# A TIME SERIES ANALYSIS OF TRENDING DENGUE CASES IN SRI LANKA

**RUVANI KURUKULASURIYA PERERA**
**Bachelor of Science Honors in Psychology, Cardiff Metropolitan University, UK, 2019**

A thesis submitted
in partial fulfillment of the requirements for the degree of

**MASTER OF SCIENCE**

in

**HEALTH SCIENCES (PUBLIC HEALTH)**

Faculty of Health Sciences
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

A TIME SERIES ANALYSIS OF TRENDING DENGUE CASES IN SRI LANKA

RUVANI KURUKULASURIYA PERERA

Date of Defense: February 28, 2025

| | | |
|---|---|---|
| Dr. Tracy Oosterbroek<br>Thesis Supervisor<br>Thesis Examination Committee Member | Associate Professor<br>Faculty of Health Sciences | Ph.D./R.N. |
| Dr. Nimesh Patel<br>Thesis Co-supervisor<br>Thesis Examination Committee Member | Instructor<br>Faculty of Health Sciences | M.D. |
| Dr. Trushar Patel<br>Thesis Committee Member<br>Thesis Examination Committee Member | Associate Dean<br>Faculty of Arts and Science<br>Associate Professor<br>Faculty, Chemistry and Biochemistry | Ph.D. |
| Dr. Sudath Samaraweera<br>Thesis External Committee Member<br>Thesis Examination Committee Member | Director,<br>National Dengue Control Unit<br>Ministry of Health, Sri Lanka | Ph.D./M.D. |
| Dr. Julia Brassolotto<br>Chair, Thesis Examination Committee | Associate Dean<br>Associate Professor<br>Faculty of Health Sciences<br>University of Lethbridge | Ph.D. |

## Dedication

I dedicate this work to my beloved and supportive husband of 25 years, the pillar of unity and strength in our family. For three years, he not only stood by me in spirit when I lived in solitude overseas but also stepped into the role of both mom and dad and guided our sons through the *storm and tide* of their adolescent transition. Also, I dedicate my work with much affection to my twin sons, now young men, who united in supporting their dad despite missing their mom's presence while she *followed the light* of knowledge in a faraway land!

My sacrifices were not in vain, as they have rewarded me profoundly. I am deeply grateful for the personal investment and the rewards I have received from my sacrifices.

**Abstract**

The study aimed to predict dengue case numbers in Sri Lanka from January 2024 to December 2025. The prediction will assist the National Dengue Control Unit of Sri Lanka in assessing the potential dengue case numbers before a seasonal dengue crisis. This allows the Ministry of Health of Sri Lanka to plan effective healthcare mobilization and manage its resources during dengue seasons. Secondary data on all island dengue cases was obtained from the National Dengue Control Unit's national surveillance system from 2015 to 2023. A seasonal ARIMA(0,1,1)(0,0,2)[12] model was generated in R software by the *forecast* package's time series function based on the Box-Jenkins method. The ARIMA model was validated as a good fit for prediction with the Ljung-Box ($p$-value >0.05), Shapiro-Wilk ($p$-value >0.05), and ADF ($p$-value <0.05) tests. The prediction's MAPE was estimated as accurate for forecasting (4.46). The seasonal ARIMA model demonstrated the ability to make a short-term prediction in univariate analyses.

***Keywords:*** *Sri Lanka, Time Series, Prediction, National, Dengue, Unit, ARIMA*

## Ethics Statement

The work described in this thesis received research ethics approval from the University of Alberta Research Ethics Board under the following applications:

| **Title** | **Number** | **Date** |
|---|---|---|
| A Time Series Analysis of Trending Dengue Cases in Sri Lanka | **Pro00142813** | June 6, 2024 |

## Acknowledgments

I am deeply grateful to Dr. Tracy Oosterbroek, who took up supervising me when I faced many challenges. Her guidance in laying a solid foundation for my thesis proposal, followed by invaluable feedback when writing the research, was the cornerstone of my academic journey that took me to the end. Her empathy and kindness were remarkable humane qualities I admired throughout my educational journey. I am sincerely grateful to Dr. Nimesh Patel, whose instruction in R software I extensively used to build up the research's data analysis. I am honored to have the Director of the National Dengue Control Unit of the Ministry of Health, Sri Lanka, on the thesis committee, Dr. Sudath Samaraweera, who guided me through my thesis project. In addition, I am grateful to him for obtaining approval from the Public Health Services of the Ministry of Health, Sri Lanka, to permit me to work on the national notification dengue data. His expertise in dengue epidemiology empowered my research thesis and its practical application. Furthermore, I thank Dr. Trushar Patel, who was always ready to assist me with queries regarding my research work and being a part of the supervisory committee in exigency. Finally, but importantly, I am grateful to Dr. Julia Brassolotto for considering chairing the thesis defense committee at such short notice.

## Table of Contents


Dedication ..... iii

Abstract ..... iv

Ethics Statement ..... v

Acknowledgments ..... vi

List of Tables ..... ix

List of Figures ..... x

List of Boxes ..... xi

List of Abbreviations ..... xii

Introduction ..... 1

Materials ..... 5

- Secondary Data ..... 5
- Ethics Approval ..... 6

Methodology ..... 6

- Time Series Properties ..... 6
- ARIMA Model's Mathematical Components ..... 10
- Data Analysis ..... 11
  - Automated ARIMA Analysis ..... 11
  - Data Transformation and Partition ..... 12

Results ..... 17

Discussion ..... 29

Limitations and Recommendations ..... 33

Conclusion ..... 36

Supporting Information: All Island Dengue Cases from 2015 to 2023 ..... 37

References ..... 38

## List of Tables

**Table 1.** Stationarity hypothesis with the ADF test in R software. ..... 8
**Table 2.** Ljung-Box test of ARIMA model validation. ..... 9
**Table 3.** Model's accuracy test with MAPE. ..... 9
**Table 4.** Table of p-values of ARIMA(0,1,1)(0,0,2)[12] model. ..... 22

## List of Figures

**Figure 1.** Time series plot for all-island dengue data from January 2015 to December 2023. .... 13
**Figure 2.** Boxplot of the all-island dengue time series. .... 14
**Figure 3.** Boxplot of the log-transformed all-island time series data. .... 14
**Figure 4.** Log transformed all-island dengue time series. .... 15
**Figure 5.** ACF plot of the training data component. .... 16
**Figure 6.** PACF plot of the training data component. .... 16
**Figure 7.** Mean and variance of the training dataset with the first-order differencing. .... 17
**Figure 8.** ACF plot of the time series training data component with the first-order differencing. .... 18
**Figure 9.** PACF plot of the time series training data component with the first-order differencing. .... 18
**Figure 10.** ADF test of residuals of the modified automated All_Auto model ARIMA(0,1,1)(0,0,2)[12]. .... 20
**Figure 11.** ACF plot of All_Auto model ARIMA(0,1,1)(0,0,2)[12] residuals. .... 21
**Figure 12.** PACF plot of All_Auto model ARIMA(0,0,1)(0,0,2)[12] residuals. .... 21
**Figure 13.** Ljung-Box test p-value plot of the Auto_Model ARIMA(0,1,1)(0,0,2)[12] residuals. .... 23
**Figure 14.** Forecast from Auto_Model ARIMA(0,1,1)(0,0,2)[12] on the test data component. . 24
**Figure 15.** ADF test residual plot of the Best_Model ARIMA(0,1,1)(0,0,2)[12]. .... 25
**Figure 16.** Histogram of residuals of the Best_Model ARIMA(0,1,1)(0,0,2)[12]. .... 25
**Figure 17.** Boxplot of residuals of the Best_Model ARIMA(0,1,1)(0,0,2)[12]. .... 26
**Figure 18.** Forecast of dengue cases from the Best_Model ARIMA(0,1,1)(0,0,2)[12] from January 2024 to December 2025 (without scaling). .... 27
**Figure 19.** Forecast of dengue cases from the Best_Model ARIMA(0,1,1,)(0,0,2)[12] from January 2024 to December 2025 (scaled). .... 28
**Figure 20.** Forecast of dengue cases from the Best_Model ARIMA(0,1,1)(0,0,2) from January 2024 to December 2025 (without prediction intervals). .... 28

## List of Boxes

**Box 1.** Automated ARIMA coefficients for the model ARIMA(1,0,1)(2,0,0)[12] with non-zero mean. ..... 19
**Box 2.** Automated ARIMA coefficients for the modified All_Auto model ARIMA(0,1,1)(0,0,2)[12]. ..... 20
**Box 3.** All_Auto_Forecast's MAPE forecasting accuracy test. ..... 24
**Box 4.** Predicted dengue case number from January 2024 to December 2025 by the Best_Model ARIMA(0,1,1)(0,0,2)[12]. ..... 26

## List of Abbreviations

Akaike Information Criterion
AIC ..... 8
Akaike Information Criterion-Corrected
AICc ..... 8
Augmented Dicky Fuller
ADF ..... 7
Auto-Correlation Function
ACF ..... 6
Auto-Regression
AR ..... 7
Autoregression Integrated Moving Average
ARIMA ..... 4
Integrated
I…… ..... 10
International Classification of Diseases
ICD ..... 1
Mean Absolute Percentage Error
MAPE ..... 9
Moving Average
MA ..... 7
Partial Auto-Correlation Function
PACF ..... 6

## Introduction

Dengue is a mosquito-borne viral infection that poses a significant public health crisis during outbreaks.[1] Dengue is transmitted by Aedes mosquitoes, specifically *Aedes aegypti* and *Aedes albopictus*, which begin their lifecycle by breeding in water.[2 3 4 5] A seasonal pattern of mosquito breeding co-occurs with the cyclic monsoon rainfall pattern, and the virus eventually transmits rapidly.[5 3 6] Consequently, the interaction of seasonal climatic variations and vector dynamics substantially impacts the disease's transmission patterns.[5 6] Dengue transmission depends on the contact between the vector and the human host.[5 6] The geographic range of dengue is increasing due to several other factors, including unplanned rapid urbanization and construction, high population densities, and ineffective vector-control strategies.[6 7]

The World Health Organization classifies dengue fever in the International Classification of Diseases (ICD) 11$^{th}$ edition, under the codes 1D20 – 1D22.[8] According to the World Health Organization guidelines, hospital admission is recommended in the presence of severe dengue or dengue warning signs, which include persistent vomiting, abdominal pain, plasma leakage, and thrombocytopenia, which can eventually result in death.[9] In the context of hospital admission, individuals displaying clinical symptoms are typically identified and reported as cases.[5] Further, the virus's aggressive capacity to cause critical disease by interfering with specific functions of the host cells is called virulence.[5]

DENV-1, DENV-2, DENV-3, and DENV-4 are the four dengue serotypes circulating within Sri Lanka.[4 10 11 12] Recovery from the primary infection by one serotype provides lifelong immunity against that serotype and is temporary for the other.[4] If an individual gets infected with a different serotype in a subsequent infection, the risk of developing severe

dengue increases.[4] [10] [13] Primary and secondary dengue infection-differentiation has an immunoglobulin antibody marker ratio of 1.59 IgM: IgG.[14] Further, there is no specific vaccine or medication for dengue fever.[4] [3] Thus, the management methods are mainly consistent surveillance and vector control measures.[1] [15] [4]

Dengue is considered a public health risk in Sri Lanka, and management guidelines have been adopted to optimize therapy and ensure patient safety.[15] The public health burden proliferates, and hospitals are often overwhelmed due to high patient admissions during dengue epidemics.[16] Thus, strengthening dengue control activities and optimizing existing hospital-based care resources are essential to reducing dengue's economic impact on the public health framework.[16]

Annual dengue epidemics started in 1989 in Sri Lanka.[15] Dengue was designated a notifiable disease in Sri Lanka in 1996.[17] Due to endemic transmission, dengue epidemics occurred once every few years from 1991 to 2008.[2] From 2010 to 2016, dengue became a significant public health issue when it disproportionately affected the most populated Western province.[2] Cases increased steadily nationwide from 28,473 in 2011 to 55,150 in 2016.[2] An unprecedented dengue outbreak in 2017 reported 186,101 suspected cases and 440 dengue-related deaths were reported to the national surveillance system of the National Dengue Control Unit.[2] This trend continued into 2020 with a steady increase in dengue cases during the same months of the year compared to the cases reported in the preceding years.[2] The dengue situation in Sri Lanka reported a total of 42,729 and 39 deaths from January to July 2022.[2] Another outbreak occurred from 15 May to 14 June 2023, when 11,685 dengue patients were reported to the national surveillance system.[18]

Disease forecasting is vital for disease control, where the disease's public health burden

can be estimated in advance.[19 20 21] Although climate variables have been considered to predict dengue incidence, the weather changes due to unpredictable atmospheric changes such as El Nino heatwaves and unexpected torrential rains in a tropical region of >2500 mm could not very well be influential variables in predicting dengue cases.[2] Also, out of the four dengue strains, DENV-3 has now been commonly co-circulating with other dengue strains.[22] Thus, a new dengue strain dominating a particular season is another unpredictability, such as in the 2017 unprecedented outbreak.[2 22]

Previous studies on dengue forecasting are scarce in Sri Lanka, and they have yet to secure a reliable dataset to make their predictions. Ministry of Health divisional data for Machine Learning predictions had not been the actual manifestation of national notification data.[23] Such predictive models cannot be considered accurate. Some dengue predictions have only been confined to the Colombo district of the Western Province.[5] Nevertheless, the National Dengue Control Unit's surveillance system compiles the national notification data covering the entire island.[2] Thus, a considerable deficit in dengue prediction and modeling exists in Sri Lanka.

