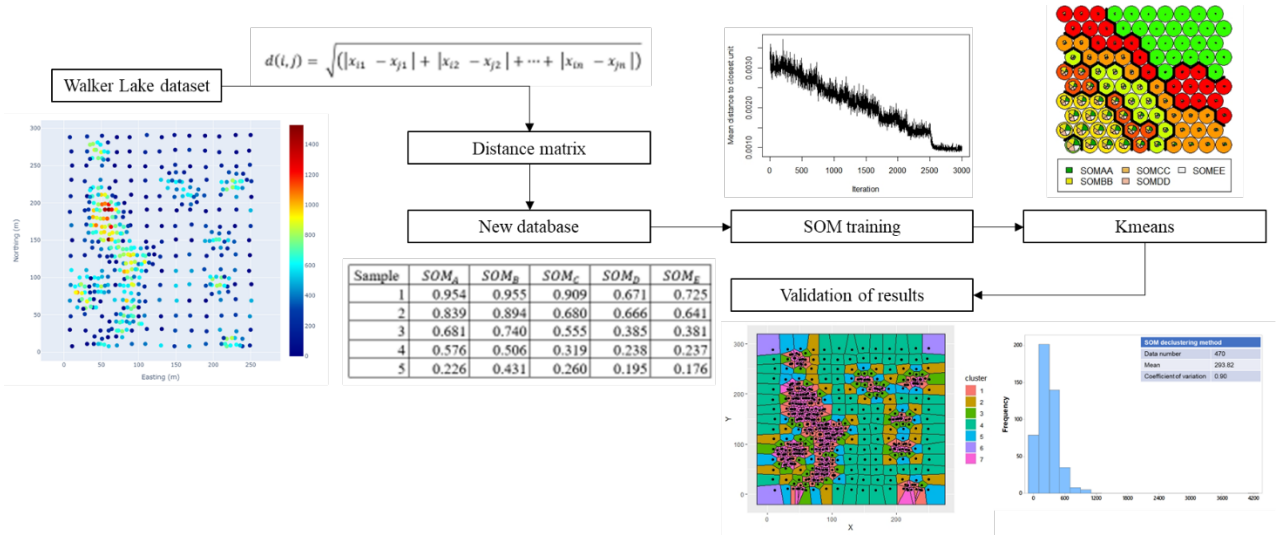
3	0.681	0.740	0.555	0.385	0.381
4	0.576	0.506	0.319	0.238	0.237
5	0.226	0.431	0.260	0.195	0.176

For the development of the distance matrix, the Rgeos package (Interface to Geometry Engine - Open Source) was used because it allows the creation of several applications and techniques for spatial data. The Rgeos package was developed by Bivand and Colin (2017).

2.3 General methodology

Kohonen networks were trained on the distance database, and the SOM map was generated. The K-means algorithm (MacQueen, 1967) was applied to generate the groups. In each group generated, the formula of Deutsch (1989) was applied to obtain the weights of declustering.

The validation of the methodology worked from the comparison of the results obtained with the results of the traditional disaggregation methods: Cell declustering (Deutsch, 1989) and polygonal method or nearest neighbor (Cover and Hart, 1967). According to this comparison it was possible to extract the metrics to assessment of the new approach studied. Figure 2 presents the flowchart of the general methodology.



Sample	SOM_A	SOM_B	SOM_C	SOM_D	SOM_E
1	0.954	0.955	0.909	0.671	0.725
2	0.839	0.894	0.680	0.666	0.641
3	0.681	0.740	0.555	0.385	0.381
4	0.576	0.506	0.319	0.238	0.237
5	0.226	0.431	0.260	0.195	0.176

Figure 2: Gerenal methodology.

2.4 Implementation of the model

The Kohonen package was used, developed by Wehrens and Kruisselbrink (2018), it has the ability to implement various forms of SOMs. The basis used for development comes from studies developed by Kohonen et al. (1995).

The grid created is 10x10 with neurons of circular topology. This choice was made from the premise of optimizing the processing time by increasing the segregation of samples in each neuron. With the grid of these dimensions, 100 neurons were created as can be seen in Figure 3. SOM training was done with 1000 iterations.

