



BRUNA CUNHA

**IDENTIFICAÇÃO DE MODELOS NARX PARA POÇOS
SURGENTES DE PETRÓLEO USANDO PROGRAMAÇÃO
GENÉTICA MULTI-GÊNICA**

LAVRAS – MG

2023

BRUNA CUNHA

**IDENTIFICAÇÃO DE MODELOS NARX PARA POÇOS SURGENTES DE
PETRÓLEO USANDO PROGRAMAÇÃO GENÉTICA MULTI-GÊNICA**

Projeto apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

Prof. DSc. Bruno Henrique Groenner Barbosa
Orientador

LAVRAS – MG
2023

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Cunha, Bruna.

Identificação de modelos NARX para poços surgentes de
petróleo usando programação genética multi-gênica / Bruna Cunha.
- 2023.

66 p. : il.

Orientador(a): Bruno Henrique Groenner Barbosa.

Dissertação (mestrado acadêmico) - Universidade Federal de
Lavras, 2023.

Bibliografia.

1. Soft-sensor. 2. Modelos NARX. 3. Poços surgentes de
petróleo. I. Barbosa, Bruno Henrique Groenner. II. Título.


BRUNA CUNHA

**IDENTIFICAÇÃO DE MODELOS NARX PARA POÇOS SURGENTES DE
PETRÓLEO USANDO PROGRAMAÇÃO GENÉTICA MULTI-GÊNICA**

Projeto apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

APROVADA em 31 de Outubro de 2023.

| | |
|--------------------------------------|--------|
| DSc. Bruno Henrique Groenner Barbosa | UFLA |
| DSc. Daniel Augusto Pereira | UFLA |
| DSc. Danton Diego Ferreira | UFLA |
| DSc. Giovani Bernardes Vitor | UNIFEI |

Documento assinado digitalmente
 BRUNO HENRIQUE GROENNER BARBOSA
Data: 31/01/2024 11:26:02-0300
Verifique em <https://validar.iti.gov.br>

Prof. DSc. Bruno Henrique Groenner Barbosa
Orientador

**LAVRAS – MG
2023**

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que me apoiaram durante minha pesquisa. Em especial, aos meus pais que sempre me incentivaram e apoiaram na busca pelo conhecimento. Aos amigos próximos que me acompanharam durante esse percurso.

Agradeço ao professor Bruno pelo apoio e orientação durante a realização da pesquisa, e aos membros e suplentes da banca, Danton, Daniel e Giovani que gentilmente se propuseram a contribuir com o trabalho.

Meus agradecimentos ao Programa de Pós-Graduação em Engenharia de Sistemas e Automação da Universidade Federal de Lavras pela oportunidade, e à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à UFLA (Universidade Federal de Lavras) pelos recursos financeiros destinados e toda a infraestrutura necessária para execução do projeto.

RESUMO

Técnicas de inteligência artificial aplicadas em contextos reais como predição de variáveis de processo, detecção de anomalia ou mudança em linhas de produção podem implicar em redução de custo de manutenção, ações de prevenção de acidentes e falhas, auxílio em tomada de decisão, evitar perdas de produção e financeira e identificar pontos para melhoria do processo. Este trabalho apresenta uma proposta de aplicação de um *soft sensor*, utilizando um algoritmo evolutivo, para identificação de comportamento futuros de um sensor de temperatura durante a operação de poços surgentes de petróleo. Para isso, primeiramente, são realizadas diversas execuções do algoritmo de programação genética multigênica (MGGP) a fim de identificar quais valores de hiper-parâmetros contribuem para um melhor resultado. Após essa análise e determinação dos parâmetros para o MGGP, o *soft sensor*, baseado em modelo NARX polinomial, é implementado sobre uma base de dados real pública, denominada 3W, com medições de sensores presentes em poços surgentes de petróleo, entre os períodos de 2017 e 2018. O desempenho do *soft sensor* é avaliado por meio do MAPE (Mean Absolute Percentage Error) que fornece a informação de adequação do modelo por meio do erro relativo entre os dados reais e a saída do modelo expresso em percentual. É realizado o treinamento do modelo e validação dos resultados que se mostram satisfatórios sobre os dados de dois poços disponíveis, resultando em uma função capaz de representar a dinâmica de operação de cada um dos poços surgentes de petróleo em operação normal. Já sobre os períodos de medição que possuem a transição ao estado de anomalia, o modelo não é capaz de explicar tal comportamento e, à medida que evolui no período transitório, o erro apresenta uma tendência de aumento. Este resultado possibilita seu uso no contexto de detecção de anomalias em poços de petróleo, como um classificador *one class*.

Palavras-chave: *Soft-sensor*. Modelos NARX. Poços surgentes de petróleo. Programação genética.

ABSTRACT

Artificial intelligence techniques applied in real contexts such as prediction of process variables, detection of anomalies or changes in production lines can result in reduced maintenance costs, actions to prevent accidents and failures, decision-making support, avoid losses of production and financial aspects and identify points for process improvement. The text presents a proposal for the application of a soft sensor, using an evolutionary algorithm, to identify future behavior of a temperature sensor during the operation of oil wells. The approach involves multiple executions of the multi-gene genetic programming algorithm (MGGP) to identify hyperparameter values contributing to the best results. After analyzing and determining the parameters for MGGP, the soft sensor, based on a polynomial NARX model, is implemented on a real public dataset called 3W. This dataset contains sensor measurements from oil wells between the periods of 2017 and 2018. The performance of the soft sensor is evaluated using the Mean Absolute Percentage Error (MAPE), which indicates the model's adequacy through the relative error between real data and model output expressed as a percentage. The model is trained and validated on data from two available wells, showing satisfactory results in representing the normal operation dynamics of each oil well. However, during periods of measurement transitioning to an anomalous state, the model is unable to explain such behavior. As it evolves through the transitional period, the error tends to increase. This outcome suggests the model's potential use in the context of anomaly detection in oil wells, functioning as a one-class classifier.

Keywords: Soft-sensor. NARX models. Offshore oil wells. Genetic programming.

LISTA DE FIGURAS

| | |
|--|----|
| Figura 2.1 – Fluxo de fluídos do meio poroso à plataforma | 15 |
| Figura 2.2 – Esquema de um poço marítimo surgente de petróleo | 16 |
| Figura 2.3 – Exemplo simbólico de um modelo genético multigênico | 22 |
| Figura 3.1 – Localização dos sensores de fundo e na superfície | 28 |
| Figura 3.2 – Mapa de dispersão das instâncias reais históricas do 3W dataset | 29 |
| Figura 3.3 – Visualização das medições realizadas entre 26/06/2017 a 27/06/2017, operação normal do poço 1 | 31 |
| Figura 3.4 – Visualização das medições realizadas entre 04/04/2017 a 07/04/2017, operação normal do poço 2 | 32 |
| Figura 3.5 – Visualização das medições realizadas entre 26/02/2017 a 27/02/2017, operação anormal do tipo 7 do poço 1 | 33 |
| Figura 3.6 – Visualização das medições realizadas entre 24/01/2014 a 25/01/2014, operação anormal do tipo 1 do poço 1 | 34 |
| Figura 4.1 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de treinamento. Fonte: Própria. | 40 |
| Figura 4.2 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de validação. Fonte: Própria. | 41 |
| Figura 4.3 – Comparativo dos dados reais e estimados em predição de 100 passos à frente pelo modelo nos três períodos de validação. Fonte: Própria. | 43 |
| Figura 4.4 – Comparativo entre os erros percentuais das janelas em predição 100 passos à frente pelo modelo nos três períodos de validação. Fonte: Própria. | 44 |
| Figura 4.5 – Simulação livre dada pelo modelo de período 9 de validação com anomalia. Fonte: Própria. | 45 |
| Figura 4.6 – Simulação de predição de 100 passos à frente e erro percentual por janela dada pelo modelo de período 10 de validação com anomalia. Fonte: Própria. | 46 |
| Figura 4.7 – Simulação de predição de 200 passos à frente e erro percentual por janela dada pelo modelo de período 10 de validação com anomalia. Fonte: Própria. | 47 |
| Figura 4.8 – Simulação livre do modelo no período 10 de validação com anomalia. Fonte: Própria. | 47 |
| Figura 4.9 – Visualização das medições do período 3, operação normal do poço 1. Fonte: Própria. | 49 |

| | |
|--|----|
| Figura 4.10 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de treinamento. Fonte: Própria. | 51 |
| Figura 4.11 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de validação. Fonte: Própria. | 52 |
| Figura 4.12 – Comparativo dos dados reais e estimados em simulação de predição de 100 passos à frente pelo modelo de três períodos de validação. Fonte: Própria. . | 54 |
| Figura 4.13 – Comparativo os erros percentuais das janelas em simulação de predição 100 passos à frente pelo modelo de três períodos de validação. Fonte: Própria. . | 55 |
| Figura 4.14 – Simulação livre do modelo no período 10 de validação com anomalia. Fonte: Própria. | 56 |

LISTA DE TABELAS

| | |
|---|----|
| Tabela 3.1 – Informações das amostras do poço 1 | 29 |
| Tabela 3.2 – Informações das amostras do poço 2 | 30 |
| Tabela 3.3 – Separação de períodos de treinamento e validação | 30 |
| Tabela 4.1 – MAPE das variações dos parâmetros do MGGP aplicado ao poço 1 na pre- dição de temperatura de <i>choke</i> | 37 |
| Tabela 4.2 – Parâmetros do MGGP | 38 |
| Tabela 4.3 – Resultados do modelo NARX obtido para predição de T-JUS-CKP. | 39 |
| Tabela 4.4 – Resultado de simulações de predição de k passos à frente para os períodos normais | 42 |
| Tabela 4.5 – Resultados do modelos NARX obtido para predição de T-JUS-CKP. | 53 |

LISTA DE SIGLAS

ANM *Árvore de Natal Molhada*

BSW *Basic Sediment and Water*

CKP *Choke de Produção*

MAPE *Mean Absolute Percentage Error*

MGGP *Multi-Gene Genetic Programming*

PDG *Permanent Downhole Gauge*

P-MON-CKP Pressão do fluido montante à válvula CKP

P-TPT Pressão do fluido no TPT

T-JUS-CKP Temperatura do fluido jusante à válvula CKP

T-TPT Temperatura do fluido no TPT

SUMÁRIO

| | | |
|----------|--|-----------|
| 1 | INTRODUÇÃO | 11 |
| 1.1 | Objetivos | 13 |
| 1.2 | Organização do trabalho | 13 |
| 2 | REFERENCIAL TEÓRICO | 14 |
| 2.1 | Poços marítimos de produção | 14 |
| 2.2 | Tipos de anomalias em <i>offshores</i> | 16 |
| 2.3 | Reconhecimento de padrões | 18 |
| 2.3.1 | Sensores virtuais | 19 |
| 2.3.1.1 | <i>Nonlinear polynomial autoregressive models with exogenous inputs (NARX)</i> | 20 |
| 2.3.2 | <i>Multi-gene genetic programming (MGGP)</i> | 21 |
| 2.4 | Aprendizado de máquina no ramo petrolífero | 23 |
| 3 | MATERIAIS E MÉTODOS | 26 |
| 3.1 | Base de dados | 26 |
| 3.2 | Algoritmo MGGP | 34 |
| 3.3 | Medidas de Desempenho | 35 |
| 4 | RESULTADOS | 36 |
| 4.1 | <i>Soft sensor</i> | 36 |
| 4.1.1 | Poço 1 | 36 |
| 4.1.1.1 | Hiperparâmetros do MGGP | 36 |
| 4.1.1.2 | Desempenho do modelo | 38 |
| 4.1.1.3 | Anomalia 1 (Aumento Abrupto de BSW) | 43 |
| 4.1.1.4 | Anomalia 7 (Incrustação em CKP) | 45 |
| 4.1.1.5 | Representação do modelo | 47 |
| 4.1.2 | Poço 2 | 50 |
| 4.1.2.1 | Anomalia 6 (Restrição Rápida em CKP) | 55 |
| 4.1.2.2 | Representação do modelo | 56 |
| 4.2 | Detecção de anomalia | 57 |
| 5 | CONSIDERAÇÕES FINAIS E CONCLUSÃO | 59 |
| | REFERÊNCIAS | 61 |

1 INTRODUÇÃO

Há anos, monitorar processos é uma atividade feita em diversos setores da indústria e trata-se de uma etapa muito importante para a garantia de um funcionamento correto, seja para reduzir custos de manutenção, evitar perdas de produção, identificar melhorias no processo, possibilitar ações de prevenção de acidentes e falhas. Quando um processo envolve ambiente hostil, esse monitoramento passa a ser indispensável, como é o caso de poços de extração de petróleo.

A segurança operacional dos processos é importante e, portanto, deve ser o primeiro objetivo do controle do processo. Em 2007, estimava-se que ações inadequadas em situações anormais de processo chegaram a causar perdas anuais de US \$ 20 bilhões na indústria petroquímica dos EUA (JäMSä-JOUNELA, 2007). Em 2013, acidentes nas atividades *upstream*, ou seja, em aplicações que antecedem às atividades de refino dentro da cadeia produtiva da indústria de óleo e gás, registrados nos EUA e na Angola causaram perda aproximada de 140 milhões e 240 milhões de dólares, respectivamente (PORTELA, 2015). Evitar ocorrências de determinadas anomalias significa evitar perdas de produção durante dias ou até semanas. Em certos casos, um procedimento para correção e/ou manutenção de equipamentos pode exigir sonda marítima cujo custo diário em 2016 podia ultrapassar 500 mil dólares (ANDREOLLI, 2016).

Dessa forma, buscar por recursos embasados em dados tem um papel essencial no setor industrial, uma vez que possibilita o estudo de ambientes e processos específicos a cada cenário, tornando um aliado na mitigação de problemas. Técnicas de aprendizado de máquina vêm ganhando espaço em cenários de classificação de falhas e detecção de anomalias em poços de petróleo, modelagem de sensores virtuais que simulam o comportamento de sensores de alto custo localizados em ambiente hostil e detecção de falhas. O aprendizado de máquina é a abordagem mais explorada para a tomada de decisões.

Sensor virtual, ou *soft sensor*, vem sendo amplamente estudado e aplicado nos processos industriais ao longo dos últimos anos (FREITAS; BARBOSA; AGUIRRE, 2021) (de Moraes et al., 2019). Trata de um modelo de predição baseada em grandes quantidades de dados disponíveis nestes processos e que pode ser classificado como modelo caixa-branca, caixa-cinza ou caixa-preta. Essa classificação é baseada em dados e conhecimento especialista prévio de um sistema, em que o primeiro considera apenas o conhecimento das equações que regem a física do sistema, o segundo considera tanto dados dinâmicos de entrada e a saída do sistema,

quanto o conhecimento prévio sobre o sistema e o terceiro considera apenas os dados de entrada e saída do sistema. A modelagem caixa-preta é bastante utilizada neste contexto e totalmente orientada a dados, ou seja, fornece modelos empíricos baseados nos dados históricos coletados no processo industrial. Diversas técnicas de inferência estatística e técnicas de aprendizado de máquina têm sido empregadas em *soft sensors* orientados a dados, tendo como exemplos a máquina de vetor de suporte (SVM), rede neural artificial (ANN) e modelo polinomial não linear autorregressivo com entradas exógenas (NARX) (AGUIRRE et al., 2017) (CAPRIGLIONE et al., 2020).

NARX é um modelo recursivo de entrada-saída onde a saída atual é obtida por uma expansão funcional não linear de termos de entrada e saída defasados, com adição de ruído. Existem várias representações deste modelo como, por exemplo, polinomial e redes neurais. Nos modelos NARX polinomiais, uma ampla gama de comportamentos pode ser representada de forma concisa usando apenas alguns termos do vasto espaço de busca formado pelos regressores candidatos e geralmente um pequeno conjunto de dados é necessário para estimar um modelo, o que pode ser crucial em aplicações onde é desafiador adquirir uma grande quantidade de dados (JUNIOR; MARTINS; NEPOMUCENO, 2021). Além disso, um grande número de dinâmicas não lineares pode ser caracterizada usando o modelo NARX polinomial e, por isso, um *soft sensor* projetado com modelo NARX polinomial se torna uma combinação promissora nos cenários de processos industriais.

O reconhecimento de padrões é parte integrante da maioria dos sistemas de inteligência de máquina construídos para tomada de decisão e está no centro de várias áreas de aplicação, incluindo análise de imagem, reconhecimento de fala e áudio, biometria, bioinformática, mineração de dados, descoberta e recuperação de informações (THEODORIDIS; KOUTROUMBAS, 2009). Técnicas de reconhecimento de padrões se mostram oportunas no contexto de detecção de falha, uma vez que possibilitam tanto a classificação de falhas em categorias específicas quanto em classificação de classe única, apenas categorizando em comportamento normal ou anormal. Este trabalho utiliza técnicas de reconhecimento de padrões e identificação de sistemas para análise de dados reais obtidos em operação de poços de petróleo para elaboração de proposta para tomada de decisão orientada a dados, conforme detalhado na próxima subseção.

1.1 Objetivos

O objetivo deste trabalho é o desenvolvimento de um *soft sensor* baseado em modelos NARX polinomiais para caracterização da operação normal de extração de petróleo em poços surgentes *offshore* utilizando técnicas de reconhecimento de padrões e identificação de sistemas, a partir de uma base de dados pública e real, a base de dados 3W-dataset (VARGAS, 2019b). Dentre as contribuições principais pode-se destacar a análise do algoritmo *Multi-Gene Genetic Programming* na obtenção de modelos NARX para a temperatura de *choke* de produção e os efeitos do uso de abordagens de múltiplos passos à frente e também do emprego de diferentes atrasos nas variáveis de processo a fim de compor o *soft-sensor*, característica importante para representar longas dependências temporais entre as variáveis.

Os objetivos específicos consistem em:

1. Analisar o algoritmo MGGP (do inglês, *multi-gene genetic programming*) na obtenção de modelos NARX e efeitos de mudanças em parâmetros de entrada no modelo final;
2. Desenvolver um *soft sensor* usando modelo NARX polinomial capaz de prever o comportamento de uma variável dos poços surgentes de petróleo com base em outras variáveis disponíveis;
3. Analisar os resultados do modelo aplicado a dados com ponto de operações distintos, incluindo dados obtidos durante ocorrência de anomalia.

1.2 Organização do trabalho

Este documento está organizado da seguinte maneira. O Capítulo 2 apresenta a revisão bibliográfica a respeito de poços marítimos de petróleo e técnicas de reconhecimento de padrões, apresentando definição teórica e pesquisas realizadas na área. No Capítulo 3 tem-se a metodologia de realização da pesquisa. Os resultados e discussões se encontram no Capítulo 4. Por fim, no Capítulo 5 são apresentadas as considerações finais.

2 REFERENCIAL TEÓRICO

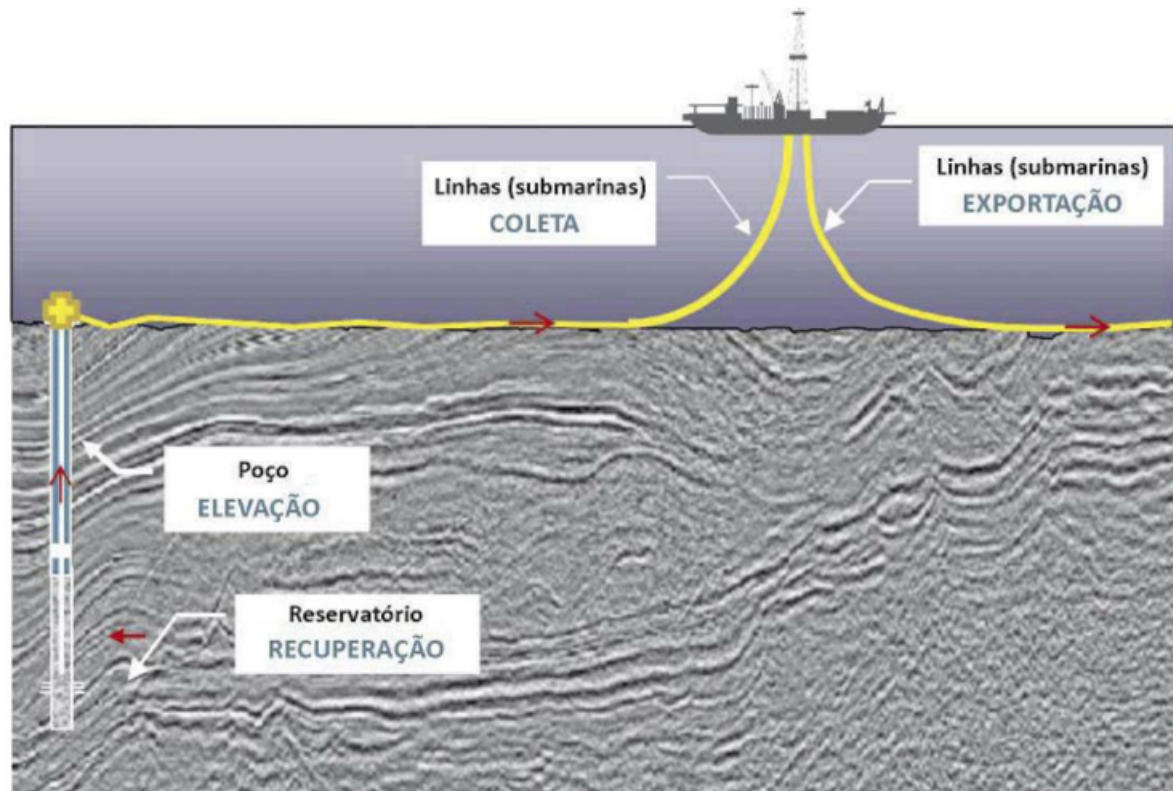
2.1 Poços marítimos de produção

As reservas mundiais mais significativas de petróleo encontram-se no mar (*offshore*) e a extração no meio marítimo envolve diversos processos controlados do escoamento como as etapas de recuperação no meio poroso, de elevação no poço, de coleta nas linhas de produção e de exportação. O processo de exploração se inicia com a prospecção do petróleo. A identificação de uma área favorável à acumulação de petróleo é realizada por meio de métodos geológicos e geofísicos que, atuando em conjunto, conseguem indicar o local mais propício para a perfuração (THOMAS, 2004). Em seguida, é feita a perfuração das rochas e criação dos poços por onde escoará o fluido composto por óleo, gás, água e eventualmente sedimentos (ANDREOLLI, 2016) ou injetará gases para elevação artificial.

