JOÃO GUILHERME DOS SANTOS PRUDENTE DO AMARAL

IMPUTAÇÃO COM SELF-ORGANIZING MAPS: COMPARAÇÃO DE ABORDAGENS APLICADAS À DENSIOMETRIA GAMMA

Trabalho de Formatura em Engenharia de Petróleo apresentado à Escola Politécnica da Universidade de São Paulo

SÃO PAULO

2019

JOÃO GUILHERME DOS SANTOS PRUDENTE DO AMARAL

IMPUTAÇÃO COM SELF-ORGANIZING MAPS: COMPARAÇÃO DE ABORDAGENS APLICADAS À DENSIOMETRIA GAMMA

Trabalho de Formatura em Engenharia de Petróleo apresentado à Escola Politécnica da Universidade de São Paulo

Área de concentração: Engenharia de Petróleo

Orientador: Prof. Dr. Rafael dos Santos Gioria

SÃO PAULO

2019

Catalogação-na-publicação

Amaral, João Guilherme
Imputação com Self-Organizing Maps: Comparação de abordagens aplicadas à densiometria gamma / J. G. Amaral, R. Gioria -- São Paulo, 2019.
51 p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Minas e Petróleo.

1.Engenharia de Petróleo 2.Redes Neurais 3.Escoamento Multifásico I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Minas e Petróleo II.t. III.Gioria, Rafael

AGRADECIMENTOS

Devo agradecer antes de tudo a meus pais, Cristiane e Astor, que apesar de todas as dificuldades que apareceram sempre prezaram pela boa educação, me aconselharam e apoiaram pelos caminhos nessa jornada de aprendizado. Aos meus avós, Sueli e João, que compartilharam com eles os ensinamentos e ajudaram a construir uma base sólida. Ao amigo Rafael, que após uma breve conversa pós-aula tem me orientado em trabalhos acadêmicos e desenvolve uma brilhante carreira na educação. E à Universidade de São Paulo e todos seus funcionários, que me proporcionaram uma educação de altíssima qualidade.

RESUMO

A Inteligência Artificial tem recebido amplo destaque nos últimos anos, sobretudo pela capacidade de análise de dados e identificação de padrões. Na indústria do petróleo, seu uso expande com a vasta gama de experimentos e formulações que a envolvem, assim como a exploração de novos horizontes. O trabalho traz uma abordagem estatística e experimental que avalia a capacidade da utilização de Redes Neurais Artificiais do tipo Self-organizing Maps (SOM) na imputação de dados para identificação de topologia de escoamentos multifásicos de água-óleo-gás e cáculo de suas frações volumétricas em medições de densiometria *gamma*, o que ressalta sua importância para a indústria do petróleo, onde um dos grandes desafios ainda é medir propriedades de escoamentos multifásicos de óleo, gás e água. Os dados são imputados seguindo 4 diferentes metodologias, sendo elas imputação simples com o BMU (*Best Matching Unit*, ou neurônio mais representativo), proporcional com o BMU, com a média entre o BMU e o segundo candidato a BMU, e com a média entre o BMU e seus vizinhos, os resultados são comparados com base em coeficiente de determinação, Bias de Correlação, e acurácia na classificação. As metodologias baseadas no BMU, substituição simples e proporcional, se mostram mais adequadas.

Palavras-chave: Engenharia de Petróleo. Redes Neurais. Self-organizing Maps. Medições em escoamento multifásico.

ABSTRACT

Artificial Intelligence has received a great deal of attention in recent years, mainly due to the capacity of data analysis and identification of patterns. In the oil industry, its use expands with the wide range of experiments and formulations involving it, as well as exploring new horizons. The assignment brings out a statistical and experimental approach that evaluates the ability of using Artificial Neural Networks of the Self-organizing Maps (SOM) type to data imputation and classification of multiphase flows by gamma densiometry, which highlights its importance for the petroleum industry, where one of the great challenges is still to measure properties of multiphase flows of oil, gas and water. The data are imputed following 4 different methodologies, BMU replacement (most representative neuron), proportional BMU replacement, BMU and 2nd mean replacement, and BMU and neighbors mean replacement, the results are compared based on determination coefficient, Correlation Bias, and classification accuracy. BMU-based methodologies, BMU replacement and proportional BMU replacement, are the most appropriate.

Keywords: Petroleum Engineering. Neural Networks. Self-Organizing Maps. Measurements in multiphase flow.

SUMÁRIO

1 INTRODUÇÃO

A IA tem recebido amplo destaque nos últimos anos, sobretudo pela capacidade de análise de dados e identificação de padrões. Na Engenharia de Petróleo, com o recente interesse e entusiasmo da indústria em análises em tempo real de poços e campos inteligentes, a IA tem sido centro de atenções (BRAVO et al., 2014). Pode-se citar, a título de exemplo : redes neurais sendo utilizadas para gerar regressões otimizadas que permitem predizer produções de óleo (WEISS; BALCH; STUBBS, 2002); técnicas de fuzzy ranking para rankear os dados de treinamento da rede em ordem de importância e aumentar sua taxa de acerto; e o aprendizado de máquina aplicado em métodos de análise real-time de poços em produção com integração de fuzzy e Multiphase Flow Metering, associados a bases de dados históricas por Knowledge Discovery in Databases (ALIMONTI; FALCONE, 2002). Incorporado a tal tendência, o objetivo do presente trabalho é desenvolver um estudo prático e estatístico comparando diferentes modelos de imputação utilizados em redes neurais do tipo SOM (*Self-Organizing Maps*), aplicados a medições de densiometria *gamma*.

Muitos dos estudos exploratórios na indústria do petróleo sofrem perdas de dados por falhas de equipamentos ou carecem de uma relação analítica entre medições diretas e informação desejada. Este é um problema que poderia ser amenizado utilizando técnicas de imputação de dados. Os SOM têm sido amplamente utilizados em machine learning, pelo fato de serem baseados em aprendizagem não supervisionada e identificarem relações pouco triviais. Seu algoritmo mapeia o conjunto de dados de treinamento por competitividade e resulta numa superfície que representa a distribuição da amostra num espaço bidimensional.

O conceito de imputação envolve uma estimativa para preenchimento de variável faltante num vetor de dados n-dimensional de acordo com um critério baseado em suas demais (n-1) variáveis conhecidas. O uso de técnicas matemáticas clássicas permite preencher as lacunas com estimativas como regressões dos demais vetores e substituição pela média. Contudo, tomar decisões com base em dados gerados por métodos restritos como esses pode não ser seguro, uma vez que a conclusão apresentada carrega problemas como tendências e linearizações de dados. Os valores encontrados são tendenciosos, pois a abordagem para análise da amostra

não representa a população (GELMAN; HILL, 2007). As redes neurais aparecem então como um mecanismo que permite não só produzir estimativas livres de subjetividade, como também aproveitar a alta capacidade de processamento de dados das máquinas, dando a elas a habilidade de aprender sem serem explicitamente programadas, o que aumenta a escabilidade dos experimentos.

O trabalho é dividido em 3 principais etapas. No capítulo de Revisão Bibliográfica são introduzidos os conhecimentos necessários para entendimento da Inteligência Artificial e suas bases aplicadas. Apresenta-se ainda o que consiste e os motivos para a realização de medições em escoamentos multifásicos e do funcionamento de redes neurais artificiais do tipo Self-Organizing Maps . Em seguida, já na seção de metodologia, se descreve como se desenvolve a geração de dados simulados para densiometria *gamma* em um escoamento multifásico de óleo, gás e água, divididos em grupos para treinamento de rede e imputação. Com a rede treinada, realiza-se um estudo prático de comparação entre diferentes métodos de imputação de dados que têm como base esse tipo de rede. Ao final do estudo, na discussão dos resultados, métricas estatísticas são utilizadas para comparar a capacidade das diferentes metodologias abordadas.

1.1 Objetivo

O objetivo geral do presente estudo é testar diferentes metodologias de imputação de dados faltantes utilizando Redes Neurais do tipo SOM em medições de fração de fases em escoamentos multifásicos com densiometria *gamma* e identificação de topologia do escoamento. Nesse contexto, é gerado um conjunto de dados sintéticos simulando valores de intensidade que seriam medidos, incluindo ruido, de alguns padrões de escoamentos multifásicos para treinar uma rede neural do tipo SOM para classificação e imputação. Com a rede treinada, é possível testar formas distintas para imputação e comparar seus resultados com métricas estatísticas.

1.2 Justificativa

O presente trabalho apresenta quatro grandes motivos para sua realização :

1. A imputação de dados permite a transformação de leituras de medidores em dados úteis sem utilização de formulações complexas, por simples comparação com padrões aprendidos.
2. Soluciona problemas gerados por falhas em sensores, que poderiam exigir repetição de experimentos (MCCULLOCH; PITTS, 2017). Alguns exemplos com falhas em sensores mitigadas na indústria do petróleo são: em estudos de sísmica, sensores frequentemente falham e comprometem a análise de especialistas; Na oceanografia, cápsulas para coleta de água para análise são perdidas em alto mar, dadas as condições a que são submetidas. Uma repetição destes experimentos afetaria muito seus fluxos de caixa. Com métodos confiáveis em mãos para estimativa dos dados faltantes, o problema seria resolvido.
3. Além de sua capacidade para estimativa de dados faltantes, a imputação pode ser utilizada como forma de predição de dados, esta é a terceria justificativa. É o que ocorre nos estudos meteorológicos onde, dadas as barreiras naturais que determinadas regiões impõem, há dificuldades em predizer o que pode acontecer entre as longas distâncias inter-estações. Nesse caso, confiabilidade também é um fator-chave a ser considerado.
4. Por último, mas não menos importante, pode-se citar a contribuição da densiometria *gamma* para a indústria do petróleo, uma vez que medir propriedades de escoamentos multifásicos de óleo, água e gás em tubulações ainda é um dos grandes desafios (THORN; JOHANSEN; HJERTAKER, 2013). Destacam-se os benefícios de um perfeito conhecimento do escoamento em termos de: testes de poço, monitoramento da produção e medição submarina na cabeça de poços (MERINI, 2011).

2 REVISÃO BIBLIOGRÁFICA

2.1 Aprendizado de máquinas

Com a chegada dos computadores na década de 1940, a capacidade de cálculo para qualquer estudo foi aumentada a níveis que o cérebro humano não consegue acompanhar. As máquinas eram pré-programadas para executar processos rigorosamente sistemáticos até que Alan Turing publica seu artigo *Computing machinery and intelligence*, onde levanta a questão "*Can machines think?*" (TURING, 2009). O estudo responde à pergunta com uma pesquisa onde o entrevistado deve distinguir uma conversa com uma pessoa e outra com um computador que lhes são apresentadas. Machine Learning pode ser definida como "O campo de estudos que dá aos computadores a habilidade de aprender sem serem explicitamente programados" (SAMUEL, 2000).

Podemos dividir os algoritmos em dois grandes grupos, de acordo com o tipo de aprendizado: supervisionados e não supervisionados. No *dataset* de entrada do aprendizado supervisionado, já é conhecido como o *output* deve parecer, assumindo um relacionamento entre entradas e saídas. Por esse motivo, podemos separá-los em algoritmos de regressão, onde o objetivo é descobrir uma função contínua que mapeie a relação entre *inputs* e *outputs* com a atribuição de coeficientes para cada uma das variáveis, e algoritmos de classificação, onde o intuito é mapear os vetores *inputs* em categorias discretas pré-estabelecidas.

Já no aprendizado não supervisionado, não é explícito o que fazer com o *dataset*, e nem é dito o que cada vetor representa. O algoritmo deve identificar e separar os dados em diferentes clusters, cujo critério para agrupamento é a relação de semelhança entre vetores aprendida.

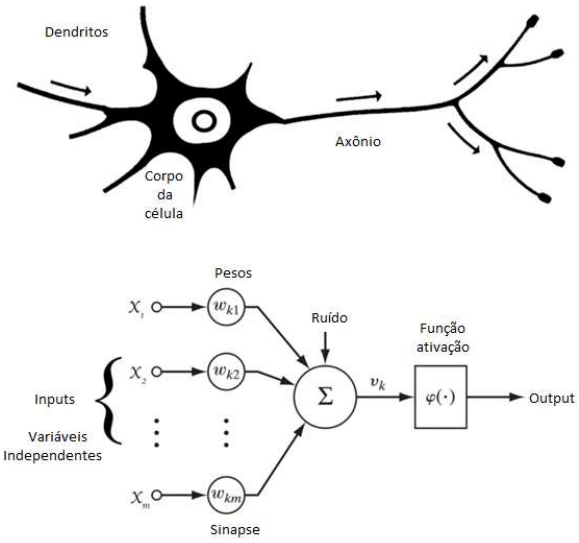
2.1.1 Redes Neurais Artificiais (RNA)

As chamadas Redes Neurais apareceram a primeira vez em 1943, num artigo onde o neurofisiologista Warren McCulloch e o matemático Walter Pitts estudam a atividade nervosa e suas relações com a comunicação entre neurônios no cérebro humano,

formulando a lógica por trás do processo e aplicado-a em um protótipo de circuito elétrico (WARREN; WALTER, 1943). Desde então, o algoritmo vem ganhando novas variações que o personalizam para diferentes tipos de aprendizado.

Sua base é imitar como o cérebro humano funciona. Analogamente, uma Rede Neural possui neurônios que receberão inputs de uma camada de variáveis dependentes. Os inputs devem necessariamente passar por uma etapa de normalização que os colocará numa base entre 0 e 1. Isso garante a convergência da rede (LECUN et al., 2012). A comunicação entre a camada de entradas e o neurônio é feita pela sinapse, etapa onde cada uma das variáveis é ponderada por um peso que melhor representa sua contribuição para a função ativação, que estima seu output contínuo, binário, ou categórico, com base nas variáveis inseridas.

Figura 1 - Neurônio humano e neurônio artificial.



Fonte: Adaptado de HAYKIN (2009).

2.1.2 SOM

Os *Self-Organizing Maps* (SOM) são redes neurais de aprendizado não supervisionado introduzidas por Teuvo Kohonen (KOHONEN, 1982). Os SOM aprendem por competição. Seus neurônios são organizados em uma camada uni ou bidimensional. É possível, mas não comum, o trabalho com camadas de maiores dimensões por questão de dificultar visualização dos dados em dimensões superiores.

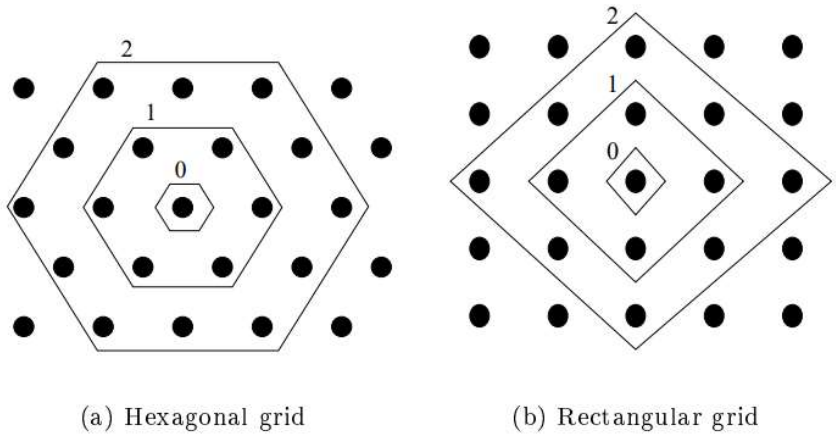
Os neurônios são ordenados de maneira a criar um novo sistema de coordenadas para as diferentes variáveis de entrada (KOHONEN, 1998). Ainda, no SOM ocorre o mapeamento dos dados de entrada, de forma que cada vetor é representante de variáveis estatísticas intrínsecas da camada de entrada.

A formulação de modelos neurais computacionais como os SOM se explica pela habilidade do cérebro humano de organizar informações sensoriais em mapas topológicos, o que abre espaço para um novo horizonte de métodos e ferramentas a serem estudados. Entre as informações sensoriais dessa forma mapeadas no córtex humano, pode-se citar o tato (MERZENICH et al., 1983), a visão (HUBEL; WIESEL; STRYKER, 1977), e a audição (SUGA; TSUZUKI, 1985). Assim dispostos, os blocos ficam disponíveis a nível de aprendizado para que, numa situação posterior, sejam processados.

Um mapa computacional é definido por uma matriz de neurônios, que representam conjuntos de variáveis cujos valores passaram por diferentes filtros ou processos, e que agem paralelamente no processamento de informações sensoriais (HAYKIN, 2008). Consequentemente, o mapa transforma a amostra de entrada em uma distribuição de probabilidade implícita no posicionamento de vetores na rede (KNUDSEN, 1987). Pode-se citar duas características dos mapas computacionais que destacam o valor que agreegam ao processamento: em qualquer estágio, cada um dos vetores é preservado em seu próprio contexto, e dessa forma, interagindo com o ambiente envolto de informação relacionada, se realizam conexões sinápticas próximas que contribuem para sua preservação topológica.

