UNIVERSIDADE DE SÃO PAULO

Escola de Engenharia de São Carlos

Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA).

Aluno: Fabrizio Silva Campos
Orientadora: Prof. Dra. Maria Mercedes Gamboa Medina

EESC · USP

UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE SÃO CARLOS

FABRIZIO SILVA CAMPOS

Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA).

São Carlos

2023

FABRIZIO SILVA CAMPOS

Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA).

Dissertação apresentada à Escola de Engenharia de São Carlos da Universidade de São Paulo, como requisito parcial para a obtenção do Título de Mestre em Ciências: Engenharia Hidráulica e Saneamento.

Área de concentração: Hidráulica e Saneamento

Orientadora: Prof. Dra. Maria Mercedes Gamboa Medina

VERSÃO CORRIGIDA

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Campos, Fabrizio Silva
C198p Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA). / Fabrizio Silva Campos; orientadora Maria Mercedes Gamboa-Medina. São Carlos, 2023.

Dissertação (Mestrado) - Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento e Área de Concentração em Hidráulica e Saneamento -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2023.

1. Sinais de pressão. 2. Modelos autorregressivos. 3. Abastecimento de água. 4. Smart water. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE JULGAMENTO

Candidato: Engenheiro **FABRIZIO SILVA CAMPOS.**

Título da dissertação: "Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA).".

Data da defesa: 01/12/2023.

Comissão Julgadora	**Resultado**
Profa. Dra. Maria Mercedes Gamboa Medina (Orientador) (Escola de Engenharia de São Carlos/EESC-USP)	Aprovado
Profa. Associada Airlane Pereira Alencar (Instituto de Matemática e Estatística/IME-USP)	Aprovado
Prof. Dr. Bruno Melo Brentan (Universidade Federal de Minas Gerais/UFMG)	Aprovado

Coordenador do Programa de Pós-Graduação em Engenharia Hidráulica e Saneamento:
Prof. Assoc. **Juliano Jose Corbi**

Presidente da Comissão de Pós-Graduação:
Prof. Titular **Carlos De Marqui Junior**

Às futuras gerações da família Silva Campos.

AGRADECIMENTOS

A gratidão é o bom sentimento que devemos cultivar durante a caminhada da vida, com a certeza de que tudo que acontece coopera para um propósito maior, assim, agradeço a Deus por cada detalhe desenhado para esta etapa de minha vida.

Aos meus pais José Carlos Campos e Eliade Silva Campos, por seus sacrifícios para sempre proporcionar os estudos aos seus filhos.

As minhas irmãs Sabrina Silva Campos, por ouvir atentamente meus devaneios, e Maria Eduarda Campos, por incentivar a cada momento.

Ao meu companheiro José Henrique Pereira, que sorridente em todo tempo, foi minha fonte de energia para seguir em frente nos momentos difíceis e sempre paciente me aconselhou e apoiou no desenvolvimento desse trabalho.

À professora Maria Mercedes Gamboa Medina, pela orientação deste trabalho e por auxiliar a fazer todas as peças do quebra-cabeças se encaixarem com sua genialidade. Por ser gentil e compreensiva, demonstrando sua humanidade desde os períodos de isolamento social no contexto de pandemia. Por compartilhar parte de sua pesquisa de doutorado para viabilização desta pesquisa, e por ensinar com tamanho talento.

À professora Airlane Pereira Alencar, pelo oferecimento da disciplina de análise de séries temporais, a qual desmistificou tantos conceitos até então nebulosos para este engenheiro e por toda contribuição ao longo da pesquisa.

Ao professor Bruno Melo Brentan, por suas contribuições e pelo aceite na participação da banca de avaliação.

A todos meus amigos de São Carlos, Miguel Campos, Ching Yu, Lizeth, Xiana, Vinícius, Emanuel, Greice, Miguel Sampaio, Alexandre e Iris, que fizeram com que o período de desenvolvimento deste projeto fosse mais leve, e por se tornarem uma família para mim, me ensinando a encontrar irmandade onde quer que esteja. Em especial a minha amiga Ianca Peixoto Miranda, por todos os chás, cafés com grãos moídos na hora, almoços, e todos os outros momentos que pudemos compartilhar mútuo apoio e risadas.

Aos funcionários e professores do Departamento de Hidráulica e Saneamento da USP São Carlos por manterem por tantos anos um Programa de Pós-Graduação de excelência, e por receberem sempre de braços abertos.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), a quem também agradeço.

“All models are wrong, but some are useful”

George Box (1979)

RESUMO

CAMPOS, F. S. **Previsão de séries temporais de pressão em redes de distribuição de água aplicando modelos generalizados autorregressivos de médias móveis (GARMA).** 2023. Dissertação (Mestrado) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

A eficiência na gestão do sistema de distribuição de água é essencial para fornecer água potável de forma contínua e adequada. Companhias de saneamento estão investindo em automação da coleta de dados em campo por meio de sensores, como vazão e pressão, mas ainda não exploram totalmente as informações disponíveis nos dados históricos. Este estudo visa aplicar análise de séries temporais e o modelo GARMA para modelar séries de pressão em redes de abastecimento de água real. Foram descritos o comportamento das séries temporais de pressão e ajustados modelos ARIMA, SARIMA e GARMA a dados de pressão classificados como de comportamento normal. Comparando o desempenho dos modelos em diferentes sensores, períodos e horizontes de previsão, pode-se concluir que o modelo GARMA(2,0) com densidade de probabilidade gama e séries de Fourier é adequado para essas séries temporais. A aplicação da função de densidade de probabilidade gama permitiu lidar com a heterocedasticidade inerente ao mecanismo gerador dos dados e a transformação da série de pressão em "Energia Consumida" permitiu generalizar o modelo para a maioria dos sensores do setor estudado. Isso proporcionou previsões mais precisas em vários sensores do mesmo setor hidráulico. Além disso, essa abordagem mostrou-se útil para identificar anomalias e estabelecer um arcabouço técnico para futuras pesquisas na área de controle operacional. Assim, conclui-se que a análise de séries temporais aplicada a séries de pressão em redes de abastecimento é uma abordagem robusta que pode melhorar a modelagem das séries históricas obtidas pelas companhias de saneamento, transformando uma massa de dados em informações úteis, fornecendo previsões precisas e insights valiosos para o processo de tomada de decisão.

Palavras-chave: Sinais de pressão. Modelos autorregressivos. Abastecimento de água. Smart water.

ABSTRACT

CAMPOS, F. S. **Forecasting of pressure time series in water distribution networks using generalized autoregressive moving average models (GARMA)**. 2023. Dissertação (Mestrado) – São Carlos School of Engineering, University of São Paulo, 2023.

Efficiency in the management of water distribution systems is crucial to ensure a continuous and adequate provision of potable water. Sanitation companies are investing in field data collection automation through sensors, such as flow and pressure, yet they have not fully harnessed the available information in historical data. This study aims to apply time series analysis and the GARMA model to model pressure series in real water supply networks. The behavior of pressure time series was described, and ARIMA, SARIMA, and GARMA models were fitted to pressure data classified as normal behavior. By comparing the performance of the models across different sensors, time periods, and forecast horizons, it can be concluded that the GARMA(2,0) model with gamma probability density and Fourier series is suitable for these time series. The application of the gamma probability density function enabled the handling of the inherent heteroscedasticity in the data-generating mechanism. The transformation of the pressure series into "Consumed Energy" allowed for generalizing the model to most sensors in the studied sector, providing more accurate forecasts for multiple sensors in the same hydraulic sector. Furthermore, this approach has proven to be useful for anomaly detection and establishing a technical framework for future research in operational control. In conclusion, the analysis of time series applied to pressure series in water supply networks represents a robust approach that can enhance the modeling of historical series obtained by sanitation companies, transforming a wealth of data into valuable information, and offering precise forecasts and valuable insights for the decision-making process.

Keywords: Pressure signals. Autorregressive models. Water supply. Smart Water

LISTA DE ILUSTRAÇÕES

Figura 1 - Componentes de uma série temporal. 18
Figura 2 - Ilustração da diferença entre as abordagens realizadas pela função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP). 22
Figura 3 - Esquema de uma modelo ARIMA Sazonal ou SARIMA. 28
Figura 4 - Estrutura do esquema metodológico da pesquisa 32
Figura 5 - Esquema do modelo hidráulico da rede de abastecimento estudada com a localização dos sensores e dos pontos de simulação de eventos. 34
Figura 6 - Séries temporais de pressão (mca) para os 9 sensores monitorados na área de estudo. 35
Figura 7 - Curva adimensional média da vazão de consumo e da pressão em uma rede de abastecimento de água em função da hora do dia. 38
Figura 8 - Gráfico de uma função de densidade de probabilidade gama genérica 40
Figura 9 - Gráfico de uma função de densidade de probabilidade normal genérica. 41
Figura 10 - Gráfico de trecho da série original para observação de padrões de comportamento. 42
Figura 11 - Gráfico de trecho da série transformada na forma de Energia Consumida (mca). 42
Figura 12 - Estrutura de inovação - modelo diamante duplo 44
Figura 13 - Esquema de separação da série temporal em conjuntos de treino e teste de acordo com a metodologia de *rolling analysis* e *backtest.* 47
Figura 14 - Detalhamento das séries temporais de pressão de cada um dos sensores monitorados. 50
Figura 15 - Diagrama de caixas dos valores de pressão para cada um dos sensores monitorados. 52
Figura 16 - Diagrama de caixas dos valores de Energia Consumida após transformação de dados para cada um dos sensores monitorados. 53
Figura 17 - Gráfico da série de pressão (mca) observados referente ao sensor s2, em todo período de monitoramento. 54
Figura 18 - Gráfico da série de pressão (mca) observados referente ao sensor s5, em todo período de monitoramento. 54
Figura 19 - Matriz de correlação entre os dados observados dos sensores monitorados. 55
Figura 20 - Gráfico de correlação de elipses das séries originais. 56

Figura 21 - Gráfico da função de autocorrelação e autocorrelação parcial referente ao sensor s7. 57
Figura 22 - Variação da pressão na rede de abastecimento em função da hora do dia e do dia da semana de forma adimensional referente ao sensor s7. 58
Figura 23 - Variação da pressão na rede de abastecimento de água em função do dia da semana e do horário do dia plotado em eixo radial. 59
Figura 24 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo SARIMA ajustado para a série da primeira janela móvel. 61
Figura 25 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo SARIMA ajustado para a série da primeira janela de análise. 62
Figura 26 - Análise de resíduos do modelo ajustado SARIMA para a base de treino da JM#1. 63
Figura 27 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo ARIMA+Harmônico ajustado para a série da primeira janela móvel. 66
Figura 28 - Análise de resíduos do modelo ajustado ARIMA+Harmônico para a base de treino da JM#1. 67
Figura 29 - Energia Consumida (mca) e série ajustada com modelo GARMA(2,0) com distribuição gama. 68
Figura 30 - Resíduos do modelo GARMA(2,0) com distribuição gama ao longo do tempo. 68
Figura 31 – Gráficos de análise dos resíduos do modelo GARMA(2,0) com distribuição gama. a) Autocorrelograma e b) Curva densidade de probabilidade dos resíduos. 69
Figura 32 - Diagrama de disponibilidade de dados. 71
Figura 33 - Frequência de melhor desempenho de cada modelo em função da métrica e do horizonte de previsão. 72
Figura 34 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo SARIMA. 75
Figura 35 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo ARIMA + Harmônico. 77
Figura 36 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo GARMA(2,0) com distribuição gama. 79
Figura 37 - Gráfico de detecção de observações de pressão fora do intervalo de confiança da previsão feita com modelo GARMA(2,0) 82

SUMÁRIO

1 INTRODUÇÃO 14

1.1 Objetivo Geral 15

1.2 Objetivos Específicos 15

2 FUNDAMENTAÇÃO TEÓRICA 16

2.1 Fundamentos de análise de séries temporais 16

2.1.1 Tendência, sazonalidade e estacionariedade 18

2.1.2 Sazonalidade complexa 20

2.1.3 Autocorrelação 21

2.2 Aplicações de análise de séries temporais a dados de monitoramento de redes de distribuição de água 22

2.2.1 Perfil de variação de consumo de água 22

2.2.2 Previsão e classificação na detecção de eventos anormais em redes de distribuição de água 23

2.3 Modelos de previsão para séries temporais 26

2.3.1 Modelos autorregressivos 26

2.3.2 Modelos combinados autorregressivos e de médias móveis 27

2.3.3 Modelo SARIMA 27

2.3.4 Identificação de modelos do tipo ARIMA 28

2.3.5 Modelos Harmônicos 29

2.3.6 Modelo Generalizado Autorregressivo de Médias Móveis - GARMA 30

3 MATERIAIS E MÉTODOS 32

3.1 Área de estudo e monitoramento de dados 33

3.2 Pré-processamento e modelagem de dados 35

3.3 Descrição do comportamento da Série 36

a) Tendência, sazonalidade e estacionariedade 36

b) Autocorrelação 37

c) Perfil de variação de consumo de água 37
3.4 Ajuste de modelos de previsão 38
3.4.1 Ajuste do modelo SARIMA 38
3.4.2 Ajuste do modelo ARIMA + Harmônico 39
3.4.3 Processo de modelagem das séries temporais de pressão com modelo GARMA 39
3.4.3.1 Pré-processamento de dados para modelagem com GARMA 40
3.4.3.2 Ajuste do modelo GARMA 43
3.4.3.3 Seleção do conjunto de ajuste e do horizonte de previsão do modelo GARMA 43
3.5 Organização de testes de previsão com os modelos 45
3.6 Métricas de avaliação de modelos 47
3.7 Exploração da resposta da previsão em cenários de vazamentos 48
4 RESULTADOS E DISCUSSÕES 50
4.4 Descrição do comportamento da série 50
4.5 Ajuste dos modelos e Análise de Resíduos 60
4.2.1 Ajuste do modelo SARIMA 60
4.2.2 Ajuste do modelo ARIMA+Harmônico 64
4.2.3 Ajuste do modelo GARMA 67
4.3 Seleção da combinação entre conjunto de dados de ajuste e horizonte de previsão 70
4.3.1 Avaliação do desempenho entre modelo completo e modelo mensal 71
4.3.2 Avaliação comparativa entre horizontes de previsão 73
4.4 Análise das previsões com aplicação da metodologia de janelas móveis 73
4.4.1 Previsões em janelas móveis com modelo SARIMA 73
4.4.2 Previsões em janelas móveis com modelo ARIMA+Harmônico 76
4.4.3 Previsões em janelas móveis com modelo GARMA 78
4.5 Detecção de vazamentos em eventos simulados 81
5 CONCLUSÃO 83
REFERÊNCIAS 87

1 INTRODUÇÃO

A gestão eficiente do sistema de abastecimento de água é fundamental para garantir o fornecimento contínuo e adequado de água potável à população. Desta forma as companhias de saneamento tendem a dedicar seus esforços em busca de melhorar seus processos de tomada de decisão, que por sua vez esbarra na barreira do entendimento sobre o comportamento das redes de distribuição de água.

Em geral, os investimentos recentes de diversas cidades do Brasil em automação da coleta de dados em campo a partir de sensores, como monitoramento da vazão e da pressão, por si só não refletem um aumento significativo do nível de conhecimento sobre os padrões de comportamento da rede de abastecimento de água. O que se tem muitas vezes são centros de controle operacional com baixo nível de maturidade, que ainda dependem de profissionais habilitados analisando visualmente gráficos nas telas para operar o sistema com base em sua experiência (HAMILTON et al., 2021). Ou seja, há muita informação escondida nos próprios históricos de dados que não são aproveitadas.

Diante da complexidade que envolve o gerenciamento e operação de um sistema de abastecimento de água, contar com modelos capazes de extrair informações de qualidade das séries temporais monitoradas e que sejam capazes de compor suporte ao processo de tomada de decisão, tornando-o mais rápido e assertivo, tende a colaborar com o aumento da eficiência operacional por parte das companhias de saneamento (GAMBOA-MEDINA, 2017). Neste sentido, diversas são as iniciativas por parte da academia na construção desses modelos, seja para séries de vazão ou pressão, com objetivos diversos como previsão da demanda de água, detecção e localização de vazamentos, preenchimento de falhas em séries de monitoramento ou otimização de conjuntos moto bombas.

Contudo, o fato de tais pesquisas terem um foco maior em trabalhos de aplicação, uma lacuna que se encontra é a necessidade de aprofundar a descrição das séries, identificação de padrões e o relacionamento com seu mecanismo gerador. Pois, durante os esforços de buscar a aplicação de seus métodos, os pesquisadores encontram desafios e limitações em seus modelos, tais como citados por Yipeng et al. (2017) em sua revisão da literatura: variações periódicas e não-estacionárias de demanda se tornam fontes de incerteza, quando deveriam ser fonte de informação; modelos determinísticos que guiam a resultados falhos, quando as séries temporais provém de sistemas estocásticos e sua distribuição de probabilidade adequada deveria ser

levada em consideração; necessidade de grandes transformações de dados, tais como suavização da série para redução de sua complexidade, quando os modelos deveriam considerar a complexidade da série como resultante da variância intrínseca ao mecanismo que a gerou.

Com o objetivo de contribuir para superar essas limitações, a presente pesquisa apresenta uma abordagem guiada pela mineração de conhecimento a partir de séries temporais de pressão em uma rede real de abastecimento de água. Isso se dará sob abordagem da análise de séries temporais, que possibilita descrição da série em tendências e sazonalidades, exploração da correlação entre múltiplas séries, previsão de valores futuros e controle estatístico de processo. Não é visado aqui esgotar o assunto, mas colaborar com arcabouço técnico na discussão de melhores abordagens a um problema presente na realidade de todos os sistemas de abastecimento de água, e consequentemente todas as cidades e pessoas.

1.1 Objetivo Geral

Esta pesquisa tem por objetivo geral descrever, analisar, modelar e prever sinais de pressão em uma rede de abastecimento de água real aplicando conceitos de análise de séries temporais e através do modelo GARMA.

1.2 Objetivos Específicos

a) Descrever estatisticamente o comportamento das séries temporais de pressão em redes reais de abastecimento de água;
b) Ajustar modelos autorregressivos, ARIMA[1], SARIMA[2] e GARMA[3], em série de dados de pressão classificada como de comportamento normal da rede;
c) Comparar o desempenho dos modelos quanto a acurácia na previsão, considerando séries adquiridas com diferentes sensores e períodos do ano, e diferentes horizontes de previsão;
d) Explorar viabilidade de técnica de detecção de vazamentos baseada no desvio entre os dados de monitoramento e a previsão realizada pelo modelo.

[1] Modelo Autorregressivo Integrado de Médias Móveis
[2] Modelo Sazonal Autorregressivo Integrado de Médias Móveis
[3] Modelo Generalizado Autorregressivo de Médias Móveis

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são abordados os principais temas que englobam o escopo deste trabalho, através de uma descrição detalhada sobre os fundamentos que levaram a proposição do tema e levantamento da hipótese de pesquisa. Como por exemplo, a abordagem dos sinais de carga de pressão de uma rede de distribuição para abastecimento de água através de séries temporais, fatores que devem ser levados em consideração na descrição do mecanismo gerador e como a análise de séries temporais auxilia neste processo. Serão introduzidos os conceitos principais sobre os modelos utilizados no estudo.

2.1 Fundamentos de análise de séries temporais

As séries temporais podem ser definidas como “uma sequência de observações tomadas sequencialmente no tempo” (BOX et al., 2008) ou ainda, segundo Chatfield (1996) “uma coleção de observações feitas sequencialmente no tempo”, o que pode-se observar em comum na definição dos autores é a ideia de sequência no tempo, que é fundamental para o estudo das séries temporais pois neste ponto está contida uma das principais características das séries temporais, sua dependência entre observações adjacentes.

Os autores Box et al. (2008) afirmam que a dependência entre observações adjacentes são um recurso intrínseco das séries temporais, e a área de estudo de análise de séries temporais consiste em técnicas para analisar esta dependência. Sobre essa dependência, Chatfield (1996) afirma que “quando sucessivas observações são dependentes, pode-se prever valores futuros a partir de observações passadas”. Novamente há o apontamento da dependência como característica essencial entre os pares de observações, e uma indicação: quando essa dependência é estudada e conhecida você pode predizer valores futuros.

As técnicas citadas por Box et al. (2008) podem ser entendidas como ferramentas e modelos estocásticos e dinâmicos, classificados assim como referência ao tipo de sistema que buscam representar. Diferentemente de sistemas/séries temporais determinísticos, onde os valores futuros podem ser exatamente preditos, em sistemas/séries temporais estocásticos, o futuro é apenas parcialmente determinado por valores passados, e então as predições exatas são impossíveis e devem dar lugar a ideia de que os valores futuros possuem uma distribuição de probabilidade condicionada ao conhecimento dos valores passados.

A análise de séries temporais se concentra em séries do tipo discreta, que tem por principal característica observações tomadas em intervalos igualmente espaçados (CHATFIELD, 1996). As séries discretas podem ser amostradas, quando são realizadas leituras em intervalos de tempo pré-definidos, por exemplo o monitoramento de vazão de um rio. Outro tipo é definido quando a variável não possui um valor instantâneo, mas pode-se agregá-lo, por exemplo a precipitação diária.

Segundo Chatfield (1996), há diversos objetivos em se analisar uma série temporal, e podem ser resumidos e classificados em descrição, explicação, predição e controle.

A descrição consiste em obter informações sobre a série com base na observação de sua representação gráfica, com uma simples plotagem é possível visualizar a existência de tendência de crescimento ou decrescimento, se há padrões sazonais, além de permitir detectar possíveis *outliers* ou pontos de inflexão.

A explicação possui foco de análise multivariada, quando se possui duas ou mais variáveis e busca-se por utilizar uma variável para explicar a outra. É uma análise muito útil quando se busca maior entendimento sobre o mecanismo gerador das séries.

A predição resume as etapas de conhecimento sobre a série temporal, valendo-se dessa informação para ajustar modelos que, através das observações passadas, sejam capazes de prever os valores futuros.

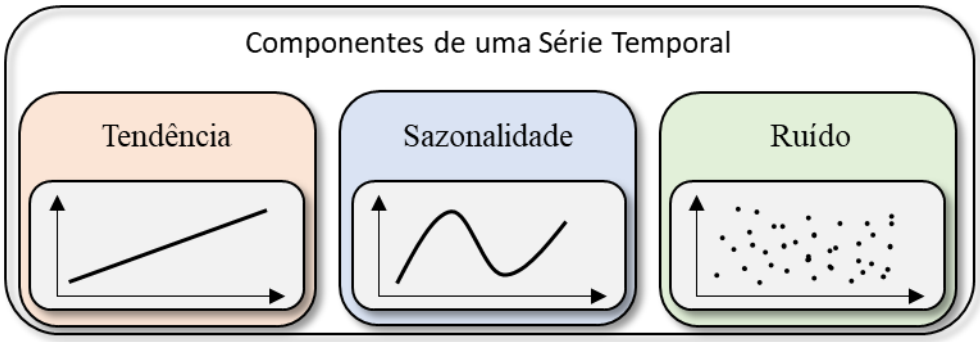
O controle está relacionado com medidas de qualidade e infere sobre o controle de um processo. Utiliza-se do conhecimento sobre a série temporal para instituir controle estatístico de processo, proporcionando avaliações de qualidade e funcionando como suporte à tomada de decisão.

Chatfield (1996) esclarece que “predição está diretamente relacionada com problemas de controle em diversas situações, pode-se por exemplo, prever falhas em mecanismos geradores de algum sistema e então realizar ações corretivas apropriadas”. O autor faz menção a diversas aplicações de estratégias sofisticadas de controle envolvendo previsão, como um modelo estocástico ajustado à série, em que os valores futuros da série são previstos e, em seguida, as variáveis do processo de entrada são ajustadas para manter o objetivo do processo.

2.1.1 Tendência, sazonalidade e estacionariedade

Para o estudo do comportamento de séries temporais é comum propor uma decomposição. Como ilustrado na Figura 1 , existem três componentes principais que forma uma série temporal, tendência, sazonalidade e ruído (ou resíduo).

Figura 1 - Componentes de uma série temporal.



A tendência pode ser entendida como uma mudança de longo termo no nível da média da série temporal (CHATFIELD, 1996). Ela pode assumir as formas linear, quadrática ou exponencial, ou outro comportamento, em resumo são mudanças de direção no sentido de crescimento (ou decrescimento).

Sazonalidade está relacionada com flutuações temporais periódicas, com uma frequência fixa e conhecida. Como citado por Lazzeri (2020), geralmente está condicionada a periodicidade da coleta de dados, por exemplo em frequências horárias, diárias, semanais e mensais. A sazonalidade pode ainda ter uma subdivisão, chamada de ciclos, que representam comportamento de subidas e descidas na série com um certo padrão, mas que não ocorrem em períodos fixos, são em geral mais difíceis de serem observados nas séries.

O ruído ou resíduo da série temporal é a componente resultante da remoção de tendência e sazonalidade da série, que pode ou não ser aleatória (CHATFIELD, 1996). Para análise de séries temporais são realizadas diversas suposições sobre os resíduos, uma modelagem de séries temporais é dita adequada quando consegue extrair o comportamento determinístico da série, tendência e sazonalidade, e alcança valores de resíduos cuja média seja zero, a variância constante e que não exista autocorrelação (BROCKWELL & DAVIS, 2010).

Quando a variância dos resíduos é constante, denomina-se homocedasticidade. Wooldridge (2015) define que a homocedasticidade se dá quando o erro apresenta a mesma

variância dado qualquer valor de variável explanatória, em outras palavras, Var(u) =σ^2. Porém, A homocedasticidade falha sempre que a variância dos fatores não observados varia em diferentes segmentos da população, onde os segmentos são determinados pelos diferentes valores das variáveis explicativas (WOOLDRIGDE, 2015). Neste caso, atribui-se o conceito heterocedasticidade, ou seja, a variância não é constante no tempo em função de um determinado fator. Por exemplo, no tema de consumo de água residencial, em uma equação de demanda de abastecimento, a heterocedasticidade está presente se a variância dos fatores não observados que afetam o consumo aumenta com o número de moradores.

O modelo aditivo de decomposição de uma série temporal pode ser expresso conforme a Equação 1.

$$Z_t = T_t + S_t + a_t \qquad \text{Eq. 1}$$

Sendo Z_t uma série temporal, ela é formada pelas componentes T_t de tendência, S_t que representa a sazonalidade, além da componente a_t para o resíduo. Este modelo é em geral não estacionário segundo Morettin e Tolloi (2006), e então técnicas apropriadas devem ser empregadas para remover a tendência e a sazonalidade da série para se obter uma série estacionária.

Um processo estacionário é caracterizado por suas propriedades estatísticas não dependerem do tempo. Assim, séries temporais com tendência, ou com sazonalidade, não são estacionários, com tendência e sazonalidade afetando o valor da série em momentos diferentes (DE MATOS, 2018). Deve-se atentar para a questão temporal e amostral, que podem influenciar no processo de determinação da estacionariedade de uma série, como por exemplo comportamentos cíclicos que não caracterizam a não-estacionariedade.

Segundo Chatfield (1996), uma série pode ser dita estacionária se não há mudança sistemática na média (sem tendência), sem mudança sistemática na variância, e se as variações de período foram estritamente removidas (sem sazonalidade). Ou seja, estas componentes estatísticas não possuem dependência com o tempo, sendo essa característica definida como estacionariedade fraca. Na prática, em análise de séries temporais costuma-se trabalhar com o formato de estacionariedade fraca porque ela é mais operacional e atende satisfatoriamente às caracterizações de processos estocásticos (DE MATOS, 2018).

