Developing a Trend Equation and Compararive Analysis of ARIMA and Random Forest Models for Forecasting Dengue Fever Incidence Using Time Series Data : A Case Study of Health District 12, Thailand

Wafeeqismail Noipom, Varees Adulyasas, Rattifa Kasa, Marusdee Yusoh
Islamic Science Demonstration School, Faculty of Islamic Science, Prince of Songkhla University Pattani Campus, Thailand.

ABSTRACT

Dengue fever outbreaks have been ongoing in Thailand in recent years. This study focuses on the number of dengue fever patients under investigating Health Region 12, which comprises the provinces of Songkhla, Satun, Trang, Phatthalung, Pattani, Yala, and Narathiwat provinces. The aim of this study is to formulate trend equations and forecastimg models for dengue fever patient data spanning from 2015 to 2023. Three methodologies were studied, including the Decomposition Method, Box-Jenkins Method, and Random Forest Method. The suitability of the forecasting models is evaluated based on three accuracy metrics: Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The study reveals that the Component Separation Method, coupled with the Seasonal Index, emerges as the optimal approach for trend identification, characterized by the formula $y_t = -0.2797t^2 + 2664.5478 + S_s$. Furthermore, the Box-Jenkins Method emerges as the most suitable technique for developing forecasting models, particularly with the $ARIMA(1,1,0)$ model proving to be the most accurate in predicting future data trends. These models demonstrate superior performance across all three accuracy metrics, showcasing MAPE at 61.84, MSE at 504,046.04, and RMSE at 709.96.

Keywords:

Dengue fever
Time Series
Trend
Box-Jenkins
Random Forest

Corresponding Author:

Marusdee Yusoh,
Islamic Science Demonstration School,
Faculty of Islamic Science,
Prince of Songkhla University Pattani Campus, Thailand.
Email: marusdee.y@psu.ac.th

1. INTRODUCTION

Dengue fever is currently the most significant public health problem in humid tropical countries. It typically occurs during the rainy season, and outbreaks have been documented in Thailand for more than 50 years [10]. The dengue virus, which is spread by mosquitoes, is what causes dengue fever. As per the reports issued by the Department of Disease Control, between January 1st, 2023 and December 13th, 2023, there were a total of 147,412 reported cases of dengue fever, with 174 fatalities across 57 provinces. Furthermore, it was discovered that there has been a steady rise in dengue fever cases in the Songkhla, Satun, Trang, Pattani, Yala, and Narathiwat areas that make up the 12th Public Health Region. In the 12th Public Health Region, the House Index (HI) was 24.11% and the Container Index (CI) was 9.91%, above the standard limits, according to a survey of the mosquito larval index in epidemic-prone locations. This suggests a region where the spread of

.

infectious diseases carried by mosquitoes is highly likely. The frequency of dengue fever cases has dramatically increased as a result of insufficient and inconsistent efforts to eliminate mosquito breeding sites and a lack of community participation in eliminating mosquito breeding sites within homes. In 2022, there were 8,479 cumulative cases of dengue fever, whereas in 2023, there were 10,630 cases, an increase of 9.27 times compared to the same period in the previous year.

The method forecasting of dengue fever outbreaks is a statistical approach that utilizes time series data to study and identify suitable models for explaining the patterns or variations in past data and then applies these models to forecast future data. In order to predict the incidence of different diseases in Thailand in 2015, a number of researchers have developed and compared mathematical and statistical models utilizing time series data in recent years. The Box-Jenkins method, Polynomial regression method, and Mixed model approach are the three statistical techniques used to anticipate the number of pneumonia patients in Thailand. These techniques are considered to be the most appropriate forecasting models. According to the research findings, this particular time series dataset was best suited for the Mixed model method [2]. Similarly, compared time series analysis methods to forecast Thailand's crude oil output levels. The Decomposition method, the Box-Jenkins method, and the Grey model forecasting method were the three approaches they looked at. The Box-Jenkins approach was shown to be the most suitable for forecasting [3]. Furthermore, data mining techniques are being used to create models to predict dengue fever outbreaks under the care of the second disease control office in Phitsanulok, Thailand. The Decision Tree, Bayesian, Neural Network, and Support Vector Machine methods were the four that were used [5]. The study found that the Decision Tree method was the most suitable technique for creating a model to forecast dengue fever outbreaks. Furthermore, investigated the modelling of dengue fever outbreak prediction using a technique suitable for the data format with the Grey theory. The study discovered that the GM(1,1) system's Grey theory method worked well for predicting the number of cases of dengue fever [4].

