Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Forecasting the incidence of dengue fever in Malaysia: A comparative analysis of seasonal ARIMA, dynamic harmonic regression, and neural network models

Nurakmal Ahmad Mustaffa [1,*], Siti Mariam Zahari [2], Nor Alia Farhana [2], Noryanti Nasir [2], Aishah Hani Azil [3]

[1]*School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, Sintok, Malaysia*
[2]*School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia*
[3]*Department of Parasitology and Medical Entomology, Faculty of Medicine, Universiti Kebangsaan Malaysia, Cheras, Malaysia*

ARTICLE INFO

Article history:
Received 7 August 2023
Received in revised form
11 December 2023
Accepted 18 December 2023

Keywords:
SARIMA
Dynamic harmonic regression
Neural network autoregressive
Forecasting
Dengue

ABSTRACT

Currently, no vaccines or specific treatments are available to treat or prevent the increasing incidence of dengue worldwide. Therefore, an accurate prediction model is needed to support the anti-dengue control strategy. The primary objective of this study is to develop the most accurate model to predict future dengue cases in the Malaysian environment. This study uses secondary data collected from the weekly reports of the Ministry of Health Malaysia (MOH) website over six years, from 2017 to 2022. Three forecasting techniques, including seasonal autoregressive integrated moving average (SARIMA), dynamic harmonic regression (DHR), and neural network autoregressive model (NNAR), were first fitted to the estimation part of the data. First, several SARIMA models were run, and the best seasonal model identified was SARIMA (0, 1, 2) (1, 1, 1)52. The best DHR model was obtained with a Fourier term of 2, as this corresponds to the lowest Akaike Information Criteria (AIC) value. The NNAR (9, 1, 6)52 was considered the best choice among the NNAR models due to its superior performance in terms of the lowest error measures. The comparison among the three techniques revealed that the DHR model was the best due to its lowest MAPE and RMSE values. Thus, the DHR model was used to generate future forecasts of weekly dengue cases in Malaysia until 2023. The results showed that the model predicted more than a thousand dengue cases around weeks 27 to 32. The results showed an increase in dengue cases after the end of the monsoon season, which lasted about five months. This technique is proving to be valuable for health administrators in improving preparedness.

© 2023 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In the past fifty years, the yearly incidence of dengue fever has increased by thirty percent, with between fifty and one hundred million infected and over half a million requiring hospitalization (Aguiar et al., 2022). The infectious dengue virus, which causes dengue fever, is spread by the female *Aedes aegypti* and *Aedes albopictus* mosquito species. Individuals between the ages of 15 and 49 are the most susceptible, regardless of gender, race, or ethnicity (Tantawichien, 2012). People with dengue fever are more likely than others to have fevers of up to 40 degrees Celsius, severe headaches, body aches, and rashes (Htun et al., 2021). Prior dengue sufferers, pregnant women, and infants are at a greater risk of acquiring severe dengue. Due to the likelihood of severe instances resulting in dengue hemorrhagic fever (DHF), dengue shock syndrome (DSS), internal bleeding, and possibly death, the dengue virus is considered a life-threatening disease (Wang et al., 2020). Dengue illness is associated with the four serotypes of the virus and corresponds to four distinct epidemiological trends. The four dengue virus serotypes, DENV-1, DENV-2, DENV-3, and DENV-4, are members of the Flaviviridae virus family. These serotypes can coexist in a particular location, and many countries are hyper-endemic for all four serotypes. Infected travelers frequently contribute to the geographical spread of the dengue virus from one site to another. According to the

* Corresponding Author.
Email Address: nurakmal@uum.edu.my (N. A. Mustaffa)
https://doi.org/10.21833/ijaas.2024.01.003
Corresponding author's ORCID profile:
https://orcid.org/0000-0001-6495-301X
2313-626X/© 2023 The Authors. Published by IASE.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

World Health Organization (WHO), Malaysia is home to all four serotypes of the dengue virus, which induce distinct immunological responses in the body. Vaughn (2000) and Fried et al. (2010) found that the prevailing serotype can be used to forecast the severity of a disease epidemic. They demonstrated that individuals with initial DENV-1 or DENV-3 infections experienced severe sickness but that subsequent infection with DENV-2 exacerbated the condition, resulting in DHF.

The earliest documented dengue epidemics occurred between 1780 and 1940 when the environmental instability caused by World War II increased the disease's spread across Southeast Asia and the Pacific (Gubler, 2006). According to the World Health Organization, the dengue virus threatens nearly 2.5 billion people, or one-fifth of the global population. With the alarming surge in dengue infections worldwide, the World Health Organization (WHO) has also announced that the dengue virus has infected many more rural and urban countries than it did prior to the 1970s, with 70 percent happening in Asia, including Malaysia (Johari et al., 2019). Messina et al. (2014) noted that transmission of the dengue virus has spread to new regions, with the significant geographical expansion of several subtypes during the past two decades, especially in Asia and Latin America. Due to climate change, dengue fever has also become more prevalent in Central and South America, where it was formerly absent, and the failure of the dengue control campaign in the stipulated regions in the early 1970s paved the way for reinvasion and hyperendemicity as new virus serotypes moved more freely in these countries. Earlier estimates predicted that dengue fever would infect 50-100 million people worldwide, leading to 500,000 instances of DHF and up to 20,000 deaths (Whitehorn and Farrar, 2011). A study by Bhatt et al. (2013) found worrisome estimates of the dengue burden, estimating 390 million cases yearly, which is twice as much as prior projections. If climate change continues at its current rate, 50-60% of the world's population will be exposed to these vectors in 100 years, compared to 35% today (Butterworth et al., 2017).

Currently, no vaccines or specific treatments are available to treat or prevent the rising incidence of dengue worldwide. Many strategies have been proposed to tackle dengue surveillance and control activities effectively. The strategy generally comprises putting the vector population under surveillance and strategizing the control activity. The vector surveillance methods include the processes of collecting information relating to the population of dengue vectors and environmental factors that influence the vector life expectancy stages and rate (e.g., temperature, rainfall, and wind speed), which generally be used to predict the risk of dengue outbreak (Patil and Pandya, 2021; Hamdan and Kilicman, 2021). Meanwhile, the vector control methods are designed based on the precarious nature of the situation, resulting from the outcome of the surveillance method and the occurrence of dengue cases. It includes the manual removal of potential vector breeding grounds, fogging action using adulticides and larvicides, introducing specific types of bacteria to kill the immature stages of mosquitoes, or limiting viral replication and transmission in adult mosquitoes (Nazni et al., 2019). Other than that, there is also an awareness campaign activity for society (Kamaruddin et al., 2021), and Malaysia has a Communication for Behavioral Impact (COMBI) program to reduce breeding sites (Azmawati et al., 2013). Hence, with only primary remedy treatment and without effective vaccines or specific antiviral treatments, an accurate prediction model is necessary to assist the primary anti-dengue control strategy, aiming to eliminate breeding sites to prevent dengue virus transmission by infected adult mosquitoes. Therefore, the primary objective of this study is to develop the most accurate model to forecast future dengue cases in Malaysia's environment. The prediction model will enable early prevention and preparedness to combat dengue outbreaks, including an early warning system that can detect the outbreak's onset point. It is anticipated that the forecasting model will aid the Malaysian government in formulating strategies for controlling future dengue outbreaks.

