UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE GEOCIÊNCIAS – IG
PROGRAMA DE PÓS-GRADUAÇÃO EM GEOLOGIA

MACHINE LEARNING APLICADO NA CARACTERIZAÇÃO DA ASSINATURA PETROFÍSICA, ESPECTRAL E GEOQUÍMICA DOS DEPÓSITOS AURÍFEROS DA SERRA DE JACOBINA, CRÁTON SÃO FRANCISCO

Guilherme Ferreira da Silva

Tese de Doutorado Nº 187

Brasília, DF
Abril de 2022

Universidade de Brasília
Instituto de Geociências – IG
Programa de Pós-graduação em Geologia – PPGG
Área de Concentração: Prospecção e Geologia Econômica

Título: *MACHINE LEARNING* APLICADO NA CARACTERIZAÇÃO DA ASSINATURA PETROFÍSICA, ESPECTRAL E GEOQUÍMICA DOS DEPÓSITOS AURÍFEROS DA SERRA DE JACOBINA, CRÁTON SÃO FRANCISCO

Tese submetida ao Programa de Pós-Graduação em Geologia da Universidade de Brasília, como cumprimento parcial dos requerimentos para a outorga do grau de Doutor em Geologia.

Guilherme Ferreira da Silva

Tese de Doutorado Nº 187

Orientadora: Profa. Dra. Adalene Moreira Silva

Banca examinadora

Prof. Dr. Álvaro Penteado Crosta (IG-Unicamp) – Titular
Prof. Dr. José Carlos Sicoli Seoane (DEGEO-UFRJ) – Titular
Profa. Dra. Susanne Taina Ramalho Maciel (FUP-UnB) – Titular
Prof. Dr. Augusto César Bitencourt Pires (IG-UnB) – Suplente
Profa. Dra. Roberta Mary Vidotti (IG-UnB) – Suplente

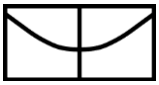
Brasília, DF
Abril de 2022

CIP - Catalogação na Publicação

Ferreira da Silva, Guilherme
FS586m MACHINE LEARNING APLICADO NA CARACTERIZAÇÂO DA ASSINATURA PETROFÍSICA, ESPECTRAL E GEOQUÍMICA DOS DEPÓSITOS AURÍFEROS DA SERRA DE JACOBINA, CRÁTON SÃO FRANCISCO / Guilherme Ferreira da Silva. -- Brasília, 2022.
208 p. Ilustrado.

Orientadora: Adalene Moreira Silva.
Tese (doutorado) - Universidade de Brasília, Instituto de Geociências, Programa de Pós-Graduação em Geologia, 2022.

1. Integração de dados Multifonte. 2. Petrofísica. 3. Geoquímica. 4. Espectrorradiometria. 5. Depósito de ouro em paleoplacer modificado. I. Moreira Silva, Adalene, orient. II. Título.



Para Jaqueline e Ulisses.

Com todo meu carinho.

"All models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind".

"Todos os modelos são aproximações. Essencialmente, todos os modelos estão errados, mas alguns são úteis. Entretanto, a natureza aproximada dos modelos deve sempre ser levada em conta".

- George P. E. Box (1919–2013)

Estatístico britânico responsável por diversos avanços nas áreas de transformação de dados, controle de qualidade e inferência bayesiana.

AGRADECIMENTOS

Este trabalho é fruto do apoio e colaboração de família, amigos e colegas de trabalho, sem o qual não seria finalizado. Trago uma tentativa de síntese da minha gratidão, na esperança de não me esquecer de nenhum nome.

Eu inicio agradecendo o apoio dos meus pais, Antonio Marcos Ferreira da Silva e Eliete Barbosa de Brito: por me ensinarem a sempre buscar aprendizado. Sem seu direcionamento e incentivo a minha jornada certamente seria outra. Não me furto de agradecer ao apoio do meu padrasto, Gilmar Elias Rodrigues, que também contribuiu neste mesmo sentido, e da minha irmã, Jéssica Ferreira, que me hospedou em sua casa no primeiro ano do doutorado.

Agradeço ao Serviço Geológico do Brasil pelo apoio logístico, financiamento dos trabalhos de campo e pela liberação para capacitação através de processo seletivo interno. Direciono este agradecimento ao ex-chefe da Divisão de Geologia Econômica, Felipe Mattos Tavares e ao Gerente de Geologia e Recursos Minerais da Superintendência de Salvador, Valter Rodrigues Sobrinho, que tantas vezes apoiaram o meu trabalho.

Agradeço a Universidade de Brasília pela sua importância em mais um estágio da minha formação profissional, pela luta contínua sob a bandeira da educação pública, gratuita e de qualidade, mantendo-se como um dos bastiões da pesquisa acadêmica em tão difíceis tempos. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

Agradeço a equipe da Jacobina Mineração e Comércio (Yamana Gold), mencionando os nomes de Artur Areal Braga, Juliano José de Souza e Wagner Pocay, além de toda a equipe de técnicos e geólogos. Agradeço por serem receptivos e generosos, viabilizando a coleta de amostras necessárias para este trabalho.

Agradeço a minha orientadora, Adalene Moreira Silva, por aceitar me acompanhar nesta caminhada e por me propor um desafio que contribuíra enormemente para meu crescimento técnico. Estendo o agradecimento aos meus coorientadores Catarina Labouré Bemfica Toledo, Evandro Luiz Klein e Farid Chemale Junior, pela convivência amiga e contribuições das mais diversas formas, dentro das possibilidades de cada um.

Agradeço a João Henrique Larizzatti, meu tutor, e Joseneusa Brilhante Rodrigues, ambos membros da Câmara Técnico Científica do Serviço Geológico do Brasil, responsáveis pela viabilização e acompanhamento da pesquisa. Agradeço a ambos pela paciência, amizade e pelos conselhos durante estes quatro anos.

Agradeço aos geólogos, amigos e colegas: Anderson Dourado, Carina Lopes, Kotaro Uchigasaki, Matheus Ferreira (meu irmão), Guilherme Teles, Fernando Almeida, Pedro Costa e Anderson Matias pelo apoio prestado durante os trabalhos de campo e levantamento de dados em laboratório.

Agradeço aos colegas de sala, Marcos Vinícius Ferreira e Iago Lima Costa, pelo incentivo contínuo no exercício das análises quantitativas em geociências, pela amizade e parceria.

Agradeço ao geólogo José Leonardo Andriotti Silva que involuntariamente redirecionou minha carreira, motivo pelo qual sou imensamente grato.

Finalizo essa seção agradecendo a minha esposa, Jaqueline, e filho, Ulisses, a quem dedico esse trabalho. Pelo apoio durante a elaboração desta tese e pela compreensão nos momentos de ausência. Sem fugir do clichê: somos o resultado dos livros que lemos, das viagens que fazemos e das pessoas que amamos.

RESUMO

Neste trabalho foram adquiridas variáveis categóricas e numéricas relacionadas a propriedades físicas das rochas, tais como densidade, susceptibilidade magnética, condutividade elétrica, concentração de radioelementos, reflectância entre outras propriedades químicas. As análises químicas de rocha foram obtidas através de medidas *in situ* de fluorescência de raios-X portátil (pXRF). Adicionalmente, foram analisadas descrições petrográficas e análises de química mineral quantitativas e semiquantitativas em amostras chave para a compreensão do sistema mineral. Ao todo, foram processadas 1950 análises de pXRF, 2484 medidas de espectrorradiometria, 7490 medidas de susceptibilidade magnética, 5720 medidas de condutividade elétrica, 598 medidas de densidade, 541 análises de química mineral (ablação de laser de espectrômetro de massa, LA-ICP-MS) e 304 medidas de radioelementos, além de 20 análises petrográficas por microscópio óptico e 5 análises por microscópio eletrônico. Utilizamos abordagens supervisionadas para fazer previsões e fornecer informações sobre as mineralizações auríferas em rochas do Grupo Jacobina, Cráton do São Francisco, usando os parâmetros petrofísicos e litogeoquímicos em escala de amostra. Um modelo de aprendizado de máquina baseado no algoritmo *Random Forests* foi aplicado para prever a mineralização em amostras de testemunho de sondagem. As acurácias médias foram de 0,87 para treinamento de validação cruzada, 0,91 para os dados de teste e 0,86 para previsão de todas as amostras. O resultado permitiu estimar a importância das variáveis de entrada para a predição e essas estimativas foram validadas por uma interpretação petrográfica de microscopia óptica e eletrônica de varredura, que foram realizadas para esclarecer a relação entre minerais de diferentes estágios com a mineralização do ouro. Paralelamente, utilizamos abordagens não-supervisionadas para extrair informações sobre a estruturação das amostras nos dados de LA-ICP-MS e de reflectância espectral. Usamos métodos de Agrupamento Aglomerativo (*Hierarchical Clustering*) para avaliar os padrões de elementos traços de acordo com o tipo de pirita (detrítica ou epigenéticas) e níveis estratigráficos. Em seguida, implementamos a técnica *Uniform Manifold Approximation and Projection* (UMAP) para reduzir a dimensionalidade avaliada para uma projeção bidimensional buscando inspecionar a estrutura interna dos dados. Elementos como Cu, Zn, Ag, Sb, Te, Au, Pb e Bi são mobilizados durante a alteração mineral e foram cristalizados em minerais recém-formados, como calcopirita, pirrotita e esfalerita, que estão espacialmente associados à pirita epigenética e ouro. O padrão das piritas do Grupo Jacobina parece não variar ao longo da estratigrafia, o que sugere uma manutenção da fonte de sedimento ao longo da história de sedimentação ou um posterior reequilíbrio químico. Relativo

às análises de reflectância espectral, aplicamos o algoritmo de *Self-Organizing Maps* (SOM) para segmentar dados em vários agrupamentos baseados na matriz de distância das unidades e, em seguida, usamos a projeção UMAP para compactar a estrutura de dados para um gráfico bidimensional, mantendo os principais padrões de dados e comparando com os espectros de minerais conhecidos descritos nos metaconglomerados. Assim, estimamos a composição mineral com base na distância de cada medição dos minerais conhecidos e validamos essa inferência usando dados geoquímicos. Os resultados da inferência mineral corresponderam ao esperado pela análise geoquímica, validando a estimativa da composição mineral das amostras. Baseado nestes resultados, separamos as assinaturas das propriedades físicas e químicas nas zonas mineralizada, proximal e estéril e indicamos critérios que podem ser utilizados para a prospecção de ouro em zonas de paleoplacer modificado, como presença de calcopirita, esfalerita e outros sulfetos na matriz, além de pirita, teores de cromo, potássio e enxofre, susceptibilidade magnética, densidade e a presença de argilominerais.

PALAVRAS-CHAVE: Machine-learning aplicado a geociências; Prospecção mineral, Ouro em paleoplacer modificado; Integração de dados multifonte.

ABSTRACT

This thesis aims to characterize the signature of gold mineralization of the Serra do Córrego Formation, the basal unit of the Jacobina Group, using multisource data (petrophysics, spectroradiometrics, geochemistry, and mineral chemistry) through data integration and pattern verification using machine learning. Categorical and numerical variables related to the physical properties of rocks were acquired, such as density, magnetic susceptibility, electrical conductivity, the concentration of radioelements, and reflectance, among other chemical properties. Rock chemical analyzes were obtained by in situ portable X-ray fluorescence (pXRF) measurements. Petrographic descriptions and quantitative and semi-quantitative mineral chemistry analyses were also considered in samples for understanding the mineral system. Altogether, 1950 pXRF analyses, 2484 spectroradiometric measurements, 7490 magnetic susceptibility measurements, 5720 electrical conductivity measurements, 598 density measurements, 541 mineral chemistry analyses (mass spectrometer laser ablation, LA-ICP-MS), and 304 measurements of radio elements, in addition to 20 petrographic analyzes by optical microscope and 5 analyzes by electron microscope. We use supervised approaches to make predictions and provide information on gold mineralizations in rocks of the Jacobina Group, São Francisco Craton, using sample-scale petrophysical and lithogeochemical parameters. A machine learning model based on the Random Forests algorithm was applied to predict mineralization in drill core samples. Average accuracies were 0.87 for cross-validation training, 0.91 for testing, and 0.86 for all-sample prediction. The result allowed us to estimate the importance of the input variables for the prediction. These estimates were validated by a petrographic interpretation of optical and scanning electron microscopy, which were performed to understand better the relationship between minerals of different stages of gold mineralization. In parallel, we used unsupervised approaches to extract information about sample structuring from LA-ICP-MS and spectral reflectance data. We used Hierarchical Clustering methods to evaluate trace element patterns according to pyrite type (detrital or epigenetic) and stratigraphic levels. Then, we implemented the Uniform Manifold Approximation and Projection (UMAP) technique to reduce the evaluated dimensionality to a two-dimensional projection, seeking to inspect the internal structure of the data. Elements such as Cu, Zn, Ag, Sb, Te, Au, Pb, and Bi are mobilized during mineral alteration and crystallized into newly formed minerals such as chalcopyrite, pyrrhotite, and sphalerite, which are spatially associated with epigenetic pyrite and gold. The multivariate pattern of the pyrites of the Jacobina Group does not seem to vary along with the stratigraphy, which suggests maintenance of the sediment source throughout the

sedimentation history or a subsequent chemical rebalancing. Concerning spectral reflectance analyses, we apply the Self-Organizing Maps (SOM) algorithm to segment data into various groupings based on the best unit machine distance matrix. We then use the UMAP algorithm to compress the data structure into a two-dimensional graph, maintaining the main data patterns and comparing them with the spectra of known minerals described in the metaconglomerates. Thus, we estimate the mineral composition based on the distance of each measurement from known minerals and validate this inference using geochemical data. The results of the lithogeochemistry validate the estimate of the mineral composition of the samples. Based on all presented results, we separated the signatures of the physical and chemical properties in the mineralized, proximal and sterile zones. We indicated criteria that can be used for prospecting for gold in modified paleoplacer zones, such as chalcopyrite, sphalerite, and other sulfides in the matrix and pyrite, besides Cr, K, and S contents, magnetic susceptibility, density, and the presence of clay minerals.

KEYWORDS: Machine-learning applied to geosciences; Mineral prospecting; Gold in modified paleoplacer; Multisource data integration.

Sumário

Lista de Figuras

1 INTRODUÇÃO

1.1 Apresentação da tese e justificativas

O avanço do conhecimento sobre o conceito de sistemas minerais (Wyborn et al., 1994) vem ao encontro a um desafio da indústria mineral no mundo, ou seja, o de mapear depósitos abaixo da cobertura e em maiores profundidades. A taxa de descobertas de novos depósitos tem diminuído na última década e novas tecnologias têm sido empregadas para se mapear terrenos potenciais em diferentes escalas. Estas tecnologias associam os fatores geológicos que controlam a geração de depósitos minerais e sua preservação ao longo da história geológica (Figura 1-1).

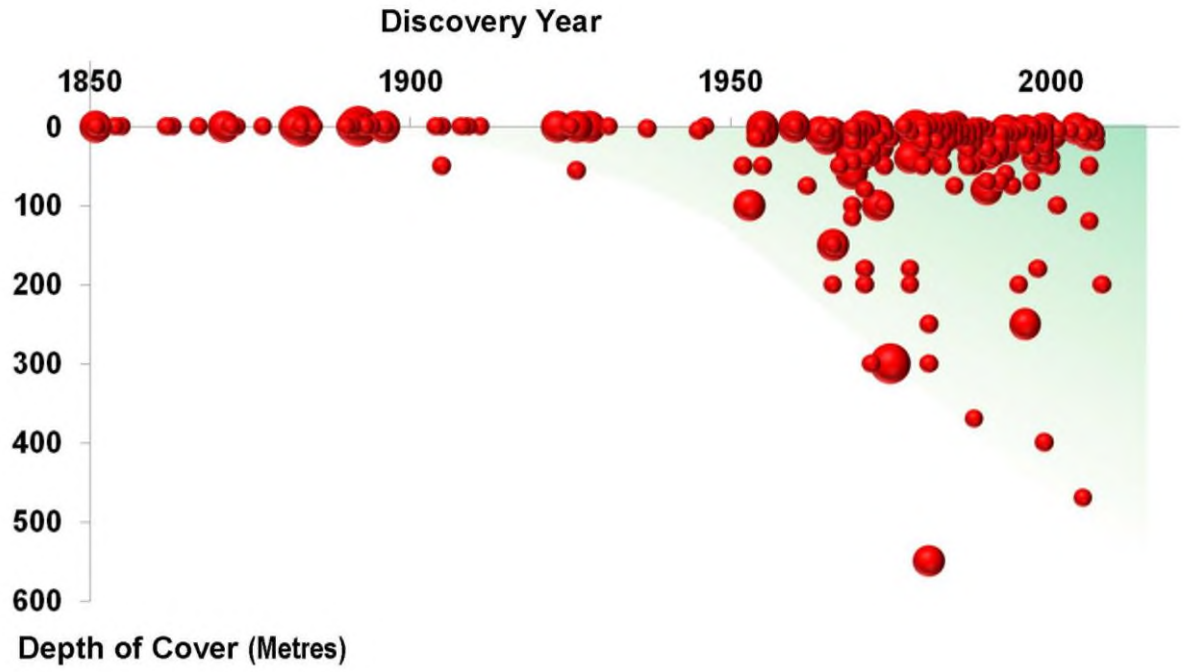


Figura 1-1 – Profundidade de cobertura para as principais descobertas de minerais na Austrália de 1850-2010 (Fonte: Richard Schodde, MinEx Consulting). Os resultados mostram a necessidade de se ampliar a utilização geofísica na prospecção sob a cobertura ("*undercover"*) e que seria a base para a compreensão das respostas geofísicas na ausência de controle geológico.

O conceito de sistema mineral sugere um novo conjunto de objetivos de exploração. Ou seja, a mudança do paradigma convencional que se baseia no mapeamento e na detecção do ambiente mineralizado em escala de depósito para uma compreensão mais abrangente do

sistema levando em conta o ambiente tectônico, as fontes de metais, os caminhos percorridos pelos fluidos e, por exemplo, os paleo-reservatórios (Wyborn et al., 1994).

O fato é que as mineralizações estão associadas a interações fluido-rocha, em maior ou menor escala, e para avançarmos com o entendimento de dados geofísicos que mapeiam abaixo da cobertura é essencial a aplicação de petrofísica para uma melhor compreensão das respostas de processos geológicos e, em especial, da alteração hidrotermal (Dentith et al., 2020; Dentith & Mudge, 2014).

Vários estudos têm sido desenvolvidos para diferentes sistemas minerais, mas as mineralizações auríferas-uraníferas hospedadas em metaconglomerados representam ainda um desafio, devido à alta complexidade dos depósitos, além do recorrente debate sobre a formação das mineralizações. Depósitos de Au (U) em metaconglomerados são conhecidos em crátons pré-cambrianos em todo o mundo. Dentre várias mineralizações, destacam-se os depósitos associados à bacia arqueana de Witwatersrand no Cráton Kaapvaal, África do Sul. Estas minas são responsáveis por quase um terço da produção global de ouro (Frimmel, 2019, 2014; Frimmel et al., 2019). Outros exemplos incluem Tarkwa no Cráton da África Ocidental (Pigois et al., 2003), o Fortescue Group, no Cráton Pilbara, Austrália (Hennigh, 2016), e o Supergrupo Huronian na Província Superior do Canadá (Whymark & Frimmel, 2018), juntamente com vários outros exemplos com maior ou menor relevância econômica.

No Brasil, os exemplares de mineralizações mais conhecidas hospedadas em metaconglomerados auríferos e uraníferos encontram-se no Supergrupo Minas, na porção meridional do Cráton do São Francisco (Guimarães et al., 2019; Minter et al., 1990), na porção basal do Grupo Jacobina, Bloco Gavião, porção setentrional do Cráton do São Francisco (Ledru et al., 1997; Milési et al., 2002; Teixeira et al., 2001) e nos metassedimentos da Formação Castelo dos Sonhos, Província Tapajós, do Cráton Amazônico (Klein et al., 2017). A área objeto

de estudo deste trabalho compreende os depósitos auríferos da Formação Serra do Córrego, unidade basal do Grupo Jacobina (Figura 1-2).

O nível de conhecimento e a quantidade de informações disponíveis em províncias geológicas historicamente conhecidas, permitem que novas abordagens sejam desenvolvidas visando a melhor compreensão da gênese dos depósitos e dos detalhes prospectivos que auxiliarão na procura por novos alvos ou em prolongar a vida útil de distritos maduros. Para tanto, precisamos de informações geoquímicas e mineralógicas para entender estas respostas e avançar nos processos de modelagem e integração de dados. É chave associar o estudo com dados de propriedades químicas, dados mineralógicos e geoquímicos "quantitativos" que irão auxiliar no entendimento do contexto geológico.

Nas últimas décadas, o uso de ferramentas de estatística computacional, *Machine Learning* e inteligência artificial (*latu* sensu) para integração de dados em geociências tem crescido, devido em parte à revolução digital e à consequente maior disponibilidade de equipamentos com boa capacidade de processamento e à tendência crescente de geração e armazenamento de dados das diversas naturezas (Davies, 2002; Flemming et al. 2021). Entretanto, a integração e interpretação conjunta deste grande volume de dados com perspectivas de gerar aplicações práticas na pesquisa mineral ainda é necessária.

Desta forma, esta tese se propõe a avaliar o uso de métodos de *Machine Learning* para integração de dados multifonte em geologia para fins de extração de informações para a pesquisa mineral, baseado em predições explícitas e no conhecimento gerado a partir da verificação da estrutura dos dados geoquímicos, espectrorradiométricos, petrofísicos e mineralógicos em várias escalas de trabalho. Assim, apresentamos os resultados na forma dos artigos técnico-científicos intitulados "*Predicting mineralization and targeting exploration criteria based on a machine learning approach in the Paleoarchean Jacobina quartz-pebble metaconglomerate Au-(U) deposit, northern of São Francisco Craton*", no Capítulo 2,

"*Machine learning applied to the analysis of mineral chemistry in pyrite grains from the Jacobina gold deposits, São Francisco Craton, Brazil: geochemical patterns and implications to mineral exploration*", no Capítulo 3, e os resultados parciais de um artigo em desenvolvimento denominado "*Unmixing spectral signal and estimating the mineral composition of metaconglomerates using dimensionality reduction and relative distance concepts*" no Capítulo 4. As considerações finais da tese são apresentadas no Capítulo 5.

1.2 Objetivos

O objetivo central desta pesquisa é caracterizar a assinatura dos da mineralização aurífera da Serra de Jacobina através do mapeamento do *footprint* petrofísico, mineralógico e geoquímico, bem como a geração de modelos de vetorização mineral através da integração de dados utilizando aprendizado de máquina.

Objetivos específicos incluem:

1. Gerar um banco de dados geofísicos, geoquímicos e de composição mineral das rochas hospedeiras e mineralizadas de amostras da Formação Serra do Córrego, base do Grupo Jacobina;
2. Adquirir dados petrofísicos qualitativos e quantitativos em furos de sondagem chaves ao longo da Formação Serra do Córrego;
3. Caracterizar a assinatura espectral das alterações hidrotermais associadas à mineralização de ouro na Formação Serra do Córrego;
4. Gerar de informações de prospecção mineral através da integração de dados multifonte que envolvam o reconhecimento de padrões e inteligência artificial (*machine learning*).

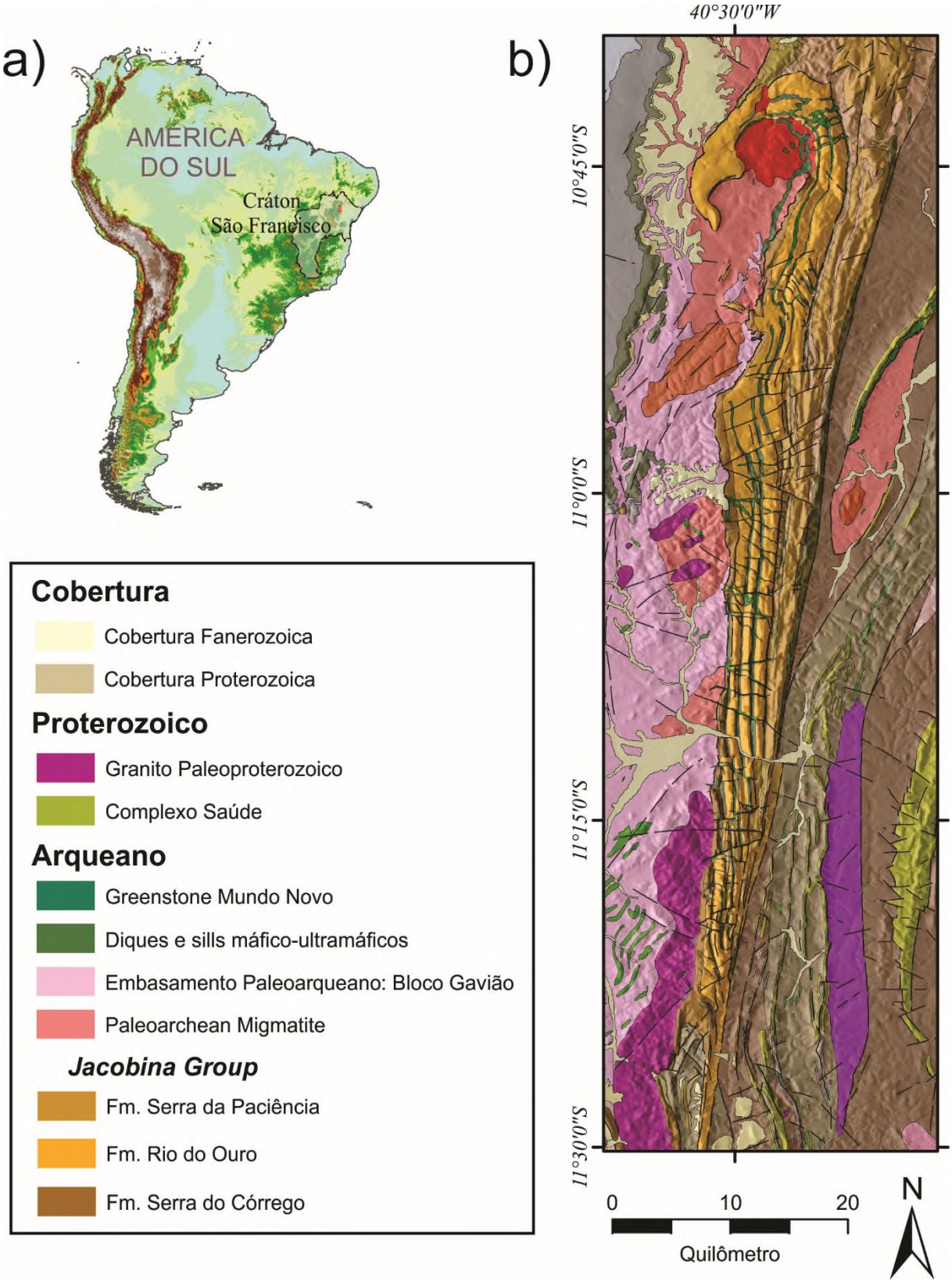


Figura 1-2 – Localização da área estudada, a) posição do cráton São Francisco dentro da plataforma Sul-Americana. O retângulo vermelho indica a localização da área de estudo, na porção setentrional do Cráton São Francisco; b) mapa geológico simplificado da Serra de Jacobina e arredores (modificado de Teles et al. 2015, Reis et al. 2020.

1.3 Materiais e métodos

Os materiais utilizados nesta pesquisa compreendem dados multifonte, como fluorescência de raios-X portátil, susceptibilímetro magnético e condutivímetro elétrico, gamaespectrômetro, balança de precisão para cálculo de densidade, com estação acoplada para medida emersa em fase líquida e espectrorradiômetro de precisão (Figura 1-3)

Com a finalidade de alcançar os objetivos propostos neste projeto, foram desenvolvidas as seguintes etapas e métodos de trabalho, discutidas em maior detalhe nas seções seguintes:

- coleta e descrição de amostras de rocha em testemunhos de sondagem representativos dos níveis mineralizados da Formação Serra do Córrego;
- aquisição de dados de propriedades físicas de rocha;
- aquisição de dados de química de rocha seguindo os protocolos de controle de qualidade indicados;
- aquisição de dados de espectrorradiometria;
- descrição petrográficas em amostras chave para melhor compreensão da dinâmica de formação e inter-relação dos minerais associados à mineralização;
- pré-processamento e análise exploratória de dados;
- integração de dados e escolha do modelo supervisionado adequado

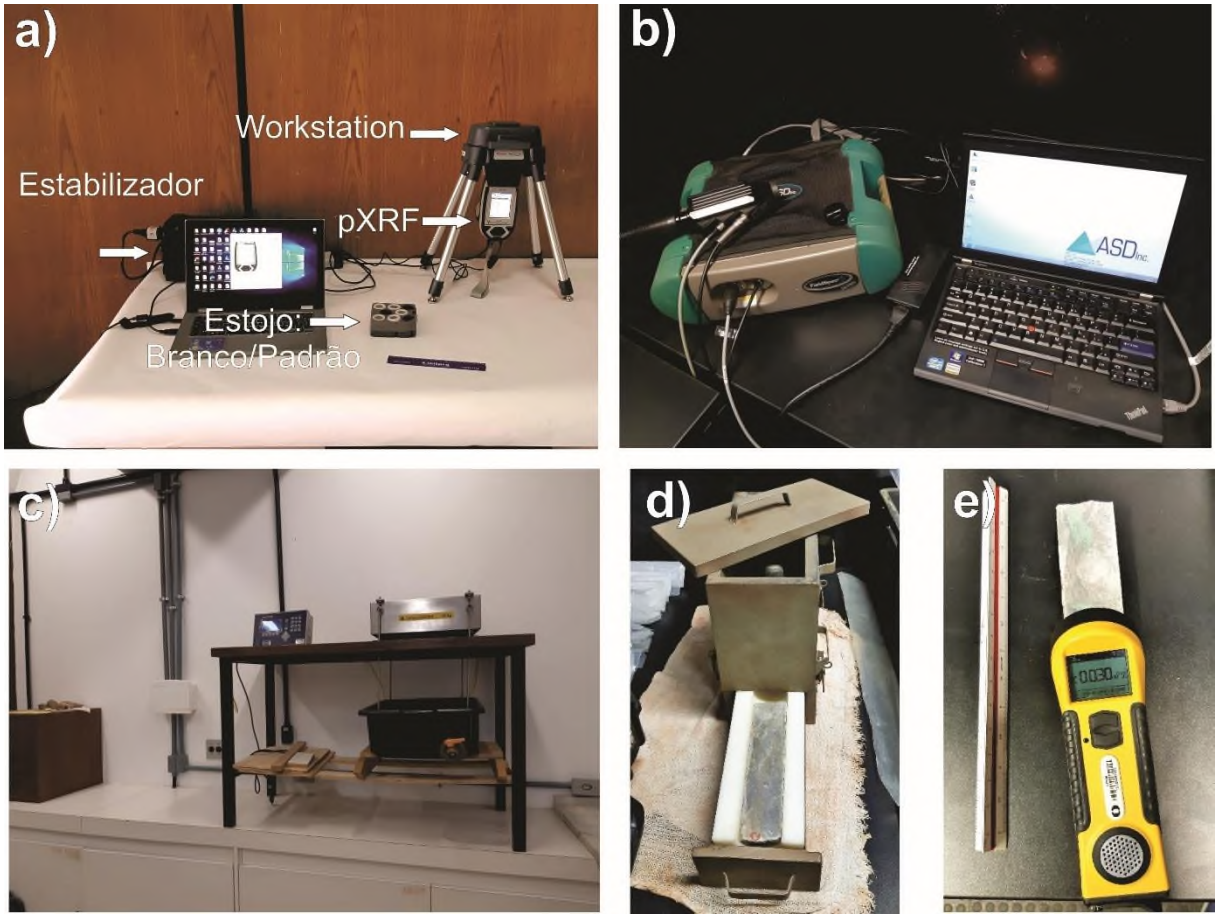


Figura 1-3 – Ilustração das ferramentas utilizadas para obtenção das variáveis numéricas: a) Fluorescência de Raios-X Portátil (pXRF) acoplada a estação de trabalho, b) espectrorradiômetro, c) balança de precisão com plataforma para medida submersa, d) caixa de chumbo para medição de radioelementos, com adaptação para amostras em testemunho de sondagem; Susceptibilímetro KT-10 plus com bobina circular.

1.3.1 Amostragem

Neste trabalho, enfocamos os depósitos de paleoplacer modificados que ocorrem na unidade inferior do Grupo Jacobina, denominada Formação Serra do Córrego. Coletamos 557 amostras de rocha de quatro diferentes testemunhos de sondagem, denominados CANIF-27, JBA-722, CAN-120 e CAN-144. Os furos interceptam toda Formação Serra do Córrego.

As amostras foram obtidas em testemunhos serrados ao meio, coletados em intervalos aproximados de 4 metros, onde foram selecionadas as amostras mais representativas. Em média, as amostras têm 15 centímetros de comprimento, variando de 8 a 50 centímetros, dependendo da representatividade e do tamanho do grão. Predominam amostras de quartzitos e

metaconglomerados, mas também estão descritas amostras de xistos, brechas e rochas meta-ultramáficas.

1.3.2 Propriedades físicas de rocha

As propriedades físicas consideradas nesta pesquisa são densidade, susceptibilidade magnética e condutividade elétrica (Figura 1-4). Os métodos de aquisição dos dados são descritos abaixo.

1.3.2.1 Densidade

As medidas de densidade foram realizadas no Laboratório de Geofísica Aplicada (LGA), da Universidade de Brasília, com base no método de pesagem hidrostática. Os valores de densidade foram calculados após a obtenção do peso da amostra em balança padrão e após a obtenção do peso imerso em água à temperatura ambiente. Em seguida, a densidade foi calculada com base no Princípio de Arquimedes.

A água no recipiente de medição foi trocada a cada 50 amostras ou após a troca da matriz da amostra. Ao final de cada rodada de medições, algumas amostras foram escolhidas aleatoriamente para repetir o teste de densidade. Em caso de repetição, foi considerada a média dos valores. Se a diferença for significativa (ou seja, superior a 0,2 g/cm^3), repetíamos o procedimento até a estabilização.

Foram coletadas 598 medidas de densidade, considerando as replicatas.

1.3.2.2 Susceptibilidade magnética e condutividade elétrica

As propriedades de suscetibilidade magnética (unidade 10-6 SI) e condutividade elétrica (unidade S/m) foram avaliadas com base em um medidor de suscetibilidade magnética Terraplus KT-10 S/C e condutividade com uma bobina circular.

Essas propriedades foram medidas pelo menos dez vezes por amostra, em diferentes pontos ao longo da amostra, e o valor mediano de cada propriedade foi considerado a medida da amostra mais representativa e foi posteriormente usado para a modelagem.

Foram coletadas 7490 medidas de susceptibilidade magnética e 5720 medidas de condutividade elétrica. Para cada amostra, a mediana foi tomada como valor mais representativo, sendo o valor considerado para as transformações subsequentes e análises multivariadas.

A distribuição dos valores das propriedades susceptibilidade magnética e condutividade elétrica tendem a se aproximar de uma distribuição logarítmica (ou seja, valores baixos são mais frequentes do que valores altos). Neste caso, ambas as propriedades foram analisadas em escala logarítmica, com o objetivo de facilitar as interpretações e reduzir a assimetria das distribuições.

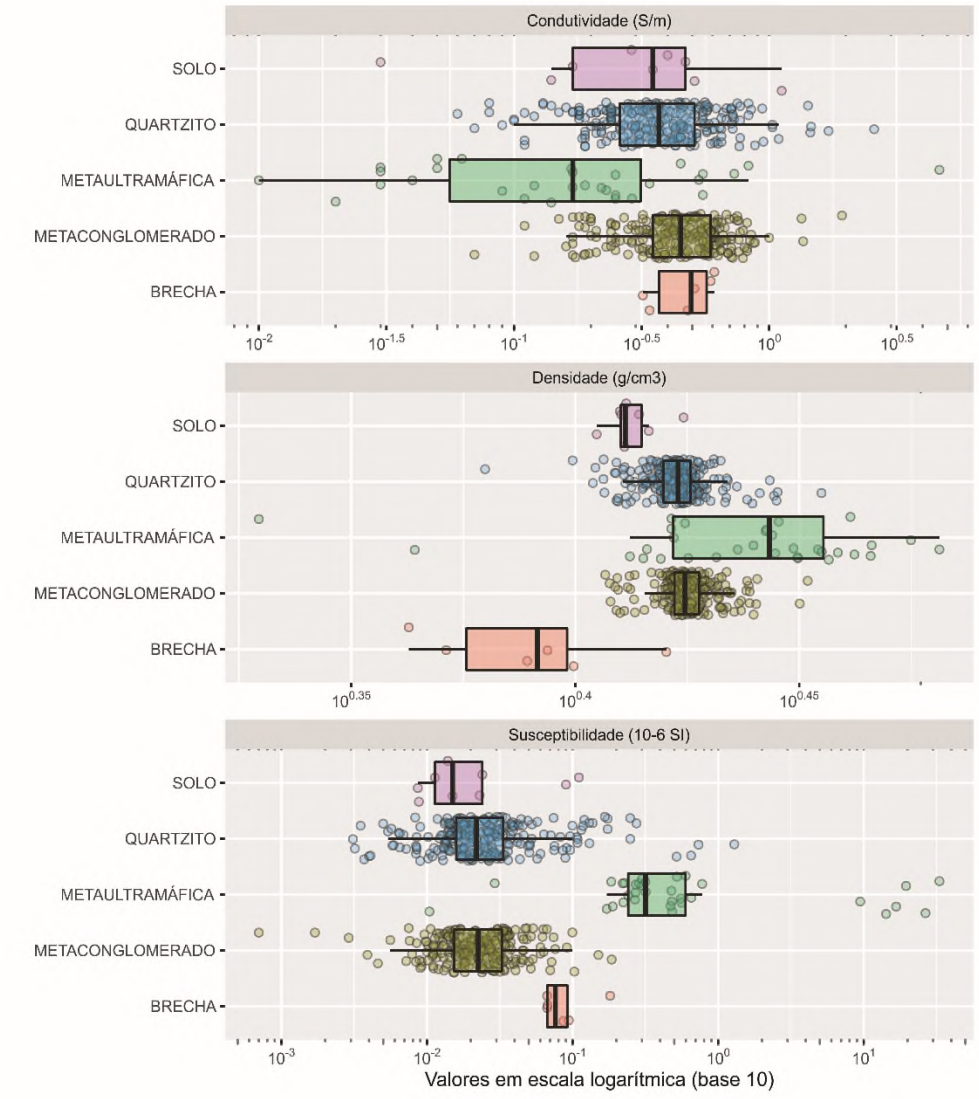


Figura 1-4 – Diagrama de caixa (*boxplot*) para as propriedades físicas adquiridas com transformação para escala logarítmica (base 10). As diferentes cores representam os litotipos considerados na descrição das amostras. Os pontos indicam a posição transformada de cada medida

1.3.2.3 Gamaespectrometria

O conteúdo de radioelementos foi medido em ensaios com uso de um gamaespectrômetro portátil RS-125, tomando os valores da amostra em medições obtidas em uma câmara de chumbo isolada por 300 segundos. Foram coletadas medidas de radioelementos em 304 amostras distintas, compreendendo parcialmente os furos CANIF-27 e JBA-722.

Contudo, os valores obtidos estão geralmente próximos do limite de detecção inferior e estas análises não foram consideradas. Uma interpretação possível é que essa resposta se deve ao baixo volume das amostras e, consequentemente, ao baixo nível de radiação, insuficiente

para excitar o cristal do instrumento de maneira adequada. Portanto, apesar de haver alguns minerais radioativos na assembleia, os valores de radioelementos não puderam ser considerados nas modelagens multiparamétricas utilizadas neste trabalho.

1.3.3 Química de rocha

Para avaliar as concentrações dos elementos nas amostras, usamos um analisador Thermo Scientific Niton XL3t Goldd + XRF, com tubo anódico Au de 2W, 50kV Au e um detector de deriva de grande área geometricamente otimizado. O instrumento foi acoplado em bancada estacionária durante as medições, onde foram colocadas as amostras. Cada medição durou 120s, com 60s de duração para cada feixe. Ao todo, foram realizadas 1950 medidas com o pXRF, considerando as leituras realizadas em padrões certificados, utilizados para o controle de qualidade.

As medições foram realizadas em amostras de testemunhos de sondagem cortados ao meio, usando o modo de ensaio "*point and shoot*". Os procedimentos adotados de controle de qualidade seguiram as sugestões apresentadas por Fisher et al. (2014) e Piercey (2014). No caso de análises cuja dissimilaridade ultrapassaram 10% do valor de referência para aquele elemento, foi realizada a retificação do valor através da aplicação de um coeficiente calculado a partir da razão da medida inacurada sobre o valor de referência. Este coeficiente então é aplicado em todas as leituras subsequentes (Figura 1.5).

Para verificar a representatividade das informações, coletamos uma segunda medição de cada amostra em um local diferente. Apesar de alguns valores atípicos, a distribuição principal é mantida tanto no primeiro quanto no segundo pontos analisados. Calculamos a média da primeira e da segunda medidas e, em seguida, pegamos o valor médio para a análise dos dados.

Os elementos foram selecionados para a análise multivariada baseado na proporção de valores não censurados (i.e., acima do limite de detecção). As variáveis selecionadas

apresentaram ao menos 75% de valores não censurados. Complementarmente, realizou-se a imputação dos valores censurados baseado na substituição por 50% dos respectivos limites de detecção inferiores.

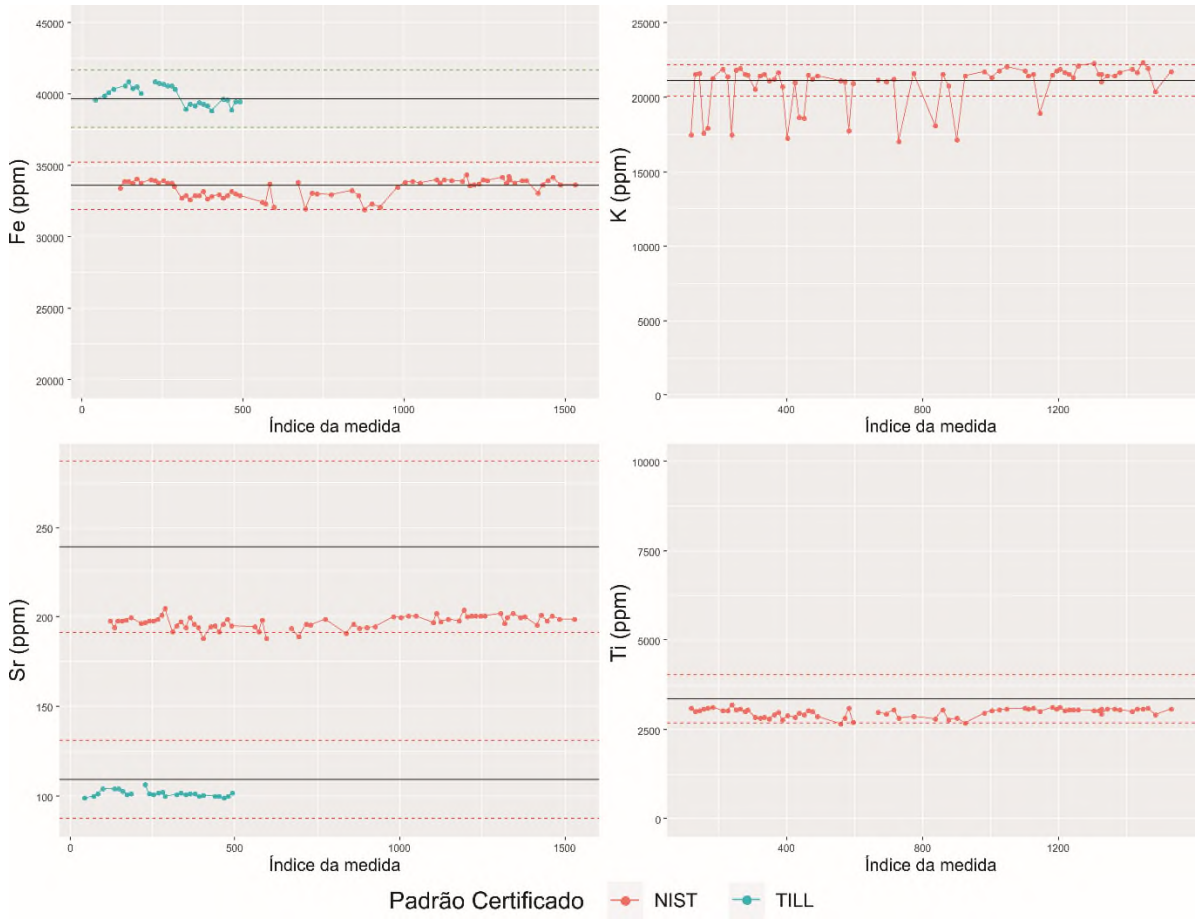


Figura 1-5 – Cartas de controle de Shewhart para as medidas do padrão certificado NIST e TILL, realizadas em determinados durante a análise. A linha pontilhada indica o limite de ± 10% do valore referenciado.