A elevação do fluido pelos poços de produção depende da pressão que o fluido está submetido no meio poroso, podendo tratar-se de elevação natural (Poços surgentes) ou artificial (Poços não surgentes) que tem como principal característica a injeção de gás natural em alta pressão, a partir da superfície, na coluna com fluido de produção por meio de uma ou mais válvulas submersas fixadas em profundidades pré-determinadas (VARGAS, 2019b). Pode haver o caso de poços inicialmente surgentes utilizarem de recurso de elevação artificial uma vez que a extração continuada de petróleo produz uma queda de pressão nos fluidos do reservatório que a engenharia de produção tenta compensar por meio da injeção de água/vapor ou de gás (gás natural, CO₂). Sendo assim o monitoramento da movimentação desses fluidos dentro do reservatório é de grande importância para a otimização da produção (THOMAS, 2004). Elevação e escoamento eficientes significam maior vazão de óleo e gás, além de menor gasto de energia e recursos (ANDREOLLI, 2016). A Figura 2.1 ilustra de forma resumida o fluxo de escoamento em poços surgentes, onde é possível identificar a área de recuperação do fluido no reservatório que se eleva pelo poço, passando pelas linhas de coleta até finalizar o processo pela passagem nas linhas de exportação.

No sistema de elevação e escoamento, sensores auxiliam no monitoramento da operação, porém, por possuírem custo elevado e estarem localizados em ambientes hostis, são poucos. Um exemplo de sensor comum em poços de produção marítimo é o sensor de pressão no fundo do poço PDG (*Pressure Dawnhole Gauge*) que pode chegar a mais de quatro quilômetros de profundidade o que evidencia a sua manutenção não ser economicamente viável, envolvendo

Figura 2.1 – Fluxo de fluídos do meio poroso à plataforma

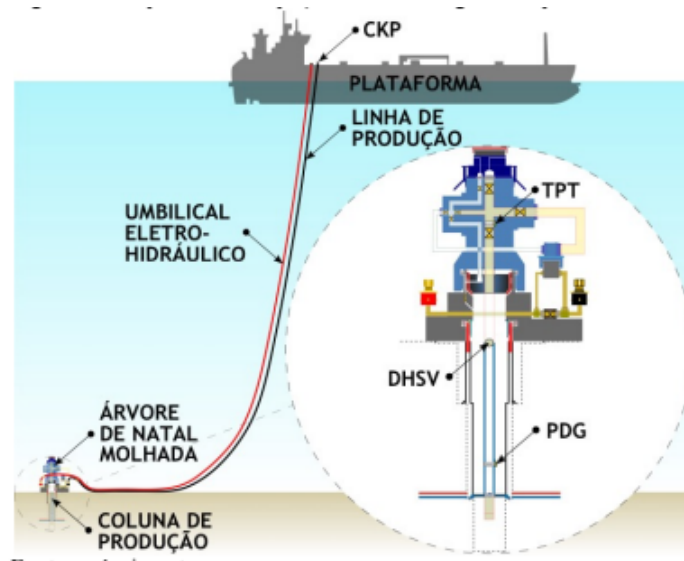


Fonte: Andreolli (2016)

ambiente de alto risco (FREITAS; BARBOSA; AGUIRRE, 2021). Além disso, são comuns sensores de temperatura e pressão do fluido localizados na árvore de natal molhada (ANM), um conjunto de válvulas no leito do oceano que permite a passagem dos fluidos para a plataforma e atua como barreira de segurança (NASCIMENTO, 2021). Já na plataforma, há sensores de pressão e temperatura do fluido montante e jusante à válvula de *Choke*, válvula que é responsável por regular a taxa de fluxo de produção. A Figura 2.2 apresenta alguns componentes comuns na operação de poços de elevação natural, sendo algumas variáveis de processo listadas a seguir:

- Pressão e temperatura do fluido no PDG (P-PDG): PDGs são sensores instalados nos poços para medir a pressão e temperatura de fundo de poço e desempenha um papel fundamental na obtenção e gestão das informações de caracterização do reservatório usando a distribuição de temperatura e pressão dos poços (HÜFFNER; TRIERWEILER; FARENZENA, 2019). Este sensor é muito útil na detecção de riscos e problemas de operação, análise de fluxo multifásico, ajuste de testes de produção, estratégia de controle, identifi-

Figura 2.2 – Esquema de um poço marítimo surgente de petróleo



Fonte: Vargas (2019b)

cação de modelo e, devido à sua localização em condições perigosas, pode ocorrer falha ou imprecisões na medição (OSSAI, 2020).

- Pressão e temperatura do fluido no TPT (P-TPT e T-TPT): Sensores localizados na árvore de natal molhada (ANM) que possuem boa confiabilidade, segundo os profissionais da área Elevação e Escoamento de Petróleo (VARGAS, 2019b).
- Pressão do fluido montante à válvula Choke de Produção (P-MON-CKP): Sensor considerado confiável quando presente (VARGAS, 2019b) e de fácil manutenção pela localização na plataforma.
- Temperatura do fluido jusante à válvula Choke de Produção (T-JUS-CKP).

2.2 Tipos de anomalias em *offshores*

Eventos indesejados podem acontecer nesses ambientes podendo ser identificados por valores de temperatura e/ou pressão, obtidos pelos sensores disponíveis durante operação, diferentes do esperado. Alguns tipos de anomalias que ocorrem prioritariamente em poços produtores marítimos operados via Elevação Natural (VARGAS, 2019b):

- **Aumento Abrupto de BSW:** O BSW, *Basic Sediments and Water*, trata-se da porcentagem de água e sedimentos em relação ao volume total do fluido medido (SAAVEDRA, 2016). Durante a vida útil de um poço, é esperado que o seu BSW aumente devido à maior

produção de água, proveniente seja do aquífero natural do reservatório ou da injeção artificial para evitar declínio da sua produção (VARGAS et al., 2019). Segundo Vargas (2019b), mudanças no BSW podem afetar diversas variáveis em um sistema de elevação e escoamento de petróleo, com efeitos positivos e/ou negativos para a produtividade do poço e que, na regra geral, as pressões medidas em pontos mais profundos aumentam e que as pressões nos pontos mais próximos à superfície diminuem, já as temperaturas costumam aumentar em todos os equipamentos.

- **Fechamento Espúrio de DHSV:** A válvula de segurança de fundo de poço (do inglês, *Downhole Safety Valve*) é um dispositivo amplamente utilizado em poços *offshore* para garantir a integridade e evitar vazamentos descontrolados de óleo e gás para o meio ambiente (COLOMBO et al., 2020), operando normalmente aberta e fechando em ocasiões de segurança operacional. Porém, esse fechamento pode vir a acontecer de forma espúria e sem sinalização na plataforma, tornando necessário uma identificação dessa falha e correção automaticamente. Essa identificação pode ser feita avaliando o aumento de pressão do PDG e redução de temperatura e pressão nos demais sensores, conforme mostrado em Vargas (2019b).
- **Intermitência Severa:** Trata-se de uma instabilidade crítica no fluxo e que, a depender da periodicidade e da intensidade, esse tipo de anomalia pode resultar em estresse ou mesmo danos em equipamentos do próprio poço e/ou da planta industrial (VARGAS et al., 2019).
- **Instabilidade no Fluxo:** Instabilidade do fluxo envolve alterações relevantes nas temperaturas e pressões, mas toleráveis, e que podem evoluir para uma intermitência severa, ressaltando a importância de um prognóstico que evite todos os aspectos negativos associados a essa anomalia mais grave (VARGAS, 2019b).
- **Perda Rápida de Produtividade:** A capacidade de fluido do poço é caracterizada pelo índice de produtividade, que relaciona a vazão, a pressão estática (ou média) do reservatório e a pressão do fluido no fundo do poço (THOMAS, 2004). Como o reservatório em produção está em constante modificação ao ser esvaziado, é esperado que essas propriedades se alterem com o decorrer do tempo (VARGAS, 2019b). Por isso, a identificação automática dessa condição em tempo hábil pode permitir que a equipe de operação mude

o ponto de operação do poço para que não perca sua produtividade (VARGAS et al., 2019).

- **Restrição Rápida em CKP:** A válvula de *Choke* faz o estrangulamento que provoca uma restrição na linha de fluxo acarretando queda de pressão ou redução da taxa de fluxo. Em geral, essa válvula é manual, e eventuais restrições rápidas indesejadas podem ocorrer devido a problemas operacionais (VARGAS et al., 2019).
- **Incrustação em CKP:** O problema da incrustação, que ocorre a partir do processo de aglomeração, consiste quando as partículas aglomeradas reduzem, total ou parcialmente, o espaço aberto ao fluxo, resultando em um aumento de perda de carga e em uma redução na produção petrolífera (CRUZ, 2018). A presença de incrustações está quase sempre associada à perda de produção, a situações de intervenção em poços e conseqüentemente à redução de lucros (VIANA et al., 2015).
- **Hidrato em Linha de Produção:** Os hidratos são formados durante o escoamento de gás ou óleo contendo obrigatoriamente gás e água, dentro dos tubos e para que ocorra a formação de hidratos é necessário que tenha alta pressão, baixas temperaturas, água e gás (VERÇOSA et al., 2018) chegando ao ponto de interromper totalmente a produção para remover o hidrato de tubulações gerando perda de produção (VERÇOSA et al., 2018) (VARGAS, 2019b) (ANDREOLLI, 2016).

Problemas que abordam o tema de detecção de falhas e anomalias em diversos contextos podem ser tratados com técnicas de reconhecimento de padrões. Por meio de técnicas computacionais de reconhecimento de padrões é possível encontrar uma maneira eficiente de, a partir das características extraídas de um conjunto de dados, organizar padrões, seja ela simbólica ou numérica, em categorias ou classes que compartilham determinadas semelhanças. Dessa forma, a seção 2.3 apresenta possibilidades para uso desta técnica no contexto de poços surgentes de petróleo.

2.3 Reconhecimento de padrões

O conceito de reconhecimento de padrões surgiu antes da década de 60 com pesquisas teóricas na área de estatística e tem como objetivo a classificação de objetos em várias categorias ou classes, podendo ser imagens, formas de onda de sinal ou qualquer tipo de medição que

precise ser classificada, a depender da aplicação (THEODORIDIS; KOUTROUMBAS, 2009). Esta técnica envolve as etapas de coleta e tratamento dos dados, a extração de características, seleção das variáveis mais relevantes, o treinamento para classificação dos conjuntos que compartilham as mesmas semelhanças e, por fim, o teste e a avaliação do modelo.

O uso do classificador do tipo classe única ou *one-class* abre possibilidade para solucionar problemas em que nem todas as categorias são conhecidas. Neste cenário, uma das classes é bem caracterizada por instâncias nos dados de treinamento, considerada a classe de operação normal de um sistema, enquanto a outra não possui instâncias ou possui em pouca quantidade, pois se trata de uma classe identificada como falha ou anomalia. O objetivo neste tipo de classificador pode ser detectar um comportamento anormal (ou novidade) e, por isso, é um bom candidato do tema em questão.

Já os modelos de regressão desempenham a tarefa de estimar ou prever um valor, fornecendo uma função que melhor representa o comportamento dos dados. A partir dessa função é possível visualizar a curva que melhor se aproxima das amostras reais. O *soft-sensor* trata-se de um modelo de regressão aplicado a casos reais, e é também conhecido por sensor virtual por sua capacidade de simular o comportamento de um sensor por meio da função resultante do treinamento dos dados. A partir desse modelo é possível realizar análises para se obter as características da operação de um sistema, bem como prever e estimar determinados comportamentos, como a ocorrência de falhas ou anomalias a partir do comportamento do modelo. Dado isso, os sensores virtuais podem ser utilizados no contexto de detecção de falhas e anomalias e serão apresentados neste capítulo.

2.3.1 Sensores virtuais

O *soft sensor*, ou sensor virtual, refere-se às abordagens e aos algoritmos usados para estimar e prever certas quantidades físicas ou qualidade do produto nos processos industriais com base nas medições e conhecimentos disponíveis (JIANG et al., 2021). Randek e Mandenius (2017) definem sensores virtuais como a combinação de dados analíticos de *hardware* (de sensores, dispositivos analíticos, instrumentos e atuadores) com modelos matemáticos que criam novas informações em tempo real sobre o processo. Lotufo e Garcia (2008) abordam sobre a possibilidade de se considerar o sensor virtual como o resultado da intersecção da tecnologia de sensores inteligentes e das técnicas de modelagem e identificação de sistemas e que o uso desse recurso para estimar uma variável da planta para a qual nenhum sensor é instalado fornece uma

oportunidade para melhorar o desempenho de uma planta uma vez que utiliza-se das medições de um sensor ou conjunto de sensores que se relacionam com a variável desejada e implementa um *software* capaz de fornecer, por meio de simulação, esta variável desejada. Por fim, Graziani e Xibilia (2020) também vão de acordo com a definição de modelos matemáticos usados para prever o comportamento de sistemas.

Para alcançar o modelo que melhor representa a dinâmica e as não-linearidades de um sistema real e definir um *soft-sensor* que seja capaz de simular o comportamento de um sensor real, existem diversas abordagens e técnicas disponíveis na literatura e comumente utilizadas para se explorar. Uma técnica comum para modelagem de *soft-sensor* é a identificação linear paramétrica e que pode ser encontrada em algumas aplicações com modelos de identificação do tipo entrada-saída tais como ARMA (Autoregressive-moving-average), ARX (Autoregressive with exogenous input), NARX (Nonlinear autoregressive with exogenous input), ARMAX (Autoregressive moving average with exogenous input) e NARMAX (Nonlinear autoregressive moving average with exogenous input) (CAPRIGLIONE et al., 2019) (CAPRIGLIONE et al., 2020) (AGUIRRE et al., 2017), (FONTES; VASCONCELOS; BRITO, 2020).

2.3.1.1 *Nonlinear polynomial autoregressive models with exogenous inputs* (NARX)

Modelos NARX (autoregressivos polinomiais não lineares com entradas exógenas) são amplamente utilizados no contexto de identificação de sistemas por sua flexibilidade e capacidade representativa. Modelo NARMAX é uma extensão de modelos NARX, onde termos residuais são inseridos na função para remover o viés dos parâmetros (CASTRO, 2021). O modelo NARX pode ser representado pela seguinte equação:

$$y(k) = f(y(k-1), \dots, y(k-n_y), x(k-1), \dots, u(k-n_u), e(k)), \quad (2.1)$$

em que $f(\cdot)$ representa uma função genérica não linear, $u(k)$, $y(k)$ e $e(k)$ representam, respectivamente, os sinais de entrada, saída e erro/incerteza/ruído no instante k , e n_u e n_y representam o atraso máximo para os sinais de entrada e saída.

Modelos NARX podem ser unidos a outras técnicas a fim de se alcançar a melhor representação de um sistema. Como exemplo, Coelho, Bora e Klein (2014) e Castro (2021) abordam o problema de seleção de estrutura de modelos NARX e NARMAX polinomiais por meio da programação genética.

Já Taqvi et al. (2020) fazem uso de rede neural NARX no contexto de predição de falhas em sistema dinâmico não linear. Redes neurais artificiais também possuem capacidade de aprendizado e adaptação através de uma arquitetura baseada na aquisição de conhecimento e habilidades organizacionais do cérebro humano. A aplicação de redes neurais na elaboração de um sensor virtual pode ser vista em Chitra e Pappa (2019), S., Mohanta e Pani (2022), Shang et al. (2014), Fan et al. (2020).

Os algoritmos genéticos, que simulam a evolução das espécies, inspirados no princípio de Darwin, podem ser utilizados na concepção dos sensores virtuais bem como a programação genética, uma técnica de aprendizado de máquina que também apresenta um processo análogo à evolução dos seres vivos envolvendo etapas de avaliação, seleção, cruzamento e mutação (KANEKO; FUNATSU, 2012) (SHARMA; TAMBE, 2014) (HUANG; LI; CAO, 2020), (KANEKO; FUNATSU, 2016), enquanto em (NASCIMENTO et al., 2021) utiliza-se no cenário de detecção de anomalias em poços de petróleo. Além disso, outras técnicas para aplicação no contexto de soft-sensors podem ser exploradas, tais como lógica fuzzy, também conhecida por lógica nebulosa (MENDES et al., 2012)(FONTES; VASCONCELOS; BRITO, 2020), e métodos estatísticos multidimensionais.

2.3.2 *Multi-gene genetic programming (MGGP)*

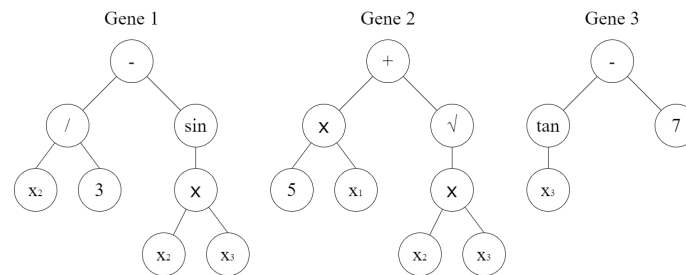
Os algoritmos evolucionários são baseados em processos naturais, onde são construídos modelos computacionais capazes de resolver problemas. Existe uma grande variedade de algoritmos que simulam a evolução das espécies por meio dos fenômenos naturais de seleção, mutação e reprodução. A programação genética trata-se de um algoritmo evolucionário onde seus indivíduos são programas de computador na forma de estruturas de árvore contendo funções e terminais, que evoluem aleatoriamente para novos programas.

O algoritmo de programação genética (GP) cria uma população inicial aleatória composta por estes indivíduos. A cada indivíduo é atribuído um valor numérico chamado aptidão, que indica o quanto a solução representada por este indivíduo é melhor em relação às outras soluções da população. A população de soluções candidatas é modificada iterativamente, cada iteração envolve a aplicação de operadores genéticos (seleção, cruzamento e mutação) na expectativa de gerar novos e melhores candidatos. O processo continua até que um dos critérios de parada é satisfeito. O *Multi-gene genetic programming (MGGP)*, ou programação genética multigênica, é uma variação da programação genética clássica, na qual, os indivíduos são estru-

turas de dados em forma de árvore. Dessa forma, o MGGP resulta em uma combinação linear ponderada das saídas de uma série de árvores GP, para as quais cada árvore pode ser considerada um gene (ÖZMEN et al., 2019). Como exemplo, a Figura 2.3 apresenta um modelo, em que x_1 , x_2 e x_3 representam os valores das entradas, y representa o valor de saída, w_0 é o termo *bias* (*offset*), e w_1 , w_2 e w_3 são os pesos dos genes, constituindo o seguinte modelo:

$$y = w_0 + w_1(x_2/3 - \sin(x_2x_3)) + w_2(5x_1 + \sqrt{x_2x_3}) + w_3(\tan(x_3) - 7)$$

Figura 2.3 – Exemplo simbólico de um modelo genético multigênico



O MGGP teve sua primeira aparição em 1996 por Hinchliffe (HINCHLIFFE et al., 1996). Segundo Hinchliffe, o MGGP possibilita variar as funções e base durante o processo evolutivo, isso significa que a medida que os cruzamentos e mutações vão acontecendo as funções irão se alterar. A principal vantagem do MGGP sobre a programação genética clássica é esta estrutura de seus indivíduos multigenes que apresentam dados em forma de árvore, compostas por nós e folhas, ou seja, cada gene é uma árvore de programação genética tradicional possibilitando a obtenção de uma equação que melhor represente um sistema.

Em Maynard et al. (2015), o MGGP é aplicado em um problema de estimação do perfil de poço sônico em poços de petróleo, utilizando dados de perfis geológicos de poço, onde apresentou um desempenho satisfatório e um erro inferior comparado à uma rede neural. É mencionado que uma das vantagens do MGGP é, caso se tenha uma ideia de comportamento da resposta do modelo, é possível a utilização de operadores específico (por exemplo, exponencial, tangente, logaritmo) para orientar a construção das soluções candidatas.

