Modelagem de Variáveis Geometalúrgicas: Uma Abordagem Comparativa entre Modelos Globais e Segmentados por Cluster

Fernando Vilela Brandão
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais Belo Horizonte, Minas Gerais
Email: fvb@ufmg.br

***Abstract*—This work evaluates predictive modeling approaches for key geometallurgical variables — metallurgical recovery (*rec*) and Bond Work Index (*BWI*) — using geochemical and mineralogical data from a mineral deposit. Three strategies are compared: a global model using the *XGBoost Regressor* trained on the entire dataset, and two segmented approaches with local models trained separately on clusters defined by *K-means* and spatially-constrained *Agglomerative Clustering*. The results show that the segmented models consistently outperform the global model, with substantial improvements in the coefficient of determination (R^2), as well as reductions in mean absolute error (MAE) and root mean square error (RMSE). The model based on *Agglomerative Clustering* achieved the best overall performance for both target variables, highlighting the contribution of spatial segmentation to capturing local patterns and enhancing predictive accuracy in geometallurgical modeling. These findings reinforce the importance of accounting for spatial heterogeneity in predictive modeling of mineral deposits.**

Geometallurgy, Predictive Modeling, XGBoost, Clustering, K-means, Agglomerative Clustering, Metallurgical Recovery, Bond Work Index, Spatial Segmentation, Machine Learning in Mining

I. Introdução

A previsão precisa de variáveis geometalúrgicas é um componente essencial no planejamento e na tomada de decisões em projetos de mineração. Em especial, a recuperação metalúrgica (REC) e o Bond Work Index (BWI) impactam diretamente a eficiência do processo de beneficiamento mineral e os custos operacionais [1], [2].

A recuperação metalúrgica representa a proporção do metal de interesse presente no minério bruto que pode ser efetivamente recuperada durante o processamento. Essa métrica é fundamental para estimar o retorno econômico de uma jazida, uma vez que perdas metalúrgicas podem comprometer significativamente a rentabilidade da operação. Além disso, podem ocorrer grandes perdas ao processar um material inadequado às condições de planta, seja por limitações no tempo de residência, escolha de reagentes ou outras variáveis operacionais. A variabilidade da recuperação está associada à composição mineralógica do minério, à presença de minerais de ganga (componentes sem valor do minério), às condições de processo e à natureza das associações entre os minerais valiosos e os minerais hospedeiros [3], [4].

Já o BWI é um indicador da moabilidade do minério, ou seja, da energia necessária para reduzir o tamanho das partículas até uma granulometria adequada para o beneficiamento. Esse índice, expresso em kWh/t, influencia diretamente o dimensionamento e o consumo energético dos moinhos utilizados nas etapas de cominuição (britagem e moagem). Materiais com altos valores de BWI exigem maior energia para serem moídos, impactando os custos energéticos da planta [2].

A estimativa antecipada dessas variáveis para blocos georreferenciados do modelo de recursos permite:

- Realizar simulações realistas de desempenho da planta de beneficiamento;
- Otimizar a mistura de minérios (blending) para maximizar a recuperação e reduzir custos energéticos;
- Apoiar decisões de investimento em tecnologias de processamento.

Neste contexto, a aplicação de técnicas de aprendizado de máquina para prever REC e BWI com base em dados geoquímicos, mineralógicos e geológicos representa uma estratégia promissora para integrar conhecimento técnico com tomada de decisões baseadas em dados [5].

Dessa forma, este trabalho tem como objetivos:

- Aplicar e comparar técnicas de clusterização multivariada em dados geometalúrgicos [4], [6];
- Avaliar a coerência estatística e espacial das segmentações obtidas [3], [7];
- Implementar modelos preditivos para *REC* e *BWI*, avaliando o impacto da segmentação geometalúrgica sobre a acurácia das previsões.

II. Referencial Teórico

A. Clusterização Multivariada

A análise de clusters é uma técnica de aprendizado não supervisionado amplamente utilizada para segmentar dados com base em suas similaridades internas. Um dos algoritmos mais tradicionais é o *K-means*, que busca minimizar a soma das distâncias quadradas intra-cluster, assumindo que os grupos formados são aproximadamente esféricos e de tamanho semelhante.

Apesar de sua eficiência, o *K-means* considera apenas o espaço estatístico das variáveis, desconsiderando a localização espacial dos dados. Em aplicações geocientíficas, essa

limitação pode comprometer a validade dos agrupamentos, gerando domínios artificialmente fragmentados e pouco condizentes com a geologia do depósito [4], [6].

Para contornar esse problema, diversas abordagens passaram a integrar informações espaciais ao processo de clusterização. Técnicas como o agrupamento hierárquico geoestatístico e os métodos com restrição de conectividade espacial buscam preservar a continuidade dos agrupamentos no espaço físico, produzindo segmentações mais coerentes com a realidade geológica [3], [6].

B. Geometalurgia e Variabilidade Espacial

Na mineração, a definição de domínios geológicos ou geometalúrgicos é uma das etapas mais críticas da modelagem de recursos minerais. Domínios mal definidos podem levar à mistura de populações geoquímicas distintas, comprometendo a acurácia das estimativas de teor e a eficácia das decisões de lavra e beneficiamento [1], [4].

A geometalurgia busca justamente integrar informações mineralógicas, geoquímicas e espaciais para aprimorar esse processo, fornecendo uma visão mais holística do depósito. Quando bem definidos, os domínios geometalúrgicos permitem estimativas mais robustas, favorecem estratégias de planejamento seletivo e viabilizam modelagens locais por regressão — prática cada vez mais adotada na indústria mineral [1], [2].

Nesse contexto, compreender a **estrutura espacial** das variáveis envolvidas é essencial. Propriedades como a recuperação metalúrgica (*REC*) e o índice de moabilidade (*BWI*) tendem a apresentar padrões de continuidade no espaço, refletindo os processos geológicos que originaram o depósito. A verificação dessa dependência espacial auxilia na escolha de técnicas mais adequadas de segmentação, como os métodos de clusterização com restrições geográficas.

Uma das métricas mais utilizadas para quantificar essa dependência é o **Índice de Moran** (I), que mede o grau de autocorrelação espacial de um atributo. Ele compara a similaridade entre valores observados em locais próximos com a variância global da variável. Sua fórmula geral é:

$$I = \frac{n}{W} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

onde:

- n é o número total de observações;
- x_i e x_j são os valores da variável nos locais i e j;
- $\bar{x}$ é a média da variável;
- w_{ij} é o peso que expressa a proximidade espacial entre os locais i e j (geralmente baseado na distância ou vizinhança);
- $W = \sum_i \sum_j w_{ij}$ é a soma total dos pesos.

Valores positivos de I indicam autocorrelação positiva (valores semelhantes próximos entre si), enquanto valores negativos sugerem dispersão espacial. Um valor próximo de zero indica ausência de estrutura espacial. A aplicação do Índice de Moran é, portanto, uma etapa importante para validar a pertinência do uso de métodos que levem em conta a localização geográfica no processo de segmentação geometalúrgica.

C. Métricas de Avaliação de Agrupamentos

A avaliação da qualidade dos agrupamentos obtidos por técnicas de clusterização pode ser realizada por diferentes métricas, que visam quantificar tanto a coerência estatística quanto a organização espacial dos grupos. Entre as mais empregadas está o *Silhouette Score*, que combina informações sobre a **coesão intra-cluster** e a **separação inter-cluster**.