1.3.4 Espectrorradiometria

As análises de assinatura espectral foram realizadas com o equipamento ASD FieldSpec 2, instalado em uma sala com controle de luz externa. As análises abrangem faixas do espectro eletromagnético desde o campo da luz visível (350-750 nm), infravermelho-próximo (751-1000 nm) e infravermelho de ondas curtas (1001-2500 nm).

Através da análise das assinaturas espectrais das amostras dos diversos litotipos, pretende-se investigar variações nos padrões relacionados às fases de alteração hidrotermal identificadas nas amostras.

Os dados de reflectância foram processados através de scripts próprios escritos na linguagem R (http://r-project.org/). Inicialmente, realizamos a segmentação do dado baseado no algoritmo de *Self-Organizing Maps* (SOM; Kohonen, 1998) que permitiu a redução da dimensionalidade para redução do custo de processamento, que é proporcional ao número de dimensões analisadas. O SOM consiste em uma grade regular bidimensional de nós (também chamados de "unidades" ou "neurônios"). A grade é automaticamente organizada baseado na estruturação dos dados de modo a agrupar as unidades semelhantes (Kohonen, 1998). Para esta finalidade, utilizamos uma grade de 400 unidades dispostas em uma malha quadrada (20 x 20). Testamos o treinamento através de diversas épocas, porém o treinamento atingiu a estabilidade em torno de 2500 iterações

Para agrupar o sinal espectral da refletância para curvas de reflectância similares, aplicamos um agrupamento hierárquico baseado na distância euclidiana das unidades do SOM. Esta etapa foi necessária devido aos padrões de refletância altamente complexos, que dificultam a interpretação direta.

1.3.5 Petrografia

Selecionamos 20 amostras mineralizadas para análise microscópica para descrever e verificar as relações texturais e minerais com a mineralização. Estas descrições permitiram o entendimento da relação textural entre as diversas fases minerais encontradas no depósito, inclusive a distinção entre piritas detríticas (da fase de sedimentação) e epigenéticas (formadas durante o metamorfismo ou evento hidrotermal subsequente).

Adicionalmente, cinco amostras foram analisadas pelo método de espectroscopia de elétron/energia dispersiva retro espalhada (BSE-EDS) para melhor avaliação das texturas petrográficas e composições dos minerais (ver Apêndice F).

1.3.6 Pré-processamento e análise exploratória de dados

Considera-se pertencente à etapa de pré-processamento e análise exploratória de dados, todos os passos de organização, estruturação e transformação de dados, testes estatísticos e geração de figuras de exploratórias sobre a distribuição e comportamento dos dados.

Todos estes estágios foram desenvolvidos através da linguagem de programação R (versão 4.1.2). As etapas de manipulação dos dados e geração de gráficos finais foram realizadas através dos pacotes disponíveis na coleção Tidyverse (Wickham, 2014).

As transformações dos dados descritas neste tópico e nos anteriores foram realizadas em linha de comando em ambiente R, através do uso da biblioteca “geoquimica", desenvolvida durante a execução desta pesquisa pelo candidato e disponibilizada através de repositório virtual (https://github.com/gferrsilva/geoquimica).

1.3.6.1 Verificação da distribuição dos dados

Uma etapa essencial na análise exploratória de dados é a determinação do tipo de distribuição. Se os dados forem paramétricos, a média e os desvios padrão são estimativas razoáveis para o centro e a dispersão dos dados e várias ferramentas podem ser usadas para analisar e inferir parâmetros populacionais. Caso contrário, os dados devem ser tratados de forma distinta, com base em métodos que não dependem de parâmetros geométricos.

Existem quase 40 testes disponíveis para verificação de normalidade da distribuição, mas vários autores (Razali & Wah, 2011; Saculinggan & Balase, 2013; Yap & Sim, 2011) mostram que o teste de Shapiro-Wilk é o teste preferido para a maioria dos tipos de distribuições e tamanho amostral. O teste de Shapiro-Wilk (Shapiro & Wilk, 1965) foi inicialmente definido para pequenas amostras ($n < 50$), e então foi aprimorado por Royston (1982) que expandiu o teste para uma faixa maior de valores ($3 \leq n \leq 5000$)

Neste trabalho, este teste foi realizado para cada elemento selecionado no banco de dados, considerando os valores da média de cada amostra. A hipótese nula consiste na

equivalência entre a distribuição dos dados e a distribuição gaussiana, enquanto a hipótese alternativa sustenta que a distribuição dos dados e a distribuição gaussiana não são equivalentes.

Para um nível de significância de 5% e analisando o parâmetro p.value, todos os elementos selecionados apresentaram p.value << 0,01 (Tabela 1.1), o que significa a rejeição da hipótese nula, e consequentemente a respectiva distribuição para cada elemento não é equivalente à distribuição gaussiana, o que se confirmada no diagrama QQ Plot (Figura 1.6).

Tabela 1-1 – Parâmetro estatístico e p.value para o teste de Shapiro-Wilk, com formulação de hipótese nula para a congruência com a distribuição gaussiana

	Dado bruto (ppm)		**Dado Transformado (clr)**	
Elemento	Estatistica	p.value	Estatistica	p.value
Zr	0.432434674	1.33E-38	0.914233291	2.77E-17
Sr	0.531424409	4.95E-36	0.971842006	6.60E-09
Cu	0.604893159	8.03E-34	0.868293075	2.45E-21
Ni	0.605874891	8.64E-34	0.818577858	1.16E-24
Fe	0.431073972	1.23E-38	0.962760516	1.02E-10
Cr	0.336519148	9.10E-41	0.982159738	2.32E-06
Ti	0.364852955	3.72E-40	0.974306805	2.35E-08
Ca	0.184832368	1.03E-43	0.951186247	1.17E-12
K	0.38565616	1.08E-39	0.913317044	2.23E-17
S	0.427819613	1.03E-38	0.959805637	3.01E-11
Ba	0.584526537	1.82E-34	0.816863879	9.17E-25
Cs	0.714989228	8.59E-30	0.807393671	2.59E-25
Te	0.734048894	5.67E-29	0.837710298	1.78E-23
Sb	0.713592928	7.51E-30	0.824527425	2.64E-24
Sn	0.750783251	3.25E-28	0.858613897	4.74E-22
Cd	0.758331722	7.36E-28	0.895298162	4.12E-19
Pd	0.73804041	8.53E-29	0.911212683	1.36E-17
Nd	0.870360028	3.53E-21	0.943938652	1.01E-13
Pr	0.891936733	2.07E-19	0.944743494	1.31E-13
Ce	0.758785709	7.74E-28	0.921160061	1.53E-16
La	0.862151281	8.56E-22	0.935599702	7.81E-15
P	0.623349799	3.25E-33	0.846132596	6.41E-23
Si	0.798403208	8.17E-26	0.667709196	1.19E-31
Cl	0.628795993	4.96E-33	0.781690128	1.06E-26

1.3.6.2 Transformação de dados composicionais

Como os dados geoquímicos são considerados variáveis composicionais (Aitchison, 1986), todas as análises geoquímicas selecionadas foram interpretadas após uma transformação da razão logarítmica centrada (*centered log-ratio* abreviada como clr). Essa transformação é uma etapa essencial para analisar dados composicionais com estatísticas multivariadas (Grunsky, 2001). Ainda assim, os dados das variáveis selecionadas não possuem conformidade com a distribuição gaussiana, e uma abordagem não paramétrica se faz necessária para a integração (Figura 1.7).

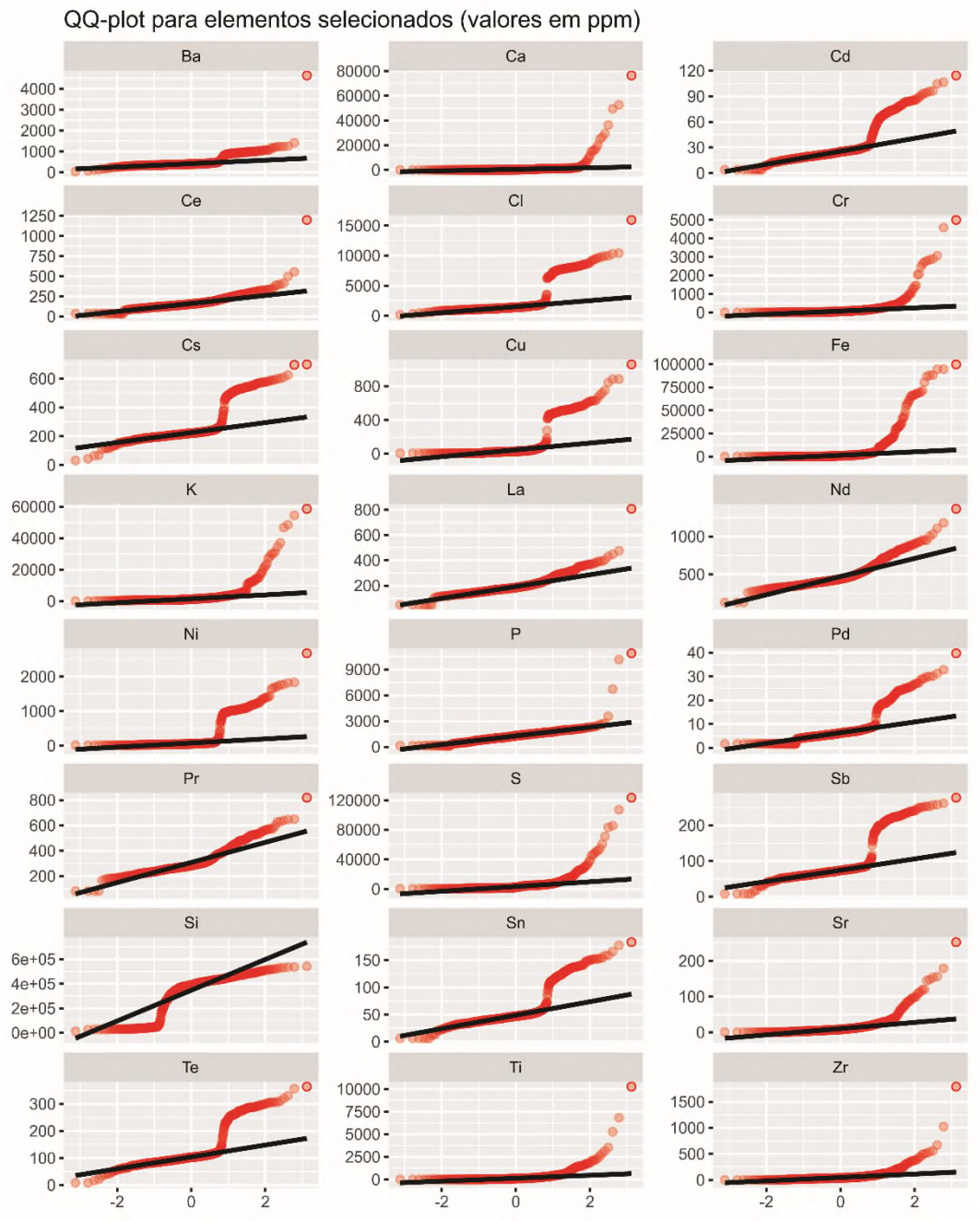


Figura 1-6 – Diagrama Quantil-Quantil (QQ-plot) para os elementos selecionados. A linha preta indica o comportamento esperado para uma distribuição gaussiana. A dispersão dos pontos nos valores extremos da distribuição sugere a presença de valores anômalos, principalmente na porção superior da distribuição. Estes valores anômalos não foram removidos da distribuição justamente por representarem um enriquecimento químico e potencialmente compor parte do objetivo desta pesquisa. A quebra de regularidade na direção de agrupamento dos pontos sugere mudanças na distribuição, o que em alguns elementos caracteriza uma distribuição polimodal. Isto é interpretado como as diferentes distribuições obtidas em litotipos distintos.

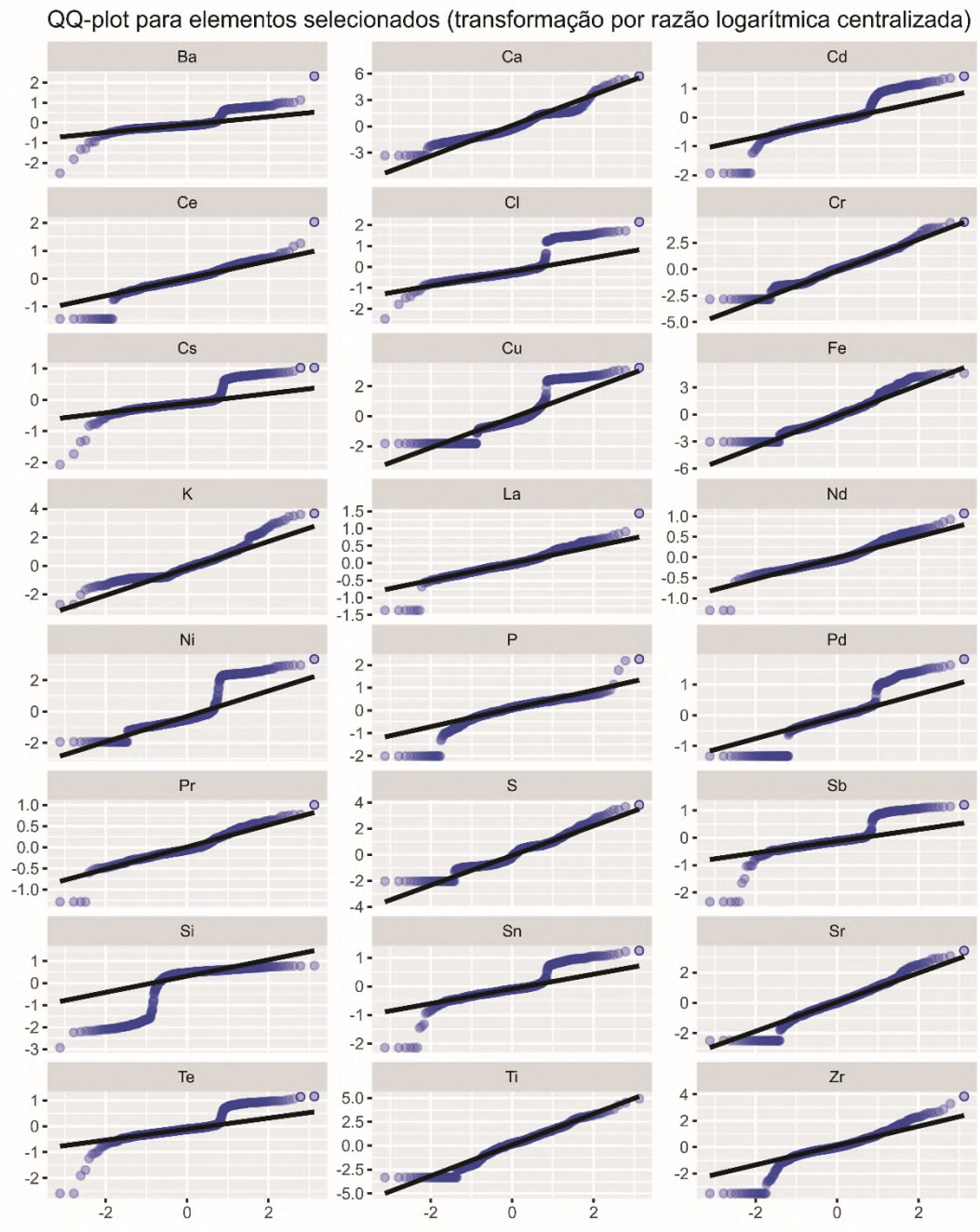


Figura 1-7 - Diagrama Quantil-Quantil (QQ-plot) para os elementos selecionados com valores transformados pelo método da razão logarítmica centralizada. Mesmo após a transformação dos dados, as distribuições não tendem a seguir a distribuição gaussiana (referenciada no gráfico pela linha preta diagonal). Da mesma forma, os valores limítrofes ainda estão em destaque em ambos os lados da distribuição, na maioria dos elementos selecionados e as quebras de tendência de direção de agrupamento também estão presentes.

2 PETROFÍSICA, GEOQUÍMICA E PREDIÇÃO DE MINERALIZAÇÃO

O presente capítulo está apresentado na forma de um artigo, intitulado "*Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil*". O artigo foi aceito para publicação na *Journal of South American Earth Sciences*, aceito para revisão e devolvido com solicitação de correções menores sob o número SAMES-D-21-00534R1.

Este trabalho discorre sobre um modelo de predição de amostras de rocha mineralizadas do Grupo Jacobina, Cráton São Francisco, através da interpretação conjunta de dados petrofísicos e geoquímicos, utilizando modelos de *Random Forests* para a integração dos dados, e geração de um modelo capaz de predizer o status de mineralização das amostras avaliadas.

Apresentamos ainda abordagens inéditas dentro do escopo de predição na prospecção mineral em escala de amostras, como a análise probabilística dos resultados de predição das *Random Forests*. Os resultados foram organizados pela posição das amostras nos respectivos furos. Adicionalmente, discriminamos ainda assembleias minerais existentes em ao menos duas fases hidrotermais presentes na história do Grupo Jacobina e sua relação com a mineralização. O código utilizado na elaboração deste trabalho está disponível no Apêndice C

Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil

Guilherme Ferreira da Silva (Corresponding author) [*,]
[1] Programa de Pós-graduação em Geologia, Instituto de Geociências, Universidade de Brasília, Brasília, DF, Brazil.
[2] Geological Survey of Brazil (SGB/CPRM), SBN, Quadra 2, Bloco H, Edifício Central Brasília, 2º andar, Brasília, DF, Brazil.
E-mail: guilherme.ferreira@cprm.gov.br,
OCID: https://orcid.org/0000-0002-3675-7289

Adalene Moreira Silva
[1] Programa de Pós-graduação em Geologia, Instituto de Geociências, Universidade de Brasília, Brasília, DF, Brazil.
E-mail: adalene@unb.br
ORCID: https://orcid.org/0000-0001-6290-2374

Catarina Labouré Bemfica de Toledo
[1] Programa de Pós-Graduação em Geologia, Instituto de Geociências, Universidade de Brasília, Brasília, DF, Brazil.
E-mail: catarinatoledo@unb.br

Farid Chemale Junior
[3] Programa de Pós-Graduação em Geologia, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, Brazil.
E-mail: faridcj@unisinos.br
ORCID: https://orcid.org/0000-0001-5003-5824

Evandro Luiz Klein
[2] Geological Survey of Brazil (SGB/CPRM), SBN, Quadra 2, Bloco H, Edifício Central Brasília, 1º andar, Brasília, DF, Brazil.
[4] Grupo de Pesquisa em Geologia Econômica, Programa de Pós-graduação em Geologia e Geoquímica, Universidade Federal do Pará, Belém, PA, Brazil
E-mail: evandro.klein@cprm.gov.br
ORCID: https://orcid.org/0000-0003-4598-9249

[*] **Corresponding author**.
Permanent address: Serviço Geológico do Brasil (SGB/CPRM), SBN, Quadra 2, Bloco H, Edifício Central Brasília, 2º andar, Brasília, DF, Brazil. CEP: 70040-904. E-mail: guilherme.ferreira@cprm.gov.br.

Abstract

Defining mineral exploration criteria is a laborious, time-consuming, and generally an empirical task often biased and limited to expert knowledge. To address this problem with a different approach, we used data-driven analysis to make predictions and provide insights about gold mineralization in rocks of the Jacobina Group, São Francisco Craton. The input variables were petrophysical parameters (density, magnetic susceptibility, and electric conductivity) and lithogeochemistry data obtained by X-Ray Fluorescence assays. A machine learning model based on the Random Forests algorithm was applied to predict mineralization in drill core samples. The database used for algorithm training was balanced using the Borderline-SMOTE technique to provide approximately the same numbers of samples of the two classes in the mineral status parameter (i.e., ore and barren samples). The quality of the predictions was assessed with different datasets (i.e., training, testing, each drill core separately, and all samples) and by parameters. The average accuracies were 0.87 for cross-validation training, 0.91 for testing, and 0.86 for all samples. Also, the model allowed us to estimate and rank the importance of the input variables to the prediction. These estimates were validated by an interpretation of optical and scanning electron microscopy petrographic analysis, which were carried out to more clearly understand the relationship between minerals of different stages and gold mineralization. As this approach can be easily replicated in mineral exploration, it is feasible to put models like this in production based on numerical and categorical variables obtained routinely.

Keywords: Gold-bearing quartz-pebble conglomerate; Modified Paleoplacer; Supervised Machine Learning; Hard rock petrophysics; Random Forests.

2.1 Introduction

As more modern exploration techniques are developed, managing, processing, and interpreting all information generated during the exploration workflow presents a significant

challenge. This does not change after the confirmation of positive results for mineralization. With the massive amount of data generated in all stages of the mineral industry, defining, and validating mineral exploration criteria turns into a laborious, time-consuming, and generally empirical task (also often biased and confined to expert knowledge).

However, in the past few years, various machine learning algorithms (MLA) have been used to deal with geological datasets and recurring tasks in several branches within geosciences, including lithology prediction and segmentation (Bérubé et al., 2018; da Silva et al., 2022; Hall, 2016; Saporetti et al., 2018), semi-automated geological mapping (Carneiro et al., 2012; Costa et al., 2019; Harris and Grunsky, 2015), mineral formula calculation (da Silva et al., 2021; Li et al., 2020) and mineral potential modeling (Carranza and Laborte, 2016, 2015a; Prado et al., 2020; Rodriguez-Galiano et al., 2015). Nevertheless, the majority of works using MLA in the mineral exploration concerns the selection of areas with potential for mineralization (i.e., finding new targets), instead of focusing on integrating information to aid in the explanation of a previously known mineralization and the ranking of mineral exploration criteria (Carranza and Laborte, 2015b, 2015a; Chen, 2015; Ford, 2019; Niiranen et al., 2019; Saljoughi and Hezarkhani, 2018; Yousefi and Nykänen, 2016; Zuo, 2017).

Therefore, this article aims to assess the quality of a predictive model based on MLA built with quantitative variables collected in split drill-core samples to predict the mineralization status (i.e., the sample prediction as Ore or Barren). We also discuss the footprint signature of the Au-(U) mineralized samples of the quartz-pebble metaconglomerates of the Jacobina Group through the analysis of petrophysical geochemical data by the variable's importance rank obtained by the data-driven model.

Furthermore, we evaluate the mineralization through a probabilistic approach, presenting the odds of a mineralized sample. This method can provide insights into probability

distribution across the drill core samples ordered by its depth. This probabilistic approach allows investigating the problem in a binary way.

Our work brings an example of the prediction of mineralization status based on quantitative petrophysical and lithochemical variables on gold-bearing rocks from the Jacobina Group. Additionally, the MLA helps estimating exploration criteria based on statistical parameters evaluated upon the evaluation metrics (e.g., model accuracy). Furthermore, those parameters were validated after field descriptions and laboratory analysis of mineralized samples.

2.2 Geology and gold mineralization in the Serra de Jacobina

2.2.1 Geological setting

The Serra de Jacobina (or Jacobina Range) contains Au deposits and is the geomorphological expression of the Jacobina Group, with more than 170 km long and up to 12 km wide, set in the northern portion of the São Francisco Craton in eastern Brazil (Figura 2-1). The Jacobina Group "lies on the eastern edge of the Paleoarchean Gavião Block, close to the suture zone derived from the Paleoproterozoic collision with the surrounding terrains" (Barbosa and Sabaté, 2004; Heilbron et al., 2017; Santos et al., 2019; Teixeira et al., 2017).

According to many authors (Alkmin et al., 1993; Barbosa and Barbosa, 2017; Santos et al., 2019; Leite et al., 2007; Leite, 2002; Teixeira et al., 2001; Teles et al., 2020) the buildup of the Serra de Jacobina occurred during the Paleoproterozoic orogeny developed by the amalgamation of the Gavião, Serrinha, and Jequié paleoplates, between 2.1 Ga and 1.91 Ga.

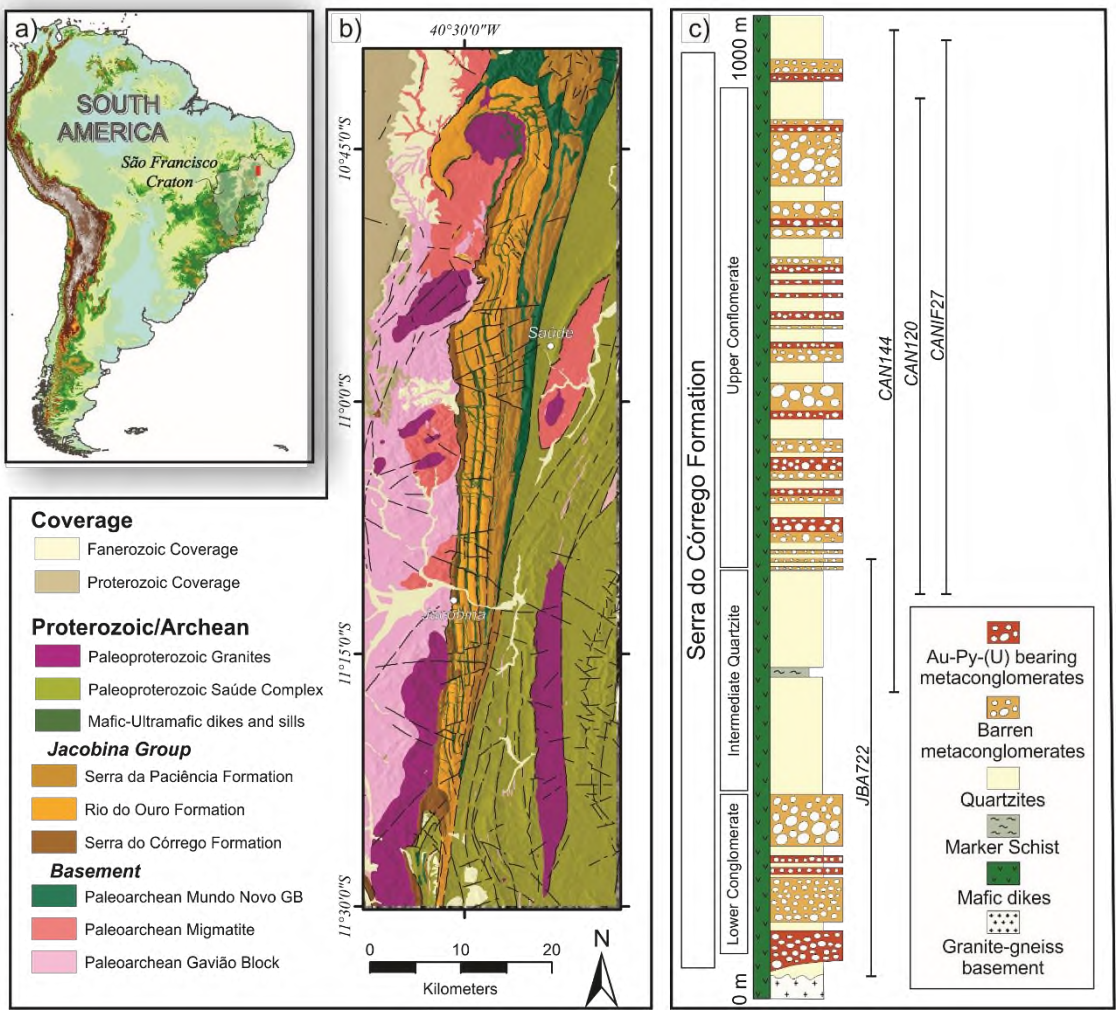


Figura 2-1 a) Localization of São Francisco Craton in South America. The red rectangle indicates the position of the Jacobina Group. b) Simplified geological map of the Jacobina Range and its surroundings (modified after Santos et al., 2019; Reis et al., 2021; Teles et al., 2015 and the references therein); c) Serra do Córrego Formation stratigraphy and an indication of the drill core approximate position (after Teles et al., 2015)

The Jacobina Group was deposited in a pre-GOE stage (> 2.3 Ga.), and the source of the sediments are TTG (Tonalite-Trondhjemite-Granodiorite) with a contribution of mafic-ultramafic rocks as indicated by high concentrations of Cr in lithogeochemistry analysis (Teles et al., 2015). Teles et al. (2015, 2020) also describe rounded sedimentary pyrite grains associated with other detrital minerals, consistent with a pre-GOE deposition. In addition, the source rocks have Paleoarchean age (between 3.3 and 3.4 Ga), indicated by the U-Pb analysis of several rounded zircons crystals (Teles et al., 2015). In the past, some authors considered the possibility of foreland basin, as the tectonic setting for the Jacobina Group (Ledru et al., 1997;

Leite and Marinho, 2012), despite that many works attributed it to inverted rift setting (Santos et al., 2019; Pearson et al., 2005; Teixeira et al., 2001; Teles et al., 2015, 2020, and others).

There is some controversy in the literature about the extension of the Jacobina Group, as some authors consider the Bananeira and Cruz das Almas formations as part of the Jacobina Group, and not consider the Serra da Paciência (Leite et al., 2007; Leite, 2002; Leo, 1964; Mascarenhas et al., 1998; Miranda et al., 2021a; Reis et al., 2021). The interpretation of Jacobina Group adopted in this work comprises quartz-pebble metaconglomerates, quartzites, and schists, from the base to the top of the stratigraphy, and includes the Serra do Córrego, Rio do Ouro and Serra da Paciência formations, following Santos et al. (2018), Teles et al. (2015) and Teles et al. (2020). Numerous mafic-ultramafic dikes and sills intersect those rocks, mainly metamorphosed and partially serpentinized.

The Serra do Córrego Formation, base unit of the Jacobina Group, consists of an alluvial-fluvial formation and comprises an association of quartz-pebble metaconglomerates and quartzites (Santos et al., 2019). The metaconglomerates are mainly oligomictic, composed primarily of quartz pebbles, but some polymictic varieties are found in the Upper Conglomerate Unit with clasts consisting of lithic fragments (granite, quartzite, and metachert of different colors). The metaconglomerates are commonly green-colored due to the presence of fuchsite but can also be yellow-greyish and red-colored, depending on the amount of fuchsite or the degree of oxidation (Mascarenhas et al., 1998).

The Rio do Ouro Formation is exposed in the central part of Serra da Jacobina and is locally 2000 m thick (Teles et al., 2015). The unit consists primarily of high-purity fine-to-medium quartzite, but also thin layers of metaconglomerates are presented in the base, making gradational contact with the lower Serra do Córrego Formation (Pearson et al., 2005; Teles et al., 2015).

149 The Serra da Paciência Formation is exposed along the eastern margin of the Jacobina
150 Range. It consists of thick packages of fine-to-coarse-grained quartzite, conglomeratic
151 quartzites, subordinate metaconglomerates with blue quartz grains of possible volcanic origin,
152 and local andalusite-quartz-graphite schist layers (Pearson et al., 2005; Teles et al., 2015).

153 ### *2.2.2 Deformation, metamorphism, and hydrothermal alteration*

154 Two main deformation phases were identified in the Jacobina Group. The first
155 deformation phase (D_1) is compressional, representing the tectonic transport from east to west,
156 which progressively evolved to a sinistral transpressive phase (D_2; Santos et al., 2019). The
157 second deformational phase (D_2) occurred due to the rotation of the compressive vector to a
158 SE-NW orientation (Santos et al., 2019). In addition, the Jacobina Group is separated from the
159 Saúde Complex by the Pindobaçu Fault System, that has transcrustal dimensions and conforms
160 to the Pindobaçu Suture (Santos et al., 2019).

161 Several metamorphic minerals are described in the rocks from the Jacobina Group and
162 in the cross-cutting mafic-ultramafic dikes (Milési et al., 2002; Miranda et al., 2021; Santos et
163 al., 2019; Teixeira et al., 2001; Teles et al., 2015). Fuchsite, chlorite, epidote, uraninite (the last
164 three are more common at the base of the Jacobina Group), serpentine (in the mafic-ultramafic
165 dikes), andalusite, and graphite (at the top of the Serra da Paciência Formation; Pearson et al.,
166 2005).

167 Additionally, sigmoidal-shaped clasts and pressure-shadows (mainly involving
168 inclusion-rich quartz, fuchsite, pyrite, or chlorite crystals) are observed in hand samples and
169 thin sections, suggesting that the metamorphic minerals were produced during a ductile
170 deformation phase (either D_1 or D_2), as shown in Figura 2-2a to f.

171 Despite da Costa et al. (2020) suggestion for the sedimentary related Witwatersrand
172 Group, Teles et al. (2020) found epigenetic pyrite crystals associated with the metamorphism
173 in the Jacobina Group. In addition, Teles et al. (2020) suggest that epigenetic pyrite grains,

locally associated with remobilized gold, are the most common variety of crystals at the Serra do Córrego and Rio do Ouro formations. Teles et al. (2020) associate this phenomenon with the Paleoproterozoic orogeny (2.1 Ga. To 1.9 Ga) that affected the Jacobina Group. The other hypothesis presented by Teles et al. (2020) is that the epigenetic pyrite could be formed by the influence of the emplacement of granitic bodies in the same era. Both events may have provided metamorphic and/or hydrothermal fluids and the heat required for the recrystallization of pre-existing pyrite. Epigenetic pyrite (Py_2) with inclusions of uraninite that overgrow detrital pyrite (Py_1) is here presented. The Py2 are spatially associated with chlorite corona texture in a shadow-pressure shape. In this texture, inclusion-bearing crystals of uraninite and tiny crystals epigenetic pyrite occur (Figura 2-2c to f).

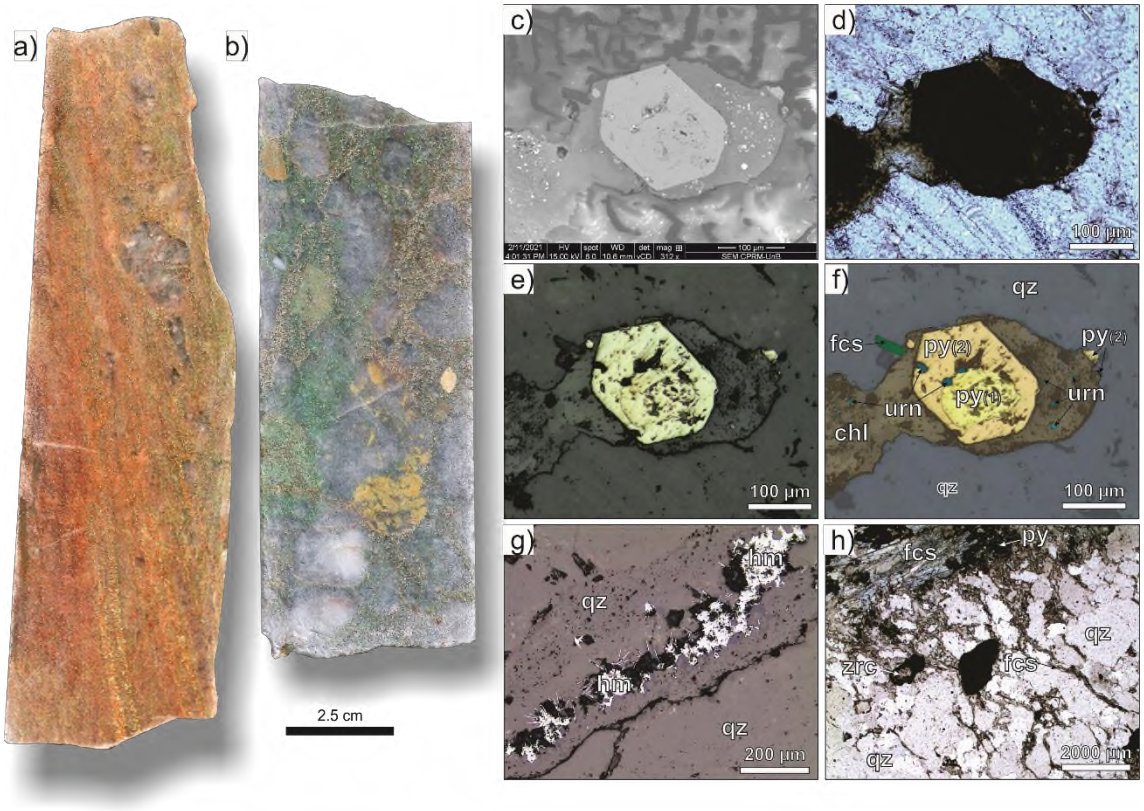


Figura 2-2 – Mineralized drill core samples and photomicrographs: a) pyrite-bearing metaconglomerate, with deformed clasts of quartz and a pervasive iron-oxide alteration (orange and reddish colors), including among the pyrite layers (light yellow); b) polymictic (clasts of quartzite, fuchsite quartzite, metachert, and lithic fragments) pyrite-bearing metaconglomerate with fuchsite (green mica); c) to f) sigmoidal-shaped chlorite in a corona texture around authigenic pyrite (Py2) overgrown on detrital pyrite (Py1), with uraninite inclusions in the chlorite and authigenic pyrite (tiny white crystals in the Back Scattered Electron image - BSE). BSE image, photomicrography (transmitted polarized light), photomicrograph (reflected polarized light) and drawn interpretations (respectively); g) euhedral hematite in a quartz vein, locally replaced by goethite; h) crystals of fuchsite (green mica) among the

recrystallized quartz grains with a euhedral aggregate of pyrite (opaque minerals) spatially associated to a large crystal of fuchsite. Abbreviations: qz – quartz, fcs – fuchsite, py – pyrite, urn – uraninite, chl – chlorite and hm – hematite.

The last alteration described is oxidation. Despite being associated with rock weathering, as suggested by the replacement of pyrite crystals by goethite and other iron oxides, this alteration also forms euhedral hematite crystals, disseminated along with some layers or crystallized within quartz veins (Figura 2-2g). Pearson et al. (2005) firstly described hematite alteration as a late alteration stage in the Serra do Córrego Formation.

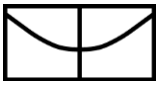
This alteration may vary the intensity depending on the proximity of some brittle structures, and in some cases, breccias and other brittle structures seem to control the alteration. In addition, euhedral crystals of hematite and goethite are more likely to be found closer to fault zones. Despite this, the relation of this alteration with gold mineralization is uncertain in the absence of more evidence.

2.2.3 Gold mineralization

Gold in the Serra do Córrego Formation quartz-pebble metaconglomerates occurs as fine-grained native gold with pyrite or hematite, predominantly in the matrix of coarser metaconglomerates (Pearson et al., 2005). There are two mineralized reefs (following the terminology used in mining) within this formation, named lower and upper conglomerate (Figure 1c; Teixeira et al., 2001).

Exploration in the Jacobina Range has occurred since the early 18th century, with numerous artisanal miners (*"garimpeiros";* Pearson et al., 2005). In modern days, three mines of the Serra de Jacobina produced approximately 700,000 ounces of gold during the years of 1983-1998, and approximately 2 million of ounces during the years of 2003-2019 (Yamana, 2020).

In addition, there are small, clearly hydrothermal gold deposits associated with quartz veins hosted in rocks of the Rio do Ouro Formation (Teixeira et al., 2001; Miranda et al., 2021). These deposits may occur in different contexts but typically contain free gold spatially



220 associated with pyrite, chlorite, epidote, tourmaline, chalcopyrite, and other minor phases,
221 hosted in meta-ultramafic rocks, quartzites, and metaconglomerates.

222 Many works have discussed the origin of the gold in the Jacobina Group in the past few
223 decades (Ledru et al., 1997; Mascarenhas et al., 1998; Milési et al., 2002; Miranda et al., 2021;
224 Pearson et al., 2005; Reis et al., 2021; Teixeira et al., 2001; Teles et al., 2020). Although some
225 authors advocate for a pure hydrothermal origin (Ledru et al., 1997; Milési et al., 2002; Pearson
226 et al., 2005), recent evidence observed in rocks of the Serra do Córrego Formation, such as age
227 of the zircons, presence of detrital pyrite, and isotopic chemistry (Teles et al., 2015, 2020)
228 favored the "modified placer model" (Frimmel, 2019; Teixeira et al., 2001; Ledru et al., 1997;
229 Frimmel et al., 1993; Robb and Meyer, 1991), in which placer mineralization was followed by
230 mobilization of ore components by post-depositional fluids.

231 ## 2.3 Materials and methods

232 ### *2.3.1 Drill core samples*

233 We collected 557 rock samples from four drill cores intersecting the Serra do Córrego
234 Formation (please refer to Fig. 1c). Quartzites and metaconglomerates predominate, but there
235 are also schists, breccias, and meta-ultramafic rocks samples (Tabela 2-1).

236 Tabela 2-1 – Lithology distribution of samples through the drill cores.

Hole	Breccia (n: 6)	Meta-Conglomerate (n: 251)	Quartzite (n: 266)	Schist (n: 02)	Meta-Ultramafic (n: 32)
CAN120	0	62	51	0	3
CAN144	0	60	78	1	3
CANIF27	0	91	27	0	11
JBA722	6	38	110	1	15
Total	6	251	266	2	32

237

The samples were obtained on split drill cores and systematically collected with 4 meters intervals, where the most representative samples were selected. On average, samples have 15 centimeters in length, and however, some samples vary from 8 to 50 centimeters in length, depending on the representativeness and grain size.

The mineralization status of samples was classified based on the discretization of the gold content in the original drill core's assays. All samples were labeled as "Ore" or "Barren" based on the given threshold of 1 ppm (please refer to Tables A1 and A2 in the supplementary data online for more details).

We described textural and mineral assemblages related to mineralization optical microscopy. Furthermore, the selected samples were analyzed by the Back-scattered Electron/Energy Dispersive Spectroscopy (BSE-EDS, FEI QUANTA 450) method to complement the petrographic texture and mineral composition study.

2.3.2 Petrophysics

The mineral exploration industry often uses petrophysical data to evaluate and characterize geological variations, including mineral changes caused by mineralization (Dentith et al., 2020).

According to Dentith et al. (2020), to understand geological controls on physical properties in hard rocks environment, it is necessary to analyze petrophysical data in terms of the properties of different rock types. Still, it is also required to interpret data based on rock and mineral characteristics, as alteration, metamorphism, and strain.

As descriptions used in geological exploration are usually categorical data, the measure of rock properties can provide resources to numerically quantify the variations across the host rocks and target mineralization. Still, the necessity of relying on quantitative variables related to the target phenomena turns the MLA into a practical approach to build the model and assess the accuracy of the predictions.