A verificação da estacionariedade no início do estudo é uma etapa necessária para aplicações de modelos estatísticos do tipo autorregressivos. Quando uma série não é estacionária é requerido que ela seja transformada através de técnicas específicas (CHATFIELD, 1996). Entre as técnicas mais aplicadas está a diferenciação. A diferenciação auxilia a estabilizar a média de uma série temporal ao remover as mudanças de nível, e, portanto, eliminando (ou reduzindo) a tendência ou sazonalidade (HYNDMAN e ATHANASOPOULOS, 2021).

Ao se tratar de remoção da tendência, para séries não sazonais, a diferenciação de primeira ordem é usualmente suficiente para tornar a série estacionária. Uma série diferenciada é então definida como o resultado de diferenças entre observações consecutivas, demonstrado pela Equação 2.

$$y_t = x_t - x_{t-1} \qquad \text{Eq. 2}$$

A série resultante possui tamanho $t - 1$, uma vez que não é possível calcular a diferença $y_1$para a primeira observação (HYNDMAN e ATHANASOPOULOS, 2021).

Ao se tratar da remoção da sazonalidade introduz-se o conceito de diferenciação sazonal, sendo ela a diferença entre uma observação e a observação passada de mesma sazonalidade (HYNDMAN e ATHANASOPOULOS, 2021), então tem-se a Equação 3.

$$y_t = x_t - x_{t-m} \qquad \text{Eq. 3}$$

Em que m é a frequência ou período dessa sazonalidade. Também pode-se chamar de "diferença de m *lags*", uma vez que representa a subtração após uma defasagem de m períodos.

2.1.2 Sazonalidade complexa

A sazonalidade complexa é um tema discutido por Hyndman e Athanasopoulos (2021) e ainda bastante amplo e aberto para novas abordagens. Os autores classificam o tema na categoria de métodos avançados de previsão, e o definem como uma característica das séries temporais com frequência de coleta semanal, diária e sub diária (horária, por exemplo), que apresentam por sua vez mais de uma sazonalidade ao mesmo tempo ou então uma sazonalidade de tamanho não inteiro.

Em alguns trabalhos desenvolvidos em conjunto com o próprio autor Prof. Hyndman foram propostas estratégias diferentes para tratar dos padrões de sazonalidades complexas

apresentados por algumas séries. Como o caso do trabalho de Gould et al. (2008) que propõe uma modificação ao modelo de Holt-Winters para lidar com mais de uma sazonalidade ao mesmo tempo, aplicado para séries de carga de energia elétrica e fluxo de veículos, ambas as séries com frequência horária.

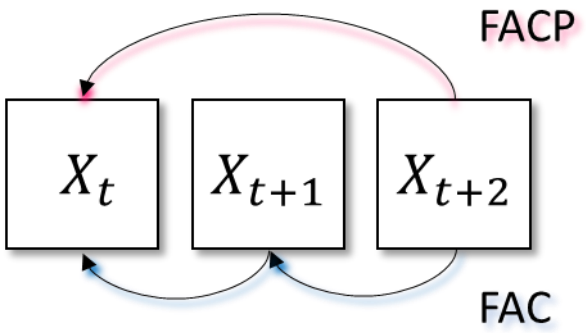
Há também a pesquisa realizada por Fan e Hyndman (2011), que avaliou a possibilidade de inclusão de variáveis como efeito de calendário, observações atuais defasadas e histórico e previsão de temperatura para a proposição de um modelo de previsão de curto prazo para séries temporais de demanda elétrica. Outro trabalho que contribui com a discussão através da proposição de novos modelos, é apresentado por De Livera et al.(2011), que formula os modelos BATS (Box-Cox transformation, ARMA erros, Trend and Seasonal components) e TBATS (Trigonometric, Box-Cox transformation, ARMA erros, Trend and Seasonal components).

2.1.3 Autocorrelação

A autocorrelação é uma medida da semelhança de uma série temporal com versões dela mesma deslocadas no tempo. Este deslocamento é uma defasagem temporal e aparecerá com terminologia em inglês (*lag*) ao longo desta pesquisa. “É utilizada para detecção de comportamento não aleatório, em particular de padrões de repetição, tais como a presença de componentes periódicos na série” (CHATFIELD, 1996). Como forma de promover o estudo da autocorrelação da série, serão utilizados gráficos chamados de autocorrelogramas, que demonstram os valores para a função de autocorrelação (FAC, ou *ACF* em inglês) e autocorrelação parcial (FACP, ou *PACF* em inglês) para diferentes *lags* de interesse. Segundo Hyndman e Khandakar (2021), inclusive, estes gráficos podem ser utilizados para determinar a ordem de modelos tanto autorregressivos, AR(p), quanto de médias móveis MA(q), pois promovem entendimento do mecanismo gerador dos dados.

Pode-se entender melhor a diferença entre as FAC e FACP através do esquema abaixo, apresentado na Figura 2, que mostra a FAC com uma consideração dos valores intermediários entre o ponto inicial e o *lag* t. A FACP considera a correlação direta de um *lag* à série original sem levar em consideração a dependência das observações intermediárias.

Figura 2 - Ilustração da diferença entre as abordagens realizadas pela função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP).



2.2 Aplicações de análise de séries temporais a dados de monitoramento de redes de distribuição de água

Neste item serão exploradas as relações entre a variação de consumo de água e as componentes das séries temporais, como forma de caracterizar o sistema gerador e se entender a origem das características dos dados de pressão na rede de distribuição. Com isso, serão apresentados trabalhos anteriores que consideram sinais de vazão e/ou pressão como séries temporais e suas diversas aplicações e limitações encontradas no processo de modelagem.

2.2.1 Perfil de variação de consumo de água

Como parte da caracterização do comportamento da série temporal de pressão em uma rede de abastecimento é importante levar em consideração características de seu mecanismo gerador. De acordo com Tsutiya (2006), para economias de água residenciais, “a quantidade de água consumida varia continuamente em função do tempo, das condições climáticas, hábitos da população, etc”. Para o abastecimento de água de uma forma geral ocorrem variações em escala anual, mensal, diária, horária e instantânea do consumo de água.

Quanto a variação anual, considerando um aumento populacional ou então aumento do consumo per capita devido a melhoria dos hábitos de higiene da população, esta variação tende a crescer, e este seria um dos principais fatores relacionados à componente de tendência das séries temporais tanto de pressão quanto de vazão de demanda. A variação mensal está relacionada com as estações do ano, é sabido que no verão o consumo supera o consumo médio, e no inverno é menor (TSUTIYA, 2006). A variação diária está intrinsecamente ligada com os hábitos da população ao longo da semana, como por exemplo trabalhar e estudar nos dias úteis

e estar em casa aos finais de semana, fazendo com que o consumo seja maior aos sábados e domingos do que em relação aos outros dias da semana. Sabe-se também que o consumo varia com as horas do dia, sendo que geralmente o maior consumo ocorre entre 10 e 12 horas, sendo estes fatores que colaboram com a caracterização sazonal da série temporal.

Ainda quanto a variação horária, Tsutiya (2006) reporta um estudo feito na região metropolitana de São Paulo (RMSP) que monitorou 22 setores durante 3,5 anos e possibilitou o levantamento do comportamento típico. Em geral "a vazão mínima ocorre às 3 horas, vazão máxima entre 10 e 12 horas e a vazão média é coincidente com o consumo do período entre as 7 e 8 horas e entre 18 e 22 horas" (TSUTIYA, 2006). Ainda afetam os valores observados de pressão os consumos instantâneos, que podem ser entendidos como ruídos nas séries temporais.

Sendo o consumo de água o fator que influencia na dinâmica operacional da rede de abastecimento, é esperado encontrar os efeitos das variações de consumo também nos sinais de pressão. Em geral, inversamente proporcionais às variações de consumo na rede, ou seja, quando a vazão de consumo aumenta em determinada hora do dia a pressão tende a decrescer, e quando a vazão é mínima, a pressão na rede é máxima.

2.2.2 Previsão e classificação na detecção de eventos anormais em redes de distribuição de água

Como visto, diversos são os fatores que influenciam no padrão de consumo de água dentro de um determinado setor de controle, soma-se aos hábitos de consumo da população a ocorrência de eventos como mudanças no clima, festivais ainda são considerados normais. Por outro lado, eventos como uso de hidrantes e ocorrência de vazamentos impactam a série de dados de forma diferente (YIPENG *et al*., 2017). Assim, há uma abundância de dados normais, provenientes do monitoramento de redes, enquanto há poucos dados *outliers* (MOUNCE et al., 2011).

Diante disso, diversas são as iniciativas de detectar esses eventos anormais, Yipeng *et al*. (2017) em sua revisão de métodos de aplicação, classificaram abordagens baseadas em dados em três categorias: métodos de classificação, métodos de previsão-classificação e métodos estatísticos. Sendo essa a abordagem que se deseja aplicar nessa pesquisa, detalha-se que, segundo os autores, os métodos de previsão-classificação podem ser entendidos como abordagens em que, quando ocorre um vazamento em um setor de abastecimento de água, os

valores medidos pelos sensores irão divergir significativamente dos valores previstos pelos modelos, porque a previsão é baseada em dados sob condições normais.

Notadamente, cada método possui suas limitações. Entre as limitações encontradas para o método de previsão-classificação pode-se citar a propagação de incertezas, pela necessidade de uma prévia classificação do que é o comportamento normal da rede, se não ocorrer da maneira correta, pode-se treinar o modelo de previsão com dados anômalos, prejudicando a detecção de vazamentos posteriormente (YIPENG *et al*., 2017). Outra limitação encontrada está em resultados errôneos de detecção de eventos em função de saídas dos modelos previsão determinísticas, pois diversos estudos não apresentam informações sobre os resíduos dos modelos. Yipeng *et al*. (2017) citam ainda que, na presença de incerteza, previsões de valores únicos guiam a decisões errôneas.

O estudo apresentado por Mason (2019) teve como objetivo desenvolver algoritmos para extrair conhecimento a partir dos dados monitorados por uma macromedidor na entrada de um setor de abastecimento. Assim, poderia propor aplicações em auxiliar a companhia de saneamento a lidar com variação do consumo de água, especialmente quanto a mudanças no clima e sazonalidade. Adicionalmente propôs a comparação da previsão com dados reais para ajudar a detectar falhas na rede. Para isso, foram utilizados 2 anos de dados para treinos dos modelos, com o objetivo de se prever um horizonte de 7 dias, sua série era composta por observações horárias de pressão. Pelo tamanho de seu conjunto de treino, foi necessário aplicar transformações aos dados para remoção da tendência. Mason (2019) aplicou três modelos de previsão, SARIMA, GAM (modelo aditivo generalizado) e Redes Neurais Artificiais (RNA).

Primeiramente Mason (2019) ajustou um modelo SARIMA (1,0,1)(2,1,0)[24], e obteve um AIC (Critério de Informação de Akaike) de 8388, e considerou como boas as previsões do modelo. Porém, na análise de resíduos realizada, os gráficos dos resíduos normalizados não demonstravam que os resíduos seguiam distribuição normal, a autora também identificou que havia um incremento da variância dos resíduos e concluiu que não poderia criar intervalos de confiança, e atribuiu limitações ao modelo SARIMA como a impossibilidade de utilização de variáveis exógenas. Posteriormente, na aplicação do modelo GAM, a autora adicionou variáveis preditoras como hora do dia, mês do ano e uma informação de tendência, além dos termos autorregressivos. Não foi indicado no trabalho a distribuição utilizada no modelo GAM. Novamente, os resultados da previsão foram bons, mas na análise de resíduos foi reportado que os resíduos não seguiam distribuição normal, assim, ao não cumprir os pressupostos, não foi

possível estabelecer intervalos de confiança para a previsão. Ficou assim evidente que a pesquisa apresentada encontrou uma das limitações descritas por Yipeng *et al.* (2017), onde por fim pode prever apenas valores únicos, como numa saída determinística do modelo.

A pesquisa apresentada por Bakker *et al.* (2014) combina uso de heurística e controle estatístico de processos para detecção de vazamentos. Foram utilizados dados do monitoramento de vazão de entrada em um setor de distribuição. A frequência de coleta dos sensores era de 5 minutos, e o autor aplicou transformação para uma escala de 15 minutos. O modelo proposto era baseado em regressão múltipla, com erros modelados por processo ARMA. O conjunto de ajuste do modelo considerava apenas dois dias de dados observados para prever 2 dias a frente, e contava com um fator de ajuste para cada dia da semana. Os autores não apresentaram análise de resíduos, nem intervalo de confiança da previsão. Assim, para aplicação em detecção de vazamentos, eles criaram o próprio critério limite da variação da vazão a ser considerada normal quando comparada com os valores previstos. Os mesmos classificaram as vazões em faixas de consumo, pois perceberam que a variância dos dados mudava com a média da vazão e assim, para cada faixa, analisaram um ano de dados monitorados (classificados como em comportamento normal da rede) e os valores de médias e desvios padrões de cada faixa, adicionaram uma margem de 5% para compor limite de confiança à previsão.

Os autores Adachi *et al.* (2017) apresentaram pesquisa que tinha como objetivo detectar vazamentos a partir do desvio entre previsão e dados observados de vazão, com modelo combinado para duas situações, um para menores intervalos e dados mais recentes para detecção de rompimentos de rede e mudanças mais bruscas no comportamento dos dados, e outro modelo ajustado apenas com valores mais antigos, classificados como normais, com objetivo de detectar vazamentos que se desenvolvem gradativamente. A estratégia para definir um limite para previsão foi a utilização do desvio padrão histórico dos dados observados, semelhante ao realizado por Bekker *et al.* (2014), porém sem margem de segurança extra e definição de faixas diferentes de consumo.

Para lidar com as limitações encontradas nesta abordagem, estratégias para assegurar estacionariedade histórica da série devem ser empregadas, tanto para tendência quanto para sazonalidade. Quanto à confiança na previsão, devem ser desenvolvidos modelos capazes expressar graus de confiança nos valores previstos, através de métodos probabilísticos, empregando-se melhores funções de densidade de probabilidade (YIPENG *et al*., 2017).

2.3 Modelos de previsão para séries temporais

Considerando as características de dependência entre observações adjacentes, se estabelece no campo da análise de séries temporais o estudo de ferramentas para estudar esta dependência (BOX et al., 2008) e a esse estudo podemos incluir o uso de modelos matemáticos. A ideia de utilizar modelos para descrever o comportamento de fenômenos físicos é bem estabelecida. Considerando que existe muitos modelos diferentes que poderiam ser utilizados para descrever o comportamento de uma série particular, deve-se atentar, durante a construção desses modelos, para vários fatores, tais como o comportamento do fenômeno ou o conhecimento a priori que temos de sua natureza e do objetivo da análise, bem como para limitações computacionais (MORETTIN e TOLOI, 2006). Nesta seção serão apresentadas descrições de modelos implementados nesta pesquisa com o objetivo de elucidar sua formulação e contribuir para o entendimento de suas características práticas.

2.3.1 Modelos autorregressivos

O processo que estuda a característica dependência temporal entre as observações de uma série denomina-se autorregressão. A autorregressão assume que as futuras observações da série temporal estão relacionadas com as observações do passado através de uma relação linear (LAZZARI, 2020). O termo autorregressão indica que esse processo é uma regressão da variável observada contra si mesma (HYNDMAN & ATHANASOPOULOS, 2021).

Assim, um modelo autorregressivo de ordem p pode ser escrito como apresentado pela Equação 4:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \qquad \text{Eq.4}$$

sendo y_t o valor atual que depende de um parâmetro ϕ_p e das observações passadas y_{t-p}. ε_t é ruído branco. Fazemos referência a esse modelo como AR(p), modelo autorregressivo de ordem p.

Considerando as seguintes notações

$$By_t = y_{t-1}\,,\; B^m y_t = y_{t-m}$$

Sendo B um operador de translação para o passado e lembrando que m refere-se ao tamanho da sazonalidade de uma série. Pode-se escrever o modelo autorregressivo na forma resumida (Equação 5):

$$\phi(B)y_t = \varepsilon_t \quad \text{Eq.5}$$

2.3.2 Modelos combinados autorregressivos e de médias móveis

Como forma de melhorar os modelos autorregressivos, pode-se acrescentar a avaliação de médias móveis. A notação que antes era AR(p) passa então a ser ARMA (p,q), sendo MA(q) o modelo de médias móveis de ordem q. De acordo com Hyndman & Athanasopoulos (2021), ao contrário de usar observações passadas para previsão, um modelo de médias móveis considera os erros das previsões anteriores em uma abordagem de regressão. Dessa forma, pode-se agregar médias móveis aos modelos autorregressivos e chegar ao que chamamos de modelo autorregressivo e de médias móveis, ARMA (p,q), definido como apresentado pela Equação 6:

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots \theta_q \varepsilon_{t-q} \quad \text{Eq.6}$$

sendo y_t o valor atual que depende de um parâmetro autorregressivo ϕ_p e das observações passadas y_{t-p}, além de depender também de um parâmetro de médias móveis θ_q e dos erros passados $\theta_{q_{t-q}}$. E ε_t é ruído branco.

Na forma resumida, apresenta-se a Eq. 7,

$$\phi(B)y_t = \theta(B)\varepsilon_t \quad \text{Eq.7}$$

Existe ainda os modelos integrados autorregressivos de médias móveis, denominados ARIMA (p,d,q), onde o termo “I” de integrado representa uma transformação de diferenciação (com ordem representada pelo termo “d”) que torna a série temporal estacionária quanto à tendência. Mais detalhes sobre essa transformação foram tratados no item 2.1.1 Tendência, sazonalidade e estacionariedade.

2.3.3 Modelo SARIMA

Os modelos ARIMA descritos anteriormente ainda possuem a capacidade de lidar com componentes sazonais presentes nos dados. Como descrito por Hyndman & Athanasopoulos

(2021), uma ARIMA sazonal ou SARIMA é formada a partir da inclusão de termos sazonais adicionais em um modelo ARIMA e pode ser escrito como segue na Figura 3,

Figura 3 - Esquema de uma modelo ARIMA Sazonal ou SARIMA.

$$ARIMA \underbrace{(p,d,q)}_{\substack{\text{Parte Não} \\ \text{Sazonal do} \\ \text{Modelo}}} \underbrace{(P,D,Q)_m}_{\substack{\text{Parte} \\ \text{Sazonal do} \\ \text{Modelo}}}$$

Fonte: adaptado de Hyndman & Athanasopoulos (2021)

Utilizando então letras minúsculas para denotar os termos da parte não sazonal e letra maiúscula para a parte sazonal do modelo. Os termos sazonais então refletem as mesmas considerações sobre a parte não sazonal, onde P representa a parte autorregressiva, D a ordem da diferenciação e Q a componente de médias móveis, mas envolvendo o período sazonal m. Assim, pode ser escrito conforme notação da Equação 8,

$$\phi(B)\Phi(B^m)(1-B)^d(1-B^m)^D y_t = \theta(B)\Theta(B^m)\varepsilon_t \qquad \text{Eq. 8}$$

Onde Φ representa os parâmetros autorregressivos da parte sazonal do modelo, assim como Θ representam os parâmetros de médias móveis da parte sazonal do modelo. O uso de modelos SARIMA pode trazer vantagens por lidar com a componente sazonal da série, fazendo com que o comportamento dela seja melhor representado.

2.3.4 Identificação de modelos do tipo ARIMA

Morettin e Toloi (2006) esclarecem que a estratégia para a construção de modelos do tipo ARIMA, propostos por Box et al. (2008), está baseada em um ciclo iterativo, com seguintes estágios: especificação, onde se define o tipo de modelo que se irá trabalhar; identificação, consiste na definição da ordem *(p,d,q)* e *(P,D,Q);* estimação, na qual os parâmetros do modelo são estimados; diagnóstico, através de uma análise de resíduos, para se saber se está adequado para os fins em vista.

Neste sentido, a definição dos tipos de modelos a serem trabalhados estarão descritas no item 3.4 Ajuste de modelos de previsão. Para identificação da ordem do modelo e para estimação dos parâmetros, será utilizado o método automático através do algoritmo

“autoarima” proposto por Hyndman & Khandakar (2008), que consiste em ajustar modelos e calcular uma medida de informação de ajuste, neste caso AIC (Critério de Informação de Akaike), e em seguida verificar o valor de AIC das combinações de parâmetros dos vizinhos e “caminhar” na direção do melhor modelo (menor AIC), até que não encontre nos vizinhos melhor combinação de parâmetros segundo seu AIC. A medida AIC por sua vez penaliza modelos complexos demais e favorece os mais simples.

2.3.5 Modelos Harmônicos

Modelos harmônicos estão ligados com análise de Fourier ou análise harmônica, e têm sido aplicados na análise de séries temporais. Resultantes da observação de processos estocásticos, o objetivo básico é o de aproximar uma função do tempo por uma combinação linear de harmônicos (componentes senoidais) (MORETTIN e TOLOI, 2006). Os autores Hyndman & Athanasopoulos (2021) recomendam também o uso de conjuntos harmônicos para representar o comportamento sazonal periódico em séries com essa característica.

Esta técnica é aplicada em casos de sazonalidade complexa da série, onde temos múltiplas sazonalidades ocorrendo ao mesmo tempo, geralmente intrínseca a frequência de coleta de dados. Com o avanço da tecnologia e dos sensores, a tendência é chegar cada vez mais próximo do monitoramento em tempo real, mas os modelos estatísticos clássicos foram desenhados para lidar com escalas menores como uma frequência igual a 12 para dados mensais ou frequência igual a 4 para dados trimestrais.

Basicamente, Fourier demonstrou que uma série de termos senos e cossenos nas frequências adequadas podem aproximar qualquer função periódica (HYNDMAN & ATHANASOPOULOS, 2021). Para um determinado período sazonal m, os primeiros termos de Fourier seriam dados por:

$$x_{1,t} = \sin\left(\frac{2\pi t}{m}\right), x_{2,t} = \cos\left(\frac{2\pi t}{m}\right) \ldots$$

Algumas vantagens do uso dessa técnica são citadas por Hyndman & Athanasopoulos (2021), como: Permite lidar com qualquer tamanho de sazonalidade; para séries com mais de um período sazonal pode-se acrescentar termos de Fourier de diferentes frequências; a suavização do comportamento sazonal pode ser controlada por uma variável K que representa o número de pares senos e cossenos.

Observou-se em trabalhos anteriores (ZAHED, 1990; ODAN & REIS, 2012; BRENTAN et al., 2017) iniciativas de modelagem de séries de vazão de demanda com uso de séries de Fourier, que por se tratar de variável proveniente do monitoramento do mesmo mecanismo gerador das séries de pressão, pode se estabelecer relações nas estratégias de modelagem. O autor Zahed (1990) apresenta análise de sensibilidade da variação do número de pares harmônicos e conclui que múltiplos de 7 (relacionado aos ciclos diários de consumo) apresentam melhores ajustes, sendo que o aumento de 7 pares para 14 aumenta o coeficiente de correlação entre dados observados e modelados em 19,8%, e o aumento de 14 para 21 pares aumenta o mesmo coeficiente em apenas 3,5%, o autor recomenda então que o número de pares não seja maior que 21. Brentan *et al.* (2017) por sua vez, encontra que o erro diminui consideravelmente a partir da inclusão de pelo menos 3 pares harmônicos, nota-se também baixos valores nos erros a partir de 7 pares harmônicos, e por fim o autor recomenda uso do máximo número de pares limitado à metade do conjunto da série.

2.3.6 Modelo Generalizado Autorregressivo de Médias Móveis - GARMA

Uma vez que para dados com distribuição normal modelos de regressão linear, com erros modelados por ARMA, são suficientes para um bom ajuste, o mesmo não ocorre com dados com distribuição não-gaussiana. Assim, Benjamin *et al.* (2003) propuseram uma nova abordagem para o problema da extensão dos modelos gaussianos autorregressivos e de médias móveis para uma estrutura não gaussiana.

Para o modelo desenvolvido a distribuição da variável dependente no tempo dado o conjunto de informações anterior é modelada por uma função de distribuição de probabilidade da família exponencial que inclui as distribuições Gaussiana, Poisson, Gama e Binomial (ALBARRACIN et al, 2019). Chamado de GARMA (*generalised autorregressive and moving average model*), ou modelo generalizado autorregressivo e de médias móveis, este modelo proposto por Benjamin et al. (2003) combina a capacidade de lidar com diversas covariáveis dos modelos lineares generalizados e a modelagem dos erros por processo autorregressivos, com a possibilidade de usar uma distribuição da família exponencial para determinar a probabilidade da observação de um valor futuro dado observações passadas. Assim o modelo é composto pelas equações 9 e 10 na seguinte forma:

$$f(y_t|H_t) = \exp\left\{\frac{y_t v_t - b(v_t)}{\varphi} + d(y_t, \varphi)\right\} \qquad \text{Eq. 9}$$

A equação 8 define a função de densidade condicional onde é representada a probabilidade de observar y_t, dado H_t. Essa função depende de parâmetros chamados v_t e φ bem como de funções específicas $b(.)$ e $d(.)$ que variam de acordo com o tipo de distribuição a ser usada.

No modelo GARMA o cálculo da média condicional μ_t é relacionado com o preditor linear η_t por meio de uma função de ligação chamada $g(.)$. Além de levar em consideração os termos autorregressivos e de médias móveis, logo temos a equação 9:

$$g(\mu_t) = \eta_t = \underline{\boldsymbol{x_t'}}\underline{\boldsymbol{\beta}} + \sum_{j=1}^{p} \phi_j \left\{ g(y_{t-j}) - \underline{\boldsymbol{x_{t-j}'}}\underline{\boldsymbol{\beta}} \right\} + \sum_{j=1}^{q} \theta_j \{ g(y_{t-j}) - \eta_{t-j} \} \qquad \text{Eq. 10}$$

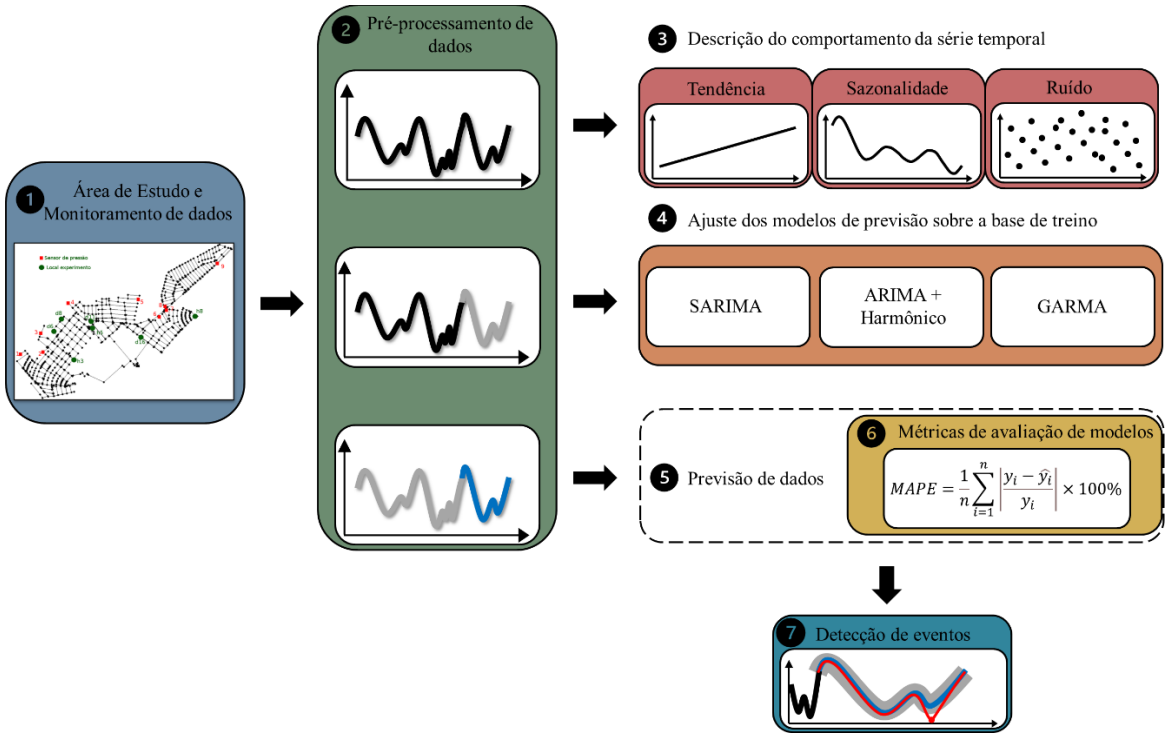
Onde, $\underline{\boldsymbol{x_t}}$ é um vetor de r variáveis explicativas e $\underline{\boldsymbol{\beta}}' = (\beta_1, \beta_2, \ldots, \beta_r)$. Os parâmetros autorregressivos são $\underline{\boldsymbol{\phi}}' = (\phi_1, \ldots, \phi_p)$. Os parâmetros de médias móveis são $\underline{\boldsymbol{\theta}}' = (\theta_1, \ldots, \theta_q)$. Os termos de resíduos considerados pela função de médias móveis podem ser de diferentes tipos como desvios resíduos de *Pearson*, resíduos na escala original dos dados ou na escala do preditor como na equação 9 (exemplo $g(y_{t-j}) - \eta_{t-j}$)

Um modelo GARMA(p,q) é definido pelas equações 9 e 10. Em resumo, os parâmetros do modelo são $\underline{\boldsymbol{\beta}}'$, $\underline{\boldsymbol{\phi}}'$ e $\underline{\boldsymbol{\theta}}'$**,** estimados a partir do método da máxima verossimilhança condicional em um processo iterativo de mínimos quadrados ponderados.

