UNIVERSIDADE FEDERAL DE MINAS GERAIS

Departamento de Engenharia de Minas

Especialização em Engenharia de Recursos Minerais

MONOGRAFIA

REVISÃO DO USO DE TÉCNICAS DE AGRUPAMENTO PARA DEFINIÇÃO DE DOMÍNIOS ESTACIONÁRIOS

Aluno: Bernardo Generoso Silveira

Orientador: Prof. Pedro Campos

Belo Horizonte, Julho de 2022

Bernardo Generoso Silveira

Revisão do uso de técnicas de agrupamento para definição de domínios estacionários

Versão Final

Dissertação apresentada ao Curso de Especialização em Engenharia de Recursos Minerais da Universidade Federal de Minas Gerais (UFMG), como requisito parcial para obtenção de título.

Orientador: Professor Pedro Campos

Belo Horizonte

2022

UNIVERSIDADE FEDERAL DE MINAS GERAIS

CURSO DE ENGENHARIA DE RECURSOS MINERAIS

UFMG

ATA DA DEFESA DA MONOGRAFIA DO ALUNO
BERNARDO GENEROSO SILVEIRA

Realizou-se, no dia 31 de agosto de 2022, às 17:00 horas, na Plataforma MEET, da Universidade Federal de Minas Gerais, a defesa da Monografia, intitulada **"Revisão do uso de técnicas de agrupamento para definição de domínios estacionários"** apresentado pelo aluno BERNARDO GENEROSO SILVEIRA, número de registro 2020721010, graduada no curso de ENGENHARIA QUÍMICA, como requisito parcial para a obtenção do certificado de Especialista em ENGENHARIA DE RECURSOS MINERAIS, à seguinte Comissão Examinadora: Prof. Pedro Henrique Alves Campos - Orientador, Prof. Alizeibek Saleimen Nader (Universidade Federal de Minas Gerais), Pedro Benedito Casagrande (Universidade Federal de Minas Gerais).

A Comissão considerou a defesa do artigo:

(X) Aprovada

( ) Reprovada

Nota: 80

Finalizados os trabalhos, lavrei a presente ata que, lida e aprovada, vai assinada por mim e pelos membros da Comissão.

Pedro Henrique Alves Campos

Prof. Pedro Henrique Alves Campos (Mestre)

Risia Magriotis Papini
Coordenadora do Curso de Especialização em Engenharia de Recursos Minerais

Prof. Alizeibek Saleimen Nader (Doutor)

pedro casagrande

Prof. Pedro Benedito Casagrande (Doutor)

Aurea Domingos
Secretaria do Curso de Especialização em Engenharia de Recursos Mineirais

AGRADECIMENTOS

- À minha família por sempre me encorajar e apoiar.
- Ao Prof. Pedro Campos pelo suporte, disponibilidade e incentivo.
- À secretaria do CEERMIN pelo incentivo.
- Aos membros da Banca Examinadora, pela leitura do texto e pelas sugestões oferecidas ao trabalho.

RESUMO

Umas das primeiras etapas na modelagem de recursos minerais é a definição de domínios estacionários. O agrupamento equivocado das amostras para tomada de decisão pode comprometer os próximos passos da modelagem e até mesmo os resultados da estimativa, gerando viés nos valores de massas e teores. A definição de domínios estacionários na maioria das vezes é confundida com a de domínios geológicos, que, além de ser subjetiva, não considera as correlações das amostras nem no espaço multivariado, nem no geográfico. Nesta monografia é feita uma revisão bibliográfica acerca de algoritmos de agrupamento que apresentam resultados promissores e contribuem para a melhor interpretação de estacionariedade do conjunto de dados e sua validação. Técnicas tradicionais de agrupamento de dados estatísticos como algoritmo hierárquico aglomerativo e *k-means* são discutidas, além de técnicas mais recentes de agrupamentos que consideram as posições espaciais das amostras (algoritmo hierárquico geoestatístico e algoritmo de aglomeração por espaço duplo). Por mais que os algoritmos mais recentes considerem a posição espacial amostral, a análise com os resultados dos algoritmos tradicionais faz-se necessária para fins comparativos, já que a validação ainda é uma medida que depende do conhecimento do geomodelador. Uma vez comparados diferentes resultados e validações dos algoritmos, o profissional terá mais embasamento para tomada de decisão mais assertiva sobre estacionariedade. Por mais que o processo possa ser laborioso, a aplicação desses algoritmos garante que os próximos passos da modelagem de recursos não sejam comprometidos, evitando, portanto, retrabalhos ou até mesmo erros significativos da estimativa final.

Palavras Chaves: Geoestatística, domínios estacionários, agrupamento de dados, modelagem geológica, recursos minerais.

ABSTRACT

The definition of stationary domains is one of the first steps in mineral resource modeling. The incorrect grouping of samples can compromise the subsequent steps of modeling and even the estimation results, generating greater uncertainties in masses and grade values. The definition of stationary domains is most often confused with geological domains, which is not only subjective, but it also does not consider the correlations of the samples on multivariate or geographical spaces. This monograph aims to provide a wide bibliographical review about cluster algorithms which present interesting results and contribute for better stationary interpretation of the geostatistical data set and its validation. From traditional grouping techniques for statistical data – such as hierarchical agglomeration algorithm and k-means – to more recent techniques of spatial clusters that consider geographic positions of the samples – geostatistical hierarchical algorithm and double space agglomeration algorithm – all are discussed. As much as spatial algorithms have more elegant applicability and support in geostatistical data, a comparison with the results of traditional algorithms is necessary for comparative purposes, since validation is still a measure that depends on the knowledge of the geomodeler. Once different results and validations of the algorithms are compared, the geomodeler will have more grounding in deciding the most appropriate stationary domains. As laborious as the process can be, the application of these algorithms ensures that the next steps of resource modeling are not compromised, thus avoiding rework or even significant errors in the final estimate.

Key Words: Geostatistics, stationary domains, cluster analysis, geologic model, mineral resources.

Índice de Figuras

Lista de siglas:

VR = Variável regionalizada

H = Entropia Espacial

wcss = Soma dos quadrados intragrupo *("within cluster sum of squares")*

Índice de Equações:

Sumário

1 INTRODUÇÃO

Um projeto de mineração eficaz e lucrativo requer uma caracterização mais fiel possível sobre a geometria e teores da zona mineralizada. Portanto é imprescindível o entendimento para representação da forma, tamanho, qualidade, variabilidade e limites destas zonas. Tal estudo é necessário para as etapas de avaliação de recursos, desenvolvimento de mina e produção (Pereira, 2017).

A geoestatística integra observações de natureza qualitativa e quantitativa de amostras para inferir propriedades do fenômeno espacial desconhecido. Essa ciência se baseia em estudos de distribuição espaciais amostrais para determinar o depósito como um todo (população) e as incertezas associadas (Yamamoto & Landim, 2013). Os valores de teores observados em um depósito mineral não são independentes uns dos outros. A dependência espacial é uma consequência da gênese do depósito mineral e dos processos geológicos subsequentes (Isaaks & Srivastava, 1989). Portanto, o detalhamento geológico na fase da exploração mineral é muito importante para a estimativa de recursos.

A caracterização geológica possui como premissa que as propriedades geológicas variam espacialmente por fatores estratigráficos e estruturais, de acordo com cada depósito. Esses condicionantes requerem uma investigação em superfície e por testemunhos de sondagem a fim de unificar unidades com mesmos atributos geológicos em volumes espaciais irregulares (Silva, 2000). Entretanto, os registros geológicos representam uma fonte complexa de informações que refletem longos períodos de aquisição de dados, envolvendo vários indivíduos e empresas distintas, coletados com possíveis diferentes interpretações, metodologias e tecnologias que mudam ao longo do tempo.

O agrupamento de amostras semelhantes para definição de domínios é a etapa inicial para a modelagem de recursos minerais. Através deste agrupamento, são criados sólidos que limitam os diferentes domínios para que a correlação entre as amostras da mesma população não tenha viés e não comprometa a estimativa (domínios estacionários).

O conceito de domínio geológico se confunde com o conceito de domínio estimativo. O domínio geológico é comumente descrito com apenas uma variável geológica, que está associado naturalmente com o processo de mineralização do depósito mineral, com

características geológicas uniformes. Por outro lado, o domínio estimativo ou estacionário é definido por um conjunto de dados sobre o controle da mineralização e pode conter mais de um domínio geológico. Portanto, um domínio estimativo deve ser definido baseado no conhecimento geológico, mas também suportado por uma profunda análise exploratória dos dados e variografia. Essa definição faz referência às zonas estacionárias dentro do depósito mineral, ou seja, é uma decisão de como agrupar informações da mesma população dentro de uma zona específica do depósito, separados por limites, ou inclusas no depósito como um todo (Rossi & Deustch, 2014).

É evidente que o agrupamento de amostras deva ser orientado pelas características geológicas do depósito, mas essa análise é subjetiva e, quando aplicada sozinha, não garante estacionariedade. Em muitos casos, os domínios geológicos se diferem em parâmetros de grau, como média e intervalo de valores, além de intervalos espaciais, como variabilidade e continuidade espacial. As consequências de misturar diferentes populações dentro de um mesmo domínio são capazes de comprometer significativamente os resultados de um modelamento geoestatístico (Sinclair & Blackwell, 2004).

Segundo Martin & Boisvert (2016), as técnicas convencionais para definição de domínios estacionários são interpretativas sobre o banco de dados de descrições geológicas ou atribuindo limites de teores dentro das unidades geológicas. Contudo, a definição de domínios de estimativas deve também levar em consideração a continuidade espacial entre as diferentes variáveis amostradas, sejam teores e/ou descrições geológicas (Rossi & Deustch, 2014). Algoritmos de análise de agrupamento *("cluster analysis")* têm sido cada vez mais utilizados para reconhecer padrões em dados mutivariados (Moreira, 2020). Um problema comum, no entanto, é que algoritmos tradicionais (aglomerativo hierárquico e *k-means*) são frequentemente usados para classificar as relações com base em parâmetros estatísticos, sem considerar os geológicos (Rossi & Deustch, 2014). Mais recentemente, técnicas têm sido desenvolvidas para tratar especificamente da análise de agrupamento de dados cuja posição no espaço é de grande relevância (Romary, *et al*., 2012; Martin & Boisvest, 2018)

Parâmetros de entrada, como número de grupos, métricas para conexões entre os grupos e o método para integrar a correlação espacial devem ser definidos pelo usuário, e são indispensáveis para o sucesso dos algoritmos de agrupamento. A aplicação e os resultados desses métodos são bastante subjetivos e por isso há a necessidade de parametrização e

validação de qualquer método aplicado (Moreira, 2020). Entretanto, validações quantitativas ainda foram pouco desenvolvidas na literatura para dados geoestatísticos (Martin, 2019)

Uma abordagem qualitativa foi sugerida por Martin & Boisvert (2018), considerando a estatística dos dados e suas respectivas posições espaciais. Os autores supracitados aplicam o conceito do algoritmo *wcss (within cluster sum of squares*, ou soma dos quadrados intragrupo) e entropia espacial para medir a relação dos grupos no espaço multivariado junto de sua relação espacial.

Portanto, a decisão de estacionariedade envolvem diferentes técnicas além do agrupamento por semelhanças apenas geológicas, que também consideram a posição espacial das amostras. Logo, cabe ao geomodelador analisar quais são mais aplicáveis ao conjunto de dados de forma a diminuir a subjetividade desta decisão.

2 OBJETIVO

Recentemente, metodologias têm sido propostas para análise de agrupamento de dados, levando em consideração a correlação espacial entre as amostras, além dos padrões multivariados. As aplicações dessas técnicas podem levar a construção de modelos de recursos mais coerentes com a realidade.

Este trabalho possui como objetivo a revisão bibliográfica de métodos de agrupamento de dados tradicional, hierárquico e *k-means*, além dos métodos de agrupamento mais recentes, que consideram a correlação espacial dos dados, como por restrição por vizinhança e hierárquico aglomerativo geoestatístico.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Estacionariedade

O conceito de estacionariedade é relacionado com homogeneidade geológica, seja ele pela gênese do depósito ou por alterações posteriores. Um domínio é considerado estacionário se em qualquer resultado de amostragem demonstra a mesma população, independentemente da localização, não existindo tendência nos dados. Entendemos melhor quando consideramos como uma suposição modelável a partir da análise geoestatística dos dados disponíveis, mais

especificamente, de momentos de primeira e segunda ordem, isto é, média, covariância e variância (Sinclair & Blackwell, 2004).

Toda jazida mineral possui sua história geológica desde sua gênese até os possíveis processos geológicos posteriores que resultaram na mineralização, como dobras, metamorfismo e lixiviação. Esses fenômenos naturais podem ser caracterizados por uma distribuição espacial de grandezas mensuráveis dependentes do espaço tridimensional, definida como "variáveis regionalizadas" (VR). Um exemplo de uma VR é a distribuição de teores no espaço e a sua forte relação de dependência de um ponto $Z(x)$ separado por outro a uma certa distância vizinha $Z(x+h)$. Essa correlação depende do módulo e da direção do vetor h, além das características singulares geológicas de cada tipo de depósito mineral. Levando em consideração a errática variabilidade espacial, anisotropias e descontinuidades espaciais, o estudo direto da função $Z(x)$ se torna muito complicado. Contudo, interpretar esses valores individualmente é inaceitável, tendo em vista que existe uma correlação espacial entre eles. Portanto, para se obter uma solução consistente que suporte tanto o contexto estruturado, quanto o aleatório da VR, se dá pela interpretação probabilística chamada de funções aleatórias (Journel & Huijbrets, 1978)

Isso significa que a inferência da continuidade espacial de uma VR pode ser feita utilizando a estatística de dois pontos como base. Ao aplicar definições das funções covariância e função variograma, verifica-se que elas dependem apenas de dois pontos (x_1) e (x_{1+h}), então cada par de pontos é considerado uma realização diferente, o que torna possível a inferência estatística dessas funções (Journel & Huijbrets, 1978)

Segundo Soares (2006), é impossível determinar as estatísticas (média e variância) no ponto x_1 dessa função com uma única realização. Portanto, o autor defende assumir diversos graus de estacionariedade da função aleatória, como por exemplo que as VR tenham a mesma média. Esta é considerada:

$$E[Z(x_1)] = E[Z(x_2)] = \cdots = E[Z(x_n)] = E[Z(x)] = m$$

Equação 3.1: Diversos graus de estacionariedade com média igual.

Onde $E[Z(x_1)]$ é a esperança matemática da *função* $Z(x)$ e *m* é a média.

Logo, a média m passa a ser independente da localização e obtida como média aritmética das realizações das variáveis aleatórias (Soares, 2006):

$$E\{Z(x)\} = m(x)$$

Equação 3.2: Esperança matemática igual à média independente da localização.

Onde $E\{Z(x)\}$ é a esperança matemática (ou primeira ordem) da função $Z(x)$ e $m(x)$ é a média do conjunto de pontos x.

Entretanto, assumir que essa hipótese é correta significa supor que a média das amostras seja representativa ao ponto de abranger toda a área estudada, assumindo a homogeneidade do depósito mineral (Soares, 2006). Contudo, sabe-se que na natureza isso raramente ocorre, impulsionando a verificação da variabilidade espacial da função aleatória. A variância da média é calculada a partir da fórmula:

$$Var\{Z(x)\} = E\{[Z(x) - m(x)]^2\}$$

Equação 3.3: Variância entre os dados.

