Desenvolvimento de *soft sensor* para poços de petróleo usando MGGP

Bruna Cunha
Departamento de Engenharia
Universidade Federal de Lavras (UFLA)
Lavras, MG, Brazil
bruna.cunha@estudante.ufla.br

Danton Diego Ferreira
Departamento de Engenharia
Universidade Federal de Lavras (UFLA)
Lavras, MG, Brazil
danton@ufla.br

Bruno H.G. Barbosa
Departamento de Engenharia
Universidade Federal de Lavras (UFLA)
Lavras, MG, Brazil
brunohb@ufla.br

***Abstract*—Técnicas de inteligência artificial aplicadas em contextos reais como, por exemplo, previsão antecipada de comportamento e detecção de anomalias, podem implicar em redução de custo de manutenção, ações de prevenção de acidentes e falhas, auxílio em tomada de decisão, redução de perdas de produção e financeira, além de 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 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 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 com medições de sensores presentes em poços surgentes de petróleo, entre os períodos de 2017 e 2018. É realizado o treinamento do modelo e validação dos resultados que se mostram satisfatórios, resultando em uma função capaz de representar a dinâmica de operação de um poço surgente de petróleo em operação normal, contribuindo para a implementação de um sistema de detecção de falhas no poço.**

***Index Terms*—*soft sensor*, poços surgentes de petróleo, programação genética, identificação de sistemas**

I. Introdução

Há anos, monitorar processos é uma atividade realizada 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 [1]–[3].

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 chegou a causar perdas anuais de 20 bilhões de dólares na indústria petroquímica dos EUA [4]. 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 [5]. 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 [6].

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 vem 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 [7] [8]. Trata-se 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 [9]. 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) [2].

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 [10]. 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 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 [11]. 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.

II. 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 [12]. Em seguida, é realizada a perfuração das rochas e criação dos poços por onde escoará o fluído composto por óleo, gás, água e eventualmente sedimentos [6] ou injetará gases para elevação artificial.

A elevação do fluído pelos poços de produção depende da pressão que o fluído 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 [11]. 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 fluídos 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 fluídos dentro do reservatório é de grande importância para a otimização da produção [12]. Elevação e escoamento eficientes significam maior vazão de óleo e gás, além de menor gasto de energia e recursos [6].

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 (do inglês, *pressure dawnhole gauge*) que pode chegar a mais de quatro quilômetros de profundidade o que evidencia a sua manutenção não ser economicamente viável, envolvendo ambiente de alto risco [7]. A figura 1 apresenta alguns componentes comuns na operação de poços de elevação natural, sendo algumas variáveis de processo listadas a seguir:

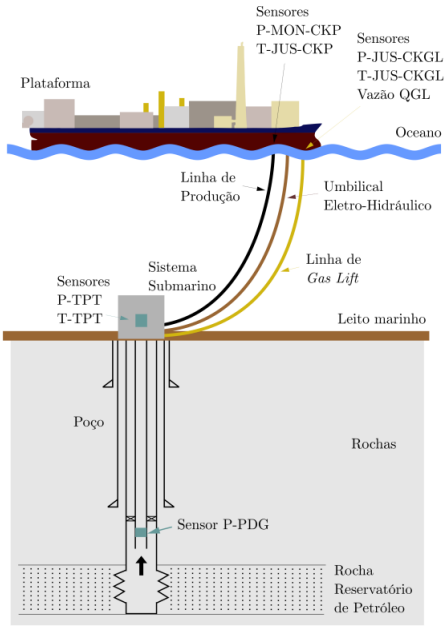


Fig. 1: Localização dos sensores de fundo e na superfície. Fonte: [13]

- 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 [14]. 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, identificação de modelo e, devido à sua localização em condições perigosas, pode ocorrer falha ou imprecisões na medição [15].
- 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 [11].
- Pressão do fluido montante à válvula Choke de Produção (P-MON-CKP): Sensor considerado confiável quando presente [11] e de fácil manutenção pela localização na plataforma.
- Temperatura do fluido jusante à válvula Choke de

Produção (T-JUS-CKP).