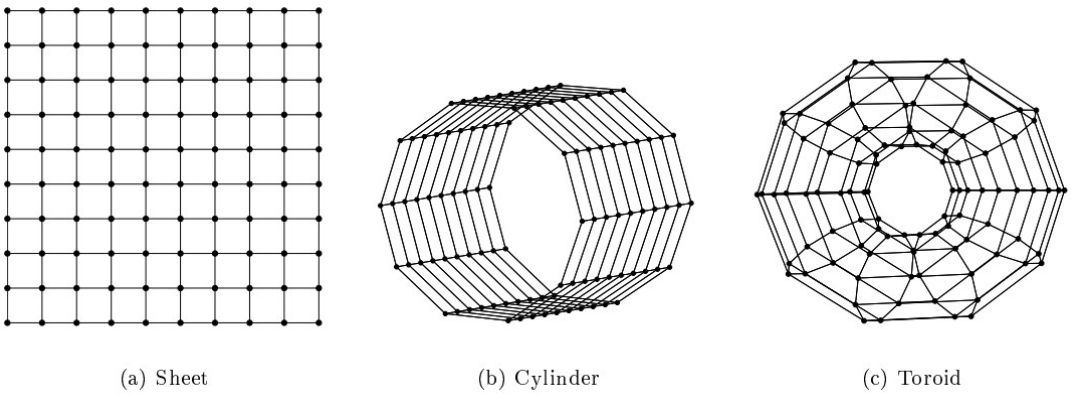
Um dos fatores cruciais para entendimento de seu funcionamento é o grid do mapa. Estes podem ser divididos em dois níveis de abrangência: os grids locais, que determinarão as vizinhanças próximas, que podem ser do tipo quadriculado ou hexagonal, apresentados na Figura 2; e o grid global do mapa, que representa uma visão mais ampla de sua estrutura dos dados (Ver figura 3). Cada quadrado ou cada hexágono terá uma vizinhança de neurônios, e o principal impacto da escolha do grid para um som está em sua sensibilidade a variações locais (SCHMIDT; REY; SKUPIN, 2011). Tal característica pode ser observada na Figura 2 ao comparar as vizinhanças 0, 1 e 2 nos dois grids: Neurônios de grid hexagonal estão sujeitos a influência de um número maior de vizinhos.

Figura 2 – Grids locais dos SOM.



Fonte: (VESANTO et al., 1999).

Figura 3 – Diferentes *shapes* de mapa.



Fonte: (VESANTO et al., 1999).

Para cada entrada de treinamento, após sua normalização são computadas as distâncias euclidianas em relação a cada vetor que compõe o mapa. O nó mais próximo ao vetor será seu BMU (*Best Match Unity*), também chamado neurônio vencedor. Computado o BMU, os pesos dos neurônios BMU e seus vizinhos são atualizados, assim cada neurônio BMU fica mais próximo de seu dado afim e a vizinhança acompanha mantendo uma hierarquia topológica. Ao terminar essa etapa, temos uma época de treinamento. A cada nova época, o raio de distância para atualização dos pesos diminui e a precisão do processo aumenta, obtendo um mapa

que se adapta à topologia do *dataset* de treinamento. Podemos separar a lógica de funcionamento dos SOM em 5 passos (HAYKIN, 2008):

1. Inicialização: Geração de pesos iniciais aleatórios para os neurônios;

2. Amostramento: Coleta de um amostra da camada de entrada para ativação do mapa;

3. Computação de similaridade: Encontro do neurônio vencedor (BMU) pelo critério da mínima distância euclidiana;

4. Atualização: Ajuste dos vetores de pesos sinápticos de todos os neurônios;

5. Continuação: Até que nenhuma mudança significativa ocorra com o mapa, retorna-se para o passo 2.

Uma vez que o SOM convergiu, obtém-se um mapa rico de informações estatísticas da camada de entrada. Biologicamente falando, a camada representa todos os neurônios e receptores que se distribuem pelo corpo humano, enquanto o mapa em si representa a camapa de neurônios que os mapeia no córtex cerebral. Pode-se citar algumas características importantes do mapa resultante: é uma redução, ou compressão, da dimensão da camada de entradas, o que justifica a comum adoção de uma função gaussiana para as sinapses entre vizinhanças; sua organização topológica destaca semelhanças entre vizinhos; dados mais recorrentes, ou seja, regiões da camada de entrada com maior número de representantes, são representadas por regiões maiores e mais nítidas no mapa; e naturalmente, como consequência de todas as demais características, o mapa seleciona as variáveis que melhor representam seus vetores de entrada.

2.1.3 K-means

Após o treinamento do SOM, o mapa é um importante *input* para a clusterização dos dados. Isto ocorre porque o próprio algoritmo dos SOM resulta numa pré-clusterização na organização topológica. O K-means auxilia na geração de grupos com características similares (FERLIN, 2008). O método k-means, agrupa os dados similares em clusters, representados por diferentes cores. É um método de

clusterização não hierárquico, ou seja, cujo número de clusters deve ser pré-definido. Podemos separar sua lógica de funcionamento em 5 passos:

1. Determinar o número K de clusters;

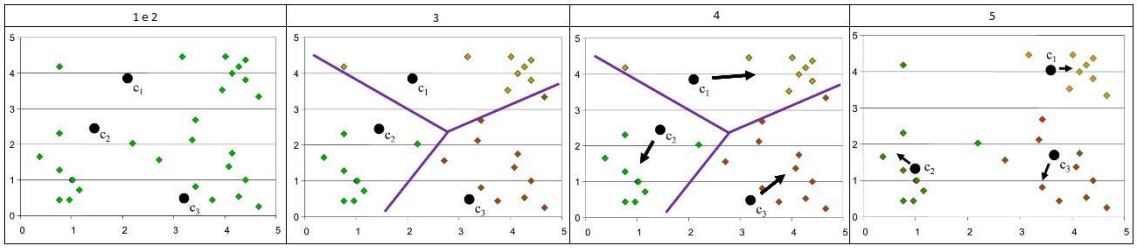
2. Selecionar aleatoriamente K pontos para serem os centróides dos clusters (não precisam estar em seu dataset);

3. Atribuir cada um dos pontos no dataset ao centróide que tenha menor distância eucidiana dele;

4. Calcular qual é o novo centróide para cada cluster de pontos (será o ponto com menor distância de todos os demais);

5. Atribua novamente cada um dos pontos a um centróide. Se houver alguma mudança de centróide para determinado ponto, repita o passo 4.

Figura 4 – Passos do algoritmo do K-means .



Fonte : Adaptado (ULLMAN et al., 2014).

A precisão do processo do k-means é muito dependente da escolha dos centros iniciais de clusters (MILLIGAN; COOPER, 1988). Para uma melhor performance, eles devem ser o máximo distintos que for possível. Uma boa estratégia para melhorar sua performance é utilizar métodos conhecidos como o de Ward, que divide os dados em grupos e utiliza como centro inicial dos clusters o vetor médio de cada um dos grupos.

2.1.4 Imputação de dados

Imputação é o nome dado para a estimativa de dados faltantes que comprometem *datasets* e dificultam análises de dados e a determinação de inferências estatísticas

(HU, M., SALVUCCI, S.M., COHEN, 1998). Os métodos mais comuns de imputação recorrem a procedimentos estatísticos, tais como médias e modas, o que gera a grande desvantagem de não considerar relações entre as diferentes variáveis, assim como diminuem a variância da amostra. Tendo isso em vista, outras metodologias mais complexas vêm sendo desenvolvidas de modo a considerar na substituição características intrínsecas ao meio em que o dado está inserido, o que se mostra muito eficiente para grandes conjuntos de dados (CARTWRIGHT; SHEPPERD; SONG, 2003).

O uso do SOM como imputador de dados tem sido explorado em diversas áreas de aplicação. Por não considerar a amostra inteira no cálculo da média para substituição, mas sim selecionar um número determinado de neurônios similares ao vetor, como seus vizinhos na rede, o algoritmo se mostra muito mais assertivo. Técnicas de substituição pelo vizinho mais próximo sao utilizadas para imputar dados pluviométricos em medições feitas na Malásia (MALEK et al., 2008) , regressão multivariada para estimar dados faltantes em datasets de qualidade do ar (JUNNINEN et al., 2004).

A confiabilidade do método utilizado para imputação dependerá de diversos fatores, tais como variabilidade da amostra e número de pontos no *dataset* de treinamento. Em um algoritmo de SOM, vetores com variáveis faltantes são computados na rede substituindo as variáveis faltantes pelos valores em seu BMU. Como as variáveis foram estimadas, pode-se dizer que o cálculo de distâncias carrega um erro induzido pela falta.

Uma vez que as variáveis dos vetores são normalizadas em um intervalo de 0 a 1, o erro é limitado a $\sqrt{n} - \sqrt{(n-k)}$, onde n é a dimensão dos vetores de entrada, e k o número de variáveis faltantes (R. RALLO, 2005). O erro de cálculo da distância é baixo para n>>k. É importante notar que após um treinamento eficiente da rede, valores semelhantes são associados a vetores vizinhos. Isso implica que, se a falta de variáveis resulta em um erro no cálculo das distâncias, o BMU calculado para o vetor pode não ser o certo, mas será um de seus vizinhos. Dessa forma, o SOM pode ser utilizado para imputação de dados, uma vez que a degradação da topografia dos dados não estará linearmente correlacionada ao percentual de variáveis perdidas (SAMAD; HARP, 1992).

No estudo aqui apresentado, quatro metodologias são analisadas e comparadas: substituir pelo BMU encontrado para o vetor na rede (Eq. 1), algo já feito na literatura; substituir pelo BMU encontrado para o vetor na rede, multiplicado pelo fator de projeção do vetor sobre o BMU (Eq. 2); encontrar o BMU e o segundo neurônio candidato a BMU, fazer a média deles e imputar (Eq. 3); e encontrar o BMU, fazer uma média do BMU com demais pontos que formam o cluster, e imputar (Eq. 4).

$$X_n = X_{n_{BMU}} \quad \text{(Eq. 1)}$$

$$X_n = X_n \ \frac{(BMU \times AMO)}{|BMU| \times |AMO|} \quad \text{(Eq. 2)}$$

$$X_n = \frac{(X_{n_{BMU_1}} + X_{n_{BMU_2}})}{2} \quad \text{(Eq. 3)}$$

$$X_n = \frac{\sum X_{n_y}}{m} \quad \text{(Eq. 4)}$$

Onde y é o conjunto do BMU e seus vizinhos, e m o número de vetores desse conjunto.

2.1.4.1 Imputação global baseada na variável com valores faltantes

O método de imputação global tem como base cálculos estatísticos que levam em consideração todos os vetores do mapa para estimar valores desconhecidos. Estes podem ser determinísticos ou estocásticos. No primeiro caso, os valores são substituidos pelo centro da distribuição, enquanto numa imputação estocástica se introduz um ruído ao centro da distribuição, procurando diminuir o viés nos dados. Para variáveis contínuas, se utiliza a média – ver Eq. 5. Enquanto categóricas são estimadas pela moda – ver Eq. 6. A opção pela imputação global carrega duas grandes desvatagens. Primeiramente, em casos onde se tem grandes outliers, a média é distorcida e seu valor não repesenta fielmente a amostra. Ainda, ao realizar substituição pela media global, se está diminuindo a variância dos dados, enviesando análises que possam ser feitas (FERLIN, 2008).

$$X = \frac{\sum_{k=1}^{n} X_k}{n} \quad \text{(Eq.)}$$

$$X = Mo(X) \quad \text{(Eq. 6)}$$

2.1.4.2 Imputação global baseada em demais variáveis

A imputação global baseada em demais variáveis engloba técnicas principalmente de regressão, onde valores de variáveis faltantes são estimados por uma relação entre as variáveis ajustada globalmente. Algumas das desvantagens dessa abordagem para imputação são: quando se tem mais de um dado faltante para o mesmo vetor, o modelo carece de informações para solucionar todos seus graus de liberdade; e a metodologia considera a premissa de que existe uma relação ente as variáveis e ainda que as amostras completas coletadas representam essa relação, o que nem sempre é verdade (FERLIN, 2008; SOARES, 2007).

2.1.4.3 Imputação local

Ao delimitar um sub-conjunto da amostra que se assemelhe ao vetor com variáveis faltantes para realização da imputação, se realiza uma imputação local. Quando da utilização da técnica, o critério de escolha do subconjunto e seu número de participantes são cruciais para uma boa estimativa (FERLIN, 2008). Aqui se enquadram o presente estudo e as três técnicas que serão testadas. A medida de similaridade normalmente é feita pela distância euclidiana entre os vetores, o que justifica as escolhas das três técnicas escolhidas, uma vez que os primeiro e segundo candidatos a BMU, assim como seus vizinhos, são vetores escolhidos pelo algoritmo como similares ao vetor a ser imputado pelo critério distância.

Entre as vantagens de uso da técnica, pode-se citar (FERLIN, 2008; MAGNANI, 2004; SOARES, 2007):

1. Obtenção de uma amostra sem dados faltantes;
2. Preservação das relações e distribuições implícitas na amostra;
3. Não toma nenhuma premissa de distribuição específica;
4. Mostra-se confiável mesmo com presença de ruídos e número elevado de dados;

Já quanto a desvantagens, pode-se citar:

1. Alto custo computacional;
2. Número de dados e escolha do sub-conjunto estão diretamente ligados à assertividade;
3. A maneira de cálculo da similaridade afeta diretamente os resultados.

2.2 Medições em escoamentos multifásicos

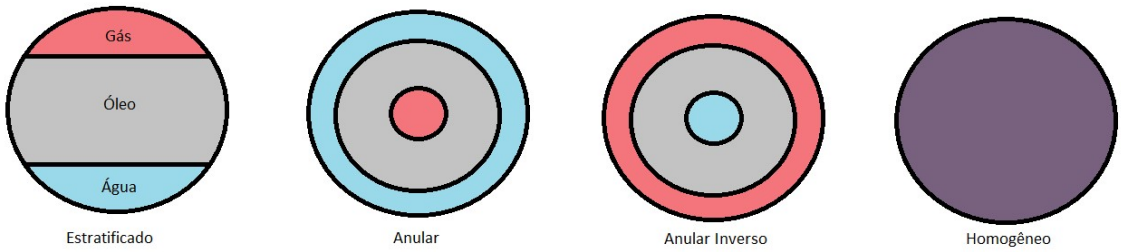
O problema de conseguir medir propriedades de escoamentos multifásicos de óleo, água e gás em tubulações ainda é um dos grandes desafios na indústria do petróleo (THORN; JOHANSEN; HJERTAKER, 2013). Existem inúmeras metodologias na literatura para medição destas propriedades, cada uma com suas peculiaridades. Podemos separá-las em dois grandes grupos: aquelas que dependem e as que não dependem de homogeneizar a representação do escoamento. Uma das metodologias, o Tubo de Venturi, permite obter a vazão do escoamento a partir da variação de pressão medida em uma contração na tubulação. O medidor de placa de orifício aproveita-se do mesmo princípio, com a inserção de uma placa com furo conhecido na tubulação, o que provoca perda de carga, mudando suas características.

Ambas as metodologias, apesar de utilizadas por seu baixo custo, são afetadas pelas frações de fases e necessitam de seu conhecimento *a priori*. Além disso, técnicas que necessitam da homogeneização não funcionam perfeitamente com escoamentos com gases, já que momentos após a mistura, o escoamento começa a heterogeneizar (FALCONE; HEWITT; ALIMONTI, 2009). A densiometria *gamma* e a impedância elétrica-magnética são duas metodologias não intrusivas utilizadas para análise de escoamentos multifásicos que não dependem de homogeneização (BELO; MENDES DE MOURA, 1999). Entre as vantagens de usá-las, pode-se citar a realização da análise sem depender da presença de um furo no trecho da tubulação a ser analisado ou da instalação de sistemas defletores que acabam mudando suas propriedades naturais. Outra metodologia não-intrusiva é a tomografia por impedância elétrica, que utiliza de correntes e potenciais elétricos. A técnica é não-linear e mal posta, o que exige sua integração com outros métodos (PELLEGRINI, 2019).

2.2.1 Análise de escoamentos multifásicos por densiometria *gamma*

Ao realizar medições por diversos raios distribuídos uniformemente no entorno de uma tubulação, a densiometria *gamma* permite calcular propriedades interessantes do fluxo que por ela passa, tal como a configuração das fases que o compõem e suas respectivas frações (Figura 5). Uma análise confiável das propriedades de escoamentos em tubulações é de extrema importância para a indústria do petróleo, uma vez que o alto custo de unidades flutuantes leva a opção dos *players* pela construção de tubulações para condução das misturas multifáscas de óleo, água e gás até o continente.

Figura 5 – Seção transversal em tubulação de transporte de óleo .



Fonte: Adaptado (BISHOP; JAMES, 1993).

