



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 DEMANDA D'ÁGUA

CURITIBA
2022

FRANCHESCO SANCHES DOS SANTOS

PREVISÃO DE SÉRIES TEMPORAIS PARA DEMANDA D'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
2022

FRANCHESCO SANCHES DOS SANTOS

PREVISÃO DE SÉRIES TEMPORAIS PARA DEMANDA D'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, 2 de fevereiro 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 abriu o caminho para mim e me deu forças para superar esse desafio, sem ele nada seria possível.

À minha família, eles sempre me apoiaram e me incentivaram a seguir em frente com a cabeça erguida e buscar um estado mais elevado.

Ao Professor Leandro dos Santos Coelho, agradeço por me dar a oportunidade de trabalhar com ele e de compartilhar seu conhecimento e experiência ao longo do Mestrado, sempre em busca do meu crescimento profissional e pessoal que tornou este trabalho possível.

A Professora Viviana Cocco Mariani, obrigada pela disponibilidade e paciência em me ajudar com minhas deficiências e por utilizar seus conhecimentos para contribuir com o desenvolvimento da pesquisa.

Agradeço à equipe da Pontifícia Universidade Católica do Paraná (PUCPR) e demais professores, em especial a secretária Denise da Mata Medeiros (PPGEPS), por cuidar de mim com paciência e carinho e me ajudar inúmeras vezes, ao invés de medir o esforço despendido.

Aos meus amigos que estiveram torcendo, assim como aos novos amigos que fiz nesta caminhada, que proporcionaram grandes momentos de alegria na batalha.

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

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

Lista de Abreviaturas e Siglas

AdaBoost	Adaptive Boosting (Impulso ou Estímulo adaptativo)
AR	Auto-Regressivo
ARIMA	Média Móvel Integrada Auto-Regressiva (do inglês autoregressive integrated moving average)
ARX	Auto-Regressivo Exogedo
BrownBoost	Algoritmo de aumento
CNN	Rede Neural Convolucional
DBN	Rede de Crenças Profundas
FT	flow transmitter (Transmissor de fluxo)
Light GBM	Máquina de Impulso de Gradiente Leve (do inglês Light Gradient Boosting Machine)
LogitBoost	Representa uma aplicação de técnicas de regressão logísticas
LPBoost	Linear Programming Boosting (Reforço da Programação Linear)
LR	Regressão linear
LSTM	Memória de curto prazo
m^3	Metros cúbicos
m^3/h	Metros cúbicos por hora
MadaBoost	Modificando o sistema de ponderação da AdaBoost
MAE	Mean Absolute Error (Erro Médio Absoluto)
MAPE	Mean Absolute Percentage Error (Erro Percentual Médio Absoluto)
<i>mca</i>	Metros coluna d'água
MSE	Mean Squared Error (Erro médio quadrático)
RBAL	Recalque Bairro Alto
RMSE	Root Mean Squared Error (Erro de Raiz Média Quadrática)
RNN	Rede Neural Recorrente
SANEPAR	Companhia de Saneamento do Paraná
SARIMA	Auto-Regressivos Integrados de Médias Móveis com Sazonalidade (do inglês Integrated Auto-Regressive Moving Averages with Seasonality)
SARIMA	integrado autorregressivo e médias móveis com sazonalidade
SARIMAX	Média Móvel Integrada Auto-Regressiva Sazonal com regressores eXogenous (do inglês Seasonal Auto-Regressive Integrated Moving Average with eXogenous regressors)
SVM-VAR	Máquinas de vetor de suporte - Vetores Auto-Regressivos
Totalboost	Impulso total
XGboost	Impulso Gradiente Extremo (do inglês eXtreme Gradient Boosting)

Resumo

Previsão de séries temporais é muito importante para a tomada de decisão. Nessa dissertação será abordado o problema de demanda d'água que ocorreu na cidade de Curitiba no estado do Paraná, houve no período dos dados coletados nos anos de 2018 a 2020, visando o ano de 2020 que foi o ano que ocorreu a maior demanda d'água, fazendo com que os reservatórios sofressem com isso, vários fatores. Na tomada de decisão desse problema em questão é usado alguns métodos encontrado na revisão que foi realizado no decorrer desse trabalho, para ser previsto em alguns horizontes de previsão, os horizonte abordado aqui é uma forma que poder resolver a questão da demanda d'água e com isso validar os modelos para ver qual deles é o mais eficiente, horizonte adotado foi de previsão de 1, 10, 30 e 60 dias a frente, assim seja cada método vai lidar com os dados no decorrer do tempo. De modo a amenizar e solucionar o problema que a empresa SANEPAR enfrentou no ano de 2020, para que não ocorra mais ou que não pegue desprevenido no próximo evento que pode surgir. Com o evento isolado que aconteceu no ano em questão e não possa se repetir nos anos futuros, esse trabalho visa a melhoria do uso d'água. Os métodos derivado dos modelos ARIMA, assim listando os modelos é AR, ARX, MA, ARMA, ARIMA, SARIMA, SARIMAX e ARIMAX, como cada modelo tem sua particularidade os modelos de variáveis exógenas pode parecer graficamente melhor de ser previsto do que os modelos de ARIMA sem a variáveis exógenas. Nos modelos de reforço de gradiente é os melhores modelos para se prever com os erros mais baixo. Os modelos chamado de reforço ou árvore de regressão de gradiente, foi usado os seguintes modelos LR, regressão florestal aleatória XGboost e Light GBM, esses modelos para série temporal é listado como os melhores modelos, pois alguns deles usa a forma de prever de gradiente. É obtido em algumas métricas de erros, quanto menor o erro melhor para a tomada de decisão. As métricas adotado nesse trabalho é MAPE, MAE e RMSE, em série temporal essas métricas são mais frequente, com modelos de previsão melhor ou mais eficaz em algumas circunstâncias com na previsão de nenhum horizonte futuro, ou apensa a previsão da série obtida nos dados o modelo XGBoost tem erro de 0,079% na métrica MAPE só analisando encima dessa métrica, e o LR com o maior erro de 21% no horizonte maior de previsão (60 dias) o modelo ARMA vem com o erro de 12,54% e o modelo LR com 11153,594%. Assim o modelo LR para um conjunto de dados menor ele pode ser mais eficiente do que os outros modelos, pois trabalha com pouco volume de dados e os erros ficam mais alto conforme vai aumentando o horizonte.

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

Abstract

Time series forecasting is very important for decision making. In this dissertation the problem of water demand that occurred in the city of Curitiba in the state of Paraná will be addressed, there was a period of data collected in the years 2018 to 2020, aiming for the year 2020 that was the year that occurred the highest water demand, 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, 10, 30 and 60 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 does not happen again or that it is not caught off guard 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 the ARIMA models, thus listing the models are AR, ARX, MA, ARMA, ARIMA, SARIMA, SARIMAX and ARIMAX, as each model has its 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 0.079% error in the MAPE metric just analyzing over this metric, and the LR model has the largest error of 21% in the longest forecast horizon (60 days), the ARMA model comes with 12.54% error and the LR model with 11153.594% 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 Tabelas

1	Cruzamento de palavras chaves aplicando os filtros de ano e idioma.	16
2	Fator de impacto.	18
3	Áreas e seus valores respetivos de artigos em cada área.	21
4	Descrição Estatística dos dados com filtro aplicado de 18 a 21h	42
5	Teste Nemenyi	51
6	Comparação dos modelos 1 dia a frente 24h Treino	60
7	Comparação dos modelos 1 dia a frente 24h Validação	61
8	Comparação dos modelos 1 dia a frente 24h Teste	61
9	Comparação dos modelos 1 dia a frente 24h Completo	62
10	Comparação dos modelos 10 dia a frente 24h Treino	62
11	Comparação dos modelos 10 dia a frente 24h Validação	63
12	Comparação dos modelos 10 dia a frente 24h Teste	63
13	Comparação dos modelos 10 dia a frente 24h Completo	64
14	Comparação dos modelos 30 dia a frente 24h Treino	64
15	Comparação dos modelos 30 dia a frente 24h Validação	65
16	Comparação dos modelos 30 dia a frente 24h Teste	65
17	Comparação dos modelos 30 dia a frente 24h Completo	66
18	Comparação dos modelos 60 dia a frente 24h Treino	66
19	Comparação dos modelos 60 dia a frente 24h Validação	67
20	Comparação dos modelos 60 dia a frente 24h Teste	67
21	Comparação dos modelos 60 dia a frente 24h Completo	68

Lista de Figuras

1	Mapa das Etapas	4
2	Estrutura da dissertação	5
3	Dados completo em frequência de 24h em média	8
4	Plotagem dos dados do ano de 2020	8
5	Exemplo de séries temporais.	9
6	Processo estocástico.	10
7	Mapa conceitual do problema de pesquisa.	11
8	Etapas da Revisão.	12
9	Palavras-chave mais populares na Scopus.	14
10	Palavras-chave mais populares na WoS.	15
11	Analise das quantidades de artigos em relação aos anos.	16
12	Autores relação entre artigos publicados.	18
13	Acoplamento bibliográfico entre os autores	19
14	Mapa mundo da publicação dos artigos pelo mundo.	20
15	Áreas de aplicação do tema.	21
16	Modelo AR(7) com um passo a frente	26
17	ARX (7) com um passo a frente	26
18	Modelo MA(7) com um passo a frente	28
19	ARMA (7,7) com um passo a frente	29
20	ARIMA (7,1,7) com um passo a frente	30
21	SARIMA (7,1,7)(2,1,1) ₁₂	31
22	ARIMAX (7,1,7) com um passo a frente	32
23	SARIMAX (7,1,7)(2,1,1) ₁₂ com um passo a frente	32
24	Corelação de Pearson	33
25	Regressão linear LT01 vs PT01 correlação 98%	34
26	Regressão linear (LR) um passo a frente	35
27	Regressão da Floresta Aleatória (RFA) um passo a frente	36
28	Esquema da Floresta Aleatória	36
29	Impulsionando gradiente com XGBoost e LightGBM	37
30	Crescimento em folha versus crescimento em nível	40
31	Solução para acionamento das bombas	43
32	Decomposição STL aditiva dos dados coletados	44
33	Decomposição STL multiplicativa dos dados coletados	45
34	Histograma do nível do reservatório	46
35	Histograma da vazão de recalque	47

36	Autocorrelação e Autocorrelação parcial	48
37	Ruído branco	49
38	Comparação dos modelos AR, ARX e MA, 1 dia a frente	68
39	Comparação dos modelos AR, ARX e MA, 10 dias a frente	69
40	Comparação dos modelos AR, ARX e MA, 30 dias a frente	69
41	Comparação dos modelos AR, ARX e MA, 60 dias a frente	70
42	Comparação dos modelos ARMA e ARIMA, 1 dia a frente	70
43	Comparação dos modelos ARMA e ARIMA, 10 dias a frente	71
44	Comparação dos modelos ARMA e ARIMA, 30 dias a frente	71
45	Comparação dos modelos ARMA e ARIMA, 60 dias a frente	72
46	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia a frente	72
47	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 10 dias a frente	73
48	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias a frente	73
49	Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 60 dias a frente	74
50	Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 1 dia a frente	74
51	Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 10 dias a frente	75
52	Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 30 dias a frente	75
53	Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 60 dias a frente	76

Sumário

1	Introdução	1
1.1	Contexto da pesquisa	1
1.1.1	Motivação da pesquisa	1
1.2	Objetivo geral	2
1.2.1	Objetivos específicos e questão de pesquisa	2
1.3	Procedimentos metodológicos	3
1.3.1	Etapas da pesquisa	3
1.4	Justificativa da pesquisa	4
1.4.1	Contribuições	4
1.5	Estrutura do trabalho	5
2	Referencial	7
2.1	Descrição do problema	7
2.2	Revisão sistemática da literatura	9
2.3	Problematização da Revisão	11
2.4	Metodologia	12
2.5	Resultados da busca da revisão	14
2.6	Conclusão da revisão	23
3	Base Teórica	24
3.1	Métricas de Erros	24
3.2	ARIMA, SARIMA e SARIMAX	25
3.2.1	Componente auto-regressivo – AR(p)	25
3.2.2	Média Móvel – MA(q)	27
3.2.3	Modelos ARMA e ARIMA	28
3.2.4	MODELOS SARIMA, ARIMAX, SARIMAX	30
3.3	Modelos Regressivo	32
3.3.1	Regressão Linear (LR)	32
3.3.2	Floresta Aleatória	35
3.3.3	LightGBM e XGboost	37
3.3.4	O Gradiente em Gradiente de Boosting (Reforço)	37
3.3.5	Algoritmos de boosting de gradiente	38
3.3.6	A diferença entre XGBoost e LightGBM	39
4	Resultados	41
4.1	Planejamento do Problema	41

4.1.1	Análise Exploratória dos dados (EDA)	41
4.1.2	Múltiplas entradas e saída única (MISO)	43
4.1.3	Decomposição STL	44
4.1.4	Separação dos Dados	49
4.1.5	Estrategia de Previsão	50
4.1.6	Horizonte	50
4.1.7	Modelos de previsão e métricas de desempenho	50
4.1.8	Teste de Significância	51
5	Conclusões	53
5.1	Limitações da pesquisa e propostas futuras	53
Referências	55	
A	Apêndice - Comparaçao dos modelos de previsão de series temporais média de 24h	60
B	Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h	68
C	Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h	70
D	Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,1,12) e SARIMAX (7,1,7) (2,1,1,12) 24h	72
E	Apêndice - Modelos Regressão linear, XGB Regressão, Ligh GBM Regressão e Regressão de Floresta Aleatória 24h	74

1 Introdução

Este capítulo apresenta a introdução quanto ao que é abordado nesta dissertação, usando modelos de aprendizado de máquina (do inglês *machine learning*), dentro desses modelos vai ser abordar a previsão futura dos dados coletados na SANEPAR de Curitiba no estado do Paraná, esses dados foram coletado no bairro alto nos anos de 2018 até 2020 houve uma falta de água que afetou todos em Curitiba.

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 máquina é frequentemente utilizado para processamento de big data, com o conjunto de dados da SANEPAR em Curitiba - PR, na cidade há algum consumo e escassez de água, é necessário avaliar os dados para ter certeza do que está acontecendo, quando há escassez d'água, e picos que ocorrem entre horas e dias.

Dentre 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 d'água. Estas análises será em *python*.

Explorar o que são séries temporais e aprendizado de máquina, séries temporais são dados armazenados ao longo do tempo que permitem ao observador analisar anomalias nos dados. Em séries temporais, ordenar os dados por ano ou dia é fundamental e, se os dados atribuídos de forma aleatória, assim podendo tornar mais difícil prever e tomar decisão baseado nos dados coletados. Analisar médias pode ser bem perigoso se não excluir pontos fora da curva também conhecidos como *outliers*. Pode gerar dados muito positivos ou negativos que não correspondem a 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 Instituto Nacional de Meteorologia (Inmet). Apesar de abaixo da média, o mínimo registrado desde então ocorreu em 2020, com 1.158 mm.

Em mediano a essa motivação pode ser melhor interpretado os dados que a SANEPAR ofertou para prever e evitar a escassez de água que foi registrada, e a anomalia que foi detectada em 2020, com a volta da chuva os reservatórios teve aumento do nível.

1.2 Objetivo geral

Objetivo para essa dissertação é encontrar o melhor modelo de séries temporais para o problema de falta d'água que houve em Curitiba. Com vários modelos coletados no decorrer da dissertação, entre modelos de regressão e aplicado ao gradiente os modelos *boosting* na literatura os melhores modelos de previsão de série temporal, e os modelos ARIMA e os ARIMA atualizado. Prevendo e analisando as anomalias nos dados, e porquê ocorrer.

1.2.1 Objetivos específicos e questão de pesquisa

Para esse trabalho é pretendente-se busca as anomalias que pode ocorrer nos dados, é porque acontece tais anomalias e responder as questões de 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 às 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 Procedimentos metodológicos

Nessa parte vai ser abordado como será conduzido a dissertação, cada etapa que foi realizado no decorrer das análises feitas.

1.3.1 Etapas da pesquisa

A pesquisa se deu seguindo as seguintes etapas:

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

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

Etapa 3 Fazer a decomposição STL (do inglês *Seasonal-Trend Decomposition*) Sazonalidade, Tendência e Resíduo

Etapa 4 Divisão do conjunto de dados em treinamento, validação e teste 70% para treinamento e validação e 30% para teste, disso tirando os 70% e dividindo em 80% para treinamento e 20% para validação. Verificar a média e desvio padrão de cada um destes conjuntos de forma que obtenha a divisão mais adequada dos dados.

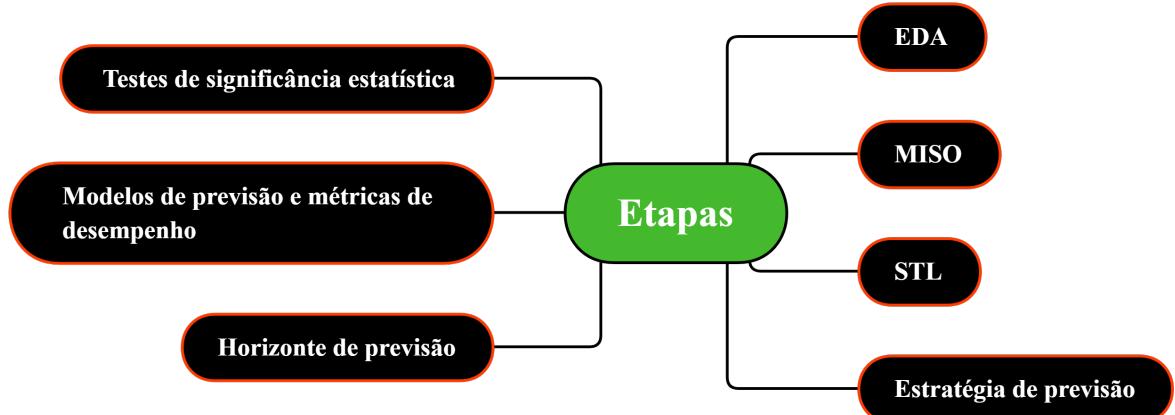
Etapa 5 Estratégia de previsão (recursiva e iterada-método direto)

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

Etapa 7 Modelos de previsão e métricas de desempenho

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

Figura 1: Mapa das Etapas



Fonte: Elaboração própria

1.4 Justificativa da pesquisa

No decorrer dessa dissertação ocorre da seguinte forma, para que possa ser previsto e para que seja evitado a efetiva falta d'água, e como pode ser solucionado esse problema para não voltar a acontecer.

1.4.1 Contribuições

Seguindo as questões de pesquisa feito na subseção 1.2.1 tem duas contribuições, a primeira levando em conta a demanda d'água na cidade de Curitiba, entre a **Q 1** a **Q 4** é feito a previsão da demanda d'água, as outras ficam em como é o consumo d'água na cidade e gasto com energia no período de pico, mostrado na **Q 5a.** a **Q 5e..**

Assim usando os métodos escolhido de previsão de series temporais, como os modelos ARIMA e ARIMA atualizado, como os modelos ARMA, SARIMA, ARIMAX e SARIMAX, outros modelos mais simples que vem do modelo ARIMA, como, por exemplo, os modelos AR, ARX e MA para previsão mais precisa como na **Q 5** em diante os modelos regressivo ou modelos de gradiente, modelos regressivo testado aqui foi os modelos LR e floresta aleatória, para os modelos de gradiente foi usado XGBoost e Light GBM se torna uma opção mais viável na hora de tomar a decisão em meio aos gastos de energia e água que a empresa SANEPAR teve e com o intuito de minimizar esses gasto.

