UFRJ

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE METODOS ESTATÍSTICOS

CLAYTON WELLINGTON BEZERRA DOS SANTOS
VICTOR YURI NOGUEIRA ALVES

ANÁLISE DA EVOLUÇÃO DE CASOS E MORTES DE COVID 19 PELO MODELO ARIMA

MONOGRAFIA

RIO DE JANEIRO
2022

Trabalho de Conclusão de Curso apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de bacharel em Ciências Atuariais.

Orientador: Marina Silva Paez

Rio de Janeiro

2022

CLAYTON WELLINGTON BEZERRA DOS SANTOS
VICTOR YURI NOGUEIRA ALVES

ANÁLISE DA EVOLUÇÃO DE CASOS E MORTES DE COVID 19 PELO MODELO ARIMA

Trabalho de Conclusão de Curso apresentado ao curso de Ciências Atuariais do Instituto de Matemática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do grau de bacharel em Ciências Atuariais.

Aprovador por:

Profa. Marina Silva Paes
(Orientadora)
Instituto de Matemática - UFRJ

Prof. João Batista de Moreira Pereira
(Professor)
Instituto de Matemática - UFRJ

Profa. Kelly Cristina de Mota Gonçalves
(Professora)
Instituto de Matemática - UFRJ

Rio de Janeiro
2022

CIP - Catalogação na Publicação

S96145a

Santos; Clayton Wellington Bezerra dos, Alves; Victor Yuri Nogueira/
ANÁLISE DA EVOLUÇÃO DE CASOS E MORTES DE COVID 19 PELO MODELO ARIMA / Alves; Victor Yuri Nogueira/ Santos; Clayton Wellington Bezerra dos. -- Rio de Janeiro, 2022.
39 f.

Orientadora: Marina Silva Paez.
Coorientador: Nei Carlos dos Santos Rocha.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Instituto de Matemática, Bacharel em Ciências atuariais, 2022.

1. Séries Temporais. 2. Modelo ARIMA. 3. Covid 19. I. Paez, Marina Silva, orient. II. Rocha, Nei Carlos dos Santos, coorient. III. Titulo.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

AGRADECIMENTOS

Clayton Wellington Bezerra dos Santos

Nada na vida, nós conquistamos sozinhos, sempre há alguém por trás de nós, servindo como um pilar de sustentação para que nós nos mantenhamos erguidos, e comigo isso não é diferente.

Gostaria de agradecer primeiramente aos meus pais, pois garanto que o sacrifício e o esforço de ambos eu sequer teria conseguido pisar no campus da UFRJ. Gostaria de agradecer ao meu irmão por aguentar todas as possíveis crises que eu tive dentro de casa causadas por noite que eu quase não dormi estudando para inúmeras provas.

Gostaria de agradecer a Thais, minha namorada de 10 anos, a qual também se manteve ao meu lado em diversos momentos que eu me via perdido dentro da faculdade, em momentos que eu deixei de sair para estudar, entre outros.

Gostaria também de agradecer a Marina, nossa orientadora neste trabalho que sempre se disponibilizou para nos orientar em dúvidas, inclusive num feriado de carnaval. Meus agradecimentos também aos amigos que a UFRJ me deu, acredito que alguns levarei para a vida inteira.

Por último, gostaria de agradecer todos os professores com quem tive aulas, pois todos eles foram fundamentais na construção do profissional que estou me tornando mesmo que de forma indireta.

Victor Yuri Nogueira Alves

Agradeço a meu pai e minha mãe por sempre estarem presentes e me apoiarem no decorrer da graduação, sem eles com certeza a jornada teria sido muito mais árdua.

Meus agradecimentos aos amigos que fiz na UFRJ que fizeram parte da minha formação e que vão continuar presentes em minha vida com certeza.

A minha orientadora Marina, pelo suporte no pouco tempo que lhe coube, pelas suas correções e incentivos

Mylla, sou muito grato por seu apoio e amor, sem você este TCC não teria terminado. Obrigado por sua gentileza e compreensão mesmo com minha ausência em diferentes momentos.

Resumo

O vírus da COVID-19 começou a se espalhar mundialmente no fim de 2019. Desde então o mundo lida com uma pandemia global sem precedentes na história. Dados sobre o volume de casos e mortes vêm sendo coletados por diversos meios para se acompanhar a evolução da doença. Este trabalho visa analisar os dados de casos e mortes de COVID-19 no Brasil agrupados por semana, mais precisamente entre a 9ª semana de 2020 até a 45ª semana de 2021. Para obter os resultados, foram utilizadas técnicas de análise de regressão e análise de séries temporais, principalmente do modelo ARIMA. A base de dados foi retirada do site ourworldindata.org, um repositório de dados gratuito que retira os dados diretamente do Ministério da Saúde.

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

SUMÁRIO

1) Introdução

A análise e gerenciamento de doenças infecciosas como a COVID-19 tem como principal objetivo entender e prever o avanço de tais doenças observando seu comportamento numa população específica. Por outro lado, a inexistência de um grupo controlado pode levar a enviesamentos e outras dificuldades experimentais (LINDEN A, ADAMS JL, ROBERTS N. AN, 2003), (SATO, R. C., 2013). O uso de séries temporais surge como alternativa para a análise e gerenciamento desse tipo de doença, pois pela dependência serial característica da análise de séries, temporais fica possível compreender e prever melhor possíveis programas de contingência e combate a essas doenças. Além disso, esse tipo de análise fornece maior usabilidade e capacidade de previsão do que os modelos não temporais (SOYIRI IN, REIDPATH DD, 2012), (MONTGOMERY, DOUGLAS C., CHERYL L. JENNINGS, MURAT KULAHCI., 2015).

A difusão de bancos de dados online gratuitos do novo coronavírus e o armazenamento de dados em nuvem sobre o avanço da doença abriram novos horizontes para a extração e tratamento de amostras. O portal *Our World in Data (ourworldindata.org)*, na qual a base de dados sobre a COVID-19 no Brasil foi extraída para esse trabalho, é um exemplo desses grandes portais para extração gratuita de dados. Sendo usado pela *Harvard University, Stanford, University of Cambridge* e outras universidades, como fonte de dados para diversas pesquisas, o site oferece uma base de dados diária sobre o avanço da COVID-19.Como o fator "tempo" se prova crucial para o estudo deste tipo de doença, o modelo de séries temporais que será usado neste trabalho também é o mais utilizado na área de saúde e se chama ARIMA (modelo auto-regressivo integrado a médias móveis) (CHOI K, THACKER SB,1981).

O modelo ARIMA foi criado em meados dos anos 70 por George Box e Gwilym Jenkins adaptando métodos de filtragem de dados em tempo discreto dos anos 30 com o objetivo de aplicá-lo a dados econômicos. Por esse fato ele também é conhecido como "modelo Box-Jenkins". Dentre alguns exemplos do uso do ARIMA no campo de doenças epidemiológicas estão: compreensão e previsão da tendência epidêmica de brucelose na China (YUAN, Z.; ZHONH-GIN, G.; PEI-FENG, L.2019) ; descrição da epidemia de influenza entre crianças de Wuhan, China e; previsão do número de leitos ocupados durante a crise de SARS (severe acute respiratory syndrome) em um hospital de Cingapura (ZHIRUI, H.; HONGBING, T.2018).

O foco principal deste trabalho é, a partir dos conhecimentos teóricos em modelagem estatística, análise e manipulação de grande volume de dados utilizando métodos computacionais adquiridos durante anos de faculdade, descrever com eficiência o comportamento das curvas de casos e mortes de COVID-19 no Brasil assumindo que estas curvas são descritas como séries temporais. Além disso, queremos também predizer o comportamento futuro das séries citadas anteriormente e analisar se nosso modelo de fato é consistente.

Para isso, será apresentada a teoria na qual o estudo foi embasado, as principais premissas que devem ser seguidas para que o resultado seja de fato consistente. Além disso, será apresentada a forma de como os dados da foram analisados, como o modelo escolhido foi implementado e quais resultados foram encontrados.

Além disso, temos como objetivo apresentar um modelo não só consistente, como simples, uma vez que o desafio do estatístico no mundo moderno é conseguir unir precisão com simplicidade. Em geral não é do interesse descrever curvas a partir de algoritmos complexos e difíceis de serem implementados, já que estes dificultam os modelos de previsão. É importante também notar que quanto mais simples é o modelo mais fácil é o entendimento do mesmo e mais difundida é a informação, implicando em maior acessibilidade ao conteúdo apresentado.

Por fim, após a obtenção dos resultados, uma análise entre o real e o estimado será feita a fim de validar o modelo preditivo em que será avaliada a assertividade do próprio. Com isso, será possível analisar se um modelo relativamente simples tanto teoricamente, quanto computacionalmente, foi suficiente para descrever os eventos de casos e mortes de COVID-19.

No capítulo a seguir apresentamos uma revisão metodológica dos conceitos estatísticos utilizados na aplicação dos modelos aos dados deste trabalho.

2) Metodologia / Fundamentação Teórica

2.1) Introdução às séries temporais

Séries temporais são conjuntos de observações quaisquer que são ordenadas em relação ao tempo. Ou seja, em séries temporais, a ordem dos dados é de extrema importância no entendimento, modelagem e previsão da série. Podemos ver o estudo de séries temporais em diversos campos da sociedade, como na área da saúde (por exemplo analisando número de casos de uma doença ao longo do tempo); nas finanças(analisando o valor de uma ação variando no tempo - ver Figura 1); ou no esporte (como o ritmo de passadas por minuto de um maratonista ao longo de uma maratona).

Figura 1: Série temporal das ações da Petrobras por dia (PETR4) no período de maio de 2009 até outubro de 2020 (fonte: Google Finanças).

