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
Universidade Federal de Ouro Preto
ORCID ID: 0000-0003-4302-3897*
allan.santos@ufop.edu.br*

Submetido 27/03/2023 - Aceito 26/12/2023

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 INTRODUÇÃO

Através da interface multidisciplinar entre geologia, geoestatística e processamento mineral é possível propor a criação de um modelo geometalúrgico (Braga e Costa, 2016). A modelagem geometalúrgica permite antever problemas nas etapas posteriores de mineração e tratamento de minérios (Motta, 2014), contribuindo para um melhor planejamento, minimizando riscos do processamento e otimizando os planos de produção nas plantas de beneficiamento (Vieira, et al. 2015). Neste contexto o processo de amostragem corresponde a uma sequência de operações sistemáticas que visam representar, por meio da coleta de uma pequena parcela denominada amostra, um determinado universo. Portanto, pode-se considerar que tal etapa pode é a chave para o sucesso da etapa de exploração mineral.

A amostragem aleatória simples ou estratificada pode ocasionar conglomerados com maior densidade amostral em áreas quando comparadas a outras áreas, para esses casos, a amostragem é dita preferencial. Souza et al. (2001) listam três situações que podem levar à amostragem preferencial em determinadas áreas: condições de acessibilidade, valores de atributos e estratégia de amostragem.

Souza et al. (2001) explicam que quando o banco de dados não contempla uma quantidade suficiente de informações para garantir confiabilidade para a inferência, é necessário realizar o desagrupamento dos dados. Este procedimento consiste em atribuir pesos aos dados, para atenuar ou moderar a influência de dados esparsos. Consequentemente, dados em áreas com amostragem densa podem receber pesos menores do que dados de áreas com amostragem esparsa.

Na prática atual da engenharia, existem métodos de desagrupamento aplicáveis a qualquer conjunto de dados amostrais, o método poligonal (Isaaks e Srivastava, 1989) e o método de desagrupamento de células (Journel, 1983; Deutsch, 1989). Nos métodos, uma combinação linear ponderada é aplicada a todos os valores de amostra disponíveis para estimar a média desagrupada. Esses métodos corrigem o peso dado para amostras em agrupamentos.

O método da poligonal atribui um polígono de influência a cada amostra. As áreas desses polígonos são então usadas como pesos para o desagrupamento. O método de desagrupamento de células usa o conceito de janela móvel para calcular quantas amostras caem em regiões ou células específicas. O peso de desagrupamento atribuído a uma amostra é inversamente proporcional ao número de outras amostras que se enquadram na mesma célula (Isaaks e Srivastava, 1989).

Este artigo apresenta uma avaliação de um desagrupamento utilizando as Self-Organizing Maps (SOMs), com um estudo de caso aplicado ao conjunto de dados Walker Laker (Isaaks e Srivastava, 1989). As SOMs podem ser definidas como uma rede neural de aprendizado não supervisionado, onde a rede busca agrupar os dados de entrada com base em suas semelhanças formando classes. Dessa forma, a partir dos resultados obtidos, podem ser mensuradas evidências iniciais de aplicabilidade como uma ferramenta alternativa aos métodos clássicos a serem utilizados no desagrupamento de dados amostrais.

A maior relevância da metodologia proposta é a definição de áreas de adensamento com margens não lineares. Assim, a metodologia é semelhante ao Método do Polígono na construção da vizinhança, e similar ao método Cell Declustering na construção dos pesos de desagrupamento.

HOLOS
ISSN 1807 - 1600

As SOMs funcionam como uma metodologia intermediária entre os dois métodos clássicos de desagrupamento, utilizando uma técnica atual de inteligência artificial. A partir disso, tem-se a principal justificativa para a escolha das SOMs como método de desagrupamento.

2 MATERIAIS E MÉTODOS

2.1 Banco de dados

A base de dados utilizada como estudo de caso foi o conjunto de dados Walker Lake, de acordo com Isaaks e Srivastava (1989), composto de 470 locais amostrados. O banco de dados do Walker Lake está localizado no distrito nordeste de Nevada e corresponde a um depósito salino e lacustre. A variável V foi utilizada como atributo de interesse. As zonas preferenciais de agrupamento estão concentradas em áreas com alto valor para a variável V (ver Figura 1). A metodologia foi desenvolvida em linguagem R (R Core Team, 2016).