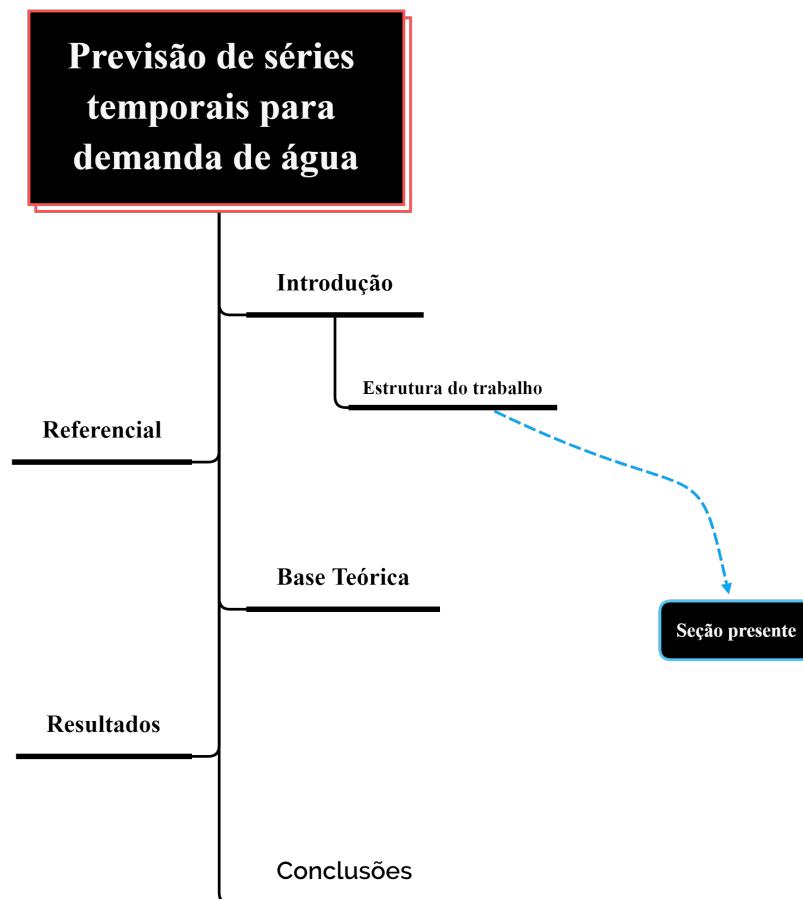
Foi estabelecido os horizonte de previsão para que possa ser tomado a melhor decisão a respeito da demanda d'água.

Em ambas das contribuições foi realizado o tabelamento tanto em curto prazo (1 a 30 dias, um mês) até longo prazo (30 a 60 dias, dois meses). Para que assim o melhor modelo tanto em curto quanto em longo prazo seja mostrado e evidenciado. Os modelos ARIMA para o problema em questão em horizonte de previsão de longo prazo se sai melhor que os modelos de reforço de gradiente, modelos de gradiente é mais viável em previsão de curto prazo, por exemplo de 1 dia a frente até uma semana. E ainda sim os modelos ARIMA ou que os modelos que se vem dele supera os gradiente.

1.5 Estrutura do trabalho

Este trabalho 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,

motivação, objetivo geral, os objetivos específicos, a metodologia utilizada, a justificativa da pesquisa, Contribuições, Publicações e a organização do trabalho. O capítulo 2 apresenta a descrição do problema, revisão teórica do trabalho, fazendo um apanhado geral dos principais pesquisadores nos temas abordados na pesquisa. O capítulo 3 apresenta os modelos que será trabalhado nos dados coletado. O capítulo 4 apresenta os resultados da pesquisa, bem como uma análise dos resultados gerado. O capítulo 5, por fim, apresenta as considerações finais da pesquisa e algumas propostas de pesquisas futuras.

2 Referencial

Nesse capítulo vai ser exposto a base da literatura que foi coletado durante a elaboração dessa dissertação, mesmo com um tanto de resultado mais baixo do que em uma tese, ainda é relevante para o trabalho que foi realizado aqui.

2.1 Descrição do problema

Nessa subseção vai ser abordado as variáveis do conjunto de dados e como vai ser previsto.

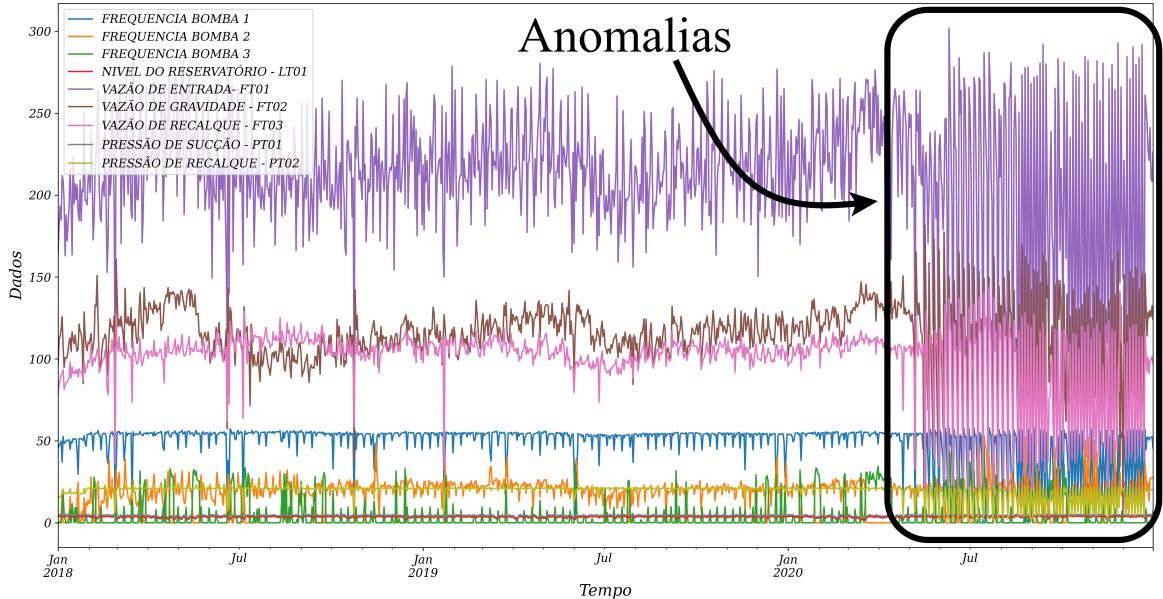
- Bombas de sucção (B1, B2 e B3) – valor máximo da frequência 60 Hz

Variáveis importantes: Vazão, 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)

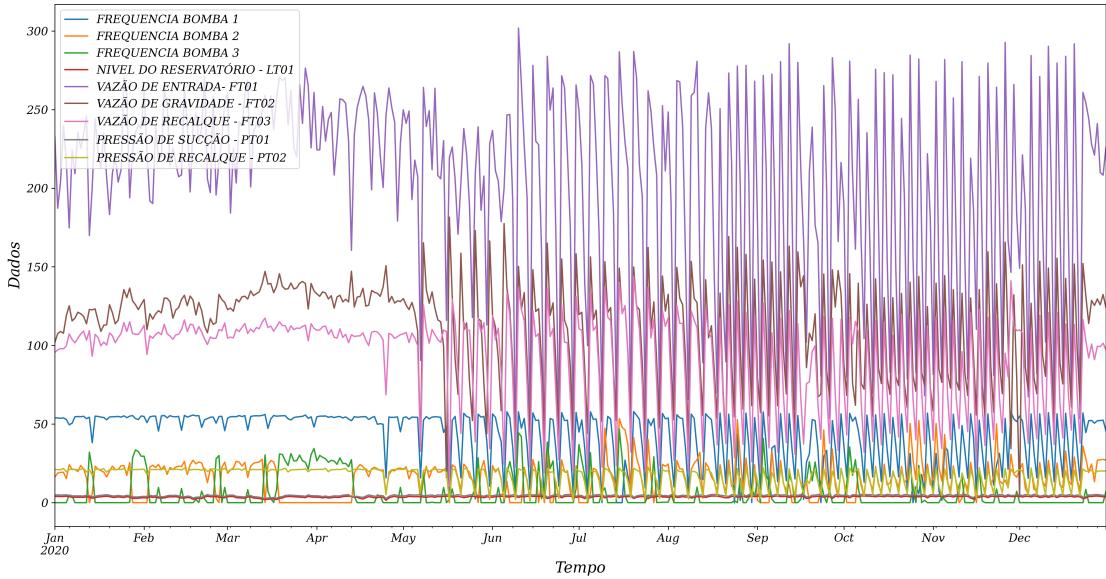
Na pesquisa vai ser usado a variável LT01 que é o nível do reservatório, esse nível é de grande importância, como visto nas Figuras 3 e 4

Figura 3: Dados completo em frequência de 24h em média



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

Figura 4: Plotagem dos dados do ano de 2020



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

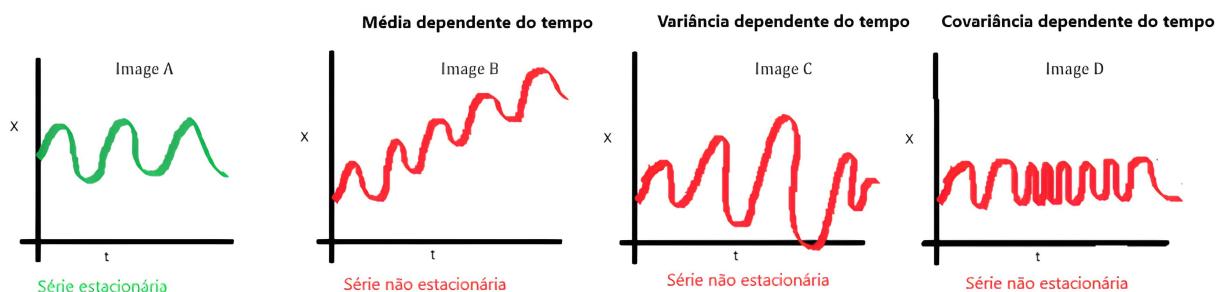
Os dados coleatos tem o tamanho de 26306 linhas \times 9 colunas, para tanta relação que vai ser usado nos modelos da subseção 1.3 para prever e analisar as anomalias, como apresentado nas Figuras 3 e 4.

2.2 Revisão sistemática da literatura

Séries temporais (time series) surge em vários campos do conhecimento como Economia (preços diários de ações, taxa mensal de desemprego, produção industrial), Medicina (eletrocardiograma, eletroencefalograma), Epidemiologia (número mensal de novos casos de meningite), Meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), etc. No decorrer dos anos usa ferramentas computacionais para fazer essa previsão mais eficiente, com o aprendizado de máquina e alguns recursos que pode ser aplicado em linguagem computacional por meio da linguagem *python* e *R* as melhores linguagens para se trabalhar com séries temporais atualmente.

Para entender melhor esse conceito de séries temporais, vamos supor que um maratonista que corre a vários anos e uma pessoa sedentária, seja submetido a uma corrida de no máximo 5 km, ambos saem ao mesmo instante de forma que eles tenha um medidor de batimento cardíco para que possa ser monitorado pelos médicos, se pegar os dados do começo e comparar com o final da prova o maratonista irá estar com uma série mais estacionaria, pois ele tem o hábito de correr regularmente, em quanto isso a pessoa sedentária vai ter uma série não estacionária como mostrado na Figura 5.

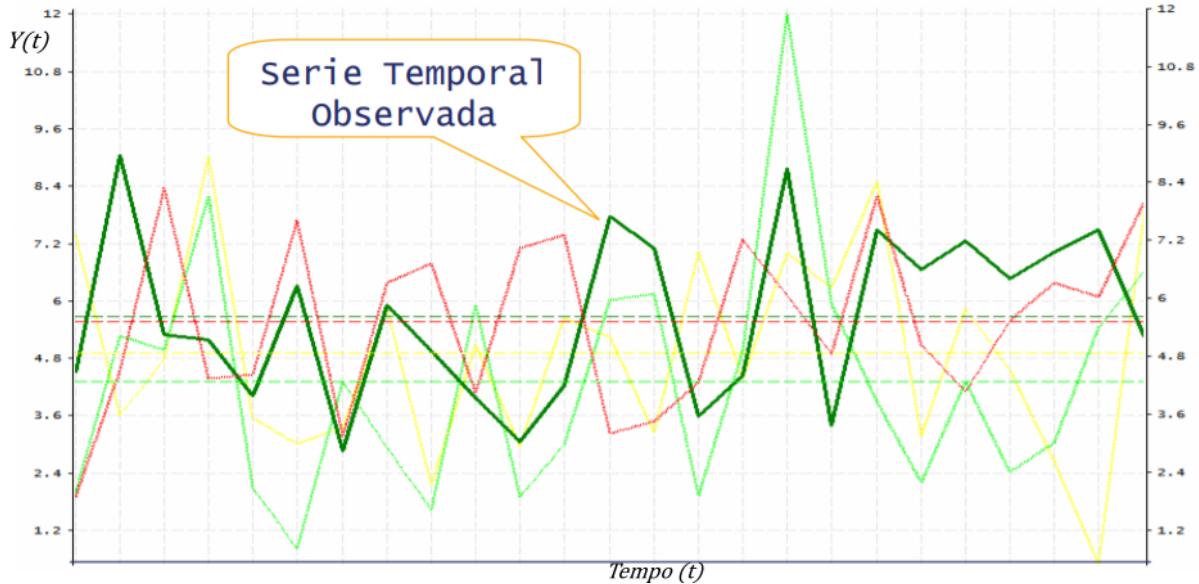
Figura 5: Exemplo de séries temporais.



Fonte: (BRANDÃO, 2020)

Na Figura 5 observar-se que o eixo x significa os dados observados e t o tempo percorrido. Além disso, séries temporais são processos estocásticos por leis probabilísticas que significa há a possibilidade de ser pensando como um conjunto de todas as possíveis trajetórias, na Figura 6 é capaz de ser observadas para uma variável alvo. Por exemplo, se você jogar um dado qualquer valor inteiro entre 1 e 6, mas apenas um número vai ocorrer. Da mesma forma, nas séries temporais existem infinitas possibilidades, dentre elas apenas uma conforme as características que assistiram naquele período e que de fato vai ocorrer.

Figura 6: Processo estocástico.



Fonte: (PINHEIRO, 2022)

Com $Y(t)$ sendo os dados fictícios e $\text{Tempo} (t)$ a linha do tempo da Figura 6.

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

Essa revisão sistemática da literatura, com o tema abordado até o momento é sobre série temporal, considerando o contexto exposto aqui, esse tema pode ser de grande relevância em várias áreas tais como mostra na Figura 15. Realizar essa análise de série temporal ao longo de 6 últimos anos para poder observar os melhores feitos nesse tema, aborda aqui um curto período, mas tendo o tempo não muito a favor por isso tive a escolha de deixar esse tempo específico de busca de artigo.

Para essa revisão tem com objetivo a análise de uma literatura menor, porém bem relevante. Como a própria série temporal procura analisar e modelar dependência, e considerando a ordem apresentada nas bases, por exemplo, os maiores autores e o ano de atuação que eles, mais publicou nos países que tem o maior número de publicação, na apresentação das palavras chaves que será mostrado, o objetivo estar em rever cada coisa que pode ser usada em uma aplicação de aprendizado de máquina.

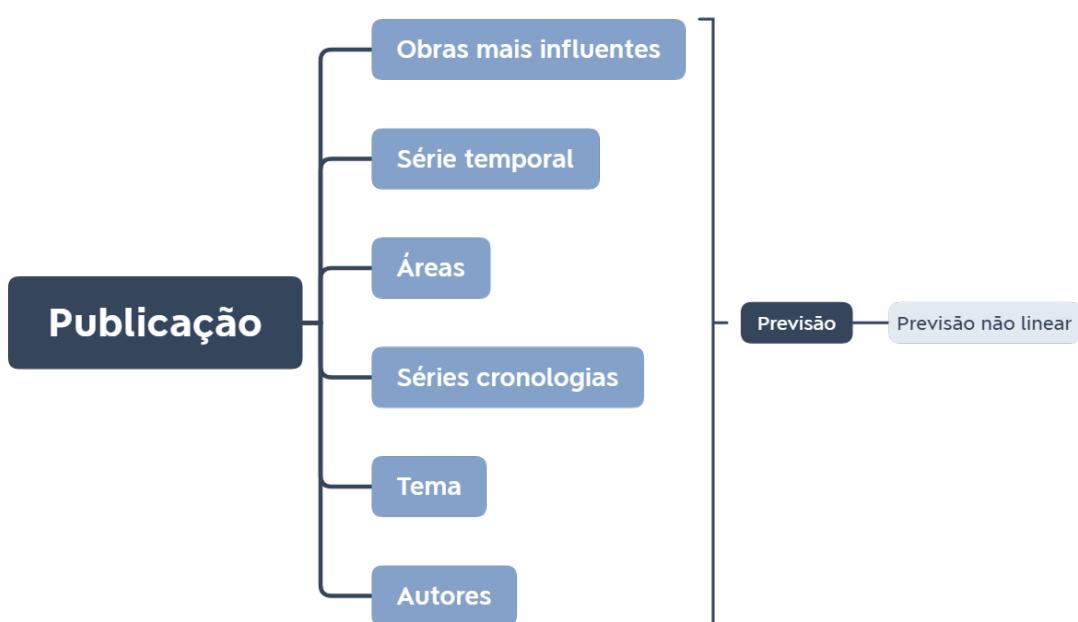
Em todos os artigos observar-se que tem uma contribuição científica, nesse trabalho é a análise de conceito de séries temporais com o melhor aproveitamento das palavras chaves, mesmo não tendo um grande relacionamento em aprendizado de máquina pode ser usado esses artigos como base para outros pesquisadores, sendo aqui algumas análises bem simples para alguns leitores. Entretanto é um ponto de partida para muitos que não

conhece o conceito de série temporal ou revisão sistemática da literatura.

2.3 Problematização da Revisão

Nessa seção é abordado um problema de pesquisa que pode ser entendido por vários leitores, na Figura 7 é apresentado um mapa conceitual de publicação e os autores são o pilar mais relevante para a revisão pois eles apresenta vários modelos que servira de base, e como está falando de série temporal a previsão que pode ser realizado nesse contexto é uma problemática devidamente de grande significância.

Figura 7: Mapa conceitual do problema de pesquisa.



Fonte: Elaboração própria

No mapa conceitual apresentado na Figura 7 é visto a problemática sendo relacionada com palavras, deixando evidente o que vai ser abordar no decorrer do trabalho deixando as questão de pesquisa em tópicos, logo adiante.

Q 1 Quais os autores que mais pública sobre o assunto de série temporal?

Q 2 Quais os países que mais pública sobre o assunto?

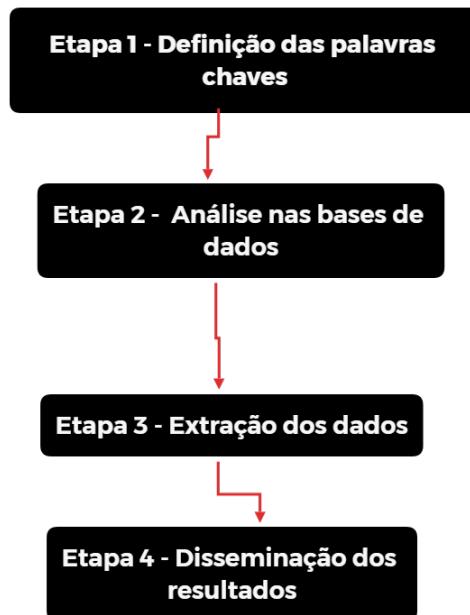
Q 3 Quais as áreas que mais pública sobre o tema?

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

2.4 Metodologia

Nessa seção é esclarecido como foi conduzindo a revisão, desde análise das bases de dados até como concluir a revisão.

Figura 8: Etapas da Revisão.



Fonte: Adaptado de Martins e Gorschek (2016)

Etapa 1 Na Figura 8 usa uma adaptação de Martins e Gorschek (2016) para essa revisão sistemática que está sendo analisado. Logo mais tem as buscas nas bases da Scopus, Web of Science e Lens. A princípio foi usado algumas base no meio de tantas na literatura para melhor atende no tema da pesquisa.

Scopus campo de busca

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))

Web of Science campo de busca

"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)

Lens campo de busca

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 realizado nos últimos 6 anos, apenas no site do lens que optou-se para colocar 6 anos, pois nos retornou poucos artigos. Nessa etapa é usado as palavras chaves que mais se adéquam na pesquisa *time series forecasting and time series analysis and nonlinear forecasting*.

Etapa 2 No cruzamento de palavras obter um número considerável de artigos sem restringir a área que cada artigo pode estar publicado. Na Tabela 1 foi realizado um tabelamento dos resultados obtidos sem excluir a duplicada, isso vamos tratar na seção 2.5.

