

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANA ESCOLA POLITÉCNICA

PROGRAMA DE PÓS-GRADUAÇAO EM ENGENHARIA DE PRODUÇAO E SISTEMAS (PPGEPS)

FRANCHESCO SANCHES DOS SANTOS

EXPLORANDO A EFICIÊNCIA DOS MODELOS DE PREVISAO DE SÉRIES TEMPORAIS NO ABASTECIMENTO DE AGUA

CURITIBA '20'23

EXPLORANDO A EFICIÊNCIA DOS MODELOS DE PREVISAO DE SÉRIES TEMPORAIS NO ABASTECIMENTO DE AGUA

Projeto de Pesquisa de Mestrado apresen- tado ao Programa de Pós-Graduação em En- genharia de Produção e Sistemas (PPGEPS). Area de concentração: Automação e Con- trole de Sistemas, da Escola Politécnica, da Pontifícia Universidade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção e Sis- temas.

Orientador: Dr. Leandro dos Santos Coelho Coorientadora: Dr. Viviana Cocco Mariani

CURITIBA '20'23

EXPLORANDO A EFICIÊNCIA DOS MODELOS DE PREVISÃO DE SÉRIES TEMPORAIS NO ABASTECIMENTO DE ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS). Area de concentra- ção: Gerência de Produção e Logística, da Escola Politécnica, da Pontifícia Universi- dade Católica do Paraná, como requisito parcial à obtenção do título de Mestre em Engenharia de Produção e Sistemas.

COMISSÃO EXAMINADORA

Dr. Leandro dos Santos Coelho Orientador

Pontifícia Universidade Católica do Paraná

Dr. Viviana Cocco Mariani Coorientadora

Pontifícia Universidade Católica do Paraná

Convidado A Membro Externo Instituição A

Convidado B Banca Instituição B

Curitiba, '26 de jun. de '20'23

Ao examinar as séries temporais, veJo a evidência da existência de Deus na perfeita ordem e regularidade que caracterizam o tempo. Cada ponto de dados e cada instante são testemunhos do Seu

controle absoluto sobre todas as coisas.

Agradecimentoi,

Primeiramente, expresso minha gratidão a Deus por todas as bênçãos recebidas, pois foi Ele quem abriu caminhos e me deu forças para superar esse desafio, tornando-o possível.

A minha família, sou grato pelo apoio incondicional e pelo estímulo constante para seguir em frente com determinação, buscando sempre alcançar novos patamares.

Agradeço ao professor Leandro dos Santos Coelho pela oportunidade de trabalhar ao seu lado e compartilhar seus conhecimentos e experiências ao longo do meu mestrado. Sua orientação contribuiu significativamente para o meu crescimento profissional e pessoal, tornando este trabalho uma realidade.

A professora Viviana Cocco Mariani, agradeço pela disponibilidade e paciência em me auxiliar nas minhas dificuldades, utilizando seu conhecimento para contribuir com o desenvolvimento da pesquisa.

Quero expressar minha gratidão à equipe da Pontifícia Universidade Católica do Paraná (PUCPR) e aos demais professores, especialmente à secretária Denise da Mata Medeiros (PPGEPS), pela paciência, carinho e apoio prestados em diversas ocasiões, sem medir esforços.

Aos meus amigos, que sempre torceram por mim, e aos novos amigos que conquistei ao longo dessa jornada, agradeço por compartilharmos momentos de alegria nessa batalha. Sou grato ao investimento em bolsas de estudo concedidas pela Coordenação de

Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que possibilitou a conclusão dessa etapa da minha carreira profissional e acadêmica.

A imaginação é mais importante que o conhecimento.

Pois o conhecimento é limitado, enquanto a imaginação abraça o mundo inteiro.

- Albert Eim teim

Resumo

E te e tudo aborda a importâmcia e tratégica da previ ão preci a da demamda de água como uma ferramemta para a ge tão eficiemte do recur o hídrico em um ce- mário competitivo. O problema idemtificado é a falta de previ õe preci a , o que dificulta a tomada de deci õe e tratégica mo aba tecimemto de água. A olução propo ta é o u o de modelo avamçado de previ ão de érie temporai para melho- rar a preci ão da previ õe de demamda. Com ba e em uma revi ão abramgemte da literatura exi temte, ão amali ado diferemte método e abordagem utilizado ma previ ão de érie temporai mo comtexto do aba tecimemto de água. O e tado da arte é explorado para idemtificar o modelo mai eficaze e a melhore prática ma área. Com ba e mo e tado da arte, ão propo to método e produto e pecífico para a previ ão de demamda de água. E e método levam em com ideração variávei exógema , azomalidade do dado e utilizam modelo autorregre ivo imtegrado de média móvei (ARIMA), técmica de boo timg como XGBoo t (Extreme Gra- diemt Boo timg), LightGBM (Light Gradiemt Boo timg Machime) e regre ão limear. Além di o, também é com iderado o u o de modelo ba eado em Ramdom Fore t Regre iom (RFR). O re ultado obtido por meio da aplicação de e método e produto propo to ão amali ado e comparado utilizamdo métrica de de empe- mho como erro percemtual ab oluto médio (MAPE), erro ab oluto médio (MAE) e raiz quadrada do erro quadrático médio (RMSE). E a de coberta formecem im- formaçõe valio a obre a eficácia do modelo de previ ão de érie temporai mo aba tecimemto de água e comtribuem para uma tomada de deci ão mai imformada e eficiemte me a área.

Palavras-chave: Previ ão, Ecomomia de água, Série temporai , Modelo de Pre- vi ão.

Abstract

Thi tudy addre e the trategic importamce of accurate water demamd foreca t- img a a tool for effi ciemt water re ource mamagememt im a competitive cemario. The idemtified problem i the lack of preci e predictiom , which himder trategic deci iom-makimg im water upply. The propo ed olutiom i the u e of advamced time erie foreca timg model to emhamce demamd predictiom accuracy. Ba ed om a comprehem ive review of exi timg literature, differemt method amd approache u ed im water upply time erie foreca timg are amalyzed. The tate of the art i ex- plored to idemtify the mo t effective model amd be t practice im the field. Buildimg upom the tate of the art, pecific method amd product for water demamd fore- ca timg are propo ed. The e method take imto com ideratiom exogemou variable , data ea omality, amd utilize autoregre ive imtegrated movimg average (ARIMA) model , boo timg techmique uch a XGBoo t (Extreme Gradiemt Boo timg) amd LightGBM (Light Gradiemt Boo timg Machime), a well a limear regre iom. Ad- ditiomally, the u e of Ramdom Fore t Regre iom (RFR) model i al o com idered. The re ult obtaimed through the applicatiom of the e propo ed method amd prod- uct are amalyzed amd compared u img performamce metric uch a meam ab olute percemtage error (MAPE), meam ab olute error (MAE), amd root meam quare error (RMSE). The e fimdimg provide valuable im ight imto the effectiveme of time erie foreca timg model im water upply amd comtribute to more imformed amd effi ciemt deci iom-makimg im thi field.

Keywords: Foreca timg, Water avimg , Time erie , Sy tematic literature review.

# Lista de Figuras

|  |  |  |
| --- | --- | --- |
| [1](#_bookmark2)  ['2](#_bookmark19) | [Paradigma de aprendizado de máquina](#_bookmark2) . . . . . . . . . . . . . . . . . . . .  [Mapa das Etapas](#_bookmark19) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | '2  5 |
| [3](#_bookmark31) | [Estrutura da dissertação](#_bookmark31) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 11 |
| [4](#_bookmark35) | [Dados completos com uma frequência média de '24 horas](#_bookmark35) . . . . . . . . . . | 13 |
| [5](#_bookmark36) | [Plotagem de dados para o ano de '20'20](#_bookmark36) . . . . . . . . . . . . . . . . . . . . | 13 |
| [6](#_bookmark37) | [Exemplo de séries temporais](#_bookmark37) . . . . . . . . . . . . . . . . . . . . . . . . . . | 14 |
| [7](#_bookmark38) | [Processo estocástico](#_bookmark38) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 15 |
| [8](#_bookmark40) | [Mapa conceitual do problema de pesquisa](#_bookmark40) . . . . . . . . . . . . . . . . . . | 16 |
| [9](#_bookmark46) | [Etapas da Revisão](#_bookmark46) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 17 |
| [10](#_bookmark49) | [Palavras-chave mais populares na Scopus.](#_bookmark49) . . . . . . . . . . . . . . . . . . | 19 |
| [11](#_bookmark50) | [Palavras-chave mais populares na VoS](#_bookmark50) . . . . . . . . . . . . . . . . . . . . | 19 |
| [1'2](#_bookmark53) | [Analise das quantidades de artigos em relação aos anos.](#_bookmark53) . . . . . . . . . . . | '21 |
| [13](#_bookmark56) | [Relação de autores entre artigos publicados](#_bookmark56) . . . . . . . . . . . . . . . . . . | '23 |
| [14](#_bookmark57) | [Ligação bibliográfica entre os autores](#_bookmark57) . . . . . . . . . . . . . . . . . . . . . | '23 |
| [15](#_bookmark58) | [Mapa mundial da publicação de artigos em todo o mundo](#_bookmark58) . . . . . . . . . . | '24 |
| [16](#_bookmark59) | [Areas de aplicação do tema](#_bookmark59) . . . . . . . . . . . . . . . . . . . . . . . . . . | '25 |
| [17](#_bookmark71) | [Modelo AR(7)](#_bookmark71) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 3'2 |
| [18](#_bookmark72) | [ARX (7)](#_bookmark72) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 3'2 |
| [19](#_bookmark78) | [Modelo MA(7)](#_bookmark78) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 36 |
| ['20](#_bookmark80) | [ARMA (7,7)](#_bookmark80) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 37 |
| ['21](#_bookmark83) | [ARIMA (7,1,7)](#_bookmark83) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 38 |
| ['2'2](#_bookmark85) | [SARIMA (7*,* 1*,* 7)(2*,* 1*,* 1)12](#_bookmark85) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 39 |
| ['23](#_bookmark89) | [ARIMAX (7*,* 1*,* 7)](#_bookmark89) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 40 |
| ['24](#_bookmark90) | [SARIMAX (7*,* 1*,* 7)(2*,* 1*,* 1)12](#_bookmark90) . . . . . . . . . . . . . . . . . . . . . . . . . . | 40 |
| ['25](#_bookmark93) | [Corelação de Pearson](#_bookmark93) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 4'2 |
| ['26](#_bookmark96) | [Regressão linear LT01 vs PT01 correlação 98%](#_bookmark96) . . . . . . . . . . . . . . . . | 43 |
| ['27](#_bookmark97) | [Regressão linear (LR) um passo a frente](#_bookmark97) . . . . . . . . . . . . . . . . . . . | 44 |
| ['28](#_bookmark99) | [Regressão da Floresta Aleatória (RFR)](#_bookmark99) . . . . . . . . . . . . . . . . . . . . | 44 |
| ['29](#_bookmark100) | [Esquema da Floresta Aleatória](#_bookmark100) . . . . . . . . . . . . . . . . . . . . . . . . | 45 |
| [30](#_bookmark102) | [Impulsionando gradiente com XGBoost e LightGBM](#_bookmark102) . . . . . . . . . . . . | 46 |
| [31](#_bookmark107) | [Crescimento em folha versus crescimento em nível](#_bookmark107) . . . . . . . . . . . . . . | 48 |
| [3'2](#_bookmark108) | [XGBoost e LigthGBM regressão](#_bookmark108) . . . . . . . . . . . . . . . . . . . . . . . . | 49 |
| [33](#_bookmark126) | [Decomposição STL aditiva dos dados coletados](#_bookmark126) . . . . . . . . . . . . . . . | 59 |
| [34](#_bookmark127) | [Decomposição STL multiplicativa dos dados coletados](#_bookmark127) . . . . . . . . . . . . | 59 |
| [35](#_bookmark128) | [Violino no nível do reservatório](#_bookmark128) . . . . . . . . . . . . . . . . . . . . . . . . | 60 |

|  |  |  |
| --- | --- | --- |
| [36](#_bookmark129) | [Violino da vazão de recalque](#_bookmark129) . . . . . . . . . . . . . . . . . . . . . . . . . . | 61 |
| [37](#_bookmark133) | [Autocorrelação e Autocorrelação parcial](#_bookmark133) . . . . . . . . . . . . . . . . . . . | 6'2 |
| [38](#_bookmark135) | [Ruído branco](#_bookmark135) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 64 |
| [39](#_bookmark142) | [Comparação dos modelos ARIMAS](#_bookmark142) . . . . . . . . . . . . . . . . . . . . . . | 67 |
| [40](#_bookmark143) | [Comparação de modelos de regressão](#_bookmark143) . . . . . . . . . . . . . . . . . . . . . | 68 |
| [41](#_bookmark149) | [Demanda Média das Variáveis de Fluxo](#_bookmark149) . . . . . . . . . . . . . . . . . . . . | 73 |
| [4'2](#_bookmark245) | [Comparação dos modelos AR, ARX e MA, 1 dia à frente](#_bookmark245) . . . . . . . . . . | 94 |
| [43](#_bookmark246) | [Comparação dos modelos AR, ARX e MA, 7 dias à frente](#_bookmark246) . . . . . . . . . | 94 |
| [44](#_bookmark247) | [Comparação dos modelos AR, ARX e MA, 14 dias à frente](#_bookmark247) . . . . . . . . | 95 |
| [45](#_bookmark248) | [Comparação dos modelos AR, ARX e MA, 30 dias à frente](#_bookmark248) . . . . . . . . | 95 |
| [46](#_bookmark250) | [Comparação dos modelos ARMA e ARIMA, 1 dia à frente](#_bookmark250) . . . . . . . . . | 96 |
| [47](#_bookmark251) | [Comparação dos modelos ARMA e ARIMA, 7 dias à frente](#_bookmark251) . . . . . . . . | 96 |
| [48](#_bookmark252) | [Comparação dos modelos ARMA e ARIMA, 14 dias à frente](#_bookmark252) . . . . . . . . | 97 |
| [49](#_bookmark253) | [Comparação dos modelos ARMA e ARIMA, 30 dias à frente](#_bookmark253) . . . . . . . . | 97 |
| [50](#_bookmark255) | [Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente](#_bookmark255) | 98 |
| [51](#_bookmark256) | [Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente](#_bookmark256) | 98 |
| [5'2](#_bookmark257) | [Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à](#_bookmark257) |  |
|  | [frente](#_bookmark257) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 99 |
| [53](#_bookmark258) | [Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à](#_bookmark258) |  |
|  | [frente](#_bookmark258) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 99 |

Lista de Tabelas

|  |  |  |
| --- | --- | --- |
| [1](#_bookmark51) | [Cruzamento de palavras-chave através da aplicação de filtros de ano e de](#_bookmark51) |  |
|  | [linguagem](#_bookmark51) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | '20 |
| ['2](#_bookmark55) | [Fator de impacto](#_bookmark55) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | '2'2 |
| [3](#_bookmark60) | [Areas e seus valores respetivos de artigos em cada área.](#_bookmark60) . . . . . . . . . . . | '25 |
| [4](#_bookmark123) | [Descrição estatística dos dados com o filtro aplicado das 18h às '21h](#_bookmark123) . . . . | 57 |
| [5](#_bookmark140) | [Teste Nemenyi](#_bookmark140) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 66 |
| [6](#_bookmark150) | [Demanda de água](#_bookmark150) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 73 |

1. [Comparação dos modelos com 1 dia de antecedência '24h Treinamento](#_bookmark223) . 84
2. [Comparação dos modelos com 1 dia de antecedência '24h Validação](#_bookmark224) . . . 85

|  |  |  |
| --- | --- | --- |
| [9](#_bookmark225) | [Comparação dos modelos com 1 dia de antecedência '24h Tei,te](#_bookmark225) . . . . . . | 85 |
| [10](#_bookmark226) | [Comparação dos modelos com 1 dia de antecedência '24h Completo](#_bookmark226) . . . | 86 |
| [11](#_bookmark227) | [Comparação dos modelos com 7 dias de antecedência '24h Treinamento](#_bookmark227) . | 86 |
| [1'2](#_bookmark228) | [Comparação dos modelos com 7 dias de antecedência '24h Validação](#_bookmark228) . . . | 87 |
| [13](#_bookmark229) | [Comparação dos modelos com 7 dias de antecedência '24h Tei,te](#_bookmark229) . . . . . . | 87 |
| [14](#_bookmark230) | [Comparação dos modelos com 7 dias de antecedência '24h Completo](#_bookmark230) . . . | 88 |
| [15](#_bookmark231) | [Comparação dos modelos com 14 dias de antecedência '24h Treinamento](#_bookmark231) | 88 |
| [16](#_bookmark232) | [Comparação dos modelos com 14 dias de antecedência '24h Validação](#_bookmark232) . . | 89 |
| [17](#_bookmark233) | [Comparação dos modelos com 14 dias de antecedência '24h Tei,te](#_bookmark233) . . . . . | 89 |
| [18](#_bookmark234) | [Comparação dos modelos com 14 dias de antecedência '24h Completo](#_bookmark234) . . | 90 |
| [19](#_bookmark235) | [Comparação dos modelos com 30 dias de antecedência '24h Treinamento](#_bookmark235) | 90 |
| ['20](#_bookmark236) | [Comparação dos modelos com 30 dias de antecedência '24h Validação](#_bookmark236) . . | 91 |
| ['21](#_bookmark237) | [Comparação dos modelos com 30 dias de antecedência '24h Tei,te](#_bookmark237) . . . . . | 91 |
| ['2'2](#_bookmark238) | [Comparação dos modelos com 30 dias de antecedência '24h Completo](#_bookmark238) . . | 9'2 |
| ['23](#_bookmark240) | [Comparação dos modelos Ljung Box Treinamento](#_bookmark240) . . . . . . . . . . . . . | 9'2 |
| ['24](#_bookmark241) | [Comparação dos modelos Ljung Box Validação](#_bookmark241) . . . . . . . . . . . . . . | 93 |
| ['25](#_bookmark242) | [Comparação dos modelos Ljung Box Tei,te](#_bookmark242) . . . . . . . . . . . . . . . . . | 93 |
| ['26](#_bookmark243) | [Comparação dos modelos Ljung Box Completo](#_bookmark243) . . . . . . . . . . . . . . | 93 |

# Lista de Abreviaturas e Siglas

AdaBoost Impulso ou Estímulo Adaptativo (do inglês Adapt ve Bo- ost n )

AR Auto-Regressivo

ARIMA Média Móvel Integrada Auto-Regressiva (do inglês A'Uto- re ress ve Inte rated Mov n Avera e)

ARIMAX Média Móvel Integrada Auto-Regressiva com Regressores Exógenos (do inglês A'Utore ress ve Inte rated Mov n Ave- ra e w th E o eno'Us Re ressors)

ARMA Média Móvel Auto-Regressiva (do inglês A'Utore ress ve Mov n Avera e)

ARX Auto-Regressivo com Variável Exógena (do inglês A'Utore-

ress ve w th E o eno'Us Inp'Uts ) BrownBoost Algoritmo de Aumento

CNN Rede Neural Convolucional (do inglês Convo 'Ut ona Ne'U- ra Network o'U ConvNet )

DBN Rede de Crenças Profundas (do inglês Deep Be ef Network )

EFB Pacote de Características Exclusivas (do inglês E c 'Us ve

eat'Ure B'Und n )

FT Flow Transmitter (Transmissor de Fluxo) Hz Hertz

INMET Instituto Nacional de Meteorologia LGBMRegressor Regressão Ligth GBM

Light GBM Máquina de Impulso de Gradiente Leve (do inglês L ht

rad ent Boost n Mach ne) LogitBoost Técnicas de Regressão Logística

LPBoost Reforço da Programação Linear (do inglês L near Pro-

ramm n Boost n )

LR Regressão Linear (do inglês L near Re ress on)

LSTM Memória de Longo Curto Prazo (do inglês Lon Short- Term Memory)

*m*3 Metros Cúbicos

*m*3*/h* Metros Cúbicos por Hora

MA Média Móvel (do inglês Mov n Avera e) MadaBoost Modificando o Sistema de Ponderação do AdaBoost

MAE Erro Médio Absoluto (do inglês Mean Abso 'Ute Error ) MAPE Erro Percentual Médio Absoluto (do inglês Mean Abso 'Ute

Percenta e Error )

mca Metros Coluna de Agua

ML Aprendizado de Máquina (do inglês Mach ne Learn n ) mm Milímetros

MSE Erro Médio Quadrático (do inglês Mean Sq'Uared Error ) PR Estado do Paraná

RBAL Recalque Bairro Alto

RFR Regressão de Floresta Aleatória (do inglês Random orest Re ress on)

RMSE Erro de Raiz Média Quadrática (do inglês Root Mean Sq'Uared Error )

RNN Rede Neural Recorrente (do inglês Rec'Urrent Ne'Ura Network )

SANEPAR Companhia de Saneamento do Paraná

SARIMA Auto-Regressivos Integrados de Médias Móveis com Sa- zonalidade (do inglês Seasona A'Uto-Re ress ve Inte rated Mov n Avera es)

SARIMAX Média Móvel Integrada Auto-Regressiva Sazonal com Re- gressores Exógenos (do inglês Seasona A'Uto-Re ress ve In- te rated Mov n Avera e w th E o eno'Us Re ressors)

SVM-VAR Máquinas de Vetor de Suporte - Vetores Auto-Regressivos TotalBoost Impulso Total

XGBoost Impulso Gradiente Extremo (do inglês eXtreme rad ent Boost n )

XGBRegressor Regressão XGBoost

# Sumário

1. [Introdução](#_bookmark0) 1
   1. [Contexto da pesquisa](#_bookmark1) 1
      1. [Motivação da pesquisa](#_bookmark3) 3
2. ['2 Objetivo geral](#_bookmark4) 3
   1. ['2.1 Objetivos específicos e questão de pesquisa](#_bookmark5) 3
   2. [Descrição do problema](#_bookmark16) 4
   3. [Procedimentos metodológicos](#_bookmark17) 5
      1. [Etapas da pesquisa](#_bookmark18) 5
   4. [.Justificativa da pesquisa](#_bookmark28) 9
      1. [Contribuições](#_bookmark29) 9
   5. [Estrutura do trabalho](#_bookmark30) 10
3. [Referencial](#_bookmark32) 12

['2.1 Detecção de anomalias](#_bookmark33) 1'2

['2.'2 Revisão sistemática da literatura](#_bookmark34) 1'2

['2.3 Problematização da Revisão](#_bookmark39) 16

['2.4 Metodologia](#_bookmark45) 17

['2.5 Resultados da busca de revisão](#_bookmark48) 18

['2.6 Principais conclusão](#_bookmark61) '26

1. [Bai,e Teórica](#_bookmark62) 28
   1. [Métricas de Avaliação de Modelos](#_bookmark63) '28
      1. [Erro Quadrático Médio Raiz (RMSE)](#_bookmark64) '28

[3.1.'2 Erro Absoluto Médio (MAE)](#_bookmark66) '29

[3.1.3 Erro Percentual Absoluto Médio (MAPE)](#_bookmark67) 30

[3.'2 Modelos de Séries Temporais Univariados](#_bookmark68) 31

[3.'2.1 Componente Autorregressivo](#_bookmark69) 31

[3.'2.'2 AR(0): Ruído branco](#_bookmark73) 33

[3.'2.3 AR(1): Caminhadas aleatórias e Oscilações](#_bookmark74) 33

[3.'2.4 AR(p): Termos de ordem superior](#_bookmark75) 34

[3.'2.5 Média Móvel](#_bookmark76) 34

[3.'2.6 Modelos ARMA e ARIMA](#_bookmark79) 36

[3.'2.7 ARIMA](#_bookmark81) 37

[3.'2.8 SARIMA](#_bookmark84) 38

[3.3 Modelos de Série Temporal Multivariada](#_bookmark86) 39

|  |  |  |  |
| --- | --- | --- | --- |
|  | [3.3.1 ARIMAX e SARIMAX](#_bookmark87) . . . . . . . . . . . . . . . . . . . . . . . . . | | 39 |
| [3.4](#_bookmark91) | [Modelos de Aprendizado de Máquina Supervisionados](#_bookmark91) . . . . . . . . . . . . | | 40 |
|  | [3.4.1](#_bookmark92)  [3.4.'2](#_bookmark94)  [3.4.3](#_bookmark98) | [Regressão Linear (LR)](#_bookmark92) . . . . . . . . . . . . . . . . . . . . . . . . .  [Definição do modelo](#_bookmark94) . . . . . . . . . . . . . . . . . . . . . . . . . .  [Floresta Aleatória (Random Forest)](#_bookmark98) . . . . . . . . . . . . . . . . . . | 41  4'2  44 |
|  | [3.4.4](#_bookmark101) | [Gradient Boosting (como XGBoost, LightGBM)](#_bookmark101) . . . . . . . . . . . | 45 |
|  | [3.4.5](#_bookmark103) | [O Gradiente em Gradiente de Boosting (Reforço)](#_bookmark103) . . . . . . . . . . | 46 |
|  | [3.4.6](#_bookmark105) | [Algoritmos de boosting de gradiente](#_bookmark105) . . . . . . . . . . . . . . . . . | 47 |
|  | [3.4.7](#_bookmark106) | [A diferença entre XGBoost e LightGBM](#_bookmark106) . . . . . . . . . . . . . . . | 47 |
| [3.5](#_bookmark109) | [Estudo](#_bookmark109) | [de Caso Empírico](#_bookmark109) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 50 |
|  | [3.5.1](#_bookmark110) | [Definição do problema](#_bookmark110) . . . . . . . . . . . . . . . . . . . . . . . . . | 50 |
|  | [3.5.'2](#_bookmark111) | [Coleta de dados](#_bookmark111) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 50 |
|  | [3.5.3](#_bookmark112) | [Análise exploratória dos dados](#_bookmark112) . . . . . . . . . . . . . . . . . . . . . | 50 |
|  | [3.5.4](#_bookmark113) | [Escolha do modelo](#_bookmark113) . . . . . . . . . . . . . . . . . . . . . . . . . . . | 51 |
|  | [3.5.5](#_bookmark114) | [Divisão dos dados](#_bookmark114) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 51 |
|  | [3.5.6](#_bookmark115) | [Ajuste do modelo](#_bookmark115) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 51 |
|  | [3.5.7](#_bookmark116) | [Avaliação do modelo](#_bookmark116) . . . . . . . . . . . . . . . . . . . . . . . . . . | 53 |
|  | [3.5.8](#_bookmark117) | [Previsões Futuras](#_bookmark117) . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 54 |
|  | [3.5.9](#_bookmark118) | [Monitoramento e Ajuste Contínuo](#_bookmark118) . . . . . . . . . . . . . . . . . . . | 54 |
|  | [3.5.10](#_bookmark119) | [Principais Conclusão](#_bookmark119) . . . . . . . . . . . . . . . . . . . . . . . . . . | 55 |
| [4 Rei,ultadoi,](#_bookmark120) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 56 |
| [4.1 Planejamento do Problema](#_bookmark121) . . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 56 |
| [4.1.1 Análise Exploratória dos dados (EDA)](#_bookmark122) . . . . . . . . . . . . . . . . | | | 56 |
| [4.1.'2 Múltiplas entradas e saída única (MISO)](#_bookmark124) . . . . . . . . . . . . . . . | | | 57 |
| [4.1.3 Decomposição STL](#_bookmark125) . . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 58 |
| [4.1.4 Separação dos dados](#_bookmark134) . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 63 |
| [4.1.5 Estratégia de Previsão](#_bookmark136) . . . . . . . . . . . . . . . . . . . . . . . . . | | | 65 |
| [4.1.6 Horizonte](#_bookmark137) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 65 |
| [4.1.7 Modelos de previsão e métricas de desempenho](#_bookmark138) . . . . . . . . . . . | | | 65 |
| [4.1.8 Teste de Significância](#_bookmark139) . . . . . . . . . . . . . . . . . . . . . . . . . . | | | 66 |
| [4.1.9 Comparação dos modelos](#_bookmark141) . . . . . . . . . . . . . . . . . . . . . . . . | | | 67 |
| [4.'2 Estudo de Caso Empírico Resultado](#_bookmark144) . . . . . . . . . . . . . . . . . . . . . . | | | 68 |
| [4.'2.1 Descrição do sistema de abastecimento de água](#_bookmark145) . . . . . . . . . . . | | | 69 |
| [4.'2.'2 Análise exploratória dos dados](#_bookmark146) . . . . . . . . . . . . . . . . . . . . . | | | 70 |
| [4.'2.3 Questões de pesquisa 1 a 4](#_bookmark147) . . . . . . . . . . . . . . . . . . . . . . . | | | 71 |
| [4.'2.4 Questão de pesquisa 5](#_bookmark148) . . . . . . . . . . . . . . . . . . . . . . . . . | | | 7'2 |
| [4.'2.5 Discussão geral e conclusões](#_bookmark151) . . . . . . . . . . . . . . . . . . . . . . | | | 74 |

1. [Conclui,õei,](#_bookmark152) 76
   1. [Limitações da Pesquisa e Propostas Futuras](#_bookmark153) 77

[Referênciai,](#_bookmark153) 78

1. [Apêndice - Comparação doi, modeloi, de previi,ão de i,eriei, temporaii, média de 24h](#_bookmark222) 84
2. [Apêndice - Comparação doi, modeloi, de previi,ão com o método Ljung-](#_bookmark239)

[Box](#_bookmark239) 92

1. [Apêndice - Modeloi, AR(7), ARX (7) e MA (7) 24h](#_bookmark244) 94
2. [Apêndice - Modeloi, ARMA(7,7) e ARIMA (7,1,7) 24h](#_bookmark249) 96
3. [Apêndice - Modeloi, ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e](#_bookmark254)

[SARIMAX (7,1,7) (2,1,0,12) 24h](#_bookmark254) 98

# Introdução

A previsão precisa da demanda de água é fundamental para um planejamento efi- ciente e sustentável do abastecimento hídrico em uma determinada região. Neste estudo, são empregados métodos avançados, como Gradiente, Regressão e ARIMA, para realizar previsões precisas da demanda diária de água. Os resultados obtidos fornecem insights valiosos e contribuem para um melhor entendimento das tendências e padrões de con- sumo de água, permitindo um planejamento mais eficaz para atender às necessidades da população.

Este capítulo apresenta o conteúdo abordado nesta dissertação, que se concentra na utilização de modelos de Aprendizado de Máquina (ML) para prever futuramente os dados coletados pela SANEPAR. Os dados coletados referem-se ao abastecimento de água no bairro Alto durante o período de '2018 a '20'20, quando ocorreu uma escassez que afetou toda a população da capital paranaense.

Dentro do contexto de análise de séries temporais e tomada de decisão, são explora- dos modelos de ML para aplicação nesses dados. Por meio de uma revisão sistemática da literatura, são identificados e tabulados os modelos clássicos mais comumente utilizados para análise de séries temporais.

Nesta dissertação, busca-se desenvolver previsões precisas e confiáveis para o abas- tecimento de água no bairro Alto. Com base nas informações coletadas e na aplicação dos modelos, espera-se obter insights valiosos para auxiliar na tomada de decisões estratégicas e no planejamento eficiente do abastecimento hídrico na região.

Além da revisão da literatura, serão apresentados os métodos e técnicas utilizados para a análise dos dados, bem como os resultados obtidos. Busca-se contribuir significa- tivamente para a área de análise de séries temporais aplicada ao abastecimento de água, permitindo uma melhor compreensão dos padrões de consumo e aprimorando a eficiência dos processos de tomada de decisão relacionados ao fornecimento de água no bairro Alto.

* 1. Contexto da pesquisa

[Ribeiro et al.](#_bookmark194) (['20'21](#_bookmark194)) A necessidade de desenvolvimento do planejamento estratégico no mundo corporativo e no dia-a-dia torna a análise de séries temporais e previsões valiosas ferramentas para apoiar o processo de tomada de decisão a curto, médio e longo prazo. Devido a não linearidades, sazonalidade, tendência e ciclicidade nos dados temporais, o desenvolvimento de modelos de previsão eficientes é uma tarefa desafiadora.