Para exemplificar o conceito de séries temporais, suponhamos que um atleta esteja interessado em saber o volume de batimentos cardíacos em uma futura prova de velocidade a partir de uma estimação de dados passados. Se por acaso, os dados fossem coletados de forma independente de hora em hora em relação ao dia a dia do atleta, claramente a previsão dos seus batimentos cardíacos para as futuras provas de velocidade seria totalmente diferente dos dados reais. Por outro lado, se a coleta dos dados focasse na média de batimentos durante provas de velocidade semelhantes feitas pelo atleta, certamente a previsão seria muito mais precisa. Isso nos mostra o quão importante são a coleta e ordenação dos dados.

Uma série temporal também pode ser vista como um processo estocástico, uma vez que processos estocásticos podem ser definidos como uma família de variáveis aleatórias X(.,t) indexada no tempo.

Então, X(.,t) é uma variável aleatória para algum determinado t e X(ω,t) é uma realização desse processo observado no tempo t em que ω é um elemento do espaço amostral. Uma população que engloba todas as possíveis realizações é denominada processo estocástico gerador de dados. Portanto, uma série temporal pode ser definida como uma realização ordenada de um processo estocástico qualquer.

Estudar e entender a estrutura de uma série é de grande importância uma vez que podemos descobrir alguma espécie de tendência em eventos considerados aleatórios auxiliando na previsão de valores futuros para auxiliar em tomadas de decisão. Por exemplo, podemos ter como objetivo estudar a curva de uma determinada ação a fim de saber quando é o momento mais propício de vender tudo ou comprar mais da mesma.

Podemos definir as principais propriedades estudadas em uma série temporal como:

- Tendência: Segundo o (MORETTIN, P. A.; M. C. TOLOI, C., 2006), tendência é o nome dado ao comportamento que uma série temporal descreve. Os comportamentos mais comuns de uma série temporal são: tendência constante, tendência linear e tendência quadrática, estes ainda podendo ser positivos (crescem com o tempo) ou negativos (diminuem com o tempo). Como vemos a seguir na Figura 2.

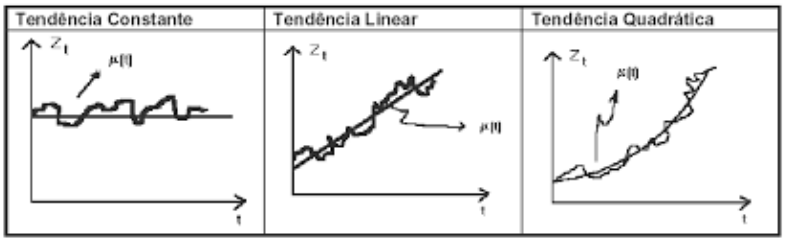

Figura 2: Ilustração dos tipos de tendência de uma série (fonte:https://medium.com/lica-ufal/uma-breve-jornada-em-s%C3%A9ries-temporais-pt-1-introdu%C3%A7%C3%A3o-5d6d581ba803).

- Sazonalidade: Segundo (CRYER J. D., CHAN, K. S., 2008), sazonalidade é um conjunto de padrões que se repetem a cada período, necessariamente idênticos. Por exemplo, a cada mês, ano, trimestre etc. Como vemos a seguir na Figura 3.

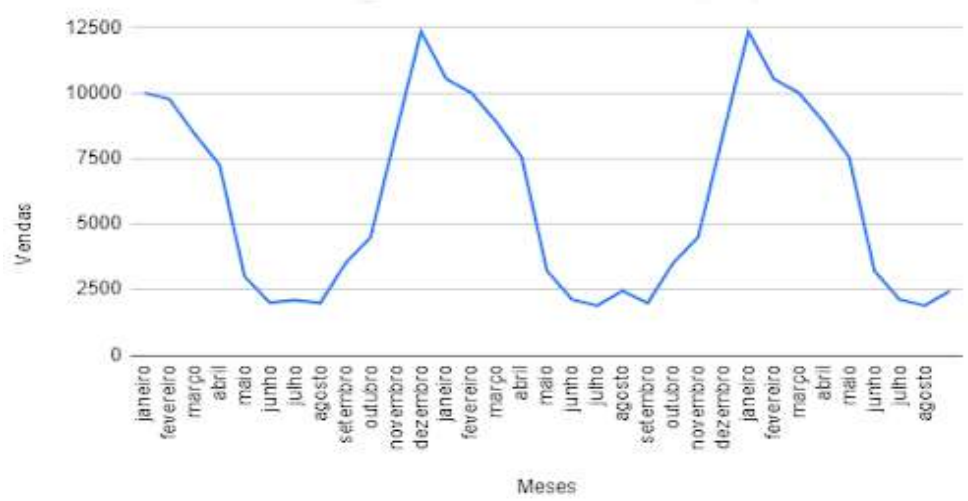

Figura 3: Exemplo ilustrativo de uma série temporal com sazonalidade nas vendas onde os meses de junho, julho e agosto apresentam queda (fonte: https://www.alura.com.br/artigos/series-temporais-tipos-de-sazonalidade).

Outro conceito importante para a compreensão de séries temporais é o de estacionariedade de uma série. De acordo com (MORETTIN, P. A.; M. C. TOLOI, C., 2006). uma série é dita estacionária se a sua média, variância e autocorrelação se mantém constantes no tempo (Figura 4).

Sendo:

- média definida por:

$$\mu(t) = E[X(t)] = \int_{-\infty}^{\infty} x_t \, f(x_t) \, dx_t$$

- variância definida por:

$$\sigma^2_{(t)} = \text{Var}\,[X(t)] = \int_{-\infty}^{\infty} (x_t - \mu_t)^2 \, f(x_t) \, dx_t$$

- autocovariância definida por:

$$\gamma(\tau) = E[X(t) - \mu][X(t + \tau) - \mu] = \text{Cov}[X(t), X(t + \tau)]$$

Onde τ é um tempo diferente de t.

- autocorrelação:

$$\rho(\tau) = \gamma(\tau)/\gamma(0) = \gamma(\tau)/\sigma 2$$

A estacionariedade é muito importante pois diversas técnicas analíticas de séries temporais necessitam que a série seja estacionária para serem funcionais. Porém, fora do mundo

teórico, é mais comum trabalharmos com séries temporais não estacionárias (Figura 5), nestes casos muitas vezes é possível utilizar mecanismos para transformá-las em estacionárias, já que para aplicarmos o modelo ARIMA, a série precisa necessariamente ser estacionária.

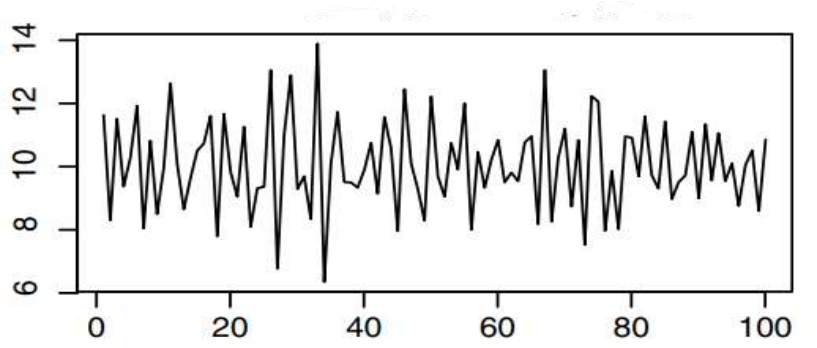

Figura 4: Exemplo de uma série temporal estacionária (fonte: BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. , 2015).

Figura 5: Ilustração da contração de uma substância qualquer em um processo químico fictício que apresenta uma série temporal não estacionária (fonte: BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. , 2015).

Caso, a série não seja estacionária existem algumas formas de tentar torná-la estacionária e ou diminuir o ruído da série causado pela aleatoriedade. Aqui, será apresentado o conceito de diferenciação para tentar tornar a série estacionária e o conceito de transformação logarítmica para tentar diminuir o ruído da série.

- Diferenciação: Basicamente consiste em fazer a diferença do termo atual com o anterior. Exemplo: Uma série $X = (X_1, X_2, X_3, ..., X_n)$ ao ser diferenciada dá origem a uma nova série que será $X' = (Z_i = X_i - X_{(i-1)}$, para $i = 2,.....,n)$
- Transformada Logarítmica: Consiste em trabalhar com o log dos valores da série, diminuindo assim a variância entre eles.

2.2) Modelos de séries temporais para séries

Podemos tentar representar o comportamento de uma dada série através de modelos estatísticos. Os modelos mais utilizados para séries temporais estacionárias são os modelos autorregressivos (AR) e os modelos de médias móveis (MA). Para séries não estacionárias uma possível solução é utilizar modelos autorregressivos integrados e de médias móveis (ARIMA). A seguir esses métodos serão apresentados.

I) Modelo AR:

Os modelos autorregressivos (AR) são modelos que fazem uso de observações passadas ou presente da própria variável para estimar um valor futuro. Ou seja, o valor presente de uma série que possui característica autorregressiva depende unicamente dos valores passados.

Escrevendo em fórmula, temos que um modelo do tipo AR(p) pode ser escrito da seguinte forma:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \phi_3 Z_{t-3} + ... + \phi_p Z_{t-p} + a_t$$

Onde:

- Z_t é a série temporal no tempo *t*
- *p* é a ordem do modelo AR.
- ϕ_i é o peso atribuído à variável no tempo *i*. Estes parâmetros são o que se busca estimar. Para $i = 1,...,p$
- a_t é o ruído branco com média zero e variância $\sigma_{a_t}^2$ para o tempo t.

Desta forma, se um modelo é do tipo AR(1), ele é descrito da seguinte forma:

$$Z_t = \phi_1 Z_{t-1} + a_t$$

II) Modelo MA:

Os modelos de médias móveis são modelos que fazem uso de observações passadas dos ruídos da variável. Ou seja, o valor presente da variável que temos o objetivo de estimar depende apenas de uma combinação linear dos valores de seus ruídos nas observações passadas.

