UNIVERSIDADE FEDERAL DA BAHIA
Curso de Pós-graduação em Geofísica

DISSERTAÇÃO DE MESTRADO

PREDIÇÃO DO CARBONO ORGÂNICO TOTAL A PARTIR DE PERFIS GEOFÍSICOS DE POÇOS DA BACIA DE SANTOS

CARLOS ALBERTO CAMPOS DA PURIFICAÇÃO

SALVADOR – BAHIA
12 – 2019

Documento preparado com o sistema LaTeX.

Documento elaborado com os recursos gráficos e de informática do CPGG/UFBA

Predição do carbono orgânico total a partir de perfis geofísicos de poços da Bacia de Santos

por

Carlos Alberto Campos da Purificação

Geofísico (Universidade Federal da Bahia – 2016)

Mestre em Geofísica (Universidade Federal da Bahia – 2019)

Orientador: Prof. Dr. Marcos Alberto Rodrigues Vasconcelos

DISSERTAÇÃO DE MESTRADO

Submetida em satisfação parcial dos requisitos ao grau de

MESTRE EM CIÊNCIAS

EM

GEOFÍSICA

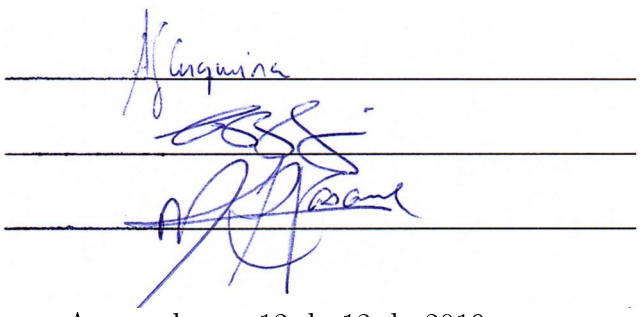
ao

Conselho Acadêmico de Ensino

da

Universidade Federal da Bahia

Comissão Examinadora



Dr. Alexsandro Guerra Cerqueira

Dr. Carolina Barros da Silva

Dr. Marcos Alberto Rodrigues Vasconcelos

Aprovada em 12 de 12 de 2019

A presente pesquisa foi desenvolvida no Centro de Pesquisa em Geofísica e Geologia da UFBA, com recursos próprios, da CAPES, da CNPq, CTPETRO, ANP, PETROBRAS

Ficha catalográfica elaborada pelo Sistema Universitário de Bibliotecas (SIBI/UFBA), com os dados fornecidos pelo(a) autor(a).

Campos da Purificação, Carlos Alberto,
Predição do carbono orgânico total a partir de perfis geofísicos de poços da Bacia de Santos / Carlos Alberto Campos da Purificação. - - , 2019.
106 f.: il., mapas, fotos.

Orientador: Prof. Dr. Marcos Alberto Rodrigues Vasconcelos
Dissertação (Mestrado - Geofísica) - - Universidade Federal da Bahia, Universidade Federal da Bahia, 2019.

1. Aprendizado de Máquina. 2. Carbono Orgânico Total. 3. Perfilagem Geofísica de Poços. 4. Bacia de Santos. 5. Máquina de Vetores de Suporte. I. Rodrigues Vasconcelos, Marcos Alberto. II. Título.

A todos que amo!

Resumo

O conteúdo de Carbono Orgânico total (COT) é a medida da quantidade de carbono disponível em um composto orgânico e é parâmetro essencial na avaliação de rochas potencialmente geradoras de hidrocarbonetos. O uso de perfis geofísicos de poços na avaliação geoquímica de tais rochas é uma técnica importante, não apenas por sua utilidade na capacidade de rápida varredura, mas também pela sua capacidade de identificar o teor de COT dessas rochas. Além disso, os perfis geofísicos podem ser usados ainda nos primeiros estágios da perfuração do poço para identificar intervalos de rochas geradoras. Dentre os perfis geofísicos utilizados para avaliações de rochas geradoras e cálculo de COT, os mais largamente observados na literatura são o densidade (ρ_b), sônico (Δt), raios gama (GR), porosidade neutrônica (ϕ_N) e resistividade profunda (ILD), através de equações empíricas e algoritmos de aprendizagem de máquina. A ideia deste trabalho é obter perfis contínuos de COT, através de uma suíte de perfis geofísicos de poços da Bacia de Santos, utilizando algoritmos de aprendizagem supervisionada, de modo que seja possível identificar rochas potencialmente geradoras onde há ausência ou escassez de dados de COT. Em geral, rochas geradoras apresentam textura fina, como folhelhos, siltitos, calcilutitos e margas. Dependendo da composição das matérias orgânicas nelas presentes e nas condições de temperatura e pressão às quais são submetidas, as condições propícias para a geração de petróleo ou gás poderão se constituir. Além da disponibilidade de dados de COT medidos em rochas potencialmente geradoras, houve uma abundância de dados obtidos a partir de amostras de formações arenosas (arenitos e calcarenitos). Sendo assim, foi possível estimar o COT nestas regiões também. Três algoritmos foram aplicados e comparados entre si, dentre eles a clássica Regressão Linear Múltipla (RLM), além do algoritmo de Máquina de Vetores de Suporte (SVM, do inglês: *support vector machine*) e Florestas Aleatórias (RF, do inglês: *random forest*). O maior poder preditivo dos dois últimos algoritmos frente à RLM está na capacidade de conseguirem ótimos ajustes, sem a necessidade de se ter relações lineares entre as variáveis de entrada e saída, além de não ser necessário aplicar transformações monotônicas (logaritma, inversa, de potência, etc.) nos atributos, sendo também menos influenciados por valores anômalos ou fora da curva (*outliers*) que a RLM. Além disso, não é necessário gerar atributos extras

a partir da iteração das curvas disponíveis, passos estes que são muito significativos para o algoritmo de RLM. Desta forma, faremos uma análise abrangente sobre as vantagens e desvantagens de cada algoritmo aqui utilizado, comparando os resultados obtidos por cada um no cálculo de COT.

Palavras-chave: Carbono Orgânico Total, Aprendizado de Máquina, Perfilagem Geofísica de Poços, Bacia de Santos, Máquina de Vetores de Suporte.

Abstract

Total Organic Carbon (TOC) content is a measure of the amount of carbon available in an organic compound and is an essential parameter in the evaluation of potentially hydrocarbon source rocks. The use of well logs in geochemical evaluation is an important technique, not just for its usefulness as a quick scan for identification of such rocks, but also by their ability to identify the TOC content of these rocks. Geophysical log data can be used to identify source rocks formations in the early stages of well drilling. Consequently, the records used for source rock assessments and TOC calculation commonly includes density, sonic, gamma rays, neutron and resistivity, through several widely spread techniques. In this work, a suite of well logs along with laboratory-measured TOC data of core samples from 10 boreholes in Santos Basin, were used. to obtain TOC profiles continuously over the entire profiled range so that it is possible to identify potentially generating rocks where there is no measured TOC data. Given the abundance (45 % of the data available) of TOC data measured in sandy formations (sandstones and calcarenites), it was possible to make predictions in these regions as well, with purely scientific interest, since the TOC content is not a parameter of industry interest in conventional reservoirs (sandstones and calcarenites, for example). Three algorithms were applied and compared, among which include the classic multiple linear regression (MLR), in addition to the Support Vector Machine (SVM) and Random Forest (RF). Finally, the mean of the results given by the three algorithms was used as a meta-regressor. Although multiple linear regression is an often underrated algorithm compared to others more complex, it returned interesting results in this research, with the smallest errors in clayey rocks and the second smallest error in sandy rocks (after SVM). However, a closer assessment shows that the MLR model displays exaggerated values in a specific formation, when it should not, as oppose to SVM and RF.

Key-words: Total Organic Carbon Content, Machine Learning, Well Logging, Santos Basin, Support Vector Machine.

Índice

Índice de Tabelas

Índice de Figuras

Introdução

Uma das principais aplicações da Geoquímica Orgânica na exploração de hidrocarbonetos é a identificação das rochas potencialmente geradoras de um sistema petrolífero. E dentre as ferramentas Geoquímicas disponíveis, a determinação da quantidade de Carbono Orgânico Total (COT) é considerada uma das mais efetivas na discriminação entre rochas potencialmente geradoras e não-geradoras, a partir de amostras de rochas (Peters e Cassa, 1994).

O COT, geralmente expresso em percentual de massa, é um indicador da quantidade de matéria orgânica presente em uma rocha sedimentar. Por exemplo, um COT de 1% denota que a cada 100g de rocha, 1g é devido somente ao carbono orgânico presente. O COT engloba tanto o querogênio, como a fração de hidrocarbonetos móveis (óleo e/ou gás) presentes nas rochas, mas esses geralmente correspondem a menos de 1 % do COT em rochas geradoras (Jarvie, 1991).

As rochas geradoras são comumente folhelhos e alguns carbonatos (normalmente margas e calcários argilosos) que contêm quantidades significativas de matéria orgânica (Tissot e Welte, 1984). A viabilidade de interpretação da matéria orgânica a partir de perfis geofísicos de poços vem de suas propriedades físicas, que diferem consideravelmente dos componentes minerais de sua rocha hospedeira: menor densidade, maior tempo de trânsito da onda compressional, frequentemente maior teor de urânio, maior resistividade e maiores concentrações de hidrogênio e carbono. Consequentemente, os perfis geofísicos utilizados para avaliações de rochas geradoras, muito comumente incluem o densidade, sônico, raios gama, neutrônico e de resistividade (Serra, 1986; Herron et al., 1988; Luffel et al., 1992).

Numerosos estudos já ilustraram o potencial da utilização da perfilagem geofísica de poços na avaliação de rocha geradoras. Beers (1945), Swanson (1960), Fertl et al. (1980), Schmoker (1981) e Hertzog et al. (1989), por exemplo, fizeram uso do perfil espectral de raios gama para identificar rochas ricas em matéria orgânica. Schmoker e Hester (1983), propuseram o uso do perfil de densidade para estimar o conteúdo de matéria orgânica. Dellenbach et al. (1983) e Hussain et al. (1987) desenvolveram um método usando as curvas de tempo de trânsito e de raios gama para fornecer um parâmetro que se relaciona

linearmente com o teor de matéria orgânica. (Meyer e Nederlof, 1984) introduziram uma técnica combinando os perfis de resistividade, densidade e tempo de trânsito. Este método permite discriminar rochas geradoras de não geradoras, sem no entanto quantificar o teor de matéria orgânica. Mendelzon et al. (1985) lançaram mão da análise multivariada dos dados de perfis, na tentativa de caracterização das rochas geradoras.

Para tempos de trânsito longos e alta resistividade em folhelhos orgânicos, Bessereau et al. (1991) propuseram o método CARBOLOG para obter uma estimativa *in situ* do COT. No entanto, isso requer calibração com base nos dados de COT medidos em laboratório (Yun et al., 2000; Liu et al., 2003). Além disso, Passey et al. (1990), inventaram uma nova técnica chamada ΔLogR. Essa técnica emprega a sobreposição dos perfis de porosidade (sônico, densidade e neutrônico) e o perfil de resistividade, para identificar e quantificar o teor de carbono orgânico total.

Songnian (1998) selecionou a linha base de lamas carbonáticas, no perfil sônico e no perfil resistividade, calculando assim o conteúdo de COT numa dada formação alvo. Kamali e Mirshady (2004) combinaram o método ΔLogR com sistemas Neuro-Fuzzy para determinar o conteúdo de COT presente; técnica esta que também pode ser usada em folhelhos preenchidos com gás (Gás de Folhelho) (Renfang et al., 2009).

Passey et al. (2010) revisaram a calibração do método ΔLogR para incluir rochas altamente maturadas, ricas em matéria orgânica e assim poder identificar rochas fonte de hidrocarbonetos em reservatórios não convencionais do tipo Gás de Folhelho. Amiri Bakhtiar et al. (2011) aplicaram os métodos ΔLogR e redes neurais para estimar o conteúdo de COT na avaliação de rochas geradoras. Dentre as técnicas que utilizam perfis geofísicos de poços, é salutar notar que a técnica do ΔLogR é a mais utilizada na quantificação do COT. No entanto, esta técnica exige uma seleção manual da linha de base nos perfis, mas sabe-se que o nível de COT varia regionalmente em termos de *background* e é difícil de determinar.

Nos últimos anos, sistemas inteligentes e redes neurais foram aplicados na predição de COT. Kadkhodaie-Ilkhchi et al. (2009), aplicaram um comitê de máquinas para estimar o teor de COT em dados petrofísicos. Khoshnoodkia et al. (2011) também usaram um método inteligente para investigar e determinar o COT da Formação Gadvan, usando perfis geofísicos convencionais. No entanto, estimativas de COT usando redes neurais são complexos e envolvem muitos parâmetros; escolher os parâmetros que são relevantes é tipicamente difícil. Na maioria dos casos, a estimativa de COT é obtida através da construção de métodos de regressão multivariada. Adicionalmente, nos últimos anos, algumas técnicas de estimativa de COT diretas, *in situ*, foram introduzidas, incluindo a perfilagem geoquímica espectral. Tal tecnologia poderia fornecer medições diretas de COT, sem exigir algoritmos complexos

(Radtke et al., 2012). No entanto, esta tecnologia não foi amplamente aplicada porque as medições foram feitas apenas em alguns poços.

Devido à alta complexidade existente entre os registros de poços e o teor de COT, nenhuma metodologia é unânime sobre a outra, devendo-se sempre levar em conta a geologia, as condições geoquímicas e os perfis geofísicos disponíveis na abordagem do problema. Apesar de todas essas metodologias empregadas, os algoritmos de árvore foram pouco explorados na predição do COT. Desta forma, o algoritmo Floresta Aleatória foi escolhido aqui para compor a suíte de três algoritmos utilizados nesta pesquisa.

1

Contextualizando o Trabalho e a Área de Estudo

1.1 Bacia de Santos

A Bacia de Santos está localizada na margem sudeste do Brasil, entre os Altos de Cabo Frio - paralelo 23°30'S - e Florianópolis - paralelo 28°00'S (Caldas, 2007). Limita-se ao norte com a Bacia de Campos e ao sul com a Bacia de Pelotas, a oeste com a Serra do Mar e a leste com o limite oriental do Platô de São Paulo (Gamboa et al., 2008). Engloba o território marítimo dos estados de Santa Catarina, Paraná, São Paulo e Rio de Janeiro, além de território marítimo internacional (Figura 1.1). Numa escala mais ampla, a área de estudo envolveria além do sudeste brasileiro, o sudoeste Africano, incluindo suas respectivas áreas *offshore*.

Segundo Clemente (2013), as informações disponíveis sobre a bacia de Santos são relativamente escassas por questões de confidencialidade. A exploração, descoberta e explotação de hidrocarbonetos desta bacia vem sendo realizada há um intervalo de tempo relativamente curto (menos de 10 anos). Antes das descobertas de petróleo na bacia de Santos, a bacia de Campos representava o maior percentual da produção de petróleo no Brasil (entre 50-75 %). Agora, acredita-se que a bacia de Santos represente mais de 50 % nos aumentos das reservas brasileiras de óleo.

O tamanho da Bacia de Santos, suas semelhanças geológicas e proximidade com a Bacia de Campos, justificavam a condição de bacia promissora. Entretanto, não representou nenhum sucesso exploratório até o início dos anos 1980 (Pereira e Macedo, 1990).

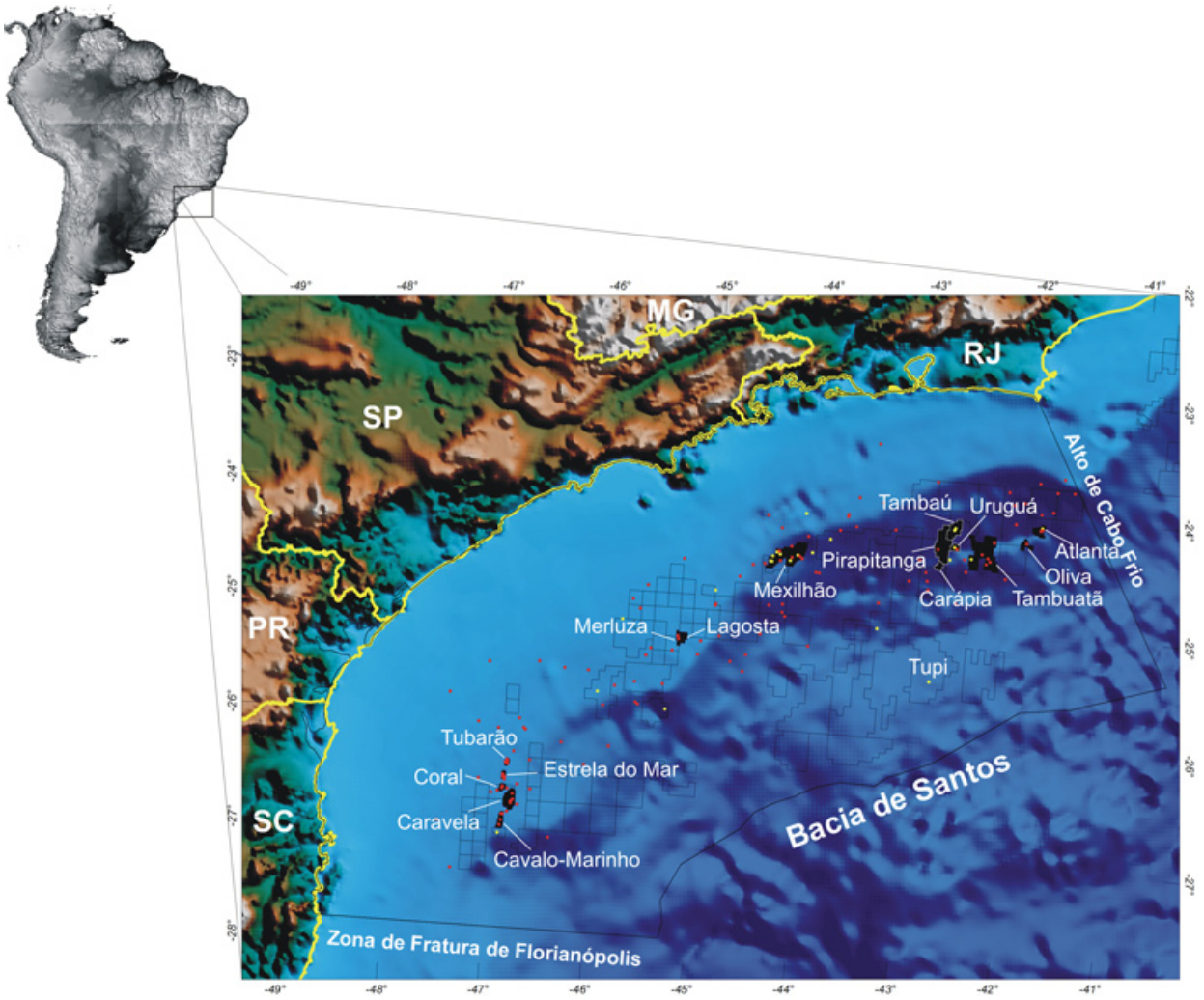


Figura 1.1: Mapa de localização da Bacia de Santos e áreas adjacentes. A imagem colorida resulta da integração do modelo digital de terreno (GTOPO30) e da batimetria da bacia. Os polígonos preenchidos correspondem aos campos produtores e os polígonos sem preenchimento aos blocos exploratórios. Os pontos em vermelho correspondem aos poços exploratórios públicos e os amarelos a poços exploratórios confidenciais. Fonte: Chang et al. (2008)

As atividades de exploração e produção da Petrobras nas bacias da margem leste brasileira nas décadas de 1960 e 1970 apontaram para o potencial de reservas de petróleo e gás natural na Bacia de Santos (Pereira et al., 1986), sobretudo, depois do início da produção de hidrocarbonetos na Bacia de Campos no final dos anos de 1970 (Pereira e Macedo, 1990).

Essa condição apresentou mudanças significativas com a implementação dos polos de produção de gás natural Merluza, na década de 1990 e Mexilhão, início dos anos 2000 (Assine et al., 2008). Inclusive, segundo Sauer (2016), estimativas apontam que o volume das reservas do Pré-sal na Bacia de Santos seja superior a 100 bilhões de barris, posicionando o país entre as cinco maiores reservas mundiais, condição esta que atraiu as atenções da indústria mundial do petróleo.

As descobertas dos campos de Tubarão, Coral, Estrela do Mar e Caravela, em reservatórios carbonáticos albianos, no sul da Bacia de Santos, aumentaram as expectativas de ser uma grande bacia petrolífera. Após a criação da nova lei do Petróleo (Lei no 9.478, 3 de 1997), a Bacia de Santos recebeu novamente atenção exploratória da Petrobrás e de outras companhias estrangeiras na aquisição de dados geológicos e geofísicos, que culminaram na descoberta de novas jazidas: de óleo em Oliva e Atlanta; e de gás natural em Lagosta, Tambuatá, Tambaú, Mexilhão, Carapiá, Uruguá e Pirapitanga.

Com relação à litoestratigrafia, focaremos nas formações que cortam os perfis de poço que nos foram disponibilizados pela ANP, nas regiões onde há medidas de COT. Assim sendo, as formações em questão são: Marambaia, Juréia, Itajaí-Açu, Itanhaém, Guarujá e Ariri, que estão ilustradas na Figura 1.2. Nesta imagem da seção esquemática da Bacia de Santos, ilustra-se, além das Formações, algumas estruturas geológicas que as compõem.

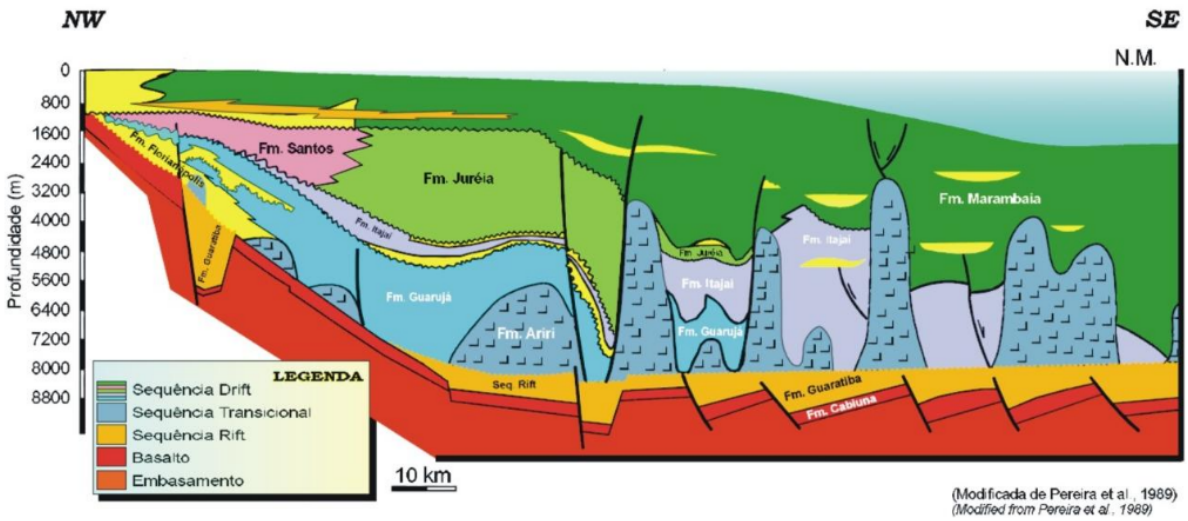


Figura 1.2: Seção geológica esquemática da Bacia de Santos. Fonte: Mohriak (2003)

Segundo Moreira et al. (2007), a Formação Marambaia é depositada nas regiões de plataforma distal, talude e bacia, predominando os siltitos e folhelhos, além de diamictitos e margas. Expressivos cânions surgem cortando esses sedimentos. Ainda segundo o mesmo autor, em seu interior e nas regiões batiais ocorrem os arenitos resultantes de fluxos turbidíticos densos, fortemente canalizados, que compõe o Membro Maresias da Formação Marambaia.

A formação Ariri é composta por evaporitos. Diferente das cartas anteriores, o tempo estimado de deposição é de 0,7 a 1 Ma. Geralmente, os evaporitos são compostos por halita e anidrita. Entretanto, constatou-se a presença de sais mais solúveis, tais como taquidrita, carnalita e, localmente, silvinita (Dias et al., 1998).

Segundo De Souza et al. (1993), a Formação Guarujá é caracterizada pela implantação de uma plataforma carbonática ao longo do Albiano e corresponde às fácies distais. Os carbonatos do Albiano apresentam-se basculados, devido à movimentação tectônica do sal aptiano subjacente, originou falhamentos em blocos rotacionados e estruturas do tipo turtleback (casco de tartaruga). Além disso, segundo Vivier (1987), a Formação Guarujá apresenta três seções bem definidas, que compreendem calcarenitos, calcilutitos e margas.

A formação Itanhaém é caracterizada por folhelhos e, mais raramente, margas de origem marinha distribuídas desde a plataforma até as regiões bacinais. Interacamadados na formação Itanhaém encontram-se os depósitos arenosos de sistemas originados por fluxos gravitacionais densos que compõem o Membro Tombo. Estes arenitos geralmente ocorrem encaixados em baixos deposicionais gerados e controlados pela tectônica salífera albiana. Esta seqüência apresenta um padrão retrogradante e seus depósitos são resposta a uma progressiva subida relativa do nível do mar com afogamento da plataforma rasa pelos sedimentos pelágicos (Moreira et al., 2007).

Ainda segundo Moreira et al. (2007), a Formação Juréia ocorre sob a forma de sedimentos arenosos, folhelhos, siltitos e argilitos depositados desde os ambientes continentais até as porções mais distais da plataforma. Interacamado a estes depósitos ocorre intenso vulcanismo extrusivo, assim como níveis de coquinas e calcilutitos podem ocorrer intercalados.

A formação Itajaí-Açu é caracterizada por folhelhos e argilitos cinza-escuros, depositados nos ambientes de plataforma distal, talude e bacia. A porção inferior da Formação Itajaí-Açu corresponde aos folhelhos escuros depositados durante o Evento Global Anóxico ocorrido durante o Médio Cretáceo (Araing, 1988). Estes folhelhos anóxicos podem atingir espessuras superiores a 500 m, o que é consideravelmente maior às observadas nos furos do DSDP 364 (na Margem Continental de Angola) e do DSDP 356 (no Platô de São Paulo). Na Figura 1.3 logo abaixo, temos a carta estratigráfica da Bacia de Santos.

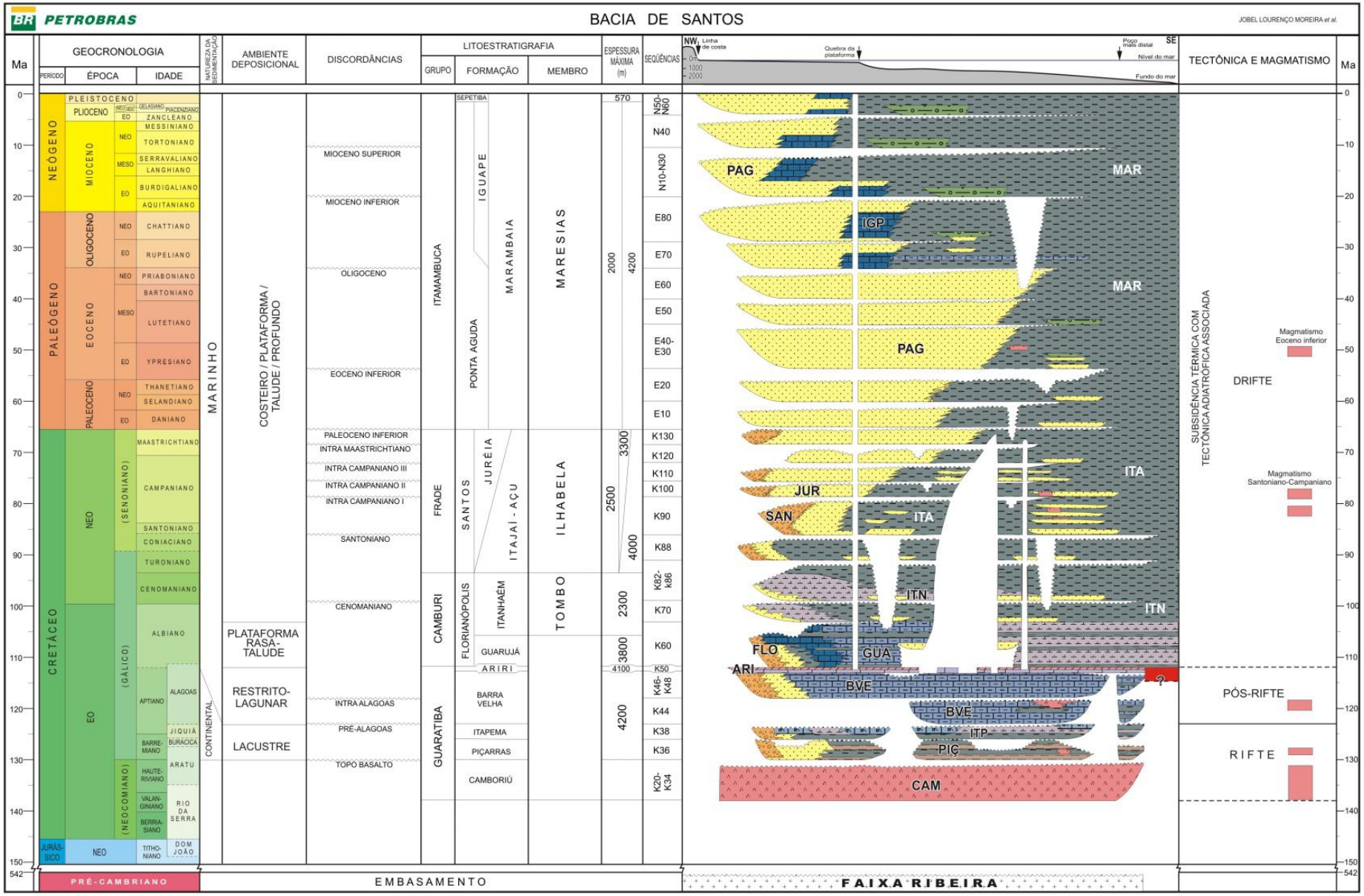


Figura 1.3: Diagrama estratigráfico da Bacia de Santos. Fonte: Moreira et al. (2007)

1.2 Geração e Migração

De acordo com Chang et al. (2008), análises de biomarcadores em amostras de óleo provindas de 15 amostras selecionadas ao longo da bacia caracterizaram a provável fonte de óleo como sendo lacustre salino, com contribuição marinha siliciclástica.

Há na Bacia de Santos dois intervalos geradores de hidrocarbonetos: a Formação Piçarras e a Formação Itajaí-Açu. As rochas geradoras da Formação Piçarras foram depositadas em ambiente lacustre salino no estágio final da fase rifte, no Aptiano. Supõe-se que nesse ambiente, o sistema de lagos passou a receber influência de águas salinas do sul, tendo se tornado salinizado devido ao acréscimo de aridez ao final do Cretáceo Superior (Chang et al., 2008). Já de acordo com Moreira et al. (2007), a Formação Piçarras corresponde a depósitos de leques aluviais compostos por conglomerados e arenitos polimíticos, nas porções proximais, e por arenitos, siltitos e folhelhos de composição talco-estevensítica, nas porções lacustres. Os valores para a concentração de Carbono Orgânico Total (COT) para as rochas geradoras da Formação Piçarras variam entre 2 a 6 %. O Índice de Hidrogênio é superior a 900 mg de HC/g COT, o que indica a formação de querogênio do tipo I. As rochas da Formação Itajaí-Açu são representadas por folhelhos e argilitos cinza-escuros depositados nos ambientes de plataforma distal, talude e bacia (Moreira et al., 2007).

O valor de COT médio para as rochas da Formação Itajaí-Açu é próximo a 1 %, com máximo de 6 %, e análise de amostras de rochas dessa formação indicou que sua matéria orgânica é composta por uma mistura dos tipos II e III, ou seja, de origem marinha depositada em ambientes redutores e de origem terrestre. Segundo resultados geoquímicos sobre a origem dos óleos, as rochas da Formação Itajaí-Açu entraram na janela de geração em diferentes locais da bacia (Chang et al., 2008).

1.3 Rochas Reservatório

A Bacia de Santos tem um conjunto diversificado de rochas reservatório, como os carbonatos oolíticos de águas rasas da Formação Guarujá, os arenitos turbidíticos eocênicos da Formação Marambaia e do Membro Ilha Bela, da Formação Itajaí-Açu (Chang et al., 2008). Além deles, há também os carbonatos das formações Itapema e Barra Velha, que constituem os principais reservatórios da seção pré-sal, com gigantescos volumes de óleo descobertos nos campos de Lula, Sapinhoá, Búzios, entre outros.

Os reservatórios das formações Itapema e Barra Velha são compostos por rochas carbonáticas formadas por coquinas e/ou microbialitos, além de coquinhas de ostracodes e clastos

de etromatólitos. Os carbonatos microbiais ocorrem nas seções rifte superior (sin-rifte) e sag (pós-rifte), podendo estar sobrepostos a depósitos de coquinas da Formação Itapema (rifte superior), de idade neobarremiana-eoaptiana. As coquinas são calcirruditos constituídos de fragmentos de conchas e pelecípodes frequentemente dolomitizados ou silicificados.

Os carbonatos da Formação Guarujá representam os reservatórios mais importantes da seção pós-sal devido ao grande volume de óleo descoberto nessas rochas nos campos de Tubarão, Estrela do Mar, Coral, Caravela e Cavalo-Marinho (Chang et al., 2008). Correspondem a calcarenitos oolíticos e foram depositados em águas rasas, em uma extensa plataforma carbonática durante o Albiano Médio- Inferior.

