A comparative study of forecasting models for the number of Malaria patients in Phanom District, Surat Thani Province

Sujaree Damsri$^{1,2}$ and Ketsuda Maneewong$^1$

$^1$Department of Mathematics, Faculty of Science and Technology, Surat Thani Rajabhat University, Surat Thani 84100, Thailand

Abstract

The purpose of this study was to compare 5 forecasting models: Ratio-to-Trend, Holt-Winters Exponential Smoothing, Regression Dummy Variables, Theta, and Combined Forecasting. These methods were used for forecasting the number of Malaria patients in Phanom District, Surat Thani Province. The monthly time series data were collected from Surat Thani Provincial Health Office during January 2011 through December 2015 and divided into two groups. The first group of data covered the period of time from January 2011 to December 2014 and has been used for constructing the forecasting models. The secondary group of data covered the period of time from January 2015 to December 2015 and has been used for checking the accuracy of the forecasting models. The accuracy characteristics that we used were Mean Absolute Deviation (MAD) and Mean Square Error (MSE). The smallest values of MAD and MSE indicate the optimal forecasting model.

Based on the study, the optimal model was found to be Combined Forecasting and the equation of the forecasting model was $\hat{Y}_t = 0.907 F_0 + 0.093 F_1; t = 1, 2, \ldots, 12$ where each $F_0$ and $F_1$ represents the single forecast at time $t$, from the Regression Dummy Variables model and Holt-Winters Exponential Smoothing model, respectively.

**Keywords:** forecasting model, Malaria, Surat Thani Province, combined forecasting model

**Article history:** Received 11 January 2017, Accepted 11 May 2017

1. Introduction

Forecasting is the process of making predictions of the future based on past and present data. The categories of forecasting are qualitative and quantitative. Qualitative forecasting is the approach based on judgments and opinions [1]. Quantitative forecasting can be applied when two conditions are satisfied: numerical information about the past was available and some aspects of the past patterns will continue into the future [2]. Many years ago, forecasting techniques have been continuously developed due to their usefulness in planning and decision making. Both the short-term and long-term forecasting techniques were a tool to obtain future information for planning. It is widely accepted that forecasting plays a key role both in the public and private sectors [3]. At the present time, forecasting techniques are used to predict the number of patients in epidemic diseases [4, 5, 6, 7, 8]. The suitable forecasting model depends on many factors such as the amount of historical data available, the experience of the forecaster, the degree of accuracy desirable, the time period of forecasting etc.

Malaria is caused by five species of parasites belonging to the genus *Plasmodium*: *P. falciparum*, *P. vivax*, *P. malariae*, *P. knowlesi* and *P. ovale*. Of these, *P. falciparum* and *P. vivax* are the most important [9]. Major *Plasmodium* that are spread from one person to another by the bite of female *Anopheles* mosquitoes in Thailand were *P. falciparum* and *P. vivax* [9, 10, 11]. From 1991 to 2002, the incidence of malaria in 14 southern provinces of Thailand has been present as important health problem especially in Ranong, Krabi, Surat Thani and Yala Province [10]. In 2015, Surat Thani Province is the region with the ninth Malaria patients of country especially in Phanom district has the highest number of malaria cases in Surat Thani province [11, 12].

In this paper, we constructed 5 forecasting models to determine the optimal models for the number of Malaria patients in Phanom District, Surat Thani Province. Forecasting values that we obtain will be used as the information in the preliminary plan for disease prevention and control.

2. Materials and methods

2.1 Data collection

The data under investigation were the secondary time series data from Phanom District. These time series data contain the number of Malaria patients and were collected by Surat Thani Provincial Health Office from January 2011 to December 2015. The data were monthly time series which amounts to 60 values that we divided into two groups. The data collected from January 2011 to December 2014...
amounts to 48 values, are considered to be the first group. The data collected from January 2015 to December 2015 amounts to 12 values, are considered to be the second group.

2.2 Exploring data pattern

The identification and understanding of historical patterns in the data is a major factor influencing the selection of a forecasting model [13, 14]. There are four general patterns of time series data: horizontal, trend, seasonal and cyclical [1, 13]. We investigated the data pattern by graphing the time series data and the autocorrelation for various lags of a time series which is called correlogram. When time series data fluctuate around a constant level or mean, a horizontal pattern exists. When time series data grow or decline over an extended period of time, a trend pattern exists. When time series data exhibit rises and/or falls that were not over an extended period, a cyclical pattern exists. When time series data display the pattern of change that repeats itself year after year, a seasonal pattern exists [13, 15, 16].

