Comparative Analysis of ARIMA and ANN Models for Forecasting Dengue Fever Incidence: A Case Study of Health District 12, Thailand

Wafeeqismail Noipom[1], Marusdee Yusoh[2], Varees Adulyasas[3], Rattifa Kasa[4]

Islamic Sciences Demonstration School, Faculty of Islamic Sciences, Prince of Songkhla University Pattani Campus, Pattani, Thailand

wafeeqismail.n@ids.ac.th[1], marusdee.y@psu.ac.th[2], vareesadulyasas@gmail.com[3], rattifa.k@ids.ac.th[4]

Abstract

In recent years, Thailand has faced recurring outbreaks of dengue fever, prompting a focused investigation into the southern region's Health district 12. This study concentrates on the incidence of dengue fever patients within this region, encompassing Songkhla, Satun, Trang, Phatthalung, Pattani, Yala, and Narathiwat provinces, aiming to develop forecasting models using data spanning from 2015 to 2023. Two distinct methodologies were explored: the Box-Jenkins Method and Artificial Neural Network (ANN) Technique. These methods were applied to a training set that was previously partitioned from the main dataset, containing patient counts. Furthermore, the main dataset underwent cross-validation for partitioning. Evaluation of forecasting models was conducted using three key accuracy metrics: Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). The results indicate that the Artificial Neural Network technique, particularly the $NNAR(2,1,2)$ model, emerges as the most suitable for constructing forecasting models. This model demonstrates superior performance, exhibiting the lowest values across all three-accuracy metrics: MAPE at 83.590, MSE at 2,402,303, and RMSE at 1,539.936.

Keywords*: Dengue fever; Time Series; Artificial Neural Network; Box-Jenkins; Cross-validation*

Introduction

Currently, dengue fever is the most significant public health issue in humid tropical countries. Generally, it prevails during the rainy season. Moreover, Thailand has reported dengue fever outbreaks for over 50 years (Piyarat et al., 2015). Dengue fever is caused by the dengue virus, transmitted by mosquitoes. According to the reports from the Department of Disease Control, from the beginning of 2023 to the end of the year (December 13, 2023), there were a total of 147,412

reported cases of dengue fever, with 174 fatalities across 57 provinces. Additionally, it was found that the number of dengue fever cases in the 12th Public Health Region, comprising Songkhla, Satun, Trang, Pattani, Yala, and Narathiwat, has been continuously increasing. A survey of the mosquito larval index in epidemic-prone areas showed that in the 12th Public Health Region, the House Index (HI) was 24.11% and the Container Index (CI) was 9.91%, exceeding the standard criteria. This indicates a high-risk area for the transmission of mosquito-borne infectious diseases. Due to inadequate and inconsistent efforts in mosquito breeding site elimination and lack of cooperation from the community in controlling mosquito breeding sites within their homes, the number of dengue fever cases has risen significantly. 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. Over the past years, numerous researchers have proposed and compared mathematical and statistical models using time series data to forecast the occurrence of various diseases in Thailand in the year 2015. Vorangkanat (2015) developed the most appropriate forecasting model for predicting the number of pneumonia patients in Thailand. They created forecasting models using three statistical methods: Box-Jenkins method, polynomial regression method, and combined forecasting method. The research findings indicated that the combined forecasting method was the most suitable for this specific time series dataset. Similarly, Piraoratan (2019) conducted a study to compare time series analysis techniques for predicting crude oil production quantities in Thailand. They studied three methods: decomposition method, Box-Jenkins method, and Grey model forecasting method. The results revealed that forecasting using the Box-Jenkins method was the most appropriate. Additionally, in 2016, Jiraroj (2021) sought to find models for predicting the outbreak of dengue fever within the jurisdiction of the second disease control office in Phitsanulok province, Thailand, using data mining techniques. Four methods were applied: decision tree method, Bayesian method, neural network method, and support vector machine method. The research concluded that the decision tree method was the most suitable

technique for creating a model to forecast dengue fever outbreaks. Furthermore, in 2024, Pricha et al. investigated the modeling of dengue fever outbreak prediction using a technique suitable for the data format with the Grey theory. The research found that the Grey theory method within the GM (1,1) system was suitable for forecasting the number of dengue fever cases.

The objective of this study 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 centers on comparing the efficacy of two forecasting methodologies: Box-Jenkins and Artificial Neural Network methods.

