UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

ANTÔNIO VICTOR ALVES SILVA

IMPACTO DA PANDEMIA DE COVID-19 NA IBOVESPA: UMA ANÁLISE ESTATÍSTICA COM MODELOS DE MACHINE LEARNING PROPHET E AUTOARIMA

CAMPINA GRANDE - PB
2023

ANTÔNIO VICTOR ALVES SILVA

IMPACTO DA PANDEMIA DE COVID-19 NA IBOVESPA: UMA ANÁLISE ESTATÍSTICA COM MODELOS DE MACHINE LEARNING PROPHET E AUTOARIMA

Trabalho de Conclusão de Curso (artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Sílvio Fernando Alves Xavier Júnior

CAMPINA GRANDE - PB
2023

S586i Silva, Antonio Victor Alves.
Impacto da pandemia de Covid-19 na Ibovespa [manuscrito] : uma análise estatística com modelos de *machine learning* Prophet e AutoARIMA / Antonio Victor Alves Silva. - 2023.
34 p.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2023.
"Orientação : Prof. Dr. Sílvio Fernando Alves Xavier Júnior, Coordenação do Curso de Estatística - CCT. "

1. Índice Bovespa. 2. Análise estatística. 3. Prophet. 4. AutoARIMA. I. Título

21. ed. CDD 519.5

Elaborada por Talita M. A. Tavares - CRB - CRB 15/971 BC/UEPB

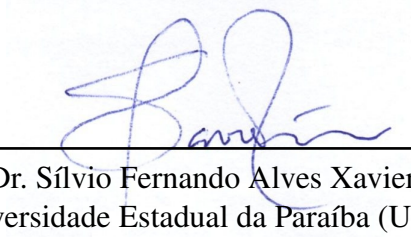
ANTÔNIO VICTOR ALVES SILVA

IMPACTO DA PANDEMIA DE COVID-19 NA IBOVESPA: UMA ANÁLISE ESTATÍSTICA COM MODELOS DE MACHINE LEARNING PROPHET E AUTOARIMA

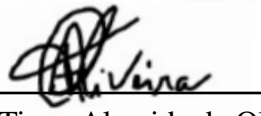
Trabalho de Conclusão de Curso (artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Trabalho aprovado em 27 de fevereiro de 2023.

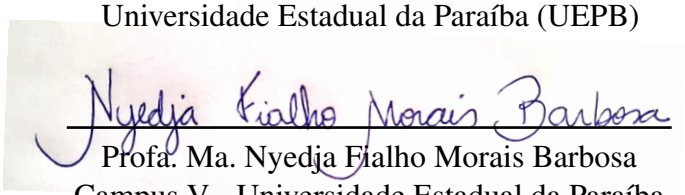
BANCA EXAMINADORA



Prof. Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba (UEPB)



Profa. Ma. Nyedja Fialho Morais Barbosa
Campus V - Universidade Estadual da Paraíba (UEPB)

Dedico a minha família pelo amor e apoio incondicional que me deram durante toda a minha jornada. Em especial, à minha mãe, por ter sido meu porto seguro e ter acreditado em mim sempre. Aos meus professores pelo conhecimento e orientação valiosos que me ajudaram a crescer e chegar até aqui. E ao meu orientador, Professor Sílvio, que foi um guia constante e dedicado durante todo este processo. Obrigado por tudo.

“O aprendizado de máquina é uma das principais tecnologias que mudarão o mundo na próxima década e além. Acho que é uma das coisas mais importantes que alguém pode estudar se quiser ter um impacto positivo na vida das pessoas.”
(Elon Musk)

SUMÁRIO

IMPACTO DA PANDEMIA DE COVID-19 NA IBOVESPA: UMA ANÁLISE ESTATÍSTICA COM MODELOS DE MACHINE LEARNING PROPHET E AUTOARIMA

Antônio Victor Alves Silva *
Dr. Sílvio Fernando Alves Xavier Júnior †

RESUMO

Esse trabalho analisou o impacto da pandemia de covid-19 no ano de 2020 nas ações brasileiras utilizando o indice bovespa e identificou valores atípicos nos dados, também observou uma tendência de estabilidade nos anos seguintes indicando recuperação econômica. A sazonalidade nos padrões regulares foi identificada e representada em um gráfico de linha, destacando as piores medianas nos meses de junho e julho. Foram utilizados modelos Prophet e autoARIMA para previsões, e os resultados foram avaliados usando várias métricas de erro, entre eles o RMSE, MAE, SMAPE, MAPE, MASE e RSQ. Embora o modelo Prophet tenha apresentado melhor desempenho com os dados diferenciados, o modelo AutoARIMA apresentou melhor desempenho com os dados originais e transformados com log1p. O estudo é relevante para entender o impacto da pandemia nas ações brasileiras e como os modelos de previsão podem ser usados para ajudar na tomada de decisões.

Palavras-chaves: Indice Bovespa; Análise estatística; Prophet; AutoARIMA.

ABSTRACT

This work analyzed the impact of the covid-19 pandemic in the year 2020 on Brazilian stocks using the Bovespa index and identifying atypical values in the data, it also observed a trend of stability in the subsequent years indicating economic recovery. Seasonality in regular patterns was detected and illustrated in a line graph, highlighting the worst medians in the months of June and July. Forecasting models, including Prophet and autoARIMA, were utilized and the results were evaluated using several error metrics, such as RMSE, MAE, SMAPE, MAPE, MASE, and RSQ. Although the Prophet model showed better performance with differenced data, the AutoARIMA model performed better with the original and log1p-transformed data. This research is significant in comprehending the impact of the pandemic on Brazilian stocks and how forecasting models can be utilized to aid decision-making.

Keywords: Bovespa Index; Statistical analysis; Prophet; AutoARIMA.

1 INTRODUÇÃO

O impacto da pandemia sars-covid19 nas ações brasileiras pode ser avaliado por meio do Índice Bovespa. De acordo com as informações disponibilizadas no site da B3 (s.d.), o Bovespa é uma medida importante do desempenho das ações negociadas na B3, juntando as empresas mais significantes do mercado de capitais brasileiro, recalculado a cada quatro meses e representando cerca de 80% do volume financeiro do mercado. O indicador financeiro utiliza o retorno total

* Aluno do curso de estatística, Depto de ciências e tecnologia, UEPB, Campina Grande, PB, antonio.victor@aluno.uepb.edu.br
† Prof. do curso de estatística, Depto de ciências e tecnologia, UEPB, Campina Grande, PB, silvio@servidor.uepb.edu.br

das ações como critério, refletindo variações e distribuição de proventos das empresas, sendo amplamente utilizado como referência para rentabilidade de fundos de ações e para avaliar o desempenho da Bolsa.

Para analisar o índice Bovespa, é importante considerar a utilização de séries temporais, como apontado por Nielsen (2021) no livro "Análise Prática de Séries Temporais". As séries temporais são ferramentas estatísticas utilizadas para prever eventos futuros a partir de dados passados. Sua simplicidade e transparência tornam seu uso intuitivo e acessível. Morettin e Toloi (2006) apontam que a análise de séries temporais tem como principais objetivos descrever o comportamento da série, investigar o mecanismo gerador dos dados, procurar periodicidades nos dados e fazer previsões dos valores futuros. Entre esses objetivos, a previsão é uma das mais utilizadas, especialmente em séries econômicas e financeiras.

Usando as séries temporais do Ibovespa, é possível treinar modelos de machine learning como Prophet e AutoARIMA. De acordo com o IBM Cloud Education (2022), o objetivo do machine learning é simular a forma como os humanos aprendem, melhorando a precisão gradualmente através da utilização de dados e algoritmos. Ao usar as séries temporais do Ibovespa como dados de entrada para o machine learning, é possível treinar modelos que possam prever o comportamento futuro do mercado de ações.

De acordo com um estudo realizado pela Faculdade Getúlio Vargas (FGV), houve uma queda significativa no desempenho do índice Bovespa entre janeiro e abril de 2020, registrando uma baixa de 32,4%, a maior entre os principais mercados de ações do mundo. Nesse contexto, surge a questão sobre a capacidade dos algoritmos de machine learning em prever com precisão o comportamento do mercado de ações após uma queda tão acentuada.

O Prophet foi criado pelo *Facebook (s.d.)* e, segundo ele, é uma ferramenta utilizada para prever informações em séries temporais. Ele usa um modelo aditivo para ajustar as tendências não lineares, como sazonalidade diária, semanal, anual e de feriados. É efetivo em séries temporais com fortes padrões sazonais e muitos dados históricos, e é capaz de lidar com dados faltantes e mudanças nas tendências, incluindo outliers. O *Facebook (s.d.)* afirma que muitos aplicativos usam o Prophet para obter previsões precisas e que ele tem desempenho superior a outras técnicas, com previsões realizadas rapidamente graças ao ajuste de modelos.

O pacote forecast de Hyndman e Khandakar (2008), escrito na linguagem de programação R (R CORE TEAM, 2021), disponibiliza a função "auto.arima"para o ajuste automático de modelos ARIMA. Essa função utiliza um método de busca exaustiva que considera todas as combinações possíveis de termos autoregressivos e de médias móveis, dentro de certos limites, para selecionar o melhor modelo que minimize uma função de perda. No entanto, quando se lida com uma grande quantidade de séries simultaneamente, esse método pode se tornar impraticável, especialmente para dados sazonais que podem gerar centenas ou até milhares de modelos alternativos. Para resolver esse problema, Hyndman e Khandakar (2008) desenvolveram um algoritmo de busca em etapas (stepwise) que reduz significativamente a quantidade de combinações testadas.

Este trabalho tem como objetivo analisar o impacto da pandemia de covid-19 no ano de 2020 nas ações brasileiras, especificamente no índice bovespa, por meio de técnicas de análise estatística. Além disso, a identificação da sazonalidade nos padrões regulares do índice também será realizada. Será aplicada uma estratégia de previsão com as ferramentas de machine learning Prophet e AutoARIMA, avaliando os resultados com métricas de erro, como RMSE, MAE, SMAPE, MAPE, MASE e RSQ. Por fim, será feita uma comparação do desempenho dos modelos com os dados originais e transformados para avaliar qual apresenta melhor precisão na previsão de 15 dias.

2 METODOLOGIA

Neste estudo, foi utilizado um conjunto de dados contendo informações sobre o Índice Bovespa obtido por meio do *Yahoo! Finanças*. Os dados foram coletados de maneira metódica e rigorosa do período de 02 de janeiro de 2020 a 29 de dezembro de 2022, a partir do fechamento diário dos dados, totalizando 670 observações.

2.1 Estacionariedade

Existem duas formas de se avaliar a estacionariedade de um processo estocástico: a estacionariedade forte (ou estrita) e a estacionariedade fraca (ou de segunda ordem). De acordo com Morettin e Toloi (2006), temos:

Estacionariedade Forte: Um processo estocástico $Z = \{Z_t, t \in T\}$ é considerado estritamente estacionário se todas as distribuições finito-dimensionais $F(z_1,...,z_n;t_1,...,t_n) = P\{Z_{t_1} \leq z_1,...,Z_{t_n} \leq z_n\}$ permanecem inalteradas no tempo. Em outras palavras, a média e a variância do processo são constantes nas translações do tempo, ou seja:

$$E(t) = \mu \quad \text{e} \quad Var(t) = \sigma^2, \quad \forall t \in T. \tag{1}$$

Estacionariedade Fraca: Um processo estocástico $Z = \{Z_t, t \in T\}$ é considerado fracamente estacionário se e somente se:

1. $E[Z_t] = \mu$, constante $\forall t \in T$;
2. $E^2[Z_t] < \infty$;
3. $\gamma_{t,s} = Cov[Z_t, Z_s]$ é uma função de $t - s$, chamada de defasagem.