2. Literature review

Many studies have been conducted on dengue prediction, employing various methods, including time series, machine learning, and mathematical approaches. Among the recent studies are Aziz and Aziz (2021), Keshavamurthy et al. (2022), Roster et al. (2022), and Kaur and Sharma (2023), who applied machine learning, Juraphanthong and Kesorn (2021), Othman et al. (2022), and Waseem et al. (2023) considered time series approach, and Ewing et al. (2021), Vásquez et al. (2022), and Wang et al. (2022) who deployed mathematical modeling.

Diverse comparisons of time series techniques have been used to examine the pattern of dengue occurrence, and it has been determined that the disease exhibits a seasonal pattern. For instance, Autoregressive Integrated Moving Average (ARIMA) models offer a different perspective on time series prediction. Mekparyup and Saithanu (2016) found that the seasonal ARIMA model best-predicted dengue haemorrhagic fever cases in Rayong, Thailand. Seasonal ARIMA was also the best model for predicting dengue cases in other countries such as Brazil, India, the Philippines, and Indonesia. On the other hand, Khashei and Bijari (2011) utilized a hybrid ARIMA-ANN model to improve forecasting accuracy. Chakraborty et al. (2019) compared models like ARIMA, ANN, NNAR, SVM, and LTSM and discovered that the hybrid ARIMA-NNAR model outperformed the others. A study by Nayak and Narayan (2019) in Kerala, India, predicted an increase in future dengue cases using SARIMA$(1,0,0)(0,1,1)_{12}$. Another study in Asahan District, Indonesia, showed that seasonal ARIMA

outperformed other models in predicting dengue haemorrhagic fever outbreaks. Several other studies have also compared the ARIMA approach with the Holt-Winters method. Abas et al. (2018) compared the Double Exponential Smoothing and Holt-Winters Method to forecast the number of dengue cases in Malaysia from 2010 to 2015. Even though both models demonstrated good performance, the Holt-Winters technique exhibited the lowest error measure, making it the most appropriate choice. In another study by Anggraeni et al. (2017) that aimed to forecast dengue fever patients in different groups in Malang, Indonesia, the Multiplicative Holt-Winters model outperformed other models in the low-Malang and medium-Malang groups. Elhassan et al. (2020) employed spatial methodologies to investigate mosquito reproduction patterns within University Municipality. Their objective was to assist Jeddah Municipality decision-makers in understanding the geographic areas where mosquito concentrations are prevalent. The ARIMA model has also been applied in other research areas. The model has been utilized to model the spread of COVID-19 cases (Jabeen et al., 2022) and mortality forecasting (Hamid et al., 2019). Moffat and Akpan (2017) identified and characterized deterministic and stochastic trends in their study in the context of revenue data. They analyzed revenue data to understand the underlying patterns and trends.

Harmonic Regression is the popular machine learning used to identify and examine recurring patterns or seasonal effects in time series data modeling. This modeling technique is beneficial when dealing with data that exhibit cyclical patterns with known or estimated periods. For instance, studies by Naumova et al. (2007), Wenger and Naumova (2010), and Ramanathan et al. (2020) have utilized harmonic regression to capture and model the seasonal variations in infectious diseases. By decomposing the time series into harmonic components, these studies were able to identify the underlying seasonal patterns and estimate their frequencies and amplitudes.

3. Methodology

The process of selecting the best forecasting model for predicting dengue fever incidence in Malaysia comprises four general processes. The first process is collecting the data, followed by examining the presence of seasonality component in the number of dengue cases using three different time series techniques, namely Seasonal Autoregressive Integrated Moving Average (SARIMA), Dynamic Harmonic Regression (DHR), and Neural Network Autoregressive Model (NNAR) in the second process. Then, in the third process, the outcomes from the three techniques are compared to select the best model, and finally, the forecasting analysis is performed using the best model. Detailed explanations of the processes are presented in the following section.

3.1. Data collection

This study utilizes a dataset comprising 306 weekly dengue cases reported in Malaysia from January 2017 to October 2022. These data points provide a historical record of dengue cases reported weekly over these six years. This data was sourced from the Ministry of Health Malaysia's (MOH) official website (idengue.mysa.gov.my). Initially, the data was split into estimation (in-sample) and evaluation (out-of-sample) parts as it is a critical step in time series analysis in ensuring that models are developed, tested, and evaluated in a rigorous and unbiased manner, with a focus on accurate forecasting and generalization to new data. This practice will help to build reliable models to make meaningful predictions for future time series observations. Generally, the data splitting ratio can range from 0.60-0.9:0.1-0.4. In this study, the estimation is 207 weeks, and the evaluation is 99 weeks to consider the existence of seasonality effect the dengue cases.

This study captures the presence of seasonality component in the number of dengue cases using three different time series techniques, namely seasonal Autoregressive Integrated Moving Average, Dynamic Harmonic Regression, and Neural Network models, which will be pointed as SARIMA, DHR, and NNAR, respectively, throughout this article. These three-time series techniques are employed to model the in-sample part of the data, and then the best model for each technique will be selected based on a few statistical criteria. Next, the three models are then compared using an out-of-sample dataset. The best-performing model of three different time series techniques will be utilized to forecast future dengue cases in Malaysia.

3.2. Seasonal autoregressive integrated moving average (SARIMA) model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model was first introduced by Box et al. (1994) and is also known as the Box-Jenkins method (Lazim, 2011). It serves the purpose of capturing and describing the presence of autocorrelation in time series data. The SARIMA model incorporates additional components or extensions to address specific features or challenges in time series data, such as seasonal patterns. The time plot of the dengue series will indicate whether there is a need to address the presence of autocorrelation in the data. A unit root test was performed to examine the stationarity of the data. If the data is found to be non-stationary, differencing is applied until the data becomes stationary. This study deployed several alternative models to identify the best model to represent the dynamic behavior of historical dengue cases in Malaysia. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots were used to aid model identification. The seasonal ARIMA (*p, d,*

q) (P, D, Q)s model, based on the Box-Jenkins approach, can be expressed as follows:

$$y_t = \mu + \frac{\theta_q(B)\theta_Q(B^s)}{\phi_p(B)\phi_P(B^s)(1-B)^d(1-B^s)^D} \quad (1)$$

where, y_t represents the weekly dengue cases at time t=1, 2,..., N, which spans 207 weeks. μ is the constant term. The model consists of three parameters divided into two parts: Non-seasonal (p, d, q) and seasonal (P, D, Q). The ARIMA order (p, d, q) denotes the autoregressive parameter(AR), p, the order of non-seasonal differencing, d, moving average (MA), q. The seasonal autoregressive (SAR), P, the order of seasonal differencing, D, and the seasonal moving average (SMA), Q. To eliminate the presence of seasonal component, seasonal differencing, $(1-B^s)^D$ is performed where s corresponds to 52 weeks, the length of the seasonality. Non-seasonal differencing is carried out to remove the presence of trends, expressed as $(1-B)^d$. The polynomials of the backward shift operator, $\phi_p(B)$, $\theta_q(B)$, $\phi_P(B^s)$, and $\theta_Q(B^s)$ denotes for AR(p), MA(q), SAR(P), and SMA(Q), respectively. The stages in the SARIMA model development can be summarized in Fig. 1.

