Ibsen Pereira da Silva Gomes

O método K-Means na classificação de litofácies a partir de perfis geofísicos sintéticos inspirados no campo de Namorado, Bacia de Campos.

Niterói, RJ - Brasil

2 de junho de 2021

Ibsen Pereira da Silva Gomes

O método K-Means na classificação de litofácies a partir de perfis geofísicos sintéticos inspirados no campo de Namorado, Bacia de Campos.

Projeto Final de Graduação em Geofísica apresentado à Universidade Federal Fluminense como exigência parcial para obtenção do título de Bacharel em Geofísica

Universidade Federal Flumiense

Orientador: Victor Ribeiro Carreira

Coorientador: Rodrigo Bijani

Niterói, RJ - Brasil

2 de junho de 2021

Ibsen Pereira da Silva Gomes

O método K-Means na classificação de litofácies a partir de perfis geofísicos sintéticos inspirados no campo de Namorado, Bacia de Campos.

Projeto Final de Graduação em Geofísica apresentado à Universidade Federal Fluminense como exigência parcial para obtenção do título de Bacharel em Geofísica

Trabalho avaliado pela seguinte comissão:

Victor Ribeiro Carreira
Orientador - DIPPG/ON

Rodrigo Bijani
Co-orientador - GGO/UFF

Fernando Vizeu
GIECAR/UFF

Mario Martins Ramos
GIECAR/UFF

Niterói, RJ - Brasil
2 de junho de 2021

Agradecimentos

Agradeço a minha querida e amada família, por apoiar de várias maneiras, o meu sonho de frequentar a faculdade em uma cidade distante. Em espacial à minha mãe Corina, pelo apoio e confiança incondicionais durante todos esses anos. Minha força de vontade foi testada em vários momentos nesses anos na UFF e ela me incentivou incontáveis vezes, sem ela, não sei o que seria de mim e minha jornada universitária. Agradeço à minha tia Eliana pelos depósitos emergenciais no banco me salvando de perrengues financeiros. À minha falecida tia Carmem Heloísa, pelas aulas particulares de matemática em vários domingos.

Agradeço aos professores do pré-vestibular social de Resende, pelo clima de descontração durante as excelentes aulas. Sem eles, meu sonho de entrar na UFF seria um caminho muito mais árduo. Agradeço também aos professores e amigos que tive o prazer de conhecer, durante toda a minha jornada dentro da UFF.

Um agradecimento especial ao Mário Martins pelo grande suporte e conselhos durante todos esses meses. Nos bastidores desse projeto, ele foi de grande ajuda durante esses dias difíceis. Muito obrigado pelo apoio Sr. Mário.

Por fim, mas não menos importante, aos meu orientadores Vitor Carreira e Rodrigo Bijani. Agradeço a eles pelo voto de confiança no projeto, por tudo que aprendi com eles e também pela paciência deles em momentos onde não estava nos meus melhores dias. Não posso deixar de mencionar o clima amistoso das reuniões semanais, que se tornaram bons momentos de escape nesses dias de Pandemia. Muito obrigado pela excelente orientação durante todos esses meses.

Resumo

A inteligência artificial tem sido, nas últimas três décadas, uma poderosa aliada tecnológica das diversas áreas da sociedade. Com o propósito de transformar o computador em um automatizador de tarefas, muitos serviços de grande relevância contam com a celeridade dos modernos e sofisticados computadores. Na geofísica, a inteligência artificial tem contribuído diretamente para o avanço na interpretação de uma grande demanda de dados sísmicos. Na perfilagem geofísica, ferramentas de aprendizado de máquina se destacam como excelente alternativas para suprir as ausências ou corrompimento dos dados durante a aquisição. Adicionalmente, são métodos cada vez mais robustos na classificação de eletrofácies, imprescindíveis para a industria de óleo e gás. No entanto, a definição de banco de dados robustos e rotulados nem sempre são consistentes para que a classificação litológica seja mais eficaz. A metodologia desenvolvida nesse trabalho consiste na reconstrução de Perfis Geofísicos sintéticos utilizando o *K-Means*, um método de aprendizado de máquina estocástico não-supervisionado. Para execução deste trabalho, primeiramente foi realizado uma modelagem de perfis geofísicos, inspirados em uma seção geológica interpretada referente ao Campo de Namorado, Bacia de Campos. Em seguida, utilizou-se o *K-Means* para reconstrução de um perfil sintético abordando três tipos de inicialização. Sendo, duas aleatórias e uma baseada em uma iniciação de centroides determinística. As inicializações aleatórias desempenharam bons resultados na delimitação de topo e base de várias litologias. Porém, não recuperaram todas as litologias do poço simulado. A inicialização determinística possibilitou a inserção de informação a priori no método *K-Means*, o que aprimorou a classificação litológica obtida. A ausência de litologias nas outras inicializações foi mitigada pela introdução de alguns centroides deterministicamente no processo de inicialização do método. Com isto, percebe-se a robustez de aprendizado de máquina aplicados em perfis geofísicos, mesmo para métodos simples como o *K-Means*.

Palavras chave: Aprendizado de máquina, K-Means, Perfilagem Geofísica, Perfis sintéticos, Campo de Namorado, Bacia de Campos.

Abstract

Recently, artificial intelligence (AI) has been a powerful technological element of different society fields. To transform the computer into a decision maker, many sophisticated methods and algorithms are considered. In geophysics, AI has contributed to the development of interpretation of a great demand for seismic data. In well logging, machine learning tools also stand out as an excellent alternative to supply absences or data corruption. In addition, they are increasingly robust methods in the classification of electrofacies, which is essential for the oil and gas industry. In this work we use the K-means clustering method for classification of electrofacies by synthetic well-log data. The geophysical logs are computed from an interpreted geological section comprising the Namorado Oilfield, in Campos Basin. To mitigate the K-means clustering method, we run three initialization of clusters (e.g., random, k-means++ and deterministic) to bespeak the produced classifications. Both random and k-means++ inputs performed well in recovering most of simulated lithologies. Despite, none of them are able to locate all lithologies, due to severe data overlapping analyzed in crossplots. To overcome such limiting aspect, we manually initialize four clusters, which improved the classification results. This is a valuable setup offered by K-means method implemented in scikit-learn python library.

Keywords— Artificial Intelligence, K-means, Synthetic well-log data, classification of electrofacies

Sumário

Lista de ilustrações

1 Introdução

Nos últimos anos, a inteligência artificial apresenta-se como uma poderosa ferramenta computacional, cuja influencia nos diversos ramos da sociedade é notória. Pode-se mencionar as investigações policiais através do sistema de reconhecimento facial, na robótica, medicina, computação e até mesmo nos aplicativos das redes sociais (BARSTUGAN; OZKAYA; OZTURK, 2020; PETERS, 2008; WU et al., 2019; BIAMONTE et al., 2017). A filosofia da inteligência artificial consiste em transformar o computador em um tomador de decisões, especialmente na automatização de tarefas repetitivas (RUVINI; DONY, 2000). Dentro do universo da inteligência artificial, dois grandes grupos podem ser destacados: o aprendizado de máquina (*machine learning*, em inglês) e aprendizado profundo (*deep leraning*, em inglês). O primeiro permite que o sistema aprenda os padrões e modelos estruturais úteis a partir de dados de treinamento (XIE et al., 2018). Já o segundo, reconhece padrões nos dados a partir de redes neurais com multi-camadas (LAUZON, 2012).

No universo de aprendizado de máquinas há duas vertentes importantes: o supervisionado e o não-supervisionado, relativos à forma de aplicação do método e do algoritmo utilizados. No primeiro, diversas informações rotuladas, chamadas de dados de treinamento, devem ser disponibilizados ao programa de computador. Este realiza o processamento dessas informações baseado em algum algoritmo competente, na etapa chamada de treinamento. Em seguida, informações não rotuladas chamadas de dados de classificação, são considerados pelo algoritmo devidamente treinado. Por fim, o programa de computador rotula os dados de classificação (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007; SCHRIDER; KERN, 2018). Porém, nem sempre é possível adquirir e rotular um conjunto muito elevado de dados, devido ao alto custo financeiro e até mesmo de esforço para rotulá-los. Dessa forma, o aprendizado de máquina não-supervisionado surge como uma poderosa alternativa. Neste caso, a etapa de treinamento não é necessária, sendo o algoritmo capaz de encontrar padrões diretamente a partir dos dados de classificação (DAYAN; SAHANI; DEBACK, 1999). O principal interesse dos algoritmos não-supervisionados está na busca por características específicas nos dados de classificação. Esses padrões de similaridade são chamados de grupos, que podem ser representados por centroides (ROBERTS, 1997). Com isso, o algoritmo então torna-se hábil em separar os dados em grupos afins, o que permite uma classificação dos dados (ZHUANG et al., 1998).

Existe uma gama significativa de métodos e algorítimos de aprendizado de máquina desenvolvidos e em desenvolvimento. Dentre eles, podemos mencionar as Redes Neurais Artificiais desenvolvidas a partir do modelo de McCulloch e Pitts (1943) e os Mapas auto-organizáveis proposto por Kohonen (1982). O primeiro é totalmente inspirado nos processos de funcionamento do cérebro humano, tais como interpretação perceptual,

abstração e aprendizagem (CALDERÓN-MACÍAS; SEN; STOFFA, 2000). Já o segundo é basedo no mapa estrutural do córtex humano, exemplificado pelo homúnculo de Penfield (CRICK; KOCH, 2000). Neste caso, as características mais importantes do corpo humano, como a língua, as mãos e os orgãos genitais ocupam maior espaço no cortex cerebral. Há ainda alguns algoritmos que, embora tenha aplicação supervisionada em alguns casos, são mais comumente aplicados como não-supervisionados na literatura. Dentre os quais pode-se destacar o *K-means* (LLOYD, 1982), o *Support Vector Machine* (SVM) (CORTES; VAPNIK, 1995), o *Linear Discriminant Analysis* (FISHER, 1936) e o *Gaussian Mixture* (REYNOLDS, 2009). Existem até mesmo métodos que mesclam a supervisão e não-supervisão, podendo ser chamados de métodos híbridos (FRITZKE, 1994; DOUGHERTY; KOHAVI; SAHAMI, 1995).

Na geociências, o aprendizado de máquina é bastante prolífico, especialmente com o avanço da tecnologia de aquisição e armazenamento dos dados, que permite um aumento significativo no número de dados geofísicos disponíveis para pesquisa. Diversos trabalhos tem contribuído direta e indiretamente para o avanço desta linha de pesquisa, especialmente nas áreas de sísmica (GUILLEN et al., 2015; MATOS; OSORIO; JOHANN, 2007; JIA; MA, 2017) e de perfilagem geofísica (KUMAR; KISHORE, 2006; BESTAGINI; LIPARI; TUBARO, 2017; ZHANG; YUNTIAN; JIN, 2018). Por exemplo, Kuroda et al. (2012) utilizam os mapas auto-organizáveis (*Self-Organizing Maps* (SOM), em inglês) na classificação de eletrofácies a partir de perfis de poços do campo de Namorado, Bacia de Campos. Para a etapa de treinamento, diversos poços com descrição litológica foram considerados. Seguindo a mesma abordagem, Neyamadpour, Taib e Abdullah (2009) investigam a aplicabilidade de uma Rede Neural Artificial (RNA), implementada em linguagem MATLAB, para inverter dados geoelétricos obtidos através da configuração Wenner - Schlumberger. Já Konaté et al. (2015) desenvolvem uma RNA para a previsão da porosidade na crosta cristalina utilizando dados de perfis geofísicos. Adicionalmente, este estudo mostra a eficiência de dois tipos de RNAs, que incluem a rede neural de retropropagação *feed-forward* (FFBP) e a de função de base radial (RBF) para resolver problemas de porosidade em situações reais. Du et al. (2015) utilizam o SOM para melhorar a interpretação de sismofácies. Adicionalmente, os autores utilizaram um método empírico de decomposição de sinal sísmica para facilitar o treinamento do SOM. Já Carreira, Neto e Bijani (2018) apresentam um estudo comparativo de métodos supervisionados, entre eles o SOM, para a classificação de eletrofácies em um ensaio sintético. Adicionalmente, várias metodologias foram utilizadas para solucionar problemas específicos. Kuyuk et al. (2012) utilizam o *K-Means* e o *Gaussian Mixture* para classificar eventos sísmicos nas imediações de Istanbul, na Turquia. Niknam e Amiri (2010) apresenta um método híbrido para aprimorar a análise de agrupamentos envolvendo algoritmos evolucionários e o *K-Means*. A avaliação de performance do método proposto é realizada em testes controlados. Dias et al. (2018) compara diferentes técnicas, dentre elas o *K-Means*, para automatizar o processo de segmentação de fraturas. Para isso, imagens de fraturas simuladas são utilizadas para verificar a qualidade da segmentação. Santos

(2016) aplica e analisa diversos métodos não-supervisionados no problema da classificação de eletrofácies utilizando perfis geofísicos de poços da bacia de Campos.

Este trabalho apresenta uma avaliação da performance do método *K-Means* mediante diferentes testes de inicialização quando o problema da classificação de eletrofácies é considerado. Iniciamos nossa investigação através da modelagem dos perfis de raios-gama (GR), densidade (RHOB), resistividade (ILD) e sônico (DT) a partir de uma seção geológica interpretada por BARBOZA (2005) referente ao campo de Namorado, Bacia de Campos. Um banco de dados é estabelecido a partir de pesquisa bibliográfica, com as médias das propriedades físicas de cada uma das litologias que compõem o Campo de Namorado. As eletrofácies simuladas são associadas a uma cor da paleta de cores da Petrobrás. Em seguida define-se o poço sintético por meio da interação entre o usuário e o programa de modelagem. Nesta etapa, os perfis são simulados por meio de sorteios com distribuição Gaussiana das propriedades ao longo da profundidade do poço. Diante dos dados sintéticos, utilizamos o método *K-Means*, disponível na biblioteca *Sckit-learn* (PEDREGOSA et al., 2011) com diferentes parâmetros de entrada para verificar a qualidade deste método na classificação de eletrofácies. Primeiramente inicializamos o *K-Means* de forma totalmente aleatória, sem qualquer inferência sobre a localização inicial dos centroides. A seguir acionamos o modo *K-means++* para inicialização dos centroides, que considera o critério da inércia para acelerar o processo de localização dos centroides promissores. Na sequência, foi realizada a inicialização determinística[1] de alguns centroides, a fim de induzir o método na busca por uma classificação mais assertiva. Por fim, analisamos cada resultado através de *logplots* e *crossplots*. Adicionalmente, computamos os erros percentuais da classificação com relação ao perfil litológico sintético para cada um dos testes mencionados.

[1] Determinístico significa que os centroides são inicializados a partir de informação a priori introduzida pelo usuário.

2 Perfilagem Geofísica

A Perfilagem é um método Geofísico amplamente estudado tanto na academia, como na indústria petrolífera, devido a sua importância na caracterização de reservatórios (LUCIA; KERANS; JENNINGS, 2003; SUBHAKAR; CHANDRASEKHAR, 2016). Um perfil é definido como uma representação gráfica de registro contínuo em função da profundidade, de observações feitas em rochas e fluidos na seção geológica exposta em um poço (BATES; JACKSON, 1980). A combinação de diversos perfis geofísicos pode auxiliar na interpretação geológica da área de estudo.

Neste trabalho, foram utilizados os perfis de densidade (RHOB), raios-gama (GR), resistividade (ILD) e sônico (DT). Eles oferecem grande robustez na investigação sobre a relação rocha-perfil. Adicionalmente, são perfis muito utilizados na caracterização das variações litológicas, eletrofácies e identificação de fluidos no poço (SERRA, 1983).

2.1 Densidade (RHOB)

O perfil de Densidade é amplamente utilizado na identificação litológica, quantificação de parâmetros como a porosidade, além da identificação de zonas de gás e outros minerais. Adicionalmente, é possível observar efeitos da compactação, da idade e da composição mineralógica das rochas (RIDER; KENNEDY, 2002). A figura 1 mostra um registro continuo de densidade em rochas na formação em subsuperfície. É interessante notar características como valores altos de densidade em rochas ígneas; além do aumento progressivo da densidade no folhelho, com o aumento da compactação devido a alteração de profundidade.

