UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Lucas Felippe Mateus

# Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020

Araranguá
2022

Lucas Felippe Mateus

# Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020

Trabalho de Conclusão de Curso do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.
Orientador: Prof. Fabrício de Oliveira Ourique, Dr.

Araranguá
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Mateus, Lucas Felippe
Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020 / Lucas Felippe Mateus ; orientador, Fabrício de Oliveira Ourique, 2022.
31 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Engenharia de Computação, Araranguá, 2022.

Inclui referências.

1. Engenharia de Computação. 2. Séries temporais. 3. Modelo preditivo. 4. Excesso de mortes. 5. Subnotificação de mortes por COVID-19. I. Ourique, Fabrício de Oliveira. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Computação. III. Título.

Lucas Felippe Mateus

**Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de "Bacharel em Engenharia de Computação" e aprovado em sua forma final pelo Curso de Graduação em Engenharia de Computação.

Araranguá, 21 de Março de 2022.


Documento assinado digitalmente
Analucia Schiaffino Morales
Data: 23/03/2022 13:57:24-0300
CPF: 622.256.420-87
Verifique as assinaturas em https://v.ufsc.br

---

Profª Analúcia Schiaffino Morales, Dra.
Coordenadora do Curso

**Banca Examinadora:**


Documento assinado digitalmente
Fabricio de Oliveira Ourique
Data: 24/03/2022 06:10:49-0300
CPF: 916.167.860-00
Verifique as assinaturas em https://v.ufsc.br

---

Prof. Fabrício de Oliveira Ourique, Dr.
Orientador
Universidade Federal de Santa Catarina


Documento assinado digitalmente
Analucia Schiaffino Morales
Data: 23/03/2022 13:57:45-0300
CPF: 622.256.420-87
Verifique as assinaturas em https://v.ufsc.br

---

Profª Analúcia Schiaffino Morales, Dra.
Avaliadora
Universidade Federal de Santa Catarina


Documento assinado digitalmente
Alexandre Leopoldo Goncalves
Data: 23/03/2022 21:21:34-0300
CPF: 786.740.349-34
Verifique as assinaturas em https://v.ufsc.br

---

Prof. Alexandre Leopoldo Gonçalves, Dr.
Avaliador
Universidade Federal de Santa Catarina

---

Profª Luciana Bolan Frigo, Dra.
Avaliadora Suplente
Universidade Federal de Santa Catarina

# Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020

## *Implementation of a time series forecasting model to estimate excess deaths in Brazil in 2020*

Lucas Felippe Mateus [*] Fabrício de Oliveira Ourique [†]

2022, Março

## Resumo

Desde o início de 2020 o mundo vem lidando com as consequências de uma pandemia. Esta acometeu diversos aspectos da vida cotidiana, afetando a economia e os movimentos mundiais. Em prol de freia-la, foram realizados estudos que objetivam descrever e compreender o seu comportamento. Dentro desses estudos se destacou que uma característica muito importante a ser analisada é a questão de subnotificação de casos e mortes. Por isso, nesta pesquisa tem-se como objetivo implementar um modelo preditor para estimar a quantia de óbitos causados pela COVID-19 no Brasil em 2020. A implementação desse modelo conta com os conceitos da classe de modelos ARIMA e os dados extraídos da base do Sistema Unificado de Saúde. Após a previsão, estima-se que a COVID-19 contribuiu, em média, para um excedente de 713 óbitos diários. Além disso, comparando o resultado do modelo com o registro de mortes por COVID-19, obtido através do repositório de dados do Centro de Ciência e Engenharia de Sistemas da Universidade Johns Hopkins, conclui-se que há subnotificação de óbitos causados por essa doença durante o ano de 2020 no Brasil.

**Palavras-chaves**: Séries temporais. Modelo preditivo. Excesso de mortes. Subnotificação de mortes por COVID-19.

[*] lucasfelippemateus@yahoo.com.br
[†] fabricio.ourique@gmail.com

# Implementação de um modelo de previsão de séries temporais para estimar o excedente de mortes no Brasil em 2020

## *Implementation of a time series forecasting model to estimate excess deaths in Brazil in 2020*

Lucas Felippe Mateus [*] Fabrício de Oliveira Ourique [†]

2022, Março

## Abstract

Since the beginning of 2020, the world has been dealing with the consequences of a pandemic. This affected several aspects of everyday life, affecting the economy and world movements. In order to stop it, studies were carried out to describe and understand its behavior. Within these studies, it was highlighted that a very important feature to be analyzed is the issue of underreporting of cases and deaths. Therefore, this research aims to implement a predictor model to estimate the number of deaths caused by COVID-19 in Brazil during 2020. The implementation of this model relies on the concepts of the ARIMA model class and the data extracted from the base of the Unified Health System. After the forecast, it is estimated that COVID-19 contributed, on average, to a surplus of 713 daily deaths. In addition, comparing the result of the model with the registration of deaths by COVID-19, obtained through the data repository of the Center for Science and Systems Engineering at Johns Hopkins University, it is concluded that there is underreporting of deaths caused by this disease during the year 2020 in Brazil.

**Key-words**: Time series. Predictive model. Surplus deaths. Underreporting of deaths from COVID-19.

[*] lucasfelippemateus@yahoo.com.br
[†] fabricio.ourique@gmail.com

# 1 Introdução

No final do ano de 2019, a China reportou casos de pneumonia, que ocorreram na província de Hubei, mais especificamente em Wuhan, à Organização Mundial da Saúde (OMS). Tais casos aconteceram em decorrência de um novo coronavírus, o qual ficou conhecido como 2019-nCoV. Autoridades da área da saúde encontraram evidências de transmissão ao longo de quatro gerações, ou seja, a primeira pessoa infectada contraiu o vírus de uma fonte não humana e acabou infectando outra, que infectou um terceiro indivíduo e este transmitiu a infecção para mais um. Essas evidências demonstram que a transmissão de humano para humano conseguiu se sustentar. Além disso, muitos países como Austrália, Camboja, Canadá, França, Alemanha, Japão, Nepal, Cingapura, Coreia do Sul, Taiwan, Tailândia, Emirados Árabes Unidos, Estados Unidos e Vietnã confirmaram casos associados às viagens. Dessa forma, a crescente quantidade de casos e mortes começou a se mostrar um desafio à saúde pública e políticas de governo (PHELAN; KATZ; GOSTIN, 2020).

Depois que os casos fora da China aumentaram em treze vezes, fazendo com que a quantia de países com casos confirmados triplicasse, a Organização Mundial da Saúde (OMS), preocupada com os níveis alarmantes de propagação, declarou o surto do novo coronavírus como uma pandemia em março de 2020. No momento da declaração, ainda haviam estudos de projeções que indicavam crescimento no número de casos (CUCINOTTA; VANELLI, 2020).

A pandemia da COVID-19 afetou rapidamente a vida cotidiana e os negócios, perturbando o comércio e os movimentos mundiais. Sem esquecer de mencionar o impacto na vida de milhares de pessoas, que ou estão doentes, ou estão morrendo por causa dessa doença. Na área da saúde surgiram grandes desafios como diagnósticos, tratamentos para os casos suspeitos e confirmados, sobrecarga em médicos e outros profissionais, além da interrupção na cadeia de suprimentos hospitalares. Já na economia, o mundo se deparou com a desaceleração significativa no crescimento da receita em paralelo a perdas de negócios nacionais e internacionais, sendo que os que ainda se mantiveram acabaram sendo afetados pelo interrompimento na cadeia de fornecimento de produtos. Enquanto no âmbito social ocorreu o distanciamento entre colegas, amigos e familiares, o fechamento de estabelecimentos de entretenimento e restaurantes, o estresse indevido entre a população e outras dificuldades (HALEEM; JAVAID; VAISHYA, 2020).

Para Morato *et al.* (2020), um dos pontos importantes para a análise dessa pandemia é a compreensão sobre a subnotificação de casos. De acordo com estes autores, trabalhos recentes alertam contra a grande quantia de subnotificação no Brasil, um país no qual, devido à falta de testes, está considerando apenas pacientes com sintomas graves ou falecidos. Essa abordagem acaba fomentando um grande percentual de subnotificação. Quanto maior o número de infectados, mais subnotificações ocorrem, o que causa incerteza frente a casos confirmados e óbitos, ou seja, para a margem de pessoas infectadas, a subnotificação faz com que a mortalidade pela doença diminua enquanto quantia de mortes aumenta. Caso essas incertezas não sejam consideradas, pode-se chegar a decisões precipitadas.

Segundo Orellana *et al.* (2021), o elevado excedente de mortes, devido às que não são explicadas diretamente pela COVID-19 e às que ocorrem fora do hospital, sugerem alta subnotificação de óbitos por COVID-19. Essa subnotificação reforça a dispersão do SARS-CoV-2, além de atribuir incerteza as mortes associadas a sintomas respiratórios. Isso faz com que o excesso de mortalidade seja um indicador importante durante a definição de prioridades na hora de tomar decisões para combater a pandemia, especialmente nos

países em desenvolvimento, onde existem dificuldades para o diagnóstico adequado e o enfrentamento da doença.

De acordo com Shumway e Stoffer (2000), um epidemiologista que estivesse interessado em entender um vírus gripal durante algum período poderia utilizar a análise de séries temporais para ter uma melhor compreensão sobre o número de casos observados. A análise de séries temporais é uma ferramenta de grande impacto em diversas aplicações científicas que buscam analisar dados experimentais observados ao longo de diferentes pontos no tempo com o intuito de alcançar uma modelagem e uma inferência estatística.

Uma série temporal é um conjunto de valores de uma medida obtida em tempos sucessivos, geralmente com intervalos iguais entre eles. Em outras palavras, uma série temporal é uma coleção de observações feitas sequencialmente no tempo que descrevem o comportamento de determinada variável (CHATFIELD, 1996).

Esta pesquisa propõe analisar uma série temporal, que representa a quantidade de mortes diárias no Brasil desde 2014 até 2019, para implementar um modelo de previsão com a intenção de estimar o número de óbitos diários no Brasil em 2020. A partir dessa estimativa, é possível compará-la com os dados reais e extrair, dessa diferença, uma noção do excedente de mortes neste ano.

Neste trabalho, será abordada a fundamentação teórica na Seção 2, a qual trará conceitos teóricos importantes para o entendimento das seções seguintes. A seguir, serão apresentados os trabalhos correlatos na Seção 3. Depois, a metodologia estará disposta na Seção 4, antecedendo os resultados que estarão descritos na Seção 5. Enfim, na Seção 6, serão explanadas as conclusões extraídas deste estudo.

# 2 Fundamentação Teórica

## 2.1 Séries temporais

Uma série temporal é uma coleção de observações de uma determinada variável feita sequencialmente no tempo. As séries temporais podem ser subdivididas em contínuas ou discretas, isto é, se as observações da série foram feitas continuamente em uma determinada faixa de tempo, esta série é dita contínua, enquanto que se as observações da série foram feitas em momentos específicos, geralmente igualmente espaçados em uma determinada faixa de tempo, esta série é dita discreta. Ressaltando que mesmo que a variável observada na coleta contínua assuma um conjunto de valores discretos, a série continua sendo considerada contínua. O mesmo serve para a série discreta, ela permanecerá discreta mesmo que a variável observada assuma valores contínuos. Há várias formas de se obter uma série discreta, uma delas é através da amostragem de uma série contínua, outra é através da agregação dos valores da variável observada em intervalos de tempo igualmente espaçados. Um dos pontos importantes na análise de séries temporais é o fato de que a ordem das observações deve ser levada em conta e que as observações, usualmente, não são independentes. Quando as observações são dependentes, é possível utilizar valores passados para prever valores futuros. Nesse cenário, se for possível prever os valores exatos de uma série, ela é classificada como determinística, entretanto, a maioria das séries cai na classificação estocástica, a qual representa séries que foram parcialmente previstas (CHATFIELD, 1996).

