Estimação do teor de óleos e graxas em água descartada no mar usando redes neurais recorrentes

José Marques Oliveira Júnior [*,**] **Karina dos Reis Teixeira** [*,**]
Ricardo Emanuel Vaz Vargas [*] **Patrick Marques Ciarelli** [**]
Celso J. Munaro [**]

[*] *Petróleo Brasileiro S.A., ES (e-mails: josemarques.oliveira@petrobras.com.br; karinareisteixeira@gmail.com; ricardo.vargas@petrobras.com.br)*
[**] *Departamento de Engenharia Elétrica, Universidade Federal do Espírito Santo, ES (e-mails: patrick.ciarelli@ufes.br; celso.munaro@ufes.br)*

Abstract: Produced water in offshore oil rigs is an effluent recovered from wells of oil and natural gas, and is the main waste generated in this process. The Total Oil and Grease (TOG) is considered one of the main parameters for controlling the disposal of water produced in sea, with daily and monthly limits defined by current legislation. The TOG measurements used as reference by IBAMA is made using the gravimetric method, with samples of produced water collected daily and send to a licensed laboratory, which provides the results about 20 days after the sampling date. The need for corrective actions in case of values above the limit has motivated the use of alternative methods that generate estimates more frequently. In this work, two models based on recurrent neural networks are proposed to obtain TOG estimates. Variables from the produced water treatment process, information about chemicals used, and data about daily production from an offshore platform were collected, processed, and used to train, validate, and test these models. The comparison between the estimation error of these models and those of the existing estimation methods shows the advantage of the proposed models.

Resumo: Água produzida, em plataformas marítimas, é um dos efluentes recuperados de poços em conjunto com petróleo e gás natural, sendo o principal resíduo gerado nesse processo. O Teor de Óleos e Graxas (TOG) é considerado um dos principais parâmetros de controle do descarte de água produzida no mar, com limites diários e mensais definidos pela legislação vigente. A medição de TOG usada como referência pelo IBAMA é feita pelo método gravimétrico, com amostras de água coletadas diariamente e enviadas para laboratório acreditado, que fornece os resultados após cerca de 20 dias a partir da data de amostragem. A necessidade de ações corretivas em caso de valores acima do limite tem motivado o uso de métodos alternativos que gerem estimativas com maior frequência. Neste trabalho, dois modelos baseados em redes neurais recorrentes são propostos para obter estimativas de TOG. Variáveis do processo de tratamento de água produzida, informações sobre produtos químicos utilizados e dados sobre produção diária de uma plataforma *offshore* foram coletados, tratados e utilizados para treinar, validar e testar esses modelos. A comparação entre o erro de estimativa desses modelos com os dos métodos existentes de estimação mostra a vantagem do uso dos modelos propostos.

Keywords: Total oil and grease; Environment monitoring; Process monitoring; Data Science; Digital Transformation; Recurrent neural networks

Palavras-chaves: Teor de óleos e graxas; Monitoramento ambiental; Monitoramento de processo; Ciência de Dados; Transformação Digital; Redes neurais recorrentes

1. INTRODUÇÃO

O petróleo é uma das maiores fontes da matriz energética mundial. Entretanto, a sua produção pode afetar o meio ambiente de diversas maneiras, sendo a água produzida (AP) o maior efluente desse processo (Amini et al., 2012). A AP é escoada até a superfície junto ao óleo e ao gás durante a produção desses fluidos e possui contaminantes como o próprio óleo, produtos químicos e gases dissolvidos (Bayati et al., 2011).

Uma das principais métricas de qualidade da AP descartada no mar é o Teor de Óleos e Graxas (TOG), por vezes chamado de *Total Oil and Grease* ou *Oil-in-Water*. Descartes de AP no mar com TOG acima do permitido nas licenças ambientais podem implicar em sanções pecuniárias impostas pelos órgãos reguladores. Além dos aspectos financeiros, podem incorrer prejuízos à imagem e ao meio ambiente (Gagnon, 2011; Meier et al., 2010).

No Brasil, a Resolução CONAMA 393/2007 (Brasil, 2007) estabelece que o método de referência é o gravimétrico (Rice et al., 2017). Em plataformas de produção *offshore*, o balanço inerente às embarcações dificulta a execução das análises por esse método (Yang, 2011). Assim, as amostras são enviadas a laboratórios *onshore* e, devido à logística, os resultados são disponibilizados com defasagem, o que impossibilita o controle da planta de processamento em função desses resultados.

Para permitir um melhor acompanhamento do processo, a água é analisada a bordo da unidade por meio de análises de laboratório e instrumentos *online*, estes últimos instalados no processo. Porém, vale frisar que o TOG é uma medida método-dependente, ou seja, os valores encontrados só têm significado quando é informado o método aplicado (Yang, 2011).

Os processos relacionados à produção de petróleo estão cada vez mais instrumentados, utilizando diversos sensores de pressão, temperatura, vazão, nível e analisadores que visam a monitorar e controlar a planta. Com isso, modelos de predição baseados em dados têm se tornado viáveis e são cada vez mais populares, demonstrando amplo potencial de aplicações (Kadlec et al., 2009).