Escrevendo em fórmula, temos que um modelo do tipo MA(q) pode ser escrito da seguinte forma:

$$Z_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \theta_3 a_{t-3} + ... + \theta_q a_{t-q}$$

Onde:

- Z_t é a série temporal no tempo t.
- q é a ordem do modelo MA.
- a_t é um ruído branco com média zero e variância $\sigma_{a_t}^2$ para o tempo t.
- θ_i é o peso atribuído ao resíduo no tempo i. Estes parâmetros são o que se busca estimar. Para $i = 0,...,q$
- μ é a média das observações de Z_t

Desta forma, se um modelo é do tipo MA(1), ele é descrito da seguinte forma:

$$Z_t = \mu + a_t + \theta_1 a_{t-1}$$

III) Modelo ARIMA(p,d,q):

Como citado no início da seção, os modelos autoregressivos e de médias móveis são utilizados sob a premissa de que a série é estacionária. Porém, é muito mais comum na prática encontrarmos séries que não satisfazem esta premissa.

Uma das formas de obter uma série estacionária a partir de uma não estacionária é diferenciando, conforme já foi citado na seção anterior. Esta abordagem unida com os dois modelos citados anteriormente, se transforma numa ferramenta poderosa chamada de modelo ARIMA. Assumindo que obtivemos uma série estacionária a partir de "*d"* diferenciações, um modelo ARMA (união dos 2 modelos anteriormente citados) é proposto para ela.

Considerando que teremos que tomar uma diferença de ordem "*d*", para tornar a série estacionária. Seja Z_t uma série temporal para t = 1, 2, 3..., uma diferença de ordem $d = 1$ é dada por:

$$\tilde{Z}_t = (1-B)(Z_t) = Z_t - Z_{t-1},$$

onde B é o operador de defasagem da forma $B(Z_t) = Z_{t-1}$.

A junção dos modelos AR(p), MA(q) com integrações para tornar no modelo estacionário, recebe o nome de ARIMA(p,d,q) chamado de modelo autorregressivo, integrado

e de médias móveis onde os parâmetros "*p"* e "*q"* representam a ordem de cada modelo e "*d*" é o número de diferenciações.

Uma série temporal Z_t, um modelo AR(p) pode ser representado meio do operador de defasagem da seguinte forma:

$$Z_t = \phi_1 Z_{t-1} + ... + \phi_p Z_{t-p} + a_t$$

$$\Rightarrow Z_t - \phi_1 Z_{t-1} - ... - \phi_p Z_{t-p} = a_t$$

$$\Rightarrow (1 - \phi_1 B - ... - \phi_p B^p) Z_t = a_t$$

$$\Rightarrow \phi(B) = a_t,$$

onde $\phi(B)$ é o operador autorregressivo.

De forma análoga, um modelo MA(q) pode também ser representado a partir do operador de defasagem da seguinte forma:

$$Z_t = a_t - \theta_1 a_{t-1} - ... - - \theta_q a_{t-q}$$

$$\Rightarrow Z_t = (1 - \theta_1 B - ... - \theta_1 B^q) a_t$$

$$\Rightarrow Z_t = \theta(B),$$

onde, $\theta(B)$ é um operador de médias móveis.

Unindo o que foi dito anteriormente, uma série diferenciada *d* vezes, $(1-B)^d Z_t$, o modelo ARIMA pode ser escrito de forma geral como:

$$\phi (B)(1-B)^d Z_t = \theta(B) a_t,$$

onde, $\phi(B)$ é o operador autorregressivo não estacionário de ordem p+q e Z_t é uma série com número finitos de diferenciações *d*.

Para fazer o ajuste do modelo, serão utilizados os estágios de ciclo iterativo proposto pelo método de Box-Jenkins (BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. , 2015) que consiste na identificação, estimação e previsão do melhor modelo para a série.

2.2.1) Determinando o modelo e os parâmetros do modelo.

O grande desafio da modelagem descrita no início da seção 2.2 é determinar as ordens do modelo escolhido quando temos apenas observações sem uma função matemática especificada. Com isso, a fase mais complexa para o modelo de Box-Jenkins é a identificação

do melhor modelo. Segundo (CRYER J. D. E CHAN, K. S., 2008), o melhor modelo é aquele que representa o menor número de parâmetros e ainda assim representa a série de forma adequada.

Uma possível forma de identificação do modelo é entendendo e analisando o comportamento das funções de autocorrelação (ACF) e de autocorrelação parcial (PACF), empíricas e investigar se elas têm o comportamento parecido com os modelos teóricos e, dependendo da ordem do modelo em questão, as ACF e PACF com características específicas.

O objetivo desta etapa de identificação é determinar quais são os parâmetros p,d,q do modelo ARIMA. Esta etapa em questão é dividida em 3 processos:

a) Fazer a transformação da série original, caso haja necessidade
b) Tomar diferenças da série em questão, quantas vezes forem necessárias para obter uma série que de fato seja estacionária, caso ela não seja.
c) Identificar a ordem de p e q a partir das funções de autocorrelação (ACF) e autocorrelação parcial (PACF).

2.2.2) Características do ACF:

Processo AR(p): apresenta ACF que decai de acordo com exponenciais e/ou senoides amortecidas, infinitas em extensão.

Processo MA(q): tem ACF finita, apresentando um corte após o "lag" q.

Processo ARMA(p,q): apresenta ACF infinita por extensão, com decaimento de acordo com exponenciais e/ou senoides amortecidas após o "lag" q – p.

2.2.3) Características do PACF:

Processo AR(p): apresenta um corte após a defasagem p, é finita em extensão.

Processo MA(q): apresenta uma PACF infinita e decai de acordo com exponenciais e/ou senoídes amortecidas.

Processo ARMA(p,q): apresenta o mesmo comportamento de uma PACF no processo MA.

A Tabela 1, a seguir, apresenta um resumo do comportamento das funções de autocorrelação em cada modelo citado:

Tabela 1: Resumo do comportamento das funções de autocorrelação para cada modelo citado anteriormente.

Função	AR	MA	ARIMA/ARMA
PACF	Decai Exponencialmente	Corte Brusco na defasagem de q	Corte Brusco na defasagem de q
ACF	Corte Brusco na defasagem de p	Decai Exponencialmente	Corte Brusco na defasagem de p

Apesar da análise das ACF e PACF nos ajudar a identificar o melhor modelo em questão, nem sempre a identificação deste é fácil através dessa análise uma vez que esta possa ter um olhar muito subjetivo e não necessariamente preciso. Por conta disso, uma alternativa matemática para a identificação é usar os critérios de comparação de modelos, como os critérios de Akaike (AIC) e o critério da informação Bayesiana (BIC) (CRYER J. D., CHAN, 2008) que minimizam funções penalizadoras.

Quanto menor forem os AIC e o BIC, melhor é o modelo para ajustar a série em questão.

O critério de Akaike (AIC) pode ser escrito da seguinte forma:

$$AIC = - 2\log(L(\Theta:z_1, ..., z_n)) + 2m,$$

onde L é a função de verossimilhança do modelo, m é o número de parâmetros e Θ representa o vetor paramétrico.

O critério de informação Bayesiana (BIC), pode ser escrito da seguinte forma:

$$BIC = - 2\log(L(\theta:z_1, ..., z_n)) + m\log(n),$$

onde n é o número de observações.

2.2.4) Estimação

Após identificar o modelo, o próximo passo do ciclo é a estimação do modelo que pode ser feita a partir do método da máxima verossimilhança.

Para determinar o EMV foi suposto que o $a_t \sim N(0, \sigma_a^2)$.

Seja Θ o vetor paramétrico do modelo em questão a ser ajustado. Sabemos que $f(z_1,..,z_n| \Theta)$ corresponde a função de probabilidade dado o parâmetro Θ, a função de verossimilhança é dada por:

$$L(\Theta) = f(z_1,..,z_n| \Theta),$$

agora vista como função de Θ.

O estimador de máxima verossimilhança é a função da amostra que maximiza a função de verossimilhança e é dado por:

$\Theta_{estimado} = argmaxL(\Theta;z_1,..,z_n)$.

2.2.5) Análise dos resíduos:

Após finalizada a estimação do modelo, é importante verificar se de fato ele representa os dados de forma coesa, ou seja, se ele realmente explica de forma adequada. Ele deve explicar a dependência temporal dos dados e satisfazer suas suposições. Qualquer insuficiência revelada sugere que pode haver a necessidade de utilizar um modelo alternativo.

No modelo em estudo, existe a suposição de que os erros se comportam como um ruído branco normal, portanto, é necessário identificar se os resíduos apresentam evidências de dependências temporais.

Este tipo de verificação pode ser feito a partir de uma análise dos resíduos. Neste estudo, serão focados os testes de Ljung- Box e Box-Pierce. Mais detalhes sobre outros testes, são encontrados em (MORETTIN, P. A.; M. C. TOLOI, C.,2006).

Teste de Ljung- Box:

Este teste é utilizado para verificar a hipótese de independência dos resíduos do ajuste em questão. A hipótese nula (H_0) é de que uma série de observações ao longo do tempo são independentes enquanto a hipótese alternativa é de que não são.

Se a autocorrelação tender a zero, a hipótese nula não será rejeitada. A rejeição da hipótese nula, significa que o modelo não é adequado para explicar a série temporal.

A estatística deste teste é dada por:

$$L = n(n+2) \sum_{K=1}^{m} \frac{\rho_K^2}{n-k},$$

sendo a estatística de teste uma qui- quadrado com m graus de liberdade, ρ_K é a autocorrelação amostral de lag k, n é o tamanho da amostra e m é o número de lags que estão sendo testados. Sob H_0, a estatística deste teste, segue uma distribuição qui- quadrado com m graus de liberdade.