Methods and Experimental Details

1. Dataset

The dataset of the number of patients with dengue fever is collected as a monthly time series from the year 2015 to 2023, sourced from the Ministry of Public Health's website at https://pnb.hdc.moph.go.th/hdc/main/index.php. The dataset comprises a total of 100 entries. For this research, the data was divided into two parts: a Training Set and a Testing Set, in an 80:20 ratio, utilizing the R programming language for data analysis. Additionally, two methods were selected for analysis: the Box-Jenkins method and the Artificial Neural Network (ANN) technique as depicted in Figure 1. The details of each method are as follows:

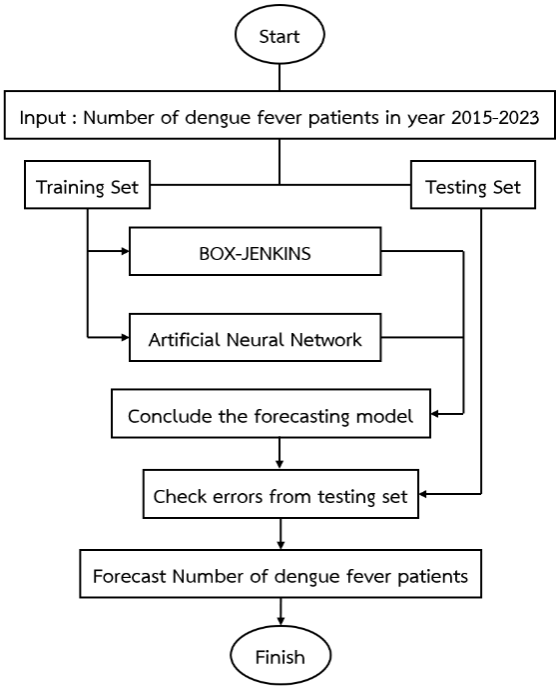

Figure 1: An overview illustrating the process of trend equation formulation and forecasting model development.

2. Box-Jekins

The Box-Jenkins method is a technique widely recognized for analyzing 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:

2.1 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 is found that the time series is non-stationary, it is necessary to transform the time series into a stationary one by differencing. Differencing involves finding the differences between consecutive observations. This process of making the time series stationary through differencing yields the value $I(d)$, which represents the number of differencing steps required to achieve stationarity.

2.2 Model Identification

- $ARMA(p,q)$ is used when the time series is stationary.
- $ARIMA(p,d,q)$ is used when the time series is non-stationary.

The values of $AR(p)$ and $MA(q)$ can be determined from the PACF and ACF graphs. The number of significant components in the ACF graph indicates the value of $MA(q)$, while in the PACF graph, it indicates the value of $AR(p)$. Therefore, the model term of the time series can be determined by estimating the parameters of the model, as shown in Equations (1) and (2) as follows.

For, $ARMA(p,q)$ model :

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

For, $ARIMA(p,d,q)$ model :

$$\Delta Y_t = \alpha_0 + \sum_{i=1}^{p} \phi_i \Delta Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (2)$$

Where: Y_t represents the data at time t. ε_t denotes the error or residual. α_o signifies the parameter. and $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$

$\theta_p(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_p B^p$

2.3 Model selection

Once a plausible model has been obtained, it is then examined to find the most suitable model by comparing the Akaike information criterion (AIC) values. The model that yields the lowest AIC value is considered the best-fit model. This comparison can be made using the Equation (3) as follows:

$$AIC = N \times \ln(\frac{ss_e}{N}) + 2K \quad (3)$$

Where: N represents the number of data points. ss_e denotes the sum of squared errors. K signifies the number of parameters.

3.Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a mathematical computing system that mimics the structure and functioning of the human

nervous system. It is a machine learning technique used to analyze data. The components of an ANN are divided into three parts: Input layer, Hidden layer, and Output layer. They consist of weights (w), biases (b), and activation functions, which simulate the properties of nerve cells. In this research, one variable input is imported, which is the number of patients with Dengue fever. Then, a feedforward process is carried out to calculate the results in each node according to the relations between the values of w and b. Subsequently, the residuals between the error values and the obtained results with the expected results are calculated. This is followed by backpropagation to adjust the values of w and b in the ANN and iterate this process until the minimum residual value is achieved, as shown in Figure 2. The forecasting model of the ANN method can be represented by Equations (4)

$$y_t = w_0 + \sum_{j=1}^{Q} w_j g(w_{0j} + \sum_{i=1}^{P} w_{i,j} y_{t-i}) \quad (4)$$

In this equation, y_t represents the output value, $y_{t-i}(i = 1, 2, ..., P)$ P , Q represents the number of input and hidden nodes respectively, where g is the Sigmoid transfer function. w_j denotes the weights from the hidden layer to the output node, while $w_{i,j}$ represents the weights from the input to the hidden node. w_0 and w_{0j} represent the bias terms.

