Time Series Analysis for Forecasting Prevalence of Dengue Hemorrhagic Fever in Thailand

Suppawan Promprao[1], Suphat Rittirat[2], Siriluk Sintupachee[3] and Yhodpha Ratmanee[4]

[1,2,3,4] Mosquitoes Operational Research Unit (MORU), Faculty of Science and Technology, Nakhon Si Thammarat Rajabhat University, Mueang, Nakhon Si Thammarat, Thailand

Abstract

Dengue Hemorrhagic Fever (DHF) outbreaks was one of the lethal health problems in Thailand. *Aedes aegypti* was the main mosquito vector that transmitted the dengue virus from person to person. There was no available treatment other than supportive care. Responsibly forecasting where DHF outbreaks occur before dengue occurrence season could help public health officials prioritize public health activities to prevent and control this disease. To develop a forecasting model in anticipating the number of DHF cases in Thailand using time series analysis. We developed autoregressive integrated moving average (ARIMA) models on the data collected between January 2004–August 2019. The models were validated over the testing period (September–December 2019) and forecasted DHF cases from January–June 2020. The results revealed that the regressive forecast curves were consistent with the pattern of actual values. The ARIMA $(1,1,1)(1,1,1)_{12}$ model fitting was adequate for the data with the Q-statistic (Q=22.14, p-value =0.076). The observed and predicted DHF cases from January 2004–August 2019 matched reasonably well. Time-series forecasting of DHF cases in Thailand may offer the potential for improving planning, control and prevention by public health intervention.

Keywords: Dengue Hemorrhagic Fever, ARIMA, forecasting, Time Series

Introduction

People at risk for dengue disease inhabit tropical and subtropical regions around the world with an estimated 40% of the global population at risk for dengue infection and 390 million cases annually (WHO, 2017). Dengue Hemorrhagic Fever (DHF) outbreaks was one of the lethal health problems in Thailand. It was caused by the dengue virus, belonging to the genus Flavivirus, family Flaviviridae. DHF can occur when someone was bitten by a mosquito or exposed to blood infected with the dengue virus. Infected mosquitoes were the most common causes. *Aedes aegypti* was the main mosquito vector that transmitted the dengue virus from person to person in South-East Asia, including Thailand. (Halstead, 1990; Barbazan et al, 2002). Researchers are working on a vaccine to prevent dengue fever. However, it is currently unavailable. The best way to prevent dengue fever is to protect yourself from being bitten by mosquitos. There was no available treatment other than supportive care so that responsibly forecasting where DHF outbreaks occur before dengue occurrence season could help public health officials prioritize public health activities to prevent and control this disease (Gubler, 1998; Phun-Urai et al., 1995; Severson et al., 2004; Lauer et al., 2017). The objective of this study was to develop a forecasting model in anticipating the number of DHF cases in Thailand using time series analysis.

Materials and methods

We developed autoregressive integrated moving average (ARIMA) models on the data collected between January 2004–August 2019. The models were validated over the testing period (September–December 2019) and forecasted DHF cases from January–June 2020. This monthly DHF data were obtained from the Bureau of Epidemiology, Ministry of Public Health, Thailand.

Study Area

Thailand is located at 99° 58' E, 5° 37' N and 101° 8' E, in the middle of mainland Southeast Asia. It has a total size of 513,120 km^2 (198,120 sq mi) which is the 50^{th} largest in the world. The land border is 4,863 km (3,022 mi) long with Myanmar, Cambodia, Laos and Malaysia.

Statistical analysis

Autoregressive integrated moving average (ARIMA) models have been used for analyzing time series data containing seasonal trends to develop a forecasting model by Box and Jenkins (Wei, 1989). This method consisted of identification, estimation, diagnostic checking and forecasting. ARIMA model may include autoregressive (p) terms, differencing (d) terms and moving average (q) operations and is represented by ARIMA (p, d, q) (Box and Jenkins, 1976). Seasonal ARIMA (SARIMA) is an extension of the ARIMA method to a series which a pattern repeats seasonally over time. It usually causes the series to be nonstationary whose statistical properties such as mean and variance are not constant over time. The Seasonal autoregressive (P), seasonal differencing (D), and seasonal moving average parameters (Q) and s defines the number of time periods until the pattern repeats again for a monthly data was represented as SARIMA (p, d, q) (P, D, Q)s. SPSS for Windows version 24.0 was used to analysis and determine the fitting model.