## 2.2 Componentes de uma séries temporal

Em prol de facilitar a interpretação e o entendimento do conjunto de dados de uma série temporal, é possível decompor esta série em quatro partes, isto é, nível, tendência, sazonalidade e ruído. O nível é o valor médio da série. A tendência é o comportamento crescente ou decrescente da série ao longo do tempo, que em muitos dos casos aparece de forma linear. A sazonalidade representa os padrões de repetição ou ciclos de comportamento ao longo do tempo. O ruído é a variabilidade nas observações. Por fim, pode-se dizer que as séries temporais têm nível, a maioria delas possuí ruído, mas a tendência e a sazonalidade são opcionais (BROWNLEE, 2020).

## 2.3 Análise de séries temporais

Existem diversas intenções por trás da análise de séries temporais como, por exemplo, obter a descrição das medidas das principais propriedades da série ou obter a previsão de valores futuros. A análise de séries temporais se preocupa com a avaliação das propriedades do modelo probabilístico que gerou a série temporal observada. Ressaltando que tais propriedades podem sofrer alterações provenientes de diferentes fontes como: efeito sazonal, onde a série temporal exibe uma variação periódica; variação cíclica em um período específico devido a um fenômeno físico; tendência de variação da média ao longo do tempo; e flutuações. De forma abrangente, uma série que não possui mudanças sistemáticas nem na sua média, nem na sua variância e passou por um processo de remoção de variações periódicas, é conhecida como uma série estacionária. É importante ter o conhecimento das séries estacionárias, pois, a maior parte da teoria de probabilidade em torno das séries temporais é direcionada para as séries estacionárias. Portanto, antes de aplicar modelos probabilísticos como Média Móvel ou Autorregressivo, é preciso fazer um tratamento na série temporal com o objetivo de retirar essas fontes de variação, tornando-a estacionária. Por fim, existe uma propriedade muito importante na análise de séries temporais conhecida como autocorrelação, que mede a correlação entre observações em diferentes momentos dentro da série. Os coeficientes da autocorrelação proveem esclarecimento sobre o modelo probabilístico que gerou a série observada. Estes coeficientes podem ser melhor observados através do correlograma, o qual nem sempre é de fácil interpretação, pois, apresenta diferentes comportamentos dependendo da série que pode ser randômica, não estacionária, possuir flutuações ou até mesmo ter um período de correlação curto (CHATFIELD, 1996).

## 2.4 Previsão de séries temporais

A previsão de séries temporais consiste em fazer a extrapolação no tratamento estatístico clássico sobre os dados das séries temporais. Sendo que a previsão envolve ajustar dados históricos em modelos e utilizá-los para prever observações futuras. Na previsão é importante destacar que o futuro está completamente indisponível e só deve ser estimado a partir do que já aconteceu. Portanto, conclui-se que a habilidade de um modelo de previsão de séries temporais é determinada por seu desempenho em prever futuras observações (BROWNLEE, 2020).

Para BOX *et al.* (2015), as previsões geralmente são necessárias quando se deseja fazer projeções, as quais variam de acordo com cada problema. O objetivo das previsões é conseguir uma função que utilize as observações disponíveis até um determinado momento do tempo para estimar os valores que a série vai assumir nos momentos subsequentes com o menor desvio quadrado médio possível entre o valor estimado e o valor que de fato assumiu.

Muitos procedimentos de previsão são baseados em um modelo de série temporal. Portanto, é útil estar familiarizado com uma gama de modelos e ter uma compreensão completa do processo de modelagem aplicado aos dados da série temporal antes de começar a olhar para métodos de previsão. A classe de modelos ARIMA é uma importante ferramenta de previsão além de ser a base de muitas ideias fundamentais na análise de séries temporais. A referência original para esse estudo é Box e Jenkins (1970), por isso que os modelos ARIMA são às vezes chamados de modelos Box-Jenkins (CHATFIELD, 1996).

Agora serão introduzidos os diferentes componentes desta classe geral de modelos, começando pelo modelo autorregressivo. Em seguida apresentando o modelo de médias móveis. Partindo do conhecimento desses dois modelos, será trabalhado o modelo autorregressivo de médias móveis. Enfim, finalizando com o modelo autorregressivo integrado à média móvel.

### 2.4.1 Modelo Autorregressivo

Uma série temporal $X_t$ é dita ser um processo autorregressivo de ordem p, AR(p), se for uma soma linear ponderada dos p valores anteriores mais uma componente de aleatoriedade, ou seja,

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t.$$

O exemplo mais simples de um processo autorregressivo é o caso de primeira ordem, AR(1), dado por

$$X_t = \phi_1 X_{t-1} + Z_t$$

o qual tem uma única solução estacionária desde que $|\phi| < 1$ (CHATFIELD, 2000).

O modelo autorregressivo usa observações anteriores como entrada para uma equação de regressão para prever o valor na próxima observação. Apesar de ser uma ideia simples, pode resultar em previsões precisas em diversos problemas de séries temporais. Um modelo de regressão, modela um valor de saída com base em uma combinação linear de valores de entrada, isto é,

$$X_{t+1} = b_0 + b_1 X_t + b_2 X_{t-1} + \ldots + b_{p+1} X_{t-p}$$

onde $b_0, \ldots, b_{p+1}$ são coeficientes obtidos através da otimização do modelo de treinamento, $X_t, \ldots, X_{t-p}$ são os valores de entrada e p representa a quantia de regressões realizadas. Essa característica de utilizar dados da mesma variável de entrada em etapas de tempo anteriores é o que deu a esse modelo o nome de autorregressivo (BROWNLEE, 2020).

### 2.4.2 Modelo Média Móvel

Uma série temporal $X_t$ é dita ser um processo de média móvel de ordem q, MA(q), se for uma soma linear ponderada dos últimos q choques aleatórios, isto é,

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}.$$

onde $Z_t$ denota um processo puramente aleatório com média zero e variância constante, $\theta_1, \ldots, \theta_q$ são coeficientes que ponderam os choques, $Z_t, \ldots, Z_{t-q}$ são os choques aleatórios e q representa a quantia de choques aleatórios que foram considerados.

O processo de média móvel é de grande relevância na teoria matemática, pois, consegue trabalhar com qualquer processo estacionário expressando-o como a soma de

dois tipos de processos, um dos quais é não determinístico enquanto o outro é linearmente determinístico (CHATFIELD, 2000).

O valor da média móvel de uma série também pode ser usado diretamente para fazer previsões. Apesar de ter o mesmo nome, o processo de previsão utilizando média móvel é diferente do processo de suavização por média móvel, este é utilizado para remover o ruído entre as observações e expor melhor o sinal. O modelo de previsão por média móvel é um modelo ingênuo e só aplicável quando novas observações são disponibilizadas, por exemplo, faz-se a primeira previsão com os dados existentes, em seguida coleta a nova observação, atualiza o modelo, faz a previsão seguinte e assim em diante (BROWNLEE, 2020).

### 2.4.3 Modelo Autorregressivo de Média Móvel

É possível fazer a combinação dos modelos trabalhados anteriormente para obter um modelo autorregressivo de médias móveis, ARMA(p,q), com p termos autorregressivos e q termos de média móvel. Esta combinação pode ser escrita da seguinte forma:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \ldots + \theta_q Z_{t-q}.$$

O interessante dessa abordagem é que muitos conjuntos de dados reais podem ser aproximados com menos parâmetros por um modelo ARMA misto do que processos AR ou MA puros (CHATFIELD, 2000).

### 2.4.4 Modelo Autorregressivo Integrado a Média Móvel

Este modelo é amplamente utilizado para análise e previsão de séries temporais, pois, captura um conjunto de diferentes estruturas temporais na série e fornece um método simples para fazer as previsões. O modelo ARIMA é visto como uma generalização do modelo ARMA, isto é, mantém os componentes autorregressivo e média móvel e adiciona a noção de integração (BROWNLEE, 2020).

Na prática, muitas séries temporais não são estacionárias, portanto, não podemos aplicar processos do tipo AR, MA ou ARMA estacionários diretamente. Nesses casos, é necessário torná-las estacionárias e uma das formas de se alcançar isso é aplicando a diferenciação. Caso a série original seja diferenciada $d$ vezes antes de ajustar um processo ARMA(p, q), então o modelo para a série original não diferenciada é dito ser um processo ARIMA(p, d, q) sendo que o novo parâmetro $d$ denota o número de diferenças tomadas. Ao ajustar os modelos AR e MA, a principal dificuldade é avaliar a ordem do processo. Com os modelos ARIMA, há um problema adicional na escolha da ordem necessária de diferenciação (CHATFIELD, 2000).

## 2.5 Verificando o desempenho da previsão

Medidas de desempenho para previsão de séries temporais fornecem um resumo da habilidade e capacidade do modelo de realizar tais previsões. Tendo em vista que as séries temporais geralmente se concentram na previsão de valores reais, adotaremos as medidas de desempenho para avaliar previsões de valores reais. O erro de previsão, ou erro residual, é calculado como o valor esperado menos o valor previsto. Este cálculo pode ser feito para cada previsão, o que resultará em uma série temporal de erros. Além disso, ainda é possível calcular a média dos erros de previsão. Como estes erros podem ser tanto positivos como negativos, a média desses valores pode ser zero, ou seja, caso ideal onde a previsão

ocorreu sem erros, pode ser maior que zero (positiva), o que indica que o modelo tem uma tendência a subestimar, ou pode ser menor que zero (negativa), mostrando que o modelo tem uma tendência de superestimar as observações. Podemos forçar o cálculo do erro dar positivo elevando a diferença entre valor esperado e previsto ao quadrado. Em seguida, é possível tirar a média. E assim se obtém o erro quadrático médio. Entretanto, deve-se estar ciente que este cálculo coloca mais peso em erros maiores e não se encontra na mesma escala que as observações, pois, foi elevado ao quadrado. Para uma melhor comparação, é possível resgatar o erro quadrático médio para a mesma unidade que as observações. Fazendo a extração da raiz quadrada do erro quadrático médio obtém-se o erro quadrático médio na mesma escala em que as observações foram feitas. Lembrando que quanto mais próximo de zero este erro for, melhor é o modelo de previsão (BROWNLEE, 2020).

### 2.6 Modelos de previsão de séries temporais em Python

Python é uma linguagem bem adequada para o desenvolvimento e prototipagem rápida com arcabouço para apoiar o desenvolvimento de grandes aplicações. Além disso, Python é amplamente utilizado para ciência de dados e aprendizado de máquina devido ao excelente conjunto de bibliotecas que dão suporte à estas aplicações. Um complemento ao Python importante para a previsão de séries temporais é o SciPy, que é um ecossistema de bibliotecas para matemática, ciência e engenharia. Dentro deste ecossistema, é possível encontrar bibliotecas tais como NumPy para operações eficientes com matrizes, Matplotlib para traçar os dados, Pandas para a manipulação de dados, Statmodels para a modelagem de séries temporais e scikit-learn para o aprendizado de máquina (BROWNLEE, 2020).

De acordo com Brownlee (2020), a biblioteca Pandas oferece ferramentas de alta desempenho para carregar e manipular dados em Python, retornando estruturas convenientes e fáceis para representar os dados como DataFrame e Series. Enquanto a biblioteca Statsmodels provê ferramentas para modelagem e testes estatísticos, além de também contar com ferramentas dedicadas à análise e à previsão de séries temporais. Dentre as funcionalidades do Statsmodels, vale destacar aquelas que são relevantes para a previsão de séries temporais, que são testes estatísticos, como o teste de Dickey-Fuller aumentado, gráficos de análise de séries temporais, como função de autocorrelação (ACF) e função de autocorrelação parcial (PACF), e modelos de séries temporais, como o autorregressivo (AR).