Para cada ponto i, o valor do *Silhouette* $s(i)$ é definido como:

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

em que:

- $a(i)$ é a distância média entre o ponto i e os demais pontos do mesmo cluster (medida de coesão);
- $b(i)$ é a menor distância média entre o ponto i e os pontos pertencentes a outros clusters (medida de separação).

Os valores de $s(i)$ variam entre -1 e 1, sendo que:

- $s(i) \approx 1$ indica alocação adequada do ponto ao seu cluster;
- $s(i) \approx 0$ sugere sobreposição entre clusters vizinhos;
- $s(i) < 0$ indica possível alocação incorreta.

A média dos valores $s(i)$ para todo o conjunto de dados constitui o **Silhouette Score médio**, indicador global da qualidade da segmentação. No entanto, essa métrica apresenta limitações em cenários onde a continuidade espacial é relevante, pois assume implicitamente que os agrupamentos possuem forma convexa e estão bem separados no espaço vetorial [7].

Em aplicações geocientíficas, outras duas métricas complementam a análise da qualidade dos agrupamentos: o **Within-Cluster Sum of Squares (WCSS)** e a **entropia espacial (H)**.

- **Within-Cluster Sum of Squares (WCSS)**: esta métrica quantifica a variabilidade interna dos clusters no espaço das variáveis. Para um dado número de clusters k, ela é definida como:

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2,$$

  em que:
  - C_i representa o conjunto de pontos atribuídos ao cluster i;
  - μ_i é o vetor centróide do cluster i;
  - $\|x - \mu_i\|^2$ é a distância euclidiana quadrática entre o ponto x e o centróide do seu grupo.

  Valores baixos de WCSS indicam que os pontos estão bem concentrados ao redor de seus centróides, o que representa alta coesão estatística. No entanto, essa métrica não considera a distribuição espacial dos dados.
- **Entropia espacial (H)**: essa métrica avalia o grau de dispersão geográfica dos clusters, medindo o quão misturados os diferentes grupos estão ao longo do espaço. É

definida com base na distribuição de rótulos de cluster em uma vizinhança local (por exemplo, os k vizinhos mais próximos de cada ponto):

$$H = -\sum_{c=1}^{k} p_c \log(p_c),$$

em que:

- p_c é a proporção de vizinhos do ponto que pertencem ao cluster c;
- a soma percorre todos os clusters presentes na vizinhança local.

Esse valor é então calculado para todos os pontos e, posteriormente, uma média é tomada para representar a entropia espacial global do agrupamento.

Valores baixos de H indicam que, em geral, os vizinhos de um ponto pertencem ao mesmo cluster — ou seja, há continuidade espacial. Já valores mais altos refletem mistura entre clusters em regiões próximas, indicando fragmentação espacial e menor plausibilidade geológica.

Essas três métricas fornecem visões complementares sobre a qualidade dos agrupamentos. Enquanto o *Silhouette Score* e o WCSS avaliam a estrutura estatística no espaço multivariado, a entropia espacial foca na coerência geográfica dos domínios. Na prática, é comum observar um comportamento compensatório entre essas medidas: estratégias que maximizam a coesão estatística podem gerar agrupamentos espacialmente desconexos, ao passo que métodos que priorizam continuidade espacial tendem a apresentar maior variabilidade interna. O ideal, conforme proposto por Martin e Boisvert (2018) [8], é buscar configurações que alcancem um equilíbrio entre essas duas dimensões, assegurando tanto a robustez estatística quanto a plausibilidade geológica dos domínios definidos.

D. Métricas de Avaliação de Modelos Preditivos

A avaliação do desempenho dos modelos de regressão neste estudo foi realizada com base em três métricas amplamente utilizadas na literatura: o coeficiente de determinação (R^2), o erro quadrático médio (*Root Mean Squared Error – RMSE*) e o erro absoluto médio (*Mean Absolute Error – MAE*). Essas métricas fornecem perspectivas complementares sobre a acurácia e a qualidade das predições.

- **Coeficiente de Determinação (R^2)**: mede a proporção da variância da variável dependente que é explicada pelas variáveis independentes no modelo. É definido como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

onde y_i representa os valores reais, $\hat{y}_i$ os valores preditos pelo modelo, e $\bar{y}$ é a média dos valores observados. Valores próximos de 1 indicam forte capacidade explicativa, enquanto valores negativos apontam para um desempenho pior do que o modelo nulo (que prediz sempre a média).

- **Erro Quadrático Médio (RMSE)**: representa a raiz da média dos quadrados dos erros de predição, penalizando mais fortemente erros grandes. É definido por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Quanto menor o RMSE, mais próximas as previsões estão dos valores reais. É uma métrica sensível a outliers.

- **Erro Absoluto Médio (MAE)**: calcula a média dos módulos das diferenças entre os valores reais e preditos, sem penalizar quadraticamente os erros:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

O MAE fornece uma interpretação direta do erro médio, sendo menos influenciado por valores extremos em comparação ao RMSE.

A combinação dessas três métricas permite avaliar tanto a *precisão global* quanto a *robustez* dos modelos ajustados, oferecendo uma visão mais completa do seu desempenho preditivo em diferentes cenários.

E. Trabalhos Correlatos

A definição de domínios geometalúrgicos por meio de técnicas de clusterização tem sido amplamente explorada na literatura, com ênfase crescente na integração entre variáveis geoquímicas, mineralógicas e espaciais. Em particular, diversos autores têm destacado a importância de considerar simultaneamente a coerência estatística e a continuidade geológica na segmentação de depósitos minerais.

Moreira et al. (2021) [6] propuseram um fluxo de trabalho que combina *machine learning* e geoestatística para definição de domínios, ressaltando que abordagens puramente estatísticas, como o *k*-means tradicional, podem falhar em representar a realidade geológica quando não incorporam informações espaciais [6]. Essa crítica é reforçada por Boroh et al. (2021) [3], que demonstraram que modelos baseados em domínios geológicos e geoquímicos apresentam desempenho superior à segmentação global na estimativa de recursos minerais, especialmente no que se refere à redução de viés e variância em áreas heterogêneas [3].

Silveira (2022) [4] enfatiza o papel dos algoritmos de agrupamento na construção de domínios estacionários, argumentando que a interpretação geológica isolada pode ser insuficiente para garantir a representatividade estatística necessária à modelagem confiável [4]. Nessa mesma direção, Mohammadi et al. (2022) [9] investigaram o impacto da incorporação de medidas de distância espacial em algoritmos de clusterização e constataram melhorias substanciais na continuidade dos agrupamentos e na delimitação de domínios geoestatisticamente coerentes [9].

Avanços mais recentes apontam para abordagens híbridas. Abildin et al. (2023) [10] propuseram um modelo que combina algoritmos de aprendizado de máquina com métodos geoestatísticos para segmentar domínios litológicos, alcançando maior alinhamento com as estruturas geológicas observadas em campo [10]. Esse movimento é corroborado por Jung e

Choi (2021) [11], que, em uma revisão sistemática, mapearam o crescimento do uso de técnicas de *machine learning* em mineração, abrangendo desde a exploração até o fechamento de mina, com destaque para aplicações em modelagem preditiva e otimização de processos [11].