First, a check for stationary was made with the aid of a control chart, which was a useful graphical device for detecting the lag of stationary in a time-series analysis. With 188 monthly values used for model synthesis, only correlations at the first 188/4 ≈ 47 lags needed to be examined. (Lim and McAleer, 2002). The basic idea was to superimpose reference lines, called control limits, on a time-series plot. A proposed definition of the limits used the mean ($\bar{x}$) and the standard deviation (S.D.) of the time series and regarded the lower control limit as LCL = -3S.D. and the upper control limit as UCL = +3S.D.

After verifying that the series was stationary, an ARIMA model was developed. An effort was made to express each observation as a linear function of the previous value of the series (autoregressive parameter) and of the past error effect (moving average parameter). The adequacy of the above model was checked by comparing the observed data (i.e. September–December 2019) with the forecasted results (i.e. January–December 2020). The requirements were confirmed by inspecting the graphs of the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). A set of lengthy time-series data was required for univariate time-series forecasting. It was usually recommended that at least 50 observations should be available (Hammon, 1960). Therefore, 188 observations were used in this study.

As a first step to model identification, the monthly DHF cases time series Y_t for 15 years and 8 months or 188 months were used for constructing the Univariate Box-Jenkins model, while data for the remaining 4 months were reserved for model evaluation. Using the general term of it was given: (Chang et al., 2012)

$$\Phi_P(B^s)\varphi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t$$

where, $\{w_t\}$ is the nonstationary time series, $\{w_t\}$ is the usual Gaussian white noise process. s is the period of the time series. The ordinary autoregressive and moving average components are represented by polynomials φ(B) and θ (B) of orders p and q. The seasonal autoregressive and moving average components are Φ_P (B^s) and Θ_Q (B^s), where P and Q are their orders. ∇^d and ∇_s^D are ordinary and seasonal difference components. B is the backshift operator. The expressions are shown as follows:

$$\begin{aligned}
&\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \\
&\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{ps} \\
&\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \\
&\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs} \\
&\nabla^d = (1 - B)^d \\
&\nabla_s^D = (1 - B^s)^D \\
&B^k x_t = x_{t-k}
\end{aligned}$$

The parameters for the tentative model from the identification step were estimated using the ARIMA module in SPSS. The model was used to forecast for 6 consecutive future months, based on the last available data point Y_{188} as the forecasting origin. At this stage, the statistical adequacy of the estimated tentative models was verified. Plotting the residuals of the estimated model was a useful diagnostic check. The autocorrelogram or autocorrelation of the residuals was plotted to check the model suitability. If the statistical properties of the sampled population were adequately modelled using the identified ARIMA processes, the residuals (ε_t) should be statistically independent (i.e. the residuals should not be correlated with each other (Abdel-Aal and Mangoud, 1998). In practice, ε_t was taken as the residuals, i.e. the differences between actual observations in the time-series model and corresponding values predicted by the estimated model.

The final test for an ARIMA model was its ability to forecast. We used the model from diagnostic checking step to forecast the number of DHF cases in Thailand. We assessed the adequacy of our model by checking whether (i) the model assumption was satisfied; (ii) the errors were normally distributed; and (iii) all residual ACF's were equal to zero by using the Q-Statistic Box-Ljung test as follow.

$$Q = [(n-d)\{(n-d)+2\}]\sum_{j=1}^{k}\frac{r_j^2}{[(n-d)-j]}$$

Results and discussion

The reported monthly DHF cases during 2004–August 2019 exhibited a seasonal pattern (Figure 1). It displayed seasonal fluctuations and therefore considered non stationary such as mean and variance are not constant over time. Autocorrelation function (ACF) and partial autocorrelation function (PACF) of the data were shown in Figure 2. Logarithm and first differencing were used to transform this DHF data and were expressed in Figure 3.

.