From the Previous sections indicate that this research is to explore techniques for predicting the number of dengue fever cases in Thailand's 12th Public Health Region, while also evaluating appropriate forecasting models for this time series dataset. Specifically, the research centres on comparing the efficacy of three forecasting methodologies: Box-Jenkins and Random methods.

2. RESEARCH METHOD

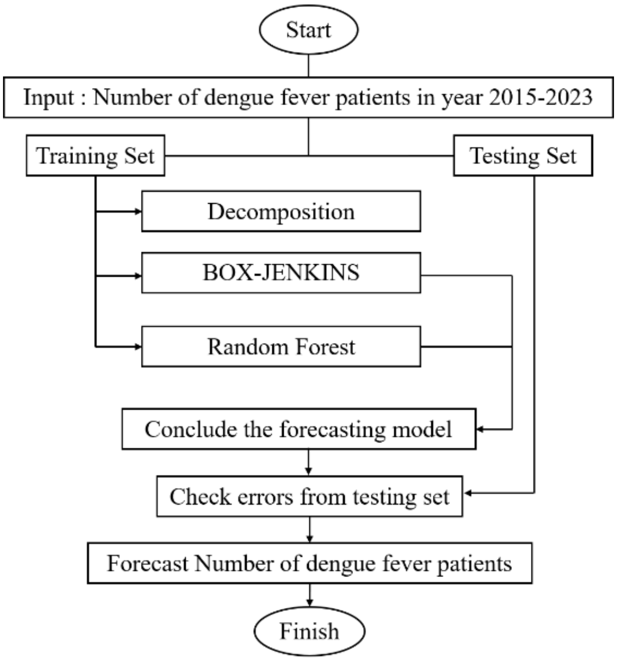
Start
Input : Number of dengue fever patients in year 2015-2023
Training Set
Testing Set
Decomposition
BOX-JENKINS
Random Forest
Conclude the forecasting model
Check errors from testing set
Forecast Number of dengue fever patients
Finish

Figure 1. An overview illustrating the process of forecasting model development.

2.1. Dataset

From 2015 to 2023, a monthly time series is gathered to create the database of dengue fever patients. The dataset was obtained from the https://pnb.hdc.moph.go.th/hdc/main/index.php. website of the Ministry of Public Health. There are one hundred entries in the dataset. The data used in this study is split in an 80:20 ratio between the Training Set and Testing Set. applying data analysis using the R programming language. In addition, as shown in Figure 1, three approaches are chosen for analysis: the Decomposition method, the Box-Jenkins method, and the Random Forest (RF) method. The following are the details of each method:

2.2. Decomposition

The decomposition method is a technique for separating the components of time series into various parts, including trend, seasonal effect, cyclical effect, and irregular events, as shown in Figure 2. Each separated component will reveal the movement characteristics of each part of the time series and can be used to create forecasting equations for future use [11].

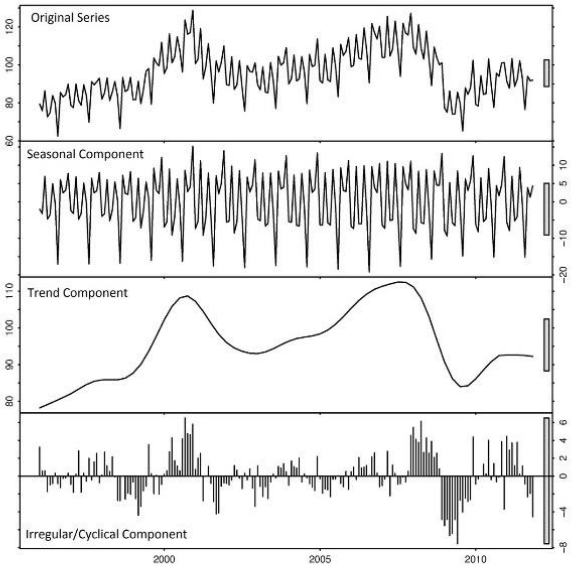

Figure 2. Data decomposition [12]

2.2.1. Create Trend equation

In this research, the researchers were interested in studying the method of trend equation creation and comparing the effectiveness of the equations, As follow:

- Linear Regression

$$y_t = mT_t + c \tag{1}$$

Where y_t represents the data at time t, T_t represents the data at time t, m represents the slope of the dataset, and c represents the intercept of the dataset.