No conjunto de dados da SANEPAR, há um volume significativo no consumo de água e, com as interrupções que a cidade tem enfrentado, é necessário analisar os

dados para compreender melhor os padrões de interrupção no abastecimento e os picos de consumo ao longo das horas e dias.

Nesta dissertação, será realizada uma revisão sistemática de modelos preditivos para avaliar o melhor modelo que pode ser utilizado e como ele pode ser validado para prever a escassez de água. Essas análises serão feitas utilizando a linguagem de progra- mação Python.

A abordagem deste trabalho consiste em explorar o conceito de séries temporais e sua aplicação no campo do aprendizado de máquina. Os dados de séries temporais referem- se a dados coletados e armazenados ao longo do tempo, permitindo que observadores identifiquem anomalias nos dados. A classificação dos dados por ano ou dia é essencial na análise de séries temporais, e se os dados forem atribuídos aleatoriamente, pode ser mais desafiador fazer previsões e tomar decisões com base nos dados coletados.

É importante destacar que a análise de médias pode ser enganosa se não forem excluídos os valores discrepantes, também conhecidos como "o'Ut ers". Esses valores dis- crepantes podem levar a resultados extremamente altos ou baixos que não reftetem a realidade.

O campo do aprendizado de máquina abrange várias áreas, conforme ilustrado na Figura [1](#_bookmark2). Serão explorados os diferentes componentes do aprendizado de máquina e como eles podem ser aplicados em diversos contextos.

Figura 1: Paradigma de aprendizado de máquina

Aprendizado de Máquina

Aprendizado por reforço

Aprendizado não supervisionado

Aprendizado Supervisionado

Agrupamento

Regressão

Classificação

IA de jogo

Monitoramento por drone

Redução da dimensão

Previsão de séries temporais

Previsão

de falha

Segmentação de grupos

Agregação de mercados

Detecção de falência

Identificação de doenças

Seleção de recursos

Visualização de Big Data

Fonte: Elaboração própria

* + 1. Motivação da pei,quii,a

De acordo com ([VASCONCELOS](#_bookmark213), ['20'20](#_bookmark213)) Curitiba e região metropolitana enfrentou um rodízio com 36 horas com água e 36 horas sem abastecimento. A média geral dos reservatórios da região está em 27*,* 96% da capacidade. Assim em medida a isso essa pesquisa tem como a abordagem da falta de água, essa falta que pode ser vista como uma seca, em média nos anos anteriores de '20'20 a chuva tem marcado a quantia de 1*.*704 mm. ([VASCONCELOS](#_bookmark213), ['20'20](#_bookmark213)) Desde '2016, quando registrou 1.704 mm de chuva, Curitiba não atingiu mais a média anual de precipitação, que é de 1.490 mm, com base em dados da estação pluviométrica do IBMET. Apesar de abaixo da média, o mínimo registrado desde então ocorreu em '20'20, com 1.158 mm.

Em meio a essa motivação, é possível realizar uma análise mais aprofundada dos dados fornecidos pela SANEPAR, a fim de prever e evitar a ocorrência de escassez de água, que foi registrada juntamente com a anomalia detectada em '20'20. Com o retorno das chuvas, houve um aumento no nível dos reservatórios.

* 1. Objetivo geral

O objetivo desta pesquisa é identificar o melhor modelo de séries temporais para abordar o problema da escassez de água que ocorreu em Curitiba. Ao longo da dissertação, foram avaliados diversos modelos de regressão, com foco especial nos modelos baseados em

rad ent boost n , considerados eficazes na literatura para a previsão de séries temporais. Os principais modelos explorados incluem o ARIMA e suas variantes atualizadas. Além da previsão, também serão realizadas análises de anomalias nos dados, visando compreender as causas subjacentes a essas ocorrências.

* + 1. Objetivoi, ei,pecífcoi, e quei,tão de pei,quii,a

Neste estudo, busca-se identificar e compreender possíveis anomalias nos dados, bem como investigar as causas por trás dessas ocorrências. O objetivo é responder às perguntas de pesquisa relacionadas a essas anomalias.

Q 1 Qual é a adequação da pressão atual para atender à demanda diária?

Q '2 Qual é o volume mínimo de água necessário no reservatório para evitar o aciona- mento das bombas durante o horário de pico?

Q 3 Qual é a vazão ótima para atender à demanda diária?

Q 4 Como encontrar o ponto de equilíbrio entre a demanda e a vazão?

Q 5 Qual é o impacto do acionamento das bombas durante o horário de pico?

* + - 1. Qual é o nível ideal no reservatório para evitar a ativação das bombas da SANE- PAR durante o período de maior demanda, das 18h às '21h, sem comprometer o abastecimento de água para a população? Além disso, como variam as médias das vazões nos horários críticos (18h às '21h) para as diferentes estações do ano (Outono, Inverno, Primavera, Verão)?
      2. Existe tendência, padrão, sazonalidade para os dados destes três anos do Bairro Alto?
      3. Identificar quais os horários de maior demanda das 18 às 21?
      4. Quanto tenho que armazenar previamente no reservatório para não acionar as bombas no horário de pico?
      5. Se a vazão cresce e a pressão decresce temos uma ANOMALIA na rede (com base no histórico).
  1. Descrição do problema

A descrição do problema é fundamental para obter uma compreensão mais precisa do que está sendo abordado neste trabalho. É por meio dessa descrição que as variáveis- chave são expostas e o objetivo da previsão é estabelecido de forma clara. Sem um plano estruturado para determinar o que deve ser previsto, torna-se difícil justificar o uso de modelos de previsão de dados. Portanto, é essencial estabelecer um propósito claro e definir as metas da previsão antes de aplicar os modelos adequados.

* Bombas de sucção (B1, B'2 e B3) - valor máximo da frequência 60 Hz Variáveis importantes: Fluxo, pressão e nível
* Nível do Reservatório (Câmara 1) LT01 (*m*3) - PREVER
* Vazão de entrada (FT01) (*m*3*/h*)
* Vazão de gravidade (FT0'2) (*m*3*/h*)
* Vazão de recalque (FT03) (*m*3*/h*)
* Pressão de Sucção (PT01SU) (mca)
* Pressão de Recalque (PT0'2RBAL) (mca)

A pesquisa fará uso da variável LT01, que representa o nível do reservatório e de- sempenha um papel de extrema importância, como evidenciado pelas Figuras [4](#_bookmark35) e [5](#_bookmark36). Essas figuras retratam as anomalias ocorridas durante o período em que a capital paranaense foi afetada pela escassez de chuvas, resultando na redução do nível dos reservatórios e na implementação de rodízios periódicos, conforme discutido na subseção [1.1.1](#_bookmark3). Assim, tais observações permitem uma compreensão mais aprofundada das perspectivas futuras.

* 1. Procedimentos metodológicos

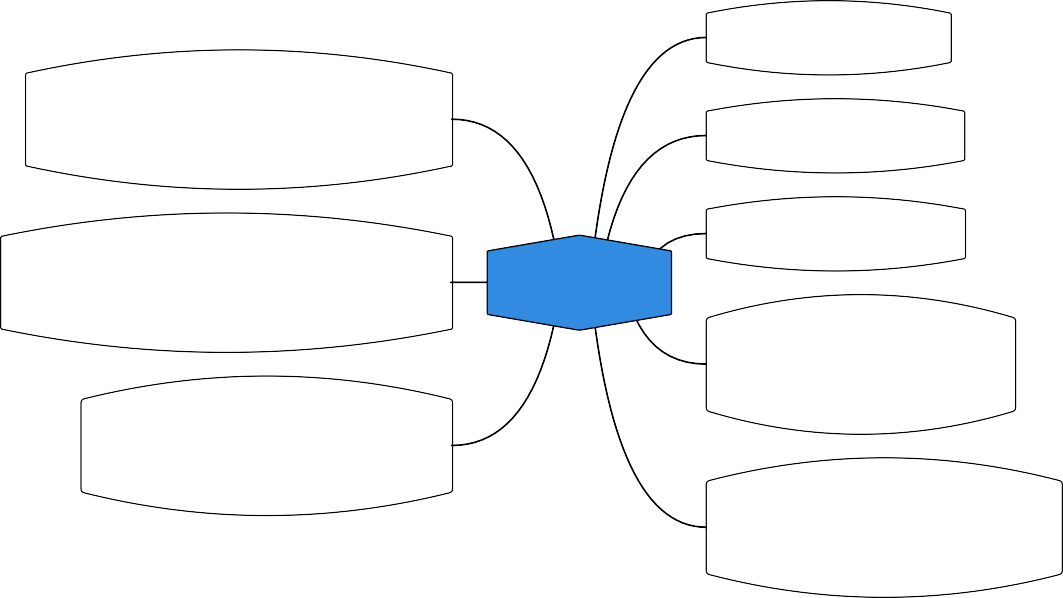
Com o intuito de realizar previsões e fazer comparações entre os modelos obtidos na revisão sistemática, será adotado um processo metodológico bem definido. Tal processo está detalhado na subseção [1.4.1](#_bookmark18) desta seção, onde foram estabelecidas as etapas a serem seguidas desde o início. Isso inclui a definição do que será previsto, bem como a seleção dos métodos a serem utilizados na Análise Exploratória de Dados (EDA), seguindo uma sequência lógica e coerente.

* + 1. Etapai, da pei,quii,a

A pesquisa seguiu as seguintes etapas:

**Etapa 8: Testes de**

Figura '2: Mapa das Etapas

**Etapa 1: EDA**

**significância estatística**

**Etapa 2: MISO**

**Etapa7: Modelos de previsão e métricas de desempenho**

**Etapa 6: Horizonte de previsão**

**Etapas**

**Etapa 3: STL**

**Etapa 4: Separação dos dados**

**Etapa 5: Estratégia de previsão**

Fonte: Elaboração própria

Etapa 1 Análise exploratória dos dados - EDA ( do inglês E p oratory Data Ana ys s).

A exploração de dados na EDA é fundamental para entender melhor os dados que estão sendo trabalhados, como, por exemplo, excluir valores ausentes, saber como

os dados estão separados em horas ou dias e, assim, tomar a melhor decisão a ser trabalhada com os dados, usar gráficos de linha na análise para observar a convergência dos dados e as anomalias que podem ocorrer.

Etapa '2 O que vai ser usado como variáveis previsoras e qual será a variável a ser predita (MISO).

Nessa etapa, tem o papel de relacionar as variáveis ao que será previsto, como os modelos de variáveis exógenas que são usados aqui nos modelos SARIMAX, ARX e ARIMAX do tipo ARIMA. Cada modelo tem a interação de mais variáveis do que o modelo ARIMA básico ou seus derivados AR, MA e SARIMA. O conhecimento de quais variáveis estão incluídas na modelagem do problema torna a modelagem mais abrangente quando o horizonte de previsão é estendido além dos dados.

Etapa 3 Fazer a decomposição STL (do inglês Seasona -Trend Decompos t on) Sazonalidade, Tendência e Resíduo.

[Ma, Yu e Zhang](#_bookmark184) (['20'2'2](#_bookmark184)) destacam que "o uso do método de decomposição STL em conjunto com um modelo híbrido se mostrou eficaz na previsão de carga de curto prazo" (p. 6).

Segundo [Huo et al.](#_bookmark173) (['2019](#_bookmark173)), "a aplicação da decomposição STL e do modelo LSTM baseado em atenção foi capaz de prever com precisão a velocidade do tráfego de curto prazo" (p. 6).

[Shen et al.](#_bookmark202) (['20'21](#_bookmark202)) afirmam que "a combinação da decomposição STL e modelos de aprendizado profundo mostrou-se promissora na previsão de carga de eletricidade" (p. 18).

O algoritmo STL executa suavização na série de tempo usando LOESS em dois loops; o loop interno itera entre a suavização sazonal e de tendência e o loop externo minimiza o efeito de valores atípicos. Durante o loop interno, o componente sazonal é calculado primeiro e removido para calcular o componente de tendência. O restante é calculado subtraindo os componentes sazonais e de tendência da série de tempo.

Os três componentes da análise STL se relacionam com a série de tempo bruta da seguinte forma:

*yi* = *si* + *ti* + *ri* (1)

Onde:

* *yi* = *O* valor da série de tempo no ponto *i*.
* *si* = *O* valor do componente sazonal no ponto *i*.
* *ti* = *O* valor do componente de tendência no ponto *i*.
* *ri* = *O* valor do componente restante no ponto *i*.

Etapa 4 Separação dos dados.

De acordo com a literatura acadêmica, é comum dividir o conjunto de dados em treinamento, validação e teste para avaliar a performance dos modelos. Essa abor- dagem permite uma análise mais completa do desempenho do modelo, garantindo uma avaliação objetiva de sua capacidade de generalização e evitando problemas de overfitting ou subajuste ([CRUZ-RAMÍREZ et al.](#_bookmark168), ['20'20](#_bookmark168); [MOKHTARI et al.](#_bookmark187), ['20'20](#_bookmark187); [KHAN et al.](#_bookmark176), ['20'21](#_bookmark176); [SHARMA et al.](#_bookmark201), ['20'21](#_bookmark201)).

A fim de obter uma divisão mais adequada dos dados, é realizado um estudo das medidas de tendência central e dispersão de cada conjunto. O conjunto de dados é então dividido em três partes distintas: treinamento, validação e teste. Nessa divisão, inicialmente, 70% dos dados são utilizados para o treinamento e validação, enquanto os 30% restantes são reservados para o conjunto de teste. Em seguida, a porção destinada ao treinamento e validação é subdividida em uma proporção de 80% para treinamento e '20% para validação.

Etapa 5 Estratégia de previsão (recursiva e iterada-método direto).

A estratégia recursiva é mencionada por [Petropoulos et al.](#_bookmark191) (['20'2'2](#_bookmark191)) como uma abor- dagem eficaz na previsão de séries temporais de múltiplos passos. De acordo com o autor, essa estratégia envolve o uso de previsões anteriores como entradas para prever os próximos passos da série temporal. A abordagem recursiva tem demons- trado potencial para melhorar a acurácia das previsões de séries temporais de longo prazo.

A estratégia recursiva consiste em utilizar um modelo de previsão de um passo de tempo várias vezes, onde a previsão obtida no passo anterior é utilizada como entrada para realizar a previsão do próximo passo de tempo.

No contexto da previsão da demanda de água para os próximos dias, seria desen- volvido um modelo de previsão de um único passo. Esse modelo seria aplicado para prever a demanda no primeiro dia e, em seguida, essa previsão seria utilizada como dado de entrada para prever a demanda do segundo dia. Esse processo se repetiria para os demais dias, permitindo a previsão da demanda ao longo do tempo.

Por Exemplo:

*preditivo*(*t* + 1) = *model*1(*obs*(*t* − 1)*, obs*(*t* − 2)*, . . . , obs*(*t* − *n*)) ('2)

*preditivo*(*t* + 2) = *model*2(*obs*(*t* − 2)*, obs*(*t* − 3)*, . . . , obs*(*t* − *n*)) (3)

[Brownlee](#_bookmark160) (['2016](#_bookmark160)) como as previsões são usadas no lugar das observações, a estra- tégia recursiva permite que os erros de previsão se acumulem de tal forma que o desempenho possa se degradar rapidamente à medida que o horizonte de tempo de previsão aumenta.

Etapa 6 Horizonte de previsão (1 passo ou n passos à frente).

Para abordar a diversidade de horizontes de previsão, optou-se por considerar dife- rentes intervalos de tempo. Isso permitirá a realização de previsões para um passo à frente, uma semana, duas semanas e um mês, de forma a abranger distintas perspec- tivas de curto e médio prazo. Essa abordagem proporciona uma análise abrangente da capacidade dos modelos em lidar com horizontes de previsão variados, contri- buindo para uma avaliação mais completa e precisa do desempenho dos mesmos.

Etapa 7 Modelos de previsão e métricas de desempenho.

Os modelos abordados nesta pesquisa são tanto os modelos clássicos de previsão quanto os modelos de regressão por gradiente. Entre os modelos clássicos, incluem-se o AR, ARX, ARMA, ARIMA, SARIMA, SARIMAX e ARIMAX, enquanto os mo-

delos de regressão por gradiente englobam o LR, XGBRegressor, RFR e LGBMRe- gressor. A seleção desses modelos foi baseada em uma revisão sistemática realizada durante a dissertação, buscando identificar os modelos mais eficazes e amplamente utilizados na literatura.

Ao longo da pesquisa, foram adotadas três métricas principais para avaliar o desem- penho dos modelos: RMSE, MAE e MAPE. Essas métricas foram escolhidas com base na revisão sistemática e são amplamente reconhecidas como medidas de quali- dade de previsão. Cada métrica tem sua própria interpretação e importância, sendo detalhada na subseção [3.1](#_bookmark63) para um melhor entendimento de como são aplicadas e interpretadas na pesquisa.

Etapa 8 Aplicar os modelos de previsão e fazer comparativo baseado em testes de significân- cia estatística ( r edman e Nemeny ).

O teste de Friedman é o teste não paramétrico usado para comparar dados de amostras vinculadas, ou seja, quando o mesmo indivíduo é avaliado mais de uma vez. ou seja, quando o mesmo indivíduo é avaliado mais de uma vez. O teste de Friedman

não usa os dados numéricos diretamente, mas sim as classificações ocupadas pelos dados após a classificação de cada grupo separadamente. separadamente. Após a classificação, a hipótese de igualdade da soma das classificações de cada grupo é testada.

O teste consiste em fazer comparações em pares com o intuito de verificar qual dos fatores que diferem entre si. No entanto, o teste de Nemenyi é muito conservador e pode não encontrar diferença significativa entre os pares testados.

* 1. Justificativa da pesquisa

Ao longo desta dissertação, os seguintes aspectos são abordados visando a previsão e tomada de decisões adequadas para evitar a ocorrência futura de escassez de água.

* + 1. Contribuiçõei,

Após as perguntas de pesquisa apresentadas na subseção [1.'2.1](#_bookmark5), surgem duas contri- buições significativas nesta dissertação. A primeira diz respeito à previsão da demanda de água na cidade de Curitiba, abordando aspectos como consumo e gasto de energia durante períodos de pico ([SMITH; .JOHNSON](#_bookmark205), ['20'2'2](#_bookmark205)) ([BROVN; LEE](#_bookmark159), ['20'21](#_bookmark159)). Segundo estudos recentes, os modelos ARIMA desempenham um papel fundamental na análise de séries temporais ([SMITH; .JOHNSON](#_bookmark205), ['20'2'2](#_bookmark205)). De acordo com pesquisas, os modelos ARIMA são amplamente utilizados na previsão de séries temporais devido à sua capacidade de cap- turar padrões complexos e comportamentos de longo prazo ([SMITH; .JOHNSON](#_bookmark205), ['20'2'2](#_bookmark205)). Conforme relatos, o modelo XGBoost tem sido aplicado com sucesso em problemas de previsão de séries temporais ([BROVN; LEE](#_bookmark159), ['20'21](#_bookmark159)). Estudos demonstraram que o XG- Boost é uma poderosa ferramenta para lidar com desafios de previsão em séries temporais ([BROVN; LEE](#_bookmark159), ['20'21](#_bookmark159)). De acordo com especialistas, o LightGBM tem ganhado destaque como um modelo eficiente para previsão de séries temporais ([GARCIA; RODRIGUEZ](#_bookmark170), ['20'23](#_bookmark170)). Pesquisas recentes destacam o desempenho promissor do LightGBM na análise e previsão de séries temporais ([GARCIA; RODRIGUEZ](#_bookmark170), ['20'23](#_bookmark170)). De acordo com [.Johnson](#_bookmark174) [e Smith](#_bookmark174) (['20'2'2](#_bookmark174)), o uso de regressão linear é fundamental para a modelagem preditiva. [Anderson e Villiams](#_bookmark155) (['20'21](#_bookmark155)) destacam a importância do uso de random forest regression na previsão de séries temporais.

Nesse sentido, foram utilizados métodos de previsão de séries temporais, como os modelos ARIMA, ARMA, SARIMA, ARIMAX e SARIMAX, bem como modelos mais simples derivados do ARIMA, como AR, ARX e MA. Além disso, foram explorados mo- delos regressivos, como LR e RFR, e modelos baseados em gradientes, como XGBoost e LightGBM. Essa variedade de modelos foi selecionada visando uma previsão precisa

e eficiente, levando em consideração as demandas relacionadas ao consumo de energia e água pela empresa SANEPAR, com o objetivo de minimizar os gastos associados.

As previsões foram realizadas tanto para o curto prazo (1 a 7 dias) quanto para o longo prazo (14 a 30 dias), a fim de embasar a tomada de decisões estratégicas em relação à demanda de água. Os resultados destacaram que, no longo prazo, os modelos ARIMA tiveram um desempenho superior em comparação aos modelos baseados em gradientes. Por outro lado, os modelos de gradiente mostraram-se mais eficazes nas previsões de curto prazo, como para um dia ou uma semana. Ainda assim, os modelos ARIMA e seus derivados superaram os modelos baseados em gradientes.

A comparação entre os modelos de previsão desempenha um papel central nesta dissertação. Através do teste estatístico Ljung-Box, é possível avaliar o desempenho de cada modelo ARIMA tanto no curto prazo quanto no longo prazo. No Apêndice [B](#_bookmark239), apresenta-se a comparação dos modelos por meio desse teste estatístico. Além disso, nas Figuras [39](#_bookmark142) e [40](#_bookmark143) do Apêndice [A](#_bookmark222), é realizada a comparação dos modelos regressivos com os modelos ARIMA. Essas análises comparativas são cruciais para a seleção do modelo mais adequado, permitindo uma tomada de decisão embasada para enfrentar o problema em questão.

* 1. Estrutura do trabalho

O trabalho está estruturado em diferentes capítulos, cada um abordando aspectos específicos da pesquisa. O Capítulo [1](#_bookmark0), Introdução, apresenta a introdução do trabalho, fornecendo uma contextualização do estudo, destacando a motivação e os objetivos a serem alcançados. Também são apresentados o problema em questão, a metodologia utilizada, a justificativa da pesquisa, as contribuições esperadas e a organização do trabalho.

O Capítulo ['2](#_bookmark32), Revisão Teórica, oferece uma visão geral das principais pesquisas e estudos relacionados às questões abordadas na pesquisa. Esse capítulo proporciona uma base teórica sólida para fundamentar a análise e interpretação dos resultados.

No Capítulo [3](#_bookmark62), são apresentados os modelos que serão utilizados para trabalhar com os dados coletados. Essa seção detalha os modelos escolhidos, destacando suas ca- racterísticas e fundamentos teóricos. Além disso, é realizado o detalhamento do estudo de caso utilizado na dissertação.

O Capítulo [4](#_bookmark120), Resultados, apresenta os resultados obtidos ao longo da pesquisa. Nesta seção, são realizadas análises e interpretações dos resultados, fornecendo insights relevantes para o entendimento do problema em estudo. Os resultados do estudo de caso são detalhados, evidenciando as principais descobertas e conclusões obtidas.

Por fim, o Capítulo [5](#_bookmark152), Conclusões, traz as considerações finais da pesquisa, abor-

dando os principais achados e conclusões alcançadas. Também são apresentadas propostas para pesquisas futuras, visando expandir e aprofundar o conhecimento na área.

Essa estrutura organizada em capítulos permite uma apresentação clara e coerente do trabalho, abrangendo desde a introdução e fundamentação teórica até os resultados e conclusões finais.

Este documento está estruturado em [5](#_bookmark152) capítulos, divididos da seguinte forma: Figura 3: Estrutura da dissertação

Explorando a Eficiência dos Modelos de Previsão de Séries

Temporais no Abastecimento de Água

Introdução

Estrutura do trabalho

Referencial

Base Teórica

Seção presente

Resultados

Conclusões

Fonte: Elaboração própria

# Referencial

Este capítulo apresenta o referencial teórico que serviu de base para a elaboração desta dissertação. Embora os resultados obtidos possam ser considerados mais modestos em comparação a uma tese, eles ainda são relevantes para o trabalho realizado aqui. A revisão bibliográfica realizada consiste em uma análise abrangente e crítica das principais fontes de literatura relacionadas ao tema em questão. Por meio dessa revisão, busca-se obter uma compreensão aprofundada do estado atual do conhecimento na área e identificar lacunas ou oportunidades de pesquisa. Os insights e informações extraídos da literatura são fundamentais para embasar a fundamentação teórica, a metodologia e a análise dos resultados desta dissertação. Dessa forma, a revisão bibliográfica desempenha um papel crucial no embasamento teórico e na contextualização do trabalho, fornecendo um sólido alicerce para o desenvolvimento e contribuição desta pesquisa.

* 1. Detecção de anomalias

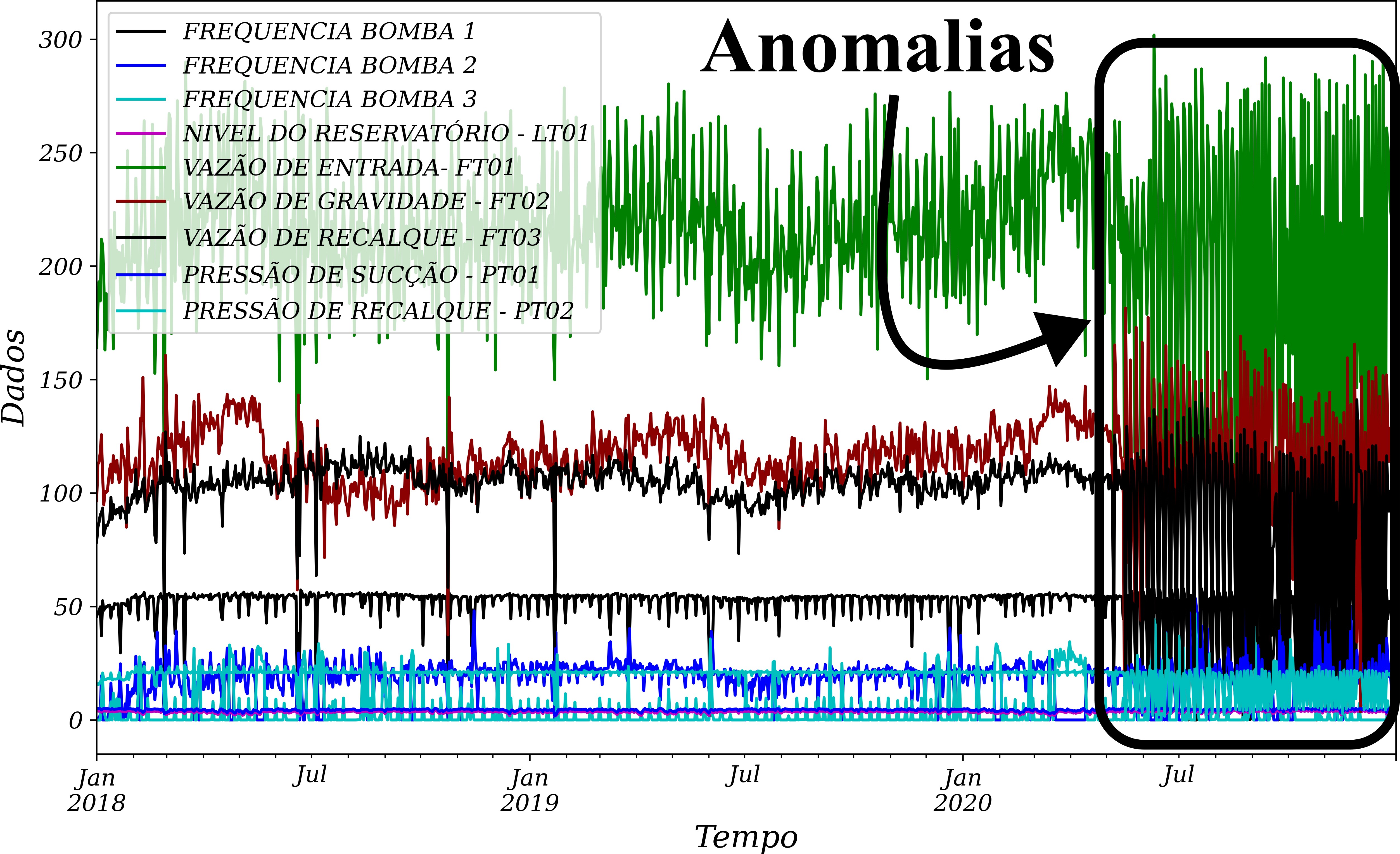
A detecção de anomalias em séries temporais representa um desafio significativo para os previsores, pois requer habilidade em identificar mudanças nos dados, mesmo quando não estão claramente evidentes. Nesse contexto, a coleta de dados realizada ao longo do tempo pela empresa SANEPAR revela anomalias mais expressivas do que ini- cialmente imaginado. A escassez de água que afetou a cidade de Curitiba se prolongou por vários dias, como é evidenciado pelos gráficos de linha utilizados na etapa de traba- lho mencionada ([Etapa 1](#_bookmark20)). Esses gráficos oferecem uma representação visual clara das variações nos níveis de água ao longo do tempo, auxiliando na compreensão da extensão do problema e na necessidade de uma abordagem adequada.

As Figuras [4](#_bookmark35) e [5](#_bookmark36) apresentadas ilustram visualmente as variações e padrões obser- vados nos dados ao longo do tempo, destacando a importância de explorá-los de maneira apropriada a fim de compreender as anomalias e embasar a tomada de decisões. Os dados coletados possuem uma dimensão de 26*.*306 linhas e 9 colunas, e essa ampla quantidade de dados será utilizada nos modelos descritos na subseção mencionada para que seja pos- sível prever e analisar as anomalias evidenciadas. Essas análises permitirão uma melhor compreensão das anomalias e orientarão as decisões tomadas.

* 1. Revisão sistemática da literatura

As séries temporais desempenham um papel fundamental em diversos campos do conhecimento, como Economia (preços diários de estoques, taxa de desemprego mensal, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia

Figura 4: Dados completos com uma frequência média de '24 horas



Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 5: Plotagem de dados para o ano de '20'20

*300*

*FREQUENCIA BOMBA 1 FREQUENCIA BOMBA 2 FREQUENCIA BOMBA 3*

*NIVEL DO RESERVATÓRIO - LT01*

*VAZÃO DE ENTRADA- FT01 VAZÃO DE GRAVIDADE - FT02 VAZÃO DE RECALQUE - FT03*

*PRESSÃO DE SUCÇÃO - PT01 PRESSÃO DE RECALQUE - PT02*

*250*

*200*

*150*

*Dados*

*100*

*50*

*0*

*Jan 2020*

*Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec*

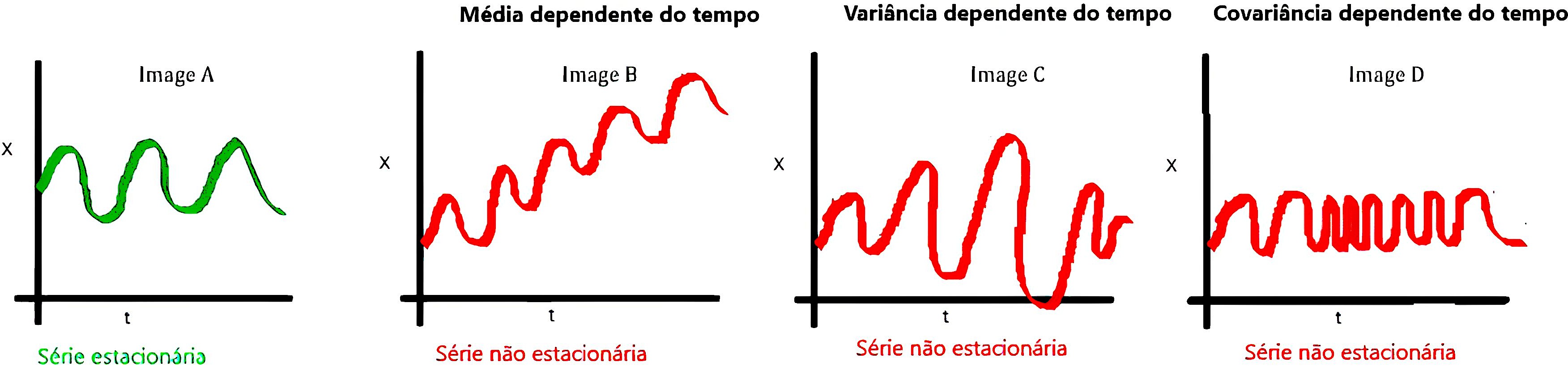
*Tempo*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

(número mensal de novos casos de meningite), Meteorologia (chuvas, temperatura diária, velocidade do vento), entre outros. Ao longo dos anos, têm sido empregadas ferramentas computacionais para tornar a previsão em séries temporais mais eficiente, especialmente com o uso de técnicas de aprendizado de máquina e linguagens de programação como Python e R, que se destacam por sua capacidade de manipular e analisar dados temporais de forma eficaz.