As redes neurais recorrentes (RNNs - *recurrent neural networks*) possuem a capacidade de lidar com dados sequenciais, como linguagem natural, áudio e séries temporais (LeCun et al., 2015). As RNNs processam a entrada sequencial, um elemento por vez, mantendo em suas células informações sobre o histórico de todos os elementos anteriores. Um dos tipos de RNN, as LSTMs (*Long Short-Term Memory*), propostas inicialmente por Hochreiter e Schmidhuber (1997), mostraram-se mais efetivas que as RNNs convencionais (LeCun et al., 2015), uma vez que solucionam o problema do desaparecimento/explosão de gradientes (Géron, 2019).

Neste trabalho, propõe-se o desenvolvimento de modelos baseados em dados que auxiliem o monitoramento do TOG gravimétrico a fim de garantir que a água descartada no mar esteja dentro dos limites legais. O objetivo não é propor um novo método de referência, mas sim apresentar modelos que estimam o TOG com maior frequência e aderência.

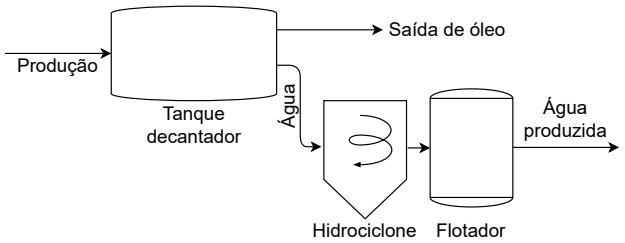

Figura 1. Diagrama simplificado de um típico sistema de tratamento de água produzida. Fonte: Oliveira Júnior e Pereira (2020).

O restante deste artigo está organizado da seguinte forma: a Seção 2 discorre sobre TOG em sistemas de produção de petróleo *offshore*; a Seção 3 apresenta a metodologia utilizada durante a coleta, o tratamento e a análise dos dados; os métodos propostos são detalhados na Seção 4; a Seção 5 apresenta os resultados obtidos neste trabalho e algumas discussões; a Seção 6 traz as conclusões e trabalhos futuros sugeridos.

2. TRATAMENTO DE ÁGUA PRODUZIDA EM INSTALAÇÕES OFFSHORE E MEDIÇÃO DE TOG

Um típico sistema de tratamento de água produzida em instalações *offshore* é apresentado na Figura 1. Nesse sistema, o fluido produzido (composto basicamente por óleo, água e gás residual), que já passou por processos de separação de gás, é enviado primeiramente para um tanque decantador (*settling tank*), onde ocorre a separação por gravidade. As principais influências nesta etapa são o nível da interface água/óleo, o tamanho das gotículas de óleo e a temperatura do fluido (Stewart e Arnold, 2011).

Posteriormente, a água é bombeada para hidrociclones, onde uma força centrífuga é aplicada para separar líquidos de diferentes densidades (Veil, 2011). O desempenho desse equipamento é influenciado sobretudo pela taxa de rejeito e pela vazão de entrada (Husveg et al., 2007). Além disso, destaca-se que incrustações podem ocorrer no interior das linhas, sendo necessárias manutenções periódicas e injeção de inibidor de incrustação, a fim de manter a eficiência do sistema, a qual também pode ser verificada por variáveis de processo obtidas em tempo real.

Por fim, a água contendo menos contaminantes chega a etapa de flotação, última parte do tratamento. Pequenas bolhas de gás são injetadas na água na entrada do vaso flotador. Ao subirem, tais bolhas associam-se a gotículas de óleo e partículas sólidas e, na superfície, esses elementos são removidos por um processo chamado *skimming* (Veil, 2011). Dessa forma, a eficiência do tratamento está vinculada ao tamanho das bolhas e à vazão de gás, bem como à vazão de entrada de água (Stewart e Arnold, 2011).

Ademais, o TOG é afetado diretamente por produtos químicos utilizados, tais como: inibidores de incrustação, desemulsificantes, antiespumantes e polieletrólitos (Al-Ghouti et al., 2019; Jiménez et al., 2018).

De acordo com Yang (2011), os métodos para medição de TOG podem ser separados em três grupos: referência, bancada (ou laboratório de bordo) e em linha (instrumento *online*). Os métodos de referência são utilizados por insti-

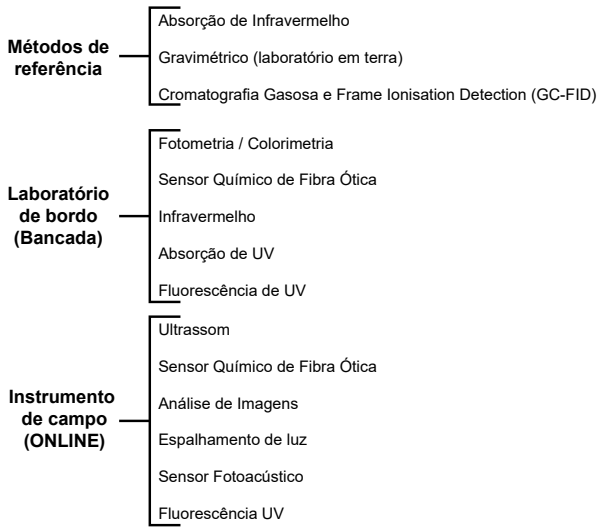

Figura 2. Métodos para determinação do Teor de Óleos e Graxas. Adaptado de: Oliveira Júnior e Pereira (2020).

tuições governamentais e estabelecem medidas de referência abrangentes e padronizadas. Já os métodos de bancada e *online* têm ganhado espaço por possuirem procedimentos menos complexos, menores custos e geração de estimativas com menor tempo (Cirne et al., 2016).