2.3 Constructing the forecasting model

We used the data from January 2011 to December 2014 for constructing 5 forecasting model. Considering the model to constructing the forecasting model is as follows: appropriate for the pattern of data and used to predict the number of patients in epidemic diseases [4, 17, 18]. We selected 5 forecasting models: Ratio-to-Trend, Holt-Winters Exponential Smoothing, Regression Dummy Variables, Theta, and Combined Forecasting by using Mintab software. The equations of forecasting models are as follows:

1. Ratio-to-Trend [4, 15]
   The forecasting equation used in Multiplicative Ratio-to-Trend is as follows:
   \[
   \hat{Y}_i = h_i b_i \hat{S}_i; i = 1, 2, ..., 12
   \]  

   This method has three parameters: \( \alpha, \beta, \gamma \) where the values of these parameters are between 0 to 1. The four equations used in Multiplicative Holt-Winters’ Exponential Smoothing are as follows:
   - Exponentially smoothed series:
     \[
     L_t = \alpha \frac{Y_t}{S_{t-s}} + (1 - \alpha)(L_{t-1} + T_{t-1})
     \]  
   - The trend estimate:
     \[
     T_t = \beta(L_t - L_{t-1}) + (1 - \beta)T_{t-1}
     \]  
   - The seasonality estimate:
     \[
     S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-s}
     \]  
   - The forecasting model:
     \[
     \hat{Y}_{t+p} = (L_t + pT_t)S_{t+s+p}
     \]

Where:
- \( L_t \) means the new smoothed value or the current level estimate.
- \( \alpha \) means the smoothing constant for the level,
- \( Y_t \) means the actual value in period \( t \),
- \( \beta \) means the smoothing constant for trend estimate,
- \( T_t \) means the trend estimate,
- \( \gamma \) means the smoothing constant for seasonality estimate,
- \( S_t \) means the seasonality estimate,
- \( p \) means the period to be forecasted into the future,
- \( s \) means the length of seasonality,
- \( \hat{Y}_{t+p} \) means the forecast for \( p \) periods into the future.

3. Regression Dummy Variables Models

Regression Dummy Variables Models are regression models for time series with seasonal pattern that involve dummy variables. A seasonal model for monthly data with the time trend in multiplicative model is given below [13, 15]:

\[
\ln \hat{Y}_t = \ln a_0 + \ln a_t + \ln a_i L_t + \ln b_1 S_t
\]

\[
+ \ln b_2 S_1 + \cdots + \ln b_{12} S_{12}
\]

(6) where:
- \( \hat{Y}_t \) means the forecast value for time period \( t \),
- \( a_t, a_i, b_1, b_2, b_3, b_4, b_5 \) means the coefficients to be estimated,
- \( S_t \) means the dummy variable which is 1 for month \( i \); 0 otherwise.

4. Theta

The Theta model is based on the concept of modifying the local curvatures of the time series. This change is obtained by a coefficient, called Theta-coefficient (\( \theta \)). When the value of Theta-coefficient equals zero, the time series is equivalent to a linear regression line. When the Theta-coefficient takes a positive value, then the time series is dilated [19]. In this paper, we used Theta model for Theta-coefficient equals 0 or 2. The steps of Theta model are given below [20, 21]:

Step 1: Seasonality testing by \( t \)-test for autocorrelation function.

Step 2: Deseasonalisation by the classical decomposition method.

Step 3: Decomposition of the time series into two theta lines, \( \theta = 0 \) and \( \theta = 2 \). The equations are as follows: [9, 10].

The theta line for \( \theta = 0 \):

\[
\hat{Y}_{nh}(0) = \hat{a}_0 + \hat{b}_n(n + h - 1)
\]

(7) The theta line for \( \theta = 2 \):

\[
\hat{Y}_{nh}(2) = \alpha \sum_{i=0}^{12} (1 - \alpha)^{i} Y_{s+i+2} + (1 - \alpha) Y_{i}
\]

(8)

Step 4: Combination two theta lines with equal weight.

The forecasting model:

\[
\hat{X}_{nh} = \frac{\hat{Y}_{nh}(0) + \hat{Y}_{nh}(2)}{2}
\]
choose the optimal model for forecasting. The forecasting model which had the smallest values of MAD and MSE gives the optimal forecasting model. The equations of forecasting accuracy are given as follows: [13]

\[
MAD = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i| \\
MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2
\]

where:
\( Y_i \) means the actual value in time period \( t \),
\( \hat{Y}_i \) means the forecasted value for time period \( t \).

3. Results and discussion

3.1 Data pattern

From Figure 1, the patterns of the number of Malaria patients showed that there exist both trend and seasonal variations because the series grew over an extended period of time and a repeated calendar pattern can be observed.

Figure 2 shows the same result of data pattern as Figure 1 because of autocorrelation coefficients \( r_k \) were typically large for the first several time lags and
then gradually drop toward zero as the number of periods increased. This means that the number of Malaria patients seems to indicate a trend and seasonal variation.

### 3.2 Constructing the forecasting models

In this section, we present the equations of five forecasting models which are given in Table 1.

