

# CAPÍTULO 4

## UMA APLICAÇÃO DO MODELO SARIMA PARA PREVER O NÚMERO DE NOTIFICAÇÕES DE DENGUE NA REGIÃO DE PARANAÍ - PARANÁ

Gabriel Manoel da Silva  
Eduardo dos Santos Zdepski  
Camila Celestrino de Castro  
Franchesco Sanches dos Santos  
Marcos Vinicius de Oliveira Peres

### RESUMO

Dentre as diversas doenças de áreas tropicais e subtropicais, a dengue está em destaque entre elas. Considerando esse fato, nesse estudo que surgiu a partir de um projeto de ensino dentro do curso de matemática da Universidade Estadual do Paraná – *campus* de Paranavaí, buscou-se modelar o número de notificações de casos de dengue na região de saúde de Paranavaí - Paraná. Para realizar a modelagem foi considerado as abordagens tradicionais das séries de tempo em dados com sazonalidade, por meio dos modelos SARIMA. Diversos parâmetros foram considerados para o modelos SARIMA e testado sua aderência aos dados, tal que, uma modelo melhor se ajustou. Considerando esse modelo, comparações de previsões com dados reais foram realizadas, e assim, verificou-se que o modelo encontrando é útil para fazer a previsão do número de notificações de casos de dengue na região de saúde de Paranavaí.

**PALAVRAS-CHAVE:** Box-Jenkins. Previsão. Software R. Séries Temporais. UNESPAR.

### 1. INTRODUÇÃO

A dengue é uma das principais doenças para a saúde pública em áreas tropicais e subtropicais do mundo. A dengue é uma doença viral transmitida principalmente pelo mosquito *Aedes aegypti*, que é mais comum em regiões tropicais e subtropicais. Ela é caracterizada por febre alta, dores de cabeça e no corpo, náuseas, vômitos e manchas vermelhas na pele. Em casos mais graves, pode levar à dengue hemorrágica, que pode ser fatal. A quantidade média de casos de dengue no mundo varia de ano para ano e de região para região. Segundo a Organização Mundial da Saúde e Organização Pan-Americana da Saúde (OPAS, 2020), estima-se que haja cerca de 390 milhões de casos de dengue a cada ano no mundo, sendo que cerca de 96 milhões de casos manifestam sintomas clínicos da doença. No entanto, esses números podem variar bastante de acordo com fatores como clima, urbanização, densidade populacional e a presença do mosquito transmissor. É importante destacar que a dengue é uma doença endêmica em muitos países tropicais e subtropicais, e que medidas de prevenção e controle são fundamentais para minimizar sua incidência e impacto na saúde pública.

Existe na literatura diversas ferramentas matemáticas e estatísticas que podem serem utilizados para descrevem e compreender a dinâmica da dengue. Uma dessas ferramentas são

os Modelos de Séries Temporais. Segundo Morettin e Tolo (2006), uma série temporal é qualquer conjunto de observações ordenadas no tempo. Modelos de séries temporais são modelos estatísticos utilizados para analisar e prever padrões e tendências em dados que variam ao longo do tempo. Esses modelos são amplamente utilizados em diversas áreas, como finanças, economia, engenharia, biologia, entre outras, para prever comportamentos futuros com base em padrões históricos. Os modelos de séries temporais geralmente levam em consideração fatores como sazonalidade, tendências, efeitos de eventos externos e autocorrelação dos dados, entre outros, para criar previsões e estimativas precisas.

Dentre os modelos de séries temporais, um modelo de destaque é o SARIMA. O modelo SARIMA (Seasonal Autoregressive Integrated Moving Average - Autorregressivo Integrado de Médias Móveis Sazonal) é uma classe de modelos de previsão de séries temporais que estendem o modelo ARIMA (Autoregressive Integrated Moving Average - Autorregressivo Integrado de Médias Móveis) para lidar com dados sazonais. Ele é usado para modelar e prever séries temporais com padrões sazonais que se repetem regularmente ao longo do tempo, ou seja, quando os dados apresentam flutuações regulares em intervalos de tempo fixos. Os parâmetros do modelo SARIMA são estimados a partir dos dados históricos da série temporal. A partir daí, o modelo pode ser usado para fazer previsões futuras, que serão influenciadas tanto pelos padrões de longo prazo quanto pelos padrões sazonais da série.

