Gabriel de Castro Moreira

Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais

Porto Alegre

2020

Gabriel de Castro Moreira

Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais

Documento submetido ao programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e Materiais da Escola de Engenharia da UFRGS, para a obtenção ao título de Mestre em Engenharia.

Universidade Federal do Rio Grande do Sul

Escola de Engenharia

Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais

Orientador: Prof. Dr. João Felipe Coimbra Leite Costa

Coorientador: Prof. Dr. Diego Machado Marques

Porto Alegre

2020

Gabriel de Castro Moreira

Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais/ Gabriel de Castro Moreira. – Porto Alegre, 2020-
101 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. João Felipe Coimbra Leite Costa

Dissertação de Mestrado – Universidade Federal do Rio Grande do Sul
Escola de Engenharia
Programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e de Materiais, 2020.

1. Aprendizado de máquina; 2. Agrupamento de dados; 3. Geoestatística; 4. Análises multivariadas; 5. Recursos minerais.
I. Prof. Dr. João Felipe Coimbra Leite Costa. II. Universidade Federal do Rio Grande do Sul. III. Faculdade de Engenharia. IV. Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais

CDU 02:141:005.7

Gabriel de Castro Moreira

Análise de agrupamento aplicada à definição de domínios de estimativa para a modelagem de recursos minerais

Documento submetido ao programa de Pós-Graduação em Engenharia de Minas, Metalúrgica e Materiais da Escola de Engenharia da UFRGS, para a obtenção ao título de Mestre em Engenharia.

Trabalho aprovado. Porto Alegre, 27 de maio de 2020:

Prof. Dr. João Felipe Coimbra Leite Costa
Orientador

Prof. Dr. Diego Machado Marques
Coorientador

Prof. Dr. Marcel Antônio Arcari Bassani
PPGE3M - UFRGS

Dr. Áttila Leães Rodrigues
UFRGS

Dr. Luciano Nunes Capponi
Mosaic Fertilizantes

Porto Alegre
2020

À todos.

Agradecimentos

Difícil expressar tanto em apenas alguns poucos parágrafos...

Acredito que a vida seja um processo integrado, no qual as trajetórias de todos se entrelaçam de alguma forma em algum momento...ou em vários. Mesmo que indiretamente, muita gente esteve envolvida na construção desta Dissertação, no passado recente ou no remoto, mesmo que anteriormente ao próprio período que me dediquei ao mestrado, nos últimos dois anos. Alguns compartilhando conhecimento e tempo, outros, fornecendo meios e serviços. Sou grato a todos, mas aqui gostaria de expressar alguns agradecimentos especiais.

À minha família, sempre compreensiva e disposta a ouvir, com quem compartilhei momentos bons e ruins, meu verdadeiro porto seguro em todas as ocasiões. Sem seu apoio não conseguiria chegar aos meus objetivos, profissionais ou pessoais.

Ao meu orientador, Prof. João Felipe Costa, um verdadeiro exemplo de pessoa e de profissional, sempre receptivo e solícito. Ao meu coorientador, Prof. Diego Marques, pelas ótimas ideias e pelo tempo compartilhado em longas conversas sobre variados assuntos. Aos demais professores, por disseminarem seu conhecimento.

Aos meus amigos e colegas do LPM, com quem aprendi muito, sempre dispostos a compartilhar seu tempo e conhecimento. Aos meus "irmãos" do ap 501, com quem dividi o tão sagrado espaço de convívio em Porto Alegre nesses últimos dois anos.

À Mosaic Fertilizantes, por disponibilizar os dados utilizados neste trabalho e ao Dr. Luciano Capponi, por permitir uma visita técnica muito relevante para o progresso do trabalho. Também aos funcionários que me acompanharam durante aquela visita, com atenção e paciência.

À CAPES e à Fundação Luiz Englert, pelo indispensável apoio financeiro.

Por fim, aos membros da banca, Prof. Dr. Marcel Bassani, Dr. Áttila Rodrigues e Dr. Luciano Capponi, por aceitarem o convite para avaliar este trabalho.

"Quanto mais sabemos, melhor entendemos a vastidão de nossa ignorância e mais perguntas somos capazes de fazer, perguntas que, previamente, nem poderiam ter sido sonhadas."

(Marcelo Gleiser – A Ilha do Conhecimento)

Resumo

A definição de domínios de estimativa é uma das primeiras etapas a se cumprir na modelagem de recursos minerais e uma das decisões mais importantes em todo o processo. Uma definição inadequada de domínios pode complicar desnecessariamente a modelagem ou, pior, comprometer os resultados das estimativas, o que pode levar a uma avaliação imprecisa de massas e teores. O conceito de domínio de estimativa está relacionado à noção de estacionariedade, e existem várias abordagens para se tratar o assunto. No campo do aprendizado de máquina, a análise de agrupamento fornece algumas técnicas interessantes que podem ser aplicadas nesse contexto. No entanto, tradicionalmente, esses métodos são próprios para se lidar com dados no espaço multivariado, sem considerar a posição das amostras no espaço geográfico. Mais recentemente, técnicas específicas têm sido apresentadas a fim de realizar o agrupamento de dados geoposicionados. A validação da análise de agrupamento também é uma tarefa um tanto complexa, já que não existem rótulos predefinidos para referência, e diversos métodos devem ser utilizados simultaneamente para que as conclusões sejam mais assertivas. Nesta Dissertação, é feita uma ampla discussão acerca da análise de agrupamento e das técnicas de validação. Como demonstração, são apresentados e discutidos os resultados de quatro algoritmos de agrupamento e alguns métodos de validação, aplicados a um conjunto de dados de um depósito de fosfato e titânio. Também é verificada a possibilidade de se utilizar algoritmos de aprendizado supervisionado para a classificação automatizada de novas amostras, baseado nos grupos definidos na análise de agrupamento. A automatização de procedimentos permite aumentar significativamente a reprodutibilidade do processo de modelagem, uma condição essencial na avaliação de recursos minerais, principalmente para fins de auditoria. No entanto, embora muito eficazes no processo de tomada de decisão, os métodos apresentados ainda não são totalmente automatizados, exigindo conhecimento prévio e muito bom senso.

Palavras-chaves: Aprendizado de máquina; Análise de agrupamento; Geoestatística; Análises multivariadas; Recursos minerais.

Abstract

The definition of estimation domains is one of the first steps to be taken in mineral resource modeling and one of the most important decisions in the entire process. An inadequate definition of domains can unnecessarily complicate the modeling or worse, compromise the results of the estimates, which can lead to an inaccurate evaluation of grades and tonnages. The concept of estimation domain is related to the notion of stationarity and there are several approaches to deal with this matter. In the field of machine learning, cluster analysis provides some interesting techniques that can be applied in this context. However, traditionally, these methods are constructed for dealing with data in the multivariate space only, without considering the position of the samples in the geographic space. More recently, specific techniques have been presented in order to perform the grouping of spatial data. Validating the results of cluster analysis can also be challenging because there are no predefined labels for reference, and several methods must be used simultaneously so that the conclusions are more accurate and precise. In this thesis, an extensive discussion on cluster analysis and validation techniques is presented. As an illustration, a dataset from a phosphate and titanium deposit is used in order to demonstrate the application of four clustering algorithms and some validation methods. The possibility of using supervised learning algorithms for the automatic classification of new samples, based on the results of the cluster analysis, is also verified. The automation of methods and procedures can significantly increase the reproducibility of the modeling process, an essential condition in the evaluation of mineral resources, especially for auditing purposes. However, although very effective in the decision-making process, the methods herein presented are not yet fully automated, requiring prior knowledge and good judgment.

Key-words: Machine learning; Cluster analysis; Geostatistics; Multivariate analysis; Mineral resources.

Lista de ilustrações

Figura 1 – Fluxograma geral do método proposto, incluindo as etapas de análise de agrupamento e classificação automática de novas amostras. Conforme indicado pela linha tracejada, esporadicamente, a análise de agrupamento deve ser revisitada, de modo a incorporar a totalidade dos dados. O classificador deverá ser, então, atualizado 23
Figura 2 – Possível fluxo para a definição de domínios estacionários considerando as propriedades das variáveis regionalizadas, a configuração espacial e outros tipos de informações geológicas (Adaptado de Martin (2019)). . 28
Figura 3 – Diferentes maneiras de se agrupar um mesmo conjunto de dados. Em (A), os pontos originais, não agrupados. Em (B), conjunto dividido em dois grandes grupos e, em (C) e (D), duas maneiras distintas de se subdividir os dois grandes grupos de (B) (TAN et al., 2006). 31
Figura 4 – Exemplos das diferenças nos resultados que podem ser obtidos, para os mesmos conjuntos de dados, com distintos algoritmos de análise de agrupamento, disponíveis na biblioteca de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011). Cada coluna corresponde a um algoritmo diferente, sendo, da esquerda para a direita: *k-means*, aglomerativo hierárquico, *DBSCAN*, *OPTICS* e Modelos Gaussianos Mistos. 33
Figura 5 – Cenário ilustrativo de um agrupamento aglomerativo hierárquico com uma variável, no qual 20 amostras (representadas pelos pontos na parte superior da figura) são sucessivamente aglomeradas, par a par, até a formação de cinco grupos. Cada etapa do algoritmo é representada por uma linha transversal que corta o dendrograma, e cada nó representa uma aglomeração entre amostras e/ou grupos. Após quinze etapas (S_1 a S_{15}), as 20 amostras foram aglomeradas em cinco grandes grupos (adaptado de Faber (1994)). 34
Figura 6 – Exemplo do método aglomerativo hierárquico para duas variáveis (Z_1 e Z_2) em que são exibidos o gráfico de dispersão de sete amostras e o respectivo dendrograma. Amostras dos grupos A, B1 e B2 são identificadas com cores diferentes, que também estão presentes nas linhas e conexões do dendrograma (BARNETT; DEUTSCH, 2015). . . 35
Figura 7 – Representação gráfica das distâncias usadas em alguns dos critérios de proximidade. Os pontos representam amostras. (A) Distância mínima, ou “single link”, (B) Distância máxima, ou “complete link”, (C) Média grupal (adaptado de Tan et al. (2006)). 36

Figura 8 – Representação do funcionamento do método *k-means* para definição de três grupos (adaptado de Tan et al. (2006)). 37
Figura 9 – Representação de agrupamento de dados usando modelos Gaussianos mistos com quatro componentes (grupos). MGMs podem ser aplicados para definir grupos com contornos circulares (A) ou não-circulares (B) (VANDERPLAS, 2016). 38
Figura 10 – Ilustração da aplicação do algoritmo de agrupamento *DBSCAN*. (A) Representação gráfica da classificação das amostras como 'de núcleo' (vermelho), 'de borda' (amarelo) e 'ruído' (azul). Os círculos em torno das amostras representam o raio de busca (Eps) (SCHUBERT et al., 2017); (B) Resultado da aplicação do *DBSCAN* em um banco de dados bidimensional, onde cada grupo é representado por uma cor diferente, sendo as amostras de núcleo como pontos maiores, as de borda como pontos menores e o ruído como pontos pretos em zonas de baixa densidade (PEDREGOSA et al., 2011). 40
Figura 11 – Descrição conceitual do funcionamento do algoritmo para agrupamento em espaço duplo. Na etapa 1, os minigrupos são formados pelos pares com menor distância no espaço multivariado, dentre os n vizinhos mais próximos; na etapa 2, os minigrupos são combinados para formar os agrupamentos-alvo e; na etapa 3, a matriz de proximidade é construída para que seja extraída a configuração de agrupamentos consensual final (adaptado de Martin & Boisvert (2018)). 43
Figura 12 – Ilustração dos elementos envolvidos no cálculo do coeficiente de silhueta (s) do objeto i, pertencente ao grupo A, sendo B o grupo mais próximo e C, um grupo qualquer, mais distante que B. Assim, para o cálculo de $s(i)$ são computadas as distâncias entre i e as demais amostras pertencentes ao grupo A e as distâncias entre i e as amostras pertencentes ao grupo B (ROUSSEEUW, 1987). 46
Figura 13 – Exemplo adaptado de Rousseeuw (1987) de gráficos de silhuetas. (A) Relação dos dados no espaço (bivariado), com a delineação natural dos grupos A, B, C e D; (B) Gráficos de silhuetas para quatro grupos (esquerda), para dois grupos (centro), e para seis grupos (direita). Percebe-se que a configuração natural, de quatro grupos, é aquela que apresenta o maior valor médio do índice s, ao passo que o agrupamento dos dados em dois grandes grupos, bem como sua subdivisão em seis subgrupos leva a valores mais baixos do s médio. 47

Figura 14 – Demonstração gráfica dos elementos envolvidos no cálculo do índice Davies-Bouldin para o grupo C_i, do qual C_j é o grupo mais próximo. C_k representa um grupo qualquer, mais distante. d_i representa as distâncias internas do grupo C_i, d_j, as distâncias internas do grupo C_j e, d_{ij}, a distância entre os centroides de C_i e C_j. 49
Figura 15 – Diferentes configurações de agrupamento no espaço multivariado e suas relações com a soma dos quadrados das distâncias intragrupo ($wcss$). Quanto mais coesos os grupos, menor é o $wcss$ (adaptado de Martin & Boisvert (2018)). 51
Figura 16 – Diferentes configurações de agrupamento no espaço geográfico e suas relações com a entropia espacial (H). Quanto maior a conectividade espacial dos grupos, menor é H (adaptado de Martin & Boisvert (2018)). 51
Figura 17 – Valores de $wcss$ e H para diferentes configurações de agrupamentos, plotados em um gráfico de dispersão, evidenciando sua relação inversa. Quanto maior a coesão dos grupos no espaço multivariado (menor $wcss$), mais desorganizados eles são no espaço geográfico (maior H). Nos detalhes podem ser observadas as configurações no espaço multivariado e no espaço geográfico de algumas das configurações de agrupamentos (ver Figs. 15 e 16) (adaptado de Martin & Boisvert (2018)). 52
Figura 18 – Exemplo de definição de indicadores para uma variável categórica do banco de dados Jura, de Goovaerts (1997) 53
Figura 19 – Quatro tipos comuns de semivariogramas de fenômenos naturais (adaptado de Matheron (1963)). 54
Figura 20 – Exemplo de matriz de confusão para validação de classificadores supervisionados. As colunas representam os valores preditos pelo modelo, as linhas, os valores reais. É desejável que a diagonal principal apresente valores relativamente altos. 57
Figura 21 – Esboço geológico da porção sul da Faixa Brasília (SILVA et al., 2006) . 60
Figura 22 – Amostras de rochas frescas obtidas a partir de testemunhos de sondagem: (A) piroxenito (bebedourito); (B) carbonatito; (C) brecha magmática. . 61
Figura 23 – Perfil esquemático da jazida relacionada a este estudo de caso. No detalhe, foto de uma porção da mina. As espessuras são aproximadas, sendo que cada banco mede 10 metros de altura. 62
Figura 24 – Quadro com as principais tipologias relacionadas aos tipos de rochas encontradas no depósito. 63
Figura 25 – Quadro com as tipologias referentes aos padrões de alteração intempérica no contexto do depósito. 63

Figura 26 – Histogramas do comprimento das amostras. (A) Conjunto de dados pré-regularização (após as etapas (*i*) e (*ii*) do tratamento); (B) Conjunto de dados após a regularização. 65
Figura 27 – Mapa com a distribuição espacial dos furos de sondagem usados neste estudo. 65
Figura 28 – Seções verticais de direção NE-SW (N30°E) mostrando parte representativa dos dados, com amostras simbolizadas de acordo com as variáveis categóricas: (A) intemperismo e (B) litologias. 66
Figura 29 – Seções verticais de direção NE-SW (N30°E) mostrando parte representativa dos dados, com amostras simbolizadas pelos teores de P_2O_5 (A), TiO_2 (B) e CaO (C). 66
Figura 30 – Histogramas das variáveis contínuas. 67
Figura 31 – Gráficos de dispersão das variáveis contínuas presentes no banco de dados, plotadas duas a duas. Na diagonal principal, os histogramas de cada variável. 68
Figura 32 – Matriz com os coeficientes de correlação de Pearson entre as variáveis contínuas. Células azuis representam correlações negativas e células vermelhas, correlações positivas. 69
Figura 33 – Gráficos de barras com a quantificação de amostras classificadas por intemperismo (A) e por litologia (B). 69
Figura 34 – *Boxplots* de teores por tipologia definida pelo intemperismo. 70
Figura 35 – *Boxplots* de teores por litologia. 70
Figura 36 – Diagramas de dispersão das principais variáveis contínuas (P_2O_5, TiO_2 e CaO) presentes no banco de dados, plotadas duas a duas. As cores se referem às tipologias: intemperismo (A) e litologia (B). 71
Figura 37 – Variogramas experimentais das principais variáveis contínuas (P_2O_5 (A), TiO_2 (B) e CaO (C)) em diferentes direções: plano horizontal (*i*); direção vertical (*ii*). 72
Figura 38 – dendrograma obtido com a aplicação do agrupamento aglomerativo hierárquico, mostrando a tendência natural que os dados têm de se agrupar (no espaço multivariado). As cores indicam as linhas e conexões caso os dados fossem aglomerados em oito grupos 73
Figura 39 – Índices de Davies-Bouldin, Silhueta e Calinski-Harabasz e métricas *wcss* e H (nos eixos verticais) para diferentes números de grupos (k) (nos eixos horizontais). Para Davies-Bouldin, *wcss* e entropia espacial são desejáveis valores baixos, já para Silhueta e Calinski-Harabasz, valores altos. 75

Figura 40 – Gráfico de dispersão *wcss versus H*, que permite avaliar simultaneamente a conectividade espacial e a organização multivariada dos grupos nas diferentes configurações de agrupamentos. Cada ponto corresponde a uma configuração distinta (a cor indica o algoritmo, e o ícone, o número de grupos). 76
Figura 41 – Gráficos de dispersão entre os índices de Davies-Bouldin, Silhueta, Calinski-Harabasz e métricas *wcss* e *H*, plotados dois a dois. Para Davies-Bouldin, *wcss* e entropia espacial são desejáveis valores baixos, já para Silhueta e Calinski-Harabasz, valores altos. 77
Figura 42 – Parâmetros de busca para o cálculo dos correlogramas experimentais dos indicadores. 78
Figura 43 – Correlogramas experimentais dos indicadores no cenário com quatro agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical. 78
Figura 44 – Correlogramas experimentais dos indicadores no cenário com cinco agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical. 79
Figura 45 – Correlogramas experimentais dos indicadores no cenário com seis agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical. 79
Figura 46 – Correlogramas experimentais dos indicadores no cenário com sete agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical. 80
Figura 47 – Seções verticais mostrando parte dos dados, com amostras simbolizadas de acordo com os grupos aos quais pertencem nos cenários de agrupamento por *dsclus* em quatro (A), cinco (B), seis (C) e sete (D) grupos. 81
Figura 48 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com quatro agrupamentos por *dsclus*. 82
Figura 49 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com cinco agrupamentos por *dsclus*. 82
Figura 50 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com seis agrupamentos por *dsclus*. 83
Figura 51 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com sete agrupamentos por *dsclus*. 83

Figura 52 – Gráficos de barras mostrando os quantitativos de amostras por intemperismo (A) e litologia (B) em cada cenário de agrupamento pelo algoritmo *dsclus*: quatro (*i*), cinco (*ii*), seis (*iii*) e sete (*iv*) grupos. 84
Figura 53 – Quadros com a caracterização de cada um dos grupos, em cada um dos cenários, caso utilizados para constituir domínios para modelagem. . . 85
Figura 54 – Métricas globais para a validação do classificador para cada *fold*. 89
Figura 55 – Matrizes de confusão para a validação do classificador para cada *fold*. . 89
Figura 56 – Métricas e matriz de confusão para a validação final do classificador com os 2.902 dados de teste. 90
Figura 57 – Seções verticais mostrando parte dos 2.902 dados de teste, simbolizados com as categorias atribuídas pelo classificador automático. 90
Figura 58 – *Boxplots* mostrando a distribuição estatística dos 2.902 dados de teste, após a classificação com o modelo obtido com o algoritmo de florestas aleatórias. 91

Lista de abreviaturas e siglas

ReV	Variável regionalizada (“regionalized variable”)
RV	Variável aleatória (“random variable”)
RF	Função aleatória (“random function”)
MGMs	Modelos Gaussianos Mistos
AM	Aprendizado de máquina
DBSCAN	Agrupamento espacial baseado em densidade de aplicações com ruído (“Density-Based Spatial Clustering of Applications with Noise”)
dsclus	Algoritmo para o agrupamento de dados espaciais em espaço duplo
acclus	Algoritmo para o agrupamento de dados espaciais por estatísticas de autocorrelação
H	Entropia espacial
wcss	Soma dos quadrados intragrupo (“within cluster sum of squares”)
vmp	Vizinhos mais próximos

Lista de símbolos

Σ = somatório

σ = desvio padrão

$\rho(h)$ = função correlograma

$C(h)$ = função covariograma

$\gamma(h)$ = função semivariograma

Cov = covariância

Sumário

1 INTRODUÇÃO 19
1.1 Motivação e meta principal 21
1.2 Objetivos específicos 21
1.3 Metodologia 22
1.4 Fluxo de trabalho proposto 22
1.5 Estrutura da Dissertação 23

2 FUNDAMENTAÇÃO TEÓRICA 24
2.1 Variáveis regionalizadas e funções aleatórias 24
2.2 Estacionariedade 25
2.2.1 Esperança matemática, ou momento de primeira ordem 26
2.2.2 Momentos de segunda ordem 26
2.2.3 A decisão de estacionariedade 27
2.3 Análise de agrupamento ("cluster analysis") 30
2.3.1 Métodos tradicionais de agrupamento de dados 33
2.3.1.1 Método aglomerativo hierárquico 34
2.3.1.2 *k-means* 36
2.3.1.3 Modelos Gaussianos Mistos 37
2.3.1.4 Agrupamento espacial baseado em densidade de aplicações com ruído (*Density-Based Spatial Clustering of Applications with Noise – DBSCAN)* 38
2.3.2 Agrupamento de dados espaciais 40
2.3.2.1 Agrupamento espacial por restrição de vizinhança 41
2.3.2.2 Agrupamento espacial por estatísticas de autocorrelação 43
2.3.3 Sobre a parametrização e as escolhas do número de grupos e do algoritmo . 44
2.4 Avaliação dos resultados de agrupamentos 45
2.4.1 Método das silhuetas 46
2.4.2 Índice Davies-Bouldin 48
2.4.3 Índice Calinski-Harabasz 49
2.4.4 Avaliação em espaço duplo 50
2.4.5 Abordagem dos indicadores na validação dos agrupamentos 53
2.5 Sobre a aplicação do aprendizado supervisionado na classificação automática de novas amostras 56
2.5.1 O aprendizado supervisionado 56

3 ESTUDO DE CASO 59
3.1 Contextualização geológica 59

3.2 Apresentação e validação dos dados . . . 64
3.3 Análise exploratória de dados . . . 65
3.4 Metodologias para a aplicação dos algoritmos de agrupamento e das técnicas de validação . . . 73
3.5 Apresentação de resultados . . . 75
3.5.1 Cálculos de índices e métricas e seleção dos cenários mais promissores . . . 75
3.5.2 Verificação da continuidade espacial dos grupos através de correlogramas dos indicadores . . . 78
3.5.3 Avaliação visual . . . 80
3.5.4 Avaliação da distribuição estatística dos grupos e comparação de agrupamentos com as tipologias . . . 81
3.6 Discussão dos resultados da análise de agrupamento . . . 86
3.7 Aplicação de um classificador supervisionado para a inclusão de novas amostras . . . 87

4 CONSIDERAÇÕES FINAIS . . . 92
4.1 Conclusões . . . 92
4.2 Sugestões para Trabalhos Futuros . . . 93

REFERÊNCIAS . . . 95

APÊNDICES 99

APÊNDICE A – PARÂMETROS DE ENTRADA DOS ALGORITMOS UTILIZADOS . . . 100
A.1 Algoritmos de agrupamento de dados . . . 100
A.1.1 *k-means* . . . 100
A.1.2 Aglomerativo hierárquico . . . 100
A.1.3 Agrupamento em espaço duplo (*dsclus*) . . . 100
A.1.4 Agrupamento por estatísticas de autocorrelação (*acclus*) . . . 101
A.2 Classificação por Florestas Aleatórias . . . 101

1 Introdução

Como já apontado por Sinclair & Blackwell (2004), a definição de estacionariedade pode ser de difícil compreensão. Nesta dissertação, o conceito apresenta estreita relação com a homogeneidade de corpos geológicos. Simplificando a definição de Journel & Huijbregts (1978), assumiremos que um fenômeno é estacionário quando apresenta média, variância e estrutura de autocorrelação constantes através da área de estudo.

Na prática, a hipótese de estacionariedade não pode ser adotada para um depósito mineral como um todo, devendo ser considerados modelos de funções aleatórias diferentes para cada domínio, que devem ser estimados ou simulados separadamente. No entanto, definir domínios estacionários pode ser uma tarefa complexa, principalmente na geoestatística multivariada (MARTIN; BOISVERT, 2017).

Via de regra, em geociências, domínios estacionários estão associados a características geológicas. No entanto, apenas a partir de um conjunto de dados que reproduzam as características do fenômeno com certa fidelidade pode-se acessar essa informação, sendo necessário um criterioso processo de análise. A rigor, cada domínio estacionário apresenta sua própria função de distribuição de probabilidades e continuidade espacial, que são distintas de outros domínios.

Métodos não supervisionados de aprendizado de máquina, especialmente algoritmos de análise de agrupamento ("cluster analysis"), têm sido empregados para reconhecer padrões em dados multivariados, e assim auxiliar na definição de domínios geoestatísticos, o que pode possibilitar a obtenção de modelos de recursos mais aderentes aos fenômenos que descrevem.

Tan et al. (2006) definem a análise de agrupamento como um processo que agrupa dados com base somente nas características dos dados e em suas relações. O objetivo é que elementos de um determinado grupo sejam semelhantes entre si e, ao mesmo tempo, diferentes daqueles pertencentes a outros grupos.

Os algoritmos tradicionais de agrupamento (e.g. aglomerativo hierárquico, *k-means*, *DBSCAN*), em geral, tratam propriedades puramente estatísticas, e suas aplicações em bancos de dados geoestatísticos são limitadas, uma vez que a correlação espacial e as propriedades geológicas das amostras não são consideradas (ROSSI; DEUTSCH, 2014). Mais recentemente, técnicas têm sido desenvolvidas para tratar especificamente da análise de agrupamento de dados cuja posição no espaço é de grande relevância (e.g. Romary et al. (2012), Scrucca (2005), Martin & Boisvert (2018)).

São inúmeras as técnicas de agrupamento apresentadas na literatura e muitas

já se encontram implementadas em bibliotecas de aprendizado de máquina, como a do projeto *Scikit-learn* (PEDREGOSA et al., 2011). Esta dissertação aborda algumas delas no capítulo destinado à fundamentação teórica. Em um estudo de caso, são aplicados os seguintes algoritmos, de particular interesse: (*i*) método aglomerativo hierárquico (SOKAL; SNEATH, 1963); (*ii*) *k-means* (MACQUEEN, 1967); (*iii*) agrupamento espacial por estatísticas de autocorrelação (SCRUCCA, 2005) e (*iv*) agrupamento em espaço duplo (MARTIN; BOISVERT, 2018).

Diferentes algoritmos podem fornecer resultados consideravelmente distintos, mas uma questão comum a todos eles é a necessidade de parametrização e validação, uma vez que a aplicação e os resultados são um tanto subjetivos. Fatores como o número de grupos, as métricas para as conexões entre grupos e o método para integrar a correlação espacial são exemplos de parâmetros que devem ser definidos *a priori* pelo usuário.

A validação dos resultados também pode ser um desafio, já que os dados não são rotulados e, logo, não há um "gabarito" para referência, como é característico dos métodos não supervisionados de aprendizado de máquina. Há algumas técnicas que servem a esse propósito, como o método das silhuetas (ROUSSEEUW, 1987) e o índice Davies-Bouldin (DAVIES; BOULDIN, 1979) que, no entanto, são mais adequadas a agrupamentos com formatos convexos, o que quase nunca é o caso dos dados geoestatísticos, aos quais devem ser aplicadas com cuidado.

Ainda nesse contexto, Martin (2019) menciona que a validação dos agrupamentos de dados espaciais não é tipicamente discutida na literatura. Segundo aquele autor, os domínios gerados podem até fazer sentido estatístico, de modo que as características estacionárias de primeira e segunda ordem sejam satisfeitas, mas, em geral, medidas quantitativas de validação ainda não foram desenvolvidas.

Martin & Boisvert (2018) sugerem uma abordagem qualitativa para medir a eficiência do agrupamento de dados geoestatísticos, de modo que tanto seu caráter estatístico multivariado, quanto seu caráter espacial sejam considerados. Em seu trabalho, aqueles autores aplicam os conceitos de *wcss* (*within cluster sum of squares*, ou soma dos quadrados intragrupo) e entropia espacial para medir a coesão dos grupos no espaço multivariado e sua conectividade espacial, respectivamente.

Assim, não existe algo como um "algoritmo definitivo" para o agrupamento de dados, cabendo ao usuário uma série de decisões baseadas na experiência e no bom-senso, quase sempre através da comparação de resultados de diferentes métodos e parâmetros.

Esta dissertação apresenta os resultados da investigação de diversos algoritmos de agrupamento e da aplicação de alguns deles a dados espaciais. Além disso, discorre sobre técnicas de validação, indo um pouco além e sugerindo uma abordagem adicional para a validação da conectividade espacial dos agrupamentos gerados, por meio da medida da

continuidade espacial dos indicadores.

Por fim, é discutida a possibilidade do emprego de algoritmos de aprendizado supervisionado para a incorporação de novas amostras, de modo que a classificação dessas seja condizente com os grupos definidos na análise de agrupamento. A ideia é que as mesmas relações sejam seguidas, sem que, para isso, haja a necessidade de executar todo o processo de análise novamente, cada vez que os dados são atualizados.

1.1 Motivação e meta principal

Em geociências, e ainda mais especificamente, na indústria mineira, a determinação de domínios geoestatísticos utilizados na definição de recursos e reservas é quase sempre feita de maneira muito subjetiva, de acordo apenas com o conhecimento geológico prévio, sem considerar apropriadamente as relações estatísticas e espaciais dos atributos e amostras. Muitas vezes, os processos de formação, ou processos posteriores, como a alteração intempérica, podem levar a grupos geoestatísticos distintos das litologias que ocorrem em determinada região.

Essa abordagem pode levar à definição insatisfatória de domínios para modelagem. Quando exagerada, ou seja, sendo definido um número excessivo de domínios, aumentam-se desnecessariamente o tempo e a energia despendidas pelo geomodelador e, quando imprópria, pode ocasionar a mistura de populações estatísticas, comprometendo as estimativas e levando a complicações no controle de qualidade dos materiais provenientes da mina.

Este trabalho tem como meta principal a investigação de métodos de agrupamento de dados, considerando suas implicações nas relações estatísticas e posicionais, a fim de se definir domínios para a estimativa de recursos minerais. Como meta adicional, propõe-se a verificação do uso de algoritmos supervisionados para a classificação de novas amostras, a serem incorporadas ao banco de dados após o processo de agrupamento.

1.2 Objetivos específicos

Para cumprir as metas propostas, foram traçados objetivos específicos, de modo a tornar os problemas mais tangíveis:

(i) Investigar métodos de agrupamento de dados aplicados a Geociências;

(ii) Avaliar os agrupamentos definidos, através de técnicas de validação estatísticas e geoestatísticas;

(iii) Verificar a possibilidade do uso de algoritmos supervisionados para a classificação automática de novas amostras;

(iv) Aplicar a metodologia através de um estudo de caso em um banco de dados real e avaliar os resultados.

1.3 Metodologia

Para se atingir os objetivos propostos, a seguinte metodologia foi aplicada:

(i) Revisão da literatura e investigação dos aspectos teóricos de variados métodos de agrupamento;

(ii) Testes com diferentes algoritmos de agrupamento de dados, bem como de técnicas específicas para a validação e seleção dos métodos mais relevantes;

(iii) Calibração de um classificador por florestas aleatórias para inclusão de novas amostras ao banco de dados com classificação automática;

(iv) Compilação dos códigos em linguagem Python® para operacionalização de uma rotina computacional em *Jupyter Notebook*;

(v) Aplicação da rotina em um banco de dados real 3D multivariado;

(vi) Considerações e sugestões para trabalhos futuros.

1.4 Fluxo de trabalho proposto

Em linhas gerais, esta Dissertação propõe um método que pode ser melhor compreendido em duas etapas principais: (*i*) aplicação da análise de agrupamento para a definição de domínios e, (*ii*) emprego de aprendizado supervisionado para a classificação automática de novas amostras. O fluxograma da Figura 1 mostra o fluxo de trabalho proposto.

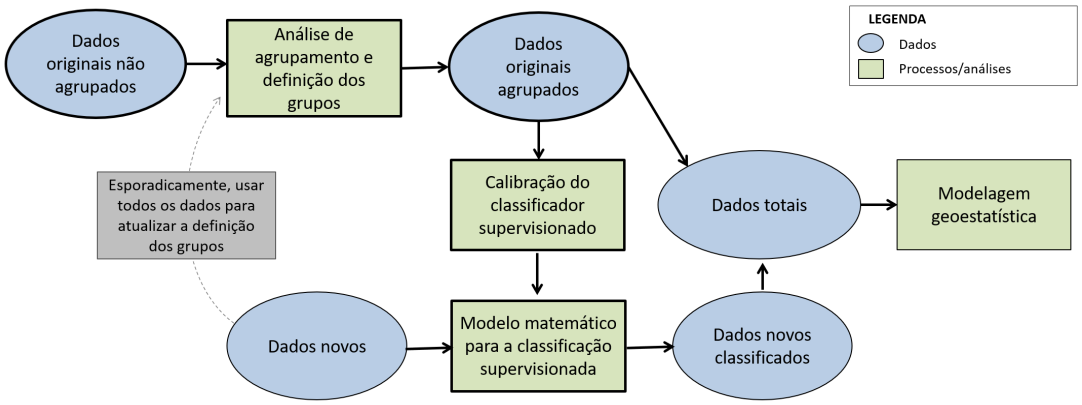

Figura 1 – Fluxograma geral do método proposto, incluindo as etapas de análise de agrupamento e classificação automática de novas amostras. Conforme indicado pela linha tracejada, esporadicamente, a análise de agrupamento deve ser revisitada, de modo a incorporar a totalidade dos dados. O classificador deverá ser, então, atualizado

.

