UNIVERSIDADE FEDERAL DE ALAGOAS
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL

IGOR DE MELO NERY OLIVEIRA

TÉCNICAS DE INFERÊNCIA E PREVISÃO DE DADOS COMO SUPORTE À ANÁLISE DE INTEGRIDADE DE REVESTIMENTOS

Maceió-AL

Dezembro de 2020

IGOR DE MELO NERY OLIVEIRA

TÉCNICAS DE INFERÊNCIA E PREVISÃO DE DADOS COMO SUPORTE À ANÁLISE DE INTEGRIDADE DE REVESTIMENTOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Civil do Centro de Tecnologia da Universidade Federal de Alagoas.

Orientador: Prof. Dr. Eduardo Toledo de Lima Junior

Maceió-AL

Dezembro de 2020

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

O48t Oliveira, Igor de Melo Nery.
Técnicas de inferência e previsão de dados como suporte à análise de integridade de revestimentos / Igor de Melo Nery Oliveira. - 2020.
84 f. : il.

Orientador: Eduardo Toledo de Lima Junior.
Dissertação (Mestrado em Engenharia Civil) – Universidade Federal de Alagoas. Centro de Tecnologia. Maceió, 2021.

Bibliografia: f. 83-84.

1. Inferência estatística. 2. Análise de séries temporais. 3. Modelos Autorregressivos, Integrados e de Médias Móveis. 4. Monitoramento estrutural - Transformação digital. I. Título.

CDU:624.04

TÉCNICAS DE INFERÊNCIA E PREVISÃO DE DADOS COMO SUPORTE À ANÁLISE DE INTEGRIDADE DE REVESTIMENTOS

IGOR DE MELO NERY OLIVEIRA

Dissertação submetida à banca examinadora do Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Alagoas e aprovada no dia 21 do mês de dezembro do ano de 2020.

Banca Examinadora:

Prof. Dr. Eduardo Toledo de Lima Junior
(Orientador – PPGEC/UFAL)

Prof. Dr. João Paulo Lima Santos
(Avaliador Interno- PPGEC/UFAL)

Pesquisador Dr. Ricardo Emanuel Vaz Vargas
(Avaliador Externo – PETROBRAS)

AGRADECIMENTOS

À minha esposa Camila Farias por me apoiar nos momentos difíceis, sempre me motivando para a conclusão deste trabalho.

À minha mãe Maria Cícera, por sempre acreditar em mim e me dar todo o suporte para concluir mais esta fase da minha vida.

Ao meu pai Esequiel Nery e minha vó Maria Vitória, que não estão mais aqui mas são minha fonte de inspiração, por sempre terem se preocupado com minha educação e formação.

Ao professor Eduardo Toledo por sua imensa paciência, atenção, apoio, e orientação incansável. Este trabalho não seria possível sem sua ajuda.

Ao professor William Lira por todas as oportunidades que me foram dadas, imprescindíveis para que eu trilhasse esta jornada.

Aos demais professores que contribuíram com minha formação acadêmica, em especial os professores Eduardo Nobre, Flávio Lima, Márcio André e Adeildo Soares, por compartilharem seus conhecimentos e serem fonte de inspiração e exemplo.

A todos os amigos que me apoiaram durante os contratempos, em especial Felipe Pedrosa, Lucas Omena, Weverton Marques, Lucas Gouveia, Wellington Pedro, Thiago Barbosa, Tarciso, Ricardo, Catarina, Tiago, Heleno e Emerson.

Ao Laboratório de Computação Científica e Visualização (LCCV), por sua infraestrutura e ambiente propício para desenvolvimento do trabalho.

À PETROBRAS, pelo suporte financeiro concedido por meio do projeto de pesquisa, desenvolvimento e inovação identificado pelo número ANP 20601-1.

RESUMO

O projeto de um poço de petróleo é uma atividade complexa e multidisciplinar, que tem como uma de suas principais premissas a adequada previsão de integridade do poço ao longo de seu ciclo de vida. Apesar de todos os cuidados no dimensionamento da sua estrutura, eventualmente o poço pode ser exposto a condições de carregamento não previstas. No tocante ao monitoramento de poços em serviço, a aquisição de dados referentes a variáveis como pressão e temperatura permite identificar se o poço está operando dentro dos parâmetros previstos em projeto. Por meio de técnicas de previsão de séries de dados temporais, as informações desses sensores têm potencial de serem utilizadas não só para diagnosticar um problema já ocorrido, mas também para prevenir a sua ocorrência, criando um sistema supervisório em tempo real que seja capaz de antecipar estados futuros de carregamento. Em outra perspectiva, no contexto de projetos de poços, com a evolução das normas de projeto de revestimento, sugere-se o uso de métodos probabilísticos em seu dimensionamento, evidenciando a importância de um melhor conhecimento acerca das variáveis de projeto. A inferência estatística sobre dados de fabricação é motivada pela demanda por um melhor entendimento das incertezas sobre esses parâmetros, em termos das dimensões do tubular e das propriedades da liga metálica que o constitui. Esta dissertação versa sobre um conjunto de técnicas de inferência estatística e previsão de dados, como suporte a práticas de projeto e de monitoramento de integridade estrutural de poços.

Palavras-chave: Transformação Digital; Inferência Estatística; Monitoramento Estrutural; Séries Temporais; ARIMA.

ABSTRACT

The design of an oil well is a complex and multidisciplinary activity, which has as one of its main premises the adequate prediction of the well's integrity throughout its life cycle. Despite all the effort taken in designing its structure, eventually, the well may be exposed to unpredicted loading conditions. Regarding the well monitoring, the acquisition of data referring to variables such as pressure and temperature allows identifying whether the well is operating within the parameters predicted in the design. By using time series forecasting techniques, the information gathered from the sensors can be used to diagnose an anomaly that has already occurred and prevent its occurrence, creating a real-time supervisory system capable of anticipating future loading states. Furthermore, in the context of well design, with the evolution of the rules for casing design, it is suggested to use probabilistic methods, highlighting the importance of better knowledge about the design variables. The statistical inference about manufacturing data is motivated by the demand for a better understanding of these parameters' uncertainties in terms of the dimensions of the tubular and the properties of the metallic alloy that constitutes it. This dissertation approaches a set of statistical inference and data prediction techniques, aiming to support designing practices and monitor the structural integrity of wells.

Keywords: Digital Transformation; Statistical inference; Structural Monitoring; Time Series; ARIMA .

LISTA DE FIGURAS

Figura 1 – Combinações de enviesamento e variância. 25
Figura 2 – Histograma de avgWT do tubo 2 e suas distribuições candidatas. 40
Figura 3 – Histograma de avgWT do tubo 18 e suas distribuições candidatas. 41
Figura 4 – Histograma de avgOD do tubo 29 e suas distribuições candidatas. 42
Figura 5 – Histograma de avgOD do tubo 46 e suas distribuições candidatas. 43
Figura 6 – Histograma dos dados de YS e suas distribuições candidatas. 47
Figura 7 – Histograma dos dados de UTS e suas distribuições candidatas. 47
Figura 8 – Histograma de avgWT do conjunto global e suas distribuições candidatas. . 49
Figura 9 – Histograma de avgOD do conjunto global e suas distribuições candidatas. . 50
Figura 10 – Histograma de avgWT do conjunto global e todas as distribuições candidatas. 53
Figura 11 – Histograma de avgOD do conjunto global e as distribuições candidatas. . . . 53
Figura 12 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados avgWT. 55
Figura 13 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados avgOD. 55
Figura 14 – Histograma de YS do conjunto global e todas as distribuições candidatas. . . 56
Figura 15 – Histograma de UTS do conjunto global e as distribuições candidatas. 56
Figura 16 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados YS. 58
Figura 17 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados UTS. 58
Figura 18 – Histograma de avgOD do conjunto global padronizado e suas distribuições candidatas. 59
Figura 19 – Histograma de avgWT do conjunto global padronizado e suas distribuições candidatas. 59
Figura 20 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados global padronizado avgOD. 60
Figura 21 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados global padronizado avgWT. 60
Figura 22 – Pressão no PDG durante um evento de fechamento espúrio da DHSV. 64
Figura 23 – Pressão no PDG logo após um evento de fechamento espúrio da DHSV. . . 64

Figura 24 – Gráfico ACF para $d = 0$. 65
Figura 25 – Gráfico PACF para $d = 0$. 65
Figura 26 – Gráfico ACF para $d = 1$. 65
Figura 27 – Gráfico PACF para $d = 1$. 66
Figura 28 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA. 66
Figura 29 – Previsão da série temporal via modelo ARIMA(1,0,10). 67
Figura 30 – Previsão da série temporal via modelo ARIMA(5,0,13). 68
Figura 31 – Esquema simplificado de um poço típico *offshore*. 69
Figura 32 – Pressão no P-MON-CKP com amplitude do sinal. 69
Figura 33 – Janela utilizada para treinamento do modelo. 70
Figura 34 – Gráfico ACF para $d = 0$. 70
Figura 35 – Gráfico PACF para $d = 0$. 71
Figura 36 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA. 71
Figura 37 – Previsão da série temporal via modelo ARIMA (11,0,19). 72
Figura 38 – Previsão da série temporal via modelo ARIMA (11,0,17). 72
Figura 39 – Janela utilizada para treinamento do modelo. 74
Figura 40 – Gráfico ACF para $d = 0$. 74
Figura 41 – Gráfico PACF para $d = 0$. 74
Figura 42 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA. 75
Figura 43 – Previsão da série temporal via modelo ARIMA (2,0,0). 76
Figura 44 – Dados do poço número 1 e previsões via ARIMA. 78
Figura 45 – MAE dos valores previstos via ARIMA para o poço número 1. 78
Figura 46 – Janela de operação normal do poço número 2 e previsões via ARIMA. . . . 79
Figura 47 – MAE dos valores previstos via ARIMA para o poço número 2. 80
Figura 48 – Dados do poço número 2 e previsões via ARIMA. 80

LISTA DE TABELAS

Tabela 1 – Funções de probabilidade de modelos usuais de distribuição. 20
Tabela 2 – Momentos dos modelos de distribuição candidatos. 21
Tabela 3 – Dados de espessura média do tubo 2. 37
Tabela 4 – Dados de espessura média do tubo 18. 37
Tabela 5 – Dados de diâmetro médio do tubo 29. 38
Tabela 6 – Dados de diâmetro médio do tubo 46. 38
Tabela 7 – Modelos candidatos ajustados sobre os dados de avgWT do tubo 2. 38
Tabela 8 – Modelos candidatos ajustados sobre os dados de avgWT do tubo 18. 39
Tabela 9 – Modelos candidatos ajustados sobre os dados de avgOD do tubo 29. 39
Tabela 10 – Modelos candidatos ajustados sobre os dados de avgOD do tubo 46. 39
Tabela 11 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do tubo 2. 40
Tabela 12 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do tubo 18. 41
Tabela 13 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do tubo 29. 42
Tabela 14 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do tubo 46. 43
Tabela 15 – Resumo da inferência dos 50 conjuntos amostrais de seções de avgWT. . . . 44
Tabela 16 – Resumo da inferência dos 50 conjuntos amostrais de seções de avgOD. . . . 44
Tabela 17 – Modelos candidatos parametrizados sobre os dados de resistência YS. . . . 45
Tabela 18 – Modelos candidatos parametrizados sobre os dados de resistência UTS. . . . 45
Tabela 19 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de YS. 46
Tabela 20 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de UTS. 46
Tabela 21 – Modelos candidatos parametrizados sobre os dados de avgWT do conjunto global. 48
Tabela 22 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global. 48
Tabela 23 – Modelos candidatos parametrizados sobre os dados de avgOD do conjunto global. 50
Tabela 24 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global. 50
Tabela 25 – Grau de incerteza na aferição dos dados do conjunto global de avgWT. . . . 51
Tabela 26 – Grau de incerteza na aferição dos dados do conjunto global de avgOD. . . . 51

Tabela 27 – Lista das 94 distribuições candidatas e seu correspondente número de parâmetros. 52
Tabela 28 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global incluindo modelos de maior complexidade. . 54
Tabela 29 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global incluindo modelos de maior complexidade. . . 54
Tabela 30 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de YS para os modelos expandidos. 57
Tabela 31 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de UTS para os modelos expandidos. 57
Tabela 32 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global padronizado. 61
Tabela 33 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global padronizado. 62
Tabela 34 – Modelos indicados segundo seu AIC. 67
Tabela 35 – Medidas de acurácia das estimativas dos valores para treino da Série de Dados 1. 68
Tabela 36 – Medidas de acurácia das previsões dos valores futuros da Série de Dados 1. 68
Tabela 37 – Modelos indicados segundo seu AIC. 71
Tabela 38 – Medidas de acurácia das estimativas dos valores para treino da série de dados 2. 73
Tabela 39 – Medidas de acurácia das previsões dos valores futuros da série de dados 2. . 73
Tabela 40 – Modelos indicados segundo seu AIC. 75
Tabela 41 – Medidas de acurácia das estimativas dos valores para treino da série de dados 3. 76
Tabela 42 – Medidas de acurácia das previsões dos valores futuros da série de dados 3. . 76

LISTA DE ABREVIATURAS E SIGLAS

ACF	Função de Autocorrelação (*Autocorrelation Function*)
AIC	Critério de Informação de Akaike (*Akaike Information Criterion*)
AR	Modelos Autorregressivos (*Auto Regressive Models*)
ARIMA	Modelos Autorregressivos, Integrados e de Médias Móveis (*Auto Regressive Integrated Moving Average Models*)
ARMA	Modelos Autorregressivos e de Médias Móveis (*Auto Regressive Moving Average Models*)
avgOD	Diâmetro Externo Médio (*Average Outer Diameter*)
avgWT	Espessura de Parede Média (*Average Wall Thickness*)
BIC	Critério de Informação Bayesiano (*Bayesian Information Criterion*)
CDF	Função de Distribuição Acumulada (*Cumulative Distribution Function*)
Cov	Covariância (*Covariance*)
CV	Coeficiente de Variação (*Coefficient of Variation*)
CVRMSE	Coeficiente de Variação da Raiz Quadrada do Erro Quadrático Médio (*Coefficient of Variation of the Root Mean Square Error*)
DHSV	Válvula de Segurança de Subsuperfície (*Downhole safety valve*)
I	Modelos Integrados (*Integrated Models*)
IID	Independente e Identicamente Distribuída (*Independent and Identically Distributed*)
KS	Teste Kolmogorov-Smirnov (*Kolmogorov-Smirnov Test*)
MA	Modelos de Médias Móveis (*Moving Average Models*)
MAE	Erro Absoluto Médio (*Mean Absolute Error*)
MAPE	Erro Absoluto Médio Percentual (*Mean Absolute Percentage Error*)
PACF	Função de Autocorrelação Parcial (*Partial Autocorrelation Function*)
PCK	Válvula de Bloqueio de Produção (*Production Choke*)
PDF	Função de Densidade de Probabilidade (*Probability Density Function*)

PDG	Medidor de Pressão de Fundo (*Pressure Downhole Gauge*)
P-MON-CKP	Pressão à Montante da PCK (*Pressure Upstream of PCK*)
RMSE	Raiz Quadrada do Erro Quadrático Médio (*Root Mean Square Error*)
RNA	Redes Neurais Artificiais (*Artificial Neural Network*)
UTS	Resistência à Tração (*Ultimate Tensile Strength*)
YS	Tensão de Escoamento (*Yield strength*)

SUMÁRIO

1	**INTRODUÇÃO**	**14**
1.1	**Objetivos**	**16**
1.2	**Metodologia e Organização do Trabalho**	**16**
1.3	**Delimitação do Trabalho**	**17**
2	**INFERÊNCIA ESTATÍSTICA DE DADOS**	**18**
2.1	**Seleção de um Modelo de Distribuição Estatística**	**19**
2.2	**Parametrização dos Modelos Estatísticos Hipotéticos**	**21**
2.3	**Testes de Aderência**	**22**
2.4	**Critérios de Informação**	**24**
2.5	**Quantificação de Incertezas no Processo de Aquisição de Dados**	**26**
2.6	**Imposição de Limites num Modelo de Distribuição Teórico**	**27**
3	**PREVISÃO DE DADOS POR MODELOS ARIMA**	**29**
3.1	**Quantificação da Dependência dos Termos de uma Série Temporal**	**29**
3.2	**Modelos Autorregressivos, Integrados e de Médias Móveis (ARIMA)**	**31**
3.3	**Medidas de Acurácia**	**34**
4	**ESTUDOS DE CASO DE INFERÊNCIA E PREVISÃO DE DADOS COMO SUPORTE À ANÁLISE DE INTEGRIDADE DE REVESTIMENTOS**	**36**
4.1	**Inferência Estatística de Dados de Produção de Tubulares**	**36**
4.2	**Previsão de Dados para Detecção de Anomalias**	**63**
5	**CONCLUSÃO E SUGESTÕES PARA TRABALHOS FUTUROS**	**81**
	REFERÊNCIAS	**83**

1 INTRODUÇÃO

O projeto de um poço de petróleo é uma atividade complexa e envolve a análise de diferentes aspectos, tais como a confiabilidade do revestimento, o comportamento do sal durante a perfuração, dentre outros. Cada um desses aspectos é analisado por profissionais com especialidades distintas, tornando, assim, o projeto de um poço uma tarefa multidisciplinar, composto de vários subprojetos, que devem ser analisados de forma integrada.

O gerenciamento de um poço de petróleo é fundamentado em analisar, de forma eficiente, uma grande massa de dados oriundos das diversas partes do sistema que compõem o poço de petróleo. Partes estas relativas à sua composição - formação geológica, maquinário, fluidos, componentes estruturais do poço, estado de pressão e temperatura atuantes - e à sua temporalidade - etapa de exploração, projeto, perfuração, produção, intervenção e possível fechamento.

A integração dos diversos subprojetos, através de um sistema inteligente e automatizado, capaz de tomar decisões e auxiliar o gerenciamento de um poço, de forma eficiente e segura, está no ápice da inovação tecnológica em termos de transformação digital para o setor de óleo e gás. Diversas são as referências à transformação digital e à indústria 4.0 no setor. No que diz respeito a equipamentos e instalações *offshore*, existem iniciativas para monitoramento estrutural e supervisão operacional em plataformas marítimas (GERKE, 2018 (acessado em 24/07/2018)). Mayani et al. (2018) apresentam casos de utilização de modelos *Digital Twin* em operações de perfuração nos últimos dez anos, em diferentes operadoras. Nadhan et al. (2018) mostram como a aquisição de dados em tempo real e a análise preditiva em operações de perfuração podem trazer benefícios no processo de tomada de decisão.

Em sua maioria, as metodologias referenciadas são baseadas no entendimento dos efeitos das condições operacionais e anômalas sobre a integridade estrutural, segurança e desempenho do sistema em estudo. Normalmente, esse entendimento se concretiza na forma de fatores de segurança utilizados durante o projeto. Estes asseguram um determinado nível de segurança para cenários de carregamento previamente experimentados e, portanto, já conhecidos pelos projetistas, não contemplando situações extremas ainda não previstas.

Apesar de todos os cuidados no dimensionamento da estrutura de um poço (considerando a pior situação possível com a máxima carga e a mínima resistência), eventualmente pode ocorrer algo em sua operação que venha a submetê-lo a uma condição não prevista, como, por exemplo, uma manobra de válvulas que pode pressurizar demasiadamente um revestimento. Essa situação pode comprometer toda a estrutura do poço, podendo causar um prejuízo financeiro ou um dano ambiental expressivo. Há casos registrados onde mais de 60% dos poços de um campo tiveram falha no revestimento durante a sua produção (YUAN et al., 2013) .

Mesmo com todos esses impactos, o monitoramento estrutural de um poço apresenta grandes dificuldades, a começar pelo número reduzido de sensores que podem ser instalados,

limitados ao interior da coluna de produção e ao anular adjacente. Cada sensor de monitoramento tem objetivos diversos, podendo ainda ser aproveitados para monitoramento estrutural dos poços de petróleo, atualizando continuamente os cenários de carregamento aos quais os revestimentos estão submetidos, e verificando se algum elemento estrutural está próximo da falha.

Para garantir a integridade do poço e promoção da segurança operacional, propõe-se uma estratégia de monitoramento do estado de carregamento do poço para identificar se ele está operando dentro dos parâmetros previstos em projeto. Medições fora do esperado podem indicar diversas situações, como falha de equipamento, previsões equivocadas de comportamento do reservatório, erros humanos na operação do poço ou mesmo falhas no projeto.

Através de técnicas de previsão de séries de dados temporais, as informações desses sensores têm potencial de serem utilizadas não só para diagnosticar um problema já ocorrido, mas também para prevenir a sua ocorrência, criando um sistema preditivo em tempo real que seja capaz de antecipar o estado de carregamento futuro.

Neste contexto, a demanda por técnicas que realizem a previsão de estados futuros de carregamento em um poço é uma realidade corrente da Petrobras, sendo o tema explorado em um projeto de pesquisa em execução pelo grupo em que o autor está inserido (LCCV, 2020).

Não obstante, no contexto de projetos de poços de petróleo, com o advento da API TR 5C3 (2008), que preconiza uso de modelos de resistência de tubulares em Estado de Limite Último e sugere o uso de métodos probabilísticos para o projeto de revestimentos, evidencia-se a importância de um melhor conhecimento acerca das variáveis de projeto, incluindo parâmetros dos tubulares e carregamentos atuantes, possibilitando projetar poços que satisfaçam um nível de confiança alvo.

Estudos referentes à comparação dos níveis de segurança embutidos nas formulações de projeto sugeridas pela API TR 5C3 (2008) são apresentados em Gouveia (2014). Ainda, o autor verifica a probabilidade de falha de tubos de revestimento em poços de petróleo com base em dados estatísticos de tubos reais compilados pela API. Oliveira (2017) contribui ao tema ao incorporar variáveis aleatórias de carregamento à análise. Já Silva (2018) utiliza a teoria de confiabilidade de sistemas para analisar modos de falha associados em sistemas de revestimento.

A inferência estatística sobre dados de produção é motivada pela demanda por um melhor entendimento das incertezas sobre os parâmetros de projeto de revestimento, em termos das dimensões do tubular e das propriedades da liga metálica que o constitui.

Quantificar estas variabilidades decorrentes dos sistemas de produção, somadas à imprecisões na aquisição de dados, permite avaliar a influência de cada variável aleatória sobre o desempenho estrutural do tubo, bem como sua conformidade, levando a uma definição mais robusta dos valores admissíveis de cada um desses parâmetros. Assim, contribui-se com o projeto de sistemas de revestimento, de forma que estes satisfaçam níveis de segurança especificados, sem abrir mão da economicidade.

Técnicas de inferência estatística sobre dados de fabricação de tubulares figuram como um tema de interesse por parte da Petrobras, sendo objeto de estudo em um projeto de pesquisa em andamento no grupo do autor (LCCV, 2019), de forma que tais técnicas auxiliem a análise de confiabilidade estrutural de poços, assim como apontado por Silva (2018).

Em suma, esta dissertação versa sobre um conjunto de técnicas de inferência estatística e previsão de dados. Entende-se que as duas frentes de trabalho, por mais que distintas, são campos complementares de ciência de dados.

As aplicações apresentadas e discutidas nesta dissertação são concentradas em poços de petróleo, devido aos projetos de pesquisa supracitados. No entanto, ressalta-se que tais técnicas podem ser aplicadas em outros contextos com tratamento de séries de dados.

1.1 Objetivos

Estudar e aplicar técnicas de ciência dos dados no contexto do projeto e do monitoramento de integridade estrutural de poços de petróleo.

1.1.1 Objetivos Específicos

- Estudar e codificar técnicas de inferência estatística, e posterior aplicação a dados de manufatura de tubulares;
- Codificar e analisar técnicas de previsão de séries temporais, tendo em vista o monitoramento e antecipação do estado de carregamento de poços em produção;
- Criar uma rotina de monitoramento em tempo real para detecção e antecipação de anomalias como apoio à análise de integridade do poço.

1.2 Metodologia e Organização do Trabalho

Esta dissertação foi desenvolvida em cinco capítulos. O primeiro capítulo consiste da presente introdução sobre os temas abordados neste trabalho.

No Capítulo 2 expõe-se um conjunto de técnicas para inferência estatística de dados. Uma revisão bibliográfica é realizada sobre conceitos de probabilidade e estatística e as principais distribuições de probabilidade contínuas; seguida por métodos de parametrização de distribuições de probabilidade hipotéticas, são eles: Método do Momentos (HANSEN, 1982) e Método da Máxima Verossimilhança (SEVERINI, 2000); os testes de aderência Chi-Quadrado, Kolmogorov-Smirnov e Anderson-Darling (GIBBONS; CHAKRABORTI, 2014) são apresentados; ao final do capítulo, técnicas complementares para inferência estatística são expostas: seleção de modelos por meio de critérios de informação, quantificação de incertezas no processo de aquisição de dados e uma metodologia de imposição de limites numa distribuição hipotética.

No Capítulo 3 são abordadas técnicas para previsão de séries temporais de carregamento. O método Autorregressivo Integrado de Médias Móveis (VU, 2007), do inglês *Auto Regressive Integrated Moving Average* (ARIMA) é explorado.

O Capítulo 4 traz exemplos, analisa os resultados e discute o emprego das técnicas expostas nos Capítulos 2 e 3, como também apresenta as rotinas de inferência estatística e de monitoramento em tempo real para detecção de anomalias.

Por fim, o Capítulo 5 conclui o texto, sintetizando e sugerindo aplicações para trabalhos futuros que utilizem os resultados concebidos e apresentados.

1.3 Delimitação do Trabalho

Este trabalho versa sobre um conjunto de técnicas de inferência estatística e previsão de séries temporais. Por mais que estas técnicas sejam contemplados em ciência de dados, atenta-se para a sutil diferença entre os dois termos.

Segundo James et al. (2013), existem duas abordagens para estimar um conjunto de dados: um problema de inferência pretende estimar quanto cada variável precisamente influencia o resultado da observação, isto é, a inferência almeja identificar a formulação que origina o conjunto de dados; enquanto que uma abordagem de previsão de dados entende que existem erros irredutíveis na formulação e não é possível identificar a formulação original do conjunto de dados, seu objetivo é aproximar um modelo de forma a minimizar seu erro redutível.

No escopo deste trabalho, as técnicas de inferência apresentadas visam identificar o comportamento de séries de dados de produção que, por sua natureza, são séries cujos elementos não apresentam uma relação de dependência entre si. Tais técnicas empenham-se em detectar qual distribuição de probabilidades hipotética melhor se adéqua a determinado conjunto de dados de produção. Os modelos estatísticos testados são limitados à sua disponibilidade na biblioteca Scipy Stats (Virtanen et al., 2020), observando-se a aplicabilidade ao conjunto de dados em estudo.