2.2 Modelagem Box-Jenkins

A metodologia de Box-Jenkins é amplamente utilizada na construção de modelos paramétricos para séries temporais univariadas. O processo envolve quatro etapas principais, descritas por Morettin e Toloi (2006): especificação, identificação, estimação e verificação ou diagnóstico.

Durante a especificação, uma classe geral de modelos é considerada. Em seguida, um modelo é identificado por meio da análise de autocorrelações, autocorrelações parciais e outros critérios. Os parâmetros do modelo são então estimados na etapa de estimação.

Finalmente, o modelo escolhido é verificado para ver se se ajusta adequadamente aos dados, através da análise de resíduos. Se o modelo não for adequado, o processo é repetido, retornando à fase de identificação. Vários modelos podem ser estimados em uma série e, se a finalidade da estimação for a previsão, busca-se o modelo com o menor erro quadrático médio de previsão.

2.2.1 Modelo autorregressivo (AR)

Considerando que x_t seja um processo puramente aleatório com média zero e variância σ^2, um processo Z_t é designado como um processo autoregressivo de ordem p, ou AR(p), se:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t, \tag{2}$$

em que, ϕ_i para todo $i = 1,2,...,p$ são parâmetros do modelo e ε_t é o ruído branco no tempo t.

2.2.2 *Modelo de médias móveis (MA)*

Se considerarmos que x_t seja um processo discreto puramente aleatório com média zero e variância σ^2, então o processo Z_t é designado como uma média móvel de ordem q, ou MA(q), se:

$$Z_t = x_t - \theta_1 x_{t-1} - \theta_2 x_{t-2} - ... - \theta_q x_{t-q}, \tag{3}$$

onde há q defasagens na média móvel e $\theta_1, \theta_2, ..., \theta_q$ (com $q \neq 0$) são os parâmetros.

2.2.3 *Modelo autoregressivo de médias móveis (ARMA)*

O modelo ARMA(p,q) é um modelo estatístico que combina os modelos autorregressivos e de médias móveis. Ele é uma combinação linear dos valores passados (p) da série Z_t com o ruído branco (x_t) no tempo presente e nos q tempos passados. Esse modelo é utilizado para representar séries temporais e procura utilizar uma quantidade menor de parâmetros na sua modelagem, tornando a análise mais simples e eficiente (BOX et al., 2016;Morettin & Toloi, 2006). O ARMA é descrito pela equação:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + x_t - \theta_1 x_{t-1} - \theta_2 x_{t-2} - ... - \theta_q x_{t-q} \tag{4}$$

onde os ϕ's são os parâmetros autoregressivos e os θ's são os parâmetros das médias móveis, com $\phi \neq 0$, $\theta \neq 0$ e $\sigma_x^2 > 0$.

2.2.4 *Modelo autoregressivo integrado de médias móveis (ARIMA)*

O modelo ARIMA é uma ampliação do modelo ARMA. Ele é apropriado para situações em que a série temporal não apresenta características de estacionariedade, mas pode ser transformada em uma série estacionária ao se aplicar diferenciações, sendo que no máximo duas diferenciações podem ser realizadas (BOX et al., 2016).

A notação ARIMA(p,d,q) é usada para identificar o modelo, onde "p"representa o número de termos autoregressivos, "d"denota o número de diferenciações aplicadas e "q"representa o número de termos de média móvel. A equação geral para o modelo ARIMA pode ser expressa como:

$$W_t = \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}, \tag{5}$$

onde W_t representa a série diferenciada, ϕ e θ são os parâmetros autoregressivos e de média móvel, respectivamente, e ε_t é o termo de erro.

Esta equação também pode ser reescrita como:

$$\phi(B)((1-B)^d Z_t - \alpha) = \theta(B)\varepsilon_t, \tag{6}$$

onde Z_t representa a série original de dados e B é o operador de diferenciação.

2.2.5 *Modelo autoregressivo de médias móveis sazonal (SARIMA)*

O modelo SARIMA representa uma extensão do modelo ARIMA que tem em consideração a sazonalidade estocástica dos dados (MORETTIN; TOLOI, 2006). O modelo é caracterizado pela notação SARIMA$(p,d,q) \times (P,D,Q)$, em que P representa o número de coeficientes sazonais autoregressivos necessários para obter a estacionariedade, D é o número de diferenças sazonais

que precisam ser tomadas para que a série seja estacionária, com relação ao número de períodos, e Q é o número de coeficientes sazonais de médias móveis que precisam ser considerados para garantir a invertibilidade. A equação geral para o modelo SARIMA pode ser expressa como:

$$\phi(B)\Phi(B)[(1-B)^d(1-B^s)^D - \alpha]Z_t = \theta(B)\Theta(B)\varepsilon_t \tag{7}$$

2.2.6 AutoARIMA

Para evitar a subjetividade e a complexidade teórica necessária para especificar um modelo ARIMA, Morettin e Toloi (2006) recomendam o uso de um método baseado em uma função penalizadora que determina a combinação de p, q, P e Q que minimiza uma função que indica a falta de ajuste do modelo. Os critérios mais utilizados para essa finalidade são o critério de informação de Akaike corrigido (AICc) e o critério de informação bayesiano (BIC).

Quando se analisa uma grande quantidade de séries simultaneamente, acaba não sendo prático, por isso Hyndman e Khandakar (2008) criaram um algoritmo de busca em etapas que reduz a quantidade de combinações a serem testadas. O algoritmo consiste em ajustar quatro modelos iniciais e selecionar o que tem o menor AICc ou BIC. Em seguida, considera-se até treze variações no modelo corrente, sendo selecionado o novo modelo com menor AICc ou BIC. O processo é repetido até que não seja possível encontrar um novo modelo com menor AICc ou BIC. O algoritmo tem restrições para evitar problemas de convergência e garantir o retorno de um modelo final válido.

2.2.7 Função de autocorrelação (FAC)

Segundo Montgomery, Jennings e Kulahci (2015), se uma série temporal for estacionária, onde W_t e $W_t + k$ são duas observações, então a distribuição conjunta de probabilidade é a mesma para qualquer período de tempo t e $t + k$ que sejam separados por um intervalo comum (lag) k. A função de autocorrelação amostral (FAC), rk, no lag k é definida por:

$$r_k = \frac{\sum_{t=k+1}^{n}(W_t - \bar{W})(W_{t-k} - \bar{W})}{\sum_{t=1}^{n}(W_t - \bar{W})^2}, \quad k = 1, 2, ..., \tag{8}$$

onde n é o número de observações e $\bar{W}$ é a média de W.

É importante conhecer a distribuição de probabilidade de r_k para uma análise satisfatória. Se n for suficientemente grande, então r_k segue uma distribuição normal padrão $N(0, \frac{1}{n})$. Com a distribuição de r_k, pode-se criar intervalos de confiança e testar hipóteses para verificar a hipótese nula da correlação (LÚCIA, 2000).

Para testar as hipóteses em que os primeiros coeficientes de autocorrelação k são conjuntamente iguais a zero, utiliza-se a equação:

$$Q_K = n(n+2)\sum_{k=1}^{K}\frac{r_k^2}{n-k}, \tag{9}$$

onde $Q_k \sim \chi^2$ com k graus de liberdade.

2.2.8 Função de autocorrelação parcial (FACP)

De acordo com Shumway e Stoffer (2017), a função de autocorrelação parcial (FACP), representada por ϕ_{kk}, mede a correlação entre W_t e W_{t-k} após remover o efeito das variáveis intervenientes W_{t-1}, W_{t-2}, ..., W_{t-k+1}. A FACP, assim como a função de autocorrelação (FAC),

é utilizada para identificar modelos para séries temporais e auxiliar na identificação de sua estacionariedade.

O modelo de séries temporais pode ser representado de forma geral como:

$$\begin{cases} W_t = \phi_{11} W_{t-1} + \varepsilon_t \\ W_t = \phi_{11} W_{t-1} + \phi_{22} W_{t-2} + \varepsilon_t \\ \quad \vdots \\ W_t = \phi_{k1} W_{t-1} + \phi_{k2} W_{t-2} + \cdots + \phi_{kk} W_{t-k} + \varepsilon_t \end{cases} \tag{10}$$

Uma alternativa para estimar os valores de ϕ_{kk} é o sistema de equações de Yule-Walker:

$$\begin{cases} r_1 = \phi_{11} + \phi_{22} r_1 + \cdots + \phi_{kk} r_{k-1} \\ r_2 = \phi_{11} r_1 + \phi_{22} + \cdots + \phi_{kk} r_{k-2} \\ \quad \vdots \\ r_k = \phi_{11} r_{k-1} + \phi_{22} r_{k-2} + \cdots + \phi_{kk} \end{cases} \tag{11}$$

Assim como a FAC, se n for suficientemente grande, tem-se que $\phi_{kk} \sim N(0, \frac{1}{n})$ para $k > p$. O uso da FACP e FAC possibilita a utilização de intervalos de confiança e testes de hipóteses para verificar a hipótese nula da correlação entre os valores de uma série temporal.

2.3 Etapas do modelo SARIMA

A metodologia Box-Jenkins, segundo Walter et al. (2013), é compreendida como um processo composto de três etapas intercaladas: identificação do modelo, estimação dos parâmetros e verificação.

2.3.1 Identificação

Inicialmente, a tarefa de identificar os valores adequados de p, d, e q é complexa, uma vez que determiná-los de forma equivocada pode prejudicar os resultados obtidos.

O objetivo da identificação do modelo é encontrar aquela que melhor representa a série temporal, sendo essa fase realizada por meio da análise das funções de autocorrelação (FAC) e autocorrelação parcial (FACP) (LIMA et al., 2019).

Para simplificar a tarefa, é recomendável iniciar com a determinação do parâmetro de integração (d) por meio de testes de estacionariedade, visando identificar a ordem de integração.

2.3.2 Teste de estacionariedade de Dickey-Fuller

A partir do modelo $Z_t = \rho Z_{t-1} + \varepsilon_t$, onde ε_t é um ruído branco, a presença de uma raiz unitária na série Z_t impede sua estacionariedade. No entanto, se $|\rho| < 1$, a série Z_t é estacionária, não possuindo raiz unitária. De acordo com Dickey e Fuller (1979), a série Z_t é descrita por três modelos diferentes, em que o processo gerador pode ser expresso como:

1. $\Delta Z_t = \alpha + \beta t + \lambda_3 Z_{t-1} + \varepsilon_t$

2. $\Delta Z_t = \alpha + \lambda_2 Z_{t-1} + \varepsilon_t$

3. $\Delta Z_t = \lambda_1 Z_{t-1} + \varepsilon_t$

onde $\lambda_i = \rho - 1$, para todo $i = 1,2,3$, e α e β são constantes a serem estimadas. A hipótese a ser testada é apresentada na Tabela 1, onde $H_0 : \rho = 1$ equivale a $H_0 : \lambda = 0$. se pelo menos uma das hipóteses não for rejeitada, pode-se concluir que a série possui pelo menos uma raiz unitária e, portanto, não é estacionária. A Tabela 1 apresenta as regras de decisão para o teste da raiz unitária de Dickey-Fuller.