3 MATERIAIS E MÉTODOS

A metodologia proposta está representada através do esquema da Figura 4, onde o início se dá através do monitoramento de dados realizado por Gamboa-Medina (2017), descrito em 3.1, que resultou na construção de séries temporais de pressão, em seguida no item 3.2 são descritas as etapas de pré-processamento de dados e modelagem, onde promove-se alteração da frequência de dados da série e classificação de cada período de acordo com ocorrências na rede de distribuição. A seção 3.3 retrata os pontos principais levados em consideração para a descrição do comportamento da série temporal e entendimento do mecanismo gerador. Na sequência são ajustados e treinados os modelos de previsão, item 3.4. A estratégia metodológica da etapa de previsão de dados é apresentada na seção 3.5. É realizada a avaliação dos resultados de acordo com métricas de acurácia de acordo com o descrito no item 3.6, e por último, a exploração da viabilidade das técnicas de previsão para detecção de vazamentos no item 3.7.

Figura 4 - Estrutura do esquema metodológico da pesquisa



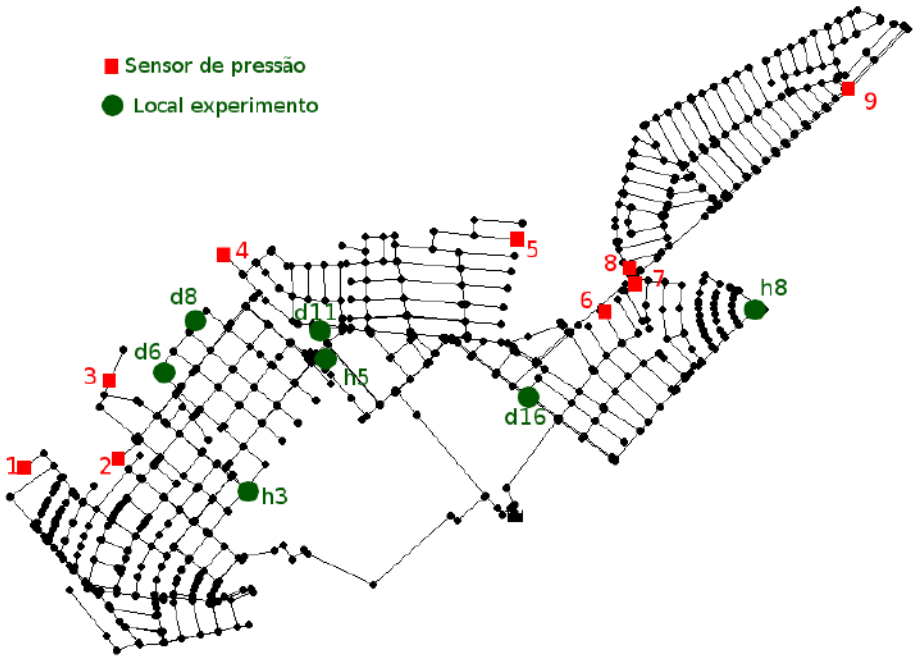
3.1 Área de estudo e monitoramento de dados

A base de dados deste estudo provém dos ensaios de campo realizados por Gamboa-Medina (2017), a qual monitorou por 357 dias as cargas de pressão da rede de abastecimento de água de um setor no município de Araraquara, SP – Brasil, com resolução temporal de 1 dado a cada 2 minutos,. Como características do setor tem-se uma área de 2,75 km², com aproximadamente 6 mil ligações, sendo 90% de categoria residencial. A extensão da rede é de aproximadamente 57 km, formada principalmente por tubulações de PVC, com diâmetros de 50 mm (72%) até 250 mm (7%). Alguns trechos de rede mais antigos possuem cerca de 20 anos e diversos são os chamados para reparo de vazamentos no local feitos para a companhia de saneamento local (DAAE – Departamento Autônomo de Água e Esgoto da cidade de Araraquara).

Ao todo, 9 sensores foram utilizados por Gamboa-Medina (2017), os quais sua localização foi determinada por processo de otimização de modo a representar a melhor cobertura do monitoramento. Durante o período de monitoramento, foram realizadas diversas simulações de vazamentos na rede, com o objetivo de analisar posteriormente o impacto destas simulações na série de dados.

A Figura 5 apresenta o esquema do modelo hidráulico da rede de abastecimento estudada com a localização dos sensores e dos pontos em que foram realizadas simulações de eventos. Ao todo foram realizadas 12 simulações de vazamentos mediante abertura controlada de válvulas de descarga e hidrantes na rede, identificados na Figura 5 através do círculo verde seguido da nomenclatura específica, sendo “d” para válvulas de descarga e “h” para hidrantes. Mais detalhes sobre os eventos estarão na sessão 3.7 .

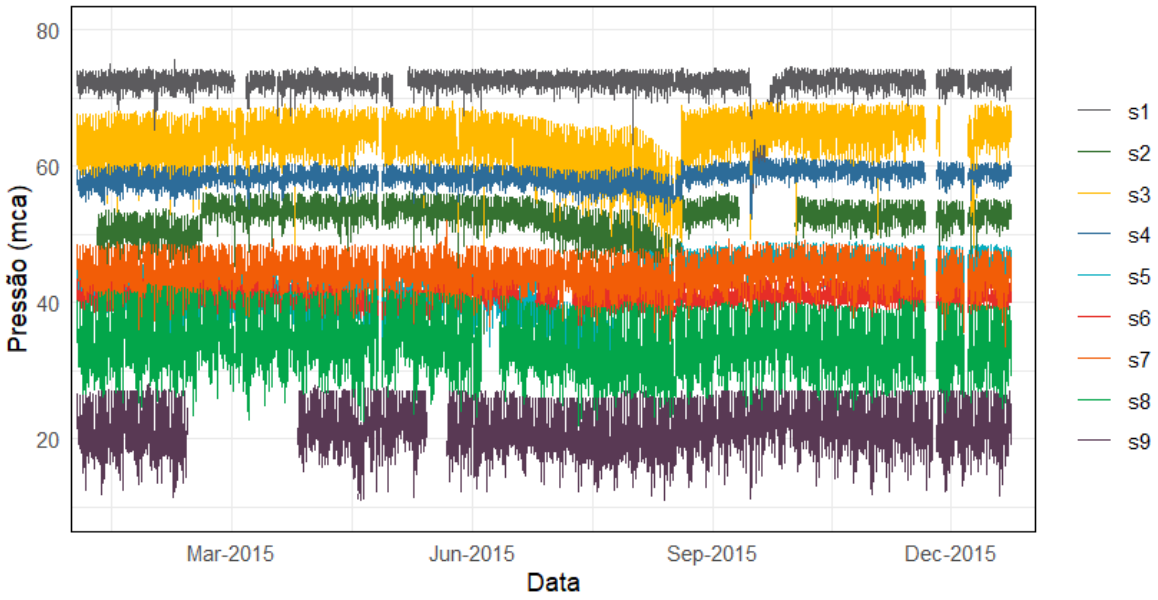
Figura 5 - Esquema do modelo hidráulico da rede de abastecimento estudada com a localização dos sensores e dos pontos de simulação de eventos.



Fonte: Gamboa-Medina (2017)

Como resultado deste monitoramento foi construída uma base de dados com os valores de carga de pressão em metros para cada um dos 9 sensores, sendo a resolução de coleta de 2 minutos, com 357 dias de campanha chegou-se a um total de 257040 linhas de observações. A indexação temporal se deu através da contagem de dias a partir do início do monitoramento. Através da Figura 6 pode-se ter uma visão geral sobre os dados.

Figura 6 - Séries temporais de pressão (mca) para os 9 sensores monitorados na área de estudo.



3.2 Pré-processamento e modelagem de dados

A resolução de 2 minutos dos sensores de pressão instalados na rede de abastecimento real monitorada no estudo de Gamboa-Medina (2017) favorece a velocidade de detecção de vazamentos. Sabe-se, porém, que a maior resolução torna a base de dados mais pesada do ponto de vista computacional. Diante disso, foi proposto a mudança para a resolução de 10 minutos, considerado um intervalo satisfatório para não causar perda de informação relevante para o estudo aqui proposto, tornar a base mais leve para o processamento computacional e ainda por se adequar a resolução de outros sensores de mercado que trabalham nesta escala. Como a série original de 2 minutos é formada pela média de medidas pontuais a cada 10 segundos, optou-se por calcular a partir dela a média a cada 5 pontos, obtendo assim o novo intervalo de 10 minutos.

Gamboa-Medina (2017) realizou em seu trabalho uma classificação da base de dados, identificando para cada um dos pontos medidos o status de 0 – normal, 1 – simulado, 2 – reporte, 3 – visual e 4 – extremo. Sendo,

> Classe 1 ou “simulado”: dados adquiridos durante os experimentos de simulação de vazamentos e rupturas, identificados a partir dos horários de realização dos experimentos; Classe 2 ou “reporte”: dados adquiridos durante o período entre a notificação de vazamento e seu reparo, segundo registro da companhia de saneamento da cidade; Classe 3 ou “visual”: dados que apresentam descontinuidade visualmente identificáveis, reconhecidos por inspeção visual dos gráficos de cada uma das séries; Classe 4 ou “extremo”: dados maiores ou menores a 99,8% dos valores, e que não foram identificados

> nas classe anteriores, e; Classe 0 ou “normal”: dados considerados normais, por não se enquadrarem em nenhuma das classes anteriores. (GAMBOA-MEDINA, 2017)

Esta classificação proposta auxiliou na separação de trechos de interesse na base, que levou em consideração alguns critérios, como: não existência de dados faltantes e todos os dados classificados como 0 – “normal”. Fez-se avaliação destes critérios para um conjunto que considerasse os 9 sensores simultaneamente, com o objetivo de proporcionar validação dos modelos posteriormente.

3.3 Descrição do comportamento da Série

A etapa de descrição da série temporal, que ocorre antes do ajuste de modelos, como apresentado anteriormente, busca através de técnicas conjuntas de análise gráfica e aplicação de testes estatísticos, obter informações sobre as componentes de autocorrelação dos dados, identificação de tendências e sazonalidades, assim como outliers e pontos de inflexão que promovem mudanças no comportamento da série.

a) Tendência, sazonalidade e estacionariedade

No caso das séries de sinais de pressão, foi reportado por Gamboa-Medina (2017) que se deve atentar para escalas de curto e longo prazo, justamente pois no curto prazo pode-se encontrar trechos não-estacionários, e quando se analisa a partir da ordem de dias, a série tende a apresentar comportamento estacionário quanto à tendência. Acrescenta-se ainda que nesta escala de dias, é possível que ciclos anuais não caracterizem a não-estacionariedade para a série de pressão em redes de abastecimento.

Além da análise gráfica promover uma boa ideia sobre a existência de tendência na série, inclusive sendo recomendada por Box *et al.* (2008), nesta pesquisa optou-se por aplicar um teste estatístico para auxiliar nesta verificação, o teste ADF (Dickey-Fuller Aumentado). Os testes aplicados para verificação da condição de estacionariedade, quanto à tendência da série, são chamados de testes de raízes unitárias, que são testes de hipóteses estatísticos de estacionariedade desenhados para determinar se é necessário aplicar métodos de diferenciação na série temporal ou não (HYNDMAN e ATHANASOPOULOS, 2021). Em caso da hipótese de estacionariedade ser rejeitada, aplica-se então a diferenciação e repete-se o teste.

De forma similar ao tratado sobre a tendência, para sazonalidade analisará graficamente sua identificação, além de serem realizados testes estatísticos que comprovem a presença dessa

componente na série. A metodologia de teste de sazonalidade utilizada consiste em combinar os resultados dos testes QS e Kruskall-Wallis, ambos calculados sobre os resíduos de um modelo ARIMA não sazonal automaticamente gerado. Essa combinação é chamada de teste de Ollench-Webel (WEBEL e OLLECH, 2019), e seu resultado é uma soma booleana que classifica a sazonalidade presente na série original como verdadeira ou falsa.

b) Autocorrelação

Propõe-se então que a série temporal seja analisada quanto a sua autocorrelação através de análise gráfica do autocorrelograma, seguido de análise interpretativa. O autor Chatfield (1996) propõe diversas interpretações para o autocorrelograma de uma série, constituindo pontos de atenção ao analista, como por exemplo: se uma série é completamente aleatória a autocorrelação será próxima de zero. Séries estacionárias frequentemente exibem uma correlação de curto-termo, fazendo com que o valor da autocorrelação seja razoavelmente grande nos primeiros *lags*, seguido por valores que tendem a ser sucessivamente menores. Séries alternadas, como o nome diz, possuem uma tendência de alternância com observações em diferentes lados da média global. Séries não-estacionárias, se a série possui tendência então o valor da autocorrelação não chegará a zero (autocorrelação só tem significado para séries temporais estacionárias), assim, toda tendência deve ser removida antes de seu cálculo. Flutuações sazonais, se a série possuir sazonalidade, o correlograma exibirá uma oscilação na mesma frequência.

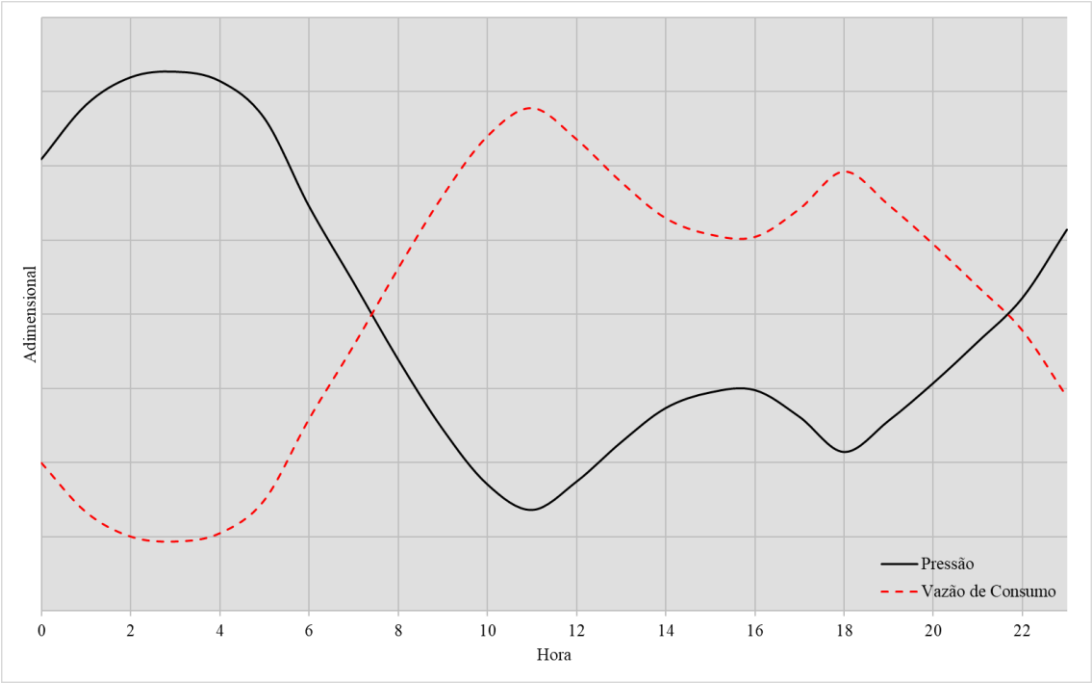
c) Perfil de variação de consumo de água

A Figura 7 apresenta uma curva adimensional média da vazão de consumo e da pressão em uma rede de abastecimento de água em função da hora do dia. Os valores adimensionais foram tomados a partir da divisão dos valores observados pela pressão média daquele sensor. A partir da análise visual do gráfico é possível observar os conceitos anteriormente apresentados, com uma variação de vazão de consumo e de pressão ocorrendo de forma inversamente proporcionais, com os pontos de vazão máxima coincidindo com os de pressão mínima (entre as 10 e 12 horas do dia) e com o ponto de vazão mínima coincidindo com o ponto de pressão máxima (próximo das 3 horas).

Com o objetivo de aplicar os conceitos aqui apresentados, como forma contribuir para o entendimento do comportamento da série temporal serão construídas curvas típicas de variação da pressão na rede de distribuição de água monitorada. Através dos gráficos de curva

adimensional construídos para possibilitar a visualização da pressão na rede em função da hora do dia e do dia da semana.

Figura 7 - Curva adimensional média da vazão de consumo e da pressão em uma rede de abastecimento de água em função da hora do dia.



3.4 Ajuste de modelos de previsão

Em termos práticos metodológicos, valendo-se das constatações teóricas apresentadas no item 2.3 Modelos de previsão para séries temporais, foram combinadas estratégias para ajustar os seguintes modelos: SARIMA, seguido de ARIMA adicionado de variáveis exógenas do tipo harmônicas, denominado aqui de ARIMA Harmônico, e o terceiro modelo implementado GARMA. Os dois primeiros modelos foram aplicados primeiramente para um único sensor da série selecionada, o sensor 7, com o objetivo de avaliar seu desempenho e limitações antes de se escalar os testes para os demais sensores.

3.4.1 Ajuste do modelo SARIMA

Uma das características do modelo SARIMA é que ele pode lidar com apenas uma sazonalidade e não múltiplas, assim, para sua aplicação definiu-se trabalhar com a sazonalidade diária. Para lidar com a componente sazonal, neste caso em que não há preditores como no modelo ARIMA Harmônico, deve-se aplicar a transformação de diferenciação sazonal, descrita no item 2.1.1 Tendência, sazonalidade e estacionariedade.

Utilizou-se o algoritmo descrito na seção 2.3.4 para identificação da ordem das componentes sazonais e não sazonais do ARIMA. Foram considerados valores de “d” igual a zero e “D” igual a 1, ou seja, não foi necessária a diferenciação para a parte não sazonal do modelo e para a parte sazonal foi aplicada 1 diferenciação. Com o modelo ajustado segue-se para as etapas de previsão e avaliação de métricas de performance.

3.4.2 Ajuste do modelo ARIMA + Harmônico

A estratégia metodológica neste caso consiste em utilizar uma série de Fourier como preditores no modelo, ou seja, esses valores entram como variáveis exógenas para representar a componente sazonal e os erros são modelados pelo processo ARIMA. Como a utilização das séries de Fourier como preditores do modelo permite que se trabalhe com mais de uma sazonalidade, definiu-se então um conjunto de pares senos e cossenos para a sazonalidade diária e um conjunto para a sazonalidade semanal.

Neste método o número de pares senos e cossenos são definidos manualmente antes da etapa de identificação do modelo. Conforme apresentado por Hyndman & Athanasopoulos (2021), não há uma regra clara sobre a definição deste número de pares (k), então adotou-se o número de 6 termos senos e cossenos, sendo valores próximos utilizados pelos autores em uma modelagem realizada com dados de escala semelhante a que se trabalha nesta pesquisa e pelos resultados apresentados por Brentan et al. (2017) que demonstram que a partir de 3 pares os erros entre dados observados e previstos reduz consideravelmente.

Após essa definição, é utilizado o algoritmo descrito na seção 2.3.4, para identificação e estimação dos componentes autorregressivos e de médias móveis do modelo. Com o modelo ajustado segue-se para as etapas de previsão e avaliação de métricas de performance.

3.4.3 Processo de modelagem das séries temporais de pressão com modelo GARMA

A seleção dos modelos a serem aplicados nesta pesquisa se deu pela sequência lógica de complementação das limitações encontradas, ou seja, a modelagem da série com modelo SARIMA apresentou limitações quanto da modelagem da sazonalidade complexa, fenômeno descrito nas seções anteriores que se manifesta no presente conjunto de dados e que levou ao uso da estratégia de modelar a sazonalidade complexa com séries de Fourier (modelo ARIMA+Harmônico). Essa segunda estratégia por sua vez, apesar de lidar bem com as questões de sazonalidade, não foi capaz de extrair toda informação presente na série, principalmente

quanto à heterocedasticidade dos dados, fazendo com que os resíduos do modelo não obedecessem a todos os pressupostos teóricos.

Assim, para lidar com essa limitação, é estudada e aplicada a modelagem com modelo GARMA, por sua característica principal de modelar dados a partir de diferentes funções de densidade de probabilidade. Para isso, antes do ajuste do modelo, será realizada nova etapa de pré-processamento de dados.

3.4.3.1 Pré-processamento de dados para modelagem com GARMA

Diferentemente da distribuição normal, que possui uma única forma simétrica, e parâmetros média e desvio padrão, a distribuição gama é assimétrica, podendo apresentar diferentes formas e caudas, e possui como parâmetros alfa (forma) e beta (escala) o que a torna adequada para modelar fenômenos com valores sujeitos à variabilidade não normal. As Figura 8 e Figura 9 abaixo apresentam exemplos da forma que essas funções assumem.

Figura 8 - Gráfico de uma função de densidade de probabilidade gama genérica

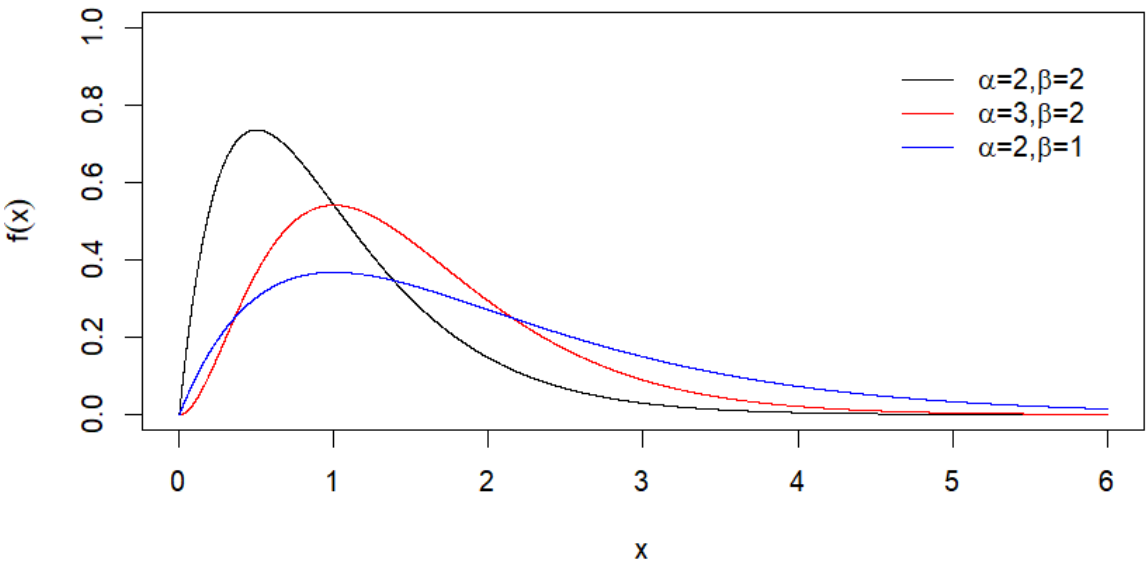


Figura adaptada a partir de Mattos (2018).

Figura 9 - Gráfico de uma função de densidade de probabilidade normal genérica.

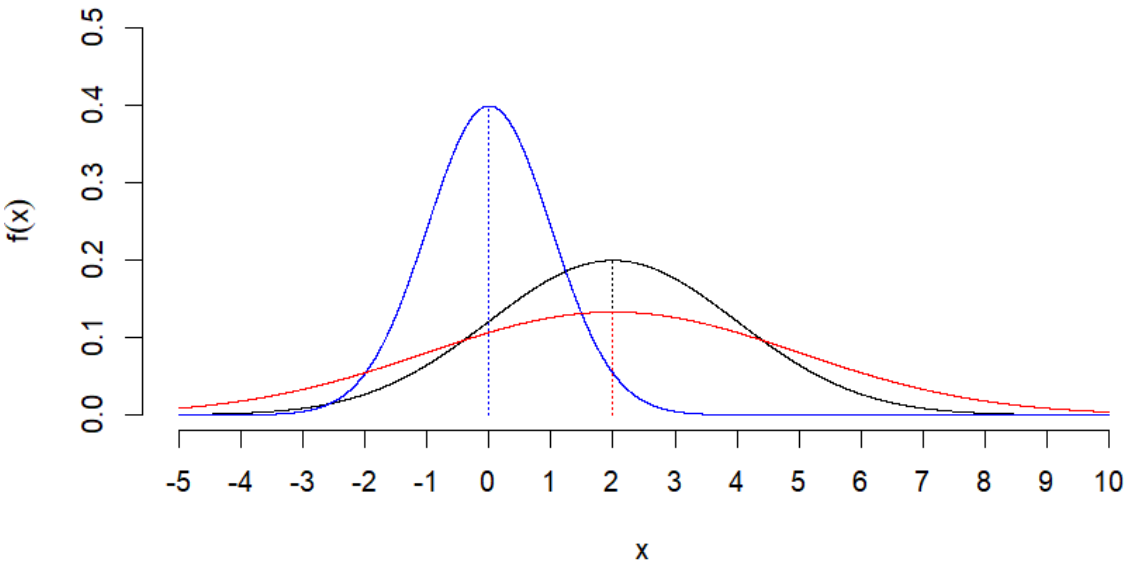


Figura adaptada a partir de Mattos (2018).

O fato de a função de densidade de probabilidade gama assumir formas assimétricas resulta em uma melhor representação dos picos das séries de dados. Fazendo uma analogia com os nossos dados, e sua ligação com o padrão de consumo de água típico em setores de abastecimento com características residenciais, como apresentado anteriormente, temos que a vazão nas redes de abastecimento possui maior variação durante o dia quando há os maiores consumos, e baixa variação de seus valores durante a noite quando o consumo é também baixo. Como a pressão possui comportamento inversamente proporcional à vazão, durante a noite quando assume valores mais altos (picos) sua variação é baixa, e durante o dia quando a pressão é baixa na rede (vales) coincide com a maior variação de valores, em função do consumo.

Dessa forma, procedeu-se com a inversão da série de pressões, de forma a calcular para cada sensor a diferença entre a pressão estática máxima e o valor observado, com correção do nível médio do reservatório do setor. Ao resultado dessa diferença atribui-se o nome de “Energia consumida” (mca). Desta forma obteve-se uma série de valores com maior variância nos picos, formato adequado para modelagem com função de densidade de probabilidade gama.