A técnica de perfilagem da densidade consiste em submeter as rochas a um bombardeio de raios gama de média ou alta energia (0.2 - 2.0 MeV) colimados, e medir sua atenuação entre a fonte da ferramenta e os detectores (RIDER; KENNEDY, 2002). A densidade é então estimada com a medição da radiação gama que retorna para o detector, sendo que a resposta quando o feixe se propaga pelo material depende de sua densidade eletrônica. O espalhamento sofrido pelo raio-gama ao interagir com os elétrons da formação, é chamado de Espalhamento Compton (RIDER; KENNEDY, 2002).

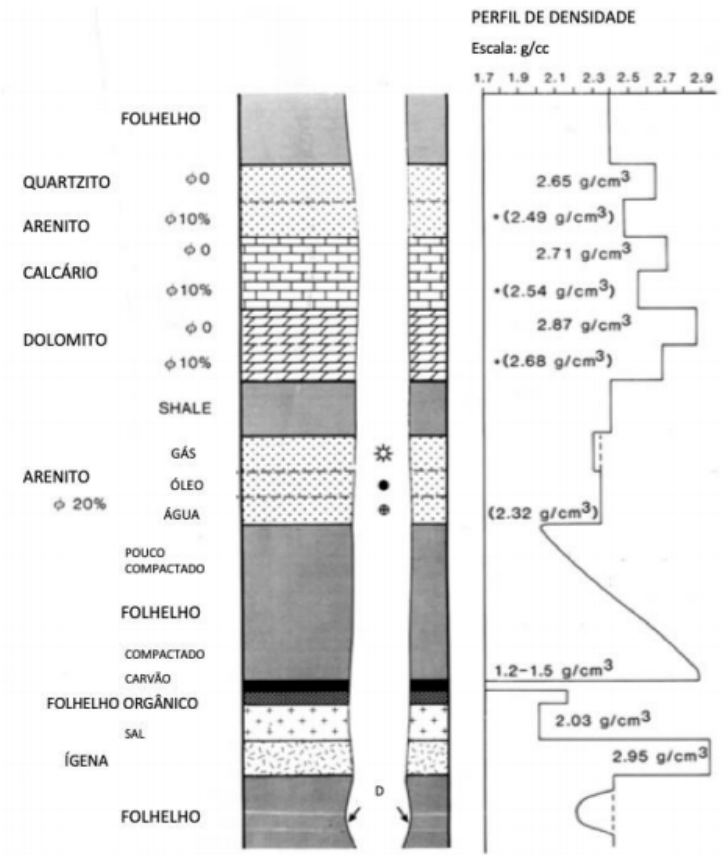


Figura 1 – Perfil litológico com os valores característicos de densidade, retirado de OLIVEIRA (2019)

Sendo assim, o perfil RHOB, é uma ferramenta que responte diretamente a densidade do material e inversamente a sua porosidade. Este aspecto está em acordo com o fato de que o Espalhamento Compton é diretamente proporcional à densidade eletrônica e esta é proporcional à densidade da formação (NERY, 2004).

2.2 Raios-Gama (GR)

O perfil de Raios-Gama manifesta a radioatividade natural e o comportamento de decaimento espontâneo dos elementos pesados e instáveis. Isso acontece devido a perda de energia com tempo por emissão de partículas alfa, beta, gama e calor (SERRA; SERRA, 2004). Geralmente, os elementos radioativos mais encontrados na natureza são o ^{239}U, ^{232}Th e ^{40}K. Estes que contribuem significativamente para a detecção de radioatividade nos Cintilômetros, ferramenta esta, usada na medição de radiação em perfis de poços.

Na natureza, o comportamento radioativo nas rochas ígneas e metamórficas apresentam maior intensidade em comparação com as rochas sedimentares. E dentro da classe de rochas sedimentares, as mais argilosas apresentam uma radioatividade superior (STEVANATO, 2011). O perfil de Raios-Gama também é caracterizado pela boa distinção de

argilas e folhelhos em relação a outras litologias, ja que estes são uns dos elementos mais radioativos encontrados naturalmente em perfilagens (ARAGÃO, 2017). A figura 2, mostra um perfil com variações de valores de raio-gama na formação com valores mais altos para rochas, com maior presença de matéria orgânica como no folhelho e valores baixos para o Calcário e Arenito.

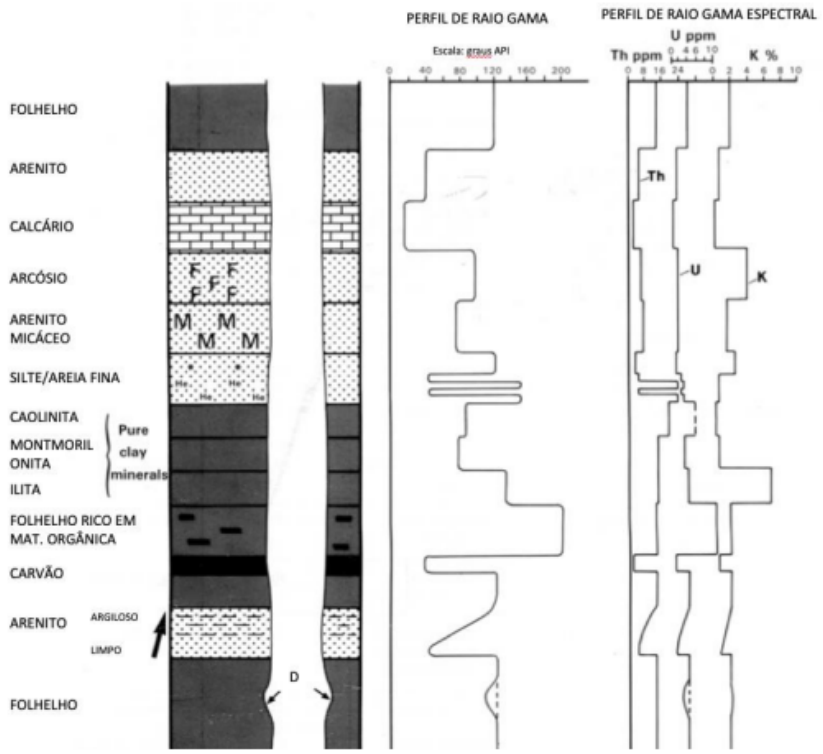


Figura 2 – Perfil litológico com os valores característicos de raios-gama, retirado de OLIVEIRA (2019)

2.3 Resistividade (ILD)

Este perfil mede a resistividade de formação das rochas em subsuperfície. Existem várias técnicas de medição e todas funcionam mediante um sistema básico comum: um emissor (eletrodo ou bobina) envia um sinal (corrente elétrica, campo eletromagnético) para a formação. Um receptor (eletrodo ou bobina) mede a resposta da formação a uma certa distância do emissor (SERRA, 1983).

O perfil de resistividade pode ser usado para identificação das camadas que possuem hidrocarbonetos, identificação das zonas saturadas com óleo a partir do cálculo de saturação de água (Sw), definição do contato óleo-água, correlação entre poços, identificação do tipo de fluido ou rocha presentes (ARAGÃO, 2017). A figura 3 mostra a resposta do perfil mediante a formação em subsuperfície, mostrando valores mais elevados para Arenitos porosos com saturação de óleo ou gás; além do baixo valor para o Folhelho.

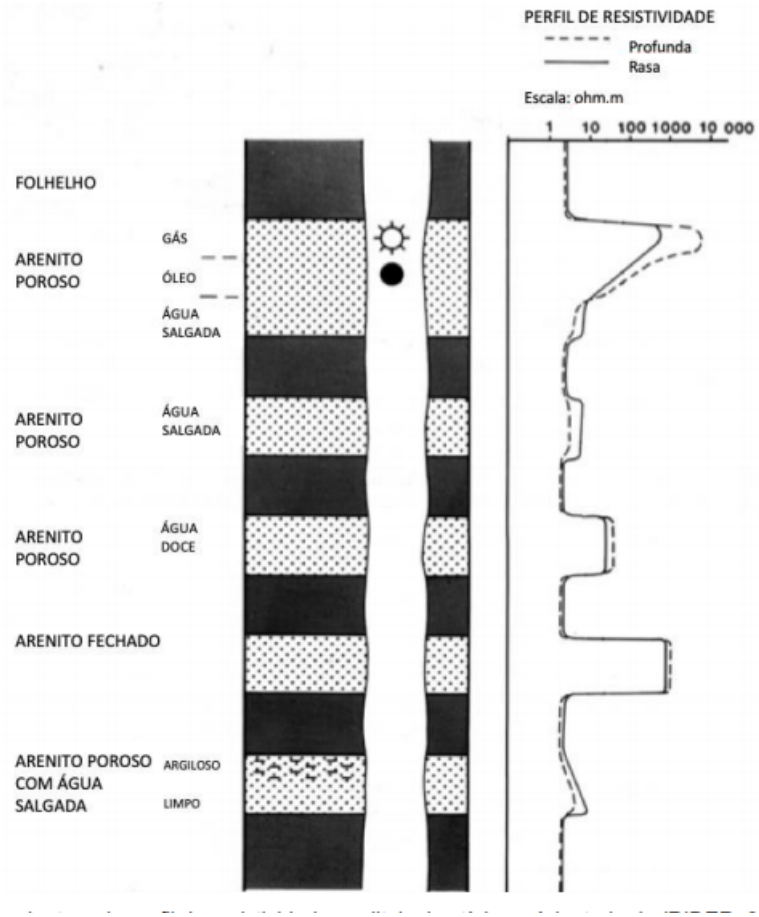


Figura 3 – Perfil litológico com os valores característicos de resistividade, retirado de OLIVEIRA (2019)

Os fluidos, quando penetram rochas porosas, alteram o comportamento de corrente elétrica na mesma. Os Hidrocarbonetos, por exemplo, elevam a intensidade dos valores de resistividade no perfil (RIDER; KENNEDY, 2002). A figura 4, mostra como uma corrente se comporta dentro de fluidos no interior de uma rocha matriz não condutora. O método ainda conta com uma gama de diversidade de arranjos, com a finalidade de coletar melhor os dados em diferentes zonas ao redor do poço, além da possibilidade de diferentes tipos de perfis de resistividade que podem ser medidos durante uma única passagem de sonda (KEAREY et al., 2009).

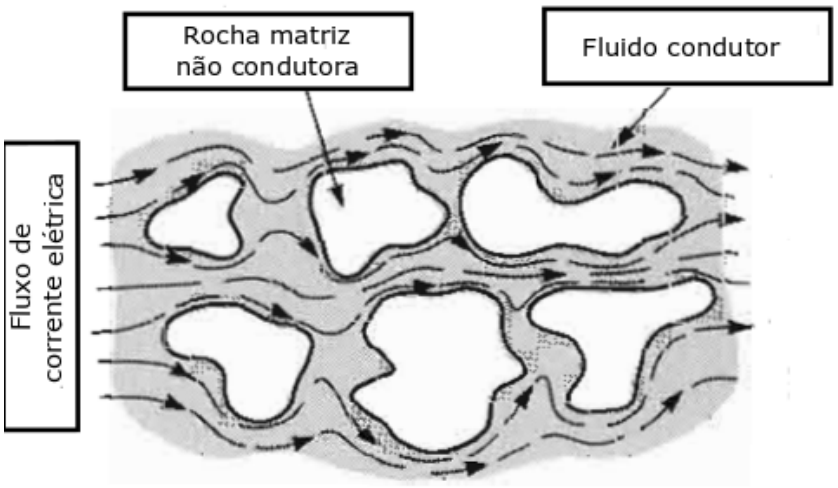


Figura 4 – Fluxo de corrente em fluidos no interior de rochas, adaptado de Rider e Kennedy (2002)

2.4 Sônico (DT)

O perfil sônico tem a finalidade de medir a velocidade de propagação de ondas ao longo de um poço, além da capacidade da formação de transmitir ondas acústicas. Geologicamente, esse perfil responde bem a variações de litologia e de porosidade das rochas (RIDER; KENNEDY, 2002). Na figura 5, pode-se notar valores altos para velocidade de onda sonora em Calcário e Arenito compactados e uma progressão positiva da velocidade no Folhelho, relativo ao aumento de compactação.

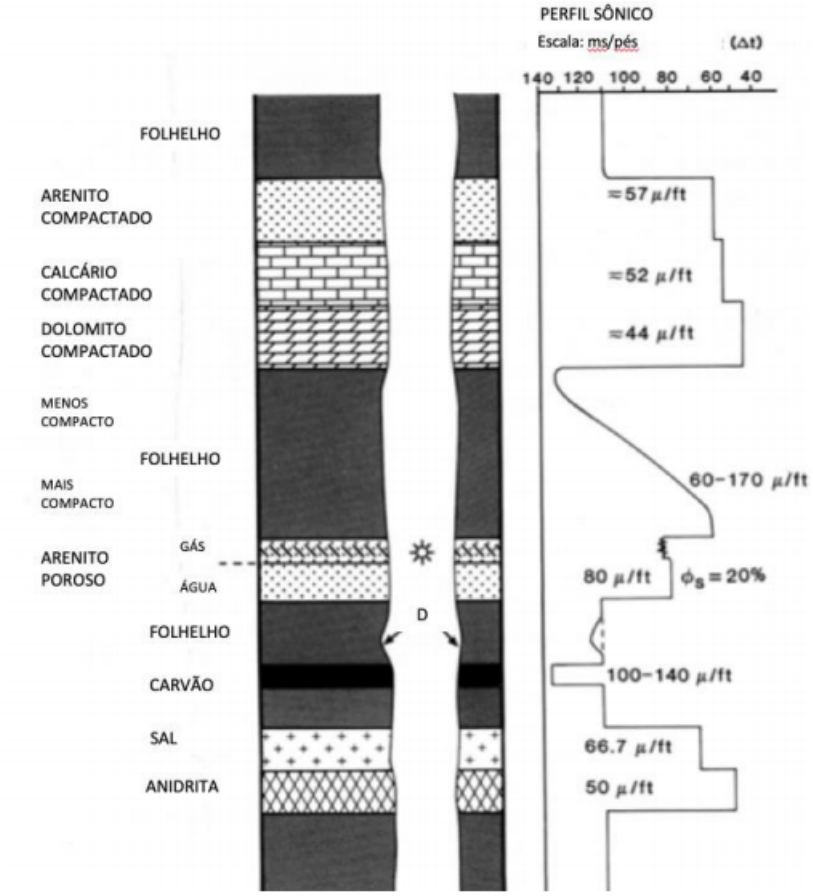


Figura 5 – Perfil litológico com os valores característicos de velocidade, retirado de OLIVEIRA (2019)

As ferramentas de medição deste perfil consistem em um transdutor magnético que quando excitado, emite uma onda acústica cuja frequência média é da ordem de 20 a 40 kHz. A duração da emissão é curta, mas é repetida várias vezes por segundo (10 a 60 vezes dependendo da ferramenta) (SERRA, 1983). Numa aquisição, esse perfil pode medir a grandeza tempo de trânsito com unidade de (μs/ft).

3 O método K-Means

O *K-Means* é um método de agrupamento de dados muito popular desenvolvido por Hartigan e Wong (1979), e utilizado em diversos estudos e áreas do conhecimento (CHANG; ZHANG; ZHENG, 2009; SIDDIQUI et al., 2020). As bases conceituais deste método visam o particionamento dos dados em k-grupos, representados por pontos chamados de centroides. Estes, que localizam-se na média das amostras do k-ésimo grupo, são totalmente análogos ao centro de massa do grupo (CAPÓ; PÉREZ; LOZANO, 2017; ARTHUR; VASSILVITSKII, 2006). Na figura 6, pode-se observar a dinâmica do *K-Means*. Na figura 6-A, tem-se uma distribuição inicial das amostras, subsequentemente, na figura 6-B, as amostras são separadas em dois grupos com médias representadas pelos centroides C1 e C2. A figura 6-C mostra o processo para agregar o ponto P em vermelho a um dos dois grupos, utilizando a distância Euclidiana. Como D1<D2, o ponto é atribuído ao Grupo em azul e a nova posição do centroide C1 é atualizada como mostrado na figura 6-D. Então, o processo de repete em um *loop* até alcançar um critério de convergência estabelecido.