Onde Var é a variância a priori de $Z(x)$, definida pela esperança matemática da diferença quadrática entre o valor de $Z(x)$ e sua média $m(x)$

Ao levarmos em consideração a distância do vetor *h* entre as VR, deve-se considerar a covariância entre a mesma variável, obtidos em pontos distintos pela distância *h*, assumindo uma certa direção. Logo, ao modificar a direção, a covariância também não será a mesma, indicando anisotropia no conjunto de dados. Portanto, a covariância é função da direção e distância entre os pares de pontos, sendo definida por:

$$C(x_1, x_2) = E\{[Z(x_1 - m(x_1)][Z(x_2) - m(x_2)]\}$$

Equação 3.4: Covariância entre os dados.

Onde, $C(x_1, x_2)$ é a covariância entre as amostras no espaço amostral x_1 e x_2.

Por fim, o variograma, que é uma função definida como a variância do incremento $\{Z(x_1) - Z(x_2)\}$, pode ser escrito como:

$$2\gamma(x_1, x_2) = Var\{Z(x_1) - Z(x_2)\}$$

Equação 3.5: Variograma.

O semivariograma é denominado pela função $\gamma(x_1, x_2)$.

Para determinar o modelo de correlação espacial da variável regionalizada (VR), calcula-se experimentalmente a correlação usando os pontos vizinhos e, em seguida, ajusta-se um modelo teórico através de uma função. Esse modelo teórico permite determinar o valor da correlação espacial para qualquer distância dentro do espaço amostrado (Yamamoto & Landim, 2013).

3.2 Definição de domínios estacionários

Em um cenário ideal, o depósito mineral como um todo seria homogêneo ao ponto de considerar as mesmas médias, variâncias e covariâncias. Contudo, na maior parte das vezes, cada porção da jazida mineral apresenta propriedades únicas, que refletem na gênese e nas alterações posteriores da mineralização. Dessa forma, a estacionariedade é uma variável subjetiva, que depende da experiência do geoestatístico (Moreira, 2020).

Rossi & Deutsch (2014) defendem que em uma estimativa de recursos bem suportada, deve-se combinar as análises geológicas e estatísticas para definir os domínios estacionários. O conceito é baseado na descrição e modelagem das relações entre as diferentes variáveis geológicas. O resultado é uma matriz que classifica os controles de teores identificados pelos dados e suportados pelo conhecimento geológico. Os mesmos autores recomendam a validação inicial dos domínios através de ferramentas de estatística clássica, como histogramas, scartterplots (gráfico de dispersão), QQ plots e variogramas.

Segundo McLennan (2007), as seguintes decisões prévias devem ser tomadas para a eficiência na estimativa de recursos:

I. Definir o número e os tipos de domínios nos quais serão modeladas as propriedades de interesse;
II. Modelar os limites desses domínios;
III. Quantificar a variáveis de interesses desses domínios;
IV. Quantificar tendências determinísticas de grande escala dentro dos domínios;
V. Ajustar um modelo teórico que permita determinar o valor da correlação espacial para qualquer distância dentro do espaço amostrado.

Martin (2017) descreve alguns critérios que são mais comumente aplicados na indústria para definição de domínios estacionários (Figura 3.1):

- *Definição geológica de domínios estacionários*

Dependendo do tipo de jazida mineral, o controle da distribuição espacial da mineralização está fortemente ligado às unidades geológicas e pode naturalmente definir-se como um domínio estacionário. Para isso, as litologias devem ser evidentemente distintas tanto no mapeamento de superfície, quanto nas descrições dos furos de sondagem, e exige a experiência do geológico de campo para dominar o conhecimento sobre a gênese e os processos que ocorreram na sequência da mineralização, podendo resultar em múltiplos estilos de alterações e mineralizações sobrepostas.

- *Domínios estacionários com mais de uma litologia*

A definição de domínio depende da disponibilidade de dados suficientes para inferir parâmetros estatísticos de forma confiável dentro de cada domínio (Rossi & Deustch, 2014)

Na prática, podem ocorrer muitas unidades litológicas com pouca representatividade nos dados, ocasionando um problema de suporte na descrição da população daquele domínio. Pode acontecer também que o depósito apresenta uma quantidade exagerada de litologias e torna a modelagem pouco prática.

Logo, é comum unir diferentes litologias com o mesmo contexto geológico a fim de corroborar com a praticidade da modelagem e aumentar o número de amostras no domínio estacionário (Martin, 2019).

- *Domínios estacionários por intervalos de teor*

Mesmo após um longo tratamento da correlação entre os dados e o contexto geológico, algumas variáveis são agrupadas mesmo com uma diferença evidente que a separariam em diferentes domínios. Isso acontece devido às limitações práticas, como considerações metalúrgicas e econômicas (Rossi & Deustch, 2014).

Os teores de corte, ou *"cut-offs"*, influenciam na definição de domínios estacionários e os intervalos de teores são definidos especificamente para cada projeto (Martin, 2019).

- *Análise de agrupamento ou "cluster analysis"*

A análise de agrupamento é realizada por algoritmos computacionais e separa um conjunto de dados multivariados em grupos com padrões semelhantes entre si (Martin, 2019).

Contudo, sabe-se que os dados geoestatísticos possuem uma relação entre si e com sua posição espacial. Os algoritmos tradicionais (aglomerativo hierárquico e *k-means*) são frequentemente usados para classificar as relações apenas com base em parâmetros estatísticos, sem considerar os geológicos (Rossi & Deustch, 2014)

Recentemente, metodologias têm sido propostas para análise de agrupamento de dados, levando em consideração a correlação espacial entre as amostras, além dos padrões multivariados (Moreira, 2020)

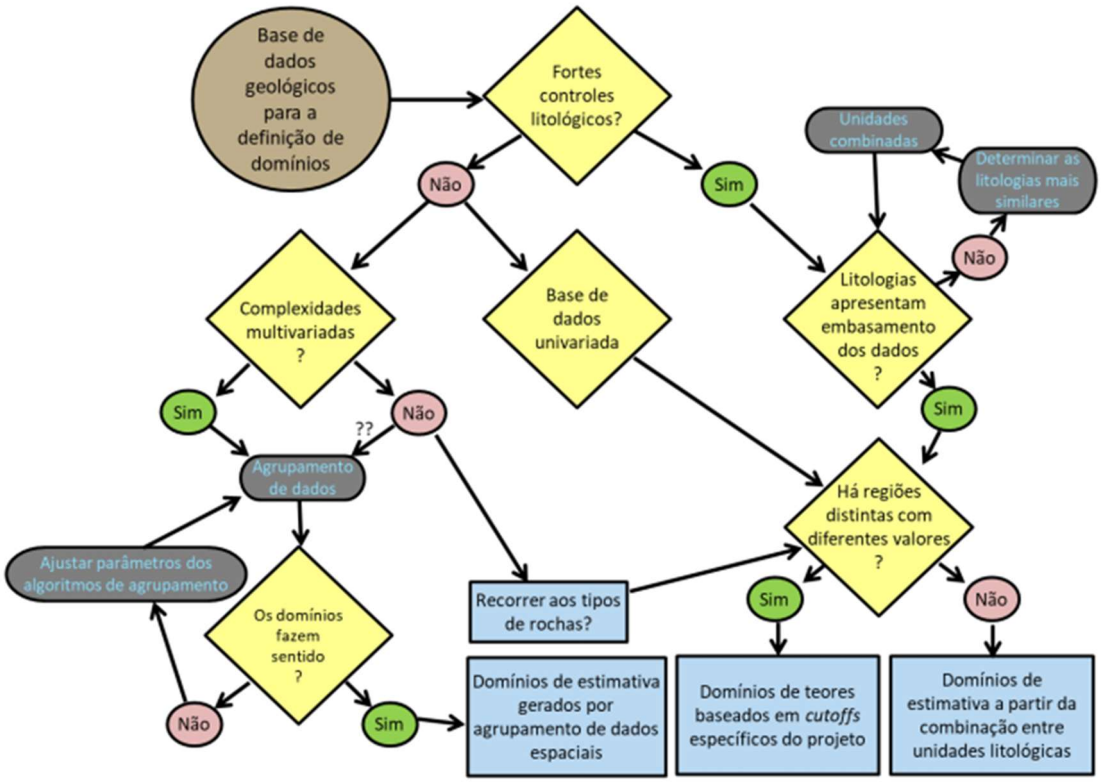

Figura 3.1: Fluxo de definição de domínios estacionários, levando em consideração análises estatísticas multivariadas e geológicas espaciais. Fonte: Moreira (2020) adaptado de Martin (2019).

3.3 Análise de agrupamento ou *"cluster analysis"*

Análise de agrupamento separa grupos de dados ou objetos utilizando unicamente a informação encontrada no banco de dados e descreve a relação entre os dados. O agrupamento de dados é utilizado por várias ciências, pois essa ferramenta eficiente é capaz de trazer um melhor entendimento sobre variáveis abstratas dos dados, sumarizá-las e encontrar grupos semelhantes de forma automatizada (Tan *et al.,* 2006)

A análise de agrupamento ou *"cluster analyses"* é um tema inserido no contexto de aprendizado de máquina não supervisionado, no qual são encontrados padrões em dados que não apresentam classificação prévia. Em outras palavras, os algoritmos de aprendizado não supervisionado são capazes de encontrar semelhanças não evidentes, e, a partir disso, relacionam e rotulam grupos de dados (Moreira, 2020)

Contudo, a noção de agrupamento pode não ser muito bem definida. Para melhor entender a dificuldade de decidir o que constitui um *"cluster"*, considere a Figura 3.2, que mostra vinte pontos e três maneiras diferentes de dividi-los em subgrupos. No entanto, a aparente divisão de cada um dos grupos maiores em subgrupos pode ser simplesmente uma questão humana. Além disso, pode ser não é razoável dizer que os pontos formam quatro agrupamentos, como mostra a Figura 3.2 (C), por exemplo. Portanto, essa figura ilustra que a definição de um cluster é imprecisa e que a melhor definição depende da natureza dos dados e dos resultados desejados.

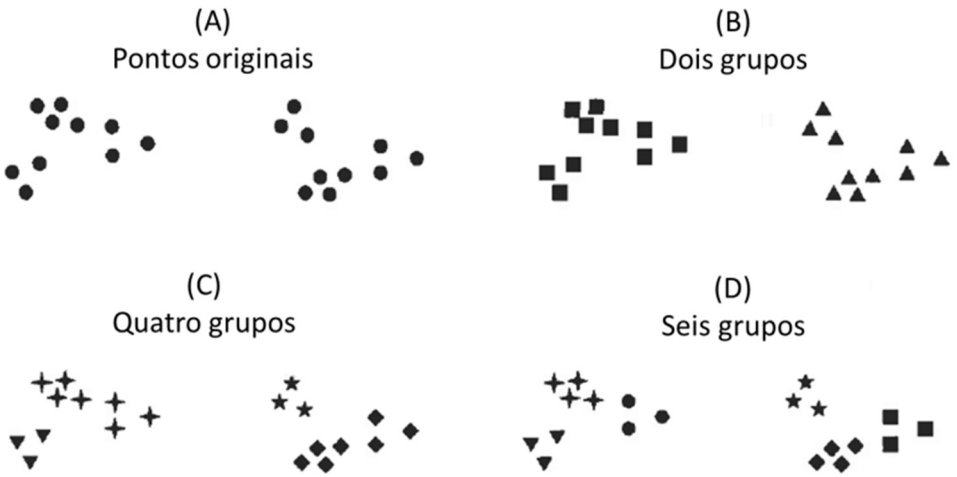

Figura 3.2: Dados originais ilustrados por (A) e diferentes maneiras de agrupá-lo. Em (B), os pontos são divididos em 2 grandes grupos. Em (C) e (D), duas maneiras de subdividir o subgrupo (B). Fonte: Tan *at al.,* (2006).

Basicamente, o grupo é um conjunto de dados em que cada dado é mais similar a um outro que pertence ao mesmo grupo, do que algum outro que não pertence ao grupo. Existem diversos tipos de algoritmos de agrupamento e cabe ao usuário escolher qual adapta melhor ao objetivo da análise do conjunto de dados em questão (Tan *et al.,* 2006)

Devido a essa imensa gama de alternativas, neste trabalho foram escolhidos os que ilustram o desenvolvimento das aplicações na indústria mineira para definição de domínios estacionários. Dois são considerados métodos tradicionais por considerarem apenas análises estatísticas do espaço multivariado: aglomerativo hierárquico (Sokal & Sneath, 1963) e *k-means* (Macqueen, 1967) comumente utilizado na indústria para auxiliar os geomodeladores a tomar

decisões sobre os domínios geológicos. Outros dois consideram a continuidade espacial entre os dados geoestatísticos e sua aplicabilidade aparenta ser promissora para a rotina de uma estimativa de recursos mais aderente à realidade. São eles: agrupamento espacial por restrição de vizinhança (Martin & Boisvert, 2018) e aglomerativo hierárquico geoestatístico (Romary *et al.,* 2012)

3.4 Validação dos métodos de agrupamento de dados

O objetivo do agrupamento espacial de dados geoestatísticos para definição de domínio estacionário deve ser minimizar o erro de uma validação cruzada de uma modelagem geoestatística completa. Em outras palavras, a validação do agrupamento em um contexto geoestatístico deve incluir, de alguma forma, uma medida de quão bem o valor real foi pVRisto em um local não amostrado (Martin & Boisvert, 2018). Esse processo é necessário, uma vez que as consequências são comprometedoras no resultado de massas e teores do modelo geológico (Rossi & Deustch, 2014)

Tanto nos métodos de agrupamentos de dados estatísticos tradicionais, quanto nos métodos de agrupamento espaciais, parâmetros de entrada devem ser definidos pelo usuário, como número de grupos, métricas para conexões entre os grupos e o método para integrar a correlação espacial, que são indispensáveis para o sucesso dos algoritmos de agrupamento. A aplicação e os resultados desses métodos são bastante subjetivos e por isso há a necessidade de validação de qualquer método aplicado (Moreira, 2020). Todos os algoritmos de classificação não supervisionados requerem algum conhecimento de domínio para justificar os resultados e garantir que as classes resultantes sejam razoáveis. Além disso, os resultados de dois agrupamentos de dados com parâmetros distintos podem ser significativamente diferentes (Martin & Boisvert, 2018).

No entanto, a validação de técnicas de agrupamento espacial é uma questão ainda não muito desenvolvida (Martin & Boisvert, 2018) . Em casos de aprendizado de máquina supervisionado, a avaliação do resultado é direta por existir um rótulo comparativo. Os métodos de agrupamento de amostras para definição de domínios estacionários não é supervisionado e se torna um problema avaliar o resultado gerado, levando em consideração que não se conhece um gabarito (Moreira, 2020). Logo, algum conhecimento do domínio é esperado para que os resultados gerados possam ser avaliados (Romary *et al.,* 2012).

Uma abordagem qualitativa foi sugerida por Martin & Boisvert (2018), considerando a estatística dos dados e suas respectivas posições espaciais. Sumariamente, dois critérios são considerados para aferir a eficiência do agrupamento e serão descritos em seguida:

(i) A classificação de populações no espaço multivariado *(wcss)*.

(ii) A continuidade espacial no plano cartesiano (entropia espacial);

O algoritmo calcula cada métrica individualmente, mas a análise deve ser feita de forma simultânea para concluir sobre o agrupamento. O primeiro cálculo é baseado pela soma dos quadrados intragrupos *("within cluster sum of squares" – wcss)* que mede a distância entre as populações no espaço multivariado. O valor mais baixo de *wcss* indica uma maior coesão entre os dados dentro de cada domínio (Figura 3.3):

$$\boldsymbol{wcss} = \sum_{\boldsymbol{k=1}}^{\boldsymbol{k}} \sum_{\boldsymbol{x_i \epsilon k_k}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{M}} (\boldsymbol{x_{ij}} - \boldsymbol{x_{kj}})^2$$

Equação 3.6: Algoritmo *(wcss).*

Onde: *M* representa as variáveis, *k* os grupos, *i* as amostras e *j* as posições. Ou seja, $(x_{ij} - x_{kj})$ representa a distância entre uma determinada amostra e o centróide da distribuição multivariada de seu respectivo grupo.