As figuras abaixo Figura 10 e Figura 11 representam um exemplo de trechos da série original, e o mesmo trecho da série após transformação, respectivamente. Nota-se que, a partir do novo conjunto de dados a série possui o comportamento esperado, onde os vales são mais uniformes e as maiores variações nas observações estão nos picos.

Figura 10 - Gráfico de trecho da série original para observação de padrões de comportamento.

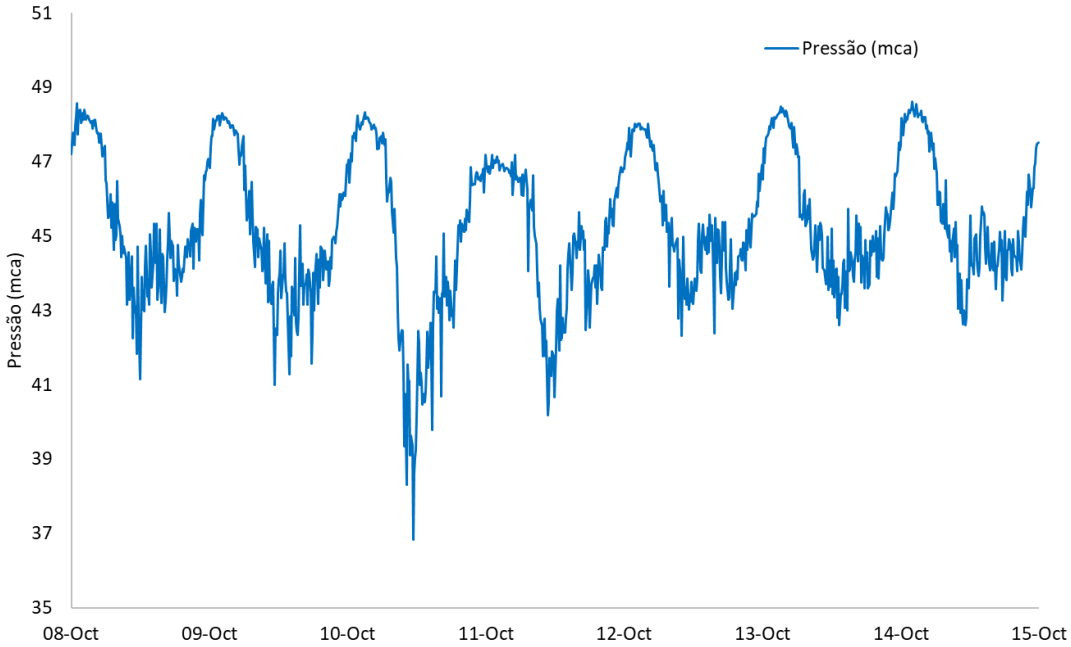
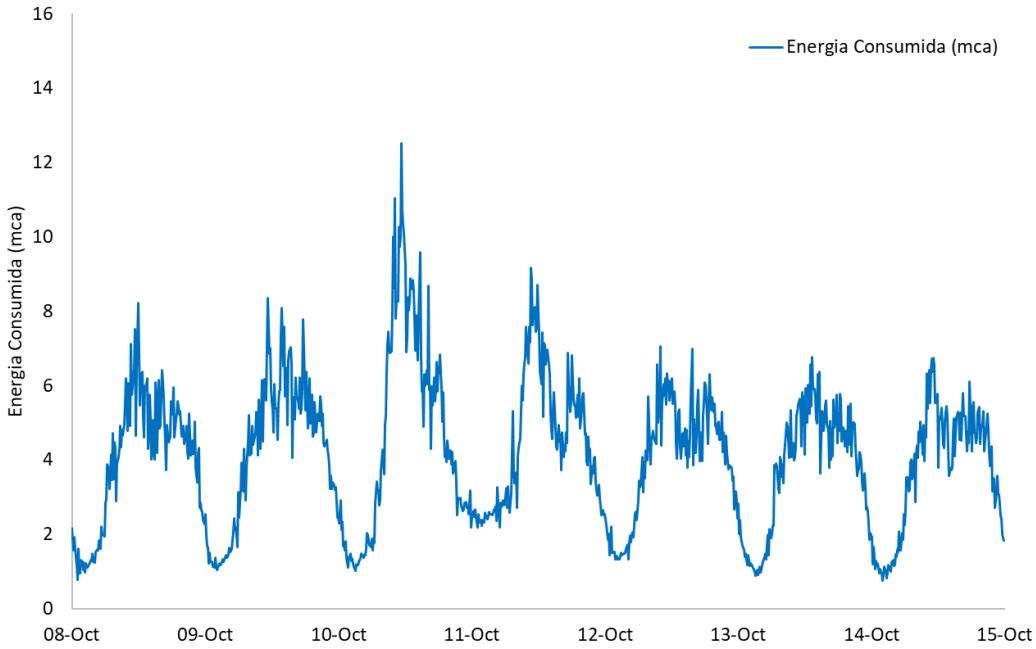


Figura 11 - Gráfico de trecho da série transformada na forma de Energia Consumida (mca).



3.4.3.2 Ajuste do modelo GARMA

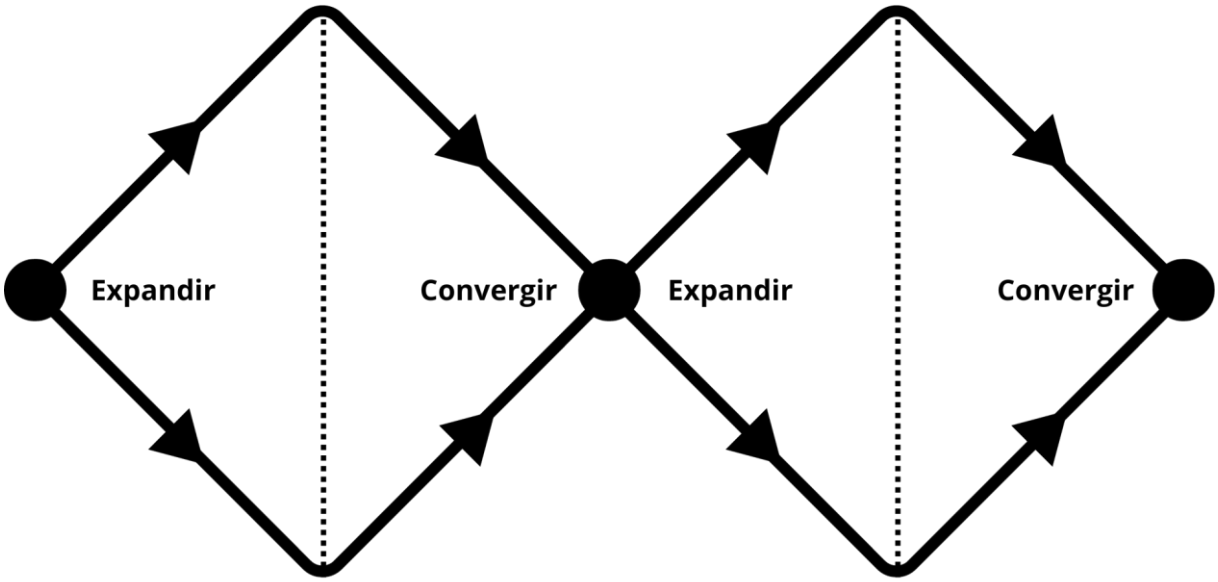
Para o ajuste do modelo GARMA se define a série a ser modelada e as variáveis exógenas. Nesse caso foram incluídas *dummies* de finais de semana, uma para representação do sábado e outra para representação do domingo. A inclusão de *dummies* é uma estratégia para de forma numérica representar variáveis categóricas, como os dias da semana, dentro do modelo. As mesmas consistem em um conjunto binário, sendo denominadas de d_01 a variável *dummy* que representa o sábado, recebendo em cada observação da série o valor de 1 quando verdadeiro (o dia da semana é sábado) e 0 caso contrário, e d_02 a variável *dummy* que representa o domingo, de igual forma recebendo em cada observação da série o valor de 1 quando verdadeiro (o dia da semana é domingo) e 0 caso contrário.

Define-se também o número de pares senos e cossenos, da mesma forma que apresentado anteriormente na modelagem da sazonalidade com séries de Fourier. Quanto a parte ARMA(p,q) do modelo, são testadas diversas combinações de modelos, variando os valores da ordem de 0 a 3. Novamente aqui, o valor de AIC é um primeiro apontamento para melhor combinação de parâmetros e em um segundo momento é realizada a análise de resíduos, possibilitando a verificação do atendimento aos pressupostos do modelo. Nesta etapa é também definida a função densidade de probabilidade adequada, no caso, a distribuição gama. O ajuste se dá através do uso da função “garmafit()” no software R, desenvolvida por Rigby et al. (2019).

3.4.3.3 Seleção do conjunto de ajuste e do horizonte de previsão do modelo GARMA

Por diversas vezes, para viabilizar a verificação das hipóteses de pesquisa precisa-se realizar o movimento de aumentar e diminuir a dimensionalidade das análises, seja através de análises de sensibilidade de parâmetros, seleção de variáveis ou comparação de métricas de acurácia. Assim, foi proposta uma análise inspirada na ferramenta de design de soluções proposta pelo Design Council (2004), denominada diamante duplo, onde os dois diamantes representam um processo de explorar um problema de forma mais ampla ou profunda (pensamento divergente) e, em seguida, tomar ações direcionadas (pensamento convergente) (Figura 12).

Figura 12 - Estrutura de inovação - modelo diamante duplo



Adaptado de Design Council (2004).

Desta forma a etapa de seleção do conjunto de ajuste do modelo e do horizonte de previsão do modelo GARMA, é considerada uma etapa de expansão (primeira expansão conforme Figura 12), onde se abre um leque de análises combinando os diversos fatores relativos à seleção da modelagem mais adequada aos propósitos requeridos. O resultado desta etapa irá convergir na seleção do modelo desejado (primeira etapa de convergência conforme Figura 12). A partir deste ponto expande-se (segundo movimento de expansão conforme Figura 12) novamente a análise comparando o desempenho do modelo GARMA com os demais modelos propostos, até que haja convergência final da comparação, etapa final da análise e que será descrita com mais detalhes no item 3.5. Este duplo movimento de expansão e convergência dá-se o nome de diamante duplo.

Por ora, com o objetivo de avaliar quantitativamente qual combinação de conjunto de ajuste, e horizonte de previsão, apresenta melhor desempenho na acurácia de previsão para cada sensor, com modelo GARMA, foi desenhada uma estratégia iterativa que possibilitou a combinação de todas as variáveis. Em resumo, foram calculadas as métricas de acurácia de previsão (a serem apresentadas no item 3.6) para cada um dos 9 sensores, nos horizontes de previsão de 24 e 48 horas, para cada um dos 12 meses do ano, comparando-se modelos ajustados com conjuntos de dados denominado “completo” - quando ajustado com todos os dados observados para aquele sensor durante o ano- e modelo denominado “mensal” - quando ajustado com 2 semanas de dados no mês presente à comparação.

Esta análise resultou nas respostas para as seguintes perguntas: qual conjunto de treino de ajuste é melhor, dentro das métricas propostas? Quão pior é fazer previsão de 48 horas? A combinação selecionada é válida para todos os sensores, e em todos os meses do ano? Quais características das séries podem influenciar nos resultados? E as respostas a estes questionamentos convergem para a próxima etapa de expansão.

3.5 Organização de testes de previsão com os modelos

De posse dos modelos ajustados e escopo de ajuste definido conforme seção 3.4 Ajuste de modelos de previsão, segue-se para a etapa comparativa entre modelos. Segundo definição de Lazzeri (2020), o conjunto de treino é o montante de dados utilizados para estimação dos parâmetros do modelo, é a partir dele que são extraídos os padrões de comportamento da série. O conjunto de validação oferece avaliação não enviesada da estimação e pode ser utilizado para ajuste fino dos parâmetros. O conjunto de teste por sua vez é utilizado para determinar se o modelo apresenta *"underfitting*", quando a performance é ruim na base de treino, ou "*overfitting",* quando o modelo tem boa performance na base de treino e uma má performance na base de teste, através da avaliação dos erros de previsão sobre os dados de teste.

A aplicação do conceito de *rolling analysis* (análise móvel) para um modelo de séries temporais é geralmente utilizado para avaliar a estabilidade do modelo ao longo do tempo. Bem como, segundo Zivot e Wang (2003), para avaliar a acurácia de previsão através de um *backtest*. O *backtest* geralmente funciona da seguinte forma. O modelo é ajustado para a base de treino e faz-se previsões *h-passos* à frente sobre a base de teste, assim é possível avaliar os erros entre os valores preditos e a base de teste composta por valores observados. Então, move-se a base de treino para a frente (conceito janela móvel (JM)) dado um valor incremental e então repete-se o ajuste até que não haja mais amostras para testes.

Com isso, seguiu-se para a separação de trechos de interesse na base completa, que levou em consideração alguns critérios, como: não existência de dados faltantes e todos os dados classificados como 0 – "normal". Fez-se avaliação destes critérios para um conjunto que considerasse os 9 sensores simultaneamente, com o objetivo de proporcionar validação dos modelos posteriormente.

Desta forma, o maior conjunto obtido foi de 2736 pontos, equivalente a 19 dias contínuos de monitoramento. O início dessa série se dá no dia 280 (08/10/2015 00:00) de monitoramento e vai até o dia 299 (23/10/2015 23:50). De posse desta série procedeu-se com a

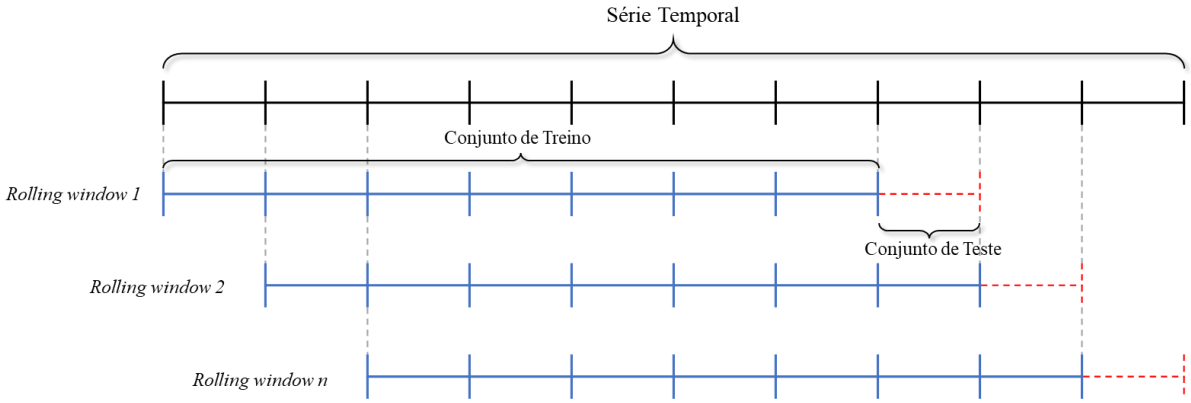
separação da base de dados em conjuntos de treino e teste no formato de janelas móveis. Assim, foram criadas 5 janelas de análise, sendo o conjunto de treino constituído de 14 dias de observações e o conjunto de teste 1 dia, mas detalhes constam na Tabela 1. Para fins de análise, é importante saber que o período de teste, onde se faz a previsão da série, a partir da JM#1 corresponde ao dia da semana de quinta-feira e vai até a JM#5 segunda-feira.

Tabela 1 - Detalhamento do conjunto de dados de treino e teste segundo a respectiva janela móvel de análise.

Janelas Móveis	**Início da base de treino (dias)**	**Fim da base de treino (dias)**	**Dia referente a base de teste**	**Dia da semana correspondente**
#1	280	294	295	Quinta-feira
#2	281	295	296	Sexta-feira
#3	282	296	297	Sábado
#4	283	297	298	Domingo
#5	284	298	299	Segunda-feira

A Figura 13 apresenta o esquema de separação da série temporal em conjuntos de treino e teste de acordo com a metodologia de janelas móveis para realização do *backtest*. É possível observar que ao mover-se a janela selecionada h-passos à frente tem-se a possibilidade de avaliar a performance do modelo em prever sobre um conjunto de teste diferente, fazendo com que se aproveite melhor a série temporal completa. Outra vantagem está relacionada com a semelhança dessa estratégia ao se pensar em um sistema de automação em produção, é comum que para promoção da escalabilidade do modelo para um sistema autônomo de monitoramento dado um conjunto de dados recentes passados, faz-se a previsão para um dia a frente, para que se execute os algoritmos uma única vez ao dia reduzindo custo computacional, por consequência, no dia seguinte, atualiza-se a base de treino acrescentando as novas medições do dia.

Figura 13 - Esquema de separação da série temporal em conjuntos de treino e teste de acordo com a metodologia de *rolling analysis* e *backtest.*



Adaptado de Larrubia (2021).

Para o modelo SARIMA, o algoritmo de identificação automática dos modelos descritos nas seções 2.3.4 ajustaram-se os modelos somente para a primeira janela de análise e então a mesma ordem e parâmetros foram replicados nas demais janelas móveis para se fazer a previsão.

Para o modelo ARIMA com termos da série de Fourier como preditores, o algoritmo de identificação automática dos modelos descritos nas seções 2.3.4, ajustou o modelo para a primeira janela de análise e apenas a sua ordem identificada foi replicada nas demais janelas. Foi necessário reestimar os parâmetros e os termos de Fourier devido ao deslocamento realizado h-passos à frente, que altera o tempo de referência entre cada análise.

Para o modelo GARMA, o processo é semelhante ao ajuste do modelo ARIMA+Harmônico, sendo diferente apenas a etapa de identificação automática da ordem do modelo, pois a ordem adequada provém do teste de diferentes combinações conforme descrito na seção anterior.

3.6 Métricas de avaliação de modelos

A etapa de avaliação dos modelos tem o objetivo de validar os parâmetros ajustados e sua performance. Enquanto a base de treino é utilizada para calibrar o modelo, pode-se também aplicar as medidas de acurácia nessa etapa como forma de validação dos parâmetros inicialmente ajustados.

A avaliação sobre a base de teste mensura a performance preditiva do modelo ajustado e validado. Busca-se então mensurar e comparar o erro entre os valores originais e preditos, a meta é minimizar esta diferença (LAZZARI, 2020). Para isso existem medidas que são relevantes para séries temporais, que foram adotadas neste trabalho. Sendo elas MAPE (Equação 11) (*mean absolute percentage error*), que representa o erro percentual médio absoluto e RMSE (Equação 12) (*root mean squared error*), que representa a raiz do erro quadrático médio, dadas pelas seguintes fórmulas:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100\% \quad \text{Eq. 11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \quad \text{Eq. 12}$$

Ressalta-se que são medidas complementares na análise da qualidade do ajuste do modelo e da previsão, pois a medida MAPE é percentual, não possuindo assim dependência de escala com a base de dados podendo ser utilizada para comparar a acurácia dos modelos nos diferentes sensores que existem na base desta pesquisa, bem como inclusive com os resultados encontrados na literatura sobre aplicações em outros sistemas. Diferentemente da medida RMSE que possui dependência de escala e é mais útil na avaliação de diferentes modelos somente para uma mesma base de dados. Ambas as métricas são avaliadas de acordo com quanto menor for o seu valor, melhor é o resultado.

3.7 Exploração da resposta da previsão em cenários de vazamentos

De posse da previsão, com seu respectivo intervalo de confiança, foram exploradas as possibilidades de detecção de eventos a partir das simulações realizadas por GAMBOA-MEDINA (2017). Assim, selecionou-se um evento onde foi simulado um vazamento no setor monitorado através da abertura de um hidrante. O evento ocorreu em 28/10/2015 a partir das 22:00h, e teve duração de 30 minutos, localizado no nó h5, conforme a Figura 5. A vazão simulada foi de 28 l/s.

Uma vez que vazamentos impactam o comportamento dos dados com quedas de pressão, criou-se função para destacar a partir de gráficos os pontos cujo valor observado estivesse abaixo do intervalo de confiança mínimo. Assim, foi possível avaliar dentre os nove sensores

quais aqueles que conseguiram detectar o evento, para que se pudesse explorar a resposta dos sensores em cenários de vazamentos.

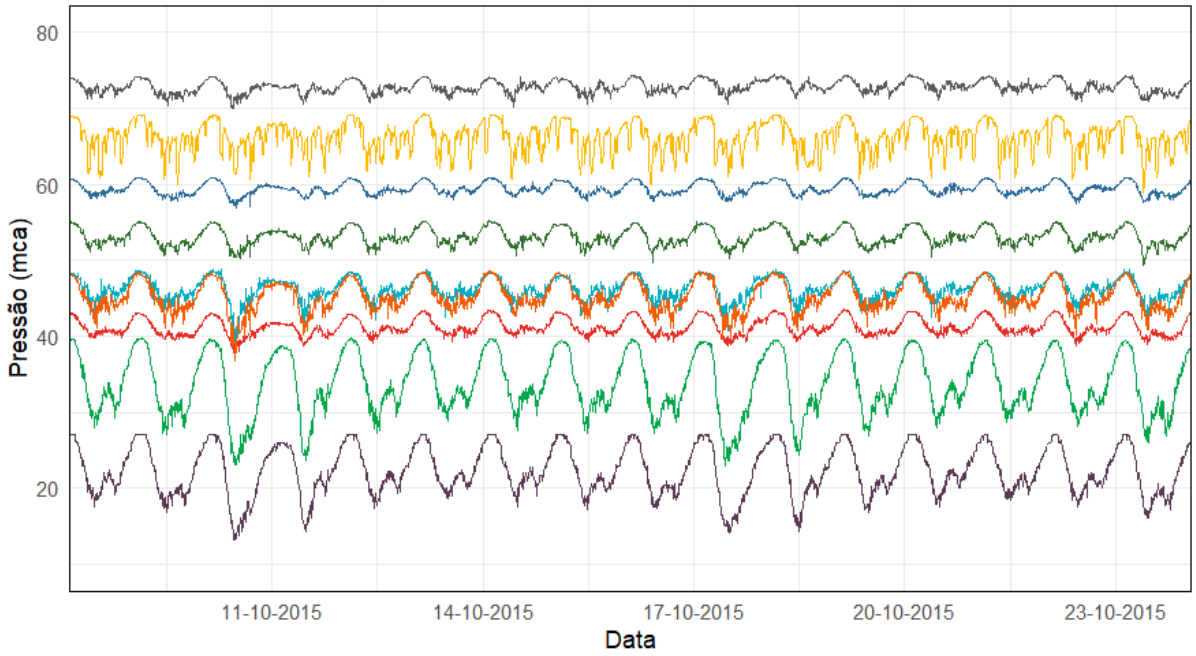
4 RESULTADOS E DISCUSSÕES

Nesta seção são reportados os resultados encontrados durante o desenvolvimento da pesquisa com foco principal na descrição da série temporal estudada, nos ajustes dos modelos e na acurácia de suas previsões. Dessa forma, estabelecendo relação entre as características das séries com seus respectivos mecanismos geradores e consequente desempenho dos modelos.

4.4 Descrição do comportamento da série

A seção 3.2 “Pré-processamento e modelagem de dados” descreveu o processo de classificação dos trechos da base de dados original, sobre os trechos classificados como em condição de normalidade se direciona esta análise exploratória e descritiva. Os gráficos apresentados anteriormente, como na Figura 6, devido à escala, transmitem mais uma ideia geral sobre a disponibilidade de dados, a distribuição dos sinais nos diferentes períodos do ano, porém não se consegue extrair muitas informações sobre o comportamento das séries. Buscando por um maior detalhamento, a Figura 14 apresenta as séries temporais de pressão de cada um dos sensores monitorados com foco no período que compreende o trecho mais longo sem falhas do monitoramento.

Figura 14 - Detalhamento das séries temporais de pressão de cada um dos sensores monitorados.



Através da Figura 14 é possível visualizar algumas nuances sobre o comportamento geral das séries, percebe-se que cada um dos sensores trabalha em uma média diferente, e isso se dá por conta da diferença de altitude entre o local em que foram instalados e a localização dos reservatórios. Fica evidente também os ciclos de repetição diários e semanais, que estão relacionados com a variação do consumo de água por parte da população ao longo da rede, outro detalhe é que mesmo seguindo o mesmo ciclo, os sensores capturam sinais em amplitudes diferentes.

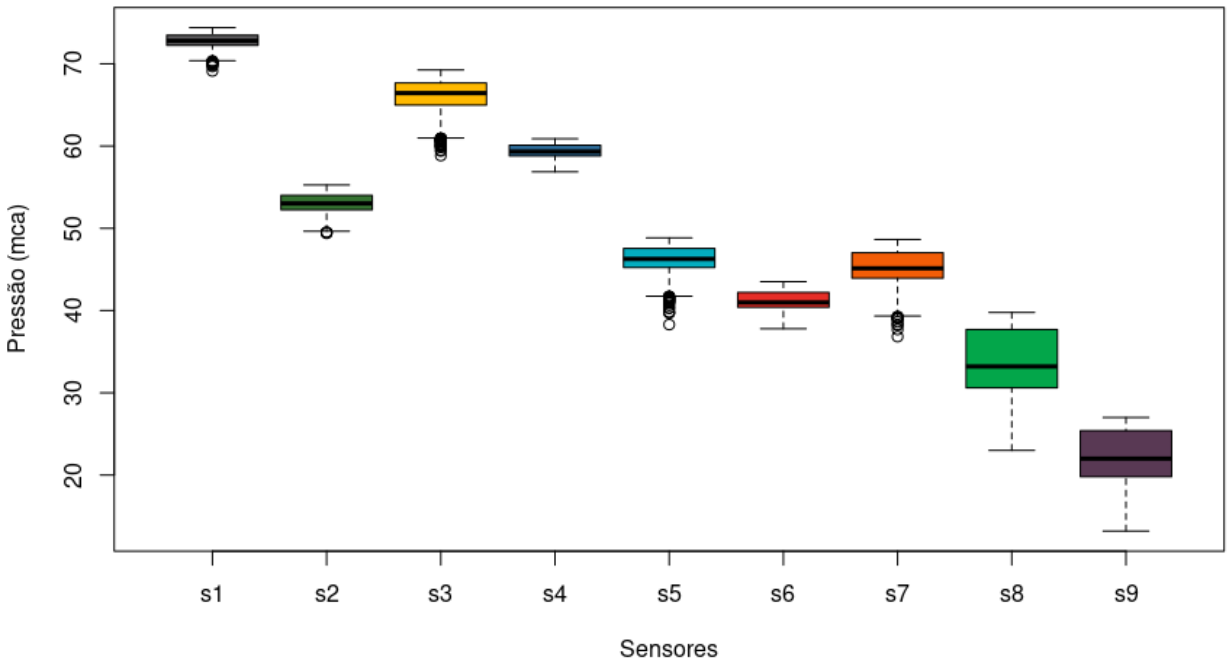
Para esclarecer estas diferenças, a Tabela 2 resume as principais medidas estatísticas das séries. A medida de desvio padrão apresenta uma dimensão da dispersão dos dados em torno da média. Neste sentido, o sensor s4 é o que possui menor desvio padrão, com valor de 0.81, muito próximo do s1, que é de 0.84. O sensor s8, além de possuir uma grande amplitude de medições, evidenciadas pela diferença entre valores máximos e mínimos, é o que possui maior desvio padrão, com valor de 4.13.

Tabela 2 - Resumo de medidas estatísticas de pressão (mca) de cada um dos sensores monitorados.

