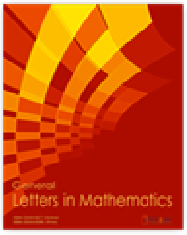
Available online at www.refaad.com
Gen. Lett. Math., 7(2)2019, 100-107
Research Article

General Letters in Mathematics (GLM)

Journal Homepage: http://www.refaad.com/views/GLM/home.aspx
ISSN: 2519-9277 (Online) 2519-9269 (Print)



Forecasting international tourist arrivals in zanzibar using box – jenkins SARIMA model

Zulkifr Abdallah Msofe [a*] **, Maurice Chakusaga Mbago** [b]

[a,b] Department of Statistics, University of Dar es Salaam, Tanzania
Emails :[a] zully.abdallah2010@gmail.com , [b] mmbago49@gmail.com

Abstract

The arrival of international tourists contributes to the generation of foreign currencies and creates employment opportunities to the local people. Modelling and forecasting tourist arrivals plays a major role in tourism planning and marketing and therefore crucial for policy decision-making towards sustainable tourism development. In this paper an attempt has been made to forecast international tourist arrivals in Zanzibar using Seasonal Autoregressive Integrated Moving Average (SARIMA) model. Data from January 1995 to December 2017 covering 276 observations were used. The SARIMA $(1, 1, 1) \times (1, 1, 2)_{12}$ model was found to be the best fitted model on the basis of Akaike`s Information Criterion (AIC). The adequacy of the fitted model was confirmed by Ljung-Box test statistic and the model was used to generate monthly forecasts from January 2018 to December 2019 with 95% confidence interval. The forecasting performances of candidate models were evaluated on the basis of Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). The forecasts indicate that the number of tourists visiting Zanzibar is likely to keep on increasing with seasonal pattern similar to that of the original data.

Keywords: *Modelling, Forecasting, Seasonal Autoregressive Integrated Moving Average (SARIMA), International Tourist Arrivals.*

1. Introduction

Tourism has been one of the rapidly growing and important economic sectors across the world. The sector contributes largely to the foreign exchange earnings, provides employment and investment opportunities to the destination countries due to global increase in the number of tourist arrivals [7]. Zanzibar, the semi-autonomous island within the United Republic of Tanzania and one of the Indian Ocean islands, has numerous tourist attractions which are nature-based such as tropical beaches, forest reserves, palm trees, historical and cultural sites and the historical stone town. In 2000, UNESCO declared Zanzibar's stone town the world heritage site due to diverse of the race on its architecture and tradition [11]. The island has been attracting a great number of tourists from various areas of the world such as Europe, Asia, America, Africa, Oceania and few of them from emerging markets such as Russia, Poland, China, India and Israel [13]. Tourist arrivals being vital to the economic growth, has made modelling and forecasting tourist arrivals to gain huge consideration among researchers and policy makers globally.

[*] Corresponding author

Email address: zully.abdallah2010@gmail.com (Zulkifr Abdallah Msofe)
doi:10.31559/glm2019.7.2.6
Received Date: *11 Nov 2019*
Accepted Date: *27 Dec 2019*

This paper reviews few recent studies which applied different time series models in forecasting different variables including tourist arrivals. A study carried out in India applied Holt-Winters Exponential Smoothing (HWES) and ARIMA models in forecasting foreign tourist arrivals [12]. Another study carried out in Thailand attempted to forecast international visitor arrivals by using different time series methods [6]. Also the use of SARIMA and GARCH models in forecasting tourist arrivals was applied in Sri-Lanka [10], while [2] applied ARIMA model and Double exponential smoothing to forecast tourist arrivals in Kenya. [9] Applied ARIMA model in an attempt to forecast wholesale price of maize in Tanzania. [1] Used hybrid method and ARIMA model in forecasting annual data of rain precipitation in the Province of Erbil-Iraq.

In this paper an attempt has been made to model and forecast international tourist arrivals in Zanzibar using Seasonal Autoregressive Integrated Moving Average (SARIMA) model suggested by Box-Jenkins.

2. Material and Methods

This study used time series data of international tourist arrivals in Zanzibar from January 1995 to December 2017 covering 276 observations and data were obtained from the Office of Chief Government Statistician Zanzibar (OCGS). Data in the training set (from January 1995 to December 2016) were used for model fitting while the remaining data in the test set (from January 2017 to December 2017) were used for validation purposes. Box-Jenkins procedures were used to obtain an appropriate SARIMA model for forecasting international tourist arrivals in Zanzibar. Data analysis was carried out using R version 3.5.1 and Microsoft-Excel 2013.