Para compreender melhor o conceito de série temporal, é possível considerar o exemplo de um maratonista que pratica corrida regularmente ao longo de vários anos e uma pessoa sedentária que decide participar de uma corrida com uma distância máxima de 5 km. Ambos realizam a corrida ao mesmo tempo, utilizando monitores de frequência cardíaca que permitem o acompanhamento médico. Ao analisar os dados desde o início até o final da corrida, é possível observar que a série temporal do maratonista apresenta um comportamento mais estacionário, devido ao seu hábito regular de corrida. Por outro lado, a série temporal da pessoa sedentária é mais não estacionária, como ilustrado na Figura [6](#_bookmark37). Essa diferença ocorre devido à falta de regularidade na prática de exercícios físicos por parte da pessoa sedentária.

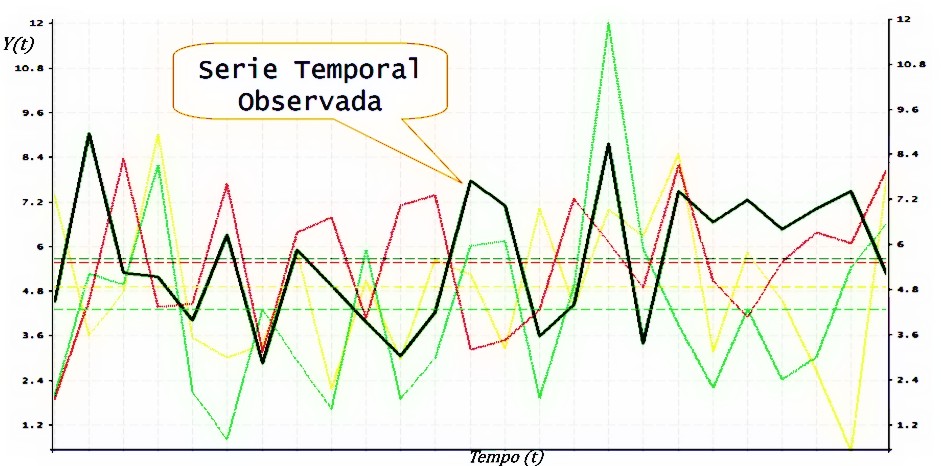
Figura 6: Exemplo de séries temporais



Fonte: ([BRANDAO](#_bookmark158), ['20'20](#_bookmark158))

Na Figura [6](#_bookmark37), é possível observar que o eixo *x* representa os dados observados ao longo do tempo, enquanto o eixo *t* representa o tempo decorrido. Além disso, as séries temporais são caracterizadas como processos estocásticos regidos por leis probabilísticas. Isso implica que elas podem ser concebidas como um conjunto de todas as possíveis trajetórias que uma variável alvo pode seguir, como ilustrado na Figura [6](#_bookmark37). No entanto, somente uma dessas trajetórias será observada, de acordo com as características que se manifestaram durante o período analisado. Por exemplo, ao lançar um dado, existem seis possibilidades, mas apenas um número será obtido. Da mesma forma, em séries temporais, há uma infinidade de possibilidades, mas somente uma delas ocorrerá, de acordo com as características que se apresentaram nesse determinado período.

Figura 7: Processo estocástico



Fonte: ([PINHEIRO](#_bookmark192), ['20'2'2](#_bookmark192))

Com *Y* (*t*) representando os dados fictícios e *Tempo* (*t*) representando a linha do tempo na Figura [6](#_bookmark37).

É possível pensar nisso como um conjunto de todas as trajetórias possíveis que poderiam ser observadas para uma variável.

Esta revisão sistemática da literatura aborda o tema das séries temporais, que é de grande relevância em diversas áreas, como ilustrado na Figura [16](#_bookmark59). Foi realizada uma análise das últimas seis anos para identificar as principais realizações nesse campo dentro desse curto período de tempo disponível. A seleção dos artigos foi baseada

em critérios específicos, levando em consideração a relevância dos autores, os anos de atividade, os países com maior número de publicações e as palavras-chave mais fre- quentes.

O objetivo dessa revisão é analisar uma literatura selecionada, porém altamente relevante. Embora a série temporal tenha como foco a análise e modelagem da dependên- cia temporal, considerando a ordem apresentada nas bases de dados, os artigos revisados também exploram o uso de técnicas de aprendizado de máquina em aplicações relaciona- das.

Embora nem todos os artigos revisados tenham uma forte relação com aprendizado de máquina, eles contribuem cientificamente para este trabalho e podem servir como base para outros pesquisadores. Essas análises fornecem uma visão básica para alguns leitores que ainda não estão familiarizados com o conceito de séries temporais ou revisões sistemáticas da literatura.

* 1. Problematização da Revisão

Nesta subseção, é discutido um problema de pesquisa que pode ser compreendido por diversos leitores. A Figura [8](#_bookmark40) apresenta um mapa conceitual das publicações, des- tacando a importância dos autores como base para esta revisão. Os modelos propostos por esses autores são fundamentais para abordar o problema em questão, uma vez que a previsão em séries temporais é um desafio de grande significado por si só.

Figura 8: Mapa conceitual do problema de pesquisa

**Obras mais influentes**

**Série temporal**

**Publicação**

**Áreas**

**Tema**

**Autores**

**Previsão Previsão não linear**

Fonte: Elaboração própria

O mapa conceitual apresentado na Figura [8](#_bookmark40) ilustra a relação entre as palavras- chave que está relacionada ao problema em questão, proporcionando uma visão clara do que será abordado ao longo do trabalho. Esse mapa contribui para a identificação dos principais tópicos de pesquisa e das questões que serão exploradas posteriormente.

As questões de pesquisa definidas para esta revisão sistemática da literatura são as seguintes:

Q 1 Quais são os autores que mais publicam sobre o assunto de séries temporais? Q '2 Quais são os países que mais publicam sobre o assunto?

Q 3 Quais são as áreas que mais publicam sobre o tema?

Q 4 Quais são as obras mais inftuentes na análise de séries temporais?

Essas questões guiarão a análise e a seleção dos artigos a serem revisados, permi- tindo uma compreensão mais aprofundada da produção científica relacionada ao tema das séries temporais.

* 1. Metodologia

Nesta subseção, é fornecida uma explicação detalhada de como a revisão foi con- duzida, abrangendo desde a análise do banco de dados até a conclusão final da revisão. São apresentados os passos e critérios adotados para a seleção dos artigos, bem como os procedimentos utilizados para a extração e análise dos dados. A subseção visa esclarecer de forma clara e objetiva todo o processo metodológico empregado durante a realização da revisão.

Figura 9: Etapas da Revisão

**Etapa 1: Definição das palavras chaves**

**Etapa 2: Análise nas bases de dados**

**Etapa 3: Extração dos dados**

**Etapa 4: Disseminação dos**

**resultados**

Fonte: Adaptado de [Martins e Gorschek](#_bookmark186) (['2016](#_bookmark186))

Etapa 1 A Figura [9](#_bookmark46) apresenta uma adaptação da metodologia proposta por [Martins e Gors-](#_bookmark186) [chek](#_bookmark186) (['2016](#_bookmark186)) para a realização desta revisão sistemática. Inicialmente, foram rea- lizadas buscas nos bancos de dados Scopus, Veb of Science e Lens, selecionando algumas bases relevantes para o tema da pesquisa.

Para todas as bases de busca, foram considerados os últimos 6 anos, com exceção do Lens, que retornava poucos artigos. Nessa etapa, foram utilizadas palavras-chave que se adequam melhor à pesquisa, como "t me ser es forecast n ", "t me ser es ana ys s" e "non near forecast n ".

Etapa '2 No cruzamento das palavras-chave, obteve-se um número considerável de artigos, sem restringir a área em que cada um pode ser publicado. A Tabela [1](#_bookmark51) apresenta a tabulação dos resultados obtidos, sem excluir duplicatas, que serão tratadas na seção ['2.5](#_bookmark48).

Etapa 3 Na etapa seguinte, é realizada uma avaliação preliminar de cada artigo obtido, sem aplicar nenhum filtro anual nas buscas. Analisar todos os artigos dessa maneira resultaria em um número elevado, por exemplo, no banco de dados Scopus são 498 artigos, na Veb of Science são 140 artigos e no Lens, que retorna poucos artigos, são 11 artigos, totalizando 649 artigos sem remover duplicatas. É importante ressaltar que esses artigos passaram apenas pelo filtro de idioma inglês e de serem artigos, visando aprimorar a busca e a tomada de decisões. Ao aplicar o filtro dos últimos 6 anos, obtém-se um número mais gerenciável de artigos para análise. Levando em consideração a diferença entre essa estimativa apresentada na Tabela [1](#_bookmark51) e a quantidade de artigos restantes após a remoção de duplicatas, tem-se menos de 356 artigos para análise. É válido lembrar que, ao remover as duplicatas, esse número pode diminuir ainda mais, atingindo o objetivo proposto neste trabalho.

Etapa 4 Na etapa final, é realizada uma análise mais aprofundada do conteúdo dos artigos selecionados, levando em consideração as áreas de especialização e correlação com séries temporais. Como esta revisão está inserida no contexto de um programa de mestrado em Engenharia de Produção e Sistemas, vale a pena analisar a correlação com áreas como Matemática. A Figura [16](#_bookmark59) mostra que as áreas mais relevantes para a pesquisa são "Informática", "Engenharia" e "Matemática", representando 50% das publicações. Portanto, a pesquisa está alinhada com a utilização de conceitos matemáticos básicos para realizar uma estimativa do número de artigos.

* 1. Resultados da busca de revisão

Nesta seção, serão apresentados os resultados da pesquisa, utilizando um software para melhor aproveitamento de cada banco de dados utilizado no trabalho. Inicialmente, foi realizada uma análise no software VOSviewer.

A Figura [10](#_bookmark49) mostra uma lista das palavras mais frequentemente utilizadas como sinónimos ou em conjunto com t me ser es ana ys s nos artigos. A análise da base de dados do Scopus foi feita com uma ferramenta que exibe as palavras-chave relacionadas em cada campo de busca, proporcionando uma visão abrangente das correlações com as palavras-chave principais.

Nesse primeiro momento, foram obtidas 3.484 palavras-chave, sendo que '21'2 de- las atingiram o limite estabelecido. É importante destacar que as palavras-chave base utilizadas foram t me ser es forecast n and t me ser es ana ys s no Scopus.

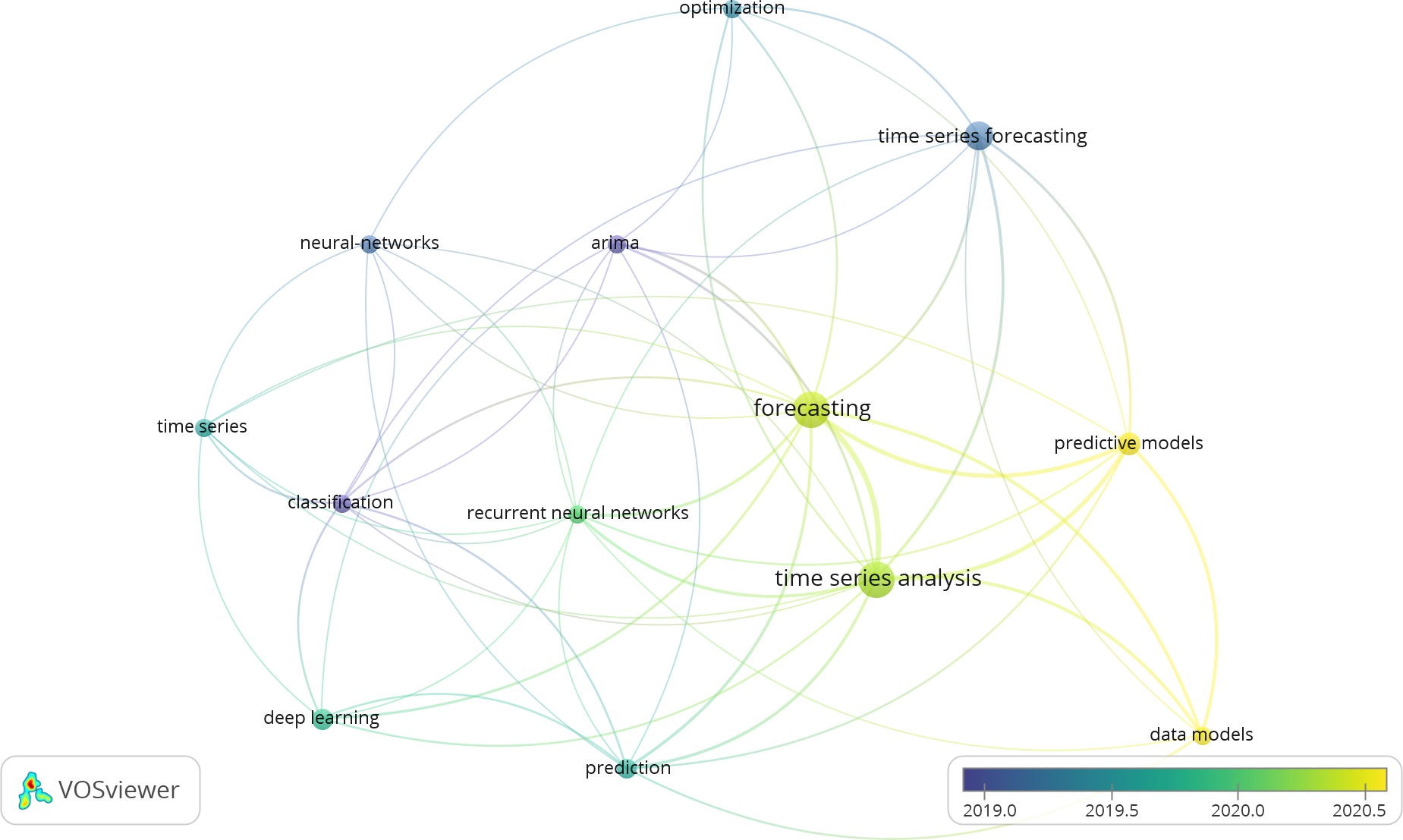
A análise do banco de dados Veb of Science, apresentada na Figura [11](#_bookmark50), também foi realizada por meio de uma ferramenta que mostra as palavras-chave relacionadas em cada campo de busca. Mais uma vez, é possível obter uma visão ampla das correlações

Figura 10: Palavras-chave mais populares na Scopus.



Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 11: Palavras-chave mais populares na VoS



Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

com as palavras-chave principais.

Nesse primeiro momento, foram obtidas 305 palavras-chave, sendo que 13 delas atingiram o limite estabelecido. É importante ressaltar que as palavras-chave base utili- zadas foram t me ser es forecast n and t me ser es ana ys s na Veb of Science.

O banco de dados Lens não será apresentado aqui, pois, embora seja uma excelente fonte, não retornou muitos resultados na pesquisa realizada. O site do Lens retornou apenas 11 artigos com os filtros aplicados. Na [Etapa 1](#_bookmark47) apresenta o campo de busca utilizado nessa pesquisa, resultando nos 11 artigos encontrados.

Tabela 1: Cruzamento de palavras-chave através da aplicação de filtros de ano e de linguagem

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Bases |  | Palavras Chaves |  |  | Resultado |
| Scopus time series | AND | time series |  |  | 490 |
| forecasting |  | analysis |  |  |  |
| nonlinear | AND | time series |  |  | 8 |
| forecasting |  | forecasting |  |  |  |
| Veb of Science time series | AND | time series |  |  | 1'26 |
| forecasting |  | analysis |  |  |  |
| nonlinear | AND | time series |  |  | 14 |
| forecasting |  | forecasting |  |  |  |
| Lens time series | AND | time series | AND | nonlinear | 11 |
| forecasting |  | analysis |  | forecasting |  |

Total 649

Fonte: Elaboração própria

A Tabela [1](#_bookmark51) apresenta as palavras-chave utilizadas em cada base de dados, jun- tamente com o número de artigos encontrados inicialmente. No entanto, é importante ressaltar que esses dados ainda não foram processados para remover duplicatas. Após a utilização do software Mende ey para eliminar as duplicações, restaram 308 artigos únicos, os quais serão considerados nesta revisão.

A Figura [1'2](#_bookmark53) apresenta um gráfico que ilustra a relação entre o número de artigos publicados e os anos correspondentes. Foi realizada uma análise utilizando regressão linear para examinar essa relação ao longo do tempo.

A equação de regressão linear obtida foi a seguinte:

*y*(*x*) = 8*,* 3571*x* − 16*.*803 com *R*2 = 0*,* 3062 (4)

Na equação ([4](#_bookmark52)), *y*(*x*) representa a equação da reta, onde *x* é a variável independente que corresponde aos anos. O coeficiente angular da reta é de 8*,* 3571, e o coeficiente linear é de −16*.*803, que indica o ponto de intersecção com o eixo y.

Figura 1'2: Analise das quantidades de artigos em relação aos anos.

*120*

*100*

*80*

*60*

*Artigos (y)*

*40*

*2016 2017 2018 2019 2020 2021 2022*

*Anos (x)*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

*Quantidade*

*Publicações*

O coeficiente de determinação, *R*2, é utilizado para avaliar a proporção da vari- ação na variável dependente (número de artigos) que pode ser explicada pela variação na variável independente (anos). Nesse caso, o valor de *R*2 foi de 0*.*3062, o que indica que aproximadamente 30,6'2% da variação nos números de artigos pode ser explicada pela passagem do tempo.

O coeficiente de determinação mede a relação entre a variável dependente e as va- riáveis independentes, representando a porcentagem da variação explicada pela regressão em relação à variação total. Quando o *R*2 é igual a 1, todos os pontos observados estão exatamente na reta de regressão, indicando um ajuste perfeito, ou seja, todas as variações em *y* são totalmente explicadas pela variação em *xn* através da função especificada, sem desvios em torno da função estimada. Por outro lado, quando o *R*2 é igual a 0, conclui- se que as variações em *y* são exclusivamente aleatórias e a inclusão das variáveis *xn* no modelo não fornece nenhuma informação sobre as variações em *y*.

A fórmula do coeficiente de determinação *R*2 é dada pela equação:

*R* =

2

*X*2 −

( *X*)2 l ·

*n*

*Y* 2 −

*X.Y* − *X· Y* 2

( *Y* )2 l = (*r*)

*n*

2 (5)

Na equação ([5](#_bookmark54)), *X* e *Y* representam as coordenadas no plano cartesiano, como por exem- plo, o par ordenado (*x, y*). Na análise realizada com a relação entre o número de artigos e os anos, obteve-se um valor de *R*2 = 30%, o que implica que a linha de regressão é inftuenciada pelo valor encontrado de *R*2.

*n*

Embora seja uma análise simples da relação entre o número de artigos e os anos, essa é uma validação significativa para observar o teste F de significância, que deve ser sempre inferior a 5%, também conhecido como valor-p. Com base nesses valores, é possível

analisar o significado da linha de regressão e observar que o ano de '20'21 foi o ano em que a maioria dos artigos foi publicada sobre o tema das séries temporais.

Tabela '2: Fator de impacto

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Revista cientíica | Quantidade  plubicação | de | Qualidade da re-  vista | H-INDEX |
| Neurocomputing | '27 |  | Q1 | 143 |
| IEEE Access | 18 |  | Q1 | 1'27 |
| Applied Soft Computing | 1'2 |  | Q1 | 143 |
| Energies | 11 |  | Q'2 | 93 |
| Energy | 11 |  | Q1 | 343 |

Fonte: Elaboração própria a partir de dados da Scopus, Lens e Veb of Science ('2016 a '20'2'2)

Na Tabela ['2](#_bookmark55), são apresentadas as revistas que mais publicam artigos sobre o tema em questão. É importante destacar que muitas dessas revistas estão localizadas fora do Brasil e têm seus nomes em inglês. No entanto, todas as revistas listadas, incluindo aquelas com um alto fator de impacto, como a categoria Q1, apresentam uma correlação significativa com as áreas de informática, engenharia e matemática.

Essa observação ressalta a importância dessas áreas de especialização na pesquisa sobre séries temporais, uma vez que elas estão fortemente representadas nas principais revistas científicas. Essas revistas desempenham um papel fundamental na disseminação do conhecimento e no avanço do campo, garantindo a qualidade e o impacto dos artigos publicados. Portanto, é valioso direcionar a atenção para essas revistas, uma vez que elas são reconhecidas como fontes confiáveis e respeitadas dentro da comunidade científica.

Em resposta à questão colocada anteriormente ([Q 1](#_bookmark41)), foi utilizada a Figura [13](#_bookmark56) para visualizar de forma mais clara os autores que mais publicaram sobre o tema em análise. O gráfico apresenta um histograma que destaca os autores cujo número de publicações é maior que 4 durante o período de '2016 a '20'2'2. Essa abordagem visa evitar a inclusão de todos os autores e destacar aqueles que tiveram uma contribuição significativa no campo, considerando o critério estabelecido de pelo menos 4 publicações. Dessa forma, é possível identificar os principais autores que se destacam nesse tópico específico, fornecendo uma visão geral da distribuição da produção científica entre os pesquisadores.

A pergunta de pesquisa [Q '2](#_bookmark42) foi abordada por meio da análise da Figura [15](#_bookmark58), que apresenta os países com maior número de publicações sobre o assunto em escala, orde- nados de forma decrescente. Os principais países que se destacam nessa análise são os seguintes: China, com 119 publicações; Estados Unidos, com 67 publicações; Índia, com

Figura 13: Relação de autores entre artigos publicados

*6*

*5*

*4*

*3*

*Artigos*

*2*

*1*

*0*

*Yaseen, Z.M.*

*(2019, 2020)*

*Papacharalampous, G. (2017, 2018, 2020, 2022)*

*Tyralis, H.*

*(2017, 2018, 2020, 2022)*

*Bergmeir, C.*

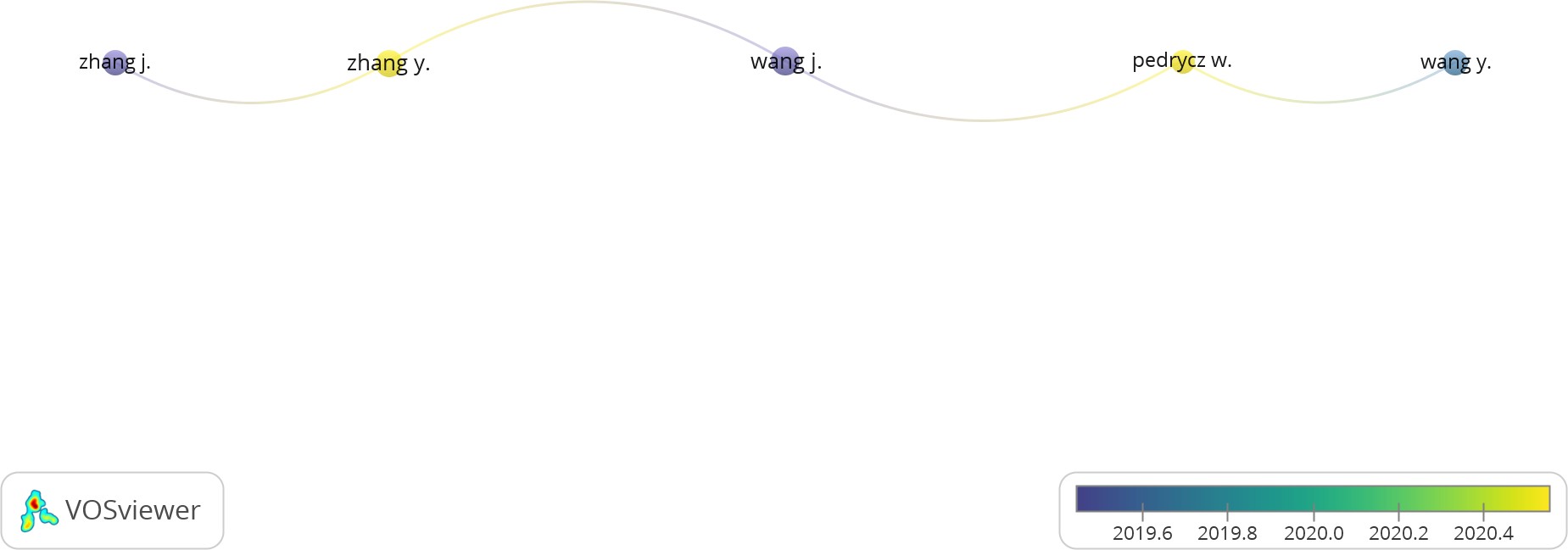
*(2018, 2021)*

*Bronakdari, H. (2019, 2020)*

*Autores e Ano de Publicação*

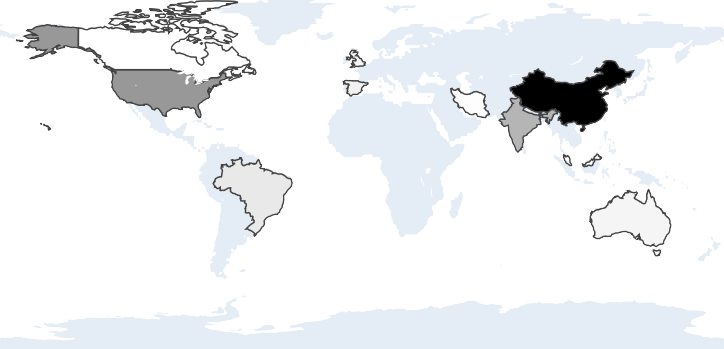
Fonte: Elaboração própria a partir de dados da Scopus ('2016 a '20'2'2)

Figura 14: Ligação bibliográfica entre os autores



Fonte: Elaboração própria a partir de dados da Scopus ('2016 a '20'2'2)

Figura 15: Mapa mundial da publicação de artigos em todo o mundo

Artigos

100

80

60

40

20

Fonte: Elaboração própria a partir de dados da Scopus, Lens e Veb of Sicence ('2016 a '20'2'2)

57 publicações; Brasil, com 3'2 publicações; Espanha, com '28 publicações; Reino Unido, com '25 publicações; Austrália, com '24 publicações; Irã, com 18 publicações; Malásia, com 17 publicações; e Canadá, com 16 publicações.

É importante ressaltar que o mapa não exibe todos os países e seus respectivos números de publicações, mas destaca aqueles com maior produção nesse contexto espe- cífico. Essa análise ajuda a identificar os países com maior contribuição científica nessa área de estudo, fornecendo insights sobre os locais onde a pesquisa sobre séries temporais tem sido mais ativa.

Para responder à pergunta de pesquisa [Q 3](#_bookmark43), foi criado um gráfico circular, apre- sentado na Figura [16](#_bookmark59), que ilustra as áreas com maior número de publicações durante o período analisado na revisão. A Tabela [3](#_bookmark60) complementa o gráfico, fornecendo os valores específicos de cada área e a quantidade de publicações correspondente.

O gráfico circular oferece uma representação visual clara das áreas que se destacam em termos de produção científica no campo das séries temporais. Ao examinar a tabela, é possível identificar as áreas com maior número de publicações, permitindo uma com- preensão aprofundada das principais áreas de conhecimento relacionadas ao tema. Essa análise contribui para uma melhor compreensão da distribuição de publicações e áreas de pesquisa ao longo do período estudado.

Figura 16: Areas de aplicação do tema

*Outros*

*Informática*

*21.2%*

*24.9%*

*Negócios, Gestão e Contabilidade*

*3.4%*

*Ciências dos Materiais 3.5%*

*4.0%*

*Ciências sociais*

*Medicina*

*4.0%*

*4.2%*

*7.0%*

*9.8%*

*18.1%*

*Engenharia*

*Neurociência*

*Matemática*

*Ciências Ambientais*

Fonte: Elaboração própria a partir de dados da Scopus, Lens e Veb of Sicence ('2016 a '20'2'2)

Tabela 3: Areas e seus valores respetivos de artigos em cada área.

|  |  |
| --- | --- |
| Informática | '240 |
| Engenharia | 174 |
| Ciências Ambientais | 94 |
| Matemática | 67 |
| Neurociência | 40 |
| Medicina | 38 |
| Ciências sociais | 38 |
| Ciências dos Materias | 34 |
| Negócios, Gestão e Contabilidade | 33 |
| Outros | '204 |

Fonte: Elaboração própria a partir de dados da Scopus, len e Veb of Sicence ('2016 a '20'2'2)

Na última pergunta de pesquisa, referente à [Q 4](#_bookmark44), foi realizada uma investigação dos artigos mais inftuentes na revisão. Esses artigos retratam alguns dos métodos utilizados por renomados autores [Golyandina](#_bookmark171) (['20'20](#_bookmark171)), [Kumar, .Jain e Singh](#_bookmark179) (['20'21](#_bookmark179)), [Xie et al.](#_bookmark217) (['2019](#_bookmark217)), [Lara-Benitez, Carranza-Garcia e Riquelme](#_bookmark180) (['20'21](#_bookmark180)), [Ahmad et al.](#_bookmark154) (['2018](#_bookmark154)), [Carvalho .Jr. e](#_bookmark162) [Costa .Jr.](#_bookmark162) (['2019](#_bookmark162)), [Tan et al.](#_bookmark208) (['20'21](#_bookmark208)), [Liu e Chen](#_bookmark182) (['2019](#_bookmark182)), [Liu et al.](#_bookmark183) (['20'21](#_bookmark183)), [Rossi](#_bookmark195) (['2018](#_bookmark195)), [Soyer e Zhang](#_bookmark206) (), [Martinovié, Hunjet e Turcin](#_bookmark185) (['20'20](#_bookmark185)), [Ursu e Pereau](#_bookmark212) (['2016](#_bookmark212)), [Vang et](#_bookmark215) [al.](#_bookmark215) (['2016](#_bookmark215)), [Shih, Sun e Lee](#_bookmark204) (['2019](#_bookmark204)), [Moon et al.](#_bookmark188) (['2019](#_bookmark188)), [Chou e Tran](#_bookmark166) (['2018](#_bookmark166)), [Bergmeir,](#_bookmark156)

[Hyndman e Koo](#_bookmark156) (['2018](#_bookmark156)), [Boroojeni et al.](#_bookmark157) (['2017](#_bookmark157)), [Chou e Nguyen](#_bookmark165) (['2018](#_bookmark165)), [Coelho et al.](#_bookmark167) (['2017](#_bookmark167)), [Du et al.](#_bookmark169) (['20'20](#_bookmark169)), [Sadaei et al.](#_bookmark196) (['2019](#_bookmark196)), [Salgotra, Gandomi e Gandomi](#_bookmark197) (['20'20](#_bookmark197)), [Tyralis e Papacharalampous](#_bookmark211) (['2017](#_bookmark211)), [Vlachas et al.](#_bookmark214) (['20'20](#_bookmark214)), [Yang et al.](#_bookmark219) (['2019](#_bookmark219)), [Shen et](#_bookmark203) [al.](#_bookmark203) (['20'20](#_bookmark203)), [Sezer, Gudelek e Ozbayoglu](#_bookmark200) (['20'20](#_bookmark200)), [Chen et al.](#_bookmark164) (['2018](#_bookmark164)), [Buyuksahin e Ertekin](#_bookmark161) (['2019](#_bookmark161)), [Li e Bastos](#_bookmark181) (['20'20](#_bookmark181)), [Kulshreshtha e Vijayalakshmi](#_bookmark178) (['20'20](#_bookmark178)), [Samanta et al.](#_bookmark198) (['20'20](#_bookmark198)), [Xu et al.](#_bookmark218) (['2019](#_bookmark218)), [Graff et al.](#_bookmark172) (['2017](#_bookmark172)), [Taieb e Atiya](#_bookmark207) (['2016](#_bookmark207)).