A análise feita parte do fenômeno de atenuação dos raios *gamma* ao passar por diferentes meios, que dependerá do tamanho da camada, do comprimento de onda com que se está trabalhando e da natureza do meio onde está propagando. A intensidade de um raio *gamma* após passar por uma camada do meio de comprimento d é dada por:

$$I = I_0 \times e^{-\mu \times \rho \times d} \qquad \text{(Eq. 7)}$$

Onde ρ é a densidade do material, μ o coeficiente mássico de absorção do material para o comprimento de onda utilizado e I_0 a intensidade do raio-gamma antes de

adentrar o meio. Tendo conhecidos μ, ρ e coletando as itensidades observadas no experimento com raios *gamma*, encontra-se o comprimento da camada atravessada.

Para o caso de uma tubulação onde escoa uma mistura de óleo, água e gás na vertical, e com o raio sendo emitido na direção diametral, temos três parcelas do termo exponencial, representadas respectivamente pelos sub-índices O, A e G, uma para cada fase:

$$I = I_0 \times e^{-\mu_O \times \rho_O \times d_O} \times e^{-\mu_A \times \rho_A \times d_A} \times e^{-\mu_G \times \rho_G \times d_G} \quad \text{(Eq. 8)}$$

A equação 8 possui três variáveis desconhecidas, os comprimentos de cada camada por onde o raio passa. Sendo assim, para determiná-las faz-se necessária a eliminação de mais dois graus de liberdade. Emitindo um segundo raio, de comprimento de onda diferente, de mesma direção e percorrendo o mesmo caminho na tubulação, obtém-se a equação 9. A última equação, que completa a solução do problema vem da propriedade geométrica de que a soma dos comprimentos de cada uma das fases deve ser igual ao diâmetro (2R) da tubulação:

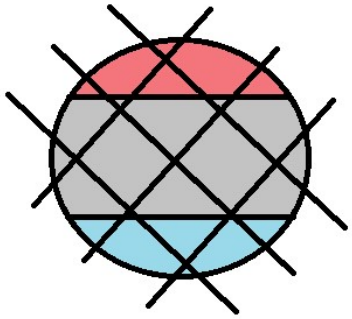
$$I' = I'_0 \times e^{-\mu'_O \times \rho_O \times d_O} \times e^{-\mu'_A \times \rho_A \times d_A} \times e^{-\mu'_G \times \rho_G \times d_G} \quad \text{(Eq. 9)}$$

$$2R = d_O + d_A + d_G \quad \text{(Eq. 10)}$$

As variáveis μ e ρ são calibradas previamente. Com as equações 8, 9 e 10, ao obter as intensidades medidas para ao menos dois raios distintos passando pela tubulação, obtém-se os comprimentos das camadas . As frações para cada fase do escoamento F_O, F_A *e* F_G observadas pelo raio serão dadas pela divisão de seu comprimento pelo diâmetro total da tubulação atravessada. Como o objetivo da análise é obter a fração de óleo por exemplo, e não o comprimento da camada, uma vez que essa dependerá

da configuração, são necessárias mais informações. Indica-se então um procedimento que permite obter as informações necessárias sem interferir no escoamento (BISHOP; JAMES, 1993).

Figura 6 – Padrão de distribuição dos 6 raios do experimento.



Fonte: BISHOP (1993).

A alternativa consiste na utilização de múltiplos raios distribuídos numa configuração padronizada no entorno da tubulação (por exemplo Figura 6), de forma que 12 medições sejam feitas em 6 caminhos distintos, adotando dois comprimentos de onda para cada raio. Com os dados obtidos, uma rede neural do tipo SOM, com a habilidade de reconhecer padrões previamente treinada é utilizada para reconhecer o tipo de escoamento observado. Aqui se tem mais uma motivação para a utilização da rede neural. Como ainda não é conhecida a distribuição das fases, o SOM, além de calcular suas frações, indica qual a topologia do escoamento, apenas com os dados da densiometria gamma e o treinamento adequado da rede.

2.3 Validação de modelos

2.3.1 Coeficiente de determinação

O coeficiente de determinação, ou R^2, é uma métrica de validação de modelos que diz quanto o modelo encontrado representa a amostra (ZHANG, 2017). Pode ser entendido também como o grau em que a variabilidade do modelo explica a variabilidade da amostra (Eq. 11). No caso de comparação de representatividade de modelos, o modelo com maior coeficiente de determinação consegue representar

melhor a amostra. Portanto, no trabalho em questão, busca-se a metodologia que gere maior coeficiente de determinação.

$$R^2 = 1 - \frac{SQ_{mod}}{SQ_{mod}} \quad \text{(Eq. 11)}$$

$$SQ_{mod} = \sum_1^n (y_i - mr_i)^2 \quad \text{(Eq. 12)}$$

$$SQ_{amo} = \sum_1^n (y_i - mt_i)^2 \quad \text{(Eq. 13)}$$

Onde SQ_{mod} e SQ_{amo} são as somas dos desvios quadráticos do modelo e da amostra, respectivamente, em relação a suas médias.

2.3.2 Bias de Correlação

Enquanto o coeficiente de determinação se apresenta como métrica para comparação de modelos em questão de variabilidade, o Bias de Correlação é uma métrica que indica quão distorcidas foram as correlações entre variáveis da amostra na construção do modelo (FERLIN, 2008). Portanto, busca-se no presente estudo a metodologia com menor Bias de Correlação em módulo (Eq. 16).

$$OC(K_1) = \frac{\sum \rho(K_1, K_n)}{n-1} \quad \text{(Eq. 14)}$$

$$ACB\left(K_{original}, K_{imputado}\right) = OC\left(K_{imputado}\right) - OC(K_{original}) \quad \text{(Eq. 15)}$$

$$CB = \sum_{i=1}^n ACB(K_{imputado}, K_{original}) \quad \text{(Eq. 16)}$$

Onde ρ é o coeficiente de correlação entre as variáveis, ACB o Bias de Correlação do atributo em questão, e CB o Bias de Correlação do modelo.

3 MATERIAL E MÉTODOS

A metodologia proposta segue uma ordem que garante congruência entre suas etapas interdependentes. Para que a Rede Neural seja treinada, é necessária uma base de dados confiável e não tendenciosa, gerada por simulações de densiometria gamma. Quatro tipos diferentes de escoamento multifásico são abordados: homogêneo, estratificado, anular e anular inverso. Treina-se então a rede até que seu mapa represente a amostra. Por fim, parte dos dados é ocultada e as capacidades de diferentes técnicas de imputação são comparadas, cumprindo então o objetivo do trabalho.

3.1 Geração de *datasets*

A formulação toma como base o embasamento teórico apresentado na seção 2.2 e adiciona a simulação ruídos, como a estatística de fótons para os dados medidos, aproximando-a a um experimento real. A geração de dados se dá em quatro passos:

1. Escolha aleatória de uma das quatro configurações de escoamento adotadas;
2. Escolha de números aleatórios entre 0 e 1 para $F_1, F_2 \; e \; F_3$, de modo que $F_O, F_A \; e \; F_G$ serão dados por:

$$F_O = \frac{F_1}{F_1+F_{2+F_3}}; \; F_A = \frac{F_2}{F_1+F_{2+F_3}}; \; F_G = 1 - F_O - F_A.$$

3. Para cada um dos seis caminhos de raios, como ilustrado na figura 5, calcular os comprimentos das camadas para as configurações e fases escolhidas;
4. Adicionar ruído aos comprimentos das camadas para considerar o efeito das estatísticas de fótons, que simula a distribuição de detecção de fótons por um sensor, tornando o dataset mais próximo de algo medido e não simulado.

A definição de parâmetros a serem utilizados é etapa crucial do processo. Para garantir proximidade das simulações a medições reais, se adota os parâmetros apresentados na tabela 1 (BISHOP; JAMES, 1993). Cabe levantar que tais valores

ainda são sintéticos e arredondados. O escopo do estudo é mostrar que os SOM e a imputação de dados funcionam, o que abre espaço para utilizar os parâmetros de referência, mesmo que esses não respeitem padrões de significância. Dado que padrões de comportamento são estudados, a única premissa é que os valores de ρ sejam distintos entre as três fases. Quanto ao intervalo de tempo de medição, definido em 10 segundos, destaca-se a razão para o valor ser ótimo. Dado que as medições de intensidade de raios gamma depende diretamente da estatística de fótons, tal efeito deve ser considerado no planejamento do experimento, de forma a evitar que os dados sejam enviesados. Tendo isso em conta, experimentos analisando a predição de dados por redes neurais em função do intervalo de experimentação levam a um Δt ótimo de 10 segundos. A estatística de fótons entra ainda como diferenciadora de um experimento real para aquele simulado. De modo a aproximar a simulação em questão a um caso real, se adiciona ruído à amostra, simulando uma distribuição de Poisson (BISHOP; JAMES, 1993).

Tabela 1 –Parâmetros utilizados para simulação (BISHOP; JAMES, 1993).

Diameter (cm)	15.00
Decay for gamma 1 (cm^2/g) - water	0.220
Decay for gamma 1 (cm^2/g) - oil	0.197
Decay for gamma 1 (cm^2/g) - gas	0.213
Decay for gamma 2 (cm^2/g) - water	0.058
Decay for gamma 2 (cm^2/g) - oil	0.062
Decay for gamma 2 (cm^2/g) - gas	0.068
Density (g/cm^3) - water	1.05
Density (g/cm^3) - oil	0.9
Density (g/cm^3) - gas	0.2
Δt (s)	10

A metodologia descrita é programada e simulada em Python para os dados de treinamento da rede e, posteriormente, para geração da amostra com dados faltantes. Foi escolhido Python dada sua gama de módulos de cálculo já implementados de maneira eficiente para utilização. Essa eficiência é crucial, dada a grande quantizade de dados que será gerada.

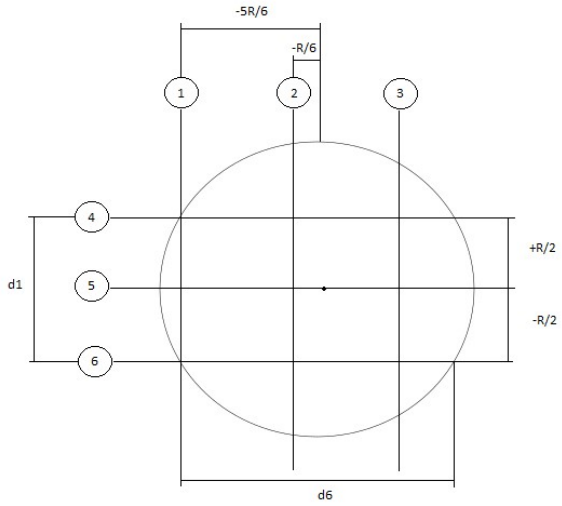
Para ambos os casos, são simulados dois valores iniciais de intensidade I_0^1 e I_0^2, dados pela equação 17, onde mu_g representa o decaimento gamma em gás eρ_ga densidade do fluido. A formulação assume que para cada raio emitido a intensidade

máxima enxergada seja de 60000 s^{-1}, num cenário onde o fluxo é monofásico, de gás.

$$I^{n}_{0}=\frac{6\times e^{4}}{e^{(-mu_{gn}\times\rho_{g}\times d)}} \quad \text{(Eq. 17)}$$

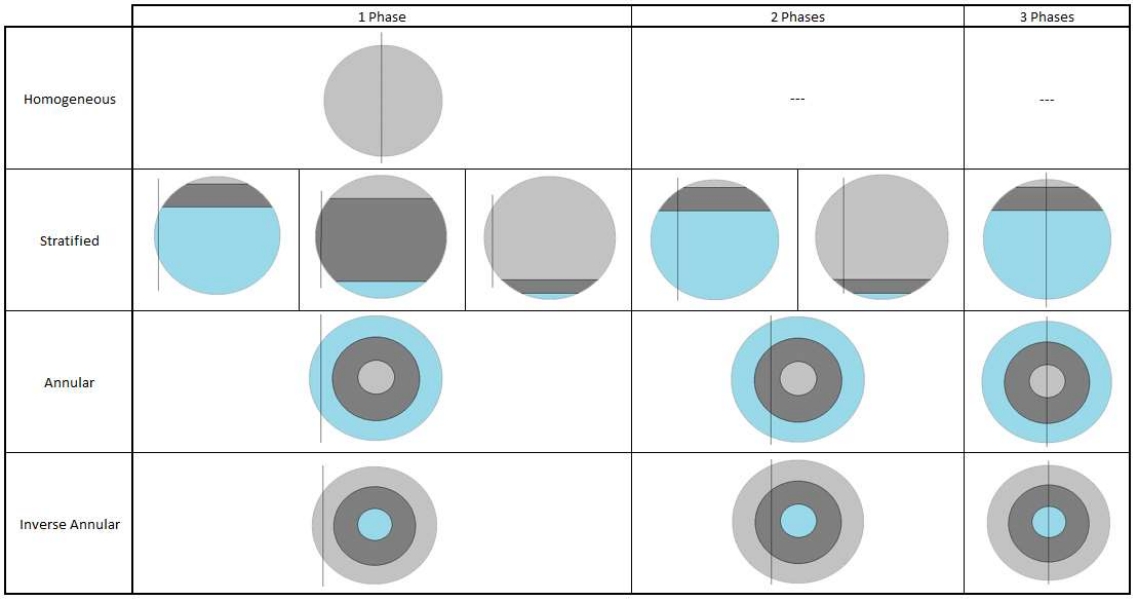
No desenvolvimento da formulação, preza-se pela utilização de medidas referenciais ao raio de tubulação simulado. Ainda, em busca de maior nitidez na distinção entre fases, posiciona-se os feixes de diferentes direções em seções distintas: horizontais posicionados a (+ $\frac{R}{2}$; 0 ; - $\frac{R}{2}$) em referência ao diâmetro, e verticais a (- $\frac{5R}{6}$; 0 ; - $\frac{R}{6}$) , conforme Figura 7.

Figura 7 – Esquema representativo do sistema de feixes de raios *gamma* simulados



Sorteadas as topologias, passa-se a tratar cada um dos casos e suas particularidades. A etapa 3 do processo de geração de dados deve ser tratada diferentemente, conforme apresentam as seções 3.1.1 a 4.1.3. Pada cada topologia encontra-se distintas possibilidades de caminhos dos feixes: passando por uma, duas ou três fases (Tabela 2).

Tabela 2 – Diferentes possibilidades para caminho dos feixes de raios *gamma* para cada topologia.



Os valores de $F_O, F_A \; e \; F_G$ obtidos são transformados e inseridos na rede como intensidades observadas, utilizando as equações 8, 9 e 10, e não só a classificação dada, como os outputs calculados pelas redes, são comparados com os dados simulados. O algoritmo desenvolvido para simulação compreende todas as particularidades descritas para cada um dos tipos de escoamento. É gerado um *dataset* com 10000 amostras. Com a validação finalizada, 20% dos dados gerados são utilizados para treinamento da rede e os restantes para testes de imputação.

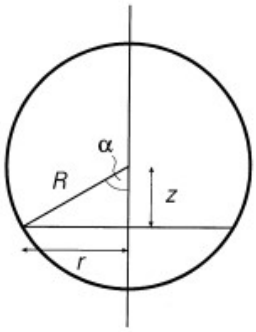
3.1.1 Caso homogêneo

Para escoamentos homogêneos, os comprimentos das camadas de óleo, gás e água serão dados em relação ao comprimento d da tubulação que o raio atravessa por:

$$x_O = F_O \times d, x_A = F_A \times d, x_G = F_G \times d \qquad \text{(Eq. 18)}$$

3.1.2 Caso estratificado

Figura 8 – Definição de variáveis para o escoamento estratificado .



Fonte: (BISHOP; JAMES, 1993).

Para escoamentos estratificados, faz-se necessário utilizar propriedades trigonométricas que relacionam as variáveis apresentadas na Figura 8, de forma a obter a equação transcendental Eq. 22. Para sua solução, foi utilizado um modulo de Python chamado scipy.optimize.brentq, que encontra uma raiz para a equação dentro de um intervalo entre 0 e R através do método de Brent de interpolação quadrática inversa. O método foi escolhido por otimizar a convergência de seu método iterativo de interpolação.

$$z^2 + r^2 = R^2 \qquad \text{(Eq. 19)}$$

$$z = R \times \cos\alpha \qquad \text{(Eq. 20)}$$

$$\left(\frac{2\alpha}{2\pi}\right)\pi R^2 = f\pi R^2 + 2\left(\frac{zr}{2}\right) \qquad \text{(Eq. 21)}$$

$$G_f(z) \equiv f\pi + \frac{z}{R}\left(1 - \frac{z^2}{R^2}\right)^{\left(\frac{1}{2}\right)} - \cos^{-1}\left(\frac{z}{R}\right) \qquad \text{(Eq. 22)}$$

Além do respeito dessa formulação, como nem todos os raios atravessarão a direção radial da tubulação, deve ser consideradas para simulação 6 possibilidades: 3 onde o raio atravessa apenas uma das fases, óleo, água ou gás; duas onde o raio atravessa duas das fases, óleo e água ou óleo e gás; e uma onde o raio atravessa todas as fases conforme ilustrado na Tabela 2.