2.1. SARIMA Model

The SARIMA model combines both seasonal and non-seasonal components. The model is abbreviated as SARIMA$(p,d,q) \times (P,D,Q)_s$ and can be written as:

$$\varphi_p(Z)\Phi_P(Z^s)\nabla^d\nabla_s^D X_t = \theta_q(Z)\Theta_Q(Z^s)\mathcal{E}_t \qquad (1)$$

Model (3) can further be rewritten as:

$$\varphi_p(Z)\Phi_P(Z^s)W_t = \theta_q(Z)\Theta_Q(Z^s)\mathcal{E}_t \qquad (2)$$

where: $W_t = \nabla^d\nabla_s^D X_t$

The seasonal components are:

$$\mathrm{AR}(P)\colon \Phi_P(Z)^s = 1 - \Phi_1 Z^s - \Phi_2 Z^{2s} - \cdots\cdots\cdots - \Phi_P Z^{sP}$$

$$\mathrm{MA}(Q)\colon \Theta_Q(Z)^s = 1 - \Theta_1 Z^s - \Theta_2 Z^{2s} - \cdots\cdots\cdots - \Theta_Q Z^{sQ}$$
$$\nabla_s^D = (1 - Z^s)^D$$

The non-seasonal components are:

$$\mathrm{AR}(p)\colon \varphi_p(Z) = 1 - \varphi_1 Z - \varphi_2 Z^2 \ldots\ldots\ldots\ldots\ldots - \varphi_p Z^p$$
$$\mathrm{MA}(q)\colon \theta_q(Z) = 1 - \theta_1 Z - \theta_2 Z^2 \ldots\ldots\ldots\ldots\ldots - \theta_q Z^q$$
$$\nabla^d = (1 - Z)^d$$

where Z is the backshift operator such that

$$Z^j X_t = X_{t-j},\ Z^j\mathcal{E}_t = \mathcal{E}_{t-j},\quad j=0,1,2,\ldots\ldots$$

p= non-seasonal AR order, d= degree (order) of non-seasonal differencing, q= non-seasonal MA order, P= seasonal AR order, D= degree (order) of seasonal differencing, Q= seasonal MA order, s= the number of seasons per year, $\mathcal{E}_t$ = the white noise.

2.2. Stationarity of the Time Series

SARIMA models are defined for stationary time series, thus there was a need to check whether the data are stationary or not [3]. Time plot and Augmented Dickey Fuller (ADF) test were used to test for stationarity of the series.

2.2.1. Differencing Technique

The differencing technique with both seasonal and non-seasonal differencing were used to transform the series from non-stationary to stationary. Seasonal differencing of the first order was employed to remove seasonality in the given time series data while non-seasonal differencing was employed to get rid of the trend. Seasonal and non-seasonal differencing of the first order can be expressed as given in equations (1) and (2) respectively.

$$(1-Z^{12})X_t = X_t - X_{t-12} \tag{3}$$
$$(1-Z)X_t = X_t - X_{t-1} \tag{4}$$

where Z is the backshift operator such that

$$Z^j X_t = X_{t-j}\ ,\ \text{j=0, 1, 2, ………….}$$

2.3. Box-Jenkins Procedure

Box and Jenkins model building procedure which involved four steps was employed in order to find the best model to fit within the class of SARIMA models.

2.3.1. Step 1: Model Identification

This step involved the tentative identification of the model order, that is identifying the values of p, P, q, Q, d, and D. According to [4], model identification in seasonal time series is very difficult and normally trial and error is used. Usually the values of p, P, q, Q, d, and D selected should be less than or equal to two. Plots of ACF and PACF were used for this purpose.

2.3.2. Step 2: Model estimation and Selection

This step involved the estimation model parameters identified in the first step. The method of maximum likelihood was used in this case. The model with the lowest Akaike Information Criterion (AIC) was selected as the best model to fit among the candidate SARIMA models. The Akaike information criterion (AIC) is defined by the formula:

$$AIC(k) = T\log(\hat{\sigma}^2(k)) + 2k \tag{5}$$

where k is the number of parameters estimated and T is the number of observations.

