UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Rafael Andrello Rubo

Agrupamento do tipo de porosidade em rochas carbonáticas a partir de imagens segmentadas de lâminas petrográficas delgadas

São Carlos

2022

Rafael Andrello Rubo

Agrupamento do tipo de porosidade em rochas carbonáticas a partir de imagens segmentadas de lâminas petrográficas delgadas

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Afonso Paiva Neto

São Carlos

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

S856m

Rubo, Rafael Andrello
Agrupamento do tipo de porosidade em rochas carbonáticas a partir de imagens segmentadas de lâminas petrográficas delgadas / Rafael Andrello Rubo ; orientador Afonso Paiva Neto. – São Carlos, 2022.
47 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Ciências de Dados) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Ciência de Dados. 2. Petrografia. 3. Porosidade. 4. Óleo e Gás. I. Neto, Afonso Paiva, orient. II. Título.

Rafael Andrello Rubo

Agrupamento do tipo de porosidade em rochas carbonáticas a partir de imagens segmentadas de lâminas petrográficas delgadas

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Data de defesa: 22 de janeiro de 2022

Comissão Julgadora:

Prof. Dr. Afonso Paiva Neto
Orientador

Jadson Jose Monteiro Oliveira
Especialista

São Carlos
2022

RESUMO

RUBO, R. A. **Agrupamento do tipo de porosidade em rochas carbonáticas a partir de imagens segmentadas de lâminas petrográficas delgadas**. 2022. 47p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

A caracterização da porosidade em rochas reservatório de óleo e gás é essencial para o direcionamento da atividade exploratória e para o planejamento e desenvolvimento da produção. A análise petrográfica através de microscopia óptica permite detalhar o tipo de porosidade, o que auxilia os geocientistas a compreenderem e teorizarem a respeito da deposição e diagênese das bacias sedimentares, viabilizando a definição de parâmetros para modelos computacionais estáticos e dinâmicos. Neste trabalho, imagens de lâminas petrográficas delgadas adquiridas em microscópios de luz transmitida foram segmentadas de forma a destacar os poros de rochas carbonáticas, um dos principais litotipos em reservatórios. A partir dos poros segmentados, foram extraídos sete atributos geométricos: 1. área, 2. perímetro, 3. comprimento e 4. largura do retângulo de seleção, 5. comprimento dos eixos principal e 6. secundário da melhor elipse ajustável, e o 7. ângulo entre este eixo principal e uma linha paralela ao eixo x da imagem. Com estes atributos, foram calculados outros quatro atributos adimensionais: 1. diâmetro equivalente, 2. circularidade, 3. elongação e 4. retangularidade. Com estes atributos, um *dataset* de onze dimensões foi consolidado. Com este *dataset*, foi realizado um agrupamento utilizando o algoritmo K-Means em que todas as dimensões foram utilizadas. Um novo agrupamento foi realizado somente com os quatro atributos adimensionais calculados. Posteriormente, duas técnicas de redução de dimensionalidade foram aplicadas e testadas: *Principal Component Analysis* – PCA – e *t-distributed stochastic neighbor embedding* - t-SNE. Os componentes obtidos pelos dois métodos foram utilizados para dois novos agrupamentos, um para cada tipo de técnica utilizada. Os quatro agrupamentos realizados são então analisados em relação à anotação original referente ao tipo de porosidade. Através da análise, foi possível comparar os resultados obtidos e concluir que o agrupamento utilizando K-Means após a aplicação da t-SNE apresentou um agrupamento com maior correspondência em relação à anotação original dos dados, além da simplificação na visualização dos dados e redução do custo computacional.

Palavras-chave: porosidade, rochas carbonáticas, K-Means, PCA, t-SNE

ABSTRACT

RUBO, R. A. **Porosity type clustering in carbonate rocks from petrographic thin section segmented images**. 2022. 47p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

Porosity characterization of oil and gas reservoir rocks is essential for guiding exploration and for planning and development of production. Petrographic analysis through optical microscopy allows describing the type of porosity, which helps geoscientists to understand and theorize about the deposition and diagenesis of sedimentary basins, enabling the definition of parameters for static and dynamic computational models. In this work, images of petrographic thin sections acquired in transmitted light microscopes were segmented in order to highlight the pores of carbonate rocks, one of the main lithotypes in reservoirs. From the segmented pores, seven geometric features were extracted: 1. area, 2. perimeter, 3. length and 4. width of the bounding box, 5. length of the main and 6. secondary axes of the best adjustable ellipse, and the 7. angle between this main axis and a line parallel to the x axis of the image. With these features, four other dimensionless attributes were calculated: 1. equivalent diameter, 2. circularity, 3. elongation and 4. rectangularity. With these features, an eleven-dimensional dataset was consolidated. With this dataset, clustering was performed using the K-Means algorithm in which all dimensions were used. A new clustering was performed with only the four calculated dimensionless attributes. Subsequently, two dimensionality reduction techniques were applied and tested: Principal Component Analysis – PCA – e t-distributed stochastic neighbor embedding - t-SNE. The components obtained by the two methods were used for two new clustering, one for each dimensionality reduction technique used. The four obtained clustering are then analyzed in relation to the original labels regarding the type of porosity. Through the analysis, it was possible to compare the results obtained and to conclude that clustering using K-Means after the application of t-SNE showed greater correspondence in relation to the original labels of the data, in addition to the simplifying visualization and reducing computational cost.

Keywords: porosity, carbonatic rocks, K-Means, PCA, t-SNE

SUMÁRIO

1 INTRODUÇÃO

A caracterização de reservatórios de óleo e gás tem como um de seus principais objetivos a descrição da distribuição espacial de parâmetros petrofísicos: porosidade e permeabilidade (LUCIA, 1995). A utilização de variáveis mais próximas da realidade nos modelos de reservatório orienta uma exploração com maior sucesso geológico e viabiliza a otimização do desenvolvimento da produção.

A definição dos melhores parâmetros a serem utilizados nos modelos está vinculada a integração de estudos de rocha, perfis e sísmicos. A partir de estudos de rocha, são realizados upscales para perfis e, finalmente, para sísmica, extrapolando e correlacionando os dados às respostas dos métodos indiretos.

Além da descrição macroscópica de amostras laterais auxiliada por lupa petrográfica, são realizadas medidas petrofísicas com o auxílio de porosímetro e permeâmetro. Estas medidas são complementadas por estudos microscópicos de lâminas petrográficas delgadas, embasados nas propriedades ópticas dos minerais.

A análise microscópica possibilita a obtenção de dados muito particulares referentes à textura, processos diagenéticos (cimentação, substituição, compressão, percolação de fluidos, entre outros), sistema poroso e ocorrência de argilominerais. Entretanto, seus resultados são, em sua maioria, qualitativos e dependem da experiência do petrógrafo, além da área de interesse do projeto em questão. Eventualmente, podem ser aplicados métodos estatísticos, como a descrição por pontos da lâmina em intervalos regulares de espaçamento. Porém, estas análises costumam ser demoradas e, apesar de apresentarem um componente quantitativo, ainda dependem em grande parte de fatores qualitativos.

Muitos trabalhos abordam a aplicação de métodos quantitativos alternativos na descrição de lâminas petrográficas através de análise espacial auxiliada por computador com a aplicação de processamento sequencial tradicional (EHRLICH; DAVIES, 1984; TOMUTSA; BRINKMEYER; RAIBLE, 1984; ANJOS et al., 1995; ANSELMETTI; LUTHI; EBERLI, 1998, entre outros). Estas técnicas têm se mostrado muito úteis, principalmente na identificação e quantificação de atributos muito distintos em lâminas, como a porosidade petrográfica total. Entretanto, eles são insuficientes para a aplicação de funções mais complexas, como a segmentação e classificação de porosidade.

Com o avanço de técnicas de análise espacial e do aumento na capacidade de processamento dos computadores, além da aplicação do processamento em paralelo utilizando múltiplos núcleos (GPU, *Graphics Processing Unit*), modelos baseados na aplicação de redes neurais artificiais estão tornando-se mais aplicáveis em análise de imagens (GURGEL, 2014; ŁADNIAK; MŁYNARCZUK, 2015; BUDENNYY et al., 2017; IZADI; SADRI;

BAYATI, 2017).

Os algoritmos *Random Forest* e *K-means* são aplicados neste trabalho, de forma a auxiliar na descrição de lâminas petrográficas, tornando-as mais quantitativas, confiáveis e rápidas. Especificamente, são segmentados os poros das imagens de lâminas petrográficas de rochas carbonáticas. Na sequência, os poros são agrupados conforme suas características geométricas.

Esta sequência de segmentação e agrupamento em imagens de lâminas petrográficas já foi aplicada de forma similar e integrada, utilizando métodos simples e consolidados, porém que apresentam limitações relevantes. Novos métodos mais robustos foram propostos e aplicados de forma desconectada para ambas as técnicas. A evolução da aplicação dos métodos na literatura científica especializada é apresentada na próxima seção. Neste trabalho, os métodos mais recentes são aplicados, tanto para segmentação quanto para agrupamento, de forma integrada, viabilizando a predição de parâmetros de porosidade mais acurados.

2 REVISÃO BIBLIOGRÁFICA

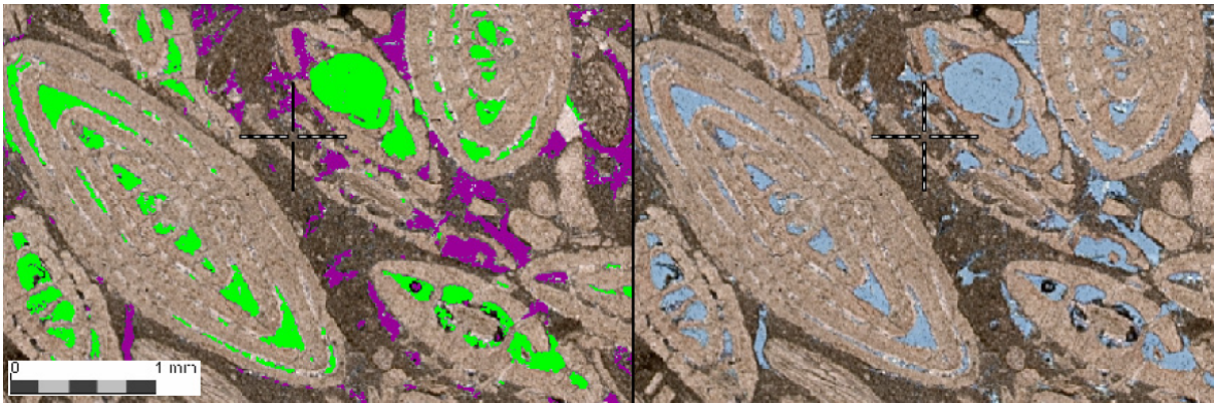
A análise de imagens de lâminas petrográficas delgadas é proposta na literatura científica há bastante tempo. Anjos et al. (1995) já apresentaram uma caracterização do sistema poroso de rochas-reservatório utilizando técnicas de segmentação por limiar (*threshold*) seguida por extração de atributos geométricos. Neste método de segmentação, o usuário define os limites inferior e superior do histograma dos canais das imagens e seleciona somente os pixels de interesse. Este método de segmentação pode ser limitado do ponto de vista operacional, pois depende do ajuste manual do usuário.

Roduit (2007) expande a segmentação de imagens de lâminas petrográficas por limiar para a utilização do método *watershed* (BEUCHER; LANTUEJOUL, 1979), no qual o gradiente da imagem atua como uma superfície topográfica com a intensidade de cor de cada *pixel* representando o seu peso. Os equivalentes a fontes de água são então posicionados em mínimos locais e "inundam" uma feição até atingirem o divisor de águas, delimitando bordas de grãos, por exemplo. Neste trabalho, o autor vai além da segmentação de poros nas imagens e propõe uma classificação do tipo de porosidade utilizando contexto. Esta classificação por contexto considera o ambiente ao redor do poro para classifica-lo. Por exemplo, janelas nos histogramas são definidas e, caso a feição de interesse estiver em contato com outras feições que atendam esta restrição, ela é classificada como sendo de uma categoria ou de outra. Este é um método muito simples e, por vezes, pode não ser suficiente para diferenciar algumas das categorias de interesse.