Os reservatórios turbidíticos do Membro Ilhabela têm ocorrência intercalada com pelitos de águas profundas da Formação Itajaí-Açu. Essas rochas funcionam como reservatórios para os campos de Merluza, Lagosta e Mexilhão.

Além dos reservatórios citados, outras unidades litoestratigráficas compõem reservatórios siliciclásticos, como os arenitos das Formações Santos e Juréia e os arenitos turbidíticos do Paleoceno, Eoceno e Oligoceno da Formação Marambaia, que são reservatórios para os campos de Oliva, Atlanta e Baúna.

1.4 Sobre a Pesquisa

1.4.1 Objetivo

O principal objetivo desta pesquisa é a obtenção do conteúdo de carbono orgânico total (COT) a partir de perfis geofísicos de poços, valendo-se de algoritmos de aprendizado de máquina (AM) e após, fazer um estudo comparativo, ressaltando as vantagens e desvantagens de cada algoritmo utilizado.

1.4.2 Base de Dados

A base de dados utilizada neste estudo foi cedida pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) e contam com dados de perfis geofísicos de poços e dados geoquímicos. Os dados de perfilagem são compostos por 12 poços no formato *.dlis*, com a suíte de curvas Resistividade Indutiva Profunda (ILD, em Ω.m); Raios Gama (GR, em unidades API); Densidade (ρ_b, em g/cm^3); Porosidade Neutrônica (ϕ_N, em %) e Sônico (Δt, em μs/pé), entre outros que não foram utilizados aqui, como a curva de cáliper, por exemplo. Os dados de geoquímica consistem em uma planilha no formato *.xlsx* com dados coletados

em amostras de calha, testemunho lateral, pontual e intervalar. Estes dados consistem em medidas de COT (%), Resíduo insolúvel (%), S1-Hidrocarbonetos livres (mg HC/g rocha), S2-Hidrocarbonetos liberados (mg HC/g rocha), S3-CO2 liberado (mg CO2/g rocha), Temperatura máxima (°C), Índice de hidrogênio S2/COT (mg HC/g COT) e Índice de oxigênio S3/COT (mg CO2/g COT). Dentre os dados geoquímicos disponíveis, apenas o COT foi utilizado. Dos 12 poços disponíveis, apenas 10 possuem dados de COT, formando uma base de dados com 817 medidas de COT. Dados advindos de 9 destes poços foram utilizados na etapa de treinamento e validação dos algoritmos, formando um total de 730 medidas. O poço teste (poço "cego") continha 87 medidas de COT e foi utilizado para comparação dos ajustes das curvas dadas pelos algoritmos em dados não processados a priori pelos mesmos. Os outros 2 poços sem dados de COT foram utilizados para avaliar o poder preditivo dos algoritmos de aprendizado de máquina (após treinados, validados e testados) na ausência de tais medidas.

Dados de COT medidos em calhas intervalares foram descartados, por apresentarem um único valor médio de COT dentro de um volume muito grande de rocha. O escopo do trabalho é estabelecer relações entre medidas pontuais do COT com os respectivos valores dos perfis geofísicos naquela profundidade específica; assim sendo; utilizou-se apenas os dados de COT medidos em calhas pontuais (de onde veio o maior volume de dados) e alguns poucos medidos em amostras laterais de rochas e testemunhos pontuais.

As proporções litológicas onde há dados de COT, no conjunto de treinamento e no poço "cego", estão representadas respectivamente pelas Figuras 1.4 e 1.5 a seguir. Em termos de medidas de COT realizadas nas litologias de marga e calcilutito, as proporções entre o conjunto de treinamento e o poço teste são semelhantes. Já em relação a arenitos e calcarenitos, há relativamente mais dados de COT medidos nestas litologias no poço teste que nos de treinamento. Folhelhos e arenitos são as litologias onde mais se tem valores medidos de COT, compondo juntos 73 % dos dados de treinamento e quase 67 % do poço teste.

Na Figura 1.6 está representado o mapa da região trabalhada, com as localizações e nome dos poços, bem como a discriminação entre aqueles que foram utilizados para treinamento e validação (círculos verdes), teste (círculo vermelho), além de dois outros poços sem dados de COT (círculos azuis). Apenas o poço 1-BRSA-211-RJS faz parte do Campo de Uruguá, enquanto todos os outros fazem parte do Campo de Tambuatá.

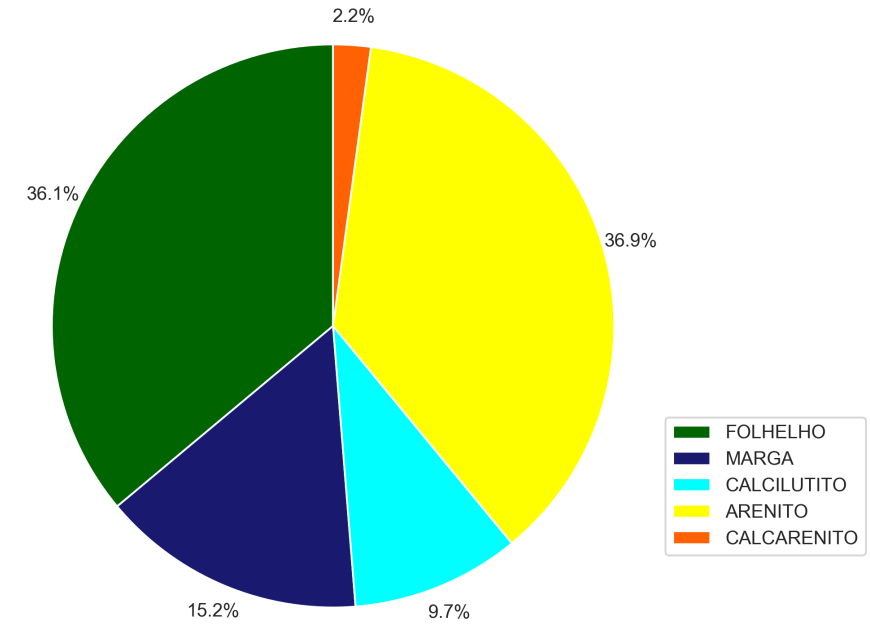


Figura 1.4: Gráfico exibindo as proporções litológicas para os dados de treinamento, onde há valores de COT medido. Folhelhos e arenitos compõem 73 % das litologias onde foi medido o conteúdo de COT.

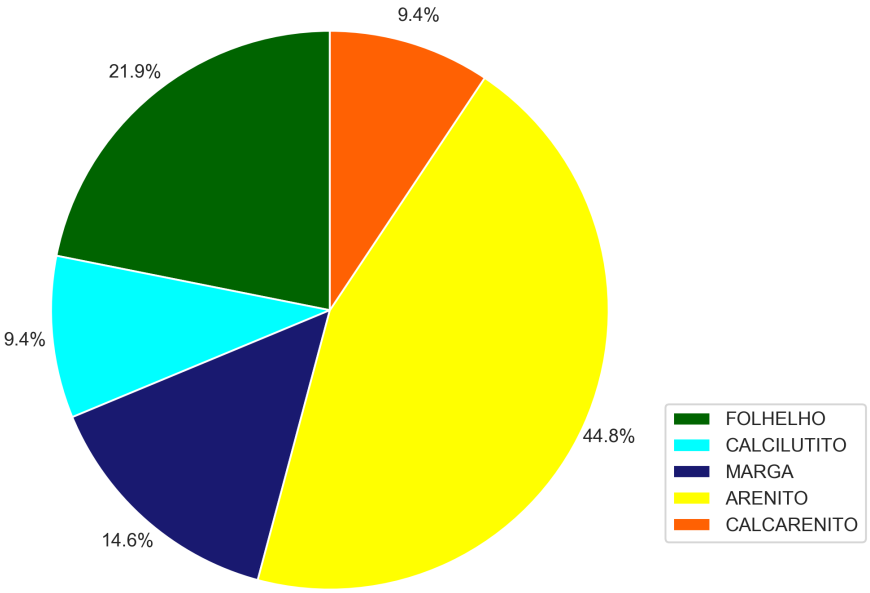


Figura 1.5: Gráfico exibindo as proporções litológicas para os dados do poço teste, onde há valores de COT medido. Folhelhos e arenitos compõem quase 67 % das litologias onde foi medido o conteúdo de COT.

O sumário executivo externo da ANP de 28/07/2011, afirma que o Campo de Tambuatá

possui área de 738 km^2 e é oriundo da "Rodada Zero", localizado a aproximadamente 170 km do litoral do Rio de Janeiro, em lâmina d'água entre 1050 e 1750 m. Foi descoberto pelo poço 1-RJS-539 (localizado também no mapa) em agosto de 1999 e, até o momento, foram perfurados um total de treze poços dentro de suas adjacências. Os reservatórios do Campo de Tambuatá estão representados por arenitos de idade Eoceno Inferior (Fm. Marambaia), Campaniano e Santoniano (Fm. Itajaí-Açu/Mb. Ilha Bela) e Albiano Superior (Fm. Itanhaém). Os hidrocarbonetos encontrados na área campo são de origem lacustre, gerados em sedimentos finos da Fm. Guaratiba.

Os reservatórios do campo de Uruguá são formados principalmente por arenitos da Formação Itajaí-Açu, gerados através de sistemas turbidíticos com idades que variam do turaniano ao campaniano e situam-se entre 4000 a 5000 metros abaixo do nível do mar.

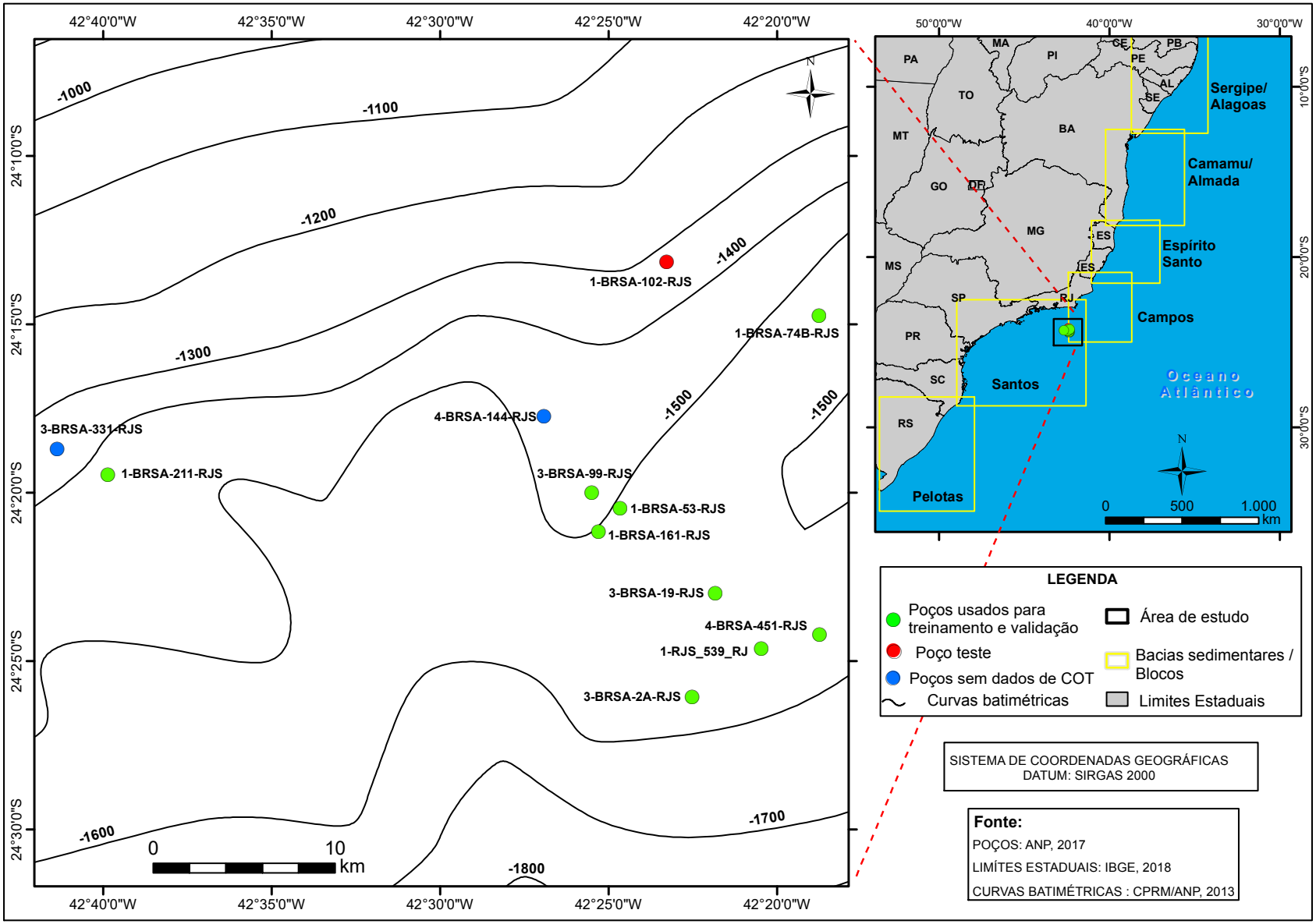


Figura 1.6: Mapa da região em estudo, exibindo a posição dos poços de treinamento e teste.

1.4.3 Metodologia

Uma vez que os perfis geofísicos de cada poço vieram fragmentados em diversos arquivos no formato *.dlis*, foi necessário escrever códigos em *Python* de modo que fosse feita a concatenação desses vários intervalos em um só. Primeiramente, estes pequenos intervalos de perfis em formato *.dlis* foram transformados em arquivos no formato *.las*, que eram então lidos e trabalhados no *Python*. Isso foi alcançado plotando-se todos os fragmentos em cores distintas, num único *track*. Uma seleção com base em análise visual foi feita, comparando-se os fragmentos de perfis plotados, com os perfis compostos ofertados pela ANP, para os poços em questão. Logo após, foram concatenados, formando um perfil único da grandeza física em questão em função da profundidade (Figura 1.7). Em seguida, os perfis foram plotados juntos aos dados de COT e das litologias, em função da profundidade. Finalmente, a sequencia abaixo foi seguida como aplicação metodológica:

1. Agrupamento dos dados de COT em dois grandes grupos, com base no seguinte critério litológico:
   I - Rochas argilosas: calcilutitos, margas e folhelhos.
   II - Rochas arenosas: arenitos e calcarenitos.
   Obs.: Os dados de COT medidos nas demais litologias foram descartados, seja por escassez, ou por serem dados medidos em rochas sem relevância para esta pesquisa, relativamente ao fato de não serem rochas reservatório nem geradoras da bacia em questão (diabásio e halita, por exemplo);

2. Remoção de valores de COT discrepantes (*outliers*)[1], e/ou inconsistentes, com base em análise visual de vários gráficos de dispersão (*cross-plots*) plotados como matriz de correlação (perfis×COT);

3. O algoritmo Máquina de vetores de Suporte (SVR) foi escolhido para fazer a seleção da melhor combinação de perfis na predição de COT, servindo como etapa de seleção de atributos (*feature selection*). Nesta fase, realizou-se combinações dos 5 perfis 2 a 2, 3 a 3 e 4 a 4 para cada grupo (argilosas e arenosas), além de se utilizar cada um dos perfis como dados de entrada, separadamente. Deste modo, foi utilizado o objeto *k-fold cross-validator* da biblioteca *sklearn* do *Python*, para seleção dos hiperparâmetros (C e γ) ótimos;

4. A combinação de perfis que retornou o menor erro quadrático médio (MSE) com os

[1] *Outliers* são dados que se diferenciam drasticamente de todos os outros, são pontos fora da curva. Em outras palavras, um *outlier* é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos.

hiperparâmetros ótimos, foi utilizada como dados de entrada para o modelo de Regressão Linear Múltipla (RLM) e para o Floresta Aleatória (RF), onde neste último, aplicou-se também validação cruzada para seleção dos hiperparâmetros ótimo, etapa esta não necessária à RLM, posto que não possui hiperparâmetros;

5. *Crossplots* exibindo o Coeficiente de Correlação de Pearson (R) entre os valores de COT medidos e os resultados das predições nas rochas argilosas e arenosas, além da correlação global (juntando os dois grupos), além de tabelas explorando os erros nas fases de treinamento, validação e no poço teste, para cada um dos algoritmos;

6. Por fim, os resultados foram apresentados como perfis de COT, gerados por cada algoritmo, para todo o intervalo perfilado do poço teste onde houvesse arenitos, calcarenitos, folhelhos, margas ou calcilutitos, no mesmo *track* dos dados de COT medidos. Além disso, apresentou-se o resultado em dois poços sem dados de COT, para efeito de se avaliar e mostrar o objetivo final para o qual os algoritmos de AM se propõem: predição onde não se tem dados da variável resposta.

O fluxograma exibido na Figura 1.8 ilustra os passos para a aplicação dos algoritmos. Estes passos foram aplicados separadamente nos dois grupos de rochas previamente separados (rochas argilosas e arenosas) e quando aplicados no poço teste, levou-se em consideração os mesmos grupos de rochas dos poços de treinamento; ou seja; o algoritmo treinado nas formações argilosas foi aplicado nas formações argilosas do poço teste, analogamente para as rochas arenosas, e o resultado das duas predições foi exibido em um perfil único para cada algoritmo.

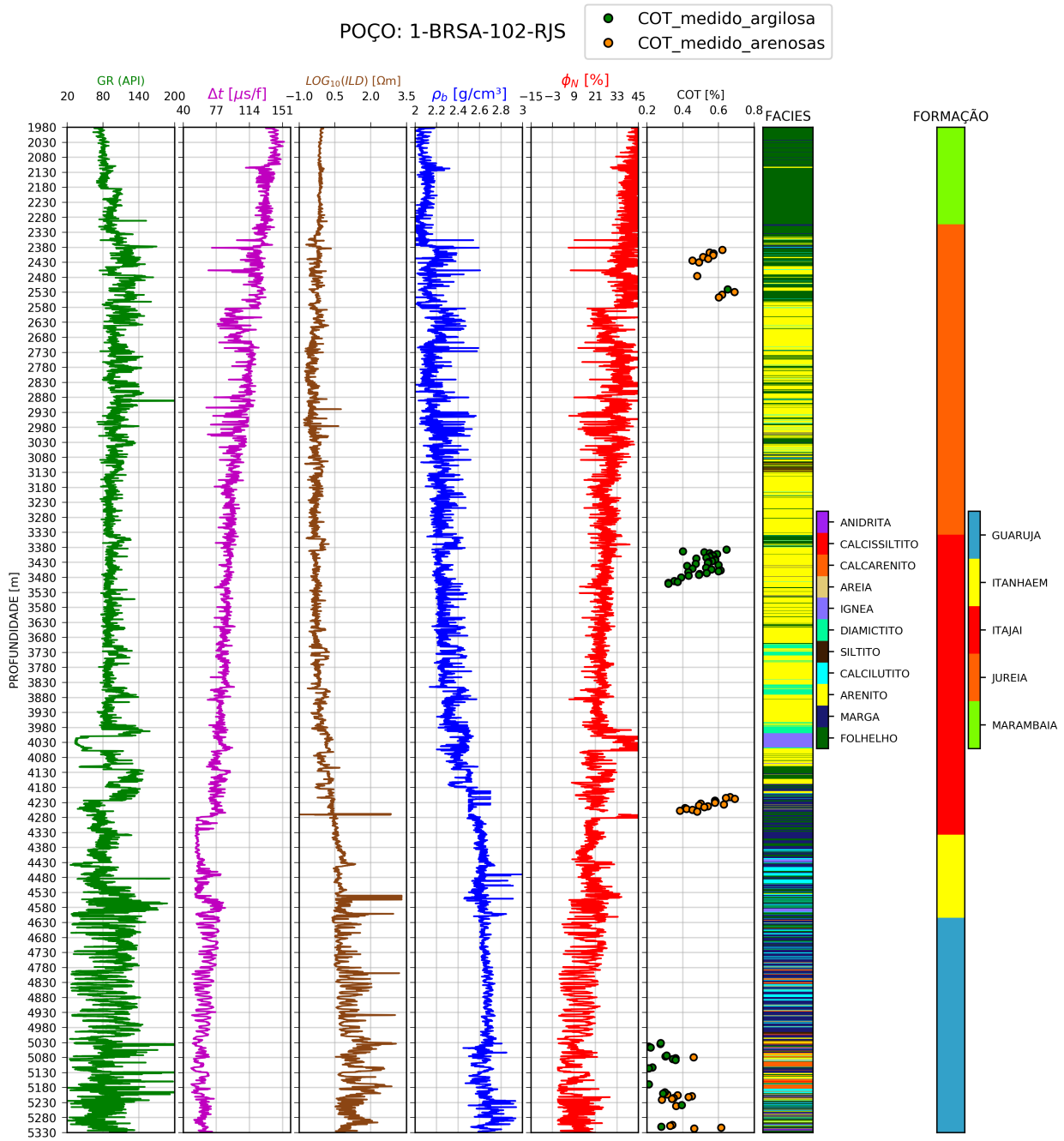


Figura 1.7: Dados do poço teste, utilizado para avaliação dos algoritmos.

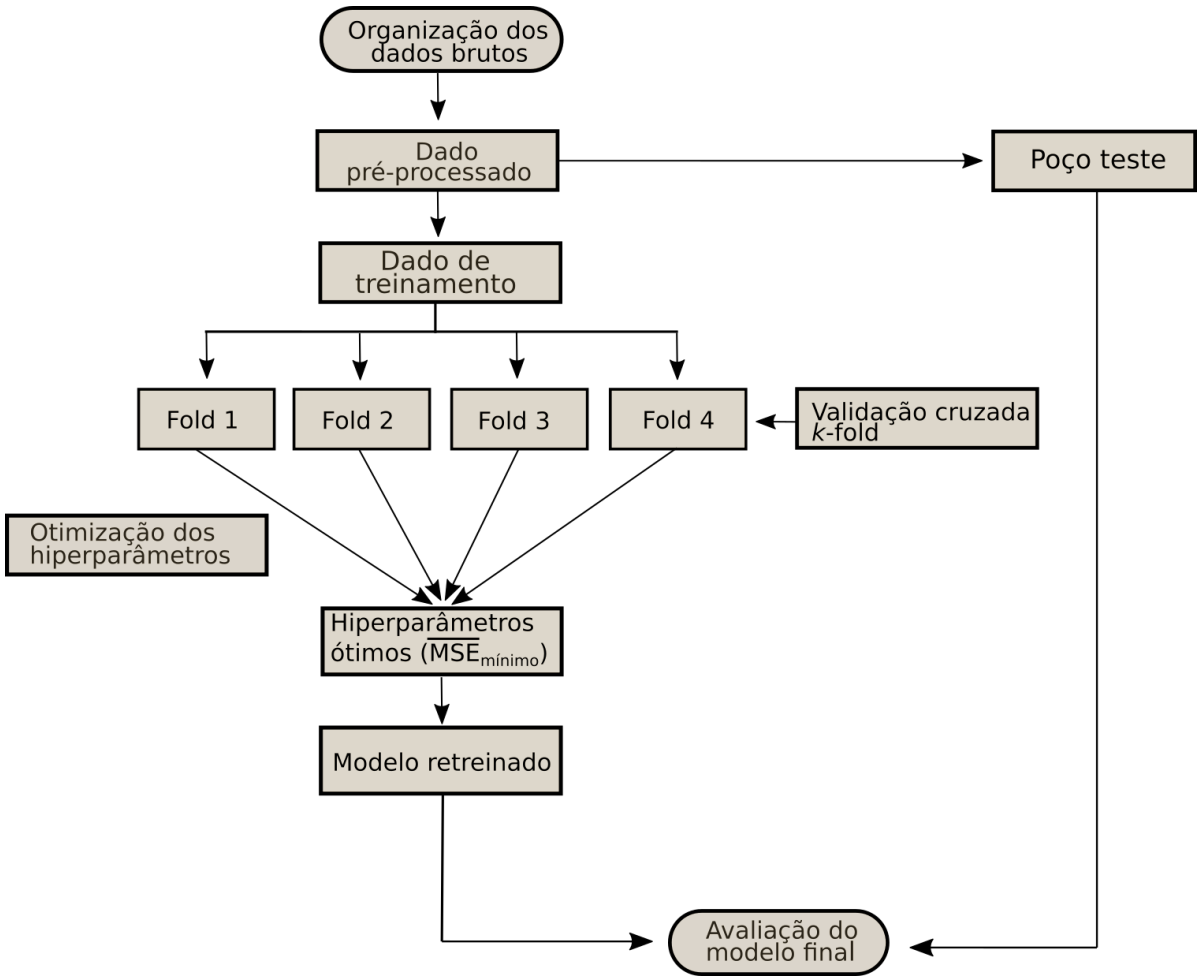


Figura 1.8: Floxograma exibindo o passo-a-passo da metodologia aplicada.

2

Rochas geradoras

Uma rocha geradora de petróleo é definida como qualquer rocha com capacidade de gerar e expelir quantidade suficiente de hidrocarbonetos para formação e acumulação de petróleo, na forma de óleo ou gás. Na verdade, qualquer rocha que contenha matéria orgânica (MO) produz hidrocarbonetos. No entanto, uma rocha com potencial para geração de hidrocarbonetos é aquela que se apresenta em estado imaturo de geração de petróleo em seu estado natural, mas é capaz de liberar quantidades significativas quando seu grau de maturação é "acelerado", ao ser submetida a processos de hidropirolização. Uma rocha geradora matura ou efetiva é caracterizada por apresentar petróleo formado e expulso para um reservatório, ativo ou inativo (Hunt, 1995). Tais rochas são, em sua maioria, rochas com granulação fina e ricas em matéria orgânica, tais como: folhelhos, calcilutitos, siltitos ou margas.

Os mecanismos pelos quais gás e óleo são formados variam de bacia para bacia, dependendo do tipo e abundância da matéria orgânica, fácies sedimentares, tectônica e condições físicas e químicas do paleoambiente deposicional, por exemplo. O modelo genérico que explica o processo de geração é bastante simples, contudo.

São três os fatores que qualificam e quantificam o potencial gerador de uma rocha fonte: seu volume, a porcentagem de matéria orgânica em massa e sua maturidade termal. O volume é uma função da espessura e da área de extensão da rocha geradora e está diretamente relacionado com a quantidade de matéria orgânica da rocha. Logo, quanto maior o volume da rocha fonte, uma maior quantidade de matéria orgânica poderá gerar um maior volume de hidrocarbonetos, seja ele gás ou óleo.

2.1 Teor de Carbono Orgânico Total

Em função das características geoquímicas necessárias para avaliar um sedimento como sendo de uma rocha geradora de hidrocarbonetos, a quantificação da matéria orgânica se apresenta como o primeiro parâmetro analisado. A quantidade de matéria orgânica é medida através do teor de carbono orgânico total (COT), expresso na forma de percentual em relação ao extrato seco, que reflete as condições de produção e preservação no ambiente deposicional (Espitalie et al., 1977), (Milner, 1982). A quantidade de matéria orgânica presente nos sedimentos ou rochas inclui tanto a matéria orgânica insolúvel, denominada de querogênio, como a matéria orgânica solúvel em solventes orgânicos, denominada de betume (Kvenvolden, 2006).

O procedimento experimental consiste inicialmente em tratar a amostra pulverizada com ácido clorídrico a fim de se liberar o carbono inorgânico (na forma de carbonato) e, após, medir a quantidade de CO_2 gerada quando a amostra é submetida ao processo de combustão. Os valores médios de COT para folhelhos geradores de hidrocarbonetos são de 2,0 % (em massa). No entanto, segundo Tissot e Welte (1984), o valor mínimo aceitável de COT para que uma rocha seja considerada como potencialmente geradora de petróleo é de 1 % para folhelhos e 0,5 % para carbonatos. Apesar do carbono orgânico estar associado majoritariamente à folhelhos ou folhelhos siltosos, pode também estar presente em rochas relativamente "limpas", como siltitos, arenitos e carbonatos (Crain, 2002).

Tabela 2.1: Avaliação da qualidade de Rochas Geradoras através do COT. Fonte: Peters e Cassa (1994).

COT (%)	**Potencial de Geração de Hidrocarbonetos**
$< 0{,}5$	Nenhum
0,5	Pobre
1 – 2	Razoável
2 – 5	Bom
> 5	Excelente

2.2 Predição de COT através de Perfis Geofísicos

Relações entre COT e dados petrofísicos, que incluem os perfis GR, ρ_b, Δt, ϕ_N e ILD, foram tratadas por Schmoker (1979), Fertl et al. (1980), Schmoker (1981), Meyer e Nederlof (1984), Fertl, Chilingar et al. (1988), Herron et al. (1988), Passey et al. (1990) e Bessereau et al. (1991). Assim sendo, os perfis acima mencionados, além de serem utilizados para inferirem na física das rochas, podem ser utilizados para informar ao geocientista sobre a geoquímica

das mesmas. Estes dados serão utilizados como dados de entrada para a estimativa dos valores de COT, tanto das rochas geradoras quanto das rochas reservatório, para as rochas da Bacia de Santos, ES.

3

Aprendizado de Máquina (AM)

Um eminente estatístico, George E.P. Box, escreveu em seu livro *Empirical Model Building and Response Surfaces*: "Essencialmente, todos os modelos (estatísticos) estão errados, mas alguns são úteis". Modelos estatísticos são uma descrição de um fenômeno real, utilizando-se de conceitos matemáticos; e como tal, eles são apenas uma simplificação da realidade.

A Teoria do Aprendizado Estatístico se refere a um vasto conjunto de ferramentas para a compreensão de dados. Estas ferramentas podem ser classificadas como supervisionadas e não supervisionadas, de modo que a maior parte dos modelos de aprendizado de máquina estão na categoria dos supervisionados. A aprendizagem não supervisionada nos permite abordar problemas com pouca ou nenhuma ideia do que esperar nos resultados, de modo que não temos saídas rotuladas a priori. Um exemplo é o agrupamento dos dados com base em relações entre as variáveis. Além disso, também pode ser usada para reduzir o número de dimensões em um conjunto de dados, ou ainda na detecção de tendências.

Em AM supervisionado, algoritmos são utilizados para induzir modelos preditivos por meio da observação de um conjunto de objetos rotulados (Von Luxburg e Schölkopf, 2011), tipicamente referenciado como conjunto de treinamento. Os rótulos contidos em tal conjunto correspondem a classes ou valores obtidos por alguma função desconhecida. Desse modo, um algoritmo de classificação buscará produzir um classificador capaz de generalizar as informações contidas no conjunto de treinamento, com a finalidade de classificar, posteriormente, objetos cujo rótulo seja desconhecido.

Formalmente, um conjunto de dados de treinamento pode ser definido como uma coleção de tuplas $\{(x_i, y_i)\}_{i=1}^{n}$, onde, em cada tupla, x_i indica um vetor descrito por m características e y_i indica o rótulo correspondente a x_i. Quando os valores de y_i são definidos por uma

quantidade limitada de valores discretos, tem-se um problema de classificação. Quando tais valores são contínuos (foco desta pesquisa), tem-se um problema de regressão.

À seguir, serão apresentados os algoritmos aqui utilizados, descrevendo-se de maneira sucinta a matemática e os parâmetros empregados na construção de cada um deles.

3.1 Máquina de Vetores de Suporte

O algorítimo Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*) é uma generalização não-linear do algoritmo *Generalized Portrait*, desenvolvido na Rússia nos anos sessenta (Vapnik, 1963; Vapnik e Chervonenkis, 1964). Assim sendo, está firmemente alicerçado no quadro de aprendizado da estatística teórica, ou *teoria VC* (Vapnik-Chervonenkis), que foi desenvolvida ao longo das últimas três décadas por Vapnik e Chervonenkis (1974), Vapnik (1982), Cortes e Vapnik (1995).

Em suma, a *teoria VC* é caracterizada por propriedades de aprendizado de máquina que lhe permite generalizar bem em dados fora do conjunto de treinamento. A teoria foi originalmente desenvolvida com base em um problema de bipartição separável, nos Laboratórios AT & T Bell. O SVM implementa um algoritmo de aprendizado útil para reconhecer padrões sutis em conjuntos de dados complexos.

O SVM pode ser utilizado tanto para resolver problemas de classificação (variável alvo categórica), quanto para problemas de regressão (variável alvo contínua). Em problemas de classificação, é comum utilizar-se a notação SVC (*Support Vector Classification*), ao passo que em problemas de regressão, é comum utilizar-se a notação SVR (*Support Vector Regression*). Uma vez que a variável alvo deste estudo é contínua, utilizaremos a notação SVR para fazer referência ao algoritmo SVM daqui em diante.

Como não são conhecidas as não-linearidades presentes e a complexidade intrínseca da maior parte dos problemas, os algoritmos de otimização e as ferramentas estatísticas utilizadas para a seleção de modelos podem induzir ao uso de algoritmos com baixa capacidade de generalização. Sendo assim, as principais vantagens do SVR em suas aplicações são:

- Elevada capacidade de generalização, evitando o sobreajuste[1];
- Robustez em grandes dimensões, possibilitando aplicação de SVRs em espaços de características de grandes dimensões;

[1]Ocorre quando o algoritmo performa muito bem no conjunto de treinamento, mas não é capaz de performar bem em dados novos; ou seja; o modelo perde a capacidade de generalização.

- Convexidade da função objetivo, o que implica na otimização de uma função quadrática; ou seja; que possui apenas um mínimo;
- Teoria bem estabelecida dentro da Matemática e Estatística, posto que foi desenvolvida por cientistas destas áreas.

Uma das principais características do SVR é que, em vez de minimizar o erro no conjunto de treinamento, tenta minimizar o erro generalizado associado, a fim de obter um "desempenho generalizado". Este erro de generalização associado é a combinação do erro de treinamento e um termo de regularização, que controla a complexidade do espaço de hipóteses.

Ao contrario do SVC, o SVR propõe determinar um hiperplano ótimo em que as amostras de treinamento estejam tão próximas quanto possível, não importando em qual dos lados da superfície os pontos se localizam e sim que a distância para a superfície seja a mínima possível. Porém, mesmo com propósitos opostos, ambos buscam estabelecer uma função com máxima capacidade de generalização (Lima et al., 2004).

