



PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ  
ESCOLA POLITÉCNICA  
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO E  
SISTEMAS (PPGEPS)

**FRANCHESCO SANCHES DOS SANTOS**

PREVISÃO DE SÉRIES TEMPORAIS PARA A DEMANDA DE ÁGUA

CURITIBA  
2023

**FRANCHESCO SANCHES DOS SANTOS**

**PREVISÃO DE SÉRIES TEMPORAIS PARA A DEMANDA DE ÁGUA**

Projeto de Pesquisa de Mestrado apresentado ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS). Área de concentração: Automação e Controle 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 Sistemas.

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

CURITIBA  
2023

**FRANCHESCO SANCHES DOS SANTOS**

**PREVISÃO DE SÉRIES TEMPORAIS PARA A DEMANDA DE ÁGUA**

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas (PPGEPS). Área de concentração: Gerência de Produção e Logística, da Escola Politécnica, da Pontifícia Universidade 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, 8 de maio de 2023

Com gratidão, dedico este trabalho a Deus.  
Devo a ele tudo o que sou.

## **Agradecimentos**

Em primeiro lugar, agradeço a Deus por tudo o que ele tem a oferecer, pois ele abriu o caminho para mim e me deu forças para superar esse desafio, sem ele nada seria possível.

À minha família, que sempre me apoiou e incentivou a seguir em frente com a cabeça erguida e buscar um status mais elevado.

Ao professor Leandro dos Santos Coelho, agradeço por ter me dado a oportunidade de trabalhar com ele e compartilhar seus conhecimentos e experiências ao longo do mestrado, sempre em busca do meu crescimento profissional e pessoal que tornou este trabalho possível.

À Professora Viviana Cocco Mariani, obrigado pela disponibilidade e paciência em me ajudar com minhas deficiências e por usar seu conhecimento para contribuir com o desenvolvimento da pesquisa.

Agradeço à equipe da Pontifícia Universidade Católica do Paraná (PUCPR) e aos demais professores, em especial à secretária Denise da Mata Medeiros (PPGEPS), por me atenderem com paciência e carinho e me ajudarem inúmeras vezes, e nem mesmo medirem o esforço despendido.

Aos meus amigos que torceram por mim, bem como os novos amigos que fiz nessa trajetória, que proporcionaram grandes momentos de alegria na batalha.

Graças ao investimento em bolsas de estudo concedidas pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), essa etapa da minha carreira profissional e acadêmica foi concluída.

*“As leis da natureza não são, senão, os  
pensamentos matemáticos de Deus”.*  
(Euclides)

## Resumo

**CONTEXTO;** Diante de um cenário competitivo, a previsão assertiva da demanda tem se tornado cada vez mais uma ferramenta estratégica para vários ramos organizacionais. **PROBLEMA; SOLUÇÃO; ESTADO-DA-ARTE; MÉTODO(S)/PRODUTO(S) PROPOSTO(S); RESULTADOS.**

Diante de um cenário competitivo, a previsão assertiva da demanda tem se tornado cada vez mais uma ferramenta estratégica para vários ramos organizacionais. Dentro deste contexto, a previsão de séries temporais tem desempenhado um papel muito crítico na tomada de decisões. Recentemente, a capital do Paraná enfrentou uma grave crise de saúde, com períodos de escassez que geraram grave instabilidade no abastecimento das casas de muitas famílias. Esta dissertação aborda o problema da previsão da demanda de água que ocorreu na cidade de Curitiba, coletando os períodos dos anos 2018 a 2020, visando o ano 2020, que foi o ano em que ocorreu a maior demanda de água, fazendo com que os reservatórios sofressem com isso, vários fatores. Na tomada de decisão deste problema em questão, alguns métodos encontrados na revisão que foi realizada durante este trabalho são utilizados, para serem previstos em alguns horizontes de previsão, o horizonte aqui abordado é uma forma de resolver a questão da demanda de água e assim validar os modelos para ver qual deles é o mais eficiente, o horizonte adotado foi a previsão de 1, 7, 14 e 30 dias à frente, de modo que cada método irá lidar com os dados ao longo do tempo. A fim de mitigar e resolver o problema que a SANEPAR enfrentou no ano 2020, para que não ocorra novamente ou para que não nos pegue despreparados no próximo evento que possa surgir. Com o evento isolado que aconteceu no ano em questão e que pode não se repetir em anos futuros, este trabalho tem como objetivo melhorar o uso da água. Os métodos derivados dos modelos ARIMA, listando assim os modelos são AR, ARX, MA, ARMA, ARIMA, SARIMA, SARIMAX e ARIMAX, pois cada modelo tem sua particularidade os modelos com variáveis exógenas podem parecer graficamente melhores de serem previstos do que os modelos ARIMA sem variáveis exógenas. Os modelos com aumento gradual são os melhores modelos para prever com os menores erros. Os modelos chamados boosting ou árvore de regressão de gradiente, os seguintes modelos LR, XGboost random forest regression e Light GBM foram usados, estes modelos para séries temporais são listados como os melhores modelos porque alguns deles usam a forma de previsão de gradiente. É obtido em algumas métricas de erro, quanto menor o erro, melhor para a tomada de decisão. As métricas adotadas neste trabalho são MAPE, MAE e RMSE, em séries temporais estas métricas são mais freqüentes, com modelos de previsão melhores ou mais eficazes em algumas circunstâncias sem previsão de horizonte futuro, O modelo XGBoost tem erro de 0,013% na métrica MAPE apenas analisando sobre esta métrica, e o modelo LR tem o maior erro de 21% no maior horizonte de previsão (30 dias), o modelo MA vem com erro de 11,57% e o modelo LR com erro de

548,59%. Assim, o modelo LR para um conjunto de dados menor pode ser mais eficiente do que os outros modelos, já que trabalha com um pequeno volume de dados e os erros ficam maiores à medida que o horizonte aumenta.

**Palavras-chave:** Previsão, Economia de água, Séries temporais, Análise de séries temporais.

## Abstract

Facing a competitive scenario, assertive demand forecasting has increasingly become a strategic tool for several organizational branches. Within this context, time series forecasting has played a very critical role in decision making. Recently, the capital city of Paraná faced a serious health crisis, with periods of shortages that generated severe instability in the supply of many families' homes. This dissertation addresses the problem of forecasting water demand that occurred in the city of Curitiba by collecting the periods from the years 2018 to 2020, aiming at the year 2020 which was the year that the highest water demand occurred, causing the reservoirs to suffer from this, several factors. In the decision making of this problem in question, some methods found in the review that was conducted during this work are used, to be predicted in some forecast horizons, the horizon addressed here is a way to solve the issue of water demand and thus validate the models to see which one is the most efficient, the horizon adopted was the forecast of 1, 7, 14 and 30 days ahead, so that each method will deal with the data over time. In order to mitigate and solve the problem that SANEPAR faced in the year 2020, so that it doesn't occur again or that it doesn't catch us unprepared in the next event that may arise. With the isolated event that happened in the year in question and may not be repeated in future years, this work aims to improve the use of water. The methods derived from ARIMA models, thus listing the models are AR, ARX, MA, ARMA, ARIMA, SARIMA, SARIMAX and ARIMAX, as each model has its own particularity the models with exogenous variables may seem graphically better to be predicted than the ARIMA models without exogenous variables. Gradient boosting models are the best models to predict with the lowest errors. The models called boosting or gradient regression tree, the following models LR, XGboost random forest regression and Light GBM were used, these models for time series are listed as the best models because some of them use the gradient way of predicting. It is obtained in some error metrics, the smaller the error the better for decision making. The metrics adopted in this work is MAPE, MAE and RMSE, in time series these metrics are more frequent, with forecasting models better or more effective in some circumstances with in forecasting no future horizon, The XGBoost model has 0.013% error in the MAPE metric just analyzing over this metric, and the LR model has the highest error of 21% in the longest forecast horizon (30 days), the MA model comes with 11.57% error and the LR model with 548.59% error. Thus, the LR model for a smaller data set can be more efficient than the other models, since it works with a small volume of data and the errors get higher as the horizon increases.

**Keywords:** Forecasting, Water economy, Time series, Time series analysis.

# **Lista de Abreviaturas e Siglas**

AdaBoost	Impulso ou Estímulo adaptativo (do inglês <i>Adaptive Boosting</i> )
AR	Auto-Regressivo
ARIMA	Média Móvel Integrada Auto-Regressiva (do inglês <i>auto-regressive integrated moving average</i> )
ARX	Auto-Regressivo com variável Exógena (do inglês <i>auto-regressive with exogeneous inputs</i> )
BrownBoost	Algoritmo de aumento
CNN	Rede Neural Convolucional (do inglês <i>Convolutional Neural network ou ConvNet</i> )
DBN	Rede de Crenças Profundas (do inglês <i>Deep Belief Network</i> )
EFB	Pacote de características exclusivas (do inglês <i>Exclusive Feature Bundling</i> )
FT	flow transmitter (Transmissor de fluxo)
Hz	Hertz
INMET	Instituto Nacional de Meteorologia
LGBMRegressor	Regressão Ligh GBM
Light GBM	Máquina de Impulso de Gradiente Leve (do inglês <i>Light Gradient Boosting Machine</i> )
LogitBoost	Representa uma aplicação de técnicas de regressão logísticas
LPBoost	Reforço da Programação Linear (do inglês <i>Linear Programming Boosting</i> )
LR	Regressão linear (do inglês <i>linear regression</i> )
LSTM	Memória de longo curto prazo (do inglês <i>Long short-term memory</i> )
$m^3$	Metros cúbicos
$m^3/h$	Metros cúbicos por hora
MadaBoost	Modificando o sistema de ponderação da AdaBoost
MAE	Erro Médio Absoluto (do inglês <i>Mean Absolute Error</i> )
MAPE	Erro Percentual Médio Absoluto (do inglês <i>Mean Absolute Percentage Error</i> )
<i>mca</i>	Metros coluna d'água

ML	Aprendizado de máquina (do inglês <i>machine learning</i> )
mm	Milímetros
MSE	Erro médio quadrático (do inglês <i>Mean Squared Error</i> )
PR	Estado do Paraná
RBAL	Recalque Bairro Alto
RFR	Random Forest Regression
RMSE	Erro de Raiz Média Quadrática (do inglês <i>Root Mean Squared Error</i> )
RNN	Rede Neural Recorrente (do inglês <i>Recurrent Neural Network</i> )
SANEPAR	Companhia de Saneamento do Paraná
SARIMA	Auto-Regressivos Integrados de Médias Móveis com Sazonalidade (do inglês <i>Integrated Auto-Regressive Moving Averages with Seasonality</i> )
SARIMAX	Média Móvel Integrada Auto-Regressiva Sazonal com regressores eXogenous (do inglês <i>Seasonal Auto-Regressive Integrated Moving Average with eXogenous regressors</i> )
SVM-VAR	Máquinas de vetor de suporte - Vetores Auto-Regressivos
Totalboost	Impulso total
XGBoost	Impulso Gradiente Extremo (do inglês <i>eXtreme Gradient Boosting</i> )
XGBRegressor	Regressão XGBoost

## **Lista de Tabelas**

1	Cruzamento de palavras-chave através da aplicação de filtros de ano e de linguagem . . . . .	18
2	Fator de impacto . . . . .	20
3	Áreas e seus valores respetivos de artigos em cada área. . . . .	23
4	Descrição estatística dos dados com o filtro aplicado das 18h às 21h . . . . .	45
5	Teste Nemenyi . . . . .	55
6	Comparação dos modelos com 1 dia de antecedência 24h <b>Treinamento</b> . . . . .	64
7	Comparação dos modelos com 1 dia de antecedência 24h <b>Validação</b> . . . . .	65
8	Comparação dos modelos com 1 dia de antecedência 24h <b>Teste</b> . . . . .	65
9	Comparação dos modelos com 1 dia de antecedência 24h <b>Completo</b> . . . . .	66
10	Comparação dos modelos com 7 dias de antecedência 24h <b>Treinamento</b> . . . . .	66
11	Comparação dos modelos com 7 dias de antecedência 24h <b>Validação</b> . . . . .	67
12	Comparação dos modelos com 7 dias de antecedência 24h <b>Teste</b> . . . . .	67
13	Comparação dos modelos com 7 dias de antecedência 24h <b>Completo</b> . . . . .	68
14	Comparação dos modelos com 14 dias de antecedência 24h <b>Treinamento</b> . . . . .	68
15	Comparação dos modelos com 14 dias de antecedência 24h <b>Validação</b> . . . . .	69
16	Comparação dos modelos com 14 dias de antecedência 24h <b>Teste</b> . . . . .	69
17	Comparação dos modelos com 14 dias de antecedência 24h <b>Completo</b> . . . . .	70
18	Comparação dos modelos com 30 dias de antecedência 24h <b>Treinamento</b> . . . . .	70
19	Comparação dos modelos com 30 dias de antecedência 24h <b>Validação</b> . . . . .	71
20	Comparação dos modelos com 30 dias de antecedência 24h <b>Teste</b> . . . . .	71
21	Comparação dos modelos com 30 dias de antecedência 24h <b>Completo</b> . . . . .	72
22	Comparação dos modelos Ljung Box <b>Treinamento</b> . . . . .	72
23	Comparação dos modelos Ljung Box <b>Validação</b> . . . . .	73
24	Comparação dos modelos Ljung Box <b>Teste</b> . . . . .	73
25	Comparação dos modelos Ljung Box <b>Completo</b> . . . . .	73

# **Lista de Figuras**

1	Mapa das Etapas . . . . .	4
2	Estrutura da dissertação . . . . .	8
3	Dados completos com frequência média de 24h . . . . .	9
4	Plotagem dos dados para o ano 2020 . . . . .	10
5	Exemplo de séries temporais . . . . .	11
6	Processo estocástico . . . . .	11
7	Mapa conceitual do problema de pesquisa . . . . .	13
8	Etapas da Revisão. . . . .	14
9	Palavras-chave mais populares na Scopus. . . . .	16
10	Palavras-chave mais populares na WoS . . . . .	17
11	Analise das quantidades de artigos em relação aos anos. . . . .	18
12	Relação de autores entre artigos publicados . . . . .	20
13	Ligaçāo bibliográfica entre os autores . . . . .	21
14	Mapa mundial da publicação de artigos em todo o mundo . . . . .	22
15	Áreas de aplicāo do tema . . . . .	23
16	Modelo AR(7) . . . . .	28
17	ARX (7) . . . . .	28
18	Modelo MA(7) . . . . .	30
19	ARMA (7,7) . . . . .	31
20	ARIMA (7,1,7) . . . . .	32
21	SARIMA (7,1,7)(2,1,1) <sub>12</sub> . . . . .	33
22	ARIMAX (7,1,7) . . . . .	34
23	SARIMAX (7,1,7)(2,1,1) <sub>12</sub> . . . . .	34
24	Corelação de Pearson . . . . .	35
25	Regressão linear LT01 vs PT01 correlação 98% . . . . .	36
26	Regressão linear (LR) um passo a frente . . . . .	37
27	Regressão da Floresta Aleatória (RFA) . . . . .	38
28	Esquema da Floresta Aleatória . . . . .	38
29	Impulsionando gradiente com XGBoost e LightGBM . . . . .	39
30	Crescimento em folha versus crescimento em nível . . . . .	42
31	XGBoost e LighGBM regressão . . . . .	42
32	Solução para o acionamento das bombas . . . . .	46
33	Decomposição STL aditiva dos dados coletados . . . . .	47
34	Decomposição STL multiplicativa dos dados coletados . . . . .	48
35	Violino no nível do reservatório . . . . .	49

36	Violino da vazão de recalque . . . . .	49
37	Autocorrelação e Autocorrelação parcial . . . . .	51
38	Ruído branco . . . . .	52
39	Comparação dos modelos ARIMAS . . . . .	56
40	Comparação de modelos de regressão . . . . .	56
41	Comparação dos modelos AR, ARX e MA, 1 dia à frente . . . . .	74
42	Comparação dos modelos AR, ARX e MA, 7 dias à frente . . . . .	74
43	Comparação dos modelos AR, ARX e MA, 14 dias à frente . . . . .	75
44	Comparação dos modelos AR, ARX e MA, 30 dias à frente . . . . .	75
45	Comparação dos modelos ARMA e ARIMA, 1 dia à frente . . . . .	76
46	Comparação dos modelos ARMA e ARIMA, 7 dias à frente . . . . .	76
47	Comparação dos modelos ARMA e ARIMA, 14 dias à frente . . . . .	77
48	Comparação dos modelos ARMA e ARIMA, 30 dias à frente . . . . .	77
49	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente . . . . .	78
50	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente . . . . .	78
51	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à frente . . . . .	79
52	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à frente . . . . .	79

# Sumário

<b>1</b>	<b>Introdução</b>	<b>1</b>
1.1	Contexto da pesquisa	1
1.1.1	Motivação da pesquisa	2
1.2	Objetivo geral	2
1.2.1	Objetivos específicos e questão de pesquisa	2
1.3	Descrição do problema	3
1.4	Procedimentos metodológicos	4
1.4.1	Etapas da pesquisa	4
1.5	Justificativa da pesquisa	7
1.5.1	Contribuições	7
1.6	Estrutura do trabalho	8
<b>2</b>	<b>Referencial</b>	<b>9</b>
2.1	Detecção de anomalias	9
2.2	Revisão sistemática da literatura	10
2.3	Problematização da Revisão	12
2.4	Metodologia	13
2.5	Resultados da busca de revisão	16
2.6	Conclusão da revisão	24
<b>3</b>	<b>Base Teórica</b>	<b>26</b>
3.1	Métricas de Erros	26
3.2	ARIMA, SARIMA e SARIMAX	27
3.2.1	Componente auto-regressivo – AR(p)	27
3.2.2	Média Móvel – MA(q)	29
3.2.3	Modelos ARMA e ARIMA	30
3.2.4	Modelos SARIMA, ARIMAX e SARIMAX	32
3.3	Modelos Regressivo	34
3.3.1	Régressão Linear (LR)	34
3.3.2	Floresta Aleatória	37
3.3.3	LightGBM e XGboost	38
3.3.4	O Gradiente em Gradiente de Boosting (Reforço)	39
3.3.5	Algoritmos de boosting de gradiente	40
3.3.6	A diferença entre XGBoost e LightGBM	41
<b>4</b>	<b>Resultados</b>	<b>44</b>

4.1	Planejamento do Problema . . . . .	44
4.1.1	Análise Exploratória dos dados (EDA) . . . . .	44
4.1.2	Múltiplas entradas e saída única (MISO) . . . . .	46
4.1.3	Decomposição STL . . . . .	47
4.1.4	Separação dos dados . . . . .	52
4.1.5	Estratégia de Previsão . . . . .	53
4.1.6	Horizonte . . . . .	53
4.1.7	Modelos de previsão e métricas de desempenho . . . . .	54
4.1.8	Teste de Significância . . . . .	54
4.1.9	Comparação dos modelos . . . . .	55
<b>5</b>	<b>Conclusões . . . . .</b>	<b>58</b>
5.1	Limitações da pesquisa e propostas futuras . . . . .	58
<b>Referências . . . . .</b>	<b>60</b>	
<b>A</b>	<b>Apêndice - Comparação dos modelos de previsão de séries temporais média de 24h . . . . .</b>	<b>64</b>
<b>B</b>	<b>Apêndice - Comparação dos modelos de previsão com o método Ljung Box . . . . .</b>	<b>72</b>
<b>C</b>	<b>Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h . . . . .</b>	<b>74</b>
<b>D</b>	<b>Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h . . . . .</b>	<b>76</b>
<b>E</b>	<b>Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e SARIMAX (7,1,7) (2,1,0,12) 24h . . . . .</b>	<b>78</b>

# 1 Introdução

Este capítulo apresenta a introdução do que é abordado nesta dissertação, usando modelos ML, dentro destes modelos será abordada a previsão futura dos dados coletados na SANEPAR Curitiba no estado do Paraná, estes dados foram coletados no bairro superior nos anos 2018 a 2020 houve uma escassez de água que afetou a todos em Curitiba.

No uso de séries temporais pensando neste contexto de tomada de decisão, pode-se pensar como a aplicação de modelos ML em séries temporais, usando os modelos mais clássicos encontrados durante uma revisão sistemática do conteúdo, para tabular alguns modelos que são usados na literatura.

## 1.1 Contexto da pesquisa

Ribeiro et al. (2021) 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.

Em séries temporais o aprendizado de máquinas é frequentemente utilizado para grandes processamentos de dados, com o conjunto de dados SANEPAR da cidade de Curitiba no estado do Paraná há um volume significativo no consumo de água e com a escassez que a cidade experimentou é necessário avaliar os dados para ter certeza do que está acontecendo, quando há escassez de água e picos que ocorrem entre horas e dias.

Entre os modelos preditivos que serão apresentados em uma revisão sistemática, avaliar o melhor modelo que podemos utilizar e validar quando e como ocorre a escassez de água. Estas análises estarão em *python*.

Explorando o que são séries temporais e aprendizagem de máquinas, séries temporais são dados armazenados ao longo do tempo que permitem a um observador analisar anomalias nos dados. Nas séries cronológicas, a classificação dos dados por ano ou dia é crítica, e se os dados forem atribuídos aleatoriamente, pode tornar mais difícil prever e tomar decisões com base nos dados coletados. A análise das médias pode ser bastante perigosa se você não excluir pontos fora da curva também conhecidos como “*outliers*”. Isto pode gerar dados muito positivos ou negativos que não correspondem à realidade.

### 1.1.1 Motivação da pesquisa

De acordo com (VASCONCELOS, 2020) 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 2020 a chuva tem marcado a quantia de 1.704 mm. (VASCONCELOS, 2020) 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 2020, com 1.158 mm.

Em mediana a esta motivação podem ser melhor interpretados os dados que a SANPEAR ofereceu para prever e evitar a escassez de água que foi registrada e a anomalia que foi detectada em 2020, com o retorno das chuvas os reservatórios tinham aumentado de nível.

## 1.2 Objetivo geral

O objetivo para esta dissertação é encontrar o melhor modelo de série temporal para o problema da escassez de água que ocorreu em Curitiba. Com vários modelos coletados no decorrer da dissertação entre modelos de regressão e aplicados ao gradiente, os modelos *boosting* na literatura os melhores modelos de previsão de séries temporais e os modelos ARIMA e o ARIMA atualizado. Predição e análise das anomalias nos dados e por que ocorrem.

### 1.2.1 Objetivos específicos e questão de pesquisa

Para este trabalho, pretende-se procurar anomalias que possam ocorrer nos dados e por que tais anomalias ocorrem e responder às perguntas da pesquisa.

**Q 1** A pressão é suficiente para a demanda diária?

**Q 2** Quanta água deve ter no reservatório para evitar o acionamento das bombas no horário de pico (18 as 21 h)? Quanto maior a frequência de funcionamento da bomba maior a demanda. Valor máximo 60 Hz.

**Q 3** Qual a vazão ótima para atender a demanda? Quanta pressão para atender a demanda?

**Q 4** Ponto de equilíbrio entre demanda e vazão e ter um armazenamento sem necessidade de acionar as bombas no período do custo energético mais caro (18 às 21 horas).

**Q 5** Se a SANEPAR ativar as bombas de sucção das 18 às 21 horas ela tem o maior custo energético, isto é, ela paga mais caro pela energia neste período.

- a. Qual o nível que deve estar no reservatório para não ser necessário a SANEPAR ativar as bombas das 18 às 21 horas sem faltar água para a população? Verificar a média das vazões nos horários críticos (onde tem a maior demanda 18 às 21 horas) para as diferentes estações do ano (Outono, Inverno, Primavera, Verão).
- b. Existe tendência, padrão, sazonalidade para os dados destes 3 anos do Bairro Alto?
- c. Identificar quais os horários de maior demanda das 18 às 21?
- d. Quanto tenho que armazenar previamente no reservatório para não acionar as bombas no horário de pico?
- e. Se a vazão cresce e a pressão decresce temos uma ANOMALIA na rede (com base no histórico).

### 1.3 Descrição do problema

Esta subseção discute as variáveis no conjunto de dados e como elas serão previstas.

- Bombas de sucção (B1, B2 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 (FT02) ( $m^3/h$ )
- Vazão de recalque (FT03) ( $m^3/h$ )
- Pressão de Sucção (PT01SU) (mca)
- Pressão de Recalque (PT02RBAL) (mca)

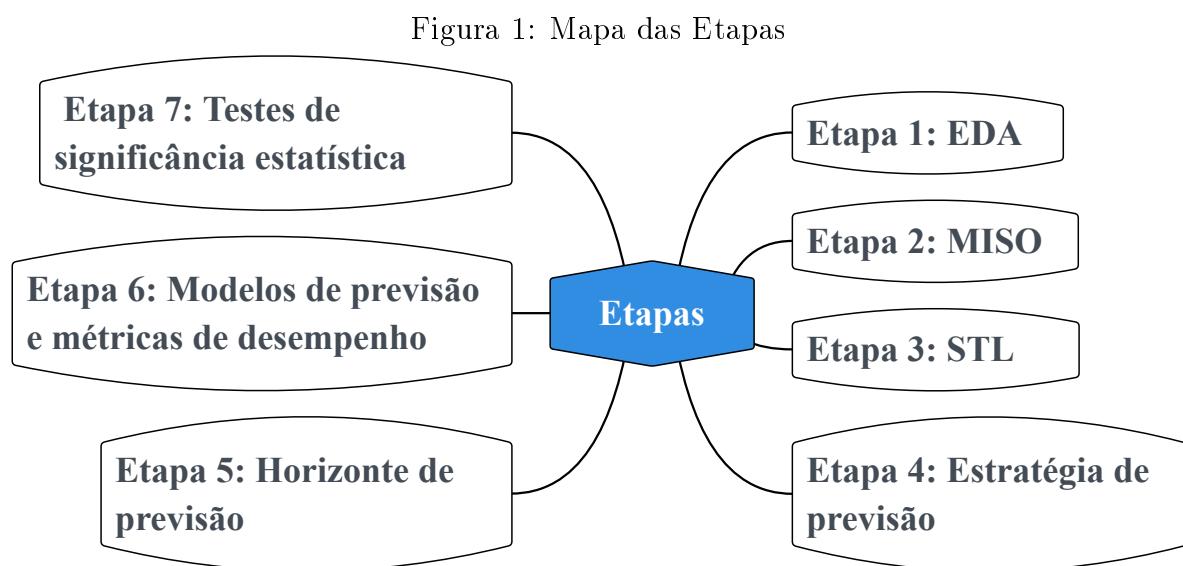
Na pesquisa será utilizada a variável LT01, que é o nível do reservatório, este nível é de grande importância como mostram as Figuras 3 e 4.