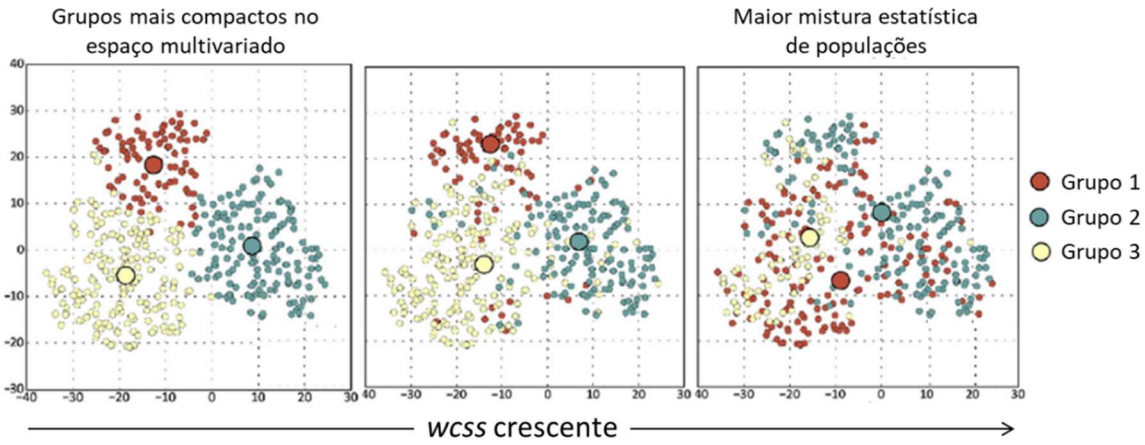

Figura 3.3: Diferentes configurações de agrupamento no espaço multivariado de acordo com o crescimento da soma dos quadrados das distâncias intragrupo *(wcss).* Quanto mais coesos os grupos, menor é o valor de *wcss*. Fonte: Adaptado de Martin & Boisvert (2018)

Em seguida, é necessário medir a interconectividade dos grupos no espaço multivariado, levando em consideração a continuidade espacial no plano cartesiano. Para isso, mede-se a entropia espacial, onde os menores valores representam maior aderência espacial entre os grupos (Figura 3.4):

$$H_{total} = \sum_{i=1}^{N} \sum_{k=1}^{K} p_{i,k} \, ln \, p_{i,k}$$

Equação 3.7: Algoritmo de entropia espacial *(H).*

Onde, $p_{i,k}$ é a probabilidade de se encontrar outra amostra da categoria ou grupo *k* nos arredores de localizações no espaço geográfico *i* e *N* é o número de amostras.

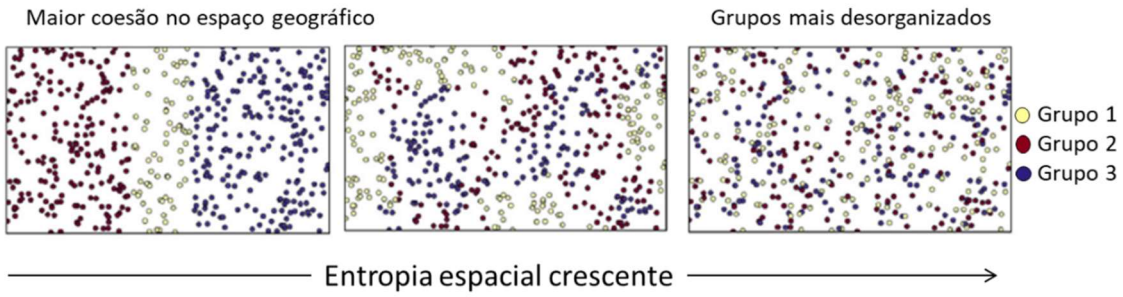

Figura 3.4: Diferentes configurações de agrupamento no espaço geográfico de acordo com o crescimento da entropia espacial (H). Quanto maior a coesão entre os grupos, menor é o valor de H. Fonte: Adaptado de Martin & Boisvert (2018)

No entanto, a Figura 3.5 mostra uma maior coesão no espaço multivariado *(wcss)* implica em um menor valor deste indicador e, consequentemente, um maior valor de entropia espacial *(H).* A combinação dessas medidas fornece um método mais aderente para escolher um agrupamento de maneira mais objetiva do que apenas com uma inspeção visual. Para um conjunto de dados cuja ordenação espacial é mais evidente, sugere-se um *wcss* mais baixo ou com alta coesão no espaço multivariado (Martin & Boisvert, 2018). Segundo Moreira (2020), a melhor configuração para o agrupamento de dados espaciais é aquela que apresenta valores intermediários (Figura 3.5).

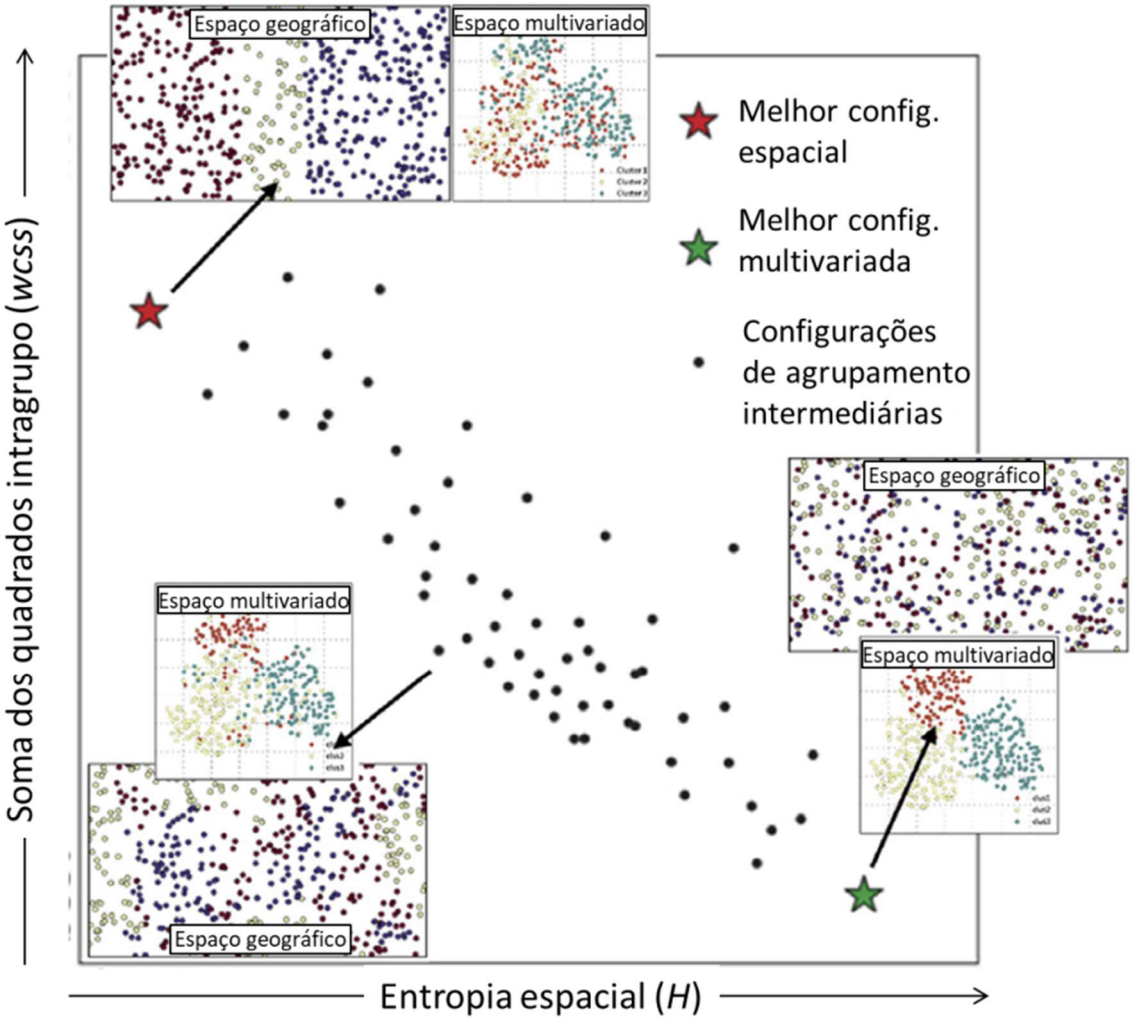

Figura 3.5: Valores de *wcss* e *H* para diferentes configurações de agrupamentos, plotados em um gráfico de dispersão, evidenciando sua relação inversa. Fonte: Adaptado de Martin & Boisvert (2018)

3.5 Métodos tradicionais de agrupamento de dados

Segundo Rossi & Deustch (2014) os algoritmos tradicionais são frequentemente usados para classificar as relações com base em parâmetros estatísticos, sem considerar as relações espaciais geológicas.

Teoricamente, esses métodos podem ser aplicados em qualquer tipo de conjunto de dados e tendem a produzir grupos coesos no espaço multivariado com fronteiras bem definidas no espaço de dispersão (Moreira, 2020)

Contudo, o mesmo conjunto de dados pode fornecer agrupamentos distintos quando aplicados diferentes métodos. Pedregosa *et al.* (2011) apresenta os diferentes resultados propostos pelos diferentes algoritmos de agrupamentos, aplicados para um mesmo banco de dados (Figura 3.6).

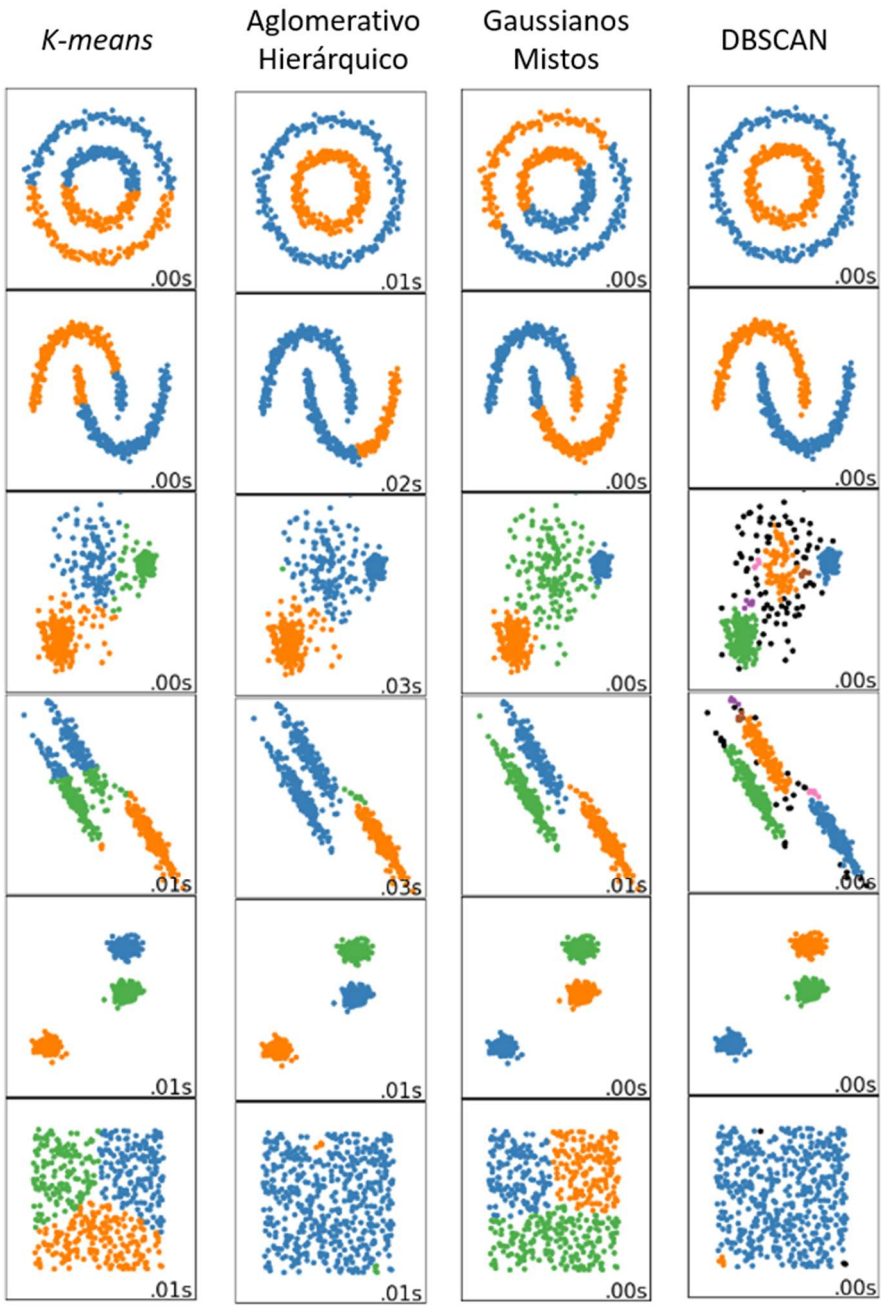

Figura 3.6: Comparativo de diferentes algoritmos de agrupamento tradicionais da biblioteca scikit-learn. Da esquerda para a direita, os algoritmos são apresentados respetivamente: k-

means, aglomerativo hierárquico, gaussiano misto, DBSCAN. Fonte: Adaptado de (Pedregosa, 2011).

Neste trabalho, serão discutidos de forma mais aprofundada os métodos tradicionais comumente aplicados na indústria de mineração: *k-means* e aglomerativo hierárquico.

3.5.1 Método de agrupamento hierárquico

O método de agrupamento hierárquico, segundo Tan *et al.* (2006), é representado graficamente através de nós que representam as conexões entre os dados. Caso os dados não se conectarem por algum nó, significa que não pertencem ao mesmo grupo (Figura 3.7). Em especial, método aglomerativo hierárquico foi originalmente desenvolvido na área das ciências biológicas, para definir os grupos de organismos com base em suas características comuns (taxonomia) (Sokal & Sneath, 1963). Este é um método tradicional já amplamente utilizado em diversas áreas do conhecimento, inclusive para definição de domínios estacionários.