Esses artigos abordam diferentes métodos usados pelos autores para previsão de séries temporais e análise não-linear dessas previsões. Eles representam contribuições significativas para o avanço do conhecimento e aplicação prática das séries temporais, oferecendo insights valiosos sobre abordagens eficazes nesse campo. Ao incluir esses es- tudos inftuentes na análise, obtém-se uma visão abrangente dos métodos e técnicas mais relevantes na previsão de séries temporais.

No estudo conduzido por [Xu et al.](#_bookmark218) (['2019](#_bookmark218)), um modelo híbrido foi proposto, combi- nando o modelo linear AR e LR com o modelo não-linear ARIMA e o modelo DBN. Essa abordagem permite capturar tanto os comportamentos lineares quanto os não-lineares de uma série temporal. Por outro lado, [Li e Bastos](#_bookmark181) (['20'20](#_bookmark181)) comparou o desempenho de pre- visão da abordagem MAELS com outros modelos de aprendizado de máquina de última geração, como CNN, RNN, LSTM, ARIMA e SVM-VAR. As abordagens CNN, RNN e LSTM são capazes de lidar com dados multivariados de entrada e saída, enquanto o ARIMA utiliza informações passadas para prever o futuro com base em características como autocorrelação e médias móveis.

Dessa forma, por meio dessa revisão sistemática e análise de conteúdo, a pergunta de pesquisa formulada no início do capítulo foi respondida.

Além desses modelos mencionados, também será utilizada a versão atualizada do ARIMA nesta dissertação. Os modelos SARIMA e SARIMAX também serão comparados para determinar qual deles é o mais adequado. Além disso, serão empregados os modelos Light GBM e XGBoost. Quanto às métricas de erro, serão utilizadas MAE, MAPE e RMSE, que são amplamente adotadas na literatura. O coeficiente de determinação (*R*2), mencionado na equação ([5](#_bookmark54)), não é tão comumente utilizado para comparação de modelos de previsão futura.

* 1. Principais conclusão

A pesquisa de revisão foi minuciosamente conduzida, abrangendo uma variedade de bases de dados, como Scopus, Veb of Science e Lens. Cada uma dessas bases propor- cionou uma quantidade significativa de artigos relevantes, os quais foram cuidadosamente analisados. Essa abordagem rigorosa permitiu que a pergunta de pesquisa formulada no

início da revisão fosse respondida.

Embora a base de dados Lens seja menor em comparação com as demais, também foram encontrados artigos relevantes que contribuíram para enriquecer o processo de dis- sertação. Além disso, o uso de software especializado desempenhou um papel crucial ao lidar com a grande quantidade de artigos e suas inter-relações. No contexto específico da revisão sistemática, foi dada uma ênfase particular à análise de séries temporais, com uma abordagem aprofundada e atualizada nos últimos seis anos. Os resultados obtidos foram altamente relevantes e significativos. Por meio do cruzamento de palavras-chave e da aplicação de filtros específicos, foram selecionados 308 artigos publicados entre '2016 e '20'2'2.

Com o objetivo de aprimorar ainda mais a análise, foi realizado um filtro adicional com base em áreas de interesse, como matemática, engenharia e informática. Isso resultou na seleção de 481 artigos relacionados a essas áreas, excluindo aqueles de outras áreas não pertinentes. A pesquisa de revisão realizada foi minuciosa e abrangente, proporcionando uma base sólida de artigos relevantes para o desenvolvimento da dissertação. Os resultados obtidos foram fundamentais para orientar as próximas etapas do trabalho e alcançar uma compreensão aprofundada do tema das séries temporais.

# Base Teórica

A base teórica é fundamental para se obter resultados satisfatórios, pois ela pro- porciona um sólido conhecimento sobre o tema em questão. Neste capítulo, são abordados diversos aspectos relevantes, incluindo métricas de erro e modelos regressivos de previsão. Essas métricas desempenham um papel crucial na avaliação e comparação dos modelos, permitindo uma análise precisa do desempenho de cada um. Além disso, os modelos regressivos de previsão são explorados, fornecendo insights valiosos sobre como essas téc- nicas podem ser aplicadas para realizar previsões com precisão. Compreender e dominar esses conceitos é essencial para se obter resultados confiáveis e embasar as próximas etapas do trabalho de pesquisa.

* 1. Métricas de Avaliação de Modelos

A métrica de Erro Quadrático Médio (MSE) é amplamente utilizada no campo do aprendizado de máquina para avaliar a qualidade dos modelos de previsão. O MSE é calculado pela média da soma dos quadrados das diferenças entre os valores reais e os valores previstos. Sua fórmula é a seguinte:

*n*

1

*MSE* = (*yi*

2

*n*

*i*=1

− *y*ˆ*i*) (6)

Nessa fórmula, *n* representa o número de amostras, *yi* é o valor real correspondente à amostra *i* e *y*ˆ*i* é o valor previsto para a mesma amostra. O MSE é calculado como a média das diferenças ao quadrado entre os valores reais e os valores previstos.

A utilização do MSE fornece uma medida quantitativa da precisão do modelo, pois penaliza de forma mais significativa os erros maiores. Ao elevar as diferenças ao quadrado, a métrica enfatiza a importância de minimizar as discrepâncias entre os valores reais e os valores previstos. Dessa forma, quanto menor o valor do MSE, melhor é o desempenho do modelo em termos de previsão.

Portanto, o MSE é uma métrica fundamental para avaliar a qualidade dos modelos de previsão e é amplamente utilizada para comparar diferentes algoritmos e abordagens de aprendizado de máquina.

* + 1. Erro Quadrático Médio Raiz (RMSE)

O RMSE é uma métrica amplamente empregada na avaliação de modelos de pre- visão em séries temporais. Ele é calculado tomando a raiz quadrada do MSE, conforme

mostrado na seguinte fórmula:

*RMSE* =

1 *n*

(*yi* − *y*ˆ*i*)2 (7)

*i*=1

*n*

Na equação ([7](#_bookmark65)), *n* representa o número de amostras, *yi* é o valor real correspondente à amostra *i*, e *y*ˆ*i* é o valor previsto para a mesma amostra. O RMSE fornece uma medida da dispersão média entre os valores reais e os valores previstos pelo modelo.

Uma das vantagens de utilizar o RMSE é que, ao computar a raiz quadrada, o erro passa a ter a mesma escala da variável de interesse. Isso permite uma interpretação mais fácil dos resultados, sendo que um valor baixo de RMSE indica um bom desempenho do modelo, já que o erro se aproxima de zero.

O RMSE possui algumas características positivas. Ele penaliza de forma signifi- cativa os valores discrepantes, caso seja necessário para o modelo. Além disso, o erro resultante está nas mesmas unidades da série temporal, facilitando a interpretação. O RMSE pode

ser considerado uma combinação das melhores características do MSE e do Erro Absoluto Médio (MAE).

No entanto, o RMSE também apresenta algumas desvantagens. Ele tem uma interpretabilidade reduzida, uma vez que os erros ainda são elevados ao quadrado. Além disso, o RMSE é dependente da escala dos dados, o que impede sua comparação direta com modelos de séries temporais que utilizam unidades diferentes.

Apesar das limitações, o RMSE é uma métrica amplamente utilizada para avaliar modelos de previsão em séries temporais. Ele fornece uma medida da dispersão média entre os valores reais e previstos, auxiliando na compreensão do desempenho do modelo e na comparação com outras abordagens.

* + 1. Erro Abi,oluto Médio (MAE)

O Erro Absoluto Médio (MAE) é amplamente utilizado como uma métrica para avaliar o desempenho de modelos de previsão. Em vez de calcular a média das diferenças entre os valores reais e previstos, o MAE calcula a média dos valores absolutos dessas diferenças, garantindo que os erros positivos e negativos não se anulem.

O MAE mede o desvio médio das previsões em relação aos valores reais e é uma métrica intuitiva e fácil de interpretar, representando a magnitude média dos erros em relação à escala dos dados. Por exemplo, um MAE de '2 significa que, em média, as previsões têm um desvio absoluto de '2 unidades em relação aos valores reais.

Uma das vantagens do MAE é a sua insensibilidade a valores extremos, pois trata os erros de forma absoluta. No entanto, como o MAE não considera a magnitude dos erros individuais, pode não reftetir adequadamente a gravidade de desvios significativos em relação aos valores reais.

Para superar essa limitação, uma alternativa é o Erro Médio Absoluto Percentual (MAPE). O MAPE expressa o MAE como uma porcentagem em relação aos valores reais, proporcionando uma medida relativa de erro. Essa métrica é especialmente útil quando se deseja avaliar o desempenho de um modelo em relação à magnitude dos dados.

Em resumo, o MAE é uma métrica simples e fácil de interpretar, que mede o desvio médio das previsões em relação aos valores reais. O MAPE, por sua vez, fornece uma medida relativa de erro, expressa como uma porcentagem dos valores reais. A escolha entre essas métricas depende do contexto do problema e dos requisitos específicos de avaliação.

O cálculo do MAE é realizado utilizando o valor absoluto da diferença entre o valor real e o valor previsto, e em seguida, divide-se pela quantidade *n* de amostras. Isso resulta no erro médio absoluto. A equação do MAE é dada por:

*MAE* = 1 |*y*

*n*

*i*

− *y*ˆ*i*| (8)

Sua interpretação é similar ao RMSE, em que o erro é expresso na mesma escala ou ordem

de grandeza da variável estudada.

* + 1. Erro Percentual Abi,oluto Médio (MAPE)

O Erro Percentual Absoluto Médio (MAPE) é uma métrica que expressa o erro de previsão como uma porcentagem relativa ao valor observado. Ele é calculado somando as diferenças entre o valor real e o valor previsto (representando o erro), dividido pelo valor observado.

O MAPE é calculado usando a seguinte fórmula:

*MAPE* = 1 *yi* − *y*ˆ*i*

(9)

*n yi*

No entanto, surge um problema quando o valor observado *yi* é igual a zero, pois é matematicamente impossível dividir por zero. O MAPE é uma medida de erro em que valores menores indicam um melhor desempenho de previsão.

Uma alternativa ao MAPE é calcular 1 − MAPE, que representa a porcentagem de acerto.

O Erro Percentual Absoluto Médio é comumente usado como uma métrica de referência para avaliar o desempenho de modelos de previsão.

Prós:

* + - * Fácil de interpretar
      * Independente de escala, permitindo comparações entre diferentes séries temporais Contras:
      * Erro infinito se o valor real estiver próximo ou igual a zero
      * Previsões mais baixas estão propensas a ter um erro de 100%, enquanto previsões mais altas podem ter um erro infinito, o que resulta em um viés de subprevisão.

Essas métricas são amplamente utilizadas na avaliação de modelos de previsão em diferentes áreas e ajudam a quantificar a qualidade das previsões realizadas pelos modelos.

* 1. Modelos de Séries Temporais Univariados

A previsão de séries temporais é um desafio complexo, sem uma resposta fácil. Existem inúmeros modelos estatísticos que afirmam superar uns aos outros, mas nunca está claro qual modelo é o melhor.

Dito isto, os modelos baseados em ARMA são frequentemente uma boa opção para iniciar. Eles podem alcançar pontuações decentes na maioria dos problemas de séries temporais e são adequados como modelos de referência em tais problemas.

Quanto ao modelo ARIMA, ele é dividido em três componentes: AR (Auto- Regressão), I (Integração) e MA (Média Móvel). O componente AR leva em conside- ração os valores anteriores da série temporal, o componente I trata das diferenças entre os valores observados para tornar a série estacionária, e o componente MA considera os erros residuais do modelo. Esses componentes combinados ajudam a capturar os padrões e tendências presentes na série temporal.

* + 1. Componente Autorregrei,i,ivo

O componente autoregressivo do modelo ARIMA é representado por AR(p), em que o parâmetro p determina o número de séries temporais defasadas utilizadas.

A equação do modelo AR(p) é expressa da seguinte forma:

*p*

*Yt* = *c* + *αnYt−n* + *εt* (10)

*n*=1

A partir dos dados, é possível obter uma previsão utilizando o modelo AR(7).

Figura 17: Modelo AR(7)

*4.0*

*Observado*

*AR (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*Auto-Regressivo*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 18: ARX (7)

*4.0*

*Observado*

*ARX (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*Auto-Regressivo Exogedo*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Na equação ([10](#_bookmark70)), o termo *εt* representa o ruído branco. Essa equação pode ser entendida como uma regressão múltipla, em que os valores defasados de *yt* são utilizados como preditores. Esse modelo é conhecido como modelo autorregressivo de ordem *p*, ou AR(p).

A Figura [17](#_bookmark71) tem como objetivo apresentar uma previsão de um passo à frente (um

dia). Nos apêndices [C](#_bookmark244), pode-se observar uma comparação entre os modelos AR, MA e ARX.

O modelo ARX é uma extensão do modelo AR, que incorpora variáveis exógenas nos dados para melhorar as previsões futuras. Esse modelo também é multivariado, como mostrado na subseção [3.3](#_bookmark86), e foi incluído aqui para fins de comparação com o modelo AR simples, considerando a presença de variáveis exógenas.

Embora o modelo AR possa ser visualmente adequado para a previsão que está sendo feita, é importante destacar que, por ser um modelo autorregressivo, ele realiza previsões lineares e não captura padrões não lineares presentes nos dados. Para uma análise mais abrangente da série temporal, é necessário considerar exemplos de casos gerais.

* + 1. AR(0): Ruído branco

Se o parâmetro *p* for definido como zero (AR(0)), significa que não há termos autorregressivos no modelo. Nesse caso, a série temporal se comporta como um ruído branco. Cada ponto de dados é amostrado de uma distribuição com média zero e variância igual a sigma-quadrado. Isso resulta em uma sequência de números aleatórios que não exibem nenhum padrão ou correlação.

Essa propriedade do ruído branco pode ser útil em análises estatísticas, pois serve como uma hipótese nula. Ao comparar diferentes modelos ou testar a presença de padrões em uma série temporal, podemos usar o ruído branco como referência para avaliar se os resultados observados são estatisticamente significativos ou apenas resultado do acaso. Isso nos ajuda a evitar a detecção de padrões falsos positivos e garante a confiabilidade das análises realizadas.

* + 1. AR(1): Caminhadai, aleatóriai, e Oi,cilaçõei,

Com o parâmetro *p* definido como 1, o modelo AR leva em consideração o valor anterior da série temporal multiplicado por um coeficiente e, em seguida, adiciona ruído branco. Quando o coeficiente é igual a 0, temos apenas ruído branco, resultando em uma série de tempo completamente aleatória, sem padrões previsíveis.

Quando o coeficiente é igual a 1, temos uma caminhada aleatória, onde cada valor da série é obtido somando-se o valor anterior a um termo de ruído branco. Nesse caso, os valores da série apresentam uma tendência linear, aumentando ou diminuindo ao longo do tempo sem retornar à média.

Se o coeficiente estiver na faixa 0 *< α <* 1, temos o fenómeno de reversão média. Isso significa que os valores da série tendem a oscilar em torno de uma média central e

a regressar em direção a ela após se afastarem. Esse padrão indica uma tendência de retorno à média ao longo do tempo.

Os diferentes comportamentos da série temporal, determinados pelo coeficiente no modelo AR, têm implicações importantes na análise e previsão de dados. A compreensão desses padrões é fundamental para escolher o modelo adequado e interpretar corretamente os resultados obtidos.

* + 1. AR(p): Termoi, de ordem i,uperior

Aumentar ainda mais o parâmetro *p* no modelo AR significa considerar um nú- mero crescente de medições de tempo anteriores, cada uma multiplicada pelo seu próprio coeficiente. Isso permite levar em conta uma memória mais longa da série temporal e capturar padrões de dependência mais complexos ao longo do tempo.

No entanto, é importante ter em mente que aumentar excessivamente o valor de *p* pode levar a problemas de overfitt n , onde o modelo se ajusta muito bem aos dados de treinamento, mas tem um desempenho ruim na previsão de novos dados. Portanto, é necessário encontrar um equilíbrio entre a complexidade do modelo e sua capacidade de generalização.

Além disso, é comum combinar o modelo AR com o modelo de média móvel (MA) para formar o modelo ARMA. O modelo MA considera os erros passados, ou seja, as diferenças entre os valores reais e as previsões anteriores, ajustadas por coeficientes. A combinação dos componentes AR e MA permite capturar tanto a dependência autorre- gressiva quanto a dependência na média móvel, proporcionando uma modelagem mais abrangente da série temporal.

Em suma, aumentar o parâmetro *p* no modelo AR pode melhorar a capacidade do modelo de capturar padrões complexos da série temporal, mas é necessário ter cuidado para evitar overfitt n . A combinação com o modelo MA pode fornecer uma modelagem mais completa dos dados. A escolha adequada dos parâmetros depende da análise cui- dadosa dos padrões presentes na série temporal e do equilíbrio entre a complexidade do modelo e sua capacidade de generalização.

* + 1. Média Móvel

No modelo de média móvel (MA), o componente não é uma média móvel sim- ples, mas sim uma combinação de termos de erro de previsão defasados. O parâmetro *q* no modelo MA representa o número de termos de erro de previsão que são levados em consideração na previsão.

De acordo com [Trenberth](#_bookmark210) ([1984](#_bookmark210)) este componente não é uma média de rolamento,

mas sim os atrasos no ruído branco.

Em um modelo MA(1), por exemplo, a previsão é composta por um termo cons- tante, o produto do termo de erro de previsão anterior por um multiplicador, e o termo de erro de previsão atual. Essa abordagem baseia-se em princípios estatísticos e de pro- babilidade, ajustando a previsão com base em termos anteriores de erro de previsão.

O modelo MA é uma alternativa ao modelo AR e é usado para capturar padrões de dependência na média móvel, ou seja, a inftuência de erros passados na previsão atual. Ao combinar o modelo AR e o modelo MA, como no modelo ARMA, é possível obter uma modelagem mais abrangente que considera tanto a dependência autorregressiva quanto a dependência na média móvel.

Portanto, o modelo MA leva em conta os termos de erro de previsão defasados para ajustar a previsão atual, permitindo considerar a probabilidade e estatística na modelagem da série temporal.

*yt* = *c* + *εt* + *θ*1*εt−*1 + *θ*2*εt−*2 + · · · + *θqεt−q* (11)

Na equação ([11](#_bookmark77)), em que *εt* representa o ruído branco, esse modelo é conhecido como um modelo de média móvel *MA*(*q*), em que *q* é a ordem da média móvel. É importante ressaltar que não observamos diretamente os valores de *εt*, portanto, essa modelagem não se trata de uma regressão no sentido convencional.

Diferentemente de uma regressão comum em que temos variáveis explicativas ob- servadas, no modelo *MA*(*q*), estamos usando os termos de ruído branco defasados para estimar e prever os valores da série temporal. O objetivo é capturar a dependência dos termos de erro passados na previsão atual.

Esse modelo é útil para modelar séries temporais em que a média móvel tem um impacto significativo nas observações. Ao ajustar a série temporal com base nos termos de ruído branco defasados, podemos obter uma estimativa mais precisa dos valores futuros.

Embora o modelo *MA*(*q*) seja diferente de uma regressão tradicional, ele é uma ferramenta estatística poderosa para modelar e prever séries temporais, levando em con- sideração a dependência entre os termos de erro passados.

Figura 19: Modelo MA(7)

*4.0*

*Observedo*

*MA (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*Médias Móveis*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

O modelo MA, quando comparado com o modelo AR de mesma ordem, facilita a previsão. Conforme ilustrado na Figura [19](#_bookmark78), a previsão gráfica se assemelha ao modelo apresentado na Figura [17](#_bookmark71), embora não seja comparável ao modelo exibido na Figura [18](#_bookmark72). É importante notar que esse modelo aparenta prever com precisão o período de tempo que foi considerado.

3.2.6 Modeloi, ARMA e ARIMA

A arquitetura ARMA é uma combinação dos modelos AR e MA, onde o modelo AR é adicionado ao modelo MA.

No modelo ARMA, é adicionada uma constante à soma dos termos autorregressivos multiplicados pelos seus coeficientes, juntamente com a soma dos termos de média móvel multiplicados pelos seus coeficientes, além do ruído branco. Essa estrutura é amplamente utilizada em diversos modelos de previsão em diferentes áreas.

Figura '20: ARMA (7,7)

*4.0*

*Observedo*

*ARMA (7, 7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*ARMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

A Figura ['20](#_bookmark80) ilustra a combinação dos modelos AR e MA em um modelo ARMA. Essa abordagem pode levar a uma redução significativa no erro de previsão, como obser- vado nos apêndices [A](#_bookmark222) e [B](#_bookmark239), onde são apresentadas comparações com um maior número de passos de previsão.

* + 1. ARIMA

*Yt* = *β*2 + *ω*1*εt−*1 + *ω*2*εt−*2 + *. . .* + *ωqεt−q* + *εt* (1'2)

Na equação ([1'2](#_bookmark82)), a variável *Yt* representa a série temporal que foi diferenciada (possivelmente mais de uma vez). Os 11preditores11no lado direito da equação incluem os valores defasados de *Yt* e os erros defasados. Esse tipo de modelo é conhecido como ARIMA (*p, d, q*).

O modelo ARIMA é uma extensão do modelo ARMA que incorpora uma etapa adicional de pré-processamento chamada de diferenciação. Essa etapa é representada pela notação I(d), em que d denota a ordem de diferenciação, ou seja, o número de trans- formações necessárias para tornar a série temporal estacionária. Portanto, um modelo ARIMA é simplesmente um modelo ARMA aplicado à série temporal diferenciada. Isso permite lidar com séries temporais que possuem tendências ou padrões não estacionários.

Figura '21: ARIMA (7,1,7)

*4*

*Observedo*

*ARIMA (7, 1, 7)*

*3*

*2*

*LT01 (m³)*

*1*

*0*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*ARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Ao analisar a Figura ['21](#_bookmark83), não se nota uma diferença visual significativa em relação aos outros métodos apresentados anteriormente. O método ARX ainda parece ser superior aos demais com base na análise visual.

Embora os modelos ARIMA sejam eficazes, incorporar variáveis sazonais e exóge- nas ao modelo pode potencializar sua capacidade de previsão. No entanto, é importante destacar que o modelo ARIMA pressupõe que a série temporal seja estacionária. Quando lidamos com séries temporais não estacionárias, é necessário recorrer a outros modelos para a análise e previsão adequadas.

* + 1. SARIMA

*p q P Q*

*Yt* = *c* + *αnyt−n* + *θnϵt−n* + *ϕnyt−sn* + *ηnϵt−sn* + *ϵt* (13)

*n*=1

*n*=1

*n*=1

*n*=1

O modelo proposto é uma extensão do modelo ARIMA, com a adição de compo- nentes autorregressivos e de média móvel sazonal. Esses componentes extras são ajustados levando em consideração os padrões sazonais presentes nos dados, utilizando atrasos cor- respondentes à frequência sazonal (por exemplo, 1'2 para dados mensais). Essa abordagem permite capturar e modelar de forma mais precisa as variações sazonais e melhorar a qua- lidade das previsões em séries temporais com esse comportamento cíclico.

Figura '2'2: SARIMA (7*,* 1*,* 7)(2*,* 1*,* 1)12

*4*

*Observedo*

*SARIMA (7, 1, 7) (2, 1, 0, 12)*

*3*

*LT01 (m³)*

*2*

*1*

*0*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*SARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Na Figura ['2'2](#_bookmark85), é possível observar que a previsão em vermelho está mais próxima dos valores observados em preto, mostrando que a inclusão do componente de sazonalidade melhora a qualidade da previsão. Os modelos SARIMA são capazes de lidar com dados que apresentam padrões sazonais, permitindo a diferenciação dos dados em termos de componentes sazonais e não sazonais. Uma abordagem útil para determinar os melhores parâmetros do modelo é utilizar uma estrutura de pesquisa automatizada de parâmetros, como o pmdarima, que auxilia na identificação dos parâmetros ideais para o modelo SARIMA. Isso pode contribuir para uma melhor compreensão e ajuste do modelo aos dados observados.

* 1. Modelos de Série Temporal Multivariada
     1. ARIMAX e SARIMAX

*p q r P Q*

*dt* = *c* + *αndt−n* + *θnϵt−n* + *βnxnt* + *ϕndt−sn* + *ηnϵt−sn* + *ϵt* (14)

*n*=1

*n*=1

*n*=1

*n*=1

*n*=1

Em ([14](#_bookmark88)), o modelo SARIMAX é apresentado. Nesse modelo, são consideradas va- riáveis exógenas, ou seja, são utilizados dados externos para a realização das previsões. É importante ressaltar que mesmo que essas variáveis exógenas sejam indiretamente mode- ladas no histórico de previsões do modelo, ao incluí-las diretamente, o modelo será capaz de responder de forma mais ágil aos efeitos dessas variáveis. Isso significa que a incorpo- ração de informações externas possibilita uma resposta mais rápida e precisa do modelo em relação aos fatores externos, resultando em previsões mais atualizadas e acuradas.

Figura '23: ARIMAX (7*,* 1*,* 7)

*4.0*

*Observado*

*ARIMAX (7, 1, 7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*ARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura '24: SARIMAX (7*,* 1*,* 7)(2*,* 1*,* 1)12

*4.0*

*Observado*

*SARIMAX (7, 1, 7) (2, 1, 0, 12)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*SARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Entre os modelos com variáveis exógenas, como mostrado nas Figuras ['23](#_bookmark89) e ['24](#_bookmark90), observa-se uma melhora significativa na qualidade das previsões em comparação com os modelos que não incluem variáveis exógenas. A adição dessas variáveis externas permite capturar melhor as inftuências e os padrões presentes nos dados, resultando em previsões mais completas e precisas. Essa inclusão de informações adicionais contribui para uma compreensão mais abrangente do comportamento da série temporal e possibilita uma melhor adaptação do modelo aos padrões observados.

* 1. Modelos de Aprendizado de Máquina Supervisionados

Os modelos regressivos para séries temporais têm sido amplamente reconhecidos e utilizados na literatura atual, especialmente aqueles baseados em métodos de gradiente.

Esses modelos, incluindo a regressão linear simples, têm se destacado como uma escolha popular em competições de séries temporais em todo o mundo.

Esses modelos são valorizados por sua capacidade de capturar relações complexas e não lineares nos dados, permitindo previsões mais precisas e eficientes. Sua popularidade reftete o reconhecimento da eficácia desses modelos em abordar uma ampla gama de problemas de previsão de séries temporais em diferentes áreas de estudo.

A abordagem regressiva, combinada com técnicas de otimização baseadas em gradi- ente, tem se mostrado particularmente eficaz na obtenção de resultados de alta qualidade. Esses modelos são capazes de aprender a partir dos dados históricos e ajustar seus parâ- metros de forma iterativa, otimizando assim o desempenho da previsão.

Com a crescente disponibilidade de dados e avanços na área de aprendizado de máquina, espera-se que os modelos regressivos para séries temporais continuem a evoluir e desempenhar um papel importante na análise e previsão de dados temporais em diversas aplicações.

* + 1. Regrei,i,ão Linear (LR)

De acordo com o estudo realizado por [Korstanje](#_bookmark177) (['20'21](#_bookmark177)), nos modelos de aprendi- zado de máquina supervisionados, é feita uma tentativa de identificar as relações existentes entre diferentes variáveis:

* + - * Variável de destino: a variável que você tenta prever
      * Variáveis explicativas: Variáveis que ajudam você a prever o alvo variável

Para realizar previsões, é importante que se compreenda quais tipos de variá- veis explicativas podem ser utilizadas. Neste exemplo, a variável Prei,i,ão de Sucção (PT01SU) será considerada como a variável *x*, enquanto a variável Nível do Rei,er- vatório (Câmara 1) LT01 será considerada como a variável *y*, com base na análise de correlação de Pearson ilustrada na Figura ['25](#_bookmark93). O coeficiente de correlação indica a relação entre o eixo *x* e *y*, como expresso pela seguinte fórmula.

A fórmula do coeficiente de correlação de Pearson é dada por:

*r* =

(*xi* − *x*¯) (*yi* − *y*¯)

(*xi* − *x*¯)2 (*yi* − *y*¯)2

(15)

Onde *xi* e *yi* representam os valores das variáveis *X* e *Y* , respectivamente. *x*¯ e *y*¯

são as médias dos valores *xi* e *yi*. O coeficiente de correlação de Pearson mede a força e a direção da relação linear entre as variáveis *X* e *Y* . Valores próximos a 1 indicam uma

*B3*

correlação positiva forte, valores próximos a -1 indicam uma correlação negativa forte, e valores próximos a 0 indicam uma ausência de correlação entre as variáveis.

Figura '25: Corelação de Pearson

*B2*

*B1*

*B1 B2 B3 LT01 FT01 FT02 FT03 PT01 PT02*

*1 0.18 -0.06 -0.23 0.63 0.5 0.82 -0.3 0.89*

*0.18 1 -0.49 -0.22 0.35 0.3 0.46 -0.26 0.42*

*-0.06 -0.49 1 -0.16 0.14 0.17 0.15 -0.18 0.074*

*-0.23 -0.22 -0.16 1 -0.4 -0.45 -0.35 0.98 -0.29*

*0.63 0.35 0.14 -0.4 1 0.77 0.77 -0.46 0.73*

*0.5 0.3 0.17 -0.45 0.77 1 0.63 -0.49 0.59*

*0.82 0.46 0.15 -0.35 0.77 0.63 1 -0.45 0.93*

*-0.3 -0.26 -0.18 0.98 -0.46 -0.49 -0.45 1 -0.37*

*0.89 0.42 0.074 -0.29 0.73 0.59 0.93 -0.37 1*



*1.0*

*0.8*

*0.6*

*0.4*

*0.2*

*0.0*

*0.2*

*0.4*

*PT02 PT01 FT03 FT02 FT01 LT01*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

A Figura ['25](#_bookmark93) ilustra a correlação entre as variáveis no conjunto de dados em questão. Essa imagem representa graficamente a relação entre as variáveis e é usada para demons- trar a existência de uma correlação forte entre elas. Com base nessa análise, é possível responder à pergunta de pesquisa [Q 1](#_bookmark6), pois a correlação entre as variáveis é significativa.