## 3 Trabalhos Correlatos

Após pesquisar nas bases do Google Acadêmico e IEEEXplore, utilizando as buscas: *'time series AND Prediction'*, *'time series AND (Moving Average OR Auto Regressive) AND Prediction'* e *'time series AND (Moving Average OR Auto Regressive) AND Prediction AND COVID-19'*, foram elencadas cerca de vinte pesquisas que expunham alguma aplicação com ou alguma análise feita sobre séries temporais. Dentro desse conjunto, priorizaram-se os trabalhos que abordavam o estudo de séries temporais direcionando-o para a área da saúde. Em última análise, considerou-se que os mais relevantes para o contexto desta pesquisa, eram aqueles realizados dentro do cenário da pandemia da COVID-19.

Finalizando a leitura, encontram-se trabalhos que propõem a utilização de modelos de previsão de séries temporais para modelar o cenário da pandemia, os quais objetivam auxiliar a administração pública no remanejamento de recursos para que o sistema de saúde consiga suprir a necessidade da população acometida pela COVID-19. Vale destacar

os trabalhos que conseguiram uma previsão com um fator de erro baixo utilizando dados dos países mais afetados pela pandemia.

- Liu *et al.* (2021) propõem a aplicação de quatro modelos para resolver o problema de previsão e análise da série temporal da epidemia da COVID-19. Série que possui dados amostrais limitados, características não lineares e de alta dimensão. O primeiro modelo é o de regressão logística, o qual é uma generalização do modelo de análise de regressão linear. Este modelo difere de outros métodos de previsão utilizados para explorar fatores de risco. O segundo é modelo de séries temporais Autoregressivo Integrado a Média Móvel, do inglês *Autoregressive Integrated Moving Average* (ARIMA), que tem como principal aplicação a análise de séries temporais que não são estacionárias e não têm tendências de variação sazonal. Tal modelo ainda conta com casos especiais como o modelo Autorregressivo ou *Autoregressive* (AR), o modelo de Média Móvel, também encontrado como *Moving Average* (MA), e o modelo Autorregressivo de Médias Móveis, derivado da literatura em inglês *Autoregressive Moving Average* (ARMA). Em terceiro lugar, temos os modelos de dinâmica de doenças infecciosas: o *Susceptible, Infectious and Recovered* (SIR) e sua extensão o *Susceptible, Exposed, Infectious and Recovered* (SEIR). Estes modelos consideram o modo de transmissão da doença, por exemplo, patógeno e rota, além de algumas características da população tais como suscetibilidade, infectividade e imunidade. Por fim, o último modelo é de rede neural dinâmica que combina o modelo não linear autorregressivo, *Nonlinear Auto-Regressive* (NAR), com um perceptron de várias camadas, visando fazer um mapeamento não linear do sinal com vários atrasos para ajustar a saída do modelo em um determinado momento. Dessa forma concluem que a rede neural dinâmica NAR é melhor na previsão da nova epidemia, com o erro máximo de 3,6% e mínimo de -0,3%.

- Kumar *et al.* (2021) destacam que analisar e tomar decisões de controle frente a pandemia se tornaram tarefas difíceis devido a um crescimento exponencial nos casos de coronavírus, às limitações no âmbito de recursos humanos e à necessidade de dispor de tratamentos para atender uma grande quantia de pacientes dentro de um período adequado. Sabendo que um modelo computacional automatizado poderia auxiliar na tomada de decisões frente a atendimentos médicos e que a análise estatística de padrões e a visualização dos dados com séries temporais são amplamente aceitas, propuseram uma modelagem de previsão, que seria aplicada sobre os dados do centro de pesquisa John Hopkins, para entender o curso das ocorrências provenientes do surto da COVID19 e assim verificar se esses números aumentariam ou diminuiriam na Índia em um futuro próximo. Com isso, concluíram que os modelos de séries temporais como *Autoregressive integrated moving average* (ARIMA), *Moving Average* (MA) e *Autoregressive* (AR) foram capazes de estimar a situação atual do surto de COVID-19.

- Bertozzi *et al.* (2020) detalham três modelos básicos de transmissão de doenças para previsão e avaliação do curso da pandemia em escala regional. Também mostram que estes modelos podem ser adequados aos dados que emergiram de governos locais e nacionais, o que evidencia a utilidade de modelos parcimoniosos, ou seja, modelos que envolvam o mínimo possível de parâmetros, para dados iniciais. Estes autores trabalham os modelos macroscópicos de crescimento exponencial, de processo de ramificação auto-excitante e de compartimento *Susceptible Infected Resistant* (SIR), destacando como esses modelos podem ser conectados uns aos outros e à

séries temporais para uma determinada região. Além disso, o fato desses modelos serem parcimoniosos faz com que eles sejam particularmente adequados para isolar características-chaves da pandemia. Mesmo após uma variedade de suposições com o objetivo de aumentar a compreensão e evitar o excesso de adaptação dos dados disponíveis, que eram limitados ou incompletos, os autores concluem que esses modelos possuem várias fontes de incerteza, incluindo incerteza de parâmetros, variação baseada em dados ou tipo de modelo utilizado, e, o mais importante, a incerteza na gravidade e duração das medidas de distanciamento social, que podem alterar a data de pico por meses ou até mesmo criar múltiplos picos. Essa variabilidade nos resultados destaca os desafios da modelagem e previsão do curso de uma pandemia durante seus estágios iniciais e com apenas dados limitados.

- Ceylan (2020) formulou diferentes modelos *Auto-Regressive Integrated Moving Average* (ARIMA), onde cada tinha uma parametrização única, e os implementou sobre um conjunto de dados obtidos através da coleta na base da Organização Mundial da Saúde. O conjunto de dados reflete os casos de COVID-19 desde 21 de fevereiro de 2020 até 15 de abril de 2020. Assim, o objetivo do autor era estimar a prevalência de COVID-19 nos três países mais afetados da Europa na época: Itália, Espanha e França. Após a aplicação dos modelos com diferentes valores de parâmetro e a seleção deles de acordo com a sua precisão, Ceylan concluiu que os modelos de séries temporais são significativos e desempenham um papel importante na análise de surtos e na previsão de doenças. Os resultados obtidos através destes modelos auxiliam no planejamento e fornecimento de recursos com eficácia, incluindo pessoal, leitos e unidades de terapia intensiva para gerenciar a situação de surto nesses países. Para trabalhos futuros, com a intenção de alcançar uma comparação mais precisa, recomendou que os dados sejam atualizados em tempo real.

- Maleki *et al.* (2020) modelaram o número total de casos confirmados e recuperados de COVID-19 no mundo utilizando o modelo de série temporal *two-piece scale mixture normal autoregressive* (TP-SMN-AR), que é proveniente de uma família de modelagem flexível e que inclui os modelos de distribuição clássicos gaussianos, simétricos e assimétricos, e de cauda leve ou pesada. Inicialmente, os autores ajustaram os modelos propostos aos dados que refletem os casos confirmados e recuperados de COVID-19 pelo mundo. Depois, selecionaram a série temporal que melhor se ajustou para cada um dos conjuntos de dados. Por fim, utilizaram os modelos selecionados para prever o número de casos mundiais de COVID-19, confirmados e recuperados, entre 20 e 30 de abril de 2020. Para medir o desempenho dos modelos foi feita a diferença entre os valores reais e previstos. Com isso concluíram que os modelos de séries temporais autorregressivas são úteis para modelar dados indexados ao longo do tempo, mas que alguns modelos padrões de séries temporais partem da suposição de que o termo de erro ou resíduos são simétricos, o que não é satisfatório para a modelagem de muitas situações do mundo real, por isso utilizaram modelos baseados em *two-piece scale mixture normal* (TP-SMN). Além disso, os resultados indicaram que o método proposto teve bom desempenho na previsão desejada. Portanto, os principais pontos de destaque foram um modelo de série temporal autoregressiva, melhorado com base nas distribuições TP-SMN, e um novo modelo preditivo eficiente aplicado para prever e estimar os casos COVID-19, confirmados e recuperados, no mundo utilizando dados passados e atuais.

Dentre os trabalhos coletados durante a pesquisa, é importante destacar o trabalho

que trata sobre o ressurgimento da COVID-19 em uma cidade brasileira. Como foi colocado na introdução, o ressurgimento pode estar associado às subnotificações.

- Sabino *et al.* (2021) abordaram quatro possíveis explicações, não mutuamente exclusivas entre si, para o ressurgimento da COVID-19 em Manaus. Primeira, o ressurgimento poderia ser explicado pela maior mistura de indivíduos infectados e suscetíveis durante dezembro de 2020. Esta explicação parte da hipótese de que a taxa de infecções por SARS-CoV-2 poderia ter sido superestimada durante a primeira onda, desse modo a população permaneceu abaixo do limiar de imunidade de rebanho até o mês que antecedeu os picos de internações. Segunda, a imunidade contra a infecção pode já ter começado a diminuir em dezembro de 2020, devido a uma diminuição geral da proteção imunológica contra SARS-CoV-2 após uma primeira exposição, o que é embasado em um estudo realizado com profissionais de saúde do Reino Unido. Este estudo mostra que a reinfecção com SARS-CoV-2 é incomum até 6 meses após a infecção primária. Como a maioria das infecções pelo SARS-CoV-2 em Manaus ocorreu de 7 a 8 meses antes do ressurgimento em janeiro de 2021, levanta-se a possibilidade de reinfecções. Entretanto, é improvável que a diminuição da imunidade por si só explique completamente o ressurgimento. Terceira, diferentes linhagens SARS-CoV-2, isto é, uma infecção produz anticorpos apenas para aquela linhagem específica. Entretanto, foram detectadas três linhagens de SARS-CoV-2 extraordinariamente divergentes (B.1.1.7, B.1.351 e P.1), onde cada uma possui uma constelação única de mutações. Duas destas linhagens (B.1.1.7 e P.1) estavam circulando no Brasil, sendo que a P.1 foi detectada em Manaus em 12 de janeiro de 2021, ou seja, próximo à época onde ocorreu o crescimento abrupto de internações por COVID-19. Quarta, as linhagens SARS-CoV-2 que circulam na segunda onda podem ter maior transmissibilidade em relação às que estavam circulando anteriormente. Em estudo preliminar, a linhagem P.1 atingiu alta frequência entre amostras obtidas a partir de casos de COVID-19 em dezembro de 2020, mas esteve ausente em 26 coletadas entre março e novembro de 2020. Dessa forma, concluem que as novas linhagens de SARS-CoV-2 podem impulsionar um ressurgimento de casos nos locais onde circulam se tiverem maior transmissibilidade em comparação com as que circularam anteriormente. Por outro lado, se o ressurgimento em Manaus se deu em maior parte pela diminuição da imunidade, então cenários semelhantes devem ser esperados em outras localidades.

Os autores que foram destacados aqui propõem formas de analisar e prever o comportamento da pandemia com a intenção de auxiliar nas tomadas de decisões. Este trabalho almeja contribuir com esses estudos, porém, através de outra abordagem, isto é, ao invés de forcar na previsão de casos de COVID-19, a ideia é estimar quanto que a pandemia da COVID-19 impactou no excedente de mortes.