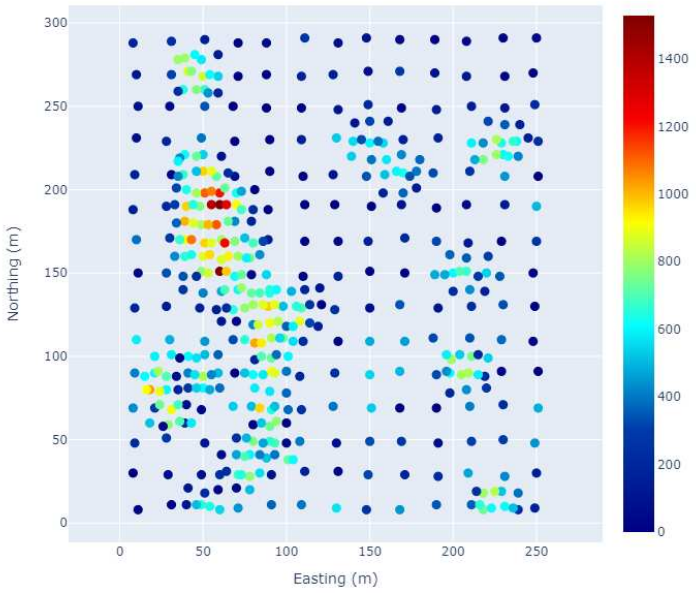

Figura 1: Mapa da variável V.

2.2 Métricas de similaridades

A distância de Manhattan foi aplicada, cuja métrica é tal que a distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas conforme a Equação 1, onde *d(i,j)* é a distância de Manhattan entre as amostras *i* e *j*.

$$d(i,j) = \sqrt{\left(\left|x_{i1} - x_{j1}\right| + \left|x_{i2} - x_{j2}\right| + \cdots + \left|x_{in} - x_{jn}\right|\right)} \quad (1)$$

De acordo com os dados da matriz de distâncias, foi feita uma seleção das 5 menores distâncias obtidas para cada amostra do banco de dados (SOM_A, SOM_B, SOM_C, SOM_D e SOM_E). Uma vez selecionado, foi possível gerar um novo banco de dados com 5 variáveis que representavam, em ordem crescente, as 5 menores distâncias para cada amostra. Essas novas variáveis foram

usadas nas Redes de Kohonen (Kohonen, 1981a; Kohonen, 1981b; Kohonen, 1981c). O cabeçalho do banco de dados utilizado é apresentado na Tabela 1.

Tabela 1: Cabeçalho do banco de dados.

Amostra	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

Para o desenvolvimento da matriz de distâncias, foi utilizado o pacote Rgeos (Interface to Geometry Engine - Open Source) por permitir a criação de diversas aplicações e técnicas para dados espaciais. O pacote Rgeos foi desenvolvido por Bivand e Colin (2017).

2.3 Metodologia geral

As redes de Kohonen foram treinadas no banco de dados de distância e o mapa SOM foi gerado. O algoritmo K-means (MacQueen, 1967) foi aplicado para gerar os grupos. Em cada grupo gerado, a fórmula de Deutsch (1989) foi aplicada para obter os pesos de desagrupamento.

A validação da metodologia foi realizada a partir da comparação dos resultados obtidos com os resultados dos métodos tradicionais de desagrupamento: Cell declustering (Deutsch, 1989) e método da poligonal ou vizinho mais próximo (Cover & Hart, 1967). De acordo com esta comparação foi possível extrair as métricas para avaliação da nova abordagem estudada. A Figura 2 apresenta o fluxograma da metodologia geral.