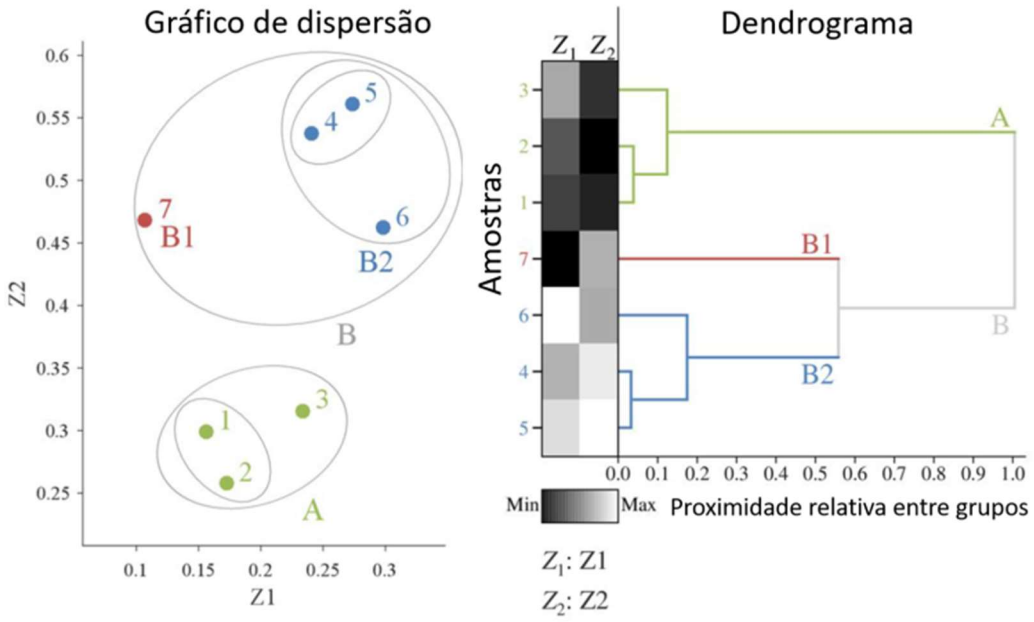

Figura 3.7: Método aglomerativo hierárquico. Gráfico de dispersão das variáveis (Z1 e Z2) para sete amostras à esquerda e seu respectivo dendrograma a direita. O algoritmo agrupou as amostras em 3 diferentes grupos: A, em verde. B1, em vermelho e, por fim, B2 em azul. Observe que a matriz de escala de cinza colorida ao longo do eixo y do dendrograma é um

recurso opcional que exibe o valor relativo das variáveis associadas a cada observação. Fonte: Adaptado de Martin & Boisvert (2018)

O algoritmo começa com os pontos como grupos individuais e em cada passo, se junta com o par de grupo mais similar, baseado na métrica de similaridade escolhida. As junções continuam até que reste apenas um único grupo restante no topo. O resultado é uma matriz de similaridade, base para a construção do gráfico chamado dendrograma (Figura 3.8) que revela os grupos e subgrupos formados através dos dados originais (Martin, 2019).

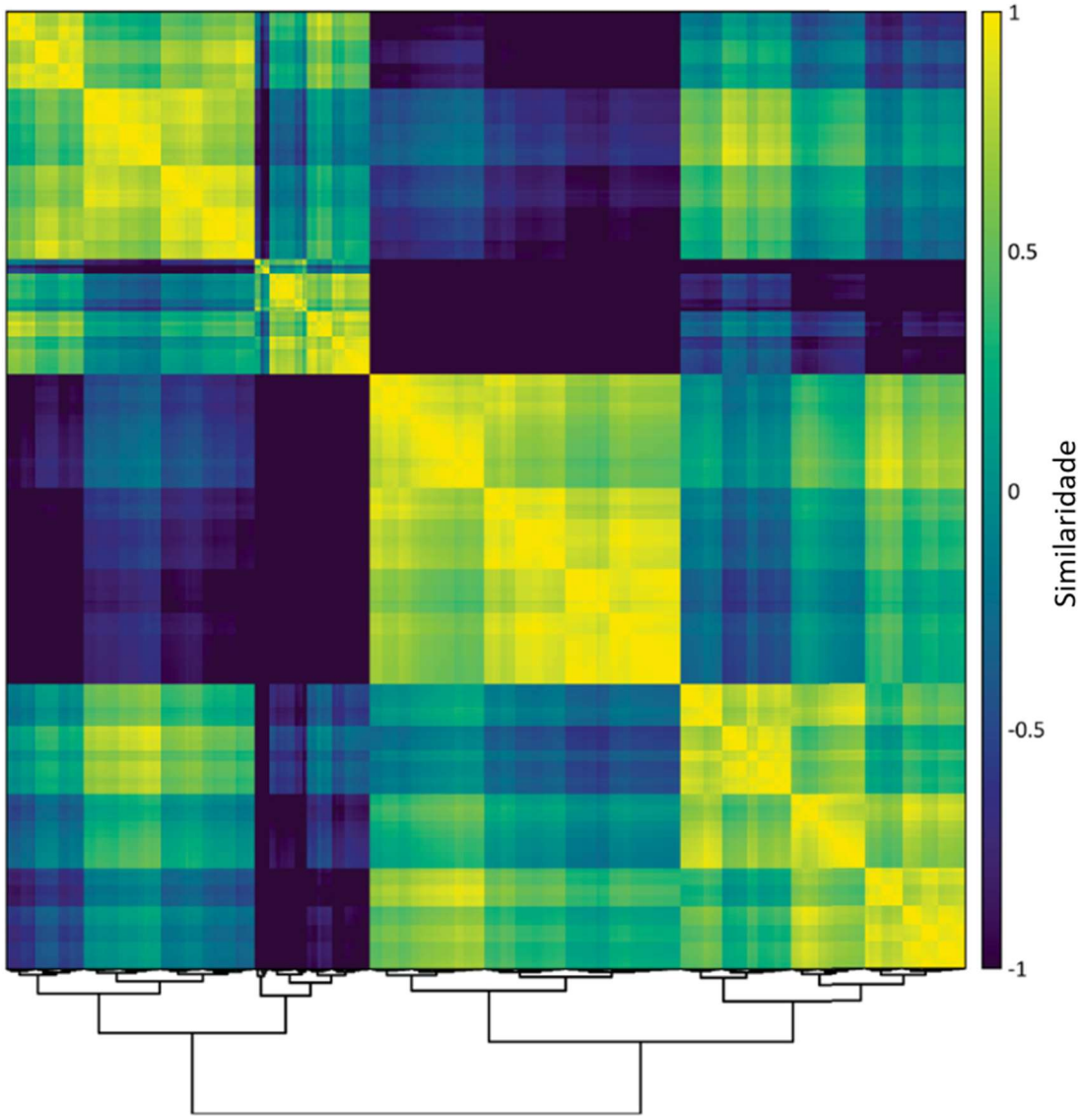

Figura 3.8: Método aglomerativo hierárquico: matriz de similaridade (computa o número de vezes que cada local *i* se encontra no mesmo grupo que cada local *j*) exagerada propositalmente com seu respectivo dendrograma na porção inferior. Fonte: Martin (2019)

Como subproduto, o dendrograma registra a distância relativa entre grupos misturados em cada nível hierárquico do agrupamento (Martin, 2019). Tan *et al.* (2006), descreve que o critério (métrica) aplicado para a definir a distância necessária para que os dados sejam aglomerados, se baseia em uma noção gráfica de proximidade, que pode ser definida pela distância mínima, distância máxima, média grupal ou *Ward* (Figura 3.9).

(A) A distância mínima ou "*minimun link*" é definida entre os pontos mais próximos que se encontram em grupos diferentes. Deve ser considerado onde a distância que existe entre as observações é julgada como importante para a separação de clusters. Essa métrica é muito sensível a ruídos e valores extremos (Tan, *et al.,* 2006; Barnett & Deustch, 2015) (Figura 3.9- A).

(B) A distância máxima ou "*complete link*" é definida pela mais longa distância entre dois pontos de diferentes grupos. Este método é menos sensível a valores extremos do que o "*minimun link*", mas geralmente resulta em formas globulares (Tan, *et al.,* 2006; Barnett & Deustch, 2015) (Figura 3.9 - B).

(C) A média grupal ou "*group average*" é definida pela média das distâncias de cada ponto em um grupo com todos os outros de outro grupo, par a par. Essa métrica é considerada uma opção intermediária em termos de balanceamento de características de "*complete link*" e "*minimun link*" (Tan, *et al.,* 2006; Barnett & Deustch, 2015) (Figura 3.9- C).

(D) No método "*Ward"*, cada grupo é representado por um centróide e a medida da distância entre os grupos é feita em termos do aumento da soma dos erros ao quadrado que resulta na aglomeração desses grupos. Essa métrica é muito semelhante ao método *k-means.*

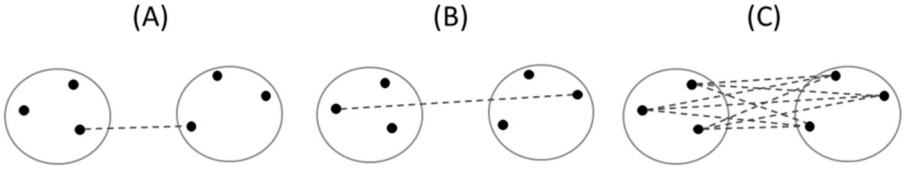

Figura 3.9: Representação gráfica das distâncias das métricas de proximidade. Cada ponto representa uma amostra e as circunferências, os grupos. (A) Distância mínima. (B) Distância máxima. (C) Média grupal. (Adaptado de Tan et al., 2006)

As quatro métricas de agrupamentos hierárquicos serão demonstrados a seguir, onde Barnett & Deutsch (2015) agruparam primeiramente um conjunto de dados tradicional estatístico (Figura 3.10) e posteriormente foram utilizados dados geoestatísticos ou que possuem correlação espacial (Figura 3.11).

Para o primeiro banco de dados os autores utilizaram as latitudes e longitudes das cidades mundiais como variáveis e definiram 4 grupos para o agrupamento (ilustrados pelas cores roxo, verde, azul e vermelho). É evidente a sensibilidade a valores ruidosos no *"minimun link"*, que classificou a cidade Auckland em um grupo diferente de seu país Nova Zelândia. O *"complete link"* revela a maior parte dos grupos em formatos globulares, enquanto o *"group average"* é o resultado intermediário esperado quando comparado aos outros métodos. Por fim, o método de *"Ward"* é o único que divide a América do Norte e do Sul (Figura 3.10).

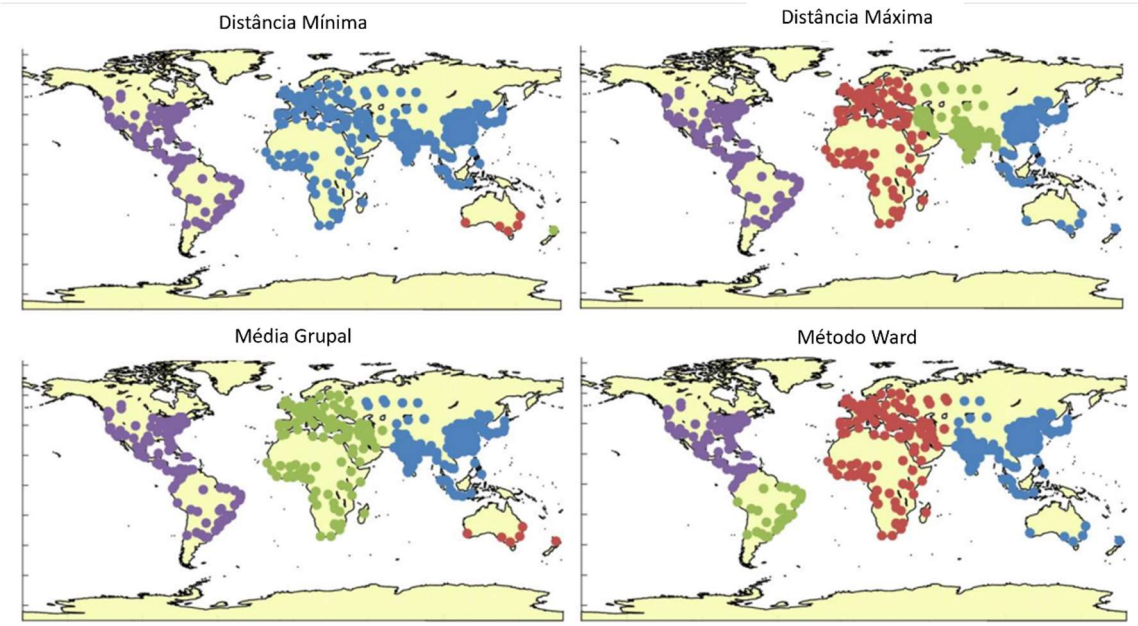

Figura 3.10: Aplicação do algoritmo de agrupamento hierárquico com diferentes métricas de aplicações: minimun link ou distância mínima, complete link ou distância máxima, group average ou média grupal e Wards. Como parâmetro de entrada, definiu-se 4 grupos. Fonte: Modificado de (Barnett & Deutsch, 2015)

Transferindo para o conjunto de dados geoestatístico (Figura 3.11) de um depósito de níquel laterítico, os autores consideraram as variáveis níquel (Ni) e óxido de magnésio (MgO) para aplicar as quatro métricas de agrupamento hierárquico. Primeiramente, é apresentado o

gráfico de dispersão *(“scartterplot”),* com a variável Ni no eixo X e a variável MgO no eixo Y, evidenciando a legenda de densidade de pontos através das cores da classificação KDE *(“kernel density estimation”)* ou estimativa de densidade do tipo kernel (Figura 3.11). Foram utilizados então três grupos como parâmetro de entrada para o agrupamento, representados pelas cores verde, azul e vermelho (Figura 3.12). Nota-se que os quatro métodos se mostraram sensíveis aos elevados valores extremos de Ni, principalmente o método *“minimun link”*. Além disso, nenhuma das quatro métricas apresentaram um resultado compatível com os valores de maiores e menores densidades da nuvem de pontos do gráfico de dispersão (Figura 3.11), que podem ser considerados como aglomerados naturais que devem ser isoladas como estruturas. Portanto, os resultados se mostram insatisfatórios.

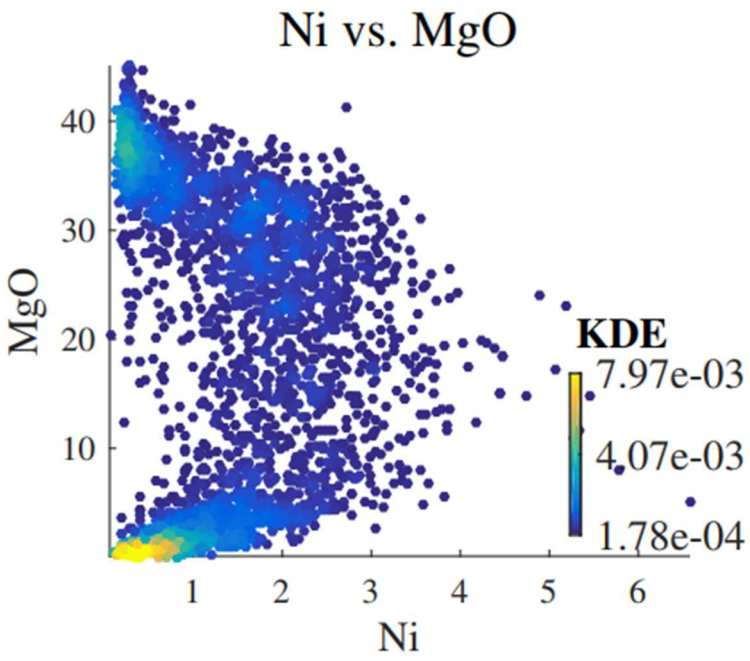

Figura 3.11: Gráfico de dispersão das variáveis Ni e MgO de um depósito geoestatístico de níquel laterítico. Aplicou-se uma legenda KDE, onde é possível analisar por cores onde se concentram as maiores e menores densidades de pontos na nuvem. Em outras palavras, os valores altos e baixos podem representar agrupamentos naturais dos dados. Fonte: (Barnett & Deutsch, 2015)

