Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 4 (2021)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i4.235

A Comparison of Fuzzy Time Series and ARIMA to Forecast Tourist Arrivals to Homestay in Pahang

Maizatul Akhmar Jafridin[1*], Nur Fatihah Fauzi[2], Rohana Alias[3], Huda Zuhrah Ab Halim[4], Nurizatul Syarfinas Ahmad Bakhtiar[5], Nur Izzati Khairudin[6], Nor Hayati Shafii[7]
[1,2,3,4,5,6,7] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA Perlis Branch, Arau Campus, 02600 Arau, Perlis, Malaysia

*Corresponding author: *fatihah@uitm.edu.my*
Received Date: 15 August 2021
Accepted Date: 9 September 2021
Revised date: 19 September 2021
Published Date: 1 October 2021

HIGHLIGHTS

- The application of Fuzzy Time Series and ARIMA forecasting techniques, to predict number of tourist arrivals at homestays in Pahang.
- The data was obtained from Tourism Pahang Malaysia dating January 2015 until December 2018.
- Fuzzy Time Series seems to be the best method for forecasting tourist arrivals than ARIMA.
- Forecasting plays a major role in tourism to analyse current and past tourist traffic and predict the nature of changes in tourism demand.

ABSTRACT

Predictions of future events must be incorporated into the decision-making process. For tourism demand, forecasting is very important to help directors and investors to make decisions in operational, tactical, and strategic decisions. This study focuses on forecasting performance between Fuzzy Time Series and ARIMA to forecast the tourist arrivals in homestays in Pahang. The main objective of this study is to compare and identify the best method between Fuzzy Time Series and Autoregressive Integrated Moving Average (ARIMA) in forecasting the arrival of tourists based on the secondary data of tourist arrivals to homestay in Pahang from January 2015 to December 2018. ARIMA models are flexible and widely used in time-series analysis and Fuzzy Time Series which do not need large samples and long past time series. These two methods have been compared by using the mean square error (MSE) and mean absolute percentage error (MAPE) as the forecast measures of accuracy. The results show that Fuzzy Time Series outperforms the ARIMA. The lowest value of MSE and MAPE was obtained from using the Fuzzy Time Series method at values 2192305.89 and 11.92256, respectively.

Keywords: *tourist arrivals, forecast, tourism, domestic tourist, fuzzy time series, ARIMA*

INTRODUCTION

Tourism is defined by the World Tourism Organization (UNTWO) as the activity of person(s) travelling somewhere outside their usual places for a purpose and staying there in within a year. People travel for different reasons that might include one of the followings: for leisure, recreation, holiday, business, visiting relatives, or for medical purposes. There is more to tourism than people's general idea about it; it is closely

connected to various other sectors like politics, economics, religion, agriculture, environment, health, finance, transportation, society, immigration, and education. In fact, the tourism industry has been acknowledged to be one of the world's major significant service industries (Bhuiyan et al., 2013). This industry plays the key role in economic growth as an important foreign exchange mediator, and it will continue to be so to many countries across the globe (Salman & Hasim, 2012).

Homestays are one of the important accommodation services, apart from hotels and resorts, in the tourism industry where the tourists can experience the country's rich culture. Since homestays has potential in attracting both domestic and international tourists, the Ministry of Tourism Malaysia has considered the Homestay Program a top priority. Homestays can greatly help in the promotion of local attractions to international tourists because it focusses on the lifestyle and experience of the culture and economic activities. Besides that, there is also the opportunity of cross-cultural exchange between hosts and guests and the potential impact of cultural differences on selection of destination attributes (Fathilah et al., 2011).

There are several methods, which have been applied to compare the accuracy of many univariate and multivariate models in forecasting the international city tourist arrival in Paris from its most important foreign source markets, namely Germany, Italy, Japan, the United Kingdom, and the United States (Gunter & Onder, 2014). These varying methods are EC-ADLM, classical and Bayesian VAR, TVP, ARMA, ETS, and I model. The RMSFE and the MAE were used to assess the accuracy of the methods used in their study, and they found that the univariate models of ARMA(1,1) and ETS are the most accurate in predicting the foreign source markets from US and UK.

Claveria & Torra (2013) identified the best method for forecasting Catalonia's tourist arrivals by comparing artificial neural network (ANN) and two-time series models, ARIMA and self-exciting threshold autoregression (SETAR) models. They used the statistical data from 2001 to 2009 of tourist arrivals and the overnight stay from all different countries to Catalonia. The value of the root mean squared forecast error (RMSFE) was compared during the process of analysing the best method, and it was found that compared to ANN and SETAR, the ARIMA model is the most accurate method to forecast the tourist arrivals. This is because the model showed a significant lower forecasting error in most of the countries, but SETAR and ANN showed significant lower forecasting error for six and two countries.

