UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE ENGENHARIA DE MATERIAIS

Vito Francisco Chiarella

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE FATIAS DE SISTEMAS POROSOS EM IMAGENS TOMOGRÁFICAS 3D POR MÉTODOS DE CLUSTERIZAÇÃO

Florianópolis

2018

Vito Francisco Chiarella

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE FATIAS DE SISTEMAS POROSOS EM IMAGENS TOMOGRÁFICAS 3D POR MÉTODOS DE CLUSTERIZAÇÃO

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Materiais para a obtenção do Grau de Mestre em Ciência e Engenharia de Materiais.
Orientador: Prof. Dr. Celso Peres Fernandes

Florianópolis

2018

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Chiarella, Vito Francisco
Caracterização e classificação de fatias de sistemas porosos em imagens tomográficas 3D por métodos de clusterização / Vito Francisco Chiarella ; orientador, Celso Peres Fernandes, 2018.
76 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós Graduação em Ciência e Engenharia de Materiais, Florianópolis, 2018.

Inclui referências.

1. Ciência e Engenharia de Materiais. 2. Materiais porosos. 3. Visão computacional. 4. Análise de características. 5. Análise de clusters. I. Fernandes, Celso Peres. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência e Engenharia de Materiais. III. Título.

Vito Francisco Chiarella

CARACTERIZAÇÃO E CLASSIFICAÇÃO DE FATIAS DE SISTEMAS POROSOS EM IMAGENS TOMOGRÁFICAS 3D POR MÉTODOS DE CLUSTERIZAÇÃO

Esta Dissertação foi julgada aprovada para a obtenção do Título de "Mestre em Ciência e Engenharia de Materiais", e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Materiais.

Florianópolis, 29 de Novembro 2018.

Prof. Dr. Guilherme Mariz de Oliveira Barra
Coordenador

Banca Examinadora:

Prof. Dr. Celso Peres Fernandes
Orientador

Dr. rer. nat. Eros Comunello

Prof. Dr. Anderson Camargo Moreira

AGRADECIMENTOS

Agradeço à minha familia que sempre me empurrou para ser o melhor eu que pudesse ser.

À minha namorada, que me apoiou a cada passo ao longo desta jornada.

Ao meu orientador e demais pessoas da UFSC, por sua paciencia ensinamentos e ter acreditado em mim.

E finalmente agradecimentos especiais a Petrobras por ceder parte dos dados usados neste trabalho.

An education was a bit like a communicable sexual disease. It made you unsuitable for a lot of jobs and then you had the urge to pass it on.

Terry Pratchett

RESUMO

Rochas porosas apresentam características físicas que podem ser preditas a partir da analise de suas imagens tomográficas (MIRABOLGHASEMI et al., 2015), tradicionalmente as medições realizadas nas imagens são as de porosidade, permeabilidade e autocorrelação, que permitem prever propriedades como escoamento de fluxo através do material, ou a topologia dos poros (ADLER; JACQUIN; QUIBLIER, 1990). Porém estas características podem variar, as vezes drasticamente, dentro de uma mesma rocha, especialmente ao longo do seu eixo de formação, o qual implica em diferentes características petrofísicas. Considerando esta heterogeneidade das rochas é possível intuir que rochas completamente diferentes possam ter fatias que se comportem de forma semelhante entre si, e portanto uma forma de caracterizar e classificar estas fatias semelhantes pode trazer uma nova perspectiva ao campo de analise destes materiais. No presente trabalho um conjunto de doze imagens tomográficas de raio-X de plugs de rochas porosas foram selecionadas pela sua aparente homogeneidade, as características de suas camadas foram avaliadas, parte destes resultados foram classificados utilizando diferentes algoritmos de clusterização, e o restante dos dados foram utilizados para validar a classificação. O trabalho mostra resultados promissores, permitindo classificar com um elevado grau de certeza a qual rocha pertence uma camada.

Palavras-chave: Materiais porosos. Visão computacional. Análise de características. Análise de clusters.

ABSTRACT

Porous rocks present physical characteristics that can be predicted based on the analysis of their tomographic images (MIRABOLGHASEMI et al., 2015). Traditionally those images are analyzed to extract values for porosity, permeability and auto-correlation, which allows to foresee properties such as the flow of a given fluid through the material, or its pore topology (ADLER; JACQUIN; QUIBLIER, 1990). However these characteristics might change, sometimes drastically, inside the same rock, especially along its formation axis, which implies in different petrophysical characteristics. Given that these rocks are heterogeneous it's possible to realize that very different rocks can have slices that behave similarly, and thus a way to characterize and classify these similar slices can bring new insights to the way these materials are analyzed. In the current work a set of twelve x-ray tomographic images of porous rock plugs were selected for their apparent homogeneity, their characteristics were extracted by slices, part of these results were classified using different clusterization algorithms, and the remaining of the data was used for validation of the classification. The work presents promising results, allowing the classification with a high degree of certainty of which sample a given slice belongs to.

Keywords: Porous materials. Computational vision. Characteristic analysis. Cluster analysis

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES

CT	Computational Tomography (Tomografia computacional)
CSV	Comma Separated Values (Valores separados por coma)

LISTA DE SIMBOLOS

ϕ	Porosidade
R_H	Raio hidráulico
$\prec \succ$	Média estatística para o domínio da imagem em consideração

SUMÁRIO

1 INTRODUÇÃO

A estimativa de propriedades petrofísicas de um material poroso a partir de medições obtidas de imagens tomográficas é um assunto que tem sido muito estudado recentemente. Isto se deve em parte a que imagens tomográficas, mesmo custosas, permitem a extração de varias características de um material com apenas um experimento (MOHAGHEGH et al., 1996). Outros métodos experimentais para a medição direta destas características podem ser destrutivos e requererem varias amostras, cuja obtenção pode ser ainda mais custosa. Um exemplo é o no caso do pré-sal onde "a exploração dessas reservas encontra grandes desafios, como a profundidade da lâmina d'água e a espessura de coluna de rochas a serem atravessadas, as enormes pressões e temperaturas a serem encontradas, o comportamento do sal e da porosidade dos reservatórios face à perfuração, e as heterogeneidades dos reservatórios carbonáticos, dentre outras" (RICCOMINI et al., 2012)

As estimativas de propriedades descritas anteriormente normalmente assumem a homogeneidade do material poroso em questão, porém muitas vezes uma rocha pode ser composta de diversas camadas que apresentem propriedades diferentes devido ao seu processo de formação.

É de interesse a criação de um método para classificação destas camadas, este trabalho propõe uma solução baseada em algoritmos de clusterização aplicados a parâmetros extraídos de imagens tomográficas de diversos tipos de rochas.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Propor um conjunto de medições simples que podem ser obtidos de imagens tomográficas capazes de permitir que um algoritmo de clusterização reconheça a qual amostra pertence uma determinada fatia.

1.1.2 Objetivos específicos

- Definir um conjunto de medições capazes de extrair informações sobre a amostra.
- Encontrar algoritmos de clusterização capazes de distinguir os clusters esperados.
- Validar o algoritmo selecionado com um conjunto de dados cujo valor de referencia é conhecido.

1.2 ESTRUTURA DO TRABALHO

O trabalho consiste em 6 capítulos, incluindo a presente introdução que presenta o tema a do trabalho de forma mais ampla.

O Capitulo 2 apresenta o estado da arte dos vários métodos utilizados no decorrer deste trabalho bem como o motivo de sua escolha para a analise.

O Capitulo 3 apresenta trabalhos similares, bem como uma comparação dentre eles e com este trabalho.

O Capitulo 4 apresenta o processo cientifico utilizado neste trabalho.

O Capitulo 5 apresenta os resultados obtidos dos dados avaliados.

Finalmente o Capitulo 6 apresenta a conclusão do trabalho bem como a interpretação dos dados apresentados anteriormente.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 IMAGENS TOMOGRÁFICAS

Diferente de métodos tradicionais de captura de imagens o tomografo de raios-X captura volumes e não apenas imagens bidimensionais, por este motivo a unidade básica de medição destas imagens é conhecida como voxel ao invés da tradicional nomenclatura de pixel (KETCHAM; CARLSON, 2001).

Estas imagens tridimensionais apresentam várias vantagens em relação as tradicionais laminas delgadas, dentre as quais pode-se destacar a existência de informação de profundidade, permitindo mensurar continuidade de poros, como pode ser visto na Figura 1, onde podem ser observados poros aparentemente desconexos na lamina delgada que foram classificados como parte do mesmo poro pelo algoritmo de segmentação (Números foram adicionados a alguns dos poros para permitir a visualização mesmo em impressões preto e branco).

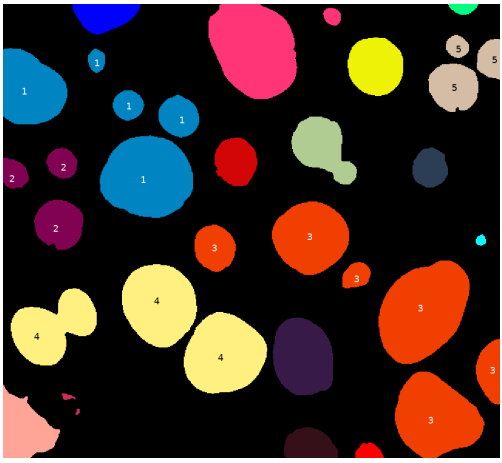


Figura 1: Poros aparentemente desconexos

Mais importante para o trabalho atual é que imagens tomográficas nos permitem observar mudanças do material ao longo desta terceira dimensão, a qual usualmente corresponde com o eixo de formação da rocha, devido ao seu processo de extração.

Alem disso, diferentemente das laminas delgadas, o processo de captura de imagens tomográficas não danifica a amostra original, garantindo assim que a estrutura interna dos poros não foi alterada por um processo físico como o da injeção do material que acontece para a captura de laminas delgadas.

2.2 PREPARAÇÃO DA IMAGEM

Uma vez obtida a imagem tomográfica ela deve ser preparada para a extração de dados. Alguns dos dados extraídos dependem do conhecimento sobre quais voxels da imagem se referem ao arcabouço solido e quais se referem a poros, e outros dependem ainda de informação de continuidade entre os diversos poros da imagem.

2.2.1 Binarização

As imagens obtidas pelo tomografo não distinguem entre material solido e espaço oco, para tanto é necessário classificar cada voxel da imagem como solido ou vazio. O processo de classificar cada voxel de uma imagem em um grupo é conhecido como Segmentação (IASSONOV; GEBRENEGUS; TULLER, 2009). Quando exitem apenas dois grupos de interesse algumas simplificações podem ser feitas, este subgrupo de algoritmos de segmentação é conhecido como Binarização.

Existem vários métodos de binarização na literatura, e todos eles são sensíveis a parametrização, podendo mudar drasticamente seu resultado a depender da mesma.

O algoritmo mais simples e tradicionalmente utilizado envolve a determinação de um limiar, onde apenas voxels com valor superior a dito limiar são considerados sólidos. Porém a escolha deste valor para limiar deve mudar a depender da amostra e do equipamento utilizado para captura, fazendo da escolha deste valor o ponto mais importante na binarização de uma imagem.

No presente trabalho foi utilizado o método de Otsu (OTSU, 1979) para a determinação do limiar, esta escolha se baseia na sua ampla aceitação nos trabalhos avaliados como um dos métodos que produz os resultados mais confiáveis.