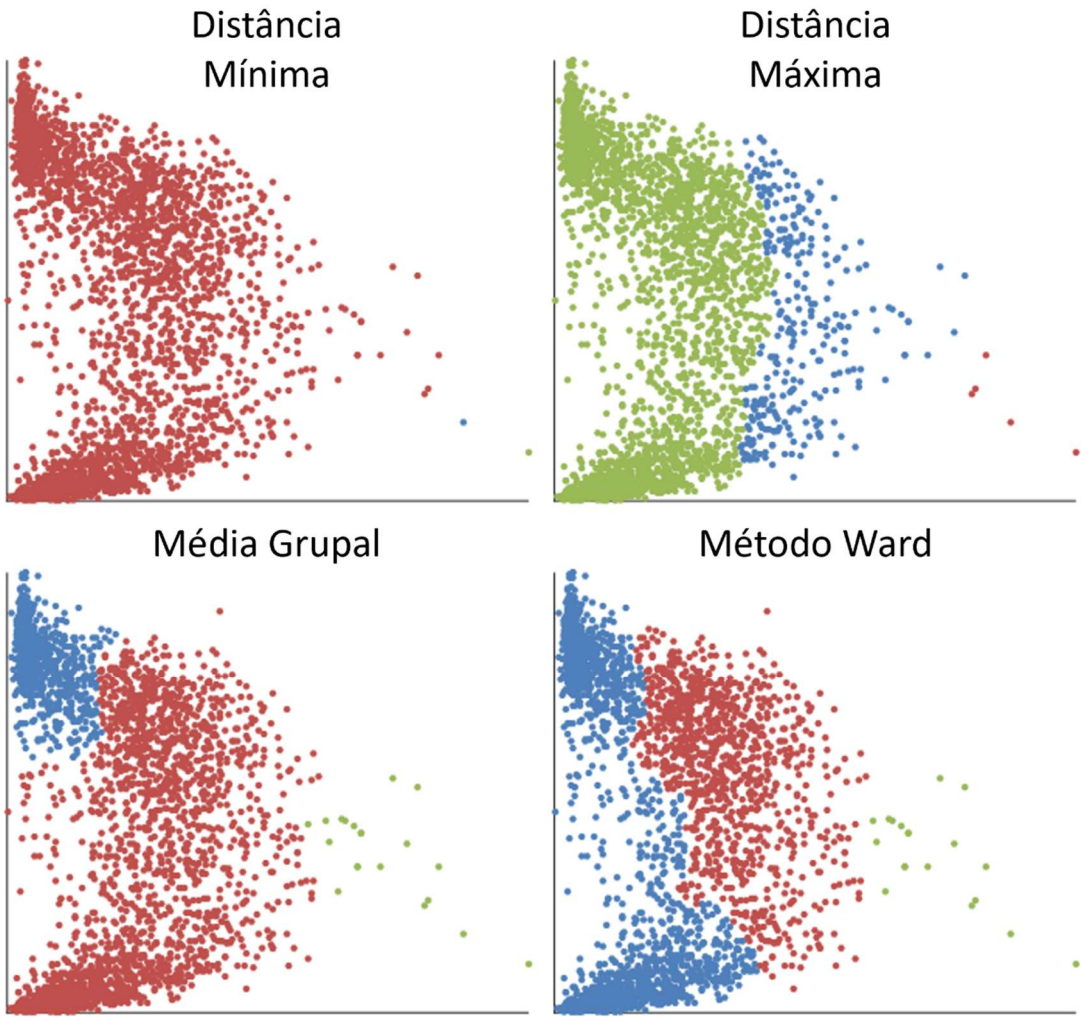

Figura 3.12: Gráficos de dispersão das variáveis Ni e MgO do mesmo depósito de níquel laterítico apresentado na Figura 3.11, aplicando as quatro diferentes métricas do algoritmo hierárquico. Percebe-se que os resultados se mostraram insatisfatórios tendo em vista que foram sensíveis aos valores extremos de Ni e não retratam os agrupamentos naturais apresentados no gráfico de dispersão da Figura 3.11. Fonte: Modificado de (Barnett & Deutsch, 2015)

A inspeção visual deste gráfico permite a interpretação da estrutura de aglomerações do banco de dados, o que facilita na escolha do número de grupos. Por essas razões, este gráfico é bastante utilizado para investigações iniciais do conjunto de dados (Martin, 2019). Moreira (2020), por exemplo, utilizou essa ferramenta na análise exploratória de dados geoestatísticos, investigando a tendência natural que os dados têm de se agrupar no espaço multivariado (Figura 3.13). Segundo o autor, a análise preliminar da estrutura do dendrograma permite inferir o

número de grupos mais adequado para o agrupamento. Ao extrapolar esse número, as distâncias no espaço multivariado são muito reduzidas, o que é ilustrado pelo curto comprimento das linhas do dendrograma, passando a não justificar nenhum agrupamento.

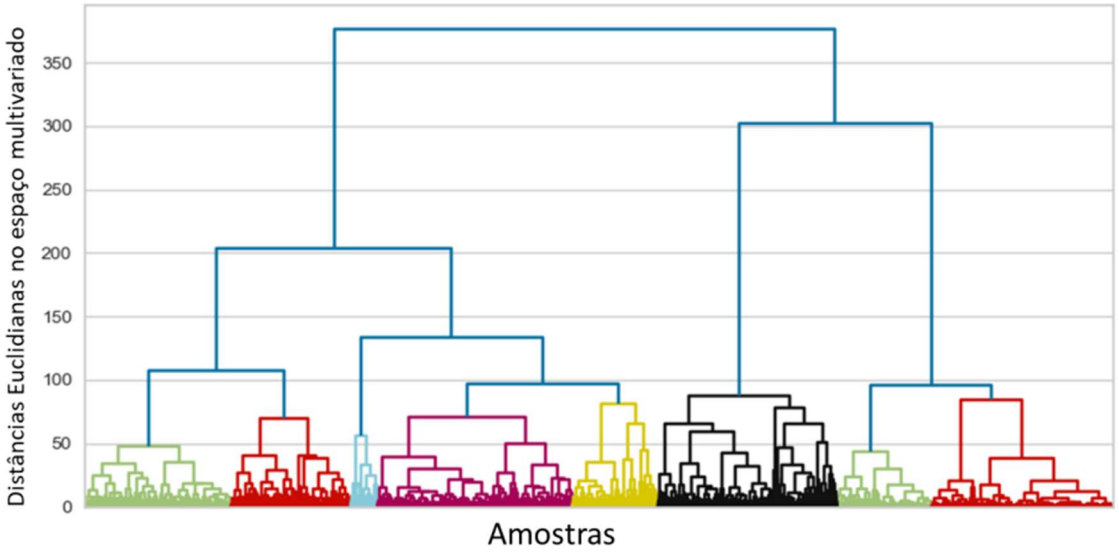

Figura 3.13: Aplicação do agrupamento aglomerativo hierárquico para um depósito de fosfato e titânio na região central do Brasil. A estrutura do dendrograma ilustra a tendência natural dos dados de se agrupar no espaço multivariado (entre 2 e oito grupos). As cores e os nós representam as conexões caso os dados fossem agrupados em oito grupos. Fonte: (Moreira, 2020).

3.5.2 Método de Agrupamento k-means

Originalmente, o método de agrupamento *k-means* foi proposto por MacQueen (1967), é um método tradicional de agrupamento de dados baseado em protótipos definidos por centroides (ponto mais representativo do grupo de pontos, que geralmente equivale a média) dos *k* grupos definidos pelo usuário. O algoritmo consiste em dividir uma população no espaço multivariado em *k* subconjuntos de dados nas distribuições de probabilidades das variáveis consideradas (Tan *et al.,* 2006).

Ao atribuir o parâmetro *k* (número de grupos), o algoritmo cria o equivalente de centroides aleatórios, onde os pontos ao seu redor serão representados pelo centroide mais

próximo. Em seguida, as posições dos centroides são atualizadas de acordo na configuração dos grupos e o processo é repetido até que os centroides permaneçam fixos (Tan *et al.,* 2006).

A técnica do método é ilustrada na Figura 3.14, onde cada grupo corresponde a uma cor (vermelho, verde e azul) e os centróides são representados pelo símbolo de uma cruz. Neste caso, o agrupamento possui quatro atualizações, denominadas de iterações 1 a 4. Na primeira, os pontos são posicionados de acordo com os centróides iniciais, todos localizados dentro do grande grupo de pontos. Na segunda iteração, as posições dos centróides são atualizadas e consequentemente, a disposição dos pontos também. Logo, as atribuições dos pontos e aos centróides são revistas por mais duas iterações, até chegar em uma configuração final, onde dois dos centróides se deslocam para conjuntos menores de dados (Tan *et al.,* 2006).

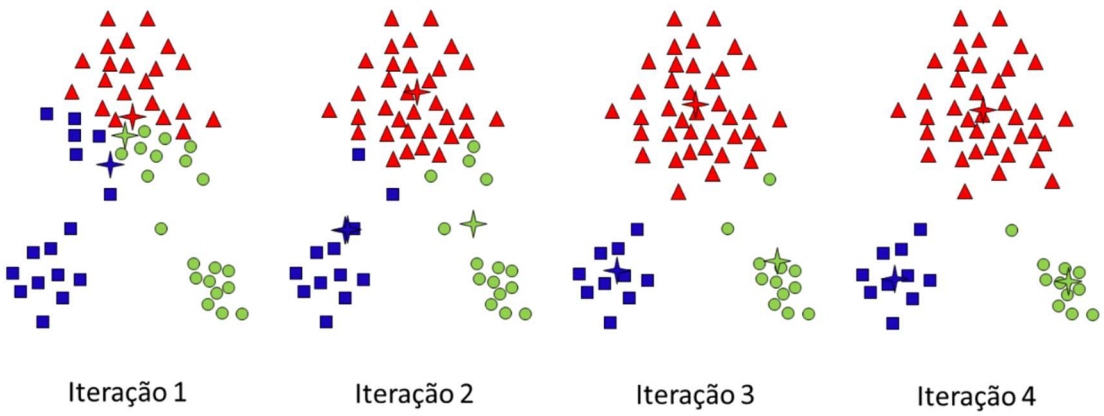

Figura 3.14: Atualização de interações entre centroides e dados para formação de três subconjuntos através do método k-means. Fonte: Moreira (2020), adaptado de Tan *et al.* (2006).

Assim como no método de aglomeração hierárquico, Barnett & Deutsch (2015) também utilizaram este método para agrupar o conjunto de dados referentes às cidades mundiais e o depósito de níquel laterítico. O conjunto multivariado das cidades é apresentado na Figura 3.15, ilustrando quatro atualizações de iterações que o algoritmo realizou para revelar o resultado final, com quatro grupos (vermelho, verde, roxo e azul) de parâmetro de entrada e seus respectivos centróides representados por quadrados. Nota-se que os resultados intermediários 2 e 3 são idênticos e estáveis, se mostrando bem diferente do inicial. No entanto, a configuração inicial influenciou negativamente no resultado final, pois o conjunto verde iniciou

no meio do oceano atlântico que presumivelmente ocorreu porque o centróide deveria estar situado na borda ocidental da África ou na borda leste da América do Sul.

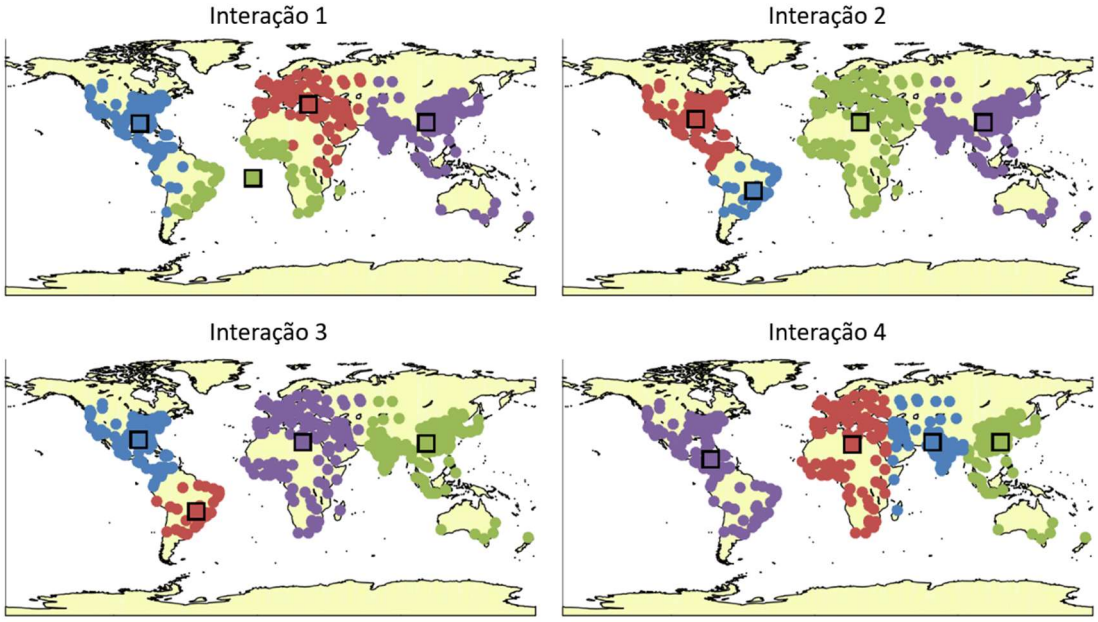

Figura 3.15: Aplicação do algoritmo k-means para o conjunto de dados multivariado das cidades mundiais. As iterações representam atualizações dos dados e seus respectivos centroides. Fonte: adaptado de (Barnett & Deutsch, 2015)

Para os dados bivariados geoestatísticos do depósito de níquel laterítico, também foi aplicado o algoritmo k-means para análise dos agrupamentos. O resultado é revelado na Figura 3.16, onde foram determinados três grupos (destacados pelas cores vermelho, verde e azul) como parâmetro inicial. Percebe-se que o resultado gerado é mais fracionado do que quando aplicado o agrupamento hierárquico. Isso acontece, pois, a soma do quadrado do erro é menor neste método. Contudo, ainda sim é um método sensível aos valores extremos de Ni.

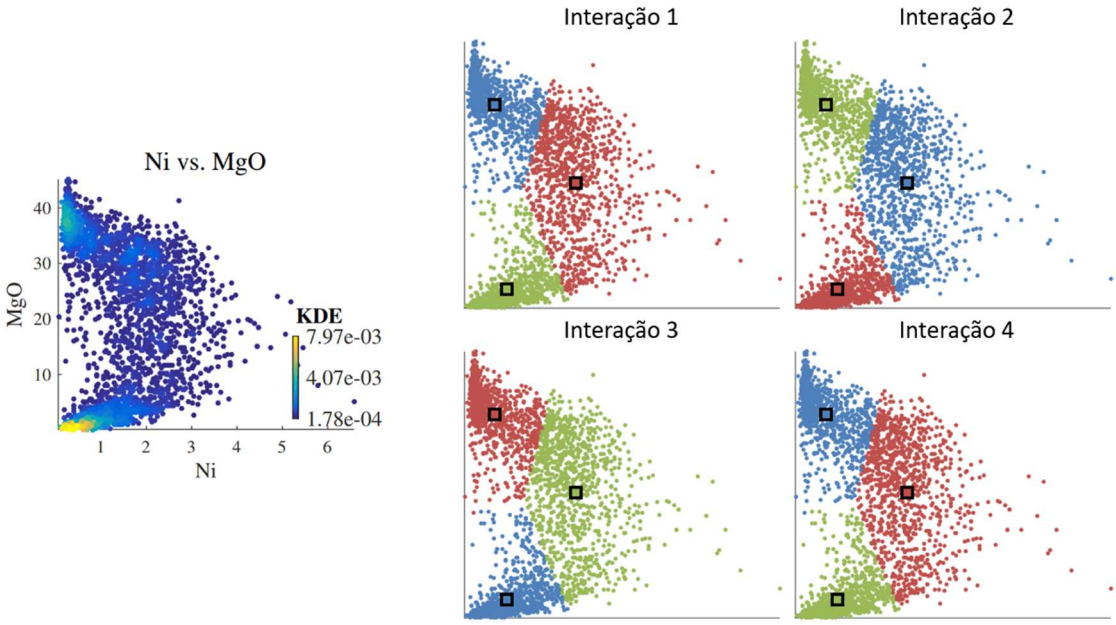

Figura 3.16: À esquerda podemos observar o gráfico de dispersão (Ni x MgO) referente ao banco de dados geoestatístico de níquel laterítico com sua legenda de densidade de pontos evidenciando um possível agrupamento natural dos dados. À direta, aplicação do método k-means, gerando grupos fracionados e sensíveis aos valores extremos. Fonte: adaptado de (Barnett & Deutsch, 2015)

3.6 Agrupamento de dados espaciais

Técnicas de agrupamento de dados multivariado tradicionais podem produzir domínios com amostras misturadas no espaço geográfico, limitando a utilidade desses algoritmos para a estima de recursos. O conhecimento de técnicas de agrupamentos de dados espaciais vem crescendo bastante ultimamente e o seu principal objetivo é produzir domínios com continuidade espacial e significância no espaço multivariado (Martin, 2019).