Esses trabalhos reforçam a relevância da clusterização como etapa crítica no fluxo de modelagem mineral e validam o uso de técnicas multivariadas e espacialmente conscientes, como as aplicadas neste estudo. A literatura evidencia que a simples aplicação de algoritmos clássicos, sem adaptações ao contexto geológico, pode comprometer a qualidade dos modelos preditivos e a interpretação dos domínios, justificando a adoção de métodos que integrem conhecimento espacial desde as etapas iniciais da modelagem. Além da definição de domínios geometalúrgicos, incorporar técnicas de regressão preditiva para estimar variáveis como recuperação metalúrgica e consumo energético. Ortiz et al. (2022) [5], em uma revisão abrangente, destacam que algoritmos baseados em *Gradient Boosting*, como o XGBoost, oferecem desempenho superior a métodos tradicionais em tarefas de predição de variáveis de processo. A robustez frente a outliers e a capacidade de modelar não linearidades complexas são apontadas como diferenciais importantes.

Abildin et al. (2023) [10] reforçam esse posicionamento ao demonstrarem que modelos treinados separadamente em domínios litológicos apresentam desempenho mais acurado do que abordagens globais, mesmo quando utilizados algoritmos de menor complexidade. Tal evidência sugere que a combinação entre segmentação geometalúrgica e regressão avançada, como o *Gradient Boosting*, pode ser uma estratégia eficaz na modelagem mineral.

Esses estudos corroboram a proposta adotada neste trabalho, que combina a definição de domínios geometalúrgicos via clusterização com a modelagem preditiva supervisionada, buscando avaliar se abordagens locais podem superar modelos globais em termos de acurácia e robustez estatística.

III. Metodologia

A metodologia adotada neste estudo foi estruturada com o objetivo de segmentar um depósito mineral em domínios geometalúrgicos internamente homogêneos, a partir de atributos geoquímicos, mineralógicos e espaciais, e posteriormente utilizar essa segmentação como base para a modelagem preditiva de variáveis metalúrgicas. O trabalho foi conduzido em duas etapas principais: (i) segmentação geometalúrgica por meio de técnicas de clusterização e (ii) regressão das variáveis de interesse *REC* e *BWI* com modelos globais e locais.

A sequência de etapas metodológicas foi a seguinte:

- **Caracterização estatística descritiva:** análise exploratória das variáveis contínuas por meio de boxplots e de medidas de tendência central e de dispersão, com o objetivo de compreender a distribuição dos atributos e identificar possíveis outliers e assimetrias relevantes.
- **Análise de correlação e estrutura espacial:** cálculo da matriz de correlação de Pearson para avaliar dependências lineares entre as variáveis, além da aplicação do Índice de Moran para quantificar a autocorrelação espacial das variáveis.
- **Padronização dos dados:** aplicação do *StandardScaler* para transformar todas as variáveis em uma escala com média zero e desvio padrão um, garantindo que nenhuma variável domine a análise de agrupamento devido à sua magnitude.
- **Clusterização multivariada:** aplicação de dois algoritmos de agrupamento não supervisionado:
  - *K-means*: técnica baseada na minimização da variância intra-cluster.
  - *Agglomerative Clustering* com restrição espacial: método hierárquico que incorpora conectividade geográfica por meio de grafos de vizinhança, promovendo agrupamentos mais contínuos no espaço físico.
- **Segmentação geometalúrgica:** utilização dos clusters obtidos para dividir o depósito em domínios estatística e espacialmente homogêneos, servindo como base para a regressão segmentada das variáveis metalúrgicas.
- **Modelagem preditiva global:** ajuste de modelos de regressão para *REC* e *BWI* utilizando o conjunto completo de dados, sem considerar a segmentação geometalúrgica. Essa abordagem visa capturar padrões gerais nas relações entre os atributos geoquímicos/mineralógicos e as variáveis de saída. O algoritmo utilizado foi o *XGBoost Regressor*, com otimização de hiperparâmetros via *Random Search*. Os dados foram divididos em 60% para treino, 20% para validação e 20% para teste, por meio de amostragem aleatória simples, sem estratificação. Cada metodologia (modelo global, segmentado por *K-means* e segmentado por *Agglomerative Clustering*) utilizou partições independentes, garantindo que não houvesse sobreposição entre os conjuntos de dados usados em cada experimento.
- **Modelagem preditiva local por cluster:** ajuste de modelos de regressão específicos dentro de cada domínio identificado pelos algoritmos de clusterização (*K-means* e *Agglomerative Clustering*). Para cada cluster, o conjunto de dados foi novamente particionado em 60% para treino, 20% para validação e 20% para teste, de forma independente, permitindo a criação de modelos especializados para contextos estatísticos e espaciais distintos. Na etapa de avaliação, as predições foram realizadas individualmente em cada subconjunto de teste e, em seguida, todos os resultados de *predict* foram concatenados. As métricas globais de desempenho (R^2, RMSE e MAE) foram então calculadas considerando o conjunto combinado de todas as previsões, permitindo uma comparação direta com o modelo global. É importante destacar que os conjuntos de teste utilizados nos modelos segmentados não são os mesmos do modelo global. Cada metodologia (global, K-means e Agglomerative) foi particionada de forma independente, o que implica que as métricas de desempenho foram calculadas sobre amostras distintas. Essa decisão

metodológica foi intencional, pois o objetivo central do estudo não é comparar valores absolutos de erro sob um mesmo conjunto de teste, mas avaliar o comportamento e a capacidade explicativa de cada abordagem em seus respectivos contextos. Assim, a comparação entre os modelos é interpretada em termos de tendência de desempenho — isto é, verificar se a segmentação local aumenta a acurácia relativa em relação à modelagem global — e não como uma comparação direta ponto a ponto sobre o mesmo conjunto de dados.

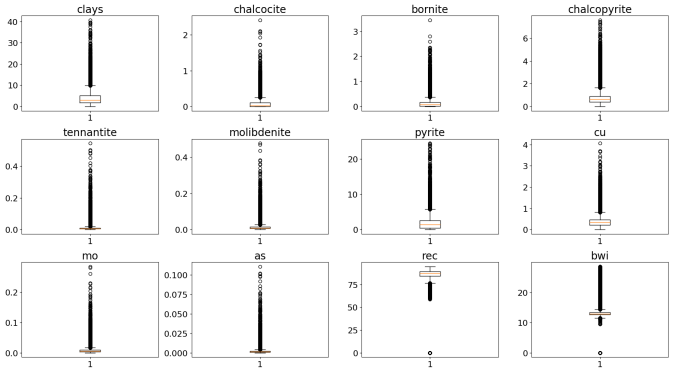

Fig. 1. Boxplots das variáveis mineralógicas, geoquímicas e metalúrgicas. Nota-se a presença de assimetrias e *outliers* relevantes em praticamente todas as distribuições.

A. Dados Utilizados

O conjunto de dados analisado neste estudo é composto por 153.076 amostras válidas e sem valores ausentes, provenientes de um modelo de blocos tridimensional que representa a distribuição espacial de variáveis geológicas e metalúrgicas em um depósito mineral. Esse modelo foi construído com base em dados de furos de sondagem e outros levantamentos geotécnicos, sendo que cada linha da base de dados corresponde a um bloco georreferenciado com propriedades estimadas por interpolação. Todos os dados disponíveis foram considerados na análise, sem exclusões, e serão utilizados nas próximas etapas de regressão das variáveis geometalúrgicas.

As variáveis disponíveis incluem coordenadas espaciais (*x*, *y*, *z*), teores de minerais (como *clays*, *chalcocite*, *bornite*, *chalcopyrite*, *tennantite*, *molibdenite*, *pyrite*) e elementos químicos (*cu*, *mo*, *as*). Além disso, o conjunto contempla duas variáveis-alvo de interesse para modelagem posterior: a recuperação metalúrgica (*REC*) e o índice de moabilidade (*BWI*).