2.3.3. Step 3: Diagnostic checking

This involved checking whether the fitted model has adequately captured the information in the time series data. The Ljung-Box test statistic defined in equation (6) was used. Ljung-Box statistic tests the null hypothesis that residuals are white noise. We reject null hypothesis that the residuals are white noise if $Q > x^2_{\alpha,m-k}$ where α is the level of significance [8].

$$Q = n(n+2)\sum_{i=1}^{m}\frac{r_i^2}{n-i} \sim x^2_{m-k} \tag{6}$$

where:
k = the number of parameters estimated
m = the lag-length chosen in the range 15-30
r_i = the sample residual autocorrelation at lag i
n = Number of observations after any differencing

2.3.4. Step 4: Forecasting

This involved predicting future number of international tourist arrivals using the fitted model selected. In this case two years ahead monthly forecasts were generated. That is the forecasts from January 2018 to December 2019.

2.4. Model Validation and Performance Evaluation

One of the most important tests of any model is how well it forecasts. The comparison of the forecasted values and observed values was made in the validation period to see how close the two values were. To understand the magnitude of the forecast error, the statistics which measure the performance of forecast by using techniques of minimizing forecast errors were used. The Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) were employed in this case to evaluate the performances of the forecasting model. A good model should have small RMSE and MAPE value which is close to zero [8].
RMSE is defined by the formula

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t - F_t)^2} \tag{7}$$

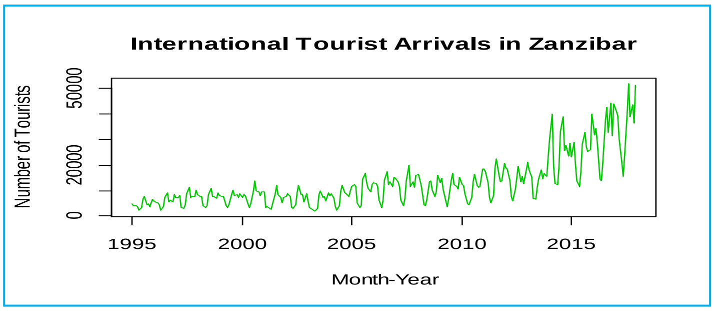
while .MAPE is defined by the formula

$$MAPE = \frac{100}{n}\sum_{i=0}^{n}\left|\frac{X_t - F_t}{X_t}\right| \tag{8}$$

where: X_t is the monthly observed number of international tourists, F_t is the forecasted number of international tourists in the corresponding months,n is the number of observations in the validation period.

3. Empirical Results

The time plot in Fig 1 shows a general upward trend over a specified period of time. This upward trend which is in fact the consistently long term increasing in the number of international tourist arrivals may be attributed partly by the efforts that have been taken by the revolutionary government of Zanzibar and stakeholders in tourism industry to improve tourism sector in the Island. A strong seasonal variations with some peaks and troughs that appear regularly every year are also indicated.



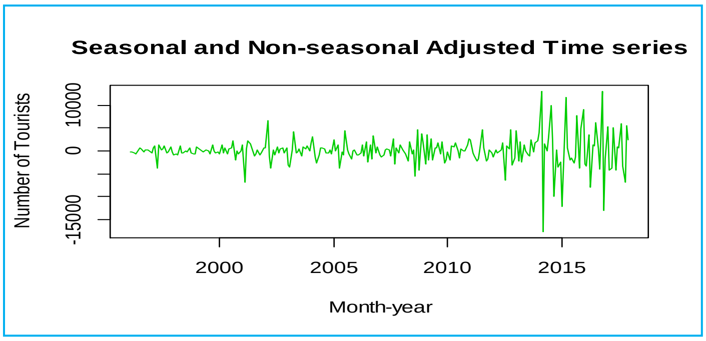
Fig(1): Time Plot of International Tourist Arrivals in Zanzibar

3.1. Stationarity of the Time series

An upward trend and seasonality indicated in Fig 1 implies that the time series is non-stationary. The ADF test which was used to supplement the graphical methods, had a p-value of 0.4317 at lag order 6 which is not less than α=0.05, indicating that the series is non-stationary. Both seasonal and non-seasonal difference of first order were made to transform the series from non-stationary to stationary. The ADF test, after differencing, had the p-value of 0.01 which is less than α=0.05 indicating a stationary series as shown in Table 1. Fig 2 is the time plot after first order difference which shows that the trend and seasonal variation are greatly reduced if not eliminated at all.

