AVALIAÇÃO DE TÉCNICAS DE AGRUPAMENTO PARA DEFINIÇÃO DE DOMÍNIOS ESTACIONÁRIOS COM O AUXÍLIO DE GEOESTATÍSTICA*

Rudi César Comiotto Modena[1]
Gabriel de Castro Moreira[1]
Diego Machado Marques[2]
João Felipe Coimbra Leite Costa[3]

Resumo

A definição de domínios geológicos/geoestatísticos é a primeira etapa na modelagem de recursos minerais. Uma subdivisão adequada desses domínios requer certo conhecimento *a priori* sobre as características geológicas do depósito e pode estar apoiada em uma cuidadosa análise estatística. É de vital importância o agrupamento de dados cujas características sejam semelhantes, a fim de se evitar a mistura de populações estatísticas, definindo assim os chamados domínios estacionários. Buscando auxiliar nesta definição, foram aplicados dois algoritmos de agrupamento não supervisionados: *Otsu* e *K-means*. O primeiro é muito utilizado na segmentação de imagens e realiza uma busca exaustiva para determinar o melhor limite de separação dos dados. Já o *K-means* é um dos mais utilizados em *machine learning*, fazendo o agrupamento com base na análise iterativa de sua distribuição estatística. Entretanto, algoritmos de agrupamento podem apresentar algumas deficiências quando aplicados a dados geológicos, pois estão fundamentados em análises puramente estatísticas, não considerando sua distribuição espacial ou localização dos dados. Determinar a quantidade mais adequada de domínios também pode ser um desafio. Assim, são propostos métodos para se estabelecer o melhor número de grupos com base em análises de variâncias entre/intra grupos, aliados à aplicação da variografia de indicadores para verificar essa definição.

Palavras-chave: Domínios de Estimativa; Estacionariedade; Geoestatística; Agrupamento de Dados.

EVALUATION OF CLUSTERING TECHNIQUES FOR DEFINING STATIONARY DOMAINS SUPPORTED BY GEOSTATISTICS

Abstract

The definition of geological/geostatistical domains is the first step in building a mineral resource model. A suitable division for these domains requires some prior knowledge about the deposit geology and can be supported by a careful statistical analysis. It is crucial that data with similar characteristics are grouped together, to avoid the mixing of statistical populations, defining the so-called stationary domains. In order to assist in this definition, two unsupervised clustering algorithms were applied: Otsu and K-means. The first one is widely used in image segmentation and it is based on the exhaustive search of the data to determine the best threshold for separating them. K-means is one of the most used techniques in machine learning, and it is based on the iterative analysis of the statistical distribution. Clustering algorithms may present some shortcomings when applied to geological data, since they are based on pure statistical analysis, not considering spatial distribution or data location. Choosing the most appropriate number of domains can also be challenging. Some methods for defining the best number of groups are presented, based on the analysis of variances between/inside groups, supported by indicators variography in order to verify this definition.

Keywords: Estimation Domains; Stationarity; Geostatistics; Clustering Algorithms.

[1] *Geólogo, mestrando, Programa de Pós-Graduação em Eng. de Minas, Metalúrgica e de Materiais, PPGE3M, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brasil.*

[2] *Engenheiro de Minas, Doutor, professor, Departamento de Geologia (Instituto de Geociências - IGEO), Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, RS, Brasil.*

[3] *Engenheiro de Minas, Doutor, professor, Departamento de Engenharia de Minas, UFRGS, Porto Alegre, Rio Grande do Sul, Brasil.*

** Contribuição técnica ao 20º Simpósio de Mineração, parte integrante da ABM Week 2019, realizada de 01 a 03 de outubro de 2019, São Paulo, SP, Brasil.*

1 INTRODUÇÃO

A modelagem de variáveis espacialmente correlacionadas resulta da associação entre a componente natural referente às disciplinas das ciências da terra e os fundamentos teóricos da matemática e da estatística, em particular da teoria das funções aleatórias (FA) [1]. A geoestatística [2] é um conjunto de técnicas que tem por objetivo a caracterização da dispersão e a avaliação das incertezas das funções aleatórias que definem os recursos naturais, ou de outros fenômenos espaciais em que os atributos manifestem uma certa estrutura no espaço e/ou no tempo [1].