Tabela 1 – Testes de raiz unitária de Dickey-Fuller

Modelos	Hipótese nula (H_0)	Critérios de decisão
1	$\lambda 3 = 0$	$\tau 3 >$ Valor crítico $\Rightarrow$ Aceitar H_0
	$(\alpha, \beta, \lambda 3) = (0,0,0)$	$\delta 2 <$ Valor crítico $\Rightarrow$ Aceitar H_0
	$(\alpha, \beta, \lambda 3) = (\alpha, 0, 0)$	$\delta 3 <$ Valor crítico $\Rightarrow$ Aceitar H_0
2	$\lambda 2 = 0$	$\tau 2 >$ Valor crítico $\Rightarrow$ Aceitar H_0
	$(\alpha, \lambda 2) = (0,0)$	$\delta 1 <$ Valor crítico $\Rightarrow$ Aceitar H_0
3	$\lambda 1 = 0$	$\tau 1 >$ Valor crítico $\Rightarrow$ Aceitar H_0

Fonte: Adaptado de Dickey e Fuller (1979)

As estatísticas presentes na Tabela 1, τ_1, τ_2, τ_3, δ_1, δ_2, e δ_3, são calculadas a partir das equações a seguir:

$$\tau_3 = \frac{\lambda_3}{\alpha \lambda_3} \tag{12}$$

$$\delta_3 = \frac{\text{SQR}(1) - \text{SQR}(1)}{3\text{SQR}(1)} \cdot \frac{1}{n} \tag{13}$$

$$\tau_2 = \frac{\lambda_2}{\alpha \lambda_2} \tag{14}$$

$$\delta_2 = \frac{\text{SQR}(2) - \text{SQR}(2)}{3\text{SQR}(2)} \cdot \frac{1}{n} \tag{15}$$

$$\tau_1 = \frac{\lambda_1}{\alpha \lambda_1} \tag{16}$$

$$\delta_1 = \frac{\text{SQR}(3) - \text{SQR}(3)}{3\text{SQR}(3)} \cdot \frac{1}{n} \tag{17}$$

onde,

1. $\text{SQR}(1)$ É a soma dos quadrados dos resíduos do modelo $\Delta Z_t = \alpha + \beta t + \lambda_3 Z_{t-1} + \varepsilon_t$;
2. $\text{SQR}(2)$ É a soma dos quadrados dos resíduos do modelo $\Delta Z_t = \alpha + \lambda_2 Z_{t-1} + \varepsilon_t$;
3. $\text{SQR}(3)$ É a soma dos quadrados dos resíduos do modelo $\Delta Z_t = \lambda_1 Z_{t-1} + \varepsilon_t$;
4. σ_{λ_1}, σ_{λ_2}, e σ_{λ_3} são as variâncias de λ_1, λ_2, e λ_3, respectivamente.

2.3.3 Análise gráfica

Após a definição da série estacionária e do seu parâmetro d, o próximo passo é identificar os parâmetros p e q. De acordo com Morettin e Toloi (2006), a escolha dos parâmetros é baseada na análise gráfica das funções de autocorrelação (FAC) e autocorrelação parcial (FACP) dos modelos ARMA. A Tabela 2 apresenta uma síntese do comportamento das FAC e FACP para os modelos ARMA.

Tabela 2 – Comportamento da função de autocorrelação e autocorrelação parcial para modelos ARIMA

	Função de Autocorrelação (FAC)	Autocorrelação Parcial (FACP)
Modelo AR(p)	Decaimento assintótico	Decaimento abrupto
Modelo MA(q)	Decaimento abrupto	Decaimento abrupto
Modelo ARMA(p,q)	Decaimento assintótico	Decaimento abrupto

Fonte: Adaptada de Shumway e Stoffer (2017)

2.3.4 Estimação

Depois de encontrar o modelo preliminar para a série, o próximo passo é calcular seus parâmetros, que envolvem termos de autorregressão e termos de média móvel. Para isso, é possível utilizar o procedimento interativo de Mínimos Quadrados Condicionais ou o Método de Máxima Verossimilhança (MORETTIN; TOLOI, 2006).

Neste estudo, optou-se por utilizar o Método de Máxima Verossimilhança por ele maximizar a verossimilhança de um conjunto de observações, além do software usado ser bem equipado para cálculos das estimativas dos parâmetros.

A função de máxima verossimilhança utilizada é dada por Brockwell et al. (2016) como:

$$L(\phi_i, \theta_j, \sigma_\xi^2 / w) = \left(2\pi\sigma_\xi^2\right)^{-\frac{n}{2}} |\sum|^{-\frac{n}{2}} \exp\left(\frac{-\frac{1}{2}x'\sum^{-1} w}{\sigma_\xi^2}\right) \tag{18}$$

Ao maximizar esta função, são obtidas as estimativas dos parâmetros ϕi, θj e σ_ξ^2.

2.3.5 Validação

Após a estimação do modelo, é importante verificar sua adequação através de um processo denominado diagnóstico. De acordo com LIMA et al. (2019), esse processo consiste em duas etapas principais: a análise dos resíduos e a avaliação do modelo.

A análise dos resíduos (ε_t) é realizada com o objetivo de verificar se o comportamento dos resíduos é de um ruído branco, ou seja, se os resíduos são independentes entre si e possuem a mesma variância. Caso contrário, o processo deve ser reiniciado, identificando um novo modelo e repetindo as etapas até que um modelo adequado seja encontrado. Para avaliar essa condição, utiliza-se o teste de Ljung-Box, cujo objetivo é testar se os coeficientes de autocorrelação residual são estatisticamente iguais a zero.

Quanto à avaliação do modelo, alguns modelos podem apresentar poucos parâmetros (modelos parcimoniosos), que podem ser correlacionados ou não-significativos. Não há uma solução específica para determinar quantos parâmetros são necessários para o modelo, mas é possível utilizar softwares estatísticos que estimam diferentes combinações de p e q, descartando

as não-significativas. Alguns critérios comuns utilizados para avaliar a adequação do modelo incluem:

FPE (Erro Padrão Final):

$$FPE = \hat{\sigma}^2 x \frac{M+p}{M-p} \tag{19}$$

AIC (Critério de Informação de Akaike):

$$AIC = 2\log\hat{L} + 2(p+q), \tag{20}$$

onde $\hat{L}$ é o valor máximo da verossimilhança.

AICC (Critério de Informação de Akaike Corrigido):

$$AICC = -2\log\hat{L} + \frac{2(p+q)M}{M-p-q-1} \tag{21}$$

BIC (Critério de Informação Bayesiano):

$$BIC = -2\log\hat{L} + (p+q)\log M \tag{22}$$

Após a seleção e validação do modelo adequado, é possível prever os valores futuros da série temporal modelada.

2.4 Prophet

Prophet é uma ferramenta oferecida pelo *Facebook* para automatizar a previsão de séries temporais. Ele pode ser usado tanto no R quanto no Python e é capaz de trabalhar com dados em diferentes níveis de detalhamento, incluindo dados de minuto a minuto ou por hora e também dados diários. Além disso, é capaz de compreender tendências que crescem de forma não linear, como ao alcançar um ponto de estabilidade (NIELSON, 2021).

O *Prophet* inclui componentes de efeitos sazonais anuais e semanais, uma lista de feriados e uma curva de tendência linear. A fórmula do Prophet é dada por:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t, \tag{23}$$

em que, $y(t)$ é o valor observado na série temporal no tempo t, $g(t)$ é o componente tendencial, $s(t)$ é o componente sazonal, $h(t)$ é o componente de feriados (fornecidos pelo usuário) e ε_t é o erro de previsão.

Segundo informações retiradas no site da B3, o índice bovespa não opera em feriados nacionais como o finados, proclamação da república, entre outros. Com essa informação, não foi inserido uma lista de feriados no *prophet*. No entanto, é importante manter-se atualizado sobre qualquer mudança na política de operação do índice Bovespa em relação aos feriados e reavaliar a necessidade de incluir uma lista de feriados no futuro.

De acordo com Taylor e Letham (2017), o *Prophet* torna muito mais fácil criar previsões razoáveis e precisas. O pacote de previsão inclui muitas técnicas diferentes de previsão (ARIMA, suavização exponencial, etc.), cada uma com suas próprias forças, fraquezas e parâmetros de ajuste. Eles descobriram que escolher o modelo ou os parâmetros errados pode frequentemente produzir resultados pobres e até mesmo analistas experientes podem ter dificuldade em escolher o modelo e os parâmetros corretos de maneira eficiente, dada essa variedade de escolhas.

Ainda de acordo com Taylor e Letham (2017), as previsões do *Prophet* são personalizáveis de maneiras que são intuitivas para não especialistas. Existem parâmetros de suavização para

sazonalidade que permitem ajustar quão bem se encaixam os ciclos históricos, bem como parâmetros de suavização para tendências que permitem ajustar quão agressivamente seguir as mudanças de tendência históricas. Para curvas de crescimento, é possível especificar manualmente as "capacidades"ou o limite superior da curva de crescimento, permitindo inserir informações prévias sobre como a previsão irá crescer (ou declinar). Por fim, é possível especificar feriados irregulares a serem modelados.

2.5 Métricas de avaliação

Para avaliar a qualidade das previsões do índice Ibovespa, foram utilizadas seis métricas de erro: RMSE, MAE, MAPE, SMAPE, MASE e RSQ.

RMSE (Root Mean Squared Error) é dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}, \tag{24}$$

onde n é o número de previsões, $\hat{y}t$ é a previsão para o período t e yt é o valor real do índice no período t.

MAE (Mean Absolute Error) é dado por:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \tag{25}$$

MAPE (Mean Absolute Percentage Error) é dado por:

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{\hat{y}_t - y_t}{y_t}\right| \tag{26}$$

SMAPE (Symmetric Mean Absolute Percentage Error) é dado por:

$$SMAPE = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{(|\hat{y}_t| + |y_t|)/2} \tag{27}$$

MASE (Mean Absolute Scaled Error) é dado por:

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{\frac{1}{h-1}\sum_{i=2}^{h}|y_i - y_{i-1}|}, \tag{28}$$

onde h é o número de períodos de atraso na série de tempo.

RSQ (R Squared) é dado por:

$$RSQ = 1 - \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}, \tag{29}$$

onde $\bar{y}$ é a média dos valores observados na série de tempo.

3 RESULTADOS

3.1 Análise exploratória dos dados

A Figura 1 é utilizada para ilustrar o comportamento do fechamento diário do Índice Bovespa durante o período analisado.

Figura 1 – Comportamento do Fechamento Diário do Índice Bovespa por Gráfico de Linha

Na Figura 1 é possível observar que no início de 2020 houve uma grande mudança de nível devido à pandemia (sars-covid19). Após esse período, o mercado começou a se recuperar, apresentando uma tendência de alta. A partir de junho-julho de 2021, o índice começou a apresentar estabilidade, com variações ocorrendo em um intervalo situado entre aproximadamente 100.000 e 120.000 pontos.