Table(1): ADF test for Original and Differenced series

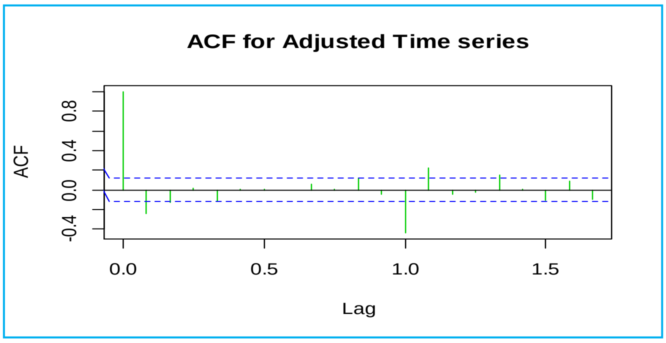
ADF Test		
	Original series	After 1st order Differencing
Lag order	6	6
ADF Test Statistic	-2.3427	-8.4487
Significance level (α)	5%	5%
P-Value	0.4317	0.01
Alternative hypothesis	Stationary	



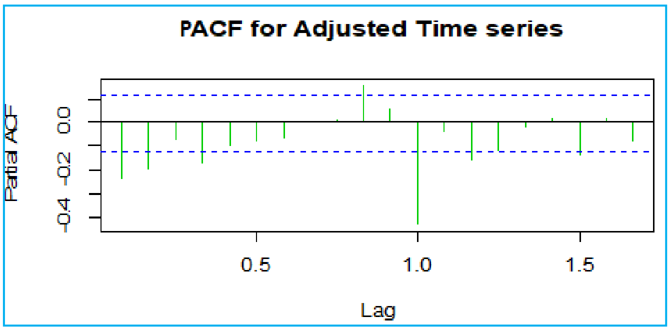
Fig(2): Time plot for differenced Series

3.2. Model Identification

The plots of ACF and PACF of the stationary time series were examined in order to identify the values of p, q, P and Q. The values of d and D which are the number of non-seasonal and seasonal differencing respectively are both equal to one. That is d=1 and D=1 transformed the time series from non-stationary to stationary.



Fig(3): The Plot of ACF for the Differenced series



Fig(4): The Plot of PACF for the Differenced series

In Fig 3 ACF has significant spike at non-seasonal lag 1 suggesting the possible non-seasonal moving average of the first order, MA (1) to be included in the model. Also ACF has significant spike at seasonal first seasonal lag (at 1.0) suggesting seasonal moving average of the first order SMA (1). The significant spike of PACF at first lag (at 0.0833) and at the second lag (at 0.1667) in Fig 4 suggests possible Autoregressive terms of the respective order which are AR (1) and AR (2). Also PACF has significant spike at first seasonal lag (at 1.0) suggesting possible seasonal Autoregressive terms which are SAR (1).

3.3. Model Estimation and Selection

The *SARIMA (1,1,1)×(1,1,2)*$_{12}$ model with the lowest Akaike Information Criteria (*AIC=4637.752*) was selected as the best model amongst the candidate SARIMA models. The selected model also perform best in term of forecasting with the lowest RMSE and MAPE compared to other candidate SARIMA models as shown in Table 2. The coefficients of the best model estimated by using the method of Maximum Likelihood (ML) and all the coefficients are statistically significant at 5% as shown in Table 3.

Table(2): AIC, RMSE and MAPE of Candidate SARIMA models

S/N	Candidate Model	AIC	RMSE	MAPE
1	SARIMA(1,1,1)×(2,1,2)$_{12}$	4639.005	2320.503	12.30974
2	SARIMA(1,1,1)×(1,1,2)$_{12}$	4637.752	2313.578	12.14104
3	SARIMA(1,1,1)×(0,1,1)$_{12}$	4640.555	2369.73	12.24227
5	SARIMA(1,1,1)×(1,1,1)$_{12}$	4640.785	2360	12.33368
4	SARIMA(1,1,2)×(0,1,2)$_{12}$	4639.31	2355.391	12.46675