A popularização de algoritmos de classificação baseados em aprendizado de máquina permitiu uma grande quantidade de pesquisas e propostas de fluxos analíticos que viabilizaram uma análise de imagem muito mais bem embasada e aplicada. O trabalho de Roduit (2007), apesar de não utilizar estes algoritmos, apresenta o diferencial de aplicar uma segmentação seguida por classificação na imagem de forma integrada. A Figura 1 evidencia uma imagem de lâmina petrográfica que teve, inicialmente, seus poros segmentados e, então, classificados quanto às suas características de contexto.

O primeiro trabalho que apresenta uma ferramenta de segmentação baseada em algoritmos de aprendizado de máquina foi publicado por Schindelin et al. (2012). Através de uma integração com o trabalho de Hall et al. (2009), Schindelin et al. (2012) apresentam a utilização de filtros convolucionais discretos como forma de ampliar o número de atributos de um *pixel*. Estes *pixels*, que passam a ter sua dimensionalidade aumentada, são classificados utilizando diferentes algoritmos de aprendizado de máquina (*support vector machine*, *random forest*, redes neurais artificiais, entre outros) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Figura 1 – Imagem à esquerda: imagem segmentada e classificada, com os poros classificados em intrapartícula (verde) e interpartícula (roxo) utilizando contexto. Imagem à direita: imagem original, com os poros na coloração azul.



Fonte: Roduit (2007)

Apesar do trabalho de Schindelin et al. (2012), muitos autores permaneceram utilizando métodos como limiar e *watershed* para segmentação de imagens de lâminas petrográficas, como Budennyy et al. (2017). Entretanto, da mesma forma como observado fora do contexto das geociências, segmentação assistida por algoritmos de aprendizado de máquina se consolidou por apresentar resultados de acurácia melhores, demostrado por trabalhos como os de Izadi, Sadri e Bayati (2017) e Rubo et al. (2019).

Para classificação de imagens de lâminas petrográficas, diversos trabalhos foram realizados, considerando diferentes objetivos: desde busca por similaridade entre imagens utilizando aprendizado não supervisionado (ŁADNIAK; MŁYNARCZUK, 2015), até utilização de modelos pré-treinados de redes neurais convolucionais (LIMA et al., 2020; SU et al., 2020) para classificação de litologias. A Figura 2 apresenta um resumo dos métodos utilizados pelos diferentes trabalhos apresentados.

Figura 2 – Resumo dos métodos utilizados pelos trabalhos apresentados.

	segmentação por threshold	segmentação por watershed	segmentação por machine learning	classificação por contexto	classificação por machine learning	integração segmentação-classificação	enfoque
Anjos et al., 1995	X						microscopia óptica de rochas siliciclásticas e carbonáticas
Roduit, 2007	X	X		X		X	microscopia óptica de rochas carbonáticas
Schindelin et al., 2012			X				microscopia eletrônica de varredura em amostras biológicas
Ładniak & Młynarczuk, 2015					X		microscopia óptica de rochas siliciclásticas, carbonáticas e ígneas
Budennyy et al., 2017		X					microscopia óptica de rochas siliciclásticas e carbonáticas
Izadi et al., 2017			X				microscopia óptica de rochas ígneas
Rubo et al., 2019			X				microscopia óptica de rochas carbonáticas
Lima et al., 2020					X		microscopia óptica de rochas siliciclásticas
Su et al., 2020					X		microscopia óptica de rochas ígneas, sedimentares e metamórficas
Este trabalho			X		X	X	microscopia óptica de rochas carbonáticas

Fonte: Elaborada pelo autor

3 CONCEITOS BÁSICOS

Neste tópico é realizada uma revisão de temas essenciais para melhor compreensão do método. São eles: 1. petrografia óptica; 2. porosidade em rochas carbonáticas; 3. convoluções; 4. *Random Forest*; 5. *K-means*; e 6. redução de dimensionalidade.

Aspectos Geológicos

- Petrografia óptica

A petrografia óptica estuda a interação dos minerais com a luz e as propriedades dos minerais em lâmina delgada. As propriedades ópticas dos minerais variam de acordo com sua composição e estrutura cristalina (NESSE, 2004). Um princípio fundamental da mineralogia óptica é que a maioria dos minerais transmitem luz se cortados em fatias muito finas, inclusive os mais escuros e outros que parecem opacos em amostras de mão (FLEISCHER; WILCOX; MARZKO, 1984; PERKINS; HENKE, 2004). Para o estudo destes minerais, é utilizado o microscópio óptico de luz transmitida.

Os minerais de brilho metálico, e alguns outros, são chamados minerais opacos, pois não transmitem luz mesmo que cortados em fatias finas. Para o estudo destes minerais é utilizado o microscópio óptico de luz refletida. Muitas vezes, o mesmo microscópio petrográfico tem as opções de luz transmitida e refletida, podendo diferentes minerais serem analisados no mesmo equipamento.

As lâminas de amostras de rocha derivadas da perfuração de poços de petróleo são confeccionadas a partir de parte das amostras laterais (chamada de “tijolinho”), em que é realizado processo de limpeza e extração de óleo e sais, auxiliados por solventes (tolueno para óleo e metanol para sais). Este processo tem como objetivo eliminar fluidos remanescentes do sistema poroso da amostra de tal forma a viabilizar o estudo do mesmo através das propriedades ópticas da rocha, comumente realizado para o estudo de rochas sedimentares com ênfase na exploração de óleo e gás.

Após esta etapa, a amostra é impregnada por resina e catalisador epóxi, juntamente com corante azul de ceres que irá representar o sistema poroso na lâmina petrográfica. A cor azul de ceres é escolhida uma vez os minerais observados na natureza não apresentam esta coloração quando submetidos à luz branca, distinguindo o sistema poroso da lâmina de qualquer mineral.

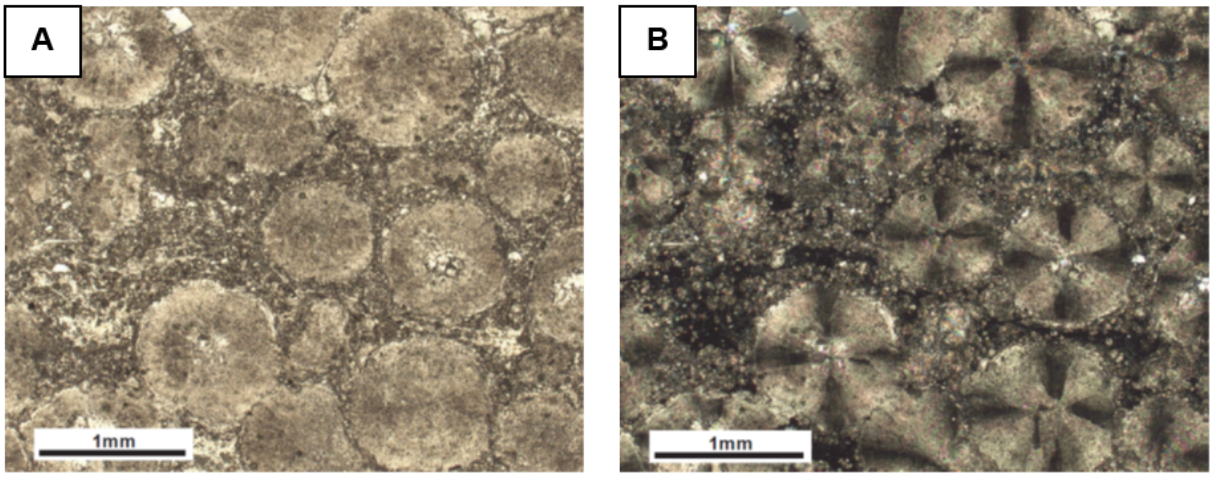
A amostra impregnada é, então, colada à lâmina de vidro, cortada e desbastada até

atingir a espessura de 30 a 35 µm. A lâmina é analisada em microscópio petrográfico de luz transmitida (para estudo das propriedades ópticas de minerais translúcidos) e refletida (para estudo das propriedades ópticas de minerais opacos).

No caso de rochas carbonáticas, a mineralogia costuma ser muito simples, constituindo-se principalmente de calcita (carbonato de cálcio - $CaCO_3$) e dolomita (carbonato de cálcio e magnésio - $CaMg(CO_3)_2$). Podem ocorrer outros minerais de forma subordinada, tais como grãos siliciclásticos (quartzo, muscovita, feldspatos, entre outros), fragmentos de rochas vulcânicas e diferentes tipos de argilas. Devido à esta simplicidade mineralógica, o estudo das texturas em rochas carbonáticas é fundamental para a interpretação de ambientes deposicionais e evolução diagenética, assim como a avaliação das características do sistema poroso.

Em feixes de luz não polarizados, as ondas de luz vibram em muitas direções diferentes. Um polarizador filtra o feixe de luz para que todas as ondas vibrem em uma única direção (polarizadores paralelos). Com o auxílio de um segundo polarizador, algumas das texturas analisadas no estudo de petrografia óptica somente são evidenciadas (polarizadores cruzados). Existem diversos outros filtros e lentes que podem ser utilizados na atividade de descrição microscópica de lâminas delgadas, que serão utilizados de forma subordinada na pesquisa (Figura 3).

Figura 3 – Fotomicrografias obtidas com microscópio óptico petrográfico. Esferulitito com argila proveniente da Fm. Barra Velha, Cretáceo da Bacia de Santos, com polarizadores paralelos (A) e polarizadores cruzados (B). Notar a textura fibro-radial dos esferulitos evidenciadas em luz polarizada.

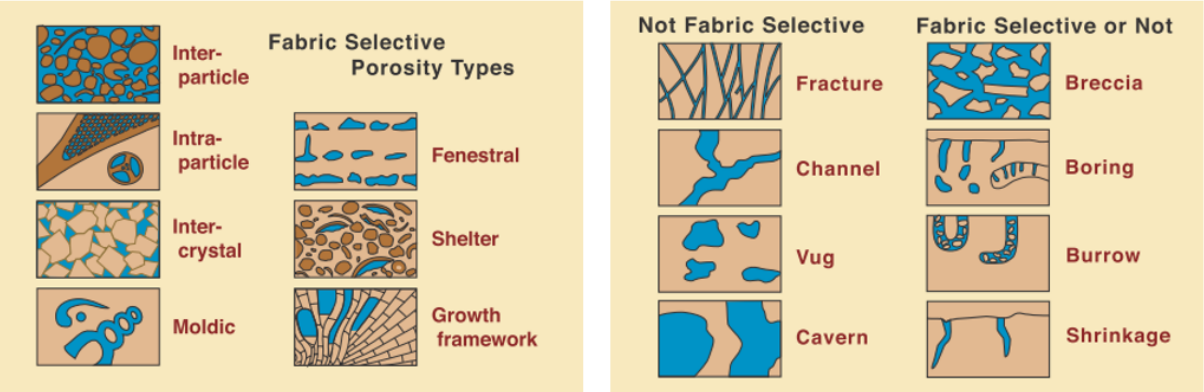


Fonte: adaptado de Terra et al. (2010)

- Porosidade em rochas carbonáticas

O estudo de sistemas porosos em carbonatos é mais complexo do que em siliciclásticos devido às características relacionadas à origem biológica e à alta reatividade química. Muitas classificações de porosidade para rochas carbonáticas foram propostas (JODRY, 1972; LUCIA, 1983; EHRLICH et al., 1991; LUCIA, 1995, entre outras), mas a classificação de Choquette e Pray (1970) consolidou-se como a mais utilizada pela academia e indústria (Figura 4).

