DOI: 10.20472/IAC.2018.935.007

SOMSRI BANDITVILAI

King Mongkut's Institute of Technology Ladkrabang, Thailand

SIRILUCK ANANSATITZIN

King Mongkut's Institute of Technology Ladkrabang, Thailand

COMPARATIVE STUDY OF THREE TIME SERIES METHODS IN FORECASTING DENGUE HEMORRHAGIC FEVER INCIDENCE IN THAILAND

Abstract:

Accurate incidence forecasting of infectious disease such as dengue hemorrhagic fever is critical for early prevention and detection of outbreaks. This research presents a comparative study of three different forecasting methods based on the monthly incidence of dengue hemorrhagic fever. Holt and Winters method, Box-Jenkins method and Artificial Neural Networks were compared. The data were taken from the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health starting from January, 2003 to December, 2016. The data were divided into 2 sets. The first set from January, 2003 to December, 2015 were used for constructing and selection the forecasting models. The second set from January, 2016 to December, 2016 were used for computing the accuracy of the forecasting model. The forecasting models were chosen by considering the smallest root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to measure the accuracy of the model. The results showed that Artificial Neural Networks obtained the smallest RMSE in the modeling process and the MAPE in the forecasting process was 14.05%

Keywords:

Dengue hemorrhagic fever, Time Series Forecasting, Holt-Winters method, Box-Jenkins method, Artificial Neural Networks

JEL Classification: C22, C45

Introduction

In rainy season, there is a rapid growing epidemic of mosquitoes which carried both serious and not serious diseses. Dengue hemorrhagic fever is an international important tropical infectious diseses and the most important public health problems in Thailand (Supat Chamnanchanunt, 2012) Dengue hemorrhagic fever transmitted to humas via the mosquito vector, Aedes aegypti. The rate of illness in 2012 to 2013 was an average of 8.65 per 100,000 population in Thailand (Bureau of Epidemiology, 2013). Forecasting predicts future events based on available information. Forecasting plays a very important role in planning. The Department of Disease Control, Ministry of Public Health, Thailand has seen the importance of prognosis and filled it in the strategic plan and also plans to establish a center for epidemiology at the Bureau of Epidemiology. The accurate forecasting of dengue hemorrhagic fever outbreak could improve the efficiency of prevention and control the spread out the epidemic. Prediction of disease incidence can be both regression analysis and time series analysis. But time series analysis is more popular since no need to collect independent variables that are associated with incidence of disease. Time series forecasting is an effective non-explanatory mean to predict future epidemic behavior based on historical data (Xingyu Zhang, 2013).

Like other infectious diseases, dengue hemorrhagic fever incidence time series exhibit seasonal behavior, secular trend and rapid fluctuations. Therefore it is reasonable to forecast epidemic incidence with time series methods. There are two methods that are commonly used in epidemic time series forecasting. First, Holt-Winters method estimates three smoothing paramethers, associated with level, trend and seasonal factors. This method is widely used and proved to perform good prediction in time series data with linear trend and seasonal behavior. (Ferbar Tratar, L., 2010) work showed that Holt-Winters method gave the best forecasting result in case the time series data have strong seasonal factor. Second, the Box-Jenkins method is reputed to have high accuracy in short-term forecasting. This method has also demonstrated an effective linear model that can grasp the linear trend and seasonal effect of the time series. Many researchers included (Gharbi, M. et al., 2011) employed Box-jenkins method in forecasting infectious disease. Artificial neural networks is the method that simulated the human brain by the computer program. It can be trained and can bring knowledge and skills to solve problems such as aviation, automotive, management, banking, military, entertainment and public health. This method can effectively extract nonlinear trend and relationship in data. This is a popular method that produce a very accurate result and widely used in forecasting. (Sigauke C., 2014 and Xingyu Zhang, 2013) works confirmed that Artificial Neural Networks provided the best prediction result.