# 4 Metodologia

O trabalho aqui proposto foi realizado em diversas etapas, as quais serão melhor descritas a seguir. De modo geral, pode-se compreender essa pesquisa em cinco estágios: seleção; pré-processamento; transformação; processamento; e, visualização. A seleção tem por objetivo coletar o conjunto de dados que fará parte do estudo. Já o pré-processamento visa preparar o conjunto de dados através de limpeza, correção ou remoção para que eles

estejam de acordo com o objetivo do estudo. Depois, na transformação, busca-se aplicar técnicas para sintetizar os dados, por exemplo, normalização e agregação. Em seguida, no estágio de processamento, ocorre a aplicação de técnicas para a verificação de uma hipótese ou para a descoberta de possíveis padrões. Por fim, utiliza-se do estágio de visualização para auxiliar na interpretação e avaliação das técnicas aplicadas com os dados coletados.

## 4.1 Seleção

Para esta pesquisa, os dados foram selecionados a partir da base de dados aberta do Sistema Unificado de Saúde (SUS), que é conhecida como openDataSus. Mais especificamente, dentre as bases disponíveis, utilizou-se a base de Sistema de Informações de Mortalidade (SIM). Além da base de dados, foi necessário utilizar o TABWIN, um software de análise exploratória, utilizado para descompactar e tabular os arquivos obtidos, pois, os dados disponibilizados no openDataSus não estão no formato desejado para a análise.

A partir do conjunto de dados obtido através do openDataSus, selecionou-se o subconjunto de interesse, ou seja, as datas dos óbitos desde o ano de 2014 até o ano de 2020. Este subconjunto é basicamente uma série que contém 9.361.816 registros das datas supracitadas.

## 4.2 Pré-processamento

Na série obtida, as datas não estão padronizadas, ou seja, há diferença na quantia de algarismos destinados para o dia. Isso pode ser um empecilho na hora do processamento e posteriormente na visualização dos dados. Por isso, aplicou-se sobre o subconjunto de dados uma rotina que faz a verificação de cada uma das datas e as padroniza.

Neste caso, a padronização das datas consiste em destinar quatro algarismos para o ano, dois para o mês, dois para o dia, adicionar um separador entre eles e formatar nessa respectiva ordem.

Fora a padronização dos dados, também foi realizada a verificação de dados nulos, procurando identificar discrepâncias entre os registros disponibilizados pela base e os obtidos após o pré-processamento. Como essa verificação não retornou nenhum caso discrepante, o subconjunto de dados foi considerado pronto para a transformação.

## 4.3 Transformação

Tendo em vista que um dos objetivos dessa pesquisa é fazer a aplicação do modelo preditor de séries temporais, é necessário, em primeiro lugar, obter a série temporal, justamente o que é feito nessa etapa. Isto é, partindo dos dados pré-processados, foi realizada a contagem de ocorrências de cada data no subconjunto em paralelo ao agrupamento delas, o que retorna a quantia de mortes por dia.

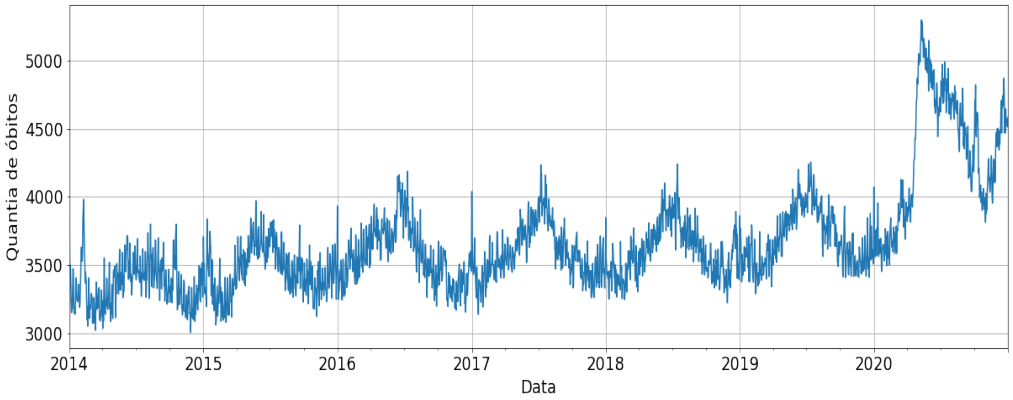
Este agrupamento por data, onde cada data traz consigo um valor, representa para nós uma série temporal discreta, pois, temos a observação de uma variável feita sequencialmente no tempo em faixas igualmente espaçadas.

É preciso reforçar que a sequência das observações importa. Isto pode ser alcançado transformando os dados da coluna de datas para o tipo 'datatime', definindo a frequência como diária e colocando a coluna de datas como index do DataFrame.

## 4.4 Processamento

O primeiro passo do processamento é fazer a análise da série temporal. Para começar essa investigação, faz-se a carga dos dados e a construção do seu gráfico.

Figura 1 – Série temporal que representa a quantia de mortes no Brasil por dia desde 2014 até 2020.




Fonte: Elaborado pelo autor (2022).

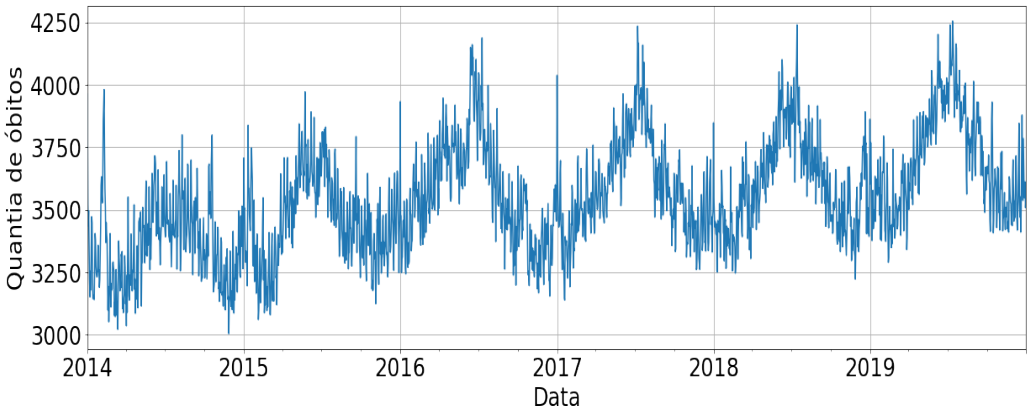
Na figura 1, observa-se a série temporal que descreve a quantia de mortes diárias no Brasil desde o início de 2014 até o final de 2020. É possível notar que há um crescimento no número de mortes com o passar dos anos, mesmo sendo suave, este representa uma tendência. Além disso, também nota-se um comportamento sazonal entre os anos, isto é, um padrão que se repete. Por fim, o ano de 2020 se destaca por seu comportamento anômalo, que ocorreu devido à pandemia da COVID-19.

Em seguida, é preciso definir a parte da série que será usada para treinar o modelo e a parte que servirá como comparativo para verificar o desempenho do mesmo, tais partes são conhecidas respectivamente como conjunto de treino e conjunto de teste. Lembrando que a separação de conjunto treino e conjunto de teste para séries temporais não se dá da mesma forma como em conjuntos que são destinados a outros algoritmos de aprendizado de máquina, pois, a sequência cronológica das observações é relevante. Para muitos métodos de aprendizado de máquina, embaralhamos os dados antes de dividi-los para tornar os dois conjuntos igualmente representativos, mas os dados de séries temporais precisam manter a ordem cronológica dos valores dentro do conjunto. Logo, o embaralhamento não é aplicável neste contexto. Assim, os conjuntos de teste e de treino também teriam que ser sequências ininterruptas de valores. Portanto, o conjunto de treinamento deve incluir todos os valores desde o início dos dados até um ponto específico no tempo, enquanto o conjunto de teste deve incluir o restante.

Tendo em mente que o objetivo dessa pesquisa é utilizar o modelo de previsão de séries temporais para entender o excesso de mortes no ano de 2020, o ponto de separação, para obter ambos conjuntos, é claro. Como serão utilizados seis anos para prever o sétimo, o conjunto de treino representa 86% do conjunto original enquanto o conjunto de teste representa os 14% restantes.

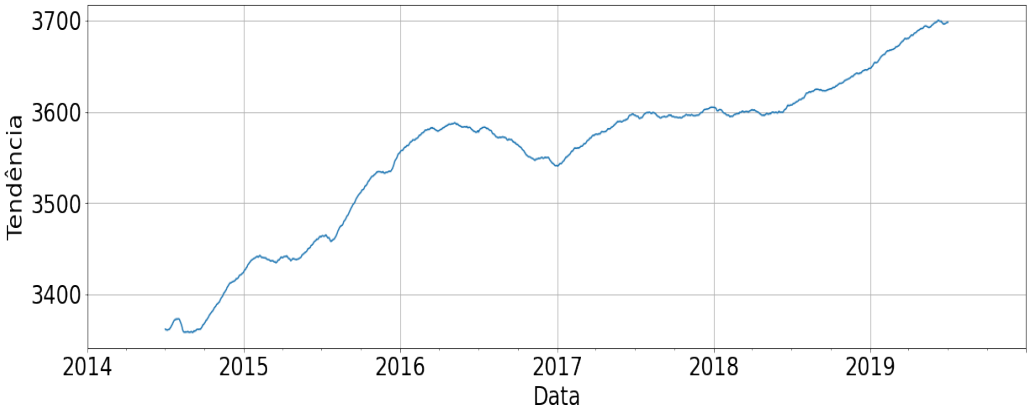
Depois de definir a série a partir da qual será feita a previsão, isto é, a série do conjunto de treino, é necessário fazer uma análise sobre ela. Para se ter uma boa descrição do seu comportamento, é possível construir seu gráfico em paralelo aos gráficos de suas componentes.

Figura 2 – Série de treino.



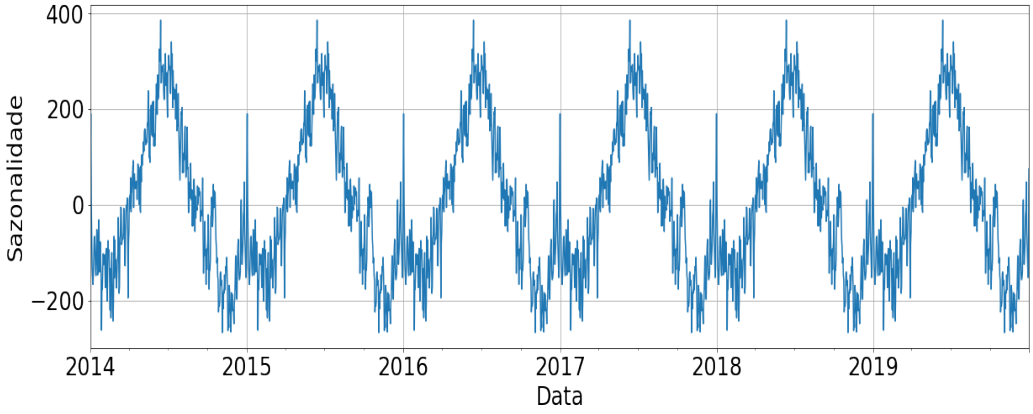

Fonte: Elaborado pelo autor (2022).

Figura 3 – Tendência da série de treino.