Etapa 3 Essa etapa serve para avaliar cada dado que obter sem nenhum filtro no começo da pesquisa, a extração desses dados sem usar nenhum filtro de ano nas buscas, ficaria muitos artigos para analisar, como por exemplo, na base de dados da Scopus ficaria com 498 artigos, na Web of Science ficaria com 140 artigos, e no Lens como não retornara muitos artigos, fica com 11 dando em um total de 649 sem remover duplicada. É certo lembrar que nesses artigos tem somente o filtro do idioma inglês e de artigo, para melhora a busca e tomada de decisão usando o filtro de anos, nos últimos 6 anos é um valor de artigos mais agradável de ser utilizado com pouco tempo para analisar, e usar a diferença entre essa estimativa que foi realizado na Tabela 1 são menos de 356 artigos para analisar. Lembrando que se foi feito a remoção dos duplicados esse número que foi obtido no resultado das bases todas pode chegar a um número menos ainda do que é pretendido nesse trabalho.

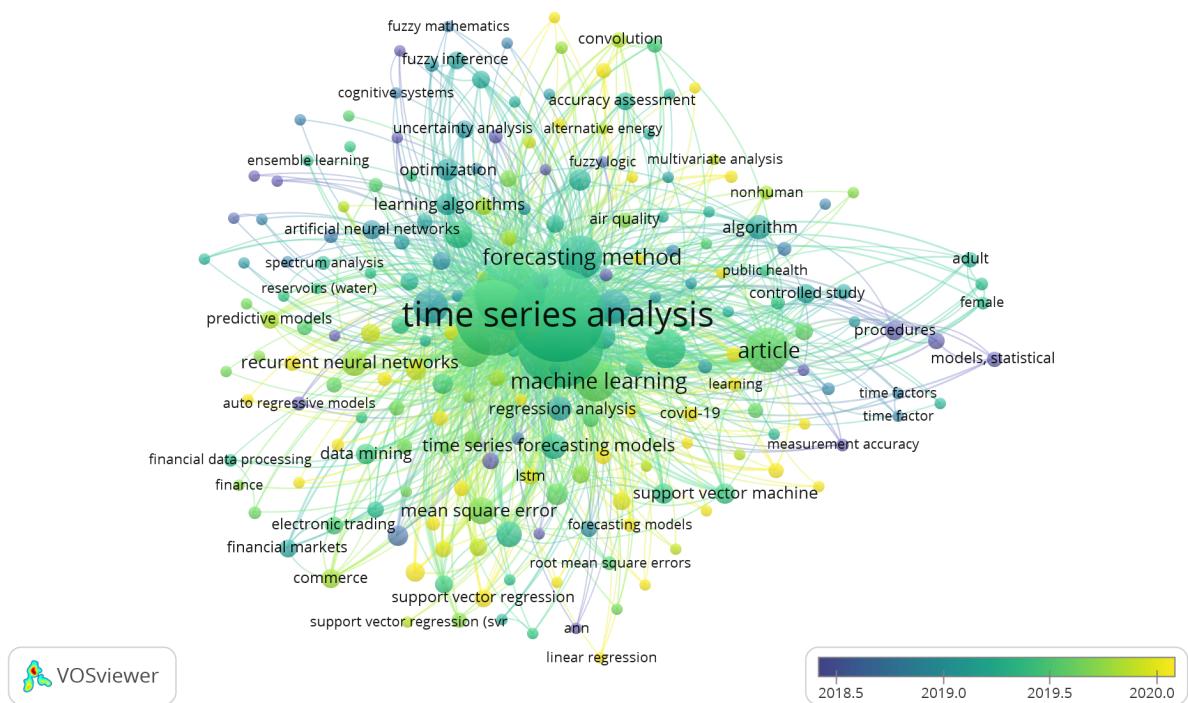
Etapa 4 Nessa etapa é mais para analisar a dimensão do que está sendo trabalhado, fazendo a análise das áreas e ler os artigos que são realmente importantes para a revisão. Como essa revisão é voltado a séries temporais em um programa de mestrado de engenharia de produção e sistemas é valido analisar a correlação. Dessa forma um das áreas é voltado a matemática assim sendo selecionado nesses artigos que pode ter um resultado de uma análise mais profunda dos artigos de séries temporais, se olhar para as áreas de atuação dos artigos pesquisados pode ser visto na Figura 15 que as áreas que foi citado aqui com grande relevância é **informática, engenharia e matemática** tem um número de publicação bem elevado, representando 50% da busca, então a pesquisa esta no caminho certo, usando a matemática básica para ter

uma estimativa de quantos artigos pode ser eliminado seria por volta de 481 artigos, mas isso sem muito fundamento de que esse número tenha uma precisão. Usando o *software mendeley desktop* para estipular o valor exato de quantos artigos usar, sem duplicado fica com um número de 308 artigos.

2.5 Resultados da busca da revisão

Nessa seção vai ser apresentado os resultados da pesquisa usando alguns *softwares* para conseguir estipular o melhor aproveitamento de cada base de dados usado no decorrer do trabalho. Dessa forma pode começar com a análise no *software VOSviewer*

Figura 9: Palavras-chave mais populares na Scopus.



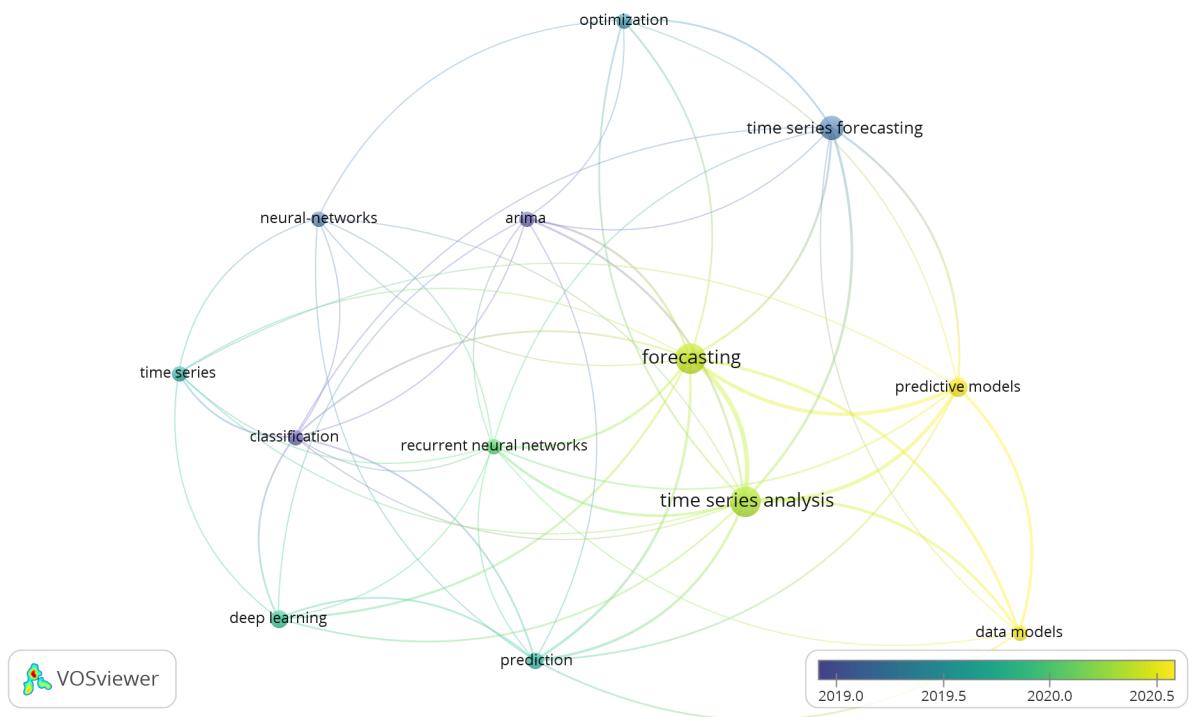
Fonte: Elaboração própria a partir de dados da Scopus (2016 a 2022)

Na Figura 9 é uma relação das palavras mais utilizadas como sinônimos da palavra *time series analysis*, ou em conjunto no corpo do texto dos artigos. A análise da base de dados na scopus foi feita na ferramenta que mostra as palavras chaves que pode ser relacionado em todo campo de pesquisa, com isso tem uma ampla visão do que pode ter correlação com as palavras-chave mãe da pesquisa.

Na relação entre as palavras chaves nesse primeiro momento, obteve um resultado de 3484 palavras-chave, 212 cumprem o limiar, lembrando que as palavras de base para

resultar foi *time series forecasting and time series analysis* na 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 da base de dados na Web of Science foi feita na ferramenta que mostra as palavras chaves que estão relacionadas em todo campo de pesquisa, com isso pode ter uma ampla visão do que tem correlação com as palavras-chave mãe da pesquisa.

Na relação entre as palavras chaves nesse primeiro momento, teve um resultado de 305 palavras-chave, 13 cumprem o limiar, lembrando que as palavras de base para resultar foi *time series forecasting and time series analysis* na web of science.

A única base de dados que não será mostrado aqui é a base da Lens, pois a mesma sendo uma base ótima ainda não teve tanto retorno na pesquisa que foi feito. O site lens retornou apenas 11 artigos com os filtros aplicados. Na **Etapa 1** é observado o campo de busca que foi utilizado nessa pesquisa que deu 11 artigos somente.

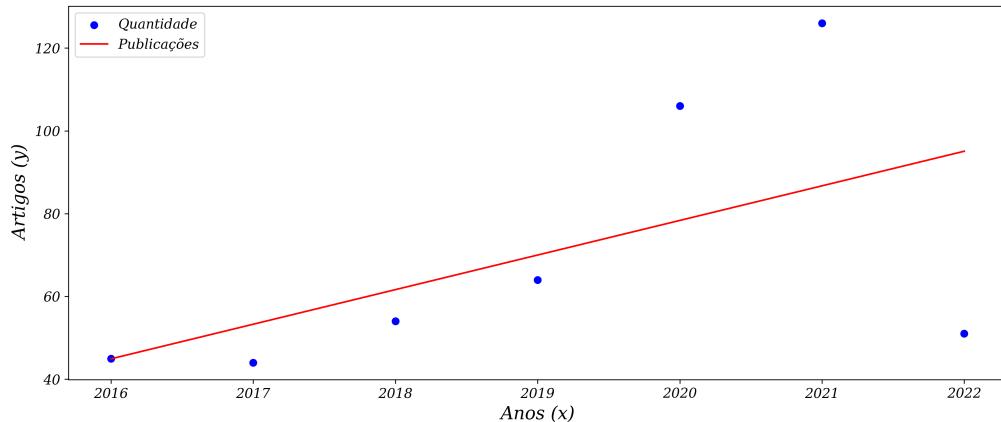
Na Tabela 1 relaciona as palavras chaves para cada base e aumenta a quantidade de artigo em todas as bases, mas essa Tabela está com os dados brutos que não foi eliminado os duplicados, então usando o software mendeley para remoção dos duplicados retorna apenas 308 artigos.

Tabela 1: Cruzamento de palavras chaves aplicando os filtros de ano e idioma.

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	forecasting	AND	time series		14
	forecasting		forecasting		
time series	forecasting	AND	time series	AND	11
	forecasting		analysis	nonlinear	
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 abscissa e ordenada, anos e artigos sendo assim a relação entre a data de publicação dos artigos no decorrer do tempo.

Número considerável de artigo para analisar, na Figura 11 foi feito uma análise baseado em uma regressão linear dos artigos em decorrer dos anos desde 2016 até 2022 nessa análise obteve a seguinte equação de regressão linear:

$$y(x) = 8,3571x - 16803 \quad \text{Com } R^2 = 0,3062 \quad (1)$$

Sendo $y(x)$ a equação da reta na equação (1). 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 estatisticamente atribuída ao conhecimento de uma ou mais variáveis independentes Quinino, Reis e Bessegato (1991).

O coeficiente de determinação mensura a relação existente entre a variável dependente e as variáveis independentes, indicando qual o percentual de variação explicada pela regressão, representa da variação total. Quando:

$R^2 = 1$: todos os pontos observados se situam exatamente sobre a reta de regressão (ajuste perfeito), ou seja, as variações de y são 100% explicadas pela variação dos x_n através da função especificada, não havendo 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 informação alguma sobre as variações de y .

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

Na equação (2) X, Y é dado pelas coordenadas no plano cartesiano, como por exemplo o par ordenado (x, y) . Na equação (1) observa-se que obteve o $R^2 = 30\%$ isso acarreta que a reta de regressão será influenciada pelo R^2 que foi achado.

Apesar de ser uma análise bem simples que foi realizada com a relação entre quantidade de artigos e anos, ainda sim é uma ótima validação de se olhar no teste de significância F que é dado a significância tem que estar sempre $F < 5\%$ esse teste também é chamado valor-p (p-value).

Tendo esses valores pode ser analisado a extrema significância, na reta de regressão observar-se que em 2021 foi o ano que mais foi publicado artigos com esse tema de séries temporais, essa análise pode nos trazer o pico maior de publicação feito.

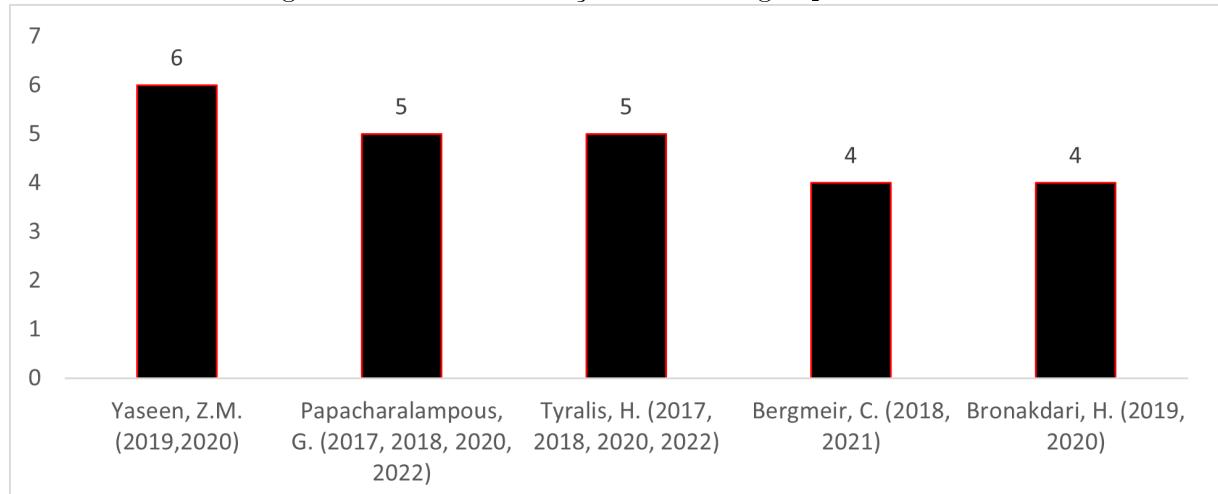
Tabela 2: Fator de impacto.

Revista científica	Quantidade de publicação	Qualidade da revista	H-INDEX
Neurocomputing	27	A1	143
IEEE Access	18	A1	127
Applied Soft Computing	12	A1	143
Energies	11	A2	93
Energy	11	A1	343

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

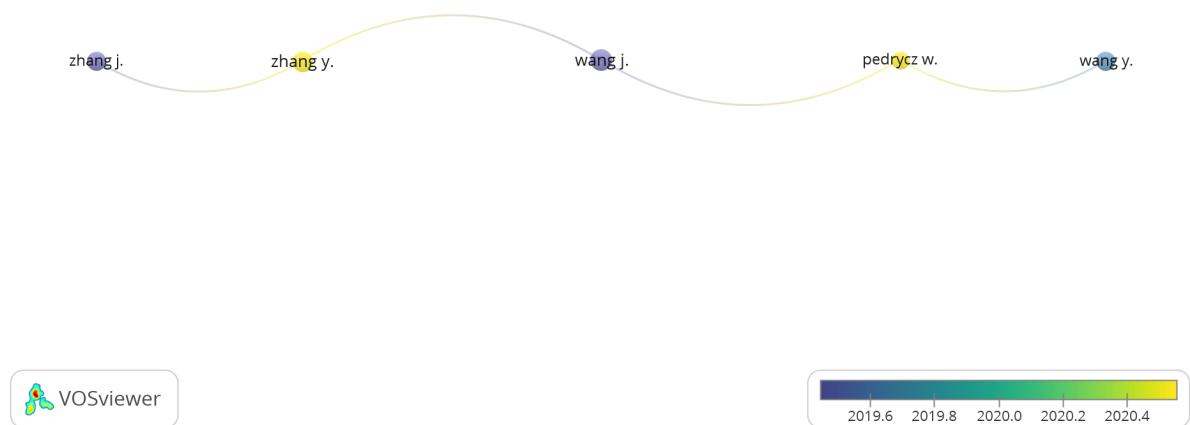
Na Tabela 2 mostra algumas revistas que mais publica, artigos nesse tema, como muita revista não se localiza no Brasil tem o nome em inglês, mas todas as revistas com um fator de impacto bem elevado como **A1** tem uma correlação com as áreas de **informática, engenharia e matemática**.

Figura 12: Autores relação entre artigos publicados.



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

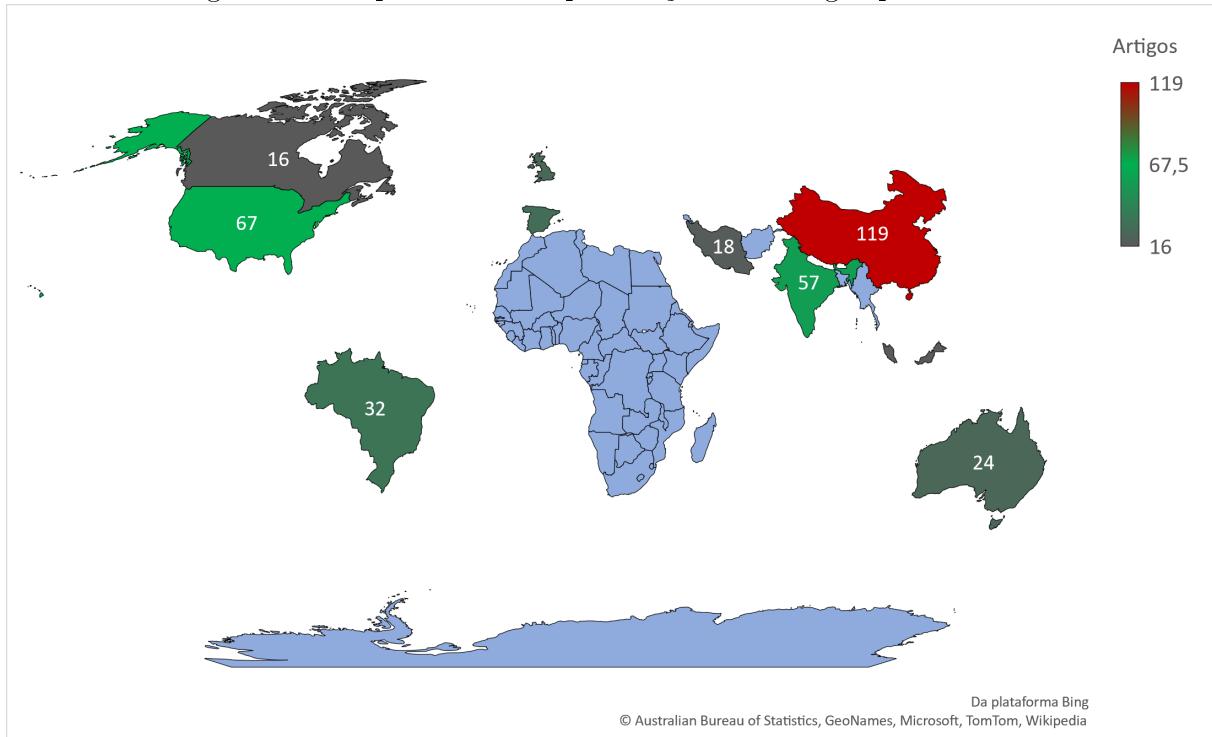
Figura 13: Acoplamento bibliográfico entre os autores



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

Respondendo um problema de questão feito aqui a **Q 1** utiliza a Figura 12 com um gráfico de histograma, como que fique mais visível os autores que mais publica nesse tema, no gráfico coloca os autores que tiveram publicação maior que 4, e com isso não coloca todos os autores, levando em consideração os autores que publicaram acima de 4 artigos nesse tema de 2016 até 2022.