## 1.4 Procedimentos metodológicos

Esta parte trata de como a dissertação será realizada em cada etapa da análise.

### 1.4.1 Etapas da pesquisa

A pesquisa seguiu as seguintes etapas:



Fonte: Elaboração própria

#### **Etapa 1** Análise exploratória dos dados – EDA ( do inglês *Exploratory Data Analysis*)

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 ARIMAS. 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 *Seasonal-Trend Decomposition*) Sazonalidade, Tendência e Resíduo

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:

$$y_i = s_i + t_i + r_i \quad (1)$$

Onde:

- $y_i$  = O valor da série de tempo no ponto  $i$ .
- $s_i$  = O valor do componente sazonal no ponto  $i$ .
- $t_i$  = O valor do componente de tendência no ponto  $i$ .
- $r_i$  = O valor do componente restante no ponto  $i$ .

**Etapa 4** Verifique a média e o desvio padrão de cada um desses conjuntos para obter a divisão mais apropriada dos dados. Dividir o conjunto de dados em treinamento, validação e teste. 70% para treinamento e validação e 30% para testes a partir daí, dos 70% divididos em 80% para treinamento e 20% para validação.

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

A estratégia recursiva envolve o uso de um modelo de uma etapa várias vezes, onde a previsão para a etapa de tempo anterior é usada como uma entrada para fazer uma previsão na etapa de tempo seguinte.

No caso de prever a demanda de água para os próximos dias, desenvolveríamos um modelo de previsão de uma etapa. Este modelo seria então usado para prever o dia 1, então esta previsão seria usada como um *input* de observação para prever o dia 2.

Por Exemplo

$$prediction(t+1) = model_1(obs(t-1), obs(t-2), \dots, obs(t-n)) \quad (2)$$

$$prediction(t+2) = model_2(obs(t-2), obs(t-3), \dots, obs(t-n)) \quad (3)$$

Brownlee (2016) como as previsões são usadas no lugar das observações, a estratégia 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 a frente)

Nessa etapa, o tipo de horizonte foi escolhido de forma a mudar entre os dias, prevendo um passo à frente, uma semana, duas semanas e um mês.

#### **Etapa 7** Modelos de previsão e métricas de desempenho

Os modelos discutidos aqui são os modelos clássicos de previsão, juntamente com os modelos de regressão gradiente. Os modelos são AR, ARX, ARMA, ARIMA, SARIMA, SARIMAX e ARIMAX, seguidos pelos modelos de regressão LR, XGBRegressor, Random Forest Regressor e LGBMRegressor. Esses modelos adotados foram escolhidos pela revisão sistemática realizada na dissertação.

As métricas usadas em toda a dissertação são as métricas RMSE, MAE e MAPE, encontradas na revisão e uma das mais usadas até hoje, na subseção 3.1 é explicado em mais detalhes cada uma delas.

#### **Etapa 8** Aplicar os modelos de previsão e fazer comparativo baseado em testes de significância estatística (*Friedman* e *Nemenyi*)

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.

## 1.5 Justificativa da pesquisa

No decorrer desta dissertação o seguinte para que se possa prever e tomar a decisão mais correta para evitar a real escassez de água e como este problema pode ser resolvido para que ele não ocorra novamente.

### 1.5.1 Contribuições

Após as perguntas de pesquisa feitas na subseção 1.2.1 tem duas contribuições, a primeira considerando a demanda de água na cidade de Curitiba entre **Q** 1 a **Q** 4 é feita a previsão da demanda de água, as outras estão em como é o consumo de água na cidade e o gasto de energia no período de pico mostrado em **Q** 5a. a **Q** 5e.

Assim, utilizando os métodos escolhidos de previsão de séries temporais, tais como os modelos ARIMA e ARIMA atualizados, tais como os modelos ARMA, SARIMA, ARIMAX e SARIMAX, outros modelos mais simples que vêm do modelo ARIMA, tais como os modelos AR, ARX e MA para uma previsão mais precisa como no **Q** 5 em diante os modelos regressivos ou modelos gradientes. Os modelos regressivos testados aqui foram o LR e RFR, para os modelos de gradiente XGBoost e Ligth GBM foi usado para se tornar uma opção mais viável no momento de tomar a decisão em meio aos gastos de energia e água que a empresa SANEPAR tinha e a fim de minimizar esses gastos. O horizonte de previsão foi estabelecido para que a melhor decisão pudesse ser tomada em relação à demanda de água.

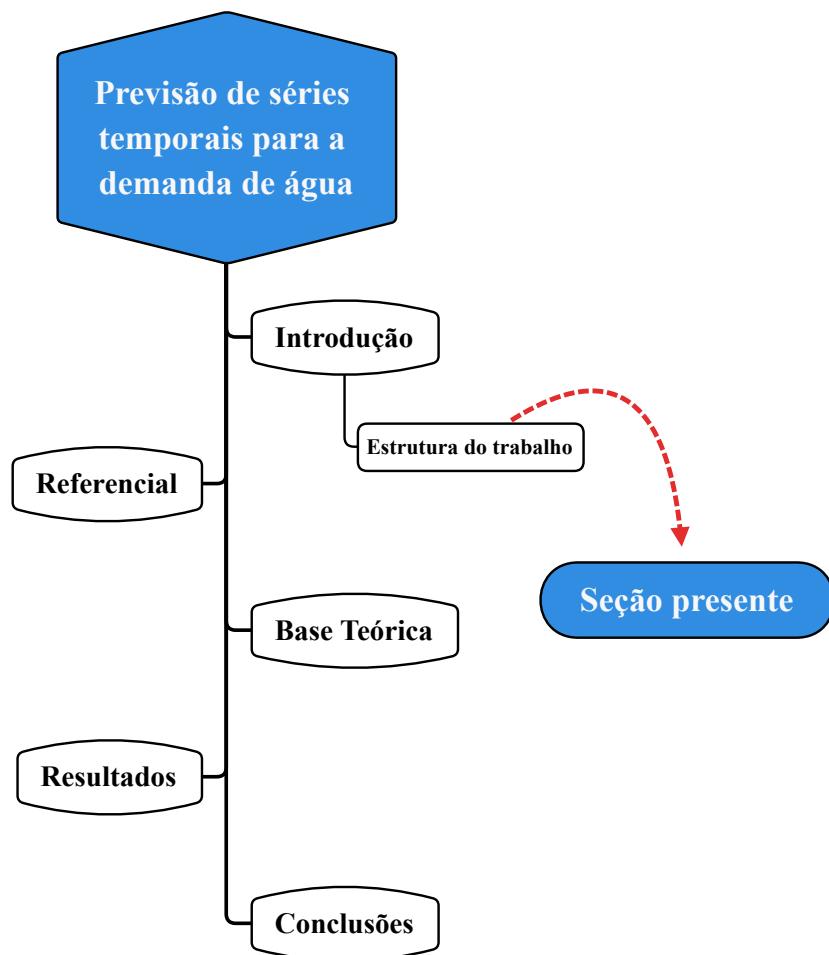
Em ambas as contribuições a tabulação foi feita tanto a curto prazo (1 a 7 dias, uma semana) quanto a longo prazo (14 a 30 dias, um mês). Para que o melhor modelo, tanto a curto prazo como a longo prazo, seja mostrado e evidenciado. Os modelos ARIMA para o problema em questão no horizonte de previsão a longo prazo têm melhor desempenho do que os modelos de aumento de gradiente os modelos de gradiente são mais viáveis na previsão a curto prazo, por exemplo, de um dia para uma semana. E ainda assim, os modelos ARIMA ou os modelos que provêm dele superam os modelos de gradiente.

A comparação de modelos de previsão pode ser uma das coisas em que esta dissertação mais busca, pois o método Ljung Box tem a medida em que cada modelo ARIMA funciona no curto e no longo prazo, no Apêndice B tem a comparação dos modelos por meio desse teste estatístico, os modelos de regressão feitos têm a comparação deles e dos modelos ARIMAS nas Figuras 39 e 40 no Apêndice A há a comparação dos modelos feitos no decorrer da dissertação a comparação dos modelos toma a melhor decisão para o problema.

## 1.6 Estrutura do trabalho

Este documento está estruturado em 5 capítulos, divididos da seguinte forma:

Figura 2: Estrutura da dissertação



Fonte: Elaboração própria

O capítulo 1 apresenta a introdução do trabalho, contendo a contextualização, a motivação, o objetivo geral, os objetivos específicos, a metodologia utilizada, a justificativa da pesquisa, as contribuições, as publicações e a organização do trabalho. O capítulo 2 apresenta a descrição do problema, revisão teórica do trabalho, fazendo uma visão geral dos principais pesquisadores sobre as questões abordadas na pesquisa. O capítulo 3 apresenta os modelos que serão trabalhados nos dados coletados. O capítulo 4 apresenta os resultados da pesquisa, assim como uma análise dos resultados gerados. O capítulo 5, finalmente, apresenta as considerações finais da pesquisa e algumas propostas para pesquisas futuras.

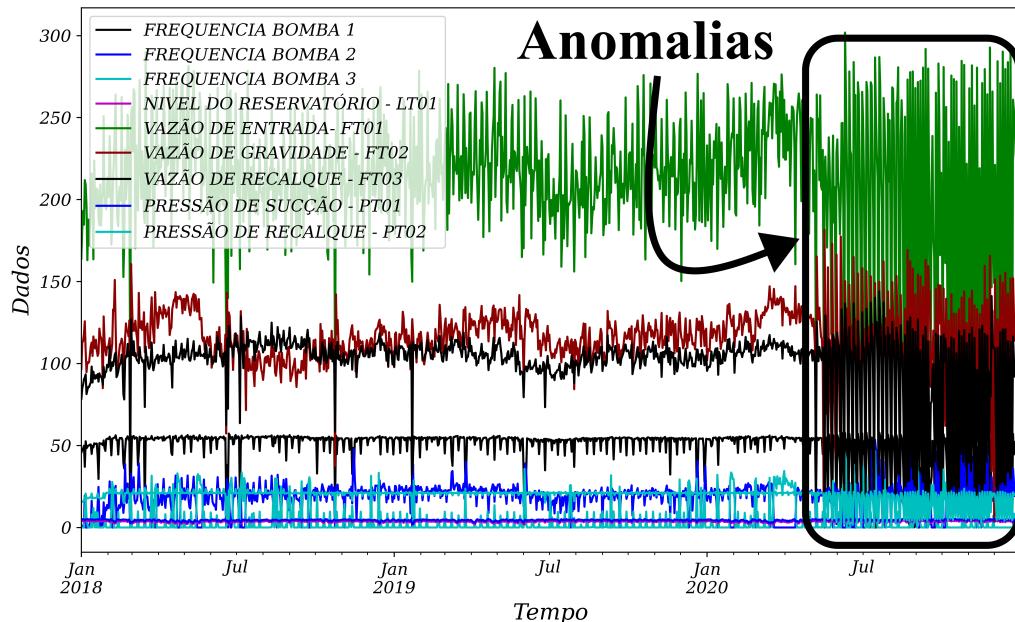
## 2 Referencial

Este capítulo apresentará a base da literatura que foi coletada durante a preparação desta dissertação, embora os resultados sejam um pouco menores do que os de uma tese, ainda são relevantes para o trabalho realizado aqui.

### 2.1 Detecção de anomalias

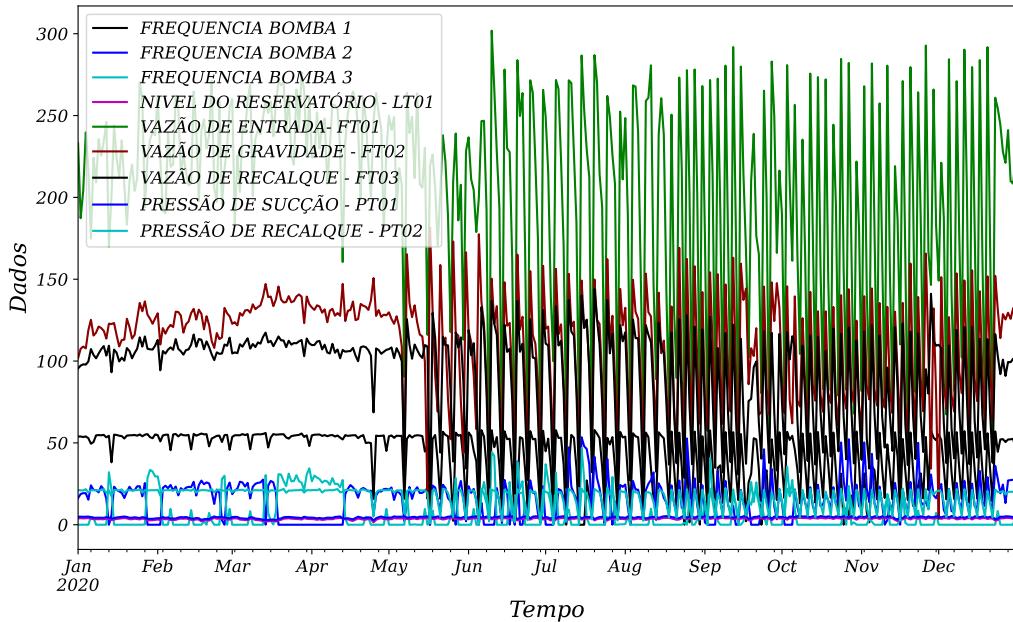
As anomalias em séries temporais é um desafio muito grande para os previsores saber quando os dados sofrem uma mudança se ela não estiver muito evidente nos dados é um exercício de concentração, com isso os dados coletados ao longo do tempo pela empresa SANEPAR traz as anomalias mais claras do que se imaginava, pois a falta de água que sofreu a cidade de Curitiba se alastrou por dias logo a frente é mostrado os gráficos de linha utilizados para a **Etapa 1** do trabalho em questão.

Figura 3: Dados completos com frequência média de 24h



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 4: Plotagem dos dados para o ano 2020



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Os dados coletados têm o tamanho de 26306 linhas  $\times$  9 colunas, para tanta relação que será usada nos modelos da subseção 1.4 para prever e analisar as anomalias apresentadas nas Figuras 3 e 4.

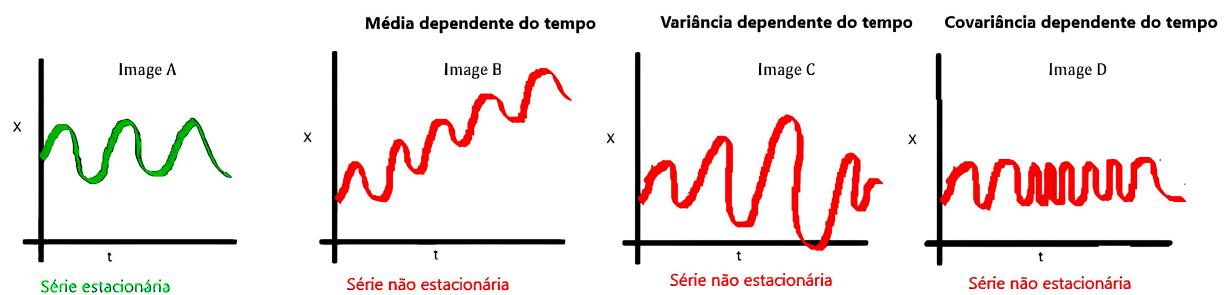
## 2.2 Revisão sistemática da literatura

As séries temporais aparecem em vários campos do conhecimento, tais como Economia (preços de estoque diários, taxa de desemprego mensal, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (chuvas, temperatura diária, velocidade do vento), etc. Ao longo dos anos tem usado ferramentas computacionais para tornar esta previsão mais eficiente, com aprendizagem de máquinas e algumas características que podem ser aplicadas em linguagem computacional através da linguagem *python* e *R*, as melhores linguagens para trabalhar com séries temporais hoje em dia.

Para entender melhor este conceito de série temporal, suponhamos que um maratonista que esteja correndo há vários anos e uma pessoa sedentária se submeta a uma corrida de, no máximo, 5 km, ambos corram ao mesmo tempo para que tenham um monitor de frequência cardíaca para que possa ser monitorado por um médico se você pegar os dados desde o início e compará-los com o final da corrida, o maratonista terá uma

série mais estacionária porque ele tem o hábito de correr regularmente enquanto a pessoa sedentária terá uma série não estacionária como mostrado na Figura 5.

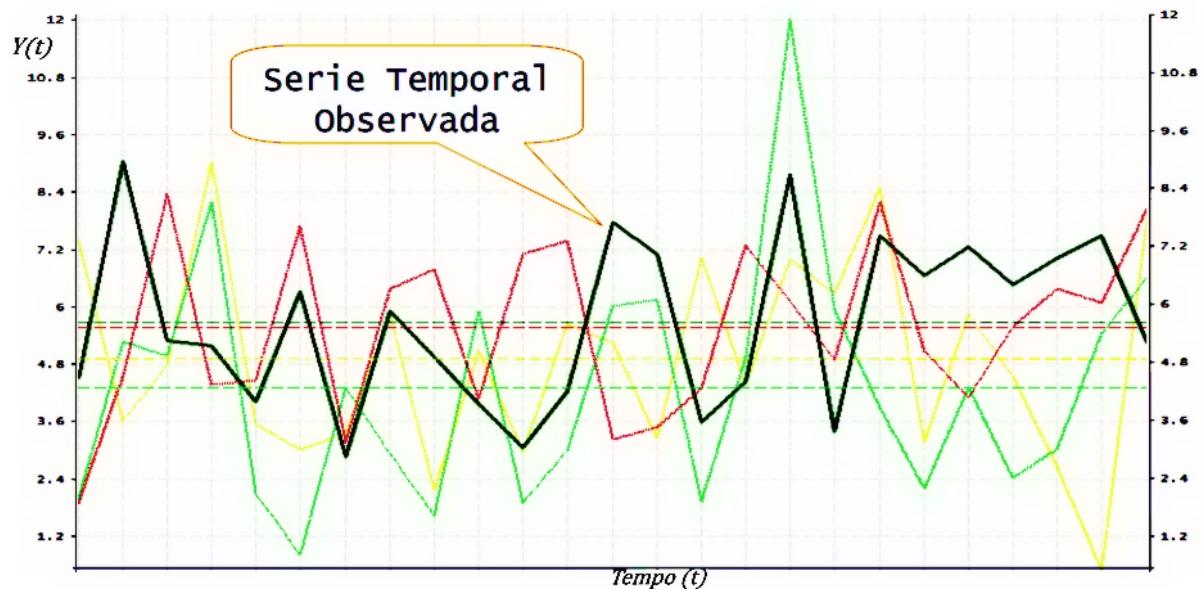
Figura 5: Exemplo de séries temporais



Fonte: (BRANDÃO, 2020)

Na figura 5 observa-se que o eixo  $x$  representa os dados observados e  $t$  para o tempo percorrido. Além disso, as séries temporais são processos estocásticos por leis probabilísticas, o que significa que há a possibilidade de ser pensado como um conjunto de todas as trajetórias possíveis na Figura 5 é capaz de ser observado para uma variável alvo. Por exemplo, se você lançar um dado qualquer valor inteiro entre 1 e 6, mas apenas um número ocorrerá. Da mesma forma, em séries temporais existem infinitas possibilidades, entre elas apenas uma de acordo com as características que atenderam a esse período e que de fato ocorrerão.

Figura 6: Processo estocástico



Fonte: (PINHEIRO, 2022)

Com  $Y(t)$  os dados fictícios e  $\text{Tempo} (t)$  a linha do tempo da Figura 5.

De repente é pensado como um conjunto de todas as trajetórias possíveis que poderiam ser para observar uma variável.

Esta revisão sistemática da literatura, com o tema abordado até agora é sobre séries temporais, considerando o contexto aqui exposto este tema pode ser de grande relevância em diversas áreas, como mostrado na Figura 15. Realizando esta análise de séries temporais nos últimos 6 anos para poder observar as melhores realizações neste tema abordado aqui um curto período, mas tendo o tempo não muito a favor, então teve a opção de deixar este tempo específico para buscar artigos.

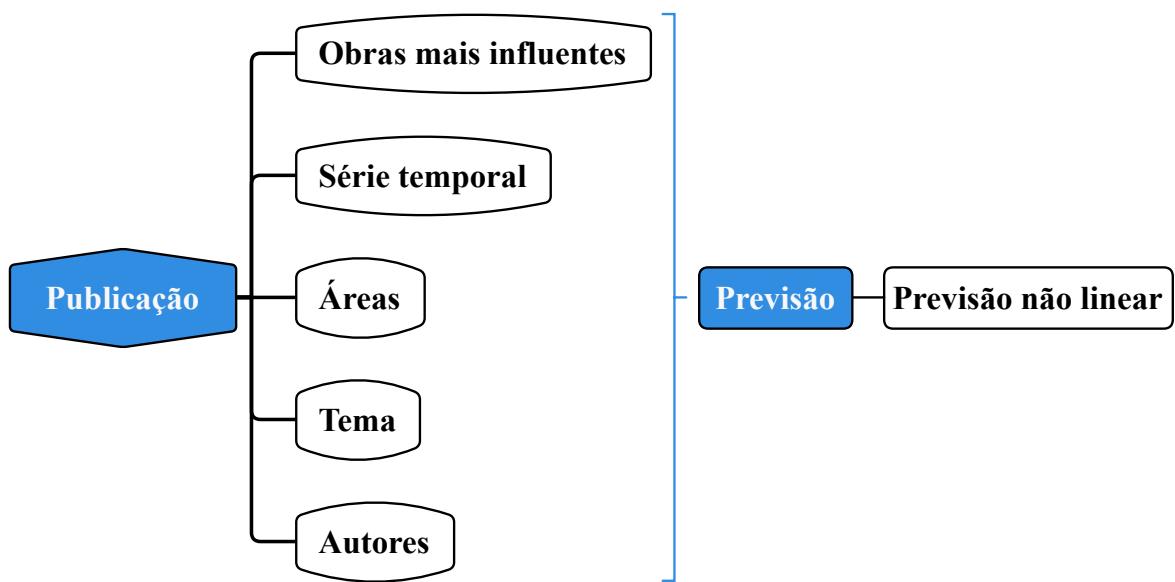
O objetivo desta revisão é analisar uma literatura menor, mas muito relevante. Como a própria série temporal procura analisar e modelar a dependência e considerando a ordem apresentada nas bases, por exemplo, os maiores autores e o ano de atividade que mais publicaram nos países que têm o maior número de publicações na apresentação das palavras-chave que serão mostradas, o objetivo é rever cada coisa que pode ser usada em uma aplicação de aprendizagem de máquina.

Em todos os artigos observados que tem uma contribuição científica neste trabalho é a análise do conceito de série temporal com o melhor uso das palavras-chave mesmo não tendo uma grande relação na aprendizagem de máquinas podem ser usados estes artigos como base para outros pesquisadores, aqui algumas análises muito simples para alguns leitores. Entretanto, é um ponto de partida para muitos que não conhecem o conceito de séries cronológicas ou revisão sistemática da literatura.

### **2.3 Problematização da Revisão**

Nesta seção é abordado um problema de pesquisa que pode ser compreendido por vários leitores na Figura 7 é apresentado um mapa conceitual de publicação e os autores são o pilar mais relevante para a revisão porque apresentam vários modelos que servirão de base e como se trata de séries temporais a previsão que pode ser feita neste contexto é um problema de grande significado em si mesmo.

Figura 7: Mapa conceitual do problema de pesquisa



Fonte: Elaboração própria

No mapa conceitual apresentado na Figura 7 é visto o problema sendo relacionado com palavras, tornando evidente o que será abordado durante o trabalho, deixando as questões de pesquisa em tópicos logo à frente.

**Q 1** Quais os autores que mais publicam sobre o assunto de séries temporais?

**Q 2** Quais os países que mais publicam sobre o assunto?

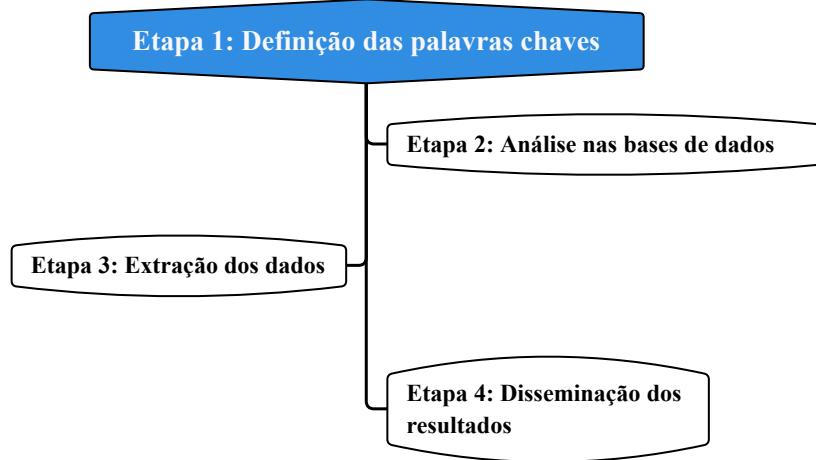
**Q 3** Quais as áreas que mais publicam sobre o tema?