- Quadratic Regression

$$y = \alpha x^2 + \beta x + \kappa \tag{2}$$

The coefficients of α, β, and κ can be calculated from equation (3) as follows.

$$\begin{bmatrix} T_t^4 & T_t^3 & T_t^2 \\ T_t^3 & T_t^2 & T_t \\ T_t^2 & T_t & n \end{bmatrix} \times \begin{bmatrix} \alpha \\ \beta \\ \kappa \end{bmatrix} = \begin{bmatrix} \sum_{t=1}^{n} T_t^2 y_t \\ \sum_{t=1}^{n} T_t y_t \\ \sum_{t=1}^{n} y_t \end{bmatrix} \tag{3}$$

Where y is the data at time t, T is the time data at time t, and a, B, and k are constants.

.

- Seasonal effect

$$S_s = \bar{Y}_s - \bar{Y} \tag{4}$$

Where S_s is the seasonal index, $\bar{Y}_s$ is the seasonal influence, and $\bar{Y}$ is the average of all data.

2.3. Box-Jenkins

The Box-Jenkins method is a technique widely recognized for analysing and forecasting time series data using Autoregressive Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) models. These models can be represented as $ARMA(p,q)$ or $ARIMA(p,d,q)$ where $AR(p)$ is Autoregressive, $I(d)$ is Integrated and $MA(q)$ is Moving average. The steps involved in building a forecasting model using the Box-Jenkins method consist of four steps, namely:

2.3.1. Stationarity test in time series

A stationarity test in time series can be determined from the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series whether the time series is stationary or non-stationary. If it be discovered that the time series is non-stationary, differencing must be used to convert the time series into a stationary one. Finding the differences between successive observations is the process of differencing. The value $I(d)$, which represents the number of differencing steps required to achieve stationarity.

2.3.2. Model Identification

- if Time series is stationary: $ARMA(p,q)$
- if Time series is non-stationary: $ARIMA(p,d,q)$

The values of $AR(p)$ and $MA(q)$ can be determined from the PACF and ACF graphs. Whereas it shows the value of $MA(q)$ in the PACF graph, and The number of significant components in the ACF graph indicates the value of $AR(p)$. Therefore, the model format of the time series can be determined by estimating the parameters of the model, as shown in Equations (5) for $ARMA(p,q)$ and (6) for $ARIMA(p,d,q)$ as follows.

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{5}$$

$$\Delta Y_t = \alpha_0 + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{6}$$

Where: ε_t stands for the error or residual, and Y_t is the data at time t. α_0 denotes parameters $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$ and $\theta_p(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_p B^p$

2.3.3. Model selection

After a credible model has been identified, its values are compared using the Akaike information criterion (AIC) to determine which model is best. The best-fit model is the one with the lowest AIC score. Equation (7) can be used to make this comparison. Where: N is the number of information points. SS_e stands for the total squared error. The number of parameters is symbolized by K. as follows:

$$AIC = N \times \ln\left(\frac{SS_e}{N}\right) + 2K \tag{7}$$

2.4. Random Forest (RF)

Forest is a learning method developed from Decision Trees (DT) for prediction in the form of multiple DT, created by randomly sampling data with replacement to construct tree models. Each tree within the forest does not exhibits redundant characteristics. Each DT predicts outcomes, and the final prediction results from averaging the outcomes of each DT, as illustrated in Figure 3

