UNIVERSIDADE FEDERAL DE OURO PRETO

ESCOLA DE MINAS

DEPARTAMENTO DE ENGENHARIA DE MINAS - DEMIN

MARIA TERESA FERNANDES MATOS ALVES

CONTRIBUIÇÕES INICIAIS À MODELAGEM GEOMETALÚRGICA DE UMA MINA DE FERRO POR MEIO DE ANÁLISES ESTATÍSTICAS MULTIVARIADAS

OURO PRETO
2025

MARIA TERESA FERNANDES MATOS ALVES

CONTRIBUIÇÕES INICIAIS À MODELAGEM GEOMETALÚRGICA DE UMA MINA DE FERRO POR MEIO DE ANÁLISES ESTATÍSTICAS MULTIVARIADAS

Trabalho de Conclusão de Curso apresentado à banca examinadora do departamento de Engenharia de Minas da Universidade Federal de Ouro Preto, como parte dos requisitos necessários para a obtenção de grau de Bacharel em Engenharia de Minas.

Orientador: Prof. Dr. Ivo Eyer Cabral

OURO PRETO
2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
ESCOLA DE MINAS
DEPARTAMENTO DE ENGENHARIA DE MINAS

FOLHA DE APROVAÇÃO

Maria Teresa Fernandes Matos Alves

Contribuições iniciais à modelagem geometalúrgica de uma mina de ferro por meio de análises estatísticas multivariadas

Monografia apresentada ao Curso de Engenharia de Minas da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Minas

Aprovada em 25 de agosto de 2025

Membros da banca

Doutor Ivo Eyer Cabral - Orientador (Universidade federal de Ouro Preto)
Doutor Carlos Enrique Arroyo Ortiz - (Universidade federal de Ouro Preto)
Doutor Allan Erlikhman Medeiros Santos - (Universidade federal de Ouro Preto))

Ivo Eyer Cabral, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 03/09/2025

Documento assinado eletronicamente por **Ivo Eyer Cabral**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 03/09/2025, às 21:12, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0972975** e o código CRC **3DC63A3D**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: 3135591590 - www.ufop.br

Aos meus pais, Gilmar e Cida, por reconhecer e honrar os esforços de vocês

AGRADECIMENTOS

Ao Senhor, meu Deus, que me guia em suas perfeitas promessas, devo toda minha gratidão. Aos meus pais, Gilmar e Cida, pelo esforço e dedicação para me proporcionar essa oportunidade e ser lar nesses anos de graduação.

Aos meus irmãos, Ana e João, por todo amor, sacrifício, incentivo e amizade durante todas as etapas da minha trajetória. Por compreenderem, melhor que ninguém, a ausência necessária em prol da dedicação de anos que uma formação exige.

Ao Ygor pelo companheirismo, incentivo e amor. Estendo a minha gratidão aos meus amigos da Engenharia de Minas, por fazerem parte de um caminho que será lembrado com muito carinho para toda a vida. Em especial ao Lucas, Victor, Leonardo, Mateus e Lóren. Ao pretanas, minhas amigas de longa data, que foram apoio durante toda a jornada.

Ao meu orientador, Ivo Eyer, por transmitir e compartilhar comigo seus conhecimentos e pela paciência em minha orientação. Ao Carlão, pelo incentivo, parceria e confiança durante toda a graduação.

Aos professores, técnicos e funcionários do Departamento de Engenharia de Minas pela vasta contribuição em minha formação. À Gloriosa Escola de Minas e à UFOP, pela educação gratuita, de qualidade e enriquecedora. À Fundação Gorceix pelo grande incentivo e contribuição em todas as etapas da minha graduação, inclusive pelo reconhecimento de toda dedicação, tempo e esforços colocados em meus estudos através da bolsa Mérito Acadêmico.

RESUMO

A inserção de variáveis relacionadas à geometalurgia em um modelo de blocos caracteriza a modelagem geometalúrgica, que por sua vez, possibilita identificar previamente a variabilidade do comportamento no processamento mineral. Para lidar com as diversas variáveis estudadas em um programa de geometalurgia, é comum a aplicação de técnicas de estatística multivariada, com o objetivo de identificar domínios geometalúrgicos e padrões relevantes que auxiliem na previsibilidade do minério. Este estudo explora os resultados preliminares de um banco de dados geometalúrgico em desenvolvimento utilizando técnicas estatísticas multivariadas como contribuição à modelagem ao desenvolver modelos preditivos, por meio de regressão linear multivariada, para a recuperação mássica na concentração de itabiritos de uma mina de ferro. Utilizando os teores globais das amostras de itabiritos, foi possível aplicar a análise de componentes principais (PCA) que revelou uma variabilidade de aproximadamente 81% explicada pelas duas primeiras componentes principais. A técnica de agrupamento (clusterização) pelo algoritmo do k-means, evidenciou padrões importantes em três agrupamentos - ao relacioná-los com a litologia e qualidade do material. O grupo 1 apresentou a maior média para a variável de interesse, a recuperação mássica, enquanto o grupo 3 apresentou a menor média e pior desempenho. A análise estatística de variância (ANOVA) revelou diferenças estatisticamente significativas em relação à média da recuperação mássica e a análise multivariada de variância (MANOVA) revelou o mesmo padrão para os teores de Fe e SiO_2. Os modelos desenvolvidos para os agrupamentos 1 e 2 indicam que as variações da recuperação mássica para as litologias presentes nesses grupos podem ser explicadas pelos teores globais, com destaque para o cluster 1, que apresentou o melhor resultado (com R^2 de 99,42% e R^2 preditivo de cerca de 92%), embora seja limitado pela quantidade de amostras. Considerando todos os itabiritos de maneira geral obteve-se um modelo com com R^2 de 82%, R^2 ajustado de aproximadamente 80% e R^2 preditivo de 77%, tendo SiO_2_gl e PPC_gl como variáveis preditoras. O modelo sugere que os teores explicam bem a variável geometalúrgica de interesse nesse trabalho, porém recomenda-se que novos estudos sejam feitos com outras características do material como a granulometria e mineralogia.

Palavras-Chave: Geometalurgia; Análise de Componentes Principais (PCA); Clusterização por K-means; Regressão Linear Multivariada; Recuperação Mássica

ABSTRACT

The inclusion of geometallurgical related variables in a block model characterizes geometallurgical modeling, which in turn allows for the early identification of variability in mineral processing behavior. To handle the various variables studied in a geometallurgical program, multivariate statistical techniques are commonly applied with the aim of identifying geometallurgical domains and relevant patterns that aid in ore predictability. This study explores preliminary results from a developing geometallurgical database, employing multivariate statistical techniques as a contribution to modeling by developing predictive models, through multivariate linear regression, for the mass recovery during itabirite concentration in an iron ore mine. Using the global grades of itabirites samples, principal component analysis (PCA) was applied, revealing approximately 83% of variability explained by the first two principal components. Clustering using the k-means algorithm revealed important patterns across the three clusters when related to lithology and material quality. Group 1 exhibited the highest mean for the variable of interest, mass recovery, while group 3 presented the lowest. Analysis of variance (ANOVA) revealed statistically significant differences in the mean mass recovery, and multivariate analysis of variance (MANOVA) revealed a similar pattern for Fe and SiO_2 grades. The models developed for clusters 1 and 2 indicate that the variations in mass recovery for the lithologies present in these groups can be explained by global grades, with cluster 1 showing the best result (R^2 of 99.42% and predictive R^2 of about 92%), although limited by sample size. Considering all itabirites collectively, a model was obtained with R^2 of 82%, adjusted R^2 of approximately 80%, and predictive R^2 of 77%, with SiO_2_gl and PPC_gl as predictors. The model suggests that the grades explain the geometallurgical variable of interest well, but it is recommended that future studies incorporate other material characteristics such as particle size distribution and mineralogy.

Keywords: Geometallurgy; Principal Component Analysis (PCA); K-means Clustering; Multivariate Linear Regression; Mass Recovery

LISTA DE FIGURAS

Figura 1 - Mapa geológico da mina em estudo 16
Figura 2 - Principais fases de um programa geometalúrgico 19
Figura 3 - Conjunto de dados antes e depois do agrupamento 24
Figura 4 - Método do cotovelo aplicado a um conjunto de dados 25
Figura 5 - Informações de um gráfico boxplot 26
Figura 6 - Parâmetros a serem analisados em um modelo de regressão 28
Figura 7 - Intensidade e direção da correlação de Pearson 30
Figura 8 - Fluxograma geometalúrgico dos Itabiritos 31
Figura 9 - Fluxograma das análises aplicadas ao subconjunto de dados selecionados 32
Figura 10 - Histograma das variáveis de teor 36
Figura 11 - Gráfico de autovetores das variáveis originais 38
Figura 12 - Gráfico biplot para a PCA 38
Figura 13 - Método do cotovelo aplicado para o agrupamento 39
Figura 14 - Gráfico de dispersão 2D do agrupamento por K-means 40
Figura 15 - Gráfico de silhueta para o agrupamento por k-means 42
Figura 16 - Testes de normalidade das recuperações mássicas por cluster e geral 43
Figura 17 - Boxplot da recuperação mássica por cluster 44
Figura 18 - Matriz de correlação do cluster 1 46
Figura 19 - Matriz de correlação do cluster 2 47
Figura 20 - Matriz de correlação do cluster 3 47
Figura 21 - Matriz de correlação do subconjunto total de dados 47
Figura 22 - Estatísticas do modelo de regressão do cluster 1 48
Figura 23 - Gráficos de resíduos do modelo de regressão do cluster 1 49
Figura 24 - Estatísticas do modelo de regressão do cluster 2 50
Figura 25 - Gráficos de resíduos do modelo de regressão do cluster 2 51
Figura 26 - Estatísticas do modelo de regressão geral 52
Figura 27 - Gráficos de resíduos do modelo de regressão geral considerando todos os itabiritos 53
Figura 28 - Estatísticas do modelo de regressão considerando os clusters 1 e 2 54

LISTA DE TABELAS

Tabela 1 - Materiais presentes na mina, processos de adequação e produtos 17
Tabela 2 - Parâmetros estatísticos de uma regressão 28
Tabela 3 - Variância do subconjunto de variáveis originais 32
Tabela 4 - Estatística descritiva das variáveis de teor 36
Tabela 5 - Autovalores e variância explicada pelas principais componentes 36
Tabela 6 - Carga (autovetores) das variáveis originais em cada componente principal 37
Tabela 7 - Scores das amostras após aplicação da PCA 37
Tabela 8 - Descrição das amostras por litologia e cluster 41
Tabela 9 - Teste paramétrico da análise de variância (ANOVA) para recuperação mássica 43
Tabela 10 - Teste paramétrico da análise multivariada de variância (MANOVA) para Fe_gl e SiO2_gl 44
Tabela 11 - Boxplot das variáveis originais por cluster 45

LISTA DE ABREVIATURAS E SIGLAS

HC – Hematita Compacta

HCS – Hematita Compacta Silicosa

PCA – *Principal Component Analysis*

PC1 – Componente Principal 1

PC2 – Componente Principal 2

IDOL – Itabirito Dolomítico

IF – Itabirito Friável

IC – Itabirito Compacto

IC2 – Itabirito Compacto Pobre

IFP – Itabirito Friável Pobre

SUMÁRIO

1. INTRODUÇÃO 13

2. OBJETIVOS 15

2.1 Objetivo geral 15

2.2 Objetivos específicos 15

3. REVISÃO BIBLIOGRÁFICA 16

3.1 Descrição do empreendimento 16

3.1.1 Operações e Materiais 32

3.2 Geometalurgia 17

3.2.1 Definição e objetivos 32

3.2.2 Domínios geometalúrgicos 32

3.3 Técnicas de análises estatísticas multivariadas 20

3.3.1 Análise de componentes principais (PCA) 32

3.3.2 Agrupamento (clusterização) por k-means 32

3.3.3 Avaliação estatística e comparativa entre os agrupamentos 23

3.3.4 Regressão linear multivariada 24

3.3.4.1 Correlação de Pearson 30

4. METODOLOGIA 31

4.1 Definição de fluxograma geometalúrgico 31

4.2 Análises estatísticas multivariadas 32

4.2.1 Análise de componentes principais 32

4.2.2 Agrupamento (clusterização) por k-means 33

4.2.3 Avaliação estatística e comparativa entre os agrupamentos 34

4.2.4 Modelo de regressão linear 34

5. RESULTADOS E DISCUSSÕES 36

5.1 Análise de componentes principais 36

5.2 Agrupamento por k-means..... 39

5.3 Avaliação estatística e comparativa entre os agrupamentos 43

5.4 Modelos de regressão linear 46

5.4.1 Matrizes de correlação..... 46

5.4.2 Desenvolvimento do modelo de regressão – cluster 1 48

5.4.3 Desenvolvimento do modelo de regressão – cluster 2 50

5.4.4 Desenvolvimento do modelo de regressão – cluster 3 51

5.4.5 Desenvolvimento do modelo de regressão geral..... 52

6. CONSIDERAÇÕES FINAIS E SUGESTÃO DE ESTUDOS FUTUROS 55

REFERÊNCIAS BIBLIOGRÁFICAS 57

1. INTRODUÇÃO

Ao longo das décadas, a complexidade dos minérios de ferro aumentou substancialmente. Segundo Policarpo (2012), o tipo de minério ofertado no mercado foi variando com o passar dos anos devido à exaustão dos materiais de melhores qualidade, fazendo com que a indústrias de mineração e siderurgia se adaptassem a uma nova realidade. Materiais que eram descartados passam a ser viáveis economicamente com o avanço da tecnologia e da crescente demanda da humanidade por este recurso mineral. Somado a isso, os desafios do setor minerário são diversos. Heider e Fonseca (2022), elencam uma série de questões que representam demandas atuais no setor.

A evolução de soluções capazes de englobar tantas mudanças se faz essencial, e uma das alternativas para o setor pode ser o desenvolvimento de um programa geometalúrgico, uma abordagem multidisciplinar que permite conhecimento detalhado das características técnicas do minério, auxiliando na previsibilidade do comportamento desses materiais, bem como na tomada de decisão e visão estratégica da vida do empreendimento.

Conforme destaca Fernandes (2013), durante a avaliação de recursos minerais é comum a construção de modelos de blocos com estimativa dos teores *in situ* do depósito em estudo, mas outras informações, principalmente ligadas a performance de processo, como a recuperação mássica, recuperação metalúrgica, energia requerida por moinhos, dentre outros, na maioria das vezes não são incorporadas a esse modelo devido ao fato de que, no passado, os materiais não apresentavam complexidade como atualmente e performavam bem com o mínimo de informação antecipada como guia para as equipes operacionais.

Na mina objeto desse trabalho, localizada no Quadrilátero Ferrífero, os diferentes tipos litológicos de itabiritos apresentam comportamento distintos como resposta ao processamento, o que dificulta a previsibilidade de variáveis de processo como a recuperação mássica, já que as litologias possuem diferentes tipos de contaminantes que podem impactar nessa variável. Sendo assim, um dos estágios de um programa geometalúrgico é analisar as relações existentes entre as diferentes variáveis estudadas e incorporar as que forem de interesse no modelo de blocos para que sejam utilizadas desde a etapa de planejamento de lavra.

O objetivo geral desse estudo é utilizar de técnicas estatísticas multivariadas para analisar as relações existentes entre variáveis de teor obtidas em um programa de geometalurgia

em fase inicial. Para isso, foi proposto a identificação de domínios geometalúrgicos através de agrupamento de dados e de desenvolvimento de modelos preditivos através de regressão para a recuperação mássica da concentração de itabiritos.

2. OBJETIVOS

2.1 Objetivo geral

Apresentar contribuições iniciais à modelagem geometalúrgica de uma mina de ferro aplicando técnicas de estatística multivariada em um banco de dados de um projeto de geometalurgia em desenvolvimento.

