Forecasting model for the number of long stay Japanese tourist arrivals in Chiang Mai

PRADTHANA MINSAN ; KUNANON JOMTOUR , WATHA MINSAN

1, 2 Faculty of Science and Technology, Chiang Mai Rajabhat University, Chiang Mai, Thiland
3 Faculty of Science, Chiang Mai University, Chiang Mai, Thailand

Abstract

The aging phenomenon of the elderly occurs worldwide, especially Japan that is the most of the highest average age. Therefore, long stay tourism is alternative tourism for the elderly Japanese tourists. The aim of this research was to construct the appropriate forecasting model for the number of long stay Japanese tourist arrivals in Chiang Mai, Thailand. The data in this study gathered from the Chiang Mai Immigration Office that recorded in monthly during from January 2014 to July 2017 a total of 43 months. Then the data were classified into two sets. The first data set from January 2014 to December 2016 for 36 months were used to build the forecasting model by the methods of Classical decomposition, Seasonal simple exponential smoothing, Box-Jenkins and Combining. The second data set from January 2017 to July 2017 a total of 7 months were used to compare the earlier three methods of the forecasting accuracy model via the criteria of Root Mean Square Error: RMSE. Research results indicated that combining forecasts was the most suitable for forecasting the number of long stay Japanese tourist arrivals in Chiang Mai.

Keywords: Long Stay Tourism, Classical Decomposition, Seasonal Simple Exponential Smoothing, Box-Jenkins, Combining Forecasts, Root Mean Square Error (RMSE)

Received: 20 June 2018 / Accepted: 18 July 2018 / Published: 27 August 2018

INTRODUCTION

According to the Annual Report on the Aging Society in Japan: 2016, the elderly rises to 26.7%, moreover, by 2060, 1 in 2.5 people will be 65 years old and over, and 1 in 4 will be 75 years old and over (Cabinet Office Government of Japan, 2016; Fahmida, Kulsuma, & Reza, 2016). Consequently, Japanese elders move to the other countries after retirement has a new possibility of solving the social problems brought by ageing, this is call International Retirement Migration (IRM) or Long stay (Chan, 2018; Ramanauskaite & Vaisnys, 2017; Yoshida, 2015). Meanwhile, Thailand is one of the top-ten favorite countries for Japanese long-stay tourism, which following by a survey from the Long Stay Foundation in Japan (Khaokhrueamuang, 2014; Komaladewi, Mulyana, & Jatnika, 2017; Sinh, Nga, Linh, & Tuan, 2016). In 2011, Chiang Mai had 2,854 long stay Japanese tourists, increase 43% from 2007, accounting for 80% of the Japanese population living in the northern part of the country (3,573). Additionally, Chiang Mai is ranked third after Bangkok (35,935) and Chonburi (3,695) (Kongprasert, 2013). Although there are a lot of studies of the factors influencing the Japanese tourists decision for long stay in Chiang Mai, or the study the decision’s factors of duration extension of Japanese long stay residents in Chiang Mai. Nevertheless, there is no study of forecasting model for the number of long stay Japanese tourist arrivals in Chiang Mai. Hence, this is the reason for this study. The aimed of this research was to predict the number of long stay Japanese tourist arrivals to Chiang Mai, Thailand in the period January to July 2017. Additionally, construct the appropriate forecasting model for the number of long stay Japanese tourist arrivals in Chiang Mai, Thailand.
LITERATURE REVIEW