Figura 4 – Sistema de classificação de porosidade de rochas carbonáticas proposto por Choquette e Pray (1970)



Fonte: Choquette e Pray (1970), adaptado por Scholle e Ulmer-Scholle (2003)

Neste sistema de classificação de porosidade, a mesma é subdividida em seletiva pela trama, não seletiva pela trama, e que pode ou não ser seletiva pela trama. Esta subdivisão está relacionada com a perspectiva de a porosidade ser ou não controlada por fronteiras primárias delimitadas por grãos, cristais ou outras estruturas físicas na rocha.

Aspectos Computacionais e Numéricos

- Convoluções

A operação de convolução em imagens permite o acréscimo de informação à um pixel referente aos seus pixels vizinhos. Tem como objetivo extrair feições das imagens a partir de operações matemáticas entre funções, representadas pela Equação:

$$\mathrm{g(x)=f(x)*h(x)=}\int_{-\infty}^{\infty} f(s)h(x-s)\,ds.$$

A operação de convolução também pode ser representada através de matrizes. A matriz convolvida resultante é a soma da multiplicação de uma matriz maior por uma matriz menor, também chamada de filtro convolucional (DUMOULIN; VISIN, 2016). Cada canal de cor de uma imagem colorida pode ser representado por uma matriz. Estas matrizes seriam equivalentes à matriz maior de uma operação de convolução. Há um grande número de filtros convolucionais discretos (a matriz menor) que podem ser aplicados à uma imagem, cada um evidenciando diferentes características, tais como as bordas de feições (KHAN et al., 2018).

Uma vez que diferentes filtros convolucionais tenham sido aplicados à uma imagem, cada pixel passa a conter muito mais informação do que somente os três canais de um espaço de cor. Eles passam a expressar sua relação com os pixels vizinhos através de novos atributos. Estes atributos convolucionais representam nova informação adicionada ao dado original referente à sua relação espacial. Desta forma, uma classificação pixel a pixel pode gerar uma imagem segmentada com base em muito mais dimensões e considerando o contexto espacial do dado.

Entretanto, para processar esta grande quantidade de informação adicionada juntamente com o dado original e obter a melhor classificação, algoritmos de machine learning de classificação são utilizados.

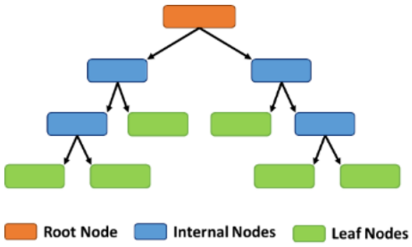
- *Random Forest*

Estes são algoritmos baseados em processos decisórios, chamados de classificadores de conjunto (*ensemble*). Isto porque são, na realidade, uma coleção de classificadores muito mais simples: árvores de decisão. Árvores de decisão são sequências de premissas estruturadas de uma maneira organizada (BREIMAN et al., 1984). Se o dado de entrada estiver em concordância com a primeira premissa (nó raíz), ele segue para a próxima premissa (nó interno), até o momento em que passa a ser classificado (folhas). Se o dado não estiver em concordância com uma das premissas pelo caminho, ele segue por uma ramificação diferente na árvore, alcançando uma posição em que é rotulado como uma classe diferente do que a da ramificação alternativa (Figura 5).

A estrutura nas quais as árvores de decisão são organizadas depende da impureza de uma premissa. Os fatores de impureza são especificados para cada premissa e são chamados impureza Gini. São uma medida do quanto decisiva uma premissa é para a classificação de um dado de entrada, calculados de acordo com a Equação:

$$\text{Impureza Gini} = 1 - (\text{probabilidade de “sim”})^2 - (\text{probabilidade de “não”})^2$$

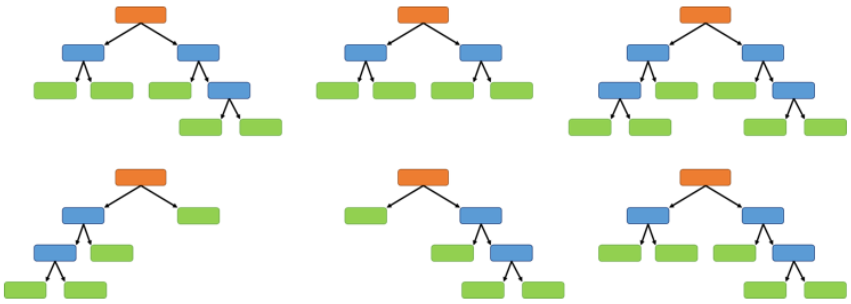
Figura 5 – Exemplo de árvore de decisão



Fonte: Elaborada pelo autor

Para analisar os dados de entrada, os algoritmos Random Forest criam um grande número de árvores de decisão com estruturas aleatórias (Figura 6). Eles também selecionam parte dos dados aleatoriamente para servirem como entrada de cada árvore criada, em um processo chamado de bootstrap aggregation, ou bagging (BREIMAN, 1996). A classe que recebe mais votos das árvores com estruturas aleatórias é a classe que será selecionada pelo algoritmo Random Forest (BREIMAN, 2001). Segundo Hastie, Tibshirani e Friedman (2009), algoritmos *Random Forest* estabilizam o erro de classificação com cerca de 200 árvores, ao menos para os dados de treinamento que eles utilizaram.

Figura 6 – Exemplo de árvores de decisão com diferentes estruturas criadas aleatoriamente



Fonte: Elaborada pelo autor

- *K-Means*

O K-Means é um algoritmo de agrupamento de dados. Não requer que os dados estejam rotulados para realizar o agrupamento – aprendizado não-supervisionado – pois utiliza similaridades entre as observações para definir os grupos. A partir de um centroide representativo para cada grupo e que tem sua posição atualizada a cada iteração, o algoritmo dá um rótulo para todas as observações com base nas distâncias entre elas e estes centroides ((STEINHAUS, 1967; MACQUEEN, 1967)).

Inicialmente, o número de centroides a ser utilizado é definido, ou seja, o número de grupos que o algoritmo irá criar – parâmetro k. Para isso, é comum a utilização da técnica chamada de curva do cotovelo, em que a variância dos dados é testada em relação ao número de grupos. O número ideal é obtido a partir do momento em que o aumento do número de grupos não implica em aumento significativo de ganho. Os centroides são distribuídos aleatoriamente e, então, é calculada a distância euclidiana entre cada observação e cada um dos centroides. Um rótulo referente ao centroide mais próximo é atribuído para cada observação. Após cada iteração, a média dos valores de um grupo é calculada e os centroides têm sua posição atualizada. As distâncias entre as observações e os centroides são novamente calculadas e novos rótulos são atribuídos. Os centroides têm sua posição novamente atualizadas, até que permaneçam estacionários e não seja mais necessário realizar iterações ((LLOYD, 1982)).

- Redução de Dimensionalidade

A redução de dimensionalidade de um conjunto de dados é importante pois elimina redundâncias e otimiza os algoritmos de classificação, regressão e agrupamento. Consiste na aplicação de diferentes técnicas que reduzem o número de atributos, seja através da seleção dos atributos que melhor representam o conjunto de dados considerando uma característica específica, ou da criação de novos espaços com número de atributos reduzidos em relação ao espaço original (MAATEN; HERIK; POSTMA, 2007; CUNNINGHAM; GHAHRAMANI, 2015). Os novos atributos criados para este espaço de dimensão reduzida geralmente são chamados de componentes. Estas componentes pretendem representar o dado original em um espaço de dimensão menor, eliminando somente redundâncias, ou dados que representam baixa contribuição na variabilidade dos atributos de interesse.

Neste trabalho, o algoritmo de agrupamento K-Means é utilizado tanto no *dataset* original quanto em um dataset com dimensões reduzidas. São utilizadas duas técnicas de redução de dimensionalidade: 1. *Principal Component Analysis* – PCA; e 2. *t-distributed stochastic neighbor embedding* - t-SNE.

O PCA, ou Análise dos Componentes Principais, é uma das técnicas mais utilizada para redução de dimensionalidade. Inicialmente, é realizada uma padronização dos dados, redimensionando-os de tal forma que todos os atributos tenham média 0 e desvio padrão 1. Após essa padronização, a correlação entre os atributos é avaliada através de uma matriz de covariância. Os autovalores e autovetores são calculados a partir da matriz de covariância de forma a obter-se os componentes principais do conjunto de dados. Estes componentes principais são os novos atributos do dataset de dimensionalidade reduzida ((KAMBHATLA; LEEN, 1997; DING; HE, 2004)).

A técnica t-SNE para redução de dimensionalidade é mais recente. Inicialmente, é calculada a distribuição de probabilidades com base em pares de observações do dataset. Desta forma, observações similares apresentam probabilidades mais altas – similaridade originalmente calculada pela distância euclidiana entre os pontos. Então, a t-SNE define uma distribuição de probabilidade similar entre os pontos de um espaço de menor dimensão e minimiza a distância entre as duas distribuições. Realiza este ajuste até obter um espaço de menor dimensão representativo do espaço original ((MAATEN; HINTON, 2008)).

4 METODOLOGIA

A criação de um *dataset* de imagens de microscopia óptica de luz transmitida é o primeiro passo para obtenção da classificação do tipo de porosidade baseada em imagens segmentadas. Para essa atividade, utiliza-se a base de dados públicos de bacias terrestres brasileiras disponibilizada pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis – ANP, além de diversas publicações científicas públicas. São selecionadas imagens de lâminas petrográficas delgadas correspondentes a rochas carbonáticas de diferentes contextos.

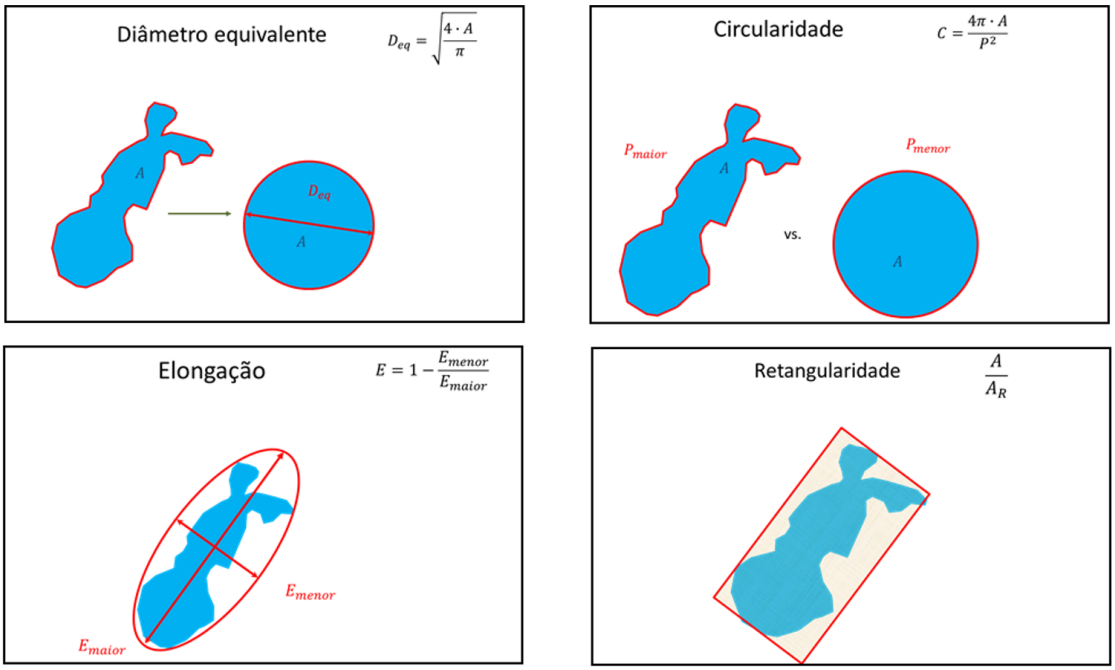
Um modelo de segmentação binária é treinado com o objetivo de destacar os poros das imagens. São aplicados filtros convolucionais discretos seguidos pela classificação *pixel* a *pixel* utilizando algoritmo *Random Forest*, conforme proposto por Rubo et al. (2019).