1.5 Estrutura da Dissertação

O Capítulo 1 apresenta os aspectos introdutórios dos assuntos abordados, a motivação e a meta principal do trabalho, bem como seus objetivos, a metodologia aplicada e o fluxo de trabalho proposto.

O Capítulo 2 apresenta os resultados da revisão bibliográfica, com a descrição dos fundamentos teóricos dos tópicos abordados. Trata dos conceitos de estacionariedade, aprendizado de máquina, análise de agrupamento e validação de resultados do agrupamento. Algumas das técnicas consideradas mais relevantes são descritas com mais detalhes.

No Capítulo 3, são apresentados e discutidos os resultados da aplicação e validação de quatro técnicas de agrupamento em um conjunto de dados provenientes de furos de sondagem de um depósito de fosfato e titânio, localizado na região sudeste do Brasil. Os resultados são discutidos e comparados com os domínios para estimativa atualmente aplicados segundo as tipologias definidas tradicionalmente na mina. Por fim, são apresentados os resultados da aplicação de um classificador por florestas aleatórias para a classificação de novas amostras.

O Capítulo 4 apresenta discussões e conclusões gerais da dissertação, sugerindo trabalhos futuros relacionados ao tema abordado.

2 Fundamentação teórica

Este capítulo apresenta a fundamentação teórica dos assuntos abordados nesta dissertação. Não somente a análise de agrupamento, tema central do trabalho, mas também tópicos relacionados, cuja compreensão é essencial para o entendimento do tema principal e seus desdobramentos.

Primeiramente, são delineados os conceitos de variáveis regionalizadas e funções aleatórias, fundamentais para a apresentação do conceito de estacionariedade, descrito na sequência. O tema análise de agrupamento é então apresentado em detalhe, com a descrição de alguns algoritmos específicos, julgados mais relevantes no contexto deste trabalho. Logo depois, são abordados conceitos e métodos para a validação dos agrupamentos e, por fim, delineia-se a aplicação de métodos supervisionados para a classificação automática de novas amostras.

2.1 Variáveis regionalizadas e funções aleatórias

A definição de variáveis regionalizadas (ReV) foi introduzida por Matheron (1963), sendo diferente do conceito de variáveis aleatórias, que se refere à noção da estatística clássica, na qual uma variável pode assumir qualquer valor, dependendo de fatores puramente aleatórios.

Diferentemente, uma variável regionalizada apresenta um aspecto espacial, dependente da posição em que a amostra se encontra no espaço. Essa variável pode ser determinada por uma função, que assume um valor definido em cada ponto desse espaço.

De acordo com a definição de Journel & Huijbregts (1978), de um ponto de vista matemático, a ReV é simplesmente uma função $f(x)$ que assume um valor em cada ponto x de coordenadas $(x_{\mathrm{u}}, x_{\mathrm{v}}, x_{\mathrm{w}})$ em um espaço tridimensional.

Em quase todos os depósitos minerais, existem zonas mais ricas que outras, e amostras retiradas de zonas ricas serão, em média, mais ricas que aquelas retiradas de zonas mais pobres. Assim, o valor da ReV $f(x)$ depende da localização de x e apresenta duas características aparentemente contraditórias (JOURNEL; HUIJBREGTS, 1978):

(i) um aspecto local, errático, que remete à noção de variável aleatória (RV);

(ii) um aspecto geral, estruturado, que requer alguma representação funcional.

Uma formulação apropriada deve considerar esse aspecto duplo, de maneira a fornecer uma representação da variabilidade espacial para que se obtenha uma solução

consistente. Uma dessas formulações é a interpretação probabilística proporcionada pelas funções aleatórias (JOURNEL; HUIJBREGTS, 1978).

Consideremos o teor de determinado elemento em um ponto x_1 de um depósito mineral qualquer, que pode ser considerado como uma realização de uma variável aleatória RV $Z(x_1)$, definida no ponto x_1. Assim, o conjunto de teores $z(x)$ de todos os pontos x dentro do depósito (ou seja, a ReV $z(x)$) pode ser considerado como uma realização da RV $[Z(x), x \in \text{depósito}]$. Esse conjunto de RV é denominado como a função aleatória (RF), $Z(x)$.

Essa definição de função aleatória expressa os aspectos aleatório e estruturado da variável regionalizada (JOURNEL; HUIJBREGTS, 1978):

(i) localmente, no ponto x_1, $Z(x_1)$ é uma variável aleatória;

(ii) $Z(x)$ também é uma função aleatória no sentido de que, para cada par de pontos x_1 e $x_1 + h$, as RV correspondentes $Z(x_1)$ e $Z(x_1 + h)$ não são, em geral, independentes, mas relacionadas a uma correlação expressa pela estrutura espacial da variável regionalizada original $Z(x)$.

2.2 Estacionariedade

O conceito de estacionariedade pode ser de difícil definição mas, assim como em Sinclair & Blackwell (2004), no âmbito desta dissertação apresenta estreita relação com o termo "homogeneidade", usado para se determinar domínios com características geológicas similares como, por exemplo, estilos de mineralização, tipos de rochas e estados de alteração. Nesse sentido, um domínio de dados é dito estacionário se a mesma população está sendo amostrada dentro de todo aquele mesmo domínio, independente da localização, ou seja, não existe tendência ("trend") nos dados.

A estacionariedade não é uma propriedade intrinsecamente geológica, mas sim uma suposição feita acerca do fenômeno a ser modelado a partir da observação das características dos dados coletados, mais especificamente de seus momentos de primeira e segunda ordem, isto é, média, variância e covariância. Essas distinções entre determinados grupos de dados podem ser devidas a diversos fatores, como:

- Ocorrência de litotipos distintos;
- Padrões de alteração hidrotermal e/ou intempérica;
- Estratificação das rochas;
- Sobreposição de estruturas geológicas.

Considere a RF $Z(x)$. Para cada conjunto de k pontos no R^n (espaço n-dimensional) $x_1, x_2, ..., x_k$, corresponde um componente vetorial k-dimensional de variáveis aleatórias $\{Z(x_1), Z(x_2), ..., Z(x_k)\}$.

Essa RV vetorial é caracterizada pela função de distribuição k-dimensional:

$$Fx_1, x_2, ..., x_k(z_1, z_2, ..., z_k) = Prob\{Z(x_1) < z_1, ..., Z(x_k) < z_k\}$$

O conjunto com todas essas funções de distribuição, para todos os k inteiros e positivos e para todas as escolhas possíveis de pontos no R^n, constitui a "lei espacial" da RF $Z(x)$ (JOURNEL; HUIJBREGTS, 1978).

Ainda segundo os autores acima, em aplicações na mineração, nunca é necessário caracterizar toda a lei espacial, já que apenas os dois primeiros momentos da lei são suficientes para se obter uma solução aproximada aceitável na maioria dos problemas encontrados. Além disso, a quantidade de dados disponível é geralmente insuficiente pra inferir a lei espacial em sua totalidade.

A seguir, são apresentadas as definições de Journel & Huijbregts (1978) para os momentos de primeira e segunda ordem.

2.2.1 Esperança matemática, ou momento de primeira ordem

Considere a RV $Z(x)$ no ponto x. Se a função de distribuição de $Z(x)$ tem uma esperança matemática (e suponhamos que tenha), então ela é função de x, sendo definida como:

$$E\{Z(x)\} = m(x) \tag{2.1}$$

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

2.2.2 Momentos de segunda ordem

Os três momentos de segunda ordem considerados em geoestatística são:

(i) A variância ou, mais precisamente, a variância *a priori* de $Z(x)$, definida como o momento de segunda ordem em relação a esperança matemática $m(x)$ da RV $Z(x)$:

$$Var\{Z(x)\} = E\left\{[Z(x) - m(x)]^2\right\} \tag{2.2}$$

(ii) A covariância. Pode ser demonstrado que, se duas RV $Z(x_1)$ e $Z(x_2)$ apresentam variâncias nos pontos x_1 e x_2, elas também apresentam uma covariância, que é função das localidades x_1 e x_2, sendo definida como:

$$C(x_1, x_2) = E\left\{[Z(x_1) - m(x_1)]\,[Z(x_2) - m(x_2)]\right\} \tag{2.3}$$

(iii) O variograma, que é uma função definida como a variância do incremento $[Z(x_1) - Z(x_2)]$, e pode ser escrito como:

$$2\gamma(x_1, x_2) = Var\left\{Z(x_1) - Z(x_2)\right\} \tag{2.4}$$

A função $\gamma(x_1, x_2)$ é, então, denominada "semivariograma".

2.2.3 A decisão de estacionariedade

Assim, consideremos que um fenômeno é estacionário quando apresenta média, variância e estrutura de autocorrelação (covariância) constantes através da área de estudo. Entretanto, raramente toda a área pode ser definida como pertencente a um único domínio estacionário. Geralmente, cada porção apresenta propriedades singulares, sejam elas geológicas ou ambientais, atributos que podem refletir tipos de rochas, teores, estado de alteração intempérica, entre outros.

A estacionariedade é uma característica subjetiva, que depende da avaliação e de uma série de decisões tomadas pelo geoestatístico/geomodelador. Há diversas técnicas que permitem a modelagem de funções aleatórias não estacionárias, como usar um modelo de tendência ou modelos de anisotropias locais. Além disso, é interessante subdividir as amostras em grupos e, então, modelar porções do depósito separadamente.

Rossi & Deutsch (2014) afirmam que domínios de estimativa são os equivalentes geológicos às zonas estacionárias geoestatísticas, definidas como volumes de rochas cujos controles de mineralização resultam em distribuições aproximadamente homogêneas daquela mineralização. A compreensão das características estatísticas dos dados, juntamente com o conhecimento geológico, leva à subdivisão do depósito em domínios para estimativa.

De acordo com McLennan (2007), inferências geoestatísticas têm como pré-requisito a decisão da estacionariedade, que assegura a homogeneidade geológica dentro dos domínios definidos. Para que essa decisão seja razoável, cinco fases devem ser consideradas:

(i) Escolher o número e os tipos de domínios nos quais serão modeladas as propriedades petrofísicas de interesse;

(ii) Modelar os contornos desses domínios;

(iii) Quantificar a natureza das transições entre domínios;

(iv) Quantificar tendências determinísticas de grande escala dentro dos domínios;

(v) Prever com um modelo de tendência.

Esta Dissertação aborda a primeira fase sugerida por McLennan (2007) no processo de decisão da estacionariedade, na qual são definidos o número e os tipos de domínios a serem modelados.

Diversas metodologias são comumente usadas para a definição de domínios estacionários para a subsequente modelagem geoestatística, como pode ser visto no fluxograma da Figura 2, de Martin (2019).

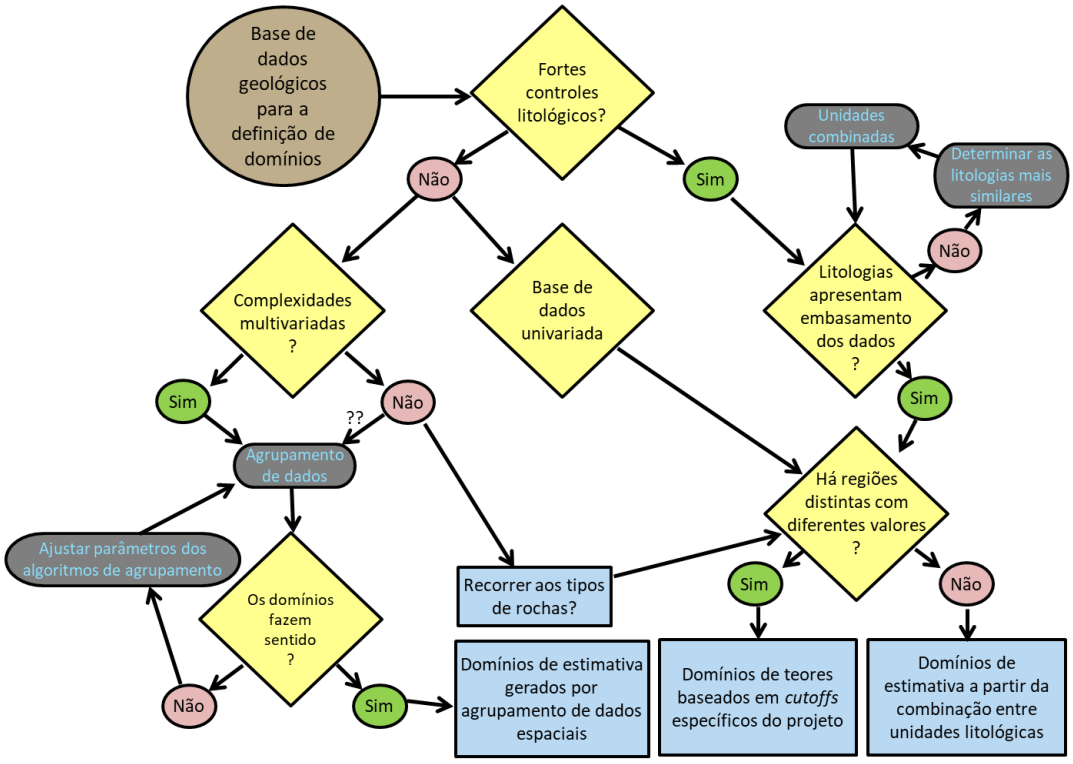

Figura 2 – Possível fluxo para a definição de domínios estacionários considerando as propriedades das variáveis regionalizadas, a configuração espacial e outros tipos de informações geológicas (Adaptado de Martin (2019)).

A seguir, descrições dos critérios mais comuns para a definição de domínios estacionários, de acordo com Martin (2019).

(i) **Configuração geológica**

Quando a distribuição espacial da mineralização é fortemente relacionada às unidades litológicas, os domínios podem ser naturalmente definidos a partir da descrição de litologias em superfície e em testemunhos de sondagem.

A determinação desses domínios pode ser mais ou menos óbvia, uma vez que a sobreposição de eventos geológicos, como tectonismo ou alterações hidrotermais e intempéricas, pode elevar a complexidade do sistema. Por exemplo, em um depósito do tipo porfirítico, a mineralização está associada a sistemas hidrotermais de grande escala que abrangem diferentes tipos de rochas, o que resulta em múltiplos estilos de alteração e mineralização sobrepostos.

(ii) **Combinação de litologias**

Um problema comum em estimativa de recursos é quando ocorrem muitas unidades litológicas, mas amostras insuficientes para apoiar a descrição estatística/geoestatística de cada uma delas, ou mesmo quando a quantidade de litologias presentes no depósito é tamanha que torna a modelagem de todas elas uma questão pouco prática.

Nesse caso, é interessante combinar a grande quantidade de unidades litológicas em um número razoável de domínios de estimativa, que sejam geologicamente relacionados e contenham quantidades razoáveis de dados.

(iii) **Intervalos de teores**

Também é comum na estimativa de recursos minerais o particionamento do corpo de minério em domínios determinados por intervalos de teores (EMERY; ORTIZ, 2005), definindo as chamadas "grade shells". Esse método é tipicamente aplicado em unidades litológicas razoavelmente homogêneas.

Os teores de corte ("cutoffs") que definem os intervalos são específicos de cada projeto, sendo determinantes para a destinação dos materiais lavrados. Esses teores de corte podem ser, também, baseados em uma combinação de fatores que influenciam no tratamento dos minérios.

(iv) **Análise de agrupamento ("cluster analysis")**

Basicamente, a análise de agrupamento separa dados ao identificar padrões e estruturas no espaço multidimensional. Essa análise é feita por algoritmos computacionais, de acordo com as relações entre as variáveis do sistema. Os dados são, então, agrupados em domínios e estarão mais relacionados a outros dados do mesmo domínio do que a elementos de outros domínios.

No entanto, um problema comum à aplicação das técnicas tradicionais de agrupamento (e.g. *k-means*, aglomerativo hierárquico, *DBSCAN*) na modelagem de recursos minerais é que somente parâmetros estatísticos são utilizados, sem levar em conta aspectos geológicos (ROSSI; DEUTSCH, 2014).

Mais recentemente, metodologias têm sido propostas para a análise de agrupamento de dados espaciais, considerando, assim, além das relações multivariadas, a correlação espacial das amostras.

Esse assunto é o tema central desta Dissertação e será abordado em maiores detalhes nas próximas seções.

2.3 Análise de agrupamento ("cluster analysis")

A análise de agrupamento tem um papel importante não só nas Ciências da Terra, mas em uma variedade de temas: ciências sociais, biologia, estatística, mineração de dados, entre outros.

O simples ato de agrupar objetos com características similares desempenha um papel importante na maneira como analisamos o mundo à nossa volta. Como já dito por Tan et al. (2006), nós, seres humanos, somos muito eficientes ao dividir objetos em diferentes grupos, processo que pode ser definido como agrupamento. Também o somos ao classificar novos elementos de acordo com grupos predeterminados, o que, por sua vez, pode ser definido como classificação.

Técnicas para o agrupamento de dados com base em suas características, sob o grande tema denominado análise de agrupamento, têm sido amplamente investigadas desde a década de 1960 (e.g. Sokal & Sneath (1963), MacQueen (1967)). Mais recentemente, esses métodos têm sido empregados para definir domínios estacionários de dados espacialmente posicionados (e.g. Oliver & Webster (1989), Ambroise et al. (1997), Scrucca (2005), Martin & Boisvert (2018)). A aplicação dessas técnicas na indústria mineral pode possibilitar a obtenção de modelos de recursos mais coerentes com a realidade.

A análise de agrupamento é um tema inserido no contexto de aprendizado de máquina, mais especificamente, no aprendizado não supervisionado, no qual devem ser encontrados padrões em dados que não apresentam rotulação prévia. Em outras palavras, os algoritmos de aprendizado não supervisionado são capazes de encontrar estruturas internas ocultas e, a partir disso, relacionar grupos de dados. Diferentemente, os algoritmos de aprendizado supervisionado constroem modelos matemáticos capazes de reconhecer padrões a partir de dados rotulados, correlacionando variáveis de entrada ("inputs") às variáveis de saída ("outputs") desejadas.

Tan et al. (2006) definem a análise de agrupamento como um processo que agrupa dados com base somente em suas características e relações. O objetivo é que objetos de um determinado grupo sejam semelhantes entre si e, ao mesmo tempo, diferentes de objetos pertencentes a outros grupos. Quanto maior a semelhança (ou homogeneidade) dentro de um grupo, e a diferença entre grupos, mais eficiente é o processo de agrupamento.

Em análise exploratória, pode ser muito útil entender como os dados podem ser agrupados. Por exemplo, para definição de tipologias e domínios de estimativa em um depósito mineral, como já dito na Seção anterior.

No entanto, em muitas aplicações a noção de análise de agrupamento pode não ser muito clara. Por exemplo, a Figura 3 mostra vinte pontos e três maneiras distintas de dividi-los em grupos. No entanto, a divisão aparente dos dois grandes grupos (Figura 3(B)) em sub-grupos (Figura 3(C) e (D)) pode ser apenas uma questão da visão humana. Assim, a Figura 3 mostra que a análise de agrupamento é imprecisa, e que a melhor definição depende da natureza dos dados e dos resultados almejados (TAN et al., 2006).

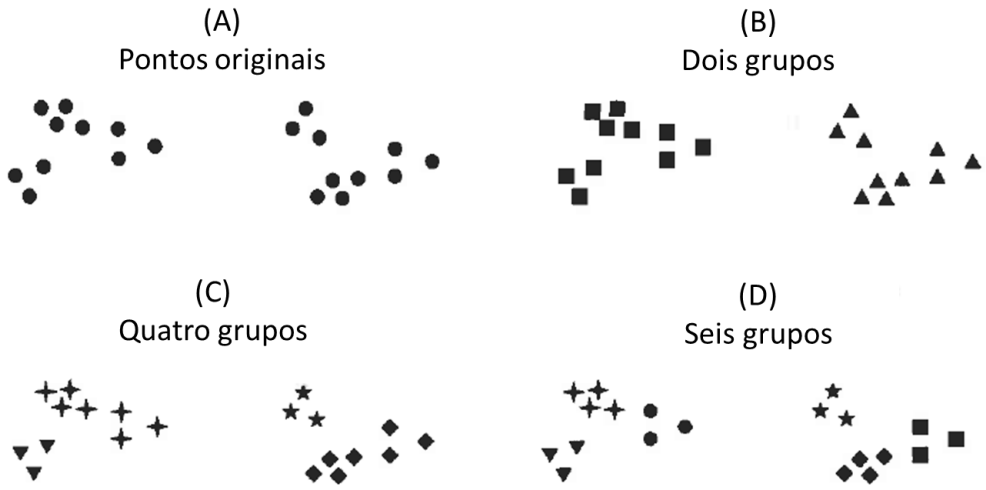

Figura 3 – Diferentes maneiras de se agrupar um mesmo conjunto de dados. Em (A), os pontos originais, não agrupados. Em (B), conjunto dividido em dois grandes grupos e, em (C) e (D), duas maneiras distintas de se subdividir os dois grandes grupos de (B) (TAN et al., 2006).

Técnicas de agrupamento podem ser facilmente compreendidas com exemplos bivariados, uma vez que somos visualmente capazes de identificar esses grupos quando em baixa dimensionalidade. No entanto, o problema pode se tornar um tanto desafiador quando em espaço multidimensional, fazendo com que a aplicação de algoritmos computacionais seja necessária.

A análise de agrupamento pode ser aplicada a qualquer tipo de dados e, a depender da abordagem, os algoritmos podem ser organizados em, basicamente, dois tipos (TAN et al., 2006):

(i) **Baseados em protótipos** ("Prototype-based"): o equivalente ao que Schubert et al. (2017) chamam de métodos de particionamento ("partitioning methods"). Nesta definição, cada grupo é formado por um conjunto de dados no qual cada um deles é mais similar ao protótipo (ponto central) que define esse grupo do que ao protótipo de qualquer outro grupo. Para variáveis contínuas, geralmente esse protótipo é o centro de gravidade de sua distribuição (e.g. a esperança matemática, ou média de todos os pontos). Esses grupos tendem a apresentar formato globular, com contornos convexos.

(ii) **Baseados em gráfico** ("Graph-based"): compatível ao que Schubert et al. (2017) definem como métodos hierárquicos. Neste caso, as relações entre os dados podem ser representadas em um gráfico no qual as observações são nós, e as conexões, os *links* entre elas. O agrupamento, nesta definição, organiza os dados de acordo com esses nós e conexões. Dados representados por nós interconectados pertencem a um determinado grupo, já que não estão conectados a dados de outro grupo no gráfico. Os grupos vão sendo aglomerados entre si em cada iteração do algoritmo se os nós que os definem são considerados suficientemente similares.

A perspectiva gráfica será ilustrada mais adiante, quando for apresentado o método aglomerativo hierárquico, um exemplo clássico e de ampla aplicação de algoritmo do tipo baseado em gráfico, tendo sido originalmente desenvolvido no campo da taxonomia por Sokal & Sneath (1963). O *k-means* e os modelos gaussianos mistos (MGMs) são exemplos de aplicações baseadas em protótipos.

A ampla gama de algoritmos disponíveis na literatura e nas bibliotecas de aprendizado de máquina torna a escolha do método, em si, uma tarefa complexa, e novos algoritmos surgem a cada momento. É tentador, porém impraticável, listar e discorrer sobre todos eles, o que, na verdade, extrapolaria o escopo deste trabalho. Ao invés disso, foram selecionadas algumas das técnicas julgadas mais relevantes neste contexto, que serão detalhadas nas seções seguintes. São elas:

(i) Método aglomerativo hierárquico (SOKAL; SNEATH, 1963);

(ii) *k-means* (MACQUEEN, 1967);

(iii) Modelos Gaussianos Mistos (como descrito por VanderPlas (2016));

(iv) *DBSCAN* (ESTER et al., 1996);

(v) Agrupamento por estatísticas de autocorrelação local (SCRUCCA, 2005);

(vi) Agrupamento em espaço duplo (MARTIN; BOISVERT, 2018).

Outro aspecto importante é quanto à natureza dos dados. Tradicionalmente, a análise de agrupamento tem sido utilizada para se particionar amostras em um espaço multivariado, com base em parâmetros puramente estatísticos, sem considerar sua posição no espaço geográfico, questão essencial a bancos de dados geoestatísticos. Nas últimas décadas, trabalhos têm sido desenvolvidos para tratar bancos de dados cuja localização é de fundamental importância na definição dos grupos.

Os quatro primeiros algoritmos da lista acima podem ser reunidos no que se define como algoritmos tradicionais, que podem ser aplicados, pelo menos à princípio, a qualquer

tipo de dado. Os dois últimos foram concebidos especialmente para tratar dados cuja posição no espaço é relevante, sendo definidos como algoritmos de agrupamento de dados espaciais.

A escolha da técnica a ser aplicada é de responsabilidade do usuário e deve ser feita de acordo com a natureza dos dados e do objetivo do estudo. É uma escolha flexível, subjetiva, e diferentes algoritmos podem servir bem a um mesmo conjunto de dados, não existindo algo como um algoritmo definitivo de análise de agrupamento.

Além disso, cada algoritmo pode fornecer resultados consideravelmente distintos, mesmo quando aplicados à mesma base de dados. A Figura 4 mostra os resultados de alguns métodos empregados aos mesmos bancos de dados, gerados através da biblioteca de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011).

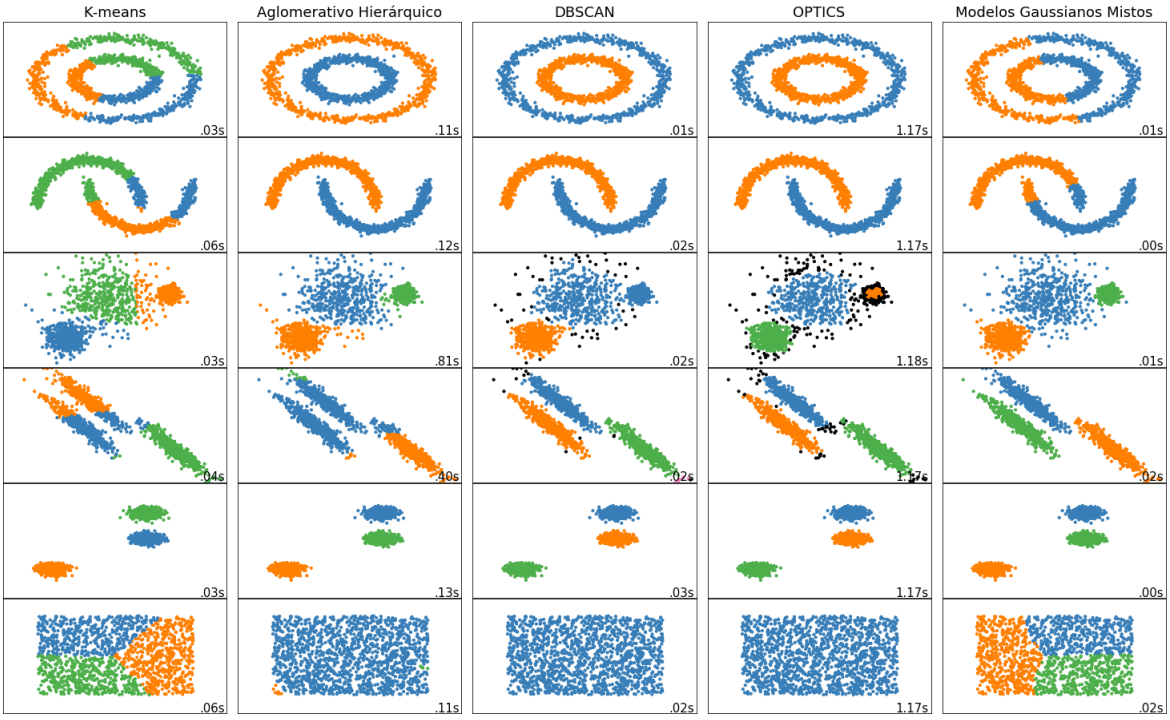

Figura 4 – Exemplos das diferenças nos resultados que podem ser obtidos, para os mesmos conjuntos de dados, com distintos algoritmos de análise de agrupamento, disponíveis na biblioteca de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011). Cada coluna corresponde a um algoritmo diferente, sendo, da esquerda para a direita: *k-means*, aglomerativo hierárquico, *DBSCAN*, *OPTICS* e Modelos Gaussianos Mistos.

2.3.1 Métodos tradicionais de agrupamento de dados

Métodos tradicionais são aqueles que podem ser aplicados, pelo menos teoricamente, a qualquer tipo de dado. Consideram puramente as relações estatísticas dos atributos no espaço M-dimensional, ignorando as relações das amostras no espaço geográfico.

Em geral, tendem a produzir grupos coesos no espaço multivariado, o que pode ser visualizado como fronteiras bem definidas em gráficos de dispersão.

2.3.1.1 Método aglomerativo hierárquico

O agrupamento aglomerativo hierárquico inicialmente categoriza cada observação como um único grupo, antes de aglomerar iterativamente os grupos mais próximos, até que apenas um grande grupo permaneça. O método foi originalmente desenvolvido no campo da taxonomia (SOKAL; SNEATH, 1963) antes de se popularizar em outros campos científicos.

Neste método, cada ponto é inicialmente definido como o centro de um grupo individual e então, a cada etapa, os pontos vão sendo aglomerados de acordo com suas proximidades. O conceito de proximidade será abordado mais adiante. Os resultados podem ser mais intuitivamente visualizados através de gráficos denominados dendrogramas (daí sua definição como método gráfico), como pode ser observado na Figura 5, de Faber (1994), na qual um cenário simples com apenas uma variável é considerado.

Quanto mais atributos são considerados, mais complexo o problema se torna. A Figura 6, de Barnett & Deutsch (2015), ilustra um cenário com duas variáveis. Em cenários com mais variáveis, não é mais possível observar as relações em um gráfico de dispersão bidimensioinal, mas os dendrogramas ainda podem ser muito úteis.

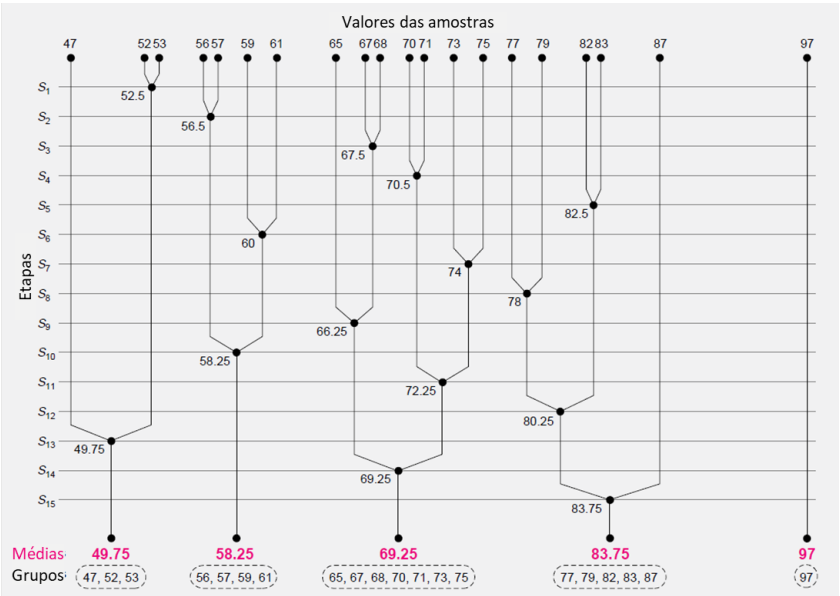

Figura 5 – Cenário ilustrativo de um agrupamento aglomerativo hierárquico com uma variável, no qual 20 amostras (representadas pelos pontos na parte superior da figura) são sucessivamente aglomeradas, par a par, até a formação de cinco grupos. Cada etapa do algoritmo é representada por uma linha transversal que corta o dendrograma, e cada nó representa uma aglomeração entre amostras e/ou grupos. Após quinze etapas (S_1 a S_{15}), as 20 amostras foram aglomeradas em cinco grandes grupos (adaptado de Faber (1994)).

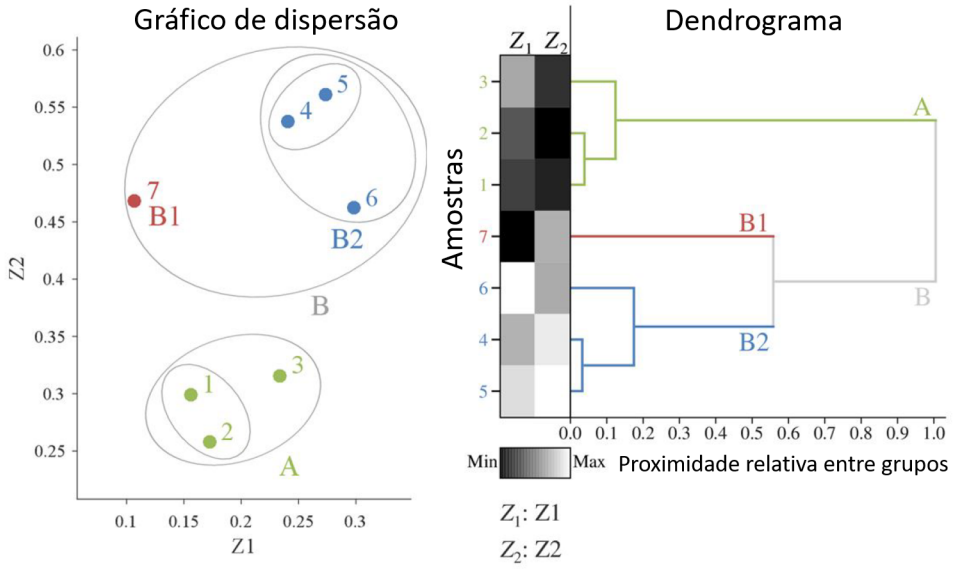

Figura 6 – Exemplo do método aglomerativo hierárquico para duas variáveis (Z_1 e Z_2) em que são exibidos o gráfico de dispersão de sete amostras e o respectivo dendrograma. Amostras dos grupos A, B1 e B2 são identificadas com cores diferentes, que também estão presentes nas linhas e conexões do dendrograma (BARNETT; DEUTSCH, 2015).