Teste de Box-Pierce:

Neste teste, o objetivo e as premissas são iguais ao de Ljung-Box, diferenciando apenas a estatística de teste que é definida por:

$$Q = n \sum_{k=1}^{m} \rho_k^2,$$

onde n é o tamanho da amostra, ρ_K é a autocorrelação amostral de lag k e m é o número de lags que estão sendo testados. Sob H_0 a estatística também segue uma distribuição qui-quadrado com m graus de liberdade.

Com isso, em ambos os testes buscamos ajustes com maior p-valor.

2.2.6) Previsão

Uma vez que os parâmetros do modelo são estimados, é possível fazer uma previsão sobre um determinado tempo além do período obtido dos dados e verificar uma possível tendência de crescimento ou decaimento dos dados.

Na equação de previsão, o tempo t passa a ser fixo, passando a ser considerado o maior valor observado e *h* passa a ser a varável objetivo. Para entender a função de previsão procurar sobre o os estágios de ciclo iterativo proposto pelo método de Box-Jenkins (BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. , 2015).

3) Aplicação

Neste capítulo serão explicitados a procedência e tratamento da base de dados usada no experimento, assim como os modelos estatísticos, conceitos e suposições necessárias para a adequação do modelo escolhido.

3.1) Dados

Todos os dados deste trabalho foram retirados de um repositório de dados globais gratuito dos mais diversos temas como saúde, meio ambiente, guerras, educação, política etc. O site *ourworldindata.org* foi criado em 2011 por Marx Roser com o objetivo de ser uma plataforma que reunisse diversas pesquisas sobre temas e problemas variados do planeta e tornasse a consulta a esses dados fácil e prática.

Em relação à COVID-19, na metade de 2020 foram criados 207 perfis de países que permitem aos usuários explorar os dados e estatísticas da pandemia desde então. Cada perfil possui visualizações interativas, explicações sobre as métricas utilizadas e fontes de dados. Cada perfil de país é atualizado diariamente. Os dados brutos sobre casos e mortes de todos os países são provenientes do *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE)* da Universidade Johns Hopkins. Abaixo temos o exemplo de uma figura sobre a evolução de casos de COVID-19 no Brasil, no período de 02/03/2020 à 15/02/2022, disponibilizados no site (Figura 6).

Figura 6: Curva de casos de covid no Brasil com dados retirados do site onde foi extraída a base de dados entre março de 2020 até 2022 (fonte:Johns Hopkins University CSSE COVID 19 data retirado do site Own Word in Data).

Para este trabalho, foi utilizado o conjunto de dados *the complete Our World in Data COVID-19 dataset* que contém todos os dados da COVID-19 consolidados pelo portal no formato .csv. Como este trabalho foca-se no cenário brasileiro, os dados foram tratados e filtrados no Microsoft Excel e Microsoft Power BI para que retornassem apenas os dados de novos casos e novas mortes por dia no Brasil.

O intervalo de datas para a análise foi do dia 24/02/2020 até 31/12/2021 (sendo as 5 últimas semanas usadas apenas para comparação com a previsão). Os dados foram agrupados por semana do ano, indo de 1 até 52, juntamente com o ano correspondente. Um exemplo deste agrupamento é 25|2020 que representa os dados da semana 25 do ano de 2020.

Após os tratamentos iniciais dos dados, toda a modelagem, análise e conclusões obtidas foram feitas usando o *R Project.R* é uma linguagem e ambiente para computação estatística e gráficos. É um projeto GNU que é semelhante à linguagem e ambiente S que foi desenvolvido nos Laboratórios Bell (anteriormente *AT&T*, agora *Lucent Technologies*) por John Chambers e colegas.

A linguagem R fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento etc.) e gráficas, é altamente extensível e gratuito. Por causa dessa liberdade de edição e complementação da plataforma, o programa *RStudio*, ideal para linguagem R foi escolhido para esse trabalho.

Como trabalhamos com séries temporais e modelos ARIMA, alguns pacotes complementares foram instalados no R. São eles: "ggplot2" (H. WICKHAM, 2016.) para questões estéticas dos gráficos apresentados, "*lubridate*" (GARRETT GROLEMUND, HADLEY WICKHAM, 2011) que traz uma sintaxe amigável para trabalho com datas, "*readr*" (HADLEY WICKHAM, JIM HESTER AND JENNIFER BRYAN,2021) que facilita a leitura de arquivos.csv, "*lmtests*" (ACHIM ZEILEIS, TORSTEN HOTHORN, 2002) uma coleção de testes, conjuntos de dados e exemplos para verificação de diagnóstico em modelos de regressão linear, "*nortest*" (JUERGEN GROSS AND UWE LIGGES, 2015) são testes para testar a hipótese composta de normalidade, "*moments*" (LUKASZ KOMSTA AND FREDERICK NOVOMESTKY, 2015) que são funções para calcular momentos, curtose de Pearson, curtose e assimetria de *Geary* e testes relacionados a eles (ANSCOME-GLYNN, D'AGOSTINO, BONETT-SEIER; 2010) e por fim o pacote "*forecast*" (HYNDMAN R, ATHANASOPOULOS G, BERGMEIR C, CACERES G, CHHAY L, O'HARA-WILD M, PETROPOULOS F, RAZBASH S, WANG E, YASMEEN F,2021) que será melhor abordado no parágrafo abaixo.

O pacote "*forecast*" do R traz métodos e ferramentas para exibir e analisar previsões de séries temporais univariadas, incluindo suavização exponencial por meio de modelos de espaço de estado e modelagem ARIMA automática. Estes métodos e ferramentas serão cruciais para este trabalho.

Uma das principais ferramentas do pacote *Forecast* é a função `auto.arima()`. Essa função utiliza uma variação do algoritmo de *Hyndman-Khandakar* (HYNDMAN & KHANDAKAR, 2008), que combina testes de raiz única, minimização do Critério de Informação de *Akaike* corrigido (AICc) e de estimação de máxima verossimilhança (MLE)

identificar e ajustar um modelo ARIMA usando a técnica de stepwise.(J. HYNDMAA, R.; ATHANASOPOULOS, G.2018)

O algoritmo Hyndman-Khandakar para modelagem *auto.arima()* pode ser descrito pelos seguintes passos:

Para nossos dados, consideramos o processo ARIMA não-sazonal:

$\phi(B)\ (1-B)^{d}Y_{t}\ (1-B)^{d}\ \phi(B)a_{t}$

1. O número de diferenças $0 \leq d \leq 2$ é determinado usando testes KPSS (*Kwiatkowski Philips Schmidt e Shin*) (HYNDMAN, R. J., & KHANDAKAR, Y., 2008) repetidas vezes.

2. Os valores de p e q são escolhidos minimizando os AICc após diferenciar os dados d vezes. Em vez de considerar todos os valores possíveis para p e q , o algoritmo usa o método *stepwise* para percorrer o espaço do modelo.
    a) Quatro modelos iniciais são ajustados: ARIMA (0,d,0), ARIMA (2,d,2), ARIMA (1,d,0), ARIMA (0,d,1).

Uma constante é incluída a menos que $d = 2$. Se $d \leq 1$ um modelo adicional é montado: ARIMA (0,d,0) sem a constante

    b) O melhor modelo (com o menor valor de AICc) ajustado na etapa (a) é definido como o "modelo atual").
    c) Variações no modelo atual são consideradas:
        - variar p e/ou q do modelo atual por ± 1.
        - incluir/excluir c do modelo

O melhor modelo considerado até agora (seja o modelo atual ou uma dessas variações) passa a ser o novo modelo atual.

    d) Repete-se a etapa 2(c) até que nenhum AICc inferior possa ser encontrado.

3.2) Análise da série de casos por semana

Inicialmente, importamos os dados para o R e extraímos os dados que interessavam ao trabalho: novos casos de COVID-19 por semana e novas mortes de COVID-19 por semana. Realizamos uma pequena análise exploratória destes dados (Tabela 2).

Tabela 2: Resumo dos dados utilizados.

Estatística	Quantidade de casos diários	Quantidade de mortes diárias
Min	2	15
1º Quartil	119.560	3.648
Mediana	237.536	6.543
Média	241.711	7.007
3º Quartil	350.020	7.698
Max	539.839	21.171

Para este conjunto de dados, verificamos que a variância apresentou uma ordem de grandeza extremamente alta, mais precisamente de 2.204.122.097 . Dessa forma, preferimos trabalhar na escala logarítmica. O comportamento da série na escala logarítmica pode ser visto nas Figuras 7 e 8 onde podemos ver uma comparação do comportamento da série em sua escala original e na escala logarítmica tanto na visão de série (Figura 7) quanto na visão de histograma (Figura 8).

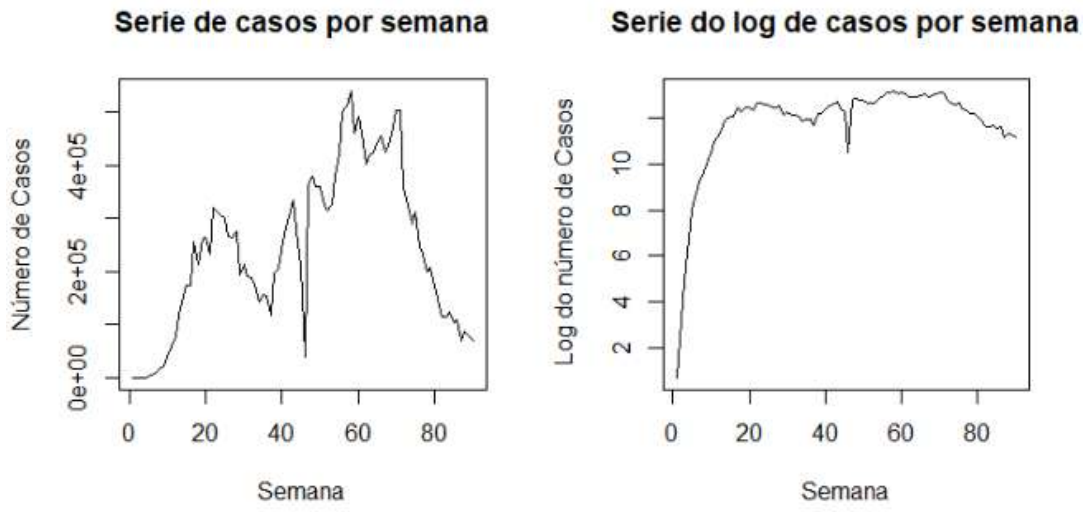

Figura 7: Número de casos de COVID-19 por dia no Brasil na escala original (direita) e na escala logarítmica (esquerda).