A. Tipos de anomalias

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. Não se tem muitas informações e dados sobre anomalias neste contexto, muitas vezes este monitoramento de poços é feita por profissionais que analisam janelas temporais com tamanhos diferentes a fim de compreender a dinâmica de cada tipo de anomalia, por isso, treinar qualquer modelo nessa condição é desafiador.

Alguns tipos de anomalias que ocorrem prioritariamente em poços produtores marítimos operados via Elevação Natural [11]:

- **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 [16]. 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 [17]. 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 diminuam, já as temperaturas costumam aumentar em todos os equipamentos [11].
- **Incrustação em CKP:** A incrustação ocorre a partir do processo de aglomeração de partículas que reduz ou obstrui o espaço de passagem do fluído. A presença de incrustações está quase sempre associada à perda de produção, a situações de intervenção em poços e consequentemente à redução de lucros [18].

Problemas que abordam o tema de detecção de anomalias em diversos contextos podem ser tratados com técnicas de reconhecimento de padrões e identificação de sistemas. Por meio de técnicas computacionais é possível encontrar uma maneira eficiente de, a partir das características extraídas de um conjunto de dados, organizar padrões em categorias ou classes que compartilham determinadas semelhanças ou realizar a predição de comportamentos futuros. Dessa forma, a seção III apresenta possibilidades para uso de *soft sensors* no contexto de poços surgentes de petróleo.

III. Soft sensors

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 [19].

[20] 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. [21] 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, [22] 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 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 autogressive with exougenous input*), ARMAX (*Autoregressive moving average with exogenous input*) e NARMAX (*Nonlinear autoregressive moving average with exogenous input*)

A. Nonlinear Autoregressive 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 [23]. O modelo NARX pode ser representado pela seguinte equação:

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

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, [24] e [23] abordam o problema de seleção de estrutura de modelos NARX e NARMAX polinomiais por meio da programação genética.

Já [25] 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 [9], [26], [27], enquanto em [3] utiliza-se no cenário de detecção de anomalias em poços de petróleo.

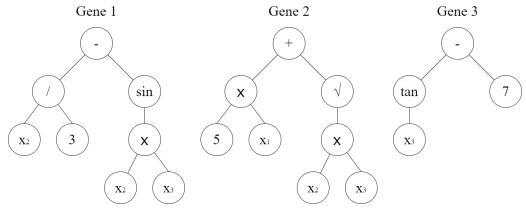


Fig. 2: Exemplo simbólico de um modelo genético multigênico

Os algoritmos genéticos, que simulam a evolução das espécies, inspirados do 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 [28]–[31]. Além disso, outras técnicas para aplicação no contexto de *soft sensors* podem ser exploradas, tais como lógica *fuzzy*, também conhecida por lógica nebulosa [32] [33], e métodos estatísticos multidimensionais.

IV. 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 estruturas 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 [34]. Como exemplo, a Figura 2 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: $y = w_0 + w_1(x_2/3 - \sin(x_2 x_3)) + w_2(5x_1 + \sqrt{x_2 x_3}) + w_3(\tan(x_3) - 7)$

O MGGP teve sua primeira aparição em 1996 por Hinchliffe [35]. 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 [36], 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 [37] cujo objetivo foi alcançar um modelo capaz de predizer surtos de COVID-19 em alguns países. Uma vantagem mencionada neste 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 [38], 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.

Metodologia

A. Base de Dados

As amostras reais presentes na base de dados foram coletadas 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 [39] e possui observações de 5 variáveis de processo comuns em poços de elevação natural, sendo elas: medições de pressão do fluído no PDG (*Permanent Downhole Gauge*), pressão e temperatura do fluído no TPT (*Temperature and Pressure Transducer*), pressão do fluído montante à válvula de *Choke* de produção e temperatura do fluído jusante à válvula de *Choke* de produção.

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 presentes na configuração do poço e/ou apresentar falhas. As medições reais foram obtidas durante o período do ano de 2012 a meados de 2018.

