UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA
DEPARTAMENTO DE ENGENHARIA NAVAL E OCEÂNICA

LUCAS ABREU BLANES DE OLIVEIRA

Modelagem geoquímica e mineralógica dos reservatórios carbonáticos do pré-sal da Bacia de Santos através de perfis de poços e inteligência artificial

São Paulo
2022

LUCAS ABREU BLANES DE OLIVEIRA

Modelagem geoquímica e mineralógica dos reservatórios carbonáticos do pré-sal da Bacia de Santos através de perfis de poços e inteligência artificial

Versão Corrigida

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para obtenção de título de Mestre em Ciências.

Área de concentração:
Engenharia Naval e Oceânica

Orientador:
Prof. Dr. Cleyton de Carvalho Carneiro

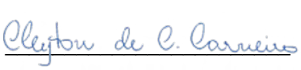
São Paulo

2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 7 de fevereiro de 2022

Assinatura do autor: 

Assinatura do orientador: 

Catalogação-na-publicação

Oliveira, Lucas Abreu Blanes de
Modelagem geoquímica e mineralógica dos reservatórios carbonáticos do pré-sal da Bacia de Santos através de perfis de poços e inteligência artificial / L. A. B. Oliveira -- versão corr. -- São Paulo, 2022.
180 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Naval e Oceânica.

1.Perfilagem de poços 2.Perfis geoquímicos 3.Modelo mineralógico 4.Aprendizado de máquina I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Naval e Oceânica II.t.

Dissertação de autoria de Lucas Abreu Blanes de Oliveira, sob o título **"Modelagem geoquímica e mineralógica dos reservatórios carbonáticos do pré-sal da Bacia de Santos através de perfis de poços e inteligência artificial"**, apresentada à Escola Politécnica da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação do Departamento de Engenharia Naval e Oceânica, na área de concentração de Óleo e Gás Natural, aprovada em ____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr.
Instituição
Presidente

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Prof. Dr.
Instituição

Dedico este trabalho às pessoas que estiveram comigo e me apoiaram ao longo dessa jornada, contribuindo direta ou indiretamente para o meu desenvolvimento como pesquisador científico, profissional e ser humano.

À minha esposa Luísa, por dividir sua vida comigo, me apoiando e incentivando durante todo o período do mestrado. Sua presença em minha vida me dá a força necessária para continuar buscando meu desenvolvimento profissional e pessoal.

Aos meus pais Jomar e Elaine, por fornecerem os alicerces tão essenciais desde a minha infância, sem os quais não seria uma fração do que me tornei. Espero que a finalização dessa etapa seja uma pequena forma de retribuir os incalculáveis sacrifícios feitos para me proporcionar a educação base que me trouxe até aqui.

Aos meus amigos e familiares próximos, os quais não cito nominalmente para não cometer a injustiça de esquecer alguém. Vocês estiveram sempre ao meu lado e, mesmo não atuando diretamente neste trabalho, proporcionaram os momentos leves e descontraídos, tão necessários na manutenção da sanidade e equilíbrio mental.

Agradecimentos

Ao Prof. Dr. Cleyton de Carvalho Carneiro pela orientação ao longo da pesquisa. Sua atuação foi fundamental no meu desenvolvimento como pesquisador.

Aos amigos da Petrobras e da Universidade de São Paulo que estiveram presentes em inúmeras discussões sobre a pesquisa. O privilégio de poder dividir as minhas ideias com vocês foi essencial no amadurecimento do presente trabalho.

À Agência Nacional do Petróleo, Gás Natural e Biocombustíveis pela disponibilidade dos dados e autorização para publicação dos resultados da pesquisa.

Ao Laboratório de Caracterização Tecnológica da Universidade de São Paulo pela realização e disponibilização de parte dos ensaios laboratoriais utilizados.

À Petrobras, pela aquisição e disponibilização dos dados utilizados nesta pesquisa e pelo incentivo ao desenvolvimento profissional.

E por fim, agradeço à Universidade de São Paulo, minha casa, onde tive o privilégio de me graduar como bacharel em geologia e agora como mestre em engenharia.

"Here we stand, feet planted in the earth, but might the cosmos be very near us, only just above our heads?"

(Autor desconhecido)

Resumo

OLIVEIRA, Lucas Abreu Blanes de. **Modelagem geoquímica e mineralógica dos reservatórios carbonáticos do pré-sal da Bacia de Santos através de perfis de poços e inteligência artificial**. 2021. 221 f. Dissertação (Mestrado em Ciências) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2021.

Conhecer a geoquímica e mineralogia das rochas é essencial para a avaliação de formações e caracterização dos reservatórios. Usualmente, modelos geoquímicos e mineralógicos são criados usando os perfis geoquímicos, utilizando as concentrações dos elementos químicos presentes na matriz da rocha para calcular as frações minerais. Entretanto, incertezas observadas na criação desses modelos e a disponibilidade dos perfis geoquímicos e de amostras de rocha em cenários de cortes de gastos torna essa modelagem desafiadora. Nos campos do pré-sal da Bacia de Santos operados pela Petrobras, a aquisição da ferramenta geoquímica e de amostras de rocha é mais frequente na fase exploratória, sendo diminuída nos poços da fase de desenvolvimento. Com essas complexidades, algoritmos de aprendizado de máquina representam uma solução para a criação de modelos geoquímicos e mineralógicos alinhados com o cenário de redução de custos da companhia. Para o modelo geoquímico, uma base de dados foi criada com os perfis de 19 poços. Os dados de entrada foram os perfis de raios gama, espectroscopia de raios gama, densidade, fator fotoelétrico, nêutrons, ressonância magnética nuclear e sônico. Os dados de saída foram as concentrações de Al, Ca, Fe, Mg, Na, Si, S e Ti. O algoritmo de aprendizado de máquina XGBoost foi treinado para gerar perfis geoquímicos sintéticos, e os resultados foram avaliados usando um conjunto de validação e validação cruzada. Com exceção do Na, com R^2 acima de 0,70, os modelos dos demais elementos apresentaram R^2 acima de 0,80. Para o modelo mineralógico, duas metodologias foram criadas. Na primeira, uma base de dados foi criada com as análises de FRX e DRX de 1.376 amostras de rocha coletadas no pré-sal. Os dados de entrada foram as concentrações de Al, Ca, Fe, K, Mg, Na, Si e Ti, e os dados de saída foram as frações de calcita, dolomita, quartzo, K-feldspato, argilas detríticas, plagioclásio e piroxênio. O algoritmo XGBoost foi utilizado através de aprendizado escalonado, melhorando o resultado dos modelos de argilas detríticas, quartzo e calcita quando comparada com técnicas tradicionais. O segundo modelo mineralógico utilizou uma modelagem híbrida, criada através da integração dos algoritmos do aprendizado escalonado com um modelo probabilístico. A etapa probabilística utilizou as informações dos algoritmos de aprendizado de máquina em conjunto com os perfis de densidade, fator fotoelétrico, frações de fluido da ressonância magnética nuclear e geoquímicos, para estimar também as frações de pirita, barita e argilas magnesianas, que não haviam sido contemplados na base de dados. Os modelos geoquímicos e mineralógicos foram aplicados a dados de perfis de poços não utilizados no treinamento e validação, para testar sua qualidade em situações reais. Os modelos foram capazes de honrar os perfis geoquímicos reais e as frações minerais observadas em análises de DRX, confirmando sua robustez e capacidade de generalização. Ficou demonstrado que as metodologias propostas são capazes de gerar modelos geoquímicos e mineralógicos de alta qualidade, alinhado com as iniciativas de otimização e redução de custos.

Palavras-chaves: Perfilagem de poços, Perfis geoquímicos, Modelo mineralógico, Aprendizado de máquina, Inteligência artificial.

Abstract

OLIVEIRA, Lucas Abreu Blanes de. **Geochemical and mineralogical modeling of the pre-salt carbonate reservoirs in the Santos Basin through well logs and artificial intelligence**. 2021. 221 p. Dissertation (Master of Science) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2021.

Knowing the geochemistry and mineralogy of rocks is essential for formation evaluation and reservoir characterization. Usually, geochemical and mineralogical models are created using geochemical logs, using the abundance of chemical elements present in the rock matrix to calculate mineral fractions. However, uncertainties observed when creating these models and the availability of geochemical logs and rock samples in scenarios of cost reduction makes this modeling challenging. In the pre-salt fields in the Santos Basin operated by Petrobras, the acquisition of the geochemical tool and rock samples is more frequent in the exploratory phase, decreasing in wells in the development phase. With these complexities, machine learning algorithms represent a solution for creating geochemical and mineralogical models aligned with the company's cost reduction scenario. For the geochemical model, a database was created with the logs of 19 wells. The input data were gamma-ray, gamma-ray spectroscopy, density, photoelectric factor, neutron porosity, nuclear magnetic resonance, and acoustic logs. The output data were Al, Ca, Fe, Mg, Na, Si, S, and Ti's concentrations. The machine learning algorithm XGBoost was trained to generate synthetic geochemical logs, and the results were evaluated using a validation set and cross-validation. Except for Na, with R^2 above 0.70, the models of the other elements showed R^2 above 0.80. For the mineralogical model, two methodologies were created. In the first one, a database was created with the FRX and DRX analyzes of 1,376 rock samples collected in the pre-salt. The input data were the concentrations of Al, Ca, Fe, K, Mg, Na, Si, and Ti, and the output data were the fractions of calcite, dolomite, quartz, K-feldspar, detrital clays, plagioclase, and pyroxene. The XGBoost algorithm was used through stepped machine learning, improving the result of the models of detrital clays, quartz and calcite when compared to traditional techniques. The second mineralogical model used a hybrid model, created through the integration of the stepped machine learning algorithms with a probabilistic model. The probabilistic phase used the estimates from the machine learning algorithms together with density, photoelectric factor, nuclear magnetic resonance, and geochemical logs to also estimate the fractions of pyrite, barite and magnesian clays, which had not been included in the database. The trained geochemical and mineralogical models were applied to data from well logs not used in training and validation to test their quality in real situations. The models were able to honor the real geochemical logs and mineral fractions observed in XRD analyzes, confirming their robustness and generalization capacity. It was demonstrated that the proposed methodologies can generate high-quality geochemical and mineralogical models in line with optimization and cost reduction initiatives.

Keywords: Wireline logs, Geochemical logs, Mineralogical model, Machine learning, Artificial intelligence.

Lista de figuras

Lista de algoritmos

Lista de tabelas

Lista de abreviaturas e siglas

AdaBoost	*Adaptive boosting*
API	*American petroleum institute*
Bar	Barita
BVI	*Bound volume irreducible*
Cal	Calcita
Carb	Carbonatos (calcita + dolomita)
CBW	*Clay bound water*
Clay	Argila detrítica
DEN	Perfil de densidade
DesvPad	Desvio padrão
Dol	Dolomita
DRX	Difratometria de raios X
DTP	Perfil da vagarosidade da onda P
DTS	Perfil da vagarosidade da onda S
EQ	Elementos químicos
EQM	Erro quadrático médio
EQA	Erro quadrático absoluto
FF	Fluido livre (*free fluid*)
FRX	Fluorescência de raios X
GR	Perfil de raios gama
Kfd	K-feldspato
LCT	Laboratório de caracterização tecnológica

LWD	*Logging while drilling*
Mgclay	Argila magnesiana
MIN	*Minerais*
MLP	*Multilayer perceptron*
NEU	Perfil de nêutrons
Onda P	Onda compressional
Onda S	Onda cisalhante
PEF	Perfil de fator fotoelétrico
PF	Perda ao fogo
PhiT	Porosidade total
PhiE	Porosidade efetiva
Pir	Pirita
Plg	Plagioclásio
POL	Polegadas
Prx	Piroxênio
Qtz	Quartzo
R^2	Coeficiente de determinação
RMN	Ressonância magnética nuclear
RNA	Redes neurais artificiais
RNR	Redes neurais recorrentes
SVM	*Support-vector machines*
Sw	Saturação de água
T_2	Tempo de relaxação

T_{2B}	Tempo de relaxação *bulk*
T_{2D}	Tempo de relaxação por difusão
T_{2S}	Tempo de relaxação de superfície
T2LM	Média geométrica da distribuição de T_2
U	Índice de absorção fotoelétrica volumétrica
UCS	*Unconfined compressive strength*
Wat	Água capilar
XGBoost	Extreme gradient boosting

Lista de símbolos

Φ	Fluxo de radiação
ρ_b	Densidade *bulk*
Z	Número atômico
A	Número de massa
N	Número de átomos por volume
σ	Seção de choque
ϕ	Porosidade
ρ_f	Densidade do fluido
ρ_{ma}	Densidade da matriz
P_e	Fator fotoelétrico
E	Energia
Σ	Seção de choque macroscópica
N_{Av}	Número de Avogadro
L_s	Distância de desaceleração
L_d	Distância de difusão
L_m	Distância de migração
Y	Coeficientes (*yields*) do espectro de referência de um elemento
S	Sensibilidade de detecção de um elemento
X	Razão entre a massa do óxido/carbonato e a massa de um elemento
F	Fator de normalização
w	Concentração mássica absoluta de um elemento
Δt	Vagarosidade

Δt_f	Vagarosidade do fluido
Δt_{ma}	Vagarosidade da matriz
B_0	Campo magnético constante
B_1	Campo magnético de pulso de radio frequência
$M(t)$	Magnetização em função do tempo
$A(T_2)$	Amplitude do tempo de relaxação
$\frac{S}{V}$	Razão entre superfície e volume dos poros
f	Perfil reconstruído
c	Número de componentes
V	Fração volumétrica de um componente
e	Valor de referência de um componente para uma ferramenta
k	Índice referente ao perfil reconstruído
l	Índice referente ao componente
p	Perfil adquirido
t	Número de equações de reconstrução
ϵ	Desvio padrão associado a incerteza atribuída a um perfil
y	Valor de uma instância
n	Total de instâncias
m	Iteração
M	Total de iterações
$f(x)$	Árvore de decisão
r	Resíduo
j	Número de folhas da árvore de regressão

R	Folha da árvore de decisão
γ	Valor da folha da árvore de decisão
U	Índice de absorção fotoelétrica volumétrica de um componente
U_T	Índice de absorção fotoelétrica volumétrica da formação
ϕ_T	Porosidade total
ϕ_{CBW}	Porosidade de argila
ϕ_{FF}	Fluido livre (*free fluid*)
W	Fração mássica absoluta de um mineral
w_T	Concentração mássica total de um elemento na matriz da rocha

Sumário

1 Introdução

Na indústria de óleo e gás, o modelo geoquímico e mineralógico representa as bases da caracterização petrofísica dos reservatórios e avaliação de formações. Quando confiáveis, esses modelos podem melhorar significativamente os cálculos de porosidade, saturação de hidrocarbonetos e volume de argila (FREEDMAN *et al.*, 2015). Além da aplicação direta na avaliação de formações, o conhecimento da composição geoquímica e mineralógica podem auxiliar em operações de acidificação (JIN *et al.*, 2019) e no monitoramento da variação do contato hidrocarboneto/água durante a produção de um reservatório (WESTAWAY; HERTZOG; PLASEK, 1983; NORTH, 1987; ULLOA *et al.*, 2016).

A modelagem geoquímica dos reservatórios consiste na determinação da concentração dos elementos químicos que compõem a matriz da rocha. Essas concentrações são obtidas pela ferramenta geoquímica, através da radiação gama proveniente da interação dos nêutrons emitidos pela ferramenta e a formação (ELLIS; SINGER, 2007). Já a modelagem mineralógica consiste na determinação das frações minerais presentes na matriz da rocha. Diferente da modelagem geoquímica, não existe uma ferramenta de perfilagem capaz de detectar esses minerais, demandando diferentes técnicas para sua quantificação.

As frações minerais em escala de poço podem ser estimadas a partir de duas principais metodologias. Uma abordagem tradicional, denominada de “modelo probabilístico”, consiste na definição de um sistema de equações em que os valores correspondentes a cada perfil de poço são aproximados pela combinação das respostas individuais dos diferentes componentes – minerais e fluidos – constituintes da formação, em função de suas respectivas frações volumétricas (MITCHELL; NELSON, 1988). Outra abordagem amplamente empregada, denominada de “modelo direto”, consiste no cálculo das frações minerais a partir das concentrações de elementos químicos adquiridas pela ferramenta geoquímica (HERRON; HERRON, 1996), calibrados com as concentrações de elementos químicos obtidas por fluorescência de raios X (FRX) e as frações minerais obtidas por difratometria de raios X (DRX) adquiridas em amostras de rochas (HERRON *et al.*, 2014).

No entanto, essas metodologias possuem limitações. Os modelos probabilísticos restringem a quantidade máxima de componentes ao número de perfis de entrada mais um. Como esses modelos precisam considerar a porosidade e os fluidos presentes na formação, o número de minerais utilizados pode ficar restrito, impactando sua aplicação

às formações de mineralogia complexa. O uso das concentrações de elementos químicos em modelos diretos requer a determinação precisa da composição química dos minerais, que podem variar em função das condições geológicas. Além disso, esses modelos são altamente dependentes da qualidade dos teores elementares obtidos a partir da ferramenta geoquímica.

As limitações apresentadas são particularmente significativas na criação de um modelo mineralógico para as rochas do pré-sal da Bacia de Santos. Além da complexidade de fácies observada nos reservatórios carbonáticos (GOMES *et al.*, 2020), a presença de rochas siliciclásticas e ígneas (MOREIRA *et al.*, 2007) gera uma assembleia mineral desafiadora, tanto para modelos diretos quanto probabilísticos. A representatividade de alguns minerais em modelos diretos é impactada pelo baixo número de amostras de rocha com frações expressivas desses minerais, uma vez que a coleta de amostras é limitada e concentrada em intervalos reservatórios de boa permo-porosidade, constituídos majoritariamente de carbonatos. Além disso, as argilas magnesianas encontradas apresentam composição química peculiar, dificultando sua estimativa através de abordagens tradicionais (HERLINGER *et al.*, 2020). As argilas magnesianas podem estar associadas tanto a porosidade secundária, através da sua dissolução, quanto a barreiras ao fluxo, tornando sua quantificação de extrema importância para a caracterização de reservatórios.

Além das complexidades observadas na criação de um modelo mineralógico, outros desafios relacionados à sua utilização em poços podem ser encontrados. Os modelos diretos demandam que a ferramenta geoquímica tenha sido adquirida durante a perfilagem. Entretanto, isso pode não acontecer em todos os poços. A busca por otimização e maior rentabilidade em projetos de desenvolvimento de um campo gera redução de custos, impactando a quantidade dos perfis adquiridos. Nos campos do pré-sal da Bacia de Santos operados pela Petrobras, a aquisição da ferramenta geoquímica é restrita apenas a fase exploratória. Durante a fase de desenvolvimento dos campos, a quantidade de perfis de poço é reduzida e a ferramenta geoquímica é excluída das operações de perfilagem.

Sob a ótica das complexidades apresentadas, a inteligência artificial representa uma solução para a criação de modelos geoquímicos e mineralógicos. Os poços que possuem os perfis geoquímicos podem ser usados no treinamento de algoritmos de aprendizado de máquina. Esses algoritmos seriam então aplicados nos demais poços, gerando perfis geoquímicos sintéticos que abrangessem todo o reservatório. Adicionalmente, algoritmos de aprendizado de máquina podem ser treinados com os dados obtidos das análises de FRX e

DRX para a criação de um modelo mineralógico. A principal vantagem é a independência de conhecimento prévio quanto à composição química dos minerais presentes na rocha, reduzindo a subjetividade interpretativa. Em poços com mineralogia complexa e minerais pouco representados pelas análises de FRX e DRX, abordagens híbridas que unam modelos probabilísticos com o aprendizado de máquinas podem ser uma alternativa para lidar com o enviesamento da base de dados. Com isso, seria possível obter um modelo geoquímico e mineralógico confiável de um reservatório de hidrocarbonetos, alinhado com as políticas de redução de custos dos projetos.

1.1 Motivação e objetivos

O principal objetivo desse trabalho é criar modelos geoquímicos e mineralógicos confiáveis para as rochas do pré-sal da Bacia de Santos. Os modelos visam um cenário de otimização e redução de custos de perfilagem e coleta de amostras de rocha, com base em rotinas de inteligência artificial aplicadas em dados de perfis de poços. Os objetivos específicos são:

1. Desenvolver um modelo de aprendizado de máquina para a geração de perfis geoquímicos sintéticos. Esses perfis serão usados na modelagem geoquímica;
2. Estruturar uma rotina sistemática baseada em aprendizado de máquina capaz de produzir modelos mineralógicos através da integração de análises por FRX e DRX em amostras de rochas;
3. Desenvolver um modelo mineralógico específico para os carbonatos do pré-sal da Bacia de Santos, tendo como base a rotina sistemática de aprendizado de máquina que se utiliza dos dados de FRX e DRX;
4. Desenvolver um modelo mineralógico híbrido, reunindo o modelo de aprendizado de máquina e as informações obtidas por perfis de poços;
5. Avaliar a aplicação dos modelos propostos em poços perfurados no pré-sal da Bacia de Santos.

1.2 Organização do texto

O trabalho está estruturado da seguinte forma:

1. **Capítulo 2 - Fundamentos teóricos**: é apresentada uma revisão da geologia dos carbonatos do pré-sal, dos princípios físicos e de aquisição das ferramentas de perfilagem, das análises de FRX e DRX, do algoritmo de aprendizado de máquina e de trabalhos que aplicam inteligência artificial a perfis de poço.
2. **Capítulo 3 - Metodologia**: são descritas as metodologias empregadas na criação dos modelos geoquímico e mineralógico.
3. **Capítulo 4 - Resultados**: são apresentados os resultados obtidos, com foco na qualidade e representatividade dos modelos gerados e na aplicação em cenários práticos.
4. **Capítulo 5 - Discussão**: são discutidos as melhorias observadas nos modelos geoquímicos e mineralógicos frente às demais técnicas.
5. **Capítulo 6 - Conclusão**: são apresentadas as conclusões do presente trabalho.
6. **Anexos**: são apresentados três artigos publicados com resultados desta pesquisa de mestrado.

2 Fundamentos teóricos

2.1 Carbonatos do pré-sal da Bacia de Santos e seus desafios tecnológicos

O pré-sal brasileiro pode ser considerado uma das descobertas de hidrocarbonetos mais relevantes das últimas décadas (BELTRAO *et al.*, 2009). Localizado na Bacia de Santos (figura 1), o pré-sal consiste nas rochas encontradas estratigraficamente abaixo da camada de evaporitos da Formação Ariri, de idade Neoaptiana (MOREIRA *et al.*, 2007). Essa sequência compreende os estromatólitos, laminitos microbiais, microbiolitos ricos em argilas magnesianas e folhelhos carbonáticos da Formação Barra Velha, os grainstones, wackestones e packstones da Formação Itapema, rochas siliciclásticas da Formação Piçarras e o embasamento cristalino da Formação Camboriú. Basaltos e diabásios podem aparecer intercalados ao longo de toda a sequência.

Figura 1 – Localização dos reservatórios carbonáticos do pré-sal da Bacia de Santos

As Formações Barra Velha, Itapema e Piçarras foram depositadas entre o Hauteriviano e o Eoaptiano (entre 110 e 130 milhões de anos), e o embasamento cristalino possui idade pré-cambriana. A figura 2 apresenta a carta estratigráfica da Bacia de Santos.

Argilas magnesianas são encontradas principalmente na Formação Barra Velha e representam uma mistura de kerolita, Mg-esmectita e sepiolita (HERLINGER *et al.*, 2020). Essas argilas podem estar associadas tanto a porosidade secundária, por conta da sua dissolução, quanto a barreiras ao fluxo, formando acumulações que podem chegar a mais de 200 metros de espessura, sendo consideradas peça-chave no entendimento do fluxo de fluidos dentro do reservatório. As argilas magnesianas são importantes também para os estudos de sedimentologia e estratigrafia dos carbonatos, com estudos recentes propondo seu uso como um dos três principais componentes para a classificação de fácies carbonáticas do pré-sal (GOMES *et al.*, 2020).

A composição química peculiar das argilas magnesianas dificulta sua quantificação através de metodologias tradicionais. A quase ausência de elementos radioativos como o potássio impossibilita o uso da radiação gama (ELLIS; SINGER, 2007) e os padrões de densidade e nêutrons não apresentam uma clara distinção, como a observada em rochas siliciclásticas (HERLINGER *et al.*, 2020). Apesar dos perfis acústicos e de ressonância magnética nuclear serem capazes de identificar as argilas da formação, eles não são usados de forma quantitativa para estimar a fração de argilas magnesianas.

O contexto dos reservatórios do pré-sal apresenta alta complexidade tecnológica, em um cenário de perfuração de poços de mais de 5.000 m de profundidade em lâminas d'água de mais de 2.000 m (BELTRAO *et al.*, 2009). Isso induz os especialistas a utilizarem todo e qualquer recurso necessário para minimizar riscos durante o gerenciamento dos reservatórios. Com isso, os perfis de poço se destacam como ferramentas importantes na avaliação de formações. Ferramentas de perfilagem adquirem informações que podem ser usadas para calcular porosidade, saturação de hidrocarbonetos e a composição da matriz da rocha.

Mesmo nesse cenário desafiador, a busca por otimização e redução de custos faz com que certas ferramentas deixem de ser adquiridas a medida que um campo passa da fase exploratória para a fase de desenvolvimento. A tabela 1 apresenta uma comparação entre os perfis adquiridos em uma perfilagem completa e uma perfilagem reduzida realizadas em poços perfurados no pré-sal e operados pela Petrobras. A quantidade de corridas reduz

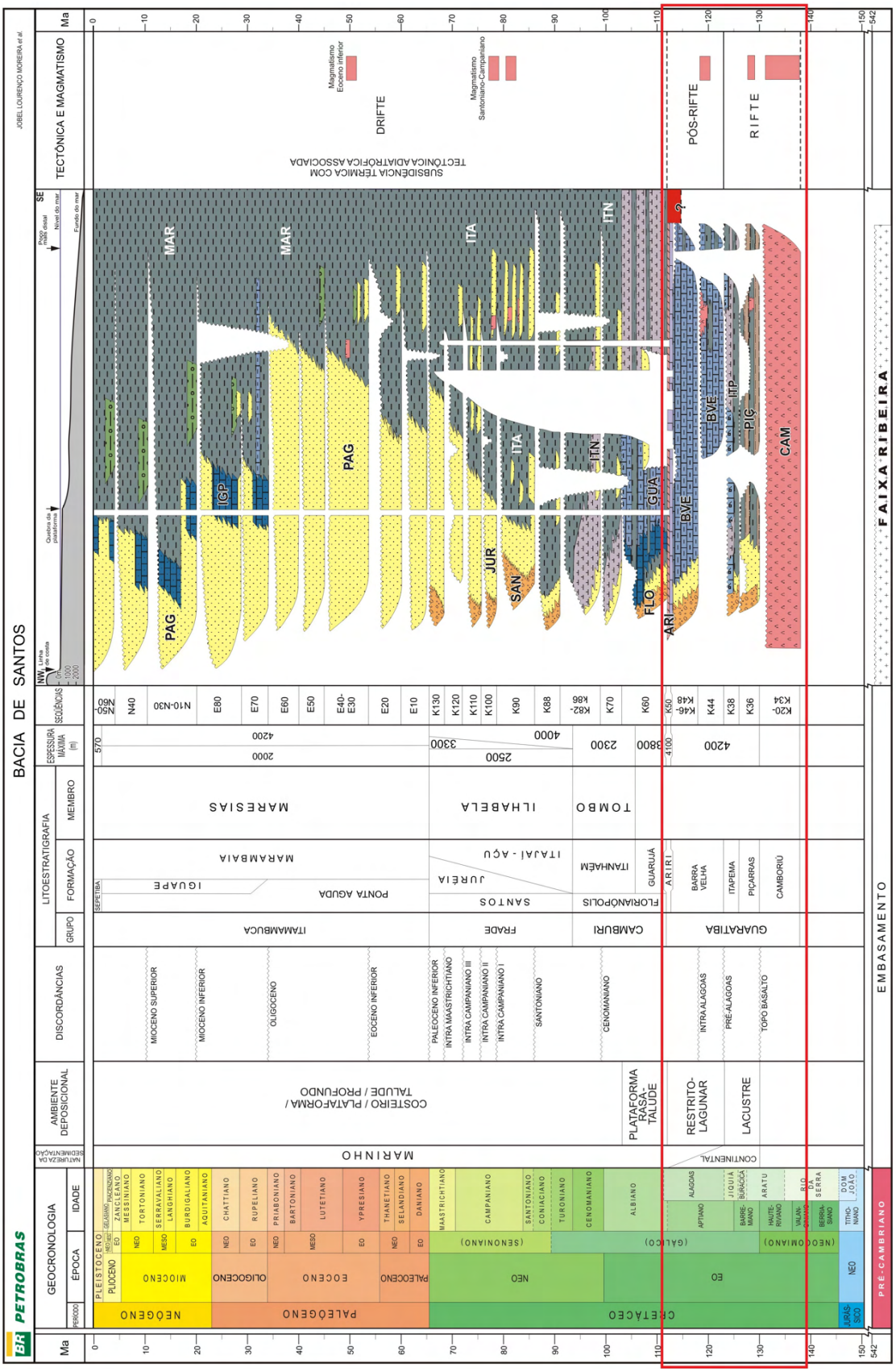


Figura 2 – Carta estratigráfica da Bacia de Santos. O destaque em vermelho demarca a seção que compreende o pré-sal.

Fonte – Moreira *et al.* (2007)

de seis para duas e o tempo total de perfilagem diminui em dez dias, equivalente a uma economia de milhões de dólares por poço.

Tabela 1 – Comparação entre perfilagens completas e reduzidas realizadas nos campos do pré-sal da Bacia de Santos operados pela Petrobras.

Perfilagem completa	
Corrida 1	Raios gama Resistividade Densidade e fator fotoelétrico Porosidade de nêutrons
Corrida 2	Ressonância magnética nuclear Geoquímico Raios gama espectral
Corrida 3	Amostragem de fluido e pressão da formação
Corrida 4	Amostragem de rocha
Corrida 5	Imagem acústica Imagem resistiva Perfis acústicos
Corrida 6	Perfil sísmico vertical
Tempo total de operação - 10 a 12 dias	

Perfilagem reduzida	
Corrida 1	Raios gama Resistividade Densidade e fator fotoelétrico Porosidade de nêutrons Ressonância magnética nuclear Raios gama espectral
Corrida 2	Imagem acústica Perfis acústicos Pressão de formação
Tempo total de operação - 2 a 3 dias	

Fonte – Lucas Oliveira, 2021

A economia proporcionada pela substituição da perfilagem completa pela reduzida obriga que a ferramenta geoquímica não seja mais adquirida. Conforme apresentado na tabela 1, a remoção dessa ferramenta da corrida 2 em uma perfilagem completa permite a combinação das corridas 1 e 2, reduzindo o número total de corridas na perfilagem reduzida. Os perfis geoquímicos permitem a detecção dos elementos químicos presentes na matriz da rocha (ELLIS; SINGER, 2007), que podem ser usados para a criação de um modelo mineralógico da formação, importante no cálculo de porosidade, refinamento estratigráfico, detecção de diagênese e útil nas operações de acidificação de poços.

Também é possível observar na tabela 1 que a amostragem de rocha presente na corrida 4 da perfilagem completa não ocorre na perfilagem reduzida. Como a quantidade de poços com perfilagem reduzida é maior do que os com perfilagem completa, a representatividade da base de dados de análises laboratoriais realizadas em amostras de rocha fica limitada a apenas alguns poços. Isso se torna um desafio em reservatórios heterogêneos como os do pré-sal da Bacia de Santos, uma vez que técnicas de avaliação dependentes de amostras de rocha pode não representar a complexidade das formações encontradas em todos os poços. Isso motiva o desenvolvimento de metodologias capazes de lidar com o enviesamento da base de dados frente às políticas de otimização e redução de custos de projetos de desenvolvimento de reservatórios.

2.2 Perfis de poços

Segundo Serra (1984), perfis de poço são registros de quaisquer características da formação rochosa tomadas em relação a profundidade, utilizando uma ferramenta que faz as medidas conforme ela viaja por dentro do poço. Os perfis de poço podem ser adquiridos após o término da perfuração do poço, utilizando ferramentas de perfilagem a cabo, ou durante a perfuração (*logging while drilling*, LWD), utilizando ferramentas que se conectam a coluna de perfuração.

De maneira geral, as medidas adquiridas pelas ferramentas de perfilagem consistem na emissão de algum tipo de sinal ou perturbação na formação por um transmissor, e posterior leitura da resposta desse sinal ou perturbação por um receptor após ele interagir com a formação. A diferença entre o que foi emitido e o que foi lido permite estimar alguma propriedade da formação, como densidade ou índice de hidrogênio, que terá correlação com alguma propriedade petrofísica de interesse.

2.2.1 Perfis de raios gama

Os perfis de raios gama são adquiridos utilizando ferramentas que fazem a leitura da radiação natural da formação utilizando um cintilômetro de NaI(Tl) (SERRA, 1984). Ao contrário das demais ferramentas, ela não emite qualquer sinal para a formação e apenas lê a radiação emitida. A principal fonte de radiação natural encontrada em rochas são os

isótopos de K, Th e U, elementos presentes principalmente em formações argilosas (ELLIS; SINGER, 2007). Por conta disso, em ambientes siliciclásticos, os perfis de raios gama são muito utilizados na identificação de rochas reservatório e não reservatório. Em carbonatos, a radiação gama está mais relacionada à quantidade de matéria orgânica.

Como a radiação é impactada pelo tipo de detector usado e pela geometria da ferramenta, alguns padrões de calibração foram estipulados pelo *American Petroleum Institute* (API). Essa calibração utiliza uma rocha artificialmente construída pela Universidade de Houston, que simula um folhelho contendo aproximadamente 4% de K, 24 ppm de Th e 12 ppm de U, referente a 200 unidades API (ELLIS; SINGER, 2007). Sendo assim, a resposta da radiação gama em API pode ser dada pela equação 1.

$$GR_{API} = \alpha^{238}U_{ppm} + \beta^{232}Th_{ppm} + \gamma^{39}K_{\%} \tag{1}$$

Onde os coeficientes α, β e γ são calibrados para diferentes ferramentas.

A radiação gama emitida pela formação sofre atenuação a medida que ela viaja pelo interior do poço e atinge o detector. A intensidade da atenuação é relacionada a distância da parede do poço e do receptor, ou seja, o diâmetro do poço, e a densidade do fluido de perfuração. Logo, correções ambientais visam corrigir o perfil de raios gama levando em consideração o diâmetro do poço e a densidade do fluido de perfuração (SCHLUMBERGER, 2009). Outras correções podem ser feitas para levar em consideração a radioatividade do fluido de perfuração, caso ele possua algum material radioativo em sua composição.

Ferramentas modernas são capazes de adquirir um espectro da radiação gama natural proveniente da formação (figura 3). Com esse espectro, é possível definir janelas de energia e calcular as quantidades relativas de K, Th e U presentes na rocha. Isso permite melhorar a caracterização litológica da formação, diferenciando folhelhos a partir da composição e permitindo a detecção de arenitos arcóseos.

2.2.2 Perfis de densidade e fator fotoelétrico

A ferramenta que adquire os perfis de densidade e fator fotoelétrico utiliza uma fonte de radiação gama (tipicamente ^{137}Cs) e dois detectores de NaI(Tl), um mais próximo e outro mais distante da fonte radioativa (ELLIS; SINGER, 2007). A radiação gama emitida

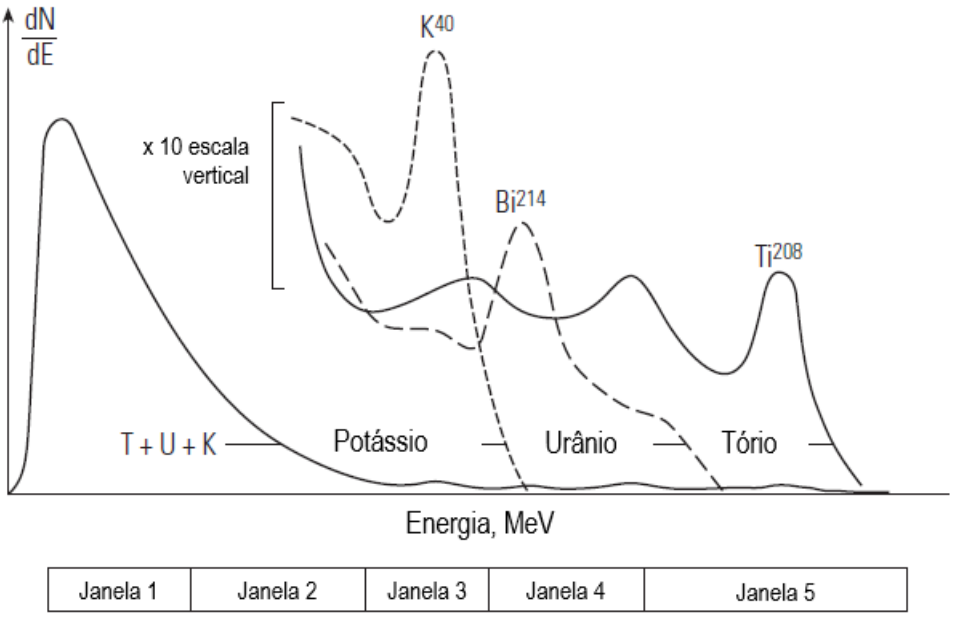


Figura 3 – Espectro de radiação gama e as respectivas janelas relacionadas ao K, Th e U.

Fonte – Ellis e Singer (2007)

pela ferramenta interage com a formação e é atenuada pelo espalhamento Compton, que apresenta correlação com a densidade *bulk* da formação, conforme a equação 2.

$$\Phi = \Phi_0 e^{-\rho_b \frac{Z}{A} N_0 \sigma x} \tag{2}$$

Onde Φ é o fluxo de radiação, $\rho_b \frac{Z}{A} N_0$ é o número de densidade de elétrons de um material de densidade ρ_b, σ é a seção de choque do espalhamento Compton e x é a espessura do material. A razão $\frac{Z}{A}$ é igual a aproximadamente 0.5 para a maioria dos elementos químicos. Como x é a distância entre a fonte radioativa e o detector da ferramenta, a densidade da formação pode ser facilmente calculada. Normalmente, a densidade é fornecida em g/cm^3.

A ferramenta de densidade é adquirida junto a parede do poço (figura 4). Isso visa minimizar distorções causadas pela presença de fluido de perfuração, caso o poço esteja arrombado, ou reboco. Mesmo assim, é possível que as leituras ainda sofram com esses fatores externos. A presença de dois detectores na ferramenta permite a correção da leitura de densidade nesses cenários. Em uma situação ideal, o valor de densidade lido pelos dois receptores seria o mesmo. Caso haja algum arrombamento ou reboco espesso, a leitura do detector próximo será mais afetada do que a do detector distante, já que a quantidade de

formação lida por ele é menor. Logo, a relação entre as leituras dos dois receptores gera um fator de correção que fornece a densidade real da formação.

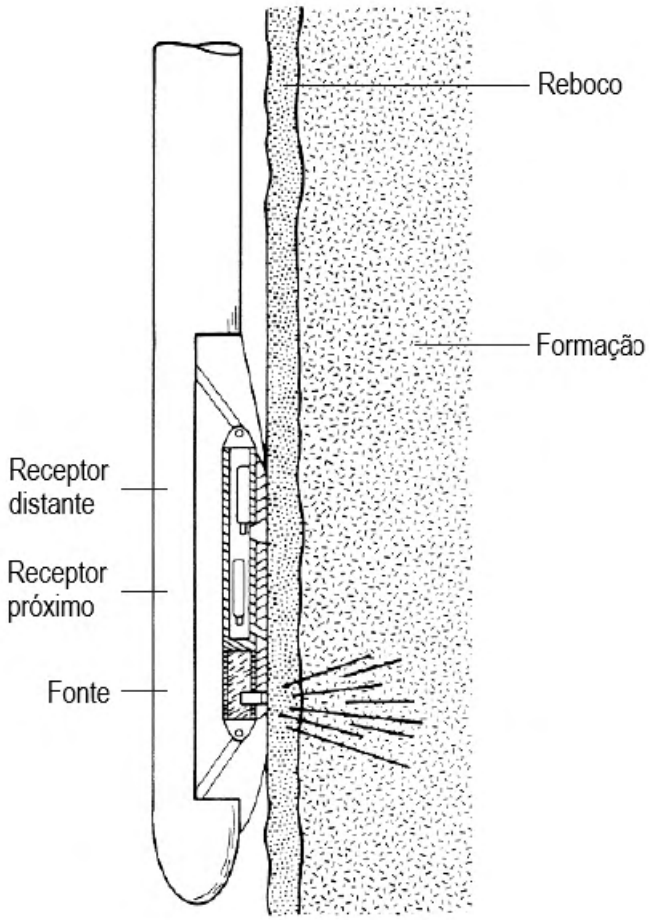


Figura 4 – Exemplo de ferramenta de aquisição do perfil de densidade, sendo adquirida junto a parede do poço. A radiação emitida pela fonte é afetada pelo reboco, afetando a leitura de densidade final. A relação entre as densidades medidas pelos receptores próximo e distante corrige essa distorção.

Fonte – Ellis e Singer (2007)

O principal uso do perfil de densidade é estimar a porosidade da formação através da equação 3.

$$\rho_b = \phi\rho_f + (1-\phi)\rho_{ma} \tag{3}$$

Onde ϕ é a porosidade, ρ_f é a densidade do fluido presente nos poros e ρ_{ma} é a densidade da matriz da rocha. As densidades da matriz e do fluido podem ser estimadas usando outros perfis de poço.

A equação 2 é válida somente para a radiação gama de alta energia. Em baixas energias, o efeito da absorção fotoelétrica também é observado (ELLIS; SINGER, 2007). A absorção fotoelétrica é proporcional ao número atômico Z do material que interage com a radiação (equação 4). Logo, a equação 2 pode ser reescrita como a equação 5.

$$P_e \equiv \frac{Z}{10}^{3.6} \tag{4}$$

$$\Phi = \Phi_0 e^{-N_0 \rho_b (a(E) P_e + b(E)) x} \tag{5}$$

Onde E é a energia da radiação gama e a e b são coeficientes conhecidos e praticamente constantes. Assumindo que a densidade ρ_b pode ser obtida na região de alta energia do espectro de radiação, o fator fotoelétrico P_e pode ser estimado. Ferramentas de densidade modernas permitem a leitura da radiação gama em altas e baixas energias (figura 5), entregando também o perfil de fator fotoelétrico.

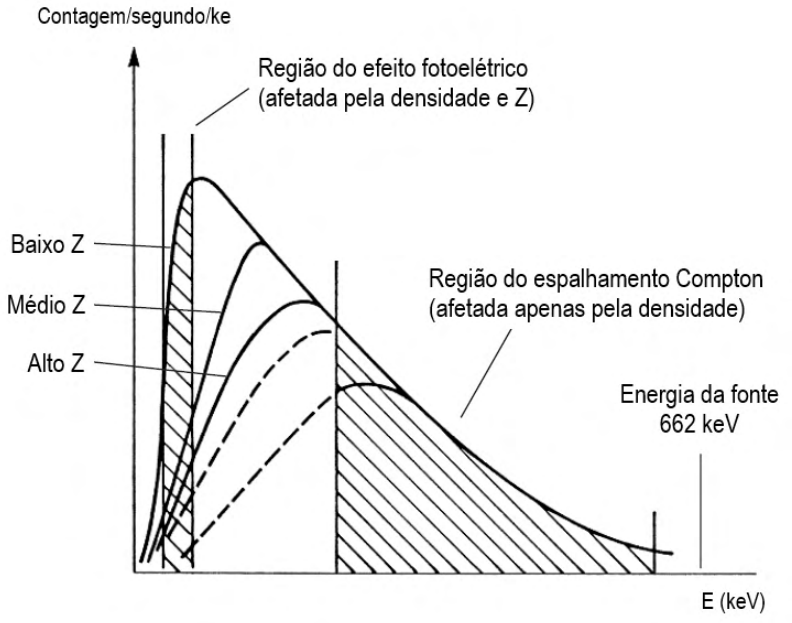


Figura 5 – Representação esquemática do espectro de raios gama lido em três formações compostas por átomos de diferentes números atômicos. A atenuação causada pelo espalhamento Compton é similar, uma vez que a densidade do material pode ser igual. Já a região de baixa energia apresenta diferentes graus de atenuação, uma vez que essa região é afetada pelo efeito fotoelétrico.

Fonte – Ellis e Singer (2007)

Por ter relação com o número atômico dos átomos que compõem a formação e ser pouco sensível a variações de porosidade, o perfil de fator fotoelétrico permite estimar a composição da matriz da rocha (ELLIS; SINGER, 2007). O fator fotoelétrico comumente encontrado em rochas varia entre 1 e 6, com o quartzo apresentando 1,83, a dolomita 3,1 e a calcita 5,1. Materiais como a barita, muito utilizada em fluidos de perfuração, podem apresentar fator fotoelétrico superiores a 100, gerando valores anômalos.

Em análises quantitativas, o fator fotoelétrico normalmente é multiplicado pela densidade *bulk*, gerando o parâmetro U, um índice de absorção fotoelétrica volumétrica (ELLIS; SINGER, 2007). Ele especifica a absorção radioativa de um determinado volume de material, e é medido em seção de choque por cm^3.

2.2.3 Perfis de nêutrons

O princípio de aquisição dos perfis de nêutrons é similar ao do perfil de densidade. Porém, ao invés de usar as interações entre radiação gama e formação, a ferramenta utiliza as interações entre nêutrons e formação. Os nêutrons podem ser emitidos por uma fonte química de AmBe ou por uma fonte eletrônica de D-T (ELLIS; SINGER, 2007). Assim como na ferramenta de densidade, um detector próximo e um distante são utilizados e a ferramenta é adquirida junta a parede do poço.

As fontes emitem nêutrons rápidos de alta energia e, à medida que eles interagem com a formação, vão sendo desacelerados até atingirem as faixas epitermal ou termal de energia (figura 6). As três principais interações responsáveis por essa desaceleração são: espalhamento elástico, espalhamento inelástico e captura radioativa.

O espalhamento elástico é mais eficiente quando os nêutrons interagem com elementos de massa muito próxima a sua. Logo, o hidrogênio é o principal elemento químico que irá causar a desaceleração dos nêutrons de alta energia. A figura 7 apresenta a redução de energia após uma única colisão elástica com elementos químicos de diferentes massas atômicas. Para a maioria dos elementos químicos presentes na matriz da rocha (C, O, Si, Ca) a perda de energia não é superior a 25%. Para o hidrogênio, a perda de energia é quase total.

No espalhamento inelástico, uma parte da energia do nêutrons excita o núcleo do átomo atingido. A captura radioativa acontece com nêutrons de baixa energia (termais),

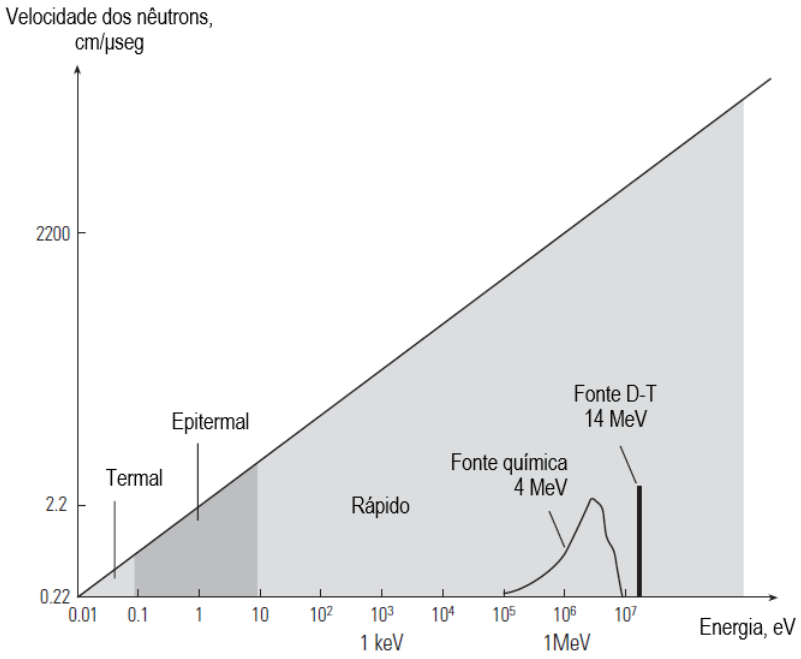


Figura 6 – Classificação dos nêutrons de acordo com suas energias e velocidades.

Fonte – Ellis e Singer (2007)

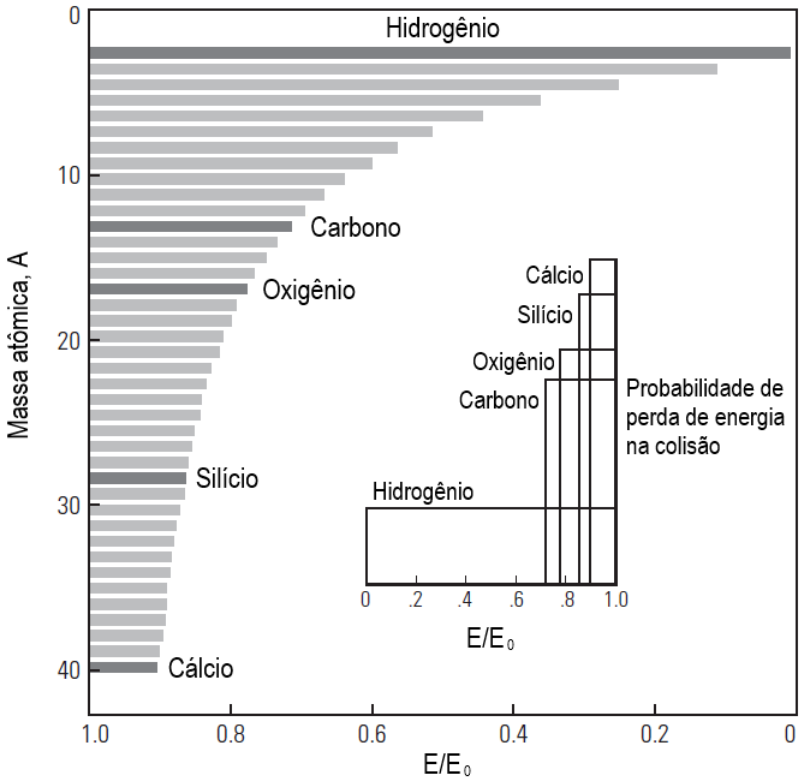


Figura 7 – Relação entre a perda de energia e a massa atômica do elemento com o qual o nêutron interage após uma única colisão elástica. A perda de energia é quase total quando o nêutron interage com o hidrogênio.

Fonte – Ellis e Singer (2007)

quando os nêutrons são absorvidos pelos átomos da formação. Ao contrário do espalhamento elástico, essas interações produzem radiação gama. A seção de choque de cada interação irá conduzir o regime ao qual os nêutrons estarão submetidos. A seção de choque σ pode ser multiplicada pelo número de átomos por centímetro cúbico N para se obter a seção de choque macroscópica Σ da interação i, segundo a equação 6.

$$\Sigma_i = N\sigma_i = \frac{N_{Av}\rho_b}{A}\sigma_i \tag{6}$$

Onde N_{Av} é o número de Avogadro, ρ_b é a densidade *bulk* e A é a massa atômica. A seção de choque de cada interação é função do nível de energia do nêutron. A captura radioativa acontece em baixas energias e o espalhamento inelástico acontece principalmente em altas energias. O espalhamento elástico pode acontecer em qualquer faixa de energia.

Apesar da complexidade das interações dos nêutrons com os átomos da formação, o processo como um todo pode ser resumido da seguinte maneira: nêutrons rápidos de alta energia são emitidos de uma fonte. Inicialmente, as interações são dominadas pelos espalhamentos elásticos e inelásticos, diminuindo a velocidade e energia dos nêutrons para a faixa epitermal e, em seguida, termal de energia. Nessas baixas energias, o espalhamento inelástico é reduzido e fenômenos de captura radioativa se tornam relevantes. As colisões elásticas continuam, diminuindo a energia dos nêutrons até que eles possam ser capturados pelos átomos da formação.

Nesse processo, duas distâncias características são importantes: a distância de desaceleração L_s e a de difusão L_d. A distância de desaceleração é aquela em que o nêutron precisou percorrer para reduzir sua energia da faixa rápida para a epitermal e termal, e é muito impactada pelo índice de hidrogênio da formação. A distância de difusão é a distância que o nêutron precisou percorrer em baixas energias até ser capturado, e terá relação com o índice de hidrogênio e a composição química da formação. A soma quadrática dessas duas distâncias será igual ao quadrado da distância de migração L_m (equação 7).

$$L_m^2 = L_s^2 + L_d^2 \tag{7}$$

Uma ferramenta de aquisição de perfis de nêutrons genérica é apresentada na figura 8. Os receptores realizam medidas em uma faixa de energia fixa (termal ou epitermal) e, em seguida, fazem a contagem de nêutrons daquela determinada energia. A razão entre

essas leituras terá relação com a distância de migração dos nêutrons (ELLIS; SINGER, 2007). Como a distância de migração é afetada principalmente pelo espalhamento elástico e pela captura radioativa, a razão entre as leituras dos detectores será afetada pelo índice de hidrogênio e composição química da formação. O índice de hidrogênio tem relação direta com a porosidade. Essas relações são demonstradas na figura 9.

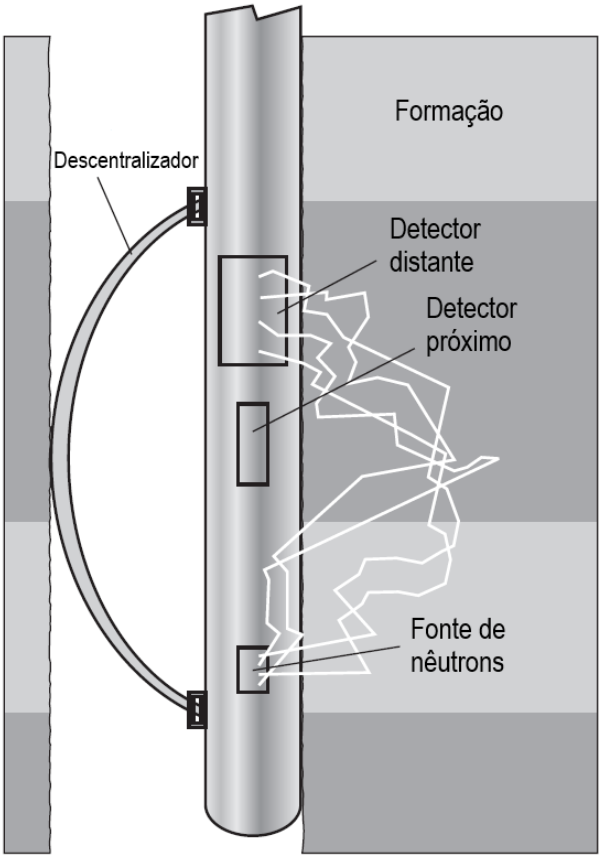


Figura 8 – Exemplo de ferramenta de aquisição do perfil de nêutrons, sendo adquirida junto a parede do poço.

Fonte – Ellis e Singer (2007)

Em posse das medidas feitas pelos detectores e suas relações com o índice de hidrogênio e com a composição química da rocha, torna-se possível calibrar as ferramentas de nêutrons. Essa calibração normalmente é feita em uma rocha calcítica de porosidade conhecida, cujo fluido possui índice de hidrogênio igual a 1. Para formações com essas características, o perfil de nêutrons representará a porosidade total da formação. Caso a composição da rocha seja diferente (arenitos e dolomitos) ou o índice de hidrogênio do fluido presente na formação seja muito diferente de 1 (gás), correções precisarão ser aplicadas à porosidade de nêutrons.

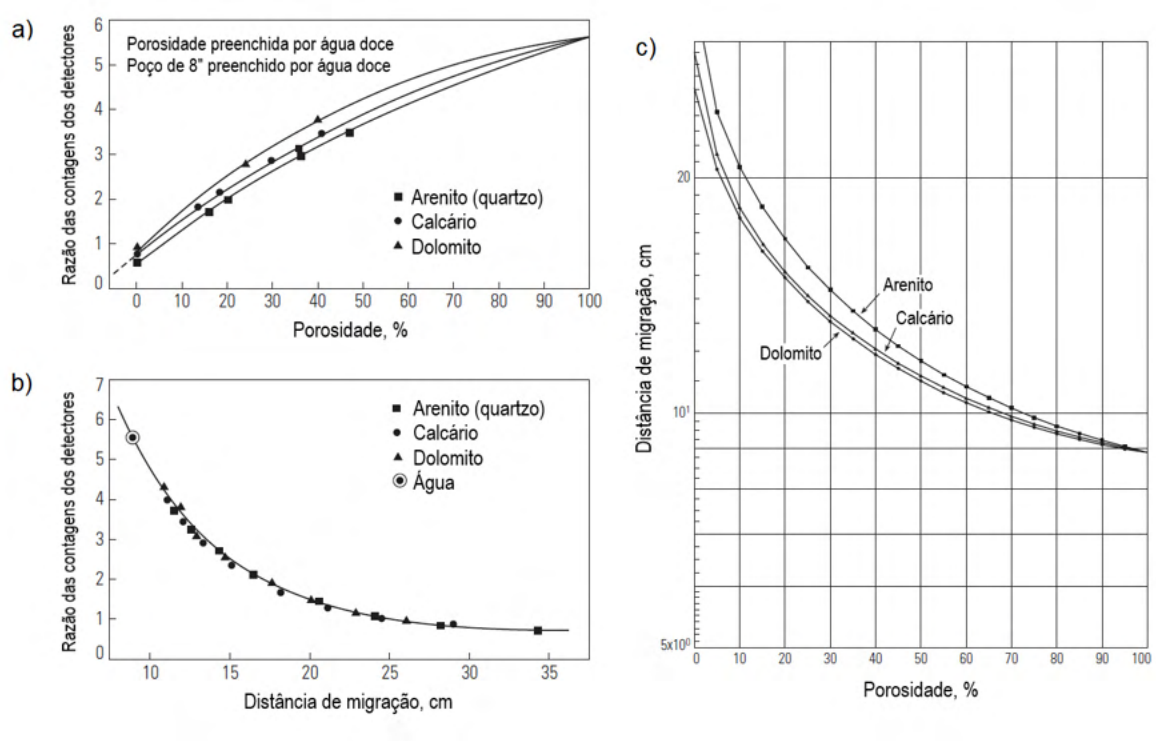


Figura 9 – Gráficos mostrando (a) a relação entre a razão das contagens dos detectores e a porosidade para diferentes tipos de rocha, (b) a relação entre a razão das contagens dos detectores e a distância de migração e (c) a relação entre a distância de migração e a porosidade para diferentes tipos de rocha.

Fonte – Ellis e Singer (2007)

Diversos fatores ambientais podem impactar a leitura da ferramenta de nêutrons (SCHLUMBERGER, 2009). Os principais são o diâmetro do poço, a salinidade e o peso do fluido de perfuração. Outros fatores como espessura do reboco, temperatura e pressão também podem ter impacto nas leituras. Por conta disso, o perfil de nêutrons precisa passar por diversas correções ambientais.

2.2.4 Perfis geoquímicos

Ferramentas que adquirem os perfis geoquímicos, também conhecidos como perfis de radiação induzida por nêutrons, exploram o espectro de radiação gama gerado pelas interações entre nêutrons emitidos e os elementos químicos da formação (ELLIS; SINGER, 2007). Como visto no capítulo 2.2.3, o espalhamento inelástico e a captura radioativa emitem raios gama de energia relacionada ao átomo e ao tipo de interação. Essa radiação é lida por cintilômetros de NaI(Tl), BGO ou $LaBr_3(Ce)$ na forma de um espectro que representa a soma dessas diversas emissões. Esse espectro é então decomposto em espectros elementares

de referência, representando as assinaturas típicas de cada elemento químico presentes na formação e no poço (figura 10).

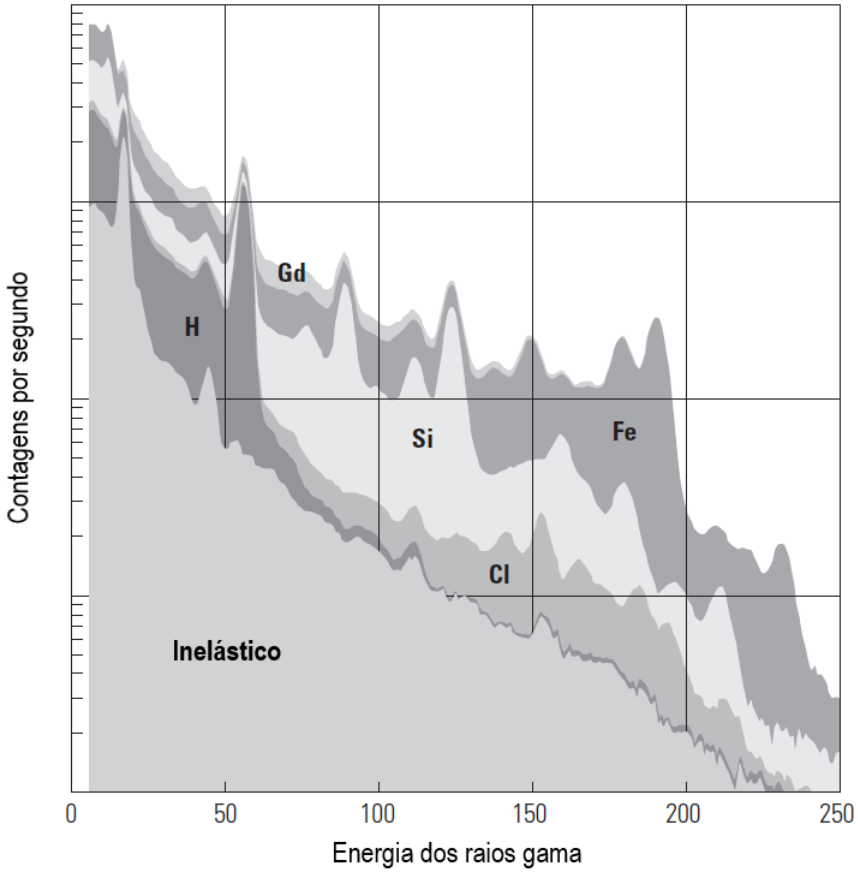


Figura 10 – Um exemplo de espectro de radiação gama e sua decomposição em espectros elementares de referência para diversos elementos químicos.

Fonte – Ellis e Singer (2007)

Os coeficientes relacionados aos espectros de referência, chamados de rendimentos ou *yields*, representam os pesos das contribuições de cada elemento químico no espectro medido. Porém, a determinação da concentração dos elementos químicos apresenta diversas dificuldades no ambiente de perfilagem (HERTZOG *et al.*, 1989), impossibilitando o uso direto dos rendimentos no cálculo das composições mássicas dos elementos. Isso se deve principalmente pelas seguintes questões:

1. Cada elemento possui um grau de resposta intrínseco (por exemplo: para o mesmo teor na rocha de dois elementos diferentes, a quantidade de radiação emitida é diferente);
2. Os rendimentos calculados envolvem respostas tanto de fluido quanto de matriz, e normalmente a decomposição espectral é feita de maneira que esses rendimentos

sejam relativos, ou seja, que a soma de todos os rendimentos seja igual à 1. Assim, para uma mesma composição de formação, os rendimentos de elementos da matriz (como Ca, Mg e Si) dependerão de fatores como a porosidade e salinidade;

3. Elementos como O e C estão presentes tanto no fluido quanto na matriz da rocha, não havendo uma maneira direta de se obter a contribuição exclusiva da matriz.

Essas questões são contornadas pela remoção dos rendimentos correspondentes a elementos presentes exclusivamente ou em grande parte nos fluidos, como H, O, C e Cl. Os rendimentos restantes são utilizados nas equações 8 e 9 (HERTZOG *et al.*, 1989) para calcular as concentrações mássicas dos elementos químicos presentes na matriz da rocha. Essas equações são conhecidas como modelo de óxidos.

$$F\left(\sum_i X_i \frac{Y_i}{S_i}\right) = 1 \tag{8}$$

$$w_i = F\frac{Y_i}{S_i} \tag{9}$$

Na equação 8, o rendimento Y do elemento i é dividido pela sensibilidade de detecção S desse elemento. Essa sensibilidade visa normalizar a resposta da radiação gama emitida entre os elementos químicos, levando em consideração diferenças na detecção da radiação gama, nas seções de choque das interações, nos pesos atômicos, entre outros. O rendimento corrigido é multiplicado por X, que representa a razão da massa do respectivo óxido ou carbonato para a massa do elemento i. Com isso, o modelo leva em consideração os elementos O e C, removidos anteriormente. A soma das massas é então igualada a 1 através de um fator de normalização F, representando a totalidade da matriz da rocha.

Na equação 9, o fator F é usado para calcular as concentrações mássicas absolutas w dos elementos químicos. Como essas concentrações foram calculadas apenas para os elementos presentes na matriz da rocha, elas são chamadas de concentrações mássicas. Essas concentrações podem ser utilizadas para a criação de modelos mineralógicos, úteis em reservatórios de litologia complexa (FLAUM; PIRIE, 1981; JR *et al.*, 1982; QUIREIN; VIGNE; CHAPMAN, 1987; ANDERSON *et al.*, 1988; GALFORD *et al.*, 2009; MACDONALD *et al.*, 2010; AJAYI; TORRES-VERDIN; PREEG, 2015; FREEDMAN *et al.*, 2015; ZHANG *et al.*, 2017). Esses modelos podem fornecer parâmetros como capacidade de troca catiônica (HERRON, 1986) ou carbono orgânico total (GONZALEZ *et al.*, 2013), e auxiliar em operações de acidificação (JIN *et al.*, 2019). Os perfis geoquímicos podem ser adquiridos

em poços revestidos, fornecendo informações sobre a razão C/O, úteis para mapear o contato hidrocarboneto/água em poços produtores (WESTAWAY; HERTZOG; PLASEK, 1983; NORTH, 1987; ULLOA *et al.*, 2016).

Ferramentas geoquímicas modernas podem usar fontes eletrônicas pulsantes de nêutrons (PEMPER *et al.*, 2006; RADTKE *et al.*, 2012), que permitem a aquisição de dois espectros de radiação gama: o primeiro adquirido durante a emissão de nêutrons, dominado pelo espalhamento inelástico, e o segundo adquirido entre emissões, dominado pela captura radioativa. O uso integrado desses dois espectros fornece inversão mais acurada, permitindo a detecção de um número maior de elementos químicos com melhor qualidade.

2.2.5 Perfis acústicos

Os perfis acústicos compreendem medidas das vagarosidades das ondas compressional (P) e cisalhante (S), propagadas na formação. A vagarosidade é o inverso da velocidade, normalmente fornecida em µs/pés. Em situações específicas, a vagarosidade da onda Stoneley também pode ser lida. Com esses perfis, outras informações de interesse podem ser adquiridas, como a razão de Poisson e os módulos *bulk*, cisalhante e de Young. O principal uso dessas informações é na calibração do poço com a sísmica, mas o cálculo de parâmetros petrofísicos como a porosidade também podem ser adquiridos.

A figura 11 apresenta o esquema de uma ferramenta simples de aquisição dos perfis acústicos. Ferramentas modernas podem possuir mais de dez receptores, diminuindo efeitos de parede de poço, mas piorando a resolução vertical. Ao contrário das ferramentas nucleares, a ferramenta acústica é adquirida centralizada em relação ao poço.

O transmissor emite uma onda acústica que se propaga no interior do poço até chegar na formação (ELLIS; SINGER, 2007). Essa onda continua a se propagar pela formação até formar uma frente de onda com um ângulo de 90º em relação à parede do poço (figura 12). Isso gera pequenos distúrbios na parede do poço que, por sua vez, promovem uma nova frente de onda que se propaga pelo fluido de perfuração do poço com a mesma velocidade da onda na formação. Essa nova frente de onda é lida pelos receptores, e a diferença do tempo de detecção entre receptores fornece a velocidade da onda acústica na formação. Esse fenômeno explica como a onda cisalhante, que se propaga apenas em sólidos, também pode ser lida apesar do fluido de perfuração: a onda cisalhante propagada

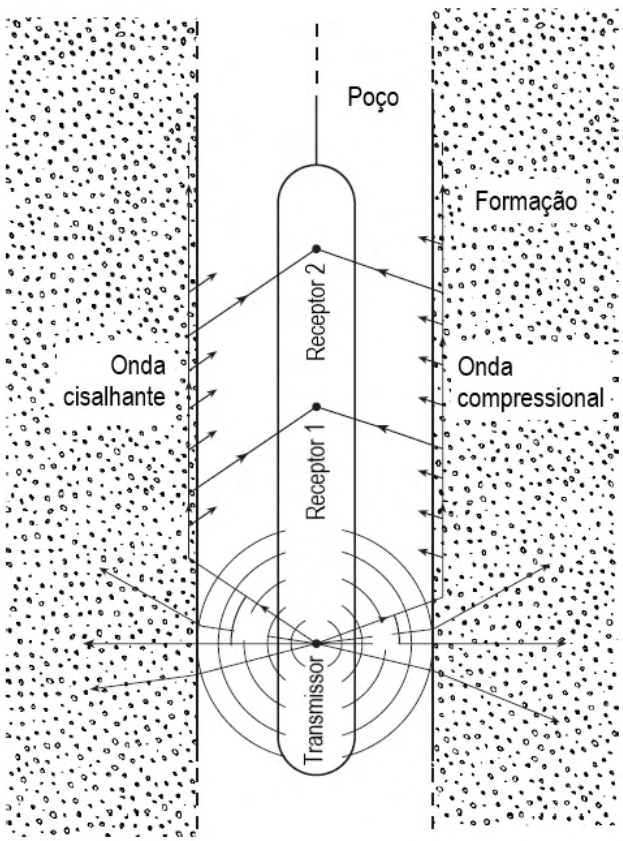


Figura 11 – Exemplo de ferramenta de aquisição dos perfis acústicos, sendo adquirida centralizada em relação ao poço. Ferramentas modernas podem possuir mais de dez receptores.

Fonte – Ellis e Singer (2007)

na formação gera uma frente de onda compressional que viaja pelo poço com a mesma velocidade da onda cisalhante.

A porosidade da formação pode ser calculada através da vagarosidade da onda compressional utilizando a equação 10 (WYLLIE; GREGORY; GARDNER, 1956).

$$\Delta t = \phi \Delta t_f + (1 - \phi)\Delta t_{ma} \tag{10}$$

Onde ϕ é a porosidade, Δt_f é a vagarosidade do fluido presente nos poros e Δt_{ma} é a vagarosidade da matriz da rocha. Entretanto, a equação 10 só é válida em formações muito compactadas. Posteriormente, Raymer, Hunt e Gardner (1980) propuseram uma equação para levar em consideração efeitos de baixa compactação.

Além do uso no cálculo de porosidade, os perfis acústicos podem ser usados para a identificação litológica, identificação e quantificação de gás, estimar propriedades mecânicas

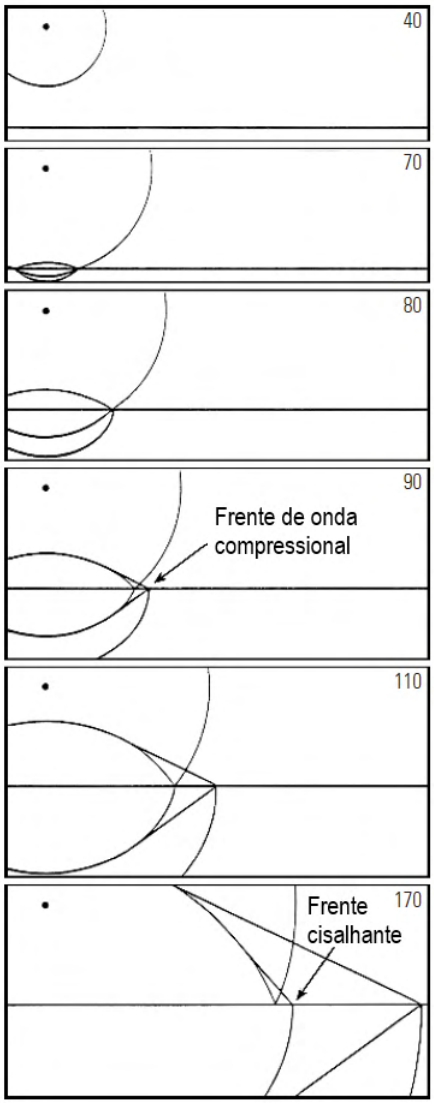


Figura 12 – Simulação 2D de uma onda acústica se propagando do poço para a formação. Em 90 µs, a frente de onda da formação faz um ângulo de 90º com a parede do poço, formando uma nova frente de onda que se propaga pelo fluido de perfuração do poço com a mesma velocidade da onda na formação.

Fonte – Ellis e Singer (2007)

da formação, estimar a pressão da formação, calibrar a sísmica e fazer a avaliação da cimentação de poços revestidos (ELLIS; SINGER, 2007).

2.2.6 Perfis de ressonância magnética nuclear

Os perfis de ressonância magnética nuclear (RMN) iniciaram uma nova era da avaliação de formações (COATES; XIAO; PRAMMER, 1999). Além de fornecer uma porosidade total independente da matriz, a análise da distribuição do tempo de relaxação (T_2, medida em milissegundos) permite estimar propriedades dos fluidos presentes na formação, determinar tamanho de poros e calcular de maneira mais acurada a permeabilidade absoluta da rocha.

Segundo Victor (2017), o procedimento de leitura da ferramenta de RMN pode ser resumido em: (i) polarização do momento magnético dos átomos de hidrogênio presentes na formação através de um campo magnético constante $\overrightarrow{B_0}$; (ii) excitação dos átomos de hidrogênio com um campo magnético $\overrightarrow{B_1}$ gerado por pulsos de rádio frequência oscilando na frequência ressonante dos átomos; e (iii) remoção de $\overrightarrow{B_1}$ e leitura do decaimento do sinal magnético emitido pelos átomos enquanto eles retornam para o estado original.

O procedimento descrito também é conhecido como experimento Carr-Purcell-Meiboom-Gill (CARR; PURCELL, 1954; MEIBOOM; GILL, 1958). O decaimento magnético medido é modelado como a combinação linear de diversos decaimentos exponenciais, conforme a equação 11.

$$M(t) = \int_0^\infty A(T_2) e^{-t/T_2} dT_2 \tag{11}$$

Onde $M(t)$ é a magnetização em função do tempo, T_2 é o tempo de relaxação e A é a amplitude relacionada ao tempo de relaxação. A figura 13 apresenta um decaimento magnético e a distribuição de T_2 adquirida após sua inversão usando a equação 11.

O valor da magnetização inicial $M(0)$ será igual a integral da distribuição de T_2, e dependerá do índice de hidrogênio do fluido presente na formação. Dessa forma, as ferramentas de RMN são calibradas com água de índice de hidrogênio igual a 1, e o valor lido em $M(0)$ dará a porosidade total da formação (PhiT). Caso os fluidos presentes na formação possuam índice de hidrogênio inferiores a 1, como gás, correções serão necessárias.

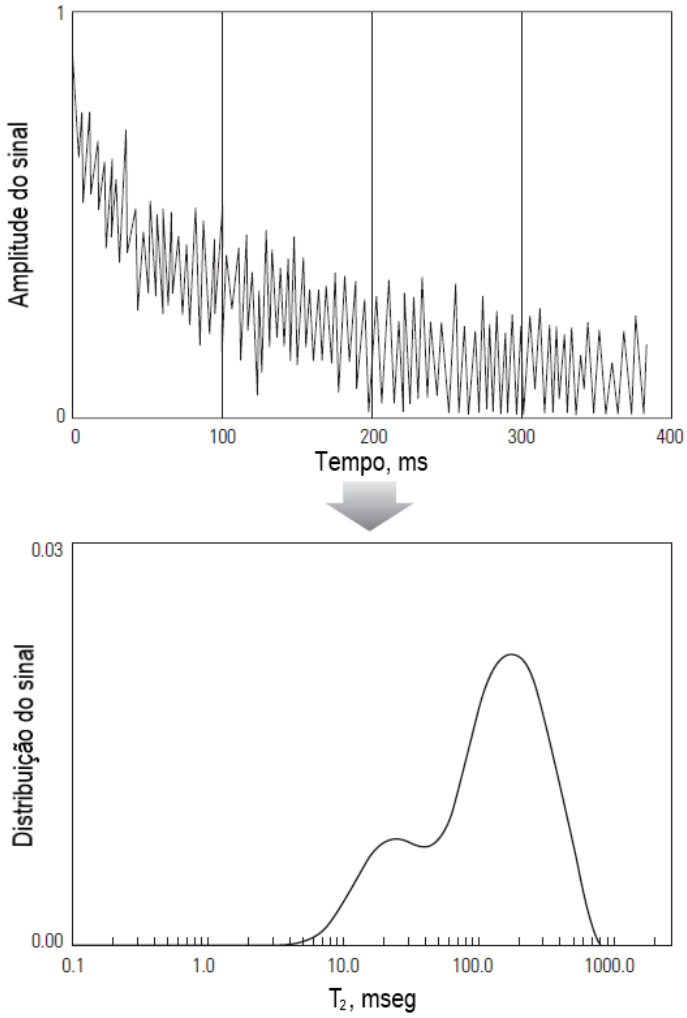


Figura 13 – Um exemplo de decaimento magnético ruidoso e a distribuição de T_2 adquirida após a inversão desse decaimento.

Fonte – Ellis e Singer (2007)

A distribuição de T_2 irá depender do tempo de relaxação *bulk* do fluido (T_{2B}), da relaxação de superfície (T_{2S}) e da relaxação por difusão (T_{2D}), segundo a equação 12 (COATES; XIAO; PRAMMER, 1999).

$$\frac{1}{T_2} = \frac{1}{T_{2B}} + \frac{1}{T_{2S}} + \frac{1}{T_{2D}} \tag{12}$$

O T_{2B} é uma propriedade intrínseca do fluido. O T_{2S} é resultado da interação dos átomos de hidrogênio com as paredes dos poros, relacionado a composição da matriz da rocha e do tamanho dos poros, conforme a equação 13.

$$\frac{1}{T_{2S}} = \rho_2 \frac{S}{V} \tag{13}$$

Onde ρ_2 é a constante de relaxação da matriz da rocha e $\frac{S}{V}$ é a razão da superfície por volume dos poros. Através dessa razão, percebe-se que poros maiores terão pouca influência no tempo de relaxação. O T_{2D} é a relaxação causada principalmente por campos magnéticos de gradiente não uniforme, e pode ser ignorado na maioria dos casos.

Em um meio poroso saturado apenas por um fluido, o mecanismo de relaxação por superfície T_{2S} tende a ser o dominante. Dessa forma, a distribuição de T_2 é relacionada à distribuição de tamanho de poros. Em rochas saturadas por um fluido molhante e outro não molhante, o mecanismo de relaxação *bulk* T_{2B} controla a relaxação do fluido não molhante, pois ele não está em contato com as paredes dos poros. O espaço ocupado pelo fluido não molhante diminui o volume de fluido molhante sem necessariamente diminuir sua superfície, deslocando os valores de T_2 do fluido molhante para tempos baixos. A figura 14 exemplifica esse fenômeno.

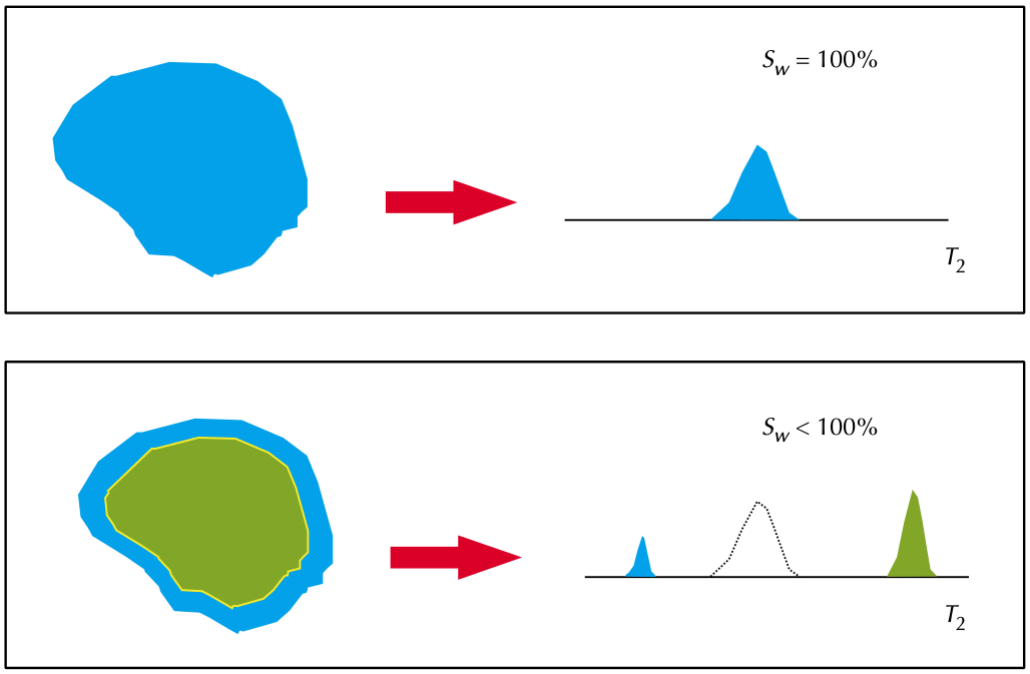


Figura 14 – Representação de uma distribuição de T_2 em um poro com apenas um fluido (acima) e após a inclusão de um fluido não molhante (abaixo). O T_{2B} domina a relaxação do fluido não molhante, deslocando o T_2 para tempos altos, enquanto que a diminuição da razão da superfície por volume do fluido molhante desloca seu T_2 para tempos baixos. Sw: saturação de água.

Fonte – Coates, Xiao e Prammer (1999)

Com a distribuição de T_2, é possível somar as amplitudes relacionadas a valores de T_2 acima e abaixo de um determinado corte para definir fluidos de interesse. Experimentos realizados por Straley *et al.* (1997) demonstram que o fluido livre (*Free fluid*, FF) em arenitos

está acima do corte de T_2 de 33 ms. Em carbonatos, está entre 92 e 100 ms, este último usado nos carbonatos do pré-sal da Bacia de Santos. O FF seria a fração de fluido que pode ser produzido. Abaixo desses cortes, o fluido é imóvel e não flui (*Bound volume irreducible*, BVI). Segundo Herlinger *et al.* (2020), no pré-sal da Bacia de Santos os valores de T_2 abaixo de 3 ms estão relacionados a água presente em argilas (*Clay Bound Water*, CBW). A porosidade acima do corte de argila é chamada de porosidade efetiva (PhiE). A figura 15 apresenta uma distribuição genérica de T_2 e sua interpretação.

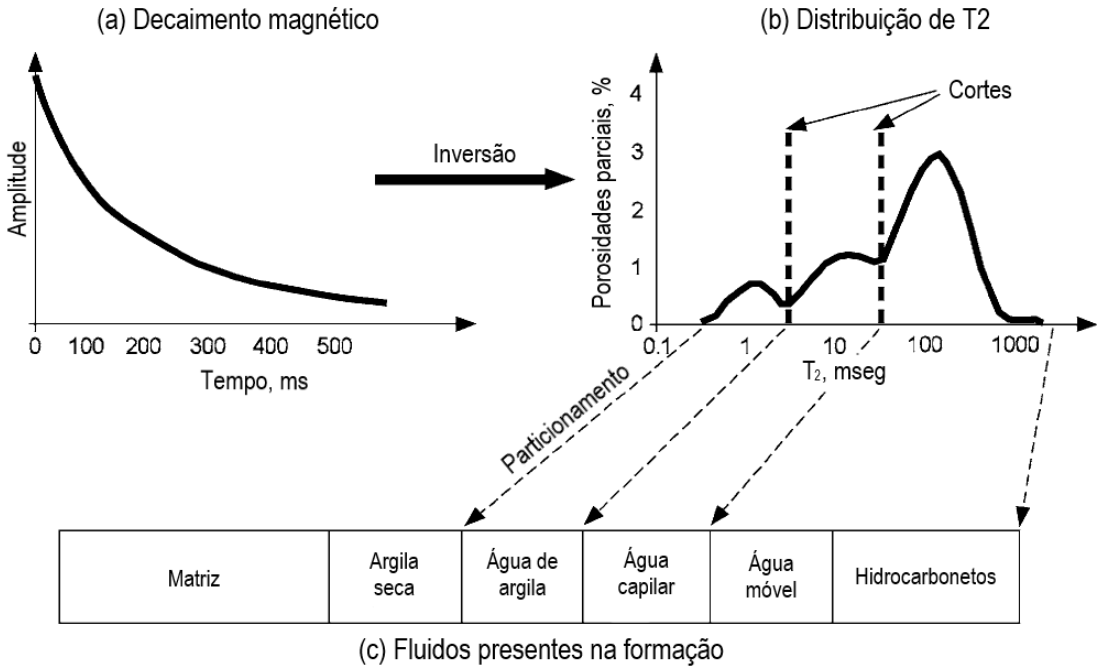


Figura 15 – Exemplo do uso de cortes para divisão e cálculo de volumes de fluido em uma distribuição de T_2. Em (a) decaimento magnético; (b) distribuição de T_2 após a inversão do decaimento, com seus respectivos cortes; e (c) interpretação dos diferentes fluidos relacionados aos valores da amplitude de T_2 entre os cortes.