O objetivo do SVR é encontrar uma função $f(\vec{x})$, com uma margem de erro caracterizada pelo intervalo $[y_i - \varepsilon,\ y_i + \varepsilon]$, onde, desvios são permitidos desde que não ultrapassem a margem especificada. Assim, seja $x_i \in \mathbb{R}^p$, $i = 1, 2, \ldots, n$ e um vetor y $\in \mathbb{R}^n$, ε-SVR, e assumindo funções da forma $f(\vec{x}) = \langle \vec{w} \cdot \vec{x} \rangle + b$, as seguinte restrições devem ser satisfeitas:

$$\begin{aligned} y_i - \varepsilon \leq \langle \vec{w} \cdot \vec{x} \rangle + b \Rightarrow y_i - \langle \vec{w} \cdot \vec{x} \rangle - b \leq \varepsilon, \\ y_i + \varepsilon \geq \langle \vec{w} \cdot \vec{x} \rangle + b \Rightarrow \langle \vec{w} \cdot \vec{x} \rangle + b - y_i \leq \varepsilon \end{aligned} \tag{3.1}$$

Uma vez que a função $f(\vec{x})$ deve satisfazer as restrições de erro $|f(\vec{x_i}) - y_i| \leq \varepsilon$, para todo $i = 1, 2, ..., n$, resolve-se o problema de otimização em sua forma primal como:

$$\begin{aligned} \min \frac{1}{2}\|\vec{w}\|^2 & \\ \text{sujeito a} \quad & y_i - \langle \vec{w} \cdot \vec{x_i} \rangle - b \leq \varepsilon \\ & \langle \vec{w} \cdot \vec{x_i} \rangle + b - y_i \leq \varepsilon \\ & \forall i = 1, \ldots, n, \end{aligned} \tag{3.2}$$

onde $\vec{w} \in \mathbb{R}^n$ e $b \in \mathbb{R}$ são as incógnitas do problema.

Este problema supõe a existência de uma função $f(\vec{x})$ que aproxima todos os pares $(\vec{x_i}, y_i)$ com uma precisão ε que deve ser especificada a priori. No entanto, nem sempre é

possível garantir que isto ocorra, visto que existem pontos que poderão violar tais restrições. Assim, introduz-se uma função, conhecida como função de perda (*loss function*), de modo que seja possível empreender variáveis de folgas (*slack variables*) não-negativas (ξ_i e ξ_i^*), cuja finalidade é penalizar dados que se situem fora da margem $|f(\vec{x_i}) - y_i| \leq \varepsilon$. Considerando essa função é possível trabalhar com um número limitado de erros, que em outras condições tornariam o problema de otimização inviável (Dias, 2007).

A função de perda, denominada $\varepsilon - Insensitive$, é descrita por:

$$\begin{cases} |\xi|_\varepsilon = 0, & \text{se } |\xi| \leq \varepsilon. \\ |\xi|_\varepsilon = |\xi| - \varepsilon, & \text{caso contrário.} \end{cases} \tag{3.3}$$

A Figura 3.1 mostra a representação gráfica da situação proposta pela função (3.3) na qual apenas os pontos localizados fora da região tracejada contribuem para o valor da função de custo.

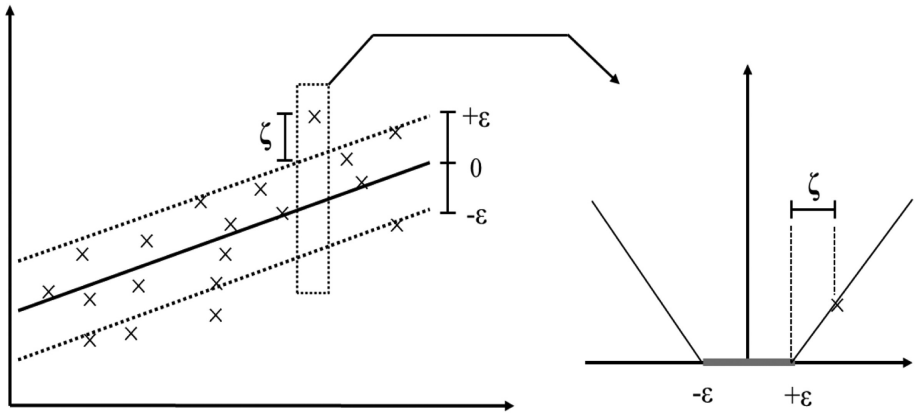


Figura 3.1: Smola e Sclkopf (2004)

As variáveis de folga ξ_i e ξ_i^* estão associadas aos dados que se situam fora das margens, tanto da margem inferior quanto da superior. Assim, pode-se reescrever o problema primal como sendo:

$$\min \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3.4}$$

$$\begin{aligned} \text{sujeito a} \quad & y_i - \langle \vec{w} \cdot \vec{x_i} \rangle - b \leq \varepsilon + \xi_i \\ & \langle \vec{w} \cdot \vec{x_i} \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, \quad \forall i = 1, \ldots, n, \end{aligned} \tag{3.5}$$

onde $\vec{w} \in \mathbb{R}^n$ e $b \in \mathbb{R}$ são as incógnitas do problema e C é denominada constante de regularização, pois faz o balanço entre a função de minimização e a margem (valor dos desvios), de modo que, desvios maiores que ε são tolerados, e desta forma, sendo tolerados erros. Os pontos entre as margens ($-\varepsilon$ e $+\varepsilon$) não sofrem esta penalização, apenas os valores fora das margens são penalizados, conforme a função de perda $\varepsilon - Insensitive$ (Smola e Sclkopf, 2004).

Para resolver o problema de otimização apresentado em 3.4, sujeito às restrições em 3.5, o modelo primal é expresso em sua forma dual, por meio da função de Lagrange, proporcionando assim eficiência e flexibilidade ao algoritmo. Assim, segundo Smola e Sclkopf (2004), para transformar a Equação 3.4 (forma primal) na formulação dual, são introduzidos Multiplicadores de Lagrange não negativos, resultando em:

$$\begin{aligned} L := \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) - \sum_{i=1}^{n}(\alpha_i\xi_i + \alpha_i^*\xi_i^*) \\ - \sum_{i=1}^{n}\lambda_i(\varepsilon + \xi_i - y_i + \langle\vec{w}.\vec{x_i}\rangle + b) \\ - \sum_{i=1}^{n}\lambda_i^*(\varepsilon + \xi_i^* + y_i - \langle\vec{w}.\vec{x_i}\rangle - b) \end{aligned} \tag{3.6}$$

Em que L é a função de Lagrange e $\lambda_i, \lambda_i^*, \alpha_i, \alpha_i^*$ são os Multiplicadores de Lagrange (Smola e Sclkopf, 2004). Para encontrar a solução ótima, é necessário minimizar a função de Lagrange, realizando a maximização dos seus multiplicadores ($\lambda_i, \lambda_i^*, \alpha_i, \alpha_i^*$) e minimização de $\vec{w}$ e b (Chamasemani e Singh, 2011), resultando no problema de otimização. Assim, chega-se à formulação apresentada abaixo em sua forma dual (Gunn et al., 1998):

$$\min -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i.x_j\rangle - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n}y_i(\alpha_i + \alpha_i^*) \tag{3.7}$$

$$\text{sujeito a} \quad \begin{cases} \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \quad \forall i = 1, \cdots, n \\ \alpha_i, \alpha_i^* \in [0,\ C] \end{cases} \tag{3.8}$$

A computação de b e $\vec{w}$ pode ser realizada explorando as chamadas condições de Karush-Kuhn-Tucker (KKT) (Karush, 1939), que são condições utilizadas na solução de problemas computacionais não-lineares. Tais condições afirmam que no ponto de solução ótima do problema de otimização, o produto interno entre variáveis duais e as restrições deve desaparecer. Assim, adaptando para o problema do SVR:

$$\alpha_i(\varepsilon + \xi_i - y_i + \langle \vec{w}.\vec{x_i} \rangle + b) = 0, \qquad \forall i = 1, \cdots, n, \tag{3.9}$$

$$\alpha_i^*(\varepsilon + \xi_i^* - y_i + \langle \vec{w}.\vec{x_i} \rangle + b) = 0, \qquad \forall i = 1, \cdots, n, \tag{3.10}$$

$$(C - \alpha_i)\xi_i = 0, \qquad \forall i = 1, \cdots, n, \tag{3.11}$$

$$(C - \alpha_i^*)\xi_i^* = 0, \qquad \forall i = 1, \cdots, n, \tag{3.12}$$

$$\alpha_i \alpha_i^* = 0, \qquad \forall i = 1, \cdots, n, \tag{3.13}$$

$$\xi_i \xi_i^* = 0, \qquad \forall i = 1, \cdots, n \tag{3.14}$$

Pela equação 3.13, observa-se que não é possível a existência de um conjunto de variáveis duais α_i e α_i^*, em que ambos os valores sejam não nulos.

Das relações 3.9 e 3.10, obtém-se que apenas os pontos de treinamento em que $|f(\vec{x_i}) - y_i| \geq \varepsilon$ estão associados aos multiplicadores de Lagrange não nulos. Tais pontos estão localizados sobre as margens $+\varepsilon$ e -ε ou fora da região delimitada. Estes valores são os únicos a serem utilizados no cálculo do vetor dos pesos $\vec{w}$ e por isso são chamados de vetores suporte.

Das condições 3.11 e 3.12, obtêm-se que, os vetores suportes cujo $\alpha_i = C$ ou $\alpha_i^* = C$, têm variáveis de folga respectivamente ξ_i e ξ_i^*, não nulas, caracterizando pontos localizados fora das margens, relacionados à erros associados ao modelo.

O vetor de pesos pode então ser calculado pela expressão (Gunn et al., 1998):

$$\vec{w^*} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\vec{x_i} \tag{3.15}$$

e a função de decisão é dada por:

$$f(\vec{x}) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\langle \vec{x_i} \cdot \vec{x} \rangle + b \tag{3.16}$$

onde o viés (*bias*) é determinado por meio da relação:

$$b^* = -\frac{1}{2}\langle \vec{w^*}, (\vec{x_r} + \vec{x_s}) \rangle \tag{3.17}$$

em que $\vec{x_r}$ e $\vec{x_s}$ são quaisquer vetores suporte de cada classe, satisfazendo $\alpha_r, \alpha_s > 0$.

O SVR pode ser utilizado tanto para ajustes lineares quanto para ajustes não lineares. No último caso, é necessário o mapeamento (ϕ) dos dados (conjunto de treinamento) do

seu espaço original para o espaço de características (*feature space*), de maior dimensão. Este procedimento é feito através de funções *kernel* (núcleo), de modo que,a única coisa necessária é o produto escalar dos pontos do conjunto de dados. Segundo (Gunn et al., 1998), para obter o produto escalar dos pontos, são utilizadas funções que permitem calcular este procedimento. Essas funções são chamadas de funções *Kernel*.

Uma função é dita ser uma função *Kernel*, se ela satisfaz as condições estabelecidas pelo Teorema de Mercer (Carvalho, 2005). Assim, seja K uma matriz positiva definida, o Teorema de Mercer afirma que K é função uma função *Kernel*, se a matriz K é obtida por:

$$K = K_{ij} = K(\vec{x_i}, \vec{x_j}) \tag{3.18}$$

Não se pode dizer que uma função *Kernel* é melhor que outra. O desempenho de uma função vai depender do tipo de banco de dados utilizado na aplicação. Sendo assim, é possível testar diversos tipos de *Kernel* a fim de encontrar aquela que melhor se adapte ao problema estudado (Nguyen et al., 2006).

Com a aplicação da função *Kernel* no problema de otimização demonstrado na Equação 3.7, tem-se (Basak et al., 2007):

$$\min -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\vec{x_i}, \vec{x_j}) - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*) \tag{3.19}$$

Deste modo, o cálculo do vetor dos pesos ($\vec{w}$) é redefinido:

$$\vec{w} = \sum_{i=1}^{n} y_i(\alpha_i + \alpha_i^*).\phi(\vec{x_i}) \tag{3.20}$$

Enquanto isso, o cálculo do viés é reformulado para:

$$b^* = -\frac{1}{2}(\alpha_i - \alpha_i^*).k(\vec{x_r}, \vec{x_s}) \tag{3.21}$$

Que, conforme Smola e Sclkopf (2004), resulta na seguinte função de regressão da SVR para dados não lineares: :

$$f(\vec{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*).k(\vec{x_i}, \vec{x}) + b \tag{3.22}$$

Na predição do COT a partir dos perfis, utilizando a SVR, a camada de entrada consiste nos perfis geofísicos, a camada escondida consiste nas função do *Kernel*, e a camada de saída consiste nos resultados da predição do COT. A figura 3.2 abaixo mostra o diagrama esquemático do processo acima citado, quando utilizados os 5 perfis para o treinamento. Primeiramente, os dados dos perfis ($\vec{x_i}$) são projetados em um espaço de características multidimensional (*i.e.*, espaço de Hilbert), através de uma função não-linear $\phi(\vec{x_i})$; assim, um problema linear ou não linear é transformado em uma estrutura multidimensional no espaço de características, com uma função de decisão de regressão do hiperplano. Finalmente os valores de COT são gerados.

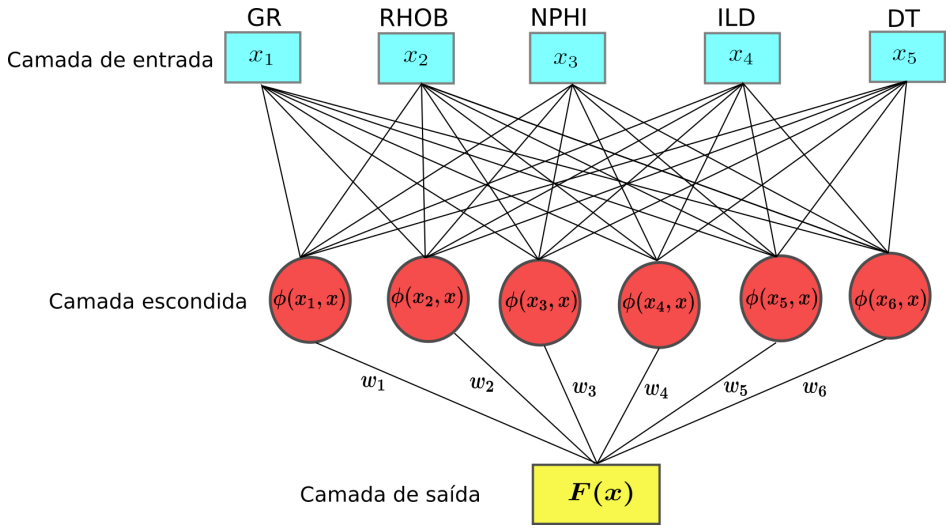


Figura 3.2: Diagrama esquemático da predição de COT a partir dos 5 perfis, utilizando o algoritmo SVR.

A função kernel utilizada neste trabalho será a Função de Base Radial (RBF, do inglês *Radial Basis Function*). Está função é descrita matematicamente como segue:

$$K(\vec{x_i}, \vec{x}) = e^{\frac{-\|\vec{x_i} - \vec{x}\|^2}{\sigma^2}} = e^{-\gamma\|\vec{x_i} - \vec{x}\|^2} \tag{3.23}$$

Ao treinar o algoritimo SVR com o *kernel* RBF, dois parâmetros devem ser considerados: C e γ. O parâmetro C, comum à todos os kernels SVM, negocia os erros de classificação dos exemplos no treinamento com a simplicidade da superfície de decisão. Um C baixo facilita a superfície de decisão, enquanto um C alto visa classificar todos os exemplos de treinamento corretamente. O parâmetro γ define quanta influência um único exemplo de treinamento

tem. Quanto maior o valor de γ, mais próximos devem ser outros exemplos, de modo que sejam afetados.

3.2 Regressão Linear Múltipla

A análise de regressão foi desenvolvida por Sir Francis Galton no final do século XIX. É um método estatístico que utiliza a relação entre duas ou mais variáveis quantitativas, de modo que uma variável resposta ou resultado possa ser predita a partir de outra, ou outras. Esta metodologia é amplamente utilizada no mundo dos negócios, em ciências comportamentais e sociais, ciências biológicas, e muitas outra disciplinas (Neter et al., 1996).

O modelo de regressão linear múltipla (RLM) é um dos métodos estatísticos mais utilizados entre todos (Neter et al., 1996). Deve ser aplicada quando apenas uma variável regressora não é capaz de explicar totalmente o fenômeno estudado, devendo-se portanto, introduzir no modelo outras variáveis dependentes, de modo que:

$$Y_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + ... + \alpha_{p-1} X_{i,p-1} + \epsilon_i, \qquad i \in N^*, \tag{3.24}$$

é chamado modelo de regressão linear múltipla com $p-1$ variáveis regressoras; onde α_0, α_1, α_2 ...α_{p-1} representam os coeficientes de regressão (α_0 é também chamado de intercepto). A função Y_i no modelo linear múltiplo é uma superfície de resposta, e descreve um hiperplano no espaço p-dimensional das variáveis de entrada X_i.

Emprega-se esta técnica em problemas cuja relação ente as variáveis explicativas e a dependente é linear. Para achar a solução, é comum utilizar o método dos mínimos quadrados ordinários (MMQ). Este método é uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados, buscando minimizar a soma dos quadrados dos resíduos (diferença entre valor observado e estimado). Esta técnica será melhor discutida adiante.

3.2.1 Ajuste de curvas pelo Método dos Mínimos Quadrados

O método dos mínimos quadrados ordinários (MMQ) é um método de ajuste de pontos a uma reta, e se baseia em que a reta resultante do ajuste seja tal que a soma dos quadrados das distâncias verticais dos pontos à reta seja mínima. Esta reta recebe o nome de retas dos mínimos quadrados, reta de regressão, ou reta de regressão estimada.

Seja e_i o resíduo ou desvio que a regressão linear obtém no ponto i, então:

$$e_i = Y_i - \hat{Y}_i \tag{3.25}$$

onde Y_i é o valor observado e $\hat{Y}_i$ é o melhor valor estimado pela regressão linear. Assim, segundo Neter et al. (1996), o MMQ requer que consideremos a soma dos quadrados dos n desvios. Tal critério é denotado por Q:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{3.26}$$

Pra o caso de regressão linear com $p-1$ variáveis regressoras (3.24) e escrevendo de maneira diferencial em relação aos coeficientes, temos:

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha_0} = -2 \displaystyle\sum_{i=1}^{n} [Y_i - \alpha_0 - \alpha_1 X_{i1} - \alpha_2 X_{i2} - ... - \alpha_{p-1} X_{i,p-1}] \\ \dfrac{\partial Q}{\partial \alpha_1} = -2 \displaystyle\sum_{i=1}^{n} [Y_i - \alpha_0 - \alpha_1 X_{i1} - \alpha_2 X_{i2} - ... - \alpha_{p-1} X_{i,p-1}] X_{i1} \\ \vdots \\ \dfrac{\partial Q}{\partial \alpha_j} = -2 \displaystyle\sum_{i=1}^{n} [Y_i - \alpha_0 - \alpha_1 X_{i1} - \alpha_2 X_{i2} - ... - \alpha_{p-1} X_{i,p-1}] X_{ji} \end{cases}$$

com $j = 0, 1, 2, ..., p$. Após igualar as derivadas a zero para obter os "pontos críticos" e rearrumá-las, obtêm-se o sistema de equações normais:

$$\begin{cases} \mathrm{n}\alpha_0 + \alpha_1 \displaystyle\sum_{i=1}^{n} X_{i1} + \alpha_2 \sum_{i=1}^{n} X_{i2} + ... + \alpha_{p-1} \sum_{i=1}^{n} X_{i,p-1} = \sum_{i=1}^{n} Y_i \\ \alpha_0 \displaystyle\sum_{i=1}^{n} X_{i1} + \alpha_1 \sum_{i=1}^{n} X_{i1}^2 + \alpha_2 \sum_{i=1}^{n} X_{i1} X_{i2} + ... + \alpha_{p-1} \sum_{i=1}^{n} X_{i1} X_{i,p-1} = \sum_{i=1}^{n} X_{i1} Y_i \\ \vdots \\ \alpha_0 \displaystyle\sum_{i=1}^{n} X_{i,p-1} + \alpha_1 \sum_{i=1}^{n} X_{i,p-1} X_{i1} + \alpha_2 \sum_{i=1}^{n} X_{i,p-1} X_{i2} + ... + \alpha_{p-1} \sum_{i=1}^{n} X_{i,p-1}^2 = \sum_{i=1}^{n} X_{i,p-1} Y_i \end{cases}$$

Resolvendo-se este sistema de equações encontra-se os coeficientes $\alpha_0, \alpha_1, ...\alpha_{p-1}$, das variáveis regressoras.

Nota: Assume-se que o erro (e_i) tem distribuição normal (Gaussiana), com média zero e variância σ^2 constante; ou seja; $erro \approx N(0, \sigma^2)$, e que valores extremos *outliers* são raros, mas ainda assim podem ocorrer. O método de ajuste por MMQ possui alta sensibilidade

de valores extremos. Estes pontos extremos possuem grande influência no ajuste porque ao se elevar os resíduos ou quadrado a magnitude do efeito é potencializada. Para minimizar a influência de *outliers*, pode-se ajustar os dados por regressão linear robusta.

3.3 Floresta Aleatória

Algoritmos de Floresta Aleatória (RF, do inglês *Random Forests*) são métodos de aprendizado em conjunto, aplicados tanto para classificação, quanto regressão, além de outras tarefas que operam na construção de uma multiplicidade de árvores de decisão no momento do treinamento e gerando a classe que é o modo das classes (classificação) ou predição média (regressão) das árvores individuais (Ho, 1995; Barandiaran, 1998). Algoritmos de RF corrigem o "hábito" de sobreajustes (*overfitting*) do seu conjunto de treinamento, muito comum nos algoritmos de Árvores de Decisão (*decision trees*).

O primeiro algoritmo de Floresta Aleatória foi criado por Ho (1995), usando o método do subespaço aleatório, apresentado por (Barandiaran, 1998), que na formulação de Ho é uma maneira de implementar a abordagem de "discriminação estocástica" à classificação, proposta por Eugene Kleinberg (Kleinberg, 1990; Kleinberg et al., 1996). Foi posteriormente aprimorada por Breiman (2001), com base na combinação de um grande conjunto de árvores de decisão.

Nos algoritmos de árvore padrões, cada nó é dividido usando a melhor divisão entre todas as variáveis. No algoritmo de RF, cada nó é dividido usando o melhor entre um sub-conjunto de preditores escolhidos aleatoriamente nesse nó. Esta estratégia um tanto quanto contra-intuitiva, acaba por performar muito bem em comparação com muitos outros classificadores, incluindo análise discriminante, SVM e até mesmo redes neurais, além de ser robusta contra sobreajuste (Breiman, 2001). Além disso, é muito fácil de usar, no sentido de que possui apenas dois parâmetros principais: o número de variáveis no subconjunto aleatório em cada nó e o número de árvores na floresta, e geralmente não é muito sensível aos seus valores. O algoritimo RF (tanto para classificação quanto para regressão) funciona como segue:

1. Extrai-se n conjuntos de amostras *Bootstrap* [2] do dado original.

2. Para cada um dos conjuntos de amostras *Bootstrap*, uma árvore de classificação ou regressão podada é construída, com a seguinte modificação: em cada nó, ao invés de escolher a melhor divisão entre todos os preditores, h amostras de variáveis são

[2]Na literatura estatística, o método *bootstrap* de amostragem remonta ao trabalho de Efron (1982).

selecionadas aleatoriamente e a melhor divisão é escolhida entre elas, onde n é o número de árvores na floresta e h é o número de atributos utilizados para construir cada árvore.

3. Dados novos (chamados de *out-of-bag*) são preditos pela média das predições das n árvores e são utilizados para estimar uma taxa de erro, chamada estimativa de erro *out-of-bag* (ERR_{OOB}), numa técnica similar ao método de validação cruzada *leave-one-out*, mas sem nenhum custo computacional extra. Este erro é calculado como segue:

   i. Em cada iteração *Bootstrap*, os elementos *out-of-bag* são preditos pela árvore construída, utilizando-se a amostra *Bootstrap* X_i ;

   ii. para o i-ésimo elemento (y_i) do conjunto de dado de treinamento X, todas as árvores são consideradas, nas quais o i-ésimo elemento é *out-of-bag*. Na média, cada elemento de X é *out-of-bag* em 36 % das n iterações. Com base nas árvores aleatórias, uma previsão agregada g_{OOB} é desenvolvida. A estimativa do erro *out-of-bag* é computada como: $ERR_{OOB} = (1/n)\sum_{i=1}^{n}[y_i - g_{OOB}(X_i)]^2$.

   A explicação para a quantidade de 36 % de dados deixados de fora do treinamento (*out-of-bag*), advém da probabilidade de que em t seleções aleatórias e com reposição, a probabilidade de um indivíduo não ser selecionado nenhuma vez é dada por:

$$\begin{gathered}\prod_{i=1}^{t}\frac{t-1}{t} = (1-\frac{1}{t})^t \\ \text{quando } t \to \infty,\ \ (1-\frac{1}{t})^t \to e^{-1} \approx 0,367\end{gathered} \tag{3.27}$$

Apesar do algoritmo RF possuir algumas dezenas de hiperparâmetros, os principais a considerar e que foram utilizados neste trabalho são:

- n: número de árvores na floresta. Quanto maior, melhor, mas também exigirá mais tempo de processamento. Além disso, observa-se que os resultados deixarão de melhorar significativamente a partir de um número crítico de árvores. O ideal é começar com um número de árvores 10 vezes maior que o número de variáveis (atributos) e ir ajustando à medida em que se ajusta os outros hiperparâmetros.

- m: representa a profundidade de cada árvore na floresta. Quanto mais profunda a árvore, mais divisões ela terá e, consequentemente, também será capturada mais informação sobre os dados.

- l: número mínimo de amostras necessárias para estar em um nó da árvore. Um ponto de divisão em qualquer profundidade só será considerado se deixar pelo menos "l" amostras de treinamento em cada um dos ramos esquerdo e direito. Isso pode ter o efeito de suavizar o modelo, especialmente em regressão.

- s: número mínimo de amostras necessárias para dividir um nó interno. Pode variar de uma até todas as amostras em cada nó.

- h: número de atributos (variáveis) a serem considerados ao procurar a melhor divisão (nó) para cada árvore. Quando se tem poucos preditores relevantes, aumentar o valor de h geralmente melhora o desempenho do modelo, pois aumenta-se a chance de variáveis com assinaturas mais relevantes serem consideradas em cada nó. No entanto, se há muitos preditores relevantes, um menor valor de h talvez performe melhor, pois aumenta a diversidade das árvores individuais, que é uma das principais característica do Floresta Aleatória.

3.4 Pré-processamento

É comum e necessário que, antes da aplicação dos algoritmos, se faça um trabalho de seleção dos dados na fase de pré-processamento. Nesta fase, organiza-se os dados e aplica-se transformações não lineares (logarítmica, inversa, etc) e/ou redimensionamentos (padronização e normalização), quando necessários. Tanto a padronização quanto a normalização possuem o mesmo objetivo: redimensionar as variáveis, colocando-as na mesma ordem de grandeza. Na padronização, a ideia é modificar a escala de um atributo, sem contudo alterar sua distribuição . Na normalização, subtrai-se os valores dos atributos da média e divide-os pelo desvio padrão, tornando os dados maios ou menos normalmente distribuídos (média 0 e desvio padrão 1).

De modo geral, os dados em sua forma bruta não irão resultar na melhor performance de um algoritmo, sendo necessário considerar características relacionadas à dimensão do conjunto de dados (número de observações e de preditores disponíveis), se a variável resposta é categórica ou contínua, balanceada ou desbalanceada, simétrica ou assimétrica e se os preditores são contínuos ou categóricos. Além disso, é necessário verificar se as variáveis são correlacionadas, se as escalas são diferentes, presença de valores faltantes e dados esparsos (Kuhn e Johnson, 2013; Raschka e Mirjalili, 2017). Muitos elementos utilizados na função objetivo dos algoritmos de AM, como o kernel RBF do SVM ou regularização L1 e L2 de modelos lineares, assumem que todos os atributos estão centrados em 0 e possuem variância[3]

[3]O desvio padrão é a raiz quadrada da variância; ou seja; são proporcionais.

de mesma ordem de grandeza. Deste modo, se um atributo possui variância que é algumas ordens de magnitude maior que outro, pode dominar sobre a função objetivo e tornar o estimador inapto a aprender corretamente a partir de outros atributos, o oposto do que seria esperado. O algoritmo floresta aleatória não necessita destas transformações monotônicas para uma boa convergência.

Nesta dissertação, foi utilizado o método *StardardScaler*, da biblioteca *sklearn*. Este método trás a média das distribuições das variáveis para 0 e o desvio padrão para 1, além de normalizar os dados, fazendo com que cerca de 68% dos valores ficam entre -1 e 1. A Figura 3.3 exibe os dados não padronizados (a) e os dados padronizados (b). É possível ver em (a) que as distribuições e as distâncias relativas estão muito divergentes. O perfil ρ_b, por exemplo, varia de 2 a 3 g/cm^3, ao passo que o GR passa dos 170, em unidades API; ou seja; são números com quase duas ordens de grandeza de diferença.

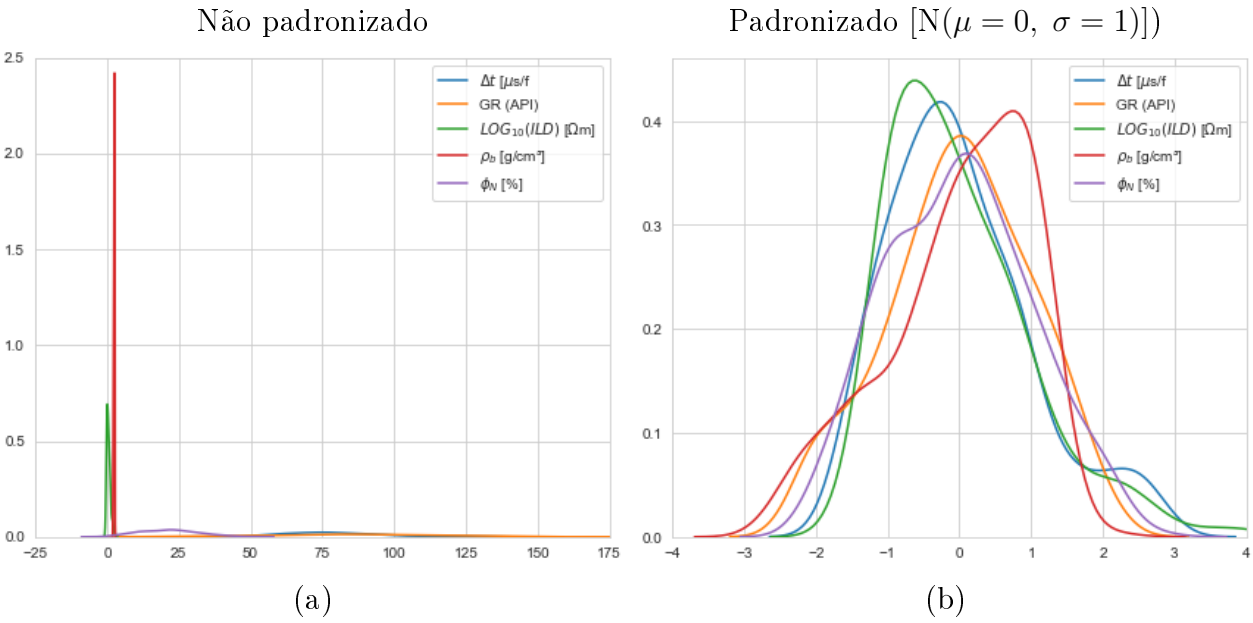


Figura 3.3: Em (a) estão representados os dados não padronizados, ao passo que em (b) são exibidos os dados padronizados, com cada variável assumindo média 0 e desvio padrão 1.

3.4.1 Seleção de Atributos

A seleção de atributos (SA) é um tema de pesquisa e desenvolvimento constante desde os anos 70, nas áreas de reconhecimento de padrões, aprendizado de máquina e mineração de dados (Yu e Liu, 2003), representando um papel importante na etapa de pré-processamento. Executar a seleção de atributos, permite, por exemplo, a ordenação das variáveis segundo algum critério de importância, redução da dimensionalidade do espaço de busca de atributos,

além da remoção de dados contendo ruídos, entre outras aplicações.

Medir o quão bom um determinado atributo é segundo um critério de avaliação, é uma tarefa importante na SA. Em outras palavras, avaliar como ele interage com o algoritmo de aprendizado. Essa interação pode ser subdividida em, basicamente, três abordagens (Kohavi, John et al., 1997):

1. *embedded*, a qual é embutida no próprio algoritmo de aprendizado;
2. filtro, a qual é utilizada para filtrar atributos durante um passo de pré-processamento, sem levar em conta o algoritmo de aprendizado que utilizará esse subconjunto de atributos, e
3. *wrapper*, a qual emprega o próprio algoritmo de aprendizado como uma "caixa preta".

A técnica de SA utilizada nesta dissertação foi um misto entre "força bruta" e *wrapper*. O termo força bruta aqui, refere-se ao fato de que foram feitas combinações de N atributos disponíveis, de modo a tentar achar o subconjunto de atributos que melhor se relacionasse com os dados de COT. Assim, dados N atributos, é possível fazer uma combinação de 2^N subconjuntos distintos de atributos. No entanto, este método é impraticável em conjuntos de dados com grande número de observações e de atributos. Devido à pouca quantidade de atributos utilizados nesta pesquisa (5 perfis geofísicos), bem como à relativamente pequena quantidade de dados de COT, esta metodologia foi possível ser empregada. Assim, ficou-se com uma combinação total de 32 (2^5) subconjuntos disponíveis para serem avaliados. Para realizar as combinações de atributos, foi empregado o algoritmo SVR; ou seja; aqui entra o método *wrapper*.