As técnicas de previsão expostas são aplicadas em séries de dados temporais estáveis com padrões visíveis, de forma que observações futuras possam ser estimadas a partir das observações passadas, não sendo contemplados conjuntos de dados que apresentem, por exemplo, sazonalidade pronunciada.

2 INFERÊNCIA ESTATÍSTICA DE DADOS

Em processos estocásticos, um evento aleatório é um experimento o qual pode gerar resultados diferentes (diz-se aleatórios) mesmo que satisfeitas as mesmas condições iniciais, isto é, não é possível garantir resultados idênticos ao se reproduzir com exatidão um mesmo experimento.

Parte da aleatoriedade se dá pelo fato de não se conhecer (ou não ser possível controlar) todos os fatores que determinam o resultado de um experimento. Erros de medida são os mais comuns, especialmente quando determinada precisão é almejada.

Define-se variável aleatória como o resultado quantitativo de um evento aleatório, população como o conjunto de todos os resultados de um determinado evento aleatório e amostra como um subconjunto da população, geralmente criado a partir de uma investigação dos resultados do evento aleatório.

Na indústria de óleo e gás, pode-se exemplificar tais conceitos no processo de fabricação de um tubo de revestimento: por mais automatizado que seja o processo de fabricação do tubo, de forma a garantir um padrão nas peças produzidas, há uma variação entre os diâmetros dos tubos de um mesmo lote. De fato, há inclusive uma variação no diâmetro ao longo de um tubo, quando este é aferido ao longo de diferentes seções transversais, como habitualmente utilizado no sistema de controle de qualidade.

Pode-se, assim, considerar o processo de fabricação desse tubular como um evento aleatório, suas variáveis geométricas e mecânicas (como espessura, diâmetro, tensão de escoamento do aço, dentre outras) são variáveis aleatórias e o conjunto de resultados de um lote é uma amostra da população total de todos os tubos que podem ser produzidos.

Para caracterizar uma variável aleatória a partir de uma amostra representativa da população, é necessário encontrar um modelo de distribuição estatística que se adéque aos dados. Esta busca envolve três etapas:

1. Seleção de um modelo de distribuição estatística candidato;
2. Determinação de parâmetros para o modelo candidato, de forma que este mimetize o comportamento aleatório dos dados amostrais;
3. Testes para avaliar a fidelidade do ajuste realizado no modelo candidato parametrizado (testes de aderência).

Estas etapas são exploradas nas seções seguintes. De forma geral, quanto maior o conjunto de dados, mais ruído é adicionado à amostra, além disso os testes de aderência tornam-se mais rigorosos. Um pré-tratamento dos dados pode ajudar na redução destes ruídos.

Após a exposição dos métodos para caracterização estatística dos dados, via testes de aderência, alguns outros conceitos são apresentados, no que diz respeito a: seleção de modelos por meio de critérios de informação; quantificação de incertezas no processo de aquisição de dados, e; uma metodologia de imposição de limites numa distribuição hipotética.

2.1 Seleção de um Modelo de Distribuição Estatística

Diversos modelos teóricos de distribuição podem ser empregados na representação de variáveis aleatórias, a depender da natureza e do comportamento destas. Alguns modelos usuais de fenômenos em engenharia são descritos a seguir, sendo suas equações de densidade de probabilidade (PDF) e de distribuição acumulada (CDF) listadas na Tabela 1.

2.1.1 Modelo de Distribuição Uniforme

É a mais simples das distribuições contínuas, e preconiza que a probabilidade de se gerar qualquer ponto em um intervalo contido no espaço amostral é inversamente proporcional ao tamanho do intervalo.

2.1.2 Modelo de Distribuição Normal (Gaussiano)

Modelo amplamente difundido, frequentemente utilizado para prever e modelar fenômenos da natureza graças ao Teorema do Limite Central. A partir deste, propõe-se que toda soma de variáveis aleatórias independentes (de média finita e variância limitada) tem distribuição aproximadamente Normal, desde que o número de termos da soma seja suficientemente grande.

Trata-se de um modelo simétrico em relação à média, com decaimento exponencial de probabilidades nas caudas, amplamente utilizado na modelagem de dados de diversos setores do conhecimento. Na escassez de dados amostrais, pode ser utilizado como uma primeira estimativa e comportamento aleatório de variáveis em engenharia.

2.1.3 Modelo de Distribuição Log-Normal

O logaritmo de uma variável com distribuição Log-Normal é uma distribuição Normal. O teorema do limite central prova que todo produto de variáveis aleatórias independentes e positivas de média finita e variância limitada tem distribuição aproximadamente Log-Normal, desde que o número de termos do produto seja suficientemente grande. Isto é, a distribuição Log-Normal é o resultado estatístico do produto de várias distribuições independentes.

2.1.4 Modelo de distribuição Logístico

Modelo simétrico, semelhante à distribuição Normal, que se diferencia desta pelos valores extremos mais pronunciados, o que resulta em caudas com valores de probabilidade mais representativos. Utilizada em estudos demográficos e em diferentes problemas de engenharia.

2.1.5 Modelo de Distribuição Gamma

Muito utilizada na análise de tempo de vida de produtos e materiais, sendo uma das distribuições mais gerais em sua definição. Muito utilizada como base geral para criação de distribuições mais específicas, como casos particulares desta temos os modelos Exponencial e Chi-Quadrado.

2.1.6 Modelo de Distribuição Weibull

Distribuição bastante utilizada devido à sua versatilidade, sendo capaz de reproduzir comportamentos diversos. Mostra-se apropriada para a representação de valores de carregamentos ambientais, de resistência de materiais e a vida útil de produtos industriais. Proposta originalmente em estudos relacionados ao tempo de falha devido à fadiga de metais.

2.1.7 Modelo de Distribuição Gumbel

Modelo de distribuição de extremos, geralmente empregado para representar valores extremos de eventos ambientais, na previsão de desastres naturais como enchentes, tornados e terremotos, por exemplo. Porém, também é usada para modelar a distribuição de diversas variáveis em engenharia, incluindo parâmetros geométricos, de material, além de solicitações.

A variável aleatória modelada costuma ser um valor de máximo ou mínimo. A distribuição Gumbel é também conhecida como Log-Weibull, uma outra relação entre a distribuição Gumbel com outras distribuições é que a diferença entre duas distribuições Gumbel é uma distribuição Logística.

Tabela 1 – Funções de probabilidade de modelos usuais de distribuição.

Distribuição	**PDF(x)**	**CDF(x)**	**Limites**
Uniforme	$\frac{1}{b-a}$	$\frac{x}{b-a}$	$[a, b]$
Normal	$\frac{1}{\sigma\cdot\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}$	$\frac{1}{\sigma\cdot\sqrt{2\pi}} \int_{-\infty}^{x} e^{\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]} dt$	$[-\infty, \infty]$
Log-Normal	$\frac{1}{\sigma x\sqrt{2\pi}} e^{\left[-\frac{1}{2}\left(\frac{ln(x)-\mu}{\sigma}\right)^2\right]}$	$\Phi\left(\frac{ln(x)-\mu}{\sigma}\right)$	$[0, \infty]$
Logístico	$\frac{e^{-\left(\frac{x-\mu}{s}\right)}}{s\cdot\left(1+e^{-\left(\frac{x-\mu}{s}\right)}\right)^2}$	$\frac{1}{1+e^{-\left(\frac{x-\mu}{s}\right)}}$	$[-\infty, \infty]$
Gamma	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{1}{\Gamma(\alpha)}\gamma(\alpha, \beta x)$	$[0, \infty]$
Weibull	$\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$	$1 - e^{-\left(\frac{x}{\lambda}\right)^k}$	$[0, \infty]$
Gumbel	$\frac{1}{\beta} e^{-\left(\frac{x-\mu}{\beta}+e^{-x}\right)}$	$e^{-e^{-\left(\frac{x-\mu}{\beta}\right)}}$	$[-\infty, \infty]$

Fonte: próprio autor.

2.2 Parametrização dos Modelos Estatísticos Hipotéticos

Os modelos teóricos de distribuição de probabilidade são modelos genéricos para descrição de um comportamento estatístico, sua forma pode ser modificada de acordo com a determinação de seus parâmetros. A seguir são dispostos métodos para determinação dos parâmetros de um modelo de distribuição que represente o comportamento de uma população, a partir de seus valores amostrais.

2.2.1 Método dos Momentos

Neste método os parâmetros são determinados de forma que os momentos do modelo sejam iguais aos da amostra. Para um modelo de dois parâmetros, isto significa que a média e a variância devem ser iguais às da amostra. Caso o modelo possua mais de dois parâmetros há a necessidade de determinar momentos de graus superiores, *skewness* e *kurtosis* são os momentos de terceira e quarta ordem de uma variável aleatória.

A Tabela 2 apresenta a relação entre os parâmetros e os momentos dos modelos usuais de fenômenos em engenharia apresentados anteriormente.

Tabela 2 – Momentos dos modelos de distribuição candidatos.

Distribuição	**Média**	**Desvio Padrão**
Uniforme	$\frac{1}{2}(b+a)$	$\frac{1}{12}(b-a)^2$
Normal	μ	σ
Log-Normal	$e^{\left(\mu+\frac{\sigma^2}{2}\right)}$	$\sqrt{(e^{\sigma^2}-1)\cdot e^{2\mu+\sigma^2}}$
Logistico	μ	$\frac{s^2\pi^2}{3}$
Gamma	$\frac{\alpha}{\beta}$	$\frac{\sqrt{\alpha}}{\beta}$
Weibull	$\lambda\cdot\Gamma\left(1+\frac{1}{k}\right)$	$\lambda\cdot\sqrt{\Gamma\left(1+\frac{2}{k}\right)-\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2}$
Gumbel	$\mu+\beta\cdot\gamma$	$\frac{\beta\cdot\pi}{3}$

Fonte: próprio autor.

É importante mencionar que o modelo está sendo aproximado para a amostra, a qual possui um erro intrínseco em relação a sua população total. Ainda assim, este é um método bastante preciso e utilizado. A depender do modelo de distribuição, a obtenção analítica de seus parâmetros em função dos momentos da amostra pode não ser direta e o uso de algum método de otimização é necessário (WOOLDRIDGE, 2001).

2.2.2 Método da Máxima Verossimilhança

Diferente do método anterior onde os parâmetros do modelo são ajustados para se adequarem à amostra de dados, neste método o modelo é ajustado de forma que uma amostragem futura do modelo tenha uma chance ótima de coincidir com os valores amostrais já observados (WALPOLE, 2011).

O método propõe adequar o modelo à população, diferente do método dos momentos que visa a adequação aos valores amostrais. Ressalta-se que há um erro intrínseco entre a população total e seus dados amostrais, qualquer que seja o seu tamanho.

Seja x_i o i-ésimo de um total de n valores amostrais, e seja $f(x|\theta)$ a PDF do modelo, de parâmetros θ, aplicada no ponto x. O funcional de verossimilhança $\mathcal{L}(\theta)$ a ser maximizado (Equação 2.1) é o produtório da PDF em todos os pontos amostrais, ou seja:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{2.1}$$

Assim, o método da máxima verossimilhança busca o vetor θ que maximize o funcional de verossimilhança, é comum aplicar o logaritmo natural sobre a função de maximização para simplificar o termo produtório (Equação 2.2) (PAWITAN, 2001). O resultado apresenta benefícios pois a maioria das PDF dos modelos teóricos de distribuição são funções exponenciais, para muitas das quais é possível encontrar uma expressão analítica de seus parâmetros em função dos valores amostrais.

Evidencia-se que o artifício de aplicar o logaritmo natural só é possível pois a função logarítmica é estritamente crescente para bases superiores à unitária.

$$\max_{\theta} \mathcal{L}(\theta) = \max_{\theta} \sum_{i=1}^{n} \ln\left[f(x_i|\theta)\right]. \tag{2.2}$$

Em termos estatísticos, este método é um dos mais eficientes, por alcançar o limite inferior da variância entre os estimadores não-viesados de um parâmetro determinístico. Ou seja, qualquer estimador não-viesado terá um erro médio de estimação maior que o estimador de máxima verossimilhança (ANG; TANG, 2007).

2.3 Testes de Aderência

O ajuste de um modelo estatístico não garante que ele é uma boa representação dos valores observados, garante apenas que os melhores parâmetros foram determinados, dadas as características do modelo. Após a parametrização, testes de aderência são usados para avaliar a fidelidade do ajuste realizado no modelo de distribuição hipotético.

A qualidade do ajuste do modelo hipotético é estimada por um valor denominado valor observado. Os testes de aderência realizam um teste de hipótese, onde avaliam se o valor hipotético (valor observado) referente ao ajuste está abaixo do valor esperado. O valor esperado, por sua vez, quantifica o valor máximo admissível que o modelo hipotético pode apresentar de forma a ser fiel à amostra de dados.

Os testes são pautados num nível de significância α, que pode ser visto como a margem de erro admissível do teste, usualmente adotado como 5% em problemas de engenharia. Esse

valor significa que existe não mais do que 5% de chances de que um bom modelo seja rejeitado no teste, ou seja, a adoção de níveis de significância maiores indica a imposição de testes mais rigorosos. Encontra-se na literatura valores usuais entre 1% e 10%, a depender do nível de precisão requerida e da natureza dos dados da amostra. Dentre os testes de aderência amplamente difundidos na literatura, destacam-se os três descritos a seguir.

2.3.1 Teste Chi-Quadrado

O Teste Chi-Quadrado avalia a qualidade do ajuste do modelo por meio da comparação entre sua PDF e a frequência de ocorrências da amostra (histograma de densidade de probabilidade), sendo portanto sensível ao número de intervalos em que o histograma é dividido. A formulação de seu valor esperado é apresentada na Equação 2.3:

$$\chi_o^2 = \sum_{i=1}^{k} \frac{\left(O_i - n \cdot \left(CDF\left(x_{i+1}\right) - CDF\left(x_i\right)\right)\right)^2}{n \cdot \left(CDF\left(x_{i+1}\right) - CDF\left(x_i\right)\right)} \tag{2.3}$$

sendo k o número total de intervalos do histograma e O_i o total de observações no i-ésimo intervalo do histograma de ocorrências. O termo no denominador fornece o total de indivíduos observados, sendo então o produto do número de observações totais n com a probabilidade do i-ésimo intervalo do histograma (este por sua vez é a diferença entre as funções de probabilidade acumulada nos extremos do intervalo do histograma).

A distribuição da estatística χ_o^2 se aproxima do modelo de distribuição Chi-Quadrado conforme o tamanho da amostra cresce. O valor esperado do teste é determinado a partir da distribuição Chi-Quadrado com $p - 1$ graus de liberdade, sendo p o número de parâmetros da distribuição a ser testada.

2.3.2 Teste Kolmogorov-Smirnov

O teste de aderência Kolmogorov-Smirnov realiza uma comparação entre o modelo e a amostra pelo cálculo da máxima distância entre a função de probabilidade acumulada e o histograma cumulativo, fornecendo um resultado mais confiável por não depender do número de intervalos do histograma, manipulando os dados de forma direta. A formulação de seu valor observado é apresentada na Equação 2.4:

$$ks_o = \max_{1 \leq i \leq n} \parallel \frac{i}{n} - CDF(x_i) \parallel \tag{2.4}$$

sendo x_i o i-ésimo termo da amostra ordenada e $\frac{i}{n}$ a função cumulativa da amostra, razão entre a ordem da observação e o número total de amostras.

O valor esperado da estatística de teste Kolmogorov-Smirnov, para amostras com menos de 50 indivíduos, costuma ser apresentado por meio de valores tabelados. Para amostras com mais de 50 elementos, o valor esperado pode ser definido por uma função assintótica.

2.3.3 Teste Anderson-Darling

Semelhante ao teste Kolmogorov-Smirnov, o teste de aderência Anderson-Darling avalia o modelo ajustado em termos de densidade acumulada, sua diferença está em realizar a quantificação a partir do somatório de distâncias quadradas ponderadas. Sua apresentação formal é apresentada na Equação 2.5:

$$A_o^2 = -n - \frac{1}{n}\sum_{i-1}^{n}\left[(2i-1)\cdot\ln\left(CDF\left(x_i\right)\right) + \left(2\left(n-i\right)+1\right)\cdot\ln\left(1-CDF\left(x_i\right)\right)\right] \quad (2.5)$$

sendo x_i o i-ésimo termo de um conjunto com n valores amostrais.

A estatística de teste A_o assume que a população possui distribuição normal e seus parâmetros são conhecidos. Caso a média ou variância da população tenha sido obtida por meio de observações amostrais, ou o modelo de distribuição a ser testado não seja normal, uma modificação deve ser realizada na estatística de teste, a depender do modelo utilizado.

2.4 Critérios de Informação

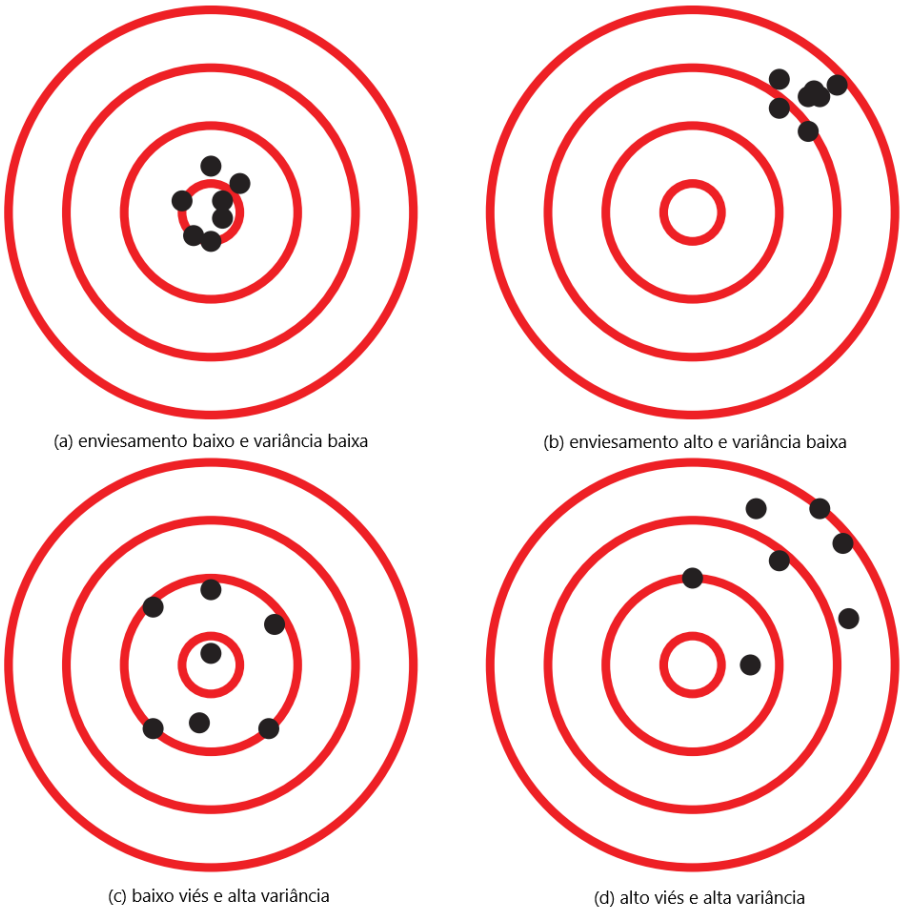
Sempre que se discute inferência ou previsão de dados, deve ser dada atenção aos erros inerentes a estes processos: erros de viés, erros de variância e erros irredutíveis (este último será discutido posteriormente).

O viés é a diferença entre os valores gerados pelo modelo substituto e o valor real da população ou que se tenta prever. Um modelo com alto viés absorve pouca informação dos dados que ajustaram o modelo, resultando em modelos mais simples e com poucos parâmetros.

Variância é a variabilidade da previsão do modelo para um determinado ponto de dados ou um valor que nos informa a distribuição de nossos dados. Um modelo com alta variância absorve muita informação dos dados que ajustaram o modelo e não generaliza sobre os dados que não viu antes.

A Figura 1 ilustra graficamente as combinações de enviesamento e variância. A compensação entre viés e variância (Bias-Variance Tradeoff) se refere ao fato de que ao tentar fazer uma previsão estatística, há uma compensação entre a exatidão da previsão e sua precisão, ou equivalentemente entre seu enviesamento e variância (DOROUDI, 2020). Obter uma compreensão adequada desses erros ajuda não apenas a construir modelos precisos, mas também a evitar os erros de sobreajuste (*overfitting*) e subajuste (*underfitting*).

Figura 1 – Combinações de enviesamento e variância.



Fonte: Doroudi (2020), editado.

Uma solução para avaliar a compensação entre viés e variância dos modelos estatísticos ajustados, que se sobressai ao teste de aderência por não se basear em testes de hipótese, é o Critério de Informação de Akaike (AIC).

Baseado no funcional de verossimilhança (Equação 2.1), este critério busca maximizar o logaritmo da verossimilhança enquanto minimiza o número de parâmetros do modelo, sua fórmula é exposta na Equação 2.6:

$$AIC = -2 \cdot \ln\left[\mathcal{L}(\theta)\right] + 2 \cdot p \tag{2.6}$$

sendo p o número de parâmetros do modelo e $\mathcal{L}$ o funcional de verossimilhança para parâmetros θ.

Existem outros critérios de informação, semelhantes ao AIC. Destaca-se o Critério de Informação Bayesiano (BIC), sua formulação é expressa na Equação 2.7:

$$BIC = -2 \cdot \ln\left[\mathcal{L}(\theta)\right] + \ln(n) \cdot p \tag{2.7}$$

sendo p o número de parâmetros do modelo, n o número de dados amostrais e $\mathcal{L}$ o funcional de verossimilhança para os parâmetros θ.

O BIC costuma penalizar distribuições com mais parâmetros de forma mais rígida, mas depende do termo adicional do tamanho amostral e sua relação com o número de parâmetros da distribuição.

Note-se que os critérios de informação servem apenas para realizar comparações entre modelos, de forma que, quanto menor o seu valor, melhor o modelo deve se adequar aos dados. Estas métricas fornecem uma análise qualitativa da aderência, sendo útil em uma pré-seleção de modelos estatísticos mais adequados, por meio do ranqueamento de uma lista de modelos candidatos.

2.5 Quantificação de Incertezas no Processo de Aquisição de Dados

A aquisição dos dados amostrais para posterior inferência estatística deve ser realizada com cuidado quanto à sua precisão. O registro de dados imprecisos diminui a acurácia no cálculo dos momentos da amostra, podendo originar uma análise estatística inconclusiva ou incorreta.

De forma geral, a margem de erro δf do valor de uma função f propagada a partir das suas n variáveis x_i é avaliada por meio da expressão da Equação 2.8, originada a partir da expansão em série de Taylor de primeira ordem da função f:

$$\delta f = \sum_{i=1}^{n} \left| \frac{\partial f}{\partial x_i} \right| \delta x_i \tag{2.8}$$

onde δx_i corresponde à margem de erro da variável x_i.

Desta forma, assumindo que todos os dados foram registrados com uma mesma precisão δx, a margem de erro propagada para a média, variância e desvio padrão da amostra podem ser avaliados, respectivamente, por meio das Equações 2.9, 2.10 e 2.11:

$$\delta \mu = \delta x \tag{2.9}$$

$$\delta \nu = \sum_{i=1}^{n} \left| \frac{2\left(x_i - \mu\right)}{n-1} \right| \cdot \delta x \tag{2.10}$$

$$\delta \sigma = \sum_{i=1}^{n} \left| \frac{x_i - \mu}{n-1} \right| \cdot \frac{\delta x}{\sigma}. \tag{2.11}$$

Observa-se pelas equações acima que a média amostral possui uma incerteza semelhante à dos dados amostrais. A margem de erro da variância é amplificada de acordo com o número de elementos da amostra e de sua variância em torno do seu valor médio, já a incerteza no desvio padrão é semelhante à incerteza da variância mas atenuada pelo seu próprio valor.

2.6 Imposição de Limites num Modelo de Distribuição Teórico

Grande parte dos modelos contínuos de distribuição de probabilidades não possui valores limites, ou seus limites não condizem com a natureza da variável estudada. Assim, por mais que um determinado modelo emule o comportamento de uma variável aleatória, o comportamento de suas caudas pode ser decisivo no estudo, sendo adequado impor limites nos valores da variável. A título de ilustração, destaca-se a tolerância preconizada pela API TR 5C3 (2008) para o valor do diâmetro externo de tubulares de revestimento. Segundo o documento, pode-se ter uma variação de -0,5% a 1,0% em torno do valor nominal da variável.

Ao ser imposto um limite $[a, b]$ para os valores de uma variável aleatória de distribuição X, de CDF F_X e PDF f_X, temos que a probabilidade total dos eventos deixa de ser unitária (Equação 2.12), o que não é possível para distribuição de probabilidades segundo os axiomas da teoria da probabilidade:

$$\int_a^b f_X \, dx = 1 - [F_X(b) - F_X(a)] \tag{2.12}$$

Uma solução proposta por Du e Hu (2012), aplicada em rotinas de confiabilidade estrutural, é utilizar uma constante multiplicadora que corrija a probabilidade total para seu valor unitário. A Equação 2.13 apresenta a PDF $\tilde{f}$ do modelo corrigido proposto:

$$\tilde{f}(x) = \begin{cases} 0 & \text{se } x < a \\ \frac{1}{F_X(b) - F_X(a)} \cdot f_X(x) & \text{se } a \leq x \leq b \\ 0 & \text{se } x > b \end{cases}. \tag{2.13}$$

A CDF $\tilde{F}$ do modelo corrigido proposto pode ser avaliada por meio da Equação 2.14, avaliada a partir da aplicação de integral sobre os termos da Equação 2.13:

$$\tilde{F}(x) = \begin{cases} 0 & \text{se } x < a \\ \frac{F_X(x) - F_X(a)}{F_X(b) - F_X(a)} & \text{se } a \leq x \leq b \\ 1 & \text{se } x > b \end{cases}. \tag{2.14}$$

Observa-se que a imposição de limites numa distribuição hipotética pode acarretar em resultados indesejados, como a modificação dos momentos da distribuição. Sugere-se realizar esta imposição em casos onde a região negligenciada tenha uma probabilidade de ocorrência pouco relevante:

$$1 - [F_X(b) - F_X(a)] < 1\%. \tag{2.15}$$

Caso a variável possua valores limites com uma probabilidade de ocorrência superior à condição imposta na Equação 2.15, é provável que os dados, em si, não respeitem estes limites. Assim, as caudas pronunciadas não são um mero produto da distribuição teórica adotada, mas um comportamento dos dados da amostra.