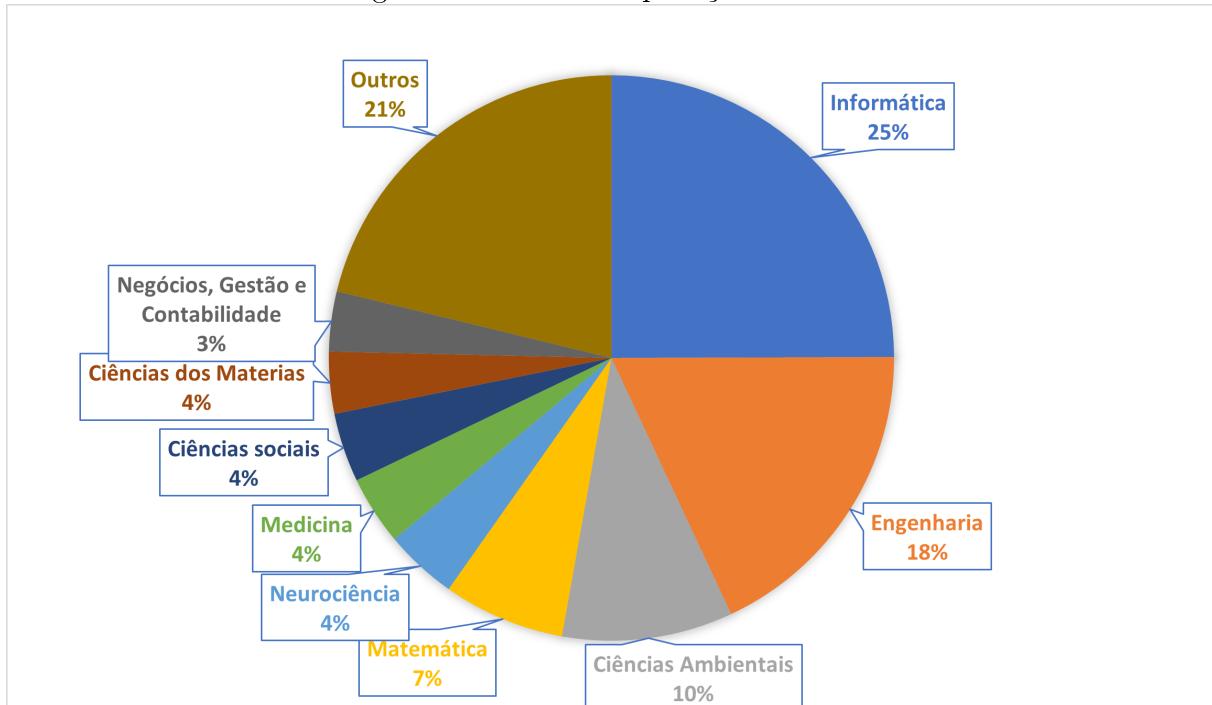
Figura 14: Mapa mundo da publicação dos artigos pelo mundo.



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

Na questão de pesquisa **Q 2** é respondida com a Figura 14, os países que mais publica sobre o assunto, em escala de maior publicação para o menor em escalar da seguinte forma China - 119, Estados Unidos - 67, Índia - 57, Brasil - 32, Espanha - 28, Reino Unido - 25, Austrália - 24, Irã - 18, Malásia - 17, Canadá - 16. No mapa não aparece todos os países com seus números de publicações.

Figura 15: Áreas de aplicação do tema.



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

Questão de pesquisa **Q 3** para responder essa questão foi feito um gráfico circular de modo a rotular as áreas que mais tem publicação no tempo escolhido na revisão. Na 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 questão **Q 4** de pesquisa foi feito um levantamento de alguns dos artigos mais influentes da revisão esses artigos retrata de alguns métodos sobre séries temporais os 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 utilizado pelos autores para previsão de séries temporais, e alguns modelos de análise do mesmo, 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 captar os comportamentos lineares e não lineares de uma série temporal, respectivamente. Li e Bastos (2020) o desempenho de previsão da abordagem MAELS é comparado aos predecessores baseados na aprendizagem de máquinas do estado para a arte, tais como CNN, RNN, LSTM, ARIMA, e SVM-VAR. As abordagens, CNN, RNN e LSTM permitem o tratamento de entrada e saída multivariada, o ARIMA utiliza dados passados para prever o futuro, usando dois principais recursos: a autocorrelação e médias móveis.

Então com essa revisão sistemática e análise do conteúdo obteve a resposta da responder à questão de pesquisa feita no começo do capítulo, com essa revisão sistemática pode perceber haver muitos métodos em séries temporais.

Fora esses modelos também tem a atualização do ARIMA, que vai ser utilizado nessa dissertação, como SARIMA, SARIMAX, ambos desses modelos vai ser comparado para obter o melhor modelo entre eles, fora esse também sera utilizado o Light GBM e XGBoost. Para as métricas de erros nessa dissertação será utilizado as seguintes métricas e explicado no capítulo 3 MAE, MAPE e RMSE, na literatura é uma das mais usadas entre várias, com por exemplo o R^2 citado (2) para as previsões futuras sempre foi separado com essas métricas de erros. o R^2 não é tão utilizado para comparação.

2.6 Conclusão da revisão

Nessa seção relatar o que foi abordado durante a pesquisa de revisão sendo ela em algumas bases como Scopus, Web of Science e Lens, cada base retornou vários artigos que foi analisado e com isso responde à questão de pesquisa feita na revisão, a pesar da base Lens foi a menor entre todas ainda encontra alguns artigos que foi de relevância no processo da dissertação, também com ajuda dos *softwares* para analisar muitos arquivos e suas relações entre cada um. Séries temporais sendo uma análise profunda e mais atual na revisão sistemática, fazendo a busca nos 6 últimos anos.

Na busca realizada foi obtido alguns resultados bem relevante, como no cruzamento das palavras em cada base com o filtro aplicado obteve 308 artigos de 2016 até 2022, com isso foi preciso filtrar mais sobre cada área de atuação dos artigos, como matemática, engenharia e informática nesse filtro teve um total de 481 artigos excluindo que seria das outras áreas.

3 Base Teórica

Para um bom resultado é fundamental uma base solida, nesse capítulo vai ser abordado as métricas de erros e os modelos de previsão modelo regressivo 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:

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2 \quad (3)$$

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

$$RMSE = \sqrt{\frac{1}{n} \sum (y_i - \hat{y}_i)^2} \quad (4)$$

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:

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (5)$$

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.

$$MAPE = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (6)$$

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.

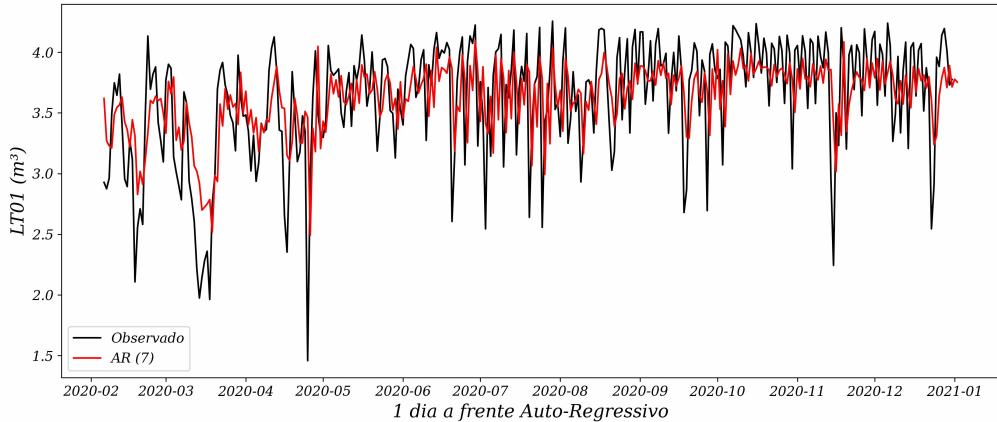
3.2.1 Componente auto-regressivo – AR(p)

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.

$$Y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \varepsilon_t \quad (7)$$

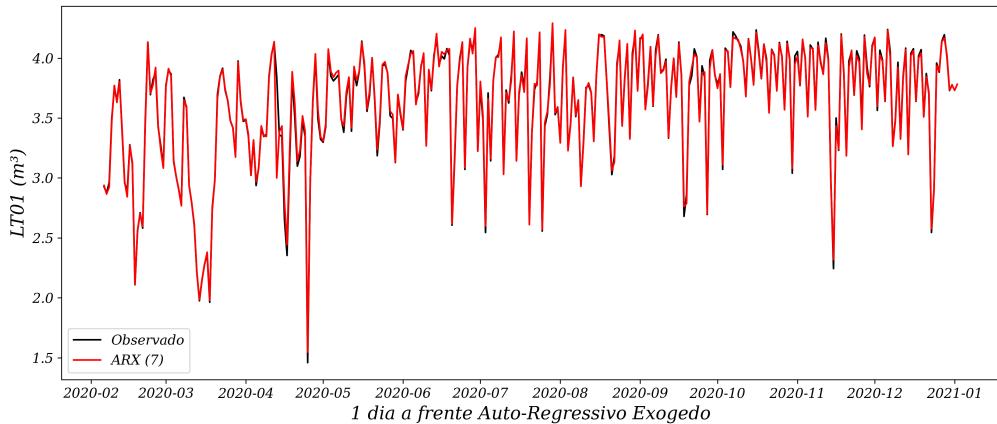
Dos dados pode ser obtido a seguinte previsão no modelo AR(7)

Figura 16: Modelo AR(7) com um passo a frente



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

Figura 17: ARX (7) com um passo a frente



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

Onde em (7) o ε_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 ??, B 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 α , 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)

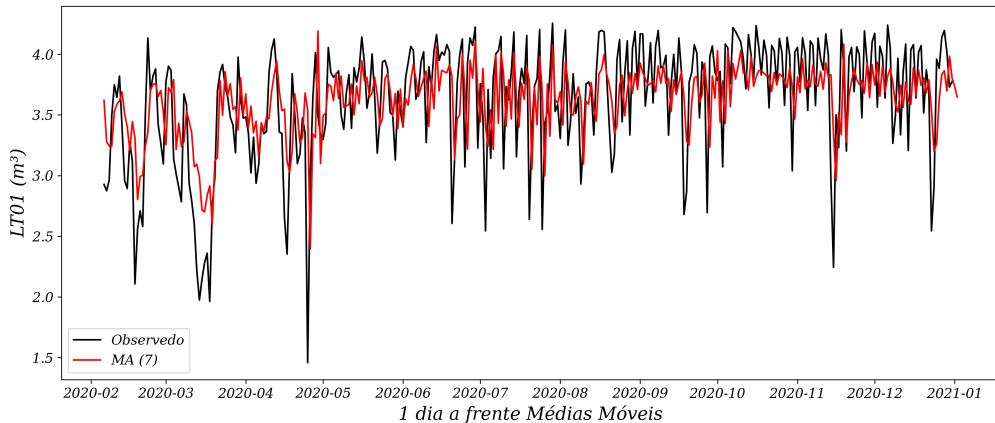
Este componente não é uma média de rolamento, mas sim os atrasos no ruído branco. Trenberth (1984)

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.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (8)$$

De (8) onde ε_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 ε_t , por isso não é realmente uma regressão no sentido habitual.

Figura 18: Modelo MA(7) com um passo a frente



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.

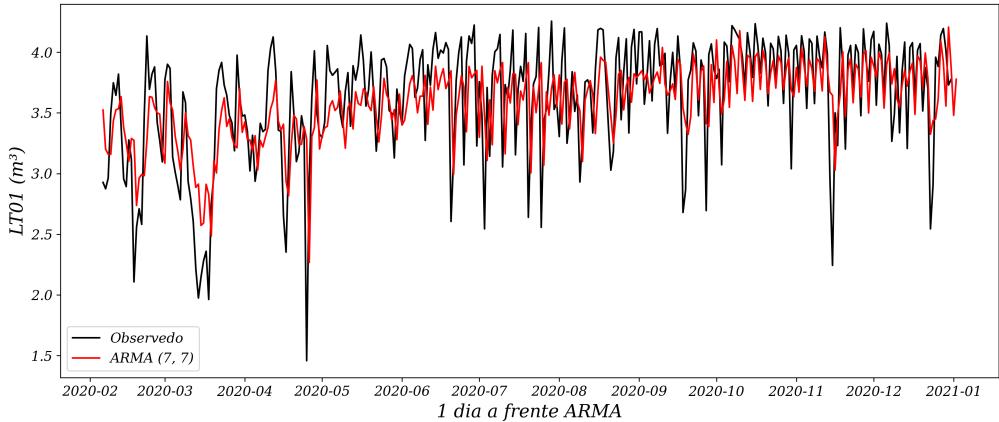
3.2.3 Modelos ARMA e ARIMA

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) com um passo a frente



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 ?? pode ser notado a comparação de alguns passos a mais do que mostrado aqui.

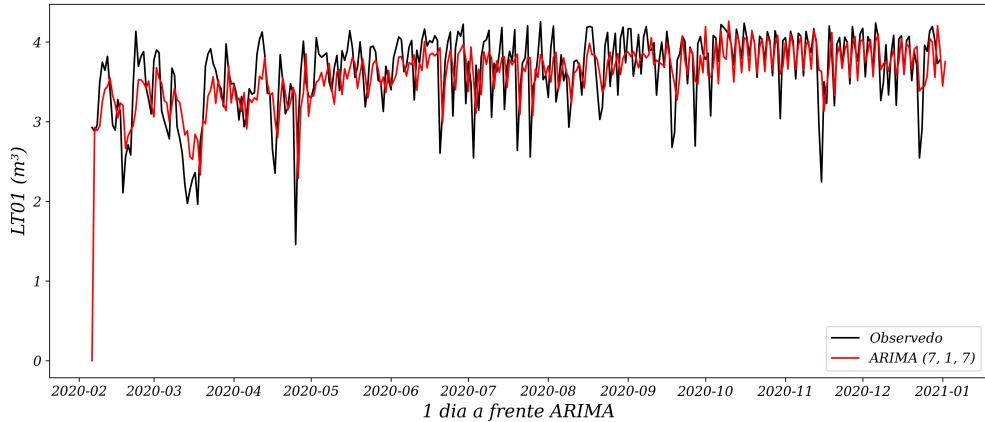
ARIMA

$$Y_t = \beta_2 + \omega_1 \varepsilon_{t-1} + \omega_2 \varepsilon_{t-2} + \dots + \omega_q \varepsilon_{t-q} + \varepsilon_t \quad (9)$$

Onde em (9) 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) com um passo a frente



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

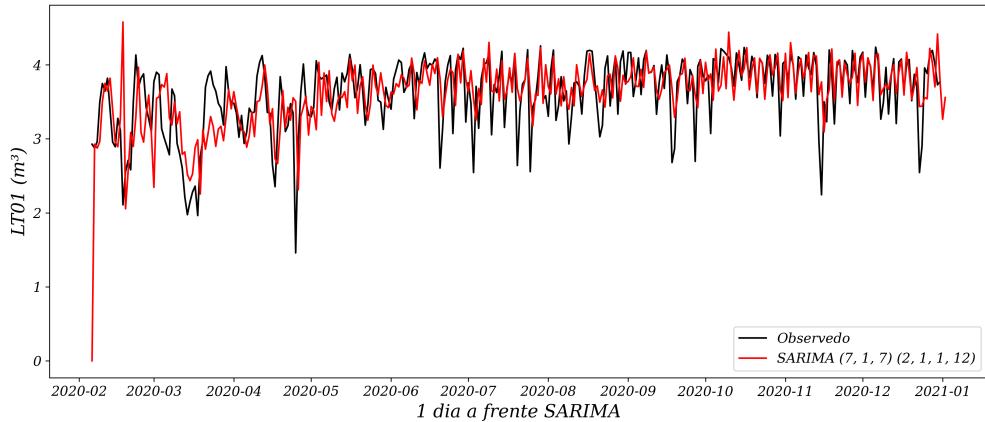
Olhando a Figura 20 podemos perceber que não tem muita diferença visual com os outros métodos mostrados até agora, visualmente o método ARX ainda esta melhor que os outros.

3.2.4 MODELOS SARIMA, ARIMAX, 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**

$$Y_t = c + \sum_{n=1}^p \alpha_n y_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^P \phi_n y_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (10)$$

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.

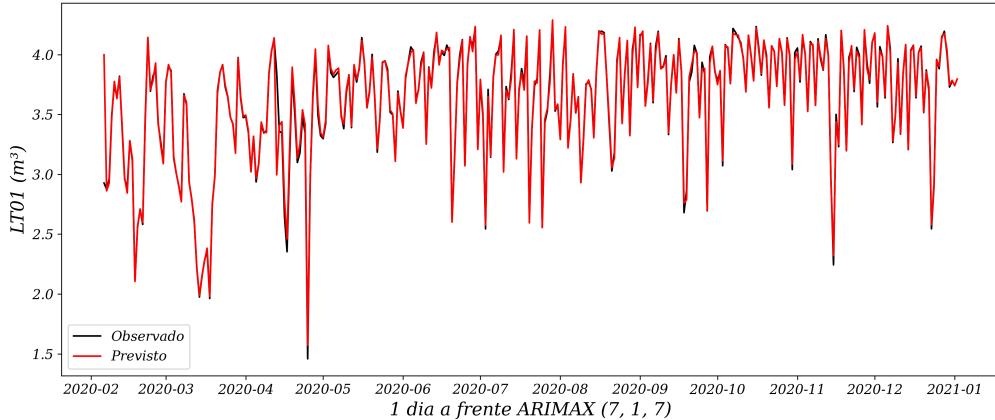
Os modelos SARIMA permitem diferenciar dados por frequências sazonais e não sazonais. Uma estrutura de pesquisa automatizada de parâmetros, como pmdarina, pode ajudar a entender quais são os melhores parâmetros.