O pseudo-código representado em Algorithm 1 foi aplicado em todos os subconjuntos de forma iterativa, de modo que o subconjunto de atributos que retornou o menor $\overline{MSE}$ foi selecionado. Um problema que surge é que o método *wrapper* é totalmente dependente do algoritmo utilizado, de modo que os melhores atributos para o SVR podem não ser para o RF ou para a RLM; ou seja; ocorrerá um enviesamento para o algoritmo SVR. Ainda assim, para efeito de facilidade na comparação final, os perfis mais representativos para o algoritmo SVR foram também utilizados nos algoritmos RF e RLM.

3.5 Hiperparâmetros de algoritmos de aprendizagem

Um mesmo modelo de AM pode exigir diferentes restrições, vetores de pesos ou taxas de aprendizado, para generalizar padrões de dados diferentes. As restrições representadas por

essas medidas são chamadas de hiperparâmetros e precisam ser ajustadas para que o modelo possa resolver da melhor maneira possível o problema em questão. Assim, hiperparâmetros controlam a complexidade (ou o equilíbrio entre viés e variância) do modelo ajustado. Em outras palavras, hiperparâmetros de um algoritmo de AM são configurações externas ao modelo, cujos valores não podem ser estimados a partir dos dados, mas sim, são definidos manualmente (a priori) pelo usuário. No processo de otimização, uma tupla de hiperparâmetros produz um modelo ideal que minimiza uma função de perda predefinida em dados independentes fornecidos (Claesen e De Moor, 2015). É usual fazer uso de validação cruzada para estimar a performance e poder de generalização do modelo (Bergstra e Bengio, 2012).

Para o algoritmo SVR, os hiperparâmetros são os valores C e γ (quando se utiliza o *kernel* RBF), enquanto que os vetores suporte são os parâmetros (obtidos após treinamento do algoritmo). As notações n, m, l, s e h (definidas em 3.3), presentes no algoritmo de RF, são seus hiperparâmetros, ao passo que as variáveis e os limites (*thresholds*) usados para dividir cada nó durante o treinamento, são seus parâmetros. O MMQ utilizado para realizar a regressão linear produz uma solução de forma fechada, isto é, retorna de maneira direta um único resultado. Sendo assim, a regressão linear por MMQ não possui hiperparâmetros, possuindo apenas os parâmetros do modelo, que são os coeficientes da regressão (eq. 3.24).

Como os hiperparâmetros apresentam relação com a complexidade (flexibilidade) de um modelo preditivo, escolhas inadequadas para o seu valor podem resultar, por exemplo, em sobreajuste *(overfitting)* ou ainda subajuste *(underfitting)*, o que leva à performances ruins do modelo e diminui sua capacidade de generalização em dados não vistos.

Logo abaixo, é exibido o pseudo-código utilizado na busca de hiperparâmetros ótimos do algoritmo SVR, num processo chamado de validação cruzada (*cross-validation*). O mesmo pseudo-código foi utilizado para achar os hiperparâmetros ótimos do algoritmo RF.

- $Kfold$ é um objeto da biblioteca *sklearn*, utilizado para realizar validação cruzada, dividindo os dados de treinamento em subconjuntos (*folds*) de treino e teste, de modo que em cada iteração, um subconjunto será utilizado para validação e o restante para treinamento, até que todos os *folds* tenham sido percorridos,
- cs e gammas são listas que armazenam os valores dos hiperparâmetros C e γ, respectivamente,
- g e f são listas com valores fornecidos a priori para C e γ, respectivamente,
- var_1 é uma lista onde serão armazenados os valores do erro quadrático médio em cada iteração,

Algorithm 1 SVR

kf = KFold($n\_folds = 4$)
$\overline{MSE}$ = [ ]
gamas = [ ]
cs = [ ]
g = [ ]
f = [ ]
for γ in g:
    for c in f:
    var_1 = [ ]
        for $train_i$, $test_i$ in kf.split(x_norm, y):
        svr = SVR(kernel='rbf', C=c, gamma=γ)
        svr.fit($x\_norm[train_i]$, $y[train_i]$)
        $\hat{y}$ = svr.predict($x\_norm[test_i]$)
        $mse = \frac{1}{n}\sum_{j=test_i}^{n}(y_j - \hat{y}_j)^2$
        var_1.append(mse)
    $\overline{MSE}$.append($\overline{var_1}$)
    gamas.append(γ)
    cs.append(c)

- $train_i$ e $test_i$ são índices que referenciam os valores de cada entrada e saída, nos subconjuntos de treino e teste, respectivamente,
- x_norm são os dados de entrada já normalizados e y são os dados de saída,
- mse é o erro quadrático médio, e
- $\overline{MSE}$ é uma lista que recebe os valores médios dos erros quadráticos médios retornados pela validação cruzada nos subconjuntos.

3.6 Equilíbrio entre viés e variância

Determinados valores do hiperparâmetro irão implicar modelos mais simples, ao passo que outros irão resultar em modelos mais complexos. De modo geral, modelos simples apresentam variância baixa e viés (*bias*) alto; ou seja; sua função estimada não irá apresentar modificações substanciais entre diferentes conjuntos de treinamento (apresenta poucos parâmetros estimados - baixa variância), porém pode não ser uma boa aproximação para o padrão presente nesses dados (viés alto). Por outro lado, modelos mais complexos apresentam variância alta e viés baixo: sua função estimada pode ser composta por muitos parâmetros, de modo que pequenas modificações no conjunto de treinamento podem implicar funções estimadas muito diferentes (variância alta). No entanto, para um conjunto de treinamento em

particular, tal função se aproxima muito bem do padrão presente nos dados e, dessa forma, apresenta viés reduzido. Tais características estão relacionadas a modelos que se ajustam bem aos dados de treinamento, mas que apresentam desempenho ruim quando aplicados a novas observações, ou seja, apresentam baixa capacidade de generalização (*overfitting*)

3.7 Métricas para avaliação de performance

A avaliação da performance de um algoritmo de AM em um determinado conjunto de dados é realizada por meio da mensuração da qualidade dos ajustes em relação à resposta de interesse conhecida. Problemas de classificação, regressão ou misto exigem métricas diferentes. Assim, existe uma diversidade de técnicas de avaliação de performance que variam de acordo com o tipo de problema envolvido. Em problemas de regressão, é comum utilizar as técnicas empregadas à seguir.

3.7.1 Raiz do Erro Quadrático Médio

No contexto de regressão, o Erro Quadrático Médio (MSE - *mean squared error*) é a medida utilizada com mais frequência, sendo definido matematicamente por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{3.28}$$

onde y_i representa o valor do COT observado (medido), $\hat{y}_i$ é o valor do COT ajustado (calculado) e n é o número de observações.

O MSE será pequeno se as respostas preditas pelo modelo forem muito próximas das observadas e será grande se, para algumas observações, a resposta predita e a observada diferirem substancialmente. Na etapa de treinamento dos algoritmos, o MSE pode ser utilizado para comparar modelos com diferentes preditores, hiperparâmetros ou modelos decorrentes de algoritmos distintos. Adicionalmente, como o objetivo final da modelagem preditiva é obter predições acuradas em dados que não foram utilizados para o ajuste do modelo (dados novos), a performance preditiva do modelo selecionado deve ser avaliada a partir da mensuração de seu MSE em dados de teste (James et al., 2013).

A raiz dessa quantidade resulta em um valor na mesma unidade que os dados originais. Sua interpretação refere-se à distância média entre os valores observados e os preditos pelo modelo. Justamente por ser dado na mesma dimensão dos dados originais, será a métrica aqui utilizada na validação do algoritmo no poço teste. Tem-se então a expressão para o

RMSE como segue:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.29}$$

3.7.2 Erro Absoluto Médio

A outra métrica utilizada neste trabalho é o Erro Absoluto Médio (MAE - *Mean Absolute Erro*). A escolha deste método se deve ao fato de ser uma excelente métrica para comparar modelos com diferentes preditores. O MAE representa o desvio padrão do ajuste em relação à média, nas mesmas unidades dos dados. Por exemplo, se estamos ajustando uma série temporal de visitas durante o tempo e encontramos um MAE de 72, quer dizer que o nosso ajuste possui um desvio padrão da média de 72 dias.

Para encontrarmos os MAE, realizamos o seguinte cálculo:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)| \tag{3.30}$$

em que, novamente, y_i representa o valor do COT observado (medido), $\hat{y}_i$ é o valor do COT ajustado (calculado) e n é o número de observações. Note que se $y_i = \hat{y}_i$ obtem-se $MAE = 0$; ou seja; quanto menor o MAE melhor é o nosso ajuste.

Uma vantagem desta métrica é a de ser menos sensível à presença de pontos discrepantes do que o MSE, por exemplo, uma diferença de predição de 10 unidades retornará um MAE duas vezes maior que uma diferença de 5 unidades, ao passo que para a métrica do MSE a diferença na predição de 10 unidades é 4 vezes pior do que 5 unidades. Assim, como os dados de COT apresentam alguns valores discrepantes (mesmo após pré-processamento), essa métrica será de grande valia.

3.7.3 Coeficiente de Correlação de Pearson

O Coeficiente de Correlação de Pearson (R) ou Coeficiente de Correlação Produto - Momento, mede o grau da correlação linear entre duas variáveis quantitativas. É um índice adimensional com valores situados ente -1 e 1, que reflete a intensidade de uma relação linear entre dois conjuntos de dados. Assim, quando $R = 1$ significa que há uma correlação positiva perfeita entre as duas variáveis; caso $R = -1$ ocorre correlação negativa perfeita; e quando $R = 0$ significa que as duas variáveis não dependem linearmente uma da outra, podendo existir

entretanto, uma outra dependência que não seja linear. Assim, o resultado $R = 0$ deve ser investigado por outros meios. O valor de R pode ser calculado como segue:

$$R = \frac{\sum_{i=1}^{n}(y_i - \overline{y}_i)(\hat{y}_i - \overline{\hat{y}}_i)}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}}_i)^2}}, \tag{3.31}$$

onde y_i representa o valor do COT observado (medido), $\hat{y}_i$ é o valor do COT ajustado (calculado), $\overline{y}_i$ é a média dos valores observados de COT, $\overline{\hat{y}}_i$ é a média dos valores ajustados de COT e n é o número de observações.

4

Efeitos Causados Pela Matéria Orgânica nas Diferentes Ferramentas da Perfilagem

A importância de quantificar a matéria orgânica das rochas geradoras, impulsionou uma nova linha de estudos na perfilagem no final do século passado: o cálculo do COT através dos perfis geofísicos de poços. Em diversos artigos publicados a partir de trabalhos feitos, utilizando-se de dados geológicos de diversas regiões ao redor do globo, verificou-se como a presença de matéria orgânica influencia nos perfis geofísicos de poços convencionais. Algumas dessas propriedades incluem elevados valores de GR associados às maiores concentrações de urânio. As menores densidades da matéria orgânica implicam em leituras mais baixas no perfil de densidade (ρ_b) e maiores na curva do tempo de trânsito compressional (Δt). A contagem de hidrogênio também sofre um aumento em zonas geradoras, que influência a leitura do perfil neutrônico (ϕ_N).

4.1 Perfil de Raios Gama (GR)

O perfil de Raios Gama, ou GR (*Gamma Ray*) é um método nuclear baseado na medição da radiação gama natural emitida pelas rochas, que é inofensiva para os seres humanos mas que pode ser detectada por sensores sensíveis. Tal radiação é oriunda principalmente de rochas que contém em sua composição isótopos radioativos do potássio ^{40}K, ^{238}U e ^{232}Th. Entretanto, segundo Kaplan (1964) apud Ribeiro et al. (2014), isótopos ^{238}U e ^{232}Th não emitem radiação gama, porém seus produtos decorrentes do decaimento radioativo (^{214}Bi e

^{208}Tl, respectivamente) emitem raios gama com energias centradas em 1,76 e 2,61 MeV. Ideal para o cálculo do volume de folhelho ou argilosidade[1] e volume de rocha do reservatório que podem conter argila.

Com relação à radioatividade das rochas sedimentares, cristais de quartzo têm alto grau de organização estrutural, o que impede a presença de elementos radiativos na sua estrutura. Em vista disso, arenitos "limpos"[2] e a maioria dos carbonatos apresentam baixos níveis de radiação, enquanto que argilas e folhelhos exibem os mais altos valores de radiação gama, depois dos evaporito potássicos. Por esse motivo, a curva de raios gama diferencia rochas reservatório em potencial (calcários, dolomitos e arenitos) dos folhelhos (Asquith e Gibson, 1982). De maneira geral, altas concentrações de K podem ser causadas pela presença de feldspatos potássicos ou micas. Altos valores de Th podem estar associados com a presença de minerais pesados, particularmente em depósitos de canais ou à presença de argilas terrígenas. Altos valores de U estão relacionados com a presença de matéria orgânica.

Com o surgimento das ferramentas de raios gama espectral, as quais também medem separadamente os canais de K, Th e U, cresceu o número de pesquisas que usam estas ferramentas para quantificar a matéria orgânica em rochas geradoras. Por conta da relação empírica entre a matéria orgânica e o urânio (Swanson, 1960), esperava-se que a utilização do GR espectral facilitaria a identificação de rochas fonte, e que as estimativas de COT fossem diretas e simples. Todavia, ainda não existe uma relação universal entre a contagem específica de urânio e a quantidade de matéria orgânica numa dada rocha fonte. Ainda assim, quando disponível, esta ferramenta possui um valor significativo na interpretação.

4.2 Perfil de Indução (ILD)

Por conta dos complexos e nem sempre compreendidos processos físicos que ocorrem nos folhelhos, estimar o teor de matéria orgânica através das respostas dos perfis resistivos foi a menos aplicada das técnicas.

Nixon (1973), Meissner (1978) e Schmoker e Hester (1989), observaram a ocorrência de um aumento significativo nas leituras dos perfis de indução relacionado com a geração de hidrocarbonetos (não condutores) que não sofreram migração primária e permaneceram na rocha fonte. Posteriormente, Passey et al. (1990) corroboram, mostrando que, em rochas fonte cuja matéria orgânica já foi maturada, a resistividade aumenta por conta da presença de hidrocarbonetos. Assim, a resposta da resistividade de rochas fonte será afetada pelo

[1] Os minerais de argila são ricos em ^{40}K, emitindo grande quantidade de raios gama.
[2] O termo "limpo" refere-se à ausência de argilominerais na estrutura do arenito.

conteúdo de COT.

4.3 Perfil Densidade (ρ_b)

O Perfil de Densidade é um perfil nuclear assim como o GR, entretanto, ao invés de captar a radiação gama natural, a ferramenta emite radiação gama nas formações rochosas e capta a radiação emitida de volta pelo Espalhamento Cômpton. O perfil mede a densidade aparente das formações, uma resposta combinada da densidade dos fluidos e da matriz dos minerais constituintes. Em outras palavras, quanto maior a quantidade de fluidos contidos numa formação, mais porosa ela é. Em folhelhos com um grau semelhante de compactação, matriz semelhante e fluidos de mesma densidade, a saturação da água deve ser igual. A matéria orgânica sólida tem uma densidade semelhante à da água ($\approx$ 1,0 g/cm^3) e, portanto, menor que a densidade da matriz da rocha circundante. Se a densidade lida em folhelhos geradores é menor que a densidade lida em folhelhos não-geradores, pode ser um indicativo da presença de matéria orgânica.

4.4 Perfil Neutrônico (ϕ_N)

O Perfil de Porosidade Neutrônica (ϕ_N) é baseado no espalhamento elástico de nêutrons à medida que colidem com núcleos atômicos das formações. Cada nêutron se dispersa de um núcleo com energia cinética menor, de modo que nas colisões elásticas as conservações de energia e momento ditam que a presença de hidrogênio nas formações dominam o processo de desaceleração dos nêutrons. A razão para isso é que a massa do núcleo de Hidrogênio é aproximadamente igual à do nêutron incidente. Consequentemente, formações com altas concentrações de H^+ exibem baixa concentração de nêutrons termais, epitermais e gama de captura. Inversamente, formações com baixa concentração de H^+ exibem altas concentrações de nêutrons termais, epitermais e gama de captura. Na figura 4.1, tem-se uma ilustração deste processo.

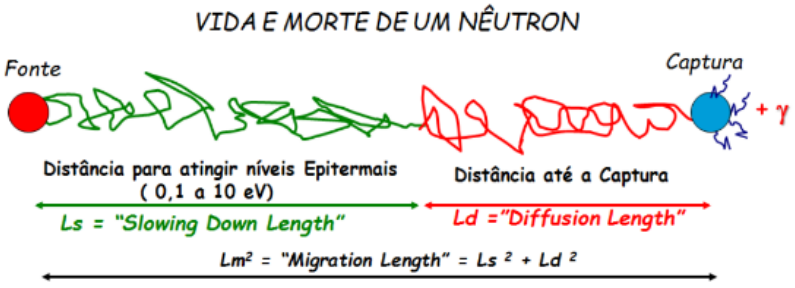


Figura 4.1: Ilustração da trajetória randômica realizada por um nêutron. Fonte: extraída de notas de aula do professor Geraldo Girão Nery, em 2015.1.

A presença de matéria orgânica na formação tem relação direta com o conteúdo de átomos de hidrogênio e porosidade das rochas. Assim, o perfil ϕ_N aumenta nos intervalos ricos em matéria orgânica.

4.5 Perfil Sônico (Δt)

O tempo de trânsito compressional (Δt) é o recíproco da velocidade da onda compressional e é função da litologia e porosidade das formações, além do tipo e modelos de distribuição de fluidos (água, gás, petróleo, querogênio, etc.). Como resultado, com o aumento aparente do valor de Δt, o conteúdo de COT tende a aumentar (Kamali e Mirshady, 2004).

Myers e Jenkyns (1992), observaram que o tempo de trânsito compressional medido em rochas geradoras de hidrocarbonetos são mais longos, quando comparados à folhelhos pobre em componentes orgânicos. Isso se deve à baixa densidade da matéria orgânica. Dessa forma, pode-se identificar possíveis rochas fonte através de uma análise do perfil sônico, pois a diminuição da velocidade de propagação das ondas nestas rochas torna as leituras maiores no perfil. É necessário, no entanto, para minimizar os erros, identificar os saltos de ciclo, que alteram as leituras, aumentando os valores.

5

Resultados do pré-processamento

Os resultados exibidos nesta seção são apresentados em gráficos *boxplots*, mostrando a amplitude de variação do COT nas cinco litologias selecionadas antes e após o pré-processamento. Em seguida, são apresentados gráficos *crossplots* exibindo o valor do Coeficiente de Correlação de Person (R) entre os parâmetros analisados, além de tabelas mostrando os resultados da estatística descritiva básica dos dados antes e após serem processados. Os gráficos das predições são avaliados conforme a tendência da dispersão dos pontos, comparando-se o COT medido com o COT predito no poço teste, além de tabelas mostrando os erros e correlações para cada algoritmo no conjunto de treinamento, conjunto de teste e no poço teste. Posteriormente, os resultados das predições são apresentados em perfis de COT no poço teste.

O gráfico de *boxplot* exibido na Figura 5.1, mostra que os valores do conteúdo de COT medidos nos folhelhos são maiores que nas demais litologias, o que é esperado. Os losangos pretos são considerados *outliers*; ou seja; estão a mais ou menos duas vezes o desvio padrão em relação à média, considerando uma distribuição Normal de probabilidades. Apesar destes valores extremos serem considerados *outliers*, sabe-se que estão condizentes com a realidade da geoquímica de poços do local, visto que o COT está detalhado como tendo "picos de 6 %" na descrição da carta estratigráfica da Bacia de Santos. Apesar disso, optou-se pela remoção de valores de COT acima de 3 %, pois os resultados das predições foram bastante afetado pelos mesmos. Após a remoção, os algoritmos retornaram erros menores na etapa de validação cruzada e no poço teste. Apenas um ponto possuía COT acima de 6 %). Foram removidos também outros pontos discrepantes selecionados visualmente nos vários *crossplots* realizados. Tais pontos podem ter sido oriundos de erros na leitura dos perfis, ou durante o processo de concatenação dos fragmentos de perfis, visto que foram feitos vários códigos em

linguagem de programação *Python* para montagem dos perfis em *track* único e além disso, não tivemos controle sobre a coleta dos dados.

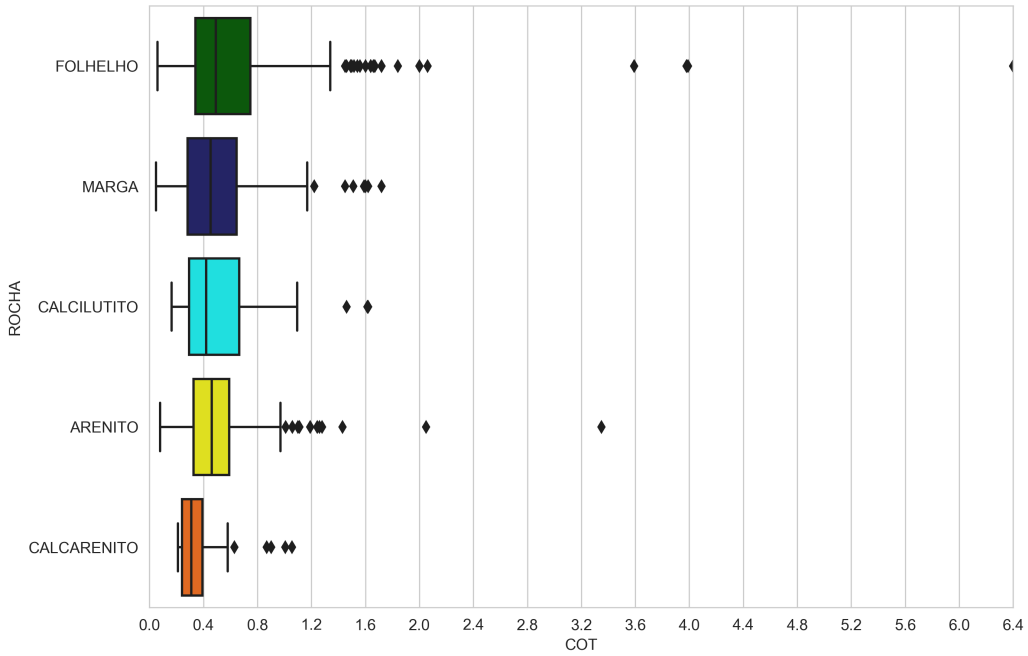


Figura 5.1: *Boxplot* exibindo valores de COT por litologia, antes da remoção de *outliers*, com COT variando de 0.04 a 6.4 %

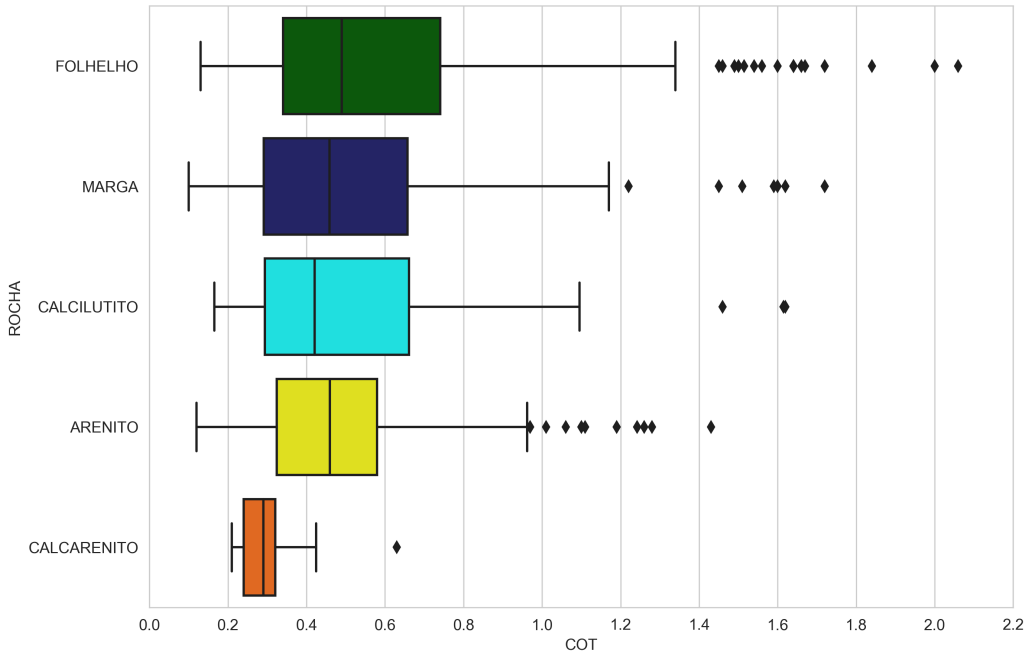


Figura 5.2: *Boxplot* exibindo valores de COT por litologia, após remoção de *outliers* e valores não confiáveis, com COT variando de 0.04 a 2.06 %

Na Figura 5.1 nota-se claramente dois pontos destoantes nos arenitos, ambos marcando acima de 2 %. É possível que descreveram erroneamente como arenito o que deveria ser folhelho ou marga, e assim, optamos pela remoção destes valores também. Já na Figura 5.2 é exibido o *boxplot* após remoção dos valores discrepantes de COT e ainda assim, uma grande quantidade de *outliers* se mantêm. Optou-se por deixar esses valores, já que há uma quantidade razoável representativa dos mesmos, variando de 1 à 2 %, aproximadamente.

Os valores de COT seguem uma tendência esperada, mostrando que as rochas argilosas registram os maiores valores, com os folhelhos registrando valor médio de 0.59 %, seguido das margas (0.53 %) e calcilutito (0.52 %), contrastando com menores valores registrados nas rochas arenosas, com os arenitos registrando média de 0.48 % e os calcarenitos de 0.30 %.

Valores de estatística descritiva básica são exibidos nas tabelas 5.1 e 5.2, antes e após pré-processamento, respectivamente. Como esperado, a média dos valores de COT se aproxima da mediana após remoção dos picos de COT, como mostra a Tabela 5.2, além disso, o desvio padrão também foi diminuído consideravelmente; ou seja; em sendo o desvio padrão uma medida do grau de dispersão dos dados em torno da média, sua diminuição implica uma maior homogeneidade dos dados. Em relação às variáveis de entrada, apenas os perfis GR e ILD sofreram alguma mudança descritiva nos dados, com seus valores máximos razoavelmente diminuídos.

Tabela 5.1: Estatística descritiva dos dados de treinamento antes do pré-processamento, apenas onde há dados de COT.

	COT	$\boldsymbol{\Delta t}$	**GR**	$\boldsymbol{log_{10}}$**(ILD)**	$\boldsymbol{\phi_N}$	$\boldsymbol{\rho_b}$
Média	0.55	84.33	84.0	0.37	23	2.41
Mediana	0.47	81.12	85.8	0.30	23.24	2.43
Desvio padrão	0.42	19.62	24.76	0.67	9.83	0.20
Mínimo	0.04	47.80	14.0	-0.67	0.53	1.90
Máximo	6.4	144.93	171.81	4.73	48.64	2.95

Tabela 5.2: Estatística descritiva dos dados de treinamento, após pré-processamento, apenas onde há dados de COT.

	COT	$\boldsymbol{\Delta t}$	**GR**	$\boldsymbol{log_{10}}$**(ILD)**	$\boldsymbol{\phi_N}$	$\boldsymbol{\rho_b}$
Média	0.52	84.5	83.97	0.36	23	2.41
Mediana	0.465	81.40	85.83	0.29	23.30	2.43
Desvio padrão	0.30	19.50	24.27	0.64	9.78	0.20
Mínimo	0.04	49.11	20.81	-0.67	0.53	1.90
Máximo	2.06	141.30	148.0	3.30	48.64	2.90

Fazendo uma análise mais aprofundada dos valores médios de COT por litologia, por poço, chegamos à conclusão de que os arenitos em alguns poços podem assumir valores maiores que as margas, ou ainda maiores que os folhelhos do mesmo poço, como observado na Figura 5.3. Além disso, deve-se notar na mesma imagem que valores de COT medidos em calcarenitos só foram registrados em apenas três poços.

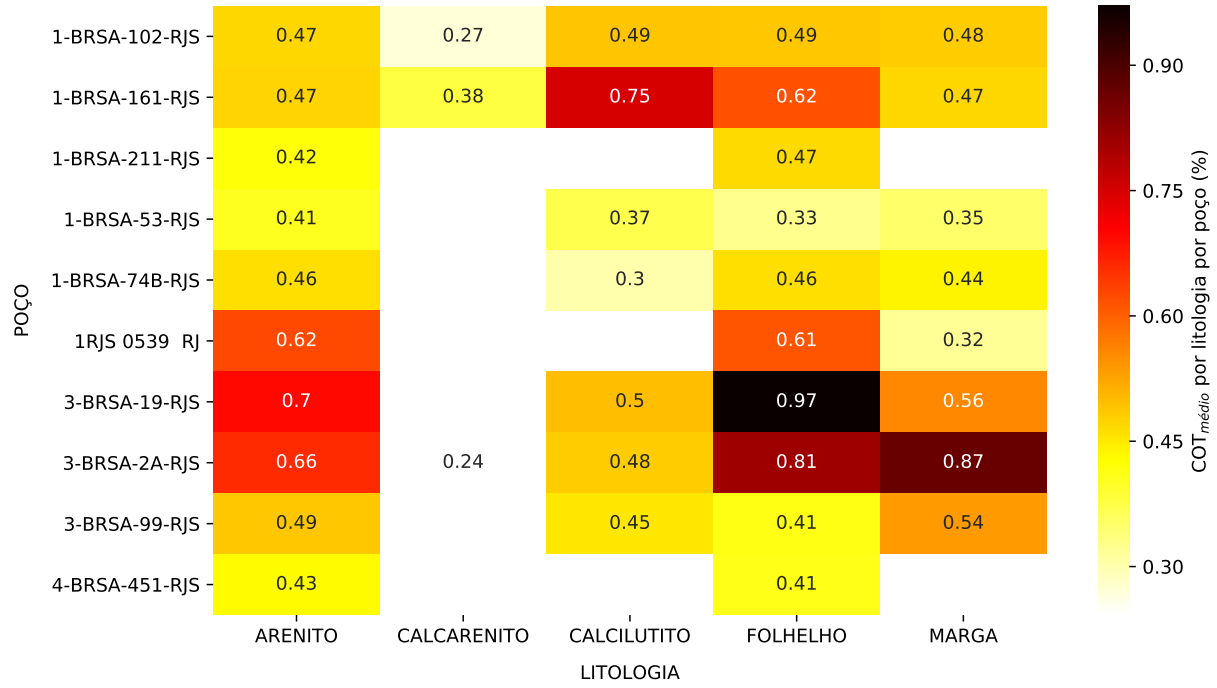


Figura 5.3: Valores médios de COT por litologia, por poço.

Ao fazer a leitura da Figura 5.3 é imprescindível levar em conta que, do conjunto de dados utilizados após pré-processamento, 315 medidas foram feitas em arenitos, 285 medida em folhelhos, 124 medidas em margas, 80 medidas em calcilutitos e apenas 21 medidas em calcarenitos. Assim, o banco de dados utilizado está desbalanceado em relação à dados de COT por litologia.

Com relação aos valores médios de COT para cada Formação, por poço, observa-se que a Formação Guarujá assume os menores valores (Figura 5.4), e são justamente os carbonatos (calcarenitos) da Formação Guarujá que representam os reservatórios mais importantes da seção pós-sal, como já descrito na seção 1.3. Enquanto isso, as Formações Itajaí-Açú e Marambaia assumem os maiores valores médios de COT, corroborando com o fato já descrito na seção 1.2, onde descreve-se que a Formação Itajaí-Açú possui intervalos de rochas geradoras caracterizados por folhelhos e argilitos cinza-escuros e a que a Formação Marambaia é composta basicamente de siltitos e folhelhos. Portanto, faz sentido esses altos valores de COT, ao passo que a Formação Marambaia é repleta de folhelhos e margas, que resultam

também nesses mais altos valores de COT em relação às demais Formações. Os dados de COT medidos nas Formações, após pré-processamento, estão divididos na seguinte ordem: 325 valores medidos na Formação Itajaí-Açú, 303 medidos na Formação Itanhaem, 89 medidos na Formação Marambaia, 80 medidas na Formação Guarujá e 28 medidas feitas na Formação Juréia.

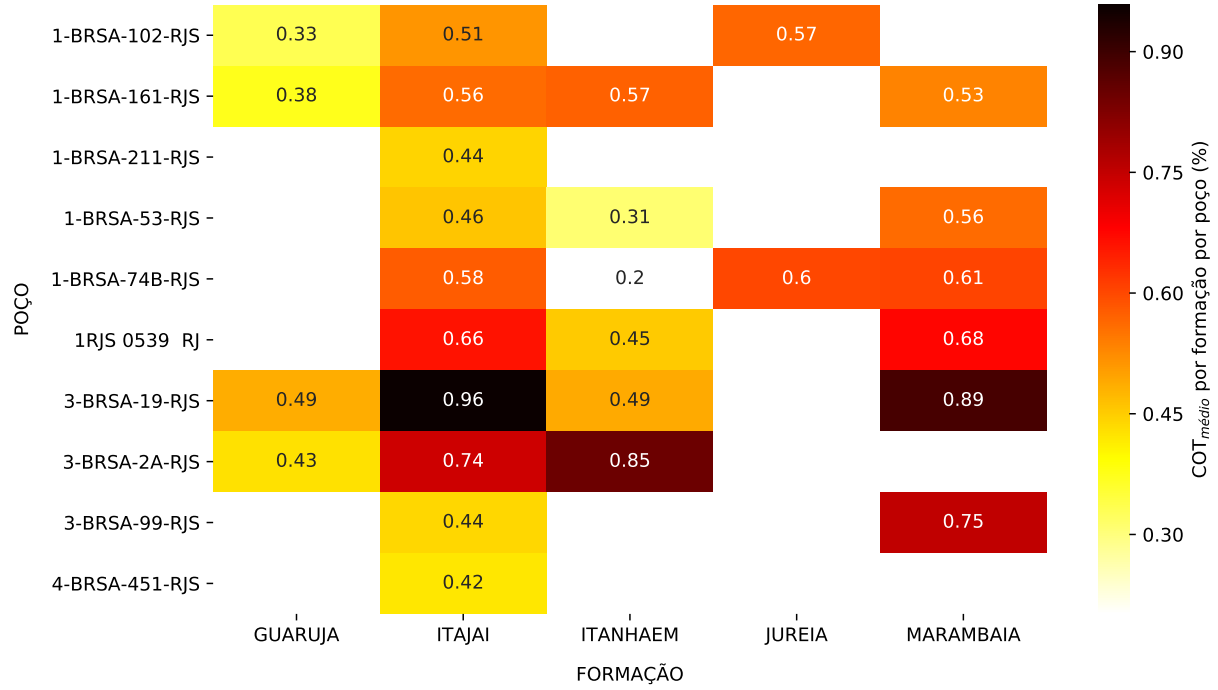


Figura 5.4: Valores médios de COT por Formação, por poço.