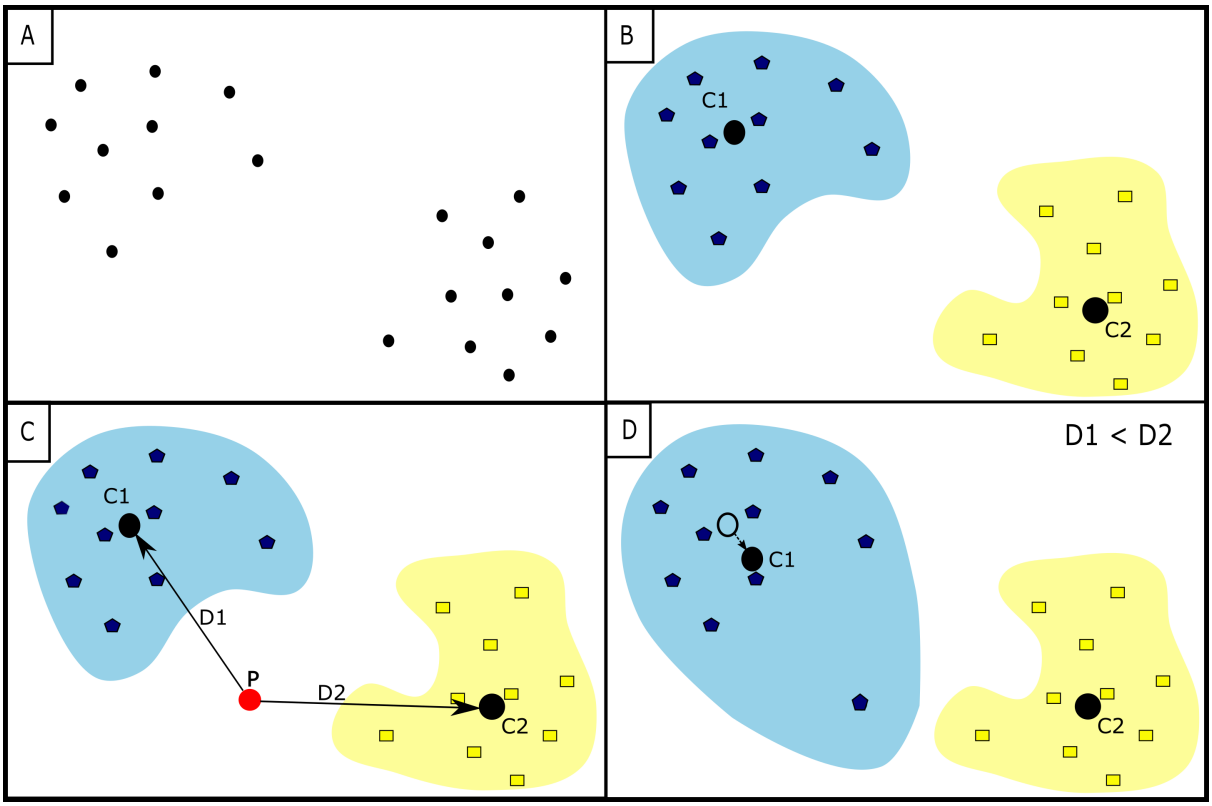


Figura 6 – A - Dados de entrada a serem agrupados; B - Dados agrupados com seus devidos centroides C1 e C2; C - Cálculo de distância Euclidiana no ponto P; D - O ponto agora pertence ao agrupamento em azul, e a posição do centroide C1 é atualizada para o novo centro de massa

A movimentação dos centroides em cada iteração é controlada por alguma métrica, como a Manhattan ou Euclidiana, por exemplo (SANTOS, 2016). Como os centroides buscam uma posição ideal na média de um agrupamento, é necessária a aplicação de uma minimização, ou seja, um critério para se procurar o mínimo de uma função de custo. Para o método aqui apresentado, a métrica usada é a Euclidiana e o critério de minimização utilizado é o da inércia, ou a regra que minimiza a soma dos quadrados das distâncias dos pontos de um agrupamento ao centroide que pertence (FAHIM et al., 2006; SCULLEY, 2010; SANTOS, 2016):

$$\sum_{k=1}^{m} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \tag{3.1}$$

em que, m é o número de grupos em que o dado será dividido, C_k é o conjunto de amostras que pertencem ao k-ésimo grupo e μ_k é o centroide que representa a média do k-ésimo grupo. Esse processo é realizado para todos os agrupamentos e seus devidos centroides.

A figura 7 mostra as principais etapas que uma implementação do método *K-Means* geralmente apresenta. No algoritmo deve-se inserir primeiramente, o número de grupos que o usuário quer dividir seus dados. Posteriormente, fornecer as coordenadas dos centroides e consequente agrupamento das amostras mediante Norma Euclidiana. Após o critério de minimização pode ou não haver uma alteração nos agrupamentos, se houver retorna-se para etapa de cálculo dos centroides, se não, o algoritmo converge.

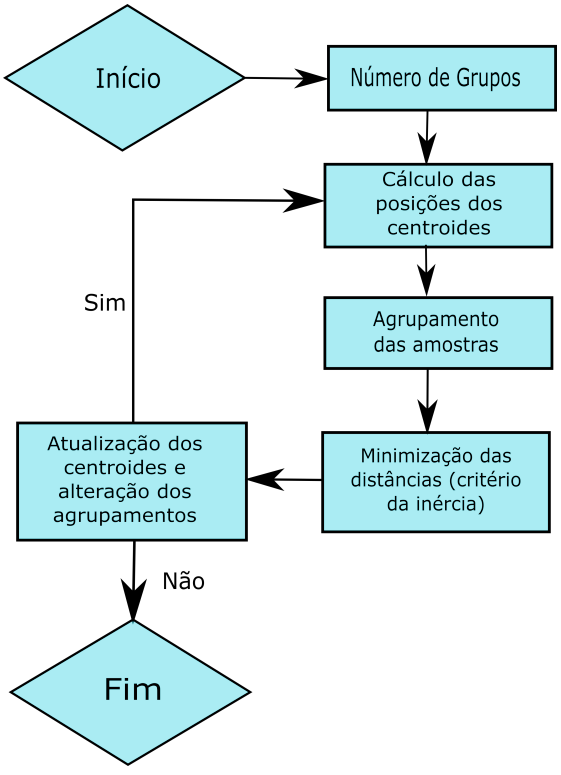


Figura 7 – Fluxograma para o algoritmo do K-Means

O *K-Means*, além de ser um algoritmo popular e amplamente usado, é conhecido também pela rápida convergência. No entanto, é possível que a sua celeridade leve a um minimo local indesejável (PEDREGOSA et al., 2011). Isso significa, agrupar dados não correlacionados, e isso pode causar problemas em interpretações posteriores. Alguns estudos (ARORA; VARSHNEY et al., 2016; WHANG; DHILLON; GLEICH, 2015) ainda mostram que o método também é considerado "rígido", ou seja, pode não se comportar bem diante de uma distribuição de dados com forte sobreposição. Além disso, ele depende muito da posição inicial dos centroides, ou seja, diferentes inicializações acarretam em diferentes agrupamentos com o mesmo conjunto amostral. Uma forma de contornar este problema é realização de várias inicializações (SANTOS, 2016). O algoritmo disponibilizado pela biblioteca do *Sckit-learn* (PEDREGOSA et al., 2011) disponibiliza a realização várias inicializações internas, com diferentes sementes (coordenadas iniciais) para os centroides, o que contorna a dependência mencionada anteriormente. A figura 8 mostra um esquema com os principais parâmetros de entradas e saída do *K-Means*, cujo o entendimento é fundamental para o melhor uso do método.

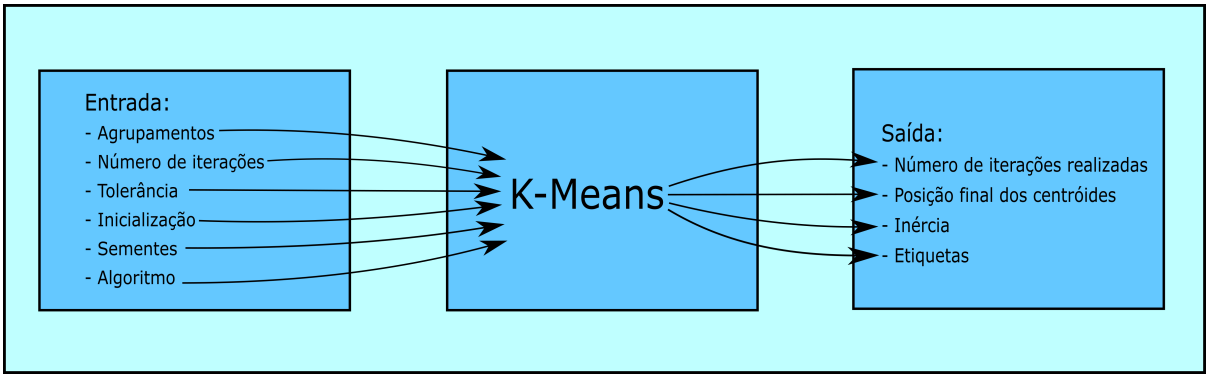


Figura 8 – Informações de entrada e saída do *K-means* do *scikit-learn*

3.1 K-Means: Parâmetros de entrada

Nessa sessão será apresentado os parâmetros de entrada do *K-Means* na biblioteca do *scikit-learn*.

3.1.1 Agrupamentos (*n_clusters*)

Este parâmetro define quantos agrupamentos os dados devem ser divididos. O números de grupos está diretamente ligado ao número de centroides, pois cada um desses centros define um agrupamento formado pelos dados mais próximos a ele.

3.1.2 Número máximo de iterações (*max_iter*)

O *k-Means* do *Sckit-Learn* tem a opção de estabelecer um limite máximo de iterações que podem ser utilizadas. Vale salientar que a convergência adequada pode ser alcançada antes do número máximo de iterações ser atingido, fazendo com que o K-means finalize o processo antes que *max_iter* seja contemplado.

3.1.3 Tolerância (*tol*)

A convergência é alcançada quando há apenas um deslocamento mínimo dos centroides em relação à posição anterior durante as iterações. Por isso, é necessário um critério de parada quando essa distância mínima for alcançada e a tolerância tem esse papel. Neste algoritmo, é calculada a distância entre a posição atual C_k e anterior C_{k-1} dos centroides. Se o deslocamento D_k, destacado na equação 3.2, for menor que o valor da tolerância o processo é encerrado.

$$D_k = \sqrt{\|C_k - C_{k-1}\|^2} \tag{3.2}$$

A escala dos dados pode ser um fator importante na escolha de uma tolerância. Neste trabalho, foram utilizados dados em diferentes escalas. Para controlar essa variabilidade, foi realizado um uma normalização chamada *min-max*, que será apresentado com mais detalhes no capítulo 4.

3.1.4 Inicialização (*init*)

Esse parâmetro controla as coordenadas iniciais dos centroides no método. O *k-Means* do *Sckit-Learn* apresenta três formas de inicializações sendo duas totalmente randômicas, utilizando pontos do conjunto de amostras, e outra utilizando um vetor com dimensão (n^o grupos, n^o propriedades), informando as posições dos centroides.

O *k-means++* é uma forma de inicializar o *K-Means* desenvolvido por Arthur e Vassilvitskii (2006). Ele se diferencia por ser uma inicialização que pondera os pontos de dados, de acordo com sua distância ao quadrado do centro mais próximo já escolhido, para aumentar a velocidade de convergência (DAVID; SERGEI, 1996).O aumento da velocidade de convergência pode ser medido através do cálculo da complexidade "O". (PAKHIRA, 2010; ARTHUR; VASSILVITSKII, 2006).

Outra possibilidade é realizar a inicialização personalizada, onde alguns ou mesmo todos os centroides são colocados de forma determinística no algoritmo. Dessa forma, podemos dizer que essa mecanismo de inicialização impõe informações a priori ao método. Neste trabalho, mostramos a importância desta inicialização no capítulo de Resultados.

3.1.5 Número de inicializações (*n_init*)

O número de inicializações indica quantas inicializações o *K-Means* deve processar. As sementes escolhidas serão o melhor resultado dentre execuções consecutivas em termos de inércia (PEDREGOSA et al., 2011), e a partir dessa escolha, realizar as futuras iterações. E é importante salientar que esse processo só tem sentido utilizando uma inicialização aleatória com *init* sendo "random"ou "k-means++".

3.1.6 Algoritmo (*algorithm*)

O método K-means do Scikit-Learn é baseado em dois algoritmos que podem ser utilizados pelo usuário. São eles o algoritmo "Full"e o "Elkan"(PEDREGOSA et al., 2011). O algoritmo "Full"é baseado em um método estocástico chamado de *Expectation Maximization* desenvolvido por Dempster, Laird e Rubin (1977). Trata-se de um algoritmo iterativo que se alterna entre dois passos E-passo e M-passo, após a determinação de um estado inicial (SANTOS, 2016). Os trabalhos Moon (1996), Dellaert (2002) mostram com mais detalhes os passos citados, além da dedução das equações usadas neste algoritmo.

O algoritmo *Elkan* utiliza desigualdade triangular para otimizar alguns cálculos de distâncias redundantes relativos a uma amostra e seu centroide, como proposto por Elkan (2003). As bases dessa operação são estabelecidas segundo dois lemas matemáticos:

Sendo P um ponto e C1 e C2 os centroides, representados na figura 9, então:

- Se D(C1, C2) $\leq$ 2D(P, C1) então: D(P,C2) $\leq$ D(P, C1) e;
- D(P, C2) $\geq$ max(0,D(P,C1) - D(C1, C2))

Se um ponto P da amostra for mais próximo de um centroide C1 em relação a outro centroide C2, ou seja $D2(P, C2) > D1(P, C1)$, a distância D2 não é calculada (ELKAN, 2003). Isto acarreta numa diminuição do custo computacional do algoritmo.

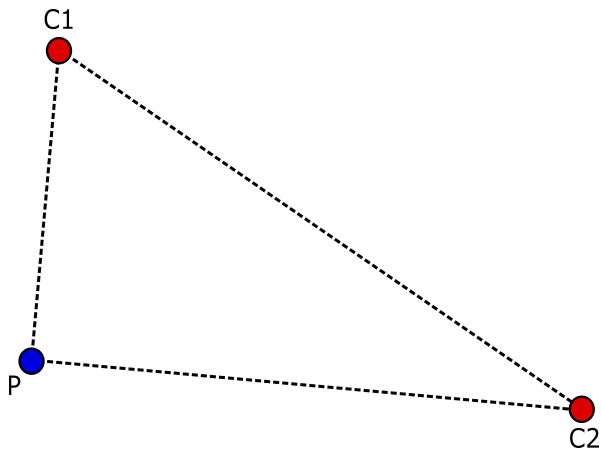


Figura 9 – Exemplo pictórico da Desigualdade triangular usado por Elkan

3.2 K-Means: Parâmetros de saída

Apresentamos aqui cada um dos parâmetros de saída do *K-Means* na biblioteca do *scikit-learn*, após a convergência.

3.2.1 Posição final dos centroides (*cluster_centers_*)

Este parâmetro fornece coordenadas, no espaço dos atributos das posições finais dos centroides, após a convergência.

3.2.2 Número de iterações realizadas (*n_iter_*)

Como ja mencionado, o *scikit-learn* tem como uma das informações de entrada o número máximo de iterações realizadas. Porém, com a saída pode-se saber quantas iterações foram efetivamente realizadas. Aqui entra o senso crítico do usuário para analisar o número de iterações realizadas. Por exemplo, o *K-Means* é um método que converge rápido, se o número de iterações final for muito alto, pode ser indício de que o método se perdeu e pode ser necessário um ajuste em algum parâmetro de entrada.

3.2.3 Inércia (*inertia_*)

Neste parâmetro, temos a informação sobre a soma do quadrado das distâncias das amostras até o centro do agrupamento mais próximo, após a parada do algoritmo. A figura 10 mostra as distâncias entre as amostras (xi) e os centroides (C_k), além dos agrupamentos representados pelas diferentes cores. A inércia é um indicativo da convergência do *K-Means*, já que, um valor relativamente baixo desta grandeza indica que os centroides estão na posição ideal (mínimo global) do agrupamento. Outros fatores como quantidade de amostras podem interferir nessa grandeza, isso porque, quanto mais amostras mais distâncias são calculadas e agregadas à inercia.