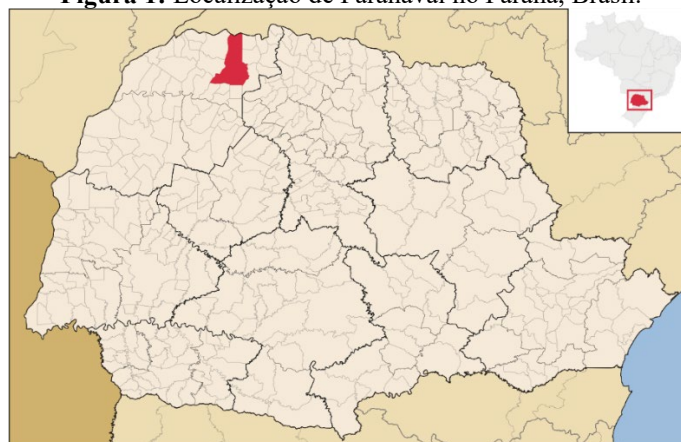
Esse trabalho surgiu durante um estudo de um projeto de ensino realizado pelo Colegiado de Matemática da Universidade Estadual do Paraná, *campus* de Paranavaí. Ao estudar sobre modelagem de séries de tempos com sazonalidade, foi considerado os dados locais sobre o número de notificações de casos de dengue na regional de saúde de Paranavaí. Com intuito de aprender, praticar e interpretar as técnicas de modelagem de séries temporais considerando um modelo SARIMA e o uso da metodologia proposta por Box & Jenkins, elaborou-se esse trabalho e os resultados são apresentados no decorrer no texto.

## 2. MATERIAIS E MÉTODOS

A cidade de Paranavaí é a maior cidade da região noroeste do estado do Paraná, no Brasil, possui uma população de cerca de 90 mil habitantes (IBGE, 2023). O clima predominante em Paranavaí e região é o subtropical úmido, com chuvas bem distribuídas ao longo do ano e temperaturas amenas. Durante o verão, a temperatura média é de cerca de 25°C, com máximas que podem chegar a 30°C ou mais. Já no inverno, as temperaturas ficam em torno

de 15°C, com mínimas que podem chegar a 5°C ou menos em alguns dias (WEATHER SPARK, 2023). Veja na Figura 1, a localização da cidade de Paranavaí.

**Figura 1:** Localização de Paranavaí no Paraná, Brasil.



**Fonte:** Abreu (2006).

A regional de saúde de Paranavaí é constituída pelos 28 municípios do Noroeste do Paraná, possuindo uma população total aproximada de 250.000 habitantes (veja a Figura 2). Nomeada de 14º Regional de Saúde do estado, sendo uma instância administrativa intermediária da Secretaria de Estado da Saúde (SESA) e as Secretarias Municipais de Saúde dos municípios.

**Figura 2:** Mapa de abrangência da Regional de Saúde de Paranavaí.



**Fonte:** SESA (2020).

O número mensal de caso notificados de dengue na regional de Saúde de Paranavaí foi obtido do Sistema de Informação de Agravos de Notificação (SINAN) dos registros de dengue pertencentes ao DATASUS do Ministério da Saúde, podendo ser consultado gratuitamente em DATASUS (2022). O conjunto de dados foi separado em duas partes: os dados mensais

observados de notificações de janeiro de 2010 a dezembro de 2020, que foram usados para desenvolver o modelo de série temporal, e os dados mensais de notificações do ano de 2021, que foi usado para validar o modelo. Veja na Tabela 1, os dados completos do número mensal de notificações de dengue na regional de Saúde de Paranavaí no período de janeiro de 2010 até dezembro de 2021.