A Figura 2 e Figura 3 será utilizada para extrair informações, como a estacionariedade dos dados e a identificação de padrões de autocorrelação.

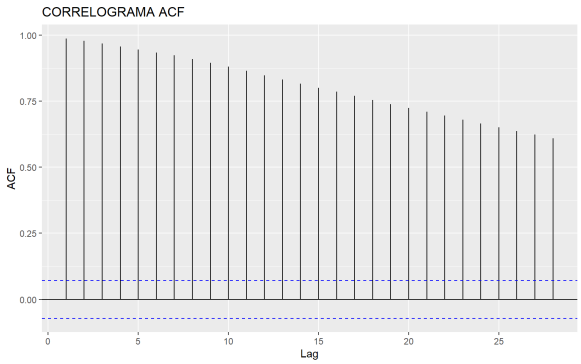


Figura 2 – FAC

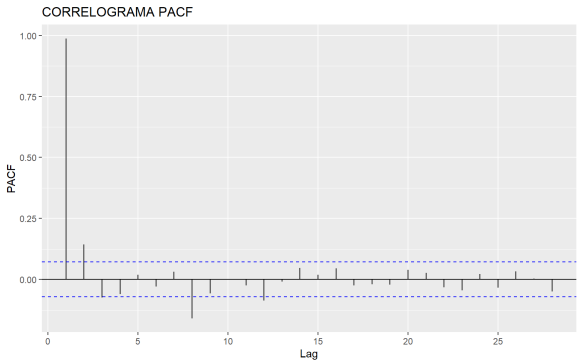


Figura 3 – FACP

Perante a análise do gráfico da FAC (Figura 2) observa-se um decaimento lento dos valores de autocorrelação conforme os lags aumentam (forte indicação de uma não estacionariedade).

O FACP (Figura 3) revela que os dois primeiros atrasos apresentam correlação significativa com o ponto atual, enquanto os demais atrasos não apresentam uma correlação forte. Isso sugere um modelo com poucos parâmetros.

A Figura 4 refere-se a um histograma de densidade, uma ferramenta estatística que permite visualizar a distribuição de dados.

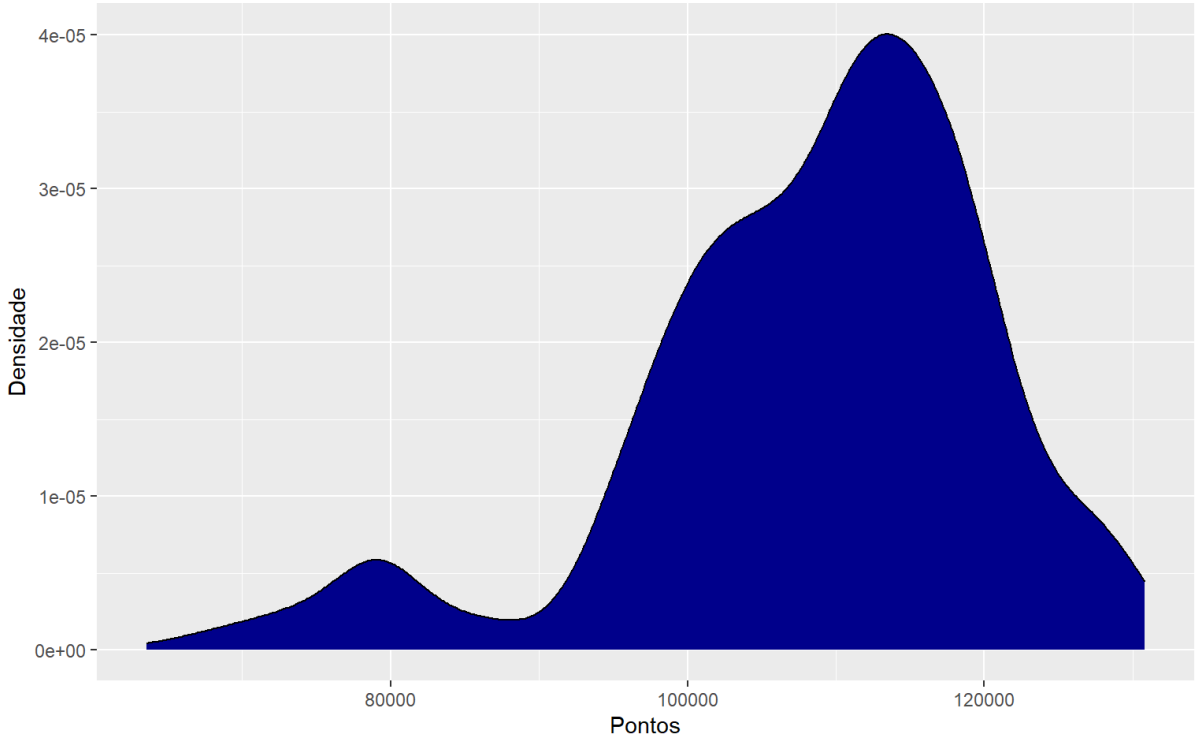


Figura 4 – Histograma de densidade

Ao realizar a análise da Figura 4, é possível observar uma assimetria para a direita nos dados, indicando que os dados podem não seguir uma distribuição normal. Tal suposição pode ser confirmada por meio do teste de Shapiro-Wilk ($W = 0{,}93588$; valor-p $< 0{,}001$).

A Figura 5 apresenta a decomposição dos dados. Através desta técnica é possível separar a série em suas componentes, como tendência, sazonalidade e resíduos, permitindo assim uma melhor compreensão dos padrões e comportamentos apresentados pelos dados ao longo do tempo.

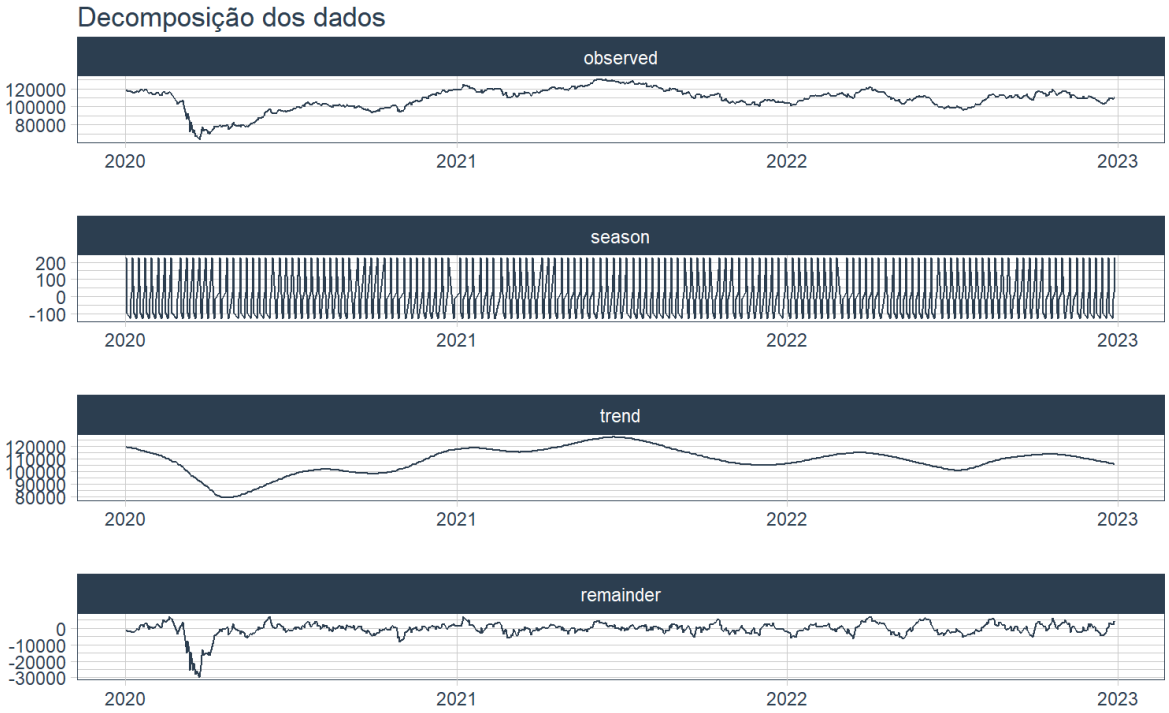


Figura 5 – Decomposição dos dados

Através da decomposição (Figura 5), foi possível identificar a presença de sazonalidade e tendência. A sazonalidade foi observada através dos padrões regulares presentes na série, enquanto a tendência foi representada através de um gráfico de linha, mostrando como a série se comportou ao longo do tempo. Além disso, a análise dos resíduos permitiu detectar variações significativas na série temporal, como a queda brusca no início de 2020, relacionada a pandemia, permitindo assim avaliar a magnitude dessa diferença.

Na Figura 6, Figura 7, Figura 8 e Figura 9, são apresentados boxplots referentes a diferentes períodos de tempo.

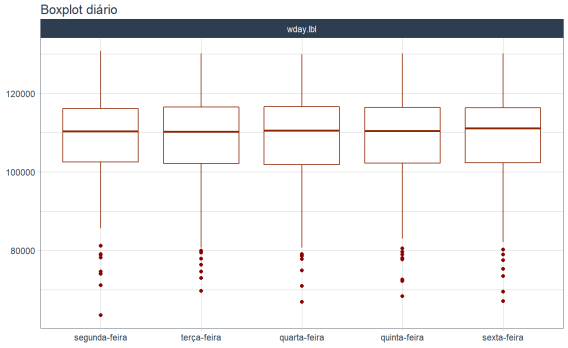


Figura 6 – Boxplot diário

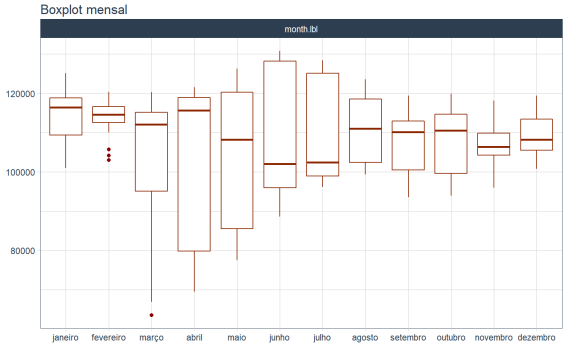


Figura 7 – Boxplot mensal

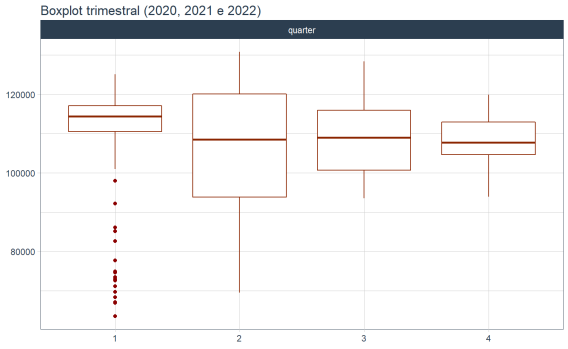


Figura 8 – Boxplot trimestral

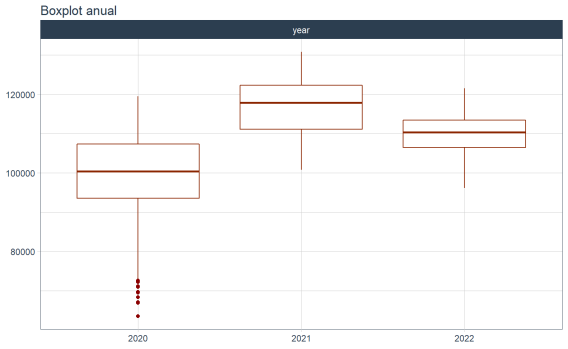


Figura 9 – Boxplot anual

Ao analisar os dados diários (Figura 6), é possível observar a presença de outliers e uma variação pequena nas medianas. Apesar da possível existência de sazonalidade nos dados, esta não foi detectada visualmente por meio do uso de boxplots.