The monthly data from the years 2003 to 2013 were used in a study to forecast the tourism arrivals in Singapore (Kumar & Sharma, 2016) and the application of SARIMA, ARIMA, and Holt winters models were conducted in forecasting the tourist inflow. For SARIMA, the initial stages of the time series showed a non-stationary with seasonality that later became trend, seasonal and irregular components. Finally, with Mean Absolute Percentage Error (MAPE) of 3.21, it was concluded that SARIMA is the most accurate model compared to ARIMA and Holt Winter. Thus, the estimated highest seasonal factor value is for July and the smallest seasonal factor value is January. It was also found that the tourist arrival in Singapore has an increasing trend.

The forecasting of the tourism demand modelling in Malaysia was researched by Loganathan et al. (2010), with the purpose of forecasting the one-period-ahead of international arrivals in the country using the quarterly data from 1995 to 2008 to forecast the year 2009. The applications of the ARIMA models were used to generate the forecast of international tourist arrivals. Next, a formal stationary test was conducted that showed that the series is stationary. After that, the autoregressive and moving average was identified. The study also forecasts future tourist arrivals using the ARMA model, which is a combination of the AR and MA. Finally, the research has concluded that the ARIMA model (1,0,1) cannot be used to predict the tourism demand because the tourism demand is not affected by seasonality.

Other than that, Muainuddin & Hasan (2018) conducted a research that aims to forecast the number of domestic tourist's homestay use in Pahang, using data sets from January 2009 to December 2016 of the number of domestic tourists at homestays in Pahang. The single exponential technique, Box-Jenkins (ARIMA), and the Holt's method were used to forecast Pahang's tourist arrivals. The research has concluded that the single exponential and Holt's methods were more suitable compared to Box-Jenkins model, while the best method based on the MAPE values is single exponential method. For future studies, the researchers have recommended that more forecasting methods be used.

A study on the forecasting of the total number of tourist flow to Xi'an Musuem was conducted by Li et al. (2016). They used the monthly tourist arrival from 2011 to 2014 as the study's secondary data. The researchers compared the existing Fuzzy Time Series model, traditional Grey Model, and time series models (ARMA(2,1)) with the proposed method, Hybrid Fuzzy Time Series model based on entropy and Markov chain optimization method for the study. They concluded that the entropy-based method is the most suitable method to forecast the tourist arrival to Xi'an Musuem.

Based on the previous research on forecasting tourism demand based on Improved Fuzzy Time Series Model by Chou et al. (2010), the proposed model is verified by using tourist datasets and comparing forecasting accuracy. The results showed that the Improved Fuzzy Time Series Model approach outperforms with lower mean absolute percentage error.

In addition, Lee et al. (2012) made study about the performance of forecasting between SARIMA, Holt Winters and Fuzzy Time Series. The methods have been compared by using data of tourist arrivals to Bali and Soekarno-Hatta gate in Indonesia. The result showed Fuzzy Time Series give more accurate forecasts compare to other methods. The Fuzzy Time Series outperformed all other methods, where FTS is the best according to RMSE.

Lastly, Sarahintu & Tarmudi (2015) had studied the application of Fuzzy Time Series method to predict the tourist arrivals to Sabah. Steps of this method were defining fuzzy sets based on the universe discourse, fuzzification, establishing fuzzy logical relationship groups, defuzzification and computing the forecasted results. Based on average forecasting errors, the forecasting accuracy was determined. As a result, Fuzzy Time Series is a suitable method to forecast the number of tourist arrivals.

In this study, tourist arrivals at homestays in Pahang are considered. Ultimately, the comparison of the forecasting methods that is Fuzzy Time Series and ARIMA presented here may allow people to make more accurate forecasts of tourism and help develop planning for various tourism activities.

METHODOLOGY

This study compares Fuzzy Time Series and ARIMA in forecasting. The number of domestic tourists who used homestays in Pahang Malaysia presents the data for this study. The following describes the selected forecasting methods.

Fuzzy Time Series

In previous research, the forecasting of real-world situation has been done by using the Fuzzy Time Series. The fuzzy time series concept is a popular choice for research and application of social science (Chou, 2018). The data used for Fuzzy Time Series was divided into two variables which are Date and Tarrivals (Table 1).