ARIMAX e SARIMAX

$$d_t = c + \sum_{n=1}^p \alpha_n d_{t-n} + \sum_{n=1}^q \theta_n \epsilon_{t-n} + \sum_{n=1}^r \beta_n x_{nt} + \sum_{n=1}^P \phi_n d_{t-sn} + \sum_{n=1}^Q \eta_n \epsilon_{t-sn} + \epsilon_t \quad (11)$$

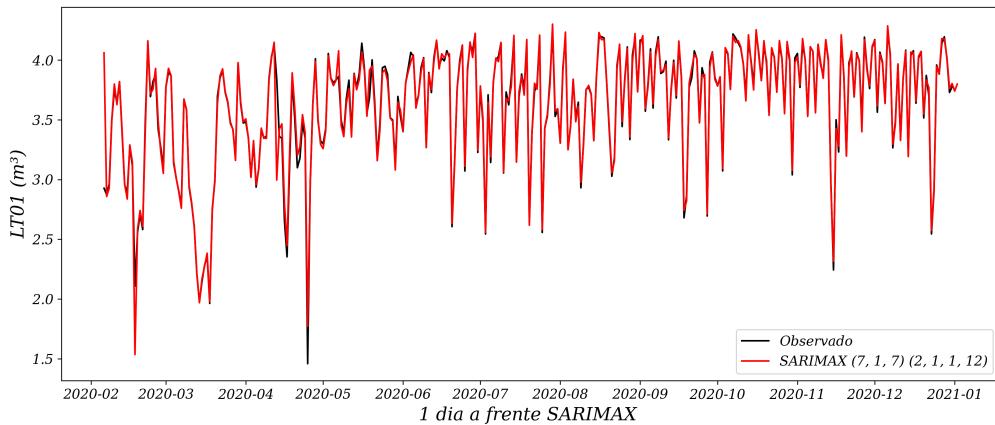
Em (11) está o modelo SARIMAX. Este modelo tem em conta variáveis exógenas, ou por outras palavras, utiliza dados externos na nossa previsão. É interessante pensar que todos os factores exógenos ainda são tecnicamente indiretamente modelados na previsão histórica do modelo. Dito isto, se incluirmos dados 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) com um passo a frente



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

Figura 23: SARIMAX (7, 1, 7)(2, 1, 1)₁₂ com um passo a frente



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

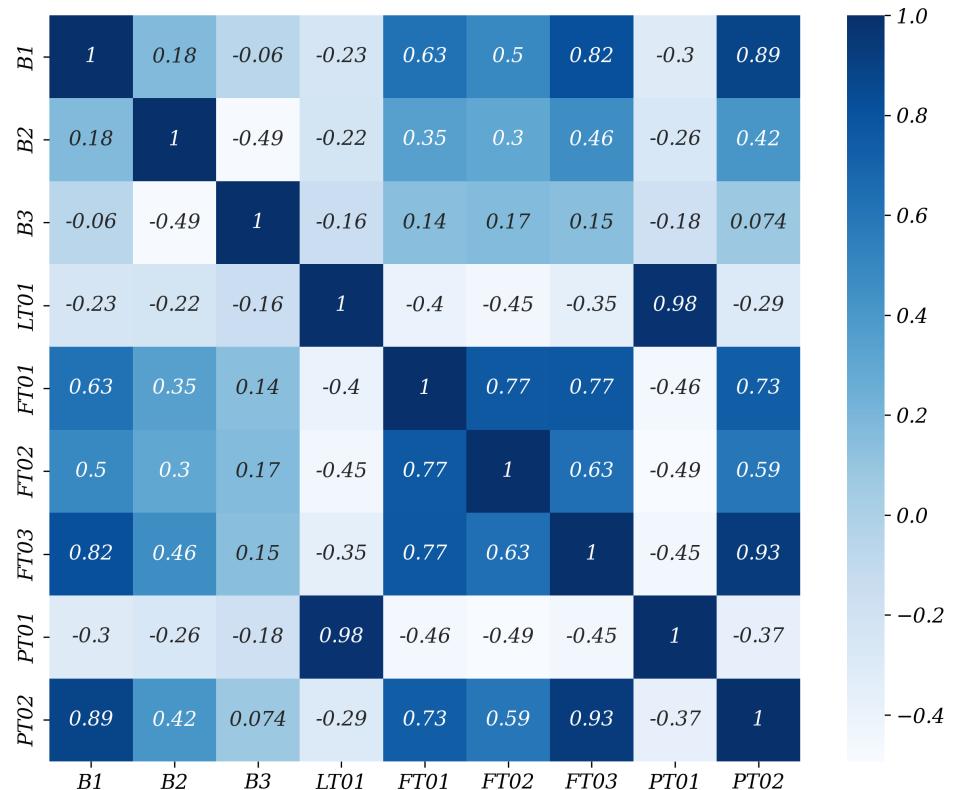
Para a previsão, é importante entender quais tipos de variáveis explicativas você pode ou não usar. Como exemplo, aqui vai ser usado as variáveis **Pressão de Succção (PT01SU)** como variável x e **Nível do Reservatório (Câmara 1) LT01** como variável y pois na correlação de Pearson mostrado na Figura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

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum (x_i - \bar{x})^2)(\sum (y_i - \bar{y})^2)}} \quad (12)$$

De (12) 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

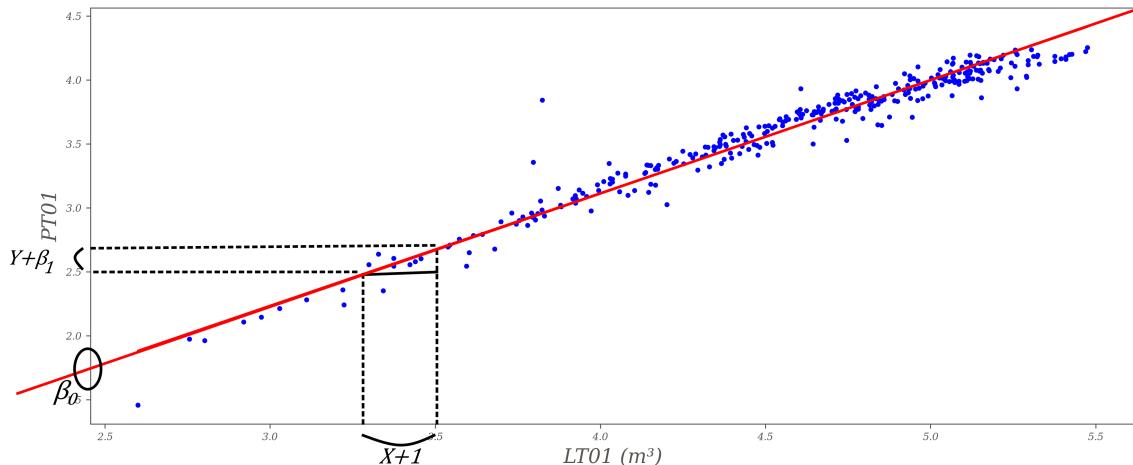
A regressão linear é definida da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \quad (13)$$

Da (13) 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 (β_0) mais os valores do x variáveis multiplicadas pelos seus coeficientes β_1 para β_p .

Figura 25: Regressão linear LT01 vs PT01 correlação 98%



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

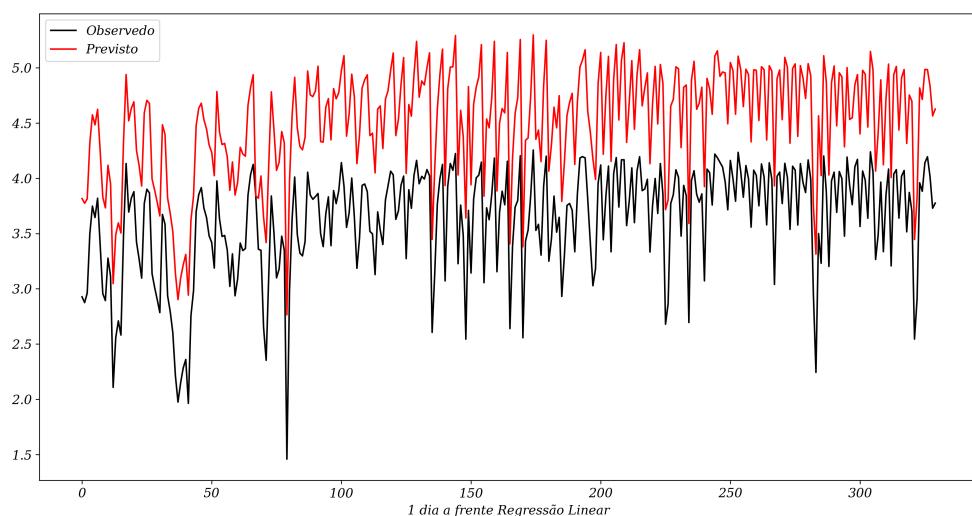
A Figura 25 mostra como interpretar β_0 e β_1 visualmente. Mostra que para um aumento de 1 na variável x , o aumento na variável x representa β_1 . O valor para 0 é o valor para x quando y é 0.

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

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (14)$$

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

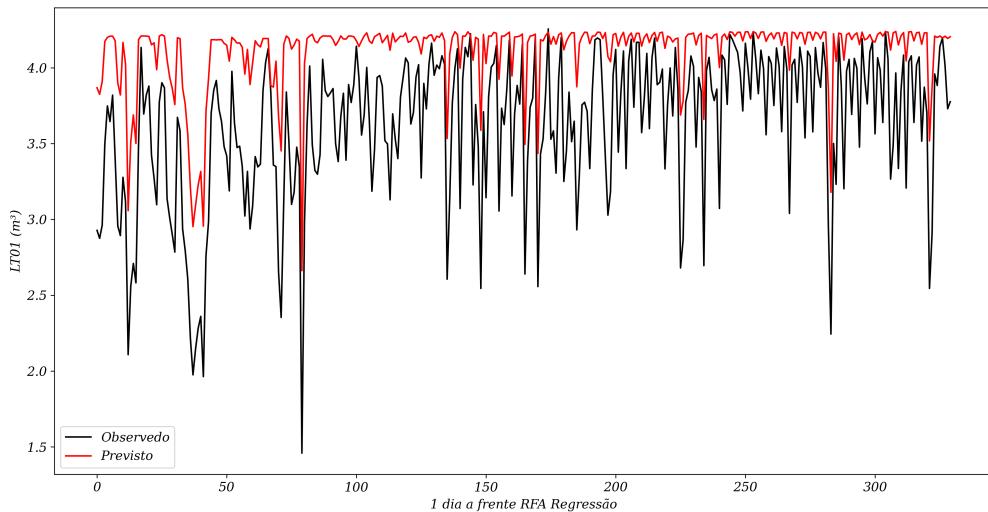


Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

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

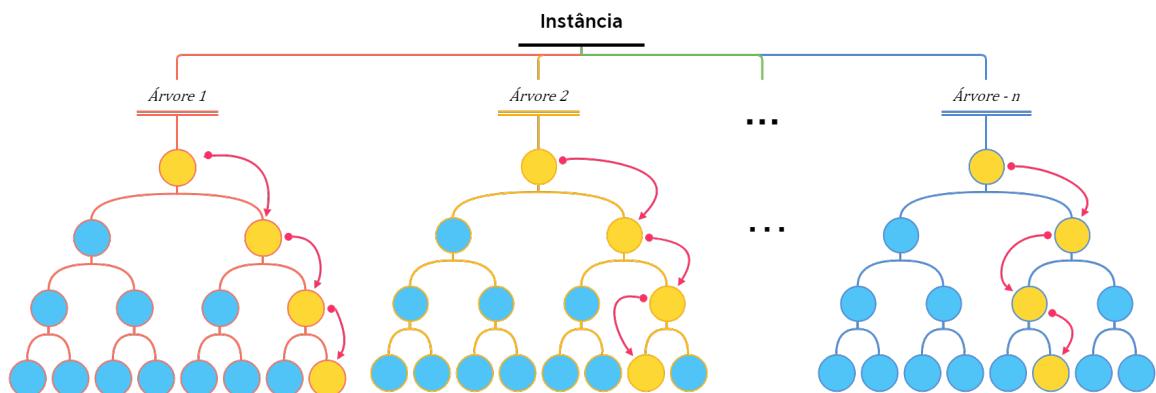
Figura 27: Regressão da Floresta Aleatória (RFA) um passo a frente



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

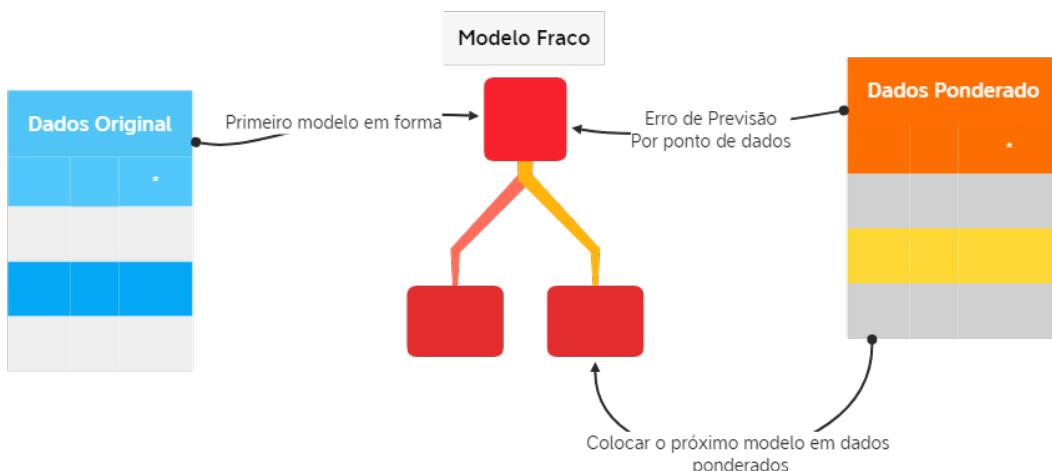
3.3.3 LightGBM e XGboost

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 diferentes umas 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 algoritmo 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 acente: 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.

$$y - \hat{y} \tag{15}$$

De (15) isso pode ser denotado reescrito como a derivada parcial negativa da função de perda em relação às previsões de y :

$$y - \hat{y} = -\frac{\partial L}{\partial \hat{y}} \tag{16}$$

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.

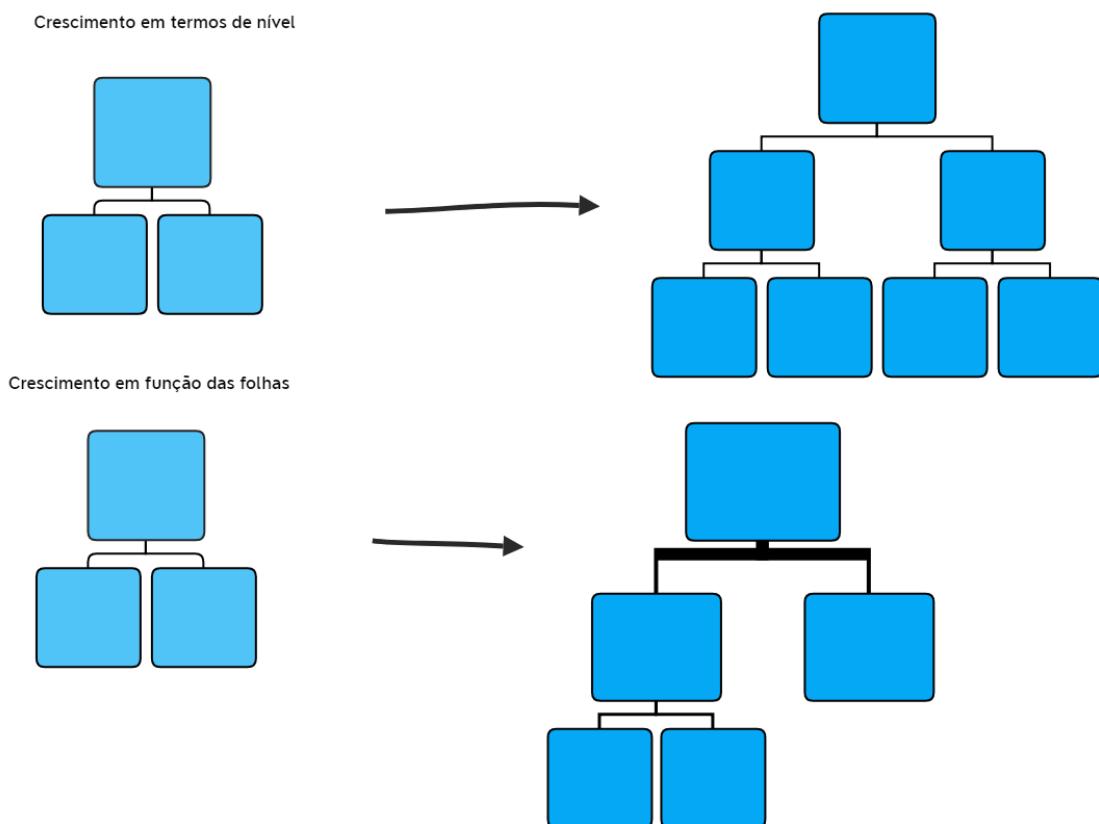
A ideia intuitiva e mais simples para encontrar a melhor divisão é iterar todos os ajustes possíveis e encontrar a melhor divisão. No entanto, isso leva muito tempo e algoritmos recentes apresentam alternativas melhores. Uma alternativa proposta pelo XGBoost é usar a segmentação baseada em histograma. 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.

4 Resultados

Neste capítulo é mostrado um breve resultado do que foi realizado até o presente momento.

4.1 Planejamento do Problema

Assim como foi mostrado na seção 1.3.1 as etapas da dissertação, com isso cada modelo e os métodos que pode ser utilizado para responder as Questões de pesquisas abordado na seção 1.2.1. Com as etapas podemos dar uma cronologia logica do que foi adquirido ao longo do tempo com os dados da SANEPAR.

4.1.1 Análise Exploratória dos dados (EDA)

Da **Etapa 1** é realizado o EDA para os processamento de dados que obteve até aqui, com o EDA vai ser respondido. Segundo Yu (2016) Na era do big data, coletamos volumes de dados de massa caóticos, não estruturados e multimídia por meio de vários canais. Como descobrir as regras, os modelos analíticos e as hipóteses nesses dados tornou-se o novo desafio. A análise exploratória de dados foi promovida por John Tukey para incentivar os estatísticos a explorar os dados e, possivelmente, formular hipóteses que poderiam levar a uma nova coleta de dados e experimentos. Diferente da 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 foram adotadas na análise de big data.

Olhando para **Q 1** relacionando a demanda com a variável que esta sendo prevista, e a pressão com a variável PT01 da Figura 24 pode se notar que ambas estão trabalhando por igual, quase uma correlação perfeita de $r = 1$ então para essa questão basta olhar a correlação de Pearson na Figura 24.

Para **Q 2** vai ser feito uma tabela para que seja respondido melhor essa questão

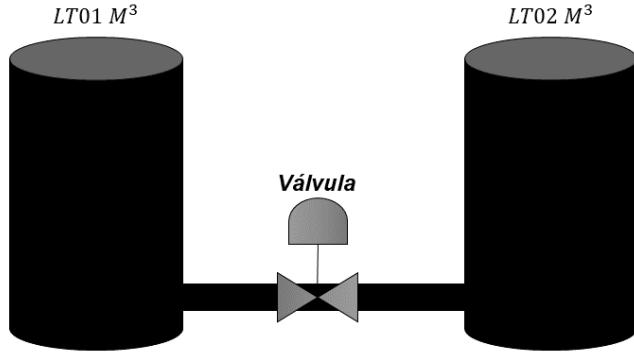
Tabela 4: Descrição Estatística dos dados com filtro aplicado de 18 a 21h

18 a 21h	B1	B2	B3	LT01	FT01	FT02	FT03	PT01	PT02
Contagem	1096	1096	1096	1096	1096	1096	1096	1096	1096
Média	48,830	17,538	4,299	3,545	211,771	113,805	100,139	4,485	19,424
STD	12,354	9,282	8,976	0,438	44,496	21,486	19,822	0,487	4,323
Min	0,000	0,000	0,000	1,459	22,854	3,584	10,383	1,925	0,831
25%	50,491	13,289	0,000	3,345	195,944	105,096	99,239	4,245	19,805
50%	54,204	19,965	0,000	3,644	215,791	115,733	104,846	4,571	21,015
75%	54,818	22,792	2,376	3,843	238,325	126,246	110,006	4,813	21,140
Max	57,885	53,488	46,841	4,256	301,863	181,565	143,988	5,475	23,679

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

Na Tabela 4 o desvio padrão é dado pela sigla de STD que vem do inglês *standard deviation*, observando também para responder a Q 2, assim como toda companhia de tratamento de água é feito um acionamento automático, chamando de trava de segurança, para que o tanque não chegue a zerar e faltar água em todos os lugares adjacente que é abastecido por essa água, esse mínimo que o tanque pode chegar é de $1,459m^3 \iff 1459$ litros e as bombas serão acionadas em sua potencia máxima, para evitar o acionamento das bombas o nível do reservatório tem que estar no intervalo de $[3.843, 4.256] m^3$ ainda sim, a bomba 1 estaria em funcionamento para completar o nível. Em casos de horários de picos o mais ideal, mas não o mais lucrativo é outro tanque de reserva nesses horários, e instalar uma tubulação para ligar um ao outro. Durante o dia estaria abastecendo os dois e a noite pela gravidade eles ficariam com o mesmo nível até o consumo chegar em um nível de acionamento das bombas.

Figura 31: Solução para acionamento das bombas



Fonte: Elaboração própria

Na Figura 31 um esquema prático para evitar a falta de água e o consumo em horários de pico. Esse esquema é bem simples de como pode ser melhorado o aproveitado de tempo no período do dia para armazenamento de água.

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

Na **Q 4** o ponto de equilíbrio para não ser acionado as bomba seria de as vazão FT01 $211m^3/h$ FT02 $114m^3/h$ FT03 $100m^3/h$ e o nível do tanque com $3,545m^3$.

Na a. o tanque deve estar com o nível de $4,00m^3$ para que não precise acionar bombas no horário de pico.