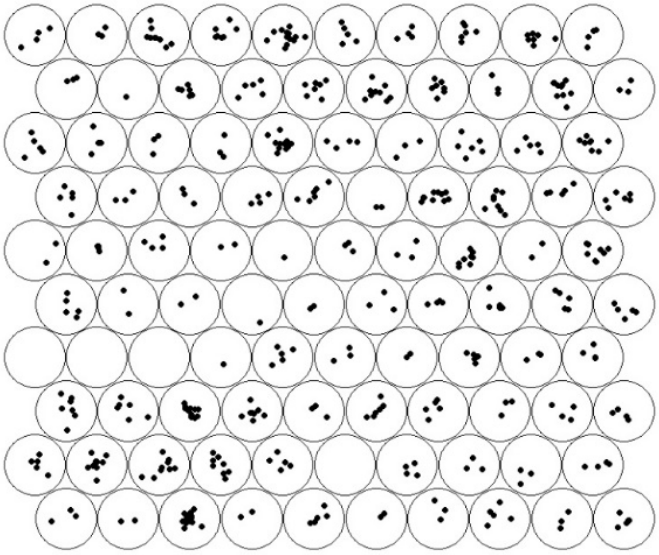


Figure 3: Grid of neurons.

With the samples classified in different neurons, it was possible to group them and determine the sets with the greatest similarity. Input data were neurons created by SOM and output information were clusters created by K-means. The algorithm used in K-means was Hartigan-Wong (1979) (Hartigan and Wong, 1979). In addition, other parameters used in K-means were the 50 iterations, and the number of 1000 random starts to minimize the variation of the output results obtained. To determine the number of clusters, the Elbow method was applied.

After obtaining the number of groups the result obtained determined which neurons, and thus the samples, would be grouped in each cluster, generating the final grouping. With the groups determined, they were submitted to the equations developed by Deutsch (1989) to calculate the weights. Equations 2 and 3 represent the calculations used to determine the weights and ungrouped means respectively, according to Deutsch (1989).

$$w_i = \frac{1}{n_l . l_o} \tag{2}$$

$$\bar{Z} = \sum_{i=1}^{n} w_i . z_i \tag{3}$$

Where w_i is the weight of the samples in the group, n_l is the number of groups, l_o the number of samples in the group, $\bar{Z}$ the declustered mean and z_i a sample *i* of the data group.

2.5 Repository of codes for reproducing the applied methodology

The repository with the codes can be found in GitHub, see the link: https://github.com/MrColugo/Kohonen-Self-Organizing-Maps-applied-to-declustering-in-preferential-sampling

3 RESULTS AND DISCUSSIONS

Figure 4 shows the number of samples selected for each neuron in the grid created. The ideal of this step is that there are not many empty neurons or overloaded neurons, the balance is necessary so that there is no unnecessary excess processing or low processing rate, respectively the conditions mentioned above. The result obtained was satisfactory, demonstrating the efficiency in the selection of the developed neuron grid.

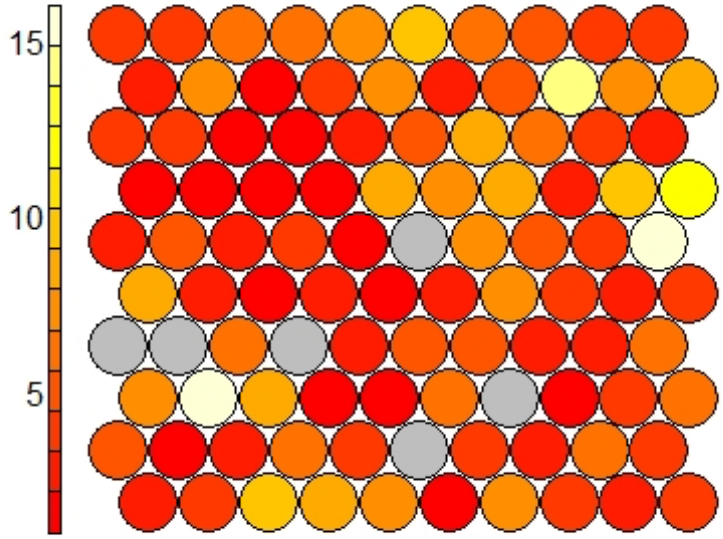


Figure 4: Sample density in Walker Lake grid neurons.

