Advances and Applications in Statistics
© 2025 Pushpa Publishing House, Prayagraj, India
https://pphmjopenaccess.com
https://doi.org/10.17654/0972361725004
Volume 92, Number 1, 2025, Pages 77-87 **P-ISSN: 0972-3617**

TIME SERIES FORECASTING FOR TOURISM INDUSTRY IN MALAYSIA

Noratikah Abu[1], Siti Aishah @ Tsamienah Taib[1], Nurul Amira Zainal[2,*], Nor Azuana Ramli[1] and Clark Kendrick Go[3]

[1]Universiti Malaysia Pahang Al-Sultan Abdullah
Lebuh Persiaran Tun Khalil Yaakob
26300 Kuantan, Pahang
Malaysia

[2]Universiti Teknikal Malaysia Melaka
76100 Durian Tunggal
Melaka, Malaysia

[3]Department of Mathematics
Ateneo de Manila University
Philippines

Received: September 1, 2024; Accepted: October 7, 2024

2020 Mathematics Subject Classification: 62M10, 62P25.

Keywords and phrases: Box-Jenkins model, SARIMA model, time series.

*Corresponding author

Communicated by K. K. Azad

How to cite this article: Noratikah Abu, Siti Aishah @ Tsamienah Taib, Nurul Amira Zainal, Nor Azuana Ramli and Clark Kendrick Go, Time series forecasting for tourism industry in Malaysia, Advances and Applications in Statistics 92(1) (2025), 77-87.
https://doi.org/10.17654/0972361725004

Published Online: November 9, 2024

Abstract

This study is conducted to forecast the future tourism demand in Malaysia by applying Box-Jenkins modelling. The time series data of tourist arrivals volume in Malaysia before MCO retrieved from MOTAC Malaysia database is implemented in this study. The forecast evaluation methods used to validate the best Box-Jenkins model before proceeding to forecasting stage are MAPE and RMSE, and the analysis was performed by using Python. The findings show that $SARIMA(2, 1, 1)(0, 1, 1)_{12}$ was considered as highly accurate forecasting model based on its least error produced.

1. Introduction

In early 2020, due to the COVID-19 pandemic, the tourism industry has been greatly affected causing enormous losses in all aspects. In 2020, there was a decrease of 83.4% of the tourist arrivals volume compared to the previous year, and its contribution towards the total Gross Domestic Product (GDP) fell to just 14.1% [1]. In response to this situation, the government is dedicated to redeveloping this industry after post-pandemic.

Based on Frechtling [2], a good forecasting model greatly contributes to the process of decision-making. According to Peng et al. [3], in time series forecasting, the common quantitative methods used are time-series, econometric and artificial intelligence (AI). According to Hamzah et al. [4], $SARIMA(1, 1, 1)(1, 1, 4)_{12}$ is able to provide accurate prediction on the number of international tourist arrivals to Malaysia since it gives the lowest value of error of MSE, MAD and MAPE. In 2019, Thushara et al. [5] implemented SARIMA model to forecast total number of international tourist arrivals in Sri Lanka. They claimed that SARIMA method was appointed as the best method due to the forecasting accuracy examined.

Abdul Halim and Nora [6] made a comparison between Box-Jenkins models, which are ARIMA and SARIMA with singular spectrum analysis (SSA) to identify the best model to forecast the international tourist demand to Malaysia. Blanco and Ronald [7] demonstrated SARIMA model to predict the international tourist demand in Puno-Peru. From the empirical results, they found that $SARIMA(6, 1, 2, 4)(1, 0, 1)_{12}$ model outperforms the other models, attaining the highest forecasting accuracy. Based on the literature search stated above, the traditional Box-Jenkins time series model (ARIMA and SARIMA) still can be the most accurate forecasting model, especially in tourism demand forecasting. Therefore, in this study, Box-Jenkins model will be implemented using the data of tourist arrivals volume in Malaysia to support the above statement.

2. Methodology

2.1. Data collection

The time series data used in this study is the tourist arrivals volume in Malaysia, consisting of 242 data (January 2000 until February 2020) before the movement control order commencement. By considering ratio 80:20, there are 192 training data and 50 testing data. The training data will be used to develop the forecasting model. Meanwhile, the testing data evaluate the model's accuracy level.