O conceito de estacionariedade do modelo das funções aleatórias, apesar de ser teoricamente imprescindível para qualquer ato de inferência estatística, não é validável ou refutável *a priori*, uma vez que se conhece uma só realização da função aleatória. No entanto, a geoestatística não interpreta essa limitação como impeditiva, mas parte do princípio de que a hipótese de estacionariedade pode ser julgada apropriada para um determinado conjunto de dados. São os modelos subsequentes de inferência, continuidade espacial e simulação, que devem ser validados *a posteriori* pelo maior ou menor afastamento dos seus resultados em relação à realidade que se pretende modelar [1].

Uma função aleatória estacionária (SRF) é uma representação probabilística de uma propriedade petrofísica com um valor esperado constante e com os momentos de covariância independentes da localização. Porém, raramente os recursos minerais apresentam essas características. Dessa forma, é necessário identificar domínios mais homogêneos para que possam ser mais consistentes com as premissas matemáticas de uma SRF [3].

Na área das geociências, os domínios são definidos como corpos geológicos ou parte destes que apresentem distribuições aproximadamente homogêneas, tanto físicas como químicas. A compreensão das estatísticas dos dados, em conjunto com a geologia, permite a subdivisão do depósito em domínios para estimativa, sendo mais razoável do que utilizar todo depósito como uma unidade. A definição dos domínios depende da disponibilidade de dados, que devem ser suficientes para as etapas posteriores (modelagem, estimativa, simulação geoestatística, etc.). Além disso, os domínios devem ter alguma previsibilidade espacial e não serem excessivamente misturados com outros domínios [4].

Ferramentas estatísticas tais como histogramas, gráficos de probabilidade, *box plots*, gráficos de dispersão, gráficos Q-Q, gráficos de efeito proporcional e variogramas são utilizados para comparações das distribuições de teores dentro de cada um dos domínios identificados. Essa análise estatística apresenta certa subjetividade, uma vez que é necessário definir um grau aceitável de similaridade [4]. Outras técnicas estatísticas multivariadas podem ser utilizadas para descrever a relação entre a geologia e os teores, como a análise de agrupamentos, ou *cluster analysis*, em inglês (e.g. *k-means, hierarchical clustering e Mixed Gaussian Models*)[4].

Foi avaliada a utilização da técnica denominada “Otsu”, introduzida por [5], que é uma das técnicas mais utilizadas e de maior sucesso na segmentação de imagens [6], e se baseia em uma análise exaustiva para encontrar a configuração com a máxima variância entre grupos [6].

Outro método utilizado foi o *k-means* [7]*, que* é um dos mais utilizados algoritmos de agrupamento não supervisionados, cujas principais vantagens são a simplicidade e a velocidade de execução, porém, a utilização de centroides iniciais aleatórios pode gerar resultados consideravelmente diferentes a cada execução do algoritmo, não

garantindo uma solução ótima. Assim, algumas metodologias já foram propostas para melhorar a classificação dos grupos com relação a seleção inicial dos centroides. Uma das alternativas mais utilizadas, e que se encontra implementada na biblioteca de aprendizado de máquina *scikit-learn* é a *k-means*++ [8], onde a posição inicial entre os centroides busca ser a mais afastada possível, melhorando assim os resultados.
Assim, buscando confrontar a aplicação dos algoritmos Otsu e *K-means* para o agrupamento de dados espacialmente posicionados, o presente trabalho utilizou um banco de dados univariado bidimensional, e comparou os resultados obtidos através de análises estatísticas e geoestatísticas. Os domínios gerados tiveram sua conectividade espacial medida através do variograma dos indicadores de cada domínio. O critério básico proposto para avaliar se um determinado número de domínios é adequado, foi verificar se todos os domínios apresentam variogramas dos indicadores bem estruturados. Se algum domínio apresentar comportamento de efeito pepita puro, esse número de domínios, ou um número superior, é inadequado.