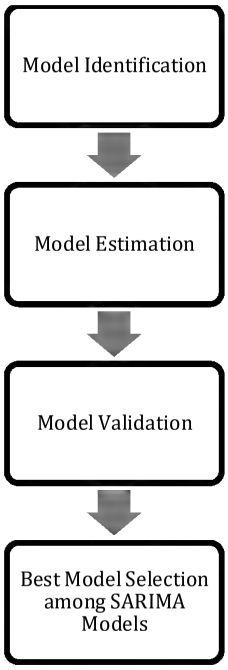


Fig. 1: Stages in SARIMA modeling

The SARIMA modeling approach begins with model identification. This step determines the most suitable model class for the time series data. Various statistical analyses are conducted during this process, including examining the autocorrelation function (ACF) and partial autocorrelation function (PACF). These analyses provide crucial insights into the data's characteristics. An appropriate model class can be identified based on the information gathered from ACF and PACF plots. Several statistical tests are employed in SARIMA model analysis, such as unit root tests, to ascertain whether the time series data exhibits a unit root, a sign of non-stationarity. A differencing approach can be performed if non-stationarity is detected, and the process continues until stationarity is achieved. In this study, three tests were employed, namely, Augmented Dickey-Fuller (ADF), Phillips-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests. The key differences among these tests lie in their null and alternative hypotheses and their primary purposes. The ADF and PP tests are used to test for unit roots and distinguish between stationary and non-stationary time series, while the KPSS test is used to test for the presence of a stationary trend within a time series. A combination of these tests is used to thoroughly analyze the stationarity properties of time series data.

Once the model is identified, the model fitting process is performed. The model fitting process assesses the estimated model's validity through various diagnostic and statistical tests, including the Portmanteau test called Box-Pierce, Akaike Information Criteria (AIC), and log-likelihood (LL) criteria. In this study, the best SARIMA model has been selected based on the criteria of the lowest AIC value, the highest LL, and the presence of white noise or the absence of correlation among the model's residuals, as determined by the Box-Pierce test. According to Box and Pierce (1970), the statistic value can be expressed as follows:

$$Q_{BP} = T\sum_{m=1}^{h}\hat{\rho}_m^2 \quad (2)$$

where, T is the number of sample size, $\hat{\rho}_k^2$ represents the squared of the estimated autocorrelation coefficient for each lag m, m=1, 2, ..., h. If the calculated statistic value exceeds the critical value of a $\chi_{\alpha,h}^2$ distribution, it leads to rejecting the null hypothesis that the residuals are independent.

Next, the AIC is also employed to validate the SARIMA models. It is a standard measure of the fitness of a SARIMA model. The equation can be expressed as follows:

$$AIC = e^{\frac{2k}{T}}\left(\frac{\sum_{t=1}^{T}e_t^2}{T}\right) \quad (3)$$

where, T represents the sample size, k is the sum of the number of parameters estimated in the model (p+P+q+Q), and e is the residual of the model. This equation aims to determine the values of p, P, q, and Q that minimize the value of the AIC. The SARIMA model with the lowest AIC value is chosen as the best-fitted model.

LL is also considered part of the model validation process among SARIMA models. In general, this equation can be expressed as follows:

$$\log(L) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}e_t^2 \quad (4)$$

where, L is the likelihood function, T is the number sample size, σ^2 is the error variance, and e is the residual of the model. The model with the highest LL is the best-fitting model.

3.3. Dynamic harmonic regression (DHR)

This study employed a DHR approach with Fourier terms to model the seasonal effect of dengue cases. Among the early studies that utilized this technique was Young et al. (1999). This technique is considered superior in handling more extensive seasonal lengths than SARIMA models. Different numbers of Fourier terms, K=1, 2, ..., 6, were allowed to capture the smoothness of the seasonal effect. The appropriate terms were determined based on the lowest AIC. This model has short-term dynamics that can be addressed by incorporating ARIMA errors into the DHR model. Following the approach of Young et al. (1999) and Ramanathan et al. (2020), the basic Harmonic regression model with ARIMA errors can be expressed as follows:

$$y_t = \mu + \beta_S \sin(2\pi\omega t) + \beta_C \cos(2\pi\omega t) + \eta_t \quad (5)$$

The ARIMA errors can be expressed using backward shift operator, B as:

$$(1 - B)\eta_t = (1 - \theta B)\varepsilon_t \quad (6)$$

The periodic component has a frequency of ω, which is equal to $1/52$ weeks. The model parameters β_S and β_C are defined as $\beta_S = -\gamma sin\varphi$ and $\beta_C = \gamma cos\varphi$, respectively. γ represents the amplitude that controls the fluctuation between two extreme points (peak and the lowest point), and φ is the phase angle that determines the location of the two extreme points. The error terms, ε_t are assumed to be independently and identically distributed with $E[\varepsilon_t] = 0$ and $Var[\varepsilon_t] = \sigma^2$.

3.4. Neural network autoregressive (NNAR) model

The Neural Network Autoregressive (NNAR) model is a nonlinear autoregressive model that is believed to provide more accurate forecasts, especially when dealing with complex and highly nonlinear data patterns. It can capture intricate dependencies and better adapt to changing conditions in the data. Hyndman and Athanasopoulos (2021) discussed the NNAR model in their book "Forecasting: Principles and Practice" and developed the *nnetar* function in the *forecast* package of the R programming language. In the case of seasonal data, an NNAR model can be theoretically expressed as NNAR (p, P, k)$_s$ model, which is equivalent to an ARIMA $(p, 0, 0)$ $(P, 0, 0)_s$ model but with nonlinear functions. The model can be expressed as:

$$y_t = f\left(y_{t-1}, y_{t-2}, \cdots, y_{t-p}, y_{t-s}, y_{t-2s}, \cdots, y_{t-Ps}\right) + \varepsilon_t \quad (7)$$

where, k represents the number of neurons in the hidden layer and the inputs are represented by $f()$. The order of p and P represents the number of non-seasonal lags and seasonal lags, respectively. All equations involved in this study assume that the error terms, ε_t are independently and identically distributed with $E[\varepsilon_t] = 0$ and $Var[\varepsilon_t] = \sigma^2$. The best NNAR model is selected using the Box-Pierce test and two error measures: The root mean square error (RMSE) and Mean Absolute Percentage Error (MAPE). Further discussions on RMSE and MAPE are elaborated in the next section.

3.5. Model selection among SARIMA, DHR, and NNAR techniques

In the estimation part, the best model was selected for each type of time series model used based on the AIC and LL values. The lowest AIC values and LL (in absolute value) indicate the best estimation model for each. A Portmanteau test was conducted to ensure diagnostic checking on the residuals of the best model.

Subsequently, the best model identified in the estimation part for each time series approach is fitted on the evaluation part (out-of-sample dataset) to evaluate the forecasting models' performance. This study utilized error measures such as Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) to obtain quantitative assessments of the accuracy and quality of the model's predictions. MSE, a widely used error measure, calculates the average of the squared differences between the predicted values and the actual values. It assigns greater weight to larger errors due to the squaring operation. MSE is valuable for assessing the overall performance of a model, with lower values indicating greater accuracy. On the other hand, MAPE is a relative error measure that calculates the average percentage difference between the predicted and actual values. It expresses the prediction error as a percentage of the actual value, making it easier to interpret.