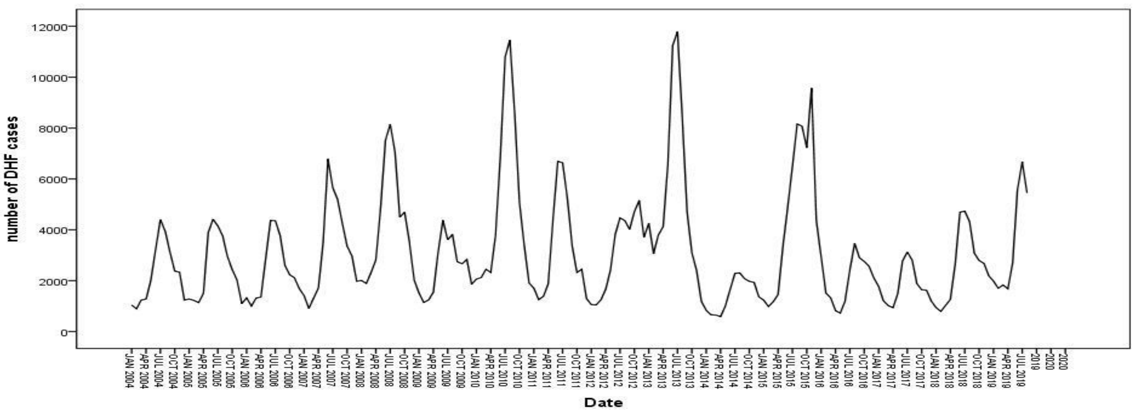

Figure 1 The monthly DHF cases time series (non stationary time series)