Fonte – Westphal *et al.* (2005)

2.3 Análises laboratoriais

As análises laboratoriais consistiram nas análises de fluorescência de raios X e difratometria de raios X utilizadas para a caracterização geoquímica e mineralógica das amostras de rocha coletadas no pré-sal.

2.3.1 Fluorescência de raios X

A fluorescência de raios X (FRX) é uma das técnicas mais usadas para analisar a composição química de materiais. Após ser bombardeado por raios X de alta energia, um material emite radiação X secundária. Como a radiação emitida depende das transições que ocorrerão entre elétrons e átomos de um determinado material, ela pode ser usada para sua caracterização.

Átomos serão ionizados após serem bombardeados por radiação eletromagnética X de alta energia. Durante a ionização, os elétrons que orbitam as camadas interiores dos átomos serão expulsos, causando instabilidade e promovendo sua substituição por elétrons que orbitam camadas exteriores. Essas mudanças de órbitas emitem fótons de energia igual a diferença de energia entre órbitas. Logo, essa energia é uma assinatura específica do átomo envolvido na mudança de órbita.

A energia secundária emitida, diferente daquele inicialmente absorvida, recebe o nome de fluorescência (SIMON, 2018). Essa energia é adquirida na forma de um espectro com diferentes picos em diferentes faixas de energia. A localização desses picos está relacionada ao elemento químico, enquanto que suas intensidades terão relação com a concentração desse elemento químico presente no material, permitindo sua quantificação.

2.3.2 Difratometria de raios X

O fenômeno de difratometria de raios X (DRX) representa o espalhamento coerente relacionado a interação entre um feixe de raios X e os elétrons dos átomos que compõem um determinado material. Segundo Cullity e Stock (2001), o DRX é uma das principais ferramentas usadas na caracterização qualitativa e quantitativa de materiais cristalinos.

Quando os átomos que compõem um material possuem um arranjo sistemático, como é o caso da estrutura cristalina de um mineral, o espalhamento proveniente da incidência de um feixe de raios X se torna periódico. O fenômeno de difração pode ser observado em vários ângulos de incidência. A intensidade da difração será específica de uma estrutura cristalina, já que os planos que constituem essa estrutura possuem diferentes densidades de átomos e elétrons. Logo, cada material cristalino possui um padrão difratométrico característico, identificado pela angulação e intensidade relativa do

feixe difratado. O fenômeno de DRX ocorre somente quando a Lei de Bragg é satisfeita (JENKINS; SNYDER, 1996).

O espectro de difração obtido é então comparado com padrões de referência disponíveis em diversas bases de dados de organizações internacionais. Assim como na análise de FRX, o posicionamento e intensidade dos picos no difratograma permite a identificação e quantificação dos minerais que compõem uma determinada rocha.

2.4 Modelos mineralógicos

A quantificação das frações minerais em escala de poço se dá através de duas principais metodologias: modelagem probabilística e modelagem direta. Os modelos probabilísticos estimam as frações volumétricas dos componentes constituintes da formação, sejam minerais ou fluidos, através de quaisquer perfis de poços disponíveis. Já os modelos diretos estimam as frações mássicas dos minerais constituintes da matriz da rocha através da concentração de elementos químicos adquiridas pela ferramenta geoquímica.

2.4.1 Modelos minerais probabilísticos

Antes do desenvolvimento de ferramentas geoquímicas capazes de fornecer a concentração de elementos químicos presentes na matriz da rocha na década de 1990 (PEMPER, 2020), a mineralogia das formações em escala de poço era estimada somente através de modelos probabilísticos. Esses modelos usavam principalmente os perfis de raios gama, densidade, nêutrons, fator fotoelétrico e sônico (FREEDMAN *et al.*, 2015). Como esses perfis são afetados tanto pelos minerais quanto pelos fluidos, a porosidade e a saturação de água também precisam ser consideradas.

Os modelos probabilísticos calculam os componentes da formação através da minimização da diferença observada entre os perfis adquiridos e os perfis reconstruídos. Os perfis reconstruídos são estimados por equações de reconstrução, podendo ser aproximadas através de relações lineares de mistura (MITCHELL; NELSON, 1988), segundo a equação 14.

$$f_k = \sum_{l=1}^{c} e_{kl} V_l \tag{14}$$

Onde f é o valor do perfil reconstruído k, c é o número de componentes, V é a fração volumétrica do componente l e e é a resposta esperada do componente puro l para a ferramenta k, denominado valor de referência. Além das equações de reconstrução, a equação 15 obriga que a soma dos volumes de todos os componentes seja igual a um. Essa equação é conhecida como restrição de unidade.

$$V_1 + V_2 + ... + V_c = 1 \tag{15}$$

A função custo, dada pela diferença observada entre os perfis adquiridos e os perfis reconstruídos, é normalmente escrita como a equação 16.

$$\Delta = \sqrt{\frac{1}{t}\sum_{k=1}^{t}\left(\frac{p_k - f_k}{\epsilon_k}\right)^2} \tag{16}$$

Onde Δ é a diferença entre os perfis adquiridos e reconstruídos, t é o número de equações, p é o valor do perfil adquirido k, f é o valor do perfil reconstruído k, e ϵ é o desvio padrão associado a incerteza atribuída ao perfil k. Essa incerteza está associada a diversos fatores como a precisão da ferramenta, condições de poço e incertezas em relação à equação de reconstrução. O sistema de equações pode ser escrito no formato matricial e rapidamente resolvido profundidade a profundidade. Uma série de restrições podem ser adicionadas ao modelo, sendo a mais comum a exigência de não negatividade para as frações volumétricas dos componentes.

A incerteza ϵ serve para padronizar os valores dos perfis. Isso faz com que perfis como o de densidade, que varia entre 2 e 3 g/cm^3, e a vagarosidade, que varia entre 40 e 140 µs/pés, tenham pesos semelhantes na função custo. Caso alguma equação de reconstrução apresente maior ou menor confiabilidade, sua incerteza pode ser reduzida ou aumentada para que ela tenha maior ou menor importância na resolução do sistema, segundo os critérios do analista.

O valor Δ, que reflete o grau das diferenças entre os perfis medidos e reconstruídos, pode ser usado como controle de qualidade dos resultados do modelo. Idealmente, espera-se que esse valor seja igual ou muito próximo de zero. Na prática, diferenças muito maiores podem ser observadas. Além de problemas relacionados a baixa qualidade de aquisição de perfis devido a condições de poço, essas diferenças podem ser impactadas principalmente pela escolha incorreta de valores de referência, presença de minerais não considerados no modelo, e equações de reconstrução inadequadas.

Para manter o sistema matematicamente determinado ou sobre-determinado, é necessário que o número de componentes escolhidos seja no máximo igual ao número de perfis de poços mais um. Como as equações precisam considerar a porosidade da formação e podem também considerar os diferentes fluidos, o número de minerais utilizados no modelo pode ficar muito restrito. Essa limitação impacta principalmente os modelos aplicados às formações de mineralogia complexa, como é o caso das rochas do pré-sal da Bacia de Santos. Além disso, a escolha de valores de referência para alguns perfis pode não ser trivial para minerais complexos, como é o caso das argilas magnesianas.

Os modelos probabilísticos podem ser usados em conjunto com as concentrações de elementos químicos fornecidas pela ferramenta geoquímica. Entretanto, a transformação das frações mássicas elementares da matriz para as frações volumétricas dos minerais na rocha necessita da porosidade, obrigando o sistema a se tornar não linear. Ademais, não há uma forma direta de incorporar as análises laboratoriais de composição química e mineral de rocha nesses modelos.

Apesar das limitações, os modelos minerais probabilísticos ainda são muito empregados para a estimativa de frações minerais em diversos softwares de avaliação de perfis de poços, como o Interactive Petrophysics™, Geolog™ e Techlog™. Esses modelos foram usados como base da caracterização de diversos tipos de formações, como reservatórios siliciclásticos de gás hidrato (COLLETT *et al.*, 2011), folhelhos ricos em matéria orgânica (STADTMULLER; LIS-SLEDZIONA; SŁOTA-VALIM, 2018) e arenitos argilosos com óleo (EL-BAGOURY, 2020).

2.4.2 Modelos minerais diretos

A partir da década de 1990, a concentração dos elementos químicos presentes na matriz da rocha pela ferramenta geoquímica permitiu o desenvolvimento de modelos minerais diretos (PEMPER, 2020). Esses modelos são calibrados com análises laboratoriais de composição química e mineral de amostras de rochas das formações de interesse. A modelagem direta tem como meta diminuir a subjetividade da criação dos modelos probabilísticos, que dependem muito da quantidade e forma das equações de reconstrução escolhidas e dos perfis de poços disponíveis. Além disso, podem ser considerados mais

simples do que modelos probabilísticos na medida em que tratam apenas da matriz da rocha, independentemente dos fluidos.

Os primeiros modelos diretos consistiam em equações lineares que estimavam as frações minerais através das concentrações de elementos químicos, calibrados a partir de uma base de dados de análises de composição química e mineral de amostras de rocha de diversas formações (HERRON; HERRON, 1996). Como a quantidade de elementos químicos fornecida pela ferramenta pode ser insuficiente para estimar as frações de uma assembleia mineral complexa, modelos diretos subsequentes propuseram a inclusão de condições de contorno geoquímicas e geológicas para lidar com não singularidade. Essas condições podem ser restrições de balanço de massa e estequiometria mineral (FRANQUET; BRATOVICH; GLASS, 2012) ou restrições relacionadas ao tipo de rocha esperada (QUIREIN *et al.*, 2010).

O grande desafio na criação de modelos minerais diretos para as rochas do pré-sal da Bacia de Santos é a ambiguidade na composição química de alguns minerais. Os carbonatos das Formações Barra Velha e Itapema podem possuir alterações diagenéticas de dolomitização e silicificação, apresentando diferentes frações de calcita, dolomita e quartzo em sua composição (SILVA *et al.*, 2020). Enquanto isso, as principais espécies de argilas magnesianas encontradas nessas rochas são kerolita, Mg-esmectita e sepiolita (HERLINGER *et al.*, 2020). Como pode se observar na tabela 2, os principais elementos que compõem essas argilas são Mg e Si. Como as concentrações de O e H de matriz não são fornecidas nas ferramentas geoquímicas, as concentrações de Ca, Mg e Si não são suficientes para quantificar as argilas magnesianas, uma vez que sua fração pode ser escrita como qualquer combinação de calcita, dolomita e quartzo.

Tabela 2 – Composição química dos principais minerais que compõem os carbonatos das Formações Barra Velha e Itapema. Valores para as argilas magnesianas extraídos de Herlinger *et al.* (2020).

Mineral	Composição química
Calcita	$CaCO_3$
Dolomita	$CaMg(CO_3)_2$
Quartzo	SiO_2
Kerolita	$(Na_{0.10}Sr_{0.03}K_{0.03}Ca_{0.01})Si_{8.00}(Mg_{5.78}Al_{0.06}Fe_{0.03})O_{20}(OH)_4{\cdot}nH_2O$
Mg-esmectita	$(Na_{0.08}Sr_{0.03}K_{0.12}Ca_{0.06})(Si_{8.00}Al_{0.08})(Mg_{5.68}Al_{0.08}Fe_{0.04})O_{20}(OH)_4{\cdot}nH_2O$
Sepiolita	$(Sr_{0.05}K_{0.03}Ca_{0.02})Si_{12.11}(Mg_{7.51}Al_{0.03}Fe_{0.09})O_{30}\ (OH)_4{\cdot}(OH_2)_4$

Fonte – Lucas Oliveira, 2021

Um exemplo da ambiguidade das composições de calcita, dolomita, quartzo e argilas magnesianas é apresentado na figura 16. Essa figura apresenta os perfis adquiridos em um intervalo na Formação Barra Velha. Apesar da identificação de água de argila nos perfis de RMN abaixo de aproximadamente X482 m, nenhuma alteração é observada nos perfis geoquímicos. Esse fenômeno é típico da passagem do carbonato dolomitizado e silicificado para um carbonato composto por calcita, dolomita, quartzo e argilas magnesianas. Apesar de identificável, a quantificação dessas argilas por modelos diretos é bastante desafiadora.

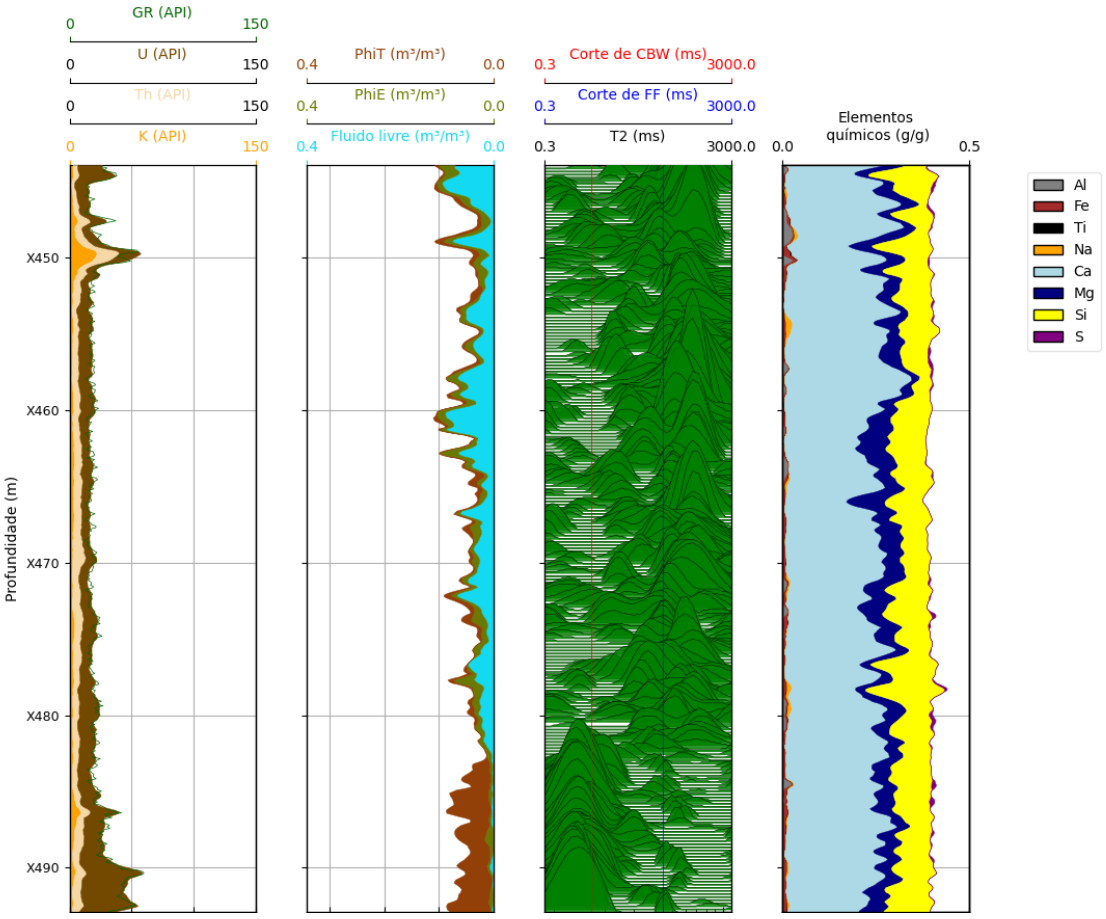


Figura 16 – Perfis adquiridos na Formação Barra Velha, mostrando a ambiguidade composicional dos carbonatos do pré-sal. Acima de aproximadamente X482 m, os perfis de RMN indica um carbonato sem argilas magnesianas. Abaixo dessa profundidade, a água de argila dos perfis de RMN (cor marrom) aponta para um espesso intervalo com argilas magnesianas. Apesar dessas diferenças, os perfis geoquímicos não apresentaram mudança significativa nas concentrações de Ca, Mg e Si.

2.5 Algoritmos de aprendizado de máquina

Algoritmos de aprendizado de máquina são excelentes ferramentas para achar correlações e padrões estruturais em extensas bases de dados (WITTEN; FRANK, 2005). A partir das correlações e padrões, esses algoritmos podem ser usados para fazer predições com novos valores de entrada.

Antes de detalhar os algoritmos específicos usados neste trabalho é importante revisar uma série de conceitos específicos da área de aprendizado de máquina. As bases para a explicação desses conceitos foram retiradas de Bishop (2006).

Base de dados: a base de dados agrupa um conjunto de instâncias que servirão de exemplos para os algoritmos de aprendizado de máquina. Diversas propriedades podem ser extraídas dessas instâncias, e serão consideradas as variáveis utilizadas nos algoritmos.

Variáveis: as variáveis se dividem em dependentes e independentes. As independentes são as utilizadas como entrada dos algoritmos de aprendizado de máquina, com as quais eles irão estipular correlações, e as dependentes são as saídas ou alvos dos algoritmos, que deverão ser preditas. As variáveis independentes são obrigatórias em todas as bases de dados, enquanto as variáveis dependentes podem ou não existir a depender do tipo de aprendizado, supervisionado ou não supervisionado.

Aprendizado supervisionado: quando o objetivo da aplicação do aprendizado de máquina é predizer uma variável de saída a partir de uma série de variáveis de entrada, o aprendizado é chamado de supervisionado. O algoritmo irá calibrar uma série de parâmetros próprios para determinar as formas como as variáveis de entrada se correlacionam entre si, e que geram estimativas mais próximas das variáveis de saída. Esse processo de calibração é chamado de treinamento. As variáveis de saída podem ser valores contínuos ou discretos, subdividindo o aprendizado supervisionado em regressão ou classificação, respectivamente.

Aprendizado não supervisionado: quando a base de dados não possui variáveis de saída, o aprendizado é considerado não supervisionado. O algoritmo irá calibrar uma série de parâmetros para subdividir as instâncias em grupos que apresentem alguma familiaridade entre si. Essa familiaridade pode se dar através de distâncias médias ou ângulos entre os vetores formados pelas variáveis das instâncias. O objetivo do aprendizado não supervisionado é o de agrupar instâncias ou reduzir a dimensionalidade das variáveis de entrada.

Viés e variância: em aprendizado de máquina, o viés e variância podem ser entendidos como o grau de complexidade dos modelos gerados pelos algoritmos. O viés de um modelo é a diferença entre a predição e o valor real da variável de saída, enquanto a variância é o quanto as predições se distanciam de uma predição média para um conjunto de instâncias. Modelos muito simples apresentam alto viés, enquanto modelos muito complexos apresentam alta variância. É impossível aumentar um sem diminuir o outro, fenômeno esse conhecido como troca viés-variância.

Troca viés-variância: é esperado que modelos com alto viés apresentarão alto erro durante o treinamento, uma vez que sua simplicidade não é capaz de capturar as complexidades das correlações entre as variáveis de entrada. A esse fenômeno é dado o nome de *underfitting*. Já modelos com alta variância apresentarão baixo erro durante o treinamento, mas terão péssimo desempenho em novas instâncias uma vez que sua alta complexidade irá memorizar as correlações da base de dados de treinamento, impactando sua capacidade de generalização. A esse fenômeno é dado o nome de *overfitting*. A troca viés-variância representa o fato de que os modelos com menores erros serão os que apresentarem um equilíbrio adequado entre viés e variância, exemplificado na figura 17.

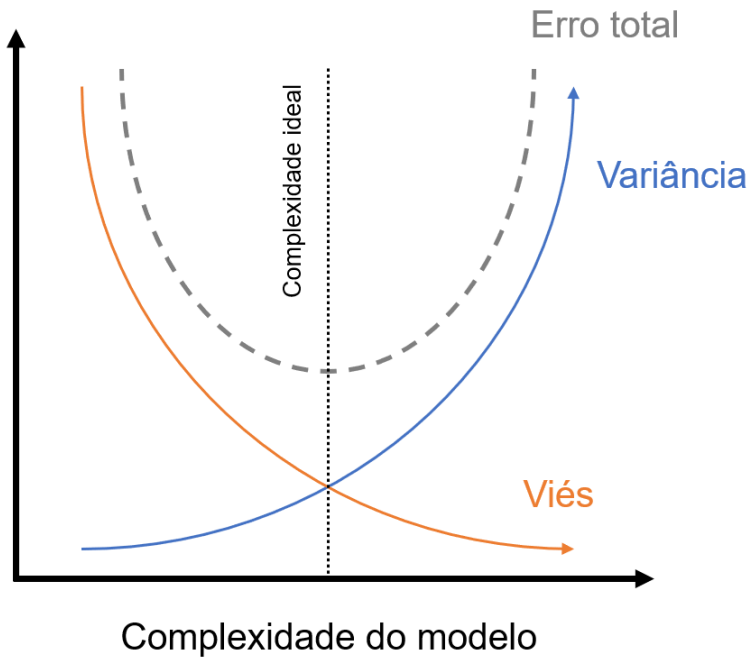


Figura 17 – Gráfico exemplificando a troca viés-variância. Observa-se que é impossível diminuir o viés (ou complexidade) de um modelo sem aumentar sua variância, e modelos com alto viés ou alta variância apresentarão os maiores erros. A complexidade ideal de um modelo é aquela que apresenta um equilíbrio entre essas duas propriedades.

A utilização de todas as instâncias de uma base de dados durante o treinamento de um algoritmo de aprendizado de máquina impede a identificação da complexidade ideal de um modelo. Sendo assim, as bases de dados normalmente são subdivididas pelo menos em conjuntos de instâncias para treinamento e teste dos modelos.

Conjunto de treinamento: antes do treinamento dos algoritmos de aprendizado de máquinas, as instâncias da base de dados precisam ser divididas aleatoriamente em conjuntos de treinamento e teste. O conjunto de treinamento será utilizado na calibração dos parâmetros do algoritmo, onde ele buscará ativamente o ajuste desses parâmetros para minimizar a diferença entre as estimativas e os valores reais das variáveis de saída. Essa diferença pode ser medida através do erro médio quadrático (EQM) ou do coeficiente de determinação (R^2), entre outros.

Conjunto de teste: após o treinamento, o algoritmo treinado faz estimativas utilizando as variáveis de entrada do conjunto de teste e essas estimativas são comparadas aos valores reais das variáveis de saída do conjunto de teste. Um modelo com alto viés apresentará resultados ruins tanto no conjunto de treinamento quanto no de teste (*underfitting*). Um modelo com alta variância apresentará ótimos resultados no conjunto de treinamento, porém péssimos resultados no conjunto de teste (*overfitting*). O modelo ideal apresentará resultados semelhantes tanto no conjunto de treinamento quanto no de teste. Para controlar o viés e variância dos modelos, hiperparâmetros precisam ser ajustados, sendo necessário um conjunto de validação ou validação cruzada no conjunto de treinamento.

Parâmetros e hiperparâmetros: em um algoritmo de aprendizado de máquina, os parâmetros podem ser entendidos como os pesos atribuídos aos diferentes elementos que compõem a estrutura do algoritmo. Esses pesos são modificados e calibrados durante o treinamento do algoritmo. Já os hiperparâmetros controlam característica relacionadas ao treinamento do algoritmo, como número de preditores de base ou taxa de aprendizado, precisando ser definidos antes do treinamento. Ao contrário dos parâmetros, eles não são modificados durante o treinamento. Como os hiperparâmetros modificam o viés e a variância dos modelos, é importante que eles sejam calibrados em um conjunto de instâncias separadas do conjunto de treinamento, definido como conjunto de validação. Caso a quantidade de instâncias seja insuficiente para a criação desse conjunto, a validação cruzada no conjunto de treinamento pode ser utilizada.

Conjunto de validação e validação cruzada: a escolha dos hiperparâmetros se dá em uma etapa anterior ao treinamento do modelo de aprendizado de máquina. Parte das

instâncias do conjunto de treinamento são aleatoriamente separadas em um conjunto de validação. Uma série de hiperparâmetros são definidos, o algoritmo é treinado e aplicado ao conjunto de validação. Os hiperparâmetros que proporcionarem os melhores resultados no conjunto de validação são escolhidos para o treinamento definitivo do algoritmo. Caso a base de dados contenha uma quantidade limitada de instâncias, inviabilizando a criação de um conjunto específico para a validação, o conjunto de treinamento pode ser subdividido em diversos conjunto menores. O treinamento então se dá revezando parte desses conjuntos ora para treinamento, ora para validação, e a média dos resultados obtidos é calculada. Esse procedimento é chamado de validação cruzada.

Base de dados enviesada: apesar dos esforços realizados para garantir um bom treinamento dos algoritmos de aprendizado de máquina, é fato que as estimativas de um modelo serão tão representativas quanto for a base de dados utilizada para treinar o modelo. Quando as instâncias utilizadas no treinamento e teste forem representativas apenas de uma parte do fenômeno que se deseja investigar, diz-se que a base de dados está enviesada. Mesmo apresentando bons resultados no conjunto de teste, um modelo treinado em uma base de dados enviesada terá dificuldades em generalizar suas estimativas, demandando soluções fora das rotinas de aprendizado de máquina para lidar com esses desafios.

Os conceitos definidos acima são comuns a quaisquer classes de algoritmos de aprendizado de máquina. As principais diferenças observadas são relacionadas aos diferentes tipos de instâncias e a arquitetura ideal do algoritmo para cada tipo de instância. Por exemplo, redes neurais convolucionais são ideais para trabalhar com imagens e redes neurais recorrentes são indicadas para séries temporais; porém, a troca viés-variância e as rotinas de treinamento, teste e validação se mantêm as mesmas.

Os principais algoritmos de aprendizado de máquina utilizados atualmente são construídos a partir de dois principais elementos: os neurônios artificiais, base das redes neurais (WENDEMUTHT, 1995), e as árvores de decisão, base dos algoritmos conhecidos como *ensemble* (HO, 1995). Este capítulo irá focar nos algoritmos *ensemble*.

As árvores de decisão são algoritmos estruturados em camadas de nós de decisão e folhas, conforme a figura 18 (WITTEN; FRANK, 2005). As folhas são as predições finais das árvores, e os nós de decisão são os pontos onde as variáveis de entrada são divididas com base em algum critério. Durante o treinamento de uma árvore de decisão, as variáveis de entrada são continuamente subdivididas com base em seus valores até que as estimativas

finais sejam as mais próximas das variáveis de saída. Elas podem ser usadas tanto em problemas de regressão como de classificação.

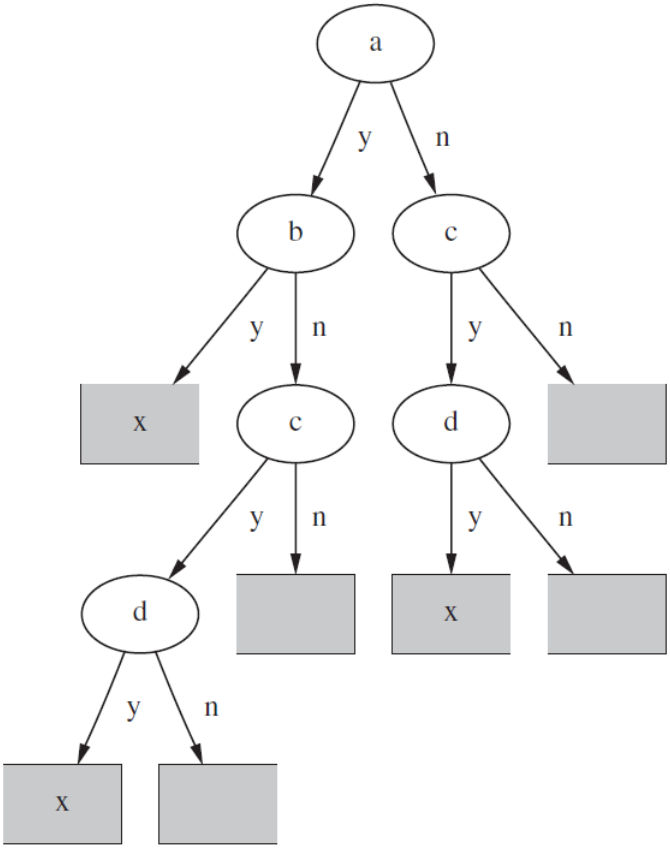


Figura 18 – Exemplo de árvore de decisão, com as folhas marcadas como quadrados cinzas e os nós de decisão são os circulos brancos.

Fonte – Witten e Frank (2005)

Apesar de úteis em alguns cenários, as árvores de decisão normalmente apresentam baixo desempenho em bases de dados complexas. Isso se dá pela limitação dos seus hiperparâmetros: para aumentar a complexidade, é necessário essencialmente aumentar o número de camadas da árvore de decisão. Árvores com poucas camadas irão apresentar alto viés, enquanto árvores com muitas camadas irão apresentar alta variância.

Para lidar com as limitações das árvores de decisão, duas principais técnicas são utilizadas. No *bootstrap aggregation*, diversas árvores de decisão com muitas camadas são ajustadas a grupos aleatórios do conjunto de treinamento (WITTEN; FRANK, 2005). A estimativa final é a combinação da estimativa de todas as árvores, podendo ser a média ou mediana dos valores em problemas de regressão ou votação majoritária em problemas de classificação. Essa combinação diminui a variância do modelo final. Já no *boosting*, árvores de decisão com poucas camadas são ajustadas sequencialmente no conjunto de

treinamento, diminuindo o erro da estimativa final gradativamente. O treinamento sequencial diminui o viés do modelo. Os modelos que usam o agrupamento de diversas árvores de decisão, seja por *bootstrap aggregation* ou *boosting*, são chamados de *ensemble*.

Uma vantagem dos algoritmos *ensemble* é a possibilidade de calcular a importância de cada uma das variáveis de entrada do modelo durante o treinamento. A importância das variáveis é uma métrica que indica o quanto uma determinada variável foi importante durante o treino. Essa métrica pode ser calculada (i) contando quantas vezes uma variável é utilizada na divisão do nó de uma árvore de decisão; (ii) permutando variáveis e analisando os efeitos no ajuste de uma árvore; ou (iii) avaliando como a remoção aleatória de uma variável afeta o ajuste de uma árvore de decisão.

O *boosting* é uma técnica de treinamento extremamente poderosa, apresentando alta capacidade de previsibilidade e generalização (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Pela sua natureza sequencial, é possível dizer que esses algoritmos aprendem com os próprios erros. Esse trabalho irá focar no algoritmo *Gradient Boosting*.

2.5.1 Gradient Boosting

Inicialmente proposto por Friedman (2001), o *Gradient Boosting* utiliza o treinamento sequencial e seus preditores simples são árvores de decisão. A cada iteração, o algoritmo ajusta uma árvore de decisão aos resíduos da iteração anterior, fazendo com que o algoritmo caminhe gradativamente na direção do menor erro. Por esse motivo, ele se assemelha a um processo de gradiente descendente. O algoritmo 1 apresenta o pseudo-código do *Gradient Boosting*.

Algoritmo 1 Pseudo-código do algoritmo *Gradient Boosting*.

procedure GRADIENT BOOSTING
  Considere y_i como o valor y da instância i, totalizando n instâncias.
  Ajuste uma primeira árvore $f_0(x)$ a y_i como $f_0(x) = argmin \sum_{i=1}^{n} L(y_i, \gamma)$.
  Para $m = 1$ até M faça:
    Para as n instâncias, calcule o resíduo $r_{im} = -\partial L(y_i, f_{m-1}(x_i))/\partial f_{m-1}(x_i)$.
    Ajuste uma árvore de regressão a r_{im} resultando em j folhas R_{jm}.
    Para as j folhas, calcule $\gamma_{jm} = argmin \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$.
    Atualize $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.
  Retorne as árvores $f_m(x)$, sendo $m = 1, 2, ..., M$.

Fonte – Adaptado de Kuhn e Johnson (2013)

Uma primeira árvore $f(x)$ é ajustada às instâncias e uma primeira aproximação é obtida. Essa primeira aproximação também pode ser a média das n instâncias. As iterações iniciam e o resíduo r da iteração m é calculado usando uma função de erro. Essa função pode variar em problemas de classificação. Uma nova árvore é ajustada aos resíduos; essa árvore irá possuir j terminações ou folhas. Como essa árvore foi ajustada aos resíduos, os valores contidos em suas folhas precisam ser recalculados. O valor γ é calculado e usado em conjunto com a árvore anterior para gerar uma nova árvore de regressão. Como o *Gradient Boosting* é um processo de gradiente descendente, todas as árvores são usadas de forma sequencial nas predições.

O algoritmo *Extreme Gradient Boosting* (XGBoost), proposto por Chen e Guestrin (2016), é uma evolução do *Gradient Boosting* tradicional. Além de ser mais rápido em termos computacionais, ele utiliza a segunda derivada da função dos resíduos.

2.5.2 Outros algoritmos de aprendizado de máquina

A lista de algoritmos de aprendizado de máquina é vasta, e novas técnicas são constantemente criadas para lidar com diferentes tipos de dados. Entretanto, alguns algoritmos se destacam por serem muito utilizados nos mais diferentes problemas nos últimos anos. Entre eles, é possível mencionar o *Support-Vector Machines* (SVM), redes neurais artificiais (RNA), o *Random Forest* e o *Adaptive Boosting* (AdaBoost).

O algoritmo SVM aplicado para problemas de regressão foi inicialmente proposto por Drucker *et al.* (1997), uma evolução do trabalho de Cortes e Vapnik (1995). Quando aplicado em problemas de classificação, o SVM encontra hiperplanos no espaço n-dimensional capazes de separar as instâncias nos seus devidos valores discretos. Em regressão, esses hiperplanos vão se ajustar aos dados, minimizando a distância entre eles e as instâncias. Os primeiros algoritmos de SVM eram capazes de honrar apenas relações lineares ou dados linearmente separáveis. Posteriormente, a introdução de operações de transformação dos dados através de um *kernel* permitiu introduzir não-linearidade ao SVM.

RNA são um dos algoritmos de aprendizado de máquina mais populares dos últimos anos, ganhando notoriedade na aplicação em visão computacional e problemas de classificação de imagem. O principal elemento das RNA é o neurônio artificial (MCCULLOCH; PITTS, 1943). Esse neurônio recebe informações de uma fonte externa e gera uma

saída usando uma função de ativação, capaz de introduzir não-linearidade ao modelo. A estrutura mais comum de RNA é a *Multilayer Perceptron*, proposta por Rosenblatt (1957) e aprimorada por autores como Wendemutht (1995) e Freund e Schapire (1999). A MLP consiste em um arranjo sequencial de diversas camadas, cada uma contendo múltiplos neurônios. O treinamento é realizado através da atualização dos pesos de cada uma das conexões entre neurônios, através de um processo de minimização de erro conhecido como *back-propagation*.

Introduzido por Ho (1995) e aprimorado por Breiman (2001), o *Random Forest* é um algoritmo *ensemble* que utiliza *bootstrap aggregation* para diminuir a alta variância das árvores de decisão. Durante o treinamento, as instâncias são divididas em diversos conjuntos menores, e diversas árvores são ajustadas a esses conjuntos. A predição final será a combinação dessas árvores, podendo ser votação majoritária em problemas de classificação, ou a média em regressão. Para reduzir ainda mais a variância e a possibilidade de *overfitting*, o *Random Forest* aleatoriamente seleciona as variáveis que serão usadas para treinar cada um dos conjuntos de instâncias.

O algoritmo *Adaptive Boosting* (AdaBoost) foi proposto por Freund e Schapire (1997) e utiliza o treinamento sequencial de preditores simples, sendo que o mais utilizado é a árvore de decisão. Um peso é atribuído a cada uma das instâncias da base de dados e, durante o treinamento, esses pesos são atualizados. As instâncias que apresentaram os maiores erros receberão os maiores pesos, incentivando os preditores a se ajustarem às instâncias de maior erro. O preditor final é calculado como o resultado de todos os preditores simples ponderados pelo erro global de cada iteração.

A principal diferença entre o *Random Forest* e o AdaBoost e *Gradient Boosting* é a de que o *Random Forest* não utiliza o erro observado em uma árvore de decisão para treinar as demais. Sendo assim, é esperado que o número de árvores necessário para treinar um modelo seja muito maior no *Random Forest*, aumento o tempo e custo computacional.

2.6 Aprendizado de máquina aplicado a perfis de poços

A aplicação de algoritmos de aprendizado de máquina tem aumentado ao longo dos últimos anos na indústria do petróleo como um todo, e o uso em perfis de poço é uma das

principais aplicações. Em geral, essas aplicações podem ser divididas em três categorias: predição, interpretação e imputação de dados.

Os problemas que visam calcular parâmetros petrofísicos de interesse através de perfis de poços e aprendizado de máquina podem ser considerados de predição. O objetivo é aproximar uma solução quando modelos analíticos não podem ser claramente definidos. Usualmente, esses modelos utilizam os perfis de poços como dados de entrada e parâmetros petrofísicos medidos em laboratório como saída. Exemplos de problemas de predição para calcular saturação de água em reservatórios podem ser encontrados em Al-Bulushi *et al.* (2012) e Hamada, Ahmed e Y (2018). Estimativa de porosidade e permeabilidade podem ser encontrados em Aminian *et al.* (2003), Nashawi e Malallah (2009), Verma *et al.* (2012) e Shabab *et al.* (2016). Problemas de predição para calcular carbono orgânico total podem ser encontrados em Alizadeh, Najjari e Kadkhodaie-Ilkhchi (2012), Zhao *et al.* (2015) e Negara, Jin e Agrawal (2016).

Problemas de interpretação podem ser divididos em classificação litológica ou petrofísica. Quando o objetivo é propagar litotipos identificados em amostras de rocha para um ou demais poços, algoritmos de aprendizado de máquina supervisionados podem ser aplicados (BESTAGINI; LIPARI; TUBARO, 2017; HOEINK; ZAMBRANO, 2017; GUARIDO, 2018). Quando amostras de rocha não estão disponíveis, uma classificação de eletrofácies pode ser feita através de algoritmos não supervisionados (YE; RABILLER, 2005; ASFAHANI; AHMAD; GHANI, 2018). Na classificação petrofísica, o aprendizado de máquina visa aumentar a objetividade e consistência da interpretação de perfis. A identificação automática de zonas de interesse permite a divisão do poço em aquíferos, zonas produtivas e zonas portadoras de óleo e gás (BELOZEROV *et al.*, 2018; WU *et al.*, 2018).

Imputação de dados usa aprendizado de máquina para substituir os dados de perfis em regiões em que eles não puderam ser adquiridos ou apresentaram baixa qualidade. Segundo Rolon *et al.* (2009), o uso de aprendizado de máquina para essa função é melhor do que métodos tradicionais, como regressão linear. Porém, Bahrpeyma, Golchin e Cranganu (2013) apontam para o fato que essas técnicas podem demandar alto custo computacional. Akkurt *et al.* (2018) propõem um fluxo totalmente automatizado que é capaz de fazer o controle de qualidade, detectar anomalias em perfis e fazer a reconstrução de regiões com baixa qualidade, usando os algoritmos *Support-Vector Machines* e *Quantile Regression Forest*.

A tabela 3 apresenta um resumo de resultados obtidos em diferentes problemas de imputação encontrados na literatura (CHEN *et al.*, 2005; ROLON *et al.*, 2009; BAHRPEYMA; GOLCHIN; CRANGANU, 2013; KORJANI *et al.*, 2016; SALEHI *et al.*, 2017; AKINNIKAWE; LYNE; ROBERTS, 2018; AKKURT *et al.*, 2018; ZHANG; CHEN; MENG, 2018). Observa-se que as referências encontradas propõem a imputação dos perfis considerados básicos (raios gama, resistividade, densidade, nêutrons, acústico e fator fotoelétrico). Nenhum trabalho propõe a aplicação de aprendizado de máquina em perfis complexos, como os geoquímicos ou os de RMN.

Também se observa que a maioria dos trabalhos de imputação de perfis faz uso de RNA. Provavelmente isso se deve ao aumento de popularidade desses algoritmos nos últimos anos, e não a resultados concretos. De fato, a maioria das competições de aprendizado de máquinas realizadas no site *Kaggle*[1] no ano de 2015 foram vencidas por algoritmos de *Gradient Boosting*, conforme relatado por Chen e Guestrin (2016).

Outra característica interessante dos trabalhos de imputação é que não há uma padronização na métrica para definir a qualidade dos resultados gerados. As métricas encontradas foram: erro quadrático médio (EQM), erro quadrático absoluto (EQA) e coeficiente de determinação (R^2). Em alguns trabalhos, erros não especificados ou apenas comparação visual foram utilizados. Como os perfis apresentam diversas escalas variando em diversas unidades, o R^2 parece ser a métrica mais recomendada.

Embora os modelos mineralógicos sejam parte essencial da avaliação dos reservatórios, ainda não há abordagens de desenvolvimento desses modelos com base em aprendizado de máquina.

2.7 Hibridização de modelos de aprendizado de máquina

No contexto de aprendizado de máquina, o termo “modelo híbrido” é usado principalmente para descrever modelos que utilizam diferentes algoritmos de aprendizado de máquina de princípios distintos (ARDABILI; MOSAVI; VáRKONYI-KóCZY, 2019). Como exemplo, é possível citar a união de algoritmos baseados em árvores de decisão com redes neurais, algoritmos baseados em aprendizado supervisionado com não-supervisionado, e fluxos de trabalho que utilizam aprendizado de máquina clássico com aprendizado profundo. Esse tipo de abordagem híbrida já foi utilizada na caracterização de reservatórios,

[1] https://www.kaggle.com/

Tabela 3 – Resumo de resultados obtidos em diferentes problemas de imputação encontrados na literatura.

Referência	Algoritmo usado	Resultados	Perfis gerados
Chen *et al.* (2005)	RNA*	EQM** = 0,02385	Resistividade Densidade Nêutrons
Rolon *et al.* (2009)	RNA	R^2 entre 0,85 e 0,95	Raios gama Resistividade Densidade Nêutrons
Bahrpeyma, Golchin e Cranganu (2013)	Lógica *fuzzy*	R^2 = 0,85 R^2 = 0,92	Densidade Acústico
Korjani *et al.* (2016)	RNA	Erro global = 0,021 EQM e R^2 não especificados	Raios Gama Resistividade Densidade Nêutrons
Salehi *et al.* (2017)	RNA	R^2 = 0,92018 R^2 = 0,97962	Resistividade Densidade
Akinnikawe, Lyne e Roberts (2018)	*Random Forest* RNA	EQA** = 0,33 EQA = 320,2	Fator fotoelétrico UCS***
Akkurt *et al.* (2018)	*Quantile Regression Forest*	Erro não especificado, mas satisfatório	Densidade Acústico
Zhang, Chen e Meng (2018)	RNR*	EQM = 0,6083	Resistividade Acústico Nêutrons

*RNA, RNR: Redes neurais artificiais e recorrentes, respectivamente.
**EQM, EQA: Erros quadráticos médio e absoluto, respectivamente.
***UCS: *Unconfined compressive strength*, resistência à compressão não confinada.

Fonte – Lucas Oliveira, 2021

majoritariamente na estimativa de porosidade e permeabilidade (ANIFOWOSE; LABADIN; ABDULRAHEEM, 2017).

Entretanto, no presente trabalho, o termo “modelo híbrido” será utilizado para descrever o uso conjunto de algoritmos de aprendizado de máquina e modelos baseados em princípios físicos. Essa abordagem também é chamada de aprendizado de máquina orientado por princípios físicos (*physics-guided machine learning*). Nesse contexto, modelos minerais probabilísticos podem ser considerados orientados por princípios físicos já que buscam reconstruir as respostas de perfis utilizando equações das respostas individuais dos componentes da formação em função de suas frações volumétricas.

Modelos híbridos dessa natureza foram aplicados no setor de óleo e gás para auxiliar na inversão sísmica (CHEN; SAYGIN, 2020; SUN; INNANEN; HUANG, 2021). Inversões

sísmicas puramente analíticas, baseadas em equações que descrevem a propagação de ondas em subsuperfície, precisam de uma série de variáveis de entrada e condições iniciais e de contorno que podem gerar resultados pouco confiáveis, além de demandarem muito tempo e recursos computacionais. Ao mesmo tempo, modelos de aprendizado de máquina puramente condicionados pelos dados demandam gigantescas bases de dados e diversos graus de liberdade para apresentarem resultados confiáveis. De maneira geral, os modelos híbridos buscam regularizar a inversão gerada por uma RNA utilizando equações analíticas da propagação de ondas em subsuperfície (CHEN; SAYGIN, 2020), penalizando os resultados que apresentam inconsistências físicas durante o treinamento da rede.

Apesar do potencial dos modelos híbridos, até o momento não existem trabalhos que explorem essa técnica para resolver problemas de caracterização de reservatórios. Não há ainda uma metodologia estabelecida de como desenvolver tais modelos, que requerem adequações às complexidades de cada aplicação. Por exemplo, modelos baseados em princípios físicos podem ser usados para direcionar o treinamento de algoritmos de aprendizado de máquina ou os resultados de algoritmos treinados podem ser usados como dados de entrada de modelos baseados em princípios físicos.

3 Metodologia

A metodologia empregada visou incorporar as melhores práticas encontradas em problemas de aprendizado de máquina às particularidades dos perfis de poço e análises laboratoriais de amostras de rocha. Três fluxos de trabalho foram desenvolvidos: o primeiro para o desenvolvimento de perfis geoquímicos sintéticos para a caracterização geoquímica do reservatório (figura 19); o segundo para a construção de uma sistemática para obtenção de um modelo mineralógico através de algoritmos de aprendizado de máquina e análises de FRX e DRX (figura 20); e o terceiro para a construção de uma sistemática de um modelo mineralógico através da hibridização dos algoritmos de aprendizado de máquina treinados e modelos probabilísticos (figura 21).

Após a construção das duas bases de dados, uma análise exploratória foi feita para entender as particularidades das variáveis e traçar as melhores estratégias para a criação dos modelos de aprendizado de máquina. Em seguida, foi realizado o treinamento e validação dos modelos. Para o modelo mineralógico híbrido, uma etapa adicional posterior a do aprendizado de máquina foi realizada, com a combinação das estimativas geradas pelos algoritmos com equações de reconstrução e suas incertezas. Uma etapa final de teste foi realizada utilizando dados que não foram adicionados às bases de treinamento e validação, com o objetivo de avaliar a qualidade e aplicação dos modelos criados em cenários reais.

3.1 Perfis geoquímicos sintéticos

3.1.1 Preparação dos dados

A base de dados para a criação dos perfis geoquímicos sintéticos foi composta por 22 poços perfurados nas rochas do pré-sal de oito campos distintos. A soma da metragem dos poços foi de aproximadamente 8.600 m e os perfis apresentaram uma taxa de amostragem média de 12 cm, totalizando 77.800 instâncias. A perfilagem adquirida em todos os poços foi a completa (rever tabela 1). A tabela 4 apresenta um resumo das principais características desses poços. Dos 22 poços, três foram separados para a fase de teste. Esses três poços não participaram das etapas de análise exploratória e treinamento e validação.

Como o objetivo é o desenvolvimento de perfis geoquímicos sintéticos em poços com redução do escopo de perfilagem, as variáveis de entrada do modelo foram os perfis

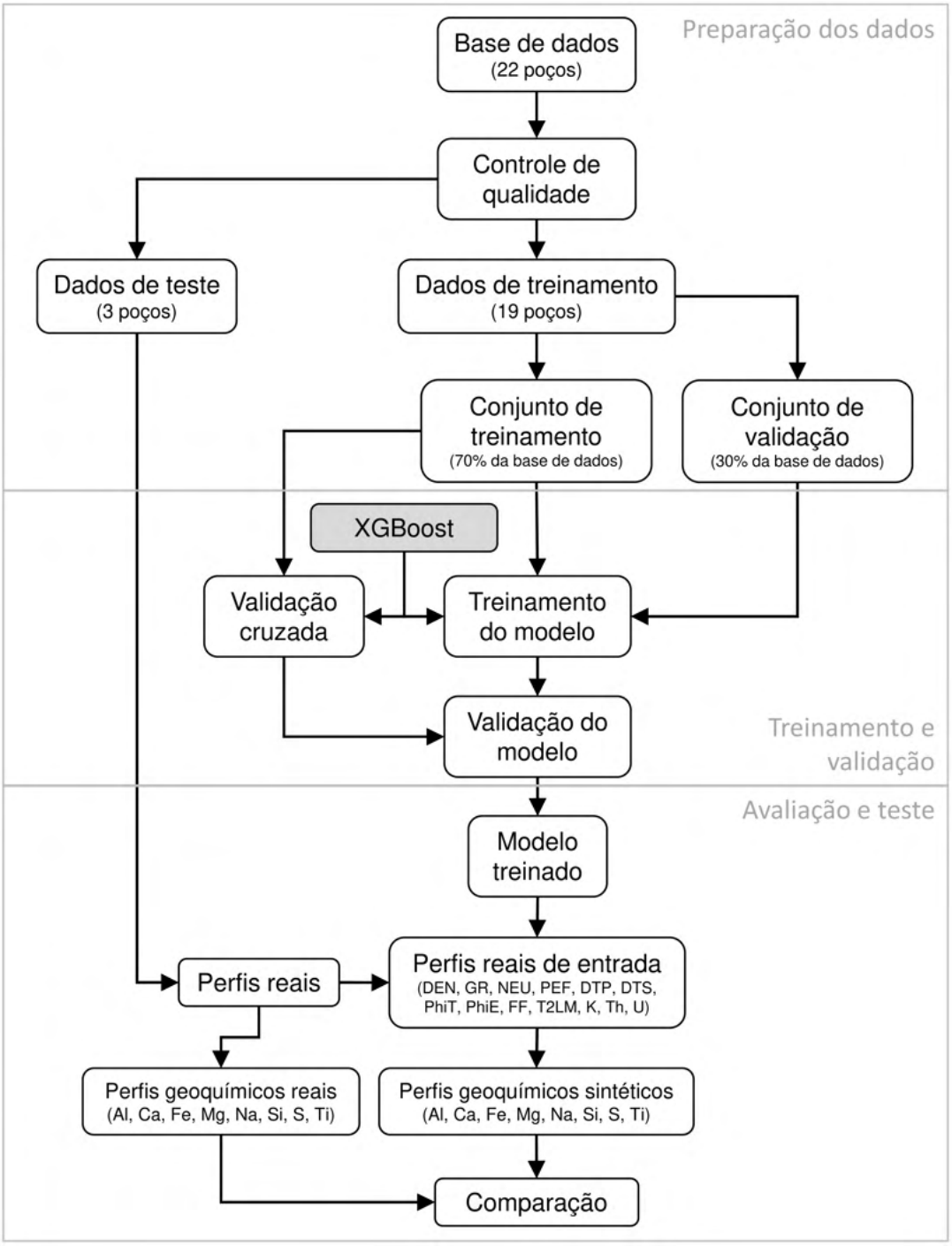


Figura 19 – Sequência proposta para o desenvolvimento dos perfis geoquímicos sintéticos.

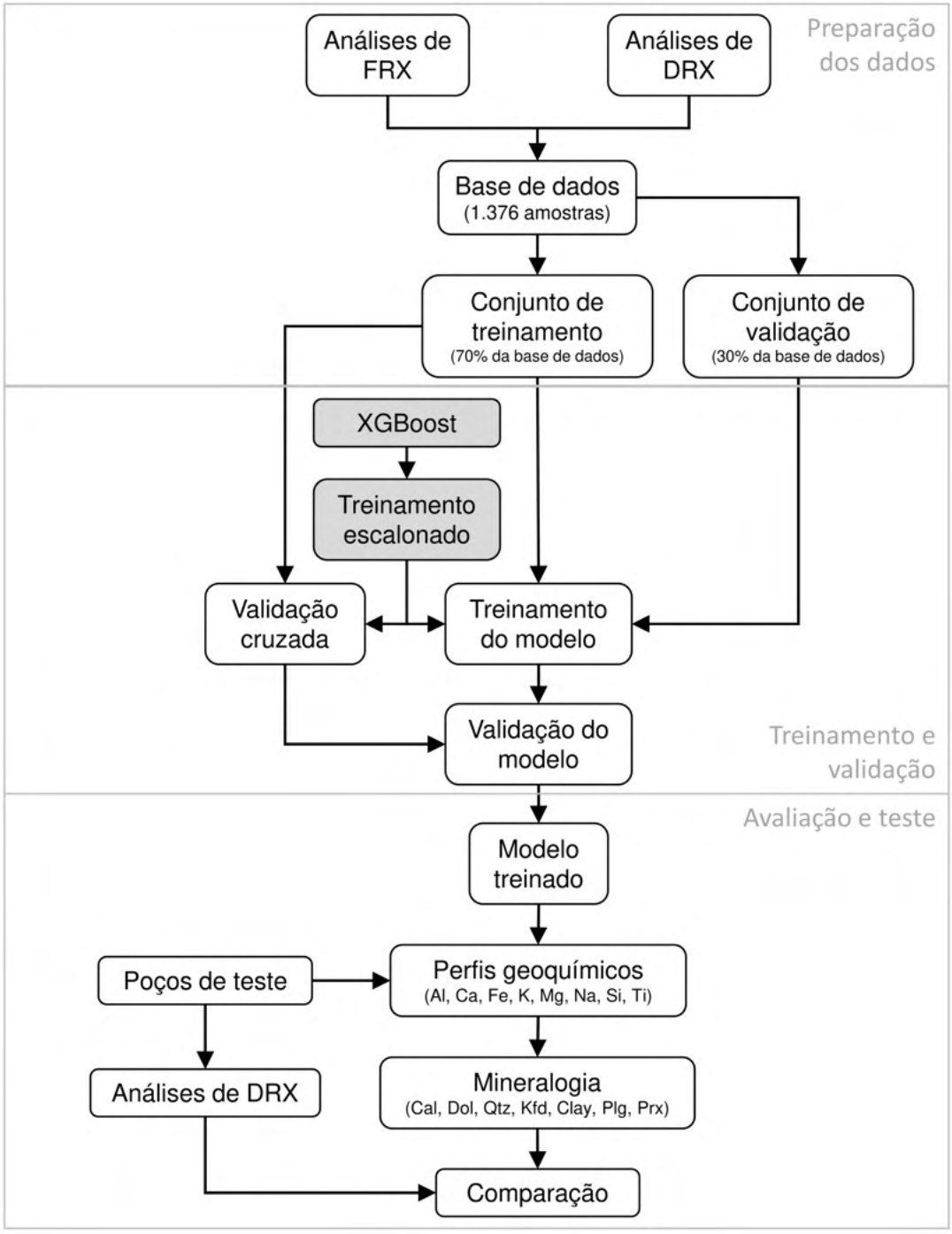


Figura 20 – Sequência proposta para a criação do modelo mineralógico por aprendizado de máquina.

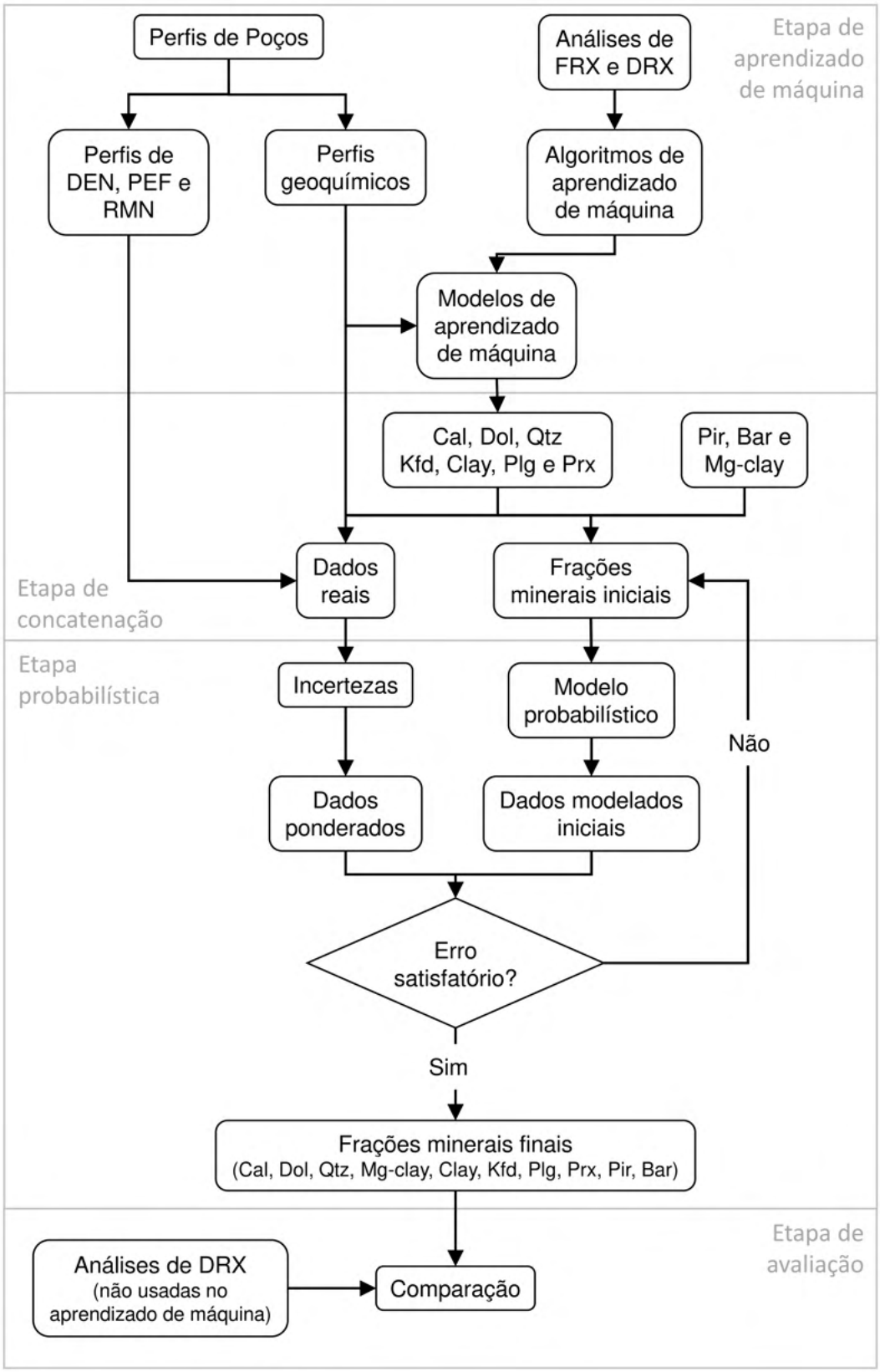


Figura 21 – Sequência proposta para a criação do modelo mineralógico híbrido.

Tabela 4 – Resumo da característica dos poços utilizados para compor a base de dados para o desenvolvimento dos perfis geoquímicos sintéticos.

Campo	Número de poços	Diâmetro dos poços	Fluido de perfuração	Formações encontradas
1	1	8,50"	Base aquosa	Barra Velha
2	3	12,25"	Base aquosa e oleosa	Barra Velha Itapema Piçarras Camboriú
3	1	12,25"	Base oleosa	Barra Velha
4	3	8,50"e 12,25"	Base aquosa e oleosa	Barra Velha Itapema
5	1	8,50"	Base aquosa	Barra Velha
6	1	12,25"	Base oleosa	Barra Velha
7	8	12,25"	Base oleosa	Barra Velha Itapema Piçarras Camboriú
8	4	12,25"	Base oleosa	Barra Velha Itapema Piçarras

Fonte – Lucas Oliveira, 2021

adquiridos em perfilagens reduzidas. As variáveis de saída foram as concentrações mássicas dos elementos químicos fornecidas pelas ferramentas geoquímicas. Essas variáveis e suas principais estatísticas estão resumidas na tabela 5.

Observa-se a grande diversidade de unidades que compõem uma aquisição de perfis, o que acarreta em um intervalo de valores bastante heterogêneo. Alguns perfis apresentam valores da ordem de fração de unidade, como é o caso das concentrações mássicas dos elementos químicos, enquanto outros apresentam valores da ordem de centenas a milhares de unidades, como é o caso das vagarosidades das ondas P e S e da média geométrica de T_2.

3.1.2 Controle de qualidade

Os perfis dos poços que compõem a base de dados passaram por um rigoroso controle de qualidade. Correções ambientais foram aplicadas aos perfis para que os efeitos de diâmetro de poço e fluido de perfuração fossem considerados. Regiões do poço com intenso arrombamento foram removidas da base de treinamento. Essas zonas foram identificadas

Tabela 5 – Resumo das principais estatísticas das variáveis de entrada e saída dos modelos de perfis geoquímicos sintéticos.

Perfis	Descrição	Unidade	Mínimo	Máximo	Média	Desvio padrão
Variáveis de entrada						
DEN	Densidade	g/cm^3	1,59	4,04	2,52	0,12
GR	Raios gama	API	3,25	238,0	34,6	22,6
NEU	Nêutrons	m^3/m^3	-0,02	1,12	0,12	0,06
PEF	Fator fotoelétrico	-	1,72	10,0	4,90	0,92
DTP	Vagarosidade da onda P	µs/pés	41,0	131,0	62,8	7,71
DTS	Vagarosidade da onda S	µs/pés	64,2	269,0	111,0	15,6
PhiT	Porosidade total do RMN	m^3/m^3	0,00	2,53	0,08	0,07
PhiE	Porosidade efetiva do RMN	m^3/m^3	0,00	2,53	0,11	0,07
FF	Fluido livre do RMN	m^3/m^3	0,00	2,53	0,12	0,07
T2LM	Média geométrica do T_2	ms	0,53	2817	242,0	262,0
K	Potássio	g/g	0,00	0,05	0,00	0,01
Th	Tório	ppm	0,00	39,3	1,70	1,47
U	Urânio	ppm	0,13	20,5	2,72	1,99
Variáveis de saída						
Al	Alumínio	g/g	0,00	0,11	0,00	0,01
Ca	Cálcio	g/g	0,00	0,40	0,29	0,07
Fe	Ferro	g/g	0,00	0,09	0,01	0,01
Mg	Magnésio	g/g	0,00	0,13	0,02	0,02
Na	Sódio	g/g	0,00	0,04	0,00	0,00
Si	Silício	g/g	0,00	0,44	0,07	0,06
S	Enxofre	g/g	0,00	0,25	0,00	0,01
Ti	Titânio	g/g	0,00	0,02	0,00	0,00

Fonte – Lucas Oliveira, 2021

pelo perfil de diâmetro do poço corrido junto da ferramenta de densidade. Essas regiões de alto arrombamento poderiam comprometer a qualidade dos perfis, principalmente aqueles que são adquiridos junto à parede do poço, como os perfis de densidade, nêutrons e RMN (figura 22).

3.1.3 Treinamento e validação

A base de dados foi aleatoriamente dividida em 70% para treinamento e 30% para validação. Um total de 49.270 instâncias foram utilizadas no treinamento e 21.117 instâncias foram utilizadas na validação.

A escolha do algoritmo levou em conta os resultados apresentados no anexo A, que demonstra a aplicação de aprendizado de máquina a perfis de poços. Cinco algoritmos foram

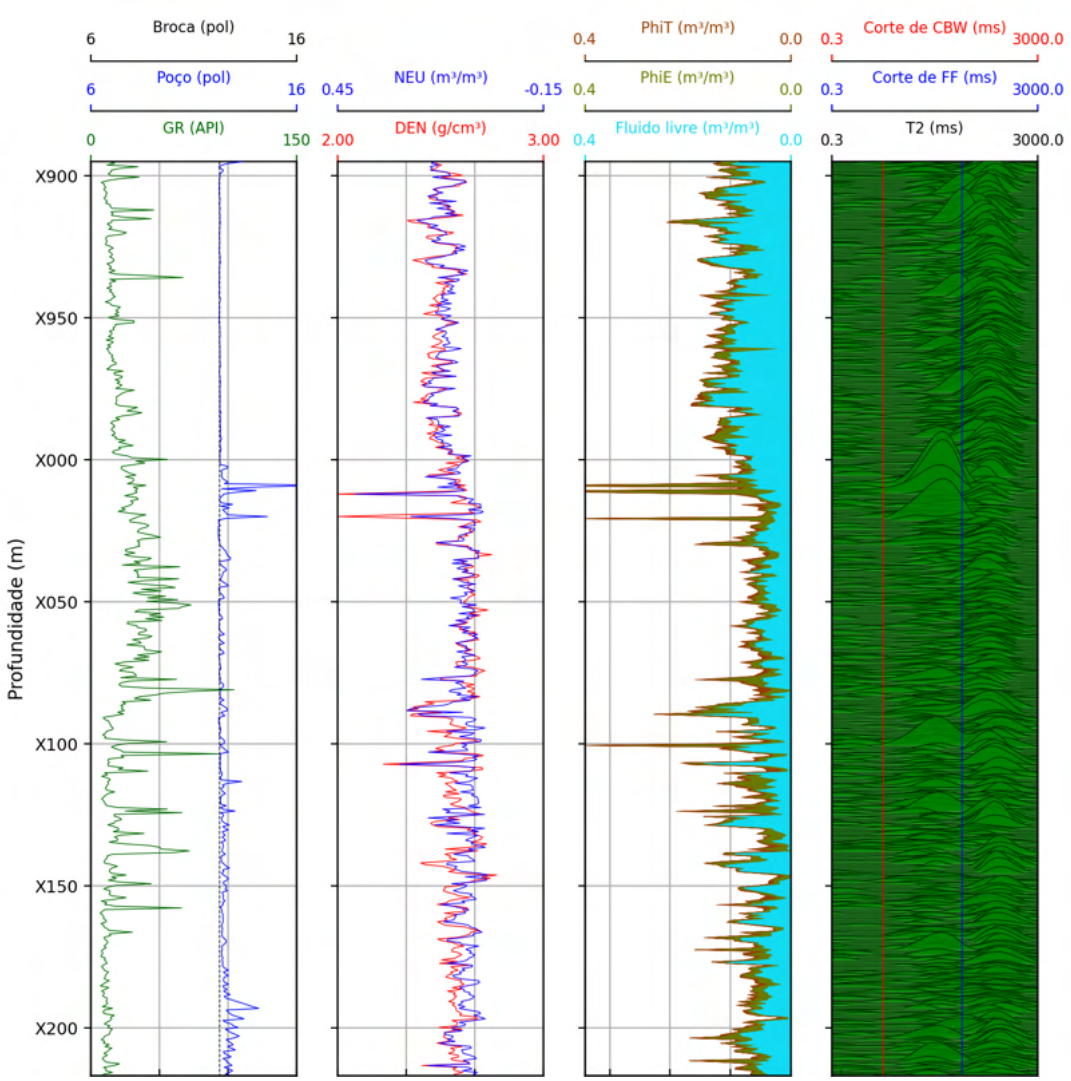


Figura 22 – Exemplo de zonas removidas do modelo no controle de qualidade. Profundidades com arrombamento intenso, evidenciadas pela diferença entre o diâmetro do poço e da broca, afetam os perfis de densidade, nêutrons e RMN.

escolhidos e testados: *Support-Vector Machine* (SVM), *Multilayer Perceptron* (MLP, um tipo de rede neural artificial), *Random Forest*, *Adaptive Boosting* (AdaBoost) e XGBoost. Essa escolha visou abranger os principais algoritmos encontrados na literatura, e as seguintes conclusões foram alcançadas:

1. Os algoritmos SVM e MLP apresentaram os piores resultados. O SVM não foi capaz de capturar a estruturação da base de dados e fazer boas predições. Em relação ao MLP, é possível que uma estrutura de rede neural mais complexa conseguisse apresentar melhores resultados, porém com um incremento de tempo e custo computacional.
2. O *Random Forest*, AdaBoost e XGBoost apresentaram pré-processamento mais simplificado. Por usarem árvores de decisão, esses algoritmos não necessitam de

uma etapa de normalização ou padronização dos dados. Tendo em vista que os perfis de poços apresentam unidades e intervalos de valores muito díspares, o fato de não necessitarem dessas etapas se constitui em uma grande vantagem. Além disso, o ajuste da importância das variáveis ao longo do treinamento desses algoritmos faz com que perfis que poderiam atrapalhar os modelos sejam descartados, simplificando ainda mais o pré-processamento.

3. O AdaBoost e XGBoost apresentaram melhores resultados do que o *Random Forest*. O uso de *boosting* durante o treinamento gerou melhores modelos com menos árvores, reduzindo tempo e custo computacional.

Dessa forma, se verifica que os algoritmos de *ensemble boosting* são os mais indicados para lidarem com informações obtidas em perfilagens. Isso não foi observado na literatura, onde a maioria dos trabalhos utiliza redes neurais artificiais. Dentre esses algoritmos, o AdaBoost e o XGBoost apresentam resultados similares em termos de qualidade de modelo. Como o XGBoost é um algoritmo mais robusto, ele foi escolhido para ser utilizado na etapa de treinamento e validação.

Um total de oito modelos foram treinados para a geração dos perfis geoquímicos sintéticos, um para cada elemento químico. Uma calibração dos hiperparâmetros do algoritmo XGBoost foi realizada na base de treinamento dos modelos, apresentada na tabela 6. Essa calibração foi feita através do algoritmo *GridSearch* da biblioteca *Scikit Learn* escrita em linguagem Python (PEDREGOSA *et al.*, 2011). Uma descrição detalhada dos hiperparâmetros do XGBoost pode ser encontrada em XGBoost-documentação (2016).

Tabela 6 – Hiperparâmetros do algoritmo XGBoost utilizado na criação dos perfis geoquímicos sintéticos.

Hiperparâmetro	Al	Ca	Fe	Mg	Na	Si	S	Ti
Profundidade máxima	50	50	50	50	50	50	25	50
Peso mínimo de prole	5	10	5	10	5	10	1	5
Gama	0	0	0	0	0	0	0	0
Razão de subamostragem	1	0,7	1	1	1	0,8	0,8	0,8
Amostragem de colunas	1	1	1	1	1	1	1	1
Alfa	0	0,01	0	0,01	0	0,01	0	0
Taxa de aprendizado	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1

Fonte – Lucas Oliveira, 2021

Os algoritmos foram treinados na base de treinamento e aplicados na base de validação, avaliados a partir do R^2 e EQM. Adicionalmente, uma validação cruzada do tipo

k-fold foi realizada, onde a base de treinamento é subdividida em *k* partes; *k*-1 partes são usadas para treinamento e a parte restante é usada para a validação do modelo. Esse processo se repete até que todas as partes tenham sido usadas como validação. A média do R^2 e do EQM desses resultados é então calculado. O *k* para o modelo dos perfis geoquímicos sintéticos foi igual a 10. A análise conjunta dos resultados da validação e validação cruzada buscaram garantir maior qualidade e robustez aos modelos, reduzindo *overfitting*.

3.1.4 Avaliação e teste

Os R^2 e EQM da validação e validação cruzada foram utilizados para avaliar a qualidade dos modelos para a geração dos perfis geoquímicos sintéticos.

O desvio padrão do erro da validação de cada modelo foi calculado. Foi considerado erro a diferença entre os valores reais e modelados. Com o desvio padrão, é possível estabelecer intervalos de confiança conforme a figura 23. Em se tratando do erro dos modelos, o desvio padrão serve para indicar que, dado um valor modelado, existe 68% de chance desse valor estar entre mais ou menos um desvio padrão do erro daquele modelo. Para intervalos de confiança mais rigorosos, é possível afirmar que o valor modelado terá 99,7% de chance de estar entre mais ou menos três desvios padrão do erro do modelo. Sendo assim, o desvio padrão do erro da validação pode ser utilizado como uma métrica de incerteza de um modelo de aprendizado de máquina.

A importância das variáveis também foi calculada contando quantas vezes uma variável é utilizada na divisão do nó de uma árvore de decisão e utilizada na compreensão dos fatores que impactaram o treinamento dos modelos dos elementos químicos. Essa informação pode indicar a possibilidade de criação de perfis geoquímicos sintéticos em cenários de maior redução de custos, com a disponibilidade de menos perfis.

Finalmente, os modelos treinados e avaliados foram utilizados na fase de teste, onde foram aplicados em dados de poços que não foram utilizados em nenhuma das etapas anteriores. Isso visou testar de fato a qualidade, robustez e capacidade de generalização dos modelos, simulando um cenário real onde apenas os perfis adquiridos em perfilagens reduzidas estariam disponíveis.

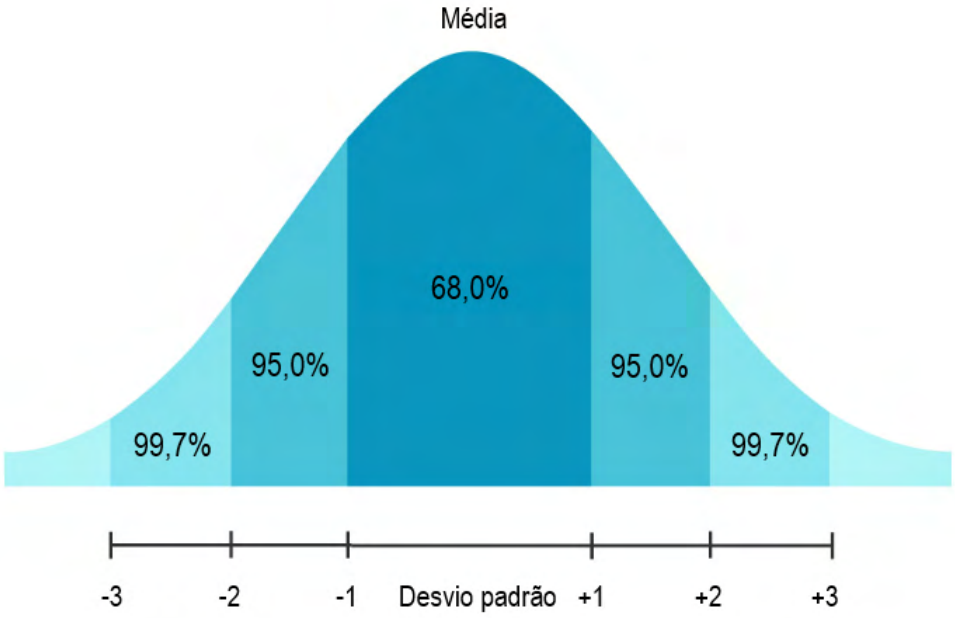


Figura 23 – Histograma mostrando a relação entre desvio padrão e intervalo de confiança. O valor de um desvio padrão representa um intervalo de confiança de 68%, indicando que existe 68% de chance de uma dada propriedade possuir um valor dentro do intervalo de mais um e menos um desvio padrão. Três desvios padrão representam um intervalo de confiança de 99,7%.

3.2 Modelo mineralógico por aprendizado de máquina

3.2.1 Preparação dos dados

A base de dados para a criação do modelo mineralógico foi composta por 1.376 instâncias de amostras de rocha coletadas nos mesmos 22 poços utilizados na geração de perfis geoquímicos sintéticos. Desse total, 77 são amostras de calha coletadas durante a perfuração, 199 são amostras de rocha coletadas durante a perfilagem e 1.100 amostras são plugues de testemunho. A tabela 7 mostra o número de amostras coletadas em cada uma das formações do pré-sal.

Tabela 7 – Resumo das características das amostras coletadas nas formações do pré-sal.

Formação	Número de amostras	Litotipos	Porcentagem
Barra Velha	1059	Carbonatos, ígneos	77,0%
Itapema	288	Carbonatos, siliciclásticos, ígneos	20,8%
Piçarras	9	Siliciclásticos, ígneos	0,7%
Camboriú	20	Ígneos	1,5%

Fonte – Lucas Oliveira, 2021

As amostras foram enviadas para análises de FRX e DRX, onde as quantificações química e mineralógica foram feitas. Os resultados foram concatenados em uma base de dados onde a composição química serviu de entrada e as frações mineralógicas serviram de saída para o modelo.

As análises de FRX foram realizadas no Laboratório de Caracterização Tecnológica (LCT) da Universidade de São Paulo e pela companhia SGS Geosol. A determinação foi realizada pelo método quantitativo, usando curvas de calibração específicas para rochas carbonáticas, com a quantificação de CaO, MgO, SiO_2, Al_2O_3, Fe_2O_3, Na_2O, K_2O, P_2O_5, TiO_2, SrO, MnO e S. A perda ao fogo (PF) foi quantificada a 1.020 ºC por duas horas. As análises foram realizadas pelo equipamento Zetium da PANalytical. A preparação das amostras consistiu de pastilhas fundidas com a adição de tetraborato de lítio.

Para as análises de DRX, um total de 20 g das amostras foi pulverizado para fração menor do que 0,04 mm usando prensa e moedor. Os difratogramas foram adquiridos pelo equipamento X'Pert da PANalytical, com um detector de tubo de cobre (radiação Cu Kα - λ = 1.54186 Å com filtro de Ni) a 45 kV x 40 mA, 2θ variando de 2 a 70º, com passo de 0.02º, 0,20 segundos de tempo entre passo e um tempo total de aquisição de 53 segundos.

A tabela 8 apresenta as principais estatísticas das variáveis usadas para a construção do modelo mineralógico. Os elementos químicos selecionados como entrada do modelo foram os mesmos disponíveis nos perfis geoquímicos, já que o objetivo é sua aplicação em poço. A única exceção foi o S, pois ele não foi adquirido em todas as amostras. Os minerais selecionados foram os mais abundantes no pré-sal, baseados nas suas frações máximas, médias e maiores do que zero. Sendo assim, os minerais selecionados foram calcita, dolomita, quartzo, K-feldspato, argila detrítica, plagioclásio e piroxênio. Esses minerais são os principais componentes das rochas carbonáticas, siliciclásticas e ígneas que compõem o pré-sal. Minerais como dawsonita, pirita, barita, fluorita, hematita e ilmenita também foram identificados nas análises de DRX, porém suas baixas frações impediram sua inclusão no modelo.

É possível observar que os valores máximos e médios dos elementos químicos são coerentes com os observados nos perfis geoquímicos (tabela 5). Essa correspondência sugere que os perfis geoquímicos adquiridos em operações de perfilagem podem ser usados na modelagem geoquímica das rochas do pré-sal. Sendo assim, para a fase de teste, três novos poços foram utilizados. Esses poços possuem perfis geoquímicos e as

amostras coletadas possuem apenas análises de DRX, tornando esses poços excelentes candidatos para a fase de teste.

Tabela 8 – Resumo das principais estatísticas das variáveis de entrada e saída dos modelos de perfis geoquímicos sintéticos.

Nome	Descrição	Unidade	Mínimo	Máximo	Média	Desvio padrão	Não zero
Variáveis de entrada							
Al	Alumínio	g/g	0,000	0,088	0,006	0,013	99,3%
Ca	Cálcio	g/g	0,003	0,402	0,247	0,078	100,0%
Fe	Ferro	g/g	0,000	0,099	0,006	0,013	100,0%
K	Potássio	g/g	0,000	0,090	0,004	0,009	97,1%
Mg	Magnésio	g/g	0,000	0,179	0,045	0,026	100,0%
Na	Sódio	g/g	0,000	0,128	0,003	0,006	99,3%
Si	Silício	g/g	0,000	0,421	0,089	0,071	100,0%
Ti	Titânio	g/g	0,000	0,021	0,001	0,002	95,6%
Variáveis de saída							
PF	Perda ao fogo	g/g	0,008	0,48	0,35	0,09	100,0%
Cal	Calcita	g/g	0,00	1,00	0,48	0,25	93,8%
Dol	Dolomita	g/g	0,00	1,00	0,29	0,21	94,7%
Qtz	Quartzo	g/g	0,00	0,98	0,16	0,14	93,4%
Kfd	K-Feldspato	g/g	0,00	0,64	0,02	0,05	26,7%
Clay	Argila detrítica	g/g	0,00	0,48	0,03	0,07	25,1%
Plg	Plagioclásio	g/g	0,00	0,70	0,01	0,07	5,1%
Prx	Piroxênio	g/g	0,00	0,34	0,00	0,03	3,3%

Fonte – Lucas Oliveira, 2021

3.2.2 Treinamento e validação

A base de dados foi aleatoriamente dividida em 70% para treinamento e 30% para validação, totalizando 963 instâncias na base de treinamento e 413 instâncias na base de validação.

O algoritmo utilizado também foi o XGBoost. Além dos minerais apresentados na tabela 8, modelos para PF, carbonatos (calcita + dolomita) e piroxênio + plagioclásio também foram treinados, totalizando dez modelos. A PF não é um propriedade adquirida em operações de perfilagem e pode oferecer informações importantes quanto a composição da rocha. Os carbonatos apresentam alta correlação positiva com a PF e podem ser mais facilmente obtidos do que apenas calcita e dolomita. Os plagioclásios e piroxênios podem aparecer associados em rochas ígneas e, já que possuem baixas frações, sua combinação pode apresentar melhores resultados durante o treinamento.

Os resultados da calibração dos hiperparâmetros do algoritmo XGBoost é apresentada na tabela 9.

Tabela 9 – Hiperparâmetros do algoritmo XGBoost utilizado na criação do modelo mineralógico.

Hiperparâmetro	PF	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Prx+Plg
Profundidade máxima	25	15	25	25	15	15	50	15	15	15
Peso mínimo de prole	10	1	5	5	5	1	10	1	10	10
Gama	0	0,01	0,001	0,01	0	0,01	0	0	0,001	0,01
Razão de subamostragem	1	0,7	0,7	0,7	0,6	0,6	0,7	0,8	1	1
Amostragem de colunas	0,7	1	1	0,8	1	1	1	1	0,7	0,7
Alfa	0	1	0	0	0,1	0	0,01	1	1	0,1
Taxa de aprendizado	0,1	0,1	0,1	0,1	0,1	0,1	0,1	1	0,1	1

Fonte – Lucas Oliveira, 2021

Os algoritmos foram treinados na base de treinamento e aplicados na base de validação, avaliados a partir do R^2 e EQM. Uma validação cruzada *k-fold* foi realizada, com *k* igual a 5.

A soma das frações minerais que compõem uma rocha é igual a 100%, fazendo com que a composição mineralógica seja um sistema fechado. Logo, a fração de um determinado mineral irá impactar na estimativa dos demais, sendo uma informação importante no treinamento dos modelos. Sendo assim, uma nova técnica chamada de aprendizado escalonado é proposta para a criação do modelo mineralógico.

O aprendizado escalonado consiste na adição dos resultados de um modelo mineral nas variáveis de entrada de um modelo subsequente. Entretanto, essa inclusão não pode ser feita de forma total e indiscriminada, uma vez que os erros de um modelo podem ser propagados para os demais. Os melhores resultados serão atingidos através da união dos conhecimentos geológico e de aprendizado de máquina, fazendo com que o aprendizado escalonado faça sentido do ponto de vista de assembleia mineral e que isso reflita na qualidade do modelo adquirido.

O passo-a-passo do aprendizado escalonado se deu da seguinte forma:

1. Modelo de PF: como a PF não é adquirida por ferramentas de perfilagem, um modelo foi treinado através do algoritmo XGBoost usando as concentrações dos elementos químicos. Em uma situação de poço real, essas concentrações serão as adquiridas pela ferramenta geoquímica.
2. Modelo de carbonatos: a PF foi adicionada às variáveis de entrada e um modelo para carbonatos foi criado. Os carbonatos foram considerados a soma de calcita e dolomita. Como a PF apresenta alta correlação com os carbonatos, é esperado que essa inclusão melhore a qualidade do modelo.
3. Modelos minerais: a fração de carbonatos foi adicionada às variáveis de entrada junto com a PF e modelos para os demais minerais foram treinados. Como os carbonatos compõem uma fração significativa das rochas do pré-sal, sua inclusão pode melhorar significativamente os demais modelos.
4. Modelo de argila detrítica: devido a complexidade das argilas detríticas, a fração de quartzo foi adicionada às variáveis de entrada junto aos carbonatos e PF para o treinamento de um modelo de argilas detríticas. É esperado que essa base de treinamento forneça informações quanto a composição das argilas detríticas e ao ambiente no qual ela foi depositado, melhorando significativamente os resultados desse modelo.