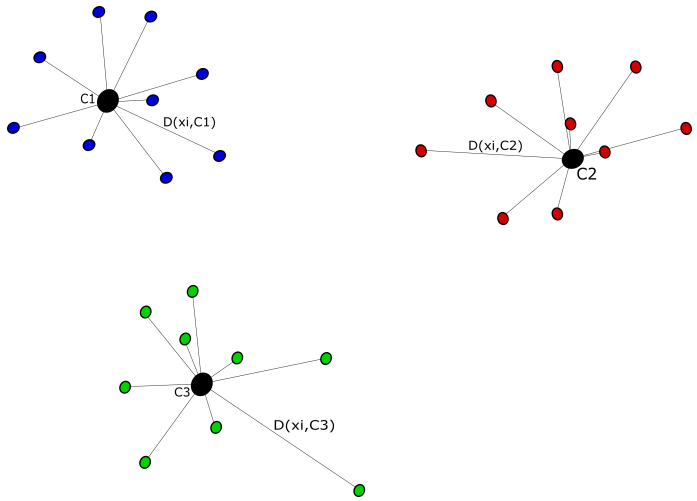


Figura 10 – Soma de todas as distâncias D(xi, C_k) das amostras em um agrupamento, ao centroide C_k que pertence

3.2.4 Etiquetas (n_samples)

As etiquetas são fundamentalmente o resultado final em termos de numéricos, sendo representado por números inteiros. Eles são atribuídos a cada amostra que representam cada um dos k-centroides, ou seja, uma associação de cada dado ao centroide do grupo ao qual pertence. Os valores inteiros vão de 0 até o número de agrupamentos - 1 (0,...,k-1), lembrando que o *scikit-learn* utiliza a linguagem Python que tem como padrão iniciar contagens a partir do índice 0.

4 Metodologia

Este trabalho conta com dois *scripts*, implementados em linguagem *python* por meio do ambiente *Jupyter Notebook* (PÉREZ; GRANGER, 2007). O primeiro consiste da etapa de modelagem dos perfis geofísicos a serem considerados. Já o segundo é utilizado para a classificação de eletrofácies em um poço sintético. Neste capítulo ambos os *scripts* são explicados e detalhados. A figura 11 mostra o fluxograma do trabalho realizado. As próximas seções são dedicadas à explicação mais aprofundada das etapas realizadas neste trabalho.

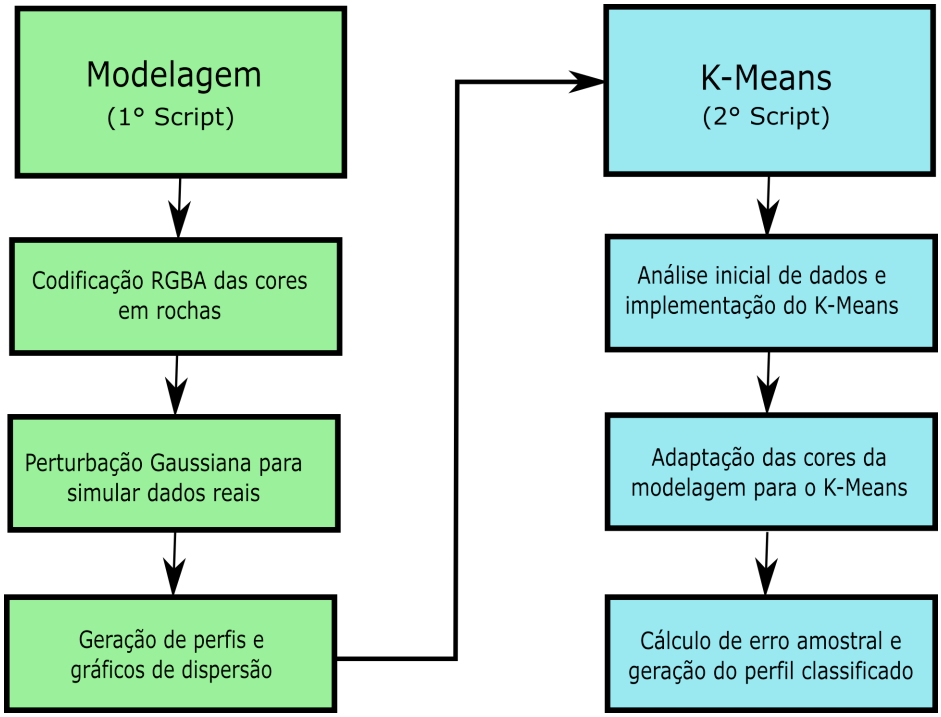


Figura 11 – Fluxo de trabalho realizado nos dois *scripts* desenvolvidos

4.1 Modelagem de perfis

A modelagem se inicia na utilização de um *software open-source* de imagem vetorizada (*Inkscape*, especificamente) para desenhar uma seção geológica previamente interpretada. Neste caso, foi escolhida uma seção do Campo de Namorado, Bacia de Campos (BARBOZA, 2005), conforme pode ser observado na figura 12-(a). Esta que apresenta três sequencias deposicionais separadas pelos limites erosivos (discordâncias), em vermelho (SOUZA, 1997). Segundo Jr, Vail e III (1977), discordâncias são definidas como uma superfície de erosão ou não deposição que separa estratos mais jovens de estratos mais antigos. Estas sequências são causadas devido a variações do nível do mar. Quando há

uma discordância, significa que o mar regrediu e deixou a plataforma exposta e suscetível a erosão. Quando o nível do mar volta a subir, ela transporta e deposita material fino como folhelhos acima da discordância na seção geológica. Segundo BARBOZA (2005), Souza (1997), os arenitos depositados são turbiditos ligados a corrente de alta densidade. Para diferenciar litologias que aparecem nessas sequências, foram atribuídas diferentes valores de propriedades físicas, a fim de diferenciá-las. O folhelho por exemplo, recebeu diferentes tonalidades de verde com diferentes conjuntos de valores de densidade (RHOB), raios gama (GR), resistividade (ILD) e sônico (DT). O mesmo foi feito para os Arenitos e Margas. A primeira coluna da tabela 1 mostra a nomenclatura adotada neste trabalho. As litologias depositadas em momentos diferentes, são representados pelas diferentes tonalidades na figura 12-(b).

Posteriormente, foi criado de um banco de dados de algumas propriedades físicas, baseadas na literatura já estabelecida para a bacia de Campos (FREITAS, 2008; RIDER; KENNEDY, 2002; STEVANATO, 2011). Para isso, foi realizada uma investigação em Winter, Jahnert e França (2007), sobre quais litologias podem ser encontradas na Bacia de Campos, englobando a região do campo de namorado. A tabela 1 mostra os valores médios, em função das litologias observadas no campo de namorado, da densidade (RHOB), raios gama (GR), resistividade (ILD) e sônico (DT). Nesta mesma tabela, foi realizada uma diferenciação das litologias, proveniente das sequências deposicionais mencionadas na figura 12-(a).

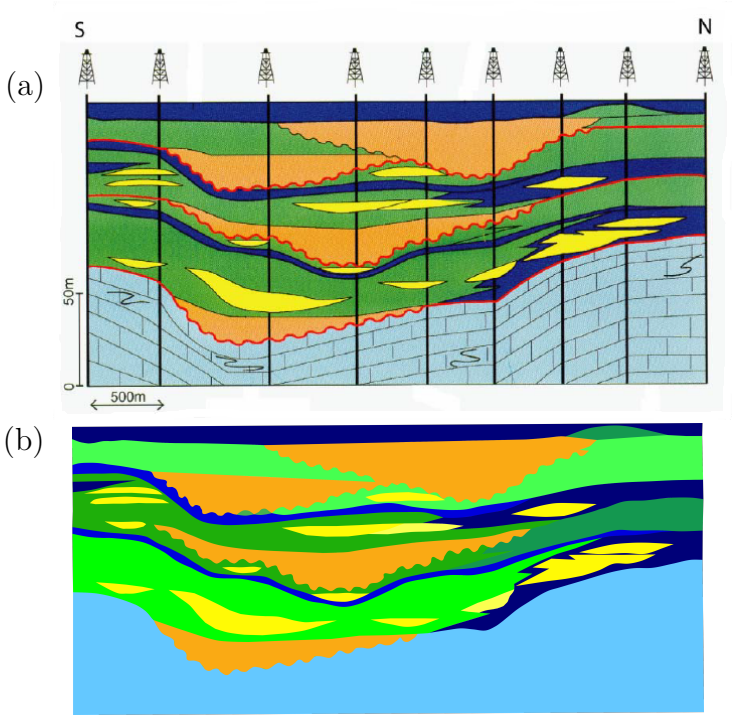


Figura 12 – Seção original de BARBOZA (2005) em (a), e seção desenhada no *Inkscape* em (b).

Tabela 1 – Banco de dados das litologias utilizado na etapa de modelagem de perfis. As litologias depositadas em diferentes sequências são representados pelas numerações em algumas litologias.

Propriedades físicas das litologias do Campo de Namorado.				
Litologia	RHOB (g/cm^3)	GR (API)	ILD ($Ohm.m$)	DT ($\mu s/ft$)
Folhelho1	2.12	115	1.0×10^3	62
Folhelho2	2.25	75	10	167
Folhelho3	2.43	40	1.0×10^5	132
Folhelho4	1.8	140	1.0×10^4	97
Arenito1	1.9	20	10	70
Arenito2	2.4	37	1.0×10^8	80
Marga1	2.75	80	55	70
Marga2	2.7	70	50	75
Conglomerado	2.55	22	1.2×10^4	60
Calcarenito	2.45	7.5	3.5×10^2	80

Após a criação do banco de dados e da seção geológica, a próxima etapa é associar as fácies com as propriedades físicas estabelecidas no banco de dados. Para essa finalidade, foi desenvolvido um *script* e utilizada a linguagem de programação *Python* na plataforma do *Jupyter Notebook*. Esta etapa, que é dividida em quatro passos:

1. Carregamento da seção geológica e da tabela de propriedades;
2. Simulação da perfuração de um poço de interesse;
3. Atribuição das cores de cada fácie aos perfis RHOB, GR, ILD e DT;
4. Determinação dos perfis em função de sorteios baseados em função distribuição de probabilidade Gaussiana.

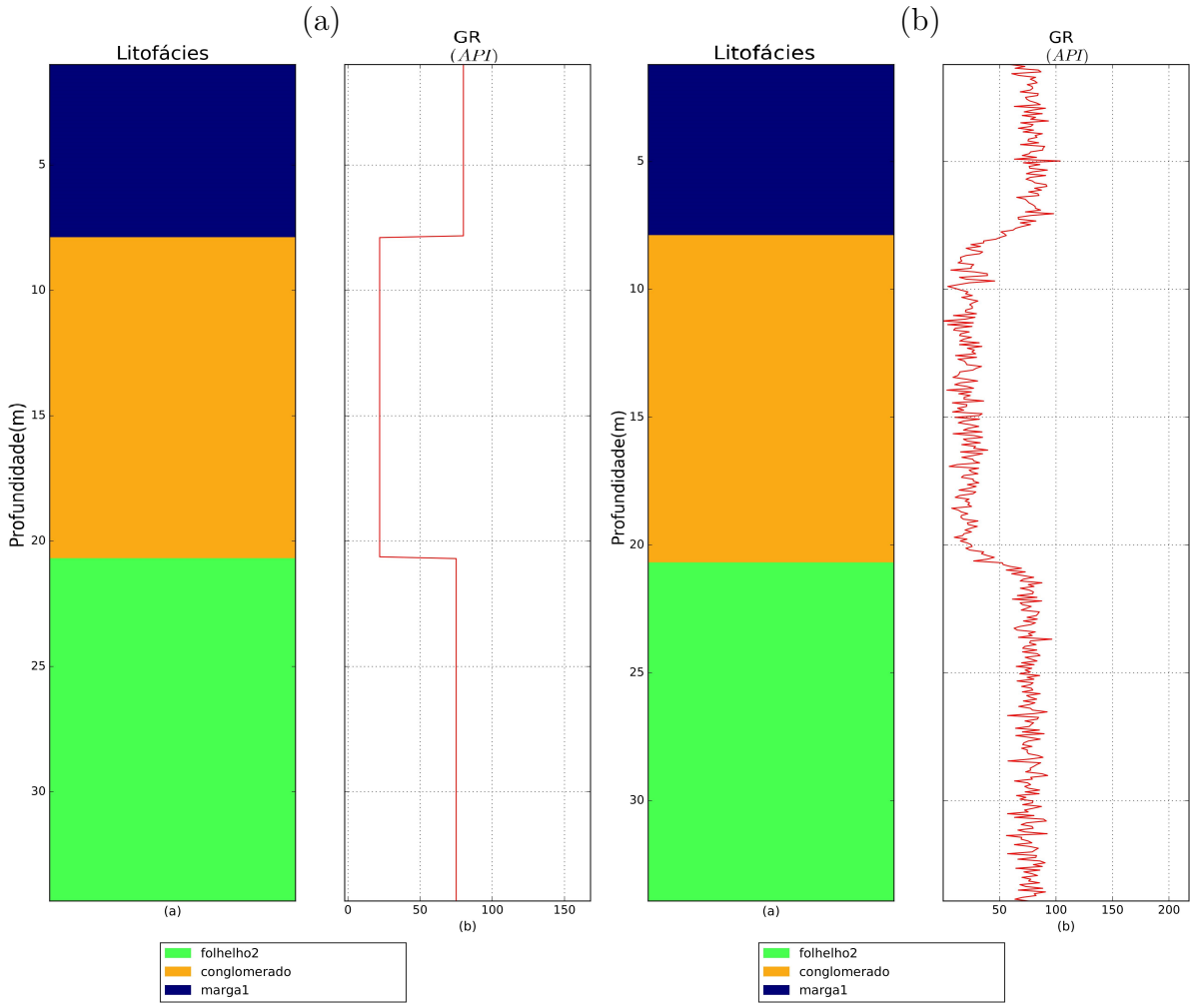


Figura 13 – Perfil de Raios Gama antes e depois da perturbação Gaussiana

Na etapa 3, utilizamos o sistema de cores RGBA (Red, Green, Blue e Alpha, em inglês) estabelecidos na etapa de criação da seção geológica. Vale salientar que a codificação das litofácies adotada neste trabalho é a mesma utilizada e disponibilizada pela Petrobras. Após a atribuição de cores, foi realizada uma perturbação Gaussiana (VIRTANEN et al., 2020) nos valores estabelecidos na tabela, a fim de simular os perfis, conforme mostrado na figura 13. Para suavizar as curvas nas transições entre litofácies, foi aplicada uma convolução janelada em cada um dos perfis utilizados. A figura 13 (a) mostra o perfil de raios gama antes e a figura 13 (b) após a pertubação, em um poço teste.

A figura 14, sintetiza todas as etapas da modelagem de perfis estabelecidas neste trabalho. Observe que os perfis geofísicos apresentam leve suavização na transições entre litologias. Outro aspecto que merece destaque consiste na faixa de ruído Gaussiano estabelecido para os dados, cuja média é zero e o desvio padrão é 10% do máximo valor da propriedade física.