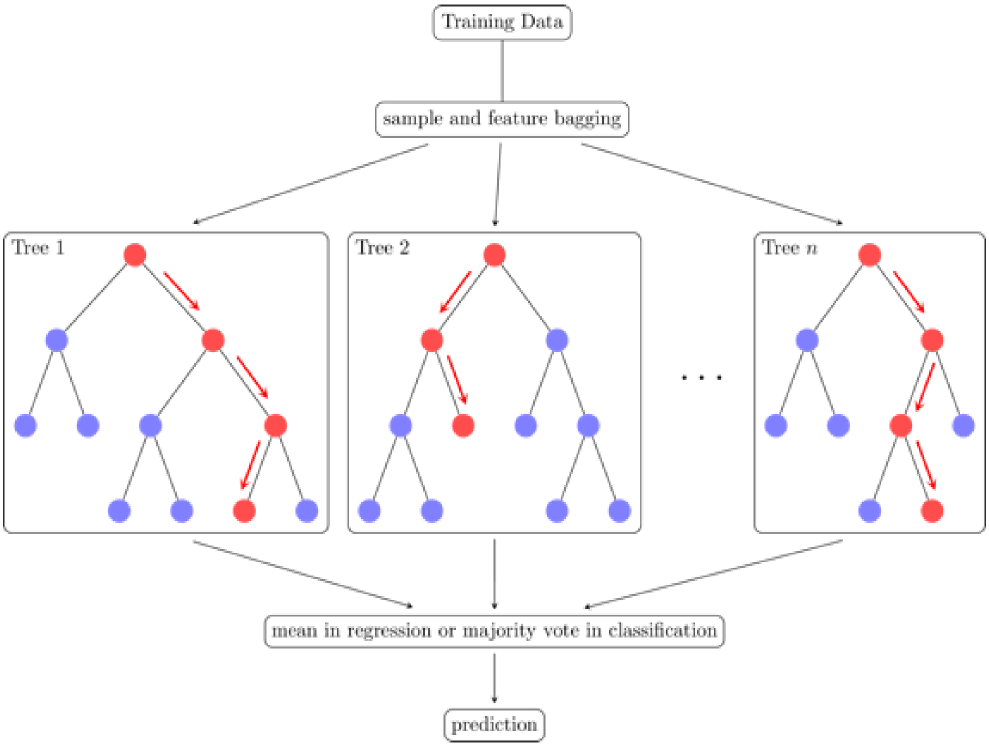
Training Data
sample and feature bagging
Tree 1
Tree 2
Tree n
. . .
mean in regression or majority vote in classification
prediction

Figure 3. Operation of the Random Forest Algorithm [11]

2.. Comparison of forecasting model performance

In this research, the researchers chose to investigate the three methods to compare the accuracy of forecasting model for the number of patients with Dengue fever which are Mean Square Error (MSE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) as shown in Equation (8), (9) and (10) Where: The total number of time periods is n, and e is the difference between the actual value and the anticipated value at time t. t stands for the time unit.

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n} \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}} \tag{9}$$

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|e_i|}{Y_i}}{n} \times 100\% \tag{10}$$

3. RESULTS AND DISCUSSIONS

3.1. Trend Result

The comparison results after using the decomposition method to find the trend using linear regression and quadratic regression equations

.

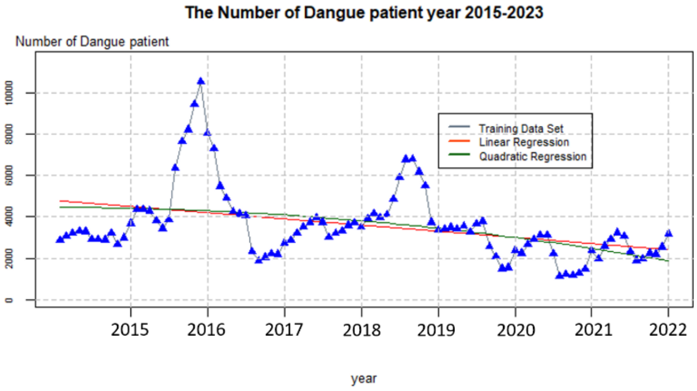

Figure 4. Graph showing the comparison of trend equations after using the decomposition method between linear regression and quadratic regression equations.

Table 1: Table showing the comparison of the efficiency of trend equations using the decomposition method.