We acquired properties related to the three types of petrophysical behavior that can respond to changes in overall mineralogy, texture, and grain-size variations (Figura 2-3). Thus, we measured the density, magnetic susceptibility, and electrical conductivity, all of them described in the following section.

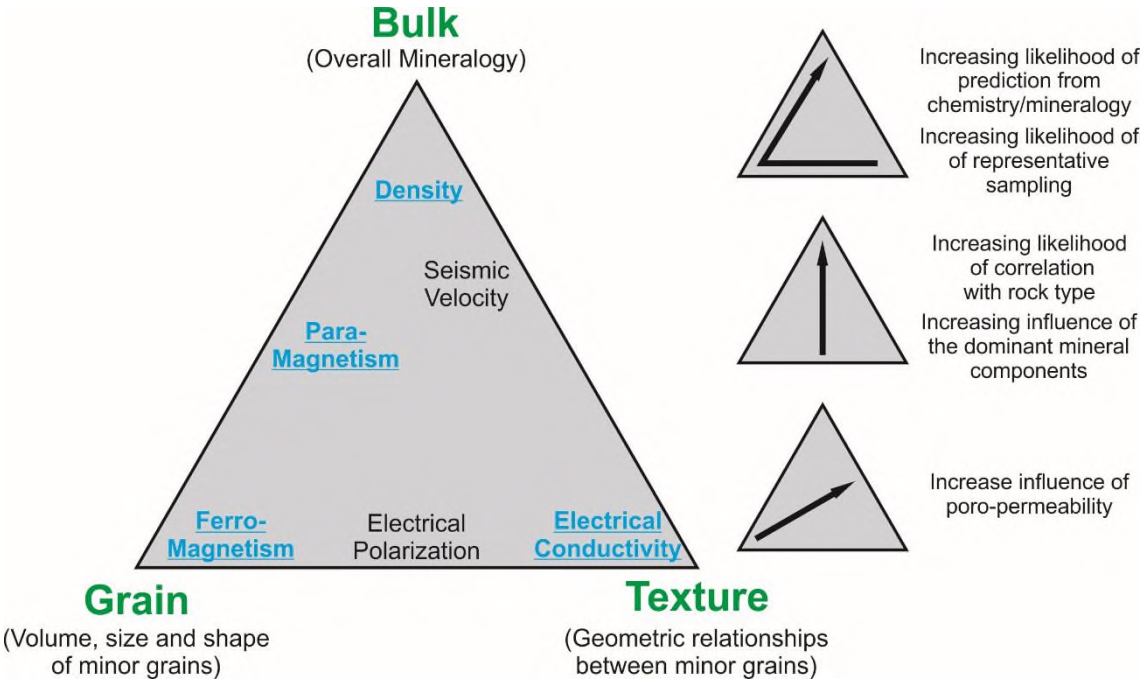


Figura 2-3 – Conceptual framework describing the behavior of various physical properties. The physical properties highlighted in the graph were considered for this work (based on Dentith et al., 2020).

2.3.2.1 Density

The density measurements were taken in the Laboratory of Applied Geophysics (LGA), at the University of Brasília, based on the hydrostatic weighing method. First, the density values were calculated after obtaining the sample weight on a standard scale and after obtaining the weight immersed in ambient temperature water. Then, the density was calculated based on the Principle of Archimedes.

The water in the measuring vessel was changed every 50 samples or after changing the sample matrix. At the end of each round of measurements, some samples were chosen randomly to repeat the density test. In the case of repetition, the average of the values was considered.

We repeated the procedure until stabilization if the difference was not significant (i.e., higher than 0.2 g/cm³).

2.3.2.2 Magnetic susceptibility and electrical conductivity

The magnetic susceptibility (unit 10^{-6} SI) and electric conductivity (unit S/m) properties were assessed using a Terraplus KT-10 S/C magnetic susceptibility and conductivity meter with a circular coil.

These properties were measured at least ten times per sample, at different points throughout the sample, and the median value of each property was considered the most representative sample's measure and was used as input for the modeling.

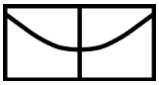
Due to the nature of the distribution of the values (i.e., low values are more frequent than high values), both properties were transformed to a logarithm scale, aiming to ease the interpretations and reduce the asymmetry of data distribution.

2.3.3 X-Ray Fluorescence

To assess the concentrations of elements in the samples, we used a Thermo Scientific Niton XL3t Goldd+ XRF analyzer, with 2W, 50kV Au anode tube, and a geometrically optimized large area drift detector. The instrument was coupled on a stationary test stand during the measurements, where the samples were placed. Each measurement took 120s, with 60s of duration for each beam.

We collected a second sample measurement in a different spot in each sample to check the representativeness of information. Despite some outlier values, the main distribution is maintained in the first and second analyzed spots. We calculated the average of the first and second measurements, and then we took the average value for the data analysis.

The QA/QC adopted procedures that followed the suggestions presented by Fisher et al. (2014) and Piercey (2014). The measurements were performed on the sawn surface of split drill core samples, using the "point and shoot" assay mode. The reference material RM 180-646 was



read between every ten samples (or 20 spot measurements). At the beginning of every analysis batch, the instrument was set to read continually for at least 40 minutes to correct the instrumental drift caused by variations on the cathode temperature (Ida, 2004; Thermo-Scientific, 2013). We collected a total of 1114 measurements, excluding the reference material analysis.

Geochemistry data are considered compositional variables (Aitchison, 1986), and all selected geochemical variables were interpreted after a centered-log ratio transformation. This transformation is essential for analyzing compositional data with multivariate statistics (Grunsky, 2001).

2.3.4 Machine learning analysis (MLA): Random Forests

One of the most employed MLA in geoscience prediction problems is the Random Forests (RF - Breiman, 2001). RF combines several independent estimators (decision trees) to build classification or regression models through bootstrap aggregation. We processed data and built the RF models in the R programming language, using the Tidyverse collection of packages for data wrangling (Wickham, 2014) and the randomForest package (https://cran.r-project.org/web/packages/randomForest) for modeling.

RF relies on the bootstrap principle (random sampling methods with reposition) and the law of large numbers statistical principle, which tells that as the sample size gets large enough, its mean gets closer to the actual average of the whole population (Breiman, 2001). That statistical principle indicates that the RF algorithm does not overfit by considering more estimators. Also, another of RF most significant advantages is that this algorithm has a high performance combined with the low required numbers of hyperparameters, easy to tune. Several geoscientific researchers have shown that the RF outperformed the other MLA, such as support vector machines, artificial neural networks, and logistic regression (e.g., Kuhn et al., 2018; McKay and Harris, 2016; Rodriguez-Galiano et al., 2015). These characteristics make RF

widely and effectively used (e.g., Carranza and Laborte, 2015b; Costa et al., 2019; Ford, 2019; Hariharan et al., 2017; Harris et al., 2015).

In this work, we employ the RF algorithm to predict the mineralized samples, presented in section 2.4.2, to evaluate the ore probability across the drill core samples, presented in section 2.4.3, and to build insights about mineralization targeting and signature by the analysis of variable importance rank, discussed in section 2.5.2.

2.3.4.1 Data balancing, train, and test splits

Imbalanced data is a common issue in the implementation of predictive models. This issue considerably reduces the capacity of the model to perform predictions, especially for the minority classes, where the recognition rate decreases considerably (Japkowicz and Stephen, 2002). Therefore, resampling data configures a mandatory pre-processing step for a successful, high-performance machine learning model (Koziarski et al., 2020).

This problem was previously discussed by (Prado et al., 2020) under the geoscience scope, and the resolution of this issue, or the implementation of a data balancing stage, significantly increased the accuracy of the discussed mineral potential model.

A reasonable solution to this problem was first introduced by Chawla et al. (2002), presenting the SMOTE algorithm (Synthetic Minority Over-sampling Technique). This technique over-samples the minority class by creating synthetic examples rather than over-sampling by replacement. The new generated synthetic data is obtained by the linear combination of previous data, considering the *k*-nearest neighbors of actual samples, until the number of minority and dominant classes is achieved.

In this work, we used a modified version of the SMOTE algorithm, called Borderline-SMOTE (Chawla et al., 2002; Han et al., 2005; Koziarski et al., 2020; Prado et al., 2020), to balance samples classified as Ore (n = 49) and Barren samples (n = 523). The Borderline-SMOTE method priorly classifies the minority data sample into two different subsets, named

DANGER and SAFE samples. This approach assumes that the samples near the limit between minority and dominant classes (DANGER zone) are more easily misclassified than those far from the borderline (SAFE zone, Chawla et al., 2002; Han et al., 2005; Koziarski et al., 2020; Prado et al., 2020). Thus, the algorithm first identifies the borderline between minority examples, and then synthetic examples are generated and added to the original dataset to strengthen the border.

To avoid overfitting, the model by resampling bias, we first randomly sample the data evaluated by the Borderline-SMOTE approach, with a 300% oversampling rate that virtually equalized the number of minority dominant samples. Then, we split the generated dataset by the proportion of 0.7 of original data to separate the data into a train (281 samples) and test dataset (120 samples), considering the original distribution of minority and dominant samples. These data were used to train and assess the model's performance according to sections 2.3.4.2 and 2.3.4.3, respectively.

2.3.4.2 Model tunning

The model tuning is a prior stage to select the optimum hyperparameters of the chosen machine learning model. For RF, the mandatory hyperparameters to tune are the number of estimators (*ntree*, or the number of considered trees) and the number of features randomly taken in each tree (*mtry*), as shown in Figura 2-4.

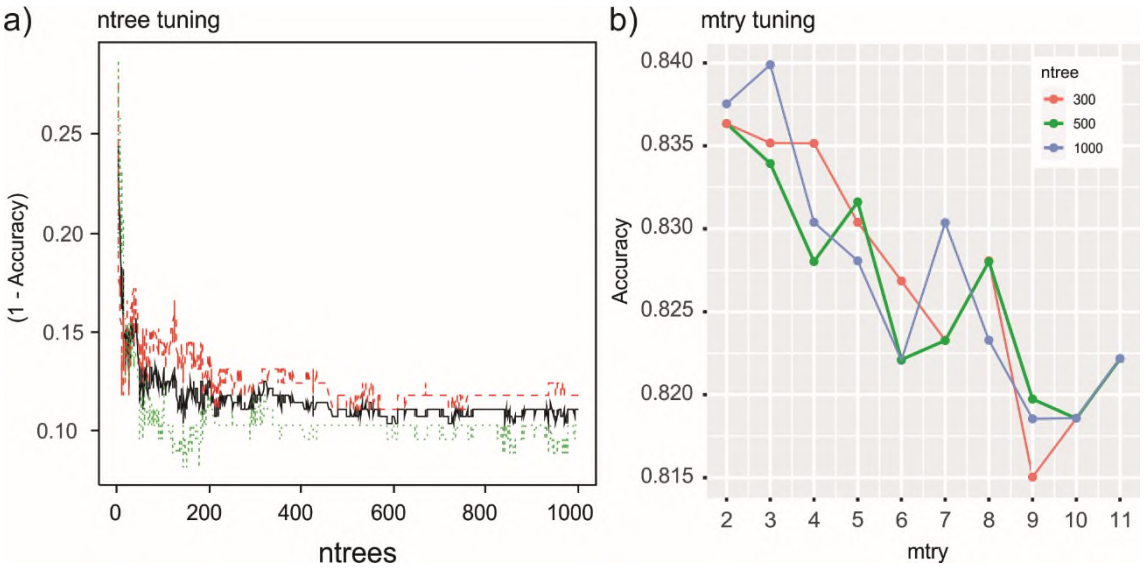


Figura 2-4 – Hyperparameters evaluation and definition of optimum values: a) ntree (number of estimators) search according to the error rate. The error starts to stabilize for ntree greater than 500. The red curve represents the error rate for the barren samples, while the green curve represents the ore samples, and the black curve is the average error; b) feature values and mean accuracy on a grid-search optimization in the cross-validation training step. Each curve represents a given number of trees. The highest accuracy value was taken with the parameters mtry (number of features taken randomly) set to 3 and ntree set to 1000.

One of the possible approaches to defining the optimum values is when an exhaustive search is performed (e.g., "grid search," Bérubé et al., 2018; Prado et al., 2020; Rodriguez-Galiano et al., 2014) and the values which the higher obtained accuracy was selected. For the model in consideration, we took the number of features randomly taken (mtry) and the number of predictors (ntree) values as 3 and 1000, respectively.

2.3.4.3 Performance evaluation

The model's performance is assessed based on quantitative parameters calculated across some subsets of data. For this work, we tested the performance for the train data split, test data split, and for each one of the drill core's samples separated (Tabela 2-2). We also compared the accuracy obtained for the unbalanced and SMOTE-balanced models, supporting the idea that a well-balanced model can make better predictions (Figura 2-5).

Tabela 2-2 – Evaluation parameters for the random forests model implemented in this work for each data subset.

Parameters	**Train Split**	**Test Split**	**CAN144 Drill Core**	**CAN120 Drill Core**	**CANIF27 Drill Core**	**JBA722 Drill Core**	**All Samples**
Number of Samples	281	120	142	116	129	170	557

NIR	0.527	0.500	0.880	0.836	0.884	0.847	0.860
Raw Accuracy	0.854	0.917	0.958	0.922	0.977	0.871	0.926
Balanced Accuracy	0.854	0.917	0.849	0.805	0.989	0.577	0.769
AUC	0.918	0.917	0.942	0.894	0.917	0.934	0.905
Final Accuracy	0.875	0.917	0.916	0.874	0.960	0.794	0.867

Abbreviations; NIR: No Information Rate, AUC: Area Under the Curve.

Raw accuracy is the proportion of the correct predictions, usually based on the analysis of previously labeled samples. We applied cross-validation accuracy for the train dataset in this work, with five repetitions with a 3-times folded analysis. Then, the accuracy obtained in the training dataset is the average of the accuracy obtained in all five repetitions. For the other data subsets, the raw accuracy is calculated as the ratio of correct predictions over the number of samples.

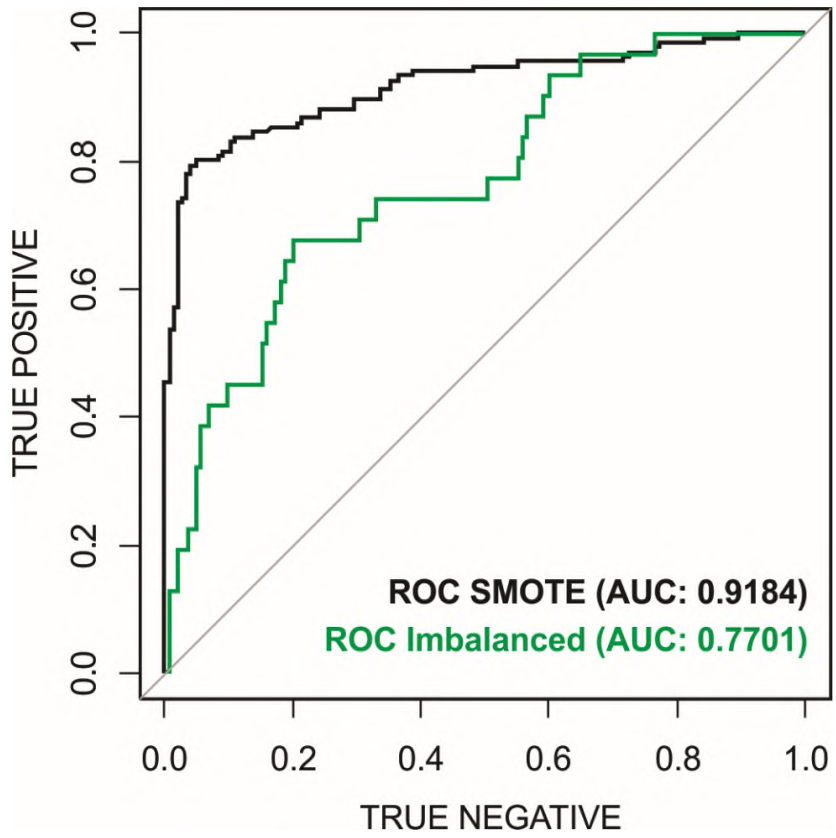
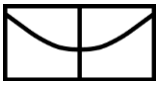


Figura 2-5 – ROC and AUC diagrams for SMOTE-balanced (AUC: 0.91) and imbalanced (AUC: 0.77) training datasets.

The “No Information Rate” (NIR) parameter is calculated based on the proportion of the dominant class in the dataset. The model is considered helpful for each data split if the



calculated raw accuracy is higher than the NIR value, and this value is closer to 0.5 if the dataset is balanced. The NIR value was calculated based on the proportion of the barren samples in each subset of data. In all considered subsets, the raw accuracy values were higher than the NIR, attesting by this criterium that the models are valid.

The balanced accuracy is calculated as the average of the specificity and sensitivity parameters (Zhu et al., 2010). In this paper, specificity is considered the proportion of correctly identified mineralized samples, and sensitivity is the proportion of correctly identified barren samples. The balanced accuracy approach is helpful to evaluate unbalanced datasets (i.e., with a different number of instances for the considered classes).

The Area Under the Curve (AUC) parameter is calculated based on the Receiver Operating Characteristic (ROC), which estimates the trade-off of True positives and False Positive rates (Torppa et al., 2019). The AUC parameter is calculated based on the ROC curve and ranges from 0.5 (i.e., a completely random model) and 1.0 (perfectly accurate model).

To level and consider all described parameters, we calculated the final accuracy as the average of the raw accuracy, balanced accuracy, and AUC values. The final accuracy values range from 0.7937 (for the JBA 722 drill core subset) to 0.9600 (for the CANIF27 drill core subset). The train and test final accuracy were 0.8755 and 0.9168. Thus, the model does not overfit, as the test and train accuracies are close to each other.

2.4 Results and data analysis

2.4.1 Petrophysics and lithochemistry

Figura 2-6 presents the variation among the measured physical properties and some selected elements using a combined graphs strategy (histograms, density plots, and boxplots) and the bivariate analysis. The graphs below are color-coded according to the mineralization status (i.e., Ore or Barren).

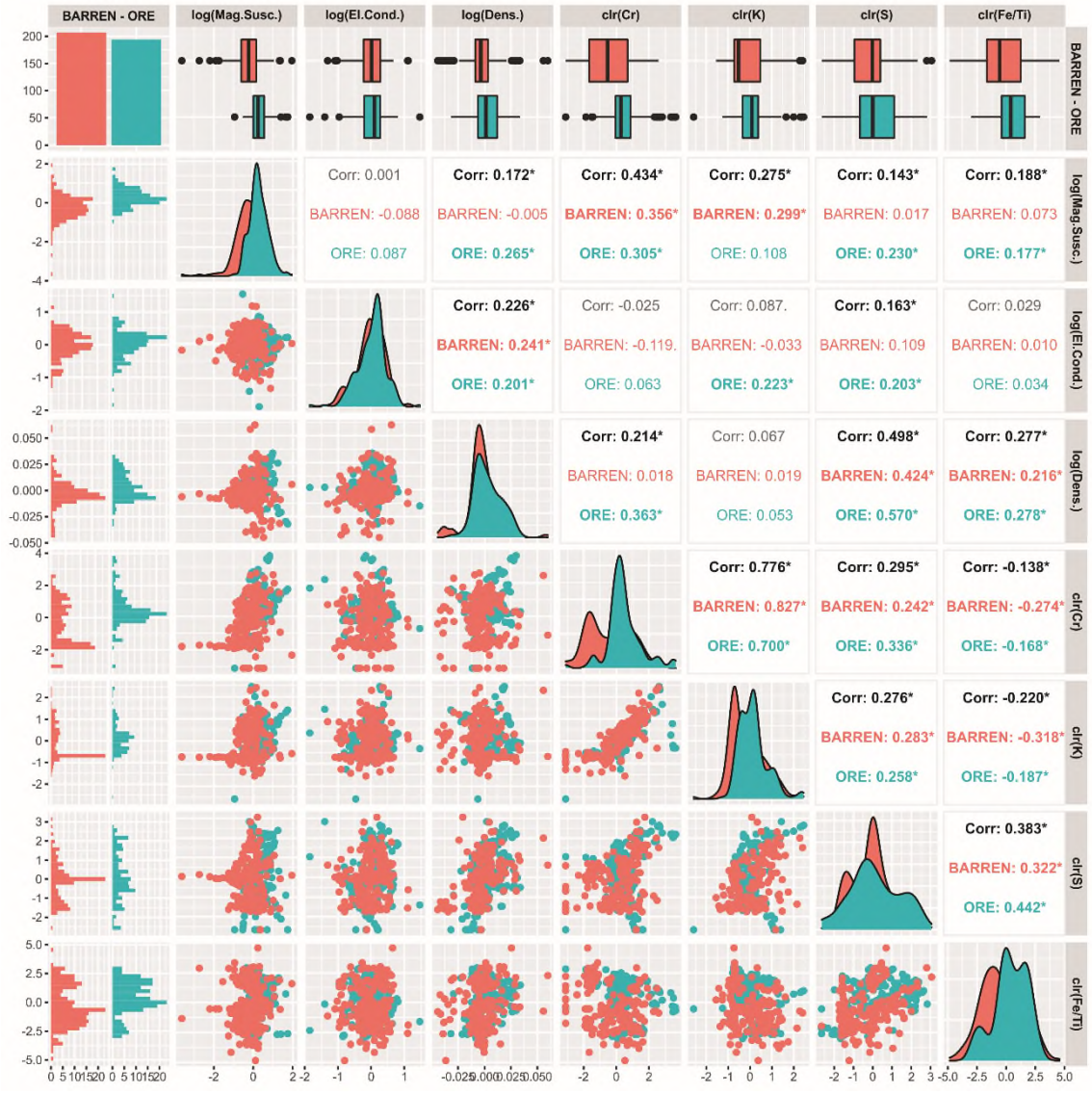


Figura 2-6 – Selected variables transformed to natural log and centered-log ratio distributions and schematic graphics at the lower portion. Histograms, scatterplots, and density plots for each variable and combinations of two variables coded according to mineralization status. At the upper portion, boxplot diagram, colored according to mineralization label and Spearman ranked correlations calculated based on barren samples (red), ore samples (blue), and both (black). The correlations values marked with asterisks were validated for a significance test at the level of 5%. Abbreviations: Mag.Suscep. – Magnetic susceptibility; El.Cond. – Electric conductivity; Dens. – Density; clr – Centered log-ratio.

We calculated Spearman's correlation for each selected variable pair to numerically assess their relationship. Additionally, we validated the correlation values by a statistical significance symmetrical t-test, with a tolerance level of 5%.

The Spearman's correlation is an index calculated according to the rank of samples and used when data follows a multimodal or non-parametrical distribution. For the dataset, the absolute value of significant correlations ranges from 0.143 (among the Magnetic Susceptibility

to with the clr(S) content, considering all samples) to 0.827 (among the values of clr(K) and clr(Cr), considering barren samples).

More important than the correlation strength between the variables is the correlation between ore and barren populations to predict mineralization status correctly. In that perspective, we point out the correlations between the following pairs of variables: log(Density) and clr(Cr), log(Magnetic Susceptibility) and clr(S), log(Density) and log(Magnetic Susceptibility), clr(Fe/Ti) and log(Magnetic Susceptibility), and clr(K) and log(Magnetic Susceptibility).

Also, by the analysis of the boxplot diagram, it is possible to notice that the position of quantiles and median values according to the mineralization status has significant contrast for some of the selected variables, mentioning the log(Magnetic Susceptibility), log(Density), clr(Cr), clr(K), and clr(Fe/Ti) variables.

2.4.2 Mineralization prediction

We predicted the mineralization status of the samples based on the Random Forests classification model described in section 2.3.4.

We prepared an alluvial plot to visualize better the relations between predictions and references across the lithology variations (Figura 2-7). In this type of plot, the reference labels are disposed of in the columns, with the inner subdivisions expressed by their proportion inside the columns. Then, a link (or alluvial) is related to each bar's portions and indicates if a set of samples were labeled as ore or barren by the reference and prediction columns. Additionally, each link is color-coded to aid the visualization of a good prediction or a disagreement with the reference.

It is possible to observe that most misclassifications (i.e., poor mineralization predictions) are associated with quartzite samples predicted as Ore but labeled as Barren samples (False Positive Type, or FP, Figura 2-7). Twenty-seven misclassified samples were

recognized, corresponding to a 5% error rate in the raw accuracy parameter considering all samples. Some other samples were wrongly classified as Barren, while the reference indicated mineralized (False Negative Type, or FN). Furthermore, the FN proportion is almost negligible and is restricted to five conglomerates, quartzite, or ultramafic mineralized samples. Considering the FP, the model predicted 50% mineralized samples than the reference, but most of these samples (n: 27) are from quartzites from the JBA-722 drill core.

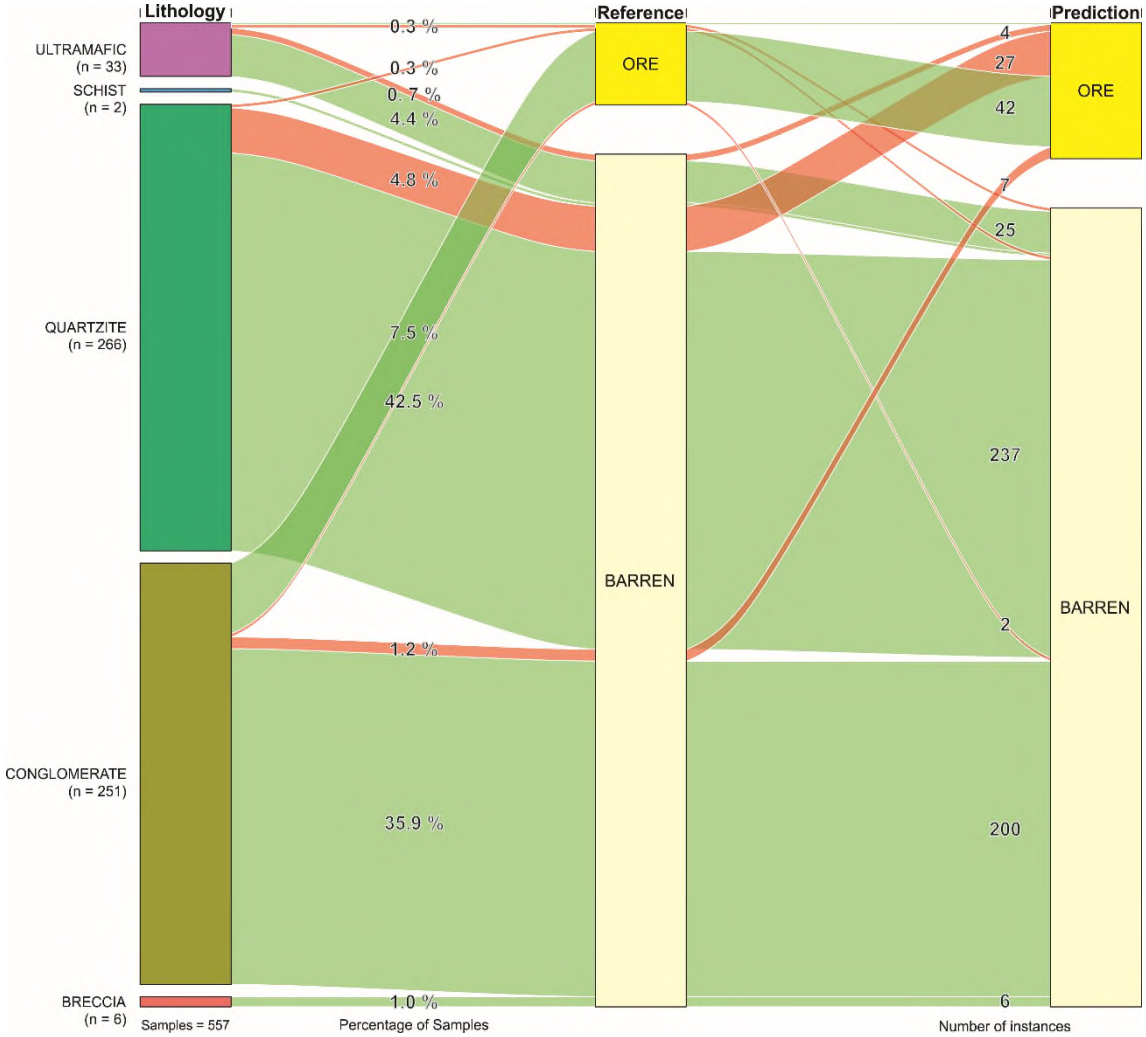


Figura 2-7 – Alluvial validation diagram for model prediction. This diagram uses column bars to identify the proportion of lithology, mineralization status for reference, and model prediction. Each link represents a relation between the three columns, and the links colored green and red represent correct and incorrect predictions, respectively. The proportion of each link through the sample space and the respective number of samples are indicated inside the chart.

2.4.3 *Probabilistic prediction approach*

The mineralization status was predicted considering the average of the votes of all estimators (decision trees in the forest). If the proportion of trees that classified the sample as mineralized (i.e., Ore) is higher than 0.5, the sample is labeled in this way. Then, by analyzing both the reference classification values and the proportion of samples that voted to the considered class, we can build insights about the probability of mineralization of a given sample, even then it is classified as non-mineralized (i.e., Barren, Figura 2-8).

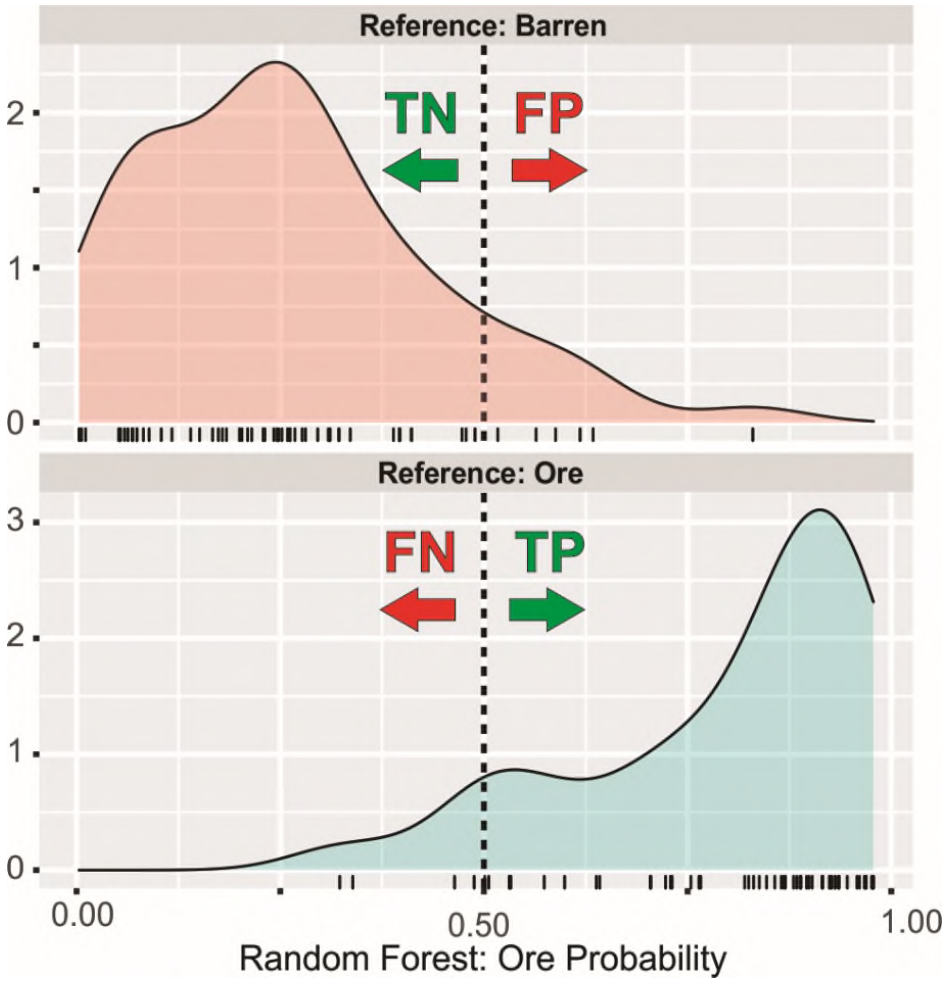


Figura 2-8 – Ore probability analysis for all samples based on the mineralization status of the test dataset (barren samples are represented in the red curve and ore samples in the blue curve). Most barren samples took a low Ore probability, and the mineralized samples got the highest probabilities. The fields of True Negative (TN, i.e., barren samples predicted as non-mineralized), False Positive (FP, barren samples predicted as mineralized), False Negative (FN, ore samples predicted as non-mineralized), and True Positive (TP, ore samples predicted as mineralized) are indicated in the plot. The ticks at the bottom of each plot indicate the calculated probability for each test dataset sample.

Most samples are grouped in the TP and TN fields for all the drill core data. Also, the field with the highest density probability of samples referenced as ore is placed around the probability of 0.9, which suggests that these samples were estimated adequately with a high probability by the model. On the other hand, for the samples classified as barren in the mineralization status, the highest density probability is concentrated around 0.1 to 0.4, indicating that some samples may have any of the considered features indicating a chance of being mineralized.

A schematic strip log color-coded mapping the lithology, mineral status, validation, and ore probabilities parameters provide a better understanding of the mineralization (Figura 2-9).

2.5 Discussion

2.5.1 Mineral Targeting

Pearce et al. (2005) stated that metaconglomerates with blue-gray quartz pebbles and fine-grained disseminated pyrite or hematite usually host gold at the Serra de Jacobina. They noted that the same type of rock with the same pebble size, packing styles, percentage of the matrix, white quartz pebbles, and fuchsite in the matrix tends to have less gold. So, the description mentioned above can be used as a mineral footprint assemblage. A mineral paragenesis schema representing the assemblage, mineral abundance, and the stage of the hydrothermal alteration events is built up using the petrographic and microtexture analysis (Figura 2-10).

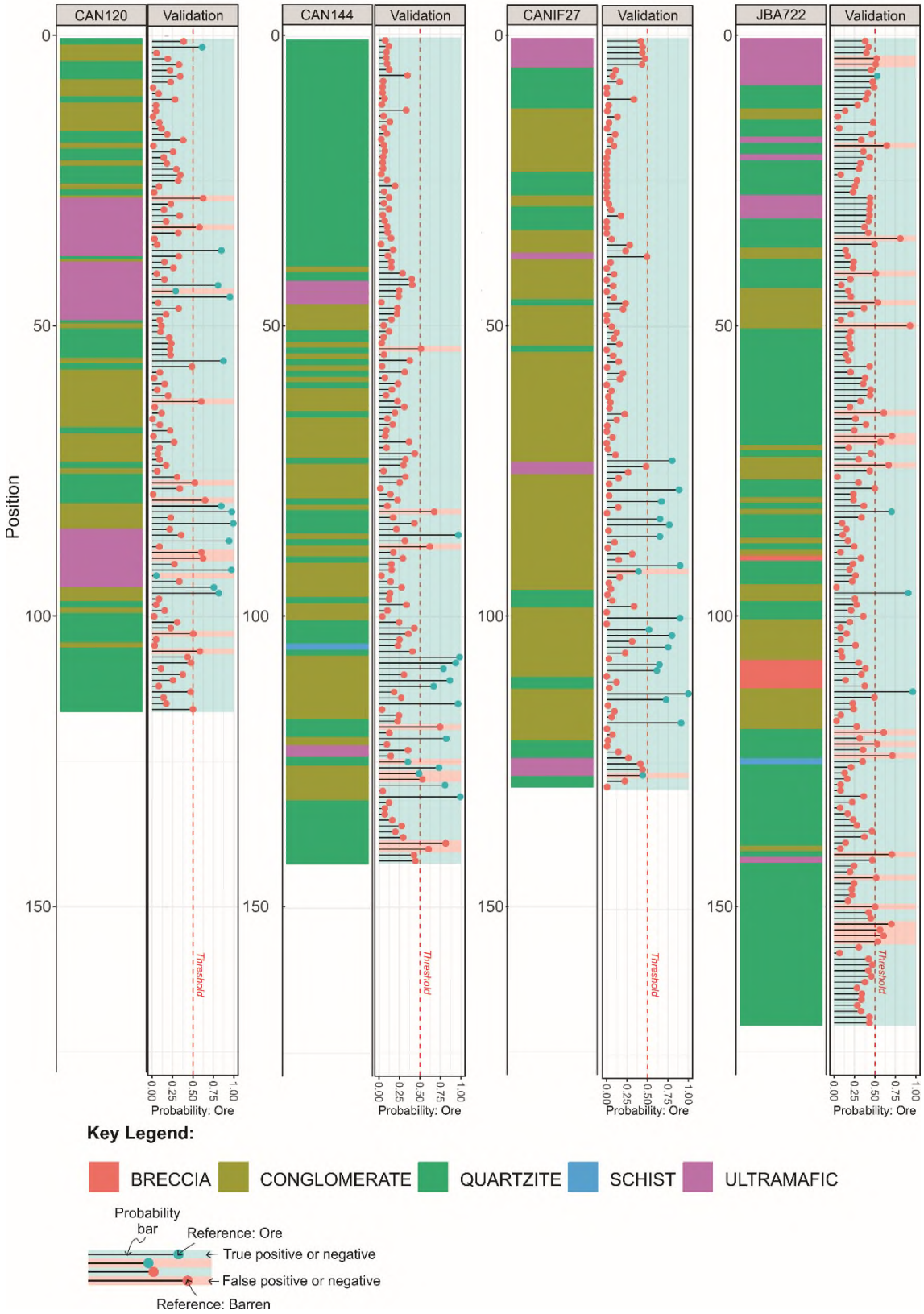


Figura 2-9 – Color-coded strip log according to the lithologies for drill cores studied in this work and respective validation column. The calculated Ore probability is indicated as a bar beside the sample position for each sample and drill core, and the threshold of probability is indicated as the red dashed line. The circle color shows the reference values of the mineralization status at the end of the probability bar. The validation column is color-coded according to the verification of the predicted and reference mineralization status.

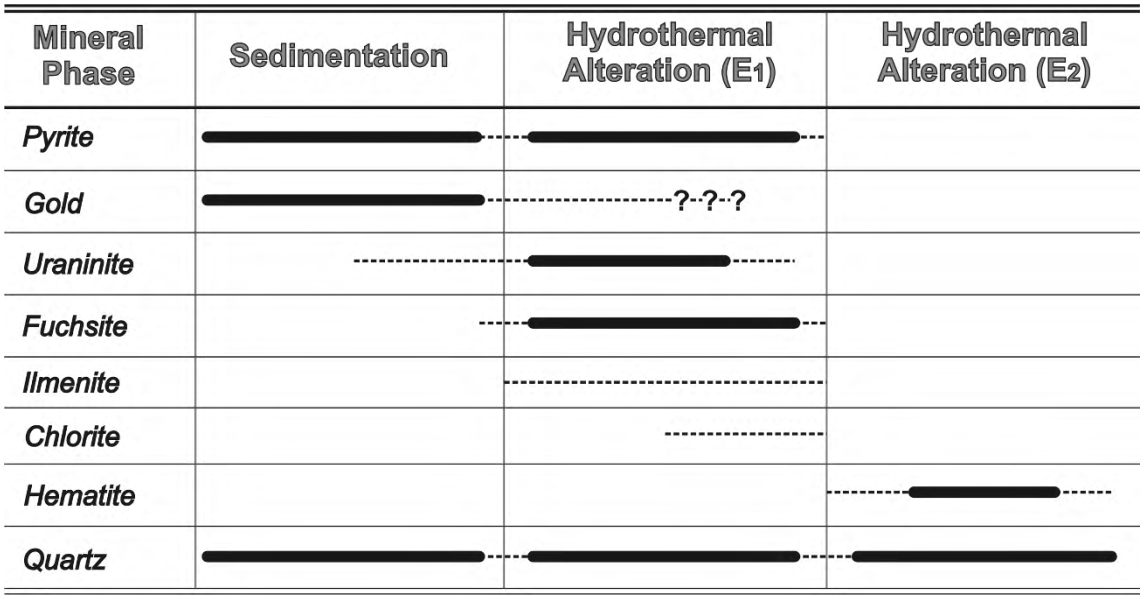


Figura 2-10 – Mineral paragenesis flowchart. Pyrite, gold, and uraninite are present at the sedimentation and first hydrothermal alteration (E1). Fuchsite, ilmenite, chlorite, and hematite were only encountered at the first hydrothermal alteration event (E1). Hematite, as euhedral to subhedral crystals, is only encountered on the second hydrothermal alteration event (E2), in quartz veins, in the middle of the recrystallized matrix, or substituting previous crystals of pyrite.

The first hydrothermal alteration event (E_1) relates to the Paleoproterozoic deformation, either D1 or D2, due to the association with pressure-shadow texture, deformation pattern, and associated metamorphic minerals (e.g., fuchsite, chlorite, remobilized uraninite, or epigenetic pyrite). In addition, some authors described remobilized gold associated with this assemblage, occurring both as free gold and spatially associated with sulfides (Pearson et al., 2005; Teles et al., 2020). This alteration is interpreted here as a product of the first modification of the paleoplacer deposit. It may be due to the syntectonic metamorphism event during the Paleoproterozoic, with or without the participation of fluids derived from the Paleoproterozoic granitic intrusions (Teles et al., 2020). Despite that, the $\Delta^{33}S$ values do not show a disturbance in the isotopic system, which does not favor the entry of external fluids on the basin during metamorphism (Teles et al., 2020).

The second hydrothermal alteration event, also described as hematitization (E_2), affected the Serra do Córrego Formation rocks and is more prominent close to large brittle

structures. Despite this fact, evidence on the relation between hematitization and gold mineralization is still lacking.

RF models estimate the importance of the variables to predict a target by analyzing the mean accuracy decrease for each variable. The raw accuracy of the prediction is compared with the accuracy obtained for the estimators that did not evaluate the variable in question to calculate the Importance parameter (Breiman, 2001). Thus, the average accuracy decrease is calculated, and the variables are ranked according to the degree of importance. The variables were normalized for better comparison. The Variable Importance rank is based on the mean accuracy decrease parameter, and the ranking is driven from the data signature (Breiman, 2001). In addition, as the variable in an estimator is taken by chance, the model bias is not significant.

For the model in question, the rank of the variables shows that properties such as Cr, Magnetic Susceptibility, Fe/Ti ratio, K, Si, Fe, and S, density are the variables of the most significant importance for the assertiveness of the predictions (Figura 2-11).

According to the minerals presented in Figure 10, the variables discussed here can be traced back to the signature of minerals related to the mineralization event (i.e., mineralization footprint). Cr and K are elements present in the mineral structure of the fuchsite, and Fe and S are present in the mineral structure of pyrite and pyrrhotite.

Thus, based on Figures 2-10 and 2-11, it is possible to infer that the mineral contents such as fuchsite $K(Al,Cr)_2(AlSi_3O_{10})(OH)_2$, pyrite FeS_2, intermediate magnetic susceptibility values (possibly associated with the presence of pyrrhotite, Fe_7S_8, or ilmenite, $FeTiO_3$, observed here in thin sections and the literature) and density (associated with the presence of the sulfides and hematite) have an essential role in the quantitative prediction of mineralization through the Random Forests model. All minerals listed above are associated with the hydrothermal alteration event E_1.

