Kamarul Ariffin MANSOR
Universiti Teknologi MARA (UiTM), Kedah, Malaysia
Wan Irham ISHAK
Universiti Teknologi MARA (UiTM), Kedah, Malaysia

FORECASTING TOURIST ARRIVALS TO LANGKAWI ISLAND MALAYSIA

Case Study

Keywords
Tourist Arrivals,
Forecasting,
Time Series,
Exponential Smoothing,
ARIMA,
ARFIMA

JEL Classification
C51, C52, C53, L83

Abstract

Tourism is the act of travelling for a person or group of people from their own locality to a specific destination in a short term or long term period either for leisure or business purposes. Tourism is an important sector in the Malaysian economy where tourism development will lead to the positive economic development of the country and in general improve the quality of life for all citizens. Therefore, forecasting tourist arrivals with high accuracy becomes important since it may ensure the development and the readiness of all tourism related industries such as hotels, transportation, food and services industries and their best shape. This study focuses on tourist arrivals in Langkawi Island as one of the major tourist attractions situated in the northerly region of Peninsular Malaysia. Importantly, this paper attempts to measure and compare the performance of forecasting with Exponential Smoothing, ARIMA and ARFIMA models using the R software package.

1. Introduction

Tourism is an important sector in the Malaysian economy. As viewed by Les Lumsdon (1997), tourism is a complex activity and geographical aspects to try to get a variety of different services and different situations from the point of origin to destination. According to the definition given by the World Tourism Organization (UNWTO), tourism comprises the activities of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes (UNWTO, 2014).Thus, tourist arrivals can either be classified as domestic or international. Importantly, tourism development will contribute to positive economic development and improve the quality of life for all citizens.

The increase in tourist arrival either domestic or international, indicates a healthy economy to Malaysia. Based on the monetary economics view, this means that increase in domestic tourist ensure less currency outflow and increase in international tourist will increase the foreign currency inflow to the country. This can be proven by past data in this study. Based on the report from the Department of Statistics, Malaysia (2013), the domestic visitors from urban areas recorded the highest spending MYR36.1 billion (75.6 per cent) of the total expenditure MYR47.8 billion while, domestic visitors from rural areas spend only 24.4 per cent or MYR11.7 billion. Total expenditure for both categories of visitors each increased by 13.7 percent and 10.1 percent compared to the year 2011.

In addition, in 1998 the number of tourist arrivals in Malaysia is only at 5.56 million with a total receipts of approximately MYR 8.6 billion. This number has increased drastically over the decades and in year 2013, the total number of tourist arrivals has reached to 25.72 million with a total receipts amounted to approximately MYR 66.44 billion, a growth of 398% over the 15-year period. It was reported by the Malaysia Tourism Promotion Board (MTPB) that the number is still in the rising trend when in the year 2014, it is recorded a total number of 27.44 million tourist arrivals which is equivalent to approximately 6.7% growth from the previous year 2013 (MTPB, 2014). Significantly, according to the World Travel and Tourism Council annual report (2014), the direct contribution of travel and tourism to Gross Domestic Product (GDP) was MYR 70.4 billion (7.2% of total GDP) in 2013, and is forecast to rise by 7.0% in 2014, and to rise by 4.4% pa, from 2014-2024, to MYR115.4 billion (7.6% of total GDP) in 2024. In addition, the total contribution of Travel and Tourism to GDP was MYR158.2 billion (16.1% of GDP) in 2013, and is forecast to rise by 6.8% in 2014, and to rise by 4.5% pa to MYR 262.5 billion (17.3% of GDP) in 2024. Thus, this reflects the importance of travel and tourism industry in boosting the economy of any country in the world including Malaysia.

In this paper, the tourist arrivals analysis is focused mainly on Langkawi Island, one of the famous tourism destination located in the utmost northern part of Peninsular Malaysia. Langkawi Island provides a spectacular escape for both business and leisure. This is true based on the historical data of the tourist arrivals which shows an increasing trend since 2005 until now. Statistics provided in Langkawi Development Authority (LADA) websites indicate that the tourist arrivals in 2005 were recorded at 1.84 million and this number has increased to 3.57 million in 2014, a percentage increase of approximately 94% (LADA, 2015). Due to the increasing trend, it is crucial to forecast the number of tourist arrivals with accuracy since it will benefit the direct and indirect activities that related to the tourism industry. Thus, the government or related organization and agencies could use the forecast figure to encourage social, economic and physical development of Langkawi, to establish a development scenario such as preservation of natural resources, to establish a development for a conducive environment, and also to create attractive opportunities for foreign investors (Marzuki, 2011).