Os dados foram separados segundo critério de argilosidade, onde o grupo de rochas argilosas é representado por folhelhos, margas e calcilutitos e o grupo de rochas arenosas, por arenitos e calcarenitos, como já explicado em 1.4.3. Com base nos dados utilizados, foram analisadas as correlações entre as resposta dos perfis (eixo x) e o conteúdo de COT medido (eixo y), juntamente com as respostas petrofísicas envolvidas. Tais relacionamentos podem ser vistos na Figura 5.5 para o conjunto de dados de treinamento com os respectivos coeficientes de correlação obtidos para cada grupo, antes do pré-processamento, enquanto a Figura 5.6 mostra os mesmos gráficos após pré-processamento. Como a curva ILD e os dados de COT possuem ambos uma distribuição de probabilidades com assimetria positiva, utilizou-se o logaritmo na base 10 para que ficassem mais normalmente distribuídos, observando assim um aumento nos valores de R entre o COT e as respostas dos perfis. Além disso, notou-se uma melhora nos resultados das predições para o poço teste em todos os três algoritmos. No entanto, para uma melhor compreensão visual, o conteúdo de COT medido foi apresentado em escala linear, o que significa que as curvas de regressão tiveram que ser exibidas na forma

de função potência dos perfis, da seguinte forma:

$$COT_i = 10^b.10^{a.X_i} \tag{5.1}$$

onde b e a são respectivamente os Coeficientes Linear e Angular da regressão linear quando um perfil geofísico X é usado como variável independente e $\log_{10}(COT)$ é usado como variável dependente. O índice i descreve cada ponto do par ordenado na curva de regressão.

É possível observar que, após a remoção de valores não confiáveis e/ou muito distantes da média, o coeficiente de correlação entre a variável de saída e as variáveis de entrada aumentou consideravelmente para o conjunto de rochas arenosas, mas permaneceu praticamente inalterado para o conjunto de rochas argilosas, chegando à diminuir um pouco para o GR e ϕ_N (Figuras 5.5 e 5.6 (a) e 5.5 e 5.6 (b)). Uma das causas possíveis e plausíveis é que, esses pontos discrepantes nas litologias de arenito e/ou calcarenito eram na verdade folhelho ou marga, erros estes propagados pelos técnicos que fizeram a análise e descrição dos testemunhos. Uma das explicações para a não melhoria das correlações para as rochas argilosas é que os calcilutitos possuem composição muito diferente dos folhelhos e margas. Além disso, os dados estão muito dispersos, pois são folhelhos de diverssas regiões dos poços e em diversos poços agrupados, e os folhelhos possuem características muito distintas uns dos outros, com grandes mudanças mineralógicas e no conteúdo de matéria orgânica em cada tipo. Ademais, o conteúdo de argila altera significativamente as respostas dos perfis geofísicos.

Analisando-se os *cross-plots* da Figura 5.6, é possível descrever detalhadamente as relações entre o conteúdo de COT e as respostas de cada ferramenta de perfilagem. A Figura 5.6 (a) mostra que a curva GR tem uma correlação positiva com o conteúdo de COT para ambos os grupos, mas o grupo de rochas argilosas exibe correlação muito mais fraca que o grupo de rochas arenosas. Isso ocorre porque a matéria orgânica (MO) dos folhelhos marinhos geralmente adsorve uma quantidade substancial de urânio, resultando em MO com alta radioatividade. No entanto, vale ressaltar que a radioatividade do conteúdo de argila também é muito alta; portanto, a correlação entre o GR e o COT é baixa para rochas ricas em argila, neste caso. Também é perceptível que o conteúdo de COT para rochas arenosas está principalmente no intervalo $60 < \text{GR} < 100$, enquanto que para rochas argilosas está espalhado por toda a amplitude do GR. Porém, cabe ressaltar que a rocha argilosa que engloba esses valores de $\text{GR} < 60$ são calcilutitos, enquanto que valores de $\text{GR} > 100$ engloba basicamente folhelhos, como pode ser visto na Figura B.1 do Apêndice B.

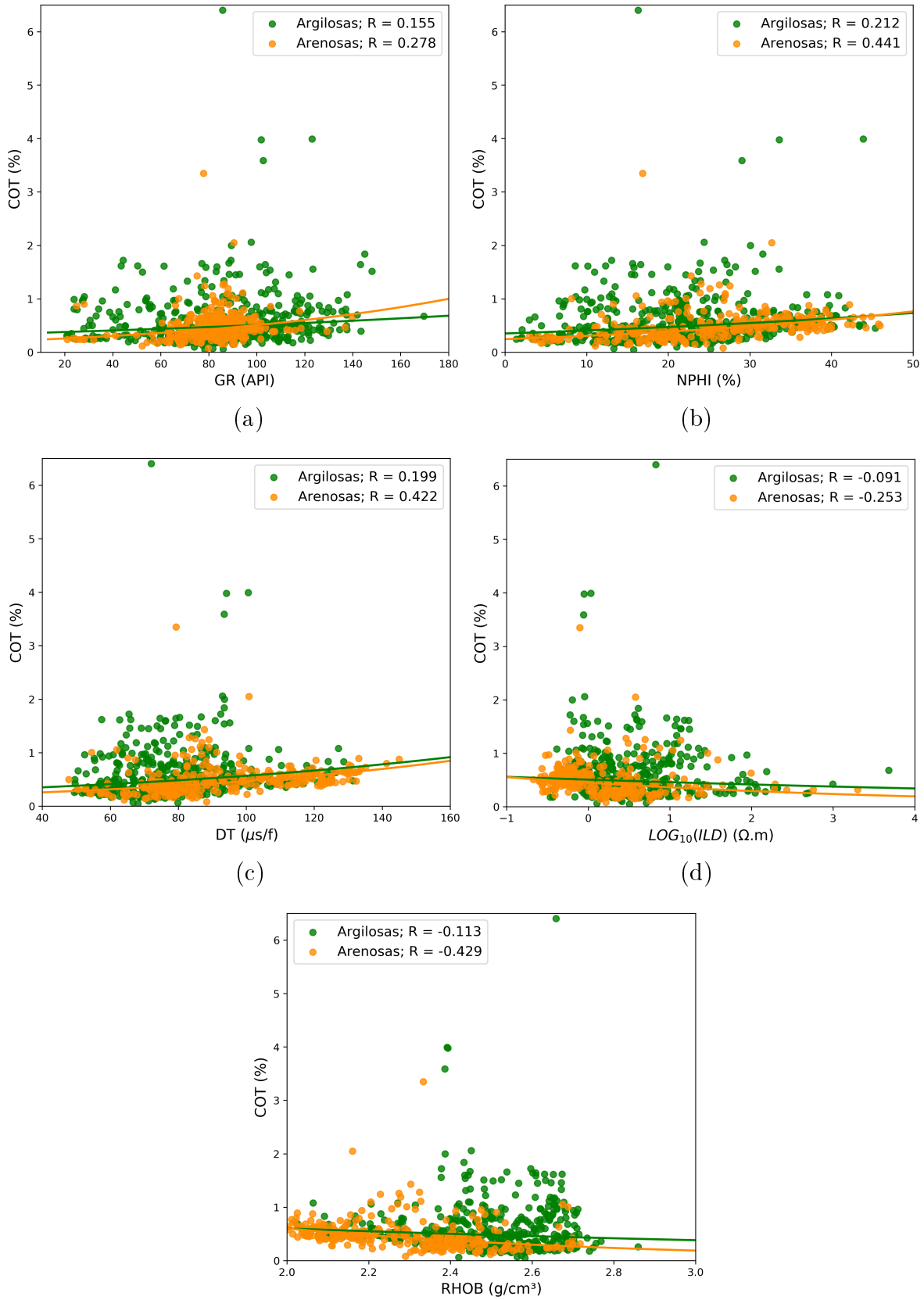


Figura 5.5: *Cross-plots* exibindo ambos os grupos de rochas para o conjunto de dados de treinamento, antes da remoção de pontos discrepantes, destacando o coeficiente de correlação (R) entre o COT medido e os perfis GR (a), ϕ_N (b), Δt (c), ILD (d) e ρ_b (e), para cada grupo.

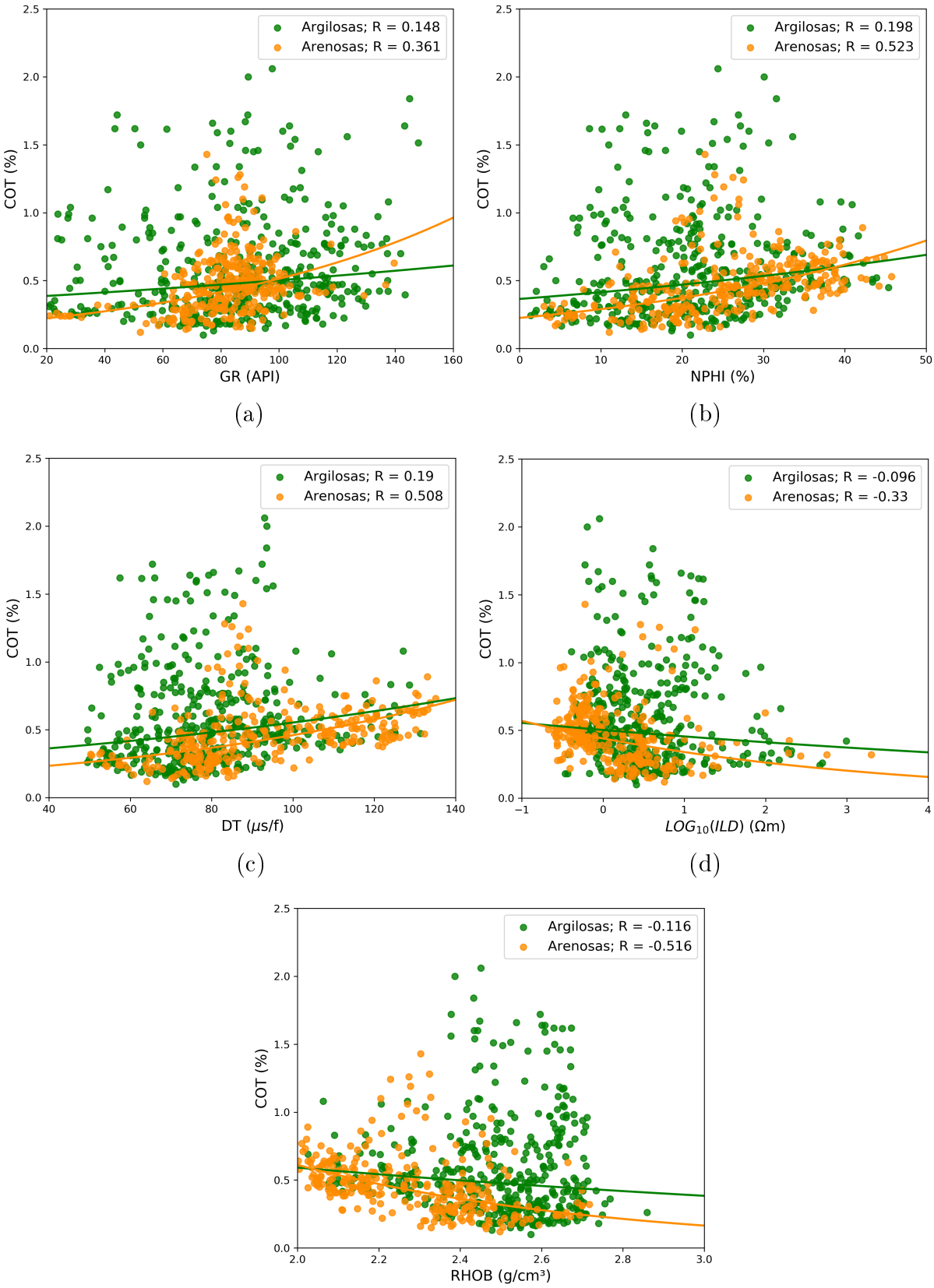


Figura 5.6: *Cross-plots* exibindo ambos os grupos de rochas para o conjunto de dados de treinamento, após a remoção de pontos discrepantes, destacando o coeficiente de correlação (R) entre o COT medido e os perfis GR (a), ϕ_N (b), Δt (c), ILD (d) e ρ_b (e), para cada grupo.

A Figura 5.6 (b) mostra o *cross-plot* ϕ_N×COT, com destaque para as mais altas correlações dentre todos os perfis, tanto para as argilosas quanto para as arenosas, indicando que o perfil neutrônico, neste caso, é o que melhor pode explicar a variabilidade do COT dentro do conjunto de dados utilizados. Uma possível razão para isso é que a MO é muito mais facilmente encontrada em rochas argilosas e sabe-se que a porosidade aparente registrada pela ferramenta ϕ_N aumenta com o aumento do conteúdo de argila, de modo que esse raciocínio pode ser válido tanto para as rochas argilosas quanto para as rochas arenosas (com alguma argilosidade). O *cross-plot* COT×ϕ_N na Figura B.1 do Apendice B corrobora com essa afirmação, mostrando que os maiores valores de COT estão situados nos folhelhos, justamente onde a resposta da ferramenta ϕ_N se amplifica (valores acima de 30 %), ao passo que os menores valores de COT situam-se nos calcilutitos, onde estão também os menores valores de ϕ_N (abaixo de 17 %). Adiante na Figura 5.6 (c), observa-se a baixa correlação entre Δt e COT para rochas argilosas, atribuível às pequenas diferenças nos tempos de trânsito da onda acústica que se propaga através de rochas sedimentares ricas em MO e outros minerais com densidades próximas à da MO. Por outro lado, as rochas arenosas mostram alto coeficiente de correlação, confirmando a ideia de que os minerais de argila são um dos fatores da extrema dispersão exibida no *cross-plot* Δt×COT para rochas argilosas. Para o *cross-plot* ILD×COT, o valor de R é até agora o mais baixo para ambos os grupos de rochas (Figura 5.6 (d)); portanto, calcular o conteúdo de COT usando apenas dados de resistividade não seria confiável, neste caso específico. Sabe-se que o valor da densidade exibida no perfil de densidade diminui com o aumento de hidrocarboneto ou matéria orgânica nas rochas, uma vez que esses compostos possuem densidade entre (1,1-1,4 g/cm^3), sendo menores que a do quartzo ($\approx$ 2,65 g/cm^3) e da argila ($\approx$ 2,77 g/cm^3). A Figura 5.6 (e) corrobora com este fato, mostrando que a correlação entre a curva de densidade (ρ_b) e o conteúdo de COT é negativa para ambos os grupos de rochas. No entanto, novamente a correlação para o grupo de rochas arenosas apresenta-se muito mair que para o grupo de rochas argilosas. É possível que isso ocorra porque, apesar da densidade da MO ser baixa em relação à de outros minerais, vai mudando com seu grau de maturidade, levando à uma maior dispersão em rochas ricas em MO do que em rochas arenosas, já que aquelas, por terem um maior conteúdo de matéria orgânica disperso, também exibirão maior variabilidade para os diferentes estágios de maturação nas diferentes Formações. Outra resposta notável é que, para $\rho_b > 2{,}5$ g/cm^3, encontra-se principalmente litologias argilosas e os mais altos valores de COT; ou seja; retratando o fato de que os minerais de argila (presentes nos folhelhos, margas e calcilutitos) são mais densos que arenitos e calcarenitos, e são nessas rochas ricas em argila que se situa a maior quantidade de MO.

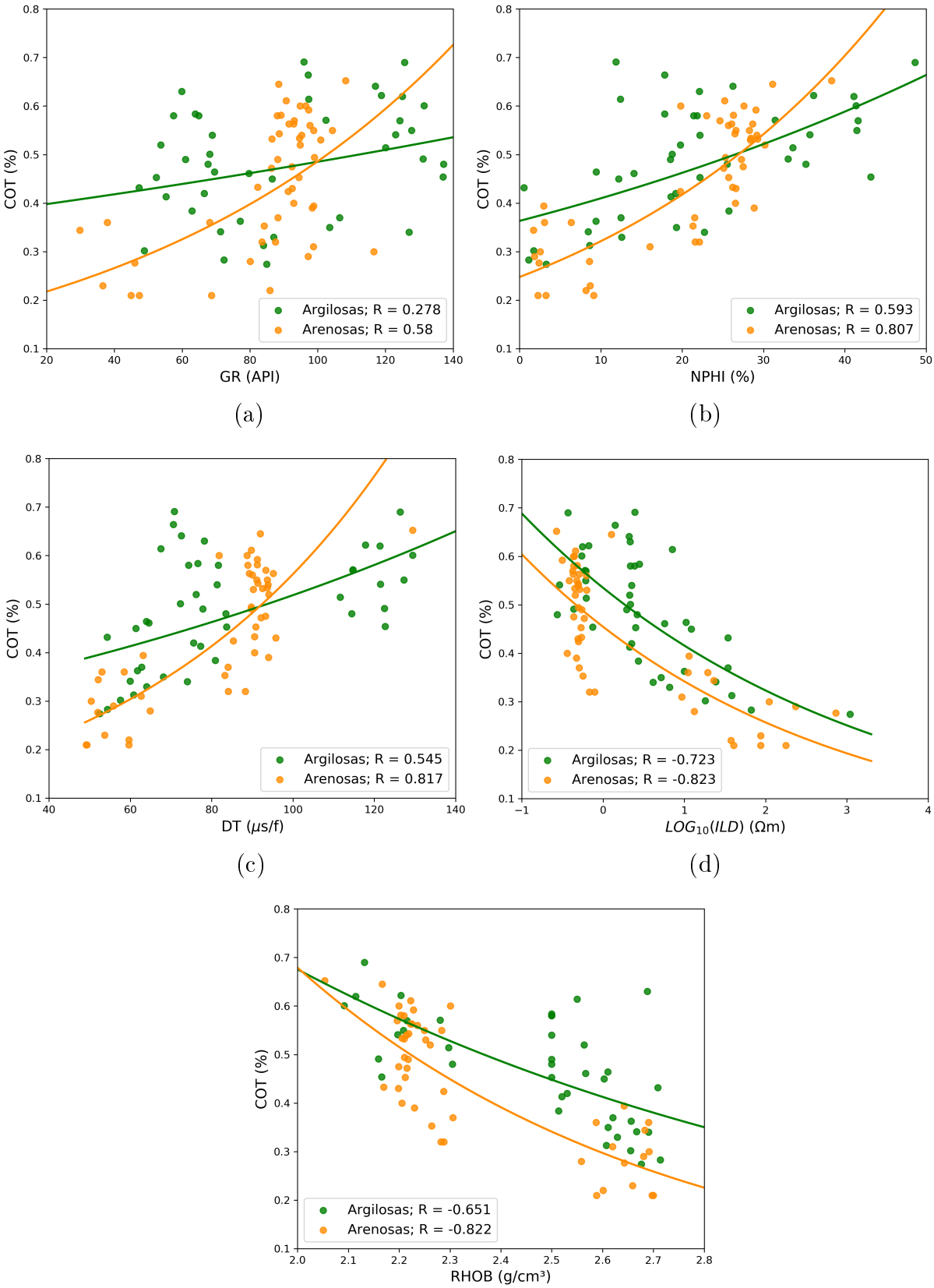


Figura 5.7: *Cross-plots* exibindo ambos os grupos de rochas para o conjunto de dados do poço teste, destacando o coeficiente de correlação (R) entre o COT medido e os perfis GR (a), ϕ_N (b), Δt (c), ILD (d) e ρ_b (e), para cada grupo.

Seguindo o mesmo princípio da Figura 5.6, é possível observar nos *cross-plots* da Figura 5.7 que os valores de correlação para os dados do poço teste são muito maiores que nos dados de treinamento, chegando inclusive a apresentar a mais alta correlação entre a curva ILD e o COT para rochas argilosas, em contraste com os dados de treinamento, onde essa curva apresentou a menor correlação para essas rochas. Assim, se tivéssemos de obter o COT através de apenas um perfil, este perfil certamente não seria o ILD.

6

Resultados dos ajustes

Analisadas as relações entre o COT e os perfis nos dados de treinamento e no poço teste, será introduzido agora os resultados dos ajustes das curvas apresentados pelos algoritmos, nos dados do poço teste, bem como a comparação entre eles. Iniciaremos mostrando os resultados da melhor combinação de perfis; ou seja; aquela que retornou os menores erros de predição nos dados de validação, que neste caso foi utilizando-se o par GR-ϕ_N para o grupo de rochas argilosas e Δt-ϕ_N para o grupo de rochas arenosas (Figura 6.1).

No eixo x estão os valores do COT medidos em laboratório e no eixo y estão os valores de COT obtidos pelos algoritmos, onde COT_predito_SVR é o resultado obtido utilizando-se o algoritmo Máquina de Vetores de Suporte, COT_predito_RF é o resultado obtido utilizando-se o algoritmo Floresta Aleatória e COT_predito_LINEAR é o resultado do algoritmo obtido utilizando-se Regressão Linear Múltipla Múltipla. Em (a), (b) e (c) exibe-se o valor global de R para todo o conjunto de dados do poço teste, sem discriminar entre argilosas e arenosas; em (d), (e) e (f) os valores de R estão discriminados entre argilosas e arenosas, para os respectivos algoritmos; ou seja; (a) está para (d), assim como (b) está para (e) e (c) está para (f).

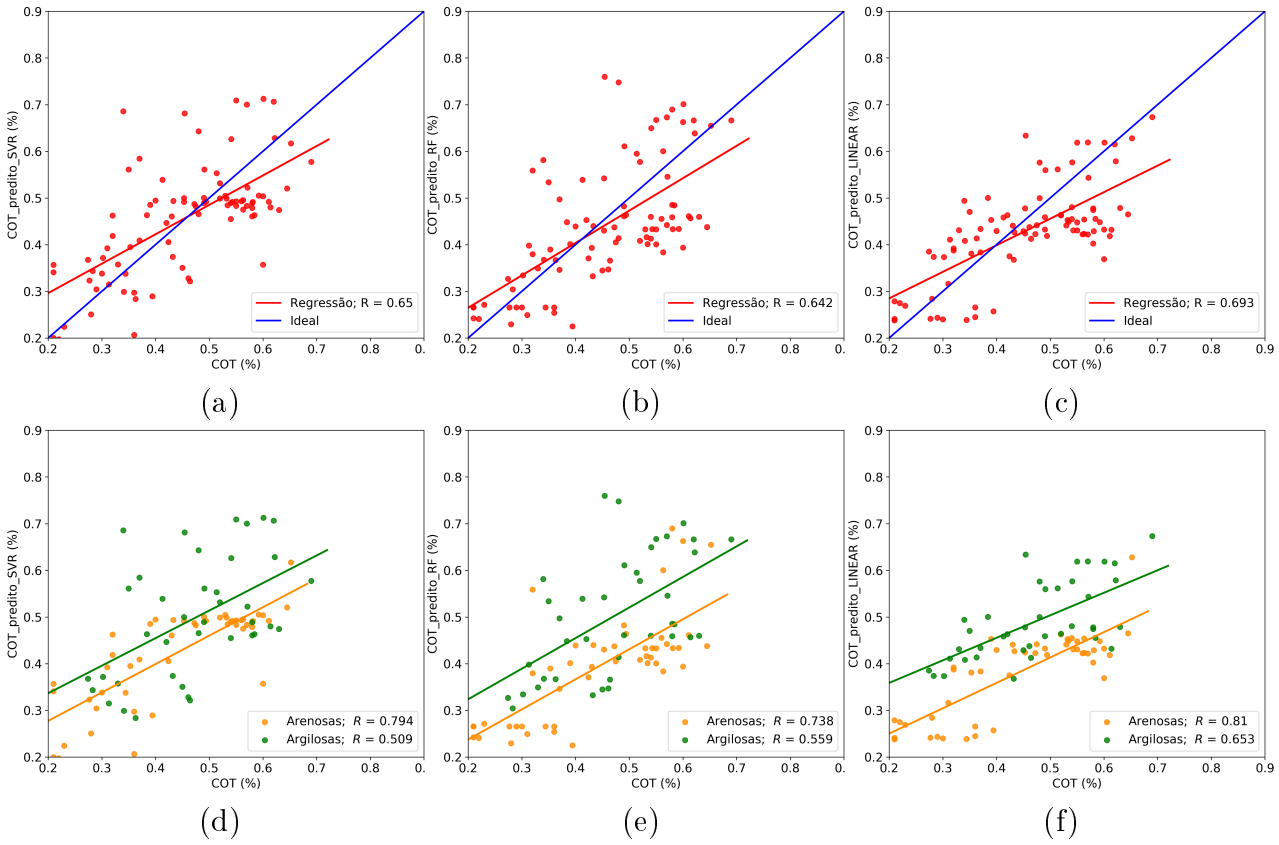


Figura 6.1: Comparação entre os resultados do COT predito e medido em laboratório, para o poço teste, utilizando os perfis GR e ϕ_N para rochas argilosas e Δt e ϕ_N para as arenosas: (a) comparação entre o COT predito por SVR e medido em laboratório; (b) comparação entre o COT predito por RF e medido em laboratório; (c) comparação entre o COT predito por RLM e medido em laboratório; em (d), (e) e (f) são exibidos os mesmos gráficos respectivos de cima, separados em grupos de rochas argilosas e arenosas.

Para ambos os grupos de rochas (argilosas e arenosas), os hiperparâmetros que retornaram o menor erro na validação cruzada para o algoritmo SVR e RF estão representados respectivamente nas tabelas 6.1 e 6.2. A notação utilizada para os hiperparâmetros do RF já foi descrita na seção 3.3.

Tabela 6.1: Hiperparâmetros ótimos para o algoritmo SVR, usando-se os perfis GR e ϕ_N como dados de entrada para rochas argilosas e Δt e ϕ_N para arenosas.

Hiperparâmetros	**Rochas argilosas**	**Rochas arenosas**
C	0.1	1
γ	1	1

Tabela 6.2: Hiperparâmetros ótimos para o algoritmo RF, usando-se os perfis GR e ϕ_N como dados de entrada para rochas argilosas e Δt e ϕ_N para arenosas.

Hiperparâmetros	**Rochas argilosas**	**Rochas arenosas**
n	153	144
m	26	25
l	7	7
s	7	5
h	2	2

As tabelas 6.3 e 6.4 exibem de modo mais detalhado os resultados já apresentados na Figura 6.1, para rochas argilosas e arenosas, respectivamente. São apresentados os resultados para os conjuntos de treinamento e teste (durante a validação cruzada) e no poço teste. É natural que os resultados no conjunto de treinamento sejam melhores que no conjunto de teste, pois os algoritmos são treinados neste dados e tendem à captar suas características, adaptando-se até mesmo à ruídos (valores anômalos) específicos do conjunto de treinamento. No entanto, foram inesperados os excelentes resultados no poço teste, visto que os dados deste poço não passaram pelo processo de validação cruzada e foram tratados como dados totalmente externos e alheios ao conjunto de treinamento inicial.

Tabela 6.3: resultado da predição dos algoritmos para as rochas argilosas, utilizando os perfis GR e ϕ_N.

	Dado de treinamento			**Dado de teste**			**poço teste**		
Algoritmo de regressão	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**
SVR	0.317	0.237	0.36	0.225	0.247	0.37	0.51	0.098	0.12
RF	0.58	0.217	0.33	0.12	0.25	0.346	0.56	0.0967	0.12
LINEAR	0.131	0.282	0.373	0.0376	0.257	0.321	0.653	0.072	0.085

Tabela 6.4: resultado da predição dos algoritmos para as rochas arenosas, utilizando os perfis Δt e ϕ_N.

	Dado de treinamento			**Dado de teste**			**poço teste**		
Algoritmo de regressão	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**
SVR	0.54	0.12	0.185	0.50	0.13	0.19	0.794	0.065	0.081
RF	0.63	0.116	0.182	0.53	0.11	0.163	0.738	0.083	0.102
LINEAR	0.394	0.15	0.214	0.54	0.11	0.153	0.81	0.082	0.10

Para rochas argilosas (Tabela 6.3), a RLM foi o algoritmo que retornou os melhores resultados no poço teste, com o maior valor de R; ou seja; ajustes menos dispersos. No entanto, no dado de teste, exibiu o maior valor de MAE (indicando menos robustez que os outros dois), mas retornou o menor RMSE; ou seja; consegue responder melhor à valores extremos que o SVR e RF.

Para as rochas arenosas (Tabela 6.4), o algoritmo SVR é o que retorna os menores erros no poço teste, mas novamente a RLM retornou um valor de R levemente maior, porém o SVR exibe os menores valores de MAE e RMSE dentre os três algoritmos. É possível que o resultado satisfatório da RLM nos dois grupos litológicos do poço teste se deva ao fato de que os dados de COT possuem altas correlações com os respectivos pares de perfis para este poço, como pode ser visto nas Figuras C.1 e C.2, do Apêndice C; ou seja; já tem um viés (*bias*) associado para o favorecimento da RLM frente aos outros dois algoritmos.

Apesar da RLM ter retornado excelentes resultados estatísticos, este método não consegue descrever bem a região de rochas argilosas das formações Itanhaem (calcilutitos e pelitos) e Guarujá (margas e calcilutitos), onde seriam esperados valores mais altos de COT, intercalados com menores valores para os calcarenitos porosos da Formação Guarujá. Em contrapartida, os algoritmos SVR e RF retornam valores mais representativos, como é mostrado na Figura 6.2. Nesta imagem, é exibido o resultado da utilização do par de perfis GR-ϕ_N para as rochas argilosas e o par Δt-ϕ_N para as rochas arenosas em *track* único para cada algoritmo, de modo que o algoritmo treinado no conjunto de dados de rochas argilosas foi, obviamente, aplicado na região de rochas argilosas do poço teste, analogamente para as rochas arenosas.

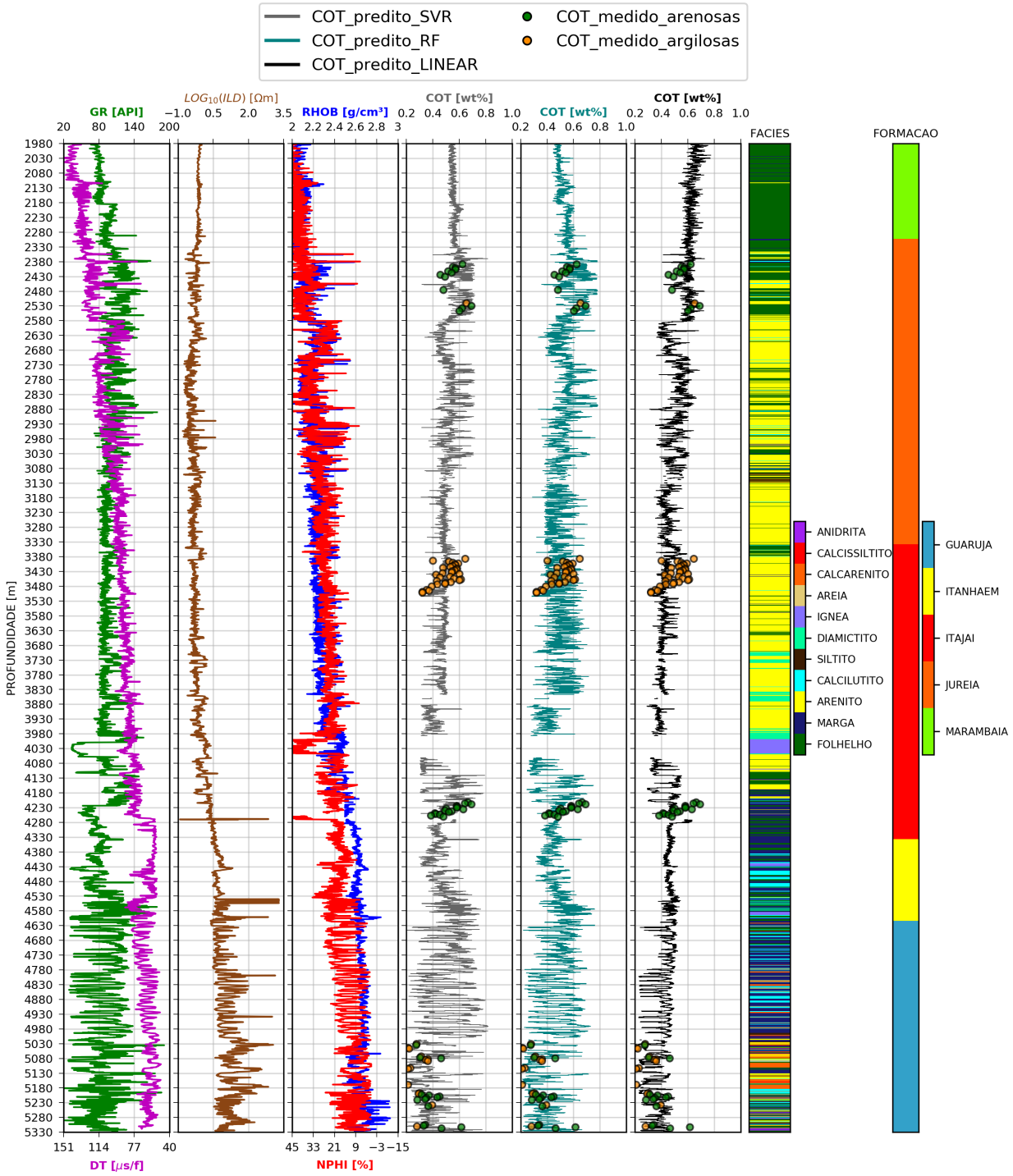


Figura 6.2: Plotagem do COT predito para todo o intervalo perfilado, nas regiões de folhelhos, margas, calcilutitos, arenitos e calcarenitos, utilizando os perfis GR e ϕ_N para rochas argilosas e Δt e ϕ_N para rochas arenosas como dados de entrada para a predição.