Surveillance data on infectious diseases benefits healthcare administrators by allowing them to monitor the trends of reported cases through statistical analysis periodically.[3 21] Several studies have attempted to make predictive assessments grounded in observable variables, such as dengue notification data of reported cases, in the context of dengue.[5 19] Statistical time series forecasting takes a more data-centric approach and is effective at modeling dengue fever's highly autocorrelated nature.[19] Further, a time series can be analyzed independently without any information from other covariates in univariate analysis.[5 24] Other predictive modeling includes General Additive Mixed models, Spatial Analysis, Non-linear methods, Multivariate

modeling, and Global Circulation models, which have all found frequent applications in modeling and predicting the occurrence of infectious diseases, including dengue.[5 3 6 21]

Autoregression Integrated Moving Average (ARIMA) modeling, developed initially by George Box and Gwilym Jenkins had designed the ARIMA model in the 1970s to define time series deviations using a mathematical method, stands out as a widely utilized approach for the statistical forecasting of time series data.[5 25] Although ARIMA models have become a popular choice to incorporate autoregressions, they are unsuitable for all data types.[19 25] The objective of this model is to modify the observed values and to minimize the difference between the values obtained in the model and the observed values as close to zero as possible.[25] Thus, if a data variable can be fitted into a time series, ARIMA modeling will successfully predict future disease patterns. Further, ARIMA modeling is a practical approach to uncovering hidden patterns in data and generating forecasts.[4 5 20] Using a time series analysis approach to analyze past trends of dengue fever outbreaks and incorporating this data into an appropriate statistical model to forecast future outbreaks can be beneficial for improving current prevention and control measures and gaining lead time for public health resource management.[5 20 6]

Our study aims to identify a suitable ARIMA model in R software to predict dengue cases in Sri Lanka, encompassing the whole island, monthly from 2024 to 2025. The potential of ARIMA modeling to improve dengue management strategies is significant and should be a source of hope and optimism.[5 6 20] The prediction model will be based on the surveillance data recorded reliably in the dataset *DenSys* at the National Dengue Control Unit under the Ministry of Health of Sri Lanka.[2] The National Dengue Control Unit is the central authority responsible for preventing and managing dengue and compiling national notification data.[2] DenSys has recorded dengue surveillance data weekly, monthly, and annually from 2015 to date. We will

specifically select DenSys data from 2015 to 2023 for our study. To enhance dengue management strategies, it is crucial to share the results of our research with the National Dengue Control Unit. This collaboration can significantly improve our understanding and management of dengue. Accordingly, the National Dengue Control Unit will benefit from our modeling procedure to make quick predictions from DenSys data in univariate analyses, even without meteorological data, which the organization does not compile. Enhancing the understanding of dengue temporal patterns ahead of seasonal dengue attacks will assist the National Dengue Control Unit in informing the Ministry of Health to implement effective strategies to manage public health resources.

## Materials

### Secondary Data

The Ministry of Health of Sri Lanka provides the study's dataset, meticulously registered under the National Dengue Control Unit. This comprehensive dataset, comprising numeric data with no patient identification details, reports dengue cases from 2015 to 2023. The government-approved surveillance data is a testament to the Ministry of Health's commitment to public health, ensuring the reliability and accuracy of the data source.[2] Each hospital diligently accumulates dengue patients' daily and weekly data upon admission when a dengue infection is suspected and reports them to the Ministry of Health of Sri Lanka.[2 5] The National Dengue Control Unit maintains national notification datasets collected from the Ministry of Health, which further ensures the dataset's reliability for analytical purposes.

The National Dengue Control Unit, in adherence to strict ethical guidelines, has anonymized all patients' data upon release of data to the principal investigator. This process

involves removing personally identifiable information, such as names and addresses, and replacing them with unique identifiers. The principal investigator obtained permission from the Director of the National Dengue Control Unit of the Ministry of Health, Sri Lanka. This meticulous process ensures the integrity of the study and the protection of patients' privacy. The anonymized dataset in Microsoft Excel is attached in the supporting information section of this paper.

## Ethics Approval

The University of Alberta Research Ethics Board officially approved the study on June 6, 2024, under the study identification number Pro00142813. This official approval, a significant milestone in our research journey, further validates the study's legitimacy and ethical soundness, instilling confidence in the reader about the robustness of our work.

## Methodology

### Time Series Properties

A time series can be analyzed independently without any information from covariates.[24 26 5 27] Time series comprises three properties: autocorrelation, non-seasonality, and non-stationarity.[25 28] When working with time series data, it is essential to analyze the autocorrelation of the data to understand the pattern and the dependencies between each time lag.[25] Autocorrelation means that each subsequent data point in the Y variable is correlated to the previous one, which gives it dependency and linearity.[28 27] The Auto-Correlation Function (ACF) estimates the correlation between each observation and previous values at various lags, where a lag is the number of time points between an observation and its prior values.[25 28] The companion to the ACF is the Partial Auto-Correlation Function (PACF), which estimates the correlation between an observation and past values that is not explained by correlations at lower-order lags.[28]

ACF and PACF plots can identify the difference in the autocorrelation of successive data points in the Y variable, which falls within the acceptable threshold of a 95% confidence interval.[25] [28] [26] Lags by months are plotted on the X axis, and each value corresponding to the lag is plotted on the Y axis. The autocorrelations within the dotted blue line bounds in the ACF and PACF lag plots can be considered insignificant, which will not affect the time series prediction.[28] [20] ACF and PACF plots visually represent autocorrelations between lags in time series data. PACF and ACF determine the Auto-Regression (AR) and Moving Average (MA) orders in manually selected ARIMA models.[26] [5] [25] [28] [29] Further, a "white noise series" suggests the time series data is stationary.[20]

Seasonality refers to the periodic data pattern along observations, a fixed or known frequency variation occurring at regular intervals, such as the time of year or the day of the week. Seasonality in time series of health data is typical and can be due to natural causes, such as weather patterns.[28] For example, in our study, dengue cases are seasonally recorded in the rainy months. ARIMA models usually deal with seasonality by taking the seasonal difference, which calculates the difference between each observation and the previous value.[28] [29]

ARIMA modeling requires the data to be stationary, meaning the data should have a constant mean and variance.[28] [25] [26] [29] [20] A stationary series is also called a white noise series.[28] Changing variance over time, also known as heteroscedasticity, can be controlled by log-transforming the data.[28] If the data has an increasing or decreasing trend, stationary differencing can control for the trend.[28] [25] [29] Also, another test on R to check if the data is stationary to perform a time series is the Augmented Dicky Fuller (ADF) test.[26] [25] Typically, the ADF test's value is presented in a *p*-value. The ADF test hypothesizes ($H_1$) that a *p*-value for stationary time series data will be $< 0.05$. A *p*-value of $\geq 0.05$ hypothesizes ($H_0$) that the time series data is non-stationary

(Table 1).[26]

**Table 1.** Stationarity hypothesis with the ADF test in R software.

| **ARIMA Function** | ***p*-Value** | **$H_0$ – Non-stationary** | **$H_1$ - Stationary** |
|---|---|---|---|
| ADF Test | ≥ 0.05 | Fail to Reject | Reject |
| ADF Test | < 0.05 | Reject | Fail to Reject |

Among the coefficients generated from the automated ARIMA model, the Akaike Information Criterion (AIC) and the Akaike Information Criterion-Corrected (AICc) are indicators that will yield their minimum values for the best model selection in an iterative process.[5 25 6 26 29] We validate the automated ARIMA model using the Ljung-Box Test of model residuals. The random errors derived from lagging and differencing in the time series analysis are critical to eventually validating the prediction model.[28 26 27 30] Suppose the *p*-values for extended lags of the Ljung-Box Test are > 0.05. In that case, the automated ARIMA model fits the all-island dengue time series data because no significant autocorrelations in the residuals affect the model.[28 26 27 30] On the other hand, if the Ljung-Box Test *p*-values < 0.05, the ARIMA model will lack fit for the time series because the autocorrelations in residuals significantly affect the model.[28 26 27 30] Thus, the null hypothesis ($H_0$) for the ARIMA modeling will be that if the *p*-values > 0.05, the model shows a significant fit, while the hypothesis ($H_1$) will be that if the *p*-values < 0.05, the model shows a significant lack of fit (Table 2).[28 26 27 30 20]

Further, ACF and PACF plots of the model's residuals can also explain if the model has significant error lags.[24 28 29 3 20] Suppose the ARIMA residuals fall within the dotted blue line bounds of the 95% confidence intervals of the ACF and PACF plots. In that case, we can conclude that the ARIMA model's errors are insignificant.[28 29 5 24 3 20] Furthermore, we validate the ARIMA model's residuals using a formal test, the Shapiro-Wilk test, to determine if they are normally

distributed.[27 31]

**Table 2.** Ljung-Box test of ARIMA model validation.[26 28 27 30]

| **ARIMA Model Fit** | ***p*-Value** | **$H_0$- No Significant Lack of Fit** | **$H_1$- Significant Lack of Fit** |
|---|---|---|---|
| Box-Ljung Test | > 0.05 | Fail to Reject | Reject |
| Box-Ljung Test | < 0.05 | Reject | Fail to Reject |

We perform the automated ARIMA model's accuracy analysis using the predictions' Mean Absolute Percentage Error (MAPE) compared with the actual data in a test data component.[4 31 26 25 6] MAPE is a statistical measure that helps determine the accuracy between the predictions and the actual values. Gauging the model's performance, such as time series analysis, is crucial in forecasting models.[4 26 25] MAPE offers a measure to express the error as a percentage, making interpreting and communicating the model's accuracy easier[4 31 26 25] MAPE determines forecast accuracy as values <10 highly accurate forecasting, 10-20 good forecasting, 20-50 reasonable forecasting, and >50 inaccurate forecasting (Table 3).[31]

**Table 3.** Model's accuracy test with MAPE.[31]

| **MAPE** | **Interpretation** |
|---|---|
| **<10** | Highly Accurate Forecasting |
| **10-20** | Good Forecasting |
| **20-50** | Reasonable Forecasting |
| **>50** | Inaccurate forecasting |

**ARIMA Model's Mathematical Components**

ARIMA models have a single dependent variable that is a function of past values of the variable and the error term.[28 5 24] Further, ARIMA models assume that the errors generated by lagging and differencing are non-significant and can accommodate continuous outcomes.[28] ARIMA's components—AR, Integrated (I), and MA—can explain it.[28] Seasonal and non-seasonal ARIMA models are similarly comprised of these components.[4 28 26 21]

- AR component: $Y_t$ is predicted by one or multiple lagged values of $Y_t$. This is represented by the equation below, where $c$ is a constant, $\varphi$ is the autocorrelation coefficient, $p$ is the number of lags, and $\varepsilon_t$ is the error.[28 25]

  $$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

- MA component: $Y_t$ is predicted by one or multiple lagged error values, $\varepsilon_t$. In the equation, $\theta$ is the coefficient of the autocorrelation of the errors, and $q$ is the number of lags.[28 25]

  $$Y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- I component: Stationarity integration by differencing refers to calculating the difference between adjacent observations.[28 25]

  $$Y'_t = Y_t - Y_{t-1}$$

- Seasonal Model: $Y_t$ is predicted by lagged values of $Y_t$ at regular intervals throughout the season. In the equation below, $\Phi$ is the autocorrelation coefficient, and $s$ is the seasonality. Since dengue is a seasonal disease, we can presume a seasonal component in the ARIMA model.[28 25]

  $$Y_t = c + \Phi Y_{t-s} + \varepsilon_t$$

The basic notation for describing the parameters of a non-seasonal ARIMA model is ($p$, $d$, $q$),

where *p*, *d*, and *q* are positive integers.[28]

- *p* = The order of the AR part of the model.
- *d* = The degree of non-seasonal differencing.
- *q* = The order of the MA part of the model.