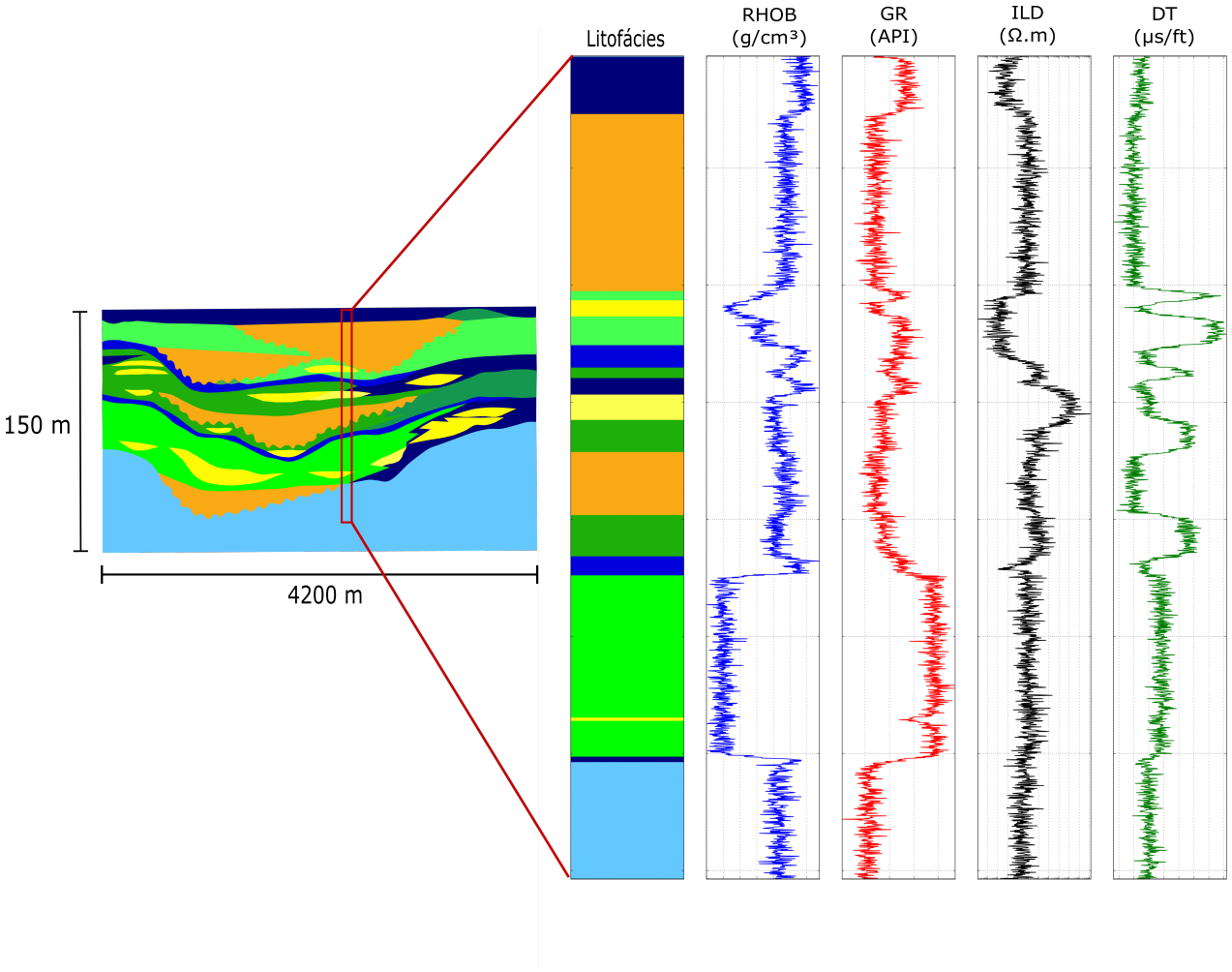


Figura 14 – Ilustração do funcionamento da modelagem de perfis utilizada neste trabalho.

A modelagem é apresentada na figura 14, onde tem-se a seção desenhada no *Inkscape* e a perfuração sintético de um poço, chamado aqui de poço de classificação. Adicionalmente, podem ser observados os perfis geofísicos simulados para o poço de classificação.

A figura 15 mostra os *crossplots* dos dados em pares de perfis, além da distribuição dos dados em histogramas. Neste resultado, notou-se um problema que poderia ocorrer na classificação. Após alguns testes realizados no *K-Means* percebeu-se que o perfil de resistividade (ILD) apresenta uma grande sobreposição dos dados, conforme visto no histograma (ILD), na figura 15. E utilizando-se da base teórica do método, como visto no capítulo 3, o *K-Means* não se comporta bem com essa sobreposição de dados. Outro fator importante é a aplicabilidade deste perfil, como visto no capítulo 2, a resistividade é bom indicador de fluidos em espaços porosos, o que não foi abordado na modelagem. Após as correções de escala com a normalização e a decisão da retirada do perfil de resistividade (ILD), os resultados se tornaram mais consistente.

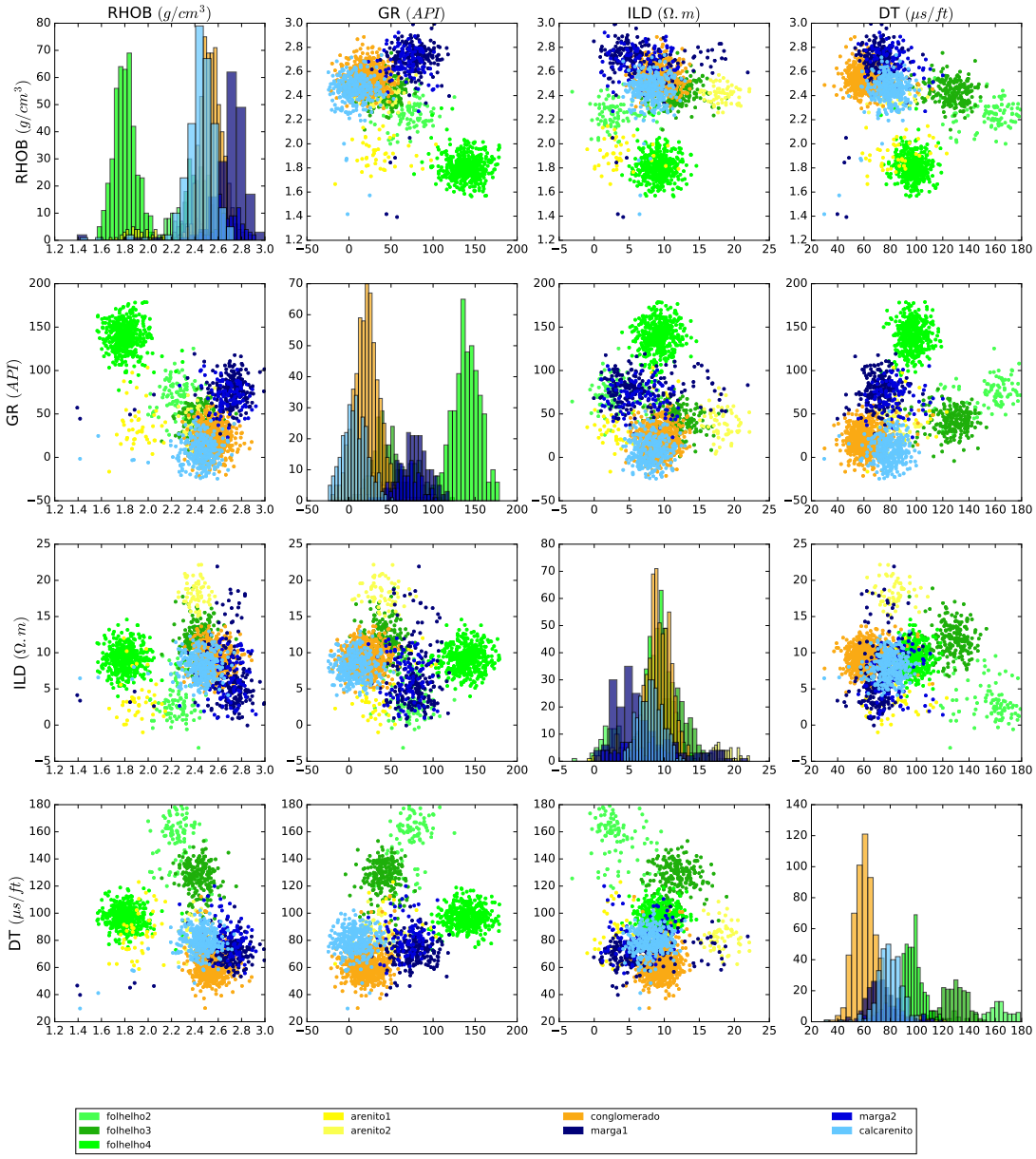


Figura 15 – Gráfico de Dispersão dos perfis sintéticos. Pode-se observar uma grande sobreposição das litologias no histograma (ILD) em relação aos outros perfis.

4.2 O K-Means na reconstrução de perfis litológicos

Findada a modelagem, o próximo passo é a aplicação do *K-Means* para reconstrução litológica. Neste *script*, foram realizados os duas etapas, que são:

1. Análise inical dos dados e implementação do *K-Means*;

2. Correlação entre os centroides do *K-Means* e os centroides verdadeiros;

4.2.1 Análise inicial dos dados e implementação do K-Means

O carregamento dos dados no segundo *script* foi realizado utilizando a biblioteca *Pandas* (MCKINNEY, 2010; TEAM, 2020) do *Python*. Estes que importam as informações dos perfis, profundidade, litologia e dos centroides litológicos, vale salientar que este último retrata as informações contidas no banco de dados inicial. Porém, antes da aplicação do método, foi realizada uma normalização dos dados para não haver tendenciosidade (SANTOS, 2016). Isso porque, os valores absolutos das propriedades variam em muitas ordens de grandeza. Algumas distâncias podem ser menores que outras dependendo do perfil, por exemplo, a densidade que tem uma variação aproximada entre 1.8 - 2.98 g/cm^3 e o perfil acústico que varia de 6 - 167 $\mu s/ft$. Além disso, o parâmetro tolerância (tol) é muito sensível pois ela também depende de cálculos de distâncias, no caso entre centroides na posição atual e anterior (i e i-1), para o critério e parada do algoritmo. Se a tolerância for ($tol = 1$), será um valor alto para o perfil densidade, porém baixo para o perfil acústico. Padronizando os perfis para valores entre 0 e 1, pode-se contornar este problema. Para isto, foi realizada uma normalização dos dados, chamada de *min-max* (equação 4.1).

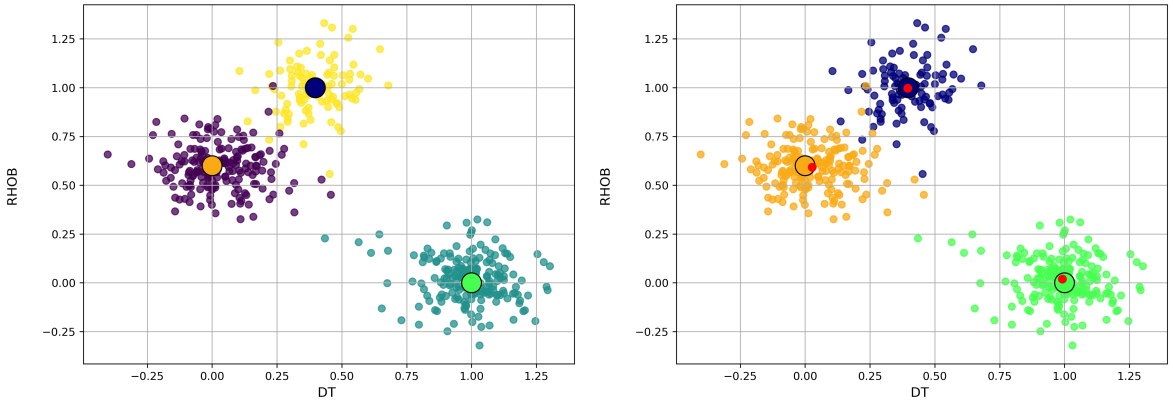
$$X = \frac{x_i - x_{(máx)}}{x_{(máx)} - x_{(min)}}, \tag{4.1}$$

em que, $x_{(máx)}$ é valor máximo do perfil, $x_{(min)}$ é o valor mímimo do perfil, x_i é a amostra que será normalizada e X é a amostra normalizada.

O passo seguinte é a definição dos parâmetros de entrada do *k-means*. No *Sckit-learn*, a função *KMeans* é onde coloca-se os valores dos parâmetros de entrada mostrados no capítulo 3. E na função *y_kmeans* é onde são armazenados so códigos de rochas. Os dados são normalizados e separados em perfis e códigos litológicos (i.e., números inteiros), para estabelecer o perfil litológico pós-classificação.

4.2.2 Correlação entre os centroides do *k-means* e os centroides verdadeiros

A última etapa relevante da nossa metodologia consiste em atribuir as cores das litologias modeladas aos centroides estimados pelo *K-means*. Dessa forma, utilizou-se o banco de dados criados para modelagem dos perfis, além dos gráficos de dispersão, para guiar essa etapa. A solução implementada é a utilização de um cálculo de distâncias entre o centroides do *K-Means* e os centroides verdadeiros, conforme pode ser observado na figura 16. Por proximidade, os grupos recebem as mesmas cores atribuídas na etapa da construção da seção geológica interpretada. A figura 16-(b) mostra o resultado desta implementação em um poço teste com três litologias. Aqui, os centroides do *K-Means*, são representados pela cor vermelha, e os grupos com as devidas cores das litologias da modelagem.



(a) (b)

Figura 16 – (a) Centroides verdadeiros e a classificação do *K-Means*, sem qualquer associação. Já em (b), temos as nuvens de dados classificadas pelo *k-Means* com a respectiva cor do centroide verdadeiro, o mais próximo daquele estabelecido pelo método (ponto em vermelho).

Observe na figura 16-(a) que os centroides verdadeiros não seguem a mesma cor dos grupos identificados pelo k-means. Para contornar esse problema, calculamos as distancias entre os centroides verdadeiros e os estabelecidos pelo k-means, conforme pode ser visto na figura 16-(b).

5 Resultados e Discussões

Neste capítulo, serão apresentados os resultados obtidos após a implementação da metodologia descrita no capítulo anterior. Primeiramente será mostrado uma justificativa da utilização dos parâmetros de entrada no *script* e, posteriormente serão apresentados os resultados das três formas de inicializações do *K-Means* utilizadas neste trabalho. A saber, o *random*, o *k-means++* e a definição determinística dos centroides, chamado aqui de *centroides a priori*. A sequência de resultados serão apresentados por gráficos de dispersão de pares de propriedades físicas, seguido da classificação litológica do *K-Means* através da comparação dos perfis. O trabalho ainda contempla um cálculo de erros percentuais das classificações obtidas através de cada uma das diferentes inicializações. Também apresenta uma análise de litofácies comparando as litologias verdadeiras das classificadas pelo *K-Means*.

5.1 Parâmetros de entrada utilizados

Foram definidos nove centroides como possíveis grupos, já que, esta é a quantidade de litologias presentes no poço de classificação. Essa premissa é fundamental para a boa utilização do *K-Means*. O número de iterações máxima usado foi 50, isso porque a teoria apresentada no capítulo 3 mostra que o *K-Means* converge rapidamente, não havendo necessidade de números exorbitantes de iterações. Além disso, nos testes realizados, quando se achou um conjunto de parâmetros ideais, esse valor oscilava entre 5 e 35 iterações, o que está de acordo com a teoria. A tolerância usada foi de 10^{-4}, isso porque, quando se usava um valor muito baixo para o método, o *K-Means* não convergia. E quando a tolerância era muito baixa (tol = 10^{-50}), não havia convergência adequada, e a quantidade de iterações realizadas era demasiadamente alta. E valores altos de tolerância como (tol = 1 - 10) o método não realizava iterações.

Depois de estabelecido alguns parâmetros, foram geradas diversas classificações com diferentes parâmetros de entrada, com o objetivo de compreender as potencialidades e limitações do método, disponível na biblioteca *sckit-learn*. Dessa forma, focou-se em reproduzir resultados com diferentes formas de inicializações, são elas: *random*, o *k-means++*, e o *centroides a priori*.

5.2 Inicialização aleatória de centroides *(random)*.

O primeiro resultado a ser analisado é relativo a inicialização *random*. Nesta inicialização, o *K-Means* estimou sete das nove litologias verdadeiras, além disso, realizou

dezoito iterações e obteve um erro amostral de 17.61%.