No método hierárquico de agrupamento, um conceito-chave é o de proximidade entre grupos, ou seja, o critério aplicado para definir a distância necessária para que dois conjuntos sejam aglomerados. Algumas opções se baseiam em uma noção gráfica de proximidade, que é o caso dos critérios de distância mínima, distância máxima e média grupal (Figura 7). Um outro critério, denominado *Ward's*, baseia-se em uma perspectiva de protótipos, em que cada grupo é representado por um centroide, e a distância para que sejam aglomerados é a distância definida entre os respectivos centroides. A seguir as descrições de cada um dos critérios mencionados.

(i) **Distância mínima** - Também referido na literatura como "single link", neste caso, a distância entre grupos é definida como a distância entre os dois pontos mais próximos que se encontram em grupos diferentes;

(ii) **Distância máxima** - Também denominado na literatura como "complete link", ao contrário do método de distância mínima, neste caso, a distância entre grupos é definida como a distância entre os dois pontos mais distantes que se encontram em grupos diferentes;

(iii) **Média grupal** - A distância entre grupos é definida pela média das distâncias de cada ponto de um grupo com todos os outros pontos de outro grupo, par a par;

(iv) ***Ward's*** - Assume que cada grupo é representado por seu centroide e a medida de distância entre grupos é feita em termos do aumento da soma dos erros ao quadrado

(*SSE*, na sigla em inglês) que resulta da aglomeração desses grupos. Similarmente ao método *k-means* de agrupamento (apresentado em seguida), busca minimizar a soma do quadrado das distâncias dos pontos aos centroides dos grupos aos quais pertencem.

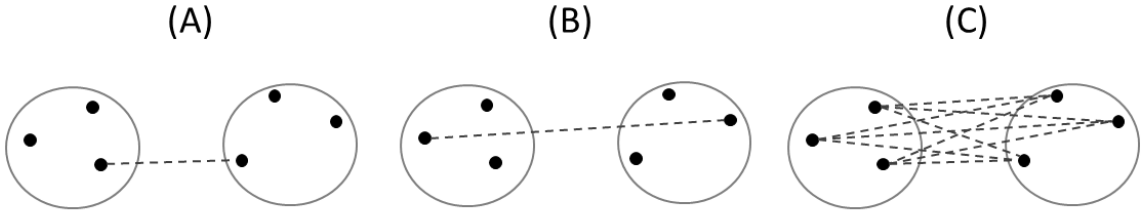

Figura 7 – Representação gráfica das distâncias usadas em alguns dos critérios de proximidade. Os pontos representam amostras. (A) Distância mínima, ou "single link", (B) Distância máxima, ou "complete link", (C) Média grupal (adaptado de Tan et al. (2006)).

2.3.1.2 *k-means*

Método do tipo baseado em protótipos, proposto originalmente por MacQueen (1967), e que leva o nome devido aos k centroides localizados no centro de distribuição (geralmente a média – "mean", em inglês) dos k grupos definidos por eles. O processo consiste em particionar uma população M-dimensional em k conjuntos de dados, baseado nas distribuições de probabilidades das variáveis consideradas.

Primeiramente, são definidas as posições dos k centroides aleatoriamente, com base no número de grupos que se deseja dividir os dados. Cada ponto amostral é então atribuído ao grupo com centroide mais próximo. A posição de cada centroide então é atualizada, com base na configuração dos pontos de cada grupo. O processo é repetido até que os centroides não mais se modifiquem (TAN et al., 2006).

Na Figura 8, adaptada de Tan et al. (2006), pode-se observar um esquema do funcionamento do algoritmo. No primeiro passo (iteração 1), os pontos são assinalados com relação aos centroides iniciais, posicionados dentro da maior aglomeração de pontos (à cada grupo corresponde uma cor: vermelha, azul ou verde). Após a atribuição dos pontos a seus respectivos centroides, estes são atualizados, de acordo com suas respectivas distribuições. Então, as atribuições dos pontos são revistas, de acordo com o centroide mais próximo, e a posição dos centroides é atualizada novamente. Nas iterações 2, 3 e 4, dois dos centroides se deslocam para os conjuntos menores, localizados na porção inferior da figura. Quando não há mais modificações significativas nas posições dos centroides, o algoritmo é finalizado.

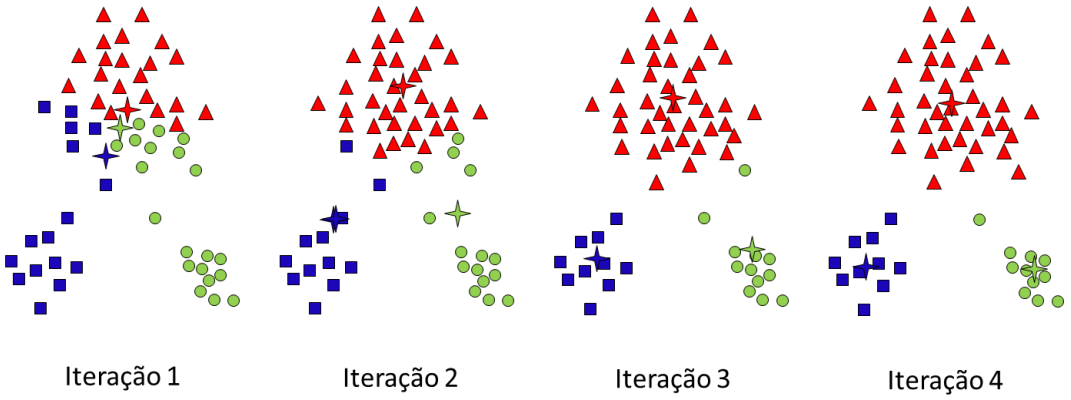

Figura 8 – Representação do funcionamento do método *k-means* para definição de três grupos (adaptado de Tan et al. (2006)).

Uma questão um tanto quanto indesejável do algoritmo é a inicialização aleatória dos centroides, que pode levar a resultados consideravelmente distintos a cada execução. Por isso, alternativas já têm sido propostas para melhorar a definição dos grupos com relação a seleção inicial dos centroides. Uma solução interessante é aquela denominada *"k-means++"* (ARTHUR; VASSILVITSKII, 2007), que se encontra implementada na biblioteca de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011), na qual os pontos de inicialização do algoritmo buscam ser mais afastados, levando a resultados mais razoáveis.

2.3.1.3 Modelos Gaussianos Mistos

Modelos Gaussianos Mistos (MGMs) podem também ser aplicados à análise de agrupamento, como bem descrito por VanderPlas (2016), de modo que os dados sejam interpretados como pertencentes a uma mistura de distribuições de probabilidades. De acordo com Barnett & Deutsch (2015), os MGMs podem ser entendidos como um método baseado em protótipos em análise de agrupamento, em que cada distribuição de probabilidade (modelo Gaussiano) é um protótipo que representa um grupo. Os dados são assinalados ao grupo ao qual apresentam a maior probabilidade de pertencer.

Segundo Pedregosa et al. (2011), um modelo Gaussiano misto é um modelo probabilístico que assume que todos os pontos são gerados a partir de uma mistura de um número finito de distribuições Gaussianas com parâmetros desconhecidos. Assim, em agrupamento de dados, pode-se pensar nos MGMs como uma generalização do algoritmo *k-means*, de modo a incorporar informações sobre a estrutura de covariância dos dados, permitindo a definição de grupos com contornos não circulares (Figura 9).

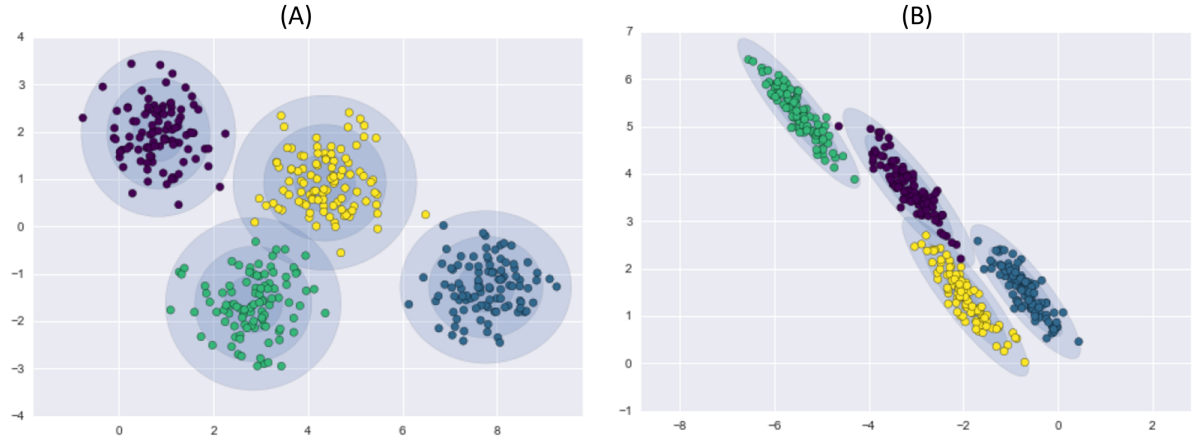

Figura 9 – Representação de agrupamento de dados usando modelos Gaussianos mistos com quatro componentes (grupos). MGMs podem ser aplicados para definir grupos com contornos circulares (A) ou não-circulares (B) (VANDERPLAS, 2016).

A chamada "Densidade *Kernel*" é um método comum de MGMs, no qual um modelo Gaussiano é ajustado para cada dado. No entanto, na presença de muitos dados, isso pode se tornar pouco prático e o algoritmo de maximização da expectativa (DEMPSTER et al., 1977) pode ser interessante de ser aplicado, uma vez que proporciona o ajuste de MGMs com menos modelos. No contexto da análise de agrupamento, o algoritmo de MGMs é aplicado para ajustar k modelos, para que sejam definidos k grupos, e seu funcionamento pode ser descrito de acordo com as seguintes etapas (VANDERPLAS, 2016):

(i) Escolha do número de componentes e forma dos grupos;

(ii) Repetir até a convergência:

a. Para cada ponto, encontrar pesos que representem as probabilidades de associação a cada grupo;

b. Para cada grupo, atualizar os parâmetros baseado nos pesos encontrados na etapa anterior.

Os resultados da aplicação de MGMs ao agrupamento de dados refletem os contornos de densidade desses dados, ou seja, a densidade de dados define os contornos das distribuições de probabilidades dos modelos, como pode ser observado na Figura 9.

2.3.1.4 Agrupamento espacial baseado em densidade de aplicações com ruído (*Density-Based Spatial Clustering of Applications with Noise – DBSCAN)*

O algoritmo denominado *DBSCAN*, introduzido por Ester et al. (1996), não requer que um número de grupos seja predeterminado. Ele o encontra automaticamente, buscando

agrupar os dados com base nas densidades das nuvens de pontos no espaço multivariado. O método compreende os grupos como áreas de alta densidade, separadas por áreas de baixa densidade.

Cabe citar que o termo 'espacial', no nome do algoritmo, não se refere ao espaço geográfico, já que este é um algoritmo do tipo tradicional.

Dado o caráter generalista da técnica, os grupos formados pelo *DBSCAN* podem assumir qualquer formato, diferentemente de alguns outros algoritmos, por exemplo o *k-means*, que tende a formar grupos de formatos convexos ("blob" é um termo informal, na língua inglesa, que se encontra na literatura para este tipo de forma).

A técnica se baseia no conceito de amostras de núcleo (em vermelho na Figura 10(A)), que se encontram em zonas de alta densidade. Assim, cada grupo é um conjunto de amostras de núcleo localizadas próximas umas das outras, circundadas por outro conjunto de amostras, denominadas amostras de borda (em amarelo na Figura 10(A)), localizadas nas proximidades de alguma amostra de núcleo, mas que não são, elas próprias, amostras de núcleo. Todas as outras amostras, que estão fora de um determinado raio de alcance (definido pelo usuário), são classificadas como ruído (em azul na Figura 10(A)).

Os parâmetros que definem cada tipo de amostra como de núcleo, de borda, ou ruído, e que calibram o conceito de densidade do modelo, são:

- Raio de alcance (Eps): é a distância em torno de cada amostra em que se faz a busca por outras amostras, ou seja, a vizinhança de busca, representada pelos círculos em torno de cada amostra na Figura 10(A);
- Número mínimo de amostras (MinPts): quantidade mínima de amostras necessárias dentro do raio de alcance para se definir aquela amostra como de núcleo.

A densidade é calibrada alterando-se esses parâmetros, sendo que quanto maior o MinPts e menor o Eps, maior é a densidade de pontos necessária para se formar um grupo. A Figura 10(B) mostra o resultado da aplicação do *DBSCAN* em um banco de dados bidimensional, tendo sido formados três grupos. As amostras representadas em preto, localizadas em zonas de baixa densidade, são aquelas classificadas como ruído.

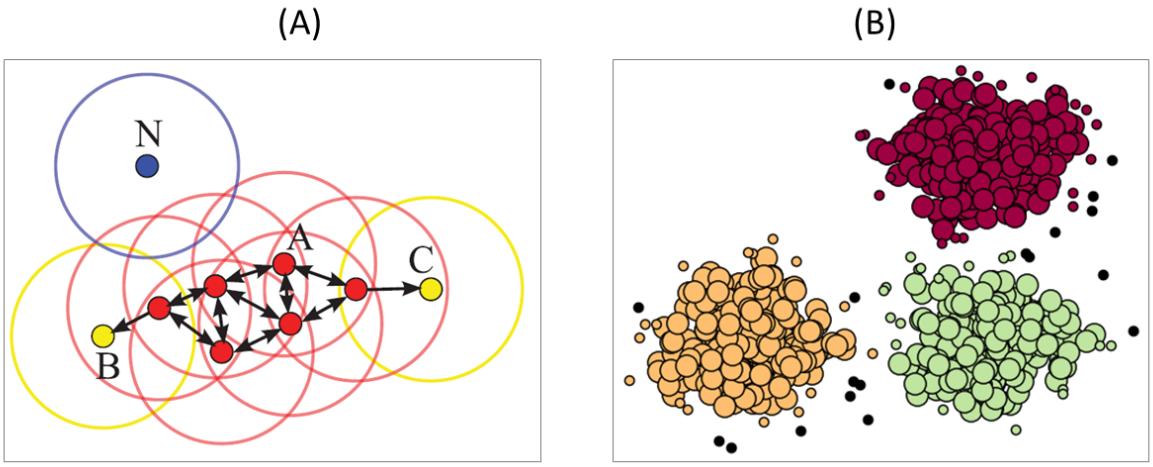

Figura 10 – Ilustração da aplicação do algoritmo de agrupamento *DBSCAN*. (A) Representação gráfica da classificação das amostras como 'de núcleo' (vermelho), 'de borda' (amarelo) e 'ruído' (azul). Os círculos em torno das amostras representam o raio de busca (Eps) (SCHUBERT et al., 2017); (B) Resultado da aplicação do *DBSCAN* em um banco de dados bidimensional, onde cada grupo é representado por uma cor diferente, sendo as amostras de núcleo como pontos maiores, as de borda como pontos menores e o ruído como pontos pretos em zonas de baixa densidade (PEDREGOSA et al., 2011).

2.3.2 Agrupamento de dados espaciais

O agrupamento espacial abrange métodos desenvolvidos para tratar não só das relações das variáveis no espaço multivariado, mas também no espaço geográfico. O agrupamento tradicional tende a produzir grupos espacialmente fragmentados, o que é indesejável na perspectiva da exploração mineral (FOUEDJIO et al., 2018). Assim, o propósito do agrupamento espacial é gerar grupos cujos elementos apresentem similaridades estatísticas e que sejam, ao mesmo tempo, geograficamente coesos, o que é mais adequado para a modelagem de fenômenos geológicos.

A primeira solução que pode vir à mente é simplesmente incorporar as coordenadas geográficas como variáveis de entrada em algoritmos tradicionais. No entanto, isso pode acabar resultando em domínios artificialmente geométricos, inadequados e sem sentido prático.

O assunto já vem sendo investigado há algumas décadas, e há distintas abordagens para se incorporar informações geográficas à análise de agrupamento. Segundo Martin & Boisvert (2018), duas estratégias gerais podem ser definidas:

(i) Alguma forma de restrição de vizinhança, de modo a modificar a relação de amostras distantes, não-correlacionadas (e.g. Oliver & Webster (1989), Ambroise et al. (1997), Romary et al. (2012), Fouedjio (2016), Martin & Boisvert (2018));

(ii) Geração de um banco de dados secundário, calculado a partir do original, com estatísticas de autocorrelação local (e.g. Scrucca (2005)).

2.3.2.1 Agrupamento espacial por restrição de vizinhança

Oliver & Webster (1989) descrevem um procedimento em que utilizam análise espacial (modelos variográficos) para determinar a escala da variação espacial, que é então aplicada na classificação e no agrupamento de pontos amostrais. A coesão dos grupos pode ser ajustada ao modificar o alcance e a forma dos modelos variográficos.

Ambroise et al. (1997) propõem uma metodologia onde restrições espaciais são aplicadas ao algoritmo de maximização da expectativa, ou seja, dentro do âmbito dos Modelos Gaussianos Mistos.

Romary et al. (2012) fazem uma revisão de alguns dos métodos tradicionais de agrupamento de dados e, assim como Fouedjio (2016), apresentam uma técnica baseada no método hierárquico, na qual consideram a dependência espacial das amostras. O algoritmo de Romary et al. (2012) foi implementando no *software Minestis*® e, posteriormente, também disponibilizado no *software Isatis.neo*®, ambos desenvolvidos pela *Geovariances*®. Um fluxo de trabalho utilizando esse algoritmo é apresentado em Mariz et al. (2016).

Martin & Boisvert (2018) apresentam uma metodologia que compreende dois aspectos: *i*) introdução de um novo algoritmo de agrupamento espacial-multivariado de caminhos aleatórios que reduz a dependência da interferência do usuário; *ii*) uma métrica espacial-multivariada combinada, que descreve a qualidade do agrupamento em espaço duplo. Essa metodologia será descrita em detalhes a seguir.

Agrupamento em espaço duplo

O método desenvolvido por Martin & Boisvert (2018) é fundamentado na combinação de amostras por vizinhança de busca em espaço geográfico, associada a medidas de distâncias em espaço multivariado para a formação dos grupos. Através de caminhos aleatórios, o algoritmo gera múltiplas realizações ("ensemble clustering", ou conjuntos de agrupamentos) e usa técnicas específicas para extrair uma configuração final, a partir do conjunto de distintas configurações de agrupamentos.

De acordo com os próprios autores, a metodologia pode ser descrita em três etapas (Figura 11):

(i) **Etapa 1: aglomeração espacial preliminar**

Na primeira etapa, uma busca por n vizinhos mais próximos gera agrupamentos baseados em distâncias no espaço geográfico. A distância Euclidiana no espaço M-dimensional é então calculada por caminhos aleatórios entre a amostra atual

e aquelas encontradas na vizinhança. As amostras com as menores distâncias no espaço M-dimensional são então aglomeradas para formar minigrupos. Esta fase se repete até que todas as amostras pertençam a um minigrupo.

(ii) **Etapa 2: aglomeração multivariada secundária**

Na etapa 2 é que é finalizada cada realização (configuração de agrupamentos), ao serem aglomerados os $k_{minigrupos}$, formados na etapa anterior, em uma quantidade-alvo de grupos (k_{alvo}). Neste ponto, não é necessário que o k_{alvo} seja o número final de grupos, já que a fase final (etapa 3) será usada para se inferir o k_{final}. Para se aglomerar os minigrupos, são consideradas relações no espaço multivariado, de acordo com métricas como a proximidade *Ward* (Ward (1963) *apud* Martin & Boisvert (2018)), a "energy distance" de Rizzo & Székely (2016), ou a divergência de Kullback-Leibler (KLD) (Hershey Olsen (2007) *apud* Martin & Boisvert (2018)).

(iii) **Etapa 3: rotulação final de grupos**

Na etapa 3, o conjunto de realizações de agrupamentos é armazenado em uma matriz C, N x L, na qual cada coluna contem rótulos para os N locais para cada realização, L. Técnicas para se lidar com conjuntos de agrupamentos ("clustering ensembles") são então usadas para extrair uma configuração final, obtida por uma função consensual (STREHL; GHOSH, 2002). A função consensual mais simples consiste no agrupamento hierárquico de uma matriz de similaridade (STREHL; GHOSH, 2002; MANITA et al., 2012).

A matriz de similaridade N x N com as amostras, denominada "matriz de proximidade", computa o número de vezes que cada local i se encontra no mesmo grupo que cada local j, {i, j = 1, ..., N}. A matriz de similaridade normalizada pelo número de agrupamentos no conjunto (L) exprime a probabilidade de que cada amostra seja combinada com cada uma das outras. O agrupamento hierárquico da matriz de similaridade com medidas de proximidade *Ward* é então usada para se obter a configuração final.

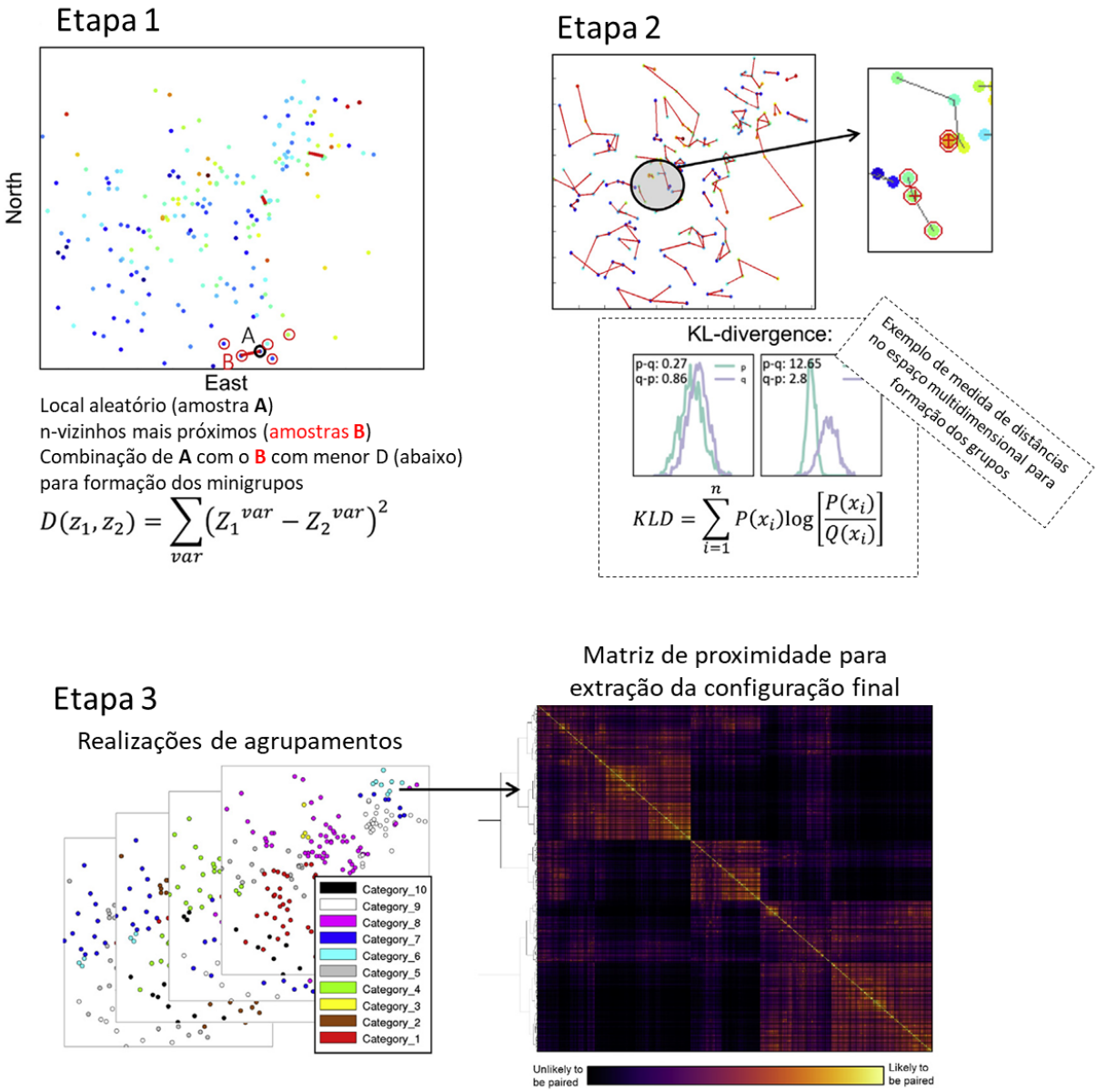

Figura 11 – Descrição conceitual do funcionamento do algoritmo para agrupamento em espaço duplo. Na etapa 1, os minigrupos são formados pelos pares com menor distância no espaço multivariado, dentre os n vizinhos mais próximos; na etapa 2, os minigrupos são combinados para formar os agrupamentos-alvo e; na etapa 3, a matriz de proximidade é construída para que seja extraída a configuração de agrupamentos consensual final (adaptado de Martin & Boisvert (2018)).

2.3.2.2 Agrupamento espacial por estatísticas de autocorrelação

Scrucca (2005), com base em conceitos introduzidos por Getis & Ord (1992) e Ord & Getis (1995) para casos univariados, aplica estatísticas de autocorrelação local para gerar grupos multivariados interconectados espacialmente, método que se dá, basicamente, em duas etapas:

(i) Um banco de dados é gerado com medidas de autocorrelações locais para cada variável, dadas as relações diretas com dados nas vizinhanças;

(ii) O algoritmo tradicional *k-means* é aplicado a esse novo banco de dados, para que sejam definidos grupos que apresentem coerência tanto espacial quanto estatística.

A etapa (i) ocorre de modo que, dada uma matriz binária W, relacionada à localização das amostras, são computadas as estatísticas padronizadas de Getis-Ord para cada variável do sistema, através da seguinte equação:

$$z(G_i) = \frac{\sum_{j=1}^{n}(w_{ij}x_j - \overline{x}w_i)}{\sqrt{\frac{s^2}{n-1}(n(\sum_{j=1}^{n} w_{ij}^2) - w_i^2)}} \tag{2.5}$$

onde $z(G_i)$ é a estatística de Getis-Ord no local i para a variável x; $w_{i,j}$ é uma medida de continuidade espacial entre os locais i e j, sendo o (i,j)-ésimo elemento da matriz binária simétrica, W, com os pesos relativos à distribuição espacial das amostras, sendo $w_{i,j} = 1$ para os locais nas vizinhanças (à uma distância d) e 0 para os demais; $w_i = \sum_j w_{ij}$; x_j é o valor da amostra no local j; n é o número de amostras e $s^2 = \sum_i (x_i - \overline{x})^2/n$.

Valores positivos de $z(G_i)$ indicam grupos com valores altos na vizinhança do local i, enquanto valores negativos indicam a existência de valores baixos nos arredores de i (SCRUCCA, 2005).

2.3.3 Sobre a parametrização e as escolhas do número de grupos e do algoritmo

Comum a todos os métodos, espaciais ou não, é a questão da parametrização. A métrica para se medir distâncias entre amostras e grupos, o método para integrar a correlação espacial, a maneira como a vizinhança espacial correlacionada é determinada e a distância máxima de correlação são exemplos de parâmetros que devem ser determinados *a priori* por um usuário experiente, ciente das implicações de cada uma dessas escolhas.

A própria definição do número de grupos (k) não é trivial, e deve ser feita com cuidado, para que não haja mistura de populações estatísticas ou que não seja definido um número excessivo de domínios, o que pode complicar desnecessariamente as etapas subsequentes de modelagem e estimativa/simulação.

A maioria das técnicas de agrupamento requer que o k seja definido *a priori*, e essa é uma decisão subjetiva, auxiliada por ferramentas específicas. A maneira mais simples e intuitiva de fazê-la é pela simples observação do dendrograma, resultante da aplicação do método aglomerativo hierárquico, e cuja construção não exige a definição prévia de k. Outra técnica útil é o cálculo da soma dos quadrados das distâncias intragrupo (inércia), e a plotagem de um gráfico da inércia *versus* k, cujo ponto de inflexão pode indicar o número mais adequado de grupos.

Ferramentas complementares para a definição de k também podem ser utilizadas na validação dos resultados do agrupamento, como o método da silhueta e os índices de Calinski-Harabasz e de Davies-Bouldin, que serão descritos em detalhes mais adiante.

A própria escolha do método de agrupamento a ser aplicado pode ser um tanto complexa, e deve ser adequada ao conjunto de dados e ao propósito do estudo. Conforme já delineado por Kaufman & Rousseeuw (2005), algumas vezes algoritmos distintos podem ser aplicados à mesma situação, sendo importante que se façam comparações cuidadosas dos resultados, inclusive por meio de análise gráfica (e.g. mapas de localização, gráficos de dispersão, dendrogramas).

Ainda de acordo com Kaufman & Rousseeuw (2005), é interessante aplicar diversos algoritmos ao mesmo banco de dados, uma vez que o agrupamento de dados é, na maior parte das vezes, usado como uma ferramenta descritiva, exploratória, em contraste com outros testes estatísticos, que são realizados com o propósito de inferir ou confirmar resultados. Ou seja, a finalidade não é provar (ou refutar) hipóteses preconcebidas, e sim investigar o que os dados têm a mostrar.

Assim, é importante ressaltar que os resultados raramente são definitivos, sendo a análise de agrupamento, em realidade, um tanto quanto subjetiva. A escolha do algoritmo, do número de grupos e a parametrização devem ser baseadas na compreensão dos dados originais e na interpretação dos resultados, baseadas no conhecimento do usuário.

2.4 Avaliação dos resultados de agrupamentos

Avaliar a qualidade dos resultados é uma das considerações mais importantes a se fazer quando se deseja implementar algoritmos de aprendizado de máquina. Para casos de aprendizado supervisionado, essa avaliação é bem direta, já que existem rótulos para que o desempenho do algoritmo seja medido.

Em casos não supervisionados, como a análise de agrupamento, essa avaliação é um tanto problemática, já que os dados não estão rotulados e, logo, não há um gabarito para referência.

Há diversas técnicas disponíveis na literatura que permitem avaliar a qualidade dos resultados. São métricas que podem fornecer uma perspectiva sobre a eficácia do processo e que permitem observar a tendência natural que os dados têm de se agrupar.

Entretanto, essas validações não devem ser dadas como verdades absolutas, já que também apresentam considerável subjetividade. Na verdade, a ideia é usar essas métricas para testar comparativamente os resultados, ao serem alterados os parâmetros, o número de grupos e o algoritmo aplicado. Em outras palavras, o que se testa não é a veracidade dos agrupamentos, mas sim sua qualidade relativa, seu sentido prático, isto é, se foi possível

agrupar os dados de maneira satisfatória.

2.4.1 Método das silhuetas

Proposto por Rousseeuw (1987), trata-se de um método gráfico, que representa o quão bem agrupados estão os objetos dentro de seus respectivos grupos. O coeficiente de silhueta (s) é calculado para cada amostra, usando a distância média entre ela e as outras amostras do mesmo grupo, bem como a distância média entre ela e amostras do grupo mais próximo (Fig. 12):

$$s = \frac{(b - a)}{max(a, b)} \tag{2.6}$$

onde b é a distância média entre a amostra em questão e as demais do mesmo grupo e a, a distância média entre aquela mesma amostra e cada uma das amostras do grupo mais próximo.

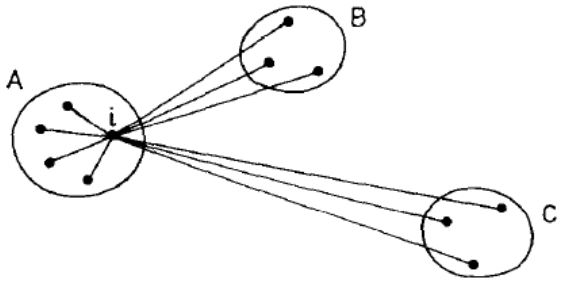

Figura 12 – Ilustração dos elementos envolvidos no cálculo do coeficiente de silhueta (s) do objeto i, pertencente ao grupo A, sendo B o grupo mais próximo e C, um grupo qualquer, mais distante que B. Assim, para o cálculo de $s(i)$ são computadas as distâncias entre i e as demais amostras pertencentes ao grupo A e as distâncias entre i e as amostras pertencentes ao grupo B (ROUSSEEUW, 1987).

Após o cálculo de s para todos os pontos, os valores obtidos são plotados em um gráfico onde a silhueta referente a cada grupo mostra os s de cada objeto pertencente ao respectivo grupo, em ordem decrescente, como mostra o exemplo da Figura 13. A largura da silhueta (eixo horizontal) corresponde ao quão bem aquele determinado objeto foi atribuído àquele grupo, de modo que silhuetas mais largas correspondem a valores mais altos de s. A outra dimensão da silhueta é a altura, que equivale ao número de objetos designados àquele grupo. Para tornar a análise mais prática, podem-se calcular as médias dos s totais para cada configuração, e aquela com o maior valor é a configuração mais adequada.