Este método procura o valor de limiar que divida o histograma de tal forma que a variância intraclasse seja a menor possível, i.e. que

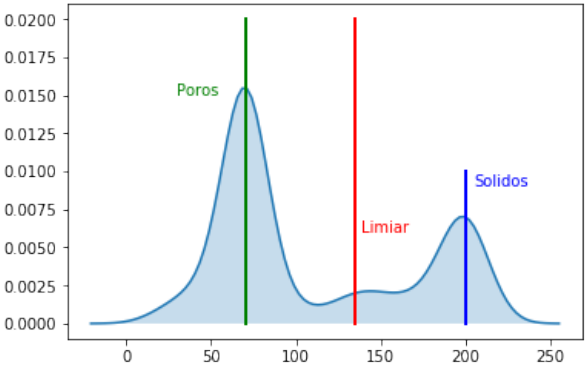


Figura 2: Limiar de Otsu para um Histograma

os valores considerados poro tenham a menor variância entre si, assim como aqueles considerados solido, como pode ser observado na Figura 2 onde os grupos de Poros e Sólidos foram identificados pelo seu centro e o Limiar que separa os dois grupos também foi demarcado.

Este algoritmo pode apresentar problemas para imagens compostas por diversas estruturas diferentes entre si, por isto o algoritmo foi aplicado para cada uma das amostras, consideradas homogêneas, individualmente ao invés de calcular um único limiar para todas as amostras.

2.2.2 Rotulação

Com a imagem binarizada é possível realizar uma nova segmentação com o intuito de rotular poros diferentes dependendo de sua conectividade.

Vários métodos existem para segmentar uma imagem (PAL; PAL, 1993), e eles normalmente são baseados em Limiar, Fronteiras, Regiões ou híbridos (SHIH; CHENG, 2005). A binarização descrita anteriormente é um exemplo de segmentação baseada em limiar, porém este tipo de segmentação não leva em conta a continuidade das regiões avaliadas, o qual é muito útil para se determinar o possível escoamento através de uma rocha.

Para o escopo deste trabalho o interesse esta na continuidade dos voxels considerados poros, mesmo que não aparentem estar

quando uma fatia do volume é observada. Os métodos encontrados na bibliografia, costumam utilizar métodos híbridos para tentar diminuir os erros causados por diferenças na imagem, como ser (TREMEAU; BOREL, 1997).

Existem na literatura diversos métodos de rotulação de uma imagem, um dos mais simples e intuitivos deles esta descrito em (ADAMS; BISCHOF, 1994) e consiste na escolha de um voxel para ser usado como semente do processo, os vizinhos deste voxel irão ser avaliados pela sua semelhança com a semente, os que forem semelhantes serão considerados parte do mesmo objeto e terão seus vizinhos avaliados. Parte deste processo ainda pode ser simplificado se a imagem binarizada for utilizada, desta forma a verificação de semelhança consiste apenas em se o voxel pertence a mesma fase.

Este método pode ser aplicado iterativamente procurando na imagem um voxel vazio que não tenha sido rotulado ainda e utilizando ele como semente até não poder adicionar mais vizinhos, então um novo voxel é escolhido. O processo pode ser simplificado da seguinte forma:

1. Cria-se uma imagem rotulada do mesmo tamanho da imagem binarizada e inicializa-se ela com 0
2. Inicia-se um contador com 0
3. Aumenta-se o contador em 1
4. Vasculha-se a imagem binarizada por um voxel que seja vazio e tenha valor 0 (não rotulado) na imagem rotulada
5. Adiciona-se este voxel a lista de voxels a verificar
6. Escolhe-se o primeiro elemento da lista de voxels a verificar
7. Este voxel é pintado na imagem rotulada com o valor do contador
8. Os seus vizinhos que forem vazios e não rotulados são adicionados a uma lista de voxels a verificar
9. Volta-se a 6 até que a lista fique vazia
10. Volta-se a 3 até que nenhum voxel não rotulado possa ser achado

Um exemplo do resultado deste algoritmo pode ser observado na Figura 1, onde os diferentes objetos foram coloridos de acordo com o numero com o qual foram pintados na imagem rotulada.

2.3 MEDIÇÕES

A seguinte seção apresenta as medições efetuadas nas imagens tomográficas bem como a relação delas com certas propriedades petrofísicas

2.3.1 Porosidade

A porosidade, representada neste trabalho pela letra grega ϕ, é uma medida percentual do espaço vazio de um material poroso comparada com o seu volume total observado (DULLIEN, 2012). Alternativamente esta medição nos indica quanto do material é de fato parte do arcabouço solido.

Experimentalmente este valor pode ser medido através da submersão de um material poroso em um fluido e comparando o peso dele com o seu peso quando seco, no caso de imagens tomográficas binarizadas pode ser obtido a partir da razão entre a contagem dos voxels considerados poro em relação ao total de voxels da imagem.

Materiais com a mesma porosidade podem ter os mais variados índices de escoamento, pois a percentagem de espaço vazio não diz nada a respeito de sua organização. Desta forma um material com uma dada porosidade pode ter uma infinidade de formatos.

Por exemplo pode ser observado na Figura 3, três diferentes imagens binarizadas representando o mesmo índice de porosidade. A Figura 3(a) representa uma imagem artificial, criada com uma porosidade especifica, 3(b) apresenta poros não conexos, portanto o seu índice de escoamento é zero, enquanto que 3(c) apresenta um único macro-poro.

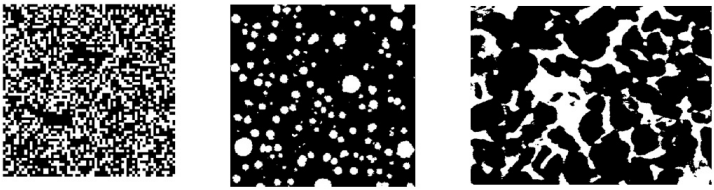


Figura 3: Três imagens com a mesma porosidade.

2.3.2 Autocorrelação

A autocorrelação representa uma medição da estrutura interna de um material, e como tanto permite predições sobre diversas outras propriedades do material, como a permeabilidade (BERRYMAN; BLAIR, 1986).

Matematicamente a autocorrelação pode ser descrita pelos momentos de uma variável aleatória, denominada de função de fase e definida por:

$$Z_\alpha(x) = \begin{cases} 1 & \text{se } x \in \alpha \\ 0 & \text{se } x \notin \alpha \end{cases} \tag{2.1}$$

Onde $x = (i, j, k)$ denota um vetor posição em relação a uma origem arbitrária, e α designa a fase, poros ou sólidos. A fração "volumétrica" da fase será dada por:

$$\varepsilon_\alpha = \prec Z_\alpha(x) \succ \tag{2.2}$$

Onde o símbolo $\prec \succ$ denota a média estatística para o domínio da imagem em consideração, sabendo-se que

$$\sum_{\alpha=1}^{n} \varepsilon_\alpha = 1 \tag{2.3}$$

Onde n é o número de fases do material, no caso de materiais porosos pode-se assumir apenas duas fases (poro e sólido) simplificando a equação para $\varepsilon_{poros} + \varepsilon_{sólido} = 1$, desta forma ε_{poros} representa a média estatística para o domínio da fase poro, i.e. a porosidade ϕ.

Com a hipótese de meio estatisticamente homogêneo, a função de autocorrelação a m pontos para cada fase é escrita como:

$$C_\alpha(u) = \prec Z_\alpha(x) Z_\alpha(x + u_1) \dots Z_\alpha(x + u_m) \succ \tag{2.4}$$

Tradicionalmente $m = 2$, pois a complexidade matemática para $m > 3$ é muito superior sem apresentar ganhos em acurácia (KARSANINA et al., 2015). Portanto:

$$C_\alpha(u) = \prec Z_\alpha(x) Z_\alpha(x + u) \succ \tag{2.5}$$

Onde u representa um deslocamento arbitrário no volume. Assumindo um meio isotrópico a autocorrelação depende apenas do deslocamento em uma direção, e não do seu sentido, em outras

palavras pode-se assumir $u = |u|$.

Pode-se imaginar $C(u)$ como a chance de duas agulhas separadas por um raio u caírem sobre a fase α ao serem arremessadas aleatoriamente sobre uma imagem com pelo menos duas fases, este método de cálculo da autocorrelação é conhecido como o método das agulhas. Sendo assim a função de autocorrelação é mais praticamente uma medição que indica qual a probabilidade de achar dois pontos separados por um raio u em uma região ocupada por uma das fases do material (BERRYMAN, 1985).

Computacionalmente este cálculo pode ser feito através da translação e sobreposição de um volume sobre ele mesmo e a verificação de quantos pontos ambos possuem numa mesma fase α de interesse, sendo não somente mais rápido que uma implementação do método das agulhas descrito anteriormente como também permitindo resultados reproduzíveis (YEONG; TORQUATO, 1998).

Na Figura 4 a área de interesse é representada por um quadrado preto, e apenas é considerado poro as áreas brancas, ou seja, onde nem a imagem original (vermelho) nem a transposta (azul) cobriram o fundo.

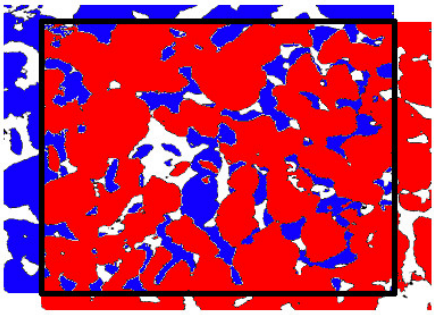


Figura 4: Uma fatia de um volume transposta sobre si mesma.

O cálculo pode ser simplificado ainda mais fazendo o deslocamento u acontecer em apenas um dos eixos principais do volume (X,Y,Z) assim obtém-se uma fórmula de autocorrelação para cada eixo.

$$
\begin{aligned}
X : C_\alpha(u) &=\prec Z_\alpha(i,j,k)Z_\alpha(i+u,j,k) \succ \\
Y : C_\alpha(u) &=\prec Z_\alpha(i,j,k)Z_\alpha(i,j+u,k) \succ \\
Z : C_\alpha(u) &=\prec Z_\alpha(i,j,k)Z_\alpha(i,j,k+u) \succ
\end{aligned}
\tag{2.6}
$$

O processo é repetido transladando o volume cada vez mais, i.e. aumentando u, e um gráfico é formado onde um eixo representa a distância que a imagem foi transladada e o outro o valor obtido pelo algoritmo. A Figura 5 representa um exemplo destes gráficos para deslocamentos nos eixos X e Y da mesma amostra.

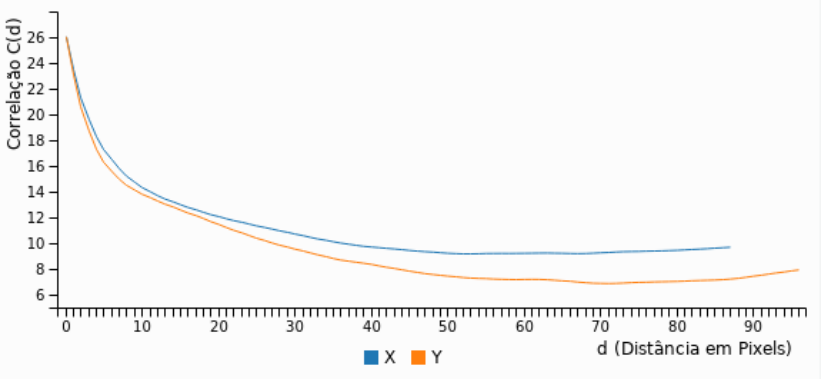


Figura 5: Gráfico de autocorrelação por volume de poro em duas direções.

Pode ser observado no gráfico da figura 5 que para a distância 0 o valor é igual à porosidade ε, o qual pode ser intuído já que para uma distância 0 a chance de ambas agulhas caírem em um poro deve ser igual à percentagem de volume que corresponde a poro. Mais interessante porém, é o outro extremo do gráfico, onde observa-se que a partir de um determinado valor de distância cada agulha tem uma chance totalmente aleatória de atingir um poro, i.e. porosidade ao quadrado ε^2.