2. Literature Review

Forecasting is the act of making predictions about the events or circumstances that will occur in the future (Bowerman and Connell, 1999). Forecasting is very important for prediction of future events and needs to be incorporated into any decision making process. Predictions of events are dependent on information concerning events which occurred in the past and using time series data to provide predictions.

Forecasting technique is based on two main methods, the explanatory methods and the extrapolation methods. Explanatory method is a method of forecasting by pointing to factors that are believed to affect the analysis of predictive value. Meanwhile, the extrapolation method is forecasting techniques based on past data for the event (Arsham, 1994). According to Kulendran and Witt (2001), the first study of forecasting in the field of tourism has been run by Martin and Witt in 1989. They use seven different methods to make comparisons, including exponential smoothing and linear regression. Since then, the study of tourism through forecasting techniques has continued. In 2008, a study on tourist arrival to Malaysia using forecasting technique was done by Mahendran Shitan. Shitan (2008) in his study, update and compare the performanceof three time series models, using ARMA and ARFIMA models, for modelling tourist arrivals to Malaysia. In the study, he found that ARFIMA(0, -0.2058, 12) model was

the best model. However, in his study, the evaluation data set used is also used in the modelling process. In another study by Nanthakumar, Subramaniam, and Kogid (2012), forecasting tourism demanded from ASEAN countries to Malaysia was evaluated using SARIMA approach and found that seasonality model does not offer any valuable insights or provide reliable forecasts of tourism demand in Malaysia by ASEAN countries.

3. Methodology

The data set used in this study is discussed in section 3.1 and continued with a brief discussion of the models used in this study in section 3.2.

3.1 Data

The data used in this study were obtained from a secondary source. Tourist arrivals statistics were obtained from the LADA websites. The data sets consist of 156 monthly statistics on tourist arrivals to Langkawi Island from January 2002 until December 2014.

Figure 1 displays a time series plot for all the 156 data from January 2002 until December 2014. It is clear that the total number of tourist arrivals shows a gradual growth every year with a consistent fluctuation showing a seasonal effect on the rise and fall on the number of tourists entering Langkawi Island. The plot also shows that the data is not a stationary time series.

3.2 Exponential, ARIMA and ARFIMA Models

The models investigated in this study come from three family models, namely Exponential Smoothing, Mixed Autoregressive Integrated Moving Average (ARIMA) with Seasonal Component, and Fractionally Integrated ARMA (ARFIMA). In this section, a brief description on each of the models is discussed.

Exponential smoothing is a forecasting technique that establishes a simple statistical model for time series. It tries to detect the "change" in a time series using the new observed time series data to estimate the parameter that gives the most "recent" time series value (Bowerman and O'Connel, 1999). The exponential smoothing model family ranges from the simplest form of simple exponential smoothing to an exponential smoothing models that include not only trends but also the seasonal effect. Please refer Bowerman, O'Connell and Koehler (2005) for the complete state space models for exponential smoothing methods. However, in this study, a total of fifteen exponential smoothing methods will be evaluated and the best method will be used for evaluation purpose. All of the fifteen exponential smoothing methods adapted from an article written by Hyndman and Khandakar (2008) are presented in Table 1.

Under the stationary assumption, a mixed autoregressive and moving average model or simply denoted as ARMA(p,q) can be modeled and defined as a random variable$\{Y_t\}$ given by

$$(1 - \phi_1 B - \phi_2 B^2 + ... + \phi_p B^p)\, Y_t = \mu + (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)\varepsilon_t$$

When the stationary assumption of the variable is not met, then the ARIMA model will be formulated. Since the data used in this study show a seasonal pattern, as discussed in section 3.1, ARIMA with seasonal component model will be estimated and evaluated. In general the model is denoted as ARIMA$(p,d,q)(P,D,Q)_m$ and is defined as

$$\phi(B^m)\phi(B)(1- B^m)^D(1-B)^d Y_t = c + \Theta(B^m)\theta(B)\varepsilon_t$$

where B is the usual backward shift operator.