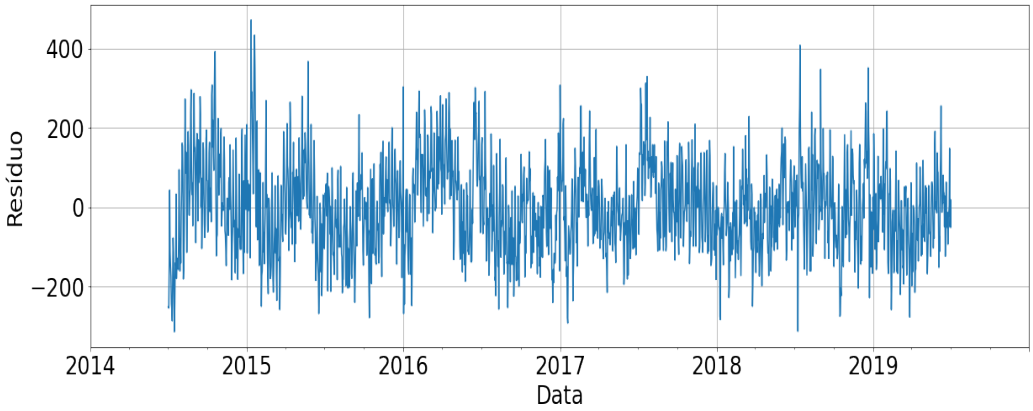

Fonte: Elaborado pelo autor (2022).

Figura 4 – Sazonalidade da série de treino.




Fonte: Elaborado pelo autor (2022).

Figura 5 – Resíduo da série de treino.




Fonte: Elaborado pelo autor (2022).

Pode-se observar nas figuras acima as componentes da série temporal, que foram descritas na fundamentação teórica. Este é o tipo mais simples de decomposição e é conhecido como decomposição ingênua, nela espera-se observar uma relação linear entre às três partes e a série observada. Essa relação linear pode ser feita de duas formas, aditiva, onde a série observada é o resultado da soma das suas três componentes, ou multiplicativa, onde a série observada é igual ao produto das suas três componentes. Aqui, optou-se por utilizar a decomposição aditiva.

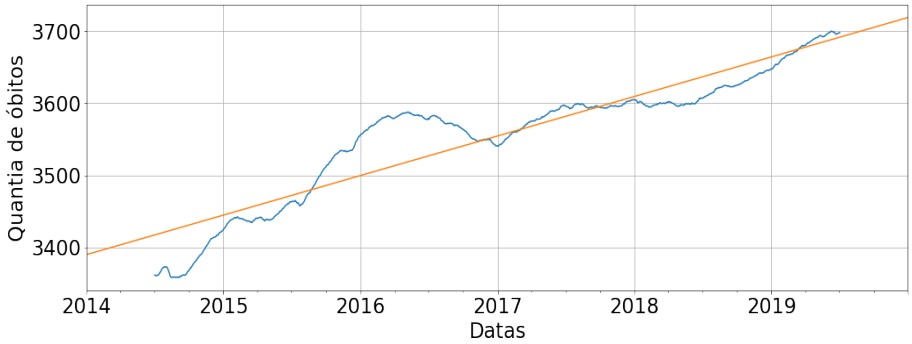
A figura 2 representa o sinal obtido através da série de treino. A tendência, que está na figura 3, representa o padrão consistente ao longo dos dados. Já a sazonalidade, presente na figura 4, expressa todos os efeitos cíclicos ao longo da série. Os resíduos, que aparecem na imagem 5, são os erros associados ao modelo, isto é, a diferença entre os

dados reais e o modelo ajustado.

Agora que se têm uma noção das componentes da série, é preciso encontrar funções que se aproximem da tendência e da sazonalidade. Essa estimativa deve ser feita para conseguir reproduzir estes comportamentos na previsão dos dados futuros. Em outras palavras, foram observados dois comportamentos ao longo do tempo no conjunto de treino e é esperado que tais comportamentos persistam em futuras observações, portanto, é necessário ter as funções que descrevem esses padrões com o objetivo de reproduzi-los.

Começando com a tendência, observa-se que dá para fazer uma regressão linear e aproxima-la com uma função de primeiro grau. Apesar de ser uma aproximação bem simples, essa função consegue atender os propósitos dessa pesquisa. Observe a figura 6, nela estão dispostas duas curvas, a azul que representa a tendência obtida pelo statsmodels do Python e a laranja que representa a função de aproximação. Vale destacar que em mais de duas mil observações, a quantia de óbitos aumentou aproximadamente em trezentos em um conjunto onde os dados descrevem valores na unidade de milhar, ou seja, essa tendência é suave ao longo do tempo.

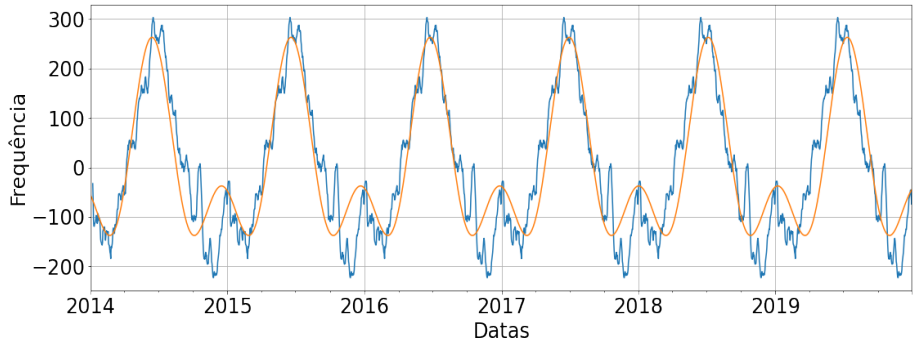
Figura 6 – Estimando a tendência do conjunto de treino.




Fonte: Elaborado pelo autor (2022).

O próximo comportamento a ser estimado é a sazonalidade. Neste caso, a função que melhor se aproximou foi uma composta por um somatório de dois cossenos. Para ajustá-los sobre a sazonalidade foi necessário adiantar seus sinais e colocar a frequência de um cosseno como a metade da frequência do outro. Já a amplitude foi regulada de acordo com a sazonalidade obtida através do Python statsmodels. Veja o estimador da sazonalidade na figura 7. Novamente, a curva em azul é o sinal de sazonalidade do conjunto de treino e a curva laranja é a função de aproximação.

Figura 7 – Estimando a sazonalidade do conjunto de treino.



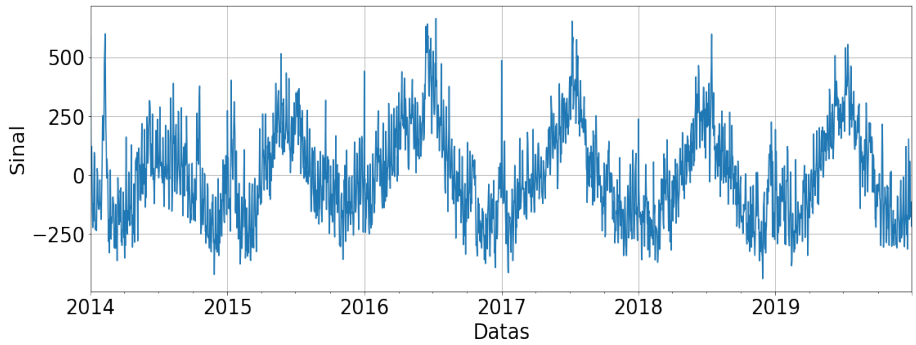

Fonte: Elaborado pelo autor (2022).

Como foi destacado, a série temporal do conjunto de treino tem padrões ao longo do tempo que causam variações em métricas estatísticas como, por exemplo, a média. Como foi discutido na seção 2, uma série temporal que tem uma média que varia com o tempo é uma série não estacionária. Além disso, os modelos que estão sendo utilizados como referenciais teóricos são destinados às séries temporais estacionárias. Portanto, é preciso tornar a série do conjunto de treino estacionária.

Para transformar a série de treino em estacionária, é preciso remover as componentes que causam as variações indesejadas. A primeira componente a ser removida será a tendência. Existe mais de uma técnica para fazer a remoção da tendência, por exemplo, diferenciação ou subtrair a tendência estimada do conjunto original. A técnica escolhida aqui foi o método detrend do SciPy signal, pois, foi a que conseguiu deixar a média do sinal mais próximo de zero.

Após a remoção da tendência, obtém-se o sinal apresentado na figura 8.

Figura 8 – Série temporal de treino sem a tendência.

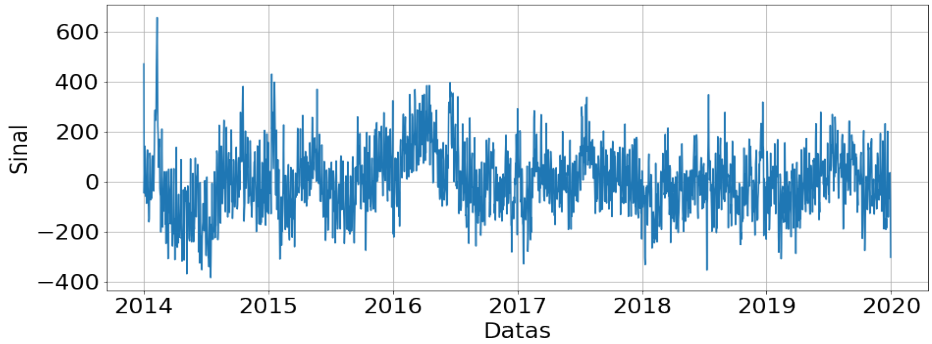



Fonte: Elaborado pelo autor (2022).

A próxima componente que precisa ser removida é a sazonalidade. Para esse caso, optou-se por remover a sazonalidade através da decomposição sazonal. Esse processo nada mais é do que pegar a série sem tendência, obtida anteriormente, e subtrair dela

a sazonalidade retornada pela decomposição da série de treino feita com o statsmodels. Assim, chega-se ao sinal apresentado na figura 9.

Figura 9 – Série temporal de treino sem a tendência e sem sazonalidade.




Fonte: Elaborado pelo autor (2022).

Agora, para se certificar de que a série que será modelada é uma série estacionária, pode-se aplicar um teste estatístico como o teste aumentado de Dick-Fuller. Este é um teste estatístico que tem por objetivo identificar raiz unitária em séries temporais.

O teste aumentado de Dick-Fuller considera duas hipóteses, a hipótese nula e hipótese alternativa. A hipótese nula assume que a série não é estacionária, enquanto a hipótese alternativa considera a série estacionária. Para saber em qual das hipóteses a série se encaixa, é preciso interpretar o valor $p$ retornado pelo teste. Se o valor $p$ for menor que 0,05, então a hipótese nula pode ser rejeitada, ou seja, a série é estacionária (MUSHTAQ, 2011; BROWNLEE, 2020)

O teste aumentado de Dick-Fuller pode ser computado em Python utilizando a função adfuller proveniente da biblioteca statstools. Observe o quadro 1 para visualizar as informações retornadas por esse teste.

Quadro 1 - Teste aumentado de Dick-Fuller.

| **Parâmetros** | **Valores** |
|---|---|
| Valor do teste estatistico | -5.30229666059677 |
| Valor P | 5.415989714115785e-06 |
| Defasagem | 26 |
| Observações | 2164 |
| Valores críticos | '1%': -3.4333754500434264<br>'5%': -2.862876536558312<br>'10%': -2.56748150557262 |

Fonte: Elaborado pelo autor (2022).

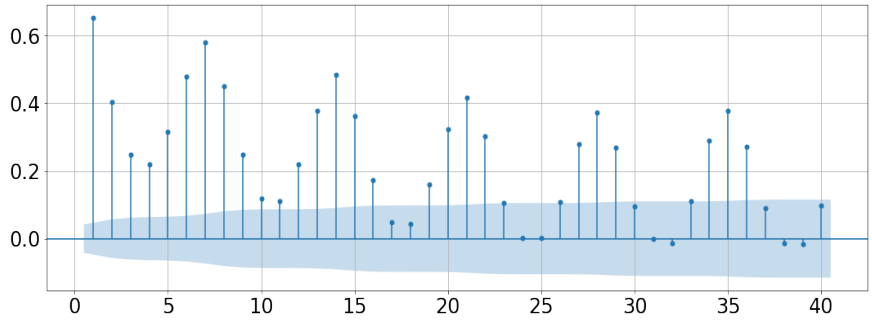
Na tabela acima, encontra-se o valor do teste estatístico, que representa o valor crítico da série submetida, o valor $p$, o qual indica a probabilidade de rejeição ou não da hipótese nula, a defasagem, que indica o número de atrasos usados na regressão para determinar os níveis de confiança, o número de observações utilizadas na análise e valores críticos dos níveis de confiança para 1%, 5% e 10%.