Segundo Moreira (2020), uma possível solução poderia ser simplesmente incorporar as coordenadas geográficas como variáveis nos métodos tradicionais. Contudo, o resultado pode gerar domínios artificialmente geométricos, não condizentes com a realidade.

3.6.1 Agrupamento Hierárquico Geoestatístico

Romary et al. (2012) apresentam uma técnica baseada no método hierárquico, na qual consideram a dependência espacial entre as amostras. Este algoritmo foi descrito com mais detalhes por Fouedjio (2016), e foi implementado no software *Minestis* e, posteriormente, no software *Isatis.neo*, ambos desenvolvidos pela *Geovariances*. O primeiro passo, o algoritmo consiste em criar conexões entre os pontos amostrais através de algoritmos eficientes de triangulações de *Delaunay*, unindo os pontos que não são necessariamente os vizinhos mais próximos (Figura 3.17). Alguns ramos desse gráfico podem parecer muito longos, principalmente nas bordas. Caso necessário, pode-se pós-processar podando as arestas mais longas para evitar conexões indesejadas. Para exemplos em duas dimensões essa tarefa é simples. Contudo, para três direções, são necessárias duas etapas (Romary *et al.,* 2012):

(i) Construção de um gráfico sobre os dados para estruturá-los em relação a sua proximidade (triangulação de *Delaunay*). No caso de um depósito em três dimensões, este gráfico é feito por sessões;

(ii) Extensão das conexões em terceira dimensão, levando a geologia em consideração, quando possível.

Fouedjio (2016) diz que a informação espacial é incluída através de um estimador kernel não paramétrico da estrutura de dependência espacial multivariada dos dados. Este estimador constrói uma medida de similaridade entre duas localizações enfatizando a posição geográfica entre os dados. Esta abordagem é aplicável em malhas irregulares de dados geoestatísticos e não inclui restrições geométricas.

Basicamente, a continuidade espacial é definida por variogramas diretos e cruzados entre duas localidades (estimador kernel), ponderando os locais de dados de acordo com a distância entre eles, atribuindo mais peso, portanto, a amostras mais próximas entre si. Então, a matriz de similaridade resultante em locais de dados serve como a entrada para um algoritmo de agrupamento hierárquico aglomerativo clássico (Fouedjio, 2016).

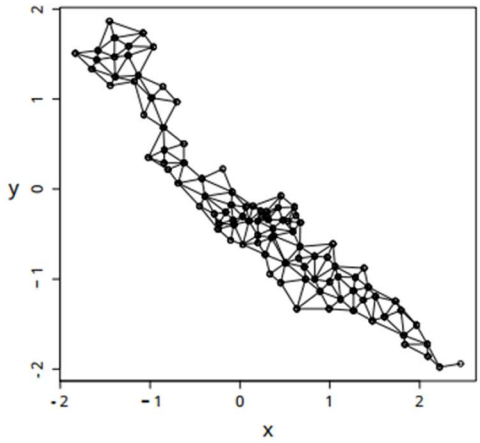

Figura 3.17: Triangulação de *Delauney* representada pelas linhas sobre amostras de um depósito mineral de urânio representadas por pontos. Fonte: Romary *et al.,* (2012).

A cada etapa do processo de agrupamento, é necessário atualizar a matriz de similaridade. Após cada agrupamento de grupos, as similaridades entre eles recém-formadas e os demais grupos são calculadas e substituídas na matriz agregada de similaridades. Diferentes estratégias são possíveis neste nível correspondendo a diferentes algoritmos de agrupamento hierárquico aglomerativo tradicional. A distinção entre essas estratégias está na forma como especificam a similaridade entre dois clusters. As variações mais populares já foram descritas neste trabalho e são conceituadas como *minimun link*, *complete link*, *group average* e *Ward*. A Figura 3.18 demonstra o resultado final do mesmo conjunto de dados mostrados na Figura 3.17, de depósito de urânio, onde cada um dos seis grupos é representado por uma cor. As variáveis consideradas foram (Romary *et al.,* 2012):

(i) Coordenadas X, Y e Z
(ii) Grau de urânio;
(iii) Fator geológico
(iv) Grau de hematização.

Após normalizações em seus valores para fins comparativos equivalentes, foram definidos pesos para as variáveis através de tentativas e erros:

(i) Coordenada: peso 1
(ii) Grau de uranio: peso 4

(iii) Fator geológico: peso 10

(iv) Grau de hematização: peso 2

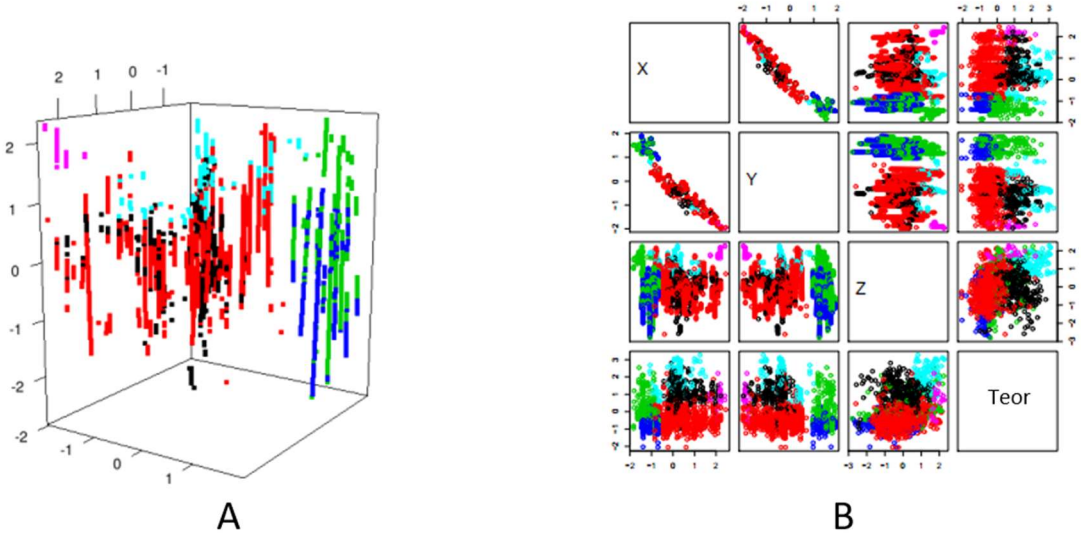

Figura 3.18: Exemplo de formação de 6 grupos pelo método aglomerativo hierarquico geoestatístico para um depósito de urânio, representados pelas diferentes cores. (A) Vista em três dimensões. (B) Gráficos de dispersão em relação às coordenadas X, Y, Z e os teores amostrados. Fonte: (Modificado de (Romary *et al.,* 2012).

O grupo representado pela cor ciano representa elevados teores de urânio e teores médios a grandes de grau de hematização, localizados na porção central e no topo do depósito e corresponde majoritariamente aos arenitos do depósito. O grupo roxo possui valores menores de urânio que o grupo ciano e baixo grau de hematização, localizados no topo e à sudeste do depósito. Não possui relevância significativa à geologia do depósito. Tanto os grupos pretos e vermelhos estão localizados no embasamento, em uma área que se estende desde a central até a porção sudeste do depósito. Possuem como característica comum o grau de hematização insignificante e o preto possui teores de urânio mais elevados que o vermelho. Finalmente, os grupos azul e verde estão localizados na área noroeste do depósito e distinguem-se dos demais pelos seus teores destoantes (Romary *et al.,* 2012).

Fouedjio (2016) também aplica este algoritmo para outro conjunto de dados geoestatísticos e compara os resultados com os métodos tradicionais de agrupamento. Os dados correspondem a oito metais pesados críticos em solos superficiais do banco de dados

geoquímico referente a 26 países que compõem o Fórum de Pesquisas Geológicas Europeias (LADO et al., 2008). As variáveis são: arsênio (As), cádmio (Cd), cromo (Cr), cobre (Cu), mercúrio (Hg), níquel (Ni), chumbo (Pb) e zinco (Zn). Foram utilizados 1498 dados no estudo que compunham todo o leque de variáveis necessários. Para todos os métodos aplicados de agrupamento, todas as variáveis foram normalizadas em escala logarítmica, inclusive as coordenadas geográficas, quando aplicadas (Figura 3.19). Assim como o autor, neste trabalho serão descritos os métodos aplicados como:

- (M1) *k-means* com coordenadas geográficas como variáveis adicionais (Figura 3.20);
- (M2) Agrupamento hierárquico com coordenadas geográficas como variáveis adicionais, utilizando *ward* como métrica (Figura 3.21);
- (M3) Agrupamento hierárquico geoestatístico (Figura 3.22).

Os resultados mostram que as aplicações dos algoritmos tradicionais (M1 e M2) não geram grupos espacialmente contínuos, enquanto o método de agrupamento hierárquico geoestatístico consegue oferecer este diferencial. Para o método (M1 e M2), quanto mais aumenta o número de grupos, mais desordenado e espalhado no espaço cartográfico os grupos são impressos.

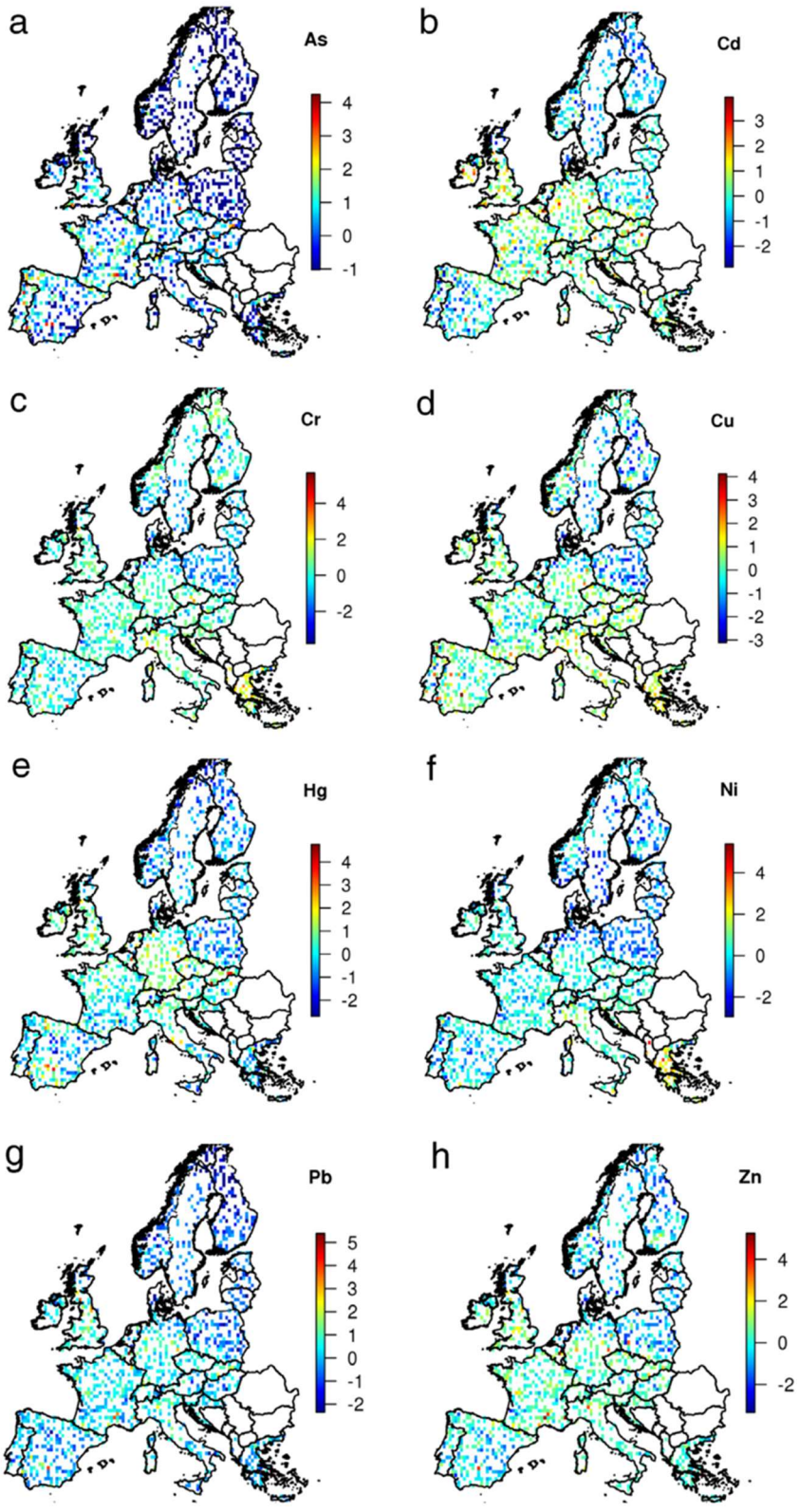

Figura 3.19: Variáveis que compões o banco de dados Fórum de Pesquisas Geológicas Europeias (Lado *et al.,* 2008) normalizadas e em escala logaritma. Fonte: (Foedjio, 2016).