Aplicação dos algoritmos nos dois poços sem dados de COT:

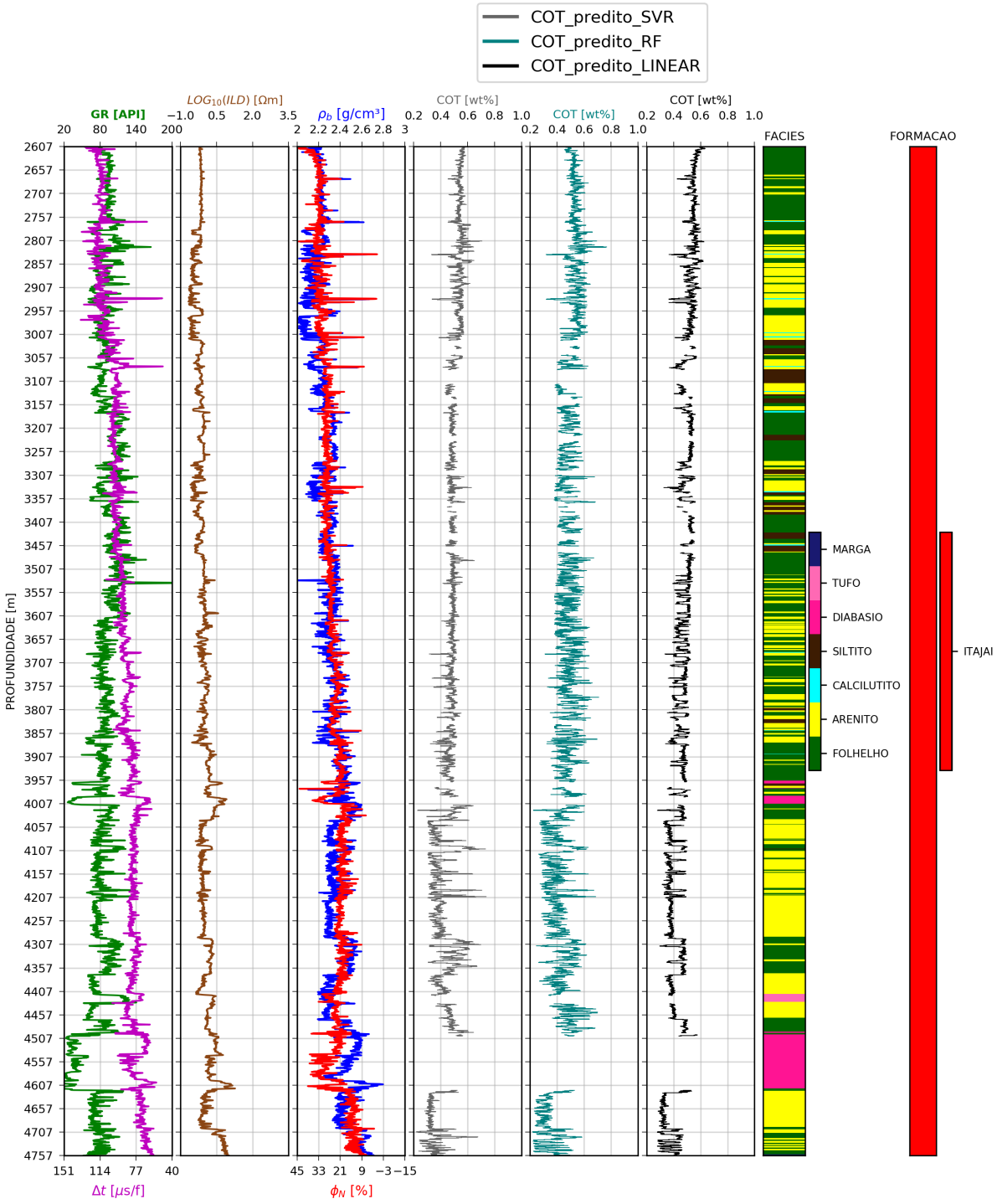


Figura 6.3: Predição do COT para o poço 3-BRSA-331-RJS.

O comportamento das predições mostra baixa variabilidade no poço 3-BRSA-331-RJS (Figura 6.3), posto que os perfis atravessam apenas a formação Itajaí, indicando que possivelmente não há folhelhos geradores nessas regiões indicadas. Enquanto isso, o COT predito para o poço 4-BRSA-144-RJS (Figura 6.4) exibe uma maior variabilidade, chegando a ul-

trapassar 0.8 % na região dos calcilutitos, indicando um maior acúmulo de matéria orgânica que no poço 3-BRSA-331-RJS.

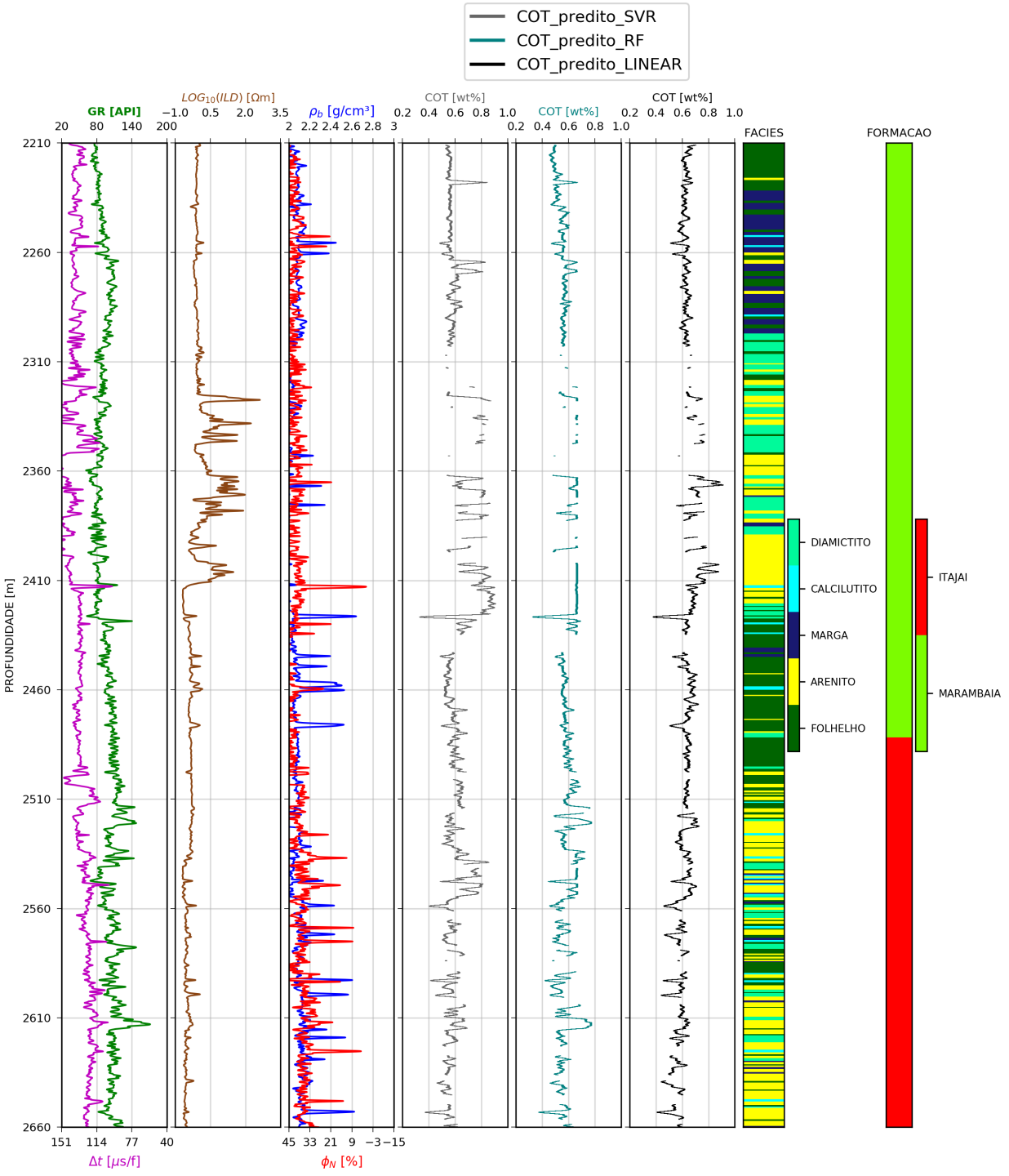


Figura 6.4: Predição do COT para o poço 4-BRSA-144-RJS.

Após exibidos os resultados da melhor combinação de perfis (aquela que retornou o menor MSE na validação dentre todas as combinações), serão descritos agora os resultados

da aplicação dos cinco perfis como dados de entrada. A Figura 6.5 exibe o resultado do COT medido com o COT calculado. Em (a), (b) e (c) estão expostos os valores de R para os grupos de rochas argilosas e arenosas, ao passo que em (d), (e) e (f) são exibidas as correlações globais para os dados do poço teste, para cada algoritmo.

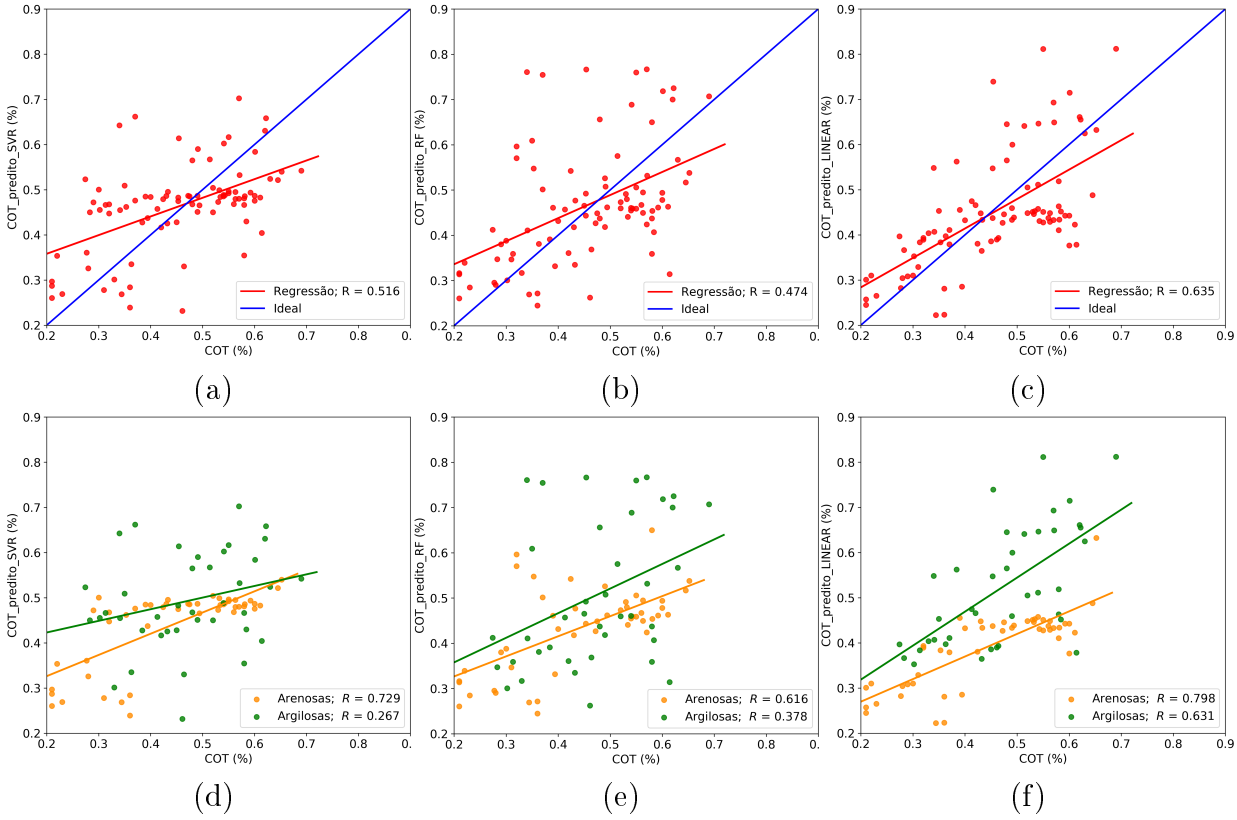


Figura 6.5: Comparação entre os resultados do COT predito e medido em laboratório, utilizando os cinco perfis disponíveis: (a) comparação entre o COT predito por SVR e medido em laboratório; (b) comparação entre o COT predito por RF e medido em laboratório; (c) comparação entre o COT predito por RLM e medido em laboratório; em (d), (e) e (f) são exibidos os mesmos gráficos respectivos de cima, separados em grupos de rochas argilosas e arenosas.

Utilizando-se todos os perfis como dados de entrada, os hiperparâmetros que retornaram o menor erro na validação cruzada para o algoritmo SVR e RF, estão representados respectivamente nas tabelas 6.5 e 6.6.

Tabela 6.5: Hiperparâmetros ótimos para o algoritmo SVR, nas dados de rochas argilosas e arenosas, usando-se os 5 perfis como dados de entrada.

Hiperparâmetros	**Rochas argilosas**	**Rochas arenosas**
C	0.1	1
γ	1	0.1

Tabela 6.6: Hiperparâmetros ótimos para o algoritmo RF, nas dados de rochas argilosas e arenosas, usando-se os 5 perfis como dados de entrada.

Hiperparâmetros	**Rochas argilosas**	**Rochas arenosas**
n	100	190
m	80	42
l	3	2
s	3	5
h	5	5

Para rochas argilosas, utilizando os cinco perfis, novamente a RLM retorna o maior valor de R e menores valores de MAE e RMSE, no dado de teste (Tabela 6.7), indicando menor dispersão na predição, no poço teste. Em contra partida, o SVR retornou o menor valor de R.

Tabela 6.7: resultado da predição dos algoritmos para as rochas argilosas, utilizando os perfis GR, ϕ_N, ρ_b, Δt e ILD.

	Dado de treinamento			**Dado de teste**			**poço teste**		
Algoritmo de regressão	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**
SVR	0.6389	0.20	0.31	0.382	0.24	0.35	0.267	0.10	0.131
RF	0.77	0.171	0.274	0.18	0.254	0.35	0.378	0.116	0.153
LINEAR	0.244	0.27	0.365	0.08	0.256	0.326	0.631	0.097	0.116

Já para as rochas arenosas Ttabela 6.8), o SVR e a RLM retornaram métricas do erro próximas para o poço teste, mas a RLM retorna um valor de R razoavelmente maior (indicando menor dispersão).

Tabela 6.8: resultado da predição dos algoritmos para as rochas arenosas, utilizando os perfis GR, ϕ_N, ρ_b, Δt e ILD.

	Dado de treinamento			Dado de teste			poço teste		
Algoritmo de regressão	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**	**R**	**MAE (%)**	**RMSE (%)**
SVR	0.56	0.123	0.18	0.504	0.128	0.19	0.73	0.08	0.089
RF	0.817	0.088	0.145	0.52	0.11	0.164	0.616	0.09	0.108
LINEAR	0.425	0.14	0.20	0.53	0.11	0.16	0.80	0.08	0.096

Algo que acontece utilizando-se todos os perfis, é uma superestimação dos valores de COT nos folhelhos da Formação Marambaia, retornado pela RLM, diferentemente do que ocorreu quando se utilizou os pares de perfis. Em contra partida, os outros dois algoritmos exibem comportamento semelhante ao obtido com os pares de perfis, como é exibido na Figura 6.6. Essa resposta da RLM não é muito confiável, visto que nos poços de treinamento exibidos no Apêndice A, os valores de COT medidos na Formação Marambaia não excedem 1 %, ficando quase sempre abaixo de 0.7 % nos parcos dados. Além disso, quando se utilizou apenas o par de perfis, a predição por SVR e RF nas rochas argilosas na região inferior da Formação Itajaí-Açú (entre 4130 e 4280 m) retornou resultados mais próximos dos dados medidos do que quando utilizou-se os cinco perfis, ao passo que a RLM não conseguiu mapear o comportamento destes folhelhos geradores. Foi perceptível também que o COT predito pelo algoritmo RF sofreu um leve deslocamento para a direita em quase todo o perfil, quando comparado com o COT predito utilizando apenas o par de perfis, em contra partida o COT predito pelo SVR permanece praticamente inalterado, garantindo maior estabilidade que o RF e a RLM, mostrando assim que os pares de perfis utilizados são suficientes para representar o COT em todo o intervalo com este algoritmo.

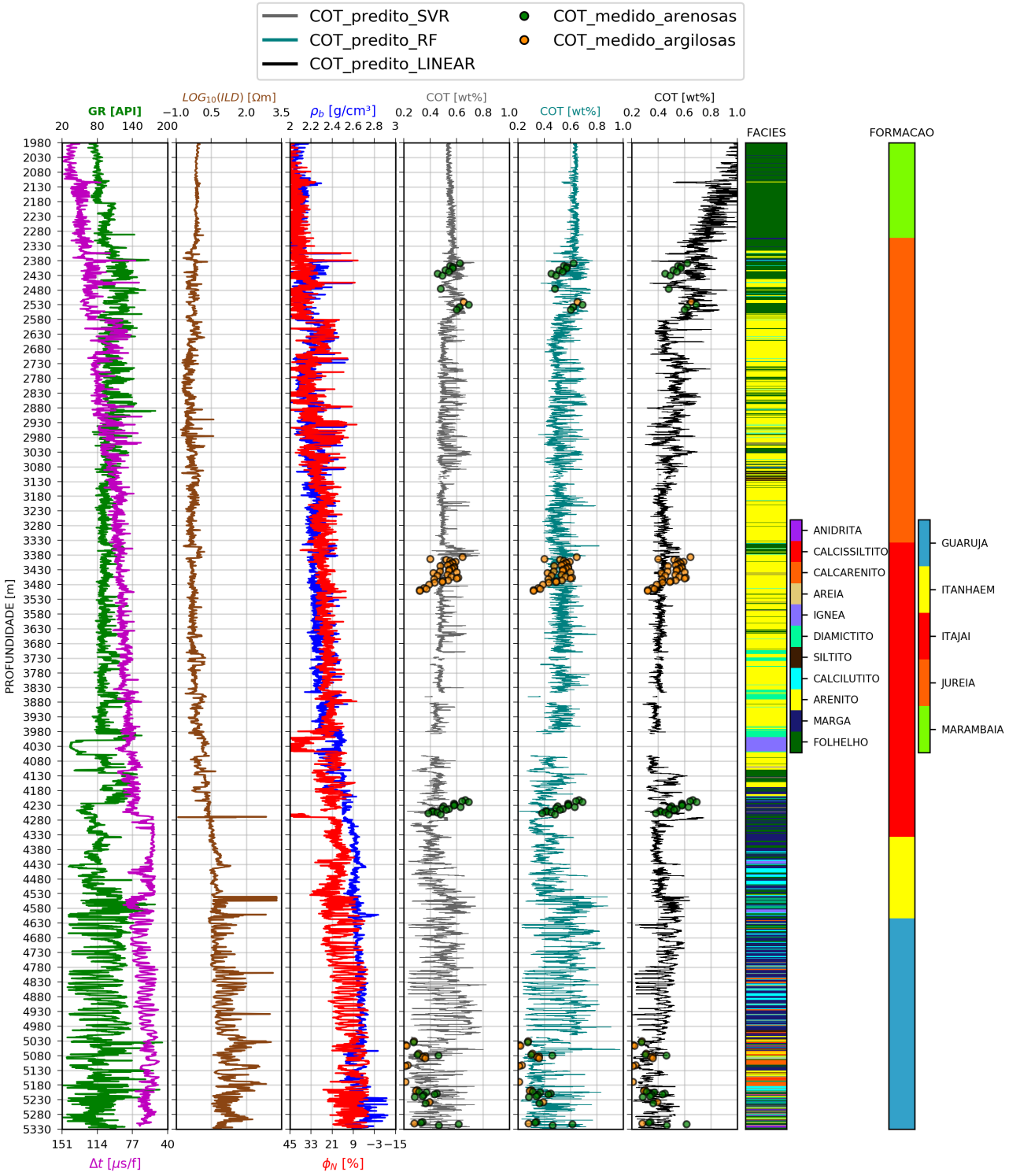


Figura 6.6: Plotagem do COT predito para todo o intervalo perfilado, nas regiões de folhelhos, margas, calcilutitos, arenitos e calcarenitos, utilizando todos os perfis GR, DT, NPHI, ILD e ρ_b rochas argilosas e arenosas como dados de entrada para a predição.

Aplicação dos algoritmos nos dois poços sem dados de COT:

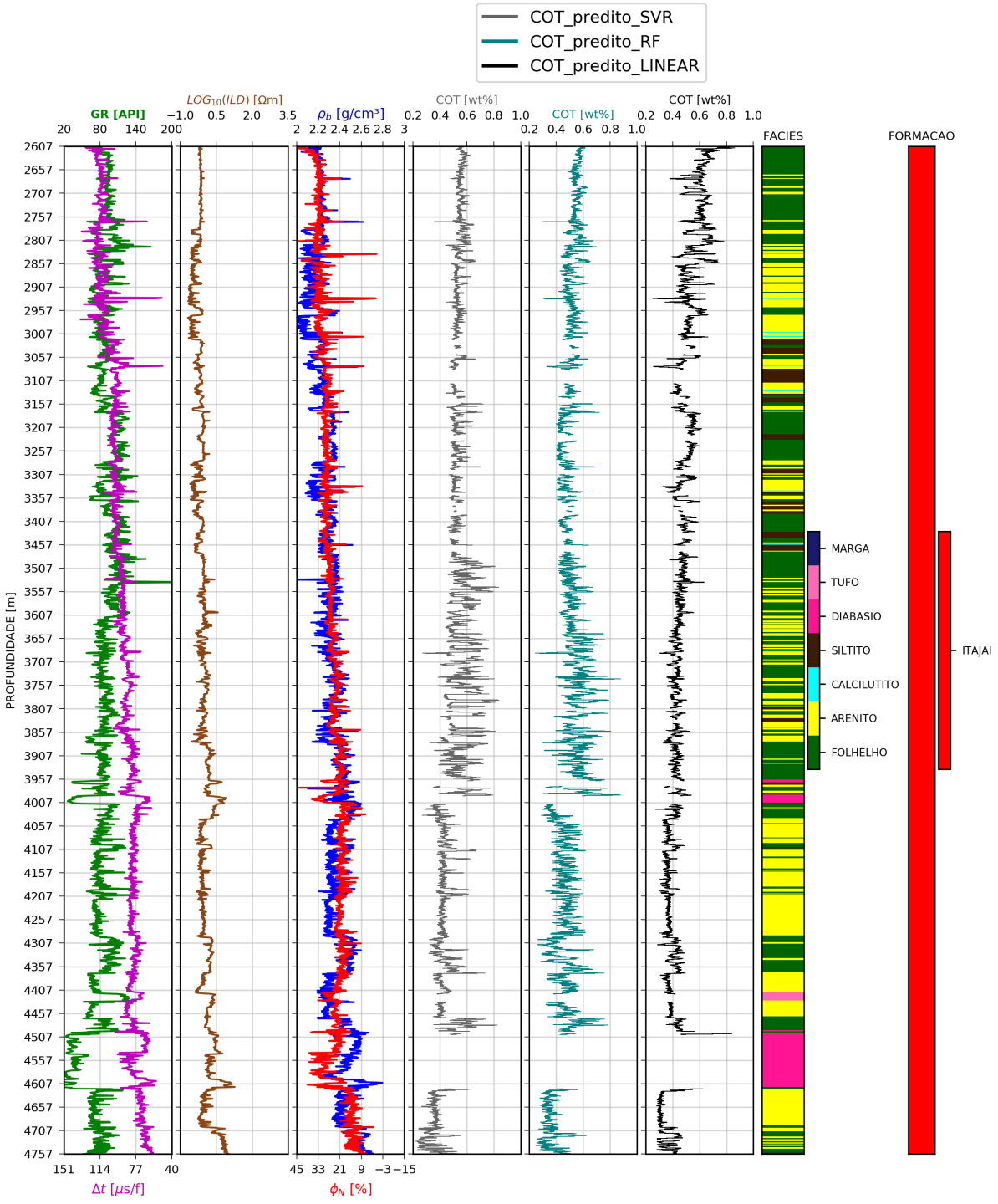


Figura 6.7: Predição do COT para o poço 3-BRSA-331-RJS.

Utilizando-se todos os perfis, a predição do COT para o poço 3-BRSA-331-RJS exibe maior variabilidade do COT do que quando se utilizou aoenas os pares de perfis, chegando

à valores em torno de 0.8 % entre 3657 e 3957m de profundidade para os algoritmos SVR e RF. No entanto, a RLM exibe comportamento semelhante ao resultado em 6.3. Já para o poço 4-BRSA-144-RJS, a RLM exibe valores acima de 1 % na base da Formação Marambaia e nas profundidades acima de 2290m, diferentemente dos resultados quando se utilizou o par de perfis. Enquanto isso, o SVR e RF não exibem variabilidade na predição do COT.

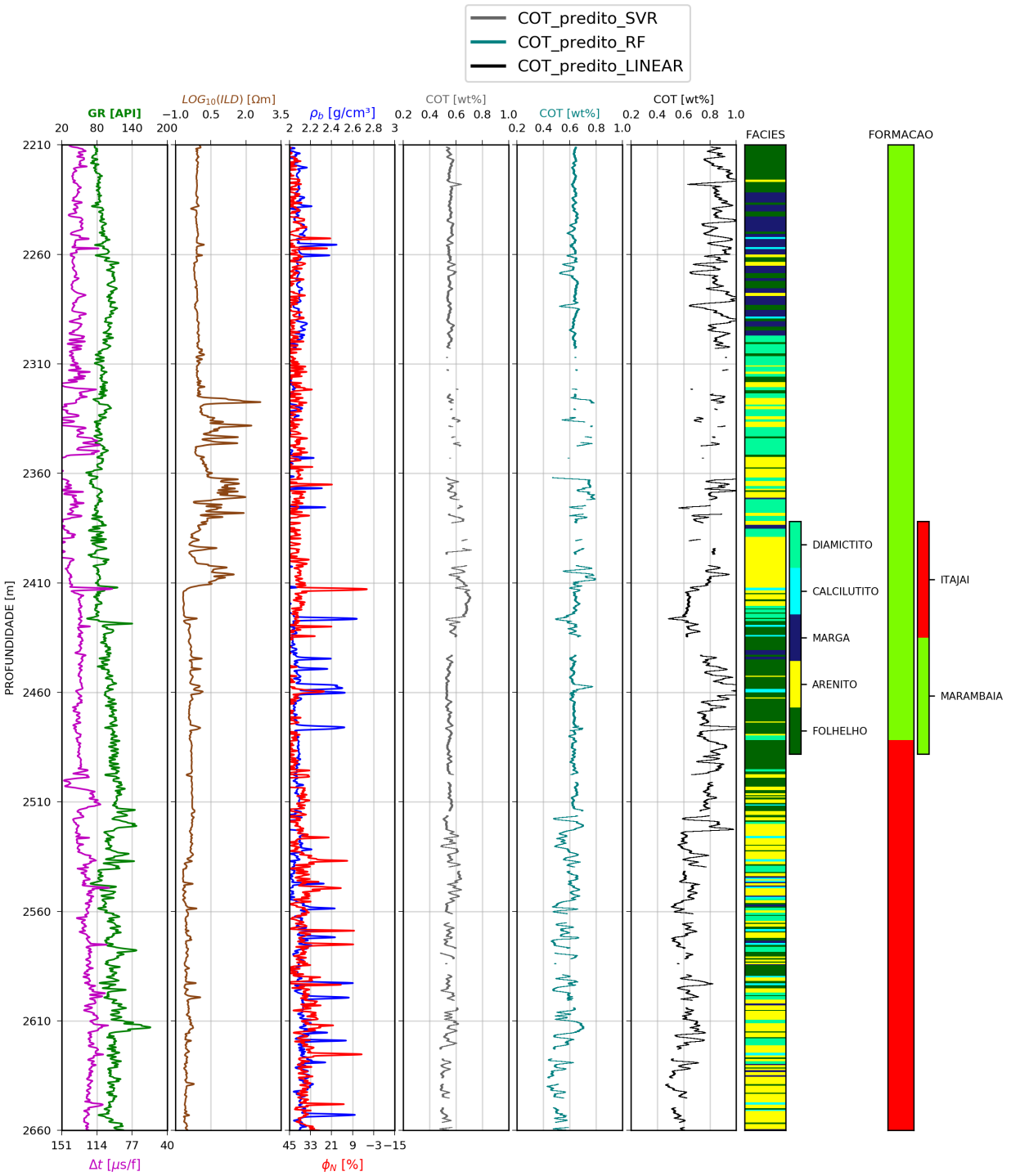


Figura 6.8: Predição do COT para o poço 3-BRSA-331-RJS.

7

Conclusões

Nossos resultados mostraram que a adição de mais perfis de poços como dados de entrada não garantem melhores resultados. Com apenas três perfis diferentes (Δt, GR e Φ_N), consegui-se um resultado melhor para a predição das formações argilosas (GR e Φ_N) e arenosas (Δt e Φ_N), retornando menores erros e maiores valores de correlação (R) do que quando se utilizou os cinco perfis. Além disso, é notório também que apesar da simplicidade da regressão linear frente aos algoritmos SVR e RF, retornou excelentes resultados, além de ser um algoritmo largamente utilizado nos diversos campos das ciências pela sua simplicidade e capacidade de interpretabilidade das relações entre as variáveis envolvidas. No entanto, nota-se que, analisar apenas a métrica do erro sem um vislumbre da predição pode levar à equívocos. Os algoritmos SVR e RF, apesar de exibirem erro levemente superior, conseguiram captar o padrão das rochas carbonáticas e dos folhelhos geradores da Formação Itajaí-Açú e da porção inferior da Formação Itanhaem, chegando à dobrar os valores de COT obtidos pela regressão linear múltipla. Enquanto a regressão linear retorna praticamente uma linha reta entre 0.4 e 0.6 % de COT, com valores rígidos oscilando para valores de COT ainda menores que em rochas arenosas (onde seriam esperados os menores valores), o SVR e RF oscilam entre 0.4 e 0.8 % (ou mais), nessas rochas. Outro ponto importante é que a combinação de perfis que retornou menor erro de predição exibe um perfil em comum: o Φ_N. Deve-se à isso o fato do perfil neutrônico ter uma relação direta com o índice de hidrogênio (IH) das formações rochosas, pois rochas argilosas normalmente possuem conteúdo de matéria orgânica associada, além de água nos poros dos folhelhos (responsável pelos altos valores de IH medidos pela ferramenta CNL). Além disso, no caso de sedimentos de granulação mais fina (folhelhos e carbonatos argilosos), o urânio concentra-se geralmente na matéria orgânica, e justamente essa radiação captada pelo perfil GR tende à ser proporcional ao teor

de carbono orgânico total nestas rochas. Assim sendo, o melhor resultado da combinação dos perfis GR e Φ_N para rochas argilosas poderia ser esperado, com base na literatura.

Nos arenitos e calcarenitos, o fato da melhor combinação de perfis ter sido Δt e Φ_N advém do fato das altas correlações com esses perfis. O mapeamento do COT nas rochas arenosas teve intuito meramente científico, visto que o estudo do COT nessas rochas é desinteressante do ponto de vista da avaliação do potencial gerador. O estudo foi realizado porque havia uma abundância de dados de COT medido nessas rochas e se pretendeu criar um perfil de COT contínuo, de modo que pudéssemos observar seu comportamento ao longo de todo o poço e quem sabe, realizar no futuro uma nova abordagem empíricas a partir destes resultados. Ainda assim, o estudo mostrou como rochas arenosas se comportam estatisticamente melhores que as rochas argilosas, quando analisados os *crossplots* com os cinco perfis e o COT (Figura 5.6).

Os corpos de prova que foram utilizados para medir o COT em laboratório foram coletados majoritariamente como calhas pontuais e alguns poucos em amostras laterais de rochas e testemunhos pontuais, que foram analisados pela Petrobrás. O ideal é que todas as amostras fossem coletadas por testemunhos e/ou amostras laterais de rocha, por serem mais representativos e confiáveis do que amostras de calha.

Observa-se que, apesar dos perfis convencionais conseguirem obter uma boa resposta na predição do COT, ainda há muita dificuldade de se relacionar a petrofísica com o conteúdo de carbono orgânico total das rochas. Por mais que diversos esforços tenham sido aplicados nessa direção, como por exemplo o uso do perfil espectral de raios gama, separando as contribuições radiológicas em canais de tório, urânio e potássio na tentativa de relacionar o COT com o canal de urânio, ainda não há uma maneira eficiente capaz de descrever perfeitamente o conteúdo de carbono orgânico total de rochas geradoras apenas com estes perfis, sendo portanto necessário o uso de técnicas geoquímicas de laboratório. Os perfis podem auxiliar, no entanto, numa abordagem geral, dando uma resposta imediata de valores aproximados do COT num poço, aplicando algoritmos de *machine learning*.