Lowest AIC=4637.752
Best Model: SARIMA(1,1,1)×(1,1,2)$_{12}$

Table(3): Coefficients and Standard Error (S.E) of the Best model

Best fitted model: SARIMA(1,1,1)×(1,1,2)$_{12}$

	AR(1)	MA(1)	SAR(1)	SMA(1	SMA(2)
Coefficients	0.4579*	-0.8493*	0.8828*	-1.6088*	0.6931*
S.E	0.0995	0.0662	0.0877	0.1013	0.079
t-value	4.602	-12.829	10.0661	-15.882	8.7734

log-likelihood=-2312.7, AIC=4637.752
*Means statistically significance at 5%

The best model, that is, SARIMA (1,1,1)×(1,1,2)$_{12}$ model can explicitly be written as:

$$(1-0.4579Z)(1-0.8828Z^{12})W_t = (1+0.8493Z)(1+1.6088Z^{12}-0.6931Z^{24})\mathcal{E}_t$$

where: $W_t = \nabla^1\nabla_{12}^1 X_t = X_t - X_{t-1} - X_{t-12} + X_{t-13}$

X_t is the number of international tourist arrivals in month t
$\mathcal{E}_t$ is the white noise
Z *is* the backshift operator such that:

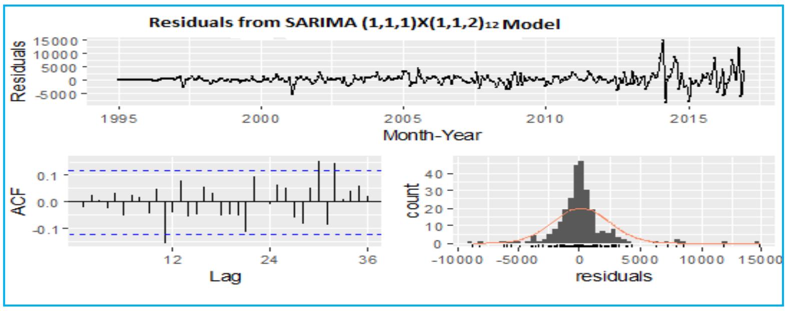
$$Z^j X_t = X_{t-j}, \quad Z^j \mathcal{E}_t = \mathcal{E}_{t-j}, \quad \text{j=0,1,2,………}$$

3.4. Diagnostic Checking

This involved checking whether the fitted model has adequately captured the information in the data. Residual analysis which involves graphical procedures and the Ljung-Box statistical test were used as shown in Table 4 and Fig 5 respectively. The residuals from the fitted model shown in Fig 5 seem to be random as they have nearly constant variance and zero mean indicating that the fitted model is adequate. The Ljung-Box statistical value, Q*=24.128 is insignificant because the p-value shown in Table 4.9 is 0.1913 which is greater than 0.05 level of significance suggesting that the residuals of the best model are not statistically significantly distinguishable from white noise. That means the model SARIMA (1, 1, 1) × (1, 1, 2)$_{12}$ is adequate.

Table(4): Ljung-Box test

Ljung-Box Test
Data: Residuals from Best model: SARIMA(1,1,1)×(1,1,2)$_{12}$
Q*=24.128, df=19, p-value=0.1913
Model df:5, Total lags used: 24



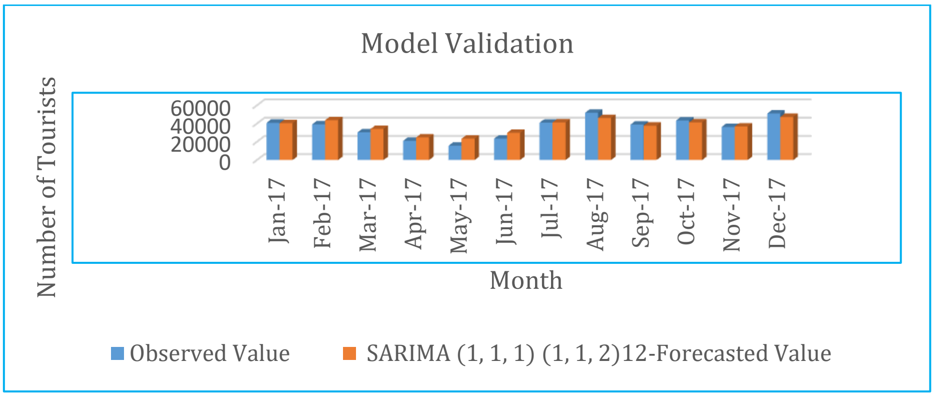
Fig(5): Residual Analysis from SARIMA (1, 1, 1) × (1, 1, 2)$_{12}$ Model