Table 1: Name and the description of variables

Number of variables	Name of variable	Description
1	Date	January 2015 – December 2018
2	Tarrivals	The number of domestic tourist arrival at homestay in Pahang, Malaysia

Step 1: Using Microsoft Excel, the monthly data was converted into a percentage change. This percentage change is computed using Equation 1:

$$y_n = \frac{(y_n - y_{n-1})}{y_{n+1}} \times 100 \tag{1}$$

where, y_n is the actual value of domestic tourist arrivals at time t, and y_{n-1} is the observed value of domestic tourist arrivals at $t-1$.

Step 2: The maximum and minimum value need to be identified from the percentage of changes. Then, define the Universe of discourse, U by using Equation 2:

$$U = [D_{min} - D_1, D_{max} + D_2] \tag{2}$$

where D_1 and D_2 are the positive number that needs to be assigned in U.

Step 3: Construct the fuzzy set U_i within the same length of the intervals where i is equal to 1 until 7. The U_i will be constructed into equal length of intervals where i equal to 1 to 7. Next, fuzzification of interval and the frequency distribution will be calculated as follows:

$$\text{Length of interval} = \frac{(D_{max} - D_2) - (D_{min} - D_1)}{7} \tag{3}$$

Next, each interval was added by the length of interval or also known as fuzzification of interval and the frequency was generated by using Microsoft Excel.

Step 4: Based on step 2, the interval of $v_1, v_2, v_3, \ldots, v_n$ was generated. The interval was done in form of trapezoidal number which can be represented as follows:

$$\begin{aligned} A_1 &= [b_0, b_1, b_2, b_3], \\ A_2 &= [b_1, b_2, b_3, b_4], \\ A_3 &= [b_2, b_3, b_4, b_5], \\ &\vdots \\ A_n &= [b_{n-1}, b_n, b_{n+1}, b_{n+2}] \end{aligned} \tag{4}$$

where b_n is the membership values.

Step 5: All the data need to be listed in terms of percentage and each data is classified according to the generated interval from step 4. Next, based on the data classification, the fuzzy logical relationship needs to be generated to which the fuzzy logical relation is symbolized as:

$$A_i \rightarrow A_j \tag{5}$$

where A_i is the actual data and A_j is the upcoming data.

Step 6: Create the fuzzy logical relationship rule by referring to the fuzzy logical relation in step 5. The rule of the fuzzy logical relation needs to be arranged in groups such as:

$$\begin{aligned} &A_n \rightarrow A_m, \\ &A_n \rightarrow A_k, \\ &A_n \rightarrow A_l. \end{aligned} \tag{6}$$

A_n can be grouped as $A_n \rightarrow A_m, A_k, A_l$.

Step 7: Every fuzzy relation rule group must be classified into one of the three different types of rules set.

Rule 1: The fuzzy group of A_j is empty, it means there is no relationship rules with others. $A_j \rightarrow \varnothing$ is a symbolized for empty A_j and it also can be represented as $A_j \rightarrow A_j$. In this rule, the forecast value is represented as:

$$F_{vt} = R\left[\text{NSTFN}\left(A_j\right)\right] \tag{7}$$

Rule 2: The fuzzy group of A_j is one to one, which means there is only one relationship rule that can be related to A_j. It can be written as $A_j \rightarrow A_m$ and the forecast value can be calculated as follows:

$$\text{NSTFN}\left(A_t\right) = \left[t_2 + \left[\frac{t_4 - t_1}{4}\right] - \frac{t_4 + t_1}{2}, t_2 + \left[\frac{t_4 - t_1}{4}\right], t_3 + \left[\frac{t_4 - t_1}{4}\right], t_3 + \left[\frac{t_4 - t_1}{4}\right] + \frac{t_4 + t_1}{2}\right] \tag{8}$$

$$F_{vt} = R\left[\text{NSTFN}\left(A_m\right)\right] \tag{9}$$

Rule 3: The fuzzy group of A_j is one to many such as $A_j \rightarrow A_{m1}, A_{m2}, A_{m3}$ or

$$\begin{aligned} &A_j \rightarrow A_{m1}, \\ &A_j \rightarrow A_{m2}, \\ &A_j \rightarrow A_{m3}. \end{aligned} \tag{10}$$