Sensor	Mínimo	Mediana	Média	Máximo	Desvio Padrão
s1	69.13	72.79	72.81	74.39	0.84
s2	49.39	53.02	53.08	55.26	1.14
s3	58.84	66.43	66.14	69.24	2.03
s4	56.86	59.33	59.41	60.89	0.81
s5	38.3	46.27	46.23	48.84	1.55
s6	37.79	40.98	41.23	43.51	1.15
s7	36.84	45.13	45.29	48.64	1.98
s8	22.99	33.20	33.62	39.77	4.13
s9	13.20	22.00	22.21	27.00	3.31

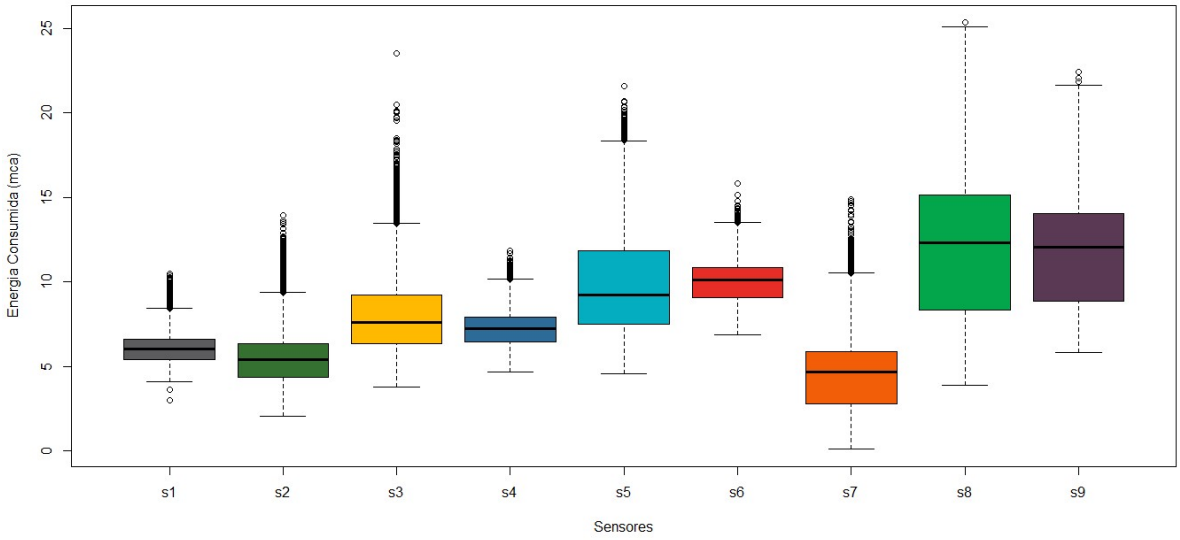
Os valores médios são diferentes para cada um dos sensores, sendo o s1 aquele que possui a maior média entre todos, 72.81 mca, e o s9 com 27 mca é o que possui a menor média. Com atenção para os sensores s5, s6 e s7, que possuem médias muito próximas. O diagrama de caixas das séries originais (sem transformação) apresentado na Figura 15 contribui com a análise tanto dos valores médios de cada sensor, quanto com a amplitude que evidencia as sobreposições, além de indicar a presença de possíveis *outliers*.

Figura 15 - Diagrama de caixas dos valores de pressão para cada um dos sensores monitorados.



Como resultado da transformação de dados descrita na seção 0 3.4.3.1 Pré-processamento de dados para modelagem com GARMA, há um novo comportamento dos dados, que pode ser visualizado na Figura 16. Com a transformação dos dados de pressão para a variável “Energia Consumida” houve uma equalização dos dados para um nível muito próximo, com sua amplitude de medições se sobrepondo na maioria dos pontos. Logo, essa transformação resultou em benefícios não restritos ao comportamento adequado para aplicação da função de densidade de probabilidade gama no modelo GARMA, mas também a possibilidade de ser testada a generalização de modelos que pode resultar em vantagens do uso da variável Energia Consumida.

Figura 16 - Diagrama de caixas dos valores de Energia Consumida após transformação de dados para cada um dos sensores monitorados.



Quanto à tendência dos dados, através da Figura 14 pode-se perceber que não há de forma clara um comportamento de crescimento ou decrescimento da média ao longo do tempo, quando se analisa a série em um período de tempo da ordem de semanas. Adicionalmente à análise gráfica, os testes de raízes unitárias de *Dikey-Fuller* Aumentado (ADF) retornaram p-valores menores do que o nível de significância $\alpha = 0.05$, rejeitando-se a hipótese de que existe raiz unitária, logo concluindo que a as séries são estacionárias quanto a tendência , nos trechos que possuem tamanho da ordem de 14 a 19 dias.

Porém, na análise das séries completas, período de um ano de monitoramento, a análise gráfica aponta que os sensores s2 e s5 apresentam uma mudança de média ao longo do tempo. Sendo a mudança no sensor 2 (s2) próxima ao dia 16/03/2015, e a mudança no sensor 5 (s5) a partir de 15/08/2015, caracterizando a não estacionariedade das séries, as Figuras Figura 17 e Figura 18 abaixo apresentam com detalhe as séries dos sensores 2 e 5 respectivamente.

Figura 17 - Gráfico da série de pressão (mca) observados referente ao sensor s2, em todo período de monitoramento.

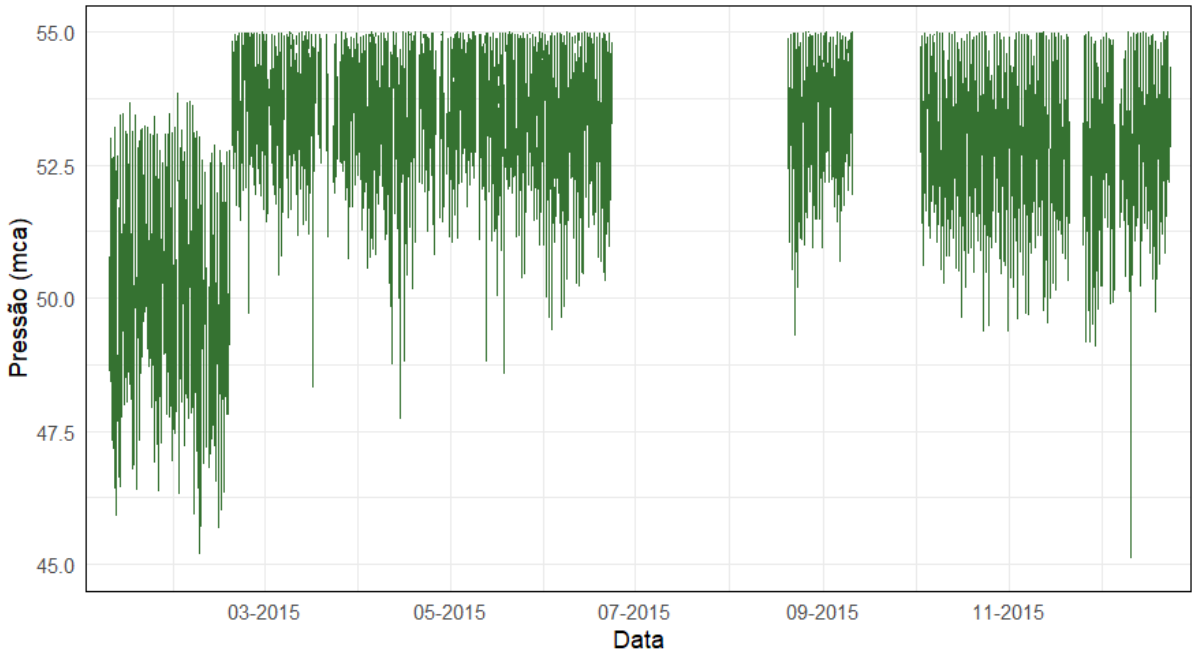
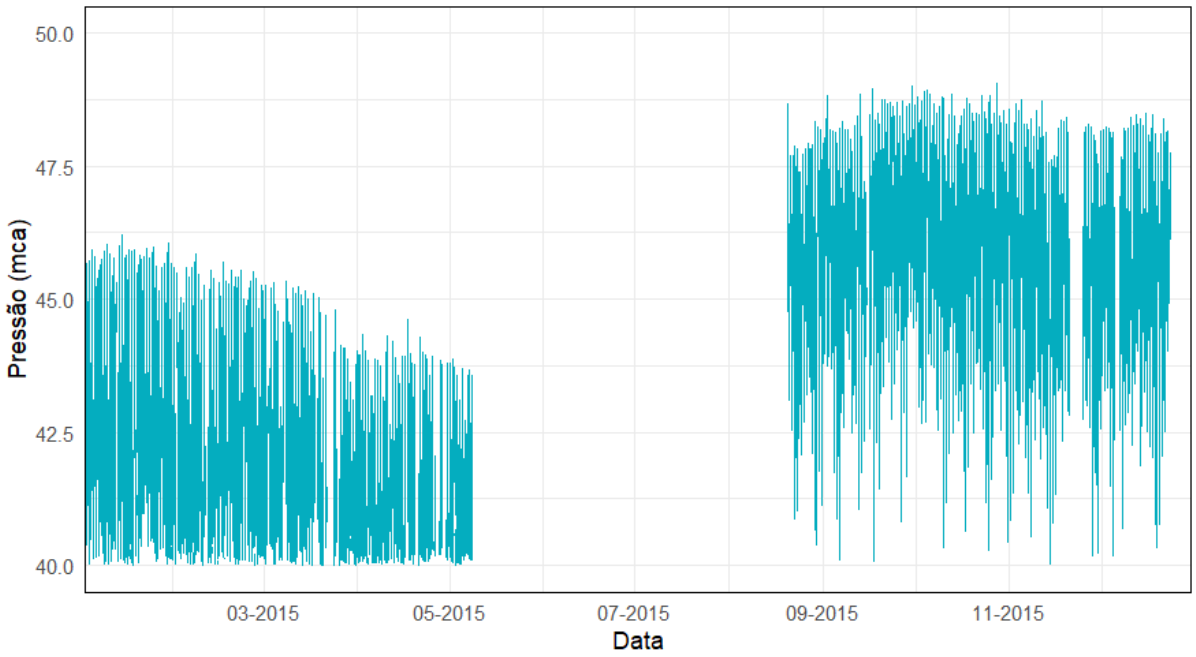


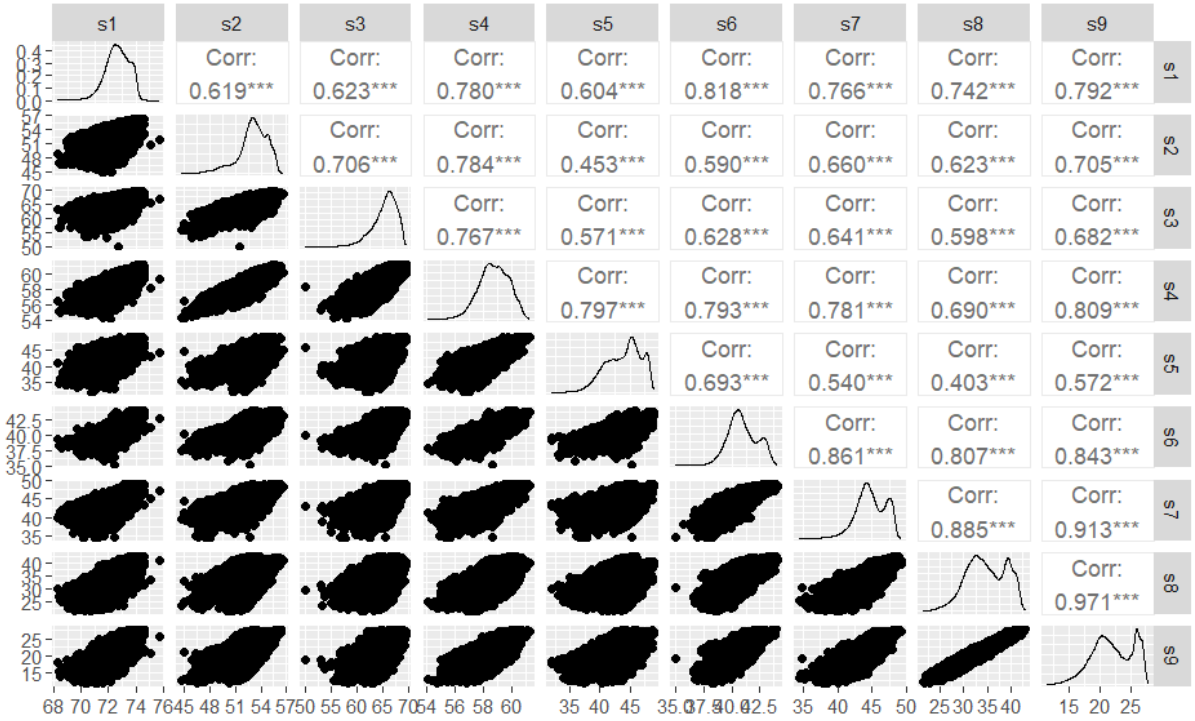
Figura 18 - Gráfico da série de pressão (mca) observados referente ao sensor s5, em todo período de monitoramento.



O fato dos sensores estarem localizados no mesmo setor de medição, e as séries possuírem comportamento periódico bem definido transmite a ideia de que os dados observados são correlacionados entre si. Procedeu-se assim, com análise da correlação entre os dados dos sensores como forma de identificar padrões de comportamento e relacionamentos entre os diferentes pontos de monitoramento, conforme Figura 19. Através da correlação, é possível

quantificar o grau de associação entre as séries temporais, determinando se elas variam de maneira semelhante, oposta ou independentemente. Essa abordagem estatística não apenas fornece uma medida objetiva da interdependência entre os pontos, mas também ajuda a revelar ideias sobre a influência de eventos ou mudanças operacionais em toda a rede.

Figura 19 - Matriz de correlação entre os dados observados dos sensores monitorados.

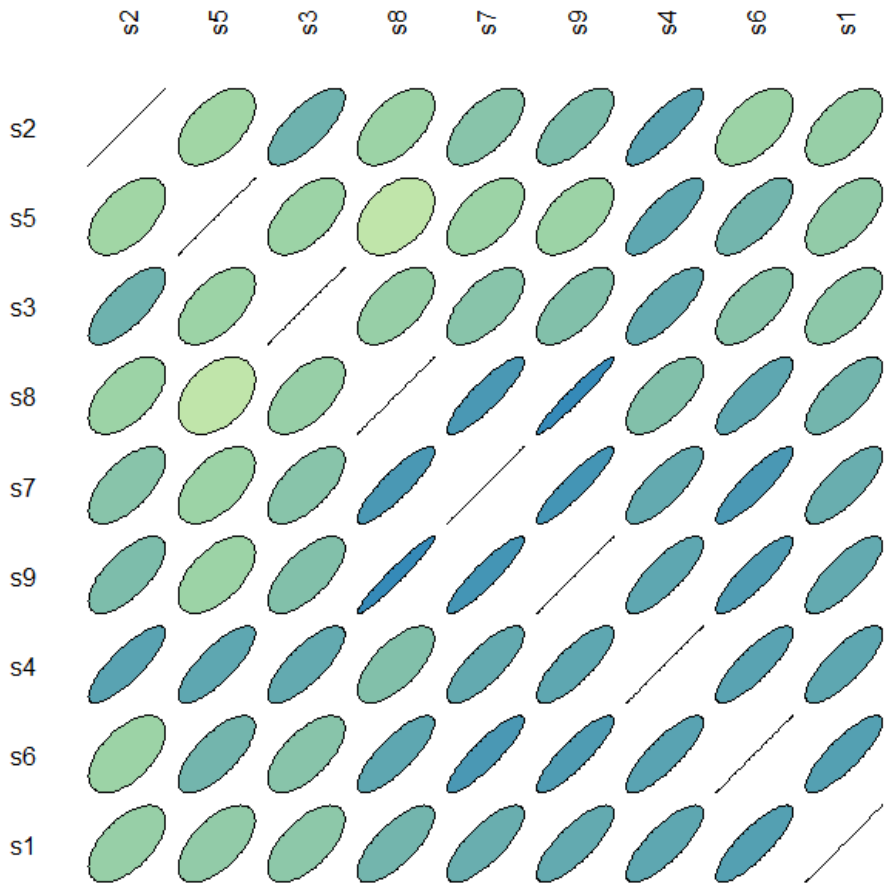


Através da Figura 19 é possível observar gráficos de dispersão entre as observações dos sensores avaliados conforme a matriz, bem como a curva de densidade de distribuição dos dados na diagonal principal, e o valor da correlação entre os sensores monitorados. Em geral as correlações são fortes e positivas, com valores superiores a 0.50, com exceção da correlação entre os sensores s8 e s5, com valor de 0.40. O sensor s8 inclusive apresenta resultado peculiar, sendo a pior correlação entre todas as combinações de sensores, com o sensor s5, e a melhor delas, com o sensor s9 no valor de 0.97.

A correlação também foi avaliada através de um gráfico correlograma de elipses, que desenha para cada correlação uma elipse que varia em formato e cor de acordo com o grau de correlação entre os sensores. De acordo com a Figura 20, pode-se interpretar os resultados como quanto mais achatada e mais azul for a elípse, significa que mais forte é a correlação e quanto

menos achatada e mais verde for a elípse, menor a correlação entre os sensores. Além de a direção da elípse apontar se a correlação é positiva ou negativa. Uma outra análise suportada por este gráfico é que a ordem dos sensores é hierarquizada, partindo daqueles com menor correlação média com os demais sensores, sensor s2, ao sensor com maior correlação média sensor s1. Pode-se observar ainda, que os dois sensores que aparecem com menores correlações médias são o s2 e s5, os mesmos apresentados anteriormente como os que possuem a série anual não estacionária, com variação brusca de seu comportamento médio ao longo do ano.

Figura 20 - Gráfico de correlação de elipses das séries originais.

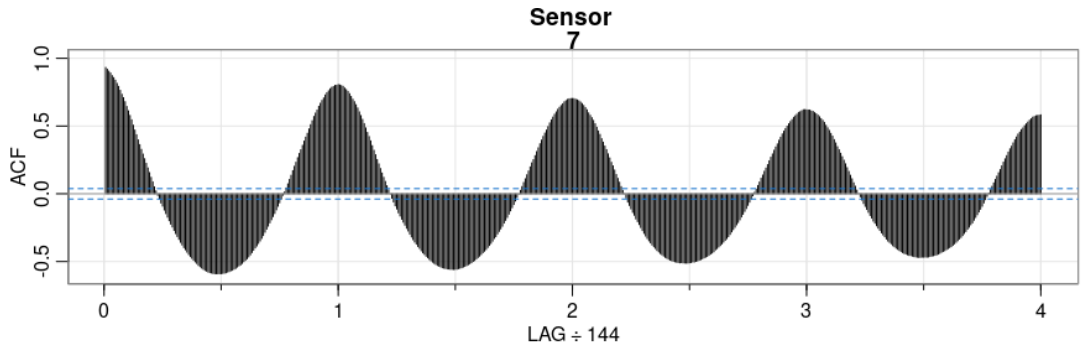


A avaliação da sazonalidade também se dá de forma combinada entre análise gráfica e testes estaísticos, as observações da série original na Figura 14 já indicam um padrão muito claro de repetição na série, tanto da curva característica de cada dia, quanto das diferenças existentes nestas curvas a cada 7 repetições. Ao observar as séries dos sensores s8 e s9 (os dois

de menores médias localizados na parte baixa do gráfico da Figura 14), que possuem maiores amplitudes de medição, pode-se visualizar melhor esse padrão.

Os gráficos das funções de autocorrelação para cada um dos sensores elucidam esse padrão de repetição que caracteriza a sazonalidade determinística presente na série. A Figura 21 apresenta estes gráficos para o sensor s7 e através dela podemos constatar os picos nos valores de autocorrelação nos *lags* que correspondem as variações diárias e intradiádias. Um dia de monitoramento, considerando resolução de 10 minutos, corresponde a 144 observações, e é essa medida que foi considerada no gráfico como lag de referência, assim no eixo x do gráfico os valores 1, 2, 3 e 4 representam a quantidade de dias observados.

Figura 21 - Gráfico da função de autocorrelação e autocorrelação parcial referente ao sensor s7.



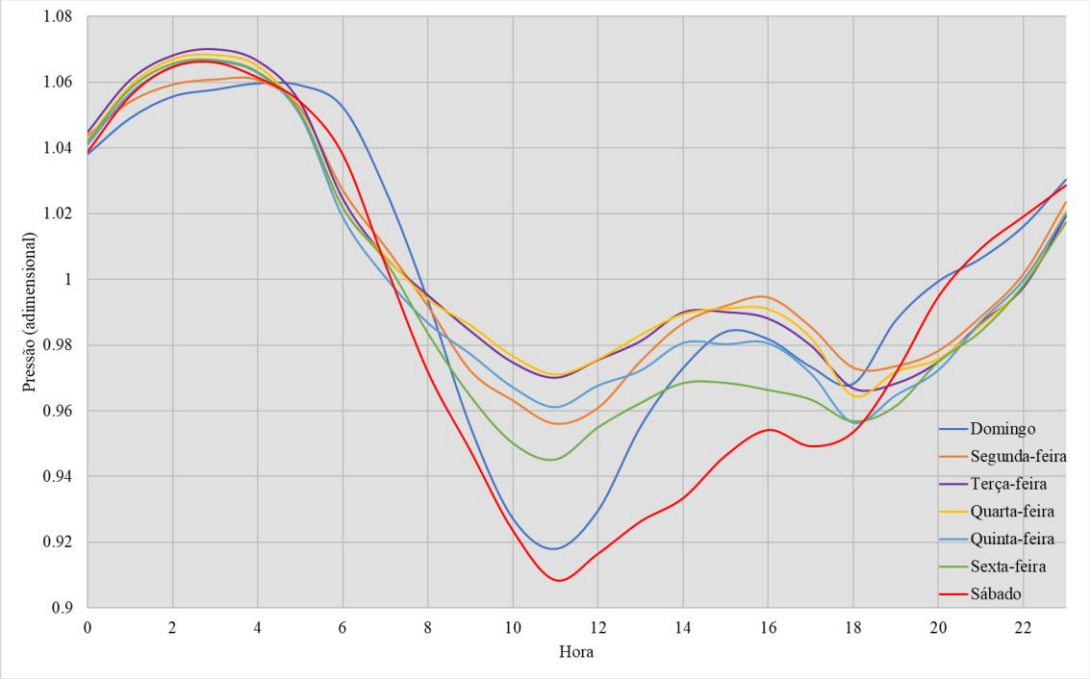
Segundo Hyndman & Athanasopoulos (2021) quando os dados são sazonais, as autocorrelações serão maiores para as defasagens sazonais (e em seus múltiplos) do que em outras defasagens. Bem como, segundo Chatfield (1996), se a série possuir sazonalidade, o correlograma exibirá uma oscilação na mesma frequência. Ao observar o autocorrelograma da série (Figura 21) percebe-se que há picos nos *lags* diários (múltiplos de 144), indicando assim a sazonalidade diária que há na série, proveniente do padrão de consumo de água. Corroboram com estas afirmações os resultados positivos obtidos para todos os sensores analisados, de acordo com os testes de Ollench-Webel (WEBEL e OLLECH, 2019), descritos na seção 3.3a) Tendência, sazonalidade e estacionariedade.

Sobre os padrões de consumo de água característicos de sistemas de abastecimento onde predominam consumidores da categoria residencial, com suas respectivas variações horárias, diárias e semanais, discutidos na seção 2.2.1 Perfil de variação de consumo de água ficou evidente seus reflexos nas séries originais, observadas na Figura 14. Como também caracterizou

a sazonalidade determinística da série, evidenciada pelo gráfico da Figura 21 e pelos testes de sazonalidade.

Estas primeiras constatações podem ser percebidas de forma mais clara para os ciclos diários de consumo, em contrapartida as variações que ocorrem em função do dia semana não são tão evidentes em alguns sensores por conta de sua média, e amplitude de medição. Neste sentido o gráfico apresentado na Figura 22 traz o detalhamento da variação de pressão na rede de abastecimento em função do horário do dia combinado com o dia da semana, de forma a elucidar as diferenças que ocorrem também neste ciclo, que foi considerado nas avaliações posteriores.

Figura 22 - Variação da pressão na rede de abastecimento em função da hora do dia e do dia da semana de forma adimensional referente ao sensor s7.

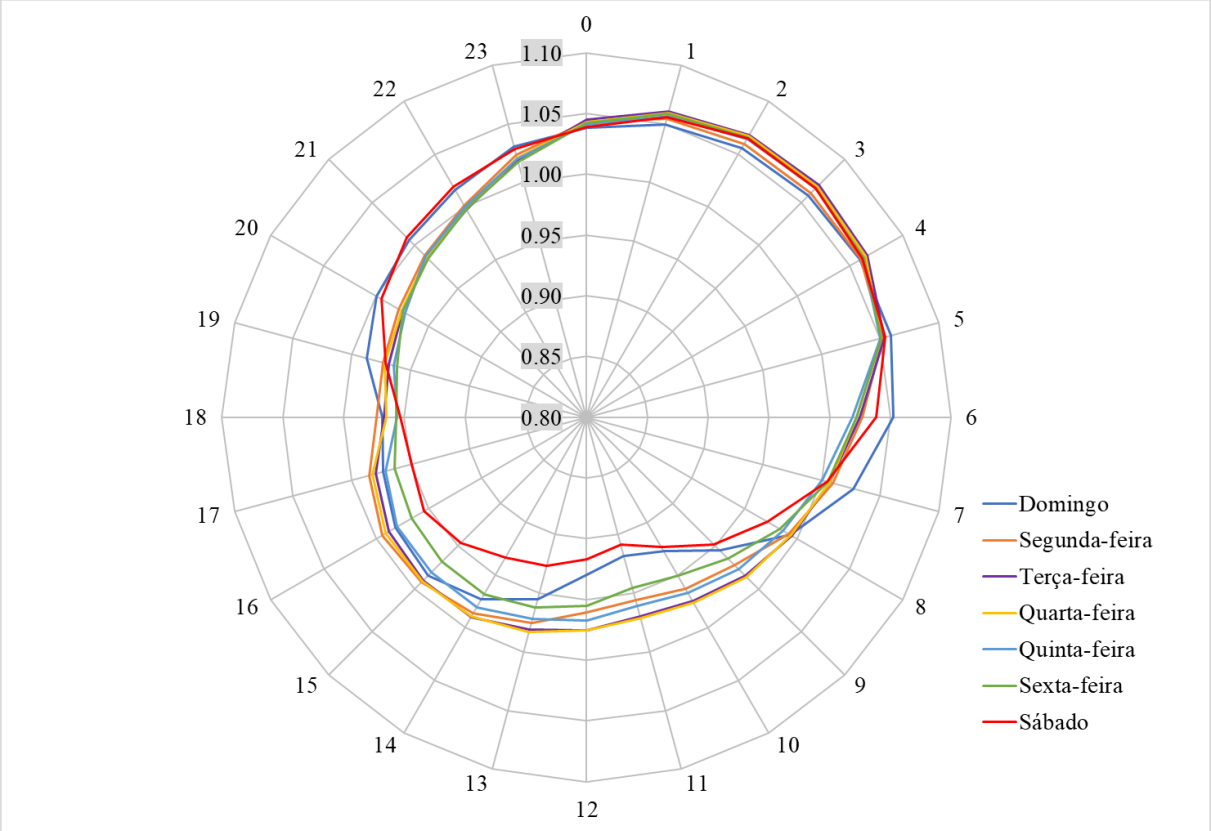


As curvas referentes aos dias da semana de segunda a quinta-feira possuem comportamento muito parecido, na sexta-feira já é possível perceber uma média menor que as demais principalmente nos horários entre 10:00 e 12:00 horas e às 18:00 horas. O sábado é o dia da semana em que as pressões atingem seus menores valores médios durante todo o período do dia, entre 8:00h e 18:00h. O domingo também possui uma média menor que os outros dias da semana, que não sábado, no intervalo das 10:00h às 12:00h, porém sua média sobe bastante entre as 13:00h e 16:00h, demonstrando uma maior volatilidade ao longo do dia, e mesmo não

sendo o dia com valores extremos, este padrão de comportamento pode se refletir em um desafio para a modelagem.