A figura 24 apresenta um resumo do aprendizado escalonado proposto nesse trabalho. O aprendizado escalonado não pôde ser empregado na criação dos perfis geoquímicos sintéticos uma vez que as concentrações dos elementos químicos fornecidas não abrangem a totalidade dos elementos que compõem a matriz da rocha, não caracterizando um sistema fechado.

3.2.3 Avaliação e teste

Os R^2 e EQM da validação e validação cruzada foram utilizados para avaliar a qualidade dos modelos. Adicionalmente, os resultados do modelo mineralógico antes e depois da aplicação do aprendizado escalonado foram comparados, para definir a melhor sequência de treinamento. Assim como nos perfis geoquímicos sintéticos, o desvio padrão do erro da validação foi utilizado como uma métrica de incerteza dos modelos minerais e a importância das variáveis calculada contando quantas vezes uma variável foi utilizada

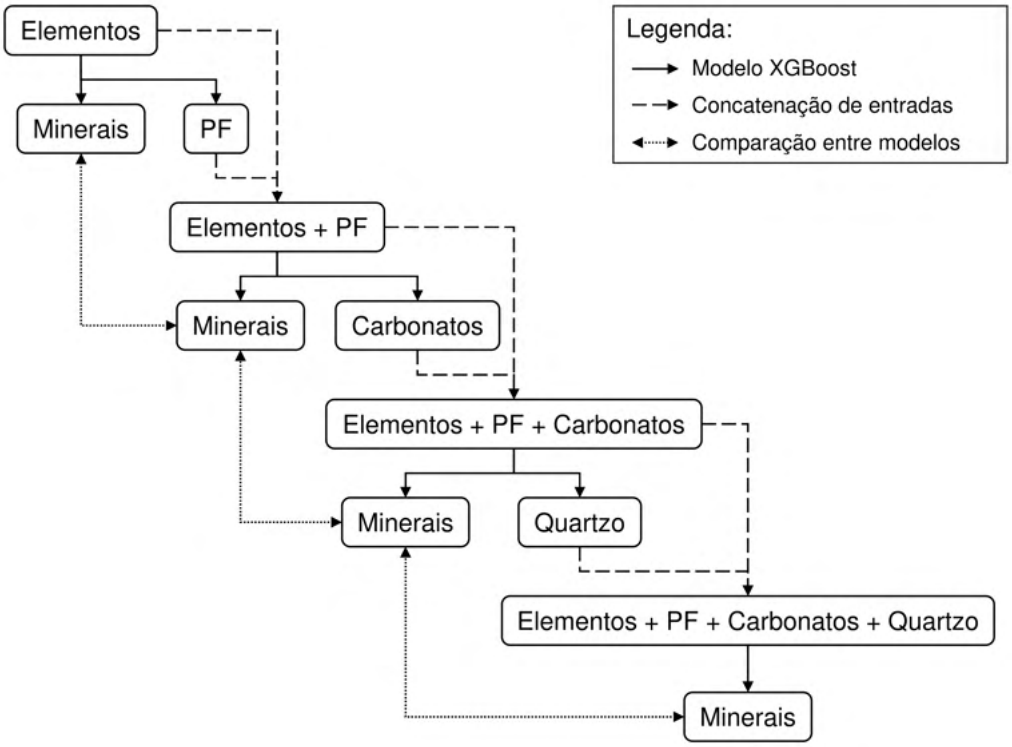


Figura 24 – Aprendizado escalonado proposto para a criação do modelo mineralógico. Setas sólidas indicam o uso do algoritmo de aprendizado de máquina para a estimativa das frações de minerais e PF. Setas tracejadas indicam a inclusão dos resultados de um modelo às variáveis de entrada do modelo anterior. Setas pontilhadas indicam a comparação entre estimativas de diferentes passos.

na divisão do nó de uma árvore de decisão foi utilizada na compreensão dos fatores que impactaram o treinamento dos modelos.

Finalmente, os modelos treinados e avaliados foram utilizados na fase de teste, onde foram aplicados nos perfis geoquímicos de poços que possuíam amostras apenas com análises de DRX, diferentes dos usados na modelagem geoquímica. As frações minerais obtidas através dos modelos de aprendizado de máquinas foram comparadas com os resultados das análises de DRX para testar a qualidade e robustez dos modelos em cenários reais.

3.3 Modelo mineralógico híbrido

Mesmo sem analisar seus resultados, é possível observar que a metodologia proposta para a criação do modelo mineralógico através de aprendizado de máquina possui limitações relacionadas ao enviesamento da base de dados. Como mais de 90% das

amostras foram coletadas nas Formações Barra Velha e Itapema (tabela 7), o modelo irá apresentar maior confiabilidade nessas formações, compostas majoritariamente por calcita, dolomita e quartzo. Outra limitação é que frações altas de plagioclásio e piroxênio terão alta incerteza, uma vez que quantidades acima de zero foram observadas em menos de 10% das amostras para esses minerais (tabela 8). Além disso, minerais como as argilas magnesianas, pirita e barita não foram contemplados por não aparecerem em quantidades expressivas na base de dados. A pirita é um importante mineral na caracterização de reservatórios por apresentar baixas resistividades, impactando os cálculos de saturação de água por ferramentas de resistividade (HAMADA; AL-AWAD, 2000; HAMADA; AL-AWAD; ALMALIK, 2001; PRATAMA; ISMAIL; RIDHA, 2017). A identificação de barita natural da formação auxilia na identificação de zonas permo-porosas danificadas pelo fluido de perfuração adensado com baritina (IBRAHIM; SAMI; BALASUBRAMANIAN, 2017; BAGERI *et al.*, 2019; IBRAHIM *et al.*, 2020).

A principal maneira de lidar com o enviesamento da base de dados seria através da coleta de mais amostras de rocha e realização de mais análises laboratoriais. Porém, essa estratégia pode gerar muitos custos, indo contra as políticas de otimização da aquisição de dados. Sendo assim, alternativas para além do universo de aprendizado de máquina se fazem necessárias.

A abordagem híbrida aqui proposta consiste na combinação dos modelos de aprendizado de máquina com modelos probabilísticos que utilizam perfis de poços. As frações minerais estimadas pelos algoritmos de aprendizado de máquina são usadas em conjunto com os perfis de densidade, fator fotoelétrico, RMN e geoquímicos em uma etapa probabilística. Essas frações são usadas como estimativa inicial da etapa probabilística, além de se tornarem equações de reconstrução. Um minimizador não linear busca a menor diferença entre as curvas reais e reconstruídas, ajustando as frações volumétricas dos componentes da formação. As equações de reconstrução são ponderadas pelas incertezas dos perfis de poço e pelo erro médio dos modelos de aprendizado de máquina.

3.3.1 Etapa de concatenação

Os modelos minerais treinados na etapa de aprendizado de máquina são aplicados às concentrações de elementos químicos adquiridos pela ferramenta geoquímica, gerando

perfis de frações mássicas dos sete minerais utilizados como saída desses modelos. Essas frações são concatenadas com frações de pirita, barita, argilas magnesianas e fluidos e são utilizadas nas equações de reconstrução da etapa probabilística. As frações iniciais de fluidos são obtidas pela ferramenta de RMN, sendo divididas em fluido livre, água capilar e água de argilas. As frações iniciais de pirita, barita e argilas magnesianas são as médias encontradas nas análises de DRX.

O fato de as frações minerais serem mássicas e as de fluido serem volumétricas não inviabiliza sua utilização em conjunto como estimativa inicial da etapa probabilística, uma vez que elas serão convertidas em frações volumétricas, conforme discutido na próxima seção. Por se tratar de um processo iterativo, essas frações de fato ajudam a minimização por serem melhores do que uma estimativa totalmente aleatória.

Além de serem usados como estimativa inicial, os perfis de frações minerais obtidos pelo aprendizado de máquina também são concatenados aos perfis de densidade, fator fotoelétrico, RMN e geoquímicos. Esses perfis são usados na função custo da etapa probabilística. A figura 25 apresenta uma sequência detalhada da etapa de concatenação.

3.3.2 Etapa probabilística

Na etapa probabilística, as frações dos componentes minerais e de fluido são usadas em equações de reconstrução para gerar perfis reconstruídos. Os perfis reais e reconstruídos são ponderados pelas suas incertezas e comparados profundidade a profundidade. A diferença entre os perfis reais e reconstruídos em função das frações volumétricas dos componentes é minimizada através de um processo iterativo até que os resultados sejam satisfatórios. As frações minerais e de fluido finais são as que geram a menor diferença entre os perfis reais e reconstruídos. Um detalhamento dos perfis e componentes utilizados na etapa probabilística é apresentado na tabela 10.

As equações de reconstrução foram escolhidas de forma que os valores de referência dos componentes sejam conhecidos com boa confiabilidade, razão pela qual perfis como os acústicos ou de nêutrons não foram utilizados. As equações de reconstrução utilizadas foram as de densidade, índice de absorção fotoelétrica volumétrica, de frações de fluido e de composição química da matriz da rocha, relacionadas aos perfis de densidade, fator fotoelétrico, de RMN e geoquímicos, respectivamente. Além dessas equações, também

Tabela 10 – Variáveis da etapa probabilística.

Perfil	Equação	Descrição	Unidade
		Perfis	
DEN	ρ_b	Densidade	g/cm^3
PEF	P_e	Fator fotoelétrico	-
U	U_T	Índice de absorção fotoelétrica volumétrica	barns/cm^3
PhiT	ϕ_T	Porosidade total (RMN)	m^3/m^3
CBW	ϕ_{CBW}	Porosidade de argila (RMN)	m^3/m^3
FF	ϕ_{FF}	Fluido livre (RMN)	m^3/m^3
Al	w_{Al}	Concentração mássica de Al	g/g
Ca	w_{Ca}	Concentração mássica de Ca	g/g
Fe	w_{Fe}	Concentração mássica de Fe	g/g
K	w_K	Concentração mássica de K	g/g
Mg	w_{Mg}	Concentração mássica de Mg	g/g
Si	w_{Si}	Concentração mássica de Si	g/g
S	w_S	Concentração mássica de S	g/g
Ti	w_{Ti}	Concentração mássica de Ti	g/g
Cal	W_{Cal}	Fração mássica de calcita	g/g
Dol	W_{Dol}	Fração mássica de dolomita	g/g
Qtz	W_{Qtz}	Fração mássica de quartzo	g/g
Kfd	W_{Kfd}	Fração mássica de K-feldspato	g/g
Clay	W_{Clay}	Fração mássica de argila detrítica	g/g
Plg+Prx	$W_{Plg+Prx}$	Fração mássica de plagioclásio e piroxênio	g/g
		Componentes volumétricos	
V_{Cal}		Fração de calcita	m^3/m^3
V_{Dol}		Fração de dolomita	m^3/m^3
V_{Qtz}		Fração de quartzo	m^3/m^3
V_{Kfd}		Fração de K-feldspato	m^3/m^3
V_{Clay}		Fração de argila detrítica	m^3/m^3
V_{Plg}		Fração de plagioclásio	m^3/m^3
V_{Prx}		Fração de piroxênio	m^3/m^3
V_{Pir}		Fração de pirita	m^3/m^3
V_{Bar}		Fração de barita	m^3/m^3
V_{Mgclay}		Fração de argila magnesiana	m^3/m^3
V_{FF}		Fração de fluido livre	m^3/m^3
V_{Wat}		Fração de água capilar	m^3/m^3
V_{CBW}		Fração de água de argila	m^3/m^3

Fonte – Lucas Oliveira, 2021

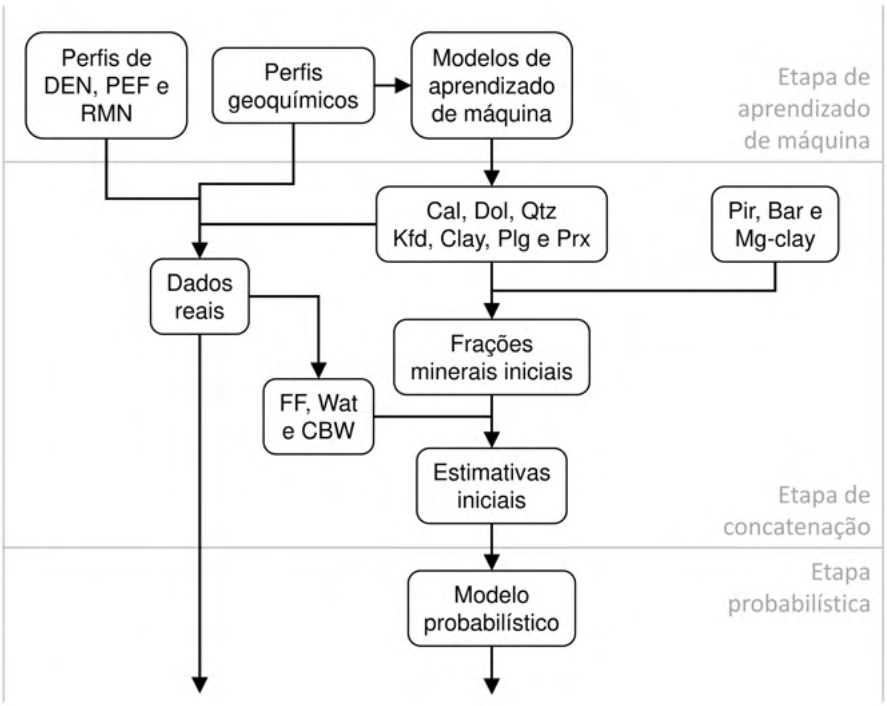


Figura 25 – Sequência focada na etapa de concatenação do modelo híbrido. As frações minerais geradas pelo aprendizado de máquina são concatenadas às frações médias de pirita, barita e argilas magnesianas e usadas como frações minerais iniciais. Em paralelo, as frações geradas pelo aprendizado de máquina são também adicionadas às curvas reais para serem comparadas às equações de reconstrução. A ferramenta de RMN fornece as frações de fluidos que são concatenadas às frações minerais, gerando a estimativa de componentes inicial usada no processo iterativo da etapa probabilística.

foram utilizadas as frações minerais da etapa de aprendizado de máquina e da restrição de unidade. Dentre os minerais, optou-se por utilizar o somatório das frações de plagioclásio e piroxênio, uma vez que essas combinação pode apresentar melhores resultados na etapa de aprendizado de máquina.

Na restrição de unidade, a soma das frações volumétricas dos minerais e fluidos deve compor um valor unitário, segundo a equação 17.

$$\sum_{l=1}^{c} V_l = 1 \tag{17}$$

Onde c é o número de componentes e, V é a fração volumétrica do componente l.

Na equação de densidade, a relação entre a densidade *bulk* da formação e as frações volumétricas dos minerais e fluidos que a compõe pode ser escrita de forma linear, conforme a equação 18.

$$\sum_{l=1}^{c} \rho_l V_l = \rho_b \tag{18}$$

Onde ρ_b é a densidade *bulk* e ρ é a densidade do componente l.

Os valores de densidade de cada componente são apresentados na tabela 11. Os valores para os minerais foram extraídos da Mineralogy-Database (2021). Análises de microssonda eletrônica em amostras de rocha indicam que as argilas detríticas encontradas no pré-sal da Bacia de Santos são filossilicatos alterados de composição ferro-magnesiana. Sendo assim, suas propriedades foram as mesmas da biotita. A densidade das argilas foi considerada em seu estado dsidratado, já que a fração de água de argila foi considerada separadamente. A densidade do fluido livre pode variar conforme a composição do fluido de perfuração, podendo ser 0,80 g/cm^3 no caso de fluido base óleo ou 1,00 g/cm^3 no caso de fluido base água. As densidades da água capilar e de argila foram calculadas para uma salinidade de 220.000 ppm de NaCl, conforme amostras de água coletadas nos reservatórios do pré-sal da Bacia de Santos.

Tabela 11 – Densidade dos minerais e fluidos usados na etapa probabilística.

Componente	Densidade (g/cm^3)
Calcita	2,71
Dolomita	2,83
Quartzo	2,65
K-Feldspato	2,54
Argila detrítica (desidratada)	2,83
Plagioclásio	2,67
Piroxênio	3,39
Pirita	4,84
Barita	3,99
Argila magnesiana (desidratada)	2,75
Fluido livre (fluido de perfuração base óleo)	0,80
Fluido livre (fluido de perfuração base água)	1,00
Água capilar (salinidade de 220.000 ppm de NaCl)	1,10
Água de argila (salinidade de 220.000 ppm de NaCl)	1,10

Fonte – Lucas Oliveira, 2021

Como o fator fotoelétrico de diferentes componentes não se combina volumetricamente, o índice de absorção fotoelétrica volumétrica foi calculada através da multiplicação do fator fotoelétrico pela densidade (equação 19).

$$\rho_b P_e = U_T \tag{19}$$

Onde P_e é o fator fotoelétrico e U_T é o índice de absorção fotoelétrica volumétrica. Essa propriedade pode ser linearmente combinada segundo a equação 20, de reconstrução do índice volumétrico.

$$\sum_{l=1}^{c} U_l V_l = U_T \tag{20}$$

Onde U é o índice de absorção fotoelétrica volumétrica do componente l.

Os valores de fator fotoelétrico e índice de absorção fotoelétrica volumétrica de cada componente são apresentados na tabela 12. Os valores para os minerais foram extraídos da Mineralogy-Database (2021). Para os fluidos, foram usados resultados de análises laboratoriais.

Tabela 12 – Fator fotoelétrico e índice de absorção fotoelétrica volumétrica dos minerais e fluidos usados na etapa probabilística.

Componente	PEF	U (barns/cm^3)
Calcita	5,06	13,70
Dolomita	3,13	8,83
Quartzo	1,80	4,77
K-Feldspato	2,85	7,26
Argila detrítica (desidratada)	3,13	8,83
Plagioclásio	2,46	6,57
Piroxênio	3,51	11,73
Pirita	16,89	79,34
Barita	265,56	1036,61
Argila magnesiana (desidratada)	1,57	4,34
Fluido livre (fluido de perfuração base óleo)	-	0,10
Fluido livre (fluido de perfuração base água)	-	1,30
Água capilar (salinidade de 220.000 ppm de NaCl)	-	1,45
Água de argila (salinidade de 220.000 ppm de NaCl)	-	1,45

Fonte – Lucas Oliveira, 2021

As equações de frações de fluido buscaram reconstruir os perfis de RMN. A primeira é a da porosidade total (equação 21), sendo o somatório de todas as frações de fluido.

$$V_{FF} + V_{Wat} + V_{CBW} = \phi_T \tag{21}$$

Onde V_{FF}, V_{Wat} e V_{CBW} são as frações de fluido livre, água capilar e água de argila, respectivamente, e ϕ_T é a porosidade total fornecida pela ferramenta de RMN, sendo calculada como a integral da distribuição do tempo de relaxação T_2 lido em uma profundidade.

A segunda equação de reconstrução é a do próprio fluido livre, fração fornecida diretamente pela ferramenta (equação 22).

$$V_{FF} = \phi_{FF} \tag{22}$$

Onde ϕ_{FF} é o fluido livre. O fluido livre em carbonatos é a soma dos valores de T_2 acima do corte de 100 ms. Essa fração é importante pois as propriedades do fluido livre variam de acordo com o filtrado do fluido de perfuração utilizado, uma vez que a ferramenta de RMN realiza medições próximas a parede do poço (SCHLUMBERGER, 2015).

A última equação de reconstrução fornecida pela ferramenta de RMN é a de água de argila, sendo o somatório das frações de argilas multiplicadas pelas suas porosidades(equação 23)

$$V_{Clay}\phi_{Clay} + V_{Mgclay}\phi_{Mgclay} = \phi_{CBW} \tag{23}$$

Onde V_{Clay} e V_{Mgclay} são as frações de argilas detríticas e magnesianas respectivamente, e ϕ_{Clay} e ϕ_{Mgclay} são as porosidades de argilas detríticas e magnesianas respectivamente. A água das argilas nos carbonatos do pré-sal da Bacia de Santos é a soma dos valores de T_2 abaixo de 3 ms e análises de relaxometria demonstram que a porosidade média das argilas do pré-sal é de 10% (JUNIOR, 2019). Dessa forma, a soma das frações de argilas detríticas e magnesianas ficam atreladas à água das argilas fornecida pela ferramenta de RMN. Como o modelo de aprendizado de máquina é capaz de estimar a fração de argila detrítica, essa equação de reconstrução é uma das principais na estimativa da fração de argilas magnesianas.

As equações de composição química da matriz da rocha buscaram reconstruir os perfis geoquímicos. Como as frações dos minerais são volumétricas e as concentrações dos elementos químicos fornecidas pela ferramenta geoquímica estão são mássicas, primeiramente é calculado a fração mássica dos minerais através da equação 24.

$$\frac{V_l \rho_l}{(1-\phi_T)\rho_{ma}} = W_l \tag{24}$$

Onde W é a fração mássica do mineral l e ρ_{ma} é a densidade da matriz da rocha. Essa densidade é calculada pelo somatório das densidades dos minerais ponderadas pelas suas respectivas frações, conforme a equação 25.

$$\sum_{l=1}^{g} \rho_l V_l = \rho_{ma} \tag{25}$$

Onde g é o total de minerais do modelo. A porosidade total ϕ_T é a mesma da equação 21.

Com as frações mássicas dos minerais, a concentração mássica de um elemento químico na matriz da rocha é dada pelo somatório da concentração mássica desse elemento em todos os minerais que compõem a rocha ponderada pela massa dos minerais (equação 26).

$$\sum_{l=1}^{g} w_l W_l = w_T \tag{26}$$

Onde w_T é a concentração mássica total de um elemento químico na matriz da rocha, w é a concentração mássica do elemento no mineral l para g minerais.

A tabela 13 apresenta a fórmula química dos minerais e as respectivas concentrações mássicas dos elementos químicos usados no modelo híbrido. Com exceção das argilas detríticas e magnesianas, cujas fórmulas químicas foram obtidas por análises de microscópio eletrônico de varredura, as fórmulas químicas dos demais minerais foram extraídas da Mineralogy-Database (2021).

Com as frações mássicas dos minerais calculadas através da equação 25, esses valores podem ser diretamente comparados com os obtidos na etapa de aprendizado de máquina. Dessa forma, além de honrar os perfis adquiridos no poço, a etapa probabilística também precisa levar em consideração as frações minerais estimadas na etapa de aprendizado de máquina. Como os algoritmos foram treinados utilizando as análises de FRX e DRX de amostras de rocha, a etapa probabilística carregará essa informação ao gerar suas estimativas.

A tabela 14 resume os perfis de referência utilizados e suas respectivas equações de reconstrução, bem como suas referências ao longo do texto.

Tabela 13 – Fórmula química dos minerais e as respectivas frações mássicas dos elementos químicos usados na etapa probabilística.

Minerais	Concentração mássica							
	Al	Ca	Fe	K	Mg	Si	S	Ti
Calcita $CaCO_3$	-	0.40	-	-	-	-	-	-
Dolomita $CaMg(CO_3)_2$	-	0.22	-	-	0.13	-	-	-
Quartzo SiO_2	-	-	-	-	-	0.47	-	-
K-Feldspato $KAlSi_3O_8$	0.10	-	-	0.14	-	0.30	-	-
Argila detrítica $(Na_{0.09}K_{1.69}Ca_{0.04})Si_{5.60}$ $(Mg_{2.15}Al_{2.64}Fe_{2.49}Ti_{0.61})$ $O_{20}(OH)_4$	0.08	0.002	0.15	0.07	0.06	0.17	-	0.03
Plagioclásio $NaCa(Si_3AlO_8)_2$	0.10	0.07	-	-	-	0.31	-	-
Piroxênio $Ca_3Mg_{6.6}Al_{0.75}Fe_{0.65}$ $(Si_{10}AlO_{32})(OH)_2$	0.04	0.10	0.03	-	0.13	0.24	-	-
Pirita FeS_2	-	-	0.47	-	-	-	0.53	-
Barita $BaSO_4$	-	-	-	-	-	-	0.14	-
Argilas magnesianas $(Na_{0.17}K_{0.06}Ca_{0.04})Si_{8.04}$ $(Mg_{5.74}Al_{0.29}Fe_{0.06})$ $O_{20}(OH)_4$	0.01	0.002	0.004	0.003	0.18	0.29	-	-

Fonte – Lucas Oliveira, 2021

Tabela 14 – Resumo das propriedades e equações de reconstrução utilizadas na etapa probabilística e suas referências no texto.

Referência	Perfil (p)	Equação (f)	Descrição
17	1	-	Restrição de unidade
18	DEN	ρ_b	Densidade
20	U	U_T	Índice de absorção fotoelétrica volumétrica
21	PhiT	ϕ_T	Porosidade total
22	FF	ϕ_{FF}	Fluido livre
23	CBW	ϕ_{CBW}	Água de argila
26	EQ	w	Concentrações de elementos químicos
24	MIN	W	Frações minerais do aprendizado de máquina

Fonte – Lucas Oliveira, 2021

Após as frações volumétricas dos componentes passarem pelas equações e gerarem perfis reconstruídos, a diferença entre eles e os perfis reais é calculada profundidade a profundidade através da equação 27

$$\Delta = \sqrt{\frac{1}{t}\sum_{k=1}^{t}\left(\frac{p_k - f_k}{\epsilon_k}\right)^2} \tag{27}$$

Onde Δ é a diferença entre os perfis adquiridos e reconstruídos, t é o número de equações, p é o valor do perfil adquirido k, f é o valor do perfil reconstruído k, e ϵ é o desvio padrão associado a incerteza atribuída ao perfil k.

A escolha da incerteza dos perfis precisa levar em consideração diversos fatores. Como um dos seus principais objetivos é a padronização dos dados, primeiramente o desvio padrão do perfil observado no poço deve ser usado. Dessa forma, nenhuma curva terá maior relevância na função custo por questões de unidade de medida. Em seguida, é aconselhável ponderar a incerteza pela precisão das ferramentas de perfilagem, que pode ser extraído do catálogo de ferramentas (SCHLUMBERGER, 2015). Essa precisão irá variar entre diferentes ferramentas e para diferentes condições de poço. Em relação às frações minerais estimadas na etapa de aprendizado de máquina, o desvio padrão do erro do conjunto de validação pode ser utilizado. Finalmente, a confiabilidade das equações de reconstrução pode ser usada como um critério adicional na definição do valor da incerteza.

Como o valor da incerteza de cada curva pode variar entre diferentes poços, valores únicos não podem ser facilmente fornecidos. Entretanto, critérios gerais podem ser sugeridos. Os critérios utilizados neste trabalho são descritos nos tópicos a seguir.

- Equação de unidade: como a equação de unidade obriga que os componentes da formação tenham somatório igual a um, ela apresentou a menor incerteza.
- Perfil de densidade: o perfil de densidade apresenta alta precisão e confiabilidade em relação às equações de reconstrução. Como a aquisição da ferramenta é realizada junta a parede do poço, ele é moderadamente impactado por arrombamento de poço. Dessa forma, ele é um dos perfis com menor incerteza.
- Perfil U: por se tratar da multiplicação entre o perfil de densidade e fator fotoelétrico, a precisão do perfil U precisa levar em conta a precisão desses dois perfis, aumentando a incerteza. Além disso, o perfil de fator fotoelétrico é muito impactado pela presença

de baritina no fluido de perfuração. A incerteza do perfil U foi então diretamente proporcional a qualidade do perfil de fator fotoelétrico.

- Perfis de RMN: a precisão da porosidade total do perfil de RMN é alta, apresentando incerteza equivalente ao da densidade. Entretanto, o fluido livre e a água de argila apresentam maior incerteza, uma vez que estão condicionadas aos cortes de T_2.
- Perfis geoquímicos: os perfis geoquímicos são fortemente impactados por condições de poço, principalmente arrombamento e salinidade do fluido de perfuração. Além disso, cada elemento químico tem sua própria incerteza, relacionada a sensibilidade da ferramenta e a concentração do elemento na formação. Sendo assim, os perfis geoquímicos apresentaram incerteza equivalente à do perfil U.
- Modelo de aprendizado de máquina: além do desvio padrão do erro observado no conjunto de validação, as frações minerais do modelo de aprendizado de máquina precisam levar em consideração a incerteza dos perfis geoquímicos, uma vez que eles são usados em sua geração. Somado a isso, o viés da base de dados também foi considerado, aumentando a incerteza dos minerais com menor representatividade. Logo, as curvas de referência minerais apresentaram as maiores incertezas da etapa probabilística.

A função custo é então minimizada até que a diferença seja considerada satisfatória. O minimizador utilizado é o *Trust Region Reflective* implementado na biblioteca Scipy (BRANCH; COLEMAN; LI, 1999; VIRTANEN *et al.*, 2020), capaz de lidar com restrições como não-negatividade. As frações minerais e de fluido obtidas são consideradas as mais representativas da formação. O resíduo da função custo gera um perfil de erro global, utilizado para identificar regiões do poço onde o modelo híbrido não apresentou bom desempenho.

3.3.3 Etapa de avaliação

O modelo mineral híbrido foi aplicado em três poços perfurados no pré-sal da Bacia de Santos, diferentes dos usados na modelagem geoquímica e nos modelo mineralógico por aprendizado de máquina. Esses poços foram diferentes dos da etapa de teste do modelo mineralógico por aprendizado de máquina e contêm quantidades significativas de argilas magnesianas. As frações minerais estimadas foram comparadas com análises de DRX de

amostras de rocha coletadas nesses poços. Como esses poços não foram utilizados na etapa de aprendizado de máquina, suas análises de DRX não enviesam os resultados do modelo híbrido. Além disso, os perfis reconstruídos foram comparados com os reais e suas respectivas incertezas, para avaliar o quanto as equações de reconstrução foram capazes de honrar os dados reais.

4 Resultados

4.1 Análise exploratória das bases de dados

A correlação entre as variáveis usadas no modelo para geração de perfis geoquímicos sintéticos é apresentada na figura 26. Observa-se que as porosidades das ferramentas de RMN e de nêutrons apresentam correlação positiva com as vagarosidades das ondas P e S e correlação negativa com a densidade, evidenciando o efeito da porosidade nessas propriedades. A alta correlação positiva entre o U e os raios gama demonstra que esse é o principal elemento químico responsável pela radioatividade dos carbonatos do pré-sal.

O Ca e Si apresentam uma alta correlação negativa, reflexo de dois processos: a substituição de rochas carbonáticas por siliciclásticas e ígneas, quando da passagem das Formações Barra Velha e Itapema para as Formações Piçarras e Camboriú, e a silicificação de rochas carbonáticas, comum na Formação Barra Velha. Essa correlação negativa também é observada entre Ca e Mg, reflexo dos processos diagenéticos de dolomitização dos carbonatos.

Os elementos K, Al, Fe e Ti apresentem alta correlação positiva entre si e correlação negativa com Ca, associado a presença desses elementos em minerais que compõem rochas siliciclásticas e ígneas. O Na e S não apresentaram alta correlação com nenhuma variável, possivelmente relacionado às baixas concentrações desses elementos nas rochas do pré-sal.

As figuras 27, 28 e 29 apresentam exemplos de perfis de alguns poços da base de dados, demonstrando as principais feições observadas nas rochas do pré-sal da Bacia de Santos.

O poço da figura 27 perfurou as Formações Barra Velha e Itapema. No contato entre elas, se observa rochas siliciclásticas de alta radioatividade e alto teor de K, Si, Fe e Al, em parte descritas como folhelhos (aproximadamente X650 m). A Formação Barra Velha apresenta carbonatos com alta concentração de Mg no topo, possivelmente reflexo de dolomitização. O Mg diminui em profundidades próximas aos folhelhos. Em aproximadamente X480 m se observa uma zona de baixo T_2 no perfil RMN, demarcada pela presença de água de argila. Não há mudança na concentração dos elementos químicos, indicando rochas carbonáticas. De fato, essa é uma zona de argilas magnesianas (HERLINGER *et al.*, 2020). Essas argilas não apresentam composição siliciclástica típica, como o aumento expressivo

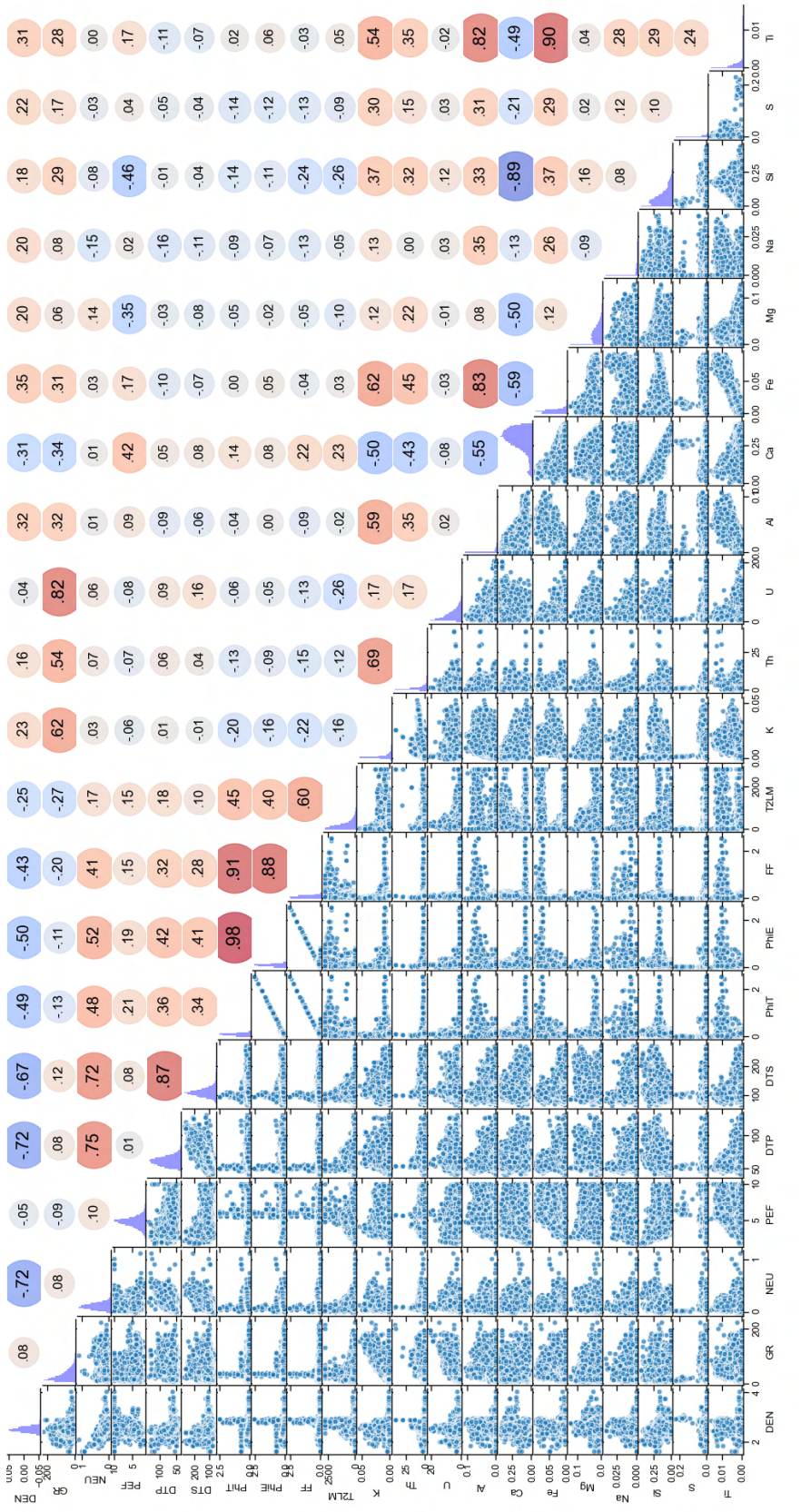


Figura 26 – Correlação entre as variáveis do modelo para geração de perfis geoquímicos sintéticos, apresentada na forma de gráficos e R^2. Cores quentes indicam correlação positiva, enquanto cores frias indicam correlação negativa. Histogramas das variáveis são apresentados na diagonal da figura.

de K, Al, Fe e Si, e podem ser facilmente confundidas com carbonatos. A Formação Itapema apresenta baixa radioatividade e composição francamente calcítica, demarcada pelas altas concentrações de Ca. Ela também possui as maiores porosidades observadas no poço.

O poço apresentado na figura 28 encontrou as Formações Barra Velha, Itapema, Piçarras e Camboriú. A Formação Barra Velha possui carbonatos pouco silicificados e dolomitizados. Em aproximadamente X360 m, observa-se uma diminuição da radiação gama e um aumento da concentração de Ca, demarcando o início da Formação Itapema, de composição calcítica. Um aumento das concentrações de K, Th, Al, Fe e Si é observado a partir de X490 m, acompanhado por uma diminuição da porosidade do perfil RMN. Essa feição marca o início da Formação Piçarras, composta por rochas siliciclásticas. Em aproximadamente X525 m se observa um aumento das concentrações de Al e Fe, acompanhado pelo aparecimento de Ti e Na. Esses elementos estão associados a minerais como plagioclásio e piroxênios que compõem as rochas ígneas da Formação Camboriú.

O poço da figura 29 apresenta as Formações Barra Velha e Itapema, novamente separadas por folhelhos de alta radioatividade em aproximadamente X690 m. No topo da Formação Barra Velha se observa altas concentrações de Si, sem um aumento de K, Al e Fe. Isso é reflexo do intenso processo de silicificação ao qual esses carbonatos foram submetidos.

A figura 30 apresenta a correlação entre as variáveis do modelo mineralógico. Observa-se uma alta correlação positiva entre carbonatos e PF, resultado da queima do dióxido de carbono, um fenômeno descrito por Dean (1974) e Heiri, Lotter e Lemcke (2001). As correlações negativas observadas entre calcita e dolomita e quartzo confirmam os processos diagenéticos de dolomitização e silicificação evidenciados pelas correlações entre os elementos químicos. A alta correlação positiva entre plagioclásio e piroxênio indica que esses minerais aparecem associados em rochas ígneas. Outras correlações óbvias observadas são: Ca e calcita, Mg e dolomita, Si e quartzo e K e K-feldspato.

As fórmulas químicas padrão dos minerais, apresentadas na tabela 8, demonstra o desafio de se construir um modelo mineralógico de forma analítica. Minerais como dolomita, K-feldspato, plagioclásio e piroxênio apresentam composição química complexa, com variações significativas nas razões de Ca-Mg, K-Al, Na-Ca e Si-Al. As fórmulas químicas reais desses minerais podem diferenciar expressivamente da fórmula padrão. Uma evidência disso é a correlação positiva entre Fe e Ti com plagioclásio e piroxênio, já que esses elementos não aparecem na fórmula desses minerais. As argilas detríticas normalmente

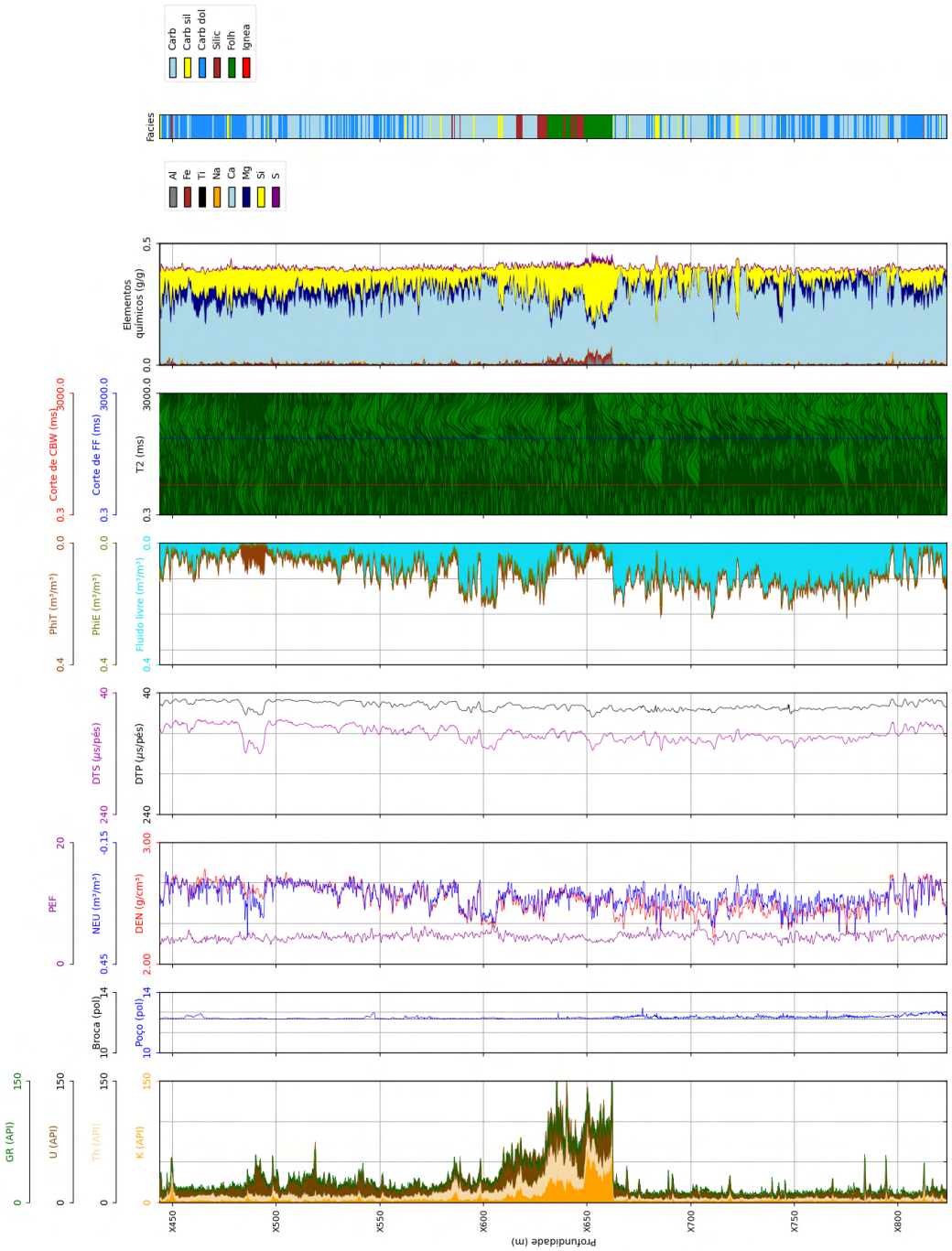


Figura 27 – Exemplo de perfis de um dos poços da base de dados. Observa-se a Formação Barra Velha na metade superior do poço e a Formação Itapema na metade inferior. Um folhelho de alta radioatividade separa as duas, em aproximadamente X650 m. Os carbonatos possuem composição calcítica com algumas regiões dolomitizadas.

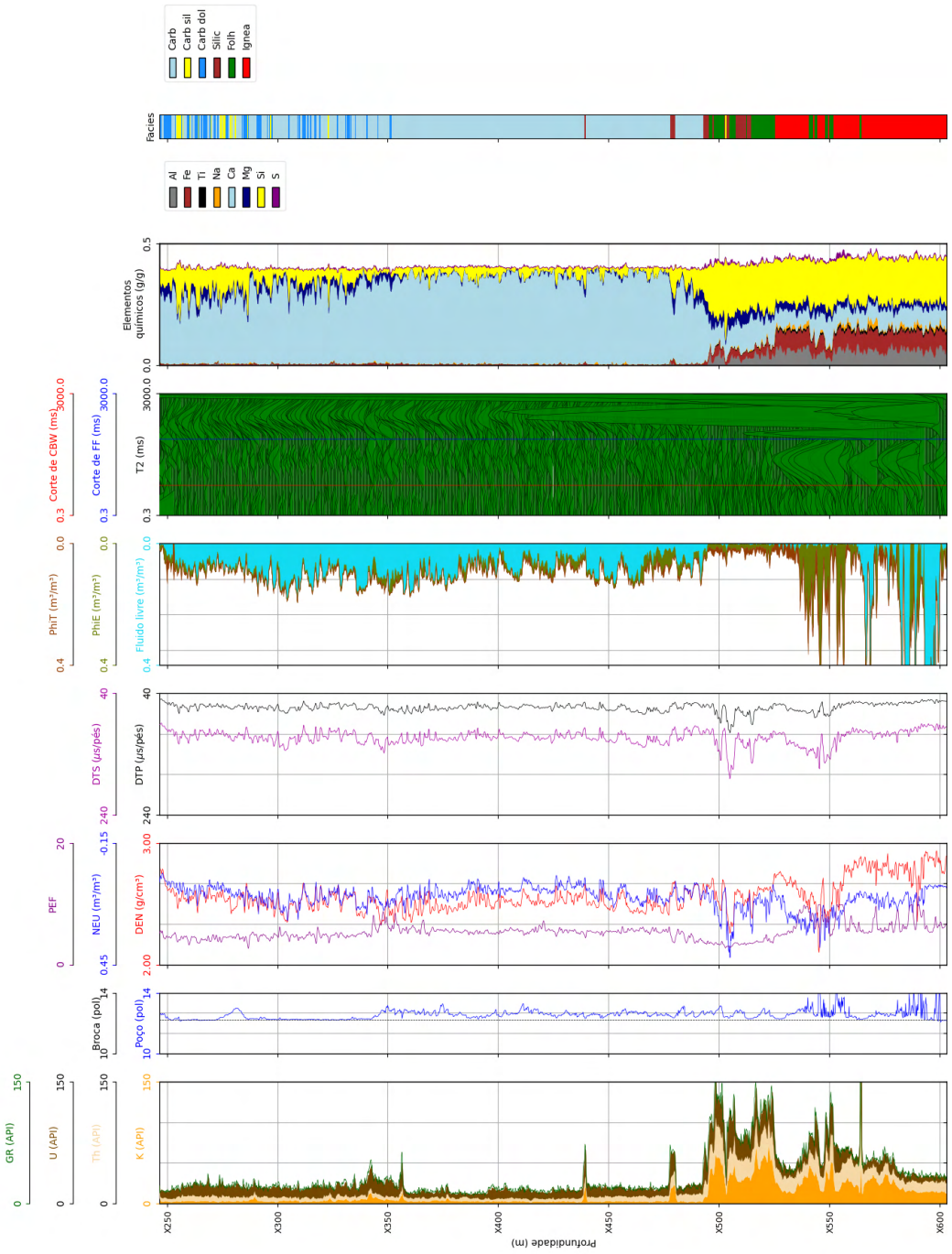


Figura 28 – Exemplo de perfis de um dos poços da base de dados. Observa-se a Formação Barra Velha no terço superior do poço, a Formação Itapema no meio e as Formações Piçarras e Camboriú no terço inferior. A Formação Barra Velha possui composição calcítica dolomitizada, enquanto a Formação Itapema é francamente calcítica. A Formação Piçarras apresenta uma intercalação de rochas siliciclásticas e folhelhos, e a Formação Camboriú é composta por rochas ígneas.

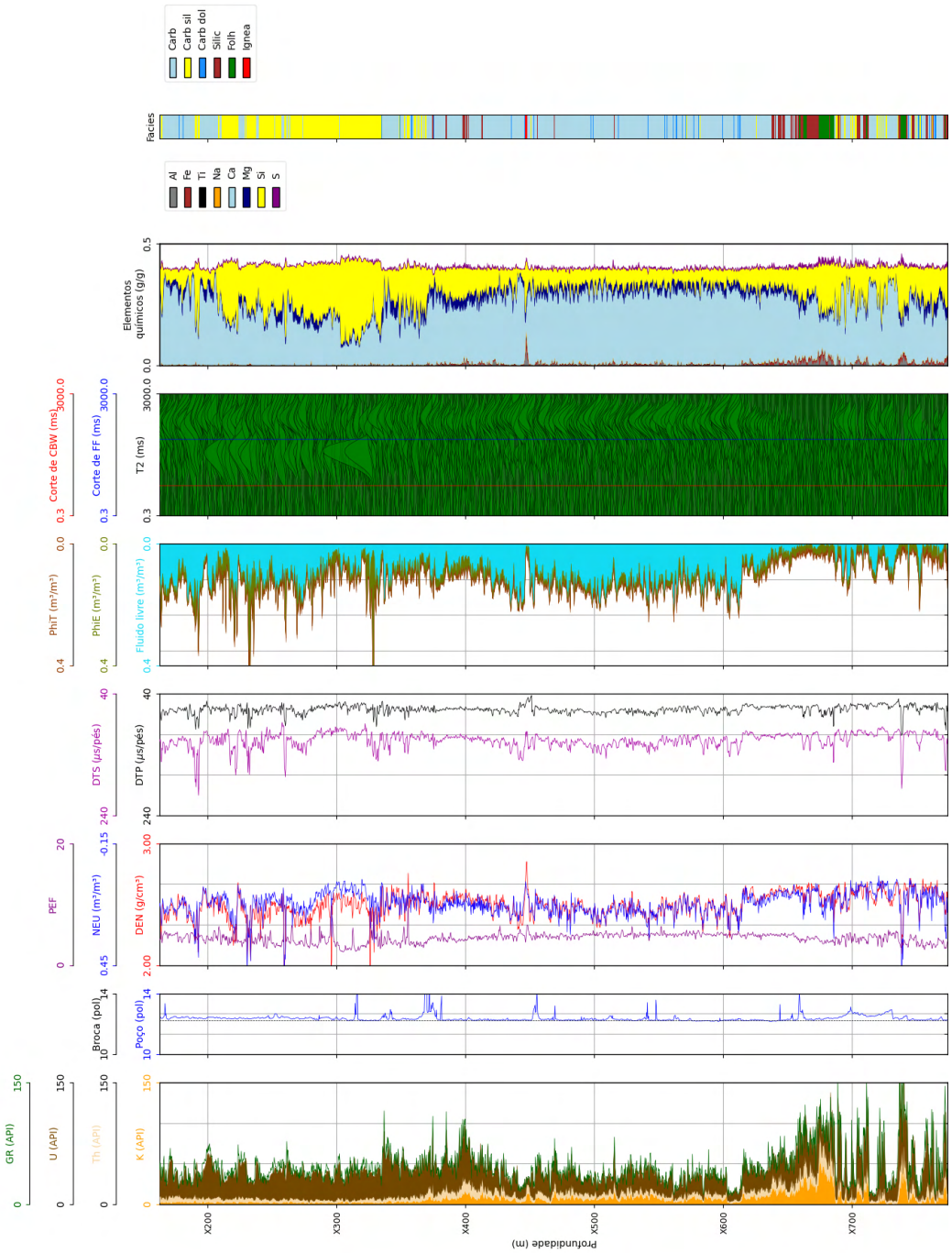


Figura 29 – Exemplo de perfis de um dos poços da base de dados. Observa-se a Formação Barra Velha nos dois terços superiores do poço e a Formação Itapema no terço inferior. Um folhelho de alta radioatividade separa as duas, em aproximadamente X690 m. O topo da Formação Barra Velha possui composição calcítica silicificada.

apresentam composição complexa, variando de acordo com a rocha fonte e o ambiente de deposição. Um reflexo disso é o fato das argilas detríticas não apresentarem correlação significativa com nenhum elemento químico.

4.2 Modelagem geoquímica

4.2.1 Modelos

As figuras 31 e 32 apresentam as variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação dos modelos para os diferentes elementos químicos. De maneira geral, observa-se que ocorre um platô em termos de melhora das métricas a partir de 50 árvores, conforme o obtido na calibração dos hiperparâmetros do algoritmo. Não foi observado *overfitting* dos modelos, que se manifestaria numa piora das métricas observadas no conjunto de validação com a adição de novas árvores.

As tabelas 15 16 apresentam os resultados de R^2 e EQM da validação e validação cruzada dos modelos treinados para a geração de perfis geoquímicos sintéticos. Os resultados são condizentes com o observado na literatura (apresentados na tabela 3) e foram considerados com alta qualidade. Os valores próximos entre os R^2 da validação e validação cruzada indicam que os modelos apresentaram boa capacidade de generalização e não apresentaram *overfitting* na base de treinamento.

O Fe apresentou os melhores resultados, seguido por Ti, Al e Ca. Esses elementos apresentaram R^2 acima de 0,90 tanto para a validação quanto para a validação cruzada. Em seguida, o Si, S e Mg, todos com R^2 acima de 0,80. O Na apresentou os menores R^2, provavelmente relacionado a baixa concentração desse elemento nas rochas do pré-sal, com leituras afetadas por ruido. Ainda assim, o R^2 do modelo de Na ficou acima de 0,70 tanto para a validação e validação cruzada.

As figuras 33, 34 e 35 apresentam gráficos dos dados reais *versus* dados modelados e histogramas dos erros das bases de validação dos modelos treinados. Apesar do grande número de instâncias dos gráficos, o histograma dos erros mostra que a grande maioria dos pontos estão alinhados com a reta 1:1, refletindo os alto valores de R^2 obtidos.

O desvio padrão do erro dos modelos é apresentado tanto nas figuras 33, 34 e 35 quanto na tabela 17. Nenhum desvio padrão ficou acima de 0,05 g/g, resultado também

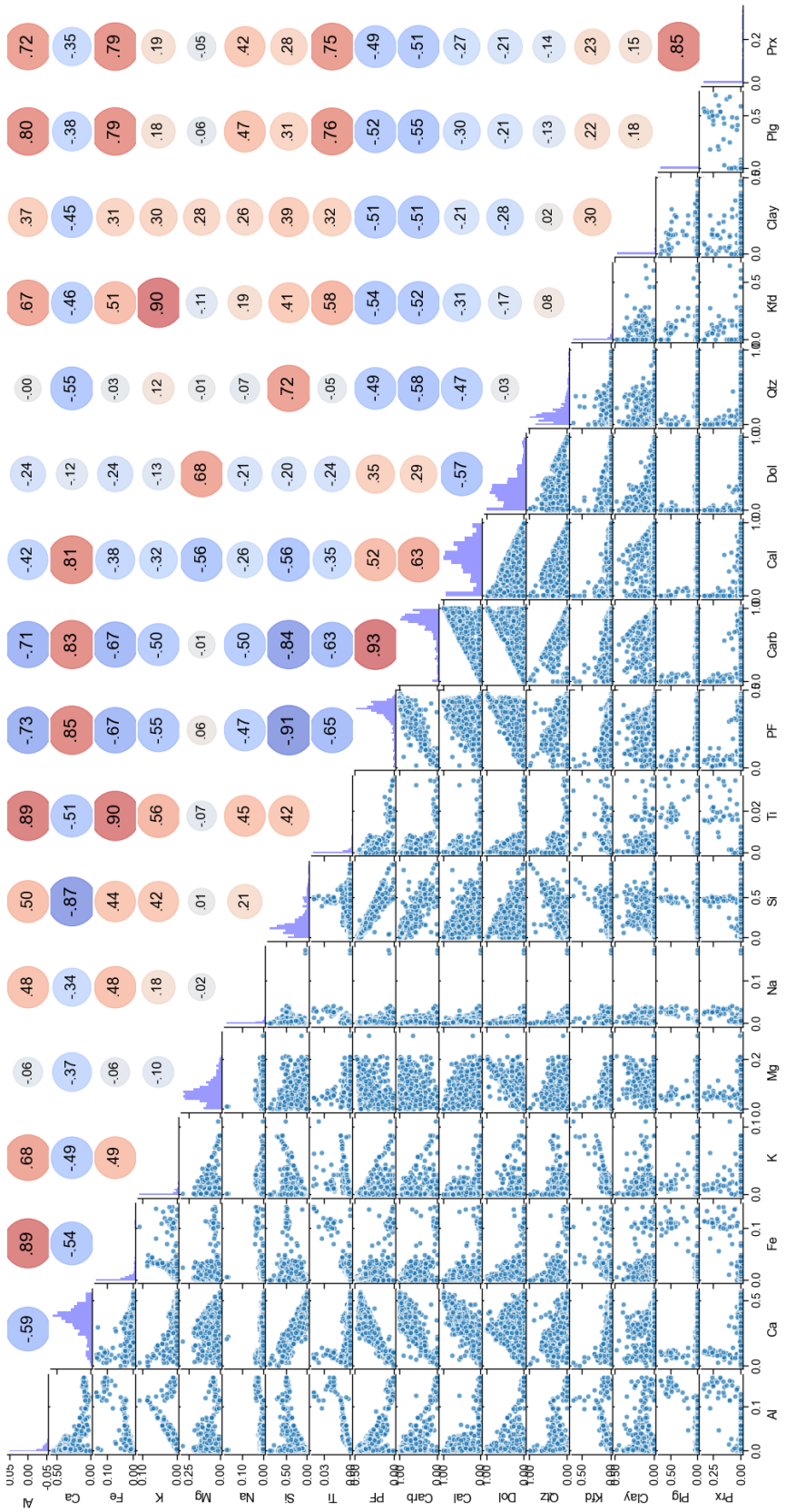


Figura 30 – Correlação entre as variáveis do modelo mineralógico, apresentada na forma de gráficos e R^2. Cores quentes indicam correlação positiva, enquanto cores frias indicam correlação negativa. Histogramas das variáveis são apresentados na diagonal da figura.

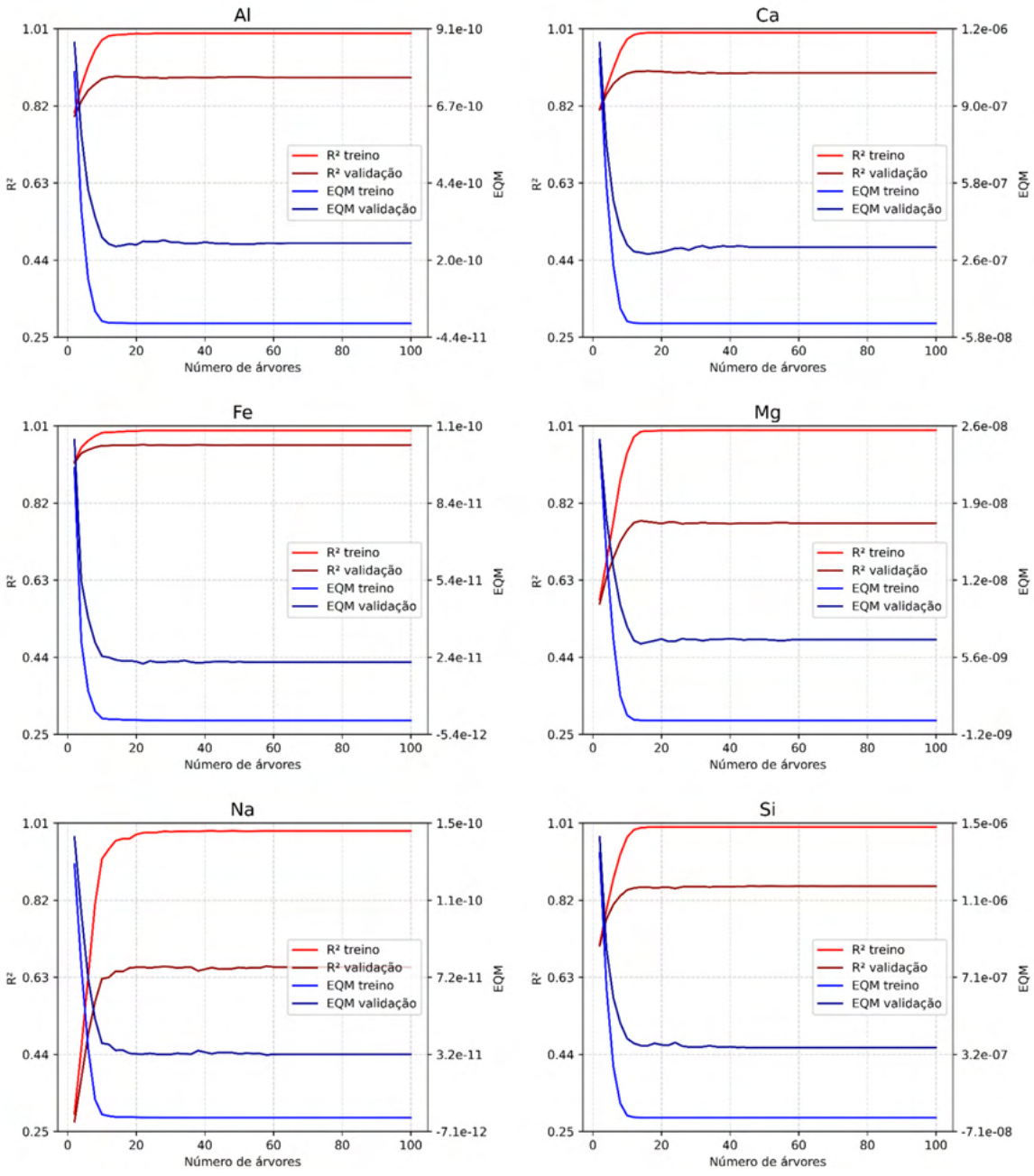


Figura 31 – Variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação para Al, Ca, Fe, Mg, Na e Si.

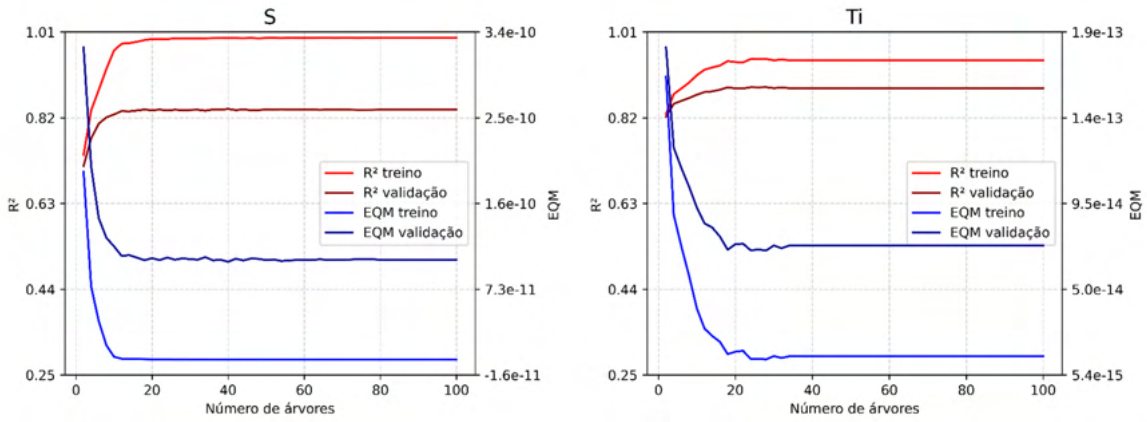


Figura 32 – Variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação para S e Ti.

considerado de alta qualidade. O seguinte exemplo ilustra como o desvio padrão pode ser usado para definir a incerteza de um modelo: considerando que o modelo de Ca estimou uma concentração de 0,350 g/g e o desvio padrão do erro desse modelo é de 0,019 g/g, é possível afirmar que existe 68% de chance do valor estimado estar entre 0,331 e 0,369 g/g. Caso se deseje aumentar o intervalo de confiança de 68 para 99,7%, basta usar três desvios padrão. Dessa forma, o valor estimado estará entre 0,293 e 0,407 g/g.

Tabela 15 – Resultados de R^2 da validação e validação cruzada dos modelos treinados para a criação dos perfis geoquímicos sintéticos.

	Al	Ca	Fe	Mg	Na	Si	S	Ti
Validação	0,93	0,94	0,97	0,86	0,76	0,91	0,89	0,89
Validação cruzada	0,91	0,92	0,97	0,81	0,71	0,88	0,83	0,89

Fonte – Lucas Oliveira, 2021

Tabela 16 – Resultados de EQM da validação e validação cruzada dos modelos treinados para a criação dos perfis geoquímicos sintéticos.

	Al	Ca	Fe	Mg	Na	Si	S	Ti
Validação	1e-5	5e-4	4e-6	7e-5	5e-6	5e-4	7e-6	3e-7
Validação cruzada	9e-7	3e-5	5e-7	2e-6	2e-7	3e-5	2e-6	1e-8

Fonte – Lucas Oliveira, 2021

4.2.2 Importância das variáveis

A figura 36 apresenta a importância das variáveis de entrada dos modelos treinados para a geração dos perfis geoquímicos sintéticos. Os modelos de Ca, Mg e Si foram

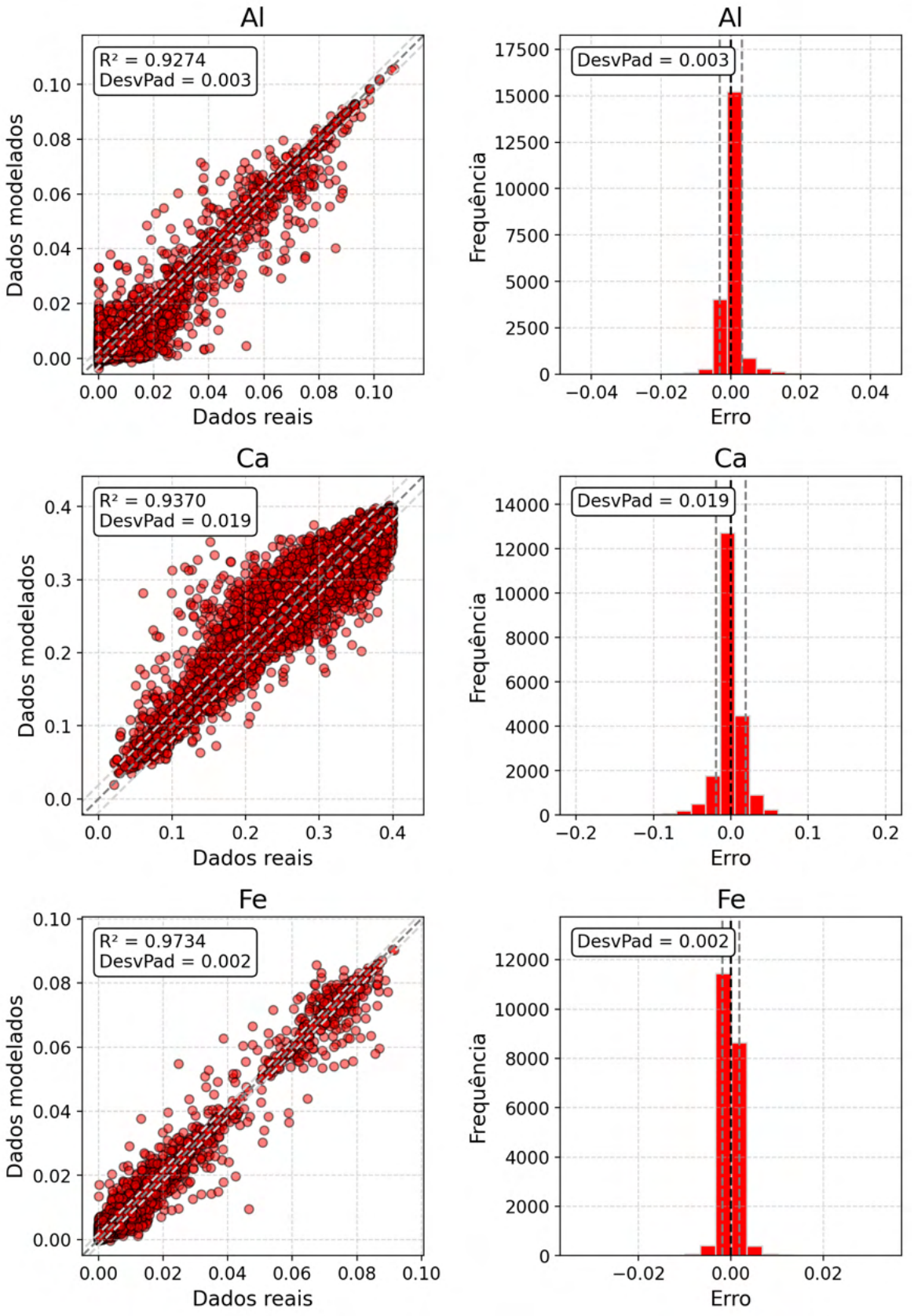


Figura 33 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de Al, Ca e Fe. DesvPad: Desvio padrão.

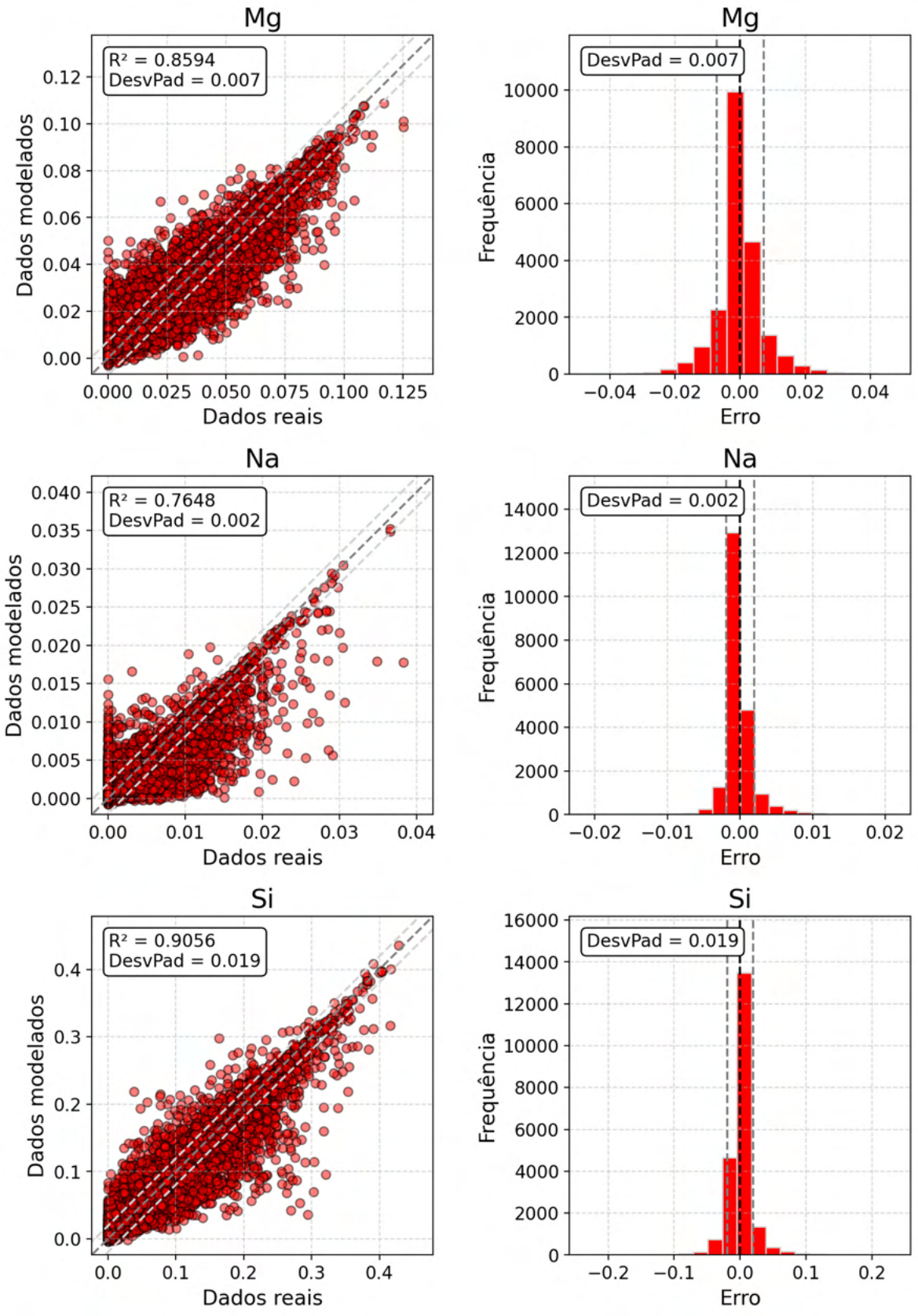


Figura 34 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de Mg, Na e Si. DesvPad: Desvio padrão.

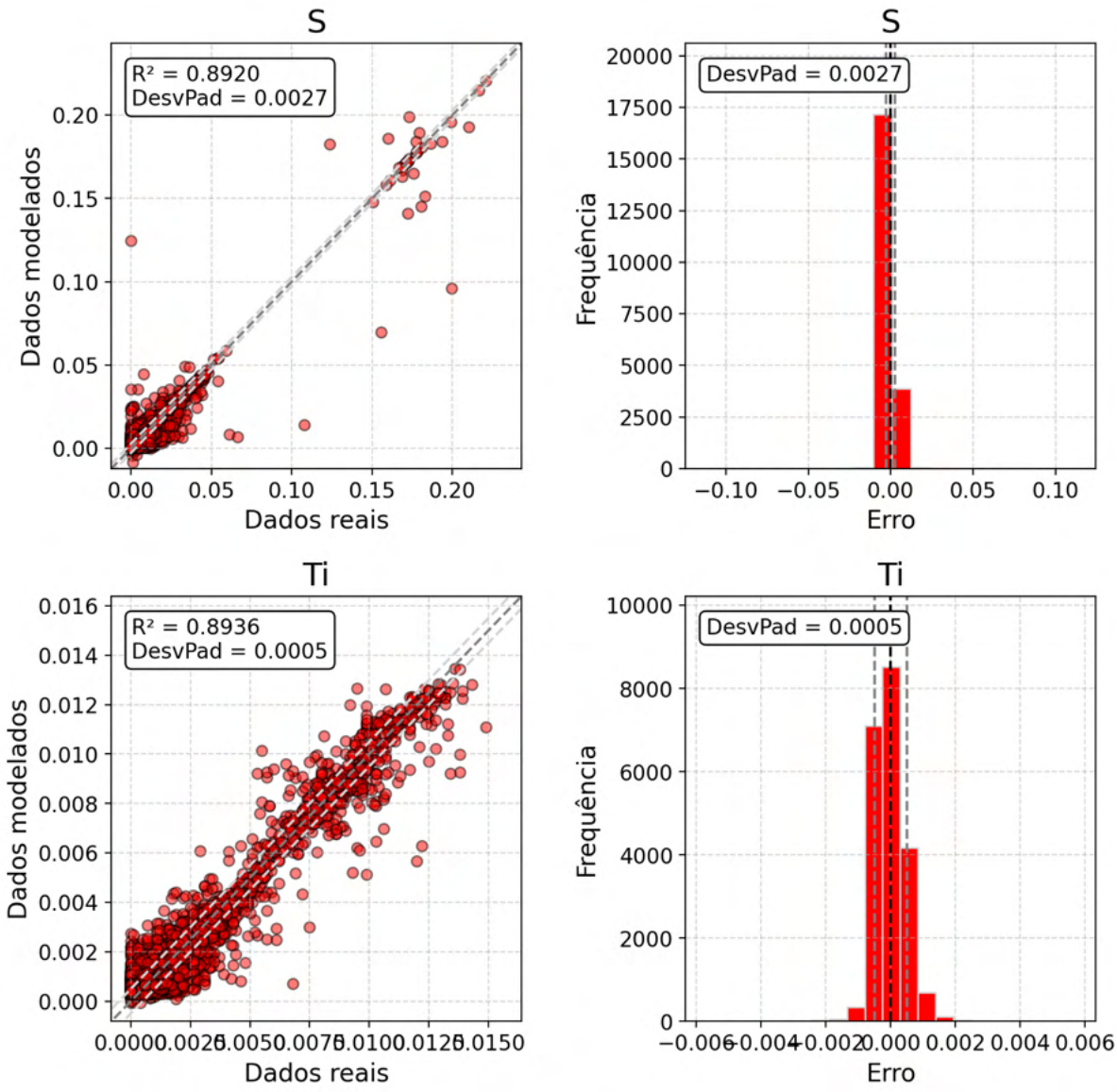


Figura 35 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de S e Ti. DesvPad: Desvio padrão.

Tabela 17 – Desvios padrão do erro da validação dos modelos treinados para a criação dos perfis geoquímicos sintéticos. Todos os valores estão em g/g.

Desvios padrão	Al	Ca	Fe	Mg	Na	Si	S	Ti
Um	0,003	0,019	0,002	0,007	0,002	0,019	0,003	0,0005
Dois	0,006	0,038	0,004	0,014	0,004	0,038	0,006	0,0010
Três	0,009	0,057	0,006	0,021	0,006	0,057	0,009	0,0015

Fonte – Lucas Oliveira, 2021

impactados principalmente pelos perfis de fator fotoelétrico, densidade, K e nêutrons. Os modelos de Al, Fe e Ti foram fortemente influenciados pelos perfis de K, densidade, fator fotoelétrico, U e Th. O modelo de S foi impactado majoritariamente pelo perfil de densidade. Nenhum perfil foi particularmente importante no treinamento do modelo de Na.

A análise da importância das variáveis sugere que perfis geoquímicos sintéticos de boa qualidade ainda poderiam ser gerados em cenários de maior redução do escopo de perfilagem. Para isso, seria necessário a aquisição dos perfis de densidade, nêutrons, fator fotoelétrico e espectroscopia de raios gama natural.

4.2.3 Perfis geoquímicos sintéticos

A fase de teste consistiu na geração de perfis geoquímicos sintéticos para três poços perfurados no pré-sal da Bacia de Santos e sua comparação com os perfis reais. Adicionalmente, os perfis de elementos químicos reais e modelados foram agrupados através de um algoritmo de clusterização aglomerativa, gerando cinco grupos distintos. Esse procedimento visou imitar a interpretação geológica desses perfis por um especialista.

Os gráficos de perfis reais *versus* modelados são apresentados nas figuras 37, 38, 39, 40 e 41. Elementos principais como Ca, Mg e Si mostram uma boa concordância com os dados modelados, apesar da alta dispersão. Elementos secundários como Al, Na e S têm baixo R^2 devido às suas leituras em baixas concentrações serem impactadas por ruído ambiental.

No topo do poço 1 (figura 42), se observa rochas carbonáticas da Formação Barra Velha cujo conteúdo de Si e Mg diminui com o aumento da profundidade. Esse padrão é observado nos dados reais e modelados. Em aproximadamente X160 m ocorre um aumento nas concentrações de Fe, Al e Si, caracterizando um grupo diferente dos carbonatos que é capturado tanto pelos perfis reais quanto pelos modelados. Provavelmente esse grupo se caracteriza por rochas siliciclásticas e folhelhos. A partir de X175 m, ocorrem carbonatos ricos em Ca intercalados com carbonatos ricos em Si e Mg da Formação Itapema, padrão também observado por ambos os perfis. Nas profundidades finais, um aumento de Fe, Al e Si é observado, acompanhado pelo aumento de Ti, Na e Mg, evidenciando um grupo diferente dos anteriores e possivelmente relacionado a rochas ígneas da Formação Camboriú. Esse grupo é capturado pelos perfis reais e modelados.

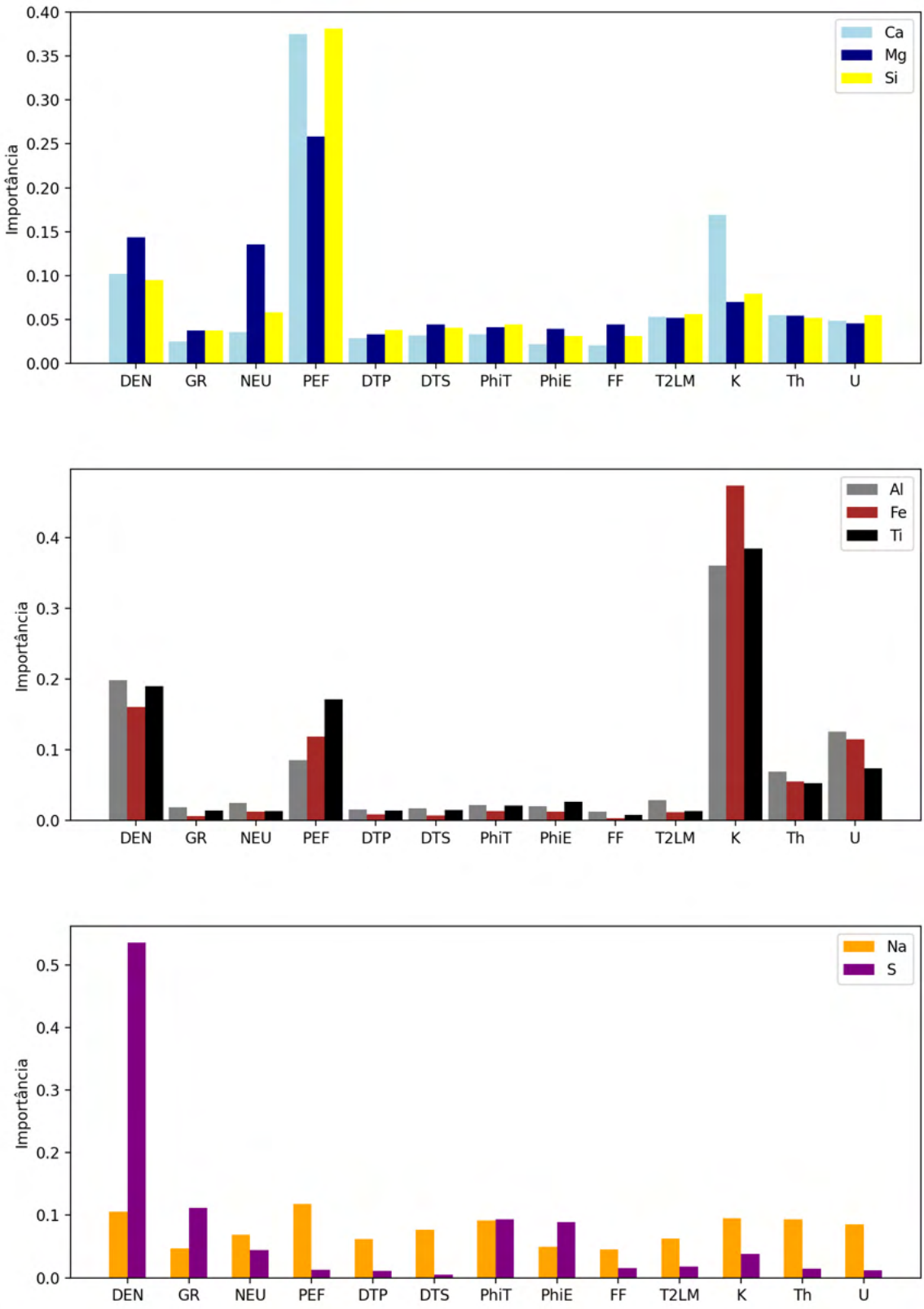


Figura 36 – Importância das variáveis de entrada dos modelos treinados para a geração de perfis geoquímicos sintéticos.