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 do poço identificado:

- Baixo percentual de dados faltantes;
- Presença de períodos com medições classificadas unicamente como normais;
- Presença de períodos com ao menos dois tipos de anomalia;
- Medições das mesmas variáveis de processo entre os períodos;

Dado tais critérios, o pré-processamento na base de dados inclui a remoção de variáveis com percentuais de medições faltantes superior a 10%, a fim de evitar preenchimento de medições que podem intervir erroneamente no treinamento do modelo, a consideração da amostragem (re-amostragem) por minuto, originalmente em segundos, uma vez que não apresentam alterações relevantes nas medições em questão de segundos, e a separação da base por períodos de medição, devido aos intervalos de tempo sem medições entre cada período. 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 Abrupto de BSW) e 7 (Incrustação em CKP) por serem as únicas presentes no poço trabalhado. Os períodos normais do poço foram separados e enumerados de 1 a 8, ordenamos por data da operação, e o critério dessa separação foi considerar períodos completos e sem interrupções nas medições. Foi feita a separação dos dados para treinamento do modelo considerando cinco dos oito períodos normais do poço, os quais apresentam maiores diferenças na dinâmica das variáveis, com o intuito de abranger o treinamento a diferentes comportamentos de um mesmo poço. Os três períodos normais remanescentes juntamente com dois períodos com anomalia foram utilizados para a validação do modelo de *soft sensor*.

B. Algoritmo MGGP

O algoritmo MGGP, apresentado na seção IV, é utilizado neste trabalho por meio da toolbox desenvolvida em [23], [40], 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.

A definição de alguns parâmetros utilizados no modelo do *soft sensor* foi realizada com base na análise das variações de seus valores aplicados aos dados do poço. Foram avaliados os parâmetros referentes à quantidade de passos a frente para predição do erro, quantidade máxima de termos e atraso máximo a fim de compreender os impactos 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 na geração aleatória na população inicial no modelo. 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 I. Os demais parâmetros do MGGP foram mantidos conforme definido em [40], por não apresentarem variações em relação a outros trabalhos que utilizam o MGGP, e não apresentarem impactos significativos no modelo em questão.

TABLE I: MAPE das variações dos parâmetros do MGGP aplicado ao poço na predição de temperatura de *choke*

Parâmetro	**Variação**	**MAPE [%]**		
		Período 5	***Período 6***	***Período 8***
Passos a frente	1	0,78	1,02	0,65
	5	0,49	0,85	0,53
	50	0,24	0,57	0,45
Quantidade máxima de termos	5	0,36	0,70	0,65
	10	0,33	0,68	0,62
	20	0,24	0,57	0,46
	30	0,27	0,61	0,50
Atraso máximo	1	0,63	0,68	0,47
	10	0,51	1,13	1,00
	100	0,31	0,56	0,81

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

TABLE II: Parâmetros do MGGP

Parâmetro	**Valor**
Tamanho da população	200
Número de gerações	50
Profundidade máxima da árvore	7
Taxa de mutação	0,1
Taxa de cruzamento	0,9

C. Medidas de Desempenho

A avaliação de desempenho do *soft sensor* utilizando o algoritmo MGGP, assim como em [8] e [9], foi definida por meio do MAPE, que fornece a informação de adequação (aptidão) do modelo por meio do erro relativo entre os dados reais e a saída do modelo expresso em percentual:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y_i}}{\hat{y_i}}\right| \tag{2}$$

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.

RESULTADOS E DISCUSSÃO

Para a 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 [40]. Como mostrado por [23], 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 utilizando os períodos 1, 2, 3, 4 e 7 de operação normal do poço, sem ocorrência de

TABLE III: Resultados do modelos NARX obtido para predição de T-JUS-CKP.

Período	MAPE [%]
1	0,50
2	0,21
3	0,16
4	0,28
5	0,21
6	0,49
7	0,35
8	0,39
9[a]	4,33
10[a]	0,70

[a]Período com transição para estado de anomalia

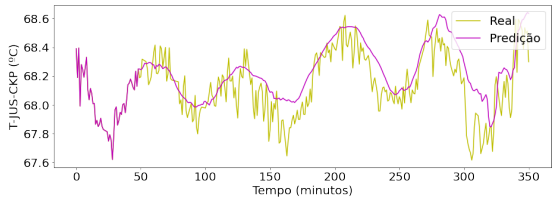
falhas, e as validações do modelo resultante foram feitas com os demais períodos disponíveis do poço (Períodos 5, 6 e 8), também sem ocorrência de anomalias. Os períodos 9 e 10, que apresentam transição para o estado de anomalia, foram utilizadas para algumas observações de desempenho do *soft sensor* dado um comportamento inesperado das variáveis. As variáveis de entrada são P-TPT, T-TPT, P-MON-CKP e T-JUS-CKP é a variável a ser estimada pelo modelo.