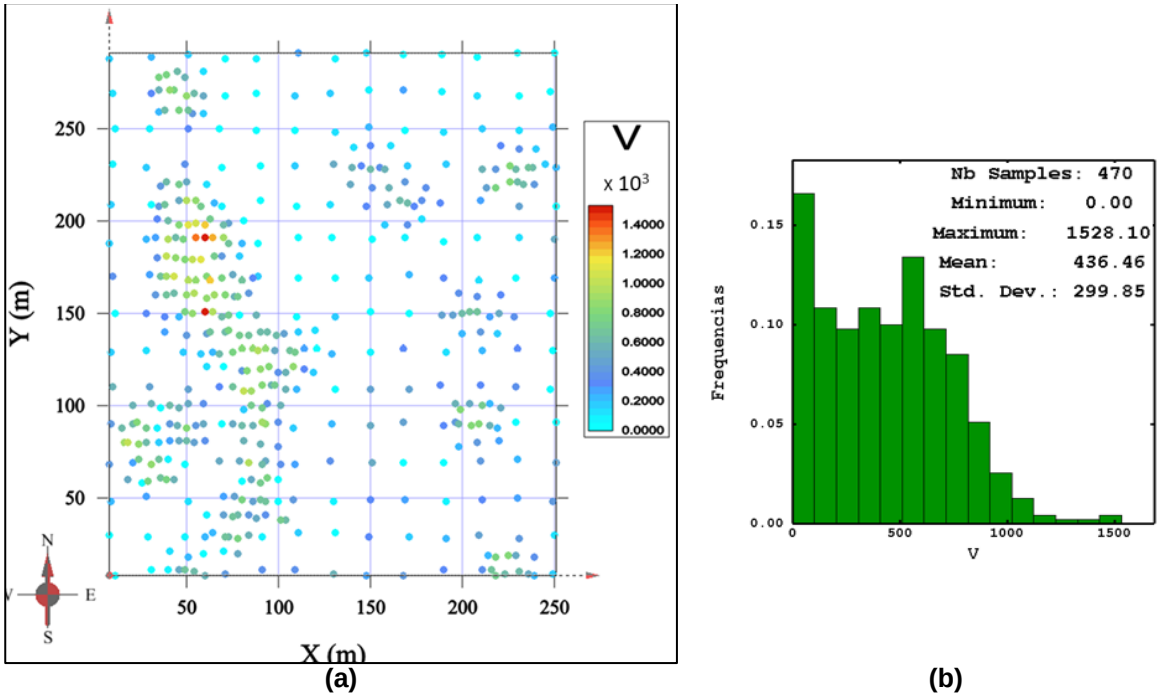
2 DESENVOLVIMENTO

2.1 Materiais e Métodos

Para realização deste trabalho foi implementado o algoritmo Otsu em linguagem Python [5] com algumas modificações, como o armazenamento das variâncias dos grupos, evitando cálculos redundantes.
Já para a execução do *k-means*, foi utilizado o algoritmo disponível na biblioteca de aprendizado de máquina *scikit-learn* e foi usado como parâmetro de inicialização dos centroides a opção "*k-means*++" [8], que busca maximizar a separação entre os centroides.
O estudo de caso deste trabalho foi realizado com o banco de dados *Walker Lake* [9], em 2D e utilizando apenas uma variável, neste caso, a variável V, com 470 dados. Na Figura 1 nota-se a disposição das amostras em uma malha semirregular com adensamento em regiões mais ricas, principalmente na porção oeste da área. A utilização de uma variável apenas foi para facilitar o entendimento dos domínios, mas recomenda-se uma abordagem multivariada quando forem utilizadas técnicas de aprendizado de máquina.
O método Otsu [5] é um método não paramétrico e não supervisionado que busca a melhor separação de grupos utilizando como critério a maximização da variância intergrupos. Como utiliza um algoritmo de busca exaustiva pela melhor solução, é muito efetivo na separação dos grupos [6]; no entanto, é de alta demanda computacional [6].
O *k-means* é um dos mais antigos e mais utilizados algoritmos de agrupamento não supervisionados, tendo sido proposto por [7]. O usuário só precisa informar a quantidade de grupos nos quais deseja dividir os dados, que o algoritmo se encarrega de buscar a melhor segmentação entre eles. As principais etapas do algoritmo são [10]:

1. O usuário seleciona o número de grupos.
2. O algoritmo escolhe aleatoriamente a posição inicial do centroide de cada grupo.
3. Cada valor do banco de dados é classificado em um grupo com base na proximidade a um determinado centroide.