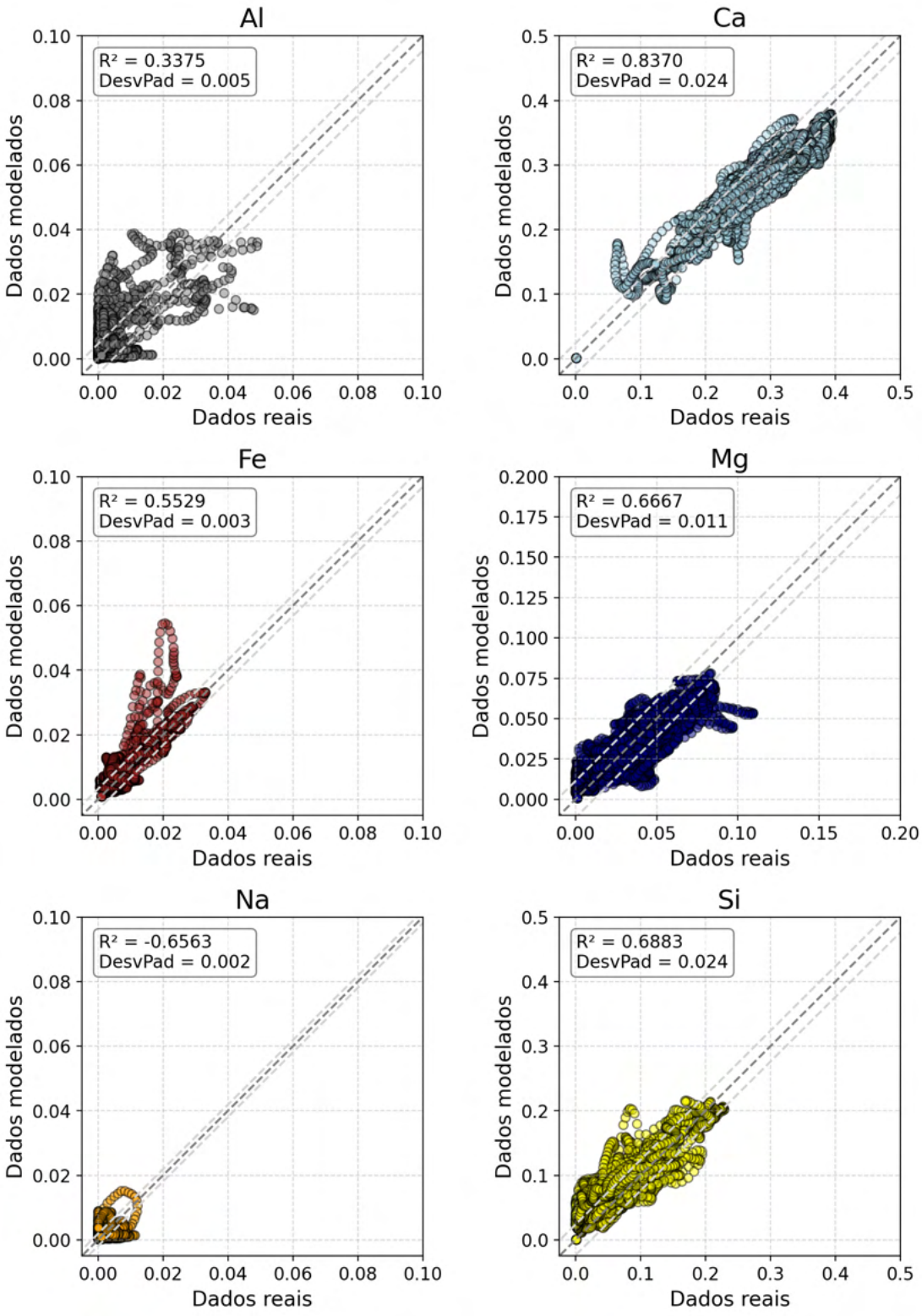


Figura 37 – Dados reais *versus* dados modelados para os elementos Al, Ca, Fe, Mg, Na e Si do poço 1. Os baixos R^2 são reflexo das baixas concentrações encontradas nas formações, com leituras muito afetadas pelo ruído ambiental. DesvPad: Desvio padrão.

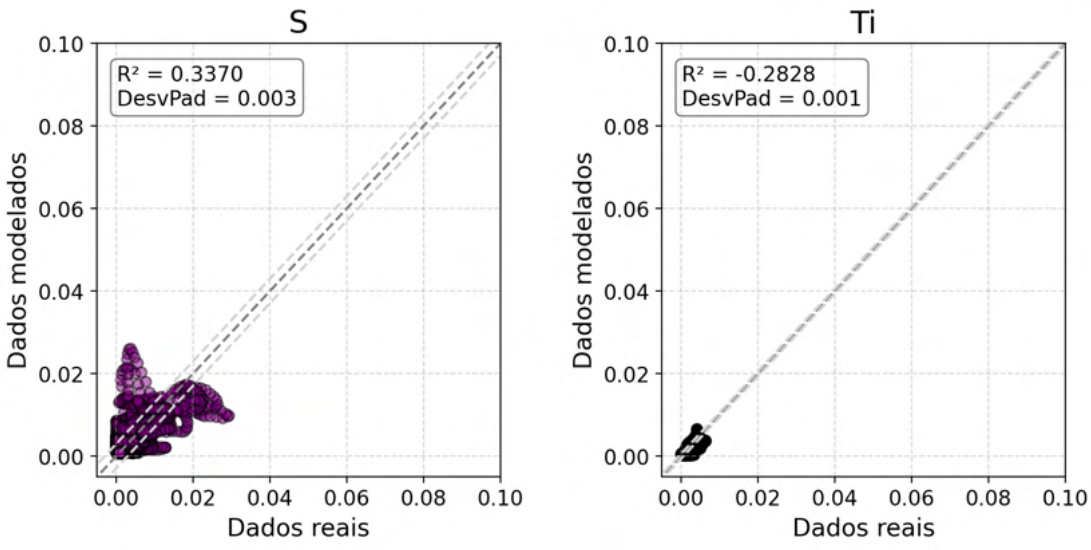


Figura 38 – Dados reais *versus* dados modelados para os elementos S e Ti do poço 1. Os baixos R^2 são reflexo das baixas concentrações encontradas nas formações, com leituras muito afetadas pelo ruído ambiental. DesvPad: Desvio padrão.

O poço 2 (figura 43) apresenta um padrão similar ao do poço 1, com intercalações de carbonatos ora ricos em Ca, ora ricos em Si e Mg, e rochas ricas em Al, Fe e Si. Entretanto, não ocorre o aparecimento de Ti e Na no final do poço, possivelmente relacionado a presença de rochas siliciclásticas da Formação Piçarras. Novamente, os mesmos padrões aparecem nos perfis reais e modelados, gerando agrupamentos similares.

No poço 3 (figura 44), não houve a aquisição de Mg e Na durante a aquisição dos perfis geoquímicos. Ainda assim, é possível observar que o agrupamento dos perfis reais e modelados é muito similar. Os perfis sintéticos foram capazes de estimar as concentrações de Mg e Na sem impactar as relações entre o Ca, Si, Al, Fe e Ti observadas nos perfis reais. Essas concentrações mostram que pode haver algum grau de dolomitização nos carbonatos da Formação Barra Velha e confirmar a presença de rochas ígneas da Formação Camboriú no final do poço, com a presença de Na.

A comparação entre os perfis reais e modelados demonstra que os perfis geoquímicos gerados por aprendizado de máquina foram capazes de reproduzir os padrões gerais das concentrações de elementos químicos observados nas rochas do pré-sal.

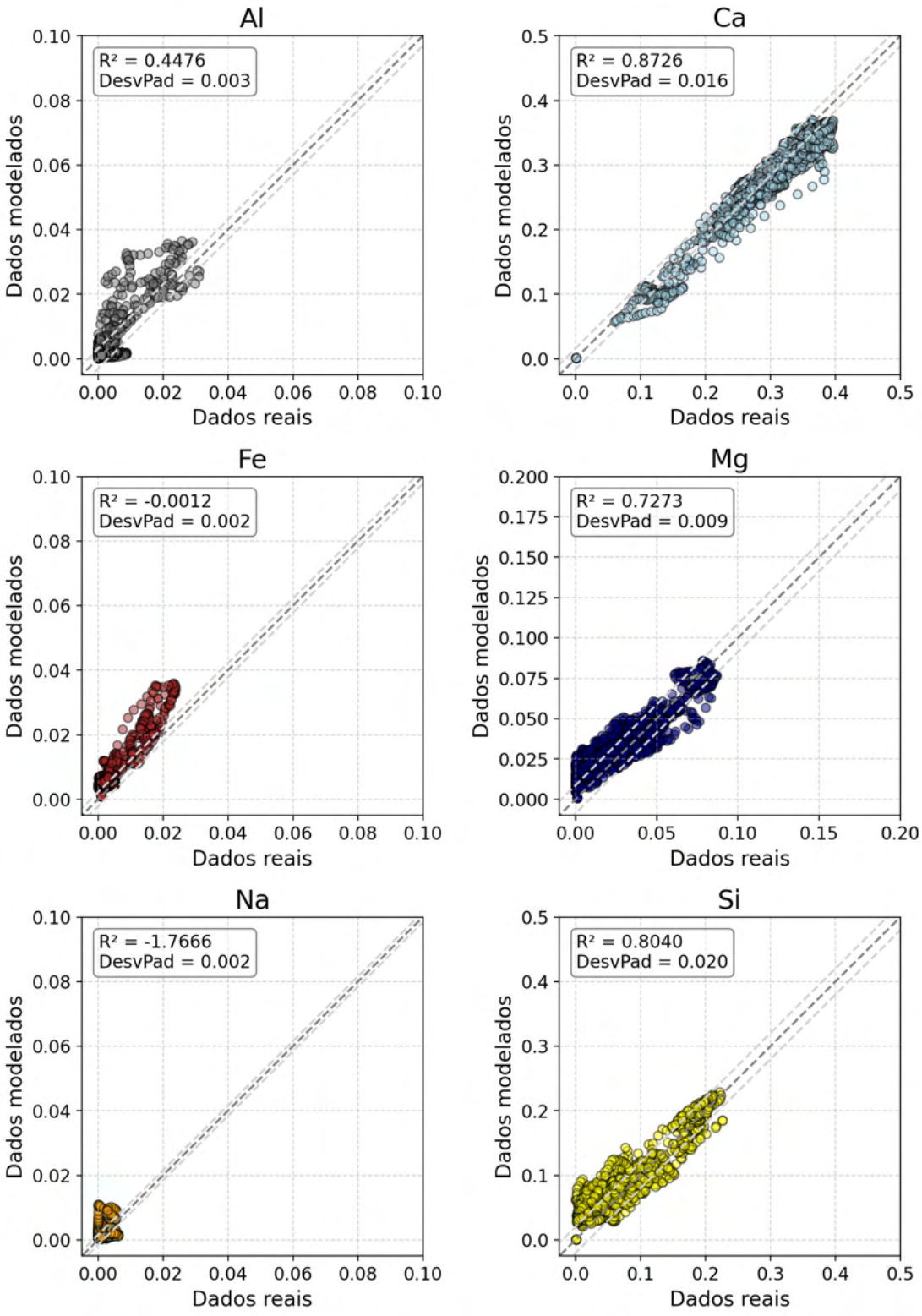


Figura 39 – Dados reais *versus* dados modelados para os elementos Al, Ca, Fe, Mg, Na e Si do poço 2. Os baixos R^2 são reflexo das baixas concentrações encontradas nas formações, com leituras muito afetadas pelo ruído ambiental. DesvPad: Desvio padrão.

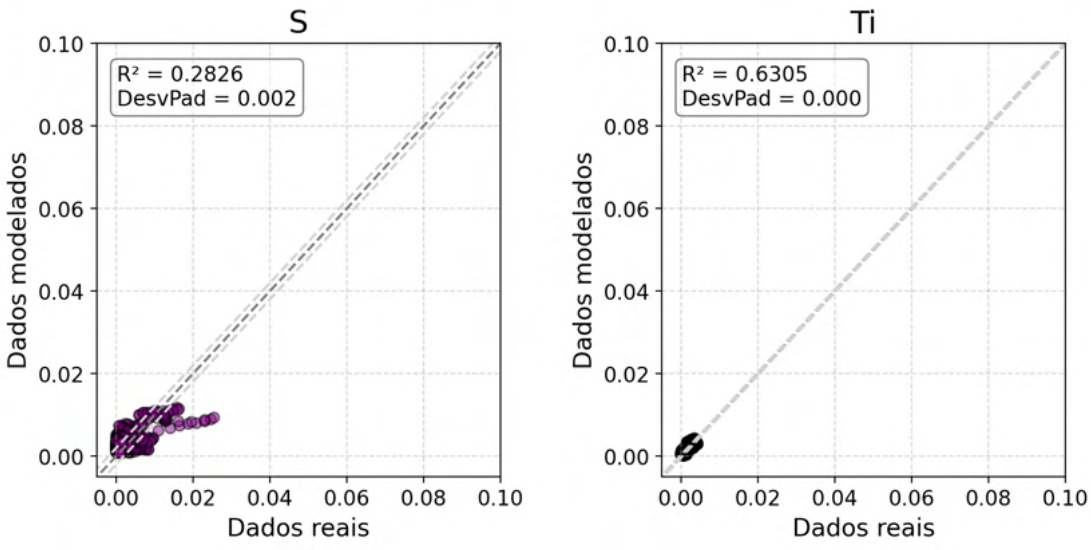


Figura 40 – Dados reais *versus* dados modelados para os elementos S e Ti do poço 2. Os baixos R^2 são reflexo das baixas concentrações encontradas nas formações, com leituras muito afetadas pelo ruído ambiental. DesvPad: Desvio padrão.

4.3 Modelagem mineralógica por aprendizado de máquina

4.3.1 Modelos

As tabelas 18 e 19 apresentam os resultados de R^2 e EQM da validação e validação cruzada dos modelos mineralógicos ao longo do aprendizado escalonado. Como era de se esperar, ocorre uma clara melhoria dos resultados a medida que novas informações são adicionadas às variáveis de entrada dos modelos.

O modelo de PF apresentou os melhores resultados observados, com R^2 de 0,96 e 0,97 para a validação e validação cruzada, respectivamente. Isso dá confiança quanto a utilização da PF nos modelos subsequentes. A inclusão da PF junto aos elementos químicos nas variáveis de entrada melhora os resultados da validação e validação cruzada do modelo de carbonatos. A junção da PF e carbonatos aos elementos químicos causa uma significativa melhora nos modelos de calcita, dolomita e quartzo. A inclusão da PF, carbonatos e quartzo nas variáveis de entrada melhorou expressivamente o modelo das argilas detríticas, aumentando o R^2 da validação e validação cruzada de valores próximos a 0,60 para valores acima de 0,80.

Os R^2 dos modelos de K-feldspato, plagioclásio e piroxênio não apresentaram grande variação com o uso do aprendizado escalonado. O modelo de piroxênio apresentou os piores resultados de forma geral. Porém, o modelo de plagioclásio + piroxênio usando os

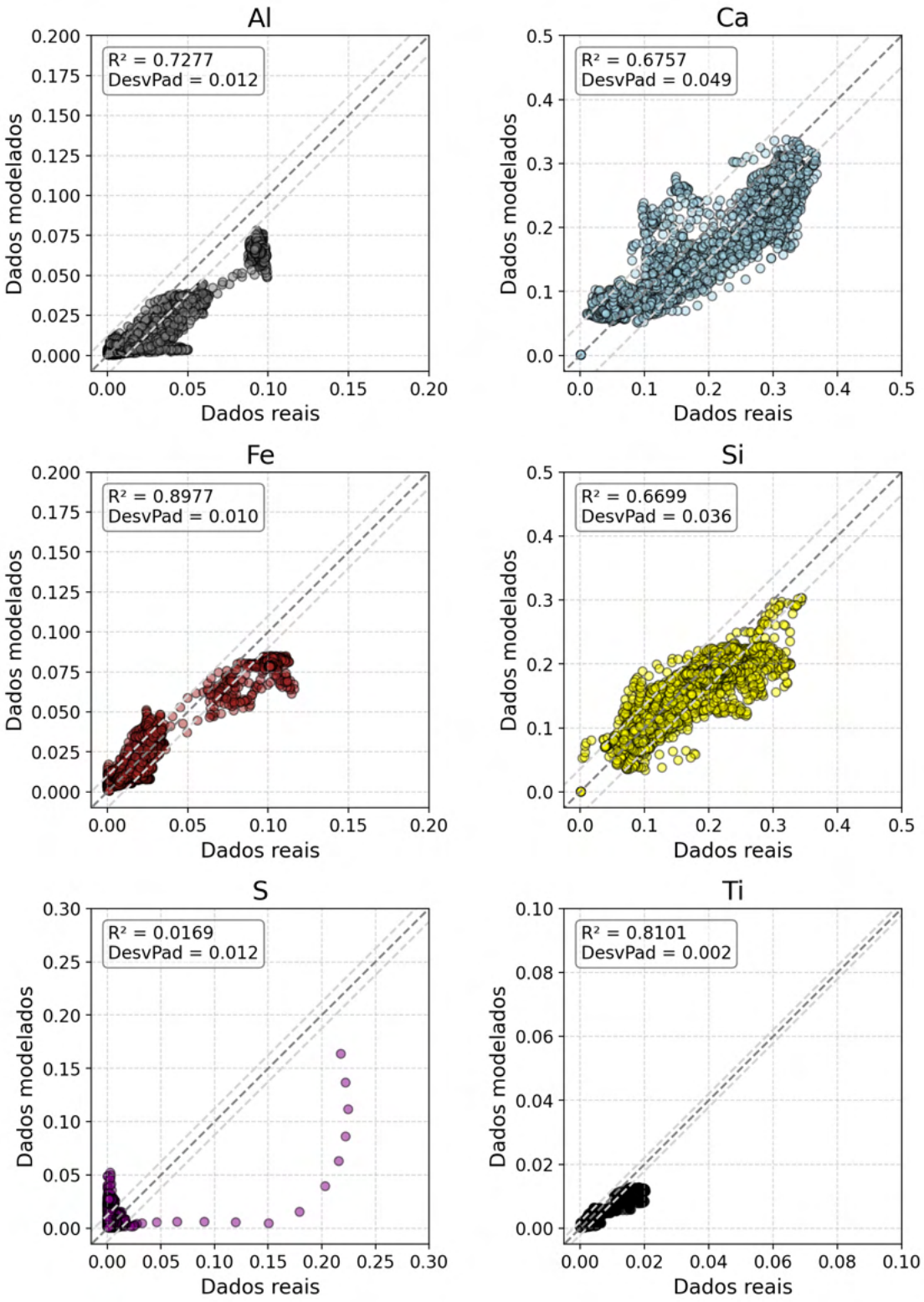


Figura 41 – Dados reais *versus* dados modelados para os elementos Al, Ca, Fe, Si, S e Ti do poço 3. Os baixos R^2 são reflexo das baixas concentrações encontradas nas formações, com leituras muito afetadas pelo ruído ambiental. DesvPad: Desvio padrão.

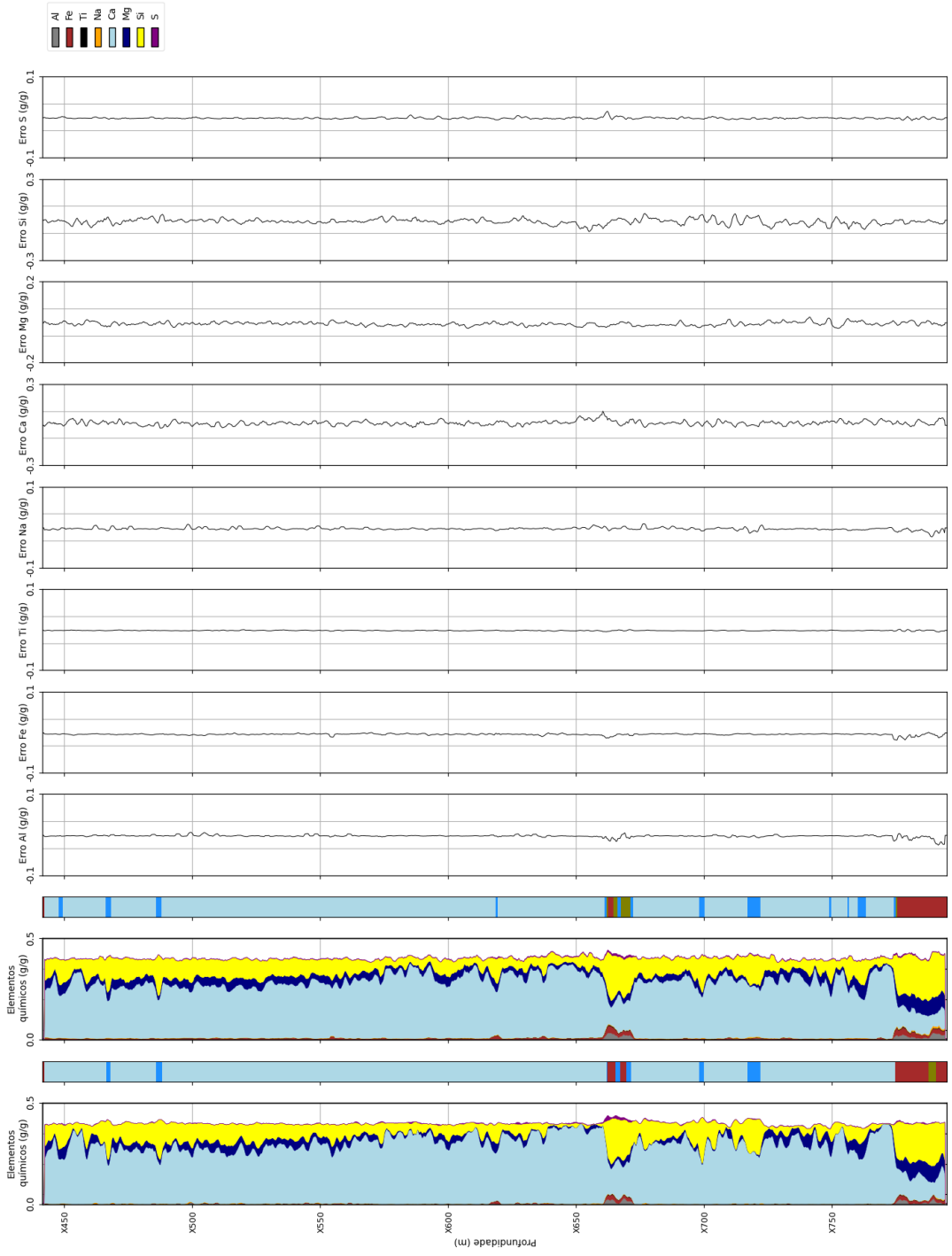


Figura 42 – Comparação entre os perfis geoquímicos reais e modelados para o poço 1. Uma evidente correspondência é observada, com padrões gerais sendo reproduzidos. Os resultados de um agrupamento aglomerativo, criados para simular uma interpretação geológica, atestam a qualidade dos perfis modelados.

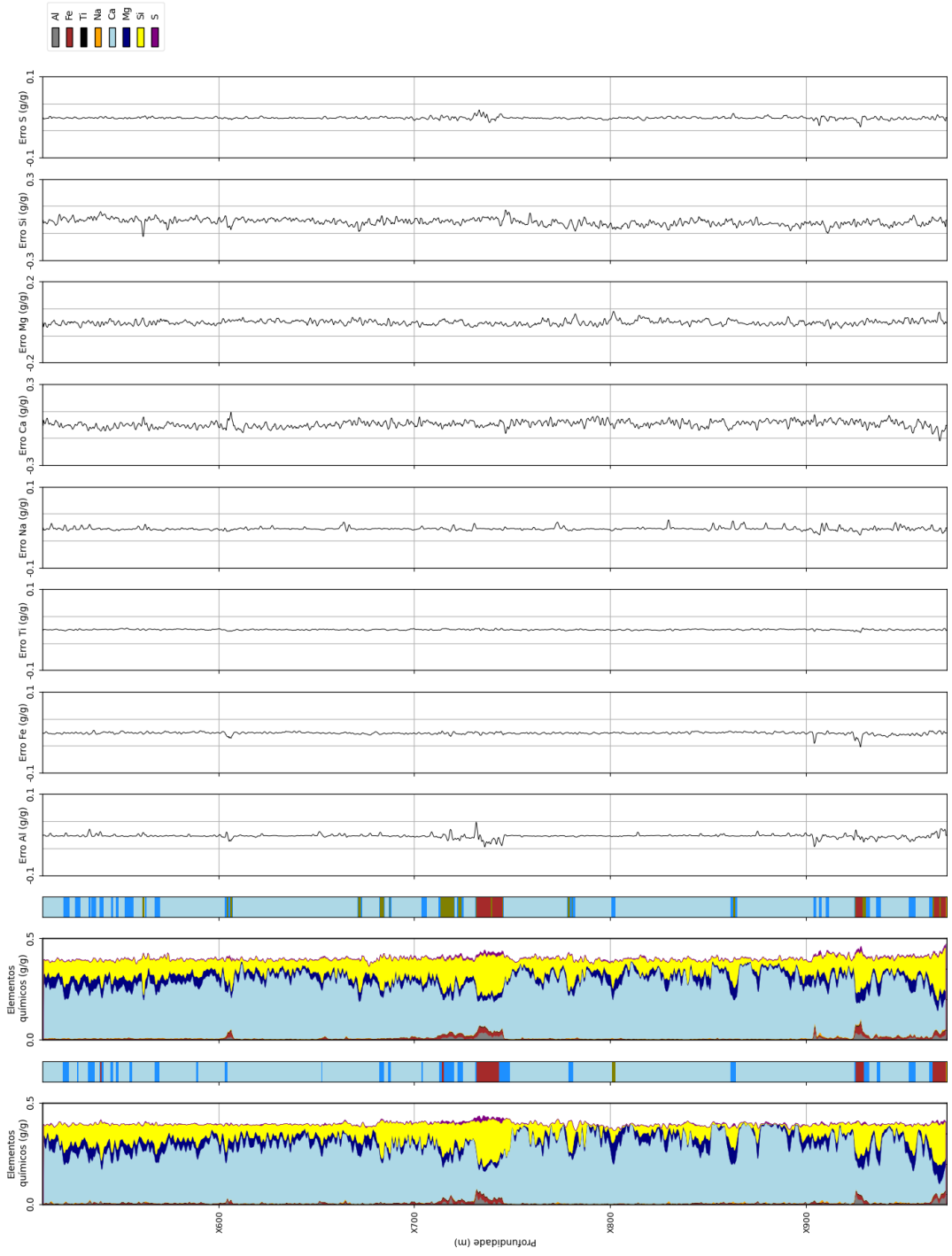


Figura 43 – Comparação entre os perfis geoquímicos reais e modelados para o poço 2. Uma evidente correspondência é observada, com padrões gerais sendo reproduzidos. Os resultados de um agrupamento aglomerativo, criados para simular uma interpretação geológica, atestam a qualidade dos perfis modelados.

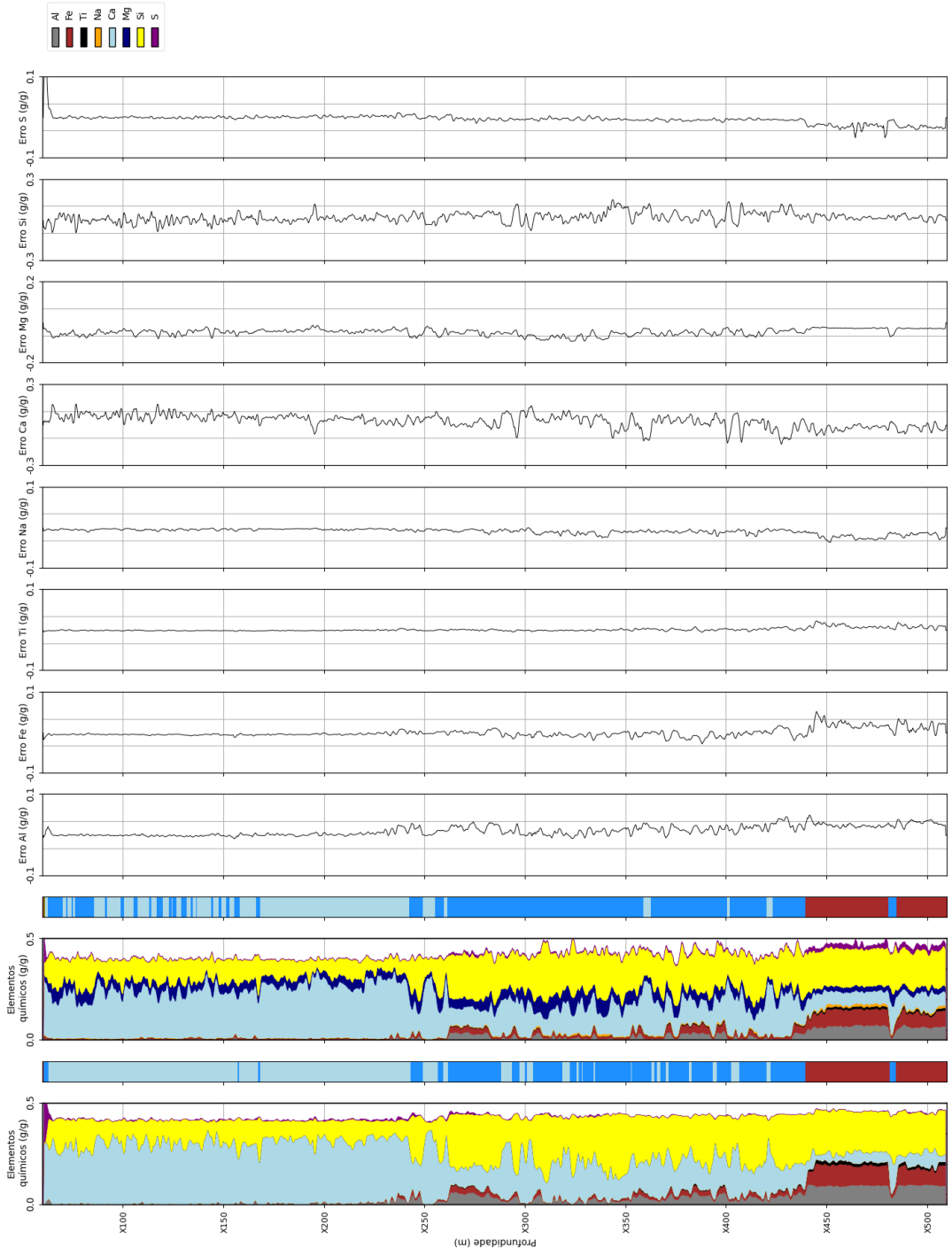


Figura 44 – Comparação entre os perfis geoquímicos reais e modelados para o poço 3. Uma evidente correspondência é observada, com padrões gerais sendo reproduzidos. Os resultados de um agrupamento aglomerativo, criados para simular uma interpretação geológica, atestam a qualidade dos perfis modelados.

elementos químicos, PF e carbonatos como variáveis de entrada obteve melhores resultados quando comparado com apenas o piroxênio.

A propagação dos resultados de um modelo para as variáveis de entrada de um modelo subsequente carrega consigo seus erros. Sendo assim, é importante que o aprendizado escalonado seja utilizado somente quando os benefícios da propagação dos resultados de um modelo superem os malefícios da propagação de erros. Dessa forma, o seguinte aprendizado escalonado foi proposto para a criação do modelo mineralógico das rochas do pré-sal:

1. Estimativa de PF, K-feldspato e plagioclásio utilizando os elementos químicos como variáveis de entrada;
2. Inclusão da PF às variáveis de entrada e estimativa de carbonatos;
3. Inclusão de carbonatos às variáveis de entrada e estimativa de calcita, dolomita, quartzo e plagioclásio + piroxênio;
4. Subtração do plagioclásio do plagioclásio + piroxênio para a estimativa de piroxênio;
5. Inclusão de quartzo às variáveis de entrada e estimativa de argilas detríticas.

A figura 45 apresenta o resumo do aprendizado escalonado final.

As figuras 46 e 47 apresentam as variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação dos modelos para os diferentes minerais. De maneira geral, observa-se que ocorre um platô em termos de melhora das métricas a partir de 15 a 25 árvores, a depender do mineral. Não foi observado *overfitting* dos modelos, que se manifestaria numa piora das métricas observadas no conjunto de validação com a adição de novas árvores.

As figuras 48, 49, 50 e 51 apresentam gráficos dos dados reais *versus* dados modelados e histogramas dos erros das bases de validação dos modelos minerais. O desvio padrão do erro dos modelos é apresentado tanto nessas figuras como na tabela 20. Nenhum desvio padrão ficou acima de 0,10 g/g, resultado considerado satisfatório. Assim como nos modelos para a geração de perfis geoquímicos sintéticos, esses desvios padrão podem ser utilizados para definir a incerteza dos modelos.

Um fenômeno interessante é observado nos desvios padrão dos modelos de K-Feldspato, plagioclásio, piroxênio e plagioclásio + piroxênio. Como a maioria das instâncias apresenta fração próxima de zero, o algoritmo de aprendizado de máquina consegue estimar essas baixíssimas frações com alta acurácia. Porém, a medida que as frações

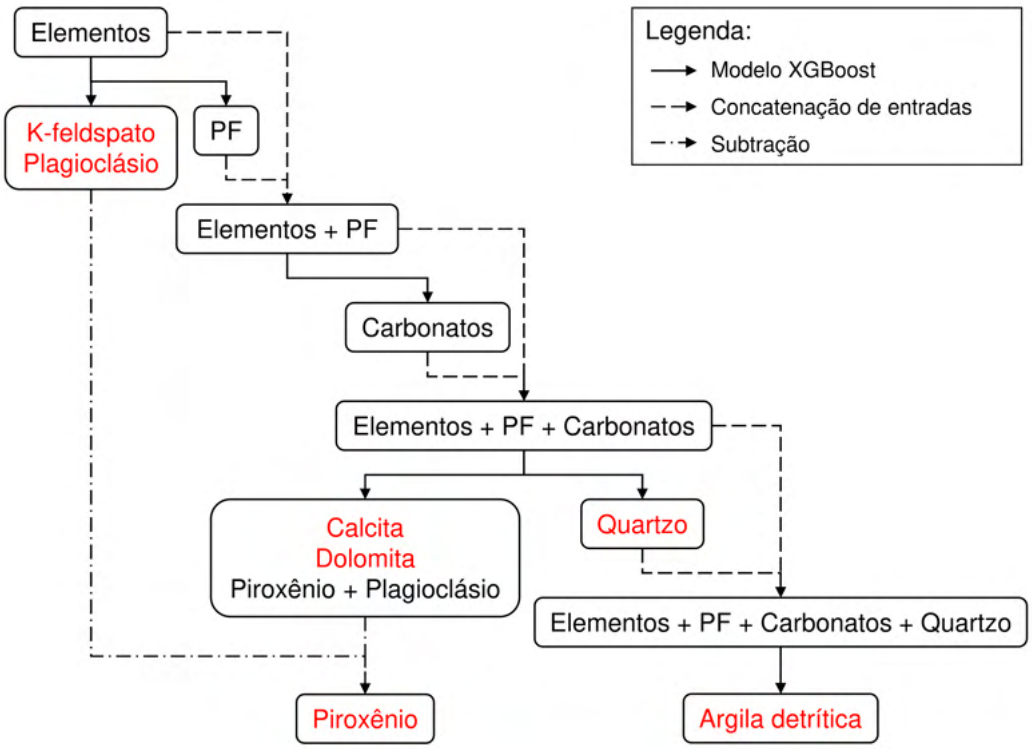


Figura 45 – Aprendizado escalonado final utilizado na criação do modelo mineralógico. Setas sólidas indicam o uso do algoritmo de aprendizado de máquina para a estimativa das frações minerais e PF. Setas tracejadas indicam a inclusão dos resultados de um modelo às variáveis de entrada do modelo anterior. Setas tracejadas e com pontos indicam a subtração do plagioclásio na fração de plagioclásio + piroxênio para a obtenção da fração de piroxênio.

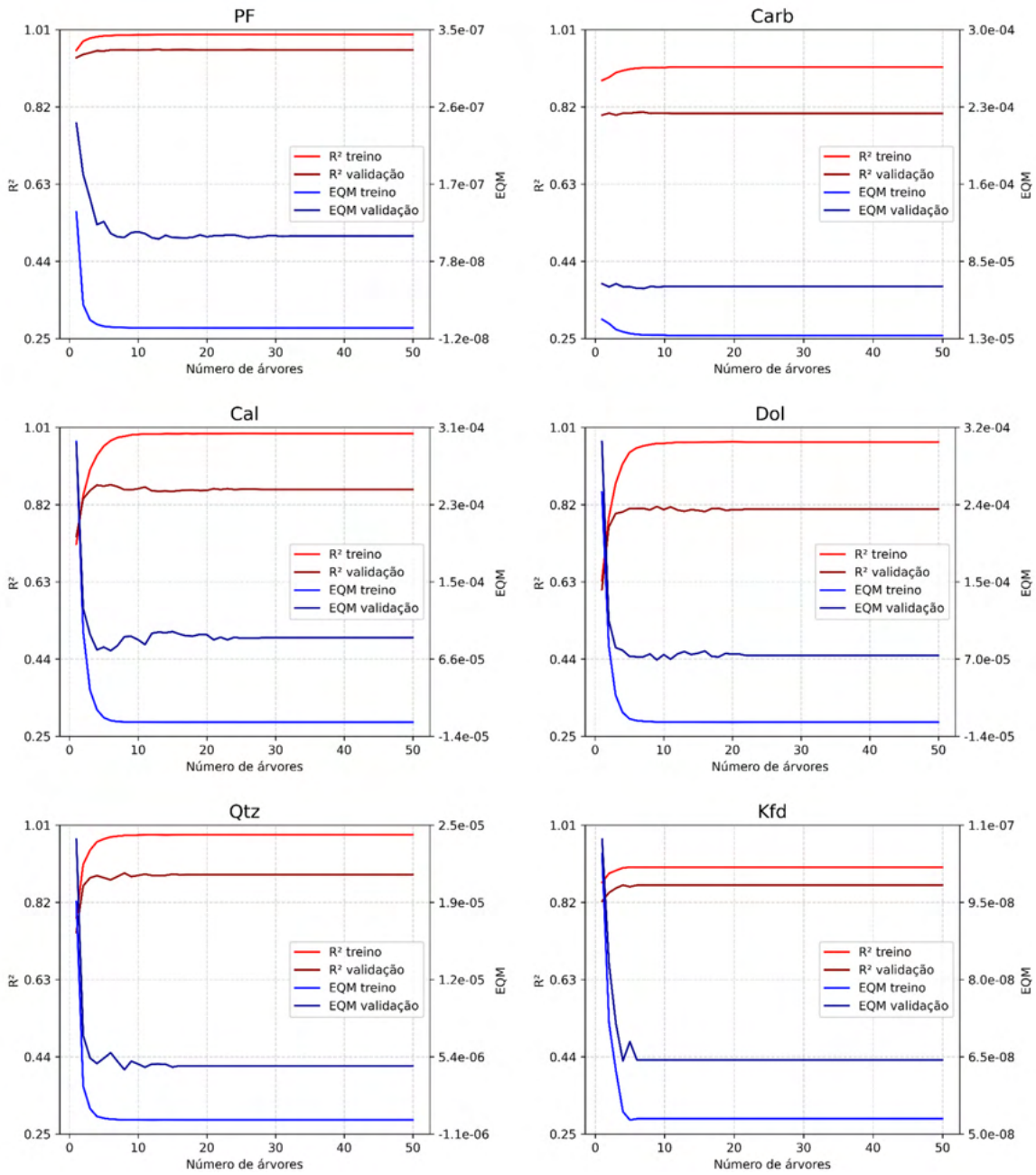


Figura 46 – Variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação para PF, carbonatos, calcita, dolomita, quartzo e K-feldspato.

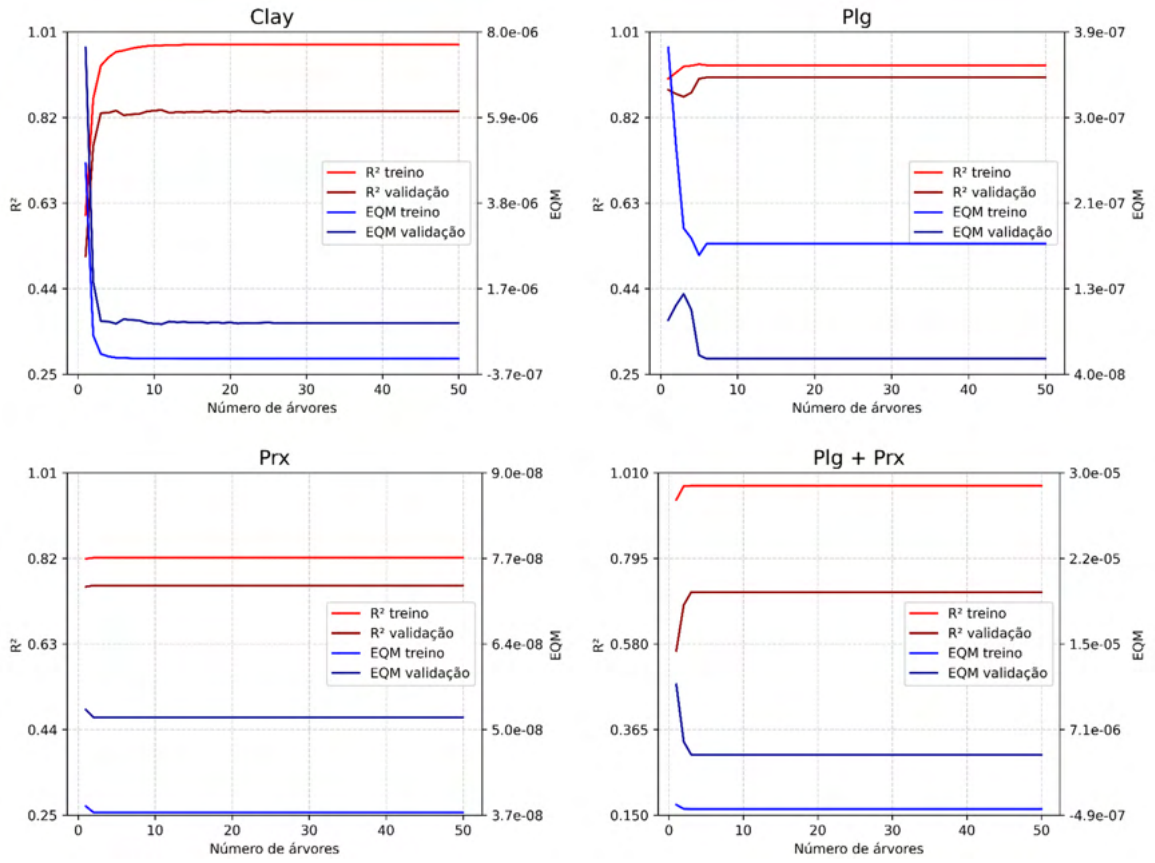


Figura 47 – Variações de R^2 e EQM com o aumento do número de árvores do algoritmo XGBoost obtidos durante o treino e validação para argilas detríticas, plagioclásio, piroxênio e plagioclásio + piroxênio.

desses minerais aumenta, o erro dos modelos também aumenta devido ao menor número de instâncias. Isso causa a falsa impressão de que a incerteza desses modelos é muito baixa. Na verdade, a incerteza desses modelos é baixa em baixas frações minerais. Em altas frações minerais, a incerteza é maior.

4.3.2 Importância das variáveis

A figura 52 apresenta a importância das variáveis de entrada dos modelos minerais. Essa importância é a relacionada aos modelos do aprendizado escalonado final; sendo assim, algumas variáveis estarão nulas (por exemplo, a variável quartzo no modelo de calcita).

Os elementos Ca e Si são os mais importantes no modelo de PF. A importância de Ca se deve a alta correlação da PF com carbonatos, cujo principal representante é a calcita. A importância do Si provavelmente se deve a substituição de carbonatos por

Tabela 18 – Resultados de R^2 da validação e validação cruzada dos modelos mineralógicos treinados.EQ: Elementos químicos.

	PF	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Plg+Prx
Validação										
EQ	0,96	0,81	0,77	0,79	0,66	0,82	0,60	0,96	0,64	0,83
EQ+PF	-	0,84	0,78	0,80	0,66	0,80	0,67	0,95	0,66	0,83
EQ+PF+Carb	-	-	0,86	0,82	0,87	0,80	0,70	0,96	0,67	0,88
EQ+PF+Carb+Qtz	-	-	0,87	0,86	-	0,79	0,85	0,95	0,68	0,85
Validação cruzada										
EQ	0,97	0,82	0,76	0,73	0,66	0,73	0,63	0,70	0,79	0,79
EQ+PF	-	0,85	0,77	0,76	0,68	0,73	0,68	0,70	0,79	0,73
EQ+PF+Carb	-	-	0,84	0,78	0,86	0,74	0,71	0,70	0,78	0,74
EQ+PF+Carb+Qtz	-	-	0,86	0,79	-	0,75	0,81	0,70	0,78	0,80

Fonte – Lucas Oliveira, 2021

Tabela 19 – Resultados de EQM da validação e validação cruzada dos modelos mineralógicos treinados.EQ: Elementos químicos.

	PF	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Plg+Prx
Validação										
EQ	3e-4	9e-3	2e-2	1e-2	7e-3	4e-4	2e-3	3e-4	5e-4	2e-3
EQ+PF	-	7e-3	1e-2	9e-3	7e-3	4e-4	2e-3	2e-4	5e-4	2e-3
EQ+PF+Carb	-	-	9e-3	8e-3	3e-3	4e-4	1e-3	3e-4	4e-4	1e-3
EQ+PF+Carb+Qtz	-	-	8e-3	7e-3	-	4e-4	7e-4	3e-4	4e-4	9e-4
Validação cruzada										
EQ	4e-4	8e-3	2e-2	1e-2	6e-3	7e-4	2e-3	2e-3	2e-4	1e-3
EQ+PF	-	7e-3	1e-2	9e-3	6e-3	7e-4	2e-3	1e-3	2e-4	1e-3
EQ+PF+Carb	-	-	1e-2	9e-3	3e-3	6e-4	1e-3	1e-3	2e-4	1e-3
EQ+PF+Carb+Qtz	-	-	9e-3	8e-3	-	6e-4	1e-3	1e-3	2e-4	1e-3

Fonte – Lucas Oliveira, 2021

Tabela 20 – Desvios padrão do erro da validação dos modelos mineralógicos treinados. Todos os valores estão em g/g.

Desvios padrão	PF	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Plg+Prx
Um	0,02	0,09	0,10	0,09	0,05	0,02	0,03	0,02	0,02	0,04
Dois	0,04	0,18	0,20	0,18	0,10	0,04	0,06	0,04	0,04	0,08
Três	0,06	0,27	0,30	0,27	0,15	0,06	0,09	0,06	0,06	0,12

Fonte – Lucas Oliveira, 2021

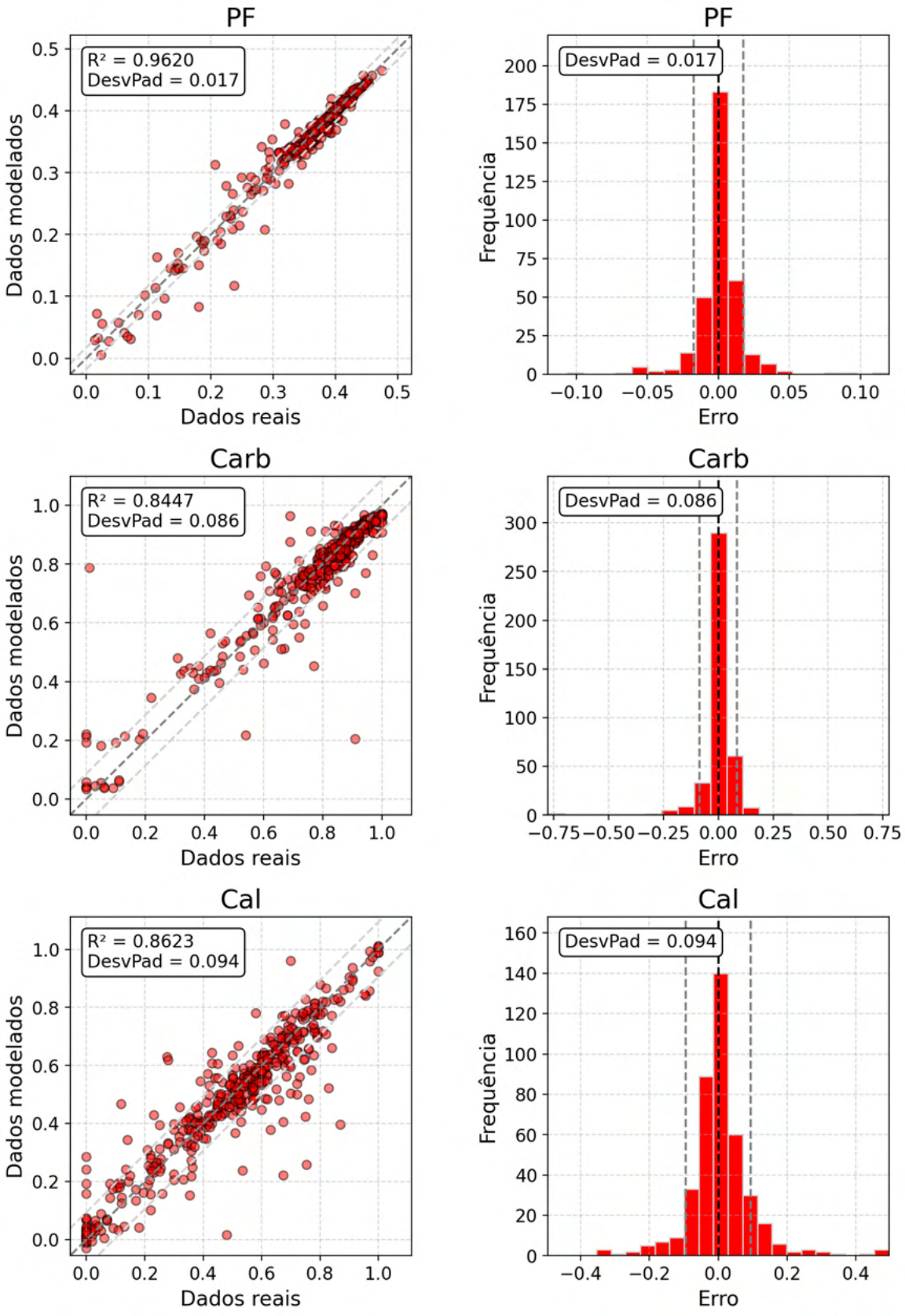


Figura 48 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de PF, carbonatos e calcita. DesvPad: Desvio padrão.

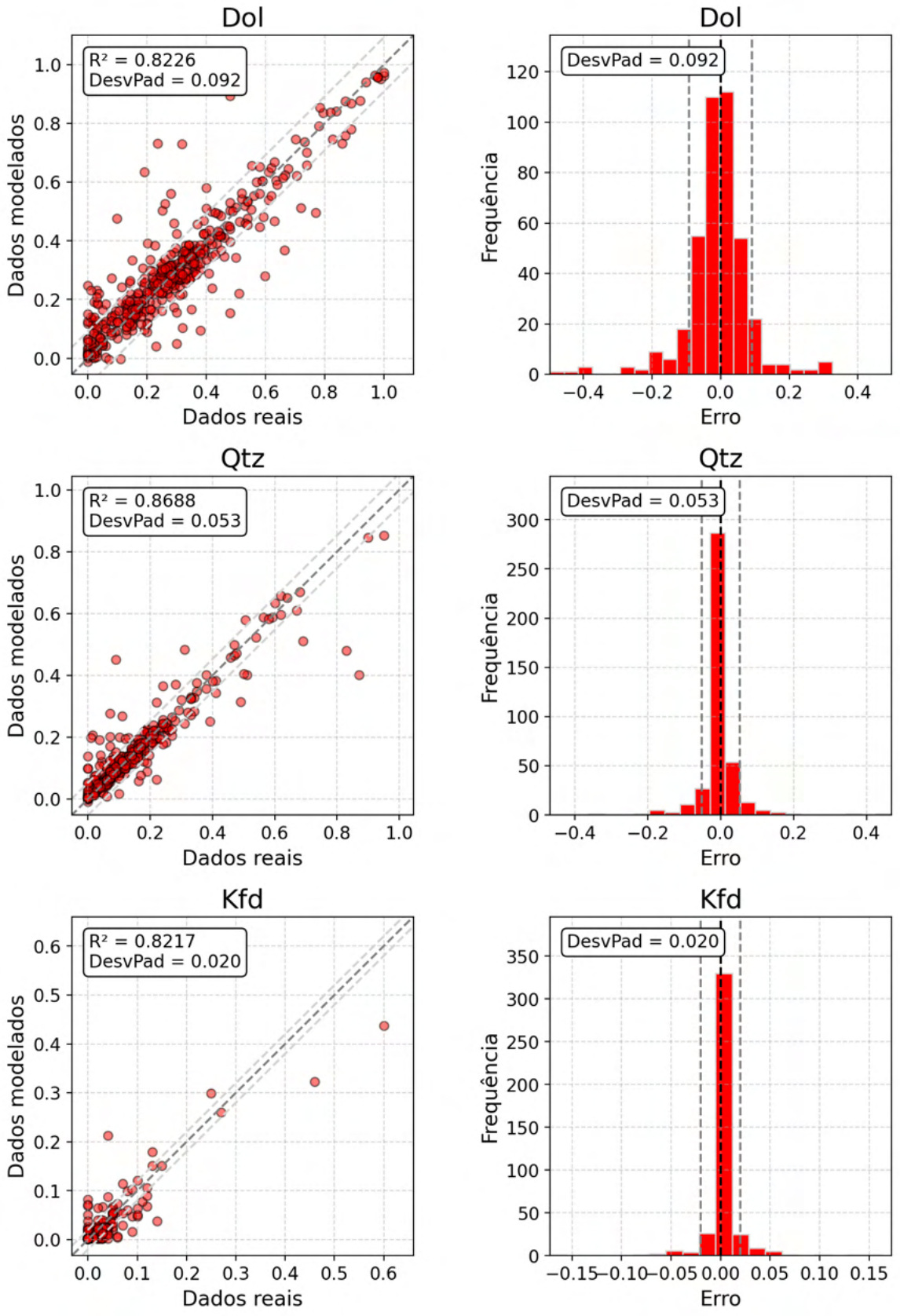


Figura 49 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de dolomita, quartzo e K-feldspato. DesvPad: Desvio padrão.

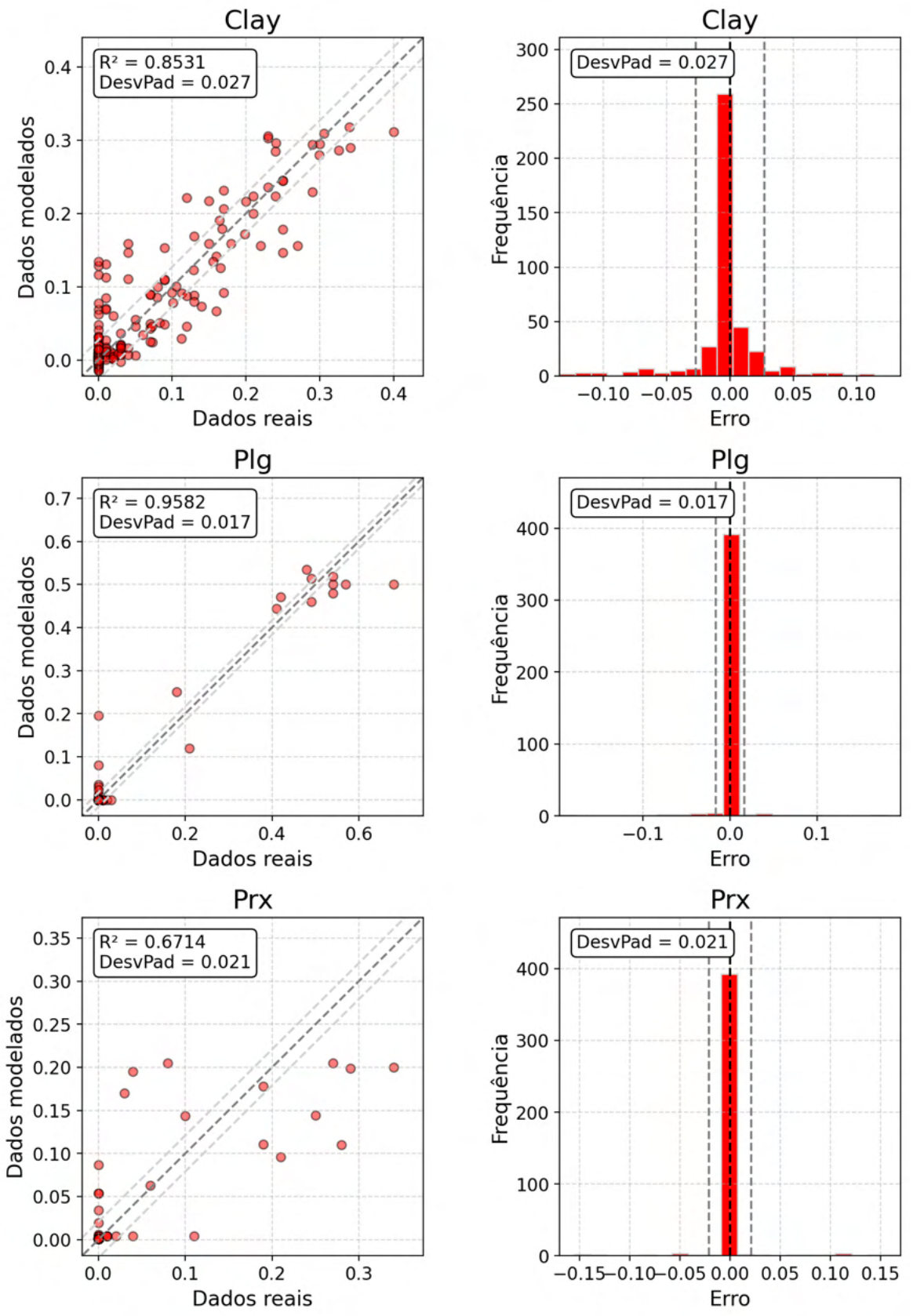


Figura 50 – Dados reais *versus* dados modelados e histograma do erro da base de validação para os modelos de argilas detríticas, plagioclásio e piroxênio. DesvPad: Desvio padrão.

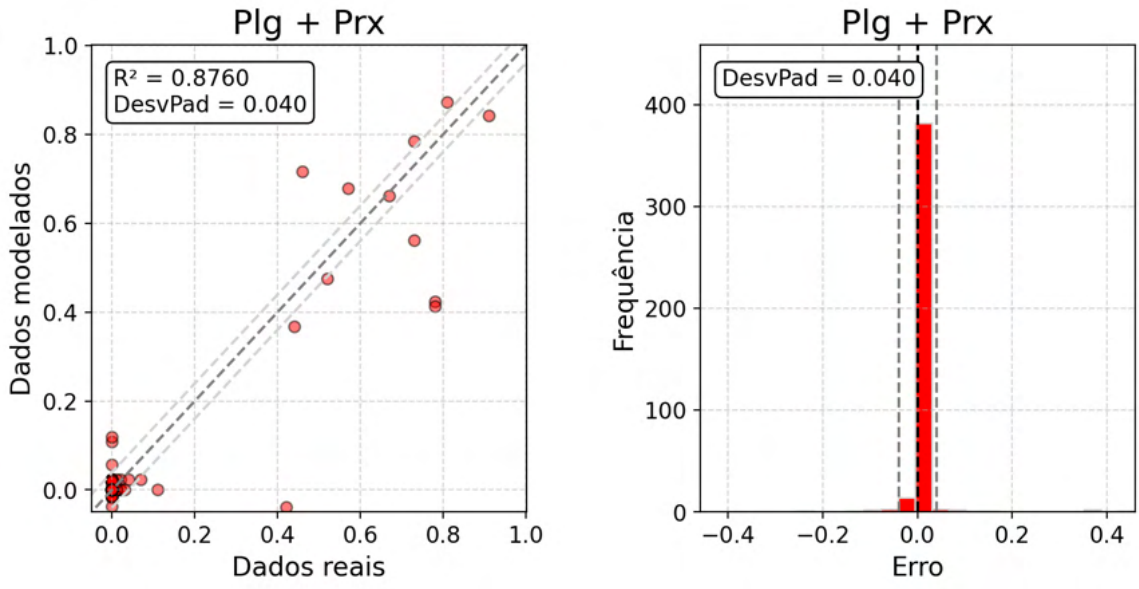


Figura 51 – Dados reais *versus* dados modelados e histograma do erro da base de validação para o modelo de plagioclásio + piroxênio. DesvPad: Desvio padrão.

rochas siliciclásticas e ígneas ou pela silicificação de carbonatos. O modelo de carbonatos é altamente influenciado pela PF. O modelo de calcita é impactado por Ca, Mg e carbonatos, e o modelo de dolomita é impactado por Mg, PF e Ca, comportamentos esperados nos carbonatos do pré-sal. A razoável importância do Na no modelo da dolomita pode se dar pela presença de Na e Mg em plagioclásios e piroxênios. O modelo identifica o aumento na concentração de Mg e Na como um indicativo da presença de plagioclásio e piroxênio, enquanto que um aumento apenas do Mg indica dolomita, fazendo do Na um elemento químico importante no modelo de dolomita.

O modelo de quartzo é influenciado por Si e carbonatos, refletindo as variações entre rochas carbonáticas e siliciclásticas ou ígneas. O K e Al são os elementos mais importantes no modelo de K-feldspato por serem os principais elementos de sua composição química. Ca e Fe, apesar de não estarem presentes na composição padrão do K-feldspato, podem ser usados pelo modelo para diferenciar plagioclásios e piroxênios. O modelo de argilas detríticas é impactado por diversas variáveis, sendo as mais importantes Mg, Na, PF, carbonatos e quartzo. Esse comportamento reflete a variada composição das argilas detríticas das rochas do pré-sal e as diferenças em seu ambiente deposicional.

Os modelos de plagioclásio e plagioclásio + piroxênio são muito impactados por Al, elemento presente na composição química padrão do plagioclásio. Os elementos Ti, Al, Fe e Na são importantes no modelo de piroxênio. A importância do Al e Na provavelmente se

deve a alta correlação entre piroxênio e plagioclásio. A importância do Ti e Fe pode estar relacionada a presença desses elementos em minerais da família do piroxênio.

4.3.3 Perfis mineralógicos

Na fase de teste, os modelos gerados pelo aprendizado escalonado foram aplicados aos perfis geoquímicos de três poços perfurados no pré-sal e seus resultados são apresentados nas figuras 53, 54 e 55. Os perfis das frações minerais foram comparados com análises de DRX de amostras de rocha coletadas nesses poços. É importante ressaltar que os perfis mineralógicos apresentam resolução vertical muito inferior às amostras de rocha, e a comparação entre esses dados não deve focar em valores exatos, mas sim em comportamentos gerais.

Carbonatos da Formação Barra Velha são observados no poço A, formados por calcita, dolomita e quartzo. As frações desses minerais estimadas pelo aprendizado escalonado estão em concordância com o observado pelas análises de DRX. A partir de X020 m as análises acusam uma diminuição da calcita, também observada no modelo de aprendizado de máquina. Nas profundidades finais, as amostras indicam um aumento das frações de quartzo, argilas detríticas e K-feldspato, possivelmente relacionado a rochas siliciclásticas que marcam a passagem para a Formação Itapema. O aprendizado escalonado foi capaz de indicar esse aumento.

As Formações Barra Velha e Itapema são observadas no poço B, separadas por uma camada de rocha siliciclástica. A Formação Barra Velha apresenta maiores frações de dolomita e quartzo quando comparado com a Formação Itapema. Essa característica é identificada pelo modelo de aprendizado escalonado e confirmado pelas análises de DRX. Entre as profundidades de X660 e X670 m, observa-se um aumento das frações de argilas detríticas e K-feldspato tanto nas análises de DRX quanto nos perfis mineralógicos.

Uma rocha composta majoritariamente por plagioclásio e piroxênio é observada na primeira metade do poço C. Esses minerais, identificados nas análises de DRX, demarcam uma espessa camada de rocha ígnea encaixada na Formação Barra Velha. O aprendizado escalonado foi capaz de identificar esses minerais. No restante do poço se observa uma concordância entre as frações minerais do modelo de aprendizado de máquina e as análises de DRX.

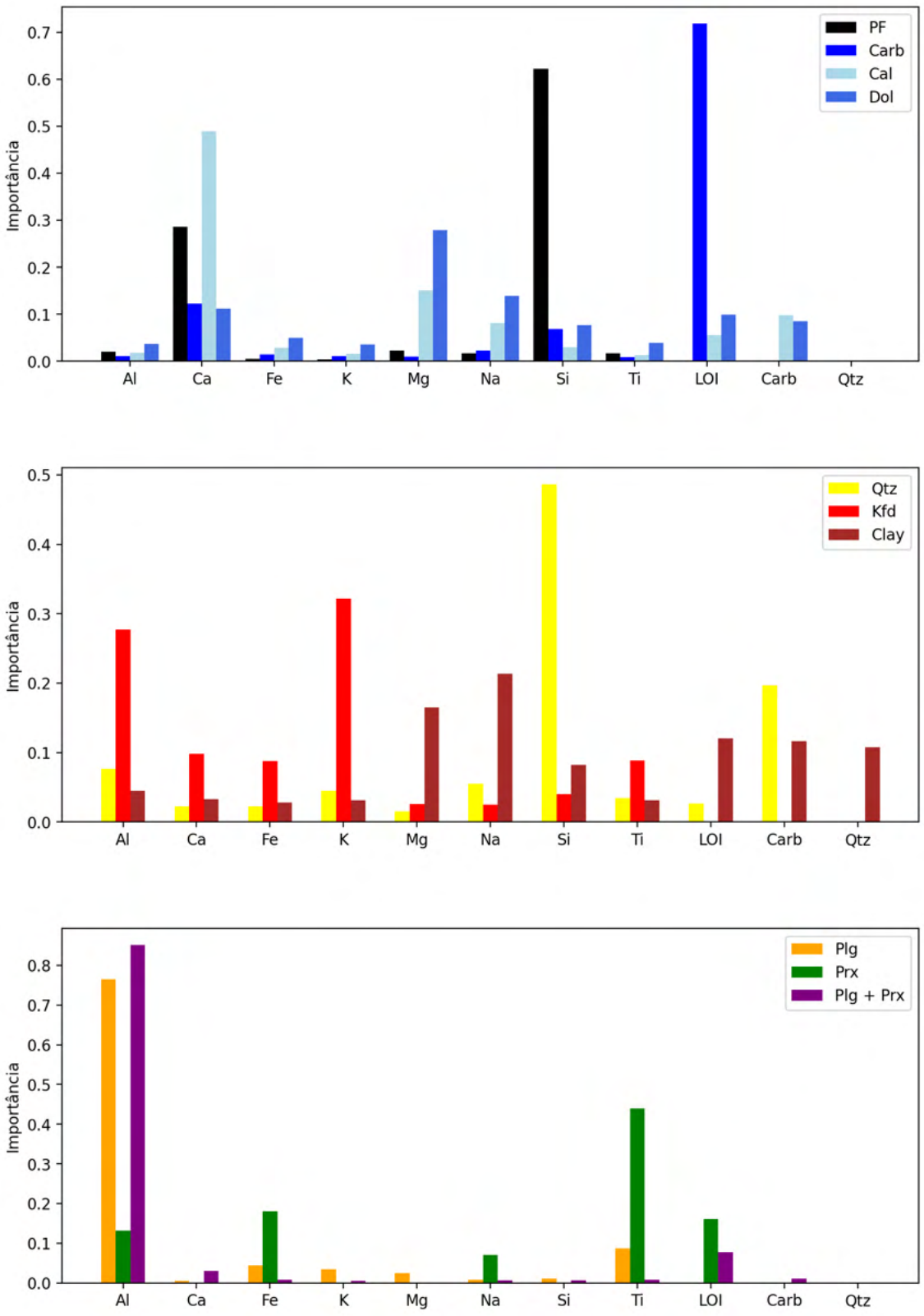


Figura 52 – Importância das variáveis de entrada dos modelos minerais.

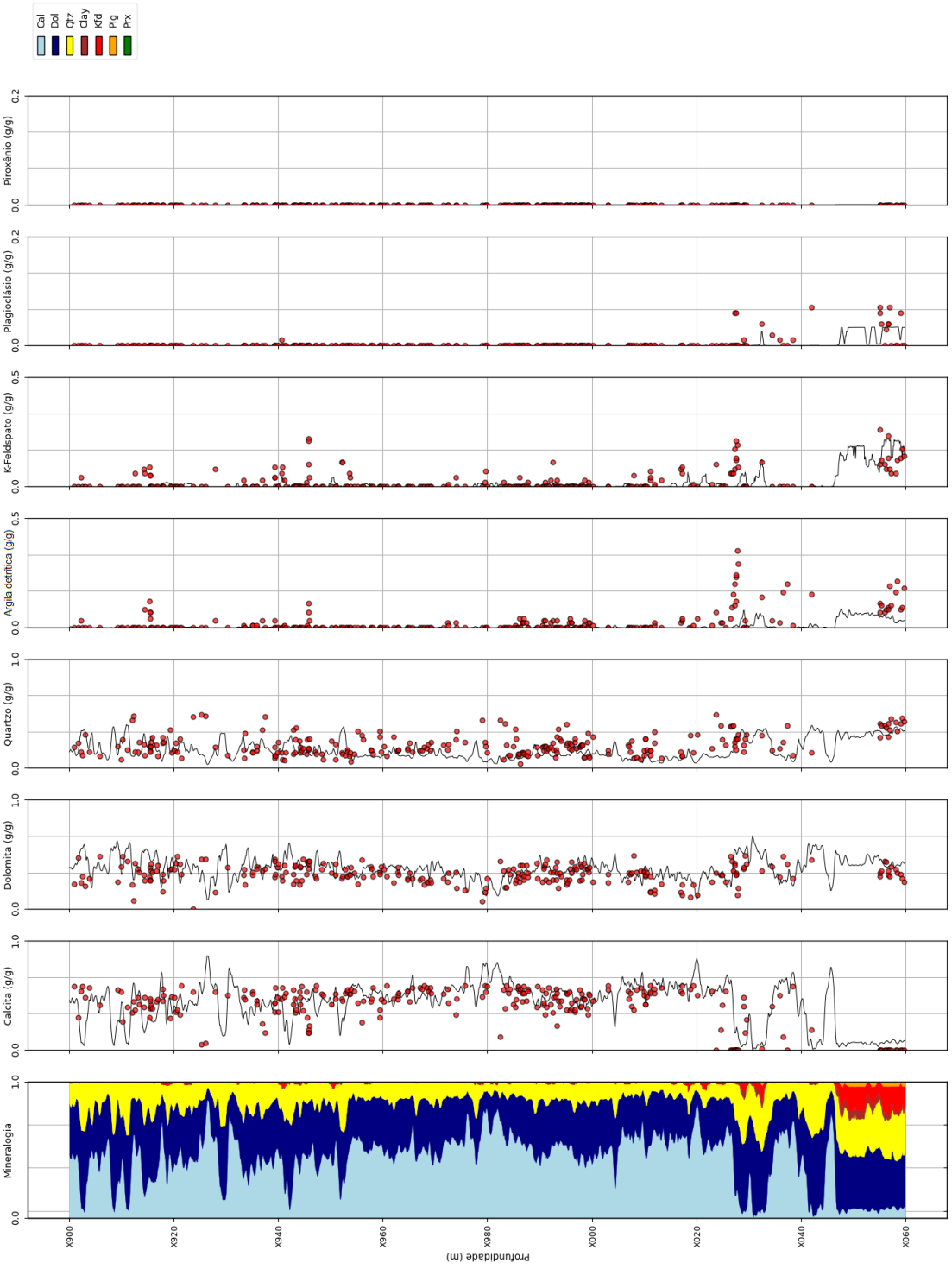


Figura 53 – Mineralogia obtida após a aplicação do aprendizado escalonado aos perfis geoquímicos do poço A e sua comparação com análises de DRX de amostras de rocha. O modelo honra as frações de calcita, dolomita e quartzo na profundidades iniciais. Ele também é capaz de estimar o aumento de argilas detríticas e K-feldspato nas profundidades finais.

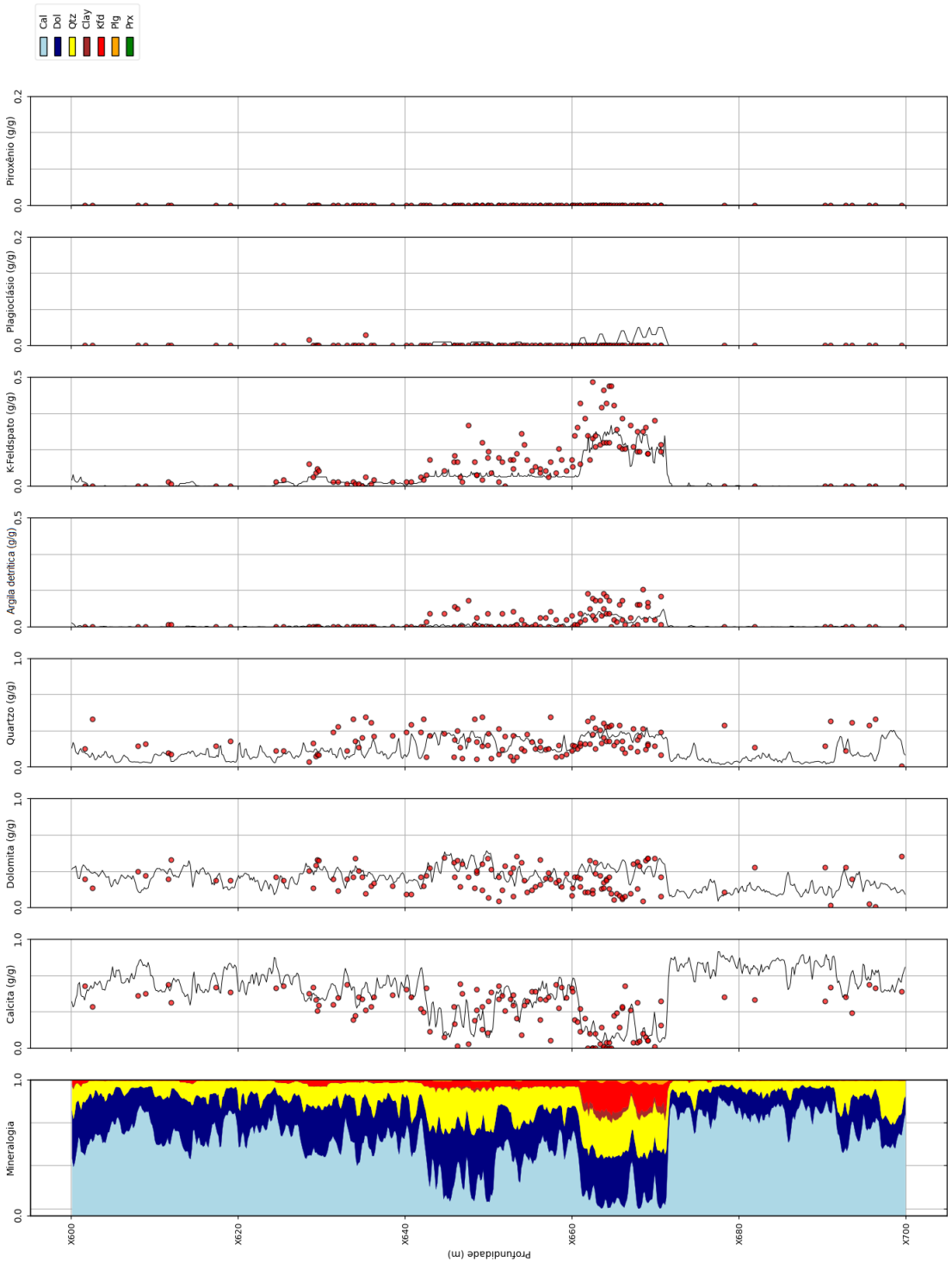


Figura 54 – Mineralogia obtida após a aplicação do aprendizado escalonado aos perfis geoquímicos do poço B e sua comparação com análises de DRX de amostras de rocha. O modelo foi capaz de detectar o aumento das frações de argilas detríticas e K-feldspato no meio do poço.

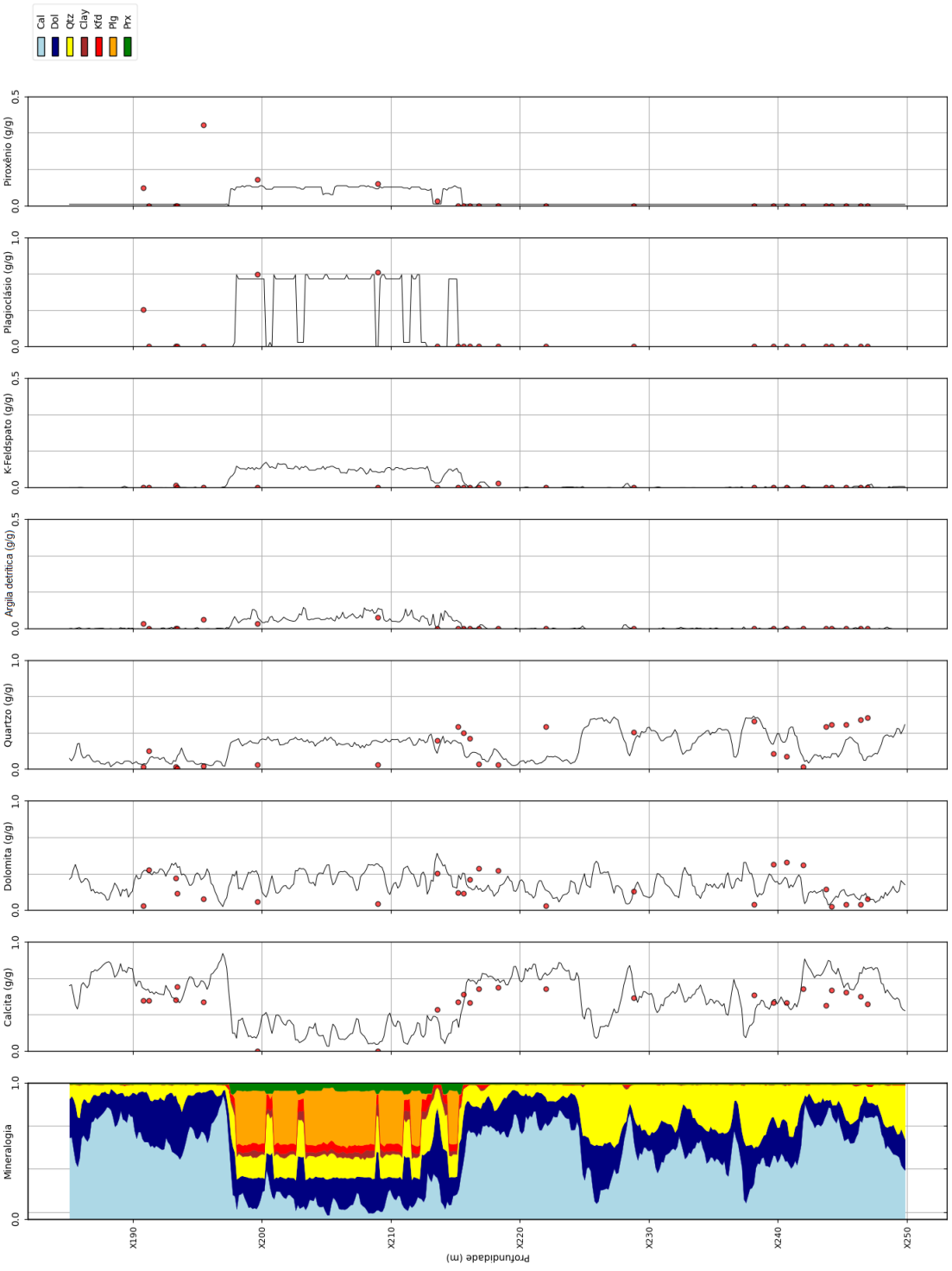


Figura 55 – Mineralogia obtida após a aplicação do aprendizado escalonado aos perfis geoquímicos do poço C e sua comparação com análises de DRX de amostras de rocha. O modelo detectou corretamente a camada de rocha ígnea, com altas proporções de plagioclásio e piroxênio.

Os gráficos de frações minerais reais *versus* modeladas pelos algoritmos de aprendizado de máquina são apresentados nas figuras 56 a 61. Apesar dos baixos R^2 observados, reflexo das diferenças de resolução vertical entre os perfis e informações obtidas em amostras de rocha, é possível observar que a maioria dos pontos está próximo da reta 1:1. Isso corrobora a qualidade dos modelos de aprendizado de máquina gerados para esses minerais.

4.4 Modelagem mineralógica pelo modelo híbrido

Conforme observado nos resultados da modelagem mineralógica por aprendizado de máquina, o enviesamento da base de dados teve impacto significante na estimativa da fração de alguns minerais. Como a maioria das amostras foi coletada nas rochas carbonáticas do pré-sal, compostas majoritariamente por calcita, dolomita e quartzo, frações altas de K-Feldspato, plagioclásio e piroxênio não puderam ser bem representadas. Sendo assim, as estimativas dos modelos para essas frações apresentou alta incerteza. Outro impacto do enviesamento da base de dados foi na própria seleção dos minerais do modelo de aprendizado de máquina: pirita, barita e argilas magnesianas não puderam ser incluídas no modelo.

Para considerar a incerteza na estimativa das frações minerais do modelo de aprendizado de máquina no modelo híbrido, o desvio padrão do erro observado no conjunto de validação apresentado na tabela 18 foi utilizado. A tabela 21 resume a incerteza dos perfis e frações minerais do modelo híbrido, apresentando a precisão das ferramentas e dos modelos de aprendizado de máquina, o impacto das condições de poço e a confiabilidade para as equações de reconstrução usadas no presente trabalho. Essas propriedades foram levadas em consideração na escolha da incerteza final.

Com as incertezas devidamente escolhidas, o modelo mineral híbrido foi aplicado a três poços perfurados no pré-sal da Bacia de Santos e seus resultados são apresentados nas figuras 62 a 73. As frações minerais estimadas foram comparadas com análises de composição mineral de amostras de rocha coletadas nesses poços e os perfis reconstruídos são comparados com os reais. O perfil de erro e o de diâmetro de poço também são apresentados. Assim como nos modelos de aprendizado de máquina, os perfis mineralógicos gerados apresentam resolução vertical muito inferior às amostras de rocha, e a comparação entre esses dados não deve focar em valores exatos, mas sim em comportamentos gerais.

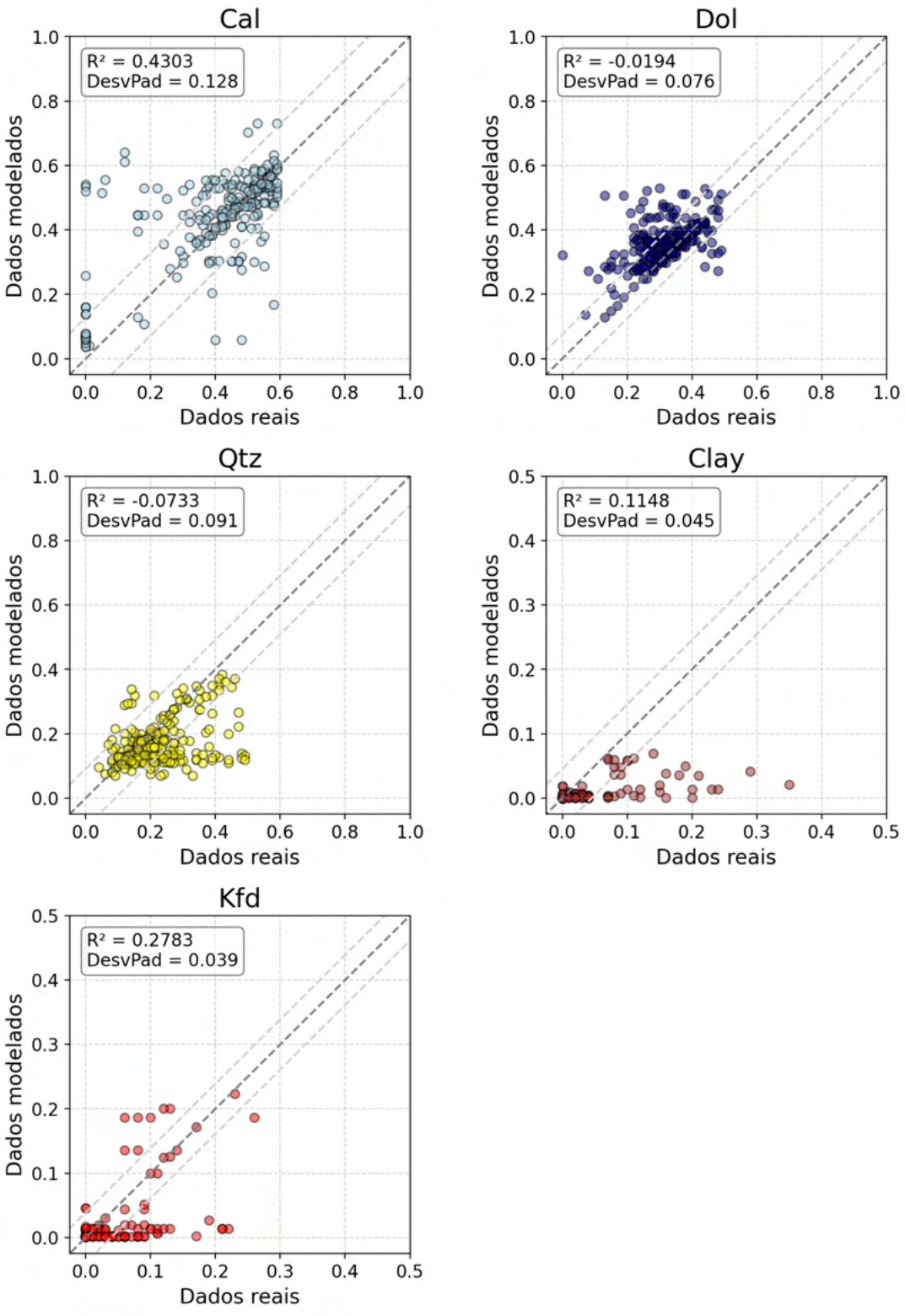


Figura 56 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica e K-feldspato do poço A. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

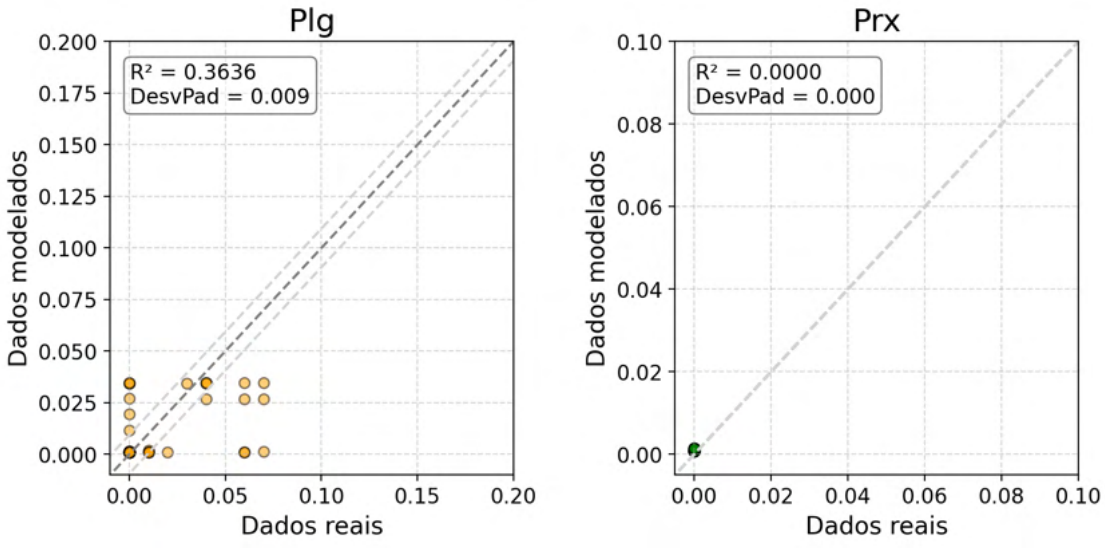


Figura 57 – Dados reais *versus* dados modelados para plagioclásio e piroxênio do poço A. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

As figuras 62 a 65 apresentam os resultados do Poço D. As análises de DRX indicaram uma quantidade expressiva de argilas magnesianas, padrão representado pelas frações estimadas pelo modelo híbrido. Além disso, a calcita, dolomita, quartzo e K-feldspato apresentaram valores coerentes com as análises de rocha. O modelo não estimou quantidades significativas de plagioclásio, mineral observado em algumas amostras em frações abaixo de 5%. Os perfis reconstruídos se assemelharam aos reais, com a ressalva de que a água de argila modelada apresentou valores levemente menores do que a medida, possivelmente causado por pequenas variações na porosidade das argilas.