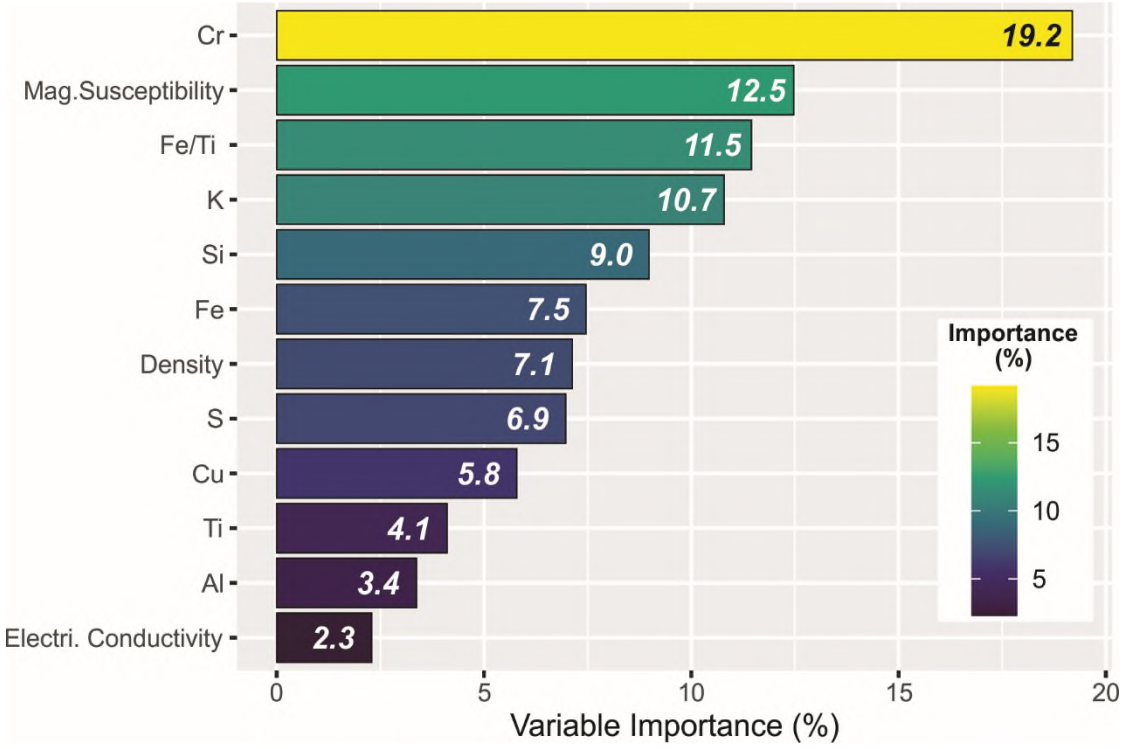


Figura 2-11– Variable importance rank based on the Mean Decrease Accuracy parameter normalized to a percentage distribution. The bars are filled with the coded colors according to their respective importance. Abbreviations: Mag.Susceptibility – Magnetic Susceptibility; Electri. Conductivity – Electric Conductivity.

Variables such as electrical conductivity, and Cu, Ti, and Al content do not significantly influence the model's accuracy and are not determinant to the predictions. This behavior may occur either because these variables do not differentiate barren and ore samples or because they were not involved in the mineralization event. For example, even though the sample has a relevant sulfide content, included in the rock matrix on the metaconglomerates, the electrical conductivity is not favored if the conductive minerals are dispersed and do not show the continuous distribution in the rocks. This interpretation could explain why electrical conductivity values did not play an essential role in the predictions.

2.5.2 Drill core prediction

We also evaluated the model's performance across the samples grouped by the drill core (see Figure 9). The final accuracy of the models significantly changes from the cores CAN120, CAN144, and CANIF27 (ranging from 0.87 to 0.96) to the core JBA722 (0.79).

The drill cores CAN120, CAN144, and CANIF27, intersects the Upper Conglomerate Unit and the Intermediate Quartzite, and the core JBA722 intersects from the Intermediate Quartzite to the Lower Conglomerate Unit (see Figura 2-1c).

This interpretation may imply that Upper Conglomerate and Lower Conglomerate units may differ in the mineralization styles. Nevertheless, this statement should be more carefully investigated in the future, and previous assumptions based on the investigation of drill core samples do not show a relevant distinction between them. In addition, even though the final accuracy is lower in the core samples from JBA722 than in the others, it stills performs satisfactorily.

Upon analyzing the probabilistic approach on all drill core samples (see Figura 2-9), we observe that the barren samples surrounding the ore samples continuously increase their Ore Probability in some cases. For example, this can be observed in core CANIF27 samples, close to positions 75 and 120, and core JBA722, close to positions 15 and 140. So, we conclude that the mineralization affects the surrounding samples, and it suggests that the observed variables are mapping the footprint signature of the mineralization.

2.6 Conclusions

We presented in this work a supervised machine learning approach used to predict the gold mineralization in the quartz-pebble metaconglomerate samples from the Serra do Córrego Formation of the Jacobina Group. The implemented predictive model was based on the Random Forests Algorithm.

The model predicted the mineralization satisfactorily, showing an accuracy of 0.87 and 0.91, respectively, for the train and test datasets. We also ran the model to predict samples grouped by the drill core label, and it resulted in minor but significant accuracy differences between samples in cores from different positions in the Serra do Córrego Formation stratigraphy.

Despite the satisfactory achieved accuracy, there are some ways to further improve the machine learning model performance results. A possible follow-up suggestion for this work, regarding the machine learning implementation, might be to separate samples into three groups (ore, barren and proximal/altered). It is expected that this might solve the false positive classification issue, as those samples might be barely altered. Also, a new modeling attempt would benefit from a larger (and originally balanced) dataset, eliminating the need to generate synthetic samples.

Regarding the interpretation of the model, the Variables Importance analysis showed that properties such as Cr, Magnetic Susceptibility, Fe/Ti ratio, K, Fe, S content, and density are the most significant. Petrographic evidence combined with probabilistic analysis (supported by the Random Forests algorithm) made it possible to explain the relevance of the variables for predicting mineralization status. Therefore, we can infer some mineral targeting criteria to understand the ore formation phenomenon, as the role of the mineral assemblage on the hydrothermal phases described in this work.

Additionally, petrographic information supported by back-scattered electron images and energy dispersion spectroscopy semi-quantitative analysis allowed the interpretation that the rocks from the Serra do Córrego Formation present at least two assemblages of hydrothermal alteration, i.e., a first alteration stage composed by epigenetic pyrite, fuchsite, and uraninite, with minor presence of other sulfides, and a second alteration stage with hematite and quartz, associated with ductile and brittle deformation. There is enough evidence of the gold and uranium remobilization during the second hydrothermal alteration stage. However, the relation between the mineralization and the second described alteration stage still needs more investigation. Therefore, the role of secondary hematite in the gold mineralization must be further investigated, as this behavior may predominate at certain mineralized levels and not be the rule.

This data-driven method is an alternative way to approach mineralization targeting and provides valuable insights in different mineral exploration stages. In addition, petrophysical measurements and geochemistry data can be obtained in the mineral exploration industry with relatively low costs, and their evaluation may be beneficial, whenever the quality control procedures are followed (avoiding undesirable bias). Thus, the use of machine learning algorithms for aiding in the understanding of a complex mineralization is feasible and can bring important practical insights for mineral exploration in many scenarios.

Further, to increase statistical representativeness and put the model in production under a more diverse scenario, the machine learning approach presents the advantage that more samples could be added to the training dataset. Thus, the updated machine learning model could "learn" some new information and theoretically evaluate and understand even minor particularities of a mineral deposit if enough data and contrasting variables are provided.

Author contribution

GFS Conceptualization, Sampling, Investigation, Data Curation, Methodology, Visualization, Writing - Original Draft. **AMS** Supervision, Investigation, Validation, Writing - Review & Editing. **CBT** Supervision, Investigation, Validation, Writing - Review & Editing. **FCJ** Resources, Validation, Writing - Review & Editing. **ELK**: Sampling, Investigation, Validation, Writing - Review & Editing

Acknowledgments

We want to thank Yamana Gold Inc. for its assistance during the collection of drill-core samples. Also, we acknowledge the Geological Survey of Brazil (SGB/CPRM) for financially supporting the fieldwork campaigns. Thanks are extended to the geologists Anderson Dourado, Carina Lopes, Kotaro Ushigasaki, João Larizzatti, Joseneusa Rodrigues, Daniel Miranda, and Valter Sobrinho for aid with field logistics, sampling, or during laboratory analysis. Lastly, the authors thank the improvements made on the original manuscript by the two reviewers, R.S.D. and P.M.P.G. The scripts used in this work may be provided by contacting the corresponding

author. This work was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. A.M. Silva, C.L.B. Toledo, F.C.J., and ELK acknowledge the Brazilian National Council for Scientific and Technological Development (CNPq) for their respective research grants.

Supplementary data

Please refer to the online version to access the supplementary files.

- Table A1 – Gold content (Fire assay) by drill-core
- Table A2 – Petrophysics and XRF data

References

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Stat. Anal. Compos. Data 44, 139–177. https://doi.org/10.1007/978-94-009-4109-0

Alkmin, F.F., Brito Neves, B.B., Alves, J.A.C., 1993. Arcabouço tectônico do Cráton do São Francisco: Uma revisão, in: Dominguez, J.M.L., Misi, A. (Eds.), O Cráton Do São Francisco. SBG - Sociedade Brasileira de Geociências, Salvador, BA, pp. 45–62.

Barbosa, J.S.F., Sabaté, P., 2004. Archean and Paleoproterozoic crust of the São Francisco Craton, Bahia, Brazil: geodynamic features. Precambrian Res. 133, 1–27. https://doi.org/10.1016/j.precamres.2004.03.001

Bérubé, C.L., Olivo, G.R., Chouteau, M., Perrouty, S., Shamsipour, P., Enkin, R.J., Morris, W.A., Feltrin, L., Thiémonge, R., 2018. Predicting rock type and detecting hydrothermal alteration using machine learning and petrophysical properties of the Canadian Malartic ore and host rocks, Pontiac Subprovince, Québec, Canada. Ore Geol. Rev. 96, 130–145. https://doi.org/10.1016/j.oregeorev.2018.04.011

Breiman, L., 2001. Random forests. Mach. Learn. 56, 5–32.

Carneiro, C.D.C., Fraser, S.J., Crósta, A.P., Silva, A.M., Barros, C.E. de M., 2012. Semi-automated geologic mapping using self-organizing maps and airborne geophysics in the Brazilian Amazon. GEOPHYSICS 77, K17–K24. https://doi.org/10.1190/geo2011-0302.1

Carranza, E.J.M., Laborte, A.G., 2016. Data-Driven Predictive Modeling of Mineral Prospectivity Using Random Forests: A Case Study in Catanduanes Island (Philippines). Nat. Resour. Res. 25, 35–50. https://doi.org/10.1007/s11053-015-9268-x

Carranza, E.J.M., Laborte, A.G., 2015. Data-driven predictive mapping of gold prospectivity, Baguio district, Philippines: Application of Random Forests algorithm. Ore Geol. Rev. 71, 777–787. https://doi.org/10.1016/j.oregeorev.2014.08.010

Chawla, N. V, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953

Chen, Y., 2015. Mineral potential mapping with a restricted Boltzmann machine. Ore Geol. Rev. 71, 749–760. https://doi.org/10.1016/j.oregeorev.2014.08.012

Costa, I., Tavares, F., Oliveira, J., 2019. Predictive lithological mapping through machine learning methods: a case study in the Cinzento Lineament, Carajás Province, Brazil. J. Geol. Surv. Brazil 2, 26–36. https://doi.org/10.29396/jgsb.2019.v2.n1.3

da Costa, G., Hofmann, A., Agangi, A., 2020. A revised classification scheme of pyrite in the Witwatersrand Basin and application to placer gold deposits. Earth-Science Rev. 201, 103064. https://doi.org/10.1016/j.earscirev.2019.103064

da Silva, G.F., Ferreira, M.V., Costa, I.S.L., Bernardes, R.B., Mota, C.E.M., Cuadros Jiménez, F.A., 2021. Qmin – A machine learning-based application for processing and analysis of mineral chemistry data. Comput. Geosci. 157, 104949. https://doi.org/10.1016/j.cageo.2021.104949

da Silva, G.F., Larizzatti, J.H., da Silva, A.D.R., Lopes, C.G., Klein, E.L., Uchigasaki, K., 2022. Unsupervised drill core pseudo-log generation in raw and filtered data, a case study in the Rio Salitre greenstone belt, São Francisco Craton, Brazil. J. Geochemical Explor. 232, 106885. https://doi.org/10.1016/j.gexplo.2021.106885

Dentith, M., Enkin, R.J., Morris, W., Adams, C., Bourne, B., 2020. Petrophysics and mineral exploration: a workflow for data analysis and a new interpretation framework. Geophys. Prospect. 68, 178–199. https://doi.org/10.1111/1365-2478.12882

Fisher, L., Gazley, M.F., Baensch, A., Barnes, S.J., Cleverley, J., Duclaux, G., 2014. Resolution of geochemical and lithostratigraphic complexity: A workflow for application of portable X-ray fluorescence to mineral exploration. Geochemistry Explor. Environ. Anal. 14, 149–159. https://doi.org/10.1144/geochem2012-158

Ford, A., 2019. Practical Implementation of Random Forest-Based Mineral Potential Mapping for Porphyry Cu – Au Mineralization in the Eastern Lachlan Orogen , NSW , Australia. Nat. Resour. Res. https://doi.org/10.1007/s11053-019-09598-y

Frimmel, H.E., 2019. The Witwatersrand Basin and Its Gold Deposits, in: Kröner, A., Hofmann, A. (Eds.), The Archaean Geology of the Kaapvaal Craton, Southern Africa. Springer-Verlag, pp. 255–275. https://doi.org/10.1007/978-3-319-78652-0_10

Frimmel, H.E., Le Roex, a P., Knight, J., Minter, W.E.L., 1993. A Case Study of the Post-depositional Alteration. Econ. Geol. 88, 249–265.

Grunsky, E.., 2001. Aspects of multivariate statistical analysis in geology, Computers & Geosciences. https://doi.org/10.1016/s0098-3004(00)00094-7

Hall, B., 2016. Facies classification using machine learning. Lead. Edge 35, 906–909. https://doi.org/10.1190/tle35100906.1

Han, H., Wang, W.Y., Mao, B.H., 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. Lect. Notes Comput. Sci. 3644, 878–887. https://doi.org/10.1007/11538059_91

Hariharan, S., Tirodkar, S., Porwal, A., Bhattacharya, A., Joly, A., 2017. Random Forest-Based Prospectivity Modelling of Greenfield Terrains Using Sparse Deposit Data: An Example from the Tanami Region, Western Australia. Nat. Resour. Res. 26, 489–507. https://doi.org/10.1007/s11053-017-9335-6

Harris, J.R., Grunsky, E., Behnia, P., Corrigan, D., 2015. Data- and knowledge-driven mineral prospectivity maps for Canada's North. Ore Geol. Rev. 71, 788–803. https://doi.org/10.1016/j.oregeorev.2015.01.004

Heilbron, M., Cordani, U.G., Alkmim, F.F., 2017. São Francisco Craton, Eastern Brazil: Tectonic Genealogy of a Miniature Continent. Springer 326. https://doi.org/10.1007/978-3-319-01715-0

Ida, H., 2004. X-ray fluorescence analysis with portable instruments. Kyoto University.

Japkowicz, N., Stephen, S., 2002. The class imbalance problem: A systematic study1. Intell. Data Anal. 6, 429–449. https://doi.org/10.3233/IDA-2002-6504

Koziarski, M., Woźniak, M., Krawczyk, B., 2020. Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise. Knowledge-Based Syst. 204, 106223. https://doi.org/10.1016/j.knosys.2020.106223

Kuhn, S., Cracknell, M.J., Reading, A.M., 2018. Lithologic mapping using Random Forests applied to geophysical and remote-sensing data: A demonstration study from the Eastern Goldfields of Australia. GEOPHYSICS 83, B183–B193. https://doi.org/10.1190/geo2017-0590.1

Ledru, P., Milési, J.P., Johan, V., Sabaté, P., Maluski, H., 1997. Foreland basins and gold-bearing conglomerates: a new model for the Jacobina Basin (São Francisco province, Brazil). Precambrian Res. 86, 155–176.

Leite, C. de M.M., Barbosa, J.S.F., Nicollet, C., Sabaté, P., 2007. Evolução metamórfica/metassomática paleoproterozóica do Complexo Saúde, da Bacia Jacobina e de leucogranitos peraluminosos na parte norte do Cráton do São Francisco. Rev. Bras. Geociências 37, 777–797. https://doi.org/10.25249/0375-7536.2007374777797

Leite, C.M.M., 2002. A evolução geodinâmica da orogênese paloproterozoica nas regiões de Capim Grosso - Jacobina e Pintadas - Mundo Novo (Bahia, Brasil): metamorfismo, anatexia crustal e tectônica. Universidade Federal da Bahia.

Leite, C.M.M., Marinho, M.M., 2012. Serra de Jacobina e Contendas-Mirante, in: Barbosa, J.S.F. (Ed.), Geologia Da Bahia: Pesquisa e Atualização. CBPM - Companhia Baiana de Pesquisa Mineral, pp. 397–441.

Li, X., Zhang, C., Behrens, H., Holtz, F., 2020. Lithos Calculating amphibole formula from electron microprobe analysis data using a machine learning method based on principal components regression. LITHOS 362–363, 105469. https://doi.org/10.1016/j.lithos.2020.105469

Mascarenhas, J.F., Ledru, P., Souza, S.L., Conceição-Filho, V.M., Melo, L.F.A., Lorenzo, C.L., Milési, J.P., 1998. Geologia e recursos minerais do Grupo Jacobina e da parte sul do Greenstone Belt de Mundo Novo. Série Arquivos Abertos, vol. 13. CBPM - Companhia Baiana de Pesquisa Mineral, Salvador, Brazil.

McKay, G., Harris, J.R., 2016. Comparison of the Data-Driven Random Forests Model and a Knowledge-Driven Method for Mineral Prospectivity Mapping: A Case Study for Gold Deposits Around the Huritz Group and Nueltin Suite, Nunavut, Canada. Nat. Resour. Res. 25, 125–143. https://doi.org/10.1007/s11053-015-9274-z

Milési, J.., Ledru, P., Marcoux, E., Mougeot, R., Johan, V., Lerouge, C., Sabaté, P., Bailly, L., Respaut, J.., Skipwith, P., 2002. The Jacobina Paleoproterozoic gold-bearing conglomerates, Bahia, Brazil: a “hydrothermal shear-reservoir” model. Ore Geol. Rev. 19, 95–136. https://doi.org/10.1016/S0169-1368(01)00038-5

Miranda, D.A., Misi, A., Klein, E.L., Castro, M.P., Queiroga, G., 2021. A mineral system approach on the Paleoproterozoic Au-bearing quartz veins of the Jacobina Range, northeastern of the São Francisco Craton, Brazil. J. South Am. Earth Sci. 106. https://doi.org/10.1016/j.jsames.2020.103080

Niiranen, T., Nykänen, V., Lahti, I., 2019. Scalability of the mineral prospectivity modelling – An orogenic gold case study from northern Finland. Ore Geol. Rev. 109, 11–25. https://doi.org/10.1016/j.oregeorev.2019.04.002

Pearson, W., Macêdo, P.M.M., Rúbio, A., Lorenzo, C.L., Karpeta, P., 2005. Geology and gold mineralization of the Jacobina Mine and Bahia Gold Belt, Bahia, Brazil and comparison to Tarkwa and Witwatersrand., in: Proceedings, Geological Society of Nevada Symposium, Vol. 1. Geological Society of Nevada Symposium, Reno, Nevada, USA., pp. 757–786.

Piercey, S.J., 2014. Modern analytical facilities 2. A review of quality assurance and quality control (qa/qc) procedures for lithogeochemical data. Geosci. Canada 41, 75–88. https://doi.org/10.12789/geocanj.2014.41.035

Prado, E.M.G., de Souza Filho, C.R., Carranza, E.J.M., Motta, J.G., 2020. Modeling of Cu-Au prospectivity in the Carajás mineral province (Brazil) through machine learning: Dealing with imbalanced training data. Ore Geol. Rev. 124, 103611. https://doi.org/10.1016/j.oregeorev.2020.103611

Reis, C., Menezes, R.C.L., Miranda, D.A., Santos, F.P. dos, Santos, R.S.V. dos, Meneses, A.R., 2021. Áreas de Relevante Interesse Mineral (Arim) Integração Geológica E Avaliação do

795 Potencial Metalogenético da Serra de Jacobina e do Greenstone Belt Mundo Novo. Serviço
796 Geológico do Brasil - CPRM, Salvador, Brazil.
797 Robb, L.J., Meyer, F.M., 1991. A contribution to recent debate concerning epigenetic versus
798 syngenetic mineralization processes in the Witwatersrands Basin. Econ. Geol. 86, 396–
799 401.
800 Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015.
801 Machine learning predictive models for mineral prospectivity: An evaluation of neural
802 networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71,
803 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001
804 Saljoughi, B.S., Hezarkhani, A., 2018. A comparative analysis of artificial neural network
805 (ANN), wavelet neural network (WNN), and support vector machine (SVM) data-driven
806 models to mineral potential mapping for copper mineralizations in the Shahr-e-Babak
807 region, Kerman, Iran. Appl. Geomatics 10, 229–256. https://doi.org/10.1007/s12518-018-
808 0229-z
809 Santos, F.P. dos, Chemale Junior, F., Meneses, A.R.A.S., 2019. The nature of the
810 Paleoproterozoic orogen in the Jacobina Range and adjacent areas, northern São Francisco
811 Craton, Brazil, based on structural geology and gravimetric modeling. Precambrian Res.
812 332, 105391. https://doi.org/10.1016/j.precamres.2019.105391
813 Saporetti, C.M., da Fonseca, L.G., Pereira, E., de Oliveira, L.C., 2018. Machine learning
814 approaches for petrographic classification of carbonate-siliciclastic rocks using well logs
815 and textural information. J. Appl. Geophys. 155, 217–225.
816 https://doi.org/10.1016/j.jappgeo.2018.06.012
817 Teixeira, J.B.G., De Souza, J.A.B., Da Silva, M. da G., Leite, C.M.M., Barbosa, J.S.F.S.F.,
818 Coelho, C.E.S., Abram, M.B., Filho, V.M.C., Iyer, S.S.S., 2001. Gold mineralization in the

819 Serra de Jacobina region, Bahia Brazil: tectonic framework and metallogenesis. Miner.
820 Depos. 36, 332–344. https://doi.org/10.1007/s001260100174
821 Teixeira, W., Oliveira, E.P., Marques, L.S., 2017. Nature and evolution of the Archean crust of
822 the São Francisco Craton, in: Heilbron, M., Alkmin, F.F., Cordani, U.G. (Eds.), The São
823 Francisco Craton and Its Margins, Eastern Brazil. Springer-Verlag, pp. 29–56.
824 Teles, G., Jr, F.C., Oliveira, C.G. De, Chemale, F., de Oliveira, C.G., 2015. Paleoarchean record
825 of the detrital pyrite-bearing, Jacobina Au-U deposits, Bahia, Brazil. Precambrian Res. 256,
826 289–313. https://doi.org/10.1016/j.precamres.2014.11.004
827 Teles, G.S., Chemale, F., Ávila, J.N., Ireland, T.R., Dias, A.N.C., Cruz, D.C.F., Constantino,
828 C.J.L., 2020. Textural and geochemical investigation of pyrite in Jacobina Basin, São
829 Francisco Craton, Brazil: Implications for paleoenvironmental conditions and formation of
830 pre-GOE metaconglomerate-hosted Au-(U) deposits. Geochim. Cosmochim. Acta 273,
831 331–353. https://doi.org/10.1016/j.gca.2020.01.035
832 Thermo-Scientific, 2013. Mining and exploration: Solutions from early-stage discovery
833 through mineral processing. San Jose, California.
834 Torppa, J., Nykänen, V., Molnár, F., 2019. Unsupervised clustering and empirical fuzzy
835 memberships for mineral prospectivity modelling. Ore Geol. Rev. 107, 58–71.
836 https://doi.org/10.1016/j.oregeorev.2019.02.007
837 Wickham, H., 2014. Tidy Data. J. Stat. Softw. 59. https://doi.org/10.18637/jss.v059.i10
838 Yamana Gold, 2020. Annual Report 2020 - NI43-101, 176p.
839 Yousefi, M., Nykänen, V., 2016. Data-driven logistic-based weighting of geochemical and
840 geological evidence layers in mineral prospectivity mapping. J. Geochemical Explor. 164,
841 94–106. https://doi.org/10.1016/j.gexplo.2015.10.008

842 Zhu, W., Zeng, N., Wang, N., 2010. Sensitivity, specificity, accuracy, associated confidence
843 interval and ROC analysis with practical SAS® implementations. Northeast SAS Users Gr.
844 2010 Heal. Care Life Sci. 1–9.
845 Zuo, R., 2017. Machine Learning of Mineralization-Related Geochemical Anomalies: A
846 Review of Potential Methods. Nat. Resour. Res. 26, 457–464.
847 https://doi.org/10.1007/s11053-017-9345-4
848

5 DISCUSSÕES E CONSIDERAÇÕES FINAIS

Os resultados apresentados no capítulo 2 mostram que propriedades como suscetibilidade magnética, densidade e teores de Cr, K, Fe, e S são as variáveis mais significativas para a predição de amostras mineralizadas (não nessa ordem de importância). Evidências petrográficas combinadas com análises probabilísticas (derivadas das inferências estatísticas utilizando o algoritmo *Random Forests*) permitiram explicar a relevância das variáveis para predição da mineralização. Portanto, podemos inferir alguns critérios de direcionamento mineral para entender o fenômeno de formação do minério, como o papel da assembleia mineral nas fases hidrotermais descritas neste trabalho.

Assim, sugerimos essa abordagem na prospecção local, com a ressalva que os modelos foram construídos para serem representativos, porém a amostragem realizada pode não abranger toda variância da população mineralizada, assim mais amostras podem ser adicionadas ao conjunto de dados de treinamento e o modelo deve ser atualizado. Desse modo, o modelo de aprendizado de máquina poderia "aprender" novas informações e avaliar particularidades de um depósito mineral se dados suficientes e variáveis contrastantes forem fornecidos.

Os resultados apresentados no capítulo 3 mostram que a investigação da assinatura de elementos traço em pirita ao longo das unidades do Grupo Jacobina sugere uma falta de contraste que pode implicar na manutenção da fonte de sedimentos durante a formação da bacia Jacobina, ou a posterior equilíbrio químico durante os estágios de alteração metamórficos/hidrotermais. No entanto, para elucidar essas questões, recomendamos a avaliação de mais amostras e análises em trabalhos futuros, pois isso pode ajudar a reduzir o viés amostral.

Para fins de exploração mineral, nossos resultados apontam para um importante papel de alguns minerais acessórios anteriormente negligenciados nas partes alteradas pelos eventos epigenéticos dos depósitos da Serra de Jacobina, como esfalerita, calcopirita, pirrotita e outros.

Esses minerais são considerados farejadores da alteração epigenética nos depósitos, e podem estar relacionados ao ouro livre, conforme indicado pela análise de dendrogramas e confirmado em lâminas delgadas. No entanto, concluímos que a alteração epigenética nos depósitos pode ter resultados positivos ou negativos sobre o conteúdo metálico (*endowment*) do depósito, pois uma forte modificação e consequente mobilização de ouro não canalizada em um mecanismo de concentração eficaz poderia espalhar o ouro em várias pequenas frações.

Os resultados parciais de reflectância espectrorradiométrica e geoquímica, apresentados no capítulo 4, indicam que a maioria das amostras mineralizadas possuem mais de 50% de pirita na composição das frações analisadas, sendo a fuchsita e a illita minerais secundários. Goethita, hematita e muscovita ocorrem como fases acessório na assembleia mineral e, exceto nas amostras mineralizadas do testemunho JBA-722, os minerais de óxido de ferro não são dominantes nos metaconglomerados auríferos. Além disso, os dados de litoquímicos validaram os valores inferidos de abundância dos minerais, destacando os teores de Cr, K, Al (no caso de fucsita) e Fe e S (no caso de minerais de óxido de ferro e pirita).

Assim sendo, utilizando os resultados obtidos pelas ferramentas descritas nessa tese, e observando a dimensão das amostragens e as escalas de trabalho, resumimos os critérios de prospecção encontrados na Figura 5-1. Separamos as informações de acordo com a assinatura das propriedades físicas e químicas nas zonas mineralizada, proximal e estéril. Finalizando, sugerimos para trabalhos futuros que as abordagens aqui apresentadas sejam traduzidas em critérios de exploração para modelamento de potencial mineral e de prospecção local para que a abordagem seja validada (ver Apêndice B).

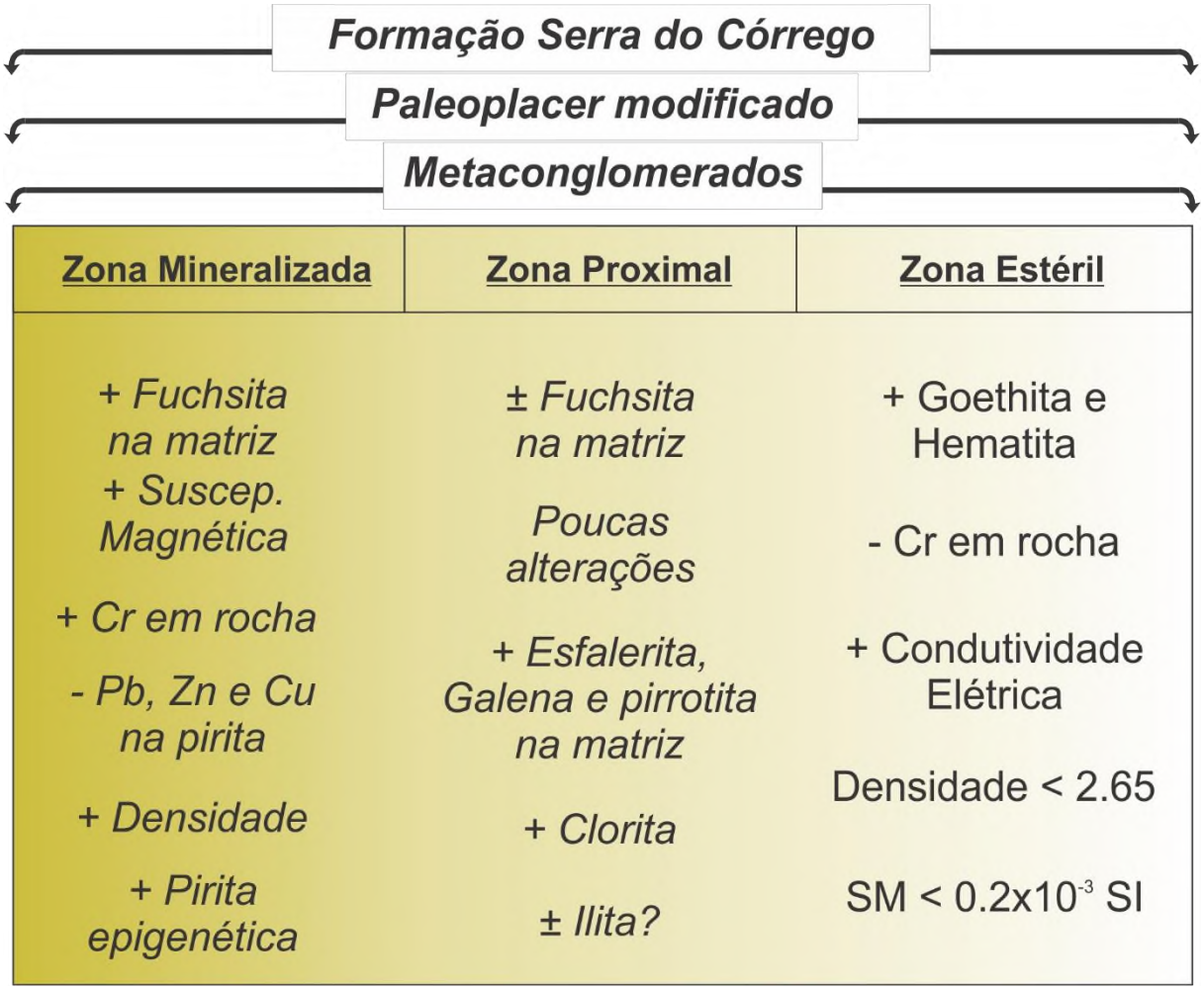


Figura 5-1: Quadro resumo dos resultados obtidos na tese a respeito das propriedades físicas e químicas das rochas e minerais analisados, no escopo das amostras de metaconglomerados com mineralização do tipo paleoplacer modificado da Formação Serra do Córrego, Grupo Jacobina. Os controles do componente sedimentar da mineralização não foram abordados neste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

Abedi, M., Norouzi, G.H., Bahroudi, A., 2012. Support vector machine for multi-classification of mineral prospectivity areas. Comput. Geosci. 46, 272–283. https://doi.org/10.1016/j.cageo.2011.12.014

Agterberg, F.P., Bonham-Carter, G.F., 2005. Measuring the Performance of Mineral-Potential Maps. Nat. Resour. Res. 14, 1–17. https://doi.org/10.1007/s11053-005-4674-0

Bérubé, C.L., Olivo, G.R., Chouteau, M., Perrouty, S., Shamsipour, P., Enkin, R.J., Morris, W.A., Feltrin, L., Thiémonge, R., 2018. Predicting rock type and detecting hydrothermal alteration using machine learning and petrophysical properties of the Canadian Malartic ore and host rocks, Pontiac Subprovince, Québec, Canada. Ore Geol. Rev. 96, 130–145. https://doi.org/10.1016/j.oregeorev.2018.04.011

Bonham-Carter, G.F., 1994. Geographic Information Systems for Geoscientists, 1st Ed. ed. Pergamon.

Breiman, L., 2001. Random forests. Mach. Learn. 56, 5–32.

Breiman, L., Cutler, A., Forests, R., Ho, T.K., Labs, B., Kleinberg, E., Breiman, L., 1995. Random forest 1–5.

Carranza, E.J.M. (Emmanuel J.M., 2009. Geochemical anomaly and mineral prospectivity mapping in GIS. Elsevier.

Carranza, E.J.M., Laborte, A.G., 2016. Data-Driven Predictive Modeling of Mineral Prospectivity Using Random Forests: A Case Study in Catanduanes Island (Philippines). Nat. Resour. Res. 25, 35–50. https://doi.org/10.1007/s11053-015-9268-x

Carranza, E.J.M., Laborte, A.G., 2015. Random forest predictive modeling of mineral prospectivity with small number of prospects and data with missing values in Abra (Philippines). Comput. Geosci. 74, 60–70. https://doi.org/10.1016/j.cageo.2014.10.004

Costa, I., Tavares, F., Oliveira, J., 2019. Predictive lithological mapping through machine learning methods: a case study in the Cinzento Lineament, Carajás Province, Brazil. J. Geol. Surv. Brazil 2, 26–36. https://doi.org/10.29396/jgsb.2019.v2.n1.3

Cover, T.M., Hart, P.E., 1967. Nearest Neighbor Pattern Classification. IEEE Trans. Inf. Theory 13, 21–27. https://doi.org/10.1109/TIT.1967.1053964

Davies, J. C. 2002. Statistics and Data Analysis in Geology, 3rd Edition. John Wiley & Sons, Nre York –USA. 656 pages.

Dentith, M., Enkin, R.J., Morris, W., Adams, C., Bourne, B., 2020. Petrophysics and mineral exploration: a workflow for data analysis and a new interpretation framework. Geophys. Prospect. 68, 178–199. https://doi.org/10.1111/1365-2478.12882

Dentith, M.C., Mudge, S.T., 2014. Geophysics for the mineral exploration geoscientist, 1st ed, Cambridge University Press. Cambridge University Press, Cambridge, UK.

Fleming, S.W., Watson, J.R, Ellenson, A., Canon, A.J., Vesselinov, V.C. 2021. Machine learning in Earth and environmental science requires education and research policy reforms. *Nature Geoscience*. 14, 878-880(2021). https://doi.org/10.1038/s41561-021-00881-3

Frimmel, H.E., 2019. The Witwatersrand Basin and Its Gold Deposits, in: Kröner, A., Hofmann, A. (Eds.), The Archaean Geology of the Kaapvaal Craton, Southern Africa. Springer-Verlag, pp. 255–275. https://doi.org/10.1007/978-3-319-78652-0_10

Frimmel, H.E., 2014. Chapter 10 A Giant Mesoarchean Crustal Gold-Enrichment Episode : Possible Causes and Consequences for Exploration 209–234.

Frimmel, H.E., Groves, D.I., Kirk, J., Ruiz, J., Chesley, J., Minter, W.E.L., 2019. The Formation and Preservation of the Witwatersrand Goldfields, the World's Largest Gold Province. One Hundredth Anniv. Vol. 769–797. https://doi.org/10.5382/av100.23

Guimarães, F.S., de Freitas, M.E., Rios, F.J., Pedrosa, T.A., 2019. Mineralogical characterization and origin of uranium mineralization in Witwatersrand-like metaconglomerate of the Moeda Formation, Quadrilátero Ferrífero, Brazil. Ore Geol. Rev. 106, 423–445. https://doi.org/10.1016/j.oregeorev.2019.01.016

Hagemann, S.G., Lisitsin, V.A., Huston, D.L., 2016. Mineral system analysis: Quo vadis. Ore Geol. Rev. 76, 504–522. https://doi.org/10.1016/j.oregeorev.2015.12.012

Hennigh, Q., 2016. Conglomerate-Hosted Gold Mineralization in the Pilbara, Western Australia, in: Association for Mineral Exploration British Columbia, Roundup 2016, Abstracts: Core Shack. pp. 49–50.

Hronsky, J.M.A., Kreuzer, O.P., 2019. Applying spatial prospectivity mapping to exploration targeting: Fundamental practical issues and suggested solutions for the future. Ore Geol. Rev. 107, 647–653. https://doi.org/10.1016/j.oregeorev.2019.03.016

Joly, A., Porwal, A., McCuaig, T.C., 2012. Exploration targeting for orogenic gold deposits in the Granites-Tanami Orogen: Mineral system analysis, targeting model and prospectivity analysis. Ore Geol. Rev. 48, 349–383. https://doi.org/10.1016/J.OREGEOREV.2012.05.004

Klein, E.L., Rodrigues, J.B., Queiroz, J.D.S., Oliveira, R.G., Guimarães, S.B., Chaves, C.L., 2017. Deposition and tectonic setting of the Palaeoproterozoic Castelo dos Sonhos metasedimentary formation, Tapajós Gold Province, Amazonian Craton, Brazil: age and isotopic constraints. Int. Geol. Rev. 59, 864–883. https://doi.org/10.1080/00206814.2016.1237311

Ledru, P., Milési, J.P., Johan, V., Sabaté, P., Maluski, H., 1997. Foreland basins and gold-bearing conglomerates: a new model for the Jacobina Basin (São Francisco province, Brazil). Precambrian Res. 86, 155–176.

Mccuaig, T.C., Beresford, S., Hronsky, J., 2010. Translating the mineral systems approach into an effective exploration targeting system. Ore Geol. Rev. 38, 128–138. https://doi.org/10.1016/j.oregeorev.2010.05.008

McCuaig, T.C., Hronsky, J., 2014. The mineral systems concept: the key to exploration targeting. Soc. Econ. Geol. - Spec. Publ. 18, 153–175. https://doi.org/10.1080/03717453.2017.1306274

Milesi, J.., Ledru, P., Marcoux, E., Mougeot, R., Johan, V., Lerouge, C., Sabaté, P., Bailly, L., Respaut, J.., Skipwith, P., 2002. The Jacobina Paleoproterozoic gold-bearing conglomerates, Bahia, Brazil: a “hydrothermal shear-reservoir” model. Ore Geol. Rev. 19, 95–136. https://doi.org/10.1016/S0169-1368(01)00038-5

Minter, W.E.L., Renger, F.E., Siegers, A., 1990. Early Proterozoic gold placers of the Moeda Formation within the Gandarela Syncline, Minas Gerais, Brazil. Econ. Geol. 85, 943–951. https://doi.org/10.2113/gsecongeo.85.5.943

Pigois, J.-P., Groves, D.I., Fletcher, I.R., McNaughton, N.J., Snee, L.W., 2003. Age constraints on Tarkwaian palaeoplacer and lode-gold formation in the Tarkwa-Damang district, SW Ghana. Miner. Depos. 38, 695–714. https://doi.org/10.1007/s00126-003-0360-5

Razali, N.M., Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. 2, 21–33.

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Royston, J.P., 1982. An Extension of Shapiro and Wilk’s W Test for Normality to Large Samples. Appl. Stat. 31, 115–124. https://doi.org/10.2307/2347973

Saculinggan, M., Balase, E.A., 2013. Empirical power comparison of goodness of fit tests for normality in the presence of outliers. J. Phys. Conf. Ser. 435. https://doi.org/10.1088/1742-6596/435/1/012041

Shapiro, S.S., Wilk, M.B., 1965. An analysis of variance test for normality (complete samples). Biometrika 52, 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Teixeira, J.B.G., De Souza, J.A.B., Da Silva, M. da G., Leite, C.M.M., Barbosa, J.S.F.S.F., Coelho, C.E.S., Abram, M.B., Filho, V.M.C., Iyer, S.S.S., 2001. Gold mineralization in the Serra de Jacobina region, Bahia Brazil: tectonic framework and metallogenesis. Miner. Depos. 36, 332–344. https://doi.org/10.1007/s001260100174

Whymark, W.E., Frimmel, H.E., 2018. Regional gold-enrichment of conglomerates in Paleoproterozoic supergroups formed during the 2.45 Ga rifting of Kenorland. Ore Geol. Rev. 101, 985–996. https://doi.org/10.1016/j.oregeorev.2017.04.003

Wickham, H., 2014. Tidy Data. J. Stat. Softw. 59. https://doi.org/10.18637/jss.v059.i10

Wyborn, L.A.I., Heinrich, C.A., Jaques, A.L., 1994. Australian Proterozoic mineral systems: essential ingredients and mappable criteria. Aust. Inst. Min. Metall. Publ. Ser. 109–115.

Yap, B.W., Sim, C.H., 2011. Comparisons of various types of normality tests. J. Stat. Comput. Simul. 81, 2141–2155. https://doi.org/10.1080/00949655.2010.520163

Zuo, R., 2017. Machine Learning of Mineralization-Related Geochemical Anomalies: A Review of Potential Methods. Nat. Resour. Res. 26, 457–464. https://doi.org/10.1007/s11053-017-9345-4

Zuo, R., Zhang, Z., Zhang, D., Carranza, E.J.M., Wang, H., 2015. Evaluation of uncertainty in mineral prospectivity mapping due to missing evidence: A case study with skarn-type Fe deposits in Southwestern Fujian Province, China. Ore Geol. Rev. 71, 502–515. https://doi.org/10.1016/j.oregeorev.2014.09.024

APÊNDICE A – Lista de Publicações

Trabalho Publicado em Eventos Científicos

Silva, G.F., Silva, A.D.R., Lopes, C.G., Klein, E.L., Silva, A.M., Toledo, C.L.B. 2019. *k-Nearest Neighbors algorithm applied to lithology prediction based on handheld geochemistry and petrophysics analysis – a case study of metasedimentary and intrusive rocks from the Jacobina Range*. Anais do IV Simpósio Brasileiro de Metalogenia, p125-126. Gramado – RS.

Artigo científico aceitos e em preparação

Silva, G. F., Silva, A.M., Toledo, C.L.B., Chemale-Junior, F., Klein, E.L. 2022. *Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil*". Manuscrito submetido ao periódico *Journal of South American Earth Sciences*, e aceito para revisão sob o código SAMES-D-21-00534R2.

Silva, G. F., Silva, A.M., Toledo, C.L.B., Teles, G.S., Chemale-Junior, F., Klein, E.L., Braga, A.A. 2022. *Machine learning applied to the analysis of mineral chemistry in pyrite grains from the Jacobina gold deposits, São Francisco Craton, Brazil: geochemical patterns and implications to mineral exploration.* Manuscrito submetido ao periódico Journal of Geochemical Exploration, e aceito para revisão sob o código GEXPLO-D-22-00084.