Outro exemplo de aplicação desse modelo pode ser visto em Niazkar e Niazkar (2020) cujo objetivo foi alcançar um modelo capaz de prever surtos de COVID-19 em alguns países. Uma vantagem mencionada nesse trabalho em relação a outros modelos de regressão não linear é que a estrutura de um modelo de previsão não precisa ser assumida antecipadamente, ou seja, permite desenvolver um modelo de previsão sem limitação de forma, enquanto o usuário pode decidir o compromisso entre a precisão e a complexidade do modelo de previsão, controlando o máximo de genes permitidos em cada indivíduo e a profundidade das árvores. Em Mota et

al. (2020), o MGGP, em uma versão multi-objetivo, é utilizado na identificação de um sistema de bombeamento de água onde a complexidade do modelo é levada em consideração durante o processo de identificação. Uma das características importantes elencadas é a flexibilidade do algoritmo em encontrar regressores com diferentes atrasos, possibilitando o tratamento de um grande espaço de busca.

2.4 Aprendizado de máquina no ramo petrolífero

O aprendizado de máquina (também conhecido por *Machine Learning*) ganhou destaque considerável nas últimas décadas. O aprendizado de máquina em um sentido mais amplo pode ser definido como aplicações de técnicas computacionais, como lógica *fuzzy*, redes neurais, algoritmos genéticos e agentes adaptativos para descoberta de conhecimento e engenharia (RECKNAGEL, 2001). Na inteligência artificial (IA), o aprendizado de máquina surgiu como o método preferido para o desenvolvimento de *software* prático para visão computacional, reconhecimento de fala, processamento de linguagem natural, controle de robô e outras aplicações (JORDAN; MITCHELL, 2015). Diversas pesquisas e aplicações envolvendo técnicas de inteligência computacional e artificial tem ganhado espaço no ramo petrolífero.

A pressão do fundo do poço é uma variável importante quando trata-se do método de elevação artificial com injeção de gás (*Gas-lift*, em inglês). O problema de previsão de falha de levantamento artificial apresenta desafios interessantes para algoritmos de mineração de dados, por causa dos muitos problemas de dados do mundo real, como ruído, dados ausentes, atraso de registros de eventos de falha e grande variabilidade entre unidades de levantamento artificial de poço funcionando normalmente (LIU et al., 2011).

Em de Moraes et al. (2019) são propostos *soft sensor* para predição da pressão no fundo do poço (PDG) fazendo uso de algoritmo evolucionário com diferenciação numérica (EAND) encarregado por encontrar os melhores atrasos nas variáveis de entrada do poço para uma boa predição. Nesse trabalho, foram estudados cinco poços *offshore* diferentes e o resultado do EAND foi aplicado às entradas de três modelos preditores não lineares: *Feed-forward Multilayer Perceptron Artificial (MPL) Neural Network*, *Radial Basis Function Network* e *case of Extreme Learning Machines (ELM)* e *Random Forest (RF)*, obtendo resultados satisfatórios o que possibilita a utilização do EAND em outros problemas de otimização.

Aguirre et al. (2017) também abordam o desenvolvimento de *soft sensor* para o PDG, citando a remoção de *outliers* e valores faltantes, a clusterização dos dados (Algoritmo *k-means*

com distância Euclidiana) para eliminação de amostras inadequadas, identificação de sistemas com modelos NARMAX neurais e polinomiais, técnica de filtragem (Filtro de Kalman), dentre outras tratativas.

Liu et al. (2011) também abordam o assunto de sensores no processo de *gas-lift*. Os autores utilizaram técnicas de *cross-correlation* e *autocorrelation* para entender melhor sobre a interdependência ao longo do tempo e a dependência do valor atual de um atributo em seus valores anteriores, respectivamente, com o objetivo de preparar os dados na predição de falhas no processo de *gas-lift*. A detecção de falha também pode envolver outros passos para a preparação dos dados como citado em Russell, Chiang e Braatz (2000), que avalia o uso de análise de variável canônica e análise dinâmica de componentes principais em processos industriais.

Marins et al. (2021) tratam da detecção automática e classificação de eventos de falha durante a operação prática de poços e linhas de petróleo e gás em poços surgentes. Aplicando o classificador *Random Forest* e com uma estratégia de otimização não convexa Bayesiana para ajustar os hiperparâmetros do sistema alcançaram uma acurácia de 94% de sucesso para detecção de falhas reais. Em Vargas (2019b) utiliza-se a mesma base de dados mostrando o comparativo entre os algoritmos de classificação *One Class SVM*, com diferentes funções de otimização (*RBF*, *Poly*, *Sigmoid*, *Linear*) e o algoritmo *Isolation Forest*, o qual apresentou melhor resultado. Em Nascimento (2021) foi empregado *Autoencoders* empilhados para redução de dimensionalidade dos dados de entrada dos modelos de classificação do trabalho de Vargas (2019b) comprovando que a utilização deste tipo de rede neural artificial (RNA) implica em uma melhor extração de características relevantes para este mesmo conjunto de dados, na redução do custo computacional dos classificadores e ainda apresentando melhor resultado comparado a técnica de análise de componentes principais (PCA).

O trabalho de Barbariol, Feltresi e Susto (2020) se baseia em algoritmos de aprendizado de máquina não supervisionados que permite ao sistema de metrologia detectar *outliers* e fornecer um nível estatístico de confiança nas medidas para fluxos de dados temporal multivariados podendo ser aplicado, por exemplo, na estimativa de quantidades que são difíceis de serem medidas, como o conteúdo de água ou gás de um fluxo multifásico, vindo de um poço de petróleo.

Aljubran et al. (2021) trabalham com algoritmo supervisionado para detecção precoce de incidentes de vazamento de fluídos durante as operações de perfuração de poços o qual mostrou que modelos de rede neural convolucional unidimensional (Modelo CNN) alcançou

bons resultados. Já Simoes et al. (2000) usam rede neural baseada em quantização de vetores (Modelo CPN) lineares para classificação a fim de estimar sinais de ondas acústicas em um sistema de extração de oleoduto que possibilitam o monitoramento de pressão e temperatura do poço.

É importante observar que os trabalhos citados envolvem dados reais de produção com alto volume de amostras e variáveis coletadas que podem ou não interferir na análise. Isso quer dizer que é preciso conhecer técnicas de pré-processamento, modelos de treinamento e análise de resultado para ser capaz de produzir um trabalho relevante. Aprendizado de máquina aplicado na detecção e diagnóstico de falhas em outras áreas pode envolver diversas outras técnicas e possibilidades de tratamento e análise dos dados que podem ser encontradas em Azhdari e Mehranbod (2010), Henmi et al. (2016), Chen, Liu e Zhou (2019), He et al. (2015), Li e Hu (2018), Liu et al. (2021), Li, Hu e Spanos (2016), Dhimish et al. (2018), Du et al. (2014), Nybø (2010), Gao, Cecati e Ding (2015).

3 MATERIAIS E MÉTODOS

Neste capítulo é apresentada a base de dados utilizada neste trabalho, suas características, como foram coletadas e algumas limitações. Também são apresentados o algoritmo a ser utilizado na implementação do *soft sensor*, a definição de seus parâmetros e as métricas de desempenho a serem utilizadas para validação do *soft sensor*.

O trabalho em questão é inspirado no desafio proposto por Vargas (2019b), o qual objetiva a classificação de eventos normais e anômalos na operação de poços surgentes de petróleo. Algumas regras do *benchmark* devem ser observadas, porém, apenas algumas serão aplicadas neste trabalho uma vez que a classificação do tipo *one-class* não será o foco principal:

- Apenas instâncias reais com anomalias de tipos que têm períodos normais maiores ou iguais a vinte minutos devem ser utilizadas. Entende-se por instâncias um conjunto de dados separados por poço e tipo (anomalia ou normal).
- As instâncias recomendadas estão relacionadas com as anomalias: Aumento Abrupto de BSW, Fechamento Espúrio de DHSV, Perda Rápida de Produtividade, Restrição Rápida em CKP, Incrustação em CKP e Hidrato em Linha de Produção, que serão apresentadas adiante, neste capítulo.
- Múltiplas rodadas de treinamento e validação devem ser realizadas.

3.1 Base de dados

Os dados reais presentes na base de dados foram coletados durante a operação de vinte e um diferentes poços de elevação natural. Os processos envolvidos nesta operação são monitorados com auxílio de diversos sensores localizados em regiões distintas do fluxo e que servem para mapear o comportamento de cada poço. Este banco de dados público é conhecido por 3W-dataset (VARGAS, 2019a) e possui observações de 5 variáveis de processo comuns em poços de elevação natural, sendo elas: medições de pressão do fluido no PDG (*Permanent Downhole Gauge*), pressão e temperatura do fluido no TPT (*Temperature and Pressure Transducer*), pressão do fluido montante à válvula de *Choke* de produção e temperatura do fluido jusante à válvula de *Choke* de produção. Além dessas variáveis, também são informadas a data do evento e a classe a qual ele se enquadra dentre os oito tipos de anomalias já identificadas ou evento normal. Esta base possui dados reais, simulados e desenhados, além de eventos normais

ou com alguma anomalia (VARGAS, 2019b) (MARINS et al., 2021) (NASCIMENTO, 2021). Para o trabalho em questão, apenas dados reais são utilizados.

Na Figura 3.1, pode ser visto o posicionamento de cada sensor na linha de produção que medem: pressão do fluido montante à válvula de *choke* (P-MON-CKP), temperatura do fluido jusante à válvula de *choke* (T-JUS-CKP), temperatura e pressão do fluido na árvore de natal molhada (T-TPT e P-TPT, respectivamente) e pressão do fluido no fundo do poço (P-PDG). Além desses, também existem sensores para medição de pressão do fluido jusante à válvula de controle de gás *lift* (P-JUS-CKGL), temperatura do fluido jusante à válvula de controle de gás *lift* (T-JUS-CKGL) e vazão de gás *lift* (Vazão GQL), e que não são considerados para este trabalho uma vez que são sensores pertencentes ao processo de elevação artificial com injeção de gás *lift*.

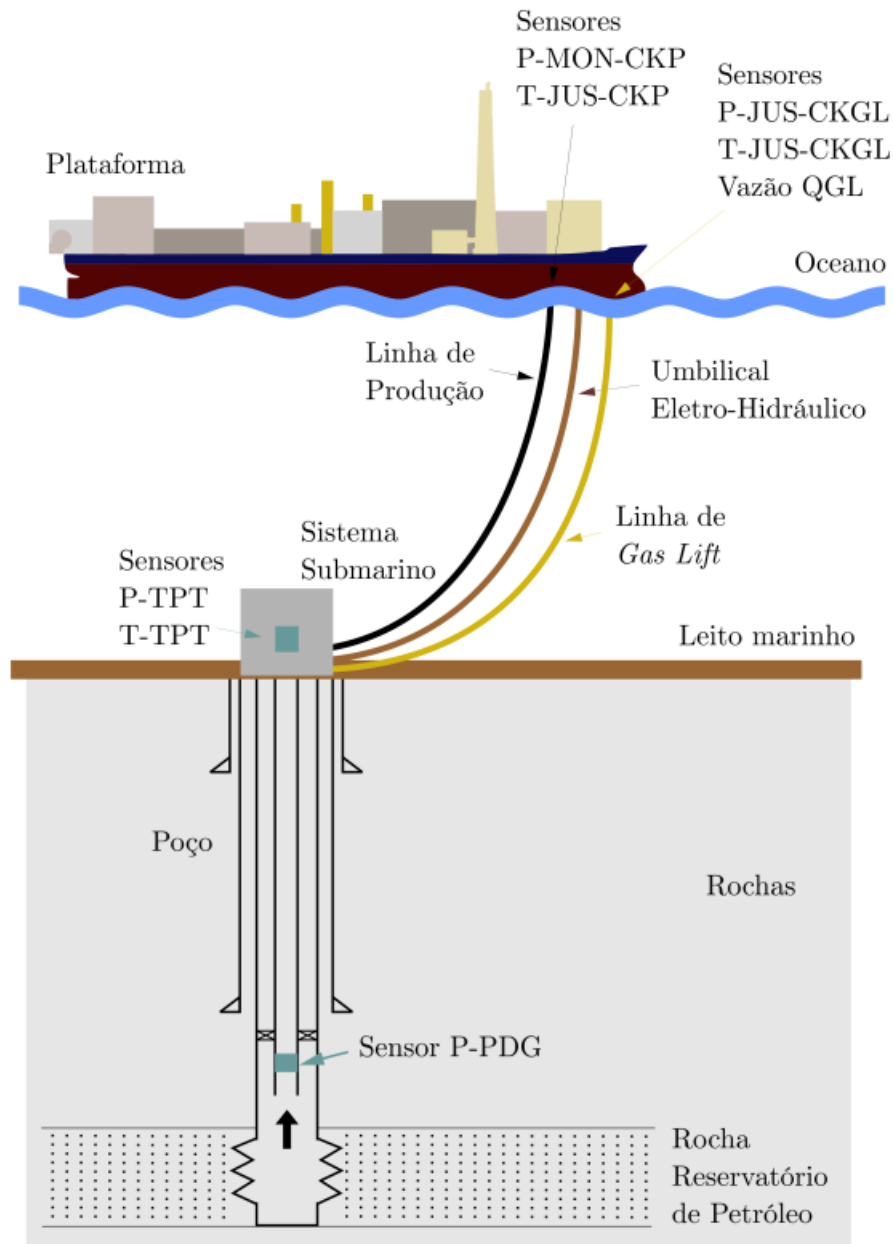
Cada poço possui características específicas, o que dificulta a utilização dos dados para análise e modelagem de mais de um poço em conjunto. Outra limitação encontrada nesta base está relacionada a dados faltantes ou constantes uma vez que algum sensor pode não estar presente na configuração do poço e/ou apresentar falhas. As medições reais foram obtidas durante o período do ano de 2012 a meados de 2018, como pode ser visto na Figura 3.2 que fornece o mapa de dispersão das instâncias reais históricas dos dados disponíveis.

Para este trabalho, restringiu-se o uso dos poços com base na avaliação dos períodos de medições e na ocorrência de anomalias. Os seguintes critérios foram utilizados para seleção dos poços 1 (WELL-00001) e 2 (WELL-00002):

- Baixo percentual de dados faltantes;
- Presença de períodos com medições classificadas unicamente como normais;
- Presença de períodos com alguma anomalia proposta no *benchmark*;
- Períodos separados por poços com médias de valores semelhantes;
- Medições das mesmas variáveis de processo entre os poços;

Dado tais critérios, o pré-processamento na base de dados inclui a remoção de variáveis com percentuais de medições faltantes superior a 10%, a reamostragem, originalmente em segundos, o preenchimento de dados faltantes com replicação do valor anterior medido e a separação da base por períodos de medição. Considerou-se as variáveis P-TPT, T-TPT, P-MON-CKP e T-JUS-CKP, presentes em todos os dados coletados, e as anomalias 1 (Aumento

Figura 3.1 – Localização dos sensores de fundo e na superfície

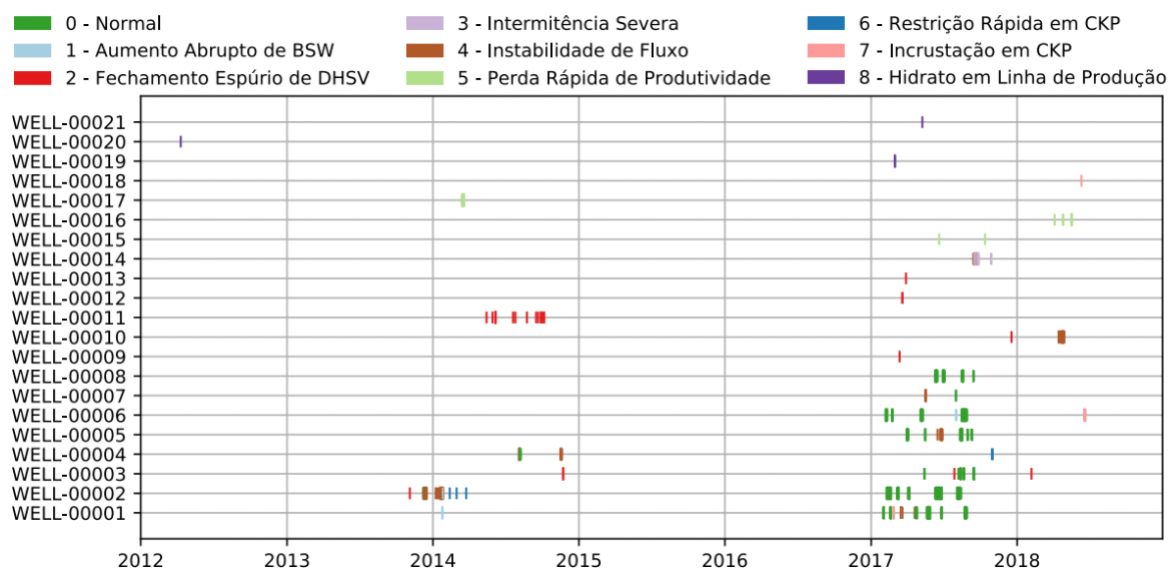


Fonte: Júnior (2022)

Abrupto de BSW), 6 (Restrição Rápida em CKP) e 7 (Incrustação em CKP), presentes nos poços trabalhados.

Os períodos normais dos poços foram separados e enumerados, ordenados por data da operação, e o critério dessa separação foi considerar períodos completos e sem interrupções nas medições. Nas Tabelas 3.1 e 3.2 é possível consultar a separação de período juntamente às informações de tamanho, intervalo e duração da medição por poço. Para cada poço, foi feita a separação dos dados para treinamento do modelo considerando alguns períodos normais do poço, os quais apresentam maiores diferenças na dinâmica das variáveis, com o intuito de

Figura 3.2 – Mapa de dispersão das instâncias reais históricas do 3W dataset



Fonte: Vargas (2019b)

abranger o treinamento a diferentes comportamentos de um mesmo poço. Os períodos normais remanescentes juntamente com os períodos com anomalia foram utilizados para a validação do modelo de *soft sensor*. A Tabela 3.3 apresenta a separação dos períodos de treinamento do modelo para os dois poços.

Tabela 3.1 – Informações das amostras do poço 1

| Período | Tamanho | Intervalo | Duração |
|---------|---------|---------------------|-----------|
| 1 | 2084 | 01/02/17 a 02/02/17 | 34h57min |
| 2 | 2983 | 18/02/17 a 20/02/17 | 49h56min |
| 3 | 353 | 21/04/17 | 5h51min |
| 4 | 356 | 22/04/17 | 5h54min |
| 5 | 1491 | 24/04/17 a 25/04/17 | 24h57min |
| 6 | 11193 | 20/05/17 a 29/05/17 | 199h54min |
| 7 | 1178 | 26/06/17 a 27/06/17 | 19h52min |
| 8 | 7756 | 23/08/17 a 28/08/17 | 117h23min |
| 9* | 209 | 24/01/14 a 25/01/14 | 3h28min |
| 10** | 735 | 26/02/17 a 27/02/17 | 12h14min |

*Presença de anomalia do tipo 1 **Presença de anomalia do tipo 7

Tabela 3.2 – Informações das amostras do poço 2

| Período | Tamanho | Intervalo | Duração |
|---------|---------|---------------------|-----------|
| 1 | 597 | 11/12/13 a 12/12/13 | 9h57min |
| 2 | 356 | 14/12/13 | 5h56min |
| 3 | 15197 | 09/02/17 a 20/02/17 | 253h17min |
| 4 | 4766 | 07/03/17 a 11/03/17 | 79h26min |
| 5 | 4171 | 04/04/17 a 07/04/17 | 69h31min |
| 6 | 23082 | 11/06/17 a 27/06/17 | 384h42min |
| 7 | 12802 | 27/06/17 a 13/08/17 | 213h22min |
| 8* | 407 | 12/02/14 | 6h47min |
| 9* | 190 | 01/03/14 | 3h10min |
| 10* | 199 | 25/03/14 | 3h19min |

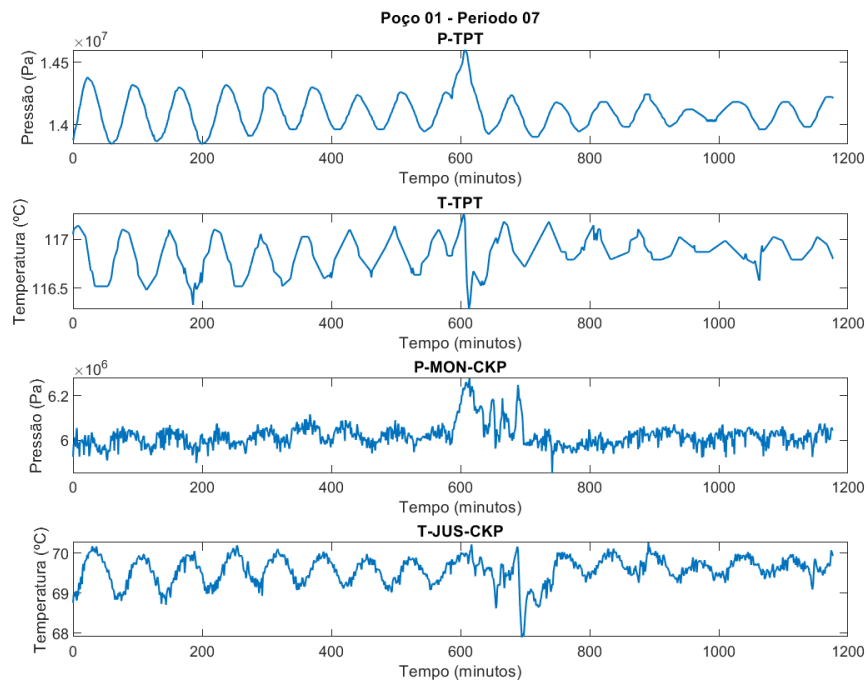
*Presença de anomalia do tipo 6

Tabela 3.3 – Separação de períodos de treinamento e validação

| Poço | Períodos de treinamento | Percentual sobre a base de dados |
|------|-------------------------|----------------------------------|
| 1 | 1, 2, 3, 4, 6 | 64,64% |
| 2 | 3, (50%) 6, 7 | 64,85% |