Na Tabela I, 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 [2], [7], [9]. A quantidade máxima de termos delimita a flexibilidade do modelo e foi observado que 20 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.

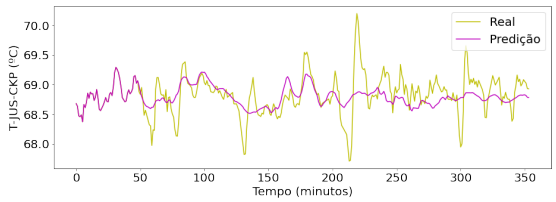
Na Tabela III é apresentado o erro percentual (MAPE) entre dados reais e estimados pelo modelo para cada período. Nota-se que, tanto períodos normais de treinamento quanto períodos de validação, os erros não ultrapassam 0,50%, variando de acordo com a dinâmica individual de cada período que pode apresentar ruído e interferências do ambiente.

As Figuras 3a, 3b e 3c mostram a predição resultante do modelo para os períodos de treinamento 3, 4 e 7, respectivamente. O resultado são curvas reais e de predição semelhantes apesar de pouco transitório e presença de ruído nos dados.

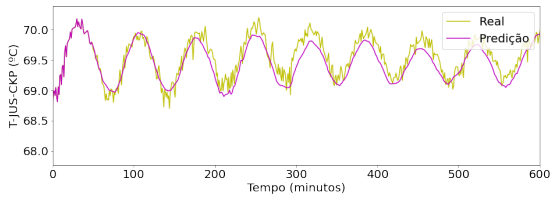
Tais resultados mostram que o modelo é capaz de capturar a relação dos dados medidos e a estimativa futura. Isso pode ser visto também na Figura 4, referente a um dos períodos de validação do MGGP. É possível visualizar como as temperaturas estimadas acompanham a curva de temperaturas reais do poço. 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.



(a) Período 3



(b) Período 4



(c) Período 7

Fig. 3: Comparativo dos dados reais e estimados em simulação livre pelo modelo de três períodos de treinamento. Fonte: Própria.

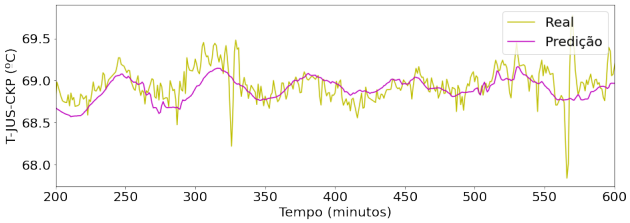


Fig. 4: Comparativo dos dados reais e estimados pelo modelo do período 5 de validação. Fonte: Própria.

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. A função está representada pela Equação 3, em que θ_n é o coeficiente de cada termo da função, 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. Os valores referentes a cada θ_n estão apresentados no vetor $theta$ a seguir:

```
theta =
    [ 6.40354684e-02,  -4.70789965e-07,
      2.11998944e-02,   1.27730667e-01,
      4.87904670e-07,   4.04490221e-08,
      1.33260000e-01,   1.71026841e-01,
     -7.99468299e-08,  -5.62155583e-08,
      3.52937369e-01,   2.30266591e-01,
      4.19993602e-08,   1.46142426e-01,
```