Uma alternativa a esta determinação, trabalhando no domínio de frequência, operando-se a transformada de Fourier na imagem. Este procedimento se apoia no teorema de Wiener-Khinchin (LIANG et al., 1998): a transformada de Fourier da função de autocorrelação de um campo (a imagem) é o power spectrum deste campo. Com este procedimento, reduz-se bastante as flutuações na função de autocorrelação. Define-se ainda uma função de autocorrelação

normalizada por:

$$R_\alpha(u) = \frac{\prec (Z_\alpha(i,j,k) - \varepsilon_\alpha)(Z_\alpha(i+u,j,k) - \varepsilon_\alpha) \succ}{\prec (Z_\alpha(i,j,k) - \varepsilon_\alpha)^2 \succ} \tag{2.7}$$

Pode-se mostrar que as funções de autocorrelação normalizadas para fase poro e para a fase sólido são iguais. Esta função normalizada assume o valor 1, para deslocamento nulo e o valor 0 para deslocamentos grandes.

Este comportamento nos permite transformar o gráfico em um único numero, i.e. a distância em que o comportamento passa a ser aleatório, o qual permite alimentar ele mais facilmente para os algoritmos de clusterização na análise estatística.

2.3.3 Conectividade

Semelhante a autocorrelação é possível interpretar uma imagem segmentada como uma imagem composta de várias fases, uma para a fase solida e uma fase para cada macro-poro conexo. Portanto é possível aplicar método descrito anteriormente na seção 2.3.2 da mesma forma, contabilizando as sobreposições de voxels da mesma fase, que no caso de uma imagem segmentada representam o mesmo poro conexo ao invés do espaço poroso ao todo.

Desta forma esta medição nos permite ter uma ideia não só da distribuição do espaço poroso, senão da conectividade que existe entre ele, sendo um valor aproximadamente igual a autocorrelação para materiais compostos por um único macro-poro, mas completamente diferente para materiais com a mesma porosidade e autocorrelação mas compostos de vários poros fechados.

Analogamente ao descrito anteriormente pode-se definir uma função de Conectividade C para um raio u como a chance de duas agulhas caírem no mesmo poro n para qualquer que seja n dos m poros presentes no volume.

$$C(u) = \sum_{n=1}^{m} \prec Z_n(x) Z_n(x+u) \succ \tag{2.8}$$

E da mesma forma este valor vai tender a ser completamente aleatório para maiores valores de u, com a diferença de que ao contrario da Autocorrelação é possível que este valor seja zero para

valores suficientemente altos de u a depender da estrutura interna do material.

2.3.4 Tamanho médio de poro

A forma de medir o tamanho médio de poro de uma imagem tomográfica pode ser feita através da distribuição de tamanhos de poros, também conhecida como granulometria (COSTER; CHERMANT, 1989). O termo vem da medição de tamanho de grãos por um método iterativo utilizando peneiras de tamanho progressivamente menores, onde o conteúdo filtrado é pesado e separado, de forma a poder obter um gráfico de correlação entre a quantia de material e o tamanho da peneira (e portanto do grão que não conseguiu passar por ela).

Tradicionalmente para materiais porosos esta medição é obtida experimentalmente através de porosimetria por intrusão de mercúrio. Neste método uma amostra é colocada em contato com mercúrio, e a pressão da câmara onde a amostra se encontra é diminuída gradualmente, permitindo que o mercúrio entre nos poros. Devido ao mercúrio ser normalmente um fluido não-molhante é possível assumir que o fluido não conseguira entrar em poros menores do que a diferença de pressão atual permitiria pela lei de Darcy (WHITAKER, 1986). Assim é possível mensurar o volume de mercúrio que foi introduzido na amostra para uma determinada pressão e determinar o volume da sua fase porosa alcançável através de cavidades de certo tamanho. O resultado deste processo é um gráfico cumulativo que correlaciona o tamanho do poro com o volume do solido.

Este método porém não determina corretamente o tamanho dos poros internos, já que podem existir grandes cavidades que apenas são acessáveis por gargantas pequenas, e o volume inteiro deste poro seria considerado como tendo o menor tamanho pelo experimento da intrusão de mercúrio.

No âmbito de processamento de imagens vários métodos exitem que permitem analisar formatos. Dentre os primeiros utilizados para a análise de imagens encontram-se métodos baseados em operadores de morfologia matemática que, quando usados corretamente, permitem analisar formatos preservando sua estrutura básica mas eliminando irrelevâncias (HARALICK; STERNBERG; ZHUANG, 1987).

Os operadores mais básicos são a Erosão e a Dilatação, que atuam sobre um dos valores binários considerado “valido”. Neles uma mascara é aplicada numa imagem binaria pixel a pixel produzindo

outra imagem resultante que pode ter adicionado ou removido pixeis de acordo com a sua vizinhança. A Erosão apenas adiciona o pixel central se, e somente se, a mascara foi preenchida por inteiro com pixeis "validos", enquanto a Dilatação adiciona todos os pixeis da sua mascara na imagem resultante se o seu pixel central foi considerado "valido". A aplicação de uma Erosão seguida de uma Dilatação quando ambos operadores utilizam uma máscara circular definida por um raio r é um processo conhecido como Abertura de raio r, que pode ser observado na figura 6.

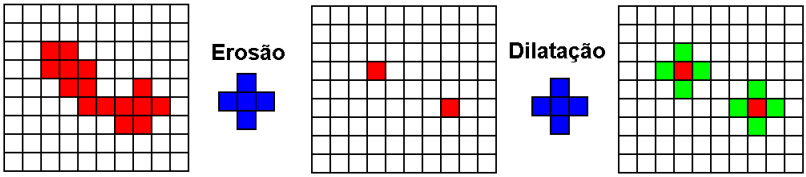


Figura 6: Abertura (Erosão seguida de dilatação).

É intuitivo perceber que aplicar uma Abertura de raio r ira remover parte do material, analogamente aos grãos sendo filtrados por uma peneira cuja malha tem tamanho r, exceto que ao invés de se medir o material filtrado contabiliza-se o total de voxels vazios antes e depois da operação normalizado pela porosidade da amostra, em outras palavras:

$$F(r) = \frac{\phi - \phi_{abertura(r)}}{\phi} \tag{2.9}$$

O processo pode ser repetido sobre a imagem original aumentando o tamanho do raio r progressivamente para obter o volume da imagem ocupado por poros de raio menor ou igual a r (VINCENT, 1994).

Um exemplo de dois passos com raio 1 e 2 podem ser vistos na figura 7, onde a porosidade da imagem é mostrada em azul, e os voxels que eram vazio no passo anterior mas foram removidos estão em vermelho.

Se o processo é repetido até a imagem ficar sem porosidade, obtém-se a probabilidade de que um poro aleatório tenha raio igual ou inferior a r. Esta distribuição é conhecida como probabilidade acumulada, e é possível converter ela em uma distribuição de probabilidade padrão subtraindo o valor do passo anterior para cada

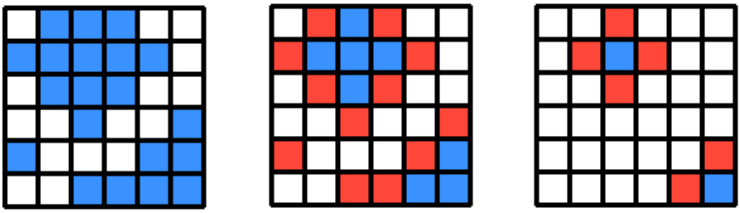


Figura 7: Resultado de Aberturas sucessivas.

elemento. Em posse desta função de probabilidade é trivial encontrar o tamanho r que tem a maior probabilidade de representar um poro aleatório, bastando procurar o valor de r para a maior probabilidade da distribuição. Um exemplo desta conversão de uma distribuição de probabilidade acumulada para uma distribuição de probabilidade pode ser vista na figura 8.

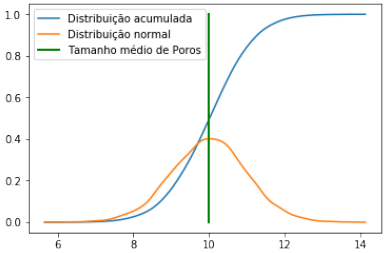


Figura 8: Distribuição de probabilidade acumulada vs Distribuição de probabilidade.

2.3.5 Raio hidráulico

O raio hidráulico é uma medição que correlaciona a área de poros com a superfície deles, dando portanto uma ideia sobre a rugosidade deles. Esta medição pode ser facilmente obtida para as imagens binarizadas contando quantos voxels pertencem a fase de poros, e quantos deles estão diretamente conectados com um voxel que não pertence a mesma fase.

Desta forma pode-se generalizar todas as mensurações de raio hidráulico como:

$$R_H = \frac{A}{P} \tag{2.10}$$

Onde A representa a área total de poros de uma dada imagem e P o perímetro total deles.

2.4 ANALISE DE CLUSTERS

A seguinte seção apresenta alguns dos métodos mais utilizados na analise de clusters, bem como suas vantagens e desvantagens para o caso do presente trabalho.

Métodos para clusterização maioritariamente pertencem a 3 categorias: Aglomeração, Divisão ou Atribuição de pontos.

Estes métodos podem representar clusters através dos seus centroides (a média dos valores dos pontos que pertencem ao cluster) ou clustoides (o ponto do cluster que melhor o representa).

2.4.1 K-Means

O algoritmo de K-Means é o algoritmo de particionamento mais popular atualmente (YADAV; SHARMA, 2013), nele cada cluster é representado pela media dos valores dos objetos pertencentes ao cluster.

O algoritmo busca particionar um grupo de n valores em k agrupamentos, tentando maximizar a similaridade intra-cluster e minimizar a similaridade inter-cluster, esta similaridade é medida a partir da media dos objetos de um cluster, i.e. diminuir o desvio padrão da media dos valores pertencentes ao mesmo cluster enquanto tenta ao mesmo tempo aumentar a diferença entre o valor médio de clusters diferentes.

A base do algoritmo de K-Means pode ser resumido como:

1. k centroides são escolhidos aleatoriamente dentre os dados disponíveis
2. Cada objeto a ser avaliado é classificado como pertencendo a um desses centroides de acordo com sua semelhança
3. As medias dos k centroides são calculadas utilizando os objetos classificados em cada um deles

4. Estas medias são usadas como centroides para o passo 2, até que os centroides de saída do passo 3 sejam iguais aos de entrada

O algoritmo é amplamente utilizado pela sua simplicidade e confiabilidade, porém existem certas desvantagens nele, dentre as principais destaca-se o fato da aleatoriedade da escolha inicial poder afetar o resultado final, sendo necessário executar o algoritmo diversas vezes sobre o mesmo grupo de dados para garantir que se obteve o melhor resultado; A alta influencia que pontos anormais tem sobre os dados; E principalmente a alta dependência do numero de clusters. Este ultimo ponto é altamente importante, porque parte do principio de que se conhece a quantidade de agrupamentos que se deseja obter, e o algoritmo irá forçar esta quantidade de agrupamentos nos dados, isto é aplicar o algoritmo de K-Means sobre um conjunto de pontos aleatórios com qualquer valor de k, irá retornar k agrupamentos diferentes, este é mais um motivo para se querer executar o algoritmo diversas vezes para garantir que os clusters extraídos sejam coerentes.

Alguns algoritmos adicionais existem para tentar prever a quantidade de clusters, calculando o algoritmo iterativamente e aumentando o numero de clusters enquanto os resultados forem melhorando, i.e. menor distancia intra-cluster e maior inter-cluster. Porém lembrando que para cada k o algoritmo deve ser executado dezenas senão centenas de vezes para garantir que o resultado seja de fato confiável.

Existem ainda mais modificações deste algoritmo, que visam compensar por alguns de seus problemas inerentes, porém estas pequenas diferenças estão fora do escopo para este trabalho.

2.4.2 Propagação de afinidade

O algoritmo de propagação de afinidade considera cada ponto dos dados como um nodo em uma rede onde nodos da borda transmitem informações sobre sua afinidade recursivamente até encontrar um bom conjunto de pontos, e portanto clusters (FREY; DUECK, 2007).