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

## 2.4 Metodologia

Nesta seção é esclarecido como a revisão foi conduzida desde a análise do banco de dados até a conclusão da revisão.

Figura 8: Etapas da Revisão.



Fonte: Adaptado de Martins e Gorschek (2016)

**Etapa 1** A Figura 8 usa uma adaptação de Martins e Gorschek (2016) para esta revisão sistemática que está sendo analisada. Depois, há as buscas nos bancos de dados Scopus, Web of Science e Lens. No início foram utilizadas algumas bases no meio de tantas na literatura para melhor atender ao tema da pesquisa.

#### Campo de pesquisa Scopus

**TITLE-ABS-KEY** ("time series forecasting") AND **TITLE-ABS-KEY** ("time series analysis") AND ( **LIMIT-TO** ( **DOCTYPE** , "ar" ) ) AND ( **LIMIT-TO** ( **LANGUAGE** , "English" ) ) AND ( **LIMIT-TO** ( **PUBYEAR** , 2022 ) OR **LIMIT-TO** ( **PUBYEAR** , 2021 ) OR **LIMIT-TO** ( **PUBYEAR** , 2020 ) OR **LIMIT-TO** ( **PUBYEAR** , 2019 ) OR **LIMIT-TO** ( **PUBYEAR** , 2018 ) OR **LIMIT-TO** ( **PUBYEAR** , 2017 ) )

#### Campo de pesquisa na Web of Science

"times series forecasting"(All Fields) and "time series analysis"(All Fields) (Publication Years: 2022 or 2021 or 2020 or 2019 or 2018 or 2017) (Document Types: Articles) (Languages: English)

#### Campo de pesquisa de Lens

**Scholarly Works (11) = (** ("time series forecasting") AND ( ("time series analysis") AND ( "nonlinear forecasting" ) ) Filters: Year Published = ( 2016 - 2022 ) Publication Type = ( journal article )

Em todos os campos de busca, foram utilizados os últimos 6 anos, com exceção do site do Lens, onde optamos por 6 anos porque ele devolvia poucos artigos. Nesta

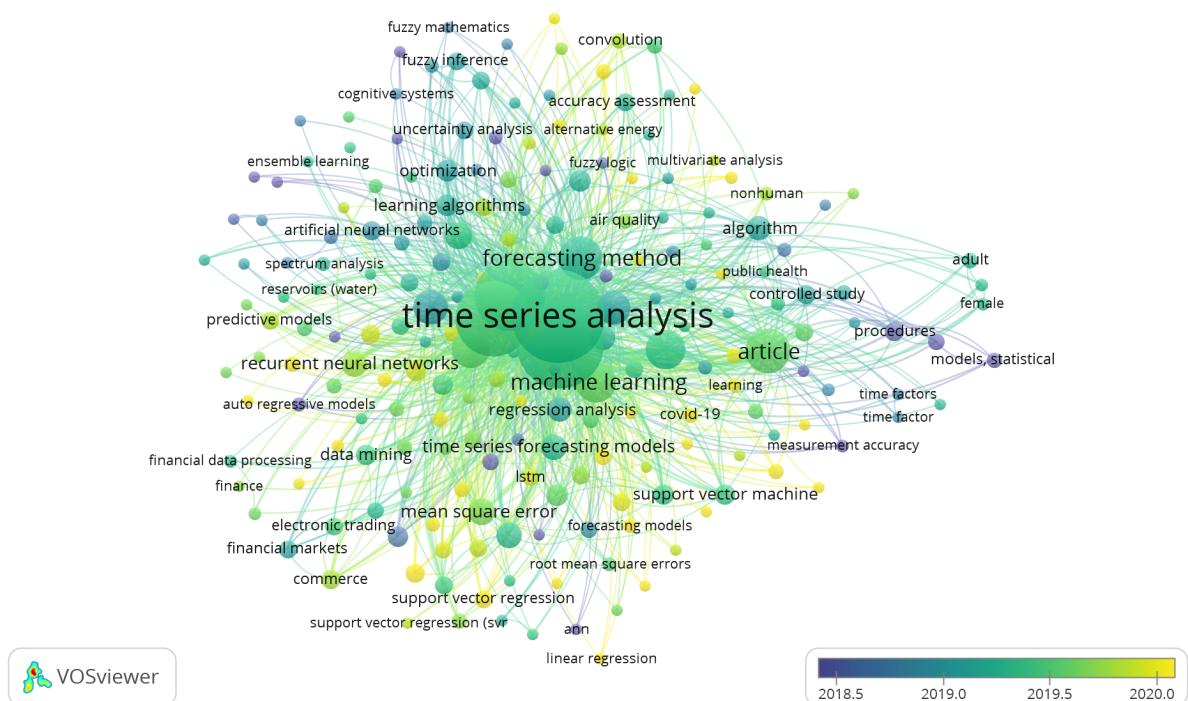
etapa, usamos as palavras-chave que melhor se adaptam à busca *time series forecasting and time series analysis and nonlinear forecasting*.

- Etapa 2** No cruzamento de palavras obtém-se um número considerável de artigos sem restringir a área em que cada artigo pode ser publicado. Na Tabela 1 foi feita uma tabulação dos resultados obtidos sem excluir a duplicata que trataremos na seção 2.5.
- Etapa 3** Esta etapa é avaliar cada dado obtido sem nenhum filtro no início da busca, a extração destes dados sem utilizar nenhum filtro anual nas buscas seria muitos artigos para analisar, por exemplo, no banco de dados Scopus seria com artigos de 498, na Web of Science seria com artigos de 140, e na Lente como não retorna muitos artigos é com 11 dando um total de 649 sem remover duplicata. É correto lembrar que estes artigos têm apenas o filtro da língua inglesa e do artigo, para melhorar a busca e a tomada de decisão usando o filtro dos anos, os últimos 6 anos é um valor mais agradável de artigos para usar com pouco tempo para analisar, e usando a diferença entre esta estimativa que foi feita na Tabela 1 são menos de 356 artigos para analisar. Lembrando que se a remoção de duplicatas foi feita, este número que foi obtido no resultado de todas as bases pode atingir um número ainda menor do que o pretendido neste trabalho.
- Etapa 4** Nesta etapa é mais analisar a dimensão do que está sendo trabalhado, fazendo a análise das áreas e lendo os artigos que são realmente importantes para a revisão. Como esta revisão está focada em séries temporais em um programa mestre de engenharia de produção e sistemas, vale a pena analisar a correlação. Desta forma, uma das áreas é a matemática, por isso foi selecionada nestes artigos que uma análise mais profunda dos artigos das séries temporais pode resultar se olharmos as áreas de especialização dos artigos pesquisados, como pode ser visto na Figura 15 que as áreas aqui citadas com grande relevância são **informática, engenharia e matemática** tem um número muito alto de publicações representando 50% da pesquisa, então a pesquisa está no caminho certo usando matemática básica para ter uma estimativa de quantos artigos podem ser eliminados seria cerca de 481 artigos, mas isto sem muita base que este número tem uma precisão. Usando o *software mendeley desktop* para estipular o valor exato de quantos artigos podem ser usados sem duplicação é deixado com um número de artigos de 308.

## 2.5 Resultados da busca de revisão

Nesta seção serão apresentados os resultados da pesquisa utilizando algum software, a fim de estipular o melhor uso de cada banco de dados utilizado durante o trabalho. Assim, pode-se começar com a análise no *software VOSviewer*.

Figura 9: Palavras-chave mais populares na Scopus.

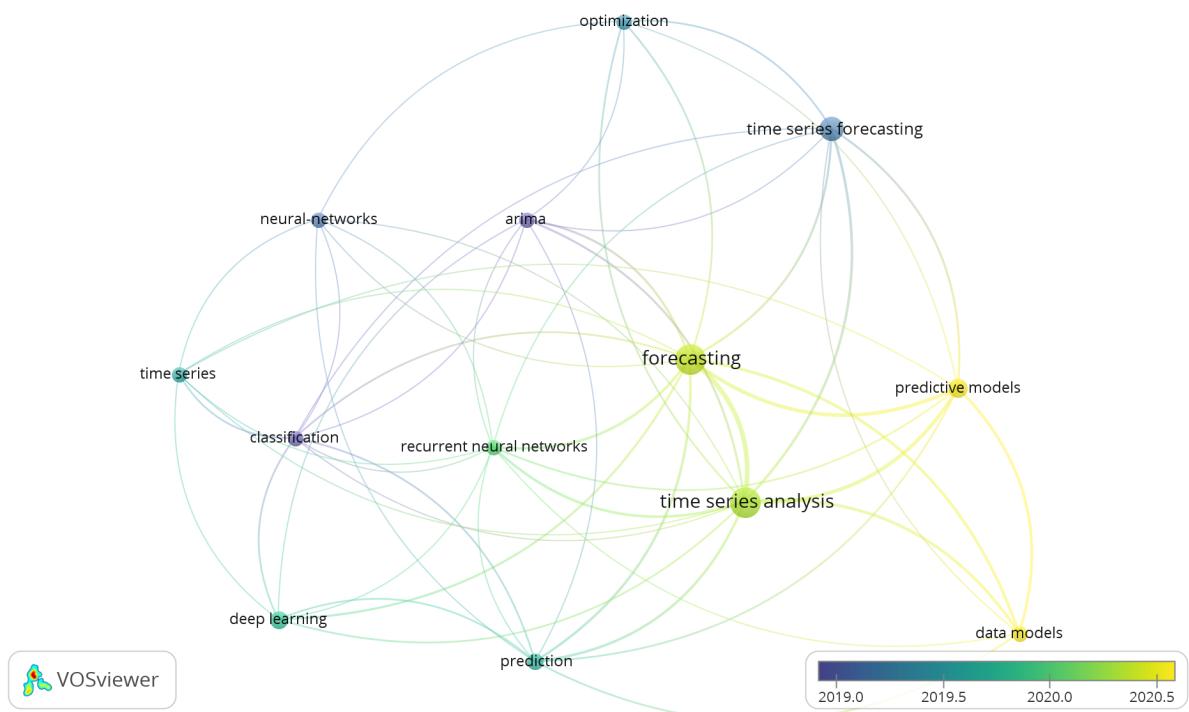


Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

Na figura 9 há uma lista das palavras mais usadas como sinônimos da palavra *time series analysis* ou juntas no corpo do texto dos artigos. A análise da base de dados no scopus foi feita na ferramenta que mostra as palavras-chave que podem ser relacionadas em cada campo de busca, com isto tem uma visão ampla do que pode ter correlação com as palavras-chave mãe da busca.

Na relação entre as palavras-chave neste primeiro momento, foi obtido um resultado de 3484 palavras-chave, 212 atingindo o limite, lembrando que as palavras base a partir das quais se deve chegar ao texto “*time series forecasting and time series analysis*” em Scopus.

Figura 10: Palavras-chave mais populares na WoS



Fonte: Elaboração própria a partir de dados da Web of Science (2018 a 2020)

Na Figura 10 a análise do banco de dados Web of Science foi feita na ferramenta que mostra as palavras-chave que estão relacionadas em cada campo de busca, com isto você pode ter uma visão ampla do que tem correlação com as palavras-chave mãe da busca.

Na relação entre as palavras-chave neste primeiro momento, teve um resultado de 305 palavras-chave, 13 atingem o limite, lembrando que as palavras base para o resultado foi “*time series forecasting and time series analysis*” na web of science.

O único banco de dados que não será mostrado aqui é o banco de dados Lente, porque é um excelente banco de dados e ainda não retornou muito na busca que foi feita. O site do lens retornou apenas 11 artigos com os filtros aplicados. Na **Etapa 1** é observado o campo de busca que foi usado nesta busca que deu apenas 11 artigos.

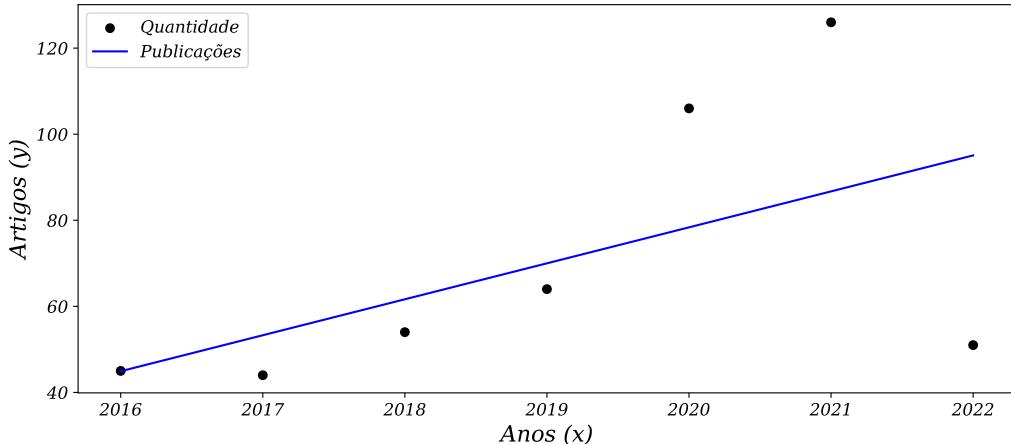
Na Tabela 1, ela lista as palavras-chave para cada base e aumenta a quantidade de artigo em todas as bases, mas esta tabela está com os dados brutos que não foram eliminados as duplicatas, portanto, usando o *software mendeley* para remover as duplicatas devolve apenas 308 artigos.

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	
Web of Science	nonlinear	AND	time series	8
	forecasting		forecasting	
Lens	time series	AND	time series	126
	forecasting		analysis	
nonlinear	AND	time series		14
	forecasting		forecasting	
Lens	time series	AND	time series	11
	forecasting		analysis	
		Total		649

Fonte: Elaboração própria

Figura 11: Analise das quantidades de artigos em relação aos anos.



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

A figura 11 tem com abcissas e ordenadas anos e artigos, assim a relação entre a data de publicação dos artigos ao longo do tempo.

Um número considerável de artigos para analisar na Figura 11 uma análise foi feita com base em uma regressão linear dos artigos ao longo dos anos de 2016 a 2022, nesta análise obteve a seguinte equação de regressão linear:

$$y(x) = 8,3571x - 16803 \quad \text{Com } R^2 = 0,3062 \quad (4)$$

Com  $y(x)$  a equação da reta na equação (4). 8,3571 é o coeficiente angular do gráfico de  $y(x)$ , 16,803 é o coeficiente linear, ou o ponto de intersecção com o eixo  $y$ ,  $x$  é a variável independente.

Este coeficiente indica a proporção da variância da variável dependente que pode ser atribuída estatisticamente ao conhecimento de uma ou mais variáveis independentes Quinino, Reis e Bessegato (1991).

O coeficiente de determinação mede a relação que existe entre a variável dependente e as variáveis independentes, indicando que porcentagem da variação explicada pela regressão representa da variação total. Quando:

$R^2 = 1$ : todos os pontos observados estão exatamente na reta de regressão (ajuste perfeito), ou seja, as variações de  $y$  são de 100% explicadas pela variação de  $x_n$  através da função especificada, sem desvios em torno da função estimada.

$R^2 = 0$ : conclui-se que as variações de  $y$  são exclusivamente aleatórias e a introdução das variáveis  $x_n$  no modelo não incorporará nenhuma informação sobre as variações de  $y$ .

$$R^2 = \frac{\left( \sum X \cdot Y - \frac{\sum X \cdot \sum Y}{n} \right)^2}{\left[ \sum X^2 - \frac{(\sum X)^2}{n} \right] \cdot \left[ \sum Y^2 - \frac{(\sum Y)^2}{n} \right]} = (r)^2 \quad (5)$$

Na equação (5)  $X, Y$  é dado pelas coordenadas no plano cartesiano como por exemplo o par encomendado  $(x, y)$ . Na equação (4) observa-se que obteve o  $R^2 = 30\%$  isto implica que a linha de regressão será influenciada pelo  $R^2$  que foi encontrado.

Embora seja uma análise muito simples que foi realizada com a relação entre número de artigos e anos, ainda é uma validação muito boa para olhar para o teste F de significância que é dado o significado tem que ser sempre  $F < 5\%$  este teste também é chamado de p-valor.

Tendo estes valores, é possível analisar o significado extremo da linha de regressão e observar que 2021 foi o ano em que a maioria dos artigos foram publicados sobre este tema da séries temporais.

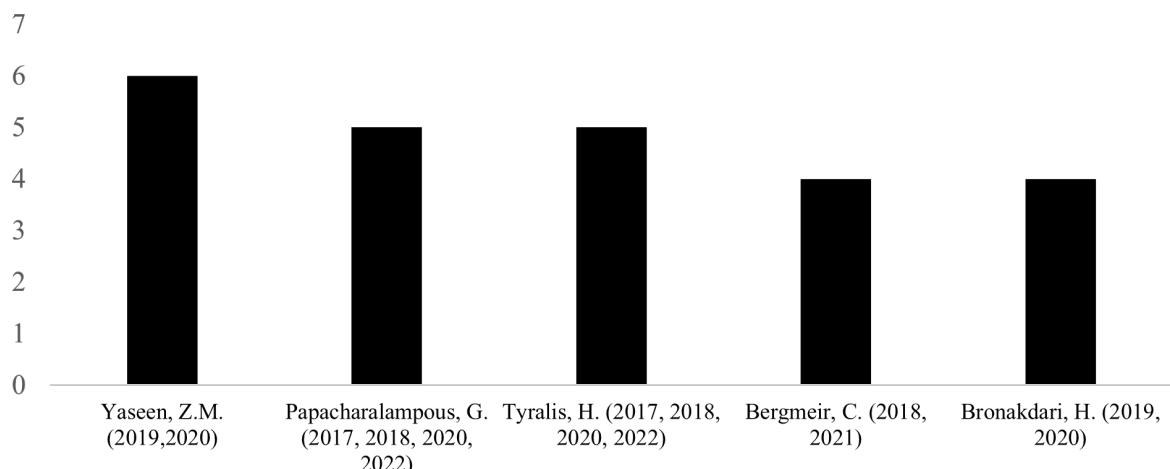
Tabela 2: Fator de impacto

Revista científica	Quantidade de publicação	Qualidade da revista	H-INDEX
Neurocomputing	27	Q1	143
IEEE Access	18	Q1	127
Applied Soft Computing	12	Q1	143
Energies	11	Q2	93
Energy	11	Q1	343

Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

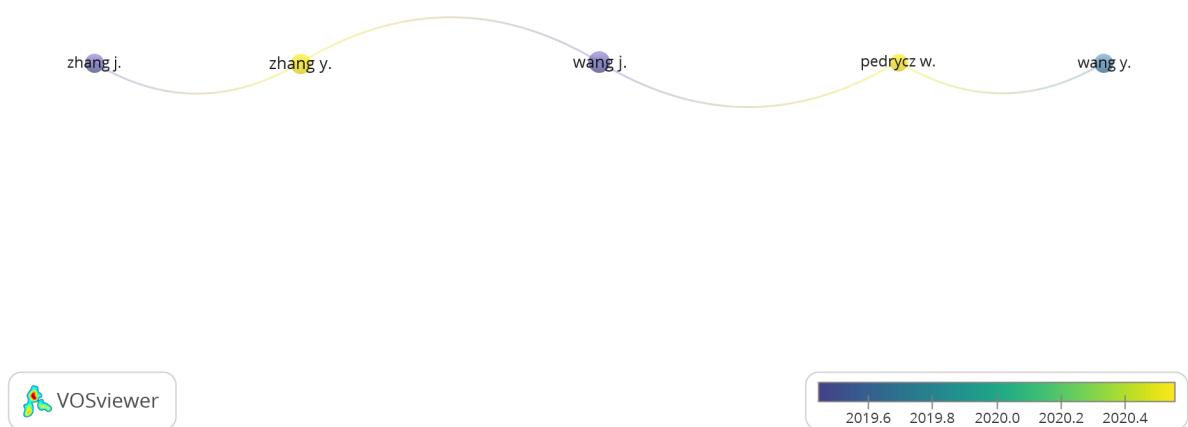
Na Tabela 2 mostra revistas que a maioria publica artigos sobre este assunto, já que muitas revistas não estão localizadas no Brasil e têm seus nomes em inglês, mas todas as revistas com um fator de impacto muito alto como **A1** têm uma correlação com as áreas de **informática, engenharia e matemática**.

Figura 12: Relação de autores entre artigos publicados



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

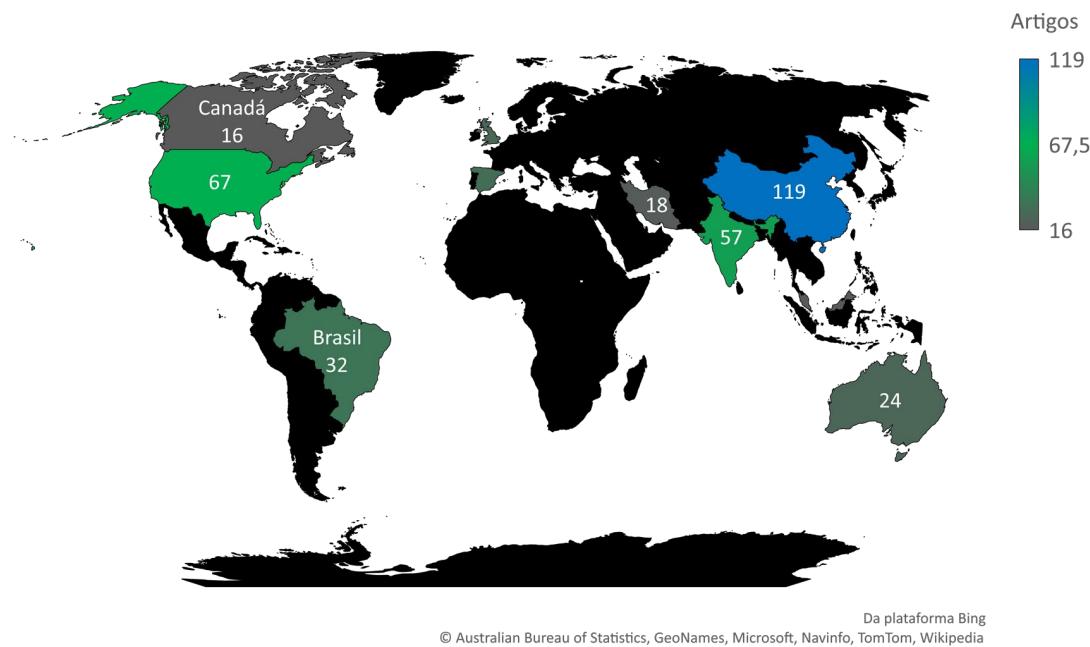
Figura 13: Ligação bibliográfica entre os autores



Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

Respondendo a um problema aqui colocado, o Q 1 usa a Figura 12 com um gráfico de histograma para ser mais visível que os autores que mais publicaram neste tópico no gráfico colocam os autores que tiveram publicação maior que 4 e com isto não coloca todos os autores considerando os autores que publicaram mais de 4 artigos neste tópico de 2016 a 2022.

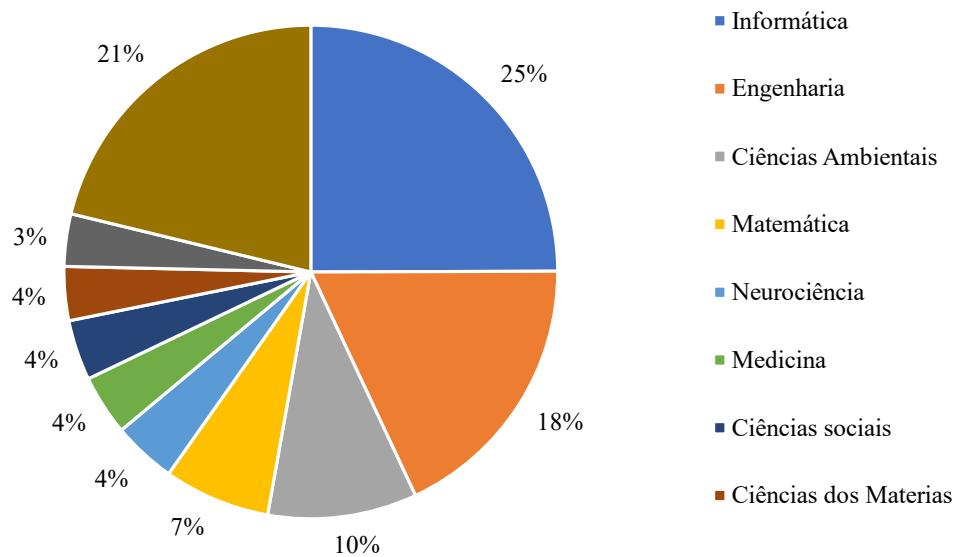
Figura 14: Mapa mundial da publicação de artigos em todo o mundo



Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

Na pergunta de pesquisa **Q 2** é respondido com a Figura 14 os países que mais público sobre o assunto em escala desde a maior publicação até a menor em escala, como segue China - 119, Estados Unidos - 67, Índia - 57, Brasil - 32, Espanha - 28, Reino Unido - 25, Austrália - 24, Irã - 18, Malásia - 17, Canadá - 16. O mapa não mostra todos os países com seus números de publicação.