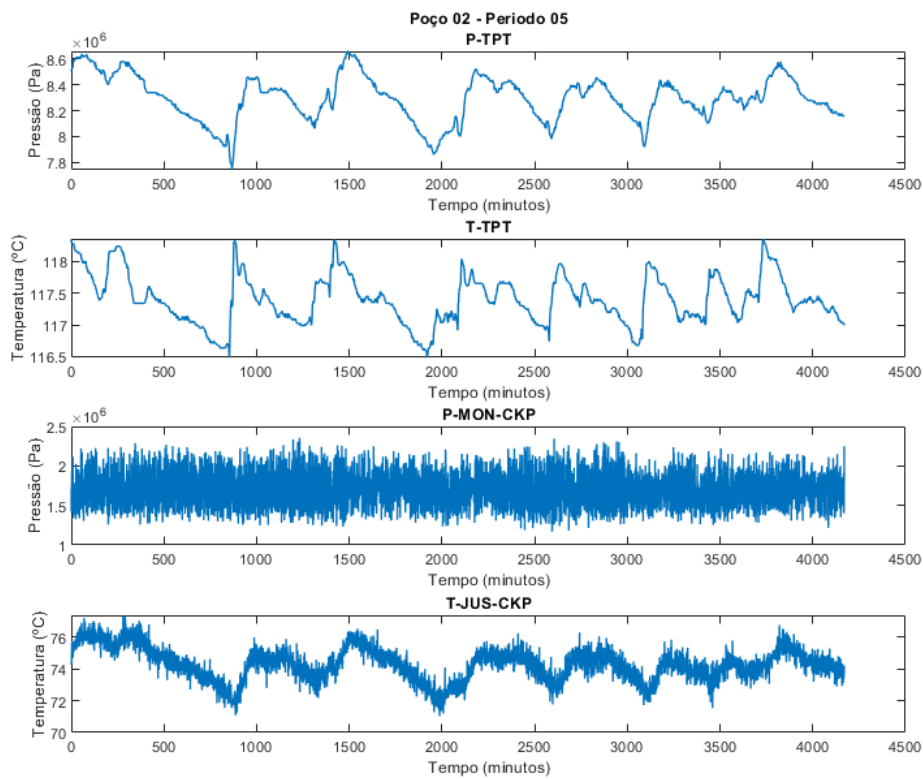
Nas Figuras 3.3 e 3.4 é possível ver o comportamento das variáveis de processo de dois períodos em operação normal dos poços 1 e 2, respectivamente. Nota-se que os dados apresentam dinâmica de variação das pressões e temperaturas, com certos padrões entre elas e presença de ruído nas medições. Por se tratarem de pontos de operação diferentes, as pressões e temperaturas médias podem variar entre períodos e poços mas, apesar disso, apresentam mesma relação entre pressões e temperaturas, aumentando e reduzindo concomitantemente. Esse comportamento dinâmico é interessante para o desenvolvimento de um *soft sensor*, já que permite a identificação de um comportamento e, conseqüentemente, viabiliza a aplicação do mesmo sobre esse conjunto de dados.

Figura 3.3 – Visualização das medições realizadas entre 26/06/2017 a 27/06/2017, operação normal do poço 1



Fonte: Própria

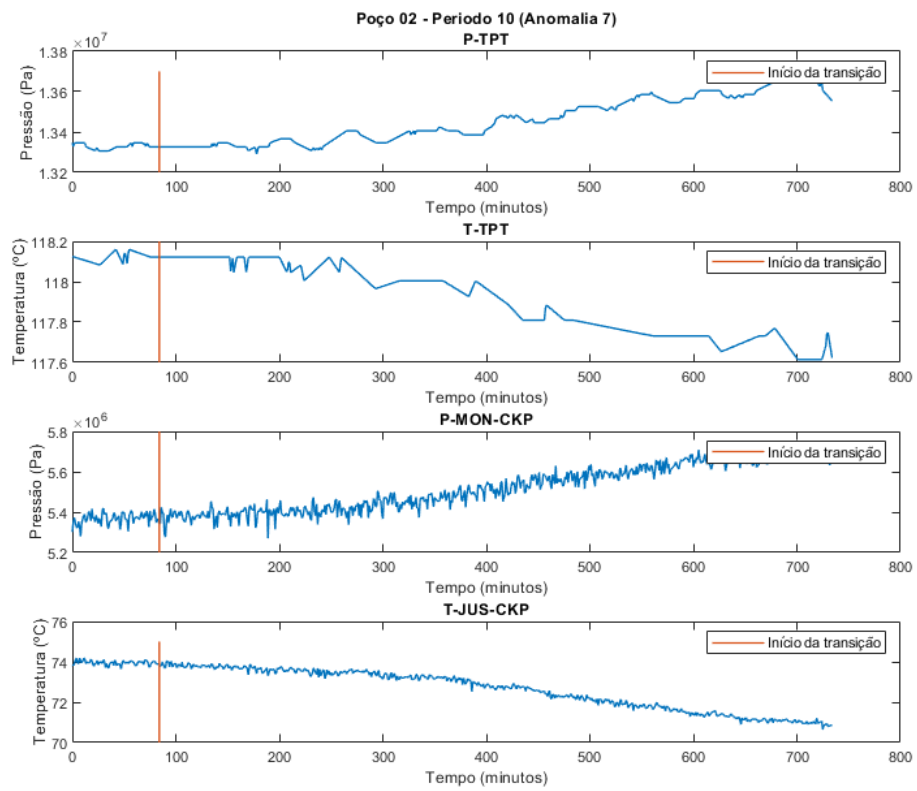
Figura 3.4 – Visualização das medições realizadas entre 04/04/2017 a 07/04/2017, operação normal do poço 2



Fonte: Própria

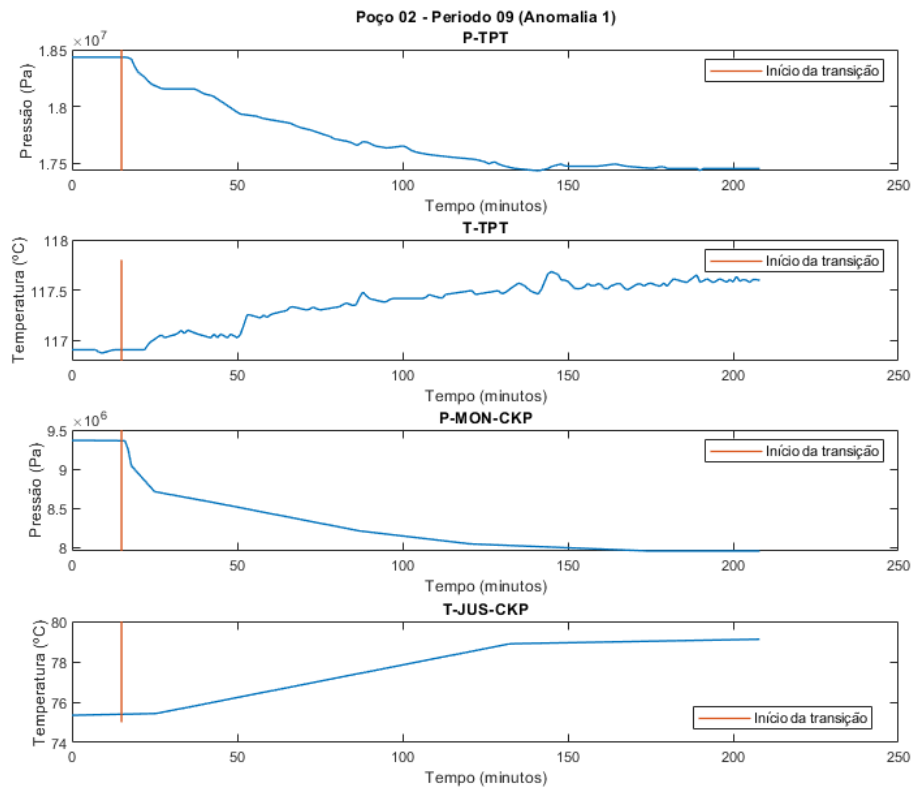
Quando se trata de períodos com transitório de anomalias, o padrão entre as variáveis de operação tende a mudar comparado à dinâmica normal, podendo variar conforme o tipo da anomalia. A Figura 3.5 mostra um período com ocorrência de incrustação de CKP (Tipo 7) em que a transição do período normal para anomalia se dá pelo aumento das pressões e redução das temperaturas. Já na ocorrência de aumento abrupto de BSW (Tipo 1), a anomalia é caracterizada pela redução das pressões e o aumento das temperaturas, como pode ser visto na Figura 3.6. O uso desses períodos para validação será adequado para entender como o *soft sensor* se comporta perante dados que apresentam mudança de dinâmica ao longo das medições, bem como a avaliação de possível detecção de anomalia a partir do modelo de predição. A linha vertical nas imagens representa o início da transição de medição normal para medição em estado anômalo, conforme definido pelos autores do banco de dados 3W.

Figura 3.5 – Visualização das medições realizadas entre 26/02/2017 a 27/02/2017, operação anormal do tipo 7 do poço 1



Fonte: Própria

Figura 3.6 – Visualização das medições realizadas entre 24/01/2014 a 25/01/2014, operação anormal do tipo 1 do poço 1



Fonte: Própria

3.2 Algoritmo MGGP

O algoritmo MGGP, apresentado na subsecção 2.3.2 é utilizado neste trabalho por meio da *toolbox* desenvolvida em Castro (2021), baseada em Python e disponível sob Licença Pública Geral.

A função responsável por carregar os atributos e funções usadas para criar e avaliar os indivíduos de uma população MGGP recebe como parâmetros o valor correspondente ao número máximo de operadores de retrocesso (atraso) a serem incluídos no conjunto primitivo, um parâmetro booleano que permite o uso da variável que representa termos residuais e um último parâmetro que refere-se à quantidade de variáveis de entrada e saída. Considera-se as variáveis P-TPT, T-TPT e P-MON-CKP entradas do modelo e T-JUS-CKP a saída. As entradas são dados de sensores que se encontram antes do sensor de saída, considerando o sentido de escoamento do fluido.

A definição de alguns hiperparâmetros utilizados no modelo do *soft sensor* foi baseada na análise das variações de seus valores aplicados aos dados do poço 1. Os demais parâmetros do MGGP foram mantidos conforme definido em (CASTRO; BARBOSA, 2022), por não apresentarem variações em relação a outros trabalhos que utilizam o MGGP e não apresentarem impactos significativos no modelo em questão. Tais resultados são apresentados na seção 4.1.1.1.

3.3 Medidas de Desempenho

A avaliação de desempenho do *soft sensor* utilizando o algoritmo MGGP foi definida utilizando o MAPE (*Mean Absolute Percentage Error*) que fornece a informação de adequação do modelo por meio do erro relativo entre os dados reais e a saída do modelo expresso em percentual, sendo recomendada em trabalhos relacionados com a identificação de sistemas (BARBOSA et al., 2011a; de Moraes et al., 2019) :

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|, \quad (3.1)$$

em que y_i é a i -ésima medida de dados reais, \hat{y}_i é o valor estimado de y_i usando o MGGP e n é o número de amostras de dados utilizadas.

4 RESULTADOS

Para alcançar os resultados desta seção, o treinamento contou com a utilização do algoritmo MGGP em conjunto com um algoritmo baseado em ERR (do inglês, *error reduction ratio*), conforme proposto em (CASTRO; BARBOSA, 2022). Como mostrado por (CASTRO, 2021), o MGGP/ERR é capaz de explorar um amplo espaço de busca para o qual um método tradicional baseado em ERR requereria um poder computacional muito alto.

Foram realizados treinamentos do MGGP utilizando medições de sensores referentes aos poços 1 e 2, em operação normal, sem ocorrência de anomalias. Para cada um deles, utilizou-se um percentual próximo de 60% da base de dados, separadas em períodos de medições sem interrupções, conforme mostrado na seção anterior. A seguir estão os resultados obtidos.

4.1 *Soft sensor*

4.1.1 Poço 1

4.1.1.1 Hiperparâmetros do MGGP

Foram avaliados os hiperparâmetros de treinamento do MGGP referentes à quantidade de passos à frente para predição do erro, quantidade máxima de termos e atraso máximo da equação do modelo, a fim de compreender os impactos dessas mudanças nos resultados do MGGP. O modelo com cada variação desses parâmetros foi executado cinco vezes a fim de validar o resultado, removendo vieses que possam estar vinculado à geração aleatória na população inicial no modelo. Os períodos 1, 2, 3, 4 e 6 foram utilizados no treinamento enquanto os períodos 5, 7 e 8 serviram para a análise do resultado. Dessa forma, a mediana do erro absoluto médio percentual (MAPE, do inglês, *mean absolute percentage error*) das execuções dessas variações foram coletadas e disponibilizadas na Tabela 4.1.

Na Tabela 4.1, os valores destacados são resultados que apresentam menor erro para cada período de validação do poço e, dessa forma, serviram de critério para escolha dos valores desses três parâmetros. Pode-se observar que o uso de vários passos à frente para cálculo do erro de predição foi benéfico para encontrar modelos mais satisfatórios, o que corrobora as discussões em (BARBOSA et al., 2011b; AGUIRRE et al., 2017; FREITAS; BARBOSA; AGUIRRE, 2021). A quantidade máxima de termos delimita a flexibilidade do modelo e foi observado que o uso de 10 termos trouxe um bom equilíbrio entre complexidade e desempenho.

Em relação ao número máximo de atrasos, foi possível observar que quanto maior o número de atrasos permitido, melhor foi o desempenho do modelo, o que mostra que informações de maior lapso temporal são importantes para predição da temperatura de *choke*.

Tabela 4.1 – MAPE das variações dos parâmetros do MGGP aplicado ao poço 1 na predição de temperatura de *choke*

| Parâmetro | Variação | MAPE [%] | | |
|----------------------------------|--------------|-----------|-----------|-----------|
| | | Período 5 | Período 7 | Período 8 |
| Quantidade máxima de termos | 5 | 0,22 | 0,33 | 4,14 |
| | 10 | 0,39 | 0,24 | 1,07 |
| | 20 | 0,71 | 0,84 | 0,39 |
| | 35 | 0,51 | 0,40 | 0,42 |
| Espaço de busca (<i>delay</i>) | 1,5,10,25 | 0,23 | 0,23 | 1,60 |
| | 1,5,10,25,50 | 0,19 | 0,20 | 1,12 |
| | 10,25,50 | 0,15 | 0,30 | 1,41 |
| Passos à frente | 1 | 5,44 | inf | 1,18 |
| | 5 | 0,18 | 0,29 | 10,60 |
| | 50 | 0,27 | 0,36 | 4,14 |
| | 100 | 0,22 | 0,25 | 2,29 |

Em cada treinamento, somente o parâmetro a ser analisado tem seu valor variado e, por padrão, assumem valores fixos enquanto não estão sob avaliação, sendo eles: máximos de 5 termos, 100 passos à frente e espaço de busca que considera 125 de máximo atraso.

Observou-se que aumentar a quantidade máxima de termos para um valor superior a 10 não traz muita vantagem ao modelo, nota-se que utilizando os valores de 5 e 10 neste parâmetro gera um melhor resultado comparado aos demais e, por isso, optou-se por um valor intermediário, definindo a quantidade máxima de 7 termos para treinamento.

Em relação a avaliação de passos à frente, o treinamento com valores menores resulta em modelos mais instáveis, elevando o erro consideravelmente quanto menor for o valor deste parâmetro, como pode ser visto nas variações 1, 5 e 50. Dessa forma, o treinamento será realizado considerando 100 passos à frente.

Por fim, a definição do vetor referente ao espaço de busca (*delays*) é essencial para a definição do atraso máximo já que o maior atraso possível é calculado por meio do maior *delay* do vetor multiplicado pela profundidade máxima da árvore. É possível ver a melhora desse resultado quando considerado o valor de 50 nas variações de espaço de busca, porém, no poço 1, há períodos curtos, com pouca quantidade de amostras disponíveis para uso como condições iniciais durante a predição, o que limita a predição quando considerado um valor de atraso elevado.

Os parâmetros considerados na função de evolução da população são apresentados na Tabela 4.2.

Tabela 4.2 – Parâmetros do MGGP

| Parâmetro | Valor |
|-----------------------------------|--------------|
| Tamanho da população | 100 |
| Número de gerações | 100 |
| Profundidade máxima da árvore | 5 |
| Número máximo de genes (termos) | 7 |
| Atraso máximo | 125 |
| Passos a frente | 100 |
| Taxa de mutação | 0,1 |
| Taxa de cruzamento | 0,8 |
| Espaço de busca (<i>Delays</i>) | [1,5,10,25] |

4.1.1.2 Desempenho do modelo

Foram realizados treinamentos considerando os hiperparâmetros definidos conforme mostrado anteriormente, utilizando os períodos 1, 2, 3, 4 e 6 de operação normal do poço, e as validações do modelo resultante foram feitas com os demais períodos disponíveis do poço (Períodos 5, 7 e 8), também sem ocorrência de anomalias. Os períodos 9 e 10, que apresentam transição para o estado de algum tipo de anomalia, foram utilizadas para algumas observações de desempenho do *soft sensor* dado um comportamento inesperado das variáveis. As variáveis de entrada são P-TPT, T-TPT, P-MON-CKP e T-JUS-CKP e a variável a ser estimada pelo modelo.

Na Tabela 4.3 é apresentado o erro percentual (MAPE) entre dados reais e estimados pelo modelo para cada período, resultado da simulação de execução livre. Nota-se que, os períodos normais de treinamento e de validação, os erros não ultrapassam 0,64%, variando de acordo com a dinâmica individual de cada período que pode apresentar ruído e interferências do ambiente.

Tabela 4.3 – Resultados do modelo NARX obtido para predição de T-JUS-CKP.

| Período | MAPE [%] |
|-----------------|-----------------|
| 1 ^t | 0,39 |
| 2 ^t | 0,64 |
| 3 ^t | 0,12 |
| 4 ^t | 0,22 |
| 5 ^v | 0,24 |
| 6 ^t | 0,47 |
| 7 ^v | 0,20 |
| 8 ^v | 0,64 |
| 9 ^a | 4,00 |
| 10 ^a | 0,99 |

^tPeríodo de treinamento

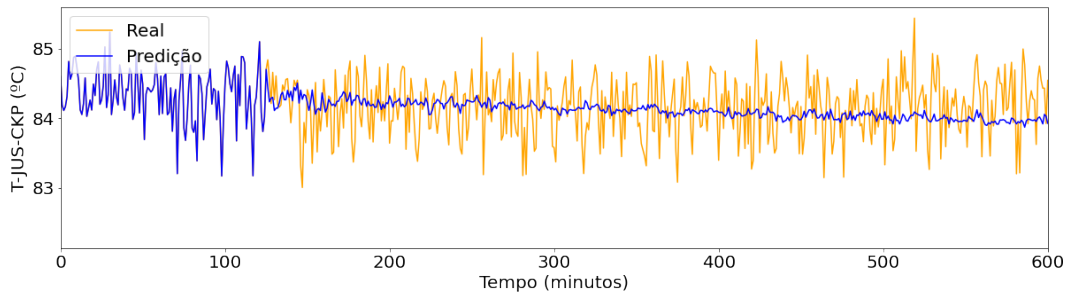
^vPeríodo de validação

^aPeríodo com anomalia

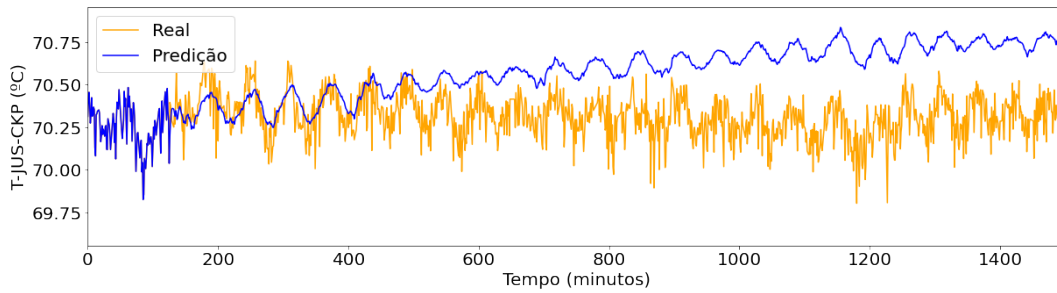
As Figuras 4.1a, 4.1b, 4.1c, 4.1d e 4.1e mostram a predição resultante do modelo para os períodos de treinamento 1, 2, 3, 4 e 6, respectivamente. Os resultados são curvas reais e de predição semelhantes apesar de pouco transitório e com presença de ruído nos dados.