2.2. Box-Jenkins modelling

Box-Jenkins modelling implements the iterative procedure to generate the best performance in forecasting which consists of four stages:

Stage 1. Model identification

There are two parts in the model identification, known as data screening and identification of the model. The first step in data screening is data plotting and the second step is data stationarity. Box-Cox transformation is used to check the stationary invariance of the data since it can convert non-normal data to more normal distribution like data, stabilizing variance and reducing homoscedasticity (constant and finite variance process) [8]. Stationarity in-mean verification will apply the Dickey and Fuller unit root test (ADF test) [9]. The autocorrelation function (ACF) and partial autocorrelation function (PACF) plots can also measure data stationarity in-mean. After obtaining a stationary series, Portmanteau test will determine serial correlation. Box-Jenkins requires serially connected data.

In this study, ARIMA model is implemented. ARIMA model is denoted by $ARIMA(p, d, q)$, where p represents the order of autoregressive (AR) part, d indicates the number of differencing and q denotes the order of moving average (MA) part. Differencing is required to transform non-stationarity in-mean data to stationary in-mean data. When a plot of the data shows that the series varies about a fixed level, the sample autocorrelations slowed down noticeably, and the differencing can stop. The general equation of ARIMA model is denoted as in equation (1):

$$\begin{aligned} y_t' &= \phi_0 + \phi_1 y_{t-1}' + \phi_2 y_{t-2}' + \cdots + \phi_p y_{t-p}' \\ &\quad + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}. \end{aligned} \tag{1}$$

The other non-stationary Box-Jenkins model is seasonal ARIMA (SARIMA) model. SARIMA model is an improved version of ARIMA model by including additional seasonal terms in the existing ARIMA model. The general form of SARIMA model is $SARIMA(p, d, q)(P, D, Q)_m$, where (p, d, q) represents a non-seasonal part, $(P, D, Q)_m$ represents a seasonal part and m represents the number of observations per year. The general equation of SARIMA model is given as follows:

$$\Phi_P(B^S)\phi_p(B)(1-B)^d(1-B^S)^D y_t = \Theta_Q(B^S)\theta_q(B)a_t. \tag{2}$$

Stage 2. Parameter estimation

The parameter estimation used is maximum likelihood method (MLE). In this study, the Box-Jenkins model with smallest Akaike information criteria (AIC) value will be chosen for the next stage.

Stage 3. Diagnostic checking

The adequacy of selected forecasting model must undergo three diagnostic checking before proceeding to the next stage. These are outcome from the analysis of residual, also known Portmanteau test or general goodness-to-fit statistic, homoscedasticity test and normality test. In analysis of residual, the Ljung-Box Q test (LBQ test) is performed on residuals. LBQ test is applied on a squared residual series to test the ARCH effects in the residuals under homoscedasticity test. To identify whether the residuals follow normal distribution or not, the normality test is carried out using Jarque-Bera test (JB test).

Stage 4. Forecasting

In this study, one-step ahead is considered. Two metrics, mean absolute percentage error (MAPE) and root mean square error are used to measure the performance of forecasting model.

3. Results and Discussion

Stage 1. Model identification

The data is divided into two subsets where 80% of the data was allocated for training, while the remaining 20% for testing data. The training data is used to develop the model, whereas the testing data is implemented to measure the model performance. The volume of tourist arrivals in Malaysia for training data can be seen in Figure 1. Since the time series plot shows upward and downward trends, it is obvious that the data is not stationary in terms of its mean, but the data may be stationary in terms of its variance. To identify whether our assumptions are correct, we conduct Box-Cox transformation and Augmented Dickey-Fuller (ADF) test. From the Box-Cox transformation, it is discovered that the value of λ is close to 1.0. Hence, there is no transformation of data required, and it is confirmed that the data used is stationary in terms of its variance. Then, the stationarity in-mean checking process was conducted using ADF test. The result shows that the data is not stationary in terms of its mean since the $p\text{-value} = 0.8517$ is greater than the level of significance, $\alpha = 0.05$. The non-stationarity in-mean also can be seen in ACF and PACF plots as given in Figure 2(a) and Figure 2(b), respectively. Based on Figure 2(a), the autocorrelation coefficient value is large and the value of ACF decreases slowly. Therefore, we can conclude that the data is not stationary in-mean.