3.1 Neural Network Autoregressive (NNAR)

The Neural Network Autoregressive (NNAR) forecasting model is an ANN model where the input layer consists of only one variable representing lagged values. It includes lag 1, lag 2, and so on, up to lag p in the model, hence it is called ANN Autoregressive (NNAR). NNAR was introduced using the R statistical application program in the "forecast" package along with the net function. This model is designed for feedforward networks with only one hidden layer and is denoted as $NNAR(p, P, k)$, where p represents lag-p as input, for seasonal time series the default values is $P = 1$ and k represents notes in the hidden layer. The NNAR method employs only one hidden layer and utilizes nonlinear functions to assign weights and generate outputs from the ANN. The activation function employs the binary sigmoid function. This study utilizes the $NNAR(p, P, k)$ model.

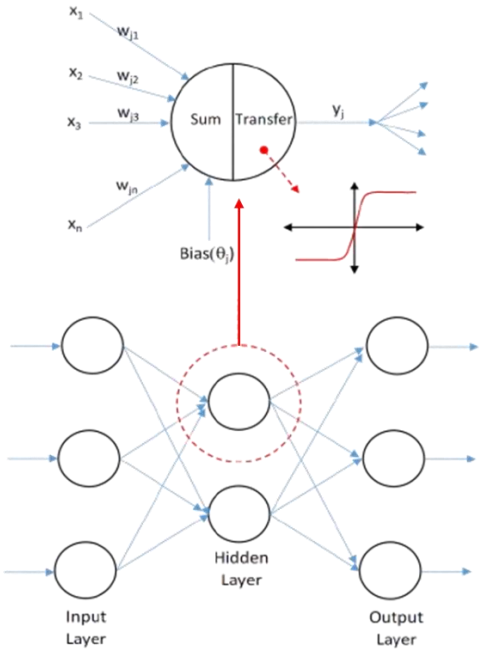

Figure 2: Artificial Neural Network process (Narongdej Intarasitthichai, 2022).

4. Cross-validation

Cross-validation is a technique used to evaluate the performance of predictive models in machine learning or statistical analysis. It involves dividing the data into K sets (folds), where each fold is used as a test set once, and all folds are iterated through until each fold has been used for testing. This ensures that every data point is used for both training and testing. As depicted in Figure 3, the performance metric of the model is processed from testing in each iteration, making the evaluation of the model challenging and producing more reliable results. In this research, K=5 folds are set to test the performance of the model.

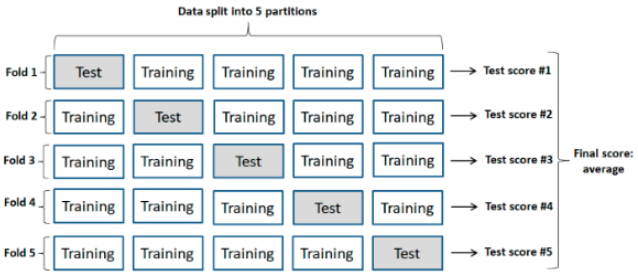

Figure 3: Procedure of Cross-Validation (Van Hiep Phung, 2019).

5. Comparison of forecasting model performance

In this research, the researchers chose to employ three methods to compare the accuracy of forecasting model for the number of patients with Dengue fever:

1. Mean Square Error (MSE), which can be calculated as shown in Equation (5)

$$MSE = \frac{\sum_{t=1}^{n} e_t^2}{n} \quad (5)$$

2. Root Mean Square Error (RMSE), represented by Equation (6)

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} e_t^2}{n}} \quad (6)$$

3.Mean Absolute Percentage Error (MAPE), which can be calculated as shown in Equation (7)

$$MAPE = \frac{\sum_{i=1}^{n} \frac{|e_i|}{Y_i}}{n} \times 100\% \quad (7)$$

Where: e_i represents the difference between the actual value and the forecast value at time t. n is the total number of time periods. t denotes the time unit