Tais resultados mostram que o modelo é capaz de capturar a relação dos dados medidos e a estimativa futura. Isso pode ser visto também nas Figuras 4.2a, 4.2b e 4.2c, referentes aos períodos de validação do MGGP. É possível visualizar como as temperaturas estimadas acompanham a curva de temperaturas reais do poço. A média das temperaturas estimadas se assemelha àquela dos dados reais e conseguem acompanhar a oscilação natural da temperatura real, apresentando erros inferiores a 0,5 graus Celsius. O período 8 apresenta mais ruído em seus dados e opera em uma média de temperatura diferente da maioria dos períodos de treinamento. Esse comportamento se assemelha ao período 1 de treinamento e, por isso, o modelo é capaz de estimar aumento e diminuição de temperatura, mas em uma amplitude menor nestes períodos.

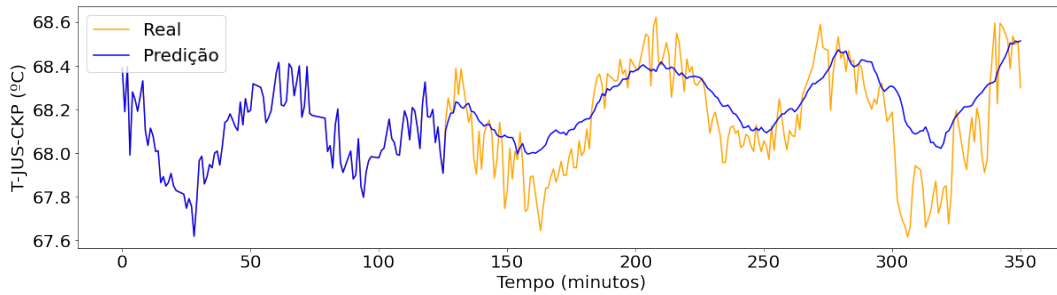
Figura 4.1 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de treinamento. Fonte: Própria.



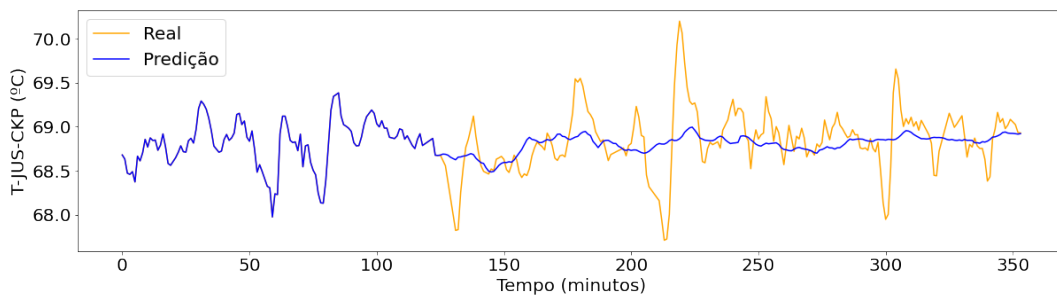
(a) Período 1



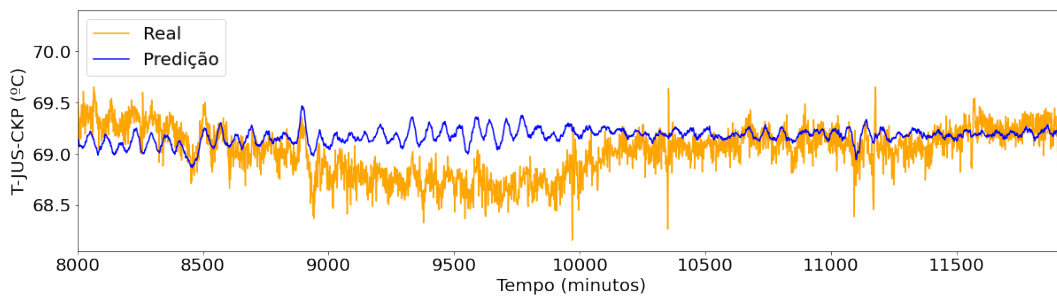
(b) Período 2



(c) Período 3

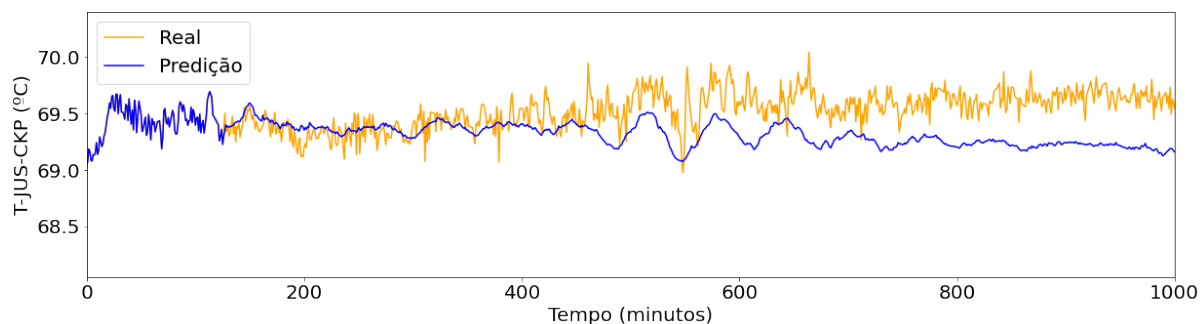


(d) Período 4

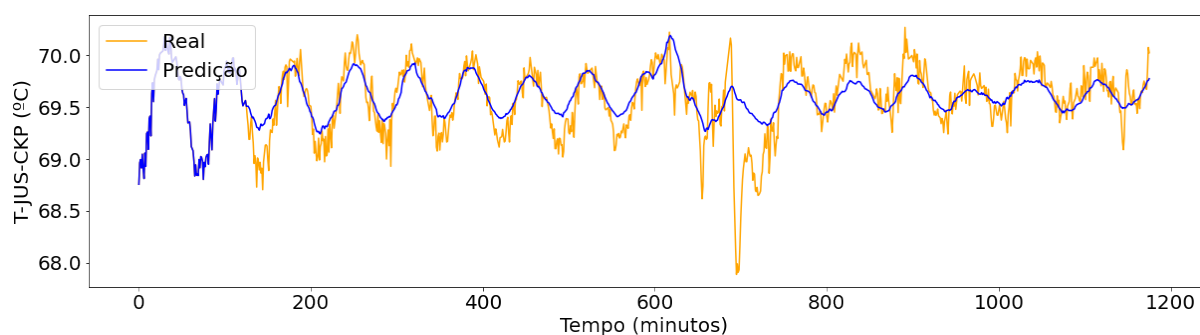


(e) Período 6

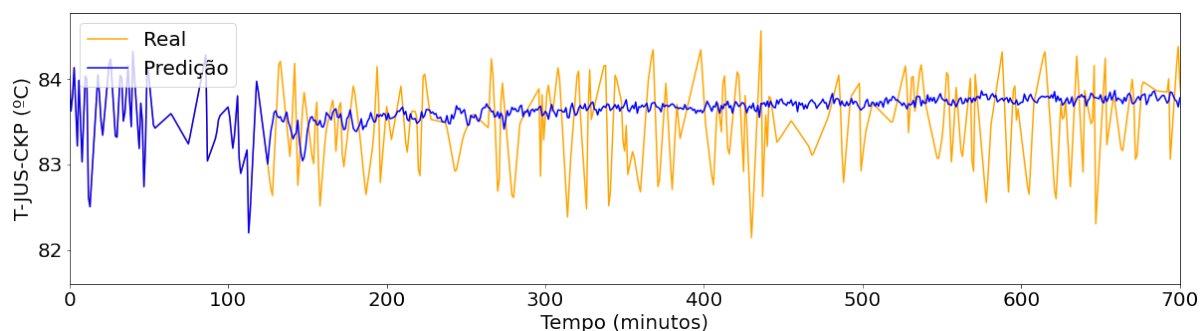
Figura 4.2 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de validação. Fonte: Própria.



(a) Período 5



(b) Período 7



(c) Período 8

Também foram realizadas simulações de predição de k passos à frente, considerando valores diferentes de k sobre os períodos de validação. Essa simulação consiste em dividir os dados de cada período em janelas curtas a serem analisadas individualmente, em que, para cada janela é considerada uma condição inicial diferente baseada nos dados reais daquele instante. Sendo assim, a predição por janela permite que o modelo considere dados de instantes anteriores mais próximos de onde se deseja prever, se ajustando a eventuais pequenas mudanças na operação normal que o período completo pode vir a ter no decorrer das medições. A divisão dos dados dos períodos é dada pela divisão do tamanho do período em questão, pela soma de passos a frente k e tamanho da condição inicial do modelo, baseada no máximo atraso resultante.

Os dados restantes são descartados. Os resultados da simulação para k igual a 20, 100 e 200 estão dispostos na Tabela 4.4, onde é possível notar que a predição do modelo para a maioria dos períodos normais, não apresenta mudança significativa entre os valores para cada k , mas há uma tendência de aumento do MAPE com o aumento do k . O desvio padrão tende a se reduzir com o aumento de k , principalmente quando comparado valores de $k = 20$ e demais valores apresentados na tabela. Os períodos 3 e 4 são menores e, por isso, para $k = 100$ e $k = 200$ é possível avaliar apenas uma janela de predição, resultando em desvio padrão igual a 0. Prever com baixo erro um comportamento de um período maior de dados é algo desejável em diversos contextos pois permite tomada de decisões antecipadas e, dado tal resultado, torna-se vantajoso o uso de mais passos a frente nas simulações deste modelo.

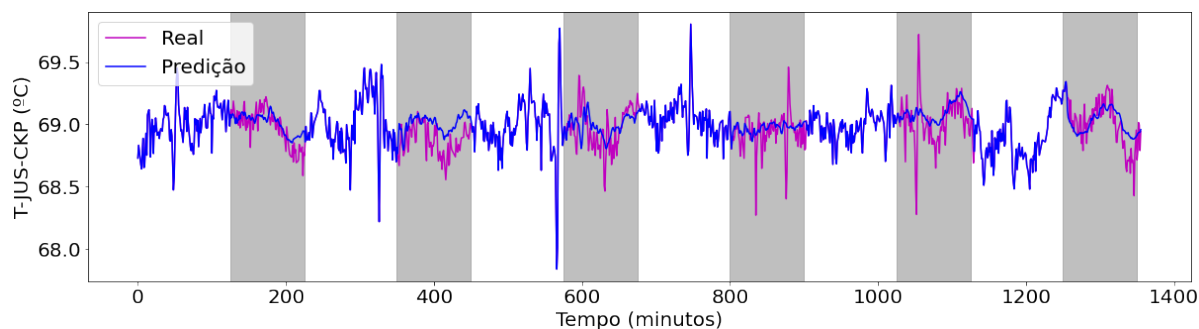
Tabela 4.4 – Resultado de simulações de predição de k passos à frente para os períodos normais

| Período | Mediana e desvio padrão do MAPE em simulação de k passos à frente | | |
|---------|---|-----------------|-----------------|
| | $k = 20$ | $k = 100$ | $k = 200$ |
| 1 | $0,41 \pm 0,09$ | $0,41 \pm 0,04$ | $0,40 \pm 0,02$ |
| 2 | $0,10 \pm 0,04$ | $0,12 \pm 0,02$ | $0,16 \pm 0,03$ |
| 3 | $0,11 \pm 0,01$ | $0,16 \pm 0,0$ | $0,19 \pm 0,0$ |
| 4 | $0,27 \pm 0,16$ | $0,42 \pm 0,0$ | $0,35 \pm 0,0$ |
| 5 | $0,12 \pm 0,11$ | $0,17 \pm 0,03$ | $0,18 \pm 0,02$ |
| 6 | $0,12 \pm 0,05$ | $0,13 \pm 0,04$ | $0,15 \pm 0,06$ |
| 7 | $0,10 \pm 0,11$ | $0,25 \pm 0,07$ | $0,21 \pm 0,02$ |
| 8 | $0,39 \pm 0,14$ | $0,39 \pm 0,06$ | $0,40 \pm 0,05$ |

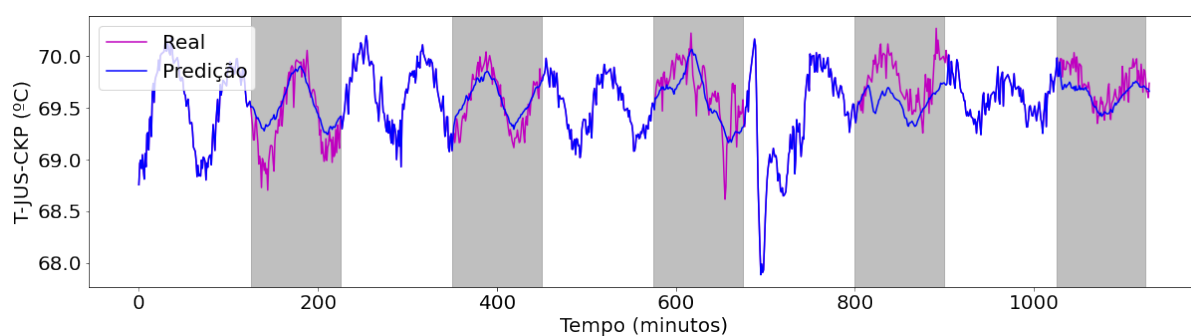
As Figuras 4.3a, 4.3b e 4.3c ilustram o comportamento da predição em janelas de 100 passos à frente em um determinado intervalo de tempo dos períodos normais de validação, em que a região destacada representa a janela de predição, as curvas de cor magenta representam o comportamento desejado e podem ser comparadas às curvas de predição em azul nesta janela. O comportamento nessas janelas apresentam um resultado satisfatório. Comparada à simulação livre, a predição de k passos à frente se mostra mais aproximada aos valores reais. Isso é mais visível quando analisado o período 8, em que possui curvas com amplitudes mais acentuadas que os demais períodos e a simulação com $k=100$ suaviza levemente essa diferença presente na Figura 4.2c, uma vez que a recursão faz com que erros anteriores sejam acumulados em um longo período de simulação.

O erro percentual para cada janela de predição pode ser visto na Figura 4.4. Para todos os períodos e janelas, o MAPE apresenta valores baixos, com picos em janelas com maior presença de ruído nos dados, sendo reduzido nas janelas seguintes.

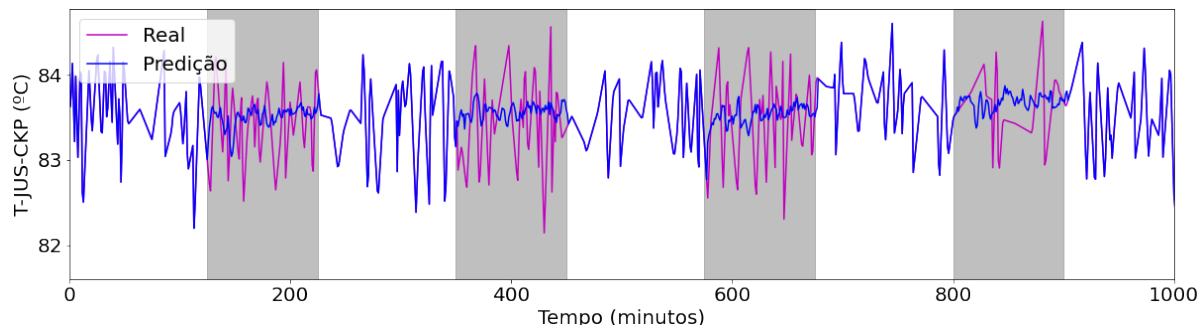
Figura 4.3 – Comparativo dos dados reais e estimados em predição de 100 passos à frente pelo modelo nos três períodos de validação. Fonte: Própria.



(a) Período 5



(b) Período 7

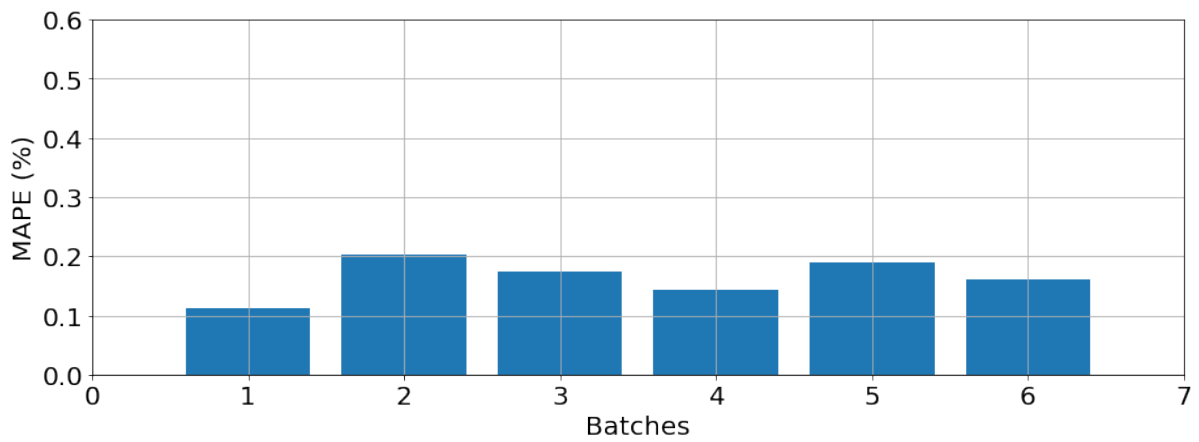


(c) Período 8

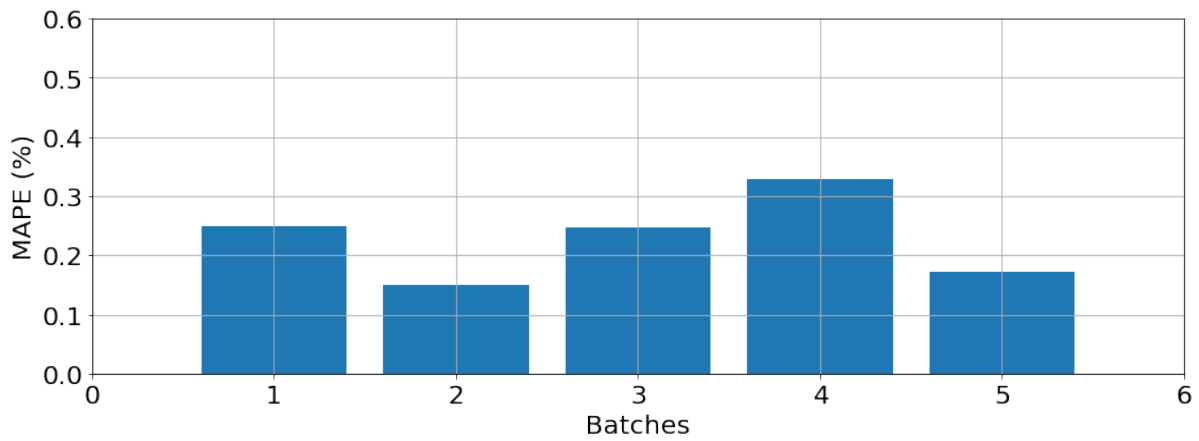
4.1.1.3 Anomalia 1 (Aumento Abrupto de BSW)

Analisando o modelo de predição aplicado aos períodos com transitório de anomalia, a simulação 100 passos à frente tende a ser diferente dos períodos normais em caso de aumento abrupto BSW (do inglês, *Basic Sediments and Water*), como pode ser visto na Figura 4.5, em que os valores de temperatura a partir da linha de início da transição sinalizam uma mudança no comportamento esperado pelo poço, ainda consideradas variações normais, até alcançar o patamar da linha de início da anomalia. BSW é uma especificação técnica de certas impurezas no petróleo bruto, um termo padrão para descrever os elementos inutilizáveis em um fluxo de poço. Por isso, monitorar essa relação é uma forma de evitar perdas, uma vez que o aumento

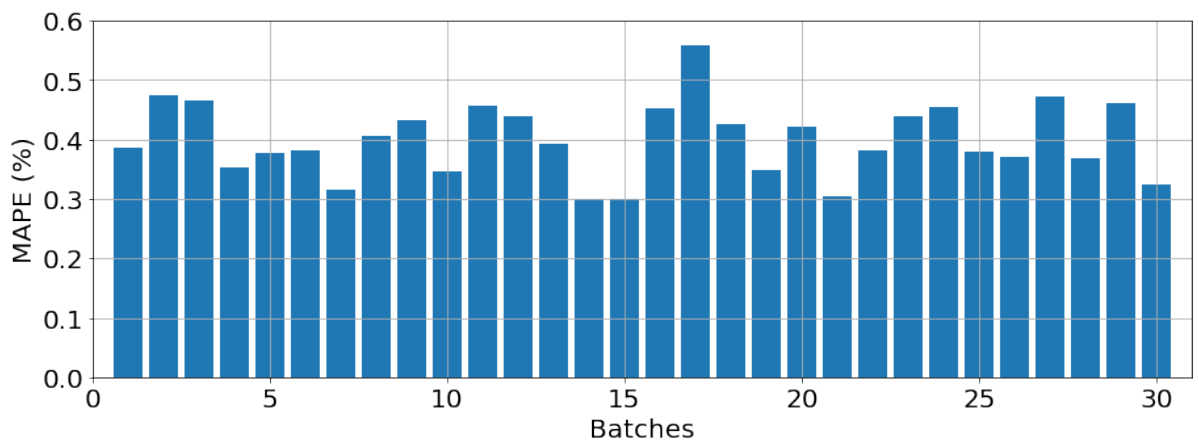
Figura 4.4 – Comparativo entre os erros percentuais das janelas em predição 100 passos à frente pelo modelo nos três períodos de validação. Fonte: Própria.