The RMSE and MAPE can be expressed as follows, respectively:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T}} \quad (8)$$

$$MAPE = \sum_{t=1}^{T} \frac{\left|\left(\frac{e_t}{y_t}\right) * 100\right|}{T} \quad (9)$$

The best model among the three is chosen based on the lowest values of MSE and MAPE. This best model is then utilized to forecast future values of dengue cases from December 2022 to December 2023.

4. Results and discussion

This section presents the analysis of determining the presence of time series patterns such as trend, irregular, and seasonal patterns in weekly dengue cases in Malaysia. Understanding the time series behavior prior to analysis is essential for accurate modeling and improving forecasting accuracy. It can assist in making informed decisions and draw meaningful insights from time series data. In this

section, the findings are presented and explained comprehensively according to the study objective.

Fig. 2 shows the weekly time series plot of the dengue cases from 2017 to 2020. From 2017 to 2020, the Vector Control Unit of the Ministry of Public Security and Justice registered 383,409 dengue fever cases. The number of dengue cases in Malaysia climbed year after year during the study period, reaching an all-time high in 2019, with 130,101 dengue cases reported to the authorities. Over the study period, the observed incidence of dengue increased in both quantity and intensity, rising from 83,848 dengue cases in 2017 to 88,845 cases in 2020 alone. The decompose function in *R* software was used to disintegrate the weekly dengue cases into their components: Trend, seasonality, and irregular (remainder). The decomposition helps understand and analyze the underlying patterns and fluctuations in the time series data. The individual component is also presented. Fig. 2 also illustrates a seasonal effect, particularly during 2018-2021, captured and portrayed in the 'seasonal' component. The 'trend' component indicates the presence of a nonlinear trend in dengue cases, while the 'remainder' component represents the remaining variation in the data after removing both the seasonal and trend components from the dengue cases series. A seasonal pattern is evident in the wave-like pattern observed in the remainder component, providing further support for seasonal variation in the data. Consequently, the weekly dengue cases series is considered non-stationary due to the non-constant mean and variance.

4.1. SARIMA modelling

SARIMA modeling was employed to capture seasonal patterns in the weekly dengue cases. Unit root tests, including the Augmented Dickey-Fuller (ADF), Phillip-Perron (PP), and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests, were conducted to confirm the attainment of stationarity of the data after performing the differencing approach.

The results of the unit root tests are presented in Table 1. The results indicated that seasonal differencing alone cannot achieve stationarity in the data. The ADF and PP tests shared the null hypothesis that the series is non-stationary, while the KPSS test assumes the null hypothesis of the series being level or trend stationary. The *p*-values obtained from the ADF and PP tests exceed the significance levels of α=5%, 10%, and 1%, indicating a failure to reject the null hypothesis of non-stationarity. In addition, the KPSS test yielded a *p*-value below α, suggesting evidence against the null hypothesis of stationarity. All tests showed that the first-order seasonal differencing is insufficient to make the data stationary. Hence, a non-seasonal differencing of order one was then performed. The *p*-values from the ADF and PP tests for non-seasonal differencing are lower than α, providing evidence to reject the null hypothesis of non-stationarity. Moreover, the KPSS test yields a *p*-value=0.10, exceeding α, supporting the hypothesis of stationarity. Thus, we conclude that after applying both seasonal and non-seasonal differencing, the results demonstrate the achievement of stationarity in the data.

In addition to the unit root tests, the behavior of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots is also examined to assess the differenced series' characteristics further. Fig. 3 presents the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots for the differenced series of the weekly dengue data. The ACF plot of the seasonal differenced series exhibits an oscillation pattern, suggesting the presence of residual seasonality. To avoid over-differencing, we limit the order to one for both seasonal and non-seasonal differencing. After applying both types of differencing, the data successfully achieves stationarity. Analyzing the ACF plot (top-right), significant autocorrelation at lags 2, 51, and 52 are identified, indicating an order of $q=2$ and a seasonal order of $Q=1$. Similarly, the PACF plot (bottom-right) reveals significant correlations at lags 1, 21, 51, and 52, suggesting an order of $p=3$ and a seasonal order of $P=1$. However, due to a non-convergence issue with estimating SARIMA (3, 1, 2) $(1, 1, 1)_{52}$, we reduce the order to SARIMA (2, 1, 1)$(1, 1, 1)_{52}$. Any spikes barely touching the bands were ignored to avoid overfitting.

To capture the seasonal effect without sacrificing parsimony, we utilized the auto.arima function from the *forecast* package. By setting the maximum parameters to five, the SARIMA (0, 1, 2) $(1, 1, 0)_{52}$ model is obtained. Considering the ACF and PACF plots of the stationary series, we explored alternative SARIMA orders, including SARIMA (0, 1, 1) $(1, 1, 1)_{52}$, SARIMA (1, 1, 1) $(1, 1, 1)_{52}$ and SARIMA (0, 1, 2) $(1, 1, 1)_{52}$, and SARIMA (0, 1, 0) $(1, 1, 1)_{52}$. These SARIMA models were constructed and fitted to the data, with the estimation models summarized in Table 2. Based on the evaluation criteria, which include the AIC value, LL, and the Box-Pierce test, the SARIMA (0, 1, 2) $(1, 1, 1)_{52}$ model emerges as the best choice with the lowest AIC value of 2149.13 and the highest LL, which is -1069.56. In addition, the p-value of the Box-Pierce test is 0.9709, higher than all three significance levels, indicating that the model's residuals are white noise.

4.2. DHR model

The dynamic harmonic regression approach with Fourier terms was employed to model the data. Table 3 presents the output, indicating that the optimal number of Fourier terms (K) is 2, which yields the lowest AIC value of -2775.057. Consequently, the model was estimated using K=2, and the results are displayed in Table 4.

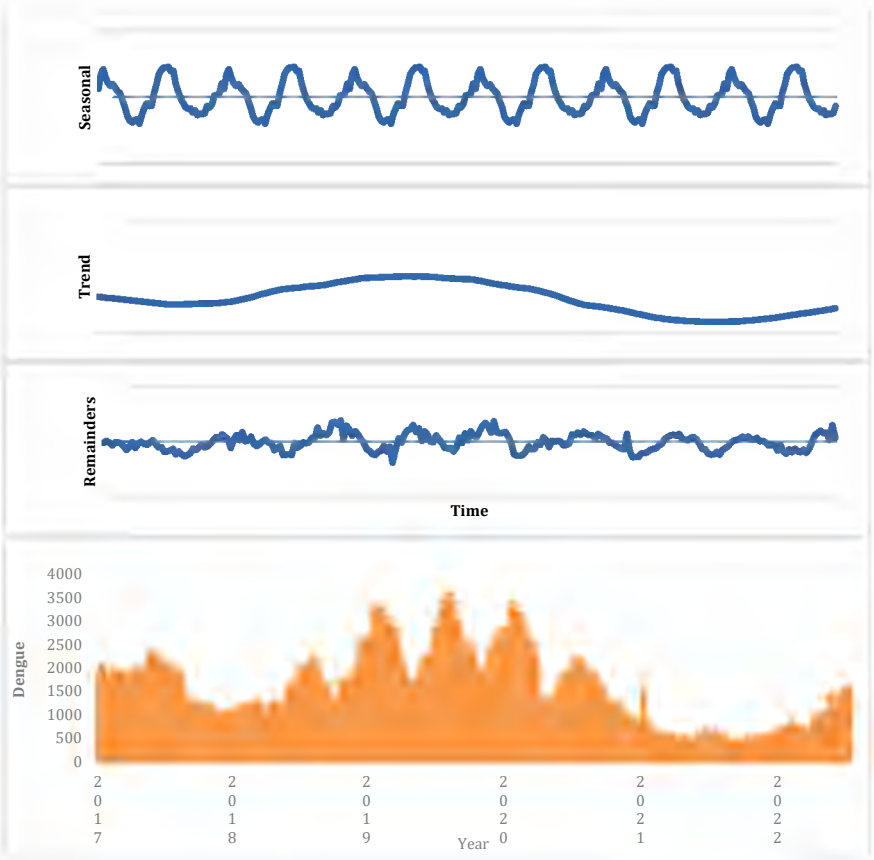


Fig. 2: Time series plot of weekly dengue cases in Malaysia

Table 1: Unit root tests for seasonal and non-seasonal differenced series of order one