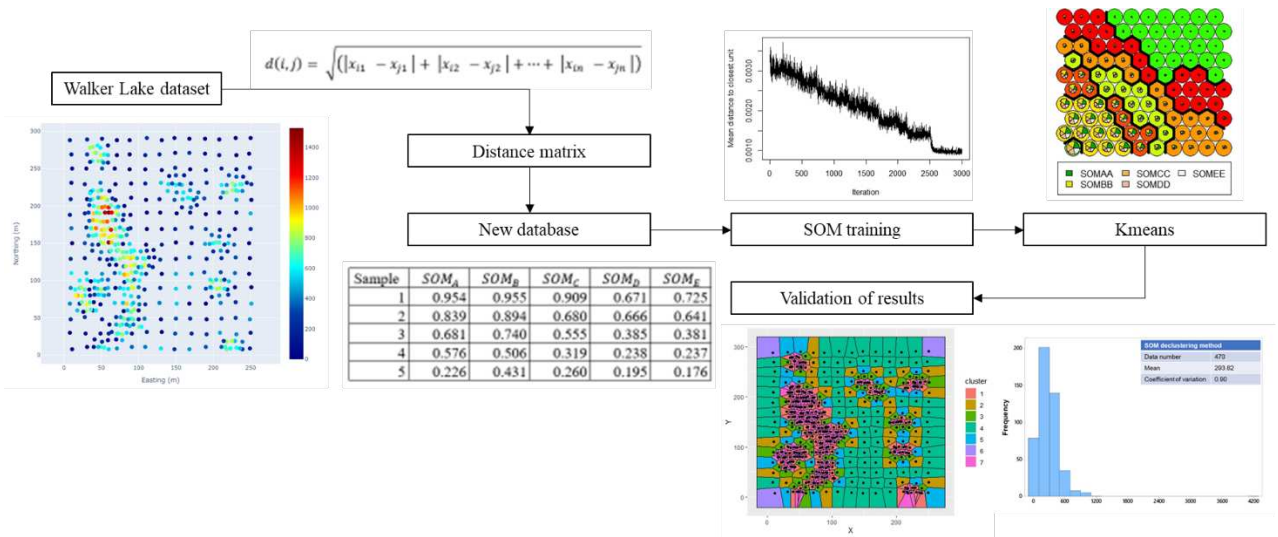

Figura 2: Metodologia geral.

2.4 Implementação do modelo

O pacote Kohonen, desenvolvido por Wehrens e Kruisselbrink (2018), foi utilizado. Este pacote tem a capacidade de implementar diversas formas de SOMs. A base utilizada para o desenvolvimento vem dos estudos desenvolvidos por Kohonen et al. (1995).

O grid criado possui dimensão de 10x10, com neurônios de topologia circular. Essa escolha partiu da premissa de otimizar o tempo de processamento aumentando a segregação das amostras em cada neurônio. Com o grid dessa dimensão, foram criados 100 neurônios como pode ser visto na Figura 3. O treinamento do SOM foi feito com 1000 iterações.

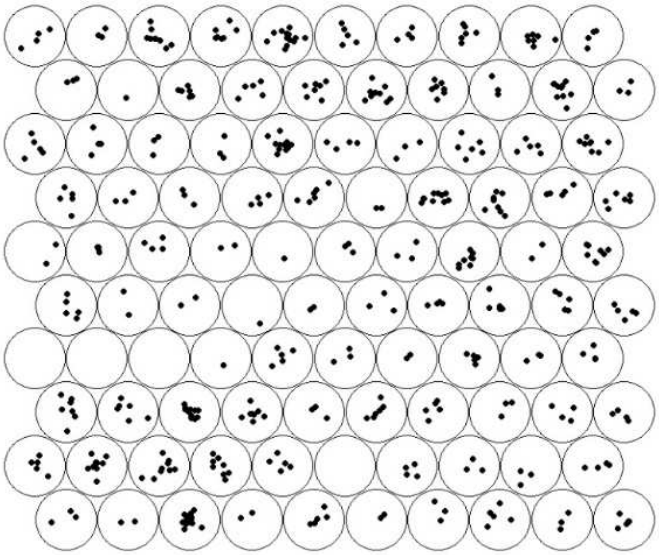

Figura 3: Grid dos neurônios.

Com as amostras classificadas em diferentes neurônios, foi possível agrupá-las e determinar os conjuntos com maior similaridade. Os dados de entrada eram neurônios criados pela SOM e as informações de saída eram agrupamentos criados pelo K-means. O algoritmo usado para o K-means foi Hartigan-Wong (1979) (Hartigan & Wong, 1979). Além disso, outros parâmetros usados no K-means foram: 50 iterações e 1000 inícios aleatórias para minimizar a variação dos resultados de saída obtidos. Para determinar o número de grupos, foi aplicado o método Elbow.