The forecast value will be calculated as follows:

$$\text{NSTFN}\left(A_t\right) = \left[t_2 + \left[\frac{t_4 - t_1}{4}\right] - \frac{t_4 + t_1}{2}, t_2 + \left[\frac{t_4 - t_1}{4}\right], t_3 + \left[\frac{t_4 - t_1}{4}\right], t_3 + \left[\frac{t_4 - t_1}{4}\right] + \frac{t_4 + t_1}{2}\right] \tag{11}$$

$$F_{vt} = \frac{R\left[\text{NSTFN}\left(A_{m1}\right) + \text{NSTFN}\left(A_{m2}\right) + \text{NSTFN}\left(A_{ms}\right)\right]}{n} \tag{12}$$

Autoregressive Integrated Moving Average (ARIMA)

The Box-Jenkins approach is synonymous with the general ARIMA modelling and for seasonal data is SARIMA. Auto Regressive (AR) is the lags of the differenced series, Moving Average (MA) is the lags of errors and (I) is the number of differences used to make the time series stationary. The procedure of ARIMA is as follows:

Step 1: The first step is to import the data from excel because the R-Programming is used as the platform to find ARIMA model. Then, convert the data into time series data format by using this command.

Step 2: Used 70% of the data of Tarrival for estimation part and the other 30% for evaluation part to identify ARIMA model. Since the data consisted of 48 sets, a total of 34 data sets are used for estimation. Another 14 data sets will be used for evaluation part.

Step 3: Plot the autocorrelation (ACF) and partial autocorrelation (PACF) to collect more conclusive evidence to identify if it is stationary or not stationary.

Step 4: Next, use the Augmented Dickey Fuller (ADF) test and KPSS test to check the stationarity. If the p-value of KPSS test is more than 5%, then the series is stationary. Meanwhile, the series is not stationary if the p-value of ADF test is more than 5%.

Step 5: If the series is not stationary based on correlogram of ACF and PACF and the result of ADF and KPSS test, perform the first differencing. Next, plot the correlogram for ACF and PACF to check the stationary. Then, test the stationary of the series by using ADF and KPSS test. If the series is stationary, develop model identification.

Step 6: Identify the equation of the ARIMA model using the estimation data. Determine the error measure of ARIMA model using the evaluation model data. Next, checking for the mis-specification using Box-Pierce Q-statistic and Ljung-Box Statistics.

Mean Square Error (MSE)

The model's forecasting performance can be compared using the Mean Square Error (MSE). Using MSE can help to prevent large errors, in addition to its being easy to calculate and understand. In this study, MSE will be calculated to measure the error and to determine the best method that gives the lowest error. The value of MSE is given by

$$e_t = y_t - \hat{y}_t$$
$$\text{MSE} = \frac{\sum_{t}^{n} e_{t+1}^2}{n} \tag{13}$$

where y_t is the actual observed value of total road accident at time t, and $\hat{y}_t$ is the forecasted value.

Mean Absolute Percentage Error (MAPE)

The popular unit for free measure is Mean Absolute Percentage Error (MAPE) that can measure the prediction accuracy of the forecasting method. (Armstrong and Collopy, 1992). MAPE will give the accuracy values in percentage, and it can be written as:

Journal of Computing Research and Innovation (JCRINN) Vol. 6 No. 4 (2021)
https://jcrinn.com : eISSN: 2600-8793
https://doi.org/10.24191/jcrinn.v6i4.235

$$e_t = y_t - \hat{y}_t$$
$$\text{MAPE} = \sum_{t=1}^{n} \frac{\left|(e_t / y_t) \times 100\right|}{n} \quad (14)$$

where e_t is the error, y_t is the actual values and $\hat{y}_t$ is the forecasted values.

Data Collection

The number of domestic tourists who used homestays in Pahang Malaysia presents the data for this study. The monthly data was obtained from Tourism Pahang Malaysia dating January 2015 until December 2018, which is four years.

Data Analysis

The Fuzzy Times Series and R-Studio for Autoregressive Integrated Moving Average (ARIMA) methods can be applied to forecast the tourism demand. The two can also be used to forecast other types of data. Thus, for this study, both methods will be analysed using Microsoft Excel and then compared using the measure of accuracy: Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE).

FINDINGS AND DISCUSSIONS

Using the Microsoft Excel and R-programming software's to analyse and forecast the data of tourist arrival that have been collected and conducted for Fuzzy Time Series and ARIMA, respectively. By identifying the smallest error measure, the comparison and selection of the best model will be determined.