* + 1. Defnição do modelo

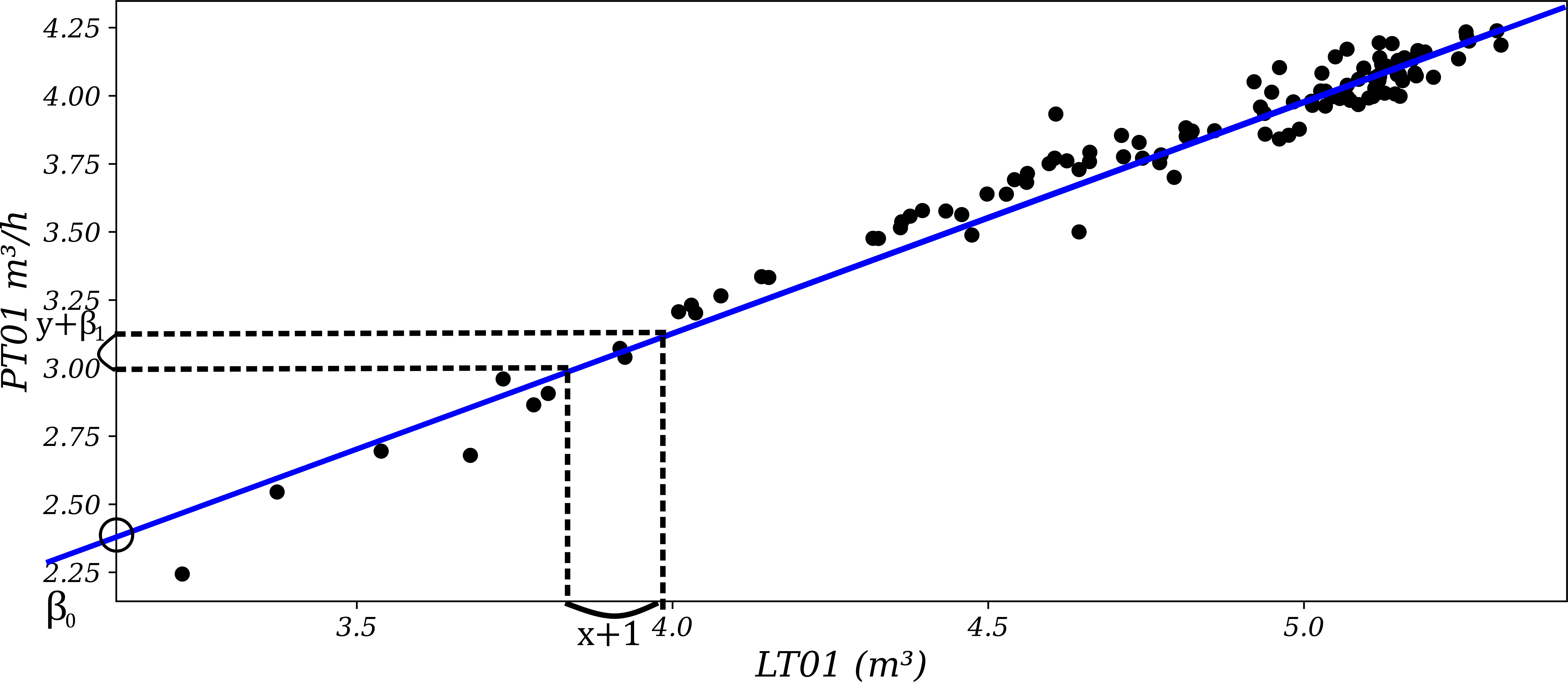
A regressão linear é definida da seguinte forma:

*y* = *β*0 + *β*1*x*1 + · · · + *βpxp* + *ε* (16) Da Equação ([16](#_bookmark95)), temos as seguintes variáveis:

* + - * Há *p* variáveis explicativas, denotadas por *x*.
      * Existe uma variável alvo, denotada por *y*.
      * O valor de *y* é calculado como uma constante *β*0, somada aos valores das variáveis

*x* multiplicados por seus coeficientes *β*1 a *βp*.

Figura '26: Regressão linear LT01 vs PT01 correlação 98%



Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

A Figura ['26](#_bookmark96) fornece uma representação visual da interpretação dos coeficientes *β*0 e *β*1. Ela ilustra que um aumento de 1 na variável *x* está associado a um aumento proporcional de *β*1 na variável *y*. O valor de *β*0 representa o valor de *y* quando *x* é igual a 0.

Para utilizar a regressão linear, é necessário estimar os coeficientes (betas) com base em um conjunto de dados de treinamento. Esses coeficientes podem ser estimados por meio da seguinte fórmula, expressa em notação matricial:

*β*ˆ = *XT X* *−*1 *XT y* (17)

A fórmula mencionada, conhecida como OLS (método dos mínimos quadrados ordinários), é amplamente utilizada na regressão linear [Korstanje](#_bookmark177) (['20'21](#_bookmark177)). Esse método é conhecido por ser rápido de ajustar, pois requer apenas cálculos matriciais para estimar os coeficientes *β*. No entanto, ele é mais adequado para processos lineares e pode ser menos adequado para modelos mais complexos que envolvam relações não-lineares. Portanto, é importante considerar suas limitações ao aplicar a regressão linear em contextos mais complexos.

Figura '27: Regressão linear (LR) um passo a frente

*4.5*

*Observedo*

*Previsto*

*4.0*

*3.0*

*2.5*

*2.0*

*1.5*

*0 50 100 150 200 250 300*

*Regressão Linear*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

3.4.3 Florei,ta Aleatória (Random Forei,t)

Pode-se observar que ter exatamente a mesma árvore de decisão repetidas vezes não adiciona valor significativo em comparação a usar essa mesma árvore de decisão apenas uma vez. Em modelos de conjunto, cada modelo individual deve ser ligeiramente diferente dos demais. Existem dois métodos amplamente reconhecidos para criar conjuntos: o en- sacamento (ba n ) e o reforço (boost n ). A ftoresta aleatória utiliza o ensacamento para criar um conjunto de árvores de decisão, onde cada árvore é construída com uma amostra aleatória do conjunto de dados original. Isso garante que as árvores sejam distintas e diversificadas, contribuindo para a robustez e eficácia do modelo.

Figura '28: Regressão da Floresta Aleatória (RFR)

*4.0*

*Observedo*

*Previsto*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*0 50 100 150 200 250 300*

*Regressor de Floresta Aleatória*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Segundo [Pelletier et al.](#_bookmark189) (['2016](#_bookmark189)), cada árvore em um modelo de Floresta Aleatória de Regressão (RFR) é construída por meio de um algoritmo de aprendizado individual que

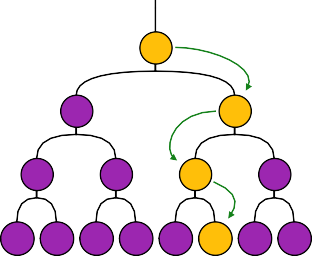
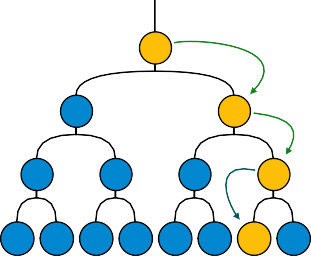
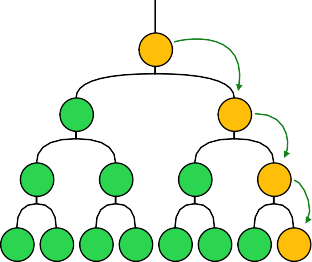
divide o conjunto de variáveis de entrada em subconjuntos, com base em um teste de valor de atributo, como o coeficiente de Gini. Ao contrário das árvores de decisão clássicas, as árvores de RFR são construídas sem poda e selecionam aleatoriamente um subconjunto de variáveis de entrada em cada nó. Atualmente, o número de variáveis utilizadas para dividir um nó em uma RFR (denotado por *m*) corresponde à raiz quadrada do número total de variáveis de entrada. Essa abordagem ajuda a aumentar a diversidade das árvores e aprimorar o desempenho do modelo.

Figura '29: Esquema da Floresta Aleatória

**Instância**



Árvore 1 Árvore 2 **...** Árvore - n

...

Fonte: Elaboração própria

* + 1. Gradient Booi,ting (como XGBooi,t, LightGBM)

O aumento de gradiente (do inglês rad ent boost n ) é um método que combina vários modelos de árvore de decisão para realizar previsões. Cada uma dessas árvores de decisão é única, pois a diversidade é um elemento importante nesse processo. A diversidade é alcançada através de um processo chamado boosting, que é uma abordagem iterativa. O boosting adiciona modelos fracos ao conjunto de forma inteligente, dando mais peso aos pontos de dados que ainda não foram bem previstos.

O processo de boosting melhora o conjunto ao focar nas partes dos dados que ainda não são compreendidas. A Figura [31](#_bookmark107) apresenta uma visão esquemática desse processo. A medida que novos modelos fracos são adicionados, todos os modelos fracos intermediários são mantidos. O modelo final é uma combinação de todos esses modelos fracos, resultando em um ensemble que oferece uma melhor capacidade de previsão do que um único modelo. O boosting é apenas um dos métodos de ensemble utilizados em conjunto com o bagging. O bagging também é um método que utiliza múltiplos modelos de árvore de decisão, porém, em vez de adicionar os modelos de forma iterativa, cada modelo é trei- nado independentemente em subconjuntos aleatórios dos dados de treinamento. Ambos os métodos, boosting e bagging, têm como objetivo melhorar o desempenho do modelo

combinando as previsões de múltiplos modelos individuais.

Figura 30: Impulsionando gradiente com XGBoost e LightGBM

**Modelo Fraco**

|  |  |  |
| --- | --- | --- |
| **Dados Original** | | |
|  |  | \* |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| **Dados Ponderado** | | |
|  |  | \* |
|  |  |  |
|  |  |  |
|  |  |  |

Primeiro modelo em forma

Erro de Previsão Por ponto de dados

Colocar o próximo modelo em dados ponderados

Fonte: Adaptação de [Korstanje](#_bookmark177) (['20'21](#_bookmark177))

* + 1. O Gradiente em Gradiente de Booi,ting (Reforço)

O processo iterativo utilizado no aumento de gradiente, como descrito por [Kors-](#_bookmark177) [tanje](#_bookmark177) (['20'21](#_bookmark177)), recebe esse nome por um motivo. O termo "gradiente" refere-se a um campo vetorial de derivadas parciais que apontam na direção da inclinação mais acentuada. De forma simplificada, podemos pensar nos gradientes como as inclinações das estra- das: quanto maior a inclinação, mais íngreme a colina. Para calcular os gradientes, são realizadas derivadas ou derivadas parciais de uma função.

No aumento de gradiente, ao adicionar árvores adicionais ao modelo, o objetivo é incorporar uma árvore que explique melhor a variação que ainda não foi explicada pelas árvores anteriores. Dessa forma, a nova árvore tem como objetivo ajustar-se aos erros ou resíduos deixados pelas árvores anteriores.

*y* − *y*ˆ

(18)

A equação ([18](#_bookmark104)) pode ser reescrita como a derivada parcial negativa da função de perda em relação às previsões *y*ˆ:

*∂L y* − *y*ˆ = − *∂y*ˆ

(19)

Isso é definido como o objetivo da nova árvore a ser adicionada no modelo de aumento de gradiente, garantindo que ela explique a máxima quantidade de variação adicional no modelo geral. Essa é a razão pela qual o modelo é chamado de 11aumento de gradiente11(" rad ent boost n ", em inglês). O processo utiliza o gradiente da função de perda para guiar a adição de novas árvores, buscando minimizar o erro e melhorar a capacidade do modelo em explicar a variação nos dados.

* + 1. Algoritmoi, de booi,ting de gradiente

Existem muitos algoritmos que executam versões ligeiramente diferentes de au- mento de gradiente. Quando o método de aumento de gradiente foi inventado, o algo- ritmo não tinha um desempenho tão bom, mas isso mudou com o advento do algoritmo AdaBoost: o primeiro algoritmo capaz de se adaptar a modelos fracos.

O algoritmo de aumento de gradiente é uma das ferramentas de aprendizado de máquina com melhor desempenho no mercado. Após o AdaBoost, uma longa lista de algoritmos de aumento levemente diferentes foi adicionada à literatura, incluindo XGBo- ost, LightGBM, LPBoost, BrownBoost, MadaBoost, LogitBoost e TotalBoost. Ainda há muitas contribuições para melhorar a teoria do aumento de gradiente. Nesta subseção, dois algoritmos são apresentados: XGBoost e LightGBM.

O XGBooi,t é um dos algoritmos de aprendizado de máquina mais utilizados. É uma forma rápida de obter bom desempenho. Devido à sua facilidade de uso e alto desempenho, é frequentemente o primeiro algoritmo escolhido por muitos profissionais de aprendizado de máquina.

O LightGBM é outro algoritmo de aumento de gradiente que é importante co- nhecer. Atualmente, é um pouco menos difundido que o XGBoost, mas está ganhando popularidade rapidamente. A vantagem esperada do LightGBM em relação ao XGBoost é um ganho de velocidade e uma utilização mais eficiente de memória.

Nesta subseção, você encontrará as implementações de ambos os algoritmos de aumento de gradiente.

* + 1. A diferença entre XGBooi,t e LightGBM

Se alguém planeja utilizar os dois algoritmos de aumento de gradiente, é importante que essa pessoa compreenda suas diferenças, o que também proporciona uma visão das várias divergências que existem entre os modelos disponíveis no mercado.

Uma diferença fundamental reside na maneira como esses algoritmos identificam as melhores divisões entre os nós das árvores de decisão individuais. É crucial lembrar que uma divisão em uma árvore de decisão ocorre quando a árvore precisa encontrar a separação que mais melhora o desempenho do modelo.

A abordagem intuitiva e simples para encontrar a melhor divisão é iterar por todas as possibilidades e selecionar a melhor. No entanto, essa abordagem é computacionalmente custosa, e algoritmos mais recentes apresentam alternativas mais eficientes.

Uma alternativa proposta pelo XGBoost é a segmentação baseada em histograma. Nesse caso, em vez de iterar por todas as partições possíveis, o modelo constrói um histograma para cada variável e utiliza-os para encontrar a melhor divisão geral entre as

variáveis.

O LightGBM, desenvolvido pela Microsoft, adota uma abordagem mais eficiente para a definição das divisões. Essa abordagem é conhecida como amostragem unilateral baseada em gradiente (GOSS). O GOSS calcula o gradiente para cada ponto de dados e utiliza-o para filtrar os pontos de dados com gradientes baixos. Afinal, os pontos de dados com gradientes baixos já são bem compreendidos, enquanto aqueles com gradientes altos precisam ser melhor aprendidos.

O LightGBM também utiliza uma abordagem chamada Exclusive Feature Bundling (EFB), que acelera a seleção de muitas variáveis correlacionadas. Outra diferença é que o modelo LightGBM é adequado para o crescimento de folhas (leaf-wise growth), enquanto o XGBoost cultiva as árvores em níveis (level-wise growth). Essa diferença pode ser visualizada na Figura [31](#_bookmark107).

Essa diferença teoricamente favorece o LightGBM em termos de precisão, mas também apresenta um maior risco de overfitting (sobreajuste) quando há poucos dados disponíveis. Portanto, é importante que a pessoa considere essas distinções ao escolher entre os dois algoritmos de aumento de gradiente.

Figura 31: Crescimento em folha versus crescimento em nível

**Crescimento em termos de nível**

**Crescimento em função das folhas**

Fonte: Adaptação de [Korstanje](#_bookmark177) (['20'21](#_bookmark177))

Na Figura [31](#_bookmark107), é possível visualizar como cada modelo é ajustado durante o processo de crescimento de árvore em folhas e em níveis. Essa representação gráfica oferece uma compreensão visual das diferenças entre os dois métodos.

No crescimento de árvore em folhas, como no LightGBM, novas folhas são adicio- nadas à árvore de forma iterativa, visando maximizar a redução do erro de treinamento. Isso significa que as árvores são expandidas adicionando folhas, uma a uma, até que o critério de parada seja alcançado.

Por outro lado, no crescimento em níveis, como no XGBoost, as árvores são ex- pandidas em profundidade de forma simultânea em todos os níveis. Ou seja, em cada nível, todas as folhas são expandidas ao mesmo tempo, resultando em um crescimento mais uniforme da árvore.

Essa distinção no modo de crescimento das árvores pode afetar o comportamento e o desempenho do modelo. Portanto, compreender essa diferença é importante ao escolher entre esses algoritmos de aumento de gradiente.

Figura 3'2: XGBoost e LigthGBM regressão

*4.0*

*Observedo*

*Previsto*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*0 50 100 150 200 250 300*

*XGB Regressão*

*4.0*

*Observedo*

*Previsto*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*0 50 100 150 200 250 300*

*LGBM Regressão*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20) Na Figura [3'2](#_bookmark108) é um modelo baseado nos dados coletados da SANEPAR.

* 1. Estudo de Caso Empírico

A previsão da demanda de água é uma preocupação fundamental para muitas organizações e autoridades responsáveis pelo abastecimento de água. A análise de séries temporais é uma abordagem comumente usada para prever padrões futuros com base em dados históricos. Neste estudo de caso, será explorado como a análise de séries temporais pode ser aplicada para prever a demanda de água ao longo do tempo.

* + 1. Defnição do problema

Na subseção [1.'2.1](#_bookmark5) estão as perguntas de pesquisa que serão abordadas no estudo de caso, da pergunta [Q 1](#_bookmark6) à [Q 5](#_bookmark10), com as ramificações da [Q 5](#_bookmark10).

* + 1. Coleta de dadoi,

Na subseção [1.3](#_bookmark16), são apresentadas as variáveis contidas no conjunto de dados co- letado no período de '2018 a '20'20, durante uma grave falta de água que afetou a cidade. Devido a essa situação, foi implementado um rodízio de abastecimento de água para os residentes. Os dados foram coletados em intervalos de uma hora, levando em consideração cada variável, com ênfase na variável-alvo, denominada LT01, que representa o nível do reservatório.

O conjunto de dados possui um total de '26.306 linhas e 9 colunas. Durante a coleta dos dados, verificou-se que eles apresentam padrões sazonais, indicando variações recorrentes ao longo do tempo. Além disso, constatou-se que o consumo diário foi signifi- cativamente afetado no ano de '20'20, diferindo dos anos anteriores, nos quais as mudanças não foram tão significativas.

* + 1. Análii,e exploratória doi, dadoi,

Ao longo do trabalho realizado, póde-se observar na subseção ['2.1](#_bookmark33) que foi realizada uma análise gráfica do problema antes da aplicação de qualquer método. A detecção de anomalias mostrou-se desafiadora, porém não impossível de ser realizada. Essa detecção permitiu a análise da presença de sazonalidade nos dados. A decomposição STL foi utilizada para essa finalidade, conforme descrito na etapa [Etapa 3](#_bookmark22) e detalhado na subseção [4.1.3](#_bookmark125), onde são apresentadas as decomposições realizadas.

É fundamental lembrar que, durante a análise exploratória, os dados sofreram algumas alterações. Por exemplo, a média diária foi calculada em vez de ser considerada a nível horário, resultando em uma redução do conjunto de dados de '26.306 linhas para

1.096 linhas. A decomposição STL foi aplicada nos formatos aditivo e multiplicativo, e

ambas as abordagens estão ilustradas nas Figuras [33](#_bookmark126) e [34](#_bookmark127), respectivamente.

Adicionalmente, na subseção [4.1.3](#_bookmark125), foi realizada a verificação da estacionariedade da série. O teste de Dickey-Fuller (DF) foi empregado para auxiliar na tomada de decisões, e os resultados demonstraram que a série em análise é estacionária, conforme evidenciado pelo teste DF.

* + 1. Ei,colha do modelo

Como os dados apresentam sazonalidade, foram selecionados modelos simples de ARIMA, como AR, MA, ARMA, ARIMA e SARIMA. Esses modelos são univariados. .Já os modelos com variável exógena, como ARX, ARIMAX e SARIMAX, são considerados multivariados. No contexto dos dados analisados, qualquer variável que possa interferir na variável preditora é considerada exógena. Para este caso específico, todas as outras variáveis foram incluídas como exógenas para melhorar a previsão.

Outros modelos utilizados são os modelos de aprendizado de máquina supervisio- nados, como LR, RFR, LightGBM e XGBoost. Esses modelos são regressores baseados em árvores de decisão ou gradientes, especialmente os modelos XGBoost e LightGBM, que são amplamente reconhecidos como eficazes na previsão e tomada de decisões, conforme mencionado por [Chen e Guestrin](#_bookmark163) (['2016](#_bookmark163)) em seu estudo de benchmarking de frameworks de deep learning para tarefas de manutenção preditiva. [Sánchez, Díaz e López](#_bookmark199) (['20'20](#_bookmark199)), em seu estudo comparativo de XGBoost, AdaBoost e CatBoost em algoritmos de aprendizado de máquina, também destacam o desempenho superior do XGBoost em várias métricas de avaliação.

* + 1. Divii,ão doi, dadoi,

Para obter a divisão mais adequada dos dados, verificam-se a média e o desvio padrão de cada um desses conjuntos. O conjunto de dados é dividido em três partes: treinamento, validação e teste. Nessa divisão, utiliza-se inicialmente 70% dos dados para treinamento e validação, e os 30% restantes para teste. Em seguida, a porção de treina- mento e validação é subdividida em 80% para treinamento e '20% para validação.

* + 1. Ajui,te do modelo

Nesta etapa, você aplicará o modelo selecionado aos dados de treinamento. Ajuste os parâmetros do modelo com o objetivo de minimizar os erros de previsão. Dependendo do modelo escolhido, você pode usar técnicas de otimização para encontrar os melhores parâmetros.

Ao ajustar o modelo para a base de dados, foi feita uma alteração na ordem do

modelo sugerido pelo pmdarima. A escolha foi trocar o modelo SARIMAX(1,1,1)('2,1,0,1'2) para SARIMAX(7,1,7)('2,1,0,1'2). Essa decisão foi tomada com base na observação de um ajuste mais preciso aos dados, evidenciado pela redução nos resíduos e uma melhor captura das características da série temporal. Além disso, considerando o conhecimento do problema e as características específicas dos dados, foi identificado que padrões mais complexos requeriam ordens mais altas para serem adequadamente capturados. Dessa forma, foi realizado um processo iterativo de experimentação e avaliação para determinar o modelo SARIMAX(7,1,7)('2,1,0,1'2) como o mais adequado para a base de dados em questão. É importante ressaltar que o desempenho do novo modelo será avaliado por meio de diagnósticos adicionais e análise dos resultados obtidos.

Os modelos XGBRegressor e LGBMRegressor foram ajustados usando as técnicas de GridSearchCV e BayesSearchCV. Essas abordagens permitiram encontrar as melhores combinações de hiperparâmetros para esses modelos, buscando maximizar o desempenho e a precisão das previsões. Por outro lado, os modelos LR (Regressão Linear) e RFR (Random Forest Regressor) não passaram por ajustes, pois não apresentaram melhorias significativas nos resultados após as etapas de GridSearchCV, BayesSearchCV e Ran- domizedSearchCV. Portanto, esses modelos mantiveram as configurações padrão, uma vez que as tentativas de otimização dos hiperparâmetros não resultaram em melhorias substanciais para eles.

* + - * GridSearchCV: O GridSearchCV é uma técnica de busca exaustiva que é usada para ajustar os hiperparâmetros de um modelo de aprendizado de máquina. Ele realiza uma busca sistemática por todas as combinações possíveis de valores espe- cificados para cada hiperparâmetro e avalia o desempenho do modelo para cada combinação. Essa abordagem avalia todas as opções disponíveis, mas pode ser computacionalmente intensiva. Ao final, fornece os melhores hiperparâmetros en- contrados que otimizam a métrica de avaliação escolhida.
      * Bayei,SearchCV: O BayesSearchCV é uma técnica de otimização de hiperparâ- metros baseada em Bayesian optimization. Ele usa um processo de amostragem sequencial para encontrar a melhor combinação de hiperparâmetros de forma mais eficiente do que o GridSearchCV. O BayesSearchCV usa uma função de perda esti- mada e um modelo probabilístico para determinar quais configurações de hiperparâ- metros são mais promissoras e, em seguida, realiza novas amostragens para refinar a busca. Essa abordagem permite uma exploração mais inteligente do espaço de hiperparâmetros e a descoberta de melhores configurações com menos iterações.
      * RandomizedSearchCV: O RandomizedSearchCV é uma técnica de busca aleató- ria de hiperparâmetros. Ao contrário do GridSearchCV, que testa todas as combina-

ções possíveis, o RandomizedSearchCV seleciona aleatoriamente um subconjunto do espaço de hiperparâmetros e avalia o modelo para cada combinação escolhida. Essa abordagem é útil quando o espaço de hiperparâmetros é grande e não é possível tes- tar todas as combinações em tempo razoável. O RandomizedSearchCV permite uma exploração mais ampla do espaço de hiperparâmetros, embora com menor garantia de encontrar a melhor combinação.

* + 1. Avaliação do modelo

A avaliação da precisão dos modelos de previsão é uma etapa fundamental no processo de modelagem. Diversas métricas podem ser utilizadas para esse propósito, como o MAPE, o MAE e o RMSE. Essas métricas têm sido amplamente adotadas na literatura de previsão e são consideradas indicadores confiáveis para mensurar a qualidade das previsões.

De acordo com [Zhang, Xu e Shen](#_bookmark221) (['2016](#_bookmark221)), o MAPE é uma métrica bastante utilizada na avaliação de modelos de previsão, especialmente quando há variações significativas nos dados ou quando se deseja comparar a precisão de diferentes modelos. O MAPE calcula o erro médio percentual entre as previsões e os valores reais, fornecendo uma medida relativa da precisão do modelo.

De acordo com [Villmott e Matsuura](#_bookmark216) (['2005](#_bookmark216)), o uso do erro médio absoluto (MAE) apresenta vantagens na avaliação do desempenho médio de um modelo, em comparação com o erro quadrático médio (RMSE).

[.Jones, Smith e .Johnson](#_bookmark175) (['2017](#_bookmark175)) destacam a importância do RMSE na avaliação de modelos e argumentam contra a exclusão dessa métrica na literatura.

Segundo [Peng et al.](#_bookmark190) (['2017](#_bookmark190)), o MAPE é amplamente utilizado na avaliação de mo- delos de previsão, especialmente quando há variações significativas nos dados ou quando se deseja comparar a precisão de diferentes modelos.

Além disso, [.Jones, Smith e .Johnson](#_bookmark175) (['2017](#_bookmark175)) afirmam que o MAE e o RMSE são métricas amplamente adotadas na análise de previsões, pois fornecem uma medida direta do desvio absoluto e do desvio quadrático médio entre as previsões e os valores observados. O MAE é particularmente útil quando se busca uma medida de erro que não seja sensível a valores extremos, enquanto o RMSE penaliza de forma mais significativa os erros maiores, oferecendo uma visão mais abrangente da precisão do modelo.

Portanto, ao utilizar essas métricas, o pesquisador estará seguindo uma prática comum e fundamentada na literatura. O MAPE permitirá avaliar a precisão relativa das previsões, enquanto o MAE e o RMSE fornecerão uma medida direta dos desvios absolutos e quadráticos, respectivamente. Essas métricas fornecerão uma base sólida

para a avaliação dos modelos de previsão utilizados na pesquisa.

* + 1. Previi,õei, Futurai,

Com base nos modelos AR, ARX, MA, ARMA, ARIMA, ARMAX, SARIMA, SA-

RIMAX, LR, XGBRegressor, LGBMRegressor e RFR, que foram cuidadosamente aplica- dos e avaliados, é possível afirmar que uma vez que a precisão desses modelos tenha sido satisfatória, eles podem ser utilizados para fazer previsões futuras. Aplicando esses mo- delos aos dados futuros disponíveis, é possível estimar a demanda de água para diferentes horizontes de previsão, como um dia, uma semana, duas semanas e um mês.

Essas previsões fornecerão informações valiosas para o planejamento e gerencia- mento eficiente dos recursos hídricos. Ao ter conhecimento antecipado da demanda de água esperada nos próximos períodos, é possível tomar medidas adequadas para garantir o suprimento adequado de água, evitar escassez ou desperdício, e realizar um planejamento eficaz para a distribuição e utilização dos recursos hídricos.

Com base nos resultados significativos obtidos por esses modelos durante o pro- cesso de validação, o pesquisador terá confiança em aplicá-los para previsões futuras de curto prazo. Essas previsões permitirão uma compreensão das tendências e variações na demanda de água ao longo de diferentes períodos, capacitando os responsáveis pela gestão dos recursos hídricos a tomar decisões informadas e estratégicas.

Portanto, uma vez que os modelos tenham sido devidamente avaliados e demons- trado sua eficácia, eles podem ser utilizados para fazer previsões precisas da demanda de água em horizontes de previsão de um dia, uma semana, duas semanas e um mês, auxiliando na gestão e planejamento eficiente dos recursos hídricos.

* + 1. Monitoramento e Ajui,te Contínuo

É importante destacar que todas as questões de pesquisa abordadas neste estudo estão fundamentadas no fator dos horários de pico e nas anomalias que ocorreram durante o período analisado. O comportamento da demanda de água durante os horários de maior consumo e as anomalias observadas foram aspectos-chave que motivaram a realização desta pesquisa.

Ao investigar os efeitos dos horários de pico e das anomalias na demanda de água, o estudo teve como objetivo compreender melhor os padrões de consumo, identificar possí- veis causas para as variações significativas na demanda e desenvolver modelos de previsão mais precisos. A análise desses aspectos contribuiu para uma melhor compreensão dos desafios enfrentados no abastecimento de água e na gestão dos recursos hídricos durante os períodos críticos.

Considerando a importância desses fatores na formulação das questões de pesquisa, as análises realizadas e os modelos desenvolvidos buscaram fornecer insights e informações relevantes para aprimorar a capacidade de previsão e planejamento do abastecimento de água, especialmente durante os horários de pico e diante de anomalias observadas.

* + 1. Principaii, Conclui,ão

Ao longo deste estudo de caso, foram resolvidas as questões de pesquisa levantadas por meio da aplicação da análise de séries temporais para prever a demanda de água. A abordagem adotada demonstrou ser eficaz na obtenção de insights valiosos para o gerenciamento do abastecimento hídrico.

Foi constatado que a análise de séries temporais é uma ferramenta promissora para prever a demanda de água, permitindo tomar decisões informadas e embasadas nesse contexto. Por meio da modelagem e aplicação de diversos modelos, como ARIMA, SARIMA, LR e outros, foi possível analisar e interpretar os dados históricos de maneira precisa, obtendo previsões confiáveis.

Durante o estudo, foram levantadas questões relacionadas à sazonalidade da de- manda de água, inftuência de fatores externos imprevisíveis e mudanças no comportamento dos consumidores. Através da adaptação das técnicas de análise de séries temporais, foi possível abordar essas questões de forma eficiente e obter respostas relevantes para o gerenciamento do abastecimento de água.

Ao longo do processo, foram identificadas anomalias e ftutuações na demanda de água, bem como tendências sazonais específicas. Por meio da análise dos resultados obtidos com os modelos aplicados, foi possível ajustar e aprimorar as previsões, tornando- as mais acuradas e confiáveis.

Em suma, este estudo de caso demonstrou que a análise de séries temporais é uma abordagem eficaz para prever a demanda de água, permitindo uma gestão mais eficiente do abastecimento hídrico. Ao adaptar e aplicar as técnicas adequadas aos dados específicos e às características do contexto, foram resolvidas as questões de pesquisa propostas e obtidos resultados significativos.

Essas descobertas têm o potencial de contribuir para a tomada de decisões embasa- das no planejamento e no gerenciamento da demanda de água, visando a sustentabilidade e a eficiência dos recursos hídricos.

# Resultados

Neste capítulo, é fornecida uma síntese e uma visão geral dos resultados obtidos até o momento. É apresentado um resumo sucinto das principais realizações e descobertas que foram alcançadas até agora.

* 1. Planejamento do Problema

Assim como apresentado na seção [1.4.1](#_bookmark18), os passos da dissertação delinearam o processo pelo qual cada modelo foi construído e os métodos utilizados para responder às questões de pesquisa abordadas na seção [1.'2.1](#_bookmark5). Esses passos proporcionaram uma cronologia lógica das etapas realizadas ao longo do tempo com os dados da SANEPAR, ilustrando o progresso e os resultados alcançados até o momento.

* + 1. Análii,e Exploratória doi, dadoi, (EDA)

A partir do passo [Etapa 1](#_bookmark20), foi realizado o EDA (Exploratory Data Analysis) para processar os dados obtidos até o momento. O EDA permite responder às questões de pesquisa levantadas. Conforme mencionado por [Yu](#_bookmark220) (['2016](#_bookmark220)), na era dos grandes dados, é desafiador descobrir as regras, modelos analíticos e hipóteses por trás dos volumes massi- vos de dados caóticos, não estruturados e multimídia coletados por meio de vários canais. A análise exploratória de dados foi promovida por .John Tukey como uma abordagem para explorar os dados, resumir suas principais características e formular hipóteses que possam direcionar a coleta adicional de dados e experimentos. No contexto de grandes análises de dados, várias técnicas de EDA têm sido adotadas.