Long stay tourism differs to mass tourism since all long stay all long stay tourists have to come back home finally (Hongsranagon, 2005; Sawatsuk, Darmawijaya, Ratchusanti, & Phaokrueng, 2018; Silva & Madushani, 2017). Thus, management to satisfactory services to tourists, destinations need to acquire reliable forecasts of future demand for accommodation, transportation, service staff and other related travel services (Wang & Lim, 2005). Additionally, Louw and Saayman (2013) stated that a lack of knowledge of future tourist arrivals may lead to missed opportunities or an overestimation of tourism demand. Therefore, forecasting long stay tourist arrivals are considerable for tourism long stay planning at all levels from government to a single tourist business and medical care service. Although time series model has various techniques suitable for forecasting, there are rarely studies applied forecasting model for predict to the number of long stay arrivals. Prior studies usually have used procedures in time series forecasting are the exponential smoothing or the ARIMA (Box-Jenkins) methodology. For instance, Gajić, Vujko, and Papić Blagojević (2015) estimated and validated an ARIMA model for forecasting long-stay visitors in Novi Sad, Serbia. Lorde and Moore (2008) used the time series modeling techniques, AR(1) and AR(2) to model and forecast the volatility in monthly international tourist arrivals to Barbados. Furthermore, most studies have applied time series modeling techniques. Song and Li (2008) reported that in the post 2000 empirical studies concentrated on the identification of the relationships between tourism demand and its influencing factors include evaluated the forecasting performance of the econometric models in addition to the identification of the casual relationship. Goh and Law (2011) studied 155 research papers published between 1995 and 2009, appears that the more advanced methods such as cointegration error correction model, time varying parameter model, and their combinations with systems of equations produce better results in terms of forecasting accuracy. Saayman and Saayman (2010) aimed to model and forecast tourism to South Africa from the countrys main intercontinental tourism markets by naive, exponential smoothing and ARIMA models. Lin, Chen, and Lee (2011) tried to build the forecasting model of visitors to Taiwan using three commonly adopted ARIMA, Artificial Neural Networks (ANNs), and Multivariate Adaptive Regression Splines (MARS) methods. Singh (2013) generate one-period-ahead forecasts of international tourism demand for Bhutan by selecting appropriate model both ARIMA and exponential smoothing. Çuhadar (2014) determine the forecasting model that provides the best performance when compared the ex post forecast accuracy of different exponential smoothing and Box-Jenkins models which were to forecast the monthly inbound tourism demand to Istanbul by the model giving best results.

On the other hand, many studies used combined forecasts to improve forecasting accuracy. For examples, Chu (1998) employ a combined seasonal nonseasonal ARIMA and sine wave nonlinear regression forecast model to predict international tourism arrivals. Before, Song, Witt, Wong, and Wu (2009) examined combination forecasting techniques for forecast tourism.

It can be noticed that there has been no published study in academic journals concerning the forecasting long stay Japanese tourist arrivals in Chiang Mai, Thailand. All this has led to point out on modeling and forecasting long stay Japanese that is main objective this study.

DATA AND METHODOLOGY

Trend Identification

A trend exists when there is a long-term increase or decrease in the data, It does not have to be linear. Sometimes the trend refers as changing direction, when it might go from as an increasing trend to a decreasing trend (Hyndman & Athanasopoulos, 2018). Therefore, Trend identification before deciding whether a stationary or non-stationary should be used to test initially. Non-parametric significance tests were used to identify statistically significant trend since they are not affected when the distribution of data is not normal, insensitive to outliers and are not affected by missing or censored data (Loftis, McBride, & Ellis, 1991). A non-parametric Daniels trend test is based on Spearman’s rank correlation coefficient, $r_s$. The null hypothesis $H_0$: There is no trend vs $H_1$: There is a trend. Firstly, the observations is ranked in ascending order (if two or more observations have the same value, their rank are average ranking. Subsequently, the rank correlation coefficient, is calculated as:

$$r_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$ (1)
where \( d_t \) is the difference between corresponding ranks of each observation, and \( n \) is number of observations \( Y_t \). However, another method for testing \( r_s \) is a Z test (Ramsey, 1989) based on the statistic:

\[
Z = r_s \sqrt{N - 1}
\]  

(2)

The Z test has been advocated provide (Berenson & Levine, 1988; Marascuilo & Serlin, 1988).

Analysis of Seasonality
Seasonal effects depict the behavior patterns of data that occur in the periods of time of less than a year. The presence of seasonal component can be confirmed by inspecting the plot of the data and by the prior knowledge of the behavior of the time series (Kurukulasooriya & Lelwala, 2014). However, in this study, a non-parametric Kruskal-Wallis test is applied for assuring more than a visual inspection of the time series plot. The test statistic takes the form of.

\[
H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)
\]  

(3)

Under the null hypothesis of the same median, statistics \( H \) is a chi-square distribution with \( k-1 \) degrees of freedom; in this test, \( n_j \) is the number of observations in collection \( j \), while \( n \) is the number of total observations in all the collections, and \( R_j \) is sum of the \( n_j \) observations of the collection \( j \) (Khalil et al., 2017).