Após a obtenção das imagens segmentadas, alguns dos poros podem permanecer conectados por gargantas de poros. Para individualizá-los, as gargantas de poros são eliminadas através de operadores morfológicos simples, como a operação de erosão – *thining*.

Uma vez obtidos os poros individualizados de uma imagem, são extraídos os seguintes atributos geométricos: área, perímetro, comprimento e largura do retângulo de seleção, comprimento dos eixos principal e secundário da melhor elipse ajustável, e o ângulo entre este eixo principal uma linha paralela ao eixo x da imagem. Com estes atributos, são calculados outros quatro atributos adimensionais: diâmetro equivalente, circularidade, elongação e retangularidade. A Figura 7 apresenta uma representação gráfica definindo estes atributos adimensionais, além das expressões utilizadas para o cálculo. Estes atributos dão origem a um *dataset* estruturado que é utilizado para realizar agrupamentos através da utilização do algoritmo K-means. O valor K ideal é obtido através do método do cotovelo e a validação do modelo é realizada a partir de *dataset* de teste anotado.

A Figura 8 apresenta um fluxograma resumido destas etapas e permite uma visualização simplificada dos métodos utilizados neste trabalho.

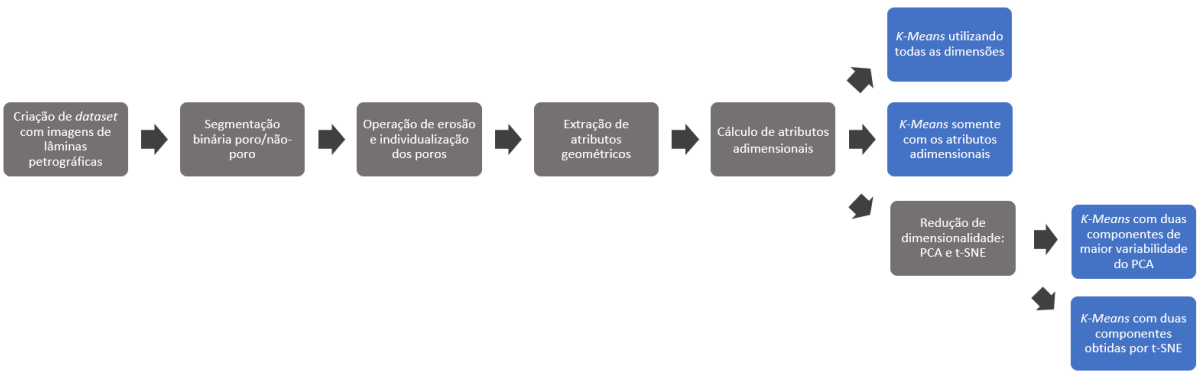
Figura 7 – Representação gráfica dos atributos adimensionais e as expressões utilizadas para os cálculos, onde A = área; P = perímetro; Emaior = comprimento do eixo principal da melhor elipse ajustável; Emenor = comprimento do eixo secundário da melhor elipse ajustável; e Ar = área do menor retângulo ajustável.



Fonte: Elaborada pelo autor

Figura 8 – Fluxograma simplificado dos métodos utilizados. Em azul, destaque para os agrupamentos realizados com o algoritmo K-Means

.

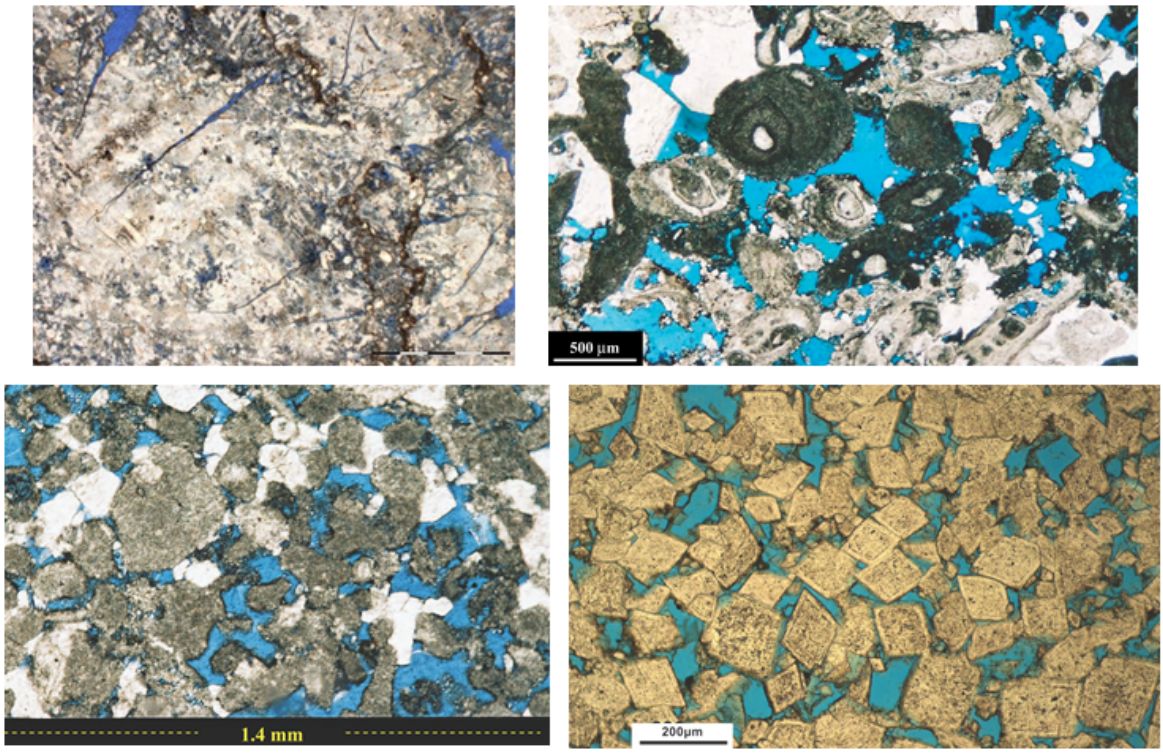


Fonte: Elaborada pelo autor

5 RESULTADOS

A primeira etapa da pesquisa foi a criação de um *dataset* de imagens de lâminas petrográficas de rochas carbonáticas apresentando diferentes tipos de porosidade. Foram obtidas 39 imagens contendo diferentes ocorrências de poros (Figura 9). Todas estas imagens são públicas, disponibilizadas para uso acadêmico pela ANP, além de diversas outras publicações em periódicos científicos (GREGG, 2001; TERRA et al., 2010; ZHAO, 2011; TONIETTO; POPE, 2016, entre outros).

Figura 9 – Exemplo de imagens de lâminas petrográficas de rochas carbonáticas apresentando diferentes tipos de porosidade, obtidas de publicações científicas.

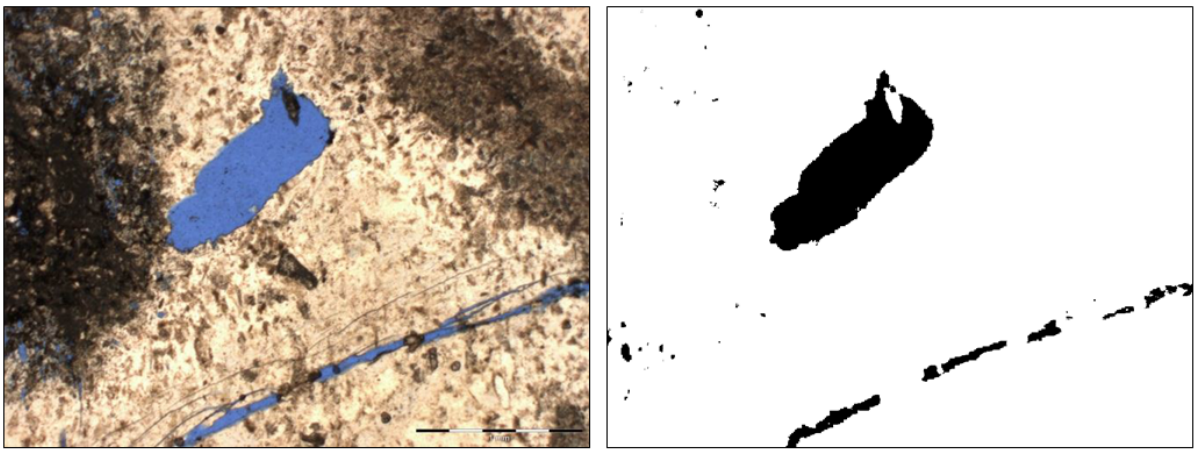


Fonte: Mosaico elaborado pelo autor, a partir de imagens provenientes de Zhao (2011), Lonoy (2016), e Terra et al. (2010)

.

Na sequência, elas foram segmentadas em duas classes: poro e não-poro. A segmentação foi realizada utilizando uma sequência de filtros convolucionais discretos seguidos de classificação *pixel* a *pixel* através de *Random Forest* com 200 árvores, conforme descrito por Rubo et al. (2019). A Figura 10 apresenta um exemplo de uma das imagens originais e de sua respectiva imagem binária, obtida após a segmentação.

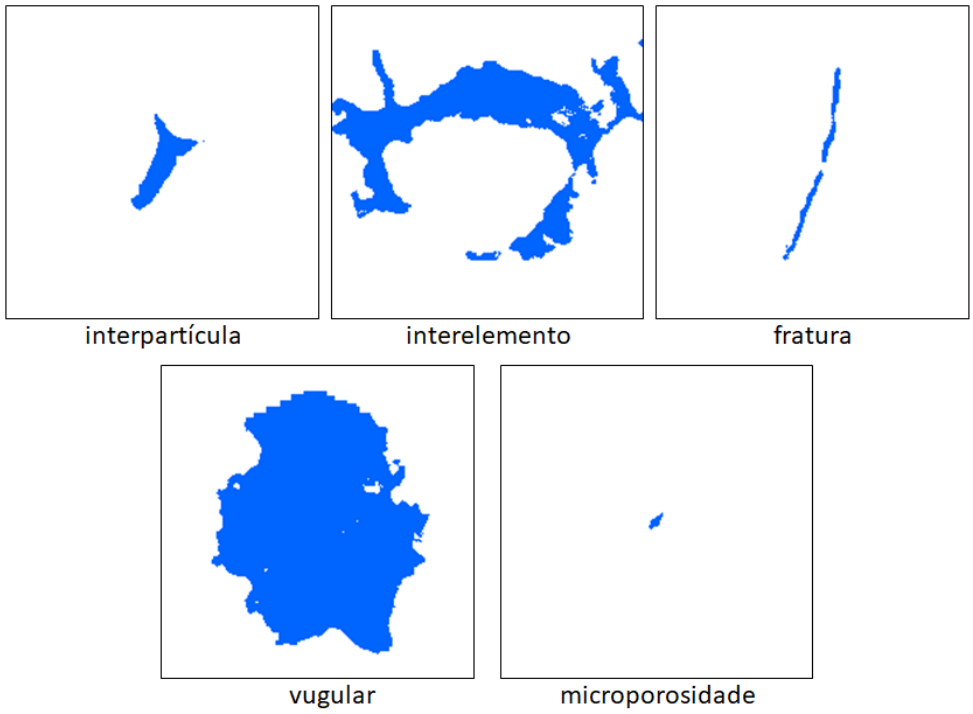
Figura 10 – Imagem de lâmina petrográfica original e sua segmentação binária, destacando a classe poros.



Fonte: Segmentação realizada pelo autor. Imagem original obtida por Zhao (2011)

Uma operação morfológica de erosão foi realizada, eliminando falsos positivos e individualizando poros distintos, uma vez que esta etapa elimina as gargantas de poros. Foram obtidos 50 poros individualizados, pertencentes a 5 classes de tipos de porosidade: fratura, interpartícula, interelemento, microporosidade, vugular. As classes estão balanceadas, cada categoria de tipo de porosidade apresenta 10 observações (Figura 11).