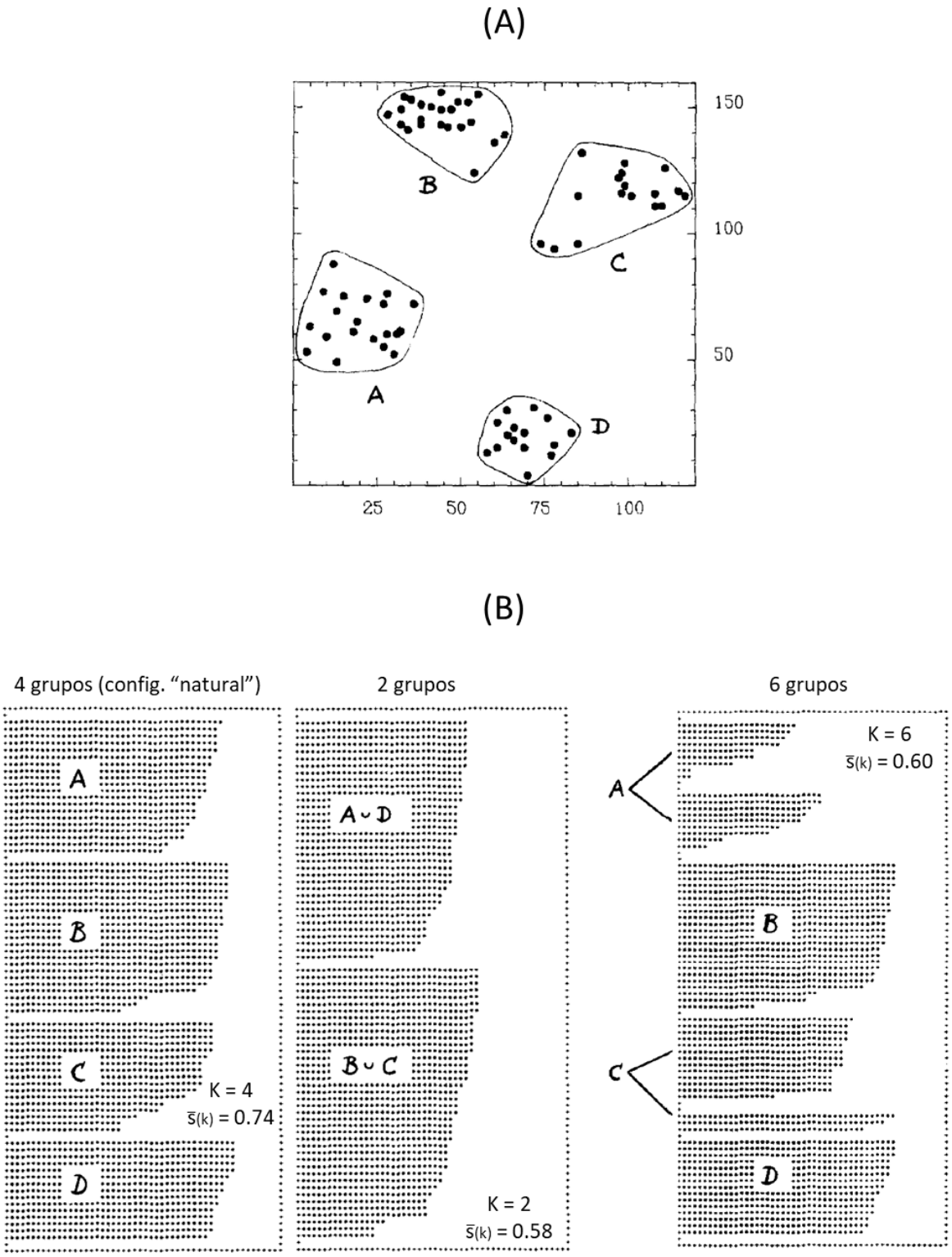

Figura 13 – Exemplo adaptado de Rousseeuw (1987) de gráficos de silhuetas. (A) Relação dos dados no espaço (bivariado), com a delineação natural dos grupos A, B, C e D; (B) Gráficos de silhuetas para quatro grupos (esquerda), para dois grupos (centro), e para seis grupos (direita). Percebe-se que a configuração natural, de quatro grupos, é aquela que apresenta o maior valor médio do índice s, ao passo que o agrupamento dos dados em dois grandes grupos, bem como sua subdivisão em seis subgrupos leva a valores mais baixos do s médio.

O índice s varia dentro do intervalo $[-1, +1]$, sendo que, quanto maior, melhor. Valores próximos a 0 indicam possíveis sobreposições de grupos. Valores negativos podem indicar que uma amostra foi indevidamente atribuída ao respectivo grupo.

Este método é mais adequado quando as distâncias estão em escala de razão (e.g. distâncias Euclidianas), e quando se buscam grupos compactos (ROUSSEEUW, 1987), que apresentam contornos convexos, ou seja, formatos mais circulares/esféricos.

2.4.2 Índice Davies-Bouldin

Apresentado por Davies & Bouldin (1979), este método envolve o cálculo da similaridade entre grupos. Valores mais baixos indicam configurações mais adequadas.

O índice Davies-Bouldin é definido como a similaridade média entre cada grupo C_i (para $i = 1, 2, ..., k$) e o grupo mais próximo, C_j. Essa similaridade é medida da seguinte maneira:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{2.7}$$

onde s_i é a distância média (Euclidiana) entre cada ponto do grupo C_i e seu respectivo centroide, s_j é a distância de cada ponto do grupo C_j e seu centroide, e d_{ij} é a distância entre os centroides dos grupos C_i e C_j (Fig. 14).

Assim, o índice Davies-Bouldin é definido como:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max R_{ij} \tag{2.8}$$

sendo $i \neq j$.

Zero é o menor valor possível, sendo que valores próximos a zero indicam agrupamentos melhores.

Assim como o método das silhuetas, o índice Davies-Bouldin se aplica melhor a agrupamentos que apresentam alguma convexidade. Além disso, a técnica também se limita ao cálculo de distâncias Euclidianas no espaço multivariado.

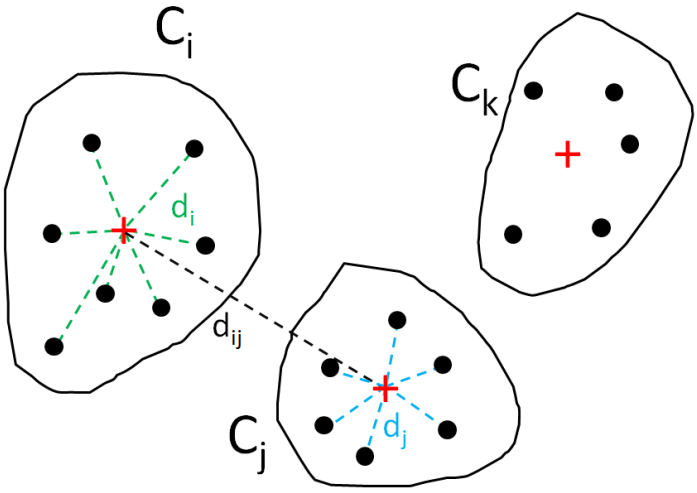

Figura 14 – Demonstração gráfica dos elementos envolvidos no cálculo do índice Davies-Bouldin para o grupo C_i, do qual C_j é o grupo mais próximo. C_k representa um grupo qualquer, mais distante. d_i representa as distâncias internas do grupo C_i, d_j, as distâncias internas do grupo C_j e, d_{ij}, a distância entre os centroides de C_i e C_j.

2.4.3 Índice Calinski-Harabasz

Também conhecido como 'critério da razão de variâncias'. Foi introduzido por Caliński & Harabasz (1974), sendo, também, um método aplicável a um cenário multidimensional no espaço Euclidiano, baseado no conceito da mínima soma de quadrados.

Segundo Pedregosa et al. (2011), para um conjunto de dados E de tamanho n_E, dividido em k grupos, o índice Calinski-Harabasz s é definido como a razão entre a média de dispersão entre grupos e a dispersão interna dos grupos:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{2.9}$$

onde $tr(B_k)$ representa o traço da matriz de dispersão entre grupos e, $tr(W_k)$, o traço da matriz de dispersão interna dos grupos, sendo:

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{2.10}$$

$$B_k = \sum_{q} n_q (c_q - c_E)(c_q - c_E)^T \tag{2.11}$$

onde C_q é o conjunto de pontos no grupo q, c_q o centroide do grupo q, c_E o centro de E e n_q o número de pontos do grupo q. T indica transposição.

Quanto melhor definidos os grupos, mais alto é o índice Calinski-Harabasz e, assim como no método das silhuetas e no índice Davies-Bouldin, a técnica é mais adequada para agrupamentos convexos.

2.4.4 Avaliação em espaço duplo

Martin & Boisvert (2018) propuseram uma abordagem alternativa para medir a qualidade do agrupamento espacial, visando auxiliar no fluxo de trabalho geoestatístico para decisões de estacionariedade. Desta maneira, dois critérios são utilizados na medida da qualidade do agrupamento: *i*) a continuidade espacial dos domínios no espaço geográfico e, *ii*) a subdivisão da população no espaço multivariado em subpopulações coesas.

O método é baseado nos seguintes conceitos: *i*) a soma dos quadrados intragrupo ("within cluster sum of squares" – *wcss*), que mede a coesão dos dados de cada grupo no espaço multivariado e, *ii*) a entropia espacial (H), que mede a interconectividade dos dados/grupos no espaço geográfico. Cada métrica é calculada independentemente, mas ambas são avaliadas simultaneamente:

$$wcss = \sum_{k=1}^{k} \sum_{x_i \in K_k} \sum_{j=1}^{M} (x_{ij} - \overline{x}_{kj})^2 \tag{2.12}$$

onde $(x_{ij} - \overline{x}_{kj})$ representa a distância entre uma determinada amostra e o centroide da distribuição multivariada de seu respectivo grupo.

$$H_{total} = -\sum_{i=1}^{N} \sum_{k=1}^{K} p_{i,k}\ ln\ p_{i,k} \tag{2.13}$$

onde $p_{i,k}$ é a probabilidade de se encontrar outra amostra da categoria k nos arredores do i-ésimo local, no espaço geográfico.

Configurações com grupos mais compactos no espaço multivariado apresentam valores mais baixos de *wcss* (Fig.15), já que as distâncias entre os elementos dentro de cada grupo são menores. Maior coesão geográfica dos elementos de cada grupo, ou seja, maior interconectividade espacial dos grupos, implica em valores mais baixos para H (Fig. 16).

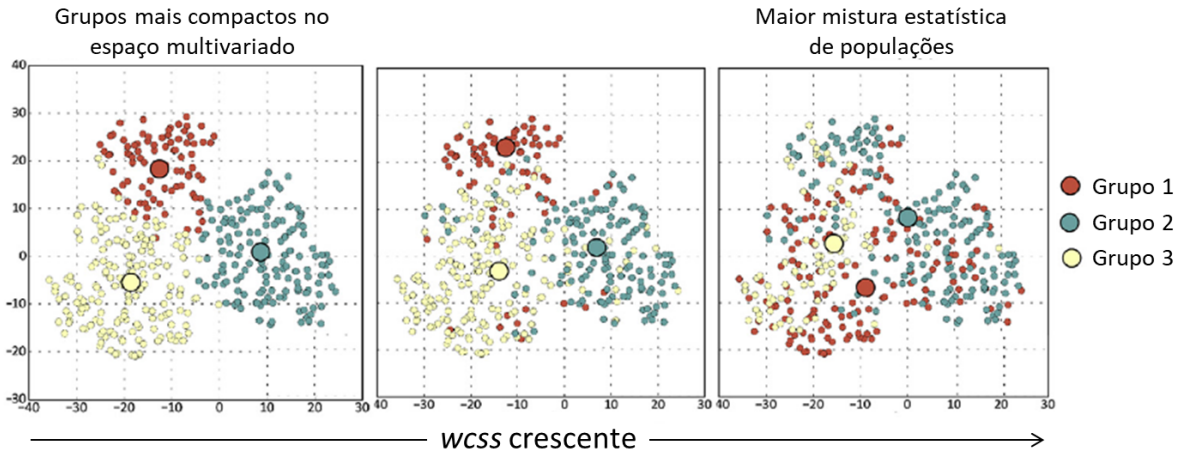

Figura 15 – Diferentes configurações de agrupamento no espaço multivariado e suas relações com a soma dos quadrados das distâncias intragrupo (*wcss*). Quanto mais coesos os grupos, menor é o *wcss* (adaptado de Martin & Boisvert (2018)).

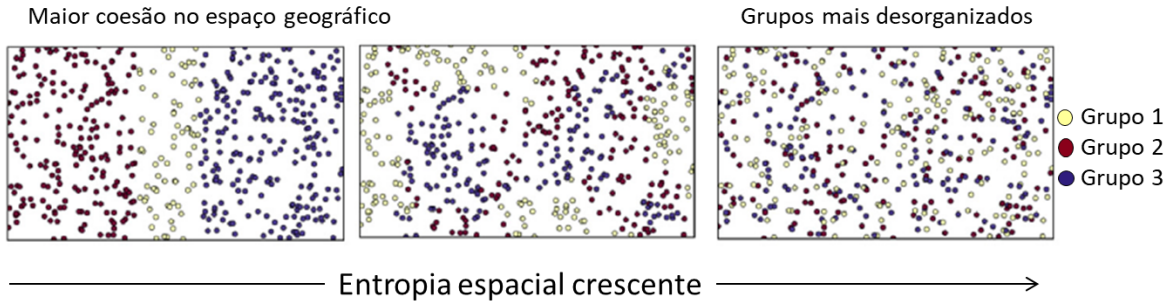

Figura 16 – Diferentes configurações de agrupamento no espaço geográfico e suas relações com a entropia espacial (H). Quanto maior a conectividade espacial dos grupos, menor é H (adaptado de Martin & Boisvert (2018)).

.

Assim, para que uma determinada configuração seja considerada satisfatória, são desejáveis valores baixos tanto de *wcss* quanto de H. No entanto, como já havia sido notado por Oliver & Webster (1989) e foi posteriormente confirmado empiricamente por Martin & Boisvert (2018), essas duas métricas são inversamente proporcionais. Ou seja, maior coesão no espaço multivariado implica em fragmentação geográfica dos agrupamentos (Fig. 17). A solução é combinar essas métricas e avaliar os resultados comparativamente, de maneira qualitativa, entre diferentes configurações. Geralmente, a melhor configuração para o agrupamento de dados espaciais não é aquela que apresenta os valores mais baixos de *wcss* ou H, mas valores intermediários.

Vale observar que os métodos tradicionais de análise de agrupamento (e.g. aglomerativo hierárquico, k-means) fornecem resultados com alta coesão no espaço multivariado (baixo valor de *wcss*), mas com pobre coerência espacial (alto H), principalmente em casos mais complexos.

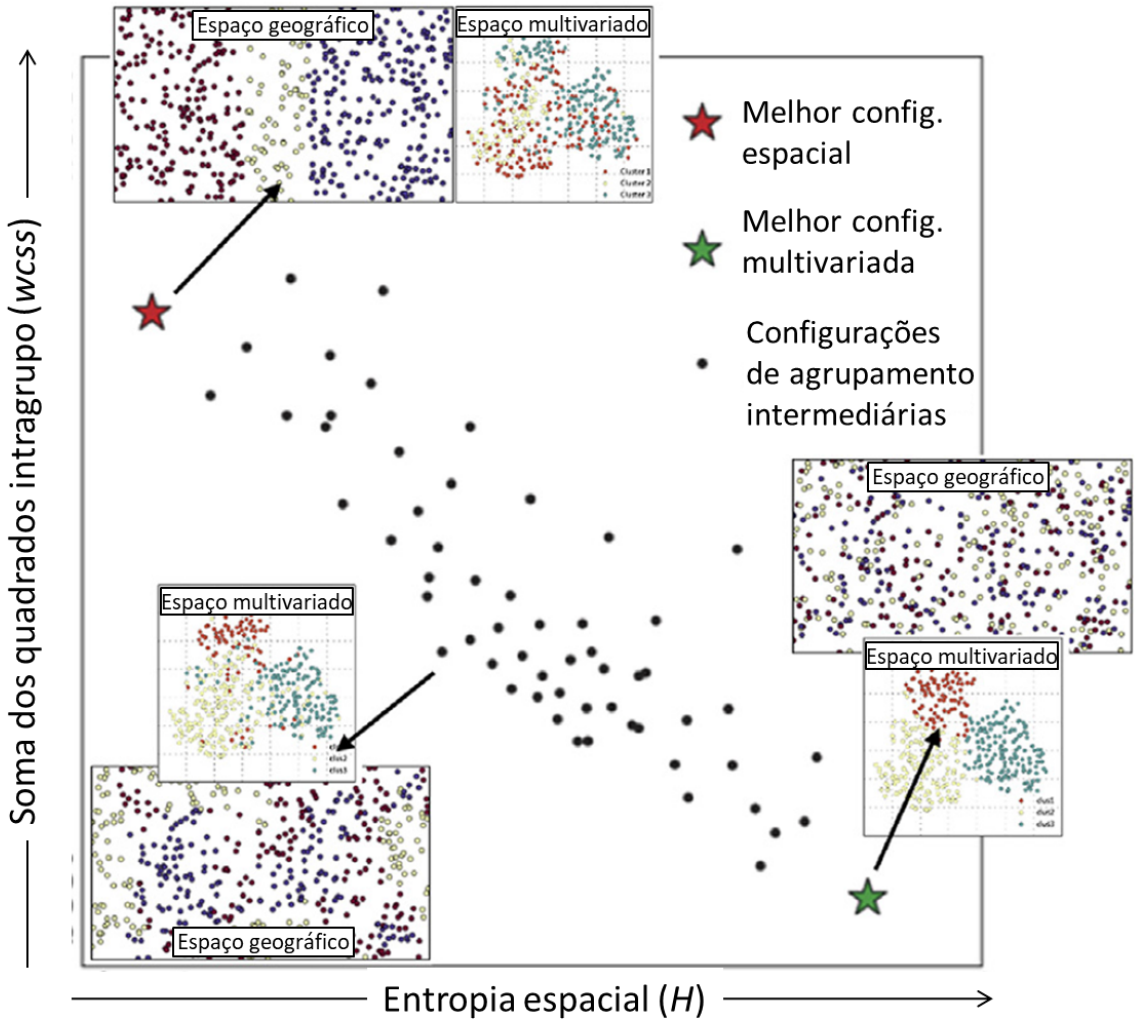

Figura 17 – Valores de *wcss* e H para diferentes configurações de agrupamentos, plotados em um gráfico de dispersão, evidenciando sua relação inversa. Quanto maior a coesão dos grupos no espaço multivariado (menor *wcss*), mais desorganizados eles são no espaço geográfico (maior H). Nos detalhes podem ser observadas as configurações no espaço multivariado e no espaço geográfico de algumas das configurações de agrupamentos (ver Figs. 15 e 16) (adaptado de Martin & Boisvert (2018)).

.

2.4.5 Abordagem dos indicadores na validação dos agrupamentos

Como uma das principais contribuições deste trabalho, e na mesma linha de Modena et al. (2019), introduz-se a aplicação das medidas de continuidade espacial de indicadores para validação das configurações dos agrupamentos no espaço geográfico.

Um indicador é uma variável binária que, quando aplicada a dados categóricos, assume valor 1 ou 0, a depender se ela pertence, ou não, a uma determinada categoria. A Figura 18 mostra um exemplo de definição de indicadores no banco de dados Jura, de Goovaerts (1997).

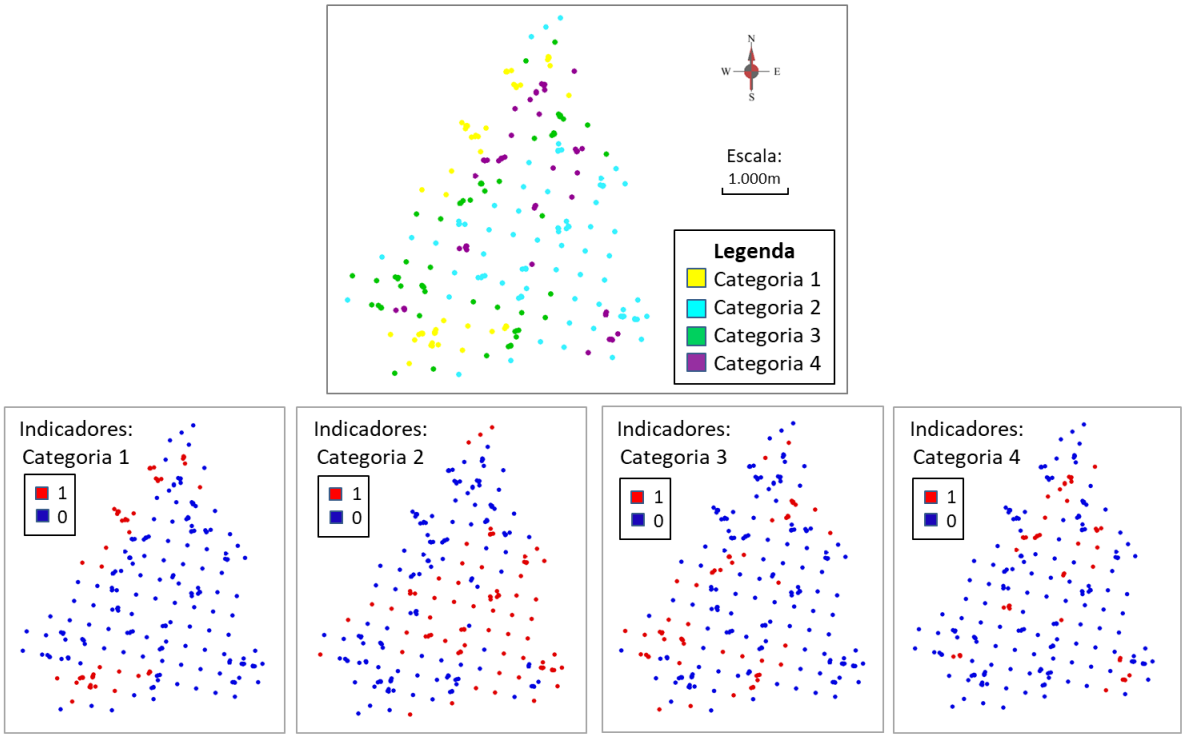

Figura 18 – Exemplo de definição de indicadores para uma variável categórica do banco de dados Jura, de Goovaerts (1997)

.

O padrão de distribuição das amostras de uma dada categoria pode ser caracterizado pela análise da continuidade espacial dos indicadores definidos por aquela categoria. A ferramenta mais amplamente utilizada para se medir a continuidade espacial é o semivariograma, mas também podem ser utilizados covariogramas e correlogramas.

O semivariograma, definido por Matheron (1963), é uma função que caracteriza o grau de continuidade espacial de um fenômeno. No eixo horizontal do gráfico plotam-se distâncias, enquanto no eixo vertical, as meias-médias das diferenças quadráticas dos valores separados por um vetor h, que representam o grau de descorrelação entre dois pontos separados no espaço. Em se tratando de indicadores, a equação do semivariograma

pode ser representada da seguinte maneira (GOOVAERTS, 1997):

$$\gamma_{I(h;s_k)} = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} (i(u_\alpha; s_k) - i(u_\alpha + h; s_k))^2 \tag{2.14}$$

onde s_k representa a categoria sendo variografada, N é o número de pares separados pelo vetor h, $i(u_\alpha; s_k)$ é o valor do indicador no local u_α, e $i(u_\alpha + h; s_k)$, o valor do indicador no local $(u_\alpha + h)$.

Assim, o semivariograma dos indicadores mede a frequência com a qual dois locais, distantes em um vetor h, pertencem a categorias diferentes. Quanto menor $\gamma_I(h; s_k)$, maior é a conectividade espacial da categoria s_k, sendo que os alcances e as formas dos semivariogramas refletem os padrões geométricos da categoria s_k (GOOVAERTS, 1997).

De acordo com Matheron (1963), podem-se distinguir, basicamente, quatro tipos de semivariogramas (Figura 19):

(i) Contínuo: parabólico na origem, indica fenômenos com alta continuidade, como a espessura de estratos sedimentares;

(ii) Linear: tangente e oblíquo à origem, representa uma variável com continuidade média. É o tipo mais comum para teores de depósitos metálicos;

(iii) Efeito pepita: revela uma descontinuidade na origem, que representa certo grau de erraticidade do fenômento a curtas distâncias;

(iv) Errático: caso extremo que corresponde ao conceito de variável aleatória. Também referido na literatura como efeito pepita puro.

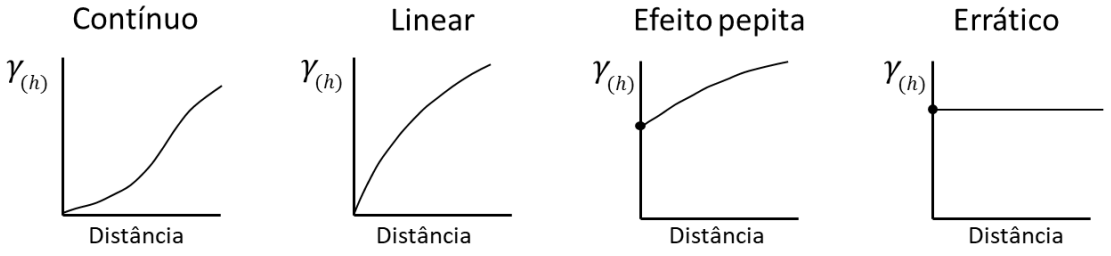

Figura 19 – Quatro tipos comuns de semivariogramas de fenômenos naturais (adaptado de Matheron (1963)).

.

Uma maneira alternativa de se medir padrões de conectividade espacial é através do covariograma, que expressa as covariâncias das amostras separadas a diferentes distâncias. Diferentemente do semivariograma, o covariograma mede a similitude entre amostras, e tende a decrescer à medida que aumentam-se as distâncias, apresentando aspecto invertido

ao semivariograma, quando plotado graficamente. Aplicada a indicadores, a equação do covariograma pode ser expressa da seguinte forma:

$$C_{I(h;s_k)} = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} (i(u_\alpha; s_k) \cdot i(u_\alpha + h; s_k) - m_{-h} \cdot m_{+h}) \tag{2.15}$$

onde m_{-h} e m_{+h} representam as médias dos valores localizados no ponto inicial (*"head"*) e no ponto à distância h do ponto inicial (*"tail"*), respectivamente, sendo:

$$m_{-h} = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} i(u_\alpha; s_k) \tag{2.16}$$

$$m_{+h} = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} i(u_\alpha + h; s_k) \tag{2.17}$$

Uma terceira forma de se expressar a continuidade espacial é pelo correlograma, a forma estandardizada do covariograma:

$$\rho_{I(h;s_k)} = \frac{C_{I(h;s_k)}}{\sqrt{\sigma^2_{-h} \cdot \sigma^2_{+h}}} \in [-1, +1] \tag{2.18}$$

sendo

$$\sigma^2_{-h} = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} [i(u_\alpha; s_k) - m_{-h}]^2 \tag{2.19}$$

$$\sigma^2_{+h} = \frac{1}{N(h)} \sum_{\alpha=1}^{N(h)} [i(u_\alpha + h; s_k) - m_{+h}]^2 \tag{2.20}$$

onde σ^2_{-h} e σ^2_{+h} são as variâncias dos valores localizados no ponto inicial e no ponto à distância h do ponto inicial, respectivamente.

Assim, por ser estandardizado pelas variâncias, o correlograma está restrito ao intervalo [-1, +1], sendo mais estável que o semivariograma e o covariograma, o que pode minimizar ruídos a curtas distâncias.

Na validação de configurações dos agrupamentos, as medidas de continuidade espacial dos indicadores podem ser utilizadas de modo que cada categoria represente um determinado grupo de amostras. Semivariogramas/covariogramas/correlogramas contínuos, bem estruturados, e com efeito pepita baixo caracterizam grupos com alta conectividade no espaço geográfico, enquanto semivariogramas/covariogramas/correlogramas ruidosos, com efeito pepita alto, indicam grupos espacialmente fragmentados, ou com baixa representatividade amostral.

2.5 Sobre a aplicação do aprendizado supervisionado na classificação automática de novas amostras

A mineração é uma atividade dinâmica, geralmente com um processo contínuo de amostragem, no qual novas informações são frequentemente adicionadas ao banco de dados. Realizar todo o procedimento de análise de agrupamento cada vez que novas amostras são incorporadas ao banco de dados seria pouco prático. Além disso, como as técnicas de agrupamento são baseadas na busca por relações complexas nos espaços multivariado e geográfico, outras configurações poderiam surgir, ligeiramente distintas daquelas já definidas, com algumas das antigas amostras sendo incluídas em grupos diferentes daqueles à que haviam sido previamente designadas.

Ao mesmo tempo, é essencial que as novas amostras sejam incorporadas ao banco de dados segundo os mesmos critérios usados para definir os grupos originais. Em outras palavras, que sejam classificadas segundo as mesmas regras, de modo que uma nova amostra, ao ser designada a um determinado grupo, seja mais parecida com as demais amostras do mesmo grupo do que com amostras de outros grupos.

Como essas designações não seguem regras simples, e levam em conta relações complexas no espaço geográfico e multivariado, a tarefa de classificar as novas amostras não é trivial, sendo interessante o uso de técnicas de aprendizado supervisionado de máquina.

Dessa maneira, o conjunto original de dados (agrupado) pode ser usado para calibrar um modelo matemático para a classificação de novas amostras por meio de um classificador supervisionado (e.g. árvore de decisão, florestas aleatórias, k-vizinhos mais próximos) (ver Figura 1).

Esporadicamente, à medida que o banco de dados cresce, a definição de grupos pode ser atualizada, conforme indicado pela linha tracejada no fluxograma da Figura 1, utilizando amostras que não haviam sido usadas anteriormente na análise de agrupamento. O classificador supervisionado deve, então, ser também atualizado, de modo que incorpore as novas informações para ser utilizado na classificação de novas amostras, em um processo contínuo.

2.5.1 O aprendizado supervisionado

O aprendizado de máquina (AM), termo introduzido por Samuel (1959), é um subcampo da ciência da computação que capacita computadores a aprender com os dados e executar tarefas, fornecendo resultados sem instruções explícitas. A ideia é, basicamente, permitir que um computador identifique padrões nos dados e faça previsões ou tome decisões a partir deles.

Como já brevemente mencionado na Seção 2.3, as tarefas de AM, em uma definição

simplificada, podem ser classificadas como aprendizado não supervisionado ou aprendizado supervisionado. Há também o chamado aprendizado por reforço, mas este não é abordado neste trabalho. O aprendizado não supervisionado busca por padrões nos dados sem recorrer a rotulações prévias, como já amplamente apresentado na Seção 2.3.

No aprendizado supervisionado, o algoritmo constrói um modelo matemático a partir de um conjunto de dados que contenha as entradas ("inputs") e as saídas ("outputs") desejadas. Se a propriedade de saída for contínua, o processo é definido como regressão, se for categórica (e.g. classes, taxonomia), classificação. Por não se tratar do tema principal deste trabalho, o aprendizado supervisionado, especificamente a classificação, é aqui apenas brevemente apresentado.

De maneira geral, esses modelos matemáticos podem ser obtidos por diversos algoritmos distintos (e.g. árvore de decisão, florestas aleatórias, k-vizinhos mais próximos), cujos parâmetros devem ser calibrados usando-se um subconjunto dos dados iniciais, denominados dados de treinamento ("training dataset"). O modelo é então validado observando-se os resultados de sua aplicação aos dados restantes, denominados dados de teste ("test dataset").

Uma maneira bastante eficiente de se calibrar os parâmetros de um modelo é através da chamada validação cruzada por *k-folds*, uma estratégia que divide o banco de dados repetidas vezes (k vezes, ou "folds") em dados de treino e de teste, utilizando métricas para avaliar os resultados de cada *fold*. Como os conjuntos de dados treino-teste são distintos em cada *fold*, a técnica permite avaliar o modelo em cenários distintos, minimizando as chances de favorecer uma determinada classe.

De maneira simplificada, essa validação pode ser feita através da chamada "matriz de confusão" e de algumas métricas globais de avaliação. A matriz de confusão (Figura 20) mostra o número de vezes que cada previsão foi feita por classe (colunas da matriz), pelo número de vezes que cada uma daquelas classes de fato ocorre nos dados (linhas da matriz). É desejável que a diagonal principal mostre números mais altos, o que atesta que valores preditos coincidem com valores reais.

		Classe predita				
		1	2	3	4	5
Classe real	1	1181	3	2	1	0
	2	3	723	0	1	2
	3	1	0	532	6	0
	4	2	0	4	873	0
	5	1	2	0	0	1293

Figura 20 – Exemplo de matriz de confusão para validação de classificadores supervisionados. As colunas representam os valores preditos pelo modelo, as linhas, os valores reais. É desejável que a diagonal principal apresente valores relativamente altos.

Além das matrizes de confusão, há diversas outras maneiras de se avaliar um modelo. No âmbito desta Dissertação, serão apresentadas as seguintes métricas globais:

(i) "Recall": razão entre o número de vezes que a previsão daquela classe está correta e o número real de amostras daquela classe:

$$Recall = \frac{Previsões\ corretas\ por\ classe}{Número\ real\ de\ amostras\ daquela\ classe}$$

(ii) Precisão: razão entre o número de vezes que a previsão daquela classe está correta e o número total de vezes que aquela classe foi predita (corretamente ou incorretamente)

$$Precisão = \frac{Previsões\ corretas\ por\ classe}{Total\ de\ previsões\ naquela\ classe}$$

(iii) "Score F1": média harmônica entre precisão e acurácia. É uma maneira de resumir precisão e *recall* em um único número:

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall}$$

(iv) Acurácia: razão entre o número de vezes que a previsão está correta e o número total de previsões.

3 Estudo de Caso

Este capítulo apresenta um estudo de caso com a aplicação de algumas das técnicas mencionadas nos capítulos anteriores aos dados de um depósito de fosfato e titânio localizado na região sudeste do Brasil.

Primeiramente, é feita a apresentação da área de estudo, com sua contextualização geológica; em seguida, são apresentados os dados e as técnicas aplicadas no seu tratamento e validação; depois, a análise exploratória dos mesmos, pós-tratamento. O fluxo de trabalho para aplicação das técnicas de agrupamento e validação é, então, exposto e os resultados, apresentados e discutidos. Por fim, são expostos e discutidos os resultados da calibração de um classificador por florestas aleatórias para a inclusão de novas amostras.