Na Figura 7, é possível notar que a semana 40 do nosso estudo apresenta uma queda repentina no número de mortes. Ao investigar foi notado que esta semana é referente a

primeira semana do ano de 2021, o que provavelmente reflete uma possível falha na coleta dos dados ou os dados foram represados e diluídos nas semanas seguintes.

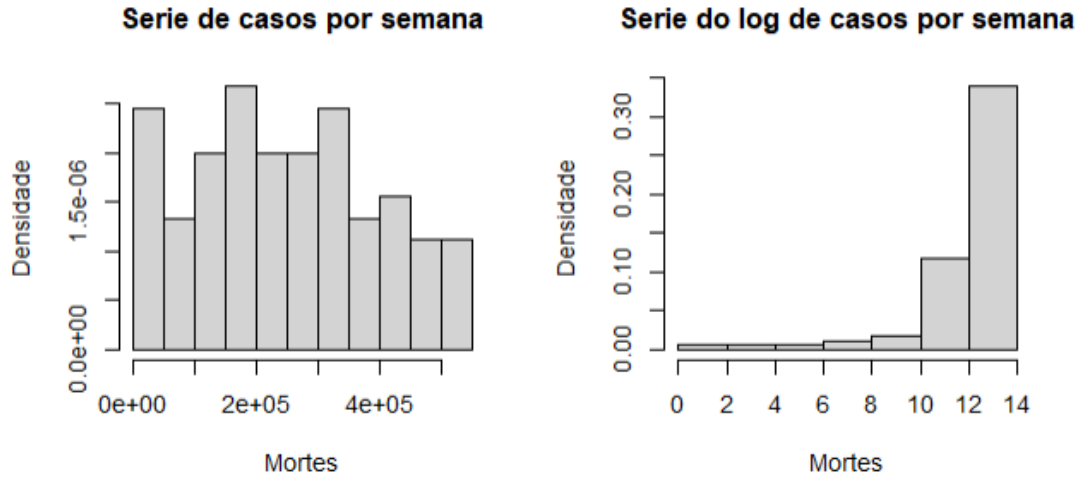

Figura 8: Histograma dos dados de casos na escala original (direita) e escala logarítmica (esquerda).

Na figura 8, é possível ver que os dados não se concentram em um valor específico, enquanto na escala logarítmica é possível ver que eles se concentram em torno do 13, tornando mais fácil uma possível modelagem em cima dessa escala. A variância para os dados nessa escala foi de 0,1359.

3.2.1) Funções de autocorrelação e modelagem ARIMA

Primeiramente iremos analisar as funções de autocorrelação (ACF) e autocorrelação parcial (PACF) a fim de escolher os parâmetros que pareçam mais apropriadas para o modelo ARIMA, ou seja, quais modelos ARIMA poderiam ajustar melhor a série temporal de COVID-19 (na escala logarítmica).

Primeiramente analisamos os gráficos sem diferenciação (Figura 9) onde obtivemos um gráfico de ACF caindo exponencialmente entre os dois primeiros "*lags*", como a PACF contendo apenas um "lag" acima do limite (indicando que q=1). Além disso, verificamos que a função PACF cai exponencialmente enquanto a ACF apresenta 9 "*lags*" acima do limite (indicando que p = 9). Nesse caso teríamos um modelo ARIMA (9,0,1). Devido ao parâmetro p estar extremamente alto, acreditamos que o modelo sem diferenciação provavelmente não é o mais adequado.

A segunda etapa foi tentar olhar um modelo para uma diferenciação (Figura 10), onde é possível ver uma ACF caindo exponencialmente nos dois primeiros "*lags*" e com dois valores acima do limite na PACF, indicando que o parâmetro q é 2 para este modelo. Porém, quando olhamos para o PACF, não vemos essa queda abrupta entre os primeiros "*lags*", indicando que

o parâmetro p desse ajuste deve ser zero. Com isso, nosso primeiro modelo a ser testado é um ARIMA(0,1,2).

O próximo ajuste a ser testado, foi o com duas diferenciações (Figura 11), onde vemos um comportamento parecido com um senóide nos primeiros "*lags*" da ACF com 2 valores passando do limite na PACF, indicando que o parâmetro q seria 2 enquanto na PACF vemos também um comportamento também de senoide nos primeiros "*lags*" com 2 valores acima do limite na ACF indicando que o melhor ajuste seria o ARIMA(2,2,2).

Por último, aplicando a função *auto.arima* do R que por métodos iterativos retorna o melhor ajuste da série vista. A resposta obtida por esta função foi o ARIMA(1,2,1).

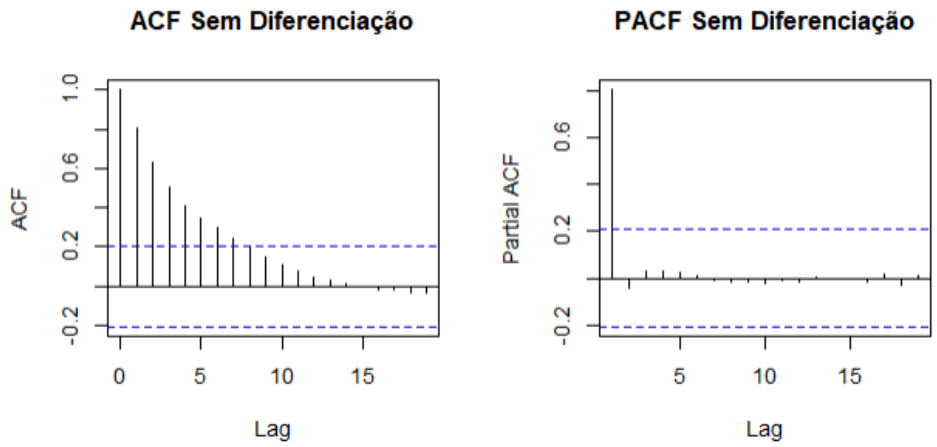

Figura 9: Funções de autocorrelação para a série de casos não diferenciada.

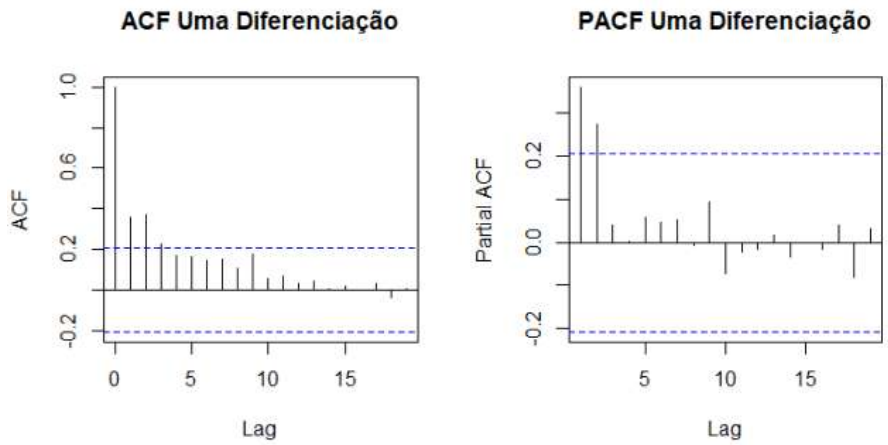

Figura 10: Funções de autocorrelação para a série diferenciada uma vez.

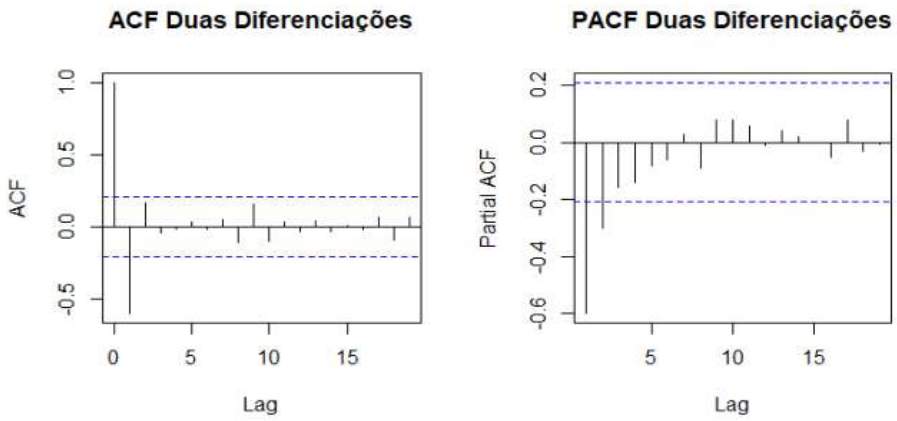

Figura 11: Funções de autocorrelação para a série diferenciada duas vezes.

Após a identificação dos melhores modelos ARIMA, iremos compará-los através do critério Akaike (AIC) e dos testes *Ljung-Box* e *Box-Pierce* para escolher o melhor modelo. Os testes de resíduos *Ljung-Box* e *Box-Pierce* testam se os resíduos do modelo se comportam como ruído branco (como deveria ser). A Tabela 3 mostra os resultados da estatística AIC e dos p-valores associados aos testes *Ljung-Box* e *Box-Pierce*.