3.5. Model Validation and Performance Evaluation

The forecasted and observed number of international tourist arrivals were compared in the validation period that is from January 2017 to December 2017 as it can be seen in Table 4 and Fig 6. The forecasting performances of SARIMA (1, 1, 1) × (1, 1, 2)$_{12}$ model was evaluated on the basis of RMSE and MAPE during the validation period as shown in Table 4.

Table(5): Comparison of Forecasted and Observed Number of International Tourist Arrivals in the Validation Period

Month-Year	Observed Value	Forecasted Value
17-Jan	40938	40500.8
17-Feb	39119	43763.2
17-Mar	30366	34136.65
17-Apr	21004	24743.21
17-May	15696	23442.99
17-Jun	23458	29864.2
17-Jul	41034	41299.06
17-Aug	51937	46186.49
17-Sep	38977	37750.39
17-Oct	43470	41355.36
17-Nov	36363	36931.02
17-Dec	51112	47322.74
	MAPE	12.14104
	RMSE	2313.578



Fig(6): Graphical Comparison of Forecasted and Observed Values in the Validation Period

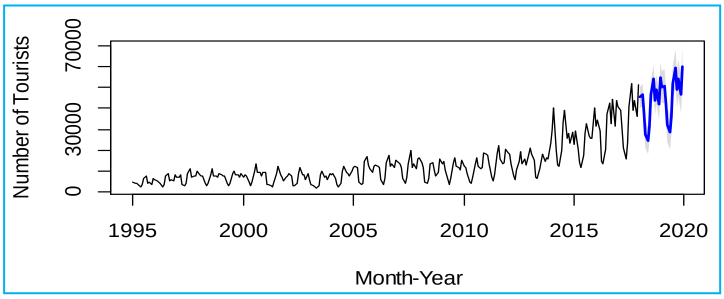
In Table 5, the values of RMSE and MAPE are 2313.578 and 12.37756 respectively. The forecasted value from the SARIMA (1, 1, 1) × (1, 1, 2)$_{12}$ model are somewhat close to the observed values as it can be seen in Fig 6.

3.6. Forecasting using SARIMA (1, 1, 1) × (1, 1, 2)12 model

The best fitted model which is SARIMA (1, 1, 1) × (1, 1, 2)$_{12}$ was used to generate forecasts with 95% level of confidence. Twenty four monthly forecasts that is from January 2018 to December 2019 were produced. Based on results shown in Table 6 as well as its plot in Fig 7, the number of international tourists visiting Zanzibar is expected to continue increasing with similar seasonal pattern as that of the original time series.

Table(6): Forecasts of International Tourist Arrivals in Zanzibar

Month-Year	Forecasts	95% Confidence Intervals	
		Lower Limit	**Upper Limit**
Jan-18	44457.00	36834.58	52079.43
Feb-18	48646.54	40742.79	56550.30
Mar-18	38654.33	30504.89	46773.77
Apr-18	28842.18	20535.59	37148.77
May-18	27598.50	19118.76	36078.25
Jun-18	34429.44	25784.34	43074.53
Jul-18	46530.87	37725.46	55336.28
Aug-18	51062.34	42640.35	60564.33
Sep-18	42782.56	33667.07	51898.06
Oct-18	47036.93	37770.64	56303.23
Nov-18	41977.84	32563.23	51392.44
Dec-18	53007.27	43446.67	62567.87
Jan-19	49709.85	39628.60	59791.10
Feb-19	54319.51	43909.13	64729.90
Mar-19	43822.12	33155.87	54488.36
Apr-19	33556.83	22666.87	44446.87
May-19	32324.83	21226.77	43422.90
Jun-19	39499.95	28202.64	50797.26
Jul-19	52181.79	40690.91	63672.67
Aug-19	57412.06	45731.80	69092.32
Sep-19	48251.91	36385.73	60118.08
Oct-19	53078.78	41029.76	65127.81
Nov-19	47458.99	35229.92	59688.05
Dec-19	59051.21	46644.73	71457.69