3.1 Contextualização geológica

A área de estudo está inserida no segmento meridional da Faixa Brasília, um cinturão de dobras e empurrões formado durante o ciclo Brasiliano (790 - 600 Ma) pela inversão tectônica de sequências sedimentares depositadas entre 900 e 800 Ma (FUCK et al., 1993) em uma bacia do tipo *rift* na margem oeste do Cráton São Francisco, conforme mencionado por Brod (1999).

Segundo Silva (2003), nessa porção da Faixa Brasília a deformação e o metamorfismo relacionados à orogênese neoproterozóica foram intensas, o que contribuiu para a obliteração das relações estratigráficas entre as várias unidades regionais. Assim, falhamentos de diferentes estilos e idades justapõem rochas de várias origens e padrões metamórficos.

No contexto da Faixa Brasília, as unidades litológicas da região correspondem a granito-gnaisses e rochas metassedimentares/metavulcânicas do Grupo Araxá; xistos carbonosos, filitos e quartzitos do Grupo Canastra e filitos esverdeados do Grupo Ibiá (Figura 21).

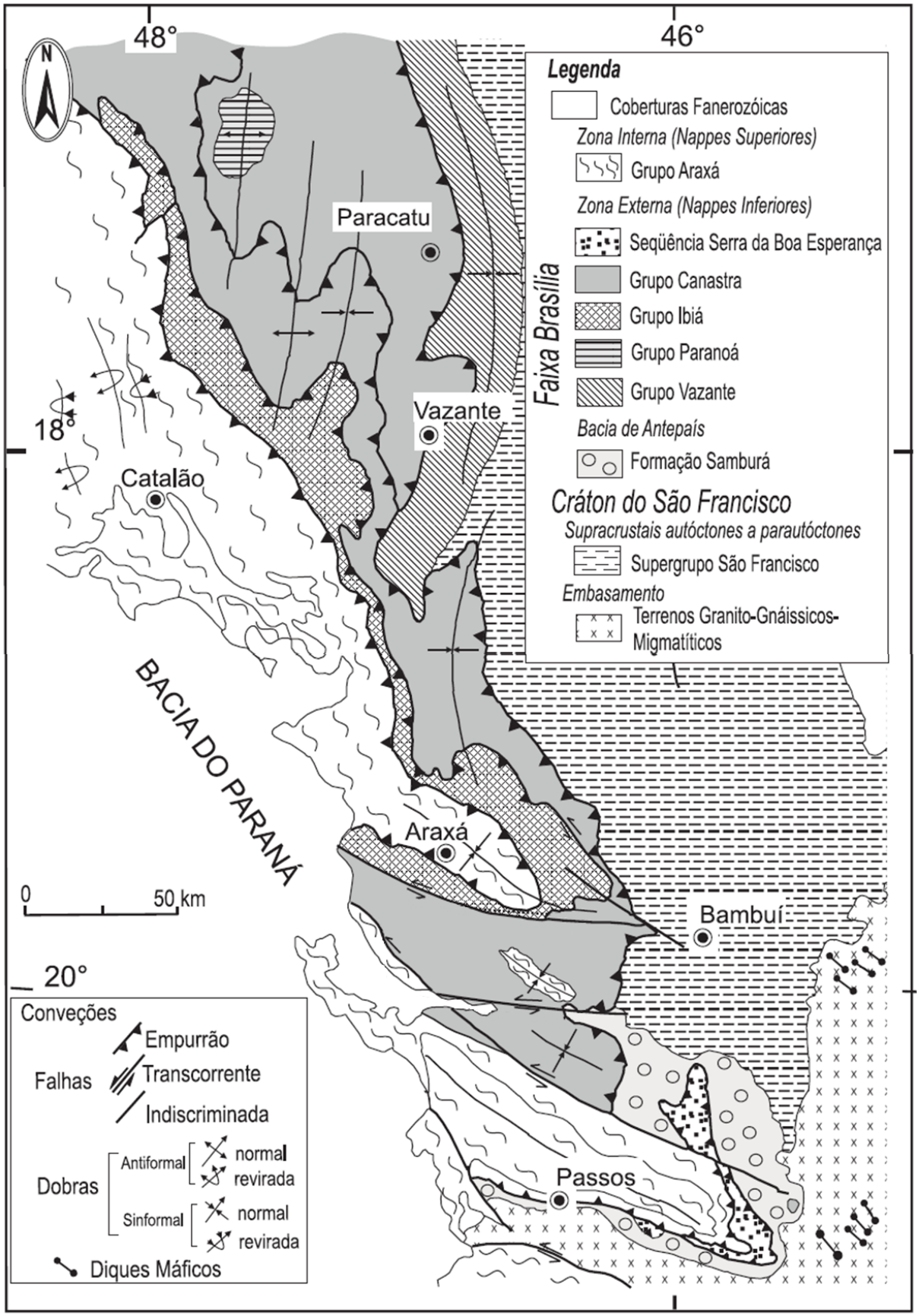

Figura 21 – Esboço geológico da porção sul da Faixa Brasília (SILVA et al., 2006)

Volumosos episódios de magmatismo ocorreram no que hoje são as porções central e sul do Brasil, entre o Cretáceo Inferior e o Eoceno. Esses eventos originaram as províncias alcalinas localizadas às margens da Bacia do Paraná, encaixadas em rochas da Faixa Brasília, bem como os extensivos derrames continentais de basaltos que cobrem grande parte da mesma bacia. Segundo Brod (1999) e autores por ele referenciados, em ambos os casos o magmatismo está frequentemente associado à influência térmica e/ou química de plumas mantélicas que impactaram a base da litosfera continental.

A Província Ígnea do Alto Paranaíba (PIAP), onde se situa o depósito relacionado a este estudo de caso, se refere a um conjunto de intrusões máficas a ultramáficas alcalinas ultrapotássicas, colocadas durante o Cretáceo Superior. O trabalho de Gibson (1995 *apud* Brod (1999)) demonstrou que os tipos de magma que ocorrem na PIAP incluem kimberlitos, lamproítos e kamafugitos, em adição a grandes complexos intrusivos, compostos por rochas máficas plutônicas (principalmente dunitos e piroxenitos) e carbonatitos.

O contexto geológico da região têm sido, há muito, intensamente estudado devido a ocorrências anômalas de substâncias como fosfato, titânio, nióbio e elementos terras raras, relacionadas à alteração intempérica dos complexos alcalinos. Muitas vezes, o manto de intemperismo alcança várias dezenas de metros de profundidade, nesse caso sendo possível observar rochas frescas somente em testemunhos de sondagem (Figura 22).

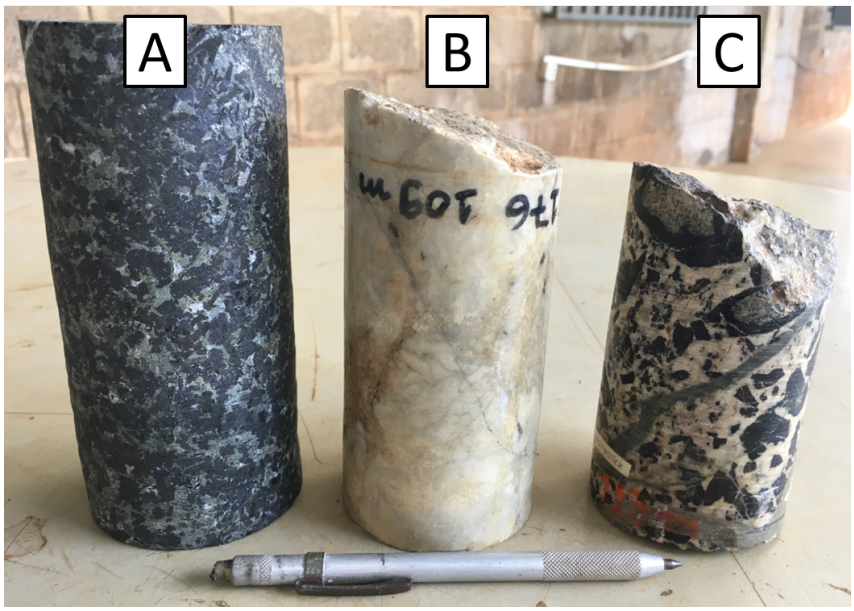

Figura 22 – Amostras de rochas frescas obtidas a partir de testemunhos de sondagem: (A) piroxenito (bebedourito); (B) carbonatito; (C) brecha magmática.

Na área de estudo, concentrações anômalas de fosfato e titânio estão relacionadas à ocorrência de apatita (de fórmula geral $Ca_5(F, Cl, OH)(PO_4)$) e anatásio (TiO_2), respectivamente. A concentração supergênica desses minerais se deu pela solubilização e lixiviação de componentes mais instáveis, como minerais máficos e carbonatos, contidos nas rochas originais, principalmente piroxenitos (bebedouritos). Nos horizontes mais

superficiais, onde a alteração foi mais intensa, a apatita foi quase totalmente lixiviada, porém, em horizontes mais profundos, permaneceu como um mineral resistato. O anatásio, formado pela descalcificação da perovskita ($CaTiO_3$), ocorre em um horizonte sobreposto às concentrações de apatita, conforme o perfil esquemático da jazida, apresentado na Figura 23. Próximo à superfície, esses elementos não ocorrem, tendo sido quase que totalmente lixiviados.

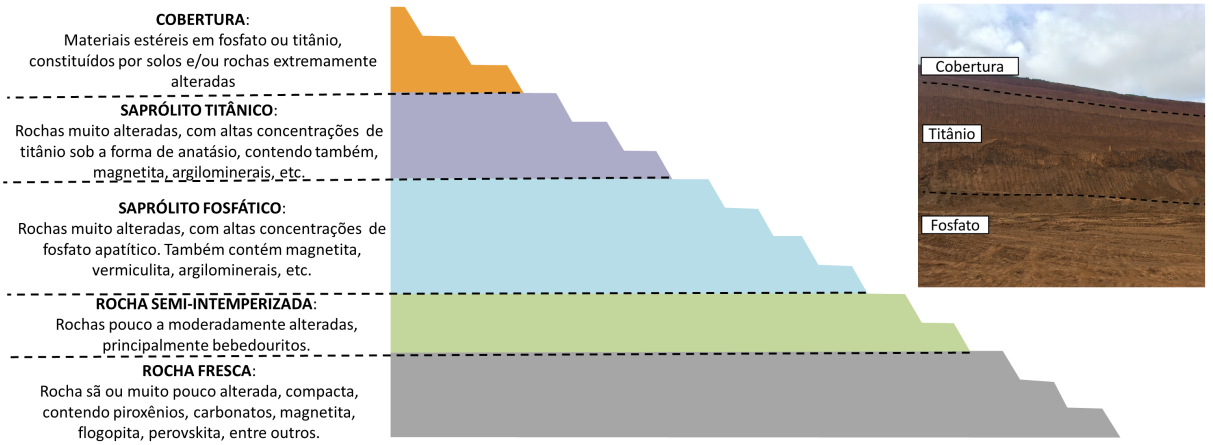

Figura 23 – Perfil esquemático da jazida relacionada a este estudo de caso. No detalhe, foto de uma porção da mina. As espessuras são aproximadas, sendo que cada banco mede 10 metros de altura.

Assim, com relação ao processo de mineralização: durante a evolução do manto intempérico, inicialmente são lixiviados elementos móveis, ocorrendo uma concentração de apatita, enquanto o titânio permanece sob a forma de perovskita, porém sem apresentar concentrações relevantes, conforme mencionado por Capponi (2012). Com o avanço do processo, a apatita é transformada em minerais secundários, do grupo da crandalita, de solubilidade reduzida, por isso de baixo valor econômico. Enquanto isso, a perovskita transforma-se em anatásio que, aos poucos, se concentra residualmente. Assim, a porção intermediária-inferior do manto de intemperismo constitui-se como minério de fosfato apatítico, com baixas concentrações de titânio, sob a forma tanto de perovskita como de anatásio, enquanto a porção intermediária-superior do perfil constitui-se de minério de titânio, sob a forma de anatásio, com baixas concentrações de fosfato apatítico.

Para a definição de unidades para modelagem e estimativa, é importante considerar tanto os tipos de rochas, quanto os padrões de alteração intempérica. Os principais litotipos encontrados na jazida estão sintetizados no quadro da Figura 24, enquanto as tipologias relativas à alteração intempérica, na Figura 25.

SIGLA	DENOMINAÇÃO	DESCRIÇÃO
COB	**Cobertura**	Denominação genérica para materiais estéreis, não mineralizados em fosfato e/ou titânio. Inclui solos, argilas, turfa, canga e laterita.
ZTI	**"Zona de titânio"**	Saprólitos oriundos da alteração sobretudo de rochas das séries bebedourítica e foscorítica. Composto principalmente por anatásio, magnetita e argilominerais. Em geral, mineralizado em titânio.
BEB	**Série bebedourítica**	Grupo de rochas silicáticas que inclui piroxenitos, bebedouritos e flogopititos, dentre outros que apresentam quantidades iguais ou superiores a 50% de piroxênio e/ou contenham perovskita. O termo flogopitito abrange a rocha ígnea ou metassomática constituída por mais de 50% de flogopita ou vermiculita, incluindo o antigo glimerito e o lamprófiro.
FCR	**Série foscorítica**	Grupo de rochas constituídas por variações modais em apatita, magnetita e silicatos magnesianos (olivina e/ou flogopita), que inclui foscorito, nelsonito, apatitito, magnetitito e dunito.
FET	**Foscrete**	Material fosfático oriundo da remobilização e recristalização da apatita. Rico em apatita e oxi-hidróxidos de Ferro.
CBN	**Série Carbonatítica**	Abrange os calciocarbonatitos e os magnesiocarbonatitos.
SIE	**Sienito**	Inclui as rochas ígneas saturadas (quartzo ausente ou subordinado) formadas por K-feldspato predominando sobre plagioclásio e minerais ferromagnesianos, como biotita e hornblenda. Sienito (plutônico) e traquito (equivalente vulcânico).

Figura 24 – Quadro com as principais tipologias relacionadas aos tipos de rochas encontradas no depósito.

SIGLA	DENOMINAÇÃO	DESCRIÇÃO
ALO	**Aloterito**	Primeiro horizonte intempérico do manto de alteração, constituído por material muito intemperizado, contendo principalmente argilominerais e sem preservação de estruturas originais da rocha. Estéril em fosfato ou titânio.
ISAT	**Isalterito de topo**	Saprólito oriundo da alteração de rochas plutônicas alcalinas, com preservação de algumas feições da rocha original. Geralmente com concentrações econômicas de titânio sob a forma de anatásio, podendo conter, também, quantidades expressivas de magnetita e argilominerais.
ISAB	**Isalterito de base**	Saprólito oriundo da alteração de rochas plutônicas alcalinas, com preservação de algumas feições da rocha original. Geralmente com concentrações econômicas de fosfato sob a forma de apatita, podendo conter, também, quantidades expressivas de vermiculita, magnetita e argilominerais.
RSI	**Rocha semi-intemperizada**	Horizonte intempérico subjacente ao isalterito de base, constituído por material rochoso semi-alterado, cujo grau de coesão permite a desagregação manual com dificuldade ou somente com o uso de martelo. Pode conter concentrações econômicas de fosfato, dependendo dos teores de P_2O_5 e da razão CaO/P_2O_5.
RSA	**Rocha sã**	Horizonte de base, constituído por material rochoso não alterado.

Figura 25 – Quadro com as tipologias referentes aos padrões de alteração intempérica no contexto do depósito.

3.2 Apresentação e validação dos dados

O banco de dados considerado neste estudo é constituído por amostras oriundas de testemunhos de sondagem rotativa diamantada, cujos furos apresentam direção vertical.

As variáveis consideradas são provenientes de resultados laboratoriais para 12 óxidos (P_2O_5, Fe_2O_3, MgO, CaO, Al_2O_3, SiO_2, TiO_2, MnO, Na_2O, K_2O, BaO e Nb_2O_5) e perda por calcinação (PPC), expressos em termos de porcentagens. Além dessas variáveis contínuas, estão presentes duas variáveis categóricas, relacionadas a características geológicas – litologia e intemperismo – conforme sintetizadas nas Figuras 24 e 25.

Para a aplicação de algoritmos de agrupamento, é necessário que o conjunto de dados seja isotópico, ou seja, apresente valores para todas as variáveis consideradas. Assim, foi feita a isotopização através da retirada de amostras que não apresentassem resultados completos. Além disso, outras tratativas se fizeram necessárias, conforme as etapas abaixo:

(i) Eliminação das amostras que não apresentassem resultados para todas as variáveis, isto é, os 12 óxidos, perda por calcinação (PPC), litologias e intemperismo;

(ii) Eliminação de amostras com fechamento estequiométrico (soma total dos teores dos 12 óxidos e PPC) fora do intervalo [95%, 103%];

(iii) Regularização do suporte amostral para 3 metros de comprimento, respeitando contatos geológicos (litologias e intemperismo).

Com relação ao item (*iii*) acima, após as etapas (*i*) e (*ii*), o conjunto de dados apresentava suporte amostral ligeiramente irregular, como mostra o histograma da Figura 26(A). A regularização do suporte amostral se dá a fim de que as análises partam de um banco de dados mais homogêneo, sendo que, de acordo com Rossi & Deutsch (2014), a regularização do comprimento amostral reduz a variabilidade, fazendo com que as análises geoestatísticas correspondentes, incluindo a variografia, sejam mais robustas. Além disso, esse procedimento reduz a quantidade de amostras, facilitando o manuseio dos dados e trazendo maior eficiência aos processos computacionais (Raymond, 1982 *apud* Sinclair & Blackwell (2004)). A tolerância mínima aplicada foi de 50% do comprimento da compósita para mais ou para menos.

Após o tratamento, de um total original de 34.295 amostras, o conjunto de dados resultante passou a contar com um total de 19.344 amostras. Na Figura 26(B), é apresentado o histograma do comprimento amostral dos dados tratados que, a partir deste ponto, serão referidos apenas como “dados”.

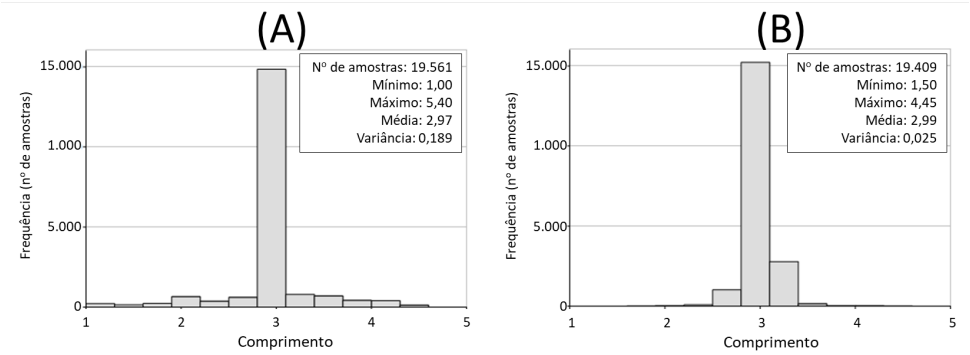

Figura 26 – Histogramas do comprimento das amostras. (A) Conjunto de dados pré-regularização (após as etapas (*i*) e (*ii*) do tratamento); (B) Conjunto de dados após a regularização.

3.3 Análise exploratória de dados

Assim, o banco de dados utilizado neste estudo de caso é constituído por 19.344 amostras, distribuídas em 527 furos verticais, dispostos em uma malha semirregular no plano horizontal, do qual ocupa uma área de aproximadamente 17km^2 (Figura 27). Há um adensamento da malha em uma faixa de direção NE-SW na porção NW da área.

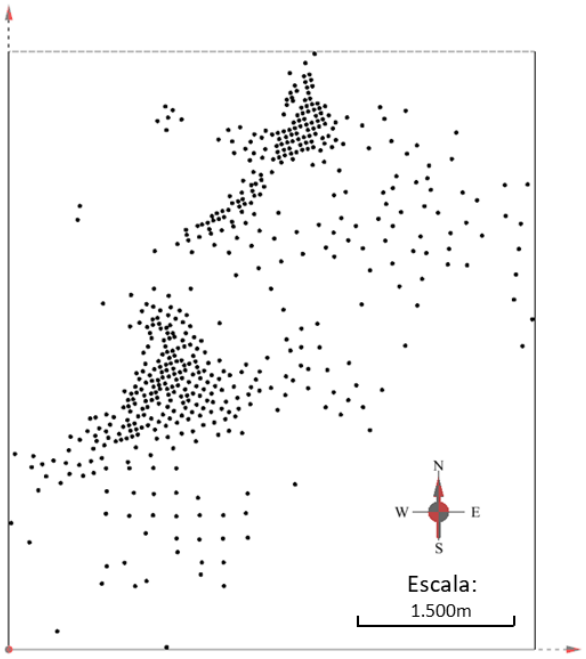

Figura 27 – Mapa com a distribuição espacial dos furos de sondagem usados neste estudo.

Nas Figura 28 e 29 são apresentadas seções verticais representativas, de direção NE-SW (N30ºE), que evidenciam o forte caráter de distribuição horizontal das litologias e dos padrões de alteração intempérica, bem como da distribuição dos teores das variáveis principais (P_2O_5, TiO_2 e CaO).

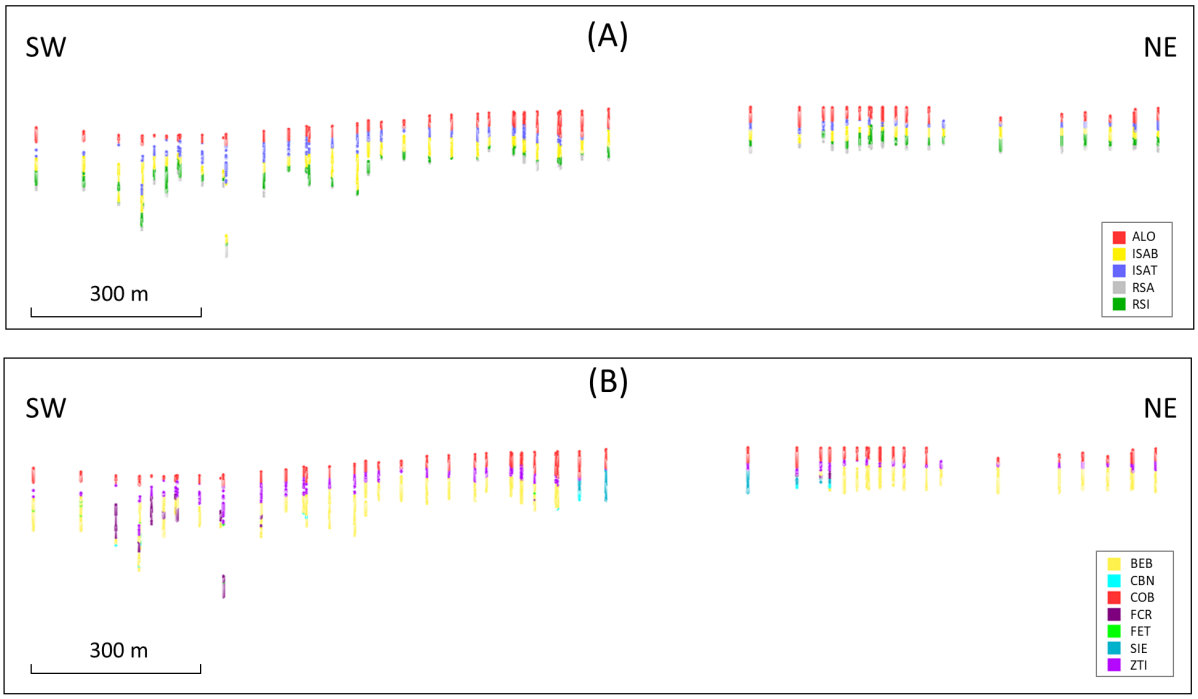

Figura 28 – Seções verticais de direção NE-SW (N30ºE) mostrando parte representativa dos dados, com amostras simbolizadas de acordo com as variáveis categóricas: (A) intemperismo e (B) litologias.

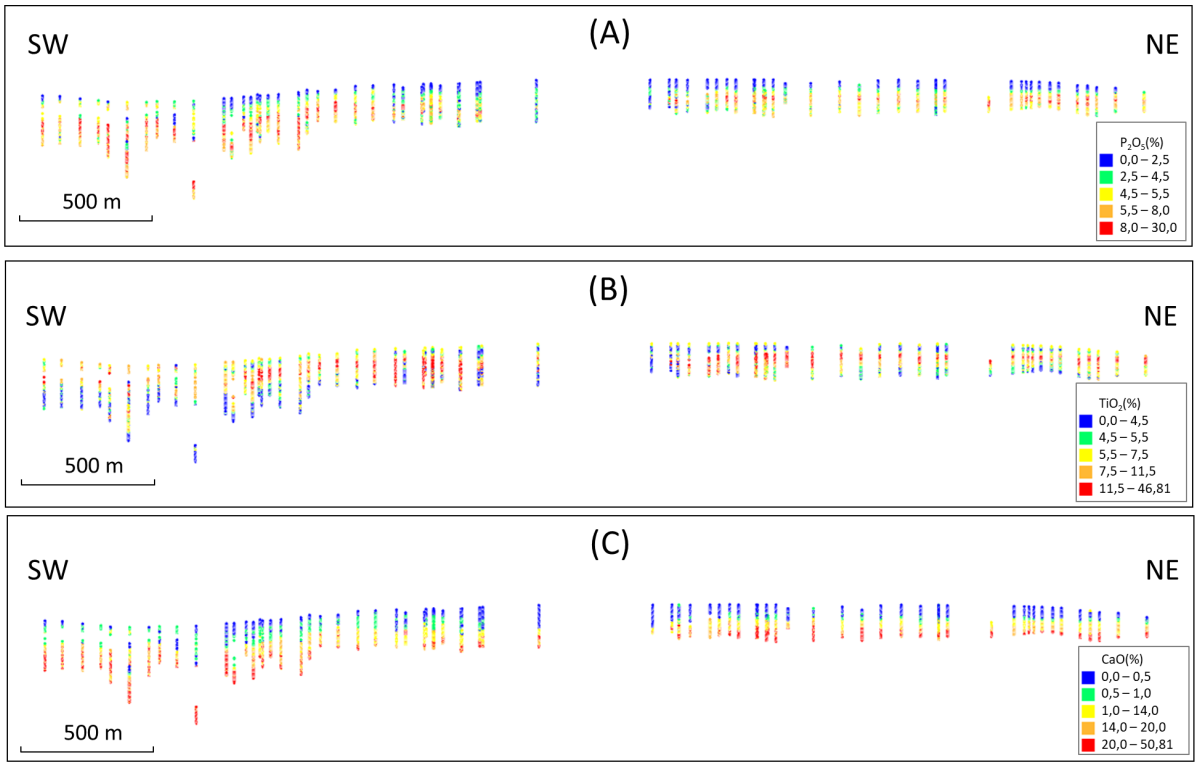

Figura 29 – Seções verticais de direção NE-SW (N30ºE) mostrando parte representativa dos dados, com amostras simbolizadas pelos teores de P_2O_5 (A), TiO_2 (B) e CaO (C).

As variáveis contínuas (óxidos e PPC) apresentam distribuições distintas, por vezes com alta assimetria, outras vezes bimodais, como pode ser observado nos histogramas da Figura 30. Além disso, as relações entre elas são um tanto complexas, o que pode ser notado através dos gráficos de dispersão da Figura 31. Na Figura 32 é apresentada a matriz com os coeficientes de correlação de Pearson entre as variáveis contínuas.

A Figura 33 apresenta gráficos de barras com as proporções de amostras correspondentes às classificações por litologia e intemperismo. Vale observar que a classificação, tanto litológica quanto por intemperismo, é feita interpretativamente pelo(a) geólogo(a), na fase de aquisição de dados, com base em descrição dos testemunhos de sondagem e análise de teores.

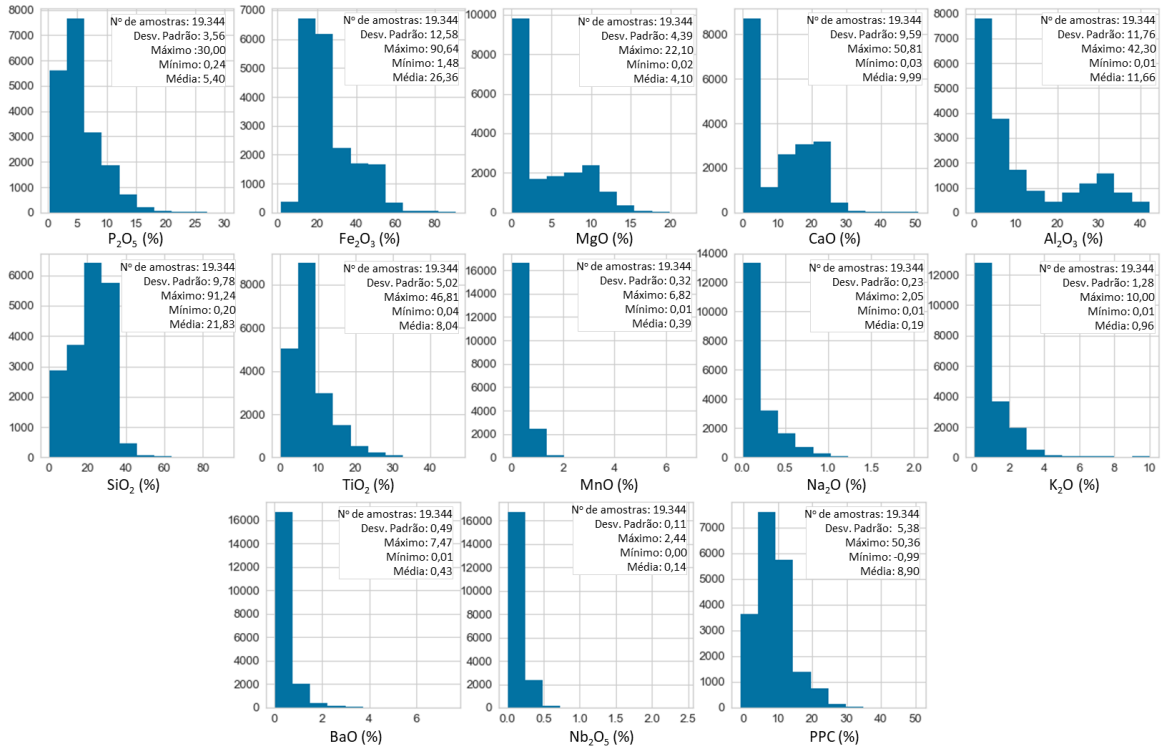

Figura 30 – Histogramas das variáveis contínuas.

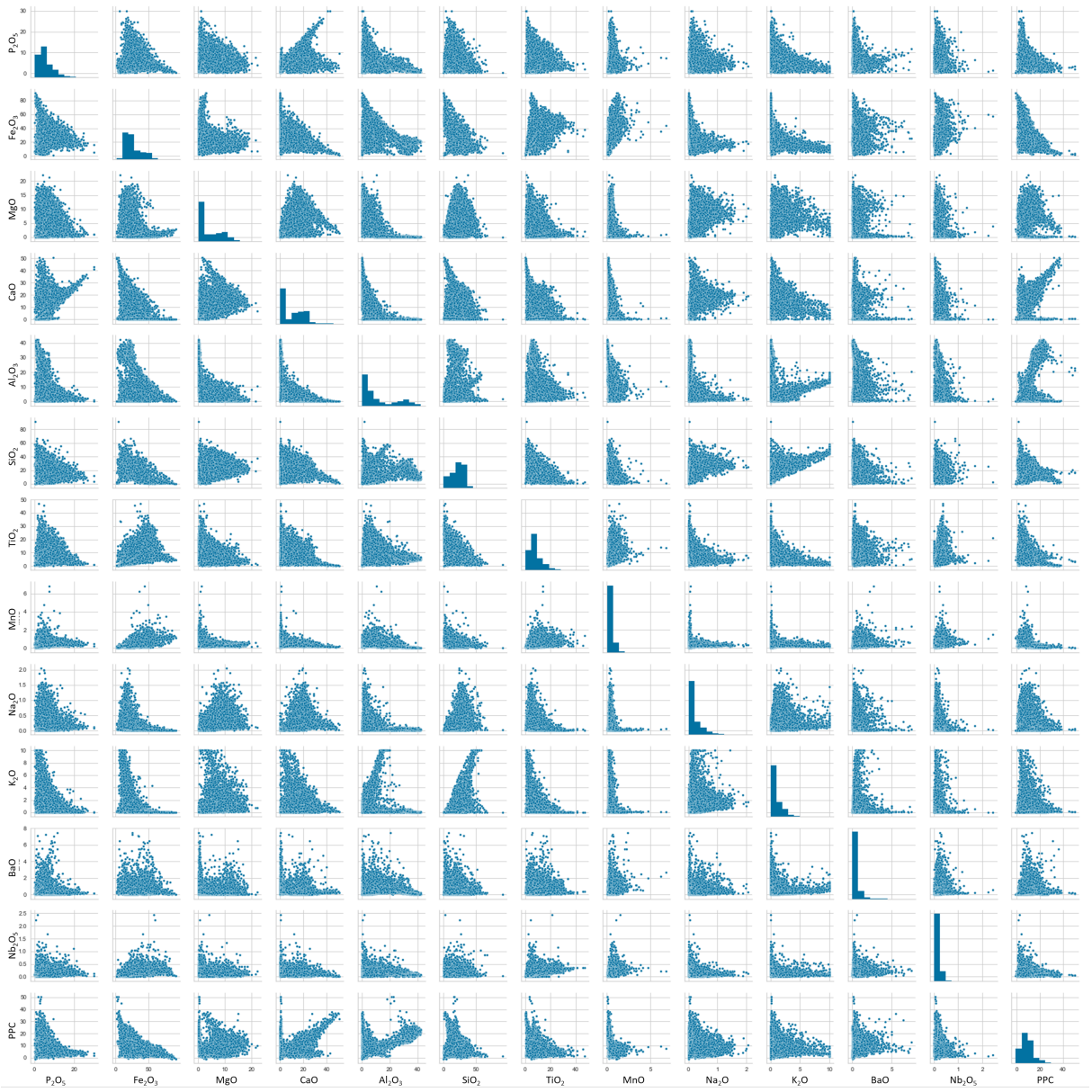

Figura 31 – Gráficos de dispersão das variáveis contínuas presentes no banco de dados, plotadas duas a duas. Na diagonal principal, os histogramas de cada variável.