Tabela 3:Estatísticas AIC e p-valor dos testes Box-Pierce e Ljung-Box para diferentes modelos de ARIMA – ajuste da série de casos.

Ajuste	AIC	Box-Pierce	Ljung-Box
Arima(0,1,2)	42,91885	0,004178	0,00305
Arima(1,2,1)	108,6213	0,8563	0,854
Arima(2,2,2)	50,67162	0,0008585	0,0005657

Pela Tabela 3 pode-se notar que embora o modelo escolhido de forma teórica tenha performado melhor no coeficiente de AIC, nos testes *Ljung-Box* e *Box-Pierce* o modelo resultante do auto.arima() performou muito melhor. Desta forma, a decisão tomada foi que embora o ARIMA(1,2,1) seja mais "complexo", seus resíduos se comportam melhor que os dos demais modelos fazendo com que seja melhor escolhê-lo.

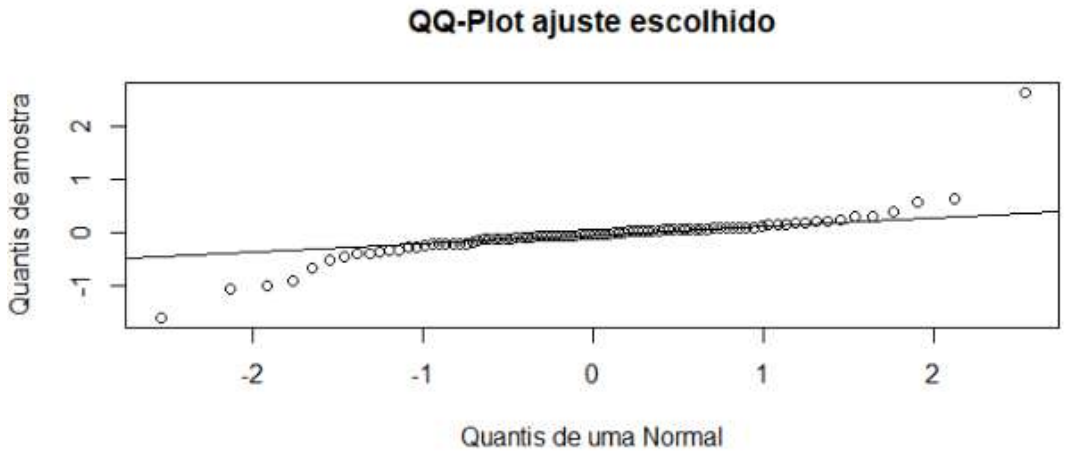

Figura 12: Gráfico QQ-Plot comparando os resíduos do ajuste de um ARIMA(1,2,1) com uma distribuição Normal a fim de comparar se os resíduos têm comportamento normal. A linha traçada representa o comportamento Normal.

A Figura 12 apresenta um gráfico QQ-Plot comparando os resíduos do ajuste do modelo ARIMA (1,2,1) com uma distribuição Normal. Pela figura vemos que o ajuste escolhido apresenta resíduos que em sua maioria se comportam de forma normal, como é desejado. Sendo assim, finalmente após garantir a independência e a aleatoriedade dos resíduos, temos um modelo escolhido para fazer previsões.

3.2.2) Previsões do modelo:

Nessa subseção faremos a previsão dos cinco últimos valores da série temporal semanal para comparar com os valores que foram de fato observados.

Como o ajuste para o modelo foi feito com os valores na escala logarítmica dos dados e o objetivo desta etapa foi calcular o valor de previsão do número semanal de casos, para isso, foi necessário retornar para a escala natural dos dados para assim calcular os valores das predições e seus intervalos de confiança a um nível de 95% usando a seguinte fórmula:

Limite Inferior = Valor Predito – Quantil 0,975 da N(0,1) x Erro quadrático.

Limite Superior = Valor Predito + Quantil 0,975 da N(0,1) x Erro quadrático.

A Figura 13 mostra a série temporal do número de casos de Covid-19 por semana série temporal com previsão do número de casos para as últimas 5 semanas e seus respectivos intervalos de confiança, além do valor real registrado para essas semanas:

Figura 13: Série Temporal dos casos semanais na escala original a predição das próximas cinco semanas e seu respectivo intervalo de confiança com nível de 95%.

A Tabela 4 apresenta essas previsões numericamente, assim como os limites dos intervalos de confiança, em comparação com os valores reais.

Tabela 4: Número de casos reais x predição para o modelo ARIMA(1,2,1) para as semanas 91 a 95.

Semana	Real	Predito	Lim Inf	Lim Sup
91	78.827	66.030	21.439	203.366
92	58.715	60.972	10.798	344.274
93	65.394	56.556	4.730	676.090
94	60.552	52.391	1.951	1.406.343
95	49.480	48.551	748	3.150.598

Para critério de comparação, vemos também a tabela das previsões do ARIMA(2,2,2) (Tabela 5) e do ARIMA(0,1,2) (Tabela 6). Onde podemos ver que de fato os outros dois ajustes não são muito consistentes, uma vez que os limite inferiores dos intervalos de confiança de ambos ficaram em zero e no ARIMA(0,1,2) a previsão chega a ser constante para 4 das 5 semanas estimadas mostrando que de fato o ajuste escolhido é o melhor dentre os três.

Tabela 5: Número de casos reais x predição para o modelo ARIMA(2,2,2) para as semanas 91 a 95.

Semana	Real	Predito	Lim Inf	Lim Sup
91	78.827	63.987	0	168.635
92	58.715	57.795	0	197.094
93	65.394	51.463	0	224.107
94	60.552	45.197	0	245.876
95	49.480	38.900	0	266.382

Tabela 6: Número de casos reais x predição para o modelo ARIMA(2,1,2) para as semanas 91 a 95.

Semana	Real	Predito	Lim Inf	Lim Sup
91	78.827	71.695	0	175.155
92	58.715	71.393	0	207.416
93	65.394	71.393	0	235.321
94	60.552	71.393	0	259.124
95	49.480	71.393	0	280.231

Podemos observar nas Tabelas 5 e 6 acima que os modelos ARIMA (2,1,2) e ARIMA (2,2,2) não se comportam tão bem quanto o ARIMA (1,2,1) para fazer previsões. O limite inferior de ambos permaneceu inalterado em 0 enquanto o limite superior crescia e os valores preditos não se aproximavam dos valores reais, o que comprova mais uma vez a eficácia do modelo escolhido em relação aos demais.

3.3) Análise da série de mortes por semana

Nesta seção faremos uma análise similar à que fizemos para o número de casos semanais de Covid-19, porém agora considerando a série de contagens de mortes por semana. Também para esse caso trabalharemos na escala logarítmica pois com a transformação os dados apresentam um melhor comportamento. A Figura 15 mostra a série diária de casos de morte por Covid-19 na escala original e na escala logarítmica.

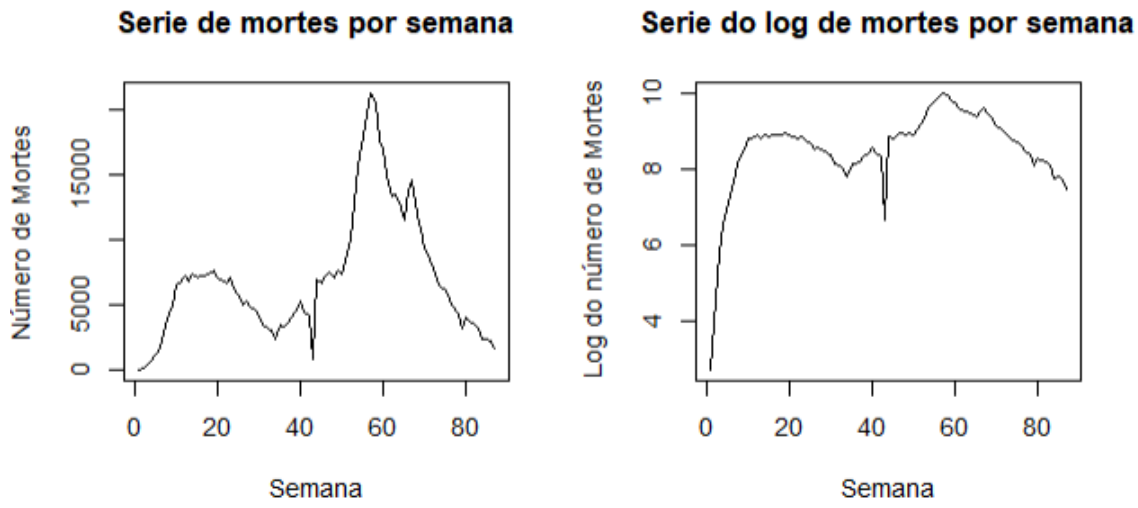

Figura 14: Distribuição das mortes na semana na escala original (direita) e na escala logarítmica (esquerda).

Na Figura 14 é possível notar que a semana 40 do nosso estudo apresenta uma queda repentina no número de mortes. Ao investigar foi notado que esta semana é referente à primeira semana do ano de 2021, o que provavelmente reflete uma possível falha na coleta dos dados ou os dados foram represados e diluídos nas semanas seguintes. Vale ressaltar que o comportamento da série de mortes semanais é bem parecido com a série de casos semanais, principalmente quando olhamos na escala logarítmica, como é de se esperar.