Para a utilização desta metodologia, é preferível realizar a parametrização dos modelos estatísticos hipotéticos pelo Método da Máxima Verossimilhança pois, como demonstrado na Equação 2.16, os parâmetros da distribuição com a imposição de limites são os mesmos da distribuição sem a imposição de limites, desde que todos os dados amostrais pertençam ao limite imposto.

$$
\begin{aligned}
\max_{\theta} \tilde{\mathcal{L}}(\theta) &= \max_{\theta} \sum_{i=1}^{n} \ln\left[\tilde{f}(x_i|\theta)\right] \\
&= \max_{\theta} \sum_{i=1}^{n} \ln\left[\frac{1}{F_X(b) - F_X(a)} \cdot f_X(x_i|\theta)\right] \\
&= \max_{\theta} \sum_{i=1}^{n} \left[\ln\left[f(x_i|\theta)\right] - \ln\left[F_X(b) - F_X(a)\right]\right] \\
&= \max_{\theta} \left[-n \cdot \ln\left[F_X(b) - F_X(a)\right] + \sum_{i=1}^{n} \ln\left[f(x_i|\theta)\right]\right] \\
&= \max_{\theta} \sum_{i=1}^{n} \ln\left[f(x_i|\theta)\right] \\
&= \max_{\theta} \mathcal{L}(\theta).
\end{aligned}
\tag{2.16}
$$

Evidencia-se pelo resultado acima que o funcional de verossimilhança é maximizado para os mesmos parâmetros, independentemente da imposição de limites para a variável, motivo pelo qual foi sugerido esta forma de correção na Equação 2.13.

3 PREVISÃO DE DADOS POR MODELOS ARIMA

De forma geral, nem todas as séries de dados temporais são previsíveis. Se todos os dados da série são independentes, então não existem modelos que garantam a previsão dos valores da série temporal. Por outro lado, caso os dados da série sejam independentes, indica-se realizar uma inferência estatística sobre os dados.

A área de previsão de dados assume que uma observação X_t, no tempo t, é passível de ser estimada por um valor $\widehat{X}_t$, determinado por meio de um modelo matemático que utiliza informações anteriores, isto é, as observações passadas X_k, com $k < t$.

Um modelo de previsão ideal deve reduzir o erro redutível (Equação 3.1) entre o valor estimado e o dado real, normalmente determinando seus parâmetros de forma a minimizar este valor segundo alguma métrica de erro (JAMES et al., 2013), sendo comum adotar a soma dos quadrados das diferenças. Em um modelo ideal, esta diferença entre o valor real e o estimado é uma série com média zero, independente e identicamente distribuída (IID), denominada ruído branco.

$$\varepsilon_t = X_t - \widehat{X}_t. \tag{3.1}$$

Para verificação das condições de IID do ruído branco, pode-se computar a autocorrelação da série. Os gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF) também são utilizados no método de previsão ARIMA e, portanto, serão apresentados a seguir.

3.1 Quantificação da Dependência dos Termos de uma Série Temporal

A dependência regular entre termos de uma série temporal pode ser avaliada por diferentes métricas, são elas a autocovariância, a autocorrelação e a autocorrelação parcial. Tais métricas são expostas a seguir.

3.1.1 Autocovariância

A dependência entre um termo X_t da série e o termo X_{t+h} pode ser computada numericamente através da covariância (Cov) entre eles. Como a covariância é avaliada entre termos da mesma série, diz-se que foi realizada a autocovariância, sendo sua formulação apresentada na Equação 3.2:

$$\gamma_X(h) = Cov(X_t, X_{t+h}) = E[(X_t - \mu) \cdot (X_{t+h} - \mu)] \tag{3.2}$$

onde E[] é o valor médio ou esperança matemática.

Para valores amostrais, como a aplicação em questão, temos a fórmula transformada para Equação 3.3:

$$\widehat{\gamma}_X(h) = \frac{\sum_{k=1}^{n-h}(X_k - \mu) \cdot (X_{k+h} - \mu)}{n - h - 1}. \tag{3.3}$$

Uma propriedade da autocovariância é que ela não excede o valor da variância da série temporal, desta forma, o valor da autocovariância pode ser normalizado.

3.1.2 Autocorrelação

A autocorrelação é estimada ao se reduzir a variância do valor da sua autocovariância, como presente na Equação 3.4:

$$\rho_X(h) = \frac{Cov(X_t, X_{t+h})}{VAR[X]} = \frac{E[(X_t - \mu) \cdot (X_{t+h} - \mu)]}{\sigma_X^2}. \tag{3.4}$$

Como a autocovariância é menor ou igual à variância de uma série, temos que a autocorrelação é compreendida entre -1 e 1. Uma correlação positiva ($\rho_X(h) > 0$) indica que os termos X_t e X_{t+h} estão crescendo ou decrescendo juntos, enquanto que uma correlação negativa ($\rho_X(h) < 0$) indica que quando um dos termos está crescendo, o outro está decrescendo.

A norma da correlação $|\rho_X(h)|$ indica quão forte a correlação entre os termos é. Quanto mais próximo de 1, maior a correlação entre os termos. Uma correlação próxima de 0 significa que não é possível associar uma tendência linear de crescimento (ou decrescimento) no termo X_{t+h} graças ao crescimento (ou decrescimento) do termo X_t.

Para valores amostrais, temos a fórmula modificada da Equação 3.5:

$$\widehat{\rho}_X(h) = \frac{\sum_{k=1}^{n-h}(X_k - \mu) \cdot (X_{k+h} - \mu)}{\sum_{k=1}^{n}(X_k - \mu)^2} \cdot \frac{n - 1}{n - h - 1}. \tag{3.5}$$

Numa série de dados com valores significativos de autocorrelação para atrasos $\rho_X(1)$, $\rho_X(2)$, $\rho_X(3)$ até $\rho_X(h)$, temos que um termo X_{t+h} é relacionado com cada um de seus termos anteriores. Assim, é interessante avaliar a correlação entre o termo X_{t+h} e seu atraso h (elemento X_t) excluindo-se as relações indiretas com os demais termos intermediários. Temos assim a autocorrelação parcial.

3.1.3 Autocorrelação Parcial

Autocorrelações parciais medem a dependência linear de duas variáveis após remover o efeito de outras variáveis que afetam as duas variáveis iniciais. A formulação para o cálculo da autocorrelação parcial não é direta e requer a solução de múltiplos sistemas de equações lineares.

De forma geral, a formulação da autocorrelação parcial pode ser apresentada pela Equação 3.6:

$$\tau_X(h) = \frac{Cov(Y_t, Y_{t+h})}{\sqrt{VAR[Y_t] \cdot VAR[Y_{t+h}]}} \tag{3.6}$$

onde Y_t é o resíduo obtido ao ajustar um modelo linear multivariado na série [X_{t+1}, X_{t+2}, ..., $X_{t+(h-1)}$] de forma a prever o termo X_t, representando a variância residual em X_t após remover a influência dos demais termos; Y_{t+h} é obtido de forma análoga ao ajustar um modelo linear multivariado na série [X_{t+1}, X_{t+2}, ..., $X_{t+(h-1)}$] de forma a prever o termo X_{t+h}.

3.2 Modelos Autorregressivos, Integrados e de Médias Móveis (ARIMA)

Modelos do tipo ARIMA agrupam uma série de estruturas lineares com forte autocorrelação de dados, encontradas em séries temporais estáveis com padrões previsíveis.

Este conjunto de modelos é formado pela combinação independente de modelos Autorregressivos (AR), Integrados (I) e de Médias Móveis (MA). A seguir é apresentada a formulação dos modelos.

3.2.1 Modelos Autorregressivos (AR)

Estes modelos são fundamentados na relação linear direta entre uma certa quantidade de observações passadas e a observação futura. O hiperparâmetro p do modelo identifica quantas observações passadas interferem no valor da observação seguinte.

Seja X_t a observação no instante t, seu valor pode ser estimado segundo a Equação 3.7:

$$(X_t - \mu) = \sum_{k=1}^{p} \phi_k \cdot (X_{t-k} - \mu) + \varepsilon_t \tag{3.7}$$

onde μ é o valor médio da série temporal, os termos ϕ_k são as constantes da combinação linear e os termos ε_k representam o ruído branco do modelo.

Uma propriedade importante dos modelos autorregressivos é que a autocorrelação parcial é nula para atrasos superiores à $p + 1$, onde p é o hiperparâmetro do modelo. A Equação 3.8 resume o resultado:

$$\tau_X(h) \begin{cases} = 0 & \text{se } h > p \\ \neq 0 & \text{se } h \leq p \end{cases}. \tag{3.8}$$

Séries autorregressivas têm, em geral, coeficientes de autocorrelação que decaem conforme o atraso entre os termos aumenta, mas nunca chega a zero. Um termo X_{t+h} está sempre relacionado ao termo atrasado X_t, mesmo para um atraso h longo.

Assim, uma série autorregressiva possui uma memória relativamente longa, pois o valor atual de uma série está correlacionado com todos os anteriores, embora que com coeficientes decrescentes.

Essa propriedade significa que é possível escrever uma série AR como uma função linear de todos os termos anteriores, com pesos que tendem a zero com o atraso. Logo, uma série AR não pode representar muito bem séries curtas, onde o valor atual da série é correlacionado apenas com um pequeno conjunto de valores anteriores.

Um modelo de previsão semelhante ao AR, mas que possui a propriedade de memória curta é o Modelo de Médias Móveis, do inglês *Moving Average Model* (MA). Uma série MA pode ser vista como uma combinação linear de um número finito, e geralmente pequeno, de seus elementos anteriores, isto é, um elemento X_t só é influenciado por q elementos anteriores, os termos x_{t-k}, onde $k > q$, não influenciam os termos atuais da série.

3.2.2 Modelos de Médias Móveis (MA)

Diferente dos modelos AR onde as observações futuras são estimadas com uma função direta das observações anteriores, modelos MA estimam observações futuras por meio dos termos da série de ruído branco das observações anteriores. A Equação 3.9 apresenta a formulação do modelo:

$$X_t = \mu + \sum_{k=1}^{q} \theta_k \cdot \varepsilon_{t-k} + \varepsilon_t \tag{3.9}$$

onde μ é o valor médio da série temporal, os termos θ_k são as constantes da combinação linear e os termos ε_k representam o ruído branco do modelo.

O principal ganho do modelo MA, em relação ao modelo AR, é que a autocorrelação de seus elementos é nula para atrasos superiores à $q + 1$, onde q é o hiperparâmetro do modelo que identifica quantos termos passados influenciam a observação seguinte. A Equação 3.10 remete ao resultado:

$$\rho_X(h) \begin{cases} = 0 & \text{se } h > q \\ \neq 0 & \text{se } h \leq q \end{cases}. \tag{3.10}$$

3.2.3 Modelos Autorregressivos de Médias Móveis (ARMA)

Um modelo mais genérico, criado a partir da combinação dos dois modelos anteriores, forma o Modelo Autorregressivo de Médias Móveis (ARMA), capaz de prever observações futuras a partir de sua relação com observações e ruídos anteriores, sua formulação é a combinação

das Equações 3.7 e 3.9, como presente na Equação 3.11, pois admite-se que os modelos AR e MA são independentes:

$$(X_t - \mu) = \sum_{j=1}^{q} \theta_j \cdot \varepsilon_{t-j} + \sum_{k=1}^{p} \phi_k \cdot (X_{t-k} - \mu) + \varepsilon_t. \tag{3.11}$$

Uma condição necessária para que uma série temporal seja bem modelada pelo modelo ARMA é que essa seja estacionária, isto é, que a média e a variância da série se mantenham constantes, independentemente do intervalo de tempo a ser analisado. Muitas vezes, as séries temporais podem apresentar uma tendência de crescimento ou decrescimento, impedindo a aplicação direta do modelo ARMA.

3.2.4 Modelos Autorregressivos, Integrados de Médias Móveis (ARIMA)

O Modelo Integrado (I) soluciona a tendência de séries temporais de crescimento ou decrescimento monótono, ampliando a quantidade de séries que podem ser modeladas pelo modelo ARMA.

Séries temporais criadas por meio da diferença entre os termos vizinhos de uma série são estacionárias, caso a série temporal original seja monótona. O parâmetro d do Modelo Integrado remete ao número de diferenciações realizadas sobre os dados originais, até que se torne estacionário.

Desta forma, a generalização dos modelos para identificação de padrões de resposta de séries temporais é denominado ARIMA, por ser a combinação dos modelos Autorregressivo, Integrado e de Médias Móveis. Este modelo possui os hiperparâmetros p, d e q relativos aos seus modelos originais e independentes AR, I e MA. A determinação de seus hiperparâmetros caracteriza o modelo utilizado para previsão dos dados.

A escolha dos hiperparâmetros é realizada mediante os indicadores de autocorrelação da série de dados. Os gráficos de autocorrelação (ACF) e de autocorrelação parcial (PACF) determinam o valor do hiperparâmetro d que garanta que os dados sejam estacionários, como também estimam os valores de p e q.

Uma vez determinado o hiperparâmetro d, a escolha dos hiperparâmetros p e q pode ser estabelecida por meio de uma otimização de inteiros. Os gráficos de ACF e PACF indicam valores de p e q, enquanto que uma busca de grade sobre os valores indicados otimiza seus valores.

O método de busca de grade é um método exaustivo de busca por um ponto ótimo. A métrica utilizada para avaliar o ajuste do modelo é a minimização do Critério de Informação de Akaike (AIC).

Assim, como apresentado no capítulo sobre inferência de dados, o AIC se baseia na função de verossimilhança, maximizando o logaritmo da verossimilhança enquanto minimiza

o número de parâmetros do modelo. A fórmula do AIC como métrica de ajuste de modelos ARIMA é exposta na Equação 3.12:

$$AIC = -2 \cdot log(\mathcal{L}) + 2 \cdot (p + q + k + 1) \tag{3.12}$$

onde $\mathcal{L}$ é a função de verossimilhança, p e q são os hiperparâmetros do modelo e $k = 0$ se a média da série amostral for nula, $k = 1$ caso contrário.

Após a parametrização adequada do modelo ARIMA e sua consequente previsão de dados, métricas que avaliam a precisão dos dados previstos são usadas para quantificar a capacidade do modelo em prever séries de dados. Algumas destas métricas são apresentadas a seguir.

3.3 Medidas de Acurácia

De forma geral, as métricas para previsão de séries temporais comparam um conjunto de observações reais (X_t) com seus valores estimados ($\widehat{X}_t$), normalmente por meio da diferença entre os dois conjuntos ou por uma análise de correlação entre eles.

3.3.1 Erro Absoluto Médio (MAE)

Uma métrica simples, da qual se derivam diversas outras técnicas, é o cálculo do valor médio entre as diferenças absolutas das observações reais e seus valores estimados, como presente na Equação 3.13:

$$MAE = \frac{\sum_{t=1}^{n} |X_t - \widehat{X}_t|}{n}. \tag{3.13}$$

Desta, derivam-se as métricas RMSE e MAPE, apresentadas a seguir.

3.3.2 Erro Absoluto Médio Percentual (MAPE)

A métrica MAPE é uma medida popular de acurácia de previsão de dados e é definida como a média percentual das diferenças absolutas normalizadas por seus valores reais, conforme apresentado na Equação 3.14:

$$MAPE = \frac{100}{n} \cdot \sum_{t=1}^{n} \left| \frac{X_t - \widehat{X}_t}{X_t} \right|. \tag{3.14}$$

A principal vantagem desta métrica é estar em sua forma percentual, de forma que esta pode ser comparada entre diferentes séries de dados. Esta medida possibilita a análise mesmo quando o tamanho da série amostral prevista é desconhecido, ao contrário de medidas em valores absolutos.

3.3.3 Raiz Quadrada do Erro Quadrático Médio (RMSE)

A raiz quadrada do erro quadrático médio (RMSE) é uma maneira padrão de medir o erro de um modelo na previsão de dados quantitativos. Formalmente, é definido segundo a Equação 3.15:

$$RMSE = \sqrt{\frac{1}{n} \cdot \sum_{t=1}^{n} \left(X_t - \widehat{X}_t\right)^2}. \tag{3.15}$$

Diferente do MAPE, que é uma métrica linear, a RMSE absorve a variância do erro de previsão. A desvantagem da métrica é seu valor absoluto ser proporcional à ordem de grandeza das séries de dados.

3.3.4 Coeficiente de Variação da Raiz Quadrada do Erro Quadrático Médio (CVRMSE)

Embora não haja meios consistentes de normalização da RMSE na literatura, indica-se a normalização por meio da média das observações reais, surgindo assim o coeficiente de variação da RMSE, descrito formalmente segundo a Equação 3.16:

$$CVRMSE = \frac{\sqrt{n \cdot \sum_{t=1}^{n} \left(X_t - \widehat{X}_t\right)^2}}{\sum_{t=1}^{n} \left(X_t\right)}. \tag{3.16}$$

Uma análise conjunta das métricas CVRMSE e MAPE deve garantir uma boa avaliação da previsão. Para mais observações e comparações sobre medidas de acurácia, cita-se Armstrong e Collopy (1992).

4 ESTUDOS DE CASO DE INFERÊNCIA E PREVISÃO DE DADOS COMO SUPORTE À ANÁLISE DE INTEGRIDADE DE REVESTIMENTOS

São apresentados neste capítulo os resultados compilados de inferência estatística e de previsão de dados, utilizando as técnicas apresentadas nos capítulos anteriores em séries de dados reais.

4.1 Inferência Estatística de Dados de Produção de Tubulares

No Anexo F da norma API TR 5C3 (2008) apresenta-se um vasto estudo estatístico para calibração de parâmetros e validação de um modelo de colapso em Estado Limite Último, proposto por Klever e Tamano (2006). As séries de dados utilizadas incluem dados de tubos de revestimento fabricados entre 1977 e 2004, por diversos fabricantes.

A fim de auxiliar a análise de confiabilidade estrutural e conhecer mais especificamente dados de produção de um material específico, permitindo projetos mais robustos e confiáveis, aplicam-se as técnicas de inferência estatística estudadas em séries históricas de dados de produção de tubos de revestimento de poços.

Por questões de proteção dos dados, estes foram modificados por um fator multiplicativo constante e uma variação aleatória foi somada ao valor dos dados, de forma que continuem confidenciais, preservando a fonte, mas sua caracterização se aproxime o máximo possível dos dados reais.

4.1.1 Dados de Produção Analisados

Os dados amostrais utilizados são o registro de 50 tubos de revestimento, aleatoriamente escolhidos pelo fabricante no período de referência. No processo de controle de qualidade, são realizadas várias medições em cada tubo, em diversos trechos ao longo deste.

Desta forma, constam medições em seções a cada 15 cm de comprimento do tubo, resultando em, aproximadamente, 73 seções por tubo, com pequenas variações. Em cada seção medem-se vários valores de espessura e diâmetro, os valores mínimo, médio e máximo destas variáveis em cada seção são registrados.

Assume-se que os resultados obtidos para séries amostrais referentes aos 50 tubos sejam representativos do comportamento das variáveis no período de produção analisado.

Ainda, foram registrados duas séries de dados de características mecânicas dos tubos de revestimento. Como os ensaios para obtenção destes parâmetros são destrutivos, o tamanho da amostragem foi reduzido para 81 dados sobre resistência em cada série. Duas séries de dados foram geradas, uma com os registros de tensão de escoamento (YS) e outra com os registros de resistência à tração (UTS), aferidos em ensaios de espécimes de aço extraídos dos tubulares no mesmo período.

4.1.2 Análise de Dados de Caracterísiticas Geométricas: Amostragem Tubo a Tubo

Como cada tubo possui um registro de cerca de 73 valores amostrais de suas variáveis geométricas, realizou-se inicialmente uma inferência estatística individual sobre cada conjunto de dados.

Para ilustração dos resultados, foram expostos nas Tabelas 3 e 4 os dados de espessura de parede média (avgWT) dos tubos 2 e 18, e de diâmetro médio (avgOD) dos tubos 29 e 46 nas Tabelas 5 e 6, respectivamente.

Tabela 3 – Dados de espessura média do tubo 2.

32,19	32,07	32,42	32,24	32,14	31,90	31,88	31,84
31,81	31,97	31,75	31,68	31,95	31,73	32,01	31,99
32,03	31,90	31,91	32,01	31,90	31,91	31,80	31,92
31,82	32,00	31,81	32,00	31,91	31,93	31,96	31,97
31,79	31,89	32,01	31,88	31,80	31,86	31,98	31,75
31,98	31,70	31,85	31,74	31,70	31,94	31,95	31,76
31,82	31,77	31,93	32,00	31,97	31,77	31,79	31,87
31,99	31,78	31,95	31,83	31,81	31,88	31,77	31,88
31,78	31,78	31,74	31,92	31,89	32,00	31,96	32,08

Fonte: próprio autor.

Tabela 4 – Dados de espessura média do tubo 18.

31,63	31,54	31,92	31,44	31,68	31,49	31,31	31,29
31,42	31,41	31,45	31,15	31,25	31,24	31,25	31,38
31,36	31,40	31,22	31,28	31,15	31,31	31,19	31,19
31,11	31,39	31,31	31,23	31,32	31,26	31,40	31,48
31,41	31,32	31,33	31,28	31,28	31,45	31,28	31,44
31,39	31,44	31,51	31,27	31,28	31,37	31,55	31,50
31,47	31,35	31,31	31,40	31,31	31,56	31,54	31,43
31,64	31,49	31,40	31,48	31,35	31,42	31,65	31,44
31,39	31,37	-	-	-	-	-	-

Fonte: próprio autor.

Desta forma, seguindo a metodologia apresentada para inferência estatística sobre os dados, foram selecionadas as distribuições Uniforme, Normal, Log Normal, Logística, Gamma, Gumbel e Weibull como candidatas, abrangendo as distribuições bastante difundidas em problemas de engenharia, todas elas possuindo dois parâmetros.

Optou-se por utilizar o Método da Máxima Verossimilhança para o ajuste das distribuições, sendo os parâmetros obtidos por meio da aplicação do método sobre os dados de espessura média do tubo 2 listados na Tabela 7, como também os momentos das distribuições candidatas parametrizadas.

Tabela 5 – Dados de diâmetro médio do tubo 29.

536,26	535,55	536,36	536,81	535,64	535,73	536,24	536,09
536,13	537,01	535,97	536,25	536,06	536,58	536,42	536,52
536,76	536,76	537,44	536,66	537,42	537,74	536,43	537,30
537,49	537,47	536,90	536,76	536,53	536,72	536,49	537,29
536,91	537,16	536,78	537,27	536,77	536,45	537,27	536,67
536,53	536,32	537,51	536,36	536,89	536,36	536,12	537,46
536,91	536,68	536,28	536,78	536,83	537,36	536,66	536,99
537,10	536,89	-	-	-	-	-	-

Fonte: próprio autor.

Tabela 6 – Dados de diâmetro médio do tubo 46.

537,46	536,64	537,27	537,93	537,14	537,32	537,22	537,51
537,82	537,42	537,54	538,47	537,50	537,91	538,45	537,92
537,14	537,53	537,36	537,65	537,55	537,89	538,26	538,23
537,72	537,89	537,63	537,85	537,41	537,92	537,88	537,27
537,30	537,25	537,30	538,04	538,11	537,87	538,10	538,69
538,24	537,83	538,45	538,53	537,47	538,32	537,55	538,33
537,39	538,24	537,59	537,08	537,06	538,19	537,79	536,89
537,06	538,00	537,65	537,83	536,77	536,38	536,52	536,97
537,00	537,39	536,84	536,88	537,22	537,45	537,57	537,28
536,60	537,68	537,67	536,56	-	-	-	-

Fonte: próprio autor.

Tabela 7 – Modelos candidatos ajustados sobre os dados de avgWT do tubo 2.

Distribuição	**Parâmetro 1**	**Parâmetro 2**	**Média**	**Desvio Padrão**
Uniforme	31,680	0,740	32,050	0,214
Normal	31,901	0,130	31,901	0,130
Log Normal	0,004	31,901	31,901	0,129
Logístico	31,894	0,070	31,894	0,127
Gamma	60802,946	0,001	31,901	0,129
Gumbel	31,842	0,104	31,902	0,134
Weibull	189,677	31,971	31,875	0,215

Fonte: próprio autor.

Evidencia-se na Tabela 7 que as distribuições parametrizadas nem sempre possuem momentos idênticos ao da amostra de dados utilizada, de média 31,90 e desvio padrão 0,13, uma vez que o Método da Máxima Verossimilhança não impõe esta restrição.

As Tabelas 8, 9 e 10 resumem os resultados da parametrização dos modelos sobre os dados avgWT 18, avgOD 29 e avgOD 46, respectivamente.

Com os modelos candidatos ajustados, o teste de aderência Kolmogorov-Smirnov foi

Tabela 8 – Modelos candidatos ajustados sobre os dados de avgWT do tubo 18.

Distribuição	Parâmetro 1	Parâmetro 2	Média	Desvio Padrão
Uniforme	31,110	0,810	31,515	0,234
Normal	31,387	0,140	31,387	0,140
Log Normal	0,004	31,387	31,387	0,139
Logístico	31,380	0,076	31,380	0,138
Gamma	50745,907	0,001	31,387	0,139
Gumbel	31,322	0,119	31,391	0,153
Weibull	179,593	31,462	31,361	0,223

Fonte: próprio autor.

Tabela 9 – Modelos candidatos ajustados sobre os dados de avgOD do tubo 29.

Distribuição	Parâmetro 1	Parâmetro 2	Média	Desvio Padrão
Uniforme	535,550	2,190	536,645	0,632
Normal	536,708	0,495	536,708	0,495
Log Normal	0,001	536,708	536,708	0,495
Logístico	536,709	0,287	536,709	0,520
Gamma	1176249,903	0,000	536,708	0,495
Gumbel	536,459	0,491	536,742	0,629
Weibull	1164,070	536,954	536,688	0,591

Fonte: próprio autor.

Tabela 10 – Modelos candidatos ajustados sobre os dados de avgOD do tubo 46.

Distribuição	Parâmetro 1	Parâmetro 2	Média	Desvio Padrão
Uniforme	536,380	2,310	537,535	0,667
Normal	537,574	0,518	537,574	0,518
Log Normal	0,001	537,574	537,574	0,518
Logístico	537,578	0,300	537,578	0,545
Gamma	1076078,928	0,000	537,574	0,518
Gumbel	537,313	0,509	537,607	0,653
Weibull	1109,453	537,831	537,551	0,621

Fonte: próprio autor.

aplicado para avaliar se as distribuições candidatas se adéquam aos dados amostrais. A Tabela 11 apresenta o resultado do teste de aderência, assim como os critérios de informação AIC e BIC.