2.2 Objetivos específicos

Para cumprir o objetivo geral, foram estabelecidos e desenvolvidos os seguintes objetivos específicos:

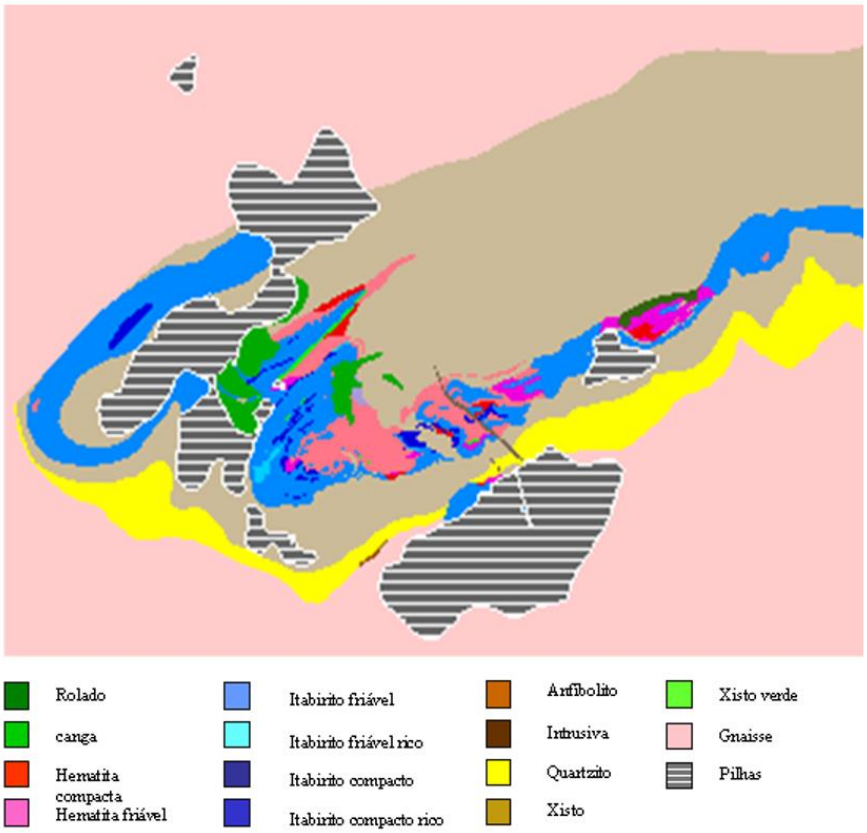
- Realizar análise descritiva das variáveis de estudo;
- Aplicar a análise de componentes principais (PCA) como técnica multivariada para redução de dimensionalidade do banco de dados e identificação das variáveis mais representativas;
- Aplicar o algoritmo de agrupamento k-means para visualização de padrões com base nas componentes principais;
- Apresentar a estatística descritiva e fundamentos estatísticos a partir dos agrupamentos obtidos;
- Desenvolvimento de modelos de regressão multivariados específicos para cada domínio identificado e para o conjunto total de dados visando estimar a variável geometalúrgica de interesse, a recuperação mássica;
- Apresentar recomendações e sugestões para os próximos passos e análises da modelagem geometalúrgica.

3. REVISÃO BIBLIOGRÁFICA

3.1 Descrição do empreendimento

A área de estudo está localizada na parte nordeste do Quadrilátero Ferrífero, região amplamente conhecida pela produção de minério de ferro. A mineralização é constituída por formações de ferro do Pré-Cambriano afetadas por muitos eventos tectônicos e por um processo de intemperismo profundo que provocou mudanças nas estruturas rochosas originais e na mineralogia. A mineralização está associada com a Formação Cauê e apresenta diferentes tipos de hematitas e de itabiritos como materiais predominantes, além de tipos menores que incluem os materiais detríticos conhecidos como canga e rolados. A Figura 1 mostra os domínios geológicos do objeto de estudo.

Figura 1 - Mapa geológico da mina em estudo



Fonte: Autoria própria

3.1.1 Operações e materiais

A mina produz diferentes tipos de materiais, que são submetidos a processos distintos de adequação conforme a qualidade, como apresentado na Tabela 1.

Tabela 1 - Materiais presentes na mina, processos de adequação e produtos

Material	Processo de adequação	Produto
HC	Peneiramento e britagem	HM1
HCS	Fração -12+1,00 mm: peneiramento, britagem	HM3
	Fração -1,00 mm: concentração da fração fina	Concentrado
Itabiritos	Peneiramento, britagem, moagem por prensa de rolos e concentração do material -1,00 mm	Concentrado

Fonte: Autoria própria

O produto final da mina é a produção de um blend a partir dos três produtos dos diferentes materiais, que servirá como matéria prima para os processos siderúrgicos subsequentes. Atualmente, esse blend é composto por 25% de HM1, 15% de HM3 e 60% de concentrado. O desafio da rotina operacional se encontra em produzir um material que atenda aos requisitos de qualidade exigidos para etapa seguinte.

Dentre os materiais presentes na mina, o itabirito destaca-se por apresentar maior complexidade em seu processamento, devido à variabilidade na qualidade e às diferenças no comportamento metalúrgico. Em função dessa complexidade e de sua representatividade para a lavra este trabalho será direcionado exclusivamente à análise do itabirito. A escolha visa garantir maior coerência nos resultados estatísticos, uma vez que a avaliação conjunta de diferentes tipos litológicos poderia mascarar padrões relevantes para o entendimento geometalúrgico deste material em particular.

3.2 Geometalurgia

3.2.1 Definição e objetivos

Nas últimas décadas muitas visões e definições sobre o significado do termo geometalurgia tem sido proposto, mas a maioria é pautada na definição de que ela funciona como uma ponte entre geologia e processamento mineral (Lishchuk et al, 2020). Para Schneider (2014), o termo é uma combinação de diversas áreas do conhecimento, que quando aplicadas em conjunto podem otimizar o planejamento de lavra, reduzindo riscos, maximizando os lucros e aproveitando melhor o recurso mineral. Gonçalves (2017) destaca que a geometalurgia pode ser entendida como uma ferramenta com abordagem interdisciplinar sendo as disciplinas de maior destaque durante um estudo geometalúrgico a geologia e mineralogia, o planejamento de mina, o processamento mineral e a geoestatística. De acordo com Michaux e O'Connor (2020), o termo apresenta mais significância do que simplesmente a relação entre testes metalúrgicos e processos mineralógicos, pois isso seria apenas parte de todo o programa. Para eles,

geometalurgia é uma apreciação da previsibilidade do comportamento de rochas, estando sujeitas a resultados de engenharia. Por fim, para Tiu et al. (2023), ela é uma ciência interdisciplinar que integra informações geológicas e metalúrgicas para o planejamento de operações mineiras.

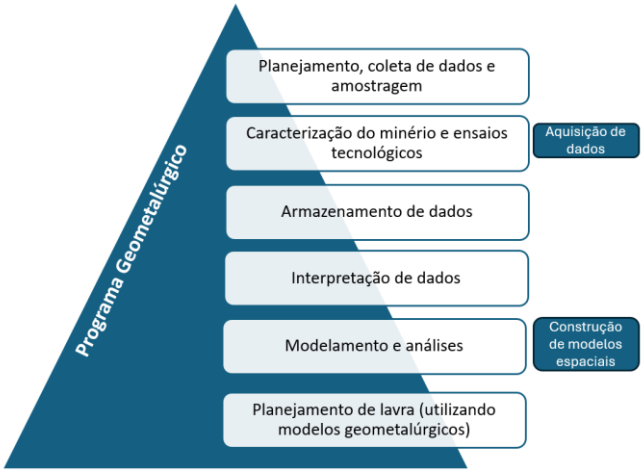
Conforme detalha Fonseca (2024), a abordagem geometalúrgica pode ser dividida em dois conceitos, i) estratégica: que utiliza ativamente todos os dados relevantes para ajudar a prever o desempenho do processamento e entender a variabilidade espacial do minério num cenário de longo prazo, buscando maximizar o valor econômico e integrando as áreas de conhecimento para decisões estratégicas e ii) tática: que visa otimização de processos específicos no curto e médio prazo e que se utiliza de amostragens frequentes e análises laboratoriais para tomada de decisões operacionais.

Considerando os objetivos e produto da ferramenta, Fernandes (2013), explica que a geometalurgia devolve um modelo de blocos que contém, além das estimativas dos teores *in situ*, estimativas das variáveis metalúrgicas de interesse para o depósito estudado. Conforme detalha Arroyo Ortiz (2014), a modelagem geometalúrgica é aplicada ao modelo de blocos, que recebem valores das variáveis estudadas e amostradas no depósito, auxiliada pela geoestatística, para inferência de variabilidade regional de algumas dessas grandezas. Segundo Gonçalves (2017), a integralização das disciplinas especialmente envolvidas na abordagem geometalúrgica visam a confecção de um modelo preditivo para descrever a variabilidade do depósito de minério de forma regional/local, o que gera ganhos e reduz riscos. A inserção de variáveis geometalúrgicas em um modelo de blocos (e.g. tempo de moagem e variação da dureza, grau de liberação, recuperações em peso e metalúrgica, dentre outras) possibilitam identificar previamente a variabilidade do comportamento no processamento mineral e produto dos diferentes tipos de materiais presentes em uma mina (Vieira, Mendonça e Costa, 2015).

Devido à demanda e aos complexos desafios enfrentados pelo setor mineral, se faz necessário a busca por inovações e soluções para otimização em todas as etapas da cadeia produtiva do minério de ferro. A geometalurgia surge, então, como uma área multidisciplinar com objetivo de contemplar as mudanças nos tempos atuais, caracterizando-se pela integração dos parâmetros geológicos com os de metalurgia na busca de maior eficiência nos processos metalúrgicos dos produtos ferrosos, representando uma ampla fonte de recursos informativos para os profissionais que atuam com minério de ferro nos seus diversos processos (Chemale Jr e Takehara, 2013).

O desenvolvimento de um programa geometalúrgico engloba todas as etapas de um projeto mineral, desde a exploração, passando pelo desenvolvimento, operação, até encerramento do negócio, se tornando uma prática constante e dinâmica para a gestão do empreendimento (Fonseca, 2024). A Figura 2 ilustra um esquema das etapas a serem seguidas para um projeto geometalúrgico.

Figura 2 - Principais fases de um programa geometalúrgico



Fonte: Adaptado de Fonseca, 2024.

3.2.2 Domínios geometalúgicos

Ao realizar os testes propostos pelo projeto geometalúrgico, se faz necessária avaliação dos resultados e análise detalhada dos dados, para que seja possível obter conclusões relevantes que serão implementadas no modelo de blocos. A análise inicial dos dados deve ser feita considerando os domínios geometalúrgicos. Esses domínios são similares aos geológicos e ajudam compreender a variabilidade metalúrgica em um depósito ao identificar amostras com propriedades parecidas como mineralogia, textura e composição, e que se espera que mostrem comportamentos semelhantes na performance do processamento mineral. Esse domínio nem sempre vai ser correlato ao geológico, considerando que diferentes tipos de rocha podem apresentar o mesmo comportamento metalúrgico e, portanto, serem agrupados no mesmo domínio geometalúrgico (Fernandes, 2013; Tiu et al., 2023).

A análise multivariada se refere a todas as técnicas estatísticas que simultaneamente analisam múltiplas medidas sobre indivíduos ou objetos sob investigação, e as técnicas deste tipo de análise proporcionam maior conhecimento auxiliando na tomada de decisão (Hair, 2009). Técnicas como análise de componentes principais (PCA) e análise de agrupamentos (CA) são duas abordagens multivariadas capazes de identificar a estrutura subjacente dos

conjuntos de dados e podem ser utilizadas para avaliar a similaridade da variável de interesse que se deseja inserir no modelo de blocos (Fernandes, 2013; Rajabinasab e Asghari, 2018), auxiliando, portanto, na definição de padrões e dos domínios geometalúrgicos.

Uma vez que os domínios estejam definidos, o passo seguinte é associar a eles os valores da variável metalúrgica de interesse. Isso pode ser feito utilizando medidas representativas, como a média ou mediana dessa variável dentro de cada domínio, ou ainda por meio de modelos empíricos que estimem seu comportamento com base em variáveis mais amplamente amostradas como o teor *in situ*, por exemplo (Fernandes, 2013).

3.3 Técnicas de análises estatísticas multivariadas

3.3.1 Análise de componentes principais (PCA)

A análise de componentes principais é uma abordagem estatística usada como uma técnica de redução de dimensionalidade linear. A PCA, do inglês, *principal componente analysis,* foi concebida no início do século por Karl Pearson (1901), tendo sido consolidada por Hottelling (1933, 1936) e é aplicada, geralmente, como a primeira técnica multivariada pois pode apresentar um papel importante na utilização de outras técnicas. O objetivo é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas com uma perda mínima de informação (Hair, 2009; Duarte, 1998).

Sendo assim, a técnica reduz o número de dimensões em grandes conjuntos de dados aos componentes principais que retêm a maior parte das informações originais. Os componentes principais são o resultado da transformação de variáveis potencialmente correlacionadas em um conjunto menor de variáveis. Estes componentes, por sua vez, apresentam algumas importantes propriedades: são variáveis independentes (ortogonais no espaço vetorial) entre si, cada componente principal é uma combinação linear de todas as variáveis originais, e são estimados com o propósito de reter, em ordem de estimação, o máximo de informação, em termos da variação total contida nos dados (IBM, 2023; Hongyu et al. 2015).

Os componentes principais são calculados na PCA de maneira que, o primeiro componente principal é a direção no espaço ao longo da qual os pontos de dados têm a variância mais alta, ou seja, quanto maior a variabilidade capturada no primeiro componente, maior será a informação retida do conjunto de dados original. O segundo componente principal, representa a segunda maior variância no conjunto de dados e não deve estar correlacionado com o

primeiro. Cada nova componente obtida, explica cada vez menos a variabilidade total, até chegar ao ponto de tal informação ser desprezível (IBM, 2023; Duarte, 1998).

Conforme pontua Duarte (1998), ao ser aplicada, é desejável obter um pequeno número de combinações lineares (componentes principais), que forneçam a maior parte das informações existentes no grande conjunto das variáveis originais. Normalmente, com as primeiras componentes pode-se explicar uma quantidade bastante alta de variação (por ex.: 80%).

De acordo com Duarte (1998) e Jolliffe (2002), o tratamento matemático aplicado aos dados pela técnica consiste, essencialmente, na determinação dos chamados autovalores e autovetores da matriz de covariância, ou correlação, considerando que a padronização prévia dos dados (média zero e variância unitária) é equivalente a usar a matriz de correlações. Cada autovalor calculado, normalmente representado por $\boldsymbol{\lambda_i}$ (**i**=1,2,...,*p*) corresponde à variância explicada pela *i-ésima* componente principal. Isso significa que ele indica quanto da variabilidade total dos dados está sendo capturada por um componente principal. Esses autovalores sempre aparecem em ordem decrescente ($\lambda 1 \geq \lambda 2 \geq ... \geq \lambda p$) pois o objetivo da técnica é condensar o máximo possível de informação (variância) nos primeiros componentes.

A Equação 1 a seguir, segundo Duarte (1998), explica que a variância total existente no conjunto de dados original pode ser expressa de três formas matematicamente equivalentes, representando a quantidade total de variância existente nas variáveis originais.

$$s^2total = \sum_{i=1}^{p} s_i^2 = \sum_{i=1}^{p} \lambda_i = TR(C) \qquad \text{Eq. 1}$$

Sendo:

s^2total = Variância total dos dados (de todas as variáveis)

s_i^2 = Variância explicada pela ***i-ésima*** variável original

λ_i = Autovalor associado à ***i-ésima*** componente principal (CP)

p = Número total de variáveis

$TR(C)$ = Operação traço (soma dos elementos da diagonal principal) para a matriz de covariâncias ou correlação (C).

Essa equivalência matemática assume que, na PCA, embora os dados sejam transformados em componentes principais, nenhuma variância é perdida, ela é apenas redistribuída entre o novo sistema de eixos, de forma a facilitar a interpretação e reduzir a dimensionalidade, quando necessário.

A medida da eficiência de uma componente principal em explicar a variabilidade contida nos dados originais, é dada pela proporção da variância total associada à essa componente, como representada na Equação 2.

$$\frac{\lambda_i}{s^2 total} = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \qquad \text{Eq. 2}$$

Já os autovetores são os coeficientes associados à cada variável original, que formam as combinações lineares. Para fins de análise exploratórias, variáveis com maior valor absoluto de autovetor são as mais influentes e explicam mais a variabilidade contida em determinado componente principal. É a partir desses valores que se pode obter os *scores de cada* componente, que são as coordenadas de cada amostra no novo espaço transformado pela PCA (Duarte, 1998).

A operação matemática para obtenção das matrizes diagonais dos autovalores, bem como dos autovetores é chamada de diagonalização (detalhada em Jolliffe, 2002 e Hongyu, et al. 2015) e a decisão sobre usar matriz de covariância ou de correlação, depende dos dados originais. Quando as variáveis estão na mesma unidade e apresentam variâncias semelhantes, a matriz de covariância é a mais adequada, pois preserva a magnitude original da variabilidade dos dados. No entanto, quando as variáveis possuem escalas diferentes ou apresentam grande diferença nas variâncias, recomenda-se o uso da matriz de correlação, que padroniza os dados, pois caso contrário as variáveis com maior variância tendem a dominar as primeiras componentes principais (Jolliffe, 2002).

Conforme explica Hongyu (2015) e Hair (2009), a análise de componentes principais não é muito recomendada quando se tem mais variáveis do que unidades amostrais, pois pode ser que tenha perda de informação ao reduzir o número de variáveis, fazendo com que a matriz de covariância ou correlação fique instável.

3.3.2 Agrupamento (clusterização) por k-means

O agrupamento é uma técnica estatística multivariada que faz parte do grupo de aprendizado não supervisionado, isto é, não há a necessidade de rótulos pré-definidos nem conhecimento prévio dos grupos, sendo uma abordagem que tem por objetivo agrupar amostras em classes baseadas em similaridade dos dados. O objetivo do agrupamento, também chamado de clusterização é identificar subclasses homogêneas dentro de um conjunto de dados heterogêneo (Rajabinasab e Asghari, 2018).

O conjunto de dados considerado para o agrupamento é uma matriz de dados multivariados, que contém os valores das variáveis pertencente a cada objeto (as amostras) a serem agrupados. A Equação 3 descreve a matriz, denominada X, onde a entrada x_{ij} representa o valor da *j-ésima* variável no objeto *i* (Everitt et al. 2011).