Com observação dos dados plotados em eixo radial, é possível perceber menor variação nos dados durante a noite, período entre às 0:00h e 5:00h, resultados que acompanham a média dos maiores valores de pressão. Durante o dia, principalmente entre 6:00h e 18:00h a variação da pressão é muito maior, com características diferentes inclusive entre os dias da semana. Neste período também, ocorre as menores medições de pressão, como pode ser observado a partir do gráfico da Figura 23.

Figura 23 - Variação da pressão na rede de abastecimento de água em função do dia da semana e do horário do dia plotado em eixo radial.



4.5 Ajuste dos modelos e Análise de Resíduos

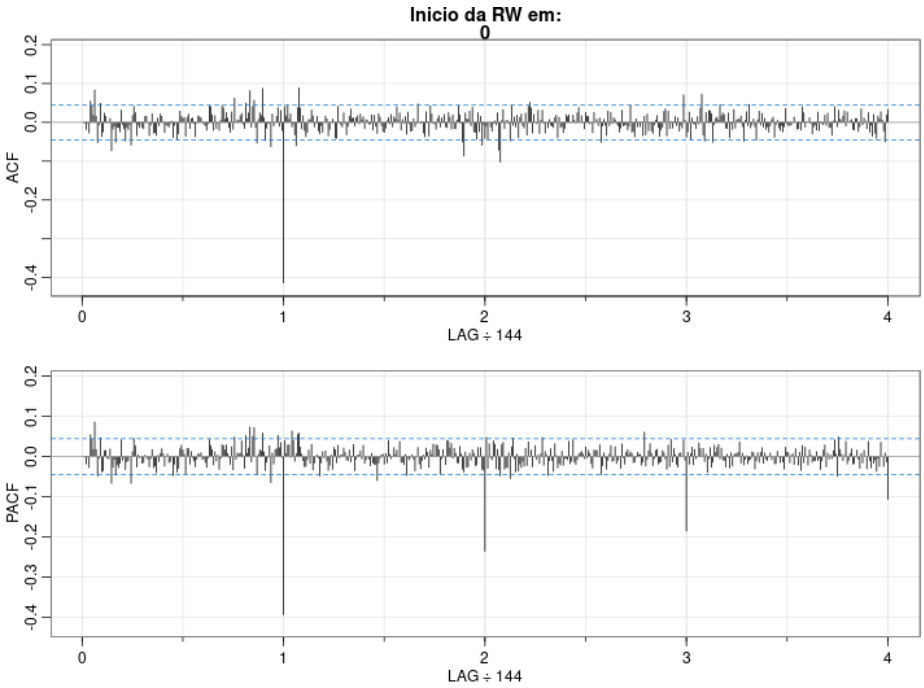
O trecho da série, mencionado na seção 3.5, serviu como ponto de partida para o ajuste inicial de cada modelo. Primeiramente, procedeu-se ao ajuste do modelo SARIMA, no qual foram identificadas algumas limitações. Como resposta a essas limitações, avançou-se na elaboração do modelo ARIMA+Harmônico. Adicionalmente, um esforço substancial foi empregado na formulação de uma abordagem de modelagem GARMA, também aplicada ao segmento da série descrito na seção 3.5, com a ressalva de que a transformação descrita na seção 0 foi aplicada. Somente após a conclusão desta fase inicial de modelagem, procedeu-se à expansão da análise para outros segmentos da série, com o intuito de selecionar a combinação mais apropriada entre o conjunto de ajustes e o horizonte de previsão.

4.2.1 Ajuste do modelo SARIMA

Os modelos SARIMA possuem formato $(p,d,q)\ (P,D,Q)m$ como especificado na Figura 3. Sendo este formato referente a ordem do modelo, e para cada ordem, estima-se um parâmetro, de forma a se obter a equação do modelo ajustado. Sobre o ajuste é realizado também a análise de resíduos, de modo a verificar se as suposições sobre os resíduos são verdadeiras.

A execução do algoritmo “autoarima” descrito na seção 2.3.4 selecionou um modelo *(1,0,3)(0,1,0)[144]*, com AIC igual a 4494,42 para a primeira janela móvel de análise, aquela sobre a qual o ajuste foi realizado. Porém, a análise de autocorrelação da série feita através do gráfico da função de autocorrelação e autocorrelação parcial demonstra que o modelo ajustado não foi capaz de capturar toda a informação sobre a série, principalmente quanto as componentes sazonais. É possível observar através da Figura 24 um pico no primeiro *lag* sazonal do gráfico ACF e um decrescimento exponencial nos *lags* sazonais no gráfico da PACF. Segundo Morettin e Toloi (2006), esse comportamento refere-se a um modelo de médias móveis, e os autores sugerem a adição de um termo MA na parte sazonal do modelo.

Figura 24 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo SARIMA ajustado para a série da primeira janela móvel.



Não havendo a priori forma de parametrizar a adição de um termo ao modelo através do algoritmo automático de identificação e estimação, neste ponto discute-se sobre a operacionalidade de manualmente realizar este procedimento. Por um lado, pode-se seguir sem alterações e assumir que o modelo não cumpre com as premissas, no tocante a autocorrelação. Por outro lado, pode-se adicionar manualmente o termo de médias móveis na parte sazonal do modelo SARIMA e reestimar os parâmetros para eliminar a autocorrelação, sendo esta segunda opção a estratégia adotada.

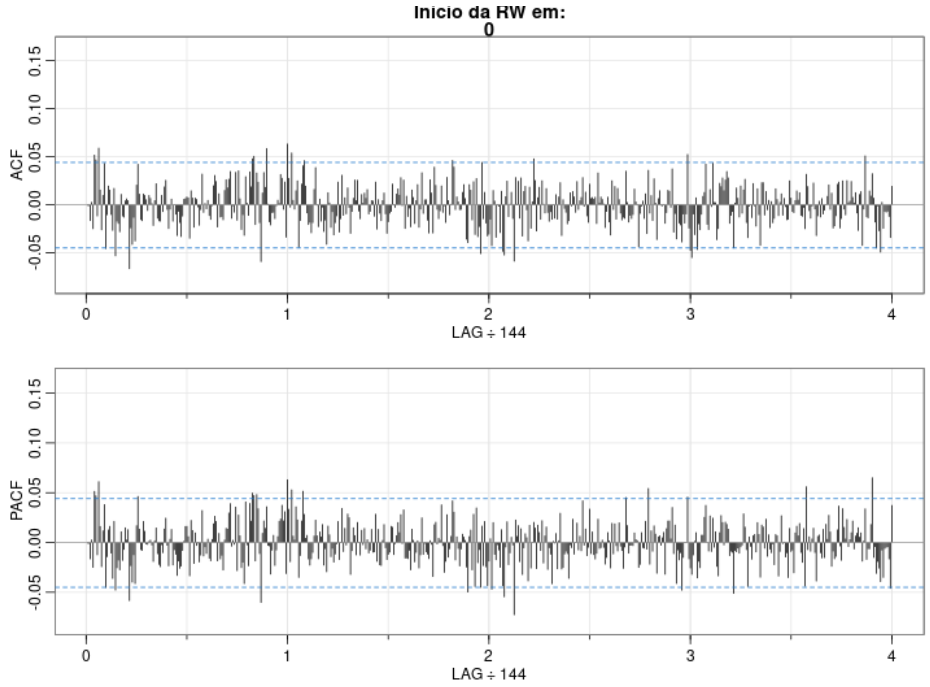
Com a realização da estimação dos parâmetros para o modelo SARIMA, obteve-se um modelo $(1,0,3)(0,1,1)[144]$ com valor de AIC igual a 3698.93, o que já representa por si só uma melhoria no modelo. O processo de estimação retornou os valores dos seguintes parâmetros para o modelo SARIMA modificado, presentes na Tabela 3, seguidos de seus respectivos valores de desvio padrão, indicando que todos são significativos.

Tabela 3 - Parâmetros estimados para o modelo estatístico SARIMA modificado.

Parâmetros	Valor	Erro Padrão
AR1	0.9530	0.0091
MA1	-0.6021	0.0249
MA2	-0.0078	0.0267
MA3	-0.0265	0.0230
MA1 - Sazonal	-0.9576	0.0764
Diferença Sazonal	1	-

A condição de não correlação entre os valores dos resíduos é verificada novamente pelo gráfico das funções de autocorrelação e autocorrelação parcial, apresentados na Figura 25. Dessa vez, é possível perceber que não há mais picos significativos nos *lags* sazonais, apenas alguns pontos aleatórios fora do intervalo de confiança, mas que na prática não possuem significado.

Figura 25 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo SARIMA ajustado para a série da primeira janela de análise.

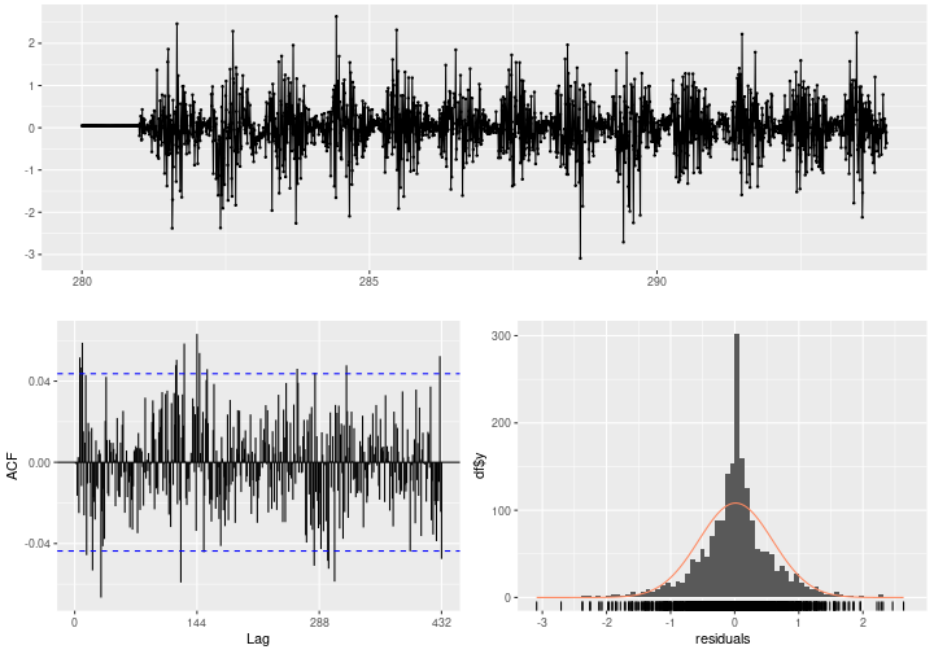


Através da análise gráfica da Figura 26 é possível verificar o comportamento dos resíduos do modelo SARIMA, ajustado para a base de treino selecionada. No gráfico superior pode-se visualizar os valores dos resíduos ao longo do tempo, e sua média igual a zero, porém também fica evidente que a variância não é constante, os resíduos estão seguindo um padrão,

na mesma escala do período sazonal diário da série. Este é um caso claro do comportamento de uma série heterocedástica, onde tem-se a variância crescendo com a diminuição da média dos dados.

Fazendo um paralelo com o mecanismo gerador, durante a noite, quando a média da pressão é alta na rede de distribuição de água, a variância é baixa, pois a vazão de consumo é mínima. Por outro lado, durante o dia há diversas variações instantâneas no consumo, descritas na seção 2.2.1, e isso faz com que haja maior variabilidade nos dados. Como o modelo SARIMA possui como premissa a modelagem de séries homocedásticas, ele não é capaz de ajustar parâmetros que eliminem essa característica da série, refletindo-se nos resíduos.

Figura 26 - Análise de resíduos do modelo ajustado SARIMA para a base de treino da JM#1.



O segundo gráfico presente na Figura 26, no canto inferior esquerdo é novamente a função de autocorrelação (*ACF*). E no canto inferior direito está presente a distribuição dos resíduos normalizados, na forma de histograma, e com uma função comparativa, há também a curva (em laranja) que representa a distribuição normal esperada para os resíduos. Mas, como pode ser visto, os resíduos não seguem a distribuição normal, justamente por conta de a variância não ser constante ao longo do tempo.

Como forma de validar o ajuste realizado nas 5 janelas móveis, além dos valores de AIC, são analisados os valores obtidos pelas métricas RMSE e MAPE. Os valores foram muito próximos para todas as JM, sendo que os melhores resultados para as duas métricas foram para a janela 1 (JM#1), com 0.57 mca e 0.89% para os valores de RMSE e MAPE respectivamente. O maior MAPE encontrado foi para a JM#4, com valor de 0.92%. A medida AIC variou bastante entre as janelas móveis, sendo seu maior valor para a JM#5, com 3750.58. Um resumo dos valores pode ser encontrado na Tabela 4.

Tabela 4 - Resumo das métricas de ajuste do modelo SARIMA modificado para cada uma das janelas móveis analisadas.

Janela Móvel	RMSE(mca)	MAPE(%)	AIC
#1	0.57	0.89	3698.94
#2	0.58	0.90	3726.38
#3	0.58	0.91	3700.67
#4	0.58	0.92	3725.05
#5	0.58	0.91	3750.58

4.2.2 Ajuste do modelo ARIMA+Harmônico

O ajuste do modelo ARIMA+Harmônico segue os mesmos princípios básicos apresentados anteriormente, principalmente quanto aos métodos de identificação e estimação dos parâmetros, e as suposições sobre os resíduos do modelo. A definição sobre as variáveis exógenas, referentes a série de Fourier com pares senos e cossenos para modelagem da sazonalidade determinística da série, consta na seção 3.4.2 “Ajuste do modelo ARIMA + Harmônico”. A estratégia adotada na estimação deste modelo foi de manter a ordem do ARIMA identificado para a primeira janela de análise, mas com reestimação dos parâmetros e dos termos senos e cossenos.

O modelo encontrado possui ordem ARIMA (2,0,4) para os erros, tendo como variáveis exógenas 6 termos senos e cossenos, onde destes, formam-se 2 pares para a sazonalidade diária (frequência 144) e 1 par para a sazonalidade semanal (frequência 1008). Os valores dos parâmetros ajustados podem ser visualizados a partir da Tabela 5.

Tabela 5 - Parâmetros estimados para o modelo ARIMA+Harmônico, termos senos e cossenos e seus respectivos erros padrões.

Parâmetros	Valor	Erro Padrão
AR1	0.8789	0.0172
AR2	-0.5698	0.0298
MA1	1.3789	0.062
MA2	1.8328	0.0617
MA3	0.7668	0.0547
MA4	-0.3114	0.0545
Sen1	-0.1376	0.0471
Cos1	0.2116	0.0469
Sen2	0.068	0.0407
Cos2	-0.09	0.0406
Sen3	-0.0046	0.0357
Cos3	0.0315	0.0356

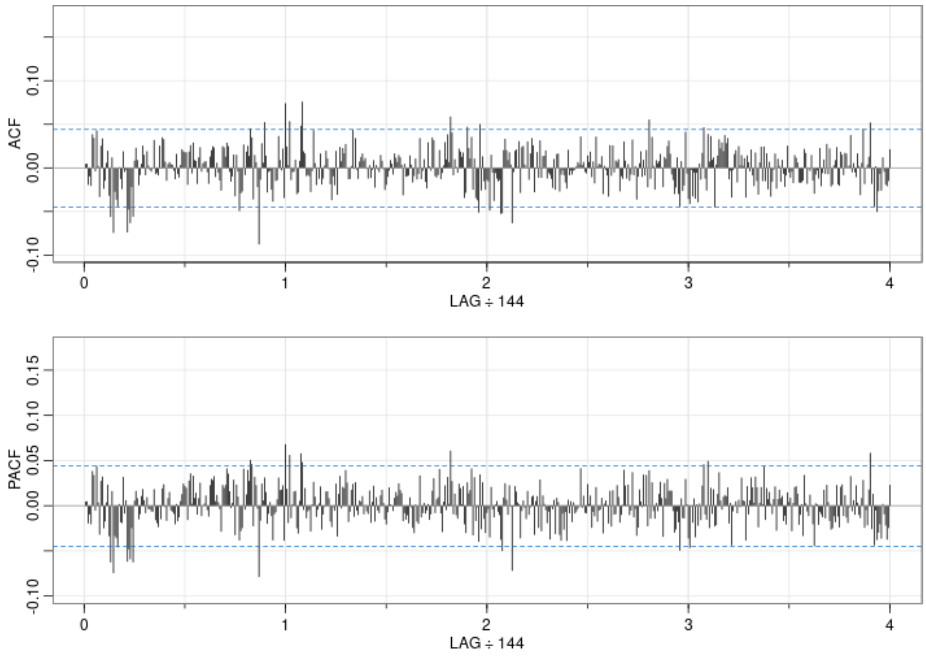
Os modelos ajustados para as demais janelas móveis reestimaram os parâmetros p e q, e calcularam novos valores de senos e cossenos. Semelhantemente ao apresentado para o modelo SARIMA, um resumo das métricas estão presentes na Tabela 6. Pode-se verificar que os valores tanto de RMSE quanto de MAPE para os respectivos períodos de treino de cada uma das janelas móveis foi o mesmo. Já o valor do AIC, variou entre as análises, sendo o maior para a JM#5 com valor de 3762.75 e o menor para a JM#3 com valor de 3724.51.

Tabela 6 - Resumo das métricas de ajuste do modelo ARIMA+Harmônico para cada uma das janelas móveis analisadas.

Janela Móvel	RMSE(mca)	MAPE(%)	AIC
#1	0.61	0.98	3750.83
#2	0.61	0.98	3752.81
#3	0.60	0.98	3724.51
#4	0.61	0.99	3747.21
#5	0.61	0.99	3762.75

O modelo ajustado possui valor de AIC igual 3750.83, para a JM#1. Após seu ajuste, faz-se a análise de resíduos. Logo, a primeira condição a ser verificada é a de não correlação. Através das funções *ACF* e *PACF* presentes no gráfico da Figura 27 pode-se observar que não há picos significativos em nenhum dos lags sazonais, apenas alguns picos em posições aleatórias e com valores pequenos que não possuem um significado. Concluindo assim que os resíduos não são autocorrelacionados.

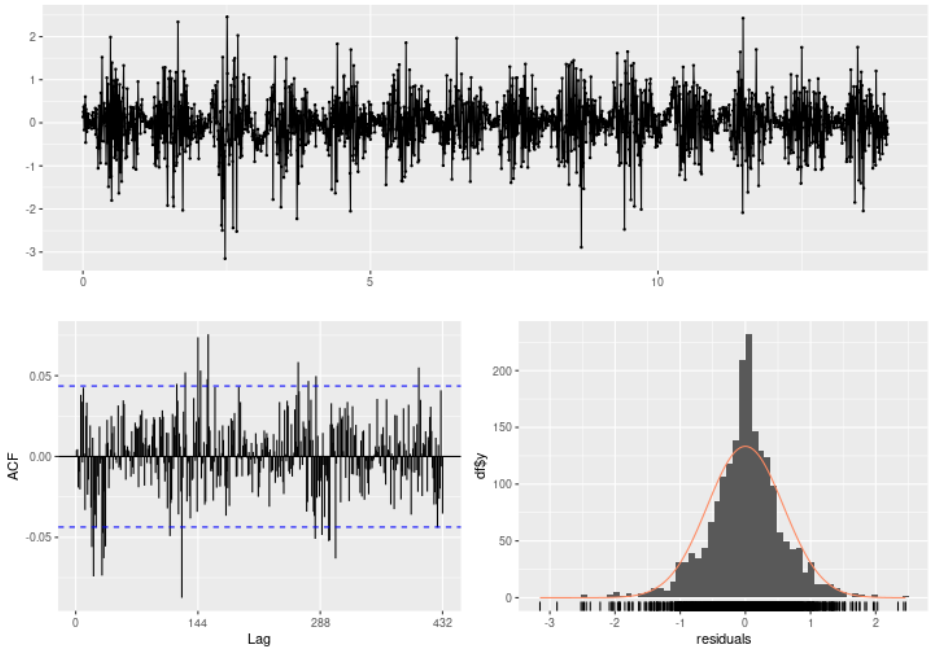
Figura 27 - Gráfico das funções de autocorrelação (*ACF*) e autocorrelação parcial (*PACF*) do modelo ARIMA+Harmônico ajustado para a série da primeira janela móvel.



A análise de resíduos para verificação do cumprimento das suposições do modelo segue com verificação das condições de média igual a zero, variância constante e distribuição normal. Neste sentido os gráficos presentes na Figura 28 auxiliam na análise. No gráfico do canto inferior esquerdo é apresentado novamente o correlograma. E no gráfico do canto inferior direito pode-se verificar a distribuição dos resíduos normalizados. Que por conta da heterocedasticidade não seguem distribuição normal, de acordo com a curva de referência laranja presente. Esgotadas as possibilidades de alteração neste modelo para adequação às suposições sobre os resíduos, segue-se para a etapa de previsão.

Sobre as premissas de média zero, variância constante e distribuição normal, como nos resultados apresentados para os dois modelos anteriores, o modelo ajustado não foi capaz de extrair toda a informação da série temporal. Obedecendo tão somente a suposição de média igual a zero e desobedecendo as demais, como pode ser visualizado na Figura 28. Onde percebe-se, mais uma vez, a variância não sendo constante ao longo do tempo e a distribuição dos resíduos não estando de acordo com a curva de referência da distribuição normal.

Figura 28 - Análise de resíduos do modelo ajustado ARIMA+Harmônico para a base de treino da JM#1.



4.2.3 Ajuste do modelo GARMA

De posse da série transformada, o processo de aplicação do modelo GARMA se caracteriza com a definição da função de densidade de probabilidade gama, o uso das variáveis exógenas composto pelas *dummies* de finais de semana e pelos 6 termos senos e cossenos. A avaliação das diferentes combinações de ordem dos parâmetros ARMA, em conjunto com análise de resíduos, apontaram para o uso de um modelo GARMA(2,0). Assim, as equações do modelo consolidaram-se como se apresenta abaixo Eq. 13,

$$\ln(\mu_t) = \eta_t = \underline{\mathbf{x}_t'}\boldsymbol{\beta} = \beta_0 + \beta_1 \cos\left(\frac{2\pi t}{144}\right) + \beta_2 \operatorname{sen}\left(\frac{2\pi t}{144}\right) + \beta_3 \cos\left(\frac{4\pi t}{144}\right) + \beta_4 \operatorname{sen}\left(\frac{4\pi t}{144}\right) + \beta_5 \cos\left(\frac{2\pi t}{1008}\right) + \beta_6 \operatorname{sen}\left(\frac{2\pi t}{1008}\right) + \tau_t$$

Eq. 13

O erro τ_t é modelado por processo AR, conforme Eq. 14,

$$\tau_t = \sum_{j=1}^{2} \phi_j \left\{ g\left(y_{t-j}\right) - \underline{\boldsymbol{x}_{t-j}'}\boldsymbol{\beta} \right\}$$

Eq. 14

A Figura 29 apresenta os valores observados de Energia Consumida (mca) no período de 08 a 22 de Outubro de 2015, primeira Janela Móvel de análise, e a série ajustada pelo modelo GARMA(2,0) com distribuição gama e processo autorregressivo como em Eq. 13 e Eq. 14. A análise de resíduos apontou que o modelo satisfaz os pressupostos, pode-se notar pela Figura 30, que apresenta os resíduos ao longo do tempo, que não há mais padrão de variação como era percebido na aplicação dos modelos anteriores, podendo assumir que os resíduos são homocedásticos.

Figura 29 - Energia Consumida (mca) e série ajustada com modelo GARMA(2,0) com distribuição gama.

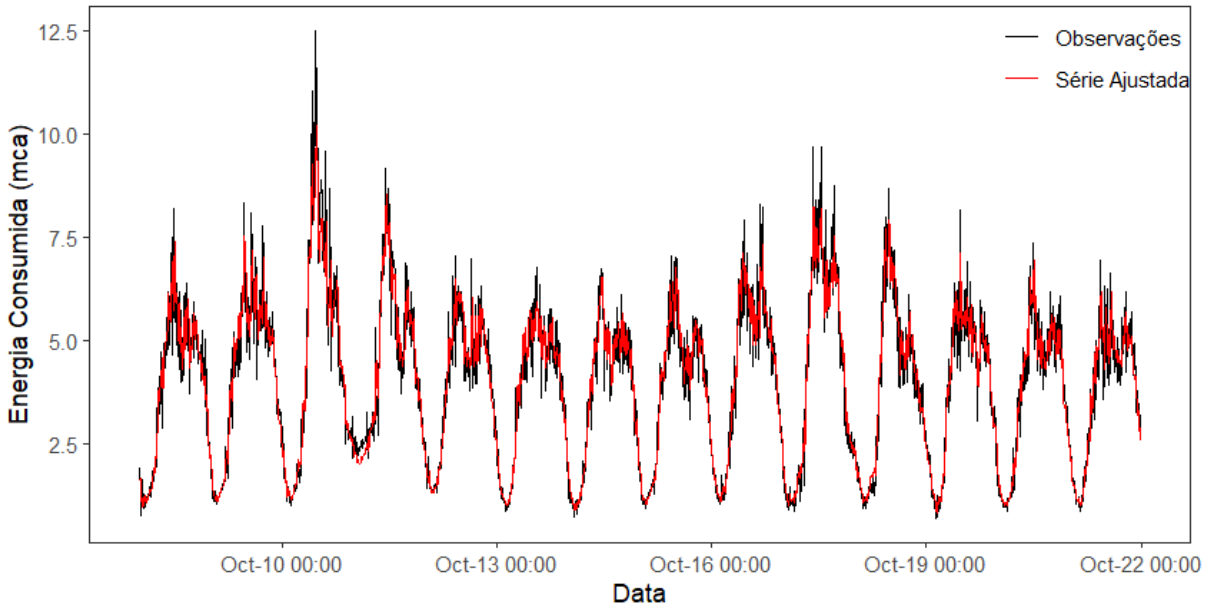
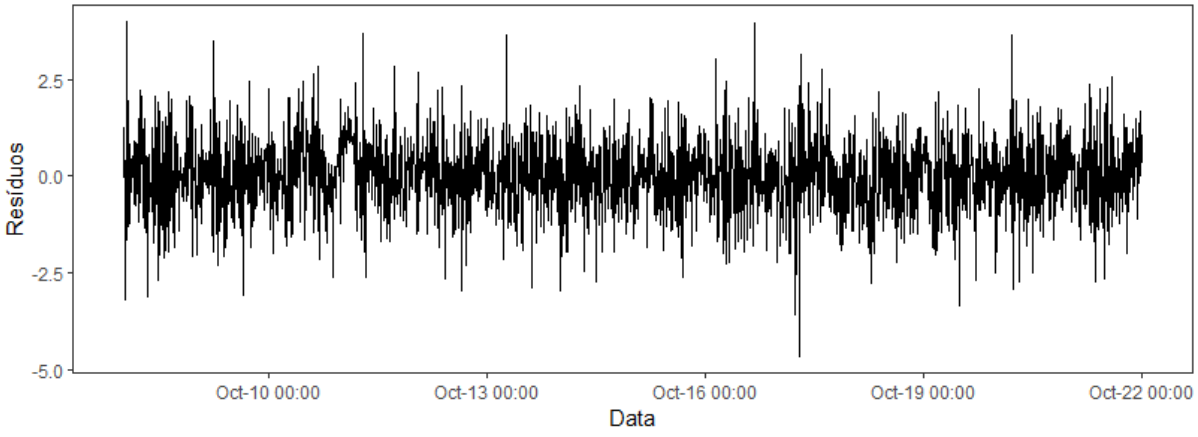
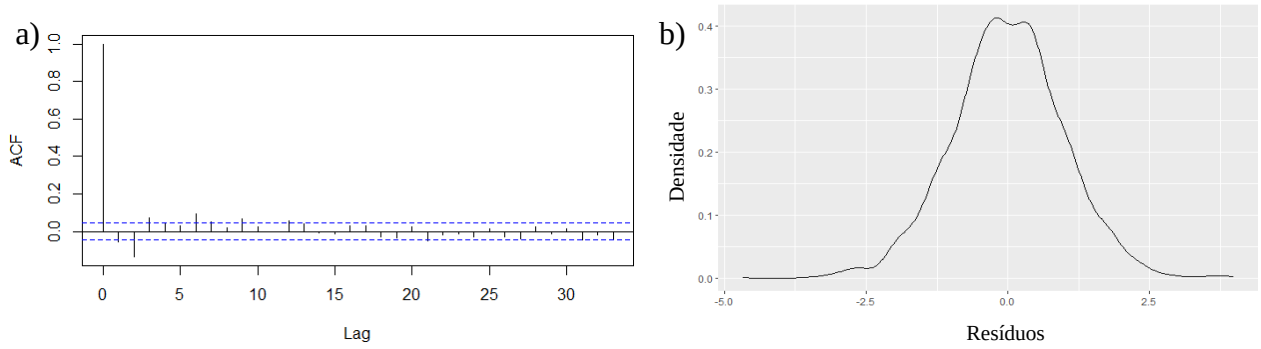


Figura 30 - Resíduos do modelo GARMA(2,0) com distribuição gama ao longo do tempo.