(a) Período 5



(b) Período 7

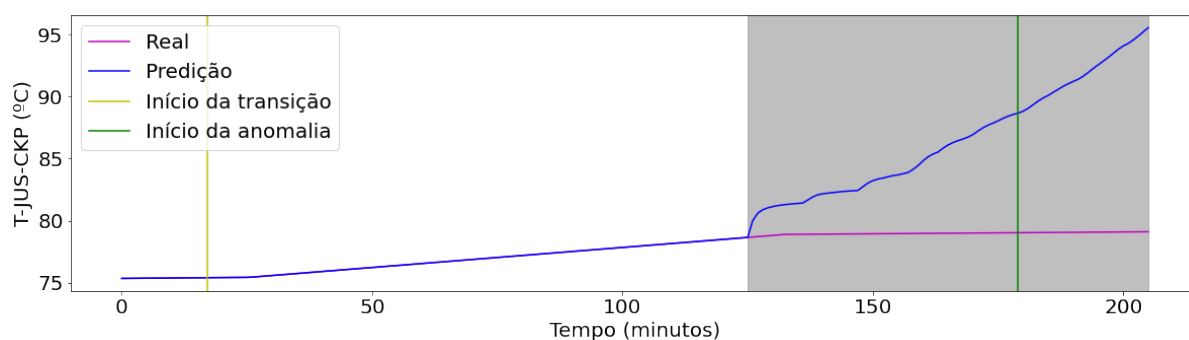


(c) Período 8

abrupto de BSW acarreta em elevação das temperaturas medidas nos equipamentos, e pode servir de critério para venda às refinarias de petróleo.

Observando a resposta do modelo a esse tipo de anomalia, nota-se que apresenta um erro de predição já na primeira janela do período, um erro mais elevado comparado a todas às janelas analisadas em períodos normais, superando 10 graus Celsius de diferença entre dado real e estimado da temperatura T-JUS-CKP. O modelo não é capaz de explicar os casos de anomalias, um indicativo da possibilidade de ser utilizado em análises de detecção desta anomalia em poços de petróleo.

Figura 4.5 – Simulação livre dada pelo modelo de período 9 de validação com anomalia. Fonte: Própria.



(a) Período 9

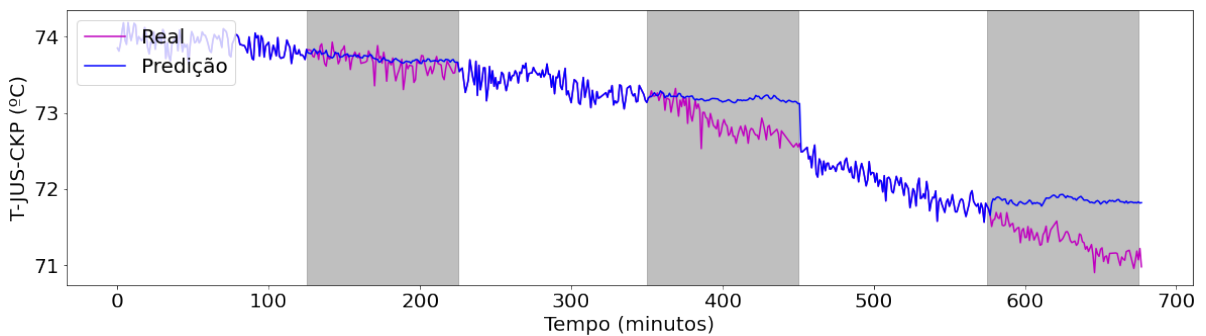
4.1.1.4 Anomalia 7 (Incrustação em CKP)

A incrustação na válvula de *choke* é ocasionada pelo acúmulo gradual de partículas nessa região que reduz ou obstrui o espaço de passagem do fluido. Com isso, há um aumento na pressão montante à válvula de *Choke* (P-MON-CKP) devido ao estrangulamento na região subsequente e uma redução na temperatura jusante à válvula de *choke* (T-JUS-CKP) pela redução no fluxo após a região estrangulada. Assim como na anomalia 1, espera-se que o modelo seja incapaz de acompanhar esse tipo de comportamento anômalo. Porém, como pode ser visto na Figura 4.6, quando simulado com 100 passos à frente, a predição das janelas iniciais resulta em erros semelhantes aos períodos normais, apresentando um aumento gradual nas janelas subsequentes do período.

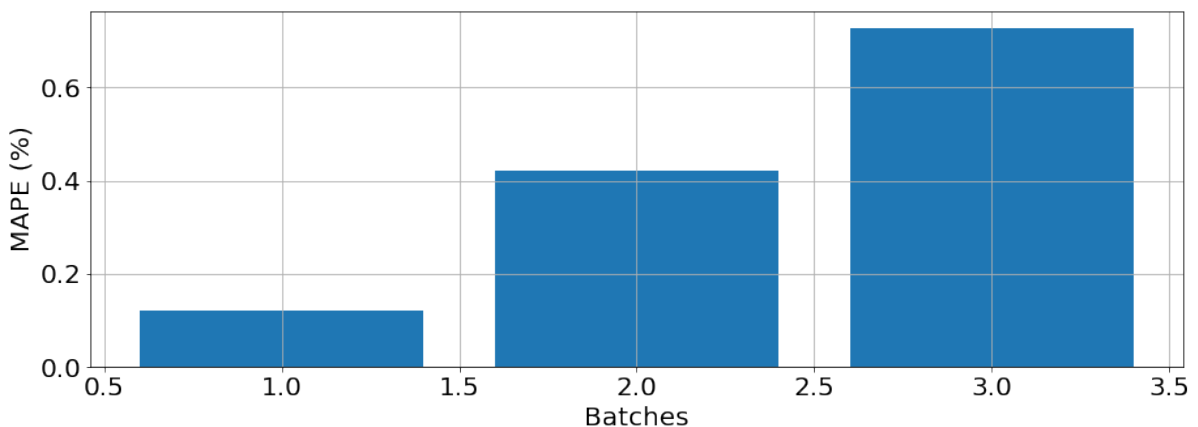
Em Vargas (2019b), é abordado como é feita a confirmação de ocorrências reais de cada anomalia da base de dados. Profissionais analisam janelas temporais com tamanhos diferentes a fim de compreenderem a dinâmica de cada tipo de anomalia, gerando uma estimativa de tamanho de janela para cada uma delas. A estimativa para esta anomalia é de cerca de 72 horas, sendo a mais elevada dentre as anomalias em questão, que variam entre 5 minutos à 12 horas. Levanto essa estimativa em questão, é possível perceber o quanto a transição para essa anomalia

lia pode ser lenta, não apresentando muita variação nas pressões e temperaturas em um curto intervalo de tempo, podendo confundir o modelo quando simulado com janelas pequenas de predição. Dado isso, simulando com mais passos à frente, é possível notar como o comportamento do modelo muda, aumentando o erro percentual a cada janela que se aproxima ao estado de anomalia e diferenciando do comportamento que ocorre nos períodos normais, conforme esperado. A Figura 4.7 mostra essa simulação considerando 200 passos à frente. O mesmo pode ser visto na simulação livre, apresentada na Figura 4.8. O erro aumenta significativamente quanto maior a janela de tempo analisada, alcançando 2 graus *Celsius* de diferença entre real e estimado no início do estado anômalo.

Figura 4.6 – Simulação de predição de 100 passos à frente e erro percentual por janela dada pelo modelo de período 10 de validação com anomalia. Fonte: Própria.

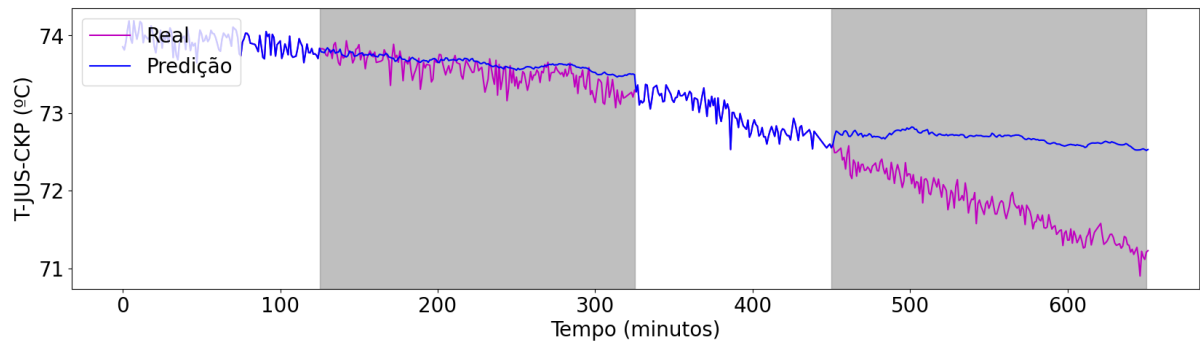


(a) Comparativo entre dados reais e predição do modelo

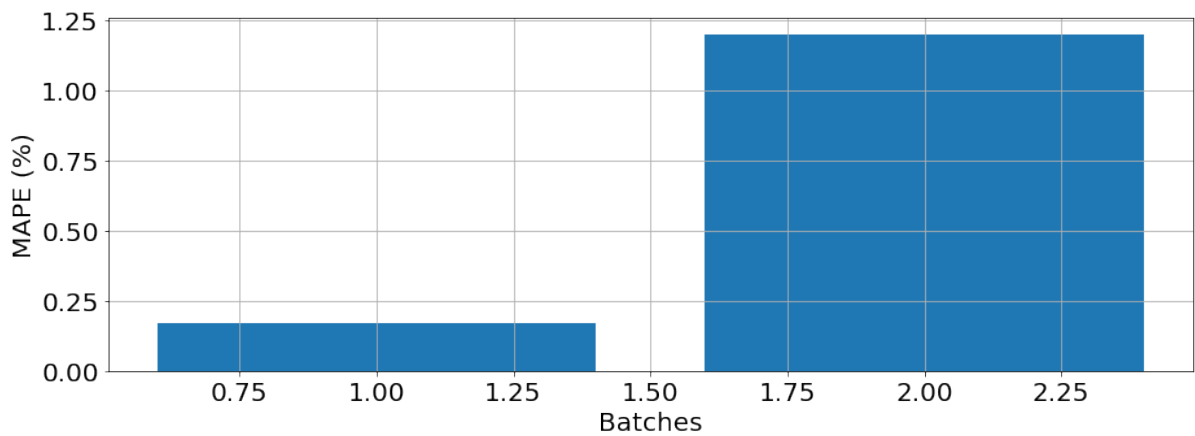


(b) Erro percentual absoluto médio por janela de predição

Figura 4.7 – Simulação de predição de 200 passos à frente e erro percentual por janela dada pelo modelo de período 10 de validação com anomalia. Fonte: Própria.

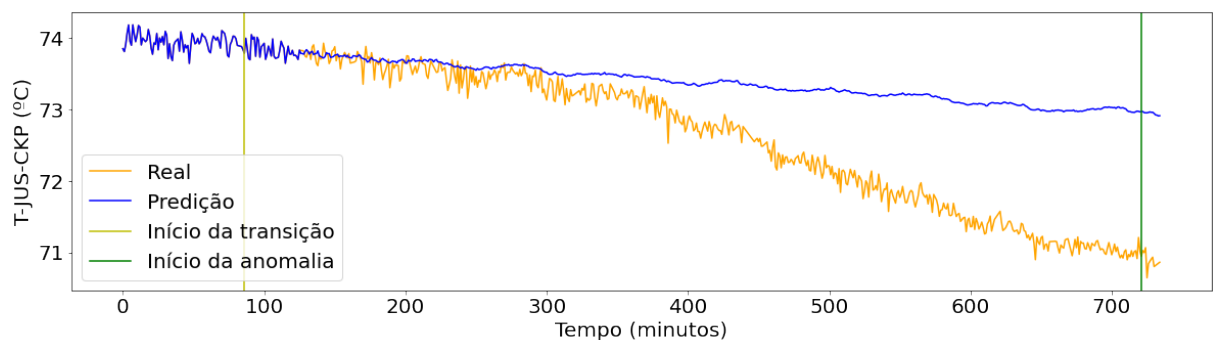


(a) Comparativo entre dados reais e predição do modelo



(b) Erro percentual absoluto médio por janela de predição

Figura 4.8 – Simulação livre do modelo no período 10 de validação com anomalia. Fonte: Própria.



4.1.1.5 Representação do modelo

O treinamento do modelo resultou em uma função que melhor representa a dinâmica da temperatura justante à válvula *Choke*, dada às variáveis de pressão e temperaturas de entrada.

O modelo encontrado é dado por:

$$\begin{aligned}
y(k) = & \theta_0 + \theta_1[y(k-38)y(k-32)] + \theta_2y(k-126) + \theta_3u_3(k-6) \\
& + \theta_4[y[(k-11)u_2(k-11)u_2(k-21)y(k-22)u_2(k-26)u_2(k-11)]] \\
& + \theta_5[u_2(k-27)u_2(k-27)u_2(k-52)] + \theta_6y(k-1) + \theta_7y(k-34)
\end{aligned} \tag{4.1}$$

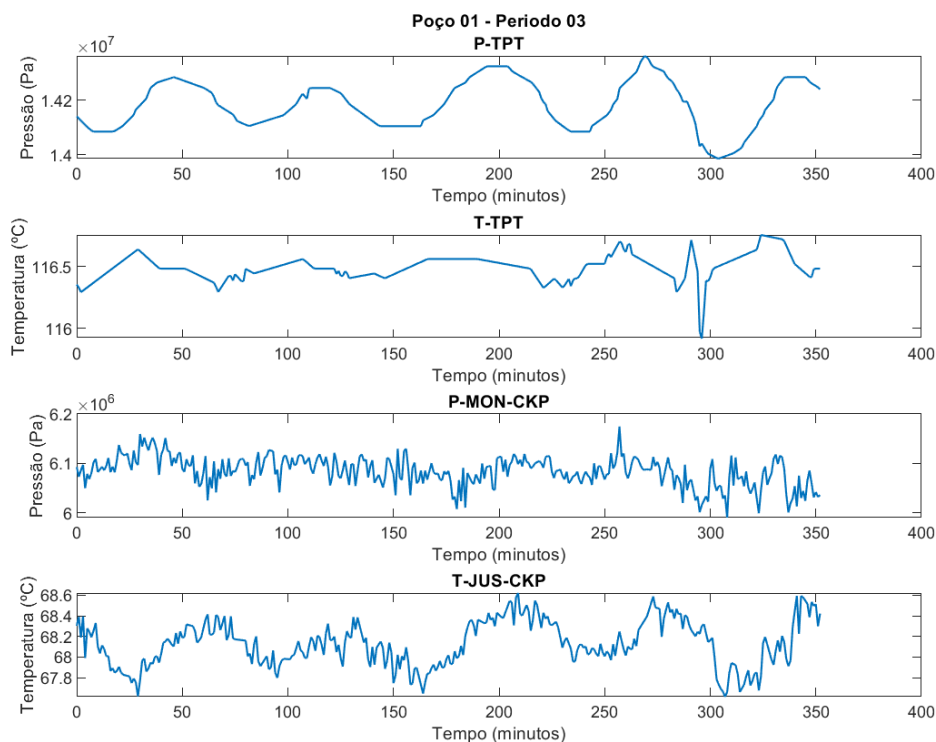
em que u_1 , u_2 e u_3 representam as entradas T-TPT, P-TPT e P-MON-CKP, respectivamente, e y representa a saída T-JUS-CKP, com atrasos. θ_n são os coeficientes de cada termo da função, apresentados no vetor Θ a seguir:

$$\begin{aligned}
\Theta = & [2,05 \times 10^1 \ 1,60 \times 10^{-3} \ 8,87 \times 10^{-2} \ 2,37 \times 10^{-7} \\
& 2,30 \times 10^{-32} \ 1,95 \times 10^{-21} \ 4,38 \times 10^{-1} \ 1,03 \times 10^{-1}].
\end{aligned} \tag{4.2}$$

A variável de entrada que representa a pressão na árvore de natal (u_2) e a de saída, que representa a temperatura justante à válvula de *choke* (y), foram as variáveis que mais apareceram no modelo final, repetindo 7 vezes cada uma. A temperatura na árvore de natal (u_1) e a pressão montante à válvula de *choke* (u_3) são as variáveis que menos influenciam na equação, sendo u_1 não utilizada pelo modelo. Além disso, a maioria dos termos possuem menos de 50 atrasos, apenas um termo explora o máximo atraso definido de 126 passos atrás. Isso ressalta a importância de medições mais próximas do período de predição para o modelo.

Dessa forma, anomalias mais lentas, como é o caso da anomalia 7, que apresenta pouca variação em um curto período de tempo podem ser interpretadas pelo modelo como medições normais nas janelas iniciais da transição da anomalia, a princípio alcançando uma predição com baixo erro em valores futuros. Porém, no decorrer da simulação, há uma tendência de aumento no erro de predição sinalizando um comportamento distinto do esperado em períodos normais. Nos períodos de treinamento, as amostras da variável T-TPT, de menor relevância para o modelo, apresentam um comportamento de variação lenta comparada às demais variáveis, não alcançando 1 grau Celsius de diferença em alguns casos. A Figura 4.9 apresenta um exemplo da dinâmica do poço no período 3. Isso mostra como a dinâmica do T-TPT não afeta o desempenho do modelo e como o treinamento do MGGP foi capaz de entender a importância das variáveis certas para um modelo de predição da T-JUS-CKP, resultando nesta equação.

Figura 4.9 – Visualização das medições do período 3, operação normal do poço 1. Fonte: Própria.



A partir dessa função, é possível verificar se a quantidade máxima de termos bem como o atraso máximo e a profundidade máxima da árvore, definidas inicialmente e apresentadas na Tabela 4.2, foi explorada pelo modelo, alcançando tais máximos. Apenas duas variáveis do processo foram mais importantes no modelo, aparecendo em diversos termos da função final.

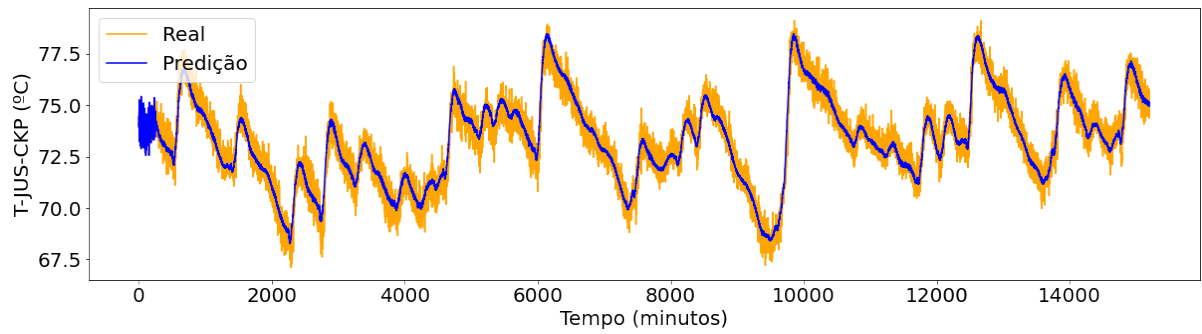
O modelo simulado em períodos normais de validação apresentara resultados satisfatórios. O algoritmo se mostra capaz de identificar a dinâmica de um poço, mesmo com presença de ruído ou condições que diferem entre os dados dos períodos de treinamento. Quando observado sobre um período com transição de anomalia, o modelo não é capaz de fornecer uma predição do comportamento dos dados reais com baixo erro. Porém, o tamanho das janelas de predição se mostra um fator importante para certas anomalias, o qual pode interferir no resultado de uma predição de passos à frente quando tratar-se de uma transição para estado anômalo lenta, podendo se confundir com o comportamento encontrado pelo modelo sobre os períodos normais.

4.1.2 Poço 2

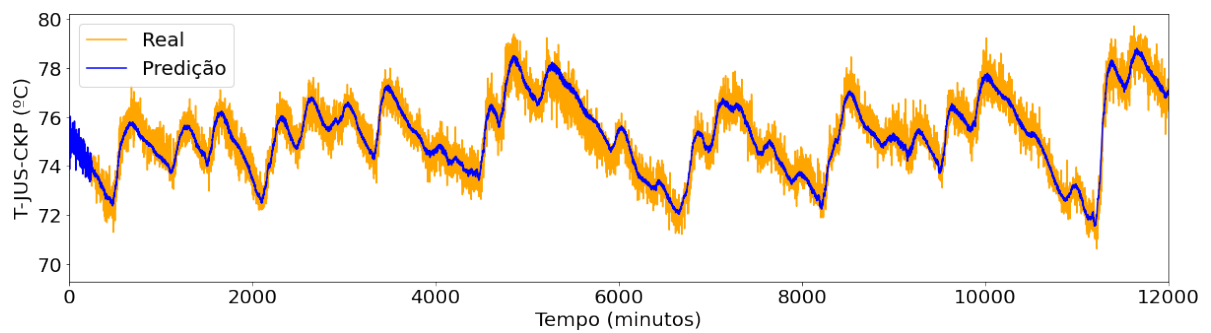
Foram realizados treinamentos utilizando os períodos 3, 7 e 50% iniciais do período 6 de operação normal do poço, e as validações do modelo resultante foram feitas com os demais períodos disponíveis do poço 2 (Períodos 4, 5 e os 50% finais do período 6), também sem ocorrência de anomalias. Os períodos 8, 9 e 10, que apresentam transição para o estado de anomalia do tipo 6 em períodos diferentes de medição, foram utilizados para algumas observações de desempenho do *soft sensor* dado um comportamento inesperado das variáveis. Os períodos 1 e 2 foram desconsiderados devido ao comportamento estacionário em suas medições. Assim como considerado no treinamento do poço 1, as variáveis de entrada são P-TPT, T-TPT, P-MON-CKP e T-JUS-CKP é a variável a ser estimada pelo modelo. A metodologia utilizada para elaboração do modelo para ambos os poços é a mesma. Comparado aos hiperparâmetros utilizados no treinamento do poço 1, apenas o *delay* é distinto, passando a considerar o máximo 50, viabilizado pelo tamanho superior dos períodos de operação no poço 2.