Data Collection and Methodology

The dengue hemorrhagic fever incidence monthly data were collected from the Bureau of Epidemiology, Department of Disease Control, Ministry of Public Health, Thailand starting from January, 2003 to December, 2016. The data were divided into 2 sets. The first set from January, 2003 to December, 2015 were used for constructing and selection the forecasting models. The second set from January, 2016 to December, 2016 were used for calculating the accuracy of the forecasting model. This research employed three forecasting methods which are Holt-Winters methods, Box-Jenkins method and Artificial Neural Networks.

Holt-Winters methods

The Holt-Winters methods of exponential smoothing involve linear trend and seasonality and are based on three smoothing equations: for level, for trend and for seasonality. The decision regarding which method to use depends on time series characteristics: the additive method is used when the seasonal component is constant, the multiplicative method is used when the size of the seasonal component is proportional to trend level (Chatfield, 1996).

Additive Holt-Winters method

If a time series has a linear trend with a fixed growth rate, β_1, and a fixed seasonal pattern, S_t , with constant variation, then the time series may be described by the model

$$Y_t = (\beta_0 + \beta_1 t) + S_t + \varepsilon_t \qquad (1)$$

For this model, the level of the time series at time t-1 is $T_{t-1} = \beta_0 + \beta_1(t-1)$ and at time t is $T_t = \beta_0 + \beta_1 t.$ Hence, the growth rate in the level from one time period to the next is β_1.

The estimate T_t for the level, the estimate b_t for the growth rate, and the estimate S_t for the seasonal factor, and e_t for the error of the time series in time period t, are given by the smoothing equations (Bowerman, O. Connell and Koehler, 2005).

$$T_t = T_{t-1} + b_{t-1} + \alpha e_t \qquad (2)$$

$$b_t = b_{t-1} + \alpha\gamma e_t \qquad (3)$$

$$S_t = S_{t-L} + (1-\alpha)\delta e_t \qquad (4)$$

where α, γ and δ are smoothing constants between 0 and 1, T_{t-1} and b_{t-1} are estimates in time period t-1 for the level and growth rate, S_{t-L} is the estimate in time period t-L for the seasonal factor, and L denotes the number of seasons in a year.

Multiplicative Holt-Winters method

If a time series has a linear trend with a fixed growth rate, β_1, and a fixed seasonal pattern, S_t , with increasing variation, then the time series may be described by the multiplicative model

$$Y_t = (\beta_0 + \beta_1 t) \times S_t \times \varepsilon_t \qquad (5)$$

The smoothing equations in the multiplicative Holt-Winters method are:

$$T_t = T_{t-1} + b_{t-1} + \frac{\alpha e_t}{S_{t-L}} \qquad (6)$$

$$b_t = b_{t-1} + \frac{\alpha\gamma e_t}{S_{t-L}} \qquad (7)$$

$$S_t = S_{t-L} + \frac{(1-\alpha)\delta e_t}{T_t} \qquad (8)$$

This research estimated the smoothing parameters α, γ and δ to minimize the root mean square error (RMSE) by using Solver module in Microsoft Excel.

Box-Jenkins method

Box-Jenkins method developed in 1974 and is widely used in forecasting the incidence of the epidemic time series. Box-Jenkins method use Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) to identify the model under the stationary condition.

Box-Jenkins method is a four-step process (Bowerman, O. Connell and Koehler, 2005):

Step 1: Tentative identification: historical data are used to tentatively identify an appropriate Box-Jenkins model.

Step 2: Estimation: historical data are used to estimate the parameters of the tentatively identified model.

Step 3: Diagnostic checking: various diagnostics are used to check the adequacy of the tentatively identified model. In some cases may need to suggest an improved model, which is then regarded as a new tentatively identified model.

Step 4: Forecasting: once a final model is obtained, it is used to forecast future time series values.