Ao analisar os dados mensais (Figura 7), constata-se que os meses iniciais do ano, janeiro, fevereiro, março e abril apresentam as melhores medianas entre todos os meses. Entretanto, é possível observar um declínio logo após esses meses. Já os meses de junho e julho apresentam as piores medianas entre todos os meses analisados, o que sugere uma tendência de queda econômica nestes períodos. Por outro lado, nos meses de agosto até dezembro é observado um equilíbrio, com o preço ficando cada vez mais estável.

Ao analisar o conjunto dos trimestres de todos os anos juntos (2020, 2021 e 2022) (Figura 8), verifica-se que:

No primeiro trimestre, apesar de apresentar a melhor mediana entre os trimestres analisados, há uma grande presença de valores atípicos (outliers) na região inferior do boxplot. Isso pode ser atribuído à crise econômica causada pela pandemia em 2020.

Já no segundo trimestre, embora apresente um boxplot mais disperso, esse trimestre tem a segunda melhor mediana entre os trimestres.

Por sua vez, o terceiro trimestre apresenta um comportamento adequado quando comparado com os outros trimestres, mas tem a pior mediana entre eles.

Finalmente, o quarto trimestre apresenta um comportamento mais estável, sem variações

significativas, e tem a terceira pior mediana entre os trimestres.

Ao analisar os dados dos anos de 2020, 2021 e 2022 (Figura 9), pode-se observar que em 2020 houve uma maior presença de valores atípicos na região inferior do boxplot (talvez devido à pandemia), indicando um desempenho econômico desfavorável quando comparado aos outros anos, com uma mediana de cerca de 100.000 pontos. Já em 2021, houve uma recuperação e não foram observados valores atípicos, apresentando valores mais elevados do que o ano anterior, com uma mediana próxima de 120.000 pontos. Por fim, em 2022, embora tenha tido um desempenho inferior ao de 2021, não foram observados valores atípicos e uma mediana de cerca de 110.000 pontos. Em resumo, os dados mostram uma recuperação ao longo do tempo após a queda brusca do ano de 2020.

3.2 Transformações dos dados

3.2.1 Log1p e uma diferenciação

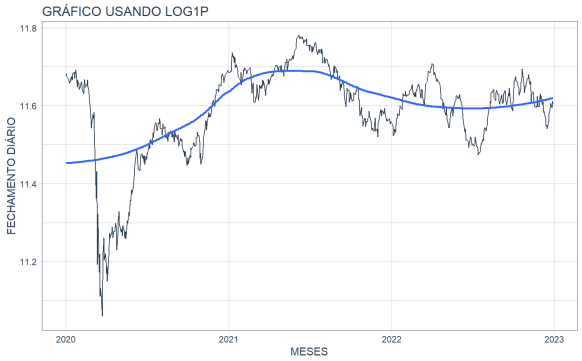


Figura 10 – Gráfico usando transformação Log1p

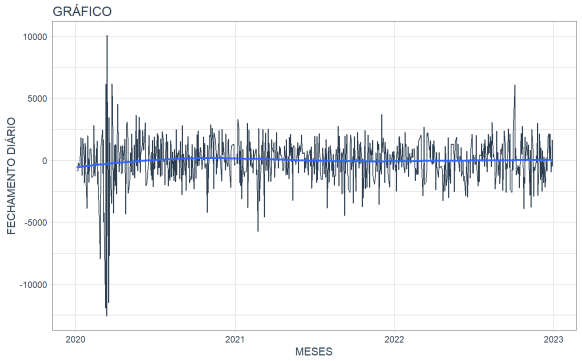


Figura 11 – Gráfico usando uma diferenciação

Devido aos valores atípicos (outliers) encontrados no banco de dados, foi utilizada também uma outra transformação, chamada log1p, que preserva a diferença, já que é importante, mas a reduz consideravelmente (Figura 10). Denota-se $\log(1+x)$ por log1p de acordo com o padrão de nomenclatura em computação científica (LIU, 1988).

Para tornar a série temporal estacionária (Figura 11), aplicou-se uma diferenciação. Após essa transformação, a análise visual dos dados sugere que eles se tornaram mais estacionários, o que é uma condição importante para usar modelos de previsão como o ARIMA. No entanto, é importante realizar outros testes estatísticos para ter certeza da estabilidade dos dados, pois a análise visual pode não ser suficiente.

No entanto, realizou-se o teste de estacionariedade (Tabela 3) para verificar a presença ou não de estacionariedade nos dados.

Tipo	lag	ADF	p-valor
1: sem desvio e sem tendência	0	-31,77	0,01
	1	-19,14	0,01
	2	-14,73	0,01
	3	-13,26	0,01
	4	-11,59	0,01
	5	-11,04	0,01
	6	-8,74	0,01
2: com desvio e sem tendência	0	-31,75	0,01
	1	-19,13	0,01
	2	-14,72	0,01
	3	-13,25	0,01
	4	-11,58	0,01
	5	-11,03	0,01
	6	-8,73	0,01
3: com desvio e com tendência	0	-31,74	0,01
	1	-19,12	0,01
	2	-14,72	0,01
	3	-13,25	0,01
	4	-11,58	0,01
	5	-11,03	0,01
	6	-8,73	0,01
Note: na verdade, p-valor = 0,01 significa que p-valor <= 0,01			

Tabela 3 – Teste ADF

Os testes ADF (Tabela 3) mostram que a série se torna estacionária quando se aplica uma diferenciação (Figura 11), pois os valores do p-valor são inferiores ao nível de significância de 0,01 em todos os tipos de hipóteses testadas (sem desvio e sem tendência; com desvio e sem tendência; e com desvio e com tendência). Isso significa que é menos provável que a rejeição da hipótese nula de não estacionariedade seja um erro.

3.3 Prophet e AutoARIMA - treino e teste

Foi adotado o procedimento de dividir os dados em proporções de 70/30 e 80/20, respectivamente para treino e teste, com o propósito de avaliar a habilidade dos modelos preditivos em generalizar a partir dos dados de treino para os de teste, a fim de determinar quais deles apresentam melhor desempenho.

3.3.1 Prophet

Com o objetivo de avaliar o desempenho do modelo Prophet, foram realizados testes utilizando diferentes proporções de dados de treinamento e teste, além de diferentes transformações. Os resultados foram apresentados na Figura 12, Figura 13, Figura 14, Figura 15, Figura 16 e Figura 17.

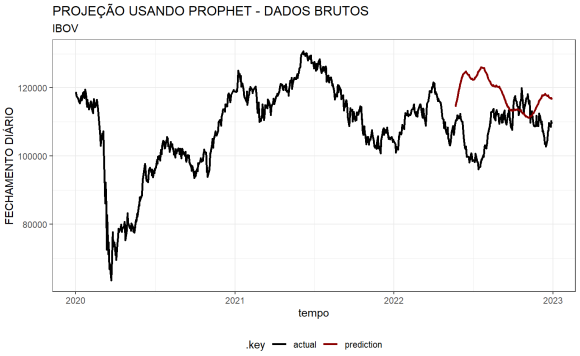


Figura 12 – Dados originais (70/30)

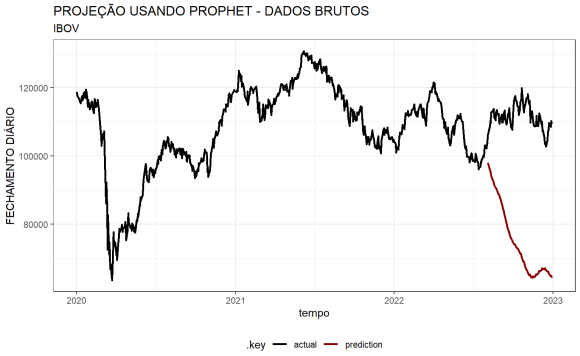


Figura 13 – Dados originais (80/20)

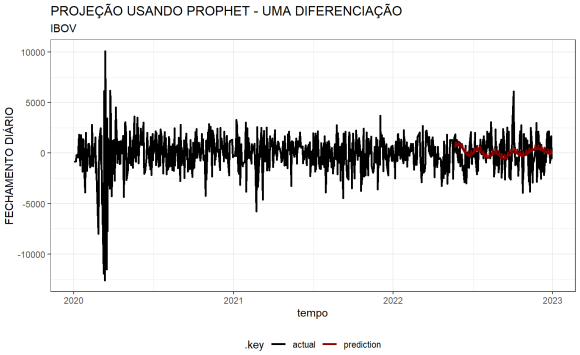


Figura 14 – Uma diferenciação (70/30)

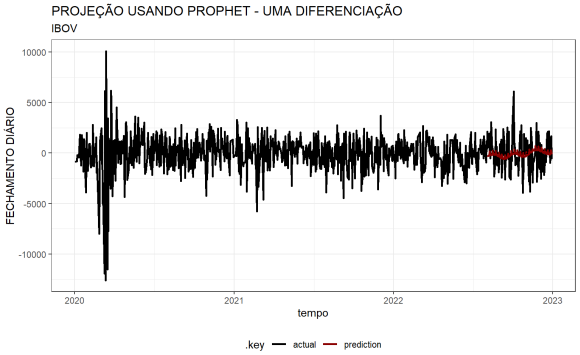


Figura 15 – Uma diferenciação (80/20)

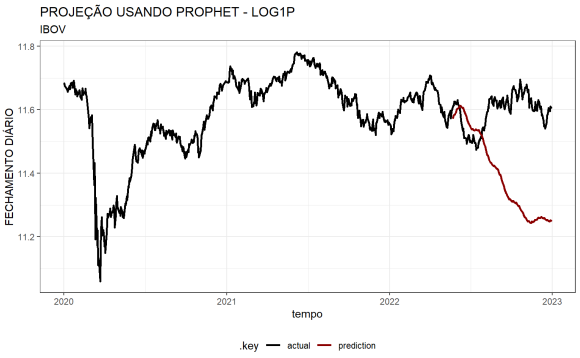


Figura 16 – Log1p (70/30)

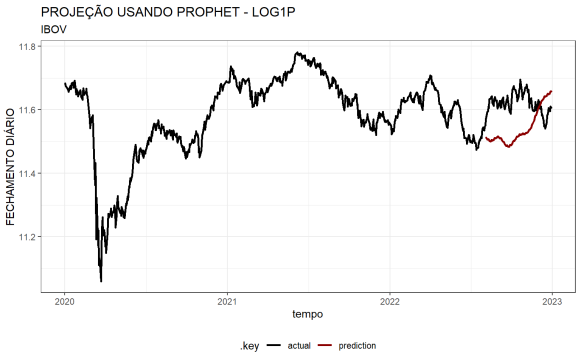


Figura 17 – Log1p (80/20)

Observa-se que, nos dados originais, a divisão de treinamento e teste na proporção de 70/30 (Figura 12) apresentou resultados superiores em relação à proporção de 80/20 (Figura 13). Em relação aos dados com diferenciação, não foi visualmente percebida uma diferença significativa entre a proporção de 70/30 (Figura 14) e a de 80/20 (Figura 15). Já para os dados com transformação log1p, a proporção de 70/30 (Figura 16) apresentou uma pior performance ao da proporção de 80/20 (Figura 17).