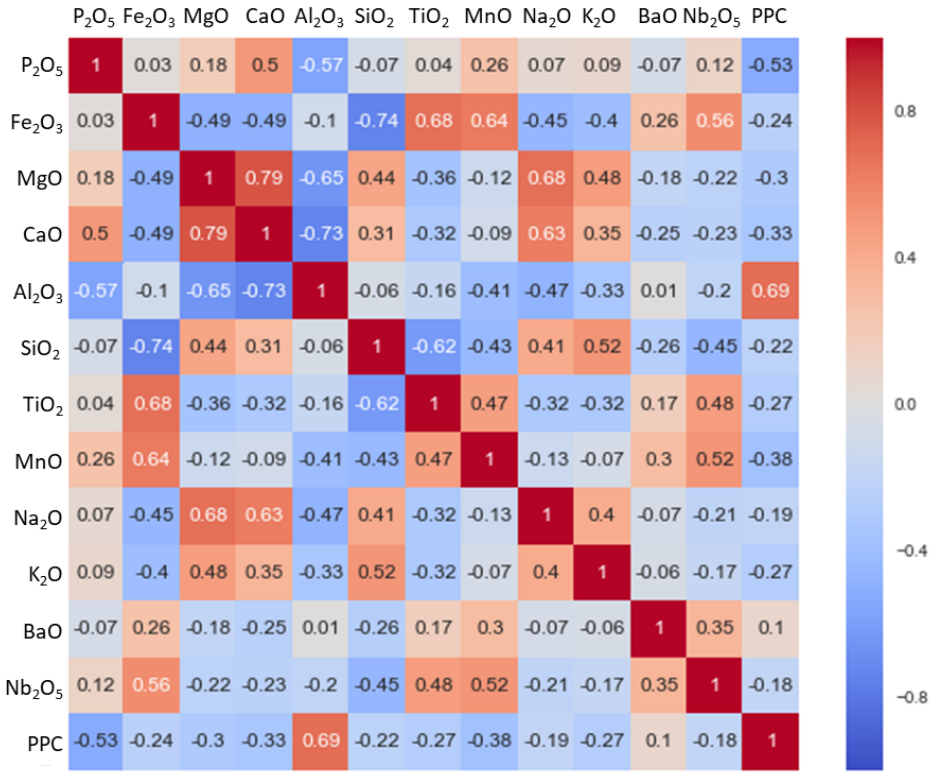

Figura 32 – Matriz com os coeficientes de correlação de Pearson entre as variáveis contínuas. Células azuis representam correlações negativas e células vermelhas, correlações positivas.

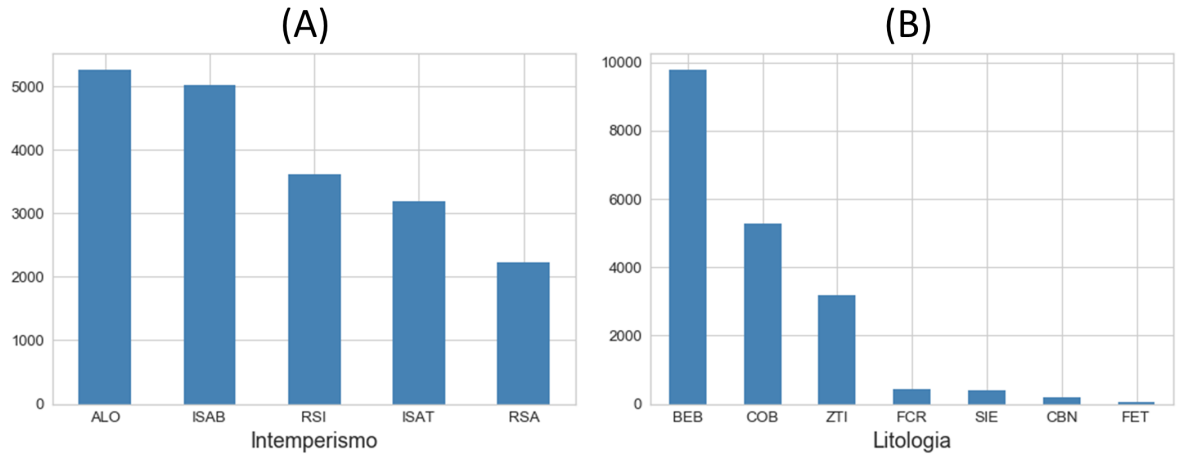

Figura 33 – Gráficos de barras com a quantificação de amostras classificadas por intemperismo (A) e por litologia (B).

Uma análise que também deve ser feita é a distribuição dos teores de acordo com a litologia e o intemperismo. Desta maneira, as Figuras 34 e 35 mostram os *boxplots* das variáveis contínuas para cada tipologia definida pelo intemperismo e pelo tipo de rocha, respectivamente. A Figura 36 apresenta as relações entre as variáveis contínuas mais relevantes (P_2O_5, TiO_2 e CaO), plotadas duas a duas, em diagramas de dispersão, com

pontos simbolizados por litologia (Figura 36(A)) e intemperismo (Figura 36(B)). Pode-se perceber que as distribuições de teores estão, de fato, intimamente relacionadas com as tipologias.

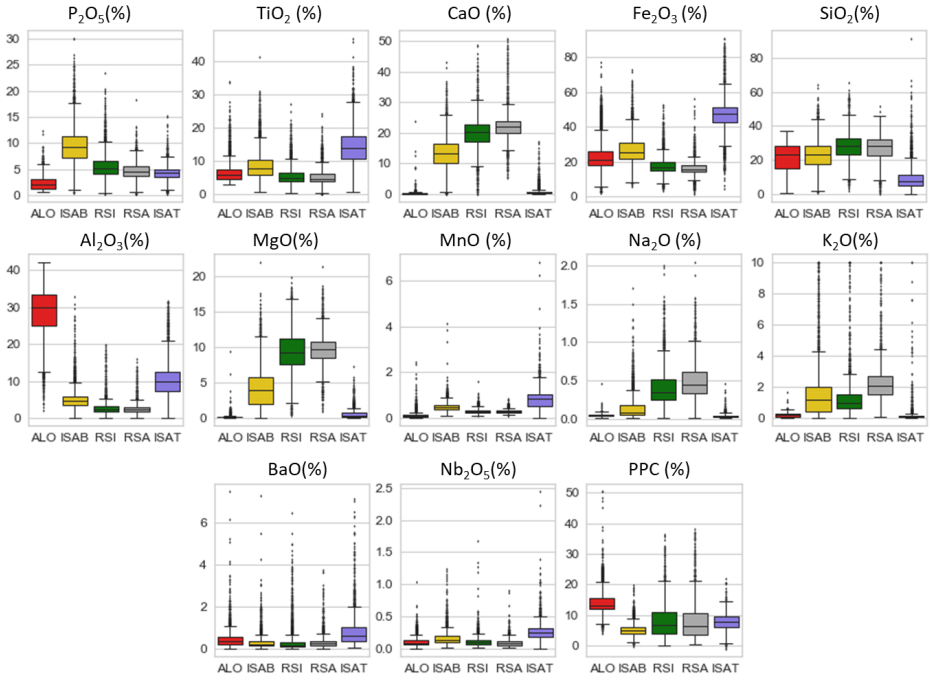

Figura 34 – *Boxplots* de teores por tipologia definida pelo intemperismo.

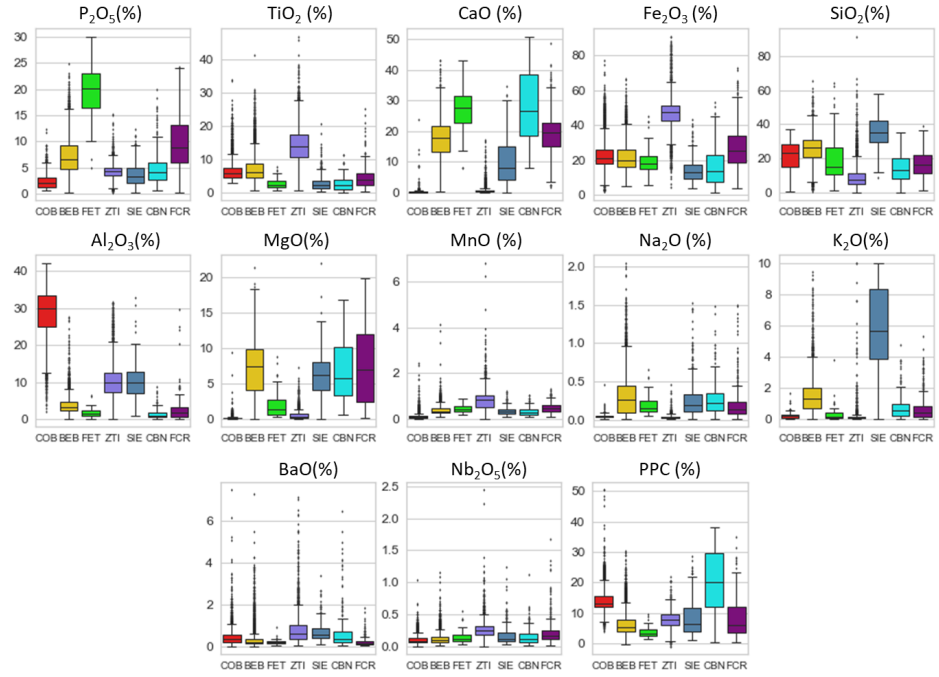

Figura 35 – *Boxplots* de teores por litologia.

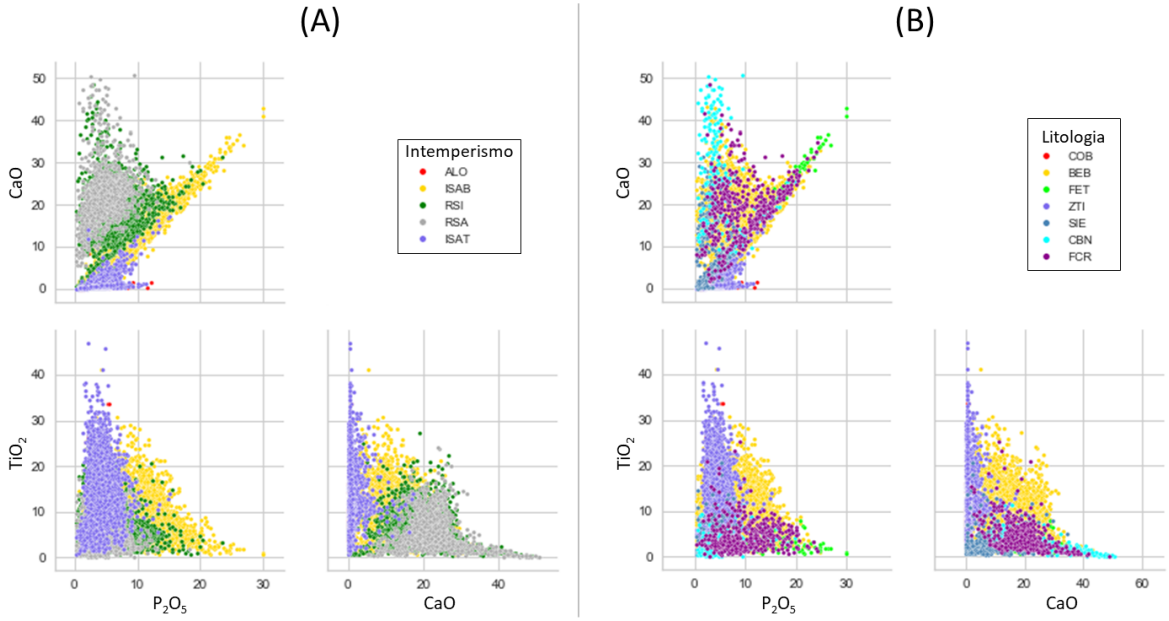

Figura 36 – Diagramas de dispersão das principais variáveis contínuas (P_2O_5, TiO_2 e CaO) presentes no banco de dados, plotadas duas a duas. As cores se referem às tipologias: intemperismo (A) e litologia (B).

Quanto à continuidade espacial, a Figura 37 mostra os variogramas experimentais das variáveis principais, nos quais pode-se perceber que P_2O_5 e CaO são aproximadamente isotrópicos no plano horizontal, com alcances na ordem de 1.000m, enquanto na direção vertical o alcance é de cerca de 30m para o primeiro e 50m para o segundo. Já o TiO_2 mostra considerável anisotropia também no plano horizontal, com alcance de 300m na direção N-S e mais de 1.000m na E-W e, na direção vertical, 25m.

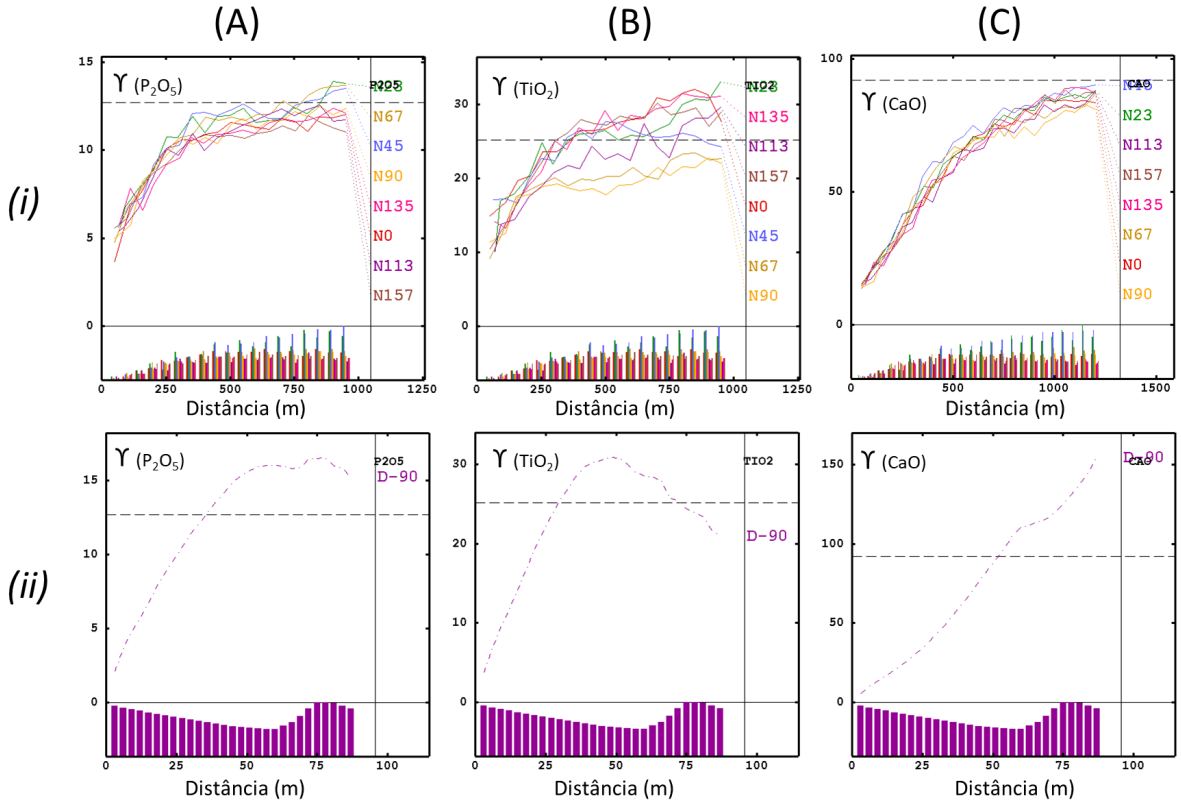

Figura 37 – Variogramas experimentais das principais variáveis contínuas (P_2O_5 (A), TiO_2 (B) e CaO (C)) em diferentes direções: plano horizontal (i); direção vertical (ii).

Outra técnica de interessante aplicação na fase de análise exploratória de dados é a observação da tendência natural que os dados têm de se agrupar, o que pode ser feito a partir do dendrograma, oriundo da aplicação do método aglomerativo hierárquico, como descrito na respectiva Seção, no Capítulo 2. Nesse caso, pela Figura 38, nota-se que o número mais adequado no qual se dividir os dados parece estar entre dois e oito. A partir de nove, as distâncias no espaço multivariado são muito reduzidas, o que é evidenciado pelo curto comprimento das linhas do dendrograma, e já não se justifica a aplicação do agrupamento.

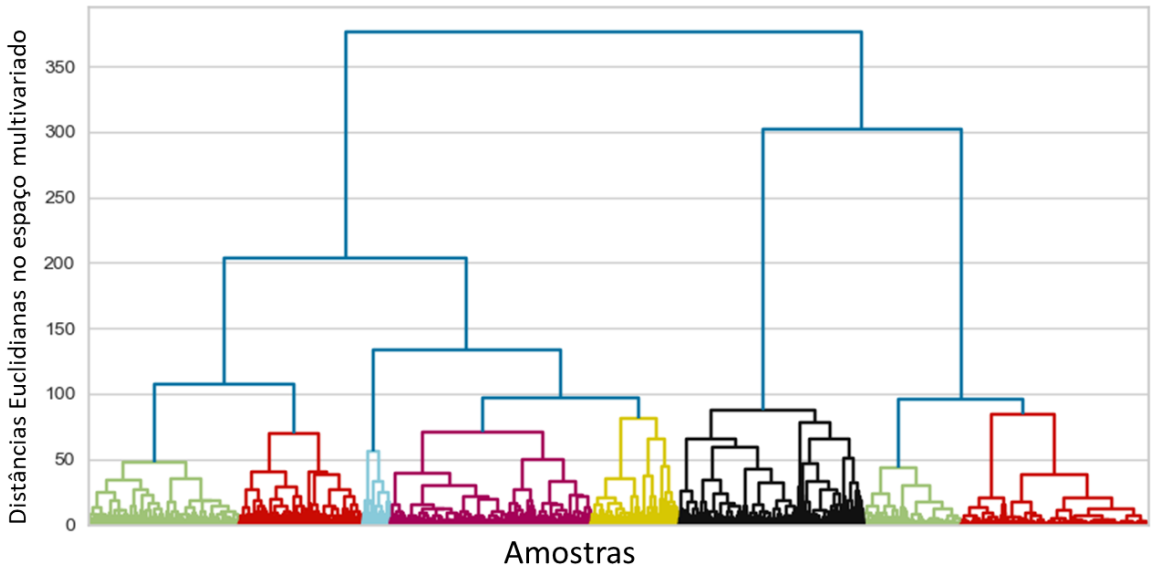

Figura 38 – dendrograma obtido com a aplicação do agrupamento aglomerativo hierárquico, mostrando a tendência natural que os dados têm de se agrupar (no espaço multivariado). As cores indicam as linhas e conexões caso os dados fossem aglomerados em oito grupos

3.4 Metodologias para a aplicação dos algoritmos de agrupamento e das técnicas de validação

A fim de observar o desempenho de diversas técnicas de agrupamento de dados, quatro algoritmos foram aplicados neste estudo de caso, dois deles do tipo tradicional e dois do tipo espacial:

(i) *k-means* (MACQUEEN, 1967);

(ii) Aglomerativo hierárquico (SOKAL; SNEATH, 1963);

(iii) Agrupamento em espaço duplo (*dsclus*) (MARTIN; BOISVERT, 2018);

(iv) Agrupamento por estatísticas de autocorrelação (*acclus*) (SCRUCCA, 2005).

Tanto para o *k-means* quanto para o hierárquico, foram usados os algoritmos disponíveis na biblioteca de aprendizado de máquina *Scikit-learn* (PEDREGOSA et al., 2011). Para o *k-means*, foi utilizada a opção *k-means++* para a inicialização dos centroides (ARTHUR; VASSILVITSKII, 2007), que busca maximizar a posição dos pontos iniciais, aumentando a acurácia e a eficiência. Já para o agrupamento hierárquico, a opção "*ward*" foi usada para medida de proximidade.

Para o *dsclus* e o *acclus*, foram aplicados os algoritmos disponíveis no *GitHub*, na conta mencionada em Martin & Boisvert (2018).

De acordo com Prades (2017), complexidades nos dados, como a presença de assimetria e *outliers*, podem levar a impactos consideráveis, tanto na modelagem geoestatística quanto na análise de agrupamento. Assim, para manter as relações na mesma escala, os dados foram padronizados, de acordo com:

$$Z = \frac{(X - m)}{s} \tag{3.1}$$

onde Z é o valor do dado padronizado, X é o valor original, m, a média dos dados originais e s, seu desvio padrão.

A flexibilidade dos parâmetros de busca dos algoritmos de agrupamento espacial (quantidade de vizinhos mais próximos – vmp – e volume de busca) permite avaliar os efeitos da quantidade de amostras utilizadas nos agrupamentos. A partir de testes realizados com diversos valores de vizinhos mais próximos, tendo um volume de busca fixo, verificou-se que quanto maior a quantidade de amostras nas vizinhanças de busca, melhores os resultados de entropia espacial, em detrimento da organização no espaço multivariado. Ao se variar o volume de busca observou-se que quanto menor, mais restrita é a distribuição geográfica dos grupos, porém piores são os resultados no espaço multivariado.

Feitos os testes, os parâmetros que geraram resultados mais consistentes foram:

(i) ***dsclus***: 30 vmp ao todo a se considerar e 20 vmp selecionados para formar os $k_{mini-grupos}$; volume de busca = (0, 0, 0, 400, 400, 12); 100 realizações;

(ii) ***acclus***: 30 vmp; volume de busca = (0, 0, 0, 400, 400, 12).

Assim, uma vez padronizados os dados e definidos os parâmetros mais adequados, foram gerados sete cenários distintos para cada algoritmo, cada um para um determinado número de grupos, de dois a oito, a fim de avaliar os resultados e selecionar o cenário e o algoritmo mais apropriados. O seguinte fluxo de trabalho foi executado:

(i) Cálculo dos índices de silhueta (ROUSSEEUW, 1987), Calinski-Harabasz (CALIŃSKI; HARABASZ, 1974) e Davies-Bouldin (DAVIES; BOULDIN, 1979) e obtenção das métricas em espaço duplo de Martin & Boisvert (2018): soma dos quadrados intragrupo ($wcss$) e entropia espacial (H);

(ii) Seleção dos cenários mais promissores a partir dos índices e métricas obtidas na etapa (i);

(iii) Verificação da continuidade espacial dos grupos através de correlogramas dos indicadores dos cenários escolhidos na etapa (ii);

(iv) Avaliação visual da distribuição espacial dos cenários selecionados nas etapas anteriores;

(v) Avaliação estatística dos grupos e comparação dos agrupamentos com as tipologias – litologia e intemperismo – através de gráficos de barras.

Todos os métodos foram executados em linguagem *Python* (versão 3.6.5, instalado via Anaconda), utilizando-se *Jupyter Notebooks*, com exceção da verificação da continuidade espacial e da análise visual, feitas com o *software Isatis*®, versão 2016.1. Quanto às configurações de sistema operacional e *hardware*: Windows 10, 64 *bits*, processador *Intel*® i7-3.20Ghz, com 24.0GB de memória RAM.

3.5 Apresentação de resultados

3.5.1 Cálculos de índices e métricas e seleção dos cenários mais promissores

Na Figura 39 podem ser observados os gráficos com os índices de Davies-Bouldin, Silhueta e Calinski-Harabasz, bem como as métricas *wcss* e entropia espacial (H), plotados nos eixos verticais, e os números de grupos (k) plotados nos eixos horizontais. Cada cor corresponde a um determinado algoritmo.

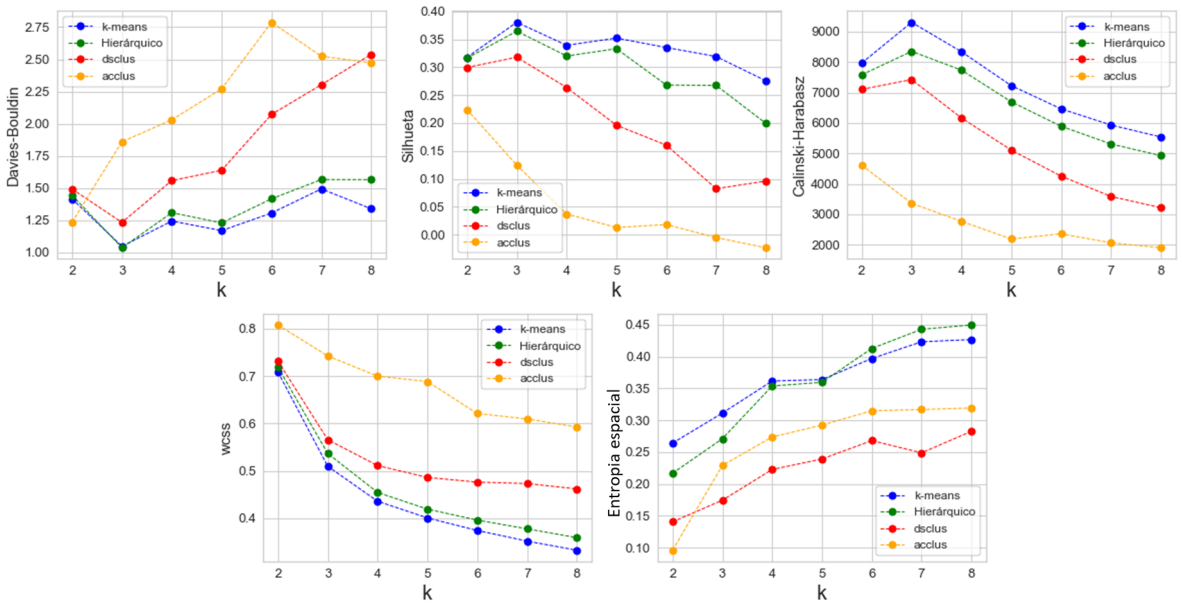

Figura 39 – Índices de Davies-Bouldin, Silhueta e Calinski-Harabasz e métricas *wcss* e H (nos eixos verticais) para diferentes números de grupos (k) (nos eixos horizontais). Para Davies-Bouldin, *wcss* e entropia espacial são desejáveis valores baixos, já para Silhueta e Calinski-Harabasz, valores altos.

Valores altos de Calinski-Harabasz e Silhueta e valores baixos de Davies-Bouldin e *wcss* indicam agrupamentos mais organizados no espaço multivariado. Portanto, percebe-se

que algoritmos tradicionais produzem melhores resultados no espaço multivariado, com o *k-means* superando o método hierárquico em todos os casos. De maneira oposta, os algoritmos espaciais mostram melhor desempenho no espaço geográfico, o que pode ser notado através do gráfico de H, sendo que o *dsclus* apresenta resultados mais interessantes do que o *acclus*, inclusive no espaço multivariado. Assim, conclui-se que o *dsclus* seja preferível dentre os demais, já que seus resultados mostram maior equilíbrio entre os aspectos geográfico e multivariado.

Para se verificar a conectividade espacial e a organização no espaço multivariado simultaneamente, as métricas em espaço duplo de Martin & Boisvert (2018), *wcss* e H, podem ser lançadas em um gráfico de dispersão (Figura 40). Valores baixos tanto para *wcss* quanto para H são desejáveis, entretanto verifica-se que suas relações são inversamente proporcionais. A solução é combinar essas métricas e avaliar os resultados comparativamente. Geralmente, os melhores resultados para o agrupamento de dados espaciais não são aqueles com o menor *wcss* ou menor H, mas aqueles intermediários, de modo que haja algum equilíbrio entre os aspectos multivariado e geográfico.

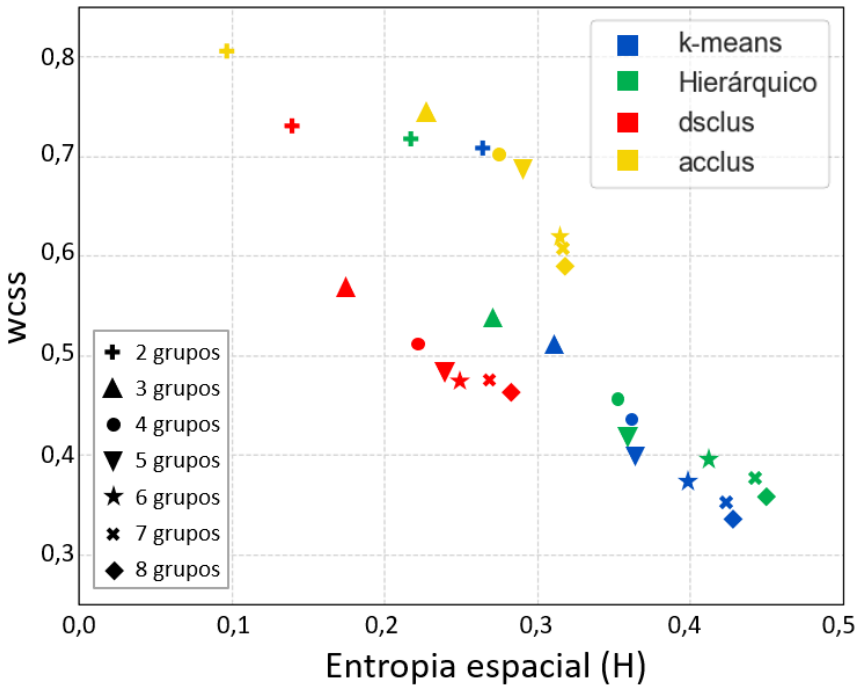

Figura 40 – Gráfico de dispersão *wcss versus* H, que permite avaliar simultaneamente a conectividade espacial e a organização multivariada dos grupos nas diferentes configurações de agrupamentos. Cada ponto corresponde a uma configuração distinta (a cor indica o algoritmo, e o ícone, o número de grupos).

A fim de se melhor explorar esses resultados, a Figura 41 apresenta gráficos de dispersão de cada um dos índices, plotados dois a dois. As nuances de cada método de

validação ficam mais evidentes ao se observar as relações entre os índices e, mais uma vez, pode-se perceber que cenários com valores intermediários parecem mais adequadas.

Quanto à escolha do número mais adequado de grupos, também são preferíveis valores intermediários dos índices e métricas utilizados. Por isso, apesar de demonstrar resultados aparentemente interessantes nos gráficos da Figura 39, o agrupamento desses dados em apenas dois ou três grupos pode levar a misturas inadequadas de populações, especialmente em se tratando de um caso com tantas variáveis e tamanha complexidade. No outro extremo, a divisão dos dados em oito grupos pode levar à complicações desnecessárias, criando grupos redundantes. Portanto, os cenários onde k =4, k = 5, k = 6 e k = 7 parecem ser mais promissores.

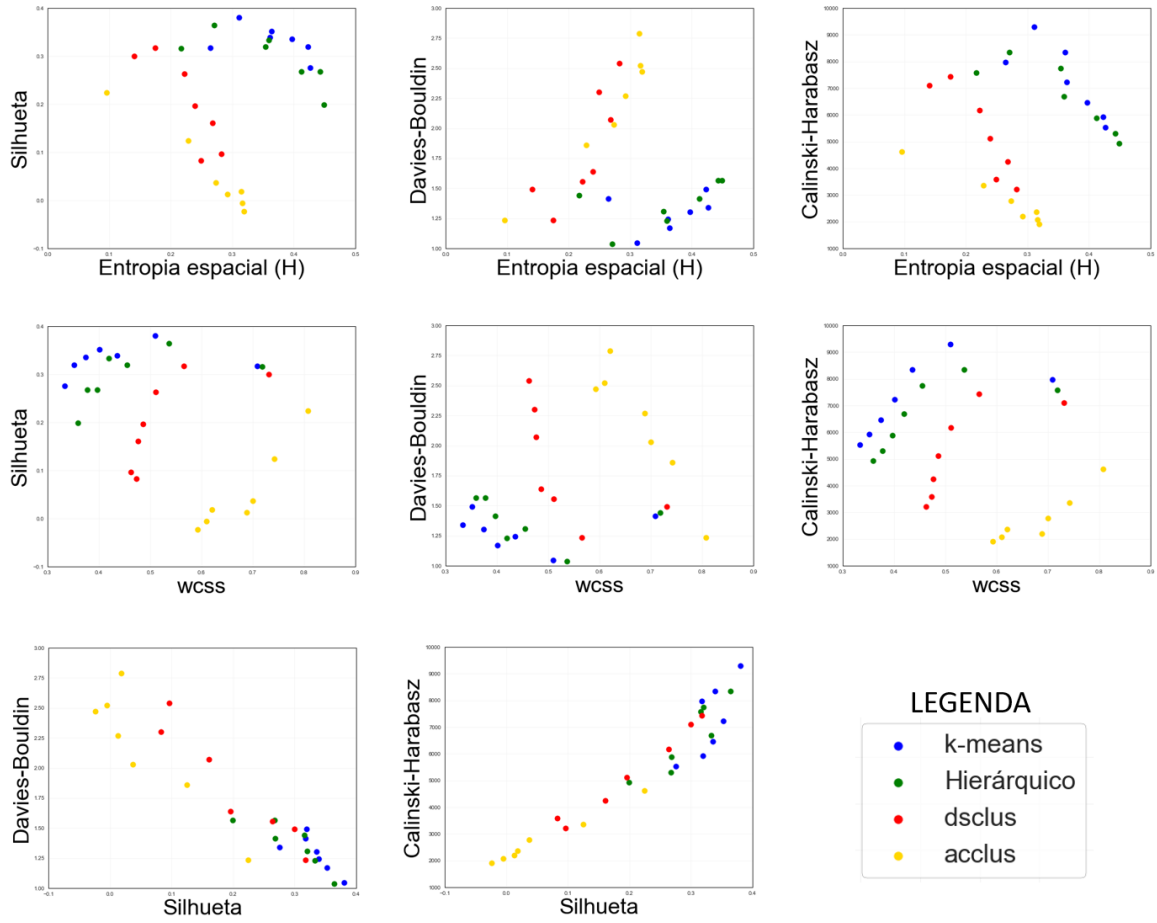

Figura 41 – Gráficos de dispersão entre os índices de Davies-Bouldin, Silhueta, Calinski-Harabasz e métricas *wcss* e *H*, plotados dois a dois. Para Davies-Bouldin, *wcss* e entropia espacial são desejáveis valores baixos, já para Silhueta e Calinski-Harabasz, valores altos.