**Tabela 1:** Número de notificações mensais na regional de saúde de Paranavaí, 2010-2021.

	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
<b>2010</b>	22	85	277	405	276	27	17	4	3	1	1	1	<b>1119</b>
<b>2011</b>	6	6	39	74	44	11	1	0	2	1	2	1	<b>187</b>
<b>2012</b>	2	2	56	42	57	20	0	1	1	3	37	239	<b>460</b>
<b>2013</b>	2001	5333	4953	2880	1194	238	71	13	14	14	37	100	<b>16848</b>
<b>2014</b>	799	905	572	766	575	137	54	34	56	101	56	63	<b>4118</b>
<b>2015</b>	651	867	1071	992	339	147	17	16	17	18	13	27	<b>4175</b>
<b>2016</b>	81	122	152	98	28	7	4	2	1	4	12	5	<b>516</b>
<b>2017</b>	9	9	17	8	3	1	5	6	2	11	8	5	<b>84</b>
<b>2018</b>	7	7	5	10	6	7	10	4	6	10	5	7	<b>84</b>
<b>2019</b>	22	44	254	1055	994	521	227	81	96	246	679	2187	<b>6406</b>
<b>2020</b>	6654	7159	3674	1657	621	218	57	27	36	26	59	76	<b>20264</b>
<b>2021</b>	62	126	224	267	138	32	12	14	20	25	15	23	<b>958</b>

Fonte: Autoria própria (2023).

A técnica desenvolvida por Box e Jenkins (1970) possibilita a previsão de valores futuros de uma série apenas com base em seus valores anteriores e presentes. Esses modelos são conhecidos como modelos autorregressivos integrados de médias móveis ou ARIMA, sendo que composto de três componentes: a componente autoregressiva (AR), a componente de média móvel (MA) e a componente de diferenciação (I). Quando consideram relações sazonais, são denominados SARIMA. Considerando uma série temporal  $Y = (y_1, y_2, \dots, y_n)$ , um modelo SARIMA com  $S$  observações por período é denotado por  $SARIMA(p, d, q)(P, D, Q)_S$  e pode ser representado da seguinte maneira:

$$\phi(L)\Phi(L^S)(1-L)^d(1-L^S)^D y_t = \theta(L)\Theta(L^S)\varepsilon_t$$

Onde:

- $p$  é a ordem do polinômio autorregressivo não sazonal,  $\phi(L) = (1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)$ ;
- $P$  é a ordem do polinômio autorregressivo sazonal,  $\Phi(L) = (1 - \Phi_1 L^S - \Phi_2 L^{2S} - \dots - \Phi_P L^{PS})$ ;
- $q$  é a ordem do polinômio de médias móveis não sazonal  $\theta(L) = (1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q)$ ;

- $Q$  é a ordem do polinômio de médias móveis sazonal  $\Theta(L) = (1 - \Theta_1 L^S - \Theta_2 L^{2S} - \dots - \Theta_Q L^{QS})$ ;
- $d$  é a ordem da diferença não sazonal;
- $D$  é a ordem da diferença sazonal;
- $L$  é a operador de defasagem temporal (lag), dado por  $L^k = y_{t-k}/y_t$ ; e
- $\varepsilon_t$  são os erros assumidos como variáveis independentes e identicamente distribuídas provenientes de uma distribuição com uma média igual a zero e a variância  $\sigma^2$ .