Unit root tests	*p*-value for seasonal differenced series	*p*-value for non-seasonal differenced series
ADF	0.9039	0.01
PP	0.8031	0.01
KPSS	0.0100	0.10

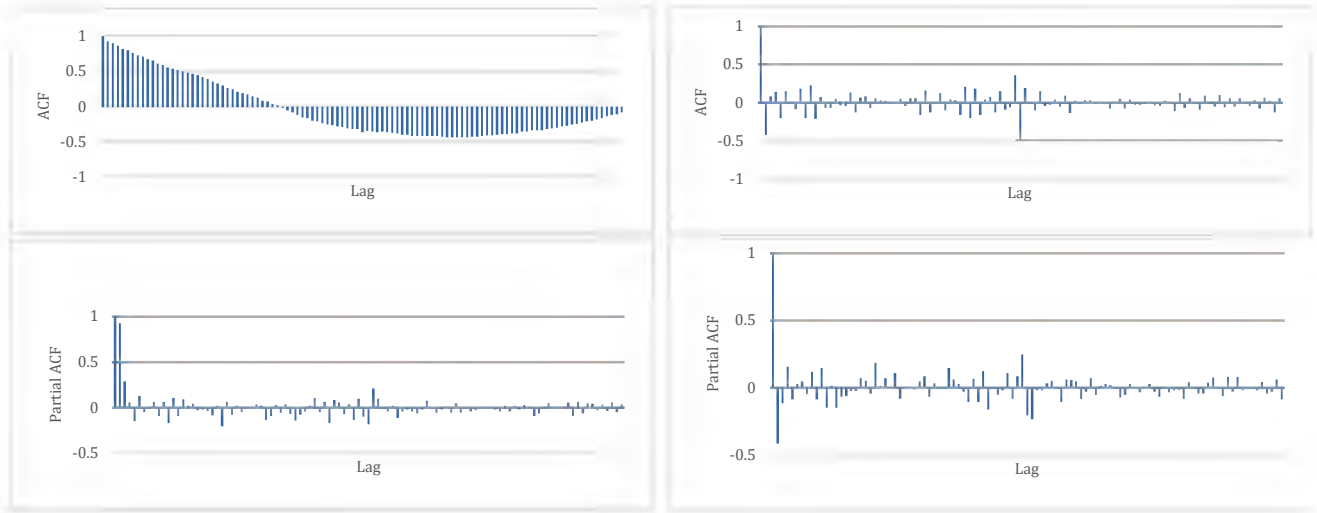


Fig. 3: ACF and PACF plots of seasonal (top-left and bottom-left) and non-seasonal differenced series of order 1 (top-right and bottom-right)

4.3. NNAR model

The NNAR model was implemented using the *nnetar* function, automatically selecting the NNAR (2, 1, 2)$_{52}$ model. However, in order to explore alternative options, we attempted to manually set the seasonal lags, *P*, ranging from 1 to 11. The error measure criteria were used to obtain the best-estimated model using NNAR. To further evaluate the model selection, we compared the correlogram plots of the residual series for all models. Among these plots, NNAR (9, 1, 6)$_{52}$ exhibited correlation coefficients that all fell within the confidence bands, indicating a white noise process, as displayed in Table 5. Based on this analysis, we concluded that the best NNAR model for our study is NNAR (9, 1, 6)$_{52}$ due to the presence of white noise residuals, as displayed in Fig. 4. Additionally, the Box-Pierce test (with a p-value greater than any level of significance) also confirms that the residuals of this model are

independent and white noise, supporting the correlogram plot's findings. Despite slightly higher RMSE and MAPE values compared to NNAR (10, 1, $6)_{52}$ and NNAR (11, 1, $6)_{52}$, we selected NNAR(9, 1, $6)_{52}$ for its parsimonious criteria among these three models.

Table 2: Empirical results of SARIMA models

SARIMA(p,d,q)(P,D,Q)	Coefficient		Value	Box-Pierce test (p-value)	AIC	LL
$SARIMA(0,1,2)(1,1,0)_{52}$	θ_1	MA(1)	-0.266 (0.084)	0.9919	2157.36	-1074.68
	θ_2	MA(2)	0.1541 (0.0796)			
	ϕ_1	SAR(1)	-0.5714 (0.0705)			
$SARIMA(1,1,1)(1,1,1)_{52}$	ϕ_1	AR(1)	-0.3502 (0.2140)	0.9348	2150.98	-1070.49
	θ_1	MA(1)	0.0922 (0.2239)			
	ϕ_1	SAR(1)	-0.1265 (0.1333)			
	θ_1	SMA(1)	-0.997 (0.4910)			
$SARIMA(2,1,1)(1,1,1)_{52}$	ϕ_1	AR(1)	-0.3502 (0.2140)	0.8976	2151.70	-1069.85
	ϕ_2	AR(2)	0.2367 (0.0988)			
	θ_1	MA(1)	-0.6171 (0.3282)			
	ϕ_1	SAR(1)	-0.1388 (0.1076)			
	θ_1	SMA(1)	-0.9967 (0.4328)			
$SARIMA(0,1,1)(1,1,1)_{52}$	θ_1	MA(1)	-0.2164 (0.0736)	0.6483	2151.10	-1071.55
	ϕ_1	SAR(1)	-0.1277 (0.1090)			
	θ_1	SMA(1)	-0.9991 (0.4348)			
$SARIMA(0,1,2)(1,1,1)_{52}$	θ_1	MA(1)	-0.2552 (0.0849)	**0.9709**	**2149.13**	**-1069.56**
	θ_2	MA(2)	0.1687 (0.0821)			
	ϕ_1	SAR(1)	-0.1322 (0.1102)			
	θ_1	SMA(1)	-0.9994 (0.4232)			
$SARIMA(0,1,0)(1,1,1)_{52}$	ϕ_1	SAR(1)	-0.2238 (0.0934)	0.0007	2157.20	-1075.60
	θ_1	SMA(1)	-0.9990 (0.4219)			

Table 3: AIC values for each Fourier term

Fourier term, K	AIC
1	2801.243
2	2775.057
3	2776.305
4	2779.074
5	2780.384

Table 4: Empirical results of the DHR model

Coefficient		Value	Box-Pierce test (p-value)	AIC	LL
θ_1	MA(1)	-0.2163 (0.0667)	0.9285	2775.057	-1381.53
β_{S1}	S(1)	23.5765 (127.8266)			
β_{C1}	C(1)	-63.5601 (126.2331)			
β_{S2}	S(2)	302.4950 (64.4958)			
β_{C2}	C(2)	265.3840 (63.7252)			