4. A posição dos centroides é atualizada, realizando o cálculo da média de todos os dados pertencentes ao grupo.
5. Os passos 3 e 4 são repetidos até que o algoritmo convirja para uma solução, ou seja, não ocorra mais mudanças na posição dos centroides.



(a) (b)

Figura 1. (a) Mapa exibindo os 470 pontos da variável V do banco de dados *Walker Lake*. (b) Histograma da variável V.

Para seleção da melhor configuração de agrupamentos (quantidade de grupos) foi utilizado o método *Elbow* [11] que consiste na análise visual do gráfico da quantidade de grupos pela variância média entre eles. De acordo com o método, deve ser selecionado o ponto a partir do qual ocorre um desvio mais abrupto na curva. A partir dele, a variância não mais sofre mudanças consideráveis [12]. Existem outros métodos de validação sobre a melhor separação de grupos, como os índices de Calinski-Harabasz, Silhouette e Davies-Bouldin [13], porém, eles não serão usados neste trabalho.
Como critério de avaliação para determinar se o número de domínios selecionados apresenta conectividade espacial, foi usado o variograma dos indicadores. Para aplicação desta técnica, foi criada uma variável binária que assume o valor 1 para as amostras dentro de um determinado domínio e 0 para as demais. Essa variável binária é então variografada.

2.2 Resultados e discussões

Os algoritmos foram aplicados para quatro cenários distintos:

- 2 domínios;
- 3 domínios;
- 4 domínios;
- 5 domínios.

Para avaliar o desempenho computacional de cada um dos algoritmos foi medido o tempo de duração de cada um deles para cada número de domínios (Tabela 1).

Tabela 1. Tempo de execução dos algoritmos para cada cenário

Método	Tempo (s)			
	2 domínios	**3 domínios**	**4 domínios**	**5 domínios**
Otsu	0,0355	1,6521	69,995	9312,2205
K-means	0,0146	0,0185	0,0210	0,0254