A análise exploratória inicial revelou forte presença de assimetrias e outliers, especialmente nos teores minerais e nos elementos químicos — um comportamento esperado em sistemas geológicos complexos. Os boxplots evidenciaram faixas interquartis estreitas e caudas superiores longas, indicando grande dispersão entre as amostras. A variável *REC* apresentou valores concentrados entre 80% e 95%, com poucas exceções, enquanto o *BWI* mostrou distribuição mais centralizada, mas também com ocorrência de valores extremos. Os teores de minerais como *chalcopyrite* e *bornite*, bem como do metal *cu*, exibiram concentrações baixas na maior parte dos blocos, com alguns picos localizados de alto teor.

A correlação entre variáveis revelou relações consistentes, destacando-se a forte associação entre cobre e *chalcopyrite*, o que reforça a coerência mineralógica dos dados. Também foram observadas correlações relevantes entre *REC* e *BWI*, sugerindo que maiores teores e maiores índices de moabilidade tendem a estar associados a melhores recuperações metalúrgicas. Apesar disso, muitas correlações entre os pares de variáveis são baixas, indicando uma alta complexidade estrutural do sistema.

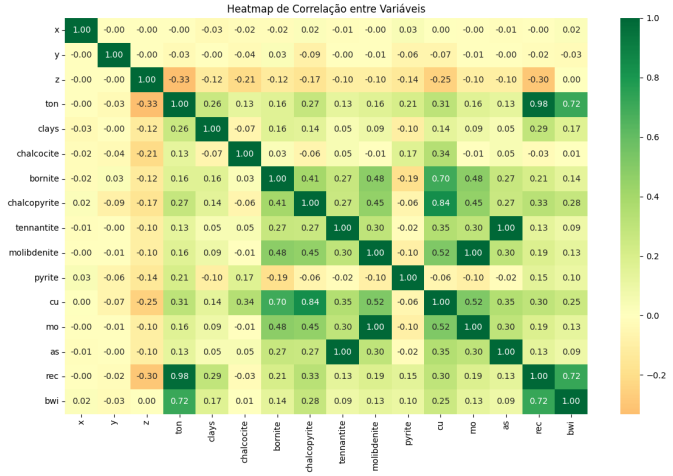

Fig. 2. Mapa de calor das correlações de Pearson entre as variáveis. Destaque para as correlações positivas entre *cu*, *chalcopyrite*, *bornite*, *REC* e *ton*.

Do ponto de vista espacial, a avaliação do índice de Moran evidenciou que várias variáveis apresentam autocorrelação espacial significativa. As variáveis *BWI*, *REC*, *ton*, *cu* e *chalcopyrite* apresentaram os maiores valores de autocorrelação, o que indica que essas propriedades tendem a se distribuir de forma estruturada no espaço. Essa característica é particularmente importante para a definição de domínios espaciais e para a aplicação de algoritmos de clusterização que levem em consideração tanto os valores das variáveis quanto a sua localização geográfica. Em contraste, variáveis como *as* e *mo* apresentaram comportamento mais aleatório, com baixa autocorrelação.

B. Pré-processamento dos Dados

Com o objetivo de uniformizar a escala das variáveis e assegurar que todas contribuam de maneira equitativa na análise de agrupamento, foi aplicada a normalização por padronização utilizando o método *StandardScaler*. Esse método transforma cada variável de forma que sua distribuição passe a ter média zero e desvio padrão igual a um, eliminando o efeito de magnitude e tornando as variáveis comparáveis.

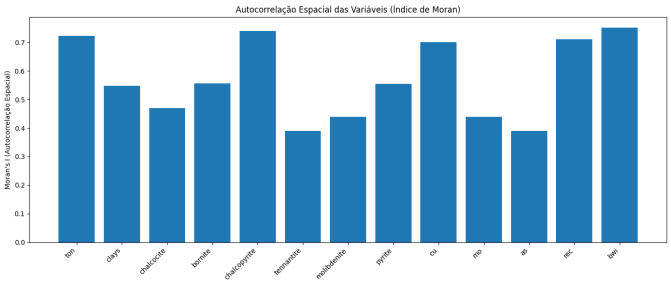

Fig. 3. Autocorrelação espacial das variáveis com base no índice de Moran. Variáveis com maior estrutura espacial incluem *BWI*, *REC*, *ton*, *cu* e *chalcopyrite*.

A transformação é realizada segundo a fórmula:

$$z_i = \frac{x_i - \mu}{\sigma}$$

em que:

- x_i representa o valor original da variável;
- μ é a média da variável;
- σ é o desvio padrão da variável;
- z_i é o valor padronizado resultante.

Como o conjunto de dados não apresenta valores ausentes, não foi necessária a aplicação de técnicas de imputação, permitindo o uso direto da base na etapa de pré-processamento.

C. Análise de Clusterização Multivariada

Para segmentar o depósito em regiões com características estatísticas e espaciais semelhantes, foram aplicadas duas técnicas de agrupamento não supervisionado: o *K-means* e o *Agglomerative Clustering* com restrição espacial.

É importante ressaltar que, com o objetivo de preservar a integridade do processo preditivo, as variáveis-alvo *recuperação metalúrgica* (*REC*) e *índice de moabilidade* (*BWI*) não foram utilizadas na etapa de clusterização. A segmentação foi realizada exclusivamente com base em variáveis independentes disponíveis no modelo de blocos:

- Teores minerais: *clays*, *chalcocite*, *bornite*, *chalcopyrite*, *tennantite*, *molibdenite*, *pyrite*;
- Elementos químicos: *cu* (cobre), *mo* (molibdênio), *as* (arsênio).

Esses atributos refletem características mineralógicas e geoquímicas que são conhecidas previamente à etapa de processamento, sendo adequados para a definição de domínios geometalúrgicos. A exclusão de *REC* e *BWI* evita vazamento de informação (*data leakage*) e garante que a regressão posterior dessas variáveis possa ser feita com base em agrupamentos genuinamente não supervisionados.

1) Seleção do Número de Clusters: A definição do número ideal de grupos (k) foi feita por meio do **método do cotovelo**, que avalia a relação entre o número de clusters e a inércia total do sistema — definida como a soma das distâncias quadradas entre os pontos e seus respectivos centróides. Quanto menor a inércia, mais compactos e coesos são os clusters.

Para essa análise, foi utilizada a função `KMeans` da biblioteca `scikit-learn`, implementada em `Python`, que calcula automaticamente a inércia ao final do ajuste do modelo. Foi executado um processo iterativo variando k entre 1 e 10, e os valores de inércia obtidos foram utilizados para construir o gráfico de cotovelo.

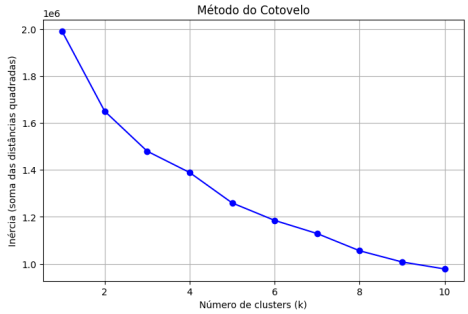

Fig. 4. Determinação do número de clusters via método do cotovelo.

A partir da curva gerada, o valor de k correspondente ao ponto de inflexão (onde a taxa de redução da inércia se estabiliza) é adotado como ótimo [7]. A análise gráfica (Figura 4) demonstrou que esse ponto ocorre em $k = 4$, justificando a escolha desse valor para ambas as técnicas de agrupamento.