O CONAMA (Brasil, 2007) estabelece que as concentrações de TOG devem ser determinadas pelo método gravimétrico (Rice et al., 2017). Além disso, a média mensal de TOG na água produzida descartada é limitada à 29 mg/l, com valor máximo diário de 42 mg/l. Esse valor médio mensal é obtido a partir de amostras diárias e estas são compostas por subamostras coletadas ao longo do dia, tipicamente sendo usadas 4 subamostras. O resultado somente é conhecido posteriormente, o que inviabiliza o controle do processo a tempo.

A medição de TOG, como enfatizado por Yang (2011), é um parâmetro método-dependente, ou seja, uma mesma amostra avaliada por diferentes métodos gerará quase sempre diferentes resultados. O autor ressalta que a forma como é realizada a amostragem pode aumentar significativamente as incertezas das análises.

Com o intuito de evitar valores altos de TOG na água descartada, controlar e otimizar o processo, são realizadas análises por outros métodos nas próprias unidades de produção *offshore*, utilizando analisadores *online* e de bancada.

No estudo em questão, dados de três métodos de análise de TOG foram utilizados como entradas do modelo: fotométrico (Hach, 2021), infravermelho (Eralytics, 2021) e um analisador *online* (Advanced Sensors, 2021), que usa a fluorescência UV como princípio de medição. Em relação aos dois primeiros, por serem de bancada, foram coletadas amostras a cada duas horas e realizada a análise no próprio laboratório de bordo. Cabe ressaltar que o monitoramento de TOG *online* tem sido um desafio nas últimas décadas, visto que apresentam dificuldades associadas a diversos fatores (principalmente em aspectos de confiabilidade e continuidade operacional) e sensibilidade a temperatura do fluido, pressão, presença de gás e particulados, tipo de óleo, tamanho das gotículas, vazão de processo e do ponto de amostragem, procedimento de coleta de amostra, entre outros.

A importância da estimativa do TOG tem resultado em contribuições recentes ao tema na literatura. Roverso (2009) propôs a criação de sensores virtuais a partir do agrupamento de redes neurais em uma unidade de produção *offshore*. Usando como referência os valores diários (mistura das quatro coletas com intervalos de seis horas entre si), o autor descreveu as dificuldades associadas às diferentes taxas de amostragem das entradas e à definição de metodologia para sincronizar os dados de processo com as medidas laboratoriais. Por fim, optou pelo emprego de médias móveis em torno dos horários de coletas.

Em outra abordagem, Oliveira Júnior e Pereira (2020) investigaram a criação de modelos capazes de estimar em tempo real o TOG fotométrico e usaram como alvo as análises realizadas no laboratório de bordo. Os autores extraíram características em diferentes janelas de tempo (5, 20 e 60 minutos) antecedentes a cada coleta, avaliaram diferentes algoritmos para regressão (árvores de regressão, *bagged trees*, SVMs e redes neurais) e alcançaram os melhores resultados com *bagged trees*.

Araujo Filho et al. (2020) criaram modelos baseados em árvores de decisão e redes neurais com o objetivo de gerar previsões do TOG gravimétrico a cada cinco minutos. Considerando que a variável alvo é resultante de análises em terra e com frequência diária, a estratégia para obtenção de valores com período de cinco minutos foi a aplicação de uma técnica de interpolação. Os resultados auferidos para dados de duas plataformas indicaram a viabilidade e utilidade dos modelos baseados em dados.

Diante do exposto, este trabalho propõe o uso de modelos baseados em redes neurais recorrentes do tipo LSTM que sejam capazes de estimar o TOG gravimétrico em uma base diária. A escolha da LSTM se baseia na sua capacidade de lidar com sequências temporais (Kadlec et al., 2009), o que tem popularizado sua aplicação em processos industriais (Zhou et al., 2020; Lyu et al., 2020). Também foi avaliada a inclusão de uma camada convolucional (CNN - *convolutional neural networks*) anterior à LSTM. Essa estratégia, diferentemente da convencional, faz com que as primeiras camadas da arquitetura extraiam características de forma automática

3. COLETA, TRATAMENTO E ANÁLISE DOS DADOS

O conjunto de dados utilizado nos experimentos é proveniente de uma plataforma de produção *offshore* e refere-se à operação entre janeiro de 2018 e fevereiro de 2021. Nas subseções seguintes são fornecidos detalhes da base de dados usada, sendo importante destacar que cada conjunto de variáveis possui uma frequência de amostragem específica.

3.1 Medições de TOG

Neste trabalho, as análises de TOG são relativas sempre a coletas da água descartada, isto é, da água produzida após tratamento. Apesar de serem realizadas análises em

outros pontos da planta, definiu-se que essas não seriam incluídas, já que ocorrem de forma ocasional.

O TOG gravimétrico, variável de interesse, é formado por subamostras, coletadas durante o dia (no geral, em intervalos de seis horas entre si), misturadas e analisadas em laboratório em terra, gerando um único valor diário. O resultado é disponibilizado com certa defasagem, justificando assim a importância de se ter outras medidas ou estimativas do TOG para eventuais ações corretivas.

No laboratório de bordo, são obtidas medidas do TOG fotométrico e infravermelho (análises individuais para cada amostra) e os resultados são gerados em questão de minutos. Entre janeiro de 2018 e março de 2020, apenas a análise fotométrica estava disponível. A partir de março de 2020, o analisador com o princípio de medição por infravermelho começou a ser testado.