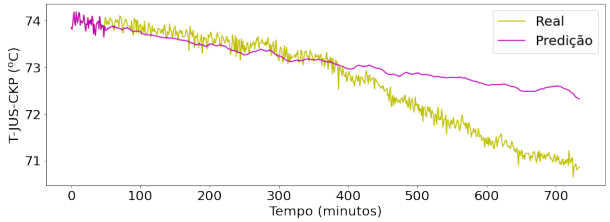


Fig. 5: Comparativo dos dados reais e estimados pelo modelo do período 10 em transição para estado de anomalia. Fonte: Própria.

```
5.64989052e-02,  2.83576743e-02,
-1.75531634e-01, 4.66381040e-02,
3.14521464e-02,  -2.38588959e-01 ]
```

$$
\begin{aligned}
y(k) &= \theta_1 y(k-67) + \theta_2 u_2(k-93) + \theta_3 u_1(k-388) \\
&+ \theta_4 y(k-3) + \theta_5 u_2(k-1) + \theta_6 u_3(k-176) \\
&+ \theta_7 y(k-65) + \theta_8 u_1(k-216) + \theta_9 u_3(k-350) \\
&+ \theta_{10} u_3(k-195) + \theta_{11} y(k-1) + \theta_{12} u_1(k-123) \\
&+ \theta_{13} u_3(k-164) + \theta_{14} y(k-14) + \theta_{15} y(k-298) \\
&+ \theta_{16} y(k-235) + \theta_{17} u_1(k-144) + \theta_{18} y(k-219) \\
&+ \theta_{19} y(k-169) + \theta_{20} u_1(k-130)
\end{aligned} \tag{3}
$$

A partir dessa função, é possível verificar que a quantidade máxima de termos bem como o atraso máximo e a profundidade máxima da árvore, definida inicialmente e apresentadas na Tabela II, foi explorada pelo modelo, alcançando tais máximos. Além disso, todas as variáveis do processo foram importantes no modelo, aparecendo em diversos termos da função final. O uso de atrasos elevados também foi bastante explorado pelo algoritmo, o que mostra que dependências de longo tempo presentes entre as variáveis foram importantes na predição de temperatura.

Quando observado sobre um período com transição de anomalia, o modelo não é capaz de explicar os casos de anomalias e, a medida que evolui no período transitório, o erro apresenta uma tendência de aumento, chegando a superar 2 graus Celsius na temperatura T-JUS-CKP. Tal comportamento é um indicativo que esse modelo poderia ser utilizado em análises de detecção de anomalia em poços de petróleo. A Figura 5 apresenta o comparativo da curva real e de predição de um dos períodos com transitório da anomalia Incrustação em CKP.

Considerações finais

Por meio deste trabalho foi possível compreender o comportamento do MGGP a partir de variações em seus hiperparâmetros, o que conferiu a escolha dos melhores valores para a realização do treinamento do modelo. 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 *soft sensor* desenvolvido apresentou um resultado satisfatório sobre os dados reais de operação de extração de petróleo em poço surgente. O modelo 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. Além disso, quando observado o desempenho desse *soft sensor* sobre períodos com transitórios de anomalia, notou-se o potencial de uso deste modelo no cenário de detecção de anomalias em poços surgentes, em razão do erro de predição nesse tipo de período.

Dado isso, os próximos passos deste trabalho será a avaliação aprofundada do *soft sensor* na detecção de anomalias em poços surgentes de petróleo como também a exploração da metodologia aplicada a outros poços disponíveis na base de dados. Outro aspecto a ser explorado em trabalhos futuros é a convergência do algoritmo evolucionário, bem como a análise paramétrica e de robustez do MGGP.

Agradecimentos

Agradecimento 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.

References

[1] B. H. Barbosa, L. P. Gomes, A. F. Teixeira, and L. A. Aguirre, "Downhole pressure estimation using committee machines and neural networks," *IFAC-PapersOnLine*, vol. 48, no. 6, pp. 286–291, 2015, 2nd IFAC Workshop on Automatic Control in Offshore Oil and Gas Production OOGP 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S240589631500912X

[2] L. A. Aguirre, B. O. S. Teixeira, B. H. Barbosa, A. F. Teixeira, M. C. M. M. Campos, and E. M. Mendes, "Development of soft sensors for permanent downhole gauges in deepwater oil wells," *Control Engineering Practice*, vol. 65, pp. 83–99, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0967066117301284

[3] R. S. F. Nascimento, R. E. V. Vargas, B. H. G. Barbosa, and I. H. F. dos Santos, "Detecção de anomalias em poços de petróleo surgentes com stacked autoencoders," in *Simpósio Brasileiro de Automação Inteligente*, vol. 1, no. 1, 2021, pp. 2049–2056.