Figura 11 – Poros individualizados após a segmentação e erosão, representativos de cada uma das classes.



Fonte: Elaborada pelo autor

A partir deste *dataset* de poros individualizados, foram extraídos 7 atributos geométricos: área, perímetro, comprimento e largura do retângulo de seleção, comprimento dos eixos principal e secundário da melhor elipse ajustável, e o ângulo entre este eixo principal uma linha paralela ao eixo x da imagem. Estes parâmetros viabilizam o cálculo, detalhado na seção Conceitos Básicos, de outros 4 parâmetros: diâmetro equivalente, circularidade, elongação, e retangularidade. A Figura 12 apresenta os valores destas 11 dimensões obtidas para as 50 observações, assim como a classe correspondente de cada observação.

Figura 12 – Valores dos 11 atributos geométricos - dimensões - obtidos a partir poros individualizados para as 50 observações. Cada observação foi, ainda, rotulada em uma classe de tipo de porosidade.

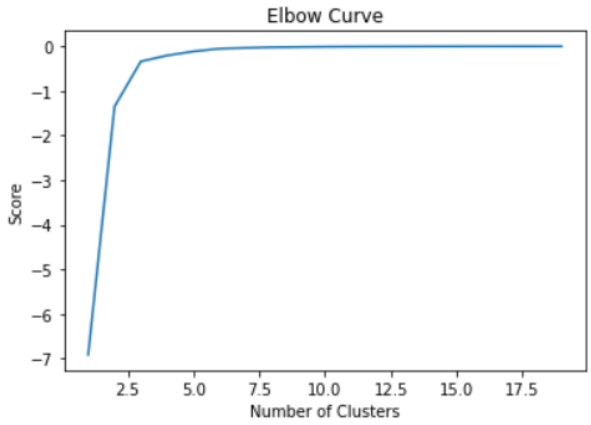
area	perim	width	height	major	minor	angle	eqdia	elongation	circularity	rectangularity	class
150.00	111.60	7.00	50.00	46.91	4.07	94.13	13.82	43.83	0.15	0.43	fratura
507.00	353.00	97.00	119.00	108.60	5.94	52.74	25.41	103.66	0.05	0.04	fratura
148.00	120.43	9.00	54.00	52.28	3.60	96.35	13.73	49.68	0.13	0.30	fratura
255.00	156.41	59.00	37.00	53.06	6.12	147.92	18.02	47.94	0.13	0.12	fratura
132.00	94.63	7.00	44.00	43.28	3.88	86.23	12.96	40.39	0.19	0.43	fratura
93.00	68.18	13.00	28.00	29.98	3.95	67.22	10.88	27.03	0.25	0.26	fratura
351.00	260.43	69.00	91.00	77.76	5.75	50.35	21.14	73.02	0.07	0.06	fratura
209.00	140.27	57.00	28.00	58.69	4.53	155.46	16.31	55.15	0.13	0.13	fratura
310.00	182.51	78.00	11.00	65.25	6.05	2.84	19.87	60.20	0.12	0.36	fratura
165.00	99.15	42.00	9.00	40.80	5.15	5.75	14.49	36.65	0.21	0.44	fratura
558.00	167.04	54.00	38.00	37.61	18.89	157.77	26.65	19.71	0.25	0.27	interpart
822.00	228.79	51.00	58.00	48.20	21.72	148.28	32.35	27.48	0.20	0.28	interpart
151.00	81.40	28.00	15.00	21.19	9.07	166.08	13.87	13.11	0.29	0.36	interpart
89.00	59.36	21.00	11.00	17.66	6.42	166.69	10.65	12.24	0.32	0.39	interpart
163.00	75.15	19.00	22.00	16.30	12.73	71.91	14.41	4.57	0.36	0.39	interpart
158.00	62.77	24.00	13.00	21.33	9.43	173.84	14.18	12.90	0.50	0.51	interpart
184.00	64.43	14.00	25.00	26.08	8.98	108.56	15.31	18.10	0.56	0.53	interpart
103.00	62.77	26.00	10.00	23.05	5.69	167.68	11.45	18.35	0.33	0.40	interpart
357.00	175.62	40.00	36.00	37.84	12.01	134.21	21.32	26.83	0.15	0.25	interpart
143.00	67.50	23.00	18.00	25.38	7.17	137.22	13.49	19.21	0.39	0.35	interpart
1139.00	273.46	75.00	56.00	51.45	28.19	20.81	38.08	24.26	0.19	0.27	interelem
6484.00	926.13	200.00	108.00	151.40	54.63	90.00	90.86	97.78	0.09	0.30	interelem
1947.00	471.53	117.00	83.00	82.09	30.20	10.89	49.79	52.89	0.11	0.20	interelem
613.00	244.25	84.00	37.00	59.00	13.23	158.82	27.94	46.77	0.13	0.20	interelem
1321.00	263.32	67.00	42.00	64.27	26.17	165.95	41.01	39.10	0.24	0.47	interelem
1324.00	374.56	50.00	70.00	48.38	34.84	92.48	41.06	14.54	0.12	0.38	interelem
554.00	233.18	60.00	47.00	43.26	16.31	33.81	26.56	27.96	0.13	0.20	interelem
747.00	247.56	52.00	75.00	59.47	15.99	123.30	30.84	44.48	0.15	0.19	interelem
677.00	207.18	46.00	47.00	38.14	22.60	58.92	29.36	16.54	0.20	0.31	interelem
448.00	175.82	42.00	29.00	32.44	17.59	10.35	23.88	15.85	0.18	0.37	interelem
40.00	24.97	9.00	7.00	9.45	5.39	161.80	7.14	5.05	0.81	0.63	micropor
49.00	34.63	7.00	12.00	11.83	5.28	66.30	7.90	7.55	0.51	0.58	micropor
34.00	21.56	7.00	7.00	8.28	5.23	50.75	6.58	4.05	0.92	0.69	micropor
21.00	18.14	6.00	5.00	5.91	4.52	28.01	5.17	2.39	0.80	0.70	micropor
26.00	19.31	8.00	4.00	8.13	4.07	10.18	5.75	5.06	0.88	0.81	micropor
40.00	24.39	7.00	8.00	8.78	5.80	136.20	7.14	3.98	0.85	0.71	micropor
21.00	16.14	5.00	6.00	6.36	4.20	133.96	5.17	3.16	1.01	0.70	micropor
30.00	22.97	7.00	8.00	8.05	4.74	127.57	6.18	4.31	0.71	0.54	micropor
24.00	17.31	6.00	5.00	6.56	4.66	156.71	5.53	2.91	1.01	0.80	micropor
28.00	19.90	7.00	5.00	7.24	4.93	179.19	5.97	3.31	0.89	0.80	micropor
10421.00	662.57	143.00	144.00	123.73	107.24	51.96	115.19	17.50	0.30	0.51	vug
4030.00	686.59	116.00	85.00	108.61	47.25	158.25	71.63	62.36	0.11	0.41	vug
1187.00	257.71	57.00	53.00	51.99	29.07	153.12	38.88	23.92	0.22	0.39	vug
15568.00	670.91	140.00	169.00	162.26	122.16	97.82	140.79	41.10	0.43	0.66	vug
6480.00	464.40	120.00	102.00	101.27	81.47	176.65	90.83	20.79	0.38	0.53	vug
8122.00	624.57	146.00	126.00	128.17	80.69	26.00	101.69	48.48	0.26	0.44	vug
1725.00	230.94	74.00	41.00	64.82	33.88	1.01	46.87	31.94	0.41	0.57	vug
16604.00	648.57	161.00	166.00	165.93	127.41	50.66	145.40	39.53	0.50	0.62	vug
1182.00	193.72	45.00	46.00	41.91	35.91	125.26	38.79	7.00	0.40	0.57	vug
7126.00	540.44	106.00	124.00	103.69	87.51	44.06	95.25	17.18	0.31	0.54	vug

Fonte: Elaborada pelo autor

Na etapa seguinte, foram realizados dois agrupamentos utilizando K-means: 1. com todas as 11 dimensões, e 2. somente com as 4 dimensões calculadas.

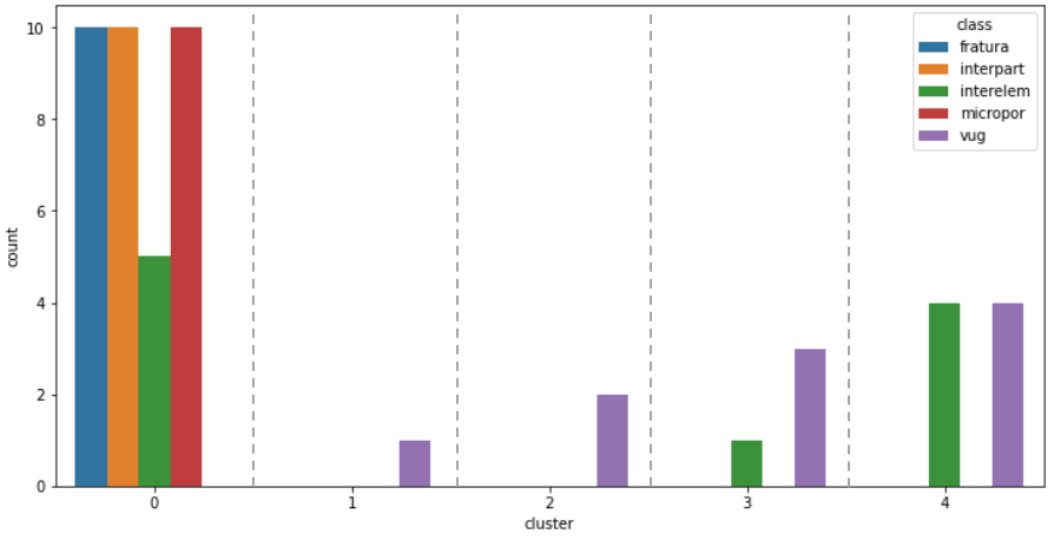
Como indicado pela curva de cotovelo apresentada na Figura 13, a utilização de todas as dimensões para o agrupamento indica um número ideal de classes abaixo do número de classes originalmente rotuladas. Idealmente, seriam necessários somente 3 grupos. Utilizando as 5 classes propostas, observa-se uma separação entre a classe vugular – agrupamentos 1, 2, 3 e 4 – e as demais classes – agrupamento 0, podendo eventualmente algumas observações da classe interelemento serem consideradas como dos agrupamentos 3 e 4 (Figura 14).

Figura 13 – Curva de cotovelo indicando um número ideal de 3 agrupamentos para o *dataset* com todas as 11 dimensões.



Fonte: Elaborada pelo autor

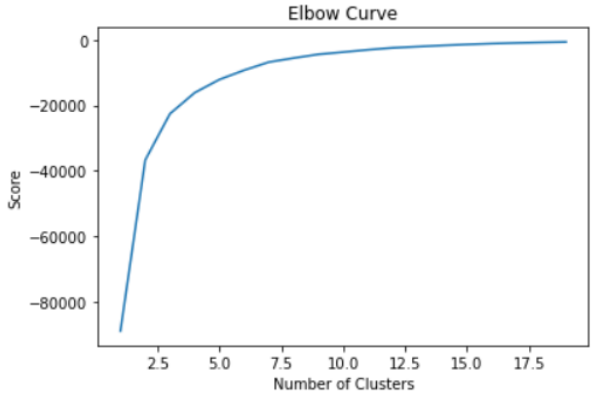
Figura 14 – Distribuição dos rótulos originais nos agrupamentos obtidos pelo método K-Means, utilizando as 11 dimensões.