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} \\ x_{21} & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & \cdots & x_{ij} \end{pmatrix} \qquad \text{Eq. 3}$$

Para aplicação da técnica é necessário a escolha de um algoritmo de agrupamento, que determinará como será feita a divisão dos casos nos grupos distintos. Todos os métodos propostos são fundamentados na ideia de distância ou similaridade entre as observações e definem a pertinência dos objetos a cada cluster segundo aquilo que cada elemento tem de similar em relação a outros pertencentes do grupo (PUC-RJ, 2015).

Existem diversos tipos de algoritmos, e para este trabalho optou-se por um método de clusterização não hierárquico, o algoritmo k-means, reconhecido por sua eficácia, simplicidade e facilidade de implementação. O k-means classifica os dados em *k* grupos pré-definidos, determinando um ponto central (ou centróide), que nesse caso, se caracteriza por ser um ponto no espaço, para cada grupo. Cada objeto é atribuído ao cluster cujo centróide é mais próximo, de acordo com critérios de distância, geralmente a distância euclidiana (Jain, 2010).

A avaliação do ajuste de cada agrupamento pode ser feita por meio da soma dos erros quadráticos dentro de cada grupo (*Sum of Squares of Errors - SSE*), Equação 4. Este erro é calculado como a soma das diferenças quadradas entre cada objeto e o centróide do grupo ao qual ele foi atribuído. O algoritmo busca minimizar esse erro, realocando objetos iterativamente até alcançar a melhor divisão possível. (Almeida et al.).

$$SSE(grupo) = \sum_{xi \in grupo} dist\ (xi - c_{grupo})^2 \qquad \text{Eq. 4}$$

Sendo:

xi = Cada objeto do grupo em questão

c_{grupo} = Centróide do grupo em questão

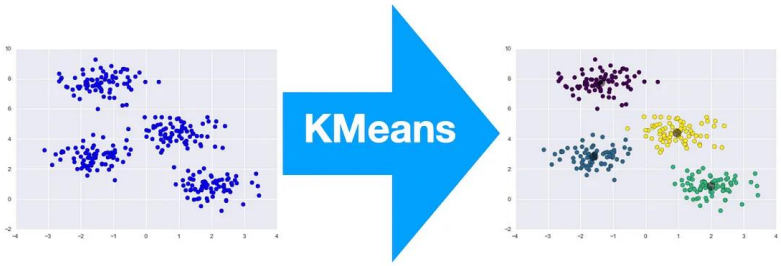
$Dist$ = Medida de distância utilizada

Pode-se dizer que o algoritmo processa o agrupamento da seguinte forma (Rajabinasab e Asghari, 2018; PUC-RJ, 2015).

- Inicialmente, os dados são divididos em *k* grupos pré-definidos pelo usuário, alocando os objetos nos *clusters*;
- Para cada objeto é calculada a distância em relação ao centróide de cada *k cluster* e, posteriormente, cada objeto é designado para o *cluster* cujo centróide é mais próximo.
- O algoritmo repete as iterações (etapa anterior), formando novos centróides, até que a função do erro não mude significativamente, ou seja, a variância dentro agrupamento seja mínima e entre os grupos seja máxima, fazendo com que nenhum objeto seja realocado.

A Figura 3 exemplifica como um conjunto de dados se apresenta após aplicação do agrupamento por k-means.

Figura 3 - Conjunto de dados antes e depois do agrupamento



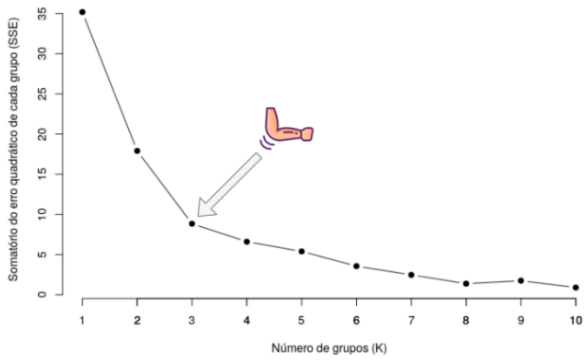
Fonte: Barros, 2021

Na combinação de utilização de técnicas estatísticas, Jolliffe (2002) apresenta as vantagens de utilizar PCA precedido da análise de agrupamentos, dentre elas se destacam a redução de dimensionalidade e a possibilidade de utilizar gráficos em 2D para representar os grupos através das componentes principais.

Existem métodos quantitativos que auxiliam na determinação do número de clusters a ser utilizado no emprego da técnica. Para este estudo, optou-se pelo emprego da abordagem denominada método do cotovelo.

O método do cotovelo usa o somatório do erro total dos grupos em cada agrupamento para determinar o número de *k*. Então, a ideia deste método é realizar o agrupamento para diferentes números de grupos e plotar um gráfico do número de *k versus* somatório do erro quadrático (SSE), como ilustrado na Figura 4. Matematicamente falando, busca-se uma quantidade de agrupamentos em que a soma dos quadrados intra-clusters (ou do inglês *within-clusters sum-of-squares* – WCSS) seja a menor possível, e que o aumento de grupos não resulta em melhorias significativas no algoritmo (Temporal, 2019).

Figura 4 - Método do cotovelo aplicado a um conjunto de dados



Fonte: Almeida et al. (2020)

Ao realizar um agrupamento pelo método k-means, é fundamental avaliar a qualidade da clusterização, garantindo que os grupos formados sejam coerentes e úteis para a interpretação dos dados. Uma das métricas mais utilizadas para essa avaliação é o coeficiente de silhueta (*Silhouette Score*), que mede o quão bem cada instância se ajusta ao seu próprio cluster em comparação com clusters vizinhos. Essa métrica combina dois aspectos essenciais: a coesão interna (intracluster), que indica o quão semelhantes estão os pontos pertencentes a um mesmo cluster, e a separação entre grupos (intercluster) que indica o quão distintos ou afastados estão os grupos, oferecendo uma visão abrangente do desempenho da clusterização.

O coeficiente de silhueta para um ponto i é calculado pela fórmula representada pela Equação 5.

$$s(i) = \frac{b_i - a_i}{\max(a_i, b_i)} \qquad \text{Eq. 5}$$

Onde:

a_i = Distância intracluster (média das distâncias para pontos do mesmo cluster)

b_i = Distância intercluster (média das distâncias para pontos do cluster mais próximo)

O valor de $s(i)$ varia entre -1 e 1. Valores próximos de 1 indicam clusters densos e bem separados, valores próximos de 0 indicam clusters sobrepostos e valores negativos sugerem que a alocação do ponto ao cluster pode estar incorreta. No gráfico de silhueta, cada barra representa o coeficiente de silhueta de um ponto, agrupado por cluster. Quanto mais largas e altas as áreas positivas, melhor definida está a separação entre os grupos (Melo, 2023; Kalimara, 2023).

3.3.3 Avaliação estatística e comparativa entre os agrupamentos

Para avaliar estatisticamente a qualidade de agrupamentos, há estudos na literatura, como o de Santos (2021), que utilizam de análises estatísticas de variância para observar se as

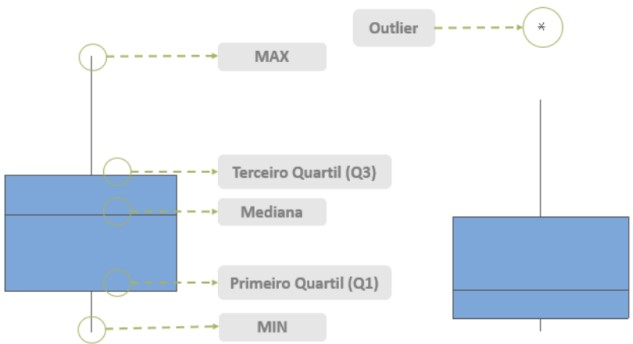
diferenças entre os agrupamentos são estatisticamente significantes. Para definição de qual teste é mais apropriado, é necessário analisar a rejeição ou não da hipótese nula de normalidade.

Os testes de normalidade são testes que examinam formalmente se um conjunto de observações de uma determinada variável possuem distribuição normal. Existem vários testes que examinam se distribuição dos dados é normal, entre os mais famosos estão o de Shapiro-Wilk, Anderson-Darling e Kolmogorov-Smirnov, sendo que todos eles possuem a hipótese alternativa que os dados possuem distribuição diferente da normal. Se o valor-p de um determinado teste for menor do que 0,05, para um nível de confiança de 95%, significa que os dados não seguem uma distribuição normal, caso contrário, se o valor-p for maior que 0,05 não se rejeita a hipótese nula de normalidade (Salemi, 2013).

Se a hipótese nula de normalidade não for rejeitada, utiliza-se de análise de variância (ANOVA), e análise multivariada de variância (MANOVA), que são testes paramétricos que comparam as médias de três ou mais grupos independentes para determinar se existe uma diferença estatisticamente significativa entre as médias da população correspondente. O teste de ANOVA é utilizado para observar diferenças existentes em relação à média de uma variável, enquanto com análise MANOVA é possível observar se grupos se diferem quanto às médias de diversas variáveis. Caso a rejeição da hipótese nula de normalidade se confirme, a alternativa é a aplicação de testes não paramétricos, como Kruskal-Wallis (MINITAB).

Para visualizar e descrever estatisticamente uma população, é comumente utilizado os gráficos de Boxplot, que permitem a observação da distribuição e variabilidade de um conjunto de dados. Trata-se de um diagrama de caixa construído utilizando as referências de valores mínimos e máximos, primeiro e terceiro quartil, mediana e outliers da base de dados (Rodrigues, 2021). A Figura 5 mostra como eles são construídos e as informações relevantes que são possíveis de se obter com a ferramenta.

Figura 5 - Informações de um gráfico boxplot



Fonte: Rodrigues, 2021

MIN é o valor mínimo do conjunto de dados, mas que respeita um certo limite. O primeiro quartil é a base da caixa retangular, uma demarcação que distingue que 25% dos dados estão abaixo e os outros 75% dos dados estão acima desse valor. A mediana ou segundo quartil demarca o valor central do conjunto de dados, 50% dos dados são maiores que esse valor e os outros 50% dos dados são menores. Já o terceiro quartil é o topo da caixa retangular, linha que informa o número que fica entre os 75% valores mais baixos e 25% valores mais altos. MAX é o valor máximo do conjunto de dados, que respeita um limite calculado. E, por fim, os *outliers* são valores discrepantes que estão fora de um intervalo definido, altos demais ou baixos demais (Noleto, 2022).

3.3.4 Regressão linear multivariada

Segundo Fávero e Belfiori (2017), a técnica de regressão linear oferece, prioritariamente, a possibilidade de que seja estudada a relação entre uma ou mais variáveis explicativas, que se apresentam na forma linear, e uma variável dependente quantitativa. Um modelo geral de regressão linear pode ser descrito pela Equação 6, que representa um modelo de regressão linear múltipla.

$$Yi = a + b_1.X_{1i} + b_2.X_{2i} + \cdots + b_k.X_{ki} + u_i \qquad \text{Eq. 6}$$

Sendo:

Yi = Variável dependente quantitativa em estudo

a = Constante ou coeficiente linear

X_j (j = 1, 2, ..., k) = Variáveis explicativas

b_j (j = 1, 2, ..., k) = coeficientes das variáveis explicativas

u = Termo de erro ou resíduo (diferença entre o valor real de *Y* e valor previsto de *Y* por meio do modelo para cada observação). Esse valor deverá ser o menor possível, caso muito alto, significa que outras variáveis que não foram consideradas afetam *Y*.

i = Cada uma das observações da amostra em análise (i = 1, 2, ..., *n*, em que *n* é o tamanho da amostra)

Ao realizar uma análise de regressão multivariada, é fundamental selecionar adequadamente as variáveis independentes que serão incrementadas no modelo. A escolha deve

considerar a existência de correlações relevantes entre as variáveis explicativas e a variável resposta. Variáveis que apresentarem baixa correlação com a resposta tendem a ter pouco poder preditivo, podendo, portanto, ser descartadas do modelo, a fim de evitar complexidade desnecessária e multicolinearidade (Pereira et al, 2017).

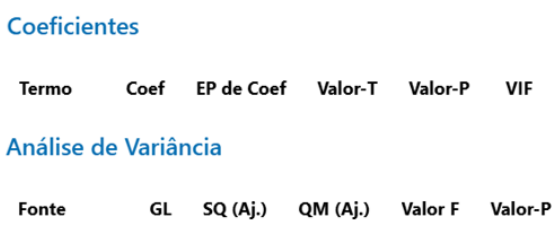
Ao interpretar um modelo de regressão, conforme destaca Fávero e Belfiori (2017), os *outputs* gerados são: estatísticas da regressão, tabela de análise de variância (*analysis of variance*, ou ANOVA), tabela de coeficientes da regressão e tabela de resíduos. A Tabela 2 descreve os parâmetros da estatística de regressão, e a Figura 6 descreve outros parâmetros importantes a serem analisados em um ajuste de modelo.

Tabela 2 - Parâmetros estatísticos de uma regressão

Parâmetro estatístico	Descrição
R^2 (coeficiente de determinação)	O R^2 explica se a relação entre as variáveis é forte ou fraca. Quanto mais perto de 1 for o resultado, mais forte será a relação.
R^2 -ajustado	O R^2 ajustado fornece uma melhor comparação entre dois modelos ou entre um mesmo modelo com tamanhos de amostras diferentes.
R(pred)	O R(pred) mede o quão bem o modelo prevê novos dados, ou seja, é a capacidade preditiva. Quando mais próximo de 1, melhor a previsão em dados que não foram usados no ajuste.
S	S mede o erro médio entre os valores previstos e os valores reais.

Fonte: Adaptado de Pereira, et al. (2017)

Figura 6 - Parâmetros a serem analisados em um modelo de regressão



Fonte: Autoria própria

Onde:

- GL (grau de liberdade da regressão) indica o número de variáveis independentes utilizadas no modelo. O grau de liberdade do resíduo é calculado como o total de

observações menos o número de variáveis independentes menos um. E o total de GL é número total de observações menos um;

- SQ (soma dos quadrados dos desvios totais), representa a dispersão da variação aleatória de *Y* em relação a sua média total;
- QM (quadrado médio) é obtido pela divisão da soma de quadrados por seus respectivos graus de liberdade;
- O Valor F mede se o modelo como um todo é significativo, comparando a variância explicada pelo modelo com a variância do erro.
- Os coeficientes são os números pelos quais as variáveis da equação serão multiplicadas;
- O erro padrão do coeficiente, EP, mede o grau de precisão com que o modelo estima o valor desconhecido do coeficiente;
- O valor-T é a divisão do coeficiente pelo erro padrão;
- O valor-P fornece a significância estatística de cada parâmetro a ser considerado no modelo de regressão. Busca saber quais variáveis estão relacionadas com a variável dependente. Para que exista significância, o valor-P deve ser < 0,05 para um nível de confiança de 95%;
- O VIF (fator de inflação da variância) mede a multicolinearidade entre as variáveis independentes. Um valor de VIF elevado pode indicar que a variável pode estar altamente correlacionada com outras, distorcendo os resultados da regressão.

A literatura relata estudos que utilizaram a técnica na modelagem de variáveis geometalúrgicas. Fernandes (2013), apresenta um exemplo prático de aplicação de um modelo empírico, utilizando regressão linear múltipla para estimar a recuperação mássica (variável resposta) de uma mina de fosfato, a partir dos teores do ROM (*Run of Mine*) como variáveis independentes. Já Rodovalho et al. (2024), utilizou regressão linear multivariada para estimar a recuperação mássica de um depósito de bauxita a partir da definição de domínios geometalúrgicos.

Neste trabalho, após a definição de grupos por meio da aplicação de Análise de Componentes Principais (PCA) e do agrupamento não hierárquico pelo algoritmo k-means, busca-se identificar, em cada grupo, quais variáveis originais mais contribuem para as componentes principais que explicam a maior parte da variância dos dados. Com base nessas variáveis, pretende-se avaliar modelos de regressão específicos para cada domínio e para o conjunto total de dados, com objetivo de prever a recuperação mássica.

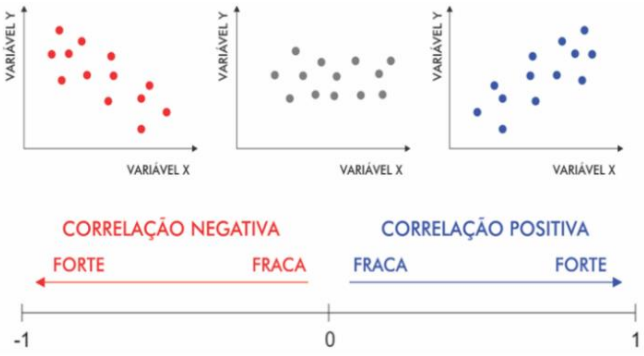
Mesmo sendo amplamente usado, o Modelo de Regressão Linear apresenta algumas limitações. As principais incluem: a) linearidade, pois quando a relação entre as variáveis não é linear, o modelo pode falhar em capturar padrões complexos; b) distribuição normal dos erros, pois caso contrário o modelo pode não ser apropriado e c) heterocesdasticidade, pois quando a variância dos erros não é constante, as estimativas dos parâmetros podem ser ineficientes (Koritiaki, 2024).