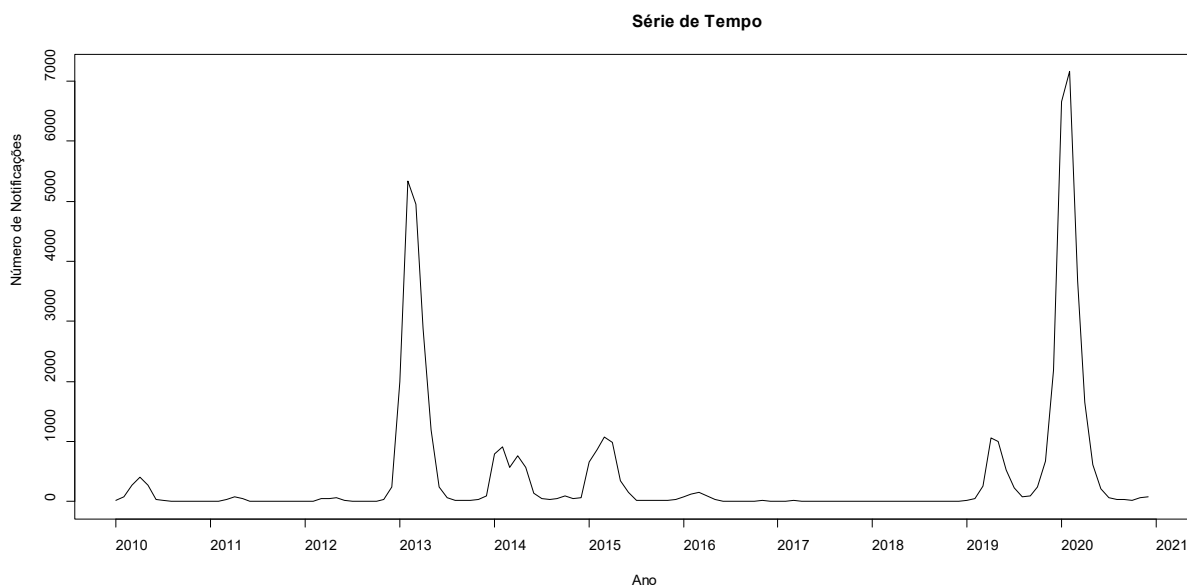
Veja que os componentes Sazonal AR e Sazonal MA são semelhantes ao AR e MA, mas modelam a dependência em relação a valores passados sazonais. Em outras palavras, elas consideram a sazonalidade dos dados e adicionam um componente de dependência que leva em conta o mesmo período de tempo da sazonalidade. A parte AR captura a dependência entre os valores passados da série, a parte MA captura a dependência entre os resíduos passados e a parte I é usada para diferenciar a série, a fim de torná-la estacionária.

No presente artigo, foi utilizado o software estatístico R (R CORE TEAM, 2022) para ajustar modelos SARIMA ao número de notificações de casos de dengue na regional de saúde de Paranavaí entre 2010 e 2021, considerando a abordagem de modelagem Box-Jenkins (BOX; JENKINS, 2008). A adequação de cada modelo foi verificada por meio de gráficos de séries temporais, de histograma e autocorrelação (ACF) dos resíduos padronizados e pelo teste de Ljung-Box, que é um teste para hipóteses de ausência de correlação em um número especificado de defasagens temporais (LJUNG; BOX, 1978). O critério de informação de Akaike (AIC) foi considerado para comparar a qualidade de ajuste de diferentes modelos (AKAIKE, 1974). Em adicional também se utilizou os valores da raiz do erro quadrático médio (RMSE, root mean square error) e valores do erro absoluto médio (MAE, mean absolute error). Os valores de AIC, RMSE e MAE menores indicam melhor ajuste.

### 3. RESULTADOS

A Figura 3 mostra o número de notificações dos casos de dengue na região de Paranavaí, entre 2010 até 2021. É possível observar que desse períodos houve alguns momentos de baixa notificações (dois ou três anos), seguidos de momentos de altas notificações (dois ou três anos). Destacando em especial os anos de 2013 e 2020, em que houve uma incidência muito grande no número de notificação de casos. Considerando a Tabela 1 e a Figura 3, os meses de março e abril são particularmente os meses com os maiores números de notificações de casos de dengue na regional de saúde de Paranavaí.