Silva, G. F., Silva, A.M., Toledo, C.L.B., Chemale-Junior, F. 2022. *Unmixing spectral signal and estimating the mineral composition of metaconglomerates using dimensionality reduction and relative distance concepts*. Manuscrito em preparação.

APÊNDICE B – Sistemas Minerais e sua aplicação na exploração mineral

Sistemas Minerais

O conceito de sistemas minerais, inicialmente desenvolvido por Wyborn et al. (1994), deriva do conceito análogo de sistemas petrolíferos, muito difundido na indústria do petróleo a partir de meados da década de 1980. Os sistemas petrolíferos categorizam todos os processos e elementos geológicos necessários para a formação e armazenamento de óleo e gás. Da mesma forma, os sistemas minerais foram definidos como "todos os fatores geológicos que controlam a geração e preservação de depósitos minerais, cujos processos são relacionados à mobilização do minério desde a fonte até a região de concentração, transporte e acumulação, e sua posterior preservação ao longo da história geológica" (Wyborn et al., 1994).

Ainda segundo Wyborn et al. (1994), a maior parte dos corpos de minério possuem menos do que 1km² de expressão, não consistindo portanto de um alvo significativo para a prospecção mineral. Felizmente, apesar de os depósitos não terem uma área muito expressiva e serem resultados de uma coincidência excepcional de determinados processos geológicos, estes processos podem ser mapeáveis em escalas regionais, sendo chave importante para o processo de prospecção. Em outras palavras, apesar de o depósito consistir em áreas de centenas de metros, o sistema total de interação entre fluido-rocha encaixante-mineralização pode se estender por áreas de até poucas dezenas de quilômetros, sendo detectáveis por determinadas ferramentas (Wyborn et al., 1994).

Os fatores críticos para a caracterização de qualquer sistema mineral incluem:

1. fonte do fluido mineralizadores e dos compostos ligantes;
2. fonte dos metais e outros componentes da mineralização;
3. caminhos de migração (*pathways*) do fluido mineralizante;
4. gradiente térmico;
5. Fonte de energia (por vezes relacionado ao item 4);
6. Estruturas ou mecanismos de concentração;
7. Condições químicas e físicas para a deposição da mineralização.

McCuaig et al., (2010) avaliam que o conceito de sistema mineral evoluiu à medida que foi aceito gradualmente pela indústria e academia nos 15 anos anteriores, apesar de que fora pouco utilizado na rotina das empresas de mineração. É enfatizado que a linha estruturante do conceito de sistema mineral está associada à compreensão dos vários processos geológicos que operam em todas as escalas, ao invés de focar na compreensão das características particulares de depósitos específicos em sua escala local. Desta forma, ressaltando que há certa dificuldade

em traduzir essa mudança de paradigma para a realidade da prospecção. McCuaig et al., (2010) propõem uma sistemática de quatro passos para guiar o *pipeline* de exploração através da ótica dos Sistemas Minerais.

Ainda nesta linha, McCuaig & Hronsky (2014) também contribuem para a evolução do conceito de Sistema Mineral. Estes autores postulam que a existência de um depósito mineral está condicionada a sobreposição de pelo menos quatro fatores críticos ao longo da história geológica, sendo eles fertilidade do terreno, favorabilidade geodinâmica, arquitetura litosférica e a posterior preservação das zonas de concentração de minério em um sistema mineral.

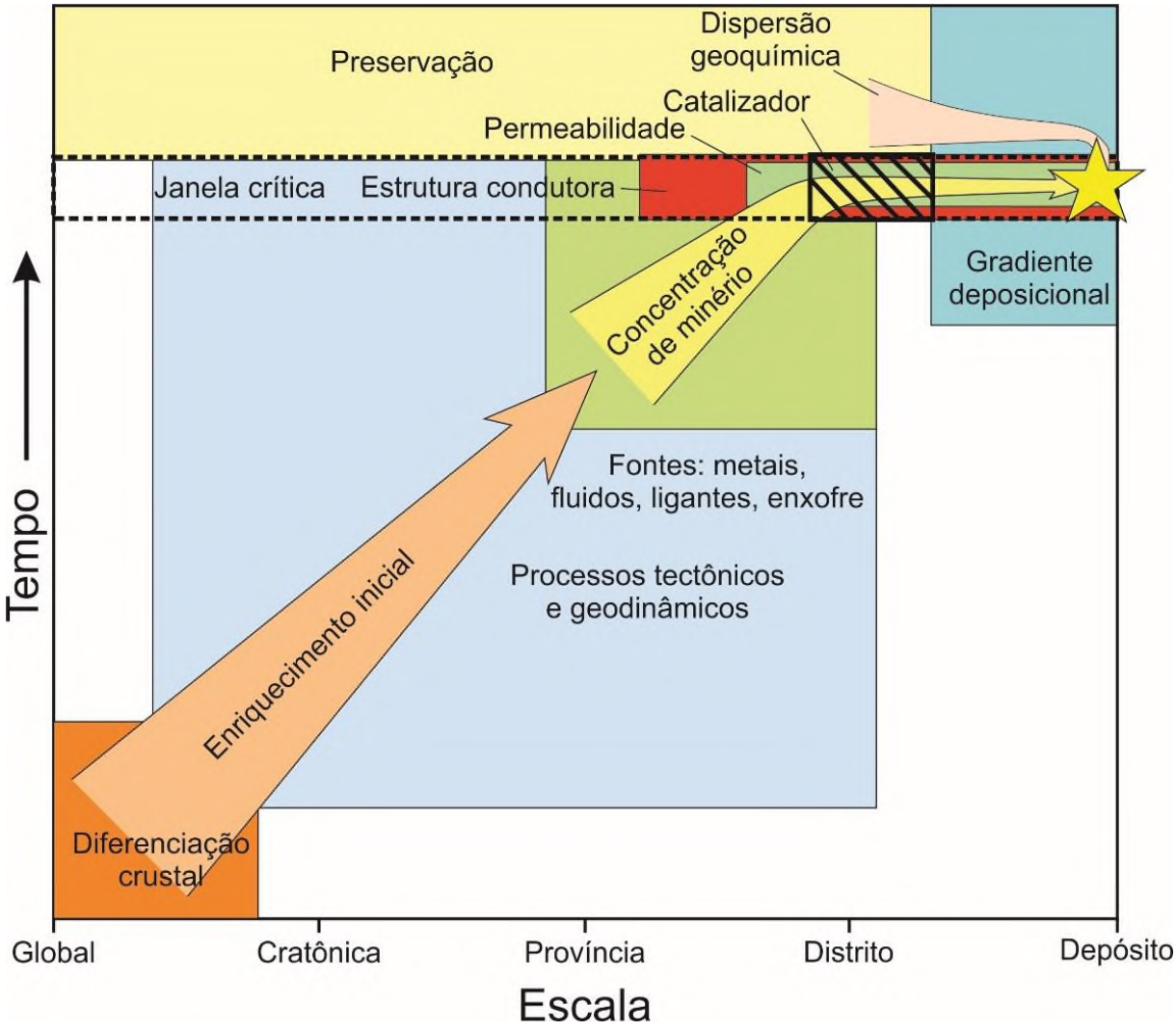


Figura B- 1 – Síntese dos processos envolvidos nos Sistemas Minerais nas suas variadas escalas, desde a diferenciação crustal até a dispersão do fluido em superfície (Fonte: www.ga.gov.au.)

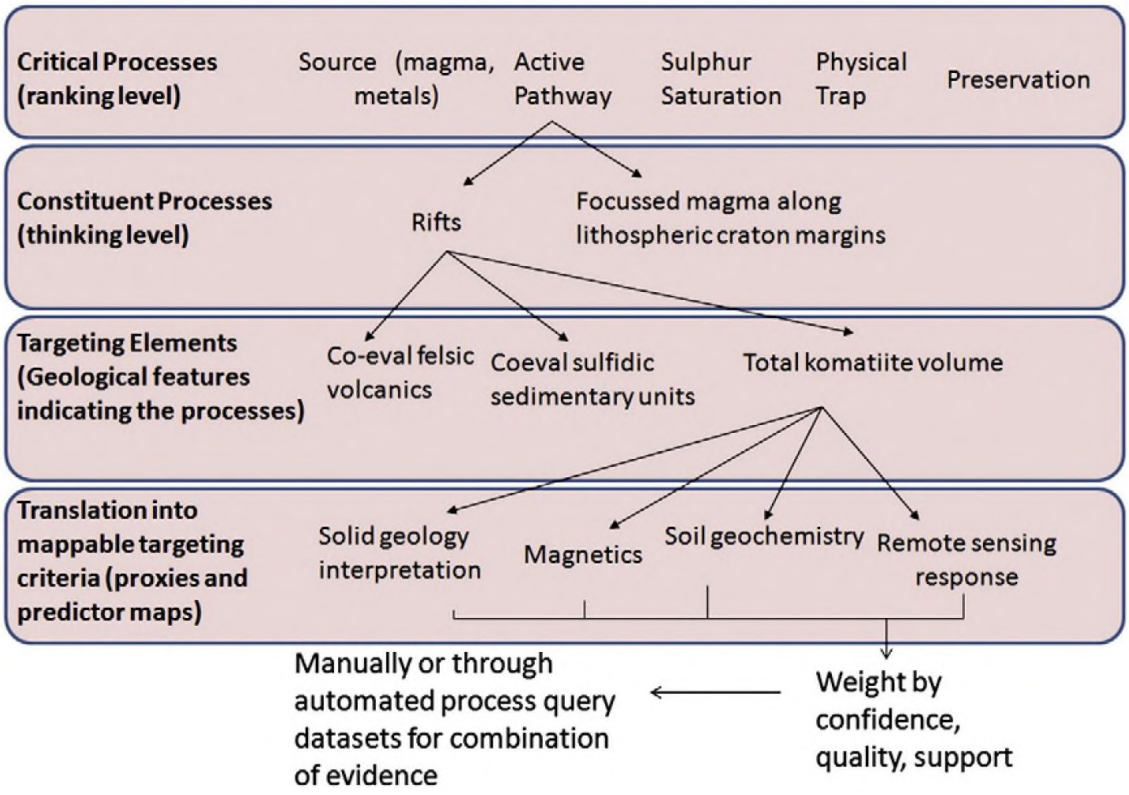


Figura B- 2 – Principais vertente de um sistema mineral do tipo Ni-Cu em komatiitos proposto por McCuaig et al., (2010). Observe o encadeamento dos processos geológicos desde a escala crustal até o mapeamento dos alvos potenciais em escala de depósito.

A fertilidade do terreno é definida como a tendência de uma região ou de uma época geológica de ser mais favorável para a formação de depósitos minerais do que outras (McCuaig & Hronsky, 2014). A fertilidade varia de acordo com a evolução da crosta em diversos momentos tais como processos de rifteamento e colisão subsequente, formação de supercontinentes, dentre outros.

A arquitetura litosférica tem relação com os padrões estruturais associados com as mineralizações, como tendências estruturais ou *oreshoots*, assim como com as estruturas de dimensão crustais, que cortam do embasamento às sequências de topo, e por muitas vezes são utilizadas como condutos pelos fluidos mineralizadores em seu processo de migração. Este critério é muito importante em processos hidrotermais, porém é também relevante na formação de depósitos minerais associados a magmatismo, como em pórfiros, *greisens* ou outros depósitos associados a intrusões (McCuaig & Hronsky, 2014).

Com a evolução das técnicas de datação e a difusão destes métodos, foi possível perceber que os grandes depósitos minerais ocorrem em momentos muito restritos da história da terra (McCuaig & Hronsky, 2014). Desta forma, a condição geodinâmica pode ser notada em depósitos distantes entre si por centenas de quilômetros, mas formados em uma mesma

época geológica e com condições similares. Os autores destacam três ambientes geodinâmicos favoráveis para a formação de grandes depósitos:

1. estágios iniciais de eventos extensionais, com ascensão de magma mantélico e/ou instalação de uma pluma sob a litosfera;
2. compressão intermitente, importante na formação de depósitos tipo pórfiro;
3. variações na direção da deformação que resultem em um campo de tensões neutro, com ausência de estruturas geradoras de permeabilidade secundária.

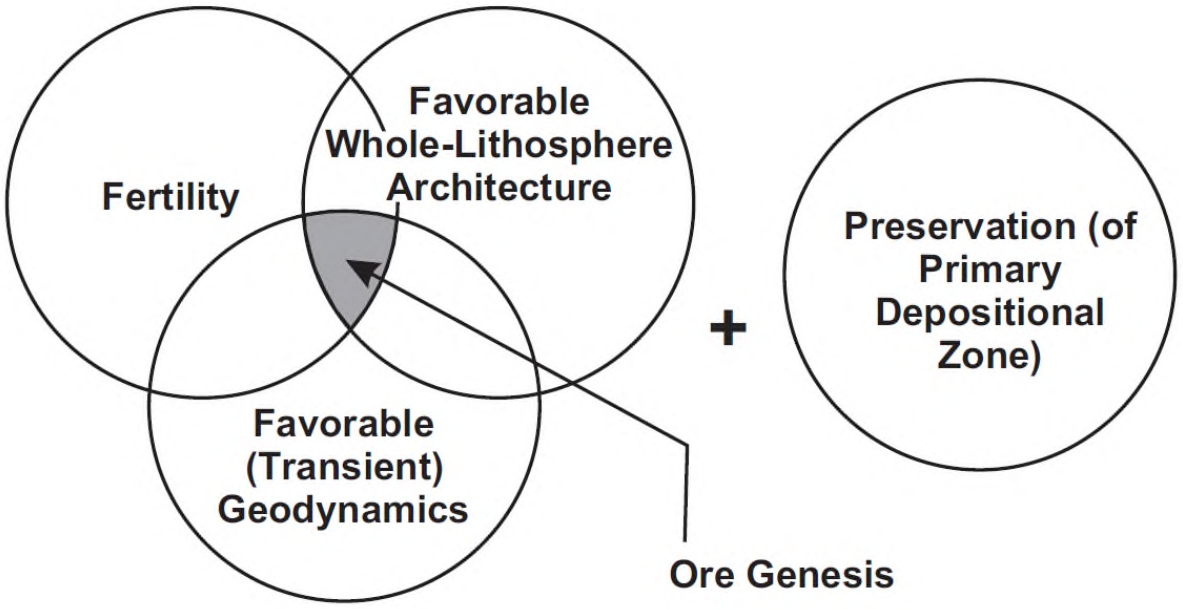


Figura B- 3 – Fatores Críticos para a formação de um depósito mineral (McCuaig & Hronsky, 2014).

Modelagem de Potencial Mineral

O desenvolvimento de uma abordagem reprodutível para identificar locais com alto potencial para exploração de uma *commodity* mineral é o objetivo central dos estudos da prospectividade mineral (Joly et al., 2012). Para tanto, se faz necessário um grande conjunto de informações consistentes em uma perspectiva multiparamétrica, aplicável à escala de interesse. Assim, os modelos de prospecção são tentativas de emular os processos formadores e dispersores de mineralizações, a fim de detectar novos alvos baseados em critérios do sistema mineral estabelecido.

O modelo de prospectividade pode ser construído com base em informações pré-definidas através de modelos orientados pelo conhecimento (ou *knowledge-driven model*) ou com base na assinatura de um depósito mineral conhecido por modelos orientados pelos dados (ou *data-driven model*). Desta forma, pode-se desenvolver uma prospecção mineral através de uma concepção regional de modelo geológico para a mineralização ou através da similaridade

de condições com um depósito de minério conhecido previamente (Carranza, 2009; Joly et al., 2012).

Os modelos orientados pelo conhecimento prévio são muitas vezes adequados para áreas onde o conhecimento geológico é limitado (*greenfield*) ou as ferramentas disponíveis são escassas. Os intérpretes se baseiam em modelos clássicos de mineralizações, e tentam estimar o tipo de resposta esperada em cada ferramenta utilizada. Se a conjunção de respostas for como esperada, gera-se um alvo a ser verificado, caso contrário, troca-se o modelo de mineralização ou a área é descartada. Estes modelos têm a vantagem de serem mais versáteis, porém são altamente tendenciosos e dependem exclusivamente da experiência da equipe de prospecção e da similaridade entre os depósitos (Agterberg & Bonham-Carter, 2005; Bonham-Carter, 1994; Carranza, 2009).

Já os modelos orientados por dados, ou também modelos empíricos, são mais complexos, sendo construídos numericamente pelo conjunto dados fornecidos e são usados para prever, não para explicar as mineralizações. Um modelo empírico simula uma função matemática que captura a tendência dos dados. Essa tendência, ou padrão, pode ser inferida pela análise e interpretação direta, ou com a ajuda de lógica computacional pré-compilada (como Algoritmos de Aprendizado de Máquina ou *Machine Learning Algorithms* – MLA), que podem dar uma resposta não tendenciosa do problema envolvido (Agterberg & Bonham-Carter, 2005; Bonham-Carter, 1994; Carranza, 2009).

Para maior eficácia, os fatores críticos de formação da mineralização deverão corresponder a pelo menos um guia prospectivo, representados como vetores em um espaço N-dimensional, onde N é o número total de variáveis envolvidas no modelo desenvolvido. Este método seria mais adequado em províncias e distritos minerais razoavelmente maduros, onde o volume de informação gerada e sistematizada poderia satisfazer estes pré-requisitos. Porém, devido ao grande volume de dados, é esperado algum problema para a integração e processamento de toda a informação gerada, uma vez que os dados podem ser coletados com metodologias, momentos e escalas distintos. Assim, integração eficaz de dados passaria a ser um novo desafio.

Para Hagemann et al. (2016), a vantagem da abordagem de sistema mineral em relação a descrição taxonômica de depósitos. Ela enfocaria os processos geológicos críticos necessários para formar grandes mineralizações, além de incluir todos os elementos descritivos de um estilo específico de mineralização. Sistemas minerais podem explicar a coexistência espacial e temporal de depósitos minerais dentro de uma província mineral específica. Também pode

explicar famílias de sistemas minerais coevos que potencialmente se formaram nos mesmos terrenos ou províncias adjacentes. Apesar disto, estes autores relacionam uma série de desafios a serem superados para que esta linha de pesquisa.

Apesar de uma aparente atração de bases científicas sólidas e consistência metodológica interna, traduzir o entendimento teórico de sistemas minerais em modelos efetivos de prospectividade mineral e alvos específicos de exploração ainda é um grande desafio em vários níveis conceituais (Hagemann et al., 2016). A dificuldade mais séria (além da compreensão inevitavelmente incompleta e evolutiva dos sistemas minerais) permanece definindo critérios mapeáveis que representam adequadamente vários elementos do sistema mineral críticos para formar, expor e preservar províncias minerais férteis e zonas dotadas e campos minerais dentro deles. Isso é particularmente problemático porque as indicações diretas desses elementos críticos e correspondentes processos geológicos de grande escala, como uma fonte de metal fértil e uma arquitetura de subsolo pré-existente favorável, geralmente não são observáveis ou têm apenas expressões muito sutis em um nível crustal de formação de depósitos minerais.

Vários desenvolvimentos recentes nas geociências podem ajudar a melhorar a eficácia da segmentação de exploração preditiva em escalas regionais (província a distrito), embora algumas delas ainda devem ser totalmente apreciados e adequadamente abordados na prática comum de direcionamento de exploração. Esses desenvolvimentos incluem o reconhecimento de: (i) dependência de escala de elementos e processos do sistema mineral; (ii) composição da arquitetura tectônica como os principais fatores que controlam a fertilidade e a operação do sistema mineral na província para a escala de distrito metalogênico; (iii) limites de domínios crustais profundos e outras estruturas pré-existentes no subsolo com alguma expressão em superfície como controle fundamental do sistema mineral na província e nas imediações dos depósitos; (iv) efeitos da incerteza na tomada de decisão sobre o objetivo da exploração (Hagemann et al., 2016).

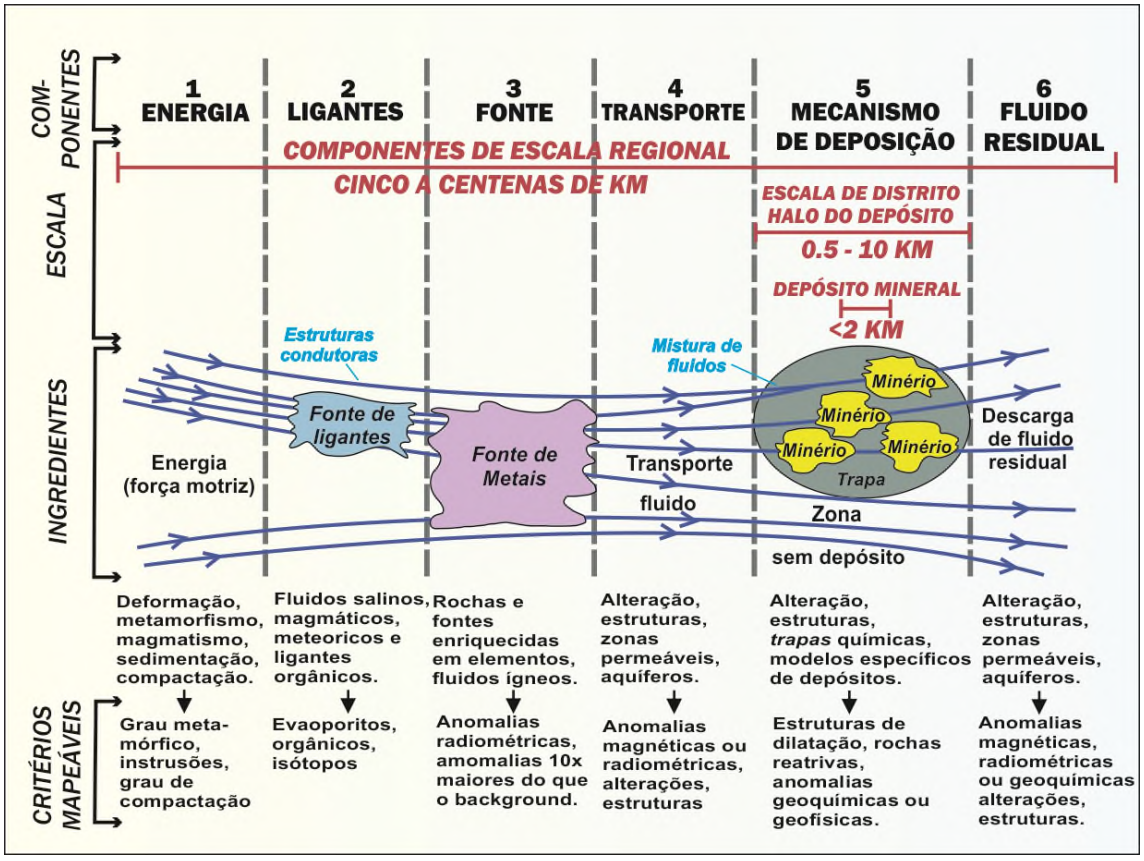


Figura B- 4 – Modelo genérico de Sistemas Minerais, com indicação de processos, escala de observação e possíveis critérios mapeáveis em levantamentos geológicos, geofísicos ou geoquímicos de exploração (basedo em Hagemann et al., 2016).

Em seu trabalho mais recente Hronsky & Kreuzer (2019) alegam que apesar de muitas décadas de desenvolvimento, a modelagem de prospectividade ainda não é amplamente utilizada ou aceita mundialmente em toda a indústria de exploração mineral, à exceção de equipes de empresas como DeBeers, Newmont e Kenex (Hronsky & Kreuzer, 2019). Uma crítica comum ao método é que ele não é praticamente útil porque tem sido utilizado para amadurecer regiões já conhecidas e muitas vezes gera áreas excessivamente grandes de alta prospectividade. Estes autores sugerem que a razão para isso não esteja primariamente relacionada a limitações nos algoritmos de mapeamento de prospectividade, mas a questões relativas ao uso de conjuntos de dados de entrada.

De acordo com Hronsky & Kreuzer (2019), críticas comuns das tentativas de modelagem prospectiva no setor de exploração relacionam-se a duas questões principais. Em primeiro lugar, a técnica é muito mais eficaz em encontrar os depósitos já conhecidos do que gerar novos alvos válidos. Isso obviamente se relaciona, pelo menos em parte, ao fato de que, na maioria das metodologias de modelagem de prospectividade comumente adotadas, os locais de depósito conhecidos são uma entrada de modelo chave. As exceções são sobreposição de

lógica difusa, sobreposição ponderada e modelos de “Sistemas de Inferência Fuzzy”, que não exigem locais de depósito como entrada (Bonham-Carter, 1994; Carranza, 2009). A segunda grande crítica é que, além de áreas adjacentes a depósitos conhecidos (e, portanto, óbvias), o próximo nível de domínios de alta prospecção gerados é tipicamente grande, relativo à área de interesse. Esta é uma questão muito importante porque, para ser praticamente útil na exploração mineral, qualquer técnica de direcionamento deve produzir pelo menos uma redução de ordem de grandeza na área de foco (Hronsky & Kreuzer, 2019) .

Finalmente, os autores apresentam algumas sugestões a serem tomadas a fim de avançar esta linha de pesquisa e torna-la mais adequada a realidade da indústria (Hronsky & Kreuzer, 2019):

• Desenvolvimento de novos fluxos de trabalho e metodologias de modelagem de prospectividade efetivos que estejam fortemente alinhados com a prática no mundo real da exploração mineral. Isso não deve ocorrer se esse desenvolvimento for deixado somente para os geoestatísticos. Ao invés disso, deve-se encorajar a formação de equipes colaborativas multidisciplinares com experiência em modelagem prospetiva e exploração mineral.

• Aplicação de uma abordagem “híbrida”, focada na amplificação da inteligência e não na inteligência artificial (IA), que aproveita poder combinado da mente humana para reconhecer, mapear e extrapolar padrões com o rigor de algoritmos baseados em máquinas. A chave para essa abordagem é a compilação dos principais mapas geológicos interpretativos que se situam intermediários entre os dados de entrada primários e os algoritmos de mapeamento de prospectividade, como os MLA. Em seguida, estes dados seriam usados como as principais entradas no processo de modelagem de prospectividade, adicionalmente a quaisquer conjuntos de dados disponíveis sistematicamente amostrados.

• Os Serviços Geológicos devem fornecer juntamente com as demais camadas de dados pré-competitivas as suas melhores interpretações de estruturas em larga escala que não são facilmente observáveis na geologia de superfície. O uso de dados gravimétricos regionais de alta resolução podem ser uma ferramenta muito importante para auxiliar nestas interpretações.

Ainda segundo Hronsky & Kreuzer (2019), embora os problemas discutidos limitem fortemente a aplicação da técnica de modelagem de prospectiva como atualmente praticada pela maioria, eles não são barreiras para a implementação bem-sucedida dessa tecnologia no futuro. Sugere-se que o método mais eficaz possa ser um híbrido de interpretação geológica humana subjetiva e análise objetiva com suporte em algoritmos, que capture os melhores aspectos dessas abordagens alternativas. Isso exigiria uma boa integração dos dados geológicos, geofísicos e

geoquímicos básicos disponíveis em camadas interpretativas que forneçam corretamente as entradas primárias para a análise de modelagem de prospectividade.

Aprendizagem de máquina para integração de dados

Os modelos orientados por dados são inferências matemáticas baseado em medidas de parâmetros quantitativos e na relação espacial destes resultados com os depósitos conhecidos (Agterberg & Bonham-Carter, 2005; Bonham-Carter, 1994; Carranza, 2009; Rodriguez-Galiano et al., 2015). Historicamente, os modelos "*data-driven"* desenvolvidos são baseados em regressões probabilísticas ou em lógica Bayesiana, o que resulta em inferências estatísticas complexas, por vezes operadas em ferramentas complexas em ambiente SIG, o que inibe o seu uso. Com o advento da popularização de técnicas de inteligência artificial para gerenciamento de dados, como os algoritmos de aprendizagem de máquina, o processamento de dados para modelagem *data driven* tem se tornado mais simples e rápido.

Algoritmos de aprendizagem de máquina (MLA na sigla em inglês, *Machine Learning Algorithms*) como redes neurais artificiais (*Artificial Neural Networks* – ANN), árvores de regressão (*Regresion Tree* – RT), florestas aleatórias (*Random Forests* – RF) e máquinas de vetores de suporte (*Support Vector Machine* – SVM) são métodos *data-driven* eficientes que podem ser usados para identificar padrões em um conjunto de dados, visando modelagem de prospectividade mineral (Abedi et al., 2012; Bérubé et al., 2018; Breiman, 2001; Breiman et al., 1995; Carranza & Laborte, 2016, 2015a; Costa et al., 2019; Rodriguez-Galiano et al., 2015; Zuo, 2017; Zuo et al., 2015). Via de regra, os métodos de MLA para classificação automática necessitam reconhecer parte dos dados como forma de "calibrar" (ou treinar) as predições.

Rodriguez-Galiano et al. (2015) comparou a performance de modelos gerados por ANN, RT, RG e SVM no Distrito Minero de Rodalquilar, no sul da Espanha, uma área com ocorrências de ouro epitermal. Os dados de entrada para elaboração dos modelos são baseados em dados geológicos, geoquímicos, geofísicos e de sensoriamento remoto hiperespectral publicados em estudos anteriores (Rodriguez-Galiano et al., 2015).

A análise comparativa dos métodos MLA para modelagem mineral prospectividade foi realizada a partir de diferentes perspectivas: facilidade de aplicação e eficácia, sensibilidade à configuração dos parâmetros do modelo e redução de dados, precisão do mapeamento das classificações, e transparência e coerência dos modelos.

Rodriguez-Galiano et al. (2015) avaliam que os modelos apresentam uma dificuldade diferente em sua formulação (ou treinamento). Algoritmos baseados em árvore de decisão (RT e RF) envolvem são mais facilmente treinados do que os demais. Os autores afirmam ainda que

o desempenho dos métodos é extremamente dependente do conjunto de dados utilizados para a calibração inicial, sendo que o método que obteve melhor desempenho com um conjunto de dados de calibração mínimo foi o RF, sendo, portanto, o método mais indicado para uso em áreas onde há pouca informação geológica prévia.

APÊNDICE C – Código utilizado no Artigo 01

NOTEBOOK

This is the source code used in the manuscript:

"Predicting mineralization and targeting exploration criteria based on machine-learning in the Serra de Jacobina quartz-pebble-metaconglomerate Au-(U) deposits, São Francisco Craton, Brazil"

authored by: "Guilherme Ferreira" date: "02/03/2022"

The following code was written in R (3.5.6)

ABSTRACT

Defining mineral exploration criteria is a laborious, time-consuming, and generally an empirical task often biased and limited to expert knowledge. To address this problem with a different approach, we used data-driven analysis to make predictions and provide insights about gold mineralization in rocks of the Jacobina Group, São Francisco Craton. The input variables were petrophysical parameters (density, magnetic susceptibility, and electric conductivity) and lithogeochemistry data obtained by X-Ray Fluorescence assays. A machine learning model based on the Random Forests algorithm was applied to predict mineralization in drill core samples. The database used for algorithm training was balanced using the Borderline-SMOTE technique to provide approximately the same numbers of samples of the two classes in the mineral status parameter (i.e., ore and barren samples). The quality of the predictions was assessed with different datasets (i.e., training, testing, each drill core separately, and all samples) and by parameters. The average accuracies were 0.87 for cross-validation training, 0.91 for testing, and 0.86 for all samples. Also, the model allowed us to estimate and rank the importance of the input variables to the prediction. These estimates were validated by an interpretation of optical and scanning electron microscopy petrographic analysis, which was carried out to understand the relationship between minerals of different stages and gold mineralization. Thus, the techniques used in this work could help to decrease the time spent in data integration and interpretation, as mineral exploration teams can easily replicate this approach.

DATA WRANGLING

Dependencies

```
library(tidyverse) # ggplot2, tidyr, dplyr
library(readxl) # open XLSX data
library(geoquimica) # Data wrangling
library(caret) # Machine Learning
library(doParallel) # Parallel Processing
library(randomForest) # RF
library(randomForestExplainer) # RF
library(pROC) # ROC and AUC
library(smotefamily) # Smote
```

Data preparation

```
set.seed(0)

setwd('~/GitHub/jacobina/data/pphy')


xrf <- read_xlsx(path = '~/GitHub/jacobina/data/xrf/pXRF_Jacobina_SAMPLES.xlsx',
                 sheet = 1)

# Petrophysics data ----
files <- list.files(pattern = '.xlsx$',path = '~/GitHub/jacobina/data/pphy')

phy <- lapply(files, read_xlsx, sheet = 1) %>%
  bind_rows() %>%
  mutate(HOLE = as.factor(HOLE),
         ID = as.factor(ID),
         FROM = as.numeric(FROM),
         TO = as.numeric(TO),
         LITHO = as.factor(LITHO),
         MINERALIZATION = as.factor(MINERALIZATION),
         SUSCEPTIBILITY = as.numeric(SUSCEPTIBILITY),
         CONDUCTIVITY = as.numeric(CONDUCTIVITY),
         DENSITY = as.numeric(DENSITY),
         COMMENTS = as.character(COMMENTS)) %>%
  arrange(ID)

phy <- phy %>%
  mutate(SAMPLE = paste(HOLE,formatC(x = phy$FROM,flag = '0',
                                     width = 6,
                                     digits = 2,
                                     format = 'f'),sep = '-'),
         ROCK = case_when(phy$LITHO %in% c('GRIT','LMPC','LVLPC','MLPC','MPC','MSPC',
                                           'SMPC','SPC','VSPC') ~ 'CONGLOMERATE',
                          phy$LITHO %in% c('QTO','QTO_SX','QZ_VEIN') ~ 'QUARTZITE',
                          phy$LITHO %in% c('ITV','UMF') ~ 'ULTRAMAFIC',
                          phy$LITHO %in% c('BRX') ~ 'BRECCIA',
                          phy$LITHO %in% c('XISTO') ~ 'SCHIST',
                          phy$LITHO %in% c('SOLO') ~ 'SOIL',
                          TRUE ~ as.character(phy$LITHO)))
# Merging dataset ----

df <- phy %>%
  left_join(xrf,by = 'SAMPLE') %>%
  mutate(min = factor(ifelse(test = phy$MINERALIZATION == 1 | phy$MINERALIZATION == 1000,
                             yes = 'ORE', no =  'BARREN')),
         Fe_Ti = Fe/Ti)
```

SMOTE

```r
set.seed(0)

conglomerate <- as.data.frame(df) %>%
  filter(ROCK == 'CONGLOMERATE') %>%
  dplyr::select(min,7:9, Cu, Fe, Cr, Ti, K, Al, Si, S, Fe_Ti) %>%
  na.omit()

# Split data to smote
index <- caret::createDataPartition(conglomerate$min,
                                    p =1,
                                    list = FALSE,
                                    times = 1)

toSmote <- conglomerate[index,]

fromSmote <- BLSMOTE(C = 5,dupSize = 0,
                     X = as.data.frame(toSmote[,-1]),
                     K = 5,
                     target = as.data.frame(toSmote['min']))

## [1] "Borderline-SMOTE done"

fromSmote$data %>%
  rename(min = class) %>%
  select(min,1:3,Cr, K, S, Fe_Ti) %>%
  elem_norm(method = 'clr') %>%
  GGally::ggpairs(mapping=ggplot2::aes(colour = min), progress = FALSE)

df_smote <- fromSmote$data %>%
  rename(min = class) %>%
  mutate(min = as.factor(min))
```

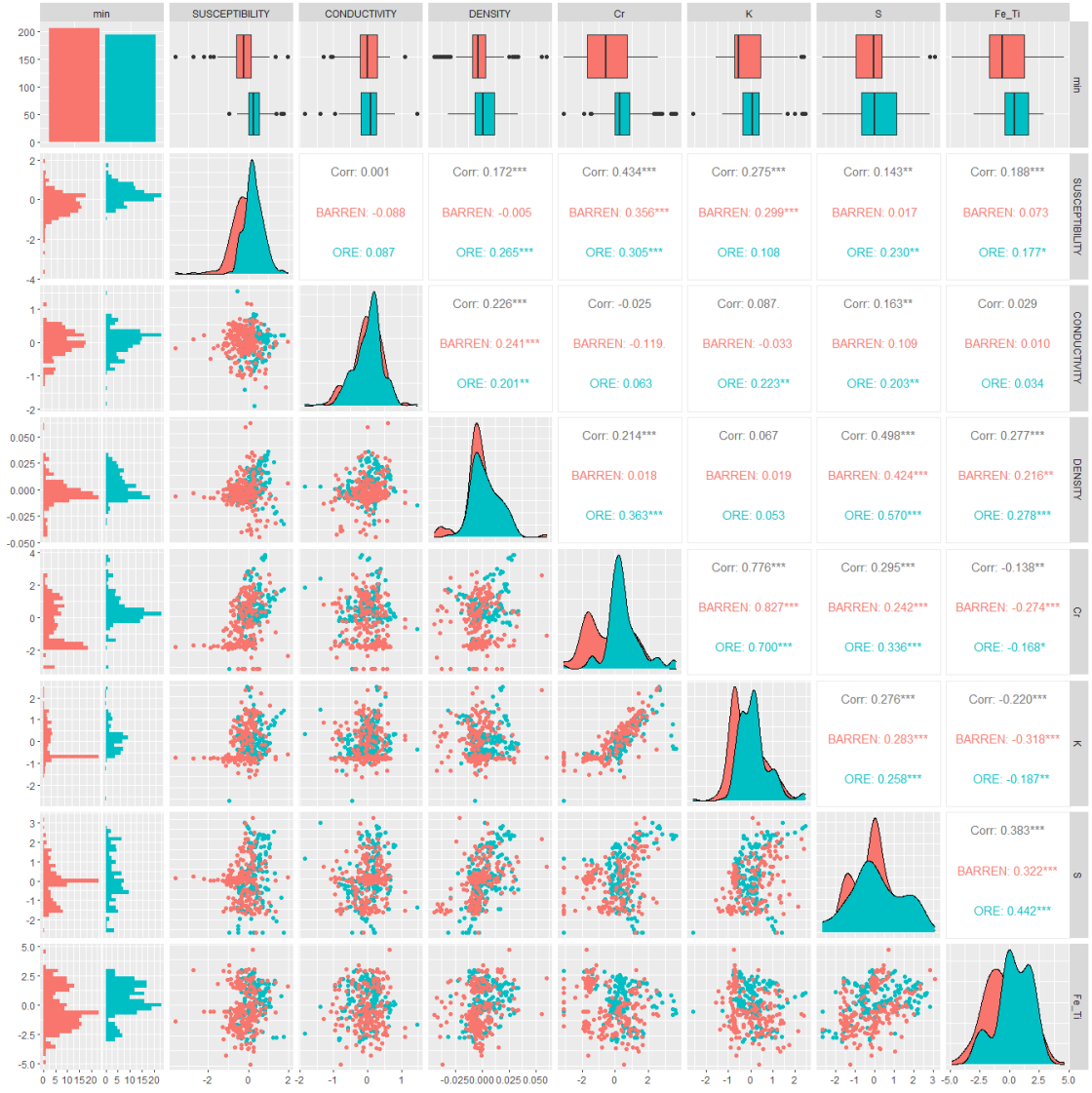


Fig. C. 1: Exploratory data analysis for selected variables after the BLSMOTE balancing. Data are color coded according to the Mineralization Status (i.e., Ore or Barren). The asterisk indicates the level of significance of the correlations. * for alpha = 0.15, ** for alpha = 0.05, and *** for alpha = 0.01.