3.3.4.1 Correlação de Pearson

Os modelos de regressão linear podem ser testados com a combinação das diversas variáveis disponíveis, entretanto, como citado no tópico acima, a escolha das variáveis independentes pode ser facilitada ao considerar a existência de correlações relevantes entre as variáveis explicativas e a variável resposta.

Essa correlação pode ser checada através da correlação de Pearson, que por sua vez, é uma ferramenta estatística que mede a relação linear entre as variáveis, indicando a força e a direção da associação. Os valores da relação variam de -1 a +1, que indicam correlação negativa e positiva, respectivamente, como demonstra a Figura 7. Uma baixa correlação não implica necessariamente ausência de correlação. Pode ser que o padrão da correlação seja do tipo não-linear (Reis, [s.d.])

Figura 7 - Intensidade e direção da correlação de Pearson



Fonte: Costa et al. (2022)

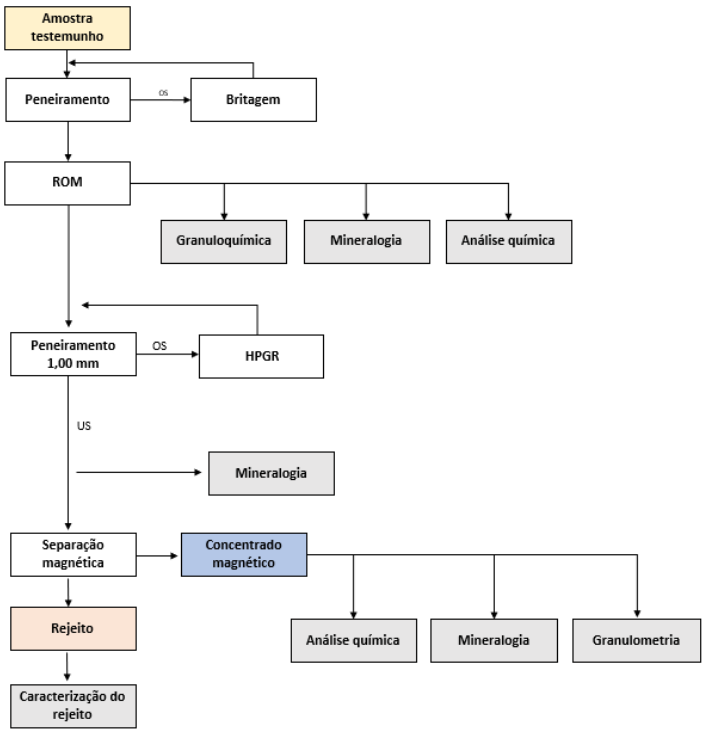
4. METODOLOGIA

4.1 Definição de fluxograma geometalúrgico

Para este programa, como etapa preliminar, elaborou-se um fluxograma geometalúrgico, apresentado na Figura 8, que engloba as principais etapas de processamento dos Itabiritos. 31 amostras do material, denominadas GMT-000, foram obtidas e selecionadas a partir de furos de sonda georreferenciados, com diferentes profundidades e classificadas previamente em diferentes litologias, de acordo com parâmetros já existentes relativos a esses furos. As amostras foram submetidas a caracterização mineralógica e ensaios tecnológicos nessa primeira campanha do programa de geometalurgia.

À medida que os testes foram sendo realizados, procedeu-se ao armazenamento dos dados obtidos, para análises posteriores.

Figura 8 - Fluxograma geometalúrgico dos Itabiritos



Fonte: Autoria própria

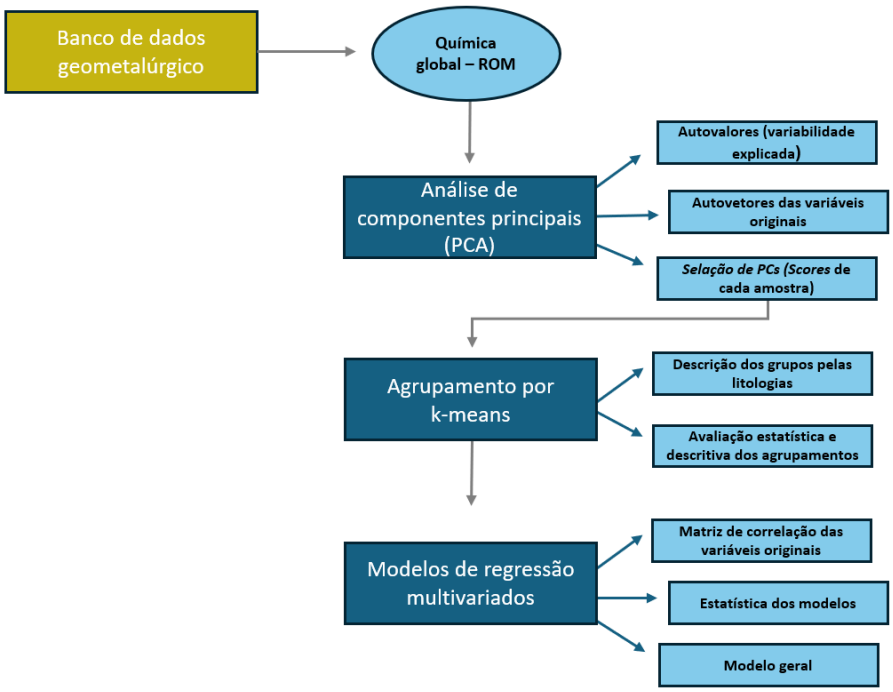
As variáveis utilizadas nas técnicas apresentadas nesse trabalho foram obtidas pelas análises químicas do ROM. Optou-se pela utilização dessas variáveis pelas análises já estarem concluídas e armazenadas no banco de dados geometalúrgico. Além disso, um dos objetivos do

presente estudo é testar modelos de regressão para a recuperação em peso em função dos teores *in situ.*

4.2 Análises estatísticas multivariadas

Essa seção é dedicada à metodologia aplicada aos dados selecionados para as análises estatísticas. Elas foram realizadas de acordo com o fluxo apresentado na Figura 9.

Figura 9 - Fluxograma das análises aplicadas ao subconjunto de dados selecionados



Fonte: Autoria própria

4.2.1 Análise de componentes principais

Para análise dos componentes principais, o subconjunto selecionado foi submetido ao teste de variância, através da função VAR.P (Excel), para comparar a "variabilidade" entre as variáveis escolhidas como entrada na aplicação da técnica. Através do teste foi possível observar como elas apresentaram alta variabilidade entre si, apesar de possuírem mesma escala (porcentagem), como apresentado na Tabela 3. Sendo assim, optou-se pela utilização da matriz de correlação, em que os dados foram normalizados (média zero e variância unitária).

Tabela 3 - Variância do subconjunto de variáveis originais

Fe_gl	SiO2_gl	Al2O3_gl	P_gl	Mn_gl	CaO_gl	TiO2_gl	MgO_gl	PPC_gl
93,785	375,392	0,481	0,005	0,013	10,360	0,001	7,742	26,449

Fonte: Autoria própria

O teste também contribui para a tomada de decisão quanto à exclusão de variáveis com baixa variabilidade ou relevância prática limitada, que podem gerar ruídos ao realizar modelagens estatísticas. Essas variáveis, por apresentarem valores quase constantes, podem exercer influência desproporcional na análise de componentes principais (PCA) devido à padronização, que atribui a elas o mesmo peso das demais, independentemente de sua contribuição real para a variância total do conjunto de dados. Levando isso em consideração, optou-se pela exclusão da variável TiO_2, que apresenta variação praticamente nula. Para além disso, é uma variável que não apresenta impacto significativo na variável resposta estudada.

A análise foi realizada pelo software MINITAB *Statistical Software* 21.1.0, onde foram obtidos os autovalores das principais componentes que explicam maior parte da variância, os autovetores (*loadings*) das variáveis originais em cada componente principal, e os valores dos *scores* (coordenadas das amostras no novo espaço) para cada componente principal, que serão aplicados na técnica de clusterização. A análise também gera um gráfico Biplot das principais componentes.

4.2.2 Agrupamento (clusterização) por k-means

Os dados de entrada para a clusterização são formados pelos *scores* de cada amostra nas componentes principais. Essa escolha é vantajosa porque reduz a dimensionalidade do conjunto de dados a ser aplicado no algoritmo, eliminando redundâncias e ruídos, além de concentrar a variância mais relevante em poucas variáveis independentes entre si, facilitando a formação de clusters mais significativos, e permitindo a visualização em 2D dos grupos formados.

Os valores desses *scores* obtidos na técnica anterior, passaram por uma renormalização. Segundo Jolliffe (2002), se é desejável dar pesos iguais para cada informação a ser empregada no agrupamento, então é importante encontrar as componentes principais e dar pesos iguais para cada uma delas que mais explicam a variabilidade. Caso essa variância não seja discrepante entre as componentes, não há mudanças significativas no resultado do k-means.

O algoritmo foi testado para k variando de 1 a 10, para que fosse possível encontrar, através de um dos métodos quantitativos (método do cotovelo) o valor ideal de grupos para os dados utilizados nessa abordagem.

Para a aplicação da técnica, recorreu-se a bibliotecas da linguagem Python. A biblioteca *scikit-learn* (Pedregosa et al., 2011) foi utilizada para importar e aplicar o algoritmo de k-means. A manipulação e organização das bases de dados foi realizada com o uso da biblioteca *pandas*

(McKinney, 2010), enquanto a biblioteca *Numpy* (Harris et al., 2020) deu suporte às operações numéricas envolvidas no processo. Para a visualização dos clusters formados, bem como para a construção do gráfico do método do cotovelo, empregou-se a biblioteca Matplotlib (Hunter, 2007), permitindo a geração de gráficos de dispersão com as amostras agrupadas por cores distintas.

Para estimar a qualidade de cada agrupamento formado, foi gerado um gráfico de silhueta, também utilizando os recursos Python, empregando novamente a biblioteca *scikit-learn* (sklearn) para importar e aplicar os coeficientes de silhueta.

4.2.3 Avaliação estatística e comparativa entre os agrupamentos

Para saber a significância dos grupos obtidos, foram aplicados testes de avaliações estatísticas a fim de verificar as diferenças, existentes ou não, em termos de recuperação mássica e dos teores globais (variáveis empregadas nas análises) para cada cluster. Para tanto, as recuperações mássicas de cada agrupamento foram submetidas à teste de normalidade de Anderson-Darling, onde apresentaram não rejeição da hipótese nula de normalidade, significando que as distribuições das recuperações podem ser consideradas normais. Posteriormente, foram submetidas ao teste paramétrico ANOVA (análise de variância), para testar a hipótese nula de que as médias de recuperação mássica são iguais entre os clusters.

Em relação às variáveis de teor global, apenas Fe_gl e SiO_2_gl apresentaram distribuição normal, sendo então, empregadas no teste de análise multivariada de variância (MANOVA).

Para visualizar as distribuições e variabilidades entre os agrupamentos gerados a partir do algoritmo de k-means, foram desenvolvidos gráficos de boxplot para todas as variáveis de teor utilizadas na técnica de análise de componentes principais e para a recuperação mássica, que é a variável geometalúrgica estudada.

Todas as análises relativas à essa metodologia foram realizadas através do Software MINITAB *Statistical Software* 21.1.0.

4.2.4 Modelo de regressão linear

Para desenvolvimento dos modelos de regressão, primeiro foram elaboradas matrizes de correlação com as variáveis originais utilizadas nas análises anteriores e a recuperação mássica, considerando que a análise de componentes principais não garante que as variáveis que mais explicam a variabilidade estrutural dos dados também expliquem bem a variável alvo.

Foram desenvolvidas matrizes de correlação para cada cluster encontrado e para o conjunto de dados total. Todas as matrizes foram obtidas através do MINITAB *Statistical Software* 21.1.0, na ferramenta de estatísticas básicas.

Diversos modelos foram testados com diferentes combinações das variáveis relacionadas aos teores das amostras, que geraram modelos significativos do ponto de vista de uma análise exploratória dos dados. Os testes foram realizados através do MINITAB *Statistical Software* 21.1.0, mantendo o nível de confiança de 95%. Foi possível realizar a análise dos resultados através dos parâmetros e gráficos dos resíduos gerados para cada modelo de regressão.

5. RESULTADOS E DISCUSSÕES

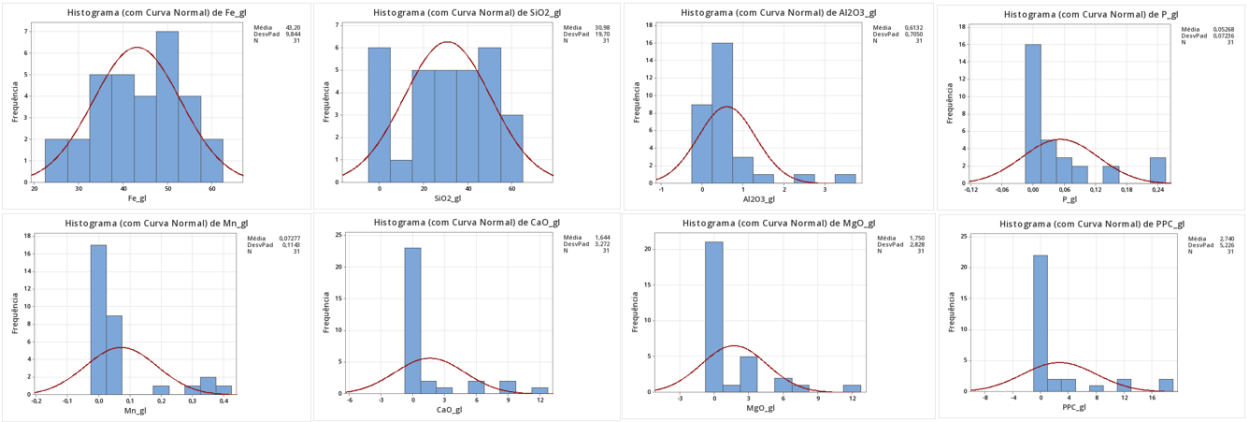
Antes da aplicação das técnicas, realizou-se uma análise estatística descritiva de cada uma das variáveis de teor, apresentadas pela Tabela 4 e pelos histogramas na Figura 10, através do MINITAB *Statistical Software* 21.1.0.

Tabela 4 - Estatística descritiva das variáveis de teor

Variáveis	Média	DesvPad	Mínimo	Q1	Mediana	Q3	Máximo
Fe_gl	43,20	9,84	24,11	34,50	44,01	51,38	60,58
SiO2_gl	30,98	19,70	0,70	19,32	29,69	49,43	64,57
Al2O3_gl	0,61	0,70	0,18	0,22	0,38	0,62	3,43
P_gl	0,05	0,07	0,00	0,01	0,01	0,07	0,24
Mn_gl	0,07	0,11	0,00	0,01	0,02	0,06	0,40
CaO_gl	1,64	3,27	0,01	0,02	0,07	1,31	11,81
MgO_gl	1,75	2,83	0,00	0,10	0,51	2,63	11,62
PPC_gl	2,74	5,23	0,00	0,13	0,35	1,43	18,84

Fonte: Autoria própria

Figura 10 - Histograma das variáveis de teor



Fonte: Autoria própria

5.1 Análise de componentes principais

A Tabela 5 mostra os resultados da análise de PCA para o subconjunto de dados em questão, utilizando a matriz de correlação.

Tabela 5 - Autovalores e variância explicada pelas principais componentes

Parâmetro	PC1	PC2
Autovalores (λ)	5,0464	1,4571

Variância	0,631	0,182
Variância acumulada	63,1%	81,3%

Fonte: Autoria própria

Juntas, as duas primeiras componentes explicam 81,3% da variância do conjunto de dados, o que é um valor consideravelmente aceito e, portanto, elas foram selecionadas para as análises seguintes. Ao analisar os autovetores, considerou-se um valor superior a 0,4 para as variáveis mais significativas em uma componente principal. Fica evidenciado que a PC1 é fortemente influenciada positivamente por Mn, CaO, MgO e PPC, enquanto PC2 é mais sensível às variações de SiO_2 (positiva) e Fe (negativa), com contribuição de Al_2O_3. Isso indica que a variação dessas variáveis originais contribui muito para a variação total explicada por cada componente principal.

Os valores obtidos dos autovetores para cada variável original são apresentados na Tabela 6, bem como os *scores* de todas as 31 amostras são mostrados na Tabela 7.