As métricas de erro serão apresentadas a seguir (Figura 18, Figura 19 e Figura 20).

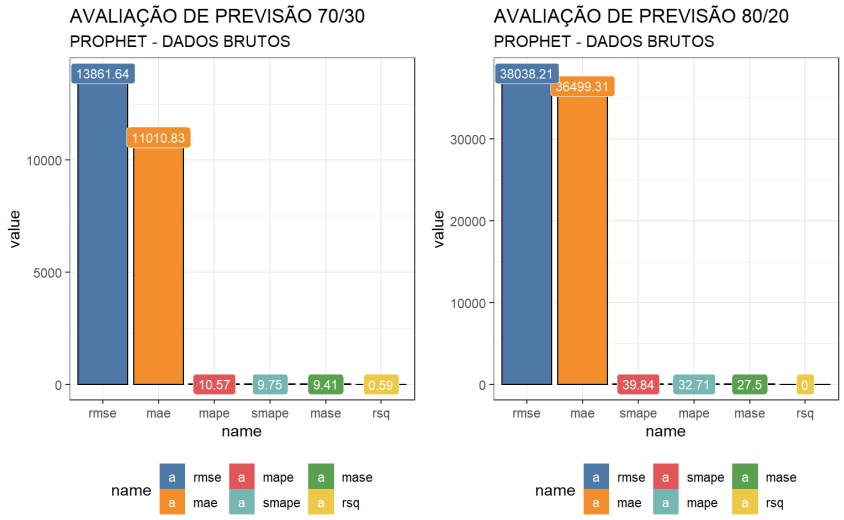


Figura 18 – Erros da predição dos dados originais

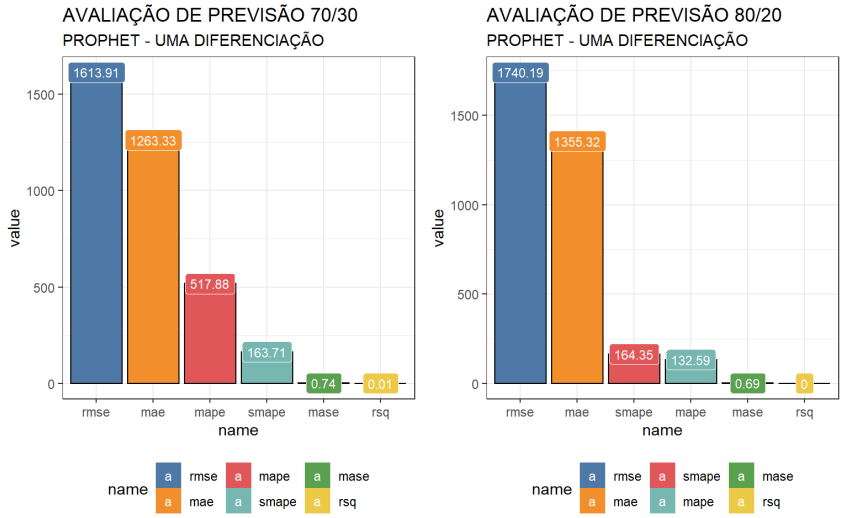


Figura 19 – Erros da predição de uma diferenciação

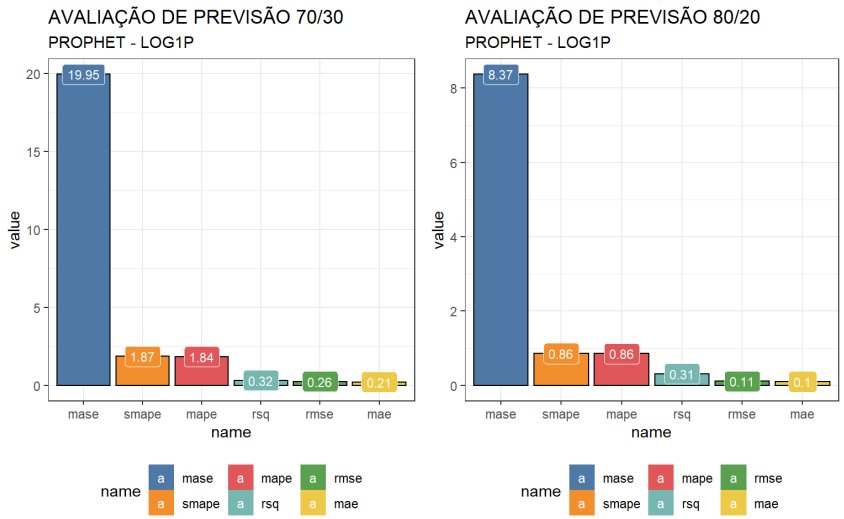


Figura 20 – Erros da predição de log1p

Após a realização de testes com diferentes divisões de treino/teste (80/20 e 70/30), se observa que o modelo Prophet que aprensentou as menores métricas de erro ao ser aplicado aos dados originais foi a divisão 70/30 para treino e teste (Figura 18), sendo, portanto, a opção mais indicada para análise do conjunto de dados em questão.

O modelo Prophet que utiliza uma diferenciação e particionamento de 70/30 para treino e teste (Figura 19) apresentou os menores erros de RMSE (Raiz do Erro Quadrático Médio) e MAE (Erro Absoluto Médio), tornando-se a melhor opção para previsões em séries temporais. No entanto, em relação aos erros de MAPE (Erro Percentual Absoluto Médio), SMAPE (Erro Percentual Absoluto Médio Simétrico), MASE (Erro Percentual Absoluto Médio Sazonal) e RSQ (Coeficiente de Determinação), o modelo não apresentou os melhores resultados.

para o modelo Prophet com transformação log1p, observou-se que a estratégia que obteve as menores métricas de erro foi a divisão de 80/20 (Figura 20). Dessa forma, conclui-se que a utilização desse modelo com essa estratégia apresenta uma opção mais adequada para a análise dos dados em questão.

Pode-se concluir que, em relação ao modelo Prophet, as configurações que apresentaram os menores erros foram: divisão 70/30 para os dados originais e com uma diferenciação, e divisão 80/20 para log1p.

3.3.2 AutoARIMA

Assim como feito com o Prophet, serão realizados testes com o AutoARIMA utilizando diferentes proporções de dados de treinamento e teste, bem como diferentes transformações (Figura 21, Figura 22, Figura 23, Figura 24, Figura 25 e Figura 26).

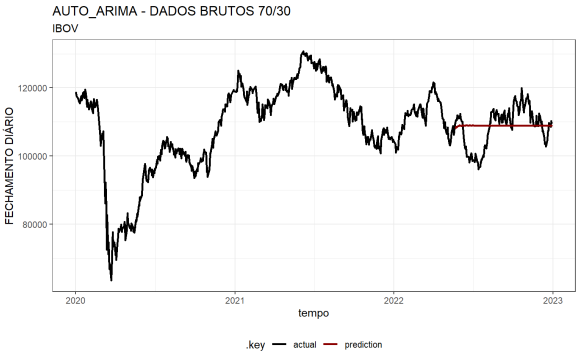


Figura 21 – Dados originais (70/30)

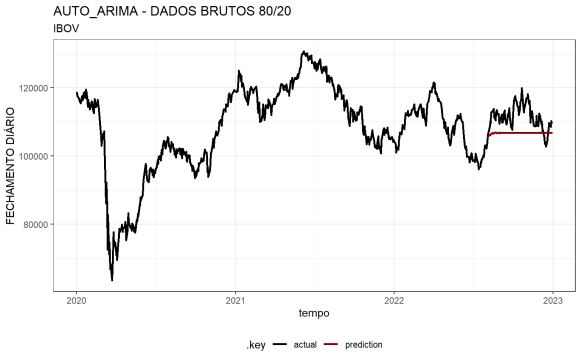


Figura 22 – Dados originais (80/20)

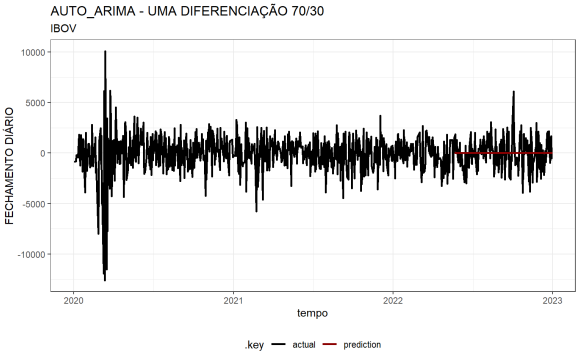


Figura 23 – Uma diferenciação (70/30)

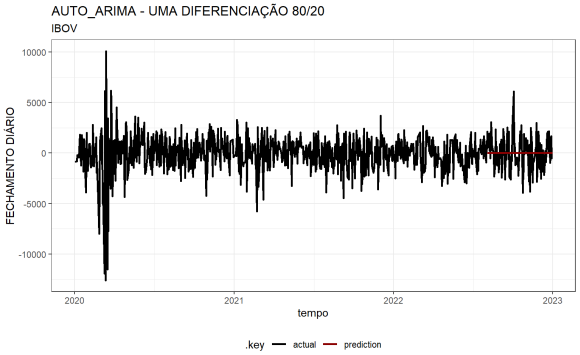


Figura 24 – Uma diferenciação (80/20)

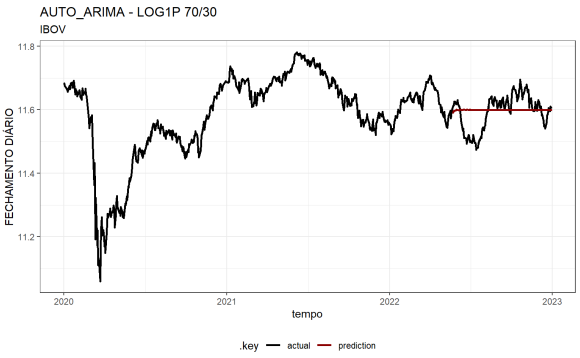


Figura 25 – Log1p (70/30)

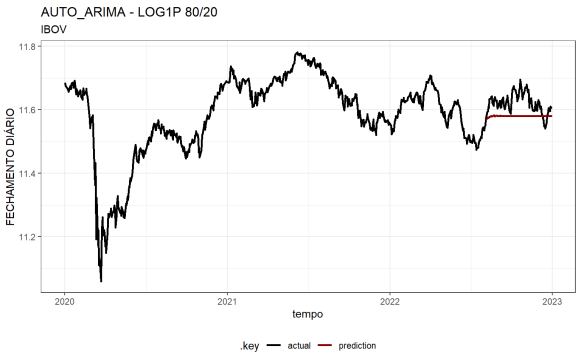


Figura 26 – Log1p (80/20)

Devido à natureza do modelo autoARIMA, é difícil determinar visualmente a melhor opção, tornando necessária a análise das métricas de erro (Figura 27, Figura 28 e Figura 29) para a escolha do modelo mais adequado.