A apresentam respectivamente o gráfico da função de autocorrelação e a densidade de distribuição dos resíduos no formato de uma curva normal. Os resultados apontam para consideração dos resíduos como independentes, sem autocorrelação significativa, e pela curva de densidade aparentam ser normalmente distribuídos.

Figura 31 – Gráficos de análise dos resíduos do modelo GARMA(2,0) com distribuição gama. a) Autocorrelograma e b) Curva densidade de probabilidade dos resíduos.



Os coeficientes estimados são apresentados na Tabela 7. Os dois parâmetros autorregressivos (ϕ) são significativos (*p-valor* < 0.001), e com valores próximos a 0.44 e 0.34 demonstram a grande influência dos valores observados passados na predição. Assim como todos os termos harmônicos, sendo o de menor significância β_5 com *p-valor* 0.086. Por outro lado, as variáveis *dummies* não foram consideradas significativas no modelo, sendo a *dummy* de Sábado com valor -0.0167 e erro padrão 0.0448, e a *dummy* de Domingo com valor 0.0036 e erro padrão de 0.04479. Ficando ambas de fora do modelo final. O valor de AIC do modelo foi de 2789.45.

Tabela 7 - Coeficientes estimados, erro padrão e p-valor do modelo GARMA(2,0).

Coeficiente		**Valor**	**Erro Padrão**	**p-valor**
beta.(Intercept)	β_0	1.292	0.018	<0.001
beta.cos1	β_1	-0.543	0.019	<0.001
beta.sin1	β_2	-0.455	0.018	<0.001
beta.cos2	β_3	0.052	0.017	0.002
beta.sin2	β_4	-0.324	0.017	<0.001
beta.cos3	β_5	-0.045	0.026	0.086
beta.sin3	β_6	0.063	0.021	0.003
phi1	ϕ_1	0.437	0.021	<0.001
phi2	ϕ_2	0.335	0.021	<0.001
sigma	σ	0.137	0.002	<0.001

As métricas para as 5 janelas móveis de análise, conforme apresentado nos modelos anteriores, segue o que se apresenta na Tabela 8. Com atenção ao valor do MAPE, que foi calculado sobre os valores da série transformada, e pela diferença de escala apresenta resultados relativamente superiores aos do modelos anteriores, porém sem prejuízo para continuidade das análises. Nota-se também, menor valor de AIC em todas as janelas de análise, considerando assim que os modelos podem ser considerados menos complexos.

Tabela 8 - Resumo das métricas de ajuste do modelo GARMA(2,0) para cada uma das janelas móveis analisadas

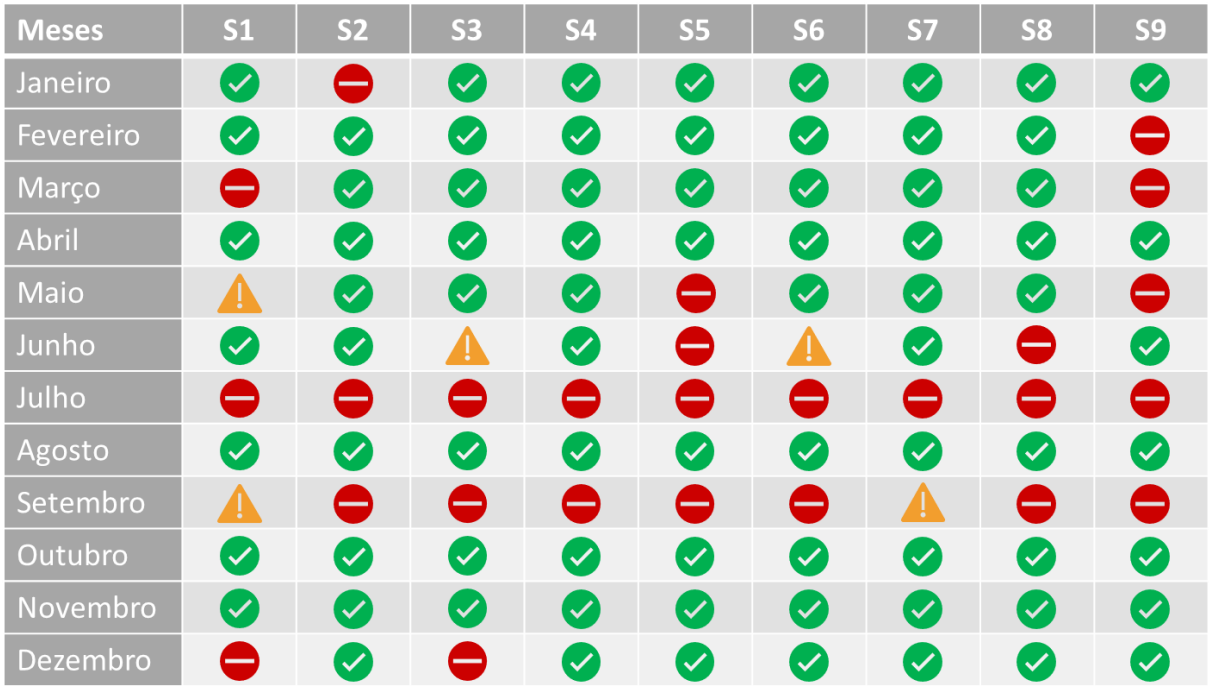
Janela Móvel	RMSE(mca)	MAPE(%)	AIC
#1	0.61	11.08	2789.45
#2	0.61	10.97	2749.79
#3	0.61	10.89	2732.78
#4	0.61	11.01	2769.08
#5	0.61	11.05	2723.36

4.3 Seleção da combinação entre conjunto de dados de ajuste e horizonte de previsão

Como descrito na seção 3.4.3.3 Seleção do conjunto de ajuste e do horizonte de previsão do modelo GARMA, para cada par sensor-mês, foram realizadas 4 previsões, referentes às combinações de conjunto de dados de ajuste e horizonte de previsão. São elas, a previsão de 24 horas à frente com modelo completo e mensal, e previsão de 48 horas à frente com modelo completo e mensal. As previsões foram realizadas de acordo com a disponibilidade de dados, apresentada na Figura 32, onde a cor verde representa que os dados estão disponíveis, o alerta laranja representa dados disponíveis com pequenas falhas (limite 15%), e o símbolo de cor vermelha representa a não disponibilidade de dados.

Após a exclusão dos meses nos quais a modelagem não pôde ser conduzida devido à escassez de dados, obtivemos um total de 82 combinações de sensor-mês. Para cada sensor, foram realizadas as quatro previsões mencionadas anteriormente e avaliadas com o uso de duas métricas, a saber, RMSE (raiz do erro quadrático médio) e MAPE (erro percentual absoluto médio). Isso resultou em um conjunto de 656 valores que foram analisados e seus resultados sintetizados são apresentados a partir deste ponto.

Figura 32 - Diagrama de disponibilidade de dados.



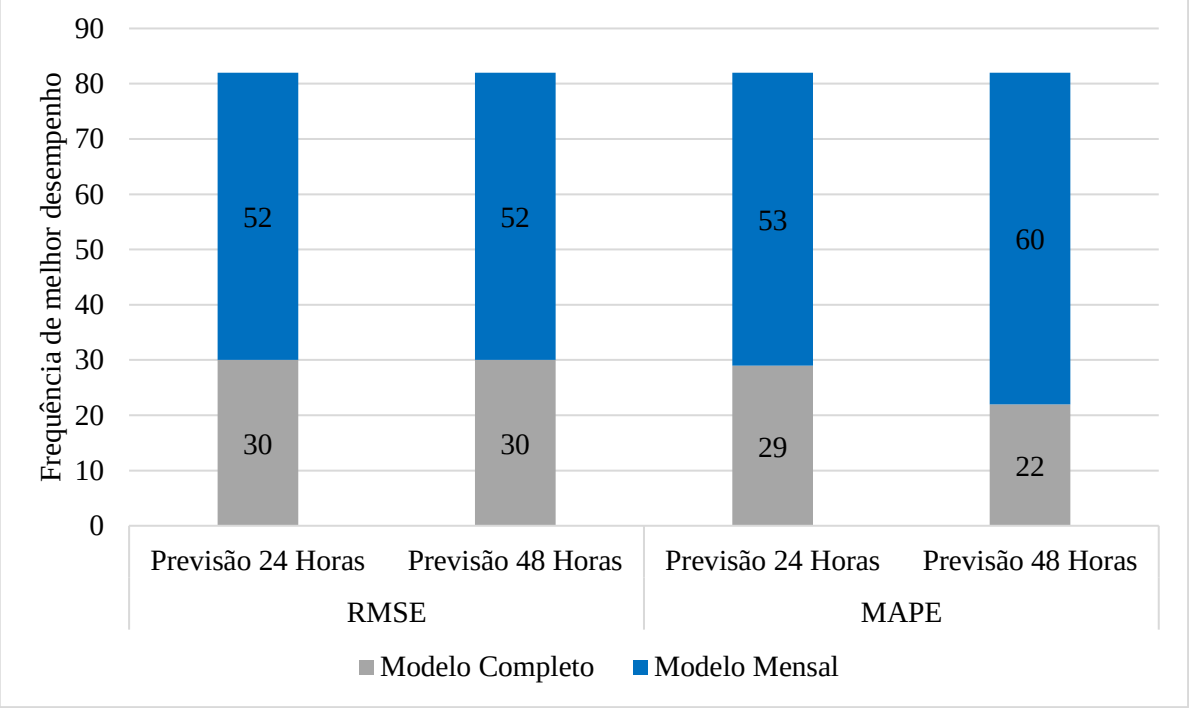
4.3.1 Avaliação do desempenho entre modelo completo e modelo mensal

Para avaliar se o melhor conjunto de dados de ajuste para previsão era o modelo completo ou modelo mensal, somou-se a quantidade de cada um deles que teve melhor desempenho quanto às métricas avaliadas em cada horizonte de previsão, conforme Figura 33. Para a métrica RMSE, independente do horizonte de previsão, em 63% das 82 combinações sensor-mês, o modelo mensal teve melhor desempenho. Para a métrica MAPE, no horizonte de previsão de 24 horas, o modelo mensal foi melhor em 53 dos 82 cenários (65%), e para previsão de 48 horas, foi melhor em 60 de 82 (73%). Por essa análise de frequência, pode-se concluir que em média o uso do modelo mensal obteve melhor desempenho em 63% dos casos para métrica RMSE e 69% dos casos para MAPE.

Além da análise da frequência, uma análise quantitativa sugere que o valor da métrica RMSE foi em média 24% maior para o modelo completo se comparada aos resultados alcançados pelo modelo mensal na previsão de 24 horas, e 29% maior na previsão de 48 horas. Para a métrica MAPE, os valores foram em média 37% maiores para o modelo completo em relação ao modelo mensal no horizonte de 24 horas, e 43% maior na previsão de 48 horas. Estes resultados auxiliam no entendimento de que em geral as métricas alcançam melhores resultados para os modelos mensais, em comparação aos completos. E que quando comparados o quão

pior é o resultado ao utilizar o modelo completo ao invés do modelo mensal, tem-se que para a métrica RMSE o modelo completo tem desempenho médio 26.5% pior que o modelo mensal, enquanto para a métrica MAPE o desempenho é 40% pior.

Figura 33 - Frequência de melhor desempenho de cada modelo em função da métrica e do horizonte de previsão.



Avaliando com maiores detalhes os resultados dos sensores, vemos que apenas para os sensores s1 e s7 a utilização dos modelos completos foi melhor na maioria das avaliações, para a métrica RMSE, tanto para o horizonte de 24 e 48 horas. Por outro lado, para a métrica MAPE, o modelo completo teve melhor desempenho apenas para o sensor s1 no horizonte de 24 horas, e para o s1 e s3 no horizonte de 48 horas. Vale ressaltar que os sensores s1 e s7 estão entre os de menores taxas de dados faltantes, 28% e 22% respectivamente, e apresentam tendência estacionária ao longo do ano.

A variação dos resultados pode ser analisada ainda quanto ao período do ano, então em uma análise mês a mês, maio foi o único mês em que, na média, tanto os valores de RMSE quanto de MAPE referente às previsões com modelo completo foram melhores que o modelo mensal. Uma possível causa para isso seria que o modelo completo tivesse capacidade de refletir melhor uma característica média nos valores, mais distante dos extremos, e o mês de maio pode se caracterizar por possuir temperaturas amenas, e não faz parte de temporada de

férias, podendo ter comportamento melhor representado pelo modelo completo em função disso. Os resultados demonstram que, em geral, o mês do ano em que se está avaliando não é grande limitante na combinação conjunto de ajuste e horizonte de previsão.

4.3.2 Avaliação comparativa entre horizontes de previsão

A avaliação da diferença nos valores das métricas, entre o desempenho dos modelos nos horizontes de previsão de 24 e 48 horas resultou que para a métrica RMSE, quanto ao modelo completo que os valores para o horizonte de previsão de 48 horas são em média 9% maiores que para 24 horas. E quanto ao modelo mensal, os valores de RMSE são em média apenas 7% maiores para o horizonte de 48 comparado ao de 24 horas. No tocante à métrica MAPE, para o modelo completo o desempenho dos modelos na acurácia de previsão no horizonte de 48 horas foi 6% pior que no horizonte de 24 horas. Para o modelo mensal, a média foi de 0%.

Sabe-se que quanto mais distante do ponto de origem há tendência de os resultados piorarem, então era esperado um melhor desempenho no horizonte de 24 horas, e os resultados apontam para isso. Porém, não se pode descartar a ideia da utilidade do uso do horizonte de previsão de 48 horas para aplicações específicas. Corrobora com esta afirmação o fato de que os valores para métrica de RMSE encontrados no horizonte de previsão de 48 horas são melhores que modelos ingênuos, conforme estudo prévio (CAMPOS & GAMBOA-MEDINA, 2023) que aplicou a modelagem da mesma série de dados e propôs comparação com previsão feita por médias e por modelo *naive* sazonal.

4.4 Análise das previsões com aplicação da metodologia de janelas móveis

De posse da melhor combinação entre conjuntos de dados de ajuste e horizonte de previsão, modelo mensal com previsão de 24 horas, seguiu-se para aplicação da metodologia de análise de janelas móveis. Como apresentado na seção 3.5, sobre modelagem de dados, cada janela de análise foi dividida entre treino e teste. A base de treino foi sobre a qual o ajuste foi realizado, e a base de teste por sua vez, utilizada para avaliar a qualidade da previsão.

4.4.1 Previsões em janelas móveis com modelo SARIMA

Ao utilizar o modelo SARIMA, obteve-se as métricas RMSE e MAPE para cada uma das janelas de análise. Todos os valores de acurácia calculados foram maiores do que na fase de ajuste do modelo, como esperado, porém, agora não se encontra mais igualdade entre as cada JM. Através da Tabela 9 é possível visualizar que o maior valor de RMSE foi para a JM#3, com

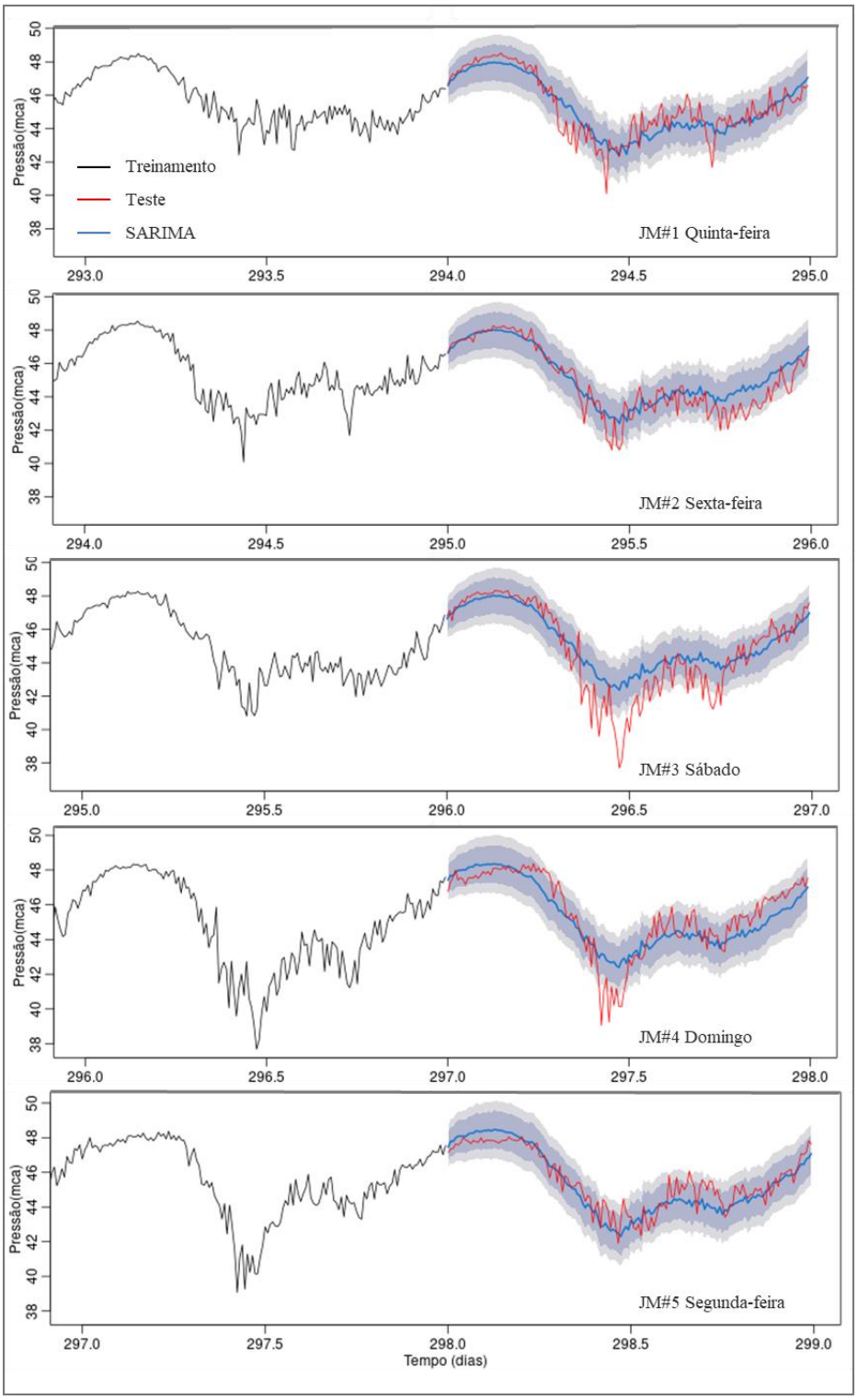
1.32 mca, referente ao sábado, e o segundo maior valor foi de 0.99 para a JM#4, domingo. A pior previsão nesses dias da semana era um resultado também esperado, pois através do estudo do comportamento da série constatou-se que esses dias possuem comportamento particularmente diferente dos demais dias da semana. O menor RMSE é referente à JM#5, segunda-feira, com 0.66. Os resultados da medida MAPE seguiram de forma parecida, sendo que os piores valores também foram para as JM#3 e JM#4, com 2.17% e 1.70% respectivamente. Com menor valor para a JM#5, 1.17%.

Tabela 9 - Resumo das métricas de acurácia de previsão segundo o modelo SARIMA para cada uma das janelas móveis analisadas.

Janela Móvel	RMSE (mca)	MAPE (%)
#1	0.82	1.32
#2	0.80	1.37
#3	1.32	2.17
#4	0.99	1.70
#5	0.66	1.17

Além das métricas apresentadas, é possível analisar os gráficos com os valores observados e preditos com o modelo SARIMA. Através da Figura 34 pode-se visualizar os dados de treinamento (limitados ao último dia) na linha de cor preta, os dados de teste na linha de cor vermelha e os dados previstos com o modelo SARIMA na cor azul, com seu respectivo intervalo de confiança de 80 e 95% nas cores cinza escuro e cinza claro, respectivamente.

Figura 34 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo SARIMA.



4.4.2 Previsões em janelas móveis com modelo ARIMA+Harmônico

O modelo ARIMA+Harmônico, que conta com variáveis preditoras da série de Fourier para modelar a sazonalidade da série com conjunto de termos senos e cossenos, demonstra maior acurácia de previsão do que o modelo SARIMA. Apesar das métricas de ajuste possuírem valores muito próximos, em se tratando da previsão fora da amostra, o modelo ARIMA+Harmônico obteve melhores resultados.

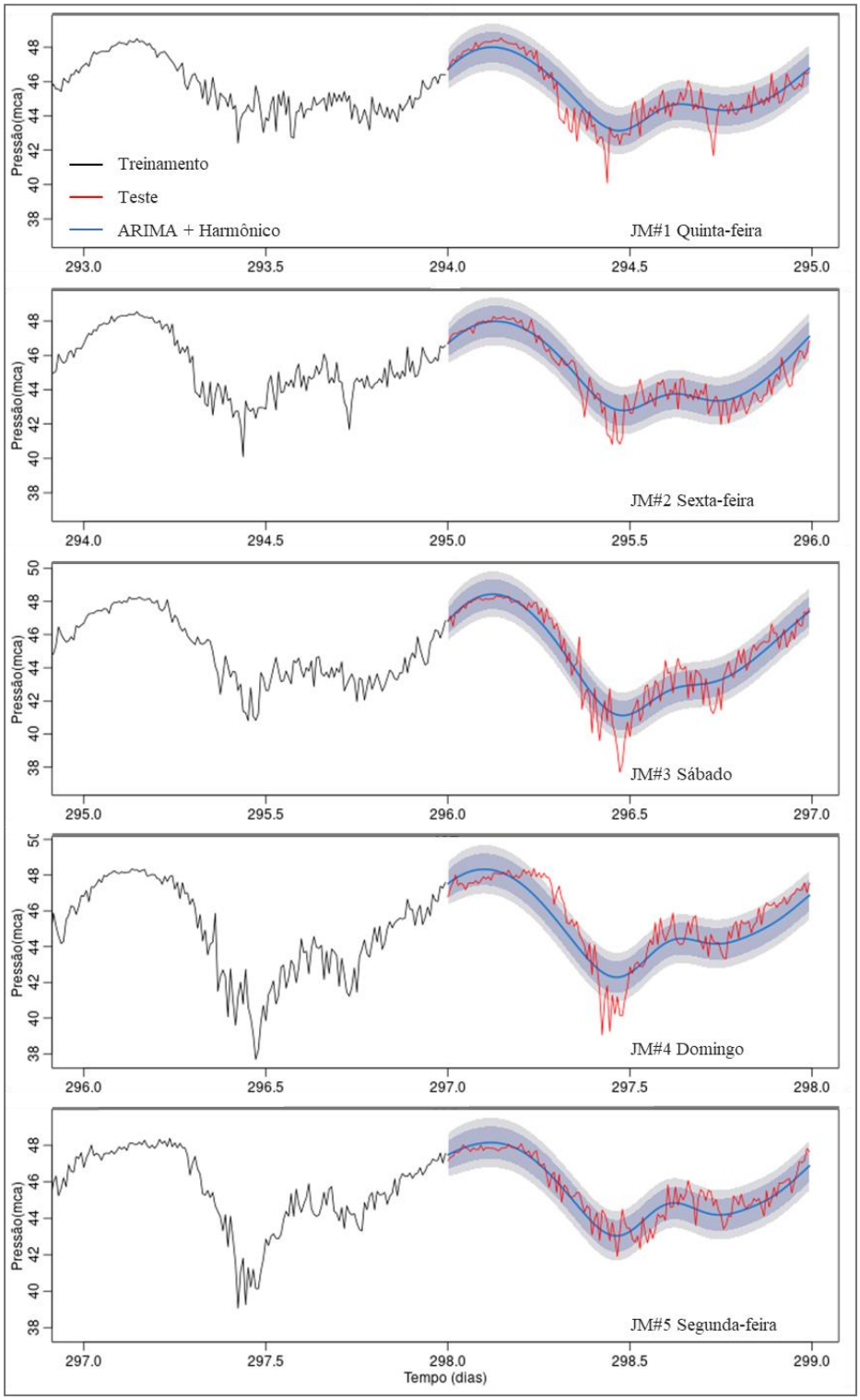
Esta conclusão é possível em função dos valores apresentados pelas métricas RMSE e MAPE quando calculadas comparando os dados previstos pelo modelo com a base de teste. Quanto ao RMSE, todos os valores ficaram abaixo 1, o que demonstra maior exatidão do modelo. O maior valor encontrado foi para a JM#4, com 0.97, que representa o domingo, resultado diferente do encontrado anteriormente em que o sábado era o dia com menor acurácia do modelo SARIMA. De igual modo, o maior valor de MAPE também foi para a JM#4. Os menores valores foram encontrados para a JM#5, com RMSE de 0.59 e MAPE de 1.03%. As métricas para todas as janelas de análise são apresentadas em Tabela 10.

Tabela 10 - Resumo das métricas de acurácia de previsão segundo o modelo ARIMA+Harmônico para cada uma das janelas móveis analisadas.

Janela Móvel	RMSE (mca)	MAPE (%)
#1	0.77	1.19
#2	0.69	1.15
#3	0.90	1.51
#4	0.97	1.71
#5	0.59	1.03

Seguindo o detalhamento da análise proposta anteriormente, através da Figura 35 é possível visualizar os gráficos com as previsões realizadas pelo modelo ARIMA+Harmônico em comparação com a base de teste.

Figura 35 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo ARIMA + Harmônico.