RANDOM FORESTS

```
# Imbalanced Model

splitIndex <- caret::createDataPartition(conglomerate$min,
                                         p =.7,
                                         list = FALSE,
                                         times = 1)

trainSPlit <- conglomerate[splitIndex,]
testSplit <- conglomerate[-splitIndex,]


imbal_minModel <- randomForest(min ~ .,
                               trainSPlit,
                               proximity = TRUE,
                               ntree = 1000,
                               localImp = TRUE)
```

```
par(pty = 's')
roc1 <- plot.roc(trainSPlit$min, imbal_minModel$votes[,1],legacy.axes=TRUE,
         percent = FALSE, print.auc = TRUE,xlab="TRUE NEGATIVE",
         ylab = 'TRUE POSITIVE')
par(pty = 'm')
```

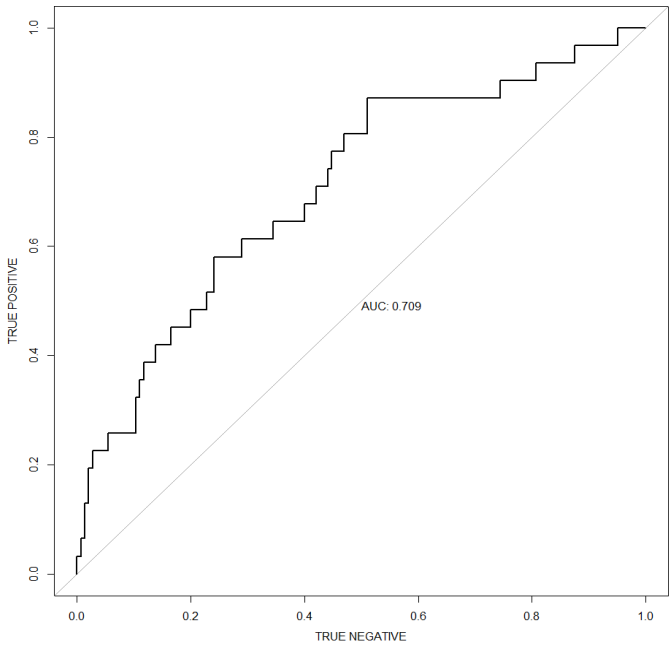


Fig. C. 2: AUC fot imbalanced dataset

```
# Balanced Model

splitIndex <- caret::createDataPartition(df_smote$min, p =.7,
                                         list = FALSE,
                                         times = 1)

trainSPlit <- df_smote[splitIndex,]
testSplit <- df_smote[-splitIndex,]

round(prop.table(table(trainSPlit$min)),digits = 3)

##
## BARREN    ORE
##  0.516  0.484

round(prop.table(table(testSplit$min)),digits = 3)

##
## BARREN    ORE
##  0.517  0.483

doParallel::registerDoParallel(cores = 3)
ctrl <- caret::trainControl(method = 'repeatedcv',
                            number = 5,
                            repeats = 3,search = 'grid',
                            allowParallel = TRUE,
                            verboseIter = TRUE)

minMod_fram <- min_depth_distribution(minModel)

print(minModel)
```

```
##
## Call:
##  randomForest(formula = min ~ ., data = trainSPlit, proximity = TRUE,      ntree = 1000,
localImp = TRUE, mtry = tunning$bestTune[[1,          1]])
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 3
##
##         OOB estimate of error rate: 14.59%
## Confusion matrix:
##        BARREN ORE class.error
## BARREN    126  19   0.1310345
## ORE        22 114   0.1617647

plot(minModel)
```

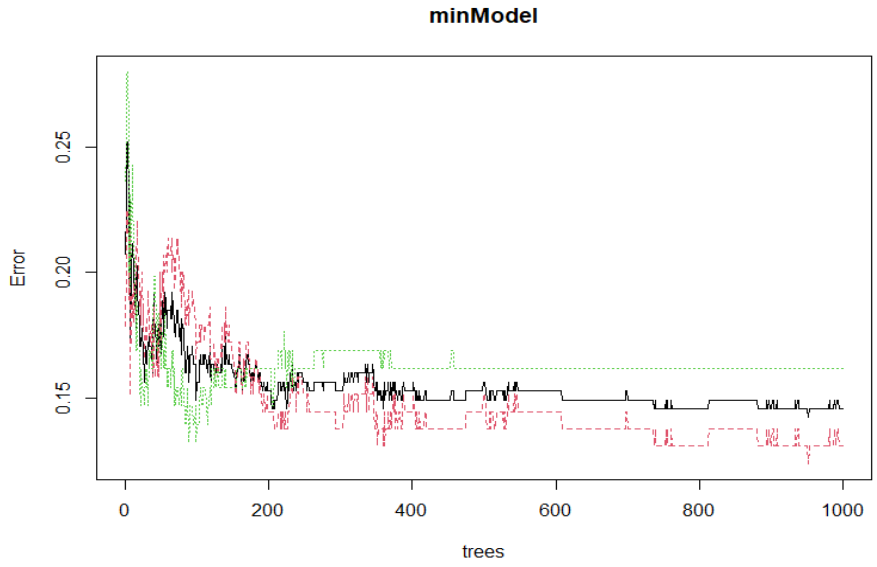


Fig. C. 3: Eror rate (1 – Accuracy) and number of estimators (tress) for the Random Forests models

```
par(pty = 's')
plot.roc(trainSPlit$min, minModel$votes[,1],legacy.axes=TRUE,
         percent = FALSE, print.auc = TRUE,xlab="TRUE NEGATIVE",
         ylab = 'TRUE POSITIVE')
par(pty = 'm')
```

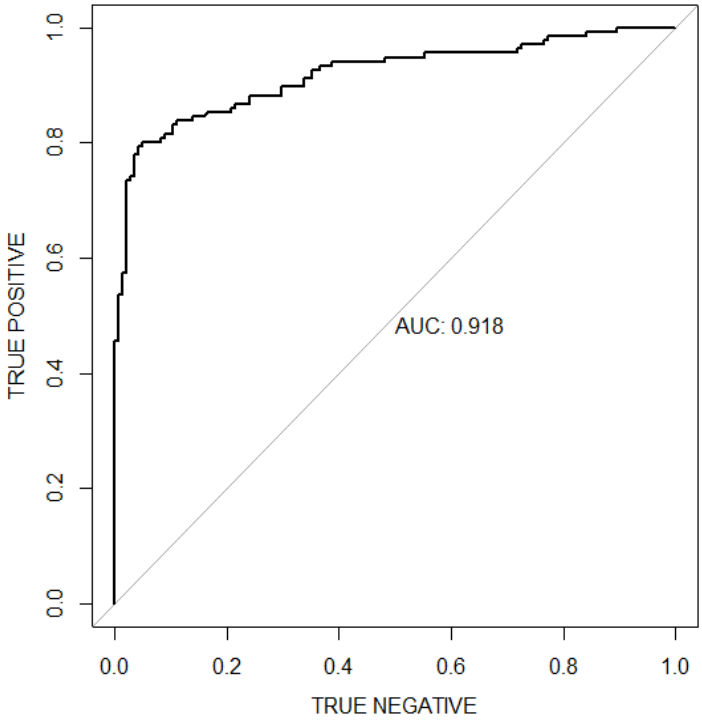


Fig. C. 4: AUC for balanced model

```
plot_min_depth_distribution(minModel)
```

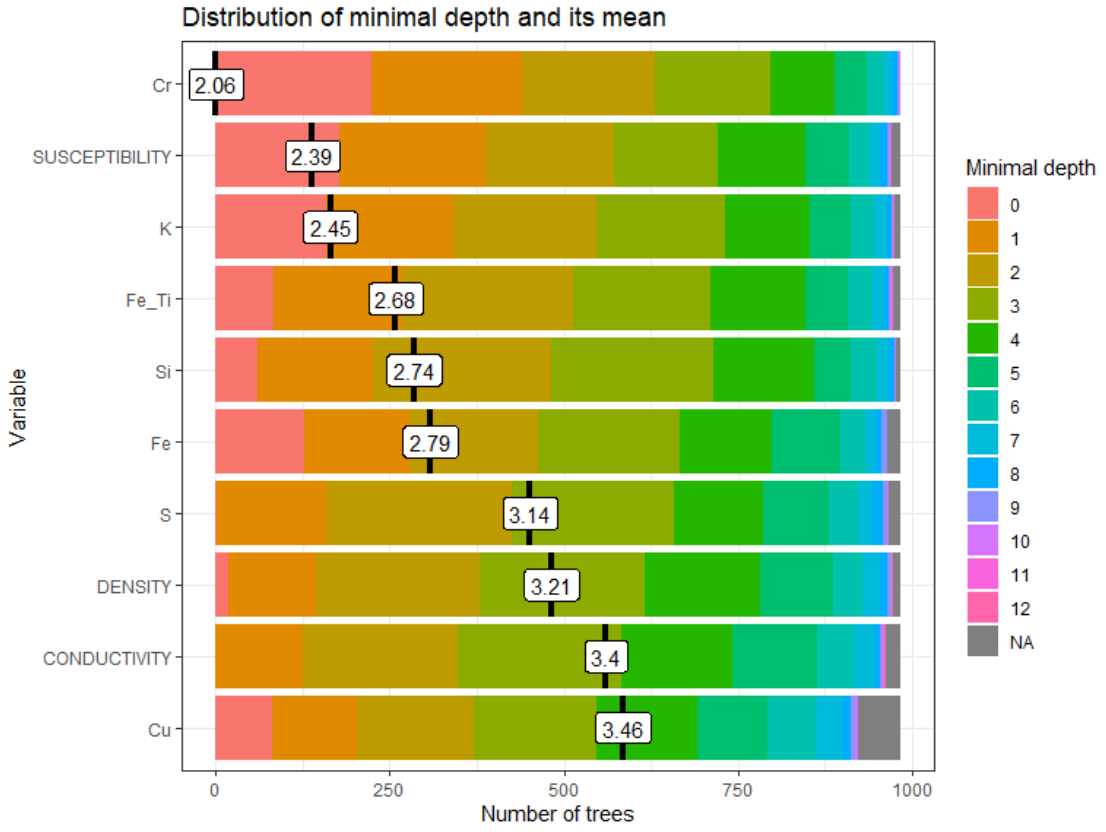


Fig. C. 5: Variable importance ranked according to the average minimal depth and number of trees

```
plot_predict_interaction(forest = minModel,
                         data = trainSPlit,
                         variable1 =  "SUSCEPTIBILITY",
                         variable2 =  "DENSITY",
                         grid = 500)
```

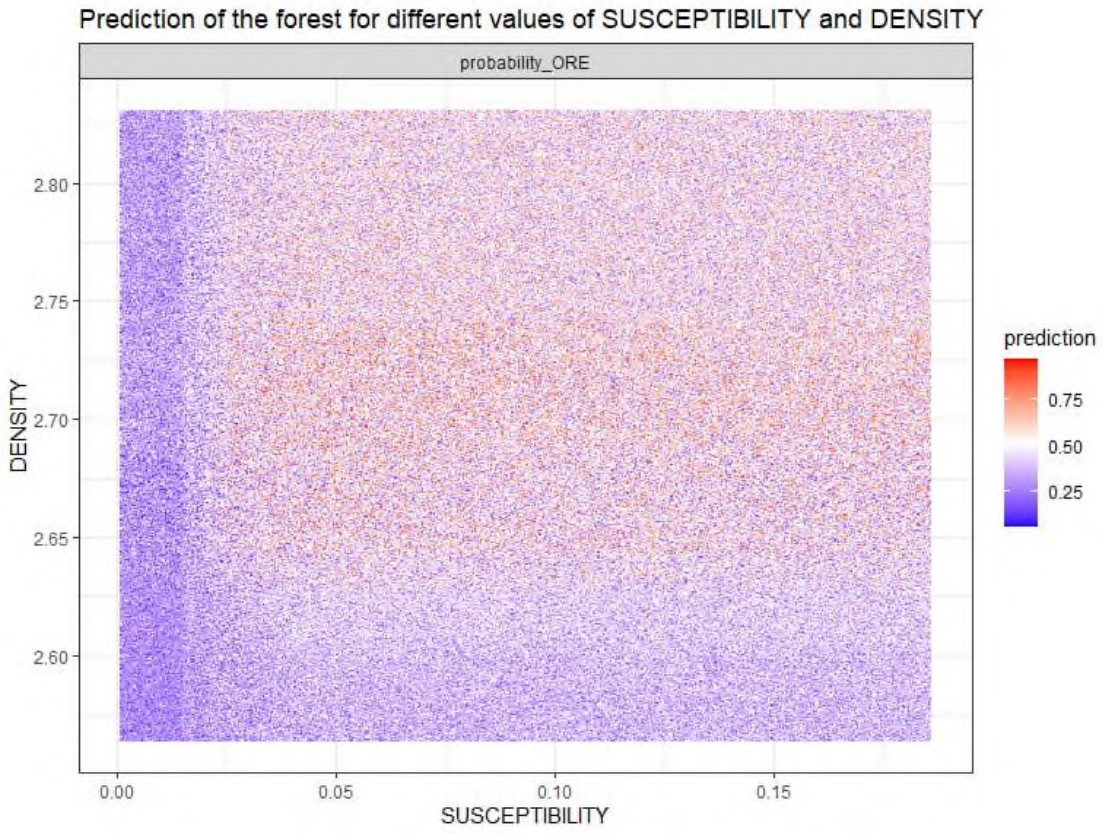


Fig. C. 6: prediction grid of the probability for been classified as ORE, according to Susceptibility and Density values

```
confusionMatrix(as.factor(trainSPlit$min), as.factor(minModel$predicted),positive = 'ORE')

## Confusion Matrix and Statistics
##
##           Reference
## Prediction BARREN ORE
##     BARREN    126  19
##     ORE        22 114
##
##                Accuracy : 0.8541
##                  95% CI : (0.8073, 0.8932)
##     No Information Rate : 0.5267
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7077
##
##  Mcnemar's Test P-Value : 0.7548
##
##             Sensitivity : 0.8571
##             Specificity : 0.8514
##          Pos Pred Value : 0.8382
##          Neg Pred Value : 0.8690
##              Prevalence : 0.4733
##          Detection Rate : 0.4057
##    Detection Prevalence : 0.4840
##       Balanced Accuracy : 0.8542
##
##        'Positive' Class : ORE
##

varImp.df <- as_tibble(varImp(minModel,
                              sort = TRUE,
                              scale = FALSE),
                       rownames = 'Variables')

varImp.df %>%
  ggplot(aes(y = reorder(Variables, ORE),
             x = round((100*ORE/sum(varImp.df$ORE)),2),
             fill = round((100*ORE/sum(varImp.df$ORE)),2))) +
  geom_col(col = 'black') +
  scale_fill_viridis_c() +
  geom_text(label = round((100*varImp.df$ORE/sum(varImp.df$ORE)),2),
            nudge_x = -1, col = 'white',aes(fontface = c('bold'))
            ) +
  theme(legend.position = c(.92,.3)) +
  labs(y = "Variables", x = "Mean Decrease Accuracy (%)", fill = 'Accuracy')
```

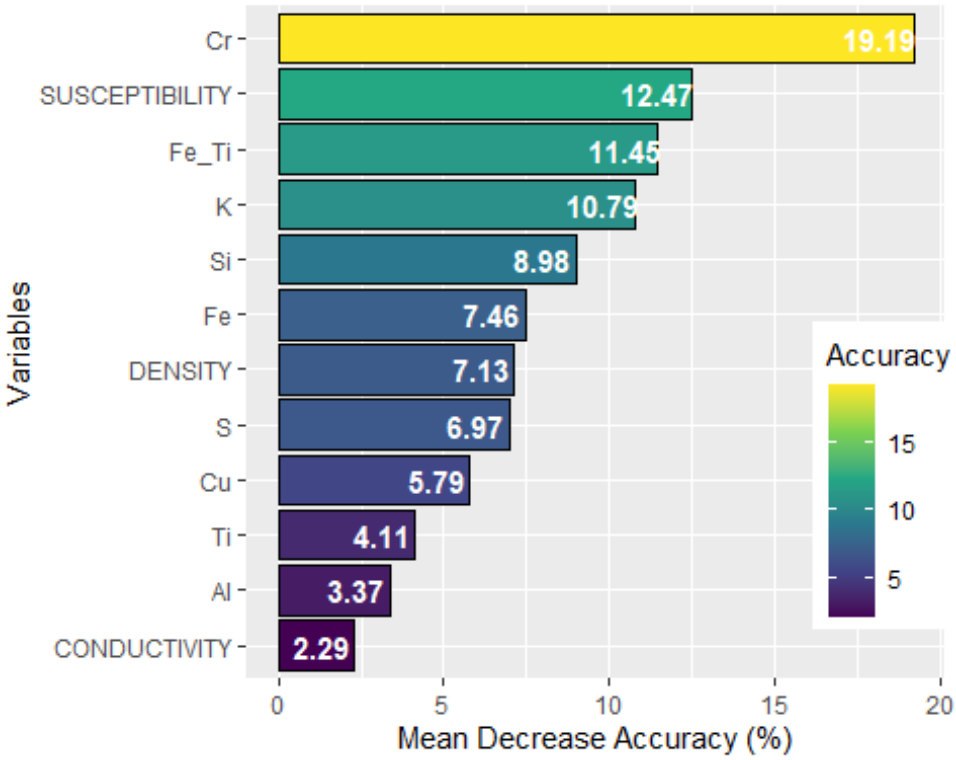


Fig. C. 7: Variable importance rank based on the Mean Decrease Accuracy parameter normalized to a percentage distribution. The bars are filled with the coded colors according to their respective importance. Abbreviations: Mag.Susceptibility – Magnetic Susceptibility; Electri. Conductivity – Electric Conductivity.

```r
x <- df %>%
  filter(!is.na(Cr))

y <- df %>%
  filter(!is.na(Cr)) %>%
  select(min)

labs <- df %>%
  select(ID:LITHO)

pred1 <- predict(minModel,newdata =  x)

pred.prob <- predict(minModel,newdata =  x,type = 'prob',norm.votes = TRUE,predict.all = TR
UE)

prob.ORE <- pred.prob$aggregate[,2]

dc_samples <- x %>%
  mutate(prob.ORE = all_of(prob.ORE),
         Prediction = all_of(pred1))

## Start by converting the proximity matrix into a distance matrix.
distance.matrix <- as.dist(1-minModel$proximity)

mds.stuff <- cmdscale(distance.matrix,
                      eig=TRUE,
                      x.ret=TRUE)

## calculate the percentage of variation that each MDS axis accounts for...
mds.var.per <- round(mds.stuff$eig/sum(mds.stuff$eig)*100, 1)

## now make a fancy looking plot that shows the MDS axes and the variation:
mds.values <- mds.stuff$points
mds.data <- data.frame(Sample=rownames(mds.values),
                       X=mds.values[,1],
                       Y=mds.values[,2],
                       Status = trainSPlit$min,
                       prob = minModel$votes)

ggplot(data=mds.data, aes(x=X, y=Y, label=Sample, shape = Status)) +
  geom_point(aes(col = prob.ORE), size = 3) +
  theme_bw() +
  xlab(paste("MDS1 - ", mds.var.per[1], "%", sep="")) +
  ylab(paste("MDS2 - ", mds.var.per[2], "%", sep="")) +
  ggtitle("MDS plot using (1 - Random Forest Proximities)") +
  scale_color_viridis_c()
```

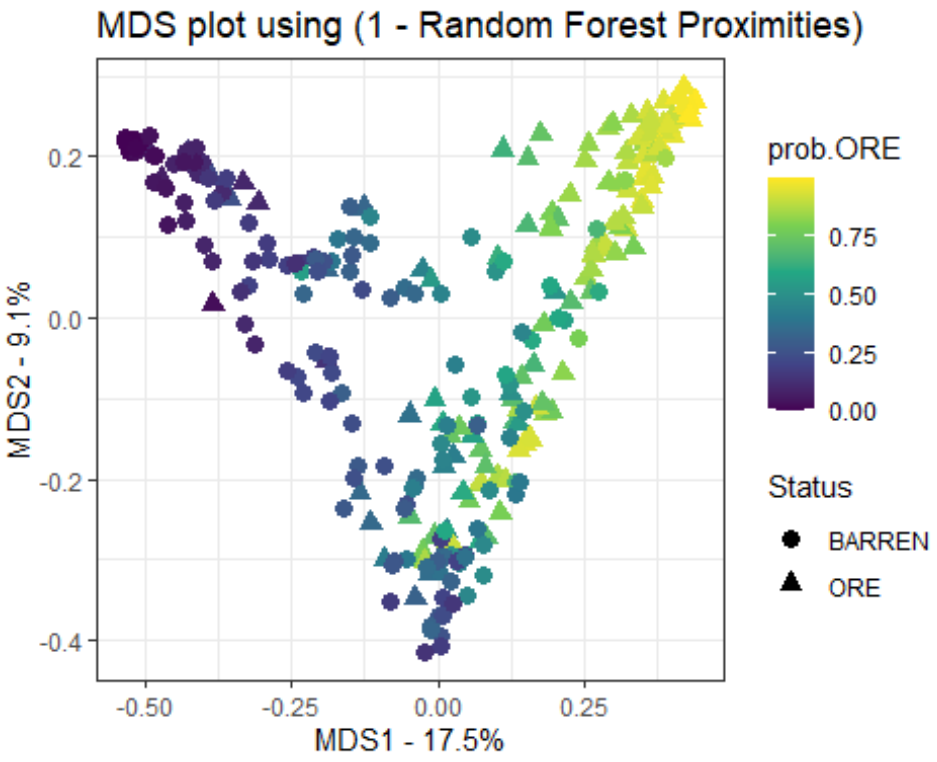


Fig. C. 8: Multidimensional Scaling graph based on all the Trees results. The points are classified according to the Mineralization Status (i.e., Ore or Barren) and color coded by the probability of been classified as ORE.

```
# Test prob

predTest <- predict(minModel,newdata =  testSplit)

predTest.prob <- predict(minModel,newdata =  testSplit,type = 'prob',norm.votes =
TRUE,predict.all = TRUE)

probTest.ORE <- predTest.prob$aggregate[,2]

test.prob <- testSplit %>%
  mutate(prob.ORE = all_of(probTest.ORE),
         Prediction = all_of(predTest))

test.prob %>%
  filter(!is.na(Fe_Ti)) %>%
  ggplot(aes(x = prob.ORE)) +
  geom_density(aes(fill = min), alpha = .3) +
  geom_rug() +
  geom_vline(xintercept = 0.5, lty = 3) +
  facet_wrap(min ~ .,ncol = 1,scales = 'free_y') +
  theme(legend.position = 'none') +
  labs(x = 'Probability: Ore')
```

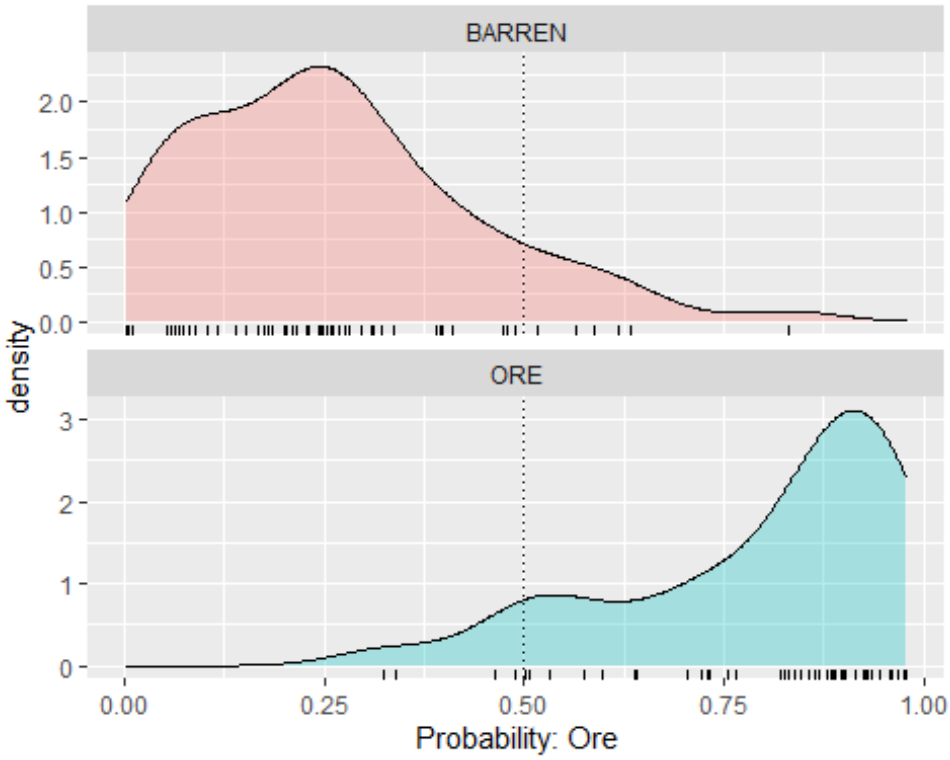


Fig. C. 9: Ore probability analysis for all samples based on the mineralization status of the test dataset (barren samples are represented in the red curve and ore samples in the blue curve). Most barren samples took a low Ore probability, and the mineralized samples got the highest probabilities. The fields of True Negative (TN, i.e., barren samples predicted as non-mineralized), False Positive (FP, barren samples predicted as mineralized), False Negative (FN, ore samples predicted as non-mineralized), and True Positive (TP, ore samples predicted as mineralized) are indicated in the plot. The ticks at the bottom of each plot indicate the calculated probability for each test dataset sample.

```
dc_samples <- dc_samples %>%
  filter(!is.na(Fe_Ti)) %>%
  mutate(number = 1) %>%
  group_by(HOLE) %>%
  mutate(fid = cumsum(number)) %>%
  ungroup()

h1 <-
  dc_samples %>%
  mutate(ROCK = as.factor(ROCK)) %>%
  filter(HOLE == 'CAN120') %>%
  arrange(HOLE, FROM) %>%
  ggplot(aes(y = fid, fill = ROCK, col = ROCK)) +
  geom_bar() +
  coord_cartesian(ylim = c(170,0)) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  labs(y = 'Position') +
  theme(
    legend.position = 'none',
        axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())


h2 <-
  dc_samples %>%
```

```
  mutate(ROCK = as.factor(ROCK)) %>%
  filter(HOLE == 'CAN144') %>%
  arrange(HOLE, FROM) %>%
  ggplot(aes(y = fid, fill = ROCK, col = ROCK)) +
  geom_bar() +
  coord_cartesian(ylim = c(170,0)) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  labs(y = 'Position') +
  theme(
    legend.position = 'none',
        axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())


h3 <-
  dc_samples %>%
  mutate(ROCK = as.factor(ROCK)) %>%
  filter(HOLE == 'CANIF27') %>%
  arrange(HOLE, FROM) %>%
  ggplot(aes(y = fid, fill = ROCK, col = ROCK)) +
  geom_bar() +
  coord_cartesian(ylim = c(170,0)) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  labs(y = 'Position') +
  theme(
    legend.position = 'none',
        axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())


h4 <-
  dc_samples %>%
  mutate(ROCK = as.factor(ROCK)) %>%
  filter(HOLE == 'JBA722') %>%
  arrange(HOLE, FROM) %>%
  ggplot(aes(y = fid, fill = ROCK, col = ROCK)) +
  geom_bar() +
  coord_cartesian(ylim = c(170,0)) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  labs(y = 'Position') +
  theme(
    legend.position = 'none',
        axis.text.x = element_blank(),
        axis.title.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())



prof1 <-
  dc_samples %>%
  filter(HOLE == 'CAN120') %>%
```

```
  filter(!is.na(Fe_Ti)) %>%
  mutate(Type = case_when(Prediction == min ~ 'True_PN',
                          Prediction != min ~ 'False_PN',
                          TRUE ~ 'ERROR!')) %>%
  ggplot(aes(y = fid)) +
  geom_vline(xintercept = 0.5, lty = 2, col = 'red') +
  geom_bar(aes(fill = Type, col = NULL),
           width = 1, alpha = .3,
           position = 'identity') +
  geom_segment(aes(y=fid, yend=fid, x=0, xend=prob.ORE)) +
  geom_point(aes(x = prob.ORE, col = min), cex = 2.5) +
  coord_cartesian(ylim = c(170,0), expand = TRUE) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  theme(legend.position = 'none',
        axis.text.x = element_text(angle = -90,
                                   hjust = .5,
                                   vjust = .5),
        axis.title.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        plot.margin = unit(c(.03,.03,.03,.03), "lines")
        ) +
  annotate(label = 'Threshold',
           x = 0.58, y = 160,
           fontface = 'italic',
           geom = 'text',
           angle = -90,
           colour = 'red',
           size = 2.5) +
  labs(y ='')

prof2 <-
  dc_samples %>%
  filter(HOLE == 'CAN144') %>%
  filter(!is.na(Fe_Ti)) %>%
  mutate(Type = case_when(Prediction == min ~ 'True_PN',
                          Prediction != min ~ 'False_PN',
                          TRUE ~ 'ERROR!')) %>%
  ggplot(aes(y = fid)) +
  geom_vline(xintercept = 0.5, lty = 2, col = 'red') +
  geom_bar(aes(fill = Type, col = NULL),
           width = 1, alpha = .3,
           position = 'identity') +
  geom_segment(aes(y=fid, yend=fid, x=0, xend=prob.ORE)) +
  geom_point(aes(x = prob.ORE, col = min), cex = 2.5) +
  coord_cartesian(ylim = c(170,0), expand = TRUE) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  theme(legend.position = 'none',
        axis.text.x = element_text(angle = -90,
                                   hjust = .5,
                                   vjust = .5),
        axis.title.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        plot.margin = unit(c(.03,.03,.03,.03), "lines")
        ) +
  annotate(label = 'Threshold',
           x = 0.58, y = 160,
           fontface = 'italic',
           geom = 'text',
           angle = -90,
           colour = 'red',
           size = 2.5) +
```

```
  labs(y ='')

prof3 <-
  dc_samples %>%
  filter(HOLE == 'CANIF27') %>%
  filter(!is.na(Fe_Ti)) %>%
  mutate(Type = case_when(Prediction == min ~ 'True_PN',
                          Prediction != min ~ 'False_PN',
                          TRUE ~ 'ERROR!')) %>%
  ggplot(aes(y = fid)) +
  geom_vline(xintercept = 0.5, lty = 2, col = 'red') +
  geom_bar(aes(fill = Type, col = NULL),
           width = 1, alpha = .3,
           position = 'identity') +
  geom_segment(aes(y=fid, yend=fid, x=0, xend=prob.ORE)) +
  geom_point(aes(x = prob.ORE, col = min), cex = 2.5) +
  coord_cartesian(ylim = c(170,0), expand = TRUE) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  theme(legend.position = 'none',
        axis.text.x = element_text(angle = -90,
                                   hjust = .5,
                                   vjust = .5),
        axis.title.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        plot.margin = unit(c(.03,.03,.03,.03), "lines")
        ) +
  annotate(label = 'Threshold',
           x = 0.58, y = 160,
           fontface = 'italic',
           geom = 'text',
           angle = -90,
           colour = 'red',
           size = 2.5) +
  labs(y ='')

prof4 <-
  dc_samples %>%
  filter(HOLE == 'JBA722') %>%
  filter(!is.na(Fe_Ti)) %>%
  mutate(Type = case_when(Prediction == min ~ 'True_PN',
                          Prediction != min ~ 'False_PN',
                          TRUE ~ 'ERROR!')) %>%
  ggplot(aes(y = fid)) +
  geom_vline(xintercept = 0.5, lty = 2, col = 'red') +
  geom_bar(aes(fill = Type, col = NULL),
           width = 1, alpha = .3,
           position = 'identity') +
  geom_segment(aes(y=fid, yend=fid, x=0, xend=prob.ORE)) +
  geom_point(aes(x = prob.ORE, col = min), cex = 2.5) +
  coord_cartesian(ylim = c(170,0), expand = TRUE) +
  scale_y_reverse() +
  facet_wrap(. ~ HOLE,ncol = 4) +
  theme(legend.position = 'none',
        axis.text.x = element_text(angle = -90,
                                   hjust = .5,
                                   vjust = .5),
        axis.title.x = element_blank(),
        axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        plot.margin = unit(c(.03,.03,.03,.03), "lines")
  ) +
  annotate(label = 'Threshold',
           x = 0.58, y = 160,
```

```
            fontface = 'italic',
            geom = 'text',
            angle = -90,
            colour = 'red',
            size = 2.5) +
  labs(y ='')

ggpubr::ggarrange(h1, prof1, h2, prof2, h3, prof3, h4, prof4,nrow = 2,
                  ncol = 4,
                  align = 'hv',widths = c(rep(c(4,4),4)),
                  common.legend = FALSE,
                  font.label = list(size = 16, face = 'bold'))
```

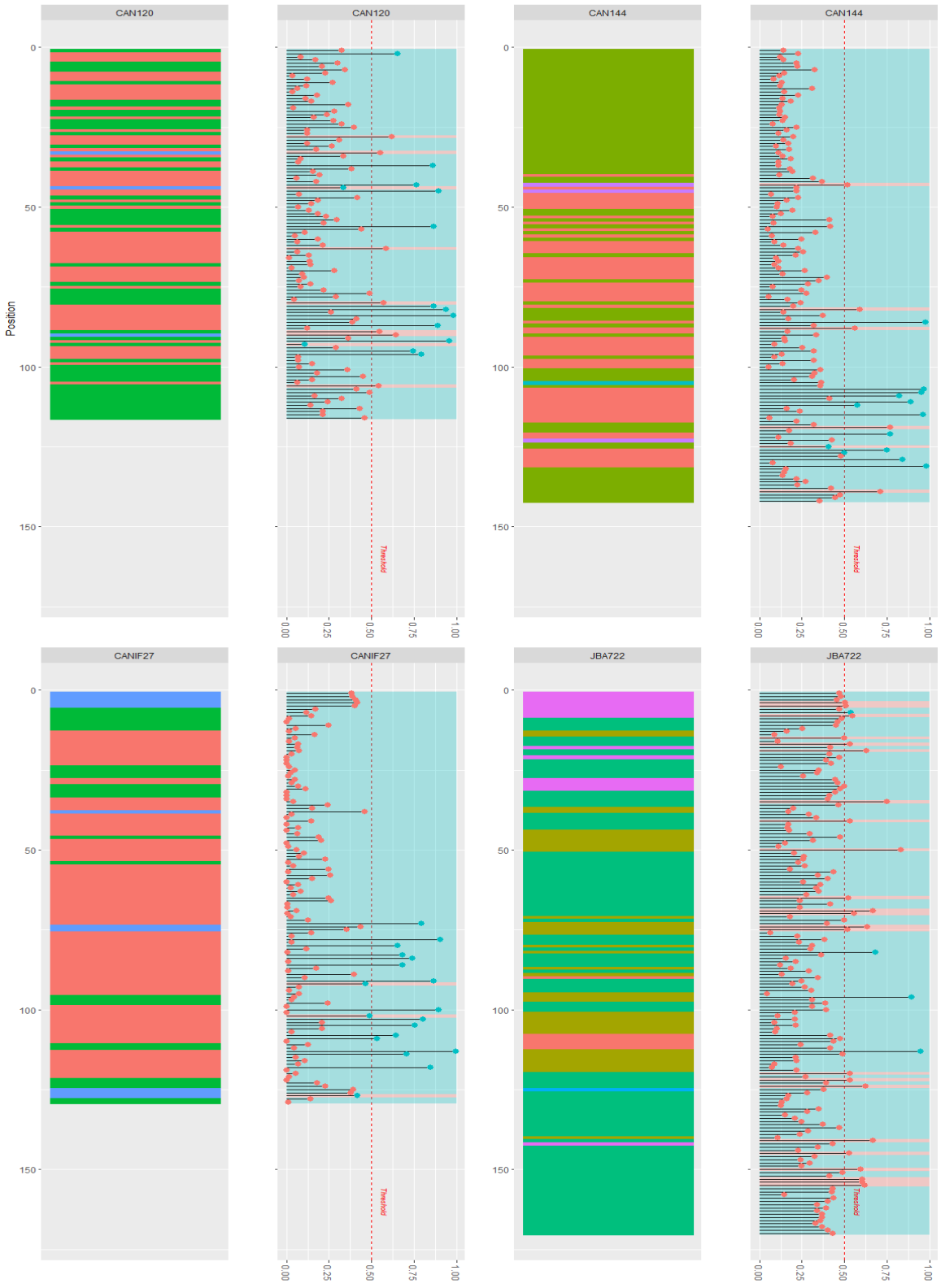


Fig. C. 10: Color-coded strip log according to the lithologies for drill cores studied in this work and respective validation column. The calculated Ore probability is indicated as a bar beside the sample position for each sample and drill core, and the threshold of probability is indicated as the red dashed line. The circle color shows the reference values of the mineralization status at the end of the probability bar. The validation column is color-coded according to the verification of the predicted and reference mineralization status.

APÊNDICE D – Código utilizado no Artigo 02

NOTEBOOK

This is the source code used in the manuscript:

'Machine learning analysis of mineral chemistry in pyrite grains from the Jacobina gold deposits, São Francisco Craton, Brazil: geochemical patterns and implications to mineral exploration'

authored by: "Guilherme Ferreira (guilherme.ferreira@cprm.gov.br)" date: "02/03/2022"

The following code was written in R (4.1.2).

The input data and complementary information can be found at:
https://github.com/gferrsilva/icpms-jacobina

ABSTRACT

We applied machine learning (ML) to process LA-ICP-MS data (45 elements) with 441 samples of pyrite from gold-bearing quartz-pebble-metaconglomerate from the Serra de Jacobina deposits in the São Francisco Craton, Brazil. First, the pyrite samples were described by optical and scanning electron microscopy to gather information about the texture differences. Then, the pyrite grains were classified according to their source and stratigraphical level: detrital and epigenetic pyrite from the mineralized Jacobina Group and pyrite from the basement or intrusive rocks. We used Agglomerative Clustering methods to evaluate the trace elements patterns according to pyrite group, mineral source, and stratigraphic levels. Then, we implemented the Uniform Manifold Approximation and Projection technique (UMAP) to reduce the dimensionality of data into a two-dimensional projection to inspect the inner structure of the data. This result was confirmed by the analysis of the dendrograms, which show different associations of elements among detrital and epigenetic pyrites. Elements such as Cu, Zn, Ag, Sb, Te, Au, Pb, and Bi are mobilized during mineral alteration and was crystallized in newly formed minerals, such as chalcopyrite, pyrrhotite, and sphalerite, which are spatially associated with epigenetic pyrite and free gold. These findings could explain the differences in the mineral assemblage in portions of the deposits that prevail sedimentary minerals or the others that were strongly modified by later alterations. In conclusion, ML is recommended in the processing of mineral chemistry data because it helps to process data without discarding significant variables, and the method allows to evidence the multivariate structure of data.

DATA WRANGLING

Dependencies

```
library(tidyverse) # ggplot2, tidyr, dplyr
library(readxl) # open XLSX data
library(geoquimica) # Data wrangling
library(umap) # Dimensionality reduction
library(pheatmap) # Distance matrices
library(dendextend) # Dendrograms
library(ggpubr) # Plot adjusts
```

Data preparation

```
set.seed(0)

df <- data.table::fread("~/piritas_jacobina_editada_v2.csv")

# Defining variable class
df[,12:142] <- lapply(X = df[,12:142],FUN = as.double)

# Creating variable of imputation control
df1 <- df %>%
  drop_na(`Pyrite Type`) %>%
  mutate(impute_ni60 = ifelse(test = is.na(Ni60),yes = 'True',no = 'False'),
         impute_co = ifelse(test = is.na(Co59),yes = 'True',no = 'False'),
         impute_ti = ifelse(test = is.na(Ti49),yes = 'True',no = 'False'),
         impute_v = ifelse(test = is.na(V51),yes = 'True',no = 'False'))


index <- df1 %>%
  select(`Source file`:`Reef`, impute_ni60:impute_v)

statistics <- df1 %>%
  select(-names(index)) %>%
  select(matches('LOD$|2SE$'))

lod <- df1 %>%
  dplyr::select(matches('LOD$'))

elems <- df1 %>%
  dplyr::select(-names(index),-V1)
```

Data selection

```
geoquimica::elem_fillrate( data.table::fread(
  "~/GitHub/jacobina/data/minchem/piritas_jacobina_editada_v2.csv",
  verbose = FALSE) %>%
    mutate_at(.vars = 12:143,.funs = as.double) %>%
    drop_na(`Pyrite Type`) %>% # Drop wrong analysis
    select(-(V1:`Source file`),-(DateTime:Comments)) %>%
    janitor::clean_names()
  ) %>%
  filter(!str_detect(string = Column.Name, pattern = 'lod$|2se$'),
```

```
         !Column.Name %in% c('datetime','generation','pyrite_type','texture','reef
')) %>%
  arrange(Fill.Rate) %>%
  mutate(Column.Name = fct_inorder(Column.Name)) %>%
  # mutate()
  ggplot(aes(x = Column.Name, y = Fill.Rate)) +
  geom_col(aes(fill = ifelse(test = Fill.Rate < 50,'Non-selected','Selected')), co
l = 'gray') +
  geom_hline(yintercept = 50, lty = 5, col = 'red', size = .7) +
  scale_y_continuous(breaks = seq(0,100,10))  +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90,vjust = 0.5, hjust = 1),
        legend.position = c(.15,.8)
        ) +
  labs(x = 'Elements',
       y = '% of non-missing values',
       fill = '')
```

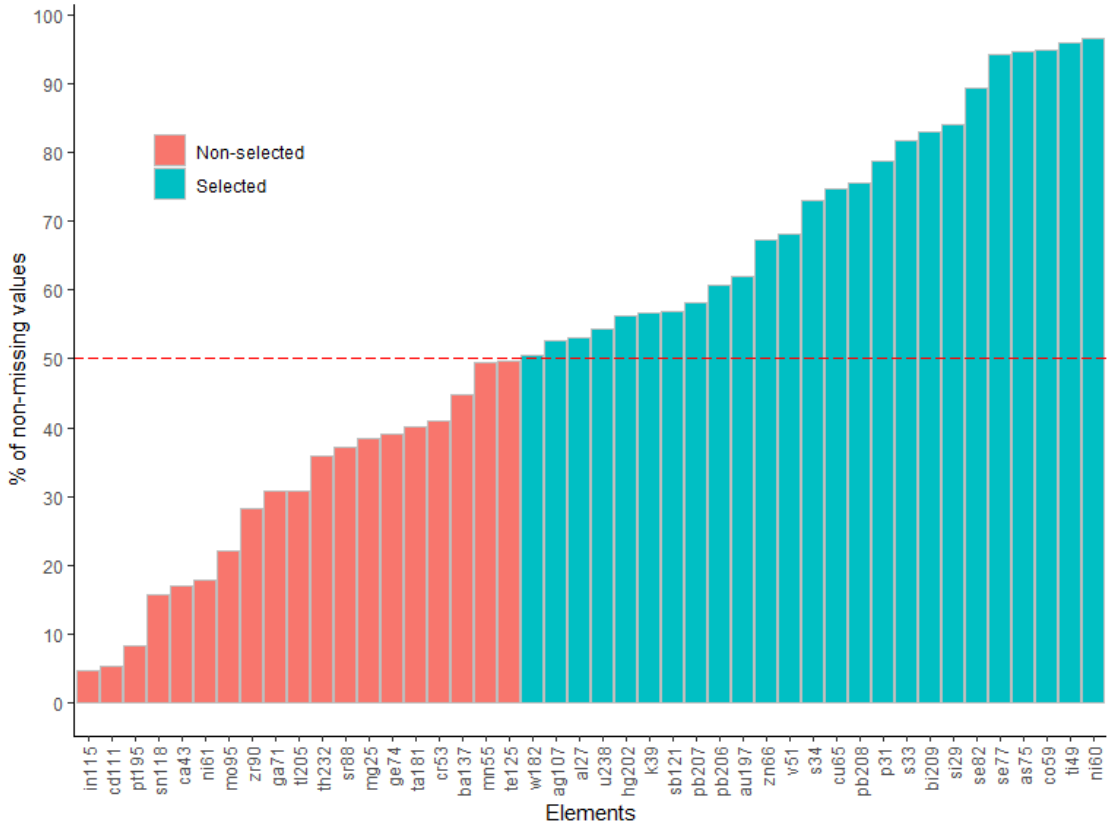


Fig. D. 1: list of elements ordered by the percentage of non-missing data. The 50% threshold (horizontal dashed line) was used to determine if a variable could be selected for multivariate analysis. The elements Al, P, and Si were not selected based on the small variability in the dataset.