4.1.2 Múltiplas entradas e saída única (MISO)

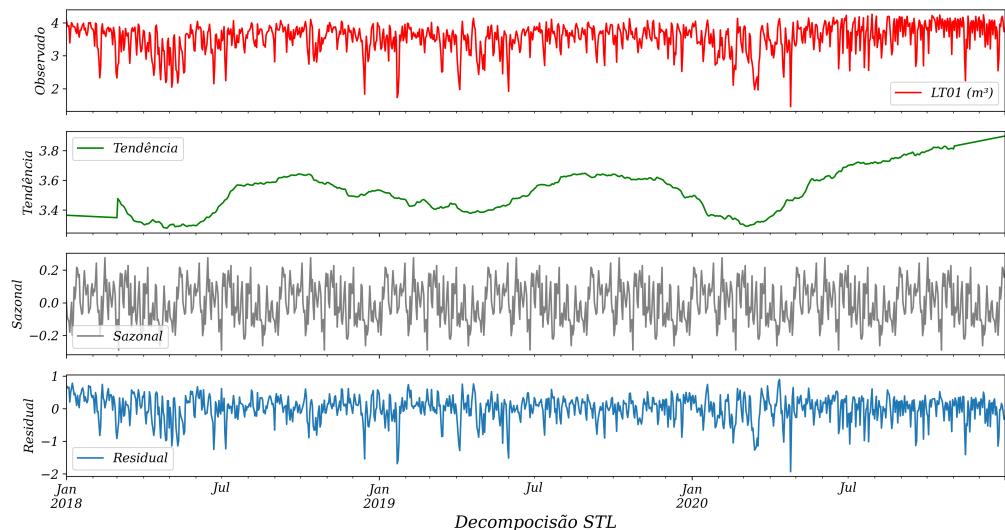
Nessa **Etapa 2** o modelos que mais foi abordado no decorrer da dissertação é os modelos ARIMA, ou os que derivam desse modelo, e os modelos regressivo fora o LR tem múltiplas entra e uma saída da variável que é prevista o LT01, as outras variáveis serve de apoio para melhorar os modelos do tipo ARX ou modelos com variáveis exógenas. Os

modelos ARIMA sem a variável exógena é apenas um entrada, semelhante com o LR.

4.1.3 Decomposição STL

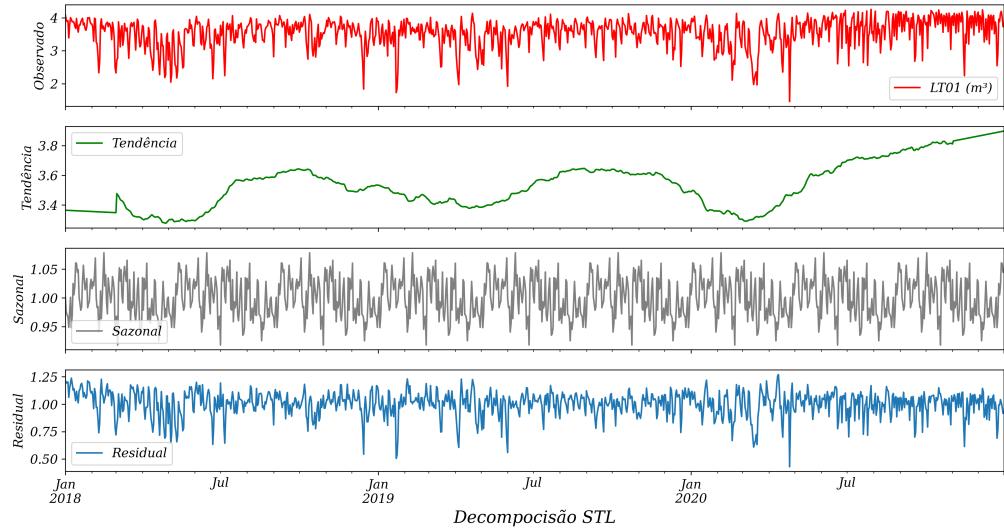
Theodosiou (2011) A decomposição sazonal e tendencial utilizando o procedimento de Loess (STL) é utilizada para a decomposição aditiva da série temporal global. O STL realiza a decomposição aditiva dos dados por meio de uma sequência de aplicações do Loess mais suave, que aplica regressões polinomiais 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 32: Decomposição STL aditiva dos dados coletados



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

Figura 33: Decomposição STL multiplicativa dos dados coletados



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

Na Q 5b. pode ser respondida pela Figuras 32 e ?? como é observado tem tenência, sazonalidade e resido.

Na decomposição o objetivo dela é analisar se há tenência, sazonalidade e resido, olhando nas Figuras 33 e 32, isso mostra que os dados tem ambas das analise. E com isso perceber que a série é estacionaria, pelo teste a seguir.

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

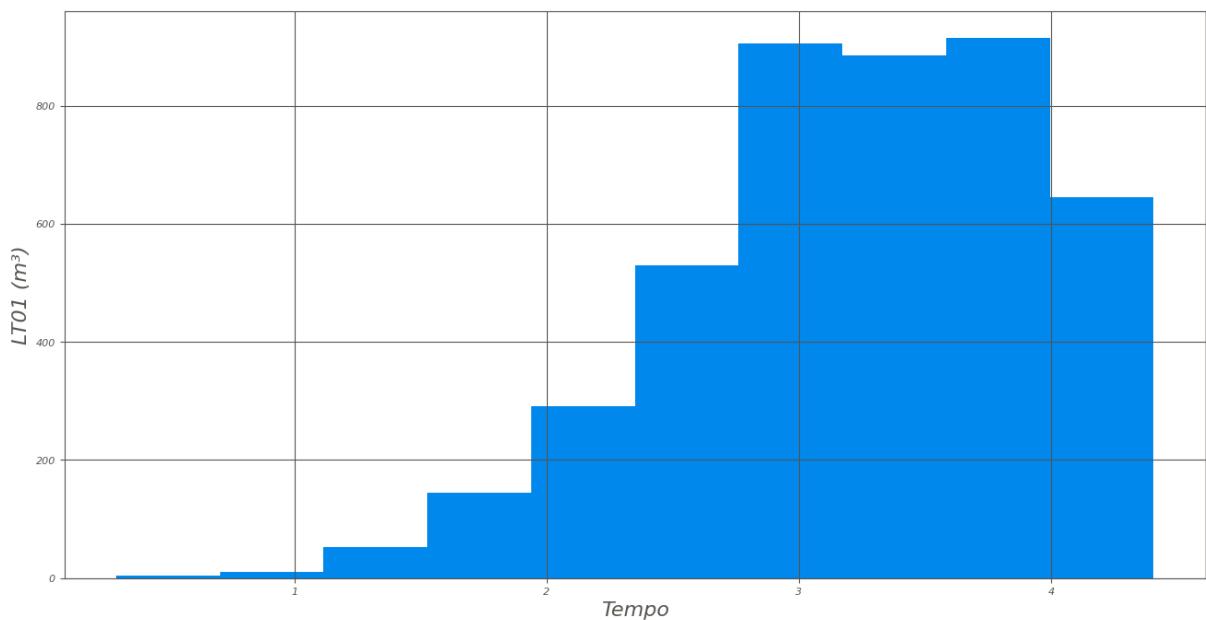
Fortes provas contra a hipótese nula

Rejeitar a hipótese nula

Os dados não têm raiz unitária e são estacionários, Na Q 5c. como a serie é estacionaria, para identificar quais os horários de pico entre as 18 até as 21h não é um

trabalho fácil, pois se pegar na Figura 34 pode perceber que no ano de 2020 teve um aumento da demanda nessas horas.

Figura 34: Histograma do nível do reservatório



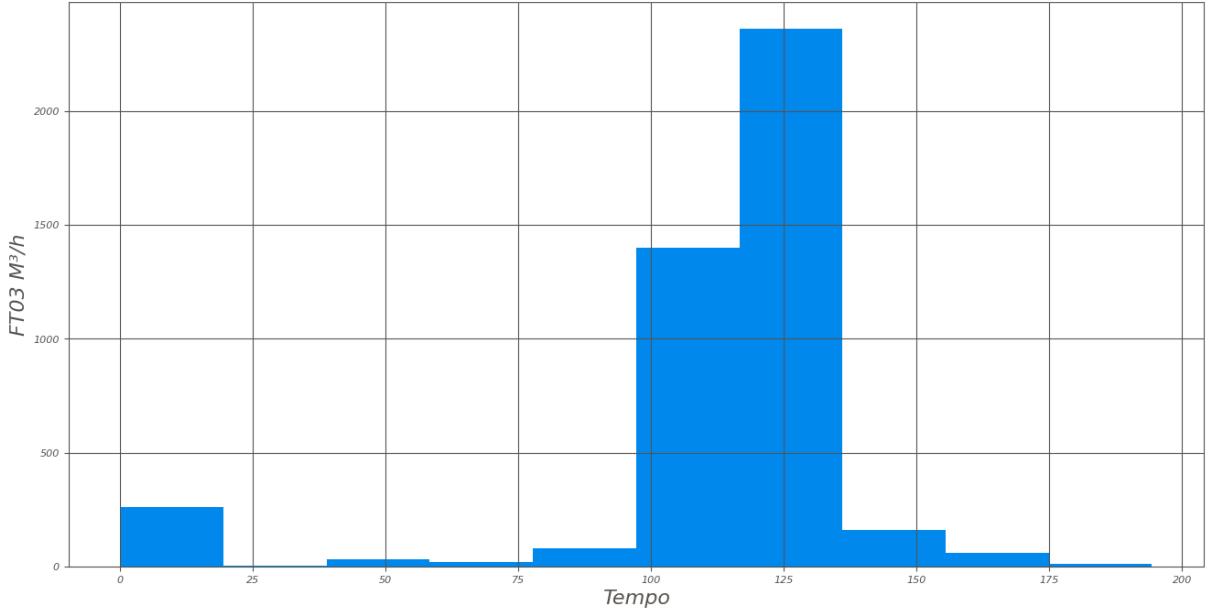
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Então como dito na seção 1.1.1 as anomalias de tempo mais ocasionada no ano de 2020 e foi devido a falta de chuva nesse período.

Na **Q 5d.** nos horários de picos deve conter no tanque por volta de $[3.545, 4.256] m^3$ para que não acione as bombas.

Para **Q 5e.** é mostrado na Figura 35 como pode ser afetado a vazão com o nível do tanque. A vazão de recalque influencia mais no nível do tanque que as outras vazão pois injeta água no tanque por meio da bomba de recalque, que fica mais próximo da base do tanque, e as outras vazão por ter alguns valores ausente, não interfere tanto na amostra.

Figura 35: Histograma da vazão de recalque



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

Segundo Reisen et al. (2017) o teste de DF tem como formula a seguinte equações

$$z_t = y_t + \theta \beta_t, \quad t = 1, \dots, T, \quad (17)$$

$$\hat{\rho}_{\text{DF}} - 1 = \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (18)$$

De (18) 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:

$$T(\hat{\rho}_{\text{DF}} - 1) = T \frac{\sum_{t=1}^T z_{t-1} \Delta z_t}{\sum_{t=1}^T z_{t-1}^2} \quad (19)$$

e

$$\hat{\tau}_{\text{DF}} = \frac{\hat{\rho}_{\text{DF}} - 1}{\hat{\sigma}_{\text{DF}} \left(\sum_{t=1}^T z_{t-1}^2 \right)^{-1/2}} \quad (20)$$

De (20) onde $\hat{\sigma}_{\text{DF}}^2 = T^{-1} \sum_{t=1}^T (\Delta z_t - (\hat{\rho}_{\text{DF}} - 1) z_{t-1})^2$.

Suponha que $(z_t)_{1 \leq t \leq T}$ são dadas por (17), então quando $\rho = 1$,

$$T(\hat{\rho}_{DF} - 1) \xrightarrow{d} \frac{W(1)^2 - 1}{2 \int_0^1 W(r)^2 dr} - \left(\frac{\theta}{\sigma}\right)^2 \frac{\pi}{\int_0^1 W(r)^2 dr}, \text{ como } T \rightarrow \infty \quad (21)$$

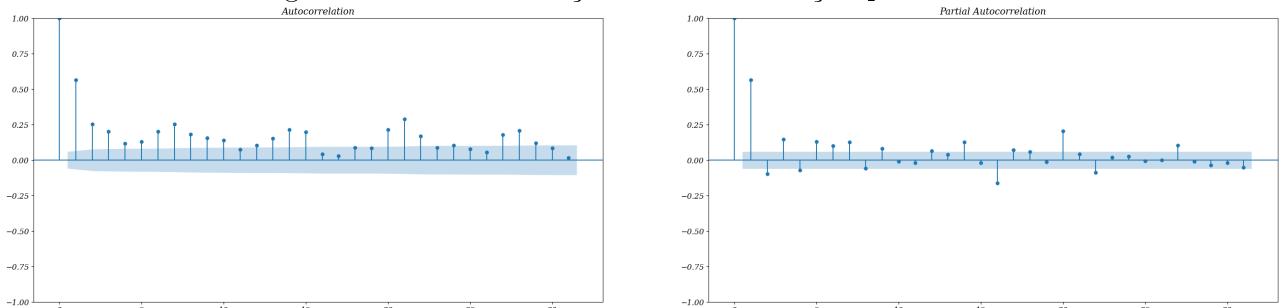
$$\hat{\tau}_{DF} \xrightarrow{d} [1 + 2(\theta/\sigma)^2 \pi]^{-1/2} \left\{ \frac{W(1)^2 - 1}{2 \left(\int_0^1 W(r)^2 dr \right)^{1/2}} - \frac{(\theta/\sigma)^2 \pi}{\left(\int_0^1 W(r)^2 dr \right)^{1/2}} \right\} \quad (22)$$

como $T \rightarrow \infty$ (23)

De (23) onde \xrightarrow{d} denota a convergência na distribuição e onde $\{W(r), r \in [0, 1]\}$ denota o movimento browniano padrão.

Esse teste na literatura é chamado de teste ACF para testar se a serie é o não estacionária, basicamente se a serie tiver um valor de raiz unitária é uma série não estacionária, do contrario como acontece com os dados coletados se torna uma série estacionária.

Figura 36: Autocorrelação e Autocorrelação parcial



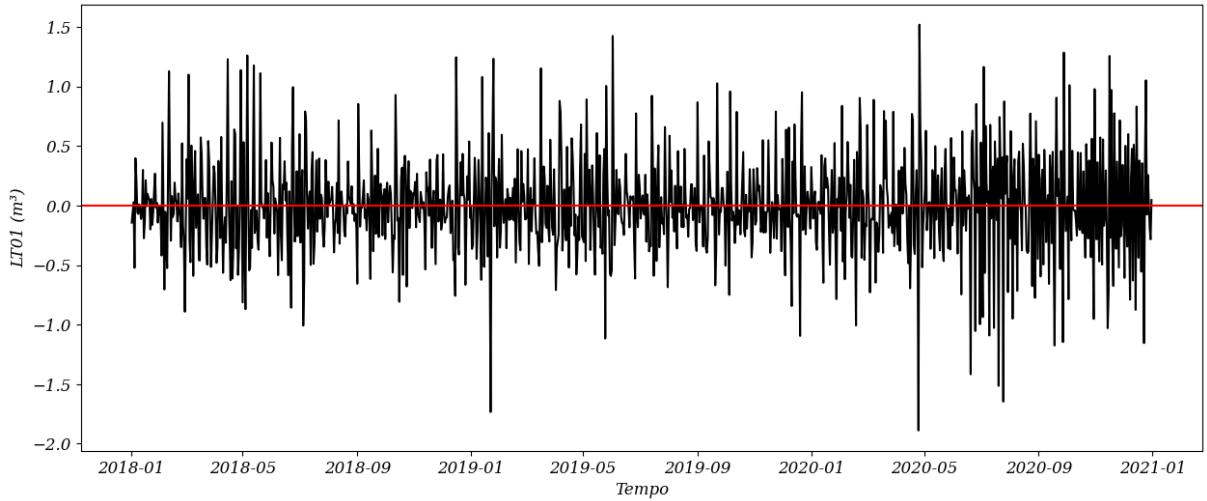
Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 2020)

Na Figura 36 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 por padrão é 95%, mostrado como essa marca azul. Observações que estão para fora da marca são consideradas estatisticamente correlacionadas.

As correlação da Figura 36 é a explicação do teste de DF, entendo isso pode ser visto o próximo passo que é a análise do ruído branco em meados a gráfico, Uma série ruído branco é uma série na qual a média 0, a variância é constante ao longo da série toda e não há correlação entre os períodos de tempo. O valores de uma série ruídos brancos são totalmente aleatórios, ou seja, essa é um tipo de série que não é previsível.

Figura 37: Ruído branco



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

Da Figura 37 uma série temporal pode ser ruído branco. Uma série temporal é ruído branco se as variáveis são independentes e distribuídas de forma idêntica com uma média de zero. Isso significa que todas as variáveis têm a mesma variância (σ^2) e cada valor tem uma correlação zero com todos os outros valores da série. Mais para frente vamos mostrar o comprimento de zeros na variável prevista. Com isso encerra a **Etapa 3**.

4.1.4 Separação dos Dados

Na **Etapa 4** tem um esquema de como foi dividido os dados em treino, teste e validação, essa prática é comum para os profissionais de aprendizado de máquina, pois assim como não é possível processar os dados todos de uma vez, se você manusear dados em uma escala menor até pode ser realizado, mas tudo depende da máquina que está sendo realizado o processamento dos dados, cada modelo em particular utiliza um certo acervo do seu computador para processar, se por exemplo você tiver trabalhando com um modelo de aprendizado profundo que é mais comum em processamento de imagem, a Nvidia tem sempre inovado com suas GPUs e trazendo mais poder para processamento, com o recente lançamento da placa de vídeo 3090 um sonho de consumo para games e os profissionais de aprendizado de máquina e profundo.

Em fim se o computador que foi realizado os processamento fosse um computador não tão bom, ainda poderia estar sendo pensado que estaria em processamento, sem as inovação que foi estabelecida aos anos, o computador que foi realizado os cálculos dos modelos foi em partes um computador de processador *i5 – 3300* e um notebook com

i7 – 5500 ambos com 4 threads e o notebook com apenas 2 núcleos o *i5* contem 4 núcleos. Cada um tem suas especificações de ser o melhor em algum certo ponto, mas sabendo que não é preciso de um de ultima geração para fazer tais processamento. E sim força de vontade para entender e aplicar em cada um.

A divisão mais básica que tem na literatura foi realizada aqui na separação dos dados, 70% para treino e os 30% restante para teste, dos 70% tem mais uma divisão pegando 80% dos 70% para treino novamente e os 20% para validação dos dados, tendo essa fórmula aplicada na linguagem de programação para que não precisa ser contado todas as vezes que for mudado o modelo.

4.1.5 Estratégia de Previsão

Na **Etapa 5** é abordado a forma que foi previsto os dados, em uma janela de horizonte de previsão bem maior do que o normal na literatura da estratégia de recursiva, sendo 1, 10, 30 e 60 dias previsto, essa estratégia para comparação dos modelos regressivo e modelos ARIMA, é bem vantajoso, pois cada modelo tem suas especificidades para prever em momentos com janela de tempo menor e com uma janela de muitos dias. Assim como explicado na seção 4.1.7 se for previsão curta, alguns vão se sobrepor em meio a outros modelos que foi feito aqui.

4.1.6 Horizonte

Na **Etapa 6** é feito o horizonte de previsão, como dito na seção 4.1.5 esse horizonte foi customizado baseado no método recursivo de prever as séries temporais e a previsão do nível do tanque LT01. Os passos para prever a frente foram de 1, 10, 30 e 60 dias, já foi realizado uma estratégia com uma janela menor, mas para comparação dos modelos essa janela foi mais adequada.

4.1.7 Modelos de previsão e métricas de desempenho

Da **Etapa 7** as métricas utilizadas aqui foram vistas na seção 3.1 foram utilizados três das métricas mais usadas na literatura para previsão de tempo e comparação de modelos ARIMA e os modelos regressores.

Em comparação com os modelos feitos, pode ser visto que o modelo LR em um passo a frente tem tanto na modelagem de 24 horas quanto no pico de horas entre as 18 e 21 horas, foi o modelo que mais se saiu bem na previsão, logo em sequência os modelos MA, AR, SARIMA, ARIMA, SARIMAX, ARIMAX, ARX, LGBMRegressor, XGBRegressor

e Random Forest Regressor, para curto prazo esses modelos estão em ordem de melhor para pior.

Já em grande espaço de tempo como foi feito de 60 dias os modelos ARMA, AR, MA, ARIMA, ARIMAX, ARX, SARIMAX, SARIMA, XGBRegressor, Random Forest Regressor, LGBMRregressor e LR, seguindo a mesma lógica do melhor para o pior. Mas se olhar graficamente nos modelos que foi feito, os modelos com variáveis exógenas aparenta prever melhor do que os outros modelos, só analisando os dados nos apêndice tanto quanto as Figuras de ?? a 53 quanto as Tabalas 6 a ??