4.4.3 Previsões em janelas móveis com modelo GARMA

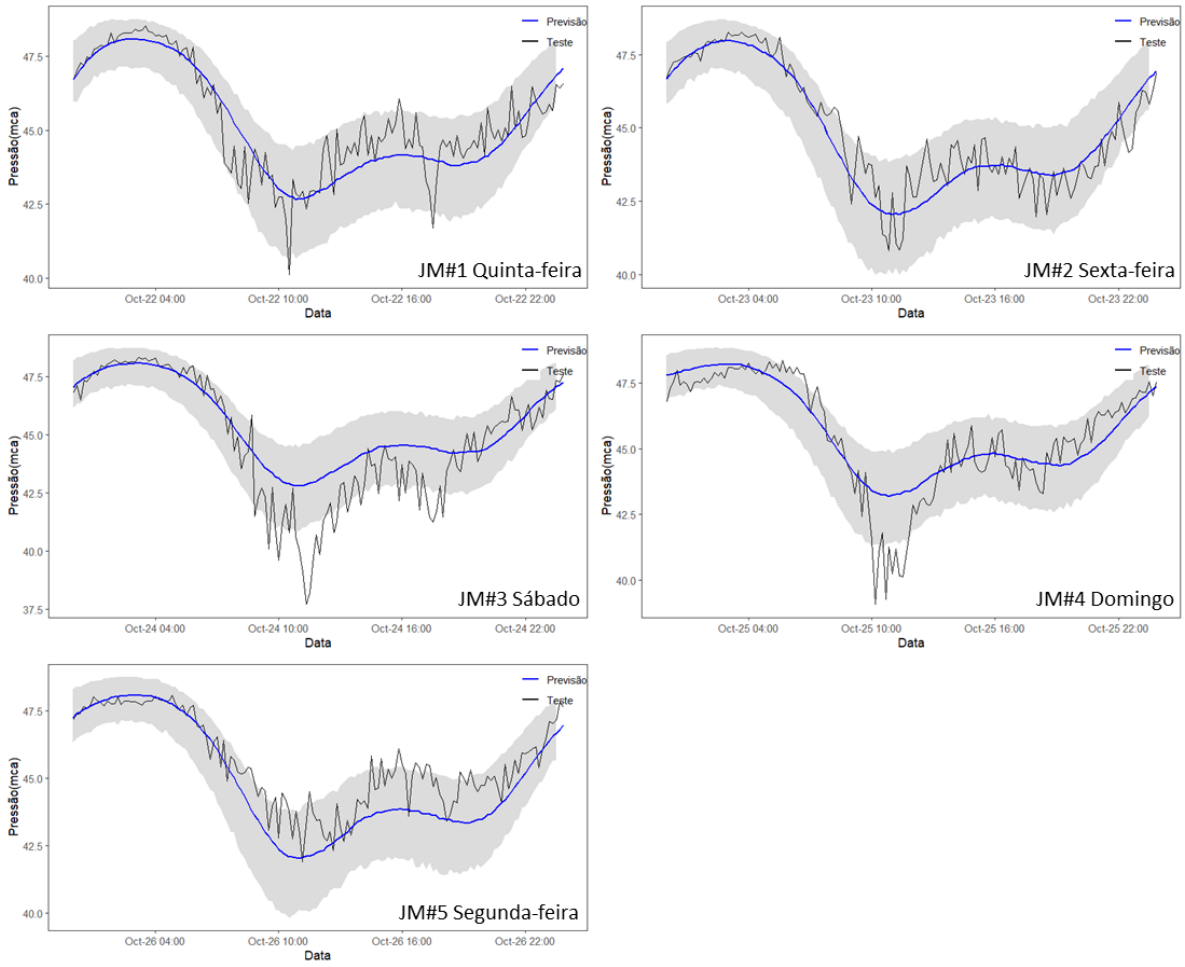
Utilizando o modelo GARMA(2,0) com distribuição gama, calcularam-se as métricas RMSE e MAPE para cada uma das janelas móveis. Verificou-se que, em comparação com a fase de ajuste do modelo, todos os valores de acurácia aumentaram, como era previsto. Contudo, observou-se uma variação nas acurácias entre as diferentes janelas móveis. Analisando os resultados na Tabela 11, notou-se que o maior valor de RMSE, 1.40 mca, correspondeu à JM#3, que representa o sábado, enquanto o segundo maior valor, 1.00 mca, foi observado na JM#4, relacionada ao domingo. Essas variações eram previsíveis, visto que a análise do comportamento da série, conforme discutido na seção 4.4, indicou que esses dias apresentam um comportamento notavelmente distinto em relação aos demais da semana. Por outro lado, a JM#2, correspondente à sexta-feira, apresentou o menor RMSE, com 0.70 mca. No que diz respeito à métrica MAPE, observou-se um padrão semelhante, com os valores mais elevados ocorrendo nas JM#3 e JM#5, atingindo 2.24% e 1.63%, respectivamente, e o valor mais baixo, 0.96%, registrado na JM#5.

Tabela 11 - Resumo das métricas de acurácia de previsão segundo o modelo GARMA para cada uma das janelas móveis analisadas.

Janela Móvel	RMSE (mca)	MAPE (%)
#1	0.80	1.30
#2	0.70	1.20
#3	1.40	2.24
#4	1.00	1.60
#5	0.96	1.63

Além das métricas apresentadas, é possível analisar os gráficos com os valores observados e preditos com o modelo GARMA(2,0) com distribuição gama. Através da Figura 36 pode-se visualizar os dados observados na linha de cor preta, a média de previsão na cor azul, com seu respectivo intervalo de confiança de 95% na cor cinza.

Figura 36 - Gráficos de previsão dos valores futuros para cada janela móvel através do modelo GARMA(2,0) com distribuição gama.



Uma comparação entre o desempenho dos modelos para cada janela de móvel é apresentada na Tabela 12. O modelo ARIMA+Harmônico obteve o melhor valor para todas as análises, se mostrando o modelo com maior exatidão independente do dia da semana em que se está prevendo. Na sequência, o modelo GARMA obteve melhor desempenho nas JM#1 e JM#2 do que o modelo SARIMA, que foi melhor nas três janelas de análise seguintes. Contudo, ressalta-se que o desempenho dos três modelos foi muito próximo.

Tabela 12 - Avaliação comparativa do desempenho das previsões dos modelos, quanto ao valor de RMSE (mca) para cada uma das janelas móveis.

Janela Móvel	SARIMA	ARIMA + Harmônico	GARMA
#1	0.82	0.77	0.80
#2	0.80	0.69	0.70
#3	1.32	0.90	1.40
#4	0.99	0.97	1.00
#5	0.66	0.59	0.96

Quanto ao resultado do MAPE, na avaliação comparativa entre modelos, novamente o modelo ARIMA+Harmônico obteve os melhores resultados para a maioria das janelas móveis analisadas, com exceção da JM#4, que representa o domingo, onde obteve desempenho pior do que os outros dois modelos. O modelo GARMA por sua vez, obteve resultados intermediários entre os modelos avaliados nas janelas móveis JM#1 e JM#2, o pior desempenho nas janelas JM#3 e JM#5, e o melhor de todos na JM#4. O modelo SARIMA não obteve destaque em nenhuma janela de análise.

Tabela 13 - Avaliação comparativa do desempenho das previsões dos modelos, quanto ao valor de MAPE (%) para cada uma das janelas móveis.

Janela Móvel	SARIMA	ARIMA + Harmônico	GARMA
#1	1.32	1.19	1.30
#2	1.37	1.15	1.20
#3	2.17	1.51	2.24
#4	1.70	1.71	1.60
#5	1.17	1.03	1.63

Foi possível perceber através desta análise que os modelos tiveram desempenho muito parecidos quanto a acurácia de previsão dentro da metodologia proposta e que as métricas atingidas satisfazem o propósito de previsão. Porém, é sabido que cada um deles possui complexidades diferentes de modelagem, e configurações que podem oferecer vantagens ou desvantagens a partir da aplicação desejada.

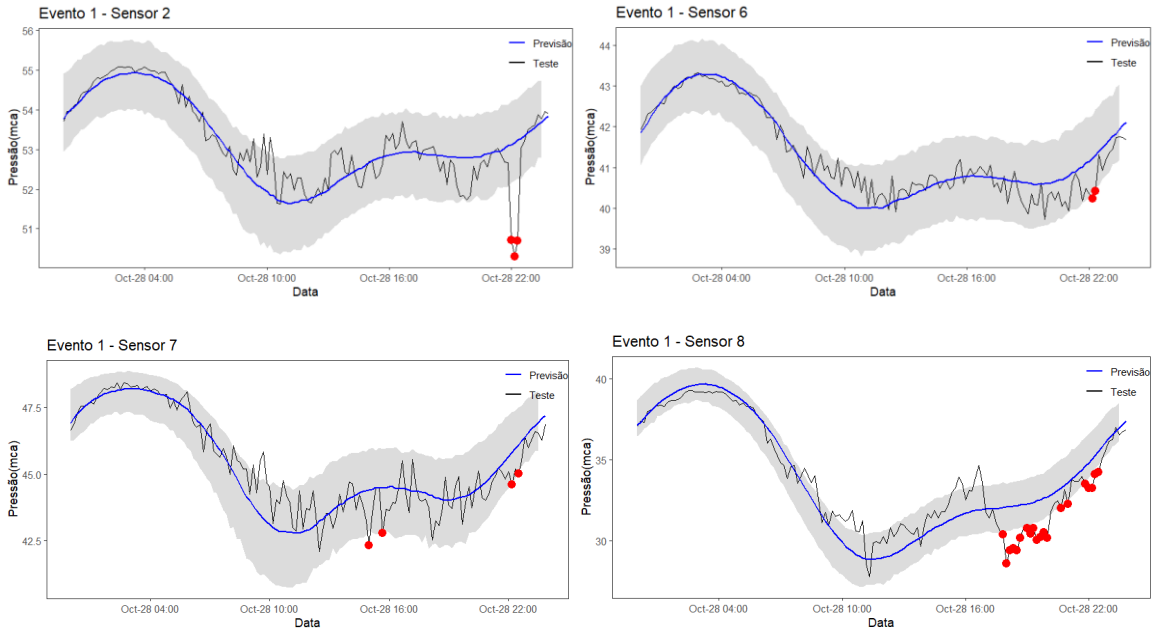
Semelhantemente aos resultados encontrados por Mason (2019), a análise de resíduos dos modelos SARIMA e ARIMA+Harmônico aqui empregadas revelou que os modelos não atenderam os pressupostos, principalmente quanto a não distribuição normal e a variância não constante. O que leva a concluir que não seria possível estabelecer intervalos de confiança realmente seguros para a previsão. Em um cenário em que se deseja um bom intervalo de confiança, que represente a realidade efetivamente, estes dois modelos estariam descartados. Comparando as figuras de previsão de cada modelo, é possível perceber que ambos SARIMA e ARIMA+Harmônico possuem um intervalo relativamente constante ao longo do tempo, sem variação das faixas, o que não representa o comportamento esperado para a série de pressões ao longo de um dia em um sistema de distribuição de água.

Diferentemente do que se encontra para os resultados do modelo GARMA(2,0) com distribuição gama. Onde seus resíduos atenderam os pressupostos e foi possível traçar intervalo de confiança útil para posteriores aplicações de controle e detecção de anomalias. Destaque para o intervalo de confiança plotado no gráfico da Figura 36, onde pode-se observar que seu comportamento segue a lógica de ter uma menor faixa de variação durante a noite e uma maior faixa durante o dia. Este ponto explicita a condição do modelo empregado em entender o comportamento da rede e consequentemente promover uma previsão de maior qualidade. A melhor representação de um intervalo de confiança exerce papel importante também nos objetivos de estabelecimento de processos de controle. Pois, fornecem informações capazes de auxiliar na quantificação do desvio das observações com o intervalo previsto, podendo ser base para implementação de técnicas de detecção de anomalias.

4.5 Detecção de vazamentos em eventos simulados

Neste sentido, uma verificação foi realizada com base em um evento simulado de vazamento durante a campanha de monitoramento de GAMBOA-MEDINA (2017), onde através da abertura de hidrantes simulou um evento de 28 l/s de vazão no dia 28/10/2015 às 22:00h em ponto localizado no nó “h5”, conforme Figura 5. A partir disso, foi ajustado modelo GARMA(2,0) para os 14 dias anteriores ao evento com o objetivo de prever todo o comportamento da rede no dia 28/10/2015, para todos os sensores, com seu respectivo intervalo de confiança. Os resultados apontam que todos os sensores detectaram pontos fora do intervalo de confiança no momento da simulação. Exemplos de comportamentos diferentes entre os sensores podem ser observados na Figura 37, onde a distância entre o valor observado e a previsão varia em cada sensor.

Figura 37 - Gráfico de detecção de observações de pressão fora do intervalo de confiança da previsão feita com modelo GARMA(2,0)



Foram detectadas anomalias nos sinais com desvios no horário relativo à simulação indicando que a abordagem previsão e controle proposta pode ser efetiva na detecção de anomalias no comportamento da rede de abastecimento de água. O sensor s2 por exemplo, detectou os três pontos referentes aos 30 minutos de simulação, e por estar próximo ao ponto simulado, fez isso com certa facilidade, pois o evento gerou uma queda brusca na pressão no nó monitorado por este sensor. Diferentemente do que aconteceu com os sensores s6 e s7 que, mais distantes, tiveram menor influência do evento, porém foram efetivos em detectar o evento. Vale ressaltar também, alguns falsos positivos encontrados nos sensores s7 e s8, onde em algumas horas antes da simulação apresentaram valores observados abaixo do limite mínimo do intervalo de confiança.

5 CONCLUSÃO

O presente estudo teve como objetivo estudar, aplicando os conceitos de análise de séries temporais, os sinais de carga de pressão de uma rede real de abastecimento de água, motivado pela importância de construir base sólida sobre a descrição das séries, e compreender mais sobre seu comportamento. Além disso, o estabelecimento da relação entre os dados observados e o mecanismo gerador faziam parte das expectativas iniciais, com foco em encontrar evidências provenientes de métodos estatísticos robustos que sustentassem as decisões estratégicas posteriores.

Isto posto, nota-se o foco da revisão bibliográfica em fundamentos de séries temporais, desde sua composição, características básicas, estimação de parâmetros, suposições e análise de resíduos, até aplicação de algoritmos automáticos de ajuste. Foram exploradas técnicas de regressão por variáveis exógenas e séries de Fourier que puderam agregar no processo de modelagem, até que se chegasse em uma configuração de modelo adequada para a série posta em desafio. De especial interesse foram as componentes de maior frequência ou ruídos das séries temporais, para os que as abordagens mais comuns de análise e modelagem não lidavam corretamente, mas que se encontrou nesta pesquisa que, com a correta modelagem, pode melhor caracterizar os intervalos de confiança das previsões. Com isso, o desafio passou de um status de obstáculo ao avanço dos estudos, para característica que confere robustez ao final do processo.

As análises de correlação e autocorrelação, sazonalidade e estacionariedade permitiram encontrar que os dados são autocorrelacionados, as séries estacionárias e com sazonalidade determinística presente, tanto pela análise gráfica, quanto por testes estatísticos. Considera-se assim que o primeiro objetivo específico desta pesquisa foi atingido, quando através da metodologia proposta foi possível identificar a relação entre cada de uma série temporal de carga de pressão em uma rede de abastecimento de água com o perfil de consumo de água típico residencial. Ou seja: tendência diretamente ligada à mudança nos padrões de consumo da população e geralmente ocorrendo a longo prazo, não estando presente em trechos curtos da ordem de dias e semanas; sazonalidade, possui comportamento complexo pela ocorrência de mais de uma sazonalidade ao mesmo tempo, em função da hora do dia e do dia da semana; ruído, fator relacionado com consumos instantâneos, que ocorre com maior variação durante o

período diurno, e com menor variação no período noturno, conferindo característica heterocedástica à série.

Uma vez analisadas as características das séries, foi possível seguir para o processo de modelagem. O modelo SARIMA demonstrou ser um ótimo primeiro passo em direção à modelagem das séries de carga de pressão, em função da boa acurácia de previsão e simplicidade no ajuste, que ocorreu sem adição de variáveis exógenas. No entanto, a limitação da consideração de uma única sazonalidade e o não atendimento dos pressupostos de variância constante e distribuição normal, sobre os resíduos, demonstrou a necessidade de melhorias. O fato de o ajuste ocorrer de forma automática pode configurar ponto positivo para futuras abordagens, a depender dos objetivos e características dos dados.

Surge a proposta da aplicação de um modelo ARIMA somado com variáveis exógenas de uma série de pares senos e cossenos capazes de representar a variação sazonal complexa presente nos dados. Essa variação complexa é definida pela ocorrência de mais de uma sazonalidade influenciando nos dados observados ao mesmo tempo, como a variação em função da hora do dia e do dia da semana, bastante exploradas ao longo do trabalho. Apesar da melhoria nos resultados, RMSE e MAPE menores do que obtidos com modelo SARIMA, o modelo construído ainda não foi capaz de atingir os pressupostos sobre os resíduos, quanto à necessidade de variância constante e distribuição normal, assim como o anterior .

Observou-se que esses primeiros modelos propostos (SARIMA e ARIMA) não conseguiram extrair todas as informações presentes nos dados, e os gráficos de resíduos ao longo do tempo demonstraram que havia padrão de repetição nos resíduos e que sua variância mudava com o tempo, diferentemente do pressuposto de homocedasticidade que sugere variância constante dos resíduos. Para lidar com essa característica, observou-se a distribuição dos dados e entendeu-se que a heterocedasticidade era intrínseca ao mecanismo gerador, e uma abordagem adequada deveria considerar uma função de densidade de probabilidade que não a normal.

Para tanto, a série de carga pressão foi transformada em uma série de “Energia consumida”, através da diferença entre a pressão estática local e as pressões observadas. Na série de energia consumida, ao contrário da de carga de pressão, as maiores variâncias ocorrem nos picos e as menores variâncias dos dados nos vales. Com isso facilita-se a aplicação da

função de densidade de probabilidade gama associada ao modelo generalizado autorregressivo e de médias móveis (GARMA).

Os resultados demonstraram que a aplicação de um modelo GARMA (2,0) com densidade de probabilidade gama e série de Fourier para as componentes sazonais, é adequada para modelagem das séries temporais de carga pressão de redes reais de abastecimento de água. Pois além de os valores de RMSE e MAPE demonstrarem boa acurácia na previsão, os resíduos do modelo podem ser considerados independentes, homocedásticos e com distribuição normal. E com a estratégia de transformação da série para Energia Consumida é possível ainda generalizar a aplicação do modelo, onde o mesmo demonstrou bons resultados em 9 sensores em um mesmo setor da rede.

De posse dos diversos ajustes e previsões realizados com modelo GARMA(2,0) aplicados nos diferentes sensores, em diferentes épocas do ano para diferentes horizontes de previsão, foi possível analisar cada um desses fatores. Os melhores resultados para as métricas RMSE e MAPE podem ser atingidos ao se utilizar modelo com 14 dias (2 ciclos semanais) para ajuste da série, ao invés do ajuste “completo”, que considerou dados observados durante um ano inteiro de observações. Uma vez que esse último obteve valores de RMSE em média de 24 a 29% maiores que o modelo mensal, e valores de MAPE de 37 a 43% também maiores que o modelo mensal.

Através de resultados alcançados com sensor s7, pode-se avaliar ainda que um baixo índice de falhas nos dados e forte estacionariedade na tendência da série em uma escala anual, pode guiar a bons resultados no ajuste de modelos que considerem séries maiores na fase de ajuste, o que pode ser útil a depender das características das séries a serem trabalhadas futuramente.

Como esperado, melhores métricas são obtidas para um horizonte de previsão de 24 horas. Com resultados para a métrica RMSE em média (a depender do conjunto de ajuste) de 7 e 6% piores para o horizonte de previsão de 48 horas, não se descarta a utilidade de previsões com esse horizonte estendido, a depender dos objetivos. O modelo de previsão se mostrou bastante estável, e horizontes de previsão ainda maiores não devem ser descartados em trabalhos futuros, inclusive com possíveis aplicações para preenchimento de falhas em séries temporais, por exemplo.

Considera-se que a modelagem GARMA pode ser utilizada para fazer previsão de sinais, considerando que apresentou bons resultados na previsão independente do período do ano e sensor considerado. E que seu ajuste, ao atender os pressupostos sobre os resíduos, garante confiabilidade no uso do intervalo de confiança, superando limitações anteriores apontadas nesta pesquisa e na revisão de literatura.

Ainda, na pesquisa provou-se que abordagens de controle de processos podem ser aplicadas de forma efetiva na identificação de anomalias a partir da previsão, com intervalo de confiança adequado, da condição de normalidade da série. E que as metodologias propostas foram capazes de direcionar a um ciclo virtuoso de erros, correções e aprendizados, onde as limitações encontradas no desenvolvimento da pesquisa demonstraram ser a base para cada degrau avançado. Assim, conclui-se que a análise de séries temporais aplicada a séries de carga de pressão de redes de abastecimento configura-se como abordagem robusta para descrição das séries, e constitui base para modelagem a partir de um modelo generalizado autorregressivo e de médias móveis que se ajusta adequadamente às características dos dados de pressão, ainda que heterocedásticos, proporcionando boas previsões de dados.

Recomenda-se que pesquisas futuras explorem a viabilidade da detecção de anomalias, como vazamentos e grandes consumos, por meio da comparação dos valores observados com as previsões geradas pelos modelos apresentados neste estudo. Além disso, é recomendável que sejam desenvolvidas funções para quantificar a magnitude das anomalias e que se investigue a relação entre os sensores em um determinado setor de distribuição de água.

REFERÊNCIAS

ADACHI, S., TAKAHASHI, S., & TAKEMOTO, T. **Online burst detection in water networks with an ensemble of flow prediction models**. CCWI2017 - Computing and Control for the Water Industry Conf., 2017.

ALBARRACIN,O.Y.E.; ALENCAR, A. P.; LEE HO,L. Generalized autoregressive and moving average models: multicollinearity, interpretation and a new modified model. **Journal of Statistical Computation and Simulation**, 2019.

BAKKER, M. et al. Detecting pipe bursts using Heuristic and CUSUM methods. **Procedia Engineering**, v. 70, p. 85-92, 2014.

BENJAMIN, M.; RIGBY, R. A.; STASINOPOULOS, D. M. **Generalized autoregressive moving average models**. J. Am. Statist. Ass., 98:214–223, 2003.

BOX, G. E. P. "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N. (eds.), **Robustness in Statistics**, Academic Press, pp. 201–236, doi:10.1016/B978-0-12-438150-6.50018-2, ISBN 9781483263366. 1979.

BOX, G.E.P.; JENKINS, G. M.; REINSEL, G. C. **Time series analysis: forecasting and control.** 4 ed. John Wiley e Sons, 2008.

BRENTAN, B.; LUVIZOTTO JR. E.; IZQUIERDO, J. & PÉREZ-GARCÍA, R. **Fourier series and Chebyshev polynomials applied to real-time water demand forecasting**. Acta Universitaria, 26(NE-3), 2016.

BROCKWELL, P.J.; DAVIS, Richard A. (2 Ed.). **Introduction to time series and forecasting.** New York, NY: Springer New York, 2010.

CAMPOS, F. S.; GAMBOA-MEDINA, M. M.. AJUSTE DE MODELOS ESTATÍSTICOS PARA SÉRIES TEMPORAIS DE PRESSÃO EM SISTEMAS DE DISTRIBUIÇÃO DE ÁGUA. **Anales del XXX Congreso Latinoamericano de Hidráulica**: INGENIERÍA E INFRAESTRUCTURAS HIDRÁULICAS, Madri, v. 5, p. 173-182, 2023.

CHATFIELD, C., **The Analysis of Time Series: An Introduction**. 5 ed. Londres: Chapman and Hall CRC, 1996.

DE LIVERA, A. M.; HYNDMAN, R. J.; SNYDER, R. D. Forecasting time series with complex seasonal patterns using exponential smoothing. **J American Statistical Association**, 106(496), 1513–1527, 2011.

DE MATTOS, R. S. **Tendências e Raízes Unitárias**. Texto Didático. UFJF: Juiz de Fora, 2018.

DESIGN COUNCIL. **Framework for Innovation: a study of the design process, the double diamond.** 2004. Disponível em: https://www.designcouncil.org.uk/our-resources/framework-for-innovation/. Acesso em: 07 ago. 2023.

DU, B. et al. Deep learning with long short-term memory neural networks combining wavelet transform and principal component analysis for daily urban water demand forecasting. **Expert Systems with Applications**, v. 171, p. 114571, 2021.

FAN, S.; HYNDMAN, R. J. Short-term load forecasting based on a semi-parametric additive model. **IEEE transactions on power systems**, v. 27, n. 1, p. 134-141, 2011.

GAMBOA-MEDINA, M. M. **Detecção de vazamentos e alterações em redes de distribuição de água para abastecimento, durante a operação, usando sinais de pressão**. 2017. Tese de Doutorado. Universidade de São Paulo.

GOULD, P. G. et al. Forecasting time series with multiple seasonal patterns**. European Journal of Operational Research**, v. 191, n. 1, p. 207-222, 2008.

HAMILTON, S.; CHARALAMBOUS, B.; WYETH, G. **Improving Water Supply Networks: Fit for Purpose Strategies and Technologies.** 2021.

HYNDMAN, R. J.; KHANDAKAR, Y., Automatic time series forecasting: The forecast package for R. **Journal of Statistical Software**, 2008.

HYNDMAN, R.J.; ATHANASOPOULOS, G. **Forecasting**: principles and practice, 2nd edition, OTexts: Melbourne, Australia, 2018. Disponível em <OTexts.com/fpp2>. Acesso em 6 de junho de 2022.

HYNDMAN, R.J.;ATHANASOPOULOS, G. **Forecasting**: principles and practice, 3ª edição, OTexts: Melbourne, Australia, 2021. Disponível em <OTexts.com/fpp3>. Acesso em 26 de abril de 2022.

LARRUBIA, L.F. (2021). **Detecção de anomalias, interpolação e previsão em tempo real de séries temporais para operação de reservatórios e distribuição de água**. Dissertação. IME, USP, São Paulo.

LAZZERI, F. **Machine learning for time series forecasting with Python**. John Wiley e Sons, 2020.

MASON, B. A. G. **Generalized Additive Model and Artificial Neural Networks for Water Demand.** Dissertação. Técnico Lisboa, 2019.

MATTOS, T. B. **Noções de Inferência no R**. Capítulo 1 Distribuições de Probabilidade. 2018. Apostila digital .Disponível em: https://bookdown.org/thalita_dobem/Apostila/. Acesso em: 01 set. 2023.

MORETTIN, P.A.; TOLOI, C.M.C. **Análise de séries temporais**. 2 ed. São Paulo: Egard Blucher, 2006.

MOUNCE, S. R.; MOUNCE, R. B.; BOXALL, J. B. Novelty detection for time series data analysis in water distribution systems using support vector machines. **Journal of hydroinformatics,** v. 13, n. 4, p. 672-686, 2011.

ODAN, F. K.; REIS, L. F. R. Hybrid water demand forecasting model associating artificial neural network with Fourier series. **Journal of Water Resources Planning and Management**, v. 138, n. 3, p. 245-256, 2012.

RIGBY, R.A., STASINOPOULOS, M.D., HELLER, G.Z., & DE BASTIANI, F. (2019). **Distributions for Modeling Location, Scale, and Shape:** Using GAMLSS in R (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9780429298547

SALLOOM, T.; KAYNAK, O.; HE, W. A novel deep neural network architecture for real-time water demand forecasting. **Journal of Hydrology**, v. 599, p. 126353, 2021.

TSUTIYA, M. T. **Abastecimento de água:** Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da Universidade de São Paulo, 643p. São Paulo, p. 42, 2006.

WEBEL, K.; OLLECH, D. An overall seasonality test. **Deutsche Bundesbank's Discussion Paper series.** 2019.

WOOLDRIDGE, J. M. **Introductory econometrics: A modern approach**. Cengage learning, 2015.

YIPENG, W.; SHUMING, L. A review of data-driven approaches for burst detection in water distribution systems, **Urban Water Journal**, 2017. DOI: 10.1080/1573062X.2017.1279191

ZAHED, K. F. (1990). **Previsão de demanda de consumo em tempo real no desenvolvimento operacional de sistemas de distribuição de água.** Tese de doutorado, Escola Politécnica da Universidade de São Paulo, São Paulo.

ZIVOT, E.; WANG, J. Rolling analysis of time series. In: **Modeling Financial Time Series with S-Plus®.** Springer, Nova Iorque 2003. p. 299-346.


EESC • USP