Further, if the ARIMA model has a seasonal component, the notations of the seasonal integers will be expressed in upper case.[28]

- *P* = The order of the AR part of the model.
- *D* = The degree of seasonal differencing.
- *Q* = The order of the MA part of the model.

A monthly ARIMA model with a non-seasonal and seasonal component can be illustrated as ARIMA(*p, d, q*) * (*P, D, Q*) [12].[4 30 28 21]

**Data Analysis**

**<u>Automated ARIMA Analysis</u>**

Relying on ACF and PACF plots can make selecting the most appropriate ARIMA model challenging, time-consuming, and subjective.[28] Identifying model orders is often not informative in such procedures.[28] Thus, we apply the automated algorithm *auto.arima()* in the *forecast* package of R software, which we chose for convenience and eventually for easy use by the National Dengue Control Unit of Sri Lanka, where dengue data is systematically compiled.[28] We use R version 4.4.1 in our data analysis. The automated algorithm, *auto.arima()*, in the *forecast* package, identifies the best ARIMA model to forecast monthly dengue data from January 2024 to December 2025. ARIMA is an automated iterative process that aims to select the optimum model.[26 29] Consequently, the time series model generated by the *auto.arima()* function in R software will be

autogenerated by identifying the essential (*p, d, q*) * (*P, D, Q*) [12] properties to be included in the time series model.[28 25 26]

**Data Transformation and Partition**

We analyzed the datasets in R software and found that the all-island dengue data fits a time series better than the sex and age-stratified datasets, as they each had only nine annually recorded data points.[28] The district-wise data is, in fact, the all-island data, and thus, we focus our time series analysis on the all-island dengue dataset.

First, we install the package under the command *install.packages* (*forecast)* on R Studio software, followed by the library for the forecast command, *library(forecast)*. Additionally, we open library commands to read the Microsoft Excel data sheets, *library(readxl),* and to analyze the dataset on the time series converted data, *library(tseries)*. Also, we open the *library(ggplot2)* to visualize the results of the time series analysis by plotting them in R Studio. Subsequently, we downloaded the dataset from the Microsoft Excel file to R studio under the *library(readxl)* command for all island dengue cases reported to the national surveillance system from 2015 to 2023, provided by the National Dengue Control Unit of Sri Lanka. The data analysis is explained step-by-step, and the R script analysis is provided for reference.

We rename the all-island dengue data Excel file *All_Island_Data* for easy reference for subsequent R commands. We then view the *All_Island_Data* using the *View(All_Island_Data)* command on R Studio to establish that it is a data frame. To perform a time series analysis on an Excel data frame, we convert it to time series data points by naming the data as *AllTS*, meaning that the time series data conversion is for the *All_Island_data*. Subsequently, when we check the class of *AllTS* with the code *class(AllTS),* the console window of R Studio derived the result that the *All_Island_data* is now a time series, *"ts."*

We plot the *AllTS*, of which the X axis plots the months under investigation from January 2015 to December 2023, and the Y axis plots the total number of dengue cases (Figure 1). The data has seasonality and a cyclic trend, with outliers in 2017 and 2019. The 2017 spike relates to the unprecedented dengue outbreak due to the DENV-2 variant, although the seasonal monsoon pattern was typically the same during that year.[2] We generate a box plot (Figure 2) to visualize the outliers and the non-constant time series variance due to cyclic monsoon patterns affecting the dengue case numbers, where certain rainy months record higher case numbers than others.

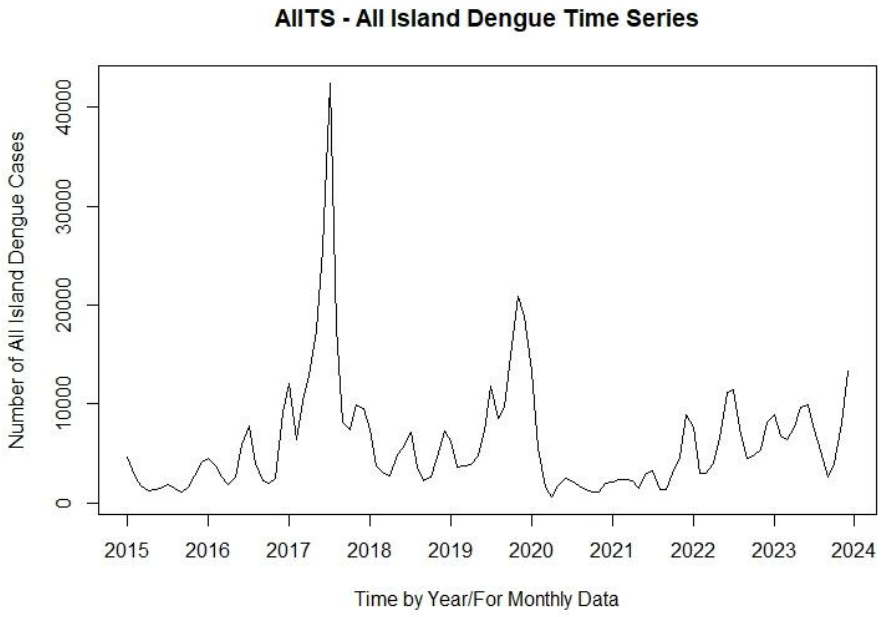


**Figure 1.** Time series plot for all-island dengue data from January 2015 to December 2023.

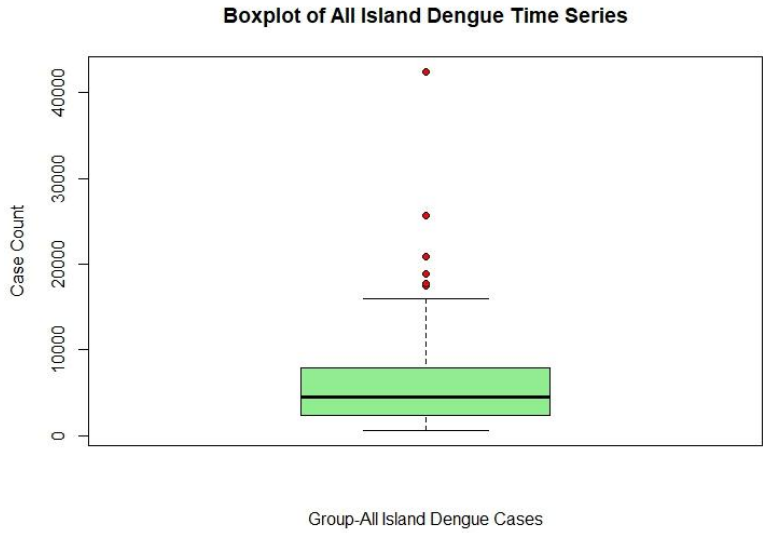


**Figure 2.** Boxplot of the all-island dengue time series.

We log-transform the *AllTS* data into *AllTS_Log* to attain a constant variance and manage the outliers and seasonality since we visualize a non-constant variance in the time series.[31] [28] [4] We re-generate a box plot of the log-transformed time series data to visualize that the outliers are under control (Figure 3).

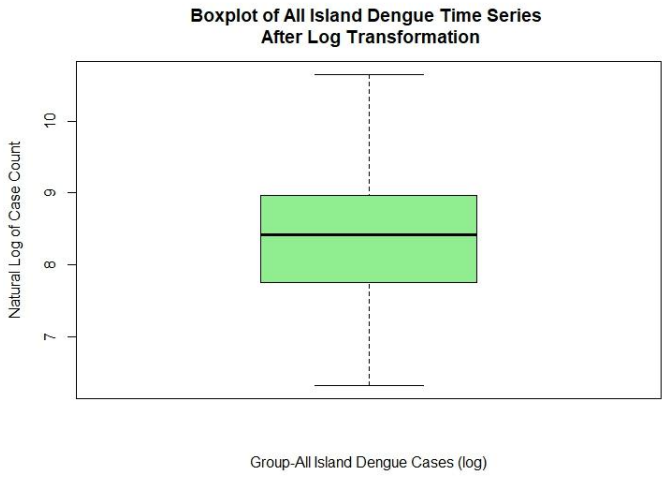


**Figure 3.** Boxplot of the log-transformed all-island time series data.

We plot the log-transformed time series to visualize the stabilized variance (Figure 4). All subsequent data analyses will be based on the log-transformed dataset until exponentiated to arrive at the original scale for case number prediction for 2024 and 2025.

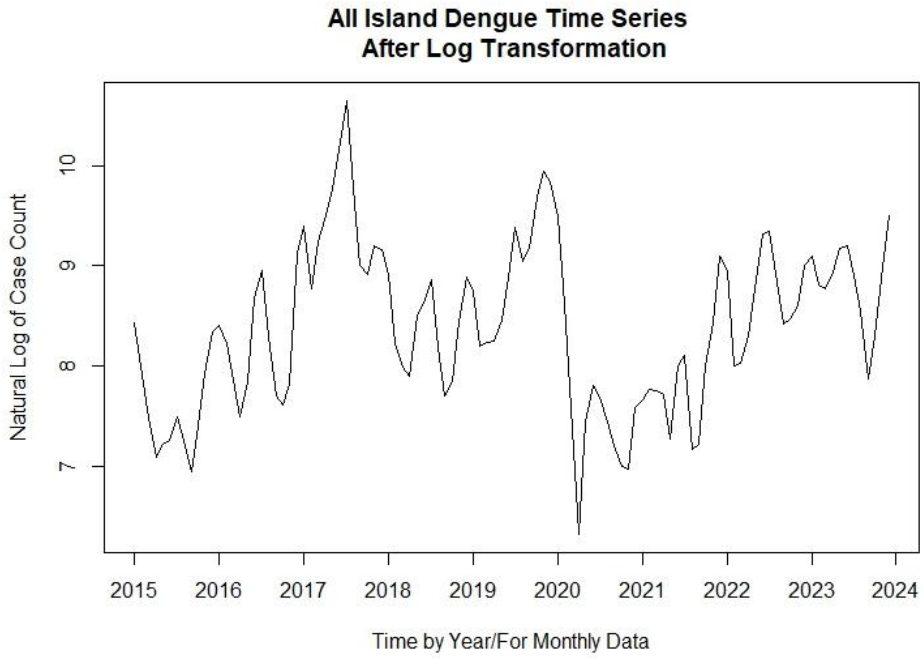


**Figure 4.** Log transformed all-island dengue time series.

Subsequently, we partition the data series into a training and a test set.[4 21 25 6 31] The training set comprises 80% of the data from 2015 to 2021, and the test set is 20% from 2022 to 2023. The main difference between training and test sets is that the training dataset is used to train the ARIMA model, and the test dataset is the hidden data from the model during training.[4] The training data window had 84 months of observations, and the remaining 24 months were the test set used to verify the model's accuracy.[4 21] The training and test components comprise log-transformed time series data values. We named the training window *All_Train* and the test window *All_Test*.

We check the order of the autocorrelation using the ACF (Figure 5) and PACF (Figure 6) on R to analyze if the *All_Train* time series needs lagging to adjust for stationarity.[26 25 28] When

we visualized the autocorrelation and partial autocorrelation in the ACF and PACF lag plots, it was evident that the training dataset had significant autocorrelation that exceeded the threshold 95% confidence interval.