3.1.3 Casos anular e anular inverso

Para os casos anular e anular inverso, a formulação é a mesma que a do caso estratificado. A diferença está no número de possibilidades de caminhos que o raio pode fazer. Enquanto para o escoamento estratificado poderiam ser encontradas 6 diferentes possibilidades, para os casos anular e anular inverso se tem apenas 3 delas: passar por uma das fases; passar por duas das fases; ou passar pelas três fases. No caso anular, passar por apenas uma das fases significa atravessar somente a água, e por duas fases a água e o óleo. Já para escoamentos anulares inversos, apenas uma das fases é atravessar apenas gás, enquanto duas fases é passar por ele e o óleo.

3.2 Implementação do SOM

Apesar de o mercado de softwares disponibilizar plataformas que oferecem algoritmos de redes do tipo SOM já implementadas, com parâmetros ajustáveis, as mais utilizadas exigem compra de assinatura. Por tal motivo, optou-se novamente pela utilização de módulos open source já desenvolvidos em Python, mas cuja estrutura de entrada de dados, pré-tratamento da amostra e a definição de parâmetros da rede deveria der implementada. O modulo Somoclu (WITTEK et al., 2017) mostrou-se mais adequado ao trabalho quanto a performance e flexibilidade para mudança de parâmetros. A topologia do mapa pode ser optada tanto como planar ou em toróide e o grid retangular ou hexagonal, assim como diversos outros parâmetros que podem ser ajustados de acordo com a necessidade do usuário.

3.2.1 Mapa

O mapa definido visa estrategicamente potencializar as análises a serem feitas ressaltando diferenças entre os diferentes tipos de topologia. Para tal, recorre-se à literatura em busca de melhores práticas e fórmas heurísticas para definição dos parâmetros a serem utilizados. O grid local escolhido é hexagonal, de forma a aumentar a sensibilidade dos neurônios a variações locais, conforme seção 2.1.2. Globalmente, opta-se pela estrutura toroidal, garantindo simetria na rede de forma que

neurônios de borda tenham todos o mesmo número de conexões. Quanto ao tamanho do mapa, este está diretemente ligado à acurácia e a interpretabilidade dos resultados. Um mapa muito pequeno resulta em baixa acurácia, enquanto um muito grande em baixa interpretabilidade (SHALAGINOV; FRANKE, 2015). Para cada grid a ser utilizado, existem na literatura exemplos de modelos que otimizam o tamanho do mapa. Para grids hexagonais, a Eq. 23 define seu número ótimo de neurônios em função no número m de amostras no *dataset* de treinamento e a Eq.24 a proporção ótima entre dimensões no mapa, onde COV_{V1} e COV_{V2} são as máximas covariâncias observadas entre variáveis ainda nos dados de treinamento (VESANTO et al., 1999). O número de amostras utilizadas para treinamento é de 8000, o que resulta em aproximadamente 500 vetores ótimos. A rede utilizada é do tamanho 20X25.

$$Nº\ de\ neurônios = 5 \times \sqrt{m} \qquad \text{(Eq. 23)}$$

$$\frac{D(X)}{D(Y)} = \frac{COV_{V1}}{COV_{V2}} \qquad \text{(Eq. 24)}$$

A função vizinhança adotada é a Gaussiana, que determinará a taxa de mudança da vizinhança ao entorno do neurônio vencedor. O coeficiente utilizado para a função é de 0.5. Ela influenciará diretamente o treinamento da rede, e em algoritmos aplicados para classificação, a função Gaussiana, combinada com uma taxa de aprendizado linear, geram ótima performance com baixo erro de quantização (NATITA; WIBOONSAK; DUSADEE, 2016).

3.2.2 Treinamento

Para treinamento da rede, faz-se necessária boa escolha de parâmetros, minimizando tedências por excesso ou falta de iterações, os chamados *overfitting* e *underfitting*. A literatura no assunto destaca a dificuldade de se criar modelos que criem padrões de otimização nessa etapa de desenvolvimento do SOM. Para cada tipo de dados, de algoritmo e variância de *dataset* se obtém um resultado. Dessa forma, opta-se pela realização de testes de forma a atingir melhor performance com ajustes finos. Para tal, se realizam testes variando número de épocas de treinamento da rede, assim como as relações de ajuste de raio de vizinhanças, linear ou exponencial, e de escala. As métricas utilizadas como parâmetro para melhora de resultados são os erros

topográfico e quantitativo obtidos ao final do treinamento. O primeiro mede o nível de preservação da topologia local após redução de dimensão, pelo percentual de vetores cujos BMUs primários e secundários não são adjacentes , enquanto o segundo resulta do cálculo da distância média entre cada um dos nós e seus BMUs. Quanto menores os números, melhor é o algoritmo (CABANES; BENNANI, 2010).

Tabela 3 – Erro topográfico observado em função da configuração de treinamento escolhida.

	Topographical error			
	Linear scale cooling		Exponential scale cooling	
	Radius cooling		Radius cooling	
Epochs	Linear	Exponential	Linear	Exponential
10	0.1995	0.1510	0.1975	0.1555
15	0.1875	0.1315	0.2295	0.2005
20	0.1960	0.1840	0.2185	0.2090
50	0.1945	0.1985	0.2195	0.1930
80	0.1795	0.2125	0.2150	0.1810
90	0.1610	0.1675	0.1900	0.1795
95	0.1775	0.2195	0.2090	0.1645
100	0.1830	0.2000	0.2180	0.1860
200	0.1895	0.2575	0.2195	0.2220
500	0.2310	0.2075	0.1990	0.1895

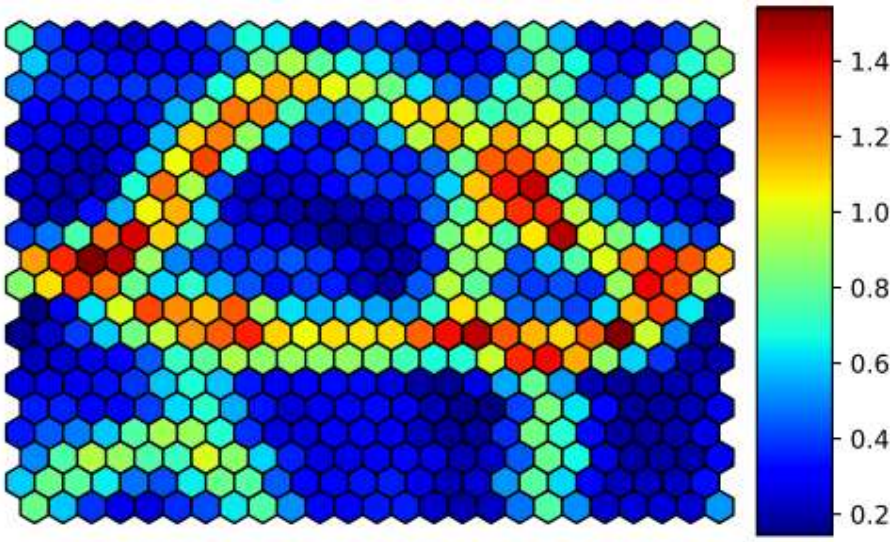
Tabela 4 – Erro quantitativo observado em função da configuração de treinamento escolhida.

	Qunatization error			
	Linear scale cooling		Exponential scale cooling	
	Radius cooling		Radius cooling	
Epochs	Linear	Exponential	Linear	Exponential
10	0.0627	0.0296	0.1410	0.0113
15	0.1034	0.1136	0.1488	0.2064
20	0.0006	0.0450	0.3220	0.3248
50	0.0563	0.1126	0.2575	0.1443
80	0.1179	0.1317	0.0798	0.1763
90	0.3830	0.1266	0.0309	0.2561
95	0.1928	0.1539	0.3987	0.0839
100	0.2021	0.0699	0.3210	0.1065
200	0.2843	0.0060	0.2596	0.3500
500	0.1301	0.2306	0.1823	0.0256

Os resultados para os testes realizados permitem observar que para números muito baixos de épocas de treinamento, tem-se alta variabilidade de resultados, consequentes de *underfitting.* A rede foi pouco treinada. Em contrapartida, para números muito altos, se observa algo semelhante, com os erros medidos assumindo valores elevados. Tendo isso em vista, opta-se por uma configuração que apresenta bom desempenho, com número de épocas de treinamento balanceado: 90 épocas, com atualização de escala linear e atualização de raios de vizinhança exponencial – vide Tabelas 3 e 4.

Plots pós-treinamento mostram como seriam visualizados os mapas gerados após seu treinamento, e as mudanças provocadas por uma alteração no *grid*. As *component planes* permitem uma visualização da distribuição de cada uma das componentes dos vetores da amostra utilizada no mapa. As U-matrix, organizam o mapa de forma a representar as distâncias entre os neurônios vizinhos no SOM – Figura 9. Para análise mais detalhada dos resultados de treinamento da rede, elenca-se no Apêndice C seus demais resultados.

Figura 9 – *U-matrix* plotada, um *heatmap* cuja formatação indica as distâncias entre neurônios.



3.2.3 Imputação

São abordados quatro tipos diferentes de imputação de variáveis faltantes. O primeiro deles segue o que é feito por muitos exemplos na literatura: a substituição do vazio por seus correspondentes do BMU encontrado para o vetor (Eq.1). Uma variável dele, ainda, é multiplicá-lo pela projeção do vetor sobre o BMU (Eq. 2). Já as duas outras metodologias combinam o SOM a métodos estatísticos para determinação dos valors imputados: substituição pela média entre os correspondentes no primeiro e no segundo candidato a BMU (Eq.3), e a substituição pela média entre os valores correspondentes no BMU e seus vizinhos (Eq.4). Para tal, exportam-se os dados da rede treinada e se realizam as estimativas.

Entre os quatro métodos utilizados, aqueles que combinam o SOM com médias estatísticas promovem a imputação de valores que representam melhor sua vizinhança como um todo, aumenta-se a variância dos dados. Por outro lado, na substituição pelo BMU e em sua variação proporcional, não há incremento relevante de variância podendo induzir uma tendência.

3.2.3.1 Validação

A validação dos dados imputados segue duas principais metodologias de comparação estatística: o Coeficiente de Determinação e o que chama-se de Bias da Correlação. No desenvolvimento do estudo, adota-se função já implementada em Python para cálculo do coeficiente de determinação (eq. 11). Já para o Bias de Correlação, adota-se metodologia apresentada na literatura (FERLIN, 2008) e também introduzida na revisão bibliográfica (eq. 14).

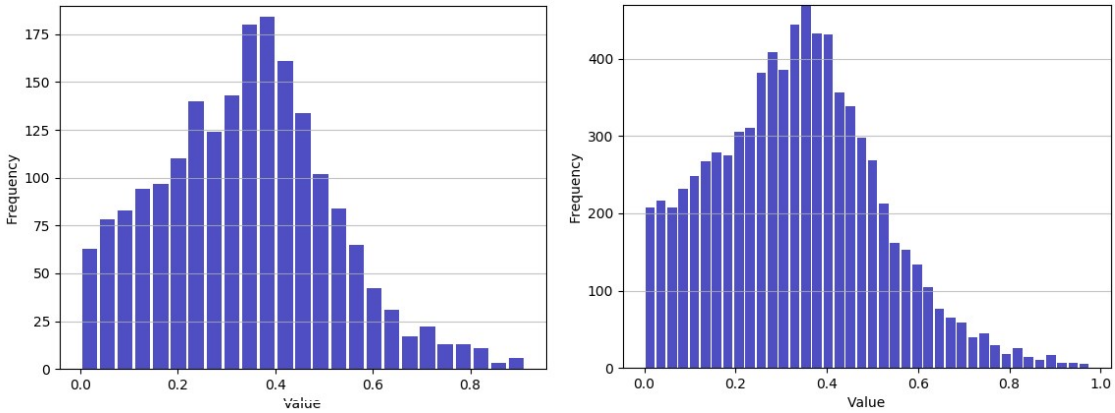
4 RESULTADOS

4.1 Geração de dados de escoamento multifásico

A geração de dados de medições de escoamento multifásico por densiometria *gamma* é etapa crucial para um bom desenvolvimento das demais etapas do estudo. Conforme metodologia, se adotam parâmetros elencados pela literatura (BISHOP; JAMES, 1993), sintéticos e arredondados. Algo importante a ser destacado é o fato de se ter dividido a amostra simulada em duas: uma para treinamento da rede, com 8000 vetores, e outra para testes, com 2000 vetores. A razão para tal divisão está em evitar o *overfitting*, um viés que resultaria em falsas constatações.

As variáveis para análise relacionadas à densiometria *gamma* são : as intensidades medidas para o primeiro comprimento de onda testado para os raios 1 a 6 (Figura 7), com nome seguindo o padrão de L11 a L16; as intensidades medidas para os mesmos raios, com o segundo comprimento de onda testado, seguindo o padrão de L21 a L26; as frações volumétricas de água, óleo e gás, respectivamente Fw, Fo e Fg; e quatro variáveis binárias que identificam a topologia do escoamento.

Gráfico 1 – Histograma dos valores simulados de fração de água para treino (a) da rede neural e para testes (b).

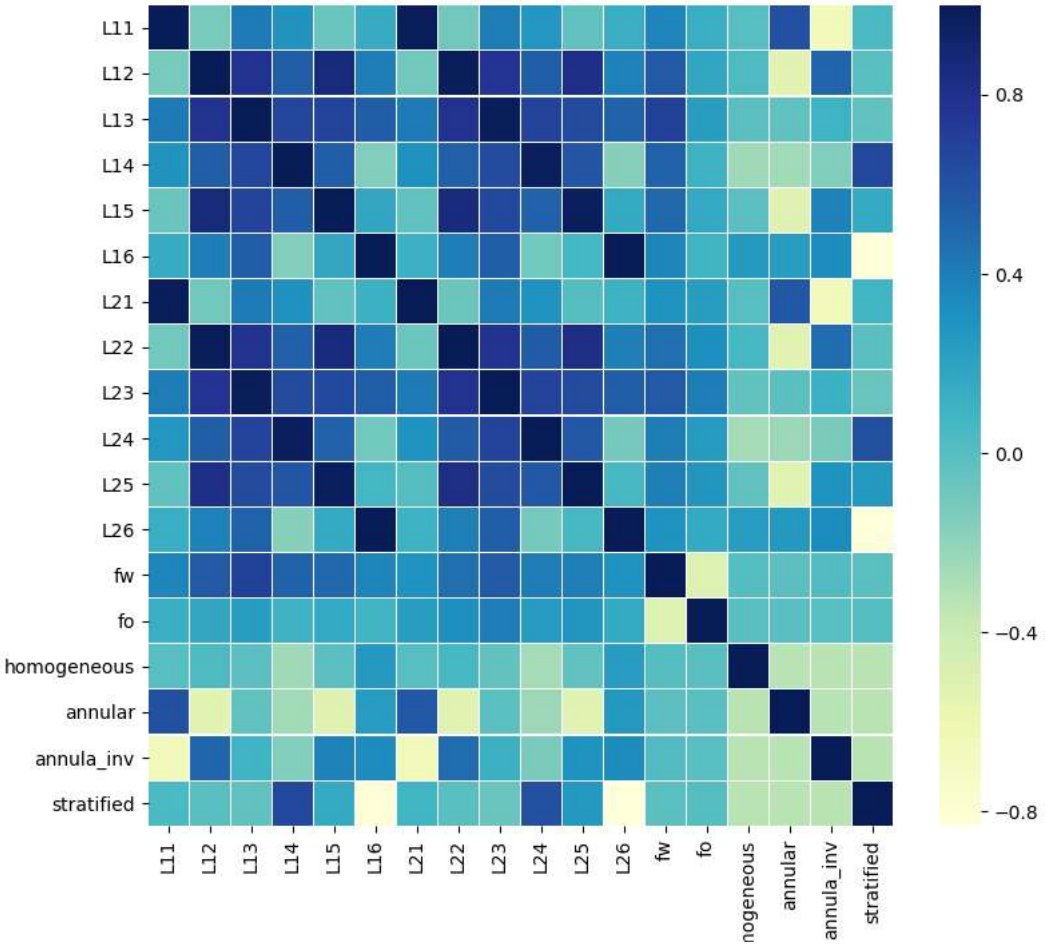


(a) (b)

Para efeito de análise, elencam-se aqui as variáveis Fw e Fo, respecivamente, as frações de água e de óleo dos vetores gerados. Por serem resultados de cálculos que envolvem as variáveis medidas na densiometria, podem servir como norte de comparação entre as duas amostras geradas. Conforme ilustram os histogramas nos

gráficos 1.a e 1.b para o caso da fração de água, ambas as distribuições, tanto para a amostra de treinamento, quanto para a amostra de teste, têm a mesma forma. Os histogramas de fração de óleo encontram-se no Apêndice B. Pode-se dizer ainda que se aproximam da distribuição de Poisson.