A informação transmitida pelos nodos para representar afinidade é o negativo da distancia euclidiana entre dois pontos, e.g. para o caso de duas dimensões x e y a similaridade entre dois pontos i e j é dada por:

$$s(i,j) = -|\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}| \tag{2.11}$$

Cada nodo da borda escolhe o nodo mais próximo para ser seu pai. O processo se repete para os nodos a distancia 1 da borda, porém desta vez caso o nodo escolhido como pai seja um nodo folha, i.e. possua um pai próprio, ambos devem ser atualizados procurando o clustoide mais próximo que não tenha um pai, de forma que qualquer nodo possa ser classificado como um nodo folha que esta ligado a um pai mas não tem filhos, ou um nodo pai que tem vários nodos filhos mas não tem um nodo pai próprio. O processo se repete até que todos os nodos sejam ou folhas ou pais, e então cada um dos nodos pais é tomado como representante de um cluster.

Um exemplo de 3 passos deste algoritmo pode ser observado na figura 9, onde estão representados as conexões iniciais, logo a escolha de cada nodo (em verde) do seu pai, e um ultimo passo de ajuste onde são escolhidos quais clustoides serão utilizados (em vermelho). Nesta ultima etapa é possível observar nodos que já decidiram que não são centroides (em azul), desta forma eles são ignorados pelos nodos que procuram um pai.

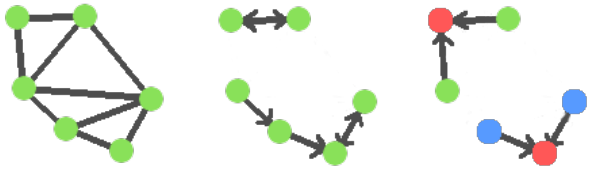


Figura 9: Exemplo de clusterização por propagação de afinidade

A grande vantagem do método de propagação por afinidade é a não dependência de conhecimento prévio do numero de clusters no sistema, além de ser muito menos sensível a inicialização já que analisa todos os nodos simultaneamente, e portanto melhor para grandes conjuntos de dados onde outros algoritmos como K-Means devem ser executados repetidas vezes para garantir que a escolha dos pontos aleatórios no começo conseguiram chegar na solução ótima. Nenhum destes é uma grande vantagem para este trabalho, já que o numero de clusters esperados é igual aos diferentes tipos de rochas analisadas, e o volume de dados é suficientemente pequeno para poder ser executado centenas de vezes nos outros algoritmos, garantindo a solução ótima deles. Porém, já que o numero de clusters não é predefinido, este algoritmo poderia encontrar subgrupos dentro de um

tipo de rochas que não seriam encontrados por outros.

Porém ele também é o algoritmo mais lento dos avaliados por vários fatores de escala.

2.4.3 Deslocamento de média

Deslocamento de média, ou Mean Shift em inglês, é um algoritmo de clusterização de dados que consiste em iterativamente deslocar cada ponto dos dados para a média dos pontos na sua vizinhança (CHENG, 1995).

Uma das vantagens deste método em relação a outros é a não dependência do conhecimento prévio da quantidade de grupos estimados, porém ele é bastante sensível à escolha da largura de banda, i.e. o tamanho da vizinhança a ser avaliada.

1. Cada ponto dos dados é usado como centroide do seu próprio cluster.
2. Para cada um dos centroides procuram-se todos os pontos dentro da vizinhança.
3. O novo centroide para este grupo é calculado como a média dos pontos obtidos no passo anterior.
4. Centroides repetidos são eliminados.
5. Verifica-se se a nova lista de centroides mudou da recebida no passo 2. Se sim, volta ao passo 2, senão terminou.

2.5 FERRAMENTAS AUXILIARES

Sempre que possível foi preferido a escolha de ferramentas conhecidas e validadas, que permitiram não somente agilizar as medições como também oferecem uma garantia maior dos nossos resultados e aumentar amplamente a quantidade de algoritmos que foi possível testar e apresentar.

2.6 IMAGO3D

As analises das imagens tomográficas foram feitas com o software Imago3D desenvolvido pelo Laboratório de Meios Porosos e

Propriedades Termo-físicas (LMPT), o qual ofereceu algumas das medições mais básicas, como porosidade ou tamanho médio de poros, e apresentou uma boa plataforma para se desenvolver as demais medidas, como ser autocorrelação e raio hidráulico.

2.7 PYTHON

A linguagem python é uma das maiores linguagens utilizadas para ciência de dados atualmente, sendo a mais amigável e a que possui a maior quantidade de recursos disponíveis. Todos as análises de clusters utilizados neste trabalho foram avaliadas utilizando a biblioteca scikit-learn (PEDREGOSA et al., 2011), o qual permitiu realizar diversas comparações entre os vários métodos visando escolher o mais eficaz para o conjunto de dados utilizado.

3 TRABALHOS CORRELATOS

(MIRABOLGHASEMI et al., 2015) utilizaram simulação de partículas em um meio poroso virtual obtido através de imagens, de CT, e comparou os resultados com medições experimentais para determinar a capacidade de se avaliar propriedades físicas através da estrutura tomografada. É importante destacar neste trabalho que ele não realizou nenhuma medição da imagem tomográfica, mas o fato de ter conseguido produzir resultados similares na simulação de partículas nestas imagens tomográficas serve para demonstrar a fidelidade das imagens CT tem com o objeto de estudo, e como o estudo da estrutura interna através deste tipo de imagens pode ser correlacionado com propriedades físicas.

(KATTAN; JAWAD; JOMAAH, 2018) utilizaram métodos de analise de Clusters para classificar rochas baseado em informações extraídas de logs de poços dentro de grupos conhecidos relacionados com a qualidade da rocha. É interessante perceber que neste trabalho os autores utilizaram apenas medidas da porosidade efetiva e da saturação de água para classificar as rochas, e ainda assim conseguiram uma classificação satisfatória.

(CHAUHAN et al., 2016) utilizaram vários métodos de aprendizado de maquina para extrair segmentar imagens tomográficas, e comparou os resultados obtidos através de métodos não supervisionados, supervisionados, e de conjuntos. Dos seus resultados é possível destacar que os métodos não supervisionados, como os que serão utilizados neste trabalho, normalmente são muito sensíveis a parametrização, porém apresentam resultados comparáveis sem a necessidade de sujeitar os resultados ao bias do supervisionamento.

(CHENG; GUO, 2017) utilizou métodos de aprendizado de maquina para classificar rochas através da analise de imagens de laminas delgadas. O método proposto apresenta uma acurácia alta (98.5%), porém por utilizar imagens de laminas delgadas a informação de interconectividade tridimensional da rocha é perdida no processo de captura, sem mencionar a destruição da rocha original para a obtenção das laminas para serem avaliadas.

(SANTOS, 2016) propôs métodos para classificar rochas através de dados de perfil ao invés de imagens tomográficas. Alem de demonstrar que é possível classificar corretamente rochas utilizando apenas dados extraídos de medições digitais destas, este trabalho

também demonstra que nem sempre aumentar o numero de variáveis avaliadas vai melhorar a classificação do sistema, e que com o mesmo dado de entrada algoritmos de classificação podem apresentar resultados muito diferentes. Neste trabalho é interessante observar que a quantidade de variáveis é pequena, mas ainda assim os algoritmos conseguem um grau de acerto elevado quando corretamente parametrizados.

(PATEL; CHATTERJEE, 2016) utilizou redes neurais para classificar rochas com propriedades heterogêneas e obteve um grau de erro de 5-6%. Porém o método proposto classifica rochas por inteiro, não considerando que uma mesma rocha pode sofrer alterações e apresentar camadas com características diferentes entre si, mas similares com certas camadas de outra rocha.

3.1 COMPARAÇÃO

Nesta seção é apresentada uma tabela comparativa entre dos trabalhos correlatos e o presente trabalho.

Trabalho	Clusterização	Tomografia	Analise 3D	Extração de parâmetros
Presente trabalho	X	X	X	X
Mirabolghasemi	-	X	X	-
Kattan	X	-	X	X
Chauhan	-	X	X	-
Cheng	-	-	-	X
Santos	X	-	X	-
Patel	-	X	X	X

Tabela 1: Comparação dos trabalhos correlatos

4 MATERIAIS E MÉTODOS

4.1 VISÃO GERAL

Um total de 12 amostras foram preparadas, binarizadas e rotuladas. Medições gerais foram feitas para cada camada ao longo do eixo Z, i.e o maior eixo da amostra. Estas amostras representam dados de diversos tipos de rochas diferentes.

Estas medições foram exportadas em diversos arquivos de texto onde cada linha do arquivo representa uma camada da amostra e os diversos valores extraídos estão separados uns dos outros por comas, seguindo o padrão do formato de arquivos CSV, por isso estes arquivos são referidos no resto do trabalho simplesmente como arquivos csv. Neste momento algumas das amostras foram separadas aleatoriamente para utilização na validação dos dados.

Os demais arquivos csv foram compilados em outro arquivo e embaralhados para garantir que não existisse uma preferencia pela ordem dos dados. Este arquivo foi utilizado como treinamento para diversos algoritmos de clusterização, e o modelo calculado por este treinamento foi utilizado para predizer o grupo apropriado para cada uma das camadas das amostras separadas anteriormente. Este resultado foi então comparado com o esperado, i.e. que os dados separados se comportem da mesma forma que a rocha da qual eles foram extraídos.

4.2 AMOSTRAS CILÍNDRICAS

Algumas imagens tomográficas referem-se a amostras cilíndricas, e portanto utilizá-las por inteiro produziria erros, já que a imagem apresentaria, do ponto de vista do algoritmo, um grande poro rodeando toda a amostra. Alem disto algumas destas amostras possuem envolucro protetor que não faz parte do material da amostra, esta "camisa" é indistinguível por métodos de limiar pois sua densidade é comparável com a da amostra no seu interior, desta forma devem ser usados métodos de analise de padrões para conseguir removê-la.

Foi desenvolvido um algoritmo capaz de extrair apenas o cilindro de interesse da amostra, e posteriormente extrair o paralelepípedo inscrito no cilindro para garantir que a imagem

resultante seja totalmente composta de voxels relevantes para a extração de dados. Porém já que nem todas as amostras são cilíndricas, foi necessário escolher manualmente em quais amostras este processo seria aplicado primeiro.

Este algoritmo é composto de 3 etapas:

1. Alinhamento
2. Removedor de camisa
3. Paralelepípedo inscrito

4.2.1 Alinhamento

Devido ao fato do posicionamento da amostra no tomografo ser realizado por humanos esta propenso a erros, estes erros, ainda que normalmente imperceptíveis, podem atrapalhar a detecção correta do cilindro pois o eixo Z da tomografia não condiz perfeitamente com o eixo do cilindro. Não foi possível encontrar algoritmos na literatura que fossem desenhados especialmente para o alinhamento de cilindros.

A forma de se alinhar os dados é encontrar a reta que determina a inclinação do cilindro no espaço e deslocar progressivamente cada camada pela posição desta reta no valor de Z correspondente para cada fatia da imagem. Uma representação gráfica disto pode ser observada na Figura 10, onde a reta vertical a direita representa o alinhamento original dos dados, as linhas horizontais representam as fatias da imagem CT com a parte grossa nelas representando a porção da imagem que contem dados, a reta inclinada a esquerda representa a inclinação do cilindro e o retângulo a imagem resultante do alinhamento.

Nesta figura também é possível observar que certos valores da imagem original serão cortados pelo retângulo resultante, porém como estes valores tratam-se de pontos fora da amostra não causa problemas para o restante das medições. Mas é importante perceber que a imagem resultante será sempre menor que a original.