IMPUTATION

LDL and Missing Value Imputation

```
# Imputation of LDL elements
half_ldl <- elems %>%
  mutate(Al27 = ifelse(test = is.na(Al27),yes = `Al27 LOD`/sqrt(2),no = Al27),
```

```
         Si29 = ifelse(test = is.na(Si29),yes = `Si29 LOD`/sqrt(2),no = Si29),
         P31 = ifelse(test = is.na(P31),yes = `P31 LOD`/sqrt(2),no = P31),
         S33 = ifelse(test = is.na(S33),yes = `S33 LOD`/sqrt(2),no = S33),
         S34 = ifelse(test = is.na(S34),yes = `S34 LOD`/sqrt(2),no = S34),
         K39 = ifelse(test = is.na(K39),yes = `K39 LOD`/sqrt(2),no = K39),
         Ti49 = ifelse(test = is.na(Ti49),yes = `Ti49 LOD`/sqrt(2),no = Ti49),
         V51 = ifelse(test = is.na(V51),yes = `V51 LOD`/sqrt(2),no = V51),
         Co59 = ifelse(test = is.na(Co59),yes = `Co59 LOD`/sqrt(2),no = Co59),
         Ni60 = ifelse(test = is.na(Ni60),yes = `Ni60 LOD`/sqrt(2),no = Ni60),
         Cu65 = ifelse(test = is.na(Cu65),yes = `Cu65 LOD`/sqrt(2),no = Cu65),
         Zn66 = ifelse(test = is.na(Zn66),yes = `Zn66 LOD`/sqrt(2),no = Zn66),
         As75 = ifelse(test = is.na(As75),yes = `As75 LOD`/sqrt(2),no = As75),
         Se77 = ifelse(test = is.na(Se77),yes = `Se77 LOD`/sqrt(2),no = Se77),
         Se82 = ifelse(test = is.na(Se82),yes = `Se82 LOD`/sqrt(2),no = Se82),
         Ag107 = ifelse(test = is.na(Ag107),yes = `Ag107 LOD`/sqrt(2),no = Ag107),
         Sb121 = ifelse(test = is.na(Sb121),yes = `Sb121 LOD`/sqrt(2),no = Sb121),
         W182 = ifelse(test = is.na(W182),yes = `W182 LOD`/sqrt(2),no = W182),
         Au197 = ifelse(test = is.na(Au197),yes = `Au197 LOD`/sqrt(2),no = Au197),
         Hg202 = ifelse(test = is.na(Hg202),yes = `Hg202 LOD`/sqrt(2),no = Hg202),
         Pb206 = ifelse(test = is.na(Pb206),yes = `Pb206 LOD`/sqrt(2),no = Pb206),
         Pb207 = ifelse(test = is.na(Pb207),yes = `Pb207 LOD`/sqrt(2),no = Pb207),
         Pb208 = ifelse(test = is.na(Pb208),yes = `Pb208 LOD`/sqrt(2),no = Pb208),
         Bi209 = ifelse(test = is.na(Bi209),yes = `Bi209 LOD`/sqrt(2),no = Bi209),
         U238 = ifelse(test = is.na(U238),yes = `U238 LOD`/sqrt(2),no = U238)) %>%
  mutate(impute_pb206 = ifelse(test = is.na(Pb206),yes = 'True',no = 'False'),
         impute_pb207 = ifelse(test = is.na(Pb207),yes = 'True',no = 'False'),
         impute_s33 = ifelse(test = is.na(S33),yes = 'True',no = 'False'),
         impute_s34 = ifelse(test = is.na(S34),yes = 'True',no = 'False')) %>%
  select(-matches('LOD$|2SE$')) %>%
  geoquimica::elem_select(cut = .5)

# Imputation of missing values based on a multivariate non-parametric regression
imputed <- missRanger::missRanger(data = half_ldl,
                                  pmm.k = 3,
                                  maxiter = 10,
                                  seed = 0,
                                  verbose = 2)

##
## Missing value imputation by random forests
##
##   Variables to impute:       P31, S33, S34, K39, Pb206, Pb207
##   Variables used to impute:  Al27, Si29, P31, S33, S34, K39, Ti49, V51, Co59, N
i60, Cu65, Zn66, As75, Se77, Se82, Ag107, Sb121, W182, Au197, Hg202, Pb206, Pb207,
Pb208, Bi209, U238, impute_pb206, impute_pb207, impute_s33, impute_s34
##  P31 S33 K39 Pb206   Pb207   S34
## iter 1:  0.9855  0.4544  0.5600  0.5095  0.2164  0.1844
## iter 2:  0.5228  0.0639  0.2971  0.3116  0.1766  0.0708
## iter 3:  0.5385  0.0658  0.3022  0.3387  0.1589  0.0599
```

Data Recode

```
# Recoding variables Generation, Reef, Reef_label and Unit
df2 <- index %>%
  bind_cols(imputed) %>%
  mutate(Unit = case_when(Reef == 'Basal Reef' ~ 'Serra do Córrego',
                          Reef == 'Main Reef' ~ 'Serra do Córrego',
                          Reef == 'SPC' ~ 'Serra do Córrego',
```

```
                            Reef == 'LU' ~ 'Serra do Córrego',
                            Reef == 'LVLPC' ~ 'Serra do Córrego',
                            Reef == 'MSPC' ~ 'Serra do Córrego',
                            Reef == 'MPC' ~ 'Serra do Córrego',
                            Reef == 'SPC' ~ 'Serra do Córrego',
                            Reef == 'MU' ~ 'Serra do Córrego',
                            Reef == 'Holandez' ~ 'Serra do Córrego',
                            Reef == 'Maneira' ~ 'Serra do Córrego',
                            Reef == 'ITV' ~ 'Intrusive',
                            Reef == 'UMF' ~ 'Intrusive',
                            Reef == 'IQL' ~ 'Serra do Córrego',
                            Reef == 'Basement' ~ 'Basement',
                            Reef == 'CAF' ~ 'Cruz das Almas'),
         Reef_label = case_when(Reef == 'Basal Reef' ~ 'LC',
                                Reef == 'Main Reef' ~ 'LC',
                                Reef == 'SPC' ~ 'UC1',
                                Reef == 'LU' ~ 'UC1',
                                Reef == 'LVLPC' ~ 'UC1',
                                Reef == 'MSPC' ~ 'UC1',
                                Reef == 'MPC' ~ 'UC1',
                                Reef == 'SPC' ~ 'UC1',
                                Reef == 'MU' ~ 'UC1',
                                Reef == 'Holandez' ~ 'UC2',
                                Reef == 'Maneira' ~ 'UC2',
                                Reef == 'ITV' ~ 'Intr.',
                                Reef == 'UMF' ~ 'Intr.',
                                Reef == 'IQL' ~ 'IQL',
                                Reef == 'Basement' ~ 'Base.',
                                Reef == 'CAF' ~ 'CdA'),
         Reef = case_when(Reef == 'Basal Reef' ~ 'Lower Conglomerate',
                            Reef == 'Main Reef' ~ 'Lower Conglomerate',
                            Reef == 'SPC' ~ 'Upper Conglomerate 1',
                            Reef == 'LU' ~ 'Upper Conglomerate 1',
                            Reef == 'LVLPC' ~ 'Upper Conglomerate 1',
                            Reef == 'MSPC' ~ 'Upper Conglomerate 1',
                            Reef == 'MPC' ~ 'Upper Conglomerate 1',
                            Reef == 'SPC' ~ 'Upper Conglomerate 1',
                            Reef == 'MU' ~ 'Upper Conglomerate 1',
                            Reef == 'Holandez' ~ 'Upper Conglomerate 2',
                            Reef == 'Maneira' ~ 'Upper Conglomerate 2',
                            Reef == 'ITV' ~ 'Intrusive',
                            Reef == 'UMF' ~ 'Intrusive',
                            Reef == 'IQL' ~ 'Intermediate Quartzite',
                            Reef == 'Basement' ~ 'Basement',
                            Reef == 'CAF' ~ 'Cruz das Almas'),
  ) %>%
  dplyr::select(-c(`Source file`:Comments, P31,Si29,Al27)) %>%
  mutate(Generation = case_when(Reef == 'Intrusive' ~ 'Intrusive',
                                Reef == 'Basement' ~ 'Basement',
                                TRUE ~ as.character(Generation)),
         Texture = case_when(Reef == 'Intrusive' | Reef == 'Basement' ~ 'Subhedral
',
                             TRUE ~ as.character(Texture)))
```

DIMENSIONALITY REDUCTION

Uniform Manifold Approximation and Projection - UMAP

```
# UMAP processing

dfumap <-
  df2 %>%
  select_if(is.numeric) %>%
  geoquimica::elem_norm(method = 'clr') %>%
  umap::umap()

# Data merging
df3 <-
  df2 %>%
  bind_cols(as_tibble(dfumap$layout))

# Making Plot a)
umap1 <-
  df3 %>%
  ggplot(aes(x = V1, y = V2,
             fill = Generation,
             shape = Generation,
             group = Generation)) +
  geom_vline(xintercept = 0, col = 'grey', size = .7) +
  geom_hline(yintercept = 0, col = 'grey', size = .7) +
  geom_point(inherit.aes = FALSE,
             data = df3 %>%
               group_by(Texture) %>%
               summarize(U1mean = mean(V1),
                         U2mean = mean(V2)),
             mapping = aes(x = U1mean, y = U2mean),
             cex = 3, shape = 3, col = 'black') +
  ggforce::geom_ellipse(inherit.aes = FALSE,
                        data = df3 %>%
                          group_by(Texture) %>%
                          summarize(U1mean = mean(V1),
                                    U1sd = sd(V1),
                                    U2mean = mean(V2),
                                    U2sd = sd(V2)),
                        mapping = aes(x0 = U1mean, y0 = U2mean, a = 2*U1sd, b = 2*
U2sd, angle = 0),
                        fill = 'grey', alpha = .05) +
  geom_point(alpha = .6, col = 'black') +
  scale_shape_manual(values = c(22,23,24, 21)) +
  facet_wrap(. ~ Texture,nrow = 2) +
  labs(x = 'UMAP1', y = 'UMAP2') +
  theme_bw()

# Making plot b)
umap2 <-
  df3 %>%
  ggplot(aes(x = V1, y = V2,
             fill = Generation,
             shape = Generation,
             group = Generation)) +
  geom_vline(xintercept = 0, col = 'grey', size = .7) +
```

```
  geom_hline(yintercept = 0, col = 'grey', size = .7) +
  geom_point(inherit.aes = FALSE,
             data = df3 %>%
               group_by(Reef) %>%
               summarize(U1mean = mean(V1),
                         U2mean = mean(V2)),
             mapping = aes(x = U1mean, y = U2mean),
             cex = 3, shape = 3, col = 'black') +
  ggforce::geom_ellipse(inherit.aes = FALSE,
                        data = df3 %>%
                          group_by(Reef) %>%
                          summarize(U1mean = mean(V1),
                                    U1sd = sd(V1),
                                    U2mean = mean(V2),
                                    U2sd = sd(V2)),
                        mapping = aes(x0 = U1mean, y0 = U2mean, a = 2*U1sd, b = 2*
U2sd, angle = 0),
                        fill = 'grey', alpha = .05) +
  geom_point(alpha = .6, col = 'black') +
  scale_shape_manual(values = c(22,23,24, 21,
                                22,23,24, 21)) +
  facet_wrap(. ~ Reef,ncol = 3) +
  labs(x = 'UMAP1', y = 'UMAP2') +
  theme_bw() #+

# Arrange plots
ggarrange(umap1, umap2,nrow = 2,
          labels = c('a)','b)'),
          align = 'hv',
          heights = c(2,3),font.label = list(size = 16, face = 'bold'))
```

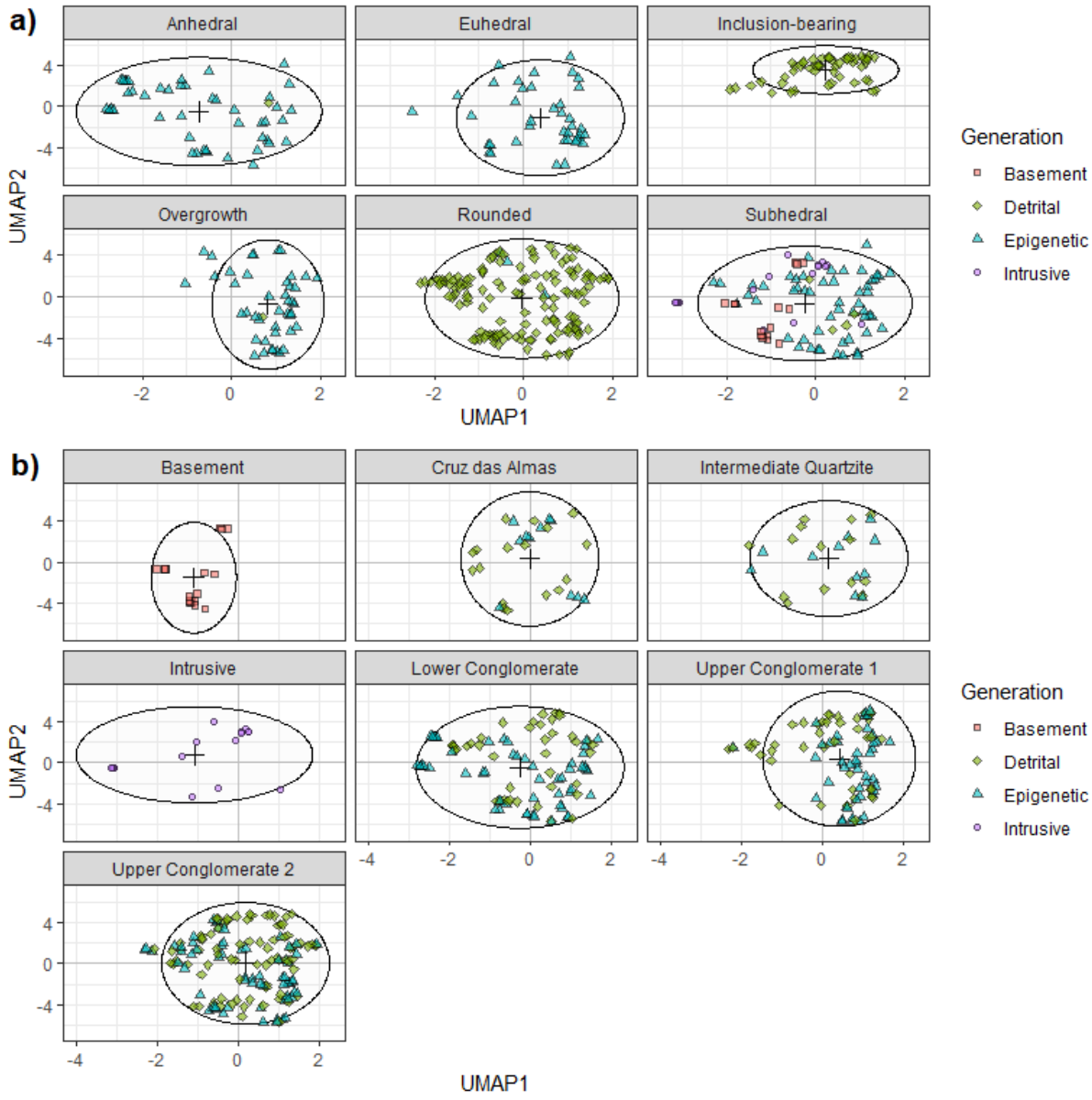


Fig. D. 2: UMAP configuration for data with samples classified according to a) pyrite texture and b) stratigraphic level. The ellipses drawn in each box are centered based on the mean of the samples, and its axes are calculated according to two times the standard deviation for each UMAP coordinate.

DISTANCE MATRICES

```
d <-
  df2 %>%
  filter(Texture == 'Inclusion-bearing') %>%
  geoquimica::elem_norm(method = 'clr') %>%
  geoquimica::elem_norm() %>%
  select(K39:U238) %>%
  t()

# Making Fig. A, Inclusion-bearing Pyrite
inbear <-
  ggplotify::as.ggplot(pheatmap(mat = df2 %>%
             filter(Texture == 'Inclusion-bearing') %>%
```

```
              geoquimica::elem_norm(method = 'clr') %>%
              geoquimica::elem_norm() %>%
              select(K39:U238) %>%
              t() %>%
              dist(method = 'manhattan'),
            labels_row = rownames(d),
            labels_col = rownames(d),
            border_color = 'white',
            clustering_distance_rows = 'manhattan',
            clustering_distance_cols = 'manhattan',
            color = pals::turbo(20),
            cutree_rows = 8,
            cutree_cols = 8,
            main = "Inclusion-bearing pyrite",
            treeheight_row = 15,
            treeheight_col = 15,))

# Making Fig. B, Overgrowth Pyrite
overg <- ggplotify::as.ggplot(pheatmap(mat = df2 %>%
              filter(Texture == 'Overgrowth') %>%
              geoquimica::elem_norm(method = 'clr') %>%
              geoquimica::elem_norm() %>%
              select(K39:U238) %>%
              t() %>%
              dist(method = 'manhattan'),
            labels_row = rownames(d),
            border_color = 'white',
            labels_col = rownames(d),
            color = pals::turbo(20),
            clustering_distance_rows = 'manhattan',
            clustering_distance_cols = 'manhattan',
            cutree_rows = 5,
            cutree_cols = 5,
            main = "Overgrowth pyrite",
            treeheight_row = 15,
            treeheight_col = 15))

# Making Fig. C, Rounded Pyrite
rounded <- ggplotify::as.ggplot(pheatmap(mat = df2 %>%
              filter(Texture == 'Rounded') %>%
              geoquimica::elem_norm(method = 'clr') %>%
              geoquimica::elem_norm() %>%
              select(K39:U238) %>%
              t() %>%
              dist(method = 'manhattan'),
            labels_row = rownames(d),
            labels_col = rownames(d),
            border_color = 'white',
            clustering_distance_rows = 'manhattan',
            clustering_distance_cols = 'manhattan',
            color = pals::turbo(20),
            cutree_rows = 8,
            cutree_cols = 8,
            main = "Rounded pyrite",
            treeheight_row = 15,
            treeheight_col = 15))

# Making Fig. D, Euhedral Pyrite
euhedral <- ggplotify::as.ggplot(pheatmap(mat = df2 %>%
```

```
              filter(Texture == 'Euhedral') %>%
              geoquimica::elem_norm(method = 'clr') %>%
              geoquimica::elem_norm() %>%
              select(K39:U238) %>%
              t() %>%
              dist(method = 'manhattan'),
            labels_row = rownames(d),
            labels_col = rownames(d),
            border_color = 'white',
            clustering_distance_rows = 'manhattan',
            clustering_distance_cols = 'manhattan',
            color = pals::turbo(20),
            cutree_rows = 6,
            cutree_cols = 6,
            main = "Euhedral pyrite",
            treeheight_row = 15,
            treeheight_col = 15))

# Arrange plots
ggarrange(inbear,overg,rounded,euhedral,
          ncol = 2,nrow = 2,
          labels = c('a)','b)','c)','d)'),align = 'hv',hjust = 0)
```

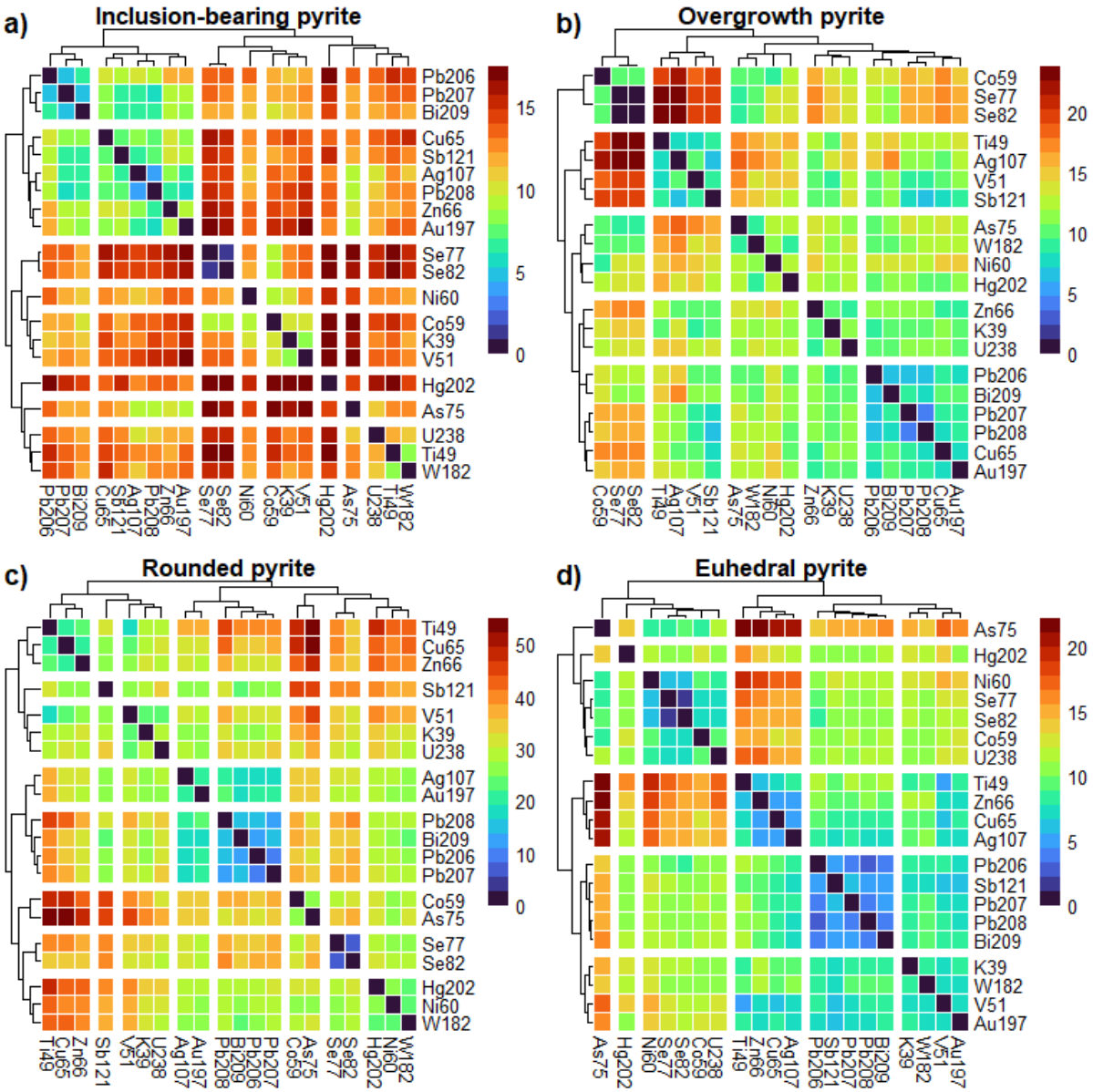


Fig. D. 3: Dendrograms and distance matrices for pyrite grains according to grain texture: a) Inclusion-bearing, b) Overgrowth, c) Anhedral, and d) Euhedral. Rows and columns are ordered according to the agglomerative clustering, calculated by the Manhattan distance.

Compared dendrograms

```
d1 <-
  df3 %>%
  filter(Texture == 'Inclusion-bearing') %>%
  select(K39:U238) %>%
  geoquimica::elem_norm(method = 'clr') %>%
  t() %>%
  dist(method = 'manhattan') %>%
  hclust(method = 'average') %>%
  as.dendrogram()

d2 <-
  df3 %>%
```

```
  filter(Texture == 'Overgrowth') %>%
  select(K39:U238) %>%
  geoquimica::elem_norm(method = 'clr') %>%
  t() %>%
  dist(method = 'manhattan') %>%
  hclust(method = 'average') %>%
  as.dendrogram()


# Custom these kendo, and place them in a list
dl1 <-
  dendextend::dendlist(
    d1 %>%
      set("branches_lty", 1) %>%
      set("labels_col",
          h = 40) %>%
      set("branches_lty", 1) %>%
      set("branches_k_color",
          h = 40),
    d2 %>%
      set("branches_lty", 1) %>%
      set("labels_col",
          h = 60) %>%
      set("branches_lty", 1) %>%
      set("branches_k_color",
          h = 60)
  )

# Plot them together
dendextend::tanglegram(dl1,
             common_subtrees_color_lines = FALSE,
             highlight_distinct_edges  = FALSE,
             highlight_branches_lwd = FALSE,
             margin_inner=4, intersecting = TRUE,
             lwd=1, k_labels = NULL, k_branches = NULL)
```

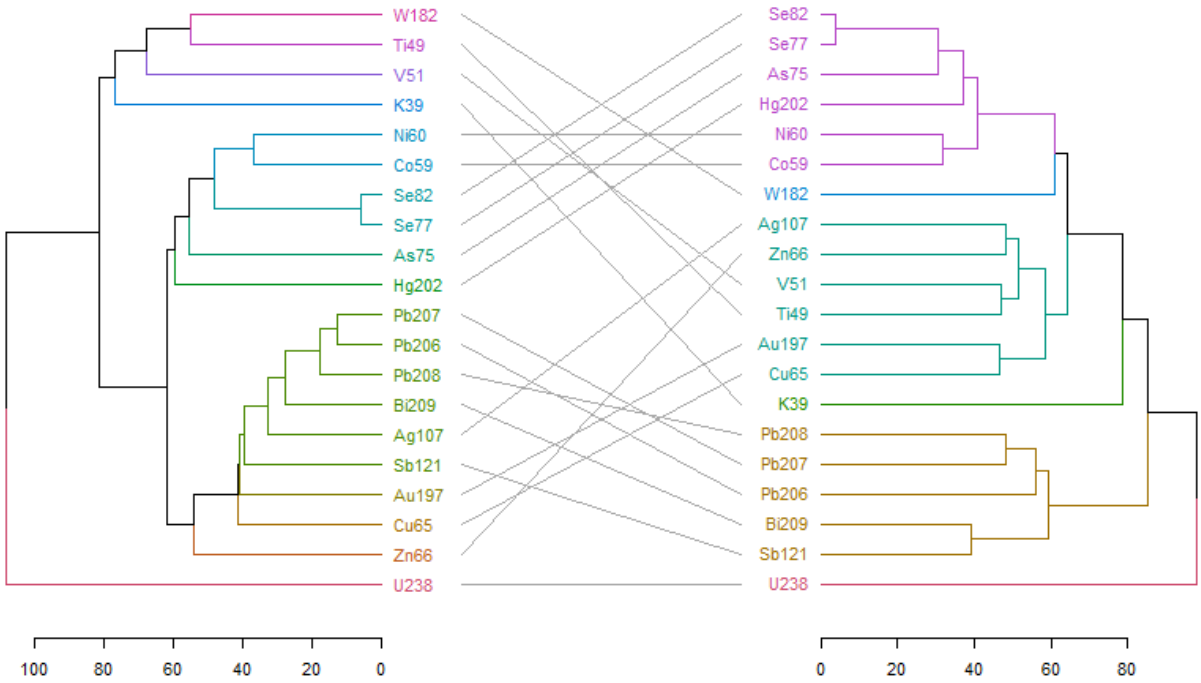


Fig. D. 4: Linked dendrograms for Detrital (left) and Epigenetic pyrites (right) mapping the relation of elements.

IMPUTATION OF LDL AND MISSING VALUES

Replacement of LDL

```
# Function to make qq plots coded by categorical variables
make_qq <- function(dd, x) {
  dd<-dd[order(dd[[x]]), ]
  dd$qq <- qnorm(ppoints(nrow(dd)))
  dd
}

# Data without imputation
p_original <-
  df3 %>%
  filter(impute_v == 'False') %>%
  make_qq(dd = .,x = 'V51') %>%
  ggplot(aes(x = qq, y = log(V51))) +
  geom_qq_line(aes(sample = qq),
               lty = 1, col = 'gray', size = 1,
               alpha = .7) +
  geom_point(aes(col = impute_v,
                 shape = impute_v),
             # col = 'white',
             alpha = .3) +
  labs(x = 'Theoretical distribution',
       y = 'log(V51)',
       col = 'Imputed?',
       shape = 'Imputed?') +
  scale_shape_manual(values = c(16,17)) +
  theme_classic() +
```

```
  theme(legend.justification = 'center',
        legend.background = element_rect(
          fill = 'grey98'))

# Data with half of LDL imputed values
p_ldl <-
  df3 %>%
  make_qq(dd = .,x = 'V51') %>%
  ggplot(aes(x = qq, y = log(V51))) +
  geom_qq_line(aes(sample = qq),
               lty = 1, col = 'gray', size = 1,
               alpha = .7) +
  geom_point(aes(col = impute_v,
                 shape = impute_v),
             # col = 'white',
             alpha = .4) +
  labs(x = 'Theoretical distribution',
       y = 'log(V51)',
       col = 'Imputed?',
       shape = 'Imputed?') +
  scale_shape_manual(values = c(16,17)) +
  theme_classic() +
  theme(legend.justification = 'center',
        legend.background = element_rect(
          fill = 'grey98'))

# Random Forest Regression without categorical variables
p_rf1 <-
  df %>%
  drop_na(`Pyrite Type`) %>%
  janitor::clean_names() %>%
  select(-matches('lod$|2se$')) %>%
  select(-`pyrite_type`,-generation,-texture,-reef,-c(source_file:comments)) %>%
  missRanger::missRanger(data = .,
                         pmm.k = 5,
                         # formula = Ni61 ~ Ni60,
                         maxiter = 10,
                         seed = 321321,
                         verbose = 0,
                         num.trees = 1000) %>%
  bind_cols(df1$impute_ti,
            df1$impute_v,
            df1$impute_co,
            df1$impute_ni60) %>%
  rename(impute_ti = `...46`,
         impute_v = `...47`,
         impute_co = `...48`,
         impute_ni60 = `...49`) %>%
  make_qq(dd = .,x = 'v51') %>%
  ggplot(aes(x = qq, y = log(v51))) +
  geom_qq_line(aes(sample = qq),
               lty = 1, col = 'gray', size = 1,
               alpha = .7) +
  geom_point(aes(col = impute_v,
                 shape = impute_v),
             alpha = .4,
             # col = 'white'
  ) +
  labs(x = 'Theoretical distribution',
```

```
      y = 'log(V51)',
      col = 'Imputed?',
      shape = 'Imputed?') +
  scale_shape_manual(values = c(16,17)) +
  theme_classic() +
  theme(legend.justification = 'center',
        legend.background = element_rect(
          fill = 'grey98'))

##
## Missing value imputation by random forests
##
##   Variables to impute:       mg25, al27, si29, p31, s33, s34, k39, ca43, ti49,
v51, cr53, mn55, co59, ni60, cu65, zn66, ga71, ge74, as75, se77, se82, sr88, zr90,
mo95, ag107, cd111, in115, sn118, sb121, te125, ba137, ta181, w182, pt195, au197,
hg202, tl205, pb206, pb207, pb208, bi209, th232, u238, ni61
##   Variables used to impute:  v1, mg25, al27, si29, p31, s33, s34, k39, ca43, ti
49, v51, cr53, mn55, co59, ni60, cu65, zn66, ga71, ge74, as75, se77, se82, sr88, z
r90, mo95, ag107, cd111, in115, sn118, sb121, te125, ba137, ta181, w182, pt195, au
197, hg202, tl205, pb206, pb207, pb208, bi209, th232, u238, ni61
##  ni60    ti49    co59    as75    se77    se82    si29    bi209   s33 p31 pb208
cu65    s34 v51 zn66    au197   pb206   pb207   sb121   k39 hg202   u238    al27
ag107   w182    te125   mn55    ba137   cr53    ta181   ge74    mg25    sr88    th
232   ga71    tl205   zr90    mo95    ni61    ca43    sn118   pt195   cd111   in11
5
## iter 1:  1.0984  0.9778  0.9965  0.8076  0.8318  0.4159  0.9323  1.0150  0.4100
0.9744  0.8928  0.0863  0.3518  0.6374  1.0521  0.7236  0.5422  0.2208  0.9814  0.
6101  0.9118  1.0325  0.8896  0.7977  0.8346  1.0043  1.0046  0.8982  0.6808  0.92
51  0.5294  0.7825  1.0083  0.8175  0.5663  0.5401  0.8608  0.4141  0.6516  0.7580
0.1158  1.0724  1.1716  0.2345
## iter 2:  0.3723  0.4508  0.4851  0.4009  0.3997  0.4136  0.6449  0.6256  0.1106
0.4962  0.6807  0.0907  0.1298  0.2900  1.0457  0.5339  0.4181  0.2509  0.8185  0.
3246  0.3550  0.8766  0.8124  0.6530  0.5592  0.9023  0.6879  0.7890  0.4692  0.69
20  0.4680  0.5910  0.9721  0.5138  0.3880  0.3131  0.4296  0.2869  0.5159  0.4987
0.0920  1.0014  1.1498  0.1896
## iter 3:  0.3508  0.4495  0.4654  0.4452  0.4003  0.3987  0.6406  0.6372  0.1052
0.5104  0.6879  0.1203  0.1243  0.2826  1.0382  0.5882  0.4123  0.2394  0.8141  0.
3171  0.3951  0.8912  0.8210  0.6645  0.6163  0.8923  0.6323  0.7860  0.4564  0.70
57  0.4843  0.5951  0.9771  0.5449  0.3799  0.2903  0.4606  0.2789  0.5045  0.4487
0.0907  1.0200  1.1339  0.2114

# Random Forests regression with caterogical variables
p_rf2 <-
  df %>%
  drop_na(`Pyrite Type`) %>%
  janitor::clean_names() %>%
  select(-matches('lod$|2se$')) %>%
  select(-`pyrite_type`,-c(source_file:comments)) %>%
  missRanger::missRanger(data = .,
                         pmm.k = 5,
                         maxiter = 10,
                         seed = 321321,
                         verbose = 2,
                         num.trees = 1000) %>%
  bind_cols(df1$impute_ti,
            df1$impute_v,
            df1$impute_co,
            df1$impute_ni60) %>%
  rename(impute_ti = `...49`,
```

```
         impute_v = `...50`,
         impute_co = `...51`,
         impute_ni60 = `...52`) %>%
  make_qq(dd = .,x = 'v51') %>%
  ggplot(aes(x = qq, y = log(v51))) +
  geom_qq_line(aes(sample = qq),
               lty = 1, col = 'gray', size = 1,
               alpha = .7) +
  geom_point(aes(col = impute_v,
                 shape = impute_v),
             alpha = .4,
             # col = 'white'
  ) +
  labs(x = 'Theoretical distribution',
       y = 'log(V51)',
       col = 'Imputed?',
       shape = 'Imputed?') +
  scale_shape_manual(values = c(16,17)) +
  theme_classic() +
  theme(legend.justification = 'center',
        legend.background = element_rect(
          fill = 'grey98'))

# Adjusting the plots
ggpubr::ggarrange(p_ldl, p_rf1, p_rf2,p_original,
                  ncol = 2,nrow = 2,align = 'hv',
                  labels = c('a)','b)','c)','d)'),
                  common.legend = TRUE,legend = 'right')
```

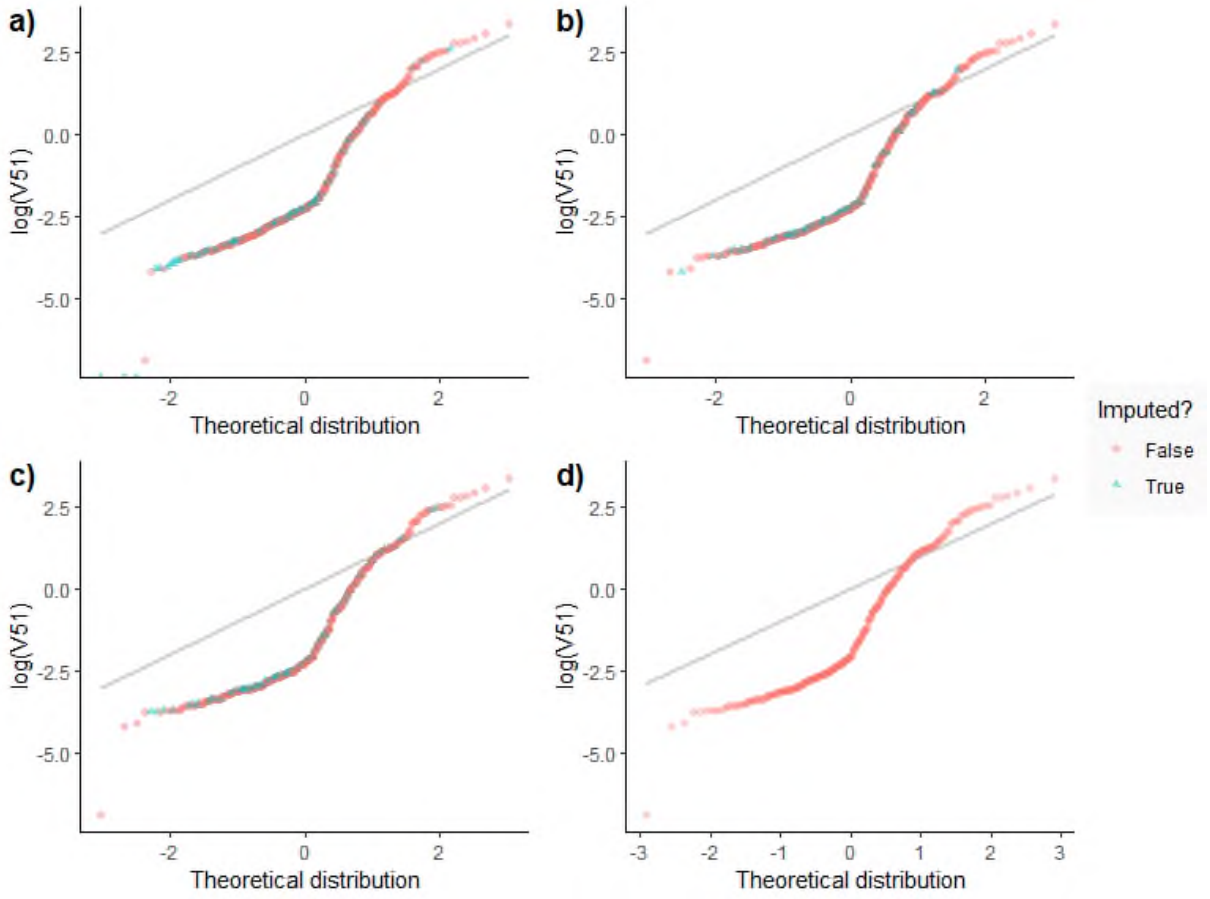


Fig. D. 5: Quantile-plot for the concentration of V51 on a logarithmic scale according to different distributions: a) imputation based on a fraction of the detection limit, b) Random Forests imputation based only in quantitative variables, c) Random Forests imputation based on quantitative and categorical variables (e.g., grain texture) and d) original distribution (without imputed values).

Replacement of Missing Values

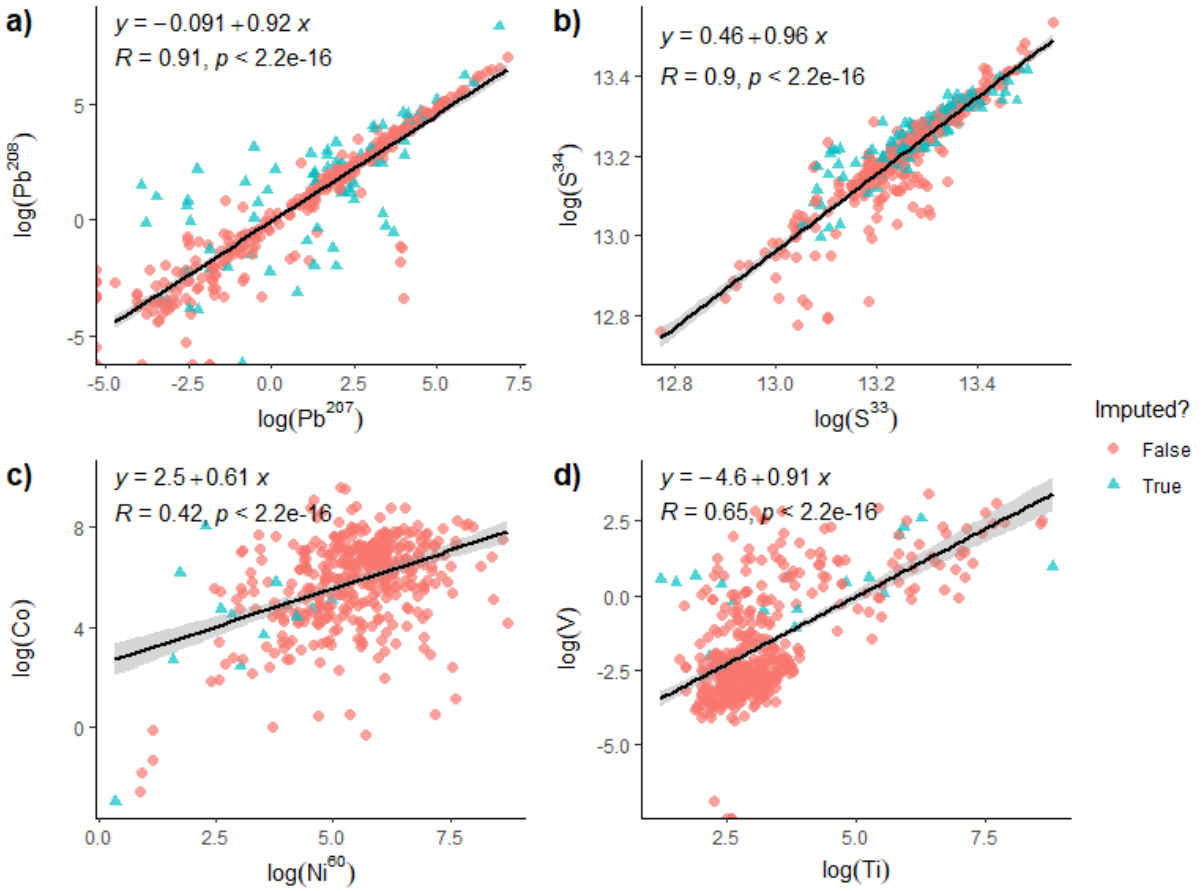
```
p1 <-
  df2 %>%
  ggplot(aes(log(Pb207), log(Pb208))) +
  geom_point(aes(col = impute_pb207,
                 shape = impute_pb206),
             cex = 2,
             alpha = .7) +
  geom_smooth(method = 'lm', col = 'black') +
  theme_classic() +
  ggpubr::stat_regline_equation(label.y = 8.5) +
  ggpubr::stat_cor(label.y = 7) +
  labs(col = 'Imputed?', shape = 'Imputed?',
       x = expression(log(Pb^207)),
       y = expression(log(Pb^208)))

p2 <-
  df2 %>%
  ggplot(aes(log(S33),
             log(S34))) +
  geom_point(aes(col = impute_s34,
```

```
                shape = impute_s34),
            cex = 2,
            alpha = .7) +
  geom_smooth(method = 'lm', col = 'black') +
  theme_classic() +
  ggpubr::stat_regline_equation(label.y = 13.5) +
  ggpubr::stat_cor(label.y = 13.4) +
  labs(col = 'Imputed?', shape = 'Imputed?',
       x = expression(log(S^33)),
       y = expression(log(S^34)))

p3 <-
  df2 %>%
  ggplot(aes(log(Ni60),
             log(Co59))) +
  geom_point(aes(col = impute_ni60,
                 shape = impute_ni60),
             cex = 2,
             alpha = .7) +
  geom_smooth(method = 'lm', col = 'black') +
  theme_classic() +
  ggpubr::stat_regline_equation(label.y = 10) +
  ggpubr::stat_cor(label.y = 8.5) +
  labs(col = 'Imputed?', shape = 'Imputed?',
       x = expression(log(Ni^60)),
       y = expression(log(Co)))

p4 <-
  df2 %>%
  ggplot(aes(log(Ti49),
             log(V51))) +
  geom_point(aes(col = impute_ti,
                 shape = impute_ti),
             cex = 2,
             alpha = .7) +
  geom_smooth(method = 'lm', col = 'black') +
  theme_classic() +
  ggpubr::stat_regline_equation(label.y = 4) +
  ggpubr::stat_cor(label.y = 2.8) +
  labs(col = 'Imputed?', shape = 'Imputed?',
       x = expression(log(Ti)),
       y = expression(log(V)))
# Adjusting the plots
ggpubr::ggarrange(p1, p2, p3,p4,
                  ncol = 2,nrow = 2,align = 'hv',
                  labels = c('a)','b)','c)','d)'),
                  common.legend = TRUE,legend = 'right')
```

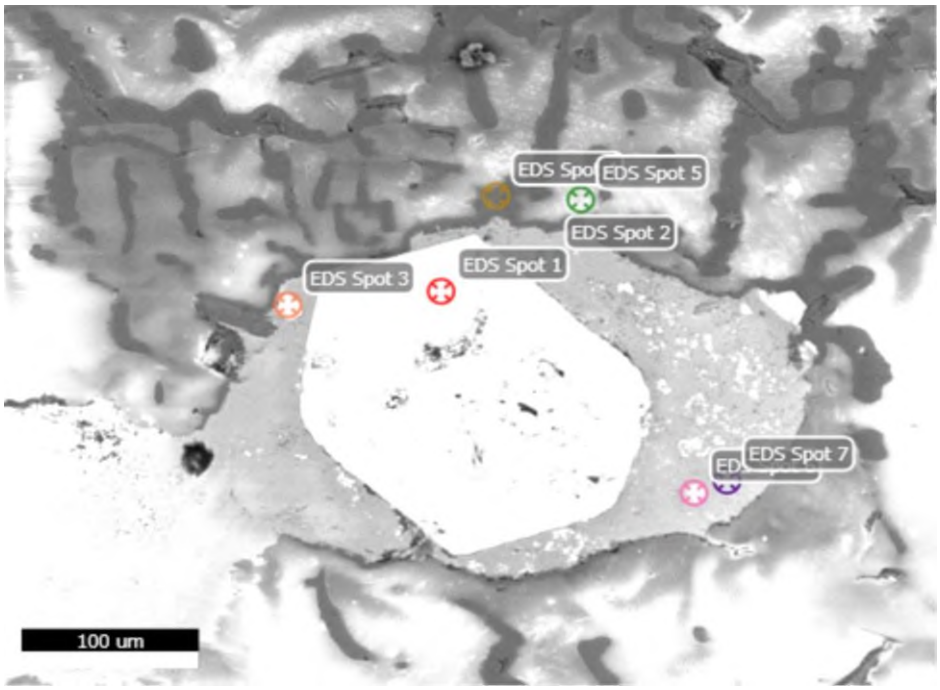


691

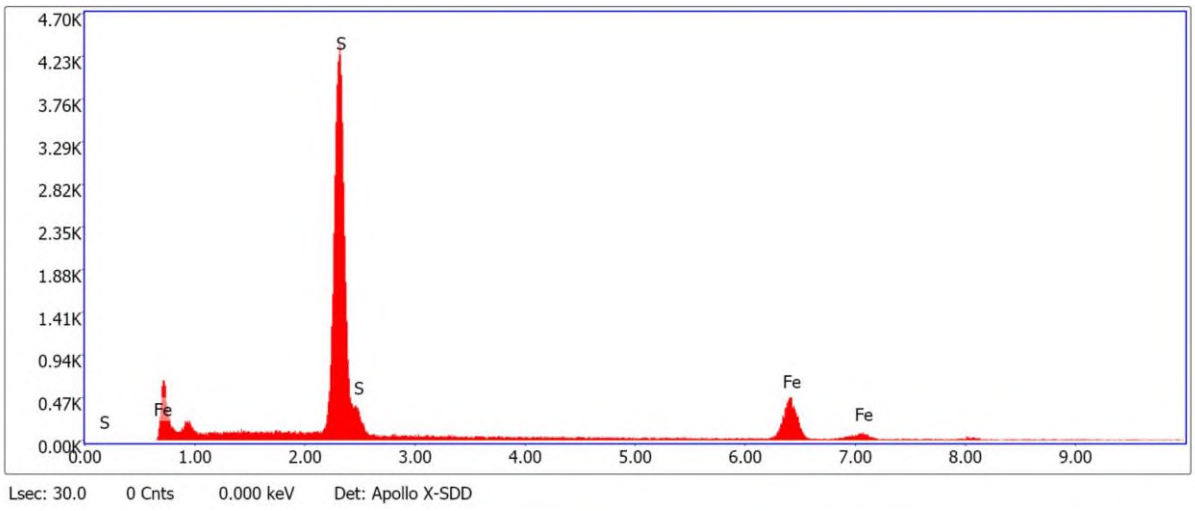
692 Fig. D. 6: Comparison of correlated log-transformed elements showing the relation of values according to whether
693 they originated from imputation (True, blue triangles) or not (False, orange circles). This plot was prepared based
694 on high correlated isotopes a) Pb207 - Pb208, b) S33 - S34, and based on moderated correlated elements c) Ni60-
695 Co and d) Ti – V.