Para um nível de significância de 5%, o valor observado do teste Kolmogorov-Smirnov deve ser inferior à 0,16 para que uma distribuição candidata com dois parâmetros seja considerada como população original da amostra utilizada com 72 elementos. Segundo a Tabela 11, as distribuições Gumbel, Logística, Log Normal, Gamma e Normal passaram no teste, enquanto que as distribuições Weibull e Uniforme falharam, isto é, não há informações suficientes para afirmar que estas distribuições são capazes de originar a amostra.

Tabela 11 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do tubo 2.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Gumbel	0,084	Sim	-96,05	-91,49
Logístico	0,074	Sim	-90,54	-85,98
Log Normal	0,103	Sim	-86,26	-81,71
Gamma	0,103	Sim	-86,16	-81,61
Normal	0,103	Sim	-85,95	-81,40
Weibull	0,187	Não	-48,46	-43,90
Uniforme	0,457	Não	-39,36	-34,81

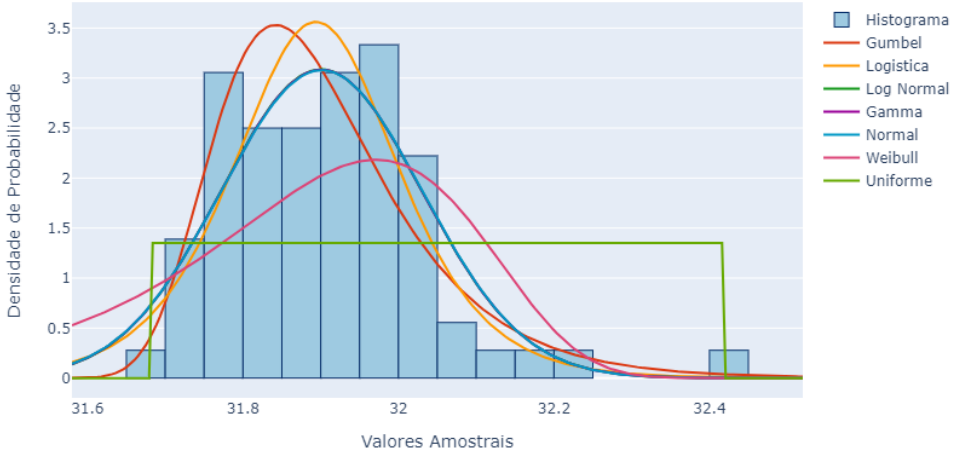
Fonte: próprio autor.

Quanto à escolha da melhor distribuição dentre as que passaram no teste, o teste Kolmogorov-Smirnov não é uma métrica ideal. Considera-se neste caso os critérios de informação AIC e BIC. Como todas as distribuições candidatas possuem o mesmo número de parâmetros, os dois critérios darão o mesmo resultado quanto à ordenação das distribuições.

Deste modo, a distribuição Gumbel é a melhor escolha para distribuição geradora dos dados de avgWT do tubo 2, seguida pela distribuição Logística, Log Normal, Gamma e Normal, nesta ordem. Ressalta-se que os critérios de informação não são capazes de avaliar se a distribuição é um bom ajuste, deve-se então utilizá-los em conjunto com os testes de aderência.

A Figura 2 corrobora os resultados do teste de aderência e critérios de informação, exibindo o histograma de densidade da amostra de avgWT do tubo 2 e a função de densidade de probabilidade das distribuições candidatas. As distribuições parametrizadas Log Normal, Gamma e Normal foram equivalentes.

Figura 2 – Histograma de avgWT do tubo 2 e suas distribuições candidatas.



Fonte: próprio autor.

Para os dados de avgWT do tubo 18, conforme resultados apresentados numericamente

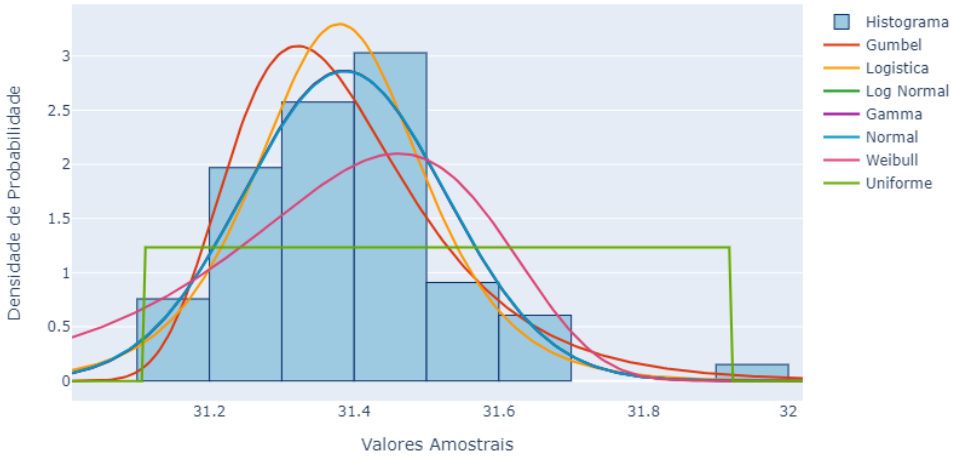
na Tabela 12 e visualmente na Figura 3, todas as distribuições candidatas passaram no teste de aderência, com exceção da distribuição Uniforme. O valor máximo para o teste Kolmogorov-Smirnov para essa amostra de dados é de 0,167, seu valor difere do máximo da amostra do tubo 2 por ser uma amostra menor, com 66 observações. Outra vez, as distribuições parametrizadas Log Normal, Gamma e Normal se comportaram de forma equivalente.

Tabela 12 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do tubo 18.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Gumbel	0,051	Sim	-73,03	-68,65
Logístico	0,065	Sim	-72,13	-67,75
Log Normal	0,083	Sim	-68,94	-64,48
Gamma	0,083	Sim	-68,86	-64,48
Normal	0,084	Sim	-68,69	-64,31
Weibull	0,150	Sim	-37,86	-33,48
Gumbel	0,151	Sim	-37,18	-32,80
Uniforme	0,370	Não	-23,82	-19,44

Fonte: próprio autor.

Figura 3 – Histograma de avgWT do tubo 18 e suas distribuições candidatas.



Fonte: próprio autor.

Um resultado interessante é que a ordem das distribuições é idêntica para o tubo 2 e tubo 18, segundo os critérios de informação AIC e BIC. Este resultado apoia a hipótese de que os diferentes conjuntos de dados de uma mesma variável podem ser inferidos por uma distribuição comum, uma vez que estes dados referem-se à produção de peças metálicas numa indústria automatizada e de alta precisão.

Para os dados de avgOD, o teste de aderência e os critérios de informação do tubo 29 indicam que a distribuição Normal é a que melhor se adéqua aos dados, seguida pelas

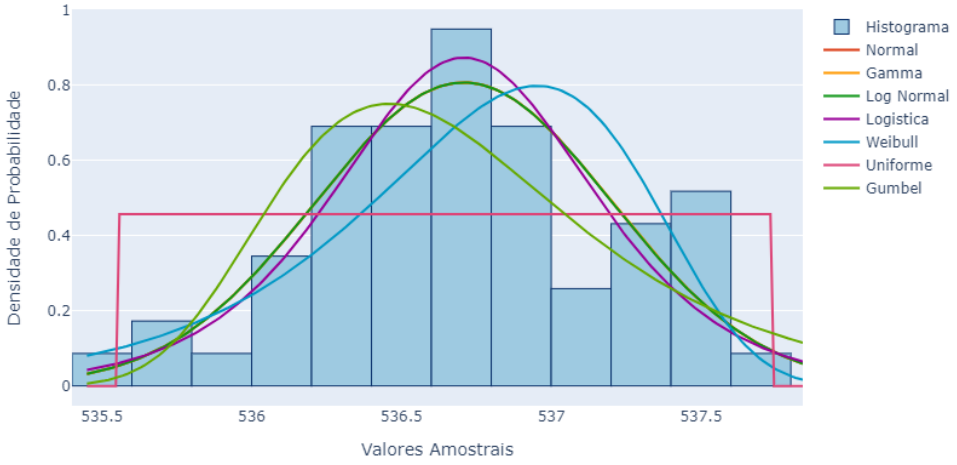
distribuições Gamma, Log Normal, Logística, Weibull, Uniforme e Gumbel, nesta ordem, como apresentado na Tabela 13 e exposto na Figura 4.

Tabela 13 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do tubo 29.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Normal	0,066	Sim	86,99	91,11
Gamma	0,066	Sim	86,99	91,12
Log Normal	0,066	Sim	87,00	91,12
Logístico	0,055	Sim	89,14	93,26
Weibull	0,127	Sim	91,92	96,04
Uniforme	0,103	Sim	94,93	99,05
Gumbel	0,094	Sim	96,36	100,48

Fonte: próprio autor.

Figura 4 – Histograma de avgOD do tubo 29 e suas distribuições candidatas.



Fonte: próprio autor.

Desta vez, as distribuições parametrizadas Normal, Gamma e Log Normal foram numericamente equivalentes. Todas as distribuições candidatas passaram no teste de aderência, o valor limite para o teste numa amostra de 58 observações é de 0,18.

De forma geral, uma razão para todas as distribuições candidatas passarem no teste de aderência se dá pelo número reduzido de observações na amostra analisada, de modo que o teste de aderência tem poucas informações e se torna menos restrito. Uma outra característica, específica para o teste Kolmogorov-Smirnov, é que o mesmo perde sua eficácia nas caudas, sendo mais competente em verificar a aderência na região em torno do centro da amostra.

A inferência sobre os dados avgOD do tubo 46 não foi diferente, a Tabela 14 mostra os resultados do teste de aderência e os critérios de informação, ao passo que a Figura 5 exibe o resultado de maneira gráfica.

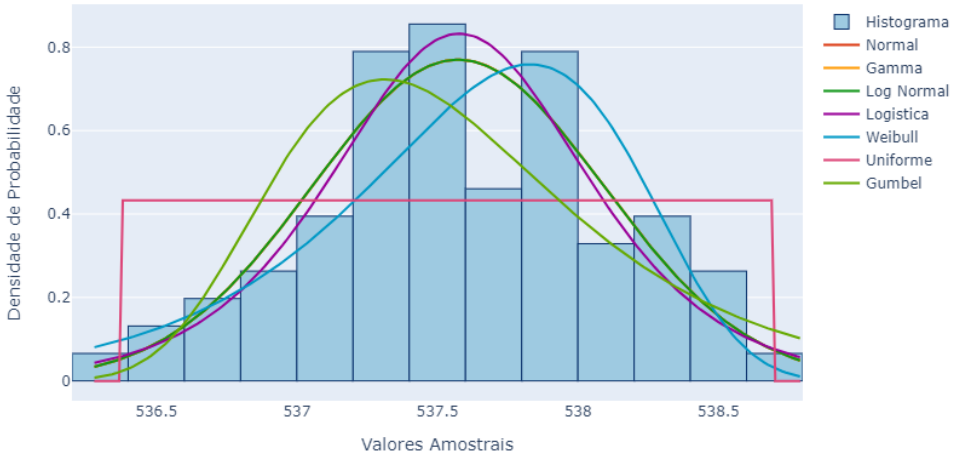
Assim como os resultados dos dados de avgOD do tubo 29, o tubo 46 aprovou todos os testes de aderência, como também a distribuição Normal é a que melhor se adéqua aos dados, seguida pelas distribuições Gamma, Log Normal, Logística, Weibull, Uniforme e Gumbel, nesta ordem.

Tabela 14 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do tubo 46.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Normal	0,036	Sim	119,76	124,42
Gamma	0,036	Sim	119,76	124,42
Log Normal	0,035	Sim	119,76	124,43
Logístico	0,036	Sim	122,66	127,32
Weibull	0,086	Sim	126,39	131,05
Uniforme	0,118	Sim	131,26	135,92
Gumbel	0,074	Sim	131,41	136,08

Fonte: próprio autor.

Figura 5 – Histograma de avgOD do tubo 46 e suas distribuições candidatas.



Fonte: próprio autor.

A inferência estatística sobre os dados expostos nas Tabelas 3, 4, 5 e 6 foi igualmente realizada para os demais tubos. A Tabela 15 resume os resultados dos critérios de informação e teste de aderência referentes à espessura média de parede (avgWT) dos tubos, nesta são contadas quantas vezes cada distribuição é a melhor das opções dentre as disponíveis segundo os critérios de informação, como também quantas vezes cada uma delas está entre as três e cinco melhores para cada amostra de dados, sendo contabilizadas apenas as distribuições que foram aprovadas no teste.

Constata-se pela Tabela 15 que a distribuição Gumbel é a melhor distribuição a ser escolhida em 29 dos 50 conjuntos de dados de avgWT, segundo os critérios de informação, e está

Tabela 15 – Resumo da inferência dos 50 conjuntos amostrais de seções de avgWT.

Distribuição	Melhor	3 Melhores	5 Melhores	Passaram no KS
Gumbel	29	32	39	50
Logístico	2	19	44	50
Log Normal	6	43	49	49
Gamma	0	29	49	49
Normal	5	16	48	49
Weibull	1	1	7	39
Uniforme	5	5	7	11

Fonte: próprio autor.

entre as três melhores opções de distribuições candidatas em 39 destes 50 conjuntos. Logo, a distribuição Gumbel é indicada como geradora dos conjuntos amostrais, seguida pela distribuição Logística que apresentou um resultado semelhante, mas inferior ao da primeira.

As distribuições Log Normal, Gamma e Normal foram satisfatórias mas um dos conjuntos amostrais não foi aprovado no teste de aderência para as distribuições, as distribuições Weibull e Uniforme tiveram um desempenho ruim quanto à sua capacidade de inferir sobre os dados amostrais.

De forma semelhante, a Tabela 16 resume os resultados dos critério de informação e teste de aderência para o diâmetro externo médio de cada seção (avgOD) nos mesmos 50 tubos.

Tabela 16 – Resumo da inferência dos 50 conjuntos amostrais de seções de avgOD.

Distribuição	Melhor	3 Melhores	5 Melhores	Passaram no KS
Normal	10	43	50	50
Gamma	0	38	50	50
Log Normal	4	21	50	50
Logístico	2	9	33	50
Weibull	17	21	37	50
Uniforme	17	18	23	47
Gumbel	0	0	7	44

Fonte: próprio autor.

Visto que a distribuição Normal está entre as três melhores distribuições em 43 de 50 conjuntos amostrais inferidos, e que está entre as cinco melhores distribuições em todos os conjuntos amostrais, a distribuição Normal é a indicada.

Diferentemente dos resultados da inferência dos dados de avgWT, os resultados do teste de aderência dos dados de avgOD foram mais irrestritos, todas as distribuições Normal, Gamma, Log Normal, Logística e Weibull foram aprovadas nos 50 casos analisados, apenas as distribuições Uniforme e Gumbel tiveram um resultado insatisfatório quanto à sua capacidade de inferir sobre os dados das amostras.

Ainda que que todas as distribuições que passaram nos testes de aderência sejam capazes de inferir sobre os dados, os resultados dos critérios de informação sofreram uma grande variância quanto à sua ordem, não sendo claro qual a melhor escolha outra que a distribuição Normal.

Observa-se então um padrão nos resultados de cada tubo, seja na análise dos valores de avgWT quanto para os valores de avgOD, o que condiz com a ideia de que todos os tubos possuem o mesmo padrão de comportamento aleatório.

Assim, realizou-se uma nova inferência estatística sobre o conjunto de dados, desta vez unindo todos os tubos em um único conjunto global. Antes da análise global, explora-se a seguir a análise da resistência dos tubos, para comparar os resultados da inferência entre características geométricas e mecânicas.

4.1.3 Análise de Dados Mecânicos

De forma semelhante à análise anterior sobre as características geométricas dos tubos, utilizou-se o Método da Máxima Verossimilhança na parametrização das distribuições selecionadas, as quais foram as mesmas da inferência anterior.

As Tabelas 17 e 18 resumem os resultados da parametrização dos modelos sobre as séries de dados de resistência YS e UTS, respectivamente.

Tabela 17 – Modelos candidatos parametrizados sobre os dados de resistência YS.

Distribuição	**Parâmetro 1**	**Parâmetro 2**	**Média**	**Desvio Padrão**
Uniforme	87,77	6,81	91,175	1,966
Normal	91,414	1,563	91,414	1,563
Log Normal	0,017	91,401	91,414	1,562
Logístico	91,386	0,900	91,386	1,632
Gamma	3424,966	0,027	91,414	1,562
Gumbel	90,642	1,475	91,494	1,892
Weibull	59,869	92,192	91,328	1,933

Fonte: próprio autor.

Tabela 18 – Modelos candidatos parametrizados sobre os dados de resistência UTS.

Distribuição	**Parâmetro 1**	**Parâmetro 2**	**Média**	**Desvio Padrão**
Uniforme	94,380	7,070	97,915	2,041
Normal	97,752	1,485	97,752	1,485
Log Normal	0,015	97,741	97,752	1,484
Logístico	97,720	0,852	97,720	1,545
Gamma	4339,109	0,023	97,752	1,484
Gumbel	97,022	1,389	97,823	1,781
Weibull	65,526	98,496	97,650	1,891

Fonte: próprio autor.

Com os modelos candidatos parametrizados, o teste de aderência Kolmogorov-Smirnov foi efetuado e computado, junto dos critérios de informação AIC e BIC. As Tabelas 19 e 20 apresentam os resultados para as duas séries de dados, o valor limite para o teste de aderência é de 0,151.

Tabela 19 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de YS.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Log Normal	0,054	Sim	306,06	310,85
Gamma	0,054	Sim	306,1	310,89
Normal	0,055	Sim	306,21	310,99
Logístico	0,041	Sim	308,73	313,51
Gumbel	0,067	Sim	313,72	318,51
Uniforme	0,121	Sim	314,78	319,57
Weibull	0,104	Sim	318,13	322,92

Fonte: próprio autor.

Tabela 20 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de UTS.

Distribuição	**KS obs.**	**Aprovada**	**AIC**	**BIC**
Log Normal	0,060	Sim	297,74	302,53
Gamma	0,061	Sim	297,8	302,59
Normal	0,063	Sim	297,94	302,73
Logístico	0,062	Sim	299,96	304,75
Gumbel	0,062	Sim	304,43	309,22
Weibull	0,104	Sim	312,42	317,21
Uniforme	0,179	Não	320,85	325,64

Fonte: próprio autor.

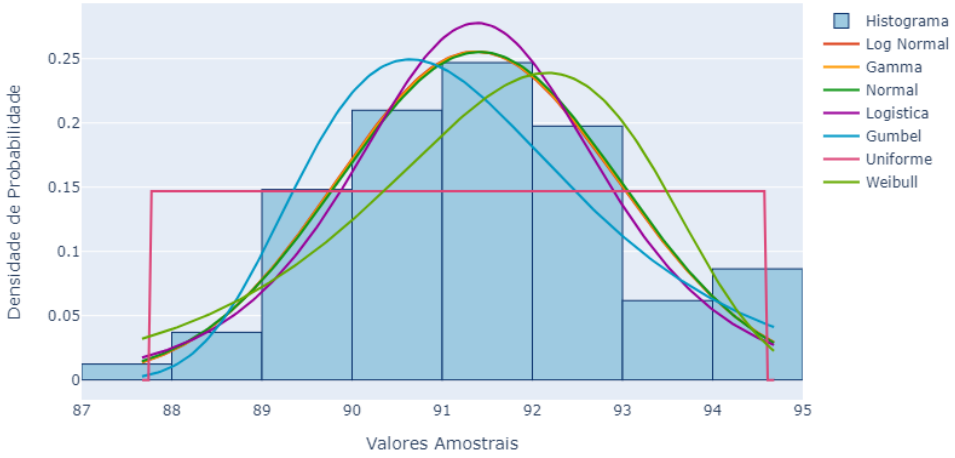
Observa-se um resultado positivo para quase todos os testes de aderência, com exceção da distribuição uniforme sobre os dados de UTS. A ordem das distribuições segundos os critérios de informação para os dois conjuntos de dados também é semelhante, indicando que as distribuições Log Normal, Gamma, Normal, Logístico e Gumbel se adéquam aos dados, nesta ordem. O resultado visual da análise pode ser visto pelas Figuras 6 e 7.

De forma diferente dos conjuntos de dados das características geométricas dos tubos, os dois conjuntos de dados de resistência contêm apenas 81 elementos, cada. Desta forma, a análise acima já é uma análise global dos tubos, tendo sido concluída de forma positiva. No entanto, estes dados serão explorados adiante para comparar seus resultados com as demais técnicas utilizadas.

4.1.4 Análise de Dados Geométricos: Amostragem Global

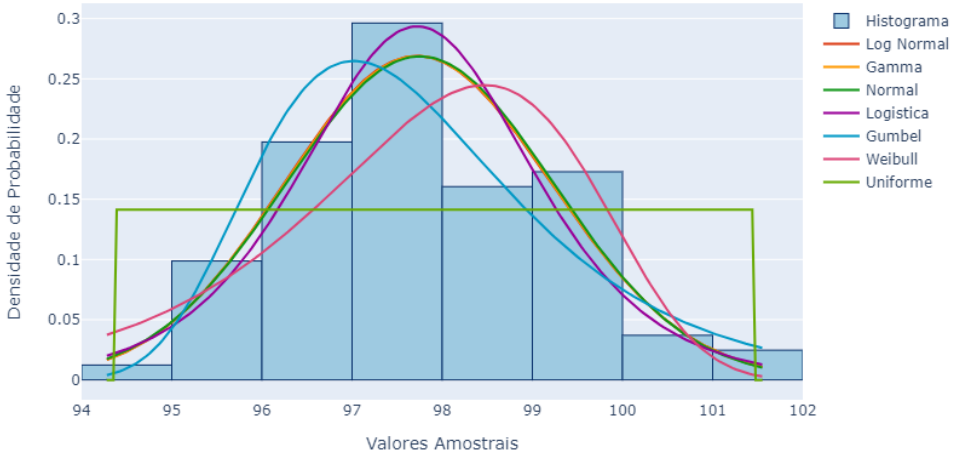
A união dos conjuntos de dados de 50 tubos resulta em dois conjuntos globais com 3669 valores amostrais cada, representando os valores aferidos de espessura de parede média (avgWT)

Figura 6 – Histograma dos dados de YS e suas distribuições candidatas.



Fonte: próprio autor.

Figura 7 – Histograma dos dados de UTS e suas distribuições candidatas.



Fonte: próprio autor.

e diâmetro externo médio (avgOD) de uma seção.

Para inferência estatística foram utilizadas as mesmas distribuições da análise tubo a tubo. A Tabela 21 apresenta o resultado da parametrização dos modelos candidatos sobre o conjunto de dados global para avgWT.

Os resultados do teste de aderência e dos critérios de informação das distribuições parametrizadas foram compilados na Tabela 22. Nesta análise, nenhuma das distribuições parametrizadas foi aprovada no teste de aderência e, portanto, não se obteve um modelo para representar os dados de produção de avgWT do conjunto global.

Mais uma vez, as distribuições Log Normal, Gamma e Normal foram equivalentes. Os critérios de informação, no entanto, foram diferentes em relação à ordem dos mesmos na análise

Tabela 21 – Modelos candidatos parametrizados sobre os dados de avgWT do conjunto global.

Distribuição	Parâmetro 1	Parâmetro 2	Média	Desvio Padrão
Uniforme	30,140	2,840	31,560	0,820
Normal	31,334	0,506	31,334	0,506
Log Normal	0,016	31,330	31,334	0,506
Logístico	31,321	0,303	31,321	0,550
Gamma	3838,687	0,008	31,334	0,506
Gumbel	31,088	0,443	31,343	0,568
Weibull	62,717	31,588	31,305	0,633

Fonte: próprio autor.

tubo-a-tubo, a distribuição Log Normal e suas equivalentes apresentaram o melhor desempenho, seguidas pelas distribuições Gumbel e Logística, que foram as melhores distribuições na análise tubo-a-tubo.

Tabela 22 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global.

Distribuição	KS obs.	Aprovada	AIC	BIC
Log Normal	0,090	Não	5408,80	5421,22
Gamma	0,090	Não	5413,01	5425,42
Normal	0,091	Não	5422,09	5434,51
Gumbel	0,047	Não	5448,89	5461,30
Logístico	0,088	Não	5684,05	5696,46
Weibull	0,091	Não	6014,58	6026,99
Uniforme	0,255	Não	7663,43	7675,85

Fonte: próprio autor.

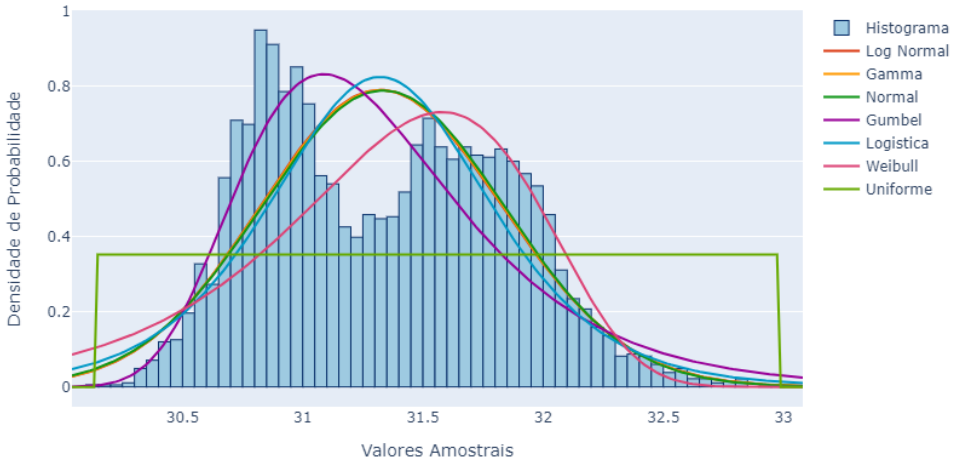
Contudo, os resultados dos critérios de informação não são suficientes para garantir que os dados são capazes de representar os dados de produção e a falha no teste de aderência indica, por ora, a desqualificação das distribuições.

A Figura 8 expõe a diferença entre o histograma de densidade de probabilidades do conjunto global dos dados de produção e a função equivalente das distribuições parametrizadas. Observa-se que o conjunto global de dados de produção possui uma característica de modelo de mistura, isto é, há mais de um pico no gráfico de densidade de probabilidades da distribuição.

Ainda que não haja registro de eventos ocorridos no processo de amostragem entende-se que este comportamento pode refletir uma mudança importante no processo de manufatura, como, por exemplo, uma recalibração ou adequação de equipamentos na linha de produção ao longo do período amostrado, caracterizando dois regimes distintos de fabricação.