The Box-Jenkins model of $ARIMA(p,d,q)\times SARIMA(P,D,Q)_L$ are defined as follows:

$$\phi_p(B)\phi_P(B^L)Z_t = \theta_0 + \theta_q(B)\theta_Q(B^L)\varepsilon_t \qquad (9)$$

when

$$\phi_p(B) = (1-\phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \cdots - \phi_p B^p)$$

$$\phi_P(B^L) = (1-\phi_{1L} B^L - \phi_{2L} B^{2L} - \phi_{3L} B^{3L} - \cdots - \phi_{PL} B^{PL})$$

$$\theta_q(B) = (1-\theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \cdots - \theta_q B^q)$$

$$\theta_Q(B^L) = (1-\theta_{1L} B^L - \theta_{2L} B^{2L} - \theta_{3L} B^{3L} - \cdots - \theta_{QL} B^{QL})$$

$$Z_t = (1-B^L)^D (1-B)^d Y_t$$

where B is backward shift operator

θ_0 is a constant

ϕ_p is the non-seasonal autoregressive model of order p

θ_q is the non-seasonal moving average model of order q

$\phi_P(B^L)$ is the seasonal autoregressive model of order P

$\theta_Q(B^L)$ is the seasonal moving average model of order Q

ε_t is the error at time t and have normal distribution which mean is equal to zero and constant variance and statistical independent

d is the number of regular difference

D is the number of seasonal difference

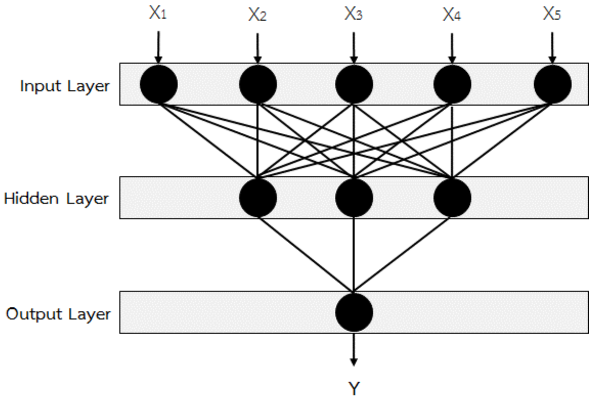
This research employed Minitab 16.0 in analyzing the Box-Jenkins model.

Artificial Neural Networks

Artificial neural networks were designed to mimic the characteristics of the biological neurons in the human brain and nervous system (Jiang L-H. et al., 2011). In the case of modeling the epidemic time series, the historical incidence are sent into the input neurons, and corresponding forecasting incidence is generated from the output neurons after the network is adequately trained. The network learns the information contained in the incidence time series by adjusting the interconnections between layers. The structure and neural networks can only be viewed in terms of the input, output and transfer characteristics. The specific interconnections cannot be seen even after the training process. There is no easy way to interpret the specific meaning of the parameters and interconnections within networks trained using the real epidemic time series data. There are two advantages of employing neural networks for forecasting time series data. First, they can fully extract the complex nonlinear relationships hidden in the time series. Second, they have no assumption of the underlining distribution for the collected data (Zhang G. et al., 1998).

Back Propagation Neural Networks are a type of feed forward artificial neural networks. In feed-forward neural networks, the data flow is in one direction and the answer is obtained solely based on the current set of inputs. Back Propagation Neural Networks consist of an input layer, a hidden layer, and an output layer. Each layer is formed by a number of nodes, and each node represents a neuron. The upper-layer and lower-layer nodes are connected by the weights. The common structure of a Back Propagation Neural Networks model is illustrated in Fig. 1

Fig 1: Schematic of Back Propagation Neural Networks



Back Propagation Neural Networks training includes three steps: (1) the forward feeding of the input training pattern, (2) the calculation and back-propagation of the associated error, and (3) the adjustment of the weights. With n input neurons, m hidden neurons, and one output neuron, the outputs of all hidden layer nodes are calculated as follows:

$$net_j = \sum_{i=0}^{n} \omega_{ij} x_i (i = 0,1,2,...,n; j = 1,2,...,m) \qquad (10)$$

$$y_j = f(net_j)(j = 1,2,...,m) \qquad (11)$$

where net_j is the activation value of the jth node, ω_{ij} the connection weight from input node i to hidden node j, x_i the ith input, y_j the corresponding output of the jth node in the hidden layer, and f the activation function of a node, which is usually a sigmoid function.