Result and Discussion

1. Model Result

1.1 Box-Jenkins

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

Table 1: 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

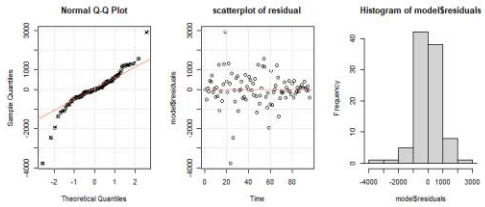

Figure 4: Scatterplot of Residuals from the $ARIMA(1,1,0)$ Model

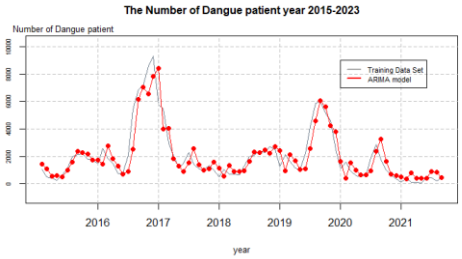

Figure 5: Graph illustrating the comparison between the $ARIMA(1,1,0)$ model and the dataset of dengue fever patients.

1.2 Artificial Neural Network (ANN)

The results from constructing forecasting models using Artificial Neural Network is shown in the graph in Figure 6

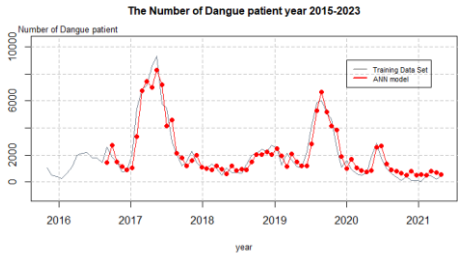

Figure 6: Graph illustrating the comparison between the $NNAR(2,1,2)$ model and the dataset of dengue fever patients.

2. 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	**ANN**
Model	$ARIMA(1,1,0)$	$NNAR(2,1,2)$
MSE	504,046.041	**439,337.643**
RMSE	709.962	**662.825**
MAPE	61.843	**68.205**

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

3. Experimental Results of Forecasting Using Artificial Neural Network Technique with $NNAR(2,1,2)$ Model

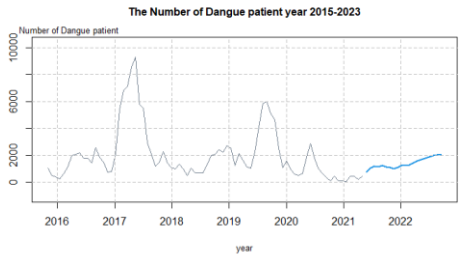

Figure 7: The illustration demonstrating the forecasting results using the ANN technique with the $NNAR(2,1,2)$ model

Conclusion

The analysis results reveal that from experimenting with the forecasting models using both the Box-Jenkins method and ANN, it was found that the ANN method, particularly the $NNAR(2,1,2)$ model, achieved the best performance in forecasting.

Acknowledgements

The researchers thanked the Ministry of Health for providing information for this study. The completion of this study could not have been possible without Mr.Marusdee Yusoh a mathematics teacher and lecturer from the Department of Mathematical Science and Technology, and Islamic Science Demonstration School for the advice and suggestions, as well as their support in equipment.

Reference

1. Ministry of Public Health. (2014). Dengue Hemorrhagic Fever Incidence Rate. October 30, 2023. Retrieved from: https://tinyurl.com/yzk695jp.
2. Warangkana Kirativibul. (2016). Forecasting Model of Pneumonia Patients in Thailand. Burapha University Journal of Public Health, 11(1), 24-38.
3. Pirawannan Nuansen, Prasit Payakkapong, and Thidaphorn Suppakorn. (2015). A Comparison of Forecasting Models for Crude Oil Production Quantity in Thailand. Bangkok: Kasetsart University.
4. Preecha Kreusom and colleagues. (2023). Grey System Model for Forecasting Dengue Fever Outbreaks: A Case Study of Bangkok. Nonthaburi Rajabhat University. Jiraroj Tosasukul. (2564).
5. Modeling Dengue Fever Outbreak Forecasting Using Data Mining Techniques. Faculty of Science, Naresuan University.
6. Narongdej Intharat Chaiyakit. (2566). Estimation of Soil CBR using Artificial Neural Network. Faculty of Engineering and Architecture, Rajamangala University of Technology East.
7. Dengue Fever. (2565). [Online]. Accessed November 12, 2566. Available from: https://ddc.moph.go.th/disease_detail.php?d=44
8. Chutimonphakdirot, Ekkasit Patcharavongsa. (2556). Product Demand Forecasting Using Machine Learning Techniques in Retail Business. Burapha University.