O desafio da geofísica e de outras ciências no futuro é criar um bom banco de dados, pois quanto menor o erro nas medidas, mais representativo e confiáveis serão os dados para os futuros profissionais que enveredarem pelo campo das geociências, incluindo aí os cientistas de dados.

Agradecimentos

Gostaria de agradecer à todos os que sempre me motivaram chegar até aqui, finalizando mais uma árdua etapa nesta vida. Ao meu pai, pelo exemplo de homem de excelente conduta moral e pelos ensinamentos desde a mais tenra idade. À minha mãe, por me ensinar o significado da palavra resiliência na prática, e por me estender as mãos com ternura, sempre que precisei. Aos queridos e amados irmãos e irmã pelas conversas, desabafos e por nunca me permitirem sentir só. Aos amigos queridos pela confiança que depositam e sempre depositaram em mim e que sei que sempre poderei contar. Gostaria de agradecer também ao apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pela concessão de bolsa de pesquisa, bem como ao Projeto PIE00005/2016 do Edital de Infraestrutura 003/2015 da FAPESB-Fundação de Amparo à Pesquisa do Estado da Bahia, pelo fornecimento de máquina computacional necessária para o fomento desta dissertação. Por fim, gostaria de agradecer ao meu orientador Marcos Vasconcelos, pela paciência e pelos ensinamentos no âmbito das tecnicidades do mundo acadêmico e por ser um exemplo de profissional para mim. Essa vitória é de todos nós!

Apêndice A

Poços utilizados no Treinamento

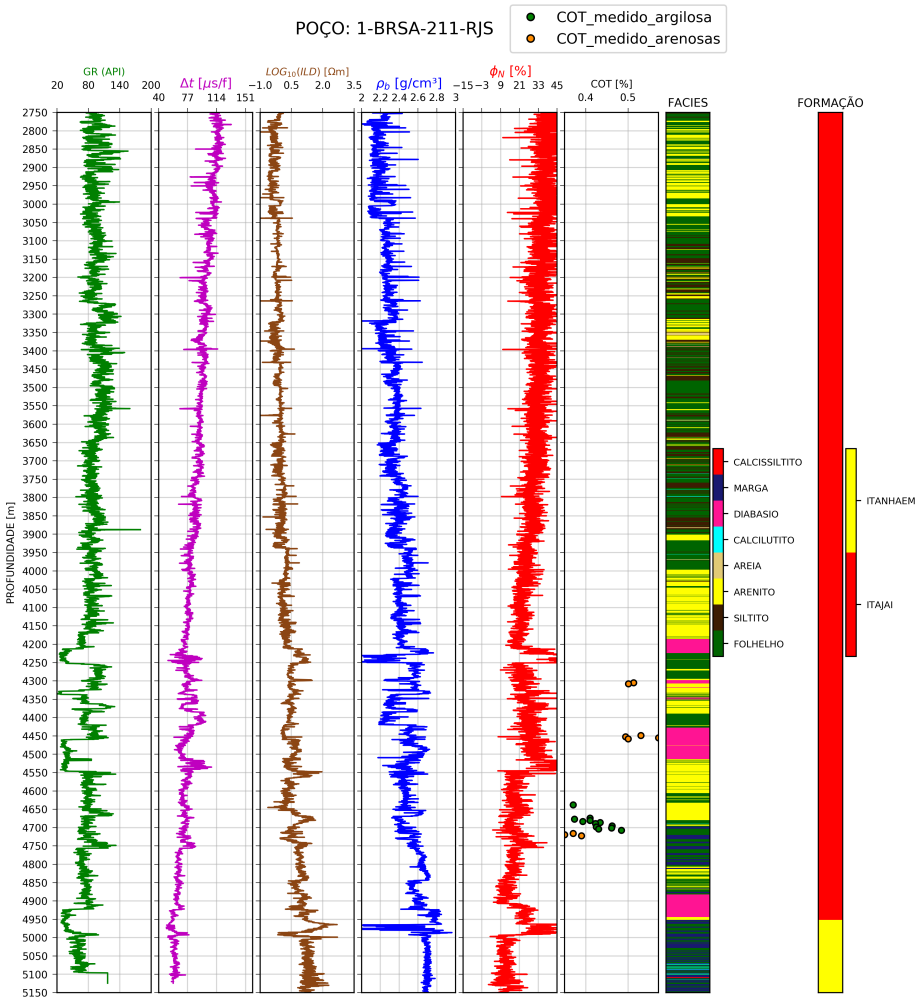


Figura A.1

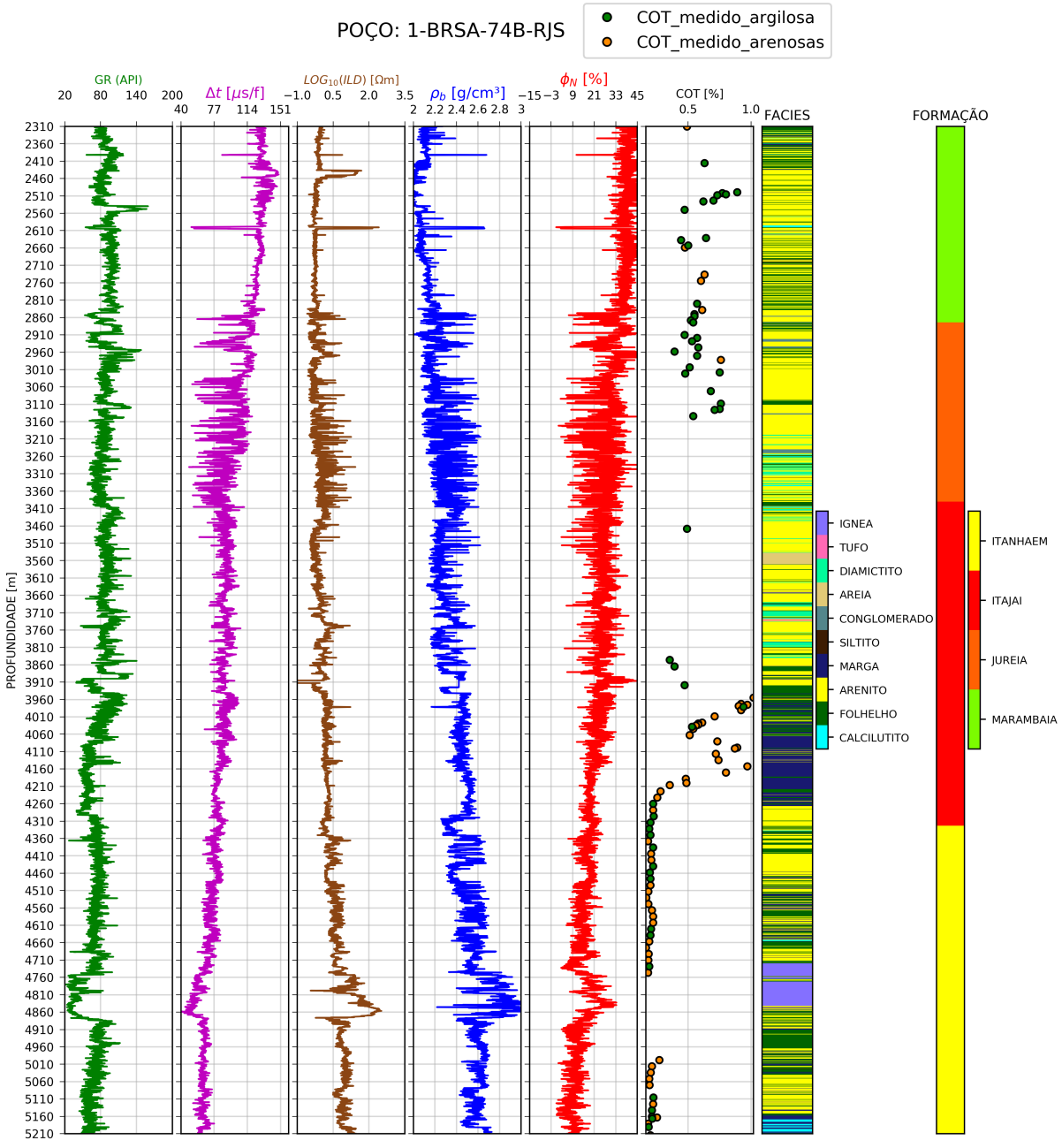

POÇO: 1-BRSA-74B-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
IGNEA
TUFO
DIAMICTITO
AREIA
CONGLOMERADO
SILTITO
MARGA
ARENITO
FOLHELHO
CALCILUTITO
ITANHAEM
ITAJAI
JUREIA
MARAMBAIA

Figura A.2

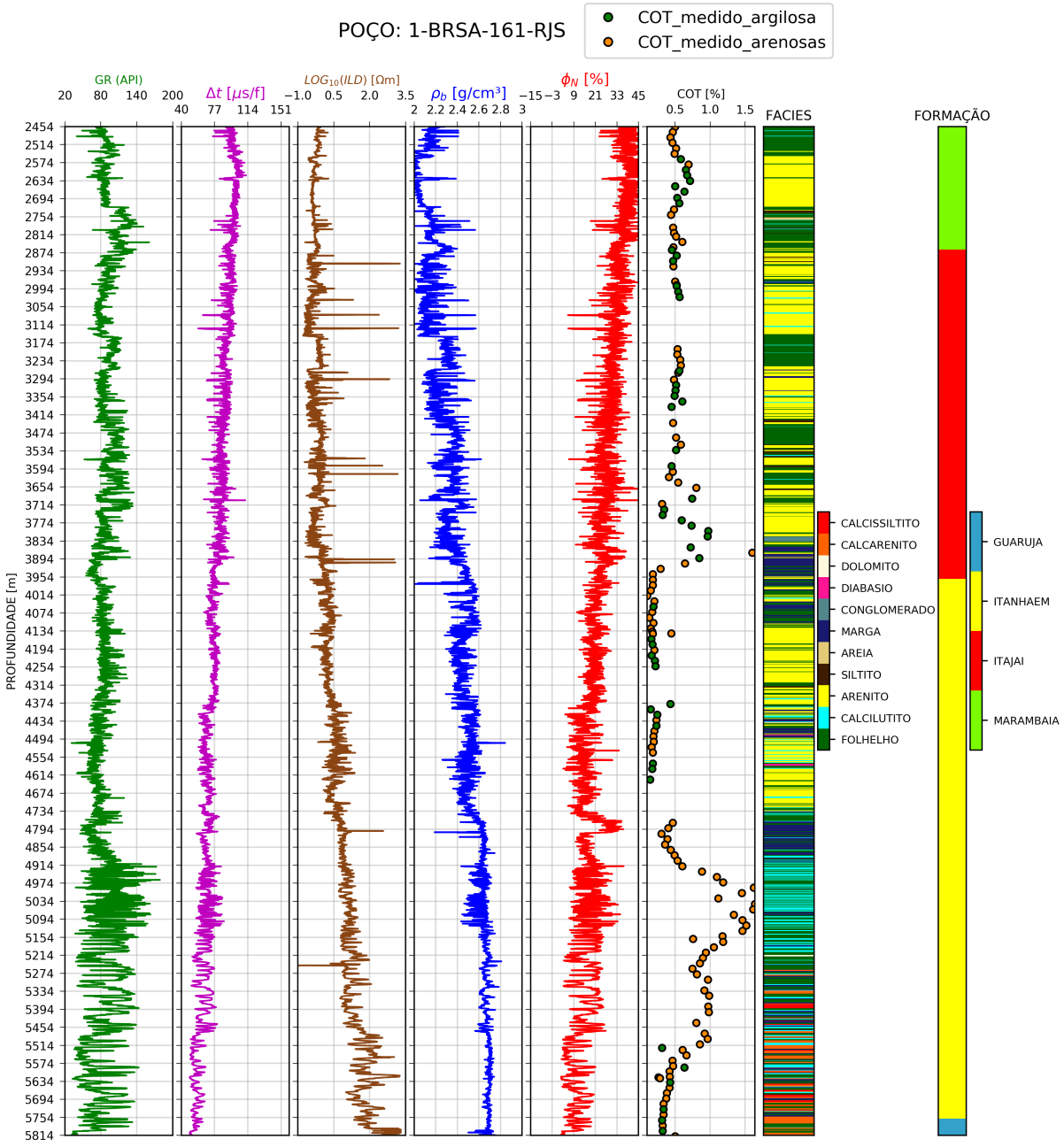

POÇO: 1-BRSA-161-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
CALCISSILTITO
CALCARENITO
DOLOMITO
DIABASIO
CONGLOMERADO
MARGA
AREIA
SILTITO
ARENITO
CALCILUTITO
FOLHELHO
GUARUJA
ITANHAEM
ITAJAI
MARAMBAIA

Figura A.3

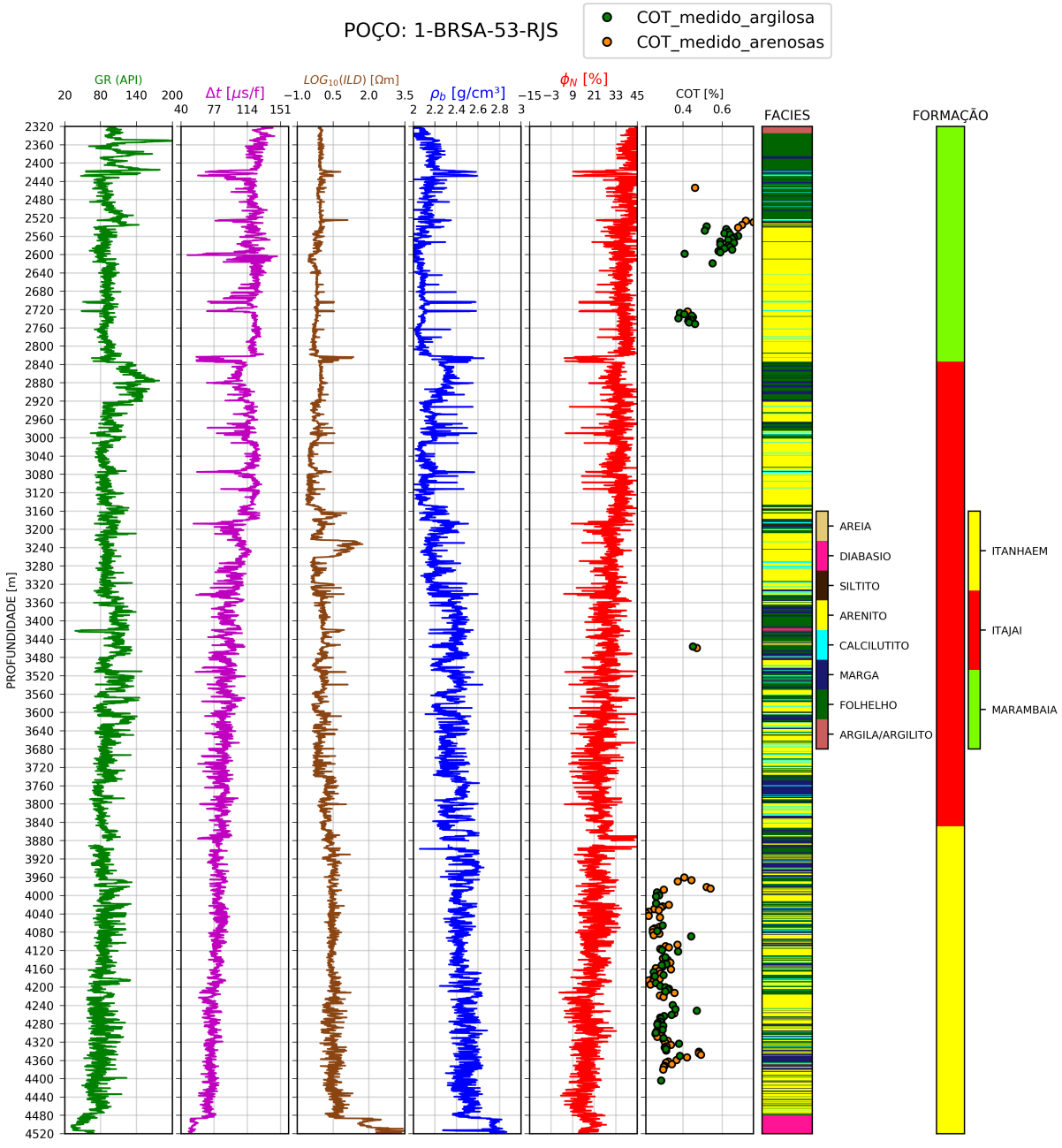

POÇO: 1-BRSA-53-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
AREIA
DIABASIO
SILTITO
ARENITO
CALCILUTITO
MARGA
FOLHELHO
ARGILA/ARGILITO
ITANHAEM
ITAJAI
MARAMBAIA

Figura A.4

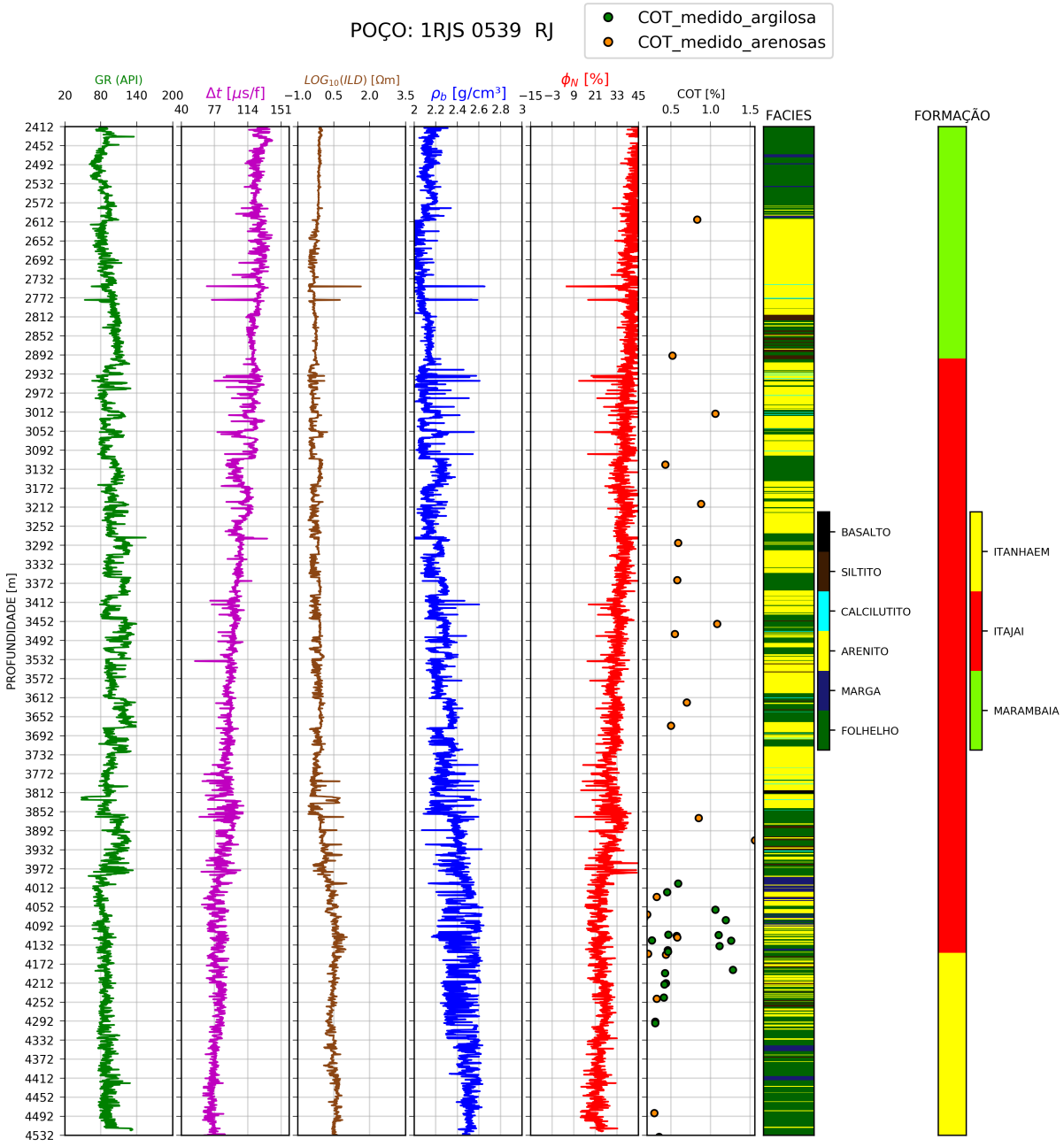

POÇO: 1RJS 0539 RJ
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
BASALTO
SILTITO
CALCILUTITO
ARENITO
MARGA
FOLHELHO
ITANHAEM
ITAJAI
MARAMBAIA

Figura A.5

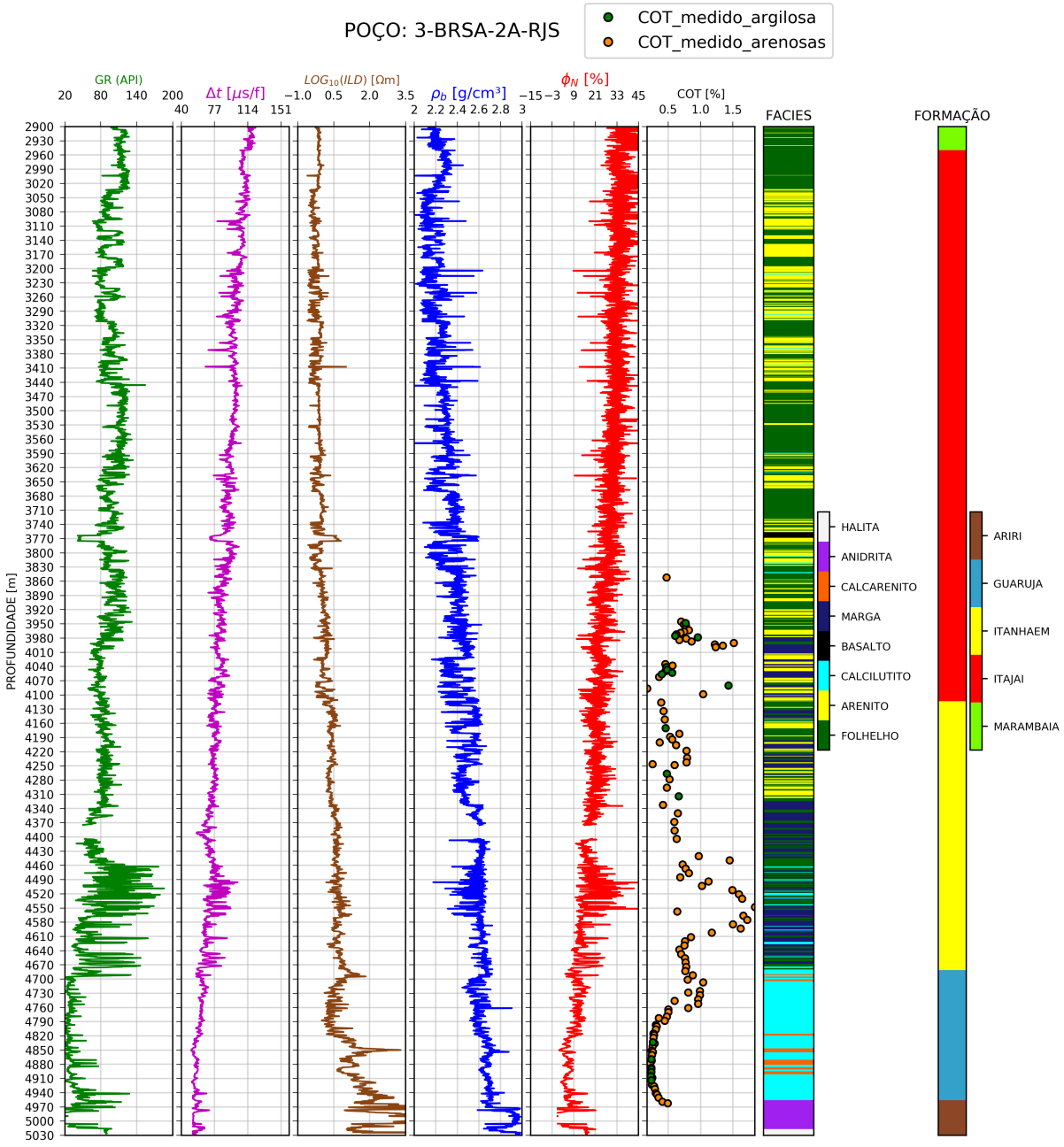

POÇO: 3-BRSA-2A-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
HALITA
ANIDRITA
CALCARENITO
MARGA
BASALTO
CALCILUTITO
ARENITO
FOLHELHO
ARIRI
GUARUJA
ITANHAEM
ITAJAI
MARAMBAIA

Figura A.6

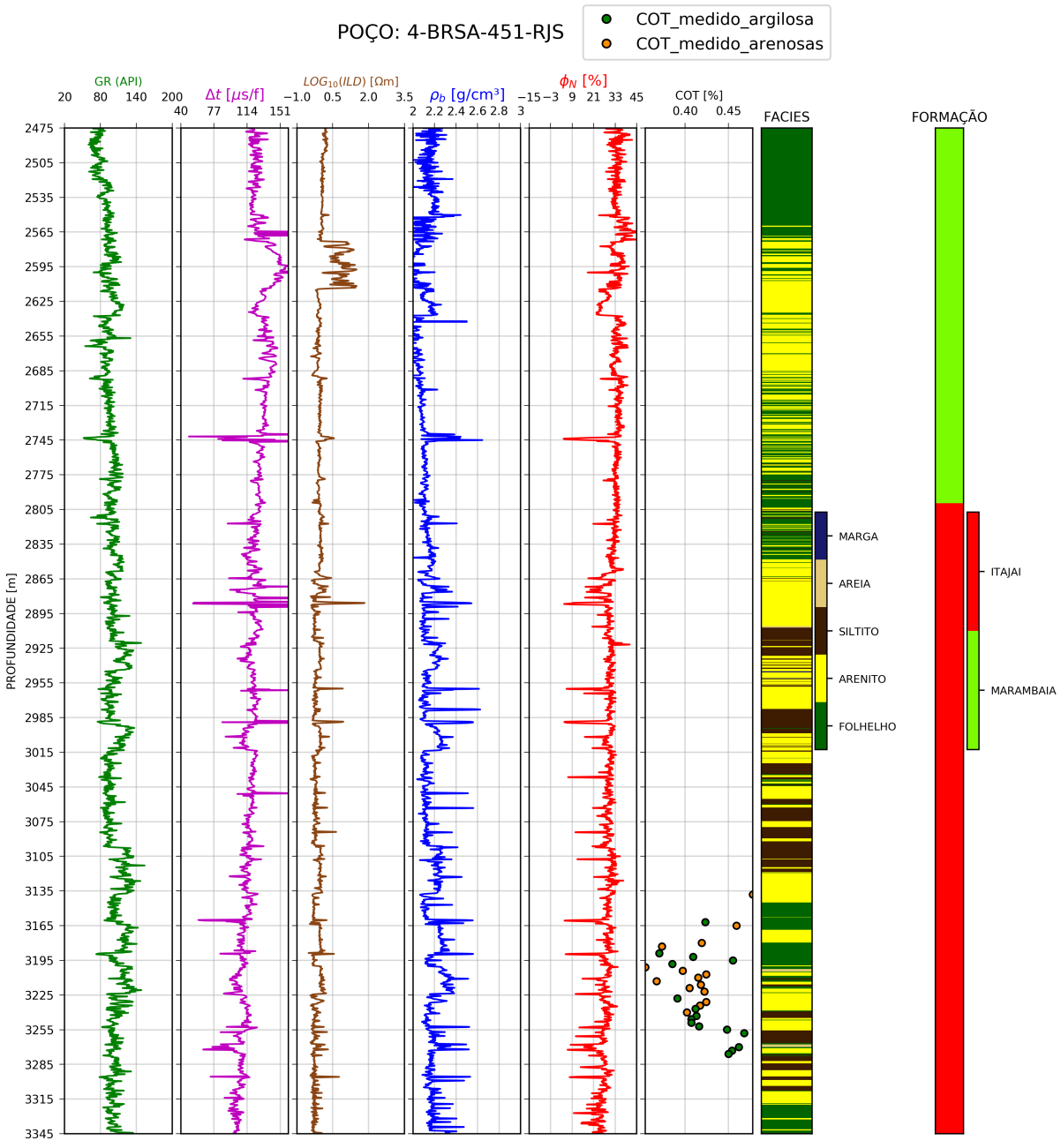

POÇO: 4-BRSA-451-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
MARGA
AREIA
SILTITO
ARENITO
FOLHELHO
ITAJAI
MARAMBAIA

Figura A.7

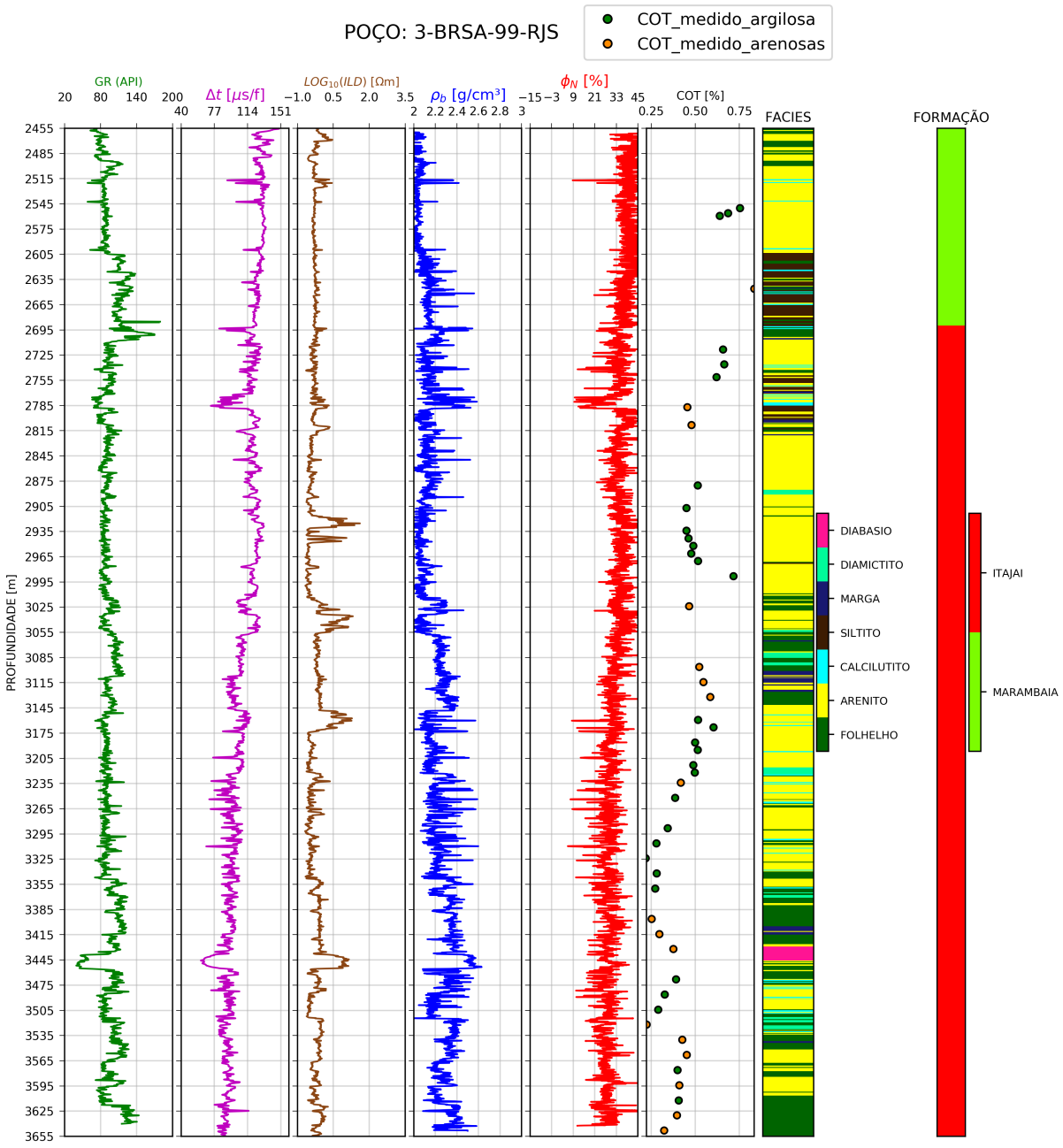

POÇO: 3-BRSA-99-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
20 80 140 200
Δt [μs/f]
40 77 114 151
LOG10(ILD) [Ωm]
−1.0 0.5 2.0 3.5
ρb [g/cm³]
2 2.2 2.4 2.6 2.8 3
ϕN [%]
−15 −3 9 21 33 45
COT [%]
0.25 0.50 0.75
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
2455 2485 2515 2545 2575 2605 2635 2665 2695 2725 2755 2785 2815 2845 2875 2905 2935 2965 2995 3025 3055 3085 3115 3145 3175 3205 3235 3265 3295 3325 3355 3385 3415 3445 3475 3505 3535 3565 3595 3625 3655
DIABASIO
DIAMICTITO
MARGA
SILTITO
CALCILUTITO
ARENITO
FOLHELHO
ITAJAI
MARAMBAIA

Figura A.8

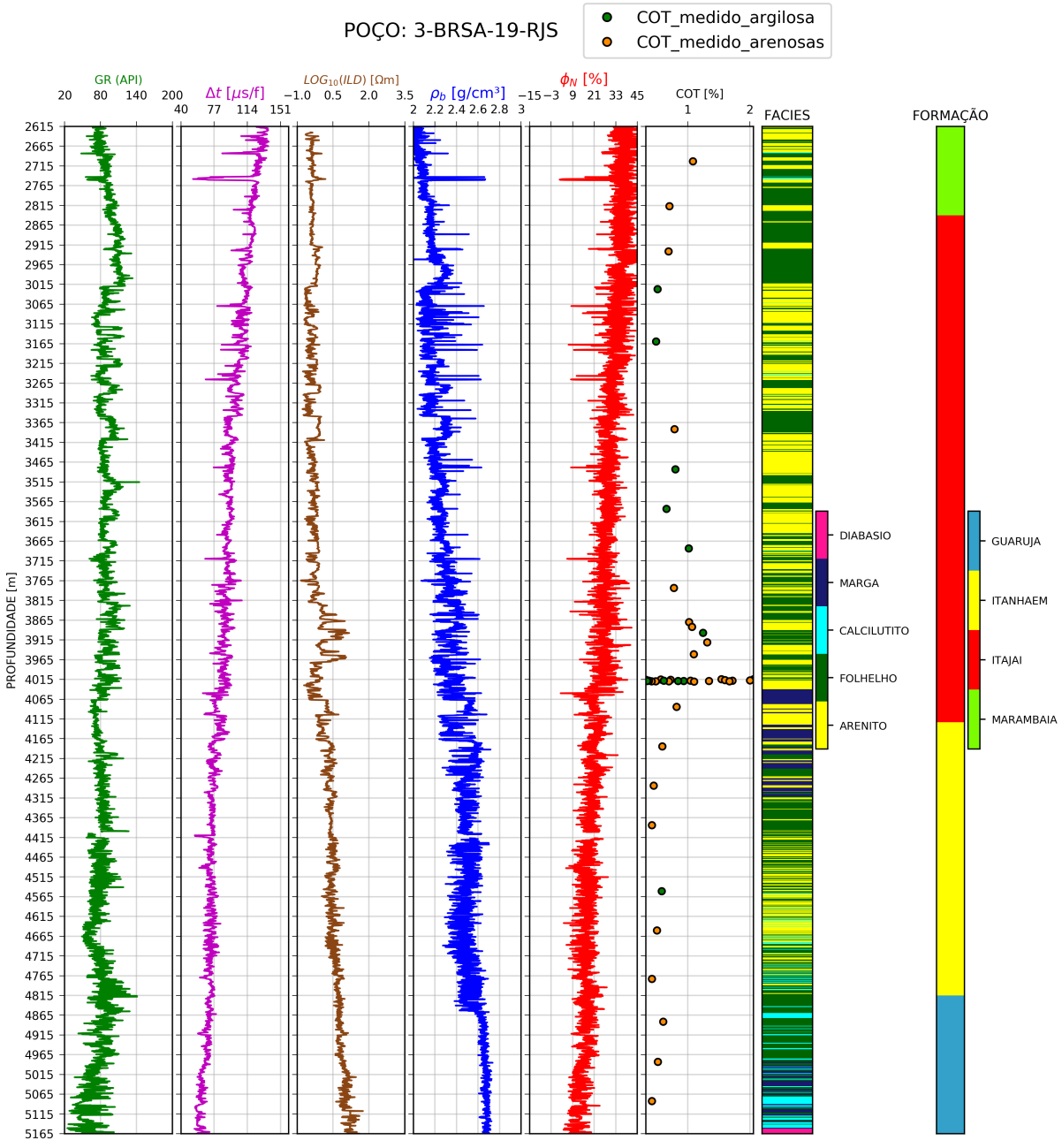

POÇO: 3-BRSA-19-RJS
COT_medido_argilosa
COT_medido_arenosas
GR (API)
Δt [μs/f]
LOG10(ILD) [Ωm]
ρb [g/cm³]
ϕN [%]
COT [%]
FACIES
FORMAÇÃO
PROFUNDIDADE [m]
DIABASIO
MARGA
CALCILUTITO
FOLHELHO
ARENITO
GUARUJA
ITANHAEM
ITAJAI
MARAMBAIA

Figura A.9

Apêndice B

cross-plots dos dados de treinamento.