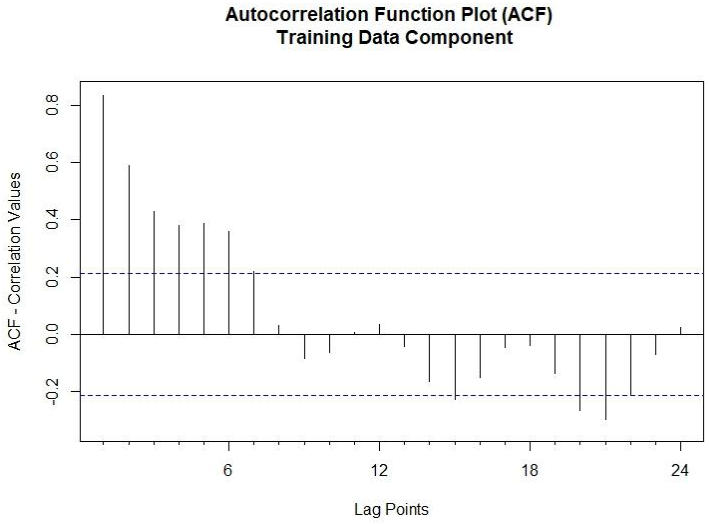


**Figure 5.** ACF plot of the training data component.

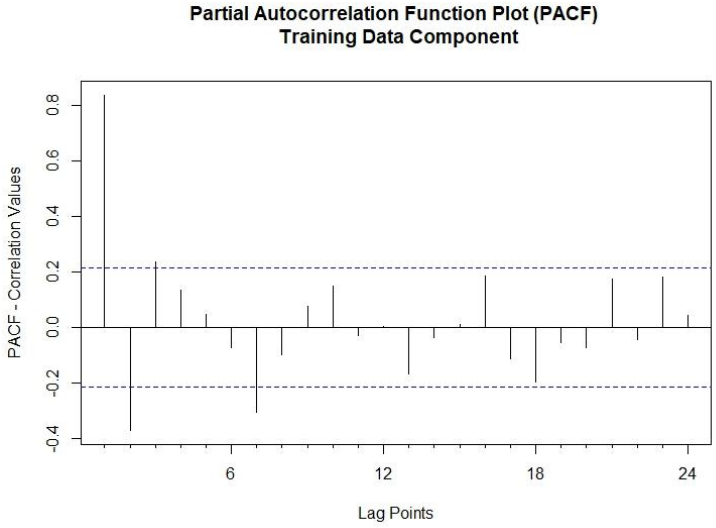


**Figure 6.** PACF plot of the training data component.

## Results

The ADF test on the training dataset derives a *p*-value of 0.5235, indicating that it is a non-stationary time series. We then apply a differencing of 1 to the training data component to determine if a first-order differencing would make the data stationary. The ADF test derives a *p*-value of 0.01, indicating that a first-order differencing will make the data stationary.[26 28]

Plotting the differenced training data shows a pattern in the early months of almost every year in the shape of the letter "M." (Figure 7). We generate ACF (Figure 8) and PACF (Figure 9) plots of the training data component to check if they have a similar pattern. We can visualize the seasonality in the ACF and PACF plots of the training data component, which alternates every three lags due to heavy and low rainfall patterns. This coincides with the shape of the letter "M" in Figure 7.

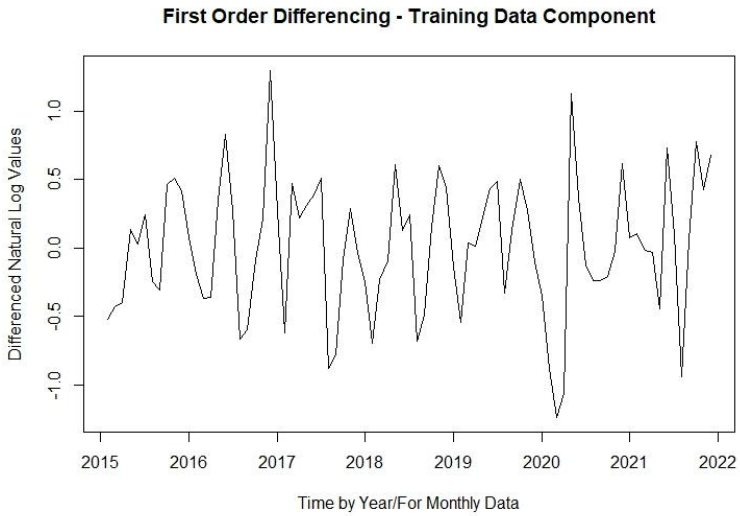


**Figure 7.** Mean and variance of the training dataset with the first-order differencing.

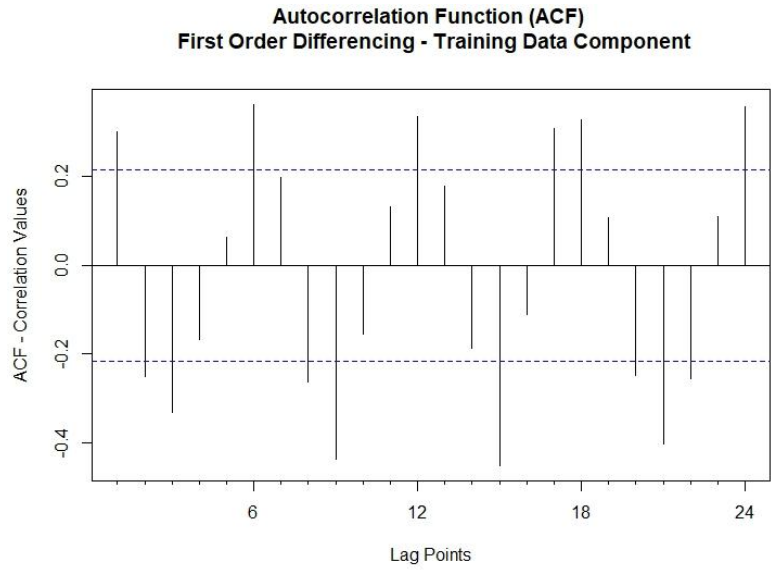


**Figure 8.** ACF plot of the time series training data component with the first-order differencing.

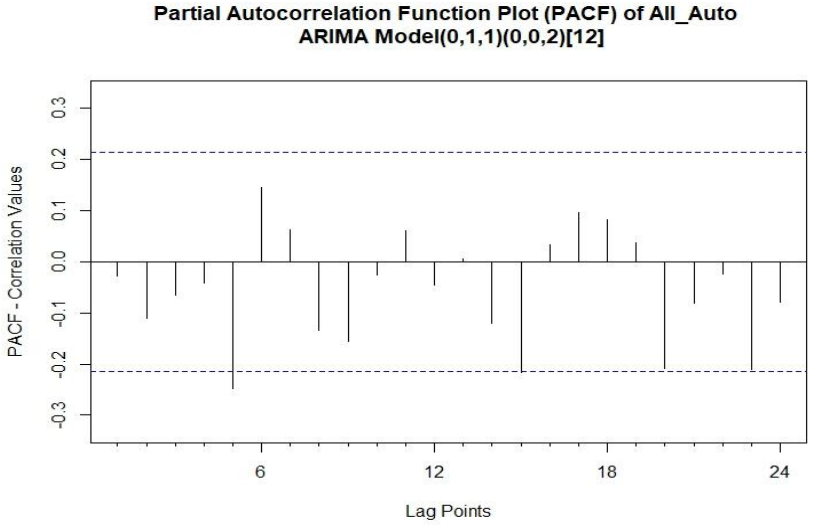


**Figure 9.** PACF plot of the time series training data component with the first-order differencing.

However, we initially tested the *auto.arima()* function in R software to determine the *p*, *d*, and *q* parameters on the training dataset to derive a suitable ARIMA model to test the prediction accuracy of the test data component, which records actual values. The model generated by *auto.arima()* is a seasonal ARIMA model with a non-seasonal and seasonal component. The model reads as ARIMA(1,0,1)(2,0,0) [12] with non-zero mean (Box 1). We also notice that the *d* and *D*

parameters for differencing essential for stationarity in the time series are zero. The ADF test on the training data component indicates that the data is non-stationary, even though *auto.arima()* has suggested an ARIMA model with a mean of non-zero value, to balance the *d* parameter being zero. Thus, we experiment with the *auto.arima()* function with a different specification. We apply the *auto.arima()* function again on the training dataset with a manually added differencing in the non-seasonal *d* component with a value of 1.[28] We did that to ensure the forecast's accuracy, as the ADF test at the start of the analysis with a first-order differencing made the data stationary.[28] The modified *auto.arima()* subsequently generates an ARIMA model with non-seasonal and seasonal components, ARIMA(0,1,1)(0,0,2)[12], on the training data (Box 2). We notice that the value-specified *d* parameter replaced the mean applied by *auto.arima()* in the previous model. In the new model, the non-seasonal component is adjusted for stationarity. The lowest value of AICc is 93.34 for ARIMA(0,1,1)(0,0,2)[12] compared to other auto-generated ARIMA models. We save the result as *All_Auto* in the R script for the modified model.

**Box 1.** Automated ARIMA coefficients for the model ARIMA(1,0,1)(2,0,0)[12] with non-zero mean.

```
 Best model: ARIMA(1,0,1)(2,0,0)[12] with non-zero mean

Series: All_Train
ARIMA(1,0,1)(2,0,0)[12] with non-zero mean

Coefficients:
         ar1     ma1    sar1    sar2    mean
      0.8170  0.4483  0.1319  0.4640  8.0837
s.e.  0.0651  0.0953  0.0944  0.1184  0.5738

sigma^2 = 0.1402:  log likelihood = -38.36
AIC=88.72   AICc=89.81   BIC=103.3
```

**Box 2.** Automated ARIMA coefficients for the modified *All_Auto* model ARIMA(0,1,1)(0,0,2)[12].

```
 Best model: ARIMA(0,1,1)(0,0,2)[12]

> All_Auto
Series: All_Train
ARIMA(0,1,1)(0,0,2)[12]

Coefficients:
          ma1     sma1     sma2
      0.4530  0.2715  0.5816
s.e.  0.0982  0.1168  0.1960

sigma^2 = 0.1491:  log likelihood = -42.41
AIC=92.82   AICc=93.34   BIC=102.5
```

After running the modified auto ARIMA on the training data component, we plot the residuals of the *All_Auto* model (Figure 10). We can visualize that the model's residuals do not reflect the recurring " M " pattern present earlier in the analysis, meaning heteroscedasticity is controlled for the model's residuals (Figure 7). We run the ADF test on the *All_Auto* residuals to achieve a *p*-value of 0.01, indicating that the model's residuals are stationary, meaning white noise.

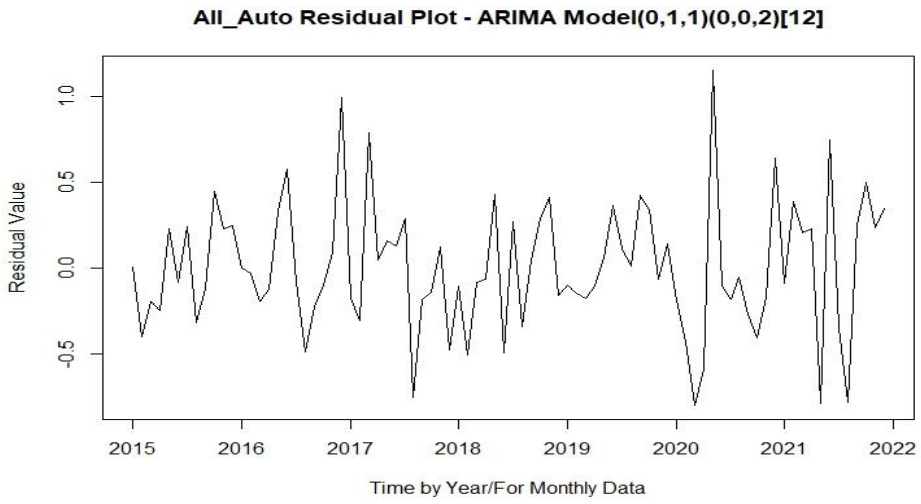


**Figure 10.** ADF test of residuals of the modified automated *All_Auto* model ARIMA(0,1,1)(0,0,2)[12].

We check the model for residuals on ACF (Figure 11) and PACF (Figure 12), which estimates whether the model has non-significant errors.[25 28 26 3] Besides two spikes at lags 5 and 15 for 24 lags in the ACF plot and two spikes at lags 5 and 15 for 24 lags in the PACF plot, the residuals for the *All_Auto* model are within the 95% confidence intervals−blue dotted line bounds, further confirming that the residuals are white noise.