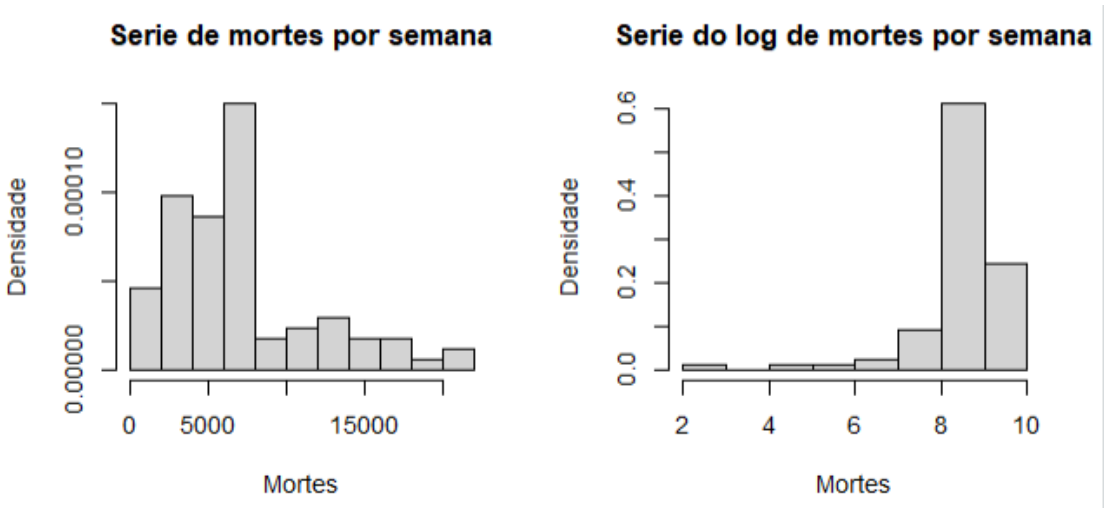

Figura 15: Histograma para densidade de mortes na escala original(direita) e logarítmica (esquerda).

Na Figura 15, vemos que na escala original as mortes semanais se concentram em torno de 7.000 em sua maioria, enquanto na escala utilizada no estudo se concentram próximas de 8.

Ao refazer a análise exploratória para os dados na nova escala, foi notada uma redução na variância de 23.155.643 para 1,15.

3.3.1) Funções de autocorrelação e modelagem ARIMA:

Repetindo o que foi feito na série de casos, aqui as ACF e PACF empíricas são olhadas e comparadas com o esperado de um modelo ARIMA, para assim, verificarmos qual ajuste do modelo ARIMA que mais se encaixa na nossa série.

De forma análoga à outra série, primeiro olhamos o gráfico sem diferenciação (Figura 16), vemos um comportamento de senoíde na ACF com 1 valor acima do limite na PACF indicando que o parâmetro *q* é 1 enquanto na PACF vemos uma queda exponencial com 7 valores acima do limite, indicando que o valor do parâmetro *p* é 7. Com isso, temos um ARIMA(7,0,1). Como o parâmetro *p* mostrou-se exageradamente alto, o modelo ARIMA (sem diferenciação) será descartado.

O segundo ajuste a ser testado foi o com uma diferenciação (Figura 17), onde vemos a ACF com queda exponencial e a PACF com 1 valor acima do limite, indicando o parâmetro *q* sendo 1, enquanto na PACF vemos uma subida exponencial com 2 valores acima do limite na ACF, indicando assim o parâmetro *p* sendo 2. Com isso temos o ajuste proposto sendo um ARIMA(2,1,1).

O outro ajuste a ser testado foi o com duas diferenciações (Figura 18), onde vemos a ACF com um comportamento senoide nos primeiros "*lags*" e 2 valores acima do limite na PACF, indicando que o parâmetro *q* é igual a 2. Na PACF vemos também um comportamento senoidal nos primeiros "*lags*" sendo 2 valores acima do limite na ACF, indicando que o valor do parâmetro *p* também é igual a 2. Sendo assim, os indícios nos levam a crer que o melhor modelo teórico para esse ajuste é o ARIMA(2,2,2).

Por último, aplicando a função "*auto.arima*" no *R*, o melhor ajuste obtido de forma iterativa foi o ARIMA(2,2,2), que ao contrário do que ocorreu para a série de casos, foi a mesma encontrada com base nas funções de ACF e PACF.

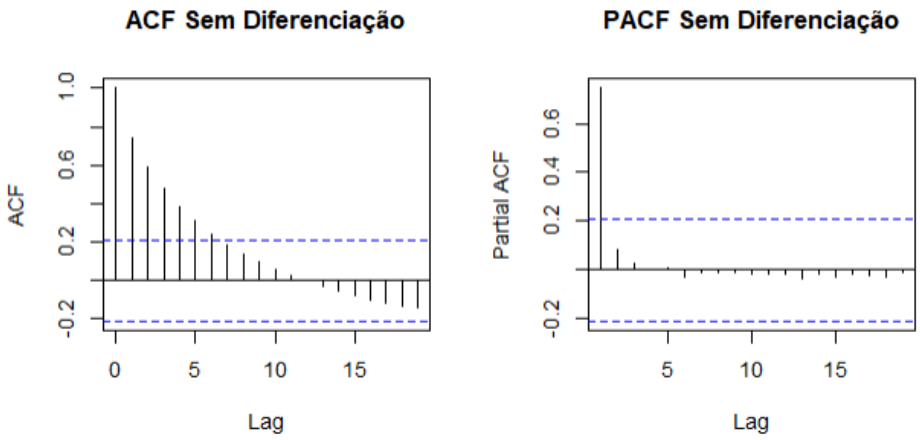

Figura 16: Funções de autocorrelação das mortes sem diferenciação.

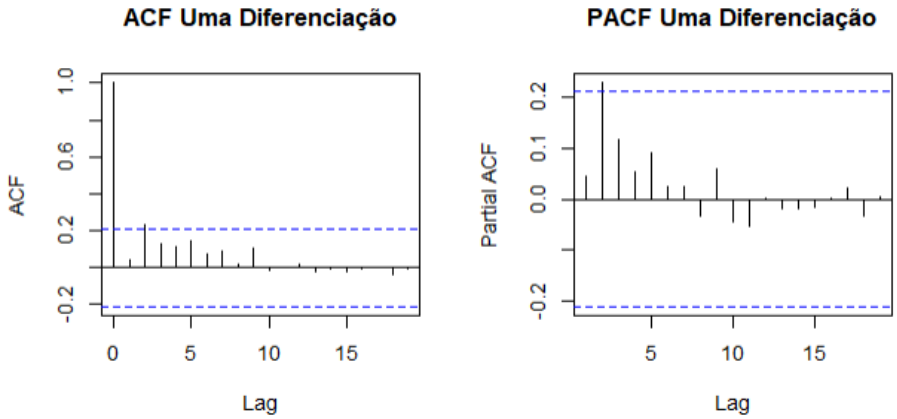

Figura 17: Funções de autocorrelação das mortes para uma diferenciação.

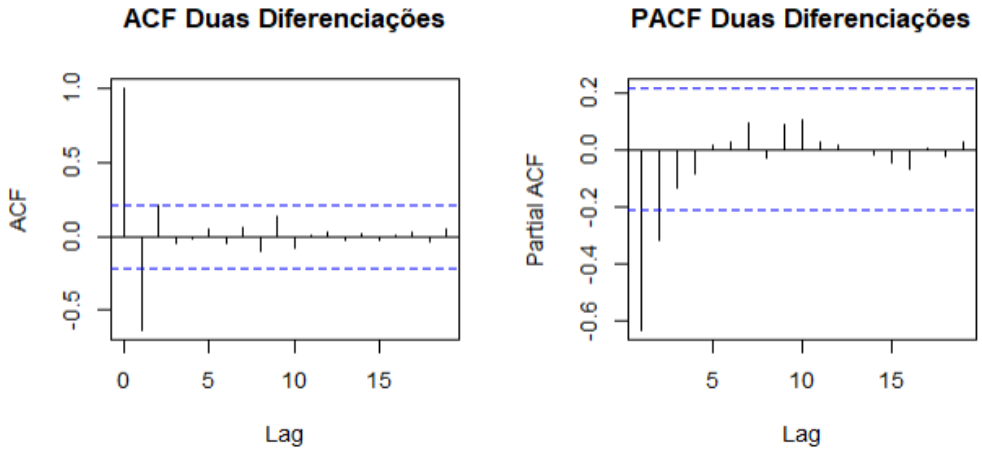

Figura 18: Funções de autocorrelação das mortes para duas diferenciações.

Na Tabela 7, podemos notar que o ajuste ARIMA(2,2,2) de fato é melhor, uma vez que possui o AIC maior indicando melhor ajuste, possui o p- valor mais alto para os testes de *Box-Pierce* e de *Ljung-Box*, indicando que os resíduos se comportam de forma mais parecida com um ruído normal do que o outro modelo. Com isso, decidimos seguir com este ajuste.

Tabela 7: Estatísticas AIC e p-valor dos testes Box-Pierce e Ljung-Box para diferentes modelos ARIMA – ajuste da série de mortes.

Ajuste	AIC	Box-Pierce	Ljung-Box
Arima(2,2,2)	88,23279	0,6259	0,62
Arima(2,1,1)	83,94661	0,5878	0,5813

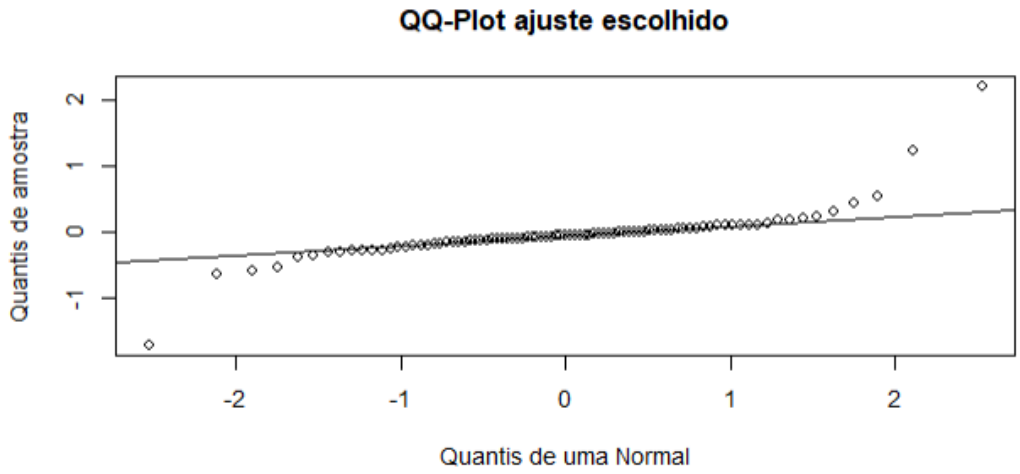

Figura 19: Gráfico QQ-Plot comparando os resíduos do ajuste de um ARIMA(2,2,2) com uma distribuição Normal.