Figura 15: Áreas de aplicação do tema



Fonte: Elaboração própria a partir de dados da Scopus, Lens e Web of Science (2016 a 2022)

Pergunta de pesquisa **Q 3** para responder a esta pergunta, foi feito um gráfico circular para rotular as áreas que têm mais publicação no tempo escolhido na revisão. A Tabela 3 mostra os valores de cada área e sua quantidade de publicação.

Tabela 3: Áreas 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 Materiais	34
Negócios, Gestão e Contabilidade	33
Outros	204

Fonte: Elaboração própria a partir de dados da Scopus, len e Web of Sicence (2016 a 2022)

Na última pergunta de pesquisa **Q 4** foi feita uma pesquisa sobre alguns dos artigos

mais influentes na revisão, esses artigos retratam alguns métodos sobre séries temporais dos artigos dos autores Golyandina (2020), Kumar, Jain e Singh (2021), Xie et al. (2019), Lara-Benitez, Carranza-Garcia e Riquelme (2021), Ahmad et al. (2018), Carvalho Jr. e Costa Jr. (2019), Tan et al. (2021), Liu e Chen (2019), Liu et al. (2021), Rossi (2018), Soyer e Zhang (), Martinović, Hunjet e Turcin (2020), Ursu e Pereau (2016), Wang et al. (2016), Shih, Sun e Lee (2019), Moon et al. (2019), Chou e Tran (2018), Bergmeir, Hyndman e Koo (2018), Boroojeni et al. (2017), Chou e Nguyen (2018), Coelho et al. (2017), Du et al. (2020), Sadaei et al. (2019), Salgotra, Gandomi e Gandomi (2020), Tyralis e Papacharalampous (2017), Vlachas et al. (2020), Yang et al. (2019), Shen et al. (2020), Sezer, Gudelek e Ozbayoglu (2020), Chen et al. (2018), Buyuksahin e Ertekin (2019), Li e Bastos (2020), Kulshreshtha e Vijayalakshmi (2020), Samanta et al. (2020), Xu et al. (2019), Graff et al. (2017), Taieb e Atiya (2016) alguns métodos usados pelos autores para previsão de séries temporais e alguns modelos de análise da mesma previsão não-linear.

Xu et al. (2019) no modelo híbrido, o modelo linear AR e LR ou o modelo ARIMA e o modelo DBN não-linear são explorados para capturar os comportamentos lineares e não-lineares de uma série temporal, respectivamente. Li e Bastos (2020) o desempenho de previsão da abordagem MAELS é comparado com os predecessores baseados na aprendizagem de máquinas de última geração como CNN, RNN, LSTM, ARIMA, e SVM-VAR. As abordagens, CNN, RNN e LSTM permitem o manuseio multivariado de entrada e saída, ARIMA usa dados passados para prever o futuro usando duas características principais: autocorrelação e médias móveis.

Assim, com esta revisão sistemática e análise de conteúdo, a resposta à pergunta de pesquisa feita no início do capítulo foi obtida.

Fora destes modelos, também há a atualização ARIMA que será utilizada nesta dissertação, pois SARIMA, SARIMAX ambos os modelos serão comparados para se obter o melhor modelo entre eles, fora disto também será utilizado o Light GBM e o XGBoost. Para as métricas de erro nesta dissertação serão utilizadas as seguintes métricas e explicadas no capítulo 3 MAE, MAPE e RMSE na literatura é um dos mais utilizados entre vários com, por exemplo, os  $R^2$  citados (5) para as previsões futuras sempre foram confrontados com estas métricas de erro. os  $R^2$  não são tão utilizados para comparação.

## 2.6 Conclusão da revisão

Nesta seção para relatar o que foi coberto durante a pesquisa de revisão sendo ela em algumas bases como Scopus, Web of Science e Lens, cada base retornou vários artigos

que foram analisados e assim responde a pergunta de pesquisa feita na revisão apesar de a base de Lens ser a menor de todas ainda encontra alguns artigos que foram relevantes no processo de dissertação também com a ajuda do *software* para analisar muitos arquivos e suas relações entre cada um. Sendo a série temporal uma análise mais profunda e mais atual na revisão sistemática fazendo a pesquisa nos últimos 6 anos.

Na busca realizada foram obtidos alguns resultados muito relevantes, pois no cruzamento das palavras em cada base com o filtro aplicado foram obtidos artigos de 308 de 2016 a 2022, com isso foi necessário filtrar mais sobre cada área de desempenho dos artigos, tais como matemática, engenharia e informática neste filtro tinha um total de artigos de 481 excluindo que seriam das outras áreas.

### 3 Base Teórica

Para um bom resultado, uma base sólida é fundamental. Este capítulo cobre métricas de erro e modelos regressivos de previsão de modelos, entre outros.

#### 3.1 Métricas de Erros

A métrica MSE é uma das mais utilizadas em aprendizado de máquina. Seu cálculo é feito da seguinte forma:

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2 \quad (6)$$

MSE é a média da somatória do erro ao quadrado. Subtraímos o que aconteceu,  $y_i$ , do valor que foi projetado,  $\hat{y}_i$ . O resultado é o cálculo do erro. Ao elevarmos o erro ao quadrado, estamos evitando que os erros fiquem negativos e, portanto, se subtraiam na somatória.

#### RMSE

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (7)$$

A vantagem de utilizarmos o RMSE é que, ao computar a raiz quadrada, o erro passa a ter a mesma escala do indicador que estamos trabalhando. Um RMSE baixo, significa que a performance do modelo foi boa, pois o erro se aproxima de zero.

#### MAE

O MAE é calculado usando o módulo da subtração, obtida entre o valor do que realmente aconteceu e o valor projetado (previsto) e dividi tudo pelo número  $n$  de amostras. Com isso, obtém o erro médio absoluto. Equação do MAE:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (8)$$

Sua interpretação é comparável ao RMSE, onde o erro se dá no mesma escala/ordem de grandeza da variável estudada.

Não é possível dizer se o MAE é um indicador melhor ou pior que os dois anteriores.

### MAPE

Conhecido como MAPE, é a porcentagem relativa ao valor observado. O cálculo é feito obtendo a somatória da diferença entre o valor que realmente ocorreu com o valor previsto (resultado do erro), dividido pelo valor observado.

$$MAPE = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (9)$$

O problema é quando o valor observado  $y_i$  é igual a 0, pois é matematicamente impossível dividir por 0. Sendo uma medição de erro, porcentagens menores são melhores.

Se fizer  $1 - MAPE$ , tem a porcentagem de acerto.

## 3.2 ARIMA, SARIMA e SARIMAX

A previsão da série temporal é um problema difícil sem resposta fácil. Existem inúmeros modelos estatísticos que afirmam superar uns aos outros, mas nunca está claro qual modelo é o melhor.

Dito isto, modelos baseados em ARMA são muitas vezes um bom modelo para começar. Eles podem alcançar pontuações decentes na maioria dos problemas de séries temporais e são bem adequados como um modelo de linha de base em qualquer problema de séries temporais.

O modelo ARIMA, vamos dividi-lo em AR, I e MA.

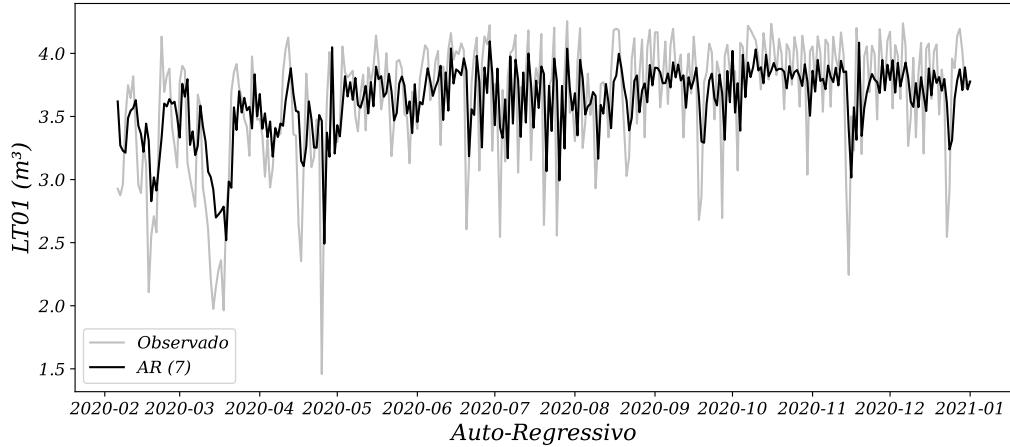
### 3.2.1 Componente auto-regressivo – AR( $p$ )

O componente auto regressivo do modelo ARIMA é representado por AR( $p$ ), com o parâmetro  $p$  determinando o número de séries defasadas que é utilizado.

$$Y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \varepsilon_t \quad (10)$$

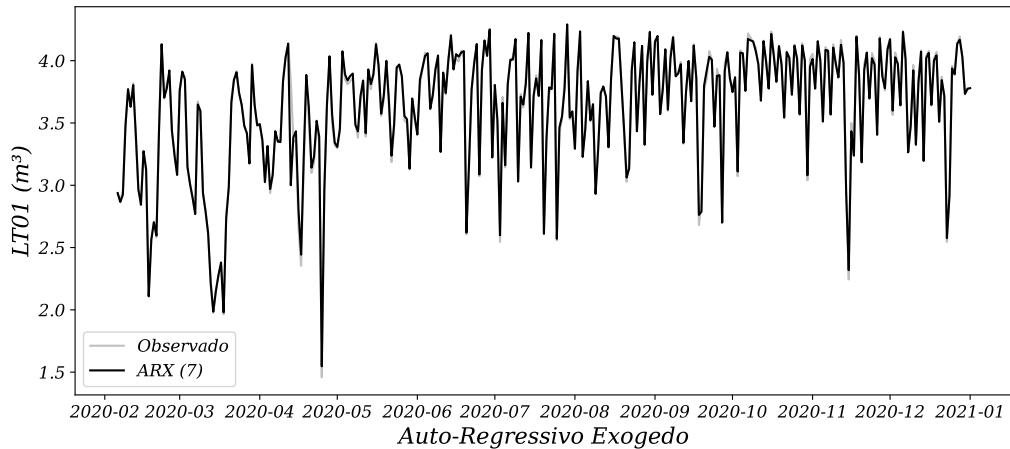
Dos dados pode ser obtido a seguinte previsão no modelo AR(7)

Figura 16: Modelo AR(7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 17: ARX (7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Onde em (10) o  $\varepsilon_t$  é ruído branco. Isso é como uma regressão múltipla, mas com valores defasados de  $y_t$  como preditores. é referido a isso como um AR( $p$ ) modelo, um modelo auto regressivo de ordem  $p$

Da Figura 16, tem como objetivo mostrar uma previsão de um passo a frente (um dia) nos apêndices C pode ver uma comparação dos AR, MA e o ARX

O modelo ARX é um modelo similar ao AR só coloca as variáveis exógenas do conjunto de dados para melhorar a previsão futura.

O modelo AR pode ser visivelmente um modelo adequado para a previsão que está sendo feito, mas como é um modelo auto regressivo ainda assim com o passar do tempo e

da previsão ele vai prever de uma forma linear e não convencional, para um analise mais rápido da série pode se considerar um modelo adequado. Logo mais adiante tem exemplos de casos gerais que pode ocorrer nesse método.

### **AR(0): Ruído branco**

Se definir o parâmetro  $p$  como zero (AR(0)), sem termos autorregressivos. Esta série de tempo é apenas um ruído branco. Cada ponto de dados é amostrado a partir de uma distribuição com uma média de 0 e uma variância de sigma-quadrado. Isso resulta em uma sequência de números aleatórios que não podem ser previstos. Isso é realmente útil, pois pode servir como uma hipótese nula, e proteger nossas análises de aceitar padrões falso-positivos.

### **AR(1): Caminhadas aleatórias e Oscilações**

Com o parâmetro  $p$  definido para 1, vai levar em conta o medidor de tempo anterior ajustado por um multiplicador  $e$ , em seguida, adicionando ruído branco. Se o multiplicador é 0, então temos ruído branco, e se o multiplicador é 1, teremos uma caminhada aleatória. Se o multiplicador estiver entre  $0 < \alpha < 1$ , então a série temporal exibirá reversão média. Isso significa que os valores tendem a pairar em torno de 0 e reverter para a média depois de regredir a partir dele.

### **AR( $p$ ): Termos de ordem superior**

Aumentar ainda mais o parâmetro  $p$  significa apenas ir mais para trás e adicionar mais medidores de tempo ajustados por seus próprios multiplicadores. Pode ir o mais longe que poder, mas à medida que aproxima é mais provável que usa parâmetros adicionais, como a média móvel (MA( $q$ )).

#### **3.2.2 Média Móvel – MA( $q$ )**

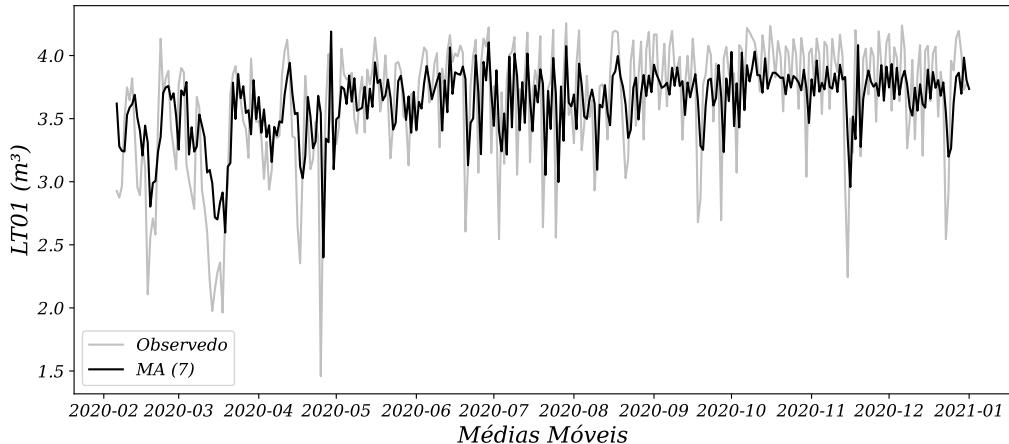
Este componente não é uma média de rolamento, mas sim os atrasos no ruído branco. Trenberth (1984)

MA( $q$ ) é o modelo de média móvel e  $q$  é o número de termos de erro de previsão defasados na previsão. Em um modelo MA(1), na previsão é um termo constante mais o termo de ruído branco anterior vezes um multiplicador, adicionado com o termo de ruído branco atual. Esta é apenas simples probabilidade mais estatísticas, pois estamos ajustando nossa previsão com base em termos anteriores de ruído branco.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots + \theta_q \varepsilon_{t-q} \quad (11)$$

De (11) onde  $\varepsilon_t$  é ruído branco. Refere nos a isto como um modelo de  $MA(q)$ , um modelo de ordem média móvel  $q$ . Claro que não observamos os valores de  $\varepsilon_t$ , por isso não é realmente uma regressão no sentido habitual.

Figura 18: Modelo MA(7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

O modelo MA com o mesmo valor do AR para comparação e torna o modelo mais fácil de ser previsto. Como observado na Figura 18 a previsão graficamente é parecido com o modelo da Figura 16, só não se compara com a Figura 17, perceba que esse modelo aparente prever perfeitamente o tempo que foi listado.

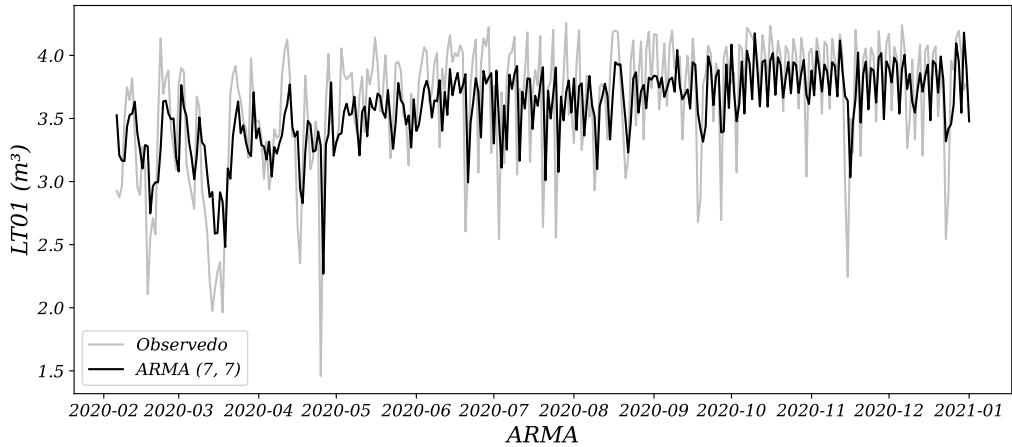
### 3.2.3 Modelos ARMA e ARIMA

As arquiteturas ARMA e ARIMA são apenas os componentes AR (Autoregressive) e MA (Moving Average) juntos.

#### **ARMA**

O modelo ARMA é uma constante mais a soma de lags AR e seus multiplicadores, além da soma dos lags ma e seus multiplicadores mais ruído branco. Esta equação é a base de todos os modelos que vêm a seguir e é uma estrutura para muitos modelos de previsão em diferentes domínios.

Figura 19: ARMA (7,7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Da Figura 19 é a junção dos modelos AR e MA esse modelos juntos pode ocorrer a redução do erro em escala mais significativa, nos apêndice A e B pode ser notado a comparação de alguns passos a mais do que mostrado aqui.

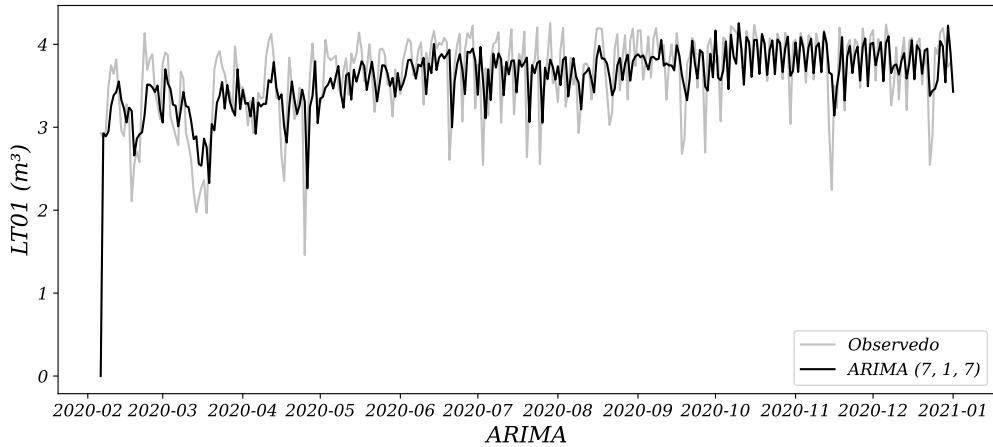
## ARIMA

$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \dots + \omega_q \varepsilon_{t-q} + \varepsilon_t \quad (12)$$

Onde em (12) o  $Y_t$  é a série diferenciada (pode ter sido diferente mais de uma vez). Os “preditores” no lado direito incluem ambos os valores defasados de  $Y_t$  e erros defasados. Chamamos isso de ARIMA( $p, d, q$ ).

O modelo ARIMA é um modelo ARMA ainda com uma etapa de pré-processamento incluída no modelo que representamos usando I(d). I(d) é a ordem de diferença, que é o número de transformações necessárias para tornar os dados estacionários. Assim, um modelo ARIMA é simplesmente um modelo ARMA na série de tempo diferente.

Figura 20: ARIMA (7,1,7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Olhando a Figura 20 podemos perceber que não tem muita diferença visual com os outros métodos mostrados até agora, visualmente o método ARX ainda está melhor que os outros.

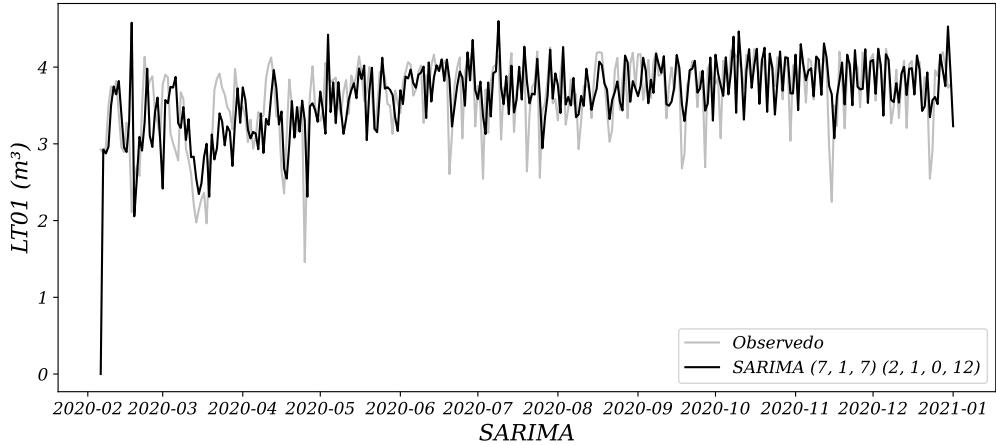
### 3.2.4 Modelos SARIMA, ARIMAX e SARIMAX

Os modelos ARIMA são ótimos, mas incluir variáveis sazonais e exógenas no modelo pode ser muito poderoso. Como o modelo ARIMA assume que a série temporal é estacionária, precisamos usar um modelo diferente. **SARIMA**

$$Y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^P \phi_n y_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (13)$$

O modelo é muito semelhante ao modelo ARIMA, com um conjunto adicional de componentes autorregressivos e de média móvel. O atraso extra é compensado pela frequência sazonal (por exemplo, 12 - mensal, 24 - por hora).

Figura 21: SARIMA  $(7, 1, 7)(2, 1, 1)_{12}$



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 21 pode ser observado como a previsão em vermelho esta mais próxima do observado em preto, só acionando o termo de sazonalidade na previsão.

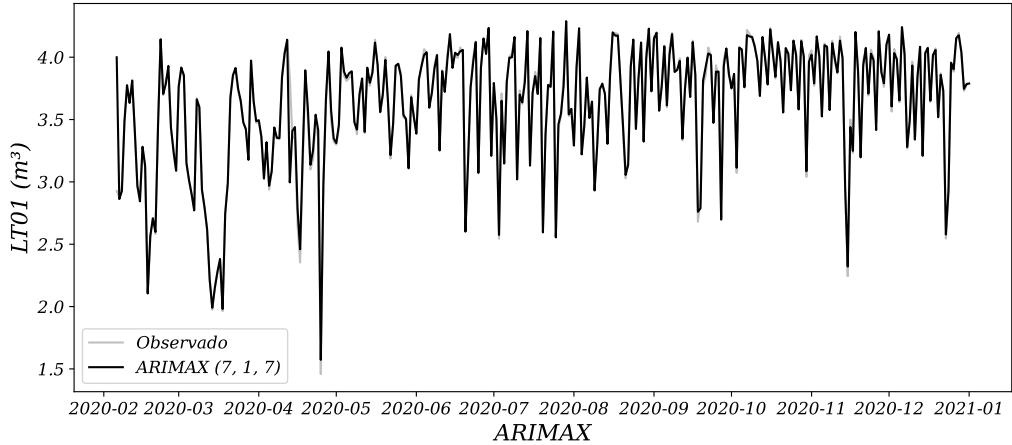
Os modelos SARIMA permitem diferenciar dados por frequências sazonais e não sazonais. Uma estrutura de pesquisa automatizada de parâmetros, como pmdarina, pode ajudar a entender quais são os melhores parâmetros.

### ARIMAX e SARIMAX

$$d_t = c + \sum_{n=1}^p \alpha_n d_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^r \beta_n x_{nt} + \sum_{n=1}^P \phi_n d_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (14)$$

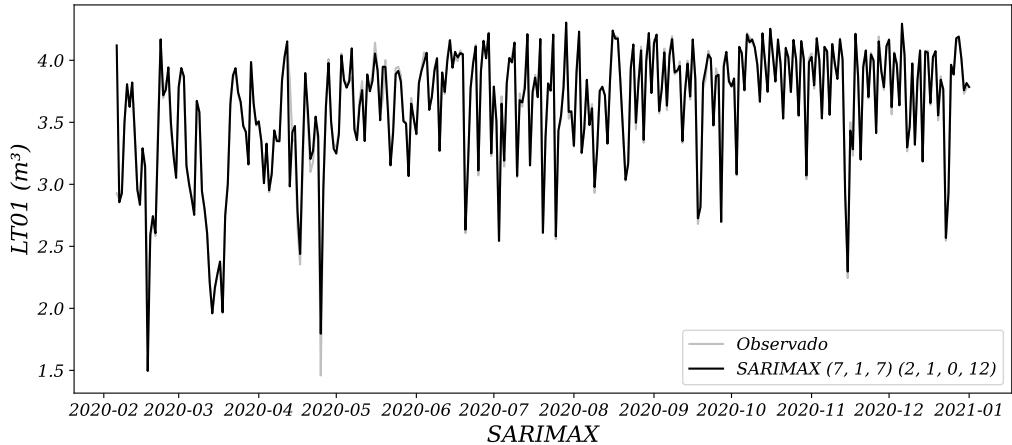
Em (14) está o modelo SARIMAX. Este modelo tem em conta variáveis exógenas, ou por outras palavras, utiliza dados externos na nossa previsão. É interessante pensar que todos os factores exógenos ainda são tecnicamente indiretamente modelados na previsão histórica do modelo. Dito isto, se incluirmos dados externos, o modelo responderá muito mais rapidamente ao seu efeito do que se confia na influência de termos desfasados.