Tabela 6 - Carga (autovetores) das variáveis originais em cada componente principal

Variáveis	**PC1**	**PC2**
Fe_gl	0,192	-0,696
SiO2_gl	-0,375	0,418
Al2O3_gl	-0,090	-0,423
P_gl	0,327	-0,164
Mn_gl	0,420	0,193
CaO_gl	0,425	0,177
MgO_gl	0,417	0,206
PPC_gl	0,420	0,153

Fonte: Autoria própria

Tabela 7 - *Scores* das amostras após aplicação da PCA

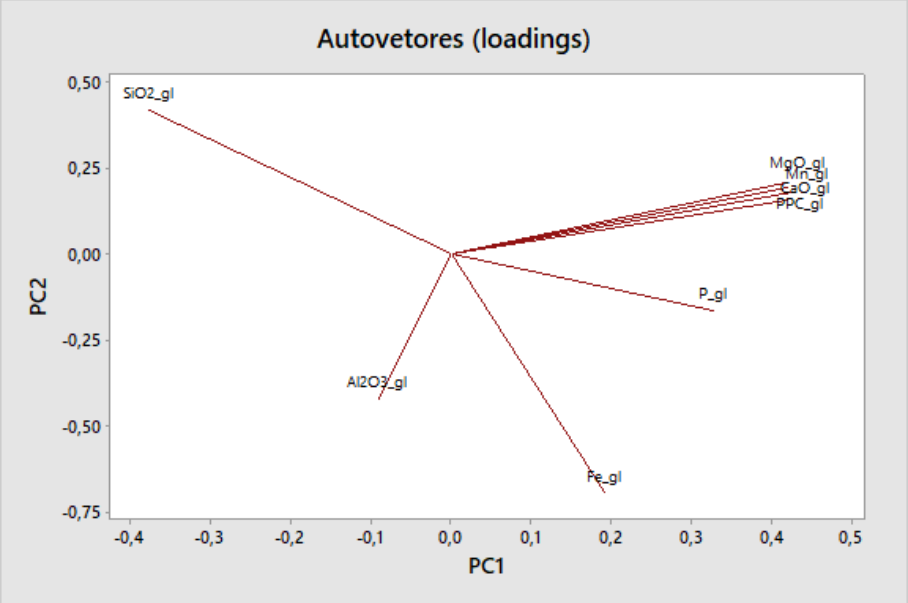
Amostra	**PC1**	**PC2**	**Amostra**	**PC1**	**PC2**
GMT-001	3,462	-1,145	GMT-017	2,262	-1,935
GMT-002	-2,003	2,014	GMT-029	-0,936	-0,636
GMT-003	-1,445	0,753	GMT-032	-1,738	1,231
GMT-004	-1,535	1,020	GMT-033	-1,603	0,684
GMT-005	-1,285	0,655	GMT-034	-1,171	-0,277
GMT-006	-1,903	1,660	GMT-035	-1,022	-1,915
GMT-007	5,019	1,594	GMT-036	-1,109	-1,904

GMT-008	-1,445	1,133	GMT-037	-0,691	-1,826
GMT-009	-1,490	0,990	GMT-038	0,291	-0,778
GMT-010	0,417	-0,253	GMT-039	-0,728	0,174
GMT-011	0,358	-0,322	GMT-040	-0,407	-0,842
GMT-012	4,233	0,599	GMT-058	-0,712	-1,123
GMT-013	4,228	-0,228	GMT-059	-1,272	0,300
GMT-014	-0,366	-1,867	GMT-062	-0,541	-0,642
GMT-015	6,052	1,726	GMT-063	-1,178	-0,156
GMT-016	-1,741	1,316			

Fonte: Autoria própria

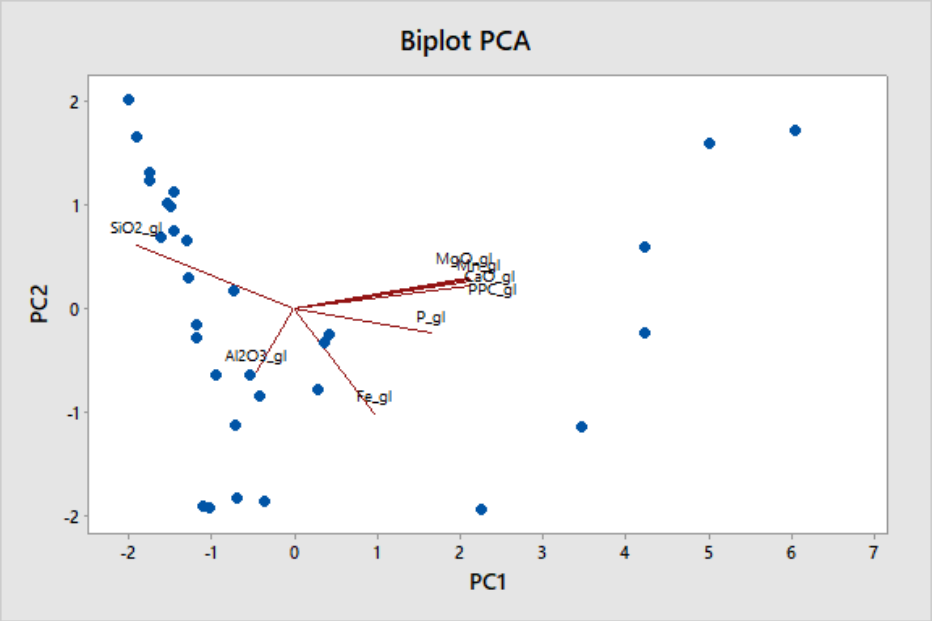
As Figuras 11 e 12 representam os gráficos de autovetores das variáveis originais e o biplot para as componentes principais, contendo as amostras plotadas de acordo com seus *scores* (as coordenadas no novo espaço) juntamente com os autovetores, respectivamente. Os gráficos também revelam a forte e positiva correlação entre MgO, CaO, Mn e PPC, que influenciam mais a PC1, e uma correlação moderada desses elementos com P. Além disso, observa-se a correlação negativa do SiO_2 com Fe e P, além de MgO, CaO, Mn e PPC.

Figura 11 - Gráfico de autovetores das variáveis originais



Fonte: Autoria própria

Figura 12 - Gráfico biplot para a PCA

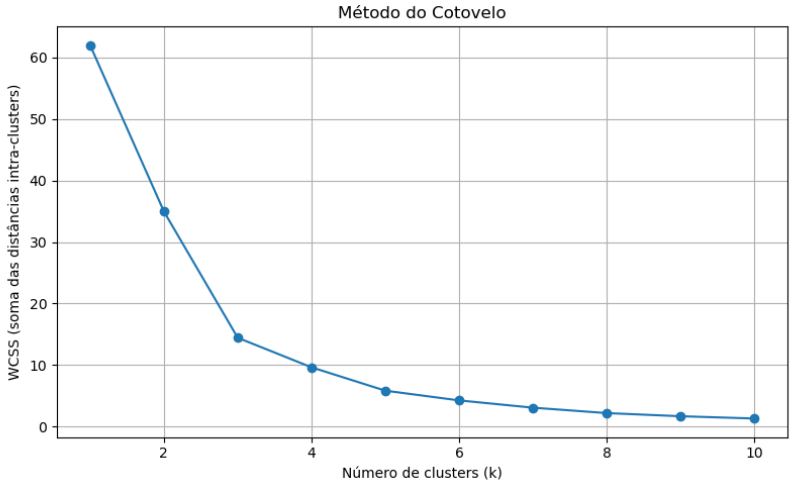


Fonte: Autoria própria

5.2 Agrupamento por k-means

Através da aplicação do método do cotovelo, empregando o algoritmo com *k* variando de 1 a 10, encontrou-se um número de clusters iguais a 3, conforme revela a Figura 13. A partir desse ponto, nota-se que o aumento da quantidade de clusters não resultaria em melhoras significativas do agrupamento.

Figura 13 - Método do cotovelo aplicado para o agrupamento



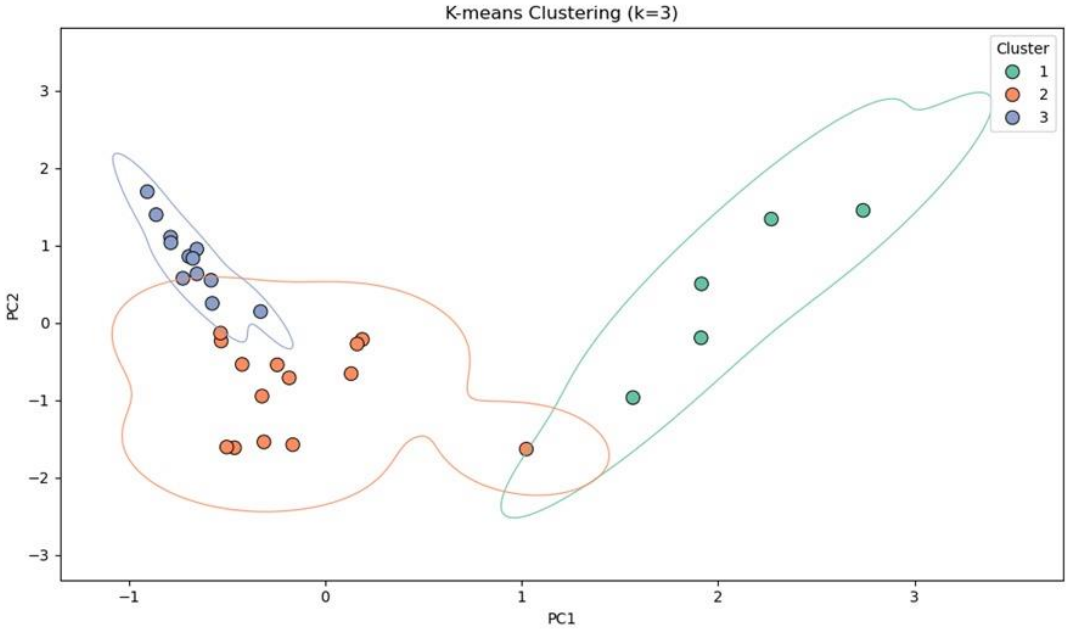
Fonte: Autoria própria

O algoritmo agrupou os valores normalizados dos *scores* obtidos através da PCA em três grupos distintos. Apesar da limitação da quantidade de amostras, que não favorece uma clara separação intra-cluster, foi possível identificar padrões relevantes entre os grupos formados. A visualização mostra que:

- O cluster 1, formado pelas variáveis mais explicativas de PC1, é composto exclusivamente por amostras do tipo litológico IDOL, que se destacam por apresentarem teores mais elevados de PPC. As variações que definem esse cluster estão mais associadas às variáveis que tem maior carga em PC1. No entanto, a dispersão das amostras dentro do cluster é mais explicada pela variabilidade do Fe, indicando que, embora não tenha sido uma variável determinante na formação do grupo, ele ainda exerce forte influência na heterogeneidade interna do cluster 1;
- O cluster 2 demonstra uma influência combinada de variáveis com cargas negativas, tanto em PC1 quanto em PC2, associadas à Fe, SiO_2 e Al_2O_3. Apesar de mais heterogêneo e com maior variabilidade, há predominância das amostras do tipo litológico IF e IC. As amostras com maiores concentrações do contaminante Al_2O_3 foram agrupadas nesse cluster. Uma amostra específica, a GMT-017, apresentou características suficientes para ser agrupada no cluster 2 segundo o algoritmo, embora pertença à litologia IDOL, o que justifica o seu maior distanciamento em relação ao centróide do grupo 2. Essa exceção pode ser explicada por estar associada, na mina, a uma região de contato entre os itabiritos dos tipos IDOL e IC.
- O cluster 3, mais influenciado pela parte positiva de PC2, revela predominância de amostras dos tipos litológicos IC2 e IFP, comumente associadas a teores mais elevados de SiO_2 e mais baixos de Fe, relevando um comportamento de acordo com as variáveis de maior influência nessa componente principal. Na análise de variância das variáveis originais para esse cluster, percebeu-se que a variabilidade interna também é mais explicada por Fe e SiO_2.

A Figura 14 revela o resultado do agrupamento, enquanto a Tabela 8 descreve cada amostra por litologia e cluster.

Figura 14 - Gráfico de dispersão 2D do agrupamento por K-means



Fonte: Autoria própria

Tabela 8 - Descrição das amostras por litologia e cluster

Amostra	Litologia	Cluster	Amostra	Litologia	Cluster
GMT-001	IDOL	1	GMT-017	IDOL	2
GMT-002	IC2	3	GMT-029	IF	2
GMT-003	IC2	3	GMT-032	IFP	3
GMT-004	IC2	3	GMT-033	IFP	3
GMT-005	IC2	3	GMT-034	IF	2
GMT-006	IC2	3	GMT-035	IF	2
GMT-007	IDOL	1	GMT-036	IF	2
GMT-008	IC2	3	GMT-037	IF	2
GMT-009	IC2	3	GMT-038	IF	2
GMT-010	IC	2	GMT-039	IF	3
GMT-011	IC2	2	GMT-040	IF	2
GMT-012	IDOL	1	GMT-058	IF	2
GMT-013	IDOL	1	GMT-059	IF	3
GMT-014	IF	2	GMT-062	IC	2
GMT-015	IDOL	1	GMT-063	IF	2
GMT-016	IC2	3			

Fonte: Autoria própria

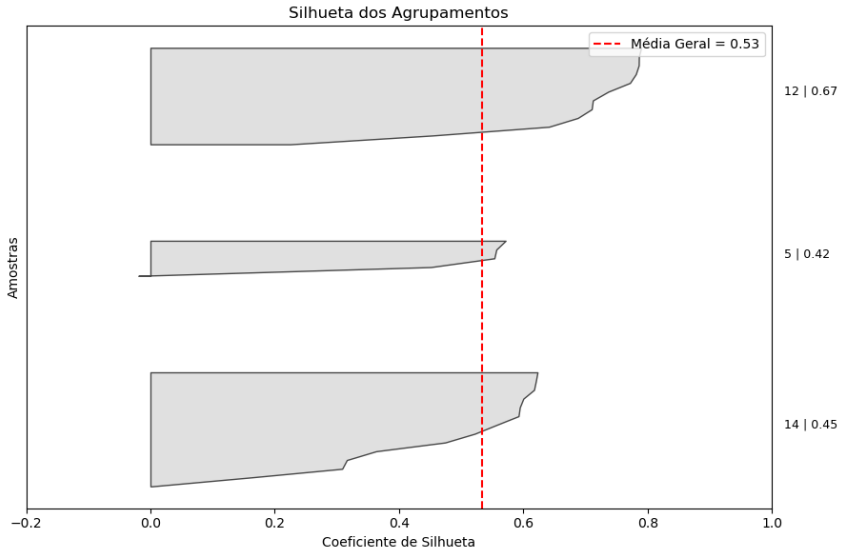
Através da Figura 14, observa-se uma certa sobreposição entre os clusters 2 e 3 no espaço das duas componentes principais, o que pode ser um indicativo que as amostras

localizadas nessa região compartilham características composicionais semelhantes. A divisão entre os grupos não é totalmente clara nessa região, o que é relativamente esperado devido ao fato de o agrupamento representar litologias de um depósito mineral, onde diversas transições e eventos geológicos nos materiais ocorrem gradualmente. Além disso, indica uma limitação da aplicação do algoritmo nesse conjunto de dados, considerando que um dos objetivos é maximizar a distância entre os clusters.

Ao analisar as amostras do cluster 3 que estão sobrepostas no cluster 2, constatou-se que elas pertencem ao tipo litológico IF. Elas apresentam teores de ferro mais baixo em relação às amostras do grupo 2, embora quando comparadas com as demais do grupo 3, se destacam com teores mais elevados desse mesmo elemento.

Para ter uma estimativa da qualidade do agrupamento, foi desenvolvido um gráfico de silhueta. A medida da silhueta demonstra melhor qualidade do agrupamento quando estão mais próximas de 1, e em contrapartida, revela que algumas amostras, ao apresentarem coeficiente negativos, podem não pertencer necessariamente ao grupo em que foram alocadas. A Figura 15 apresenta o resultado da silhueta para cada amostra dentro de cada cluster formado. Apesar da estimativa ter apresentado um valor médio de cerca de 0,5, o que significa um agrupamento regular, existem pontos com valores que chegam a 0,8. Além disso, apenas um ponto mostrou valor de silhueta negativa.

Figura 15 - Gráfico de silhueta para o agrupamento por k-means

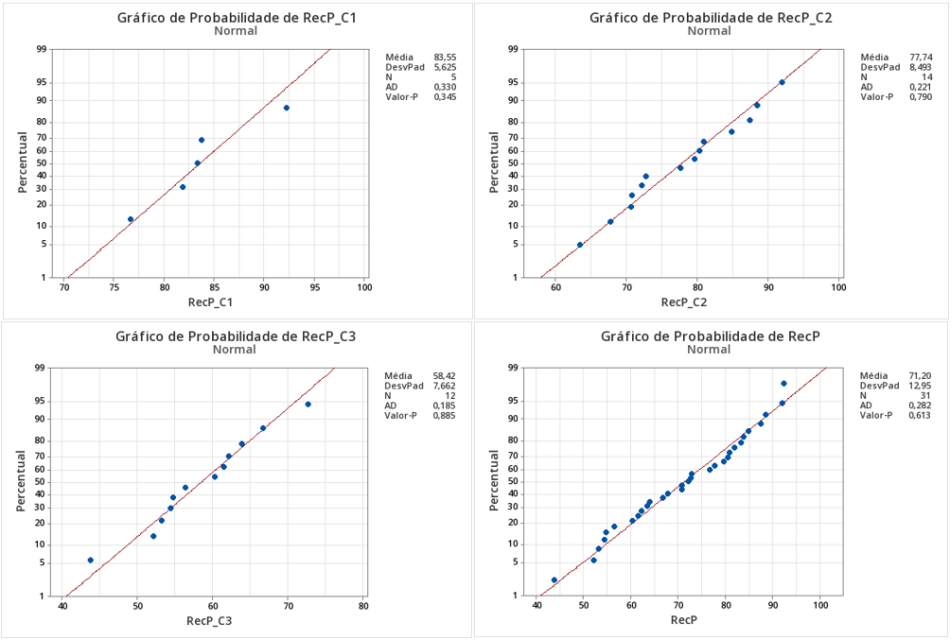


Fonte: Autoria própria

5.3 Avaliação estatística e comparativa entre os agrupamentos

Para a variável geometalúrgica de interesse, foram realizados testes de normalidade, que como citados no item 4.2.3, apresentaram não rejeição da hipótese nula de normalidade, significando que as distribuições das recuperações podem ser consideradas normais, com valores-p maiores que 0,05 para cada cluster, e de forma geral, considerando todas as amostras. A Figura 16 mostra o resultado do teste de normalidade para as recuperações de cada grupo e geral.