$$f(x) = \frac{1}{1+\exp(-x)} \qquad (12)$$

The outputs of all output layer neurons are expressed as

$$O = f_0(\sum_{j=0}^{m} \omega_j, y_j)(j = 0,1,2,..,m) \qquad (13)$$

Where f_0 is the activation function, which is usually a line function; ω_j is the connection weight from the hidden node j to the output node, and y_j is the corresponding output of the jth node in the hidden layer. All the connection weights are initialized randomly, and then modified according to the results of the Back Propagation Neural Networks training process. Several methods have been proposed for the adjustment of the connection weights, such as the steepest descent algorithm, Newton's method, Gauss-Newton's algorithm and Levenberg-Marquardt algorithm (Wilamowski BM. et al., 2001).

The available incidence of the time series was divided into three subsets. Dengue hemorrhagic fever incidence from January, 2003 to December, 2012 was employed as the training set used for training the network. The incidence from January, 2013 to December, 2015 was employed as the validation set. The remaining set of the series was used as the test set.

This research employed Weka 3.8.1 in modelling Back Propagation Neural Networks. The number of inputs of the neural networks was determined by the seasonal period of the time series. In this study, the period of the incidence of dengue hemorrhagic fever observed was twelve. Therefore, twelve and twenty four were selected as the number of input layer for the networks. The output layer of artificial neural networks contains only one neuron representing the forecast value of the incidence of the next month. This research employed Weka 3.8.1 in running artificial neural networks. The different learning rate were examined from 0.005 to 0.5 with 0.005 increments. The different momentum were examined from 0.1 to 0.8 with 0.05 increments. The number of hidden neurons were varied from 2 to 20 at an increment of 1.

Model Selection Criterion

The forecasting models were chosen by considering the smallest root mean square error (RMSE) and mean absolute percentage error (MAPE) were used to measure the accuracy of the model.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_t^2}$$

$$MAPE = \sum_{i=1}^{n} \left| \frac{e_t}{Y_t} \right| \times 100$$

Table 1: The estimating parameters α, γ and δ and RMSE of Holt and Winters model.

Model	α	γ	δ	**RMSE**
Additive	0.984442	0.988167	0.5	2868.875
Multiplicative	1	0	0	**1898.064**

Results

The first set of data from January, 2003 to December, 2015 was employed to build the Holt-Winters model and Box-Jenkins model. The forecast error of the Holt-Winters model shows random distribution. Table 1 presents the estimating parameters α,γ and δ and RMSE of Holt and Winters model. Fig 2. presents the time series of dengue hemorrhagic fever incidence from January, 2003 to December, 2015. Fig 3-4 present ACF and PACF of the time series of dengue hemorrhagic fever incidence with one regular difference and one seasonal difference of order 24. Table 2 presents Minitab output of the model $ARIMA(1,1,0)\times SARIMA(2,1,1)_{24}$. Table 2 showed that all parameters were statistically significant and errors are independent. The optimal model of artificial neural networks is 24-4-1 with learning rate of 0.005, momentum of 0.25 and 5000 iterations. Table 3 presents the RMSE of training set, validation set and testing set of the artificial neural networks model. Table 4 presents the RMSE of three forecasting models.

Fig 2: The time series of dengue hemorrhagic fever incidence from Jan, 2003 to Dec, 2015.