Para encontrar esta reta duas abordagens foram testadas, a primeira envolve a utilização de detecção de círculos utilizando uma transformada de Hough (YUEN et al., 1990), a partir desta detecção extraem-se os centros dos círculos e uma regressão linear é utilizada para calcular a reta que melhor se ajusta a inclinação do cilindro.

Porém a transformada de Hough é custosa computacionalmente e altamente dependente de parâmetros, e tendo em vista que os dados

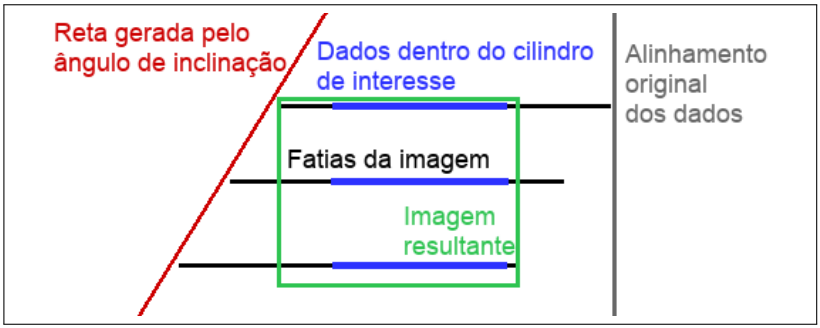


Figura 10: Representação das camadas de uma imagem cilíndrica não-alinhada

são praticamente círculos perfeitos a cada camada algumas simplificações podem ser tomadas. Primeiramente encontra-se o centroide dos dados de cada camada, mas este centroide poderia ser movimentado se a amostra possuísse porosidade elevada em uma determinada camada. Uma abordagem mais simples que produziu resultados semelhantes em segundos, em comparação com 5 a 10 minutos da abordagem descrita anteriormente, foi utilizar um histograma horizontal e vertical para encontrar o ponto onde os dados começam e onde terminam, o ponto médio entre estes dois é considerado o centroide daquela fatia. O resultado deste algoritmo, bem como os histogramas encontrados para uma fatia podem ser observados na Figura 11.

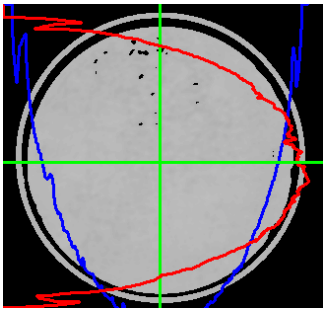


Figura 11: Centro do cilindro encontrado por histograma

Este algoritmo também poderia falhar caso a amostra apresentasse um poro suficientemente grande por diversas camadas,

porém a existência da camisa pode ser usada ao nosso favor, garantindo que sempre existirão dados que delimitem a amostra.

Uma vez em posse destes centros o mesmo algoritmo de regressão linear pode ser aplicado para determinar a reta pela qual cada fatia deve ser deslocada.

4.2.2 Removedor de camisa

Na maioria das fatias é possível visualizar a camisa como sendo um objeto externo a amostra, porém em algumas delas a amostra esta quase totalmente apoiada sobre a camisa dificultando a sua extração. Mas se parte do pressuposto de que o dado esta alinhado é possível então detectar a camisa em apenas uma das fatias e utilizar o resultado para toda a imagem.

Esta detecção pode ser realizada com uma transformada de Hough estimando parâmetros, desta vez porém a transformada é aplicada em apenas uma das fatias da imagem, diminuindo assim o custo computacional. Para tanto estima-se os parâmetros e corrigem-se algumas vezes até conseguir extrair pelo menos dois círculos concêntricos, destes o menor é selecionado para representar o total dos dados da amostra. Se depois de algumas iterações é impossível encontrar estes circulo outra fatia da imagem é selecionada.

4.2.3 Paralelepípedo inscrito

Uma vez em pose do circulo que representa os dados, extrair o retângulo inscrito nele é trivial, basta apenas recortar a imagem desde $centro - \frac{raio}{2}$ até $centro + \frac{raio}{2}$. Em relação ao eixo Z, procura-se a primeira e a ultima fatia que não sejam totalmente vazias.

4.3 EXTRAÇÃO DE PARÂMETROS

As imagens tomográficas foram alimentadas para um programa escrito encima da plataforma do Imago que permitiu utilizar as suas funções para vários cálculos, assim como permitir que as demais pudessem ser facilmente implementadas. Cada imagem foi binarizada e rotulada levando em consideração as 3 dimensões, depois para cada camada ao longo do eixo Z destas imagens os parâmetros de Porosidade, Autocorrelação em X e Y, Conectividade em X e Y,

tamanho médio de poro e raio hidráulico foram extraídos e armazenados em um csv. Desta forma cada imagem tomográfica produziu um arquivo com onde cada linha representa uma camada e cada coluna um valor extraído.

Para os valores de Autocorrelação e Conectividade, onde o resultado do algoritmo é uma distribuição estatística, o valor armazenado é o valor da distância em que o comportamento passa a ser aleatório, como definido nas suas respetivas subseções.

4.4 PREPARAÇÃO DOS DADOS

Uma vez em pose destes arquivos 10% das linhas de cada um deles foram extraídas aleatoriamente em um único arquivo csv para a validação dos resultados, o restante 90% foram extraídos para outro arquivo csv que foi utilizado na etapa de treinamento dos métodos de clusterização.

Ambos arquivos foram embaralhados para prevenir qualquer tipo de preferencia de qualquer um dos algoritmos por agrupamentos de características semelhantes devido a sua proximidade no dado de entrada. Já que os dados foram embaralhados foi preciso manter um registro indicando a qual amostra pertencia originalmente cada linha dos arquivos csv, este registro não foi utilizado na caracterização de clusters e apenas serviu para comparar os agrupamentos de uma mesma amostra tanto nos dados de treinamento quanto nos de validação.

4.5 ANALISE DE CLUSTERS

Os arquivos foram avaliados por vários algoritmos de clusterização, para cada algoritmo os seguintes passos foram realizados:

1. Clusterização do arquivo com a maioria dos dados.

2. Fitting do arquivo com os dados de validação.

3. Comparação do arquivo de registro com o Fitting.

4.5.1 Clusterização

Cada um dos algoritmos descritos anteriormente foi aplicado nos dados, aqueles que dependem de uma inicialização aleatória como o K-Means foram executados diversas vezes para tentar encontrar a melhor solução possível.

Ainda alguns algoritmos dependem de uma escolha inicial do numero de clusters, para a determinação deste numero inicia-se com o numero de arquivos originais de entrada, e iterativamente foi diminuído avaliando os resultados de forma a tentar se obter o valor ótimo para esse parâmetro, i.e. o valor que melhor classificaria os dados de fitting dentro dos mesmos conjuntos que os dados clusterizados. Esta abordagem foi escolhida para a determinação desse parâmetro por apresentar uma forma objetiva de medir a sua eficacia, já que a alternativa seria o agrupamento subjetivo das imagens de entrada em categorias de rochas.

Logo para cada arquivo utilizado verifica-se quantas linhas foram classificados como cada cluster, e este resultado é armazenado em um arquivo de saída. Idealmente seria esperado que todas as fatias de uma mesma amostra sejam classificadas como o mesmo grupo, demonstrando a homogeneidade das amostras, e que fatias de amostras diferentes sejam classificadas em grupos distintos, implicando que as amostras são diferentes entre si.

4.5.2 Fitting

É importante que os dados separados sejam similares aos utilizados para a clusterização, para isto histogramas dos parâmetros de ambos os grupos são comparados.

Nesta etapa os dados separados para validação foram adicionados ao modelo de cluster calculado anteriormente, porém não permitindo alterar os centroides/clustoides já encontrados. Desta forma cada um dos dados foi encaixado no cluster mais próximo a ele.

Assim como anteriormente para cada arquivo contabiliza-se a quantidade de classificações em cada cluster.

4.5.3 Comparação dos arquivos

O arquivo resultante do Fitting é comparado com o arquivo resultante da clusterização, espera-se que estes arquivos sejam similares. Com a pressuposição de que os dados de fitting são estatisticamente similares aos dados da amostra original, pode-se assumir que as diferenças entre a classificação da clusterização e do fitting deva-se a erros dos grupos reconhecidos pelo algoritmo de clusterização.

Para que a comparação possa ser feita diretamente, tendo em vista que os conjuntos de dados tem tamanhos diferentes, converteu-se as quantias de ambos os arquivos para percentagens do total de dados de entrada. Desta forma uma tabela pode ser diretamente subtraída da outra sobrando apenas as diferenças de classificação.

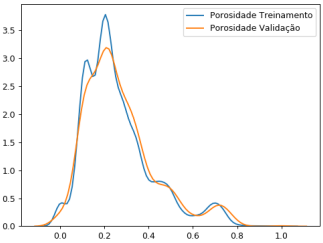
5 RESULTADOS

Para a analise dos resultados os valores de autocorrelação e conectividade, calculados em X e Y individualmente, foram agrupados usando média aritmética em um único valor por categoria. Esta decisão foi tomada pois, diferentemente de eixo Z, os eixos X e Y não tem nenhuma diferença física, e a mesma rocha poderia ter sido posicionada de forma diferente no tomografo e produzido eixos diferentes.

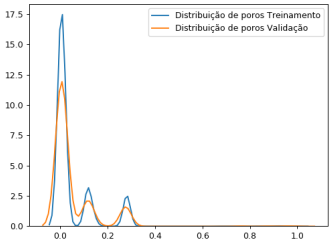
5.1 ANALISES

Nesta sessão os dados são analisados por conta própria, antes da aplicação dos métodos de clusterização. Estas analises consistem na observação das variáveis individualmente e em pares, maiores dimensões não serão avaliadas nesta sessão pois a sua visualização é extremamente difícil em um gráfico, mas as conclusões das análises realizadas em apenas duas dimensões estendem-se a qualquer numero. É necessário esclarecer também que mesmo que o trabalho não apresente gráficos destas dimensionalidades os algoritmos de clusterização trabalham em todas as dimensões das variáveis estudadas.

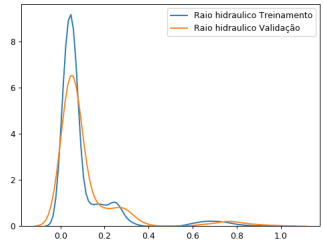
É possível estimar se os dados separados para validação dos algoritmos de clusters representam uma boa seleção aleatória dos dados originais calculando os histogramas para cada uma das variáveis analisadas tanto para os dados de treinamento quanto para os dados de validação, uma boa seleção aleatória apresentara histogramas similares para ambos os conjuntos. Como pode ser observado nas figuras 12-a até 12-e os histogramas para os dois conjuntos utilizados neste trabalho são bastante similares para todas as variáveis analisadas no trabalho. Desta forma pode-se concluir que a seleção utilizada para validação foi realmente aleatória e suficientemente grande para ser representativa dos dados utilizados para o treinamento.