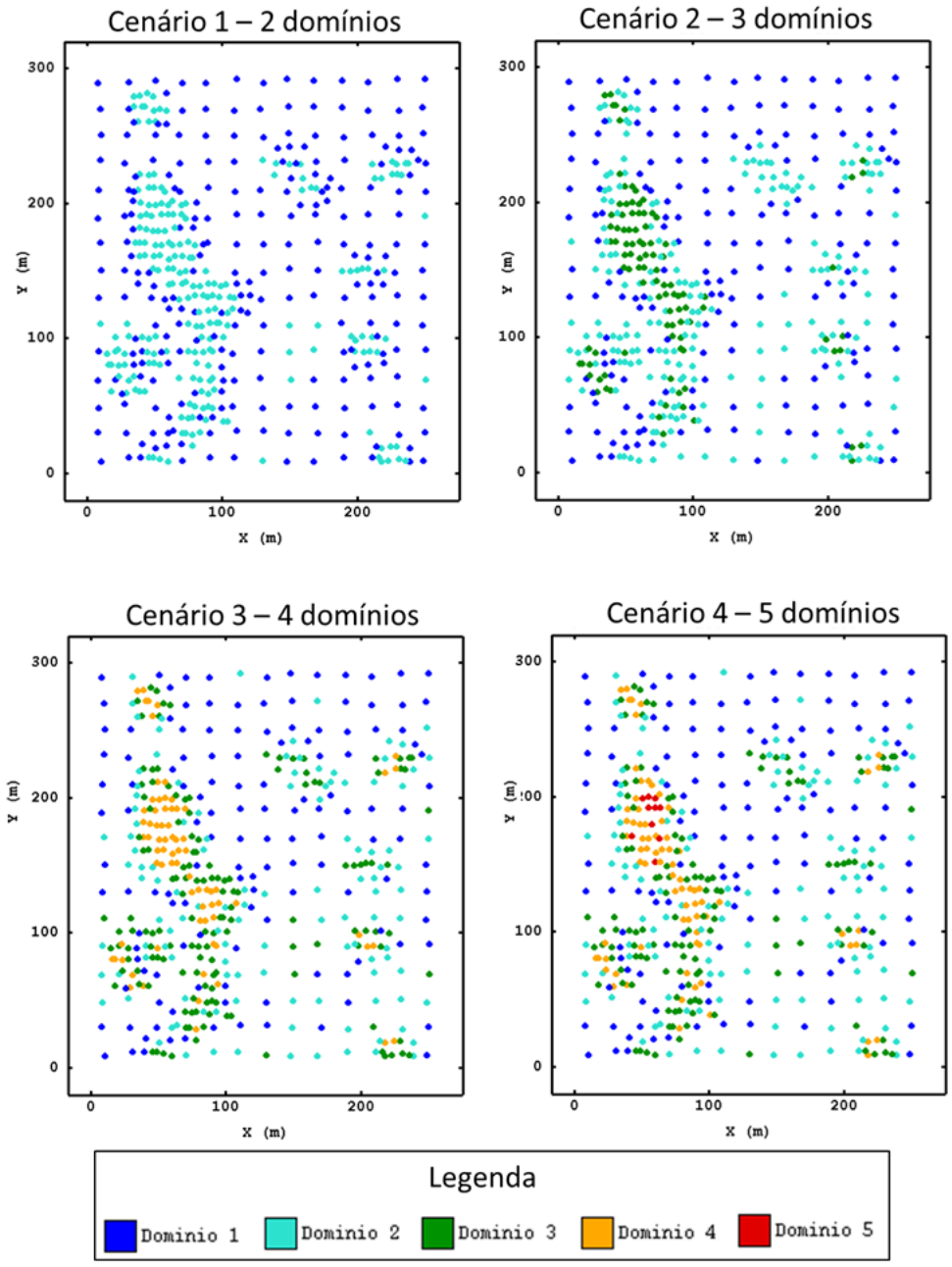


Figura 2. Agrupamentos gerados pelos algoritmos Otsu e *K-means*, cujos resultados são idênticos.

Os resultados de ambos os métodos foram idênticos, mas o tempo de execução de cada algoritmo é substancialmente diferente, especialmente quando cresce o

número de domínios. Deste modo, optou-se por usar os resultados do método *K-means*. A Figura 2 mostra os mapas de localização amostral para as diferentes configurações testadas.
A Figura 3 apresenta a distribuição das amostras nos diferentes cenários testados. Parece haver uma distribuição equilibrada das amostras nos diferentes domínios em cada caso testado.