Fig(7): Time Plot for Original and Forecasted Values

4. Conclusion

Tourism is currently one of the important sectors for economic growth in Zanzibar. Therefore, to predict future number of international tourists visiting the Island is crucial for tourism planning and marketing. In this paper Box-Jenkins procedure was applied to find appropriate SARIMA model to forecast the number of international tourist arrivals in Zanzibar using data from January 1995 to December 2017. The results show that the SARIMA (1, 1, 1) × $(1, 1, 2)_{12}$ is the best fitted model and the model was used to generate monthly forecasts from January 2018 to December 2019 with 95% confidence interval. The forecasts indicate that the number of tourists visiting Zanzibar is likely to keep on increasing with seasonal pattern similar to that of the original data.

5. Recommendations and Policy Implications

Based on the results from this paper, the following policy recommendations have been suggested.

The arrival of international tourists is characterized by upward or increasing trend but at a slow rate as indicated in Fig 1. Thus the Zanzibar Commission for Tourism (ZCT) and stakeholders in Zanzibar's tourism industry should continuously keep on upgrading the quality of tourism services and products as well as enhancing tourism marketing strategies in order to attract more international tourists from different parts of the world.

The seasonal variations was also shown in this study, thus necessary measures should be taken in order to deal with the issue of decreasing number of international tourists and tourism revenue during low season. For instance the government should encourage domestic tourism by cutting costs for tourism services and products during low seasons.

References:

[1] Q. Abdulqader, *Annual Forecasting Using a Hybrid Approach*, General Letters in Mathematics, 4(2)(2018), 86–95 , doi:10.31559/glm2018.4.2.5

[2] A. O. Akuno, M. O. Otieno, C. W. Mwangi, & L. A. Bichanga, *Statistical Models for Forecasting Tourists' Arrival in Kenya,* Open Journal of Statistics, 5(1)(2015), 60-65, https://doi.org/10.4236/ojs.2015.51008

[3] G. E. P. Box & G. M. Jenkins, *Times series Analysis Forecasting and Control*. Holden-Day San Francisco,(1970).

[4] G. E. Box & G. M. Jenkins, *Time series analysis: forecasting and control*, revised ed. Holden-Day,(1976), https://doi.org/10.1177/058310248201400608

[5] T. Chai & R. R. Draxler, *Root mean square error (RMSE) or mean absolute error (MAE)?,* Geoscientific Model Development Discussions, 7(1)(2014),1525-1534, https://doi.org/10.5194/gmdd-7-1525-2014

[6] P. Chaitip, C. Chaiboonsri, & R. Mukhjang, *Time Series Models for Forecasting International Visitor Arrivals to Thailand*, Internaonal Conference on Applied Economics-ICAE 2008. (2010), 159-163.

[7] C. Chang, T. Khamkaew, R. Tansuchat & M. McAleer, *Independence of International Tourism demand and Volatility in leading ASEAN Destinations*, Tourism Economics, 17(3)(2011),481-507, https://doi.org/10.5367/te.2011.0046

[8] J. M. Dufour & R. Roy, *Generalized portmanteau statistics and tests of randomness*, Communications in Statistics-Theory and Methods, 15(10)(1986),2953-2972, https://doi.org/10.1080/03610928608829288

[9] S. Kibona & M. Mbago, *Forecasting Wholesale Prices of Maize in Tanzania Using Arima Model*, General Letters in Mathematics, 4(3)(2018), 131–141 , doi:10.31559/glm2018.4.3.6

[10] J. H. Priyangika, Pallawala, & D. J. C Sooriyaarachchi, *Modelling and forecasting tourist arrivals in Sri Lanka*, Symposium on Statistical & Computational Modelling with Applications, (2016), 14-18.

[11] E. Rotarou, *Tourism in Zanzibar-Challenges for Pro-poor Growth, (Unpublished paper),* University of Chile. Santiago, Chile, (2014), 250-264.

[12] S. Sood, & K. Jain, *Comparative Analysis of Techniques for Forecasting Tourists' Arrival. India*, Journal of Tourism & Hospitality, 6(3)(2017), 3–6, https://doi.org/10.4172/2167-0269.1000285

[13] WTTC, *Travel & Tourism: Economic Impact 2017, Tanzania*, 1–24. https://www.wttc.org/-/media/files/reports/economic-impact-research/countries-2017/tanzania2017.pdf