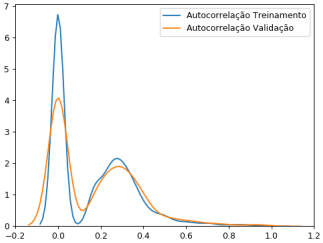
(a) Porosidade



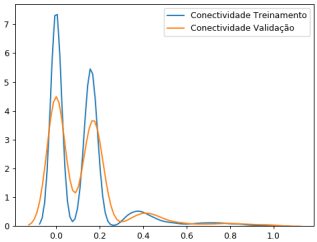
(b) Tamanho de poro



(c) Raio hidráulico



(d) Autocorrelação

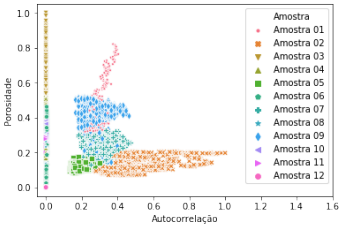


(e) Conectividade

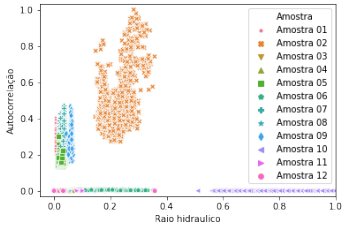
Figura 12: Histograma das variáveis analisadas

Alem disso é possível analisar os dados plotando uma variável em relação a outra, assim obtendo um mapeamento bidimensional do relacionamento entre as duas variáveis e utilizando uma cor para cada ponto que representa a amostra a qual esse dado pertence é possível ter uma ideia de como estas variáveis se relacionam entre diferentes amostras, esta técnica é conhecida como gráfico de pares. De todas as

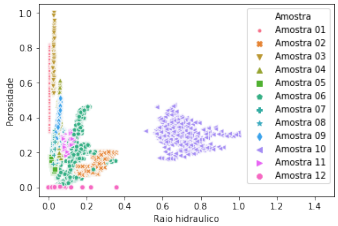
possíveis combinações, 4 foram selecionadas pela sua representatividade e estão apresentadas na Figura 13.



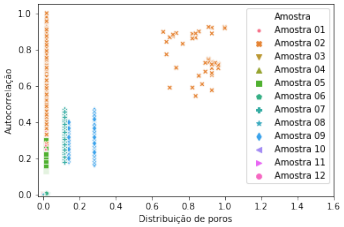
(a) Autocorrelação x Porosidade



(b) Raio hidráulico x Autocorrelação



(c) Raio Hidráulico x Porosidade



(d) Distribuição de tamanho de poro x Autocorrelação

Figura 13: Gráficos de pares de variáveis selecionadas

No gráfico 13.a observa-se a relação entre a Autocorrelação e a Porosidade, é possível observar por exemplo que a Amostra 01 apresenta um comportamento quase inteiramente vertical, implicando que existe pouca variação de autocorrelação para toda a variação de porosidade da amostra, enquanto que a Amostra 02 apresenta o comportamento contrario, isto é pouca variação de porosidade para toda a variação de autocorrelação. Isto implica que talvez seja possível diferenciar se um dado ponto pertence a Amostra 01 ou 02 apenas verificando os valores de Porosidade e Autocorrelação desta imagem.s

No gráfico 13.b observa-se a relação entre o Raio Hidráulico e a Autocorrelação, é interessante destacar que Amostra 02 pode ser inteiramente isolada utilizando apenas estes dois valores, enquanto que diversos pontos desta amostra sobrepunham-se no gráfico anterior, demonstrando a necessidade de utilizar mais valores para diferenciar amostras diferentes.

No gráfico 13.c observa-se a relação entre o Raio Hidráulico e a

Porosidade, assim como no gráfico anterior, neste também é possível observar uma amostra que pode ser inteiramente isolada, neste caso trata-se da Amostra 10, demonstrando que diferentes comparações de medições podem permitir separar grupos que outras medições considerariam iguais. Já que por exemplo a Amostra 02, tão claramente separada no gráfico 13.b, é indistinguível das outras neste gráfico.

Finalmente no gráfico 13.d observa-se a relação entre a Distribuição de tamanho de poro e a Autocorrelação, neste gráfico é possível separar as Amostras 09 e 07 do restante delas, mas não entre si.

Estas analises de duas em duas variáveis mostram que as amostras apresentam diferenças entre si que podem ser observadas apenas quando mais de uma dimensão de variáveis são analisadas em conjunto, e levam a conclusão de que um numero superior de dimensões pode ser capaz de encontrar ainda mais diferenças entre amostras que de outra forma não parecem poder ser diferenciáveis.

5.2 CLUSTERS

Para cada um dos métodos de clusters avaliados duas tabelas são apresentadas, onde as linhas representam as amostras de origem, e as colunas o percentual de fatias da amostra que foi classificado como um grupo em especifico, de forma que a soma de percentagens para cada linha seja 100%. Todos os valores são dados em percentagem do total de dados, avaliados para cada amostra, para normalizar as diferenças entre amostras com diferente quantidade de fatias dentro de uma mesma tabela, e para facilitar a comparação entre tabelas diferentes, já que os dados de fitting representam aproximadamente uma ordem de magnitude a menos que os utilizados para o treinamento, sendo 10% e 90% dos dados originais respectivamente.

A primeira tabela representa a clusterização realizada sobre os dados de treinamento, a segunda o fitting feito sobre os dados de validação. Desta forma as tabelas nomeadas “Clusters por” apresentam a percentagem de fatias de cada amostra dos dados de treinamento que foram classificadas como cada cluster, enquanto que as tabelas nomeadas “Validação por” apresentam a percentagem de fatias de cada amostra dos dados de validação que foram classificadas como cada cluster.

Idealmente as tabelas teriam 100% em uma única celula e vazio

nas outras, implicando que todas as fatias de uma determinada amostra foram classificadas como pertencendo ao mesmo grupo. E cada amostra seria classificada como um grupo diferente, implicando que o algoritmo conseguiu distinguir perfeitamente todas as fatias das amostras.

Realisticamente mesmo que as amostras tenham sido escolhidas por sua homogeneidade ainda apresentam diferenças significativas entre suas camadas, e é esperado que as amostras tenham algum erro que tenha sido detectado como outro grupo. Além disto a similaridade entre algumas amostras faz com que seja também esperado que os algoritmos classifiquem ambas amostras no mesmo grupo.

Ainda mais, a tabela dos resultados da validação é esperada ser parecida com a tabela dos dados de treinamento, não só demonstrando que o percentual de dados separados foi suficientemente similar, mas que o algoritmo de classificação comporta-se consistentemente com o seu treinamento. Se considerarmos que as amostras separadas são suficientemente similares, como demonstrado pela analise individual das variáveis demonstrada em 12, podemos assumir que toda a diferença entre estas duas tabelas se deve a erro de classificação do algoritmo, portanto na seção 5.3 é apresentado uma tabela de erros para todos os algoritmos, esta tabela nada mais é do que a diferença entre as duas tabelas apresentadas para cada algoritmo.

5.2.1 K-Means

K-Means é um dos algoritmos mais rápidos avaliados neste trabalho, porem ele depende de inicialização, por isso foi executado 100 vezes tanto nos dados de treinamento quanto nos de validação.

Por isso os valores percentuais exibidos nas tabelas 2 ate 5 foram obtidos através do calculo da moda para o resultado de cada linha dos dados.

Esta primeira analise utiliza o numero de amostras como o numero K de clusters a tentar ser encontrado. A grande similaridade entre as duas tabelas prova a robustez deste algoritmo mesmo com esta parametrização ingenua que não considera que amostras possam ser similares.

É interessante observar que certas amostras foram classificadas majoritariamente como pertencendo ao mesmo grupo, enquanto outras foram classificadas inteiramente como seu próprio grupo. Quais amostras são classificadas em conjunto com outras varia a

Cluster Amostra	1	2	3	4	5	6	7	8	9	10	11
Amostra 01	4.16%	95.19%	-	-	-	-	-	-	-	-	0.65%
Amostra 02	-	98.97%	-	-	1.03%	-	-	-	-	-	-
Amostra 03	-	100.00%	-	-	-	-	-	-	-	-	-
Amostra 04	-	-	-	-	-	-	100.00%	-	-	-	-
Amostra 05	-	96.77%	-	-	-	-	0.98%	-	-	2.25%	-
Amostra 06	0.76%	60.87%	-	-	0.43%	-	-	-	32.28%	-	5.65%
Amostra 07	-	-	96.77%	-	-	-	-	3.23%	-	-	-
Amostra 08	-	-	-	-	-	-	-	-	-	87.05%	12.95%
Amostra 09	-	9.99%	-	-	-	80.36%	-	-	-	9.66%	-
Amostra 10	-	-	-	-	-	-	-	-	100.00%	-	-
Amostra 11	-	-	-	-	100.00%	-	-	-	-	-	-
Amostra 12	-	-	-	100.00%	-	-	-	-	-	-	-

Tabela 2: Clusters por K-Means

Cluster Amostra	1	2	3	4	5	6	7	8	9	10	11
Amostra 01	1.18%	98.82%	-	-	-	-	-	-	-	-	-
Amostra 02	-	99.02%	-	-	0.98%	-	-	-	-	-	-
Amostra 03	-	100.00%	-	-	-	-	-	-	-	-	-
Amostra 04	-	-	-	-	-	-	100.00%	-	-	-	-
Amostra 05	-	94.94%	-	-	-	-	0.63%	-	-	4.43%	-
Amostra 06	-	63.73%	-	-	2.94%	-	-	-	28.43%	-	4.90%
Amostra 07	-	-	95.76%	-	-	-	-	4.24%	-	-	-
Amostra 08	-	-	-	-	-	-	-	-	-	87.13%	12.87%
Amostra 09	-	7.14%	-	-	-	80.36%	-	-	-	12.50%	-
Amostra 10	-	-	-	-	-	-	-	-	100.00%	-	-
Amostra 11	-	-	-	-	100.00%	-	-	-	-	-	-
Amostra 12	-	-	-	100.00%	-	-	-	-	-	-	-

Tabela 3: Validação por K-Means

depender do algoritmo utilizado ou até mesmo dos parâmetros do algoritmo, no caso para a parametrização ingenua do K-Means as amostras 01, 02, 03, 05 e 06 foram classificadas como o grupo 2, mas destas a amostra 06 apresenta um comportamento dividido entre vários grupos, enquanto o restante das amostras parece ter sido quase inteiramente classificado como um único grupo ao qual poucas outras fatias de outras amostras pertencem.

Em vista deste junção das amostras é possível entender que certas amostras são suficientemente similares para serem consideradas o mesmo grupo, e que talvez uma abordagem com um numero K de clusters menor do que as amostras produza melhores resultados. Tendo em vista isto, diversos valores para K foram testados e com isto foi possível obter um resultado melhor que é representado nas tabelas abaixo.

Nas tabelas anteriores o grupo 4 foi omitido por ter um numero insignificante de camadas, i.e. menor que 0.01% das fatias das imagens pertenceram a este grupo. Porém uma escolha de $K = 8$ produziu resultados piores, implicando que estes dados separados neste grupo são anomalias suficientemente grandes para atrapalhar o reconhecimento de

Cluster Amostra	1	2	3	5	6	7	8
Amostra 01	100.00%	-	-	-	-	-	-
Amostra 02	98.81%	-	-	1.19%	-	-	-
Amostra 03	100.00%	-	-	-	-	-	-
Amostra 04	-	-	-	-	100.00%	-	-
Amostra 05	-	97.54%	-	-	2.46%	-	-
Amostra 06	96.20%	-	-	3.80%	-	-	-
Amostra 07	-	96.77%	-	-	-	-	3.23%
Amostra 08	-	90.34%	-	-	-	9.66%	-
Amostra 09	-	19.11%	-	-	0.33%	80.56%	-
Amostra 10	100.00%	-	-	-	-	-	-
Amostra 11	-	-	-	100.00%	-	-	-
Amostra 12	-	-	100.00%	-	-	-	-

Tabela 4: Clusters por K-means 9 clusters

Cluster Amostra	1	2	3	5	6	7	8
Amostra 01	100.00%	-	-	-	-	-	-
Amostra 02	98.54%	-	-	1.46%	-	-	-
Amostra 03	100.00%	-	-	-	-	-	-
Amostra 04	-	-	-	-	100.00%	-	-
Amostra 05	-	98.10%	-	-	1.90%	-	-
Amostra 06	95.10%	-	-	4.90%	-	-	-
Amostra 07	-	95.76%	-	-	-	-	4.24%
Amostra 08	-	93.07%	-	-	-	6.93%	-
Amostra 09	-	18.45%	-	-	-	81.55%	-
Amostra 10	100.00%	-	-	-	-	-	-
Amostra 11	-	-	-	100.00%	-	-	-
Amostra 12	-	-	100.00%	-	-	-	-

Tabela 5: Validação por K-means 9 clusters

outras fatias caso sejam incorporados em um dos outros grupos.