Figura 10 – *Heatmap* de correlação entre as variáveis na amostra gerada para treinamento da rede neural.



Destaca-se ainda a correlação entre variáveis das amostras simuladas para treino (Figura 10). Para uma análise mais minuciosa, os valores da matriz encontram-se no Apêndice A. Conforme metodologia, para a geração de amostras de densiometria, duas medidas são simuladas para uma mesma trajetória de travessia da tubulação, com comprimentos de onda diferentes, a título de exemplo L11e L21. Entre tais variáveis , dada sua coincidência de caminho, se observa alta correlação – representada pelos pontos mais escuros do *heatmap*. Outra observação interessante está no fato de a topologia estratificada apresentar correlação relevante com as

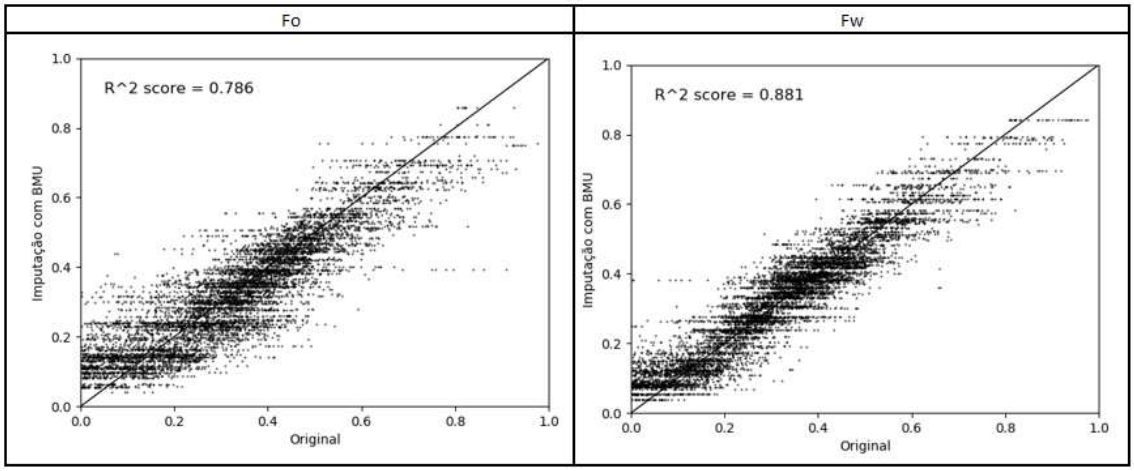
variáveis L14 e L24, referentes à medida simulada para o raio horizontal superior, o que faz sentido, uma vez que o raio que passa por tal região, no caso estratificado, estará atravessando apenas uma fase, o que permite fácil identificação da topologia.

4.2 Imputação de dados

Tem-se como principal objetivo a comparaçao de metodologias de imputação de dados faltantes de densiometria *gamma* utilizando redes neurais artificiais do tipo SOM. Quatro metodologias distintas foram implementadas: substituição pelo valor do BMU, pelo valor do BMU proporcional à projeção do vetor sobre seu BMU, pela média dos valores do primeiro BMU e do segundo, ou pela média entre o BMU e seus 6 vizinhos. Entre os quatro métodos implementados, espera-se inicialmente que aqueles que combinam o SOM com médias estatísticas promovam a imputação de valores que representam melhor sua vizinhança, no entanto aumenta-se a variância dos dados. Por outro lado, na substituição pelo BMU espera-se que não ocorra incremento relevante de variância.

As metodologias utilizadas para efeito de validação da imputação, seguindo esta lógica, são o coeficiente de determinação dos dados imputados em relação aos dados reais, que mede quanto o modelo consegue explicar os valores reais simulados (ZHANG, 2017), e o Bias de Correlação, que fornece uma medida de quanto o modelo distorce a correlação entre variáveis, em relação aos dados reais (FERLIN, 2008).

Gráfico 2 – *Plot* de frações de óleo (a) e água (b) imputados com BMU e dados originais.



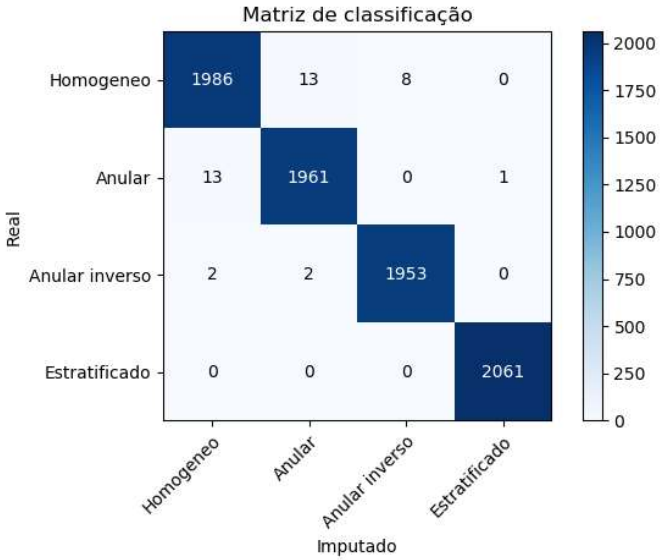
Para a imputação por substituição pelos valores do BMU encontrado, se observa altos coeficientes de determinação (Gráfico 2). Quanto mais dispersos estiverem os dados em relação à linha diagonal do Gráfico 2, menor será sua representatividade em relação aos dados originais, o que não parece ser um grande problema no caso em questão. O BMU de um vetor é definido como o vetor de menor distância euclidiana na rede em relação ao vetor a ser imputado. Sendo assim, apresenta-se como o vetor que melhor o representa. Partindo desse princípio, faz sentido imaginar que uma estimativa dos dados a serem imputados por aqueles encontrados no BMU do vetor com dados faltantes seja uma boa alternativa, dado que toda a topologia da rede, com suas relações entre vizinhos e as considerações estatísticas que estão nelas implícitas, aponta para alta semelhança entre os dois vetores. Ainda, a metodologia apresentou o menor valor para Bias de Correlação entre as abordadas pelo presente estudo (Tabela 5). Quanto menor o Bias de Correlação, menor a distorção da correlação entre variáveis do modelo em relação aos dados originais.

Tabela 5 – Bias de correlação observado em função da metodologia para imputação utilizada.

	Bias de correlação
BMU	2.29%
BMU e segundo	3.07%
BMU e vizinhos	25.74%
BMU proporcional	2.30%

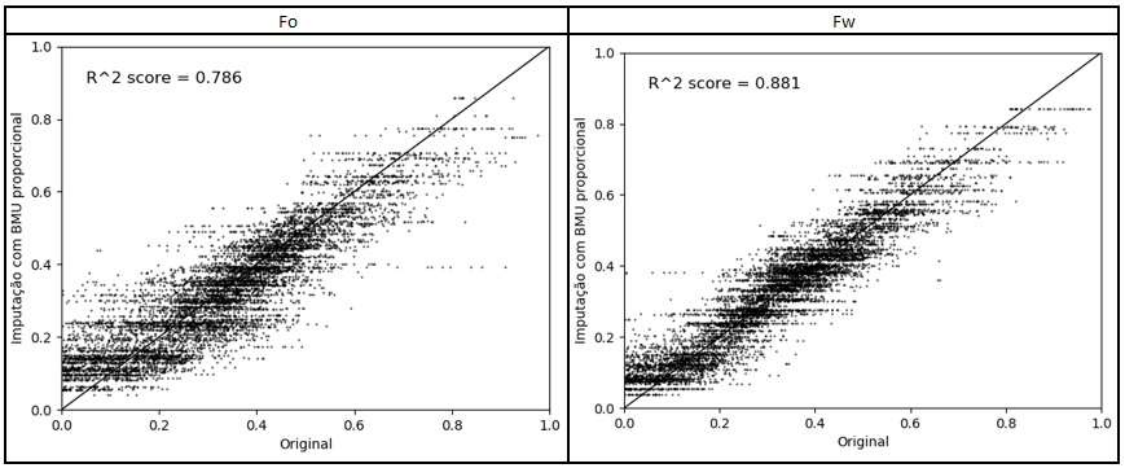
Para efeito de comparação, analisa-se ainda a assertividade da metodologia de imputação na classificação da topologia do escoamento. Neste quesito, a imputação por substituição pelo BMU apresenta alto nível de eficácia. Para a população de 2000 vetores de testes, apenas 39 receberam classificação que não condiz com sua topologia original (Figura 11). Destaca-se ainda o fato de a maior quantidade dos erros se concentrar entre os casos homogêneo e anular. De acordo com a quantidade de fases que cada raio atravessa, torna-se mais difícil de diferenciar as duas topologias. No escoamento homogêneo, os raios necessariamente atravessarão 3 fases de escoamento, enquanto no caso anular nem sempre. Num caso anular onde todos os raios atravessam 3 fases de escoamento, como exemplo, não se tem este gatilho de distinção entre as duas topologias, o que torna menor sua assertividade.

Figura 11 – Matriz de classificação para os dados imputados por substituição pelo BMU.



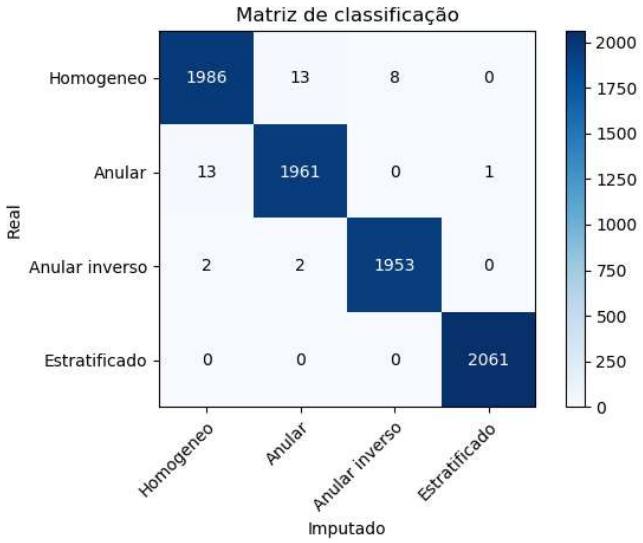
Analogamente à imputação por substituição pelo BMU, a metodologia por substituição pelos valores do BMU proporcionalmente à projeção do vetor sobre o BMU apresenta os resultados mais favoráveis. Dado que em muito dos casos o fator de proporção estimado pela projeção é aproximadamente 1, seus resultados são semelhanes. Iguais quando comparados sob o critério coeficiente de determinação (Gráfico 2) e matriz de classificação (Figura 12), e próximos ao comparar pelo critério Bias de Correlação, apresentado um valor baixo de 2.30% para a medida de distorção.

Gráfico 2 – *Plot* de frações de óleo (a) e água (b) imputados com BMU proporcional e dados originais.



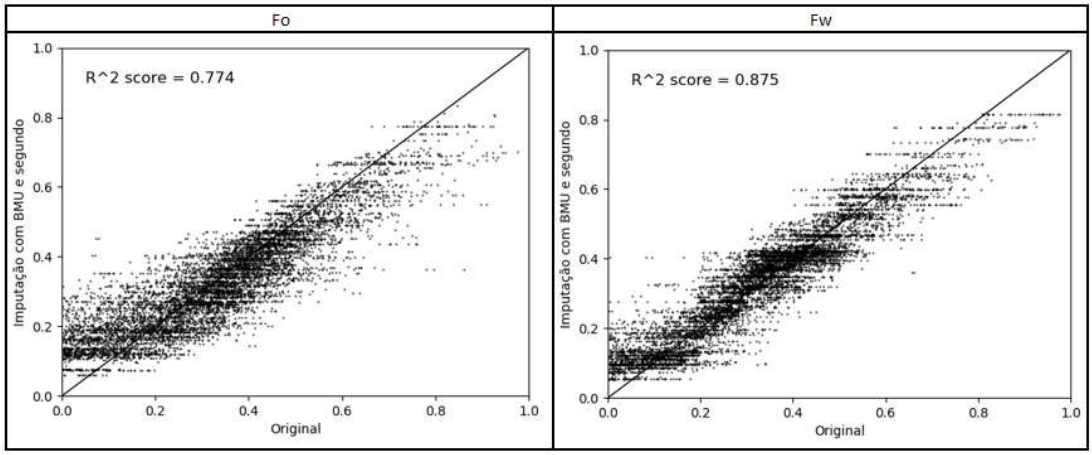
Vale discutir que, para o escopo do presente estudo, não foram encontradas diferenças relevantes entre a imputação por simples substituição pelo BMU e sua versão proporcional. No entanto, para amostras onde as variáveis apresentem maior variância, e cuja amostra de treinamento de treinamento não represente tal variância, passa-se a observar maior discrepância entre os dois métodos. Uma vez que o escopo do estudo é focado na densiometria *gamma*, que está bem representada pela abordagem tomada, opta-se por não explorar tal comparação, deixando espaço para futuros trabalhos.

Figura 12– Matriz de classificação para os dados imputados por substituição pelo BMU proporcional.



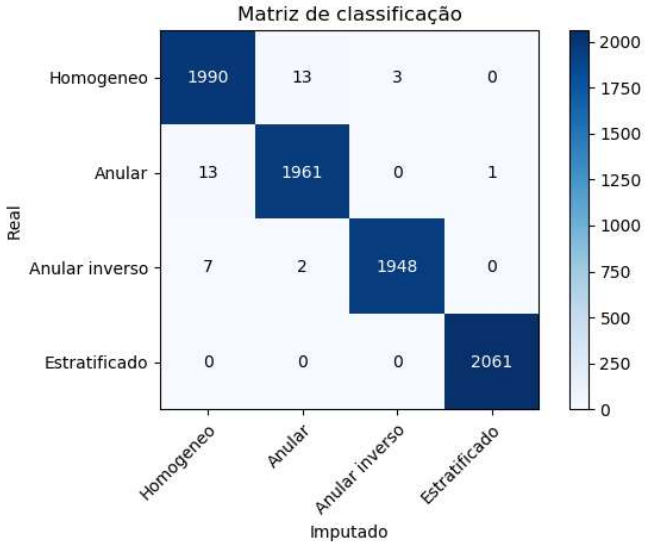
Passa-se então para as metodologias que combinam conceitos estatísticos como a média, para imputação. No caso mais próximo daqules primeiros já discutidos, a imputação pela média entre o BMU e o segundo candidato para BMU, se observa maior variância nos dados obtivos, e menor representação do modelo em relação a seus dados originais, com coeficientes de determinação menores (Gráfico 3). Uma vez que é combinada a média, e são considerados para imputação valores de um vetor que não é o que melhor o representa, do segundo candidato a BMU, aumenta-se a variância dos resultados. Quanto ao Bias de correlação, a metodologia apresentou valor levemente maior que a substituição pelo BMU, o que indica uma maior distorção da correlação original entre variáveis, e reforça a tese de aumento de variância.

Gráfico 3 – *Plot* de frações de óleo (a) e água (b) imputados com BMU e segundo BMU e dados originais.



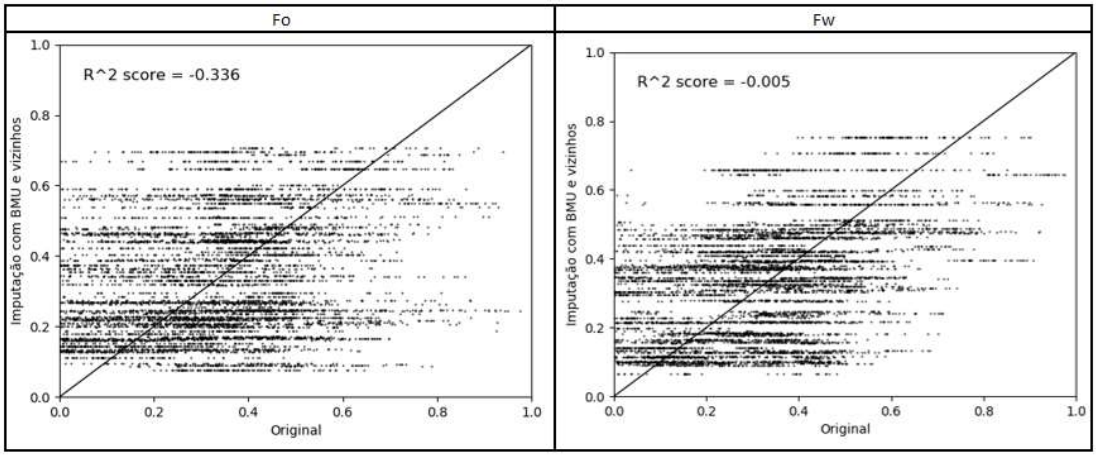
Pelo critério classificação, a soma total de erros se mantém quando comparada aos casos de substituição pelo BMU (Figura 13). Porém, destaca-se a mudança na distribuição dos erros. Como comentado anteriormente, a distinção entre topologias homogênea e anular, em casos onde todos os raios passam por três fases por exemplo, pode ser difícil, o que faz com que apareçam exemplos de erros nesse sentido. O mesmo ocorre com o caso anular inverso, por sua semelhança à topologia anular, o que pode ser observado no caso em questão, onde 38 dos 39 erros de classificação podem ser discutidos por tal argumento.