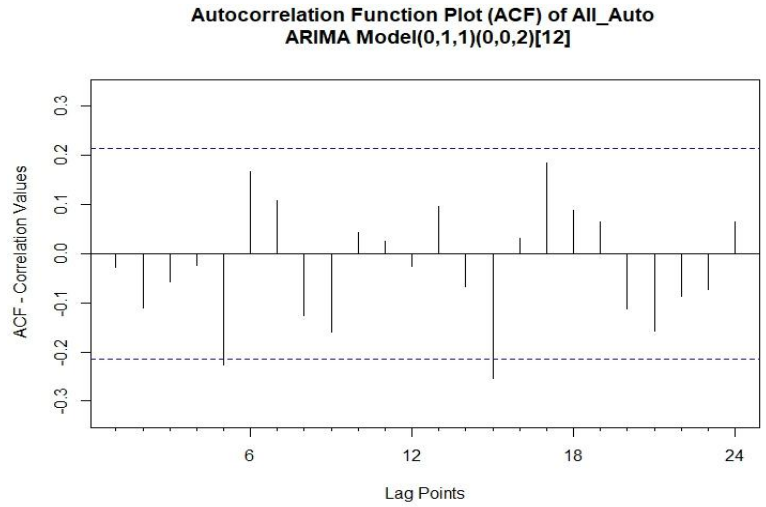


**Figure 11.** ACF plot of *All_Auto* model ARIMA(0,1,1)(0,0,2)[12] residuals.

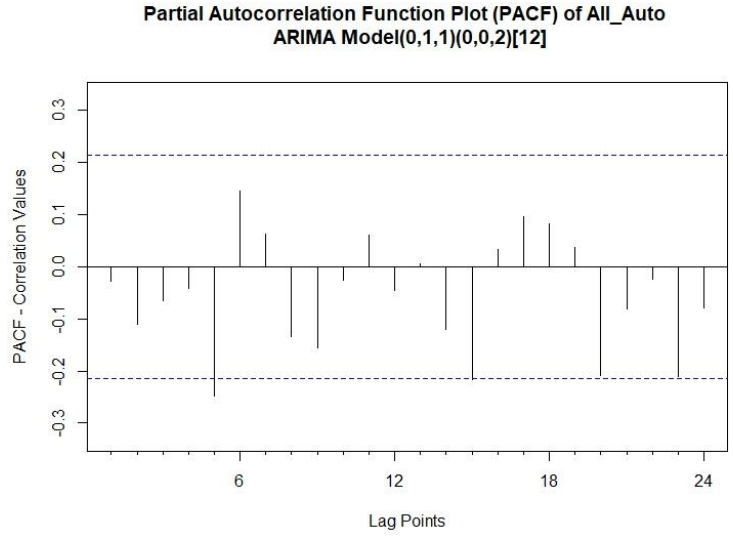


**Figure 12.** PACF plot of *All_Auto model* ARIMA(0,0,1)(0,0,2)[12] residuals.

Further, we validate ARIMA(0,1,1)(0,0,2)[12] for the lag of errors to determine whether the errors generated by the model are significant.[30 26 28 27] If the model is a good fit for the data, the errors should be non-significant, meaning the lag of errors should have a *p*-value > 0.05.[26 28 27 30] We perform the Ljung-Box test to validate the ARIMA(0,1,1)(0,0,2)[12] model and estimate that the residuals' *p*-values are > 0.05. If the residual *p*-values are > 0.05, we can conclude that the ARIMA(0,1,1)(0,0,2)[12] is a good fit for the prediction.[30 28 26 27] ACF and PACF at two lags in each plot present spikes that touched the blue dotted line bounds. Thus, we take 20 lags in the Ljung-Box test to ensure the errors are non-significant over 20 error lags (Table 4).[27] The *p*-values range from 0.091 to 0.83, indicating that the *p*-values of errors are > 0.05 and insignificant (Figure 13). This validates the ARIMA(0,1,1)(0,0,2)[12] model as a good fit for the time series.

**Table 4.** Table of *p*-values of ARIMA(0,1,1)(0,0,2)[12] model.

| Lag | *p*-Value | Lag | *p*-Value |
|---|---|---|---|
| **1** | 0.79963455 | **11** | 0.22868175 |
| **2** | 0.57067982 | **12** | 0.29151221 |
| **3** | 0.70258094 | **13** | 0.30244278 |
| **4** | 0.83239143 | **14** | 0.34172518 |
| **5** | 0.29017571 | **15** | 0.09969881 |
| **6** | 0.18955137 | **16** | 0.13006636 |
| **7** | 0.19879602 | **17** | 0.07249092 |
| **8** | 0.18259132 | **18** | 0.07973530 |
| **9** | 0.12844894 | **19** | 0.09539596 |
| **10** | 0.17231324 | **20** | 0.09080434 |

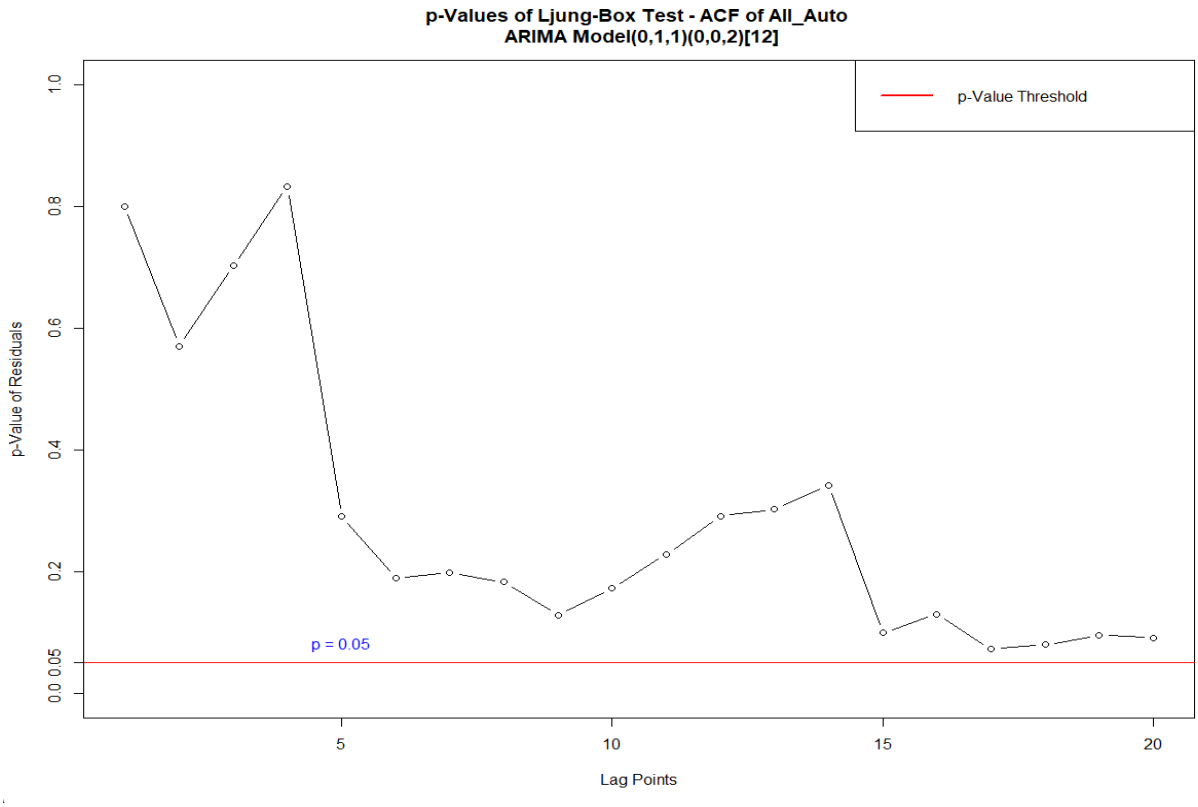


**Figure 13.** Ljung-Box test *p*-value plot of the *Auto_Model* ARIMA(0,1,1)(0,0,2)[12] residuals.

Our next step is to forecast the dengue all island cases for 2022 to 2023 on the test data to check for accuracy. The R code for ARIMA(0,1,1)(0,0,2)[12] model is given 80% and 95% prediction intervals and a 24-month frequency to predict from January 2022 to December 2025. We plot the *All_Auto_Forecast* on the test data component. The prediction results are illustrated in blue, juxtaposed with the original test data component in red (Figure 14).

Subsequently, we check the prediction accuracy percentage with the MAPE by comparing the predicted test data values of the *All_Auto_Forecast* with the actual test data component values. MAPE results in 4.46, where the expected values closely match the test data component, presenting a successful prediction (Box 3).[26 28 25]

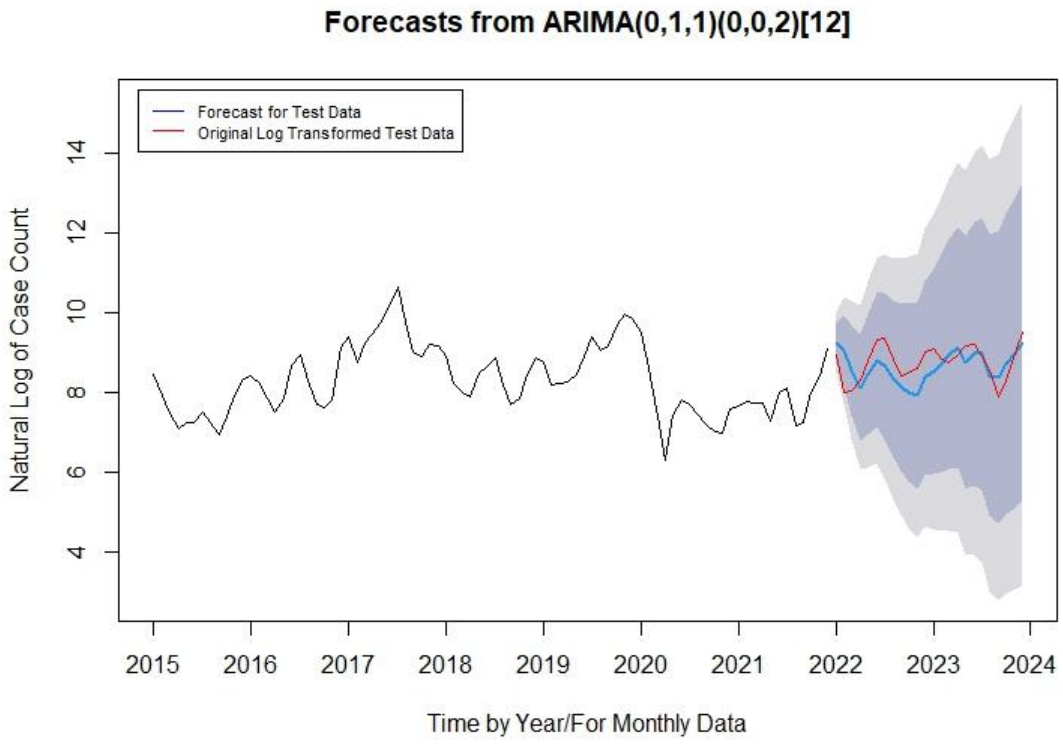


**Figure 14.** Forecast from *Auto_Model* ARIMA(0,1,1)(0,0,2)[12] on the test data component.

**Box 3.** *All_Auto_Forecast's* MAPE forecasting accuracy test.[4 31 26 25 6]

```
> accuracy(All_Auto_Forecast,All_Test)
                     ME      RMSE       MAE         MPE     MAPE      MASE        ACF1
Training set 0.005035383 0.3768254 0.2935571 -0.03541291 3.621964 0.2875969 -0.02720672
Test set     0.115344669 0.4558979 0.3866935  1.15595633 4.463851 0.3788424  0.64525933
             Theil's U
Training set        NA
Test set      1.110769
```

We name ARIMA(0,1,1)(0,0,2)[12] as the *Best_Model.* The *Best_Model* distributes the residuals normally (Figure 15 and Figure 16).[27] The boxplot of residuals (Figure 17) also shows that besides the two outliers coinciding with the two significant spikes in each of the ACF and PACF lag plots (Figure 11 and Figure 12) that touched the blue-dotted bounds, all other residuals are within the range of insignificance.[27] We formally run the Shapiro-Wilk normality test on the *Best_Model*'s residuals and derive a *p*-value of 0.2896, indicating that the residuals are

insignificant.[27]