Figura 22: ARIMAX (7, 1, 7)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 23: SARIMAX (7, 1, 7)(2, 1, 1)<sub>12</sub>



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Entre os modelos com variáveis exógenas os modelos da Figura 22 e 23, é possível perceber que a previsão está mais completa do que nos modelos sem a variável exógena.

### 3.3 Modelos Regressivo

#### 3.3.1 Regressão Linear (LR)

Segundo Korstanje (2021) nos modelos de aprendizado de máquina supervisionados, você tenta identificar relações entre diferentes variáveis:

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

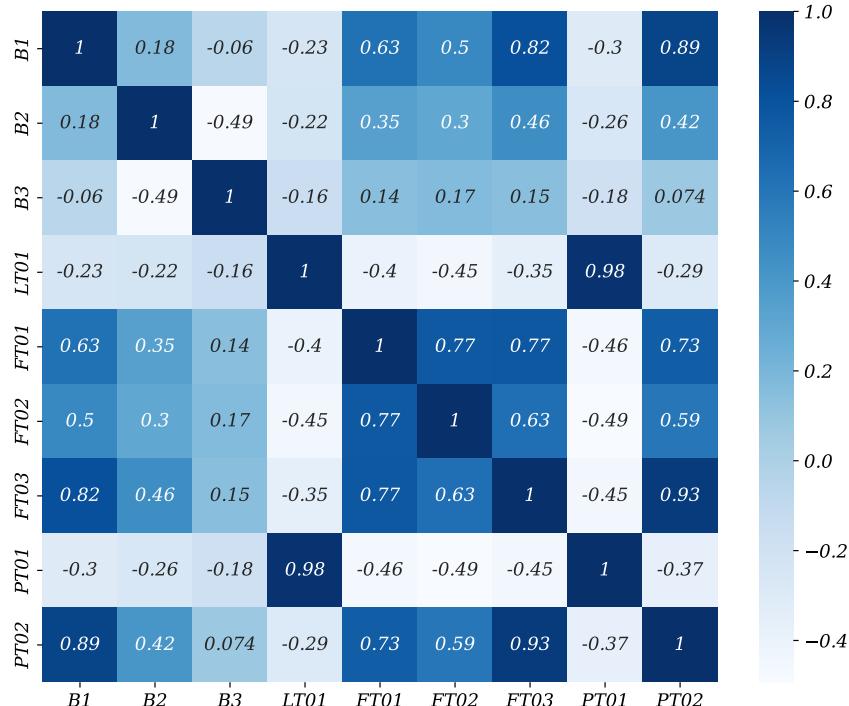
Para a previsão, é importante entender quais tipos de variáveis explicativas você pode ou não usar. Como exemplo, aqui vai ser usado as variáveis **Pressão de Succção (PT01SU)** como variável  $x$  e **Nível do Reservatório (Câmara 1) LT01** como variável  $y$  pois na correlação de Pearson mostrado na Figura 24, o coeficiente mostra a relação que tem entre o eixo  $x$  e  $y$  com a seguinte fórmula.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}} \quad (15)$$

De (15) sejam  $x_i \in y_i$  os valores das variáveis  $X$  e  $Y$ .  $\bar{x}$  e  $\bar{y}$  são respectivamente as médias dos valores  $x_i \in y_i$ .

A fórmula do coeficiente de correlação de Pearson é então,

Figura 24: Corelação de Pearson



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Como mostra a Figura 24 essa imagem é meramente ilustração da correlação que tem relação no conjunto de dados que está sendo trabalhado aqui. E com isso também pode ser respondido a Q 1 da pesquisa. porque a correlação entre essas variáveis é forte.

### Definição do modelo

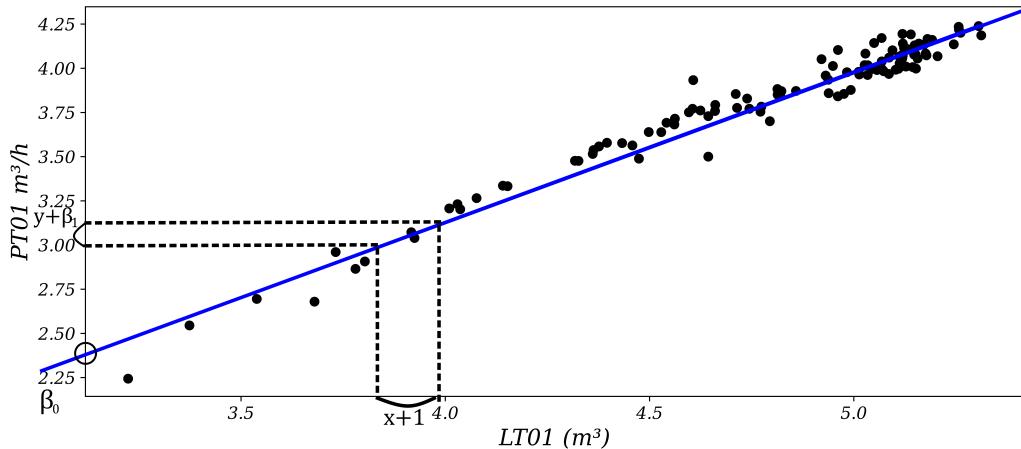
A regressão linear é definida da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (16)$$

Da (16) têm a seguinte variáveis:

- Há  $p$  variáveis explicativas, chamadas  $x$ .
- Existe uma variável alvo chamada  $y$ .
- O valor para  $y$  é calculado como uma constante ( $\beta_0$ ) mais os valores do  $x$  variáveis multiplicadas pelos seus coeficientes  $\beta_1$  para  $\beta_p$ .

Figura 25: Regressão linear LT01 vs PT01 correlação 98%



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

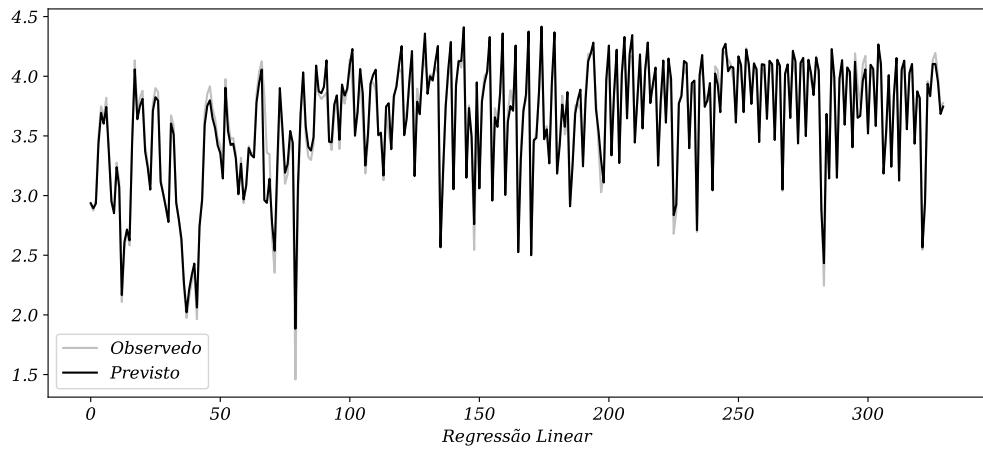
A Figura 25 mostra como interpretar  $\beta_0$  e  $\beta_1$  visualmente. Mostra que para um aumento de 1 na variável  $x$ , o aumento na variável  $y$  representa  $\beta_1$ . O valor para 0 é o valor para  $x$  quando  $y$  é 0.

Para poder utilizar a regressão linear, é necessário estimar os coeficientes (betas) sobre um conjunto de dados de formaçāo. Os coeficientes podem então ser estimados utilizando a seguinte fórmula, em notação matricial:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (17)$$

Korstanje (2021) esta fórmula é conhecida como **OLS**: o método dos mínimos quadrados ordinários (Ordinary Least Squares method). Este modelo é muito rápido para caber, uma vez que requer apenas cálculos matriciais para calcular os betas. Embora fácil para caber, é menos adequado para processos mais complexos. Afinal de contas, é um modelo linear, e pode portanto, só se encaixam em processos lineares.

Figura 26: Regressão linear (LR) um passo a frente

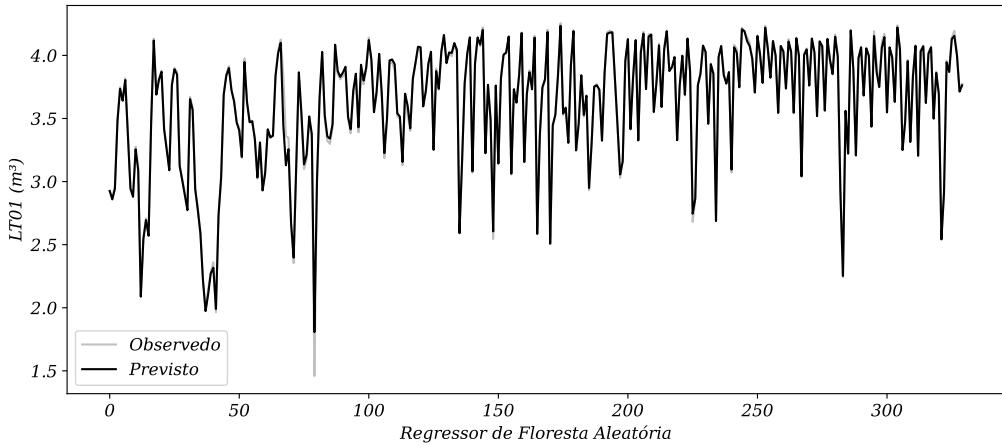


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

### 3.3.2 Floresta Aleatória

Pode entender que ter exatamente a mesma árvore de decisão 1000 vezes não tem valor agregado do que usar essa árvore de decisão apenas uma vez. Em um modelo de conjunto, cada modelo individual deve ser ligeiramente diferente do outro. Existem dois métodos bem conhecidos de criação de coleções: ensacamento e reforço. Random Forest usa ensacamento para criar um conjunto de árvores de decisão

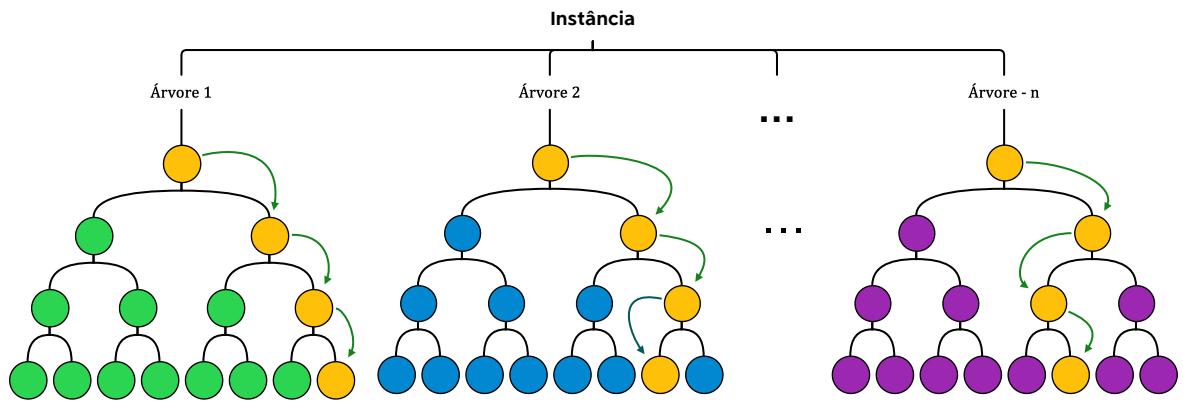
Figura 27: Regressão da Floresta Aleatória (RFA)



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Segundo Pelletier et al. (2016) Cada árvore é construída executando um algoritmo de aprendizado individual que divide o conjunto de variáveis de entrada em subconjuntos com base em um teste de valor de atributo (por exemplo, o coeficiente de Gini). Ao contrário das árvores de decisão (DT) clássicas, as árvores de RFA são construídas sem poda e selecionando aleatoriamente em cada nó um subconjunto de variáveis de entrada. Atualmente, esse número de variáveis utilizadas para dividir um nó de RFA (denotado por  $m$ ) corresponde à raiz quadrada do número de variáveis de entrada.

Figura 28: Esquema da Floresta Aleatória



Fonte: Elaboração própria

### 3.3.3 LightGBM e XGboost

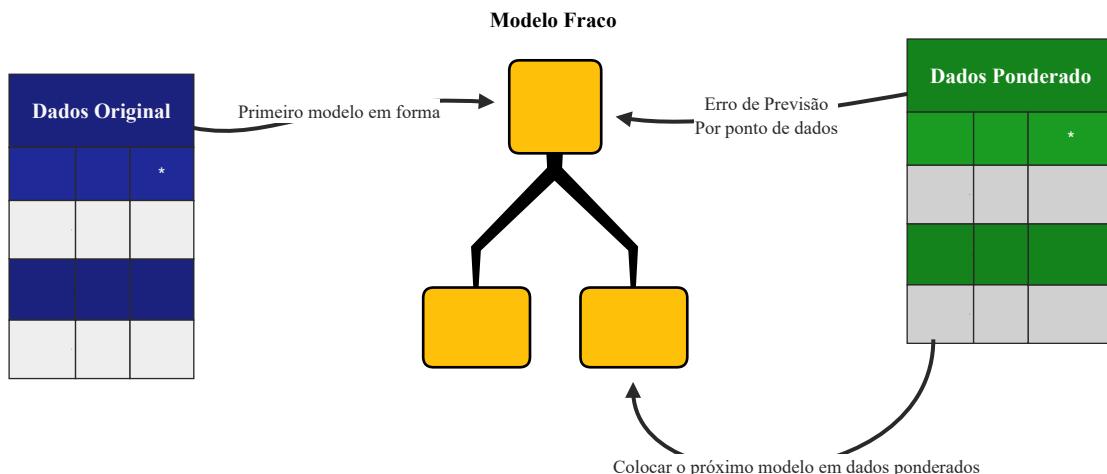
O aumento de gradiente combina vários pequenos modelos de árvore de decisão para fazer previsões. É claro que essas pequenas árvores de decisão são diferentes umas

das outras, caso contrário não há vantagem em usar mais árvores de decisão. O conceito importante a ser entendido aqui é como essas árvores de decisão se tornam diferentesumas das outras. Isto é conseguido através de um processo chamado elevação. Boosting e bagging são dois métodos principais que são aprendidos juntos. Boosting é um processo iterativo. Ele adiciona cada vez mais modelos fracos ao conjunto de modelos de maneira inteligente. Em cada etapa, pontos de dados individuais são ponderados.

Pontos de dados que já estão bem previstos não são importantes para o aluno adicionar. Portanto, novos modelos fracos se concentrarão em aprender coisas que ainda não são compreendidas, melhorando assim o conjunto.

Pode se ver uma visão geral esquemática do processo de reforço na Figura 29. Com essa abordagem, você ajusta iterativamente modelos fracos que se concentram nas partes dos dados que ainda não são compreendidas. Ao fazer isso, você mantém todos os modelos fracos intermediários. O modelo ensemble é a combinação de todos esses modelos fracos.

Figura 29: Impulsionando gradiente com XGBoost e LightGBM



Fonte: Adaptação de Korstanje (2021)

### 3.3.4 O Gradiente em Gradiente de Boosting (Reforço)

Korstanje (2021) esse processo iterativo é chamado de aumento de gradiente por um motivo. Um gradiente é um termo matemático que se refere ao campo vetorial de derivadas parciais que apontam na direção da inclinação mais acentuada. Em termos simples, muitas vezes comparamos gradientes com declives de estradas em aclive: quanto maior a inclinação, mais íngreme a colina. Os gradientes são calculados tomando derivadas, ou derivadas parciais, de uma função.

No aumento de gradiente, ao adicionar árvores adicionais ao modelo, o objetivo é

adicionar uma árvore que melhor explique a variação que não foi explicada pelas árvores anteriores. O destino de sua nova árvore é, portanto.

$$y - \hat{y} \quad (18)$$

De (18) isso pode ser denotado reescrito como a derivada parcial negativa da função de perda em relação às previsões de  $y$ :

$$y - \hat{y} = -\frac{\partial L}{\partial \hat{y}} \quad (19)$$

Você define isso como o destino da nova árvore para garantir que a adição da árvore explicará uma quantidade máxima de variação adicional no modelo geral de aumento de gradiente. Isso explica por que o modelo é chamado de aumento de gradiente boosting.

### 3.3.5 Algoritmos de boosting de gradiente

Existem muitos algoritmos que executam versões ligeiramente diferentes de aumento de gradiente. Quando o método de aumento de gradiente foi inventado, o algoritmo não Muito desempenho, mas mudou com o advento do algoritmo AdaBoost: o primeiro algoritmo que pode se adaptar a modelos fracos.

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

O **XGBoost** é um dos algoritmos de aprendizado de máquina mais usados. O XGBoost é uma maneira rápida de obter bons desempenhos. Como é fácil de usar e tem alto desempenho, é o primeiro algoritmo para muitos profissionais de aprendizado de máquina.

**LightGBM** é outro algoritmo de aumento de gradiente que é importante conhecer. No momento, é um pouco menos difundido que o XGBoost, mas está ganhando popularidade seriamente. A vantagem esperada do LightGBM sobre o XGBoost é um ganho de velocidade e uso de memória. Nesta subseção, você descobrirá as implementações de

ambos os algoritmos de aumento de gradiente.

### 3.3.6 A diferença entre XGBoost e LightGBM

Se você for usar esses dois algoritmos de aumento de gradiente, é importante entender de que maneira eles diferem. Isso também pode fornecer uma visão dos tipos de diferença que fazem um número tão grande de modelos no mercado.

É importante entender se você planeja usar os dois algoritmos de aumento de gradiente. Como eles são diferentes. Isso também fornece informações sobre as várias diferenças que acompanham tantos modelos no mercado.

A diferença aqui é a forma como eles identificam as melhores divisões entre os azarões. (árvores de decisão individuais). Lembre-se de que uma divisão em uma árvore de decisão é quando sua árvore precisa encontrar a divisão que mais melhora seu modelo.

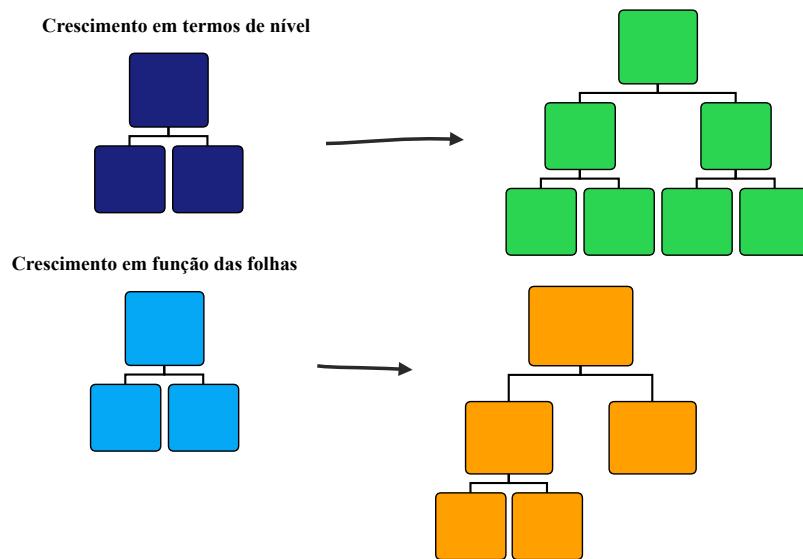
A ideia intuitiva e mais simples para encontrar a melhor divisão é iterar todos os ajustes possíveis e encontrar a melhor divisão. No entanto, isso leva muito tempo e algoritmos recentes apresentam alternativas melhores. Uma alternativa proposta pelo XGBoost é usar a segmentação baseada em histograma. Nesse caso, ao invés de iterar sobre todas as partições possíveis, o modelo constrói um histograma de cada partição. variáveis e use-as para encontrar a melhor divisão de variáveis. A melhor divisão geral é então mantida.

LightGBM foi inventado pela Microsoft e tem uma maneira mais eficiente de definir partições. Essa abordagem é chamada de amostragem unilateral baseada em gradiente (GOSS). O GOSS calcula o gradiente de cada ponto de dados e o usa para filtrar pontos de dados com gradientes baixos. Afinal, pontos de dados com gradientes baixos já são bem compreendidos, enquanto indivíduos com gradientes altos precisam ser melhor aprendidos.

O LightGBM também usa uma abordagem chamada Exclusive Feature Bundling (EFB), que permite acelerar a seleção de muitas variáveis correlacionadas. Outra diferença é que o modelo LightGBM é adequado para crescimento de folhas (preferencialmente preferido), enquanto o XGBoost cultiva árvores como árvores. A diferença pode ser vista na Figura 30.

Essa diferença é um recurso que teoricamente favoreceria o LightGBM em termos de precisão, mas apresenta um risco maior de overfitting (sobreajuste) no caso de poucos dados disponíveis.

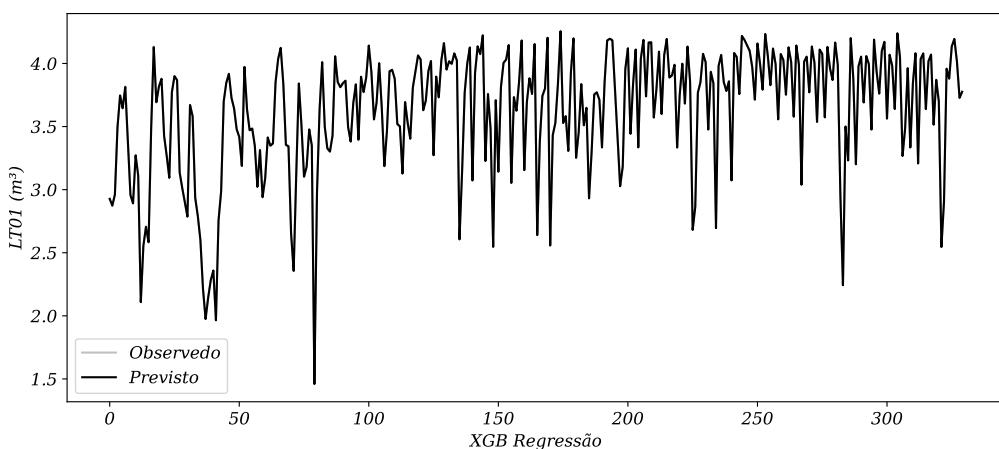
Figura 30: Crescimento em folha versus crescimento em nível

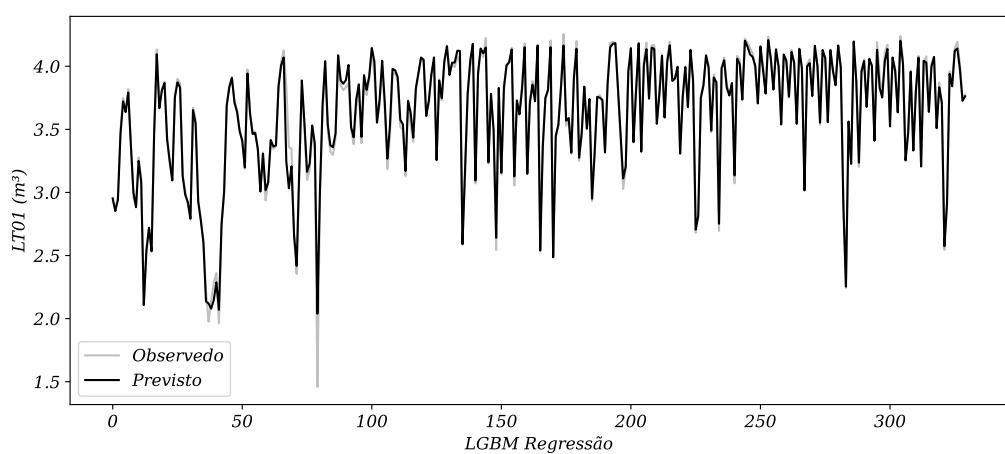


Fonte: Adaptação de Korstanje (2021)

Na Figura 30 pode ser visto como cada modelo é ajustado, no crescimento da árvore em folha e em nível.

Figura 31: XGBoost e LightGBM regressão





Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 31 é um modelo baseado nos dados coletados da SANEPAR.

## 4 Resultados

Neste capítulo é mostrado um breve resultado do que foi realizado até agora.

### 4.1 Planejamento do Problema

Assim como foi mostrado na seção 1.4.1 os passos da dissertação com que cada modelo e os métodos que podem ser usados para responder às perguntas de pesquisa abordadas na seção 1.2.1. Com os passos podem dar uma cronologia lógica do que foi adquirido ao longo do tempo com os dados SANEPAR.

#### 4.1.1 Análise Exploratória dos dados (EDA)

A partir de **Etapa 1** é realizado o EDA para o processamento de dados obtidos até agora, com EDA será respondido. De acordo com a Yu (2016) Na era dos grandes dados, coletamos volumes de dados em massa caóticos, não estruturados e multimídia através de vários canais. Como descobrir as regras, modelos analíticos e hipóteses destes dados se tornou o novo desafio. A análise exploratória de dados foi promovida por John Tukey para encorajar os estatísticos a explorar os dados e possivelmente formular hipóteses que poderiam levar a uma maior coleta de dados e experimentos. Em contraste com a análise inicial de dados, a análise exploratória de dados (EDA) é uma abordagem para analisar conjuntos de dados para resumir suas principais características, muitas vezes com métodos visuais. Muitas técnicas de EDA têm sido adotadas em grandes análises de dados.