Figure 5 shows the representation of variables in each neuron by a pie chart, each slice represents a variable and the greater the radius of the pie slice, the greater the range of acceptable values in that neuron.

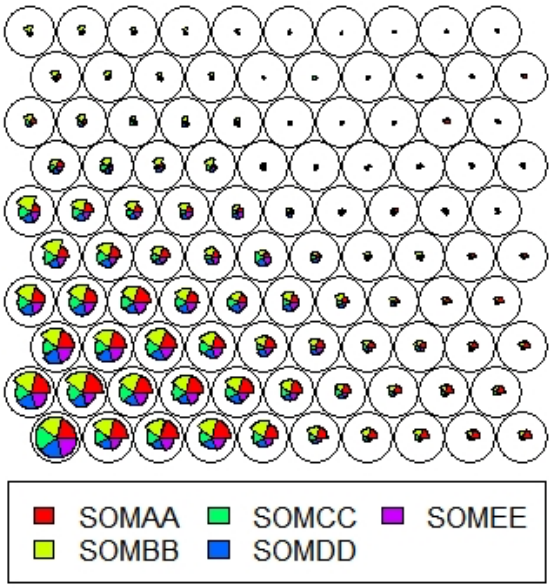


Figure 5: Relationship of variables in neurons.

The training graph of the SOM network model is shown in Figure 6. The training graph represents the average distance between input samples within the network. The smaller this distance, the better the quality of the modeling. The number of interactions between the samples determines how much this "approximation" process should be done, when this process stabilizes, it is recommended that the development of the network ends because there are no more learning gains at this point.

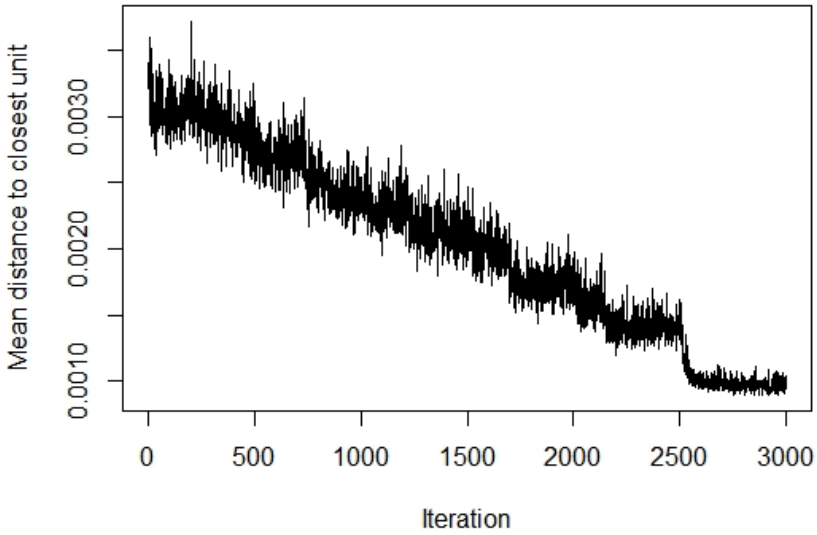


Figure 6: SOM training.

As can be seen in Figure 6, the result obtained for the training was satisfactory in 3000 iterations, reaching all the performance goals necessary for the creation of the SOM model.

The selection of the optimal number of groups (k) can be seen in Figure 7. The curve represents the average error of K-means as a function of the number of clusters, when the decrease of this error stabilizes, the number of groups is chosen.