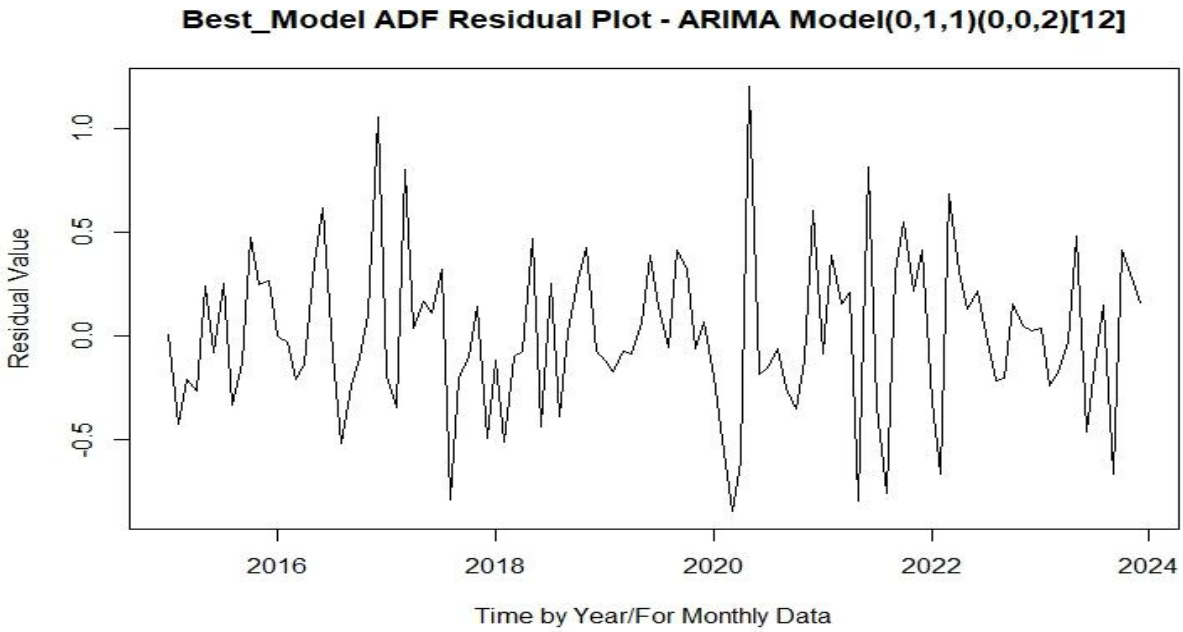


**Figure 15.** ADF test residual plot of the *Best_Model* ARIMA(0,1,1)(0,0,2)[12].

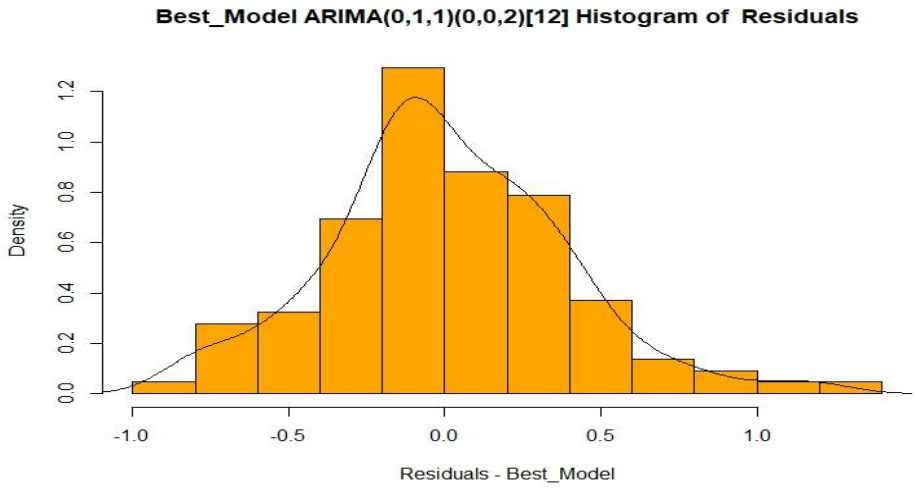


**Figure 16.** Histogram of residuals of the *Best_Model* ARIMA(0,1,1)(0,0,2)[12].

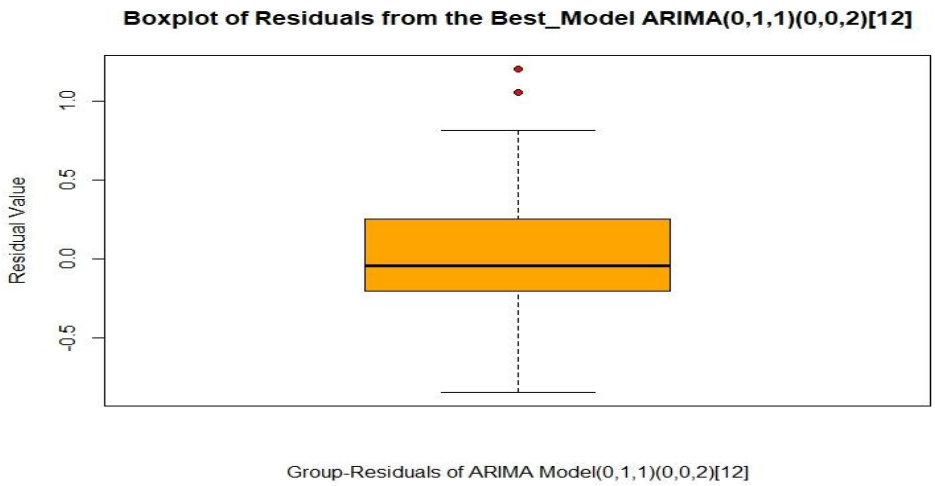


**Figure 17.** Boxplot of residuals of the *Best_Model* ARIMA(0,1,1)(0,0,2)[12].

We select the *Best_Model* ARIMA(0,1,1)(0,0,2)[12] to generate the prediction for dengue cases from January 2024 to December 2025 with 80% and 95% prediction intervals. Subsequently, we exponentiate the log-transformed values to arrive at the original data scale reflecting real-time dengue case numbers (Box 4).

**Box 4.** Predicted dengue case number from January 2024 to December 2025 by the *Best_Model* ARIMA(0,1,1)(0,0,2)[12].

```
> All_Forecast
         Point Forecast     Lo 80     Hi 80      Lo 95      Hi 95
Jan 2024       13694.847 8424.7801  22261.57 6514.22360   28790.67
Feb 2024        9189.059 3916.6319  21559.04 2493.77298   33859.86
Mar 2024       10038.954 3328.9197  30274.27 1855.80417   54305.62
Apr 2024       12657.380 3423.6757  46794.52 1713.52466   93496.92
May 2024       16046.033 3639.8649  70737.57 1659.67385  155136.01
Jun 2024       16888.519 3273.6293  87127.17 1373.47868  207663.99
Jul 2024       16194.505 2719.6912  96430.80 1057.64471  247967.96
Aug 2024       15177.067 2232.0348 103198.82  809.11272  284686.37
Sep 2024       11415.133 1482.8440  87875.23  503.35146  258875.32
Oct 2024       12136.409 1402.4930 105021.85  447.48921  329153.01
Nov 2024       14425.108 1491.9061 139475.09  448.86330  463579.31
Dec 2024       15843.847 1474.1581 170285.31  419.37068  598581.38
Jan 2025       16467.029 1330.7947 203760.25  351.37776  771713.78
Feb 2025       14972.681 1039.5189 215658.60  253.26600  885161.01
Mar 2025       13269.770  797.9799 220665.70  180.17795  977293.79
Apr 2025       12658.108  663.9439 241327.19  139.44874 1149007.89
May 2025       15412.394  709.4146 334842.12  139.04359 1708398.74
Jun 2025       13884.670  563.8250 341921.80  103.41566 1864167.01
Jul 2025       12042.484  433.4645 334563.57   74.58848 1944287.08
Aug 2025       12499.564  400.4933 390116.59   64.79789 2411175.54
Sep 2025        9708.973  277.9655 339121.82   42.37205 2224678.03
Oct 2025       10228.979  262.5842 398470.31   37.78114 2769424.47
Nov 2025       12440.831  287.2603 538794.48   39.07741 3960709.19
Dec 2025       14016.352  291.9509 672915.03   37.60711 5223962.34
```

Subsequently, we automatically plot the predicted dengue cases with the prediction intervals. The automated plot's zoomed-out version indicates the wide prediction intervals toward the end (Figure 18). However, the predicted numbers will be more explicit in a zoomed-in version of the automated plot (Figure 19). Finally, we illustrate the expected dengue case numbers without the prediction intervals for a more explicit visualization of our prediction (Figure 20).

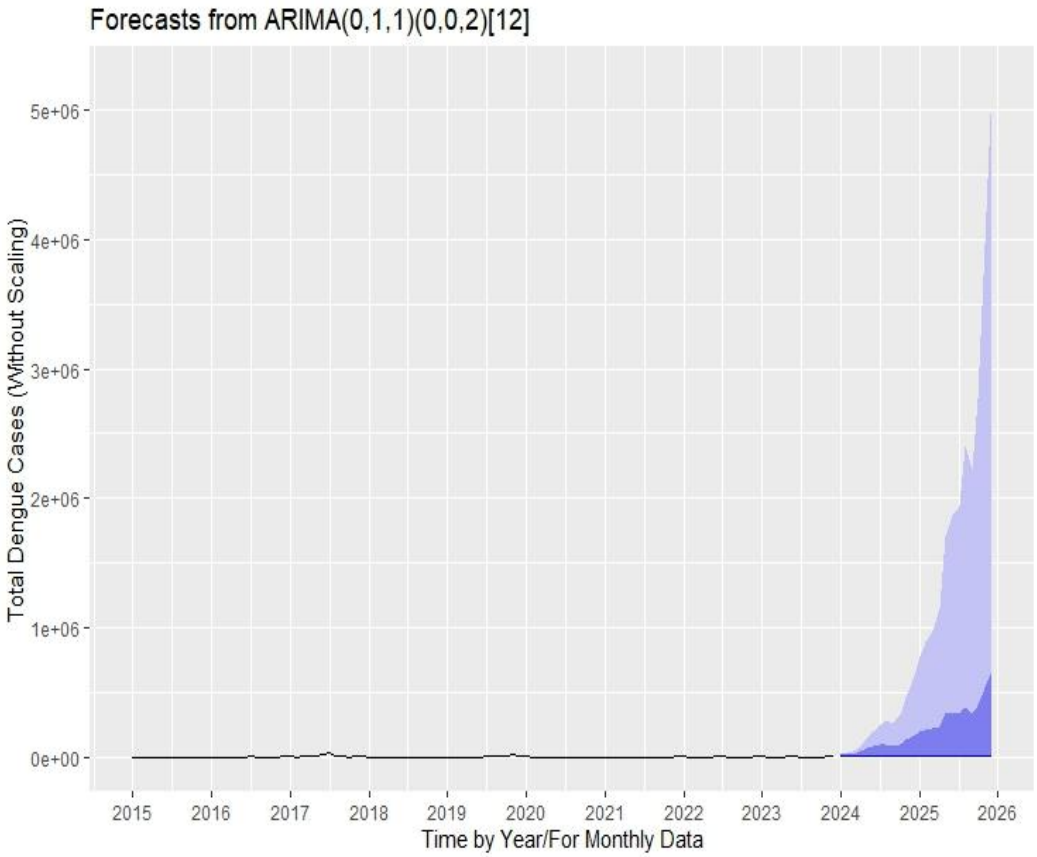


**Figure 18.** Forecast of dengue cases from the *Best_Model* ARIMA(0,1,1)(0,0,2)[12] from January 2024 to December 2025 (without scaling).

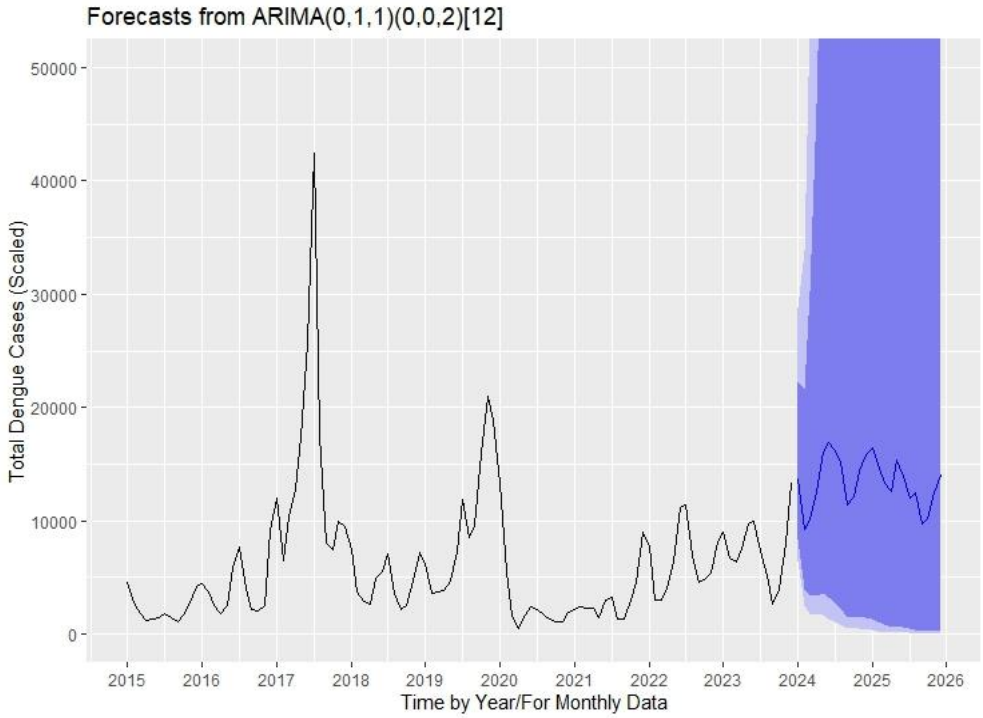


**Figure 19.** Forecast of dengue cases from the *Best_Model* ARIMA(0,1,1,)(0,0,2)[12] from January 2024 to December 2025 (scaled).