**Figura 3:** Número de notificações dos casos de dengue na região de Paranavaí.

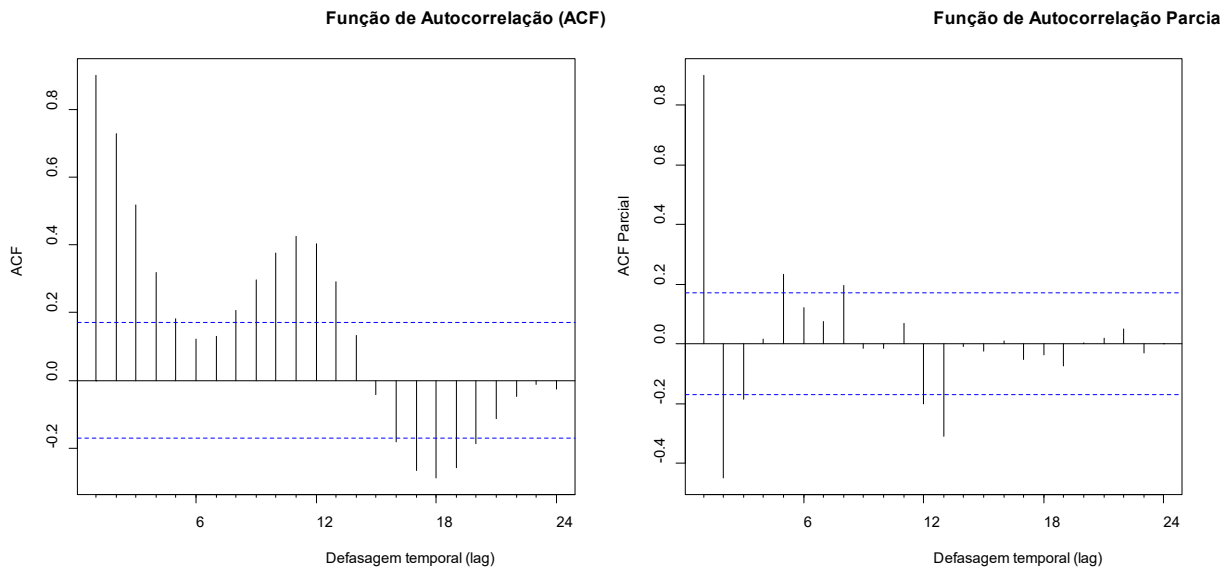


Fonte: Autoria própria (2023).

Buscando induzir uma variância constante aos dados, foi aplicado uma transformação logaritmo nos dados exibidos na Tabela 1 e na Figura 3. Considerando um gráfico da série temporal aplicado a transformação logarítmica (não mostrado aqui), observa-se que a série ainda não é estacionária. Dessa forma, para obter uma série mais estacionária foi aplicado uma primeira diferenciação simples. Após, essa primeira diferenciação foi verificada que a série aparenta estar estacionária na média, então no modelo SARIMA  $d = 1$ .

A Figura 4 mostram os gráficos da função de autocorrelação estimada (ACF) e função de autocorrelação parcial (PACF) da série transformada pelo logaritmo usando dados de 2010 a 2020. Note que a ACF da série transformada exibe uma periodicidade de comprimento próximo de 12 meses, assim  $S = 12$ . Esse fato é esperado, pois a incidência de dengue apresenta um ciclo sazonal. O PACF sugere que  $p$  deve ser igual a 2 ou 1, uma vez que as autocorrelações parciais são próximas de zero em todos os lags que excedem 2, e ainda, a ACF sugere uma média móvel de ordem  $q$  igual a 2 ou 3, uma vez que suas autocovariâncias estão próximas de zero em lags que excedem 3. Além disso, considerando o comportamento sazonal na ACF há aqui uma sugestão uma média móvel sazonal  $Q$  de ordem 1. Contudo, buscando obter um melhor ajuste também foi testado  $Q = 2$ . Em adicional, notou-se na PACF uma correlação distante do 0 próximo do lag 1, assim, foi considerando no modelo SARIMA  $P = 1$ .