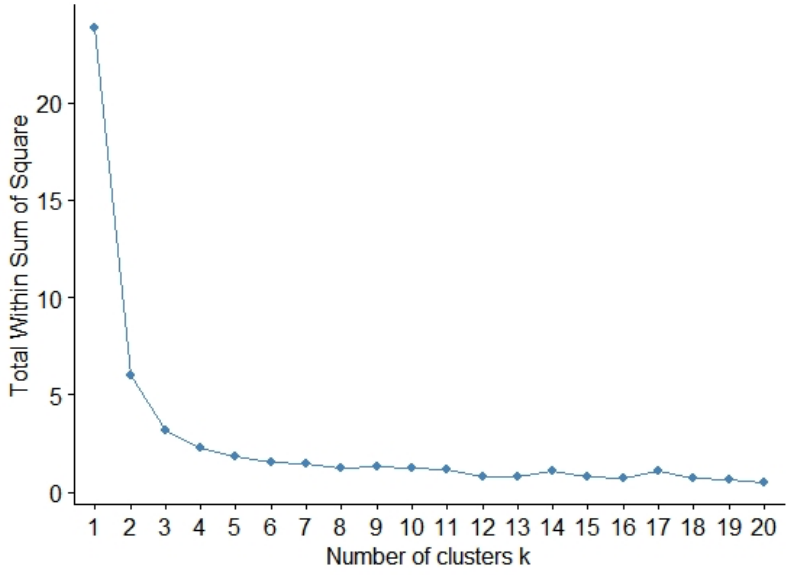


Figure 7: Elbow graph for optimal k.

HOLOS
ISSN 1807 - 1600

Parameter optimization selected 7 groups to efficiently segregate neurons as can be seen in Figure 8. Figure 9 shows the results of the groups in the spatially sampled data. This representation in Figure 9 works as a visual validation, it can be seen that the SOM networks allowed to discriminate the zones with clustering.

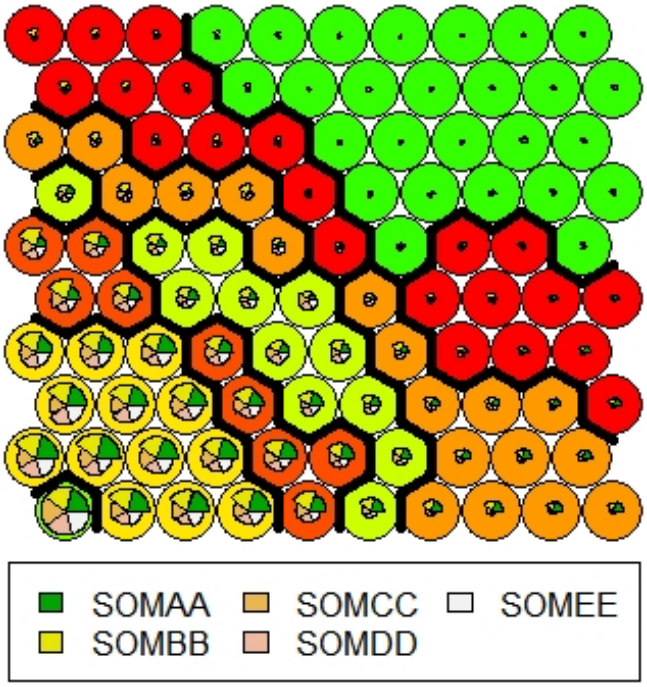


Figure 8: Selection of groups in neurons.

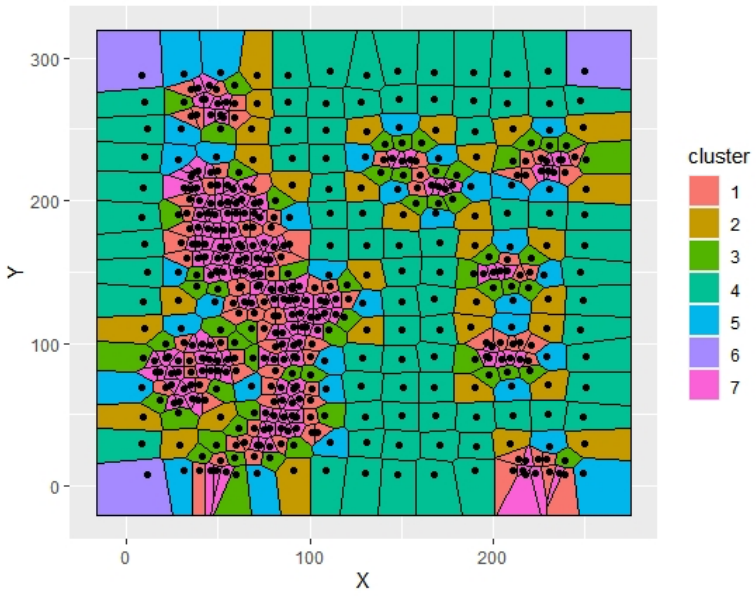


Figure 9: Representation of model groups in variable V of Walker Lake data.