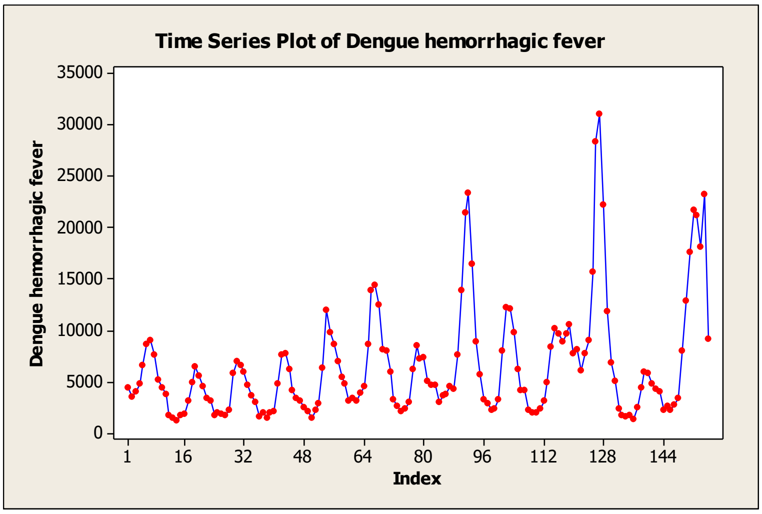


Fig 3: ACF of the time series of dengue hemorrhagic fever incidence with one regular difference and one seasonal difference of order 24.

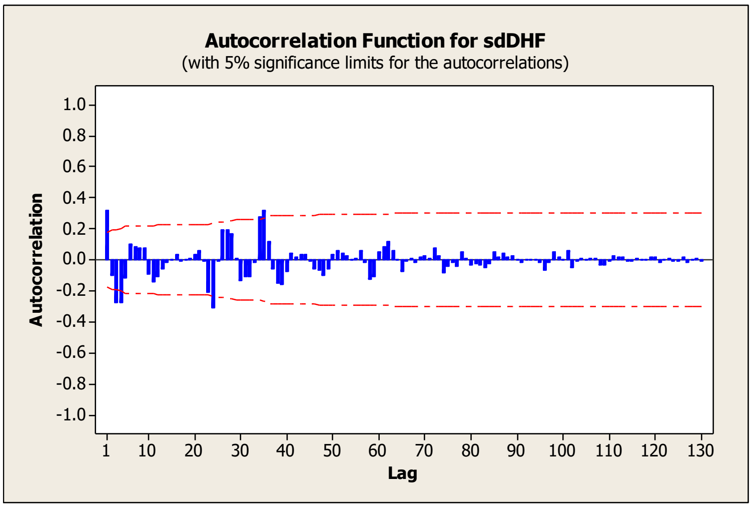


Fig 4: PACF of the time series of dengue hemorrhagic fever incidence with one regular difference and one seasonal difference of order 24.

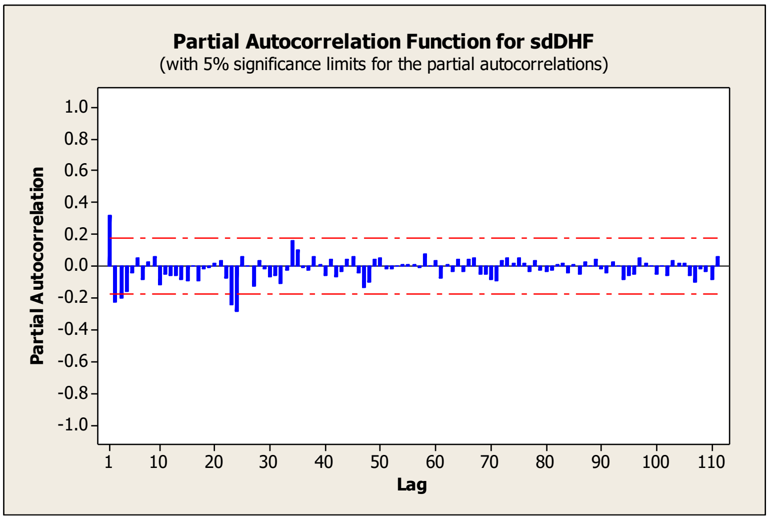


Table 2: Minitab output of the model $ARIMA(1,1,0)\times SARIMA(2,1,1)_{24}$.

```
Final Estimates of Parameters

Type          Coef   SE Coef       T       P
AR    1     0.4668    0.1064    4.39   0.000
AR    2    -0.2520    0.1072   -2.35   0.020
SMA   24    0.7613    0.1230    6.19   0.000


Differencing: 1 regular, 1 seasonal of order 24
Number of observations:  Original series 156, after differencing 131
Residuals:     SS =  659877455 (backforecasts excluded)
               MS =  5155293  DF = 128


Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square    9.2   15.2   44.8   50.6
DF              9     21     33     45
P-Value     0.421  0.811  0.082  0.261
```

Table 3: The RMSE of training set, validation set and testing set of the artificial neural networks model.