[4] P. S.-L. Jämsä-Jounela, "Future trends in process automation," *IFAC Proceedings Volumes*, vol. 40, no. 1, pp. 1–10, 2007, 8th IFAC Symposium on Cost Oriented Automation. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S147466701531199X

[5] G. Portela, *Gerenciamento de Riscos na Indústria de Petróleo e Gás: offshore e onshore*. Rio de Janeiro, RJ, Brasil: Elsevier Editora Ltda, 2015.

[6] I. Andreolli, *Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo*. Rio de Janeiro, RJ, Brasil: Editora Interciência, 2016.

[7] L. Freitas, B. H. G. Barbosa, and L. A. Aguirre, "Including steady-state information in nonlinear models: An application to the development of soft-sensors," *Engineering Applications of Artificial Intelligence*, vol. 102, p. 104253, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621001007

[8] G. A. de Morais, B. H. Barbosa, D. D. Ferreira, and L. S. Paiva, "Soft sensors design in a petrochemical process using an evolutionary algorithm," *Measurement*, vol. 148, p. 106920, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224119307778

[9] B. H. G. Barbosa, L. A. Aguirre, C. B. Martinez, and A. P. Braga, "Black and gray-box identification of a hydraulic pumping system," *IEEE Transactions on Control Systems Technology*, vol. 19, no. 2, pp. 398–406, 2011.

[10] W. R. L. Junior, S. A. M. Martins, and E. G. Nepomuceno, "Meta-model structure selection: Building polynomial narx model for regression and classification," 2021.

[11] R. Emanuel Vaz Vargas, "Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo," Ph.D. dissertation, Universidade Federal do Espírito Santo, Centro Tecnológico, Programa de pós-graduação em engenharia elétrica, Vitória, 2019.

[12] J. E. Thomas, *Fundamentos de Engenharia de Petróleo*. Rio de Janeiro, RJ, Brasil: Editora Interciência, 2004.

[13] W. Fernandes Júnior, "Detecção de anomalias em poços produtores de petróleo usando aprendizado de máquina," Master's thesis, Instituto Federal do Espírito Santo, Programa de Pós-graduação Em Computação Aplicada, Serra - Espírito Santo, 2022.

[14] L. N. Hüffner, J. O. Trierweiler, and M. Farenzena, "Are complex black-box models for permanent downhole gauge pressure estimation necessary?" *Journal of Petroleum Science and Engineering*, vol. 173, pp. 715–732, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092041051830915X

[15] C. I. Ossai, "Modified spatio-temporal neural networks for failure risk prognosis and status forecasting of permanent downhole pressure gauge," *Journal of Petroleum Science and Engineering*, vol. 184, p. 106496, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920410519309179

[16] M. C. M. d. A. D. A. L. G. Saavedra, "Avaliação do efeito da incerteza do bsw na vazão de óleo em sistemas de elevação e escoamento de petróleo," Rio de Janeiro, p. 63, 2016.

[17] R. E. V. Vargas, C. J. Munaro, P. M. Ciarelli, A. G. Medeiros, B. G. do Amaral, D. C. Barrionuevo, J. C. D. de Araújo, J. L. Ribeiro, and L. P. Magalhães, "A realistic and public dataset with rare undesirable real events in oil wells," *Journal of Petroleum Science and Engineering*, vol. 181, p. 106223, 2019.

[18] M. C. Viana, R. P. Cosmo, F. de A. Ressel Pereira, D. da C. Ribeiro, and A. L. Martins, "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.

[19] Y. Jiang, S. Yin, J. Dong, and O. Kaynak, "A review on soft sensors for monitoring, control, and optimization of industrial processes," *IEEE Sensors Journal*, vol. 21, no. 11, pp. 12 868–12 881, 2021.

[20] J. Randek and C.-F. Mandenius, "On-line soft sensing in upstream bioprocessing," *Critical reviews in biotechnology*, vol. 38, pp. 1–16, 04 2017.

[21] F. Lotufo and C. Garcia, "Sensores virtuais ou soft sensors: Uma introdução," 01 2008.

[22] S. Graziani and M. G. Xibilia, *Development and Analysis of Deep Learning Architectures (Deep Learning for Soft Sensor Design)*. Cham, Switzerland: Springer, 2020.