O analisador de TOG *online* funciona com a fluorescência UV e possui um sistema de autolimpeza na câmara de medição (Advanced Sensors, 2021). Ainda assim, é comum a necessidade de manutenções preventivas e corretivas com profissionais especializados. O sinal do instrumento é enviado em tempo real para o sistema de automação, que armazena o histórico em um PIMS (*Plant Information Management Systems*). Para a criação do banco de dados utilizado neste trabalho, optou-se por realizar a amostragem, a cada cinco minutos, desse sinal e das variáveis de processo.

3.2 Variáveis de Processo

A seleção inicial das variáveis de processo incluídas na base de dados foi realizada por um comitê multidisciplinar, incluindo especialistas em processamento de água, óleo e gás, Ciências de Dados e medições de fluidos. Além dos instrumentos relacionados à planta de tratamento da água produzida, foram coletadas medidas de variáveis relacionadas a outros trechos e equipamentos da planta de produção: chegada dos poços, separadores bifásicos, separador de teste, tratadores eletrostáticos de óleo e tanque de resíduos, totalizando 104 variáveis.

Assim como o TOG *online*, as demais variáveis de processo passam pelo sistema de automação e são armazenadas em um historiador, onde podem ser acessadas de forma segura. Apesar de todas as entradas serem atualizadas no historiador em questão de segundos, a frequência de amostragem utilizada foi de cinco minutos.

3.3 Boletins Diários de Operação

Os Boletins Diários de Operação (BDOs) contêm um resumo da operação da plataforma no dia. Eles possuem dados específicos de cada poço (produção líquida e bruta simuladas, BSW (*Basic Sediments and Water*) e vazão de água produzida por poço calculada com base nos valores anteriormente citados), além de informações totalizadas da unidade (produção bruta e líquida, água produzida e água descartada). Também estão inclusos dados sobre os produtos químicos injetados, a saber: desemulsificante (*topsides* e *subsea*), inibidor de incrustação e polieletrólitos. Os valores representam o volume adicionado ao processo durante o dia em questão.

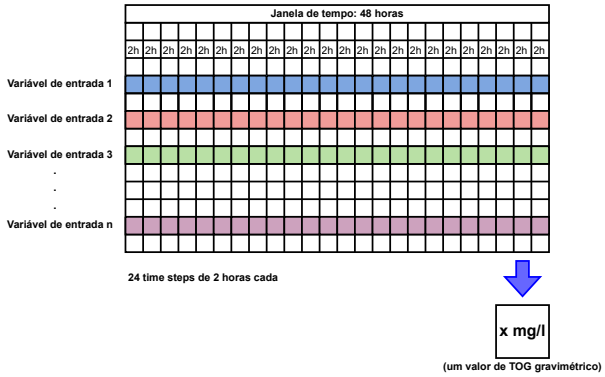

Figura 3. Exemplo de matriz de entrada contendo n atributos, divididos em 24 *time steps* de 2 horas cada, totalizando uma janela temporal de 48 horas, que se associa a um único valor de TOG gravimétrico.

3.4 Tratamento dos Dados

Nos dias em que não há descarte de água, não são coletadas amostras e, portanto, não há valores de TOG gravimétrico associados. Tais dias foram desconsiderados nas análises.

Nos momentos que o analisador de TOG *online* está em manutenção, sua saída fica constante em zero. Para estes casos, os valores zerados foram excluídos, transformando-os em valores faltantes.

4. MÉTODOS PROPOSTOS

Para treinamento dos modelos, a variável alvo escolhida foi o TOG gravimétrico. Já as entradas foram o TOG fotométrico, o TOG infravermelho, TOG online, as variáveis de processo, os dados de operação diários e os atributos calculados a partir das variáveis de processo.

Optou-se por padronizar em 2 horas a frequência de amostragem de todas as variáveis de entrada. Para as variáveis de processo, que são amostradas a cada 5 minutos, assim como para os atributos calculados a partir delas (descritos a seguir em "engenharia de atributos"), foi usada uma janela deslizante para o cálculo de suas médias a cada 2 horas. Já para os dados diários de produção, seus valores foram repetidos por 12 *time steps*, de forma a ter um valor a cada duas horas. Para cada valor de TOG gravimétrico, associaram-se os dados das 48 horas anteriores, totalizando assim 24 *time steps*, conforme exemplo apresentado na Figura 3. Cada amostra de TOG gravimétrico, que é disponibilizada diariamente, foi associada a uma matriz com tal forma.

Os experimentos para criação e avaliação dos modelos de previsão de TOG foram realizados como segue.

Divisão dos dados: a separação dos dados em treinamento, validação e teste foi realizada conforme a Figura 4, uma vez que os valores diários de TOG formam uma série temporal. Na representação gráfica, cada bloco corresponde a um conjunto de 32 dias, isto é, 32 amostras de TOG gravimétrico, e cada amostra é uma matriz de entrada (como a da Figura 3). É importante destacar que o banco de validação é sempre subsequente ao de treinamento, assim como o de teste é subsequente ao de validação, a fim que de que o algoritmo não seja treinado com dados futuros e testado

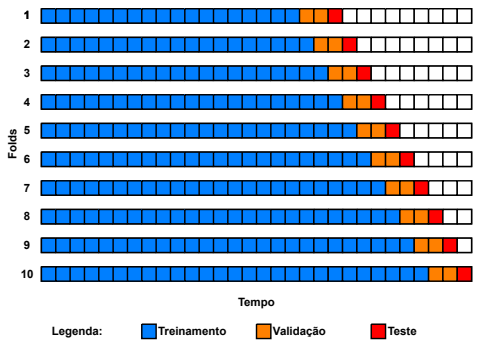

Figura 4. Separação dos dados entre treinamento, validação e teste para os 10 *folds*.

com dados passados. Além disso, o tamanho do banco de treinamento cresce ao passar para o *fold* seguinte.