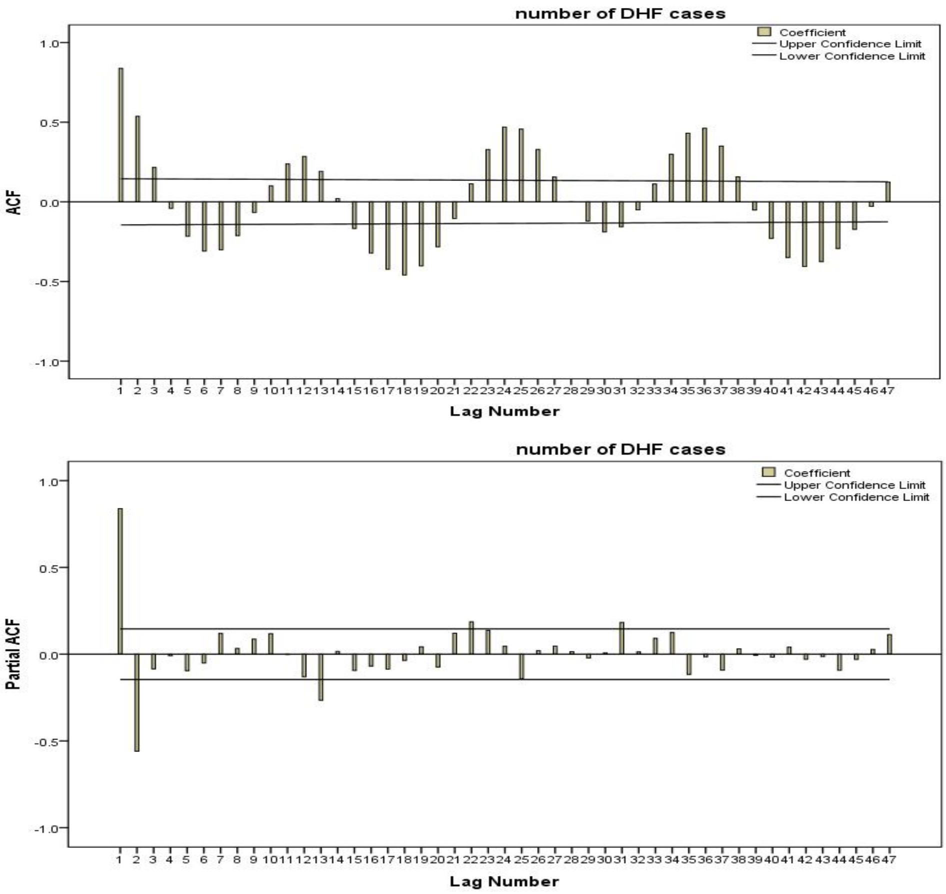

Figure 2 ACF and PACF of non stationary time series

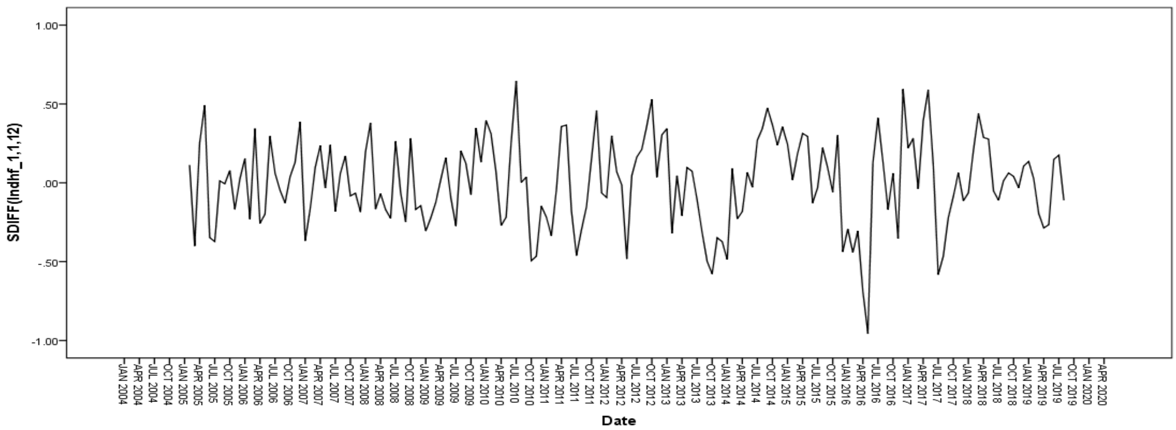

Figure 3 The monthly DHF cases time series (stationary time series)

The Autocorrelation function (ACF) and the partial autocorrelation function (PACF) of stationary time series plotted against 47-time lags were expressed in Figure 4.