	Decomposition	
	Linear regression	**Quadratic regression**
Equations	$y_t = -22.792t + 2897.165$	$y_t = -0.2797t^2 + 2664.5478$
MSE	2,531,769	**2,402,303**
RMSE	1,591.15	**1,539.936**
MAPE	86.10	**83.59**

Table 2: Seasonal indices of the data set.

month	**Seasonal indices**	**month**	**Seasonal indices**
Jan	-2.277	Jul	894.055
Feb	-734.355	Aug	1022.579
Mar	-999.765	Sep	750.138
Apr	-1250.91	Oct	946.674
May	-826.664	Nov	615.162
Jun	36.555	Dec	-451.188

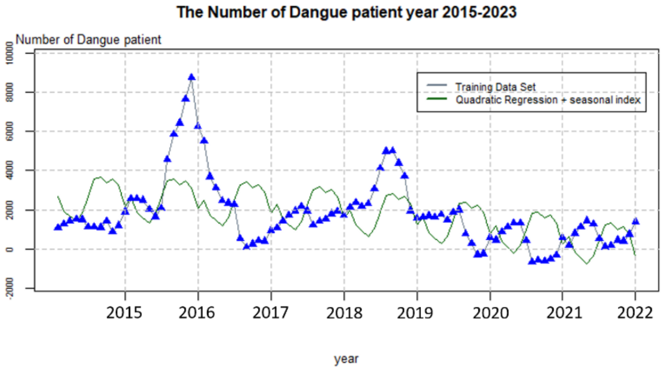

Figure 5: Graph comparing the trend equation of the combined quadratic regression equation with the seasonal index.

From Table 1, when comparing the efficiency of the trend equations using MAPE, MSE, and RMSE, it is observed that the quadratic regression equation gives lower values than the linear regression equation. Therefore, it can be concluded that the quadratic regression equation is the most suitable for creating a trend equation for the dengue fever data set, as shown in Figure 6. Then, combining the quadratic regression equation with the seasonal indices for all 12 months in Table 2, the resulting equation is (12) as follows.

$$y_t = -0.2797t^2 + 2664.5478 + S_s \tag{11}$$

From equation (11), the graph of the combined trend equation with the seasonal index is shown in Figure 5

3.2. Model Results

3.2.1. Box-Jenkins model

The results of parameter selection of p, d and q to construct possible ARIMA models and perform diagnostics are shown in table 3.

Table 3. All possible models from Box-Jenkins methods

Model	AIC
$ARIMA(1,1,0)$	**1548.605**
$ARIMA(2,1,0)$	1550.251
$ARIMA(0,1,0)$	1556.239
$ARIMA(1,1,1)$	1550.400
$ARIMA(0,1,1)$	1550.392
$ARIMA(2,1,1)$	1552.113
The best model is $ARIMA(1,1,0)$	

From Table 1, when comparing the performance of forecasting models using AIC, it is found that the $ARIMA(1,1,0)$ model is the most efficient for forecasting. Additionally, upon residual diagnostics, it is observed that the residuals exhibit a distribution close to normality, with a mean value near zero and constant variance, indicating that the random fluctuations are independent and stationary, resulting to no correlations in the residuals, as shown in Figure 4. Therefore, the $ARIMA(1,1,0)$ model is the most suitable model from the Box-Jenkins method. The prediction results of the model compared to the actual data values are illustrated in Figure 5

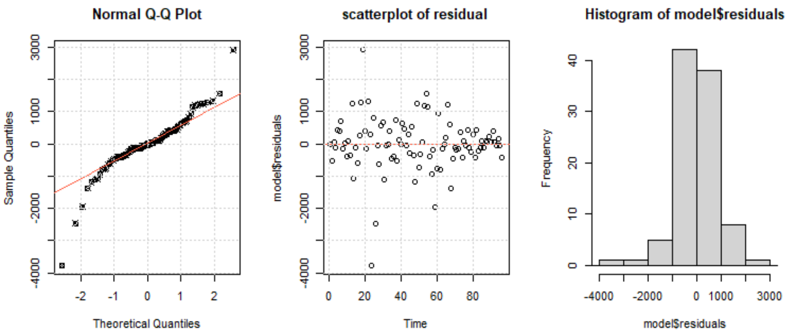

Figure 4. Scatterplot of Residuals from the $ARIMA(1,1,0)$ Model

.