Pré-processamento: os dados de entrada são padronizados com *z-score* (subtrai-se a média e divide-se pelo desvio-padrão de cada atributo). Ressalta-se que a média e o desvio-padrão são obtidos com o banco de treinamento. Na sequência, por escolha de projeto, os valores calculados são limitados a uma faixa entre -7 e 7 e, por último, os dados faltantes são substituídos pelo valor -10.

Engenharia de atributos: foram criados atributos a partir de variáveis do processo conforme detalhado a seguir.

a) Dois atributos que fazem inferência do nível da interface água/óleo, um para cada tanque (*settlings*), calculados como a média dos analisadores de BSW (cinco por tanque) em diferentes alturas dos tanques;
b) Um atributo relativo à identificação de qual tanque está em operação, que é selecionado como aquele com maior temperatura;
c) Um atributo referente à pressão de saída da bomba de água do tanque em operação. Selecionada a variável de pressão com o maior valor entre as quatro (são duas bombas por tanque);
d) Um atributo a respeito da temperatura do tanque em operação, utilizando o atributo do item b);
e) Um atributo que indica a pressão de gás do tanque em operação, também calculado com uso do atributo relatado no item b).

Seleção de grupos de variáveis de entrada: considerando-se que estão disponíveis 104 variáveis de processo, 8 variáveis de produção diária e 6 atributos calculados a partir das variáveis de processo, foram definidos grupos dessas variáveis para serem usados como entrada nos modelos. Além de reduzir o esforço computacional, tal análise visa verificar a possibilidade de reduzir o número de variáveis necessárias para estimar o TOG. Os agrupamentos foram realizados em função do equipamento que deu origem as variáveis do processo, aumentando gradativamente o número de variáveis e a complexidade do modelo. Ao todo, foram testados 9 grupos diferentes de entradas.

Modelos: foram propostos dois modelos baseados em LSTM (Figura 5), sendo que a principal diferença entre eles é a existência ou não de uma camada convolucional na entrada. O modelo com essa camada é referenciado neste documento por 'CNN + LSTM' e o outro por 'LSTM'.

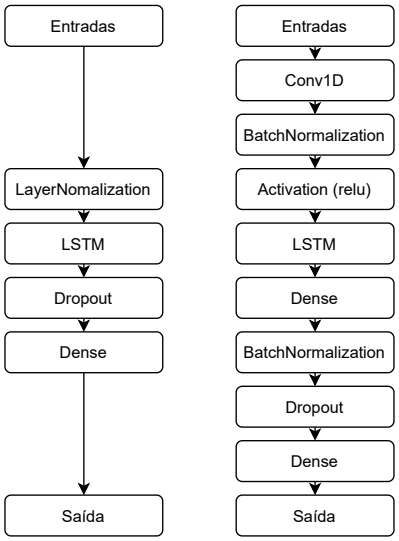

Figura 5. Modelos propostos: 'LSTM', à esquerda, e 'CNN + LSTM', à direita.

Uma breve explicação sobre as camadas utilizadas é fornecida a seguir:

- Conv1D: camada responsável pela convolução dos dados de entrada do modelo 'CNN + LSMT', buscando extrair características (Alzubaidi et al., 2021);
- *Layer Normalization* e *Batch Normalization*: normalizam os dados de entrada, tornando o treinamento das redes neurais mais estável (Alzubaidi et al., 2021);
- *Dropout*: auxilia na regularização da rede para evitar *overfitting*. O *dropout* consiste em desconectar aleatoriamente um percentual de neurônios durante o treinamento (Srivastava et al., 2014). Conforme recomendações fornecidas por Li et al. (2019), camadas de *Batch Normalization* e *Layer Normalization* foram colocadas somente antes de camadas de *dropout*, evitando assim a interferência entre elas;
- *Activation*: usada a função de ativação não-linear ReLU (*Rectified Linear Unit*);
- LSTM: camada referente à RNN em si, que recebe os dados de forma sequencial (24 *time steps*) e processa de forma a aprender os padrões nos sinais de entrada;
- *Dense*: por vezes chamada de *fully connected layer*, é uma camada de neurônios similar ao de uma rede neural clássica, que consegue se ajustar a mapeamentos não-lineares entre entrada e saída.

Para o treinamento, foi usada a função de perda MSE (*Mean Squared Error*) com um número máximo de épocas de 5000, regularização L1 e *max norm* nas camadas Conv1D, LSTM e *Dense*, com os valores de 10^{-2} e 5,0. Foi usado *early stopping* para interromper o treinamento caso o erro na validação não diminua após 500 épocas.

A busca de hiperparâmetros, por meio de *random search*, foi executada apenas para um grupo de entradas e replicada para as demais. Os melhores valores encontrados, assim como a faixa de busca, estão na Tabela 1.