Table 5: Comparison of NNAR models

NNAR(p,P,k)s	RMSE	MAPE	Box-Pierce test (p-value)
NNAR(2,1,2)52	216.426	8.635	0.7746
NNAR(1,1,2)52	221.732	8.752	0.5302
$NNAR(1,3,2)_{52}$	139.305	5.776	0.3464
$NNAR(3,1,2)_{52}$	206.248	8.220	0.5779
$NNAR(1,2,2)_{52}$	228.458	8.073	0.5543
$NNAR(4,1,3)_{52}$	174.156	7.223	0.7757
$NNAR(9,1,4)_{52}$	109.3241	4.701	0.7612
$NNAR(10,1,4)_{52}$	106.138	4.548	0.6773
$NNAR(9,1,6)_{52}$	**75.674**	**3.458**	0.5324
$NNAR(10,1,6)_{52}$	**62.686**	**2.949**	0.7801
$NNAR(11,1,6)_{52}$	**58.645**	**2.825**	0.8485

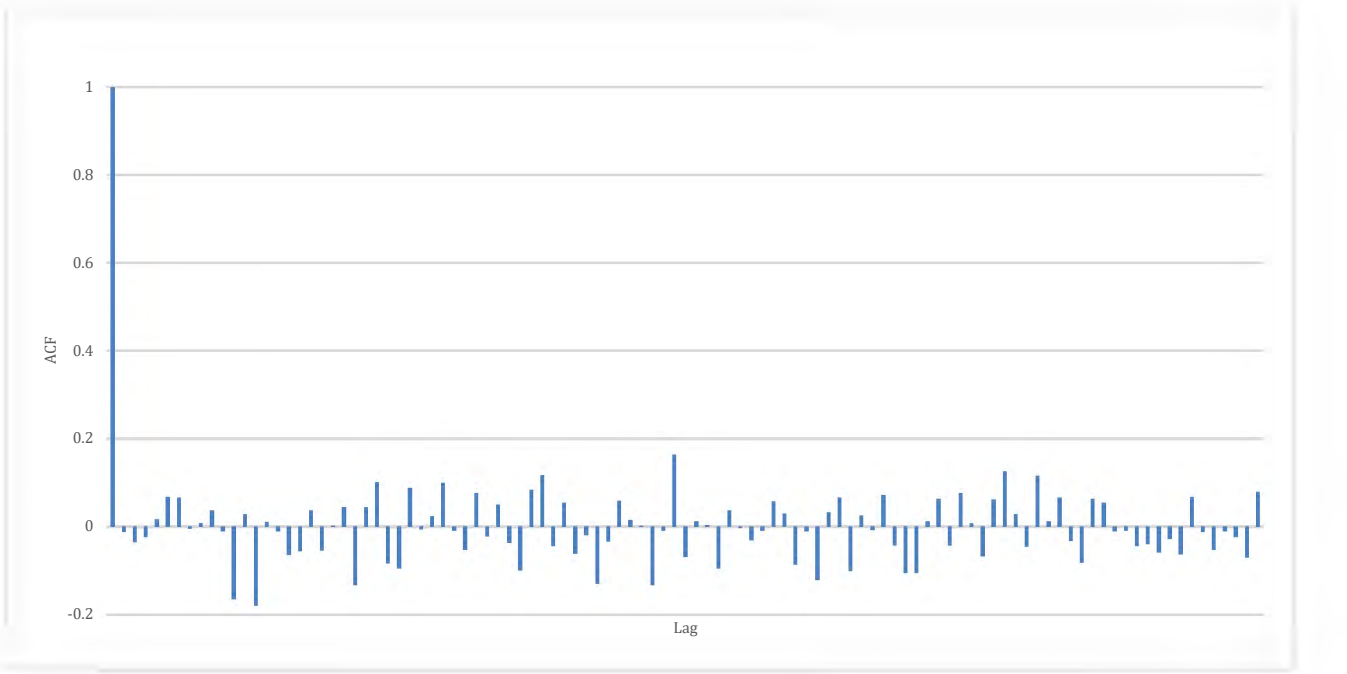


Fig. 4: ACF plot of the residuals series of NNAR(9,1,6)$_{52}$

4.4. Comparing models among the three techniques

A comparison among the best models for each modeling technique using an out-of-sample dataset was performed, and the results are presented in Table 6. The results indicate that the Dynamic Harmonic Regression with Fourier term (K=2) outperforms both the SARIMA model and the NNAR (9, 1, 6)$_{52}$ models, as evidenced by the lowest RMSE and MAPE values of 537.68 and 49.31, respectively.

Table 6: The out-of-sample model evaluation

Error measures	Models		
	SARIMA(0,1,2)(1,1,1)$_{52}$	DHR with k=2	NNAR(9,1,6)$_{52}$
RMSE	546.99	**537.68**	597.74
MAPE	58.42	**49.61**	94.84

Moving forward, the DHR model was utilized to forecast the weekly dengue cases for 2023. Table 7 presents the forecasted values for 2023, and Fig. 5 displays the plots of these forecasts. The plot of future forecasts from the DHR model reveals a slightly increasing trend and seasonal pattern in the forecasted values of 2023. The number of forecasted cases surpasses a thousand during Week 2 to Week 5, followed by Week 26 to Week 32. The forecasted values begin at approximately 956 cases in Week 1, reach their highest peak in Week 29, and then decline to below 300 cases in Weeks 41 to 45, with a minimum forecasted value of 236 cases in Week 43. There are approximately 35,664 cases of dengue occurrences forecasted using the DHR model in Malaysia, with an average of 686 cases per week in 2023.

Table 7: Forecast values for the year 2023

Week	Dengue cases	Week	Dengue cases	Week	Dengue cases
1	956	19	413	37	584
2	1005	20	481	38	485
3	1034	21	563	39	397
4	1040	22	655	40	325
5	1024	23	752	41	272
6	986	24	848	42	242
7	930	25	937	43	236
8	858	26	1015	44	253
9	775	27	1076	45	294
10	687	28	1118	46	354
11	599	29	1136	47	431
12	516	30	1131	48	520
13	444	31	1102	49	616
14	387	32	1050	50	713
15	348	33	979	51	805
16	331	34	891	52	888
17	337	35	793		
18	364	36	688		