Olhando o **Q 1** relacionando a demanda com a variável prevista e a pressão para a variável PT01 na Figura 24 pode-se ver que ambas estão trabalhando igualmente, quase uma correlação perfeita de  $r = 1$ , então para esta pergunta basta olhar para a correlação Pearson na Figura 24.

Para **Q 2** uma tabela é feita para responder melhor a esta pergunta

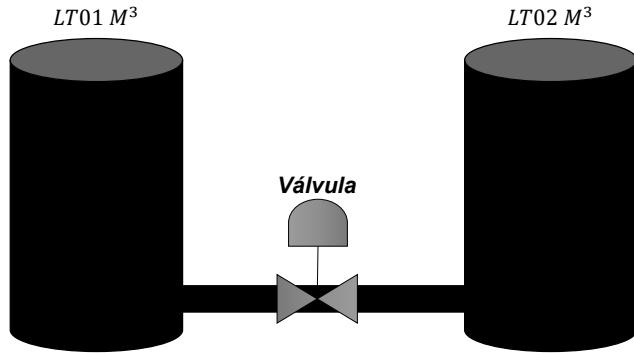
Tabela 4: Descrição estatística dos dados com o filtro aplicado das 18h às 21h

<b>18 a 21h</b>	<b>B1</b>	<b>B2</b>	<b>B3</b>	<b>LT01</b>	<b>FT01</b>	<b>FT02</b>	<b>FT03</b>	<b>PT01</b>	<b>PT02</b>
<b>Contagem</b>	366	366	366	366	366	366	366	366	366
<b>Média</b>	43,87	22,26	8,70	3,34	164,83	133,08	102,01	4,23	17,29
<b>STD</b>	23,22	18,47	17,81	0,69	114,60	67,99	47,55	0,81	8,59
<b>Min</b>	0	0	0	0,99	0,07	0	0	1,88	0
<b>25%</b>	37,93	0	0	2,87	64,31	131,06	107,92	3,69	16,77
<b>50%</b>	57,99	30,92	0	3,41	201,37	146,17	121,40	4,22	22,46
<b>75%</b>	57,99	37,25	0	3,86	268,61	158,71	127,07	4,85	22,52
<b>Max</b>	59,99	57,33	53,74	4,40	379,20	285,56	170,56	5,66	24,23

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na tabela 4 o desvio padrão é dado pela sigla STD que vem do inglês *standard deviation* também observando para responder ao Q 2 assim como toda empresa de tratamento de água é feito um acionamento automático chamado trava de segurança para que o tanque não chegue a zero e falte água em todos os lugares adjacentes que é abastecida por esta água, este mínimo que o tanque pode alcançar é  $1.459m^3 \iff 1459$  litros e as bombas serão ativadas em sua potência máxima para evitar a ativação das bombas o nível do tanque tem que estar na faixa de  $[3.843, 4.256] m^3$  bomba 1 ainda estaria funcionando para completar o nível. Em casos de pico, o mais ideal, mas não o mais rentável, é outro tanque de reserva nesses momentos e instalar uma tubulação para conectar uma à outra. Durante o dia, ambos estariam abastecendo e à noite, por gravidade, ficariam com o mesmo nível até que o consumo atingisse um nível para acionar as bombas.

Figura 32: Solução para o acionamento das bombas



Fonte: Elaboração própria

Na Figura 32 um esquema prático para evitar a escassez de água e o consumo em horários de pico. Este é um esquema muito simples de como a hora do dia pode ser melhorada para o armazenamento de água.

Na **Q 3** o tanque tem como máximo nos dados  $4,256\text{m}^3$  de dano em litros  $4256\text{L}$  para atender esta demanda e manter o tanque quase cheio ou sempre cheio o fluxo de entrada tem que estar entre  $[238, 302] \text{ m}^3/\text{h}$  fluxo de gravidade tem que estar entre  $[126, 182] \text{ m}^3/\text{h}$  fluxo de retorno entre  $[110, 144] \text{ m}^3/\text{h}$  pressão de sucção entre  $[1.92, 4.24]\text{mca}$ , pressão de retorno entre  $[21, 24] \text{ mca}$ .

Para **Q 4**, o ponto de equilíbrio para não iniciar as bombas seria o fluxo  $\text{FT01 } 211\text{m}^3/\text{h}$   $\text{FT02 } 114\text{m}^3/\text{h}$   $\text{FT03 } 100\text{m}^3/\text{h}$  e o nível do tanque a  $3,545\text{m}^3$ .

Ao **Q 5a.** o tanque deve estar a um nível de  $4,00\text{m}^3$  para que não precise funcionar com bombas nas horas de pico.

#### 4.1.2 Múltiplas entradas e saída única (MISO)

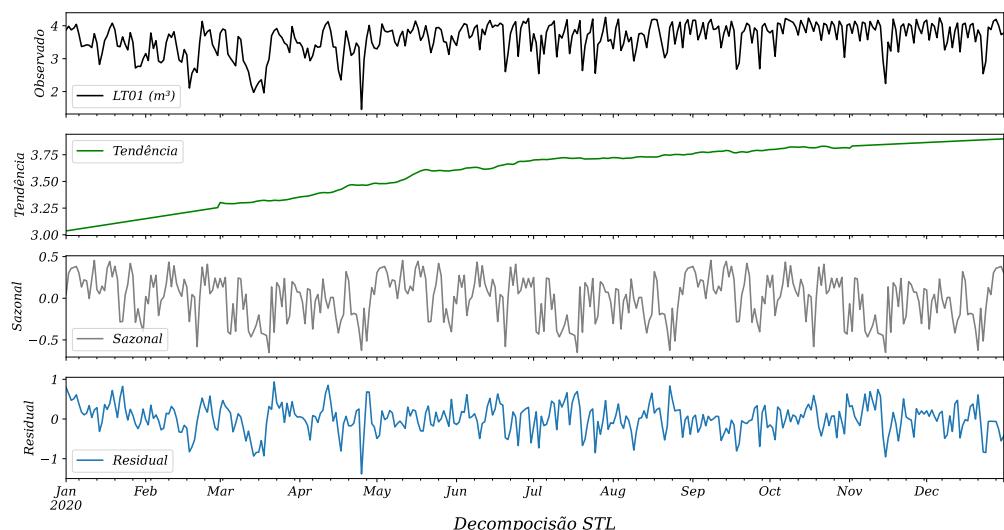
Nesta **Etapa 2** os modelos que foram mais cobertos no decorrer da dissertação são os modelos ARIMA ou aqueles derivados deste modelo e os modelos regressivos fora do LR têm múltiplas entradas e uma saída da variável que se prevê a  $\text{LT01}$ , as outras variáveis servem como suporte para melhorar os modelos do tipo ARX ou modelos com

variáveis exógenas. Os modelos ARIMA sem a variável exógena são apenas uma entrada semelhante com LR.

#### 4.1.3 Decomposição STL

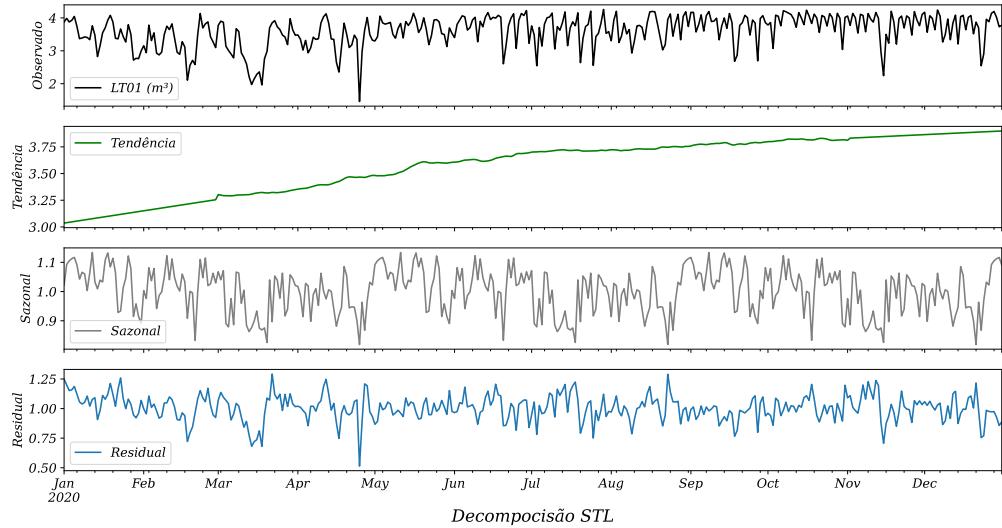
Theodosiou (2011) A decomposição sazonal e tendencial utilizando o procedimento de Loess (STL) é utilizada para a decomposição aditiva da série temporal global. O STL realiza a decomposição aditiva dos dados por meio de uma sequência de aplicações do Loess mais suave, que aplica regressões polinomiais ponderadas localmente em cada ponto do conjunto de dados, sendo as variáveis explicativas os valores mais próximos do ponto cuja resposta está sendo estimada.

Figura 33: Decomposição STL aditiva dos dados coletados



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 34: Decomposição STL multiplicativa dos dados coletados



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na **Q 5b.** resposta pode ser dada pelas Figuras 33 e 34 como é observado tem tenacidade, sazonalidade e resíduo.

Na decomposição, o objetivo é analisar se há tenacidade, sazonalidade e residência, olhando as Figuras 33 e 34, mostra que os dados têm ambas as análises. E com isso percebe-se que a série é estacionária, através do seguinte teste.

Teste de Dickey-Fuller (DF) Aumentado:

- Estatística de teste ADF  $-4.248$
- $p - valor$   $0.001$
- atrasos utilizados  $21.000$
- observações  $1074.000$
- valor crítico (1%)  $-3.436$
- valor crítico (5%)  $-2.864$
- valor crítico (10%)  $-2.568$

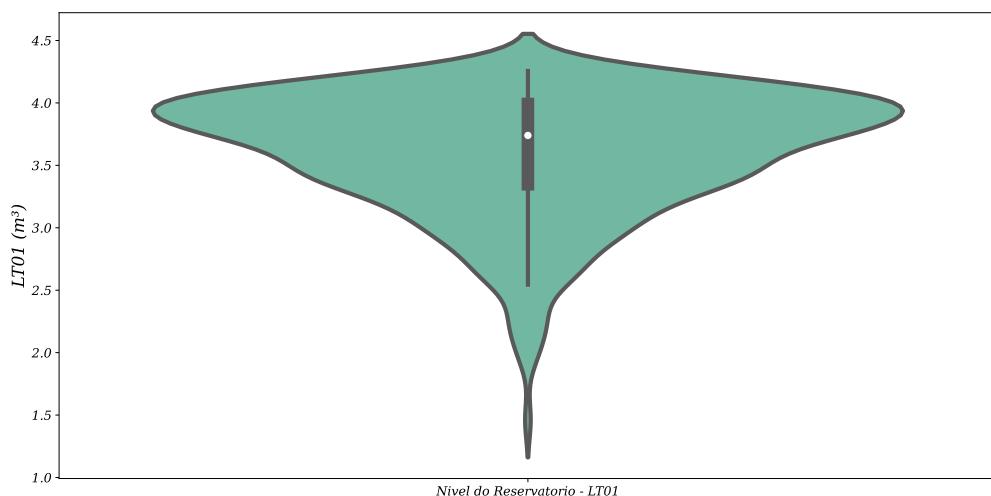
Forte evidência contra a hipótese nula

Rejeitar a hipótese nula

Os dados não têm raiz unitária e estão estacionários em **Q 5c.**, pois a série está estacionária para identificar quais são as horas de pico entre 18h e 21h não é um

trabalho fácil, como se você levasse a Figura 35 para ver que no ano de 2020 houve um aumento na demanda naquelas horas.

Figura 35: Violino no nível do reservatório

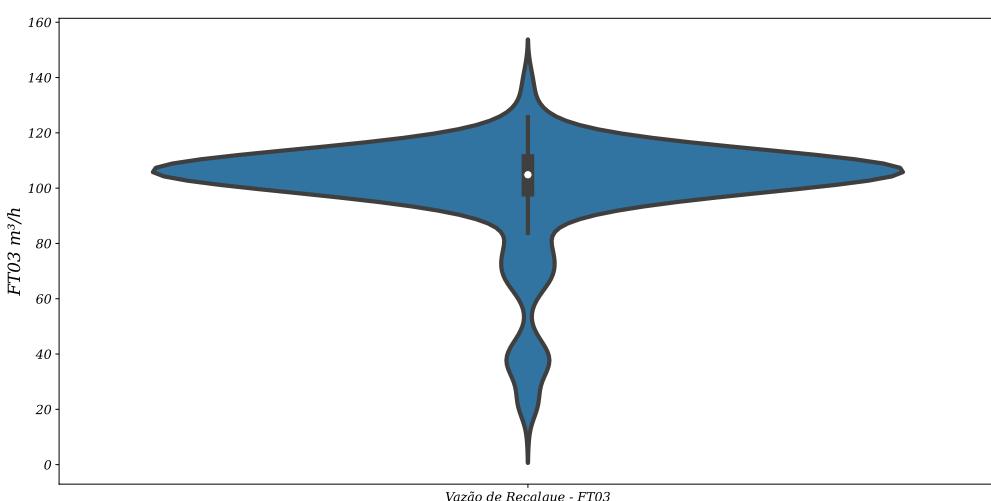


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Assim, como dito na seção 1.1.1 as anomalias climáticas mais ocasionadas no ano 2020 e foi devido à falta de chuva naquele período.

Em **Q** 5d. nas horas de pico deve conter no tanque cerca de  $[3.545, 4.256]m^3$  para que não ligue as bombas.

Figura 36: Violino da vazão de recalque



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Para **Q** 5e. é mostrado na Figura 36 como a vazão pode ser afetada com o nível do tanque. A vazão de recalque influencia mais o nível do tanque do que as outras vazões porque injeta água no tanque através da bomba que está mais próxima da base do tanque e as outras vazões por ter alguns valores ausentes não interferem tanto na amostra.

De acordo com o Reisen et al. (2017), o teste DF tem as seguintes equações

$$z_t = y_t + \theta\beta_t, \quad t = 1, \dots, T, \quad (20)$$

$$\hat{\rho}_{DF} - 1 = \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (21)$$

De (21) onde  $\Delta z_t = z_t - z_{t-1}$ . Sob a hipótese nula ( $H_0$ ) : “ $\rho = 1$ ”, as estatísticas do teste DF e suas distribuições limitantes são dadas da seguinte forma:

$$T(\hat{\rho}_{DF} - 1) = T \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (22)$$

e

$$\hat{\tau}_{DF} = \frac{\hat{\rho}_{DF} - 1}{\hat{\sigma}_{DF} \left( \sum_{t=1}^T z_{t-1}^2 \right)^{-1/2}} \quad (23)$$

De (23) onde  $\hat{\sigma}_{DF}^2 = T^{-1} \sum_{t=1}^T (\Delta z_t - (\hat{\rho}_{DF} - 1) z_{t-1})^2$ .

Suponha que  $(z_t)_{1 \leq t \leq T}$  são dadas por (20), então quando  $\rho = 1$ ,

$$T(\hat{\rho}_{DF} - 1) \xrightarrow{d} \frac{W(1)^2 - 1}{2 \int_0^1 W(r)^2 dr} - \left( \frac{\theta}{\sigma} \right)^2 \frac{\pi}{\int_0^1 W(r)^2 dr}, \text{ como } T \rightarrow \infty \quad (24)$$

$$\hat{\tau}_{DF} \xrightarrow{d} [1 + 2(\theta/\sigma)^2 \pi]^{-1/2} \left\{ \frac{W(1)^2 - 1}{2 \left( \int_0^1 W(r)^2 dr \right)^{1/2}} - \frac{(\theta/\sigma)^2 \pi}{\left( \int_0^1 W(r)^2 dr \right)^{1/2}} \right\} \quad (25)$$

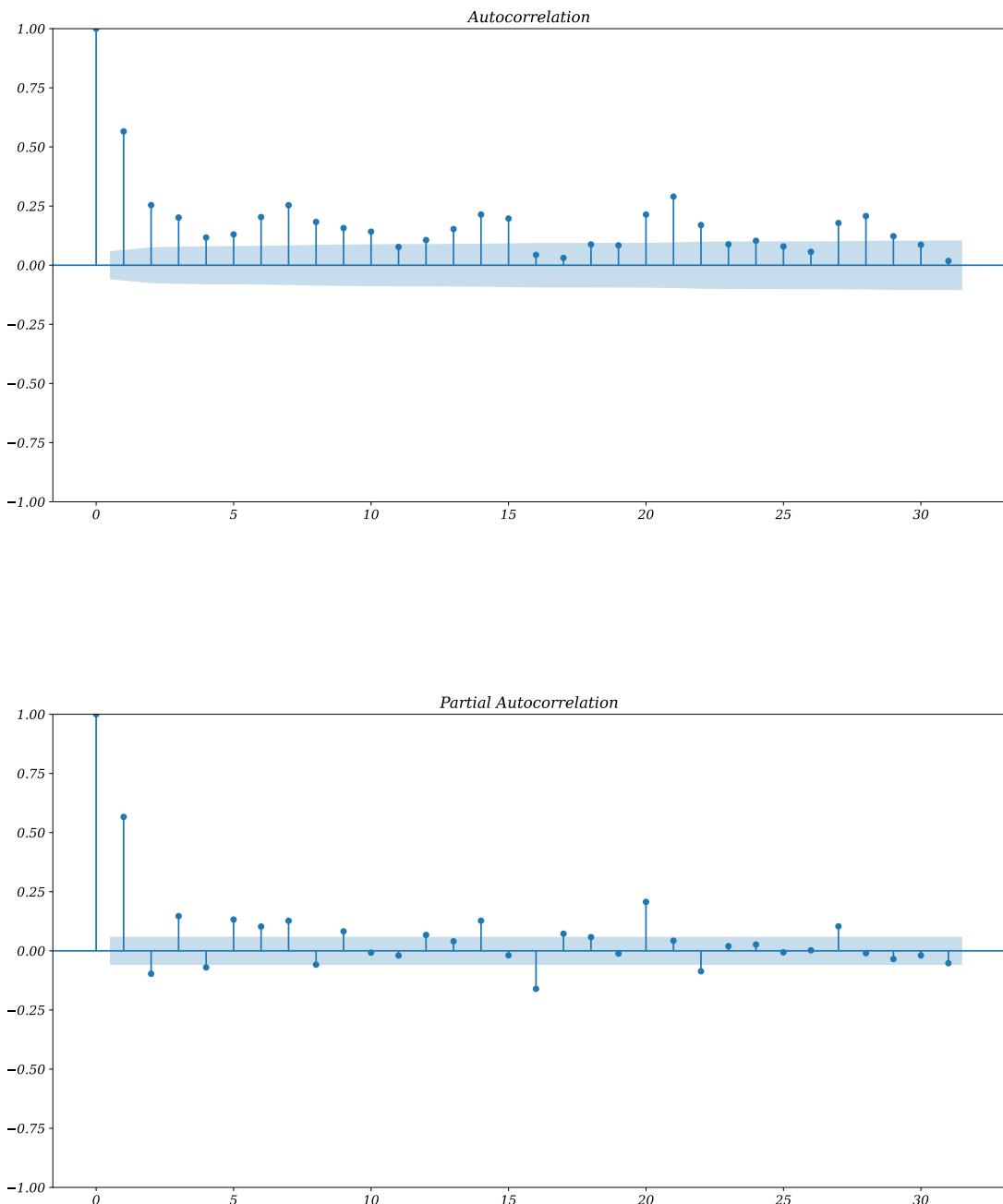
$$\text{como } T \rightarrow \infty \quad (26)$$

A partir de (26), onde  $\xrightarrow{d}$  denota convergência na distribuição e onde  $\{W(r), r \in$

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

Este teste na literatura é chamado de teste ACF para testar se a série é estacionária ou não.

Figura 37: Autocorrelação e Autocorrelação parcial



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na figura 37 temos a diferença entre autocorrelação e autocorrelação parcial (PACF) é quase um detalhe em um ACF temos a correlação direta e indireta e em um PACF so-

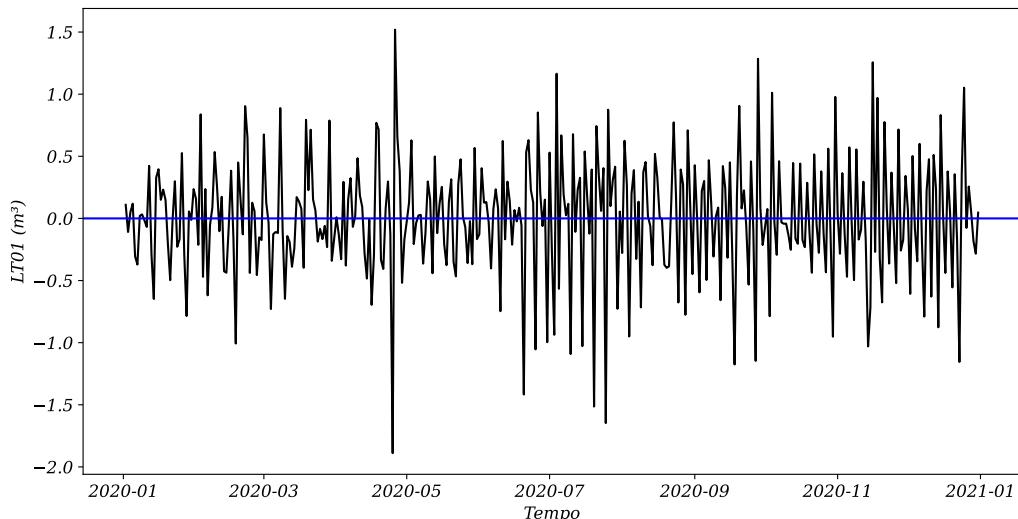
mente a correlação direta.

Na Figura 37 tem a diferença entre a autocorrelação e a autocorrelação parcial (PACF) é quase um detalhe em uma ACF temos a correlação direta e indireta e em uma PACF apenas a correlação direta.

O intervalo de confiança padrão é 95% mostrado como esta marca azul. As observações que estão fora da marca são consideradas estatisticamente correlacionadas.

A correlação na Figura 37 é a explicação do teste DF. Os valores de uma série de ruído branco são totalmente aleatórios, ou seja, este é um tipo de série que não é previsível.

Figura 38: Ruído branco



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Da Figura 38 uma série temporal pode ser ruído branco. Uma série temporal a é ruído branco se as variáveis forem independentes e distribuídas de forma idêntica com uma média de zero. Isto significa que todas as variáveis têm a mesma variância ( $\sigma^2$ ) e cada valor tem correlação zero com todos os outros valores da série. Mais adiante, é mostrado o comprimento de zeros na variável prevista. Isto conclui o **Etapa 3**.

#### 4.1.4 Separação dos dados

Na **Etapa 4** tem um esquema de como os dados foram divididos em treinamento, teste e validação, esta prática é comum para profissionais de aprendizagem de máquinas porque além de não poder processar os dados de uma só vez se lidar com dados em uma

escala menor eles podem até rodar, mas tudo depende da máquina que está rodando o processamento de dados cada modelo particular usa uma certa coleção do seu computador para processar se, Por exemplo, você está trabalhando com um modelo de aprendizado profundo que é mais comum no processamento de imagens Nvidia sempre inovou com suas GPUs e trazendo mais poder ao processamento, com o recente lançamento da placa de vídeo de 4090 um sonho do consumidor de jogos e de profissionais de aprendizado profundo e de máquinas.

Em resumo, se o computador que foi realizado o processamento era um computador não tão bom, você ainda pode estar pensando que ele estaria processando sem a inovação que foi estabelecida ao longo dos anos, o computador que foi realizado os cálculos dos modelos era em partes um processador de computador *i5 – 3330* e um notebook com *i7 – 5500* ambos com 4 fios (em português: fio de execução ou encadeamento de execução) e o notebook com apenas 2 núcleos o *i5* contém 4 núcleos. Cada um deles tem suas especificações para ser o melhor em algum momento, mas sabendo que não é preciso a última geração para fazer tal processamento. É a vontade de compreender e aplicar cada um deles.

A divisão mais básica que você tem na literatura foi realizada aqui separando os dados de 70% para treinamento e os dados restantes de 30% para testes os dados de 70% têm uma divisão adicional que leva 80% dos dados de 70% para treinamento novamente e os dados de 20% para validação tendo esta fórmula aplicada em linguagem de programação para que não precise ser contada toda vez que o modelo for alterado.

#### 4.1.5 Estratégia de Previsão

Na **Etapa 5**, discute-se como os dados foram previstos em uma janela de horizonte de previsão muito maior do que o habitual na literatura para a estratégia recursiva de 1, 7, 14 e 30 dias previstos nesta estratégia para comparação de modelos de regressão e modelos ARIMA é muito vantajosa porque cada modelo tem sua especificidade para prever em momentos com janela de tempo menor e com uma janela de muitos dias. Como explicado na seção 4.1.7, se for curta a previsão, alguns irão prever em excesso no meio dos outros modelos que foram feitos aqui.

#### 4.1.6 Horizonte

Na **Etapa 6**, o horizonte de previsão foi personalizado com base no método recursivo de previsão de série temporal e na previsão do nível do tanque LT01. Os passos para a previsão à frente foram 1, 7, 14 e 30 dias. Uma estratégia com uma janela menor

já foi executada, mas para a comparação dos modelos, esta janela foi mais adequada.