Fuzzy Time Series

The forecasted value of tourist arrival in changes for July 2016 and November 2018 were 22.06%. The calculation process has been done for all fuzzy logical relationship based on the rule types. After calculating the forecasted value by percent, the value needs to be changed to the total number of tourist arrivals at homestays in Pahang. However, the forecasted value for January 2015 and February 2015 cannot be calculated because the input data is not enough. Figure 1 shows the relationship between actual data and forecasted value of the number of tourist arrival at homestays in Pahang using Fuzzy Time Series from January 2015 to December 2018.

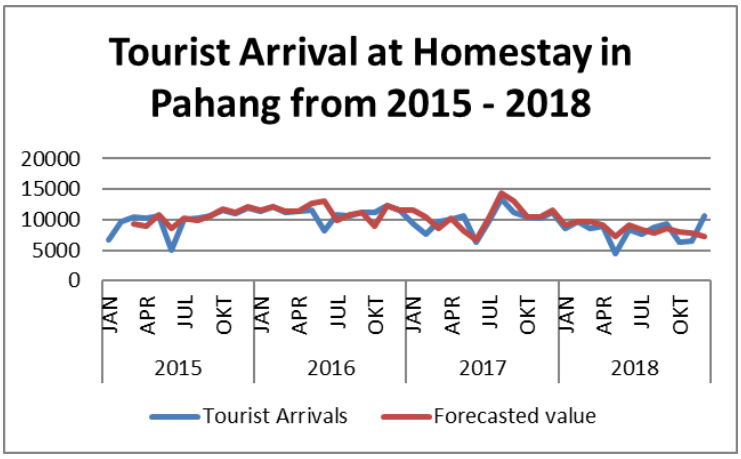

Figure 1: Forecasted value of tourist arrivals at homestays in Pahang using Fuzzy Time Series

ARIMA

The time series plot of tourist arrivals at homestays in Pahang as described in Figure 2 is executed using R-programming. The data set of Tourist Arrival at homestays in Pahang is imported and created using Microsoft Excel. In this analysis, the name of the data set was set as DataArrival and by using command View (DataArrival), the data of tourist arrival at homestay in Pahang from January 2015 to December 2018 will appear. The series requires differencing, test for ADF and KPSS and the best ARIMA model was ARIMA(1,1,0).

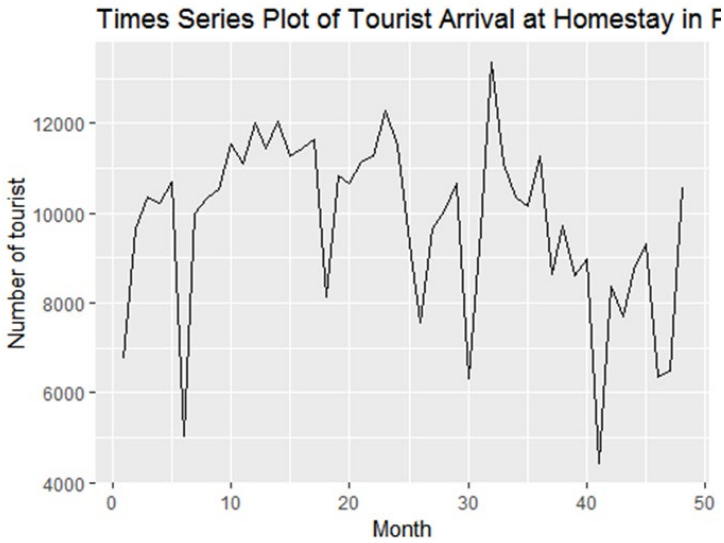

Figure 2: Time series plot of tourist arrivals at homestays in Pahang using ARIMA

Comparisons between Fuzzy Time Series and ARIMA

The predictions of visitor arrivals in the testing period are done using the two forecasting methods, to compare the prediction performance of the three approaches for the period January 2015 to December 2018, the following measures of accuracy were calculated: mean square error (MSE) and mean absolute percentage error (MAPE). The accuracy of forecasting this series by Fuzzy time series is better than forecasting by ARIMA. The reason for this is that the model has the lowest value of MSE and MAPE. Table 2 shows the comparison between two error measures for both methods.