If the value of d satisfies $0 < |d| < 0.5$, then a long memory process or Fractionally Intergrated ARMA is obtained which is often denoted as ARFIMA(p,d,q). In general, a stochastic process $\{Y_t\}$ is called an autoregressive fractionally integrated moving average process (ARFIMA(p,d,q) process) if the fractionally differenced process is an ARMA process, i.e.,

$$(1 - \varphi_1 L - \ldots - \varphi_p L^p)(\Delta^d Y_t) = (1 + \theta_1 L + \ldots + \theta_q L^q)(u_t),$$

where d is not restricted to integral values.

4. Results

This section provides the discussion on the results of our study. First of all, the data set was first divided into two sets, a training set for estimation of the model parameters, and the test or evaluation set to evaluate and compare the models in order to find the best models to forecast the tourist arrivals to Langkawi Island. The training set contains 132 data points from January 2002 until December 2012, while the evaluation set consists of two-full year cycle from January 2013 until December 2014. The three best models from each model family are given below.

Model 1

ETS(A,N,A) for monthly Langkawi Island tourist arrivals. This is an additive error model with no trend and with a monthly seasonal pattern such that the estimated parameters are given as,

(α = 0.1919, γ = 0.0001, ℓ = 162437.0914, s_{-11} = 108991.1, s_{-10} = 9222.1, s_{-9} = -24644.4, s_{-8} = -27324.2, s_{-7}= -6951.6, s_{-6} = -5666.2, s_{-5} = 19571.0, s_{-4} = -6517.1, s_{-3} = -31658.6, s_{-2} = -4538.2, s_{-1} = -3068.4, s_0 = -27415.6)

Model 2

ARIMA$(1,0,2)(0,1,1)_{12}$ model with drift for monthly Langkawi Island tourist arrivals with parameter estimates given as,

(ϕ_1 = 0.7795, θ_1 = -0.7552, θ_2 = 0.2061, Θ_1 = -0.4152, c = 789.9110)

Model 3

ARFIMA(12,0.25,11) with estimated parameters:

(φ_1 = 0.125, φ_2 = -0.063, φ_3 = -0.020, φ_4 = 0.003, φ_5 = -0.128, φ_6 = 0.191, φ_7 = -0.053, φ_8 = 0.057, φ_9 = 0.014, φ_{10} = 0.006, φ_{11} = 0.219, φ_{12} = 0.819,

θ_1 = 0.807, θ_2 = -0.251, θ_3 = -0.188, θ_4 = -0.146, θ_5 = -0.472, θ_6 = 0.088, θ_7 = -0.667, θ_8 = -0.199, θ_9 = -0.233, θ_{10} = -0.028, θ_{11} = 0.207, d = 0.253)

All analysis was carried out using R software packages (R Core Team, 2014a), namely R `forecast` package (Hyndman, 2015), R `fracdiff` package (Fraley, Leisch and Maechler, 2012) and R `stats` package (R Core Team, 2014b). We let $\{Y_t\}$ be the monthly tourist arrivals to Langkawi Island. Using function ets() and auto.arima(), the best model from each method were automatically selected based on the lowest Akaike Information Criterion (AIC) values. However, for ARFIMA model, the initial values for p and q were estimated using computer software ITSM2000.

Based on the three models, a two-year full cycle from January 2013 until December 2014 of tourist arrivals were predicted. The actual values and the predicted values together with the 80% and 95% confidence interval for Model 1, 2 and 3 were plotted in Figure 2, 3 and 4 respectively. In addition, a plot of all the predicted values for all the three models is presented in Figure 5. Clearly from Figure 5, both Model 1 and Model 2 seem to be underestimating the tourist arrival in Langkawi Island. However, predicted values obtained from Model 3 showa better estimation since the predicted line is much closer to the actual data as compared to Model 1 and Model 2.