Para este caso vemos que o teste aumentado de Dick-Fuller retornou um valor p menor do que 0,05. Além disso, o valor crítico da série é menor que o valor crítico de 1%, o que fornece um nível de confiança igual a 99%. Então, a hipótese nula é rejeitada e conclui-se que após a remoção das componentes a série tornou-se estacionária.

Seguindo com a análise da série temporal, foi abordado anteriormente que não embaralhamos os dados porque a ordem cronológica do conjunto deve ser preservada. Entretanto, há mais um motivo que indica que os dados não devem ser rearranjados. Para entender o comportamento que a série assume e até mesmo fazer previsões com ela, é preciso encontrar ligações entre as observações passadas e as observações presentes, isto é, entender como e quanto as observações passadas influenciam na observação atual. Neste caso, recorre-se a autocorrelação, que representa a correlação entre uma série e ela mesma defasada.

A função de autocorrelação pode ser observada através do correlograma apresentado na figura 10.

Figura 10 – Autocorrelação da série temporal.




Fonte: Elaborado pelo autor (2022).

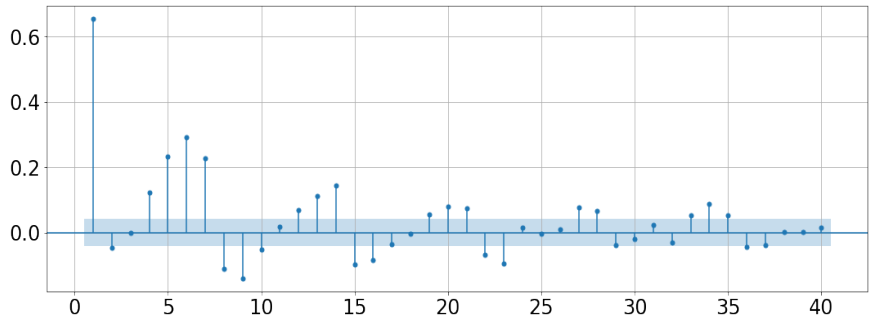
Os valores no eixo x representam defasagens enquanto o eixo y indica os possíveis valores para o coeficiente de autocorrelação. A correlação só pode assumir valores entre um e menos um, razão pela qual a amplitude máxima do gráfico é um. A linha fina ao longo do gráfico representa a autocorrelação entre a série temporal e uma cópia defasada de si mesma, a primeira linha indica a autocorrelação de um período atrás, já a segunda linha representa o valor do coeficiente para dois períodos atrás e assim por diante. A área azul ao redor do eixo $x$ representa a significância, isto é, os valores situados fora são significativamente diferentes de zero o que sugere a existência de autocorrelação para aquela defasagem específica. Essa área se expande à medida que os valores de defasagem aumentam, por isso é mais improvável que essa autocorrelação persista em defasagens elevadas. É preciso garantir que o coeficiente de autocorrelação em defasagens mais altas seja maior para ser significativamente diferente de zero. As linhas, mais altas que a região azul, sugerem que os coeficientes são significativos, por conseguinte, indicam dependência entre os dados ao longo do tempo.

Além da função de autocorrelação, existe a função de autocorrelação parcial, a qual determina apenas a relação direta entre a série temporal e sua versão defasada. Para entender melhor essa ideia, faz-se necessário ter em mente que autocorrelação mede a

semelhança entre uma série temporal e uma versão defasada de si mesma. No entanto, os coeficientes também capturam efeitos secundários, ou seja, examina o valor do coeficiente de autocorrelação para uma determinada defasagem, capturando maneiras diretas e indiretas pelas quais a série defasada afeta a original. Por indireto entende-se todos os outros canais através dos quais os dados passados afetam os dados atuais. Por exemplo, ao se examinar a correlação do período atual, $X_t$, com o valor de dois períodos atrás, $X_{t-2}$, observa-se que há duas formas de $X_{t-2}$ influenciar em $X_t$, a primeira seria direta, isto é, $X_{t-2}$ infere uma tendência para $X_t$, e a segunda seria indireta através de $X_{t-1}$, ou seja, $X_{t-2}$ influência em $X_{t-1}$ que por sua vez influência em $X_t$.

A função de autocorrelação parcial está no correlograma da figura 11.

Figura 11 – Autocorrelação parcial da série temporal.




Fonte: Elaborado pelo autor (2022).

O correlograma da função de autocorrelação parcial ficou bem diferente daquele obtido pela função de autocorrelação, isso ocorre justamente porque a função de autocorrelação parcial considera somente a correlação direta entre a série e a sua versão defasada, enquanto que a função de autocorrelação considera todas as influências sendo elas diretas ou indiretas. Neste caso, observa-se que a função de autocorrelação parcial indica uma correlação bem significativa para a defasagem de um período, além disso, vê-se que assume o mesmo valor apresentado pela função de autocorrelação visto na imagem 10, o que é esperado. Ambas funções devem retornar o mesmo valor de correlação para a defasagem de um período, pois, não existe nenhum outro canal que $X_{t-1}$ influenciaria em $X_t$. O correlograma da função de autocorrelação parcial apresenta correlações negativas para determinadas defasagens, isso significa que valores mais altos nestes períodos resultam em valores mais baixos no período atual e vice-versa. Outro ponto importante a ser destacado é que enquanto a função de autocorrelação indicou que a correlação até quarenta defasagens ainda eram significativamente diferentes de zero, a função de autocorrelação parcial mostrou que a partir da sexta defasagem a correlação começa a cair e quanto mais defasagens são consideradas mais insignificantes se tornam, ou seja, não são significativamente diferentes de zero, portanto, os valores numéricos associados a eles não são importantes, pois, podemos supor que todos são essencialmente zero, sendo positivo ou negativo, e sem efeitos duradouros.

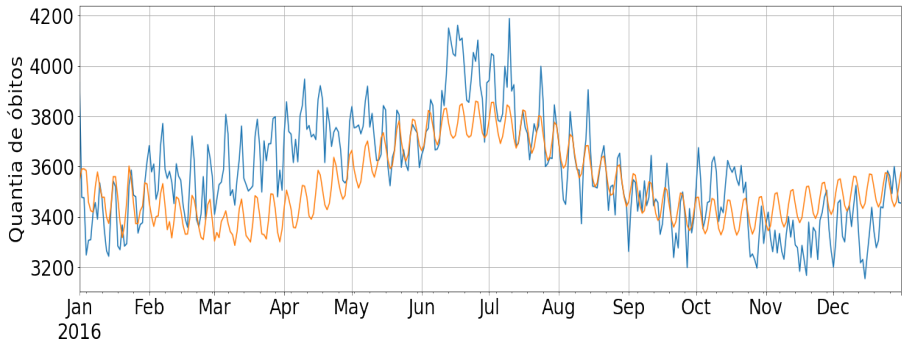
A análise realizada até aqui trouxe o conhecimento básico necessário sobre a série, com isso é possível começar a trabalhar com o modelo preditor. Lembrando que o modelo aqui proposto se baseia nos conceitos dos modelos abordados na seção 2.

Com o propósito de ajustar o modelo preditor, a ideia é realizar previsões em anos conhecidos, isto é, em anos que são abrangidos pelo conjunto de treino. Assim que o modelo estiver ajustado pode-se extrapolá-lo para o ano desconhecido, ou seja, o ano que está no conjunto de teste, e enfim compará-los.

Começando com os conceitos abordados no modelo Autorregressivo (AR), aplica-se uma regressão nos valores passados da série para prever os valores futuros. Inicialmente, o sub-conjunto de treino é dividido novamente de tal forma que os anos de 2014 e 2015 constituam as observações passadas e o ano de 2016 represente as observações futuras.

Na figura abaixo está disposta a comparação entre os óbitos e as previsões para o ano de 2016.

Figura 12 – Comparação entre os óbitos em 2016 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

Na figura 12, observam-se as mortes reais, representadas pela curva azul, comparadas com as previsões, representadas pela curva laranja, para o ano de 2016. Para essa autorregressão foi necessário utilizar valores altos de defasagem, pois, valores menores não conseguiam identificar a variabilidade semanal, mensal ou até mesmo anual. Para que o modelo começasse a entender a variabilidade semanal, foi necessário utilizar uma defasagem que representaria alguns meses no passado. É justamente a variabilidade semanal que traz essa característica senoidal para o modelo. Esse comportamento pode ser visto na função de autocorrelação parcial, figura 11.

As métricas utilizadas para avaliar o desempenho dessa previsão estão no quadro 2. Nela são apresentados o *root mean square error* (RMSE), o *mean absolute error* (MAE) e o *R2 score* ($R^2$).

Quadro 2 - Métricas da previsão do ano de 2016.

| **Métricas** | **Valores** |
|---|---|
| RMSE | 162.91509824243488 |
| MAE | 128.67603332334252 |
| $R^2$ | 0.4089393784800016 |

Fonte: Elaborado pelo autor (2022).

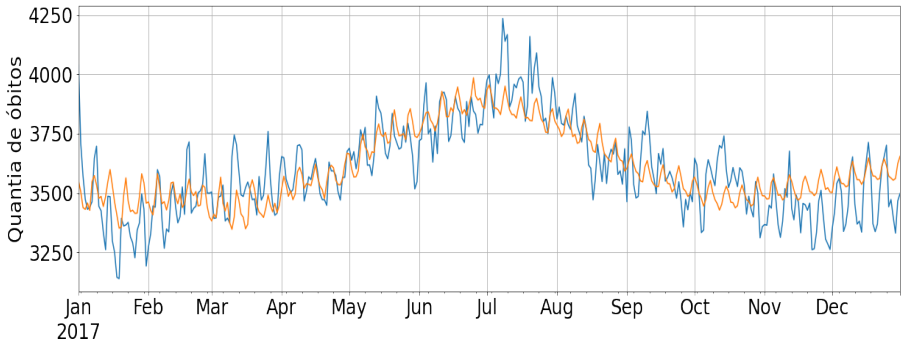
O RMSE é uma maneira útil de ver quão bem um modelo é capaz de ajustar um

conjunto de dados. Quanto maior o RMSE, maior a diferença entre os valores previstos e observados, enquanto que quanto menor o RMSE, melhor o modelo é capaz de ajustar os dados. O MAE representa a diferença entre os valores originais e previstos extraídos pela média da diferença absoluta sobre o conjunto de dados. E, o $R^2$, indica o poder de previsão do modelo, caso assuma um valor igual a um quer dizer que a previsão foi perfeita, caso assuma um valor entre um e zero quer dizer que existe poder de previsão, caso assuma um valor igual a zero quer dizer que está fazendo a previsão da média dos valores e, caso assuma um valor menor do que zero quer dizer que é um mau preditor.

Este mesmo procedimento foi realizado para os anos subsequentes, isto é, utilizaram-se os anos de 2014, 2015 e 2016 para prever o ano de 2017. Depois foram utilizados dos dados de 2014 até 2017 para prever 2018 e assim por diante.

Ao utilizar os dados de 2014, 2015 e 2016 para prever o ano de 2017, obteve-se a previsão que está na figura 13. Nela está disposta a curva real em azul e a curva das previsões em laranja.