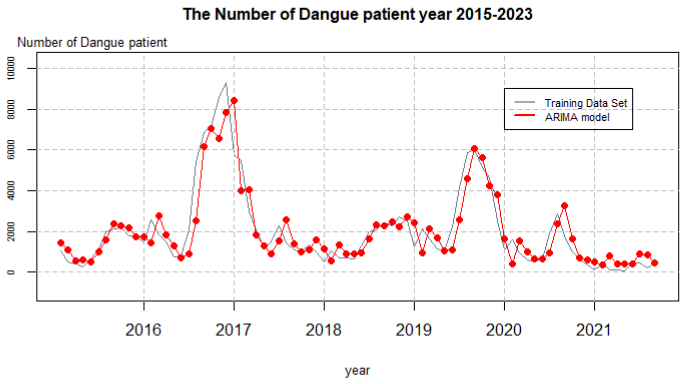

Figure 5. Graph the comparison between the dataset of dengue fever patients and the $ARIMA(1,1,0)$ model.

3.2.2. Random Forest model

The graph in Figure 6 displays the outcomes of building forecasting models with Random Forest.

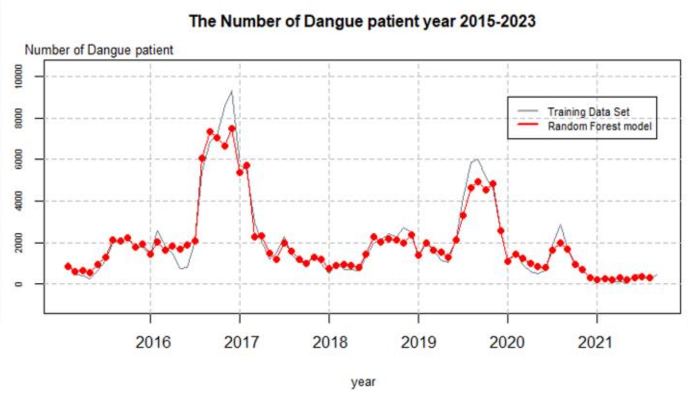

Figure 6. the comparison between the predicted and actual data of the RF prediction model.

3.3.Comparative Analysis of Forecasting Models

For forecasting the number of dengue fever patients between the Box-Jenkins method and Artificial Neural Network technique after conducting diagnostics.

Table 2. Graph compairing the performance of forecasting models

	Performance Comparison of Forecasting Models	
	Box-Jenkins	**RF**
Model	$ARIMA(1,1,0)$	$RFmodel$
MSE	504,046.041	869,002.433
RMSE	709.962	932.203
MAPE	61.843	63.465

From Table 2, when comparing the performance of models using MSE, RMSE, and MAPE, it is found that the $ARIMA(1,1,0)$ model outperforms the models obtained from the Random Forest method. Therefore, the $ARIMA(1,1,0)$ model is selected to forecast the number of dengue fever patients in the 12th Public Health Region.

3.4. Experimental Results of Forecasting Using Box-Jenkins Method with $ARIMA(1,1,0)$ model

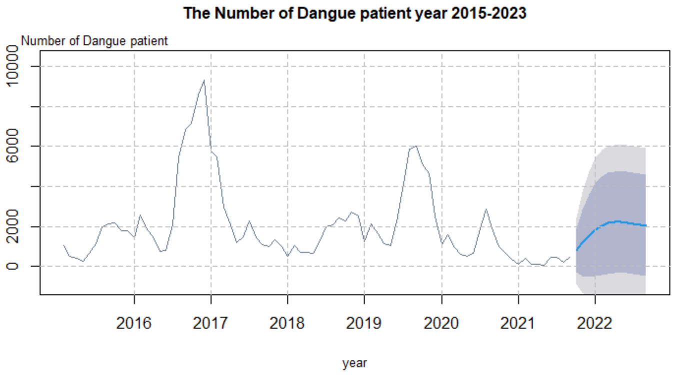

Figure 7. The illustration demonstrates the forecasting results using the ARIMA method with the $ARIMA(1,1,0)$ model

4. CONCLUSION

The analysis results reveal that the trend can be defined in the $y_t = -0.2797t^2 + 2664.5478$ equation from the Quadratic Regression. To improve the accuracy of the trend results, a seasonal index will be incorporated into the equation. From experimenting with the forecasting models using both the Box-Jenkins method and the Random Forest method, it was found that the Box-Jenkins method, particularly the $ARIMA(1,1,0)$ model, as the model shows a superior performance across all three accuracy metrics, showcasing MAPE at 61.84, MSE at 504,046.04, and RMSE at 709.96 achieved the best performance in forecasting.