Figura 16 - Testes de normalidade das recuperações mássicas por cluster e geral



Fonte: Autoria própria

A partir da não rejeição da hipótese nula de normalidade, prosseguiu-se ao teste paramétrico da análise de variância (ANOVA). A análise indicou que há diferença estatisticamente significativa entre os clusters obtidos pelo k-means em relação à média da variável resposta, apresentando valor-p menor que 0,05. Esse resultado sugere que os agrupamentos identificados podem refletir variações reais no comportamento, em termos de recuperação mássica, das amostras. A Tabela 9 revela o resultado da análise.

Tabela 9 - Teste paramétrico da análise de variância (ANOVA) para recuperação mássica

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Cluster	2	3320	1659,97	27,18	0,000
Erro	28	1710	61,08		
Total	30	5030			

Fonte: Autoria própria

Para as variáveis originais que apresentaram distribuição normal, Fe_gl e SiO_2_gl, realizou-se uma análise multivariada de variância (MANOVA), que revelou valores-p menores que 0,05 para cada variável, rejeitando, portanto, a hipótese nula. Isso constata as diferenças estatisticamente significativas, em relação à média dessas duas variáveis, para os diferentes agrupamentos formados. A Tabela 10 mostra os resultados obtidos pela MANOVA.

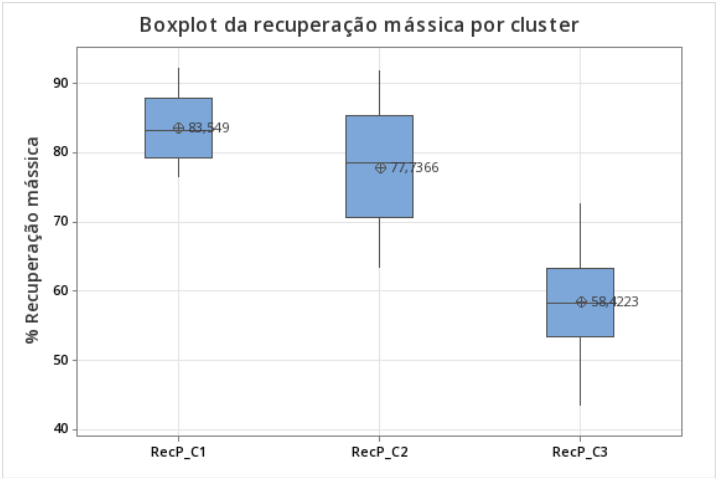
Tabela 10 - Teste paramétrico da análise multivariada de variância (MANOVA) para Fe_gl e SiO2_gl

Critério	Estatística de teste	F	NUM GL	DENOM GL	Valor-P
De Wilks	0,03791	55,839	4	54	0,000
Lawley-Hotelling	10,52435	68,408	4	52	0,000
De Pillai	1,52525	44,978	4	56	0,000
De Roy	8,84463				

Fonte: Autoria própria

Ao analisar o gráfico boxplot gerado para a recuperação mássica, notou-se que o cluster 1 apresentou a maior média de recuperação, enquanto o cluster 2 mostrou a maior variabilidade da variável em questão. A Figura 17 exibe o resultado obtido.

Figura 17 - Boxplot da recuperação mássica por cluster



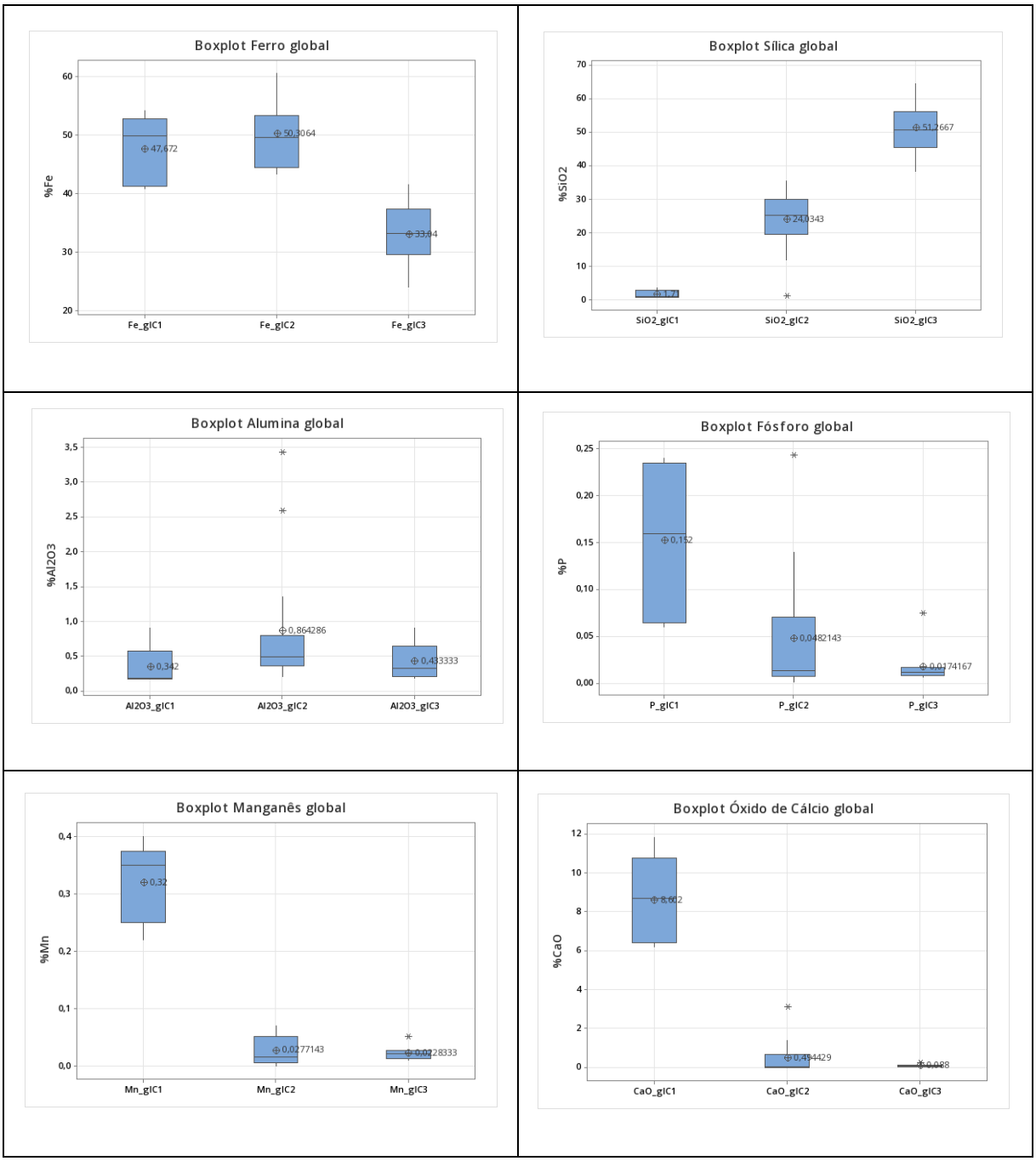
Fonte: Autoria própria

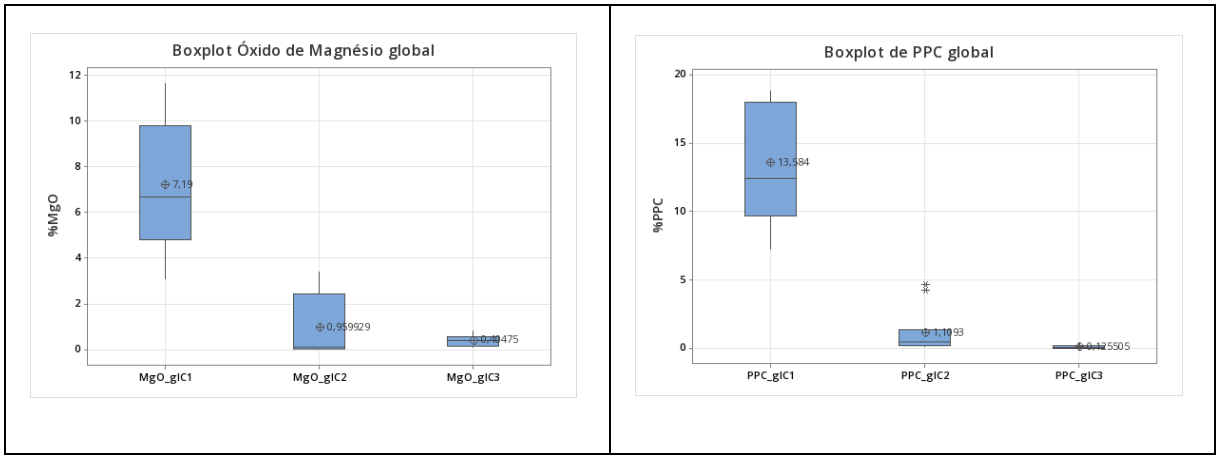
Avaliando as variáveis separadamente por cluster, foi possível observar a média e variabilidade dos teores através dos gráficos boxplot. Para a variável Fe_gl, observou-se maior média no cluster 2, e maior variabilidade no cluster 1, enquanto o cluster 3 apresentou a média mais baixa. Para SiO_2_gl, foi possível notar maior média para o cluster 3, conforme já esperado

pelo resultado do agrupamento. O cluster 2 apresentou a maior média de Al_2O_3_gl, com dois valores discrepantes desse contaminante. O cluster 1 apresentou maior média e maior variabilidade para P_gl, Mn_gl, CaO_gl, MgO_gl e PPC_gl, confirmando os resultados devolvidos pelo algoritmo do k-means.

A Tabela 11 abaixo retrata os gráficos de boxplot gerados para cada variável original envolvida nas análises.

Tabela 11 - Boxplot das variáveis originais por cluster





Fonte: Autoria própria

5.4 Modelos de regressão linear

5.4.1 Matrizes de correlação

Ao total, 4 matrizes de correlação foram desenvolvidas para as análises dos modelos de regressão, uma para cada agrupamento e uma considerando o conjunto total dos dados.

A Figura 18 revela a matriz do primeiro cluster, sendo possível a identificação de correlações fortes entre recuperação mássica e os elementos químicos. A variável geometalúrgica correlacionou-se fortemente e positivamente com Fe_gl, SiO_2_gl e Al_2O_3_gl, e por outro lado, fortemente e negativamente com MgO_gl e PPC_gl, e PPC/Fe. Como a correlação entre Fe_gl e PPC_gl é muito forte, e ambas as variáveis independentes são explicativas para a variável dependente, para evitar multicolinearidade na exploração dos modelos, criou-se uma variável, PPC/Fe, que pode ser interpretada como a contaminação por teor metálico.

Figura 18 - Matriz de correlação do cluster 1

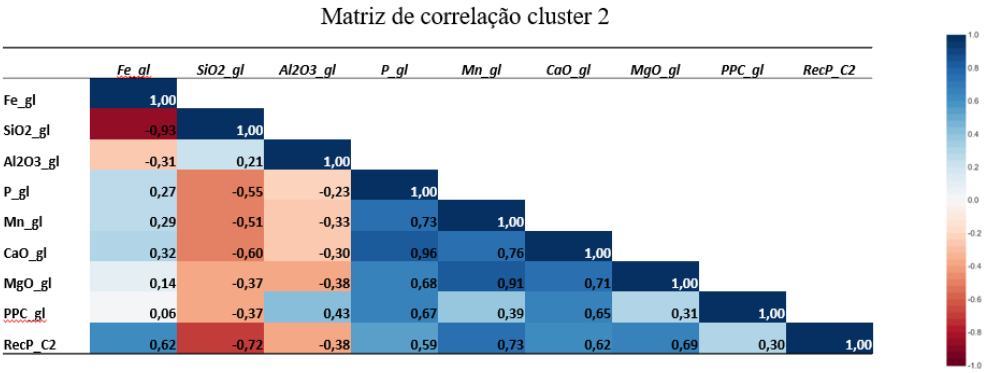
Matriz de correlação cluster 1

	Fe_gl	SiO2_gl	Al2O3_gl	P_gl	Mn_gl	CaO_gl	MgO_gl	PPC_gl	PPC/Fe	RecP_C1
Fe_gl	1,00									
SiO2_gl	0,33	1,00								
Al2O3_gl	0,58	0,92	1,00							
P_gl	0,59	0,24	0,56	1,00						
Mn_gl	-0,78	-0,11	-0,32	-0,67	1,00					
CaO_gl	-0,90	-0,16	-0,52	-0,84	0,76	1,00				
MgO_gl	-0,89	-0,66	-0,75	-0,41	0,69	0,65	1,00			
PPC_gl	-0,97	-0,55	-0,75	-0,54	0,67	0,82	0,95	1,00		
PPC/Fe	-0,99	-0,45	-0,66	-0,53	0,72	0,84	0,94	0,99	1,00	
RecP_C1	0,84	0,78	0,87	0,45	-0,59	-0,64	-0,98	-0,95	-0,91	1,00

Fonte: Autoria própria

A Figura 19 representa a matriz de correlação do cluster 2, onde foi possível identificar a forte e negativa correlação da recuperação mássica com SiO_2_gl, e a forte e positiva correlação com Mn_gl. Além disso, a recuperação mostrou correlação moderada e positiva com Fe_gl, CaO_gl e MgO_gl.

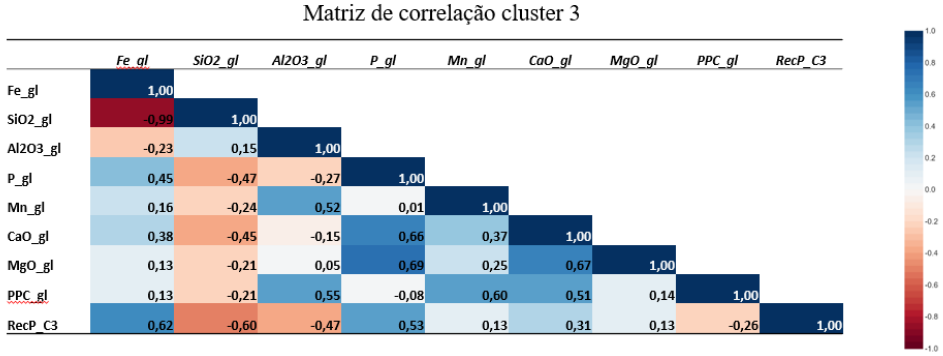
Figura 19 - Matriz de correlação do cluster 2



Fonte: Autoria própria

A Figura 20 revela a matriz de correlação do cluster 3, onde foi possível identificar que a variável geometalúrgica de interesse não mostrou correlação forte com nenhum dos elementos químicos globais, apenas correlações moderadas com Fe_gl (positiva) e SiO_2_gl (negativa).

Figura 20 - Matriz de correlação do cluster 3



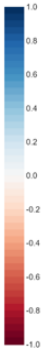
Fonte: Autoria própria

A Figura 21 mostra a matriz de correlação do subconjunto total de dados, onde foi possível identificar que a recuperação mássica apresentou correlação forte e positiva com Fe_gl, e forte e negativa com SiO_2_gl.