Na figura 17 pode-se observar os gráficos de dispersão envolvendo os perfis GR, RHOB e DT. Nestes é possível observar o comportamento da convergência dos centroides em *crossplots*. Nesta inicialização, nota-se a dificuldade de agrupamento para as Margas e Arenitos, devido a natureza de sobreposição dos dados, além da baixo conjunto amostral. Na figura 17-(a) a Marga2 aparentemente não foi contemplada na classificação. Ainda na mesma, os Folhelho2 aparece com excesso de centroides, talvez aquele que classificaria a Marga2. No gráfico 17-(b), pode-se inferir que o conglomerado também aparenta ter um centroide em excesso, adicionalmente, o Arenito2 parece não ter sido classificado. A não classificação do Arenito2 pode ter ocorrido pela pouca presença de amostras, em comparação às amostras de Calcarenito próximas. O gráfico 17-(c) mostra a dificuldade de diferenciação entre os Arenitos em relação a outras litologias próximas, devido a fenômenos relacionados à proximidade e sobreposição. Com isso, é possível observar um excessos de centroides no Conglomerado na figura 17-(b)que não é nítido na figura 17-(a). Isso porque o problema tem três dimensões e *crossplots* em pares podem não passar todas as informações necessárias, já que um centroide próximo de um grupo em um plano, pode não ser verdade quando se utiliza três dimensões. A análise inicial feita aqui, tem que ser comparada aos perfis litológicos reconstruídos, que serão mostrados a seguir.

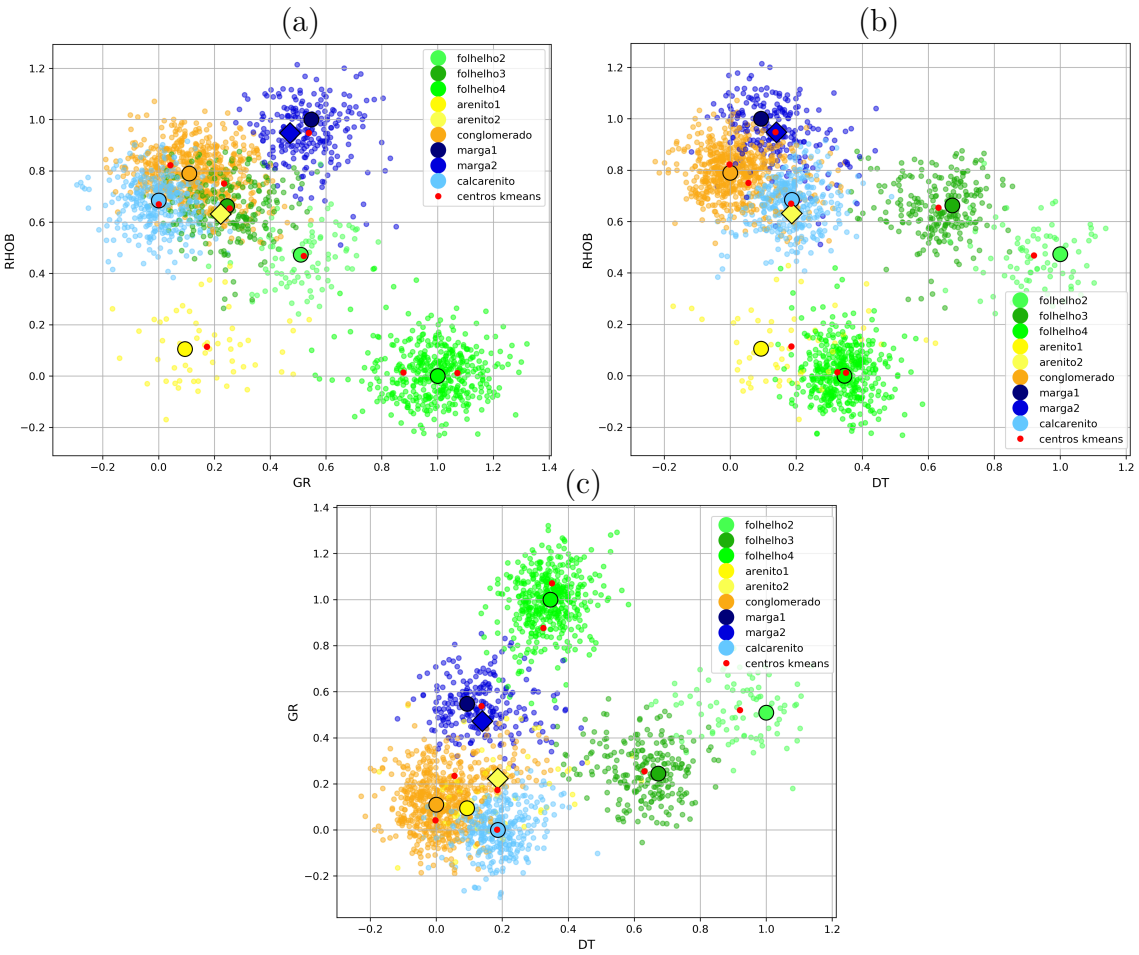


Figura 17 – Dispersão dos perfis normalizados pelo método min-max (a) RHOB vs GR, (b) RHOB vs DT e (c) GR vs DT. Os pontos coloridos indicam os dados classificados, os pontos coloridos maiores mostram os centroides verdadeiros. Já os pontos pequenos em vermelho destacam a posição dos centroides, estabelecidos pelo *K-Means* ao final da aplicação.

Na figura 18 observa-se que a classificação confundiu o pacote de Arenito1 com Calcarenito e Conglomerado, destacado em vermelho na faixa dos 60 metros de profundidade. Pode-se observar, destacado em vermelho entre 10 e 40 metros, que a classificação do Conglomerado apresenta artefatos de Calcarenito. Isto se deve talvez pela natureza sobreposta dessas litologias. Ainda neste perfil pode-se notar a não classificação da Marga1, mencionada nos *crossplots* das figuras 17, corroborando a análise inicial. Alguns outros erros pontuais como artefatos, são apontados pelas setas em vermelho. Porém, o *K-Means* se saiu bem em identificar os intervalos de várias litologias, reconstruindo bem as intercalações das eletrofácies no perfil litológico.

Vale ressaltar a boa recuperação litológica dos Arenitos1, Marga2, Folhelho2 e Folholho4, deestacados em azul entre 40 e 55 metros. Também tem-se uma boa classificação no Calcarenito na base do perfil e até mesmo camadas finas de Arenito1 e Marga2 dentro dos pacotes de Folhelhos entre 80 e 120 metros. Isso mostra que, salvo alguns erros, o

K-Means com a inicialização *random*, foi hábil na recuperação de muitas litologias. Achou e restaurou bem os intervalos de muitas litofácies, mesmo aquelas relativamente finas e difíceis.

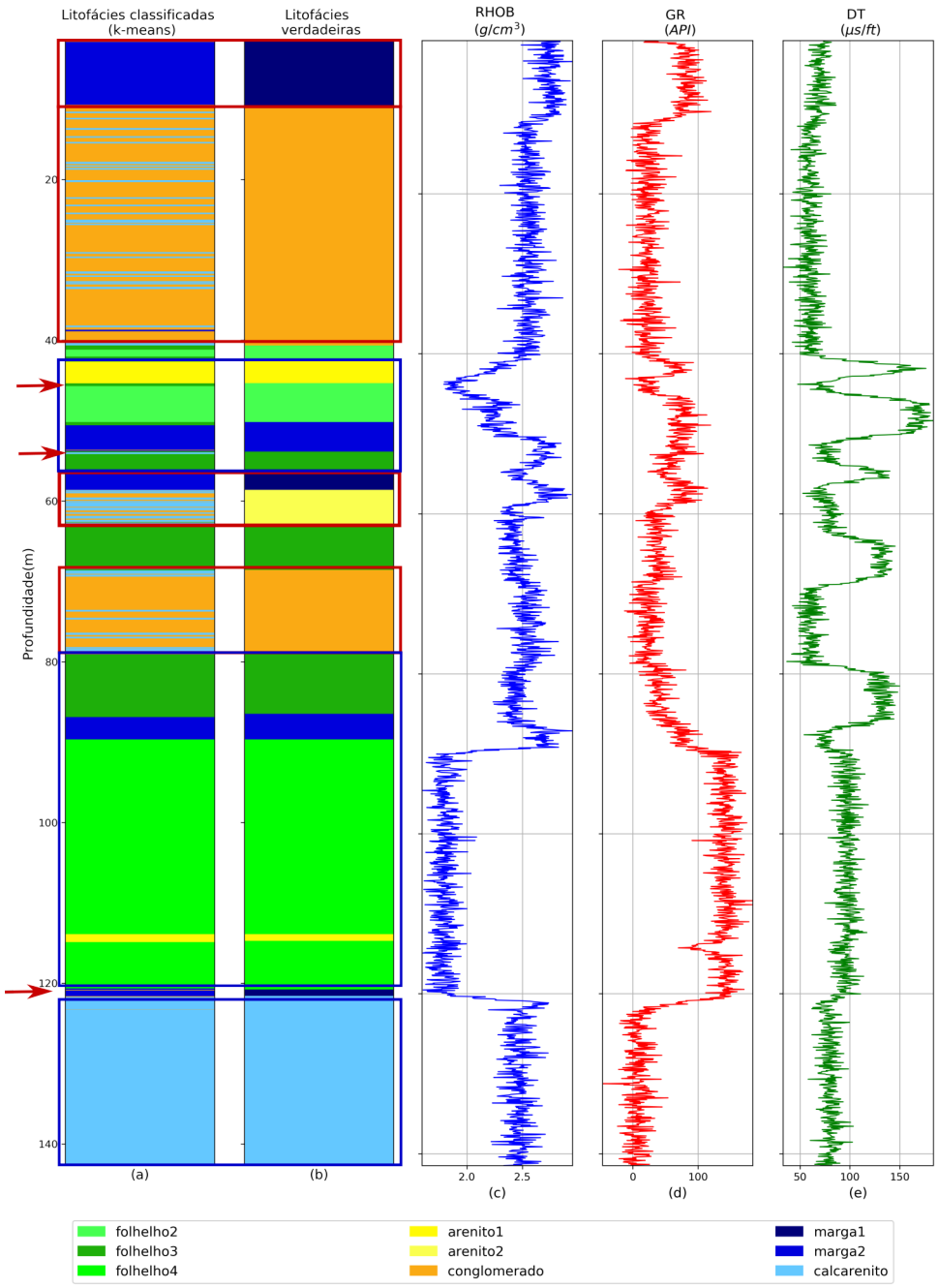


Figura 18 – Perfis (a) de classificação, (b) verdadeiro, (c) RHOB, (d) GR e (e) DT para a inicialização *random*. Os retângulos em vermelho mostram as regiões de erro de classificação, enquanto que os retângulos em azul indicam as litofácies corretamente classificadas pelo *K-Means*.

5.3 Inicialização dos centroides via método *k-means++*.

O segundo teste é referente a inicialização utilizando a entrada *k-means++*. Esta que recuperou oito das nove litologias, e de novo, com as Margas sendo uma das mais difíceis de serem diferenciadas. Porém houve um progresso em relação aos Arenitos, que foram minimamente identificados. Neste teste chegou-se a um erro amostral de 21.17% e foram realizados onze iterações. Com estes resultados, pode-se dizer que com a inicialização *k-means++* obteve-se mais sucesso nos resultados amostrais em relação a inicialização *random*. Além disso, a convergência foi alcançada em um número menor de iterações em relação ao teste anterior. E o resultado mais importante, o *k-means* inicializado pelo método *k-means++* foi capaz de classificar mais litologias do que a inicialização anterior. Ou seja, aparentemente para esse conjunto de dados, o resultados do *k-means++* foi mais satisfatório.

Analisando os gráficos de dispersão na figura 19, pode-se ter uma melhor noção de como os centroides convergiram para este resultado. A figura 19-(a) mostra que o Folhelho4 aparece com dois centroides e a Marga2, mais uma vez não é classificada pelo método. O Conglomerado que tinha aparentemente dois centroides na figura 17-(b) não apresenta ter este problema na figura 19-(b). Ainda na mesma figura, dois centroides aparecem classificando o Calcarenito e o Arenito2 o que não aconteceu na inicialização anterior. Com isso, pode-se apontar que o Arenito2 foi contemplado na classificação nesta inicialização. E novamente aparece o excesso de centroides no Folhelho4. Ainda na figura 19-(b), apesar do centroide do *K-Means* encontrar o Arenito1 ele parece relativamente distante do centroide verdadeiro, em amarelo. Isso pode provocar uma classificação errada de amostras de Folhelho4 em Arenito1. Na figura 19-(c), apesar da sobreposição de dados entre Calcarenito, Conglomerados e os Arenitos, o *K-Means* conseguiu distinguir, pelo menos parcialmente, as litologias presentes. Pode-se notar também que os centroides dos grupos bem definidos como, os Folhelhos e Conglomerados convergiram bem, em relação aos centroides verdadeiros. E em todos os *crossplots* a Marga2 parece não ter sido classificada.

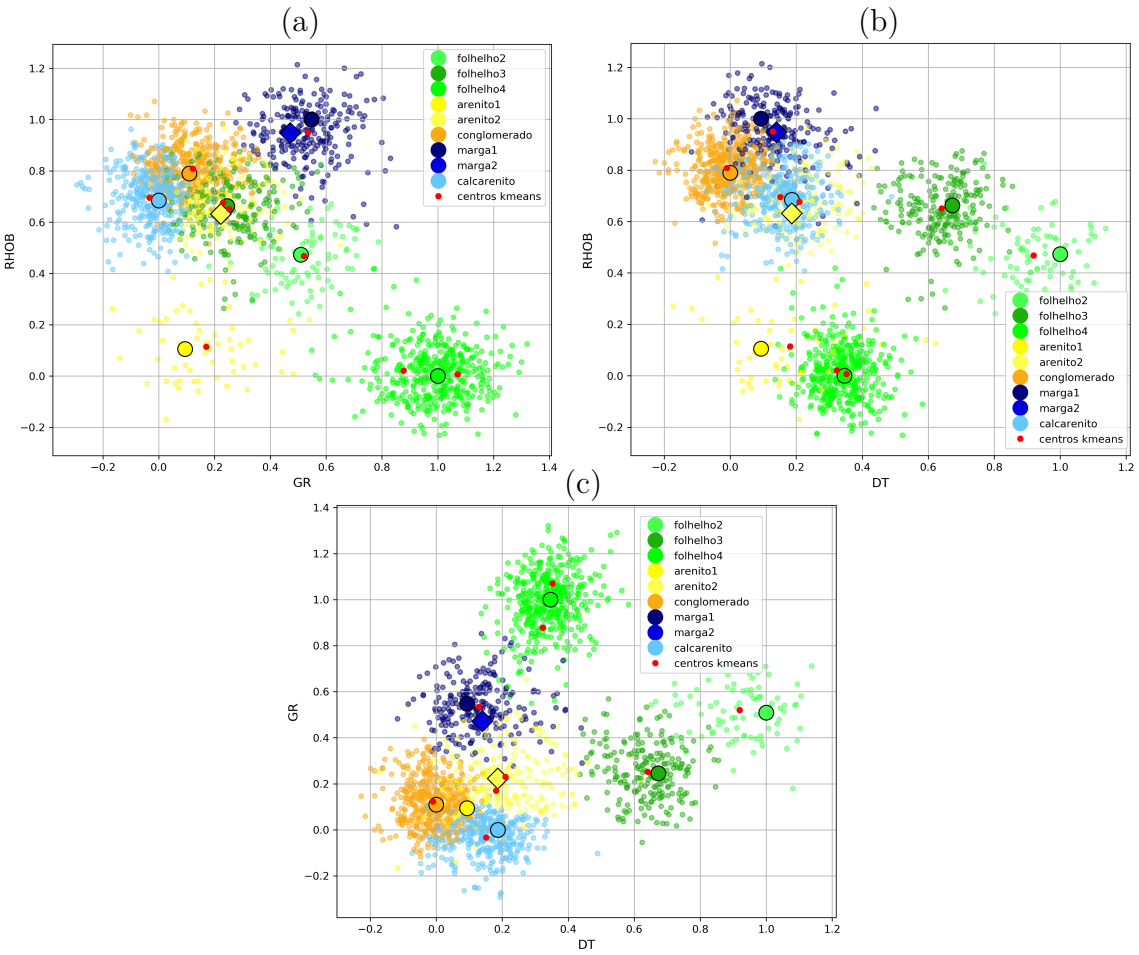


Figura 19 – Dispersão dos perfis normalizados pelo método min-max (a) RHOB vs GR, (b) RHOB vs DT e (c) GR vs DT. Os pontos coloridos indicam os dados classificados, os pontos coloridos maiores mostram os centroides verdadeiros. Já os pontos pequenos em vermelho destacam a posição dos centroides, estabelecidos pelo *K-Means* ao final da aplicação.