#### 4.1.7 Modelos de previsão e métricas de desempenho

A partir da **Etapa 7**, as métricas utilizadas aqui foram vistas na seção 3.1 e três das métricas mais utilizadas na literatura para previsão e comparação dos modelos ARIMA e modelos de regressores foram utilizadas aqui.

Em comparação com os modelos feitos, pode-se ver que o modelo LR em um passo à frente tem tanto na modelagem de 24 horas como nas horas de pico entre 18 e 21 horas foi o modelo que melhor se saiu na previsão logo após os modelos MA, AR, SARIMA, ARIMA, SARIMAX, ARIMAX, ARX, LGBMRegressor, XGBRegressor e Random Forest Regressor para o curto prazo estes modelos estão em ordem do melhor para o pior.

Já em períodos mais longos, como foi feito em 30 dias os modelos ARMA, AR, MA, ARIMA, ARIMAX, ARX, SARIMA, SARIMAX, XGBRegressor, Random Forest Regressor, LGBMRegressor e LR, seguindo a mesma lógica do melhor ao pior. Mas também olhando graficamente os modelos que foram feitos os modelos com variáveis exógenas parecem prever melhor do que os outros modelos apenas olhando os dados nos apêndices tanto quanto as Figuras de 41 a 52 como as Tabelas 6 a 21.

#### 4.1.8 Teste de Significância

Na **Etapa 8**, o teste escolhido foi de *Friedman e Nemenjy* no teste de Nemenyi precisa ser para obter a diferença entre as classificações médias (linha do meio da tabela de classificação) entre todos os classificadores (comparando pares de classificadores). Se esta diferença for maior ou igual a um CD (distância crítica), pode-se dizer que estes dois classificadores são significativamente diferentes um do outro. O CD é calculado como:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (27)$$

De (27) o termo  $q_\alpha$  é obtido de ( $\alpha = 0,05$ ):

Tabela 5: Teste Nemenyi

Nemenyi	0	1	2	3	4	5	6	7	8
<b>0</b>	1,000	0,001	0,001	0,001	0,001	0,001	0,001	0,001	0,001
<b>1</b>	0,001	1,000	0,001	0,001	0,001	0,001	0,001	0,001	0,157
<b>2</b>	0,001	0,001	1,000	0,847	0,001	0,001	0,001	0,001	0,001
<b>3</b>	0,001	0,001	0,847	1,000	0,001	0,001	0,001	0,001	0,001
<b>4</b>	0,001	0,001	0,001	0,001	1,000	0,001	0,001	0,001	0,001
<b>5</b>	0,001	0,001	0,001	0,001	0,001	1,000	0,001	0,001	0,001
<b>6</b>	0,001	0,001	0,001	0,001	0,001	0,001	1,000	0,001	0,001
<b>7</b>	0,001	0,001	0,001	0,001	0,001	0,001	0,001	1,000	0,001
<b>8</b>	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 2020)

O teste de Nemenyi (Nemenyi, 1963) é um teste *post-hoc*, ou seja, é um teste de comparação múltipla que é usado após a aplicação de teste não paramétricos com três ou mais fatores.

Para calcular a estatística de teste  $F_r$  de Friedman cria-se inicialmente uma tabela com os dados, colocando-se em cada linha uma amostra e cada coluna correspondendo a uma condição de teste. A seguir, as amostras ao longo das condições são ordenadas, da melhor situação para a pior. Se não houver empates, usa-se a equação (28) para determinar a estatística de teste  $F_r$ :

$$F_r = \left[ \frac{12}{nk(k+1)} \sum_{i=1}^k R_i^2 \right] - 3n(k+1) \quad (28)$$

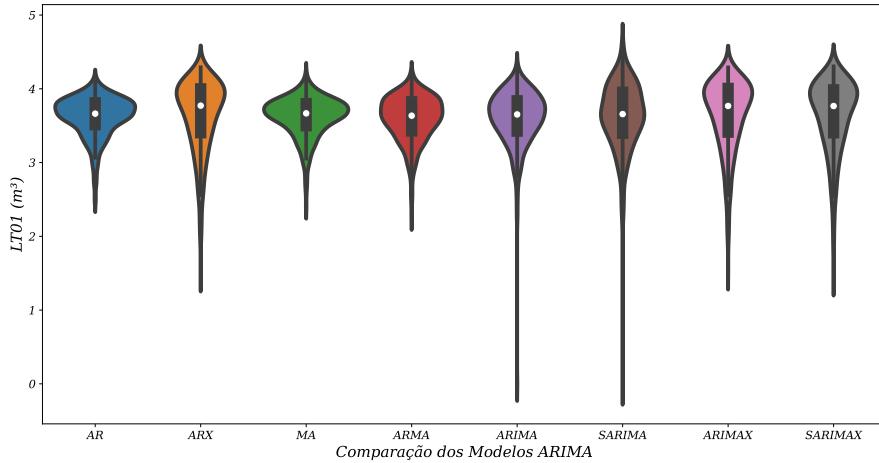
Na equação (28)  $n$  é o número de linhas (ou amostras)  $k$  é o número de colunas (ou condições) e  $R_i$  é a soma das fileiras da coluna (ou condição)  $i$ . Seguindo a equação (28) tem o seguinte resultado nos dados da pesquisa.

*statistic* = 8015.611, *pvalue* = 0.0 com o números de 26306 linhas x 9 colunas.

#### 4.1.9 Comparação dos modelos

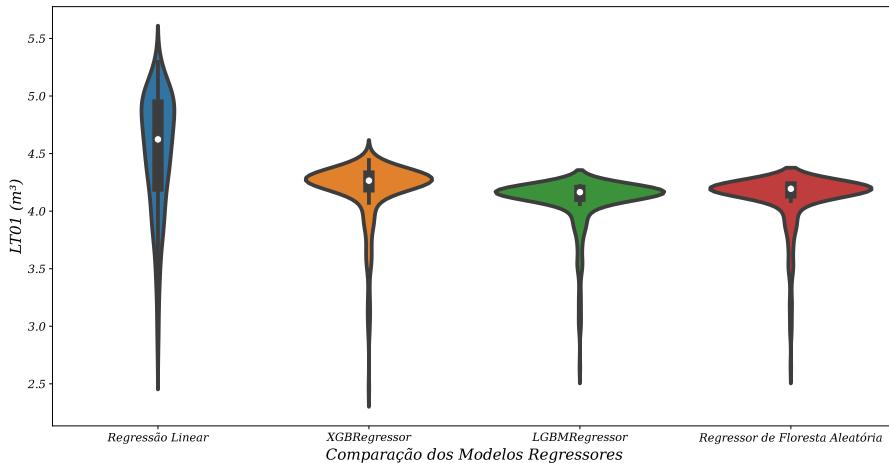
A fim de ver melhor como cada modelo se comporta, os modelos foram comparados com base em um gráfico de violino, e assim observar qual dos modelos era o melhor.

Figura 39: Comparação dos modelos ARIMAS



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 40: Comparação de modelos de regressão



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Em comparação com os modelos apresentados nas Figuras 39 e 40 os modelos que podem ser observados que são os melhores levando em conta a modelagem dos dados nos modelos ARIMA os melhores são AR, ARX, MA, ARMA, ARIMAX e SARIMAX devido aos *outliers* e ao limite inferior de alguns modelos que olham para os modelos de gradiente e regressão pode-se notar que eles eram semelhantes devido às técnicas de otimização matemática Grid Search (do inglês pesquisa grande) e Randomized Search (do inglês pesquisa aleatória) que permitiram o aperfeiçoamento do método utilizado. Em um horizonte de previsão pequeno, LR prevê melhor que os outros modelos, mas em um

horizonte de previsão maior, XGBoost e Light GBM estão prevendo com melhor precisão. A floresta aleatória também está prevendo com precisão apenas atrás do XGBoost em previsões de longo prazo.

O método Ljung box é um método que pode ser estimado nos modelos ARIMAS de longo prazo se, a longo prazo, eles ainda irão prever eficientemente nos dados de longo prazo os modelos que melhor prevêem são os modelos ARX, ARIMAX e SARIMAX com as variáveis exógenas para modelos não lineares que podem aguentar mais tempo de previsão do que os outros modelos ARIMA.

## 5 Conclusões

Nesta dissertação, o objetivo era mostrar a escassez de água que ocorreu em Curitiba, tornando possível uma decisão que foi uma adaptação do caso de 12 passos do ALMEIDA (2013), que busca e visa o ambiente para ter a visão de que não há interferência do ambiente, e se há esta interferência, ela foi listada como uma variável exógena nos modelos ARX, ARIMAX e SARIMAX, em modelos regressivos, mesmo bom para trabalhar com eles, eu não poderia incluí-los neste momento. Se o projetista está procurando anomalias nos dados como foi feito aqui, procure os dados de 2020, que foi a grande anomalia na SANEPAR, estas anomalias explicadas nos resultados no capítulo 4.

### 5.1 Limitações da pesquisa e propostas futuras

As limitações deste trabalho resultam no tempo e os modelos de aprendizagem de máquina, como vistos durante esta dissertação, têm vários modelos que podem ser trabalhados em conjunto com as séries temporais, por exemplo, os modelos de rede neural LSTM, CNN, RNN... Entre outros modelos que não foram muito bem tratados aqui porque são modelos mais complexos e exigiriam um maior intervalo de tempo para este momento, apenas os modelos que foram trabalhados no início atenderam à questão de pesquisa que foi levantada.

Mas nos próximos passos para um trabalho futuro é abordar melhor estes modelos de previsão tendo com muitos autores na literatura que trabalham com estes modelos, até competição de aprendizagem de máquinas com os modelos mais famosos como o Light GBM em comparação com o XGboost para previsão de curto prazo e para longo prazo cada modelo tem sua relevância LR como um modelo de máximo 3 variáveis para dados com poucas variáveis é muito eficiente e ágil.

No trabalho que se seguirá a este como complemento a este trabalho, tem como abordar toda a literatura, não apenas os últimos 6 anos, e também visa as outras partes que não foram abordadas como dissertações, teses e capítulos de livros, apesar de ter abordado um pequeno grupo de artigos, ainda tinha uma gama muito grande de artigos sobre o assunto.

A otimização matemática com alguns modelos como floresta aleatória, XGboost, Ligth GBM, que poderia ser usada para aumentar o gradiente e melhorar a precisão da aleatoriedade dos galhos das árvores. Os métodos de otimização para melhorar o modelo foram **Grid Search**, **Randomized Search** e **Bayesian Optimization (Bayes Search)** que vem do inglês para o português seria **Otimização Bayesiana** para a floresta

aleatória o melhor método em hipóteses sérias a busca aleatória (randomized) dos galhos mais rapidamente a árvore predizendo assim melhor o tempo, mas em teoria todos eles em algum modelo falharam em reduzir os erros listados na seção 3.1 em vez de reduzir, houve um aumento dos erros tornando a previsão ao longo do tempo pior, como por exemplo no apêndice C que teve os melhores resultados se pegar os erros entre os modelos citados anteriormente e em comparação com os modelos de otimização encontrados na literatura teve um aumento dos erros de 6 para 30 % e para uma previsão mais precisa ela precisa ser próxima de zero.

Nesta parte da otimização é relevante pesquisar ou ter mais profundidade nos hiperparâmetros para ter uma melhor utilização da árvore e modelos de gradiente.

## 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. 301–320, 2018. ISSN 0378-7788.
- ALMEIDA, A. T. D. **Processo de Decisão nas Organizações-Construindo Modelos de Decisão Multicritério. Atlas.** [S.l.]: São Paulo, 2013.
- BERGMEIR, C.; HYNDMAN, R.; KOO, B. A note on the validity of cross-validation for evaluating autoregressive time series prediction. **Computational Statistics and Data Analysis**, v. 120, p. 70–83, 2018.
- BOROOJENI, K. et al. A novel multi-time-scale modeling for electric power demand forecasting: From short-term to medium-term horizon. **Electric Power Systems Research**, v. 142, p. 58–73, 2017.
- BRANDÃO, G. A. **Séries Temporais: Parte 1.** DEV Community, 2020. Disponível em: <<https://dev.to/giselyalves13/series-temporais-parte-1-13l8>>.
- BROWNLEE, J. **Machine learning mastery with Python: understand your data, create accurate models, and work projects end-to-end.** [S.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. 226–242, 2019. ISSN 15684946 (ISSN). Disponível em: <<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, 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-Window Metaheuristic-Optimized Machine-Learning Regression. **IEEE Transactions on Industrial Informatics**, v. 14, n. 7, p. 3132–3142, 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. 709–726, 2018.
- COELHO, I. et al. A GPU deep learning metaheuristic based model for time series forecasting. **Applied Energy**, v. 201, p. 412–418, 2017.

- DU, S. et al. Multivariate time series forecasting via attention-based encoder–decoder framework. **Neurocomputing**, v. 388, p. 269–279, 2020.
- 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. 12, n. 4, 2020. ISSN 1939-0068.
- GRAFF, M. et al. Time series forecasting with genetic programming. **Natural Computing**, v. 16, n. 1, p. 165–174, 2017.
- KORSTANJE, J. **Advanced Forecasting with Python**. [S.l.]: Springer, 2021.
- KULSHRESHTHA, S.; VIJAYALAKSHMI, 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. 117–123, 2020.
- 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, 2021. ISSN 1134-3060.
- LARA-BENITEZ, P.; CARRANZA-GARCIA, M.; RIQUELME, J. C. An Experimental Review on Deep Learning Architectures for Time Series Forecasting. **INTERNATIONAL JOURNAL OF NEURAL SYSTEMS**, v. 31, n. 3, 2021. ISSN 0129-0657.
- LI, A. W.; BASTOS, G. S. Stock Market Forecasting Using Deep Learning and Technical Analysis: A Systematic Review. **IEEE ACCESS**, v. 8, p. 185232–185242, 2020. 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. 392–408, 2019. ISSN 0306-2619.
- LIU, Z. Y. et al. Forecast Methods for Time Series Data: A Survey. **IEEE ACCESS**, v. 9, p. 91896–91912, 2021. ISSN 2169-3536 J9 - IEEE ACCESS JI - IEEE Access.
- MARTINOVIĆ, M.; HUNJET, A.; TURCIN, I. Time series forecasting of the austrian traded index (Atx) using artificial neural network model. **Tehnicki Vjesnik**, v. 27, n. 6, p. 2053–2061, 2020.
- 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://www.sciencedirect.com/science/article/pii/S0950584916300568>>.
- MOON, J. et al. Temporal data classification and forecasting using a memristor-based reservoir computing system. **Nature Electronics**, 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 Sensing of Environment**, v. 187, p. 156–168, 2016. Cited By 296. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84992151859&doi=10.1016%2f.rse.2016.10.010&partnerID=40&md5=09efc79bab8e893b97fd21cb4844b98d>>.

PINHEIRO, N. M. **Introdução a Series Temporais — Parte 1.** Data Hackers, 2022. Disponível em: <<https://medium.com/data-hackers/series-temporais-parte-1-a0e75a512e72>>.

QUININO, R. C.; REIS, E. A.; BESSEGATO, L. F. O coeficiente de determinação r<sup>2</sup> como instrumento didático para avaliar a utilidade de um modelo de regressão linear múltipla. **Belo Horizonte:** UFMG, 1991.

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-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, 2021.

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. **Chaos, Solitons and Fractals**, v. 138, 2020.

SAMANTA, S. et al. Learning elastic memory online for fast time series forecasting. **Neurocomputing**, v. 390, p. 315–326, 2020.

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, 2020. ISSN 1568-4946.

SHEN, Z. et al. A novel time series forecasting model with deep learning. **Neurocomputing**, v. 396, p. 302–313, 2020.

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. 1421–1441, 2019.

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 Transactions on Neural Networks and Learning Systems**, Institute of Electrical and Electronics Engineers Inc., Department of Computer Science, Université Libre de Bruxelles, Brussels, 1050, Belgium, v. 27, n. 1, p. 62–76, 2016. ISSN 2162237X (ISSN). Disponível em: <<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, 2021. ISSN 1999-5903.

THEODOSIOU, M. Forecasting monthly and quarterly time series using stl decomposition. **International Journal of Forecasting**, v. 27, n. 4, p. 1178–1195, 2011. Cited By 86. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-80052160927&doi=10.1016%2fijforecast.2010.11.002&partnerID=40&md5=e8242471ba1ec14ada46ab567f3a364d>>.

TRENBERTH, K. E. Signal versus noise in the southern oscillation. **Monthly Weather Review**, v. 112, n. 2, p. 326–332, 1984.

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

URSU, E.; PEREAU, J. C. Application of periodic autoregressive process to the modeling of the Garonne river flows. **STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT**, v. 30, n. 7, p. 1785–1795, 2016. ISSN 1436-3240.

VASCONCELOS, F. **Falta d'água em curitiba e região metropolitana não É culpa só da estiagem**. 2020. Disponível em: <<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 Networks**, v. 126, p. 191–217, 2020.

WANG, 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. 123, p. 673–689, 2016. ISSN 0950-0618.

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 0022-1694.

XU, W. et al. Deep belief network-based AR model for nonlinear time series forecasting. **Applied Soft Computing Journal**, v. 77, p. 605–621, 2019.

YANG, W. 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. 222, p. 942–959, 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 9781509023509. 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:124077. Disponível em: <<https://www.scopus.com/inward/record.uri?eid=2-s2.0-84994079422&doi=10.1109%2fAgro-Geoinformatics.2016.7577697&partnerID=40&md5=fef861624a35632bf2d84acf63986bbe>>.

## A 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 6: Comparação dos modelos com 1 dia de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
<b>AR</b>	0,096	0,306	0,419
<b>ARX</b>	0,118	0,377	0,513
<b>MA</b>	0,093	0,296	0,403
<b>ARMA</b>	0,102	0,325	0,435
<b>ARIMA</b>	0,095	0,302	0,405
<b>SARIMA</b>	0,105	0,342	0,450
<b>ARIMAX</b>	0,119	0,378	0,511
<b>SARIMAX</b>	0,118	0,377	0,512
<b>LR</b>	<b>0,015</b>	0,069	0,077
<b>RFR</b>	0,190	0,624	0,672
<b>XGBRegressor</b>	0,207	0,683	0,720
<b>LGBMRegressor</b>	0,184	0,599	0,655

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 7: Comparação dos modelos com 1 dia de antecedência 24h **Validação**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,084	0,285	0,366
<b>ARX</b>	0,103	0,354	0,459
<b>MA</b>	0,082	0,278	0,361
<b>ARMA</b>	0,086	0,295	0,372
<b>ARIMA</b>	0,082	0,280	0,351
<b>SARIMA</b>	0,097	0,333	0,421
<b>ARIMAX</b>	0,102	0,353	0,458
<b>SARIMAX</b>	0,104	0,358	0,463
<b>LR</b>	<b>0,014</b>	0,066	0,073
<b>RFR</b>	0,172	0,587	0,633
<b>XGBRegressor</b>	0,192	0,658	0,692
<b>LGBMRegressor</b>	0,166	0,564	0,616

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 8: Comparação dos modelos com 1 dia de antecedência 24h **Teste**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,100	0,329	0,424
<b>ARX</b>	0,137	0,462	0,586
<b>MA</b>	0,102	0,336	0,431
<b>ARMA</b>	0,102	0,340	0,433
<b>ARIMA</b>	0,103	0,346	0,440
<b>SARIMA</b>	0,118	0,398	0,501
<b>ARIMAX</b>	0,137	0,461	0,587
<b>SARIMAX</b>	0,138	0,464	0,590
<b>LR</b>	<b>0,018</b>	0,087	0,098
<b>RFR</b>	0,153	0,494	0,587
<b>XGBRegressor</b>	0,170	0,560	0,643
<b>LGBMRegressor</b>	0,145	0,465	0,568

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 9: Comparação dos modelos com 1 dia de antecedência 24h **Completo**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,093	0,302	0,165
<b>ARX</b>	0,123	0,402	0,283
<b>MA</b>	0,107	0,344	0,460
<b>ARMA</b>	0,097	0,316	0,424
<b>ARIMA</b>	0,094	0,303	0,406
<b>SARIMA</b>	0,106	0,350	0,448
<b>ARIMAX</b>	0,120	0,394	0,521
<b>SARIMAX</b>	0,122	0,401	0,530
<b>LR</b>	<b>0,016</b>	0,074	0,084
<b>RFR</b>	0,176	0,579	0,642
<b>XGBRegressor</b>	0,194	0,643	0,694
<b>LGBMRegressor</b>	0,170	0,554	0,624

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 10: Comparação dos modelos com 7 dias de antecedência 24h **Treinamento**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	<b>0,093</b>	0,296	0,399
<b>ARX</b>	0,118	0,377	0,524
<b>MA</b>	0,104	0,329	0,444
<b>ARMA</b>	0,103	0,330	0,439
<b>ARIMA</b>	0,108	0,342	0,463
<b>SARIMA</b>	0,111	0,360	0,487
<b>ARIMAX</b>	0,118	0,379	0,525
<b>SARIMAX</b>	0,118	0,379	0,525
<b>LR</b>	1,197	5,230	5,230
<b>RFR</b>	0,224	0,705	0,821
<b>XGBRegressor</b>	0,260	0,823	0,934
<b>LGBMRegressor</b>	<b>0,215</b>	0,673	0,793

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 11: Comparação dos modelos com 7 dias de antecedência 24h **Validação**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	<b>0,073</b>	0,245	0,319
<b>ARX</b>	0,093	0,319	0,423
<b>MA</b>	0,080	0,269	0,353
<b>ARMA</b>	0,081	0,274	0,347
<b>ARIMA</b>	0,087	0,292	0,384
<b>SARIMA</b>	0,095	0,324	0,438
<b>ARIMAX</b>	0,093	0,318	0,422
<b>SARIMAX</b>	0,094	0,320	0,424
<b>LR</b>	1,174	5,224	5,224
<b>RFR</b>	0,188	0,630	0,712
<b>XGBRegressor</b>	0,223	0,756	0,828
<b>LGBMRegressor</b>	<b>0,179</b>	0,598	0,684

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 12: Comparação dos modelos com 7 dias de antecedência 24h **Teste**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,118	0,383	0,499
<b>ARX</b>	0,147	0,479	0,632
<b>MA</b>	0,125	0,403	0,530
<b>ARMA</b>	<b>0,117</b>	0,384	0,494
<b>ARIMA</b>	0,120	0,393	0,505
<b>SARIMA</b>	0,131	0,437	0,544
<b>ARIMAX</b>	0,148	0,480	0,632
<b>SARIMAX</b>	0,148	0,481	0,636
<b>LR</b>	1,161	5,212	5,213
<b>RFR</b>	0,187	0,578	0,755
<b>XGBRegressor</b>	0,222	0,693	0,870
<b>LGBMRegressor</b>	<b>0,177</b>	0,543	0,727

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 13: Comparação dos modelos com 7 dias de antecedência 24h **Completo**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,098	0,316	0,177
<b>ARX</b>	0,125	0,406	0,305
<b>MA</b>	0,105	0,337	0,450
<b>ARMA</b>	<b>0,097</b>	<b>0,312</b>	<b>0,418</b>
<b>ARIMA</b>	<b>0,097</b>	0,314	0,420
<b>SARIMA</b>	0,118	0,386	0,506
<b>ARIMAX</b>	0,124	0,402	0,546
<b>SARIMAX</b>	0,125	0,405	0,551
<b>LR</b>	1,183	5,224	5,224
<b>RFR</b>	0,208	0,656	0,787
<b>XGBRegressor</b>	0,243	0,775	0,901
<b>LGBMRegressor</b>	<b>0,199</b>	0,623	0,759

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

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

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	<b>0,105</b>	0,334	0,445
<b>ARX</b>	0,126	0,399	0,548
<b>MA</b>	0,106	0,336	0,447
<b>ARMA</b>	0,110	0,350	0,463
<b>ARIMA</b>	0,111	0,353	0,477
<b>SARIMA</b>	0,114	0,367	0,489
<b>ARIMAX</b>	0,126	0,401	0,547
<b>SARIMAX</b>	0,126	0,401	0,547
<b>LR</b>	2,606	11,394	11,394
<b>RFR</b>	0,221	0,696	0,812
<b>XGBRegressor</b>	0,269	0,859	0,962
<b>LGBMRegressor</b>	<b>0,215</b>	0,673	0,792

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 15: Comparação dos modelos com 14 dias de antecedência 24h **Validação**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	<b>0,078</b>	0,264	0,346
<b>ARX</b>	0,090	0,309	0,430
<b>MA</b>	0,079	0,265	0,349
<b>ARMA</b>	0,093	0,317	0,403
<b>ARIMA</b>	0,088	0,295	0,389
<b>SARIMA</b>	0,092	0,315	0,402
<b>ARIMAX</b>	0,090	0,308	0,429
<b>SARIMAX</b>	0,090	0,308	0,429
<b>LR</b>	2,558	11,388	11,388
<b>RFR</b>	0,185	0,619	0,702
<b>XGBRegressor</b>	0,233	0,790	0,859
<b>LGBMRegressor</b>	<b>0,179</b>	0,598	0,683

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 16: Comparação dos modelos com 14 dias de antecedência 24h **Teste**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,118	0,378	0,504
<b>ARX</b>	0,120	0,384	0,560
<b>MA</b>	0,120	0,385	0,509
<b>ARMA</b>	0,107	0,344	0,464
<b>ARIMA</b>	<b>0,105</b>	0,338	0,462
<b>SARIMA</b>	0,113	0,364	0,496
<b>ARIMAX</b>	0,120	0,384	0,560
<b>SARIMAX</b>	0,119	0,383	0,558
<b>LR</b>	2,531	11,376	11,377
<b>RFR</b>	0,186	0,572	0,748
<b>XGBRegressor</b>	0,227	0,710	0,889
<b>LGBMRegressor</b>	<b>0,177</b>	0,542	0,725