Table 1: Comparison of Error Measure for FTS and ARIMA

Method	MSE	MAPE
Fuzzy Time Series	2192305.89	11.92256
ARIMA(1,1,0)	424655.676	24.075

CONCLUSION AND RECOMMENDATIONS

This paper investigates Fuzzy Time Series and ARIMA methods to predict visitor arrivals at homestays in Pahang. To find the best method to forecast the number of tourist arrivals at homestays in Pahang, this study has compared the performance of two methods. The two methods used in this study are Fuzzy Time Series and Autoregressive Integrated Moving Average (ARIMA). Both data must be analysed to identify the

difference between the actual and forecasted data. Moreover, the best method to forecast the tourist arrival at homestays in Pahang can be determine by identifying the error measure for each of the data in the research. The fuzzy time series is good for predicting visitor arrivals as it gives a small MSE and MAPE values of 2192305.89 and 11.92256 respectively. Thus, the main objective of this research, which is to compare between the two methods is achieved.

There are various more methods that can be applied to forecast the tourist arrivals to homestay in Pahang. Some of methods that can be applied are Holt Winter, Artificial Neural Network, Neural Network Autoregression, SARIMA and other methods. Future researcher can use the suggested methods to study about the comparison between two or more method in forecasting the tourist arrivals to homestay in Pahang to determine which prediction method is the best. In addition, for future researcher who wish to do research regarding forecast using Fuzzy Time Series method, it is recommended to make an accuracy comparison between classical Fuzzy Time Series method and Improved Fuzzy Time Series method. Lastly, for those researchers who are interested to do research about tourist arrival, it is suggested to study the number of tourist arrival in Pahang, the number of domestic tourist arrivals in Pahang, or the effect of tourist arrival to it expenditure.

ACKNOWLEDGMENTS

The authors express sincere gratitude and thanks especially to the officers of IPD Kuala Muda for their contributions and supports in giving us information during interviews to complete our study and also people who helps us direct and indirectly in this study.

CONFLICT OF INTERESTS DECLARATION

The authors declare no conflict of interests regarding the publication of this article.

REFERENCES

Bhuiyan, M. A. H., Siwar, C. & Ismail, S. M. (2013). Tourism development in Malaysia from the perspective of development plans. *Asian Social Science*, 9(9), 11-18.

Chou, H. L., Chen, J. S., Cheng, C. H. & Teoh, H. J. (2010). Forecasting tourism demand based on Improved Fuzzy Time Series model. *Asian Conference on Intelligent Information and Database Systems,* 399-407. doi: 10.1007/978-3-642-12145-6_41

Claveria, O. & Torra, S. (2013). Forecasting tourism demand to Catalonia: Neural Networks vs. Time Series Models. *Economic Modelling*, 36 (2014), 220-228.

Fathilah, I., Roseliza M. A., Noraien, M. & Wan Hafiz, W. Z. S. (2020). A cross-cultural study of destination attributes: Impact on sustainability of island tourism. *Journal of Sustainability Science and Management*, 15(1), 1-14.

Chou, M. T. (2018). Fuzzy Forecast Based on Fuzzy Time Series. Time Series Analysis – Data, Methods, and Applications, Chun-Kit Ngan, IntechOpen. doi: 10.5772/intechopen.82843

Gunter, U. & Onder, I. (2014). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management*, 46, 123-135.

Kumar, M. & Sharma, S. (2016). Forecasting tourist in-flow in South East Asia: A case of Singapore. *Tourism & Management Studies*, 12(1), 107-119. doi: 10.18089/tms.2016.12111

Lee, M. H., Nor, M. E., Suhartono, Sadaei, H. J., Abd Rahman, N. H. & Kamisan, N. A. B. (2012). Fuzzy Time Series: An application to tourism demand forecasting. *American Journal of Applied Science*, 9(1). 132 -140.

Li, Y., Cao, H., Meng, H. Y. (2016). A hybrid tourism demand forecasting model based on Fuzzy Time Series. *International Conference on Artificial Intelligence and Computer Science (AICS 2016),* 171-177.

Loganathan, Nathankumar & Ibrahim, Y. (2020). Forecasting international tourism demand in Malaysia ising Box Jenkins Sarima application. *South Asian Journal of Tourism and Heritage*, 3(2), 51-60.

Muainuddin. M. M. A. M. & Hasan, M. N. (2018). Domestic tourism forecasting in Pahang: Comparison of selected techniques. *Global Business and Management Research*, *suppl. Special Issue*, 10(3).

Salman, A. & Hasim, M. S. (2012). Factors and competitiveness of Malaysia as a tourist destination: A study of outbound Middle East tourists. *Asian Social Science*, 8(12), 48-54.

Sarahintu, M. & Tarmudi, Z. (2015). Forecasting tourist arrivals to Sabah using Fuzzy Time Series. *Proceedings of the International Conference on Natural Resources, Tourism and Services Management 2015*, 481-488.