Na figura 20 temos a reconstrução litológica dos perfis para o *k-means++*. Aqui, tem-se novamente a ausência da Marga1 ao longo de todo o perfil. Há vários artefatos na classificação litológica do Conglomerado entre as profundidades de 10 e 40 metros, um erro mais expressivo do visto na figura 18. Além disso, na região destacada em vermelho na profundidade entre 70 m e 80 metros, onde também aparecem artefatos de Calcarenitos na classificação do Conglomerado. Também pode ser visto artefatos de Arenitos no Calcarenito. Tais intercalações podem estar associadas à proximidade entre os arenito e Calcarenito e entre Calcarenito e Conglomerados apresentados nos *crossplots* da figura 19. Ainda é possível observar alguns erros pontuais apontados pelas setas vermelhas. Apesar dos erros, alguns pacotes litológicos foram bem reconstruídos. O Arenito1 intercalado no Folhelho2 foi bem classificado e sendo bem assertivo nos limites de topo e base. A Marga1 no topo também foi bem reconstruída, uma evolução nítida em relação a inicialização anterior. A região entre 55 e 65 metros reconstruiu bem os pacotes de Arenito2, Marga1 e Folhelho3.

O Folhelho4 entre 90 e 120 metros e até mesmo a camada fina de Arenito1 em 115 metros, foram bem reconstruídos. Os perfis classificados mostram que as análises iniciais realizadas nos *crossplots* da figura 19 foram bem assertivas. Além disso, a inicialização *k-menas++* mostrou uma evolução positiva em relação a inicialização *random*.

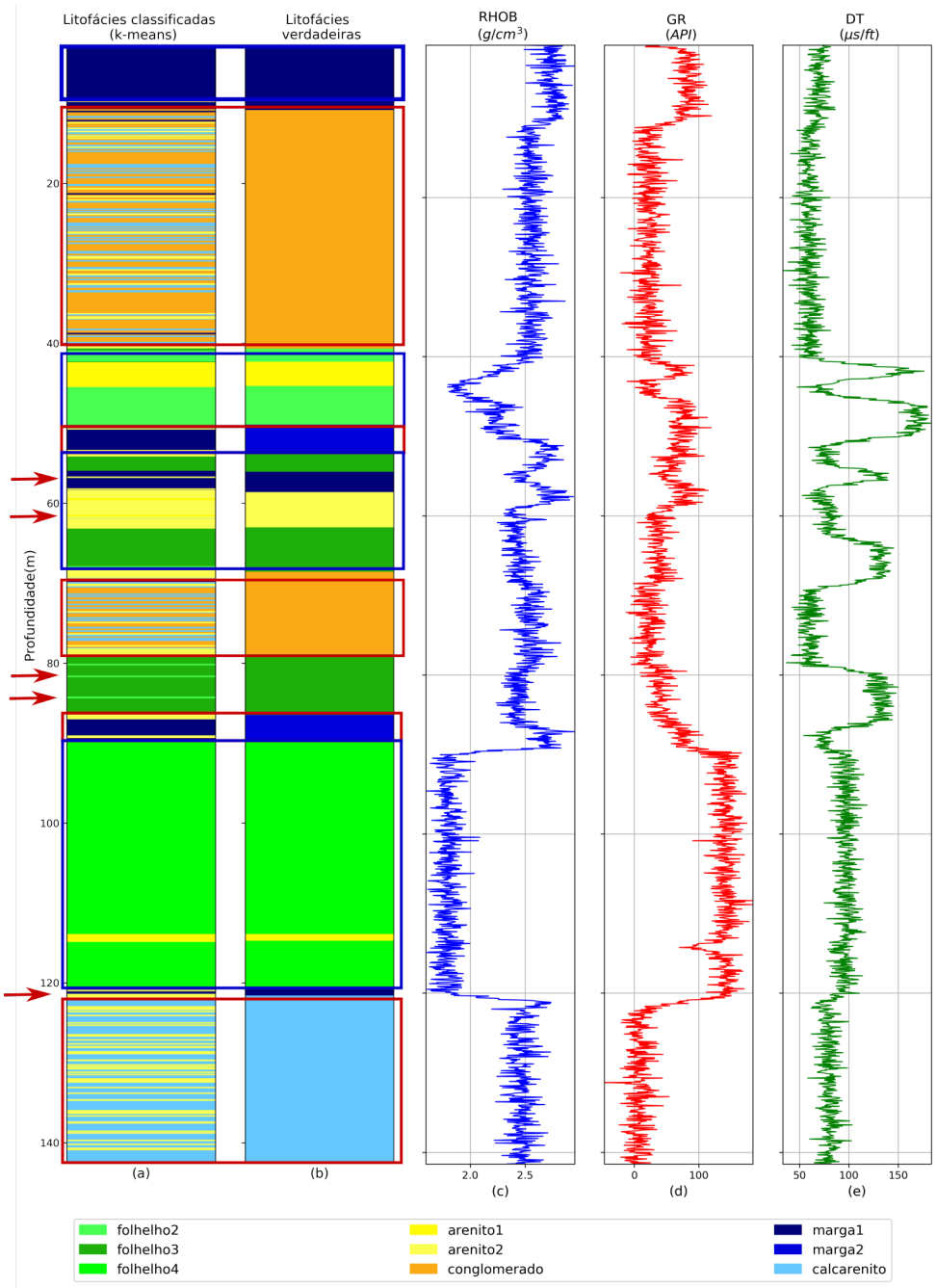


Figura 20 – Perfis (a) de classificação, (b) verdadeiro, (c) RHOB, (d) GR e (e) DT para a inicialização *k-means++*. Os retângulos em vermelho mostram as regiões de erro de classificação, enquanto que os retângulos em azul indicam as litofácies corretamente classificadas pelo *K-Means*.

5.4 Inicialização determinística dos centroides

Finalmente, neste teste é apresentado a inicialização determinística dos centroides do *K-Means*. Neste, o conjunto de centroides são formados por: 5 gerados aleatoriamente e 4 escolhidos manualmente. Estes últimos foram considerados como informação a priori, chamados aqui de *centroides a priori*. As litologias escolhidas foram as que mais causaram dificuldades ao *K-Means*, são elas: Marga1, Marga2, Arenito1 e Arenito2. Com o *centroides a priori*, o *K-Means* apresentou um erro amostral de 17.82% e realizou dezenove iterações. Nesta inicialização o erro ficou menor em comparação ao *random* e ao *k-means++*. Além disso, o método conseguiu classificar todas as nove litologias estabelecidas no poço sintético.

A figura 21, mostra o resultado da classificação com o *centroides a priori* através dos gráficos de dispersão. Nota-se que nestes gráficos, cada centroide do *K-Means*, aparentemente, conseguiu convergir para uma posição próxima dos centroides verdadeiros da modelagem. Aqui, as litologias Margas1, Marga2, Arenito1 e Arenito2 foram contempladas. Na figura 21-(a) o aglomerado formado pelas duas Margas há dois centroides do *K-Means*, o que não obteve-se nos testes anteriores. Além disso o Arenito1, que na inicialização *random* não foi classificada é restaurado nessa inicialização. O excesso de centroides em Folhelho4 presente nas figuras 19 e 17, não é observado neste teste. Na figura 21-(b), pode-se notar alguns problemas na reconstrução litológica. Isso porque, embora o Arenito1 tenha sido classificado, o centroide do *K-Means* pode ter agrupado amostras erradas. Isto pode ter ocorrido talvez pela grande quantidade de amostras de Folhelho4 próximas ao Arenito1. Consequentemente, pode ter ocorrido uma atribuição errada de amostras de Arenito1 sendo que trata-se do Folhelho4. O mesmo comportamento pode ser visto com os Calcarenito, Conglomerados, Arenito2 e as Margas, já que neste aglomerado, os centroides do *K-Means* parecem distantes de alguns centroides verdadeiros. Na figura 21-(c), também pode ter o mesmo equivoco na classificação, devido a sobreposição de amostras, entre os Arenitos, Calcarenitos e Conglomerado, como citado na figura 21-(b). Contudo, litologias bem definidas parecem ter sido bem recuperadas, resta analisar o resultado em perfis.

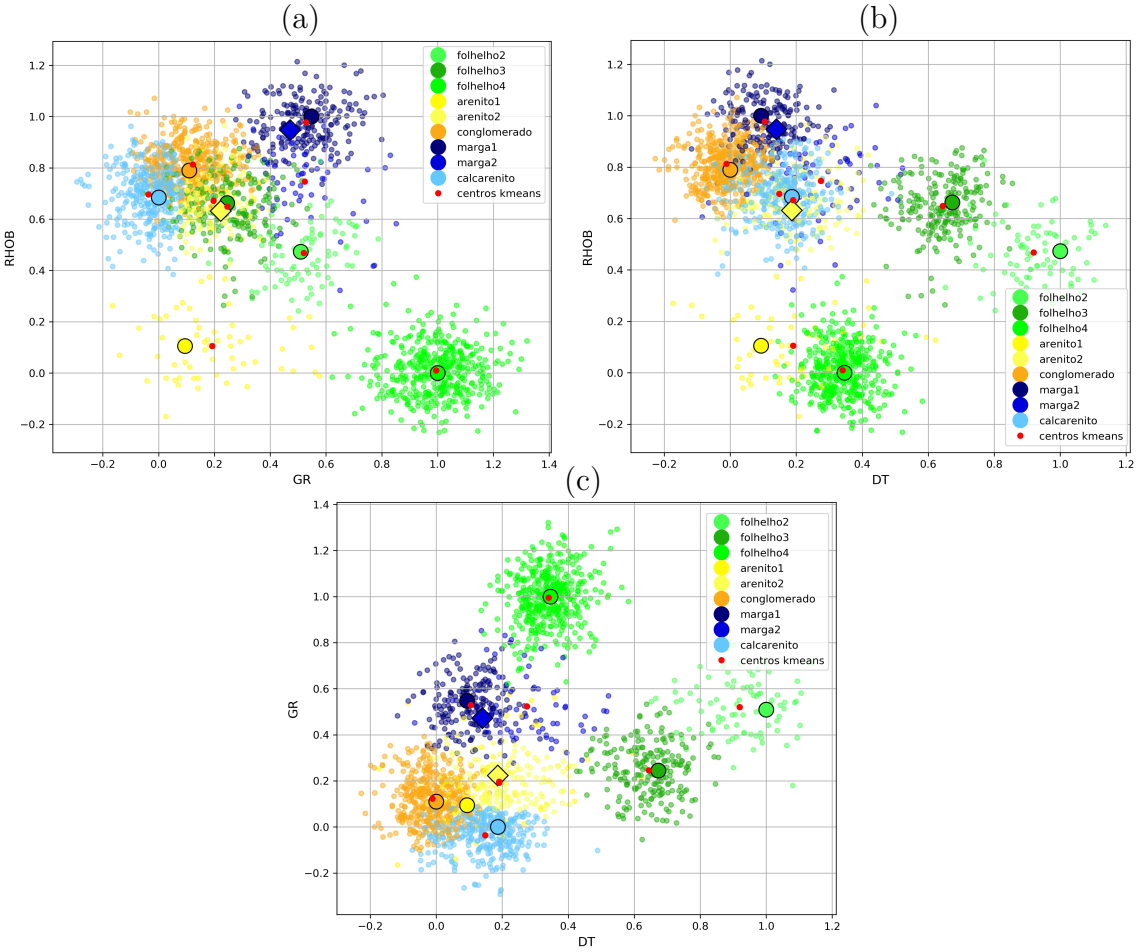


Figura 21 – Dispersão dos perfis normalizados pelo método min-max (a) RHOB vs GR, (b) RHOB vs DT e (c) GR vs DT. Os pontos coloridos indicam os dados classificados, os pontos coloridos maiores mostram os centroides verdadeiros. Já os pontos pequenos em vermelho destacam a posição dos centroides, estabelecidos pelo *K-Means* ao final da aplicação. O método de inicialização do *k-Means* em discussão é o *centroides a priori*.

A figura 22 mostra a classificação litológica com a inicialização determinística com quatro centroides, referentes à (Arenito1, Arenito2, Marga1 e Marga2). Primeiramente, pode-se notar um equivoco na classificação, em vermelho, da Marga entre 0 e 10 metros de profundidade. Pode-se inferir que esse resultado se deve a sobreposição das Margas, presente em todos os gráficos de dispersão na figura 21. Característica semelhante na classificação pode ser encontrado entre as profundidades 120 e 140 metros, destacado em vermelho, novamente tem-se a presença de artefatos na região do Calcarenito. O Conglomerado entre 10 e 40 metros, destacado em azul, teve uma evolução apresentando quase nenhum artefatos, visto muito nas inicialização *k-means++* e *random*. Outra região bem classificada está entre 45 e 50 metros, onde os limites entre Arenito1 para o Folhelho2 foram bem restaurados. Em geral os resultados destacados em azul apresentaram reconstrução bem satisfatória. A região entre 40 e 60 metros conseguiu-se recuperar todas as litologias, porém

com alguns artefatos devido a sobreposição e poucas amostras de Arenitos e Margas. A lâmina de Arenito em aproximadamente 118 metros foi reconstruída apesar da finura desta camada. Dessa forma, podemos mostrar a importância da informação a priori como alternativa para o bom uso do *K-Means*.

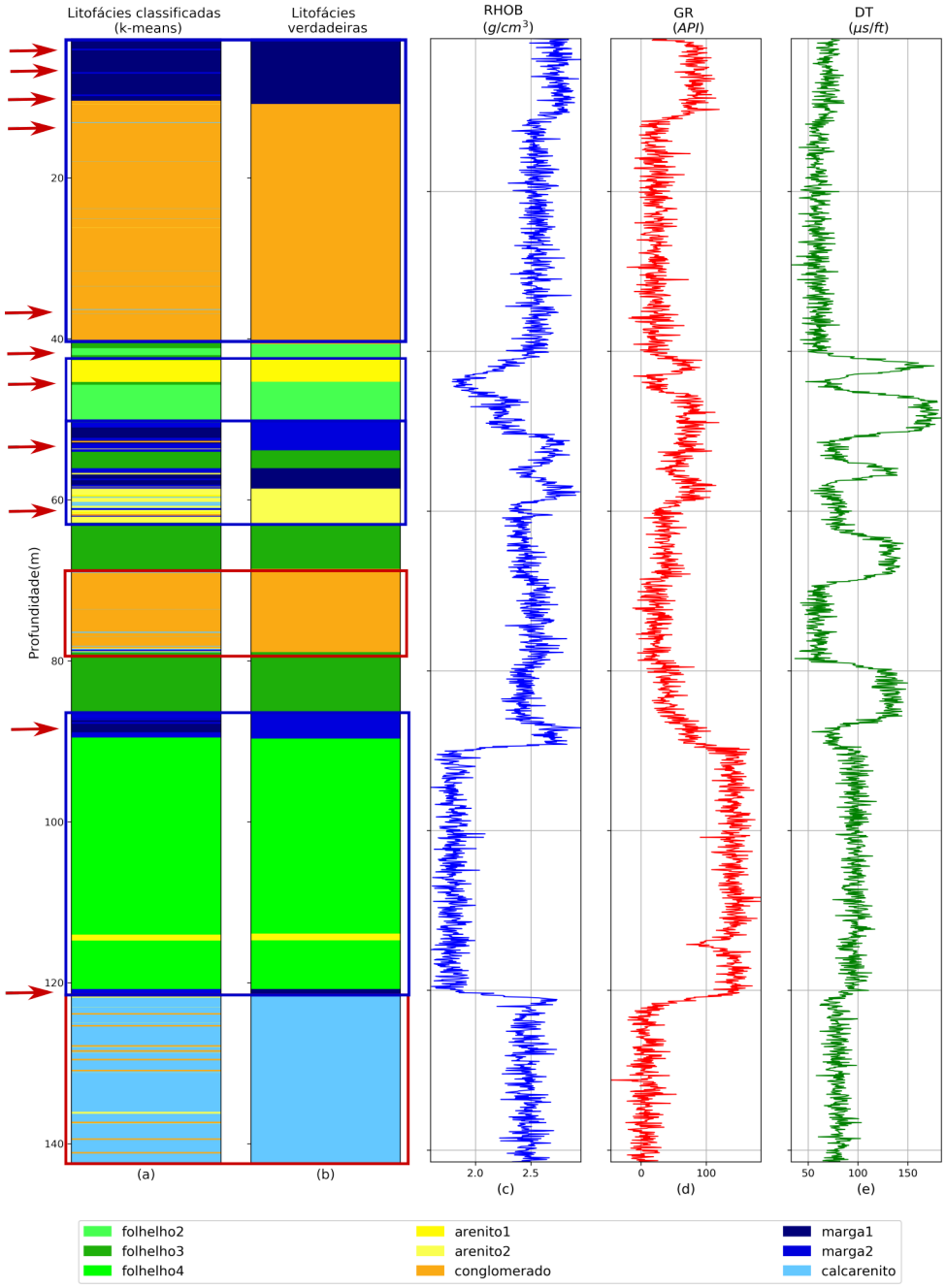


Figura 22 – Perfis (a) de classificação, (b) verdadeiro, (c) RHOB, (d) GR e (e) DT para a inicialização *centroides a priori*. Os retângulos em vermelho mostram as regiões de erro de classificação, enquanto que os retângulos em azul indicam as litofácies corretamente classificadas.