Modelos de mistura derivam da união de conjuntos de observações de distribuições distintas, as quais possuem momentos diferentes. Ou seja, modelos de mistura representam a

Figura 8 – Histograma de avgWT do conjunto global e suas distribuições candidatas.



Fonte: próprio autor.

presença de duas ou mais sub-populações distintas em uma população geral (MCLACHLAN; PEEL, 2004).

Como todas as distribuições candidatas possuem dois parâmetros, elas não são capazes de inferir sobre dados com esta característica, exigindo distribuições candidatas com mais parâmetros. Por consequência, o teste de aderência não aprovou nenhuma das distribuições.

Um outro fator que interferiu sobre o resultado do teste de aderência foi o número de observações do conjunto global, pois, quanto mais informações (observações) existem, mais rigoroso se torna o teste de aderência. A estatística de teste de Kolmogorov-Smirnov sobre 3669 observações, para um nível de significância de 5%, deve ser inferior à 0,022 para que uma distribuição candidata com dois parâmetros seja aprovada.

Lin et al. (2013) e Baird e Harlow (2016) discutem esse aumento na potência dos testes de aderência não paramétricos para grandes amostras, o que leva a uma forte tendência de resultados negativos, rejeitando a hipótese nula.

A Tabela 23 exibe os parâmetros dos modelos candidatos sobre os dados de avgOD do conjunto global. O teste de aderência dos modelos parametrizados e seus critérios de informação estão presentes na Tabela 24.

Desta vez, a distribuição Weibull se destacou nos critérios de informação, acompanhada das distribuições Normal, Gamma e Log Normal, as quais foram numericamente equivalentes, seguidas pelas distribuições Logística e Gumbel.

A Figura 9 evidencia o histograma do conjunto global e suas distribuições candidatas. Nota-se que, diferente da Figura 8 sobre avgWT, a Figura 9 sobre as medidas de avgOD não possui um comportamento multimodal aparente, que é caracterizada por mais de um pico (máximo local). Ainda assim, o teste de aderência não aprovou nenhuma das distribuições

Tabela 23 – Modelos candidatos parametrizados sobre os dados de avgOD do conjunto global.

Distribuição	Parâmetro 1	Parâmetro 2	Média	Desvio Padrão
Uniforme	534,59	4,730	536,955	1,365
Normal	537,595	0,761	537,595	0,761
Log Normal	0,001	537,595	537,595	0,761
Logístico	537,635	0,431	537,635	0,782
Gamma	499.389,493	0,001	537,595	0,761
Gumbel	537,197	0,853	537,69	1,094
Weibull	821,752	537,956	537,579	0,838

Fonte: próprio autor.

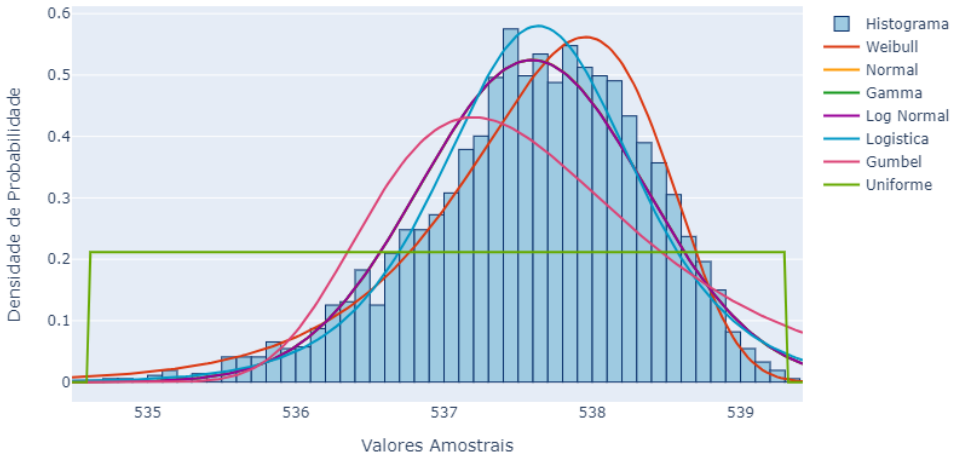
Tabela 24 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global.

Distribuição	KS obs.	Aprovada	AIC	BIC
Weibull	0,034	Não	8.276,90	8.289,31
Normal	0,026	Não	8.407,55	8.419,96
Gamma	0,026	Não	8.409,48	8.421,9
Log Normal	0,026	Não	8.410,45	8.422,87
Logístico	0,033	Não	8.435,02	8.447,43
Gumbel	0,111	Não	9.599,67	9.612,08
Uniforme	0,080	Não	11.406,70	11.419,12

Fonte: próprio autor.

candidatas.

Figura 9 – Histograma de avgOD do conjunto global e suas distribuições candidatas.



Fonte: próprio autor.

Três fatores, não excludentes entre si, são apontados como promotores deste resultado inconclusivo sobre o conjunto de dados global:

1. O grau de incerteza dos dados aferidos sobrepõe seu próprio comportamento, poluindo o conjunto de dados a ser aferido;

2. O cardápio de distribuições é insuficiente. O comportamento é melhor representado por uma distribuição que não fez parte da análise;

3. Regimes distintos de fabricação forjaram um modelo de mistura, dificultando a caracterização dos dados;

Para avaliar o quanto a incerteza na aferição dos dados afeta a análise, utilizou-se da formulação apresentada nas Equações 2.9, 2.10 e 2.11.

Os valores de avgWT e avgOD registrados possuem duas casas decimais, o que remete a uma incerteza de $\pm 0,005$ no valor de cada aferição. As Tabelas 25 e 26 exibem o grau de incerteza para os momentos dos conjuntos globais, comparando-os com os valores reais dos momentos das amostras.

Tabela 25 – Grau de incerteza na aferição dos dados do conjunto global de avgWT.

Momento	**Valor real**	**Incerteza absoluta** ($\pm$)	**Incerteza relativa** ($\pm$ **%**)
Média	31,334	0,0050	0,016
Desvio Padrão	0,5064	0,0043	0,852
Variância	0,2564	0,0044	1,704

Fonte: próprio autor.

Tabela 26 – Grau de incerteza na aferição dos dados do conjunto global de avgOD.

Momento	**Valor real**	**Incerteza absoluta** ($\pm$)	**Incerteza relativa** ($\pm$ **%**)
Média	537,595	0,0050	0,001
Desvio Padrão	0,761	0,0040	0,522
Variância	0,579	0,0060	1,043

Fonte: próprio autor.

Observa-se que a média e o desvio padrão das amostras possuem um grau de incerteza inferior à 1 % de seus valores reais, a variância não difere muito mais que 1 % também. Logo, descarta-se esta opção dentre os fatores que contribuem com a falha do teste de aderência na análise global.

O segundo fator apontado como causador do resultado inconclusivo foi o cardápio de distribuições candidatas utilizado. Desta forma, realizou-se uma nova inferência estatística sobre os conjuntos globais de avgWT e avgOD, desta vez expandindo o cardápio de distribuições para todas as distribuições disponíveis na biblioteca estatística scipy stats (Virtanen et al., 2020) que puderam ser aplicáveis, listadas na Tabela 27 com seu correspondente número de parâmetros.

Tabela 27 – Lista das 94 distribuições candidatas e seu correspondente número de parâmetros.

alpha (3)	anglit (2)	arcsine (2)	argus (3)
beta (4)	betaprime (4)	bradford (3)	burr (4)
burr12 (4)	cauchy (2)	chi (3)	chi2 (3)
cosine (2)	crystalball (4)	dgamma (3)	dweibull (3)
erlang (3)	expon (2)	exponnorm (3)	exponpow (3)
exponweib (4)	f (4)	fatiguelife (3)	fisk (3)
foldcauchy (3)	foldnorm (3)	frechet_l (3)	frechet_r (3)
gamma (3)	gausshyper (6)	genexpon (5)	genextreme (3)
gengamma (4)	genhalflogistic (3)	geninvgauss (4)	genlogistic (3)
gennorm (3)	genpareto (3)	gilbrat (2)	gompertz (3)
gumbel_l (2)	gumbel_r (2)	halfcauchy (2)	halfgennorm (3)
halflogistic (2)	halfnorm (2)	hypsecant (2)	invgamma (3)
invgauss (3)	invweibull (3)	johnsonsb (4)	johnsonsu (4)
kappa3 (3)	kappa4 (4)	kstwobign (2)	laplace (2)
levy (2)	levy_l (2)	loggamma (3)	logistic (2)
loglaplace (3)	lognorm (3)	loguniform (4)	lomax (3)
maxwell (2)	mielke (4)	moyal (2)	nakagami (3)
ncf (5)	nct (4)	ncx2 (4)	norm (2)
norminvgauss (4)	pareto (3)	pearson3 (3)	powerlaw (3)
powerlognorm (4)	powernorm (3)	rayleigh (2)	rdist (3)
recipinvgauss (3)	rice (3)	semicircular (2)	skewnorm (3)
t (3)	trapz (4)	triang (3)	truncexpon (3)
tukeylambda (3)	uniform (2)	vonmises_line (3)	wald (2)
weibull_max (3)	weibull_min (3)	-	-

Fonte: próprio autor.

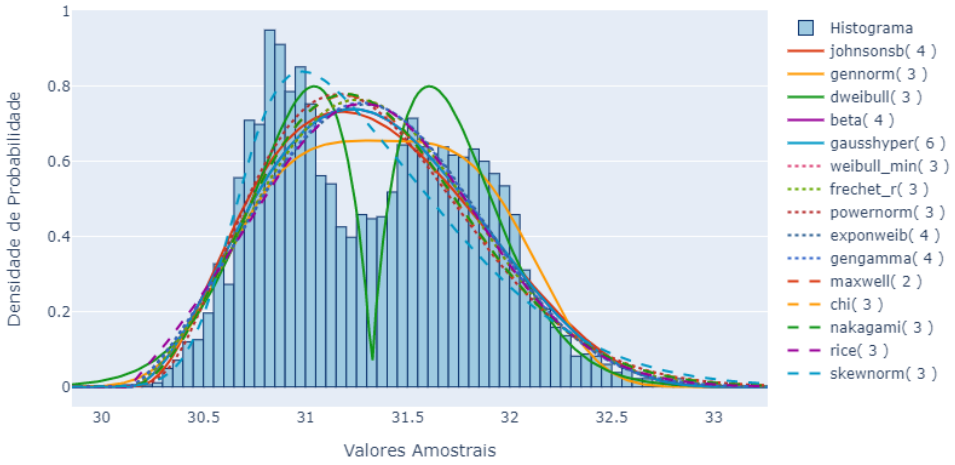
As Figuras 10 e 11 apresentam o histograma de avgWT e avgOD dos respectivos conjuntos globais e as correspondentes distribuições candidatas que melhor se adéquam a cada um dos modelos.

Os resultados do teste de aderência e critérios de informação dos 15 modelos parametrizados com melhores resultados, dos dois conjuntos globais, estão presentes nas Tabelas 28 e 29.

Observa-se pelos resultados, ordenados a partir do BIC, que a expansão do cardápio de modelos contribuiu para uma melhor inferência sobre os conjuntos de dados. Algumas distribuições do conjunto de global dados avgOD foram aprovadas no teste de aderência, enquanto que nenhuma foi aprovada no conjunto global de dados avgWT, o qual possui a característica multimodal já comentada.

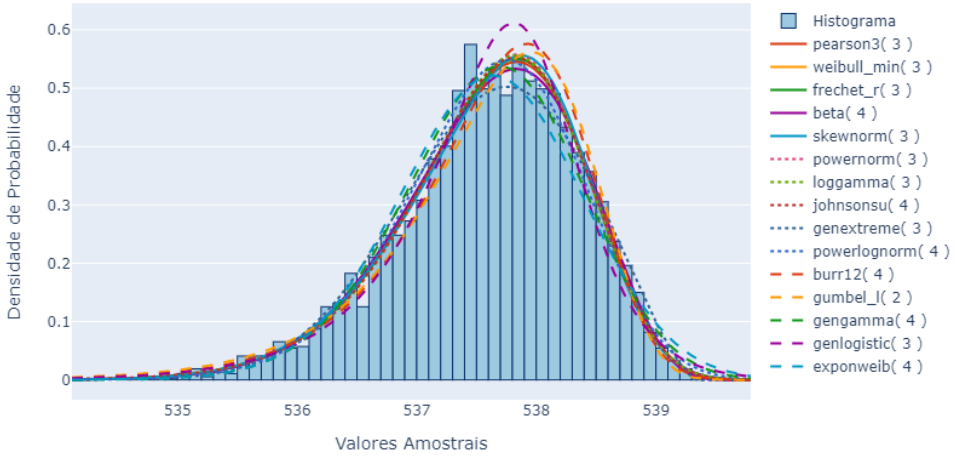
As distribuições candidatas com mais parâmetros apresentaram resultados melhores que as distribuições de dois parâmetros, mesmo utilizando critérios como o BIC, o qual se torna mais criterioso de acordo com o número de parâmetros das distribuições.

Figura 10 – Histograma de avgWT do conjunto global e todas as distribuições candidatas.



Fonte: próprio autor.

Figura 11 – Histograma de avgOD do conjunto global e as distribuições candidatas.



Fonte: próprio autor.

Para fins de comparação de desempenho entre o número de parâmetros das distribuições candidatas, avaliando a ordenação do BIC mas sem considerar os resultados do teste de aderência, as Figuras 12 e 13 destacam o desempenho das distribuições candidatas agrupadas pela sua quantidade de parâmetros.

Verifica-se um ganho de desempenho das distribuições com mais de dois parâmetros para as primeiras posições, segundo o BIC. Atenta-se que os últimos valores do gráfico de barras não são muito relevantes, uma boa quantidade das distribuições candidatas não é adequada para os conjuntos de dados utilizados, estando portanto nas últimas colocações da ordenação.

Uma metodologia semelhante foi realizada sobre os dados de resistência para comparação. As Figuras 14 e 15 apresentam o resultado visual da parametrização das distribuições candidatas, acompanhados das Tabelas 30 e 31. As Figuras 16 e 17 mostram a composição das distribuições

Tabela 28 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global incluindo modelos de maior complexidade.

Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
johnsonsb	4	0,06	Não	5150,43	5175,26
gennorm	3	0,083	Não	5160,15	5178,77
dweibull	3	0,042	Não	5174,42	5193,04
beta	4	0,067	Não	5177,21	5202,04
gausshyper	6	0,068	Não	5181,58	5218,82
weibull_min	3	0,074	Não	5232,94	5251,57
frechet_r	3	0,074	Não	5232,94	5251,57
powernorm	3	0,059	Não	5233,33	5251,95
exponweib	4	0,076	Não	5231,26	5256,09
gengamma	4	0,076	Não	5233,24	5258,07
maxwell	2	0,064	Não	5249,03	5261,44
chi	3	0,066	Não	5250,71	5269,33
nakagami	3	0,066	Não	5250,71	5269,33
rice	3	0,071	Não	5284,16	5302,78
skewnorm	3	0,031	Não	5293,58	5312,2

Fonte: próprio autor.

Tabela 29 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global incluindo modelos de maior complexidade.

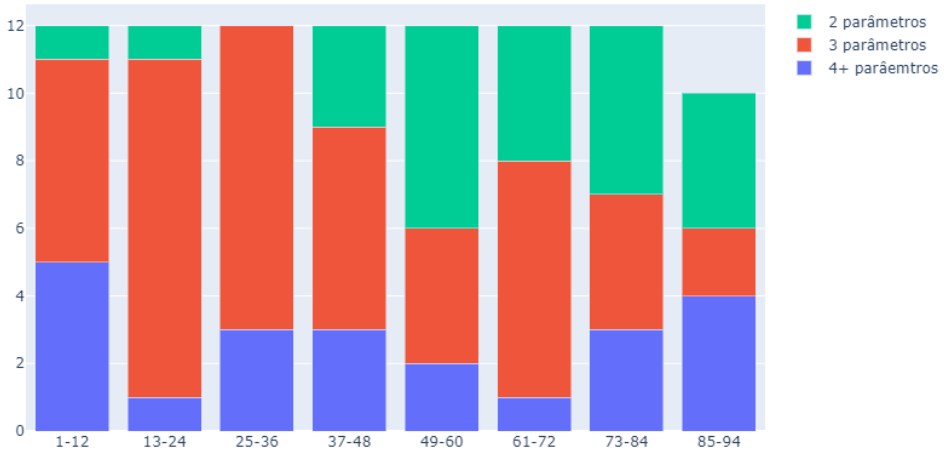
Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
pearson3	3	0,013	Sim	8191,37	8210,0
weibull_min	3	0,014	Sim	8193,99	8212,62
frechet_r	3	0,014	Sim	8193,99	8212,62
beta	4	0,01	Sim	8189,22	8214,05
skewnorm	3	0,017	Sim	8201,34	8219,96
powernorm	3	0,016	Sim	8202,0	8220,62
loggamma	3	0,014	Sim	8202,8	8221,42
johnsonsu	4	0,014	Sim	8200,31	8225,15
genextreme	3	0,015	Sim	8217,5	8236,13
powerlognorm	4	0,014	Sim	8230,86	8255,69
burr12	4	0,036	Não	8252,31	8277,14
gumbel_l	2	0,034	Não	8278,57	8290,99
gengamma	4	0,016	Sim	8273,52	8298,35
genlogistic	3	0,028	Não	8326,02	8344,65
exponweib	4	0,025	Não	8343,88	8368,71

Fonte: próprio autor.

agrupadas pela sua quantidade de parâmetros e ordenadas segundo o seu BIC.

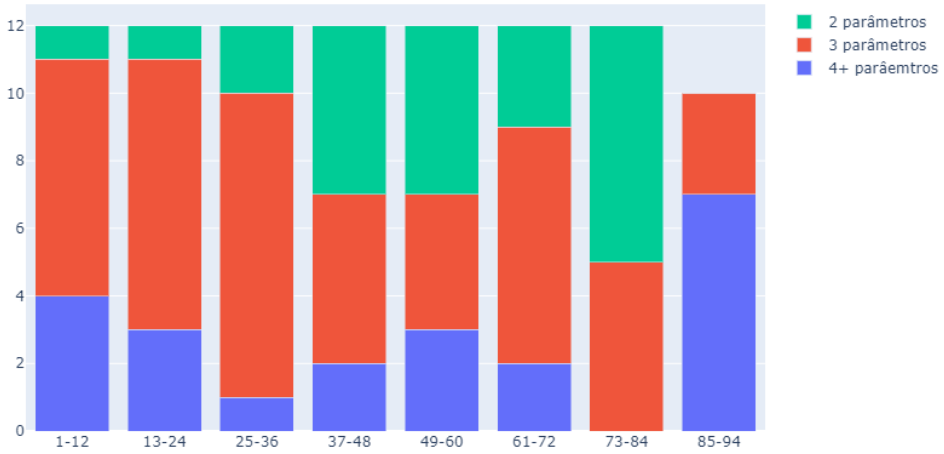
Observa-se que, por mais que distribuições candidatas com 3 parâmetros sejam bem qualificadas de acordo com os critérios de informação e os testes de aderência, há uma grande

Figura 12 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados avgWT.



Fonte: próprio autor.

Figura 13 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados avgOD.



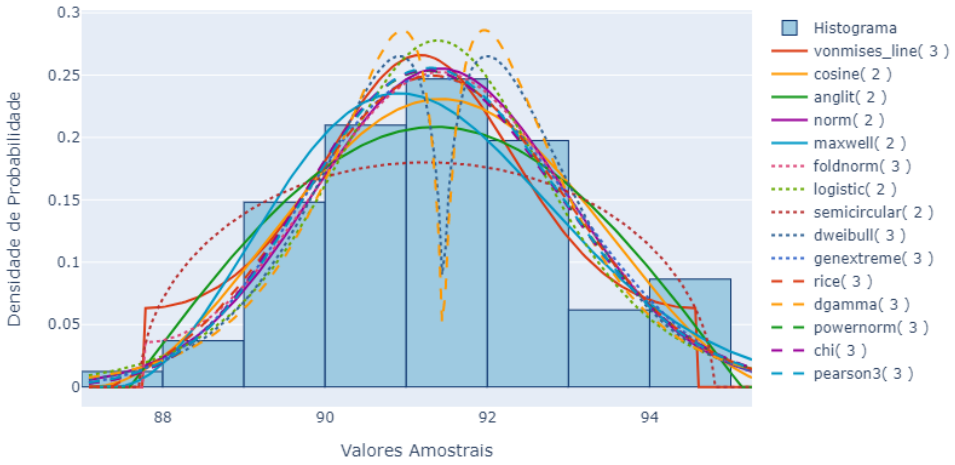
Fonte: próprio autor.

incidência de distribuições de dois parâmetros na inferência sobre ambos os dados de resistência. Ainda, evidenciou-se que todas as distribuições candidatas com 4 ou mais parâmetros não são aderentes aos dados mecânicos de resistência.

Conclui-se, portanto, que um cardápio de distribuições mais rico contribui com o resultado da inferência estatística dos dados geométricos mas não acrescentam muito para os dados mecânicos. O conjunto de dados avgOD foi aprovado no teste de aderência para algumas distribuições com mais parâmetros, enquanto que o conjunto de dados avgWT não obteve sucesso no teste de aderência de nenhuma das distribuições.

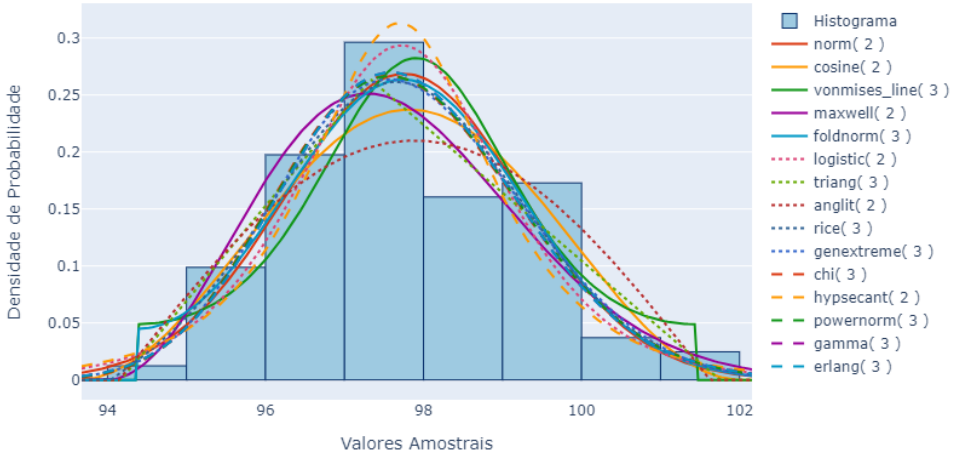
O terceiro fator apontado foi o modelo de mistura criado ao unir conjuntos com regimes

Figura 14 – Histograma de YS do conjunto global e todas as distribuições candidatas.



Fonte: próprio autor.

Figura 15 – Histograma de UTS do conjunto global e as distribuições candidatas.



Fonte: próprio autor.

distintos de fabricação. Para redução do comportamento multimodal dos dados, realizou-se uma padronização dos conjuntos de dados individuais.

A padronização dos dados retira a média amostral do conjunto de dados de cada valor aferido. Este valor é então dividido pelo desvio padrão amostral, alcançando assim conjuntos padronizados de média nula e desvio padrão unitário para cada um dos conjuntos de tubos individuais.

Um conjunto de dados padronizado global foi então criado pela união dos conjuntos padronizados individuais, espera-se que a padronização reduza o ruído gerado por diferentes padrões de fabricação. Para que os resultados sejam comparados com os resultados anteriores, optou-se por despadronizar os dados, isto é, multiplicou-se os dados do conjunto padronizado global pelo desvio padrão do conjunto global, depois foi somada a média do conjunto padronizado

Tabela 30 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de YS para os modelos expandidos.

Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
vonmises_line	3	0,013	Sim	298,12	305,3
cosine	2	0,076	Sim	304,54	309,33
anglit	2	0,084	Sim	305,13	309,92
norm	2	0,055	Sim	306,21	310,99
maxwell	2	0,055	Sim	308,09	312,88
foldnorm	3	0,057	Sim	305,95	313,14
logistic	2	0,041	Sim	308,73	313,51
semicircular	2	0,094	Sim	308,74	313,53
dweibull	3	0,066	Sim	306,84	314,02
genextreme	3	0,058	Sim	306,93	314,11
rice	3	0,055	Sim	307,22	314,4
dgamma	3	0,065	Sim	307,32	314,5
powernorm	3	0,052	Sim	307,78	314,97
chi	3	0,052	Sim	307,79	314,98
pearson3	3	0,974	Sim	307,91	315,09

Fonte: próprio autor.

Tabela 31 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de UTS para os modelos expandidos.

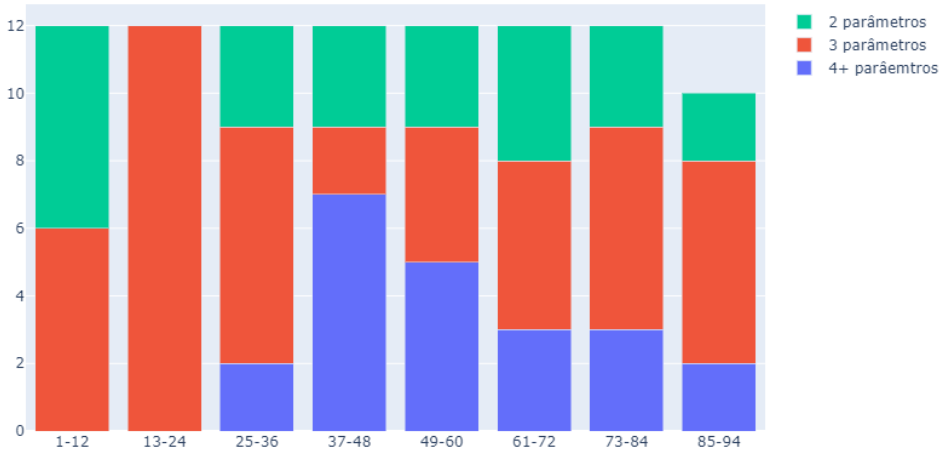
Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
norm	2	0,063	Sim	297,94	302,73
cosine	2	0,074	Sim	297,96	302,75
vonmises_line	3	0,108	Sim	295,79	302,98
maxwell	2	0,04	Sim	298,79	303,58
foldnorm	3	0,058	Sim	297,47	304,65
logistic	2	0,062	Sim	299,96	304,75
triang	3	0,056	Sim	297,7	304,88
anglit	2	0,093	Sim	300,52	305,31
rice	3	0,05	Sim	298,6	305,78
genextreme	3	0,048	Sim	298,93	306,12
chi	3	0,045	Sim	299,24	306,42
hypsecant	2	0,061	Sim	301,66	306,45
powernorm	3	0,047	Sim	299,29	306,48
gamma	3	0,049	Sim	299,38	306,56

Fonte: próprio autor.

global a estes valores.