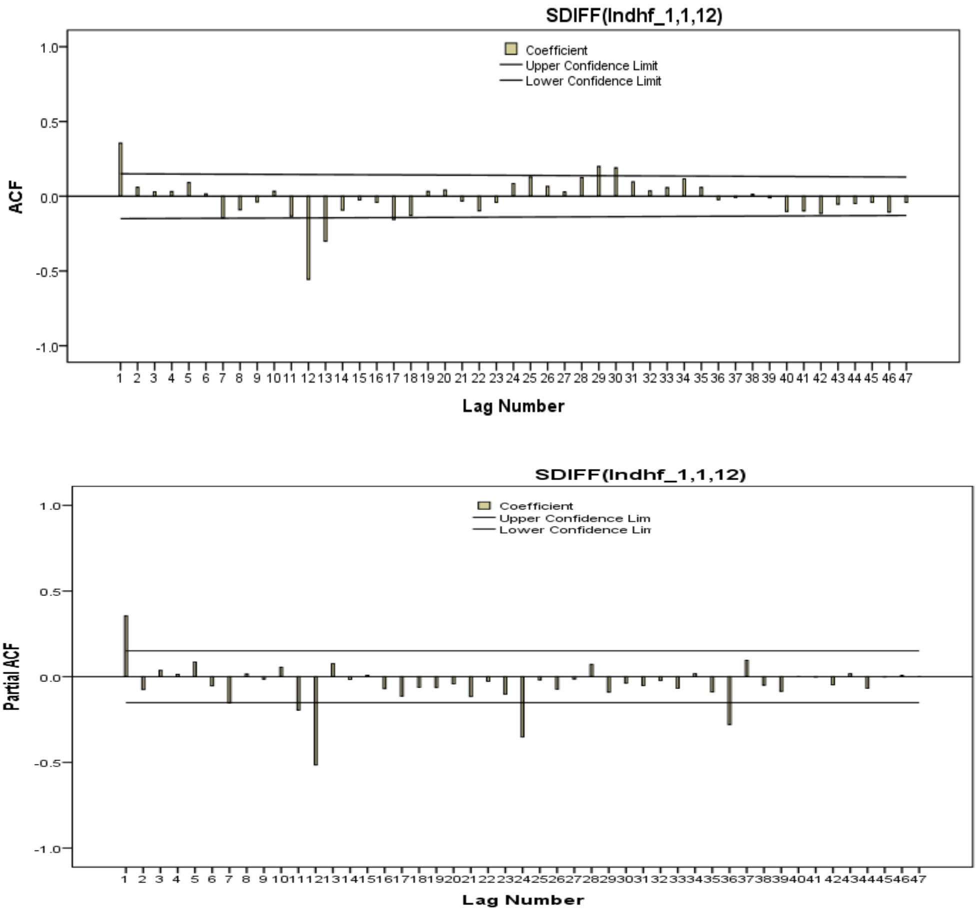

Figure 4 ACF and PACF of stationary time series

SARIMA (1,1,1) $(1,1,1)_{12}$ was selected as the best model, with the lowest BIC of -3.018 and p-value = 0.076. L jung-Box test > 0.05 suggested that there was no significant autocorrelation between residuals at different lag times (Table1). This was further confirmed by plotting the ACF and PACF of the residuals (Figure 5). The result showed the stationary series and predicted value matched reasonably well (Figure 6).

Table 1 Autocorrelation function and partial autocorrelation

Model	Const.	Q Stat	p-value	Estimates					BIC
				Const.	AR	MA	SAR	SMA	
SARIMA(0,1,1)(0,1,1)12	With	31.550	0.110	0.000		0.709*		0.858*	-2.953
	Without	31.550	0.110			0.707*		0.858*	-2.989
SARIMA(1,1,1)(1,1,1)12	With	17.978	0.208	0.000	0.366*	0.998	-0.130	0.820*	-3.022
	Without	22.140	0.076		0.487*	0.963*	-0.107	0.844*	-3.018
SARIMA(1,1,1)(0,1,1)12	With	19.610	0.187	0.000	0.373*	0.997		0.882*	-3.048
	Without	23.202	0.080		0.488*	0.963*		0.897*	-3.046