Além disso, o cálculo do *loss* durante o treinamento é ponderado pela data da amostra: dados referentes a amostras mais recentes tem um peso maior, com este diminuindo exponencialmente quanto mais antiga a amostra. Dessa

forma, o modelo tende a privilegiar à condição atual dos equipamentos da planta de processamento, fator a ser considerado quando se trata de processos industriais em que o sistema está sujeito a mudanças ao longo do tempo.

Tabela 1. Hiperparâmetros utilizados para criação dos modelos 'LSTM' e 'CNN + LSTM'.

Hiperparâmetro	LSTM	CNN + LSTM	Faixa
Learning rate	10^{-3}	5×10^{-5}	$[10^{-4}, 10^{-1}]$
LSTM - unidades	16	16	[4, 128]
Dropout	0.25	0.4	[0, 0.6]
Dense - unidades	-	16	[4, 128]
Conv1D - filtros	-	16	[4, 32]
Conv1D - *kernel size*	-	3	[3, 11]
Conv1D - *strides*	-	2	[1,3]

Comparação dos modelos: os modelos propostos foram comparados a três outros: um preditor *baseline* ingênuo, que fornece como previsão a média dos valores de TOG gravimétrico do banco de treinamento (valor constante em todas as previsões do *fold*); outro que faz a previsão usando o valor médio do TOG fotométrico do dia e, por fim, um que faz a previsão usando o valor médio do TOG infravermelho no dia. Vale ressaltar que o TOG fotométrico e o infravermelho são utilizados pela equipe de operação para monitorar o processo de tratamento de AP, sendo essas as duas melhores estimativas do TOG gravimétrico viáveis atualmente.

Métricas: as métricas de avaliação utilizadas foram o RMSE (*Root Mean Square Error)* e a qualidade do modelo (fit). O RMSE de determinado modelo é calculado ao se comparar a saída gerada pelo modelo e o valor da variável alvo (TOG gravimétrico). Já a qualidade do modelo é obtida por meio da Equação (1):

$$fit = 100 \times \left(1 - \frac{\overline{e}_M}{\overline{e}_B}\right), \qquad (1)$$

em que $\overline{e}_M$ e $\overline{e}_B$ são as médias do RMSE nos 10 *folds* de teste do modelo avaliado e do preditor *baseline* ingênuo, respectivamente. Quanto maior o valor dessa métrica, melhor é o ajuste entre o valor de referência de TOG gravimétrico e o valor gerado pelo modelo em questão.

Com o propósito de monitorar e diagnosticar o treinamento do modelo, foram traçadas as curvas do RMSE durante o aprendizado para os bancos de treinamento e validação no decorrer das épocas (Figura 6). Tal gráfico proporciona noções sobre a presença ou não de problemas como subajuste (*underfitting*) e sobreajuste (*overfitting*) ao fim do procedimento de aprendizagem da rede neural.

Após a certificação de que o treinamento do modelo foi bem sucedido, criou-se um gráfico de tendência (Figura 7) que apresenta os valores do TOG gravimétrico e da saída do modelo no decorrer dos 32 dias do banco de dados de teste de um dos *folds*.

Para avaliar a dispersão dos erros e permitir a comparação das medianas, foram gerados *boxplots* (Figura 8) dos erros RMSE calculados para os 10 *folds* de teste para os dois modelos propostos e para os três preditores atuais (preditor *baseline* ingênuo, TOG fotométrico e TOG infravermelho).

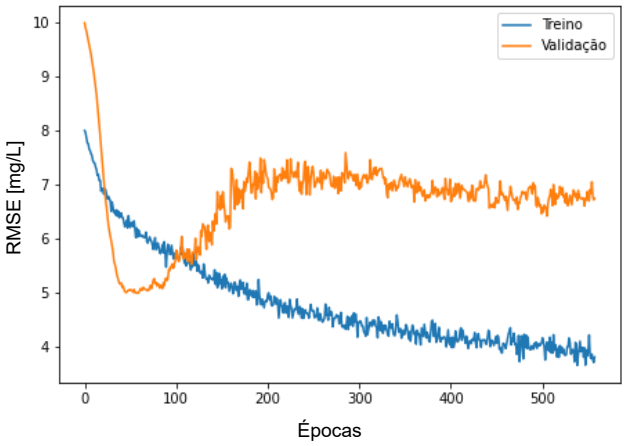

Figura 6. Exemplo de curvas de erro RMSE do modelo para os bancos de treinamento e validação ao longo das épocas de aprendizado.

5. RESULTADOS E DISCUSSÃO

São apresentados nesta seção apenas os resultados obtidos com o grupo de variáveis de entrada que proporcionou o treinamento de modelos com melhores desempenhos. Ele é composto por: os seis atributos criados referentes aos tanques *settlings* (dois referentes aos níveis das interfaces água/óleo, um relativo à identificação de qual tanque está em operação, temperatura, pressão da bomba de água e pressão de gás) e outras quatro relacionadas à injeção de produtos químicos (desemulsificante *topsides* e *subsea*, inibidor de incrustação e polieletrólito).