Após a obtenção do número de grupos, o resultado obtido determinou quais neurônios, e assim, quais amostras, seriam agrupados em cada grupo, gerando o agrupamento final. Com os grupos determinados, eles foram submetidos às equações desenvolvidas por Deutsch (1989) para cálculo dos pesos. As Equações 2 e 3 representam os cálculos utilizados para determinar os pesos e as médias não agrupadas, respectivamente, conforme Deutsch (1989).

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

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

Onde w_i é o peso das amostras no grupo, n_l é o número de grupos, l_o o número de amostras no grupo, $\bar{Z}$ a média desagrupada e z_i uma amostra *i* do grupo de dados.

2.5 Repositório de códigos para reproduzir a metodologia aplicada

O repositório com os códigos pode ser encontrado no GitHub a partir do link: https://github.com/MrColugo/Kohonen-Self-Organizing-Maps-applied-to-declustering-in-preferential-sampling

3 RESULTADOS E DISCUSSÕES

A Figura 4 apresenta o número de amostras selecionadas para cada neurônio no grid criado. O ideal dessa etapa é que não haja muitos neurônios vazios ou neurônios sobrecarregados, o balanceamento é necessário para que não haja excesso de processamento desnecessário ou baixa taxa de processamento, respectivamente as condições citadas acima. O resultado obtido foi satisfatório, demonstrando a eficiência na seleção do grid de neurônios desenvolvido.

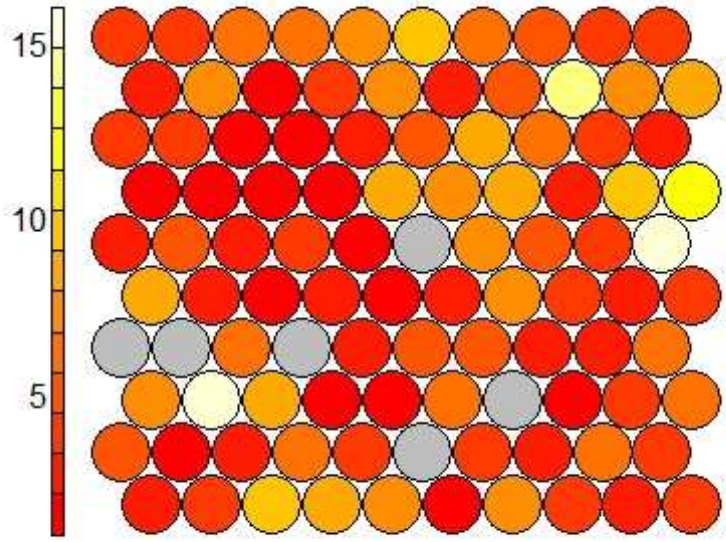

Figura 4: Densidade de neurônios em amostras do grid de Walker Lake.

A partir da Figura 5 pode-se verificar a representação das variáveis em cada neurônio por um gráfico de pizza, onde cada fatia representa uma variável e quanto maior o raio da fatia de pizza, maior a faixa de valores aceitáveis naquele neurônio.

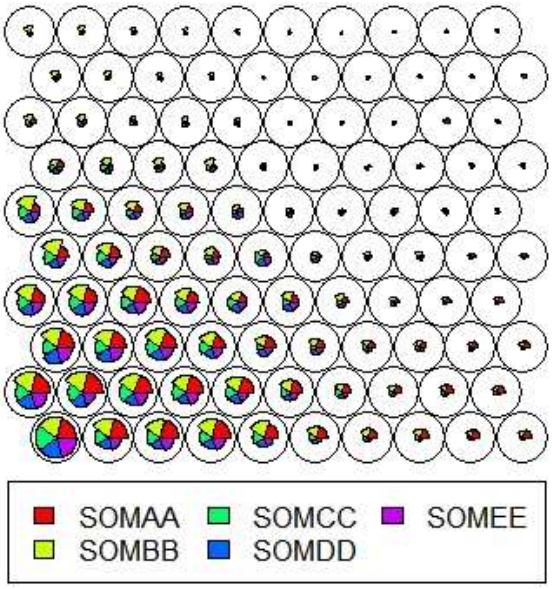

Figura 5: Relação das variáveis nos neurônios.