2) Clusterização com o Algoritmo K-means: O algoritmo *K-means* foi utilizado como técnica de referência pela sua simplicidade e eficiência computacional em grandes volumes de dados. O método busca minimizar a variância intra-cluster, associando cada ponto ao centróide mais próximo e atualizando iterativamente os centróides até a convergência.

A função objetivo minimizada é dada por:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{1}$$

em que C_i representa o conjunto de pontos do cluster i e μ_i é o centróide correspondente.

Embora eficaz do ponto de vista estatístico, o *K-means* não leva em consideração a localização espacial dos dados, podendo gerar agrupamentos fragmentados no espaço físico [6], [7].

3) Clusterização Hierárquica com Restrição Espacial: Para incorporar a estrutura espacial dos dados à análise, foi aplicada a técnica de *Agglomerative Clustering* com restrição geográfica, utilizando a implementação disponível na biblioteca `scikit-learn`, em ambiente `Python`. Embora o algoritmo em si não tenha sido desenvolvido do zero, a aplicação envolveu um nível significativo de customização no pré-processamento dos dados e na construção da estrutura de conectividade espacial.

Inicialmente, os atributos foram separados em dois grupos: variáveis geoquímicas e mineralógicas e coordenadas espaciais (x, y, z). Ambos os grupos foram padronizados separadamente e, em seguida, combinados em uma única matriz de entrada ponderada. Para isso, adotaram-se pesos $\beta = 0{,}8$ para as

variáveis geoquímicas e $\alpha = 0{,}4$ para as coordenadas espaciais, com o objetivo de balancear a influência relativa de cada dimensão no processo de agrupamento.

Esses valores foram determinados por meio de uma busca aleatória (*random search*) no intervalo contínuo entre 0 e 1, avaliando diferentes combinações de pesos e selecionando aquela que resultou na melhor coerência espacial e estatística dos agrupamentos, conforme as métricas descritas anteriormente.

Além disso, foi construído um grafo de conectividade espacial com base nos 10 vizinhos mais próximos de cada bloco, considerando apenas as coordenadas tridimensionais. Esse grafo foi fornecido como parâmetro ao algoritmo hierárquico, restringindo a fusão de clusters apenas entre blocos espacialmente próximos.

O critério de ligação adotado foi o método de Ward, que busca minimizar a variância intra-cluster a cada fusão. Esse critério calcula o aumento da soma dos quadrados das distâncias (inércia) dentro dos grupos resultantes sempre que dois clusters são unidos, e seleciona a fusão que gera o menor aumento possível. A aplicação com $k = 4$ resultou em agrupamentos espacialmente contíguos, com maior aderência à estrutura geológica do depósito [3], [4], [6].

Essa abordagem híbrida, que combina pesos explícitos, normalização independente e conectividade espacial, tem sido amplamente recomendada na literatura geometalúrgica por garantir maior coerência espacial dos domínios — o que é essencial para aplicações em modelagem mineral realista [3], [4].

D. Regressão com Gradient Boosting

A modelagem preditiva das variáveis geometalúrgicas neste trabalho foi realizada por meio do algoritmo de **Gradient Boosting**, uma técnica amplamente utilizada em problemas de regressão e classificação com dados tabulares. Essa abordagem baseia-se no princípio de aprendizado por conjunto (*ensemble learning*), combinando sequencialmente múltiplos modelos fracos — normalmente árvores de decisão — para construir um modelo forte e robusto.

O funcionamento do Gradient Boosting consiste em ajustar cada nova árvore para prever os **erros residuais** do modelo anterior. A cada iteração, é minimizada uma função de perda (como o erro quadrático médio), direcionando o aprendizado das árvores seguintes para as amostras com maior erro.

Formalmente, dado um conjunto de dados $\{(x_i, y_i)\}_{i=1}^{n}$, o modelo preditivo é construído como uma soma de funções:

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i),$$

em que cada f_m representa uma árvore de decisão ajustada para corrigir os erros residuais da predição acumulada até a iteração $m - 1$.

Neste trabalho, foi utilizada a biblioteca `XGBoost`, uma implementação eficiente e otimizada do Gradient Boosting. O XGBoost incorpora técnicas adicionais como regularização (L_1 e L_2), *shrinkage* (taxa de aprendizado), *subsampling* e paralelização de árvores, o que contribui para melhor generalização e desempenho computacional.

A escolha desse algoritmo se justifica por sua capacidade de capturar relações não lineares complexas, lidar bem com dados com outliers e oferecer ótimo desempenho mesmo com variáveis altamente correlacionadas — características presentes nos dados geometalúrgicos utilizados.

E. Busca de Hiperparâmetros com Random Search

Modelos baseados em *Gradient Boosting*, como o XGBoost, possuem diversos hiperparâmetros que controlam a complexidade do modelo, sua regularização e o comportamento da árvore durante o processo de treinamento. A escolha adequada desses hiperparâmetros é essencial para maximizar a performance preditiva e evitar problemas como sobreajuste.

Neste trabalho, a seleção dos hiperparâmetros foi realizada por meio da técnica de *Randomized Search* (*RandomSearchCV*), que consiste em amostrar aleatoriamente combinações de parâmetros a partir de distribuições previamente definidas, ao invés de testar exaustivamente todas as possibilidades como na *Grid Search*.

As distribuições adotadas para a busca foram:

```
param_dist = {
    'n_estimators': randint(100, 600),
    'max_depth': randint(3, 10),
    'learning_rate': uniform(0.01, 0.2),
    'subsample': uniform(0.7, 0.3),
    'colsample_bytree': uniform(0.7, 0.3),
    'reg_alpha': uniform(0, 1),
    'reg_lambda': uniform(1, 10)
}
```

A seguir, apresenta-se a descrição de cada hiperparâmetro:

- **n_estimators**: número total de árvores que compõem o modelo. Valores maiores aumentam o poder de aprendizado, mas também o custo computacional e o risco de sobreajuste.
- **max_depth**: profundidade máxima de cada árvore. Controla a complexidade do modelo; valores mais altos permitem capturar interações complexas, mas podem gerar sobreajuste.
- **learning_rate**: taxa de aprendizado que controla a contribuição de cada nova árvore no modelo final. Valores menores tornam o aprendizado mais lento, porém mais robusto.
- **subsample**: fração de amostras utilizada para treinar cada árvore. Introduz aleatoriedade no treinamento, ajudando a reduzir o sobreajuste (semelhante ao *bagging*).
- **colsample_bytree**: fração de variáveis (colunas) selecionadas aleatoriamente para treinar cada árvore. Ajuda a reduzir correlação entre árvores e melhorar a generalização.
- **reg_alpha**: parâmetro de regularização L_1 (Lasso), que incentiva a esparsidade dos coeficientes e pode eliminar variáveis irrelevantes.

- **reg_lambda**: parâmetro de regularização L_2 (Ridge), que penaliza grandes coeficientes e ajuda a evitar sobreajuste.

Foram testadas 10000 combinações aleatórias desses hiperparâmetros, e a melhor configuração foi selecionada com base na métrica de desempenho R^2 obtida em validação cruzada com 5 *folds*. Esse processo foi repetido separadamente para cada variável-alvo e para cada modelo treinado (global e local por cluster).