Since our data series exhibits seasonal trends, therefore the model will comprise of non-seasonal and seasonal parts. To stabilize the mean, first differencing must be carried out for both, non-seasonal and seasonal parts. After the first differencing for non-seasonal part, the ADF test is conducted once again. It is found that the *p*-value is 0.000 and it can be verified that the data is now stationary in-mean. The ACF and PACF plots after first differencing for non-seasonal part in Figure 3(a) and Figure 3(b) also show that data is stationary in-mean. According to these two correlograms, the PACF shows a "cut off" after lag 2, indicating that the non-seasonal AR(0), AR(1) or AR(2) is appropriate to fit the data. Similarly, the ACF lag "cuts off" after lag 2, suggesting that non-seasonal MA(0), MA(1) or MA(2) is appropriate to fit the data. First seasonal differencing for seasonal part was performed to eliminate the seasonality of the data series. After the differencing process, ADF test was carried out and found that the *p*-value is smaller than the significance value, $\alpha = 0.05$. Therefore, the seasonal part is already stationary in terms of its mean. The plots of ACF and PACF for seasonal part after differencing are illustrated in Figure 4(a) and Figure 4(b), respectively. According to these two correlograms, the PACF lag "cuts off" after lag 1, suggesting that the seasonal AR(0) or AR(1) is appropriate to fit the data. Similarly, the ACF lag "cuts off" after lag 3, suggesting that seasonal MA(0), MA(1), MA(2) or MA(3) is appropriate to fit the data.

Then, a hypothesis testing is carried out. The null hypothesis is the time series not serial correlated and the alternative hypothesis is the time series which is serial correlated. By using Ljung-Box Q test (LBQ test), it is found that the *p*-value is 0.000, which is smaller than significance level, $\alpha = 0.05$. Therefore, the null hypothesis is rejected, and we can say that there exists serial correlation between time series data. According to Figure 4(a) and Figure 4(b), the seasonal part of AR and MA models can be seen clearly in lags of ACF and PACF, where there is significant spike at lag 12. Therefore, there is no doubt in confirming that SARIMA model is appropriate for this study.

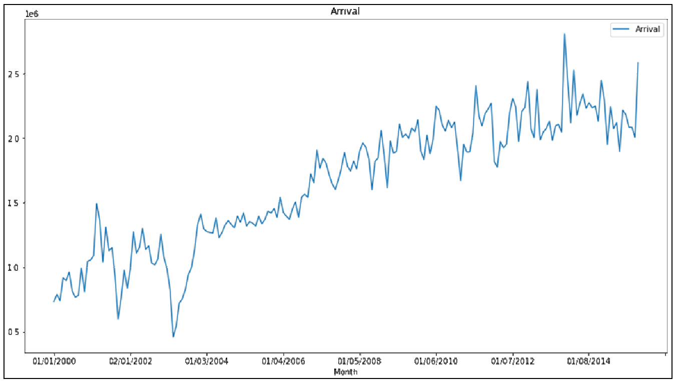
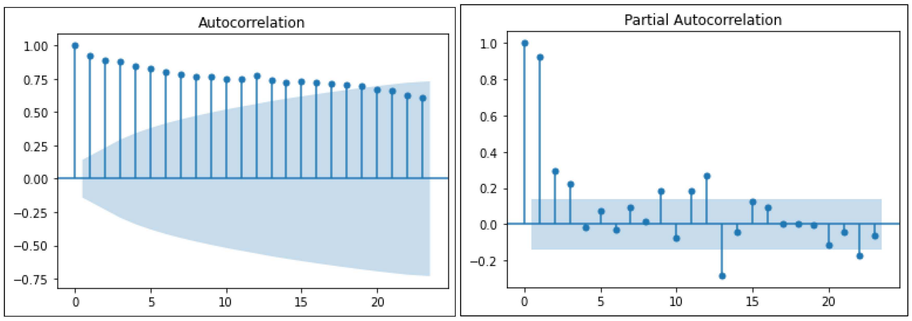
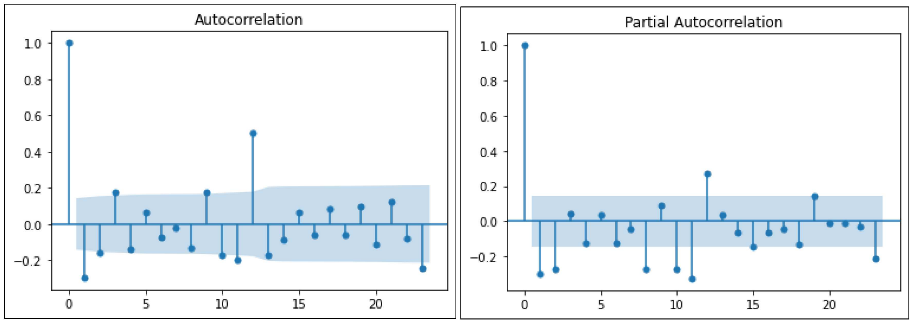