Model	Learning rate	Momentum	Iterations	RMSE		
				Training set	Validation set	Test set
24-4-1	0.005	0.25	5000	1545.9454	1012.784	712.004

Table 4: The RMSE of three forecasting models.

Forecasting model	RMSE
Multiplicative Holt-Winters model	1898.064
Box-Jenkins	2270.527
Artificial Neural Networks	**1545.9454**

Conclusion

This research presents three different forecasting methods which are Holt and Winters method, Box-Jenkins method and Artificial Neural Networks to model the incidence of dengue hemorrhagic fever in Thailand. The results found that Artificial Neural Networks give the smallest RMSE in the modeling process and the MAPE in the forecasting process was 14.05%. The performance of the three models ranked in descending order were Artificial Neural Networks, Holt Winters method and Box-Jenkins method. Artificial Neural Networks are nonparametric nonlinear models in general are tolerant to the data and less susceptible to model misspecification problems than Box-Jenkins method, and Holt and Winters method. The disadvantage of the artificial neural networks is their black-box nature, in which the specific nonlinear functions within the time series data may not be explained well in practice.

Acknowledgement

The authors would like to thank the Faculty of Science, King Mongkut’s Institute of Technology Ladkrabang for funding and supporting this research project.

Reference

BOWERMAN, O’ CONNELL AND KOEHLER. (2005) *Forecasting, Time Series, and Regression: an Applied Approach.* 4th ed. Thomson. USA.

BUREAU OF EPIDEMIOLOGY (2013). *Situation of Duegue Hemorrhagic fever in Thailand*. Department of Disease Control, Ministry of Public Health.

CHATFIELD, C. (1996). *The Analysis of Time Series*, 5th ed., Chapman & Hall, New York.

FERBAR TRATAR L. (2013) Improved Holt-Winters Method: a Case of Overnight Stays of Tourists in Republic of Slovenia. *Economic and Business Review*. 2013 Vol. 16 No.1: 5-17.

GHARBI, M., QUENEL, P., GUSTAVE. J.CASSADOU, S. RUCHE, G.L., GIRDARY, L. AND MARRAMA, L. (2011) Time Series Analysis of Dengue Incidence in Gualeloupe. French West Indies: Forecasting Models Using Climate Variables as Predictors. *BMC Infectous Diseases.* 2011, Vol.11, No. 166, 1-13.

JIANG L-H, WANG A-G, TIAN N-Y, ZHANG W-C, FAN Q-L (2011) BP Neural Network of Continuous Casting Technological Parameters and Secondary Dendrite Arm Spacing of Spring Steel. *Journal of Iron and Steel Research*. 2011, Vol 18, 25-29.

SIGAUKE C., DARIKWA T.B. AND MASEMOLA M.I. (2014) Prediction of South Africa’s Tourism Hotel Accommodation Monthly Income: Challenges in an Environment Characterized by a World Recession and a World Cup. *Mediterranean Journal of Social Science.* 2014, Vol. 5 No.20: 460-465.

SUPAT CHAMNANCHANUNT. (2012) Bleeding Disorder in Dengue Patients. *Trop Med Parasitol.* 2012, Vol 35, No. 1, 27-36.

ILAMOWSKI BM., IPLIKCI S., KAYNAK O., ELE MO. (2001) An Algorithm for Fast Convergence in Training Neural Networks. *IEEE* 3: 1778-1782

XINGYU ZHANG, YUANYUAN LIU, MIN YANG, TAO ZHANG, ALISTAIR A. YOUNG AND XIAOSONG LI. (2013) Comparative Study of Four Time Series Methods in Forecasting Typhoid Fever Incidence in China. *PLOS ONE*, 2013, Vol 8, 5:1-11.

ZHANG G., EDDY PATUWO B., Y. HU M. (1998) The State of the Art. *International Journal of Forecasting.* 1998, Vol 14, 35-62