F. Ambiente Computacional e Ferramentas

Todos os experimentos foram conduzidos em um pc Linux (Ubuntu 22.04). O código foi desenvolvido em Python 3.12, utilizando `pandas` 2.2, `scikit-learn` 1.5, `xgboost` 2.1 e bibliotecas geoespaciais (`geopandas` e `libpysal`) para cálculo do índice de Moran e operações de vizinhança. A busca de hiperparâmetros (*RandomizedSearchCV*) empregou validação cruzada 5-fold com `numpy.random.seed(42)`.

TABLE I
RESUMO DE FERRAMENTAS E VERSÕES

Componente	**Versão/Detalhe**
SO	Ubuntu 22.04
Python	3.12
pandas	2.2
scikit-learn	1.5
xgboost	2.1.x
geopandas / libpysal	0.14 / 4
ChatGPT (OpenAI)	apoio à redação/revisão (checagem humana)
Perplexity AI	triagem inicial de referências (verificação manual)

IV. RESULTADOS

A. Clusterização com o Algoritmo K-means

Após a aplicação do algoritmo *K-means* com $k = 4$, os clusters obtidos apresentaram boa separação em termos de similaridade química, porém sem garantir a continuidade espacial dos blocos. O resultado visual evidencia uma distribuição mais fragmentada, com regiões de um mesmo cluster aparecendo em diferentes zonas do depósito.

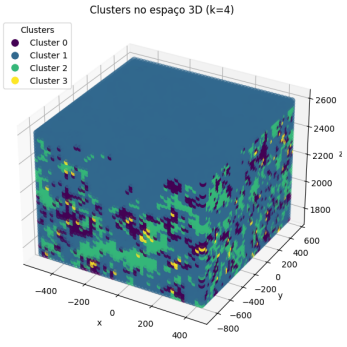

Fig. 5. Distribuição dos clusters no espaço tridimensional utilizando o algoritmo *K-means* com $k = 4$.

O número de blocos por cluster foi:

- Cluster 1: 91.760 blocos
- Cluster 0: 28.647 blocos
- Cluster 2: 28.345 blocos
- Cluster 3: 4.324 blocos

As métricas de avaliação foram as seguintes:

- **Silhouette Score médio:** 0,0060
- **WCSS (Within-Cluster Sum of Squares):** 973.715,76
- **Entropia Espacial (H):** 0,5614

O Silhouette Score muito próximo de zero indica fraca separação entre os clusters. Apesar disso, o WCSS sugere boa compactação interna. A entropia espacial aponta para uma distribuição relativamente dispersa no espaço físico.

B. Clusterização com o Algoritmo Agglomerative com Restrição Espacial

O agrupamento hierárquico com restrição espacial proporcionou uma segmentação visualmente mais contínua e coerente com a estrutura tridimensional do depósito. A imposição de conectividade geográfica resultou em regiões mais compactas e geologicamente plausíveis.

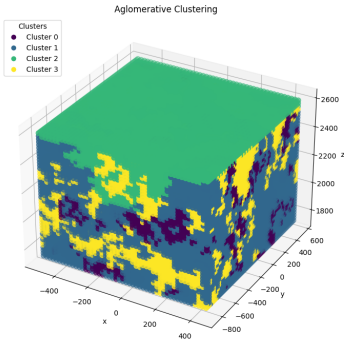

Fig. 6. Distribuição dos clusters no espaço tridimensional utilizando *Agglomerative Clustering* com restrição espacial e $k = 4$.

O número de blocos por cluster foi:

- Cluster 0: 56.854 blocos
- Cluster 3: 51.387 blocos
- Cluster 2: 26.456 blocos
- Cluster 1: 18.379 blocos

As métricas de avaliação foram:

- **Silhouette Score médio:** -0,0073
- **WCSS:** 1.177.333,40
- **Entropia Espacial (H):** 0,4855

Apesar do Silhouette Score levemente negativo — o que pode ser atribuído à limitação da métrica para reconhecer continuidade espacial —, a entropia espacial significativamente menor e a coerência visual reforçam a adequação do método para cenários com forte dependência geográfica.

C. Comparação Entre os Métodos de Agrupamento

A comparação entre os dois métodos evidencia que:

- O *K-means* apresenta clusters mais compactos do ponto de vista estatístico, mas menos coerentes do ponto de vista geológico.
- O *Agglomerative Clustering* com restrição espacial produz agrupamentos espacialmente contínuos e mais condizentes com a morfologia geológica do depósito.
- O trade-off entre compactação estatística (menor WCSS) e coerência espacial (menor entropia H) deve ser considerado conforme os objetivos da modelagem.

D. Modelagem Preditiva da Recuperação

Nesta etapa, foram ajustados modelos de regressão para a variável de interesse *recuperação metalúrgica (REC)*, com o objetivo de avaliar a capacidade preditiva dos algoritmos em diferentes contextos. As abordagens consideradas incluem tanto um modelo global, ajustado sobre todo o conjunto de dados, quanto modelos locais treinados separadamente para cada cluster geometalúrgico.

1) Modelo Global: Com os domínios definidos, foi possível dar sequência à etapa de regressão das variáveis de interesse. Inicialmente, foi ajustado um modelo global para a predição da recuperação metalúrgica (*REC*), utilizando o algoritmo *XGBoost Regressor* e todo o conjunto de dados, sem distinção por clusters.

Os resultados obtidos sobre o conjunto de teste indicam um desempenho preditivo satisfatório:

- R^2: 0,7947
- RMSE: 1,6765
- MAE: 1,2318

O coeficiente de determinação próximo de 0,80 evidencia que o modelo consegue explicar uma parcela substancial da variância da variável recuperação. O erro médio absoluto em torno de 1,23 pontos percentuais é aceitável do ponto de vista operacional para fins de planejamento de processo.

Na Figura 7, são apresentadas as visualizações correspondentes: o gráfico de dispersão entre valores reais e previstos, e o histograma da distribuição dos resíduos.

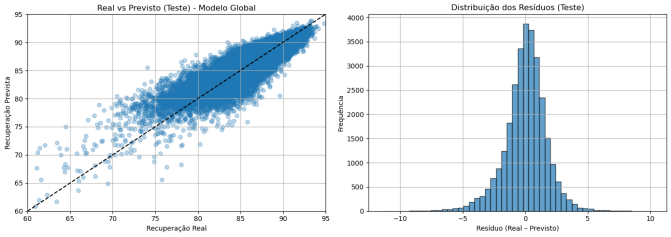

Fig. 7. Desempenho do modelo global na predição da Recuperação: (à esquerda) comparação entre valores reais e previstos; (à direita) distribuição dos resíduos.

Observa-se uma forte concentração ao longo da diagonal, indicando bom alinhamento entre predições e valores observados. A distribuição dos resíduos é aproximadamente simétrica e centrada em zero, o que sugere ausência de viés sistemático e comportamento próximo ao esperado para um modelo bem ajustado.

2) Modelo Local por K-means: Em seguida, foi ajustado um conjunto de modelos locais para a predição da recuperação metalúrgica (*REC*), considerando a segmentação prévia dos dados obtida com o algoritmo *K-means*. Para cada cluster identificado, foi treinado um modelo independente utilizando o algoritmo *XGBoost Regressor*, com particionamento interno dos dados em treino, validação e teste.

A avaliação agregada dos modelos locais, considerando o desempenho sobre os conjuntos de teste de cada cluster, revelou um leve ganho em relação ao modelo global:

- R^2: 0,8057
- RMSE: 1,6460
- MAE: 1,2092

O aumento no R^2 e a redução do erro absoluto médio indicam que a regressão segmentada por domínios gerados via *K-means* foi capaz de capturar nuances locais relevantes, resultando em predições ligeiramente mais precisas.