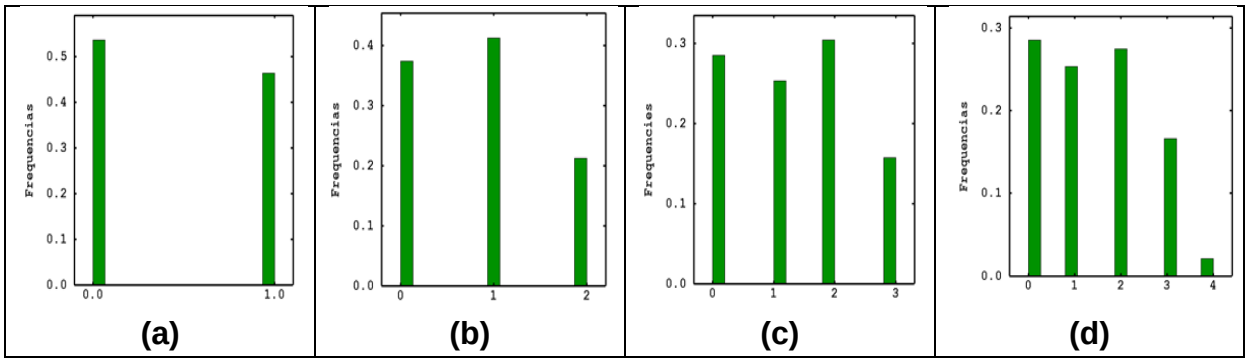
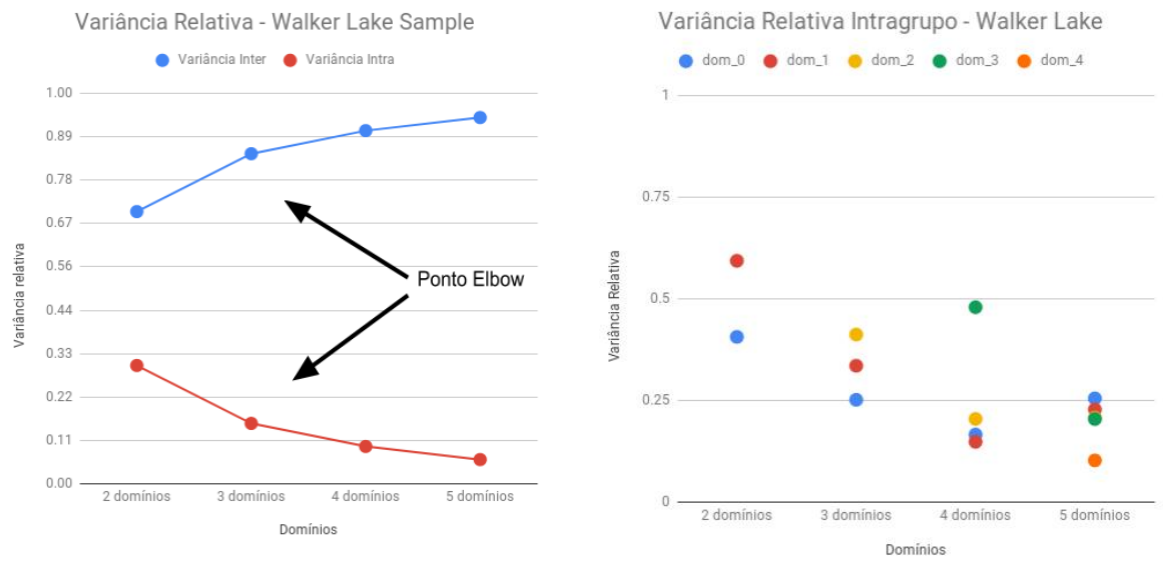


Figura 3. Histogramas de frequências de dados agrupados em cada domínio, para ambos algoritmos, Otsu e *K-means*, cujos resultados são idênticos. (a) Histograma da separação em 2 domínios, (b) 3 domínios, (c) 4 domínios e (d) 5 domínios.

A Figura 4(a) mostra que a partir de 3 domínios, não há alteração significativa da variância entre os domínios e na variância interna dos domínios. Assim, podemos dizer que a escolha de três domínios seria adequada para o caso estudado. Note que as variâncias no gráfico estão relativizadas à variância total dos dados.
O gráfico da Figura 4(b) mostra a variância dos dados dentro de cada grupo (variância intragrupo), a depender do número de domínios (eixo horizontal). Nota-se, que para três domínios, as variâncias intragrupos são mais regulares quando comparadas entre si, ao contrário dos outros cenários, onde há grandes discrepâncias entre elas.
Todos os resultados preliminares mostram que um número de domínios superior a 3 é desnecessário, mas como seria possível garantir que esses domínios apresentem a conectividade espacial característica de fenômenos naturais regionalizados? Uma alternativa encontrada é a utilização do variograma dos indicadores em cada domínio gerado pelo algoritmo de agrupamento.
A variografia dos indicadores foi aplicada aos cenários de dois e três domínios, este último sendo o caso mais adequado segundo a análise estatística preliminar das variâncias intra e intergrupo. Para aplicação da técnica, foi criada uma variável binária, que, para cada cenário, assume o valor 1 para as amostras dentro de um determinado domínio e 0 para as demais. Foi então realizada a análise da continuidade espacial (variografia omnidirecional) desta variável binária. Se existe conectividade espacial dentro de um mesmo domínio, espera-se que o variograma se apresente bem estruturado, com baixo efeito pepita e estruturas bem definidas, até que se atinja um patamar.