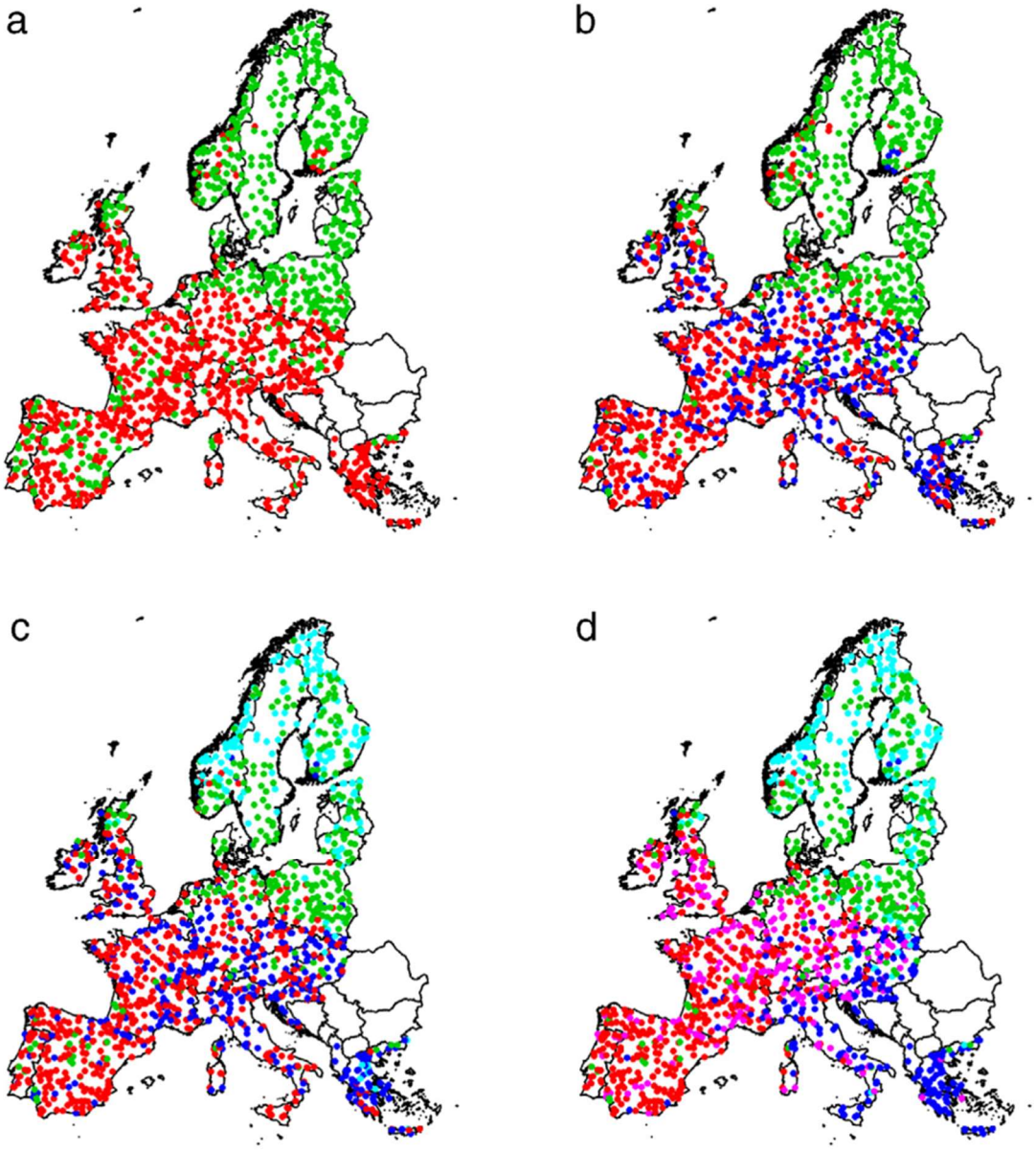

Figura 3.20: Método *k-means* aplicado ao conjunto de dados do Fórum de Pesquisas Geológicas Europeias (Lado et al., 2008). (A) 2 grupos; (B) 3 grupos; (C) 4 grupos; (D) 5 grupos. Fonte: (Fouedjio, 2016).

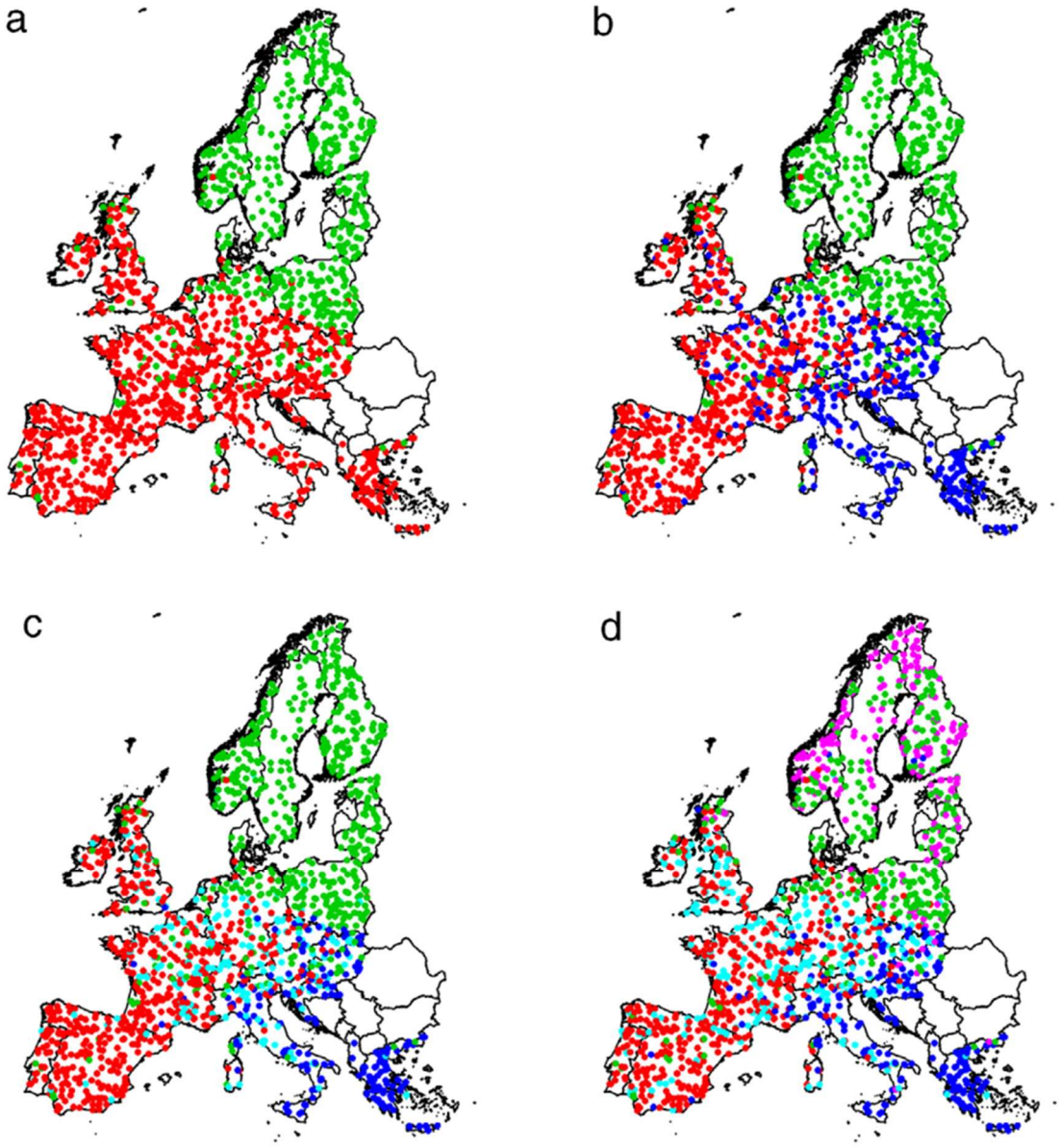

Figura 3.21: Agrupamento Hierárquico aplicado ao conjunto de dados do Fórum de Pesquisas Geológicas Europeias (Lado et al., 2008). Suas coordenadas geográficas foram utilizadas como variável para o agrupamento. (A) 2 grupos; (B) 3 grupos; (C) 4 grupos; (D) 5 grupos. Fonte: (Fouedjio, 2016)

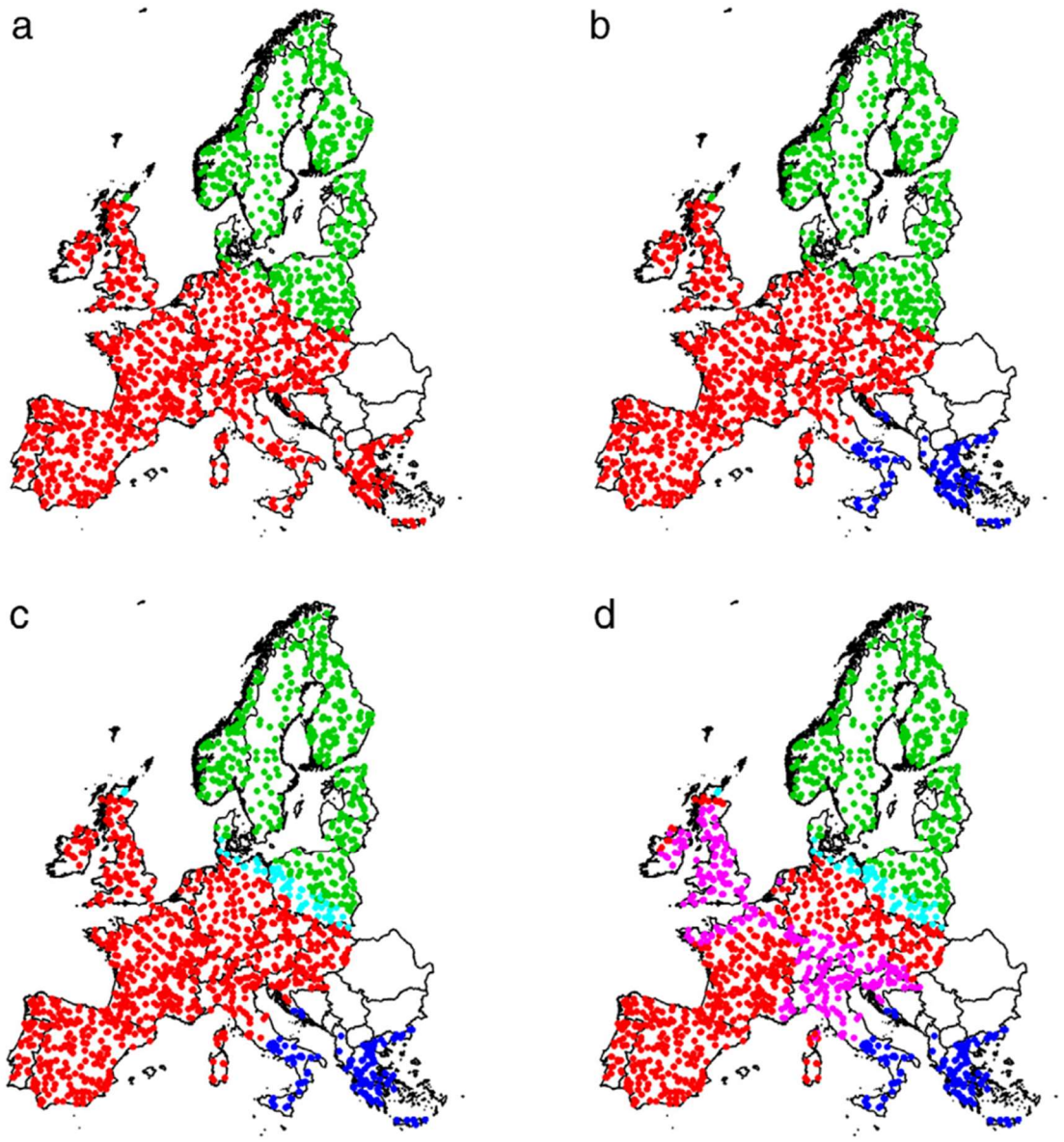

Figura 3.22: Agrupamento Hierárquico Geoestatístico aplicado ao conjunto de dados do Fórum de Pesquisas Geológicas Europeias (Lado et al., 2008). (A) 2 grupos; (B) 3 grupos; (C) 4 grupos; (D) 5 grupos. Fonte: (Fouedjio, 2016)

3.6.2 Método de agrupamento em espaço duplo

Oliver & Webster (1989) modificaram a matriz de semelhança de pares (Figura 3.8) do método de agrupamento de dados hierárquico tradicional usando uma secundária calculada a partir da interconectividade espacial das localizações das amostras. Dessa forma, eles determinaram a escala de variação espacial, efetivamente diminuindo o parentesco de pares similares para pontos separados por longas distâncias.

Mais recentemente, Martin & Boisvert (2018) publicam um algoritmo fundamentado na combinação de amostras por vizinhança de busca no espaço geográfico. Este algoritmo de agrupamento propõe uma métrica que descreve a qualidade do agrupamento tanto no espaço multivariado, quanto no cartesiano. Além disso, o algoritmo reduz a interferência do usuário, que basta parametrizar apenas a anisotropia obrigatoriamente. A busca é feita exaustivamente e aleatoriamente, até extrair uma configuração final de agrupamento. Os autores descrevem a metodologia em 3 etapas (Figura 3.23):

(i) Aglomeração espacial preliminar

Os primeiros agrupamentos são realizados por buscas de vizinhos mais próximos no espaço geográfico. No espaço multivariável, a distância Euclidiana é calculada por caminhos aleatórios entre os pares de amostras. Portanto, as amostras com as menores distâncias no espaço multivariável são aglomeradas e essa fase se repete até todas as amostras pertenciam a algum minigrupo. O número de amostras que serão juntadas é um parâmetro opcional do algoritmo.

(ii) Aglomeração multivariada secundária

Essa etapa intermediária dá sequência à anterior no sentido de agrupar os minigrupos já formados. Para isso, usa-se uma métrica de proximidade multivariada chamada *Ward.*

(iii) Rotulação final dos grupos

Nesta etapa final, todas as realizações de agrupamentos da etapa 2 são armazenados em uma matriz Local x Realização. A partir disso, técnicas são aplicadas para definir uma configuração final. A técnica mais simples consiste por agrupamento hierárquico utilizando métrica *Ward.*

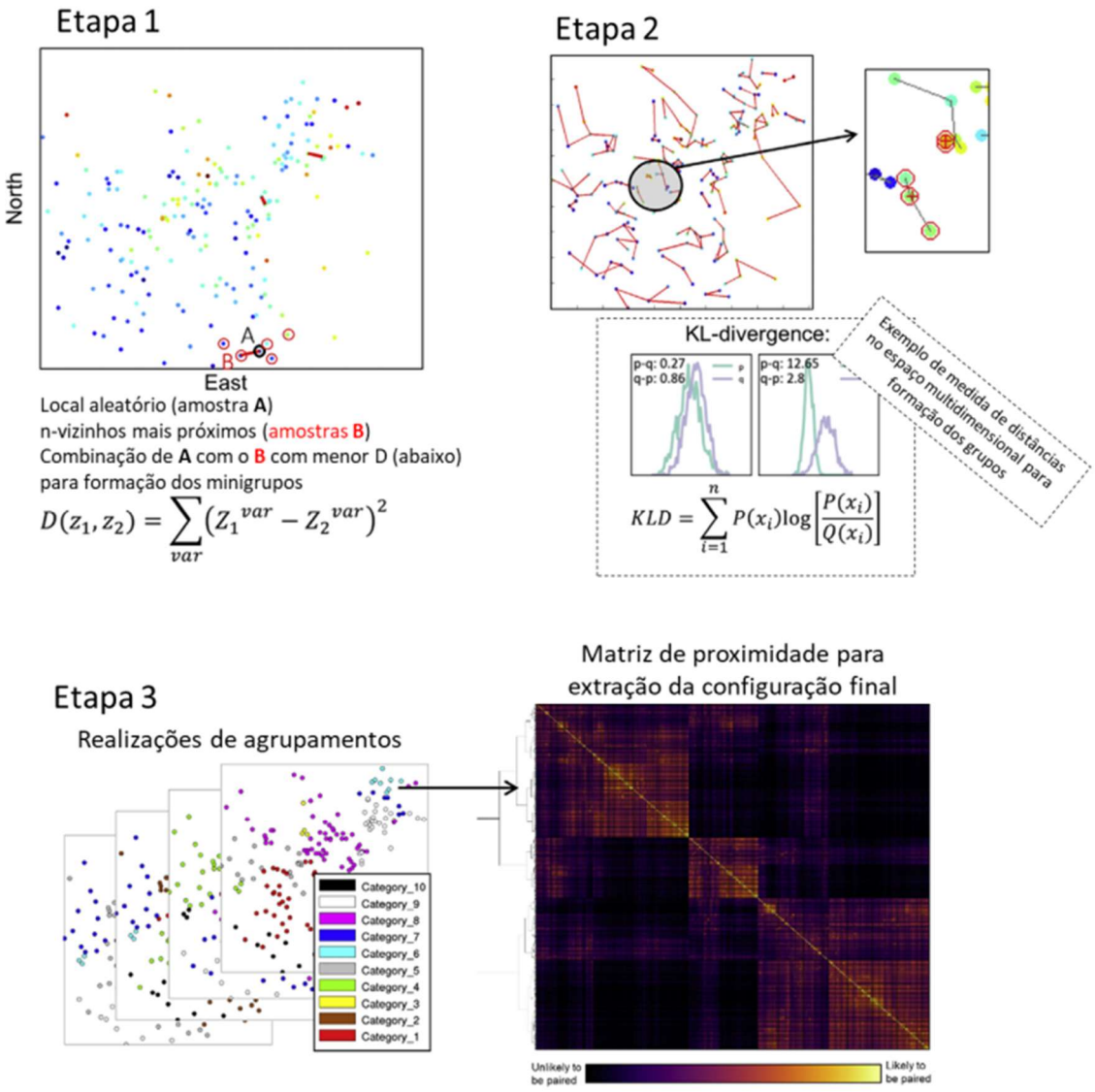

Figura 3.23: Etapas seguidas pelo algoritmo para agrupamento por restrição de vizinhança. 1) Minigrupos formados pela menor distância no espaço multivariado, dentre os vizinhos mais próximos. 2) Agrupamentos dos minigrupos. 3) Matriz de proximidade construída e agrupamento final. Fonte: Moreira (2020) adaptado de Martin & Boisvert (2018).