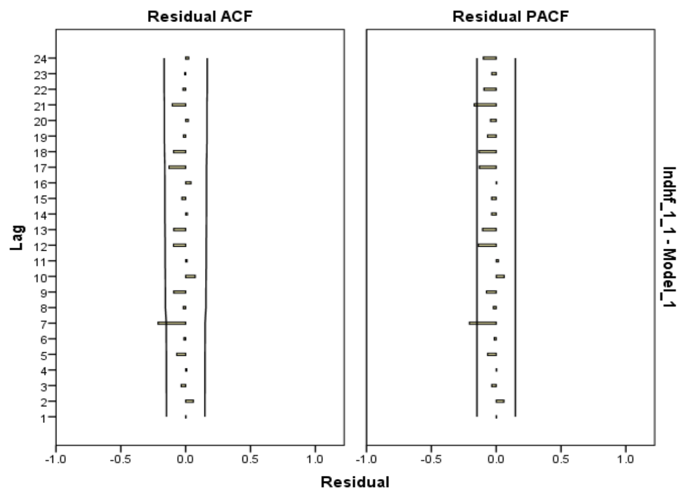

Figure 5 ACF and PACF of the residual of SARIMA (1,1,1) $(1,1,1)_{12}$

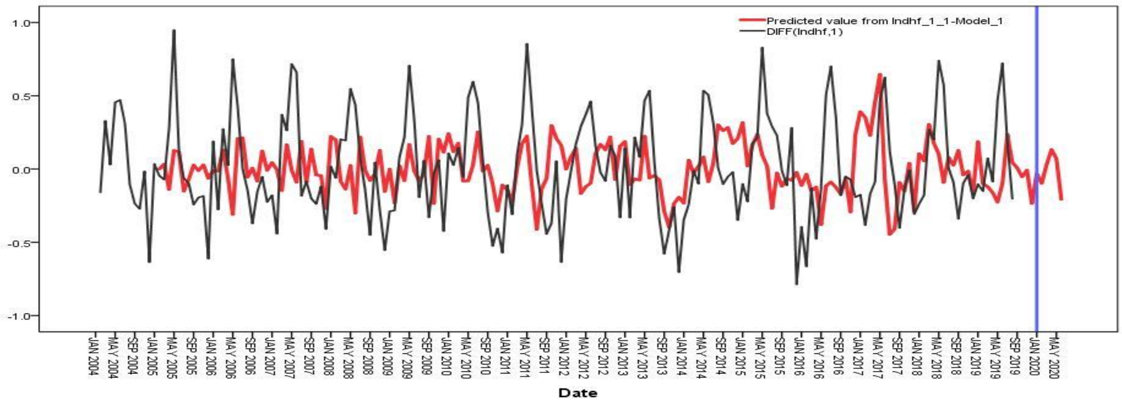

Figure 6 Stationary monthly DHF series and the predicted monthly series by using SARIMA (1,1,1) $(1,1,1)_{12}$ model

This study provides a guideline of applying a SARIMA model to forecast DHF prevalence in Thailand. These models have been applied to forecast DHF incidence in several countries. ARIMA modeling is a useful tool for revealing surveillance data and forecast DHF cases to good help guide timely prevention and control measures and may providing decision making in designing an effective prevention and control tactics for DHF prevalence in Thailand and/or nearby.

References

Abdel-Aal R.E., Mangoud A.M. (1998). Modeling and forecasting monthly patient volume at a primary health care clinic using univariate time-series analysis. Comput Programs Biomed, 56(3): 235-247.

Barbazan P, Yoksan S, Gonzalez JP. (2002). Dengue hemorrhagic fever epidemiology in Thailand: description and forecasting of epidemics. Microbes Infect, 4(7):699-705.

Box G. and Jenkins G. (1976). Time series analysis: Forecasting and Control, Holden-Day: San Francisco.

Bureau of Epidemiology, Ministry of Public Health. (2019). Surveillance Report (R 506) DHF Total. [online] Available at: http://www.boe.moph.go.th/boedb/surdata/ 506 wk.

Chang X., Gao M., Wang Y., Hou X. (2012). Seasonal autoregressive integrated moving average model for precipitation time series. J Math Stat, 8(4): 500-505.

Gubler DJ. (1998). Dengue and dengue haemorrhagic fever. Clin Microbiol Rev Jul;11(3):48096.

Halstead SB. (1990). Global epidemiology of dengue hemorrhagic fever. Southeast Asian J Trop Med Public Health, 21(4):636-641.

Hammon WM, Rudnick A, Sather GE. (1960). Viruses associated with epidemic hemorrhagic fevers of the Philippines and Thailand. Science, 131:1102-1103.

Lim C. and Mc Aleer M. (2002). Time series forecasts of international travel demand for Australia. Tourism Manag, 23: 389-396.

Lauer S.A., Sakrejdaa K., Rayb E.L., Keeganc L.T., Bic Q., Suangthod P., Hinjoyd S., Iamsirithaworne S., Suthachanad S., Laosiritawornd Y., Cummingsf D.A.T., Lesslerc J. and Reich N.G. (2017). Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014, Proceedings of the National Academy of Sciences March 2018, 115 (10) E2175-E2182; DOI: 10.1073/pnas.1714457115.

Phun-Urai P, Kong-ngamsuk W, Malainual N. (1995). Field trail of Bacillus thuringiensis H-14 (Larvitab) against Aedes aegypti larvae in Amphoe Khlung, Chanthaburi Province, Thailand. J Trop Med Parasit, 18:35-41.

Promprou S., Jaroensutasinee M. and Jaroensutasinee K. (2006). Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA models. Dengue Bull, 30**;** 99-106.

Severson DW, Knudson DL, Soares MB, Loftus BJ. (2004). *Aedes aegypti* genomics. Insect Biochem Mol Biol, 34(7):715-721.

Wei WWS. (1989). Time-series analysis. United States of America.

WHO, Dengue and severe dengue [(accessed on 9 February 2017)]; Available online: http://www.who.int/mediacentre/factsheets/fs117/en/