A Figura 8 apresenta as visualizações consolidadas dos resultados dos modelos locais: à esquerda, a comparação entre os valores reais e previstos, e à direita, a distribuição dos resíduos.

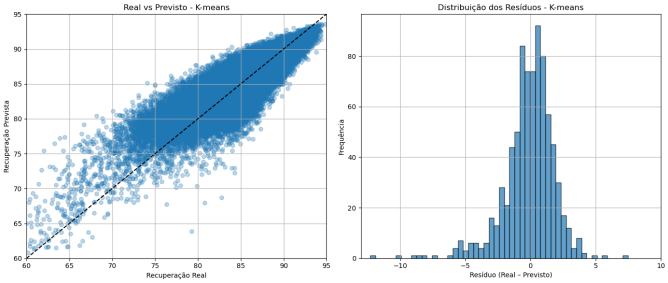

Fig. 8. Desempenho dos modelos locais por cluster (*K-means*) na predição da Recuperação.

Visualmente, observa-se uma maior concentração ao longo da reta de identidade e uma distribuição de resíduos ainda centrada em zero, porém ligeiramente mais estreita em comparação ao modelo global — sugerindo uma melhoria na robustez preditiva por meio da especialização dos modelos.

3) Modelo Local por Agglomerative Clustering: Por fim, foram ajustados modelos locais de regressão com base na segmentação geometalúrgica obtida via *Agglomerative Clustering* com restrição espacial. Essa abordagem leva em conta a conectividade geográfica entre os blocos, promovendo a formação de domínios espacialmente contíguos e geologicamente coerentes.

Assim como no caso do K-means, foi treinado um modelo *XGBoost Regressor* separado para cada cluster. O desempenho agregado obtido nos conjuntos de teste foi ligeiramente superior ao das demais abordagens, conforme resumo abaixo:

- R^2: 0,8086
- RMSE: 1,6336
- MAE: 1,2035

Os resultados indicam que a regressão segmentada por domínios definidos com restrição espacial apresenta a melhor capacidade explicativa entre as alternativas avaliadas, com o menor erro absoluto médio e o maior coeficiente de determinação.

A Figura 9 mostra os resultados consolidados da predição local por Agglomerative Clustering.

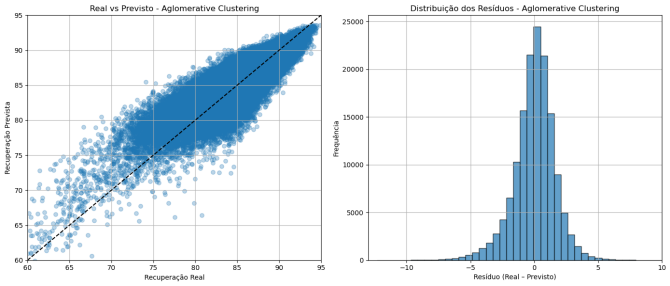

Fig. 9. Desempenho dos modelos locais por cluster (Agglomerative Clustering) na predição da Recuperação.

Observa-se uma concentração ainda mais densa ao longo da reta de identidade em relação às demais abordagens, além de uma distribuição de resíduos estreita e simétrica, com leve redução na dispersão. Esses padrões reforçam a ideia de que a especialização por clusters espacialmente coerentes contribui para a melhoria do ajuste preditivo.

E. Modelagem de Regressão para o Índice de Moabilidade (BWI)

A seguir, são apresentados os resultados da modelagem preditiva para o índice de moabilidade (*Bond Work Index* — BWI), variável que representa a energia necessária para a moagem adequada do minério. Foram testadas duas abordagens: modelo global e modelos segmentados por cluster (K-means).

Modelo Global: De forma análoga ao modelo de recuperação, também foi ajustado um modelo global para a predição do índice de moabilidade (*BWI*), utilizando o algoritmo *XGBoost Regressor* e todo o conjunto de dados, sem segmentações.

Os resultados obtidos para o conjunto de teste são os seguintes:

- R^2: 0,4440
- RMSE: 1,9305
- MAE: 0,9953

O desempenho preditivo foi inferior ao observado na variável recuperação, com um coeficiente de determinação indicando que apenas cerca de 44% da variância do *BWI* é explicada pelo modelo. Esse resultado sugere que a variável apresenta maior complexidade ou ruído, e que fatores relevantes podem não estar totalmente representados nas variáveis utilizadas.

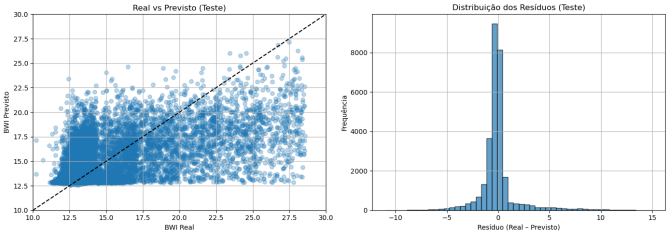

Fig. 10. Desempenho do modelo global na predição do BWI: (à esquerda) comparação entre valores reais e previstos; (à direita) distribuição dos resíduos.

Conforme mostra a Figura 10, as predições tendem a se concentrar em uma faixa centralizada, com uma dispersão considerável em relação à diagonal ideal. A distribuição dos resíduos é razoavelmente simétrica, mas com caudas mais alongadas em comparação ao modelo de recuperação, reforçando a maior variabilidade do fenômeno a ser modelado.

Esses resultados motivam a investigação de abordagens alternativas, como a segmentação espacial por cluster, visando capturar padrões mais locais que possam ser responsáveis por comportamentos distintos do índice de moabilidade no depósito.

Modelo Segmentado por K-means: Nesta abordagem, o conjunto de dados foi previamente segmentado em clusters por meio do algoritmo *K-means*, e um modelo preditivo foi treinado separadamente para cada grupo. A ideia é capturar padrões locais que possam ser perdidos em um modelo global.

Os resultados consolidados dos modelos treinados em cada cluster foram:

- R^2: 0,5505
- RMSE: 1,7475
- MAE: 0,9090

Comparando-se com o modelo global, observa-se uma melhora consistente nos três indicadores. O aumento do coeficiente de determinação indica que os modelos segmentados foram capazes de capturar melhor as variabilidades do *BWI*, provavelmente por se ajustarem a regiões estatisticamente mais homogêneas.

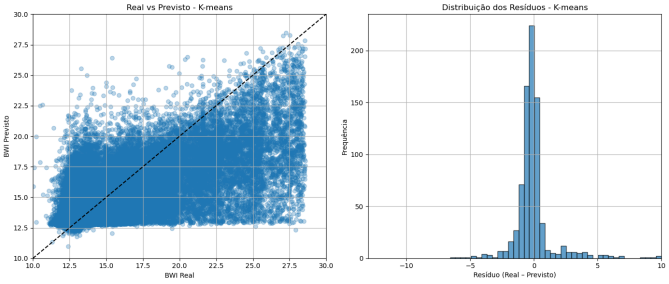

Fig. 11. Desempenho do modelo segmentado por K-means na predição do BWI. À esquerda: relação entre valores reais e previstos. À direita: histograma dos resíduos.

A Figura 11 mostra uma distribuição de resíduos bem centrada, com uma dispersão ligeiramente menor do que a observada no modelo global. Isso reforça a hipótese de que

a segmentação introduzida pelo K-means contribui para um ganho em especialização dos modelos locais.