Os resultados para um numero de clusters menores se provam bastante promissores, obtendo uma margem de erro de aproximadamente a metade que utilizando a mesma quantidade de amostras, implicando que algumas das amostras são suficientemente similares a outras para serem classificadas igualmente. É interessante observar quais amostras foram consideradas similares entre si, no caso as amostras 01, 02, 03, 06 e 10 foram quase que totalmente classificadas como grupo 1, e as amostras 05, 07 e 08 no grupo 2, enquanto as amostras 04, 09, 11, e 12 foram classificadas majoritária ou totalmente cada uma em seu próprio grupo com poucas fatias de

outras amostras sendo também classificadas nestes grupos, implicando que estas amostras são as mais diferentes pelos parâmetros avaliados.

Se este resultado é comparado com o da avaliação ingenua do K-Means pode-se perceber que a amostra 05 agora consegue ser diferenciada das amostras 01, 02, 03 e 06, mas que agora a amostra 10, que antes era seu próprio grupo, passou a ser considerada parte deste cluster. E mais importante, muitas amostras que antes conseguiam ser diferenciadas agora estão agrupadas junto, então mesmo que o erro seja menor, esta parametrização reduziu a precisão de diferenciar diferentes amostras.

5.2.2 Propagação de afinidade

O algoritmo de propagação de afinidade demorou algumas horas para encontrar um resultado para o conjunto de dados apresentados, em comparação com segundos dos outros algoritmos. Além disto, apresenta o pior resultado dos algoritmos avaliados, isto é maior diferença entre as duas tabelas presentadas nesta seção, mesmo depois de varias iterações para achar o conjunto ideal de parâmetros. Nas tabelas 6 e 7 é possível perceber a segmentação dos dados e a alta incoerência entre os dados de treinamento e validação, implicando em que o algoritmo não foi capas de encontrar padrões no conjunto de dados avaliados.

Cluster Amostra	0	1	2	3	4	5
Amostra 01	40.72%	3.29%	10.98%	15.92%	18.77%	10.32%
Amostra 02	41.45%	2.82%	10.58%	15.70%	19.22%	10.23%
Amostra 03	37.09%	4.18%	11.06%	15.46%	20.66%	11.55%
Amostra 04	40.00%	3.56%	11.93%	12.30%	22.30%	9.93%
Amostra 05	39.06%	3.37%	12.46%	15.42%	20.13%	9.56%
Amostra 06	38.95%	2.30%	13.03%	12.64%	21.33%	11.75%
Amostra 07	38.89%	4.01%	11.46%	13.92%	22.36%	9.35%
Amostra 08	38.55%	3.44%	11.07%	12.98%	23.28%	10.69%
Amostra 09	36.63%	3.48%	12.93%	14.78%	23.04%	9.13%
Amostra 10	36.80%	3.90%	9.88%	16.12%	23.02%	10.27%
Amostra 11	39.35%	3.57%	11.90%	13.89%	21.36%	9.92%
Amostra 12	41.37%	3.72%	10.67%	14.63%	19.90%	9.71%

Tabela 6: Clusters por Propagação de afinidade

Cluster Amostra	0	1	2	3	4	5
Amostra 01	-	-	10.89%	-	89.11%	-
Amostra 02	20.63%	-	-	-	-	79.37%
Amostra 03	100.00%	-	-	-	-	-
Amostra 04	-	-	-	100.00%	-	-
Amostra 05	-	33.94%	-	-	-	66.06%
Amostra 06	-	-	100.00%	-	-	-
Amostra 07	-	-	0.63%	23.42%	75.95%	-
Amostra 08	100.00%	-	-	-	-	-
Amostra 09	100.00%	-	-	-	-	-
Amostra 10	100.00%	-	-	-	-	-
Amostra 11	-	-	31.55%	0.60%	67.86%	-
Amostra 12	100.00%	-	-	-	-	-

Tabela 7: Validação por Propagação de afinidade

5.2.3 Deslocamento de média

Deslocamento de média prova-se promissório, conseguindo chegar em menos de 2.5% de erro após o ajuste iterativo do seu parâmetro de largura de banda, sem ser necessário a escolha de numero de clusters. Além disso é interessante observar que o numero de clusters detectados por este método é igual menos do que o numero de amostras, indicando que algumas amostras são muito similares a outras.

Cluster Amostra	0	1	2	3	4	5	6	7	8
Amostra 01	-	100.00%	-	-	-	-	-	-	-
Amostra 02	-	-	-	-	100.00%	-	-	-	-
Amostra 03	100.00%	-	-	-	-	-	-	-	-
Amostra 04	100.00%	-	-	-	-	-	-	-	-
Amostra 05	-	87.60%	11.09%	-	1.32%	-	-	-	-
Amostra 06	-	100.00%	-	-	-	-	-	-	-
Amostra 07	98.86%	-	-	-	1.14%	-	-	-	-
Amostra 08	-	-	-	96.36%	-	-	1.75%	1.01%	0.88%
Amostra 09	-	-	-	-	-	100.00%	-	-	-
Amostra 10	100.00%	-	-	-	-	-	-	-	-
Amostra 11	98.26%	-	-	-	1.74%	-	-	-	-
Amostra 12	-	18.85%	81.15%	-	-	-	-	-	-

Tabela 8: Clusters por Deslocamento de média

É interessante observar que para este algoritmo as amostras 03, 04, 07, 10 e 11 foram classificadas no grupo 0, as amostras 01, 05 e 06

Cluster Amostra	0	1	2	3	4	5	6	7	8
Amostra 01	-	100.00%	-	-	-	-	-	-	-
Amostra 02	-	-	-	-	100.00%	-	-	-	-
Amostra 03	100.00%	-	-	-	-	-	-	-	-
Amostra 04	100.00%	-	-	-	-	-	-	-	-
Amostra 05	-	88.12%	11.88%	-	-	-	-	-	-
Amostra 06	-	100.00%	-	-	-	-	-	-	-
Amostra 07	98.54%	-	-	-	1.46%	-	-	-	-
Amostra 08	-	-	-	93.94%	-	-	1.82%	2.42%	1.82%
Amostra 09	-	-	-	-	-	100.00%	-	-	-
Amostra 10	100.00%	-	-	-	-	-	-	-	-
Amostra 11	97.06%	-	-	-	2.94%	-	-	-	-
Amostra 12	-	17.86%	82.14%	-	-	-	-	-	-

Tabela 9: Validação por Deslocamento de média

no grupo 1, e que as amostras 02, 08, 09 e 12 foram classificadas cada uma como seu próprio grupo. Em comparação com o algoritmo de K-Means Deslocamento de média conseguiu diferenciar a amostra 02 das outras, e agrupou os clusters de forma diferente, indicando que diferentes algoritmos de clusterização podem perceber padrões diferentes nos grupos, permitindo as vezes separar grupos que antes seriam impossível, mas podendo juntar grupos que antes seriam facilmente separados.

5.3 COMPARAÇÃO DOS MÉTODOS DE CLUSTERIZAÇÃO

Pode-se sintetizar as tabelas anteriores em uma única tabela que presente o erro obtido em cada um dos algoritmos de cluster, onde cada linha representa uma amostra e cada coluna um algoritmo de clusterização. Note que o valor apresentado nesta tabela é metade da diferença absoluta entre as duas tabelas de cada algoritmo, isto acontece porque cada dado classificado errado seria contabilizado duas vezes, i.e a falta desta percentagem no grupo certo, e o excesso dela no grupo errado.

Pode-se ver que a maioria dos métodos apresenta resultados promissores, com exceção do algoritmo de Propagação de Afinidade que apresentou uma grande margem de erro, e portanto sera excluído nas analises seguintes, incluindo os cálculos de média do erro por amostra para os diversos algoritmos apresentado na tabela 10.

Uma das primeiras formas em os algoritmos podem ser avaliados é assumindo que as amostras utilizadas são homogêneas, desta forma o resultado esperado para um bom algoritmo de

Amostra	KM	KM9	PA	DM	Média sem PA
Amostra 01	3.63%	-	70.34%	-	0.91%
Amostra 02	0.05%	0.27%	69.14%	-	0.08%
Amostra 03	-	-	62.91%	-	-
Amostra 04	-	-	87.70%	-	-
Amostra 05	2.18%	0.56%	87.07%	1.32%	1.02%
Amostra 06	5.36%	1.10%	86.97%	-	1.62%
Amostra 07	1.01%	1.01%	63.08%	0.32%	0.59%
Amostra 08	0.08%	2.73%	61.45%	2.42%	1.30%
Amostra 09	2.84%	0.99%	63.37%	-	0.96%
Amostra 10	-	-	63.20%	-	-
Amostra 11	-	-	66.14%	1.20%	0.30%
Amostra 12	-	-	58.63%	0.99%	0.25%
Média Algoritmo	1.26%	0.56%	70.00%	0.52%	0.58%

Tabela 10: Erros da clusterização por algoritmo

classificação seria que 100.0% das fatias de uma mesma amostra fossem classificadas dentro do mesmo grupo. A tabela 11 apresenta a percentagem de fatias que foram classificadas dentro de um mesmo grupo para cada amostra e algoritmo de clusterização, os algoritmos de K-Means, K-Means utilizando 9 clusters, Propagação de Afinidade e Deslocamento de Média foram abreviados para KM, KM9, PA e DM respectivamente. Nesta tabela também é possível observar o erro médio de cada algoritmo bem como o erro médio de cada amostra para os diversos algoritmos.

Como é possível observar pelo alto grau de fatias sendo classificadas no mesmo grupo para cada amostra é possível assumir que as amostras são altamente homogêneas. O menor valor é observado para a Amostra 06 no método de K-Means com 12 clusters, porém é possível ver que essa mesma amostra apresenta uma homogeneidade maior em outros métodos.

Tendo em vista a alta homogeneidade das diversas fatias de um mesmo grupo, e que a mesma fatia pode ser classificada como grupos diferentes a depender do método e dos parâmetros utilizados, é interessante nomear cada um dos grupos para poder visualizar como diferentes algoritmos percebem diferentes similaridades e conseguem separar grupos que outros algoritmos consideraram iguais. Uma nomenclatura simples foi dada na tabela 12, onde começando pela

Amostra	KM	KM9	DM	Média por amostra
Amostra 01	95.19%	100.0%	100.0%	98.40%
Amostra 02	98.97%	98.81%	100.0%	99.26%
Amostra 03	100.0%	100.0%	100.0%	100.0%
Amostra 04	100.0%	100.0%	100.0%	100.0%
Amostra 05	96.77%	97.54%	87.60%	93.97%
Amostra 06	60.87%	96.20%	100.0%	85.69%
Amostra 07	96.77%	96.77%	98.86%	97.47%
Amostra 08	87.05%	90.34%	96.36%	91.25%
Amostra 09	80.36%	80.56%	100.0%	86.97%
Amostra 10	100.0%	100.0%	100.0%	100.0%
Amostra 11	100.0%	100.0%	98.26%	99.42%
Amostra 12	100.0%	100.0%	81.15%	93.72%
Média por algoritmo	93.00%	96.67%	96.85%	95.51%

Tabela 11: Agrupamento de amostras por método de clusterização

letra 'A' contando pelo grupo em que a primeira amostra foi reconhecida cada novo grupo recebeu a letra seguinte no alfabeto.

Desta tabela destaca-se o fato de que todos os algoritmos detectaram a Amostra 01 no mesmo grupo que a Amostra 06, implicando que estas duas amostras são realmente muito similares. A exceção deste contraexemplo é possível observar que nenhum outro grupo foi classificado igualmente pelos três algoritmos.