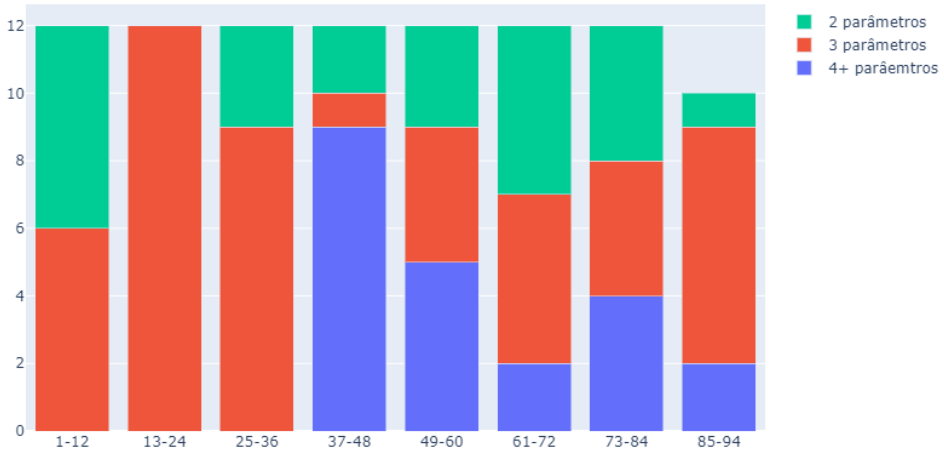
A inferência estatística utilizada foi similar à anterior, com o cardápio de distribuições expandido. As Figuras 18 e 19 apresentam o histograma de avgWT e avgOD dos respectivos conjuntos globais padronizados e as correspondentes distribuições candidatas que melhor se

Figura 16 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados YS.



Fonte: próprio autor.

Figura 17 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados UTS.



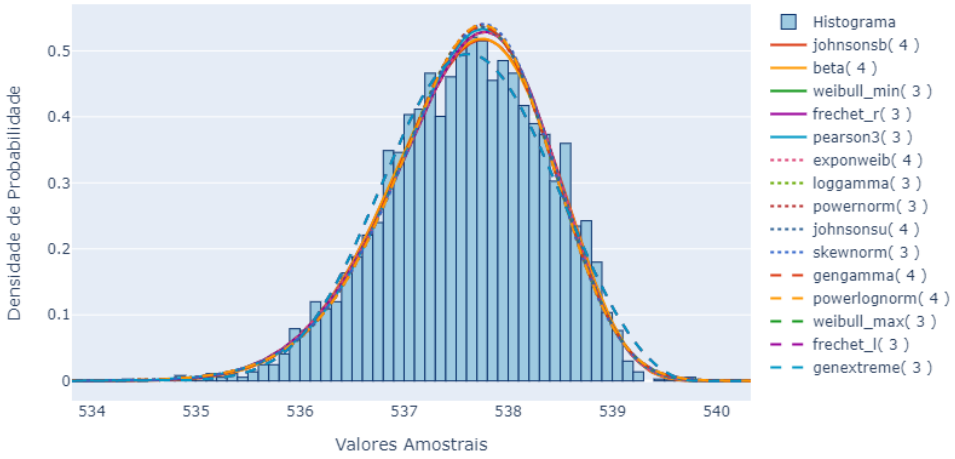
Fonte: próprio autor.

adéquam a cada um dos modelos.

Evidencia-se pelas figuras que houve uma redução significativa do efeito multimodal ao se adotar a estratégia de padronização dos dados, as Figuras 9 e 8 apresentam os conjuntos globais sem padronização, para comparação.

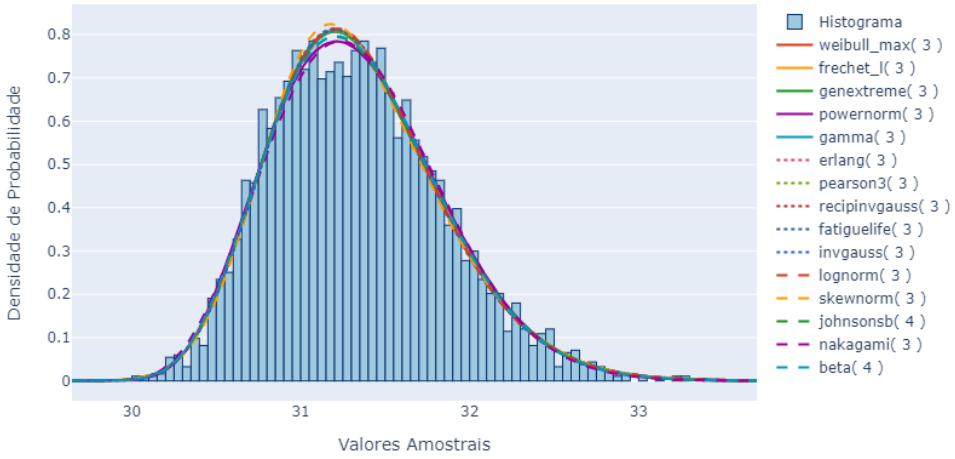
Os resultados do teste de aderência e critérios de informação dos 45 modelos parametrizados com melhores resultados, dos dois conjuntos globais padronizados, estão postos nas tabelas 32 e 33, as quais foram expandidas para exibir todos os resultados relevantes obtidos via teste de aderência. Para comparação do desempenho das distribuições em termos da quantidade de parâmetros, as Figuras 20 e 21 foram criadas.

Figura 18 – Histograma de avgOD do conjunto global padronizado e suas distribuições candidatas.



Fonte: próprio autor.

Figura 19 – Histograma de avgWT do conjunto global padronizado e suas distribuições candidatas.



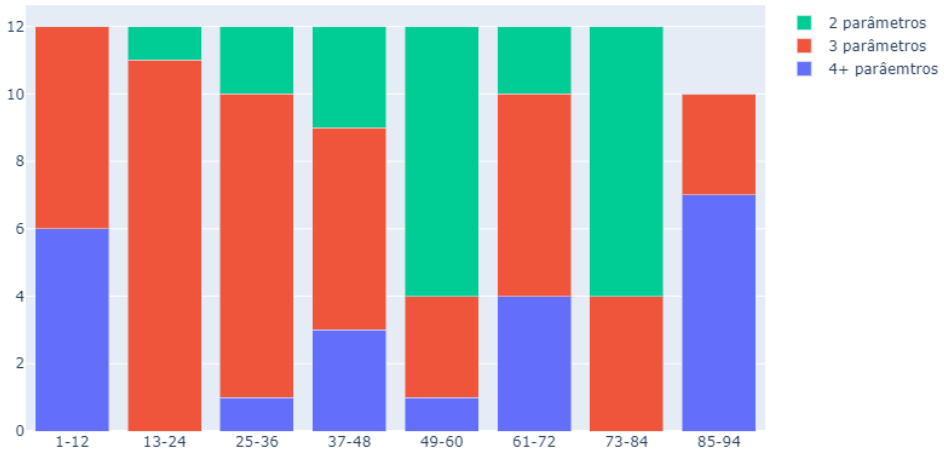
Fonte: próprio autor.

Verifica-se pelos resultados acima que a padronização dos dados reduziu perceptivelmente a multimodalidade dos dados, contribuindo para uma inferência estatística mais assertiva.

Várias distribuições foram aprovadas no testes de aderência Kolmogorov-Smirnov para um nível de significância de 5%, para os dois conjuntos de dados. 29 das 94 distribuições foram aprovadas para o conjunto global padronizado avgOD e 28 para o correspondente conjunto avgWT.

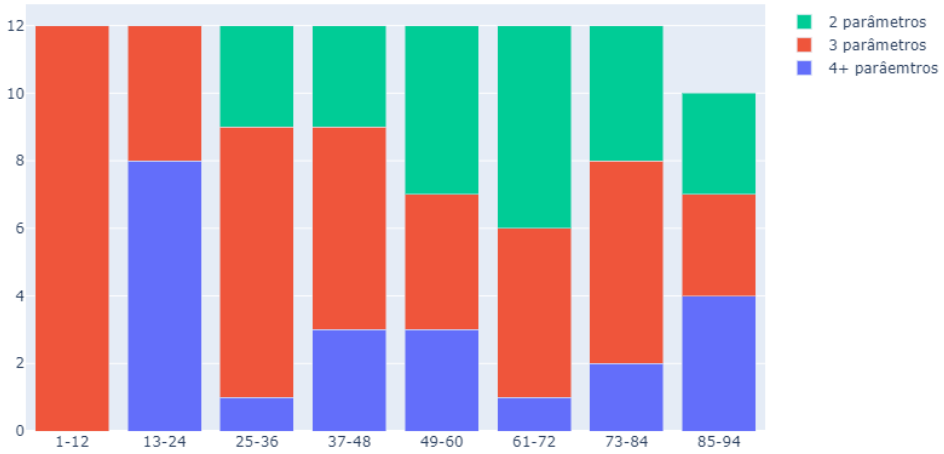
Assim, conclui-se que um cardápio de distribuições robusto e a redução do comportamento multimodal, possivelmente criado por regimes distintos do processo de manufatura, permitiram gerar resultados conclusivos. O conjunto amostral adquirido e utilizado é evidência

Figura 20 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados global padronizado avgOD.



Fonte: próprio autor.

Figura 21 – Gráfico de barras com a composição das distribuições agrupadas pela sua quantidade de parâmetros segundo a ordem do BIC para o conjunto de dados global padronizado avgWT.



Fonte: próprio autor.

suficiente para aceitar a hipótese de que este foi originado por qualquer uma das distribuições teóricas aprovadas, segundo o nível de significância utilizado.

Tabela 32 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgOD do conjunto global padronizado.

#	Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
1	johnsonsb	4	0,013	Sim	8312,557	8337,388
2	beta	4	0,013	Sim	8313,313	8338,144
3	weibull_min	3	0,018	Sim	8321,454	8340,077
4	frechet_r	3	0,018	Sim	8321,454	8340,077
5	pearson3	3	0,019	Sim	8322,422	8341,045
6	exponweib	4	0,018	Sim	8323,41	8348,241
7	loggamma	3	0,019	Sim	8330,368	8348,991
8	powernorm	3	0,02	Sim	8335,288	8353,911
9	johnsonsu	4	0,019	Sim	8329,985	8354,816
10	skewnorm	3	0,021	Sim	8336,516	8355,139
11	gengamma	4	0,018	Sim	8331,152	8355,983
12	powerlognorm	4	0,02	Sim	8336,791	8361,621
13	weibull_max	3	0,019	Sim	8366,771	8385,395
14	frechet_l	3	0,019	Sim	8366,771	8385,395
15	genextreme	3	0,019	Sim	8366,771	8385,395
16	norm	2	0,02	Sim	8407,549	8419,964
17	gennorm	3	0,018	Sim	8402,597	8421,22
18	foldnorm	3	0,02	Sim	8408,596	8427,219
19	rice	3	0,0	Sim	8409,549	8428,172
20	exponnorm	3	0,02	Sim	8409,571	8428,194
21	t	3	0,02	Sim	8409,672	8428,295
22	vonmises_line	3	0,02	Sim	8409,739	8428,362
23	lognorm	3	0,021	Sim	8414,037	8432,66
24	fatiguelife	3	0,02	Sim	8414,768	8433,391
25	tukeylambda	3	0,021	Sim	8418,889	8437,512
26	burr12	4	0,032	Não	8415,902	8440,733
27	chi	3	0,022	Sim	8427,945	8446,568
28	gamma	3	0,026	Não	8459,746	8478,369
29	erlang	3	0,025	Não	8463,913	8482,536
30	gompertz	3	0,041	Não	8477,479	8496,103
31	genlogistic	3	0,03	Não	8489,523	8508,146
32	logistic	2	0,028	Não	8508,67	8521,085
33	gumbel_l	2	0,044	Não	8509,228	8521,643
34	dweibull	3	0,022	Sim	8504,537	8523,16
35	recipinvgauss	3	0,028	Não	8508,509	8527,132
36	fisk	3	0,028	Não	8511,617	8530,24
37	chi2	3	0,03	Não	8513,786	8532,409
38	gausshyper	6	0,027	Não	8568,162	8605,408
39	invgamma	3	0,032	Não	8590,89	8609,513
40	invgauss	3	0,035	Não	8606,382	8625,005
41	dgamma	3	0,025	Não	8607,269	8625,892
42	hypsecant	2	0,034	Não	8614,489	8626,904
43	burr	4	0,05	Não	8651,942	8676,773
44	laplace	2	0,053	Não	8848,26	8860,676
45	loglaplace	3	0,053	Não	8850,833	8869,456

Fonte: próprio autor.

Tabela 33 – Resultados Kolmogorov-Smirnov e critérios de informação sobre os dados de avgWT do conjunto global padronizado.

#	Distribuição	Parâmetros	KS obs.	Aprovada	AIC	BIC
1	weibull_max	3	0,013	Sim	5280,235	5298,858
2	frechet_l	3	0,013	Sim	5280,235	5298,858
3	genextreme	3	0,013	Sim	5280,235	5298,858
4	powernorm	3	0,013	Sim	5280,242	5298,865
5	gamma	3	0,014	Sim	5283,359	5301,982
6	erlang	3	0,014	Sim	5283,359	5301,982
7	pearson3	3	0,014	Sim	5283,359	5301,982
8	recipinvgauss	3	0,015	Sim	5284,843	5303,466
9	fatiguelife	3	0,015	Sim	5284,865	5303,488
10	invgauss	3	0,012	Sim	5285,411	5304,034
11	lognorm	3	0,016	Sim	5286,372	5304,995
12	skewnorm	3	0,013	Sim	5286,6	5305,223
13	johnsonsb	4	0,013	Sim	5282,313	5307,144
14	nakagami	3	0,017	Sim	5288,614	5307,237
15	beta	4	0,014	Sim	5283,468	5308,299
16	gengamma	4	0,013	Sim	5284,391	5309,222
17	powerlognorm	4	0,014	Sim	5284,697	5309,528
18	geninvgauss	4	0,015	Sim	5285,492	5310,323
19	johnsonsu	4	0,016	Sim	5288,774	5313,604
20	f	4	0,016	Sim	5289,814	5314,644
21	invgamma	3	0,021	Sim	5298,843	5317,466
22	gausshyper	6	0,016	Sim	5291,126	5328,372
23	exponnorm	3	0,025	Não	5326,212	5344,835
24	genlogistic	3	0,017	Sim	5336,728	5355,351
25	weibull_min	3	0,021	Sim	5349,257	5367,88
26	frechet_r	3	0,021	Sim	5349,257	5367,88
27	mielke	4	0,017	Sim	5344,146	5368,977
28	rice	3	0,028	Não	5364,357	5382,98
29	alpha	3	0,034	Não	5385,54	5404,163
30	maxwell	2	0,019	Sim	5410,169	5422,585
31	foldnorm	3	0,036	Não	5413,397	5432,02
32	norm	2	0,037	Não	5422,092	5434,507
33	gumbel_r	2	0,026	Não	5423,208	5435,624
34	vonmises_line	3	0,037	Não	5422,479	5441,102
35	gennorm	3	0,037	Não	5423,891	5442,514
36	t	3	0,037	Não	5424,078	5442,701
37	rdist	3	0,037	Não	5424,198	5442,821
38	burr12	4	0,027	Não	5423,066	5447,897
39	crystalball	4	0,037	Não	5426,092	5450,923
40	burr	4	0,026	Não	5429,965	5454,796
41	loggamma	3	0,039	Não	5439,216	5457,839
42	invweibull	3	0,029	Não	5451,947	5470,57
43	kstwobign	2	0,029	Não	5459,982	5472,397
44	fisk	3	0,035	Não	5472,375	5490,998
45	logistic	2	0,035	Não	5485,13	5497,545

Fonte: próprio autor.

4.2 Previsão de Dados para Detecção de Anomalias

A fim de apresentar a aplicação da metodologia exposta de previsão de dados, foram selecionadas séries históricas reais de pressão medidas em poços de petróleo em produção. Após a análise dos resultados do emprego de técnicas de previsão de dados, discutida neste capítulo, sugere-se uma rotina de monitoramento em tempo real para detecção antecipada de anomalias.

4.2.1 Previsão de Dados Futuros

As séries de dados foram retiradas do acervo disponibilizado em Vargas et al. (2019), tornados públicos com o objetivo de servir de referência para o uso de técnicas associadas à detecção e diagnóstico de eventos anômalos em poços marítimos de petróleo e gás.

Três séries foram selecionadas, referentes à dados de pressão em eventos de fechamento espúrio da válvula de segurança de subsuperfície (DHSV), instalada na coluna de produção com o objetivo de garantir a interrupção de fluxo em casos de desconexão da coluna.

No fechamento espúrio a válvula interrompe a produção, sendo um caso de detecção falso positiva de vazamento, o que provoca uma parada inoportuna na produção e consequente prejuízo econômico.

Numa primeira abordagem de previsão de dados, testa-se o poder do modelo ARIMA para prever dados futuros após um evento que provoque uma oscilação dos dados, durante uma mudança para um novo estado estacionário. Isto é, testa-se a aptidão do modelo em prever dados futuros logo após a ocorrência de um evento.

Trata-se de uma estratégia desafiadora, que exige muito do modelo, visto que os modelos são recomendados para séries estacionárias. Entende-se que após um evento, os dados oscilam em torno de um valor médio e este tende a reduzir sua variância de modo a retornar a um valor de média constante e variância reduzida, podendo assim ser aplicado a um modelo ARIMA.

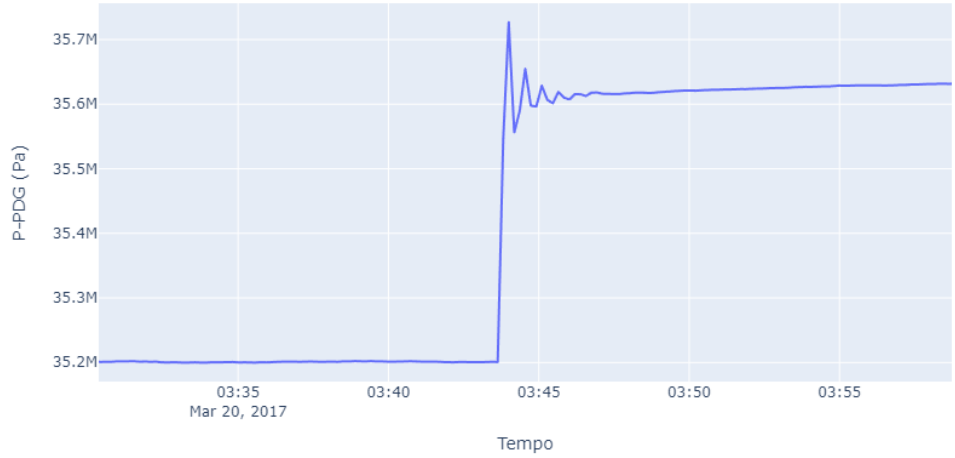
4.2.1.1 Série de Dados 1

A primeira série selecionada refere-se ao registro do poço identificado como 12, com dados coletados em 20 de março de 2017 às 03:30:22, com valores aferidos a cada segundo por uma duração de 28 minutos e 22 segundos, gerando um total de 1702 valores registrados no Medidor de Pressão de Fundo (PDG - *Pressure Downhole Gauge*). A Figura 22 ilustra a série de dados completa.

Os dados selecionados para análise e ajuste do modelo ARIMA constituem uma janela de observação de 2 minutos, nos quais verifica-se uma maior oscilação do valor da pressão logo após o fechamento espúrio da válvula, vide Figura 23.

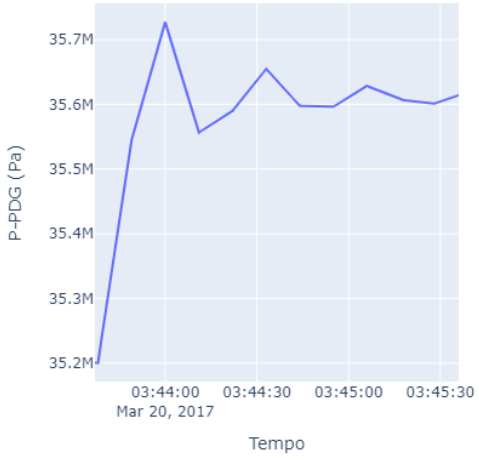
O primeiro passo para escolha dos hiperparâmetros do modelo é avaliar qual o valor de d para o qual os dados se mantêm estacionários. As Figuras 24 e 25 apresentam, respectivamente,

Figura 22 – Pressão no PDG durante um evento de fechamento espúrio da DHSV.



Fonte: próprio autor.

Figura 23 – Pressão no PDG logo após um evento de fechamento espúrio da DHSV.



Fonte: próprio autor.

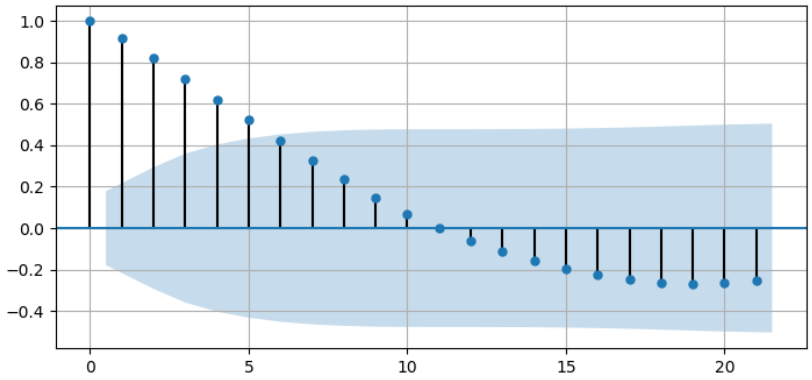
o ACF da série original (com d igual a 0) e o PACF da mesma série. A região em azul representa o intervalo de confiança para um nível de significância de 5%.

Pelos gráficos, observa-se que o ACF se aproxima de zero rapidamente. Assim, o hiperparâmetro d foi escolhido como nulo, pois os dados originais são estacionários. Como os valores dos gráficos em grande parte são nulos, as figuras apresentam apenas os seus primeiros valores.

Para comparação, as Figuras 26 e 27 apresentam o ACF da série diferenciada (com d igual a 1) e o PACF da série.

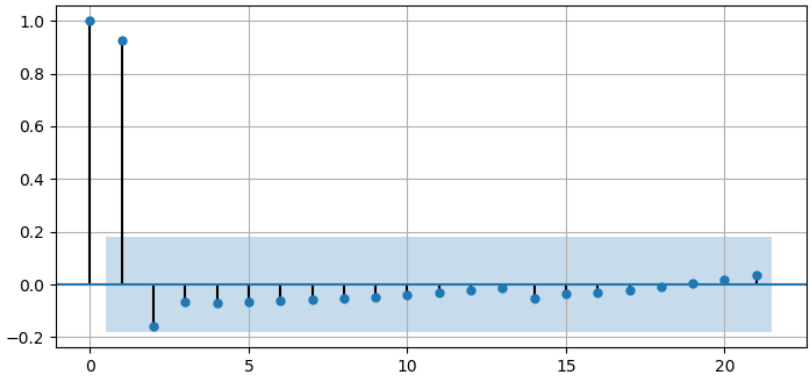
A série diferenciada não apresenta muitas diferenças em relação à série original em ambos ACF e PACF, pois ela é estacionária.

Figura 24 – Gráfico ACF para $d = 0$.



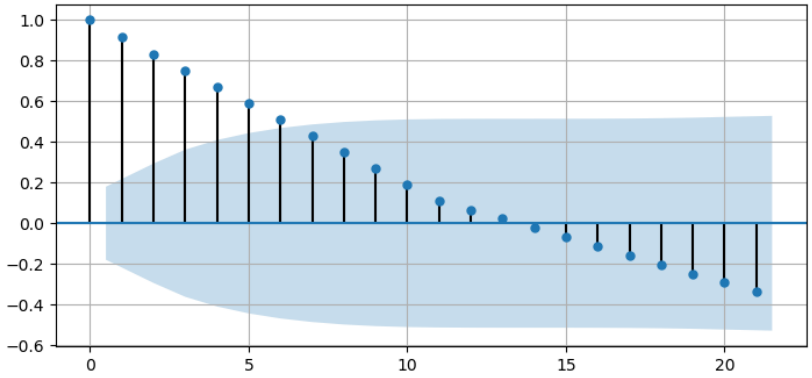
Fonte: próprio autor.

Figura 25 – Gráfico PACF para $d = 0$.



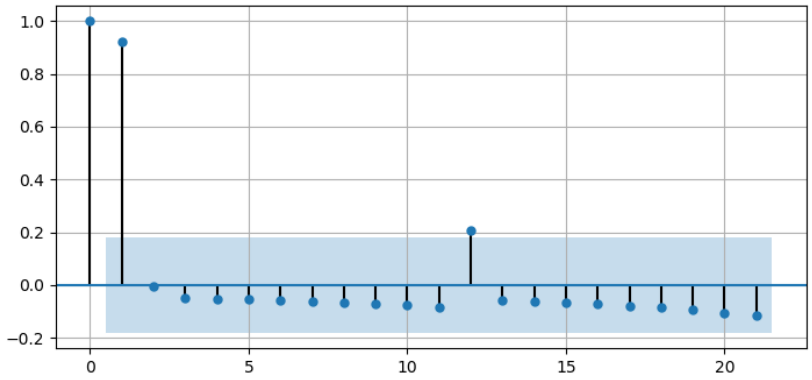
Fonte: próprio autor.

Figura 26 – Gráfico ACF para $d = 1$.



Fonte: próprio autor.

Figura 27 – Gráfico PACF para $d = 1$.



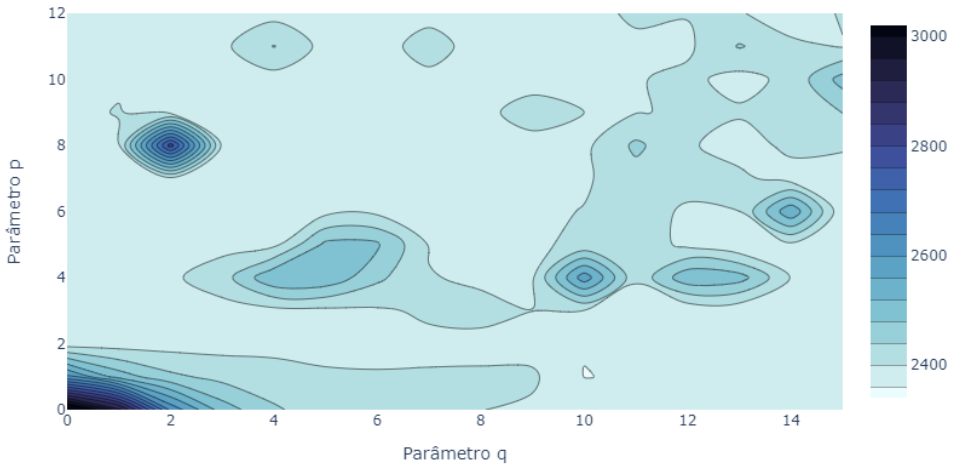
Fonte: próprio autor.