Os resultados do Poço E são apresentados nas figuras 66 a 69. O modelo híbrido foi capaz de honrar a rocha predominantemente composta por calcita, com baixas frações de quartzo, dolomita e argilas magnesianas. Os perfis reconstruídos apresentaram valores muito próximos dos reais, variando dentro do intervalo de incerteza utilizado e gerando erros no geral inferiores a 0,5.

As frações minerais e os perfis reconstruídos do Poço F são apresentadas nas figuras 70 a 73. Novamente, o modelo híbrido foi capaz de honrar os principais minerais, incluindo as argilas magnesianas. Os perfis reconstruídos se assemelharam aos reais, havendo discrepâncias significativas com as frações de calcita e dolomita estimadas pelo modelo de aprendizado de máquina. Como esse modelo não considera as argilas magnesianas, a alta concentração de Mg adquirida pela ferramenta geoquímica acarretou em frações

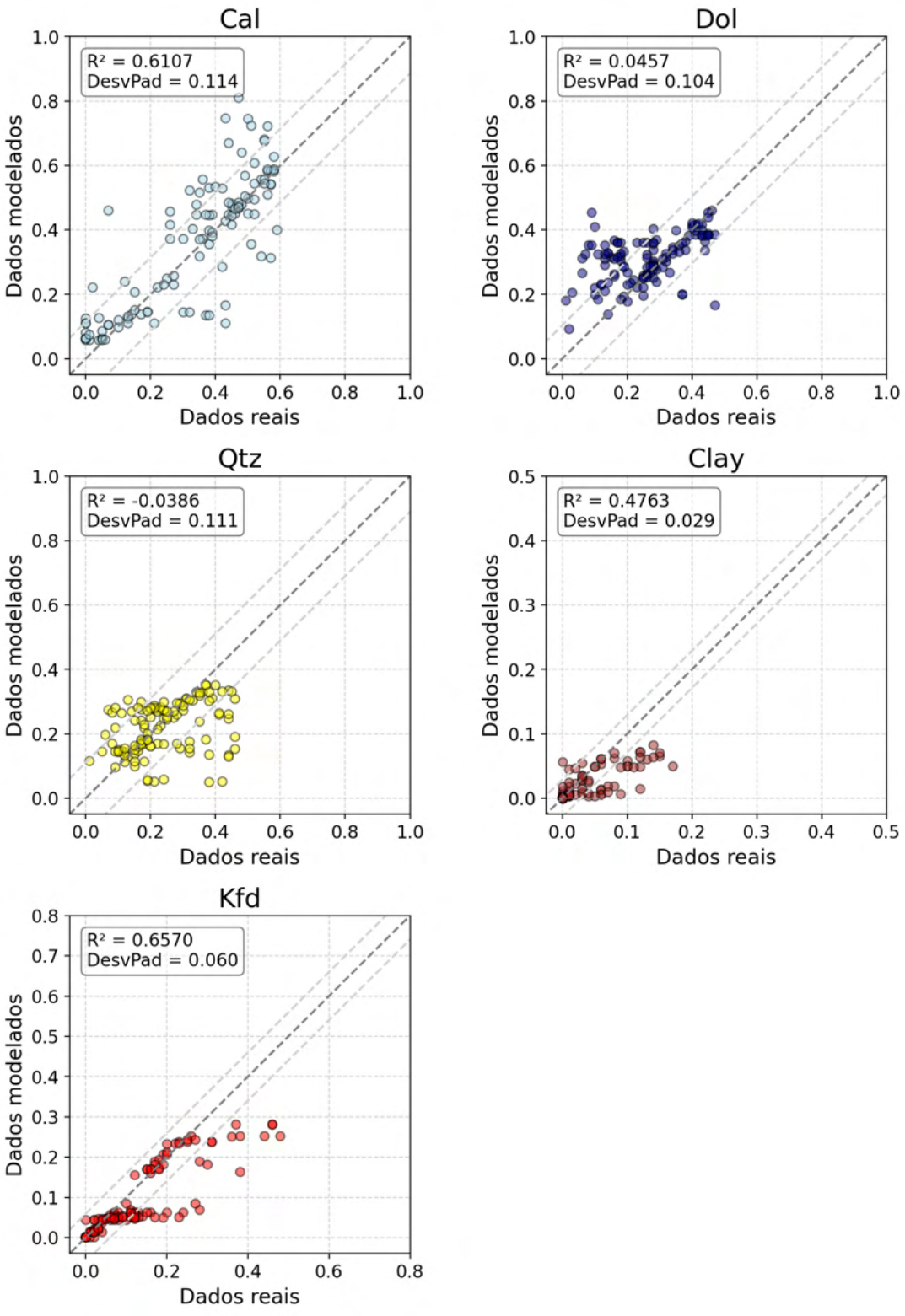


Figura 58 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica e K-feldspato do poço B. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

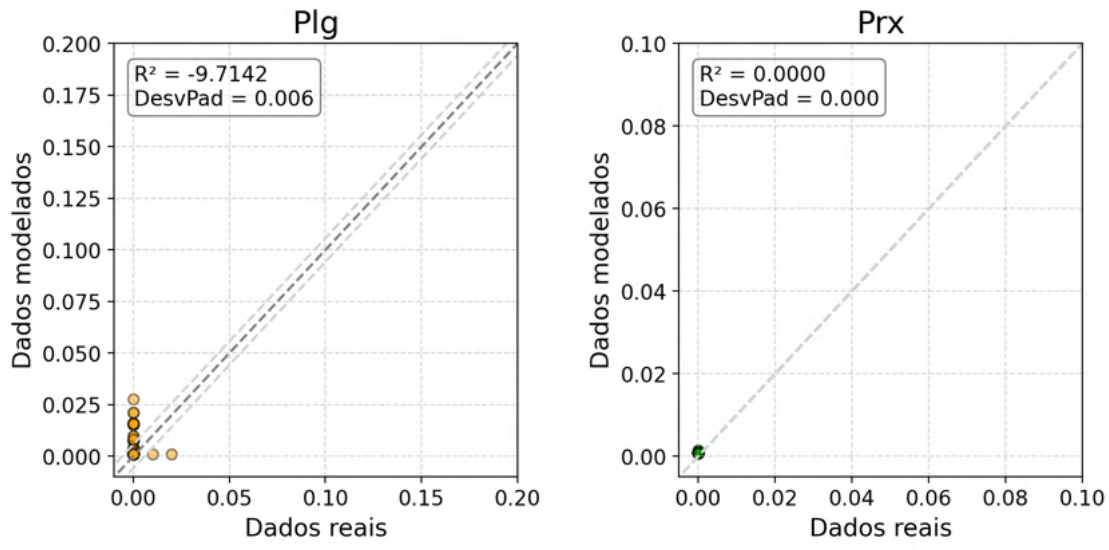


Figura 59 – Dados reais *versus* dados modelados para plagioclásio e piroxênio do poço B. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

superestimadas de dolomita. Entretanto, a inclusão dos demais perfis fizeram com que o modelo híbrido recalculasse a dolomita, agora sob a presença das argilas magnesianas. A incerteza estipulada para as frações minerais do modelo de aprendizado de máquina, principalmente a da dolomita, deram liberdade para o modelo híbrido proporcionar resultados mais coerentes.

Os gráficos de frações minerais reais *versus* modeladas pelo modelo híbrido são apresentados nas figuras 74 a 79. Assim como no modelo de aprendizado de máquina, a alta dispersão dos pontos gerada pelas diferenças de resolução vertical entre os perfis e as análises laboratoriais das amostras de rocha geraram baixos R^2. Entretanto, é possível observar que a maioria dos pontos está próximo da reta 1:1, demonstrando a qualidade das frações minerais geradas pelo modelo híbrido, em especial para a argila magnesiana.

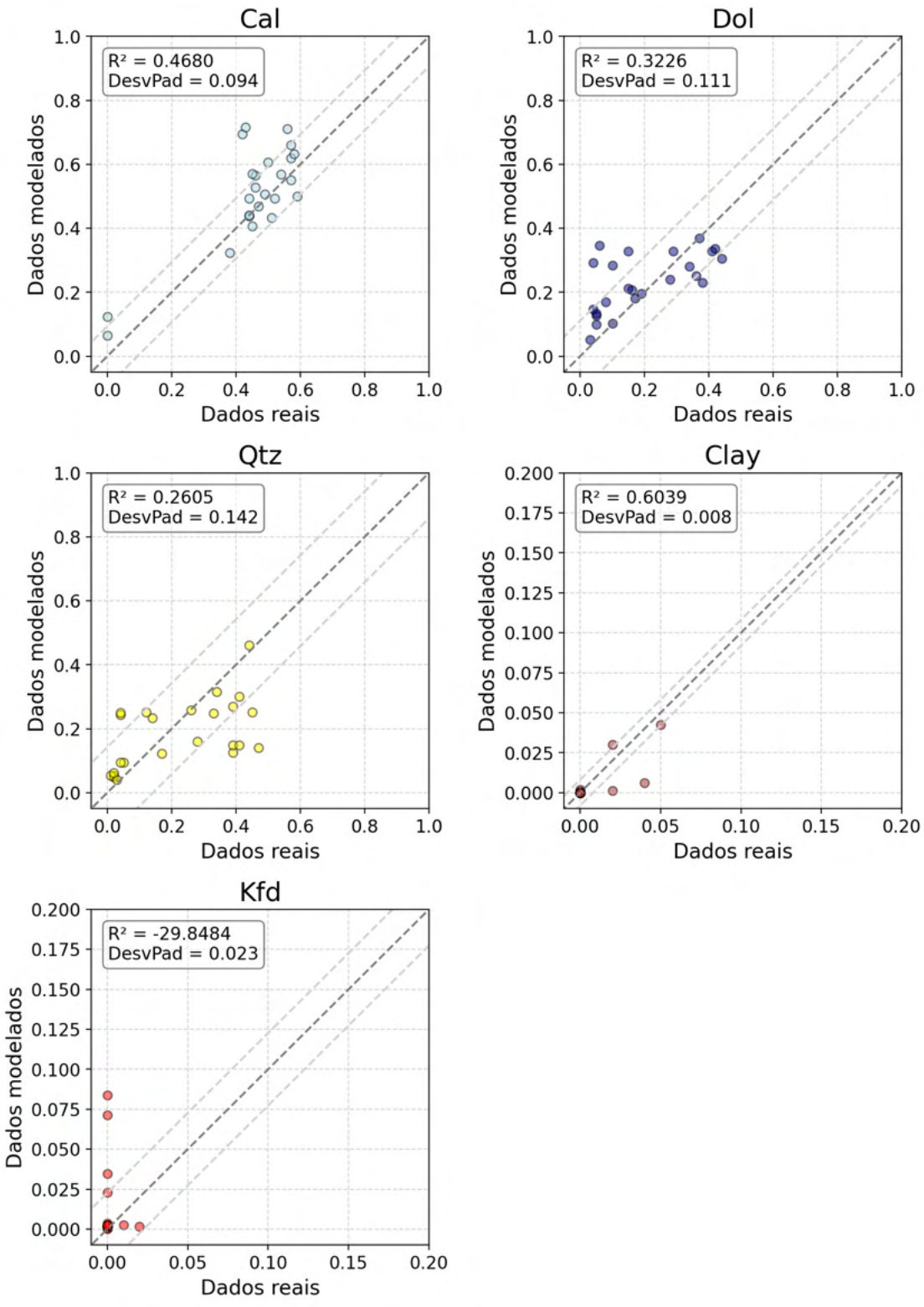


Figura 60 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica e K-feldspato do poço C. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

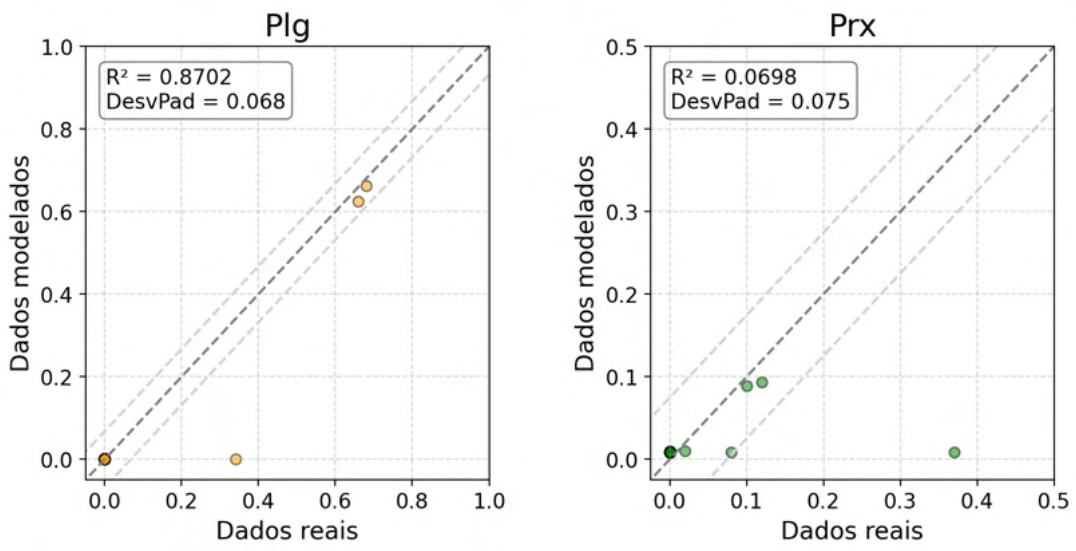


Figura 61 – Dados reais *versus* dados modelados para plagioclásio e piroxênio do poço C. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

Tabela 21 – Resumo da precisão das ferramentas e dos modelos de aprendizado de máquina, do impacto das condições de poço e da confiabilidade para as equações de reconstrução usadas na etapa probabilística do modelo híbrido. Essas propriedades foram levadas em consideração na escolha da incerteza final atribuída às equações. As precisões das ferramentas de perfilagem foram extraídas de Schlumberger (2015) e as dos modelos de aprendizado de máquina foram extraídos da tabela 18.

Perfil	Precisão da ferramenta/ modelo	Impacto das condições de poço	Confiabilidade das equações de reconstrução	Incerteza final
Unidade	-	-	Alta	0,01
DEN	±0,02 g/cm^3	Médio	Alta	0,02
U	PEF: ±0,15	Alto	Alta	5,0
PhiT	±0,01 m^3/m^3	Médio	Alta	0,02
CBW	>0,01 m^3/m^3	Médio	Média	0,03
FF	>0,01 m^3/m^3	Médio	Média	0,03
Al	±0,02 g/g	Alto	Média	0,05
Ca	±0,005 g/g	Alto	Alta	0,02
Fe	±0,01 g/g	Alto	Média	0,03
K	±0,01 g/g	Alto	Média	0,03
Mg	±0,03 g/g	Alto	Alta	0,05
Si	±0,005 g/g	Alto	Alta	0,02
S	±0,01 g/g	Alto	Baixa	0,03
Ti	±0,01 g/g	Alto	Baixa	0,03
Cal	±0,096 g/g	Alto	Alta	0,1
Dol	±0,088 g/g	Alto	Média	0,3
Qtz	±0,051 g/g	Alto	Alta	0,1
Kfd	±0,021 g/g	Alto	Média	0,2
Clay	±0,028 g/g	Alto	Média	0,2
Plg	±0,042 g/g	Alto	Baixa	0,3
Prx	±0,042 g/g	Alto	Baixa	0,3

Fonte – Lucas Oliveira, 2021

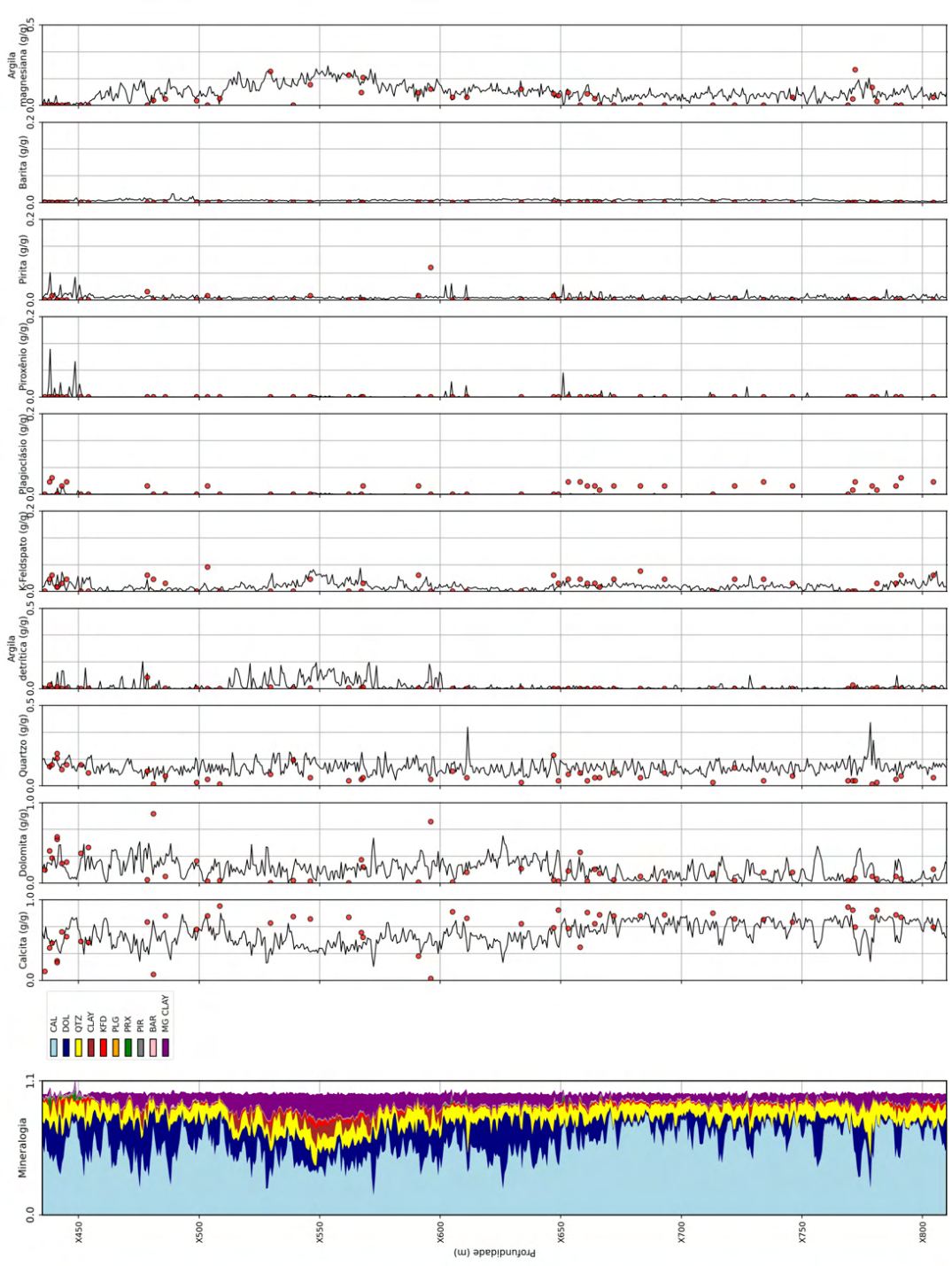


Figura 62 – Mineralogia obtida após a aplicação do modelo híbrido no poço D e sua comparação com análises de DRX de amostras de rocha. O modelo gerou equivalências para com as frações reais de calcita, dolomita, quartzo e argilas magnesianas, minerais com as mais altas frações nesse poço.

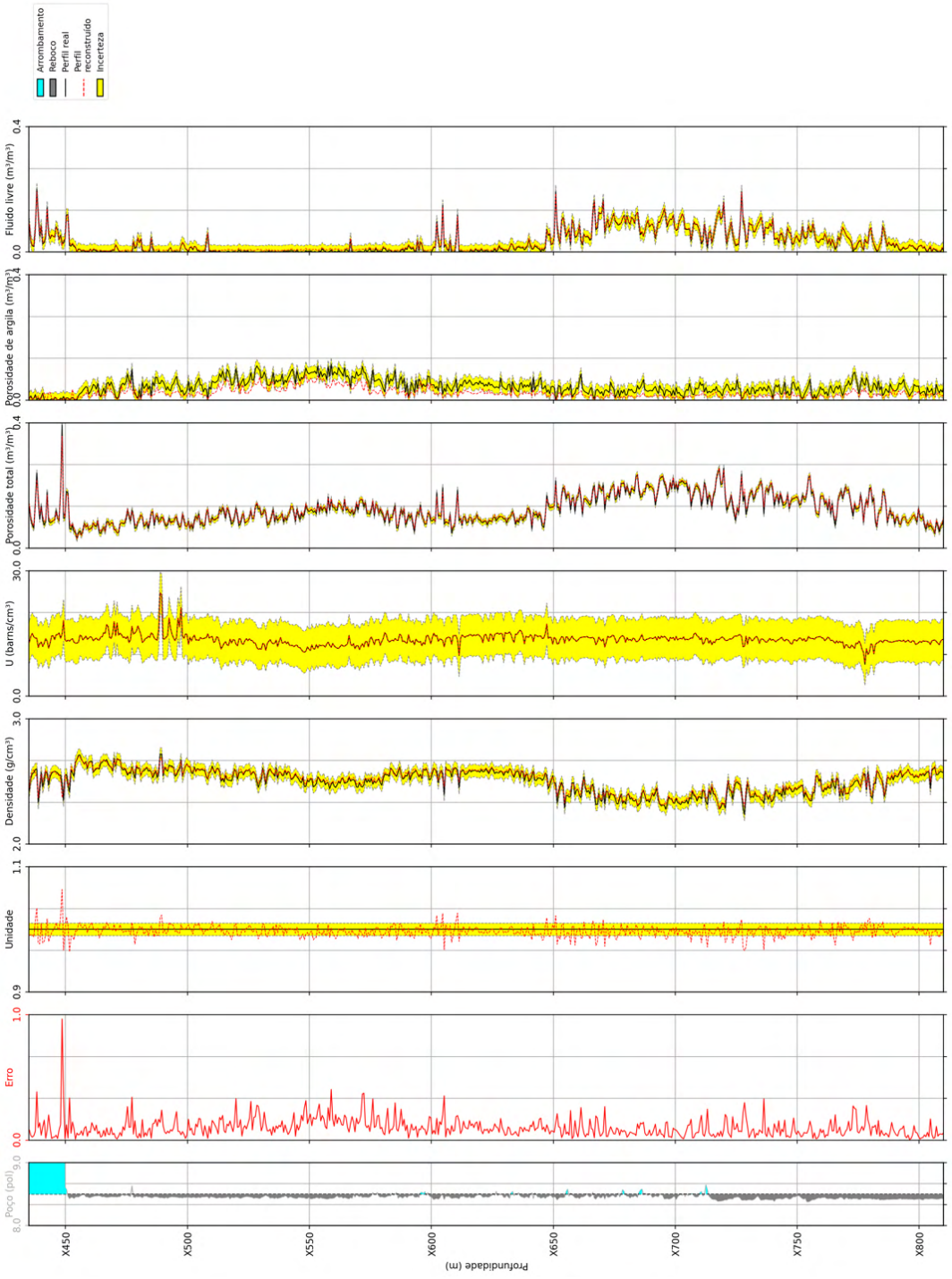


Figura 63 – Diâmetro de poço, erro e comparação entre os perfis unidade, densidade, U, porosidade total, água de argila e fluido livre reconstruídos e reais para o Poço D. A região amarela em volta dos perfis representa a incerteza.

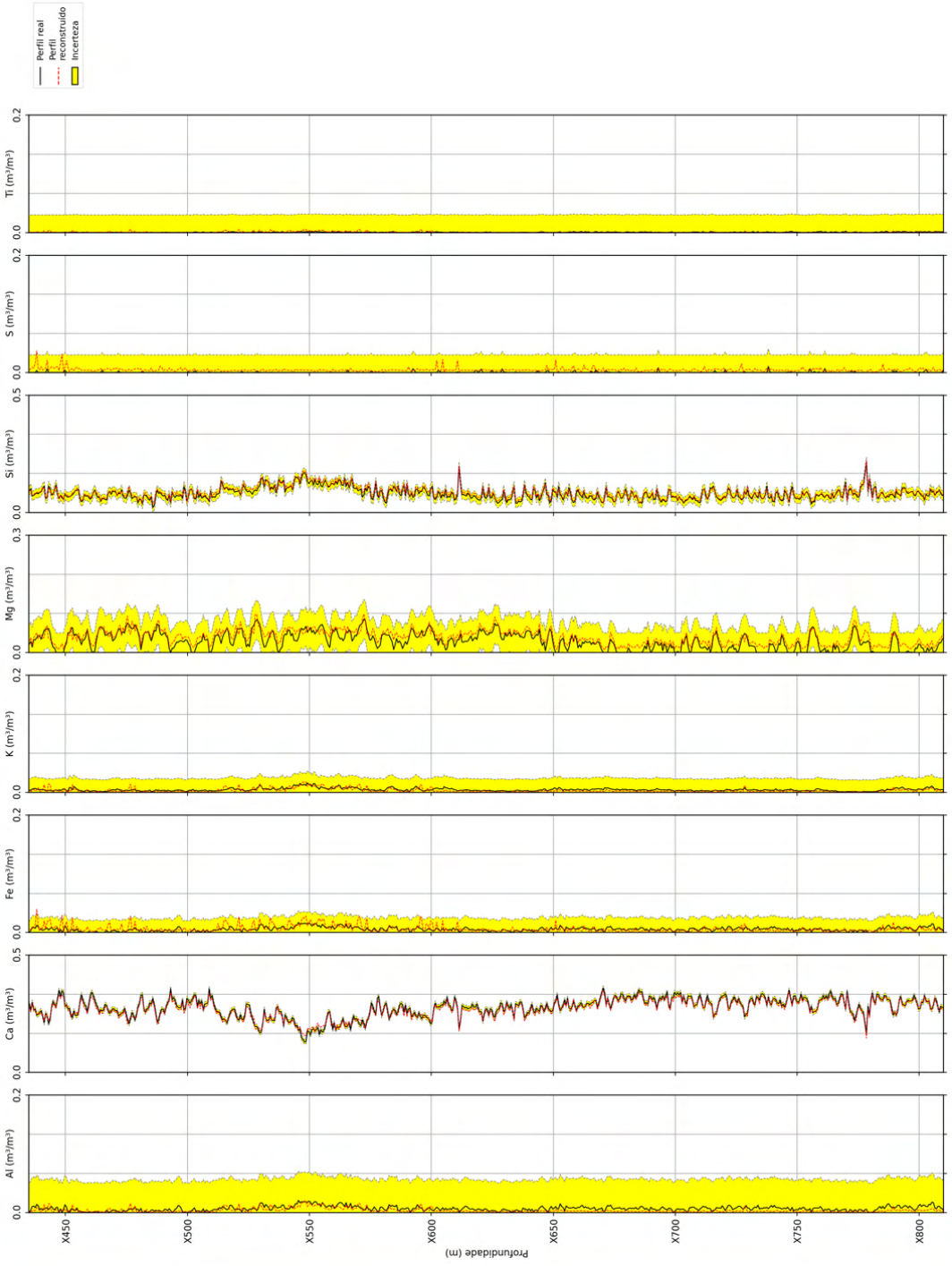


Figura 64 – Comparação entre os perfis geoquímicos reconstruídos e reais para o Poço D. A região amarela em volta dos perfis representa a incerteza.

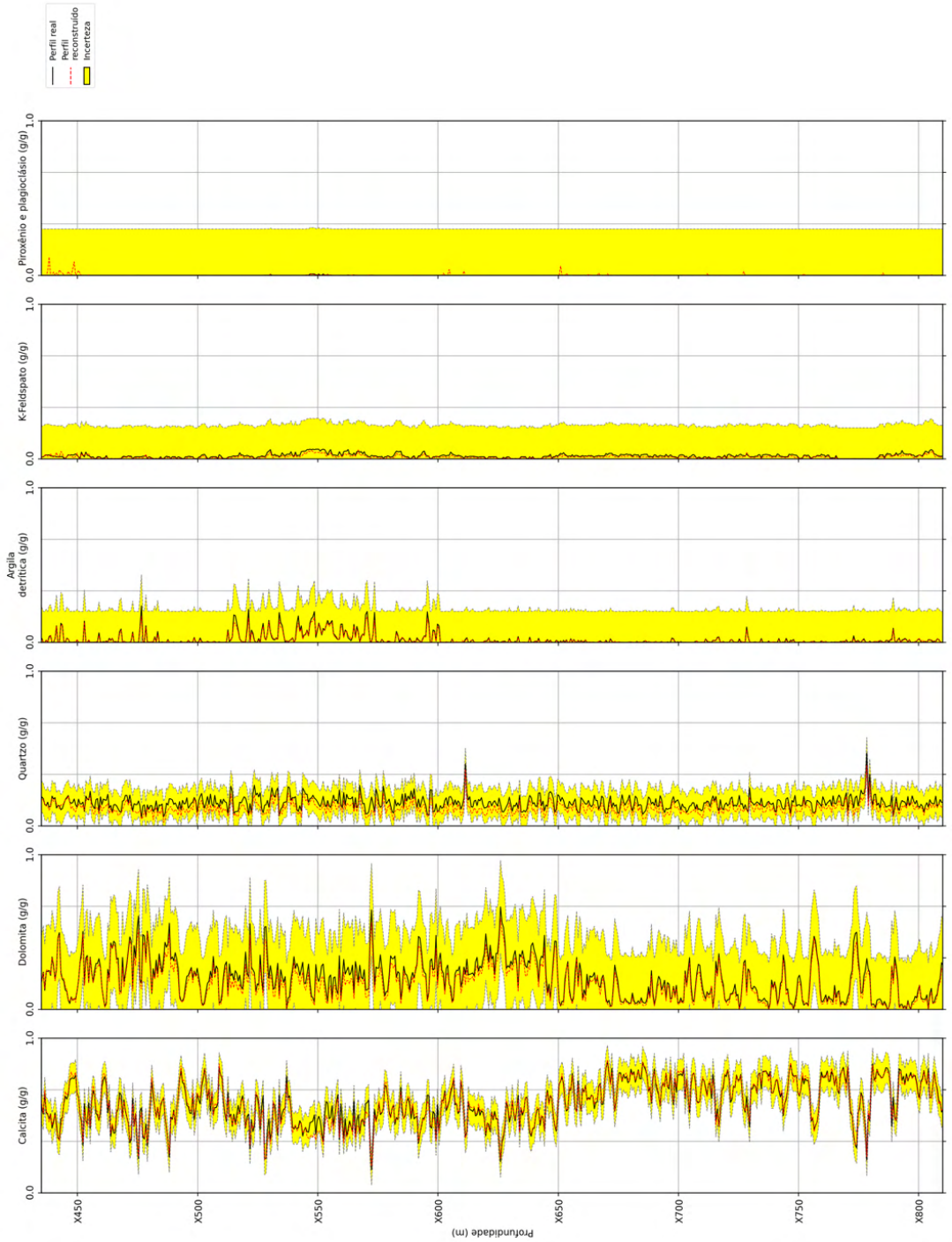


Figura 65 – Comparação entre as frações minerais estimadas pelo modelo de aprendizado de máquina reconstruídas e reais para o Poço D. A região amarela em volta dos perfis representa a incerteza.

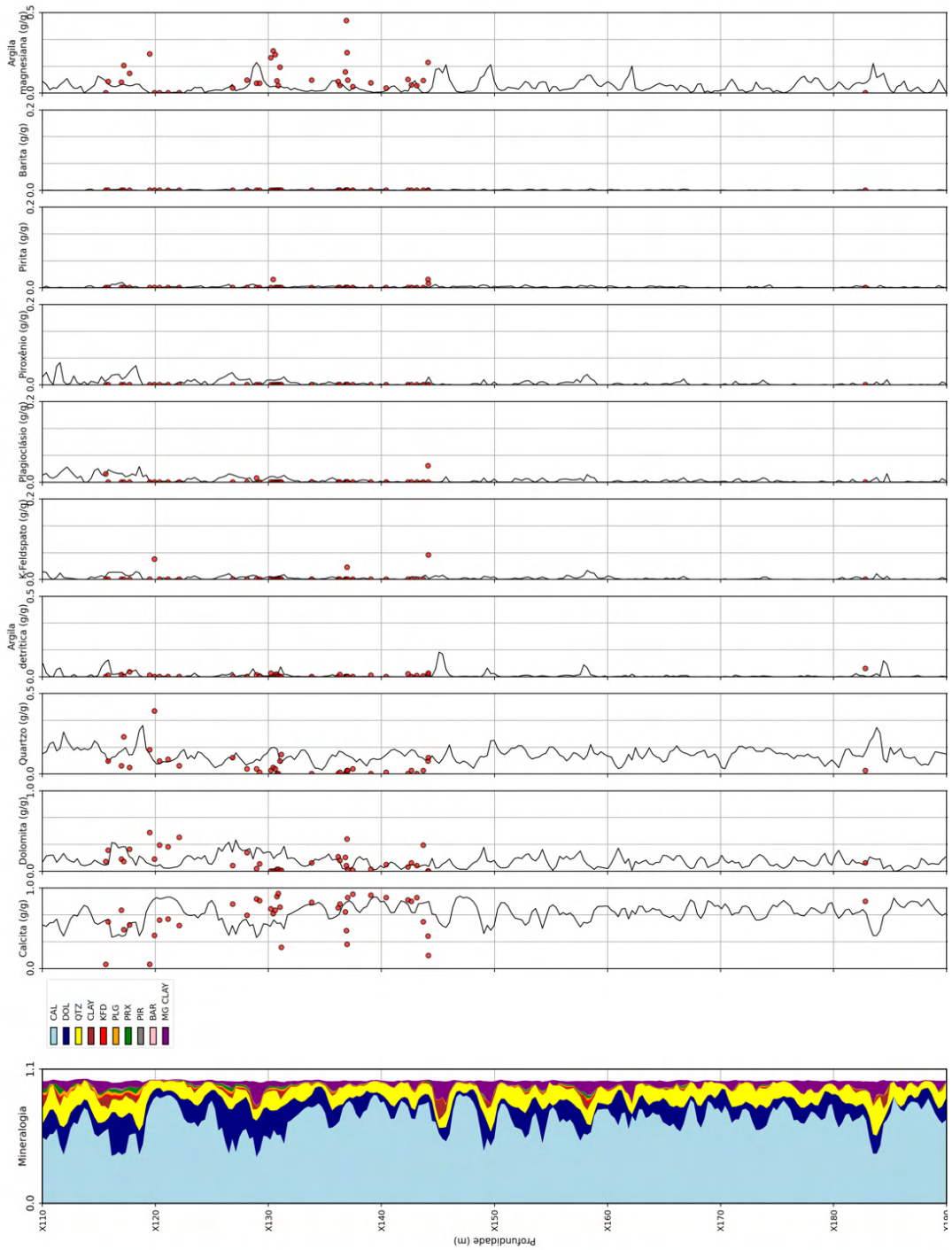


Figura 66 – Mineralogia obtida após a aplicação do modelo híbrido no poço E e sua comparação com análises de DRX de amostras de rocha. O modelo gerou equivalências para com as frações reais de calcita, dolomita e quartzo, minerais com as mais altas frações nesse poço. As altas frações de argilas magnesianas observadas nas análises de DRX não foram observadas na estimativa do modelo, devido a diferença de resolução. Porém, tendências gerais foram representadas.

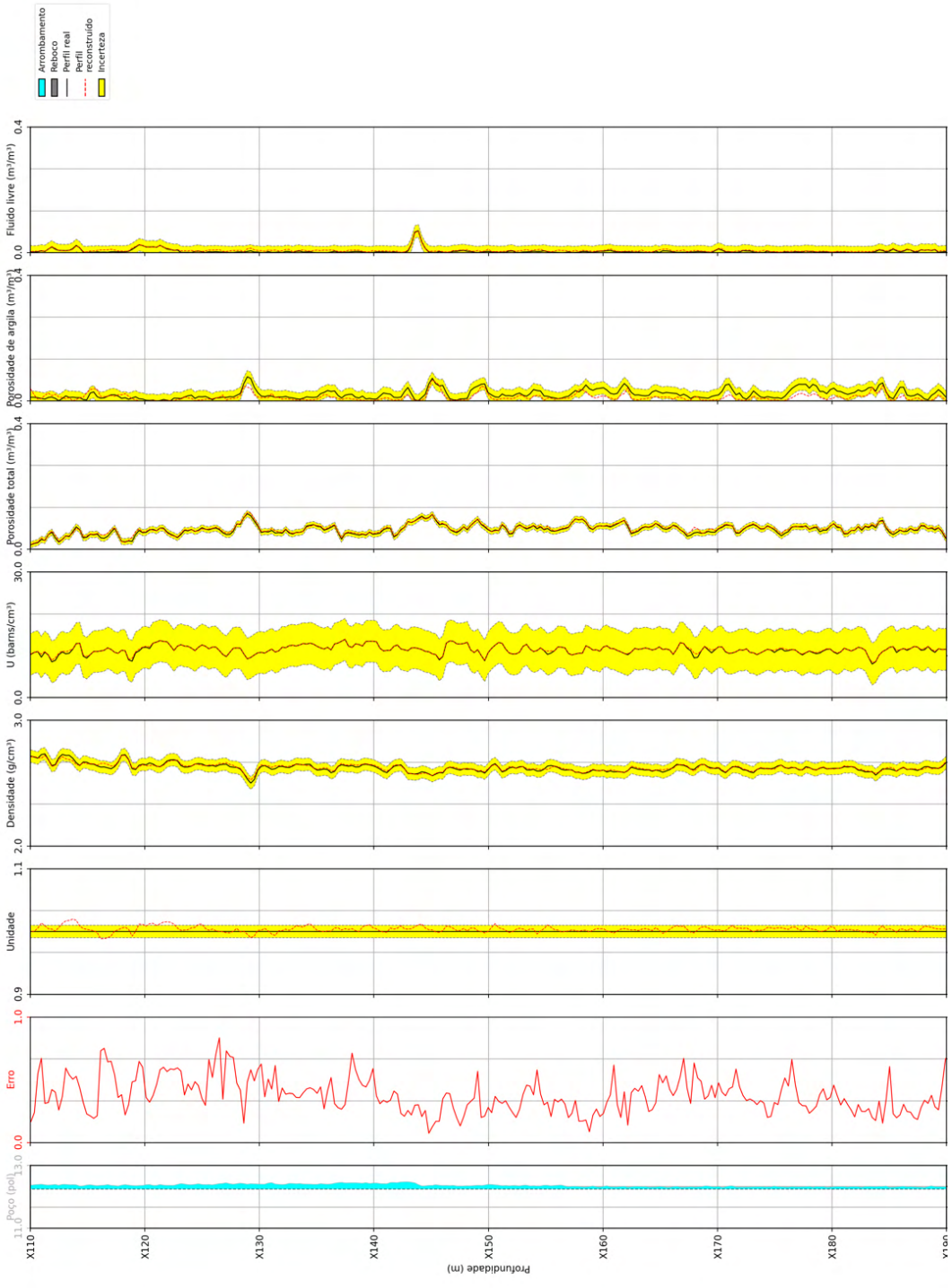


Figura 67 – Diâmetro de poço, erro e comparação entre os perfis unidade, densidade, U, porosidade total, água de argila e fluido livre reconstruídos e reais para o Poço E. A região amarela em volta dos perfis representa a incerteza.

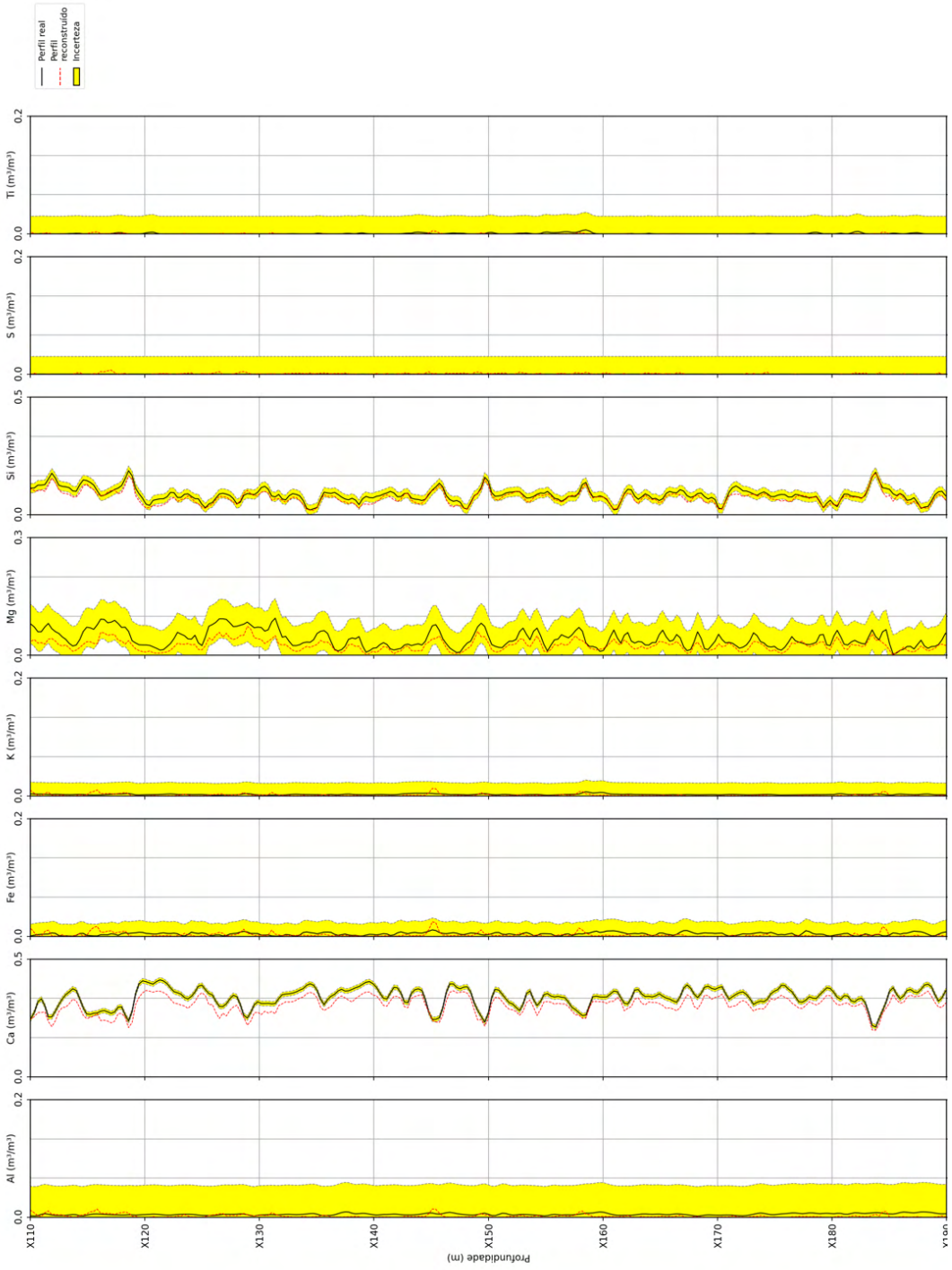


Figura 68 – Comparação entre os perfis geoquímicos reconstruídos e reais para o Poço E. A região amarela em volta dos perfis representa a incerteza.

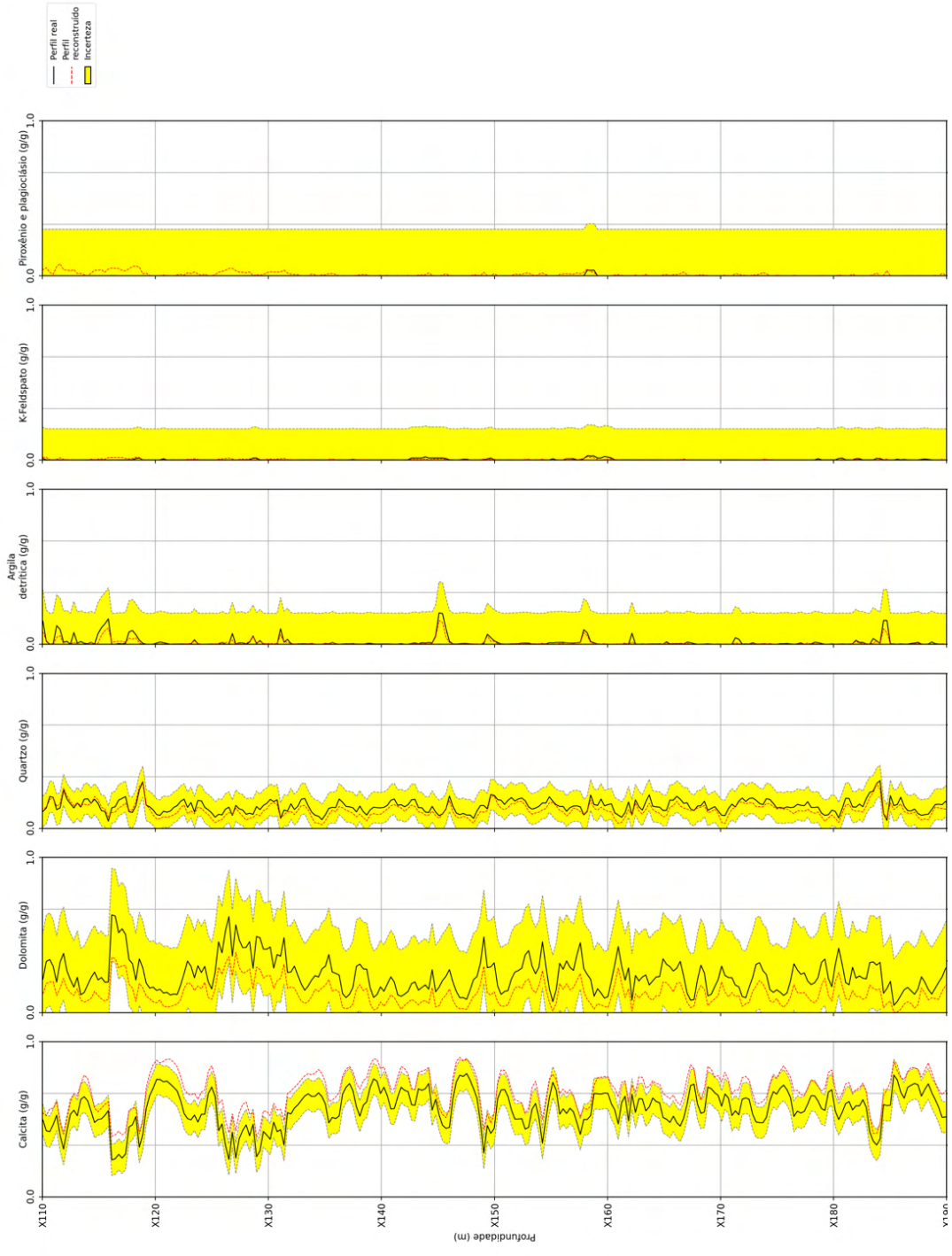


Figura 69 – Comparação entre as frações minerais estimadas pelo modelo de aprendizado de máquina reconstruídas e reais para o Poço E. A região amarela em volta dos perfis representa a incerteza.

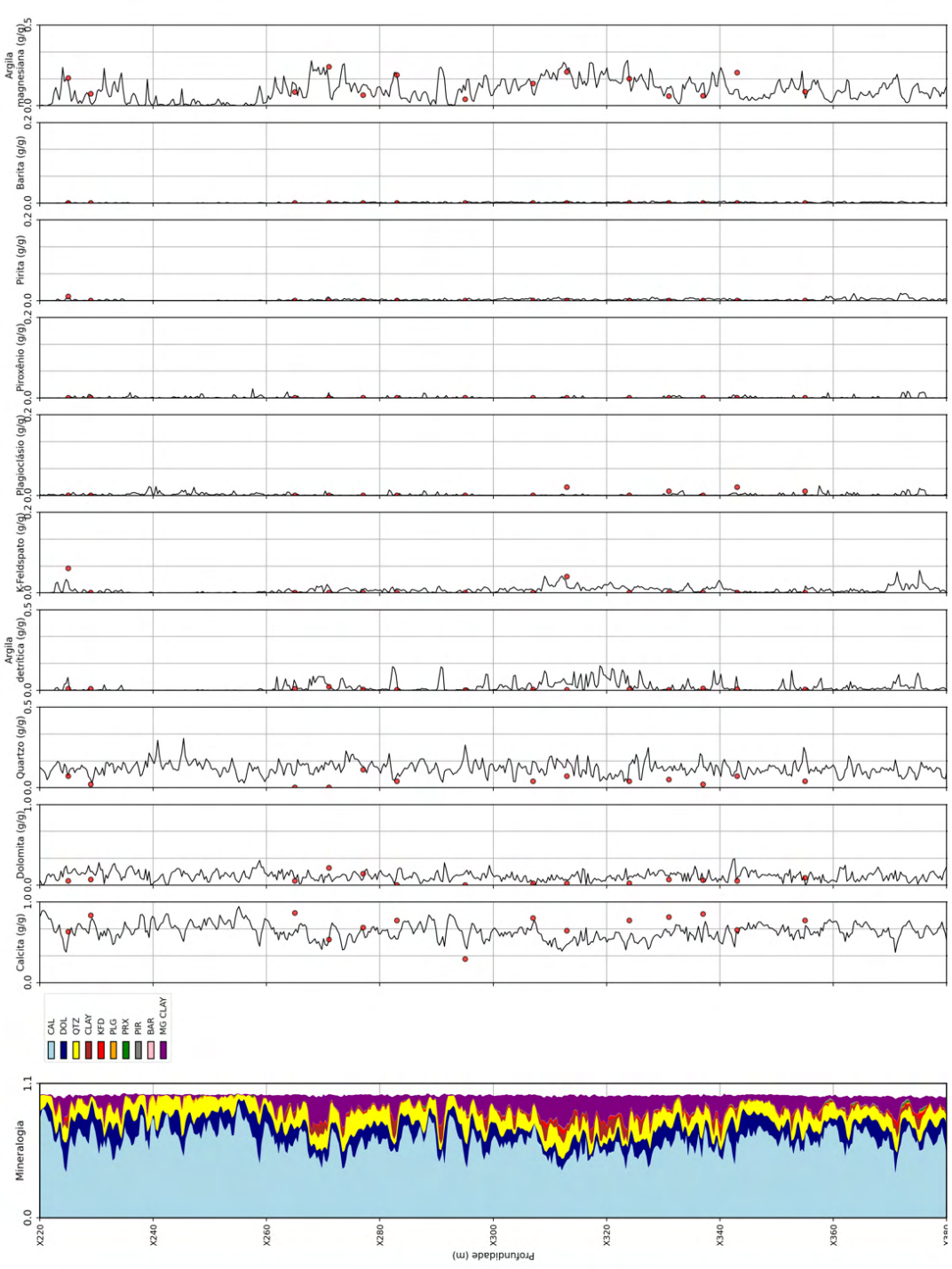


Figura 70 – Mineralogia obtida após a aplicação do modelo híbrido no poço F e sua comparação com análises de DRX de amostras de rocha. O modelo gerou equivalências para com as frações reais de calcita, dolomita, quartzo e argilas magnesianas, minerais com as mais altas frações nesse poço.

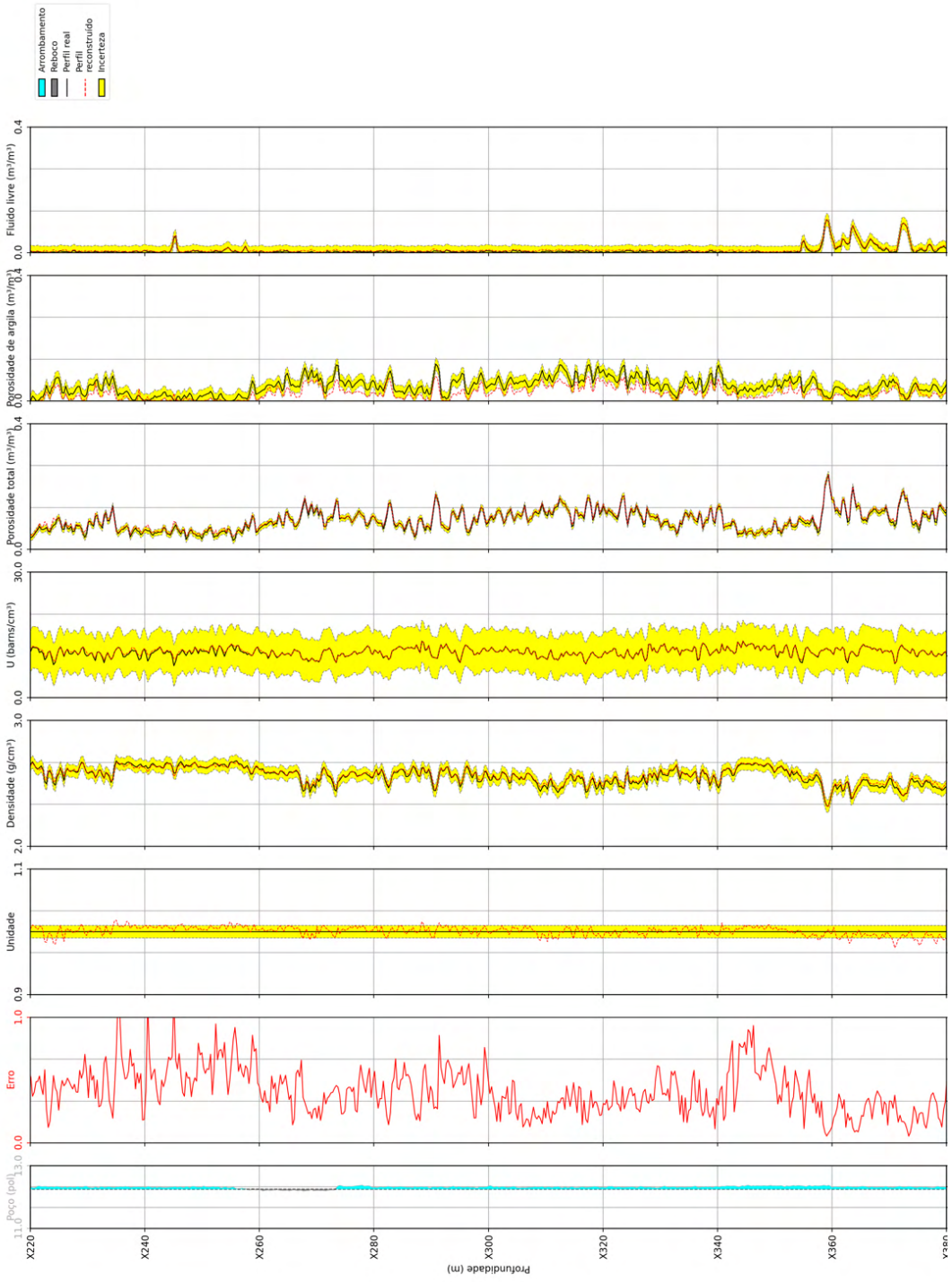


Figura 71 – Diâmetro de poço, erro e comparação entre os perfis unidade, densidade, U, porosidade total, água de argila e fluido livre reconstruídos e reais para o Poço F. A região amarela em volta dos perfis representa a incerteza.

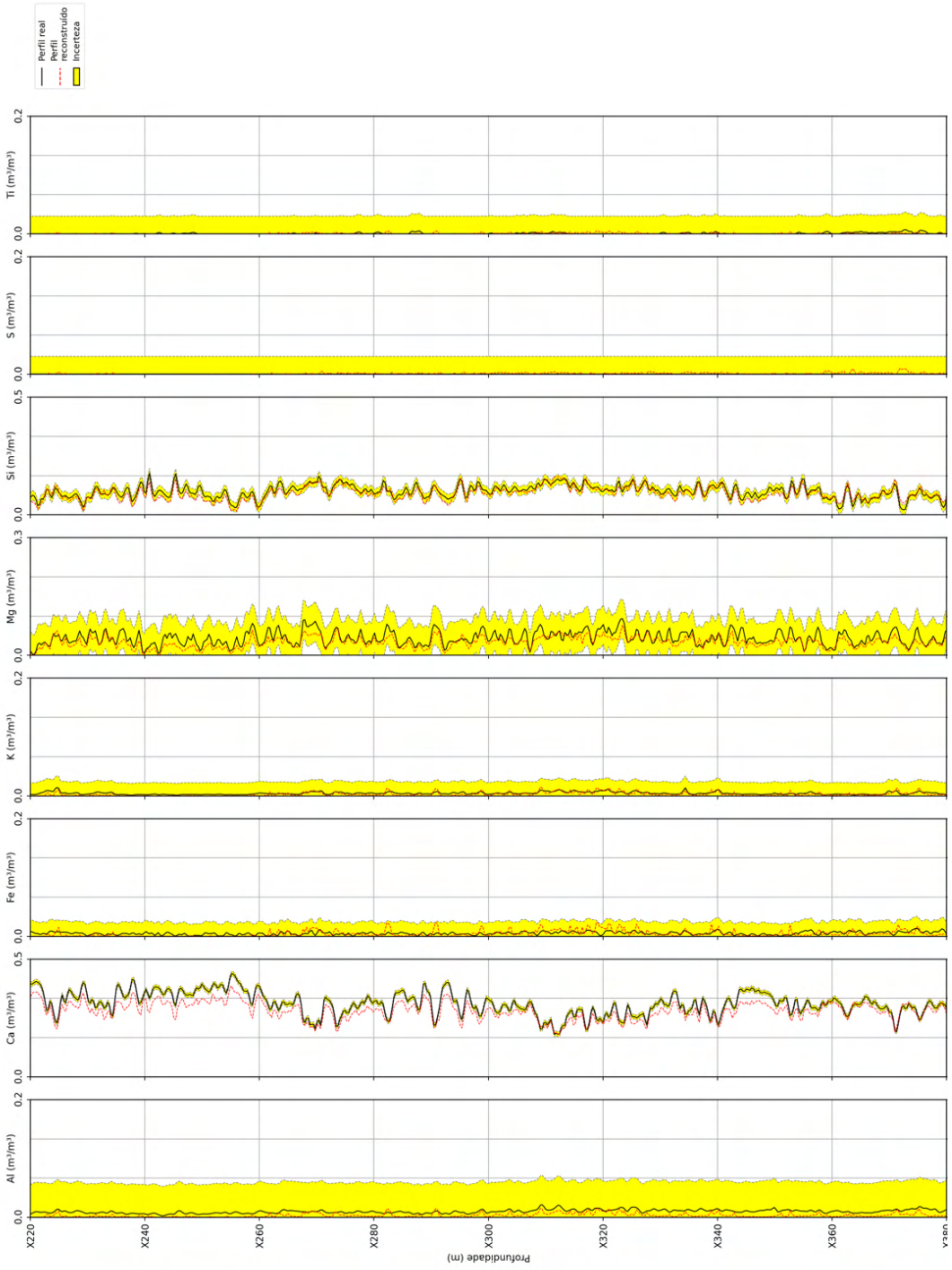


Figura 72 – Comparação entre os perfis geoquímicos reconstruídos e reais para o Poço F. A região amarela em volta dos perfis representa a incerteza.

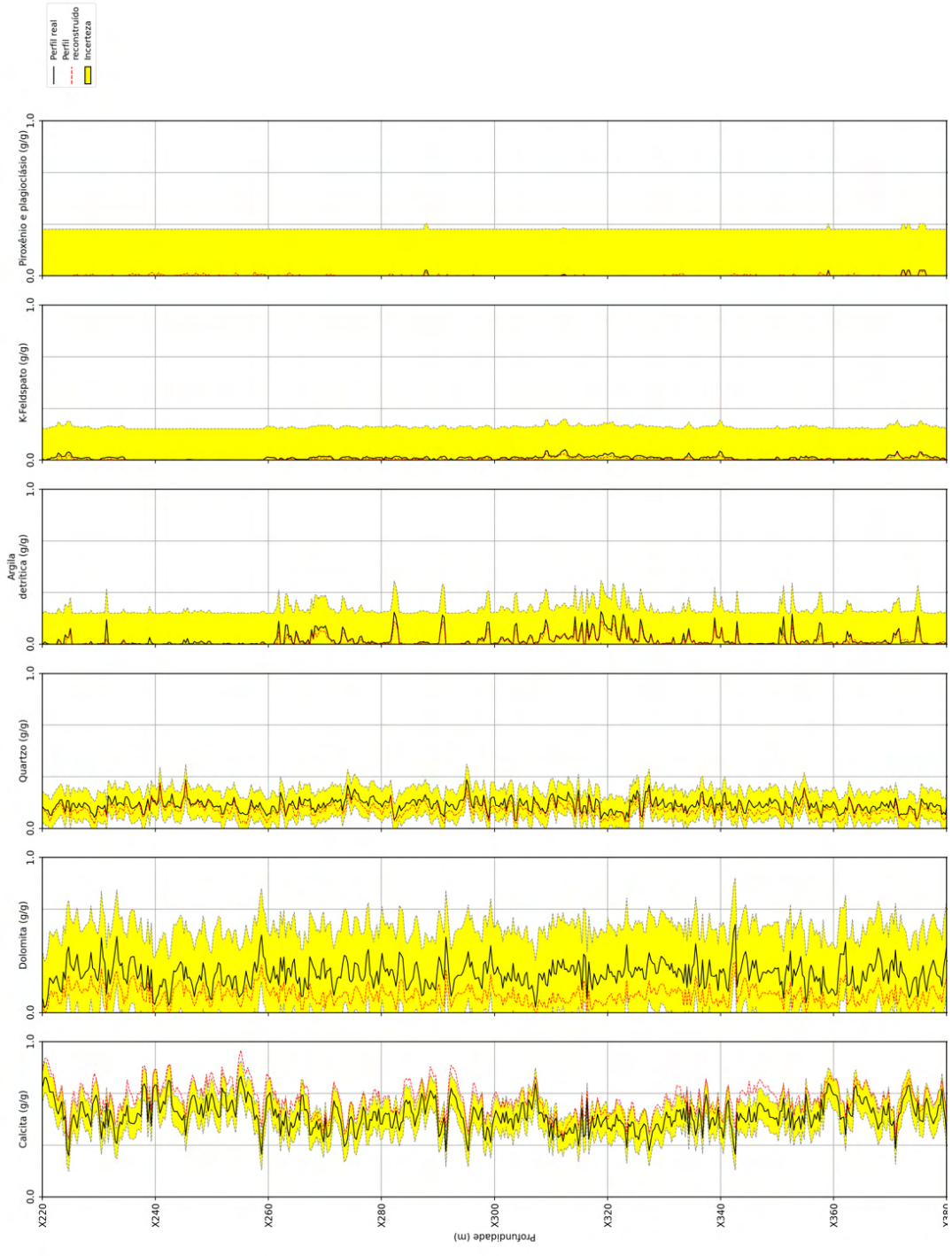


Figura 73 – Comparação entre as frações minerais estimadas pelo modelo de aprendizado de máquina reconstruídas e reais para o Poço F. A região amarela em volta dos perfis representa a incerteza.

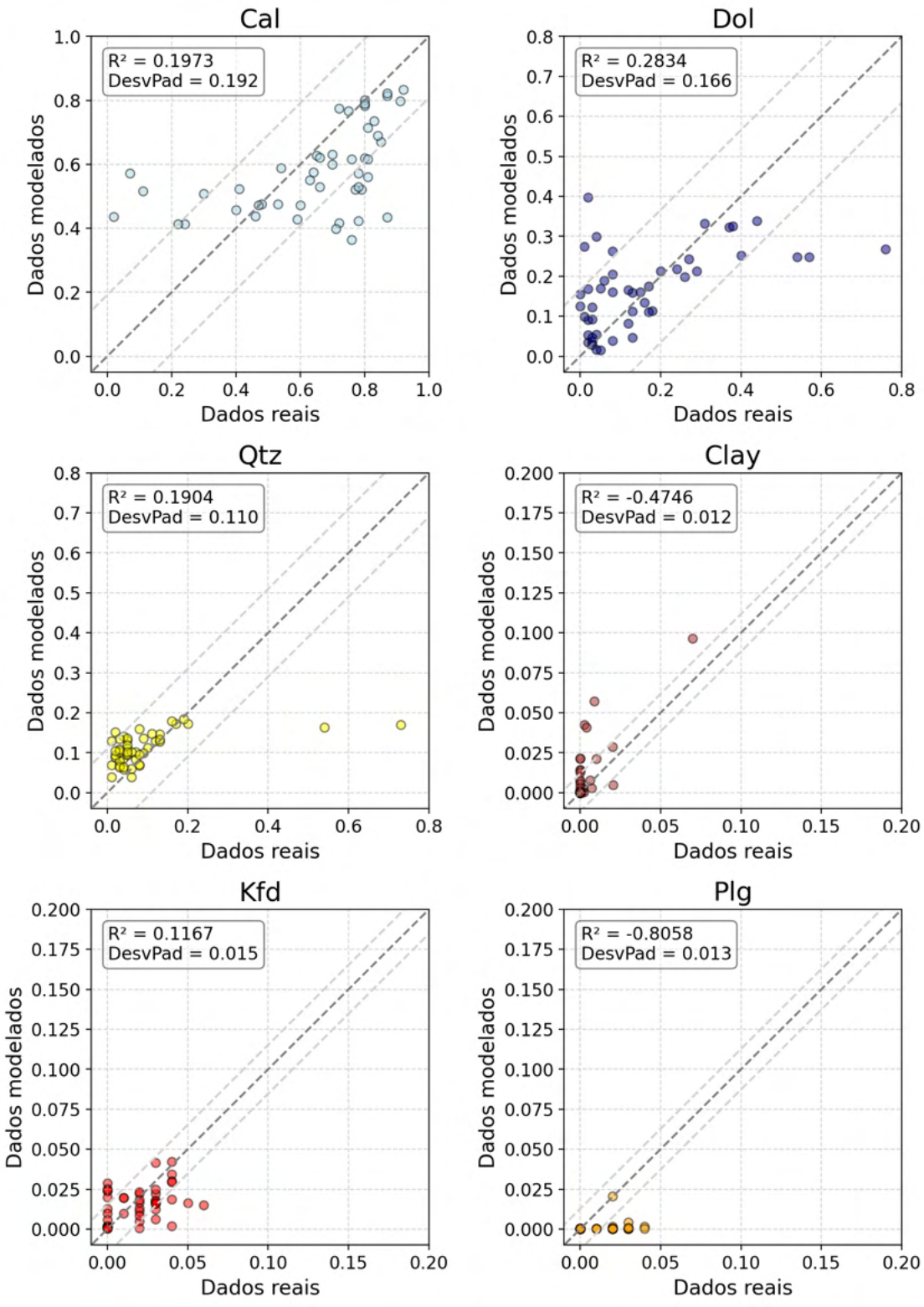


Figura 74 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica, K-feldspato e plagioclásio do poço D. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

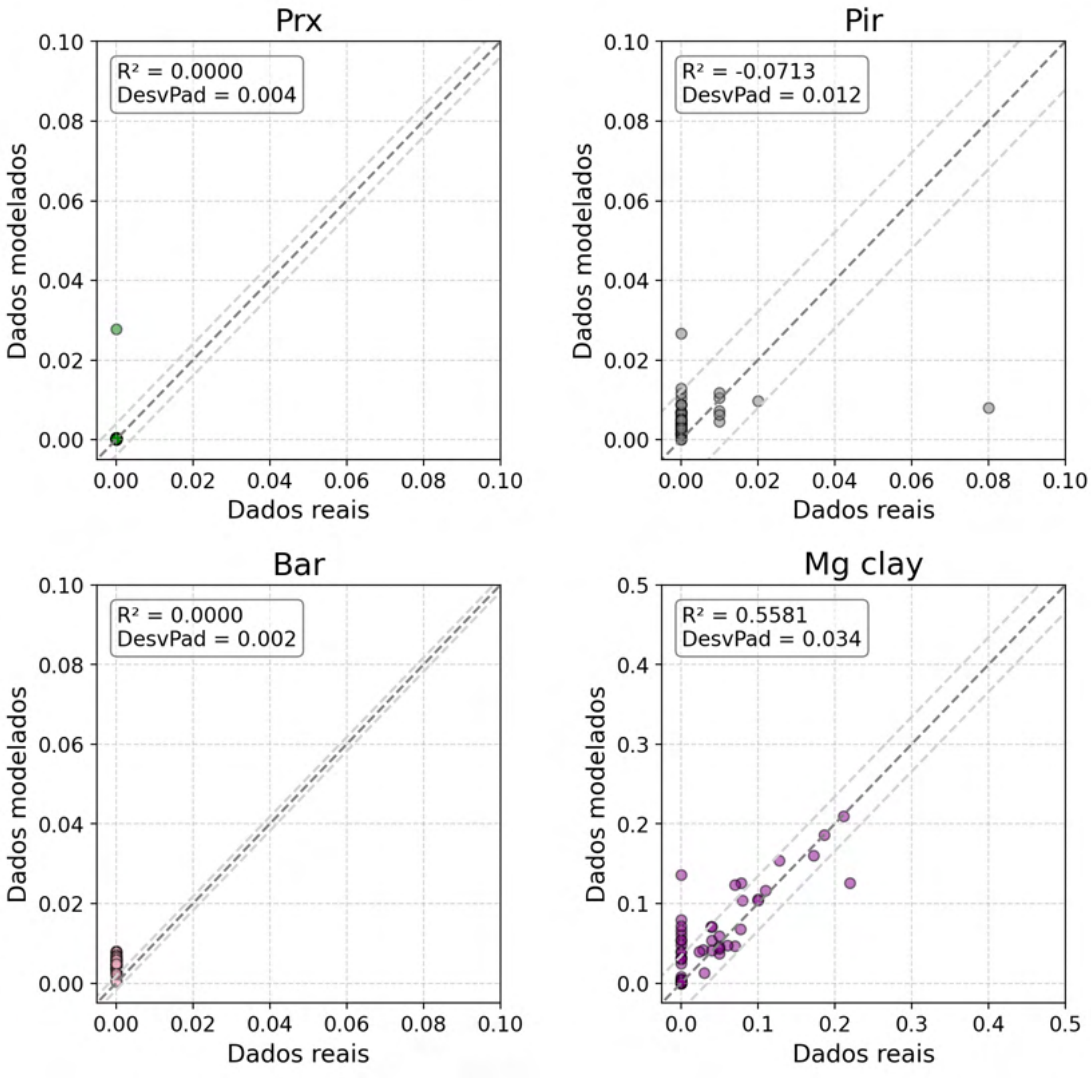


Figura 75 – Dados reais *versus* dados modelados para piroxênio, pirita, barita e argilas magnesianas do poço D. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

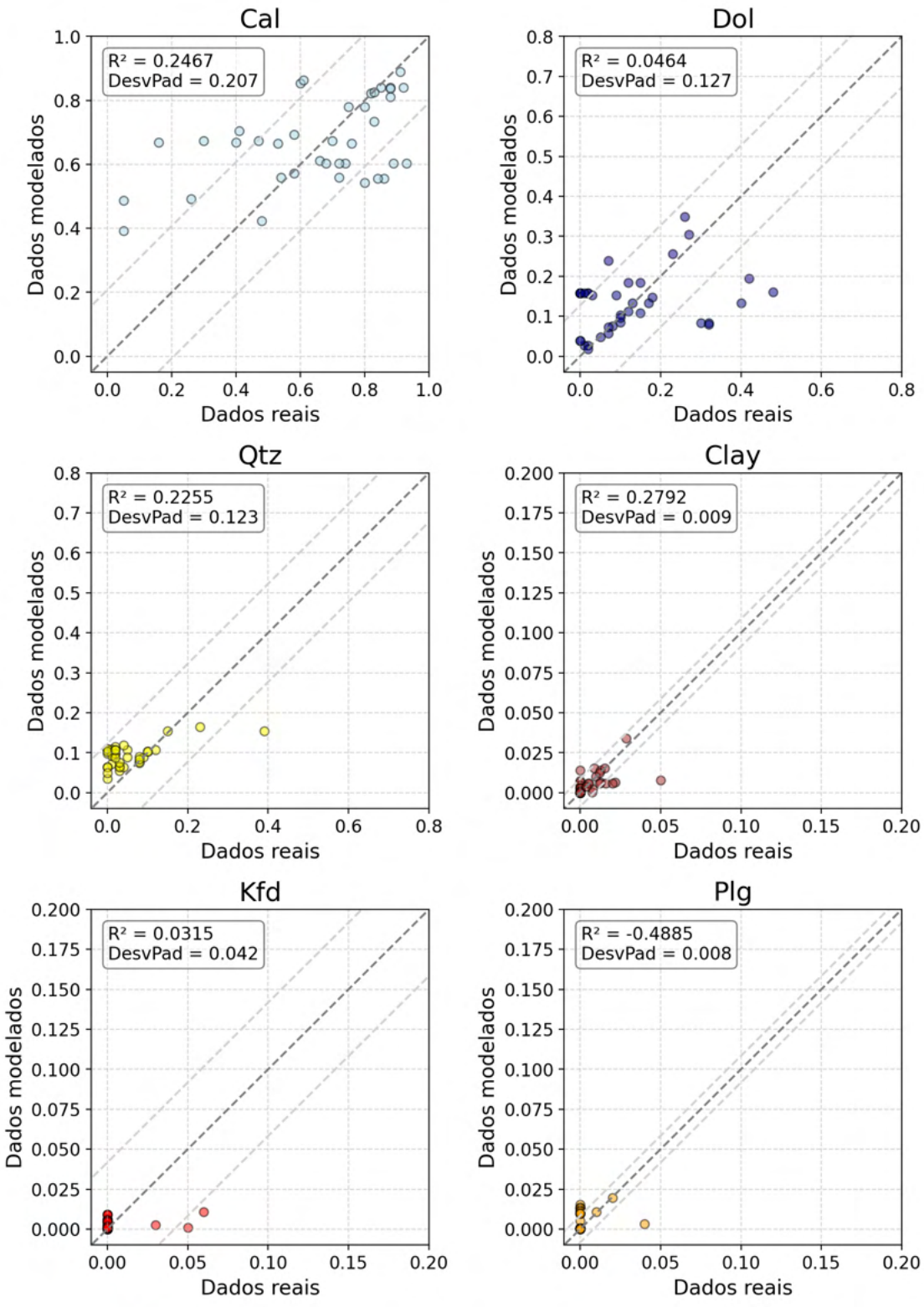


Figura 76 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica, K-feldspato e plagioclásio do poço E. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

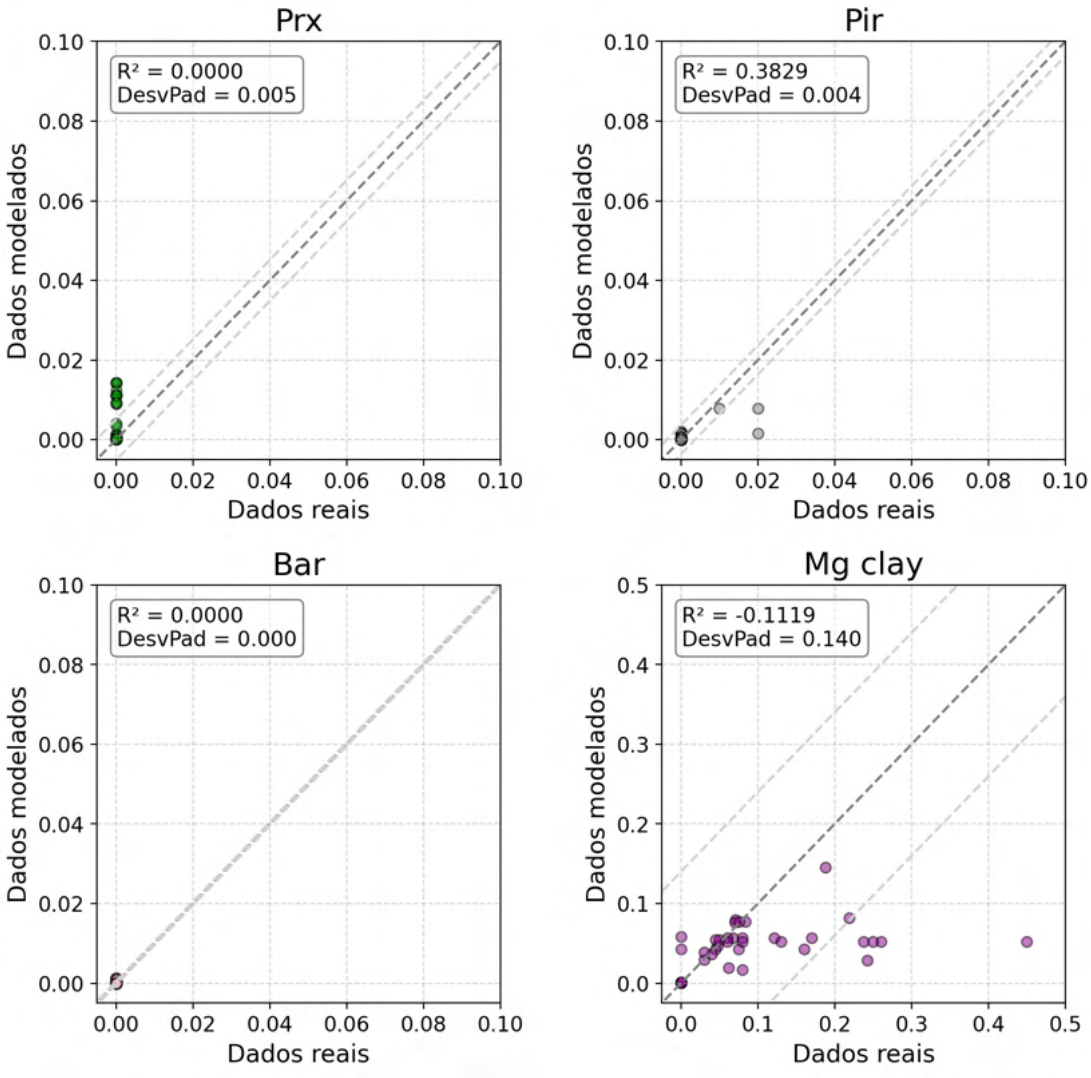


Figura 77 – Dados reais *versus* dados modelados para piroxênio, pirita, barita e argilas magnesianas do poço E. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

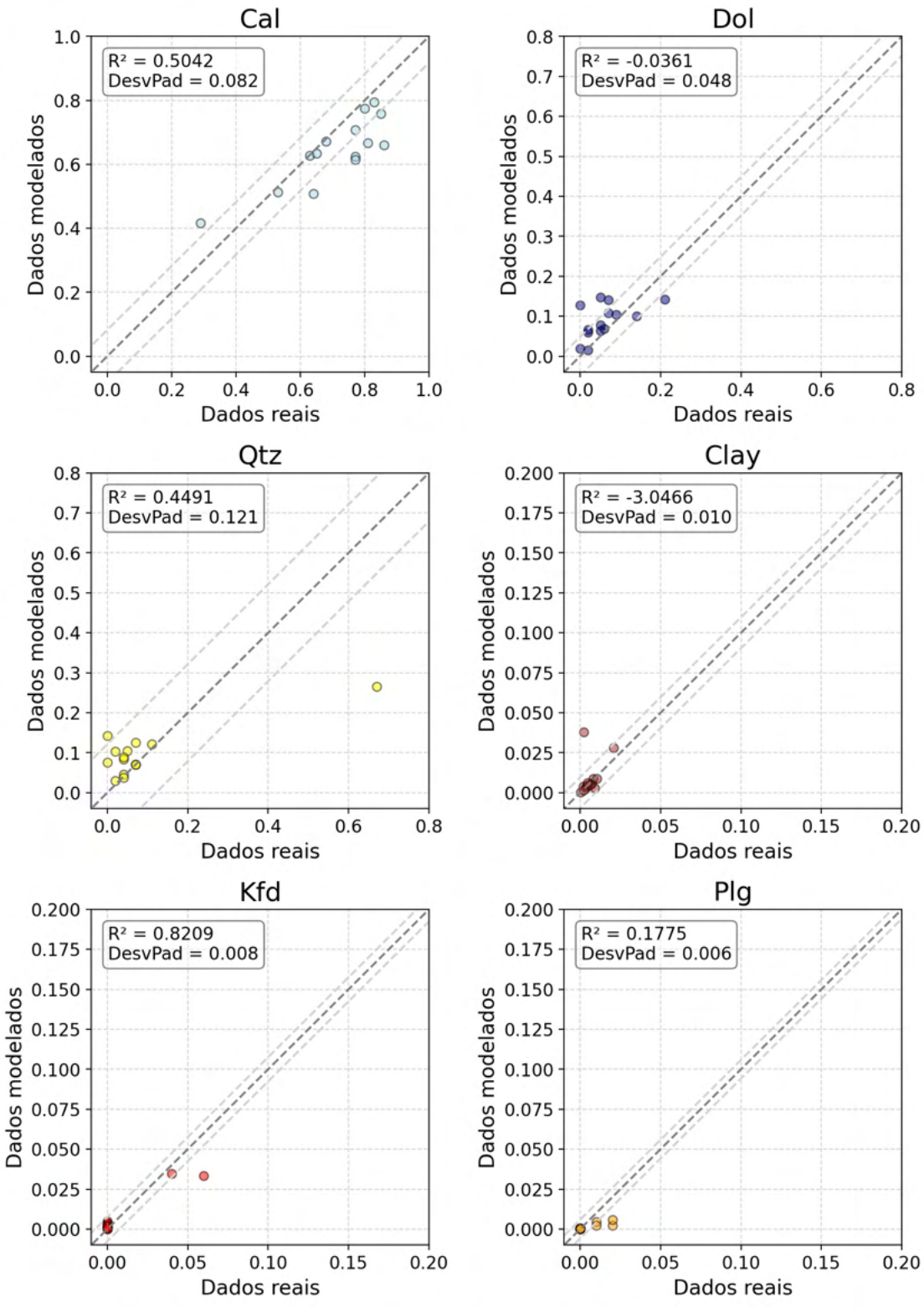


Figura 78 – Dados reais *versus* dados modelados para calcita, dolomita, quartzo, argila detrítica, K-feldspato e plagioclásio do poço F. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

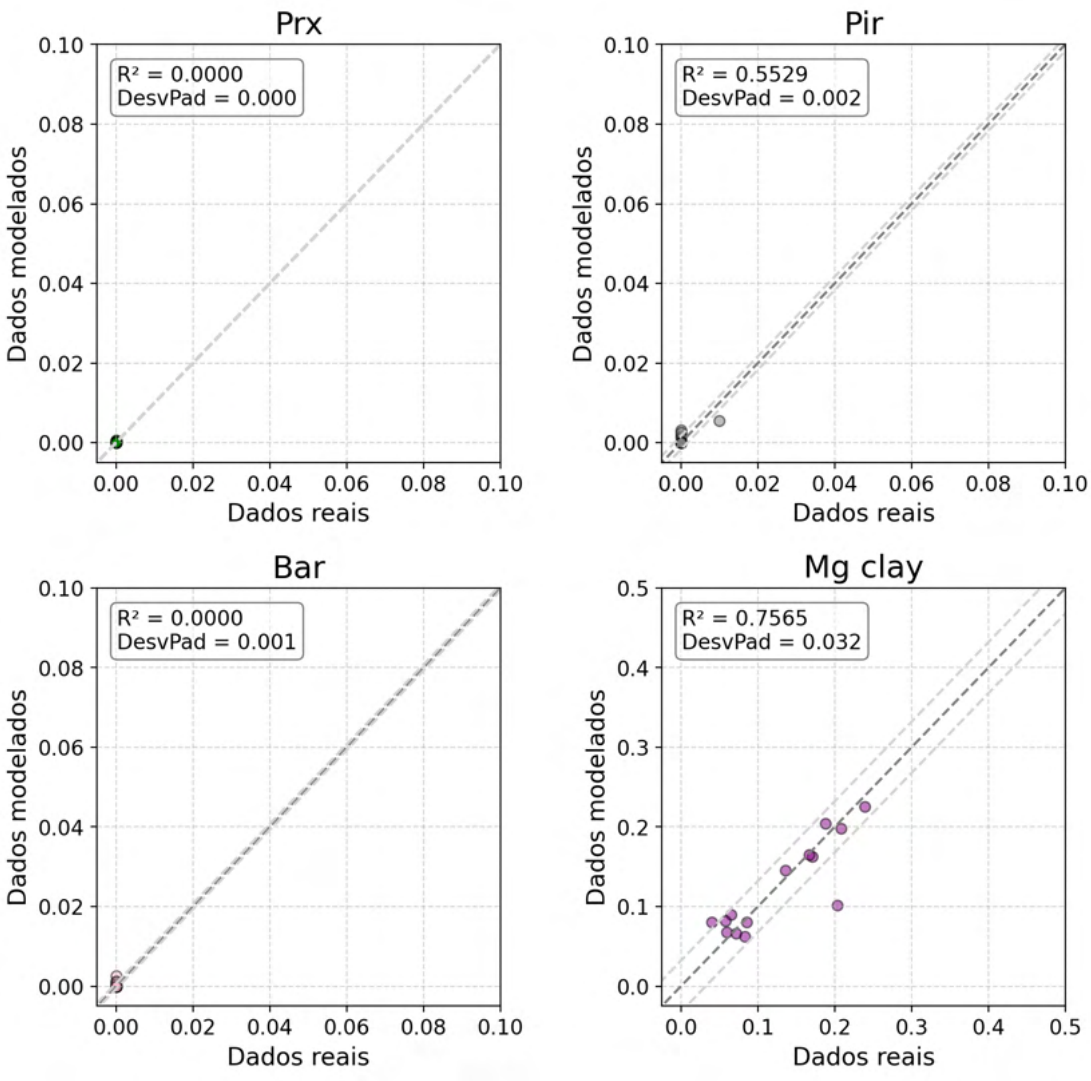


Figura 79 – Dados reais *versus* dados modelados para piroxênio, pirita, barita e argilas magnesianas do poço F. O baixo R^2 observado é reflexo da diferença de resolução vertical entre os perfis e amostras de rocha. Mesmo assim, é possível observar que a maioria dos pontos está próximo da reta 1:1. DesvPad: Desvio padrão.