4.1.8 Teste de Significância

Na **Etapa 8** os teste escolhido foi de *Friedman e Nemenyi* no teste de Nemenyi, precisamos obter a diferença entre os rankings médios (linha média da tabela de classificação) entre todos os classificadores (comparando pares de classificadores). Se essa diferença for maior ou igual a um CD (distância crítica), podemos dizer que esses dois classificadores são significativamente diferentes um do outro. O CD é calculado como:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (24)$$

De (24) o termo q_α é obtido de ($\alpha = 0,05$):

Tabela 5: Teste Nemenyi

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

Fonte: Elaboração própria a partir de dados da SANEPAR (2018 a 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 (25) para determinar a estatística de teste F_r :

$$F_r = \left[\frac{12}{nk(k+1)} \sum_{i=1}^k R_i^2 \right] - 3n(k+1) \quad (25)$$

Na equação (25) n é o número de linhas (ou amostras), k é o número de colunas (ou condições) e R_i é a soma dos postos da coluna (ou condição) i . Segundo a equação (25) têm o seguinte resultado nos dados da pesquisa.

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

5 Conclusões

Nessa dissertação teve como objetivo mostrar a escassez de água que ocorreu em Curitiba, tornado possível uma tomada de decisão que foi uma adaptação do case de 12 etapas de ALMEIDA (2013) que busca e visa o meio para ter a visão que não há interferência do meio, e se houver essa interferência listamos ela como variável exógena nos modelos do tipo ARX, ARIMAX e SARIMAX, nos modelos regressivos por mais que eles sejam bom de trabalhar não teve como incluir nesse momento. Se o previsor busca as anomalias nos dados como foi feito aqui, olhamos para os dados de 2020 que foi a grande anomalia da SANEPAR, essas anomalia explicado os resultado no capítulo 4.

5.1 Limitações da pesquisa e propostas futuras

As limitações desse trabalho resulta no tempo e os modelos de aprendizado de máquina, como visto no decorrer dessa dissertação, tem vários modelos que pode ser trabalhado em conjunto com a série temporal, por exemplo os modelos de redes neurais, LSTM, CNN, RNN... Entre outros modelos, não foi muito bem abordado aqui, pois são modelos mais complexo e exigiria uma gama maior de tempo, para esse momento apenas os modelos que foi trabalhado, a principio atendeu as questão de pesquisa que foi levantado.

Porém nos próximos passos para um trabalho futuro é abordar melhor esses modelos de previsão, tendo com muitos autores na literatura trabalhando com esses modelos, até competição de aprendizado de máquina com os modelos mais famosos, como o Light GBM em comparação com o XGboost, para previsão de curto prazo e para longo cada modelo tem sua relevância, LR como um modelo de no máximo 3 variáveis para dados com poucas viráveis ele é muito eficiente e ágil.

No trabalho que vai ser de sequência à esse como um complemento desse trabalho, tem como abordar a literatura inteira não apenas os últimos 6 anos, e também visa as outras parte que não foi abordado, como dissertação, tese e capítulos de livros, mesmo abordando um pequeno grupo de artigos ainda teve uma gama muito grande de artigos sobre o tema.

A otimização matemática com alguns modelos sendo eles, floresta aleatória, XGboost, Ligh GBM, que poderia ser usado otimização para aumentar o gradiente e melhorar a precisão de aleatoriedade dos ramos das árvores. Métodos de otimização para melhorar o modelo foi **Grid Search, Randomized Search e Bayesian Optimization (Bayes Search)** que vem do inglês, para o português seria **Pesquisa em grade, Pesquisa ale-**

atória e Otimização Bayesiana para o floresta aleatória o melhor método em hipótese séria o randomized, ele busca os galhos mais rápido da árvore, assim prevendo melhor o tempo, mas em tese todos eles em algum modelo não conseguiu reduzir os erros listado na seção 3.1 ao invés de reduzir houve um aumento dos erros, fazendo que a previsão fosse ao longo do tempo ficando pior, como por exemplo no apêndice B que teve os melhores resultado se pegar os erros entre os modelos citados anteriormente, e em comparação aos modelos de otimização encontrado na literatura teve um aumento nos erros de 20 a 50 %, e para uma previsão mais precisa é preciso que fique próximo de zero.

Nessa parte da otimização é relevante pesquisar ou ter mais aprofundamento nos hiperparâmetros, para ter um melhor aproveitamento dos modelos de árvore e dos gradientes.

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>>.
- 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.

IGNATOV, I.; MOSIN, O.; BAUER, E. Effects of calcium, magnesium, zinc and manganese in water on biophysical and biochemical processes in the human body. **Journal of Medicine, Physiology and Biophysics**, v. 25, p. 45–63, 2016.

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%2fj.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 r² 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-s2.0-84925431469&doi=10.1109%2FTNNLS.2015.2411629&partnerID=40&md5=e1c7f3c7a1136a0e0e4d2aff817b4008>>.

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

Nas tabelas do apêndice todos os valores estão multiplicado por 100 com isso sendo os valores uma porcentagem de erro.

$$(p = 7, d = 1, q = 7)(P = 2, D = 1, Q = 1)_{M=12} \text{ média } 24\text{h}$$

Tabela 6: Comparação dos modelos 1 dia a frente 24h **Treino**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	9,645	30,560	41,924
ARX	11,845	37,725	51,291
MA	9,335	29,574	40,253
ARMA	10,038	32,014	42,983
ARIMA	9,506	30,302	40,653
SARIMA	9,951	32,141	42,645
ARIMAX	11,911	37,936	51,370
SARIMAX	11,837	37,710	51,255
Linear Regression	1,546	6,865	7,716
Random Forest Regressor	19,063	62,523	67,244
XGBRegressor	19,040	62,408	67,339
LGBMRegressor	18,219	59,368	65,094

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

Tabela 7: Comparação dos modelos 1 dia a frente 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	8,397	28,469	36,552
ARX	10,262	35,365	45,938
MA	8,217	27,830	36,055
ARMA	8,625	29,546	37,165
ARIMA	8,204	28,048	35,207
SARIMA	9,293	31,932	41,394
ARIMAX	10,239	35,281	45,837
SARIMAX	10,379	35,798	46,296
Linear Regression	1,426	6,568	7,283
Random Forest Regressor	17,264	58,845	63,270
XGBRegressor	17,390	59,278	63,819
LGBMRegressor	16,393	55,609	60,941

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

Tabela 8: Comparação dos modelos 1 dia a frente 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	8,397	28,469	36,552
ARX	10,262	35,365	45,938
MA	8,217	27,830	36,055
ARMA	8,625	29,546	37,165
ARIMA	8,204	28,048	35,207
SARIMA	9,293	31,932	41,394
ARIMAX	10,239	35,281	45,837
SARIMAX	10,379	35,798	46,296
Linear Regression	1,426	6,568	7,283
Random Forest Regressor	17,264	58,845	63,270
XGBRegressor	17,390	59,278	63,819
LGBMRegressor	16,393	55,609	60,941

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

Tabela 9: Comparação dos modelos 1 dia a frente 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	9,326	30,241	16,502
ARX	12,276	40,217	28,304
MA	10,669	34,408	45,957
ARMA	9,748	31,589	42,366
ARIMA	9,501	30,766	41,175
SARIMA	9,961	32,770	42,621
ARIMAX	12,044	39,458	52,152
SARIMAX	12,250	40,123	53,073
Linear Regression	1,617	7,363	8,360
Random Forest Regressor	17,683	58,103	64,244
XGBRegressor	17,702	58,177	64,427
LGBMRegressor	16,902	55,165	62,412

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

Tabela 10: Comparação dos modelos 10 dia a frente 24h **Treino**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	11,484	36,360	48,298
ARX	12,697	40,664	56,154
MA	11,240	35,497	47,797
ARMA	11,612	36,946	48,573
ARIMA	11,800	37,485	50,025
SARIMA	11,690	37,586	49,734
ARIMAX	12,824	41,082	56,469
SARIMAX	12,797	41,017	56,332
Linear Regression	180,100	787,137	787,156
Random Forest Regressor	22,170	69,635	81,364
XGBRegressor	22,662	71,324	82,831
LGBMRegressor	22,677	71,249	83,301

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

Tabela 11: Comparação dos modelos 10 dia a frente 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	8,770	29,927	38,280
ARX	10,003	34,882	46,327
MA	8,707	29,648	37,915
ARMA	8,877	30,311	39,068
ARIMA	9,363	31,975	40,450
SARIMA	9,842	33,655	44,418
ARIMAX	9,981	34,804	46,249
SARIMAX	10,027	34,963	46,369
Linear Regression	176,691	786,562	786,573
Random Forest Regressor	18,538	62,113	70,415
XGBRegressor	19,021	63,820	71,941
LGBMRegressor	19,033	63,811	72,205

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

Tabela 12: Comparação dos modelos 10 dia a frente 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	12,031	38,705	50,716
ARX	14,899	48,829	65,680
MA	12,401	39,887	52,085
ARMA	11,709	38,210	49,248
ARIMA	11,721	38,292	49,716
SARIMA	12,691	41,520	53,713
ARIMAX	14,889	48,774	65,677
SARIMAX	14,927	48,926	65,568
Linear Regression	174,810	785,399	785,428
Random Forest Regressor	18,559	57,179	74,895
XGBRegressor	18,727	57,666	75,772
LGBMRegressor	18,993	58,395	77,394

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

Tabela 13: Comparação dos modelos 10 dia a frente 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	12,031	38,705	50,716
ARX	14,899	48,829	65,680
MA	12,401	39,887	52,085
ARMA	11,709	38,210	49,248
ARIMA	11,721	38,292	49,716
SARIMA	12,691	41,520	53,713
ARIMAX	14,889	48,774	65,677
SARIMAX	14,927	48,926	65,568
Linear Regression	174,810	785,399	785,428
Random Forest Regressor	18,559	57,179	74,895
XGBRegressor	18,727	57,666	75,772
LGBMRegressor	18,993	58,395	77,394

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

Tabela 14: Comparação dos modelos 30 dia a frente 24h **Treino**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	12,081	38,341	51,436
ARX	13,492	43,223	59,219
MA	11,965	37,924	51,020
ARMA	11,968	38,170	50,727
ARIMA	12,459	39,654	52,744
SARIMA	12,421	39,931	52,650
ARIMAX	13,570	43,451	59,456
SARIMAX	13,597	43,532	59,638
Linear Regression	582,704	2548,346	2548,352
Random Forest Regressor	22,431	70,541	82,102
XGBRegressor	22,500	70,845	82,172
LGBMRegressor	23,669	74,752	85,912

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

Tabela 15: Comparação dos modelos 30 dia a frente 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	9,116	31,080	38,977
ARX	8,628	30,245	43,415
MA	8,993	30,584	38,262
ARMA	8,868	30,399	38,414
ARIMA	10,035	34,234	42,602
SARIMA	9,590	32,836	40,747
ARIMAX	8,599	30,143	43,342
SARIMAX	8,622	30,213	43,365
Linear Regression	572,149	2547,771	2547,774
Random Forest Regressor	18,779	62,971	71,142
XGBRegressor	18,973	63,661	71,760
LGBMRegressor	20,110	67,660	75,404

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

Tabela 16: Comparação dos modelos 30 dia a frente 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	11,730	37,494	49,483
ARX	14,128	46,221	62,818
MA	11,978	38,411	50,429
ARMA	11,846	38,474	49,677
ARIMA	12,075	39,046	50,919
SARIMA	12,751	41,276	54,618
ARIMAX	14,037	45,883	62,737
SARIMAX	14,106	46,143	62,613
Linear Regression	566,324	2546,608	2546,617
Random Forest Regressor	18,799	58,005	75,607
XGBRegressor	18,548	57,215	74,750
LGBMRegressor	19,487	60,140	78,694

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

Tabela 17: Comparação dos modelos 30 dia a frente 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	11,435	36,663	23,656
ARX	13,683	44,708	36,048
MA	11,263	36,134	47,748
ARMA	11,980	38,530	50,774
ARIMA	11,724	37,686	49,853
SARIMA	11,705	37,923	49,960
ARIMAX	13,599	44,357	59,637
SARIMAX	13,680	44,669	60,103
Linear Regression	576,298	2547,743	2547,749
Random Forest Regressor	20,827	65,710	78,726
XGBRegressor	20,818	65,738	78,599
LGBMRegressor	21,913	69,362	82,380

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

Tabela 18: Comparação dos modelos 60 dia a frente 24h **Treino**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	11,435	36,663	23,656
ARX	13,683	44,708	36,048
MA	11,263	36,134	47,748
ARMA	11,980	38,530	50,774
ARIMA	11,724	37,686	49,853
SARIMA	11,705	37,923	49,960
ARIMAX	13,599	44,357	59,637
SARIMAX	13,680	44,669	60,103
Linear Regression	576,298	2547,743	2547,749
Random Forest Regressor	20,827	65,710	78,726
XGBRegressor	20,818	65,738	78,599
LGBMRegressor	21,913	69,362	82,380

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

Tabela 19: Comparação dos modelos 60 dia a frente 24h **Validação**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	8,439	28,898	36,272
ARX	6,569	23,225	37,358
MA	8,346	28,479	35,962
ARMA	9,001	31,006	38,930
ARIMA	9,066	31,015	39,006
SARIMA	9,671	33,513	41,133
ARIMAX	6,519	23,050	37,275
SARIMAX	6,569	23,218	37,262
Linear Regression	1165,336	5189,585	5189,586
Random Forest Regressor	18,374	61,534	69,907
XGBRegressor	17,607	58,834	67,490
LGBMRegressor	20,172	67,886	75,587

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

Tabela 20: Comparação dos modelos 60 dia a frente 24h **Teste**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	12,610	40,176	52,627
ARX	13,150	42,148	60,417
MA	12,704	40,641	52,995
ARMA	12,540	40,189	52,283
ARIMA	12,727	40,607	53,138
SARIMA	13,580	43,340	56,349
ARIMAX	13,006	41,636	60,147
SARIMAX	13,304	42,791	60,088
Linear Regression	1153,594	5188,422	5188,427
Random Forest Regressor	18,682	57,628	75,156
XGBRegressor	18,106	55,633	73,457
LGBMRegressor	19,832	61,419	79,294

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

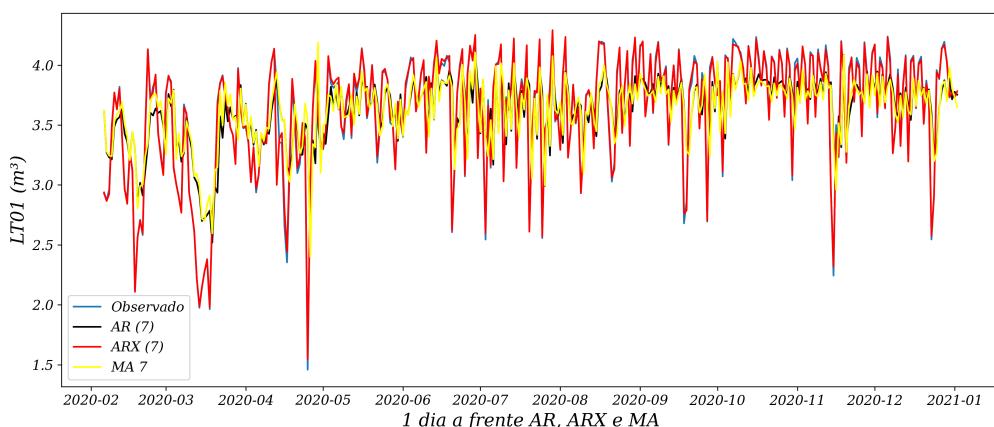
Tabela 21: Comparação dos modelos 60 dia a frente 24h **Completo**

Modelos	Erros		
	MAPE	MAE	RMSE
AR	12,610	40,176	52,627
ARX	13,150	42,148	60,417
MA	12,704	40,641	52,995
ARMA	12,540	40,189	52,283
ARIMA	12,727	40,607	53,138
SARIMA	13,580	43,340	56,349
ARIMAX	13,006	41,636	60,147
SARIMAX	13,304	42,791	60,088
Linear Regression	1153,594	5188,422	5188,427
Random Forest Regressor	18,682	57,628	75,156
XGBRegressor	18,106	55,633	73,457
LGBMRegressor	19,832	61,419	79,294

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

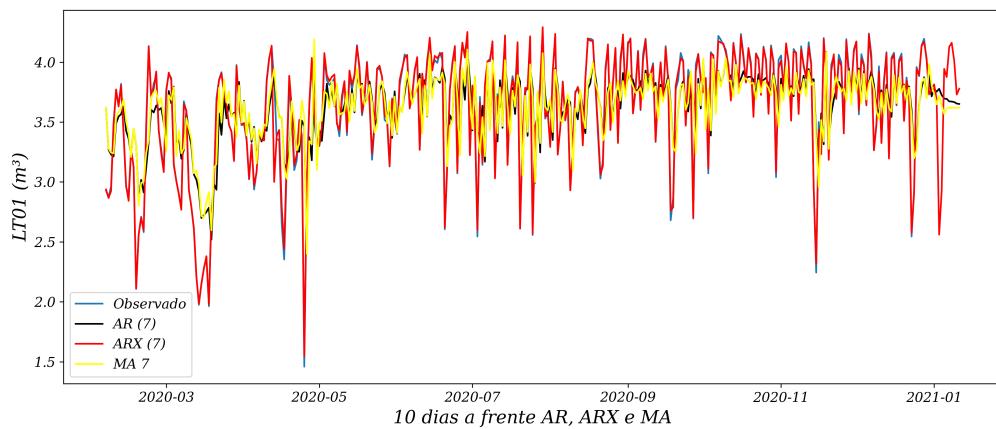
B Apêndice - Modelos AR(7), ARX (7) e MA (7) 24h

Figura 38: Comparação dos modelos AR, ARX e MA, 1 dia a frente



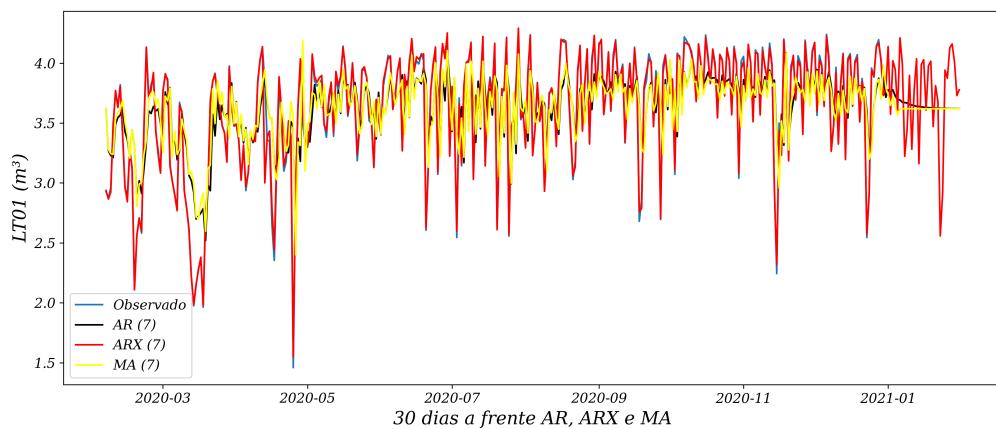
Fonte: Autoria própria.

Figura 39: Comparação dos modelos AR, ARX e MA, 10 dias a frente



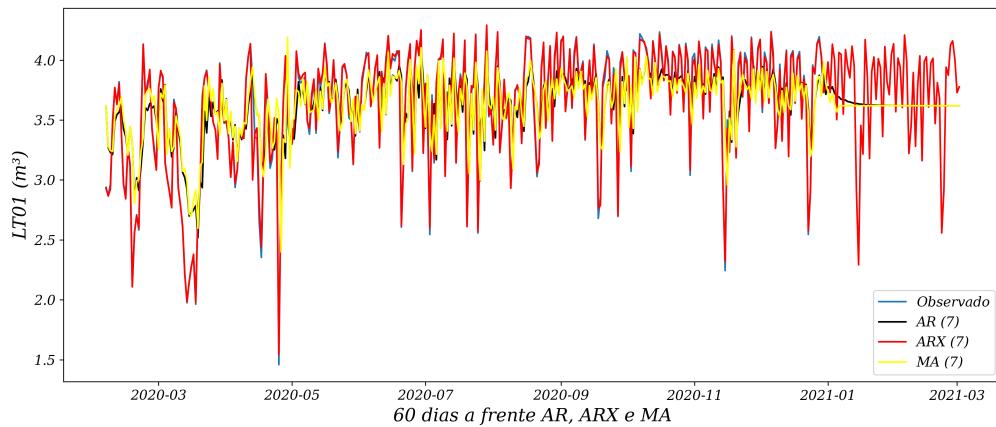
Fonte: Autoria própria.