O gráfico de treinamento do modelo das redes SOMs é apresentado na Figura 6. O gráfico de treinamento representa a distância média entre as amostras de entrada dentro da rede. Quanto menor essa distância, melhor a qualidade da modelagem. O número de iterações entre as amostras determina o quanto esse processo de "aproximação" deve ser feito, quando esse processo se estabilizar, recomenda-se que o desenvolvimento da rede termine pois não há mais ganhos de aprendizado neste ponto.

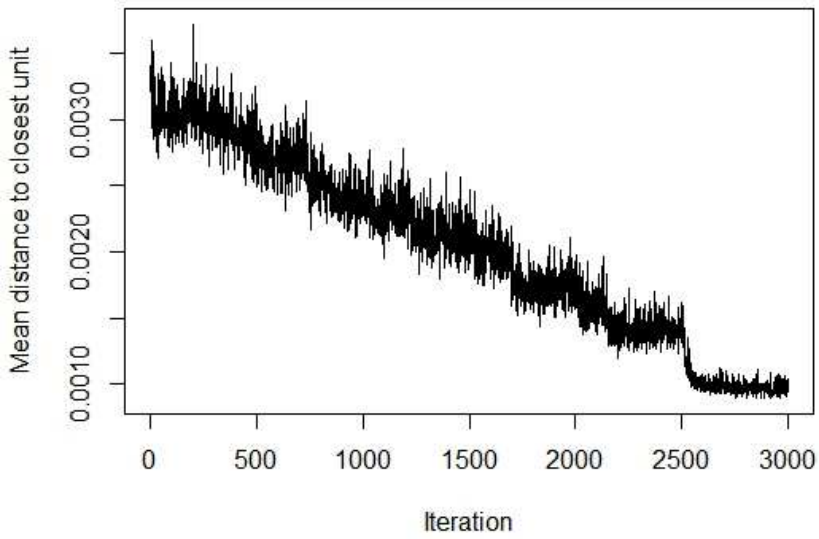

Figura 6: Treinamento das redes SOMs.

Conforme pode ser observado na Figura 6, o resultado obtido para o treinamento foi satisfatório em 3000 iterações, atingindo todas as metas de desempenho necessárias para a criação do modelo SOM.

A seleção do número ótimo de grupos (k) pode ser visualizado na Figura 7. A curva representa o erro médio de K-means em função do número de grupos. Quando a diminuição desse erro se estabiliza, o número de grupos é escolhido.

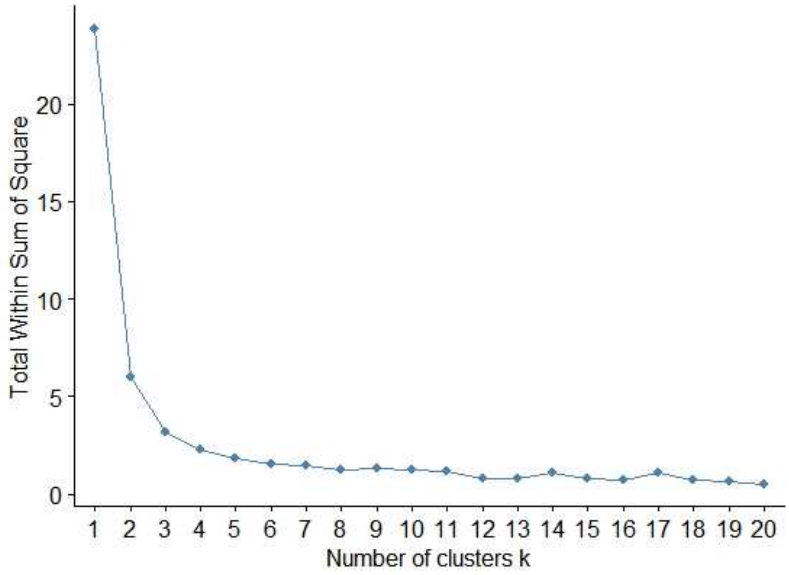

Figura 7: Gráfico de Elbow para k ideal.

A partir dos parâmetros otimizados selecionou-se 7 grupos para segregar neurônios com eficiência, como pode ser visto na Figura 8. A Figura 9 apresenta os resultados dos grupos nos dados amostrados espacialmente. A representação da Figura 9 funciona como uma validação visual e, a partir dela pode-se observar que as redes SOMs permitiram discriminar as zonas com agrupamento.