**Figura 4:** Funções de autocorrelação e autocorrelação parcial para a série transformada pelo logaritmo.



**Fonte:** Autoria própria (2023).

Ao obter as estimativas dos modelos SARIMA foram encontrados problemas de convergência ao usar  $D = 0$ , além disso, foi possível observar que resultados melhores eram obtidos quando  $D = 1$ . Isso é, quando aplicado uma diferença sazonal nos dados transformados houve uma melhor convergência e melhores resultados, do que quando não aplicado a diferença sazonal. Portanto, considerando que uma diferença sazonal de 1 foi suficiente, foi fixado que  $D = 1$ . A Tabela 2 mostra os valores para os critérios de seleção e as estimativas para a variância  $\sigma^2$  dos erros dos modelos SARIMA ajustados a série transformado considerando diferentes parâmetros. O modelo com os menores valores nos critérios de seleção para este conjunto de dados é o modelo com melhor ajuste, foi o  $SARIMA(2,1,3)(1,1,2)_{12}$ .

**Tabela 2:** Critérios de seleção para diferentes modelos  $SARIMA(p, d, q)(P, D, Q)_s$  e estimativas de  $\sigma^2$ .

Modelo SARIMA	AIC	$\sigma^2$	RSME	MAE
$(2,1,3)(1,1,2)_{12}$	<b>291.61</b>	<b>0.4693</b>	<b>0.6504</b>	0.4942
$(2,1,2)(1,1,2)_{12}$	292.94	0.4834	0.6601	0.4947
$(2,1,3)(1,1,1)_{12}$	297.41	0.5192	0.6841	0.5202
$(2,1,2)(1,1,1)_{12}$	303.72	0.5554	0.6841	0.5202
$(1,1,3)(1,1,2)_{12}$	295.72	0.4931	0.6667	0.4943
$(1,1,2)(1,1,2)_{12}$	295.42	0.4869	0.6625	<b>0.4935</b>
$(1,1,3)(1,1,1)_{12}$	301.07	0.5452	0.7010	0.5260
$(1,1,2)(1,1,1)_{12}$	300.97	0.5531	0.7061	0.5323

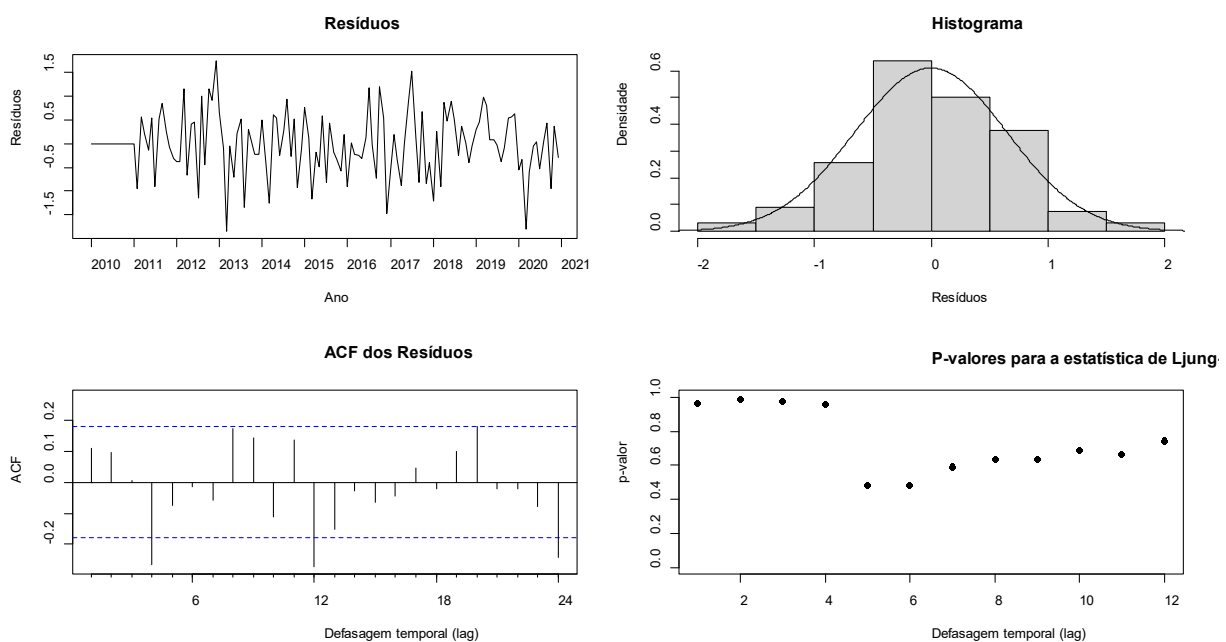
**Fonte:** Autoria própria (2023).