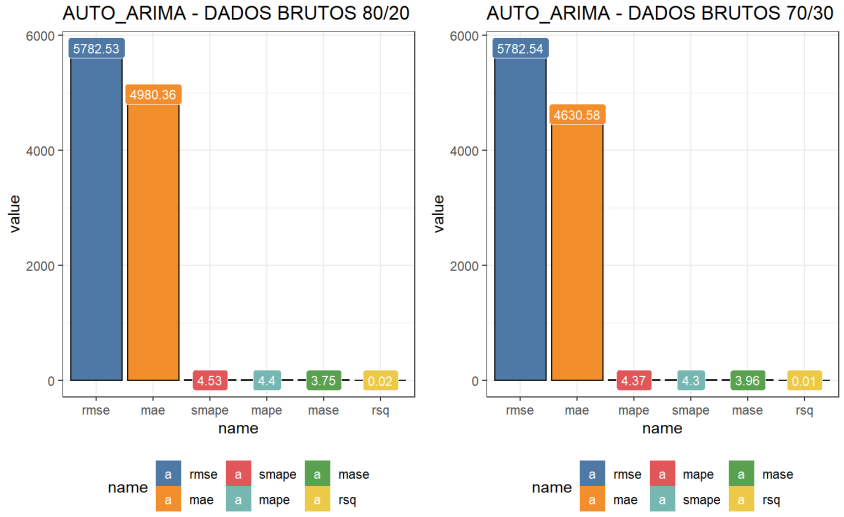


Figura 27 – Erros da predição dos dados originais

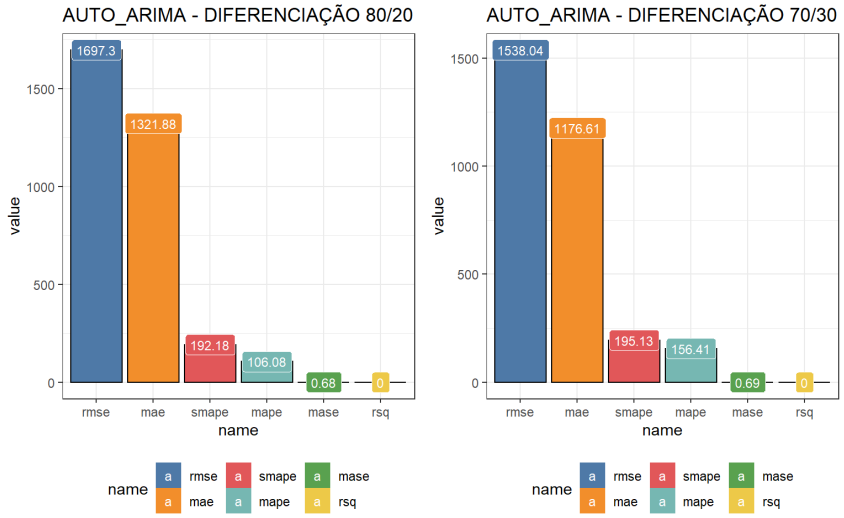


Figura 28 – Erros da predição de uma diferenciação

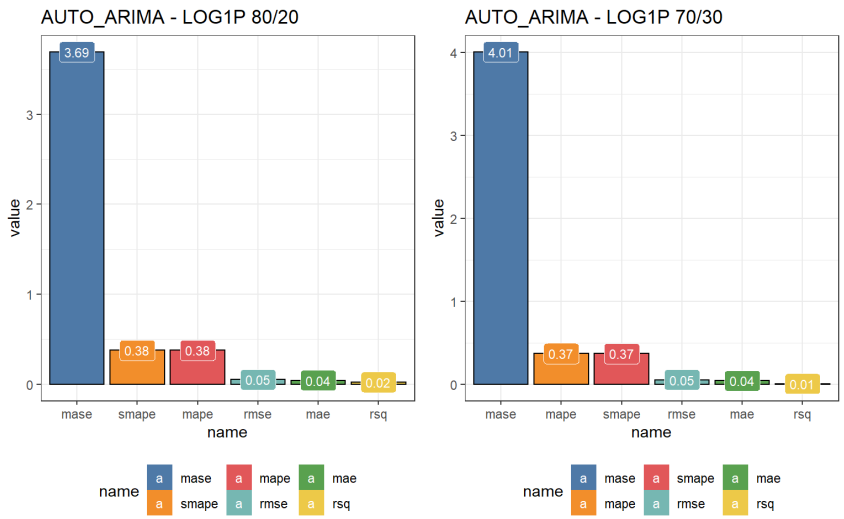


Figura 29 – Erros da predição de log1p

Uma análise das métricas dos modelos de previsão autoARIMA mostrou que as melhores performances foram alcançadas para os conjuntos de dados originais com proporção de divisão de treino e teste de 70/30 (Figura 27), bem como para com uma diferenciações com a mesma proporção (Figura 28). Já para a transformação log1p, a proporção 80/20 foi a que apresentou melhores resultados (Figura 29). É importante destacar que os resultados das melhores performances obtidos pelo autoARIMA foram semelhantes aos resultados encontrados pelo modelo Prophet.

3.4 Predição

Uma previsão de 15 dias será realizada utilizando os dados que apresentaram as melhores métricas de erro. É importante destacar que esses dados foram convertidos de volta para sua forma original, a fim de garantir a precisão e a compreensibilidade da visualização dos resultados finais. Além disso, o resultado será comparado com os valores reais para obter o erro de previsão. É importante ressaltar que, ao fazer a previsão para um número menor de dias, como 7 dias, pode haver o problema de a previsão ser semelhante à feita para 15 dias, mas reduzida para apenas 7 dias. Portanto, não é vantajoso inserir previsões para períodos menores que 15 dias, pois apresenta erros similares de previsão.

3.4.1 Dados originais

Realizando uma previsão de 15 dias com base nos dados originais divididos em proporções de 70/30 usando os modelos Prophet e AutoARIMA. As predições resultantes podem ser vistas na Figura 30 e Figura 31.

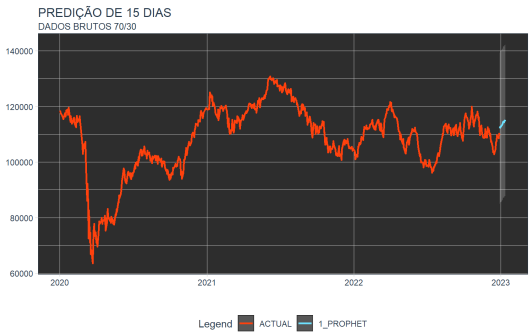


Figura 30 – Prophet predição dos dados originais 70/30

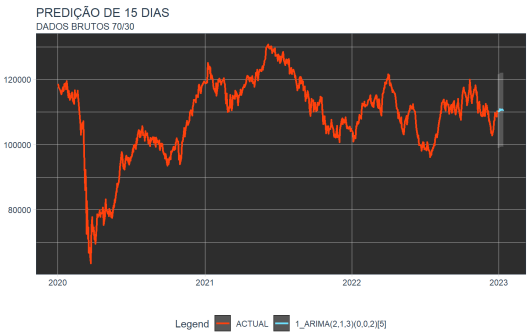


Figura 31 – AutoARIMA predição dos dados originais 70/30

A Tabela 4 e Tabela 5 mostra o erro percentual de previsão de dois modelos, o Prophet e o AutoARIMA, para um período de 15 dias com base nos dados originais divididos em proporções de 70/30.

Data	Previsto	Real	Erro de Previsão (%)
30/12/2022	112.106,20	-	-
31/12/2022	112.707,60	-	-
01/01/2023	112.919,20	-	-
02/01/2023	112.754,40	106.376,00	5,99
03/01/2023	113.092,20	104.166,00	8,56
04/01/2023	113.327,60	105.334,00	7,57
05/01/2023	113.475,00	107.518,00	5,55
06/01/2023	113.664,10	108.836,00	4,42
07/01/2023	114.295,90	-	-
08/01/2023	114.523,30	-	-
09/01/2023	114.360,00	109.227,00	4,71
10/01/2023	114.685,40	110.912,00	3,39
11/01/2023	114.895,80	111.763,00	2,80
12/01/2023	115.006,60	111.877,00	2,79
13/01/2023	115.149,00	111.036,00	3,70

Tabela 4 – Prophet (dados originais) - erro de previsão (%) de 15 dias

Data	Previsto	Real	Erro de Previsão (%)
30/12/2022	110.213,30	-	-
31/12/2022	110.262,80	-	-
01/01/2023	110.202,40	-	-
02/01/2023	110.325,40	106.376,00	3,72
03/01/2023	110.447,10	104.166,00	6,03
04/01/2023	110.852,50	105.334,00	5,00
05/01/2023	110.725,60	107.518,00	2,66
06/01/2023	110.540,70	108.836,00	1,56
07/01/2023	110.618,10	-	-
08/01/2023	110.656,00	-	-
09/01/2023	110.800,60	109.227,00	1,44
10/01/2023	110.785,70	110.912,00	-1,11
11/01/2023	110.669,30	111.763,00	-0,98
12/01/2023	110.613,20	111.877,00	-1,13
13/01/2023	110.673,00	111.036,00	-0,32

Tabela 5 – AutoARIMA (dados originais) - erro de previsão (%) de 15 dias

Ao se comparar a tabela de erros de previsão dos dados originais presente na Tabela 4 com aquela encontrada na Tabela 5, é possível constatar que o modelo AutoARIMA, cujo desempenho é ilustrado na Figura 31, obteve um resultado preditivo superior ao modelo Prophet (Figura 30).

3.4.2 Com uma diferenciação

Com base nos dados diferenciados, foram realizadas previsões de 15 dias usando os modelos Prophet e AutoARIMA, com a divisão dos dados em proporções de 70/30. As predições resultantes podem ser visualizadas na Figura 32 e Figura 33.

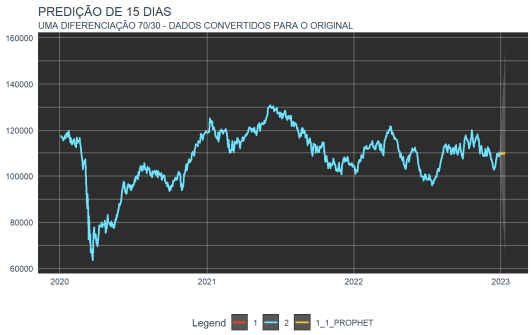


Figura 32 – Prophet predição dos dados com uma diferenciação 70/30

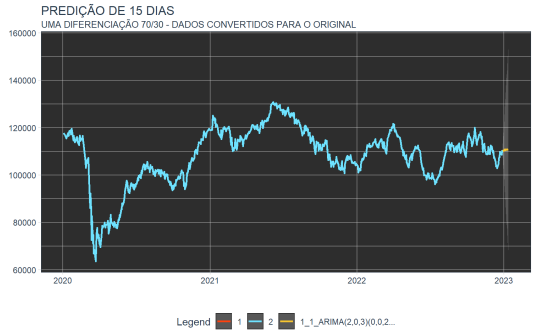


Figura 33 – AutoARIMA predição dos dados com uma diferenciação 70/30

A Tabela 6 e Tabela 7 apresenta a porcentagem de erro de previsão de dois modelos, o Prophet e o AutoARIMA, para um período de 15 dias. Essa análise foi realizada com base nos dados diferenciados, que foram divididos em proporções de 70/30.