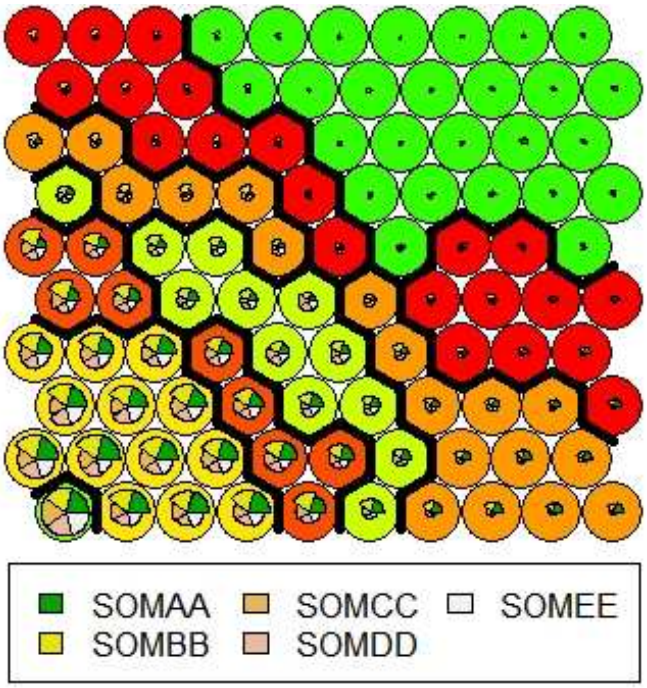

Figura 8: Seleção de grupos nos neurônios.

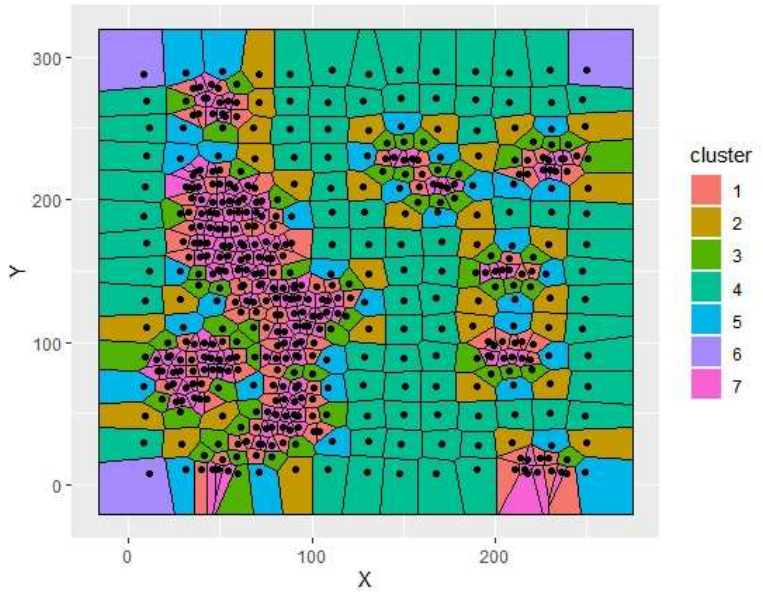

Figura 9. Representação dos grupos de modelos na variável V do banco de dados Walker Lake.

A representação do histograma com os dados desagrupados é apresentada na Figura 10. Observa-se uma aproximação com o histograma de dados exaustivos (Figura 11), quando comparado ao histograma das amostras (Figura 12). Observa-se que o método de desagrupamento trouxe o coeficiente de variação mais próximo da realidade. O coeficiente de variação das amostras foi de 0.69, o que indica menor variância devido ao agrupamento dos dados, o que não é verdade, ao observar os dados reais com coeficiente de variação igual a 0.89.