Em modelos estritamente Autorregressivos o gráfico PACF é nulo a partir do ponto p+1, o que indica que p deve ser próximo de 2 para essa série temporal. Uma análise semelhante pode ser realizada no gráfico ACF para estimar o hiperparâmetro q, o qual deve ser próximo de 6.

Para escolha aperfeiçoada dos hiperparâmetros p e q foi realizada uma otimização de inteiros por busca de grade, garantindo a escolha ótima dos hiperparâmetros de forma a minimizar o valor do AIC.

A Figura 28 ilustra o resultado da busca de grade. O modelo ARIMA(p=1, d=0, q=10) foi o que apresentou o melhor resultado para o AIC, a Tabela 34 lista os modelos com melhores indicadores.

Figura 28 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA.



Fonte: próprio autor.

Com os hiperparâmetros do modelo escolhidos, basta determinar as suas constantes através de mínimos quadrados, concluindo o ajuste do modelo ARIMA aos dados.

Tabela 34 – Modelos indicados segundo seu AIC.

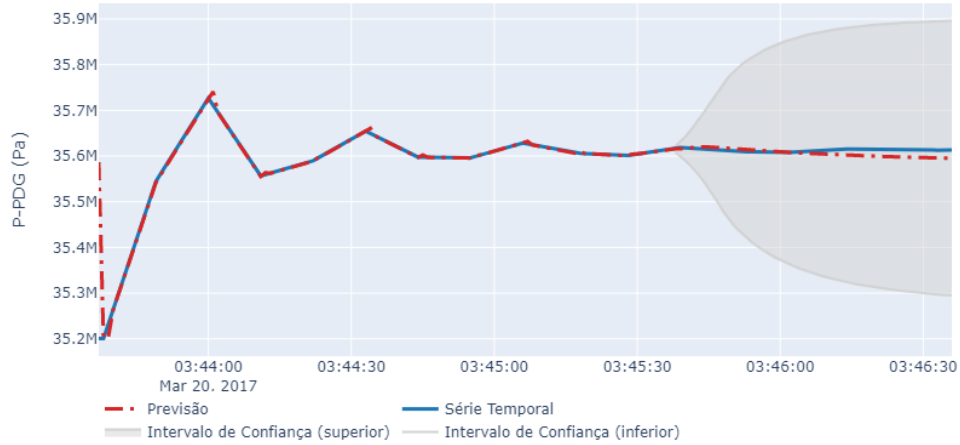
Modelo	AIC
ARIMA(p=1, d=0, q=10)	2.359,65
ARIMA(p=2, d=0, q=10)	2.360,78
ARIMA(p=1, d=0, q=11)	2.361,54
ARIMA(p=1, d=0, q=12)	2.363,42
ARIMA(p=5, d=0, q=13)	2.363,62

Fonte: próprio autor.

As Figuras 29 e 30 apresentam os resultados para os modelos ARIMA(p=1, d=0, q=10) e ARIMA(p=5, d=0, q=13) respectivamente. Os demais modelos da Tabela 34 foram omitidos por serem semelhantes ao primeiro.

Realizou-se por meio dos modelos a estimativa em vermelho dos valores utilizados para treino, como também a previsão dos valores futuros, comparando-os com os dados reais em azul. A região em cinza representa o intervalo de confiança para um nível de significância de 5%. Nota-se o aumento da largura das bandas de confiança, à medida que se infere sobre instantes de tempo mais distantes.

Figura 29 – Previsão da série temporal via modelo ARIMA(1,0,10).



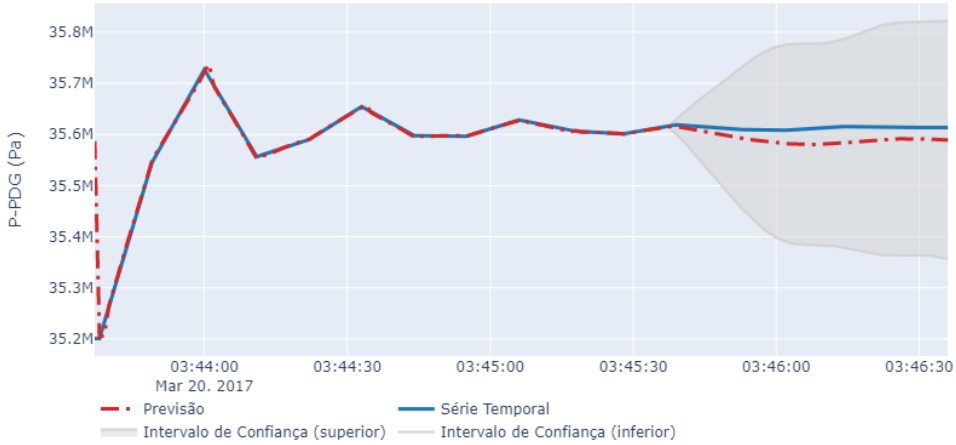
Fonte: próprio autor.

Para quantificação da qualidade das estimativas dos valores utilizados em treino e das previsões dos valores futuros pelos modelos ARIMA, medidas de acurácia foram utilizadas. As Tabelas 35 e 36 listam os resultados.

A análise de erro via MAPE leva a um resultado equivalente ao já apresentado pelo MAE, o mesmo pode ser dito para uma análise por meio do RMSE em relação ao CVRMSE.

Quanto aos resultados, observa-se que os quatro primeiros modelos são semelhantes, resultado esperado por possuírem hiperparâmetros próximos. Já a diferença entre o quinto

Figura 30 – Previsão da série temporal via modelo ARIMA(5,0,13).



Fonte: próprio autor.

Tabela 35 – Medidas de acurácia das estimativas dos valores para treino da Série de Dados 1.

Modelo	**MAE**	**MAPE (%)**	**RMSE**	**CVRMSE (%)**
ARIMA(p=1, d=0, q=10)	4.731,40	0,0134	35.467,32	0,0997
ARIMA(p=2, d=0, q=10)	4.649,08	0,0132	35.459,67	0,0996
ARIMA(p=1, d=0, q=11)	4.719,86	0,0134	35.466,15	0,0997
ARIMA(p=1, d=0, q=12)	4.709,08	0,0133	35.464,93	0,0997
ARIMA(p=5, d=0, q=13)	5.127,39	0,0145	35.439,38	0,0996

Fonte: próprio autor.

Tabela 36 – Medidas de acurácia das previsões dos valores futuros da Série de Dados 1.

Modelo	**MAE**	**MAPE (%)**	**RMSE**	**CVRMSE (%)**
ARIMA(p=1, d=0, q=10)	9.215,06	0,0259	11.135,83	0,0313
ARIMA(p=2, d=0, q=10)	10.504.81	0,0295	13.405,02	0,0376
ARIMA(p=1, d=0, q=11)	9.708,03	0,0273	12.114,69	0,0340
ARIMA(p=1, d=0, q=12)	10.479,64	0,0294	13.362,63	0,0375
ARIMA(p=5, d=0, q=13)	21.344,12	0,0599	23.138,89	0,0650

Fonte: próprio autor.

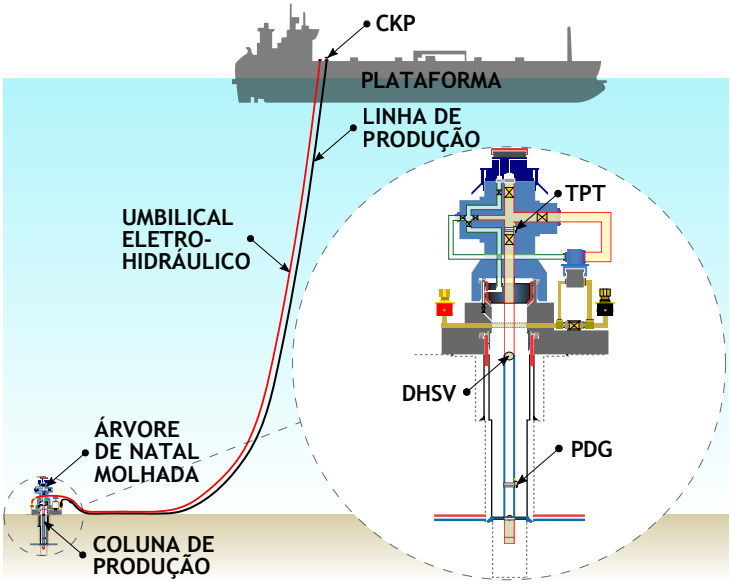
modelo (Figura 30) aos demais, perceptível na Tabela 36 que trata das medidas de acurácia das previsões futuras, foi captada no crescimento do MAPE e CVRMSE em relação aos demais modelos.

De forma geral, todos os modelos, inclusive o quinto, apresentam medidas de acurácia semelhantes. Estes valores serão utilizadas como parâmetros para os demais estudos de caso.

4.2.1.2 Série de Dados 2

A segunda série de dados contém os valores de pressão do poço 12 registrados à montante da válvula de bloqueio de produção (PCK - *Production Choke*). Um esquema da posição do PCK pode ser visualizado na Figura 31.

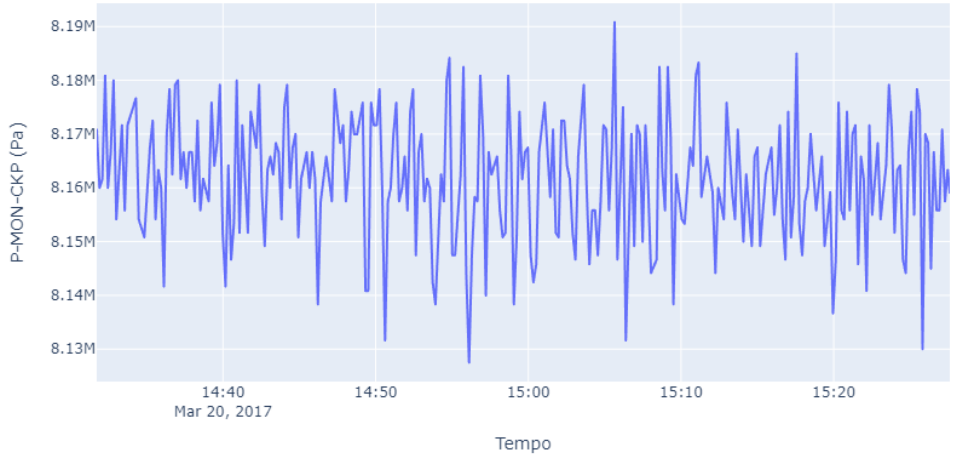
Figura 31 – Esquema simplificado de um poço típico *offshore*.



Fonte: Adaptado de Vargas et al. (2019).

Os dados foram coletados durante um momento de maior variação (amplitude) do sinal registrado às 14:31:44 do mesmo dia do evento da Série de Dados 1. A Figura 32 apresenta a série de dados completa.

Figura 32 – Pressão no P-MON-CKP com amplitude do sinal.

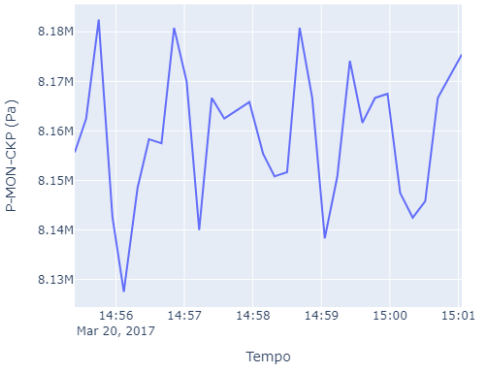


Fonte: próprio autor.

Uma janela de 340 segundos (5 minutos e 40 segundos) de operação (Figura 33) foi

escolhida para treinamento do modelo ARIMA. Como a série de dados possui uma grande variação e o modelo ARIMA é baseado em modelos lineares, espera-se que o modelo seja capaz de acompanhar o valor médio das previsões posteriores à janela de treinamento.

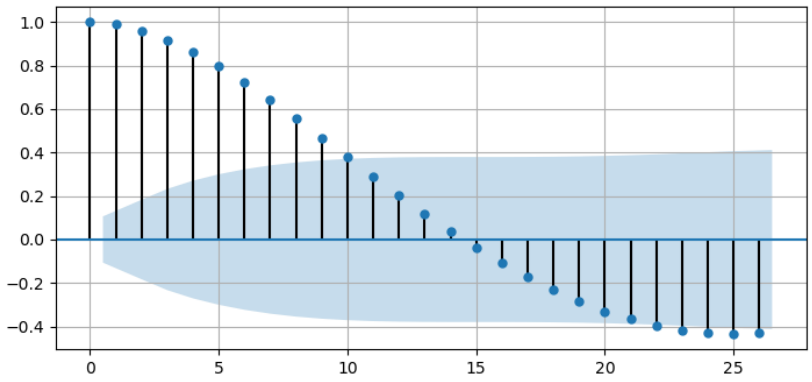
Figura 33 – Janela utilizada para treinamento do modelo.



Fonte: próprio autor.

Seguindo a metodologia para determinação dos hiperparâmetros do modelo ARIMA, as Figuras 34 e 35 apresentam um recorte dos primeiros valores do gráfico de ACF e PACF da janela de treino.

Figura 34 – Gráfico ACF para $d = 0$.

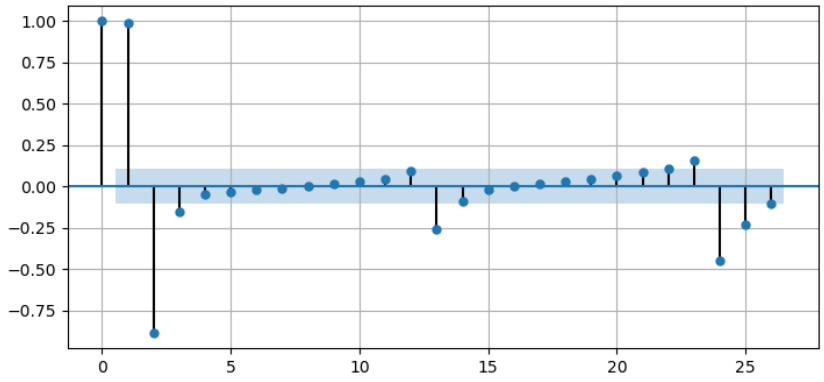


Fonte: próprio autor.

Observa-se pelos gráficos que não há necessidade de diferenciação da série, o hiperparâmetro d foi escolhido como nulo. Ainda, os gráficos revelam uma estimativa para os hiperparâmetros p e q de 2 e 10, respectivamente.

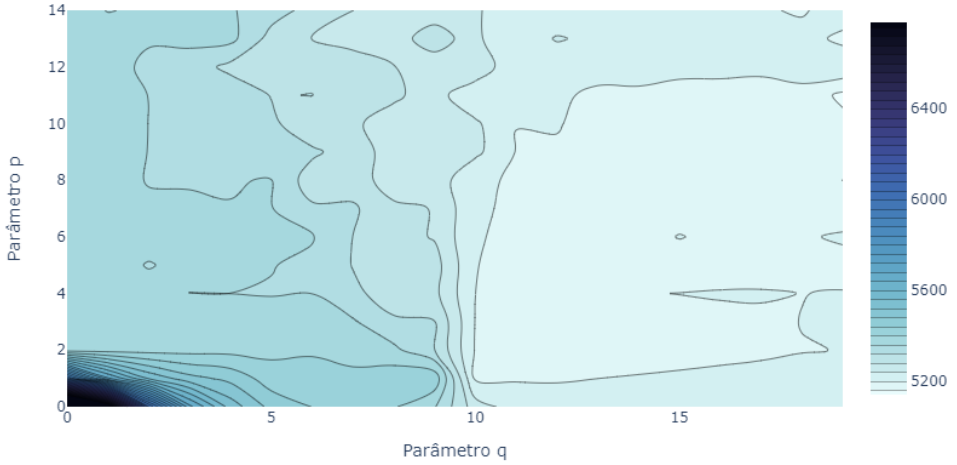
Para uma escolha otimizada dos valores dos hiperparâmetros do modelo, uma busca de grade é realizada tendo como ponto de partida a estimativa dos hiperparâmetros. A Figura 36 apresenta as isocurvas da função objetivo AIC para os hiperparâmetros do ARIMA. A Tabela 37 lista os modelos com melhores resultados.

Figura 35 – Gráfico PACF para $d = 0$.



Fonte: próprio autor.

Figura 36 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA.



Fonte: próprio autor.

Tabela 37 – Modelos indicados segundo seu AIC.

Modelo	**AIC**
ARIMA(p=11, d=0, q=19)	5.151,03
ARIMA(p=11, d=0, q=17)	5.162,84
ARIMA(p= 9, d=0, q=19)	5.163,25
ARIMA(p= 9, d=0, q=18)	5.167,53
ARIMA(p= 8, d=0, q=18)	5.170,33

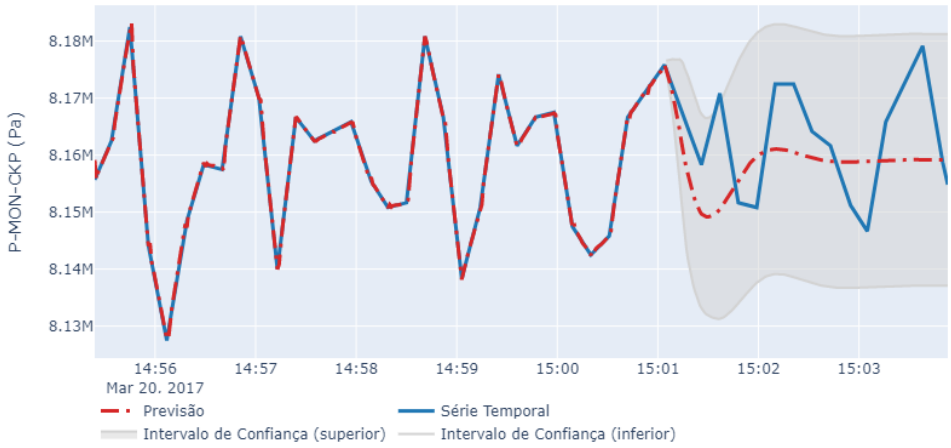
Fonte: próprio autor.

Como esperado, devido à grande variação do sinal de pressão, os hiperparâmetros ótimos do modelo ARIMA são elevados, o que indica uma dependência de termos com um maior atraso.

A previsão de valores futuros, presente nas Figuras 37 e 38 mostra uma limitação do

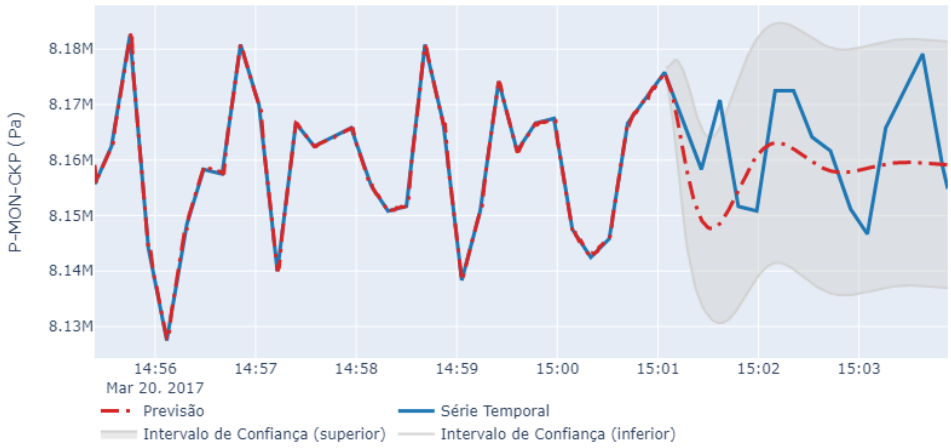
modelo ARIMA. O método não é capaz de prever com precisão dados muito erráticos. Ainda assim, o método foi capaz de manter uma média móvel capaz de acompanhar o comportamento geral dos dados.

Figura 37 – Previsão da série temporal via modelo ARIMA (11,0,19).



Fonte: próprio autor.

Figura 38 – Previsão da série temporal via modelo ARIMA (11,0,17).



Fonte: próprio autor.

Para comparação da capacidade de previsão do modelo nesta série de dados, compara-se as medidas de acurácia, listadas nas Tabelas 38 e 39, com os valores aferidos dos modelos de previsão aplicados na série de dados 1.

Atenta-se que, quando comparando séries de dados de tamanho e naturezas distintas, as medidas MAPE e CVRMSE são mais indicadas para análise do que as medidas MAE e RMSE.

Observa-se pelas medidas de acurácia que as estimativas dos valores de treino da série de dados 2 (Tabela 38) foram mais eficientes do que as estimativas da série de dados 1 (Tabela

Tabela 38 – Medidas de acurácia das estimativas dos valores para treino da série de dados 2.

Modelo	MAE	MAPE (%)	RMSE	CVRMSE (%)
ARIMA(p=11, d=0, q=19)	220,11	0,0027	491,16	0,0060
ARIMA(p=11, d=0, q=17)	229,91	0,0028	497,04	0,0061
ARIMA(p=9, d=0, q=19)	222,68	0,0027	496,89	0,0061

Fonte: próprio autor.

Tabela 39 – Medidas de acurácia das previsões dos valores futuros da série de dados 2.

Modelo	MAE	MAPE (%)	RMSE	CVRMSE (%)
ARIMA(p=11, d=0, q=19)	8.140,23	0,0997	9.571,50	0,1172
ARIMA(p=11, d=0, q=17)	7.962,77	0,0975	9.501,25	0,1164
ARIMA(p=9, d=0, q=19)	8.205,80	0,1005	9.586,71	0,1174

Fonte: próprio autor.

35), por derivar medidas de acurácia inferiores. Contudo, as previsões para valores futuros da série de dados 2 (Tabela 39) possui medidas de acurácia muito mais elevadas que as avaliadas sobre a série de dados 1 (Tabela 36).

O primeiro resultado sobre a comparação das medidas de acurácia para os dois exemplos pode ser explicado pelo uso de uma janela de dados maior para a segunda série de dados, além do fato de que o modelo que o modelo ARIMA da segunda série de dados possui mais elementos dependentes (hiperparâmetros com valores mais altos). Além disso, pressões em PDG são mais elevadas, o que pode resultar em erros absolutos maiores.

Já o segundo resultado sobre a comparação é simples e visível pelas figuras de previsão de série temporal. Isto é, os valores previstos para a segunda série de dados possuem um erro considerável, mesmo que o modelo seja capaz de prever a média móvel do sinal.

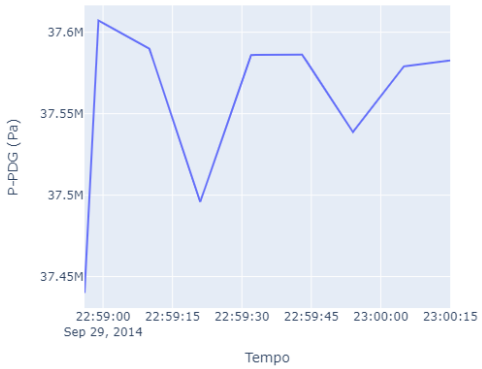
4.2.1.3 Série de Dados 3

Uma terceira série de dados, contendo os valores de pressão do poço 11 registrados no PDG logo após um fechamento espúrio da válvula, foi utilizada para teste da capacidade de previsão do modelo ARIMA. Desta vez, uma série mais curta com 80 dados (Figura 39) foi selecionada para treinamento do modelo.

Mais uma vez, como o ARIMA é baseado em modelos lineares, espera-se que ele seja capaz de acompanhar o valor médio das previsões posteriores à janela de treinamento. Seguindo a metodologia para determinação dos hiperparâmetros do modelo ARIMA, as Figuras 40 e 41 apresentam um recorte dos primeiros valores do gráfico de ACF e PACF da janela de treino.

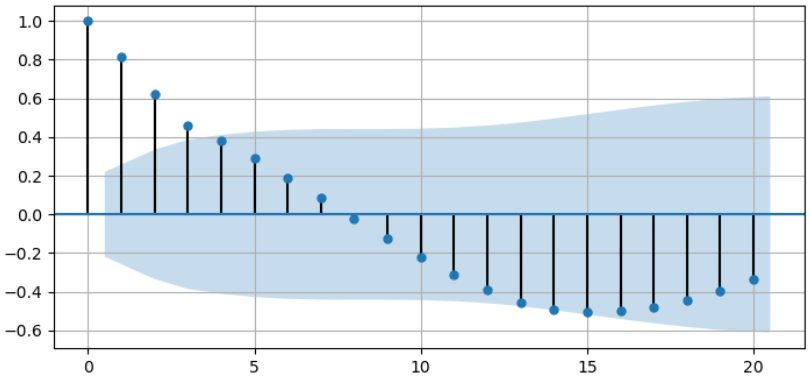
A partir dos gráficos, verifica-se que a diferenciação relativa ao hiperparâmetro d não precisa ser realizada. Ainda, os gráficos revelam uma estimativa para os hiperparâmetros p e q de 2 e 3, respectivamente.

Figura 39 – Janela utilizada para treinamento do modelo.



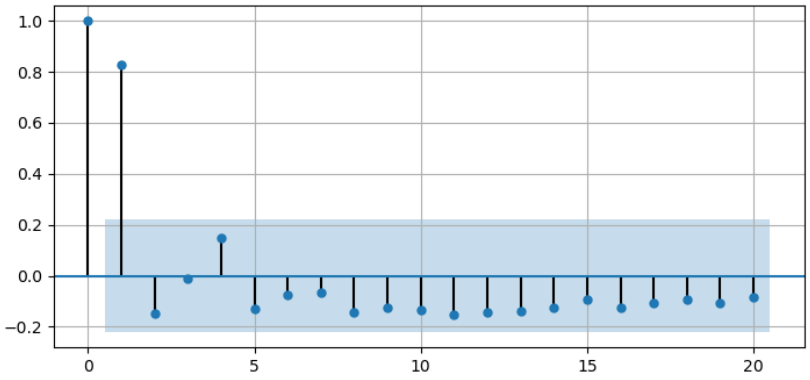
Fonte: próprio autor.

Figura 40 – Gráfico ACF para $d = 0$.