Methodologies Used

Classical decomposition
The classical decomposition method originated in the 1920s. It is a relatively simple procedure and forms the starting point for most other methods of time series decomposition. There are two forms of classical decomposition: additive decomposition and a multiplicative decomposition (Hyndman & Athanasopoulos, 2018). In the classical decomposition the variation is commonly decomposed in a trend: \( T_t \), a seasonal effect or seasonal indicators: \( S_t \), and a random error: \( \varepsilon_t \). However, this study the multiplicative decomposition used to construct the forecasting model since the size of the seasonal effect appears to increase with the mean (Khalil et al., 2017), that would be written as

\[
Y_t = T_t \times S_t \times \varepsilon_t
\]  

(4)

Traditionally \( T_t \) is the trend component that described by polynomials in the time \( t \) and a seasonal component.

\[
T_t = \theta_0 + \sum_{j=1}^{k} \theta_j \frac{t^j}{j!}
\]  

(5)

Usually \( k = 0, 1, \) or 2. A seasonal component \( S_t \) with a period of a samples may be described by seasonal indicators (or dummy variables).

\[
S_t = \sum_{j=1}^{k} \theta_j \delta_{tj}, \sum_{j=1}^{k} \theta_j = 0
\]  

(6)

where the \( \delta_{tj} = 1 \) if \( t \) corresponds to the seasonal time point \( i \), and otherwise \( \delta_{tj} = 0 \).

Seasonal simple exponential smoothing
Exponential smoothing forecasting can be considered as one of the most popular forecasting techniques since the 1950s (Osman & King, 2015). A lot of empirical studies and forecasting competitions, shows that the usefulness of the exponential smoothing technique for instance Makridakis et al. (1993). Formally, the multiplicative seasonal simple exponential smoothing with no trend equation takes the form of

\[
\hat{Y}_{t(m)} = S_t H_{t-p+m}
\]  

(7)

\[
S_t = \alpha \left( \frac{Y_t}{H_{t-p}} \right) + (1 - \alpha) S_{t-1}
\]  

(8)
\[ I_t = \delta \left( \frac{Y_t}{S_t} \right) + (1 - \delta)I_{t-p} \]  

where \( \hat{Y}_{t(m)} \) is smoothed forecast for \( m \) periods ahead from origin \( t \), \( S_t \) is smoothed level of the series, \( I_t \) is seasonal indices, \( \alpha \) is smoothing coefficient for level, \( \delta \) is smoothing coefficient for seasonality, \( p \) is number of periods in the seasonal cycle, and \( Y_t \) is the time series in period \( t \).

**ARIMA (Box-Jenkins) methodology**

The ARIMA models rely on a statistical modelling theory known as the Box-Jenkins methodology (1970 and consists of autoregressive and moving average parameters. However, the ordinary ARIMA model \((p, d, q)\) unable handle the data, since its seasonal component, therefore ARIMA \((p, d, q)(P,D, Q)s\) was used (Çuhadar, 2014). The equation of seasonal ARIMA model can be written as (Singh, 2013).

\[
(1 - \phi B - \phi_2 B^2 - \ldots - \phi_p B^p)(1 - \beta_1 B^s - \ldots - \beta_p B^{ps})Y_t = \\
\mu + (1 - \psi_1 B - \psi_2 B^2 - \ldots - \psi_q B^q)(1 - \theta_1 B - \theta_2 B^{2s} - \ldots - \theta_Q B^{Qs})\varepsilon_t
\]

where AR(p) Autoregressive part of order p, MA(q) Moving average part of order q I (d) differencing of order d ARs (P) Seasonal Autoregressive part of order P Mas (Q) Seasonal Moving average part of order Q, Is (D) seasonal differencing of order D S is the period of the seasonal pattern appearing

**Measuring Forecast Error**

The goal of the forecast is to minimize error or the predicted values have to closer the real values. Therefore, after the model specified, accuracy of the best forecasting method was determined. There are several criteria that used to compare the method. This study used by Root Mean Square Error (RMSE).

\[
RMSE = \sqrt{MSE} \\
MSE = \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t^2
\]

**RESULTS AND DISCUSSION**

All data in this study obtained from the Chiang Mai Immigration Office that recorded in monthly during from January 2014 to July 2017 a total of 43 months. Then the data were classified into two sets. The first data set for January 2014 to December 2016 for 36 months were used to build the forecasting model by the three methods of Classical decomposition, Seasonal simple exponential smoothing and Box-Jenkins. The second data set for January 2017 to July 2017 a total of 7 months were used to compare the earlier three methods of the forecasting accuracy model via the criteria of RMSE.

The data on the number of long stay Japanese tourist arrivals to Chiang Mai for period January 2014 to December 2016 is shown clearly in Figure 1. It was found that the plot exhibited no trend in the data, but may be seasonal fluctuations.