The representation of the histogram with the ungrouped data is shown in Figure 10. An approximation with the exhaustive data histogram is observed (Figure11), when compared to the histogram of the samples (Figure 12). It is observed that the declustering method brought the

HOLOS
ISSN 1807 - 1600

coefficient of variation closer to reality. The coefficient of variation of the samples was 0.69, which indicates lower variance due to the grouping of the data, which is not true, when observing the real data with a coefficient of variation equal to 0.89.

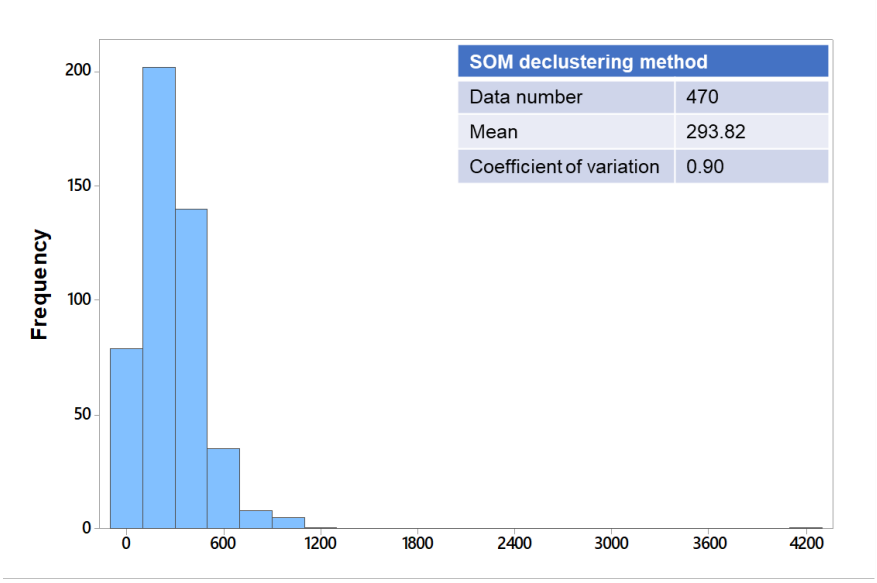


Figure 10: Histogram data declustering.

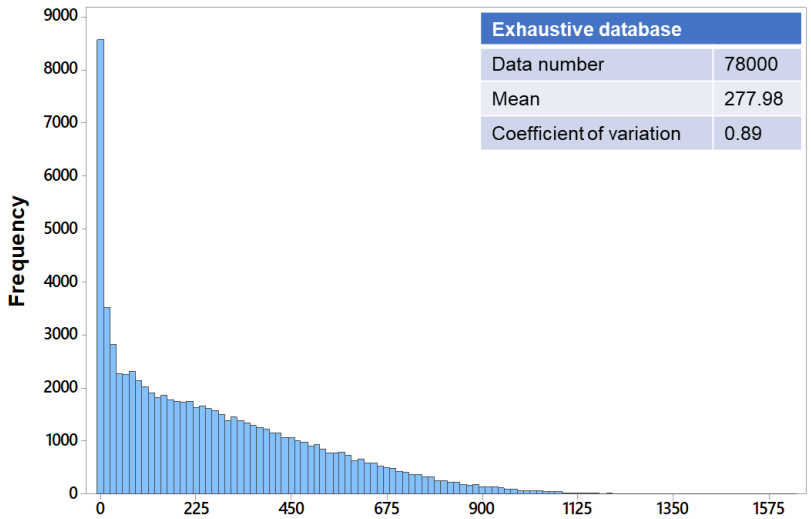


Figure 11: Histogram exhaustive data.

