Forecasting Dengue Haemorrhagic Fever Cases in Southern Thailand using ARIMA Models

S. Promprou, M. Jaroensutasinee and K. Jaroensutasinee

School of Science, Walailak University, 222 Thaiburi, Thasala District, Nakhonsithammarat 80161, Thailand

Abstract

A univariate time-series analysis method has been used to model and forecast the monthly number of dengue haemorrhagic fever (DHF) cases in southern Thailand. We developed autoregressive integrated moving average (ARIMA) models on the data collected between 1994–2005 and then validated the models using the data collected between January–August 2006. The results showed that the regressive forecast curves were consistent with the pattern of actual values. The ARIMA (1,0,1) model fitting was adequate for the data with the Q-statistic (Q=4.446). This indicated that the autocorrelation function was not significantly different for zero.

Keywords: Autoregressive integrated moving average (ARIMA) models, dengue haemorrhagic fever (DHF), time-series, disease prediction.

Introduction

Dengue haemorrhagic fever (DHF) is one of the most important public health problems in Thailand and many other tropical countries around the world. It is caused by the dengue virus, belonging to the genus Flavivirus, family Flaviviridae. It is mainly transmitted by the mosquito Aedes aegypti that is a well-known principal vector of DHF in South-East Asia, including Thailand.[1,2] In the absence of any vaccines and treatment, vector control through source reduction with active community participation is the only viable option.[3-5]

DHF has been reported in Thailand since the late 1950s.[6-8] There has been an upward trend in the incidence of DHF, an acute and severe form of dengue virus infection. Since the first DHF epidemic outbreak in 1958,[2] epidemics have been reported from almost all parts of the country. The Bureau of Epidemiology has reported that there have been several outbreak reportings regularly in Thailand. The highest number of cases was reported in 1987 when the incidence rate was as high as 325 cases per 100 000 population based on the number of cases reported. The latest epidemic was in 1998 when the incidence rate was as high as 211 cases per 100 000 population. This was the second highest incidence rate in the 40 years’ history of DHF outbreaks (Bureau of Epidemiology, 2002).

During 2000–2004, the Southern Epidemiology Department reported 113 591 cases of DHF in southern Thailand, with 251 deaths. In 2002, the DHF case incidence in southern Thailand was especially high at 33 617 cases with 64 deaths, with a fatality rate of 0.77 per 100 000 population. During January–
June 2005 (18 June), DHF cases in southern Thailand were as high as 2991 cases with 5 deaths. This indicated that DHF was a major health risk in southern Thailand. The desirability of an early warning system for DHF epidemic was considered necessary to reduce the incidence rate of dengue and the intensity of DHF in the area.

Forecasting DHF cases in southern Thailand by using time-series models would provide useful information. The main characteristic of the time-series modelling is that it only models the relationship between the observed DHF cases at time $t$ ($y_t$) from the past observations ($y_1, y_2, ..., y_{t-1}$), without using any other variables. This study aimed at developing univariate time-series models for the monthly DHF cases of southern Thailand, based on reported cases available since 1994. This forecasting offers the potential for improved contingency planning of public health intervention.

**Materials and methods**

Southern Thailand is located at 11° 42' 52" N, 5° 37' 0" N, covers an area of 70 715.2 sq km, and extends from Chumphon province to the Thai-Malaysian border comprising of 14 provinces (Figure 1). The southern region is located on the peninsula between the Andaman Sea of the Indian Ocean to the west and the South China Sea of the Pacific Ocean to the east. This study used monthly DHF data from the 14 provinces from 1994–2005. This monthly DHF data were obtained from the Office of Disease Prevention and Control, Region Nos. 11 and 12, and from the Bureau of Epidemiology, Ministry of Public Health, Thailand.

The statistical package SPSS version 12.0 for Windows was used for developing the autoregressive integrated moving average (ARIMA) models. The ARIMA models were analysed with the Box-Jenkins approach, which was appropriate for a long forecasting period. This method for selecting an appropriate ARIMA model for estimating and forecasting a univariate time-series consisted of identification, estimation, diagnostic checking and forecasting. 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 144 monthly values used for model synthesis, only correlations at the first $144/4 = 36$ lags needed to be examined. 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 (\( \bar{Y} \)) and the standard deviation (S.D.) of the time series and regarded the lower control limit as LCL = \( \bar{Y} - 3 \text{S.D.} \) and the upper control limit as UCL = \( \bar{Y} + 3 \text{S.D.} \). 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. January–August 2006) with the forecasted results (i.e. January–December 2006). The above 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.\(^6\) Therefore, 144 observations were used in this study. As a first step to model identification, the monthly DHF cases time series \( Y_t \) for 12 years or 144 months were used for constructing the Univariate Box-Jenkins model, while data for the remaining 12 months were reserved for model evaluation. Using an ARIMA specification to identify the tentative model and the general term of it was as follows:

\[
Y_t = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3} + \ldots + \epsilon_t
\]

where \( Y_t \) represented the number of DHF cases at time \( t \)

\( Y_{t-1} \) represented the number of DHF cases at time \( t-1 \)

\( Y_{t-2} \) represented the number of DHF cases at time \( t-2 \)

\( \epsilon_t \) represented the error.