Figura 13– Matriz de classificação para os dados imputados por substituição pela média entre o BMU e o segundo candidato a BMU.



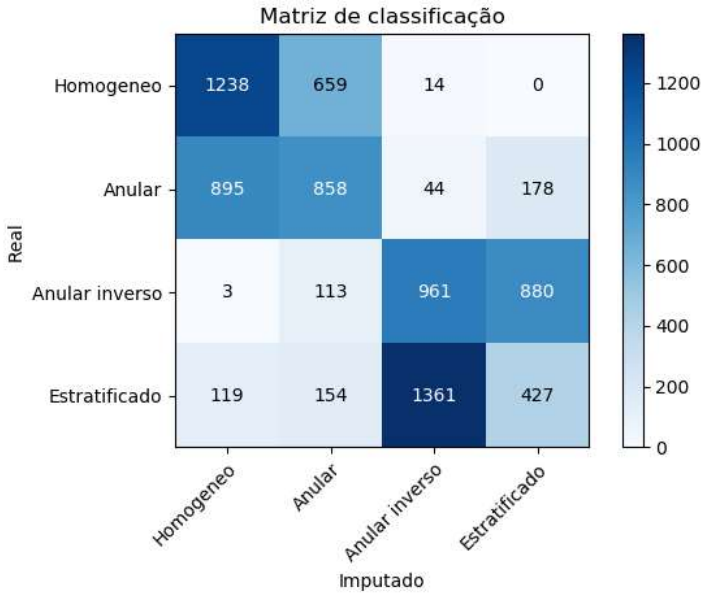
Chega-se então ao último caso abordado, e que apresenta os piores resultados: a imputação pela média entre o BMU e seus seis vizinhos, para o grid hexagonal da rede. A variâcia do modelo obtido pela metodologia é alta e sua média destoa da amostra original, o que é reforçado pelo coeficiente de determinação negativo observado (Gráfico 4). Para efeito de análise, o coeficiente de determinação negativo indica um modelo que não representa em nada a amostra original. Basicamente, o modelo se mostra pior que uma simples linha horizontal para explicar a amostra, o que exlica seu fator Ssreg maior que Sstot , resultando no valor negativo.

Gráfico 4 – *Plot* de frações de óleo (a) e água (b) imputados com BMU e seus vizinhos e dados originais.



Quanto ao critério Bias de Correlação, a metodologia também se destaca. Com o valor de 25.74% para o coeficiente, pode-se dizer que o modelo de imputação pela média entre o BMU e seus vizinhos provoca elevada distorção na amostra, as correlações entre variáveis são altamente enviesadas. Os erros de classificação também se itensificam, com um total de 4420 erros, representando aproximadamente 55% da amostra de testes (Figura 14). Conclui-se, portanto, que tal metodologia apresentou os piores resultados, de acordo com as métricas de validação utilizadas. Além de não representar a amostra original, foi incrementada alta variabilidade aos dados, assim como distorcidas as correlações entre variáveis. A taxa de erro de classificação também foi intensificada.

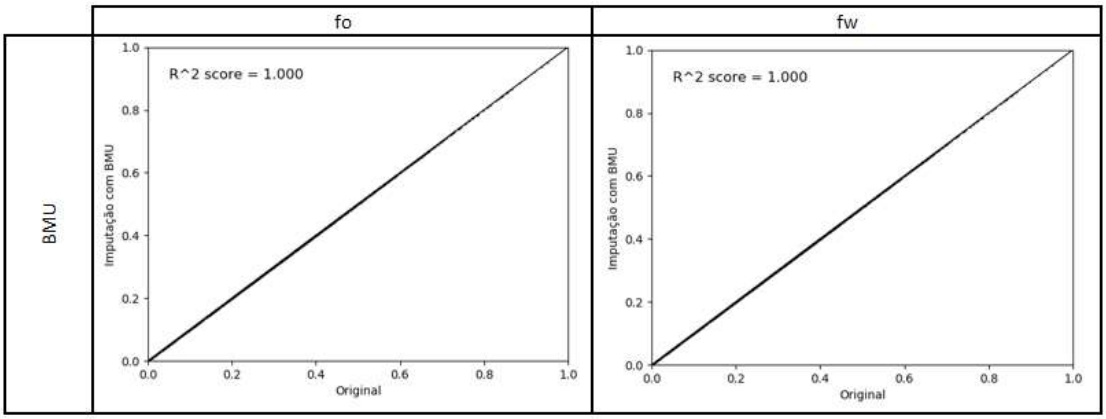
Figura 14– Matriz de classificação para os dados imputados por substituição pela média entre o BMU e seus vizinhos.



4.2.1. Verificação de *overfitting* de dados

Para efeitos de comparação e destaque ao conceito de *overfitting*, realiza-se por fim um teste de imputação utilizando a mesma amostra para o treino da rede e para os testes. Isto testa a memória da rede em relação aos dados de treinamento. Como resultado, observa-se algo que não representa um uso real: um coeficiente de determinação de valor 1 perfeito. Ao conhecer previamente por completo a amostra que seria utilizada para imputação, ao invés de conhecer a amostra para aprender a imputar, o trabalho da rede se resume a um ajuste fino, ou seja, memorizar. Em termos práticos, de nada vale utilizar o resultado, uma vez que o sentido da imputação presume um não conhecimento da amostra por completo. Para evitar tal viés no estudo, opta-se por utilizar diferentes amostras para treino da rede e realização dos testes.

Gráfico 5 – Plot de frações de óleo (a) e água (b) imputados com BMU e dados originais, treinando a rede com a mesma amostra utilizada para os testes de imputação.



5 CONCLUSÕES

As métricas que utilizam o BMU para imputação, portanto, mostram-se mais eficazes para o processo de imputação de dados faltantes em medições de densiometria *gamma*. Dado que o BMU é por definição o vetor que mais se assemelha ao vetor em questão, pelo critério distância euclidiana, faz sentido a constatação. Além de sua maior acurácia, a metodologia provoca menores distorções nas relações entre variáveis como um todo. No caso da densiometria *gamma*, a metodologia se mostra eficaz para classificação da topologia de escoamento sem a utilização da série de formulas comumente utilizadas na indústria para o mesmo fim, o que comprova a usabilidade da metodologia em larga escala. A imputação pela média entre o BMU e seus vizinhos, por outro lado, se mostra ineficaz. O modelo encontrado não representa os dados originais, assim como provoca distorções de correlações e aumento de variância.

O escopo do presente trabalho foca na comparação e aplicabilidade de quatro metodologias de imputação de dados faltantes em amostras de medições de densiometria *gamma*, em especial na indústria do petróleo, abordando escoamentos mono, bi e trifásicos de água, óleo e gás. Sendo assim, abre espaço para futuros estudos que abordem e respondam perguntas que não foram exploradas pois desviam do objetivo: a comparação de metodologias para imputação aplicadas à densiometria *gamma*. A título de exemplo: testes com tamanhos diferentes de mapas, aumentando ou diminuindo a concentração de unidades topológicas; testes com amostras de experimentos que tenham como característica uma maior variância; testes de diferentes tamanhos de amostra; e a comparação de outras metodologias encontradas na literatura.

Por ora, as métricas coeficiente de determinação e Bias de Correlação permitem elencar a substituição pelo BMU como a técnica que apresenta melhor performance e eficácia entre as 4 estudadas. Em segundo e terceiro lugares, respectivamente, encontram-se a substituição proporcional pelo BMU e a média entre o BMU e o segundo candidato a BMU. Por fim, destaca-se a pior performance da substituição pela média entre o BMU e seus vizinhos. A técnica, além de distorcer as relações entre variávies, em nada representou seus dados originais.

REFERÊNCIAS

ALIMONTI, C.; FALCONE, G. **Knowledge Discovery in Databases and Multiphase Flow Metering: The Integration of Statistics, Data Mining, Neural Networks, Fuzzy Logic, and Ad Hoc Flow Measurements Towards Well Monitoring and Diagnosis**. Proceedings - SPE Annual Technical Conference and Exhibition. **Anais**...2002

BELO, F. A.; MENDES DE MOURA, L. F. High frequency electronic transducer for multiphase flow measurements. **Revista Brasileira de Ciencias Mecanicas/Journal of the Brazilian Society of Mechanical Sciences**, 1999.

BISHOP, C. M.; JAMES, G. D. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. **Nuclear Inst. and Methods in Physics Research, A**, 1993.

BRAVO, C. et al. State of the art of artificial intelligence and predictive analytics in the E&P industry: A technology survey. **SPE Journal**, 2014.

CABANES, G.; BENNANI, Y. Learning Topological Constraints in Self-Organizing Map. In: [s.l: s.n.]. p. 367–374.

CARTWRIGHT, M. H.; SHEPPERD, M. J.; SONG, Q. **Dealing with missing software project data**. Proceedings - International Software Metrics Symposium. **Anais**...2003

FALCONE, G.; HEWITT, G. F.; ALIMONTI, C. **Multiphase Flow Metering: Principles and Applications**. [s.l: s.n.].

FERLIN, C. **Imputação Multivariada: Uma Abordagem em Cascata**. [s.l.] UFRJ, 2008.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. [s.l: s.n.].

HAYKIN, S. **Neural Networks and Learning Machines**. [s.l: s.n.].

HU, M., SALVUCCI, S.M., COHEN, M. P. **Evaluation of some popular imputation algorithms. ... American Statistical Association**. The survey research methods section of the ASA. **Anais**...1998

HUBEL, D. H.; WIESEL, T. N.; STRYKER, M. P. Orientation columns in macaque monkey visual cortex demonstrated by the 2-deoxyglucose autoradiographic technique. **Nature**, 1977.

JUNNINEN, H. et al. Methods for imputation of missing values in air quality data sets. **Atmospheric Environment**, v. 38, n. 18, p. 2895–2907, jun. 2004.

KNUDSEN, E. Computational Maps In The Brain. **Annual Review of Neuroscience**, 1987.

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

KOHONEN, T. The self-organizing map. **Neurocomputing**, 1998.

LECUN, Y. A. et al. Efficient backprop. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, 2012.

MAGNANI, M. **Techniques for Dealing with Missing Data in Knowledge Discovery Tasks**. [s.l.] University of Bologna, 2004.

MALEK, M. A et al. Imputation of Time Series Data via Kohonen Self Organizing Maps in the Presence of Missing Data. **World Academy of Science, Enginering and Technology**, 2008.

MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. In: **Systems Research for Behavioral Science: A Sourcebook**. [s.l: s.n.].

MERZENICH, M. M. et al. Topographic reorganization of somatosensory cortical areas 3b and 1 in adult monkeys following restricted deafferentation. **Neuroscience**, 1983.

MILLIGAN, G. W.; COOPER, M. C. A study of standardization of variables in cluster analysis. **Journal of Classification**, 1988.

NATITA, W.; WIBOONSAK, W.; DUSADEE, S. Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand. **International Journal of Modeling and Optimization**, v. 6, n. 1, p. 61–65, 2016.

PELLEGRINI, S. DE P. **Estimação dinâmica em tomografia por impedância elétrica com modelos adaptativos.** São Paulo: Universidade de São Paulo, 31 maio 2019.

R. RALLO, J. F. AND F. G. **Multiple imputation of missing data using self organizing map ensembles**. [s.l.] Escola Tècnica Superior d'Enginyeria, 2005.

SAMAD, T.; HARP, S. A. Self-organization with partial data. **Network: Computation in Neural Systems**, 1992.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, 2000.

SCHMIDT, C. R.; REY, S. J.; SKUPIN, A. Effects of irregular topology in spherical self-organizing maps. **International Regional Science Review**, 2011.

SHALAGINOV, A.; FRANKE, K. **A new method for an optimal SOM size determination in neuro-fuzzy for the digital forensics applications**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais**...2015

SOARES, J. D. A. **Pré-Processamento em Mineração de Dados: Um Estudo Comparativo em Complementação**. [s.l.] UFRJ, 2007.

SUGA, N.; TSUZUKI, K. Inhibition and level-tolerant frequency tuning in the auditory cortex of the mustached bat. **Journal of Neurophysiology**, 1985.

THORN, R.; JOHANSEN, G. A.; HJERTAKER, B. T. **Three-phase flow measurement**

in the petroleum industryMeasurement Science and Technology, 2013.

TURING, A. M. Computing machinery and intelligence. In: **Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer**. [s.l: s.n.].

ULLMAN, S. et al. 9.54 Class 13. **Unsupervised learning Slides**, 2014.

VESANTO, J. et al. Self-organizing map in Matlab : the SOM Toolbox. **Proceedings of the Matlab DSP Conference**, p. 35–40, 1999.

WARREN, S. M.; WALTER, H. P. A logical calculus of ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, 1943.

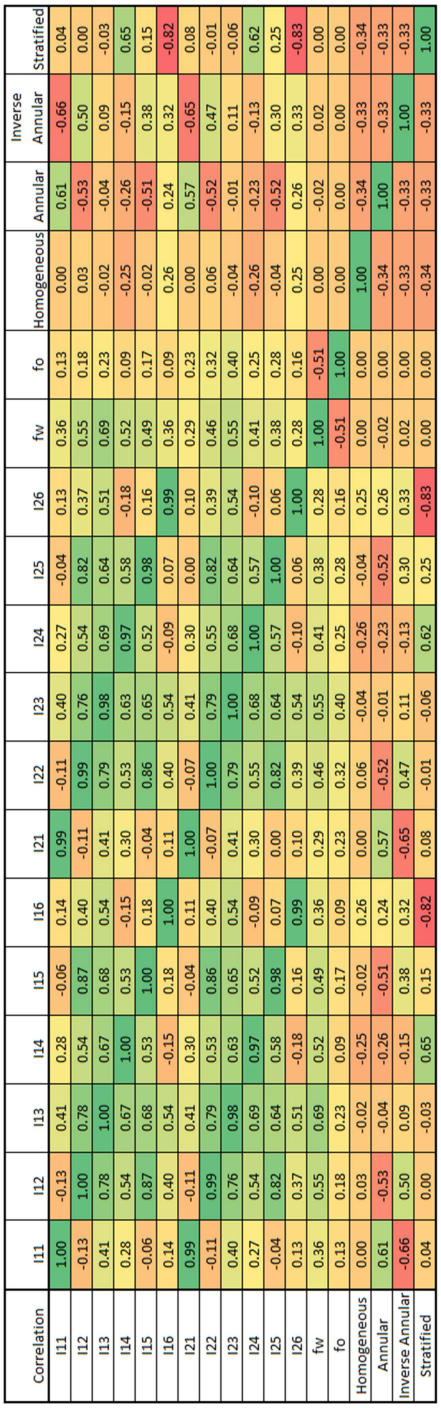
WEISS, W. W.; BALCH, R. S.; STUBBS, B. A. **How Artificial Intelligence Methods Can Forecast Oil Production**. Proceedings - SPE Symposium on Improved Oil Recovery. **Anais**...2002

WITTEK, P. et al. Somoclu: An efficient parallel library for self-organizing maps. **Journal of Statistical Software**, 2017.

ZHANG, D. A Coefficient of Determination for Generalized Linear Models. **The American Statistician**, v. 71, n. 4, p. 310–316, 2 out. 2017.

APÊNDICE A

Tabela de correlações entre variáveis para a amostra de medições de densiometria *gamma* gerada para realização do estudo. Tonalidades vermelhas indicam correlações negativas, enquanto as verdes correlações altas positivas. Conforme descrito na análise do *heatmap*, L11e L21. Entre variáveis como L11 e L21, de raios que atravessam o mesmo caminho, se observa alta correlação.