[23] H. C. de Castro and B. H. G. Barbosa, "Multi-gene genetic programming for structure selection of polynomial narmax models," in *XXIII Congresso Brasileiro de Automática, 2020, Santa Maria. CBA2020*, 2020, pp. 1–8.

[24] L. dos Santos Coelho, T. C. Bora, and C. E. Klein, "A genetic programming approach based on lévy flight applied to nonlinear identification of a poppet valve," *Applied Mathematical Modelling*, vol. 38, no. 5, pp. 1729–1736, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0307904X1300591X

[25] S. A. A. Taqvi, L. D. Tufa, H. Zabiri, A. Maulud, and F. Uddin, "Fault detection in distillation column using narx neural network," *Neural Computing and Applications*, vol. 32, 04 2020.

[26] V. V. S., H. K. Mohanta, and A. K. Pani, "Adaptive non-linear soft sensor for quality monitoring in refineries using just-in-time learning—generalized regression neural network approach," *Applied Soft Computing*, vol. 119, p. 108546, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494622000758

[27] Y. Fan, B. Tao, Y. Zheng, and S.-S. Jang, "A data-driven soft sensor based on multilayer perceptron neural network with a double lasso approach," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 3972–3979, 2020.

[28] H. Kaneko and K. Funatsu, "A new process variable and dynamics selection method based on a genetic algorithm-based wavelength selection method (vol 58, pg 1829, 2012)," *AIChE Journal*, vol. 58, 06 2012.

[29] S. Sharma and S. Tambe, "Soft-sensor development for biochemical systems using genetic programming," *Biochemical Engineering Journal*, vol. 85, 04 2014.

[30] R. Huang, Z. Li, and B. Cao, "A soft sensor approach based on an echo state network optimized by improved genetic algorithm," *Sensors (Basel, Switzerland)*, vol. 20, 09 2020.

[31] H. Kaneko and K. Funatsu, "Preparation of comprehensive data from huge data sets for predictive soft sensors," *Chemometrics and Intelligent Laboratory Systems*, vol. 153, pp. 75–81, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169743916300338

[32] J. Mendes, F. Souza, R. Araújo, and N. Gonçalves, "Genetic fuzzy system for data-driven soft sensors design," *Applied Soft Computing*, vol. 12, p. 3237–3245, 10 2012.

[33] D. Fontes, L. Vasconcelos, and R. Brito, "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*, vol. 141, p. 107028, 07 2020.

[34] Özmen, C. Sınanoğlu, T. Batbat, and A. Güven, "Prediction of slipper pressure distribution and leakage behaviour in axial piston pumps using ann and mggp," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–13, 03 2019.

[35] M. Hinchliffe, H. Hiden, B. McKay, M. Willis, M. Tham, and G. Barton, "Modelling chemical process systems using a multi-gene genetic programming algorithm," in *Late Breaking Papers at the Genetic Programming 1996 Conference Stanford University July 28-31, 1996*, J. R. Koza, Ed. Stanford University, CA, USA: Stanford Bookstore, 28–31 Jul. 1996, pp. 56–65.

[36] J. Maynard, A. Camacho, R. Oliveira, A. Talavera, and M. Pacheco, "Aplicações de programação genética em reservatórios de petróleo," 10 2015, pp. 1–6.

[37] M. Niazkar and H. R. Niazkar, "Covid-19 outbreak: Application of multi-gene genetic programming to country-based prediction models," *Electronic Journal of General Medicine*, vol. 17, p. em247, 03 2020.

[38] F. Mota, G. Carvalho, H. Castro, and B. H. G. Barbosa, "Identificação de um sistema de bombeamento hidráulico com algoritmo evolucionário multi-objetivo," *Congresso Brasileiro de Automática, Anais da Sociedade Brasileira de Automática*, vol. 2, no. 1, pp. 1–8, 2020.

[39] R. Vargas, "The first realistic and public dataset with rare undesirable real events in oil wells," 2019. [Online]. Available: github.com/ricardovvargas

[40] H. C. de Castro and B. H. G. Barbosa, "A python library for nonlinear system identification using multi-gene genetic programming algorithm," *arXiv*, vol. 2211.05723, 2022.