Fonte: Elaborada pelo autor

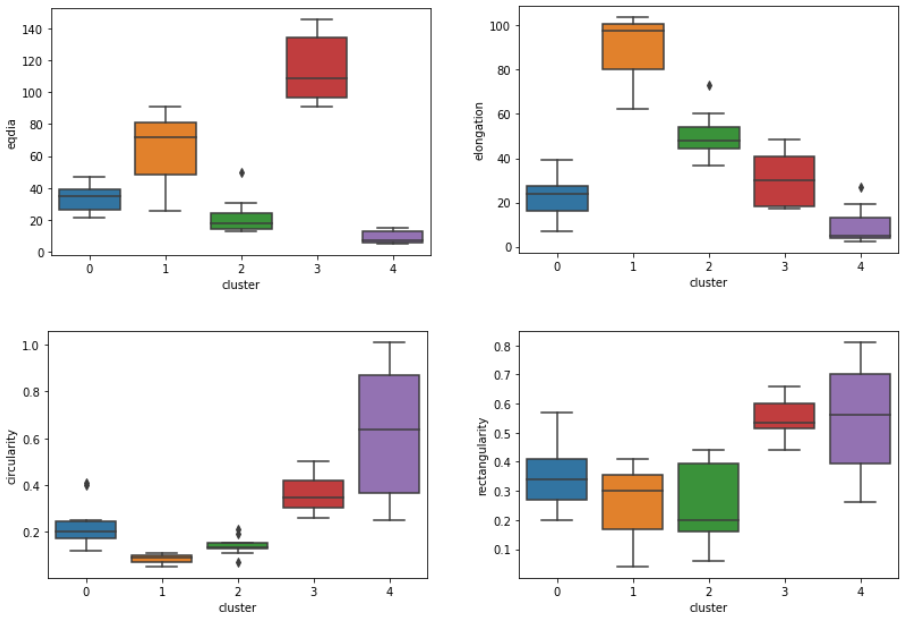
No segundo agrupamento, em que foram utilizadas somente as dimensões calculadas, a curva de cotovelo apresentada na Figura 15 indicou uma quantidade de classes mais próxima aos rótulos dos dados. Os *box plots* apresentados na Figura 16 indicam que houve uma boa separação entre os agrupamentos considerando as quatro dimensões utilizadas. Ainda assim, existem muitas regiões com sobreposição, como indicado na matriz de *crossplots* (Figura 17). Quando comparados aos rótulos originais, os agrupamentos não apresentam uma boa correspondência, sendo possível identificar diversas observações originalmente rotuladas em classes diferentes ocorrendo nos mesmos agrupamentos (Figura 18).

Figura 15 – Curva de cotovelo indicando um número ideal de 5 agrupamentos para o *dataset* com todas as 4 dimensões calculadas.



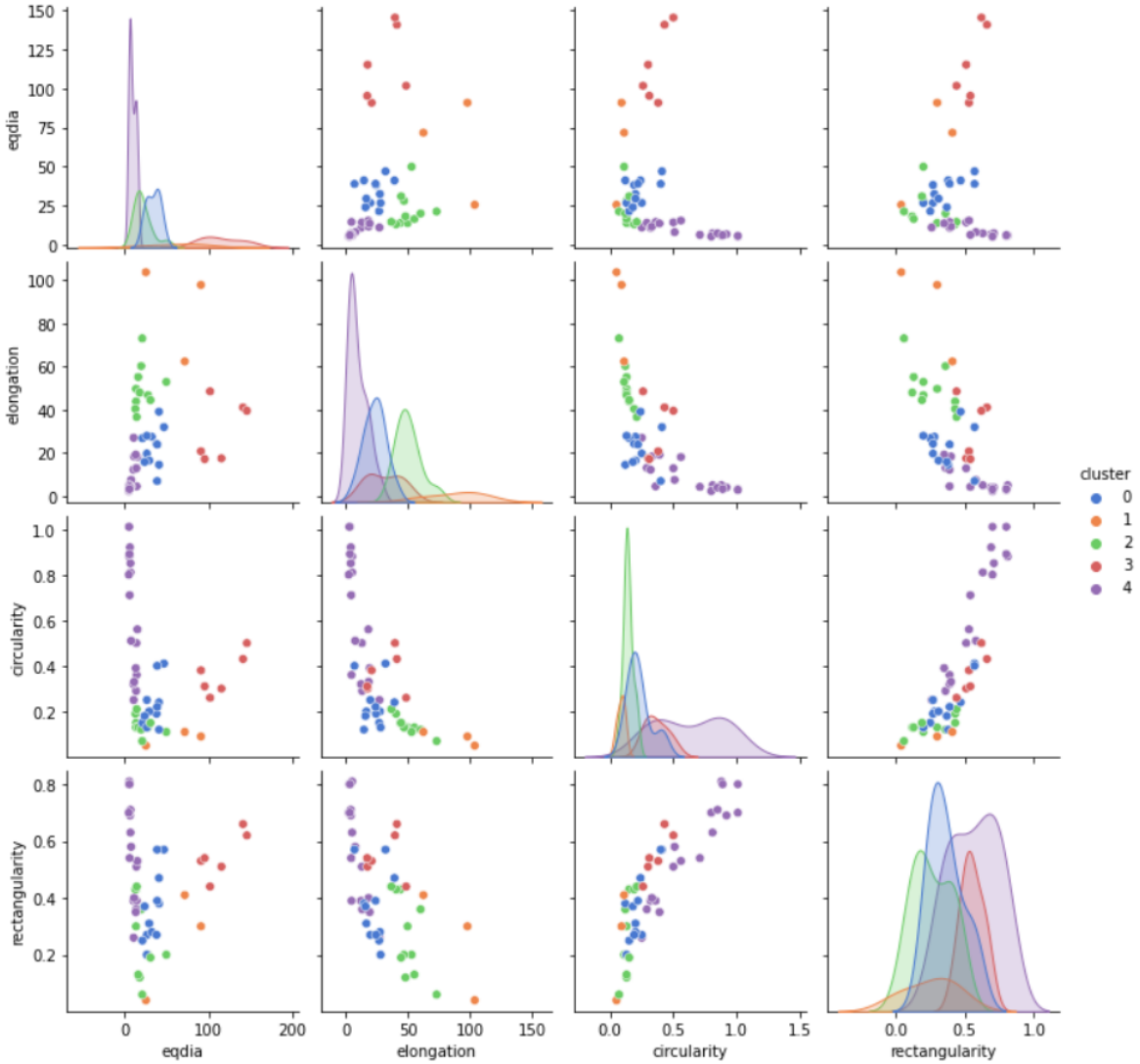
Fonte: Elaborada pelo autor

Figura 16 – *Box plots* para as quatro dimensões utilizadas, considerando a utilização de cinco grupos pelo K-Means.



Fonte: Elaborada pelo autor

Figura 17 – Matriz de *crossplots*, destacando os atributos dois a dois e seus agrupamentos correspondetes através de cores indicadas na legenda.

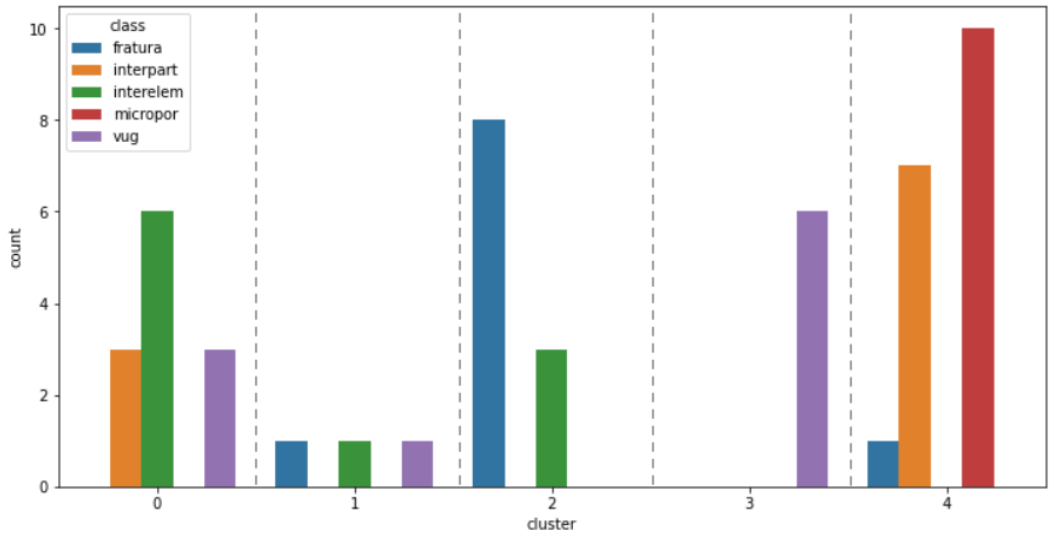


Fonte: Elaborada pelo autor

Com este resultado, uma alternativa seria realizar a proposição de uma nova classificação de tipo de porosidade baseada nestes 4 atributos geométricos calculados. Entretanto, esta classificação não apresentaria necessariamente um significado geológico. Além disso, a classificação proposta por Choquette e Pray (1970) é bastante utilizada e, apesar de diversas novas propostas, permanece como sendo a mais aceita pela comunidade científica e pela indústria.

Foram utilizados, então, dois métodos de redução de dimensionalidade: PCA e t-SNE.

Figura 18 – Distribuição dos rótulos originais nos agrupamentos obtidos pelo método K-Means, utilizando as 4 dimensões calculadas.

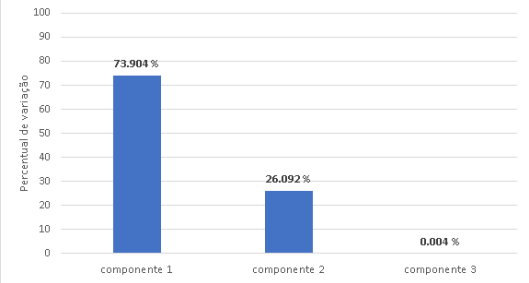


Fonte: Elaborada pelo autor

Foram obtidos 3 componentes através do PCA. Verificado que a maior parte da variância poderia ser explicada pelas duas primeiras componentes (Figura 19). Novamente o PCA foi aplicado, desta vez gerando somente 2 componentes. A primeira representa 73,9% da variação, enquanto a segunda representa 26,1%. Estas duas componentes foram utilizadas para a realização de um agrupamento em cinco categorias utilizando K-Means, em que a curva de cotovelo pode ser visualizada na Figura 20. A Figura 21 mostra dois *crossplots* entre os dois principais componentes obtidos pelo PCA destacando, através de cores, as classes originalmente rotuladas em um e os agrupamentos obtidos pelo K-Means em outro. Ao avaliar os *box plots* gerados para as duas componentes principais considerando o agrupamento encontrados, verifica-se ainda uma quantidade considerável de sobreposição entre as classes (Figura 22). Comparando o agrupamento com os rótulos originais (Figura 23), verifica-se que os agrupamentos não apresentam uma boa correspondência com esta proposta de classificação. O agrupamento utilizando as duas componentes obtidas por PCA ficou igual ao realizado para as 4 dimensões calculadas. Ambos foram mais próximos do que a correspondência obtida utilizando as 11 dimensões iniciais.

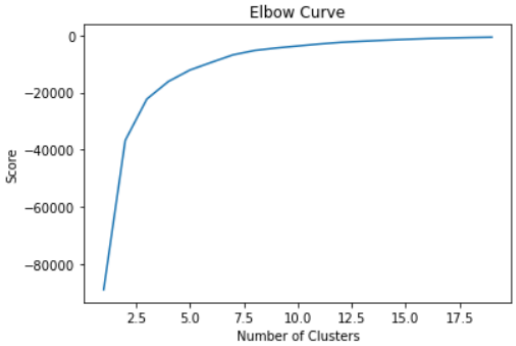
Essa correspondência é finalmente observada após a aplicação da técnica da redução de dimensionalidade chamada t-SNE, seguida pelo agrupamento utilizando o algoritmo K-Means. Considerando a elevada explicação da variância por dois componentes principais observada no PCA, foram calculados somente dois componentes utilizando o t-SNE. A curva do cotovelo é obtida (Figura 24) e são comparados os *crossplots* entre as duas componentes, um destacando as classes originalmente rotuladas e outro os agrupamentos obtidos (Figura 25). Os *box plots* para estas duas componentes considerando os agrupamentos obtidos indicam que há uma baixa sobreposição entre os agrupamentos (Figura 26).