Figura 21 - Matriz de correlação do subconjunto total de dados

Matriz de correlação geral

	Fe_gl	SiO2_gl	Al2O3_gl	P_gl	Mn_gl	CaO_gl	MgO_gl	PPC_gl	RecP
Fe_gl	1,00								
SiO2_gl	-0,84	1,00							
Al2O3_gl	0,10	0,00	1,00						
P_gl	0,42	-0,64	-0,15	1,00					
Mn_gl	0,19	-0,68	-0,21	0,58	1,00				
CaO_gl	0,21	-0,70	-0,22	0,59	0,98	1,00			
MgO_gl	0,19	-0,67	-0,28	0,60	0,93	0,92	1,00		
PPC_gl	0,19	-0,69	-0,10	0,57	0,95	0,97	0,94	1,00	
RecP	0,85	-0,87	-0,04	0,61	0,45	0,47	0,46	0,43	1,00



Fonte: Autoria própria

5.4.2 Desenvolvimento do modelo de regressão – cluster 1

Para o cluster 1, apesar de modelos com MgO_gl terem explicado bem a variável geometalúrgica, optou-se pelos modelos explicados por PPC_gl, considerando que é este o elemento químico que determina o tipo litológico contido no agrupamento. Diversas combinações foram testadas, e o modelo que melhor se ajustou aos dados contém as variáveis PPC/Fe e SiO_2_gl como variáveis independentes. A Equação 7 representa a equação de regressão encontrada para o agrupamento, e a Figura 22 mostra os resultados estatísticos obtidos para o modelo.

$$RecP_{C1} = 88{,}55 - 29{,}92\ PPC/Fe\ + 2{,}173\ SiO_{2_gl} \qquad \text{Eq. 7}$$

Figura 22 - Estatísticas do modelo de regressão do cluster 1

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
0,606285	99,42%	98,84%	92,55%

Coeficientes

Termo	Coef	EP de Coef	Valor-T	Valor-P	VIF
Constante	88,55	1,09	81,11	0,000	
PPC/Fe	-29,22	2,52	-11,58	0,007	1,25
SiO2_gl	2,173	0,280	7,75	0,016	1,25

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	**2**	**125,808**	**62,9039**	**171,13**	**0,006**
PPC/Fe	1	49,293	49,2929	134,10	0,007
SiO2_gl	1	22,070	22,0701	60,04	0,016
Erro	**2**	**0,735**	**0,3676**		
Total	**4**	**126,543**			

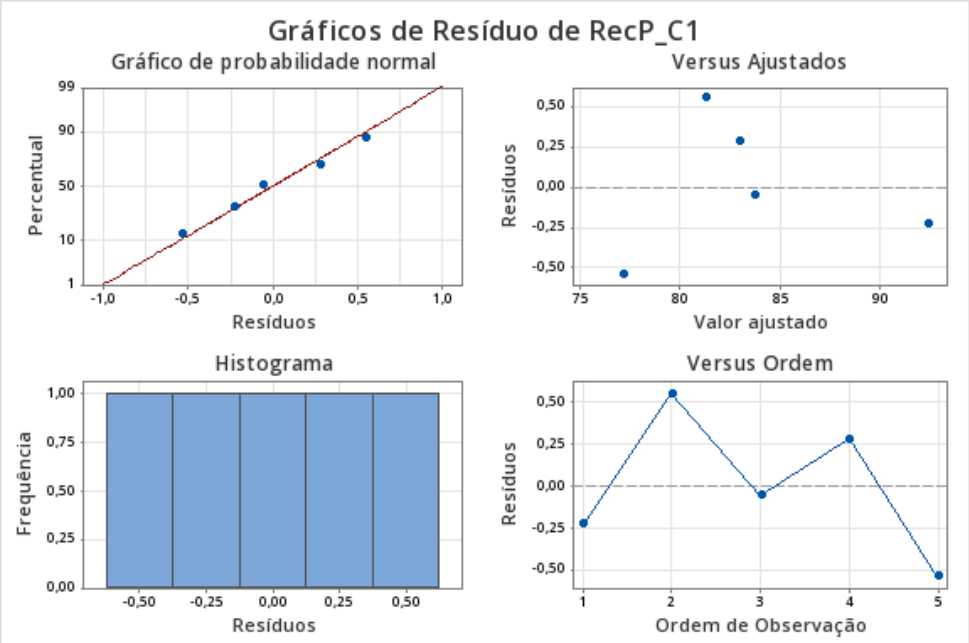
Fonte: Autoria própria

O coeficiente de determinação (R^2) de 99,42% indica que o modelo explica praticamente toda a variabilidade observada na variável resposta, a recuperação mássica. O R^2 ajustado de 98,84% reforça a qualidade do ajuste, considerando o número de variáveis no modelo. Já o R^2 preditivo de 92,55% demonstra que o modelo mantém boa capacidade de predição em novos dados.

Os coeficientes das variáveis independentes apresentaram valores-p inferiores a 0,05, evidenciando que ambas são estatisticamente significativas para explicar a recuperação. Além disso, os baixos valores de VIF indicam ausência de multicolinearidade, reforçando a estabilidade do modelo. Por fim, o valor F elevado obtido na análise de variância revela que a variância explicada pelo modelo é significativamente maior que a variância residual, confirmando a relevância estatística do ajuste.

Já a Figura 23, mostra os gráficos dos resíduos obtidos para o modelo. Apesar da pouca quantidade de dados, que limita histograma, o gráfico de probabilidade normal dos resíduos mostra que eles seguem aproximadamente uma distribuição normal, enquanto o gráfico de resíduos versus ajustados não revela formação de padrão, sendo distribuídos aleatoriamente ao redor de 0, indicando bom ajuste do modelo encontrado.

Figura 23 - Gráficos de resíduos do modelo de regressão do cluster 1



Fonte: Autoria própria

O tipo litológico presente no cluster 1, é o único que mostra correlações fortes com os elementos químicos contaminantes característicos dessa litologia. Apesar de apresentar bons resultados de recuperação mássica, é importante ressaltar que elementos como MgO, CaO, e PPC podem causar problemas relacionados ao tratamento de rejeitos, especialmente na questão de turbidez da água.

5.4.3 Desenvolvimento do modelo de regressão – cluster 2

Diversas combinações também foram testadas para o cluster 2. O único modelo que se mostrou promissor utiliza as variáveis Fe_gl e MgO_gl como preditoras, com ambas atuando positivamente no valor da variável resposta, mas não apresentou resultados tão bons quanto para o cluster 1. A Equação 8 revela o modelo de regressão para o agrupamento 2 e a Figura 24 mostra os resultados estatísticos obtidos.

$$RecP_{C2} = 33{,}9 + 0{,}798Fe_{_gl} + 3{,}884\,MgO_{_gl} \qquad \text{Eq. 8}$$

Figura 24 - Estatísticas do modelo de regressão do cluster 2

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
4,61263	75,04%	70,50%	62,86%

Coeficientes

Termo	Coef	EP de Coef	Valor-T	Valor-P	VIF
Constante	33,9	11,5	2,95	0,013	
Fe_gl	0,798	0,229	3,49	0,005	1,02
MgO_gl	3,884	0,960	4,05	0,002	1,02

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	**2**	**703,7**	**351,84**	**16,54**	**0,000**
Fe_gl	1	259,1	259,14	12,18	0,005
MgO_gl	1	348,1	348,13	16,36	0,002
Erro	**11**	**234,0**	**21,28**		
Total	**13**	**937,7**			

Fonte: Autoria própria

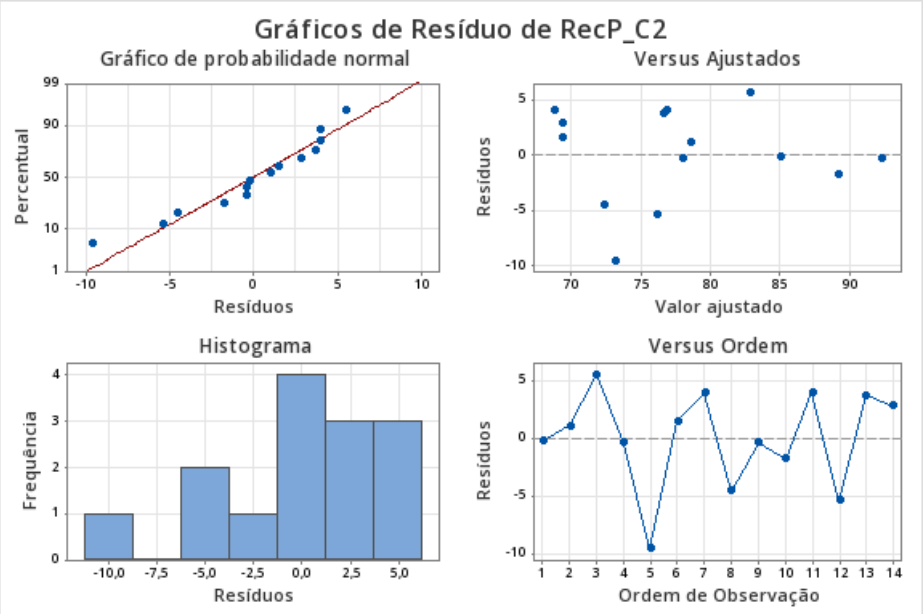
O coeficiente de determinação (R^2), indica que o modelo explica cerca de 75% da variabilidade observada para a variável resposta. O R^2 ajustado de aproximadamente 70% reforça que, embora o modelo explique uma variabilidade consideravelmente aceita, outras variáveis, que podem ser relacionadas à granulometria ou mineralogia do material, também irão

explicar as variações da recuperação mássica. O R^2 preditivo de 62,86% sugere uma capacidade preditiva moderada do modelo na predição em novos dados.

Os coeficientes das variáveis independentes apresentaram valores-p inferiores a 0,05, evidenciando que ambas são estatisticamente significativas para explicar a recuperação. Além disso, os baixos valores de VIF indicam ausência de multicolinearidade, reforçando a estabilidade do modelo. Apesar do valor F de aproximadamente 16 obtido na análise de variância não ser tão alto como o obtido para o cluster 1, ele ainda é considerado um valor forte e significativo, uma vez que apresenta valor p muito baixo, confirmando que o modelo de regressão é estatisticamente relevante.

A Figura 25, mostra os gráficos dos resíduos obtidos para o modelo. O gráfico de probabilidade normal dos resíduos mostra alguns pontos que se distanciam mais da reta, enquanto o gráfico de resíduos versus ajustados revela a aleatoriedade dos resíduos ao redor de 0. Já o histograma dos resíduos apresenta assimetria e má distribuição, comportamento que pode ser influenciado pelo número reduzido de observações, mas que também indica que o modelo não captura toda a variabilidade dos dados.

Figura 25 - Gráficos de resíduos do modelo de regressão do cluster 2



Fonte: Autoria própria

5.4.4 Desenvolvimento do modelo de regressão – cluster 3

Apesar da tentativa de ajuste de modelos de regressão para o agrupamento 3, os resultados obtidos não foram satisfatórios em termos de qualidade do ajuste (valores baixos de

R^2 e R^2 preditivo), significância estatística e análise dos resíduos. Isso pode estar relacionado a diferentes fatores, entre eles se destaca a possível influência de variáveis não incluídas no modelo, como características específicas do material. Variáveis relacionadas à mineralogia, granulometria ou outros fatores não mensurados, podem ter impacto significativo sobre a recuperação e não foram considerados nesta etapa. Além disso, o número reduzido de amostras pode comprometer o modelo, especialmente para esse grupo, que não apresentou correlações com a variável resposta tão fortes quanto os outros dois. Para os próximos testes exploratórios indica-se incluir variáveis de outras características descritivas do material, como granulometria e mineralogia, que são testes presentes no fluxograma geometalúrgico.

5.4.5 Desenvolvimento do modelo de regressão geral

Ao testar diferentes combinações para o banco de dados completo, sem separação por cluster, um bom ajuste de modelo de regressão foi encontrado contendo as variáveis SiO_2_gl e PPC_gl como preditoras, ambas atuando negativamente no valor da variável reposta. O resultado se torna representativo considerando a composição química dos itabiritos, pois SiO_2 é um contaminante presente nas amostras que compõem os clusters 2 e 3, e PPC é o contaminante que define o tipo litológico das amostras que compõem o cluster 1. A Equação 9 representa o modelo de regressão para todos os itabiritos e a Figura 26 mostra os resultados estatísticos obtidos.

$$RecP_{C2} = 95{,}99 - 0{,}7262\, SiO_{2_gl} - 0{,}838\, PPC_{_gl} \qquad \text{Eq. 9}$$

Figura 26 - Estatísticas do modelo de regressão geral

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
5,75702	81,55%	80,23%	77,32%

Coeficientes

Termo	Coef	EP de Coef	Valor-T	Valor-P	VIF
Constante	95,99	3,06	31,35	0,000	
SiO2_gl	-0,7262	0,0742	-9,79	0,000	1,93
PPC_gl	-0,838	0,279	-3,00	0,006	1,93

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	**2**	**4102,3**	**2051,13**	**61,89**	**0,000**
SiO2_gl	1	3178,5	3178,48	95,90	0,000
PPC_gl	1	298,3	298,26	9,00	0,006
Erro	**28**	**928,0**	**33,14**		
Total	**30**	**5030,3**			

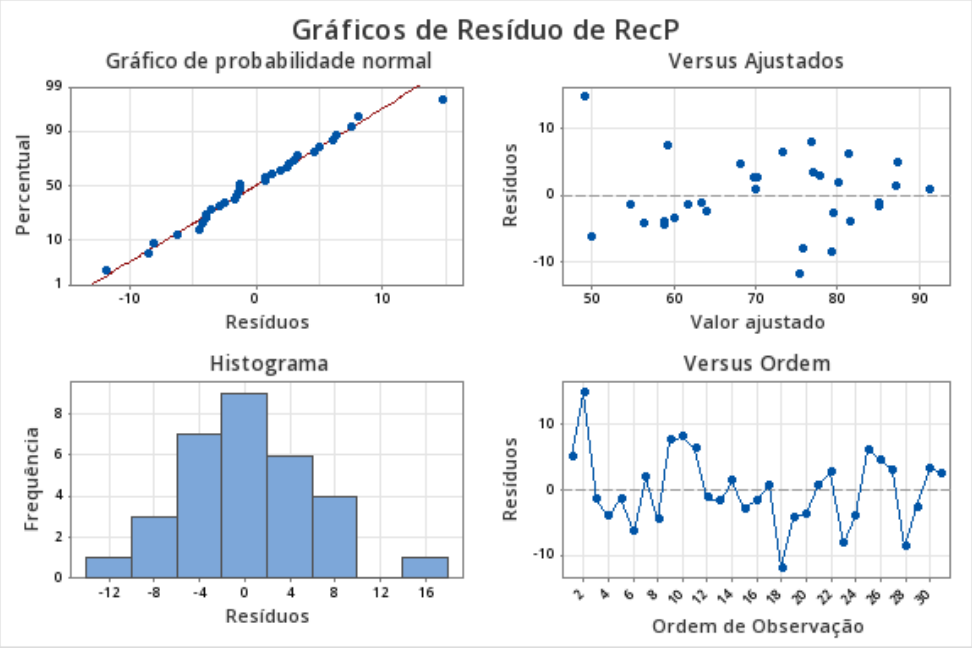
Fonte: Autoria própria

O coeficiente de determinação (R^2), indica que o modelo explica aproximadamente 82% da variabilidade observada para a variável resposta. O R^2 ajustado de 80,23%, reforça que, embora o modelo explique uma variabilidade consideravelmente aceita, pode ser que outras características do material, como granulometria e mineralogia, também irão explicar as variações da recuperação mássica. O R^2 preditivo de cerca de 77% sugere uma capacidade moderadamente alta do modelo na predição em novos dados, sendo adequado para análises exploratórias.

Os coeficientes das variáveis independentes apresentaram valores-p inferiores a 0,05, evidenciando que ambas são estatisticamente significativas para explicar a recuperação. Além disso, os baixos valores de VIF indicam ausência de multicolinearidade, reforçando a estabilidade do modelo. O valor F obtido na análise de variância é considerado um valor forte e apresenta valor p muito baixo, reforçando a relevância estatística do ajuste.

A Figura 27, mostra os gráficos dos resíduos obtidos para o modelo. O gráfico de probabilidade normal revela uma distribuição aproximadamente normal, com 2 pontos (amostras GMT-002 e GMT-029), apresentando diagnóstico de resíduo atípico. O gráfico de resíduos versus ajustados mostra a aleatoriedade dos dados, enquanto o histograma confirma que os resíduos tendem a seguir uma distribuição normal, sendo melhorado, desta vez, pela quantidade de dados inseridos na análise.

Figura 27 - Gráficos de resíduos do modelo de regressão geral considerando todos os itabiritos



Fonte: Autoria própria

O mesmo modelo foi testado considerando as amostras dos agrupamentos 1 e 2, pois são os grupos em que foi possível obter modelos consideráveis. As variáveis preditoras, SiO_2_gl e PPC_gl, se mostraram estatisticamente significativas, com valores-p menores que 0,05, entretanto, a capacidade preditiva do modelo se mostrou muito baixa, conforme mostra a Figura 28.

Figura 28 - Estatísticas do modelo de regressão considerando os clusters 1 e 2

Sumário do Modelo

S	R2	R2(aj)	R2(pred)
5,66129	56,86%	51,47%	46,34%

Coeficientes

Termo	Coef	EP de Coef	Valor-T	Valor-P	VIF
Constante	96,15	4,50	21,35	0,000	
SiO2_gl	-0,725	0,167	-4,35	0,000	2,54
PPC_gl	-0,847	0,345	-2,46	0,026	2,54

Análise de Variância

Fonte	GL	SQ (Aj.)	QM (Aj.)	Valor F	Valor-P
Regressão	**2**	**675,9**	**337,97**	**10,54**	**0,001**
SiO2_gl	1	606,0	605,99	18,91	0,000
PPC_gl	1	193,7	193,68	6,04	0,026
Erro	**16**	**512,8**	**32,05**		
Total	**18**	**1188,7**			

Fonte: Autoria própria

Além desse modelo de regressão, diversos modelos com variáveis *dummies*, representando os diferentes agrupamentos foram testados, pois existem diferenças significativas em relação à média da variável resposta. Entretanto, embora a análise de variância tenha demonstrado diferenças significativas entre os grupos, o modelo de regressão com variáveis *dummies*, não indicou efeito estatisticamente significativo dos clusters. Isso sugere que as variáveis independentes utilizadas, os teores globais, já explicam as diferenças observadas na recuperação mássica, tornando desnecessária a inclusão dos agrupamentos no modelo.