APÊNDICE F –Análises de EDS

LÂMINA HHS-443 – CAMPO 01

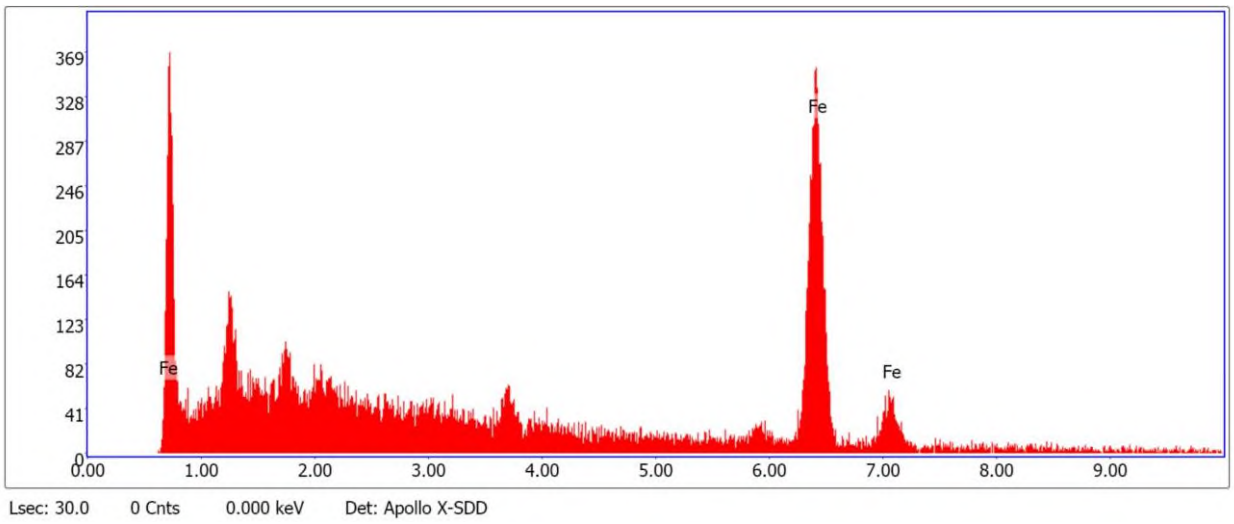


EDS Spot 1 - EDS1



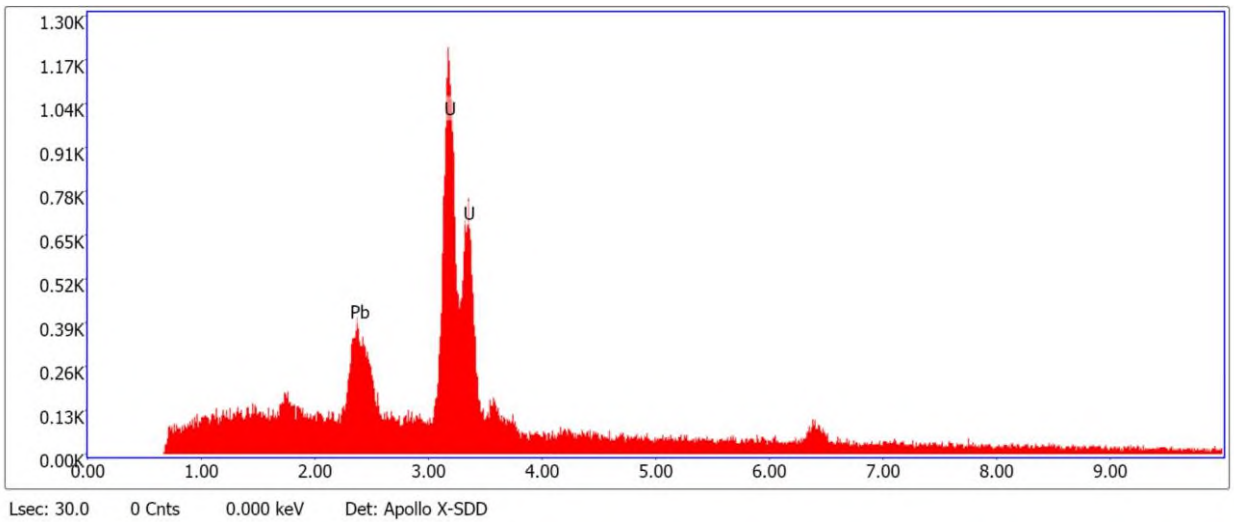
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
S K	59.95	72.28	1616.87	2.39	0.5526	0.9625	0.9534	1.0046
FeK	40.05	27.72	222.28	4.52	0.3424	0.8511	0.9944	1.0100

EDS Spot 2 - EDS1



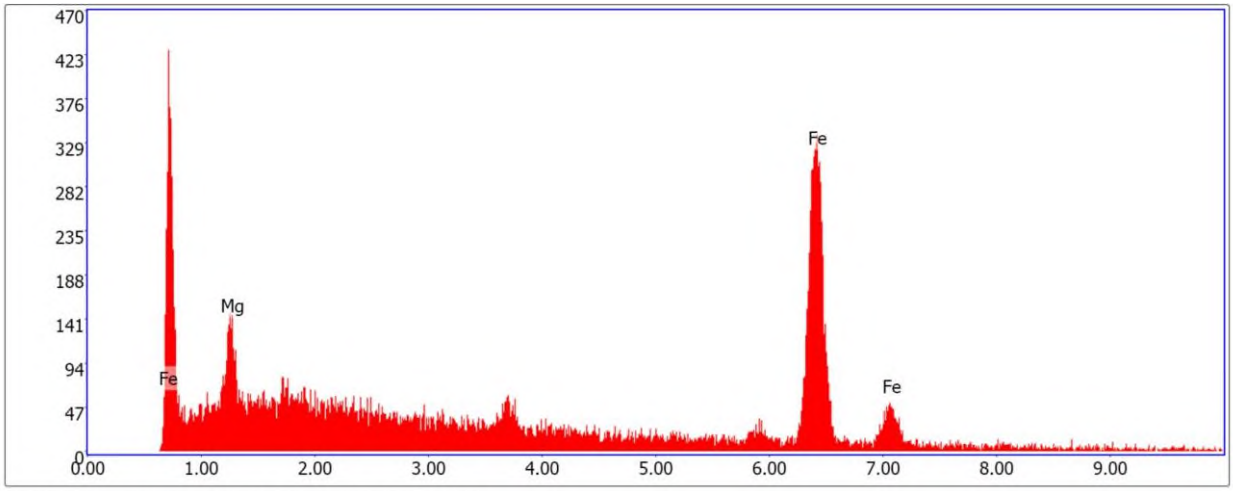
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
FeK	100.00	100.00	158.46	5.00	0.9291	0.9266	1.0027	1.0000

EDS Spot 6 - EDS1

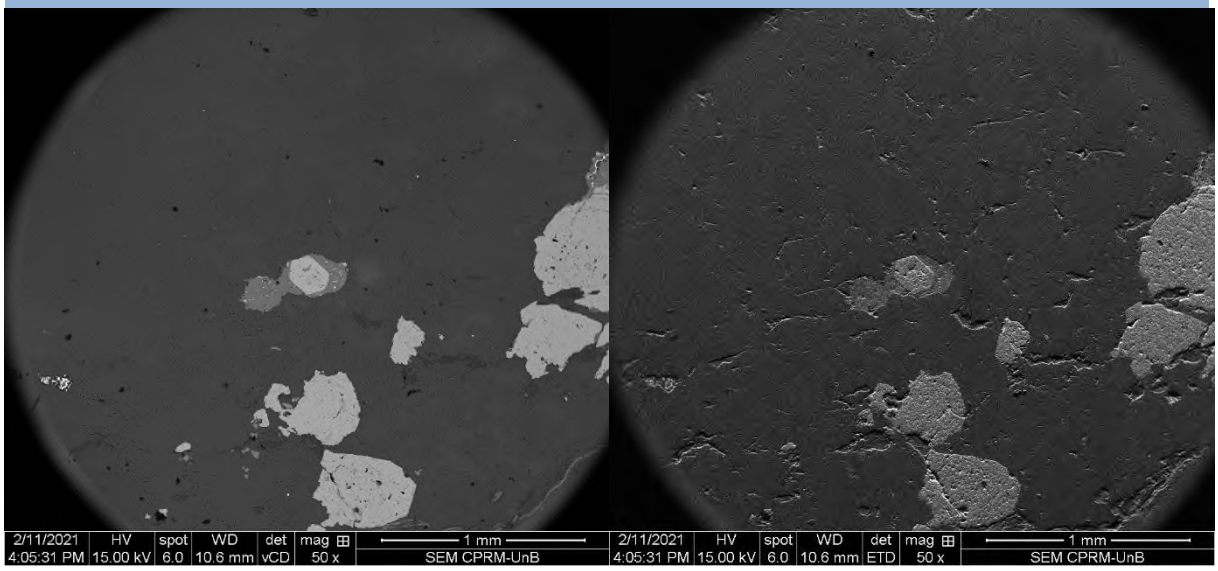


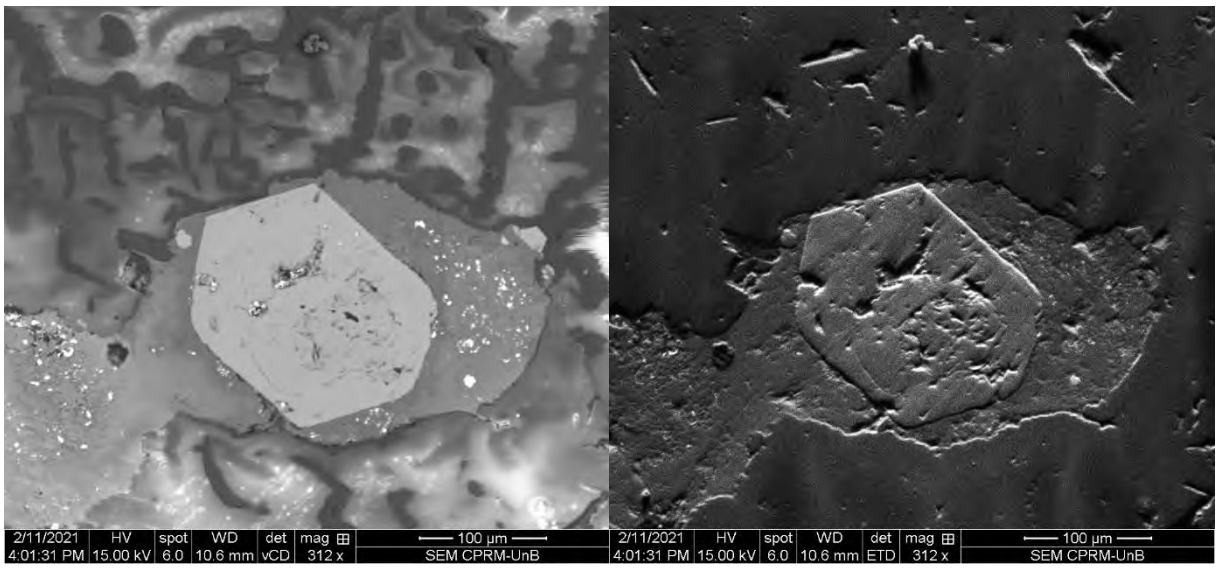
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
PbM	16.08	18.04	109.92	6.92	0.1403	0.8525	1.0187	1.0049
U M	83.92	81.96	444.74	3.76	0.6991	0.8323	1.0020	0.9990

EDS Spot 7 - EDS1

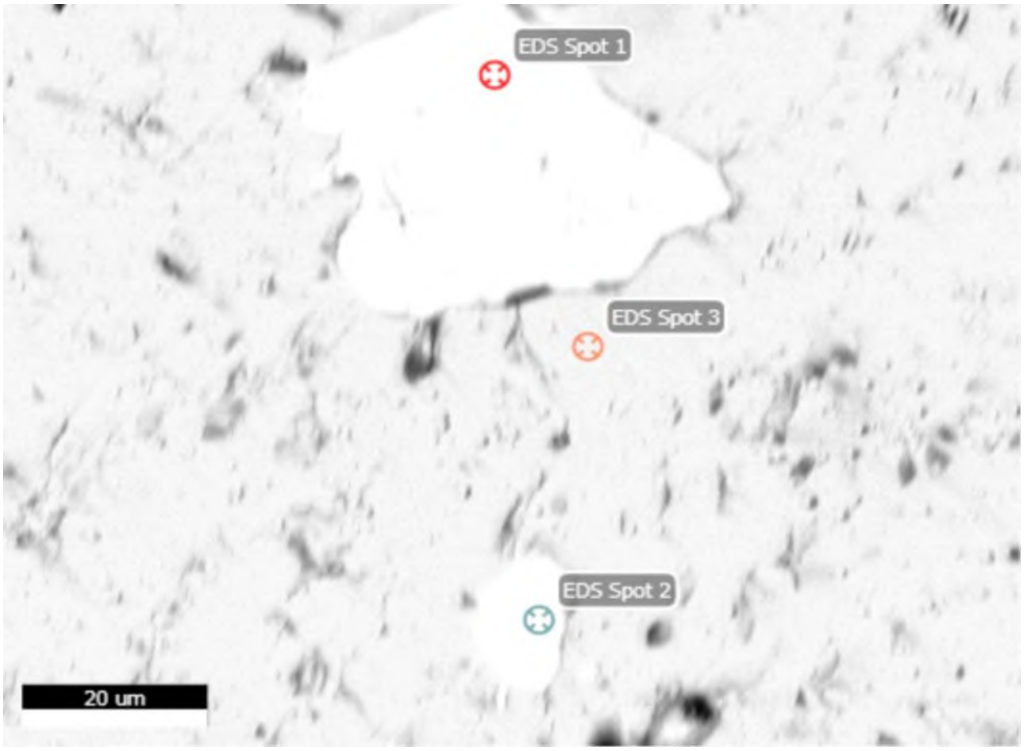


Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
MgK	3.87	8.46	25.26	21.32	0.0209	1.0570	0.5114	1.0017
FeK	96.13	91.54	163.71	4.73	0.8870	0.9199	1.0028	1.0003

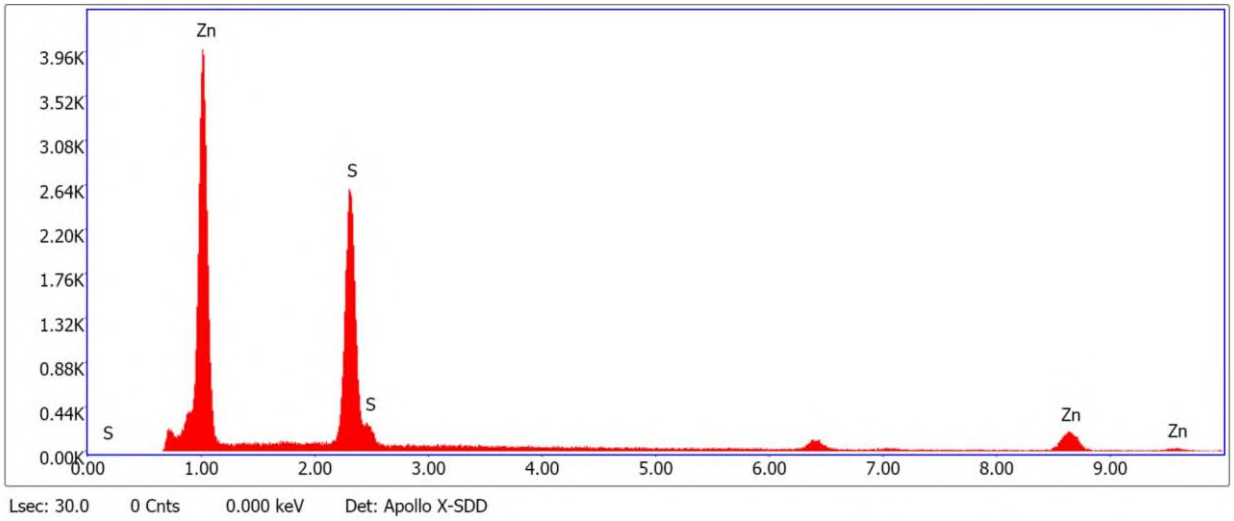




LÂMINA HHS-443 – CAMPO 02

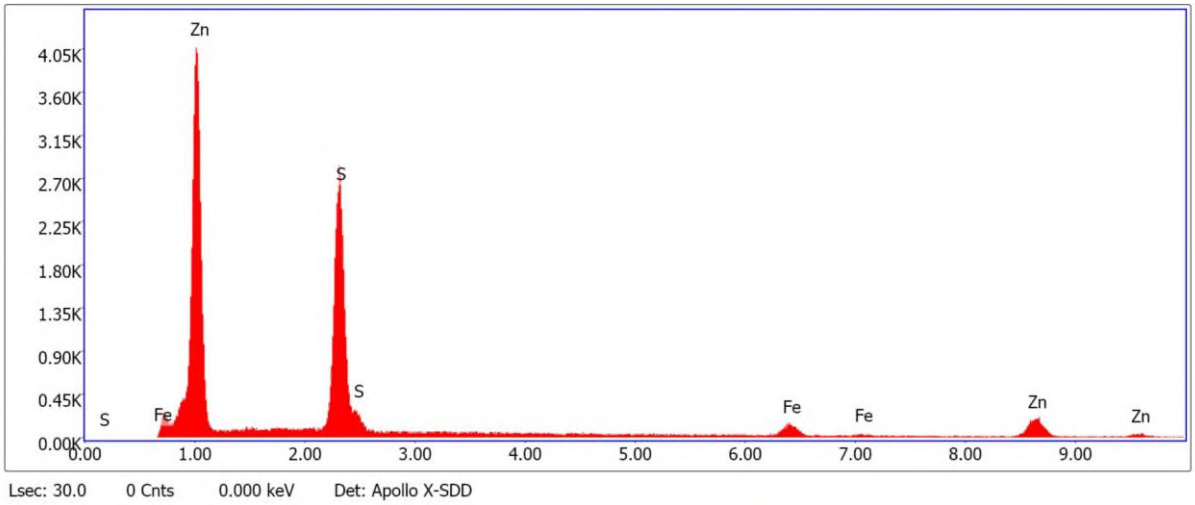


EDS Spot 1 - EDS1



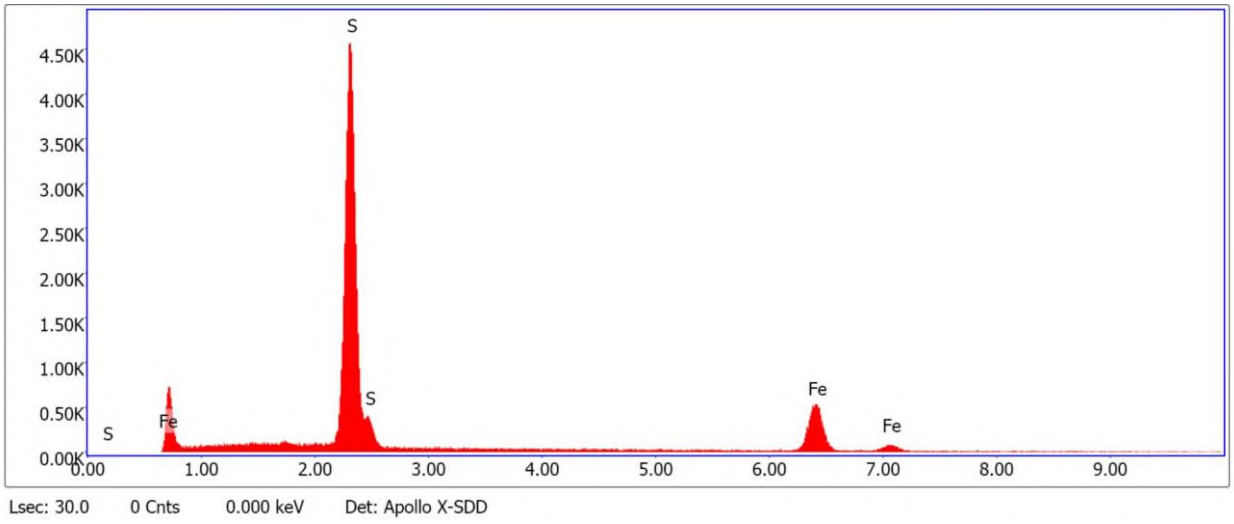
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
S K	43.06	60.65	942.84	3.42	0.3797	0.9938	0.8843	1.0034
ZnK	56.94	39.35	103.88	8.44	0.4998	0.8677	1.0006	1.0109

EDS Spot 2 - EDS1

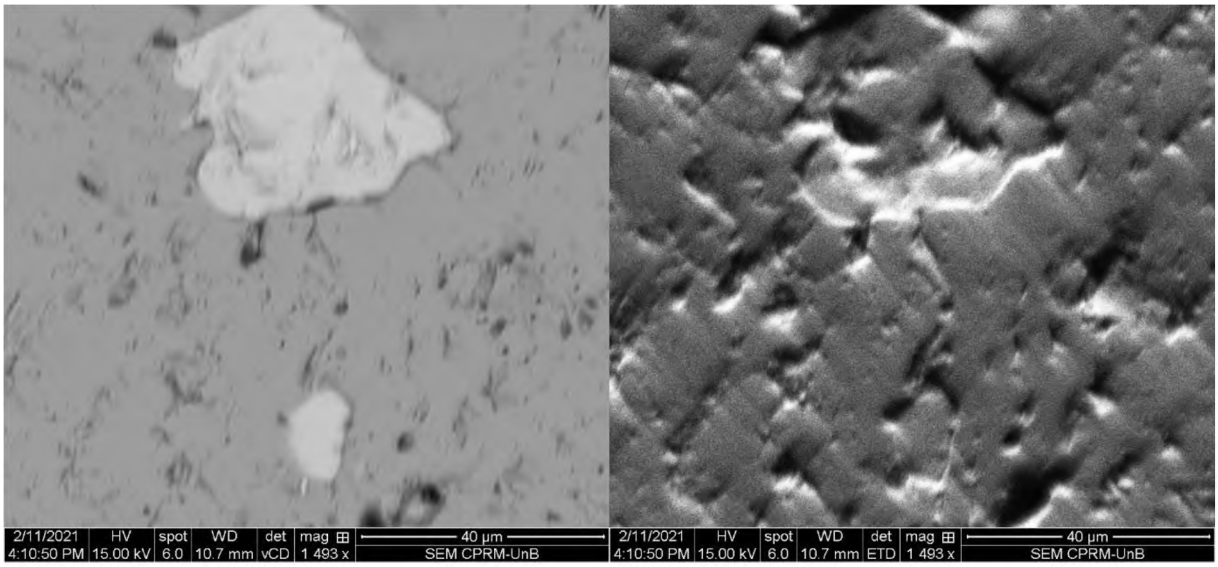


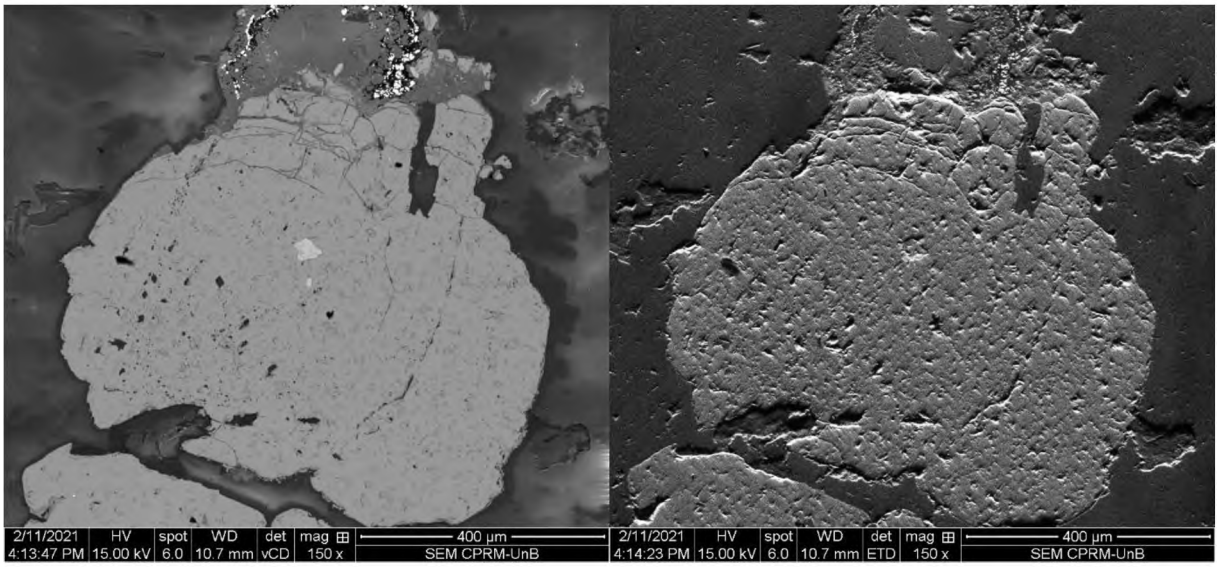
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
S K	38.84	55.74	997.17	3.41	0.3436	0.9975	0.8834	1.0039
FeK	10.07	8.30	64.83	11.50	0.1006	0.8895	0.9944	1.1300
ZnK	51.09	35.97	109.28	7.33	0.4498	0.8719	0.9984	1.0113

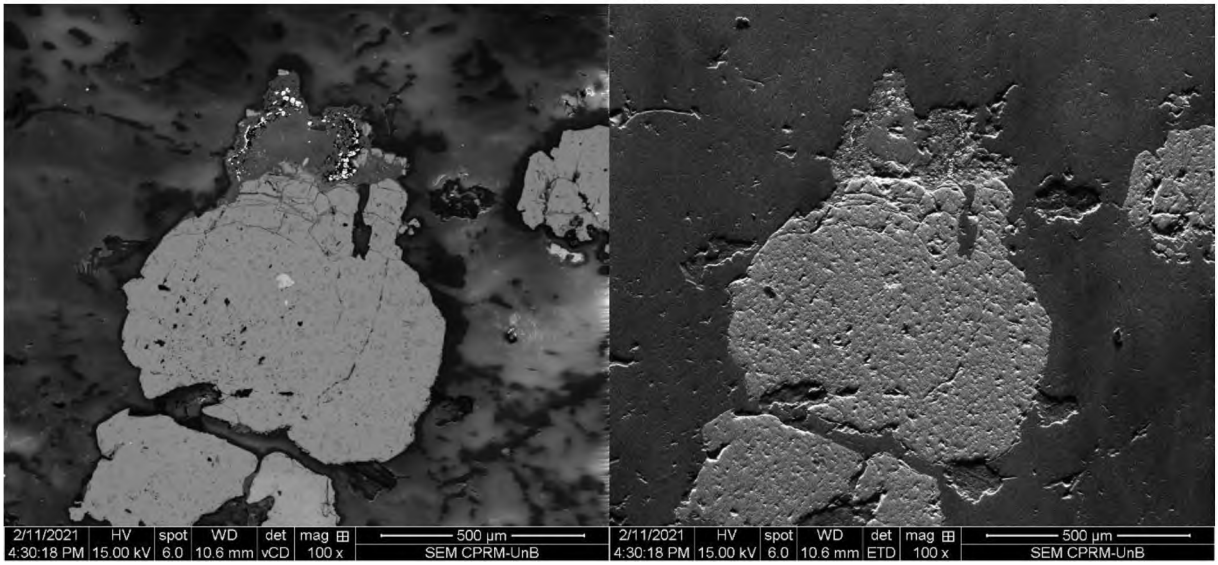
EDS Spot 3 - EDS1



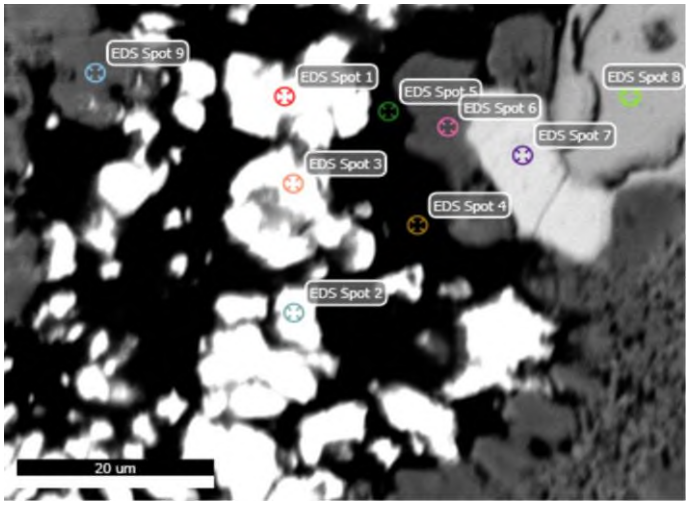
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
S K	56.21	69.10	1667.77	2.44	0.5178	0.9666	0.9481	1.0051
FeK	43.79	30.90	268.71	4.23	0.3760	0.8554	0.9949	1.0089


2/11/2021 | HV | spot | WD | det | mag | 40 µm
4:10:50 PM | 15.00 kV | 6.0 | 10.7 mm | vCD | 1 493 x | SEM CPRM-UnB
2/11/2021 | HV | spot | WD | det | mag | 40 µm
4:10:50 PM | 15.00 kV | 6.0 | 10.7 mm | ETD | 1 493 x | SEM CPRM-UnB

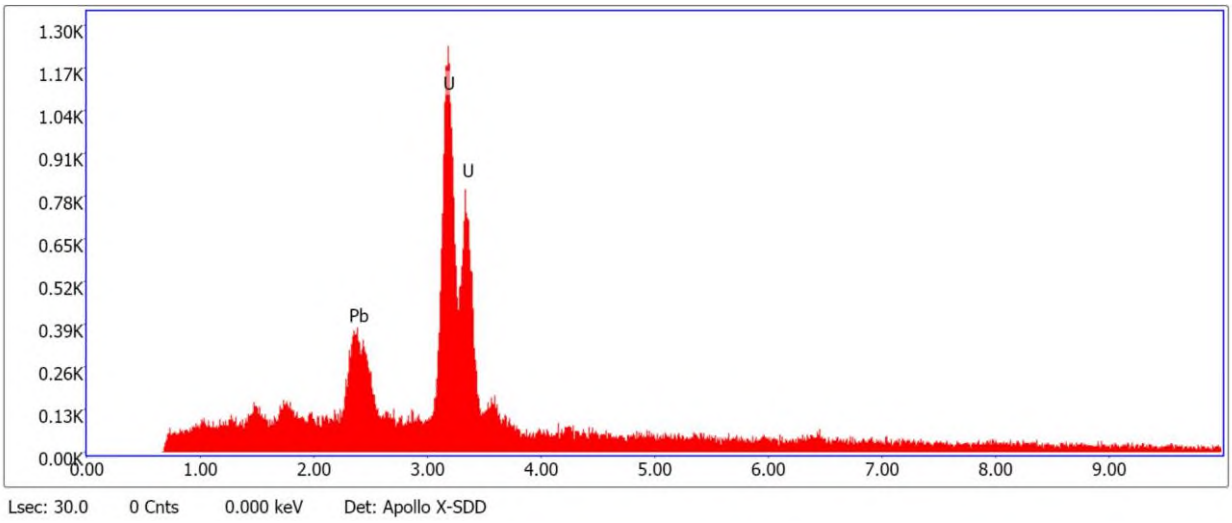

2/11/2021 | HV | spot | WD | det | mag | 400 µm
4:13:47 PM | 15.00 kV | 6.0 | 10.7 mm | vCD | 150 x | SEM CPRM-UnB
2/11/2021 | HV | spot | WD | det | mag | 400 µm
4:14:23 PM | 15.00 kV | 6.0 | 10.7 mm | ETD | 150 x | SEM CPRM-UnB


2/11/2021 | HV | spot | WD | det | mag | 500 µm
4:30:18 PM | 15.00 kV | 6.0 | 10.6 mm | vCD | 100 x | SEM CPRM-UnB
2/11/2021 | HV | spot | WD | det | mag | 500 µm
4:30:18 PM | 15.00 kV | 6.0 | 10.6 mm | ETD | 100 x | SEM CPRM-UnB

LÂMINA HHS-443 – CAMPO 03

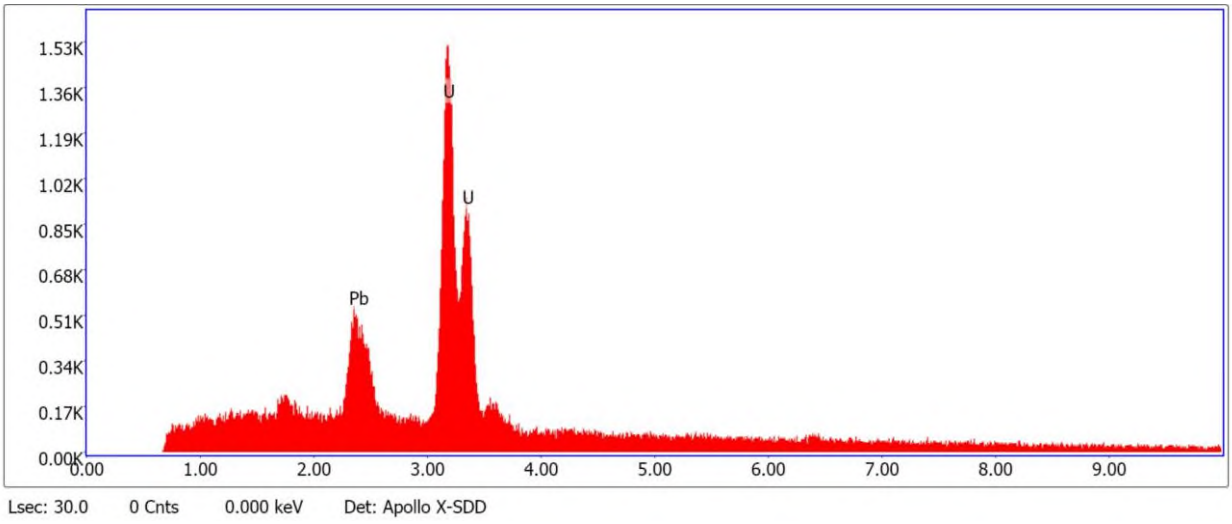


EDS Spot 1 - EDS1



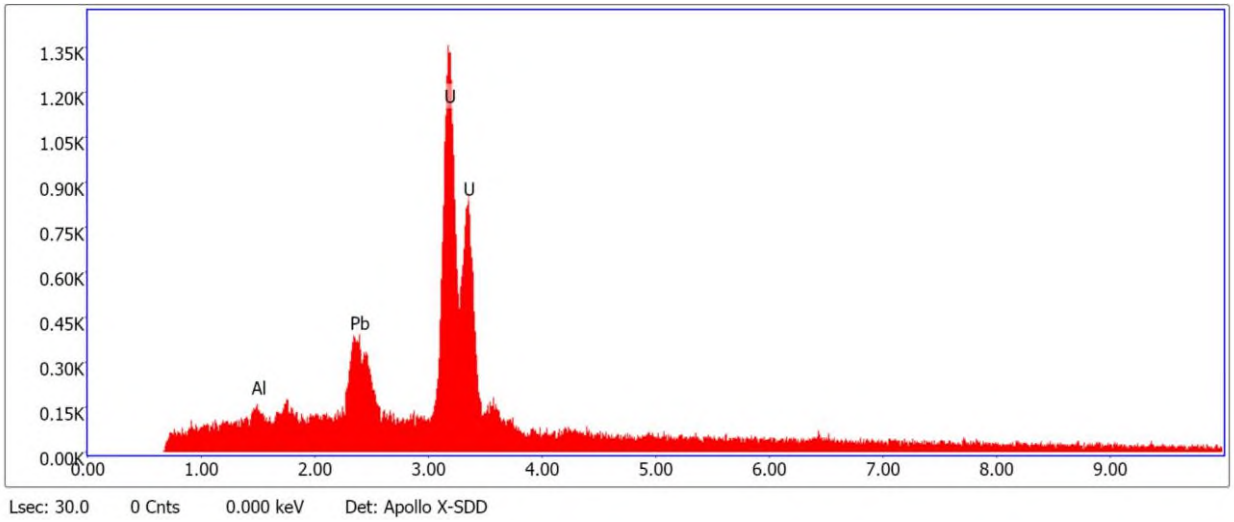
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
PbM	14.25	16.03	97.50	9.47	0.1243	0.8529	1.0181	1.0050
U M	85.75	83.97	456.98	3.53	0.7176	0.8327	1.0060	0.9991

EDS Spot 2 - EDS1



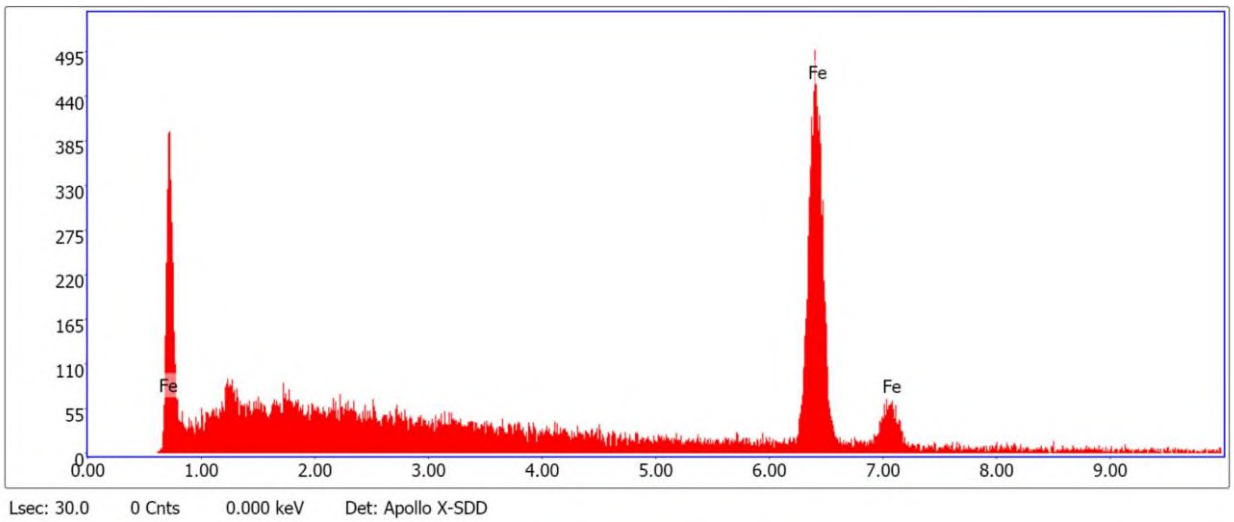
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
PbM	16.28	18.26	141.21	6.68	0.1421	0.8525	1.0188	1.0049
U M	83.72	81.74	562.59	3.28	0.6971	0.8322	1.0016	0.9990

EDS Spot 3 - EDS1



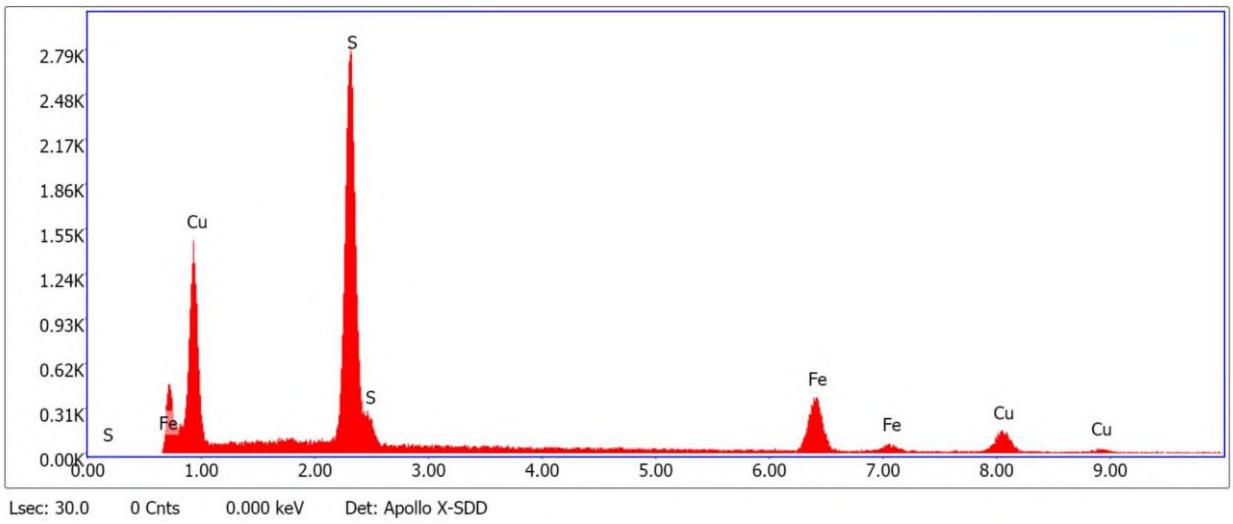
Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
AlK	0.87	7.02	17.67	40.27	0.0069	1.3337	0.5984	0.9998
PbM	13.89	14.66	109.20	8.39	0.1208	0.8497	1.0184	1.0050
U M	85.25	78.32	522.51	3.41	0.7117	0.8294	1.0075	0.9991

EDS Spot 6 - EDS1



Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
FeK	100.00	100.00	232.26	4.21	0.9292	0.9266	1.0028	1.0000

EDS Spot 7 - EDS1



Element	Weight %	Atomic %	Net Int.	Error %	Kratio	Z	A	F
S K	40.13	55.34	1032.16	3.01	0.3638	0.9918	0.9091	1.0054
FeK	31.35	24.81	183.50	5.37	0.2915	0.8822	0.9960	1.0583
CuK	28.52	19.85	78.79	9.48	0.2461	0.8596	0.9904	1.0134