ACKNOWLEDGEMENTS

The researchers would like to extend their heartfelt thanks to the Ministry of Health for supplying the essential information for this study. This study would not have been accomplished without the expert guidance, insightful suggestions, and support in equipment provided by Mr. Marusdee Yusoh, a dedicated mathematics teacher and lecturer at the Department of Mathematical Science and Technology, as well as the Islamic Science Demonstration School.

Finally, this project would not have come to completion this academic year without our family and friends having supported us throughout the entire process.

.

REFERENCES

[1] Warangkana Kirativibul. (2016). Forecasting Model of Pneumonia Patients in Thailand. Burapha University Journal of Public Health, 11(1), 24-38.

[2] Pirawan Nuansen, Prasit Payakkapong, and Thidaphorn Suppakorn. (2015). A Comparison of Forecasting Models for Crude Oil Production Quantity in Thailand. Bangkok: Kasetsart University.

[3] Preecha Kreusom and colleagues. (2023). Grey System Model for Forecasting Dengue Fever Outbreaks: A Case Study of Bangkok. Nonthaburi Rajabhat University.

[4] Jiraroj Tosasukul. (2021). Modeling Dengue Fever Outbreak Forecasting Using Data Mining Techniques. Faculty of Science, Naresuan University.

[5] Narongdej Intharat Chaiyakit. (2023). Estimation of Soil CBR using Artificial Neural Network. Faculty of Engineering and Architecture, Rajamangala University of Technology East.

[6] Dengue Fever. (2022). [Online]. Accessed November 12, 2022. Available from: https://ddc.moph.go.th/disease_detail.php?d=44

[7] Chutimonphakdirot, Ekkasit Patcharavongsa. (2556). Product Demand Forecasting Using Machine Learning Techniques in Retail Business. Burapha University.

[8] Van Hiep Phung (2019). A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets: Department of Computer Engineering, Hanbat National University, Daejeon 34158, Korea

[9] Peera Sombatdee (2015). Knowlage about Dengue Fever, Dengue Hemorrhagic Fever: srinagarind hospital khon kaen university

[10] Janosh Riebesell (2021). [Online]. Accessed May 9, 2022. Available from: https://tikz.net/random-forest/

[11] Chom Panta, & Yuphawadee Samranrut. (2017). Forecasting monthly rainfall in Nakhon Sawan province using statistical forecasting techniques.. SCIENCE AND TECHNOLOGY NAKHON SAWAN RAJABHAT UNIVERSITY JOURNAL, 9(10), 127-142.

[12] Libesa, (2014), Using decomposition to improve time series prediction, Accessed Sep 26, 2014. Available from: https://quantdare.com/decomposition-to-improve-time-series-prediction/

BIOGRAPHIES OF AUTHORS

	Name : Wafeeqismail Noipom Gender : Male Date of Birth : 29 September 2006 Address : 181/73, Moo 6, Rusamilae Subdistrict, Mueang Pattani District, Pattani Province, Pattani 94000 Thailand School : Islamic Sciences Demonstration School, Prince of Songkla University Pattani Campus, Pattani, Thailand. Email : wafeeqismail.n@ids.ac.th
	Name : Varees Adulyasas Gender : Female Date of Birth : 17 May 2007 Address : 9 Moo.2 Sub-district Thatong, District Raman, Yala 95140 Thailand School : Islamic Sciences Demonstration School, Prince of Songkla University Pattani Campus, Pattani, Thailand. Email : n varees.a@ids.ac.th
	Name : Rattifa Kasa Gender : Female Date of Birth : 26 March 2007 Address : 3 Moo.1 Sub-district Bankhuan,District Mueang, Satun 91140 Thailand School : Islamic Sciences Demonstration School, Prince of Songkla University Pattani Campus, Pattani, Thailand. Email : rattifa.k@ids.ac.th
	Name : Marusdee yusoh Gender : Male Date of Birth : 02 October 1993 Address : 181, Moo 6, Rusamilae Subdistrict, Mueang Pattani District, Pattani Province, Pattani 94000 Thailand School : Islamic Sciences Demonstration School, Prince of Songkla University Pattani Campus, Pattani, Thailand. Email : marusdee.y@psu.ac.th