According to test whether a linear trend and seasonal occur, were found the Daniels trend test shown that no significant trend. Meanwhile, non-parametric Kruskal-Wallis test was found the significance of seasonal at the significance level of 0.05.
Table 1: Model summary of trend and seasonal analysis

<table>
<thead>
<tr>
<th>Testing</th>
<th>Methodology</th>
<th>Statistical Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trend</td>
<td>Daniels trend test</td>
<td>Spearmans rho = -0.051</td>
<td>0.768</td>
</tr>
<tr>
<td>Seasonal</td>
<td>Kruskal-Wallis test</td>
<td>$\chi^2 = 27.53$</td>
<td>0.004*</td>
</tr>
</tbody>
</table>

* p-value < 0.05

Classical Decomposition Method

According to the data had indicated seasonal effect since seasonal decomposition was applied to identify. Consequence, seasonal index was obtained, the seasonal index value describes the effect of seasonal oscillation on the data periodically. Seasonal index values are shown in Table 2.

Table 2: Seasonal index of the data

<table>
<thead>
<tr>
<th>Period</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seasonal</td>
<td>130</td>
<td>115</td>
<td>127</td>
<td>74</td>
<td>99</td>
<td>72</td>
<td>73</td>
<td>104</td>
<td>104</td>
<td>85</td>
<td>103</td>
<td>115</td>
</tr>
</tbody>
</table>

Seasonal Simple Exponential Smoothing

For, the data in this study have an effect of seasonal then the seasonal simple exponential smoothing are applied. The smoothing coefficient from the equation (7) to (9) are estimated by minimizing the Sum of Square Error (SSE) and converged in probability to a global solution of RMSE. The seasonal simple exponential smoothing solution was solved under the GRG Nonlinear method in Solver add-ins of the Excel program. Finally, the model smoothing parameters which provide the smallest RMSE ware determined as $\alpha = 0$ is smoothing coefficient for level, $\delta = 0.828436$ is smoothing coefficient for seasonality.

ARIMA (Box-Jenkins) Methodology

Firstly, the ARIMA (Box-Jenkins) methodology is depend on the stationarity of the series. However, the data in this study present seasonal which is not stationary. Since using first order seasonal differencing that would be written as is employed to the data to remove seasonal before an ARIMA model can be fitted.

\[ Z_t = \nabla_{12}^1 Y_t = Y_t - Y_{t-12} \]  

(12)

The Autocorrelation (ACF) and Partial Autocorrelation (PACF) plot of first difference of long stay Japanese tourist arrivals in Chiang Mai data are shown in Figure 2.
From Figure 2 (a), the ACF drop to zero exponentially mentioning a stationary behavior (Shumway & Stoffer, 2000). Consequently, the ACF of the stationary series peak at $h = 1s, 2s, 3s$; while for PACF, it peaks at $h = 1s, 2s$. On the other hand it means that the ACF is cutting off after lag $3s$ and the PACF is cutting off after $2s$. Accordingly, the model can be (i) AR (2) (ii) MA (2) or (iii) MA (3). However, among the modes only a model that ARIMA with $p = 0, d = 1$ and $q = 2$ process has lowest MSE. Since, ARIMA (0, 0, 2)(0, 1 , 0) model has been selected for forecasting. Then, the fitted ARIMA (0, 0, 2)(0, 1 , 0) model selected for long stay Japanese tourist arrivals. Meanwhile, the model parameters are estimated using Least Square methods is shown by

$$Z_t = 1.090 - 0.697e_{t-1} - 0.549e_{t-2}$$

(13)

**Diagnostic Checking**

After parameter estimation all 3 models, the model has to assess by checking the model assumptions are satisfied whether. The basic assumption is that the assumption of residual which should be independent and normally distributed with zero mean and constant variance. Testing for independence against serial dependence is a fundamental problem in time series analysis. To determine whether a time series, $Y_t$ or $Z_t$, is independent, the ACF of the series is considered (Yürekli, Kurunç, & Öztürk, 2005).
Figure 3 (a), 3(b) and 3(c) show the ACF plots of the residual of the forecasting classical decomposition, seasonal simple exponential smoothing and Box-Jenkins method, respectively. The x-axis represents the number of lags. Dashed red lines indicate 95% confidence interval. In conclusion, all of 3 forecasting models are suitably fitted because of the ACF plots varies within 95% CI bounds ($\pm 1.96/\sqrt{N}$), where $N$ is the number of observations upon which the model is based (Hejase & Assi, 2012).