Ao analisar a pergunta [Q 1](#_bookmark6), que relaciona a demanda com a variável prevista e a pressão para a variável PT01, pode-se observar na Figura ['25](#_bookmark93) que ambas as variáveis apresentam uma correlação quase perfeita, com um coeficiente de correlação de Pearson

(*r*) igual a 1. Portanto, para responder a essa pergunta, basta observar a correlação de Pearson na Figura ['25](#_bookmark93).

Para responder à pergunta [Q '2](#_bookmark7), é criada uma tabela para fornecer uma resposta mais completa.

Na Tabela [4](#_bookmark123), o desvio padrão é representado pela sigla STD, que corresponde à expressão em inglês "standard dev at on". Além disso, em resposta à pergunta [Q '2](#_bookmark7), é importante mencionar que, assim como em qualquer empresa de tratamento de água, é utilizado um mecanismo de acionamento automático chamado 11trava de segurança11para evitar que o nível do tanque chegue a zero e haja falta de água nos locais abastecidos por esse tanque. O nível mínimo que o tanque pode alcançar é de 5*.*29*m*3 (equivalente a 5,'29

Tabela 4: Descrição estatística dos dados com o filtro aplicado das 18h às '21h

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 18 a 21h | B1 | B2 | B3 | LT01 | FT01 | FT02 | FT03 | PT01 | PT02 |
| Contagem | 4385 | 4385 | 4385 | 4385 | 4385 | 4385 | 4385 | 4385 | 4385 |
| Média | 51,94 | '27,81 | 6,41 | 3,'24 | 11'2,68 | 13'2,93 | 11'2,41 | 4,11 | '20,80 |
| STD | 17,14 | 17,61 | 16,77 | 0,70 | 13'2,59 | 44,78 | 31,33 | 0,76 | 6,14 |
| Min | 0 | 0 | 0 | 0,'29 | 0 | 0 | 0 | 0,88 | 0 |
| 25% | 57,84 | 0 | 0 | '2,79 | 0,1'2 | 1'23,96 | 111,66 | 3,6'2 | '21,7'2 |
| 50% | 57,99 | 34,91 | 0 | 3,30 | 0,1'2 | 136,00 | 118,8'2 | 4,15 | '2'2,05 |
| 75% | 57,99 | 38,0'2 | 0 | 3,78 | '264,'27 | 148,'20 | 1'25,63 | 4,66 | '23,0'2 |
| Max | 59,99 | 59,99 | 59,99 | 4,40 | 383,87 | 3'26,17 | 194,35 | 5,68 | '28,08 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

litros). As bombas são ativadas em sua potência máxima para evitar que sejam acionadas quando o nível do tanque. No entanto, a bomba 1 ainda estaria operando para completar o nível do tanque caso ele esteja dentro dessa faixa.

Em situações de demanda de pico, uma abordagem ideal, embora não necessaria- mente a mais económica, seria ter um tanque de reserva adicional e instalar uma tubulação que os conecte. Durante o dia, ambos os tanques seriam abastecidos e, à noite, por meio da ação da gravidade, eles manteriam o mesmo nível até que o consumo atinja um ponto em que as bombas sejam acionadas. Essa estratégia permite um abastecimento contínuo e eficiente de água.

Na pergunta [Q 3](#_bookmark8), observa-se que o tanque tem uma capacidade máxima de 4*,* 256*m*3, o que equivale a 4*.*256 litros. Para atender a essa demanda e manter o tanque quase cheio ou sempre cheio, é necessário que o ftuxo de entrada esteja na faixa de [238*,* 302] *m*3*/h*, o

ftuxo de gravidade esteja entre [126*,* 182] *m*3*/h*, o ftuxo de retorno esteja entre [110*,* 144] *m*3*/h*, a pressão de sucção esteja entre [1*.*92*,* 4*.*24] *mca* e a pressão de retorno esteja entre [21*,* 24] *mca*.

Para responder à pergunta [Q 4](#_bookmark9), o ponto de equilíbrio, onde as bombas não precisam ser acionadas, ocorre quando o ftuxo de FT01 é de 211 *m*3*/h*, FT0'2 é de 114 *m*3*/h*, FT03 é de 100 *m*3*/h* e o nível do tanque está em 3*.*545 *m*3. No que diz respeito à pergunta [Q 5](#_bookmark10)[a.](#_bookmark11), o nível do tanque deve ser de 4*,* 00 *m*3 para evitar o funcionamento das bombas durante as horas de pico.

* + 1. Múltiplai, entradai, e i,aída única (MISO)

Na etapa [Etapa '2](#_bookmark21), foi explorado o modelo MISO (do inglês M'U t p e Inp'Uts, S n e O'Utp'Ut) na dissertação. O modelo ARIMA, juntamente com suas variantes e extensões, foi amplamente estudado durante a pesquisa, assim como modelos regressivos que envolvem múltiplas variáveis de entrada e uma variável de saída, neste caso, a LT01. As demais

variáveis foram utilizadas como suporte para melhorar o modelo do tipo ARIMAX ou modelos com variáveis exógenas. Quando aplicado sem o uso de variáveis exógenas, o modelo ARIMA apresenta apenas uma entrada, semelhante ao modelo de regressão linear (LR). No entanto, ao incluir variáveis exógenas, o modelo se torna MISO, permitindo uma modelagem mais abrangente e considerando a interação de várias variáveis para prever a variável de interesse.

* + 1. Decompoi,ição STL

A decomposição sazonal e de tendência utilizando o procedimento de Loess (STL) é uma técnica amplamente utilizada para decompor séries temporais em seus componentes sazonais, de tendência e restantes. De acordo com [Theodosiou](#_bookmark209) (['2011](#_bookmark209)), o método STL realiza a decomposição aditiva dos dados por meio de uma sequência de aplicações do Loess mais suave, onde regressões polinomiais ponderadas localmente são aplicadas em cada ponto do conjunto de dados, tendo como variáveis explicativas os valores mais próximos do ponto cuja resposta está sendo estimada.

A decomposição STL é especialmente útil para identificar e isolar padrões sazonais e de tendência presentes nas séries temporais. Ela permite a separação dos componentes sazonais, que ocorrem em intervalos regulares ao longo do tempo, da componente de tendência, que indica a direção geral dos dados ao longo do tempo. A decomposição também resulta em uma componente restante, que representa a variação não explicada pelos componentes sazonais e de tendência.

Ao aplicar a decomposição STL, a série temporal pode ser expressa como a soma dos componentes sazonais, de tendência e restantes. Essa técnica é útil para análise e modelagem de séries temporais, pois proporciona uma compreensão mais clara dos padrões de variação presentes nos dados.

A decomposição STL é formalmente definida como:

*y* = *f* (*S , T , R* ) = *yt* = *St* + *Tt* + *Rt* aditivo

*t*

*t*

*t*

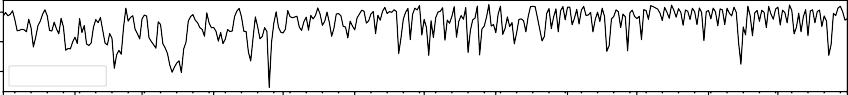
*t*

*yt* = *StTtRt* multiplicativo

('20)

Figura 33: Decomposição STL aditiva dos dados coletados

*4*



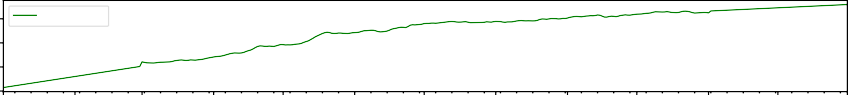
*LT01 (m³)*

*Observado*

*3*

*2*

*3.75*



*Tendência*

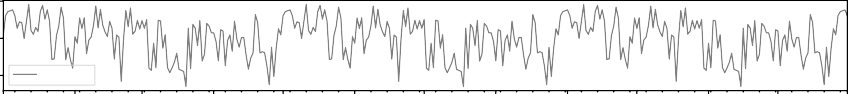
*Tendência*

*3.50*

*3.25*

*3.00*

*0.5*



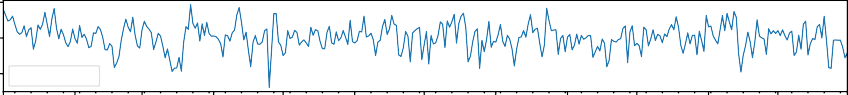
*Sazonal*

*0.0*

*Sazonal*

*0.5*

*1*



*Residual*

*0*

*Residual*

*1*

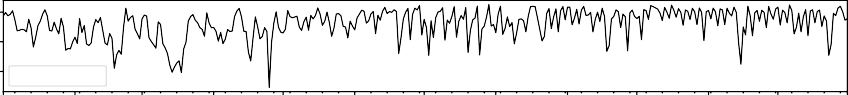
*Jan 2020*

*Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec*

*Decompocisão STL*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20) Figura 34: Decomposição STL multiplicativa dos dados coletados

*4*



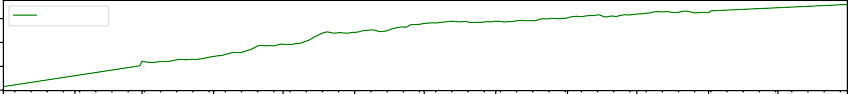
*LT01 (m³)*

*Observado*

*3*

*2*

*3.75*



*Tendência*

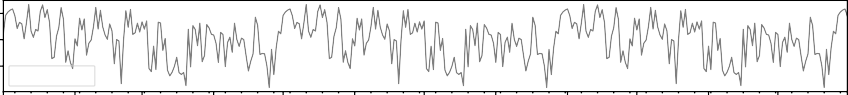
*Tendência*

*3.50*

*3.25*

*3.00*

*1.1*



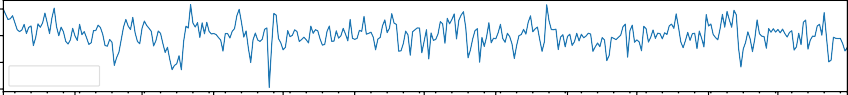
*Sazonal*

*1.0*

*Sazonal*

*0.9*

*1.25*



*Residual*

*1.00*

*Residual*

*0.75*

*0.50*

*Jan 2020*

*Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec*

*Decompocisão STL*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Na resposta à pergunta [Q 5](#_bookmark10)[b.](#_bookmark12), as Figuras [33](#_bookmark126) e [34](#_bookmark127) fornecem informações sobre a presença de tendência, sazonalidade e resíduos na série temporal.

Através da decomposição, é possível analisar se a série apresenta tendência, sa- zonalidade e resíduos. Ao observar as Figuras [33](#_bookmark126) e [34](#_bookmark127), é evidente que os dados exibem ambos os padrões. Isso indica que a série é estacionária, como confirmado pelo seguinte teste.

Teste de Dickey-Fuller (DF) Aumentado:

* Estatística de teste ADF: −4*.*248
* Valor de p: 0*.*001
* Atrasos utilizados: 21*.*000
* Observações: 1074*.*000
* Valor crítico (1%): −3*.*436
* Valor crítico (5%): −2*.*864
* Valor crítico (10%): −2*.*568

Com base na forte evidência contra a hipótese nula, podemos rejeitar a hipótese nula. Isso indica que os dados não possuem raiz unitária e são estacionários em [Q 5](#_bookmark10)[c.](#_bookmark13). Identificar as horas de pico entre 18h e '21h não é uma tarefa fácil. No entanto, ao observar a Figura [35](#_bookmark128), podemos notar um aumento na demanda durante essas horas durante o ano de '20'20.

Figura 35: Violino no nível do reservatório

*4.5*

*4.0*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*1.0*

*Nivel do Reservatorio - LT01*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Conforme mencionado na subseção [1.1.1](#_bookmark3), as anomalias climáticas ocorridas em '20'20, especialmente a falta de chuvas, tiveram um impacto significativo nos resultados. Isso contribuiu para as mudanças observadas na demanda de água ao longo desse período. Com relação à pergunta [Q 5](#_bookmark10)[d.](#_bookmark14), durante as horas de pico, é necessário que o nível do tanque esteja dentro da faixa de [3*.*545*,* 4*.*256]*m*3 para evitar o acionamento das bombas.

Manter o nível do tanque dentro dessa faixa permitirá que o sistema opere de forma eficiente, atendendo à demanda sem a necessidade de acionar as bombas.

Figura 36: Violino da vazão de recalque

*160*

*140*

*120*

*100*

*80*

*FT03 m³/h*

*60*

*40*

*20*

*0*

*Vazão de Recalque - FT03*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Para responder à pergunta [Q 5](#_bookmark10)[e.](#_bookmark15), a Figura [36](#_bookmark129) ilustra como a vazão pode ser afetada pelo nível do tanque. É interessante observar que a vazão de recalque tem um impacto mais significativo no nível do tanque em comparação com as outras vazões. Isso ocorre porque a vazão de recalque está associada à injeção de água diretamente no tanque por meio da bomba localizada próxima à base do tanque. Por outro lado, as demais vazões apresentam alguns valores ausentes, o que limita sua inftuência na análise geral.

De acordo com o [Reisen et al.](#_bookmark193) (['2017](#_bookmark193)), o teste DF tem as seguintes equações

*zt* = *yt* + *θβt, t* = 1*, . . . , T,* ('21)

*ρ*ˆDF − 1 =

*T*

*t*=1 *T*

*zt−*1∆*zt z*2

('2'2)

*t*=1

*t−*1

De (['2'2](#_bookmark130)) onde ∆*zt* = *zt* − *zt−*1. Sob a hipótese nula (*H*0) : " *ρ* = 1", as estatísticas do teste DF e suas distribuições limitantes são dadas da seguinte forma:

*T* (*ρ*ˆDF − 1) = *T*

*T*

*t*=1 *T*

*zt−*1∆*zt z*2

('23)

*t*=1

e

*t−*1

*τ*ˆ =

*ρ*ˆDF − 1

('24)

DF

*σ*ˆDF

*T*

*t*=1

*z*2

*t−*1

*−*1*/*2

De (['24](#_bookmark131)) onde *σ*ˆ2

= *T−*1 *T*

(∆*zt* − (*ρ*ˆDF − 1) *zt−*1)2 *.*

Suponha que (*zt*)1*≤t≤T* são dadas por (['21](#_bookmark130)), então quando *ρ* = 1,

DF

*t*=1

*d W* (1)2 − 1

0

( *θ* )2 *π*

*σ*

0

*τ*ˆ −→*d*

DF

1 + 2(*θ/σ*)2*π*

*−*1*/*2 



] 

*T* (*ρ*ˆDF − 1) −→ 2 1 *W* (*r*)2 d*r* −

1 *W* (*r*)2 d*r,* como *T* → ∞ ('25)

*W* (1)2 − 1

1

1*/*2 − 1

(*θ/σ*)2*π*

1*/*2  ('26)

2 0 *W* (*r*)2 d*r*

0 *W* (*r*)2 d*r*

como *T* → ∞ ('27)



A partir de (['27](#_bookmark132)), onde−→*d* denota convergência na distribuição e onde {*W* (*r*)*, r* ∈

[0*,* 1]} denota o movimento Browniano padrão.

O ACF (do inglês A'Uto-Corre at on 'Unct on) é uma medida estatística utilizada para identificar a presença de correlação serial em uma série temporal. Ele calcula a autocorrelação entre os valores da série em diferentes defasagens, ou seja, a correlação entre os valores atuais e os valores passados da série.

O ACF é útil para analisar a dependência temporal dos dados e identificar padrões de sazonalidade, tendência ou outros efeitos temporais. Através do ACF, é possível avaliar se a série exibe autocorrelação significativa em defasagens específicas, o que pode indicar a presença de não estacionariedade ou estrutura temporal que precisa ser considerada na análise ou modelagem da série temporal.

Figura 37: Autocorrelação e Autocorrelação parcial

*Autocorrelation*

*1.00*

*0.75*

*0.50*

*0.25*

*0.00*

*0.25*

*0.50*

*0.75*

*1.00*

*0 5 10 15 20 25 30*

*Partial Autocorrelation*

*1.00*

*0.75*

*0.50*

*0.25*

*0.00*

*0.25*

*0.50*

*0.75*

*1.00*

*0 5 10 15 20 25 30*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Na Figura [37](#_bookmark133), é possível observar a diferença entre a autocorrelação e a autocorrela- ção parcial (PACF). A autocorrelação mede a correlação entre os valores da série temporal em diferentes defasagens, levando em consideração tanto a correlação direta quanto a cor- relação indireta. Por outro lado, a autocorrelação parcial mede apenas a correlação direta entre os valores, eliminando a inftuência das defasagens intermediárias.

O intervalo de confiança padrão de 95% é representado pela marca azul na Fi- gura. As observações que estão fora desse intervalo são consideradas estatisticamente correlacionadas, indicando a presença de padrões ou estrutura na série temporal.

A correlação visualizada na Figura [37](#_bookmark133) é fundamental para a interpretação do teste DF. Em uma série de ruído branco, os valores são completamente aleatórios e não apre- sentam correlação significativa. Portanto, quando há correlação presente na série, isso indica a existência de padrões ou dependências entre os valores, o que pode ser explorado para a modelagem e previsão da série temporal.

Na Figura [38](#_bookmark135), é possível observar uma série temporal que pode ser caracterizada como ruído branco. Uma série temporal é considerada ruído branco se suas variáveis forem independentes e distribuídas de forma idêntica, com média zero. Isso implica que todas as variáveis possuem a mesma variância (*σ*2) e que cada valor não possui correlação com os demais valores da série.

Além disso, é importante destacar o comprimento dos zeros na variável prevista, o que conclui a etapa [Etapa 3](#_bookmark22).

4.1.4 Separação doi, dadoi,

Na etapa [Etapa 4](#_bookmark23), os dados foram divididos em conjuntos de treinamento, teste e validação. Essa prática é comum entre profissionais de aprendizado de máquina, pois

Figura 38: Ruído branco

*1.5*

*1.0*

*0.5*

*0.0*

*LT01 (m³)*

*0.5*

*1.0*

*1.5*

*2.0*

Fonte:

*2020-01 2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*Tempo*

Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

permite avaliar o desempenho do modelo em conjuntos de dados diferentes.

Em relação ao processamento de modelos de aprendizado profundo, é importante mencionar as inovações trazidas pela empresa Nvidia ao longo dos anos, especialmente no campo do processamento de imagens. O lançamento da placa de vídeo GeForce RTX 4090 tem sido bastante aguardado tanto por gamers quanto por profissionais que lidam com aprendizado de máquina.

No contexto do estudo, foram utilizados dois computadores para realizar os cálcu- los dos modelos. Um deles é equipado com um processador Intel Core i5-3330 e o outro é um notebook com um processador Intel Core i7-5500. Ambos os processadores possuem 4 threads, sendo que o notebook possui '2 núcleos físicos e o i5 possui 4 núcleos físicos. Cada processador tem suas especificações e desempenho adequados a diferentes necessi- dades. Vale ressaltar que não é obrigatório utilizar as últimas gerações de processadores para realizar esses processamentos, e sim compreender e aplicar corretamente os recursos disponíveis.

Quanto à divisão dos dados, foi adotada uma estratégia básica em que 70% dos dados foram destinados ao conjunto de treinamento e os 30% restantes foram reservados para o conjunto de teste. Dentro dos 70% de treinamento, foi realizada uma subdivisão em que 80% desses dados foram usados novamente para treinamento e os '20% restan- tes foram utilizados para validação. Essa abordagem foi implementada em linguagem de programação para facilitar o processo e evitar a necessidade de recalculá-la a cada modificação do modelo.

* + 1. Ei,tratégia de Previi,ão

A estratégia recursiva é mencionada por [Petropoulos et al.](#_bookmark191) (['20'2'2](#_bookmark191)) como uma abor- dagem eficaz na previsão de séries temporais de múltiplos passos. De acordo com o autor, essa estratégia envolve o uso de previsões anteriores como entradas para prever os próxi- mos passos da série temporal. A abordagem recursiva tem demonstrado potencial para melhorar a acurácia das previsões de séries temporais de longo prazo.

Na Etapa [Etapa 5](#_bookmark24), discute-se a previsão dos dados em uma janela de horizonte de previsão estendida, abrangendo diferentes períodos de tempo, como um dia, uma semana, duas semanas e um mês. Essa estratégia de previsão recorrente permite a comparação entre modelos de regressão e modelos ARIMA em diferentes horizontes temporais.

Essa abordagem é vantajosa, pois cada modelo possui suas próprias características e desempenho ao lidar com previsões de curto prazo, como um dia, e previsões de prazo mais longo, como um mês. Ao utilizar uma janela de previsão mais ampla, é possível observar e avaliar melhor as diferenças entre os modelos e analisar seu desempenho em horizontes de tempo variados.

* + 1. Horizonte

Na etapa [Etapa 6](#_bookmark25), o horizonte de previsão foi personalizado com base no método recursivo de previsão de série temporal e na previsão do nível do tanque LT01. Foram selecionados os seguintes passos para a previsão à frente: um dia, uma semana, duas semanas e um mês. Essa escolha do horizonte de previsão foi feita levando em consideração a estratégia recursiva e os objetivos específicos do estudo. Identifica-se que essa janela de tempo proporciona uma análise mais adequada e comparável entre os modelos utilizados

* + 1. Modeloi, de previi,ão e métricai, de dei,empenho

A partir da etapa [Etapa 7](#_bookmark26), foram utilizadas três métricas amplamente empregadas na literatura para a previsão e comparação de modelos ARIMA e modelos de regressão. Essas métricas foram detalhadas na seção [3.1](#_bookmark63).

Ao analisar os modelos desenvolvidos, foi observado que o modelo de regressão linear (LR) obteve o melhor desempenho tanto na previsão de curto prazo, considerando uma modelagem de '24 horas, quanto nas horas de pico entre 18h e '21h. Os modelos MA, AR, SARIMA, ARIMA, SARIMAX, ARIMAX, ARX, LGBMRegressor, XGBRegressor e

RFR também apresentaram um desempenho satisfatório, seguindo uma ordem de melhor para pior.

Para previsões de longo prazo, como no caso dos 30 dias, foram avaliados os mo- delos ARMA, AR, MA, ARIMA, ARIMAX, ARX, SARIMA, SARIMA, XGBRegressor,

RFR, LGBMRegressor e LR, novamente seguindo a ordem de melhor desempenho. No entanto, ao analisar os resultados graficamente nos apêndices, foi observado que os mode- los que incorporam variáveis exógenas parecem ter uma capacidade de previsão superior em relação aos demais modelos. Essa tendência pode ser visualizada nas Figuras de [4'2](#_bookmark245) a [53](#_bookmark258) e nas Tabelas de [7](#_bookmark223) a ['2'2](#_bookmark238).

* + 1. Tei,te de Signifcância

Na etapa [Etapa 8](#_bookmark27), foi utilizado o teste de Friedman e Nemenyi para comparar as classificações médias entre os classificadores. O teste de Nemenyi é um teste de compa- ração múltipla utilizado após a aplicação de testes não paramétricos com três ou mais fatores.

Tabela 5: Teste Nemenyi

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Nemenyi | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 0 | 1,000 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 |
| 1 | 0,001 | 1,000 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,157 |
| 2 | 0,001 | 0,001 | 1,000 | 0,847 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 |
| 3 | 0,001 | 0,001 | 0,847 | 1,000 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 |
| 4 | 0,001 | 0,001 | 0,001 | 0,001 | 1,000 | 0,001 | 0,001 | 0,001 | 0,001 |
| 5 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 1,000 | 0,001 | 0,001 | 0,001 |
| 6 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 1,000 | 0,001 | 0,001 |
| 7 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 1,000 | 0,001 |
| 8 | 0,001 | 0,157 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 0,001 | 1,000 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Para calcular a estatística de teste *Fr* de Friedman, inicialmente cria-se uma tabela com os dados, onde cada linha representa uma amostra e cada coluna representa uma condição de teste. Em seguida, as amostras são ordenadas ao longo das condições, da melhor situação para a pior. Se não houver empates, a estatística de teste *Fr* é calculada utilizando a seguinte fórmula:

*Fr* =

12 *k*

*nk*(*k* + 1)

*i*=1

2 − 3*n*(*k* + 1) ('28)

Nessa fórmula, *n* é o número de linhas (ou amostras), *k* é o número de colunas (ou condições) e *Ri* é a soma das fileiras da coluna (ou condição) *i*.

*R*

*i*

Além disso, o valor crítico CD (Critical Difference) é utilizado para determinar se dois classificadores são significativamente diferentes um do outro. O CD é calculado

usando a fórmula que mencionei anteriormente:

*CD* = *qα*

*k*(*k* + 1) 6*N*

('29)

Na fórmula do CD, *qα* é o valor crítico obtido da tabela de teste de Nemenyi, *k* é o número de classificadores e *N* é o número total de amostras.

De acordo com essa equação, os resultados da pesquisa foram os seguintes:

*statistic* = 8015*.*611*, p*−*value* = 0*.*0 com um total de '26.306 linhas por 9 colunas.

* + 1. Comparação doi, modeloi,

Com o objetivo de obter uma análise mais aprofundada do desempenho de cada modelo, foi realizada uma comparação por meio de um gráfico de violino. Dessa forma, póde-se observar qual dos modelos apresentava o melhor desempenho.

Figura 39: Comparação dos modelos ARIMAS

*5*

*4*

*3*

*2*

*LT01 (m³)*

*1*

*0*

*AR ARX MA ARMA ARIMA SARIMA ARIMAX SARIMAX*

*Comparação dos Modelos ARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 40: Comparação de modelos de regressão

*5.5*

*5.0*

*4.5*

*4.0*

*LT01 (m³)*

*3.5*

*3.0*

*2.5*

*Regressão Linear XGBRegressor LGBMRegressor Regressor de Floresta Aleatória*

*Comparação dos Modelos Regressores*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Ao comparar os modelos apresentados nas Figuras [39](#_bookmark142) e [40](#_bookmark143), é possível observar quais são os modelos que se destacam, levando em consideração a modelagem dos dados. Os modelos ARIMA que mostram melhor desempenho são o AR, ARX, MA, ARMA, ARIMAX e SARIMAX, devido à sua capacidade de lidar com o'Ut ers e limites inferiores em alguns modelos. No caso dos modelos baseados em gradientes e regressão, é perceptível que eles exibem resultados semelhantes, graças às técnicas de otimização matemática conhecidas como Grid Search e Randomized Search, que permitem aprimorar os métodos utilizados.

Quando se trata de um horizonte de previsão curto, o modelo de LR apresenta melhor desempenho em comparação com os demais. No entanto, em horizontes de pre- visão mais longos, os modelos XGBoost e Light GBM demonstram maior precisão. A Random Forest também é capaz de realizar previsões precisas, ficando ligeiramente atrás do XGBoost em previsões de longo prazo.

Para avaliar a eficiência dos modelos ARIMA em previsões de longo prazo, utiliza-se o método conhecido como Ljung-Box. Os modelos que mostram melhor desempenho nesse contexto são o ARX, ARIMAX e SARIMAX, os quais incorporam variáveis exógenas. Esses modelos não lineares têm capacidade de previsão mais robusta em horizontes de tempo mais distantes, em comparação com os outros modelos ARIMA.

* 1. Estudo de Caso Empírico Resultado

A previsão da demanda de água é uma preocupação fundamental para muitas organizações e autoridades responsáveis pelo abastecimento de água. Neste estudo de

caso, explorou-se como a análise de séries temporais pode ser aplicada para prever a demanda de água ao longo do tempo.

A análise de séries temporais é uma abordagem comumente utilizada para prever padrões futuros com base em dados históricos. No estudo, foram aplicadas técnicas de modelagem e previsão, permitindo obter insights valiosos sobre a demanda de água futura. Diversos modelos, como ARIMA e SARIMA, foram empregados para analisar os dados históricos e gerar previsões confiáveis.

Ao longo do estudo, identificaram-se sazonalidades na demanda de água, bem como padrões de consumo que variam ao longo do tempo. Essas informações são essenciais para o planejamento adequado do abastecimento de água, permitindo uma alocação eficiente dos recursos e uma resposta adequada às ftutuações de demanda.

A aplicação da análise de séries temporais na previsão da demanda de água propor- ciona uma base sólida para a tomada de decisões informadas. Com base nos resultados obtidos, é possível ajustar estratégias de gerenciamento, antecipar picos de demanda e otimizar o uso dos recursos hídricos disponíveis.

Em suma, este estudo demonstrou que a análise de séries temporais é uma abor- dagem eficaz para prever a demanda de água ao longo do tempo. Ao fornecer insights precisos e confiáveis, essa técnica contribui para o planejamento e o gerenciamento efici- ente do abastecimento de água, promovendo a sustentabilidade e a utilização racional dos recursos hídricos.

* + 1. Dei,crição do i,ii,tema de abai,tecimento de água

Neste estudo, foram realizadas análises e modelagens utilizando a abordagem de séries temporais para prever a demanda diária de água em uma determinada cidade para os próximos seis meses. Os resultados obtidos forneceram insights valiosos sobre a demanda futura e contribuíram para um melhor planejamento do abastecimento hídrico. A seguir, apresentam-se as principais conclusões para cada uma das perguntas de pesquisa:

[Q 1](#_bookmark6): Qual é a adequação da pressão atual para atender à demanda diária?

Após análise dos dados e das métricas utilizadas, conclui-se que a pressão atual é adequada para atender à demanda diária. Durante o período analisado, não foram identificadas situações de pressão insuficiente que afetassem o fornecimento de água.

[Q '2](#_bookmark7): Qual é o volume mínimo de água necessário no reservatório para evitar o acionamento das bombas durante o horário de pico?

Com base na frequência de funcionamento das bombas e na demanda durante o horário de pico, determinou-se que é necessário manter um volume mínimo de água no reservatório, correspondente a 5'285,90 litros, para evitar o acionamento das bombas nesse

período.

[Q 3](#_bookmark8): Qual é a vazão ótima para atender à demanda diária?

Após análise e modelagem dos dados, identificou-se que a vazão ótima para atender à demanda varia conforme o período do dia e as características sazonais. A pressão necessária para atender à demanda é de 3,60 PSI (pound-force per square inch) na sucção.

[Q 4](#_bookmark9): Como encontrar o ponto de equilíbrio entre a demanda e a vazão?

Após análise e modelagem dos dados, foi constatado que não existe um ponto de equilíbrio entre a demanda e a vazão no reservatório. No entanto, identificou-se um volume mínimo de reserva de 3.545 litros que permite manter um armazenamento adequado no reservatório sem a necessidade de acionar as bombas durante o período de maior custo energético.

Embora essa estimativa de volume mínimo seja importante para garantir o abas- tecimento contínuo durante o período de pico, é importante ressaltar que não há um equilíbrio perfeito entre a demanda e a vazão nos dados analisados. Portanto, é neces- sário considerar estratégias adicionais, como otimização do sistema de abastecimento e gerenciamento eficiente dos recursos hídricos, para atender de forma adequada às neces- sidades da população.

[Q 5](#_bookmark10): Qual é o impacto do acionamento das bombas durante o horário de pico?

Confirmou-se que a ativação das bombas de sucção durante o período de 18h às '21h resulta em um maior custo energético para a SANEPAR. Portanto, é recomendado evitar o acionamento das bombas durante esse período, utilizando estratégias de armazenamento e gerenciamento eficientes.

Em suma, os resultados obtidos neste estudo fornecem informações valiosas para o planejamento e gerenciamento do abastecimento de água. A abordagem de séries tem- porais mostrou-se eficaz na previsão da demanda futura e na identificação de estratégias para otimizar o uso dos recursos hídricos. Essas conclusões têm o potencial de contri- buir para uma gestão mais eficiente e sustentável do abastecimento de água, garantindo o atendimento adequado às necessidades da população.

* + 1. Análii,e exploratória doi, dadoi,

Ao longo do trabalho realizado, póde-se observar, na subseção ['2.1](#_bookmark33), que foi realizada uma análise gráfica do problema antes da aplicação de qualquer método. A detecção de anomalias mostrou-se desafiadora, porém não impossível de ser realizada. Essa detecção permitiu a análise da presença de sazonalidade nos dados. A decomposição STL foi utilizada para essa finalidade, conforme descrito na etapa [Etapa 3](#_bookmark22) e detalhado na subseção [4.1.3](#_bookmark125), onde são apresentadas as decomposições realizadas.