Os autores aplicaram este algoritmo para o banco de dados Jura, coletado pelo Instituto tecnológico federal em Laussane na Suíça (Geoovaerts, 1997) onde são descritas 359 amostras com 7 metais analisados (Cd, Co, Cr, Cu, Ni, Pb e Zn), 5 tipos distintos de rocha e 4 usos da terra diferentes (Figura 3.24).

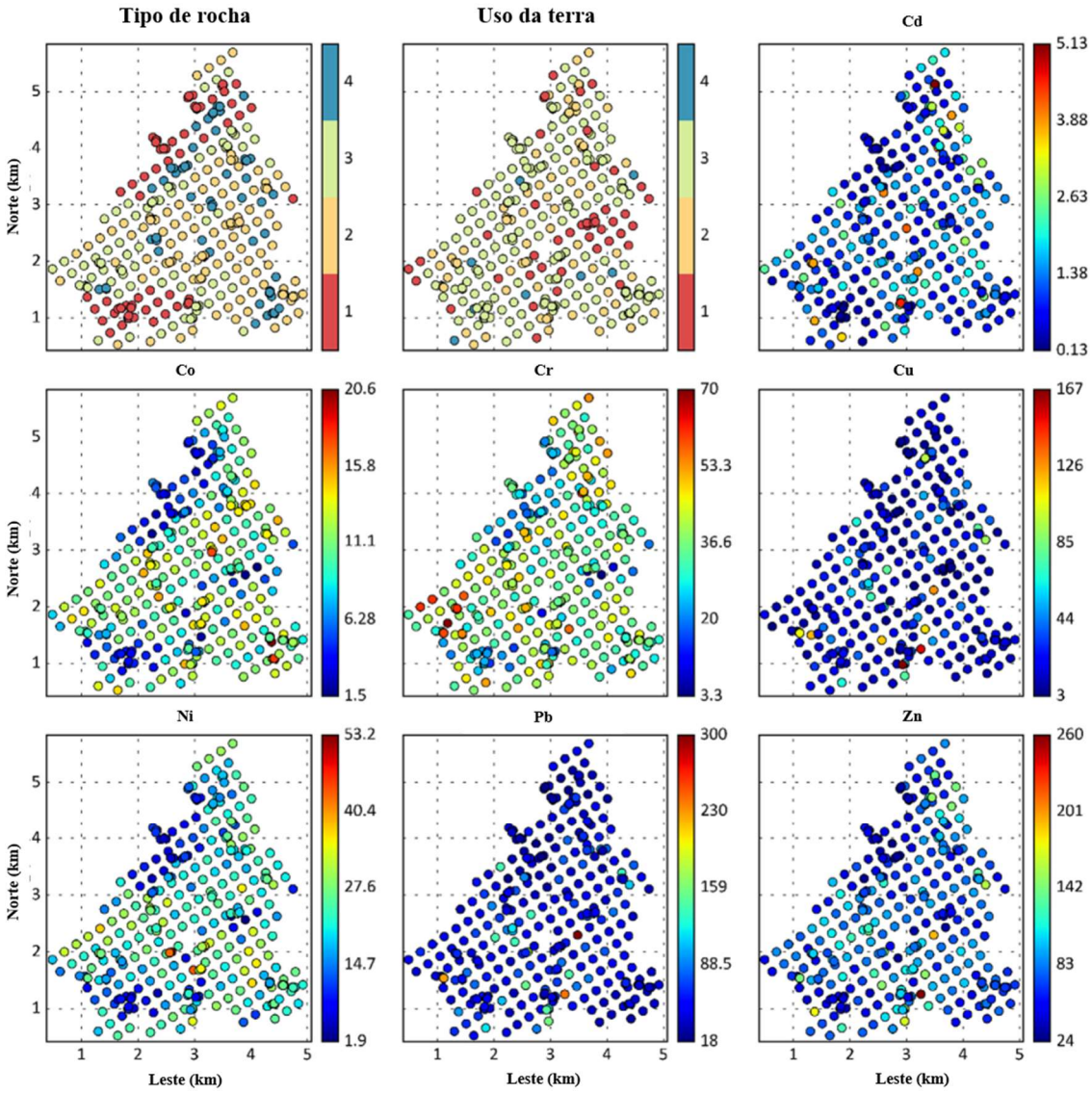

Figura 3.24: Banco de dados Jura evidenciando suas variáveis quantitativas e qualitativas. Fonte (adaptado de Martin & Boisvert (2018).

A fim de comparação entre resultados, o autor utilizou os mesmos parâmetros para agrupar esse banco de dados, aplicando diferentes metodologias: aglomerativo hierárquico, aglomerativo hierárquico com coordenadas geográficas como variável e o algoritmo de aglomeração em espaço duplo. Os autores interpretam que os tipos de rocha possuem um bom delineamento espacial, mas o pior delineamento multivariável (são muito dependentes da localização) (Figura 3.25-A). Por outro lado, os grupos do método hierárquico tradicional agrupam melhor o espaço multivariado, configurando o espaço geográfico mais aleatoriamente (Figura 3.25-B). Ao utilizar as coordenadas como parâmetro de entrada deste método

tradicional, a delimitação espacial é melhorada, mas faz com que a coordenada X tenha um maior peso para o classificador hierárquico (Figura 3.25-C). Como esperado, os métodos de agrupamentos espaciais por espaço duplo geram classes com baixa entropia espacial *(H)*, enquanto mantém pontuação multivariável *(wcss)* comparáveis com os outros métodos (Martin & Boisvert, 2018).

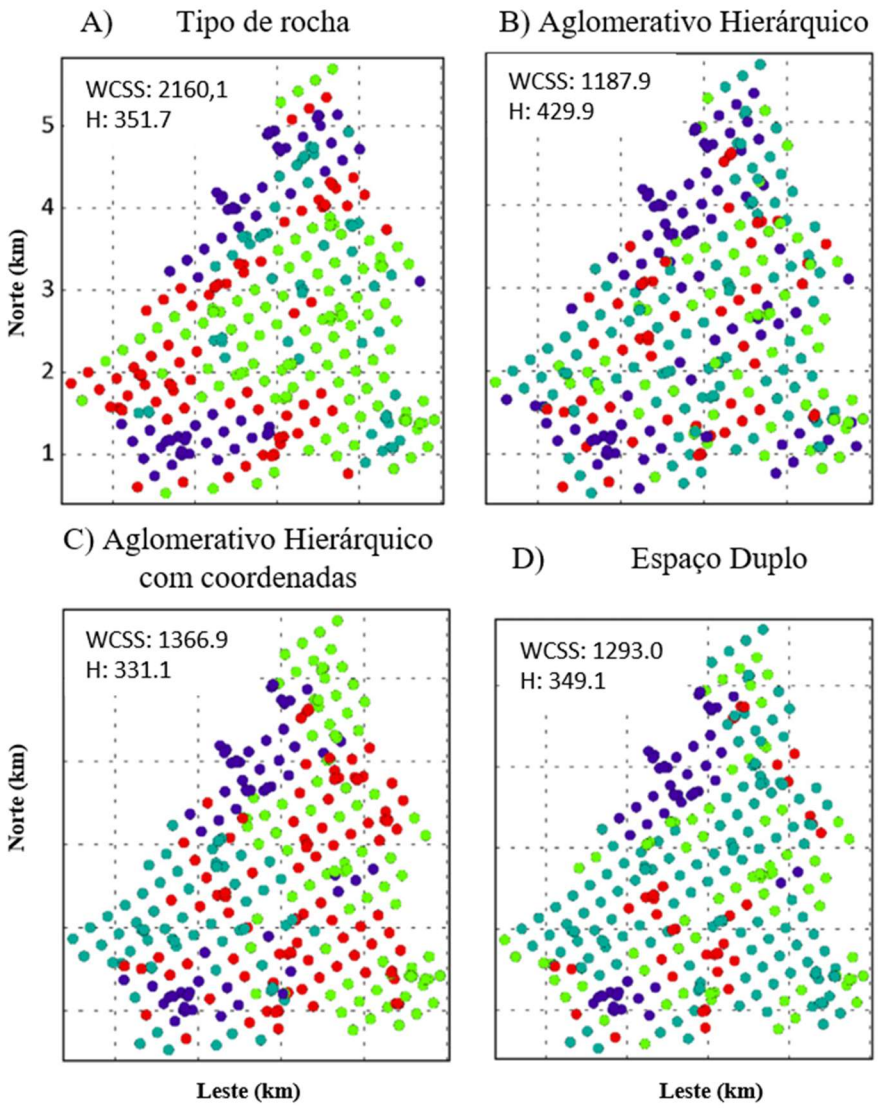

Figura 3.25: Diferentes métodos de agrupamento aplicados ao banco de dados Jura, com os valores de validação de *(wcss)* e entropia *(H).* Fonte: (adaptado de Martin & Boisvert (2018).

4 CONCLUSÃO

As ferramentas de agrupamento de amostras por algoritmos para definição de domínios ainda são pouco utilizadas no setor mineral. Mesmo diante de muitas discussões na academia sobre o assunto, a prática de grande parte das indústrias é limitada em apenas considerar as informações geológicas para a modelagem de domínios estacionários.

Os registros geológicos são uma fonte complexa de informações oriunda de diversas fontes e metodologias que podem refletir em classificações subjetivas ou até mesmo equivocadas dos domínios geológicos. Além disso, em casos complexos de alterações intempéricas, hidrotermais e/ou metamórficas, as tipologias não apresentam fortes correlações com os teores. Portanto, as características geológicas são bons guias para a classificação das amostras, mas dependem também de análises multivariadas e suas correlações espaciais para garantir estacionariedade.

Por isso, alguns poucos geomodeladores utilizam algoritmos tradicionais de agrupamento de dados para definições de domínios geoestatísticos. Embora esses algoritmos contribuam com a decisão de estacionariedade, eles consideram apenas as relações entre os dados no espaço multivariado, desconsiderando suas correlações no espaço geográfico.

Recentemente, estão sendo desenvolvidos algoritmos de agrupamentos de dados que também levam em consideração a correlação espacial entre as amostras e apresentam resultados consistentes. Entretanto, a validação deste método continua sendo interpretativa, pois não existe uma referência comparativa para os valores de *(wcss)* e entropia espacial *(H)*. Além disso, o algoritmo ainda depende de decisões subjetivas do usuário, como métricas de agrupamento e número de grupos.

Portanto, ainda cabe ao geomodelador decidir qual melhor metodologia de agrupamento de dados a ser utilizada. Para isso, sugere-se a realização de vários cenários testes, a fim de comparar os resultados para melhor entendimento do espaço multivariado e também do espaço cartesiano do determinado conjunto de dados em análise. Por mais que os algoritmos espaciais levam em consideração ambos os espaços, a comparação com os resultados dos algoritmos tradicionais faz-se necessária para entender o extremo do agrupamento puramente multivariado.

Uma vez comparados diferentes resultados e validações dos algoritmos, o geomodelador terá mais embasamento para tomada de uma decisão mais assertiva de estacionariedade. Por mais que o processo possa ser laborioso, a aplicação desses algoritmos garante que os próximos passos da modelagem de recursos não sejam comprometidos, evitando, portanto, retrabalhos ou até mesmo erros significativos da estimativa final.

Por fim, os métodos descritos neste trabalho são muito eficazes para tomadas de decisões, mas ainda dependem de um conhecimento especializado do geomodelador. De acordo com a literatura, trabalhos estão sendo desenvolvidos à fim de automatizar a decisão de estacionariedade, através de algoritmos que não exigem parâmetros de entrada do usuário e garantem melhores cenários de validação.

Referências

Barnett, R., & Deutsch, C. (2015). Conventional Clustering Algorithms and a Program for their Application. *CCG Annual Report 17*, 1-18.

Fouedjio, F. (2016). A hierarchical clustering method for multivariate geostatistical data. *Elsevier*, 2211-6753.

Isaaks, E., & Srivastava, R. M. (1989). *An introduction to applied geostatistics.* New York.: Oxford University Press.

Journel, A., & Huijbrets, C. J. (1978). *Mining Geostatistics.* Academic Press.

Lado, L. H. (2008). Heavy metals in European soils: A geostatistical analysis of the FOREGS geochemical database. *Geoderma 148*, 189-199.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY.*

Martin, R. (2019). Data driven decisions of stationarity for improved numerical modeling in geological environments. *Tese (Doutorado) — University of Alberta.*

Martin, R., & Boisvert, J. (2018). Towards justifying unsupervised stationary decisions for geostatistical. *ELSEVIER*, 82-96.

Mclennan, J. A. (2007). The decision of stationarity. *Tese (Doutorado) — University of Alberta*.

Moreira, G. (2020). Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais. *Dissertação de Mestrado - Universidade Federal do Rio Grande do Sul*.

Oliver, M., & Webster, R. (1989). A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, 15-35.

Pedregosa, F. e. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Pereira, P. E. (2017). Estimativa de Recursos Minerais e Otimização de Cava Aplicados a um Estudo de Caso de uma Mina de Calcário. *Dissertação de Mestrado – Unidade Acadêmica Especial de Matemática e Tecnologia, Universidade Federal de Goiás, Catalão, 2017.*

Romary, T., Rivoirard, J., Quinones, C., & Freulon, X. (2012). Domaining by clustering multivariate geostatistical data. *Geostatistics Oslo*, 455-466.

Rossi, M. E., & Deustch, C. V. (2014). *Mineral Resource Estimation.* Spring Science.

Scrucca, L. (2005). Clustering multivariate spatial data based on local measures of spatialautocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica, Università di Perugia*.

Silva, A. H. (2000). Modelagem Geológica e Estocástica da Porção NE da Mina de Morro do Ouro, Paracatu (MG). *Dissertação de Mestrado – Instituto de Geociências, UNICAMP, Campinas, 2000.*

Sinclair, A., & Blackwell, G. (2004). *Applied Mineral Inventory Estimation.* Cambridge University Press.

Soares, A. (2006). *Geoestatística para as ciências da terra e do ambiente.* Lisboa: Instituto Superior Técnico.

Sokal, R., & Sneath, P. (1963). *Principles of Numerical Taxonomy.* W. H. Freeman and Company.

Tan, P., Michael, S., & Kumar, V. (2006). *Introduction to data mining.* New York: Pearson Education.

Yamamoto, J. K., & Landim, P. M. (2013). *Geoestatística: conceitos e aplicações.* São Paulo: Oficina de Textos.