The parameters for the tentative model from the identification step were estimated using the ARIMA module in SPSS. Ten iterations were specified with the default tolerance of 0.001. The model was used to forecast \( Y_t \) for 12 consecutive future months, based on the last available data point \( Y_{144} \) 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 should be statistically independent (i.e. the residuals should not be correlated with each other\(^{13}\)). In practice, \( \epsilon_t \) 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 southern Thailand. We assessed the adequacy of our model by checking whether (i) the model assumption were 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.

**Results**

The ACF of DHF time-series appeared to be exponentially tailing off after lag \( t = 1 \) (Figure 2a), while the PACF of DHF time-series damped sine wave tailed off (Figure 2b).
Forecasting DHF Cases in Southern Thailand using ARIMA Models

Figure 2(a): Autocorrelation function of DHF time-series

Figure 2(b): Partial autocorrelation function of DHF time-series
When both the ACF and the PACF of time-series were tailed off, these results indicated a mixed ARIMA (1,0,1) or ARMA (1,1) (i.e. a zero order of differenced ARIMA model). The general form of this tentative model was as follows:

\[ Y_t = \phi_1 Y_{t-1} + a_t - \theta_1 a_{t-1} \]  

where \( \phi_1, \theta_1 \) were the two parameters or the model coefficients

\( a_t \) was a time-series of random shocks or white noise process at time \( t \)

\( a_{t-1} \) was white noise process at time \( t-1 \).

A process \( \{a_t\} \) was called a white noise process if it was a sequence of uncorrelated random variables from a fixed distribution with a constant mean \( E(a_t) = \mu_a \), usually assumed to be 0, a constant variance \( \text{Var}(a_t) = \sigma_a^2 \) and a covariance \( \gamma_k = \text{cov}(a_t, a_{t+k}) = 0 \) for all \( k \neq 0 \).

The coefficients were estimated as an autoregressive \( \phi_1 = 0.822 \) (\( t_1 = 16.84, P<0.001 \)), and a moving average \( \theta_1 = -0.732 \) (\( t_7 = -12.34, P<0.001 \)). The autoregressive coefficients were very close to their limit of stationarity, 1.0 (|\( \phi_1 \)|<1 and |\( \theta_1 \)|<1).[10,14] The tentative model was as follows:

\[ Y_t = 0.822Y_{t-1} + a_t + 0.732a_{t-1} \]  

It revealed that the DHF incidences at time \( t \) were approximately 82.2% of the DHF incidence at time \( t-1 \) plus a white noise process. The model in equation (2) was used to forecast DHF cases at time \( t \) (\( Y_t \)) for 12 sequential future months (i.e. January–December, 2006) based on the last obtainable data point \( Y_{144} \) (i.e. DHF incidences in December, 2005) as the forecasting origin.

The graphic analysis of residuals showed that the residuals in the model appeared to fluctuate randomly around zero with no obvious trend in variation as the predicted incidence values increased (Figure 3). A plot of the residual autocorrelation function died out after one lag and the residual autocorrelations fell within 95% confidence limit (Figure 4). This suggested that there was no significant autocorrelation between residuals at different time lags (Q Box-Ljung Statistic = 4.446, \( \chi^2_{0.05,32} = 18.5, P<0.05 \)). The observed and predicted DHF cases from 1994–2005 matched reasonably well. The predicted DHF cases for the year 2006 increased and reached a maximum predicted case level in December 2006 (Figure 5).

**Discussion**

The results from our study confirmed the existence of DHF cases in southern Thailand in 2006. It suggested that the number of DHF cases in the area would increase with a range of 403-1169 cases from January-December 2006. Based on the recent past year (i.e. 2005), our forecasting model indicated that the next expected high number of DHF cases in southern Thailand should occur in December 2006. DHF cases that we predicted were around 1169 cases with 95% confidence interval (CI) of -1881 to 4219 cases. Our results did not correspond with the DHF morbidity rate forecasting in Myanmar.[15] Their study had suggested that the next morbidity spike in Myanmar may occur in 2004, with the average morbidity rate of 127.6 and 95% CI of 22.4-232.8. In addition, their study also predicted that the morbidity rate in Myanmar should decrease in 2005.[15] The difference between this study and our study may be due to some differences in the climates, topography and
Figure 3: Scatter plot of the residuals of the ARIMA (1,0,1) model

Figure 4: Autocorrelation function of the residuals of ARIMA (1,0,1) model
health care levels between Myanmar and Thailand.

The ARIMA models are a useful tool for analysing non-stationary time-series data containing ordinary or seasonal trends.\cite{16,17} Forecasting arboviral diseases using routinely-collected data is still in its infancy, but as more data become available, forecasting offers the potential for improved contingency planning of public health interventions and more broadly-based forecasting. In this study, the model well reflected the trend in the incidence of DHF in southern Thailand. It is not yet possible to predict with precision the extent and magnitude of the alteration in the disease pattern. Time-series forecasting of DHF cases in southern Thailand may offer the potential for improving planning, control and prevention by public health intervention. The DHF cases estimated to occur calls for serious multi-sectoral preparedness to reduce the incidence in southern Thailand.

**Acknowledgements**

We thank Assistant Professor David Harding and two anonymous referees for their suggestions on the previous versions of this manuscript. This study was supported in part by the Thai Health Promotion Foundation Grant No. 47-00452, Complex System Key University Research Unit, and Graduate Research Fund, the Institute of Research and Development, Walailak University, Thailand. We thank the Office of Disease Prevention and Control 11 and 12 and the Bureau of Epidemiology, Ministry of Public Health, Thailand, for providing the data.
Forecasting DHF Cases in Southern Thailand using ARIMA Models

References