Figure 1. The training data of the number of tourist arrivals in Malaysia.



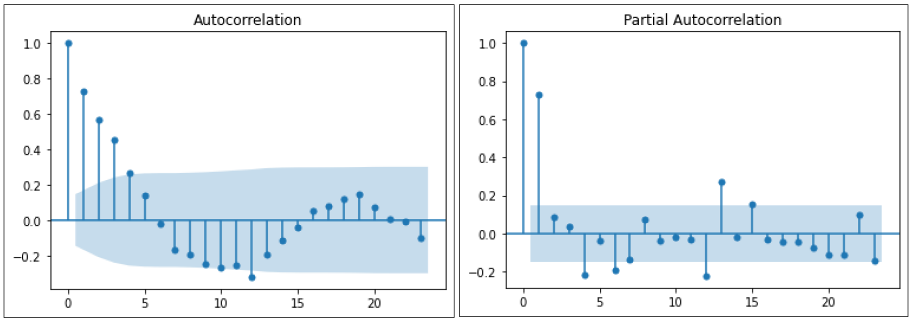
(a) (b)

Figure 2. The (a) ACF plot, and (b) PACF plot for non-stationarity in-mean.



(a) (b)

Figure 3. The plot of (a) ACF and (b) PACF for non-seasonal part after first differencing.



(a) (b)

Figure 4. The plot of (a) ACF and (b) PACF for seasonal part after first differencing.

Stage 2. Parameter estimation

Based on the output from auto.arima, the selected model for the data was $SARIMA(2, 1, 1)(0, 1, 1)_{12}$ based on its lowest AIC value. Besides that, since all the p-values are smaller than significance level, $\alpha = 0.05$, it shows that all the parameter values are significant. The equation for the best model is denoted as follows:

$$\varphi_2(B)(1-B)^1(1-B^{12})^1 y_t = \Theta_1(B^{12})\theta_1(B)a_t,$$

$$(1-\varphi_1 B-\varphi_2 B^2)(1-B)(1-B^{12})y_t = (1-\Theta_1 B^{12})(1-\theta_1 B)a_t,$$

$$(1+0.6618B+0.3217B^2)(1-B)(1-B^{12})y_t = (1+0.6057B^{12})(1-0.4908B)a_t.$$

Stage 3. Diagnostic checking

Based on the hypothesis testing, the outcome shows that the residuals are not serial correlated since the $p\text{-value} = 0.3507$ is greater than the significance level, $\alpha = 0.05$. In addition, the $p\text{-value} = 0.7$ obtained from the homoscedasticity test concluded that there are no ARCH effects in the residuals. While, for normality test, we found that the residuals do not follow the normal distribution. But, according to histogram plot (see Figure 5), the residuals follow the histogram closely. We can conclude that this phenomenon occurs due to the existing of outliers in the data. Therefore, this result for normality test is neglected.

Stage 4. Forecasting

In the forecasting stage, the selected model $SARIMA(2, 1, 1)(0, 1, 1)_{12}$ is implemented in forecasting the testing data. The comparison between actual testing data and forecasted testing data were presented in Figure 6. Based on Figure 6, the predicted testing data follows the pattern of actual testing data accurately. Since there are no zero values in this time series data, MAPE is suitable to be used as forecast evaluation method. The analysis of these forecast evaluation methods is tabulated in Table 1. Based on the results obtained, the MAPE for testing data is 5.3120%.