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 17: Comparação dos modelos com 14 dias de antecedência 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
<b>AR</b>	0,104	0,335	0,204
<b>ARX</b>	0,122	0,393	0,297
<b>MA</b>	0,106	0,340	0,452
<b>ARMA</b>	<b>0,097</b>	0,311	0,423
<b>ARIMA</b>	0,099	0,318	0,431
<b>SARIMA</b>	0,113	0,365	0,492
<b>ARIMAX</b>	0,121	0,389	0,539
<b>SARIMAX</b>	0,122	0,393	0,543
<b>LR</b>	2,577	11,388	11,388
<b>RFR</b>	0,206	0,648	0,779
<b>XGBRegressor</b>	0,251	0,804	0,927
<b>LGBMRegressor</b>	<b>0,198</b>	0,623	0,758

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 18: Comparação dos modelos com 30 dias de antecedência 24h **Treinamento**

Modelos	Erros		
	MAPE	MAE	RMSE
<b>AR</b>	0,121	0,383	0,514
<b>ARX</b>	0,135	0,432	0,592
<b>MA</b>	<b>0,120</b>	<b>0,379</b>	0,510
<b>ARMA</b>	<b>0,120</b>	0,383	<b>0,508</b>
<b>ARIMA</b>	0,124	0,395	0,527
<b>SARIMA</b>	0,126	0,405	0,538
<b>ARIMAX</b>	0,136	0,434	0,594
<b>SARIMAX</b>	0,136	0,435	0,596
<b>LR</b>	5,827	25,483	25,484
<b>RFR</b>	0,224	0,705	0,821
<b>XGBRegressor</b>	0,282	0,902	0,998
<b>LGBMRegressor</b>	<b>0,211</b>	0,659	0,780

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 19: Comparação dos modelos com 30 dias de antecedência 24h **Validação**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,091	0,311	0,390
<b>ARX</b>	0,086	0,302	0,434
<b>MA</b>	0,090	0,306	0,383
<b>ARMA</b>	0,089	0,304	0,384
<b>ARIMA</b>	0,100	0,343	0,426
<b>SARIMA</b>	0,098	0,337	0,412
<b>ARIMAX</b>	<b>0,086</b>	<b>0,301</b>	<b>0,433</b>
<b>SARIMAX</b>	<b>0,086</b>	0,302	0,434
<b>LR</b>	5,721	25,478	25,478
<b>RFR</b>	0,187	0,628	0,710
<b>XGBRegressor</b>	0,245	0,831	0,896
<b>LGBMRegressor</b>	<b>0,174</b>	0,580	0,666

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 20: Comparação dos modelos com 30 dias de antecedência 24h **Teste**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	<b>0,117</b>	0,375	0,495
<b>ARX</b>	0,141	0,462	0,628
<b>MA</b>	0,120	0,384	0,504
<b>ARMA</b>	0,118	0,384	0,496
<b>ARIMA</b>	0,120	0,390	0,509
<b>SARIMA</b>	0,132	0,431	0,570
<b>ARIMAX</b>	0,140	0,459	0,627
<b>SARIMAX</b>	0,142	0,463	0,627
<b>LR</b>	5,663	25,466	25,466
<b>RFR</b>	0,189	0,583	0,759
<b>XGBRegressor</b>	0,239	0,754	0,918
<b>LGBMRegressor</b>	<b>0,174</b>	0,532	0,716

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 21: Comparação dos modelos com 30 dias de antecedência 24h **Completo**

<b>Modelos</b>	<b>Erros</b>		
	<b>MAPE</b>	<b>MAE</b>	<b>RMSE</b>
<b>AR</b>	0,114	0,367	0,237
<b>ARX</b>	0,137	0,447	0,360
<b>MA</b>	<b>0,113</b>	0,361	0,477
<b>ARMA</b>	0,120	0,385	0,508
<b>ARIMA</b>	0,117	0,375	0,497
<b>SARIMA</b>	0,124	0,404	0,531
<b>ARIMAX</b>	0,136	0,443	0,596
<b>SARIMAX</b>	0,137	0,446	0,601
<b>LR</b>	5,763	25,477	25,477
<b>RFR</b>	0,208	0,657	0,788
<b>XGBRegressor</b>	0,264	0,847	0,961
<b>LGBMRegressor</b>	<b>0,195</b>	0,610	0,746

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

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

Modelos ARIMAS para previsão de longo prazo usando a defasagem de 10.

Tabela 22: Comparação dos modelos Ljung Box **Treinamento**

<b>Ljung Box</b>	<b>Estatística de Teste</b>	<b>Valor De p</b>
<b>ARX</b>	<b>6,30</b>	0,79
<b>AR</b>	7,13	0,07
<b>MA</b>	34,34	0,00
<b>ARMA</b>	11,60	0,31
<b>ARIMA</b>	13,01	0,22
<b>SARIMA</b>	10,17	0,43
<b>ARIMAX</b>	30,36	0,00
<b>SARIMAX</b>	11,63	0,31

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 23: Comparação dos modelos Ljung Box **Validação**

Ljung Box	Estatística de Teste	Valor De p
<b>ARX</b>	7,47	0,68
<b>AR</b>	2,43	0,99
<b>MA</b>	1,39	1,00
<b>ARMA</b>	5,42	0,86
<b>ARIMA</b>	4,04	0,95
<b>SARIMA</b>	4,45	0,93
<b>ARIMAX</b>	<b>0,02</b>	1,00
<b>SARIMAX</b>	0,04	1,00

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Tabela 24: Comparação dos modelos Ljung Box **Teste**

Ljung Box	Estatística de Teste	Valor De p
<b>ARX</b>	0,86	1,00
<b>AR</b>	7,80	0,65
<b>MA</b>	7,89	0,64
<b>ARMA</b>	19,34	0,04
<b>ARIMA</b>	9,50	0,49
<b>SARIMA</b>	3,57	0,97
<b>ARIMAX</b>	<b>0,60</b>	1,00
<b>SARIMAX</b>	3,72	0,96

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

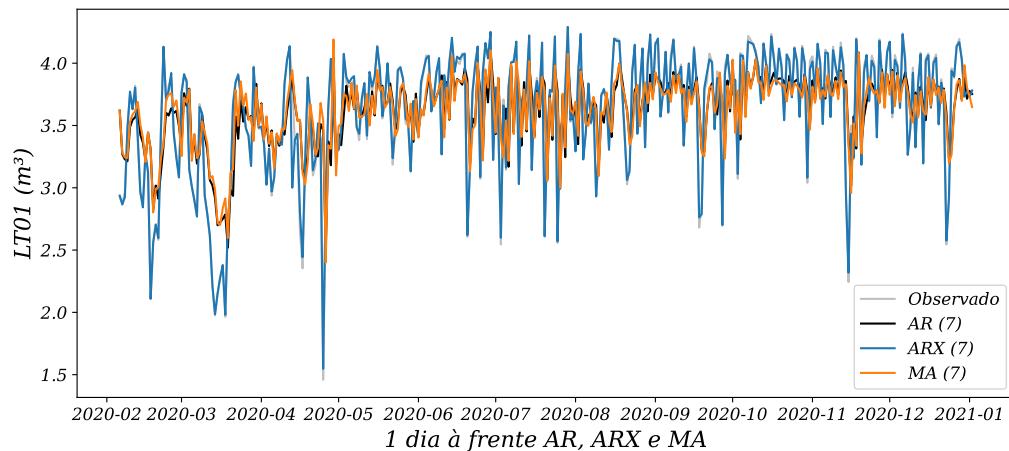
Tabela 25: Comparação dos modelos Ljung Box **Completo**

Ljung Box	Estatística de Teste	Valor De p
<b>ARX</b>	4,70	0,91
<b>AR</b>	<b>4,26</b>	0,16
<b>MA</b>	49,16	0,00
<b>ARMA</b>	40,49	0,00
<b>ARIMA</b>	40,49	0,00
<b>SARIMA</b>	40,49	0,00
<b>ARIMAX</b>	60,91	0,00
<b>SARIMAX</b>	5,83	0,83

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

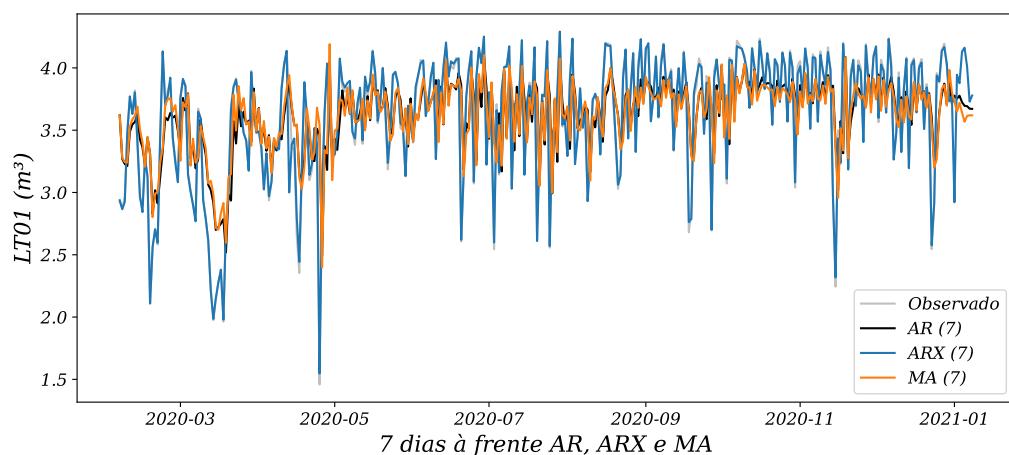
## C Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h

Figura 41: Comparação dos modelos AR, ARX e MA, 1 dia à frente



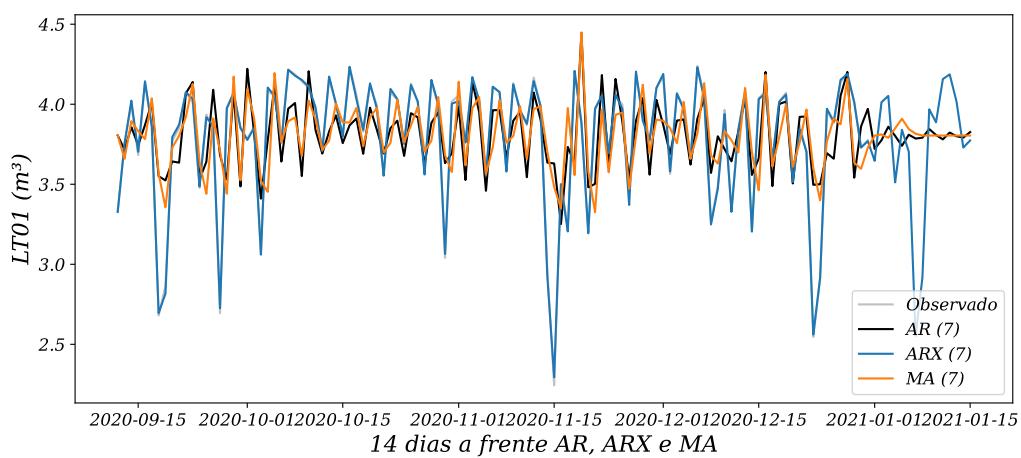
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 42: Comparação dos modelos AR, ARX e MA, 7 dias à frente



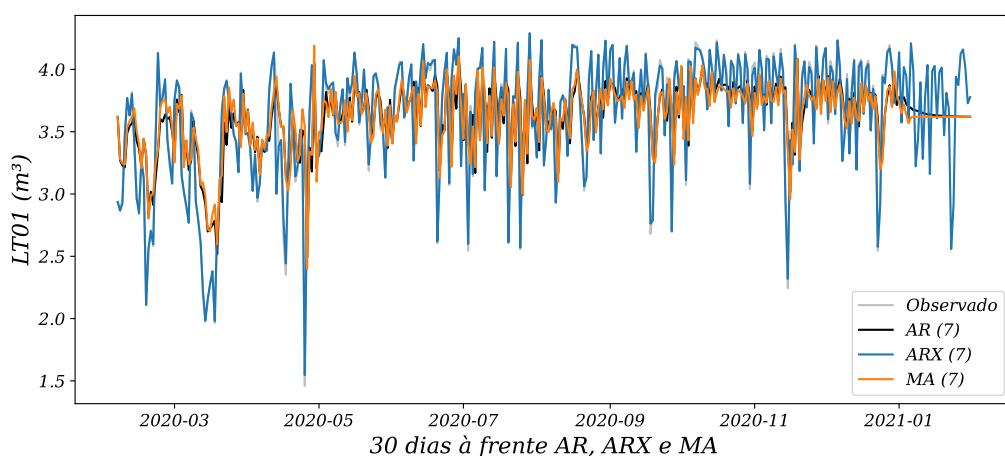
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 43: Comparação dos modelos AR, ARX e MA, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

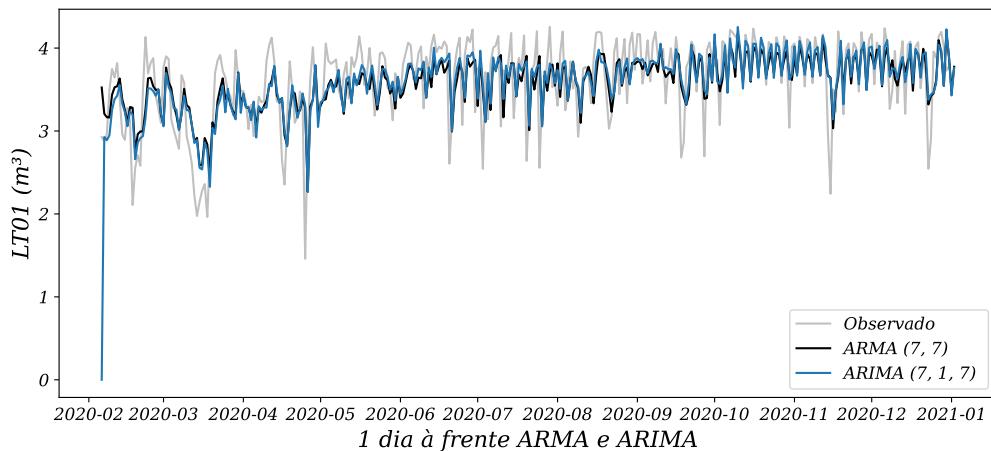
Figura 44: Comparação dos modelos AR, ARX e MA, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

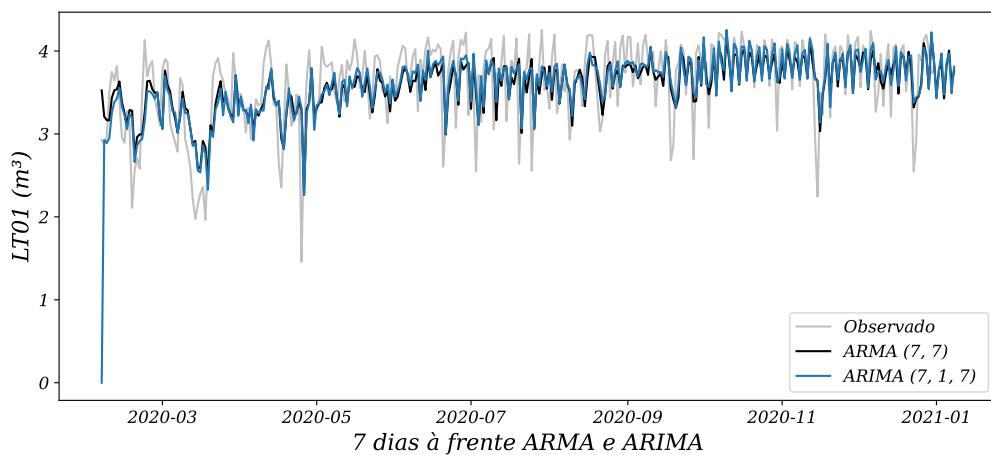
## D Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h

Figura 45: Comparação dos modelos ARMA e ARIMA, 1 dia à frente



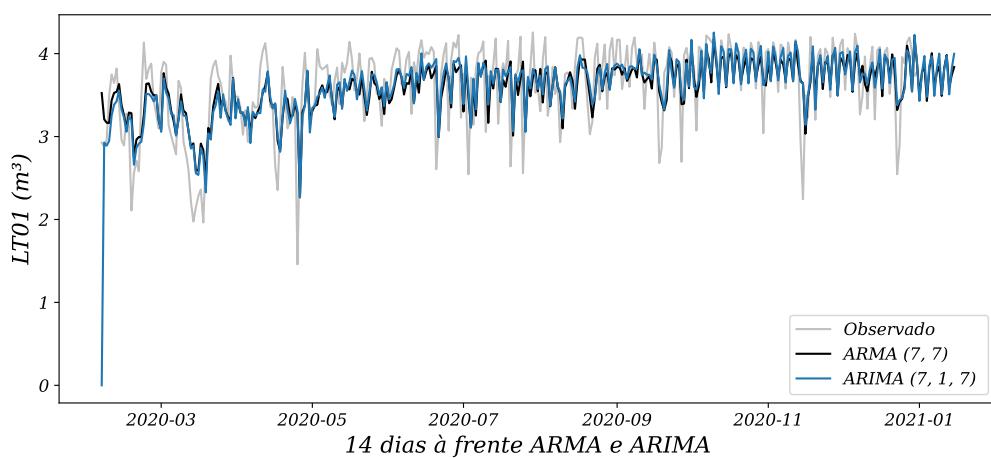
Fonte: Autoria própria.

Figura 46: Comparação dos modelos ARMA e ARIMA, 7 dias à frente



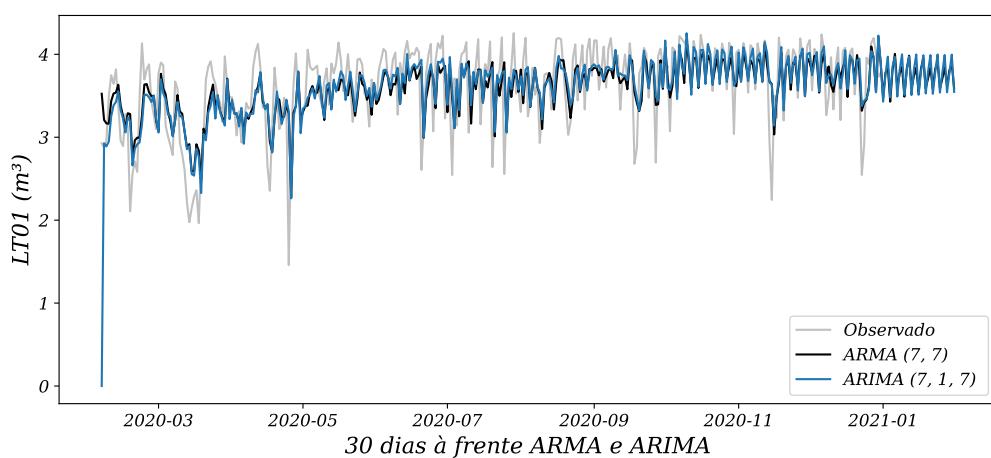
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 47: Comparação dos modelos ARMA e ARIMA, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

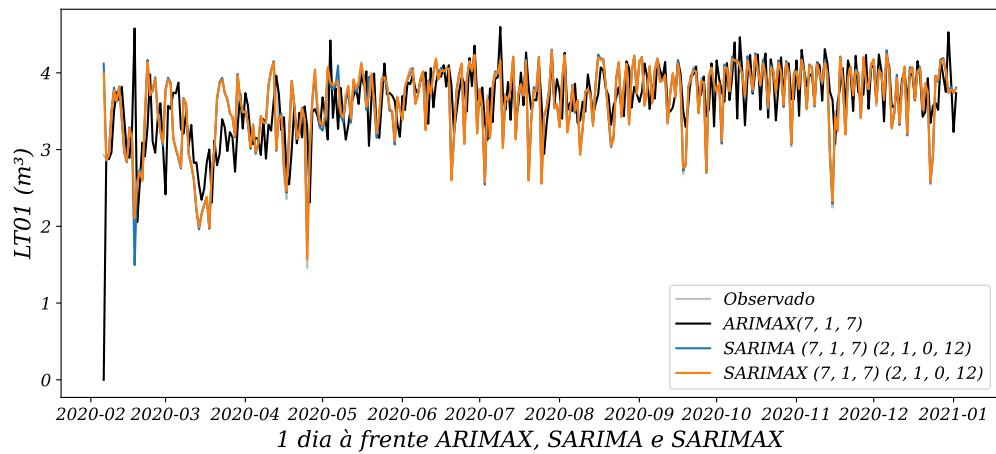
Figura 48: Comparação dos modelos ARMA e ARIMA, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

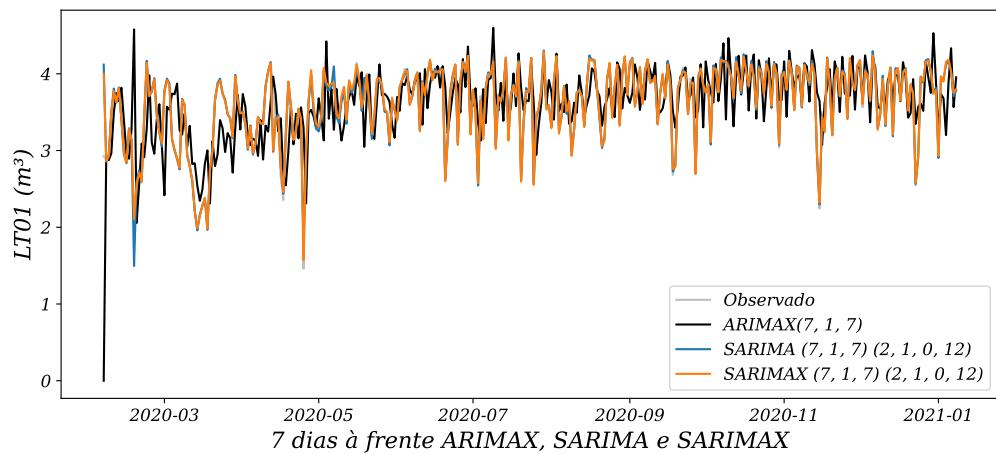
## E Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,0,12) e SARIMAX (7,1,7) (2,1,0,12) 24h

Figura 49: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia à frente



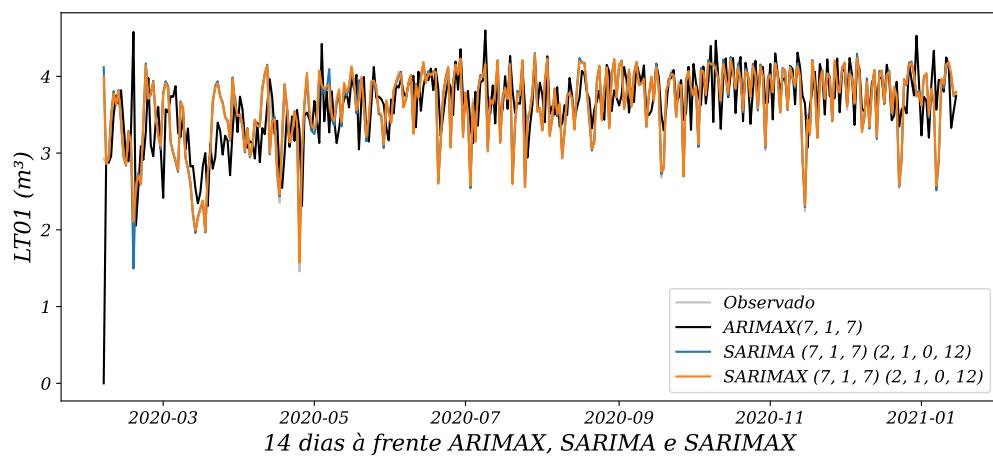
Fonte: Autoria própria.

Figura 50: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 7 dias à frente



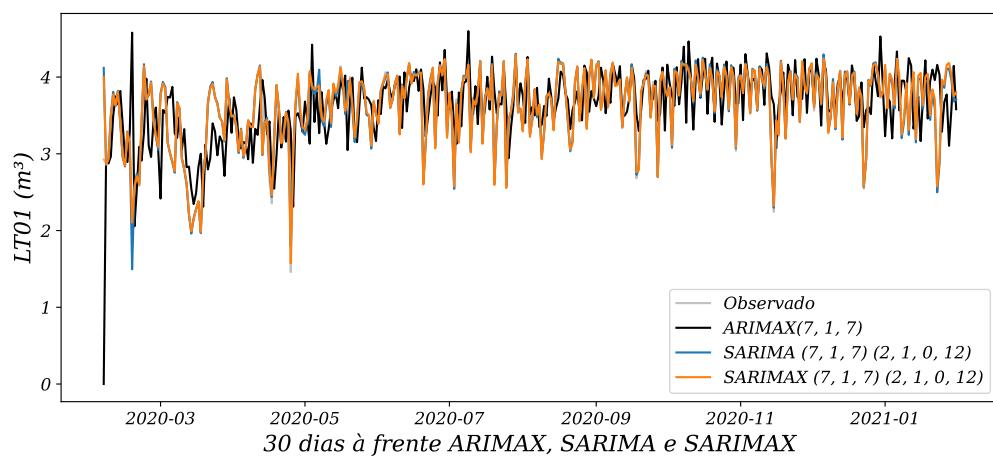
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 51: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 14 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Figura 52: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias à frente



Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)