Further investigation was carried out in order to evaluate the performance of each model and decide on the best model that suits the data. This study chose four criteria of accuracy measures which are the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE), and acceptable output percentage (**Z**). In this study, the reference point for evaluation outcome, **Z** was set for ±15%. Each of these accuracy measures is defined respectively by the following formula,

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n},\ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}},$$

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|}{n} \times 100\%,$$

$$Z = \frac{\sum_{i=1}^{n} j}{n} \times 100\% \ for \begin{cases} j = 1 \ if \ \frac{|y_i - \hat{y}_i|}{y_i} \ldots \le 0.15, \\ j = 0 \qquad otherwise \end{cases}$$

where y_i and $\hat{y}_i$ respectively are the actual observed values and the predicted values for the period of January 2013 until December 2014.

The summary of all the four accuracy measures used in this study is presented in Table 2. Based on the values presented in the summary table, we can see clearly that it supported our investigation earlier through the graphical display. Model 3 is concluded to be the appropriate forecast model for modelling tourist arrivals in Langkawi Island since it has the lowest error measure for MAE, RMSE, MAPE and the largest value for the **Z** output score. Thus, using Model 3, we forecast the monthly tourist arrivals to Langkawi Island for the year 2015. The forecasted values together with the 80% and 95% confidence level is presented in Table 3 while the plot of these values is displayed in Figure 6. Clearly from both the Table 3 and Figure 6, the number of tourist arrivals to Langkawi Island keeps increasing in the long run for the year 2015 with a total number is expected to exceed 3.9 million tourists, a growth of approximately 10.2% from the previous year.

5. Conclusion

The purpose of this paper was to compare the performance of modelling Langkawi Island tourist arrivals using exponential smoothing, ARIMA and ARFIMA model. Finding the best model for each class under study has been much easier with the automatic model builder in R package used in this paper. Based on the result obtained, we found that the ARFIMA model is the most reasonable and the most appropriate for modelling tourist arrivals to Langkawi Island. The conclusion of the best model was made based on the lowest MAE, RMSE, and MAPE measures and the largest **Z** percentage value. The approximate 95% confidence interval for the forecast value at the end of year 2015 range between as low as 345 thousand to as high as 622 thousand with a forecast value of approximately 484 thousand. Hence, with high accuracy in forecasting the tourist arrivals, tourism related industries and also the local authority like LADA can strategically plan the right action to be taken up in order to cater the increasing number of tourists visiting Langkawi Island, provided that there is no drastic change in policy by the government or any other events that may directly impact the tourism industry. For further research, multivariate analysis can be done by including other explanatory variables in order to understand what are other influences that reflect the increasing number of tourists in Langkawi Island. Specific tourist attraction areas or places around Langkawi Island could also be studied for improvement purposes to attract more tourists.

6. References

[1] Arsham, H. (1994). Time-Critical Decision Making for Business Administration. Retrieved from http://home.ubalt.edu/ntsbarsh/business-stat/stat-data/forecast.htm

[2] Bowerman, B.L. & O'Connell, R.T. (1999). *Time Series Forecasting: Unified Concepts and Computer Implementation*. Duxbury Press, Boston.

[3] Bowerman, B.L., O'Connell, R.T., & Koehler, A.B. (2005). *Forecasting, Time Series, and Regression: An Applied Approach*. (4th ed.). Thomson Brook/Cole, Belmont CA.

[4] Department of Statistics, Malaysia (2013), *Domestic Tourism Survey 2012*, Putrajaya.

[5] Fraley, C., Leisch, F., & Maechler, M. (2012). `fracdiff`: Fractionally differenced ARIMA aka ARFIMA(p,d,q) models. R package version 1.4-2, URL: http://CRAN.R-project.org/package=fracdiff

[6] Hyndman, R.J. & Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software 27*(3). pp. 1-22.

[7] Hyndman, R.J. (2015). `forecast`: Forecasting Functions for Time Series. R package version 5.9, URL: http://github.com/robjhyndman/forecast

[8] Kulendran, N. & Witt, S. F. (2001). Cointegration versus least squares regression.*Annals of Tourism Research*, Vol. 28, No.2, pp.291-311.

[9] Langkawi Development Authority (2015). Tourist Statistics. Retrieved from http://www.lada.gov.my/

[10] Lumsdon, L. (1997). *Tourism Marketing*. International Thomson Business Press, London.