Figura 40: Comparação dos modelos AR, ARX e MA, 30 dias a frente



Fonte: Autoria própria.

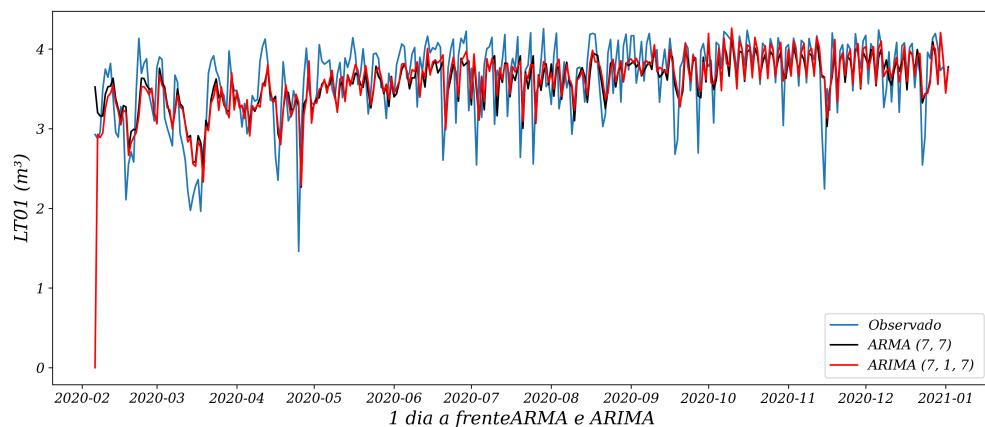
Figura 41: Comparação dos modelos AR, ARX e MA, 60 dias a frente



Fonte: Autoria própria.

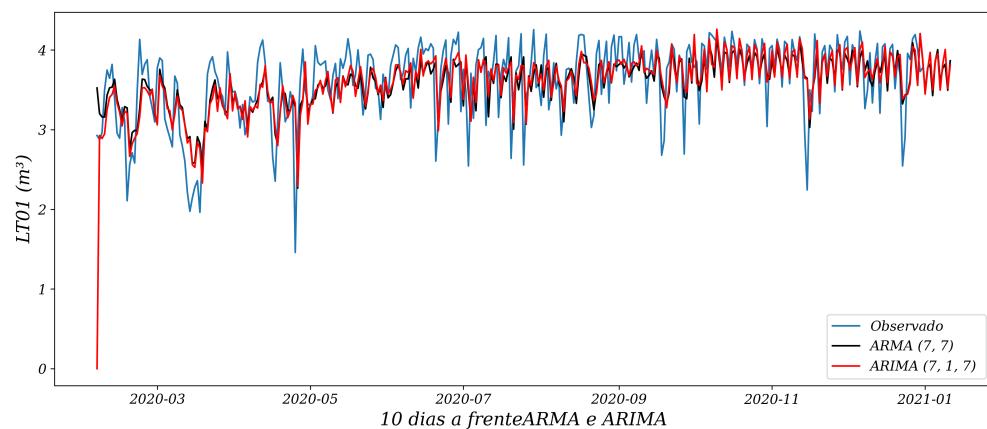
C Apêndice - Modelos ARMA(7,7) e ARIMA (7,1,7) 24h

Figura 42: Comparação dos modelos ARMA e ARIMA, 1 dia a frente



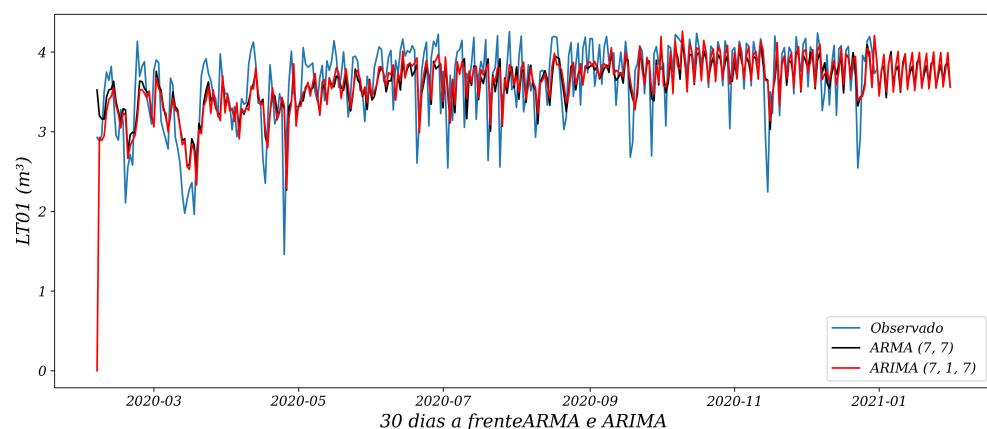
Fonte: Autoria própria.

Figura 43: Comparação dos modelos ARMA e ARIMA, 10 dias a frente



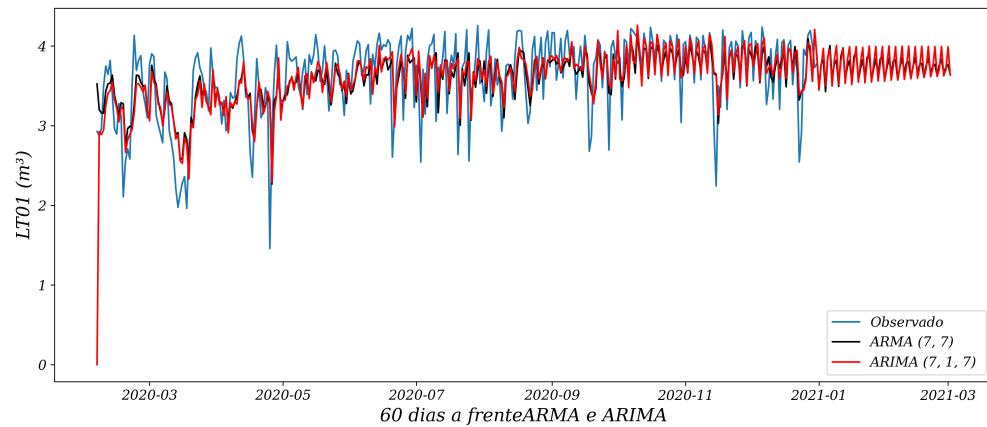
Fonte: Autoria própria.

Figura 44: Comparação dos modelos ARMA e ARIMA, 30 dias a frente



Fonte: Autoria própria.

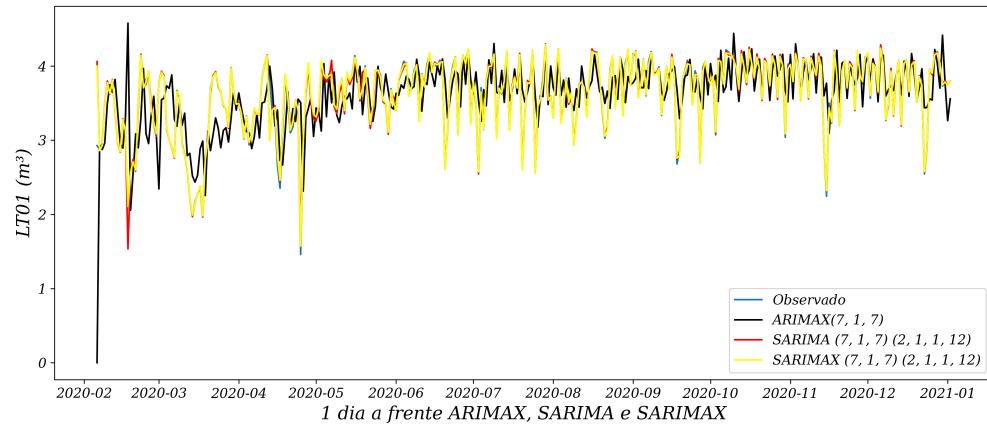
Figura 45: Comparação dos modelos ARMA e ARIMA, 60 dias a frente



Fonte: Autoria própria.

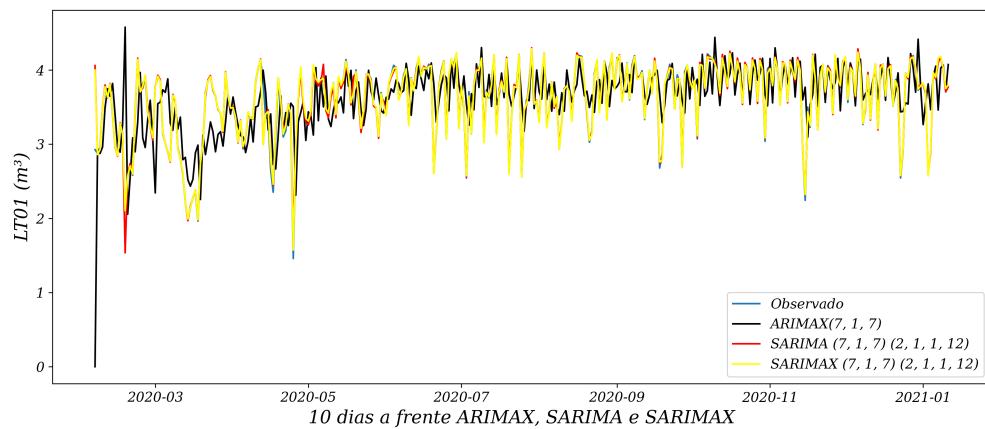
D Apêndice - Modelos ARIMAX (7,1,7), SARIMA (7,1,7) (2,1,1,12) e SARIMAX (7,1,7) (2,1,1,12) 24h

Figura 46: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 1 dia a frente



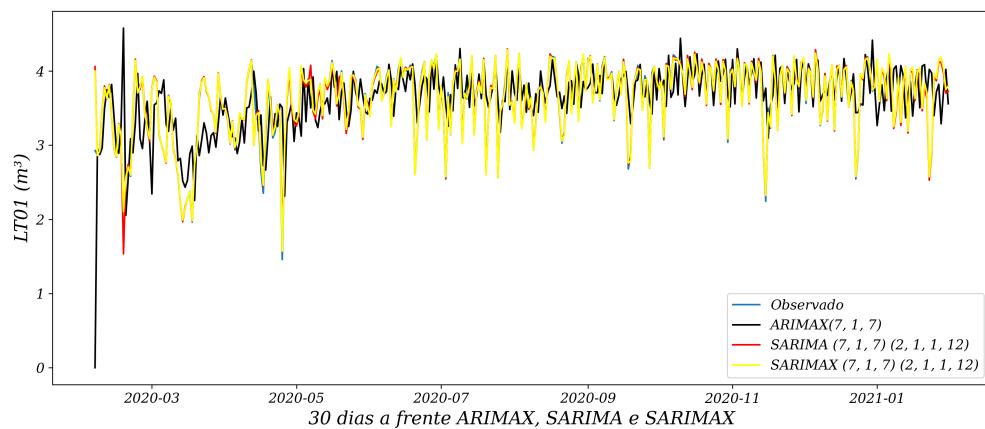
Fonte: Autoria própria.

Figura 47: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 10 dias a frente



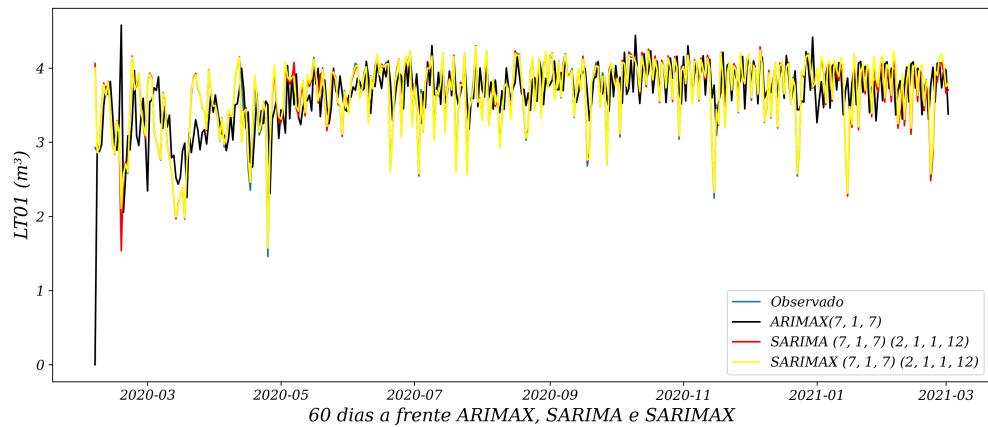
Fonte: Autoria própria.

Figura 48: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 30 dias a frente



Fonte: Autoria própria.

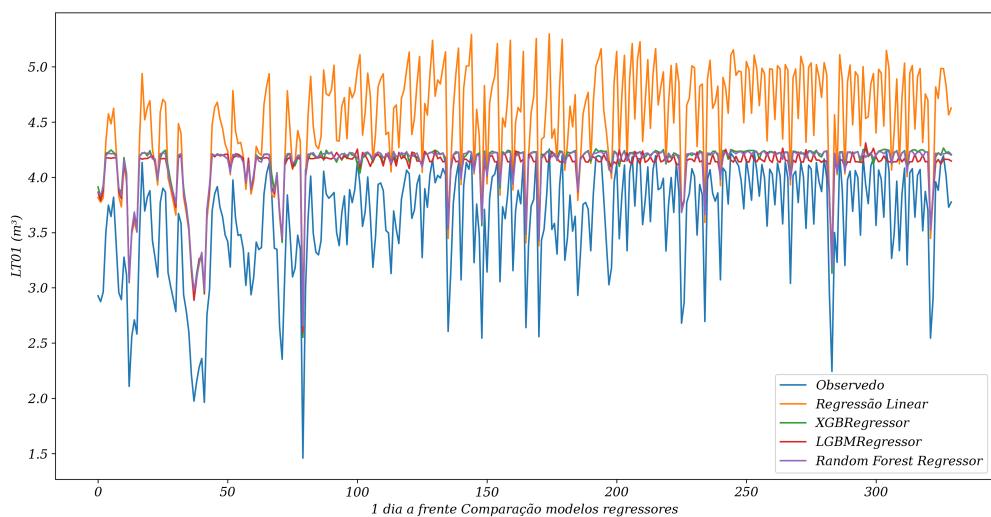
Figura 49: Comparação dos modelos ARIMAX, SARIMA e SARIMAX, 60 dias a frente



Fonte: Autoria própria.

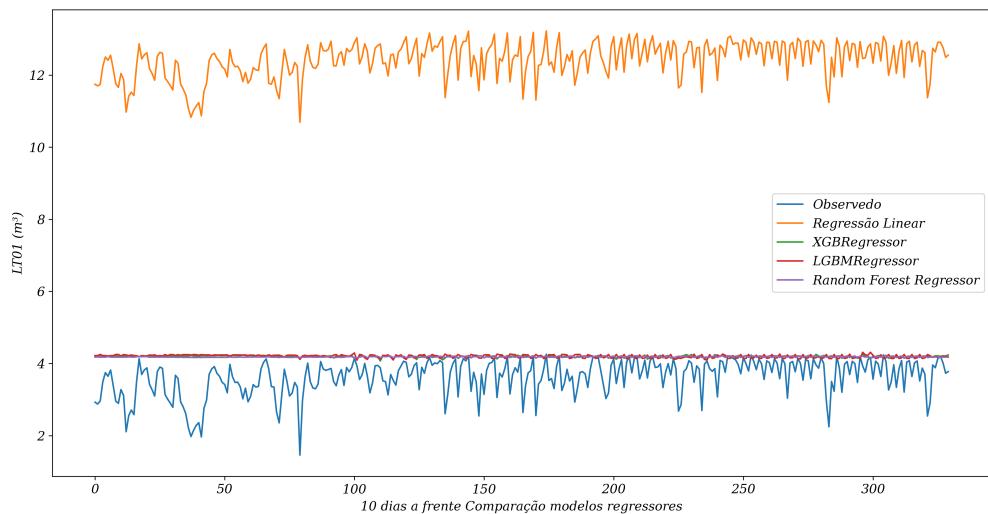
E Apêndice - Modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória 24h

Figura 50: Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 1 dia a frente



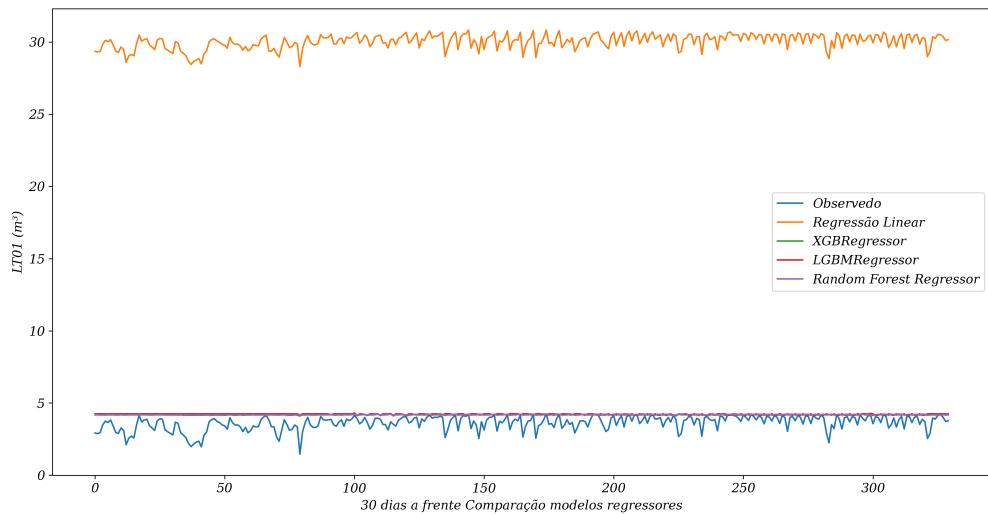
Fonte: Autoria própria.

Figura 51: Comparação dos modelos Regressão linear, XGB Regressão, Ligh GBM Regressão e Regressão de Floresta Aleatória, 10 dias a frente



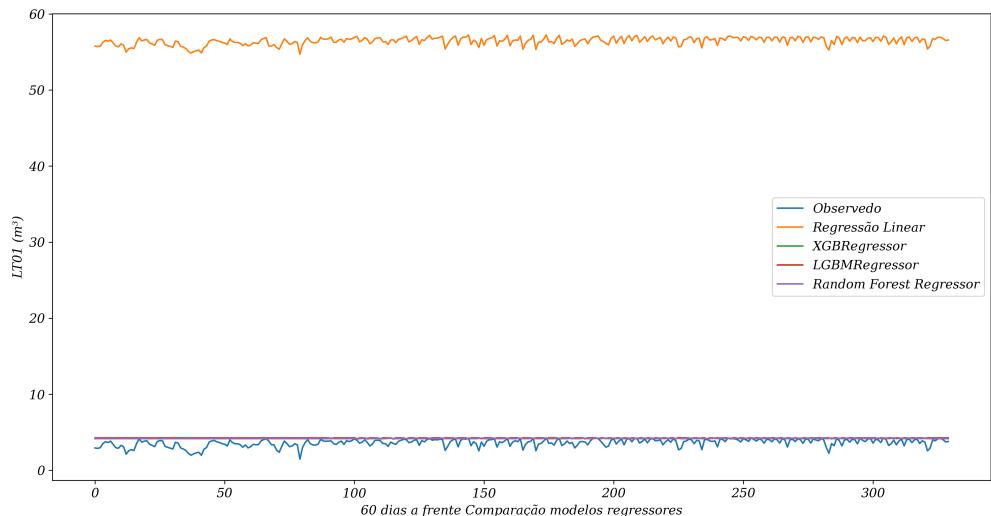
Fonte: Autoria própria.

Figura 52: Comparação dos modelos Regressão linear, XGB Regressão, Ligh GBM Regressão e Regressão de Floresta Aleatória, 30 dias a frente



Fonte: Autoria própria.

Figura 53: Comparação dos modelos Regressão linear, XGB Regressão, Ligth GBM Regressão e Regressão de Floresta Aleatória, 60 dias a frente



Fonte: Autoria própria.