Considerando o modelo escolhido, os parâmetros autorregressivos  $\phi_1$  e  $\phi_2$  foram estimados em  $-0.8998$ (erro padrão,  $0.1144$ ) e  $-0.8136$ ( $0.1441$ ), respectivamente, e os

parâmetros de média móvel  $\theta_1$ ,  $\theta_2$  e  $\theta_3$  foram estimados em -1.0609 (0.1358), 1.0862 (0.1277) e 0.2739(0.0938), respectivamente. O componente autorregressivo sazonal  $\Phi_1$  foi estimado em -0.6251(0.1034). E os com omentos de médias móveis sazonais  $\Theta_1$  e  $\Theta_2$  foram estimados em 0.1098(0.2143) e -0.8898(0.2070), respectivamente. Apesar de obter um alto erro padrão em  $\Theta_1$ , esse parâmetro é importante para a modelagem, isso pode ser colaborado pelos resultados exibidos na Tabela 2.

A Figura 5 mostra os gráficos de diagnósticos para os resíduos do modelo. É possível notar que os resíduos padronizados estimados a partir deste modelo se comportam como uma sequência independente e identicamente distribuída com média zero e variância constante. O histograma mostra que os resíduos padronizados para o modelo se aproximaram de uma distribuição normal (afirmação colaborada pelo teste de Shapiro-Wilk, p-valor = 0,4709). O ACF dos resíduos sugere que as autocorrelações estão próximas de zero, como desejado. Por fim, os altos valores-p para a estatística de Ljung-Box, indica que não há indicativos para rejeitar a hipótese nula de independência nessa série residual. Portanto, é possível dizer que o modelo  $SARIMA(2,1,3)(1,1,2)_{12}$  se ajusta adequadamente aos dados.

**Figura 5:** Gráficos de diagnósticos para os resíduos do modelo  $SARIMA(2,1,3)(1,1,2)_{12}$ .



**Fonte:** Autoria própria (2023).

A Tabela 3 mostra os valores preditos fora da amostra para 2021 considerando o modelo  $SARIMA(2,1,3)(1,1,2)_{12}$  e os valores observados do número de notificações de dengue na regional de saúde de Paranaíba. Os valores preditos estão relativamente próximos aos valores



observados. Apesar de alguns valores ficarem discrepantes, este resultado indica que o modelo fornece um ajuste aceitável para prever o número de notificações de dengue.

**Tabela 3:** Valores observados e preditos para o número de notificações de casos de Dengue na regional de saúde de Paranavaí.

2021	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out	Nov	Dez	Total
<b>Observado</b>	62	126	224	267	138	32	12	14	20	25	15	23	<b>958</b>
<b>Predito</b>	170	218	129	97	59	17	6	4	5	5	9	8	<b>727</b>

Fonte: Autoria própria (2023).