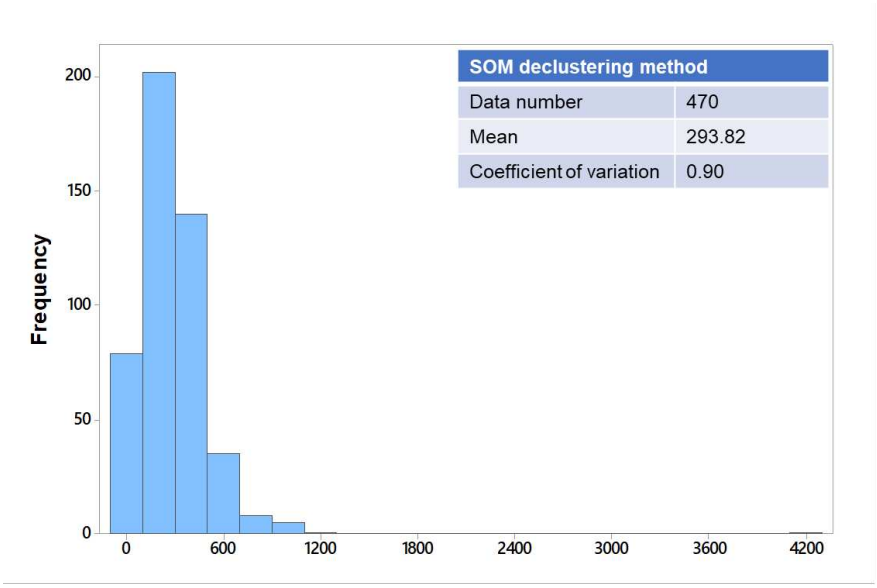

Figura 10: Histograma dos dados desagrupados.

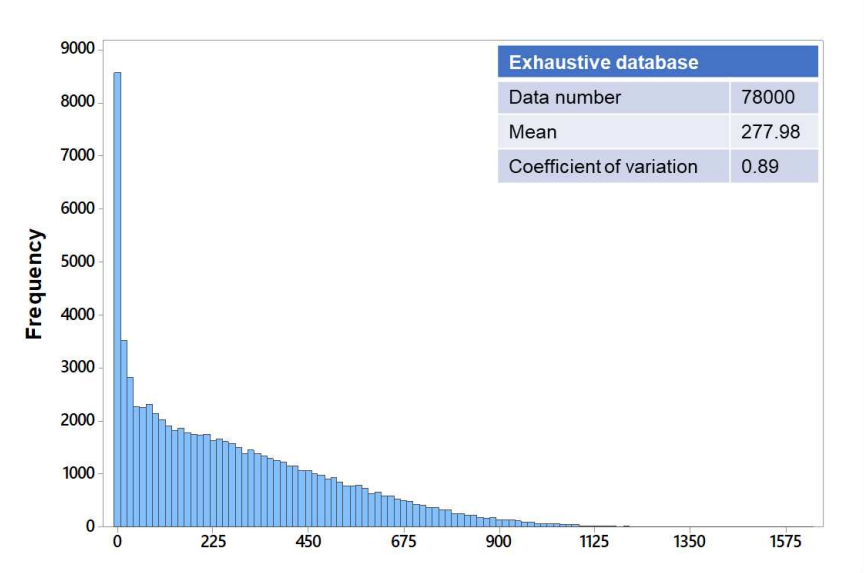

Figura 11: Histograma dos dados exaustivos.

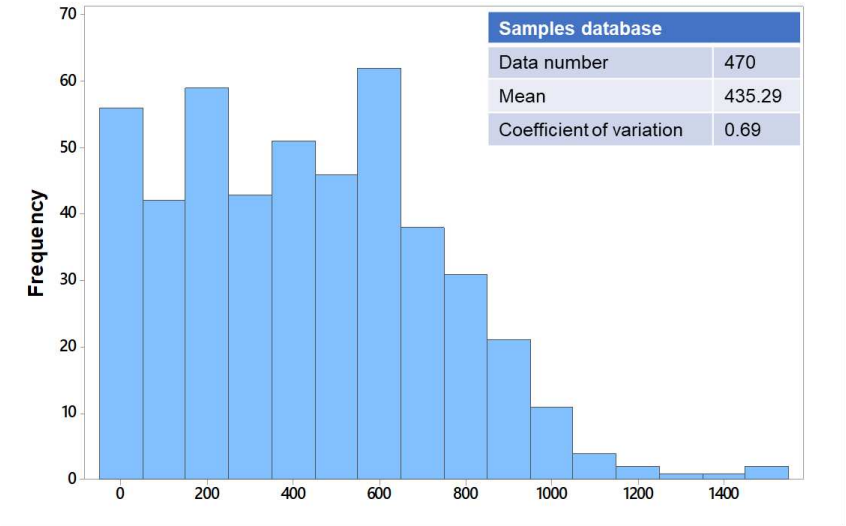

Figura 12: Histograma dos dados amostrados.