Correlation	I11	I12	I13	I14	I15	I16	I21	I22	I23	I24	I25	I26	fw	fo	Homogeneous	Annular	Inverse Annular	Stratified
I11	1.00	-0.13	0.41	0.28	-0.06	0.14	0.99	-0.11	0.40	0.27	-0.04	0.13	0.36	0.13	0.00	0.61	-0.66	0.04
I12	-0.13	1.00	0.78	0.54	0.87	0.40	-0.11	0.99	0.76	0.54	0.82	0.37	0.55	0.18	0.03	-0.53	0.50	0.00
I13	0.41	0.78	1.00	0.67	0.68	0.54	0.41	0.79	0.98	0.69	0.64	0.51	0.69	0.23	-0.02	-0.04	0.09	-0.03
I14	0.28	0.54	0.67	1.00	0.53	-0.15	0.30	0.53	0.63	0.97	0.58	-0.18	0.52	0.09	-0.25	-0.26	-0.15	0.65
I15	-0.06	0.87	0.68	0.53	1.00	0.18	-0.04	0.86	0.65	0.52	0.98	0.16	0.49	0.17	-0.02	-0.51	0.38	0.15
I16	0.14	0.40	0.54	-0.15	0.18	1.00	0.11	0.40	0.54	-0.09	0.07	0.99	0.36	0.09	0.26	0.24	0.32	-0.82
I21	0.99	-0.11	0.41	0.30	-0.04	0.11	1.00	-0.07	0.41	0.30	0.00	0.10	0.29	0.23	0.00	0.57	-0.65	0.08
I22	-0.11	0.99	0.79	0.53	0.86	0.40	-0.07	1.00	0.79	0.55	0.82	0.39	0.46	0.32	0.06	-0.52	0.47	-0.01
I23	0.40	0.76	0.98	0.63	0.65	0.54	0.41	0.79	1.00	0.68	0.64	0.54	0.55	0.40	-0.04	-0.01	0.11	-0.06
I24	0.27	0.54	0.69	0.97	0.52	-0.09	0.30	0.55	0.68	1.00	0.57	-0.10	0.41	0.25	-0.26	-0.23	-0.13	0.62
I25	-0.04	0.82	0.64	0.58	0.98	0.07	0.00	0.82	0.64	0.57	1.00	0.06	0.38	0.28	-0.04	-0.52	0.30	0.25
I26	0.13	0.37	0.51	-0.18	0.16	0.99	0.10	0.39	0.54	-0.10	0.06	1.00	0.28	0.16	0.25	0.26	0.33	-0.83
fw	0.36	0.55	0.69	0.52	0.49	0.36	0.29	0.46	0.55	0.41	0.38	0.28	1.00	-0.51	0.00	-0.02	0.02	0.00
fo	0.13	0.18	0.23	0.09	0.17	0.09	0.23	0.32	0.40	0.25	0.28	0.16	-0.51	1.00	0.00	0.00	0.00	0.00
Homogeneous	0.00	0.03	-0.02	-0.25	-0.02	0.26	0.00	0.06	-0.04	-0.26	-0.04	0.25	0.00	0.00	1.00	-0.34	-0.33	-0.34
Annular	0.61	-0.53	-0.04	-0.26	-0.51	0.24	0.57	-0.52	-0.01	-0.23	-0.52	0.26	-0.02	0.00	-0.34	1.00	-0.33	-0.33
Inverse Annular	-0.66	0.50	0.09	-0.15	0.38	0.32	-0.65	0.47	0.11	-0.13	0.30	0.33	0.02	0.00	-0.33	-0.33	1.00	-0.33
Stratified	0.04	0.00	-0.03	0.65	0.15	-0.82	0.08	-0.01	-0.06	0.62	0.25	-0.83	0.00	0.00	-0.34	-0.33	-0.33	1.00

APÊNDICE B

Gráfico 6 – Histograma dos valores simulados de fração de óleo para treino da rede neural.

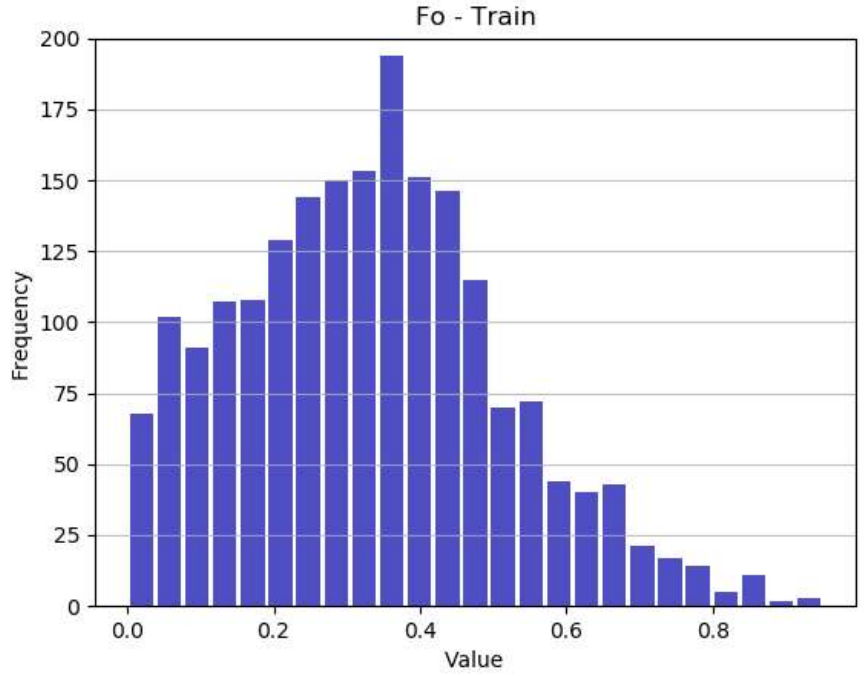
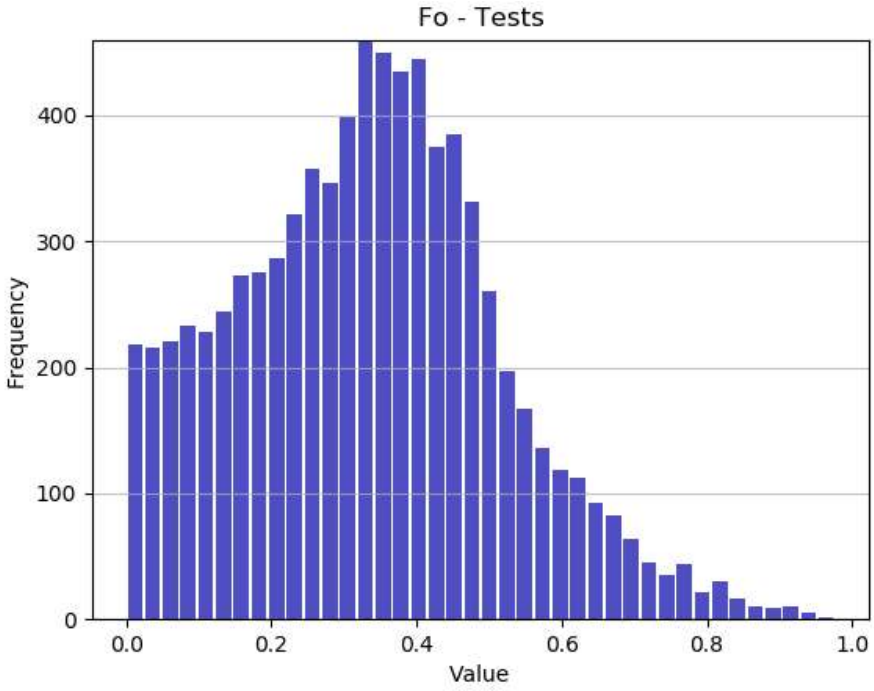


Gráfico 7 – Histograma dos valores simulados de fração de óleo para testes.



APÊNDICE C

Figura 15– Clusters identificados após o treinamento da rede.

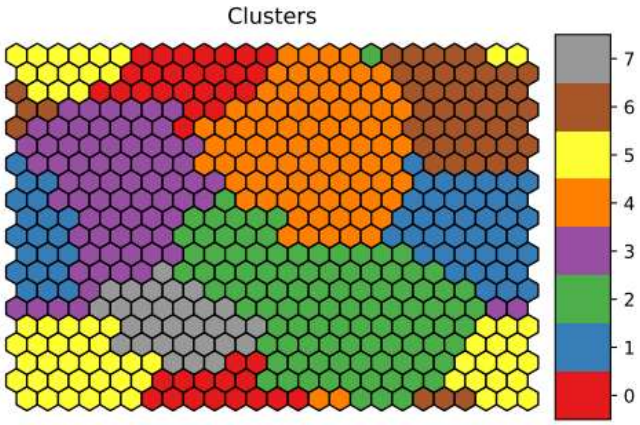
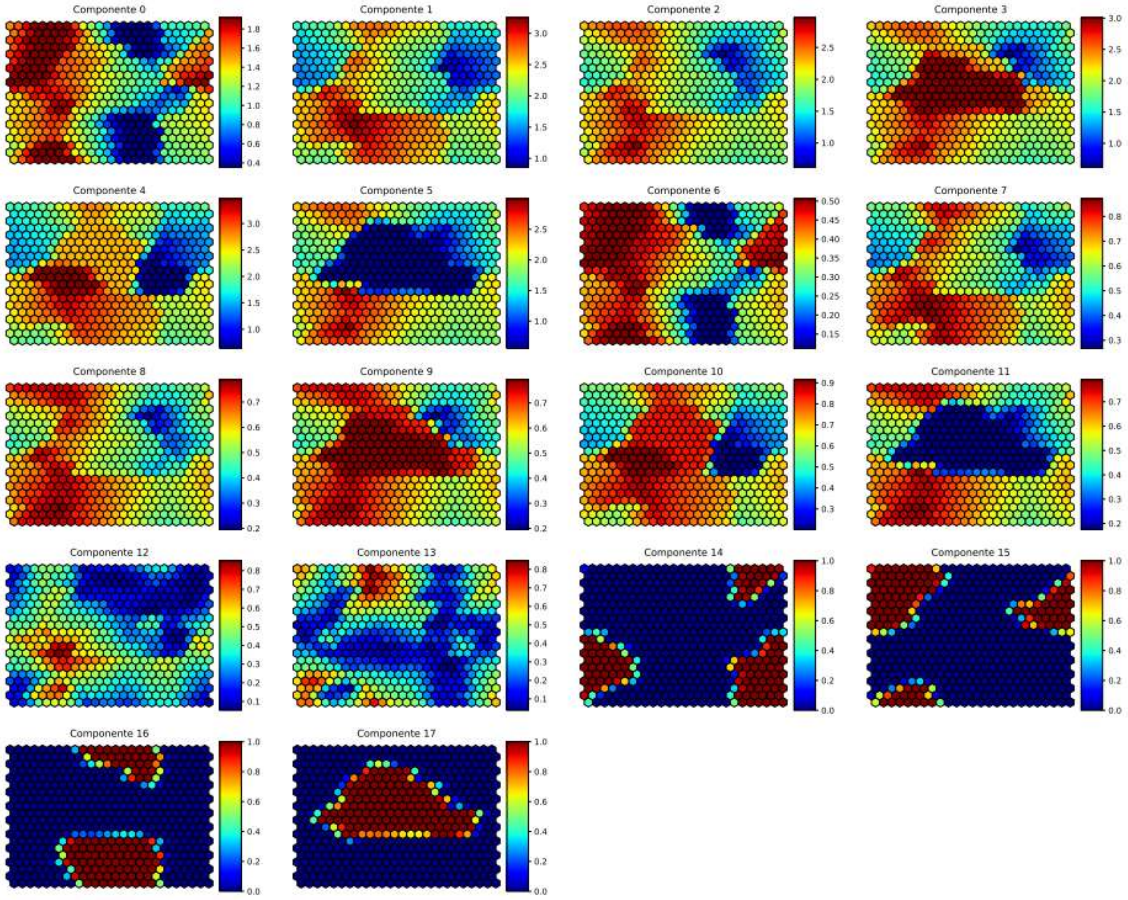


Figura 16– *Component plots* resultantes após o treinamento da rede.



Na a Figura 16, cada uma das componentes representa uma das variáveis. Da componente 1 à componente 5, as medições realizadas para o primeiro comprimento

de onda simulado conforme Figura 7, sendo a componente 0 a variável L11. Da componente 6 à 13, as medições enxergadas para o segundo comprimento de onda simulado. Por fim, as componentes 14 a 17 represenam as quatro topologias de escoamento estudadas, respectivamente homogêneo, anular, anular inverso e estratificado.

Universidade de São Paulo

Engenharia de Petróleo – Escola Politécnica

Número: 9347671USP **Data: 31/10/2019**

IMPUTAÇÃO COM SELF-ORGANIZING MAPS: COMPARAÇÃO DE ABORDAGENS APLICADAS À DENSIOMETRIA GAMMA

João Guilherme dos Santos Prudente do Amaral

Orientador: Prof. Dr. Rafael dos Santos Gioria

Artigo Sumário referente à disciplina PMI1096 – Trabalho de Formatura para Engenharia de Petróleo II
Este artigo foi preparado como requisito para completar o curso de Engenharia de Petróleo na Escola Politécnica da USP. Template versão 2018v11.

Resumo

O trabalho traz uma abordagem estatística e experimental que avalia a capacidade da utilização de Redes Neurais Artificiais do tipo Self-organizing Maps (SOM) na imputação de dados para identificação de topologia de escoamentos multifásicos de água-óleo-gás e cáculo de suas frações volumétricas em medições de densiometria gamma, o que ressalta sua importância para a indústria do petróleo, onde um dos grandes desafios ainda é medir propriedades de escoamentos multifásicos de óleo, gás e água. Os dados são imputados seguindo 4 diferentes metodologias, sendo elas imputação simples com o BMU (Best Matching Unit, ou neurônio mais representativo), proporcional com o BMU, com a média entre o BMU e o segundo candidato a BMU, e com a média entre o BMU e seus vizinhos, e seus resultados são comparados com base em coeficiente de determinação, Bias de Correlação, e acurácia na classificação. As metodologias baseadas no BMU, substituição simples e proporcional, se mostram mais adequadas.

Abstract

The assignment brings out a statistical and experimental approach that evaluates the ability of using Artificial Neural Networks of the Self-organizing Maps (SOM) type to data imputation and classification of multiphase flows by gamma densiometry, which highlights its importance for the petroleum industry, where one of the great challenges is still to measure properties of multiphase flows of oil, gas and water. The data are imputed following 4 different methodologies, BMU replacement (most representative neuron), proportional BMU replacement, BMU and 2nd mean replacement, and BMU and neighbors mean replacement, and their results are compared based on determination coefficient, Correlation Bias, and classification accuracy. BMU-based methodologies, BMU replacement and proportional BMU replacement, are the most appropriate.

1. Introdução

A Inteligência Artificial tem recebido amplo destaque nos últimos anos, sobretudo pela capacidade de análise de dados e identificação de padrões. Na Engenharia de Petróleo, com o recente interesse e entusiasmo da indústria em análises em tempo real de poços e campos inteligentes, a IA tem sido centro de atenções (BRAVO et al., 2014). Incorporado a tal tendência, o objetivo do presente trabalho é desenvolver um estudo prático e estatístico comparando diferentes modelos de imputação utilizados em redes neurais do tipo SOM (Self-Organizing Maps) aplicados a medições de densiometria gamma. Os SOM têm sido amplamente utilizados em machine learning, destacando-se pelo fato de não serem supervisionados e identificarem relações pouco triviais. Seu algoritmo mapeia o conjunto de dados de

treinamento por competitividade e resulta numa superfície que representa a distribuição da amostra num espaço bidimensional.

O conceito de imputação envolve a estimativa para preenchimento de uma variável faltante num vetor de dados n-dimensional de acordo com um critério baseado em suas demais (n-1) variáveis conhecidas. O uso de técnicas matemáticas clássicas permite preencher tais lacunas regressões dos demais vetores e substituição pela média. No entanto, os valores encontrados são tendenciosos, pois a abordagem para análise da amostra não representa a população (GELMAN; HILL, 2007). As redes neurais aparecem então como um mecanismo que permite não só produzir estimativas livres de subjetividade, como também aproveitar a alta capacidade de processamento de dados das máquinas.

O estudo é dividido em 3 principais etapas. Nas Referências Bibliográficas são introduzidos o funcionamento de redes neurais artificiais do tipo Self-Organizing Maps e o que consiste e os motivos para a realização de medições em escoamentos multifásicos. Em seguida, já na seção de metodologia, desenvolve-se como se desenvolve a geração de dados simulados para densiometria gamma em um escoamento multifásico de óleo, gás e água, divididos em grupos para treinamento de rede e imputação. Com a rede treinada, realiza-se um estudo prático de comparação entre diferentes métodos de imputação de dados que têm como base o SOM. Ao final do estudo, na discussão dos resultados, métricas estatísticas são utilizadas para comparar a capacidade das diferentes metodologias abordadas.

2. Revisão Bibliográfica

2.1. Self-Organizing Maps

Os Self-Organizing Maps (SOM) são redes neurais de aprendizado não supervisionado introduzidas por Teuvo Kohonen (KOHONEN, 1982), que aprendem por competição. Seus neurônios são organizados em uma camada uni ou bidimensional e ordenados de maneira a criar um novo sistema de coordenadas para as diferentes variáveis de entrada (KOHONEN, 1998). Ainda, em sua performance ocorre o mapeamento dos dados de entrada, de forma que cada vetor é representante de variáveis estatísticas intrínsecas da amostra. Em qualquer estágio, cada um dos vetores é preservado em seu próprio contexto, e dessa forma, se realizam conexões sinápticas próximas que contribuem para sua preservação topológica.