[11] Malaysia Tourism Promotion Board (2014). Malaysia Tourist Arrivals By Country of Nationality December 2014. Retrieved from http://corporate.tourism.gov.my/images/research/pdf/2014/arrival/Tourist_Arrivals_Dec_2014.pdf

[12] Marzuki, A. (2011). Residents Attitudes Towards Impacts from Tourism Development in Langkawi Islands, Malaysia. *World Applied Sciences Journal 12 (Special Issue of Tourism & Hospitality)*. pp.25-34.

[13] Nanthakumar, L., Subramaniam, T., & Kogid, M. (2012). Is 'Malaysia Truly Asia'? Forecasting tourism demand from ASEAN using SARIMA approach. *Tourismos, 7* (1). pp. 367-381.

[14] R Core Team (2014a). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[15] R Core Team (2014b). `stats`: The R Stats Package. R package version 3.1.2, URL http://www.R-project.org/

[16] Shitan, M. (2008). Time series modelling of tourist arrivals to Malaysia. *InterStat* (October). pp. 1-12.

[17] World Tourism Organization (2014). Glossary of tourism terms. Retrieved from https://s3-eu-west-1.amazonaws.com/staticunwto/Statistics/Glossary+of+terms.pdf

[18] World Travel and Tourism Council (2014). Travel and Tourism: Economic Impact 2014, Malaysia. Retrieved from http://www.wttc.org/

Tables

Table 1
The fifteen exponential smoothing methods

Trend Component	Seasonal Component		
	N (None)	A (Additive)	M (Multiplicative)
N (None)	N,N	N,A	N,M
A (Additive)	A,N	A,A	A,M
A_d (Additive damped)	A_d,N	A_d,A	A_d,M
M (Multiplicative)	M,N	M,A	M,M
M_d (Multiplicative damped)	M_dN	M_dA	M_dM

Source.Hyndman and Khandakar (2008)

Table 2
MAE, RMSE, MAPE and Z values

	MAE	RMSE	MAPE	Z
Model 1 ETS(A,N,A)	38228.8	47877.7	12.57%	66.7%
Model 2 ARIMA(1,0,2)$(0,1,1)_{12}$	33177.3	37134.8	11.47%	33.3%
Model 3 ARFIMA(12,0.25,11)	**29401.5**	**36833.8**	**10.93%**	**75.0%**

Table 3
Forecast of Tourist Arrivals to Langkawi Island from January 2015 to December 2015 using ARFIMA(12,0.25,11)

Month	Forecasted Value	80% Confidence Interval	95% Confidence Interval
January	343208.1	(273762.4, 412653.8)	(237000.1, 449416.1)
February	270382.2	(200812.3, 339952.1)	(163984.2, 376780.2)
March	259523.3	(189290.5, 329756.2)	(152111.5, 366935.2)
April	244793.6	(172602.2, 316985)	(134386.4, 355200.8)
May	298408.3	(224118.9, 372697.7)	(184792.5, 412024.2)
Jun	373135.8	(296622.1, 449649.5)	(256118.2, 490153.4)
July	318888.6	(238135, 399642.2)	(195386.7, 442390.6)
August	298237.1	(214610.9, 381863.3)	(170341.9, 426132.3)
September	301907.6	(215548.6, 388266.6)	(169832.9, 433982.3)
October	324198.3	(236047.6, 412349)	(189383.4, 459013.2)
November	421945.5	(332845.5, 511045.5)	(285678.8, 558212.1)
December	483736.6	(393256.2, 574217)	(345358.8, 622114.4)

Figures

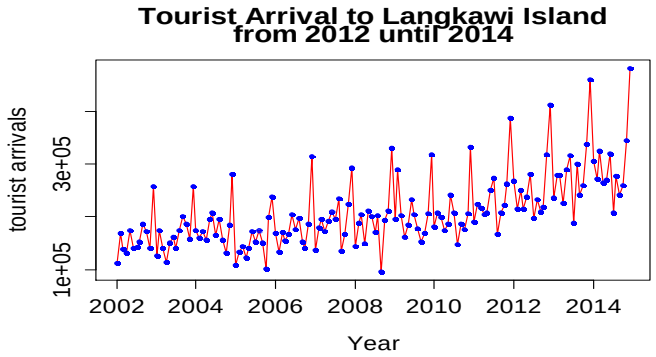

Figure 1. The time series plot of the monthly tourist arrivals to Langkawi Island data from January 2002 to December 2014