Figura 19 – Representatividade da Variação por Componente do PCA.



Fonte: Elaborada pelo autor

Figura 20 – Curva de cotovelo indicando um número ideal de 5 agrupamentos para o *dataset* com todas as 2 componentes principais obtidas por PCA.

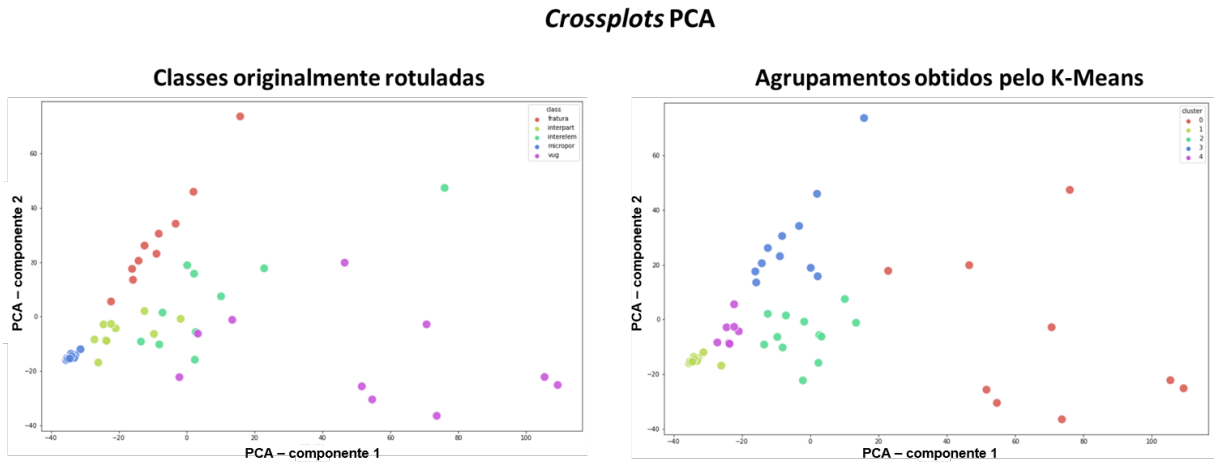


Fonte: Elaborada pelo autor

Comparando o agrupamento com os rótulos originais (Figura 27), verifica-se que algumas observações ainda são agrupadas com observações de outros rótulos. Entretanto, todos os grupos obtidos pelo K-Means parecem ser correspondentes a uma classe originalmente rotulada para os dados: o grupo 0 seria correspondente à classe vugular, o grupo 1 corresponderia à classe microporosidade, o grupo 2 se assemelha à classe interelemento, o grupo 3 à classe fratura, e o grupo 4 corresponderia à classe interpartícula. Estas correspondências não são perfeitas e algumas das observações originalmente rotuladas em outras classes ocorrem em grupos distintos, mas é o agrupamento que mais se assemelha à classificação original.

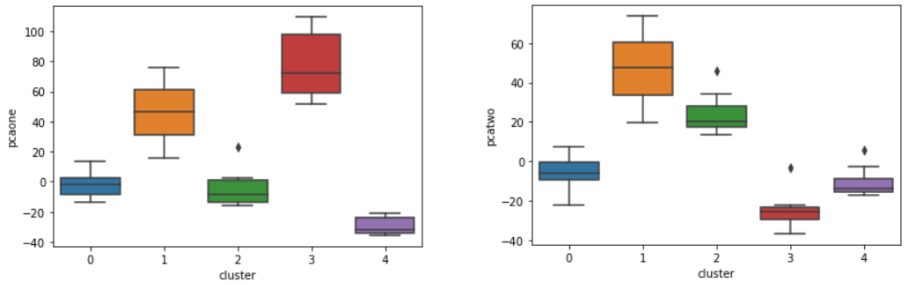
Considerando esta correspondência observada entre os grupos obtidos pelo agrupamento utilizando as componentes do t-SNE e as classes dos rótulos originais, foram geradas matrizes de confusão para todos os agrupamentos realizados respeitando esta mesma correspondência (Figura 28, Figura 29 e Figura 30). Elas permitem avaliar de forma quantitativa a evolução dos agrupamentos, desde algo próximo a um agrupamento binário obtido pelas 11 dimensões, passando por agrupamentos intermediários sem correspondência com os rótulos obtidos pelas 4 dimensões e pelas 2 componentes do PCA, até finalmente obtermos a melhor correspondência utilizando as componentes do t-SNE.

Figura 21 – *Crossplots* entre as 2 componentes principais obtidas por PCA, com destaque para as classes originalmente rotuladas e agrupamentos obtidos por K-Means.



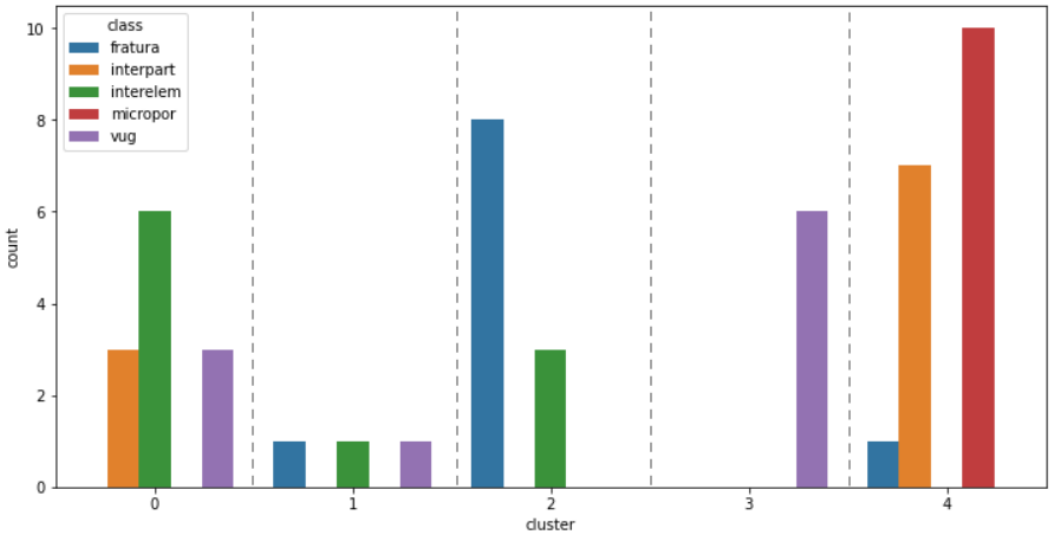
Fonte: Elaborada pelo autor

Figura 22 – *Box plots* para as duas componentes principais obtidas pelo PCA, considerando a utilização de cinco grupos pelo K-Means.



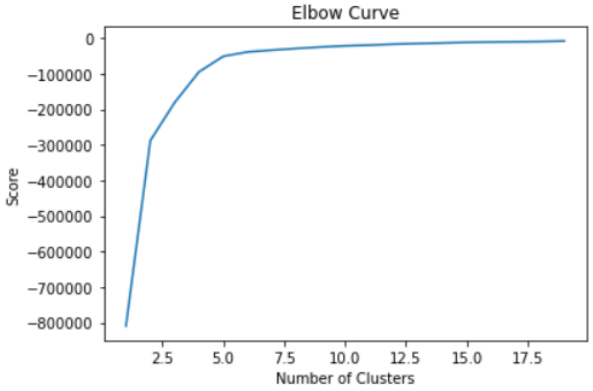
Fonte: Elaborada pelo autor

Figura 23 – Distribuição dos rótulos originais nos agrupamentos obtidos pelo método K-Means, utilizando as 2 principais componentes obtidas por PCA.



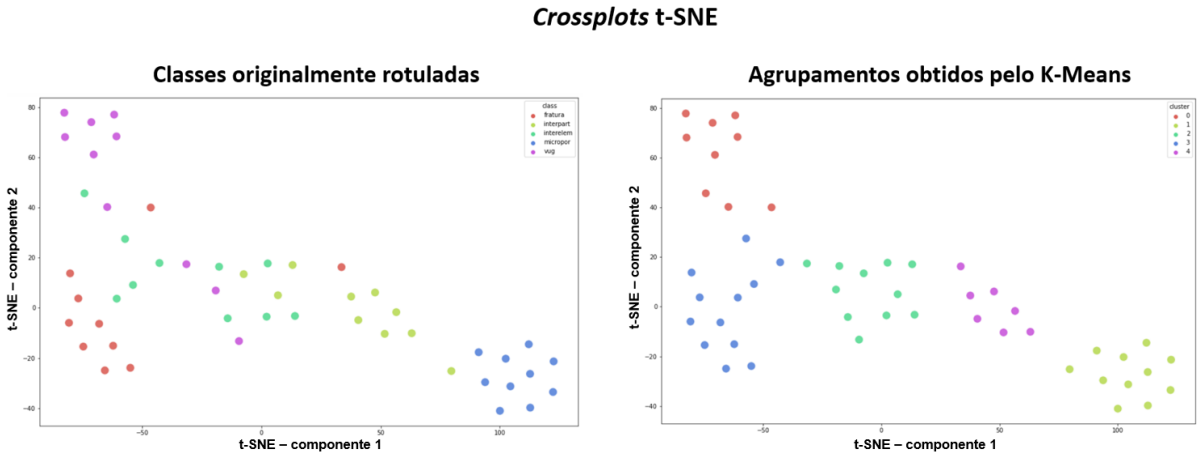
Fonte: Elaborada pelo autor

Figura 24 – Curva de cotovelo indicando um número ideal de 5 agrupamentos para o *dataset* com 2 componentes obtidas pelo t-SNE.



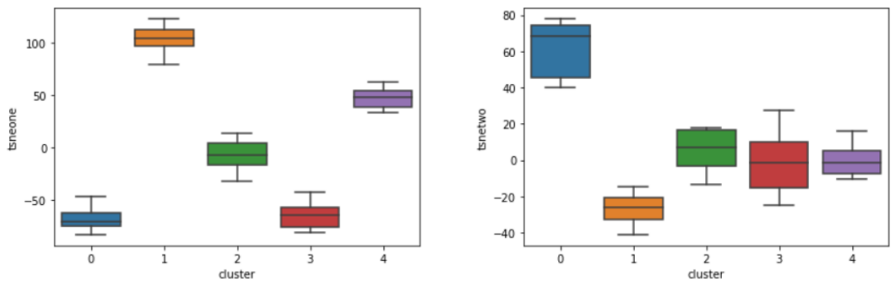
Fonte: Elaborada pelo autor

Figura 25 – *Crossplots* entre as 2 componentes principais obtidas por t-SNE, com destaque para as classes originalmente rotuladas e agrupamentos obtidos por K-Means.



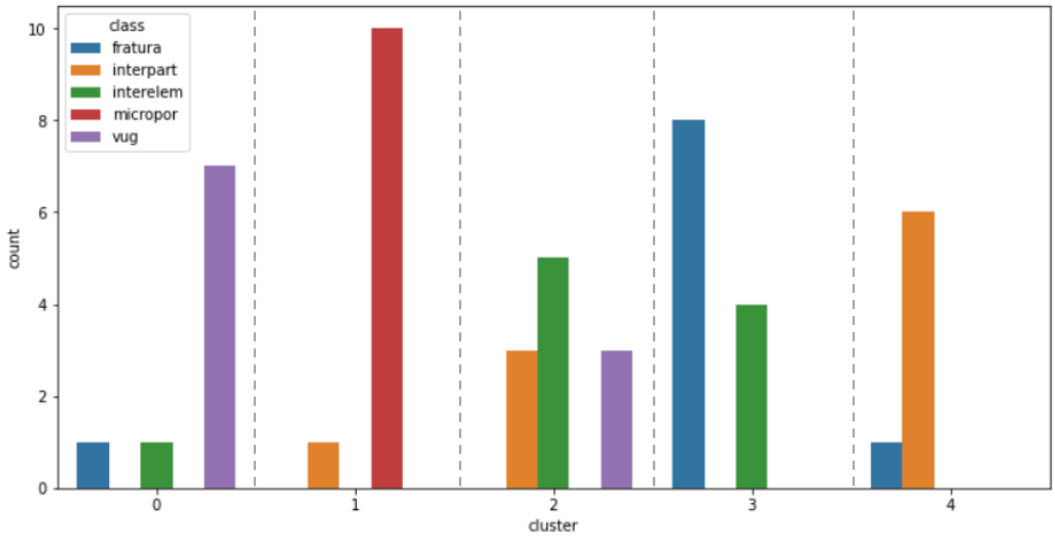
Fonte: Elaborada pelo autor

Figura 26 – *Box plots* para as 2 componentes obtidas por t-SNE, considerando a utilização de cinco grupos pelo K-Means.