A média obtida pelo desagrupamento SOM foi de 293.82, conforme mostra a Figura 10. A média obtida pelos métodos clássicos de desagrupamento foi de 290.11 para o método de desagrupamento de células e pelo método poligonal de 282.5 (Tabela 2). Esse resultado mostra que o desagrupamento das redes SOMs se aproximou do desagrupamento de células. Embora a média não tenha sido menor, o método se apresenta como uma ferramenta e pode ser mais desenvolvido futuramente.

Tabela 2: Média dos métodos.

Método	Média
SOM	293,82
Cell declustering	290.11
Método da poligonal	282.51

4 CONCLUSÕES

Uma proposta de desagrupamento usando as redes SOMs foi aplicada e estudada neste artigo. A variável V do banco de dados Walker Lake foi usada como estudo de caso.

A lógica de ponderação é semelhante ao método Cell Declustering, mas a delimitação das áreas adensadas é diferente. A SOM identifica áreas com margens não lineares, ao contrário do método Cell Declustering. Acredita-se que essa seja a principal justificativa para a escolha da SOM como método de desagrupamento.

O resultado obtido foi próximo ao Cell Declustering, comparando-se a média e o coeficiente de variação. A aplicação apresentada é uma apresentação, como um estudo inicial, podendo ser aprimorada a partir da aplicação em estudos de caso de mineração ou ambientais, onde há viés no processo de amostragem.

5 REFERÊNCIAS

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.

COMO CITAR ESTE ARTIGO:

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

SOBRE OS AUTORES

N. K. AYACHE
Engenheiro de Minas pelo Centro Federal de Educação Tecnológica de Minas Gerais; Analista de planejamento estratégico Pleno na Mosaic Fertilizantes. E-mail: naimayache98@gmail.com
ORCID ID: http://orcid.org/0000-0003-3834-6341

A. E. M. SANTOS
Doutor em Engenharia Mineral pela Universidade Federal de Ouro Preto; Mestre em Engenharia Mineral pela Universidade Federal de Ouro Preto; Engenheiro de Minas pela Universidade Federal de Ouro Preto; Professor no Departamento de Engenharia de Minas da Universidade Federal de Ouro Preto. E-mail: allan.santos@ufop.edu.br
ORCID ID: https://orcid.org/0000-0003-4302-3897

A. E. A. NASCIMENTO
Engenheiro de Minas pelo Centro Federal de Educação Tecnológica de Minas Gerais. Engenheiro de Minas na Carbonífera Cambuí. E-mail: arture.alves@gmail.com
ORCID ID: https://orcid.org/0000-0002-2199-4898

S. A. B. DE CASTRO
Doutoranda em Ciências Exatas e Tecnológicas no Programa de Pós-graduação em Ciências Exatas e Tecnológicas da Universidade Federal de Catalão; Mestre em Engenharia de Minas, Metalúrgica e de

Materiais pela Universidade Federal do Rio Grande do Sul; Engenheira Geóloga pela Universidade Federal de Ouro Preto; Professora no Centro Federal de Educação Tecnológica de Minas Gerais. E-mail: silvaniabraga@cefetmg.br
ORCID ID: http://orcid.org/0000-0002-1343-660X

D. F. S. DA SILVA
Doutoranda no Programa de Pós graduação em Geologia na Universidade Federal de Minas Gerais; Mestre em Geotecnia pela Universidade Federal de Ouro Preto; Engenheira Geóloga pela Universidade Federal de Ouro Preto; Técnica do Centro de Pesquisa Professor Manoel Teixeira da Costa/Instituto de Geociências da Universidade Federal de Minas Gerais. E-mail: denisefss@ufmg.br
ORCID ID: https://orcid.org/0000-0002-9695-2449

Editor Responsável: Franciolli Araújo

Recebido 27 de março de 2023

Aceito: 26 de dezembro de 2023

Publicado: 29 de dezembro de 2023