A predição livre dos períodos de treinamento podem ser vistas na Figura 4.10, onde é possível visualizar que, considerando uma condição inicial de amostras em 250 minutos, o modelo é capaz de fornecer um resultado semelhante aos dados reais. O mesmo acontece sobre os dados de validação, da Figura 4.11.

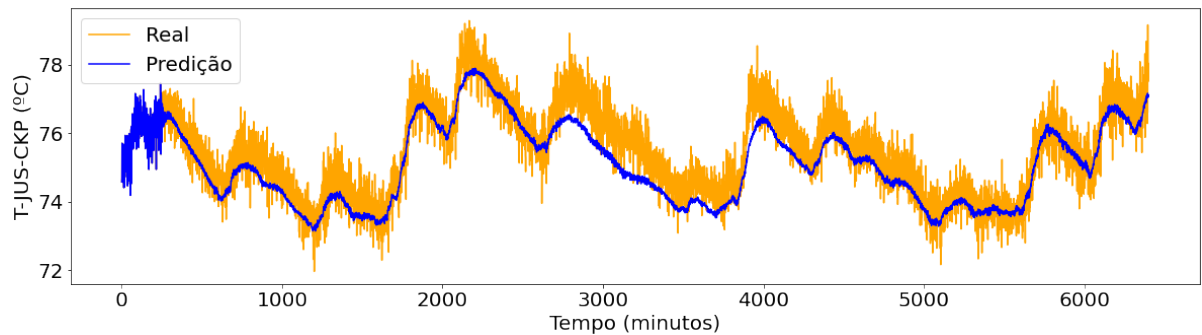
Figura 4.10 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de treinamento. Fonte: Própria.



(a) Período 3

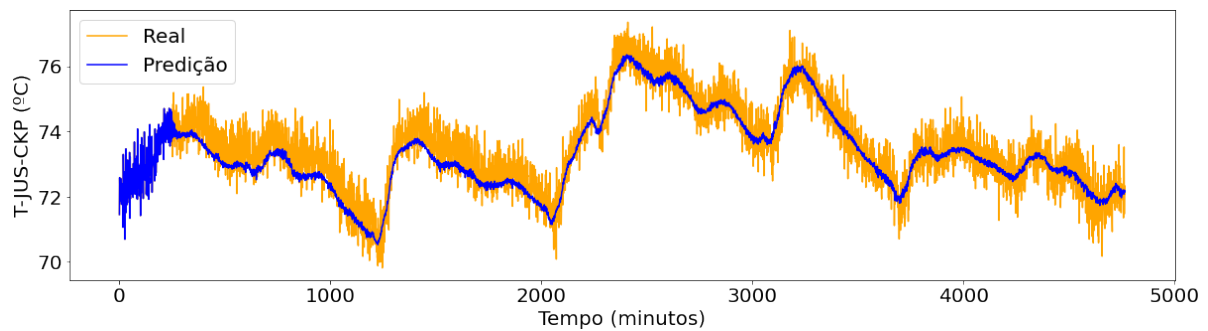


(b) Período 6 (50%)

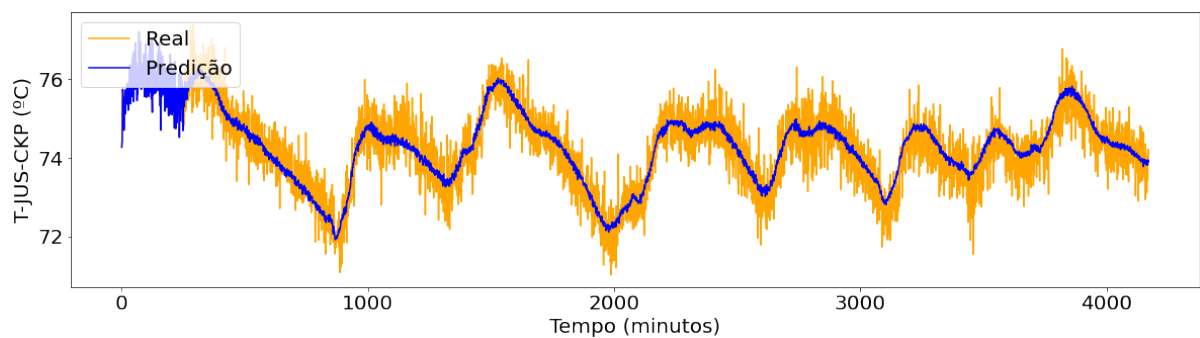


(c) Período 7

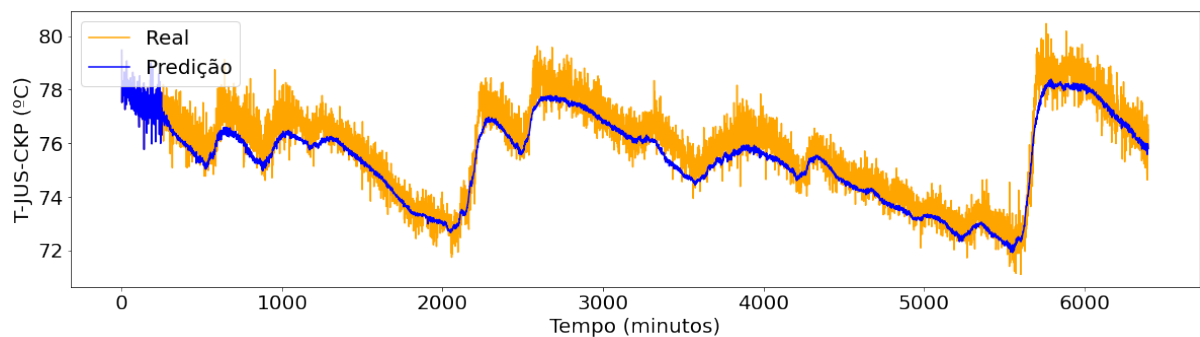
Figura 4.11 – Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de validação. Fonte: Própria.



(a) Período 4



(b) Período 5



(c) Período 6 (50%)

Na Tabela 4.5 é apresentado o erro percentual (MAPE) entre dados reais e estimados pelo modelo em simulação livre de cada período. Nota-se que, os períodos normais de treinamento e de validação, os erros não ultrapassam 0,69%, máximo semelhante ao resultado do poço 1 de 0,64%. O resultado do período com anomalia nitidamente se diferencia dos períodos normais, evidenciando a incapacidade do modelo, mesmo tratando-se de medições de um mesmo poço, de prever a temperatura de *choke* dada a dinâmica dos dados de entrada muito distinta do treinado.

As figuras em 4.12 ilustram o comportamento da previsão em janelas de 100 passos à frente em um determinado intervalo de tempo dos períodos normais de validação, em que

Tabela 4.5 – Resultados do modelos NARX obtido para predição de T-JUS-CKP.

| Período | MAPE [%] |
|----------------|-----------------|
| 3 ^t | 0,55 |
| 4 ^v | 0,55 |
| 5 ^v | 0,49 |
| 6 ^t | 0,48 |
| 6 ^v | 0,50 |
| 7 ^t | 0,69 |
| 8 ^a | 5,90 |

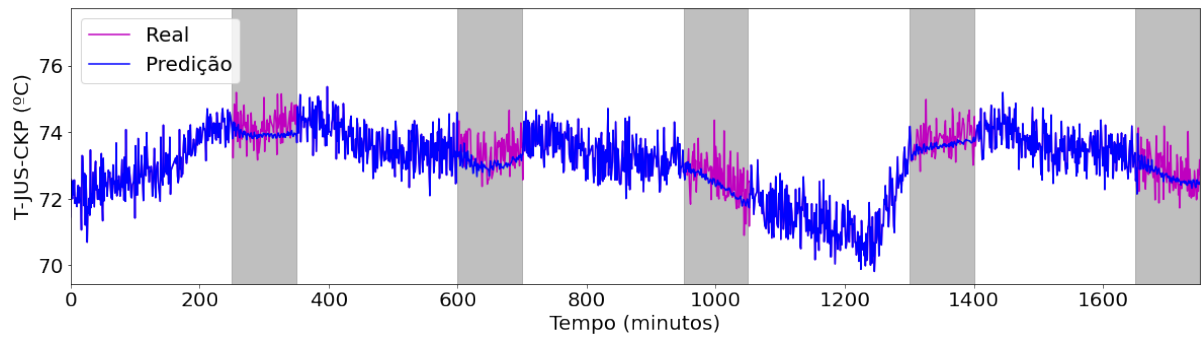
^tPeríodo de treinamento

^vPeríodo de validação

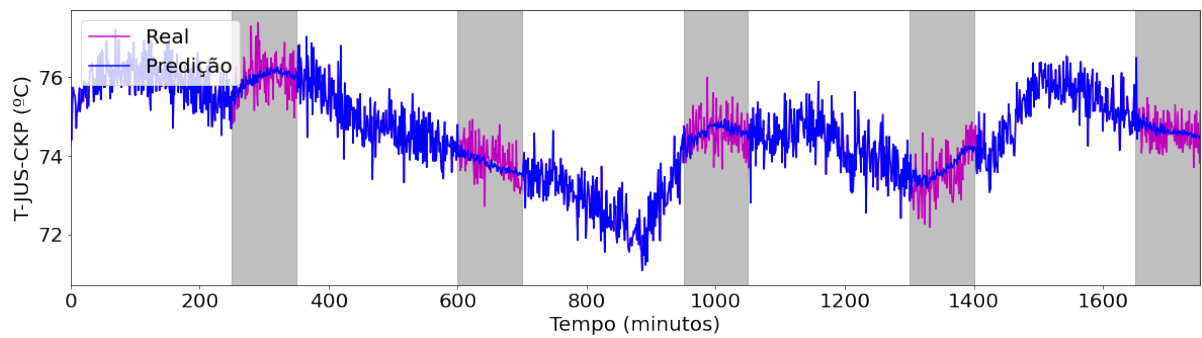
^aPeríodo com anomalia

as regiões destacadas indicam a janela de predição, as curvas de cor magenta representam o comportamento desejado e podem ser comparadas às curvas de predição em azul nesta janela. Assim como no poço 1, o comportamento nessas janelas apresentam um resultado satisfatório, em que a curva de predição das janelas, no geral, tende a ser próxima da média dos dados em cada janela e a acompanhar a dinâmica dos dados reais. Isso pode ser analisado pela Figura 4.13 que apresenta o MAPE de cada janela de predição. Nota-se que, na maioria dos casos, a predição possui MAPE inferior a 0,6%, com mediana dos períodos de validação de 0,48% e desvio padrão de 0,08%. Esse resultado se diferencia levemente do poço 1 que apresenta mediana e desvio padrão de 0,37% e 0,11%, respectivamente. Ainda assim, esses valores são satisfatórios e que dependem do tamanho da base de dados,

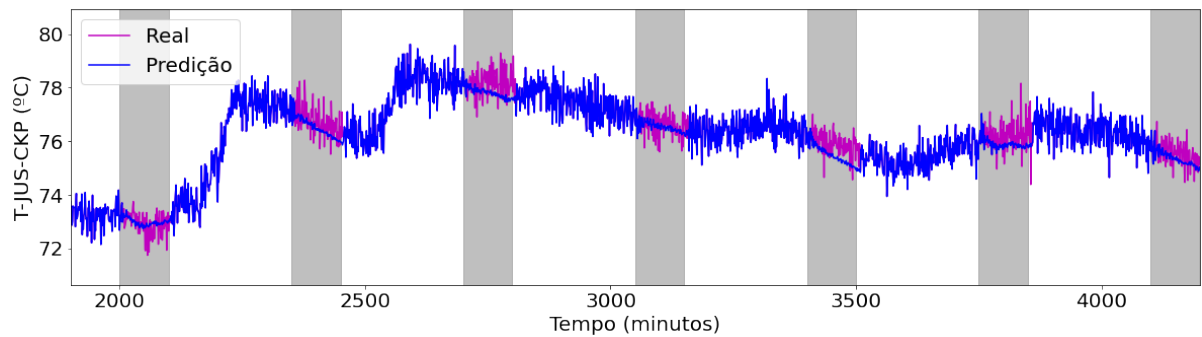
Figura 4.12 – Comparativo dos dados reais e estimados em simulação de predição de 100 passos à frente pelo modelo de três períodos de validação. Fonte: Própria.



(a) Período 4

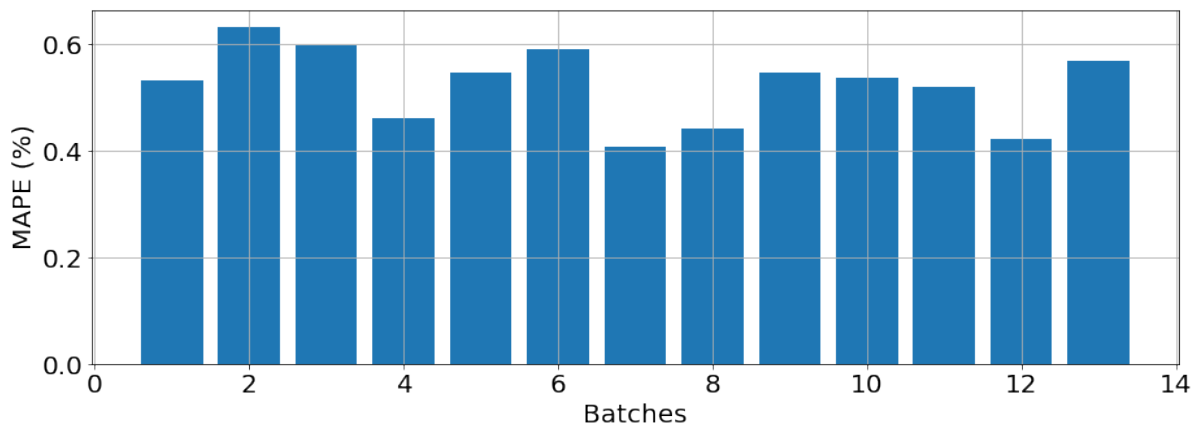


(b) Período 5

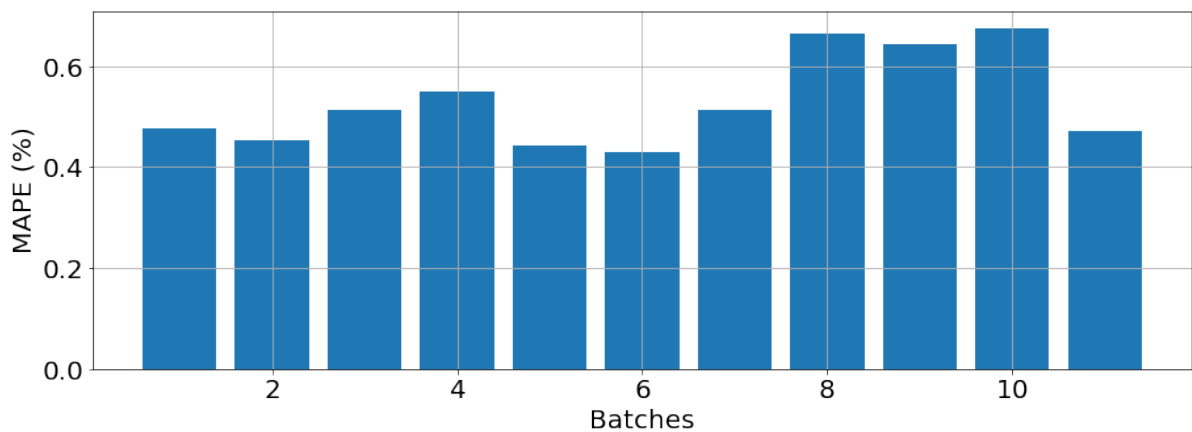


(c) Período 7

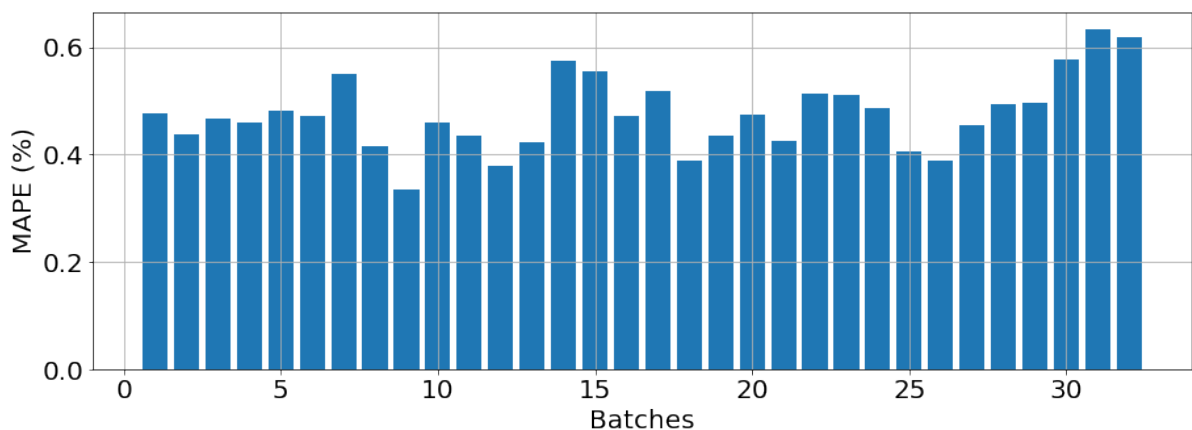
Figura 4.13 – Comparativo os erros percentuais das janelas em simulação de predição 100 passos à frente pelo modelo de três períodos de validação. Fonte: Própria.



(a) Período 4



(b) Período 5



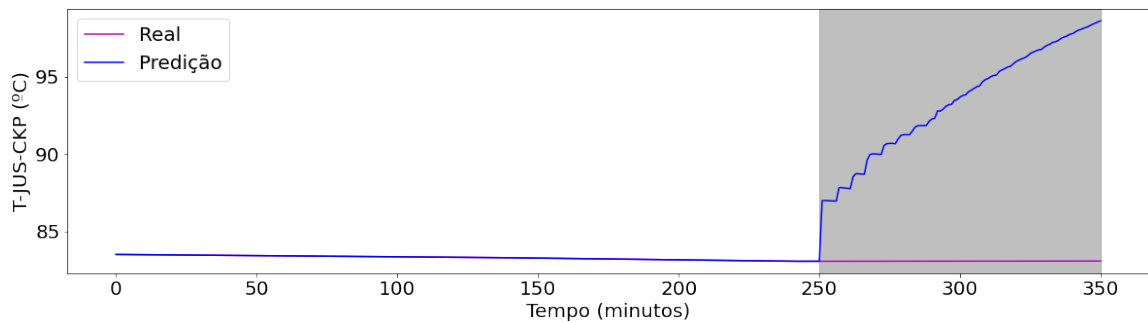
(c) Período 7

4.1.2.1 Anomalia 6 (Restrição Rápida em CKP)

O prognóstico deste tipo de anomalia é desejável para evitar perdas de produção já que, em geral, essa válvula é manual e fechamentos indesejados podem ser desfeitos mais

rapidamente (VARGAS, 2019b). O comportamento do *soft sensor* aplicado aos dados desse tipo de anomalia é o mesmo encontrado nas anomalias do poço 1, incapacidade de prever a temperatura de *choke* aproximada, já que o modelo se depara com uma dinâmica diferente das pressões e temperaturas de entrada. A predição dos primeiros 100 minutos (Figura 4.14) já apresenta um erro bem superior e um comportamento semelhante ao da anomalia 7 do poço 1 (Figura 4.6a). Tal semelhança pode estar associada às consequências de cada anomalia sobre as medições de temperaturas e pressões que são afetadas pela redução do fluxo na válvula de *choke* em ambos os casos. Porém, a consequência da restrição rápida é identificável em bem menos tempo, pois uma amplitude acima de um valor de referência e durante um período restrito de tempo já refletem em mudanças consideráveis nas medições.

Figura 4.14 – Simulação livre do modelo no período 10 de validação com anomalia. Fonte: Própria.



Os períodos 8 e 9, que também têm transição para esse tipo de anomalia, possuem menor tempo de medição dos dados, não alcançando o mínimo necessário (250 minutos) que sirva de condição inicial para predição e, por isso, não foi possível simular o comportamento do modelo sobre ambos os casos.