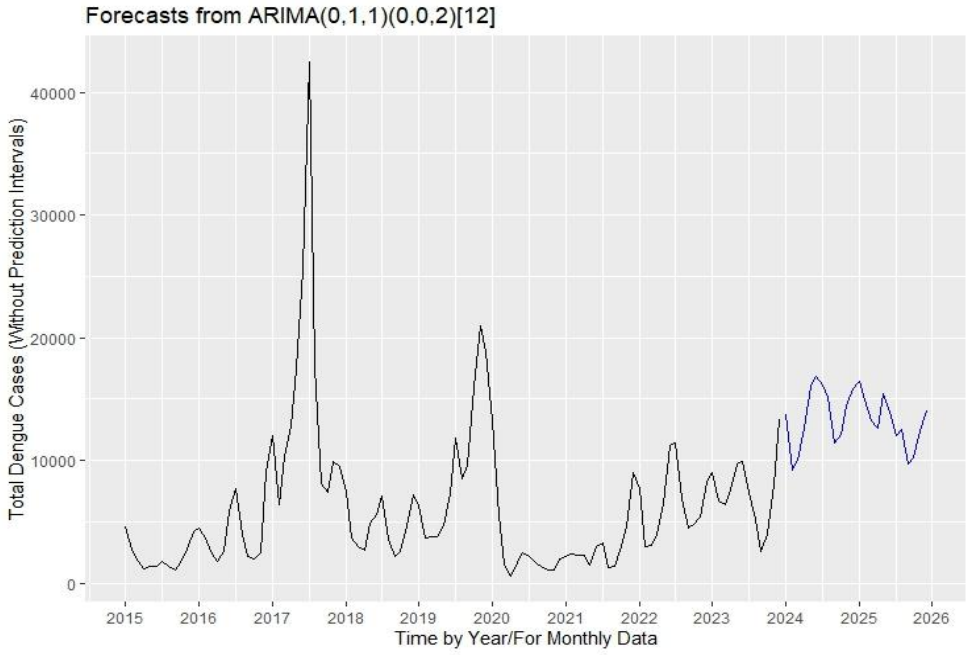


**Figure 20.** Forecast of dengue cases from the *Best_Model* ARIMA(0,1,1)(0,0,2) from January 2024 to December 2025 (without prediction intervals).

## Discussion

The study aimed to identify an ARIMA model to predict the island-wide dengue cases from 2024 to 2025 in Sri Lanka using the national surveillance notification data from 2015 to 2023 of the National Dengue Control Unit of Sri Lanka. The primary purpose of this study is to share the ARIMA model for use and replication within the National Dengue Control Unit for predictions in the imminent future with reported data during an outbreak. This assists in informing public health services about managing health resources in advance before the dengue situation becomes a crisis, as it often happens during monsoonal rainy seasons.

The potential impact of this novel concept is significant, as it will enable the National Dengue Control Unit, the pioneer government organization responsible for planning and mitigating crises, to calculate the expected case numbers through univariate analyses. This significant impact will empower the Dengue Unit to make informed decisions and take proactive measures to enhance its effectiveness in managing dengue outbreaks. The Ministry of Health of Sri Lanka, the primary government healthcare provider that distributes healthcare services free of charge across Sri Lanka, will gain lead time in determining the healthcare resource allocation during a dengue outbreak to avoid its resources running out. The potential of the ARIMA model to revolutionize dengue control instills hope and confidence in the healthcare industry.

Dengue, a vector-borne disease in Sri Lanka, is a pressing concern that demands immediate attention from health authorities. By consistently monitoring incidence trends, statistical analysis of surveillance data helps identify gaps in the reporting system and the peak period of disease incidence, thereby determining disease burden, which is where the ARIMA model comes into play. The early identification of case numbers, guided by data-driven optimization strategies, empowers health authorities to allocate resources efficiently in advance.[3] This proactive approach facilitated by the ARIMA model not only aids in effectively managing

health resources but also contributes to a more methodical and responsive healthcare delivery system during crises, instilling confidence in the system's preparedness.[3]

Tissera et al[2] reported that dengue cases in Sri Lanka exhibited the highest number in 2017 across all recorded years. Further, in the latter half of 2019, another dengue case elevation was reported. The unprecedented dengue case peak in 2017 is attributed to the DENV-2 strain that dominated the outbreak, as Malavige et al[1] suggested. Similarly, the increase in cases seen toward the latter months of 2019 coincides with the identification of the DENV-3 strain, which co-occurred with DENV-2 circulation.[1] The time series plot of all-island dengue data reflects the two peaks in Figure 1. Also, Figure 1 and Figure 4 show a notable decline in dengue cases in 2020 and 2021 due to the COVID-19 pandemic. This could have been due to the worldwide travel restrictions, which resulted in a slower upswing of dengue cases from mid-2020 to the start of 2021.

The outliers in the dataset affect data analyses and predictions.[4] As Plowiang suggested, outliers may cause a significant error in our prediction model. Two notable outliers coinciding with the 2017 and 2019 outbreaks appeared in the boxplot in Figure 2, along with several others. Like Polwiang's[4] suggestions, the outlier resulting in a non-constant variance may alter the model's accuracy. Thus, as Schaffer et al[28] suggested, log-transforming data assists us in managing the outliers by reducing the effects of a non-constant variance in the data analysis (Figure 3). Also, data partitioning is vital to train the ARIMA model to fit it to the test data component to determine the prediction accuracy, as Polwiang[4] and Baquero et al[21] suggested.

The ACF and PACF lag plots indicate that the training data component has significant spikes that cross the 95% confidence intervals, indicating high autocorrelation. Also, the ADF test's *p*-value of 0.5235 indicates a non-stationary time series in the training data component, which

requires differencing. However, with the *auto.arima()* function, we initially generated an ARIMA model in R software to test if the automated function would resolve the non-stationarity. The result was the ARIMA(1,0,1)(2,0,0) [12] with non-zero mean (Box 1), which indicated non-seasonal *d* and seasonal *D* parameters of zero value. The automated function balanced non-stationarity in the time series by applying a mean of a non-zero value.

However, according to Schaffer et al's[28] suggestions, we analyze the training data again with a specified first-order differencing in the *d* parameter, where the time series becomes stationary with a *p*-value of 0.01 in the ADF test, satisfying the test's alternate hypothesis. Furthermore, following first-order differencing, we can visualize the seasonality in the ACF (Figure 8) and PACF (Figure 9) plots of the training data component that alternates every three-month lag due to heavy and low rainfall patterns, which coincides with the shape of the letter "M" in Figure 7. Applying the first-order differencing has also adjusted the ACF plot's first lag correlation to a moderate value of approximately 0.3 from a significantly high value of 0.8 (Figure 5 and Figure 8), which affected all subsequent lags. Similarly, it adjusted the first lag's correlation value in the PACF plot from a high value of 0.8 to < -0.1 to fall within the 95% confidence interval range (Figure 6 and Figure 9); a plausible rationale for the ARIMA model to result in a zero for the AR orders of the *p* and *P* parameters in the non-seasonal and seasonal components. Afterward, the manually specified first-order differencing on the *d* parameter generated a model, ARIMA(0,1,1)(0,0,2)[12], with non-seasonal and seasonal components and the lowest AICc of 93.34 (Box 2).

ARIMA(0,1,1)(0,0,2)[12] model residuals on the ADF test are stationary with a *p*-value of 0.01, indicating the residuals are white noise, which validates the test's alternate hypothesis for residuals (Figure 10). Further, in the ACF (Figure 11) and PACF (Figure 12) residual plots, 22 of

the 24 lags fall within the white noise values. The fifth and the fifteenth lags touch and slightly exceed the 95% confidence interval marked by the dotted-blue lines in the ACF plot, and similarly, the fifth and fifteenth slightly do so in the PACF plot. Nevertheless, the Ljung-Box test vouches for the residuals to be insignificant by the *p*-values over 20 error lags (Table 4). The Ljung-Box test's error-lag plot in Figure 13 illustrates the *p*-values of residuals, which are > 0.05. Thus, the ARIMA(0,1,1)(0,0,2)[12] model is not a significant lack of fit for the dengue all-island training data component, validating our study's null hypothesis.

When we forecast the ARIMA(0,1,1)(0,0,2)[12], the *Best_Model*, on the test data component, as seen in Figure 14, we visualize that the predicted values closely follow the original test data values with a MAPE value of 4.46%. According to Moreno et al[31], our predictions are accurate. The *Best_Model*'s normal distribution pattern of residuals (Figure 16) further underscores the ARIMA model's reliability, as Cardoso et al[27] suggested, where the Shapiro-Wilk normality test derives a *p*-value of 0.2896, indicating that the residuals are insignificant. Also, the ADF test on the *Best_Model's* residuals is white noise (Figure 15).

Subsequently, we predict the dengue cases from January 2024 to December 2025 with the *Best_Model*, ARIMA(0,1,1)(0,0,2)[12], with 80% and 95% prediction intervals (Box 4). The prediction intervals of January, February, March, and April of 2024 give values with 95% prediction intervals between 6,514.22 and 28,790.67, 2,493.77 and 33,859.86, 1,855.80 and 54,305.62, and 1,713.52 and 93,496.92, respectively. However, from May 2024 to December 2025, the lower limit prediction interval decreases monthly, ranging from 1,659.67 to 37.61, while the upper limit prediction increases monthly, ranging from 155,136.01 to 5,223,962.34. As Polwiang[4] suggested, the significant outliers in the dataset in 2017 and 2019 increased the upper limit of the prediction intervals on a long-term basis (Figure 18), decreasing the

ARIMA(0,1,1)(0,0,2)[12] model's performance for a long-term prediction. However, the model is ideal for a short-term or a quarterly prediction, such as from January to April 2024, as the prediction intervals are comparatively not too wide.

Our research aimed to develop a suitable model for short-term prediction within the National Dengue Control Unit of Sri Lanka, with univariate analyses of dengue notification data readily available at the organization. The ARIMA(0,1,1)(0,0,2)[12] model we suggest thus achieves our purpose. Figures 19 and 20 illustrate the predicted dengue cases from January 2024 to December 2025 in auto-plots of the zoomed-in version.

**Limitations and Recommendations**

Schaffer et al[28] suggested that selecting AR and MA terms is an iterative process involving trial and error in formulating a robust ARIMA model. Automated algorithms in statistical packages can identify *p* and *P* and *q* and *Q* parameters that define the AR and MA terms.[28] ACF and PACF plots can also estimate the AR and MA terms manually.[28] However, estimating ACF and PACF may take time and warrant more in-depth knowledge of time series functions, a critical aspect to address within the organizations responsible for disease-related data analyses.

Automated functions typically generate errors in analyses. In our study, the stationarity assumption by the non-seasonal *d* parameter is satisfied by first-order differencing in the *auto.arima()* function. However, after the first-order differencing, the AR term that regulates the *p* parameter of autocorrelation results from one to zero, even though the PACF plot illustrates autocorrelation between lags up to the fifteenth (Figure 9). However, the lags after first-order differencing are weak to moderate, between 0.2 and 0.4 in correlation (Figure 8 and Figure 9). This ambiguity can make manual AR selection difficult. We can specify an AR term in the *auto.arima(),* which may be time-consuming as an iterative process generates it. Thus, a quick prediction could be difficult if the *p* or *P* parameters are manually specified in the computerized

function in an analysis during a crisis. Even though we use the R software algorithm to select the AR and MA terms, we must only consider that as a tool, not to derive a well-fitting model.

A training data component of only seven years, 2015 to 2021, with an 80/20 split of a 9-year dataset, provides a relatively short training period for an ARIMA model. Longer-term surveillance data can facilitate a more efficient outcome. According to Wagner and Cleland[29], there should be enough time points for highly seasonal data to identify seasonal effects and account for seasonal differencing. The all-island dengue time series in our study has 108 data points. Thus, we recommend reliable surveillance data of at least 200 data points for a time series over an extended period for more accurate analysis when accommodating 80/20 data split into training and test components.

The analysis does not explicitly address outlier handling, although the study compresses outliers by log transformation. The predicted case numbers from May 2024 to December 2025 result in wide prediction intervals due to these outliers, most pronounced by 2017 and 2019 upswings, decreasing the model's long-term performance, as Polwiang[4] suggested. However, from an epidemiological perspective, the DENV-2 and DENV-3 serotypes that dominated the two outbreaks in 2017 and 2019 account for the current island-wide dengue disease prevalence, from which we derive our predictions. Thus, we cannot eliminate these outliers to stabilize the variance, resulting in wider prediction intervals because, by doing so, our prediction will distort the accurate picture of dengue prevalence on the island.