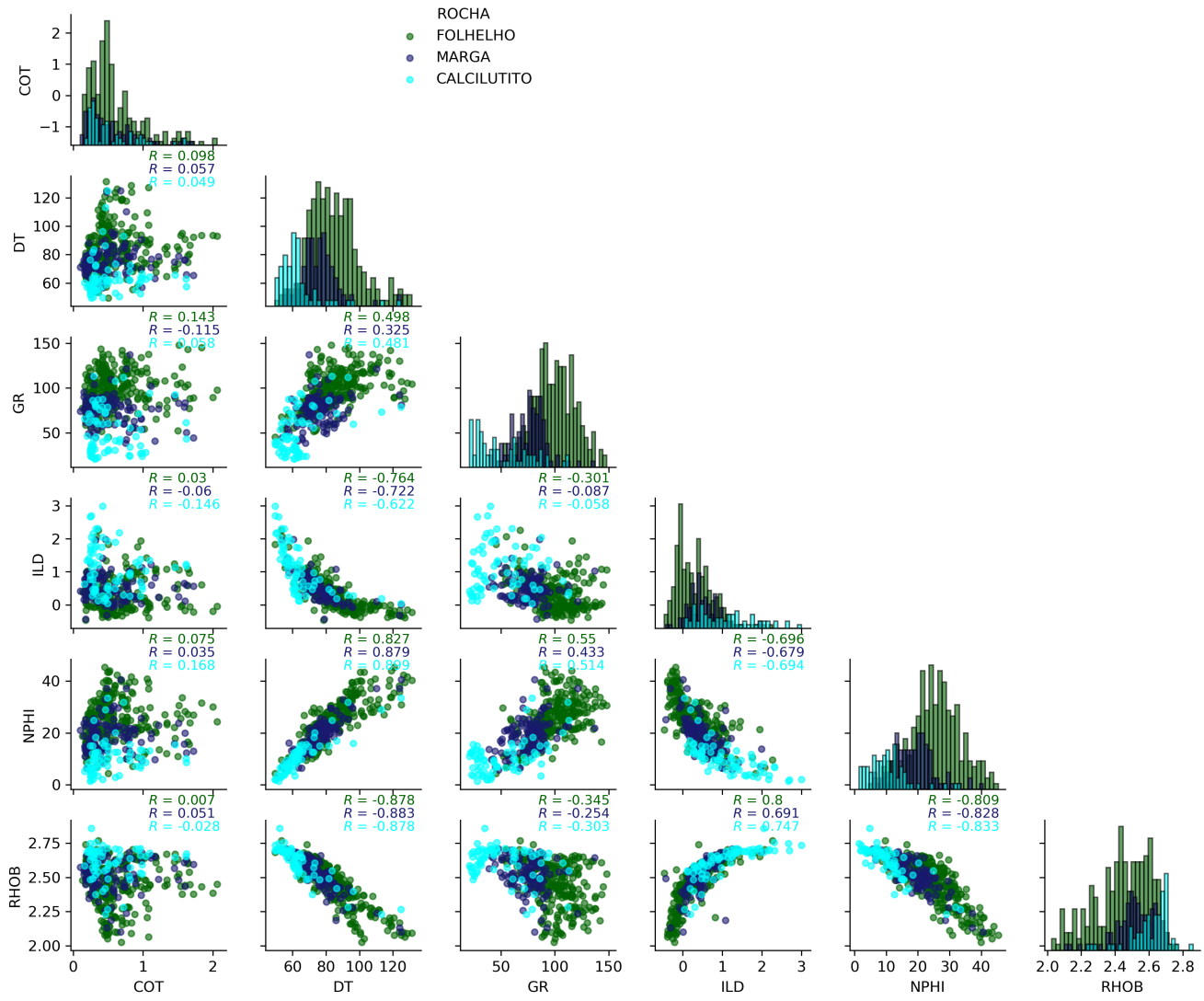


Figura B.1: Rochas argilosas

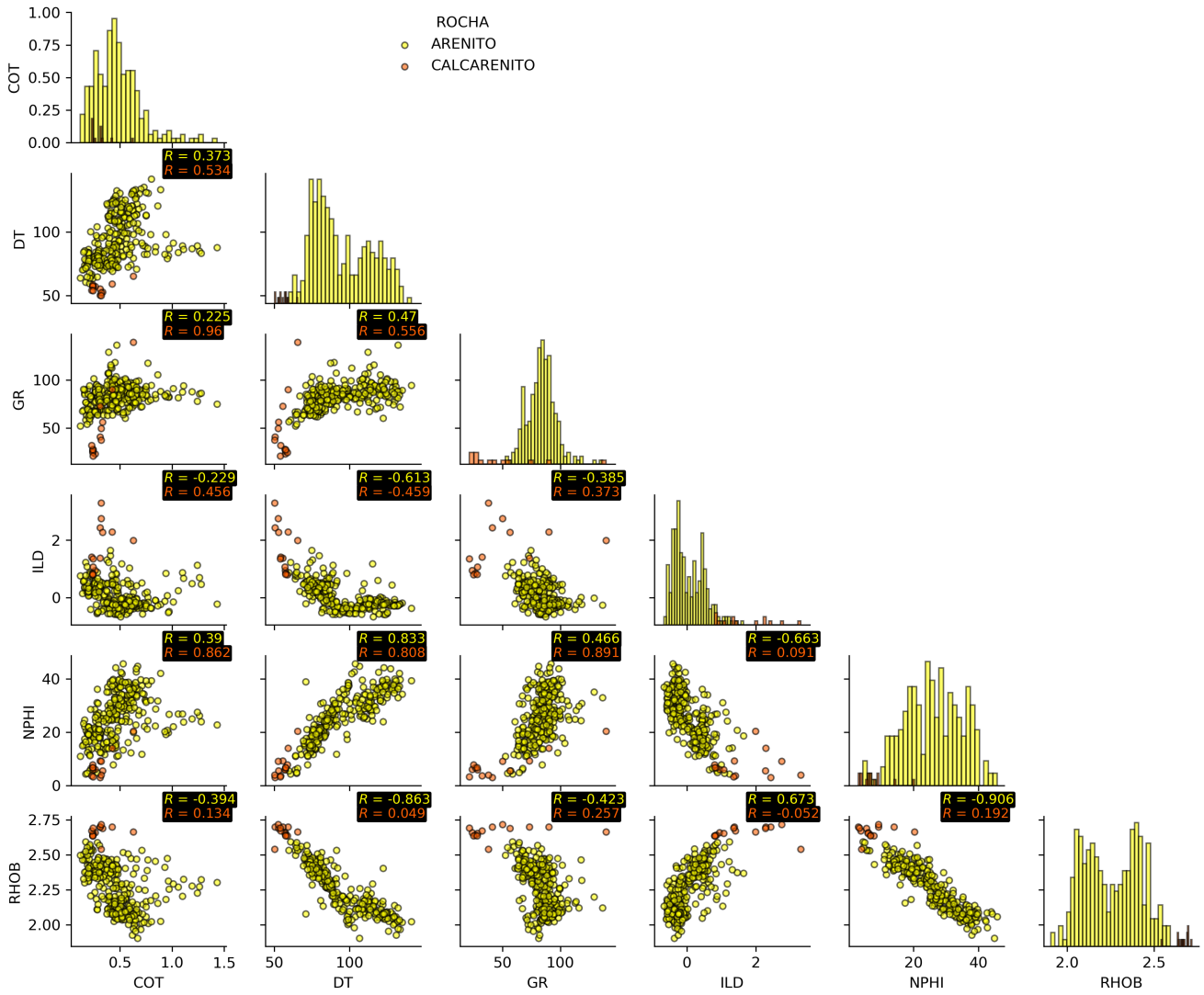


Figura B.2: Rochas arenosas

Apêndice C

cross-plots dos dados do poço teste.

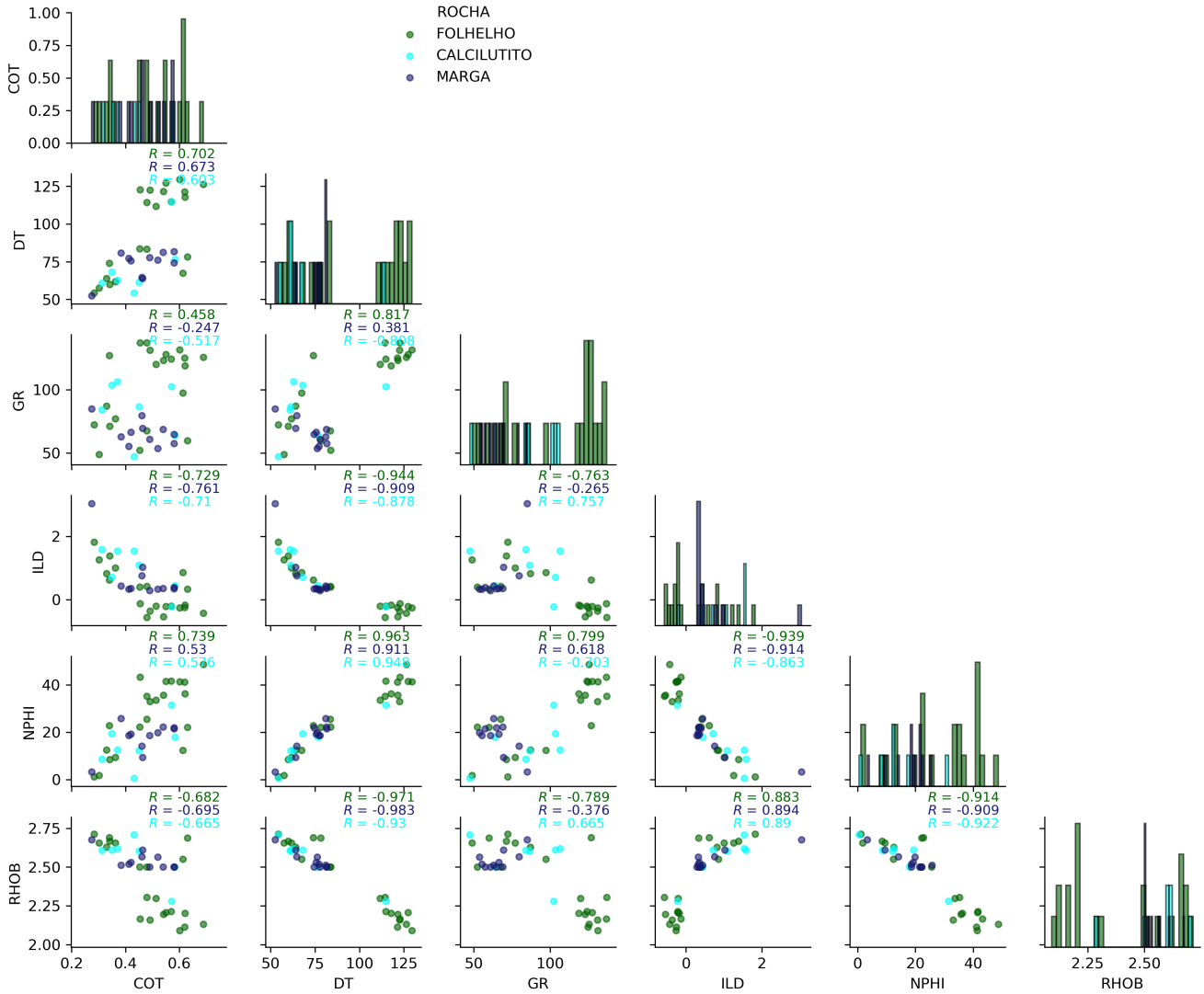


Figura C.1: Rochas argilosas

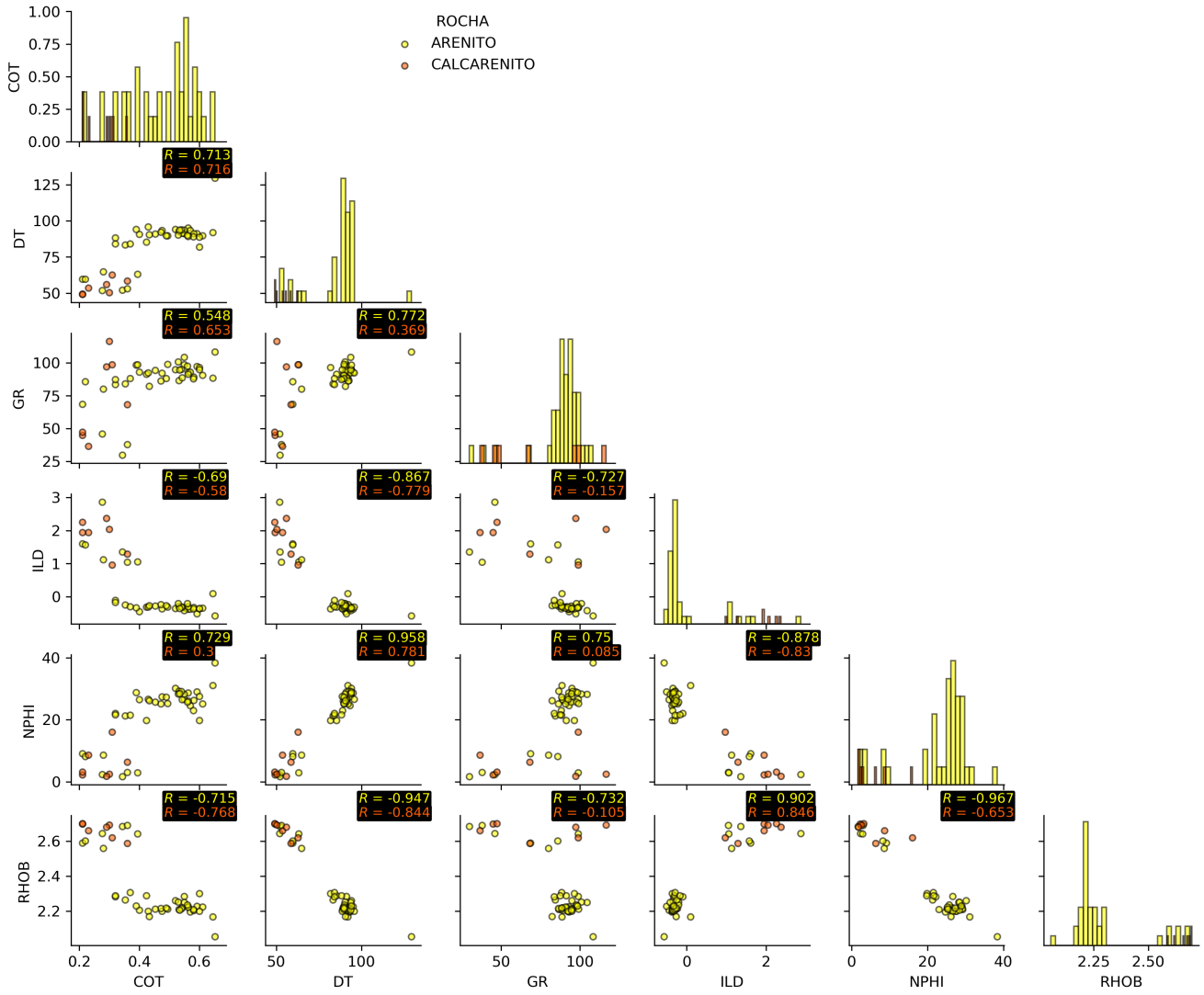


Figura C.2: Rochas arenosas

Referências Bibliográficas

Amiri Bakhtiar, H.; Telmadarreie, A.; Shayesteh, M.; Heidari Fard, M.; Talebi, H. e Shirband, Z. (2011) Estimating total organic carbon content and source rock evaluation, applying δlogr and neural network methods: Ahwaz and marun oilfields, sw of iran, Petroleum Science and Technology, **29**(16):1691–1704.

Araing, M. (1988) Geochemical reconnaissance of the mid-cretaceous anoxic event in the santos basin, brazil, Rev. Bras. Geocienc., **18**(3):273–282.

Asquith, G. e Gibson, C. (1982) Basic well log analysis for geologists, Methods in exploration series, American Association of Petroleum Geologists, ISBN 9780891816522.

Assine, M. L.; Corrêa, F. S. e Chang, H. K. (2008) Migração de depocentros na bacia de santos: importância na exploração de hidrocarbonetos, Revista Brasileira de Geociências, **38**(2 suppl):111–127.

Barandiaran, I. (1998) The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell, **20**(8):1–22.

Basak, D.; Pal, S. e Patranabis, D. C. (2007) Support vector regression, Neural Information Processing-Letters and Reviews, **11**(10):203–224.

Beers, R. F. (1945) Radioactivity and organic content of some paleozoic shales, AAPG Bulletin, **29**(1):1–22.

Bergstra, J. e Bengio, Y. (2012) Random search for hyper-parameter optimization, Journal of machine learning research, **13**(Feb):281–305.

Bessereau, G.; Carpentier, B.; Huc, A. et al. (1991) Wireline logging and source rocks-estimation of organic carbon content by the carbolbg@ method, The Log Analyst, **32**(03).

Breiman, L. (2001) Random forests, Machine learning, **45**(1):5–32.

Caldas, M. (2007) Reconstituição cinemática e tectono-sedimentação associada a Domos salinos na águas profundas da Bacia de Santos, Tese de Doutorado, Dissertação de Mestrado. Universidade Federal do Rio de Janeiro.

Carvalho, B. (2005) Novas estratégias para detecçao automática de vetores de suporte em least squares support vector machines, Tese de Doutorado, Master's thesis, PPGEE-UFMG.

Chamasemani, F. F. e Singh, Y. P. (2011) Multi-class support vector machine (svm) classifiers–an application in hypothyroid detection and classification, In: *2011 Sixth International Conference on Bio-Inspired Computing: Theories and Applications*, pp. 351–356, IEEE.

Chang, H. K.; Assine, M. L.; Corrêa, F. S.; Tinen, J. S.; Vidal, A. C. e Koike, L. (2008) Sistemas petrolíferos e modelos de acumulação de hidrocarbonetos na bacia de santos, Revista Brasileira de Geociências, **38**(2 suppl):29–46.

Claesen, M. e De Moor, B. (2015) Hyperparameter search in machine learning, arXiv preprint arXiv:1502.02127.

Clemente, P. (2013) Petroleum geology of the campos and santos basins, lower cretaceous brazilian sector of the south atlantic margin.

Cortes, C. e Vapnik, V. (1995) Support-vector networks, Machine learning, **20**(3):273–297.

Crain, E. (2002) Crain's petrophysical handbook, Spectrum 2000 Mindware Limited.

De Souza, K. G.; Fontana, R.; Mascle, J.; Macedo, J.; Mohriak, W. e Hinz, K. (1993) The southern brazilian margin: an example of a south atlantic volcanic margin, In: *3rd International Congress of the Brazilian Geophysical Society.*

Dellenbach, J.; Espitalie, J. e Lebreton, F. (1983) Source rock logging, In: *Transactions of the 8th European SPWLA Symposium, paper D.*

Dias, M. C.; VIEIRA, A. O. S.; NAKAJIMA, J. N.; PIMENTA, J. A. e LOBO, P. C. (1998) Composição florística e fitossociologia do componente arbóreo das florestas ciliares do rio iapó, na bacia do rio tibagi, tibagi, pr, Brazilian Journal of Botany, **21**(2):183–195.

Dias, M. S. (2007) O uso de Máquina de Suporte Vetorial para Regressao (SVR) na Estimaçao da Estrutura a Termo da Taxa de Juros do Brasil, Tese de Doutorado, PUC-Rio.

Efron, B. (1982) The jackknife, the bootstrap, and other resampling plans, vol. 38, Siam.

Espitalie, J.; Madec, M.; Tissot, B.; Mennig, J.; Leplat, P. et al. (1977) Source rock characterization method for petroleum exploration, In: *Offshore Technology Conference*, Offshore Technology Conference.

Fertl, W. H.; Rieke III, H. et al. (1980) Gamma ray spectral evaluation techniques identify fractured shale reservoirs and source-rock characteristics, Journal of Petroleum Technology, **32**(11):2–053.

Fertl, W. H.; Chilingar, G. V. et al. (1988) Total organic carbon content determined from well logs, SPE Formation Evaluation, **3**(02):407–419.

Gamboa, L.; Machado, M. P.; Da Silveira, D.; De Freitas, J.; Da Silva, S.; Mohriak, W.; Szatmari, P. e Anjos, S. (2008) Evaporitos estratificados no atlântico sul: interpretação sísmica e controle tectono-estratigráfico na bacia de santos, Sal: Geologia e Tectônica, Exemplos nas Basicas Brasileiras, pp. 340–359.

Gunn, S. R. et al. (1998) Support vector machines for classification and regression, ISIS technical report, **14**(1):5–16.

Herron, S.; Letendre, L. e Dufour, M. (1988) Source rock evaluation using geochemical information from wireline logs and cores, AAPG Bull.;(United States), **72**(CONF-8809346-).

Hertzog, R.; Colson, L.; Seeman, O.; O'Brien, M.; Scott, H.; McKeon, D.; Wraight, P.; Grau, J.; Ellis, D.; Schweitzer, J. et al. (1989) Geochemical logging with spectrometry tools, SPE Formation Evaluation, **4**(02):153–162.

Ho, T. K. (1995) Random decision forests, In: *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE.

Hunt, J. M. (1995) Petroleum geochemistry and geology.

Hussain, F. et al. (1987) Source rock identification in the state of kuwait using wireline logs, In: *Middle East Oil Show*, Society of Petroleum Engineers.

James, G.; Witten, D.; Hastie, T. e Tibshirani, R. (2013) An introduction to statistical learning, vol. 112, Springer.

Jarvie, D. M. (1991) Total organic carbon (toc) analysis: Chapter 11: geochemical methods and exploration.

Kadkhodaie-Ilkhchi, A.; Rahimpour-Bonab, H. e Rezaee, M. (2009) A committee machine with intelligent systems for estimation of total organic carbon content from petrophysical data: An example from kangan and dalan reservoirs in south pars gas field, iran, Computers & Geosciences, **35**(3):459–474.

Kamali, M. R. e Mirshady, A. A. (2004) Total organic carbon content determined from well logs using δlogr and neuro fuzzy techniques, Journal of Petroleum Science and Engineering, **45**(3-4):141–148.

Karush, W. (1939) Minima of functions of several variables with inequalities as side constraints, M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago.

Khoshnoodkia, M.; Mohseni, H.; Rahmani, O. e Mohammadi, A. (2011) Toc determination of gadvan formation in south pars gas field, using artificial intelligent systems and geochemical data, Journal of Petroleum Science and Engineering, **78**(1):119–130.

Kleinberg, E. (1990) Stochastic discrimination, Annals of Mathematics and Artificial intelligence, **1**(1):207–239.

Kleinberg, E. et al. (1996) An overtraining-resistant stochastic modeling method for pattern recognition, The annals of statistics, **24**(6):2319–2349.

Kohavi, R.; John, G. H. et al. (1997) Wrappers for feature subset selection, Artificial intelligence, **97**(1-2):273–324.

Kuhn, M. e Johnson, K. (2013) Applied predictive modeling, vol. 26, Springer.

Kvenvolden, K. A. (2006) Organic geochemistry–a retrospective of its first 70 years, Organic geochemistry, **37**(1):1–11.

Lima, C. A. d. M. et al. (2004) Comitê de máquinas: uma abordagem unificada empregando máquinas de vetores-suporte.

Liu, J.; Peng, P.; Huang, K. e Zhang, L. (2003) An improvement in carbolog technique and its preliminary application to evaluating organic carbon content of source rocks, Geochimica, **37**(6):58.

Luffel, D.; Guidry, F.; Curtis, J. et al. (1992) Evaluation of devonian shale with new core and log analysis methods, Journal of Petroleum Technology, **44**(11):1–192.

Meissner, F. F. (1978) Petroleum geology of the bakken formation williston basin, north dakota and montana.

Mendelzon, J.; Toksoz, M. N. et al. (1985) Source rock characterization using multivariate analysis of log data, In: *SPWLA 26th Annual Logging Symposium*, Society of Petrophysicists and Well-Log Analysts.

Meyer, B. e Nederlof, M. (1984) Identification of source rocks on wireline logs by density/-resistivity and sonic transit time/resistivity crossplots, AAPG Bulletin, **68**(2):121–129.

Milner, C. (1982) Sepm (sepm short course, 7, Houston, Texas.

Mohriak, W. U. (2003) Bacias sedimentares da margem continental brasileira, Geologia, tectônica e recursos minerais do Brasil, **2003**:87–165.

Moreira, J. L. P.; Madeira, C. V.; Gil, J. A. e Machado, M. A. P. (2007) bacia de santos, Boletim de Geociencias da PETROBRAS, **15**(2):531–549.

Myers, K. e Jenkyns, K. (1992) Determining total organic carbon contents from well logs: an intercomparison of gst data and a new density log method, Geological Society, London, Special Publications, **65**(1):369–376.

Neter, J.; Kutner, M. H.; Nachtsheim, C. J. e Wasserman, W. (1996) Applied linear statistical models, vol. 4, Irwin Chicago.

Nguyen, C.-T.; Nguyen, T.-K.; Phan, X.-H.; Le Nguyen, M. e Ha, Q. T. (2006) Vietnamese word segmentation with crfs and svms: An investigation, In: *Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation*, pp. 215–222.

Nixon, R. (1973) Oil source beds in cretaceous mowry shale of northwestern interior united states, AAPG bulletin, **57**(1):136–161.

Passey, Q.; Creaney, S.; Kulla, J.; Moretti, F. e Stroud, J. (1990) A practical model for organic richness from porosity and resistivity logs, AAPG bulletin, **74**(12):1777–1794.

Passey, Q. R.; Bohacs, K.; Esch, W. L.; Klimentidis, R.; Sinha, S. et al. (2010) From oil-prone source rock to gas-producing shale reservoir-geologic and petrophysical characterization of unconventional shale gas reservoirs, In: *International oil and gas conference and exhibition in China*, Society of Petroleum Engineers.

Pereira, M. e Macedo, J. (1990) A bacia de santos: perspectivas de uma nova província petrolífera na plataforma continental sudeste brasileira, Boletim Geociências da petrobras, **4**(1):3–11.

Pereira, M. J.; Barbosa, C.; Agra, J.; Gomes, J.; Aranha, L.; Saito, M.; Ramos, M.; Carvalho, M. d.; Stamato, M. e Bagni, O. (1986) Estratigrafia da bacia de santos: análise das seqüências, sistemas deposicionais e revisão litoestratigráfica, In: *Congresso Brasileiro de Geologia*, vol. 34, pp. 65–79.

Peters, K. E. e Cassa, M. R. (1994) Applied source rock geochemistry: Chapter 5: Part ii. essential elements.

Radtke, R.; Lorente, M.; Adolph, B.; Berheide, M.; Fricke, S.; Grau, J.; Herron, S.; Horkowitz, J.; Jorion, B.; Madio, D. et al. (2012) A new capture and inelastic spectroscopy tool takes geochemical logging to the next level, In: *SPWLA 53rd Annual Logging Symposium*, Society of Petrophysicists and Well-Log Analysts.

Raschka, S. e Mirjalili, V. (2017) Python machine learning, Packt Publishing Ltd.

Renfang, P.; Yuan, W. e Zheng, S. (2009) Geochemical parameters for shale gas exploration and basic methods for well logging analysis, China Petroleum Exploration, **14**(3):6–9.

Ribeiro, V. B.; Mantovani, M. S. e Louro, V. H. A. (2014) Aerogamaespectrometria e suas aplicações no mapeamento geológico, Terræ Didatica, **10**(1):29–51.

Sauer, I. L. (2016) O pré-sal e a geopolítica e hegemonia do petróleo face às mudanças climáticas e à transição energética, Recursos Minerais do Brasil.

Schmoker, J. W. (1979) Determination of organic content of appalachian devonian shales from formation-density logs: Geologic notes, AAPG Bulletin, **63**(9):1504–1509.

Schmoker, J. W. (1981) Determination of organic-matter content of appalachian devonian shales from gamma-ray logs, AAPG Bulletin, **65**(7):1285–1298.

Schmoker, J. W. e Hester, T. C. (1983) Organic carbon in bakken formation, united states portion of williston basin, AAPG bulletin, **67**(12):2165–2174.

Schmoker, J. W. e Hester, T. C. (1989) Formation resistivity as an indicator of the onset of oil generation in the woodford shale, anadarko basin, oklahoma.

Serra, O. (1986) Fundamentals of well-log interpretation (vol. 2): the interpretation of logging data (developments in petroleum science), Elsevier, **15**:679.

Smola, A. J. e Sclkopf, B. (2004) A tutorial on support vector regression.

Songnian, X. X. H. L. (1998) A quantitative relationship between well logging information and organic carbon content [j], JOURNAL OF JIANGHAN PETROLEUM INSTITUTE, **3**.

Swanson, V. E. (1960) Oil yield and uranium content of black shales, Rel. Téc., Geological Survey, Washington, DC (USA).

Tissot, B. e Welte, D. (1984) Petroleum formation and occurrence.

Vapnik, V. (1963) Pattern recognition using generalized portrait method, Automation and remote control, **24**:774–780.

Vapnik, V. (1982) Estimation of dependences based on empirical data berlin.

Vapnik, V. e Chervonenkis, A. (1964) A note on class of perceptron, Automation and Remote Control, **24**.

Vapnik, V. e Chervonenkis, A. (1974) Theory of pattern recognition.

Vivier, M. C. (1987) Foraminiferos planctónicos no cretáceo médio, Revista Brasileira de Geociências, **17**(2):154–161.

Von Luxburg, U. e Schölkopf, B. (2011) Statistical learning theory: Models, concepts, and results, In: *Handbook of the History of Logic*, vol. 10, pp. 651–706, Elsevier.

Yu, L. e Liu, H. (2003) Feature selection for high-dimensional data: A fast correlation-based filter solution, In: *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863.

Yun, H.; Xiang, J. e Liu, Z. (2000) Estimation method of organic carbon log and its application in shengli oilfield, Well Logging Technology, **24**:372–376.