However, there are various tests that performed for testing constant variance, then in this study used Levenes test. Meanwhile, several statistical tests used for diagnostic checking of normality. In this study Anderson-Darling was used for the diagnostic checking.
Table 3: Seasonal index of the data

<table>
<thead>
<tr>
<th>Testing</th>
<th>p-value</th>
<th>Classical Decomposition</th>
<th>Seasonal Simple Exponential Smoothing</th>
<th>Box-Jenkins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero Mean</td>
<td>$t$</td>
<td>0.823</td>
<td>0.714</td>
<td>0.714</td>
</tr>
<tr>
<td>Constant Variance</td>
<td>Levene’s test</td>
<td>0.162</td>
<td>0.166</td>
<td>0.137</td>
</tr>
<tr>
<td>Normally</td>
<td>Anderson-Darling test</td>
<td>0.45</td>
<td>0.421</td>
<td>0.453</td>
</tr>
</tbody>
</table>

* $p$-value < 0.05

Table 3 shows the computed $p$-value of the residual testing for each method. As the result $p$-value are greater than the significance level alpha = 0.05, then all of the null hypothesis could not reject.

CONCLUSION AND RECOMMENDATION

For the first time, the aim of this study was to predict the number of long stay Japanese tourist arrivals to Chiang Mai, Thailand in the period January to July 2017. And, construct the appropriate forecasting model which classical decomposition, seasonal simple exponential smoothing and Box-Jenkins was compared. However, finally a combined approach was added employ since in recent years, researchers have admitted combining as a not difficult and available approach to decreasing forecast error (Graefe, Armstrong, Jones Jr, & Cuzán, 2014). Combining forecasts (or composite forecasts) refers to the averaging of independent forecasts which these forecasts can be based on different data or different methods or both additionally the averaging is done using a rule that can be replicated, such as to take a simple average of the forecasts (Armstrong, 2001). The basic equation for combining approach is as follows:

$$\hat{Y}_4 = w_1 \hat{Y}_{1t} + w_2 \hat{Y}_{2t} + w_3 \hat{Y}_{3t}$$  \hspace{1cm} (14)

where $\hat{Y}_{1t}, \hat{Y}_{2t},$ and $\hat{Y}_{3t}$ is forecasting of the classical decomposition method, Seasonal simple exponential smoothing, and the ARIMA (Box-Jenkins) Method, respectively. $\hat{Y}_{4t}$ is forecasting of the combining forecasting $w_i$ is coefficient of each forecasting method $i = 1, 2, 3, 4$ Therefore, using time series combining forecast and parameters was estimated using least square method. Model equations are as follows:

$$\hat{Y}_{4t} = 0.9634\hat{Y}_{1t} + 0.519648\hat{Y}_{2t} - 0.4906\hat{Y}_{3t}$$  \hspace{1cm} (15)
Figure 4. Residual diagnostic checking of the combining forecasting approach
Table 4: Forecast performances and RMSE of model

<table>
<thead>
<tr>
<th>Time</th>
<th>Actual</th>
<th>Classic</th>
<th>SESS</th>
<th>Box-Jenkins</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2017</td>
<td>90</td>
<td>102</td>
<td>97</td>
<td>104</td>
<td>101</td>
</tr>
<tr>
<td>Feb 2017</td>
<td>97</td>
<td>90</td>
<td>91</td>
<td>100</td>
<td>89</td>
</tr>
<tr>
<td>Mar 2017</td>
<td>93</td>
<td>99</td>
<td>104</td>
<td>108</td>
<td>99</td>
</tr>
<tr>
<td>Apr 2017</td>
<td>51</td>
<td>58</td>
<td>61</td>
<td>63</td>
<td>57</td>
</tr>
<tr>
<td>May 2017</td>
<td>74</td>
<td>77</td>
<td>75</td>
<td>75</td>
<td>77</td>
</tr>
<tr>
<td>Jun 2017</td>
<td>50</td>
<td>56</td>
<td>59</td>
<td>60</td>
<td>56</td>
</tr>
<tr>
<td>Jul 2017</td>
<td>55</td>
<td>57</td>
<td>61</td>
<td>64</td>
<td>57</td>
</tr>
<tr>
<td>RMSE</td>
<td>6.68</td>
<td>7.94</td>
<td>10.49</td>
<td>6.35</td>
<td></td>
</tr>
</tbody>
</table>

Based on the analysis of the RMSE in Table 4, it may be concluded that the combining forecasting approach is slightly better compared to the other models. In this study, the three forecasting approaches are considered. However, another approach in data science (e.g., the neural network) which popularity that may be more accurate or suitable.

REFERENCES