A Figura 19 apresenta um gráfico QQ-Plot onde vemos o quão próximo de uma distribuição Normal estão os resíduos do ajuste do modelo ARIMA(2,2,2). Pela figura, vemos que o ajuste escolhido apresenta resíduos que em sua maioria se comportam de forma Normal, garantindo a aleatoriedade que é buscada para eles nesse tipo de modelagem. Sendo assim, após garantir a independência e a aleatoriedade dos resíduos, verificamos que o modelo ARIMA(2,2,2) parece apropriado para fazer previsões.

3.3.2) Previsões do modelo:

Nessa subseção faremos a previsão dos cinco últimos valores da série temporal semanal para comparar com os valores que foram de fato observados.

Como o ajuste para o modelo foi feito em cima dos valores na escala logarítmica dos dados e o objetivo desta etapa foi calcular o valor de previsão do número semanal de casos, foi necessário retornar para a escala natural dos dados para assim calcular os valores das predições e seus intervalos de confiança a um nível de 95% usando a seguinte fórmula:

Limite Inferior = Valor Predito – Quantil 0,975 da N(0,1) x Erro quadrático.
Limite Superior = Valor Predito + Quantil 0,975 da N(0,1) x Erro quadrático.

A Figura 20, nos mostra a distribuição da nossa série temporal com previsão do número de mortes para 5 semanas, seus respectivos intervalos de confiança, além do valor real registrado para essas semanas:

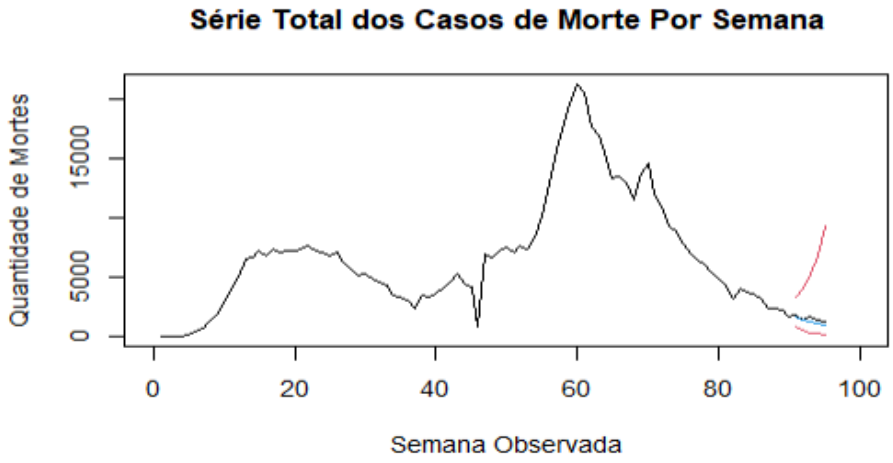

Figura 20: Série temporal das mortes semanais na escala original a predição das próximas cinco semanas e seu respectivo intervalo de confiança com nível de 95%.

Sabendo da dificuldade de se olhar numericamente os valores preditos, a Tabela 6 apresenta uma tabela na qual torna-se um pouco mais fácil a comparação dos valores preditos:

Tabela 8: Número de mortes x predição para o modelo ARIMA(2,2,2), para as semanas 91 a 95

Semana	Real	Predito	Lim Inf	Lim Sup
91	1.835	1.564	748	3.270
92	1.372	1.368	505	3.708
93	1.608	1.198	321	4.465
94	1.368	1.038	194	5.531
95	1.185	892	109	7.253

Abaixo, vemos os valores preditos pelo modelo alternativo ARIMA(2,1,1) (Tabela 9), onde vemos que embora os valores preditos por este modelo estejam tão próximos dos valores verdadeiros quanto aos obtidos pelo ARIMA(2,2,2), seus intervalos de confiança são mais largos, mostrando que de fato o modelo escolhido apresenta uma melhor performance comparado com o modelo alternativo apresentado.

Tabela 9: Número de casos reais x predição para o modelo ARIMA(2,1,1), para as semanas 91 a 95

Semana	Real	Predito	Lim Inf	Lim Sup
91	1.835	1.565	740	3.312
92	1.372	1.373	479	3.936
93	1.608	1.230	292	5.177
94	1.368	1.097	176	6.825
95	1.185	984	104	9.285

4) Conclusão

Após a discussão dos resultados encontrados na previsão do número tanto de casos quanto de mortes de COVID 19 no Brasil para as semanas 91 até 95 do nosso estudo, semanas estas que representam as semanas 41 a 45 do ano de 2021, conseguimos destacar os pontos detalhados a seguir.

Verificamos que o modelo proposto pareceu se ajustar bem aos dados e gerou previsões razoáveis a curto prazo, quando comparadas com os valores observados na realidade. O interessante é que ao olharmos as 2 últimas semanas, o intervalo de confiança aumenta muito rapidamente, indicando que os erros quadráticos estão relativamente grandes, uma vez que estes influenciam diretamente nos limites tanto inferior quanto superior do nosso modelo. Isso prova que o modelo se encaixa para previsões de 1 até 3 semanas à frente, porém ele se torna pouco eficaz para previsões acima desses intervalos semanais.

Analisando o número de mortes reais e previstas apresentadas na Tabela 6 usando um modelo semelhante ao de casos de COVID-19, podemos notar que a predição dos dados funcionou bem para as duas primeiras semanas, enquanto as demais apresentam intervalos de confiança muito amplos e suas predições estão mais afastadas dos valores observados.

De modo geral, podemos dizer que o ajuste performou muito bem para prever três semanas de casos e duas semanas de mortes. Sendo assim, concluímos ter conseguido obter um modelo bom para predição de ambas as variáveis estudadas para futuros próximos (de duas ou três semanas).

Referências Bibliográficas

ACHIM ZEILEIS, TORSTEN HOTHORN . ***DIAGNOSTIC CHECKING IN REGRESSION RELATIONSHIPS.*** R NEWS, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. ***TIME SERIES ANALYSIS: FORECAST AND CONTROL***, 2015

CHOI K, THACKER SB. ***AN EVALUATION OF INFLUENZA MORTALITY SURVEILLANCE, 1962- 1979. I. TIME SERIES FORECASTS OF EXPECTED PNEUMONIA AND INFLUENZA DEATHS. AM J EPIDEMIOL.*** 1981.

CRYER, J. D., CHAN, K. S. ***TIME SERIES ANALYSIS - WITH APPLICATIONS IN R.*** , 2008.

GARRETT GROLEMUND, HADLEY WICKHAM. ***DATES AND TIMES MADE EASY WITH LUBRIDATE***. JOURNAL OF STATISTICAL SOFTWARE, 2011. url https://www.jstatsoft.org/v40/i03/.

H. WICKHAM. GGPLOT2: ***ELEGANT GRAPHICS FOR DATA ANALYSIS***. SPRINGER-VERLAG NEW YORK, 2016.

HADLEY WICKHAM, JIM HESTER AND JENNIFER BRYAN,. ***READR: READ RECTANGULAR TEXT DATA. 2021***. R package version 2.1.1. https://CRAN.R-project.org/package=readr.

HYNDMAN, R. J.; ATHANASOPOULOS, G. ***ARIMA MODELING IN R. FORECASTING: PRINCIPLES AND PRACTICE,*** 2018

HYNDMAN, R. J., & KHANDAKAR, Y. ***AUTOMATIC TIME SERIES FORECASTING: THE FORECAST PACKAGE FOR R. JOURNAL OF STATISTICAL SOFTWARE***, 2008

LINDEN A., ADAMS J.L., ROBERTS N. ***AN ASSESSMENT OF THE TOTAL POPULATION APPROACH FOR EVALUATING DISEASE MANAGEMENT PROGRAM EFFECTIVENESS,*** 2003.

MONTGOMERY, DOUGLAS C., Cheryl L. JENNINGS, and MURAT KULAHC***I. INTRODUCTION TO TIME SERIES ANALYSIS AND FORECASTING.***, 2015.

MORETTIN, P. A.; M. C. TOLOI, C. **ANÁLISES DE SÉRIES TEMPORAIS**, 2006.

SATO, R. C. ***DISEASE MANAGEMENT WITH ARIMA MODEL IN TIME SERIES,*** 2013.

SOYIRI IN, REIDPATH DD. ***EVOLVING FORECASTING CLASSIFICATIONS AND APPLICATIONS IN HEALTH FORECASTING,*** 2012.

YUAN, Z.; ZHONH-GIN, G.; PEI-FENG, L. ***PREDICTION OF BRUCELLOSIS EPIDEMIC TREND BASED ON ARIMA MODEL. CHINESE JOURNAL OF DISEASE CONTROL & PREVENTION***, 2019

ZHIRUI, H.; HONGBING, T. ***EPIDEMIOLOGY AND ARIMA MODEL OF POSITIVE-RATE OF INFLUENZA VIRUSES AMONG CHILDREN IN WUHAN, CHINA: A NINE-YEAR RETROSPECTIVE STUDY. INTERNATIONAL JOURNAL OF INFECTIOUS DISEASES.***, 2018.