Modelo Segmentado por Agglomerative Clustering: Por fim, foi testada uma terceira abordagem com segmentação via *Agglomerative Clustering*. Assim como na estratégia anterior, um modelo foi treinado individualmente para cada cluster identificado.

Os resultados agregados dos modelos segmentados foram:

- R^2: 0,5546
- RMSE: 1,7396
- MAE: 0,9015

Esse modelo apresentou desempenho ligeiramente superior ao segmentado por K-means, consolidando-se como a melhor abordagem testada para a variável BWI. O ganho no R^2 reflete uma maior capacidade de explicação da variância local nos dados.

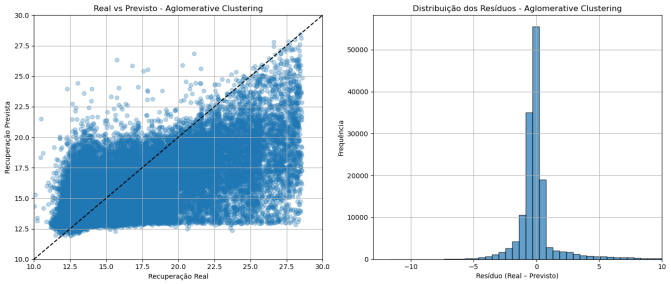

Fig. 12. Desempenho do modelo segmentado por Agglomerative Clustering na predição do BWI. À esquerda: comparação entre valores reais e previstos. À direita: distribuição dos resíduos.

Na Figura 12, observa-se uma forte concentração de resíduos em torno de zero, com distribuição simétrica e caudas estreitas. A dispersão no gráfico de predição também se mostra reduzida, indicando previsões mais consistentes em comparação às abordagens anteriores.

V. Conclusão

Desempenho na Predição da Recuperação

A Tabela II resume os resultados obtidos pelos três modelos de regressão aplicados à variável *Recuperação*.

TABLE II
Comparação dos modelos para a variável Recuperação

Modelo	R^2	RMSE	MAE
Global	0,7941	1,6787	1,2331
Segmentado (K-means)	0,8057	1,6460	1,2092
Segmentado (Aglomerativo)	**0,8086**	**1,6336**	**1,2035**

Observa-se uma melhoria progressiva nos indicadores de desempenho ao se adotar abordagens segmentadas por cluster. O modelo com *Agglomerative Clustering* obteve o melhor desempenho em todos os critérios, com ganhos discretos porém consistentes em relação ao modelo global. Isso indica que a segmentação permitiu uma especialização dos modelos locais, que conseguiram capturar com maior precisão as variações regionais da recuperação metalúrgica.

Desempenho na Predição do BWI

A Tabela III apresenta os mesmos indicadores para os modelos voltados à predição do índice de moabilidade (*BWI*).

TABLE III
Comparação dos modelos para a variável BWI

Modelo	R^2	RMSE	MAE
Global	0,4440	1,9305	0,9953
Segmentado (K-means)	0,5505	1,7475	0,9090
Segmentado (Aglomerativo)	**0,5546**	**1,7396**	**0,9015**

Os ganhos com a segmentação são ainda mais evidentes na predição do BWI. Enquanto o modelo global explicou apenas cerca de 44% da variância da variável, os modelos segmentados atingiram coeficientes de determinação superiores a 55%. Além disso, os erros (RMSE e MAE) foram consideravelmente reduzidos.

Esses resultados sugerem que o BWI é uma variável mais sensível a variações locais do depósito, e que a modelagem segmentada consegue captar nuances que se perdem na abordagem global.

Considerações Finais

A comparação entre os modelos globais e segmentados por cluster evidencia os benefícios da regionalização na modelagem de variáveis geometalúrgicas. As abordagens com *K-means* e, especialmente, com *Agglomerative Clustering* apresentaram desempenho superior em ambas as variáveis analisadas.

A principal vantagem da segmentação é permitir que os modelos se adaptem a regiões estatística e geologicamente mais homogêneas, resultando em previsões mais acuradas e confiáveis. Isso é particularmente importante em contextos de planejamento de lavra e otimização de processos, onde decisões baseadas em modelos preditivos impactam diretamente os custos e a eficiência da operação mineral.

References

[1] R. Blannin et al., "A quantitative particle-based approach for the geometallurgical assessment of tailings deposits," Earth Sci. Syst. Soc., vol. 4, Art. no. 10102, Sep. 2024. https://doi.org/10.3389/esss.2024.10102.

[2] N. J. R. Armijos and M. Calderón-Celis, "Geometallurgical simulation of the work index in a porphyry copper deposit using geostatistical techniques," Rev. Cient. de Mineralurgia y Metalurgia, vol. 8, no. 3, pp. 807–814, 2024. https://doi.org/10.37811/cl_rcm.v8i3.11288.

[3] A. W. Boroh, K. Y. Sore-Gamo, M. A. Ngounouno, I. B. G. Mbowou, and I. Ngounouno, "Implication of geological domains data for modeling and estimating resources from Nkout iron deposit (South-Cameroon)," J. Mining and Metallurgy, Sect. A: Mining, vol. 57, no. 1, pp. 1–17, 2021. https://doi.org/10.5937/JMMA2101001B.

[4] B. G. Silveira, "Revisão do uso de técnicas de agrupamento para definição de domínios estacionários," Monografia, CEERMIN/UFMG, 2022.

[5] C. R. Ortiz et al., "Predictive modelling of mineral processing performance using machine learning: A review," Minerals, vol. 12, no. 1, p. 62, 2022. https://doi.org/10.3390/min12010062.

[6] G. C. Moreira, R. C. C. Modena, J. F. C. L. Costa, and D. M. Marques, "A workflow for defining geological domains using machine learning and geostatistics," Tecnol. Metal. Mater. Min., vol. 18, 2021, Art. no. e2472. https://doi.org/10.4322/2176-1523.20212472.

[7] D. Mackenzie et al., “Inertia-based indices to determine the number of clusters in K-means: An experimental evaluation,” IEEE Access, 2024. https://doi.org/10.1109/ACCESS.2024.3350791.

[8] R. Martin and J. Boisvert, “Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics,” Comput. Geosci., vol. 120, pp. 82–96, 2018. https://doi.org/10.1016/j.cageo.2018.08.005.

[9] H. Mohammadi, S. T. Hosseini, O. Asghari, and P. A. Harouni, “Ore body domaining by clustering of multiple-point data events: a case study from the Dalli porphyry copper-gold deposit, central Iran,” Ore Energy Resource Geol., vol. 10, Art. 100018, 2022. https://doi.org/10.1016/j.oreoa.2022.100018.

[10] Y. Abildin, C. Xu, P. Dowd, and A. Adeli, “Geometallurgical responses on lithological domains modelled by a hybrid domaining framework,” Minerals, vol. 13, no. 7, p. 918, 2023. https://doi.org/10.3390/min13070918.

[11] D. Jung and Y. Choi, “Systematic review of machine learning applications in mining: Exploration, exploitation, and reclamation,” Minerals, vol. 11, no. 2, p. 148, 2021. https://doi.org/10.3390/min11020148.

[12] G. T. Nwaila et al., “An integrated geodata science workflow for resource estimation: A case study from the Merensky Reef, Bushveld Complex,” Nat. Resour. Res., vol. 34, pp. 1301–1329, 2025. https://doi.org/10.1007/s11053-025-10471-4.