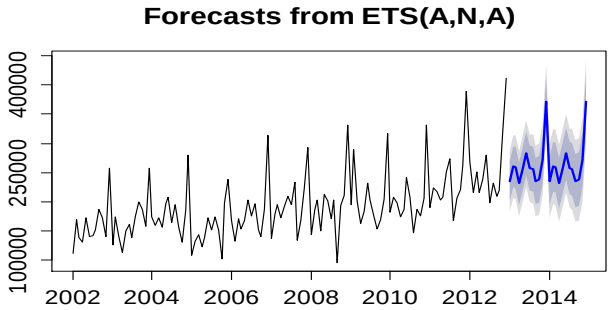

Figure 2. Plot of predicted values and actual observed values using Model 1

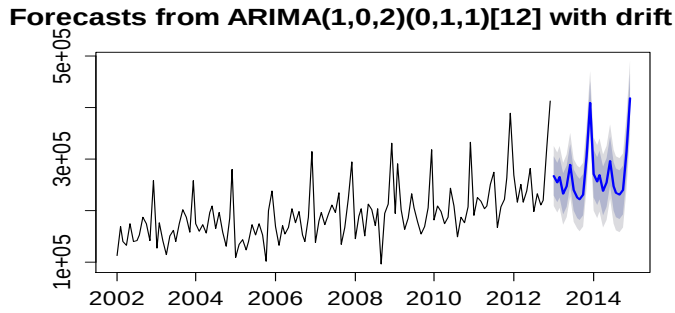

Figure 3. Plot of predicted values and actual observed values using Model 2

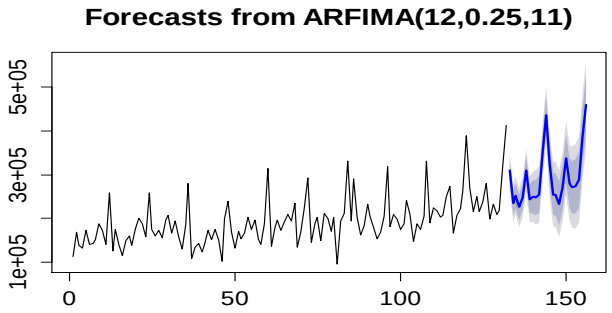

Figure 4. Plot of predicted values and actual observed values using Model 3

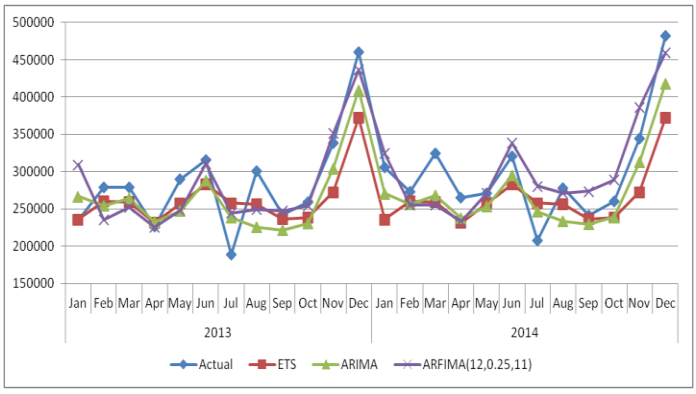

Figure 5. Plot of predicted values and actual observed values for all three models.

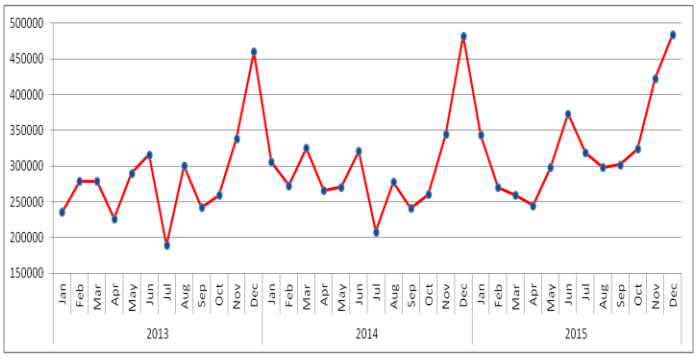

Figure 6. Forecast for January 2015 to December 2015 using ARFIMA(12,0.25,11)