#### 4. CONSIDERAÇÕES FINAIS

Considerando a proposta desse estudo em aplicar as ferramentas clássicas de séries temporais em um conjunto de dados com sazonalidade, é possível dizer que os resultados foram satisfatórios. Durante esse estudo os alunos se dedicaram em revisar conceitos que eles ainda não tinham familiaridade, demonstrando interesse nos conceitos e compreendendo a importância de estudar e modelar as séries de tempo. Para o conjunto de dados estudado, o modelo  $SARIMA(2,1,3)(1,1,2)_{12}$  refletiu bem a tendência na incidência do número de notificações de dengue na regional de saúde de Paranavaí. Mostrou-se nesse trabalho que o número de casos de dengue em um determinado mês pode ser estimado pelo número de casos de dengue ocorridos 1, 2 ( $p = 2$ ) e 12 ( $S = 12, P = 1$ ) meses antes, e ainda se descobriu que um componente de média móvel de ordem  $q$  igual a 3 é adequado para os dados. Quando usado este modelo para produzir previsões fora da amostra utilizado para estimar o modelo, observou-se que o modelo SARIMA foi capaz de representar o número de casos em um ano subsequente com relativa precisão. E por fim, ressalta-se a importância de permitir que os alunos dos cursos de Licenciatura em Matemática, conheçam e experimente ferramentas diferentes daquelas que são comuns em sua graduação, buscando expandir o conhecimento matemático e suas aplicações práticas.

#### REFERÊNCIAS

ABREU, R. L. **Map locator of Paraná's Paranavaí city**. 2006. Disponível em: [https://commons.wikimedia.org/wiki/File:Parana\\_Municip\\_Paranavai.svg](https://commons.wikimedia.org/wiki/File:Parana_Municip_Paranavai.svg). Acessado em: Mar. 2023.

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, [S.l.], v. 19, n. 6, p. 716-723, 1974. Disponível em: <https://doi.org/10.1109/TAC.1974.1100705>. Acessado em: Mar. 2023.

BOX, G.; JENKINS, G. **Time series analysis: Forecasting and control**. 4. ed. New York: John Wiley & Sons, 2008.

DATASUS. Sistema de Informação de Agravos de Notificação - TabNet: **Dengue - Casos Confirmados Notificados no Brasil**. 2022. Disponível em: <http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sinannet/cnv/denguebbr.def>. Acessado em: Mar. 2023.

IBGE. **Instituto Brasileiro de Geografia e Estatística**. Paranavaí - PR. 2023. Disponível em: <https://cidades.ibge.gov.br/brasil/pr/paranavai/panorama>. Acessado em: Mar. 2023.

LJUNG, G. M.; BOX, G. E. P. On a measure of lack of fit in time series models. **Biometrika**, [S.l.], v. 65, n. 2, p. 297-303, 1978. Disponível em: <https://doi.org/10.1093/biomet/65.2.297>. Acessado em: Mar. 2023.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de series temporais**. 2. ed. São Paulo: Blücher, 2006.

SESA. **Secretaria de Estado da Saúde do Paraná**. 14ª Regional de Saúde - Paranavaí. 2020. Disponível em: <https://www.saude.pr.gov.br/Pagina/14a-Regional-de-Saude-Paranavai>. Acessado em: Mar. 2023.

OPAS. **Organização Pan-Americana da Saúde**. Dengue. 2020. Disponível em: <https://www.paho.org/pt/topicos/dengue>. Acessado em: Mar. 2023.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, 2022. Disponível em: <https://www.r-project.org>. Acessado em: Mar. 2023.

WEATHER SPARK. **Clima e condições meteorológicas médias em Paranavaí no ano todo**. Disponível em: <https://pt.weatherspark.com/y/29639/Clima-caracter%C3%ADstico-em-Paranava%C3%AD-Brasil-durante-o-ano>. Acessado em: Mar. 2023.