3.5.2 Verificação da continuidade espacial dos grupos através de correlogramas dos indicadores

Além da entropia espacial, uma outra maneira de se avaliar a consistência geográfica dos grupos é medindo a continuidade espacial de seus indicadores, especialmente a curtas distâncias e, como já discutido na Sessão 2.4.5 desta Dissertação, os correlogramas são mais estáveis do que os variogramas, já que são padronizados pela variância, por isso foram usados. Cabe mencionar que o efeito pepita, bem como a continuidade a curtas distâncias, devem ser interpretadas a partir dos correlogramas da direção vertical, já que o espaçamento horizontal entre amostras é consideravelmente maior.

Os cenários verificados foram aqueles julgados mais promissores na etapa anterior, ou seja, os agrupamentos do algoritmo *dsclus* em quatro, cinco, seis e sete grupos. Tendo a distribuição dos dados um caráter fortemente horizontal, foram calculados os correlogramas experimentais nas direções N-S, E-W e vertical, de acordo com os parâmetros de busca apresentados na Figura 42. Os correlogramas são exibidos nas Figuras 43 a 46.

Direções	*Lag* (m)	Tolerância do *lag* (m)	Tolerância angular	Largura máxima da banda (m)	Altura da banda (m)
N-S	50	50	45°	25	5
E-W	50	50	45°	25	5
Vertical	3	3	45°	1,5	1,5

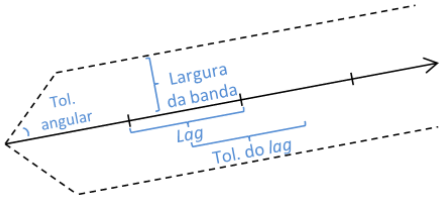

Figura 42 – Parâmetros de busca para o cálculo dos correlogramas experimentais dos indicadores.

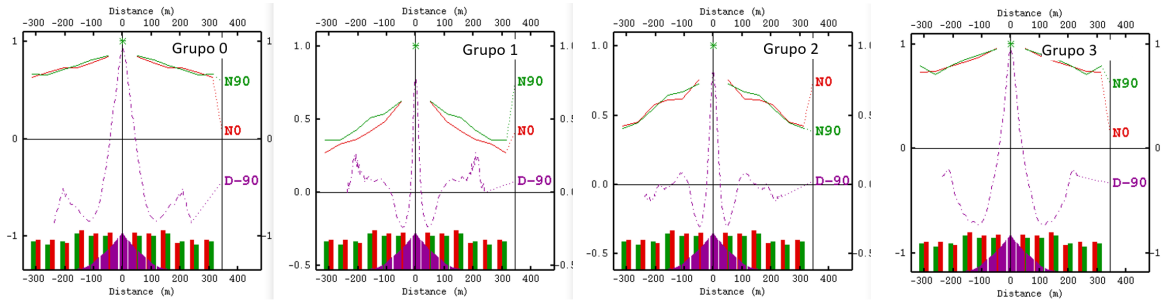

Figura 43 – Correlogramas experimentais dos indicadores no cenário com quatro agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical.

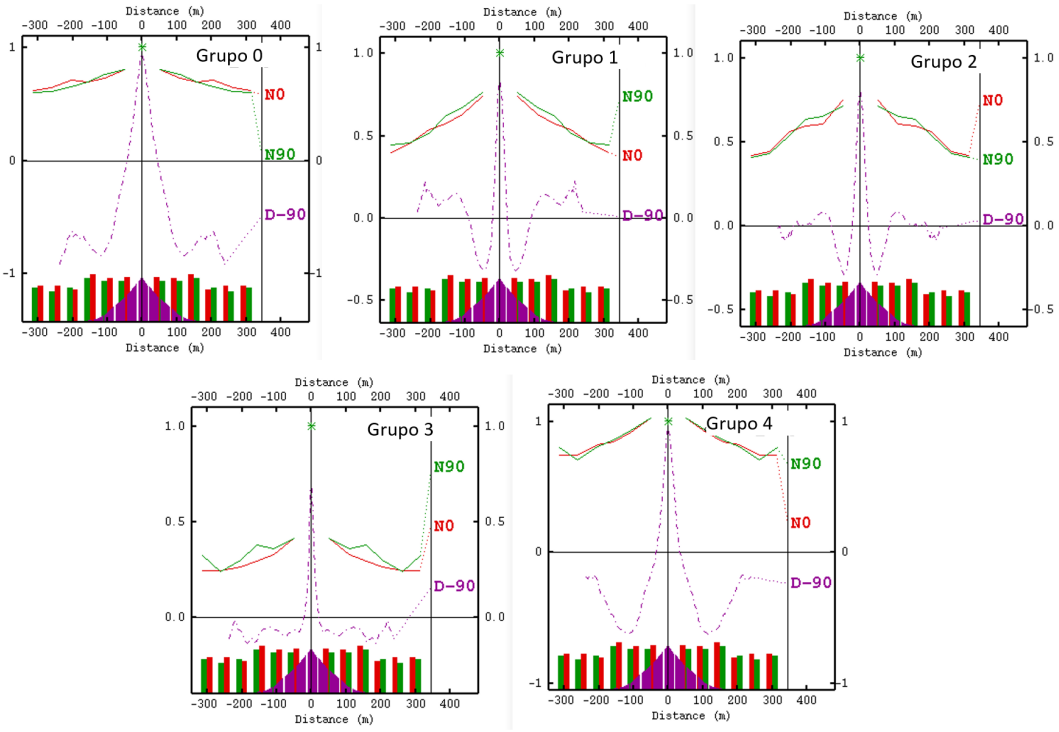

Figura 44 – Correlogramas experimentais dos indicadores no cenário com cinco agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical.

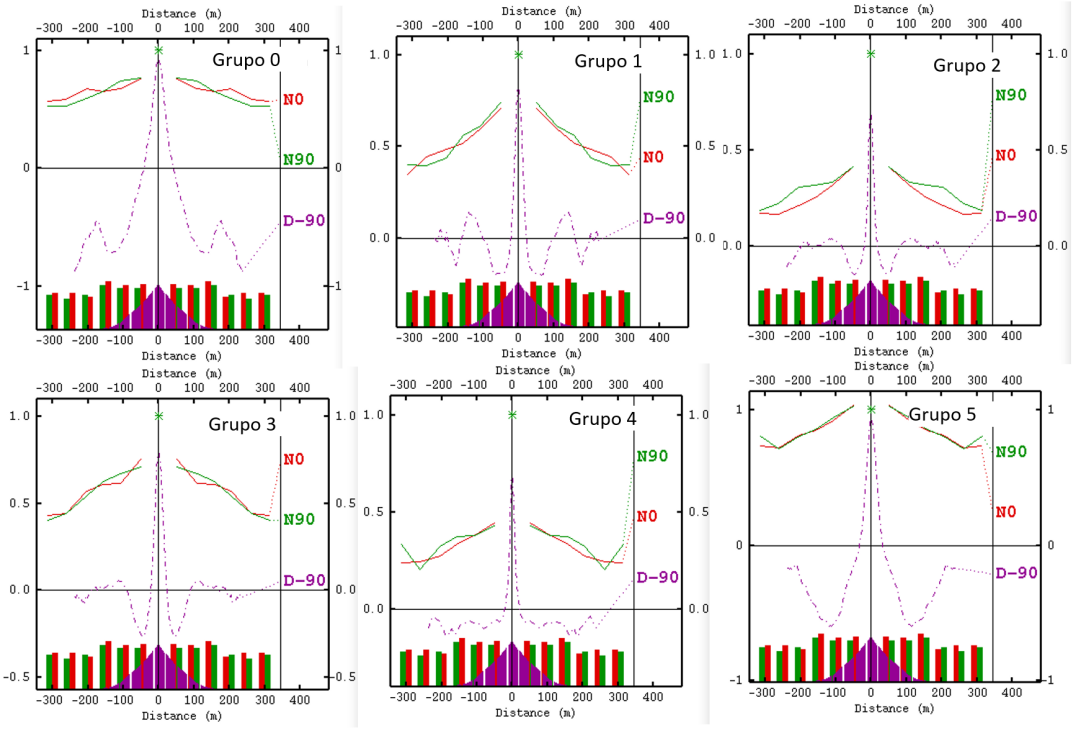

Figura 45 – Correlogramas experimentais dos indicadores no cenário com seis agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical.

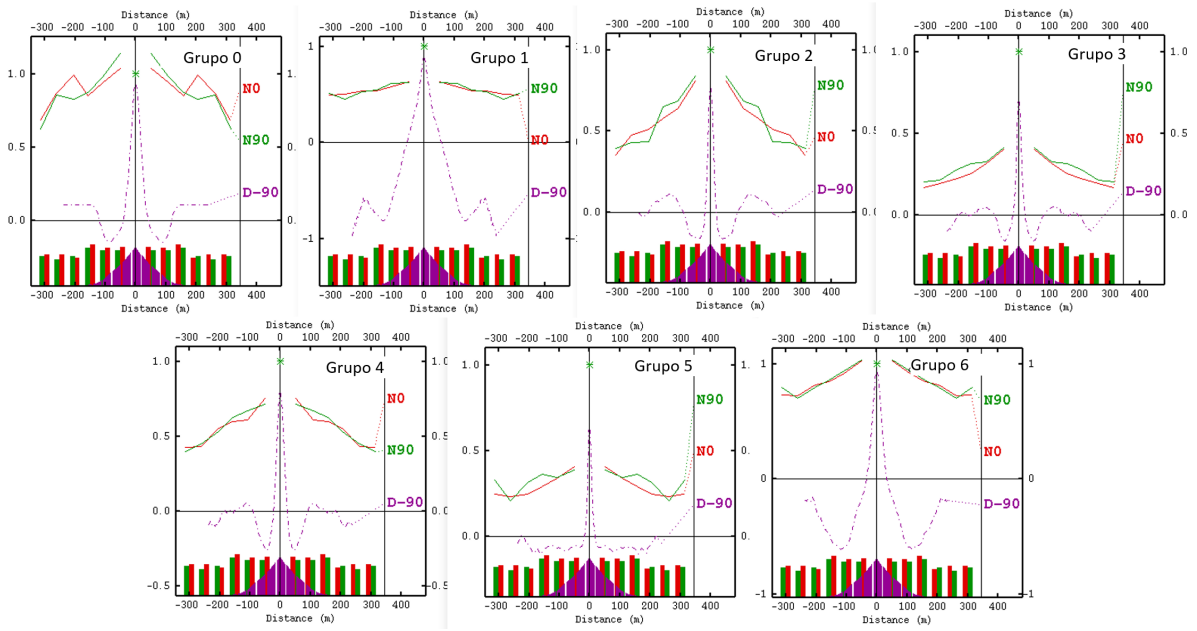

Figura 46 – Correlogramas experimentais dos indicadores no cenário com sete agrupamentos pelo algoritmo *dsclus*. Em vermelho e verde as direções N-S e E-W, respectivamente e, em magenta com linha tracejada, a direção vertical.

Todos os correlogramas das Figuras 43, 44 e 45 apresentam-se contínuos e com estruturação aceitável, além de efeitos pepitas relativamente baixos, características que indicam boa coesão geográfica dos grupos. Entretanto o correlograma do Grupo 5 no cenário com sete grupos mostra efeito pepita relativamente alto, correspondendo a quase 50% da variância total, o que leva a crer que seja um grupo fragmentado, ou com baixa representatividade amostral.

3.5.3 Avaliação visual

A Figura 47 mostra a distribuição espacial das amostras em uma seção vertical representativa de parte dos dados, de direção NE-SW (N30°E), simbolizadas de acordo com os grupos aos quais pertencem, nos cenários com quatro, cinco, seis e sete agrupamentos pelo algoritmo *dsclus*. Para que possam ser feitas comparações, a Figura 47 apresenta a mesma perspectiva das Figuras 28 e 29, que mostram as seções com as amostras simbolizadas por litologia, intemperismo e teores de P_2O_5, TiO_2 e CaO.

Assim, verifica-se que os agrupamentos mostram fortes relações com a distribuição de litologias e alteração intempérica, bem como com os teores das variáveis mostradas na Figura 29.

Nota-se, também, que ocorre alguma fragmentação de alguns grupos, como era esperado, já que a busca por uma definição ótima, considerando também a distribuição multivariada, implica em algum grau de desorganização no espaço geográfico. Na Figura

47(D), que mostra o cenário com sete grupos, nota-se a baixa representatividade do Grupo 5, que encontra-se restrito entre os Grupos 4 e 6, daí o efeito pepita alto, observado no respectivo correlograma (ver Fig. 46).

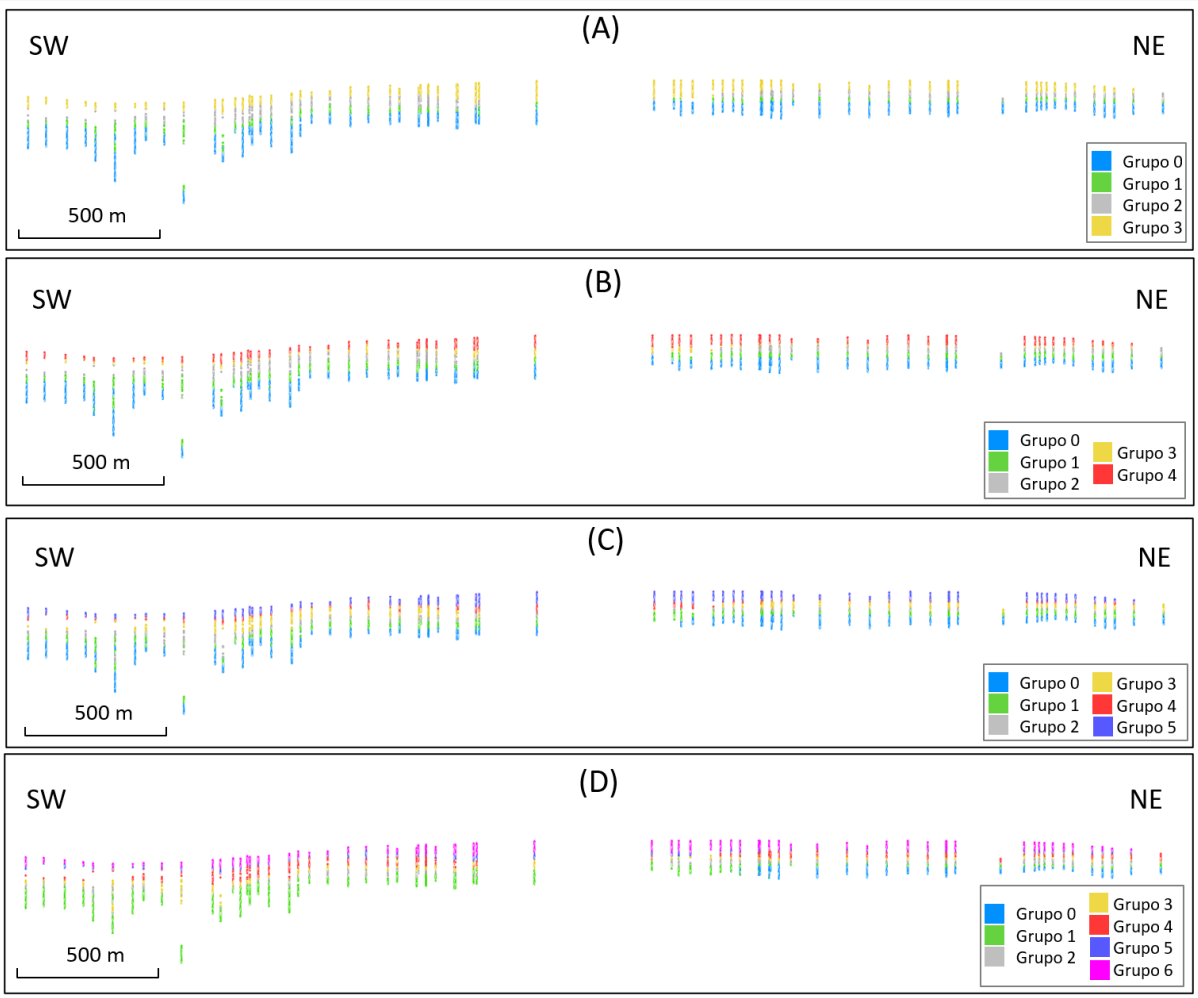

Figura 47 – Seções verticais mostrando parte dos dados, com amostras simbolizadas de acordo com os grupos aos quais pertencem nos cenários de agrupamento por *dsclus* em quatro (A), cinco (B), seis (C) e sete (D) grupos.

3.5.4 Avaliação da distribuição estatística dos grupos e comparação de agrupamentos com as tipologias

A fim de comparar as distribuições estatísticas de teores para cada grupo, são apresentados os *boxplots* nas Figuras 48 a 51. Como se pode notar, em cada caso, cada grupo apresenta características distintas, dependendo da variável considerada.

Por fim, foram feitas comparações quantitativas entre os agrupamentos e as tipologias: intemperismo e litologia. Os gráficos de barras da Figura 52 permitem verificar a correspondência entre as diferentes classificações.

A partir dessas relações, pode-se realizar a categorização de cada um dos grupos para constituir diferentes domínios para modelagem, como exposto no quadro da Figura 53.

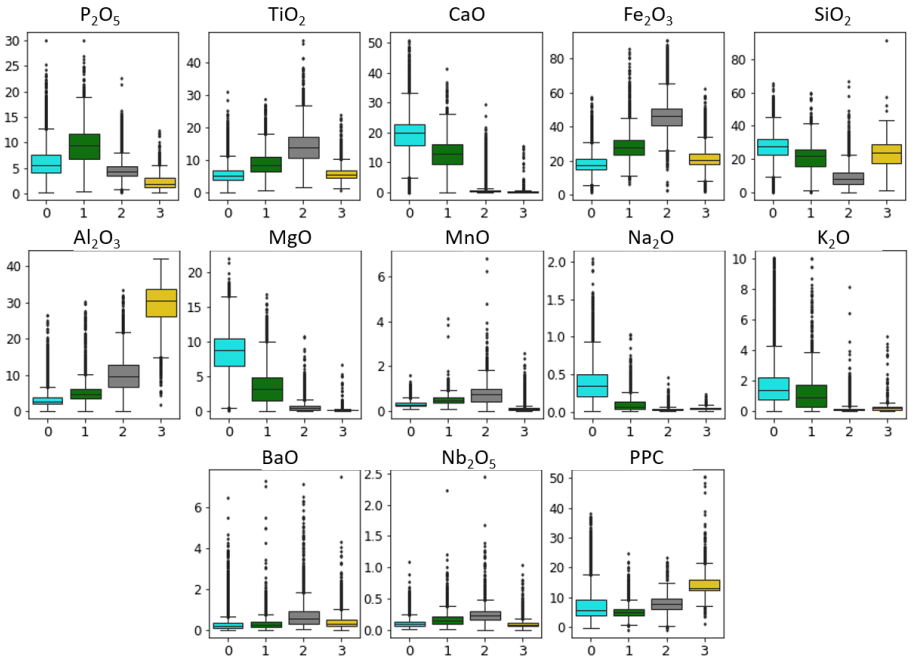

Figura 48 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com quatro agrupamentos por *dsclus*.

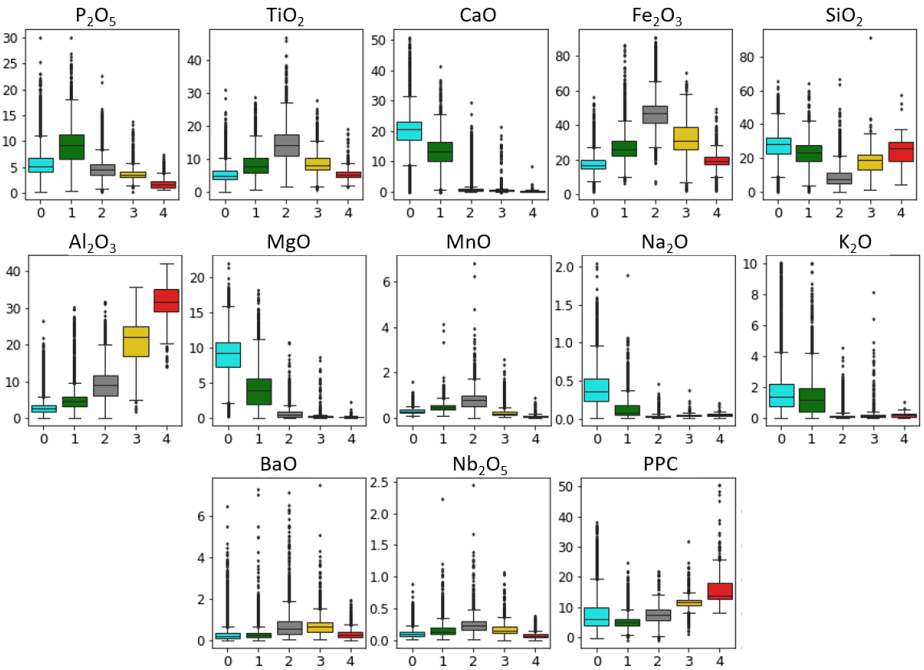

Figura 49 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com cinco agrupamentos por *dsclus*.

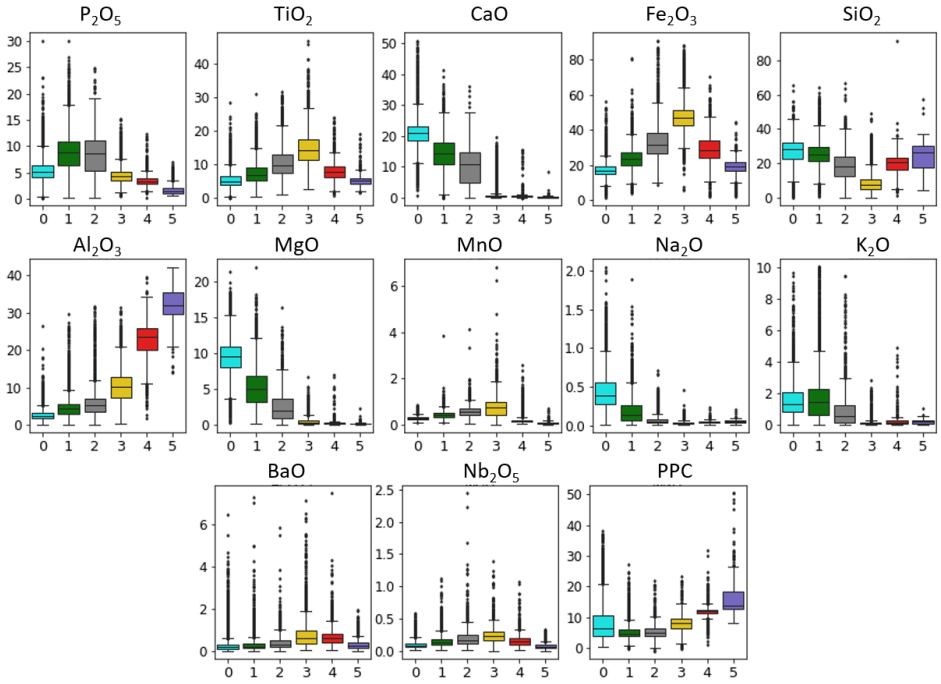

Figura 50 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com seis agrupamentos por *dsclus*.

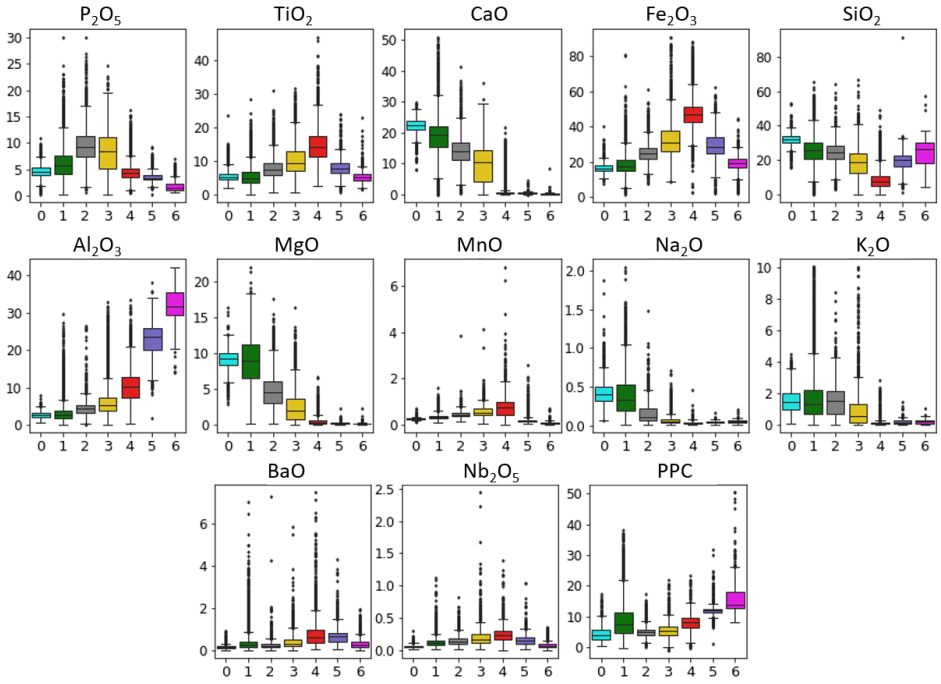

Figura 51 – *Boxplots* dos teores dos óxidos e perda por calcinação (PPC) por grupo no cenário com sete agrupamentos por *dsclus*.

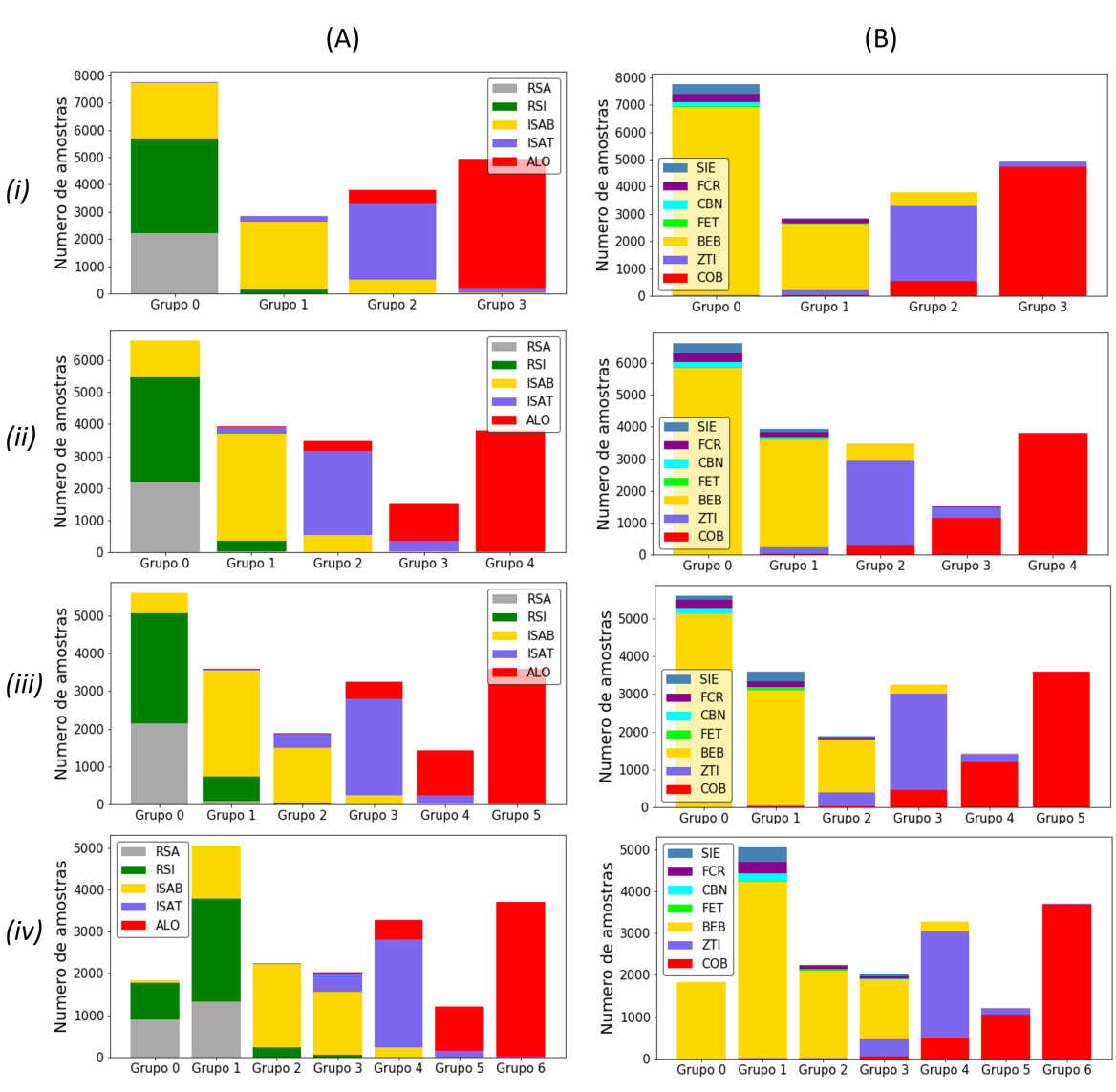

Figura 52 – Gráficos de barras mostrando os quantitativos de amostras por intemperismo (A) e litologia (B) em cada cenário de agrupamento pelo algoritmo *dsclus*: quatro (*i*), cinco (*ii*), seis (*iii*) e sete (*iv*) grupos.

Cenário	Grupo (Domínio)	Descrição	Teores médios			
			P_2O_5	TiO_2	CaO	CaO/P_2O_5
4 grupos (domínios)	0	Abrange materiais rochosos pouco ou nada alterados. Localmente, pode conter teores interessantes de P2O5, constituindo minério de fosfato. Presença expressiva de carbonatos (CaO e MgO), Na2O e K2O, esse último devido aos sienitos.	6,19	5,77	19,08	3,08
	1	Principal fonte de minério de fosfato, podendo conter, localmente, teores expressivos de TiO2, Fe2O3, CaO e MgO.	9,47	9,11	12,66	1,34
	2	Principal fonte de minério de titânio, podendo conter teores expressivos de Fe_2O_3, Al_2O_3, MnO e BaO.	4,85	14,49	1,94	0,4
	3	Primeiro horizonte intempérico, correspondendo a solos, argilas e outros materiais estéreis. Caracterizado principalmente por altos teores de Al_2O_3 e PPC.	2,24	5,99	0,38	0,17
5 grupos (domínios)	0	Abrange materiais rochosos pouco ou nada alterados, em geral estéreis em fosfato ou titânio. Localmente, pode conter teores interessantes de P_2O_5, mas também quantidades expressivas de CaO, MgO (carbonatos), Na_2O e K_2O (sienitos).	5,78	5,48	19,94	3,45
	1	Principal fonte de minério de fosfato, podendo conter, localmente, teores expressivos de TiO_2, Fe_2O_3, Na_2O, K_2O, CaO e MgO.	9,21	8,61	13,12	1,42
	2	Principal fonte de minério de titânio, podendo conter teores expressivos de P2O5, Fe2O3, MnO e BaO.	5,00	14,85	2,17	0,43
	3	Inclui, principalmente, materiais de cobertura, com teores altos de Al_2O_3, PPC e Fe_2O_3, mas pode conter teores interessantes de TiO_2.	3,80	8,77	0,77	0,2
	4	Primeiro horizonte intempérico, correspondendo a solos, argilas e outros materiais estéreis. Caracterizado principalmente por altos teores de Al2O3 e PPC.	1,81	5,36	0,29	0,16
6 grupos (domínios)	0	Abrange materiais rochosos pouco ou nada alterados, em geral estéreis em fosfato ou titânio. Localmente, pode conter teores interessantes de P_2O_5, mas também quantidades expressivas de CaO, MgO (carbonatos), Na_2O e K_2O (sienitos).	5,44	5,41	20,88	3,84
	1	Principal fonte de minério de fosfato, contendo, no entanto, teores expressivos de CaO, MgO, Na2O e K2O devido a ocorrências locais de carbonatos e sienitos.	8,85	7,33	14,29	1,61
	2	Pode constituir fonte de minério de fosfato, contendo, também, teores razoáveis de TiO2, Fe2O3 e, localmente, carbonatos.	8,49	10,56	10,04	1,18
	3	Principal fonte de minério de titânio, contendo, também, teores expressivos de Fe_2O_3, MnO e BaO.	4,56	14,91	1,26	0,28
	4	Inclui, principalmente, materiais de cobertura, com altos teores de Al_2O_3, PPC e Fe_2O_3, mas pode conter teores interessantes de TiO_2.	3,61	8,04	0,67	0,19
	5	Primeiro horizonte intempérico, correspondendo a solos, argilas e outros materiais estéreis. Caracterizado principalmente pelos altos teores de Al_2O_3 e PPC.	1,74	5,29	0,28	0,16
7 grupos (domínios)	0	Abrange materiais rochosos pouco ou nada alterados, em geral estéreis em fosfato ou titânio. Muito localmente, pode conter teores interessantes de P_2O_5, mas também quantidades expressivas de carbonatos (CaO e MgO).	4,65	5,66	22,10	4,75
	1	Abrange materiais rochosos pouco ou nada alterados, em geral estéreis em fosfato ou titânio, caracterizados por quantidades expressivas de carbonatos e teores relativamente altos de Na_2O e K_2O (sienitos). Contém, também, materiais fosfatados de baixo teor.	6,24	5,48	18,80	3,01
	2	Principal fonte de minério de fosfato, podendo, no entanto, conter quantidades locais expressivas de CaO e MgO devido à presença de carbonatos.	9,64	8,08	14,26	1,48
	3	Pode ser fonte tanto de minério de fosfato quanto de titânio, mas contendo teores significativos de CaO, MgO, Fe_2O_3 e SiO_2.	8,34	10,35	9,75	1,17
	4	Principal fonte de minério de titânio, contendo, também, teores expressivos de Fe_2O_3, MnO e BaO.	4,58	14,86	1,29	0,28
	5	Inclui, principalmente, materiais de cobertura, com altos teores de Al_2O_3, PPC e Fe_2O_3.	3,55	8,18	0,55	0,15
	6	Primeiro horizonte intempérico, correspondendo a solos, argilas e outros materiais estéreis. Caracterizado principalmente pelos altos teores de Al_2O_3 e PPC.	1,77	5,31	0,29	0,16

Figura 53 – Quadros com a caracterização de cada um dos grupos, em cada um dos cenários, caso utilizados para constituir domínios para modelagem.