5 Discussões

A modelagem geoquímica e mineralógica das rochas do pré-sal da Bacia de Santos em um cenário de otimização da aquisição de dados e redução de custos, realizada através de perfis de poços e inteligência artificial, demandou abordagens distintas. Essas abordagens estão relacionadas às características das bases de dados disponíveis para o treinamento dos algoritmos de aprendizado de máquina.

A modelagem geoquímica, realizada a partir de perfis geoquímicos sintéticos, caracterizou uma metodologia tradicional de aprendizado de máquina. Como os perfis de poço são informações baratas, ao menos quando comparados com amostras de rocha, a base de dados apresentou pouco enviesamento. A análise exploratória permitiu a identificação das características geoquímicas dos principais litotipos encontrados no pré-sal, corroborando o baixo viés da base de dados. Sendo assim, um fluxo básico de treinamento, validação e teste foi suficiente para treinar algoritmos de aprendizado de máquina capazes de gerar perfis geoquímicos sintéticos confiáveis.

Por serem informações caras, a base de dados de análises de FRX e DRX apresentou alto viés. Diferentemente dos perfis de poços, adquiridos de maneira contínua ao longo de toda a extensão do poço, as amostras de rocha são coletadas pontualmente e preferencialmente em intervalos reservatórios de boa permo-porosidade. Dessa forma, os minerais mais abundantes na base de dados são os constituintes das rochas carbonáticas do pré-sal.

O enviesamento da base de dados de amostras de rocha demandou a criação de uma nova metodologia de aplicação de aprendizado de máquina, denominada aprendizado escalonado. Ela se baseou no fato de que as frações dos minerais de uma rocha precisam somar 100%, fazendo com que as frações de um determinado mineral influenciem nas frações dos demais. Sendo assim, é possível utilizar as estimativas do modelo de um mineral como variável de entrada de outro, de forma escalonada. Como essa estratégia propaga incertezas, o escalonamento foi feito dos modelos com menor para os de maior incerteza; incerteza essa calculada a partir do desvio padrão do erro observado no conjunto de validação. O aprendizado escalonado foi capaz de melhorar significativamente a qualidade dos modelos mineralógicos obtidos por aprendizado de máquina.

Entretanto, o aprendizado escalonado ou quaisquer outras técnicas exclusivamente do universo de aprendizado de máquina não são capazes de lidar com a falta de informação da base de dados. Algoritmos treinados com a pequena quantidade de amostras onde foi observado pirita, barita e argilas magnesianas apresentariam alta incerteza, independentemente das estratégias utilizadas.

Sendo assim, uma segunda metodologia foi criada especificamente para esses cenários, denominada modelagem híbrida. Ela consistiu na combinação dos algoritmos de aprendizado de máquina treinados através do aprendizado escalonado com modelos minerais probabilísticos, que utiliza perfis de poços. Essa hibridização visou resolver os problemas relacionados ao enviesamento da base de dados através da inclusão das frações minerais estimadas pelo aprendizado escalonado em uma etapa probabilística baseada em princípios físicos. Essa metodologia foi capaz de quantificar as frações de minerais não incluídos no modelo de aprendizado de máquina, especialmente das argilas magnesianas.

A comparação entre as análises de DRX e as frações minerais estimadas pelo modelo híbrido e a comparação entre os perfis reais e reconstruídos aponta para algumas melhorias geradas pela inclusão das frações minerais da etapa de aprendizado de máquina, os perfis de RMN e o uso das incertezas no modelo híbrido. Essas melhorias são:

- O uso em conjunto da estimativa de argilas detríticas da etapa de aprendizado de máquina e da água de argila do perfil de RMN foi capaz de estimar corretamente as frações de argilas magnesianas. Os perfis de RMN tradicionalmente não são utilizados em modelos minerais, uma vez eles quantificam os fluidos presentes na formação.
- A inclusão das frações minerais da etapa de aprendizado de máquina funciona como uma regularização na etapa probabilística, direcionando-a a alcançar resultados equivalentes a essas estimativas. Como os algoritmos de aprendizado de máquina foram treinados na base de dados de análises de FRX e DRX do pré-sal da Bacia de Santos, essas informações são carregadas para dentro da etapa probabilística. Na prática, isso pode penalizar a reconstrução de alguns perfis, porém isso é compensado com informações provenientes da base de dados de análises laboratoriais incorporadas pelo aprendizado de máquina.
- Como a base de dados usada para treinar os algoritmos de aprendizado de máquina é enviesada, esses algoritmos podem gerar resultados incoerentes. A integração com os demais perfis e a inclusão de incertezas faz com que o modelo híbrido

tenha flexibilidade para lidar com todas essas informações, gerando resultados mais coerentes quando comparados com um modelo puramente probabilístico.

- A inclusão das frações minerais da etapa de aprendizado de máquina na forma de equações de reconstrução faz com que o sistema de equações da etapa probabilística se mantenha sobre-determinado para uma quantidade grande de minerais sem a inclusão de perfis de difícil interpretação.
- As frações minerais da etapa de aprendizado de máquina servem de estimativa inicial no processo de minimização da diferença entre os perfis reais e reconstruídos, importante em processos iterativos com diversas variáveis.

É importante ressaltar que os dois modelos mineralógicos propostos não competem entre si. Eles foram desenvolvidos para atender diferentes cenários. O modelo a partir de aprendizado escalonado pode ser aplicado em poços que perfuraram rochas cuja composição mineral é semelhante a da base de dados de treinamento, com rochas compostas majoritariamente por calcita, dolomita, quartzo, K-feldspato, argilas detríticas, plagioclásio e piroxênio, enquanto o modelo híbrido pode ser aplicado em poços com quantidades significativas de pirita, barita e argilas magnesianas.

Apesar das vantagens observadas na utilização dos algoritmos de *ensemble boosting* em lidar com dados de perfilagens de poços (em particular, o XGBoost), é importante entender que o surgimento de algoritmos de aprendizado de máquina é um processo dinâmico, e novos e mais robustos algoritmos podem surgir ao longo dos anos. Entretanto, isso não inviabiliza o presente trabalho, pois as metodologias definidas são o real resultado alcançado. No futuro, caso se observe que novos algoritmos desempenham melhor ao serem aplicados aos dados de perfilagens, basta substituí-los dentro dos fluxogramas das metodologias.

Um dos desafios das etapas de avaliação e teste dos modelos mineralógicos foi a comparação das frações estimadas com as análises de DRX das amostras de rocha. Os R^2 obtidos a partir dessa comparação são muito influenciados pela dispersão dos pontos, apresentando valores baixos; dispersão essa esperada por conta das diferentes resoluções verticais. Entretanto, a análise do gráfico de dados reais *versus* dados modelados demonstra que as estimativas estão no geral alinhadas à reta 1:1, indicando excelentes resultados. Dessa forma, se recomenda utilizar a análise conjunta do R^2 e dos gráficos para avaliar os resultados dos modelos. Essas questões não precisaram ser endereçadas na modelagem

geoquímica pois os perfis sintéticos são gerados a partir de perfis de poços e possuem resolução vertical semelhante a dos perfis geoquímicos reais.

6 Conclusões

Três metodologias baseadas em inteligência artificial e perfis de poços foram desenvolvidas visando viabilizar a modelagem geoquímica e mineralógica das rochas do pré-sal da Bacia de Santos em um cenário de otimização de aquisição de dados e redução de custos dos projetos de desenvolvimento de reservatórios de hidrocarbonetos.

Os modelos para a geração de perfis geoquímicos sintéticos apresentaram resultados condizentes e até superiores ao observado na literatura, com a ressalva de que nenhum trabalho encontrado aplicava aprendizado de máquina a perfis geoquímicos e não buscavam a substituição completa de uma ferramenta de perfilagem. Os modelos dos elementos Fe, Ti, Al e Ca apresentaram os melhores resultados, com R^2 acima de 0,90 na validação e validação cruzada. Os modelos de Si, S e Mg apresentaram R^2 acima de 0,80 e o modelo de Na, acima de 0,70.

A análise da importância das variáveis mostrou que os principais perfis que impactaram o treinamento dos modelos foram densidade, fator fotoelétrico, K, Th, U e nêutrons. Esse resultado sugere que perfis geoquímicos sintéticos ainda podem ser gerados em cenários de maior redução de custos de perfilagem.

A comparação entre perfis reais e modelados na fase de teste confirmou a qualidade dos perfis geoquímicos sintéticos. Eles foram capazes de reproduzir a composição química das principais rochas do pré-sal, como carbonatos, carbonatos silicificados e dolomitizados, rochas siliciclásticas e rochas ígneas. Uma clusterização aglomerativa agrupou os perfis reais e modelados de maneira similar, demonstrando que um intérprete chegaria às mesmas conclusões. Os perfis sintéticos também foram capazes de adicionar as concentrações de Mg e Na em um poço onde eles não haviam sido adquiridos, sem alterar de maneira significativa as concentrações dos demais elementos químicos.

Os resultados da geração de perfis geoquímicos sintéticos demonstraram que um modelo de aprendizado de máquina devidamente treinado e avaliado é capaz de substituir a aquisição de perfis geoquímicos com alta confiabilidade. Essa substituição permite a modelagem geoquímica das rochas do pré-sal usando apenas informações adquiridas em perfilagens reduzidas, alinhado com as políticas de redução de custo e otimização de operações de perfilagem.

O modelo mineralógico desenvolvido a partir do treinamento de algoritmos de aprendizado de máquina em análises de FRX e DRX permitiu estimar frações dos minerais mais abundantes observados na base de dados. O aprendizado escalonado proposto melhorou significativamente a qualidade dos modelos. Dentre as principais melhorias destacam-se o aumento do R^2 do modelo de carbonatos de 0,81 para 0,84, do modelo da calcita de 0,77 para 0,86, o aumento do R^2 do modelo da dolomita de 0,79 para 0,82, do modelo do quartzo de 0,66 para 0,87, e do modelo das argilas detríticas de 0,60 para 0,85. Como o aprendizado escalonado propaga erros, ele não foi aplicado aos modelos de K-feldspato e plagioclásio, já que o R^2 do modelo desses minerais apresentou pouca variação ao longo do escalonamento.

Além dos modelos minerais, o modelo de perda ao fogo apresentou alta qualidade, com R^2 de 0,96. Essa não é uma informação adquirida em perfilagem e, com esse modelo, pode facilmente ser adquirido através dos perfis geoquímicos.

A comparação entre os perfis gerados pelos modelos minerais e as análises de DRX de três poços perfurados no pré-sal confirmou a qualidade dos modelos. Eles foram capazes de identificar as variações nas frações de calcita, dolomita e quartzo nas rochas carbonáticas, de quartzo, argilas e K-feldspato em rochas siliciclásticas, e plagioclásio e piroxênio em rochas ígneas. Esses testes demonstraram que os modelos por aprendizado escalonado podem ser utilizados em poços que perfuraram rochas cuja composição majoritária apresenta minerais incluídos no treinamento desses modelos.

A quantificação de minerais importantes como pirita, barita e argilas magnesianas não pode ser feito através exclusivamente do aprendizado de máquina, uma vez que esses minerais não são devidamente contemplados na base de dados devido a limitações na aquisição de amostras de rocha. Sendo assim, um modelo mineralógico híbrido foi criado através da integração dos algoritmos do aprendizado escalonado com um modelo probabilístico. A etapa probabilística utilizou as informações dos algoritmos de aprendizado de máquina em conjunto com uma série de perfis de poço.

O modelo híbrido introduziu três principais inovações: a inclusão de informações de algoritmos de aprendizado de máquinas em um modelo probabilístico baseado em princípios físicos funcionando como um regularizador, o uso das informações dos perfis de RMN para estimar diretamente frações minerais, e a ponderação das estimativas dos algoritmos de aprendizado de máquina através de suas incertezas. Essa última inovação é particularmente importante pois, sem a hibridização, a incerteza de uma estimativa de

aprendizado de máquina serve apenas como um sinal de cautela para o usuário das informações preditas. Já sua inclusão na etapa probabilística faz com que a incerteza da etapa de aprendizado de máquina sirva ativamente na melhoria das estimativas do modelo híbrido.

O modelo híbrido estimou a mineralogia de três poços perfurados no pré-sal da Bacia de Santos. Os resultados foram equivalentes aos da mineralogia quantificada por DRX de amostras de rocha coletadas nesses poços, principalmente as de argilas magnesianas. Também se observou que o uso das incertezas deu liberdade para o modelo híbrido estimar frações minerais mais coerentes sem necessariamente reconstruir perfeitamente todos os perfis.

Apesar da robustez do aprendizado de máquina, os desafios impostos pelo enviesamento das bases de dados observados na caracterização de reservatórios exigem que soluções sejam propostas para além da melhoria ou criação de novos algoritmos ou da coleta indiscriminada de mais dados, o que geraria mais custos. O aprendizado escalonado ou os modelos híbridos, capazes de integrar o aprendizado de máquina com princípios físicos, podem ser a solução para lidar com esses desafios. Espera-se que os conceitos apresentados neste trabalho sejam usados para resolver outros desafios da caracterização de reservatórios, afastando o temor infundado relacionado à possível substituição do homem pela máquina e demonstrando que a atuação conjunta desses dois agentes é capaz de avançar o conhecimento humano.

Referências

AJAYI, O.; TORRES-VERDIN, C.; PREEG, W. E. Petrophysical interpretation of lwd, neutron-induced gamma-ray spectroscopy measurements: an inversion-based approach. *Petrophysics*, v. 56, p. 358–378, 2015. ISSN 1529-9074. Citado na página 49.

AKINNIKAWE, O.; LYNE, S.; ROBERTS, J. Synthetic well log generation using machine learning techniques. In: . [S.l.]: Unconventional Resources Technology Conference, 2018. p. 16. ISBN 2018287702. Citado 2 vezes nas páginas 72 e 73.

AKKURT, R.; CONROY, T. T.; PSAILA, D.; PAXTON, A.; LOW, J.; SPAANS, P. Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. In: . Society of Petrophysicists and Well-Log Analysts, 2018. p. 25. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85051657222&partnerID=40&md5=943a97cff3a91836bc328b998e0e1eda>. Citado 3 vezes nas páginas 71, 72 e 73.

AL-BULUSHI, N. I.; KING, P. R.; BLUNT, M. J.; KRAAIJVELD, M. Artificial neural networks workflow and its application in the petroleum industry. *Neural Computing and Applications*, v. 21, p. 409–421, 2012. ISSN 09410643. Citado na página 71.

ALIZADEH, B.; NAJJARI, S.; KADKHODAIE-ILKHCHI, A. Artificial neural network modeling and cluster analysis for organic facies and burial history estimation using well log data: a case study of the south pars gas field, persian gulf, iran. *Computers and Geosciences*, v. 45, p. 261–269, 2012. Citado na página 71.

AMINIAN, K.; AMERI, S.; OYEROKUN, A.; THOMAS, B. Prediction of flow units and permeability using artificial neural networks. In: . [S.l.]: Society of Petroleum Engineers, 2003. p. 7. Citado na página 71.

ANDERSON, R. N.; DOVE, R. E.; BOGLIA, C.; SILVER, L. T.; JAMES, E. W.; CHAPPELL, B. W. Elemental and mineralogical analyses using geochemical logs from the cajon pass scientific drillhole, california, and their preliminary comparison with core analyses. *Geophysical Research Letters*, v. 15, p. 969–972, 1988. Citado na página 49.

ANIFOWOSE, F. A.; LABADIN, J.; ABDULRAHEEM, A. Hybrid intelligent systems in petroleum reservoir characterization and modeling: the journey so far and the challenges ahead. *Journal of Petroleum Exploration and Production Technology*, Springer Berlin Heidelberg, v. 7, p. 251–263, 2017. ISSN 21900566. Citado na página 73.

ARDABILI, S.; MOSAVI, A.; VáRKONYI-KóCZY, A. R. *Advances in machine learning modeling reviewing hybrid and ensemble methods*. [S.l.], 2019. 10 p. Citado na página 72.

ASFAHANI, J.; AHMAD, Z.; GHANI, B. A. Self organizing map neural networks approach for lithologic interpretation of nuclear and electrical well logs in basaltic environment, southern syria. *Applied Radiation and Isotopes*, Elsevier Ltd, v. 137, p. 50–55, 2018. ISSN 18729800. Disponível em: <https://doi.org/10.1016/j.apradiso.2018.03.008>. Citado na página 71.

BAGERI, B. S.; ADEBAYO, A. R.; BARRI, A.; JABERI, J. A.; PATIL, S.; HUSSAINI, S. R.; BABU, R. S. Evaluation of secondary formation damage caused by the interaction of chelated barite with formation rocks during filter cake removal. *Journal of Petroleum Science and*

Engineering, Elsevier B.V., v. 183, p. 10, 2019. ISSN 09204105. Disponível em: <https://doi.org/10.1016/j.petrol.2019.106395>. Citado na página 90.

BAHRPEYMA, F.; GOLCHIN, B.; CRANGANU, C. Fast fuzzy modeling method to estimate missing logs in hydrocarbon reservoirs. *Journal of Petroleum Science and Engineering*, Elsevier, v. 112, p. 310–321, 2013. ISSN 09204105. Disponível em: <http://dx.doi.org/10.1016/j.petrol.2013.11.019>. Citado 3 vezes nas páginas 71, 72 e 73.

BELOZEROV, B.; BUKHANOV, N.; EGOROV, D.; ZAKIROV, A.; OSMONALIEVA, O. Automatic well log analysis across priobskoe field using machine learning methods. In: . [S.l.]: Society of Petroleum Engineers, 2018. p. 21. Citado na página 71.

BELTRAO, R. L.; SOMBRA, C.; LAGE, A.; NETTO, J. F.; HENRIQUES, C. Challenges and new technologies for the development of the pre-salt cluster, santos basin, brazil. In: . [S.l.]: Offshore Technology Conference, 2009. p. 11. Citado 2 vezes nas páginas 34 e 35.

BESTAGINI, P.; LIPARI, V.; TUBARO, S. A machine learning approach to facies classification using well logs. In: . [S.l.]: Society of Exploration Geophysicists, 2017. p. 5. Citado na página 71.

BISHOP, C. M. *Pattern recognition and machine learning*. First. [S.l.]: Springer, 2006. 738 p. ISSN 1098-6596. ISBN 978-0387-31073-2. Citado na página 63.

BRANCH, M. A.; COLEMAN, T. F.; LI, Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing*, v. 21, p. 1–23, 1999. Citado na página 100.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001. Citado na página 70.

CARR, H. Y.; PURCELL, E. M. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Physical Review*, v. 94, n. 3, p. 630–638, 1954. ISSN 0031899X. Citado na página 53.

CHEN, D.; QUIREIN, J.; JR, H. S.; HAMID, S.; GRABLE, J. Neural network ensemble selection using a multi-objective genetic algorithm in processing pulsed neutron data. *Petrophysics*, v. 46, p. 323–334, 2005. Citado 2 vezes nas páginas 72 e 73.

CHEN, T.; GUESTRIN, C. Xgboost: a scalable tree boosting system. In: . ACM, 2016. p. 785–794. Disponível em: <http://arxiv.org/abs/1603.02754%0Ahttp://dx.doi.org/10.1145/2939672.2939785>. Citado 2 vezes nas páginas 69 e 72.

CHEN, Y.; SAYGIN, E. Seismic inversion by hybrid machine learning. *arXiv*, p. 36, 2020. ISSN 23318422. Citado 2 vezes nas páginas 73 e 74.

COATES, G. R.; XIAO, L.; PRAMMER, M. G. *NMR logging - principles and applications*. [S.l.]: Halliburton Energy Services, 1999. 253 p. Citado 3 vezes nas páginas 53, 54 e 55.

COLLETT, T. S.; LEWIS, R. E.; WINTERS, W. J.; LEE, M. W.; ROSE, K. K.; BOSWELL, R. M. Downhole well log and core montages from the mount elbert gas hydrate stratigraphic test well, alaska north slope. *Marine and Petroleum Geology*, Elsevier Ltd, v. 28, p. 561–577, 2011. ISSN 02648172. Disponível em: <http://dx.doi.org/10.1016/j.marpetgeo.2010.03.016>. Citado na página 60.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, p. 273–297, 1995. ISSN 22502114. Citado na página 69.

CULLITY, B.; STOCK, S. *Elements of X-ray diffraction*. 3. ed. [S.l.]: Prentice-Hall, 2001. 678 p. Citado na página 57.

DEAN, W. E. J. Determination of carbonate and organic matter in calcareous sediments and sedimentary rocks by loss on ignition: comparison with other methods. *Journal of Sedimentary Petrology*, v. 44, p. 242–248, 1974. Citado na página 104.

DRUCKER, H.; BURGES, C. J.; KAUFMAN, L.; SMOLA, A.; VAPNIK, V. Support vector regression machines. *Advances in Neural Information Processing Systems*, v. 1, p. 155–161, 1997. ISSN 10495258. Citado na página 69.

EL-BAGOURY, M. Integrated petrophysical study to validate water saturation from well logs in bahariya shaley sand reservoirs, case study from abu gharadig basin, egypt. *Journal of Petroleum Exploration and Production Technology*, Springer International Publishing, v. 10, p. 3139–3155, 2020. ISSN 21900566. Disponível em: <https://doi.org/10.1007/s13202-020-00969-3>. Citado na página 60.

ELLIS, D. V.; SINGER, J. M. *Well logging for earth scientists*. Second. [S.l.]: Springer, 2007. 692 p. ISBN 9781402037382. Citado 17 vezes nas páginas 30, 35, 37, 39, 40, 41, 42, 43, 44, 46, 47, 48, 50, 51, 52, 53 e 54.

FLAUM, C.; PIRIE, G. Determination of lithology from induced gamma ray spectroscopy. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 1981. p. 16. Citado na página 49.

FRANQUET, J. A.; BRATOVICH, M. W.; GLASS, R. D. State-of-the-art openhole shale gas logging. In: . [S.l.]: Society of Petroleum Engineers, 2012. p. 663–674. ISBN 9781632667113. Citado na página 61.

FREEDMAN, R.; HERRON, S.; ANAND, V.; HERRON, M.; MAY, D.; ROSE, D. New method for determining mineralogy and matrix properties from elemental chemistry measured by gamma ray spectroscopy logging tools. *SPE Reservoir Evaluation and Engineering*, v. 18, p. 599–608, 2015. ISSN 1094-6470. Citado 3 vezes nas páginas 30, 49 e 58.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, p. 119–139, 1997. ISSN 00344885. Citado na página 70.

FREUND, Y.; SCHAPIRE, R. E. Large margin classification using the perceptron algorithm. *Machine Learning*, v. 37, p. 277–296, 1999. ISSN 08856125. Citado na página 70.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. In: . [S.l.: s.n.], 2001. p. 39. Citado na página 68.

GALFORD, J. E.; QUIREIN, J. A.; SHANNON, S.; TRUAX, J. A.; WITKOWSKY, J. Field test results of a new neutron induced gamma ray spectroscopy geochemical logging tool. In: . [S.l.]: Society of Petroleum Engineers, 2009. p. 22. Citado na página 49.

GOMES, J. P.; BUNEVICH, R. B.; TEDESCHI, L. R.; TUCKER, M. E.; WHITAKER, F. F. Facies classification and patterns of lacustrine carbonate deposition of the barra velha formation, santos basin, brazilian pre-salt. *Marine and Petroleum Geology*, Elsevier, v. 113, p. 21,

2020. ISSN 02648172. Disponível em: <https://doi.org/10.1016/j.marpetgeo.2019.104176>. Citado 2 vezes nas páginas 31 e 35.

GONZALEZ, J.; LEWIS, R.; HEMINGWAY, J.; GRAU, J.; RYLANDER, E.; SCHMITT, R. Determination of formation organic carbon content using a new neutron-induced gamma ray spectroscopy service that directly measures carbon. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 2013. p. 15. Citado na página 49.

GUARIDO, M. *Machine learning in geoscience: facies classification with features engineering, clustering, and gradient boosting trees*. [S.l.], 2018. v. 30, 1-24 p. Citado na página 71.

HAMADA, G. M.; AHMED, E.; Y, N. C. Artificial neural network (ann) prediction of porosity and water saturation of shaly sandstone reservoirs. *Advances in Applied Science Research*, v. 9, p. 26–31, 2018. Citado na página 71.

HAMADA, G. M.; AL-AWAD, M. N. Petrophysical evaluation of low resistivity sandstone reservoirs. *Journal of Canadian Petroleum Technology*, v. 39, p. 7–14, 2000. ISSN 00219487. Citado na página 90.

HAMADA, G. M.; AL-AWAD, M. N.; ALMALIK, M. S. Log evaluation of low-resistivity sandstone reservoirs. In: . [S.l.]: Society of Petroleum Engineers, 2001. p. 10. ISBN 9781555639280. Citado na página 90.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The elements of statistical learning: data mining, inference and prediction*. Second. [S.l.]: Springer, 2009. 745 p. ISBN 9780387848570. Citado na página 68.

HEIRI, O.; LOTTER, A. F.; LEMCKE, G. Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results. *Journal of Paleolimnology*, v. 25, p. 101–110, 2001. Citado na página 104.

HERLINGER, R.; FREITAS, G. do N.; ANJOS, C. D. W. D.; ROS, L. F. D. Petrological and petrophysical implications of magnesian clays in brazilian pre-salt deposits. In: *SPWLA 61st Annual Logging Symposium*. [S.l.]: Society of Petrophysicists and Well-Log Analysts, 2020. p. 13. Citado 6 vezes nas páginas 20, 31, 35, 56, 61 e 102.

HERRON, M. M. Mineralogy from geochemical well logging. *Clays and Clay Minerals*, v. 34, p. 204–213, 1986. ISSN 0009-8604. Citado na página 49.

HERRON, S.; HERRON, M.; PIRIE, I.; SALDUNGARAY, P.; CRADDOCK, P.; CHARSKY, A.; POLYAKOV, M.; SHRAY, F.; LI, T. Application and quality control of core data for the development and validation of elemental spectroscopy log interpretation. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 2014. p. 23. ISSN 1529-9074. Citado na página 30.

HERRON, S. L.; HERRON, M. M. Quantitative lithology: An application for open and cased hole spectroscopy. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 1996. p. 14. Citado 2 vezes nas páginas 30 e 61.

HERTZOG, R.; COLSON, L.; SEEMAN, B.; O'BRIEN, M.; SCOTT, H.; MCKEON, D.; WRAIGHT, P.; GRAU, J.; ELLIS, D.; SCHWEITZER, J.; HERRON, M. Geochemical logging with spectrometry tools. *SPE Formation Evaluation*, v. 4, p. 153–162, 1989. ISSN 0885-923X. Citado 2 vezes nas páginas 48 e 49.

HO, T. K. Random decision forests. In: . [S.l.: s.n.], 1995. p. 278–282. ISBN 0818671289. ISSN 15205363. Citado 2 vezes nas páginas 66 e 70.

HOEINK, T.; ZAMBRANO, C. Shale discrimination with machine learning methods. In: . American Rock Mechanics Association, 2017. p. 6. Disponível em: <https://www.onepetro.org/conference-paper/ARMA-2017-0769>. Citado na página 71.

IBRAHIM, A. F.; AL-MUJALHEM, M. Q.; NASR-EL-DIN, H. A.; AL-BAGOURY, M. Evaluation of formation damage of oil-based drilling fluids weighted with micronized ilmenite or micronized barite. *SPE Drilling and Completion*, v. 35, p. 402–413, 2020. ISSN 10646671. Citado na página 90.

IBRAHIM, D. S.; SAMI, N. A.; BALASUBRAMANIAN, N. Effect of barite and gas oil drilling fluid additives on the reservoir rock characteristics. *Journal of Petroleum Exploration and Production Technology*, Springer Berlin Heidelberg, v. 7, p. 281–292, 2017. ISSN 21900566. Citado na página 90.

JENKINS, R.; SNYDER, R. L. *Introduction to X-ray powder diffractometry*. [S.l.]: John Wiley and Sons, Inc, 1996. 432 p. ISBN 978-0-471-51339-1. Citado na página 58.

JIN, X.; ZHU, D.; HILL, A. D.; MCDUFF, D. Effects of heterogeneity in mineralogy distribution on acid fracturing efficiency. In: . [S.l.]: Society of Petroleum Engineers, 2019. p. 147–160. ISBN 9781613996294. ISSN 1930-1855. Citado 2 vezes nas páginas 30 e 49.

JR, W. A. G.; QUIREIN, J. A.; BOUTEMY, Y. L.; TABANOU, J. R. Application of gamma ray spectroscopy to formation evaluation. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 1982. p. 28. Citado na página 49.

JUNIOR, G. S. *Caracterização de argilominerais expansíveis em rochas reservatório por relaxometria*. 91 p. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2019. Citado na página 96.

KORJANI, M.; POPA, A.; GRIJALVA, E.; CASSIDY, S.; ERSHAGHI, I. A new approach to reservoir characterization using deep learning neural networks. In: . Society of Petroleum Engineers, 2016. p. 15. ISBN 978-1-61399-465-8. Disponível em: <http://www.onepetro.org/doi/10.2118/180359-MS>. Citado 2 vezes nas páginas 72 e 73.

KUHN, M.; JOHNSON, K. *Applied predictive modeling*. Springer, 2013. 600 p. ISBN 978-1-4614-6848-6. Disponível em: <http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/ref=pd_bxgy_b_img_z>. Citado na página 68.

MACDONALD, R.; HARDMAN, D.; SPRAGUE, R.; MERIDJI, Y.; MUDJIONO, W.; GALFORD, J.; ROURKE, M.; DIX, M.; KELTON, M. Using elemental geochemisty to improve sandstone reservoir characterization : a case study from the unayzah a interval of saudi arabia. In: . Society of Petrophysicists and Well-Log Analysts, 2010. v. 52, p. 16. ISSN 15299074. Disponível em: <https://www.onepetro.org/journal-paper/SPWLA-2011-v52n5a2>. Citado na página 49.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 1943. ISSN 15334406. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/20853017>. Citado na página 69.

MEIBOOM, S.; GILL, D. Modified spin-echo method for measuring nuclear relaxation times. *Review of Scientific Instruments*, v. 29, n. 8, p. 688–691, 1958. ISSN 00346748. Citado na página 53.

MINERALOGY-DATABASE. 2021. <http://webmineral.com>. Acessado: 2021-07-06. Citado 3 vezes nas páginas 94, 95 e 97.

MITCHELL, W. K.; NELSON, R. J. A practical approach to statistical log analysis. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 1988. p. 1–20. Citado 2 vezes nas páginas 30 e 58.

MOREIRA, J. L. P.; VALDETARO, C.; GIL, J. A.; MACHADO, M. A. P. Bacia de santos. *Boletim de Geociencias da Petrobras*, v. 15, p. 531–549, 2007. Citado 3 vezes nas páginas 31, 34 e 36.

NASHAWI, I. S.; MALALLAH, A. Improved electrofacies characterization and permeability predictions in sandstone reservoirs using a data mining and expert system approach. *Petrophysics*, v. 50, p. 250–268, 2009. ISSN 15299074. Citado na página 71.

NEGARA, A.; JIN, G.; AGRAWAL, G. Enhancing rock property prediction from conventional well logs using machine learning technique - case studies of conventional and unconventional reservoirs. In: . [S.l.]: Society of Petroleum Engineers, 2016. p. 13. Citado na página 71.

NORTH, R. J. Through-casing reservoir evaluation using gamma ray spectroscopy. In: . [S.l.]: Society of Petroleum Engineers, 1987. p. 329–342. Citado 2 vezes nas páginas 30 e 50.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY Édouard. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. ISSN 1467-9280. Disponível em: <https://scikit-learn.org/stable/index.html#>. Citado na página 82.

PEMPER, R. A history of nuclear spectroscopy in well logging. *Petrophysics*, v. 61, p. 523–548, 2020. Citado 2 vezes nas páginas 58 e 60.

PEMPER, R.; SOMMER, A.; GUO, P.; JACOBI, D.; LONGO, J.; BLIVEN, S.; RODRIGUEZ, E.; MENDEZ, F.; HAN, X. A new pulsed neutron sonde for derivation of formation lithology and mineralogy. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 2006. p. 13. Citado na página 50.

PRATAMA, E.; ISMAIL, M. S.; RIDHA, S. An integrated workflow to characterize and evaluate low resistivity pay and its phenomenon in a sandstone reservoir. *Journal of Geophysics and Engineering*, IOP Publishing, v. 14, p. 513–519, 2017. ISSN 17422140. Disponível em: <http://dx.doi.org/10.1088/1742-2140/aa5efb>. Citado na página 90.

QUIREIN, J.; VIGNE, J. L.; CHAPMAN, S. Enhancements to the pulsed neutron gamma ray spectroscopy interpretation. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 1987. p. 23. Citado na página 49.

QUIREIN, J.; WITKOWSKY, J.; TRUAX, J.; GALFORD, J.; SPAIN, D.; ODUMOSU, T. Integrating core data and wireline geochemical data for formation evaluation and characterization of shale gas reservoir. In: . [S.l.]: Society of Petroleum Engineers, 2010. p. 18. ISSN 01492136. Citado na página 61.

RADTKE, R.; LORENTE, M.; ADOLPH, B.; BERHEIDE, M.; FRICKE, S.; GRAU, J.; HERRON, S.; HORKOWITZ, J.; JORION, B.; MADIO, D.; MAY, D.; MILES, J.; PERKINS, L.; PHILIP, O.; ROSCOE, B.; ROSE, D.; STOLLER, C. A new capture and inelastic spectroscopy tool takes geochemical logging to the next level. In: . Society of Petrophysicists and Well-Log Analysts, 2012. p. 16. Disponível em: <http://69.18.148.110/~/media/Files/technical_papers/misc/spwla/2012_spwla_spectroscopy_tool.pdf>. Citado na página 50.

RAYMER, L. L.; HUNT, E. R.; GARDNER, J. S. An improved sonic transit time-to-porosity transform. In: *SPWLA 21st Annual Logging Symposium*. Lafayette, Louisiana: Society of Petrophysicists and Well-Log Analysts, 1980. p. 13. Citado na página 51.

ROLON, L.; MOHAGHEGH, S. D.; AMERI, S.; GASKARI, R.; MCDANIEL, B. Using artificial neural networks to generate synthetic well logs. *Journal of Natural Gas Science and Engineering*, Elsevier B.V, v. 1, p. 118–133, 2009. ISSN 18755100. Disponível em: <http://dx.doi.org/10.1016/j.jngse.2009.08.003>. Citado 3 vezes nas páginas 71, 72 e 73.

ROSENBLATT, F. The perceptron - a perceiving and recognizing automaton. *Report 85, Cornell Aeronautical Laboratory*, p. 29, 1957. Citado na página 70.

SALEHI, M. M.; RAHMATI, M.; KARIMNEZHAD, M.; OMIDVAR, P. Estimation of the non records logs from existing logs using artificial neural networks. *Egyptian Journal of Petroleum*, Egyptian Petroleum Research Institute, v. 26, p. 957–968, 2017. ISSN 20902468. Disponível em: <https://doi.org/10.1016/j.ejpe.2016.11.002>. Citado 2 vezes nas páginas 72 e 73.

SCHLUMBERGER. *Log interpretation charts*. Texas: Schlumberger, 2009. 310 p. ISBN 9781937949105. Citado 2 vezes nas páginas 39 e 47.

SCHLUMBERGER. *Schlumberger wireline services catalog*. Schlumberger, 2015. 66 p. Disponível em: <https://www.slb.com/~/media/Files/evaluation/catalogs/2015_wireline_services_catalog.pdf>. Citado 4 vezes nas páginas 21, 96, 99 e 146.

SERRA, O. *The fundamentals of well log interpretation. 1. the acquisition of logging data*. First. [S.l.]: Elsevier B.V., 1984. 423 p. ISSN 0036-8075. ISBN 0444421327. Citado na página 38.

SHABAB, M.; JIN, G.; NEGARA, A.; AGRAWAL, G. New data-driven method for predicting formation permeability using conventional well logs and limited core data. In: . [S.l.]: Society of Petroleum Engineers, 2016. p. 10. Citado na página 71.

SILVA, Y. M. P.; NEUMANN, R.; RAMNANI, C. W.; ÁVILA, C. A. Digital mineralogy: advances in pre-salt rocks characterization using automated mineralogical mapping of petrographic thin sections by sem. In: . [S.l.]: Brazilian Petroleum, Gas and Biofuels Institute, 2020. p. 13. Citado na página 61.

SIMON, A. H. *Handbook of Thin Film Deposition*. 4. ed. [S.l.]: Elsevier Inc., 2018. 392 p. Citado na página 57.

STADTMULLER, M.; LIS-SLEDZIONA, A.; SŁOTA-VALIM, M. Petrophysical and geomechanical analysis of the lower paleozoic shale formation, north poland. *Interpretation*, v. 6, p. SH91–SH106, 2018. ISSN 23248866. Citado na página 60.

STRALEY, C.; ROSSINI, D.; VINEGAR, H.; TUTUNJIAN, P.; MORRISS, C. Core analysis by low-field NMR. *Log Analyst*, v. 38, n. 2, p. 84–93, 1997. ISSN 0024581X. Citado na página 55.

SUN, J.; INNANEN, K. A.; HUANG, C. Physics-guided deep learning for seismic inversion with hybrid training and uncertainty analysis. *Geophysics*, v. 86, p. R303–R317, 2021. ISSN 0016-8033. Citado na página 73.

ULLOA, J. M.; CHAPARRO, D.; LARA, S.; ARANGO, S.; MENDEZ, F.; ALARCON, N.; GADE, S. An innovative cased-hole, oil-saturation method of utilizing excess carbon analysis of pulsed neutron measurements in a siliciclastic cenozoic formation, los llanos basin, colombia. In: . [S.l.]: Society of Petrophysicists and Well-Log Analysts, 2016. p. 12. Citado 2 vezes nas páginas 30 e 50.

VERMA, A. K.; CHEADLE, B. A.; ROUTRAY, A.; MOHANTY, W. K.; MANSINHA, L. Porosity and permeability estimation using neural network approach from well log data. In: . [s.n.], 2012. p. 6. Disponível em: <http://www.searchanddiscovery.com/documents/2014/41276verma/ndx_verma>. Citado na página 71.

VICTOR, R. A. *Multiscale, image-based interpretation of well logs acquired in a complex, deepwater carbonate reservoir*. 240 p. Tese (Doutorado) — The University of Texas at Austin, May 2017. Disponível em: <https://repositories.lib.utexas.edu/handle/2152/62228?show=full>. Citado na página 53.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; WALT, S. J. van der; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C. J.; POLAT İlhan; FENG, Y.; MOORE, E. W.; VANDERPLAS, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; MULBREGT, P. van; VIJAYKUMAR, A.; BARDELLI, A. P.; ROTHBERG, A.; HILBOLL, A.; KLOECKNER, A.; SCOPATZ, A.; LEE, A.; ROKEM, A.; WOODS, C. N.; FULTON, C.; MASSON, C.; HäGGSTRöM, C.; FITZGERALD, C.; NICHOLSON, D. A.; HAGEN, D. R.; PASECHNIK, D. V.; OLIVETTI, E.; MARTIN, E.; WIESER, E.; SILVA, F.; LENDERS, F.; WILHELM, F.; YOUNG, G.; PRICE, G. A.; INGOLD, G. L.; ALLEN, G. E.; LEE, G. R.; AUDREN, H.; PROBST, I.; DIETRICH, J. P.; SILTERRA, J.; WEBBER, J. T.; SLAVIč, J.; NOTHMAN, J.; BUCHNER, J.; KULICK, J.; SCHöNBERGER, J. L.; CARDOSO, J. V. de M.; REIMER, J.; HARRINGTON, J.; RODRíGUEZ, J. L. C.; NUNEZ-IGLESIAS, J.; KUCZYNSKI, J.; TRITZ, K.; THOMA, M.; NEWVILLE, M.; KüMMERER, M.; BOLINGBROKE, M.; TARTRE, M.; PAK, M.; SMITH, N. J.; NOWACZYK, N.; SHEBANOV, N.; PAVLYK, O.; BRODTKORB, P. A.; LEE, P.; MCGIBBON, R. T.; FELDBAUER, R.; LEWIS, S.; TYGIER, S.; SIEVERT, S.; VIGNA, S.; PETERSON, S.; MORE, S.; PUDLIK, T.; OSHIMA, T.; PINGEL, T. J.; ROBITAILLE, T. P.; SPURA, T.; JONES, T. R.; CERA, T.; LESLIE, T.; ZITO, T.; KRAUSS, T.; UPADHYAY, U.; HALCHENKO, Y. O.; VáZQUEZ-BAEZA, Y. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, v. 17, p. 261–272, 2020. ISSN 15487105. Citado na página 100.

WENDEMUTHT, A. Learning the unlearnable. *Journal of Physics A: Mathematical and General*, v. 28, p. 5423–5436, 1995. Citado 2 vezes nas páginas 66 e 70.

WESTAWAY, P.; HERTZOG, R.; PLASEK, R. E. Neutron-induced gamma ray spectroscopy for reservoir analysis. *Society of Petroleum Engineers Journal*, v. 23, p. 553–564, 1983. ISSN 0197-7520. Citado 2 vezes nas páginas 30 e 50.

WESTPHAL, H.; SURHOLT, I.; KIESL, C.; THERN, H. F.; KRUSPE, T. NMR measurements in carbonate rocks: Problems and an approach to a solution. *Pure and Applied Geophysics*, v. 162, n. 3, p. 549–570, 2005. ISSN 00334553. Citado na página 56.

WITTEN, I. H.; FRANK, E. *Data mining: practical machine learning tools and techniques*. Second. [S.l.]: Morgan Kaufmann Publishers, 2005. 525 p. ISSN 14337851. ISBN 0080890369. Citado 3 vezes nas páginas 63, 66 e 67.

WU, P.-Y.; JAIN, V.; KULKARNI, M. S.; ABUBAKAR, A. Machine learning–based method for automated well-log processing and interpretation. In: . [S.l.]: Society of Exploration Geophysicists, 2018. p. 2041–2045. ISBN 2018299697. Citado na página 71.

WYLLIE, M. R. J.; GREGORY, A. R.; GARDNER, L. W. Elastic wave velocities in heterogeneous and porous media. *Geophysics*, v. 21, n. 1, p. 41–70, 1956. Disponível em: <http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf>. Citado na página 51.

XGBOOST-DOCUMENTAçãO. 2016. <https://xgboost.readthedocs.io/en/latest/parameter.html>. Acessado: 2019-06-05. Citado na página 82.

YE, S. ju; RABILLER, P. Automated electrofacies ordering. *Petrophysics*, v. 46, p. 409–423, 2005. ISSN 15299074. Citado na página 71.

ZHANG, D.; CHEN, Y.; MENG, J. Synthetic well logs generation via recurrent neural networks. *Petroleum Exploration and Development*, Research Institute of Petroleum Exploration and Development, PetroChina, v. 45, p. 629–639, 2018. ISSN 18763804. Disponível em: <http://dx.doi.org/10.1016/S1876-3804(18)30068-5>. Citado 2 vezes nas páginas 72 e 73.

ZHANG, H.; LIANG, Y.; MA, J.; QIAN, C.; YAN, X. An milp method for optimal offshore oilfield gathering system. *Ocean Engineering*, v. 141, p. 25–34, 2017. ISSN 00298018. Citado na página 49.

ZHAO, T.; VERMA, S.; DEVEGOWDA, D.; JAYARAM, V. Toc estimation in the barnett shale from triple combo logs using support vector machine. In: . [S.l.]: Society of Exploration Geophysicists, 2015. p. 791–795. Citado na página 71.

Anexos

Anexo A – Artigo publicado: perfis geoquímicos sintéticos

Contents lists available at ScienceDirect

Journal of Petroleum Science and Engineering

journal homepage: http://www.elsevier.com/locate/petrol

Synthetic geochemical well logs generation using ensemble machine learning techniques for the Brazilian pre-salt reservoirs

Lucas Abreu Blanes de Oliveira [a,b,*], Cleyton de Carvalho Carneiro [b]

[a] Petrobras – Petróleo Brasileiro S.A., Avenida Henrique Valadares 28, Centro, Rio de Janeiro, 20231-030, Rio de Janeiro, Brazil
[b] Universidade de São Paulo, USP, Escola Politécnica, Praça Narciso de Andrade, Vila Mathias, Santos, São Paulo, 11013-560, Brazil

ARTICLE INFO

Keywords:
Neutron-induced spectroscopy
Gamma ray spectroscopy
Formation evaluation
Artificial intelligence

ABSTRACT

Geochemical logs are an essential tool for hydrocarbon reservoir characterization. The rock composition given by these logs is useful for porosity calculation, stratigraphic modeling, diagenesis estimation, and well acidification. In reduced logging operations performed in carbonate reservoirs of the Brazilian pre-salt, the geochemical tool is no longer acquired, aiming at optimization and cost reduction. This research aims the development of synthetic geochemical logs using machine learning algorithms. The database includes 22 wells with complete logging acquisition, and the input logs are natural gamma-ray, gamma-ray spectroscopy, density, photoelectric factor, neutron porosity, nuclear magnetic resonance, and sonic. The chemical elements chosen as output are Al, Ca, Fe, Mg, Na, Si, S, and Ti. Five models based on machine learning are trained using 19 wells: Support-vector machine, Multilayer perceptron, Random Forest, AdaBoost, and XGBoost. AdaBoost represents the best algorithm because, in addition to showing the best results, it allowed for a more simplified preprocessing and hyperparameter tuning. The evaluation of the models applies R^2 and root mean squared error (RMSE) in validation data and cross-validation. Robust models are acquired for Al, Ca, Fe, Mg, Si, S, and Ti, with R^2 above 0.80. Na shows R^2 slightly above 0.70. All the models have RMSE between 10^{-2} to 10^{-4}. The most important logs during training are density, photoelectric factor, K, Th, U, and neutron porosity. Synthetic geochemical logs created for three test wells show a good agreement with acquired logs, being able to reproduce general trends of the pre-salt formations. The machine learning model is capable of substituting the acquisition of geochemical logs with high confidence, representing cost reduction, and supplying engineers and geoscientists with quality data to be used in formation evaluation.

Problem statement

Well log acquisition reduces drastically from the exploratory to the development phase, aiming optimization and cost reduction. In wells drilled in the pre-salt carbonate reservoirs operated by Petrobras the number of runs is cut by one third in the reduced wireline logging operations. The geochemical logs are no longer acquired in those wells. This research proposes the development of synthetic geochemical logs using machine learning techniques.

Major results

Five machine learning algorithms were tested, ranging from Support-vector Machines to Artificial Neural Networks and decision tree-based ensemble methods. Tree-based ensemble boosting algorithms were considered better suited to deal with the creation of synthetic geochemical logs. The synthetic logs created showed a good agreement with acquired logs, being able to capture general trends in rock formation. The machine learning model was capable of substituting the acquisition of geochemical logs with high confidence, representing costs reduction and supplying engineers and geoscientists with quality data to be used in formation evaluation.

About the author

Lucas Blanes de Oliveira

Geologist graduated at the University of São Paulo, currently working in Petrobras. Started as a well site geologist working in Bahia (Reconcavo Basin), Espírito Santo (both Espírito Santo and Campos

* Corresponding author.
E-mail address: lucas.oliveira@usp.br (L.A. Blanes de Oliveira).

https://doi.org/10.1016/j.petrol.2020.108080
Received 5 February 2020; Received in revised form 18 October 2020; Accepted 22 October 2020

Basin) and Santos (Santos Basin). Works in the support of geological, geophysical and petrophysical data acquisition in Rio de Janeiro, with expertise in wells drilled in the pre-salt carbonates. Has specialization in petroleum geology and petrophysics. Started his Master's Degree in 2019 researching the applications of machine learning and artificial intelligence in well logs and petrophysical evaluation.

Cleyton de Carvalho Carneiro

Cleyton Carneiro is a Professor at the Escola Politécnica of the Universidade de São Paulo (USP) since 2013. He graduated in Geomatics with an emphasis on Remote Sensing from the Centro Federal de Educação Tecnológica do Pará (CEFET-PA) in 2002. He also graduated in Geology from the Universidade Federal do Pará (UFPA) in 2003. He became a Master in Geosciences from the Universidade Estadual de Campinas (UNICAMP) in 2005. He obtained his Ph.D. in Geology and Natural Resources from the Instituto de Geociências of UNICAMP in 2010, with a supervised period at the Australian Commonwealth Scientific and Research Organization (CSIRO) in Australia, in 2009. He completed his post-Ph.D. at the Instituto de Geociências of USP, in 2014, having served as a Visiting Scientist at the United States Geological Survey (USGS) in the United States, in 2013. His area of expertise includes Data Science and Geotechnologies applied to Natural Resources, with an emphasis on oil. He is accredited as an advisor with the Postgraduate Program in Naval and Ocean Engineering (PPGEN). He is one of the leaders of the Research, Development, and Innovation (RD&I) Integrated Technology for Rock and Fluid Analysis (InTRA) group, since 2017. His researches are related to analysis and visualization of databases with multiple variables, and digital transformation involving the integrated characterization of rocks and fluids, within the scope of sedimentary basins and oil and gas reservoirs.

1. Introduction

Among the applications of well logging, geochemical logs can generate several advantages for the characterization of oil and gas reservoirs. The records of relative abundance of chemical elements enable better compositional characterization of the formation mineralogy (Ellis and Singer, 2007). In this way, geochemical logs can help to calculate porosity and saturation, to generate stratigraphic and diagenetic models, as well as to determine zones for acidification.

Well logging investments applied to drill activities in ultra-deep waters are strategic and can be reduced by logistical and operational optimization. Because of this, several operators discontinue the acquisition of geochemical logs, acquiring only the conventional well logs sets.

In a database with multiple correlated variables, machine learning algorithms are suitable to predict trends or impute missing values in one or more variables. Several works already applied machine learning to wireline data with different objectives, among them the creation of synthetic logs (Chen et al., 2005; Rolon et al., 2009; Akkurt et al., 2018). However, this is usually applied to conventional logs, such as resistivity and density, in wells where tool failure has occurred. No research proposes the complete replacement of a wireline tool, even more so one as complex as the geochemical tool. Furthermore, the vast majority of researches use artificial neural networks as the chosen algorithm (Korjani et al., 2016; Salehi et al., 2017; Zhang et al., 2018).

Regarding ultra-deepwater reservoirs, especially those discovered on offshore southeastern Brazil, there are some wells with geochemical logs in contrast to others where this information has not been collected. The available geochemical logs could serve as training points for machine learning in the generation of synthetic data. In this way, machine learning algorithms could be able to complete the missing information based on the multivariate architecture of the integrated database.

This research aims to develop synthetic geochemical logs based on ensemble machine learning, techniques little used in the oil industry, and compare them with other algorithms. Synthetic records of geochemical logs will be obtained by tree-based ensemble boosting algorithms. The synthetic logs will be validated from real logs not used in the training database, as well as compared with analyzes from other machine learning algorithms. The cost-benefit relationship will be assessed, as well as errors associated with a model generated in carbonate reservoirs located in the Santos Basin Pre-Salt, on the southeast coast of Brazil. The criteria used in the comparison of the algorithms will take into account not only the errors observed between real and synthetic data. Simplicity in the construction and calibration of the model and computational performance will also be observed and evaluated. In this way, methods and results of simple reproduction and access will be presented in order that engineers and geoscientists can reproduce them even if they are not machine learning experts.

2. Conceptual background

2.1. Pre-salt reservoirs

The Brazilian pre-salt reservoirs are the most relevant oil discovery in recent years (Beltrao et al., 2009). Located in Santos Basin (Fig. 1), these reservoirs are mainly composed of stromatolitic carbonates from the Barra Velha Formation and grainstones, wackestones, and packstones from the Itapema Formation (Moreira et al., 2007). Microbial laminites, microbialites, and carbonate shales are also observed, as well as sandstones, pelites, and shales from the Piçarras Formation. These rocks are superimposed on the crystalline basement of the Camboriú Formation, and intercalations with basalts can also occur. Fig. 2 shows the lithology of two exploratory wells drilled in the pre-salt reservoir, exemplifying the sequences described.

The pre-salt context involves geological heterogeneity and numerous technological challenges, with water depths higher than 2200 m and targets higher than 5000 m. This context requires engineers and geoscientists to use all the necessary resources to reduce uncertainties and risks to explore and develop the pre-salt hydrocarbon. In this context, well logs are an essential tool used in formation evaluation.

Despite these needs, the search for optimization and cost reduction requires that certain information must be removed from the wireline acquisition as we move from the exploratory to the development phase of an oil field. Table 1 shows the comparison between logs acquired in an exploratory (complete wireline logging) and a development well (reduced wireline logging) drilled in the pre-salt reservoirs operated by Petrobras. The number of runs is cut by one third, and the total time is reduced by approximately ten days, generating an economy of millions of dollars per well.

Notwithstanding the best efforts of engineers and geoscientists, the decision to stop acquiring a log is taken considering the experience from the past; thus, future problems in reservoir and production management can lead to regret. An example is the geochemical logs. As shown in Table 1, removing the geochemical logs from Run 2 in an exploratory well allows the combination of Runs 1 and 2, helping to reduce the total number of runs in development wells. Geochemical logs (or neutron-induced gamma-rays spectroscopy logs) allows the detection of several chemical elements in the formation (Ellis and Singer, 2007). This information is used to create an accurate mineralogy model of the rock matrix, useful for porosity calculation, stratigraphic and diagenesis modeling, and helping in well acidification.

2.2. Geochemical logs

Geochemical logging tools explore the gamma-ray spectrum generated by interactions between emitted neutrons and atoms from the chemical elements presented in the formation (Ellis and Singer, 2007). Besides elastic scattering, which is a non-radioactive interaction, inelastic scattering and radiative capture excite atoms nuclei that, during decay, produce gamma-ray with energies dependent on the isotopes and

Fig. 1. Location of the Brazilian pre-salt and the area of study. Fields and well location were omitted due to confidentiality policies.

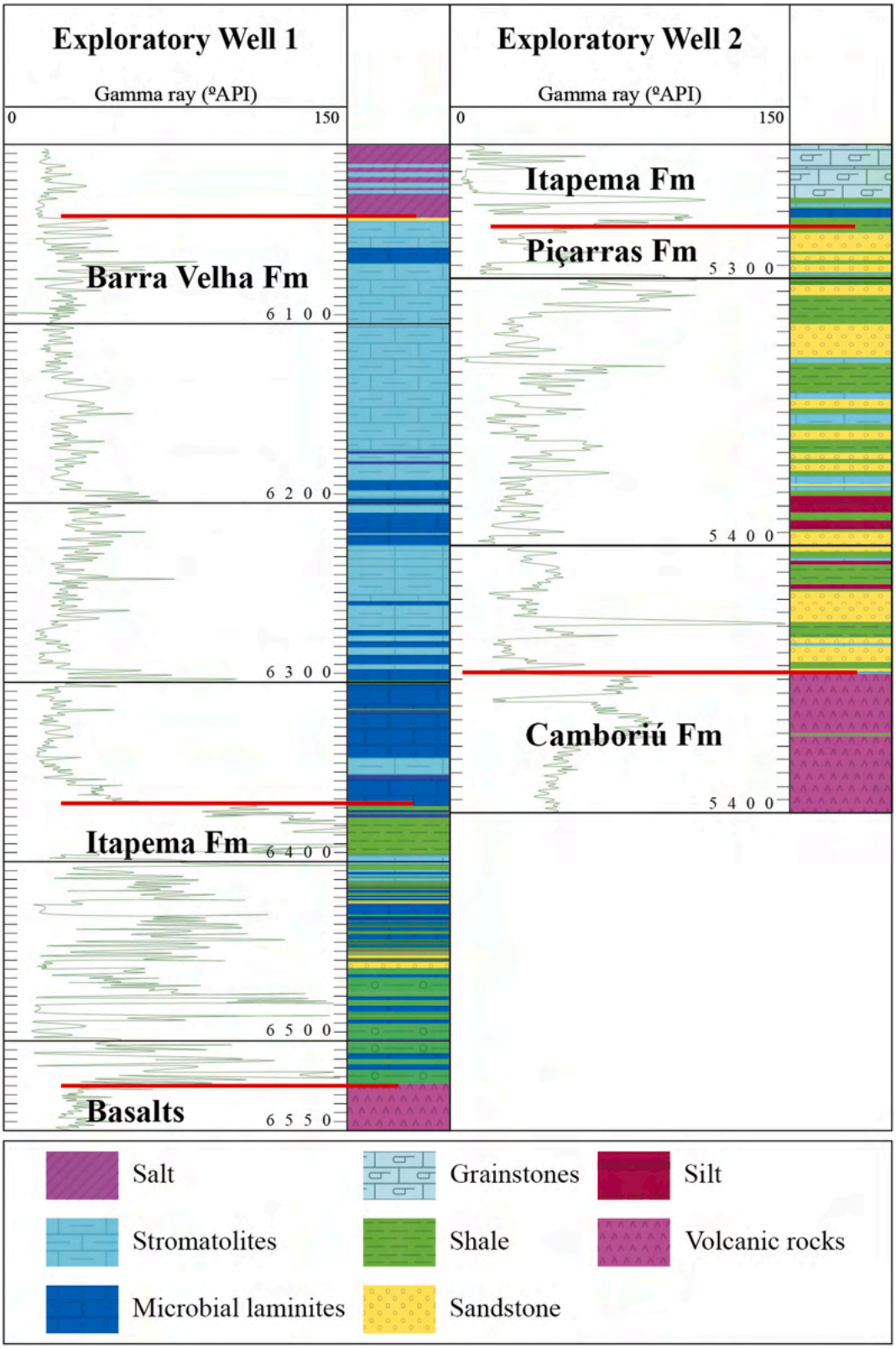


Fig. 2. Lithology observed in two exploratory wells drilled in the Brazilian pre-salt, adapted from Moreira et al. (2007).

Table 1
Comparison between wireline logging operations done in exploratory and development wells in the pre-salt fields operated by Petrobras. This information was gathered from Petrobras experts in wireline logging operations.

Logs acquired in exploratory wells	Run 1	Gamma ray
		Resistivity
		Density and photoelectric factor
		Neutron porosity
	Run 2	Nuclear magnetic resonance
		Geochemical logs
		Gamma ray spectroscopy
	Run 3	Fluid sampling and pressure measurements
	Run 4	Rock sampling
	Run 5	Acoustic image
		Resistivity image
		Sonic
	Run 6	Vertical seismic profile
	Total logging time – 10 to 12 days	
Logs acquired in development wells	Run 1	Gamma ray
		Resistivity
		Density and photoelectric factor
		Neutron porosity
		Nuclear magnetic resonance
		Gamma ray spectroscopy
	Run 2	Acoustic image
		Sonic
		Pressure measurements
	Total logging time – 2 to 3 days	

the type of interaction that promoted the excitement. The resultant spectrum, which is an overlay of several emissions, are decomposed using elementary reference spectra representing the signatures of the chemical elements present in the formation and wellbore (Fig. 3). The coefficients associated with each reference spectra are called yields and represent the contribution of each element in the gamma-ray spectrum.

The determination of absolute elemental weight fraction concentrations measured by prompt gamma-ray detection following thermal neutron capture is challenging in the borehole environment (Hertzog et al., 1989). The yields cannot be directly related to elemental weight fraction due to 1. Each element has an intrinsic radioactive emission degree (i.e., the same content of two different elements will have different amounts of radiation emitted); 2. Yields are calculated for both

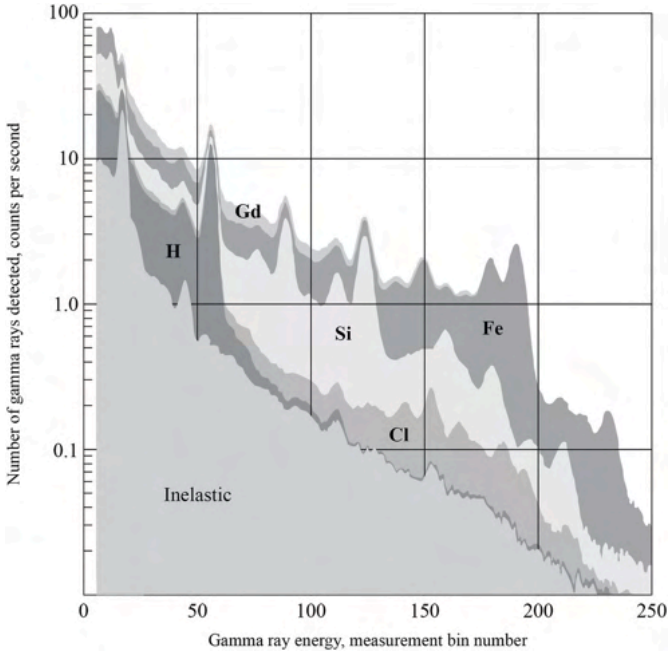


Fig. 3. An example of a gamma-ray spectrum. From Ellis and Singer (2007).

fluid and matrix alike concerning the total amount of radiation emitted; therefore, the yields of elements presented in the rock matrix will vary with parameters such as porosity and salinity; and 3. Elements like C and O are present both in the fluid and the rock matrix, being impossible to quantify the matrix contribution in the gamma-ray spectrum exclusively.

To work around these limitations, the yields of elements coming from the fluid (e.g., H, O, C, and Cl) are discarded, and the remaining elements are recalculated through an oxides closure model which is expressed as

$$F\left\{\sum_i X_i \frac{Y_i}{S_i}\right\} = 1 \tag{1}$$

$$W_i = F\frac{Y_i}{S_i} \tag{2}$$

where F = normalization factor; Y_i = fraction of the gamma-ray spectrum of element i; S_i = relative weight fraction detection sensitivity for element ; X_i = ratio of the weight of the respective oxide or carbonate to the weight of element i; and = absolute elemental weight fraction of element . Equations (1) and (2) are adapted from Hertzog et al. (1989).

Relative elemental weight fraction concentration is a division of Y_i (elemental yield) by S_i (relative spectral sensitivity). They account for several issues, such as efficiency in gamma-ray transport and detection, thermal neutron cross-section differences, gamma-ray multiplicities, and atomic weights. The relative elemental concentrations are related to (absolute concentrations) by F (normalization factor), which is calculated such that the sum of the elemental weights is equal to one. In the model, the absence of C and O is treated with an approximation: each of the quantifiable elements combines as an oxide or carbonate. The sum of these oxides and carbonate adds up to one in Equation (1). Once F is determined, Equation (2) calculates the absolute elemental weight fractions.

Modern geochemical tools use a pulsed-neutron generator (PNG) instead of chemical sources (Pemper et al., 2006; Radtke et al., 2012). The PNG allows the acquisition of two separated gamma-ray spectra in different time windows: one during the neutron burst, strongly influenced by inelastic scattering, and another one between bursts, dominated by radiative capture. These distinct spectra provide better-quality inversion and oxide closure, allowing the use of elementary reference spectra more suitable for each interaction.

After the oxide closure model, the geochemical tool provides the absolute elemental weight fraction of several chemical elements that are present in the rock matrix, also known as the elements' dry weight. The dry weights are often used in mineralogy modeling, relevant in reservoirs with complex lithology (Flaum and Pirie, 1981; Gilchrist et al., 1982; Quirein et al., 1987; Anderson et al., 1988; Galford et al., 2009; Macdonald et al., 2010; Ajayi et al., 2015; Freedman et al., 2015; Zhao et al., 2017). These mineral models must be calibrated with X-ray fluorescence and diffraction of rock samples. Integration with information acquired by other logs, such as density and nuclear magnetic resonance, can further provide enhancements in porosity and water saturation calculations. A robust mineral model can indirectly give information about high uncertainty parameters such as cation exchange capacity (CEC), as shown by Herron (1986). A comparison between the total amount of carbon read by the geochemical tool and the carbon calculated in the oxide closure model can be used to estimate total organic carbon (Gonzalez et al., 2013). Geochemical log acquisition can also be made through-casing, providing C/O ratio used to map water/oil contact in producing wells and help with production management (Westaway et al., 1983; North, 1987; Ulloa et al., 2016).

2.3. *Predictive modeling using machine learning algorithms*

Machine learning has been extensively used as a tool for finding

correlations, and identifying structural patterns in data, being able to make predictions from it (Witten and Frank, 2005). According to Bishop (2006), when the training data is comprised of input vectors and their corresponding output vectors, the application is known as supervised learning. When input vectors are assigned to discrete categories, the application is a classification problem. If the problem outputs are continuous variables, it is a regression problem. When the input vectors do not have any corresponding values, the application is known as unsupervised learning. The goal may be clustering, density estimation, and dimensionality reduction.

2.4. Ensemble boosting

Among the various existing machine learning algorithms, those using ensemble methods are particularly interesting. Its principle is to combine several weak learners to produce a powerful ensemble. This combination can be through Bayesian models, bootstrap aggregation, or boosting. The most common weak learner is the decision tree, which can bring significant advantages for the learning process, including low computational cost, minimal preprocessing, and simplicity in choosing training parameters.

Boosting is one of the most powerful learning ideas in recent years (Hastie et al., 2009). Boosting not only combines weak learners to create one optimal model, but it also acts directly on the resulting prediction of each learner, adjusting the weight of an observation based on the last prediction. This section will focus on two boosting algorithms: adaptive boosting (AdaBoost, Freund and Schapire, 1997) and gradient boosting (GBoost, Friedman, 2001).

AdaBoost uses several weak learners and, at each iteration, finds the best learner based on the sample weights (Kuhn and Johnson, 2013). The samples start with the same weight $1/n$ (n: number of samples). A weak learner is used to fit the training data and its error is calculated. The samples weights are updated proportionally to their errors to the training data: the samples with more error have their weights increased in the next iteration. Samples that are difficult to predict have their weights increased until the algorithm can fit them. Hence, at each iteration, the algorithm learns different aspects of the data. These weighted learners are combined into an ensemble with a prediction capability stronger than the individual learners.

Algorithm 1 provides an intuitive version of the AdaBoost algorithm to a classification problem. The same principle can be applied to a regression problem.

Algorithm 1. Example of the AdaBoost algorithm. Adapted from Kuhn and Johnson (2013).

1. Start with two classes: $+1$ and -1.
2. The same weight is attributed to all samples.
3. **for** $i = 1$ to I **do**
4. Calculate the i_{th} prediction ($pred_i$) by fitting a weak learner using the weighted samples.
5. Calculate the error (err_i) as $err_i = \left|\frac{correct\ predictions - n}{n}\right|$.
6. Calculate the i_{th} stage value (stv_i) as $stv_i = \ln\left(\frac{1 - err_i + \varepsilon}{err_i + \varepsilon}\right)$, with ε being a stabilizer parameter with low value, to deal with errors close to zero.
7. Update the weights by increasing it where samples were incorrectly predicted and decreasing the weight where samples were correctly predicted.
8. **End**
9. Calculate the prediction by multiplying stv_i by $pred_i$, adding these quantities across . If positive, the sample is classified as $+1$; otherwise, it is -1.

GBoost also trains several models gradually and sequentially. The difference between AdaBoost and GBoost is how the algorithms identify the deficiencies of weak learners. While AdaBoost handles the deficiencies by changing the weights of the samples, GBoost performs the same by calculating the gradient of the loss function. Having a loss function and a weak learner, GBoost will find a sequential model that minimizes the said function. After the best guess is used to start the algorithm, the residual is computed, and a model fits the residuals, minimizing the loss function. This model is added to the previous one, and the algorithm iterates until the desired error is achieved or a number of iterations is reached (Kuhn and Johnson, 2013).

Algorithm 2 shows an example of a GBoost algorithm that uses decision trees as weak learners and squared error as the loss function.

Algorithm 2. Example of a GBoost algorithm. Adapted from Kuhn and Johnson (2013).

1. Start with tree depth T, $h(x)$.
2. Choose a loss function to minimize, e.g., the least-squared error $\frac{\sum_i^n (y_i - F(x))^2}{2}$, where n is the number of samples y and $F(x)$ is the predicted value of y.
3. Calculate the first error $F_0(x)$ using the average value of y as the first guess.
4. **for** $m = 1$ to M **do**
5. Calculate the residuals as $r_m = y - F_{m-1}(x)$.
6. Fit a tree $h_m(x)$ to the residuals.
8. Update the model by adding the fitted tree to the previous model, as $F_m(x) = F_{m-1}(x) + h_m(x)$.
9. **end**

A variation of the GBoost algorithm, known as XGBoost, was proposed by Chen and Guestrin (2016). Its advantages can be attributed to several optimizations, such as (i) a tree learning algorithm to handle sparse data; and (ii) a weighted quantile sketch procedure that handles instance weights in tree learning. The XGBoost algorithm was the gradient boosting algorithm chosen for this research.

According to Kuhn and Johnson (2013), even though any algorithm can be used as a weak learner in boosting, decision trees are excellent base learners. Decision trees consist of if-then statements that partition the data, with a model predicting the output information. They are excellent base leaners because of 1. Their depths can be restricted, turning them into simple weak learners; 2. Decision trees can be added together in an ensemble to create a model; 3. It is easy to aggregate the results of many trees since they can be created very quickly; 4. Trees can handle different types of input and output data, facilitating the preprocessing. Furthermore, boosting algorithms that use decision trees are capable of handling missing data, calculate variable importance, and conduct variable selection.

2.5. Other algorithms

The list of algorithms available for machine learning is long, and each day new techniques are created to deal with different types of data. However, some stand out from the rest due to their intense application in recent decades. Among them, we can mention the Support-Vector Machines (SVM), Artificial Neural Networks (ANN), and Random Forest (RForest).

SVM applied to regression was introduced by Drucker et al. (1997), an evolution of Cortes and Vapnik (1995). The basic principle to deal with classification problems is to build a set of hyperplanes in the n-dimensional space of the variables that are capable of separate the data into their correct labels. In regression, these hyperplanes fit the data, being positioned in a way as to minimize the average distance between the data points. SVM deal with non-linearity through a kernel that transforms the data, moving to a multi-dimensional space where it is possible to perform non-linear regressions.

ANN is one of the most popular machine learning algorithms in recent years, gaining notoriety in deep learning problems, mainly for image classification. Its central element is the artificial neuron, first introduced by McCulloch and Pitts (1943). The neuron receives information from an external source and computes an output using an activation function, responsible for introducing non-linearity. The most common ANN is the Multilayer Perceptron (MLP), introduced by Rosenblatt (1957) and expanded by researches such as Wendemutht (1995) and Freund and Schapire (1999). It is a feedforward network containing multiple neurons arranged in layers. Training is done by updating the

weights assigned to each of the connections between neurons, using an error minimization technique known as back-propagation.

Introduced by Ho (1995) and expanded by Breiman (2001) and Breiman et al. (2018), RForest is an ensemble method that uses bootstrap aggregation and decision trees in its core. It is common for decision trees to overfit the data, expressed in their high variance and low bias. Bootstrap aggregation is used to reduce the variance of the model by dividing the data into subsets, fitting several trees to these subsets, and combining their prediction. Still, those trees can heavily rely on just a few variables from the training data. RForest solves this issue by also randomly selecting the variables to create the subsets, further reducing the variance of the model.

The main difference between RForest and AdaBoost and GBoost is that RForest does not use boosting, as it does not use the error of the previous prediction to create the next tree. Boosting algorithms are sequential by nature, evolving to predict the data better where the error is high. RForest does that by sheer force since the trees do not communicate with each other. In that sense, it is expected that the number of trees will be greater in RForest.

2.6. Variable importance

An advantage of tree-based ensemble algorithms is the ability to calculate the importance that each variable has in the final model. Known as variable importance, it provides a metric that indicates how valuable each variable was during training. This metric can be calculated in several ways, the main ones being (i) counting how many times a variable was used when splitting a tree node; (ii) exchanging one variable for another and analyzing how it affects the adjustment of that tree and (iii) evaluating how the adjustment of a tree is affected when a variable is randomly removed from the training. In this research, the third one was chosen.

The calculation of the variable importance reflects the advantage of tree-based algorithms over others. Variables that have little importance in improving the model will gradually be discarded during training. Thus, it is not necessary to remove any input variables in preprocessing. In other algorithms, such as SVM and ANN, all input variables will affect the training, and any unwanted variables must be removed in the preprocessing phase.

2.7. Machine learning applied to well logs

Machine learning algorithms have been applied increasingly to well log data aiming at different goals. In general, their use can be divided into three categories: (i) prediction; (ii) interpretation; and (iii) data substitution.

Prediction problems aim to calculate parameters of interest such as porosity, permeability, and water saturation with logs using mainly supervised learning algorithms. Take water saturation (Sw) as an example. Many models to calculate Sw were proposed since Archie's equation (Archie, 1942), some empirical and other analytical. However, empirical methods have low representativeness (e.g., Archie's equation does not work well in shaly sands or complex carbonates) and analytical solutions are often too complex and rely on many input parameters (e.g., see equations proposed by Garcia et al., 2017a and Garcia et al., 2017b). With machine learning, one is capable of training a model to predict Sw using the result of core analysis, which is the most reliable source, and their respective response in well logs. Coming up with an analytical solution to correlate well log data such as gamma-ray or P-wave velocity with Sw is a non-trivial, nearly impossible task to a human, but can be easily achieved by a machine learning algorithm given enough data is available.

Examples of machine learning algorithms being used to calculate Sw can be found in Al-Bulushi et al. (2012) and Hamada et al. (2018). Examples to calculate porosity and permeability are found in Aminian et al. (2003), Nashawi and Malallah (2009), Verma et al. (2012) and Shabab et al. (2016) and to calculate total organic carbon can be found in Alizadeh et al. (2012), Zhao et al. (2015) and Negara et al. (2016).

Interpretation problems can be divided into facies classification and petrophysical evaluation. Facies classification can be either supervised or unsupervised. If the goal is to propagate lithology interpreted in core samples or a region of the well, throughout the same well or to other wells in the field using logs, then supervised algorithms can be applied (Bestagini et al., 2017; Hoeink and Zambrano, 2017; Guarido, 2018). If no training data is available, then an unsupervised algorithm can be used to estimate electrofacies using well logs (Ye and Rabiller, 2005; Asfahani et al., 2018). The automatic petrophysical evaluation aims to find interest zones in a well, often dividing the well between hydrocarbon-bearing formation, aquifers, and high-productive zones. (Belozerov et al., 2018; Wu et al., 2018). It uses, generally, supervised learning algorithms. The use of machine learning improves objectivity, efficiency, and consistency of well log interpretation.

Data substitution, where this research fits in, uses supervised machine learning to create new well log data where they were not acquired

Table 2
Summary of the characteristics of the wells used in the present research.

Field	Number of wells	Well diameter	Drilling mud	Formations found
1	1	8.50”	Water-based	Barra Velha
2	3	12.25”	Oil and water-based	Barra Velha Itapema Piçarras Camboriú
3	1	12.25”	Oil-based	Barra Velha
4	3	8.50” and 12.25”	Oil and water-based	Barra Velha Itapema
5	1	8.50”	Water-based	Barra Velha
6	1	12.25”	Oil-based	Barra Velha
7	8	12.25”	Oil-based	Barra Velha Itapema Piçarras Camboriú
8	4	12.25”	Oil-based	Barra Velha Itapema Piçarras

Table 3
Statistics of the input and output variables used.

Name	Description	Unit	Min	Max	Mean	Std
	Input variables					
Rho	Formation density	g/cm^3	1.59	4.04	2.52	0.12
GR	Natural gamma-ray	°API	3.25	238.0	34.6	22.6
NPHI	Neutron porosity	m^3/m^3	-0.02	1.12	0.12	0.06
PEF	Photoelectric fator	barns/e	1.72	10.0	4.90	0.92
DTP	Compressional wave slowness	μs/ft	41.0	131.0	62.8	7.71
DTS	Shear wave slowness	μs/ft	64.2	269.0	111.0	15.6
PhiT	NMR total porosity	m^3/m^3	0.00	2.53	0.08	0.07
PhiE	NMR effective porosity	m^3/m^3	0.00	2.53	0.11	0.07
FF	NMR free fluid	m^3/m^3	0.00	2.53	0.12	0.07
T2LM	NMR T2 log-mean	ms	0.53	2817	242.0	262.0
K	Potassium from natural gamma-ray spectroscopy	m^3/m^3	0.00	0.05	0.00	0.01
Th	Thorium from natural gamma-ray spectroscopy	ppm	0.00	39.3	1.70	1.47
U	Uranium from natural gamma-ray spectroscopy	ppm	0.13	20.5	2.72	1.99
	Output variables					
Al	Aluminum dry weight	m^3/m^3	0.00	0.11	0.00	0.01
Ca	Calcium dry weight	m^3/m^3	0.00	0.40	0.29	0.07
Fe	Iron dry weight	m^3/m^3	0.00	0.09	0.01	0.01
Mg	Magnesium dry weight	m^3/m^3	0.00	0.13	0.02	0.02
Na	Sodium dry weight	m^3/m^3	0.00	0.04	0.00	0.00
Si	Silicon dry weight	m^3/m^3	0.00	0.44	0.07	0.06
S	Sulfur dry weight	m^3/m^3	0.00	0.25	0.00	0.01
Ti	Titanium dry weight	m^3/m^3	0.00	0.02	0.00	0.00

or had bad quality. Many are the reasons for data inconsistency, varying from borehole conditions, tool failure, loss of data due to problems in storage, or simply cost reduction.

According to Rolon et al. (2009), machine learning algorithms are better than traditional methods (like multi-linear regression) when used to generate synthetic well logs. Bahrpeyma et al. (2013) identified that their variation of Fuzzy Logic, the Fast-Fuzzy Modeling Method, provided results slightly inferior when compared to regular Fuzzy Logic and Artificial Neural Networks, pointing at the fact that the latter requires higher computational power. Akkurt et al. (2018) proposed an interesting automated workflow for well log quality control, outlier detection, and log reconstruction. First, a Support-Vector Machine is trained to identify outliers in a vast dataset comprising hundreds of wells. Then, similarity metrics are used to compare wells based on their petrophysical footprints. Finally, a Quantile Regression Forest can generate new data to substitute outliers by using wells with a high degree of similarity when compared to the well where the outliers were detected. The objective of their workflow is to accelerate and enhance petrophysical analysis by reducing the amount of repetitive and burdensome quality control tasks.

3. Methodology

The database includes 22 wells drilled in the pre-salt carbonate reservoir from eight different fields, with a total of 77,800 instances. Their length adds up to 8600 m of rock with a complete set of well logs acquired (see logs acquired in exploratory wells in Section 1 - Table 1). Table 2 shows a summary of the characteristics of the wells used.

Table 3 summarizes the main characteristics of the input and output variables, as well as their main statistics. The input variables chosen were the logs that are still acquired in reduced wireline logging so that the synthetic geochemical logs can still be generated in the cost optimization scenario. The output variables were the most abundant chemical elements observed in the formations from the pre-salt reservoirs. Fig. 4 shows an example of logs acquired in one of the database wells.