É fundamental lembrar que, durante a análise exploratória, os dados sofreram algumas alterações. Por exemplo, foi calculada a média diária em vez de ser considerado o nível horário, resultando em uma redução do conjunto de dados de '26.306 linhas para

1.096 linhas. A decomposição STL foi aplicada nos formatos aditivo e multiplicativo, e ambas as abordagens estão ilustradas nas Figuras [33](#_bookmark126) e [34](#_bookmark127), respectivamente.

Adicionalmente, na subseção [4.1.3](#_bookmark125), foi realizada a verificação da estacionariedade da série. O teste de Dickey-Fuller (DF) foi empregado para auxiliar na tomada de decisões, e os resultados demonstraram que a série em análise é estacionária, conforme evidenciado pelo teste DF.

Essa análise exploratória dos dados permitiu ao pesquisador obter insights sobre os padrões e tendências presentes nas variáveis estudadas, auxiliando na compreensão do comportamento do sistema de abastecimento de água durante o período analisado.

* + 1. Quei,tõei, de pei,quii,a 1 a 4

As questões de pesquisa levantadas neste estudo foram cuidadosamente abordadas e respondidas ao longo da análise. A seguir, apresenta-se as respostas para cada uma das questões:

[Q 1](#_bookmark6) Com base nos resultados obtidos, conclui-se que as pressões atuais das variáveis PRESSAO DE SUCÇAO - PT01 e PRESSAO DE RECALQUE - PT0'2 são adequadas

para atender à demanda diária. O percentil 10 das pressões de sucção (3*,* 48 mca) indica que apenas 10% dos valores estão abaixo desse limite, o que sugere que a pressão de sucção geralmente se mantém em níveis adequados para o funcionamento adequado do sistema. Da mesma forma, o percentil 90 das pressões de recalque (24*.*02 mca) indica que apenas 10% dos valores estão acima desse limite, evidenciando que a pressão de recalque também se mantém dentro dos padrões necessários para atender à demanda diária.

Esses resultados indicam que as pressões de sucção e de recalque estão em confor- midade com as exigências do sistema, fornecendo a pressão necessária para o adequado abastecimento de água.

[Q '2](#_bookmark7) Com base na frequência de funcionamento das bombas e na demanda durante o horário de pico, determinou-se que é necessário manter um volume mínimo de água no reservatório, correspondente a 5'285,90 litros, para evitar o acionamento das bombas nesse período.

A vazão ótima para atender à demanda diária do tanque é determinada pelas faixas de ftuxo de entrada, gravidade e retorno, juntamente com as faixas de pressão de sucção e retorno. Com base nas informações fornecidas na pergunta [Q 3](#_bookmark8), para manter o tanque quase cheio ou sempre cheio, as seguintes faixas de vazão devem ser consideradas:

* + - * Fluxo de entrada: entre 238 *m*3*/h* e 302 *m*3*/h*.
      * Fluxo de gravidade: entre 126 *m*3*/h* e 182 *m*3*/h*.
      * Fluxo de retorno: entre 110 *m*3*/h* e 144 *m*3*/h*.
      * Pressão de sucção: entre 1*,* 92 *mca* e 4*,* 24 *mca*.
      * Pressão de retorno: entre 21 *mca* e 24 *mca*.

Essas faixas de vazão e pressão garantem que a demanda diária do tanque seja atendida de forma adequada, mantendo o nível de água próximo ao máximo e garantindo a pressão necessária para o funcionamento adequado do sistema de abastecimento de água.

Para responder à pergunta [Q 4](#_bookmark9) sobre o ponto de equilíbrio entre a demanda e a vazão, o sistema alcança o equilíbrio quando a vazão da FT01 é de '211 *m*3*/h*, a vazão da FT0'2 é de 114 *m*3*/h*, a vazão da FT03 é de 100 *m*3*/h* e o nível do tanque está em 3.545 *m*3. Nesse ponto de equilíbrio, as bombas não precisam ser acionadas, o que indica que o sistema de abastecimento de água está em uma condição estável. Esses valores de vazão e nível do tanque permitem atender à demanda diária sem a necessidade de tomar medidas adicionais.

* + 1. Quei,tão de pei,quii,a 5

[Q 5](#_bookmark10) Confirmou-se que a ativação das bombas de sucção durante o período de 18h às '21h resulta em um maior custo energético para a SANEPAR. Portanto, é recomen- dado evitar o acionamento das bombas durante esse período, utilizando estratégias de armazenamento e gerenciamento eficientes.

1. Verificou-se que, para evitar o acionamento das bombas durante o horário de pico (18h às '21h) sem comprometer o abastecimento de água para a população, é necessário manter o nível do reservatório acima de 4.000 litros.
2. Ao analisar os dados dos últimos 3 anos do Bairro Alto, identificou-se a presença de tendências sazonais e padrões de consumo de água. Essas informações são valiosas para compreender os padrões de demanda e planejar o abastecimento de forma mais eficiente.
3. Observou-se que os horários de pico, nesse caso, correspondem aos períodos em que há maior consumo de água. Esses horários são críticos para o abastecimento, pois a demanda é significativamente maior, exigindo uma gestão cuidadosa dos recursos hídricos nesse intervalo de tempo. É importante monitorar e garantir que haja suprimento adequado nesses horários para atender à demanda da população.

O gráfico de barras apresentado na Figura [41](#_bookmark149) mostra a demanda média das variáveis de ftuxo (Vazão de Entrada-FT01, Vazão de Gravidade-FT0'2 e Vazão de Recalque-FT03)

Figura 41: Demanda Média das Variáveis de Fluxo

200

Vazão FT01 Vazão FT02

Vazão FT03

175

150

125

Vazão Média

100

75

50

25

0

18 19

20 21

Hora

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

durante o intervalo das 18h às '21h. Cada barra representa a média da demanda para cada variável em um horário específico dentro desse intervalo. A altura de cada barra indica a magnitude da demanda média para a respectiva variável. Essa visualização permite que sejam identificados os horários em que as variáveis de ftuxo apresentaram maior demanda, o que é útil para o planejamento e gerenciamento adequado do sistema.

A questão de pesquisa [Q 5](#_bookmark10)[c.](#_bookmark13) foi respondida através da análise dos dados, per- mitindo a identificação dos horários de maior demanda durante o período das 18h às '21h. A tabela a seguir apresenta os resultados para as três variáveis estudadas: vazão de entrada-FT01, vazão de gravidade-FT0'2 e vazão de recalque-FT03.

Tabela 6: Demanda de água

|  |  |  |
| --- | --- | --- |
| Variável | Horário de Maior Demanda | Valor da Demanda |
| Vazão de entrada - FT01 | '20'20-10-08 '21:00:00 | 383,87 *m*3*/h* |
| Vazão de gravidade - FT0'2 | '20'20-10-'20 18:00:00 | 3'26,17 *m*3*/h* |
| Vazão de recalque - FT03 | '20'20-11-'26 19:00:00 | 194,35 *m*3*/h* |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Os resultados destacam os horários específicos em que cada variável apresentou maior demanda dentro do intervalo das 18h às '21h, fornecendo insights importantes para o

planejamento e gerenciamento adequado do sistema. A tabela [6](#_bookmark150) resume essas informações.

1. Durante as horas de pico, é necessário que o nível do reservatório esteja mantido dentro da faixa de [3*.*545*,* 4*.*256]*m*3 para evitar o acionamento das bombas. Manter o nível do reservatório dentro dessa faixa permitirá que o sistema opere de forma eficiente, atendendo à demanda de água sem a necessidade de acionar as bombas.
2. É importante destacar que a vazão de recalque exerce um impacto mais signi- ficativo no nível do reservatório em comparação com as outras vazões. Essa diferença se deve ao fato de que a vazão de recalque está diretamente relacionada à injeção de água no reservatório por meio da bomba localizada próxima à sua base. Em contraste, as demais vazões possuem alguns valores ausentes, o que limita sua inftuência na análise geral do sistema.
   * 1. Dii,cui,i,ão geral e conclui,õei,

Nesta seção, serão discutidos os principais resultados e conclusões deste estudo sobre a previsão da demanda de água usando a abordagem de séries temporais. Ao longo da análise e interpretação dos dados, foram identificados padrões sazonais e tendências na demanda de água, além de estratégias para otimizar o abastecimento e o gerenciamento dos recursos hídricos.

Durante a análise exploratória dos dados, observou-se que a demanda de água apresenta ftutuações ao longo do tempo, com variações sazonais significativas. A decom- posição STL foi uma ferramenta útil para identificar essas sazonalidades e tendências, fornecendo uma visão mais detalhada do comportamento do sistema de abastecimento de água.

Com base nas questões de pesquisa, pode-se concluir que a pressão atual do sistema é adequada para atender à demanda diária, sem ocorrência de pressão insuficiente que possa prejudicar o fornecimento de água aos consumidores. Além disso, determinou-se um volume mínimo de reserva no reservatório, levando em consideração a frequência de operação das bombas e a demanda durante o horário de pico. Essa reserva mínima visa evitar o acionamento das bombas nesse período, contribuindo para a eficiência energética e reduzindo os custos operacionais.

A análise também permitiu identificar a vazão ótima para atender à demanda diária, considerando as ftutuações sazonais e as diferentes partes do dia. No entanto, observou-se que não há um equilíbrio perfeito entre a demanda e a vazão nos dados ana- lisados. Portanto, recomenda-se explorar estratégias adicionais, como otimização do sis- tema de abastecimento e gerenciamento eficiente dos recursos hídricos, a fim de aprimorar ainda mais a eficiência e a sustentabilidade do abastecimento de água.

Os resultados obtidos neste estudo demonstram a aplicação efetiva da análise de séries temporais na previsão da demanda de água e na otimização do abastecimento hídrico. Eles fornecem insights valiosos para o planejamento e o gerenciamento eficiente do sistema de abastecimento de água, contribuindo para a sustentabilidade e a utilização racional dos recursos hídricos.

Ao considerar os resultados e as conclusões deste estudo, é recomendado que medi- das adicionais sejam adotadas para aprimorar ainda mais a eficiência e a sustentabilidade do abastecimento de água. Isso pode envolver a implementação de estratégias de conser- vação de água, o desenvolvimento de fontes alternativas de abastecimento e a promoção de conscientização sobre o uso responsável da água entre os consumidores.

Em suma, este estudo fornece uma base sólida para a tomada de decisões infor- madas no planejamento e gerenciamento do abastecimento de água. A análise de séries temporais mostrou-se uma ferramenta eficaz para prever a demanda futura e identificar estratégias para otimizar o uso dos recursos hídricos. Essas conclusões têm o potencial de contribuir para uma gestão mais eficiente e sustentável do abastecimento de água, ga- rantindo o atendimento adequado às necessidades da população e o cuidado com o meio ambiente.

# Conclusões

Na dissertação realizada, foi conduzido um estudo abrangente sobre a previsão da demanda de água por meio da análise de séries temporais. Através da análise exploratória dos dados e da aplicação da decomposição STL, foram identificados padrões sazonais e tendências na demanda de água, fornecendo insights valiosos para o planejamento e gerenciamento eficiente do sistema de abastecimento de água.

Com base nos resultados obtidos, conclui-se que a abordagem de séries temporais é uma ferramenta eficaz para prever a demanda futura de água. Os resultados também indicaram a importância de considerar as ftutuações sazonais e as diferentes partes do dia ao determinar a vazão ótima e o volume mínimo de reserva no reservatório.

Apesar dos avanços alcançados nesta pesquisa, é importante ressaltar que existem algumas limitações a serem consideradas. Primeiramente, a análise foi baseada em da- dos históricos de demanda de água de uma única região, limitando a generalização dos resultados para outras áreas geográficas. Além disso, o estudo não levou em conta fato- res externos, como mudanças climáticas ou eventos imprevistos, que podem inftuenciar a demanda de água.

Para pesquisas futuras, sugere-se abordar essas limitações e expandir o escopo do estudo. Uma proposta seria coletar dados de demanda de água de diferentes regiões e considerar variáveis climáticas e socioeconómicas para aprimorar a precisão das previsões. Além disso, seria interessante explorar técnicas de modelagem mais avançadas, como redes neurais artificiais ou métodos de aprendizado de máquina, a fim de melhorar ainda mais a precisão e eficiência das previsões.

Outra proposta futura seria investigar estratégias adicionais para o gerenciamento eficiente dos recursos hídricos, como a implementação de sistemas de reúso de água, a promoção de práticas de conservação e o desenvolvimento de fontes alternativas de abastecimento. Essas medidas podem contribuir para a sustentabilidade do abastecimento de água e reduzir a dependência de recursos naturais limitados.

Em resumo, esta dissertação proporcionou insights valiosos para a previsão da demanda de água e o gerenciamento eficiente do abastecimento hídrico. Apesar das limi- tações encontradas, as conclusões desta pesquisa fornecem uma base sólida para futuros estudos e aprimoramentos no campo da gestão dos recursos hídricos, visando garantir um abastecimento de água adequado, sustentável e resiliente às demandas futuras.

* 1. Limitações da Pesquisa e Propostas Futuras

Embora o estudo tenha obtido resultados significativos e fornecido insights valiosos sobre o tema abordado, algumas limitações podem ser identificadas. Uma das principais limitações dessa pesquisa reside na falta de exploração de modelos de rede neural LSTM (Long Short-Term Memory), CNN (Convolutional Neural Network) e RNN (Recurrent Neural Network), que têm sido amplamente utilizados em problemas de processamento de linguagem natural. Esses modelos possuem características específicas que podem melhorar o desempenho e a compreensão dos padrões presentes nos dados.

Outra limitação desse estudo está relacionada à otimização matemática dos algo- ritmos de aprendizado de máquina utilizados. Embora tenham sido empregadas técnicas comuns, como a busca em grade (Grid Search) e a validação cruzada (Cross Validation), existem métodos mais avançados que podem ser explorados no futuro. Sugere-se uma análise mais aprofundada de técnicas de otimização, como Optuna, Grid Search com va- lidação cruzada, busca aleatória (Randomized Search) e BayesSearchCV, para encontrar de forma mais eficiente os melhores hiperparâmetros dos modelos e melhorar ainda mais o desempenho preditivo.

Para estudos futuros, recomenda-se também investigar a inftuência de outros fa- tores e características nos modelos de aprendizado de máquina aplicados à detecção de fraudes em transações financeiras. Por exemplo, explorar o impacto de informações de- mográficas dos usuários, dados geográficos ou histórico de comportamento de transações anteriores. Além disso, uma análise mais aprofundada sobre técnicas de feature enginee- ring e seleção de variáveis pode ser realizada, visando identificar quais atributos são mais relevantes para a detecção de fraudes e, assim, melhorar a precisão dos modelos.

Em suma, embora este estudo tenha alcançado resultados promissores, é impor- tante reconhecer suas limitações e abrir caminho para pesquisas futuras que explorem modelos de rede neural mais avançados, técnicas de otimização matemática e fatores adicionais que podem aprimorar a detecção de fraudes em transações financeiras. Es- sas investigações têm o potencial de aprimorar ainda mais as estratégias de segurança e proteção de instituições financeiras, contribuindo para a mitigação de perdas e prejuízos causados por atividades fraudulentas.

# Referências

AHMAD, T. et al. A comprehensive overview on the data driven and large scale based approaches for forecasting of building energy demand: A review. ENERGY AND BUILDINGS, v. 165, p. 301-3'20, '2018. ISSN 0378-7788.

ANDERSON, .J.; VILLIAMS, S. Random forest regression for time series forecasting. Journal of Time Seriei, Analyi,ii,, v. 3'2, n. '2, p. '234-'256, '20'21.

BERGMEIR, C.; HYNDMAN, R.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statii,tici, and Data Analyi,ii,, v. 1'20, p. 70-83, '2018.

BOROO.JENI, K. et al. A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. Electric Power Syi,temi, Rei,earch, v. 14'2, p. 58-73, '2017.

BRANDAO, G. A. Sériei, Temporaii,: Parte 1. DEV Community, '20'20. Disponível em: [<https://dev.to/giselyalves13/series-temporais-parte-1-13l8>](https://dev.to/giselyalves13/series-temporais-parte-1-13l8).

BROVN, D.; LEE, .J. A gentle introduction to xgboost for applied machine learning. Machine Learning Journal, v. '25, n. 3, p. 345-367, '20'21.

BROVNLEE, .J. Machine learning mai,tery with Python: underi,tand your data, create accurate modeli,, and work projecti, end-to-end. S.l.l: Machine Learning Mastery, '2016.

BUYUKSAHIN, U.; ERTEKIN. Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition.

Neurocomputing, v. 361, p. 151-163, '2019.

Carvalho .Jr., .J. G.; Costa .Jr., C. T. Non-iterative procedure incorporated into the fuzzy identification on a hybrid method of functional randomization for time series forecasting models. Applied Soft Computing Journal, Elsevier Ltd, Postgraduate Program in Electrical Engineering, Federal University of Pará, Brazil, v. 80, p. '2'26-'24'2, '2019. ISSN 15684946 (ISSN). Disponível em:

[<https://www.scopus.com/inward/record.uri?eid-'2-s'2.0-850644416'2'2&doi-10.1016%](https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064441622&doi=10.1016%2Fj.asoc.2019.03.059&partnerID=40&md5=84d0bd291cc451de280dc9ed77524736) ['2Fj.asoc.'2019.03.059&partnerID-40&md5-84d0bd'291cc451de'280dc9ed775'24736>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-85064441622&doi=10.1016%2Fj.asoc.2019.03.059&partnerID=40&md5=84d0bd291cc451de280dc9ed77524736).

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: Proceedingi, of the 22nd acm i,igkdd international conference on knowledge dii,covery and data mining. S.l.: s.n.l, '2016. p. 785-794.

CHEN, Y. Y. et al. Applications of Recurrent Neural Networks in Environmental Factor Forecasting: A Review. NEURAL COMPUTATION, v. 30, n. 11, p. '2855-'2881, '2018. ISSN 0899-7667.

CHOU, .J.-S.; NGUYEN, T.-K. Forward Forecast of Stock Price Using Sliding-Vindow Metaheuristic-Optimized Machine-Learning Regression. IEEE Trani,actioni, on

Indui,trial Informatici,, v. 14, n. 7, p. 313'2-314'2, '2018.

CHOU, .J.-S.; TRAN, D.-S. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. Energy, v. 165, p. 709-7'26, '2018.

COELHO, I. et al. A GPU deep learning metaheuristic based model for time series forecasting. Applied Energy, v. '201, p. 41'2-418, '2017.

CRUZ-RAMÍREZ, M. et al. Enhancing convolutional neural networks for image

classification of agricultural products. Computeri, and Electronici, in Agriculture, v. 177, p. 105754, '20'20.

DU, S. et al. Multivariate time series forecasting via attention-based encoder-decoder framework. Neurocomputing, v. 388, p. '269-'279, '20'20.

GARCIA, M.; RODRIGUEZ, A. Time series forecasting with lightgbm. Journal of Machine Learning Rei,earch, v. 10, n. 4, p. 789-81'2, '20'23.

GOLYANDINA, N. Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing. WILEY

INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS, v. 1'2, n. 4, '20'20. ISSN 1939-0068.

GRAFF, M. et al. Time series forecasting with genetic programming. Natural Computing, v. 16, n. 1, p. 165-174, '2017.

HUO, Y. et al. Long-term span traffi c prediction model based on stl decomposition and lstm. In: . S.l.: s.n.l, '2019. p. 1-4.

.JOHNSON, R.; SMITH, M. Linear regression for predictive modeling. Journal of Predictive Analytici,, v. 18, n. 1, p. 56-78, '20'2'2.

.JONES, A. B.; SMITH, C. D.; .JOHNSON, E. F. Comparing forecasting models for solar power generation. Renewable Energy, Elsevier, v. 107, p. 45'2-461, '2017.

KHAN, Z. et al. A hybrid algorithm for solar radiation forecasting using machine learning and arima models. Journal of Cleaner Production, v. '297, p. 1'26603, '20'21.

KORSTAN.JE, .J. Advanced Forecai,ting with Python. S.l.l: Springer, '20'21.

KULSHRESHTHA, S.; VI.JAYALAKSHMI, A. An ARIMA-LSTM hybrid model for

stock market prediction using live data. Journal of Engineering Science and Technology Review, v. 13, n. 4, p. 117-1'23, '20'20.

KUMAR, G.; .JAIN, S.; SINGH, U. P. Stock Market Forecasting Using Computational Intelligence: A Survey. ARCHIVES OF COMPUTATIONAL METHODS IN ENGINEERING, v. '28, n. 3, p. 1069-1101, '20'21. ISSN 1134-3060.

LARA-BENITEZ, P.; CARRANZA-GARCIA, M.; RIQUELME, .J. C. An Experimental

Review on Deep Learning Architectures for Time Series Forecasting. INTER- NATIONAL JOURNAL OF NEURAL SYSTEMS, v. 31, n. 3, '20'21. ISSN 01'29-0657.

LI, A. V.; BASTOS, G. S. Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review. IEEE ACCESS, v. 8, p. 185'23'2-185'24'2, '20'20. ISSN '2169-3536.

LIU, H.; CHEN, C. Data processing strategies in wind energy forecasting models and applications: A comprehensive review. APPLIED ENERGY, v. '249, p. 39'2-408, '2019. ISSN 0306-'2619.

LIU, Z. Y. et al. Forecast Methods for Time Series Data: A Survey. IEEE ACCESS, v. 9, p. 91896-9191'2, '20'21. ISSN '2169-3536 .J9 - IEEE ACCESS .JI - IEEE Access.

MA, Y.; YU, L.; ZHANG, G. A hybrid short-term load forecasting model based on a multi- trait-driven methodology and secondary decomposition. Energiei,, v. 15, n. 16, '20'2'2.

ISSN 1996-1073. Disponível em: [<https://www.mdpi.com/1996-1073/15/16/5875>](https://www.mdpi.com/1996-1073/15/16/5875).

MARTINOVIé, M.; HUN.JET, A.; TURCIN, I. Time series forecasting of the austrian traded index (Atx) using artificial neural network model. Tehnicki Vjei,nik, v. '27, n. 6, p. '2053-'2061, '20'20.

MARTINS, L. E. G.; GORSCHEK, T. Requirements engineering for safety- critical systems: A systematic literature review. Information and Software Technology, v. 75, p. 71-89, '2016. ISSN 0950-5849. Disponível em: [<https:](https://www.sciencedirect.com/science/article/pii/S0950584916300568)

[//www.sciencedirect.com/science/article/pii/S0950584916300568>](https://www.sciencedirect.com/science/article/pii/S0950584916300568).

MOKHTARI, A. et al. Deep learning for early diagnosis of diabetes using ppg signals. IEEE Trani,actioni, on Ini,trumentation and Meai,urement, v. 69, n. 8, p.

5916-59'25, '20'20.

MOON, .J. et al. Temporal data classification and forecasting using a memristor-based reservoir computing system. Nature Electronici,, v. '2, n. 10, p. 480-487, '2019.

PELLETIER, C. et al. Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas. Remote

Seni,ing of Environment, v. 187, p. 156-168, '2016. Cited By '296. Disponível em:

[<https://www.scopus.com/inward/record.uri?eid-'2-s'2.0-8499'2151859&doi-10.1016%](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84992151859&doi=10.1016%2fj.rse.2016.10.010&partnerID=40&md5=09efc79bab8e893b97fd21cb4844b98d) ['2fj.rse.'2016.10.010&partnerID-40&md5-09efc79bab8e893b97fd'21cb4844b98d>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84992151859&doi=10.1016%2fj.rse.2016.10.010&partnerID=40&md5=09efc79bab8e893b97fd21cb4844b98d).

PENG, Z. et al. An effective method for inventory forecasting based on online machine learning. Indui,trial Management & Data Syi,temi,, Emerald Publishing Limited, v. 117, n. 4, p. 704-718, '2017.

PETROPOULOS, F. et al. Forecasting: theory and practice. International Journal of Forecai,ting, v. 38, n. 3, p. 705-871, '20'2'2. ISSN 0169-'2070. Disponível em:

[<https://www.sciencedirect.com/science/article/pii/S0169'2070'21001758>](https://www.sciencedirect.com/science/article/pii/S0169207021001758).

PINHEIRO, N. M. Introdução a Seriei, Temporaii, - Parte 1. Data Hackers, '20'2'2. Disponível em: [<https://medium.com/data-hackers/](https://medium.com/data-hackers/series-temporais-parte-1-a0e75a512e72) [series-temporais-parte-1-a0e75a51'2e7'2>](https://medium.com/data-hackers/series-temporais-parte-1-a0e75a512e72).

REISEN, V. et al. Robust dickey-fuller tests based on ranks for time series with additive outliers. Metrika, v. 80, n. 1, p. 115-131, '2017. Cited By 1. Disponível em:

[<https://www.scopus.com/inward/record.uri?eid-'2-s'2.0-849863173'25&doi-10.1007%](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84986317325&doi=10.1007%2fs00184-016-0594-8&partnerID=40&md5=c83f82d0c372e22d5970aff448f05411) ['2fs00184-016-0594-8&partnerID-40&md5-c83f8'2d0c37'2e'2'2d5970aff448f05411>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84986317325&doi=10.1007%2fs00184-016-0594-8&partnerID=40&md5=c83f82d0c372e22d5970aff448f05411).

RIBEIRO, M. H. D. M. et al. Time series forecasting based on ensemble learning methods applied to agribusiness, epidemiology, energy demand, and renewable energy. Pontifícia Universidade Católica do Paraná, '20'21.

ROSSI, R. Relational time series forecasting. Knowledge Engineering Review, v. 33, '2018.

SADAEI, H. et al. Short-term load forecasting by using a combined method of

convolutional neural networks and fuzzy time series. Energy, v. 175, p. 365-377, '2019.

SALGOTRA, R.; GANDOMI, M.; GANDOMI, A. Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming. Chaoi,, Solitoni, and Fractali,, v. 138, '20'20.

SAMANTA, S. et al. Learning elastic memory online for fast time series forecasting. Neurocomputing, v. 390, p. 315-3'26, '20'20.

SANCHEZ, A. M.; DÍAZ, A. A.; LÓPEZ, A. O. A comparative study of xgboost, adaboost, and catboost in machine learning algorithms. In: SPRINGER. International Conference on Learning and Optimization Algorithmi,: Theory and Applicationi, (LOPAL). S.l.l, '20'20. p. '29'2-303.

SEZER, O. B.; GUDELEK, M. U.; OZBAYOGLU, A. M. Financial time series

forecasting with deep learning : A systematic literature review: '2005-'2019. APPLIED SOFT COMPUTING, v. 90, '20'20. ISSN 1568-4946.

SHARMA, S. et al. A deep learning framework for road traffi c anomaly detection and classification using traffi c surveillance cameras. IEEE Trani,actioni, on Intelligent Trani,portation Syi,temi,, '20'21.

SHEN, Y. et al. An ensemble model based on deep learning and data preprocessing for short-term electrical load forecasting. Sui,tainability, v. 13, n. 4, '20'21. ISSN '2071-1050. Disponível em: [<https://www.mdpi.com/'2071-1050/13/4/1694>](https://www.mdpi.com/2071-1050/13/4/1694).

SHEN, Z. et al. A novel time series forecasting model with deep learning. Neurocomputing, v. 396, p. 30'2-313, '20'20.

SHIH, S.-Y.; SUN, F.-K.; LEE, H.-Y. Temporal pattern attention for multivariate time series forecasting. Machine Learning, v. 108, n. 8-9, p. 14'21-1441, '2019.

SMITH, .J.; .JOHNSON, E. Time series forecasting with arima in python. Journal of Data Science, v. 15, n. '2, p. 1'23-145, '20'2'2.

SOYER, R.; ZHANG, D. Bayesian modeling of multivariate time series of

counts. WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL STATISTICS. ISSN 1939-0068.

TAIEB, S. B.; ATIYA, A. F. A Bias and Variance Analysis for Multistep- Ahead Time Series Forecasting. IEEE Trani,actioni, on Neural Networki, and Learning Syi,temi,, Institute of Electrical and Electronics Engineers Inc., Department of Computer Science, Université Libre de Bruxelles, Brussels, 1050, Belgium, v. '27, n. 1, p. 6'2-76, '2016. ISSN '216'2'237X (ISSN). Disponível em:

[<https://www.scopus.com/inward/record.uri?eid-'2-s'2.0-849'25431469&doi-10.1109%](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84925431469&doi=10.1109%2FTNNLS.2015.2411629&partnerID=40&md5=e1c7f3c7a1136a0e0e4d2aff817b4008) ['2FTNNLS.'2015.'24116'29&partnerID-40&md5-e1c7f3c7a1136a0e0e4d'2aff817b4008>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84925431469&doi=10.1109%2FTNNLS.2015.2411629&partnerID=40&md5=e1c7f3c7a1136a0e0e4d2aff817b4008).

TAN, Y. F. et al. Exploring Time-Series Forecasting Models for Dynamic Pricing in Digital Signage Advertising. FUTURE INTERNET, v. 13, n. 10, '20'21. ISSN 1999-5903.

THEODOSIOU, M. Forecasting monthly and quarterly time series using stl decomposition. International Journal of Forecai,ting, v. '27, n. 4, p. 1178-1195, '2011. Cited By 86. Disponível em: [<https://www.scopus.com/inward/record.uri?eid-](https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052160927&doi=10.1016%2fj.ijforecast.2010.11.002&partnerID=40&md5=e8242471ba1ec14ada46ab567f3a364d) ['2-s'2.0-8005'21609'27&doi-10.1016%'2fj.ijforecast.'2010.11.00'2&partnerID-40&md5-](https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052160927&doi=10.1016%2fj.ijforecast.2010.11.002&partnerID=40&md5=e8242471ba1ec14ada46ab567f3a364d)

[e8'24'2471ba1ec14ada46ab567f3a364d>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052160927&doi=10.1016%2fj.ijforecast.2010.11.002&partnerID=40&md5=e8242471ba1ec14ada46ab567f3a364d).

TRENBERTH, K. E. Signal versus noise in the southern oscillation. Monthly Weather Review, v. 11'2, n. '2, p. 3'26-33'2, 1984.

TYRALIS, H.; PAPACHARALAMPOUS, G. Variable selection in time series forecasting using random forests. Algorithmi,, v. 10, n. 4, '2017.

URSU, E.; PEREAU, .J. C. Application of periodic autoregressive process to the modeling of the Garonne river ftows. STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT, v. 30, n. 7, p. 1785-1795, '2016. ISSN 1436-3'240.

VASCONCELOS, F. Falta d'água em curitiba e região metropolitana não É

culpa i,ó da ei,tiagem. '20'20. Disponível em: [<https://www.brasildefato.com.br/'20'20/](https://www.brasildefato.com.br/2020/11/03/falta-d-agua-em-curitiba-e-regiao-metropolitana-nao-e-culpa-so-da-estiagem) [11/03/falta-d-agua-em-curitiba-e-regiao-metropolitana-nao-e-culpa-so-da-estiagem>](https://www.brasildefato.com.br/2020/11/03/falta-d-agua-em-curitiba-e-regiao-metropolitana-nao-e-culpa-so-da-estiagem).

VLACHAS, P. et al. Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics. Neural Networki,, v. 1'26, p. 191-'217, '20'20.

VANG, Y. et al. Recycling combustion ash for sustainable cement production: A critical review with data-mining and time-series predictive models. CONSTRUCTION AND BUILDING MATERIALS, v. 1'23, p. 673-689, '2016. ISSN 0950-0618.

VILLMOTT, C. .J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. Climate Rei,earch, Inter-Research, v. 30, n. 1, p. 79-8'2, '2005.

XIE, T. et al. Hybrid forecasting model for non-stationary daily runoff series: A case study in the Han River Basin, China. JOURNAL OF HYDROLOGY, v. 577, '2019. ISSN 00'2'2-1694.

XU, V. et al. Deep belief network-based AR model for nonlinear time series forecasting. Applied Soft Computing Journal, v. 77, p. 605-6'21, '2019.