Fonte: próprio autor.

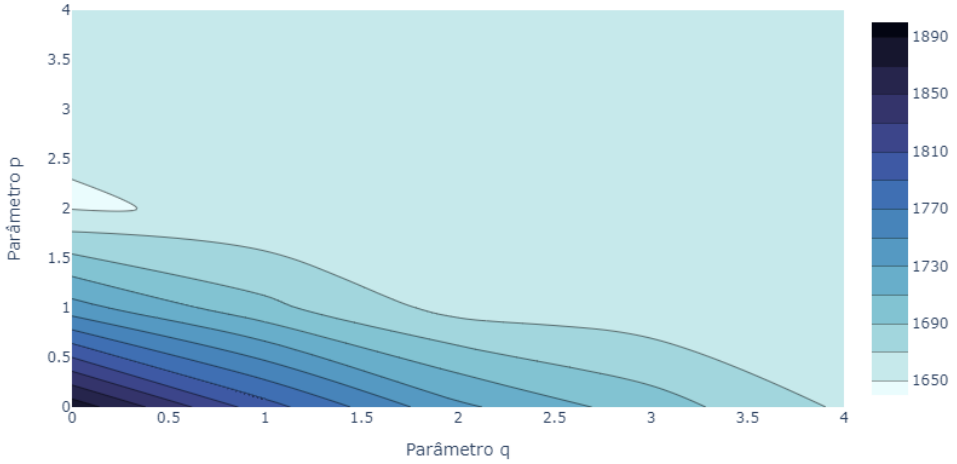
Figura 41 – Gráfico PACF para $d = 0$.



Fonte: próprio autor.

Para otimização dos hiperparâmetros do modelo, a busca de grade ilustrada na Figura 42 com as isocurvas da função objetivo AIC foi efetuada. A Tabela 40 lista os modelos com melhores resultados.

Figura 42 – Isocurvas do AIC segundo os hiperparâmetros do ARIMA.



Fonte: próprio autor.

Tabela 40 – Modelos indicados segundo seu AIC.

Modelo	**AIC**
ARIMA(p=2, d=0, q=0)	1.649,43
ARIMA(p=2, d=0, q=1)	1.651,13
ARIMA(p=3, d=0, q=0)	1.651,37
ARIMA(p=2, d=0, q=2)	1.652,08
ARIMA(p=4, d=0, q=0)	1.652,37
ARIMA(p=3, d=0, q=1)	1.653,18
ARIMA(p=2, d=0, q=3)	1.653,40

Fonte: próprio autor.

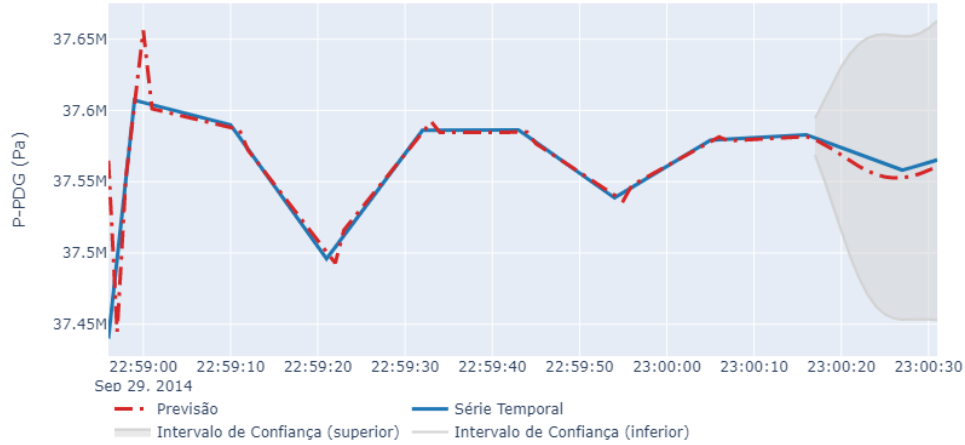
Por ser utilizada uma série mais curta e bem comportada para treinamento, os hiperparâmetros ótimos do modelo ARIMA são baixos, o que indica uma dependência de termos com um atraso menor.

A previsão de valores futuros ilustrada na Figura 43 mostra como o modelo, por possuir hiperparâmetros baixos, mantém uma média móvel capaz de acompanhar o comportamento geral dos dados.

Para avaliação da capacidade de previsão do modelo nesta série de dados, compara-se as medidas de acurácia, listadas nas Tabelas 41 e 42, com os valores aferidos dos modelos de previsão aplicados na série de dados 1.

Observa-se pelas medidas de acurácia que as estimativas dos valores de treino da série de dados 3 (Tabela 41) foram próximas das estimativas da série de dados 1 (Tabela 35) e inferiores

Figura 43 – Previsão da série temporal via modelo ARIMA (2,0,0).



Fonte: próprio autor.

Tabela 41 – Medidas de acurácia das estimativas dos valores para treino da série de dados 3.

Modelo	**MAE**	**MAPE (%)**	**RMSE**	**CVRMSE (%)**
ARIMA(p=2, d=0, q=0)	4.601,91	0,0123	16.279,26	0,0433
ARIMA(p=2, d=0, q=1)	4.624,59	0,0123	16.107,29	0,0429
ARIMA(p=3, d=0, q=0)	4.616,93	0,0123	16.267,99	0,0433
ARIMA(p=2, d=0, q=2)	4.785,98	0,0128	16.149,04	0,0430
ARIMA(p=4, d=0, q=0)	4.778,22	0,0127	16.184,69	0,0431
ARIMA(p=3, d=0, q=1)	4.665,73	0,0124	16.246,94	0,0433
ARIMA(p=2, d=0, q=3)	4.941,18	0,0132	16.269,53	0,0433

Fonte: próprio autor.

Tabela 42 – Medidas de acurácia das previsões dos valores futuros da série de dados 3.

Modelo	**MAE**	**MAPE (%)**	**RMSE**	**CVRMSE (%)**
ARIMA(p=2, d=0, q=0)	5.752,12	0,0153	6.357,24	0,0169
ARIMA(p=2, d=0, q=1)	7.454,66	0,0198	8.200,34	0,0218
ARIMA(p=3, d=0, q=0)	5.433,02	0,0145	6.016,93	0,0160
ARIMA(p=2, d=0, q=2)	3.329,73	0,0089	4.244,24	0,0113
ARIMA(p=4, d=0, q=0)	3.596,79	0,0096	4.405,09	0,0117
ARIMA(p=3, d=0, q=1)	4.697,88	0,0125	5.265,54	0,0140
ARIMA(p=2, d=0, q=3)	9.493,73	0,0253	10.355,71	0,0276

Fonte: próprio autor.

às estimativas da série de dados 2 (Tabela 38). Contudo, as previsões para valores futuros da série de dados 3 (Tabela 42) indicam uma previsão mais precisa que os indicativos dos dois testes anteriores (Tabelas 36 e 39). Ainda, há um ganho em precisão da previsão ao utilizar o modelo ARIMA(p=2, d=0, q=2), o quarto melhor indicado pela otimização dos hiperparâmetros via AIC.

Com base nos resultados dos modelos ARIMA sobre os dados de pressão dos poços, uma metodologia de detecção de anomalias em tempo real por meio de modelos ARIMA foi criada, sendo descrita e aplicada, conforme apresentado a seguir.

4.2.2 Metodologia de Detecção de Anomalias com modelos ARIMA

De acordo com os exemplos anteriores, observa-se que os modelos ARIMA são capazes de identificar mudanças pontuais no comportamento dos dados, a partir da detecção de mudanças nos valores da média móvel da série.

Logo, os modelos ARIMA podem auxiliar na identificação de mudanças iminentes de tendência, antecipando eventos indesejáveis e dando suporte aos modelos de detecção de anomalias.

A metodologia de aplicação do modelo para detecção de anomalias é realizada da seguinte forma:

- Os hiperparâmetros para determinação do modelo ARIMA (e seus parâmetros) foram avaliados por meio dos primeiros dados de pressão do poço;
- Observações futuras foram previstas pelo modelo e comparadas pelo cálculo do Erro Médio Absoluto (MAE) com os dados reais adquiridos após previsão;
- Caso o MAE esteja dentro da tolerância, os parâmetros do modelo ARIMA são atualizados com os novos dados, caso contrário uma anomalia é identificada;
- Para que o algoritmo continue sua verificação de anomalias, em um caso anômalo os novos hiperparâmetros são avaliados com os dados futuros, de forma a criar um novo modelo ARIMA que absorva a nova tendência dos dados;
- Assim, o algoritmo continua sua previsão de dados e verificação de anomalias.

A seguir, apresentam-se aplicações dos modelos ARIMA para identificação de mudanças pontuais no comportamento dos dados.

4.2.2.1 Estudo de Caso 1

Aplica-se a metodologia de identificação de anomalias por meio de modelos ARIMA sobre a pressão (PT) de um poço em produção.

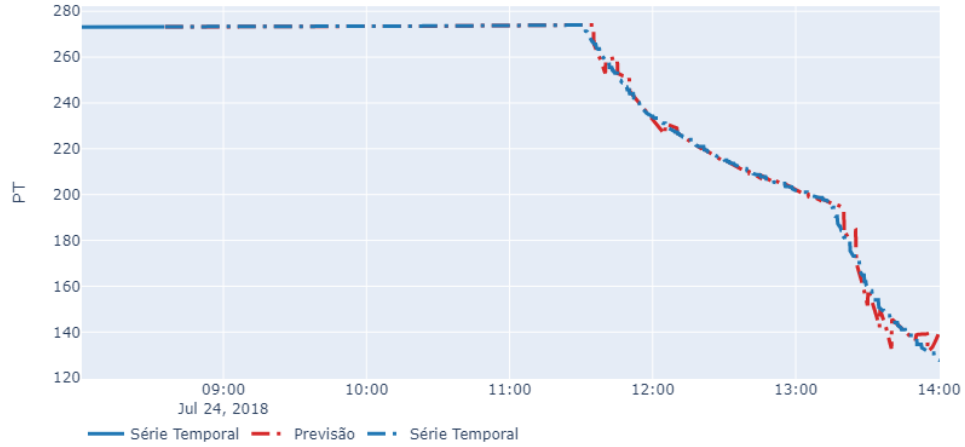
A Figura 44 apresenta os dados aferidos, as primeiras três horas e trinta minutos da janela apresentada representam um poço em estado normal de produção. Este segmento de estado normal possui aproximadamente 800 dados.

Após este momento, o poço começa a apresentar um comportamento anormal, onde o algoritmo teve que determinar novos hiperparâmetros para o modelo ARIMA, alguns outros

pontos subsequentes são identificados pelo algoritmo como anômalos, representados por uma nova mudança em seu comportamento.

A linha contínua em azul representa os primeiros 100 dados de pressão do poço, o pontilhado em azul representa os dados reais futuros, enquanto que a linha pontilhada em vermelho representa os valores previstos. Trechos de descontinuidade na linha vermelha indicam os pontos onde os hiperparâmetros do modelo ARIMA foram modificados.

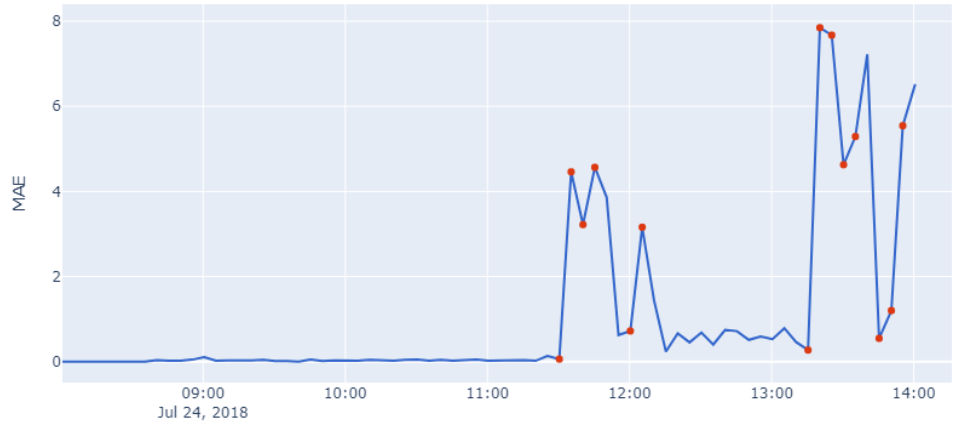
Figura 44 – Dados do poço número 1 e previsões via ARIMA.



Fonte: próprio autor.

A Fig. 45 apresenta o MAE dos valores previstos, as anomalias são identificadas pelo MAE exceder a tolerância adotada e representados por pontos vermelhos no gráfico.

Figura 45 – MAE dos valores previstos via ARIMA para o poço número 1.



Fonte: próprio autor.

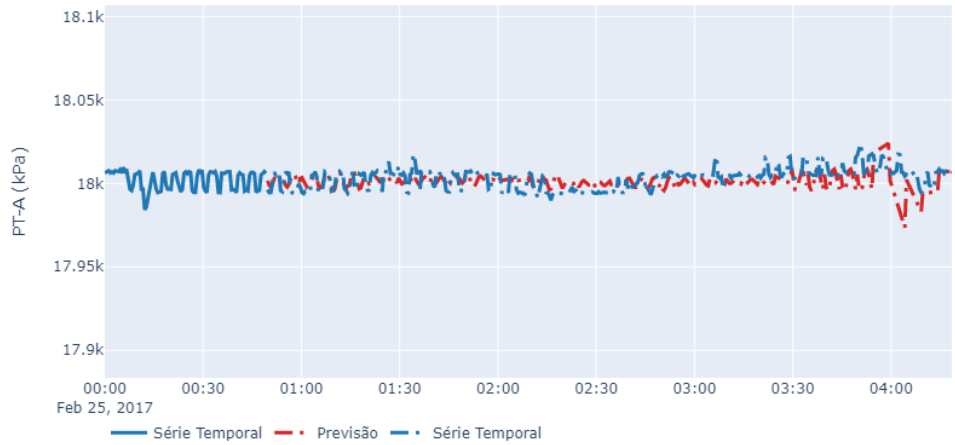
4.2.2.2 Estudo de Caso 2

Um caso de estudo para aplicação direta da metodologia é o poço número 2, apresentado como um dos motivadores da verificação de integridade em tempo real. Este poço será aplicado também com a próxima metodologia, para comparação.

Como no caso anterior, a metodologia utiliza modelos ARIMA que são atualizados conforme novos dados são adquiridos, estes dados são comparados com seus valores previstos de forma a detectar variações de tendências.

A Figura 46 apresenta o comportamento da pressão PT do poço ainda em seu estado normal, observa-se uma diferença entre os valores previstos e adquiridos do poço, diferente do caso anterior, mas o Erro Médio Absoluto se mantém dentro da tolerância, o que indica um estado normal da pressão do poço.

Figura 46 – Janela de operação normal do poço número 2 e previsões via ARIMA.

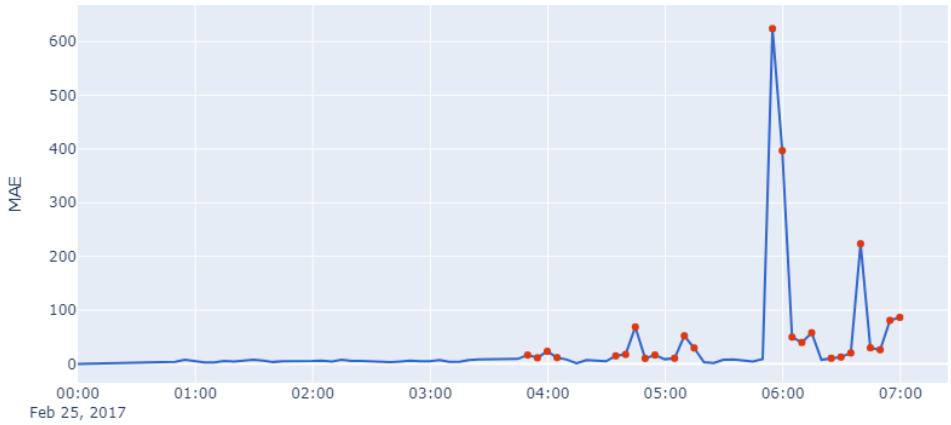


Fonte: próprio autor.

Com o decorrer do tempo, o MAE ultrapassa a tolerância adotada, indicando uma variação errática do comportamento da pressão. A Figura 47 ilustra a variação do parâmetro, enquanto que a Figura 48 expõe os dados de pressão durante sua transição para o estado anômalo, conforme verificado pela variação do MAE.

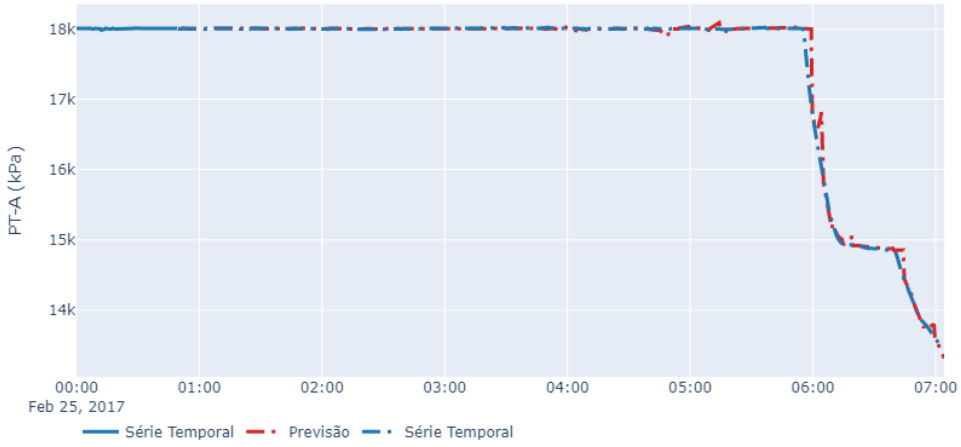
O pico do MAE na Figura 47 em torno das 06:00 horas identificou uma anomalia no sistema, novamente detectada próximo de 06:40 horas. Estas anomalias correspondem a anomalias reais visualmente perceptíveis na Figura 48, demonstrando a eficácia da metodologia em tempo real.

Figura 47 – MAE dos valores previstos via ARIMA para o poço número 2.



Fonte: próprio autor.

Figura 48 – Dados do poço número 2 e previsões via ARIMA.



Fonte: próprio autor.

5 CONCLUSÃO E SUGESTÕES PARA TRABALHOS FUTUROS

O presente texto aborda técnicas de ciência dos dados no contexto do projeto e do monitoramento de integridade estrutural de poços, no tocante à inferência estatística à previsão de séries temporais, inseridas no cenário de transformação digital do setor.

No âmbito de inferência estatística, os métodos de parametrização de distribuições de probabilidade hipotéticas e os testes de aderência são discutidos, indicando quais as vantagens da aplicação de cada um dos métodos expostos. Ainda, técnicas complementares de seleção de modelos por meio de critérios de informação, quantificação de incertezas no processo de aquisição de dados e uma metodologia de imposição de limites numa distribuição hipotética são sugeridos.

Uma extensa aplicação dos métodos indicados é realizada sobre dados de manufatura de tubulares, com vistas à quantificação de incertezas dos parâmetros de projeto de revestimento.

A inferência sobre os dados mecânicos, aferidos em ensaios destrutivos, portanto, em amostra única de tamanho regular, foi conclusiva, com resultados positivos para quase todos os testes de aderência realizados.

Já a inferência sobre os dados de características geométricas foi mais extensa, devido à amostragem dos dados e ao período de produção contemplado. Os dados se portaram de forma multimodal, demandando a aplicação de técnicas auxiliares de pré-tratamento antes de sua inferência estatística a qual, por fim, foi conclusiva em caracterizar as séries de dados.

Entende-se que a inferência é parte fundamental do projeto de revestimentos por métodos probabilísticos, os quais são recomendados em normas recentes de projetos de poços de petróleo, a exemplo da API TR 5C3 (2008). A metodologia explorada no presente texto pode ser aplicada em dados reais, os quais não puderam ser apresentados por questões de proteção de dados, para a caracterização de variáveis e posterior análise confiabilística, como apresentado em Gouveia (2014), Oliveira (2017) e Silva (2018).

No que tange às técnicas de previsão de dados, modelos do tipo ARIMA foram aplicados em séries históricas reais de pressões medidas em poços de petróleo em produção, com a finalidade de estimar os valores de pressões futuras. Após uma análise dos resultados de previsão, avaliados por meio de medidas de desempenho, uma metodologia de detecção de anomalias com modelos ARIMA é sugerida e aplicada, de forma a auxiliar na identificação de mudanças iminentes de tendência, antecipando eventos indesejáveis.

A metodologia de detecção de anomalias pode ser utilizada como suporte à técnicas mais robustas de previsão, como é o caso de previsões por meio de redes neurais artificiais (RNA). Técnicas mistas de detecção via RNA com modelos ARIMA são utilizadas em outras áreas do conhecimento, sugere-se como uma extensão deste trabalho uma investigação da qualidade de

técnicas RNA, comparadas com a precisão e velocidade de captação de anomalias por meio de modelos ARIMA, como também uma combinação das duas técnicas como alternativa, com possíveis ganhos em precisão e robustez das previsões.

Diante do exposto, reafirma-se que as técnicas de ciência dos dados são de grande valia para o cenário de transformação digital no contexto do projeto e do monitoramento de integridade estrutural de poços. Temas como os abordados aqui são uma demanda premente da indústria, e entende-se que os pontos explorados podem contribuir nesse contexto.

REFERÊNCIAS

ANG, A.; TANG, W. *Probability concepts in engineering : emphasis on applications in civil & environmental engineering*. New York: Wiley, 2007. ISBN 9780471720645. Citado na página 22.

API TR 5C3. *Technical Report on Equations and Calculations for Casing, Tubing, and Line Pipe Used as Casing or Tubing; and Performance Properties Tables for Casing and Tubing*. 1st. ed. Washington, D.C., 2008. Citado 4 vezes nas páginas 15, 27, 36 e 81.

ARMSTRONG, J. S.; COLLOPY, F. Error measures for generalizing about forecasting methods: Empirical comparisons. *International journal of forecasting*, Elsevier, v. 8, n. 1, p. 69–80, 1992. Citado na página 35.

BAIRD, G. L.; HARLOW, L. L. Does one size fit all? a case for context-driven null hypothesis statistical testing. *Journal of Modern Applied Statistical Methods*, v. 15, n. 1, p. 7, 2016. Citado na página 49.

DOROUDI, S. The bias-variance tradeoff: How data science can inform educational debates. *AERA Open*, SAGE Publications Sage CA: Los Angeles, CA, v. 6, n. 4, p. 2332858420977208, 2020. Citado 2 vezes nas páginas 24 e 25.

DU, X.; HU, Z. First order reliability method with truncated random variables. *Journal of Mechanical Design*, American Society of Mechanical Engineers Digital Collection, v. 134, n. 9, 2012. Citado na página 27.

GERKE, G. *GE Digital Twin Technology Revamps Offshore Oil Operations*. [S.l.], 2018 (acessado em 24/07/2018). Disponível em: <https://www.efficientplantmag.com/2018/03/ge-digital-twin-technology-revamps-offshore-oil-operations/>. Citado na página 14.

GIBBONS, J. D.; CHAKRABORTI, S. *Nonparametric Statistical Inference: Revised and Expanded*. [S.l.]: CRC press, 2014. Citado na página 16.

GOUVEIA, L. P. d. Avaliação da confiabilidade em tubos de revestimento de poços de petróleo. Universidade Federal de Alagoas, 2014. Citado 2 vezes nas páginas 15 e 81.

HANSEN, L. P. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, JSTOR, p. 1029–1054, 1982. Citado na página 16.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado 2 vezes nas páginas 17 e 29.

KLEVER, F.; TAMANO, T. A new octg strength equation for collapse under combined loads. *SPE Drilling & Completion*, Society of Petroleum Engineers, v. 21, n. 3, Sep 2006. Citado na página 36.

LCCV. *Modelos e Ferramentas Computacionais para Apoio ao Dimensionamento de Revestimentos de Poços*. 2019. Relatório Técnico Parcial 02. Citado na página 16.

LCCV. *Desenvolvimento de Ferramentas Computacionais para Modelagem em Tempo Real da Integridade de Estrutura de Poço*. 2020. Relatório Parcial 1. Citado na página 15.

MAYANI, M. G. et al. Drilling digital twin success stories the last 10 years. In: SOCIETY OF PETROLEUM ENGINEERS. *SPE Norway One Day Seminar*. [S.l.], 2018. Citado na página 14.

MCLACHLAN, G. J.; PEEL, D. *Finite mixture models*. [S.l.]: John Wiley & Sons, 2004. Citado na página 49.

NADHAN, D. et al. Drilling with digital twins. In: SOCIETY OF PETROLEUM ENGINEERS. *IADC/SPE Asia pacific drilling technology conference and exhibition*. [S.l.], 2018. Citado na página 14.

OLIVEIRA, F. L. Carregamentos aleatórios no dimensionamento probabilístico de revestimentos de poços de petróleo. Universidade Federal do Rio de Janeiro, 2017. Citado 2 vezes nas páginas 15 e 81.

PAWITAN, Y. *In all likelihood: statistical modelling and inference using likelihood*. [S.l.]: Oxford University Press, 2001. Citado na página 22.

SEVERINI, T. A. *Likelihood methods in statistics*. [S.l.]: Oxford University Press, 2000. Citado na página 16.

SILVA, T. B. d. Contribuição à análise de integridade em sistemas de revestimento via confiabilidade estrutural. Universidade Federal de Alagoas, 2018. Citado 3 vezes nas páginas 15, 16 e 81.

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*, Elsevier, v. 181, p. 106223, 2019. Citado 2 vezes nas páginas 63 e 69.

Virtanen, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020. Citado 2 vezes nas páginas 17 e 51.

VU, K. M. *The ARIMA and VARIMA time series: their modelings, Analyses and Applications*. [S.l.]: AuLac Technologies Inc., 2007. Citado na página 17.

WALPOLE, R. E. e. a. *Probability and Statistics for Engineers and Scientists*. [S.l.]: Pearson Education Limited, 2011. Citado na página 21.

WOOLDRIDGE, J. M. Applications of generalized method of moments estimation. *Journal of Economic perspectives*, American Society of Mechanical Engineers Digital Collection, v. 15, n. 4, p. 87–100, 2001. Citado na página 21.

YUAN, Z. et al. Casing failure mechanism and characterization under hpht conditions in south texas. In: EUROPEAN ASSOCIATION OF GEOSCIENTISTS & ENGINEERS. *IPTC 2013: International Petroleum Technology Conference*. [S.l.], 2013. p. cp–350. Citado na página 14.