Figura 13 – Comparação entre os óbitos em 2017 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

Comparando os dados reais com as previsões realizadas, calculam-se as métricas apresentadas no quadro 3. Dispondo de mais dados, foi possível utilizar mais defasagens e isso resultou em uma queda no erro quando se compara a previsão de 2016 com a previsão de 2017.

Quadro 3 - Métricas da previsão do ano de 2017.

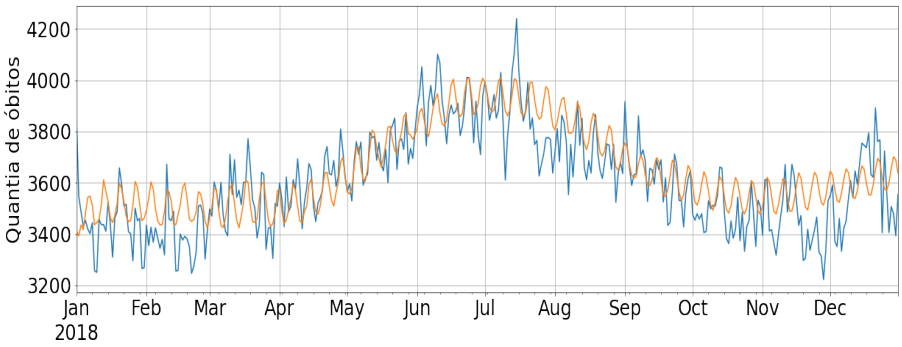
| **Métricas** | **Valores** |
|---|---|
| RMSE | 116.48898154803129 |
| MAE | 91.81878357975157 |
| $R^2$ | 0.6426137618365728 |

Fonte: Elaborado pelo autor (2022).

Para prever o ano de 2018, utilizaram-se os dados de 2014 até 2017. Observe a figura 14 para ver a comparação entre os dados reais, em azul, e a previsão, em laranja. Novamente houve uma queda no erro da previsão comparada com a previsão passada,

mas é uma queda bem menos significativa do que a que ocorreu entre 2016 e 2017. Para visualizar as métricas de previsão do ano de 2018, veja o quadro 4.

Figura 14 – Comparação entre os óbitos em 2018 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

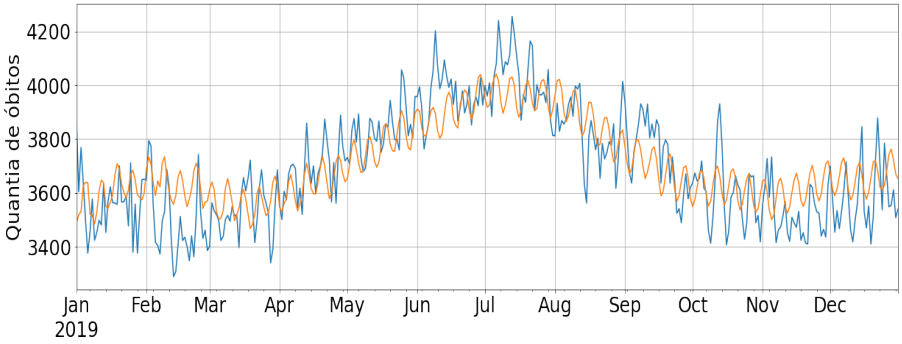
Quadro 4 - Métricas da previsão do ano de 2018.

| **Métricas** | **Valores** |
|---|---|
| RMSE | 116.44898845662787 |
| MAE | 91.70040504346765 |
| $R^2$ | 0.6261700614882884 |

Fonte: Elaborado pelo autor (2022).

O último ano que se faz a previsão para ajustar o modelo, é o ano de 2019. Visando fazer essa previsão, inclui-se o ano de 2018 ao conjunto de dados passados e se considerou os dados de 2019 como os dados futuros. Como realizado anteriormente, constrói-se o gráfico real, representado pela curva em azul, versus o gráfico de previsões, representado pela curva em laranja. Observe figura 15.

Figura 15 – Comparação entre os óbitos em 2019 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

A comparação entre os dados reais e as previsões para o ano de 2019 resultou nas métricas que estão dispostas no quadro 5.

Quadro 5 - Métricas da previsão do ano de 2019.

| **Métricas** | **Valores** |
|---|---|
| RMSE | 122.89426697577866 |
| MAE | 98.10759986125352 |
| $R^2$ | 0.6370711552740491 |

Fonte: Elaborado pelo autor (2022).

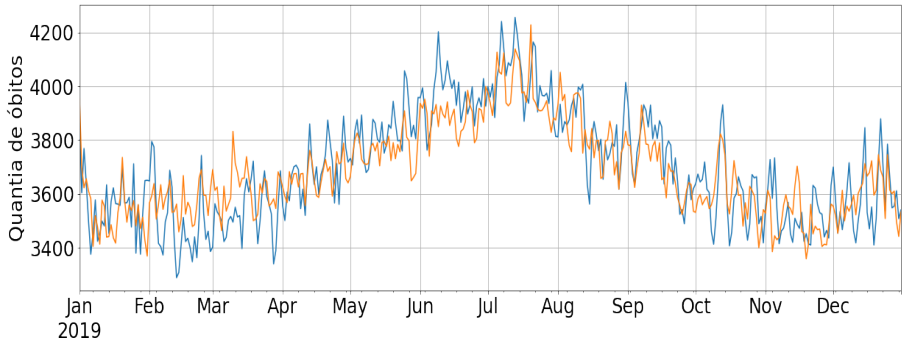
Com exceção do ano de 2019, é possível ver que ao decorrer das previsões os erros reduziram. A cada previsão feita, o grupo de valores passados aumentou o que possibilitou trabalhar com mais defasagens. O erro voltou a subir na previsão do ano de 2019 porque o estimador de sazonalidade não conseguiu acompanhar bem o comportamento dos três primeiros meses desse ano. Veja as últimas observações na figura 7 para visualizar como ocorre esse desencontro.

Continuando com a modelagem, pode-se recapitular os conceitos do modelo de médias móveis (MA). Este modelo, assim como o modelo autorregressivo, busca fazer uma combinação através da regressão, a diferença é que o modelo autorregressivo busca fazer uma autorregressão nos valores da série temporal enquanto que o modelo de médias móveis visa uma autorregressão sobre os erros cometidos em previsões passadas. Se for possível prever o erro esperado, então o modelo pode ser aperfeiçoado, onde a sua saída será o somatório dos valores e dos erros previstos.

Ao aplicar a mesma autorregressão que foi feita para os valores da série sobre os erros cometidos em previsões passadas, não se obteve um preditor de erros que melhorasse o desempenho do modelo. Entretanto, em uma análise exploratória dos erros das previsões passadas, observou-se que era possível melhorar o desempenho do modelo se o erro esperado fosse considerado como sendo a média dos dois últimos erros para cada data, por exemplo, o erro esperado para 01/01/2019 é igual à média entre o erro obtido em 01/01/2018 e o erro obtido em 01/01/2017.

Após realizar esse ajuste no modelo, o erro na previsão reduziu, por conseguinte, o seu desempenho aumentou. Observe a figura 16 para visualizar a previsão do ano de 2019 aplicando o cálculo do erro esperado. As novas métricas obtidas estão no quadro 6.

Figura 16 – Comparação entre os óbitos em 2019 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

Quadro 6 - Métricas da previsão do ano de 2019.

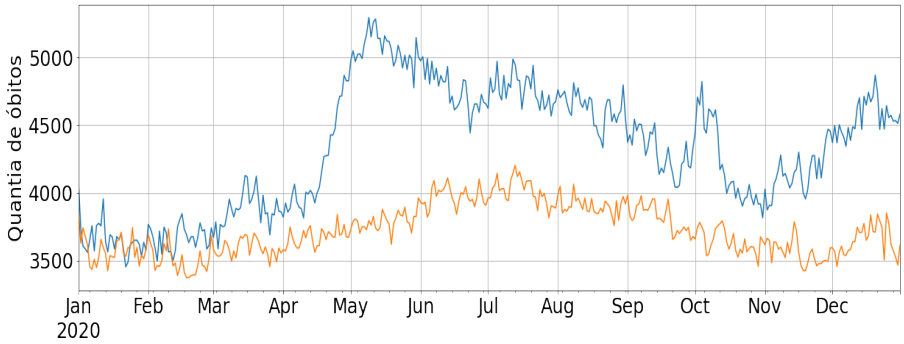
| **Métricas** | **Valores** |
|---|---|
| RMSE | 112.31513723343348 |
| MAE | 91.67120837763763 |
| $R^2$ | 0.6968658791149082 |

Fonte: Elaborado pelo autor (2022).

Enfim, os últimos conceitos a serem abordados são aqueles referentes ao modelo autorregressivo integrado à média móvel, o qual é um modelo voltado para séries não estacionárias que se utiliza da diferenciação para tornar a série estacionária antes de aplicar a autorregressão nos valores da série. Ao tornar a série estacionária por diferenciação, os erros de previsão aumentaram e o desempenho do modelo reduziu, isso indica que de fato a melhor estratégia foi remover as componentes da série temporal através do SciPy signal e da decomposição sazonal.

Agora que o modelo foi treinado e devidamente ajustado, é possível fazer a previsão para o ano de 2020. A previsão pode ser visualizada na figura 17 e as suas métricas podem ser visualizadas no quadro 7.

Figura 17 – Comparação entre os óbitos em 2020 (em azul) e a previsão de óbitos para este ano (em laranja).




Fonte: Elaborado pelo autor (2022).

Quadro 7 - Métricas da previsão do ano de 2020.

| **Métricas** | **Valores** |
|---|---|
| RMSE | 694.7765899803861 |
| MAE | 601.1173434369584 |
| $R^2$ | -1.2631215892946308 |

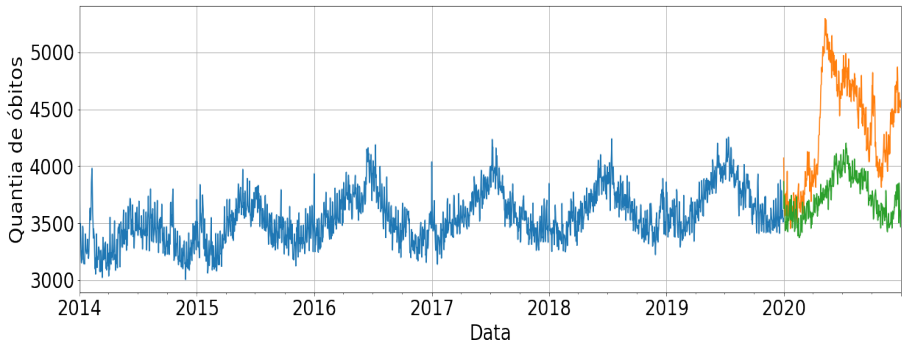
Fonte: Elaborado pelo autor (2022).

De acordo com as métricas, conclui-se que este modelo não é um bom preditor para o ano de 2020, o que já era esperado. Para ser considerado um modelo com poder preditivo, seria preciso prever o efeito da pandemia utilizando apenas os dados dos anos passados e isso é impossível. O que interessa nessa pesquisa é o erro contido nessa previsão. Em outras palavras, foi feita a previsão do que era esperado caso a pandemia da COVID-19 não tivesse acontecido. Dessa forma, é possível estimar o impacto causado pela pandemia.

## 4.5 Visualização

Durante a descrição das etapas anteriores realizadas nessa pesquisa, algumas figuras foram colocadas para acrescentar informação visual e auxiliar no entendimento. Esta seção, em específico, focará na visualização do processo como um todo, isto é, construir um gráfico com todos os dados que sumarizam o processo da pesquisa. O objetivo aqui é mostrar até que ponto foi possível evoluir partindo dos dados coletados. Observe a figura 18.