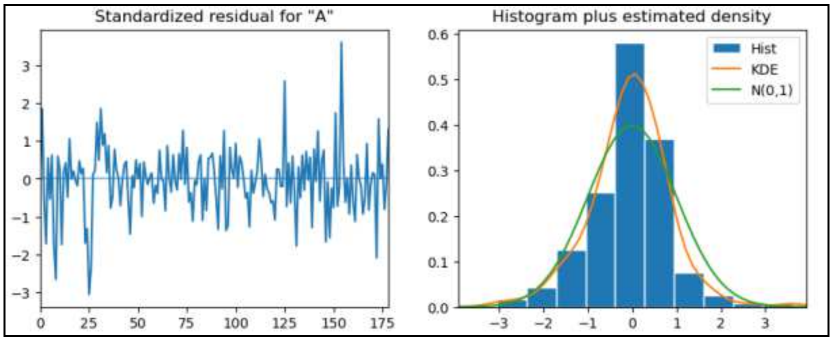


Figure 5. The residual distribution of training data.

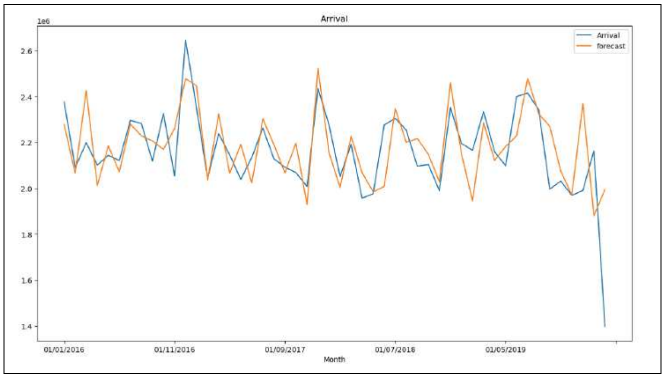


Figure 6. The comparison between actual testing data and predicted testing data.

Table 1

MAPE and RMSE analyses for $SARIMA(2, 1, 1)(0, 1, 1)_{12}$ model

Forecast evaluation	MAPE	RMSE
$SARIMA(2, 1, 1)(0, 1, 1)_{12}$	5.3120	152539.7805

4. Conclusions

Overall, the traditional Box-Jenkins model is one of the appropriate models for time series demand forecasting. The findings show that $SARIMA(2, 1, 1)(0, 1, 1)_{12}$ is highly accurate forecasting model based on its least error produced.

Acknowledgement

The authors would like to thank Universiti Teknikal Malaysia Melaka for providing financial support for this research. Also, they acknowledge the Geran Penyelidikan Kerjasama MTUN UMPSA (RDU233501).

References

[1] Department of Statistics Malaysia (DOSM), Tourism Satellite Account 2020, Last modified 2021. http://www.dosm.gov.my/portal-main/release-content/tourism-satellite-account-2020

[2] Douglas C. Frechtling, Forecasting Tourism Demand: Methods and Strategies, Butterworth-Heinemann, Oxford, 2001.

[3] Bin Peng, Hong Song and Greg I. Crouch, A meta-analysis of international tourism demand forecasting and implications for practice, Tourism Management 45 (2014), 181-193.

[4] D. I. Amir Hamzah, Maria Elena Nor, Sabariah Saharan, N. F. M. Hamdan and N. A. I. Nohamad, Malaysia tourism demand forecasting using Box-Jenkins approach, International Journal of Engineering and Technology 7(4) (2018), 454-457.

[5] S. C. Thushara, Su Jen-Je and S. B. Jayatilleke, Forecasting international tourist arrivals in formulating tourism strategies and planning: the case of Sri Lanka, Cogent Economics and Finance 7(1) (2019), Article ID 1699884.

[6] S. S. Abdul Halim and M. Nora, Modeling and forecasting of tourism demand in Malaysia, International Journal of Current Science Research and Review 3(12) (2020), 230-244.

[7] L. F. L. Blanco and W. M. H. Ronald, Modeling and forecasting international tourism demand in Puno-Peru, Brazilian Journal of Tourism Research 14(1) (2020), 34-55.

[8] G. E. P. Box and D. R. Cox, An analysis of transformations, Journal of the Royal Statistical Society: Series B (Methodological) 26(2) (1964), 211-243.

[9] Douglas C. Montgomery, Cheryl L. Jennings and Murat Kulahci, Introduction to Time Series Analysis and Forecasting, John Wiley and Sons, Inc., Hoboken, New Jersey, 2015.