Foram gerados gráficos com curvas de aprendizagem para os dois modelos. Como exemplo, o processo de treinamento do modelo 'CNN + LSTM' em um dos *folds* é exibido na Figura 6. A curva em azul, referente ao RMSE dos dados de treinamento, tem a tendência de diminuir ao longo das épocas; já o RMSE no banco de validação, representado pela curva laranja, diminui nas primeiras épocas e, a partir de determinado momento, começa a apresentar uma tendência de alta. Ao atingir o valor mínimo de erro na validação, a execução prossegue até que se passe o número de épocas de espera configurado para o *early stopping*, tornando o modelo menos sujeito a sobreajuste (*overfitting*). O modelo final possui os pesos correspondentes à época na qual foi obtido o menor erro no banco de validação.

A Figura 7 mostra um exemplo de linhas de tendência geradas comparando os valores do TOG gravimétrico verdadeiro e a saída gerada pelo modelo 'CNN + LSTM' para os dados de teste de um dos *folds*. Percebe-se que o modelo conseguiu acompanhar a tendência geral da variável desejada, mesmo havendo erros maiores em alguns pontos.

Na Figura 8 são apresentados os conjuntos de erros (RMSE) obtidos pelos modelos nos 10 *folds* de teste. Os dois que foram propostos ('LSTM' e 'CNN + LSTM') apresentaram erros inferiores ao serem comparados aos demais (*baseline* ingênuo, TOG fotométrico e TOG infravermelho). Esses resultados indicam a viabilidade do uso dos modelos baseados em LSTM para o problema tratado neste trabalho.

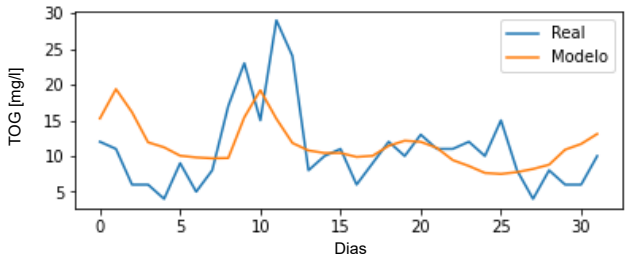

Figura 7. Linhas de tendência comparando o TOG gravimétrico real e a saída do modelo após treinamento para os dados de teste de um dos *folds*.

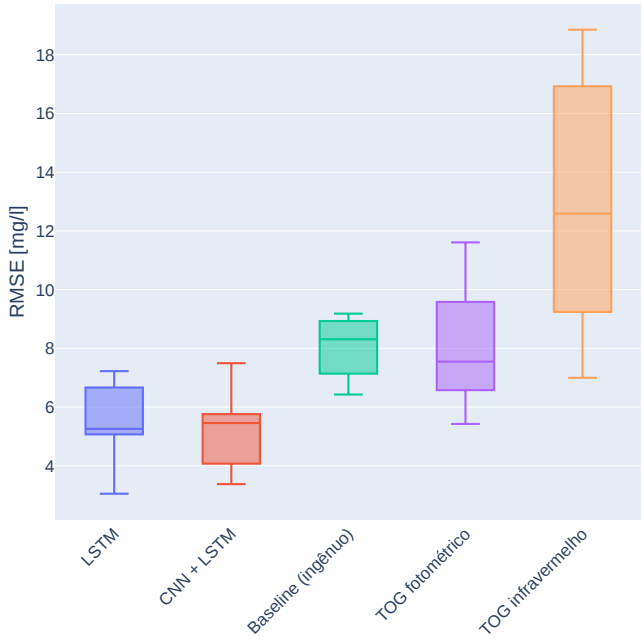

Figura 8. *Boxplot* apresentando os conjuntos do erro RMSE dos modelos em relação ao TOG gravimétrico do dados de teste para os 10 *folds*.

A Tabela 2 contém os valores do fit dos modelos avaliados nos dados de teste. Ressalta-se que tal métrica faz uma comparação entre a média dos erros de cada modelo com o erro do valor médio da variável alvo (Equação 1) e, portanto, o modelo *baseline* ingênuo obteve o valor nulo de fit. A estimativa por meio da média do TOG infravermelho obteve um percentual de fit negativo, devido a média do erro de tal modelo ser maior que a média do erro do *baseline* ingênuo (ver Figura 8). Os modelos propostos 'LSTM' e 'CNN + LSTM' alcançaram valores do fit acima de 30%. Tais resultados são coerentes com o que é observado na Figura 8, já que valores maiores de fit correspondem a melhores desempenhos.

Tabela 2. Comparação da qualidade do ajuste fit para diferentes modelos.

Modelo	fit [%]
LSTM	33,16
CNN + LSTM	36,17
Baseline (ingênuo)	0,00
TOG fotométrico	1,87
TOG infravermelho	-57,99

6. CONCLUSÃO

Neste trabalho, investigou-se o problema de estimação de TOG gravimétrico em água descartada por plataforma *offshore* de produção de petróleo, ressaltando a característica método-dependente do TOG e o tempo necessário para se obter os resultados pelo método de referência aceito pelo órgão fiscalizador devido ao envio das amostras para laboratórios *onshore*. Foram propostos dois modelos baseados em redes neurais recorrentes. O primeiro (LSTM) usa *Long Short-Term Memory* e o segundo (CNN+LSTM) adiciona camadas convolucionais ao primeiro.

Os modelos utilizaram como entradas as variáveis do processo de tratamento de água produzida, dos produtos químicos adicionados, dos dados diários de produção, dos atributos obtidos a partir das variáveis do processo e as medições de TOG obtidas a bordo por outros métodos. Os dados foram tratados e separados em conjuntos de treino, validação e teste. O valor diário do TOG foi estimado usando variáveis de entrada em um período de 48h anteriores à estimativa.