Também é interessante destacar que mesmo que o algoritmo de K-Means tenha sido parametrizado com 12 e 9 grupos as amostras foram classificadas maioritariamente em apenas 8 e 5 grupos respectivamente, já para o caso do algoritmo de Deslocamento de Média as amostras foram classificadas maioritariamente em 6 dos 9 grupos detectados pelo algoritmo. Isto implica que os dados utilizados possuem fatias com valores suficientemente diferentes do resto para serem detectados em seu próprio grupo, mas que em sua maioria ainda podem ser considerados homogêneos.

Amostra	KM	KM9	DM
Amostra 01	A	A	A
Amostra 02	A	A	B
Amostra 03	A	A	C
Amostra 04	B	B	C
Amostra 05	A	C	A
Amostra 06	A	A	A
Amostra 07	C	C	C
Amostra 08	D	C	D
Amostra 09	E	C	E
Amostra 10	F	A	C
Amostra 11	G	D	C
Amostra 12	H	E	F
Número de Grupos	8	5	6

Tabela 12: Agrupamento de amostras por método de clusterização

6 CONCLUSÃO

O presente trabalho demonstrou que há viabilidade em reconhecer fatias de imagens tridimensionais em tomografias de raio-X pertencentes a materiais porosos com um elevado grau de certeza. O método utilizado é facilmente reproduzível, dependendo apenas da extração de um numero limitado de parâmetros que são facilmente computáveis e de algoritmos de clusterização amplamente disponíveis.

A percentagem de erros apresentados na tabela 10 permite demonstrar que para a maioria dos algoritmos de clusterização os parâmetros extraídos são suficientes para classificar com um elevado grau de certeza uma dada fatia. Desconsiderando a propagação de afinidade que provou-se incapaz de encontrar os clusters para os dados avaliados, dos restantes algoritmos o maior erro ocorreu para a parametrização ingenua de K-Means, sendo de apenas 1.26%, caindo para menos da metade (0.56%) quando uma parametrização mais condicente com os dados foi utilizada. O algoritmo de deslocamento de média prova-se extremamente promissório, tendo conseguido um erro de apenas 0.52% em média.

Já no quesito dos dados serem classificados majoritariamente dentro do mesmo grupo, como pode ser observado na tabela 11, também o algoritmo de deslocamento de média apresentou os melhores resultados com uma média de 96.85% das fatias de uma mesma amostra sendo classificadas dentro do mesmo grupo. Neste quesito todos os algoritmos apresentaram resultados similares, com o menor sendo novamente o K-Means com parametrizadão ingenua, que apresentou uma média de apenas 93%. Isto nos permite concluir que não só os dados de treinamento foram extremamente homogêneos, senão que os algoritmos de classificação conseguiram reconhecer esta homogeneidade a partir dos dados extraídos das imagens.

Disto é possível perceber que o algoritmo de Deslocamento de Média foi o que melhor conseguiu classificar os dados, porém é importante ressaltar que caso o conjunto de dados tivesse sido diferente os resultados são suficientemente semelhantes que seria possível que os resultados tivessem sido diferentes.

Em vista dos resultados apresentados pelo método descrito neste trabalho para rochas homogêneas, é possível estender a sua aplicação ao reconhecimento de camadas de rochas tipicamente heterogêneas, como rochas reservatórios, se considerarmos que cada camada de uma rocha heterogênea é equivalente a uma das rochas

homogêneas estudadas neste trabalho. Para tanto é necessário a obtenção de um conjunto de imagens classificadas manualmente fatia a fatia por um geólogo em relação as diversas classificações que desejem se obter da amostra, estes conjuntos de dados são conhecidos como Padrão Ouro, e são necessários para poder determinar o nível de erro de um dado algoritmo. Pela pressuposição da homogeneidade dos dados utilizados neste trabalho é possível construir um padrão ouro sem a ajuda de um geólogo, apenas supondo que cada amostra devia ser inteiramente classificada como o mesmo grupo.

Não foi possível encontrar na literatura um conjunto de dados classificados manualmente para a construção deste padrão ouro, e a construção deste por parte dos envolvidos no trabalho foi inviabilizada pelo sigilo dos dados de rochas reservatório disponíveis e porque a construção deste grupo implicaria em uma analise subjetiva propensa a contestação por outros geólogos. Por isto o presente trabalho não apresenta uma aplicação do método apresentado para a classificação de camadas de rochas reservatório, mas o principio aplica-se igualmente desde que se tenham dados homogêneos para o treinamento do algoritmo de clusterização.

6.1 TRABALHOS FUTUROS

Trabalhos futuros poderiam expandir nas medições extraídas da imagem, utilizando, por exemplo, outros fatores de formação dos poros para tentar obter resultados mais precisos. Ou ainda tentar diferentes métodos de clusterização.

É necessário observar que para o proposito deste trabalho assumiu-se que cada imagem representa um tipo de rocha diferente, de forma que a classificação pudesse ser testada analiticamente. Portanto trabalhos futuros também poderiam explorar uma avaliação mais focada na geologia dos resultados dos algoritmos propostos. Construindo um conjunto de imagens representativas dos grupos que se desejam classificar com a ajuda de um geólogo, e avaliando o resultado da classificação contra uma imagem de referencia também construída com a ajuda de um geólogo.

Os resultados apresentados na tabela 12 nos mostram que diferentes algoritmos podem classificar a mesma fatia de forma diferente, mas que considerando o resultado de diferentes métodos simultaneamente as amostras produzem um resultado praticamente único. Desta forma um outro trabalho futuro poderia ser realizar uma

analise utilizando uma classificação fuzzy baseada nos resultados de múltiplos algoritmos de clusterização para a determinação do grupo real de uma fatia a ser avaliada.

REFERÊNCIAS

ADAMS, R.; BISCHOF, L. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 16, n. 6, p. 641–647, 1994.

ADLER, P.; JACQUIN, C.; QUIBLIER, J. Flow in simulated porous media. *International Journal of Multiphase Flow*, Elsevier, v. 16, n. 4, p. 691–712, 1990.

BERRYMAN, J. G. Measurement of spatial correlation functions using image processing techniques. *Journal of Applied Physics*, AIP, v. 57, n. 7, p. 2374–2384, 1985.

BERRYMAN, J. G.; BLAIR, S. C. Use of digital image analysis to estimate fluid permeability of porous materials: Application of two-point correlation functions. *Journal of applied Physics*, AIP, v. 60, n. 6, p. 1930–1938, 1986.

CHAUHAN, S.; RÜHAAK, W.; KHAN, F.; ENZMANN, F.; MIELKE, P.; KERSTEN, M.; SASS, I. Processing of rock core microtomography images: Using seven different machine learning algorithms. *Computers & Geosciences*, Elsevier, v. 86, p. 120–128, 2016.

CHENG, G.; GUO, W. Rock images classification by using deep convolution neural network. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2017. v. 887, n. 1, p. 012089.

CHENG, Y. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 17, n. 8, p. 790–799, 1995.

COSTER, M.; CHERMANT, J.-L. Précis d"analyse d"images. Presses du CNRS, 1989.

DULLIEN, F. A. *Porous media: fluid transport and pore structure*. [S.l.]: Academic press, 2012.

FREY, B. J.; DUECK, D. Clustering by passing messages between data points. *science*, American Association for the Advancement of Science, v. 315, n. 5814, p. 972–976, 2007.

HARALICK, R. M.; STERNBERG, S. R.; ZHUANG, X. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, n. 4, p. 532–550, 1987.

IASSONOV, P.; GEBRENEGUS, T.; TULLER, M. Segmentation of x-ray computed tomography images of porous materials: A crucial step for characterization and quantitative analysis of pore structures. *Water Resources Research*, Wiley Online Library, v. 45, n. 9, 2009.

KARSANINA, M. V.; GERKE, K. M.; SKVORTSOVA, E. B.; MALLANTS, D. Universal spatial correlation functions for describing and reconstructing soil microstructure. *PloS one*, Public Library of Science, v. 10, n. 5, p. e0126515, 2015.

KATTAN, W. A.; JAWAD, S. N. A.; JOMAAH, H. A. Cluster analysis approach to identify rock type in tertiary reservoir of khabaz oil field case study. *Iraqi Journal of Chemical and Petroleum Engineering*, v. 19, n. 2, p. 9–13, 2018.

KETCHAM, R. A.; CARLSON, W. D. Acquisition, optimization and interpretation of x-ray computed tomographic imagery: applications to the geosciences. *Computers & Geosciences*, Elsevier, v. 27, n. 4, p. 381–400, 2001.

LIANG, Z.; FERNANDES, C.; MAGNANI, F.; PHILIPPI, P. A reconstruction technique for three-dimensional porous media using image analysis and fourier transforms. *Journal of Petroleum Science and Engineering*, Elsevier, v. 21, n. 3-4, p. 273–283, 1998.

MIRABOLGHASEMI, M.; PRODANOVIĆ, M.; DICARLO, D.; JI, H. Prediction of empirical properties using direct pore-scale simulation of straining through 3d microtomography images of porous media. *Journal of Hydrology*, Elsevier, v. 529, p. 768–778, 2015.

MOHAGHEGH, S.; AREFI, R.; AMERI, S.; AMINIAND, K.; NUTTER, R. Petroleum reservoir characterization with the aid of artificial neural networks. *Journal of Petroleum Science and Engineering*, Elsevier, v. 16, n. 4, p. 263–274, 1996.

OTSU, N. A Threshold Selection Method from Gray-level Histograms. *IEEE Transactions on Systems, Man and Cybernetics*, IEEE, v. 9, n. 1, p. 62–66, jan. 1979. ISSN 0018-9472. <http://dx.doi.org/10.1109/tsmc.1979.4310076>.

PAL, N. R.; PAL, S. K. A review on image segmentation techniques. *Pattern recognition*, Elsevier, v. 26, n. 9, p. 1277–1294, 1993.

PATEL, A. K.; CHATTERJEE, S. Computer vision-based limestone rock-type classification using probabilistic neural network. *Geoscience Frontiers*, Elsevier, v. 7, n. 1, p. 53–60, 2016.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

RICCOMINI, C.; SANT, L. G.; TASSINARI, C. C. G. et al. Pré-sal: geologia e exploração. *Revista USP*, n. 95, p. 33–42, 2012.

SANTOS, F. V. *USO DE ALGORITMOS DE CLASSIFICAÇÃO PARA DETERMINAÇÃO DE ELETROFÁCIES EM POÇOS DA BACIA DE CAMPOS*. Tese (Doutorado) — UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE LABORATÓRIO DE ENGENHARIA E EXPLORAÇÃO DE PETRÓLEO, 2016.

SHIH, F. Y.; CHENG, S. Automatic seeded region growing for color image segmentation. *Image and vision computing*, Elsevier, v. 23, n. 10, p. 877–886, 2005.

TREMEAU, A.; BOREL, N. A region growing and merging algorithm to color segmentation. *Pattern recognition*, Elsevier, v. 30, n. 7, p. 1191–1203, 1997.

VINCENT, L. Fast grayscale granulometry algorithms. In: *Mathematical morphology and its applications to image processing*. [S.l.]: Springer, 1994. p. 265–272.

WHITAKER, S. Flow in porous media i: A theoretical derivation of darcy's law. *Transport in porous media*, Springer, v. 1, n. 1, p. 3–25, 1986.

YADAV, J.; SHARMA, M. A review of k-mean algorithm. *International Journal of Engineering Trends and Technology*, v. 4, n. 7, p. 2972–2976, 2013.

YEONG, C.; TORQUATO, S. Reconstructing random media. *Physical Review E*, APS, v. 57, n. 1, p. 495, 1998.

YUEN, H.; PRINCEN, J.; ILLINGWORTH, J.; KITTLER, J. Comparative study of hough transform methods for circle finding. *Image and vision computing*, Elsevier, v. 8, n. 1, p. 71–77, 1990.