6 Conclusões

O trabalho apresentou uma forma de reconstrução de litofácies a partir de alguns perfis geofísicos sintéticos inspirados no campo de namorado, Bacia de Campos. Para isso, utilizou-se o método *K-Means* com diferentes tipos de inicialização. Inicialmente, perfis geofísicos foram simulados a partir da seção geológica interpretada referente ao campo de namorado. Em seguida, utilizamos o método *k-Means*, disponível na biblioteca *sckit-learn*, com diferentes tipos de inicialização de centroides, para a reconstrução litológica de um poço sintético.

A inicialização *random* apresentou o resultado menos consistente, em relação as demais inicializações. Este teve dificuldades em classificar finas litofácies das Margas1 e Arenito2. Contudo, conseguiu recuperar algumas litofácies complexas preservando os limites de topo e base, mostrando a força do *K-Means*, mesmo numa inicialização mais simples. A inicialização *k-means++* obteve um resultado melhor que a *random*, em relação ao número de litologias recuperadas. Porém, apresentou um aumento na quantidade de artefatos na classificação, provavelmente devido à sobreposição dos dados de algumas litologias. Nesta inicialização conseguiu-se recuperar os dois lóbulos de Arenitos perfurados na simulação do poço sintético. A inicialização determinística de centroides, chamada de *priori-centros*, foi utilizado como uma forma de inserir informações a priori, com a finalidade de aprimorar a classificação. Esta inicialização permitiu ao *K-Means* recuperar todas as litologias perfuradas no poço sintético. Isto mostra a importância de inserir informações extras e refinar o método. O *K-Means* ainda conseguiu recuperar as Margas além dos lóbulos de Arenitos, feitos que não foram possíveis nas outras inicializações. Vale ressaltar que, mesmo com uma forte sobreposição de alguns dados, o *K-Means* obteve bons resultados na reconstrução de litofácies e delimitação de topos e bases no perfil. Mesmo camadas finas de Arenitos e Folhelhos foram bem classificados em alguns testes.

Como perspectivas, pode-se destacar a possibilidade de realizar um estudo controlado usando outros métodos não-supervisionados, tais como o *Expectation Maximization* ou *DBSCAN - Density-Based Spatial Clustering of Applications with Noise, em inglês.* Alternativamente, métodos supervisionados, como so *Mapas auto-organizáveis (SOM)* e as *Redes Neurais (RNAs)* podem ser consideradas. Com isto, pode-se ter diferentes referenciais, para um mesmo conjunto de perfis geofísicos e possivelmente uma classificação final mais assertiva. Por fim, o fluxo de trabalho descrito neste documento pode ser utilizado na investigação em dados reais, uma vez que podemos verificar as potencialidades e limitações do *K-Means* para a aplicação sintética.

Referências

ARAGÃO, M. M. C. A. Modelagem De Zonas De Fluxo No Campo De Namorado – Bacia de Campos, RJ. 2017.

ARORA, P.; VARSHNEY, S. et al. Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, Elsevier, v. 78, p. 507–512, 2016.

ARTHUR, D.; VASSILVITSKII, S. How slow is the k-means method? *Proceedings of the Annual Symposium on Computational Geometry*, v. 2006, p. 144–153, 2006.

BARBOZA, E. *Análise Estratigráfica do Campo de Namorado com base na interpretaçao Sısmica Tridimensional*. Tese (Doutorado) — Doctorate Thesis–UFRGS, Brazil, 2005.

BARSTUGAN, M.; OZKAYA, U.; OZTURK, S. Coronavirus (covid-19) classification using ct images by machine learning methods. *arXiv preprint arXiv:2003.09424*, 2020.

BATES, R.; JACKSON, J. Glossary of geology: Falls church. *Virginia, American Geological Institute*, v. 167, 1980.

BESTAGINI, P.; LIPARI, V.; TUBARO, S. A machine learning approach to facies classification using well logs. In: *Seg technical program expanded abstracts 2017*. [S.l.]: Society of Exploration Geophysicists, 2017. p. 2137–2142.

BIAMONTE, J. et al. Quantum machine learning. *Nature*, Nature Publishing Group, v. 549, n. 7671, p. 195–202, 2017.

CALDERÓN-MACÍAS, C.; SEN, M. K.; STOFFA, P. L. Artificial neural networks for parameter estimation in geophysics [link]. *Geophysical prospecting*, European Association of Geoscientists & Engineers, v. 48, n. 1, p. 21–47, 2000.

CAPÓ, M.; PÉREZ, A.; LOZANO, J. A. An efficient approximation to the k-means clustering for massive data. *Knowledge-Based Systems*, Elsevier, v. 117, p. 56–69, 2017.

CARREIRA, V.; NETO, C. P.; BIJANI, R. A comparison of machine learning processes for classification of rock units using well log data. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. *80th EAGE Conference and Exhibition 2018*. [S.l.], 2018. v. 2018, n. 1, p. 1–5.

CHANG, D. X.; ZHANG, X. D.; ZHENG, C. W. A genetic algorithm with gene rearrangement for K-means clustering. *Pattern Recognition*, v. 42, n. 7, p. 1210–1222, 2009. ISSN 00313203.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, Springer, v. 20, n. 3, p. 273–297, 1995.

CRICK, F.; KOCH, C. The unconscious homunculus. *Neuropsychoanalysis*, Taylor & Francis, v. 2, n. 1, p. 3–11, 2000.

DAVID, A.; SERGEI, V. k-means++: The Advantages of Careful Seeding. *NEC Research and Development*, v. 37, n. 3, p. 369–381, 1996. ISSN 0547051X.

DAYAN, P.; SAHANI, M.; DEBACK, G. Unsupervised learning. *The MIT encyclopedia of the cognitive sciences*, MIT Press, p. 857–859, 1999.

DELLAERT, F. *The expectation maximization algorithm*. [S.l.], 2002.

DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 39, n. 1, p. 1–22, 1977.

DIAS, L. O. et al. Comparação de métodos de Segmentação de Fraturas em Imagem Acústica de Perfilagem Petrofísica. *Notas Técnicas*, v. 8, n. 3, p. 7–19, 2018.

DOUGHERTY, J.; KOHAVI, R.; SAHAMI, M. Supervised and unsupervised discretization of continuous features. In: *Machine learning proceedings 1995*. [S.l.]: Elsevier, 1995. p. 194–202.

DU, H.-K. et al. Seismic facies analysis based on self-organizing map and empirical mode decomposition. *Journal of Applied Geophysics*, v. 112, p. 52–61, 2015.

ELKAN, C. Using the Triangle Inequality to Accelerate k-Means. *Proceedings, Twentieth International Conference on Machine Learning*, v. 1, p. 147–153, 2003.

FAHIM, A. et al. An efficient enhanced k-means clustering algorithm. *Journal of Zhejiang University-Science A*, Springer, v. 7, n. 10, p. 1626–1633, 2006.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936.

FREITAS, F. D. S. Modelagem geoelétrica de reservatórios em ambientes de águas doces: estudos da sensibilidade de medidas de ip-resistividade na exploração petrolífera. *Trabalho de Graduação–Universidade Federal da Bahia, Salvador*, 2008.

FRITZKE, B. Growing cell structures—a self-organizing network for unsupervised and supervised learning. *Neural networks*, Elsevier, v. 7, n. 9, p. 1441–1460, 1994.

GUILLEN, P. et al. Supervised learning to detect salt body. In: *SEG Technical Program Expanded Abstracts 2015*. [S.l.]: Society of Exploration Geophysicists, 2015. p. 1826–1829.

HARTIGAN, J. A.; WONG, M. A. A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series B Methodological*, v. 28, n. 1, p. 100–108, 1979.

JIA, Y.; MA, J. What can machine learning do for seismic data processing? an interpolation application. *Geophysics*, Society of Exploration Geophysicists, v. 82, n. 3, p. V163–V177, 2017.

JR, R. M.; VAIL, P.; III, S. T. Seismic stratigraphy and global changes of sea level: Part 2. the depositional sequence as a basic unit for stratigraphic analysis: Section 2. application of seismic reflection configuration to stratigraphic interpretation. AAPG Special Volumes, 1977.

KEAREY, P. et al. An Introduction to Geophysical Exploration. *São Paulo, Oficina de Textos*, v. 133, p. 46–73, 2009. ISSN 18791956.

KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, v. 43, 1982.

KONATÉ, A. A. et al. Prediction of porosity in crystalline rocks using artificial neural networks: an example from the chinese continental scientific drilling main hole. *Studia Geophysica et Geodaetica*, Springer, v. 59, n. 1, p. 113–136, 2015.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

KUMAR, B.; KISHORE, M. Electrofacies classification–a critical approach. In: *6th International Conference & Exposition on Petroleum Geophysics, New Delhi, India*. [S.l.: s.n.], 2006. p. 822–825.

KURODA, M. C. et al. Electrofacies characterization using self-organizing maps. *Brazilian Journal of Geophysics*, v. 30, n. 3, 2012.

KUYUK, H. S. et al. Application of k-means and Gaussian mixture model for classification of seismic activities in Istanbul. *Nonlinear Processes in Geophysics*, v. 19, n. 4, p. 411–419, 2012. ISSN 10235809.

LAUZON, F. Q. An introduction to deep learning. In: IEEE. *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*. [S.l.], 2012. p. 1438–1439.

LLOYD, S. Least squares quantization in pcm. *IEEE transactions on information theory*, IEEE, v. 28, n. 2, p. 129–137, 1982.

LUCIA, F. J.; KERANS, C.; JENNINGS, J. W. Carbonate reservoir characterization. *Journal of Petroleum Technology*, OnePetro, v. 55, n. 06, p. 70–72, 2003.

MATOS, M. C. de; OSORIO, P. L.; JOHANN, P. R. Unsupervised seismic facies analysis using wavelet transform and self-organizing maps. *Geophysics*, Society of Exploration Geophysicists, v. 72, n. 1, p. P9–P21, 2007.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, Springer, v. 5, n. 4, p. 115–133, 1943.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61.

MOON, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine*, IEEE, v. 13, n. 6, p. 47–60, 1996.

NERY, G. Perfilagem geofísica, notas de aulas. *UFBA*, p. 50, 2004.

NEYAMADPOUR, A.; TAIB, S.; ABDULLAH, W. W. Using artificial neural networks to invert 2d dc resistivity imaging data for high resistivity contrast regions: A matlab application. *Computers & Geosciences*, Elsevier, v. 35, n. 11, p. 2268–2274, 2009.

NIKNAM, T.; AMIRI, B. An efficient hybrid approach based on pso, aco and k-means for cluster analysis. *Applied soft computing*, Elsevier, v. 10, n. 1, p. 183–197, 2010.

OLIVEIRA, M. L. L. D. Reconhecimento de eletrofácies em reservatórios turbidíticos da formação carapebus no parque das baleias, bacia de campos. 2019.

PAKHIRA, M. K. An efficient distributed data clustering algorithm. *International Journal of Recent Trends in Engineering*, v. 3, n. 1, p. 81–89, 2010.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PÉREZ, F.; GRANGER, B. E. IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, IEEE Computer Society, v. 9, n. 3, p. 21–29, maio 2007. ISSN 1521-9615. Disponível em: <https://ipython.org>.

PETERS, J. Machine learning for motor skills in robotics. *KI-Künstliche Intelligenz*, v. 2008, n. 4, p. 41–43, 2008.

REYNOLDS, D. A. Gaussian mixture models. *Encyclopedia of biometrics*, Berlin, Springer, v. 741, p. 659–663, 2009.

RIDER, M.; KENNEDY, M. The geological interpretation of well logs: Rider-french consulting ltd. *Scotland, 280p*, 2002.

ROBERTS, S. J. Parametric and non-parametric unsupervised cluster analysis. *Pattern Recognition*, Elsevier, v. 30, n. 2, p. 261–272, 1997.

RUVINI, J.-D.; DONY, C. Ape: learning user's habits to automate repetitive tasks. In: *Proceedings of the 5th international conference on Intelligent user interfaces*. [S.l.: s.n.], 2000. p. 229–232.

SANTOS, F. V. *Uso de Algorítmos de Classificação para determinação de Eletrofácies em Poços da Bacia de Campos*. 64–75 p. Dissertação (Mestrado), 2016.

SCHRIDER, D. R.; KERN, A. D. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, Elsevier, v. 34, n. 4, p. 301–312, 2018.

SCULLEY, D. Web-scale k-means clustering. In: *Proceedings of the 19th international conference on World wide web*. [S.l.: s.n.], 2010. p. 1177–1178.

SERRA, O. Fundamentals of well-log interpretation. Elsevier Science Pub. Co., Inc., New York, NY, 1983.

SERRA, O.; SERRA, L. Well logging. data acquisitions and applications. 2004.

SIDDIQUI, M. K. et al. Correlation between temperature and covid-19 (suspected, confirmed and death) cases based on machine learning analysis. *J Pure Appl Microbiol*, v. 14, n. suppl 1, p. 1017–1024, 2020.

SOUZA, O. G. D. *Stratigraphie séquentielle et modélisation probabiliste des réservoirs d'un cône sous-marin profond (champ de Namorado, Brésil): intégration des données géologiques et géophysiques*. Tese (Doutorado) — Atelier national de reproduction des thèses, 1997.

STEVANATO, A. C. R. e. S. Análise Petrofísica de Reservatórios. 2011.

SUBHAKAR, D.; CHANDRASEKHAR, E. Reservoir characterization using multifractal detrended fluctuation analysis of geophysical well-log data. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 445, p. 57–65, 2016.

TEAM, T. pandas development. *pandas-dev/pandas: Pandas*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020.

WHANG, J. J.; DHILLON, I. S.; GLEICH, D. F. Non-exhaustive, overlapping k-means. In: SIAM. *Proceedings of the 2015 SIAM International Conference on Data Mining*. [S.l.], 2015. p. 936–944.

WINTER, W. R.; JAHNERT, R. J.; FRANÇA, A. B. Bacia de Campos. *B. Geoci. Petrobras, Rio de Janeiro*, v. 15, n. 2, p. 511–529, 2007.

WU, C.-J. et al. Machine learning at facebook: Understanding inference at the edge. In: IEEE. *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. [S.l.], 2019. p. 331–344.

XIE, J. et al. A survey of machine learning techniques applied to software defined networking (sdn): Research issues and challenges. *IEEE Communications Surveys & Tutorials*, IEEE, v. 21, n. 1, p. 393–430, 2018.

ZHANG, D.; YUNTIAN, C.; JIN, M. Synthetic well logs generation via recurrent neural networks. *Petroleum Exploration and Development*, Elsevier, v. 45, n. 4, p. 629–639, 2018.

ZHUANG, Y. et al. Adaptive key frame extraction using unsupervised clustering. In: IEEE. *Proceedings 1998 international conference on image processing. icip98 (cat. no. 98cb36269)*. [S.l.], 1998. v. 1, p. 866–870.