3.6 Discussão dos resultados da análise de agrupamento

Uma importante conclusão a que se pode chegar com este estudo de caso é que pode não existir uma única resposta ao se aplicar a análise de agrupamento à modelagem de recursos minerais, sendo, neste caso específico, os resultados do algoritmo *dsclus* para quatro, cinco, seis ou sete grupos igualmente apropriados, pelo menos à princípio. A escolha final deve depender de aspectos operacionais, principalmente no que diz respeito aos teores utilizados para a definição de minérios e seus contaminantes, levando sempre em consideração que, quanto menos grupos, menos laboriosas serão as etapas subsequentes da modelagem: definição de contornos, estimativa e/ou simulação geoestatística.

Em todos os cenários obtidos, pelo menos dois domínios correspondem a estéril franco, um deles equivalente a materiais de cobertura e o outro, a rochas pouco ou nada alteradas na base do depósito. A questão a ser resolvida é em quantos domínios se dividir os materiais tratados como minérios: dois, três, quatro ou cinco, tendo em consideração duas questões: (*i*) máximo aproveitamento dos materiais lavrados; (*ii*) mínima complexidade no processo de modelagem subsequente.

Com relação ao modelo praticado na mina, de maneira simplificada, há um domínio para a cobertura (ALO), um para o minério de titânio (ISAT), um para o fosfato (ISAB), e dois para os materiais rochosos. RSI e RSA, apesar de estatisticamente semelhantes em termos globais, são estimados separadamente para que a estimativa dos blocos de RSI não seja contaminada por teores de RSA, principalmente no que diz respeito a baixos valores de P_2O_5 e altos teores de CaO. Isso é praticado pois, localmente, a RSI pode ser aproveitada como minério de fosfato, dependendo dos teores de P_2O_5 e da razão CaO/P_2O_5. No entanto, na perspectiva global da análise de agrupamento, não há indicativos de que RSI e RSA devam constituir domínios distintos.

O cenário de agrupamento definido pelo algoritmo *dsclus* mais semelhante, mesmo que grosseiramente, à classificação praticada na mina é aquele com quatro domínios, porém, com RSI e RSA englobadas no Grupo 1, que inclui, também, uma considerável proporção de ISAB. Nesse caso, o Grupo 0 corresponderia ao minério de fosfato, contendo, também, partes de RSI e de ISAT. O grupo 2 corresponderia ao minério de titânio, enquanto o Grupo 4, à cobertura estéril.

Entretanto, a divisão em quatro domínios implicaria em perdas consideráveis de materiais que poderiam ser lavrados como minérios (englobados no Grupo 1). Por outro lado, o cenário com sete grupos, além de incluir parte considerável de amostras com alto teor de P_2O_5 no Grupo 1, estéril, propõe a divisão das coberturas em dois grupos (5 e 6), o que parece desnecessário, já que parte dessas coberturas com teores particularmente interessantes de TiO_2 já estão incluídas no Grupo 4, que constituiria o minério de titânio.

Assim, considerando os diversos fatores apresentados, os cenários com cinco ou

seis grupos parecem ser mais adequados. No primeiro caso, o Grupo 0 seria formado por materiais rochosos estéreis e materiais fosfatados com alto teor de carbonatos, o Grupo 1 constituiria o minério de fosfato, o Grupo 4, o minério de titânio, assim como parte do Grupo 3 e, o Grupo 2, o estéril de cobertura.

Com relação à configuração com seis grupos, os Grupos 1 e 2 constituiriam fontes de minério de fosfato, sendo que o último "aproveita" alguma proporção de RSI, apesar de uma pequena parcela de ISAB (relativamente pobre em P_2O_5) ser "perdida" no Grupo 0, constituído pelos materiais rochosos da base do depósito. Nesse caso, o Grupo 3 constituiria minério de titânio, assim como parte do Grupo 4, enquanto o Grupo 5, materiais de cobertura estéreis. Um ponto negativo dessa configuração é que o Grupo 2, considerado minério, acaba por incorporar amostras classificadas como sienitos, o que eleva substancialmente os teores globais de contaminantes, principalmente K_2O.

Conforme já mencionado, as categorias geológicas, tanto litologias quanto intemperismo, são referências essenciais. No entanto, por serem oriundas de classificações feitas de maneira subjetiva na fase de aquisição de dados, não devem ser tomadas estritamente como rótulos inquestionáveis, mesmo que feitas de maneira eficiente.

É importante salientar que o conhecimento geológico sobre o depósito é fundamental, e uma classificação geológica prévia é importante para que se tome o cuidado de que materiais particularmente desfavoráveis não sejam englobados em domínios de minérios na análise de agrupamento. Isso pode ocorrer caso esses materiais não correspondam a uma parte expressiva do banco de dados, a ponto de integrarem domínios próprios. Assim, uma abordagem híbrida, que integre classificação geológica e análise de agrupamento, parece ser mais interessante.

As diferentes técnicas de validação dos agrupamentos no espaço multivariado resultam em índices que, por definição, favorecem casos nos quais os contornos são mais arredondados. No caso das ciências da Terra, cujos dados apresentam relações e contornos mais complexos, essas métricas devem ser usadas com cuidado. O mais adequado é que várias delas sejam aplicadas simultaneamente, a fim de se comparar resultados, para que as conclusões sejam mais assertivas.

3.7 Aplicação de um classificador supervisionado para a inclusão de novas amostras

Sendo o cenário com cinco grupos mais objetivo e com maior representatividade amostral por grupo, foi ele o escolhido para o teste de classificação automática de amostras. Fez-se a opção pelo algoritmo de florestas aleatórias, um dos mais utilizados em aplicações de aprendizado de máquina, frequentemente indicado para tarefas de classificação de dados,

tendo sido utilizado o código disponível na biblioteca *Scikit-learn* (PEDREGOSA et al., 2011).

Primeiramente, os 19.344 dados agrupados foram separados em dois subconjuntos estratificados (nos quais são mantidas as proporções de cada grupo): um para o treinamento (85% dos dados totais, ou seja, 16.442 amostras) e outro para validação final (15% dos dados totais, o que corresponde a 2.902 amostras). Este último subconjunto servirá para simular uma situação em que novas amostras fossem incorporadas ao banco de dados e desejássemos atribuí-las aos grupos definidos na análise de agrupamento. Desta maneira, estando os verdadeiros rótulos disponíveis, é possível avaliar o desempenho do classificador.

Como variáveis de entrada, além de todos os atributos contínuos do sistema (óxidos e PPC), foram consideradas as coordenadas geográficas das amostras, já que a posição relativa das mesmas possui íntima relação com a configuração dos grupos, conforme já discutido. Por meio de testes, constatou-se que a inclusão das coordenadas geográficas como variáveis de entrada melhorou consideravelmente o desempenho do classificador.

As variáveis de entrada foram padronizadas, conforme a Equação 3.1, um procedimento comum ao se aplicar o aprendizado de máquina, para que diferenças entre as escalas de observação de cada atributo não interfiram nos resultados.

Para a definição de parâmetros do algoritmo, foi aplicado o método denominado "RandomizedSearchCV" (PEDREGOSA et al., 2011), que busca os melhores parâmetros através de uma série de tentativas por validação cruzada, com diferentes combinações desses parâmetros, sendo, então, possível determinar aqueles que resultam na maior acurácia do modelo.

Em uma primeira etapa de avaliação do modelo, foi executada a validação cruzada por *k-folds* estratificados nos 16.442 dados de treino, com cinco *folds*. O desempenho do treinamento em cada *fold* pode ser observado pelas métricas globais no quadro da Figura 54 e pelas matrizes de confusão da Figura 55.

	Grupo	Precisão	*Recall*	*F1-score*	Suporte
Fold 1	Grupo 0	0.93	0.96	0.94	1132
	Grupo 1	0.85	0.83	0.84	659
	Grupo 2	0.90	0.87	0.88	599
	Grupo 3	0.78	0.78	0.78	255
	Grupo 4	0.97	0.97	0.97	645
Fold 2	Grupo 0	0.94	0.96	0.95	1132
	Grupo 1	0.86	0.87	0.86	659
	Grupo 2	0.92	0.90	0.91	599
	Grupo 3	0.82	0.82	0.82	255
	Grupo 4	0.97	0.96	0.96	645
Fold 3	Grupo 0	0.96	0.95	0.95	1132
	Grupo 1	0.84	0.89	0.86	659
	Grupo 2	0.91	0.88	0.89	599
	Grupo 3	0.85	0.80	0.82	255
	Grupo 4	0.98	0.98	0.98	645
Fold 4	Grupo 0	0.94	0.94	0.94	1132
	Grupo 1	0.85	0.86	0.85	659
	Grupo 2	0.91	0.92	0.91	599
	Grupo 3	0.87	0.77	0.82	255
	Grupo 4	0.95	0.98	0.96	645
Fold 5	Grupo 0	0.95	0.97	0.96	1132
	Grupo 1	0.88	0.87	0.87	659
	Grupo 2	0.90	0.89	0.89	599
	Grupo 3	0.81	0.76	0.78	255
	Grupo 4	0.96	0.97	0.96	645

Figura 54 – Métricas globais para a validação do classificador para cada *fold*.

Fold 1

Grupo real \ Grupo predito	0	1	2	3	4
0	1085	47	0	0	0
1	82	547	26	4	0
2	1	42	524	20	0
3	0	4	34	199	18
4	0	0	1	32	624

Fold 2

Grupo real \ Grupo predito	0	1	2	3	4
0	1084	48	0	0	0
1	66	573	18	2	0
2	0	39	541	25	0
3	0	3	26	208	18
4	0	0	1	18	619

Fold 3

Grupo real \ Grupo predito	0	1	2	3	4
0	1078	53	1	0	0
1	50	588	19	1	0
2	0	52	528	16	0
3	0	7	35	203	10
4	0	0	0	19	628

Fold 4

Grupo real \ Grupo predito	0	1	2	3	4
0	1068	63	0	0	0
1	62	564	29	3	0
2	1	34	552	14	0
3	0	2	25	197	31
4	0	0	1	12	629

Fold 5

Grupo real \ Grupo predito	0	1	2	3	4
0	1093	38	0	0	0
1	55	574	28	1	0
2	0	40	534	20	0
3	0	3	30	194	28
4	0	0	0	25	624

Figura 55 – Matrizes de confusão para a validação do classificador para cada *fold*.

Como pode ser observado, o classificador apresenta resultados aceitáveis, com as diagonais principais das matrizes de confusão em destaque e uma acurácia total média de 92%. As afinidades estatísticas e geográficas entre o Grupo 1 e os Grupos 0 e 4 e as semelhanças entre o Grupo 3 e os Grupos 2 e 4 levaram a algumas classificações incorretas, mas toleráveis.

Por fim, o classificador foi aplicado aos 2.902 dados reservados inicialmente para validação final e, como pode-se notar na Figura 56, os resultados estão em conformidade

com as métricas e matrizes de confusão resultantes na fase de validação cruzada por *k-folds*, demonstrando que não ocorre *under* ou *overfitting* do modelo.

Ainda assim, uma validação humana final é desejável, com a análise visual e de ferramentas estatísticas como histogramas, *boxplots* e diagramas de dispersão. As seções verticais e *boxplots* das Figuras 57 e 58 evidenciam que o método foi aplicado adequadamente e que os resultados são coerentes com o que se deveria esperar (ver Figuras 47(B) e 49).

Domínio	Precisão	Recall	F1-score	Suporte
0	0.95	0.96	0.95	956
1	0.89	0.89	0.89	637
2	0.91	0.90	0.91	484
3	0.85	0.82	0.83	232
4	0.97	0.97	0.97	593
		Acurácia 0.93		

Grupo real \ Grupo predito	0	1	2	3	4
0	918	38	0	0	0
1	49	565	22	1	0
2	0	29	440	18	0
3	0	4	21	190	17
4	0	0	1	15	574

Figura 56 – Métricas e matriz de confusão para a validação final do classificador com os 2.902 dados de teste.

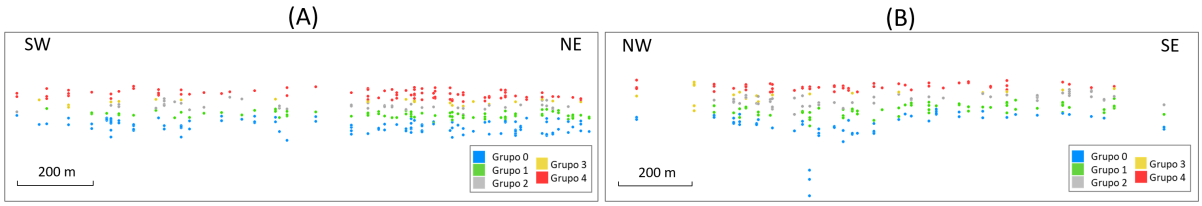

Figura 57 – Seções verticais mostrando parte dos 2.902 dados de teste, simbolizados com as categorias atribuídas pelo classificador automático.

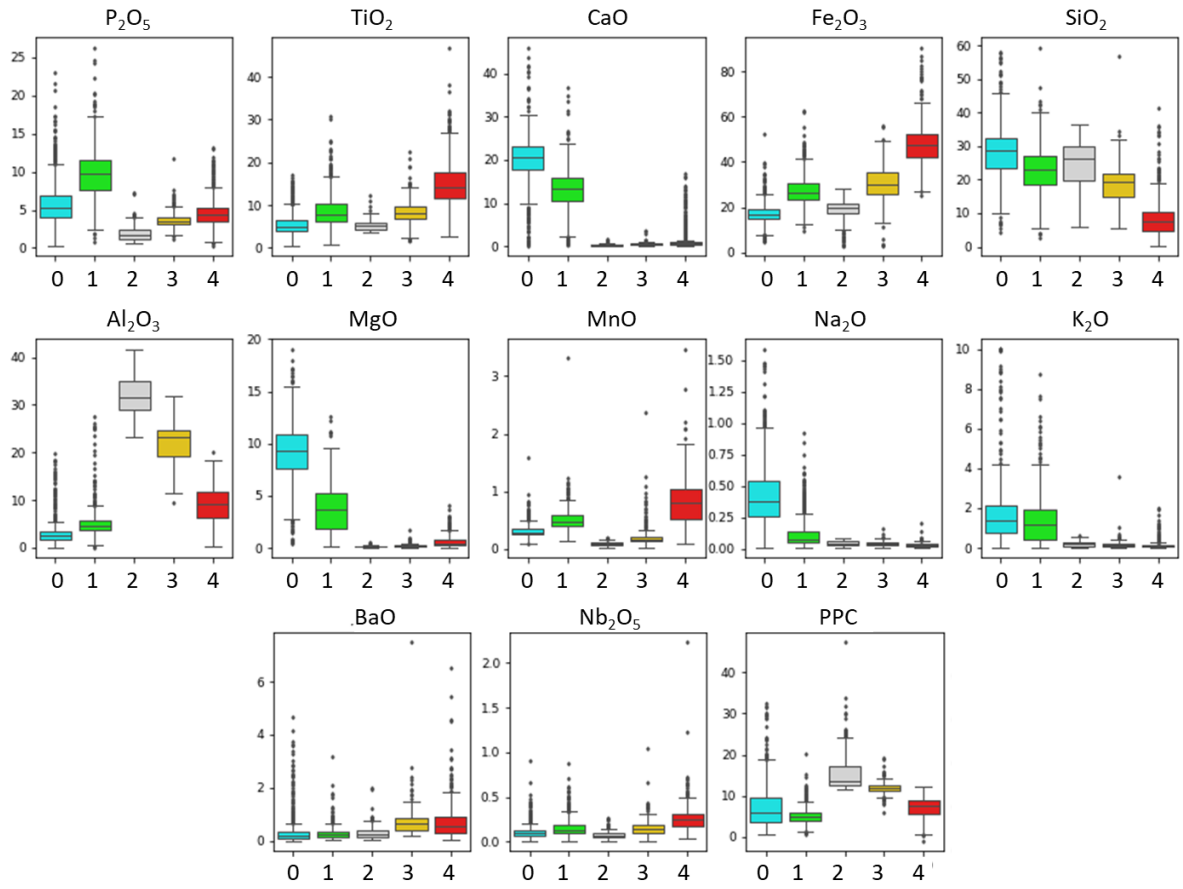

Figura 58 – *Boxplots* mostrando a distribuição estatística dos 2.902 dados de teste, após a classificação com o modelo obtido com o algoritmo de florestas aleatórias.

A tendência é que a continuidade da amostragem e a integração de novas amostras ao banco de dados aumente a representatividade dos grupos, melhorando os resultados, principalmente de grupos com menor representatividade amostral. Existem outras técnicas, específicas para se lidar com conjuntos de dados desequilibrados ("imbalanced datasets") (SUN et al., 2009; CHAWLA, 2010), nos quais as proporções entre as classes são consideravelmente distintas. No entanto, não serão abordadas aqui, já que os resultados obtidos são considerados satisfatórios para o fim aqui proposto, que é simplesmente testar a possibilidade de se usar classificadores supervisionados para a inclusão de novas amostras ao banco de dados, após a análise de agrupamento.

4 Considerações Finais

4.1 Conclusões

A aplicação de algoritmos tradicionais de agrupamento de dados, apesar de eficiente em muitos casos, se mostra bastante limitada na modelagem de recursos minerais, já que consideram apenas as relações dos dados no espaço multivariado, negligenciando sua distribuição no espaço geográfico. Assim, técnicas que também consideram a distribuição espacial das amostras são mais adequadas, conforme demonstrado no estudo de caso apresentado no Capítulo 3.

Embora esses algoritmos de agrupamento, mesmo os mais recentes, possam dividir os dados em grupos consistentes, tanto no espaço geográfico quanto no multivariado, a validação dos cenários ainda é complexa e subjetiva. Por ser uma técnica não supervisionada, não há valores "verdadeiros" para referência, e a parametrização ainda depende do usuário, que geralmente precisa definir alguns parâmetros *a priori*, como o número de grupos, as métricas para se medir distâncias, a maneira pela qual a vizinhança espacial correlacionada é determinada, entre outros.

Em geral, as características geológicas de um depósito mineral devem orientar na classificação das amostras. No entanto, tradicionalmente, essas características acabam levando a classificações subjetivas, decorrentes da interpretação do profissional responsável pela aquisição de dados. Tais classificações não devem, portanto, ser dadas como verdades inquestionáveis, e diferenças devem surgir quando comparadas aos resultados de algoritmos de agrupamento. Ainda assim, é evidente que as características geológicas devem ser, de alguma forma, refletidas nos agrupamentos, sendo fundamental o conhecimento do contexto geológico no processo de tomada de decisão quanto à seleção da configuração mais adequada de agrupamentos.

Quando as amostras são bem classificadas, especialmente em casos onde as tipologias possuem forte correlação com os teores, essas classificações "manuais" podem ser bem eficientes, como no caso apresentado no Capítulo 3. Entretanto, podem se tornar problemáticas em situações mais complexas, em que litologias e padrões de alteração se sobrepõem (e.g. intemperismo, metassomatismo, hidrotermalismo).

Há diversas técnicas disponíveis na literatura, tanto para agrupamento quanto para avaliação de resultados, cabendo ao usuário escolher cuidadosamente aquelas que forem mais adequadas ao seu caso, de acordo com os dados e os resultados almejados. Um determinado algoritmo, ideal para um dado caso, pode se mostrar inadequado em outra situação.

Uma vez definido o melhor cenário, os códigos de cada grupo podem ser alimentados como rótulos (*labels*) em algoritmos de aprendizado supervisionado (e.g. árvores de decisão, florestas aleatórias, k-vizinhos mais próximos) para a calibração de modelos matemáticos para classificação automática e inclusão de novas amostras ao banco de dados. Periodicamente, a análise de agrupamento deve ser revisitada com a totalidade das amostras e o classificador supervisionado, atualizado.

Uma das grandes vantagens da aplicação do aprendizado de máquina a bancos de dados de mineração é a capacidade de se fazer comparações objetivas no espaço multidimensional, usando relações matemáticas, uma condição praticamente impossível para um analista humano. Essa automatização proporciona um aumento significativo da reprodutibilidade no processo de modelagem, uma qualidade essencial na avaliação de recursos minerais, principalmente para fins de auditoria.

No entanto, apesar de muito eficazes no processo de tomada de decisão, os métodos apresentados não são, ainda, totalmente automatizados, exigindo conhecimento especializado e muito bom senso.

4.2 Sugestões para Trabalhos Futuros

Dadas a pertinência e a vastidão do tema, bem como sua complexidade e nuances, acredita-se que desdobramentos deste trabalho podem render estudos interessantes. Assim, é sugerido:

(i) Que sejam melhor exploradas as aplicações de classificadores supervisionados para a inclusão de novas amostras aos bancos de dados após a análise de agrupamento;

(ii) Que seja realizada uma validação adicional do classificador automático, compondo um subconjunto de dados de teste com furos específicos, a fim de simular uma situação em que os dados de uma nova campanha de sondagem são adicionados à base de dados;

(iii) Que sejam exploradas as implicações da integração de outras técnicas multivariadas à análise de agrupamento. A chamada "multidimensional scaling" (*MDS*), por exemplo, pode ser aplicada a fim de se analisar melhor as relações dos dados no espaço multivariado. Uma outra técnica muito relevante, a chamada análise das componentes principais (*PCA*), pode ser empregada para se realizar a redução da dimensionalidade em casos com muitos atributos, além de fornecer informações sobre a variabilidade de cada uma das variáveis, o que pode ter implicações importantes na análise de agrupamento;

(iv) Que sejam investigados outros métodos de transformação de dados na etapa pré-agrupamento, por exemplo, a Gaussianização;

(v) Que o fluxo de trabalho proposto seja aplicado a depósitos de outras substâncias minerais e situações ainda mais complexas, como na modelagem de cobre e ouro, metais cujo comportamento é mais errático;

(vi) Que seja conduzido um estudo considerando variáveis metalúrgicas, a fim de se construir modelos geometalúrgicos;

(vii) Que o fluxo de trabalho seja operacionalizado em *softwares* de ampla utilização na indústria, o que pode tornar sua aplicação mais prática;

(viii) Que se investigue mais a fundo as implicações da aplicação de diferentes algoritmos de agrupamento e técnicas de validação;

(ix) Sendo a análise de agrupamento uma técnica baseada em cenários, que se realize um estudo a fim de acessar as incertezas associadas a múltiplos cenários;

(x) Que seja conduzido um processo completo de modelagem de recursos minerais, incluindo a definição de contornos e a estimativa e/ou simulação geoestatística, analisando as implicações no cálculo de recursos e reservas de minérios.

Referências

AMBROISE, C.; DANG, M.; GOVAERT, G. Clustering of spatial data by the EM algorithm. In: GEOENV I — GEOSTATISTICS FOR ENVIRONMENTAL APPLICATIONS. *Proceedings...* [S.l.], 1997. p. 493–504. Citado 3 vezes nas páginas 30, 40 e 41.

ARTHUR, D.; VASSILVITSKII, S. K-means++: the advantages of careful seeding. In: EIGHTEENTH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS. *Proceedings...* [S.l.], 2007. p. 1027–1035. Citado 2 vezes nas páginas 37 e 73.

BARNETT, R. M.; DEUTSCH, C. V. Conventional clustering algorithms and a program for their application. *CCG Annual Report*, University of Alberta, v. 17, 2015. Citado 4 vezes nas páginas 9, 34, 35 e 37.

BROD, J. A. *Petrology and geochemistry of the Tapira alkaline complex, Minas Gerais state, Brazil*. Tese (Doutorado) — Durham University, 1999. Citado 2 vezes nas páginas 59 e 61.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics – Theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado 2 vezes nas páginas 49 e 74.

CAPPONI, L. N. *Introdução de parâmetros de controle de incertezas para planejamento de lavra*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2012. Citado na página 62.

CHAWLA, N. V. Data mining for imbalanced datasets: an overview. In: MAIMON, O.; ROKACH, L. (Ed.). *Data mining and knowledge discovery handbook*. [S.l.]: Springer, 2010. cap. 45, p. 875–886. Citado na página 91.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, n. 2, p. 224–227, 1979. Citado 3 vezes nas páginas 20, 48 e 74.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977. Citado na página 38.

EMERY, X.; ORTIZ, J. Estimation of mineral resources using grade domains: critical analysis and a suggested methodology. *Journal of the Southern African Institute of Mining and Metallurgy*, Sabinet, v. 105, n. 4, p. 247–255, 2005. Citado na página 29.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING. *Proceedings...* [S.l.], 1996. v. 96, n. 34, p. 226–231. Citado 2 vezes nas páginas 32 e 38.

FABER, V. Clustering and the continuous k-means algorithm. *Los Alamos Science*, n. 22, p. 67, 1994. Citado 2 vezes nas páginas 9 e 34.

FOUEDJIO, F. A hierarchical clustering method for multivariate geostatistical data. *Spatial Statistics*, Elsevier, v. 18, p. 333–351, 2016. Citado 2 vezes nas páginas 40 e 41.

FOUEDJIO, F.; HILL, E. J.; LAUKAMP, C. Geostatistical clustering as an aid for ore body domaining: case study at the Rocklea Dome channel iron ore deposit, Western Australia. *Applied Earth Science*, Taylor & Francis, v. 127, n. 1, p. 15–29, 2018. Citado na página 40.

FUCK, R. A. et al. As faixas de dobramento marginais do cráton do São Francisco: síntese dos conhecimentos. In: DOMINGUEZ, J. M. L.; MISI, A. (Ed.). *O Cráton do São Francisco*. [S.l.], 1993. p. 161–185. Citado na página 59.

GETIS, A.; ORD, J. K. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, p. 189–206, 1992. Citado na página 43.

GOOVAERTS, P. *Geostatistics for natural resources evaluation*. [S.l.]: Oxford University Press, 1997. Citado 3 vezes nas páginas 11, 53 e 54.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining geostatistics*. [S.l.]: Academic Press, 1978. 600 p. Citado 4 vezes nas páginas 19, 24, 25 e 26.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. New Jersey: John Wiley & Sons, 2005. 344 p. Citado na página 45.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: FIFTH BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY. *Proceedings...* Oakland, CA, USA, 1967. v. 1, n. 14, p. 281–297. Citado 5 vezes nas páginas 20, 30, 32, 36 e 73.

MANITA, G.; KHANCHEL, R.; LIMAM, M. Consensus functions for cluster ensembles. *Applied Artificial Intelligence*, Taylor & Francis, v. 26, n. 6, p. 598–614, 2012. Citado na página 42.

MARIZ, C. et al. Geostatistical clustering, potential modeling and traditional geostatistics: a new coherent workflow. In: SIXTH INTERNATIONAL CONFERENCE ON INNOVATION IN MINE OPERATIONS. *Proceedings...* Santiago, Chile, 2016. v. 8. Citado na página 41.

MARTIN, R. *Data driven decisions of stationarity for improved numerical modeling in geological environments*. Tese (Doutorado) — University of Alberta, 2019. Citado 3 vezes nas páginas 9, 20 e 28.

MARTIN, R.; BOISVERT, J. A random-path spatial clustering algorithm. *CCG Annual Report*, University of Alberta, v. 19, 2017. Citado na página 19.

MARTIN, R.; BOISVERT, J. Towards justifying unsupervised stationary decisions for geostatistical modeling: ensemble spatial and multivariate clustering with geomodeling specific clustering metrics. *Computers & Geosciences*, Elsevier, v. 120, p. 82–96, 2018. Citado 16 vezes nas páginas 10, 11, 19, 20, 30, 32, 40, 41, 42, 43, 50, 51, 52, 73, 74 e 76.

MATHERON, G. Principles of geostatistics. *Economic Geology*, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963. Citado 4 vezes nas páginas 11, 24, 53 e 54.

MCLENNAN, J. A. *The decision of stationarity*. Tese (Doutorado) — University of Alberta, 2007. Citado 2 vezes nas páginas 27 e 28.

MODENA, R. C. C. et al. Avaliação de técnicas de agrupamento para definição de domínios estacionários com o auxílio de geoestatística. In: FIFTH ABM WEEK. *Proceedings...* [S.l.], 2019. v. 20, p. 91–100. Citado na página 53.

OLIVER, M.; WEBSTER, R. A geostatistical basis for spatial weighting in multivariate classification. *Mathematical Geology*, Springer, v. 21, n. 1, p. 15–35, 1989. Citado 4 vezes nas páginas 30, 40, 41 e 51.

ORD, J. K.; GETIS, A. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, Wiley Online Library, v. 27, n. 4, p. 286–306, 1995. Citado na página 43.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 9 vezes nas páginas 9, 10, 20, 33, 37, 40, 49, 73 e 88.

PRADES, C. *Geostatistics and clustering for geochemical data analysis*. Dissertação (Mestrado) — University of Alberta, 2017. Citado na página 74.

RIZZO, M. L.; SZÉKELY, G. J. Energy distance. *Computational Statistics*, Wiley Online Library, v. 8, n. 1, p. 27–38, 2016. Citado na página 42.

ROMARY, T. et al. Domaining by clustering multivariate geostatistical data. In: ABRAHAMSEN, P.; HAUGE, R.; KOLBJORNSEN, O. (Ed.). *Geostatistics Oslo 2012*. [S.l.]: Springer, 2012. p. 455–466. Citado 3 vezes nas páginas 19, 40 e 41.

ROSSI, M. E.; DEUTSCH, C. V. *Mineral resource estimation*. [S.l.]: Springer Science & Business Media, 2014. Citado 4 vezes nas páginas 19, 27, 29 e 64.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado 6 vezes nas páginas 10, 20, 46, 47, 48 e 74.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, IBM, v. 3, n. 3, p. 210–229, 1959. Citado na página 56.

SCHUBERT, E. et al. DBSCAN revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, ACM, v. 42, n. 3, p. 19, 2017. Citado 4 vezes nas páginas 10, 31, 32 e 40.

SCRUCCA, L. Clustering multivariate spatial data based on local measures of spatial autocorrelation. *Quaderni del Dipartimento di Economia, Finanza e Statistica*, Università di Perugia, v. 20, n. 1, p. 11, 2005. Citado 8 vezes nas páginas 19, 20, 30, 32, 41, 43, 44 e 73.

SILVA, C. H. D. et al. Proveniência e idade do metamorfismo das rochas da Faixa Brasília, na região de Tapira (SW de Minas Gerais). *Geologia USP. Série Científica*, v. 6, n. 1, p. 53–66, 2006. Citado 2 vezes nas páginas 11 e 60.

SILVA, C. H. da. *Evolução geológica da Faixa Brasília na região de Tapira, sudoeste de Minas Gerais*. Tese (Doutorado) — Instituto de Geociências e Ciências Exatas, Universidade Estadual Paulista., 2003. Citado na página 59.

SINCLAIR, A. J.; BLACKWELL, G. H. *Applied mineral inventory estimation*. [S.l.]: Cambridge University Press, 2004. Citado 3 vezes nas páginas 19, 25 e 64.

SOKAL, R. R.; SNEATH, P. H. A. *Principles of numerical taxonomy*. [S.l.]: W. H. Freeman, 1963. Citado 5 vezes nas páginas 20, 30, 32, 34 e 73.

STREHL, A.; GHOSH, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, n. 3 (Dec), p. 583–617, 2002. Citado na página 42.

SUN, Y.; WONG, A. K.; KAMEL, M. S. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, v. 23, n. 04, p. 687–719, 2009. Citado na página 91.

TAN, P.; MICHAEL, S.; KUMAR, V. *Introduction to data mining*. [S.l.]: New York: Pearson Education, 2006. Citado 7 vezes nas páginas 9, 10, 19, 30, 31, 36 e 37.

VANDERPLAS, J. *Python data science handbook: essential tools for working with data*. [S.l.]: O'Reilly Media, Inc., 2016. Citado 4 vezes nas páginas 10, 32, 37 e 38.

Apêndices

APÊNDICE A – Parâmetros de entrada dos algoritmos utilizados

A fim de possibilitar ao leitor reproduzir os resultados obtidos neste trabalho, são aqui apresentados os parâmetros de entrada dos algoritmos de aprendizado de máquina utilizados, conforme aplicados no estudo de caso do Capítulo 3.

A.1 Algoritmos de agrupamento de dados

A.1.1 *k-means*

- $init = kmeans++$
- $n\_init = 300$
- $algorithm = full$
- $random\_state = 1$

A.1.2 Aglomerativo hierárquico

- $affinity = euclidean$
- $linkage = ward$

A.1.3 Agrupamento em espaço duplo (*dsclus*)

- $nreal = 100$
- $n\_nears = 30$
- $numtake = 20$
- $searchparams = (0, 0, 0, 400, 400, 12)$
- $method = hier$ (método para extração da configuração final)

A.1.4 Agrupamento por estatísticas de autocorrelação (*acclus*)

- $acmetric = getis$
- $cluster\_method = kmeans$
- $nnears = 30$
- $searchparams = (0, 0, 0, 400, 400, 12)$

A.2 Classificação por Florestas Aleatórias

- $n\_estimators = 700$
- $min\_samples\_leaf = 1$
- $max\_features = 13$
- $max\_depth = 20$
- $criterion = entropy$