<table>
<thead>
<tr>
<th>Forecasting models</th>
<th>Equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio-to-Trend</td>
<td>$\hat{Y}_t = 1.33(1.02)^t \hat{S}_i ; i = 1,2,\ldots,12$</td>
</tr>
<tr>
<td>Holt-Winters Exponential Smoothing</td>
<td>$\hat{Y}(p) = \left[17.52 + p(-1.08)\right]\hat{S}_i ; i = 1,2,\ldots,12, p = 1,2,\ldots$ when $\alpha, \beta, \gamma = 0.1$</td>
</tr>
<tr>
<td></td>
<td>$\hat{S}_i = 0.881, \hat{S}_1 = 0.485, \hat{S}_2 = 0.766, \hat{S}_3 = 1.412, \hat{S}_4 = 1.816, \hat{S}_5 = 1.924, \hat{S}<em>6 = 1.306, \hat{S}<em>7 = 0.966, \hat{S}<em>8 = 0.747, \hat{S}</em>{10} = 0.575, \hat{S}</em>{11} = 0.698, \hat{S}</em>{12} = 0.42$</td>
</tr>
<tr>
<td>Regression Dummy Variables</td>
<td>$\ln \hat{Y}<em>t = 0.775 + 0.176\alpha - 0.00329\beta^2 + 0.284\gamma_i + 0.054S_i + 0.450S_i + 1.01S_i + 1.05S_i + 0.696S_i + 0.278S_i - 0.061S_i - 0.303S</em>{10} - 0.014S_{11}$</td>
</tr>
<tr>
<td>Theta</td>
<td>$\hat{X}_{n/h} = \left[\alpha X_n + (1-\alpha)\hat{X}_n + 7.33165\times10^{-2}(h-1+\frac{1}{\alpha})\right] \times \hat{S}_i$ when $h = 1,2,3,\ldots$</td>
</tr>
<tr>
<td></td>
<td>and $\alpha = 0.3 \hat{S}_i = 0.834, \hat{S}_1 = 0.423, \hat{S}_2 = 0.667, \hat{S}_3 = 1.67, \hat{S}_4 = 1.857, \hat{S}_5 = 1.873, \hat{S}<em>6 = 1.374, \hat{S}<em>7 = 1.07, \hat{S}<em>8 = 0.56, \hat{S}</em>{10} = 0.725, \hat{S}</em>{11} = 0.676, \hat{S}</em>{12} = 0.272$</td>
</tr>
<tr>
<td>Combined Forecasting</td>
<td>$\hat{Y}<em>t = 0.907F_u + 0.093F</em>{2u} ; i = 1,2,\ldots,12$</td>
</tr>
</tbody>
</table>

### Table 2 Forecasting accuracy

<table>
<thead>
<tr>
<th>Forecasting model</th>
<th>Forecasting accuracy</th>
<th>MAD</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio-to-Trend</td>
<td></td>
<td>8.49</td>
<td>106.56</td>
</tr>
<tr>
<td>Holt-Winters Exponential Smoothing</td>
<td></td>
<td>6.54</td>
<td>59.18</td>
</tr>
<tr>
<td>Regression Dummy Variables</td>
<td></td>
<td>1.41</td>
<td>4.21</td>
</tr>
<tr>
<td>Theta</td>
<td></td>
<td>13.30</td>
<td>237.80</td>
</tr>
<tr>
<td>Combined Forecasting</td>
<td></td>
<td>1.24</td>
<td>3.37</td>
</tr>
</tbody>
</table>

![Figure 3](image) Comparison between actual values and forecasted values
3.3 The comparative analysis for the optimal for forecasting model

The values of forecasting accuracy are shown in Table 2 and the comparison between actual values and forecasted values are shown in Figure 3.

From Table 2, Combined Forecasting showed the lowest values of forecasting accuracy that means Combined Forecasting was the optimal model which corresponded with Manmin [3] who said that Combined Forecasting is the most accurate model for forecasting and corresponded with Taesombut [15] who said that Combined Forecasting is a method to make good prediction.

Forecasting by times series techniques is only study the difference of the data which depend on times. In fact, there are many factors that are important for the spread of disease which complicate and differ in each area. In case of some factors has changed, the number of patients may be mistaken from the forecasting. This means that the forecasting equation should usually examined by statistical analysis if forecasting values are not reasonable, we can construct new forecasting model or adjust the original forecasting model by collecting more data.

4. Conclusions

In the study of the number of Malaria patients in Phanom District, Surat Thani Province, we compare 5 models: Ratio-to-Trend, Holt-Winters Exponential Smoothing, Regression Dummy Variables, Theta, and Combined Forecasting. Two accuracy characteristics are used for forecasting: MAD and MSE. The results of the study show that Combined Forecasting was the best method to forecasting Malaria patients. The forecasted values from this model would be a useful guidance for timely prevention and control measures to define the Malaria control operation area which consists of 2 phases: transmission area (A) and non-transmission area (B) [22].

Acknowledgements

We would like to thank Surat Thani Provincial Health Office for providing the necessary information. We also would like to thank the Department of Mathematics, Faculty of Science and Technology, Surathani Rajabhat University and Department of Mathematics, Faculty of Science, Kasetsart University for providing the facilities to carry out the research.

References