(a) **(b)**

Figura 4. (a) Variâncias intragrupos e intergrupos em relação ao número de domínios. As setas indicam o melhor agrupamento escolhido pelo método *Elbow*. (b) Variâncias intragrupos de cada domínio. Os gráficos estão relativizados em relação a variância total.

A Figura 5 mostra os variogramas dos indicadores para os domínios 1 e 2 no cenário de dois domínios, que se apresentam bem estruturados, enquanto a Figura 6 apresenta o mapa variográfico, indicando a direção de maior continuidade do fenômeno, NNW-SSE, que é a direção preferencial de distribuição da variável V, original. Cabe ressaltar que devido ao banco de dados ter sido dividido em dois domínios equivalentes em quantidade de dados, os variogramas do domínio 1 e domínio 2 (Figura 5 e 6) ficaram idênticos.

A Figura 7 mostra os variogramas dos indicadores para os domínios 1, 2 e 3 no cenário de três domínios e, como se pode perceber, os variogramas dos domínios 2 e 3 não se apresentam bem estruturados, ocorrendo, inclusive, efeito pepita puro. A Figura 8 apresenta os mapas variográficos de cada domínio e, a partir deles, percebe-se que apenas o domínio 1 apresenta com clareza a existência de alguma direção preferencial.

Assim, apesar de ter sido o cenário sugerido como o mais adequado pela análise estatística, a segmentação dos dados em 3 domínios não apresenta conectividade espacial satisfatória para os domínios 2 e 3, o que sugere uma possível redundância entre eles.

Desta maneira, o cenário de 2 domínios parece ser o mais adequado neste caso.

(a) (b)

Figura 5. Variogramas omnidirecionais do (a) domínio 1 e (b) domínio 2 no cenário com 2 domínios.

(a) (b)

Figura 6. Mapa variográfico do (a) domínio 1 e (b) domínio 2 no cenário com 2 domínios.

(a) (b) (c)

Figura 7. Variograma omnidirecional do (a) domínio 1, (b) domínio 2 e (c) domínio 3 no cenário com 3 domínios.

(a) (b) (c)

Figura 8. Mapa variográfico do (a) domínio 1, (b) domínio 2 e (c) domínio 3 no cenário com 3 domínios.

Trabalhos tem sido apresentados, com análises multivariadas, para que seja considerada a posição espacial das amostras no próprio processo de definição dos agrupamentos, e não somente na escolha do número mais adequado de grupos. Simplesmente incorporar as coordenadas puras como variáveis em algoritmos de agrupamento [14], acaba gerando domínios geométricos, o que pode não ser adequado. Algumas linhas de pesquisa utilizam estatísticas de autocorrelação local [15] [16], outras aplicam alguma forma de restrição da vizinhança de busca [17] [18] [19]. Outra forma de considerar a conectividade espacial das amostras poderia ser a aplicação da tabela de covariância, que segundo [20], é uma ferramenta geoestatística que realiza um mapeamento rápido e automatizado da continuidade espacial.
Como sugestão para trabalhos futuros, poderia ser interessante criar um fluxo de trabalho interativo, unindo o algoritmo de *K-means* e a tabela de covariância [20], de modo a auxiliar na definição de domínios geológicos ou geoestatísticos com conectividade espacial.