Os erros dos modelos propostos foram comparados com os erros referentes às estimativas obtidas pelos métodos fotométrico e infravermelho executados em laboratório de bordo. Os resultados obtidos neste trabalho indicam a viabilidade e a utilidade de modelos baseados em dados para monitorar o TOG.

Os resultados podem ser melhorados aos serem revisados aspectos como seleção de atributos, variação do tamanho da janela deslizante, revisão do tempo de amostragem, uso de novos algoritmos e criação de *ensembles*.

AGRADECIMENTOS

Agradecemos à Petróleo Brasileiro S.A. (Petrobras) pelo encorajamento, pelo fornecimento dos dados necessários para a realização deste trabalho e pela permissão para a sua publicação.

REFERÊNCIAS

Advanced Sensors (2021). Advanced Sensors EX-100/1000. URL `https://www.advancedsensors.co.uk/products/ex-100-1000`.

Al-Ghouti, M.A., Al-Kaabi, M.A., Ashfaq, M.Y., e Da'na, D.A. (2019). Produced water characteristics, treatment and reuse: A review. *Journal of Water Process Engineering*, 28, 222–239.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., e Farhan, L. (2021). Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53), 1–74.

Amini, S., Mowla, D., Golkar, M., e Esmaeilzadeh, F. (2012). Mathematical modelling of a hydrocyclone for the down-hole oil–water separation (dows). *Chemical Engineering Research and Design*, 90(12), 2186–2195.

Araujo Filho, C.F., Vargas, R.E.V., Bucker, E.B., e Delaiba, V.H.B. (2020). Monitoramento de teor de óleos e graxas em água descartada no mar usando ciência de dados. *Rio Oil & Gas Conference*.

Bayati, F., Shayegan, J., e Noorjahan, A. (2011). Treatment of oilfield produced water by dissolved air precipitation/solvent sublation. *Journal of Petroleum Science and Engineering*, 80(1), 26–31.

Brasil (2007). Resolução conama nº 393/2007. *Diário Oficial da União*, (153), 72–73. URL `http://www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=541`.

Cirne, I., Boaventura, J., Guedes, Y., e Lucas, E. (2016). Methods for determination of oil and grease contents in wastewater from the petroleum industry. *Chemistry and Chemical Technology*, 10(4), 437–444.

Eralytics (2021). Eracheck Eco. URL `https://eralytics.com/instruments/oil-in-water-testers/eracheck-eco`.

Gagnon, M.M. (2011). Evidence of exposure of fish to produced water at three offshore facilities, north west shelf, australia. In *Produced Water*, 295–309. Springer.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Hach (2021). DR6000 Laboratory Spectrophotometer. URL `https://www.hach.com/spectrophotometers/dr6000-laboratory-spectrophotometer/family?productCategoryId=35547203833`.

Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

Husveg, T., Rambeau, O., Drengstig, T., e Bilstad, T. (2007). Performance of a deoiling hydrocyclone during variable flow rates. *Minerals Engineering*, 20(4), 368–379.

Jiménez, S., Micó, M., Arnaldos, M., Medina, F., e Contreras, S. (2018). State of the art of produced water treatment. *Chemosphere*, 192, 186–208.

Kadlec, P., Gabrys, B., e Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & chemical engineering*, 33(4), 795–814.

LeCun, Y., Bengio, Y., e Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444.

Li, X., Chen, S., Hu, X., e Yang, J. (2019). Understanding the disharmony between dropout and batch normalization by variance shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2682–2690.

Lyu, P., Chen, N., Mao, S., e Li, M. (2020). LSTM based encoder-decoder for short-term predictions of gas concentration using multi-sensor fusion. *Process Safety and Environmental Protection*, 137, 93–105.

Meier, S., Morton, H., Nyhammer, G., Grosvik, B., Makhotin, V., Geffen, A., Boitsov, S., Kvestad, K., Bohne-Kjersem, A., Goksøyr, A., Folkvord, A., Klungsøyr, J., e Svardal, A. (2010). Development of atlantic cod (gadus morhua) exposed to produced water during early life stages effects on embryos, larvae, and juvenile fish. *Marine environmental research*, 70, 383–94. doi:10.1016/j.marenvres.2010.08.002.

Oliveira Júnior, J.M. e Pereira, M.d.A. (2020). Forecasting Total Oil and Grease in produced water using Machine Learning methods in an oil extraction plant. *Marine Systems & Ocean Technology*, 15, 124–134.

Rice, E., Baird, R., e Eaton, A. (2017). *Standard methods for the examination of water and wastewater*. Water Environment Federation, American Public Health Association , American Water Works Association, 23 edition.

Roverso, D. (2009). Empirical ensemble-based virtual sensing-a novel approach to oil-in-water monitoring. In *Oil-in-Water Monitoring Workshop, Aberdeen, UK*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., e Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Stewart, M. e Arnold, K. (2011). *Produced water treatment field manual*. Gulf Professional Publishing.

Veil, J.A. (2011). Produced water management options and technologies. In *Produced water*, 537–571. Springer.

Yang, M. (2011). Measurement of oil in produced water. In *Produced water*, 57–88. Springer.

Zhou, X., Hu, Y., Liang, W., Ma, J., e Jin, Q. (2020). Variational LSTM enhanced anomaly detection for industrial big data. *IEEE Transactions on Industrial Informatics*, 17(5), 3469–3477.