A correlation matrix between variables is shown in Table 4. Some characteristics can be observed. Porosity logs (NPHI, FF, PhiE, and PhiT) present a positive correlation with each other and with DTP and DTS since porosity is the main responsible for the increase in the slowness in rocks. These logs have a negative correlation with Rho since porosity decreases the density of rocks. Among the radioactive elements (K, Th,

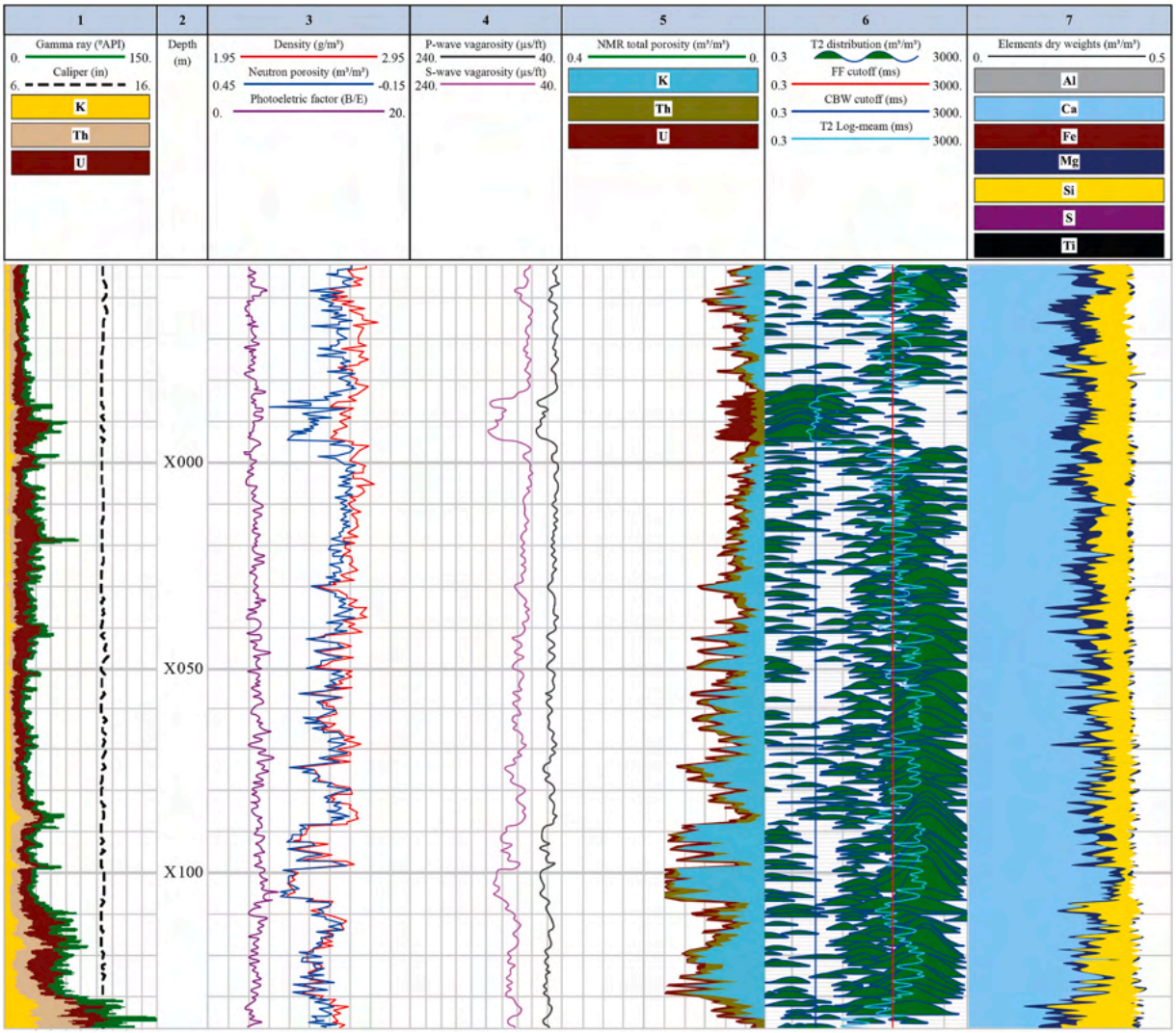


Fig. 4. Example of logs acquired in one of the database wells. Track 1: gamma-ray, caliper, and K, Th, U from gamma-ray spectroscopy, in °API. Track 2: depth. Track 3: density, neutron porosity, and photoelectric factor. Track 4: compressional and shear wave slowness. Track 5: total porosity, effective porosity, and free fluid from NMR. Track 6: T2 relaxation time from NMR, free fluid cutoff (100 ms), clay bound water cutoff (3 ms), and T2 log-mean. Track 7: cumulative elements' dry weights. True depths were omitted due to confidentiality policies.

Table 4
Correlation matrix between variables. Cold colors indicate a positive correlation, and warm colors indicate a negative correlation.

	Rho	GR	NPHI	PEF	DTP	DTS	FF	PhiE	PhiT	T2LM	K	Th	U	Al	Ca	Fe	Mg	Na	Si	Su	Ti
Rho	1																				
GR	0.087	1																			
NPHI	-0.716	0.076	1																		
PEF	-0.053	-0.087	0.103	1																	
DTP	-0.720	0.083	0.749	0.006	1																
DTS	-0.665	0.121	0.725	0.078	0.872	1															
FF	-0.453	-0.210	0.422	0.146	0.336	0.290	1														
PhiE	-0.512	-0.137	0.491	0.212	0.372	0.347	0.900	1													
PhiT	-0.521	-0.109	0.537	0.194	0.435	0.425	0.873	0.982	1												
T2LM	-0.254	-0.268	0.167	0.145	0.181	0.100	0.604	0.446	0.398	1											
K	0.226	0.624	0.035	-0.062	0.016	-0.004	-0.228	-0.205	-0.169	-0.166	1										
Th	0.159	0.534	0.065	-0.065	0.064	0.044	-0.159	-0.129	-0.092	-0.125	0.684	1									
U	-0.031	0.823	0.061	-0.079	0.089	0.159	-0.135	-0.062	-0.049	-0.257	0.176	0.173	1								
Al	0.316	0.322	0.009	0.092	-0.091	-0.061	-0.103	-0.054	-0.008	-0.035	0.592	0.352	0.019	1							
Ca	-0.311	-0.335	0.008	0.415	0.046	0.079	0.233	0.145	0.090	0.240	-0.501	-0.434	-0.084	-0.552	1						
Fe	0.345	0.311	0.036	0.172	-0.098	-0.071	-0.060	-0.008	0.037	0.016	0.621	0.448	-0.027	0.829	-0.588	1					
Mg	0.204	0.059	0.136	-0.349	-0.029	-0.074	-0.052	-0.056	-0.023	-0.101	0.116	0.217	-0.017	0.077	-0.503	0.125	1				
Na	0.200	0.082	-0.148	0.024	-0.163	-0.103	-0.136	-0.098	-0.076	-0.052	0.131	0.004	0.032	0.347	-0.126	0.259	-0.085	1			
Si	0.181	0.291	-0.083	-0.453	-0.007	-0.043	-0.253	-0.151	-0.110	-0.270	0.366	0.320	0.121	0.330	-0.888	0.365	0.160	0.081	1		
Su	0.213	0.175	-0.021	0.030	-0.049	-0.042	-0.137	-0.141	-0.127	-0.093	0.315	0.155	0.035	0.323	-0.215	0.301	0.023	0.129	0.104	1	
Ti	0.309	0.279	0.007	0.176	-0.107	-0.063	-0.050	0.015	0.056	0.033	0.542	0.349	-0.013	0.814	-0.483	0.901	0.037	0.281	0.285	0.249	1

and U), uranium is the main responsible for the increase in the natural radioactivity of pre-salt rocks, reflected by its higher correlation with GR. Among the output variables, there is a strong negative correlation between Ca and Si, demonstrating the variations in carbonate environments of the Barra Velha and Itapema Formations with the siliciclastic and igneous environments of the Piçarras and Camboriú Formations. Al, Fe, and Ti have a strong positive correlation with each other and a negative correlation with Ca, possibly being associated with minerals present in siliciclastic and igneous rocks. Mg has a negative correlation with Ca, reflecting the substitution of calcite (Ca carbonate) for dolomite (Ca and Mg carbonate), very common in the diagenetic processes of carbonate rocks.

Fig. 5 shows violin plots with the distribution of the variables. The creation of the models for the generation of synthetic logs followed the

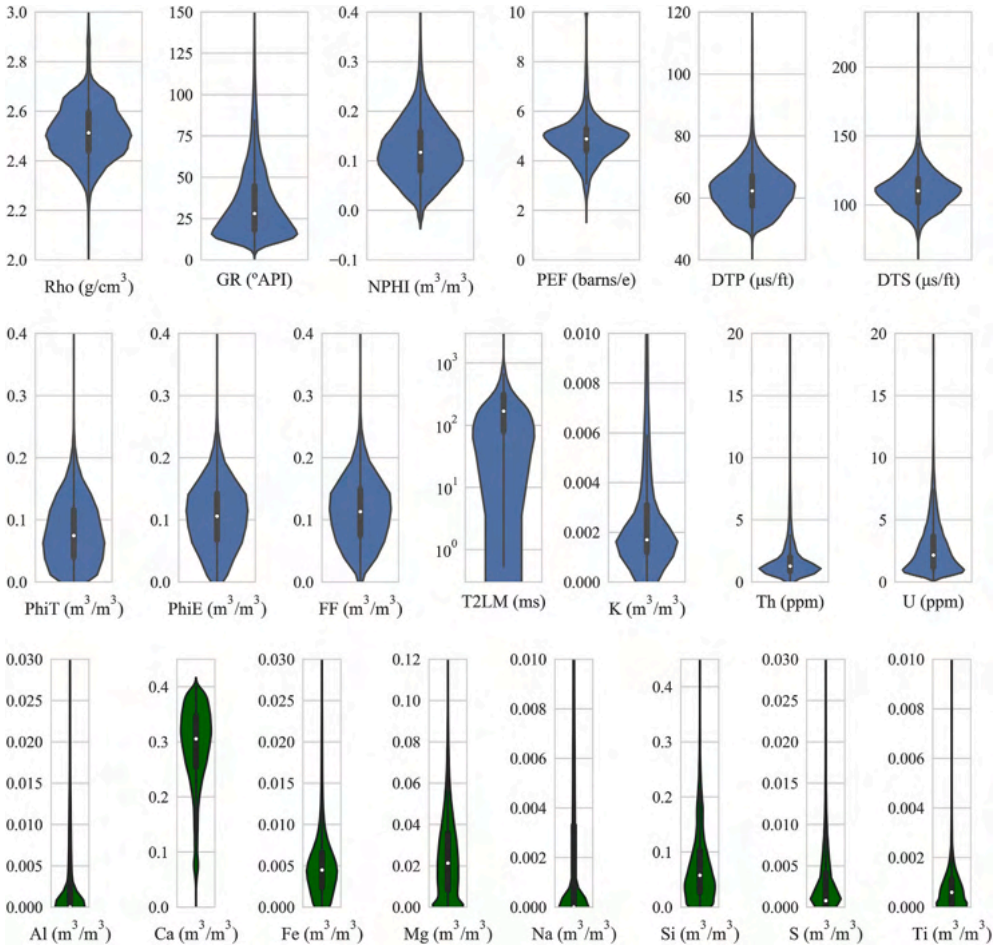


Fig. 5. Violin plots showing the variables' distributions.

workflow presented in Fig. 6, which will be detailed in the next section.

3.1. Data preparation and preprocessing

Quality control: the database was submitted to strict quality control. Environmental corrections were applied to the logs to take into account borehole and drilling fluid effects. Additionally, zones with intense rugosity or washout, identified with the caliper log, were removed from the model. These zones could compromise the quality of the logs, especially those who are acquired close to the borehole wall, like density, neutron porosity, and NMR (Fig. 7).

Training and test data: three wells were separated from the database to be used in the test phase. This is a common practice to ensure the generalizability of the model. The algorithms did not come into contact with these three wells during the training and evaluation phase. These wells were chosen based on their lithological column: in them, the main types of rock found in the pre-salt were observed - carbonate, siliciclastic and igneous rocks.

Standardization: the other 19 wells were standardized when necessary. The algorithms that need standardized data are SVM and MLP. The ensemble algorithms do not need to go through this step since the trees are based on decision-making processes. Standardization consists of transforming the data so that it has a mean of zero and a standard deviation of one.

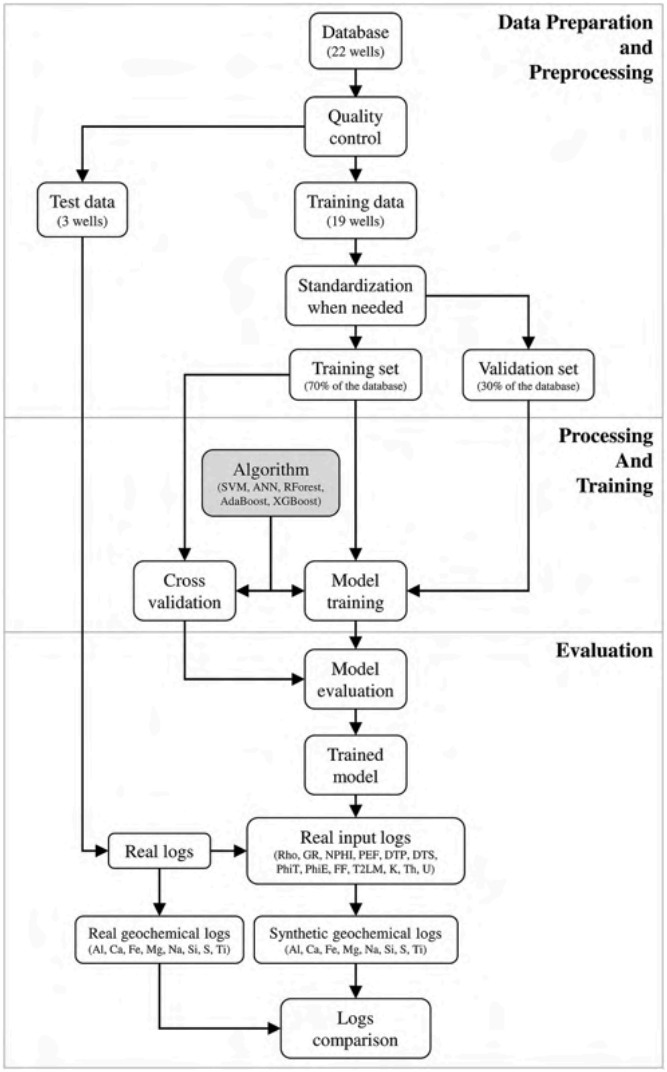


Fig. 6. Workflow used for machine learning model creation.

Training and validation sets: data from the remaining 19 wells (70,387 instances, 90% of the entire database) were grouped in a single database and randomly divided in two sets: a training set, consisting of 70% of the data (49,270 instances), and a validation set, with the remaining 30% (21,117 instances).

3.2. Processing and training

Model training: a model for each chemical element was trained using the training set, totaling eight models per algorithm. The trained models were then used to fit the validation set, and their results were evaluated using the coefficient of determination (R^2) and the root-mean-squared error (RMSE). The validation set results provide an unbiased evaluation of the models that were acquired when fitting the training set. However, an assessment based only on training and validation sets is very sensitive to how the validation set was obtained, so cross-validation is necessary.

Cross-validation: k-fold cross-validation was applied in the training set, dividing it into ten folds. At each step, nine folds are used as a training set, and one fold is used as a validation set. This process is repeated ten times, changing the folds until every fold was used at least once for validation. The resulting R^2 and RMSE are stored, and the mean and standard deviation are calculated. The cross-validation guarantees the robustness of the model and reduces overfitting.

Five machine learning algorithms were trained: SVM, MLP, RForest, AdaBoost, and XGBoost. Apart from XGBoost, the algorithms used are the ones available in Scikit Learn (Pedregosa et al., 2011). The XGBoost algorithm used is the one developed by Chen and Guestrin (2016). A hyperparameter tuning found the best parameters for each algorithm, presented in Table 5. A description of each hyperparameter can be found in Support-vector Regressor documentation (2019), Multi-layer Perceptron documentation (2015), Random Forest Regressor documentation (2019), AdaBoost documentation (2013) and XGBoost documentation (2015).

Table 5 shows that MLP is the most complex algorithm to tune. In addition to a more significant number of hyperparameters, the architecture of neuron layers is a complex property to be adjusted. Also, the higher the number of layers, and the number of neurons per layer, the higher the computational power required. SVM and RForest are the simplest algorithms to tune, followed by AdaBoost and XGBoost.

The number of trees is the most critical hyperparameter in ensemble boosting algorithms. While XGBoost finds it automatically, it needs to be determined for RForest and AdaBoost. Fig. 8 shows the evolution of the mean R^2 and RMSE of all chemical elements as new trees are added for the AdaBoost algorithm, representing the learning curve for the optimization. It can be seen that AdaBoost reaches a plateau in 50 trees, this number was chosen as the number of trees.

3.3. Evaluation

Model evaluation: the values of R^2 and RMSE of the validation set and cross-validation of all the algorithms were compared, and the best results were used to determine the best algorithm. The computational cost was also taken into account, counting the time necessary to carry out the training of the models of all chemical elements. These times will be measured on a computer with a 2.80 GHz Intel Core i5-8400 processor, 16 Gb RAM, and NVIDIA GeForce GTX 1050 Ti graphics card.

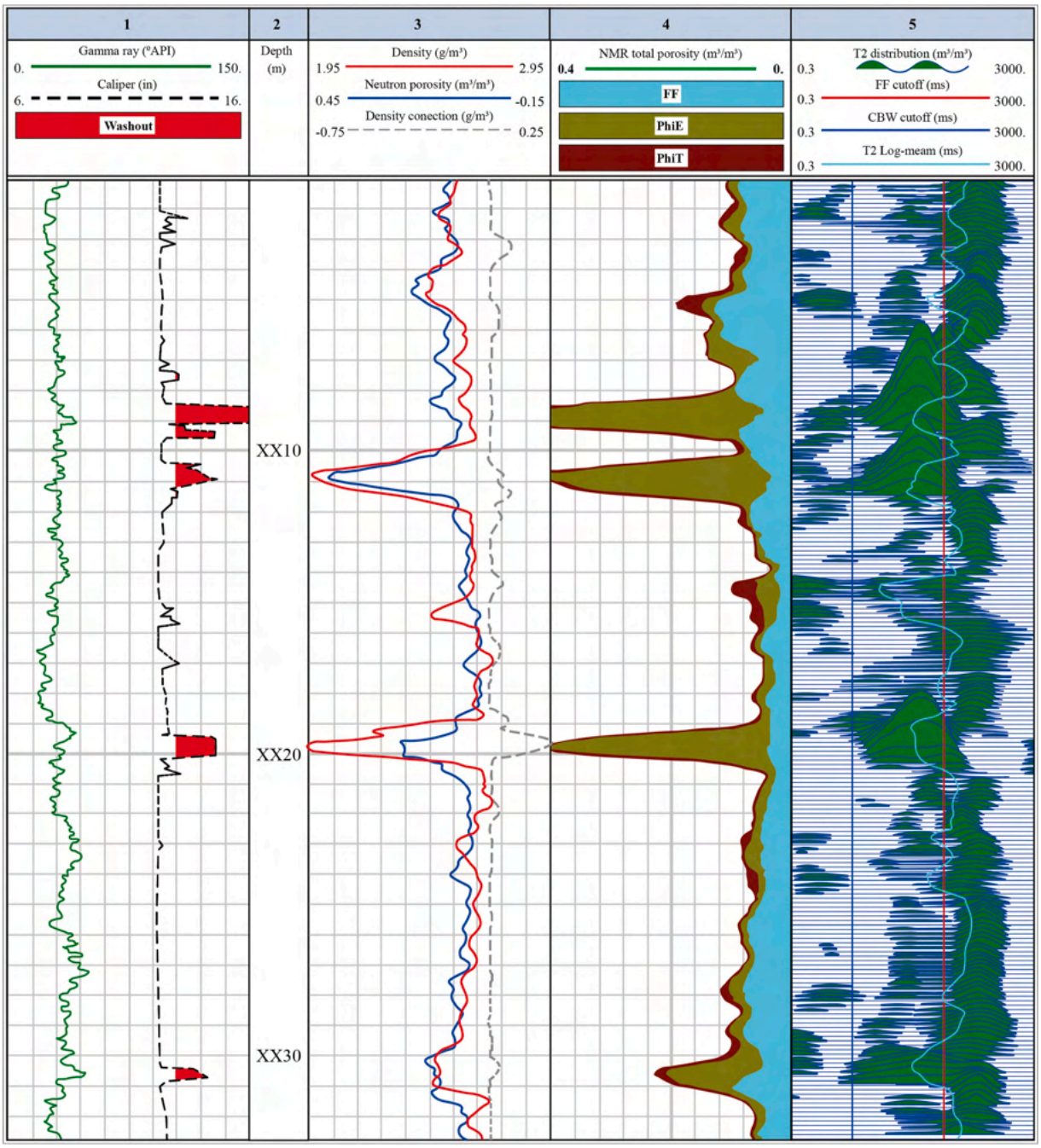


Fig. 7. Example of well zones removed in quality control. Depths with intense washout highlighted with red in the caliper log in Track 1, affect other logs such as density, neutron porosity, and NMR. These washout zones were removed from the model. True depths were omitted due to confidentiality policies. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Best hyperparameters for each machine learning algorithm used.

Algorithm	Hyperparameter type	Hyperparameter chosen
SVM	Kernel	Radial Basis Function
	Gamma	Scale
	Regularizer (C)	1
	Epsilon	0.1
MLP	Number of hidden layers	3
	Hidden layers size	100
	Activation function	ReLU
	Solver	Adam
	Alpha	0.001
	Learning rate	Adaptive
	Beta 1	0.9
	Beta 2	0.999
	Epsilon	1.00E-08
RForest	Number of trees	500
	Tree max depth	None
	Tree minimum sample leaf	1
	Tree minimum sample split	2
AdaBoost	Base learner	Decision tree
	Tree max depth	None
	Tree minimum sample leaf	1
	Tree minimum sample split	2
	Number of trees	50
	Learning rate	1
XGBoost	Base learner	Decision tree
	Tree max depth	26
	Tree minimum child weight	6
	Gamma	0
	Subsample ratio	0.8
	Columns sampling by tree	1
	Alpha	0
	Learning rate	0.1

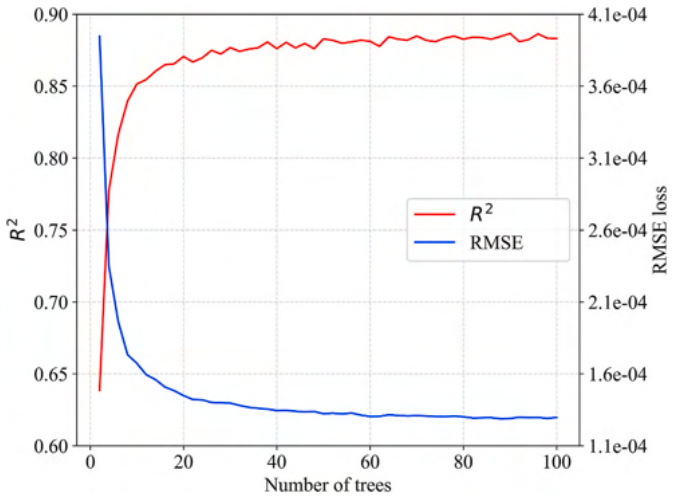


Fig. 8. Number of trees versus mean R^2 and RMSE of the validation sets for all chemical elements, showing a plateau after approximately 50 trees.

Uncertainty analysis of the results generated were made through the analysis of the standard deviation of the global error of the results of the validation set, the error being considered the difference between the real and synthetic data. For a better detailing of the uncertainties, they were also analyzed in low and high concentrations of the chemical elements.

Trained model and test: finally, the best model was used to generate synthetic geochemical logs in the test wells. Since the rocks observed in these wells are representative of the pre-salt formations, a good result attests to the model's ability to predict and generalize and can be used comprehensively in the pre-salt reservoir.

4. Results

4.1. Machine learning models

Table 6 shows the results of R^2 and RMSE obtained in the training and validation sets for all analyzed algorithms, while Table 7 shows the results of the cross-validation. The mean R^2 of all elements shows that AdaBoost had the best results, with 0.877 in the validation set and 0.872 in cross-validation. The lowest R^2 were those of the SVM, followed by the results of the MLP. Among the ensemble algorithms, RForest presented the worst results. All RMSE values were in the order of 10^{-3} or below, except for MLP, which was in the order of 10^{-1}. The low standard deviations of the cross-validation indicate that the models had excellent stability and generalizability.

Analyzing each element individually, Fe presented the best results in all algorithms, followed by Ti, Al, and Ca. In AdaBoost, the R^2 for these elements was above 0.90 both in the validation set and in cross-validation. After these elements, Si, S, and Mg showed the best results. Mg was better fitted in the two ensemble boosting algorithms, AdaBoost and XGBoost, with R^2 above 0.80 in the validation set and cross-validation. Mg is a crucial element in the pre-salt since it can detail the carbonate minerals present in the rocks and indicate zones with diagenesis. Finally, the only algorithm that obtained R^2 above 0.70 in the validation set and cross-validation for Na was AdaBoost. The other algorithms had much difficulty in adjusting a model for this element.

Fig. 9, Fig. 10, and Fig. 11 show real data *versus* synthetic data graphs for the AdaBoost results. Due to a large number of instances in the validation set, a density plot was created to facilitate visualization. It can be seen that the majority of the points fall over the 1:1 line, a reflex of the high R^2 obtained.

To analyze the uncertainty of the models, the standard deviations of the global errors obtained in the validation sets were measured, shown in Table 8. Except for Ca and Si, all the standard deviations of the global errors had values below 0.01 m^3/m^3. Even Ca and Si, elements with higher concentrations in the pre-salt rocks, showed a standard deviation just above 0.02 m^3/m^3, corroborating the excellent quality of the models.

An essential aspect of the uncertainty analysis is to assess how the uncertainties vary with the value of the variables. With this in mind, the standard deviations of the errors were calculated for low and high concentrations, showed in Table 8. It was considered low and high concentrations those below and above half of the maximum concentration observed for each chemical element. For example, the maximum concentration observed for Ca was 0.40 m^3/m^3 (as seen in Table 3); standard error deviation was calculated for concentrations below and above 0.20 m^3/m^3. Except for Ca, the standard deviations of the error increase for high concentrations. In the case of Ca, the standard deviation increases at low concentrations. It is possible to conclude that, for Ca, the uncertainty of the model tends to increase in low concentrations, and, for the other chemical elements, the uncertainty tends to increase in high concentrations.

4.2. Variable importance

Fig. 12 shows the relative importance of the input logs in the models created for each element. Ca, Mg, and Si models were profoundly affected by variations in photoelectric factor, density, K, and neutron porosity. Al, Fe, and Ti models were affected by K, density, photoelectric factor, Th, and U. S model was mainly driven by density. No particular input log had significant importance in the Na model.

The analysis of the importance of the variables shows that useful geochemical logs can still be generated in scenarios of a more significant reduction in wireline acquisition, provided that the logs of density, neutron porosity, photoelectric factor, and spectroscopy of natural gamma-rays (K, Th, and U) are acquired.

The capacity of ensemble algorithms to calibrate the importance of

Table 6
R^2 and RMSE results for the training set and validation set.

	Training set					Validation set				
	SVM	MLP	RForest	AdaBoost	XGBoost	SVM	MLP	RForest	AdaBoost	XGBoost
										R^2
Al	0.818	0.859	0.986	0.999	0.992	0.803	0.837	0.902	0.919	0.905
Ca	0.857	0.932	0.987	0.999	0.998	0.848	0.884	0.902	0.917	0.916
Fe	0.942	0.958	0.995	0.999	0.996	0.931	0.951	0.966	0.976	0.966
Mg	0.645	0.870	0.971	0.999	0.993	0.640	0.746	0.786	0.823	0.817
Na	0.239	0.764	0.956	0.999	0.972	0.204	0.597	0.691	0.734	0.708
Si	0.776	0.909	0.980	0.999	0.998	0.763	0.835	0.855	0.874	0.879
S	0.479	0.818	0.980	0.999	0.985	0.468	0.769	0.828	0.843	0.883
Ti	0.851	0.892	0.971	0.990	0.945	0.833	0.867	0.915	0.928	0.893
Mean	**0.701**	**0.875**	**0.978**	**0.998**	**0.985**	**0.686**	**0.811**	**0.856**	**0.877**	**0.871**
										RMSE
Al	1.8E-03	1.4E-01	2.0E-06	1.3E-04	1.1E-03	1.9E-03	1.6E-01	1.4E-05	3.4E-03	3.7E-03
Ca	1.4E-03	6.8E-02	7.5E-05	6.0E-04	2.9E-03	1.5E-03	1.2E-01	5.6E-04	2.2E-02	2.2E-02
Fe	5.8E-04	4.2E-02	6.0E-07	1.1E-04	6.8E-04	6.8E-04	5.0E-02	4.4E-06	1.8E-03	2.1E-03
Mg	3.5E-03	1.3E-01	1.1E-05	2.3E-04	1.6E-03	3.6E-03	2.6E-01	7.8E-05	8.1E-03	8.2E-03
Na	7.6E-03	2.4E-01	7.1E-07	1.2E-04	6.8E-04	7.8E-03	4.0E-01	5.1E-06	2.1E-03	2.2E-03
Si	2.2E-03	9.1E-02	8.1E-05	5.6E-04	3.0E-03	2.3E-03	1.7E-01	5.9E-04	2.3E-02	2.2E-02
S	5.2E-03	1.8E-01	1.2E-06	1.1E-04	9.1E-04	5.7E-03	2.1E-01	7.2E-06	3.1E-03	2.6E-03
Ti	1.5E-03	1.1E-01	6.8E-08	1.5E-04	3.6E-04	1.6E-03	1.3E-01	2.0E-07	4.1E-04	5.0E-04
Mean	**3.0E-03**	**1.2E-01**	**2.1E-05**	**2.5E-04**	**1.4E-03**	**3.2E-03**	**1.9E-01**	**1.6E-04**	**7.9E-03**	**7.9E-03**

Table 7
R^2 and RMSE results for the cross-validation.

	SVM		MLP		Rforest		AdaBoost		XGBoost	
	R^2 mean	R^2 std	R^2 mean	R^2 std	R^2 mean	R^2 std	R^2 mean	R^2 std	R^2 mean	R^2 std
Al	0.800	0.018	0.832	0.013	0.895	0.012	0.908	0.008	0.902	0.004
Ca	0.849	0.004	0.883	0.006	0.898	0.005	0.913	0.004	0.917	0.004
Fe	0.931	0.006	0.948	0.005	0.964	0.004	0.971	0.003	0.966	0.003
Mg	0.630	0.011	0.719	0.010	0.780	0.007	0.813	0.007	0.806	0.007
Na	0.215	0.017	0.550	0.022	0.666	0.014	0.724	0.014	0.691	0.009
Si	0.765	0.006	0.825	0.006	0.845	0.006	0.868	0.007	0.876	0.005
S	0.449	0.028	0.795	0.034	0.836	0.064	0.859	0.054	0.826	0.042
Ti	0.834	0.010	0.872	0.013	0.904	0.010	0.922	0.007	0.891	0.006
Mean	**0.684**	**0.012**	**0.803**	**0.014**	**0.849**	**0.015**	**0.872**	**0.013**	**0.859**	**0.010**
										RMSE
	Mean	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
Al	−2.0E-03	1.3E-04	1.7E-01	7.9E-03	−1.5E-05	1.1E-06	3.6E-03	1.1E-03	3.7E-03	9.3E-04
Ca	−1.5E-03	5.0E-05	1.1E-01	4.9E-03	−5.7E-04	1.7E-05	2.2E-02	5.0E-03	2.2E-02	5.2E-03
Fe	−6.9E-04	7.0E-05	5.2E-02	4.6E-03	−4.5E-06	3.3E-07	2.0E-03	9.0E-04	2.1E-03	6.7E-04
Mg	−3.7E-03	1.2E-04	2.8E-01	4.9E-03	−8.3E-05	2.4E-06	8.4E-03	1.5E-03	8.5E-03	1.5E-03
Na	−7.8E-03	4.3E-04	4.5E-01	2.8E-02	−5.4E-06	2.1E-07	2.1E-03	6.0E-04	2.3E-03	4.9E-04
Si	−2.3E-03	7.4E-05	1.8E-01	5.5E-03	−6.2E-04	1.8E-05	2.3E-02	5.8E-03	2.2E-02	5.2E-03
S	−5.6E-03	1.6E-03	2.0E-01	2.7E-02	−9.2E-06	3.2E-06	2.7E-03	2.0E-03	3.0E-03	1.5E-03
Ti	−1.7E-03	1.0E-04	1.2E-01	6.8E-03	−2.2E-07	1.3E-08	4.3E-04	1.1E-04	5.1E-04	1.2E-04
Mean	**−3.2E-03**	**3.2E-04**	**2.0E-01**	**1.1E-02**	**−1.6E-04**	**5.3E-06**	**8.0E-03**	**2.1E-03**	**8.0E-03**	**2.0E-03**

variables during training also demonstrates their advantage over other algorithms. As a variable proves irrelevant or even disrupts training, its importance is reduced, and it no longer impacts the model's result. In this way, preprocessing is simplified: it is not necessary to make a complex study of which variables will be used in the model since the algorithm manages this issue.

4.3. Synthetic logs

The test phase tested the generalizability of the AdaBoost model, consisting of the generation of synthetic geochemical logs in three wells not used during the training and evaluation phases. These wells presented rocks representative of the formations found in the pre-salt, being excellent candidates for test. Additionally, Well 3 does not have the acquisition of Mg and Na. The test phase objective is to evaluate the model performance in this scenario and to assess whether it can overcome missing data.

Real *versus* synthetic logs graphs are shown in Fig. 13, Fig. 14, and Fig. 15. Significant elements like Ca, Mg, and Si show a good agreement with synthetic data despite the high point dispersion, which lowers R^2. Minor elements like Al, Na, and S have low R^2 due to their low concentration readings being impacted by environmental noise. Their low RMSE verifies this. The exceptions are Fe and Ti, with high R^2 and low RMSE despite their low concentrations in formation.

Fig. 16, Fig. 17, and Fig. 18 show comparisons between real and synthetic logs plotted with depth. The curves are plotted cumulatively from left to right, useful to show the proportion between elements. Visually, the synthetic logs capture the general trends of the formations. To illustrate this, an agglomerative clustering algorithm was used to divide the logs into five different clusters. This process aimed to imitate the geological interpretation of the well by a specialist. The column of clusters is presented next to the geochemical logs. It can be seen that the same interpretations can be obtained in both real and synthetic logs. This is especially true in the separation of rocks rich in Ca and Mg

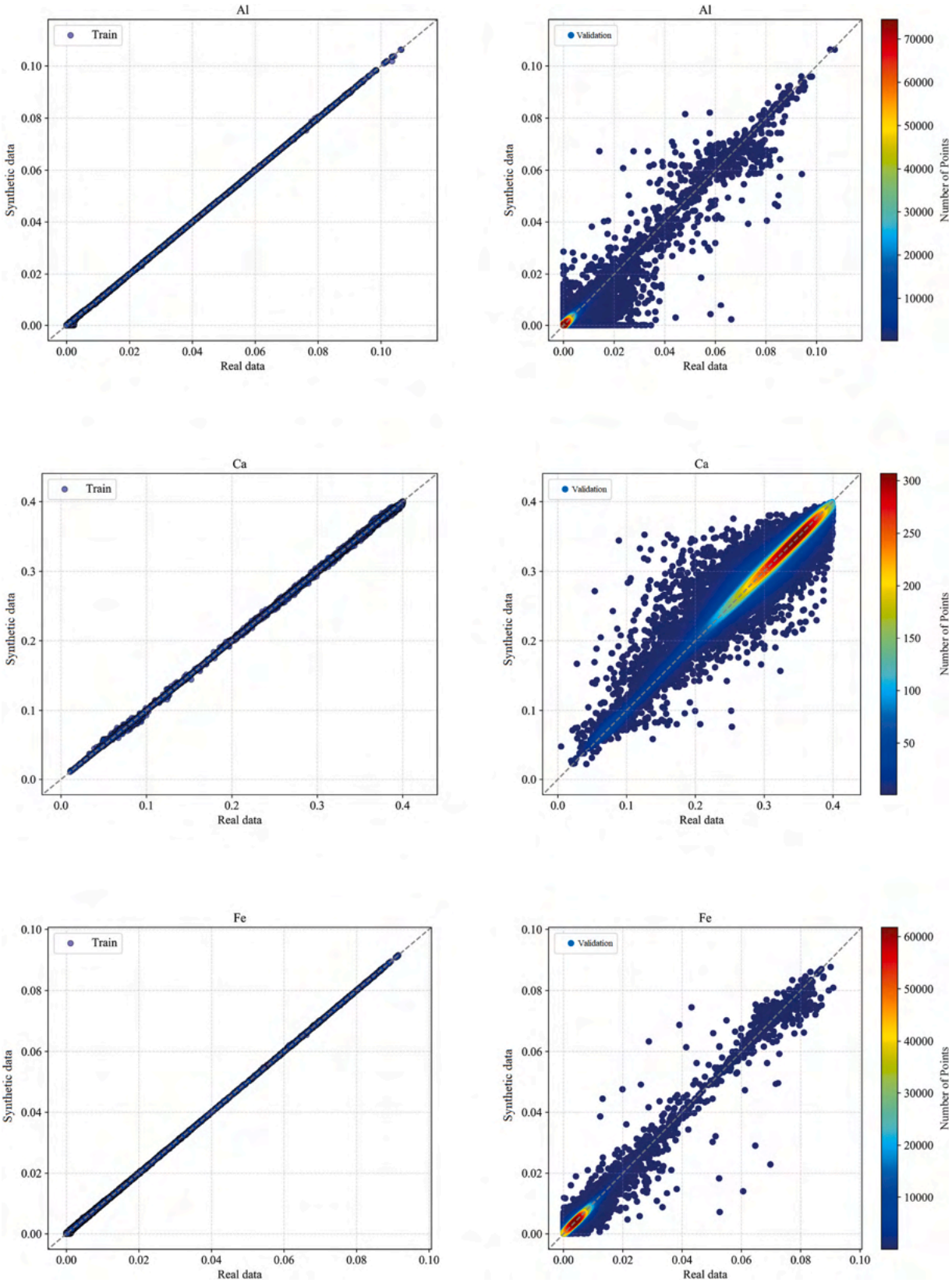


Fig. 9. Real data *versus* synthetic data graphs for AdaBoost: training and validation sets for Al, Ca, and Fe. Due to a large number of points, a density plot for the validation set is shown. Most of the points align with the 1:1 line, corroborating the high R^2 results (Al = 0.919; Ca = 0.917; Fe = 0.976).

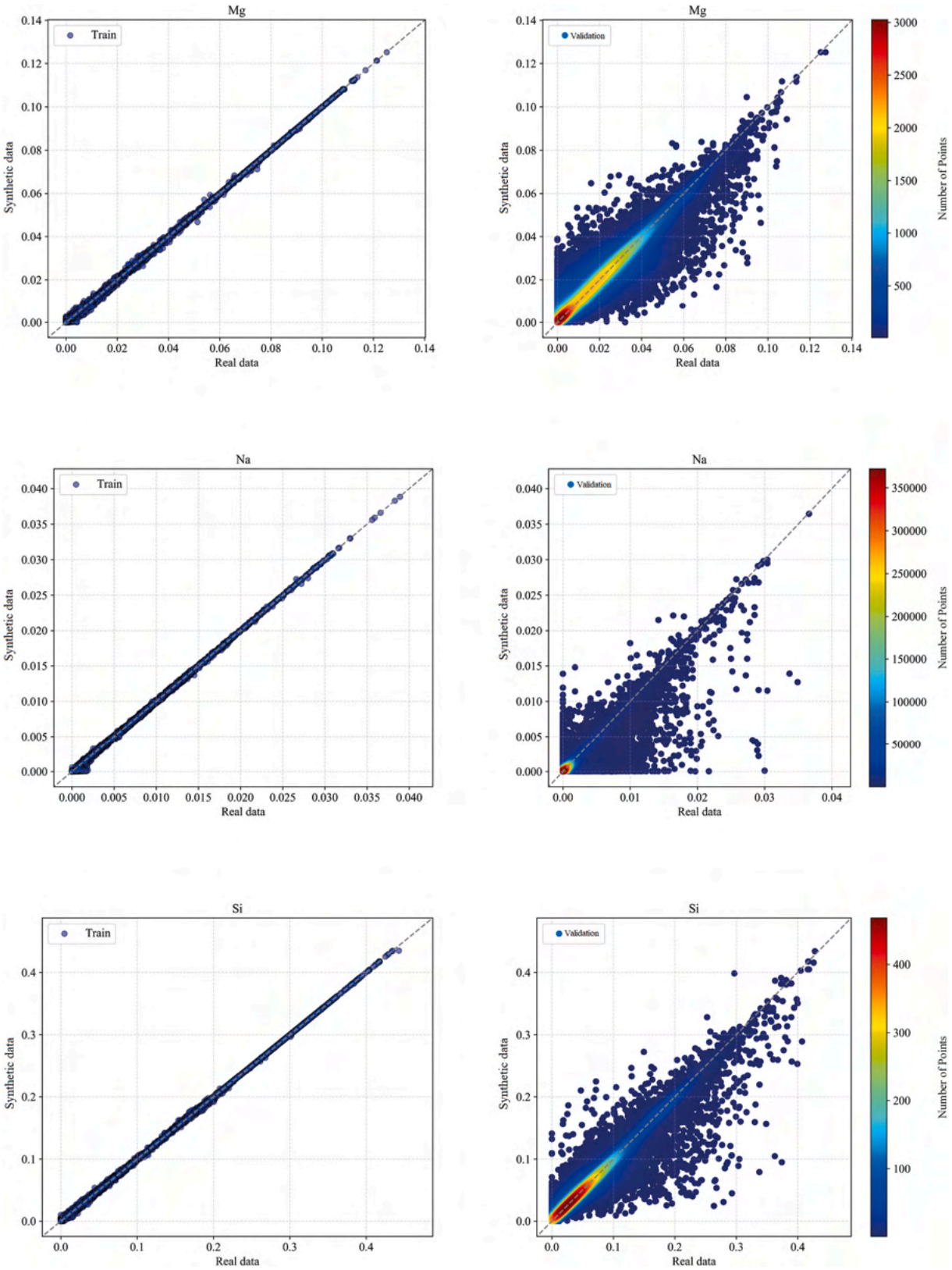


Fig. 10. Real data *versus* synthetic data graphs for AdaBoost: training and validation sets for Mg, Na, and Si. Due to a large number of points, a density plot for the validation set is shown. Most of the points align with the 1:1 line, corroborating the high R^2 results (Mg = 0.823; Na = 0.734; Si = 0.874).

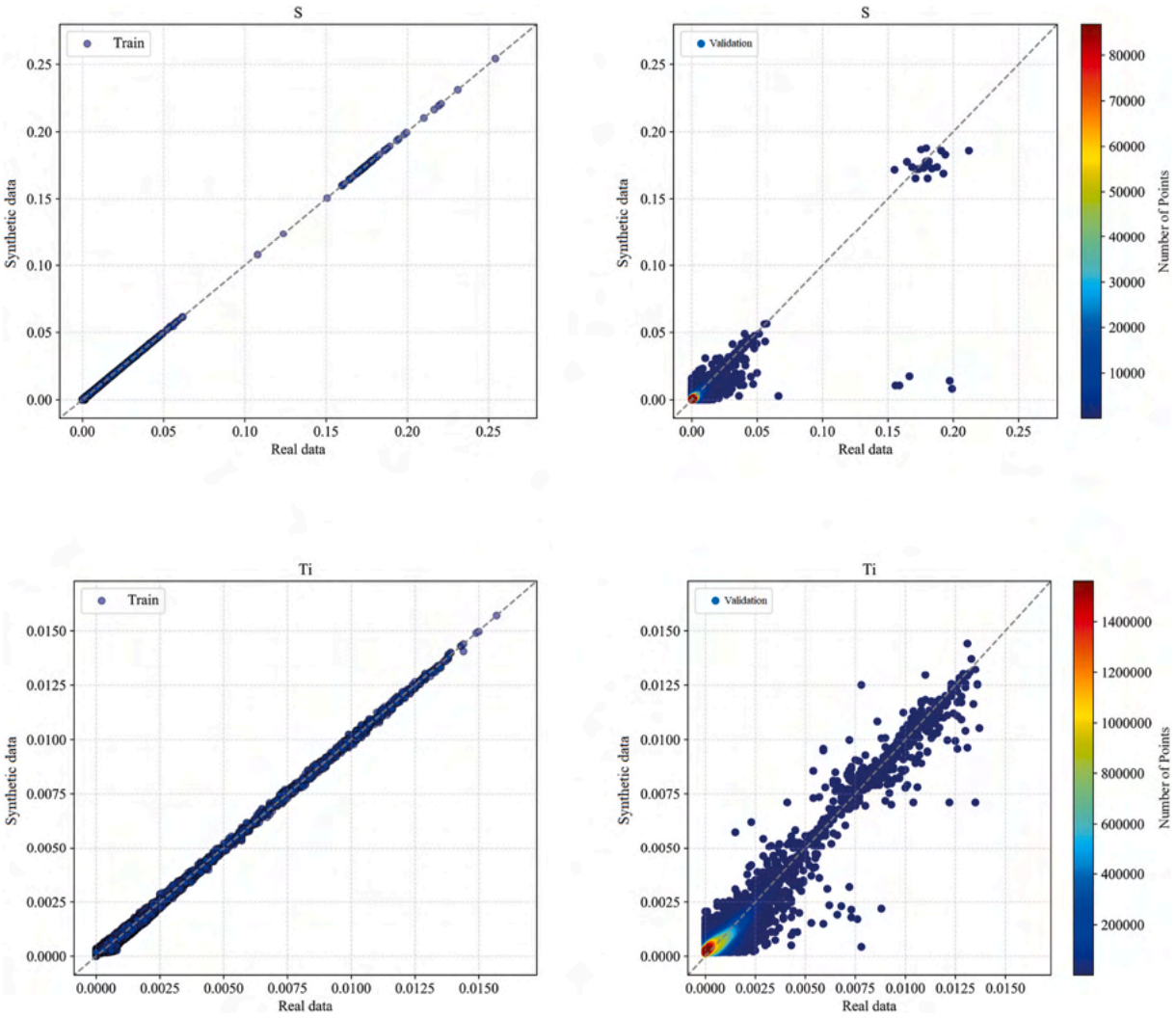


Fig. 11. Real data *versus* synthetic data graphs for AdaBoost: training and validation sets for S and Ti. Due to a large number of points, a density plot for the validation set is also shown. Most of the points align with the 1:1 line, corroborating the high R^2 results (S = 0.843; Ti = 0.928).

Table 8
Standard deviations of the errors obtained in the validation sets for each of the chemical elements.

	Global Error Std (m^3/m^3)	Error Std (m^3/m^3) for low concentrations	Error Std (m^3/m^3) for high concentrations
Al	0.0034	0.0032	0.0086
Ca	0.0211	0.0337	0.0183
Fe	0.0017	0.0015	0.0053
Mg	0.0081	0.0075	0.0124
Na	0.0020	0.0019	0.0065
Si	0.0221	0.0193	0.0526
S	0.0029	0.0025	0.0377
Ti	0.0004	0.0004	0.0011

(carbonates) from rocks with high Al, Fe, and Si content (siliciclastic rocks) and rocks with Al, Fe, Ti, and Na (igneous rocks), fundamental in the pre-salt.

In Well 1 (Fig. 16), there is a carbonate rock at the top of the well, whose proportion of Si decreases with increasing depth. This decrease is observed in both real and synthetic logs. At approximately X160 m, there is an increase in Fe, Al, and Si detected in clusters, indicating siliciclastic rocks. From X175 m, there are intercalations of carbonates with more or less Si and Mg, observed in real and synthetic logs. In the final depths, a further increase in Fe, Al, and Si are observed, now accompanied by Ti, Na, and Mg, in both geochemical logs, indicating the appearance of an igneous formation.

Well 2 (Fig. 17) presents a pattern similar to Well 1, with intercalations of carbonates with more or less Si and Mg, with rocks rich in Al, Fe, and Si. These variations are observed in both real and synthetic logs and marked on the clusters. However, in this well, there are no igneous rocks with Ti and Na, a characteristic shown in both logs.

The first half of Well 3 (Fig. 18) has carbonate rocks rich in Ca, both in real and synthetic logs. However, synthetic logs indicate significant Mg abundance, probably related to diagenetic processes. Al, Fe, and Si show an increase from X150 m, marking the beginning of intercalations between carbonate and siliciclastic rocks, observed in the clusters of both logs. Below X320 m, there is a significant increase in Al, Fe, and Ti in both logs, but the synthetic logs also point to an increase in Na, an element that was not acquired in the real logs, confirming that this is an igneous formation.

It is important to emphasize that the clustering performed was simplified, being done only to facilitate the visualization of the different groups of rocks. Any differences between the real and synthetic logs could be minimized in a more detailed clustering.

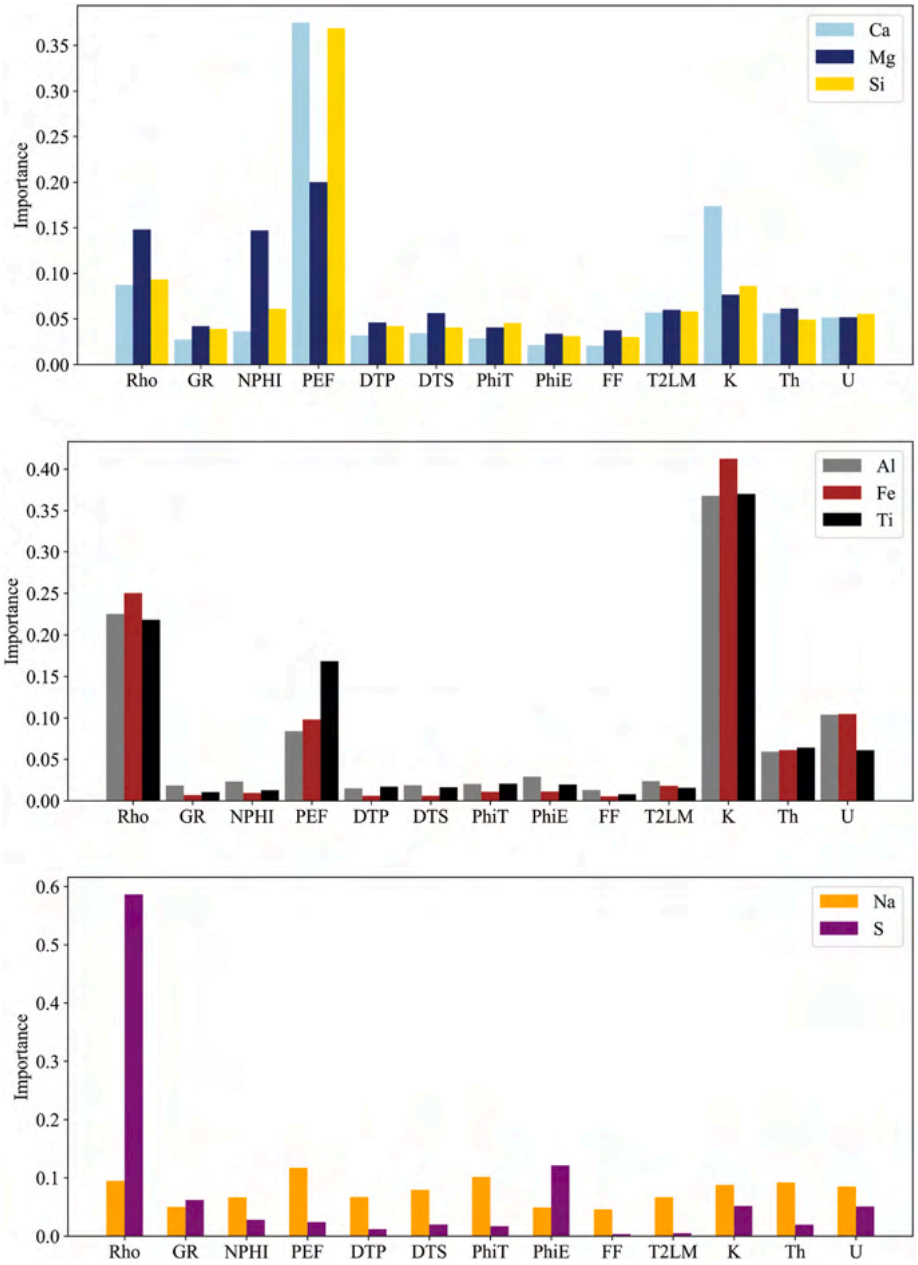


Fig. 12. Relative variable importance of the model input logs for AdaBoost.

5. Discussions

Among all algorithms, the ensemble ones showed the best results. SVM, being a simpler algorithm, was not able to present good results in essential elements, such as Mg, in addition to presenting the worst values of R^2 and RMSE as a whole. MLP was not able to present results as good as tree-based algorithms, in addition to requiring an extra pre-processing step (standardization) and not being able to deal with unimportant input variables. While a more complex neuron structure could generate better results, this would also increase time and computational cost.

When focusing on the ensemble algorithms, the boosting ones were able to achieve better results with fewer decision trees. These better results are because these algorithms actively seek the best solution using the error from previous iterations. As RForest does not do this optimization, it needed a much higher number of trees when compared to the other algorithms. Between AdaBoost and XGBoost, AdaBoost obtained slightly better R^2 and RMSE results. However, AdaBoost has an advantage: its simplicity. It has a smaller number of hyperparameters to be tuned, proving to be a great algorithm to be used in well logs.

To demonstrate the differences in computational cost between the tested algorithms, Table 9 presents the time spent on training the models of the eight chemical elements. The ensemble boosting algorithms are at least twice as fast as the others, confirming the fact that the boosting method generates significant gains in training the models.

The test phase showed that the AdaBoost was able to generate high quality synthetic geochemical logs in wells that did not participate in the training, attesting to its robustness and generalization capacity. The model was also able to add relevant information, such as estimating the presence of Mg and Na in a well that did not have these elements acquired. In addition to not impacting the presence of the other chemical elements, this information made it possible to indicate the presence of diagenesis in the carbonates and confirm the presence of igneous rocks.

Table 10 presents a comparison between the results of the present

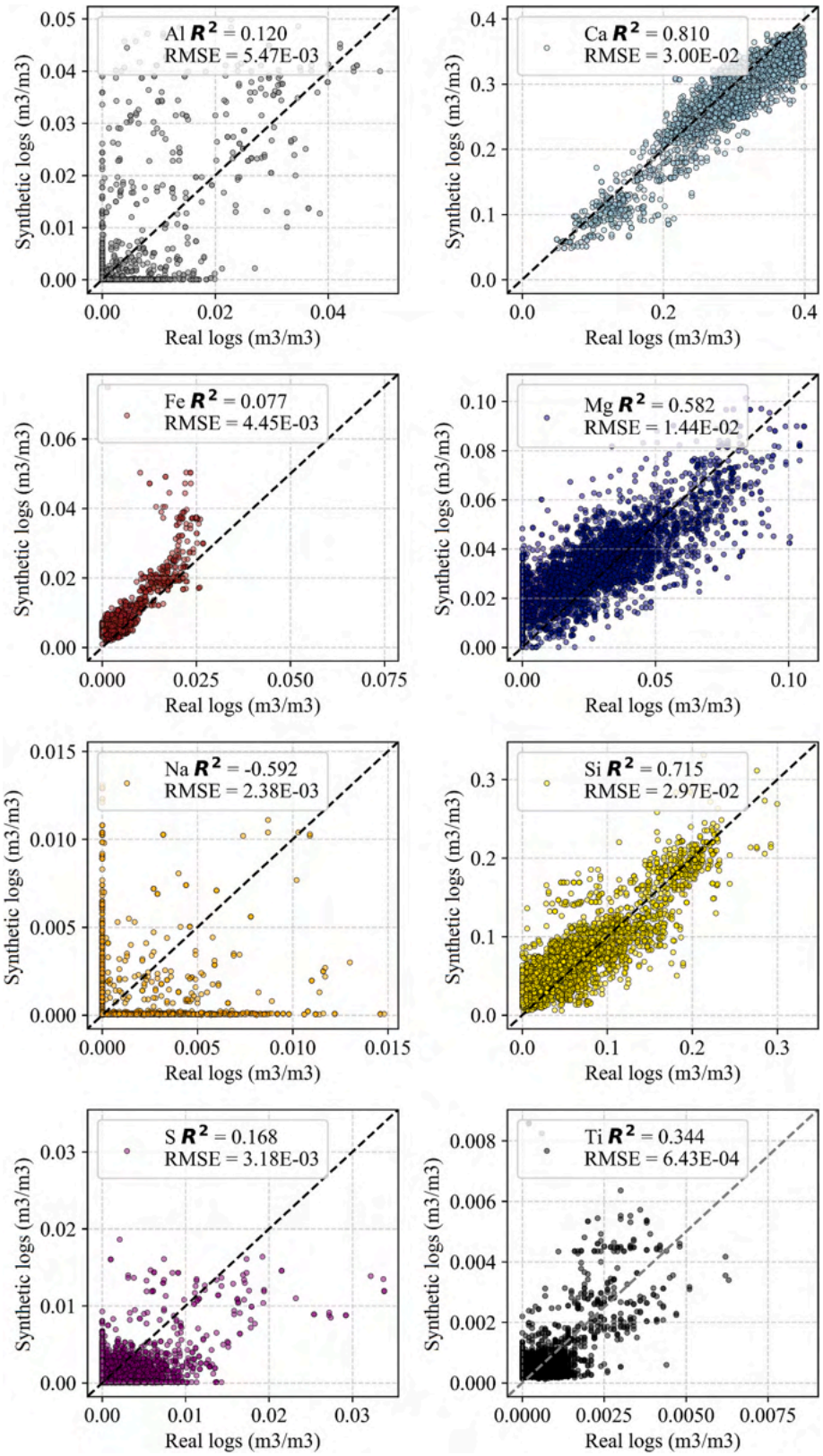


Fig. 13. Real *versus* synthetic logs graphs for AdaBoost: Well 1. Low R^2 is a reflex of low concentrations found in the formation. Probably these readings are affected by environmental noise.

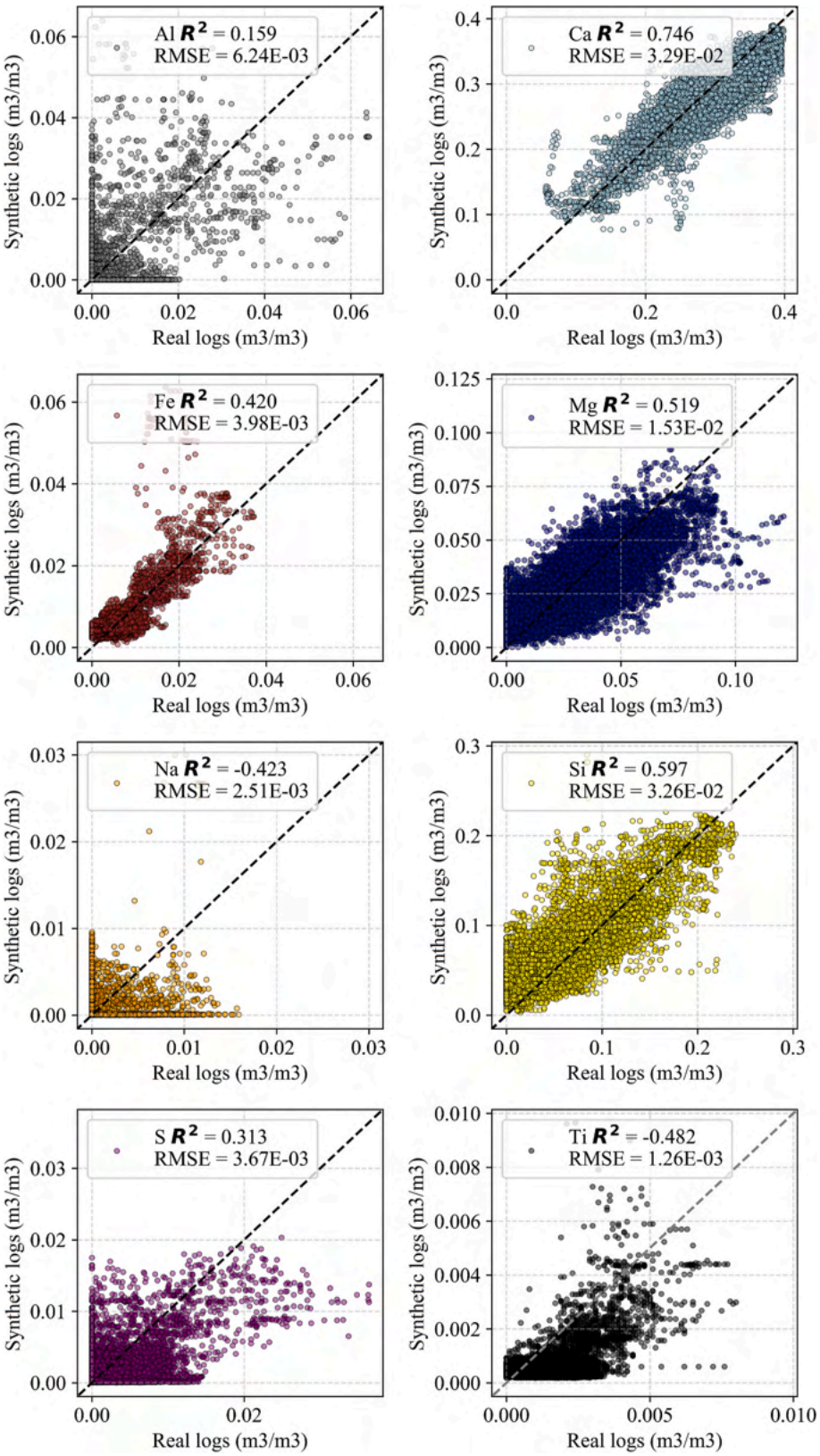


Fig. 14. Real *versus* synthetic logs graphs for AdaBoost: Well 2. Low R^2 is a reflex of low concentrations found in the formation. Probably these readings are affected by environmental noise.

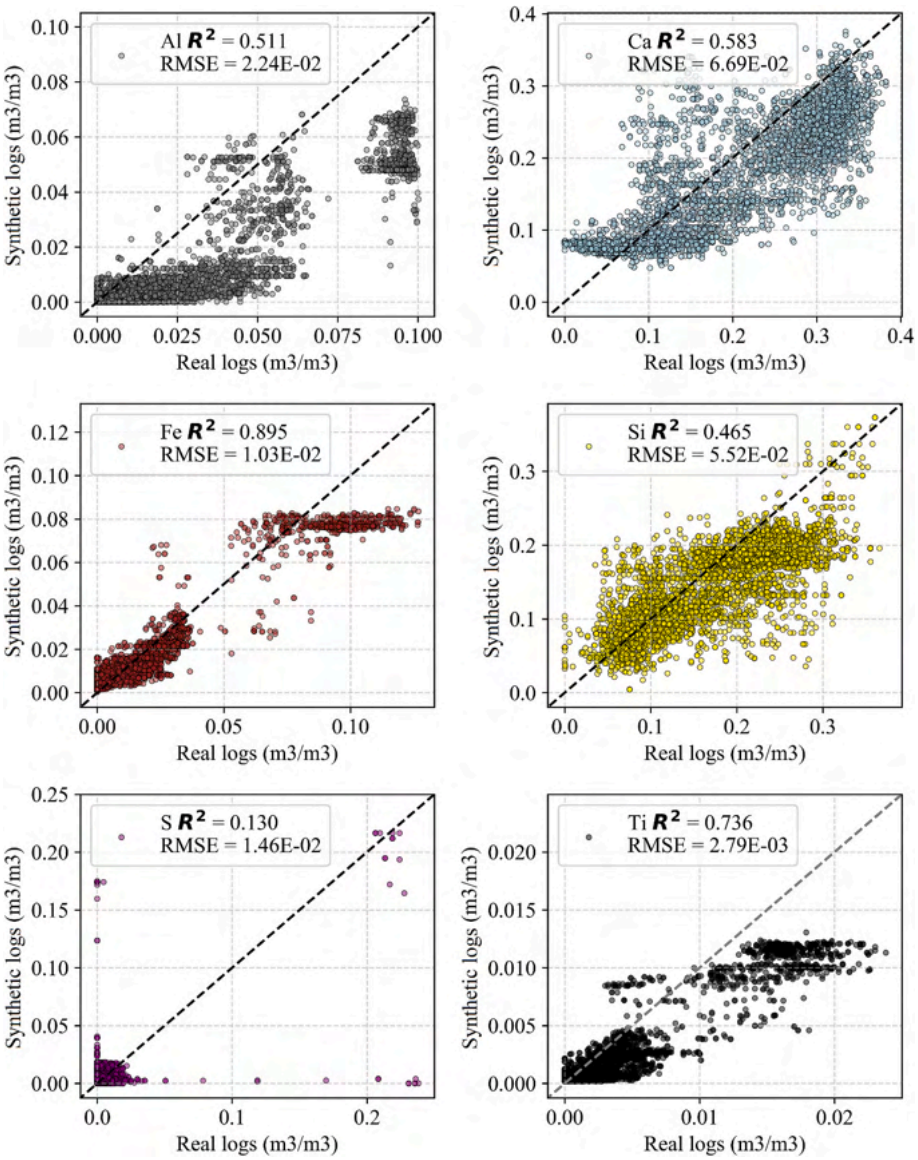


Fig. 15. Real *versus* synthetic logs graphs for AdaBoost: Well 3. Mg and Na were not acquired in this well. The other elements show a good agreement with the synthetic data, despite the high dispersion.

research and other results found in the eight bibliographic references (Chen et al., 2005; Rolon et al., 2009; Bahrpeyma et al., 2013; Korjani et al., 2016; Salehi et al., 2017; Akinnikawe et al., 2018; Akkurt et al., 2018; Zhang et al., 2018). A direct comparison is often difficult since the references focus on generating basic wireline logs and use different logs as input. However, it is possible to observe that the present research showed results as good as or even better than those found in the bibliography. Furthermore, most researches used ANN to generate synthetic logs. No research has used ensemble boosting algorithms, algorithms that are more effective in dealing with well logs.

It is crucial to notice that the models created in the present research are suitable only for the Brazilian pre-salt reservoir. Although one can use the same methodologies presented in this research, the creation of synthetic geochemical logs in different reservoirs must use a trained model with wells drilled in the respective area.

6. Conclusions

Tests developed on five algorithms, namely SVM, MLP, RForest, AdaBoost, and XGBoost, pointed out that AdaBoost models were the most consistent. These algorithms were chosen for their widespread literature and easy use, with open codes accessible to any user. A comparison between different machine learning techniques made the following conclusions possible:

1. AdaBoost showed better results when compared to SVM and MLP. SVM was not able to generate good results, especially for Mg and Na. Concerning MLP, it was possible that a more complex ANN could obtain better results but would increase time and computational cost;
2. Compared to ANN, AdaBoost has a more simplified preprocessing. By using decision trees, AdaBoost does not require data standardization. The sequential addition of trees also allows AdaBoost to calibrate the importance of variables. Therefore, it is not necessary to make a prior study to choose the input data;
3. In comparison to RForest, in addition to obtaining better results, the use of boosting employed by AdaBoost made the best models to be obtained with a much smaller number of decision trees, reducing time and computational cost;
4. Compared to XGBoost, AdaBoost has proved to be a simpler algorithm in hyperparameter tuning. Essentially, AdaBoost only needs to

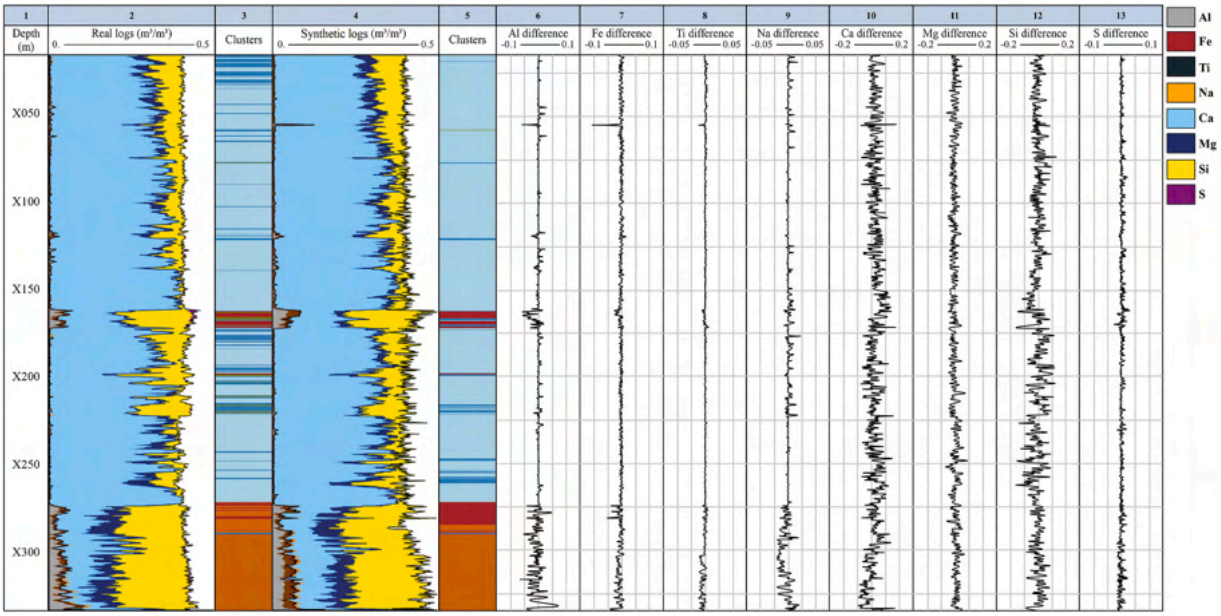


Fig. 16. Color plots (tracks 2 and 4) representing the comparison between real and synthetic logs for AdaBoost for Well 1. The curves are plotted cumulatively from left to right, useful to show the proportion between elements. An evident correspondence is observed, with general trends being respected. The results of the agglomerative clustering (tracks 3 and 5), created to simulate a geological interpretation, confirms this. Tracks 6 to 13 show the difference between real and synthetic data. True depths were omitted due to confidentiality policies. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

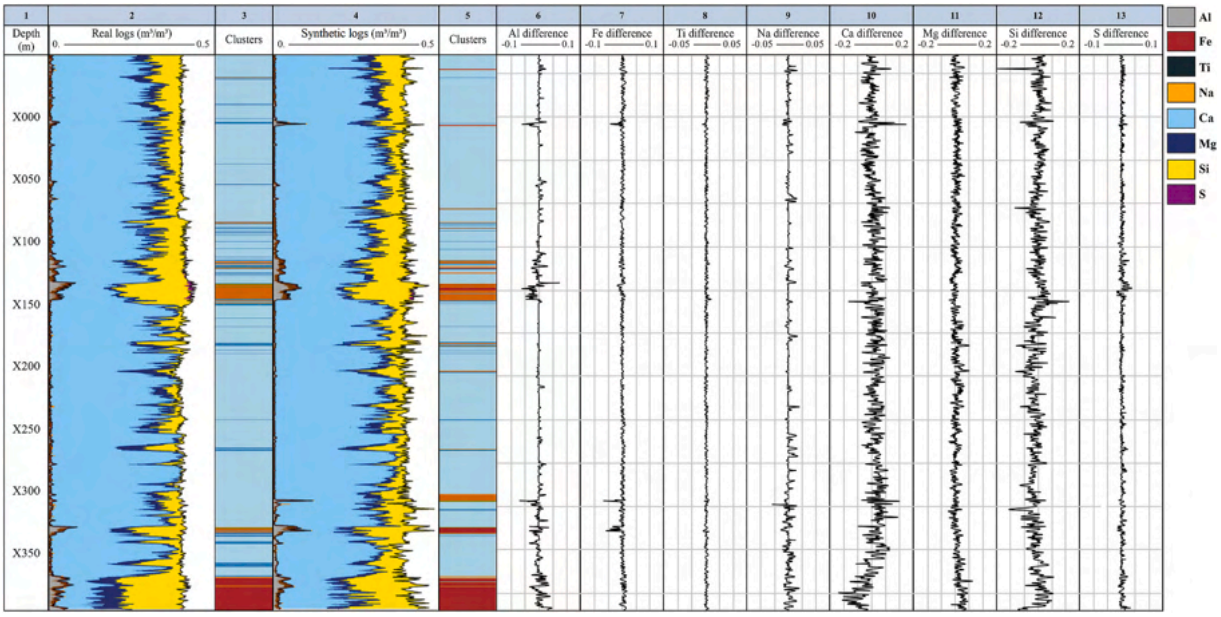


Fig. 17. Color plots (tracks 2 and 4) representing the comparison between real and synthetic logs for AdaBoost for Well 2. The curves are plotted cumulatively from left to right, useful to show the proportion between elements. An evident correspondence is observed, with general trends being respected. The results of the agglomerative clustering (tracks 3 and 5), created to simulate a geological interpretation, confirms this. Tracks 6 to 13 show the difference between real and synthetic data. True depths were omitted due to confidentiality policies. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

adjust the number and characteristics of the decision trees, while XGBoost has a higher number of hyperparameters;

5. The boosting algorithms were at least twice as fast as the others. Between AdaBoost and XGBoost, XGBoost presented itself as the fastest, but the difference was not so significant as to overcome the other advantages presented by AdaBoost.

Analyzing the models generated by AdaBoost, Fe, Ti, Al, and Ca presented the best results, with R^2 above 0.90 both in the validation set and in cross-validation. Si, S, and Mg showed R^2 above 0.80, and Na had R^2 above 0.70. A variable importance analysis showed that density, photoelectric factor, K, Th, U, and neutron porosity were the most relevant logs during model training, revealing that synthetic geochemical logs can still be generated even in more reduced wireline logging

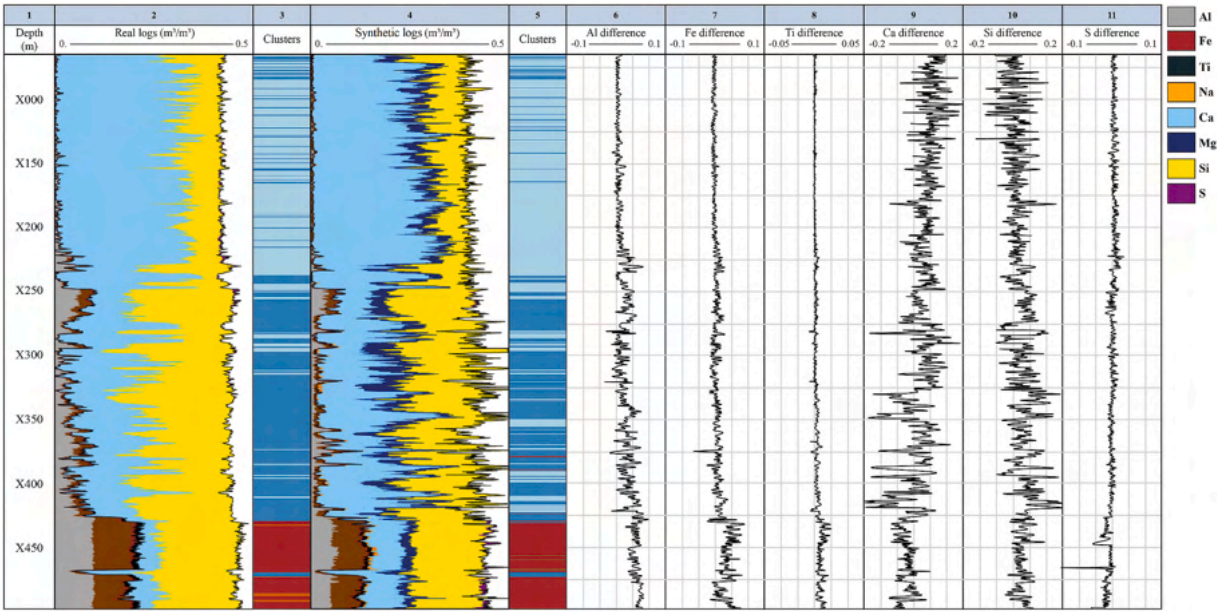


Fig. 18. Color plots (tracks 2 and 4) representing the comparison between real and synthetic logs for AdaBoost for Well 3. The curves are plotted cumulatively from left to right, useful to show the proportion between elements. An evident correspondence is observed, with general trends being respected. The results of the agglomerative clustering (tracks 3 and 5), created to simulate a geological interpretation, confirms this. Tracks 6 to 11 show the difference between real and synthetic data. True depths were omitted due to confidentiality policies. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 9
Time spent on training the models of the eight chemical elements, taken on the same computer, with a 2.80 GHz Intel Core i5-8400 processor, 16 Gb RAM, and NVIDIA GeForce GTX 1050 Ti graphics card.

	SVM	MLP	Rforest	AdaBoost	XGBoost
Time (minutes)	12.49	16.05	39.58	5.50	3.12

operations.

The comparison between real and synthetic geochemical logs in three wells not used in the training and evaluation phases attested to the quality of the AdaBoost models. The synthetic logs were able to reproduce the general trends of the pre-salt formations, such as silicified and dolomitized carbonates, siliciclastic rocks such as sandstones and shales, igneous rocks. An agglomerative clustering showed that both the real

Table 10
Summary of the results achieved in the present research and the results of different references.

Research	Machine learning algorithm used	Synthetic logs generated	Results achieved
Present research	AdaBoost	Al	$R^2 = 0.919$, RMSE = 3.4E-03
		Ca	$R^2 = 0.917$, RMSE = 2.2E-02
		Fe	$R^2 = 0.976$, RMSE = 1.8E-03
		Mg	$R^2 = 0.823$, RMSE = 8.1E-03
		Na	$R^2 = 0.734$, RMSE = 2.1E-03
		Si	$R^2 = 0.874$, RMSE = 2.3E-02
		S	$R^2 = 0.843$, RMSE = 3.1E-03
		Ti	$R^2 = 0.928$, RMSE = 4.1E-04
Chen et al. (2005)	ANN	Resistivity Density Neutron porosity	Mean MSE[a] = 0.02385
Rolon et al. (2009)	ANN	Gamma ray Resistivity Density Neutron porosity	R^2 ranging from 0.85 to 0.95
Bahrpeyma et al. (2013)	Fast fuzzy modeling method	Density Sonic	$R^2 = 0.85$ $R^2 = 0.92$
Korjani et al. (2016)	ANN	Gamma ray Resistivity Density Neutron porosity	Global error = 0.021 Error and correlation/determination coefficients not specified
Salehi et al. (2017)	ANN	Resistivity Density	$R^2 = 0.92018$ $R^2 = 0.97962$
Akinnikawe et al. (2018)	Random forest ANN	Photoelectric factor UCS[b]	ASE[a] = 0.33 ASE = 320.2
Akkurt et al. (2018)	Quantile regression forest	Density Sonic	Error not specified but considered satisfactory
Zhang et al. (2018)	Recurrent neural networks	Resistivity Neutron porosity Sonic	Mean MSE = 0.6083

[a] MSE, ASE: mean and average squared error, respectively.
[b] UCS: Unconfined compressive strength, calculated from well logs and core sample analysis.

and synthetic logs could be grouped in the same clusters, demonstrating that a specialist would interpret them in the same way.

The present research showed that tree-based ensemble boosting algorithms are better suited to deal with the creation of synthetic geochemical logs. A robust database with strict quality control and a training process incorporating a training set, a validation set, and cross-validation are essential to evaluate the model. Not only the results achieved show minimum differences compared to real logs, the simplicity in terms of construction and calibration of the model, and its computational performance, prove that these algorithms should be used more by the oil industry. A well-trained machine learning model can substitute the acquisition of geochemical logs with high confidence, align with wireline costs reduction, and supply engineers and geoscientists with quality data to be used in formation evaluation.

CRediT author statement

Lucas Blanes de Oliveira: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. Cleyton de Carvalho Carneiro: Methodology, Resources, Writing - review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Name	Description	Unit
ABBREVIATIONS		
General		
R^2	Coefficient of determination	-
RMSE	Root mean square error	-
PNG	Pulsed-neutron generator	-
CEC	Cation exchange capacity	cmolc/kg
C/O	Carbon-oxygen ratio	-
Sw	Water saturation	m^3/m^3
Algorithms		
AdaBoost	Adaptive boosting	-
Gboost	Gradient boosting	-
XGBoost	Extreme gradiente boosting	-
ANN	Artificial neural networks	-
MLP	Multilayer perceptron	-
SVM	Support-vector machine	-
Rforest	Random Forest	-
Wireline logs		
Rho	Density	g/cm^3
GR	Gamma-ray	°API
NPHI	Neutron porosity	m^3/m^3
PEF	Photoelectric factor	barns/e
DTP	Compressional wave vagarosity	µs/ft
DTS	Shear wave vagarosity	µs/ft
NMR	Nuclear magnetic resonance	-
PhiT	Total porosity from NMR	m^3/m^3
PhiE	Effective porosity from NMR	m^3/m^3
FF	Free fluid from NMR	m^3/m^3
T2LM	T2 log-mean from NMR	ms
CBW	Clay bound water	m^3/m^3
CHEMICAL ELEMENTS		
K	Potassium from natural gamma-ray spectroscopy	m^3/m^3
Th	Thorium from natural gamma-ray spectroscopy	ppm
U	Uranium from natural gamma-ray spectroscopy	ppm
Al	Aluminum dry weight	m^3/m^3
Ca	Calcium dry weight	m^3/m^3
Fe	Iron dry weight	m^3/m^3
Mg	Magnesium dry weight	m^3/m^3
Na	Sodium dry weight	m^3/m^3
Si	Silicon dry weight	m^3/m^3
S	Sulfur dry weight	m^3/m^3
Ti	Titanium dry weight	m^3/m^3
EQUATIONS		
Geochemical logs		
F	Normalization factor	-
Xi	Ratio of the weight of the respective oxide or carbonate to the weight of element i	-
Yi	Fraction of the gamma-ray spectrum of element i	-
Si	Relative weight fraction detection sensitivity for element i	-
Wi	Absolute elemental weight fraction of element i	-
Machine learning		
N	Number of samples	-
Predi	Prediction of the ith iteraction	-
Erri	Error of the ith iteraction	-
Stvi	Stage valuE of the ith iteraction	-
E	Stabilizer parameter	-

Acknowledgments

The authors are grateful to the Brazilian National Agency for Petroleum, Natural Gas and Biofuels (ANP) and Petrobras for supporting this research. This research was developed with the RDI Group Integrated Technology of Rock and Fluid Analysis (InTRA), and conducted under the Postgraduate Program from the Naval and Oceanic Engineering Department of the Escola Politécnica, Universidade de São Paulo.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.petrol.2020.108080.

References

Ajayi, O., Torres-Verdin, C., Preeg, W.E., 2015. Petrophysical interpretation of LWD, neutron-induced gamma-ray spectroscopy measurements: an inversion-based approach. Petrophysics 56 (4), 358–378.

Akinnikawe, O., Lyne, S., Roberts, J., 2018. Synthetic well log generation using machine learning techniques. Unconventional Resources Technology Conference 16. https://doi.org/10.15530/urtec-2018-2877021.

Akkurt, R., Conroy, T.T., Psaila, D., Paxton, A., Low, J., Spaans, P., 2018. Accelerating and enhancing petrophysical analysis with machine learning: a case study of an automated system for well log outlier detection and reconstruction. In: SPWLA 59th Annual Logging Symposium, vol. 25. Retrieved from. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85051657222&partnerID=40&md5=943a97cff3a91836bc328b998e0e1eda.

Al-Bulushi, N.I., King, P.R., Blunt, M.J., Kraaijveld, M., 2012. Artificial neural networks workflow and its application in the petroleum industry. Neural Comput. Appl. 21 (3), 409–421. https://doi.org/10.1007/s00521-010-0501-6.

Alizadeh, B., Najjari, S., Kadkhodaie-Ilkhchi, A., 2012. Artificial neural network modeling and cluster analysis for organic facies and burial history estimation using well log data: a case study of the South Pars Gas Field, Persian Gulf, Iran. Comput. Geosci. 45, 261–269.

Aminian, K., Ameri, S., Oyerokun, A., Thomas, B., 2003. Prediction of flow units and permeability using artificial neural networks. In: SPE Western Regional/AAPG Pacific Section Joint Meeting, vol. 7. Society of Petroleum Engineers, Long Beach, California, USA.

Anderson, R.N., Dove, R.E., Boglia, C., Silver, L.T., James, E.W., Chappell, B.W., 1988. Elemental and mineralogical analyses using geochemical logs from the Cajon Pass scientific drillhole, California, and their preliminary comparison with core analyses. Geophys. Res. Lett. 15 (9), 969–972.

Archie, G.E., 1942. The electrical resistivity log as an aid in determining some reservoir characteristics. Transactions of the AIME 146 (1), 54–62. https://doi.org/10.2118/942054-G.

Asfahani, J., Ahmad, Z., Ghani, B.A., 2018. Self organizing map neural networks approach for lithologic interpretation of nuclear and electrical well logs in basaltic environment, Southern Syria. Appl. Radiat. Isot. 137, 50–55. https://doi.org/10.1016/j.apradiso.2018.03.008.

Bahrpeyma, F., Golchin, B., Cranganu, C., 2013. Fast fuzzy modeling method to estimate missing logsin hydrocarbon reservoirs. J. Petrol. Sci. Eng. 112, 310–321. https://doi.org/10.1016/j.petrol.2013.11.019.