Fonte: Elaborada pelo autor

Figura 27 – Distribuição dos rótulos originais nos agrupamentos obtidos pelo método K-Means, utilizando as 2 componentes obtidas por t-SNE.



Fonte: Elaborada pelo autor

Figura 28 – Matriz de confusão para o agrupamento com 11 dimensões, utilizando a correspondência descrita com os rótulos originais.

clustered as -->	0	1	2	3	4	Accuracy
Vugular	**0**	1	2	3	4	0%
Microporosidade	10	**0**	0	0	0	0%
Interelemento	5	0	**0**	1	4	0%
Fratura	10	0	0	**0**	0	0%
Interpartícula	10	0	0	0	**0**	0%

Fonte: Elaborada pelo autor

Figura 29 – Matriz de confusão para o agrupamento com 4 dimensões e com as 2 componentes do PCA, utilizando a correspondência descrita com os rótulos originais.

clustered as -->	0	1	2	3	4	Accuracy
Vugular	**3**	1	0	6	0	30%
Microporosidade	0	**0**	0	0	10	0%
Interelemento	6	1	**3**	0	0	30%
Fratura	0	1	8	**0**	1	0%
Interpartícula	3	0	0	0	**7**	70%

Fonte: Elaborada pelo autor

Figura 30 – Matriz de confusão para o agrupamento com as 2 componentes do t-SNE, utilizando a correpondência descrita com os rótulos originais.

clustered as -->	0	1	2	3	4	Accuracy
Vugular	**7**	0	3	0	0	70%
Microporosidade	0	**10**	0	0	0	100%
Interelemento	1	0	**5**	4	0	50%
Fratura	1	0	0	**8**	1	80%
Interpartícula	0	1	3	0	**6**	60%

Fonte: Elaborada pelo autor

6 DISCUSSÕES E CONCLUSÕES

O agrupamento obtido com todos os atributos geométricos não foi representativo da classificação do tipo de porosidade utilizada como rótulo do *dataset*, pois é muito influenciado pelos atributos perímetro e área. Estes atributos levam o algoritmo K-Means a realizar um agrupamento simples entre porosidades vugulares e não-vugulares. Outra restrição da aplicação de um agrupamento com todos os atributos é a necessidade de extração de dados a partir de imagens que estejam na mesma escala, uma vez que que as dimensões dos poros são relevantes. Isso pode ocasionar em um trabalho adicional para ajuste de escala entre as imagens, além de potencialmente inviabilizar a aplicação em alguma nova imagem que tenha sido adquirida em condições de aproximação muito diferente das utilizadas no *dataset*.

A utilização exclusiva de atributos adimensionais, calculados com base nos atributos cujas dimensões são relevantes, viabilizou um agrupamento mais coerente com a diversidade de estilos porosos observados nas imagens. Ainda assim, estes grupos são muito distintos das classes às quais as observações foram originalmente rotuladas.

Uma nova proposta de classificação de tipos de porosidade com base nestes atributos adimensionais poderia ser realizada com estes resultados. Entretanto, considerando o caráter geológico da classificação utilizada e sua consolidação na academia e indústria, apesar de uma grande quantidade de novas propostas, alternativas que empregam técnicas de redução de dimensionalidade, nomeadamente PCA e t-SNE, foram utilizadas previamente à etapa de agrupamento pelo algoritmo K-Means.

A técnica PCA apresentou uma correspondência igual a observada para o agrupamento com 4 dimensões. A vantagem de sua utilização está, principalmente, na simplificação da visualização dos dados, uma vez que as duas componentes obtidas podem ser visualizadas através de *crossplots* bidimensionais. A técnica t-SNE viabilizou maior correspondência entre os agrupamentos e as classes originalmente rotuladas nos dados. O resultado é próximo ao da classificação original dos dados. A simplificação na visualização dos dados através da redução em duas componentes, além redução do custo computacional na posterior etapa de agrupamento, são vantagens em utilizar redução de dimensionalidade.

REFERÊNCIAS

ANJOS, S. M. C. et al. Análise de imagens no estudo do sistema poroso de rochas-reservatório. **Boletim de Geociências da Petrobras**, v. 9, n. 2-4, p. 157–173, 1995.

ANSELMETTI, F. S.; LUTHI, S.; EBERLI, G. P. Quantitative characterization of carbonate pore systems by digital image analysis. **AAPG Bulletin**, v. 82, p. 1815–1836, 1998.

BEUCHER, S.; LANTUEJOUL, C. Use of watersheds in contour detection. **International Workshop on Image Processing: Real-time Edge and Motion Detection/Estimation**, 1979.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996.

_____. Random forests. **Machine Learning**, v. 45, p. 5—-32, 2001.

BREIMAN, L. et al. **Classification and Regression Trees**. New York, EUA: Wadsworth, 1984.

BUDENNYY, S. et al. Image processing and machine learning approaches for petrographic thin section analysis. **SPE Russian Petroleum Technology Conference**, 2017.

CHOQUETTE, P.; PRAY, L. Geological nomenclature and classification of porosity in sedimentary carbonates. **American Association of Petroleum Geologists Bulletin**, v. 54, n. 2, p. 207–250, 1970.

CUNNINGHAM, J. P.; GHAHRAMANI, Z. Linear dimensionality reduction: Survey, insights, and generalizations. **Journal of Machine Learning Research**, v. 16, 2015.

DING, C.; HE, X. K-means clustering via principal component analysis. **ICML 2004 - Proceedings Of the International Conference of Machine Learning**, p. 225—-232, 2004.

DUMOULIN, V.; VISIN, F. A guide to convolution arithmetic for deep learning. **ArXiv preprint**, p. 1603.07285, 2016.

EHRLICH, R. et al. Petrography and reservoir physics, i: Objective classification of reservoir porosity. **American Association of Petroleum Geologists Bulletin**, v. 75, p. 1547–1562, 1991.

EHRLICH, R.; DAVIES, D. K. Photographic image analysis, i: analysis of reservoir pore complexes. **Journal of Sedimentary Petrology**, v. 54, n. 4, p. 1365–1378, 1984.

FLEISCHER, M.; WILCOX, R.; MARZKO, J. Microscopic determination of the nonopaque minerals. **US Geological Survey Bulletin**, n. 1627, 1984.

GREGG, J. **Dolomite Petrology Photo Gallery. Department of Geology Geophysics**. 2001. Disponível em: <http://web.mst.edu/~greggjay/carbonate_page/dologallery/pages/d-MAU2b_10.htm>.

GURGEL, S. **Análise de técnicas de implementação paralela para treinamento de redes neurais em GPU**. 2014. Dissertação (Mestrado) — Universidade Federal da Paraíba, João Pessoa, 2014.

HALL, M. et al. The weka data mining software: an update. **ACM SIGKDD Explor. Newslett**, v. 11, n. 1, p. 10–18, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. Stanford, California: Springer, 2009.

IZADI, H.; SADRI, J.; BAYATI, M. An intelligent system for mineral identification in thin sections based on a cascade approach. **Computer and Geosciences**, v. 99, p. 37–49, 2017.

JODRY, R. **Pore geometry of carbonate rocks - basic geologic concepts**. New York, EUA: Oil and Gas Production from Carbonate Rocks, Elsevier, 1972.

KAMBHATLA, N.; LEEN, T. Dimension reduction by local principal component analysis. **Neural Computation**, v. 9, p. 1493—1516, 1997.

KHAN, S. et al. **A Guide to Convolutional Neural Networks for Computer Vision. In: Synthesis Lectures on Computer Vision n. 15 (eds. Medioni, G.; Dickinson, S.)**. [S.l.]: Morgan and Claypool Publishers, 2018.

LIMA, R. et al. Petrographic microfacies classification with deep convolutional neural networks. **Computers and Geosciences**, v. 142, n. 104481, 2020.

LLOYD, S. P. Least square quantization in pcm. **IEEE Transactions on Information Theory**, v. 28, n. 2, p. 129—137, 1982.

LONOY, A. Making sense of carbonate pore systems. **AAPG Bulletin**, v. 90, 2016.

LUCIA, F. J. Petrophysical parameters estimated from visual descriptions of carbonate rocks: a field classification of carbonate pore space. **Journal of Petroleum Technology**, v. 35, p. 629–637, 1983.

_____. Rock-fabric/petrophysical classification of carbonate pore space for reservoir characterization. **AAPG Bulletin**, v. 79, n. 9, p. 1275–1300, 1995.

MAATEN, L. van der; HERIK, H. J. V. D.; POSTMA, E. Dimensionality reduction: A comparative review. **Journal of Machine Learning Research**, v. 8, n. 1, 2007.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. **Journal of Machine Learning Research**, v. 9, p. 2579–2605, 2008.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. **Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press**, p. 281—297, 1967.

NESSE, W. **Introduction to optical mineralogy**. New York, EUA: Oxford University Press, 3rd edition, 2004.

PERKINS, D.; HENKE, K. **Minerals in thin section**. New Jersey, EUA: Pearson Education, 2004.

RODUIT, N. **JMicroVision: un logiciel d'analyse d'images pétrographiques polyvalent**. 2007. Tese (Doutorado) — Faculté des sciences de l'Université de Genève, 2007.

RUBO, R. A. et al. Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. **Journal of Petroleum Science and Engineering**, v. 183, n. 106382, 2019.

SCHINDELIN, J. et al. Fiji: an open-source platform for biological-image analysis. **Nat. Methods**, v. 9, n. 7, p. 676—682, 2012.

SCHOLLE, P.; ULMER-SCHOLLE, D. **A Color Guide to the Petrography of Carbonate Rocks: Grains, textures, porosity, diagenesis**. Tulsa, EUA: The American Association of Petroleum Geologists, AAPG Memoir 77, 2003.

STEINHAUS, H. Sur la division des corps matériels en parties. **Bull. Acad. Polon. Sci.**, v. 4, n. 12, p. 801—-804, 1967.

SU, C. et al. Rock classification in petrographic thin section images based on concatenated convolutional neural networks. **Earth Science Informatics**, 2020.

TERRA, G. et al. Classificação de rochas carbonáticas aplicável às bacias sedimentares brasileiras. **Boletim de Geociências da Petrobras**, v. 18, n. 1, 2010.

TOMUTSA, L.; BRINKMEYER, A.; RAIBLE, C. **Determining petrophysical properties of reservoir rocks by image analysis**. Bartlesville: National Institute for Petroleum and Energy Research, 1984.

TONIETTO, S.; POPE, M. Modification of pore geometry and petrophysical characteristics of the upper jurassic smackover formation thrombolite reservoirs after dolomitization. **Boletim de Geociências da Petrobras**, v. 5, n. p. 275-292, 2016.

ZHAO, M. **Petrologic and Petrophysical Characteristics: Mississippian Chert, Oklahoma**. 2011. Dissertação (Mestrado) — Oklahoma State University, 2011.

ŁADNIAK, M.; MŁYNARCZUK, M. Search of visually similar microscopic rock images. **Computer and Geosciences**, v. 19, p. 127–136, 2015.