Figura 18 – Comparativo entre o conjunto de treino, conjunto de teste e a previsão realizada.




Fonte: Elaborado pelo autor (2022).

Na figura acima, observando separadamente, tem-se a série do conjunto de treino em azul, a série do conjunto de teste em laranja e a série prevista pelo modelo em verde. A curva azul somada à curva laranja representa a série do conjunto original, isto é, antes da divisão em conjuntos de treino e de teste. Vale ressaltar também que a curva verde, ou seja, a previsão, conseguiu acompanhar bem o comportamento da curva azul, a qual descreve os anos anteriores.

# 5 Resultados Experimentais

A análise da série temporal levantou informações necessárias para o ajuste do modelo, que, por sua vez, conseguiu fazer a previsão da quantia de óbitos por dia do ano desejado. O resultado a ser analisado é a diferença entre os dados reais e os previstos. Este erro representa o erro associado ao modelo mais as mortes causadas pela covid-19. Levando em conta que na equação de auto regressão existe uma componente de aleatoriedade, é difícil estimar com certeza o erro do modelo, então para simplificar a análise considera-se que a diferença entre os valores reais e a previsão foi causada apenas pelas mortes por covid-19. Dessa forma, considerando a data do primeiro registro de óbito por covid no Brasil, isto é, 17/03/2020 e extraindo a média do erro, conclui-se que a covid-19 contribuiu, em média, para 713 óbitos diários a mais do que era esperado.

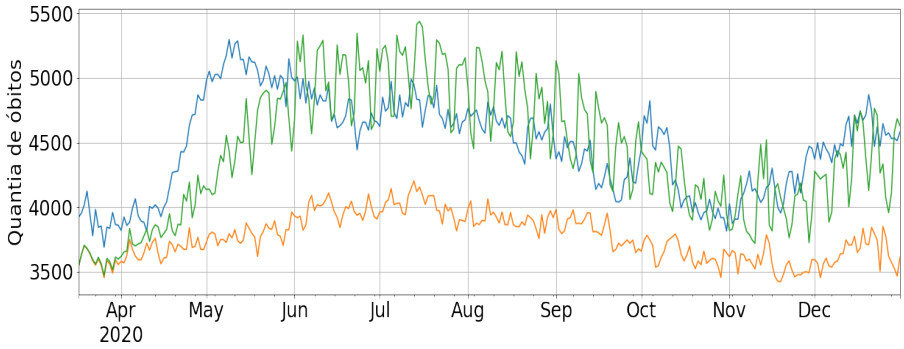
Além de conseguir observar o impacto da covid-19 nos óbitos nacionais através da diferença entre o valor real e a previsão, também é possível ter uma noção da subnotificação de mortes por covid-19. Considerando que a diferença entre as duas curvas representa os óbitos causados pela covid-19, basta somar esses óbitos a curva de previsões para chegar na curva real.

O registro de mortes diárias por COVID-19 foi obtido através Repositório de dados do Centro de Ciência e Engenharia de Sistemas (CSSE) da Universidade Johns Hopkins.

Fazendo esse somatório obtém-se uma nova curva. Observe a figura 19. A curva em verde representa a previsão de óbitos para o ano de 2020 mais os óbitos causados pela covid-19. Com essa nova curva é possível visualizar em alguns momentos que a notificação de mortes por covid-19 não explica diretamente a quantia de óbitos reais, em outros momentos a quantia de óbitos esperados mais os óbitos ela covid-19 ultrapassam a curva

real de óbitos. Então para entender se no quadro geral as mortes por covid-19 no Brasil estão sendo subnotificadas ou super notificadas, é preciso calcular a média do erro em relação a nova curva obtida. Dessa forma se a média for positiva, quer dizer que estão ocorrendo mais óbitos por covid-19 do que o que está sendo notificado, ou seja, existe subnotificação dessas mortes. Em contrapartida, se a média for negativa, quer dizer que há super notificação das mortes por covid-19.

Figura 19 – Somando os óbitos causados pela COVID-19 a previsão obtida anteriormente.




Fonte: Elaborado pelo autor (2022).

Calculando a média dos erros, obtém-se o valor arredondado de 41. Como a média retornada é positiva, estima-se que no ano de 2020 houveram mais óbitos por covid-19 do que se tem registro.

# 6 Conclusão e Trabalhos Futuros

O novo coronavírus surgiu de uma fonte não humana, mas conseguiu sustentar a sua transmissão de um ser humano para outro. Esse vírus, que é responsável pela COVID-19, se alastrou rapidamente no país de origem e, por conta de viagens, em outros países. Tal dinâmica de contágio fez com que a Organização Mundial da Saúde (OMS) declarasse essa doença como uma pandemia.

A pandemia da COVID-19 afetou a qualidade e a expectativa de vida das pessoas mundialmente. Os impactos mais evidentes ocorreram na área da saúde, onde houve uma sobrecarga no sistema de atendimento, na economia, pois, ocorreu uma desaceleração significativa no crescimento das receitas, e na vida social, que sofreu com o distanciamento e a perda de entes queridos. O cenário nacional não foi diferente, sendo possível verificar como a quantia de mortes por dia no ano de 2020 destoou do comportamento de mortes diárias nos anos predecessores.

Entender o comportamento e a tendência de crescimento da pandemia se tornou vital para tomar decisões em prol de freia-la. A análise de séries temporais, bem como o estudo sobre seus modelos preditores, se demonstraram significativos ao modelar o cenário da pandemia no Brasil. Nesta pesquisa, demonstrou-se que a quantia de mortes por dia pode ser descrita na forma de uma série temporal. Logo pode ser melhor compreendida através de duas tarefas, análise e previsão de séries temorais. A análise traz características de comportamento, enquanto a extrapolação dos dados pela previsão delimita um cenário a partir do qual decisões podem ser feitas.

Quando a previsão da quantia de mortes por dia é feita, é possível interpretar o erro associado para ter uma noção do excedente de mortes, que pode ser utilizado para estimar a subnotificação de óbitos causados pela COVID-19. O modelo aqui apresentado demonstrou ser um bom preditor para anos onde o comportamento da mortalidade era bem conhecido, entretanto, ao ser comparado com o ano de 2020 seu desempenho foi muito afetado, fazendo com que o mesmo não fosse um bom preditor para esse ano. Esta perda de desempenho era esperada e está alinhada com o contexto do estudo, que é utilizar o erro dessa previsão para estimar o impacto da COVID-19 sobre os óbitos diários. Com isso, foi possível ter uma noção de quanto a COVID-19 influenciou na quantia de óbitos diários, além de estimar a subnotificação de mortes causadas por esta doença.

Para futuros trabalhos é recomendado melhorar os estimadores de tendência e de sazonalidade. Também é possível abordar modelos que não foram contemplados nesta pesquisa como, por exemplo, o SARIMA, que é uma extensão do modelo ARIMA para englobar a sazonalidade da série temporal. Outra opção é utilizar os mesmos modelos que foram descritos aqui, mas torna-los dinâmicos de tal forma que eles aprendam com os novos dados que surgem diariamente e detectem os efeitos da pandemia. Vale destacar também a possibilidade de pesquisa no campo da Inteligência Artificial (IA), isto é, utilizar redes neurais artificiais (RNA) para ver como estas redes conseguem descrever esses dados e como suas previsões diferenciariam das previsões feitas pelos modelos de séries temporais.

# Referências


PHELAN, Alexandra L.; KATZ, Rebecca; GOSTIN, Lawrence O. The novel coronavirus originating in Wuhan, China: challenges for global health governance. Jama, v. 323, n. 8, p. 709-710, 2020.

CUCINOTTA, Domenico; VANELLI, Maurizio. WHO declares COVID-19 a pandemic. Acta Bio Medica: Atenei Parmensis, v. 91, n. 1, p. 157, 2020.

HALEEM, Abid; JAVAID, Mohd; VAISHYA, Raju. Effects of COVID-19 pandemic in daily life. Current medicine research and practice, v. 10, n. 2, p. 78, 2020.

MORATO, Marcelo M. et al. An optimal predictive control strategy for COVID-19 (SARS-CoV-2) social distancing policies in Brazil. Annual reviews in control, v. 50, p. 417-431, 2020.

SHUMWAY, Robert H.; STOFFER, David S. Time series analysis and its applications. New York: springer, 2000.

CHATFIELD, Chris. The analysis of time series: an introduction. Chapman and hall/CRC, 1996.

LIU, Zeyuan et al. Coronavirus Epidemic (COVID-19) Prediction and Trend Analysis Based on Time Series. In: 2021 IEEE International Conference on Artificial Intelligence and Industrial Design (AIID). IEEE, 2021. p. 35-38.

KUMAR, Raghavendra et al. COVID-19 Outbreak: An Epidemic Analysis using Time Series Prediction Model. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021. p. 1090-1094.

BERTOZZI, Andrea L. et al. The challenges of modeling and forecasting the spread of COVID-19. Proceedings of the National Academy of Sciences, v. 117, n. 29, p. 16732-16738, 2020.

SABINO, Ester C. et al. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. The Lancet, v. 397, n. 10273, p. 452-455, 2021.

CEYLAN, Zeynep. Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment, v. 729, p. 138817, 2020.

MALEKI, Mohsen et al. Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel medicine and infectious disease, v. 37, p. 101742, 2020.

CHATFIELD, Chris. Time-series forecasting. CRC press, 2000.

BROWNLEE, Jason. Introduction to time series forecasting with python: how to prepare data and develop models to predict the future. Machine Learning Mastery, 2020.

BOX, George EP et al. Time series analysis: forecasting and control. John Wiley  Sons, 2015.

MUSHTAQ, Rizwan. Augmented dickey fuller test. 2011.

VISHWAS, B. V.; PATEL, Ashish. Hands-on Time Series Analysis with Python: From Basics to Bleeding Edge Techniques. Apress, 2020.

ORELLANA, Jesem Douglas Yamall et al. Excesso de mortes durante a pandemia de COVID-19: subnotificação e desigualdades regionais no Brasil. Cadernos de Saúde Pública, v. 37, p. e00259120, 2021.

Como usar as bases de dados do DATASUS. Mobilidade Ativa, 202. Disponível em: <https://mobilidadeativa.org.br/como-usar-as-bases-do-datasus/>. Acesso em: 17/01/2022.

Sistema de Informação sobre Mortalidade. openDataSus, 2021. Disponível em: <https://opendatasus.saude.gov.br/dataset?organization=ministerio-da-saude>. Acesso em: 02/09/2021.

SHARMA, Parkash. Time Series Analysis in Python. Medium, 2020. Disponível em: <https://medium.com/@parkashsharma/time-series-analysis-in-python-4f2e7a453ded>. Acesso em: 03/02/2022.

SINGH, Rana. Time Series forecasting using LSTM/ARIMA/Moving Average use case(Single/Multi-variate) with code. Medium, 2021. Disponível em: <https://ranasinghiitkgp.medium.com/time-series-forecasting-using-lstm-arima-moving-average-use-case-single-multi-variate-with-code-5dd41e32d1fc>. Acesso em: 03/02/2022.

SHAH, Jinit. ARIMA Model from Scratch in Python. Medium, 2020. Disponível em: <https://medium.com/analytics-vidhya/arima-model-from-scratch-in-python-489e961603ce>. Acesso em: 21/02/2022.

Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Inf Dis. 20(5):533-534. doi: 10.1016/S1473-3099(20)30120-1. Disponível em: <https://github.com/CSSEGISandData/COVID-19>. Acesso em: 28/02/2022.