Belozerov, B., Bukhanov, N., Egorov, D., Zakirov, A., Osmonalieva, O., 2018. Automatic well log analysis across Priobskoe Field using machine learning methods. In: SPE

Russian Petroleum Technology Conference, vol. 21. Society of Petroleum Engineers, Moscow, RU.
Beltrao, R.L., Sombra, C., Lage, A., Fagundes Netto, J., Henriques, C., 2009. Challenges and new technologies for the development of the pre-salt cluster, Santos Basin, Brazil. Offshore Technology Conference 11. https://doi.org/10.4043/otc-19880-ms.
Bestagini, P., Lipari, V., Tubaro, S., 2017. A machine learning approach to facies classification using well logs. SEG Technical Program 5. https://doi.org/10.1190/segam2017-17729805.1.
Bishop, C.M., 2006. Pattern recognition and machine learning. In: Jordan, M., Kleinberg, J., Scholkopf, B. (Eds.), *Information Science and Statistics* (First). https://doi.org/10.1128/AAC.03728-14.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.
Breiman, L., Cutler, A., Liaw, A., Wiener, M., 2018. Breiman and Cutler's Random Forests for Classification and Regression. Retrieved from. https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.
Chen, D., Quirein, J., Simth Jr., H., Hamid, S., Grable, J., 2005. Neural network ensemble selection using a multi-objective genetic algorithm in processing pulsed neutron data. Petrophysics 46 (5), 323–334.
Chen, T., Guestrin, C., 2016. XGBoost: a scalable tree boosting system. 22nd KDD Conference on Knowledge Discovery and Data Mining 785–794. https://doi.org/10.1145/2939672.2939785.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. https://doi.org/10.1007/s40031-014-0099-7.
Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Adv. Neural Inf. Process. Syst. 1, 155–161.
Ellis, D.V., Singer, J.M., 2007. Well Logging for Earth Scientists, Second. Springer.
Flaum, C., Pirie, G., 1981. Determination of lithology from induced gamma ray spectroscopy. In: SPWLA 23nd Annual Logging Symposium, vol. 16. Society of Petrophysicists and Well-Log Analysts, Mexico City, Mexico.
Freedman, R., Herron, S., Anand, V., Herron, M., May, D., Rose, D., 2015. New method for determining mineralogy and matrix properties from elemental chemistry measured by gamma ray spectroscopy logging tools. SPE Reservoir Eval. Eng. 18, 599–608. https://doi.org/10.2118/170722-pa, 04.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139. https://doi.org/10.1006/jcss.1997.1504.
Freund, Y., Schapire, R.E., 1999. Large margin classification using the perceptron algorithm. Mach. Learn. 37 (3), 277–296. https://doi.org/10.1023/A:1007662407062.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 39.
Galford, J.E., Quirein, J.A., Shannon, S., Truax, J.A., Witkowsky, J., 2009. Field test results of a new neutron induced gamma ray spectroscopy geochemical logging tool. SPE Annual Technical Conference and Exhibition 22. https://doi.org/10.2118/123992-ms.
Garcia, A.P., Heidari, Z., Rostami, A., 2017a. Improved assessment of hydrocarbon saturation in mixed-wet rocks with complex pore structure. In: SPWLA 58th Annual Logging Symposium, vol. 16. Retrieved from. https://www.onepetro.org/conference-paper/SPWLA-2017-LL?sort=&start=0&q=Mixed-Wet+Rocks+With+Complex+Pore+Structure&from_year=&peer_reviewed=&published_between=&fromSearchResults=true&to_year=&rows=10#.
Garcia, A.P., Jagadisan, A., Rostami, A., Heidari, Z., 2017b. A new resistivity-based model for improved hydrocarbon saturation assessment in clay-rich formations using quantitative clay network geometry and reck fabric. In: SPWLA 58th Annual Logging Symposium, vol. 16. Society of Petrophysicists and Well-Log Analysts, Oklahoma City, Oklahoma, USA.
Gilchrist Jr., W.A., Quirein, J.A., Boutemy, Y.L., Tabanou, J.R., 1982. Application of gamma ray spectroscopy to formation evaluation. In: SPWLA 23rd Annual Logging Symposium, vol. 28. Society of Petrophysicists and Well-Log Analysts, Corpus Christi, Texas, USA.
Gonzalez, J., Lewis, R., Hemingway, J., Grau, J., Rylander, E., Schmitt, R., 2013. Determination of formation organic carbon content using a new neutron-induced gamma ray spectroscopy service that directly measures carbon. SPWLA 54th Annual Logging Symposium 15. https://doi.org/10.1190/urtec2013-112.
Guarido, M., 2018. Machine Learning in Geoscience: Facies Classification with Features Engineering, Clustering, and Gradient Boosting Trees, vol. 30 (Calgary, Alberta, CA).
Hamada, G.M., Ahmed, E., Chaw Y, N., 2018. Artificial neural network (ANN) prediction of porosity and water saturation of shaly sandstone reservoirs. Adv. Appl. Sci. Res. 9 (2), 26–31.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference and prediction. In: Springer Series in Statistics (Second). Springer.
Herron, M.M., 1986. Mineralogy from geochemical well logging. Clay Clay Miner. 34 (2), 204–213. https://doi.org/10.1346/ccmn.1986.0340211.
Hertzog, R., Colson, L., Seeman, B., O'Brien, M., Scott, H., McKeon, D., et al., 1989. Geochemical logging with spectrometry tools. SPE Form. Eval. 4, 153–162. https://doi.org/10.2118/16792-pa, 02.
Ho, T.K., 1995. Random decision forests. In: Proceedings of the International Conference on Document Analysis and Recognition. ICDAR, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994.
Hoeink, T., Zambrano, C., 2017. Shale discrimination with machine learning methods. In: 51st US Rock Mechanics/Geomechanics Symposium, 6. Retrieved from. https://www.onepetro.org/conference-paper/ARMA-2017-0769.
Korjani, M., Popa, A., Grijalva, E., Cassidy, S., Ershaghi, I., 2016. A new approach to reservoir characterization using deep learning neural networks. SPE Western Regional Meeting 15. https://doi.org/10.2118/180359-MS.
Kuhn, M., Johnson, K., 2013. Applied predictive modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3.
Macdonald, R., Hardman, D., Sprague, R., Meridji, Y., Mudjiono, W., Galford, J., et al., 2010. Using elemental geochemisty to improve sandstone reservoir characterization : a case study from the Unayzah A interval of Saudi Arabia. SPWLA 52th Annual Logging Symposium 52 (5), 16. Retrieved from. https://www.onepetro.org/journal-paper/SPWLA-2011-v52n5a2.
McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133. https://doi.org/10.1111/j.1365-2710.2009.01107.x.
Moreira, J.L.P., Valdetaro, C., Gil, J.A., Machado, M.A.P., 2007. Bacia de Santos. Bol. Geociencias Petrobras 15 (2), 531–549.
Nashawi, I.S., Malallah, A., 2009. Improved electrofacies characterization and permeability predictions in sandstone reservoirs using a data mining and expert system approach. Petrophysics 50 (3), 250–268.
Negara, A., Jin, G., Agrawal, G., 2016. Enhancing rock property prediction from conventional well logs using machine learning technique - case studies of conventional and unconventional reservoirs. In: Abu Dhabi International Petroleum Exhibition & Conference, vol. 13. https://doi.org/10.2118/183106-ms.
North, R.J., 1987. Through-casing reservoir evaluation using gamma ray spectroscopy. In: SPE California Regional Meeting, vols. 329–342. Society of Petroleum Engineers, Ventura, California, USA.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://doi.org/10.5555/1953048.2078195.
Pemper, R., Sommer, A., Guo, P., Jacobi, D., Longo, J., Bliven, S., et al., 2006. A new pulsed neutron sonde for derivation of formation lithology and mineralogy. SPE Annual Technical Conference and Exhibition 13. https://doi.org/10.2523/102770-ms.
Quirein, J., Vigne, J. La, Chapman, S., 1987. Enhancements to the pulsed neutron gamma ray spectroscopy interpretation. In: SPWLA 28th Annual Logging Symposium, vol. 23. Society of Petrophysicists and Well-Log Analysts, London, England.
Radtke, R., Lorente, M., Adolph, B., Berheide, M., Fricke, S., Grau, J., et al., 2012. A new capture and inelastic spectroscopy tool takes geochemical logging to the next level. In: SPWLA 53rd Annual Logging Symposium, vol. 16. Retrieved from. http://69.18.148.110/~/media/Files/technical_papers/misc/spwla/2012_spwla_spectroscopy_tool.pdf.
Rolon, L., Mohaghegh, S.D., Ameri, S., Gaskari, R., McDaniel, B., 2009. Using artificial neural networks to generate synthetic well logs. J. Nat. Gas Sci. Eng. 1, 118–133. https://doi.org/10.1016/j.jngse.2009.08.003.
Scikit Learn documentation, 2019. Scikit Learn documentation - Support-vector Regressor. (Accessed 25 April 2020).
Scikit Learn documentation, 2019. Scikit Learn documentation - Random Forest Regressor. (Accessed 25 April 2020).
Scikit Learn documentation, 2015. Scikit Learn documentation - Multi-layer Perceptron. (Accessed 5 June 2019).
Scikit Learn documentation, 2013. Scikit Learn documentation - AdaBoost. (Accessed 5 June 2019).
Rosenblatt, F., 1957. The Perceptron - a perceiving and recognizing automaton. Cornell Aeronautical Laboratory, Buffalo, NY. Report 85.
Salehi, M.M., Rahmati, M., Karimnezhad, M., Omidvar, P., 2017. Estimation of the non records logs from existing logs using artificial neural networks. Egyptian Journal of Petroleum 26 (4), 957–968. https://doi.org/10.1016/j.ejpe.2016.11.002.
Shabab, M., Jin, G., Negara, A., Agrawal, G., 2016. New data-driven method for predicting formation permeability using conventional well logs and limited core data. SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition 10. https://doi.org/10.2118/182826-ms.
Ulloa, J.M., Chaparro, D., Lara, S., Arango, S., Mendez, F., Alarcon, N., Gade, S., 2016. An innovative cased-hole, oil-saturation method of utilizing excess carbon analysis of pulsed neutron measurements in a siliciclastic cenozoic formation, Los Llanos Basin, Colombia. In: SPWLA 57th Annual Logging Symposium, 12. Reykjavik, Iceland. Society of Petrophysicists and Well-Log Analysts.
Verma, A.K., Cheadle, B.A., Routray, A., Mohanty, W.K., Mansinha, L., 2012. Porosity and permeability estimation using neural network approach from well log data. In: GeoConvention Vision Conference, vol. 6. Retrieved from. http://www.searchanddiscovery.com/documents/2014/41276verma/ndx_verma.
Wendemutht, A., 1995. Learning the unlearnable. J. Phys. Math. Gen. 28, 5423–5436. https://doi.org/10.1088/0305-4470/28/18/030.
Westaway, P., Hertzog, R., Plasek, R.E., 1983. Neutron-induced gamma ray spectroscopy for reservoir analysis. Soc. Petrol. Eng. J. 23, 553–564. https://doi.org/10.2118/9461-pa, 03.
Witten, I.H., Frank, E., 2005. Data mining: practical machine learning tools and techniques. In: Gray, J. (Ed.), The Morgan Kaufmann Series in Data Management Systems. Second, 0120884070, 9780120884070.

Wu, P.-Y., Jain, V., Kulkarni, M.S., Abubakar, A., 2018. Machine learning–based method for automated well-log processing and interpretation. SEG International Exposition and 88th Annual Meeting 2041–2045. https://doi.org/10.1190/segam2018-2996973.1.

XGBoost documentation, 2015. XGBoost documentation. (Accessed 5 June 2019).

Ye, S., Rabiller, P., 2005. Automated electrofacies ordering. Petrophysics 46 (6), 409–423.

Zhang, D., Chen, Y., Meng, J., 2018. Synthetic well logs generation via recurrent neural networks. Petrol. Explor. Dev. 45 (4), 629–639. https://doi.org/10.1016/S1876-3804(18)30068-5.

Zhao, J., Chen, H., Yin, L., Li, N., 2017. Mineral inversion for element capture spectroscopy logging based on optimization theory. J. Geophys. Eng. 14 (6), 1430–1436. https://doi.org/10.1088/1742-2140/aa7bfa.

Zhao, T., Verma, S., Devegowda, D., Jayaram, V., 2015. TOC estimation in the Barnett Shale from triple combo logs using support vector machine. Society of Exploration Geophysicists 85th Annual Meeting 791–795. https://doi.org/10.1190/segam2015-5922788.1.

Anexo B – Artigo publicado: modelos minerais através de aprendizado escalonado

Energy and AI 3 (2021) 100050

Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai

Stepped machine learning for the development of mineral models: Concepts and applications in the pre-salt reservoir carbonate rocks

Lucas Abreu Blanes de Oliveira [a,b,*], Luiz Felipe Niedermaier Custódio [b], Thais Bortotti Fagundes [b], Carina Ulsen [b], Cleyton de Carvalho Carneiro [b]

[a] Petrobras – Petróleo Brasileiro S.A., Avenida Henrique Valadares 28, Centro, Rio de Janeiro, 20231-030 Rio de Janeiro, Brazil
[b] Universidade de São Paulo, USP, Escola Politécnica, Praça Narciso de Andrade, Vila Mathias, Santos, São Paulo 11013-560, Brazil

HIGHLIGHTS

- Use of machine learning for the development of mineralogical models.
- Innovative stepped training to estimate mineral concentrations.
- Reliable mineralogical model to be used in formation evaluation.

ARTICLE INFO

Article history:
Received 2 December 2020
Received in revised form 18 January 2021
Accepted 20 January 2021
Available online 26 January 2021

ABSTRACT

Understanding rock mineralogy is essential for formation evaluation, improving the calculation of porosity and hydrocarbon saturation. The primary method to obtain the mineralogy from a well is by applying a model to the geochemical tool's chemical elements. However, creating a mineralogical model presents challenges such as the minerals' chemical composition and the decision to include a mineral in the model. The traditional application of machine learning can make mineral models less realistic since conventional training is developed based on a set of minerals with different occurrences, lowering some minerals' representativeness. The present research proposes the stepped machine learning (SML), a stepped way to use machine learning to create a mineralogical model from chemical and mineralogical data. A database was assembled with the elemental concentration obtained with XRF analyses and the mineral concentrations obtained with XRD analyses. The chemical elements were Al, Ca, Fe, K, Mg, Mn, Na, Si, and Ti. The minerals were calcite, dolomite, quartz, clays, K-feldspar, plagioclase, and pyroxene. Four algorithms were tested: MLP, GAN, Random Forest, and XGBoost, with XGBoost showing the best results. SML was applied, where a mineral model results are used to train a subsequent model. SML allowed for a significant improvement in some models, notably to clays with an increase in R^2 from 0.597 to 0.853, quartz an increase from 0.673 to 0.869, and calcite, from 0.758 to 0.862. A decrease in the mean squared error of these minerals' models was also observed. The model was applied to the geochemical logs from three wells drilled in the Brazilian pre-salt, and the results were compared with XRD analyzes. The SML model was able to honor the mineral concentrations for different rocks. It is demonstrated that the integration between machine learning tools and geological knowledge in SML was crucial for creating a representative mineralogical model.

Introduction

Understanding the mineralogy of the rock matrix is essential for formation evaluation. A reliable mineralogical model can significantly improve the calculation of porosity, hydrocarbon saturation, and volume of clay [10], also providing information on the cation exchange capacity [17] and total organic carbon [13]. In addition to the direct application in formation evaluation, knowledge of mineralogy can be used to assist acid-fracturing operations [21] and to monitor the variation of oil/water contact during the production of a hydrocarbon reservoir [24,28,30].

The primary way to obtain a mineralogical model of a reservoir is by using the geochemical data acquired in well logging operations. The geochemical tool detects chemical elements present in the rock matrix, exploring the spectrum of gamma-rays generated by the interaction between neutrons emitted by the tool and these elements [8]. Concentrations of Al, Ca, Fe, K, Mg, Na, and Si, can be obtained and represents the main minerals present in reservoir rocks. However, as it is an indirect measure, the concentrations of chemical elements and their respective allocations in different minerals must be marked out with information obtained from the reservoir's rock samples.

According to Herron et al. [18], elemental concentrations of rock samples can be measured by X-ray fluorescence (XRF) or by inductively coupled plasma atomic emission spectrometry (ICP-AES), while mineralogy can be analyzed by X-ray diffraction (XRD). The construction

* Corresponding author at: Petrobras – Petróleo Brasileiro S.A., Avenida Henrique Valadares 28, Centro, Rio de Janeiro, 20231-030 Rio de Janeiro, Brazil.
E-mail address: lucas.oliveira@usp.br (L.A.B. de Oliveira).

https://doi.org/10.1016/j.egyai.2021.100050

Fig. 1. Location of the Brazilian pre-salt and the area of study.

of a mineralogical model would then consist of a mathematical inversion that would transform chemical elements' concentrations into mineralogical concentrations [2,9,32]. However, creating this mathematical model is not a simple task; as pointed out by Freedman et al. [10], the minerals chemical composition, their respective endpoints, and the subjective assumption of a specific mineral's presence or absence (such as clay or trace minerals) must serve as input for the model.

In light of the difficulties presented, this research seeks to develop a mineralogical model for the carbonate, siliciclastic, and igneous rocks of the Brazilian pre-salt using machine learning algorithms. The main advantage of machine learning is that it does not need prior knowledge of the minerals' composition, reducing the model's subjectivity. The chemical concentrations of rock samples obtained through XRF measurements will be used as inputs for the model, and the mineralogy obtained in XRD will be the outputs. The model will include the main minerals observed in the pre-salt formations. Additionally, an innovative training sequence proposed in this research, called stepped machine learning (SML), will improve the models' predictive capacity. In the end, the model generated by SML will be used in the mineralogical interpretation of a well drilled in the pre-salt.

The present work is organized as follows:

- *Conceptual background:* the main aspects of the Brazilian pre-salt rocks are discussed, followed by a review of the XRF and XRD analyses and the machine learning algorithms used.
- *Methodology:* the workflow is presented from the XRF and XRD analyzes, the construction of the database, and the process of training and evaluation of the machine learning models.
- *Results:* the results of the trained models are shown, as well as the importance of each of the input variables during training. Then, the application of the models to estimate the mineralogy of rocks in a well is presented.
- *Discussion and conclusion:* the final configuration of the SML models and the main conclusions are summarized.

Conceptual background

Pre-salt rocks

Located in Santos Basin (Fig. 1), the Brazilian pre-salt consists of the rocks found stratigraphically below the thick layer of evaporites of the Ariri Formation of the Neo-Aptian age [23]. This sequence includes the Barra Velha Formation's stromatolites, the grainstones, wackestones, and packstones from the Itapema Formation, sandstones, pelites shales from the Piçarras Formation, and the igneous rocks of the Camboriú Formation. Intercalations with basalts can also occur. The sedimentary rocks were deposited between the Hauterivian and the Eo-Aptian (between approximately 110–130 million years), while the crystalline basement has Precambrian age.

XRF and XRD analyses

The X-ray fluorescence (XRF) is characterized by the emission of secondary X-rays after a particular material is bombarded by high-energy X (primary emission). XRF is widely used to analyze materials' chemical composition since the radiation energy emitted will depend on the atoms/electron transition from a given material. When exposed to high-energy electromagnetic radiation in short-wavelength X-rays, atoms may be ionized. During ionization, electrons present in the inner orbits of atoms are ejected, making them unstable and causing electrons from outer orbits to replace them. The change between orbits causes photons' emission, whose energy equals the energy difference between the two orbits, an intrinsic characteristic of the atom involved. The re-emission of radiation in energy other than that initially absorbed is called fluorescence [27]. The radiation emitted by the material after exposure to X-rays is acquired in a spectrum, where energy peaks represent different chemical elements. The intensity of these peaks allows the quantification of the elements that compose the measured material.

X-ray diffraction (XRD) corresponds to one of the main techniques for crystalline materials' microstructural characterization, responsible for providing qualitative and quantitative results of the crystalline phases present [6]. It represents the phenomenon of interaction between the incident X-ray beam and the electrons of the atoms that make up a material, related to coherent scattering. If the atoms that generate this scattering are arranged systematically, as in a crystalline structure, the phase relationships between the scattering become periodic, and the X-ray diffraction phenomenon is observed at various angles of incidence. The various planes of a crystalline structure have different atoms and electrons' densities, making the diffracted intensities specific to their different planes. Thus, each crystalline compound has a unique diffractometric pattern, and its correct identification occurs through the angular positions and the relative intensities of the diffracted beams. However, this will only occur when the incident X-ray satisfies Bragg's Law [20]. The diffractometric pattern, through angular positions and relative intensities, is compared with reference standards, available in updated databases and made available by the ICDD - International Center for Diffraction Data and ICSD - Inorganic Crystal Structure Database, obtaining crystallographic information and physical properties of the crystalline compounds (qualitative analysis).

Machine learning algorithms

Several machine learning algorithms can be found in the literature, using the most different training strategies. However, some algorithms stand out for their intense application in recent years. Among them, it is possible to mention artificial neural networks (ANN) and tree-based algorithms.

ANN can be considered the most popular machine learning algorithm today, vastly applied in deep learning problems. It uses artificial neurons at its core. Neurons receive input data and calculate outputs using activation functions capable of capturing non-linearity. The most common ANN structure is Multilayer Perceptron (MLP, [11,26,29]). It is a feedforward arrangement containing layers of multiple neurons. The training process updates the weights assigned to each neuron connection, using back-propagation to converge to the minimum error.

Generative Adversarial Networks (GAN, [14]) are a class of ANN where two networks compete against each other in a zero-sum game. A discriminator receives input data from two sources: the actual data and data created by a generator. At each iteration, the discriminator's

function is to identify which information is real and synthetic created by the generator. The function of the generator is to provide information to the discriminator to deceive it. Initially, the discriminator can easily distinguish the synthetic data. However, throughout training, both the generator and the discriminator evolve to represent the actual data. In the end, the discriminator finds it very difficult to identify the real data of the synthetic provided by the generator, which has learned the patterns of the input data and can imitate them perfectly. The discriminator and generator structures can vary, from an MLP to a convolutional or recurrent network. GANs have been applied to deep learning problems, mainly involving images. Aggarwal et al. [1] proposed a GAN architecture for regression problems, comparing its performance with other algorithms. They identified that GANs performed better on data with heteroscedasticity. In data with homoscedasticity, tree-based algorithms showed better results.

Random Forest is an ensemble algorithm that uses bagging (bootstrap aggregation) and several decision trees [3,4,19]. The technique known as bagging divides the data into subsets and fits several decision trees, combining their predictions. This combination reduces the variance that would be observed in a single tree. To further reduce variance, Random Forest randomly selects each subset's input variables, making the model less susceptible to overfitting.

Gradient Boosting uses the union of several simple predictors, also called weak learners [12]. These weak learners are trained in sequence so that a predictor learns using the previous one's error. In this way, the sequence of weak learners forms a robust committee [15]. In Gradient Boosting, the weak learner used is the decision tree. A tree fits the residuals from the previous iteration during training, converging to the smallest possible error. This sequence can be understood as a gradient descent method and can be summarized as follows [22]:

Considering y as the instances, $F(x)$ as the predicted value of y, and $h(x)$ as the decision tree.

- Calculate the first approximation for $F_0(x)$, such as the average of y.
- For i iterations, do:
- Calculate the residuals $r_i = y - F_{i-1}(x)$.
- Fit a tree $h_i(x)$ to the residuals.
- Add the tree to the model, as $F_i(x) = F_{i-1}(x) + h_i(x)$.
- Repeat the process until the error becomes constant or until the end of iterations.

An advantage of Gradient Boosting is the ability to calculate the importance of variables during training. It can be calculated by counting the number of times a variable is used to split a node in a decision tree. This importance can be used to provide insight into how each variable was used in the training stage.

Several metrics can be used to assess the quality of machine learning models. In regression problems, most commons are the coefficient of determination (R^2) and the mean squared error (MSE). The R^2 expresses how much of the dependent variable's variance can be predicted by the independent variable and can vary between 0 and 1. In machine learning, this comparison is made between real and modeled data, and a R^2 close to 1 indicates that the algorithm can reproduce the real data with high confidence.

The mathematical definition of R^2 is given by Eqs. 1, 2, and 3:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{1}$$

$$SS_{res} = \sum_i \left(y_i - f_i\right)^2 \tag{2}$$

$$SS_{tot} = \sum_i \left(y_i - \bar{y}\right)^2 \tag{3}$$

Where SS_{res} is the total sum of squares of residuals, expressed by the square of the difference between the modeled (f_i) and real (y_i) values; and SS_{tot} is the total sum of squares, expressed by the square of the difference between the real data and its mean ($\bar{y}$). If the R^2 between the real and modeled data is 1, the model perfectly predicts the real data. If R^2 is 0, the model prediction is as good as the mean of the data. Negative R^2 indicates that the model is worse than the mean.

The MSE indicates the average error of the model, calculated by Equation 4:

$$MSE = \frac{1}{n} \sum_i \left(y_i - f_i\right)^2 \tag{4}$$

Where n is the total number of instances. The MSE will always be positive, and values close to 0 indicate a high-quality model. Unlike R^2, the MSE does not have a maximum value.

Table 1
Summary of the samples collected in the pre-salt formations.

Formation	Number of samples	Lithotypes	Percentage
Barra Velha	1059	Carbonate, igneous	77.0%
Itapema	288	Carbonate, siliciclastic, igneous	20.8%
Piçarras	9	Siliciclastic, igneous	0.7%
Camboriú	20	Igneous	1.5%

Methodology

The database consists of analysis from 1,376 instances of rock samples collected from 22 wells drilled in the Brazilian pre-salt. From this total, 77 samples are cuttings collected during well drilling, 199 are rotary sidewall cores collected during wireline operations, and 1,100 are core plugs. Table 1 shows the number of samples collected in each of the pre-salt formations and the lithotypes observed in each formation.

First, the samples were sent to the XRF and XRD laboratories, where chemical and mineralogical composition measurements were carried out. The results were concatenated in a database, where the chemical composition served as input, and the mineralogy served as output for the training process. Four algorithms were tested and compared to choose the best mineral model. The training and evaluation of the model followed the best practices for creating a machine learning model. The complete workflow is shown in Fig. 2 and is detailed below.

XRF and XRD laboratory analyses

The XRF analyzes were performed at the Technological Characterization Laboratory (LCT) of the University of São Paulo and by the company SGS Geosol. For cuttings and core plugs, the chemical determination was carried out by the quantitative method, using specific calibration curves for carbonate rocks, with the quantification of CaO, MgO, SiO_2, Al_2O_3, Fe_2O_3, Na_2O, K_2O, P_2O_5, TiO_2, SrO, MnO and S. The determination of loss on ignition (LOI) was carried out at 1,020 °C for two hours. The analyzes were performed on the Zetium equipment from the PANalytical company. The preparation and analysis of the samples consisted of fused beads with the addition of lithium tetraborate.

For XRD analysis, the sample preparation comprised the pulverization of 20 g of material to below 0.04 mm, using a planetary ball mill, and manual backloading pressing in proper sample holders. The diffractograms collection was carried out in a Panalytical X'Pert equipment with a detector sensitive to the X'Celerator position with copper tube (Cu Kα radiation - λ = 1.54186 Å with Ni filter) at 45 kV × 40 mA, 2θ ranging from 2 to 70°, 0.02° 2θ step size, time per step of 0.20 s, and total collection time of 53 s.

Database preparation

Chemical (XRF) and mineralogical (XRD) data were joined in a database, where the chemical elements were considered input variables, and the mineralogical concentrations were output variables. Table 2 summarizes the main characteristics of the variables, as well as their statistics.

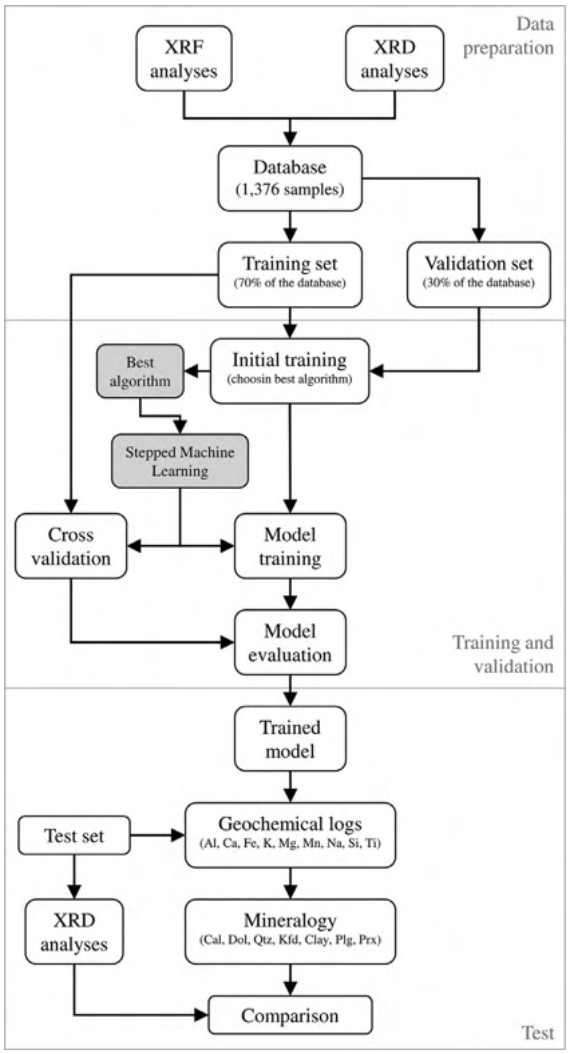


Fig. 2. Workflow used in the present research.

The elements selected as input were those acquired in wireline operations. The minerals selected as output were the most abundant in the pre-salt rocks, based on their maximum, average, and non-zero concentrations. Therefore, the minerals selected as output were calcite, dolomite, quartz, K-feldspar, clays, plagioclase, and pyroxene. These minerals are the main components of the carbonate, siliciclastic, and igneous rocks observed in the pre-salt. Minerals such as dawsonite, pyrite, barite, fluorite, hematite, and ilmenite are also present in low concentrations and were not included as output variables.

The chemical formula of these minerals, shown in Table 2, demonstrates how challenging it is to build a mineralogical model without machine learning. Minerals such as dolomite, K-feldspar, plagioclase, and pyroxene have complex compositions, with variations in the proportions of Ca–Mg, K–Al, Na–Ca, and Si–Al, making it challenging to choose endpoints. Despite having a more stable composition, calcite and quartz are subject to minor contamination by Fe, Mg, and Mn. Clays have complex compositions, which can vary according to the source rock and the environment in which the clay was deposited, making the choice of a single chemical formula impossible.

Fig. 3 presents the correlation between variables. It is possible to observe the high positive correlation between Ca and LOI, resulting from the burning of carbon dioxide from the carbonates, a phenomenon observed and described by studies such as that of Dean [7] and Heiri et al. [16]. The positive correlation between Al, Fe, K, and Ti reflects the chemical composition of siliciclastic and igneous rocks present in the pre-salt. The significant negative correlation between Ca and Si marks the passage of carbonate rocks from the Barra Velha and Itapema Formations to the siliciclastic and igneous rocks from the Piçarras and Camboriú Formations.

Concerning minerals, it is possible to observe the negative correlation between calcite and dolomite, a reflection of diagenetic dolomitization processes common in carbonates. The high positive correlation in plagioclase and pyroxene also stands out, indicating that these minerals probably compose the same rock types. Some obvious correlations are observed, such as Ca and calcite, Mg and dolomite, Si and quartz, and K and K-feldspar. The high positive correlation between Fe and Ti with plagioclase and pyroxene was not expected since these elements are not observed in these minerals' chemical composition. The clays did not significantly correlate with any chemical element, reinforcing the challenge of obtaining models that take them into account.

Table 2
Statistics of the input and output variables used in the training steps.

Input variables

Name	Description	Unit	Min	Max	Mean	Std	Non-zero
Al	Aluminum dry weight	g/g	0.000	0.088	0.006	0.013	99.3%
Ca	Calcium dry weight	g/g	0.003	0.402	0.247	0.078	100.0%
Fe	Iron dry weight	g/g	0.000	0.099	0.006	0.013	100.0%
K	Potassium dry weight	g/g	0.000	0.090	0.004	0.009	97.1%
Mg	Magnesium dry weight	g/g	0.000	0.179	0.045	0.026	100.0%
Mn	Manganese dry weight	g/g	0.000	0.003	0.000	0.000	97.4%
Na	Sodium dry weight	g/g	0.000	0.128	0.003	0.006	99.3%
Si	Silicon dry weight	g/g	0.000	0.421	0.089	0.071	100.0%
Ti	Titanium dry weight	g/g	0.000	0.021	0.001	0.002	95.6%
LOI	Loss on ignition	g/g	0.008	0.479	0.349	0.094	100.0%

Output variables

Name	Description	Formula	Unit	Min	Max	Mean	Std	Non-zero
Cal	Calcite	$CaCO_3$	g/g	0.00	1.00	0.48	0.25	93.8%
Dol	Dolomite	$CaMg(CO_3)_2$	g/g	0.00	1.00	0.29	0.21	94.7%
Qtz	Quartz	SiO_2	g/g	0.00	0.98	0.16	0.14	93.4%
Kfd	K-Feldspar	$KAlSi_3O_8$	g/g	0.00	0.64	0.02	0.05	26.7%
Clay	Clay	Variable	g/g	0.00	0.48	0.03	0.07	25.1%
Plg	Plagioclase	$NaCa(Si_3AlO_8)_2$	g/g	0.00	0.70	0.01	0.07	5.1%
Prx	Piroxene	$CaMgSi_2O_6$	g/g	0.00	0.34	0.00	0.03	3.3%
Others	Other minerals	Variable	g/g	0.00	0.31	0.01	0.02	23.5%

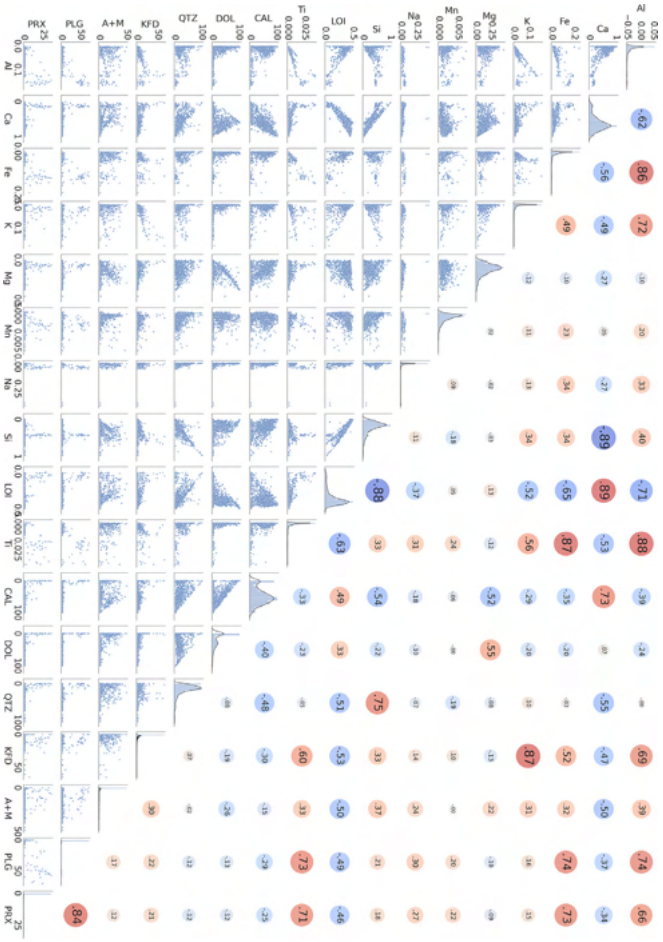


Fig. 3. Correlation between variables, presented in the form of graphs and R^2. Warm colors indicate a positive correlation, and cold colors indicate a negative correlation. Histograms of the variables are diagonally in the figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Model training

The database was randomly divided into a training set and a validation set. From the total data, 70% were separated for training (963 instances), and 30% were for validation (413 instances). As the concentrations of minerals and chemical elements vary between zero and one, data normalization/standardization was unnecessary. In addition to the minerals presented in Table 2, models for LOI, carbonates (calcite + dolomite), and pyroxene + plagioclase were also trained. LOI has a high correlation with carbonates and organic matter and is not acquired in wireline operations. The carbonate fraction of the rock, represented by calcite + dolomite, can be more easily fitted by the model since it correlates to LOI. As plagioclase and pyroxene have low abundance in the database and their concentrations have a high positive correlation, the combination of these minerals can present better results in training.

Before definitively creating the models, four machine learning algorithms were tested to assess the most suitable application in the database: Multilayer Perceptron (MLP), Generative Adversarial Networks (GAN), Random Forest, and Gradient Boosting. This choice aimed to cover both artificial neural networks and tree-based algorithms. The MLP used had the input layer, a hidden layer with 100 neurons, and the output layer. The GAN's structure was the one proposed by Aggarwal et al. [1], created to deal specifically with regression problems. Random Forest had 100 decision trees without depth restriction or limiting the number of instances for each leaf and each split. The Gradient Boosting algorithm used was the one proposed by Chen & Guestrin [5], known as XGBoost.

As the first test's objective was to evaluate each of the algorithms' initial performance and then choose the best one to be used in the research, no hyperparameter tuning was performed. The algorithms were trained using the training set, generating models for the LOI and minerals. The trained algorithms estimated mineralogical concentrations using the input data from the validation set, and these concentrations were compared with the actual data using the R^2 metric. These results are shown in Table 3.

It is possible to observe that XGBoost presented the best results for most minerals, reflecting the highest mean R^2. This result is in line with those obtained by Oliveira & Carneiro [25], who concluded that tree-based boosting algorithms present better results when applied to well data than MLP and Random Forest. GAN had the worst results, also in agreement with the results of Aggarwal et al. [1], which observed that GANs did not perform as well as algorithms like XGBoost in regression problems using tabular data. With these results, it was decided to use the XGBoost algorithm for the creation of mineralogical models. A fine-tuning found XGBoost's best hyperparameters, presented in Table 4. A detailed description of each hyperparameter can be found in XGBoost documentation [31].

The models were trained using the training set and then applied to the validation set. The results were evaluated using R^2. In addition to the application in the validation set, 5-fold cross-validation was applied to the training sets. In this cross-validation, the training set was divided into five parts, where four were used to fit the model and one as validation. The process was repeated five times until all folds have been used for validation. The validation set and cross-validation allow an unbiased evaluation of the machine learning models' quality and robustness and reduce overfitting.

Stepped machine learning

As the mineralogical composition of a rock is a closed system, where the sum of the minerals is equal to one, it is expected that prior knowledge of the concentration of one mineral will impact the estimation of the others. Thus, the present research proposes the use of an innovative learning technique called stepped machine learning (SML) to estimate the mineralogical composition of rocks.

SML consists of sequential training of mineral models, adding one mineral's concentrations to the input variables during the others' training. However, it is not desirable to include all the previous results for training a given mineral model since it can carry the previous models' errors. Therefore, it is necessary to use the integration of geological and machine learning knowledge to choose which minerals will serve as input for subsequent models.

The step-by-step of SML was as follows:

LOI model: since LOI is not acquired by any wireline tool, a model was first created using chemical element concentrations. In a real well situation, these concentrations would come from the geochemical tool. Estimates of the other minerals were made using only the chemical elements for comparison with subsequent models.

Carbonate model: LOI was added to the inputs, and a model for carbonates (calcite + dolomite) was created. As LOI has a high correlation with carbonates, it is expected that this model has better results than the previous ones. Again, estimation of the other minerals was made using the chemical elements + LOI to be compared with later steps.

Mineral model: a new input was created by concatenating chemical elements + LOI + carbonates to create the other minerals' models. As carbonates represent the most important fraction of the pre-salt mineralogy, their inclusion in the model inputs can significantly improve other minerals' estimation.

Table 3
Results of the initial training phase of the machine learning algorithms.

Algorithm	LOI	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Plg + Prx	Mean
MLP	0.836	0.817	0.711	0.682	0.663	0.703	0.409	0.954	0.614	0.864	0.725
GAN	0.816	0.625	0.308	0.457	0.313	0.143	0.225	0.187	0.285	0.317	0.367
Random Forest	0.811	0.759	0.695	0.668	0.652	0.650	0.528	0.918	0.760	0.765	0.721
XGBoost	0.895	0.811	0.774	0.791	0.656	0.802	0.602	0.959	0.641	0.833	0.776

Table 4
Best XGBoost hyperparameters for each model trained.

Hyperparameter name	LOI	Carb	Cal	Dol	Qtz	Kfd	Clay	Plg	Prx	Prx+Plg
Tree max depth	25	15	25	25	15	15	50	15	15	15
Tree minimum child weight	10	1	5	5	5	1	10	1	10	10
Gamma	0	0.01	0.001	0.01	0	0.01	0	0	0.001	0.01
Subsample ratio	1	0.7	0.7	0.7	0.6	0.6	0.7	0.8	1	1
Columns sampling by tree	0.7	1	1	0.8	1	1	1	1	0.7	0.7
Alpha	0	1	0	0	0.1	0	0.01	1	1	0.1
Learning rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	1

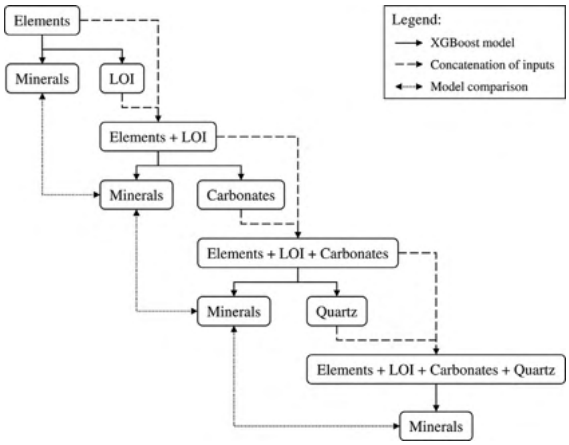


Fig. 4. SML workflow proposed in the present research. Solid arrows indicate the application of the XGBoost model to estimate minerals and LOI, while dashed arrows indicate the concatenation of the modeled output with the previous inputs. Dotted arrows indicate a comparison between estimates from different steps.

Clay model: due to the inherent challenges of estimating clay minerals, quartz was added to the input data, and a new model was created for clay. It is expected that a model created using the chemical elements + LOI + carbonates + quartz will be able to aggregate not only information on the composition but also insights into the clay's depositional environment, achieving better results. As in previous training, models for other minerals were created to be compared with previous results.

Fig. 4 presents a summary of the stepped learning proposed in this research.

Evaluation and test

R^2 and MSE values of the validation and cross-validation with and without SML were compared to determine the best models. Uncertainty of the models, measured from the standard deviation of the error, was also used to evaluate the SML. In the test phase, the trained models were applied to three wells drilled in the pre-salt, using their geochemical logs. The mineralogy acquired was compared to XRD analyses and used to make considerations about the wells' rocks.

Results

Stepped machine learning models

The results of the SML are presented in Table 5. As expected, a clear improvement in R^2 and a decrease in MSE is observed as new information is added to the models' input data.

First, the LOI model's quality is observed, with R^2 of 0.962 and MSE of 0.00028 in the validation data and R^2 of 0.956 and MSE of 0.00041 in cross-validation. These results give greater confidence in the use of LOI in subsequent models. The inclusion of LOI improves the carbonate model, both in the validation and cross-validation results. The inclusion of carbonates in the input data generates a substantial improvement in calcite, dolomite, and quartz predictions. For quartz, R^2 increases from values below 0.70 to above 0.85, and MSE decreases to less than half for both validation and cross-validation. Finally, the inclusion of quartz in the inputs generated a significant improvement in the clay model, with R^2 greater than 0.80 and MSE decreasing to almost a third for validation and cross-validation. The inclusion of quartz did not significantly improve the calcite and dolomite models.

The K-feldspar, plagioclase, and pyroxene models showed R^2 and MSE relatively constant, regardless of SML. The pyroxene model showed the worst results overall, with R^2 below 0.80. However, the model for plagioclase + pyroxene using chemical elements, LOI, and carbonates as input presented better results than pyroxene alone.

Table 6 shows the evolution of the standard deviation of errors in the models generated by stepped learning. The standard deviation of the error can be understood as a measure of the models' uncertainty since it indicates how far the predicted value is from the real value. In general, uncertainty decreases as more variables are added to the input data. This phenomenon is expected since it is a reflection of the increase in R^2 of the models. Fig. 5 shows real *versus* modeled data for LOI, carbonates, K-feldspar, and plagioclase. Figs. 6 and 7 show real *versus* modeled data for the other minerals before and after the SML application to illustrate the model's improvement. A significant decrease in the points' dispersion is observed.

Variable importance

Fig. 8 shows the importance of the variables during the training of the mineralogical model. The importance presented is related to the latest models created for each mineral and LOI; therefore, some amounts will be null (e.g., LOI in the K-feldspar model, since it was not used).

Table 5
R^2 results of the SML models.

	Validation set				Cross-validation			
	Elements	Elements + LOI	Elements + LOI + Carb	Elements + LOI + Carb + Qtz	Elements	Elements + LOI	Elements + LOI + Carb	Elements + LOI + Carb + Qtz
					R^2			
LOI	0.962	-	-	-	0.956	-	-	-
Carb	0.815	0.845	-	-	0.821	0.852	-	-
Cal	0.758	0.780	0.862	0.874	0.755	0.768	0.840	0.850
Dol	0.781	0.816	0.823	0.850	0.730	0.769	0.774	0.790
Qtz	0.673	0.663	0.869	-	0.672	0.683	0.850	-
Kfd	0.815	0.822	0.811	0.815	0.708	0.701	0.708	0.698
Clay	0.597	0.662	0.698	0.853	0.637	0.687	0.717	0.814
Plg	0.960	0.964	0.962	0.962	0.636	0.637	0.653	0.644
Prx	0.632	0.638	0.674	0.674	0.788	0.784	0.786	0.783
Plg + Prx	0.860	0.880	0.914	0.933	0.813	0.828	0.825	0.795
					MSE			
LOI	2.8E-04	-	-	-	4.1E−04	-	-	-
Carb	8.8E-03	7.4E−03	-	-	7.7E−03	6.5E−03	-	-
Cal	1.6E-02	1.4E−02	8.9E−03	8.1E−03	1.5E−02	1.4E−02	1.0E−02	9.3E−03
Dol	1.1E-02	8.8E−03	8.5E−03	7.2E−03	1.1E−02	9.6E−03	9.4E−03	8.6E−03
Qtz	6.9E-03	7.1E−03	2.8E−03	-	6.1E−03	5.8E−03	2.7E−03	-
Kfd	4.2E-04	4.0E−04	4.3E−04	4.2E−04	6.5E−04	6.7E−04	6.4E−04	6.4E−04
Clay	2.0E-03	1.7E−03	1.5E−03	7.4E−04	2.1E−03	1.8E−03	1.7E−03	1.1E−03
Plg	2.6E-04	2.4E−04	2.5E−04	2.5E−04	1.5E−03	1.4E−03	1.3E−03	1.4E−03
Prx	4.9E-04	4.8E−04	4.4E−04	4.4E−04	2.1E−04	2.2E−04	2.1E−04	2.1E−04
Plg + Prx	1.8E-03	1.5E−03	1.1E−03	8.6E−04	1.2E−03	1.3E−03	1.4E−03	1.4E−03

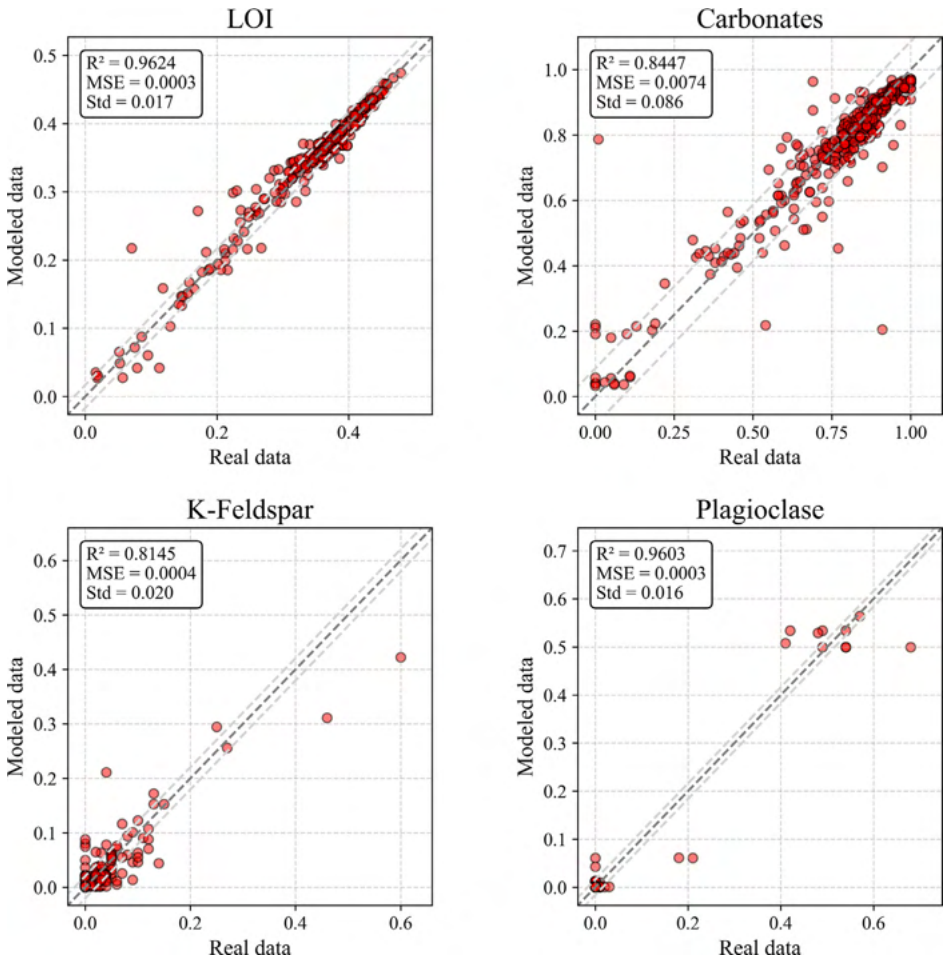


Fig. 5. Real data *versus* modeled data graphs for the test sets of LOI, carbonates, K-feldspar, and plagioclase.

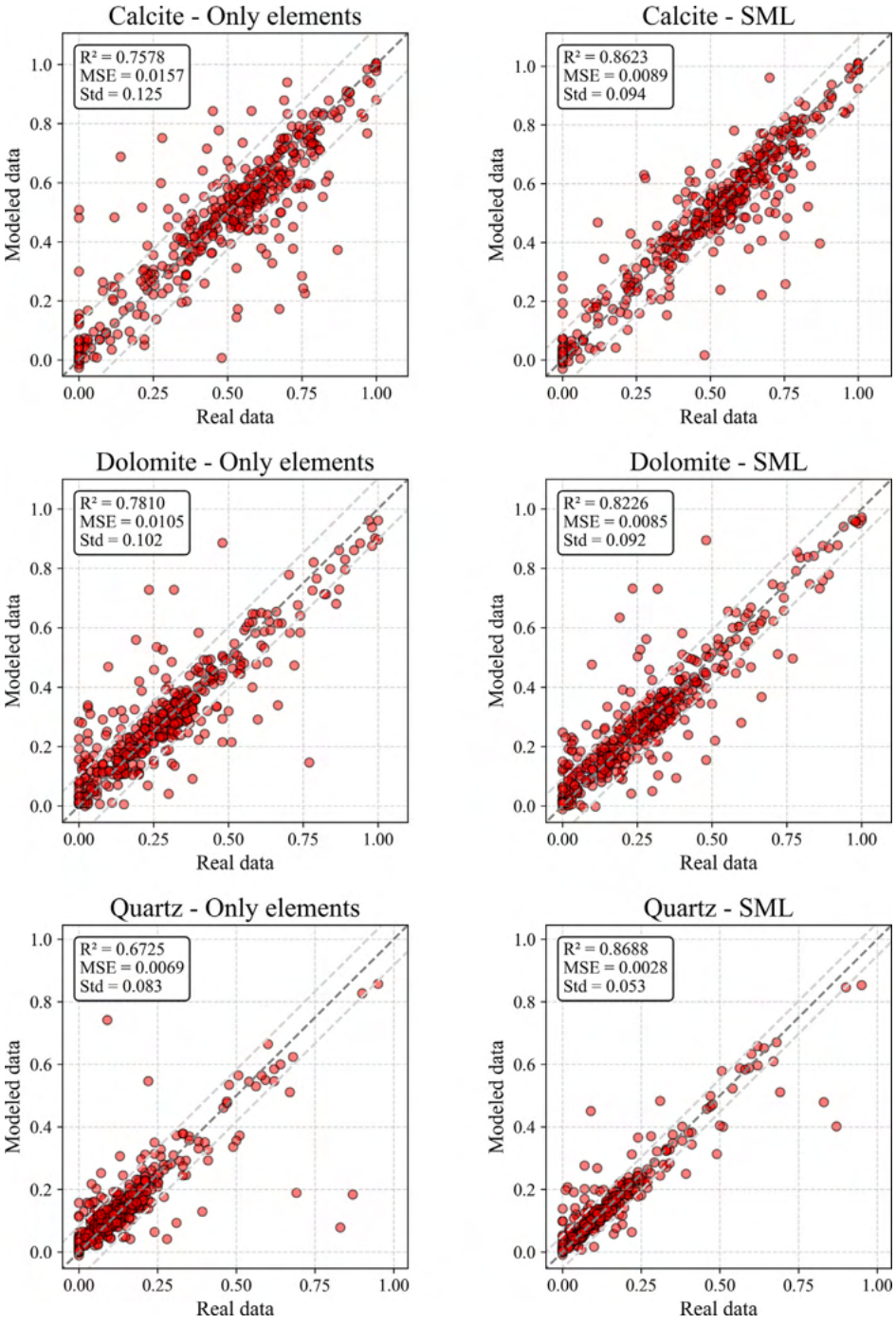


Fig. 6. Real data *versus* modeled data graphs for the test sets of calcite, dolomite, and quartz before and after stepped learning.

Ca and Si were the most critical elements for the LOI model. The importance of Ca is due to the high correlation of LOI with carbonates, whose main element is calcium. Si's importance is probably due to the replacement of carbonate minerals with siliciclastic and igneous minerals rich in silicon, a typical pre-salt pattern. As expected, the carbonate model (calcite + dolomite) is highly impacted by LOI. The calcite model is impacted by Ca, Mg, and carbonates, and the dolomite model is impacted by Mg, LOI, and Ca, something expected for pre-salt carbonates. Na's importance in the dolomite model can be explained by Na and Mg's presence in plagioclase and pyroxene. The increase in Mg together with the increase in Na may indicate the presence of plagioclase and pyroxene, whereas the increase in Mg alone means the presence of dolomite, making Na important in the dolomite model.

The quartz model is mainly impacted by Si and carbonates, reflecting carbonates' replacement by siliciclastic/igneous minerals. The elements K and Al are the most important in the K-feldspar model because they are present in its chemical composition. Although not present in K-feldspar, Ca and Fe are possibly useful to differentiate plagioclase and pyroxene. The clay model is influenced mainly by Mg, Na, LOI, carbonates, quartz, and secondarily by Si and Ca, reflecting both the varied chemical composition of the pre-salt clays and their depositional environment. This result demonstrates how complex it would be to create a clay prediction model without using SML.

Table 6
Standard deviations of the errors obtained in the validation sets of SML.

	Global Error Std (g/g)			
	Elements	Elements + LOI	Elements + LOI + Carb	Elements + LOI + Carb + Qtz
LOI	0.027	-	-	-
Carb	0.095	0.087	-	-
Cal	0.121	0.120	0.096	0.094
Dol	0.100	0.090	0.088	0.087
Qtz	0.085	0.085	0.051	-
Kfd	0.021	0.021	0.021	0.021
Clay	0.045	0.041	0.039	0.028
Plg	0.016	0.018	0.017	0.017
Prx	0.022	0.022	0.021	0.021
Plg + Prx	0.046	0.046	0.042	0.041

The plagioclase and pyroxene + plagioclase models are impacted by Al, an element present in plagioclase composition. Ti, Al, Fe, and Na are essential in the pyroxene model. Al and Na are probably related to the high correlation between pyroxene and plagioclase. Ti and Fe may be present in the pyroxene family's minerals, demonstrating the complexity of creating a traditional mineral model for pre-salt rocks.

Test

The models generated by SML were applied to the geochemical data of three wells drilled in the pre-salt, and the results are shown in Figs. 9, 10, and 11. These figures also present the estimated mineralogy using only the chemical elements, without the application of SML. The mineralogy logs were compared with the XRD analyzes of rotary sidewall cores and core plugs collected in these wells. As these samples do not have XRF analyzes, they were not used to train SML models.

Carbonates of the Barra Velha Formation are observed in Well A, formed predominantly by calcite, dolomite, and quartz. The concentrations of these minerals estimated by the SML model agree with that observed in the XRD analyzes of the rock samples. From X020 m, a decrease in the calcite concentration is observed, detected by the model and XRD analyzes. In the final depths, a layer rich in dolomite, quartz, clay, and K-feldspar is observed, related to the Piçarras Formation's siliciclastic minerals. The SML model can correctly estimate the concentration of these minerals. Regarding the estimated mineralogy using only the chemical elements, pyroxene has anomalous concentrations over the entire range, a problem corrected by the SML model.

In Well B, the Barra Velha and Itapema Formations are observed. The carbonates of the Barra Velha Formation have higher concentrations of dolomite and quartz, while the carbonates of the Itapema Formation are formed mainly by calcite. These characteristics are captured by the SML model and confirmed by the XRD analysis of the rock samples. Between the two carbonates (X660 to X670 m), a layer of siliciclastic rock is marked by the increased concentration of clay and K-feldspar in XRD analyzes. The SML model was able to identify these concentrations correctly. Comparing the mineralogy estimated by SML and only by the chemical elements, the same anomaly in the pyroxene concentration seen in Well A is also observed. Also, the concentrations of clay and dolomite show noisier values when estimated only by the chemical elements, possibly reflecting the worse stability of these models compared to that obtained by SML.

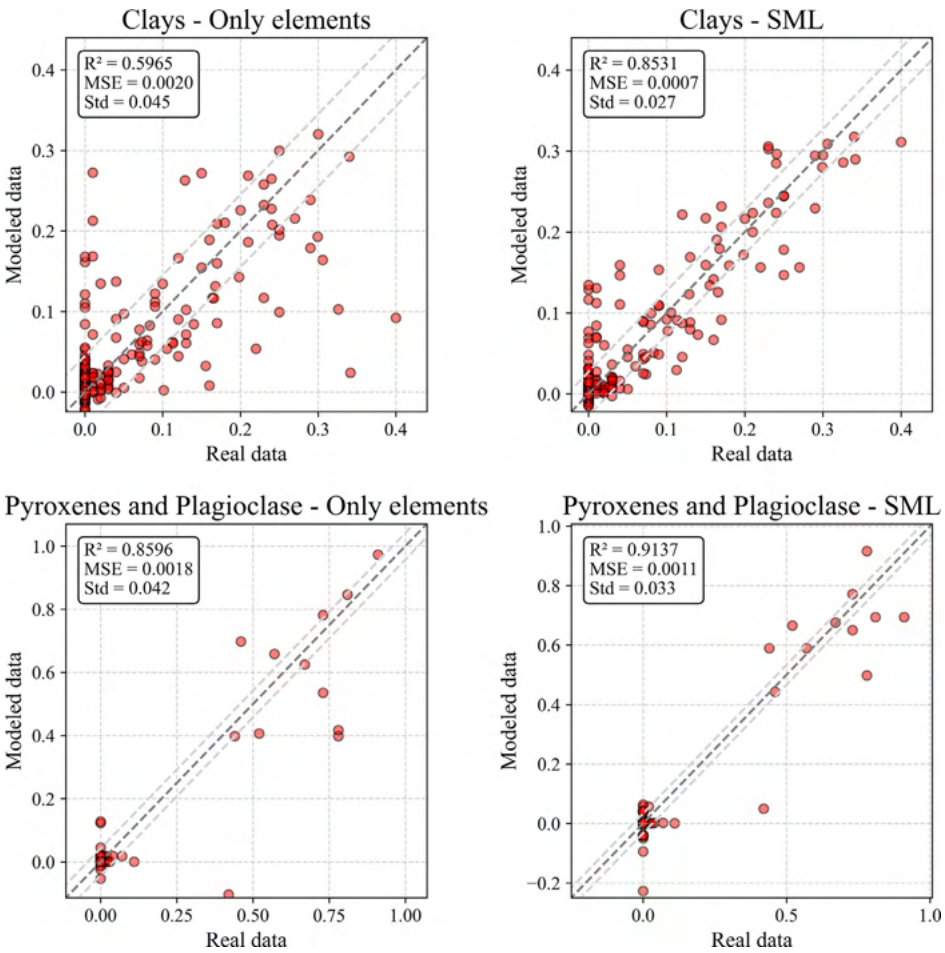


Fig. 7. Real data *versus* modeled data graphs for the test sets of clays and pyroxenes + plagioclase before and after stepped learning.

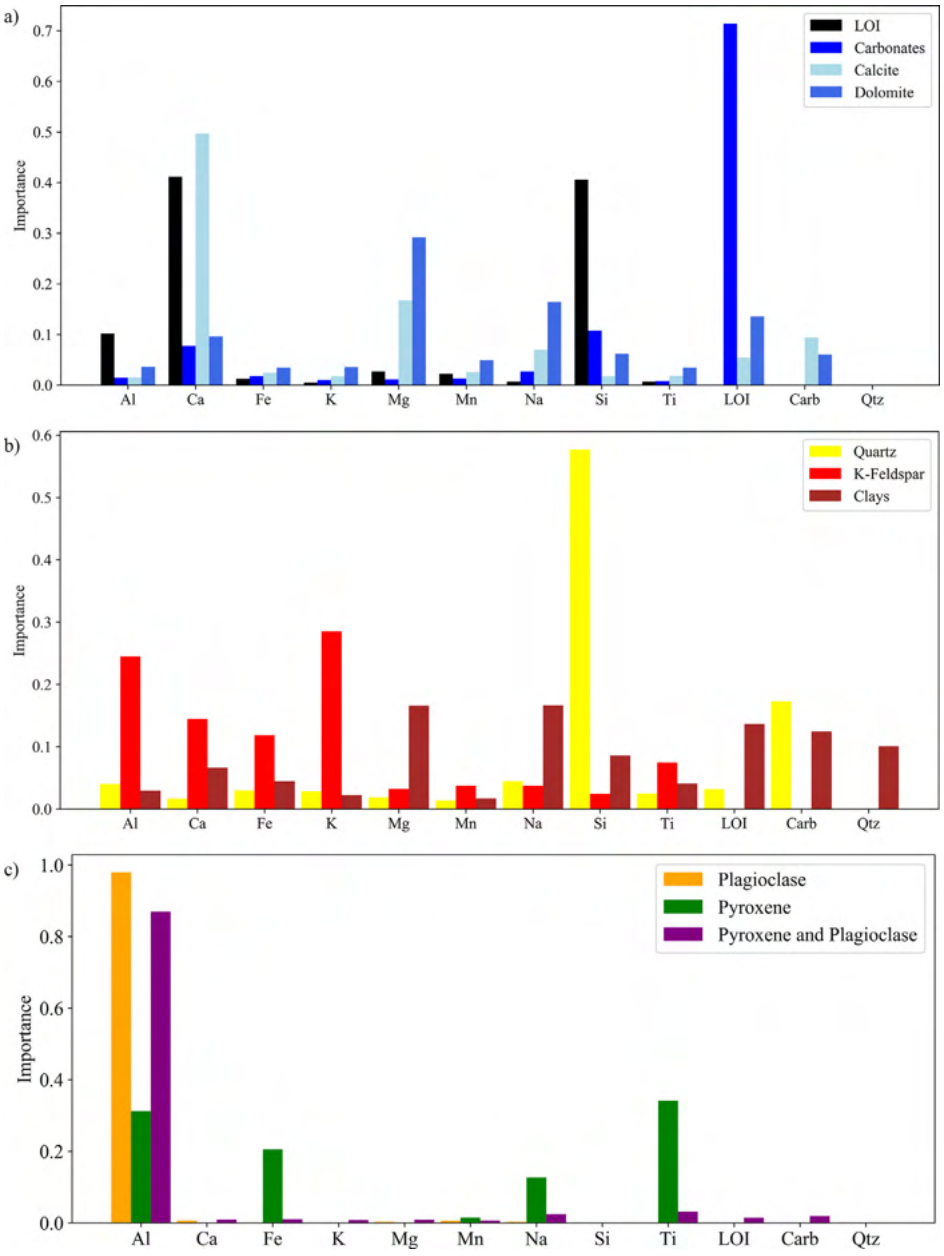


Fig. 8. Variable importance of the chemical elements during the training of the mineralogical model. a) The importance of the variables of the LOI, carbonate, calcite, and dolomite models. b) The importance of the variables of the quartz, K-feldspar, and clay models. c) The importance of the variables of the plagioclase, pyroxene, and pyroxene + plagioclase models.

A thick layer of igneous rock is observed in the first half of Well C. This rock's presence is marked by the increase in plagioclase and pyroxene observed in XRD analyzes of some rock samples. The SML model correctly identified these concentrations. In the rest of the well, the SML model correctly estimated the concentrations of calcite, dolomite, and quartz present in carbonates. Again, anomalous pyroxene concentrations are observed in the mineralogy estimated only by the chemical elements.

It is important to note that the logs acquired by wireline tools have a vertical resolution much lower than that of a core sample. Therefore, the comparison between these two pieces of information should not focus on exact values but general trends. Even so, the concentrations obtained by the SML model were very close to those observed in the core samples.

Discussion

The stepped training strategy generated better results for mineralogical estimates than the individual training of mineral models, marked by the improvement of R^2 in the comparison between real and modeled data and the decrease of the standard deviation of the models' error. Since the standard deviation of the error indicates how wrong a model can be, it can be understood as the model's uncertainty. As the models are not perfect and always have some error, it is essential to emphasize that the stepped training will propagate a model's uncertainties to the others. Therefore, the ideal is that step training is used to promote a clear improvement of the models.

After analyzing the evolution of R^2 and the uncertainty, the best SML flow was defined to create the mineralogical model for the pre-salt rocks. It is shown in Fig. 12, summarized below.

- Estimation of LOI, K-feldspar, and plagioclase using only the chemical elements.
- Inclusion of LOI to chemical elements for the estimation of carbonates (calcite + dolomite).
- Inclusion of carbonates in the chemical elements and LOI for the estimation of calcite, dolomite, quartz, and pyroxene + plagioclase.

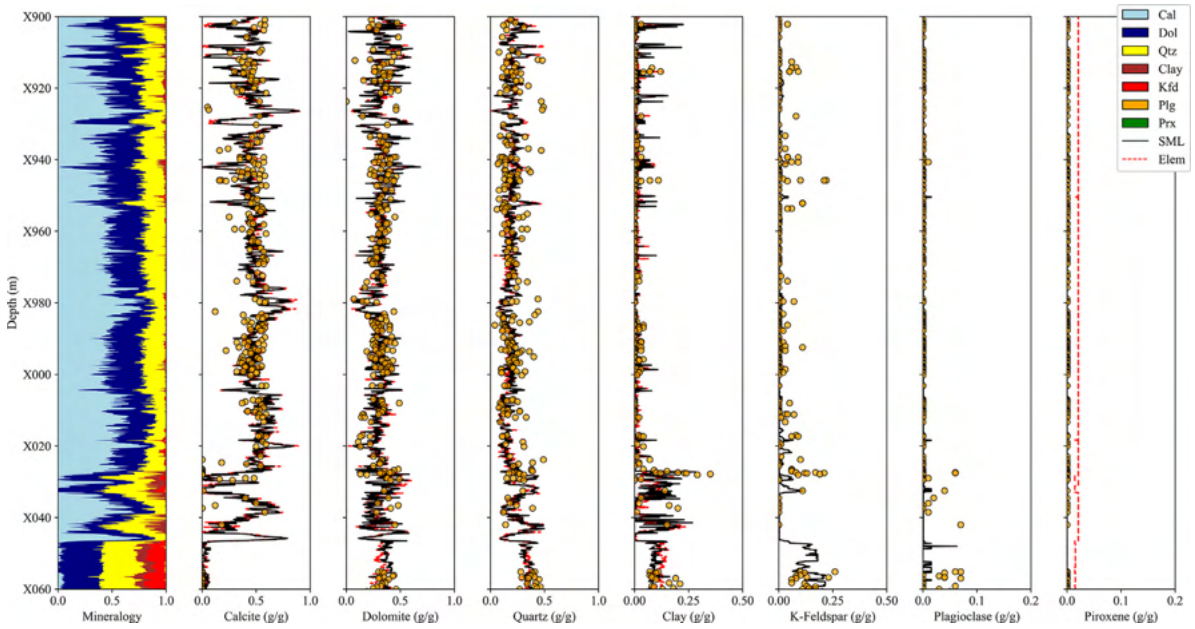


Fig. 9. Mineralogy obtained by applying the SML model to the geochemical logs of Well A and comparison with the mineralogy obtained using only the chemical elements (Elem, dashed red) and XRD analyzes of rock samples. The SML model honors calcite, dolomite, and quartz concentrations in the well's initial depths. The model was also able to correctly estimate the increase in the concentrations of dolomite, quartz, clay, and K-feldspar observed at the end of the well. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

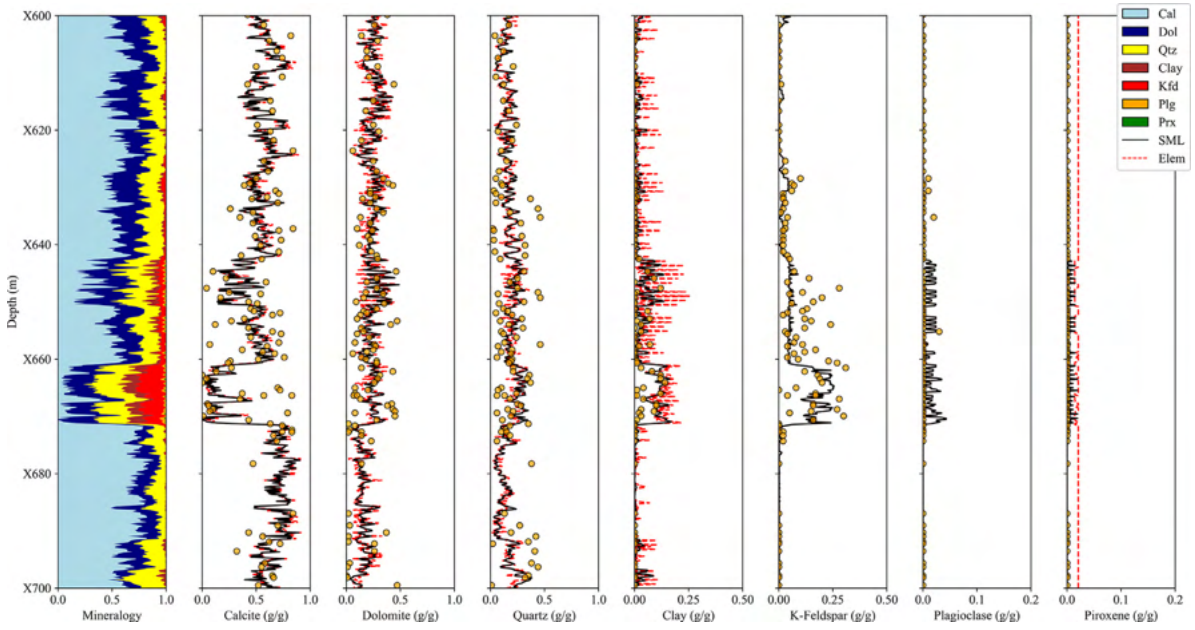


Fig. 10. Mineralogy obtained by applying the SML model to the geochemical logs of Well B and comparison with the mineralogy obtained using only the chemical elements (Elem, dashed red) XRD analyzes of rock samples. The SML model was able to detect the increase in clay and K-feldspar observed in the middle of the well. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Subtraction of plagioclase from the pyroxene + plagioclase for the estimation of pyroxene.
- Inclusion of quartz to chemical elements, LOI, and carbonates for the estimation of clay.

With the proposed SML flow, the best results were extracted from each mineral's models with minimal propagation of uncertainties. The SML models were tested in geochemical logs of wells drilled in the pre-salt rocks, demonstrating the strategy's robustness.

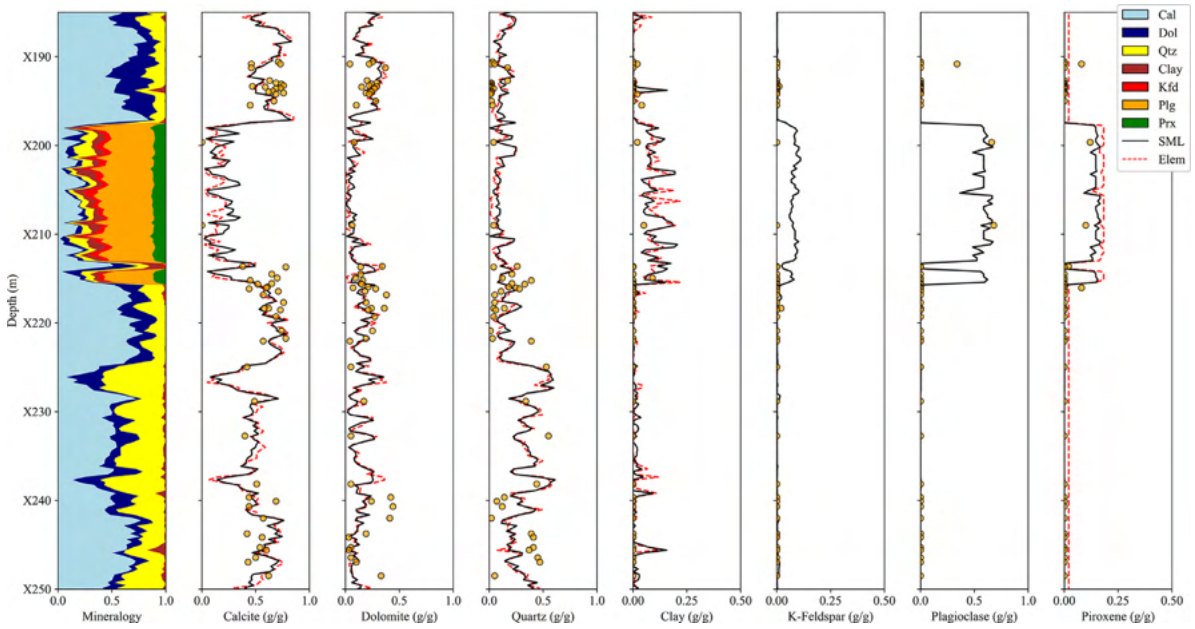


Fig. 11. Mineralogy obtained by applying the SML model to the geochemical logs of Well C and comparison with the mineralogy obtained using only the chemical elements (Elem, dashed red) XRD analyzes of rock samples. The SML model correctly detected the igneous rock layer observed in the first half of the well, with high plagioclase and pyroxene concentrations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

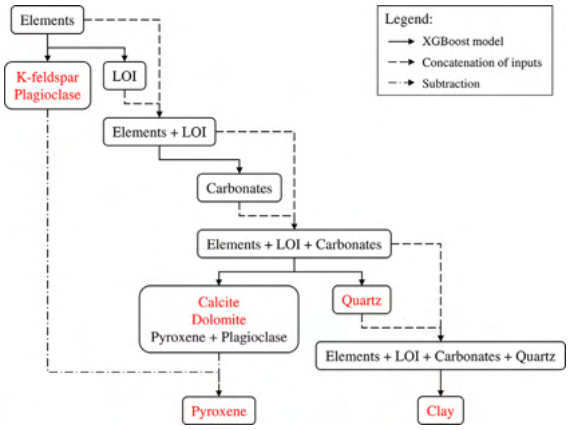


Fig. 12. The SML proposed for the mineral model for pre-salt rocks. Solid arrows indicate the application of the XGBoost model to estimate minerals and LOI, while dashed arrows indicate the concatenation of the modeled output with the previous inputs. Dash-dotted arrows indicate the subtraction of the concentration of pyroxene + plagioclase by plagioclase to obtain pyroxene.

Conclusion

Four machine learning algorithms were tested to create a mineralogical model for pre-salt rocks: MLP, GAN, Random Forest, and XGBoost. XGBoost presented the best initial results, being the chosen algorithm for the generation of mineral models.

The stepped training proposed in the present research generated significant improvements in the mineral models' quality compared to the model using only the chemical elements. The SML model increased the R^2 of the carbonate model from 0.815 to 0.845, of calcite from 0.758 to 0.862, of the dolomite model from 0.781 to 0.823, of quartz from 0.673 to 0.869, and of the clay model from 0.597 to 0.853. A decrease in MSE was also observed.

The application of stepped training propagates the errors and uncertainties of a model. Thus, the SML strategy was not applied to K-feldspar, plagioclase, and pyroxene since the stepped training did not significantly improve the quality of these minerals' models, ensuring that errors were not propagated unnecessarily.

The SML model estimated the mineralogy of three wells drilled in the pre-salt using the geochemical logs. Samples from these wells were not used in training the models. The results obtained honored the XRD analyzes of the samples from these wells, identifying calcitic carbonates with varying concentrations of dolomite and quartz, siliciclastic rocks rich in clay and K-feldspar, and igneous rocks composed of plagioclase and pyroxene. These analyses attested to the model's ability to perform well in the different scenarios found in the pre-salt.

The present research demonstrated that the integration between the machine learning tools and the geological knowledge in stepped learning was crucial for creating a mineralogical model for the Brazilian pre-salt rocks. The real gain of the machine learning application comes not only from its use alone but from what it can add to the previously existing geological knowledge. It is expected that the application of SML can be expanded to solve new challenges found in the pre-salt and other hydrocarbon reservoirs.

Acknowledgments

The authors are grateful to the Brazilian National Agency for Petroleum, Natural Gas and Biofuels (ANP), and Petrobras for supporting this research, and LCT Laboratory for performing part of the analysis. This research was developed with the RDI Group Integrated Technology of Rock and Fluid Analysis (InTRA) and conducted under the Postgraduate Program from the Naval and Oceanic Engineering Department of the Escola Politécnica, Universidade de São Paulo.

References

[1] Aggarwal K, Kirchmeyer M, Yadav P, Keerthi SS, Gallinari P. Benchmarking regression methods: a comparison with CGAN; 2019. Retrieved from http://arxiv.org/abs/1905.12868 .

[2] Anderson RN, Dove RE, Boglia C, Silver LT, James EW, Chappell BW. Elemental and mineralogical analyses using geochemical logs from the Cajon Pass scientific drillhole, California, and their preliminary comparison with core analyses. Geophys Res Lett 1988;15(9):969–72.

[3] Breiman L. Random forests. Mach Learn 2001;45(1):5–32. doi:10.1023/A:1010933404324.
[4] Breiman L, Cutler A, Liaw A, Wiener M. Breiman and Cutler's Random Forests for classification and regression; 2018. Retrieved from https://cran.r-project.org/web/packages/randomForest/randomForest.pdf .
[5] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: 22nd KDD conference on knowledge discovery and data mining; 2016. p. 785–94. doi:10.1145/2939672.2939785.
[6] Cullity B, Stock S. Elements of X-ray diffraction. 3rd ed. New York: Prentice-Hall; 2001.
[7] Dean WEJ. Determination of carbonate and organic matter in calcareous sediments and sedimentary rocks by loss on ignition: comparison with other methods. J Sediment Petrol 1974;44:242–8.
[8] Ellis DV, Singer JM. Well logging for earth scientists. Second. Springer; 2007.
[9] Flaum C, Pirie G. Determination of lithology from induced gamma ray spectroscopy. *SPWLA 23nd Annual Logging Symposium*, 16. Mexico City. Mexico: Society of Petrophysicists and Well-Log Analysts 1981.
[10] Freedman R, Herron S, Anand V, Herron M, May D, Rose D. New method for determining mineralogy and matrix properties from elemental chemistry measured by gamma ray spectroscopy logging tools. SPE Reserv Evaluat Eng 2015;18(04):599–608. doi:10.2118/170722-pa.
[11] Freund Y, Schapire RE. Large margin classification using the perceptron algorithm. Mach Learn 1999;37(3):277–96. doi:10.1023/A:1007662407062.
[12] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;39.
[13] Gonzalez J, Lewis R, Hemingway J, Grau J, Rylander E, Schmitt R. Determination of formation organic carbon content using a new neutron-induced gamma ray spectroscopy service that directly measures carbon. SPWLA 54th annual logging symposium, 15; 2013. doi:101190/urtec2013-112.
[14] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Bengio Y. Generative adversarial nets. Adv Neural Inf Process Syst 2014;3:2672–80.
[15] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. In *springer series in statistics*. Second. Springer; 2009.
[16] Heiri O, Lotter AF, Lemcke G. Loss on ignition as a method for estimating organic and carbonate content in sediments: reproducibility and comparability of results. J Paleolimnol 2001;25:101–10. doi:10.1023/A:1008119611481E.
[17] Herron MM. Mineralogy from geochemical well logging. Clays Clay Min 1986;34(2):204–13. doi:10.1346/ccmn.1986.0340211.
[18] Herron S, Herron M, Pirie I, Saldungaray P, Craddock P, Charsky A, Li T. Application and quality control of core data for the development and validation of elemental spectroscopy log interpretation. SPWLA 55th annual logging symposium, 23. Abu Dhabi: UAE: Society of Petrophysicists and Well-Log Analysts; 2014.
[19] Ho TK. Random decision forests. In: Proceedings of the international conference on document analysis and recognition, ICDAR; 1995. p. 278–82. doi:10.1109/ICDAR.1995.598994.
[20] Jenkins R, Snyder RL. Introduction to X-ray powder diffractometry; 1996. doi:101002/9781118520994.
[21] Jin X, Zhu D, Hill AD, McDuff D. Effects of heterogeneity in mineralogy distribution on acid fracturing efficiency. In: Society of petroleum engineers - hydraulic fracturing technology conference and exhibition; 2019. p. 147–60. doi:10.2118/194377-pa.
[22] Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013. doi:101007/978-1-4614-6849-3.
[23] Moreira JLP, Valdetaro C, Gil JA, Machado MAP. Bacia de Santos. *Boletim de Geociencias Da Petrobras* 2007;15(2):531–49.
[24] North RJ. Through-casing reservoir evaluation using gamma ray spectroscopy. In: SPE california regional meeting. Ventura, California, USA: Society of Petroleum Engineers; 1987. p. 329–42.
[25] Oliveira LAB, Carneiro Cde C. Synthetic geochemical well logs generation using ensemble machine learning technics for the Brazilian pre-salt reservoirs. J Petrol Sci Eng 2021;196:24 January 2021. doi:10.1016/j.petrol.2020.108080.
[26] Rosenblatt F. The Perceptron - A perceiving and recognizing automaton. In *Report*, 85. Buffalo, NY: Cornell Aeronautical Laboratory; 1957.
[27] Simon AH. Sputter processing. In: Seshan K, Schepis D, editors. Handbook of Thin Film Deposition; 2018. p. 195–230. https://doi.org/https://doi.org/. doi:10.1016/C2016-0-03243-6.
[28] Ulloa JM, Chaparro D, Lara S, Arango S, Mendez F, Alarcon N, Gade S. An innovative cased-hole, oil-saturation method of utilizing excess carbon analysis of pulsed neutron measurements in a siliciclastic cenozoic formation, Los Llanos Basin, Colombia.. SPWLA 57th annual logging symposium, 12. Reykjavik, Iceland: society of petrophysicists and well-log analysts; 2016.
[29] Wendemutht A. Learning the unlearnable. J Phys A: Math Gen 1995;28:5423–36. doi:10.1088/0305-4470/28/18/030.
[30] Westaway P, Hertzog R, Plasek RE. Neutron-induced gamma ray spectroscopy for reservoir analysis. Soc Petrol Eng J 1983;23(03):553–64. doi:10.2118/9461-pa.
[31] XGBoost documentation; 2015. Retrieved June 52019from https://xgboost.readthedocs.io/en/latest/parameter.html .
[32] Zhao J, Chen H, Yin L, Li N. Mineral inversion for element capture spectroscopy logging based on optimization theory. J Geophys Eng 2017;14(6):1430–6. doi:10.1088/1742-2140/aa7bfa.