4.1.2.2 Representação do modelo

O treinamento do modelo resultou em uma função que melhor representa a dinâmica da temperatura justante à válvula *Choke*, dada às variáveis de pressão e temperaturas de entrada. O modelo encontrado é dado por:

$$y(k) = \theta_0 + \theta_1 y(k-11) + \theta_2 [y(k-6)y(k-16)] + \theta_3 y(k-41) + \theta_4 y(k-251) + \theta_5 u_2(k-1) + \theta_6 u_1(k-56) + \theta_7 [u_3(k-2)u_3(k-1)], \quad (4.3)$$

em que u_1 , u_2 e u_3 representam as entradas T-TPT, P-TPT e P-MON-CKP, respectivamente, e y representa a saída T-JUS-CKP, com atrasos; θ_n são os coeficientes de cada termo da função, apresentados no vetor Θ a seguir:

$$\Theta = [-3.46 \ 1.94 \times 10^{-1} \ 2.72 \times 10^{-3} \ 1,19 \times 10^{-1} \ 3,02 \times 10^{-2} \ 1,28 \times 10^{-6} \ 2,29 \times 10^{-1} \ -1,01 \times 10^{-13}]. \quad (4.4)$$

Ao contrário do que ocorre na função resultante do treinamento do poço 1, neste caso todas as variáveis possuem certa relevância para o modelo em tempos passados específicos. A variável de saída y , assim como no poço 1, é a que mais se repete, estando acompanhada de atrasos que variam de 6 a 251 passos atrás. Para as variáveis de pressões P-TPT e P-MON-CKP, representados por u_2 e u_3 , respectivamente, atrasos menores e mais próximos ao instante atual, são mais importantes para a predição dos dados futuros, utilizando atrasos iguais a 1 e 2. É possível notar também que os valores menores de Θ estão relacionados com dados de pressão, multiplicados na equação por u_2 e u_3 , e que geralmente são mais elevados. Além disso, este resultado mostra como foi explorado o atraso máximo de 250 em ao menos um dos termos da equação e a quantidade máxima de 7 termos, definidos inicialmente para treinamento.

4.2 Detecção de anomalia

A predição de uma variável de processo pode ser um caminho para a detecção de anomalias no contexto de poços surgentes de petróleo. Baseado nos resultados da predição dos modelos finais de cada poço apresentado nesta seção, é possível identificar um padrão semelhante de baixo erro nas predições das janelas temporais de períodos normais de treinamento e validação. Esse erro tende a aumentar significativamente e apresentar desempenho distinto quando o modelo é aplicado sobre dados com presença de transitório de anomalia, um indicativo de seu potencial uso para detecção de comportamentos que possam afetar a operação nos poços.

No entanto, no início de uma transição de falha, para uma anomalia mais lenta, o erro percentual do modelo é baixo (como apresentado nas Figuras 4.7 e 4.8), o que faz com que este cenário seja desafiador.

A métrica utilizada para avaliação do *soft sensor*, MAPE, em alguns tipos de casos, sinaliza que há algum comportamento diferente, por meio do erro percentual entre valor estimado e real da janela de predição, porém, podem existir limitações em períodos ruidosos. Períodos

com mais ruído, como os períodos 2 e 8 do poço 1, resultam em erro de predição maior pois o modelo, ao não explicar o ruído, possui um erro absoluto maior. Nesse cenário, a métrica indica um comportamento indesejado para esses períodos e, por essa razão, pode não ser a ideal para apoio à detecção de anomalias.

Outras possibilidades a serem exploradas é o uso da variância e do erro médio das janelas de predição para critério de detecção de anomalias, pois seriam benéficos em situações onde os dados possuem muito ruído de medição. Ambos combinados podem suprir as limitações que o MAPE tem sob o aspecto de detecção e poderão ser explorados mais a fundo em trabalhos futuros. Da mesma forma, outros testes estatísticos podem ser realizados a partir de amostras medidas e preditas dentro de uma janela.

5 CONSIDERAÇÕES FINAIS E CONCLUSÃO

Por meio deste trabalho foi possível compreender o comportamento do MGGP a partir de variações em seus hiper-parâmetros, o que conferiu a escolha dos melhores valores para a realização do treinamento do modelo. Primeiramente, avaliar a distribuição dos dados para separação de treinamento e validação é crucial para que o modelo alcance bom desempenho já que a escolha dos parâmetros depende do tamanho e comportamento dos dados selecionados. Nos casos avaliados, aumentar o valor máximo de passos à frente e atraso máximo tornam o modelo mais estável e reduzem o erro final da estimativa. O aumento da quantidade máxima de termos, superior a 20 nesse caso, não trouxe ganho considerável. O aumento da população inicial do MGGP agrega um custo computacional e tempo de processamento bem superiores, e que não resultaram em melhora significativa do modelo. A partir dessa análise, o desenvolvimento do MGGP para elaboração de um *soft sensor* dos poços 1 e 2 foi iniciada apresentando resultados satisfatórios.

O modelo NARX utilizando MGGP resultou em uma estimativa da temperatura jusante à válvula de *Choke* capaz de prever a dinâmica da variável a partir de outras variáveis de sensores de pressão e temperatura disponíveis no processo. O *soft sensor* desenvolvido apresentou um resultado satisfatório sobre os dados reais de operação de extração de petróleo em poço surgente para os dois poços analisados. As simulações livres e predições por janela temporal para os períodos de operação normal resultaram em erros percentuais absoluto médio inferiores a 1% e com capacidade de acompanhar o comportamento dos dados reais apesar da individualidade de cada período, do ponto de operação e presença de ruído nos dados. As funções que representam os modelos de predição evidenciam como a própria variável de saída, a temperatura justante à válvula de *choke*, em momentos passados é de grande importância para ambos os poços enquanto a variável de entrada referente a temperatura do fluido no TPT tem pouca interferência na predição do modelo, chegando a não ser utilizada no modelo obtido para o poço 1.

Quando observado o desempenho dos *soft sensors* sobre os períodos com transitórios de anomalia, notou-se o potencial de seu uso no cenário de detecção de anomalias em poços surgentes, em razão do erro de predição elevado neste tipo de período. Porém, a métrica utilizada para avaliação sob o aspecto de detecção apresentou limitações e, por isso, avaliar melhor a distribuição dos dados medidos e preditos dentro de uma janela pode ser uma boa alternativa para fornecer uma análise mais adequada do modelo para esse fim.

Os resultados alcançados podem ser base para trabalhos futuros em que se objetiva detectar anomalias na operação dos poços de petróleo, dando abertura para uma avaliação aprofundada do *soft sensor* nesse propósito bem como a comparação destes resultados com os fornecidos por outras técnicas de identificação de sistemas e de reconhecimento de padrões aplicadas à base de dados 3W.

REFERÊNCIAS

- AGUIRRE, L. A. et al. Development of soft sensors for permanent downhole gauges in deepwater oil wells. **Control Engineering Practice**, v. 65, p. 83–99, 2017. ISSN 0967-0661. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0967066117301284>>.
- ALJUBRAN, M. et al. Deep learning and time-series analysis for the early detection of lost circulation incidents during drilling operations. **IEEE Access**, v. 9, p. 76833–76846, 2021.
- ANDREOLLI, I. **Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo**. Rio de Janeiro, RJ, Brasil: Editora Interciência, 2016. 650 p.
- AZHDARI, M.; MEHRANBOD, N. Application of bayesian belief networks to fault detection and diagnosis of industrial processes. In: **2010 International Conference on Chemistry and Chemical Engineering**. [S.l.: s.n.], 2010. p. 92–96.
- BARBARIOL, T.; FELTRESI, E.; SUSTO, G. Self-diagnosis of multiphase flow meters through machine learning-based anomaly detection. **Energies**, v. 13, p. 3136, 06 2020.
- BARBOSA, B. H. G. et al. Black and gray-box identification of a hydraulic pumping system. **Control Systems Technology, IEEE Transactions on**, v. 19, n. 2, p. 398–406, March 2011.
- BARBOSA, B. H. G. et al. Black and gray-box identification of a hydraulic pumping system. **IEEE Transactions on Control Systems Technology**, v. 19, n. 2, p. 398–406, 2011.
- CAPRIGLIONE, D. et al. Online fault detection of rear stroke suspension sensor in motorcycle. **IEEE Transactions on Instrumentation and Measurement**, v. 68, n. 5, p. 1362–1372, 2019.
- CAPRIGLIONE, D. et al. Soft sensors for instrument fault accommodation in semiactive motorcycle suspension systems. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 5, p. 2367–2376, 2020.
- CASTRO, H. C. de; BARBOSA, B. H. G. A python library for nonlinear system identification using multi-gene genetic programming algorithm. **arXiv**, v. 2211.05723, 2022.
- CASTRO, H. Carvalho de. **A modified MGGP algorithm for structure selection of NARMAX models**. Dissertação (Mestrado) — Universidade Federal de Lavras, Programa de Pós-graduação Em Engenharia de Sistemas e Automação, Lavras - Minas Gerais, 2021.
- CHEN, D.; LIU, Y.; ZHOU, J. Optimized neural network by genetic algorithm and its application in fault diagnosis of three-level inverter. In: **2019 CAA Symposium on Fault Detection, Supervision and Safety for Technical Processes (SAFEPROCESS)**. [S.l.: s.n.], 2019. p. 116–120.
- CHITRA, M.; PAPPA, N. Estimation of fungal biomass using multiphase artificial neural network based dynamic soft sensor. **Journal of Microbiological Methods**, v. 159, 02 2019.
- COELHO, L. dos S.; BORA, T. C.; KLEIN, C. E. A genetic programming approach based on lévy flight applied to nonlinear identification of a poppet valve. **Applied Mathematical Modelling**, v. 38, n. 5, p. 1729–1736, 2014. ISSN 0307-904X. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0307904X1300591X>>.

COLOMBO, D. et al. Regression-based finite element machines for reliability modeling of downhole safety valves. **Reliability Engineering System Safety**, v. 198, p. 106894, 2020. ISSN 0951-8320. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0951832018310287>>.

CRUZ, S. Freire Rigatto da. **Modelagem matemática e simulação numérica da precipitação de carbonato de cálcio em condições de poço**. Dissertação (Mestrado) — Universidade Federal do Espírito Santo, Centro Universitário Norte do Espírito Santo, Programa de Pós-graduação Em Energia, São Mateus - Espírito Santo, 2018.

de Moraes, G. A. et al. Soft sensors design in a petrochemical process using an evolutionary algorithm. **Measurement**, v. 148, p. 106920, 2019. ISSN 0263-2241. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0263224119307778>>.

DHIMISH, M. et al. Comparing mamdani sugeno fuzzy logic and rbf ann network for pv fault detection. **Renewable Energy**, v. 117, p. 257–274, 2018. ISSN 0960-1481. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0960148117310273>>.

DU, Z. et al. Sensor fault detection and its efficiency analysis in air handling unit using the combined neural networks. **Energy and Buildings**, v. 72, p. 157–166, 2014. ISSN 0378-7788. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0378778813008554>>.

FAN, Y. et al. A data-driven soft sensor based on multilayer perceptron neural network with a double lasso approach. **IEEE Transactions on Instrumentation and Measurement**, v. 69, n. 7, p. 3972–3979, 2020.

FONTES, D.; VASCONCELOS, L.; BRITO, R. Blast furnace hot metal temperature and silicon content prediction using soft sensor based on fuzzy c-means and exogenous nonlinear autoregressive models. **Computers Chemical Engineering**, v. 141, p. 107028, 07 2020.

FREITAS, L.; BARBOSA, B. H. G.; AGUIRRE, L. A. Including steady-state information in nonlinear models: An application to the development of soft-sensors. **Engineering Applications of Artificial Intelligence**, v. 102, p. 104253, 2021. ISSN 0952-1976. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0952197621001007>>.

GAO, Z.; CECATI, C.; DING, S. X. A survey of fault diagnosis and fault-tolerant techniques—part ii: Fault diagnosis with knowledge-based and hybrid/active approaches. **IEEE Transactions on Industrial Electronics**, v. 62, n. 6, p. 3768–3774, 2015.

GRAZIANI, S.; XIBILIA, M. G. **Development and Analysis of Deep Learning Architectures (Deep Learning for Soft Sensor Design)**. [S.l.]: Cham, Switzerland: Springer, 2020.

HE, S. et al. Detection and fault diagnosis of power transmission line in infrared image. In: **2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)**. [S.l.: s.n.], 2015. p. 431–435.

HENMI, T. et al. Early detection of plant faults by using machine learning. In: **2016 International Conference on Advanced Mechatronic Systems (ICAMechS)**. [S.l.: s.n.], 2016. p. 201–205.

- HINCHLIFFE, M. et al. Modelling chemical process systems using a multi-gene genetic programming algorithm. In: KOZA, J. R. (Ed.). **Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28-31, 1996**. Stanford University, CA, USA: Stanford Bookstore, 1996. p. 56–65. ISBN 0-18-201031-7.
- HUANG, R.; LI, Z.; CAO, B. A soft sensor approach based on an echo state network optimized by improved genetic algorithm. **Sensors (Basel, Switzerland)**, v. 20, 09 2020.
- HÜFFNER, L. N.; TRIERWEILER, J. O.; FARENZENA, M. Are complex black-box models for permanent downhole gauge pressure estimation necessary? **Journal of Petroleum Science and Engineering**, v. 173, p. 715–732, 2019. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S092041051830915X>>.
- JIANG, Y. et al. A review on soft sensors for monitoring, control, and optimization of industrial processes. **IEEE Sensors Journal**, v. 21, n. 11, p. 12868–12881, 2021.
- JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. ISSN 0036-8075. Disponível em: <<https://science.sciencemag.org/content/349/6245/255>>.
- JUNIOR, W. R. L.; MARTINS, S. A. M.; NEPOMUCENO, E. G. **Meta-Model Structure Selection: Building Polynomial NARX Model for Regression and Classification**. 2021.
- JÄMSÄ-JOUNELA, P. S.-L. Future trends in process automation. **IFAC Proceedings Volumes**, v. 40, n. 1, p. 1–10, 2007. ISSN 1474-6670. 8th IFAC Symposium on Cost Oriented Automation. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S147466701531199X>>.
- JÚNIOR, W. F. **Detecção de anomalias em poços produtores de petróleo usando aprendizado de máquina**. Dissertação (Mestrado) — Instituto Federal do Espírito Santo, Programa de Pós-graduação Em Computação Aplicada, Serra - Espírito Santo, 2022.
- KANEKO, H.; FUNATSU, K. A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method (vol 58, pg 1829, 2012). **AIChE Journal**, v. 58, 06 2012.
- KANEKO, H.; FUNATSU, K. Preparation of comprehensive data from huge data sets for predictive soft sensors. **Chemometrics and Intelligent Laboratory Systems**, v. 153, p. 75–81, 2016. ISSN 0169-7439. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0169743916300338>>.
- LI, D.; HU, G.; SPANOS, C. A data-driven strategy for detection and diagnosis of building chiller faults using linear discriminant analysis. **Energy and Buildings**, v. 128, 07 2016.
- LI, G.; HU, Y. Improved sensor fault detection, diagnosis and estimation for screw chillers using density-based clustering and principal component analysis. **Energy and Buildings**, v. 173, p. 502–515, 2018. ISSN 0378-7788. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0378778817337362>>.
- LIU, Y. et al. Semi-supervised failure prediction for oil production wells. In: **2011 IEEE 11th International Conference on Data Mining Workshops**. [S.l.: s.n.], 2011. p. 434–441.

LIU, Z. et al. Risk assessment on deepwater drilling well control based on dynamic bayesian network. **Process Safety and Environmental Protection**, v. 149, p. 643–654, 2021. ISSN 0957-5820. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0957582021001385>>.

LOTUFO, F.; GARCIA, C. Sensores virtuais ou soft sensors: Uma introdução. 01 2008.

MARINS, M. A. et al. Fault detection and classification in oil wells and production/service lines using random forest. **Journal of Petroleum Science and Engineering**, v. 197, p. 107879, 2021. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0920410520309372>>.

MAYNARD, J. et al. Aplicações de programação genética em reservatórios de petróleo. In: . [S.l.: s.n.], 2015. p. 1–6.

MENDES, J. et al. Genetic fuzzy system for data-driven soft sensors design. **Applied Soft Computing**, v. 12, p. 3237–3245, 10 2012.

MOTA, F. et al. Identificação de um sistema de bombeamento hidráulico com algoritmo evolucionário multi-objetivo. **Congresso Brasileiro de Automática, Anais da Sociedade Brasileira de Automática**, v. 2, n. 1, p. 1–8, 2020.

NASCIMENTO, R. S. F. et al. Detecção de anomalias em poços de petróleo surgentes com stacked autoencoders. In: **Simpósio Brasileiro de Automação Inteligente**. [S.l.: s.n.], 2021. v. 1, n. 1, p. 2049–2056.

NASCIMENTO, R. Scoralick Fontoura do. **Detecção de anomalias em poços de produção de petróleo offshore com a utilização de autoencoders e técnicas de reconhecimento de padrões**. Dissertação (Mestrado) — Universidade Federal de Lavras, Departamento de Engenharia, Programa de pós-graduação em engenharia de sistemas e automação, Lavras, 2021.

NIAZKAR, M.; NIAZKAR, H. R. Covid-19 outbreak: Application of multi-gene genetic programming to country-based prediction models. **Electronic Journal of General Medicine**, v. 17, p. em247, 03 2020.

NYBØ, R. Fault detection and other time series opportunities in the petroleum industry. **Neurocomputing**, v. 73, n. 10, p. 1987–1992, 2010. ISSN 0925-2312. Subspace Learning / Selected papers from the European Symposium on Time Series Prediction. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0925231210001049>>.

OSSAI, C. I. Modified spatio-temporal neural networks for failure risk prognosis and status forecasting of permanent downhole pressure gauge. **Journal of Petroleum Science and Engineering**, v. 184, p. 106496, 2020. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0920410519309179>>.

PORTELA, G. **Gerenciamento de Riscos na Indústria de Petróleo e Gás: offshore e onshore**. Rio de Janeiro, RJ, Brasil: Elsevier Editora Ltda, 2015. 877 p.

RANDEK, J.; MANDENIUS, C.-F. On-line soft sensing in upstream bioprocessing. **Critical reviews in biotechnology**, v. 38, p. 1–16, 04 2017.

RECKNAGEL, F. Applications of machine learning to ecological modeling. **Ecological Modelling**, v. 146, p. 303–310, 12 2001.

RUSSELL, E. L.; CHIANG, L. H.; BRAATZ, R. D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 51, n. 1, p. 81–93, 2000. ISSN 0169-7439. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0169743900000587>>.

S., V. V.; MOHANTA, H. K.; PANI, A. K. Adaptive non-linear soft sensor for quality monitoring in refineries using just-in-time learning—generalized regression neural network approach. **Applied Soft Computing**, v. 119, p. 108546, 2022. ISSN 1568-4946. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1568494622000758>>.

SAAVEDRA, M. C. M. d. A. D. A. L. G. **Avaliação do efeito da incerteza do BSW na vazão de óleo em sistemas de elevação e escoamento de petróleo**. 63 p. Monografia (Graduação) — Escola Politécnica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

SHANG, C. et al. Data-driven soft sensor development based on deep learning technique. **Journal of Process Control**, v. 24, p. 223–233, 03 2014.

SHARMA, S.; TAMBE, S. Soft-sensor development for biochemical systems using genetic programming. **Biochemical Engineering Journal**, v. 85, 04 2014.

SIMOES, M. G. et al. A novel competitive learning neural network based acoustic transmission system for oil-well monitoring. **IEEE Transactions on Industry Applications**, v. 36, n. 2, p. 484–491, 2000.

TAQVI, S. A. A. et al. Fault detection in distillation column using narx neural network. **Neural Computing and Applications**, v. 32, 04 2020.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition, Fourth Edition**. [S.l.]: Academic Press, 2009. Hardcover. ISBN 9781597492720.

THOMAS, J. E. **Fundamentos de Engenharia de Petróleo**. Rio de Janeiro, RJ, Brasil: Editora Interciência, 2004. 288 p.

VARGAS, R. **The first realistic and public dataset with rare undesirable real events in oil wells**. 2019. Disponível em: <https://github.com/ricardovargas/3w_dataset>.

VARGAS, R. E. V. **Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo**. Tese (Doutorado) — Universidade Federal do Espírito Santo, Centro Tecnológico, Programa de pós-graduação em engenharia elétrica, Vitória, 2019.

VARGAS, R. E. V. et al. A realistic and public dataset with rare undesirable real events in oil wells. **Journal of Petroleum Science and Engineering**, v. 181, p. 106223, 2019. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0920410519306357>>.

VERÇOSA, I. J. F. L. et al. Formação de hidratos em perfurações de poços em águas profundas e ultra profundas. **III Congresso Nacional de Engenharia de Petróleo, Gás Natural e Biocombustíveis**, 2018.

VIANA, M. C. et al. Modelagem e simulação da incrustação de carbonato de cálcio em condições de poço. **VI Encontro Nacional de Hidráulica de Poços de Petróleo e Gás**, 2015.

ÖZMEN, et al. Prediction of slipper pressure distribution and leakage behaviour in axial piston pumps using ann and mggp. **Mathematical Problems in Engineering**, v. 2019, p. 1–13, 03 2019.