Data	**Previsto**	**Real**	**Erro de Previsão (%)**
30/12/2022	109.824,90	-	-
31/12/2022	109.569,90	-	-
01/01/2023	109.315,50	-	-
02/01/2023	109.397,80	106.376,00	2,84
03/01/2023	109.816,80	104.166,00	5,44
04/01/2023	110.033,70	105.334,00	4,47
05/01/2023	110.152,90	107.518,00	2,46
06/01/2023	110.217,10	108.836,00	1,27
07/01/2023	109.922,20	-	-
08/01/2023	109.614,70	-	-
09/01/2023	109.631,80	109.227,00	0,37
10/01/2023	109.975,20	110.912,00	-0,85
11/01/2023	110.107,90	111.763,00	-1,48
12/01/2023	110.136,40	111.877,00	-1,56
13/01/2023	110.105,80	111.036,00	-0,84

Tabela 6 – Prophet (uma diferenciação) - erro de previsão (%) de 15 dias

Data	Previsto	Real	Erro de Previsão (%)
30/12/2022	110.213,3	-	-
31/12/2022	110.262,8	-	-
01/01/2023	110.202,4	-	-
02/01/2023	110.325,4	106.376,00	3,72
03/01/2023	110.447,1	104.166,00	6,05
04/01/2023	110.852,5	105.334,00	5,23
05/01/2023	110.725,6	107.518,00	2,89
06/01/2023	110.540,7	108.836,00	1,57
07/01/2023	110.618,1	-	-
08/01/2023	110.656,0	-	-
09/01/2023	110.800,6	109.227,00	1,44
10/01/2023	110.785,7	110.912,00	-1,02
11/01/2023	110.669,3	111.763,00	-0,98
12/01/2023	110.613,2	111.877,00	-1,12
13/01/2023	110.673,0	111.036,00	-3,27

Tabela 7 – AutoARIMA (uma diferenciação) - erro de previsão (%) de 15 dias

Ao comparar a tabela de erros de previsão com uma diferenciação presente na Tabela 6 àquela encontrada na Tabela 7, verifica-se que o modelo Prophet, cujo desempenho é ilustrado na Figura 32, apresentou um desempenho preditivo superior ao modelo AutoARIMA (Figura 33).

3.4.3 Log1p

Com base nos dados usando a transformação log1p, foi realizada uma previsão de 15 dias utilizando os modelos Prophet e AutoARIMA, onde os dados foram divididos em proporções de 80/20. As figuras (Figura 34 e Figura 35) mostram as predições resultantes.

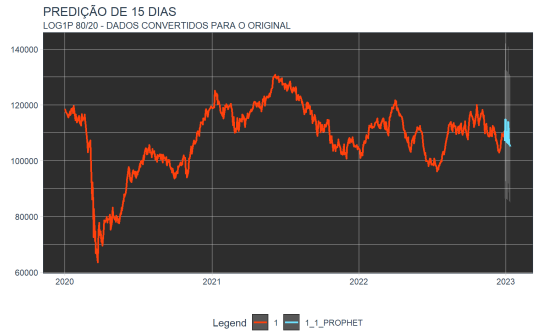


Figura 34 – Prophet predição dos dados log1p 80/20

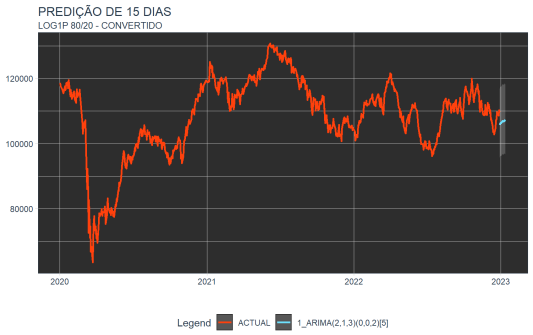


Figura 35 – AutoARIMA predição dos dados log1p 80/20

A Tabela 8 e Tabela 9, exibe a taxa de erro percentual dos modelos Prophet e AutoARIMA em prever 15 dias com base nos dados transformados com a função log1p e divididos em proporções de 80/20.

Data	Previsto	Real	Erro de Previsão (%)
30/12/2022	107.007,40	-	-
31/12/2022	114.716,50	-	-
01/01/2023	114.629,50	-	-
02/01/2023	106.714,80	106.376,00	0,32
03/01/2023	106.810,00	104.166,00	2,54
04/01/2023	106.806,40	105.334,00	1,40
05/01/2023	106.625,30	107.518,00	-0,83
06/01/2023	106.315,20	108.836,00	-2,32
07/01/2023	113.903,00	-	-
08/01/2023	113.735,90	-	-
09/01/2023	105.799,90	109.227,00	-3,14
10/01/2023	105.804,20	110.912,00	-4,61
11/01/2023	105.705,00	111.763,00	-5,43
12/01/2023	105.425,70	111.877,00	-5,79
13/01/2023	105.016,40	111.036,00	-5,42

Tabela 8 – Prophet (Log1p) - erro de previsão (%) de 15 dias

Data	Previsto	Real	Erro de Previsão (%)
30/12/2022	105.783,10	-	-
31/12/2022	106.210,40	-	-
01/01/2023	106.377,30	-	-
02/01/2023	106.255,80	106.376,00	-0,11
03/01/2023	106.466,00	104.166,00	2,21
04/01/2023	106.620,40	105.334,00	1,22
05/01/2023	106.812,50	107.518,00	-0,66
06/01/2023	106.998,00	108.836,00	-1,69
07/01/2023	106.782,80	-	-
08/01/2023	106.808,90	-	-
09/01/2023	106.864,80	109.227,00	-2,16
10/01/2023	107.097,10	110.912,00	-3,44
11/01/2023	107.146,40	111.763,00	-4,12
12/01/2023	106.994,80	111.877,00	-4,34
13/01/2023	106.870,80	111.036,00	-3,77

Tabela 9 – AutoARIMA (Log1p) - erro de previsão (%) de 15 dias

Com base na comparação das tabelas de erros de previsão de log1p apresentadas na Tabela 8 e Tabela 9, constatou-se que o modelo AutoARIMA, cujo desempenho é ilustrado na Figura 35, obteve um desempenho preditivo superior ao modelo Prophet (Figura 34).

4 CONCLUSÃO

Com base na análise realizada, conclui-se que a pandemia de covid-19 no ano de 2020 teve um impacto significativo nas ações brasileiras, gerando valores atípicos nos dados. No entanto, os anos seguintes apresentaram uma tendência de estabilidade, indicando uma recuperação

econômica. Foi identificada sazonalidade nos padrões regulares presentes na série e a tendência foi representada em um gráfico de linha, com destaque para as piores medianas observadas nos meses de junho e julho.

Foram utilizados os modelos *Prophet* e autoARIMA, que apresentaram melhor desempenho com dados originais e diferenciação na proporção 70/30, além de transformação log1p na proporção 80/20. As métricas de erro utilizadas foram RMSE, MAE, SMAPE, MAPE, MASE e RSQ.

Ao analisar as tabelas de erros de previsão (%) de 15 dias dos modelos Prophet e AutoARIMA, constatou-se que o modelo AutoARIMA obteve um desempenho preditivo superior ao modelo Prophet na previsão dos dados originais e dos dados transformados com log1p. No entanto, ao se considerar os dados diferenciados, o modelo Prophet destacou-se. Embora tenha apresentado um desempenho preditivo superior apenas nos dados diferenciados, o modelo Prophet foi considerado o melhor, indicando que a diferenciação foi uma estratégia eficaz para melhorar sua capacidade de previsão para um horizonte de 15 dias da série temporal do índice Bovespa.

REFERÊNCIAS

B3. Índice Ibovespa. Disponível em: <https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm>. Acesso em: 21 fev. 2023.

B3. Calendário de negociação. Disponível em: <https://www.b3.com.br/pt_br/solucoes/plataformas/puma-trading-system/para-participantes-e-traders/calendario-de-negociacao/feriados/>. Acesso em: 21 fev. 2023.

Box, G. E. P. et al. Time series analysis. Forecasting and control. 5th ed. 5th ed.. ed. [S.l.]: Hoboken, NJ: John Wiley & Sons, 2016. 712 p. ISSN 1940-6347. ISBN 978-1-118-67502-1/hbk.

BROCKWELL, P. J. et al. Introduction to time series and forecasting. Nova Iorque, EUA: Springer, 2016.

DICKEY, D. A.; FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. Journal of the American statistical association, Taylor & Francis, v. 74, n. 366a, p. 427–431, 1979.

FGV. COVID-19 E MERCADO FINANCEIRO. Rio de Janeiro: FGV, 2020. Disponível em:<https://fgvprojetos.fgv.br/sites/fgvprojetos.fgv.br/files/mercadofinanceiro_v07.pdf>. Acesso em: 21 fev. 2023.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic Time Series Forecasting: the forecast package for R. *Journal of Statistical Software*, v. 27, n. 3, p. 1-22, 2008.

IBM Cloud Education. Machine Learning. Disponível em: <https://www.ibm.com/br-pt/cloud/learn/machine-learning>. Acesso em: 03 jan. 2023.

LIMA, J. E. C.; CASTRO, L. F. de; CARTAXO, G. A. A. Aplicação do modelo sarima na previsão de demanda no setor calçadista/application of the sarima model in the forecast for demand in the footwear sector. ID on line REVISTA DE PSICOLOGIA, v. 13, n. 46, p. 892–913, 2019.

Liu, Z. A. Berkeley elementary function test suite. M.S. thesis, Computer Science Division, Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA, USA, December 1987.

LÚCIA, F. A. V. A. V. Manual de econometria. Vasconcelos, MAS; Alves, D. São Paulo: Editora Atlas, 2000.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. Introduction to time series analysis and forecasting. [S.l.]: John Wiley & Sons, 2015.

MORETTIN, P. A.; TOLOI, C. Análise de séries temporais. In: Análise de séries temporais. [S.l.: s.n.], 2006. p.

Nielsen, A. (2021). Análise Prática de Séries Temporais. São Paulo: Alta Books.

Facebook. Prophet: Forecasting at Scale. Disponível em: <https://facebook.github.io/prophet/>. Acesso em 13 fev. 2023.

R CORE TEAM. R: a language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria. Disponível em http:www.R-project.org, 2021.

SHUMWAY, R.; STOFFER, D. Time series analysis using the R Statistical Package.[S.l.]: free dog publishing, 2017.

Taylor, S. J., & Letham, B. (2017). Prophet: forecasting at scale. Facebook Research. Disponível em: <https://research.facebook.com/blog/2017/02/prophet-forecasting-at-scale/>. Acesso em: 21 fev. 2023.

WALTER, O. M. F. C. et al. Aplicação de um modelo sarima na previsão de vendas de motocicletas. Exacta, Universidade Nove de Julho, v. 11, n. 1, p. 77–88, 2013.

Yahoo! Finanças. *IBOVESPA* (^BVSP). Disponível em: <https://br.financas.yahoo.com/quote/%5EBVSP/history/>. Acesso em: 31 dez. 2022.