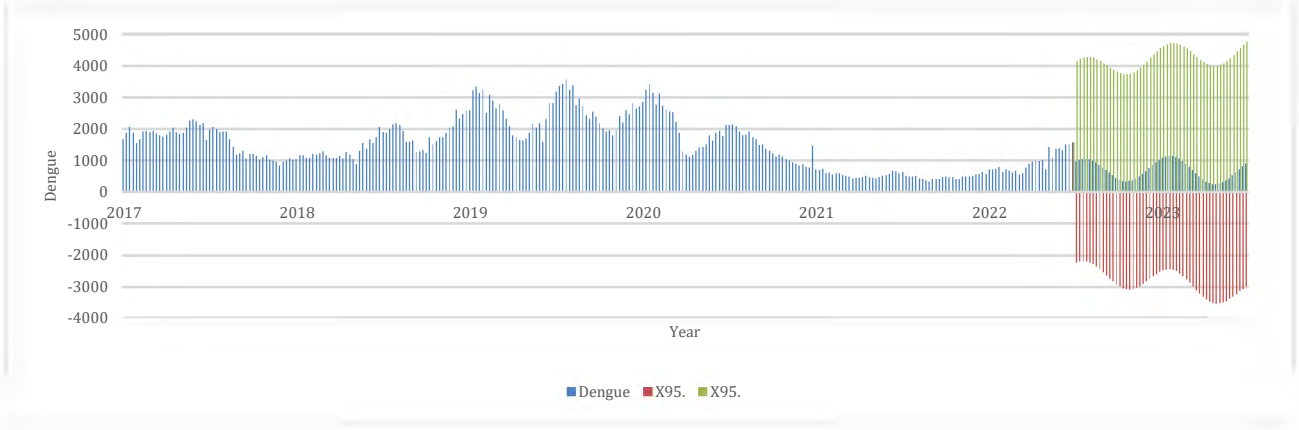


Fig. 5: Plot of forecast values up to the year 2023

5. Conclusion

This study employed three-time series modeling techniques, namely Seasonal Autoregressive Integrated Moving Average (SARIMA), Dynamic Harmonic Regression (DHR), and Neural Network Autoregressive (NNAR) models, to model and, hence, forecast weekly dengue cases in Malaysia. It was found that the DHR model outperformed the other two techniques on the evaluation dataset based on the lowest error measures, MSE and MAPE. Subsequently, this best-performing model generated future forecast values for weekly dengue cases in 2023. The analysis revealed that more than 1000 cases occurred between Week 2 and Week 5 and between Week 26 and Week 32, with the highest number of cases (i.e., 1136) recorded in Week 29. These findings suggest an increase in dengue cases lasting approximately five months following the end of the monsoon season. Applying this technique can prove valuable for healthcare administrators in enhancing preparedness. Further research could focus on improving the prediction performance of the proposed model. Additionally, more advanced techniques, such as hybrid time series and machine learning approaches, could be explored for predicting dengue cases in Malaysia.

Acknowledgment

This research was supported by the Ministry of Higher Education (MOHE) of Malaysia through the Fundamental Research Grant Scheme (FRGS/1/2022/SS0/UiTM/02/20).

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Abas N, Shamsuddin RM, and Halim SA (2018). Modelling and prediction of dengue cases at two progressive regions in Malaysia. Journal of Fundamental and Applied Sciences, 10(3S): 319-324.

Aguiar M, Anam V, Blyuss KB, Estadilla CD, Guerrero BV, Knopoff D, Kooi BW, Srivastav AK, Steindorf V, and Stollenwerk N (2022). Mathematical models for dengue fever epidemiology: A 10-year systematic review. Physics of Life Reviews, 40: 65-92. https://doi.org/10.1016/j.plrev.2022.02.001 **PMid:35219611 PMCid:PMC8845267**

Anggraeni W, Nurmasari R, Riksakomara E, Samopa F, and Wibowo RP (2017). Modified regression approach for predicting number of dengue fever incidents in Malang Indonesia. Procedia Computer Science, 124: 142-150. https://doi.org/10.1016/j.procs.2017.12.140

Aziz A and Aziz A (2021). Dengue cases prediction using machine learning approach. IRASD Journal of Computer Science and Information Technology, 2(1): 13-25. https://doi.org/10.52131/jcsit.2021.0201.0007

Azmawati MN, Aniza I, and Munawar ALI (2013). Evaluation of communication for behavioral impact (COMBI) program in dengue prevention: A qualitative and quantitative study in Selangor, Malaysia. Iranian Journal of Public Health, 42(5): 538-539.

Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL, and Hay SI (2013). The global distribution and burden of dengue. Nature, 496(7446): 504-507. https://doi.org/10.1038/nature12060 **PMid:23563266 PMCid:PMC3651993**

Box GE and Pierce DA (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association, 65(332): 1509-1526. https://doi.org/10.1080/01621459.1970.10481180

Box GEP, Jenkins GM, and Reinsel GG (1994). Time series analysis: Forecasting and control. 3rd Edition, Prentice Hall, Englewood Cliff, USA.

Butterworth MK, Morin CW, and Comrie AC (2017). An analysis of the potential impact of climate change on dengue transmission in the southeastern United States. Environmental Health Perspectives, 125(4): 579-585. https://doi.org/10.1289/EHP218 **PMid:27713106 PMCid:PMC5381975**

Chakraborty T, Chattopadhyay S, and Ghosh I (2019). Forecasting dengue epidemics using a hybrid methodology. Physica A: Statistical Mechanics and its Applications, 527: 121266. https://doi.org/10.1016/j.physa.2019.121266

Elhassan FAM, Qurashi MAE, and Elhafian MH (2020). Spatial statistic of reproduction of dengue fever mosquitoes transmitter phenomenon in the university municipality Jeddah Province Saudi Arabia during 2018. International Journal of Advanced and Applied Sciences, 7(11): 58-66. https://doi.org/10.21833/ijaas.2020.11.006

Ewing DA, Purse BV, Cobbold CA, and White SM (2021). A novel approach for predicting risk of vector-borne disease establishment in marginal temperate environments under climate change: West Nile virus in the UK. Journal of the Royal Society Interface, 18(178): 20210049. https://doi.org/10.1098/rsif.2021.0049 **PMid:34034529 PMCid:PMC8150030**

Fried LP, Bentley ME, Buekens P, Burke DS, Frenk JJ, Klag MJ, and Spencer HC (2010). Global health is public health. The Lancet, 375(9714): 535-537. https://doi.org/10.1016/S0140-6736(10)60203-6 **PMid:20159277**

Gubler DJ (2006). Dengue/dengue haemorrhagic fever: History and current status. In: Bock G and Goode J (Eds.), New treatment strategies for dengue and other flaviviral diseases: Novartis foundation symposium: 3-22. Volume 277, John Wiley and Sons, Ltd., Chichester, UK. https://doi.org/10.1002/0470058005.ch2 **PMid:17319151**

Hamdan NI and Kilicman A (2021). The development of a deterministic dengue epidemic model with the influence of temperature: A case study in Malaysia. Applied Mathematical Modelling, 90: 547-567. https://doi.org/10.1016/j.apm.2020.08.069

Hamid SI, Ofosu-Hene ED, and Ponnusamy RR (2019). A new approach to forecast Malaysian mortality rates. International Journal of Advanced and Applied Sciences, 6(10): 53-61. https://doi.org/10.21833/ijaas.2019.10.010

Htun TP, Xiong Z, and Pang J (2021). Clinical signs and symptoms associated with WHO severe dengue classification: A systematic review and meta-analysis. Emerging Microbes and Infections, 10(1): 1116-1128. https://doi.org/10.1080/22221751.2021.1935327 **PMid:34036893 PMCid:PMC8205005**

Hyndman RJ and Athanasopoulos G (2021). Forecasting: Principles and practice. 3rd Edition, OTexts, Melbourne, Australia.

Jabeen D, Rahim I, Iftikhar R, Rafiullah M, and Ansari M (2022). Regression modeling and correlation analysis spread of COVID-19 data for Pakistan. International Journal of Advanced and Applied Sciences, 9(3): 71-81. https://doi.org/10.21833/ijaas.2022.03.009

Johari NA, Voon K, Toh SY, Sulaiman LH, Yap IKS, and Lim PKC (2019). Sylvatic dengue virus type 4 in *Aedes aegypti* and *Aedes albopictus* mosquitoes in an urban setting in Peninsular Malaysia. PLOS Neglected Tropical Diseases, 13(11): e0007889. https://doi.org/10.1371/journal.pntd.0007889 **PMid:31730672 PMCid:PMC6881067**

Juraphanthong W and Kesorn K (2021). Time series data enrichment using semantic information for dengue incidence forecasting. Science, Engineering and Health Studies, 15: 21050013.