2.2. Imputação de dados

Imputação é o nome dado para a estimativa de dados faltantes que comprometem datasets e dificultam análises de dados e a determinação de inferências estatísticas (HU, M., SALVUCCI, S.M., COHEN, 1998). O uso do SOM como imputador de dados tem sido explorado em diversas áreas de aplicação. Por não considerar a amostra inteira no cálculo da média para substituição, mas sim selecionar um número determinado de neurônios similares ao vetor, o algoritmo se mostra muito mais assertivo que outras metodologias puramente estatísticas.

No estudo aqui apresentado, quatro metodologias são analisadas e comparadas: substituir pelo BMU encontrado para o vetor na rede (Eq. 1), algo já feito na literatura; substituir pelo BMU encontrado para o vetor na rede, multiplicado pelo fator de projeção do vetor sobre o BMU (Eq. 2); encontrar o BMU e o segundo neurônio candidato a BMU, fazer a média deles e imputar (Eq. 3); e encontrar o BMU, fazer uma média do BMU com demais pontos que formam o cluster, e imputar (Eq. 4).

$$X_n = X_{n_{BMU}} \tag{1}$$

$$X_n = X_n \ \frac{(BMU \times AMO)}{|BMU| \times |AMO|} \tag{2}$$

$$X_n = \frac{(X_{n_{BMU_1}} + X_{n_{BMU_2}})}{2} \tag{3}$$

$$X_n = \frac{\sum X_{n_y}}{m} \tag{4}$$

Onde y é o conjunto do BMU e seus vizinhos, e m o número de vetores desse conjunto.

2.3. Medições em escoamentos multifásicos

Conseguir medir propriedades de escoamentos multifásicos de óleo, água e gás em tubulações ainda é um dos grandes desafios na indústria do petróleo (THORN; JOHANSEN; HJERTAKER, 2013). Existem inúmeras metodologias na literatura para medição destas propriedades, cada uma com suas peculiaridades. A densiometria gamma e a impedância elétrica-magnética são duas metodologias não intrusivas utilizadas para análise de escoamentos multifásicos que não dependem de homogeneização (BELO; MENDES DE MOURA, 1999). Entre as vantagens de usá-las, pode-se citar a realização da análise sem depender da presença de um furo no trecho da tubulação a ser analisado ou da instalação de sistemas defletores que acabam mudando suas propriedades naturais. Ao realizar medições por diversos raios distribuídos uniformemente no entorno de uma tubulação, a densiometria gamma permite calcular propriedades interessantes do fluxo que por ela passa, tal como a configuração das fases que o compõem e suas respectivas frações (Figura 1).

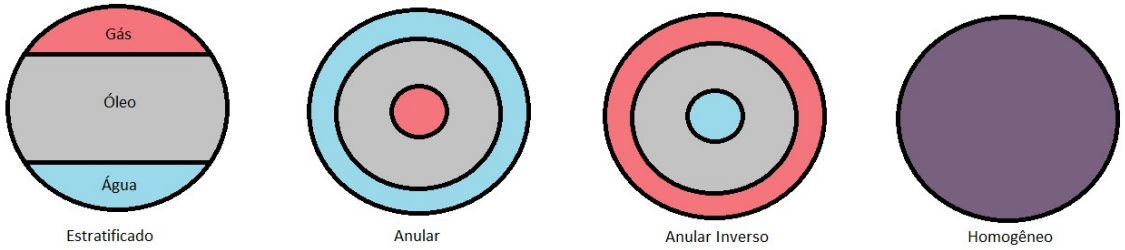


Figura 1 - Seção transversal em tubulação de transporte de óleo. Fonte: Adaptado (BISHOP; JAMES, 1993).

3. Metodologia

3.1. Geração de dados

A geração de dados de medições de densiometria *gamma* se baseia em metodologia apresentada na literatura (BISHOP; JAMES, 1993).

3.2. Escolha do Mapa

O mapa definido visa estrategicamente potencializar as análises a serem feitas ressaltando diferenças entre os diferentes tipos de topologia. Para tal, recorre-se à literatura em busca de melhores práticas e fórmas heurísticas para definição dos parâmetros a serem utilizados. O grid local escolhido é hexagonal, de forma a aumentar a sensibilidade dos neurônios a variações locais. Globalmente, opta-se pela estrutura toroidal, garantindo simetria na rede de forma que neurônios de borda tenham todos o mesmo número de conexões. Quanto ao tamanho do mapa, este está diretemente ligado à acurácia e a interpretabilidade dos resultados. Para grids hexagonais, a Eq. 5 define seu número ótimo de neurônios em função no número m de amostras no dataset de treinamento e a Eq. 6 a proporção ótima entre dimensões no mapa, onde COV_{V1} e COV_{V2} são as máximas covariâncias observadas entre variáveis ainda nos dados de treinamento (VESANTO et al., 1999). O número de amostras utilizadas para treinamento é de 2000, o que resulta em aproximadamente 230 vetores ótimos.

$$Nº\ de\ neurônios = 5 \times \sqrt{m} \tag{5}$$

$$\frac{D(X)}{D(Y)} = \frac{COV_{V1}}{COV_{V2}} \tag{6}$$

A função vizinhança adotada é a Gaussiana, que determinará a taxa de mudança da vizinhança ao entorno do neurônio vencedor. O coeficiente utilizado para a função é de 0.5. Ela influenciará diretamente o treinamento da rede, e em algoritmos aplicados para classificação, a função Gaussiana, combinada com uma taxa de aprendizado linear, geram ótima performance com baixo erro de quantização (NATITA; WIBOONSAK; DUSADEE, 2016).

3.3. Imputações

São abordados quatro tipos diferentes de imputação de variáveis faltantes. O primeiro deles segue o que é feito por muitos exemplos na literatura: a substituição do vazio por seus correspondentes do BMU encontrado para o vetor (Eq.1). Uma variável dele, ainda, é multiplicá-lo pela projeção do vetor sobre o BMU (Eq. 2). Já as duas outras metodologias combinam o SOM a métodos estatísticos para determinação dos valors imputados: substituição pela média entre os correspondentes no primeiro e no segundo candidato a BMU (Eq.3), e a substituição pela média entre os valores correspondentes no BMU e seus vizinhos (Eq.4). Para tal, exportam-se os dados da rede treinada e se realizam as estimativas.
Entre os quatro métodos utilizados, aqueles que combinam o SOM com médias estatísticas promovem a imputação de valores que representam melhor sua vizinhança como um todo, aumenta-se a variância dos dados. Por outro lado, na substituição pelo BMU e em sua variação proporcional, não há incremento relevante de variância podendo induzir uma tendência.

3.4. Validação

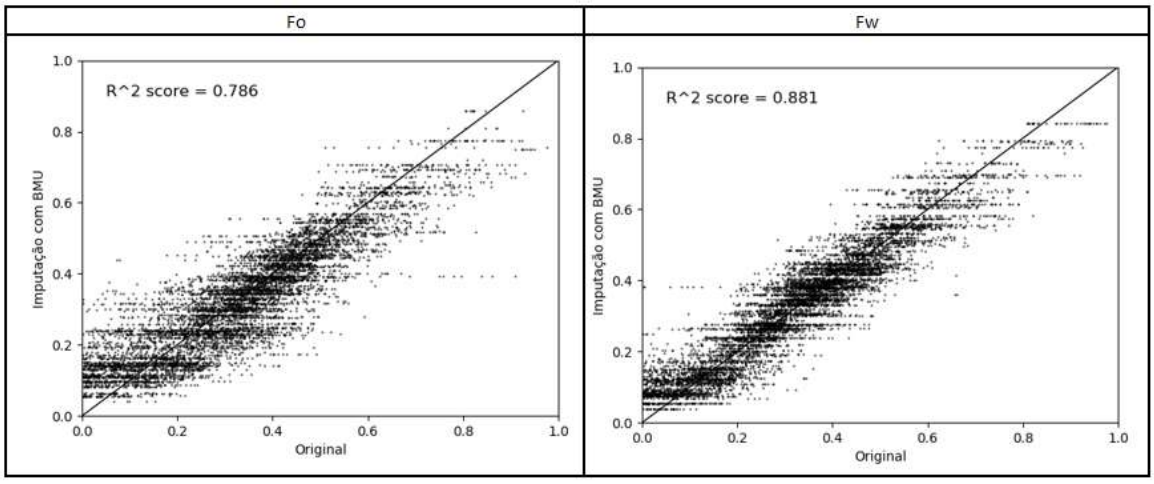
A validação dos dados imputados segue duas principais metodologias de comparação estatística: o Coeficiente de Determinação e o que chama-se de Bias da Correlação. No desenvolvimento do estudo, adota-se função já implementada em Python para cálculo do coeficiente de determinação. Já para o Bias de Correlação, adota-se metodologia apresentada na literatura (FERLIN, 2008).

4. Resultados

4.1. Imputação de dados

O presente estudo tem como principal objetivo a comparaçao de metodologias de imputação de dados faltantes de densiometria gamma utilizando redes neurais artificiais do tipo SOM. Quatro metodologias distintas foram implementadas: substituição pelo valor do BMU, pelo valor do BMU proporcional à projeção do vetor sobre seu BMU, pela média dos valores do primeiro BMU e do segundo, ou pela média entre o BMU e seus 6 vizinhos. Entre os quatro métodos implementados, espera-se inicialmente que aqueles que combinam o SOM com médias estatísticas promovam a imputação de valores que representam melhor sua vizinhança, no entanto aumenta-se a variância dos dados. Por outro lado, na substituição pelo BMU espera-se que não ocorra incremento relevante de variância.

Gráfico 1 – Plot de frações de óleo (a) e água (b) imputados com BMU e dados originais.



Para a imputação por substituição pelos valores do BMU encontrado, se observa altos coeficientes de determinação (Gráfico 1). O BMU de um vetor é definido como o vetor de menor distância euclidiana na rede em relação ao vetor a ser imputado. Partindo desse princípio, faz sentido imaginar que uma

estimativa dos dados a serem imputados por aqueles encontrados no BMU do vetor com dados faltantes seja uma boa alternativa, dado que toda a topologia da rede, com suas relações entre vizinhos e as considerações estatísticas que estão nelas implícitas, aponta para alta semelhança entre os dois vetores. Ainda, a meodologia apresentou o menor valor para Bias de Correlação entre as abordadas pelo presente estudo. Quanto menor o Bias de Correlação, menor a distorção da correlação entre variáveis do modelo em relação aos dados originais. Para a substituição pelo BMU e BMU proporcional, encontraram-se os menores valores de, respectivamente, 2.29% e 2.30%. Para a suabstituição pela media entre o BMU e o segundo, 3.07%. Já a imputação pela media entre o BMU e seus vizinhos apresentou a maior distorção, um Bias de Correlação de 25.74%.

Para efeito de comparação, analisa-se ainda a assertividade da metodologia de imputação na classificação da topologia do escoamento. Para a população de 8000 vetores de testes, apenas 39 receberam classificação que não condiz com sua topologia original. Destaca-se ainda o fato de a maior quantidade dos erros se concentrar entre os casos homogêneo e anular. De acordo com quantas fases cada raio atravessa, torna-se mais difícil de diferenciar as duas topologias. No escoamento homogêneo, os raios necessariamente atravessarão 3 fases de escoamento, enquanto no caso anular nem sempre. Num caso anular onde todos os raios atravessam 3 fases de escoamento, como exemplo, não se tem este gatilho de distinção entre as duas topologias, o que torna menor sua assertividade.

Analogamente à imputação por substituição pelo BMU, a metodologia por substituição pelos valores do BMU, proporcionalmente à projeção do vetor sobre o BMU, apresenta os resultados mais favoráveis. Dado que em muito dos casos o fator de proporção estimado pela projeção é aproximadamente 1, seus resultados são semelhanes. Iguais quando comparados sob o critério coeficiente de determinação, de 0.786 para Fo e 0.881 para Fw, e matriz de classificação , apresentando 39 erros, e próximos ao comparar pelo critério Bias de Correlação, apresentado um valor baixo de 2.30% para a medida de distorção.

Passa-se então para as metodologias que combinam conceitos estatísticos como a média, para imputação. No caso mais próximo daqules primeiros já discutidos, a imputação pela média entre o BMU e o segundo candidato para BMU, se observa maior variância nos dados obtivos, e menor representação do modelo em relação a seus dados originais, com coeficientes de determinação menores, de 0.774 para Fo e 0.875 para Fw. Uma vez que é combinada a média, e são considerados para imputação valores de um vetor que não é o que melhor o representa, do segundo candidato a BMU, aumenta-se a variância dos resultados. Quanto ao Bias de correlação, a metodologia apresentou valor levemente maior que a substituição pelo BMU, o que indica uma maior distorção da correlação original entre variáveis, e reforça a tese de aumento de variância.

Quanto ao critério classificação, a soma total de erros de classificação se mantém quando comparada aos casos de substituição pelo BMU, com 39 erros. Porém, destaca-se a mudança na distribuição dos erros. Como comentado anteriormente, a distinção entre topologias homogênea e anular pode ser difícil. O mesmo ocorre com o caso anular inverso, por sua semelhança à topologia anular, o que pode ser observado no caso em questão, onde 38 dos 39 erros de classificação podem ser discutidos por tal argumento.

Chega-se então ao último caso abordado, e que apresenta os piores resultados: a imputação pela média entre o BMU e seus seis vizinhos, para o grid hexagonal do presente estudo. A variâcia do modelo obtido pela metodologia é alta e sua média destoa da amostra original, o que é reforçado pelo coeficiente de determinação negativo observado de -0.336 para Fo e -0.005 para Fw. Para efeito de análise, o coeficiente de determinação negativo indica um modelo que não representa em nada a amostra original. Basicamente, o modelo se mostra pior que uma simples linha horizontal para explicar a amostra, o que exlica seu fator Ssreg maior que Sstot , resultando no valor negativo.

Quanto ao critério Bias de Correlação, a metodologia também destaca. Com o valor de 25.74% para o coeficiente, se pode dizer que o modelo de imputação pela média entre o BMU e seus vizinhos provoca elevada distorção na amostra, as correlações entre variáveis são altamente enviesadas. Os erros de

classificação também se itensificam, com um total de 4420 erros, representando aproximadamente 55% de nossa amostra de teste. Conclui-se, portanto, que tal metodologia apresentou os piores resultados, de acordo com as métricas de validação utilizadas. Além de o modelo obtido não representar a amostra original, foi incrementada alta variabilidade aos dados, assim como distorcidas as correlações entre variáveis. A taxa de erro de classificação também foi intensificada.

5. Conclusões

As métricas que utilizam o BMU para imputação, portanto, mostram-se mais eficazes para o processo de imputação de dados faltantes em medições de densiometria gamma. Além de sua maior acurácia, a metodologia provoca menores distorções nas relações entre variáveis como um todo. No caso da densiometria gamma, a metodologia se mostra eficaz para classificação da topologia de escoamento sem a utilização da série de formulas comumente utilizadas na indústria para o mesmo fim, o que comprova a usabilidade da metodologia em larga escala. A imputação pela média entre o BMU e seus vizinhos, por outro lado, se mostra ineficaz. O modelo encontrado não representa os dados originais, assim como provoca distorções de correlações e aumento de variância. Por ora, as métricas coeficiente de determinação e Bias de Correlação permitem elencar a substituição pelo BMU como a técnica que apresenta melhor performance e eficácia entre as 4 estudadas. Em segundo e terceiro lugares, respectivamente, encontram-se a substituição proporcional pelo BMU e a média entre o BMU e o segundo candidato a BMU. Por fim, destaca-se a pior performance da substituição pela média entre o BMU e seus vizinhos. A técnica, além de distorcer as relações entre variávies, em nada representou seus dados originais.

6. Referências

BELO, F. A.; MENDES DE MOURA, L. F. High frequency electronic transducer for multiphase flow measurements. **Revista Brasileira de Ciencias Mecanicas/Journal of the Brazilian Society of Mechanical Sciences**, 1999.

BISHOP, C. M.; JAMES, G. D. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. **Nuclear Inst. and Methods in Physics Research, A**, 1993.

BRAVO, C. et al. State of the art of artificial intelligence and predictive analytics in the E&P industry: A technology survey. **SPE Journal**, 2014.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. [s.l: s.n.].

HU, M., SALVUCCI, S.M., COHEN, M. P. **Evaluation of some popular imputation algorithms. ... American Statistical Association**. The survey research methods section of the ASA. **Anais**...1998

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

KOHONEN, T. The self-organizing map. **Neurocomputing**, 1998.

NATITA, W.; WIBOONSAK, W.; DUSADEE, S. Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand. **International Journal of Modeling and Optimization**, v. 6, n. 1, p. 61–65, 2016.

THORN, R.; JOHANSEN, G. A.; HJERTAKER, B. T. **Three-phase flow measurement in the petroleum industryMeasurement Science and Technology**, 2013.

VESANTO, J. et al. Self-organizing map in Matlab : the SOM Toolbox. **Proceedings of the Matlab DSP Conference**, p. 35–40, 1999.