HOLOS
ISSN 1807 - 1600

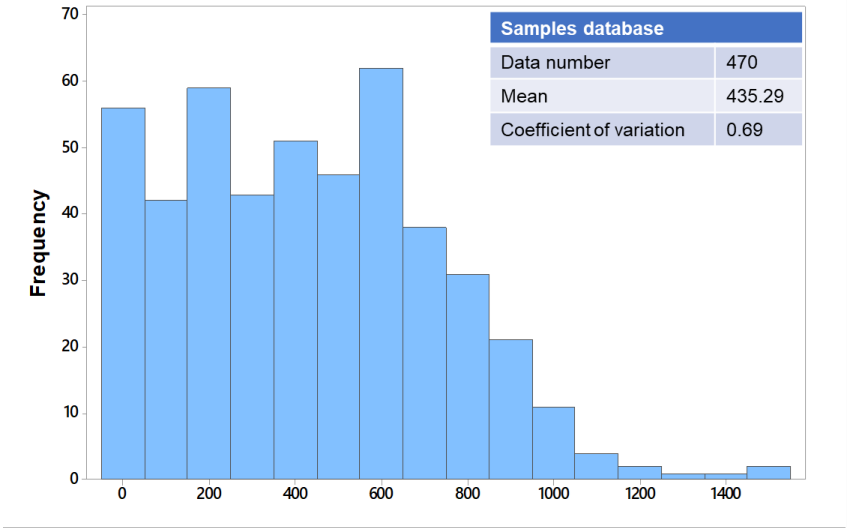


Figure 12: Histogram sample data.

The average obtained by SOM declustering was 293.82, as shown in Figure 10. The average obtained by the classical declustering methods is 290.11 for the cell declustering method, and the polygonal method 282.5 (Table 2). This result shows that SOM declustering approached cell declustering. Although the mean was not lower, the method presents itself as a tool and can be further developed in the future.

Table 2: Mean of methods.

Method	Mean
SOM method	293,82
Cell declustering method	290.11
Polygonal method	282.51

4 CONCLUSIONS

A declustering proposal using SOM was applied and studied in this paper. Variable V from the Walker Lake database was used as a case study.

The weighting logic is similar to the Cell Declustering method, but the delimitation of densified areas is different. SOM identifies areas with non-linear margins, unlike the Cell Declustering method. This is believed to be the main justification for choosing SOM as a declustering method.

The result was close to Cell Declustering, comparing the mean and the coefficient of variation. The application showed is a presentation, as an initial study, and can be improved from the application in mining or environmental case studies, where there is bias in the sampling process.

5 REFERENCES

BIVAND, R. & COLIN, R. (2017). *RGeos: Interface to Geometry Engine - Open Source ('GEOS')*. R package version 0.3–26.

BRAGA, S. A., & COSTA, J. F. C. L. (2016). KRIGAGEM DOS INDICADORES APLICADA A MODELAGEM DAS TIPOLOGIAS DE MINÉRIO FOSFATADOS DA MINA F4. *HOLOS*, *1*, 394–403. https://doi.org/10.15628/holos.2016.3870.

COVER, T. & HART, P. (1967). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21-27. Available in: <http://dx.doi.org/10.1109/TIT.1967.1053964>. Access in: 12 jan. 2022.

DEUTSCH, C.V. (1989). DECLUS: a Fortran 77 program for determining optimum spatial declustering weights. *Computers & Geosciences*, *15*, 3, 325-332.

HARTIGAN, J. A. & WONG, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 28*, 1, 100-108. https://doi.org/10.2307/2346830.

ISAAKS, E. H. & SRIVASTAVA, M. R. (1989). *An introduction to applied geostatistics.* New York: Oxford University Press, 561 p.

JOURNEL, A.G. (1983). Non-parametric estimation of spatial distributions. *Mathematical Geology, 15*, 3, 445-468.

KOHONEN, T. (1981a). Automatic formation of topological maps of patterns in a self-organizing system. E. Oja & O. Simula (eds.), *Proceedings of 2SCIA, Scand. Conference on Image Analysis*, p. 214-220, Helsinki, Finland.

KOHONEN, T. (1981b). *Hierarchical Ordering of Vectoral Data in a Self-Organizing Algorithm*. Report TKK-F-A461, Helsinki University of Technology.

KOHONEN, T. (1981c). *Construction of Similarity Diagrams for Phonemes by a SelfOrganizing Algorithm*. Report TKK-F-A463, Helsinki University of Technology, Espoo, Finland.