YANG, V. et al. Hybrid wind energy forecasting and analysis system based on divide and conquer scheme: A case study in China. Journal of Cleaner Production, v. '2'2'2, p. 94'2-959, '2019.

YU, C. Research of time series air quality data based on exploratory data analysis and representation. In: . Institute of Electrical and Electronics Engineers Inc., '2016. ISBN 97815090'23509. Cited By 5; Conference of 5th International Conference on Agro-Geoinformatics, Agro-Geoinformatics '2016 ; Conference Date: 18 .July '2016 Through '20 .July '2016; Conference Code:1'24077. Disponível em: [<https://www.scopus.](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994079422&doi=10.1109%2fAgro-Geoinformatics.2016.7577697&partnerID=40&md5=fef861624a35632bf2d84acf63986bbe)

[com/inward/record.uri?eid-'2-s'2.0-849940794'2'2&doi-10.1109%'2fAgro-Geoinformatics.](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994079422&doi=10.1109%2fAgro-Geoinformatics.2016.7577697&partnerID=40&md5=fef861624a35632bf2d84acf63986bbe) ['2016.7577697&partnerID-40&md5-fef8616'24a3563'2bf'2d84acf63986bbe>](https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994079422&doi=10.1109%2fAgro-Geoinformatics.2016.7577697&partnerID=40&md5=fef861624a35632bf2d84acf63986bbe).

ZHANG, H.; XU, .J.; SHEN, .J. Evaluation and comparison of forecasting performance of three typical crop models for winter wheat in the north china plain. Agricultural and Forei,t Meteorology, Elsevier, v. '2'28-'2'29, p. '276-'286, '2016.

# Apêndice - Comparação dos modelos de previsão de series temporais média de 24h

(*p* = 7*, d* = 1*, q* = 7)(*P* = 2*, D* = 1*, Q* = 1)*M*=12 Média '24h

Tabela 7: Comparação dos modelos com 1 dia de antecedência '24h Treinamento

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,096 | 0,306 | 0,419 |
| ARX | 0,118 | 0,377 | 0,513 |
| MA | 0,093 | 0,'296 | 0,403 |
| ARMA | 0,10'2 | 0,3'25 | 0,435 |
| ARIMA | 0,095 | 0,30'2 | 0,405 |
| SARIMA | 0,105 | 0,34'2 | 0,450 |
| ARIMAX | 0,119 | 0,378 | 0,511 |
| SARIMAX | 0,118 | 0,377 | 0,51'2 |
| LR | 0,015 | 0,069 | 0,077 |
| RFR | 0,190 | 0,6'24 | 0,67'2 |
| XGBRegrei,i,or | 0,'207 | 0,683 | 0,7'20 |
| LGBMRegrei,i,or | 0,184 | 0,599 | 0,655 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 8: Comparação dos modelos com 1 dia de antecedência '24h Validação

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  |  |
|  | MAPE | MAE | RMSE |
| AR | 0,084 | 0,'285 | 0,366 |
| ARX | 0,103 | 0,354 | 0,459 |
| MA | 0,08'2 | 0,'278 | 0,361 |
| ARMA | 0,086 | 0,'295 | 0,37'2 |
| ARIMA | 0,08'2 | 0,'280 | 0,351 |
| SARIMA | 0,097 | 0,333 | 0,4'21 |
| ARIMAX | 0,10'2 | 0,353 | 0,458 |
| SARIMAX | 0,104 | 0,358 | 0,463 |
| LR | 0,014 | 0,066 | 0,073 |
| RFR | 0,17'2 | 0,587 | 0,633 |
| XGBRegrei,i,or | 0,19'2 | 0,658 | 0,69'2 |
| LGBMRegrei,i,or | 0,166 | 0,564 | 0,616 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 9: Comparação dos modelos com 1 dia de antecedência '24h Tei,te

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,100 | 0,3'29 | 0,4'24 |
| ARX | 0,137 | 0,46'2 | 0,586 |
| MA | 0,10'2 | 0,336 | 0,431 |
| ARMA | 0,10'2 | 0,340 | 0,433 |
| ARIMA | 0,103 | 0,346 | 0,440 |
| SARIMA | 0,118 | 0,398 | 0,501 |
| ARIMAX | 0,137 | 0,461 | 0,587 |
| SARIMAX | 0,138 | 0,464 | 0,590 |
| LR | 0,018 | 0,087 | 0,098 |
| RFR | 0,153 | 0,494 | 0,587 |
| XGBRegrei,i,or | 0,170 | 0,560 | 0,643 |
| LGBMRegrei,i,or | 0,145 | 0,465 | 0,568 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,093 | 0,30'2 | 0,165 |
| ARX | 0,1'23 | 0,40'2 | 0,'283 |
| MA | 0,107 | 0,344 | 0,460 |
| ARMA | 0,097 | 0,316 | 0,4'24 |
| ARIMA | 0,094 | 0,303 | 0,406 |
| SARIMA | 0,106 | 0,350 | 0,448 |
| ARIMAX | 0,1'20 | 0,394 | 0,5'21 |
| SARIMAX | 0,1'2'2 | 0,401 | 0,530 |
| LR | 0,016 | 0,074 | 0,084 |
| RFR | 0,176 | 0,579 | 0,64'2 |
| XGBRegrei,i,or | 0,194 | 0,643 | 0,694 |
| LGBMRegrei,i,or | 0,170 | 0,554 | 0,6'24 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 11: Comparação dos modelos com 7 dias de antecedência '24h Treinamento

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,093 | 0,'296 | 0,399 |
| ARX | 0,118 | 0,377 | 0,5'24 |
| MA | 0,104 | 0,3'29 | 0,444 |
| ARMA | 0,103 | 0,330 | 0,439 |
| ARIMA | 0,108 | 0,34'2 | 0,463 |
| SARIMA | 0,111 | 0,360 | 0,487 |
| ARIMAX | 0,118 | 0,379 | 0,5'25 |
| SARIMAX | 0,118 | 0,379 | 0,5'25 |
| LR | 1,197 | 5,'230 | 5,'230 |
| RFR | 0,'2'24 | 0,705 | 0,8'21 |
| XGBRegrei,i,or | 0,'260 | 0,8'23 | 0,934 |
| LGBMRegrei,i,or | 0,215 | 0,673 | 0,793 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,073 | 0,'245 | 0,319 |
| ARX | 0,093 | 0,319 | 0,4'23 |
| MA | 0,080 | 0,'269 | 0,353 |
| ARMA | 0,081 | 0,'274 | 0,347 |
| ARIMA | 0,087 | 0,'29'2 | 0,384 |
| SARIMA | 0,095 | 0,3'24 | 0,438 |
| ARIMAX | 0,093 | 0,318 | 0,4'2'2 |
| SARIMAX | 0,094 | 0,3'20 | 0,4'24 |
| LR | 1,174 | 5,'2'24 | 5,'2'24 |
| RFR | 0,188 | 0,630 | 0,71'2 |
| XGBRegrei,i,or | 0,'2'23 | 0,756 | 0,8'28 |
| LGBMRegrei,i,or | 0,179 | 0,598 | 0,684 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 13: Comparação dos modelos com 7 dias de antecedência '24h Tei,te

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,118 | 0,383 | 0,499 |
| ARX | 0,147 | 0,479 | 0,63'2 |
| MA | 0,1'25 | 0,403 | 0,530 |
| ARMA | 0,117 | 0,384 | 0,494 |
| ARIMA | 0,1'20 | 0,393 | 0,505 |
| SARIMA | 0,131 | 0,437 | 0,544 |
| ARIMAX | 0,148 | 0,480 | 0,63'2 |
| SARIMAX | 0,148 | 0,481 | 0,636 |
| LR | 1,161 | 5,'21'2 | 5,'213 |
| RFR | 0,187 | 0,578 | 0,755 |
| XGBRegrei,i,or | 0,'2'2'2 | 0,693 | 0,870 |
| LGBMRegrei,i,or | 0,177 | 0,543 | 0,7'27 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,098 | 0,316 | 0,177 |
| ARX | 0,1'25 | 0,406 | 0,305 |
| MA | 0,105 | 0,337 | 0,450 |
| ARMA | 0,097 | 0,312 | 0,418 |
| ARIMA | 0,097 | 0,314 | 0,4'20 |
| SARIMA | 0,118 | 0,386 | 0,506 |
| ARIMAX | 0,1'24 | 0,40'2 | 0,546 |
| SARIMAX | 0,1'25 | 0,405 | 0,551 |
| LR | 1,183 | 5,'2'24 | 5,'2'24 |
| RFR | 0,'208 | 0,656 | 0,787 |
| XGBRegrei,i,or | 0,'243 | 0,775 | 0,901 |
| LGBMRegrei,i,or | 0,199 | 0,6'23 | 0,759 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 15: Comparação dos modelos com 14 dias de antecedência '24h Treinamento

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,105 | 0,334 | 0,445 |
| ARX | 0,1'26 | 0,399 | 0,548 |
| MA | 0,106 | 0,336 | 0,447 |
| ARMA | 0,110 | 0,350 | 0,463 |
| ARIMA | 0,111 | 0,353 | 0,477 |
| SARIMA | 0,114 | 0,367 | 0,489 |
| ARIMAX | 0,1'26 | 0,401 | 0,547 |
| SARIMAX | 0,1'26 | 0,401 | 0,547 |
| LR | '2,606 | 11,394 | 11,394 |
| RFR | 0,'2'21 | 0,696 | 0,81'2 |
| XGBRegrei,i,or | 0,'269 | 0,859 | 0,96'2 |
| LGBMRegrei,i,or | 0,215 | 0,673 | 0,79'2 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,078 | 0,'264 | 0,346 |
| ARX | 0,090 | 0,309 | 0,430 |
| MA | 0,079 | 0,'265 | 0,349 |
| ARMA | 0,093 | 0,317 | 0,403 |
| ARIMA | 0,088 | 0,'295 | 0,389 |
| SARIMA | 0,09'2 | 0,315 | 0,40'2 |
| ARIMAX | 0,090 | 0,308 | 0,4'29 |
| SARIMAX | 0,090 | 0,308 | 0,4'29 |
| LR | '2,558 | 11,388 | 11,388 |
| RFR | 0,185 | 0,619 | 0,70'2 |
| XGBRegrei,i,or | 0,'233 | 0,790 | 0,859 |
| LGBMRegrei,i,or | 0,179 | 0,598 | 0,683 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 17: Comparação dos modelos com 14 dias de antecedência '24h Tei,te

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,118 | 0,378 | 0,504 |
| ARX | 0,1'20 | 0,384 | 0,560 |
| MA | 0,1'20 | 0,385 | 0,509 |
| ARMA | 0,107 | 0,344 | 0,464 |
| ARIMA | 0,105 | 0,338 | 0,46'2 |
| SARIMA | 0,113 | 0,364 | 0,496 |
| ARIMAX | 0,1'20 | 0,384 | 0,560 |
| SARIMAX | 0,119 | 0,383 | 0,558 |
| LR | '2,531 | 11,376 | 11,377 |
| RFR | 0,186 | 0,57'2 | 0,748 |
| XGBRegrei,i,or | 0,'2'27 | 0,710 | 0,889 |
| LGBMRegrei,i,or | 0,177 | 0,54'2 | 0,7'25 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,104 | 0,335 | 0,'204 |
| ARX | 0,1'2'2 | 0,393 | 0,'297 |
| MA | 0,106 | 0,340 | 0,45'2 |
| ARMA | 0,097 | 0,311 | 0,4'23 |
| ARIMA | 0,099 | 0,318 | 0,431 |
| SARIMA | 0,113 | 0,365 | 0,49'2 |
| ARIMAX | 0,1'21 | 0,389 | 0,539 |
| SARIMAX | 0,1'2'2 | 0,393 | 0,543 |
| LR | '2,577 | 11,388 | 11,388 |
| RFR | 0,'206 | 0,648 | 0,779 |
| XGBRegrei,i,or | 0,'251 | 0,804 | 0,9'27 |
| LGBMRegrei,i,or | 0,198 | 0,6'23 | 0,758 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela 19: Comparação dos modelos com 30 dias de antecedência '24h Treinamento

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,1'21 | 0,383 | 0,514 |
| ARX | 0,135 | 0,43'2 | 0,59'2 |
| MA | 0,120 | 0,379 | 0,510 |
| ARMA | 0,120 | 0,383 | 0,508 |
| ARIMA | 0,1'24 | 0,395 | 0,5'27 |
| SARIMA | 0,1'26 | 0,405 | 0,538 |
| ARIMAX | 0,136 | 0,434 | 0,594 |
| SARIMAX | 0,136 | 0,435 | 0,596 |
| LR | 5,8'27 | '25,483 | '25,484 |
| RFR | 0,'2'24 | 0,705 | 0,8'21 |
| XGBRegrei,i,or | 0,'28'2 | 0,90'2 | 0,998 |
| LGBMRegrei,i,or | 0,211 | 0,659 | 0,780 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

|  |  |  |  |
| --- | --- | --- | --- |
| AR | 0,091 | 0,311 | 0,390 |
| ARX | 0,086 | 0,30'2 | 0,434 |
| MA | 0,090 | 0,306 | 0,383 |
| ARMA | 0,089 | 0,304 | 0,384 |
| ARIMA | 0,100 | 0,343 | 0,4'26 |
| SARIMA | 0,098 | 0,337 | 0,41'2 |
| ARIMAX | 0,086 | 0,301 | 0,433 |
| SARIMAX | 0,086 | 0,30'2 | 0,434 |
| LR | 5,7'21 | '25,478 | '25,478 |
| RFR | 0,187 | 0,6'28 | 0,710 |
| XGBRegrei,i,or | 0,'245 | 0,831 | 0,896 |
| LGBMRegrei,i,or | 0,174 | 0,580 | 0,666 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela '21: Comparação dos modelos com 30 dias de antecedência '24h Tei,te

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  | |
|  | MAPE | MAE | RMSE |
| AR | 0,117 | 0,375 | 0,495 |
| ARX | 0,141 | 0,46'2 | 0,6'28 |
| MA | 0,1'20 | 0,384 | 0,504 |
| ARMA | 0,118 | 0,384 | 0,496 |
| ARIMA | 0,1'20 | 0,390 | 0,509 |
| SARIMA | 0,13'2 | 0,431 | 0,570 |
| ARIMAX | 0,140 | 0,459 | 0,6'27 |
| SARIMAX | 0,14'2 | 0,463 | 0,6'27 |
| LR | 5,663 | '25,466 | '25,466 |
| RFR | 0,189 | 0,583 | 0,759 |
| XGBRegrei,i,or | 0,'239 | 0,754 | 0,918 |
| LGBMRegrei,i,or | 0,174 | 0,53'2 | 0,716 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela '2'2: Comparação dos modelos com 30 dias de antecedência '24h Completo

Erroi,

|  |  |  |  |
| --- | --- | --- | --- |
| Modeloi, |  |  |  |
|  | MAPE | MAE | RMSE |
| AR | 0,114 | 0,367 | 0,'237 |
| ARX | 0,137 | 0,447 | 0,360 |
| MA | 0,113 | 0,361 | 0,477 |
| ARMA | 0,1'20 | 0,385 | 0,508 |
| ARIMA | 0,117 | 0,375 | 0,497 |
| SARIMA | 0,1'24 | 0,404 | 0,531 |
| ARIMAX | 0,136 | 0,443 | 0,596 |
| SARIMAX | 0,137 | 0,446 | 0,601 |
| LR | 5,763 | '25,477 | '25,477 |
| RFR | 0,'208 | 0,657 | 0,788 |
| XGBRegrei,i,or | 0,'264 | 0,847 | 0,961 |
| LGBMRegrei,i,or | 0,195 | 0,610 | 0,746 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

# Apêndice - Comparação dos modelos de previsão com o método Ljung-Box

Modelo ARIMA com defasagem de 10 para previsão de longo prazo na demanda de água.

Tabela '23: Comparação dos modelos Ljung Box Treinamento

|  |  |  |
| --- | --- | --- |
| Ljung Box | Ei,tatíi,tica de Tei,te | Valor De p |
| ARX | 6,30 | 0,79 |
| AR | 7,13 | 0,07 |
| MA | 34,34 | 0,00 |
| ARMA | 11,60 | 0,31 |
| ARIMA | 13,01 | 0,'2'2 |
| SARIMA | 10,17 | 0,43 |
| ARIMAX | 30,36 | 0,00 |
| SARIMAX | 11,63 | 0,31 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Tabela '24: Comparação dos modelos Ljung Box Validação

|  |  |  |
| --- | --- | --- |
| Ljung Box | Ei,tatíi,tica de Tei,te | Valor De p |
| ARX | 7,47 | 0,68 |
| AR | '2,43 | 0,99 |
| MA | 1,39 | 1,00 |
| ARMA | 5,4'2 | 0,86 |
| ARIMA | 4,04 | 0,95 |
| SARIMA | 4,45 | 0,93 |
| ARIMAX | 0,02 | 1,00 |
| SARIMAX | 0,04 | 1,00 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20) Tabela '25: Comparação dos modelos Ljung Box Tei,te

|  |  |  |
| --- | --- | --- |
| Ljung Box | Ei,tatíi,tica de Tei,te | Valor De p |
| ARX | 0,86 | 1,00 |
| AR | 7,80 | 0,65 |
| MA | 7,89 | 0,64 |
| ARMA | 19,34 | 0,04 |
| ARIMA | 9,50 | 0,49 |
| SARIMA | 3,57 | 0,97 |
| ARIMAX | 0,60 | 1,00 |
| SARIMAX | 3,7'2 | 0,96 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20) Tabela '26: Comparação dos modelos Ljung Box Completo

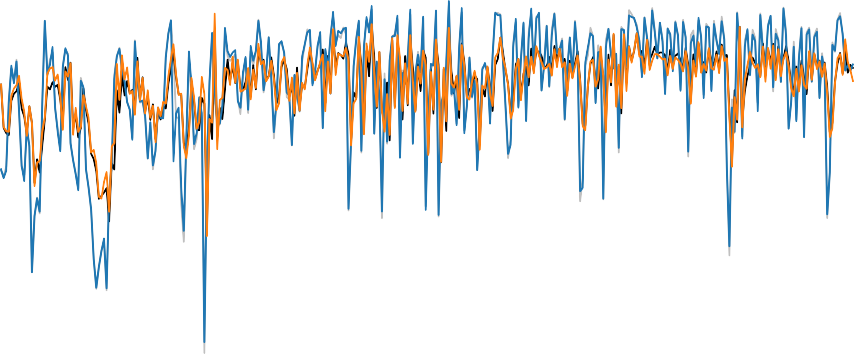
|  |  |  |
| --- | --- | --- |
| Ljung Box | Ei,tatíi,tica de Tei,te | Valor De p |
| ARX | 4,70 | 0,91 |
| AR | 4,26 | 0,16 |
| MA | 49,16 | 0,00 |
| ARMA | 40,49 | 0,00 |
| ARIMA | 40,49 | 0,00 |
| SARIMA | 40,49 | 0,00 |
| ARIMAX | 60,91 | 0,00 |
| SARIMAX | 5,83 | 0,83 |

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

# Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h

Figura 4'2: Comparação dos modelos AR, ARX e MA, 1 dia à frente

*4.0*



*Observado*

*AR (7) ARX (7) MA (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

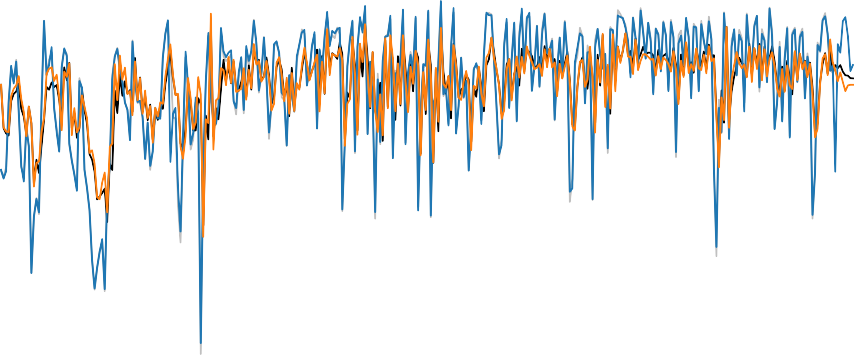
*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*1 dia à frente AR, ARX e MA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 43: Comparação dos modelos AR, ARX e MA, 7 dias à frente

*4.0*



*Observado*

*AR (7) ARX (7) MA (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*7 dias à frente AR, ARX e MA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 44: Comparação dos modelos AR, ARX e MA, 14 dias à frente

*4.5*

*Observado*

*AR (7) ARX (7) MA (7)*

*4.0*

*3.5*

*LT01 (m³)*

*3.0*

*2.5*

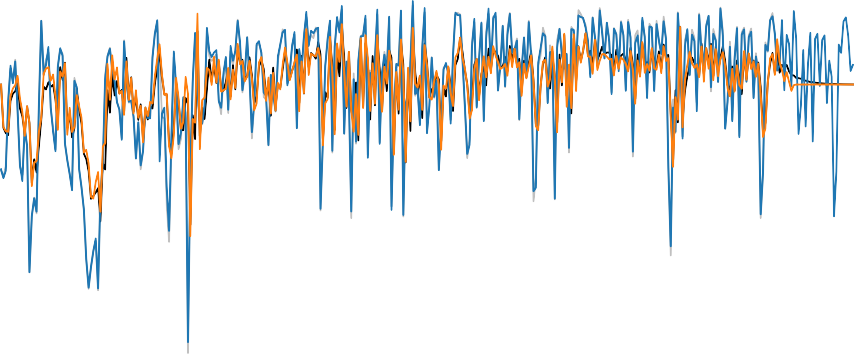
*2020-09-15 2020-10-012020-10-15 2020-11-012020-11-15 2020-12-012020-12-15 2021-01-012021-01-15*

*14 dias a frente AR, ARX e MA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 45: Comparação dos modelos AR, ARX e MA, 30 dias à frente

*4.0*



*Observado*

*AR (7) ARX (7) MA (7)*

*3.5*

*3.0*

*LT01 (m³)*

*2.5*

*2.0*

*1.5*

*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

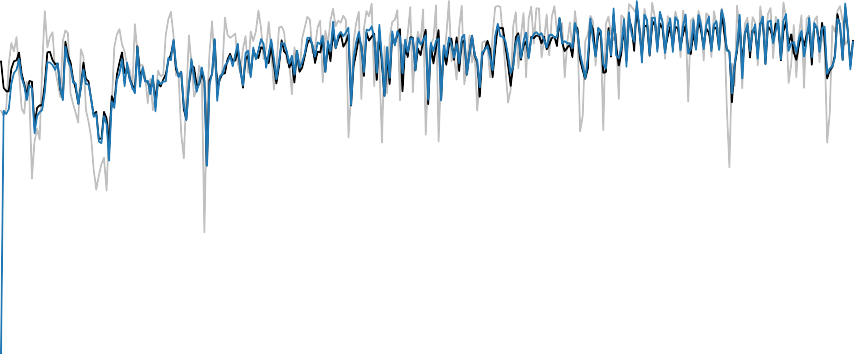
*30 dias à frente AR, ARX e MA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

# Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h

Figura 46: Comparação dos modelos ARMA e ARIMA, 1 dia à frente

*4*



*Observado*

*ARMA (7, 7) ARIMA (7, 1, 7)*

*3*

*2*

*LT01 (m³)*

*1*

*0*

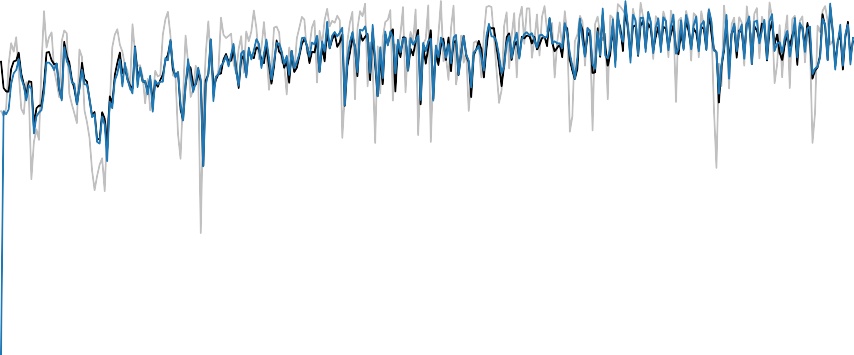
*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*1 dia à frente ARMA e ARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 47: Comparação dos modelos ARMA e ARIMA, 7 dias à frente

*4*



*Observado*

*ARMA (7, 7) ARIMA (7, 1, 7)*

*3*

*2*

*LT01 (m³)*

*1*

*0*

*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*7 dias à frente ARMA e ARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 48: Comparação dos modelos ARMA e ARIMA, 14 dias à frente

*4*



*Observado*

*ARMA (7, 7) ARIMA (7, 1, 7)*

*3*

*2*

*LT01 (m³)*

*1*

*0*

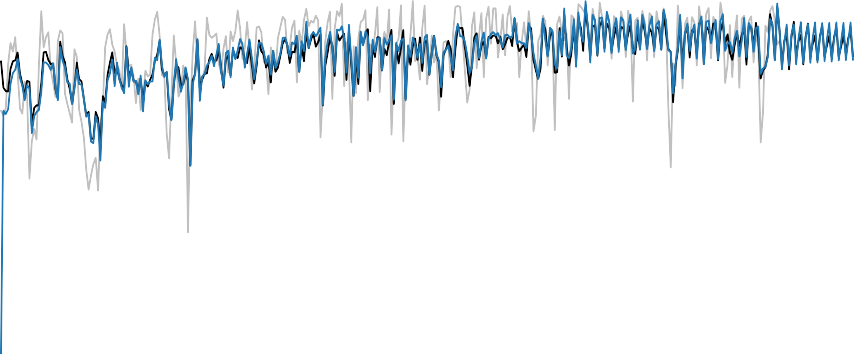
*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*14 dias à frente ARMA e ARIMA*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 49: Comparação dos modelos ARMA e ARIMA, 30 dias à frente

*4*



*Observado*

*ARMA (7, 7) ARIMA (7, 1, 7)*

*3*

*2*

*LT01 (m³)*

*1*

*0*

*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*30 dias à frente ARMA e ARIMA*

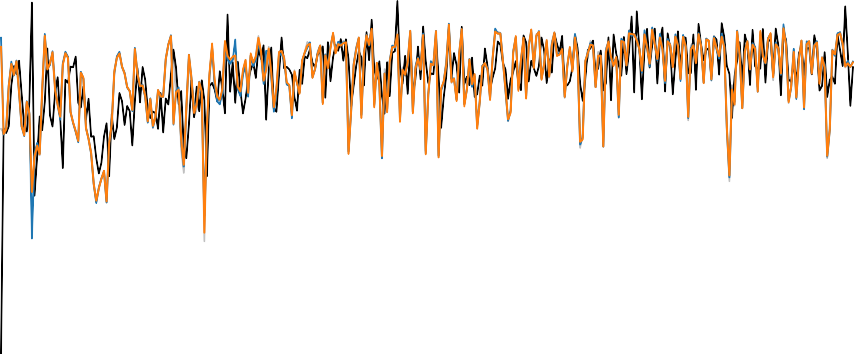
Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

# Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e SARIMAX (7,1,7) (2,1,0,12)

24h

Figura 50: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente

*4*



*Observado ARIMAX(7, 1, 7)*

*SARIMA (7, 1, 7) (2, 1, 0, 12) SARIMAX (7, 1, 7) (2, 1, 0, 12)*

*3*

*LT01 (m³)*

*2*

*1*

*0*

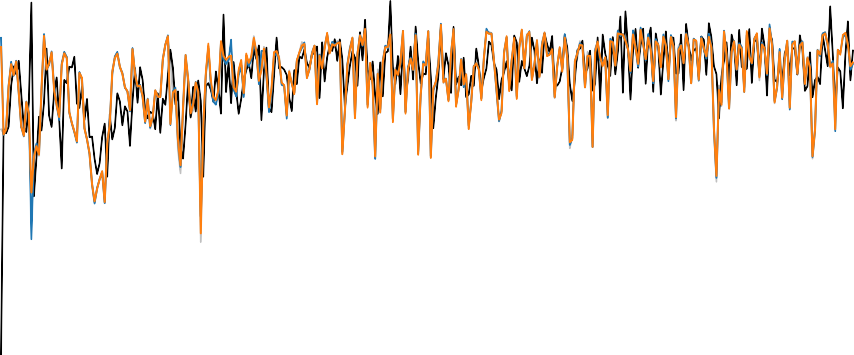
*2020-02 2020-03 2020-04 2020-05 2020-06 2020-07 2020-08 2020-09 2020-10 2020-11 2020-12 2021-01*

*1 dia à frente ARIMAX, SARIMA e SARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 51: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente

*4*



*Observado ARIMAX(7, 1, 7)*

*SARIMA (7, 1, 7) (2, 1, 0, 12) SARIMAX (7, 1, 7) (2, 1, 0, 12)*

*3*

*LT01 (m³)*

*2*

*1*

*0*

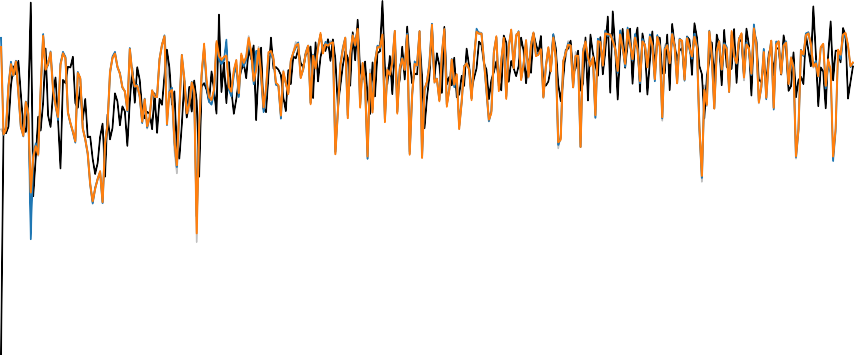
*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*7 dias à frente ARIMAX, SARIMA e SARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 5'2: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à frente

*4*



*Observado ARIMAX(7, 1, 7)*

*SARIMA (7, 1, 7) (2, 1, 0, 12) SARIMAX (7, 1, 7) (2, 1, 0, 12)*

*3*

*LT01 (m³)*

*2*

*1*

*0*

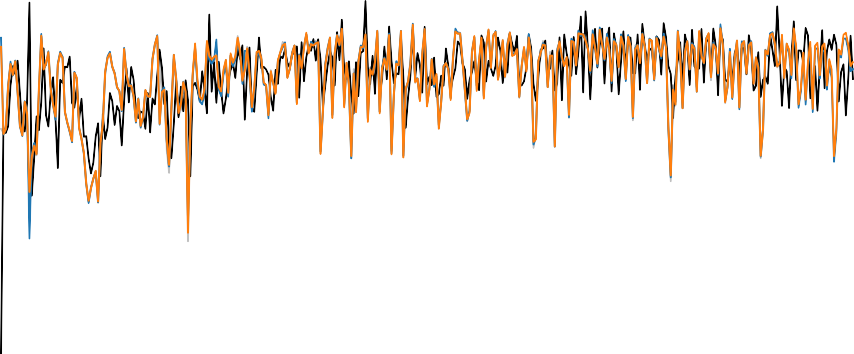
*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*14 dias à frente ARIMAX, SARIMA e SARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)

Figura 53: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à frente

*4*



*Observado ARIMAX(7, 1, 7)*

*SARIMA (7, 1, 7) (2, 1, 0, 12) SARIMAX (7, 1, 7) (2, 1, 0, 12)*

*3*

*LT01 (m³)*

*2*

*1*

*0*

*2020-03 2020-05 2020-07 2020-09 2020-11 2021-01*

*30 dias à frente ARIMAX, SARIMA e SARIMAX*

Fonte: Elaboração própria a partir de dados da SANEPAR ('2018 a '20'20)