3 CONCLUSÃO

O comparativo entre os algoritmos Otsu e *K-means,* mostrou que os resultados obtidos para ambos são idênticos, porém o tempo de processamento do Otsu é consideravelmente maior, o que dificulta a sua utilização.
Com relação a avaliação da continuidade espacial, foi possível verificar que o método proposto neste estudo, a utilização dos variogramas dos indicadores, permitiu uma escolha adequada da separação dos domínios, levando em consideração tanto a informação estatística como a continuidade espacial. Desta forma, essa técnica se mostra mais indicada para a escolha da quantidade de domínios estacionários do que as puramente estatísticas, como o *Elbow*.
Sugere-se a continuidade dos estudos com algoritmos como o *K-means* 2D, ou o *K-means* 3D, que utilizam, além dos valores da variável em si, a média e a mediana de uma janela móvel, agregando informação espacial. Da mesma forma, é possível utilizar a tabela de covariância, que auxiliaria na determinação da conectividade espacial para definir os domínios geológicos.

Agradecimentos

Os autores gostariam de agradecer ao laboratório de Pesquisa Mineral e Planejamento Mineiro (LPM) da Universidade Federal do Rio Grande do Sul, por fornecer as condições necessárias para o desenvolvimento deste trabalho. À Fundação Luiz Englert (FLE), à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

REFERÊNCIAS

1 Soares A. Geoestatística para as Ciências da Terra e do Ambiente. 2. ed. Lisboa: IST Press, 2006.
2 Matheron G. Traité de Géostatistique Appliquée. Vols 1 and 2. Paris: Technip, 1962-1963.
3 McLennan JA. The Decision of Stationarity. Edmonton. Tese [Doctor of Philosophy] - University of Alberta; 2007.
4 Rossi ME, Deutsch CV. Mineral Resource Estimation. Dordrecht: Springer, 2014.
5 Otsu N. A threshold selection method from gray-level histogram. IEEE Transactions on System Man Cybernetics. 1979;SMC-9(1):62-66.
6 Liu D, Yu J. Otsu method and K means. Ninth International Conference on Hybrid Intelligent Systems, Shenyang. 2009;9:344-349.
7 MacQueen, J Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967;1:281-296.
8 David A, Vassilvitskii S. k-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics. 2007 [acesso em 26 maio 2019]. Disponível em: http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf.
9 Isaaks HE, Srivastava MR. An Introduction to Applied Geostatistics. Oxford: Oxford University Press; 1989.
10 Tan P, Steinbeck M, Kumar V. Introduction to data mining. 2. ed. New York: Pearson Education, 2006.
11 Thorndike RL. Who Belongs in the Family? Psychometrika. 1953;18(4):267-276.
12 Kodinariya TM, Makwana PR. Review on Determining Number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies. 2013;1(6):2321–7782.
13 Aggarwal CC, Reddy CK. Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, 2014.
14 Hundelshaussen R. Otimização de Parâmetros de Krigagem Baseada na Minimização do Erro Absoluto e a Sua Incerteza. Porto Alegre. Tese [Doutorado em Engenharia] – Escola de Engenharia da UFRGS; 2018.
15 Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geographical Analysis. 1995;27(4):286–306.
16 Scrucca L. Clustering multivariate spatial data based on local measures of spatial autocorrelation. Tech. Rep., Universita a degli Studi di Perugia. 2005.[acesso em 26 maio 2019]. Disponível em: https://core.ac.uk/download/pdf/6963987.pdf
17 Oliver MA, Webster R. A geostatistical basis for spatial weighting in multivariate classification. Mathematical Geology. 1989;21(1):15–35.
18 Ambroise C, Govaert G. Convergence of an EM-type algorithm for spatial clustering. Pattern Recognition Letters. 1998;19(10):919–927.
19 Romary T, Rivoirard J, Deraisme J, Quinones C, Freulon X. Domaining by Clustering Multivariate Geostatistical Data. Geostatistics Oslo 2012. Quantitative Geology and Geostatistics, vol 17. Springer, Dordrecht. 2012;455-466.
20 Kloeckner J, Rodrigues AL, Machado PL, Costa, JFCL. Automatização da análise de continuidade espacial via tabela de covariância aplicada a estimativa e simulação de teores. Congresso Brasileiro de Minas a Céu Aberto e Minas Subterrâneas, Belo Horizonte. 2018;1:1-12.