KOHONEN, T., HYNNINEN, J., KANGAS, J., LAAKSONEN, J. SOM_PAK. (1995). *The Self-Organizing Map Program Package*. Version 3.1. Helsinki University of Technology, Laboratory of Computer and Information Science, Finland, April 7.

MACQUEEN, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *5th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, CA, USA: University of California Press, p. 281–297.

MOTTA, E.G. Definição de domínios mineralógicos de minério de ferro utilizando krigagem de indicadores. Porto Alegre, 2014. Dissertação de mestrado –Universidade Federal do Rio Grande do Sul, 2014.

R CORE TEAM. (2016). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available in: https://www.R-project.org/.

SOUZA, L. E., WEISS, A. L., COSTA, J. F. C. L., KOPPE, J. C. (2001). Impacto do agrupamento preferencial de amostras na inferência estatística: aplicações em mineração. *REM - International Engineering Journal, 54*, 257-266. https://doi.org/10.1590/S0370-44672001000400005

VIEIRA, M., MENDONÇA, A., & COSTA, J. F. C. L. (2015). MÉTODOS GEOESTATÍSTICOS APLICADOS À MODELAGEM GEOMETALÚRGICA. HOLOS, 7, 65–71. https://doi.org/10.15628/holos.2015.3727.

WEHRENS, R. & KRUISSELBRINK, J. (2018). *kohonen: Supervised and Unsupervised Self-Organising Maps*. R package version 3.0.7. Available in: https://CRAN.R-project.org/package=kohonen.

HOW TO CITE THIS ARTICLE:

Khalil Ayache, N. ., Erlilikhman Medeiros Santos, A. ., Emílio Alves Nascimento, A. ., Alves Braga de Castro, S., & de Fátima Santos da Silva, D. (2023). MAPAS AUTO-ORGANIZÁVEIS APLICADOS AO DESAGRUPAMENTO EM AMOSTRAGEM PREFERENCIAL. HOLOS, 8(39). https://doi.org/10.15628/holos.2023.15200

ABOUT THE AUTHORS:

N. K. AYACHE
Mining Engineer from the Federal Center for Technological Education of Minas Gerais; Senior Strategic Planning Analyst at Mosaic Fertilizantes. Email: naimayache98@gmail.com
ORCID ID: http://orcid.org/0000-0003-3834-6341

A. E. M. SANTOS
Ph.D. in Mineral Engineering from the Federal University of Ouro Preto; Master's in Mineral Engineering from the Federal University of Ouro Preto; Mining Engineer from the Federal University of Ouro Preto; Professor in the Department of Mining Engineering at the Federal University of Ouro Preto. Email: allan.santos@ufop.edu.br
ORCID ID: https://orcid.org/0000-0003-4302-3897

A. E. A. NASCIMENTO
Mining Engineer from the Federal Center for Technological Education of Minas Gerais. Mining Engineer at Carbonífera Cambuí. Email: arture.alves@gmail.com
ORCID ID: https://orcid.org/0000-0002-2199-4898

S. A. B. DE CASTRO
Ph.D. candidate in Exact and Technological Sciences in the Postgraduate Program in Exact and Technological Sciences at the Federal University of Catalão; Master's in Mining, Metallurgical, and Materials Engineering from the Federal University of Rio Grande do Sul; Geological Engineer from the Federal University of Ouro Preto; Professor at the Federal Center for Technological Education of Minas Gerais. Email: silvaniabraga@cefetmg.br
ORCID ID: http://orcid.org/0000-0002-1343-660X

D. F. S. DA SILVA
Ph.D. candidate in the Geology Postgraduate Program at the Federal University of Minas Gerais; Master's in Geotechnics from the Federal University of Ouro Preto; Geological Engineer from the Federal University of Ouro Preto; Technician at the Professor Manoel Teixeira da Costa Research Center/Institute of Geosciences at the Federal University of Minas Gerais. Email: denisefss@ufmg.br
ORCID ID: https://orcid.org/0000-0002-9695-2449

Editor Responsável: Franciolli Araújo

HOLOS
ISSN 1807 - 1600

Submitted March 27, 2023

Accepted December 26, 2023

Published December 28, 2023