The described case estimation process fits a model form of the order of autoregressive and moving average terms exclusively on all-island dengue cases, overlooking seroprevalence, which indicates a population's individual immunity or herd immunity. The same dengue serotype will not attack an individual twice, as Sirisena et al[10] suggested. Secondary or any subsequent dengue

infection will be from another serotype.[10] The study data does not provide the potential for a serological prediction clinically relevant to classify the infections into primary and subsequent groups.[14] Spreading an infectious disease through a population likely depends on previous infection counts. A subsequent dengue infection is more critical in individuals.[4 10 13] Non-classification of the first-time infection with subsequent infections can contaminate the case count relating to each. Thus, analyzing the data classified under the primary infected and those subsequently infected to predict each group's case counts and engage in more in-depth data analysis will be beneficial. For example, first-time dengue-infected cases are at a higher risk for three more dengue infections. In contrast, the subsequently infected cases may have a lower risk of getting infected, other than from the serotype that has not infected them, and a higher risk of infection severity. This classification will assist surveillance in assessing case counts of subsequent dengue infections and address higher healthcare costs.

Data on dengue serotypes circulating during an outbreak, such as classifying dominant-strain-infected cases and those infected from other strains, can be beneficial. Serotype surveillance determines if a different dengue variant is in circulation, which will elevate the chances for a more precise case prediction and identify case counts from the dominant and other serotypes. Suppose a new dengue variant strikes the population. In that case, our prediction will not effectively calculate the case numbers resulting in that attack. The outbreaks in 2017 and 2019 included all cases infected with DENV-1, DENV-2, and DENV-3 serotypes without a classification under each serotype infection. Our prediction is valid considering the DENV-1, DENV-2, DENV-3, and DENV-4 serotypes currently co-circulating and contributing to dengue epidemiology on the island. Although helpful in describing and discerning the observed patterns, fitting the ARIMA model on data not effectively classified with clinical relevance may limit its quality in specific contexts.

We base our study on secondary data collected from the National Dengue Control Unit. Although the prediction model identifies high-risk dengue elevation periods that coincide with weather and climatic changes, meteorological data is unavailable for this study. Meteorological variables directly affect mosquito breeding and distribution.[3 5 6] Thus, effective coordination between the National Dengue Control Unit and the Department of Meteorology of Sri Lanka for future predictive analyses will be beneficial, where weather variables can be added to bivariate and multivariate analyses to predict dengue cases.

**Conclusion**

We employed the National Dengue Control Unit's all-island dengue secondary data to derive our study's ARIMA model with a seasonal component. We test the ARIMA Model's validity using the Ljung-Box, Shapiro-Wilk, and ADF test error analyses on the model's residuals. The model is a good fit for the dengue all-island time series data. According to the MAPE estimation of 4.46, the model is accurate. ARIMA(0,1,1)(0,0,2)[12] estimated dengue case numbers from January 2024 to December 2025, but the estimate is more accurate for a short-term prediction. The National Dengue Control Unit can replicate the ARIMA model with existing data by following our study's R software analysis for an imminent case-count prediction. Sri Lanka, a country affected by dengue, will benefit from predictive analysis of dengue cases for effective and responsive public health resource management during dengue seasons from a public health resource management perspective.

**Supporting Information: All Island Dengue Cases from 2015 to 2023**

| Monthly Distribution ( All Island ) | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Year** | **Month** | | | | | | | | | | | | |
| | Jan. | Feb. | Mar. | Apr. | May | Jun. | Jul. | Aug. | Sep. | Oct. | Nov. | Dec. | **Total** |
| **2023** | 8963 | 6709 | 6419 | 7617 | 9696 | 9916 | 7369 | 5189 | 2605 | 4010 | 7995 | 13311 | 89799 |
| **2022** | 7702 | 2962 | 3040 | 4019 | 6483 | 11218 | 11437 | 6907 | 4527 | 4781 | 5416 | 8197 | 76689 |
| **2021** | 2122 | 2355 | 2312 | 2241 | 1441 | 2997 | 3292 | 1288 | 1370 | 2979 | 4561 | 8966 | 35924 |
| **2020** | 13272 | 5505 | 1592 | 551 | 1702 | 2454 | 2165 | 1700 | 1342 | 1089 | 1066 | 1973 | 34411 |
| **2019** | 6272 | 3629 | 3760 | 3804 | 4720 | 7290 | 11857 | 8526 | 9661 | 15991 | 20934 | 18857 | 115301 |
| **2018** | 7442 | 3707 | 2944 | 2677 | 4908 | 5568 | 7082 | 3592 | 2194 | 2545 | 4657 | 7216 | 54532 |
| **2017** | 12022 | 6442 | 10306 | 12810 | 17426 | 25610 | 42493 | 17690 | 8096 | 7423 | 9911 | 9536 | 179765 |
| **2016** | 4496 | 3718 | 2580 | 1797 | 2553 | 5868 | 7736 | 3978 | 2193 | 2013 | 2497 | 9156 | 48585 |
| **2015** | 4607 | 2732 | 1775 | 1192 | 1368 | 1408 | 1800 | 1410 | 1033 | 1645 | 2746 | 4176 | 25892 |

## References


1. Malavige GN, Jeewandara C, Ghouse A, Somathilake G, Tissera H. Changing epidemiology of dengue in Sri Lanka—Challenges for the future. *PLOS Neglected Tropical Diseases*. 2021;15(8):e0009624. doi:10.1371/journal.pntd.0009624

2. Tissera HA, Jayamanne BDW, Raut R, et al. Severe Dengue Epidemic, Sri Lanka, 2017. *Emerg Infect Dis*. 2020;26(4):682-691. doi:10.3201/eid2604.190435

3. Nayak SDP, Narayan KA. Prediction of dengue outbreaks in Kerala state using disease surveillance and meteorological data. *International Journal Of Community Medicine And Public Health*. 2019;6(10):4392-4400. doi:10.18203/2394-6040.ijcmph20194500

4. Polwiang S. The time series seasonal patterns of dengue fever and associated weather variables in Bangkok (2003-2017). *BMC Infectious Diseases*. 2020;20(1):208. doi:10.1186/s12879-020-4902-6

5. Karasinghe N, Peiris S, Jayathilaka R, Dharmasena T. Forecasting weekly dengue incidence in Sri Lanka: Modified Autoregressive Integrated Moving Average modeling approach. *PLOS ONE*. 2024;19(3):e0299953. doi:10.1371/journal.pone.0299953

6. Sharma V, Ghosh SK, Khare S. A PROPOSED FRAMEWORK FOR SURVEILLANCE OF DENGUE DISEASE AND PREDICTION. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. 2023;XLVIII-M-1-2023:317-323. doi:10.5194/isprs-archives-XLVIII-M-1-2023-317-2023

7. Kayesh MEH, Tsukiyama-Kohara K. Mammalian animal models for dengue virus infection: a recent overview. *Archives of Virology*. 2021;167(1):31. doi:10.1007/s00705-021-05298-2

8. ICD-11 for Mortality and Morbidity Statistics. Accessed January 21, 2024. https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/56575149

9. Ministry of Health NDCU. Dengue guidelines, for diagnosis, treatment, prevention and control. *Weekly Dengue Update*. 2024;4(1-12). Accessed October 30, 2024. https://www.who.int/publications/i/item/9789241547871

10. Sirisena PDNN, Mahilkar S, Sharma C, Jain J, Sunil S. Concurrent dengue infections: Epidemiology & clinical implications. *Indian J Med Res*. 2021;154(5):669-679. doi:10.4103/ijmr.IJMR_1219_18

11. Roy SK, Bhattacharjee S. Dengue virus: epidemiology, biology, and disease aetiology. *Can J Microbiol*. 2021;67(10):687-702. doi:10.1139/cjm-2020-0572

12. Sirisena PDNN, Noordeen F. Evolution of dengue in Sri Lanka—changes in the virus, vector, and climate. *International Journal of Infectious Diseases*. 2014;19:6-12.

doi:10.1016/j.ijid.2013.10.012

13. Shih HI, Wang YC, Wang YP, Chi CY, Chien YW. Risk of severe dengue during secondary infection: A population-based cohort study in Taiwan. *Journal of Microbiology, Immunology and Infection*. 2024;57(5):730-738. doi:10.1016/j.jmii.2024.07.004

14. Kalra C, Mittal G, Gupta P, Agarwal RK, Ahmad S. Role of IgM/ IgG Ratio in Distinguishing Primary and Secondary Dengue Viral Infections: A Cross-Sectional Study. *Cureus*. 16(8):e66714. doi:10.7759/cureus.66714

15. Bodinayake CK, Nagahawatte AD, Devasiri V, et al. Outcomes among children and adults at risk of severe dengue in Sri Lanka: Opportunity for outpatient case management in countries with high disease burden. Nacher M, ed. *PLoS Negl Trop Dis*. 2021;15(12):e0010091. doi:10.1371/journal.pntd.0010091

16. Weerasinghe NP, Bodinayake CK, Wijayaratne WMDGB, et al. Direct and indirect costs for hospitalized patients with dengue in Southern Sri Lanka. *BMC Health Serv Res*. 2022;22:657. doi:10.1186/s12913-022-08048-5

17. World Health Organization. Dengue Fact Sheet and Situation Report, 22 July 2022. July 22, 2022. Accessed October 24, 2023. https://www.who.int/srilanka/news/detail/22-07-2022-dengue-fact-sheet-and-situation-report--22-july-2022

18. Sri Lanka: Dengue Outbreak - May 2023 | ReliefWeb. October 20, 2023. Accessed October 24, 2023. https://reliefweb.int/disaster/ep-2023-000084-lka

19. Roster K, Rodrigues FA. Neural Networks for Dengue Prediction: A Systematic Review. Published online June 22, 2021. Accessed September 27, 2023. http://arxiv.org/abs/2106.12905

20. Abualamah WA, Akbar NA, Banni HS, Bafail MA. Forecasting the morbidity and mortality of dengue fever in KSA: A time series analysis (2006–2016). *Journal of Taibah University Medical Sciences*. 2021;16(3):448-455. doi:10.1016/j.jtumed.2021.02.007

21. Baquero O, Santana L, Chiaravalloti-Neto F. Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models. *PLOS ONE*. 2018;13:e0195065. doi:10.1371/journal.pone.0195065

22. Abeygoonawardena H, Dassanayake K, Kariyawasam J, et al. Identifying the strains of dengue circulating in the western province of Sri Lanka during 2019–2022. *PLOS Global Public Health*. 2024;4(7):e0003150. doi:10.1371/journal.pgph.0003150

23. Fernando L, Lokanathan S, Perera AS, Ghouse A, Tissera H. Annex 12: Improving Disease Outbreak Forecasting Models for Efficient Targeting of Public Health Resources. Published online 2018. https://idl-bnc-

idrc.dspacedirect.org/server/api/core/bitstreams/f035b268-c3e2-46d7-a8ae-210e8c8b51a5/content

24. Koh YM, Spindler R, Sandgren M, Jiang J. A model comparison algorithm for increased forecast accuracy of dengue fever incidence in Singapore and the auxiliary role of total precipitation information. *International Journal of Environmental Health Research*. 2018;28(5):535-552. doi:10.1080/09603123.2018.1496234

25. Ali N, Pazil N, Mahmud N, Jamaluddin S. Forecasting Dengue Outbreak Data Using ARIMA Model. *International Journal of Academic Research in Business and Social Sciences*. 2021;11. doi:10.6007/IJARBSS/v11-i6/10106

26. Ridwan M, Sadik K, Afendi F. Comparison of ARIMA and GRU Models for High-Frequency Time Series Forecasting. *Scientific Journal of Informatics*. 2024;10:389-400. doi:10.15294/sji.v10i3.45965

27. Cardoso F, Berri R, Giancarlo L, Borges E, Leite Dias de Mattos V. Normality tests: a study of residuals obtained on time series tendency modeling. *Exacta*. 2023;0. doi:10.5585/2023.22928

28. Schaffer AL, Dobbins TA, Pearson SA. Interrupted time series analysis using autoregressive integrated moving average (ARIMA) models: a guide for evaluating large-scale health interventions. *BMC Medical Research Methodology*. 2021;21:58. doi:10.1186/s12874-021-01235-8

29. Wagner B, Cleland K. Using autoregressive integrated moving average models for time series analysis of observational data. *BMJ*. Published online December 20, 2023:p2739. doi:10.1136/bmj.p2739

30. Mendoza A. Dengue Incidence Forecasting Model in Magalang Pampanga Using Time Series Analysis. Published online May 27, 2022. doi:10.2139/ssrn.4121536

31. Montaño Moreno JJ, Palmer Pol A, Sesé Abad A. Using the R-MAPE index as a resistant measure of forecast accuracy. *Psicothema*. 2013;(25.4):500-506. doi:10.7334/psicothema2013.23