6. CONSIDERAÇÕES FINAIS E SUGESTÃO DE ESTUDOS FUTUROS

Este trabalho teve como principal objetivo contribuir, através de análises exploratórias, para a modelagem geometalúrgica de uma mina de ferro. O projeto geometalúrgico se encontra na fase de armazenamento e interpretação dos dados, visando compreender o desempenho das diferentes litologias até que chegue ao produto. Apesar do fluxograma contemplar análises de granulometria e mineralogia, optou-se pelo uso exclusivo dos teores globais por todas as análises estarem finalizadas.

A análise de componentes principais relevou que, utilizando os teores globais, as duas primeiras componentes explicaram aproximadamente 81% da variabilidade total do conjunto de dados, sendo, portanto, as duas utilizadas na etapa de agrupamento dos dados. A variabilidade total da primeira componente foi mais explicada pelas variáveis Mn_gl, CaO_gl, MgO_gl e PPC_gl, enquanto PC2 se mostrou mais explicada pelas variações de SiO_2_gl (positiva) Fe_gl (negativa), com contribuição de Al_2O_3_gl.

Embora o agrupamento não supervisionado por k-means tenha retornado uma separação não muito clara dos grupos, comprovada pelo gráfico de silhueta com valor médio de 0,53, foi possível obter esclarecimentos das amostras presentes dentro de cada agrupamento, especialmente ao analisá-las considerando suas composições químicas e tipo de litologia. O cluster 1 apresentou composição exclusiva do tipo litológico IDOL, o cluster 2, apesar de mais heterogêneo, evidenciou amostras majoritariamente dos tipos litológicos IF e IC, e o cluster 3 mostrou que o tipo litológico IC2 e IFP formam a maioria do agrupamento.

Para verificar a significância estatística dos agrupamentos, foi constatada, por meio do teste paramétrico da análise de variância (ANOVA), que os agrupamentos apresentam diferenças significativas em relação à média da recuperação mássica, além de diferenças significativas nas médias dos teores de Fe_gl e SiO_2_gl, comprovada pela análise multivariada de variância (MANOVA), o que justificou a tentativa de buscar modelos de regressão por cluster para prever a variável geometalúrgica.

Ao analisar a composição química dos grupos formados, percebeu-se maior valor de Fe_gl para o cluster 2, maior teor de SiO_2_gl para o cluster 3 e maiores teores dos contaminantes Mn_gl, CaO_gl, MgO_gl e PPC_gl, para o cluster 1. Em termos de recuperação mássica, o cluster 1 apresentou a maior média, o cluster 2 a maior variabilidade entre os dados e o cluster 3 a menor média.

Para o grupo 1, apesar da limitação da quantidade de dados, foi obtido um bom modelo preditivo para a variável alvo, com R^2 de 99,42% e R preditivo de cerca de 92%, com as variáveis PPC/Fe e SiO_2_gl como preditoras. Para o grupo 2, um modelo com capacidade preditiva moderada foi encontrado, apresentando R^2 de 75% e R preditivo de cerca de 63%, utilizando as variáveis Fe_gl e MgO_gl como preditoras, reforçando a suposição de que outras características do material, como granulometria e mineralogia, também irão exercer influência sobre a recuperação mássica. Já para o grupo 3, nenhum modelo consideravelmente significativo foi encontrado, o que reforçou a ideia de que a inclusão de novas variáveis relacionadas a outras características desse tipo de litologia pode explicar a variável resposta estudada.

Além disso, encontrou-se um modelo de regressão considerando todos os tipos de itabirito, que resultou em um modelo com R^2 de 81,55% e R preditivo de 77,32%, apresentando as variáveis SiO_2_gl e PPC_gl como preditoras, ambas demonstrando queda no valor da recuperação mássica ao aumentar o valor desses componentes. Modelos considerando outros cenários, como junção dos clusters 1 e 2 e incluindo variáveis *dummies* também foram testados, mas não obtiveram resultados significativos.

Para os próximos estudos exploratórios, sugere-se acrescentar variáveis relacionadas à granulometria e mineralogia das amostras, além de considerar todo o material que vira concentrado e produz dados de recuperação mássica. Para além disso, outros métodos de agrupamento de dados podem ser testados, para explorar diferentes domínios geometalúrgicos, uma vez que o k-means é sensível à quantidade de dados e tende a procurar agrupamentos esféricos minimizando a distância intra cluster, forçando uma amostra pertencer a um grupo, mesmo que seja, na realidade, um *outlier*. Além disso, sugere-se investigação estatística detalhada de quais agrupamentos podem apresentar diferenças significativas em relação à média da recuperação mássica.

Considerando as limitações do conjunto de dados em função do número reduzido de amostras, pode-se recorrer a abordagens que potencializam o aproveitamento da informação disponível. Técnicas de *boosting* permitem gerar modelos mais robustos ao realçar padrões latentes nos dados e reduzir erros sucessivos, funcionando como uma forma de “amplificação” da base existente. Complementarmente, a Análise Discriminante mostra-se adequada nesse contexto, pois, ao projetar as variáveis em combinações que maximizam a separação entre classes, favorece a classificação mesmo em cenários de menor volume amostral.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, A.; CARVALHO, F.; MENINO, F. Introdução ao Machine Learning. DataAt. 2020. Disponível em: https://dataat.github.io/introducao-ao-machine-learning/regress%C3%A3o.html. Acesso em 30 maio 2025.

ARROYO ORTIZ, C. E. Caracterização geometalúrgica e modelagem geoestatística da Mina de Brucutu – Quadrilátero Ferrífero (MG). Tese de Doutorado – Programa de Pós-Graduação em Engenharia Mineral (PPGEM), Universidade Federal de Ouro Preto, v. Contribuição às Ciências da Terra - Vol. 234, 2014.

BARROS, A. Introdução à técnica de clusterização K-means. Medium. 2021. Disponível em: https://ealexbarros.medium.com/introdu%C3%A7%C3%A3o-a-t%C3%A9cnica-de-clusteriza%C3%A7%C3%A3o-k-means-89caf4f56bb3. Acesso em 30 maio 2025.

CHEMALE JR, F.; TAKEHARA, L. Livro Minério de Ferro: Geologia e Geometalurgia. REM: Revista Escola de Minas, v. 66, n. 4, p. 412-415. Ouro Preto, 2013. Disponível em: http://www.redalyc.org/articulo.oa?id=56429245002. Acesso em: 20 junho 2025.

COSTA, D. O. et al. O mercado de trabalho para o engenheiro industrial num ambiente pós pandemia: uma abordagem a partir do método dos cenários prospectivos. Revista SIMEP, João Pessoa, v. 2, n. 2, p. 68-90, 2022.

CORDEIRO, R. A. C.; SOUZA, J. A. H.; VIANA, B. H. C. A Geometalugia e a Geoestatística: melhores práticas para estimativas de recursos e reservas, controle de produção e elaboração de projetos. In: 39º Seminário de Redução de Minério de Ferro e Matérias-primas e 10º Seminário Brasileiro de Minério de Ferro. Ouro Preto, 2009.

DUARTE, J. B. Introdução à análise de componentes principais (com aplicação do SAS – Statitical Analysis System). Goiânia: Universidade Federal de Goiás (UFG), 1998. Disponível em: https://files.cercomp.ufg.br/weby/up/396/o/ACP.pdf. Acesso em: 18 maio 2025.

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. Cluster Analysis. 5th ed. New Jersey: John Wiley & Sons, 2011. 348p.

FÁVERO, L. P.; BELFIORI, P. Manual de Análise de Dados: Estatística e Modelagem Multivariada com Excel, SPSS e Stata. 1a ed. Rio de Janeiro: Elsevier, 2017. 1219p.

FERNANDES, F. G. Estudo do melhor método de extrapolação de regressão múltipla para construção do modelo geometalúrgico de uma mina de fosfato brasileira. Dissertação de Mestrado - Programa de Pós-Graduação em Engenharia Mineral (PPGEM), Universidade Federal de Ouro Preto. Ouro Preto, 2013. 160p.

FONSECA, E. Geometalurgia ao longo do desenvolvimento de Projetos Minerais. Apresentação oral – 1º Simpósio de Geometalurgia, Belo Horizonte, 2024. Disponível em: https://geomet.eng.ufmg.br/assets/files/Anais-doISimposiodeGeometalurgia-2024-VersoFinal.pdf. Acesso em: 16 maio 2025.

GOLÇALVES, P. C. P. Geometalurgia: Panorama de Aplicação no Setor Mínero-Metalúrgico. Monografia – Universidade Federal de Ouro Preto. Ouro Preto, 2017. 90p.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Análise Multivariada de Dados. Tradução de A. S. Sant'anna e A. Cloves Neto, 6.ed. Porto Alegre: Bookman, 2009, 593 p.

HARRIS, C. R.; MILLMAN, K. J.; VAN DER WALT, S. J. et al. Array programming with NumPy. Nature, 585(7825), p. 357–362, 2020.

HEIDER, M.; FONSECA, D. S. Desafios e riscos do setor mineral brasileiro. In the Mine, 2022. Disponível em: https://www.inthemine.com.br/site/desafios-e-riscos-do-setor-mineral-brasileiro/. Acesso em: 15 junho 2025.

HONGYU, K.; SANDANIELO, V. L. M.; JUNIOR, G. J. O. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. Engineering and Science, v. 1, 5. ed, 2015.

HUNTER, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, n 9(3), p 90–95, 2007.

IBM. O que é a análise de componentes principais (PCA)? 2023. Disponível em: https://www.ibm.com/br-pt/think/topics/principal-component-analysis. Acesso em: 16 maio 2025.

JAIN, A. K. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, v. 31, p. 686-695, 2010. DOI: https://doi.org/10.1016/j.patrec.2009.09.024

JOLLIFFE, I. T. Principal Component Analysis. 2nd ed. New York: Springer, 2002. 487 p. (Springer Series in Statistics).

JUNIOR, A. A. B. Análise de componentes principais (ACP). In: Programa de Pós-graduação em Economia Aplicada - disciplina: Análise Multivariada I. Universidade Federal de Juiz de Fora (UFJF), 2016. Disponível em: https://www.ufjf.br/lates/files/2016/12/Conte%C3%BAdo-2-%E2%80%93-N_An%C3%A1lise-de-Componentes-Principais-ACP.pdf. Acesso em: 16 maio 2025.

KALIMARA. Métricas de Agrupamento: Coeficiente de Silhueta, Índice de Davies-Bouldin e Índice de Calinski-Harabasz. Medium. 2023. Disponível em: https://medium.com/@kalimarapeleteiro/m%C3%A9tricas-de-agrupamento-coeficiente-de-silhueta-%C3%ADndice-de-davies-bouldin-e-%C3%ADndice-de-9462b87ce676. Acesso 11 agosto 2025.

KORITIAK, N. A. Modelos de Regressão Linear e suas limitações. Substack. 2024. Disponível em: https://natliaalbierikoritiaki.substack.com/p/modelos-de-regressao-linear-e-suas. Acesso em 06 julho 2025.

LISHCHUK, V. et al. Towards integrated geometallurgical approach: Critical review of current practices and future trends. Minerals Engineering, v. 145, 2020. DOI: https://doi.org/10.1016/j.mineng.2019.106072

McKINNEY, W. Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference. p. 51–56, 2010.

MELO, T. P. Silhouette score para avaliar qualidade dos clusters no algoritmo k-means. Artigo LinkedIn. 2023. Disponível em: https://pt.linkedin.com/pulse/silhouette-score-para-avaliar-qualidade-dos-clusters-algoritmo-parra. Acesso em 11 agosto 2025.

MICHAUX, S.; O'CONNOR, L. How to set up and Develop a Geometallurgical Program. Espoo: Geological Survey of Finland (GKT), 2020. Disponível em: https://www.researchgate.net/publication/341459317_How_to_set_up_and_Develop_a_Geometallurgical_Program_FINAL_v6. Acesso em: 16 maio 2025.

MINITAB, LLC. Interpretar todas as estatísticas – Teste de Kruskal-Wallis. In: MINITAB Support. [s.l.], 2025. Disponível em: https://support.minitab.com/pt-br/minitab/help-and-how-to/statistics/nonparametrics/how-to/kruskal-wallis-test/interpret-the-results/all-statistics/. Acesso em: 26 julho 2025.

NOLETO, C. Boxplot: o que é e como analisar esse tipo de gráfico. Trybe. 2022. Disponível em: https://blog.betrybe.com/estatistica/boxplot/. Acesso em: 25 julho 2025.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, n 12, p 2825–2830, 2011.

PEREIRA, A. S; Manosso, T. W. S; FOSSATTI, E. C; BERTI. S. M. Regressão Linear Múltipla como simplificar por meio do Excel e SPSS. Faculdade de Ciências Econômicas, Administrativas e contábeis (CEPEAC), Universidade de Passo Fundo, 67p., 2019.

POLICARPO. F. F. Minério de ferro: desafios para as indústrias mineral e siderúrgica. Monografia: Curso de Especialização em Engenharia de Recursos Minerais (CEERMIN) – Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, 2012. Disponível em: https://repositorio.ufmg.br/bitstream/1843/BUOS-9CAH4A/1/monografia___vers_o_final___fl_vio_ferreira_policarpo.pdf. Acesso em 25 junho 2025.

PONTIFÍCIA Universidade Católica do Rio de Janeiro (PUC-Rio). Clusterização de Dados. Sistema Maxwell, p. 59-96, 2015. Disponível em: https://www.maxwell.vrac.puc-rio.br/24787/24787_5.PDF. Acesso em: 30 maio 2025.

RAJABINASAB, B.; ASGHARI, O. Geometallurgical Domaining by Cluster Analysis: iron ore deposit case study. Natural Resources Research. v. 28, n.3, p. 665–684, 2019. DOI: https://doi.org/10.1007/s11053-018-9411-6

REIS, M. M. Correlação e Regressão Linear. Nota de aula, [s.l.], [s.d.]. Disponível em: https://www.kufunda.net/pubview.php?id=11153. Acesso em: 05 julho 2025.

RODOVALHO, E.; MAZZINGHY, D. B.; LIMA, J. D. O. et al. Uso de Regressão Linear Multivariada para Modelagem Geometalúrgica de Depósito de Bauxita. In: 1° Simpósio de Geometalurgia. *Anais* Belo Horizonte: UFMG, 2024. p. 57–72. Disponível em: https://geomet.eng.ufmg.br/assets/files/Anais-doISimposiodeGeometalurgia-2024-VersoFinal.pdf. Acesso em: 15 julho 2025.

RODRIGUES, L. O que é boxplot e como analisar esse gráfico. Voitto. 2021. Disponível em: https://voitto.com.br/blog/artigo/boxplot. Acesso em: 25 julho 2025.

SALEMI, 2013. Testes de normalidade: hipótese nula e alternativa. WebArtigos. 2013. Disponível em: https://www.webartigos.com/artigos/testes-de-normalidade-hipotese-nula-e-alternativa/111051/. Acesso em 26 julho 2025.

SANTOS, A. E. M. Classificação de maciços rochosos por meio de técnicas da estatística multivariada e inteligência artificial. Tese de Doutorado – Programa de Pós-graduação em Engenharia Mineral (PPGEM), Universidade Federal de Ouro Preto. Ouro Preto, 2021. 222p.

SCHNEIDER, C. Geometalurgia, mineração de precisão e sustentabilidade. Revista Mineração e Sustentabilidade, p.47, 2014.

TEMPORAL, J. Como definir o número de clusters para o seu KMeans. Medium. 2019. Disponível em: https://medium.com/pizzadedados/kmeans-e-metodo-do-cotovelo-94ded9fdf3a9. Acesso em: 30 maio 2025.

TIU, G. et al. Quantifying the variability of a complex ore using geometallurgical domains. Minerals Engineering, v. 203, 2023. DOI: https://doi.org/10.1016/j.mineng.2023.108323

VIEIRA, M.; MENDONÇA, A.; COSTA, J. F. C. L. Métodos Geoestatísticos Aplicados à Modelagem Geometalúrgica. HOLOS, v. 7, n. 31, p. 65-71, 2015. DOI: http://dx.doi.org/10.15628/holos.2015.3727

WILLIANS, S. R.; RICHARDSON, J. M. Geometallurgical Maping: A new approach that reduces technical risk. SGS Minerals Services, [s.l], 2004. Disponível em: https://www.sgs.co.uk/~/media/Global/Documents/Technical%20Documents/SGS%20Technical%20Papers/SGS%20MIN%20TP2004%2001%20Geomet%20Mapping%20to%20Reduce%20Risk.pdf. Acesso em 17 maio 2025.