Kamaruddin NIK, Said SM, Shahar HK, and Lim PY (2021). Socio-ecological determinants of dengue prevention practices: A cross-sectional study among wet market traders in a selected district in Perak, Malaysia. Asian Pacific Journal of Tropical Medicine, 14(12): 555-563. https://doi.org/10.4103/1995-7645.332810

Kaur S and Sharma S (2023). Comparative analysis of machine learning classifiers on forecasting dengue fever infection. In: Murthy KVSR, Kumar S, and Singh MK (Eds.), Recent developments in electronics and communication systems: 492–4979. IOS Press, Amsterdam, Netherlands. https://doi.org/10.3233/ATDE221302

Keshavamurthy R, Dixon S, Pazdernik KT, and Charles LE (2022). Predicting infectious disease for biopreparedness and response: A systematic review of machine learning and deep learning approaches. One Health, 15: 100439. https://doi.org/10.1016/j.onehlt.2022.100439 **PMid:36277100 PMCid:PMC9582566**

Khashei M and Bijari M (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Applied Soft Computing, 11(2): 2664-2675. https://doi.org/10.1016/j.asoc.2010.10.015

Lazim MA (2011). Introductory business forecasting: A practical approach. 3rd Edition, UiTM Press, Shah Alam, Malaysia.

Mekparyup J and Saithanu K (2016). A new approach to detect epidemic of DHF by combining ARIMA model and adjusted Tukey's control chart with interpretation rules. Interventional Medicine and Applied Science, 8(3): 118-120. https://doi.org/10.1556/1646.8.2016.3.6 **PMid:28203393 PMCid:PMC5283758**

Messina JP, Brady OJ, Scott TW, Zou C, Pigott DM, Duda KA, and Hay SI (2014). Global spread of dengue virus types: Mapping the 70 year history. Trends in Microbiology, 22(3): 138-146. https://doi.org/10.1016/j.tim.2013.12.011 **PMid:24468533 PMCid:PMC3946041**

Moffat IU and Akpan EA (2017). Time series modeling of the interaction between deterministic and stochastic trends. International Journal of Advanced and Applied Sciences, 4(9): 96-100. https://doi.org/10.21833/ijaas.2017.09.012

Naumova EN, Jagai JS, Matyas B, DeMaria A, MacNeill IB, and Griffiths JK (2007). Seasonality in six enterically transmitted diseases and ambient temperature. Epidemiology and Infection, 135(2): 281-292. https://doi.org/10.1017/S0950268806006698 **PMid:17291363 PMCid:PMC2870561**

Nayak MSDP and Narayan KA (2019). Forecasting dengue fever incidence using ARIMA analysis. International Journal of Collaborative Research on Internal Medicine and Public Health, 11(3): 924-932.

Nazni WA, Hoffmann AA, NoorAfizah A, Cheong YL, Mancini MV, Golding N, and Sinkins SP (2019). Establishment of Wolbachia strain wAlbB in Malaysian populations of *Aedes aegypti* for dengue control. Current Biology, 29(24): 4241-4248. https://doi.org/10.1016/j.cub.2019.11.007 **PMid:31761702 PMCid:PMC6926472**

Othman M, Indawati R, Suleiman AA, Qomaruddin MB, and Sokkalingam R (2022). Model forecasting development for dengue fever incidence in Surabaya City using time series analysis. Processes, 10(11): 2454. https://doi.org/10.3390/pr10112454

Patil S and Pandya S (2021). Forecasting dengue hotspots associated with variation in meteorological parameters using regression and time series models. Frontiers in Public Health, 9: 798034. https://doi.org/10.3389/fpubh.2021.798034 **PMid:34900929 PMCid:PMC8661059**

Ramanathan K, Thenmozhi M, George S, Anandan S, Veeraraghavan B, Naumova EN, and Jeyaseelan L (2020). Assessing seasonality variation with harmonic regression: Accommodations for sharp peaks. International Journal of Environmental Research and Public Health, 17(4): 1318. https://doi.org/10.3390/ijerph17041318 **PMid:32085630 PMCid:PMC7068504**

Roster K, Connaughton C, and Rodrigues FA (2022). Machine-learning–based forecasting of dengue fever in Brazilian cities using epidemiologic and meteorological variables. American Journal of Epidemiology, 191(10): 1803-1812. https://doi.org/10.1093/aje/kwac090 **PMid:35584963**

Tantawichien T (2012). Dengue fever and dengue haemorrhagic fever in adolescents and adults. Paediatrics and International Child Health, 32(Sup1): 22-27. https://doi.org/10.1179/2046904712Z.00000000049 **PMid:22668446 PMCid:PMC3381442**

Vásquez P, Sanchez F, Barboza LA, García YE, Calvo JG, Chou-Chen SW, and Mery G (2022). Mathematical and statistical models for the control of mosquito-borne diseases: The experience of Costa Rica. Revista Panamericana de Salud Pública, 46: e113.

https://doi.org/10.26633/RPSP.2022.113 **PMid:36060201 PMCid:PMC9426954**

Vaughn DW (2000). Invited commentary: Dengue lessons from Cuba. American Journal of Epidemiology, 152(9): 800-803. https://doi.org/10.1093/aje/152.9.800 **PMid:11085390**

Wang WH, Urbina AN, Chang MR, Assavalapsakul W, Lu PL, Chen YH, and Wang SF (2020). Dengue hemorrhagic fever–A systemic literature review of current perspectives on pathogenesis, prevention and control. Journal of Microbiology, Immunology and Infection, 53(6): 963-978. https://doi.org/10.1016/j.jmii.2020.03.007 **PMid:32265181**

Wang Y, Li Y, Ren X, and Liu X (2022). A periodic dengue model with diapause effect and control measures. Applied Mathematical Modelling, 108: 469-488. https://doi.org/10.1016/j.apm.2022.03.043

Waseem MD, Nawaz A, Rasheed U, Raza A, and Albarka MO (2023). A time series regression-based model for predicting the spread of dengue disease. In the International Conference on Robotics and Automation in Industry, IEEE, Peshawar, Pakistan: 1-6. https://doi.org/10.1109/ICRAI57502.2023.10089545 **PMCid:PMC9936943**

Wenger JB and Naumova EN (2010). Seasonal synchronization of influenza in the United States older adult population. PLOS ONE, 5(4): e10187. https://doi.org/10.1371/journal.pone.0010187 **PMid:20419169 PMCid:PMC2855366**

Whitehorn J and Farrar J (2011). Dengue. Clinical Medicine, 11(5): 483-487. https://doi.org/10.7861/clinmedicine.11-5-483 **PMid:22034713 PMCid:PMC4954247**

Young PC, Pedregal DJ, and Tych W (1999). Dynamic harmonic regression. Journal of Forecasting, 18(6): 369-394. https://doi.org/10.1002/(SICI)1099-131X(199911)18:6<369::AID-FOR748>3.0.CO;2-K