Time series modeling of tourist accommodation demand in Kenya

George Otieno[1*] Joseph Mung'atu[1] George Orwa[1]

1 Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, P. O. Box 62000 - 00200, Nairobi, Kenya

* E-mail of the corresponding author: otieno.george@gmail.com

Abstract

Tourism is a very important sector in the world economy and contributes significantly to foreign exchange earnings. Earnings from tourism in Kenya increased annually from Kenya Shillings 24.3 billion in 2001 to 73.7 billion in 2010 (ROK, 2012). The number of tourists coming into the country increased from 1,146,102 in the year 2003 to 1,822,885 in the year 2011. The major tourist zones in Kenya are: Nairobi, Beach, Mombasa, Coast Hinterland, Maasailand, Nyanza basin, Western, Central and North (ROK, 2012). These can further be reduced to three: Nairobi, Coastal and Others. Tourism in Kenya relies on many other sectors and industries, one of which is the hotel and accommodation. In order to enable these related industries match the specific accommodation needs for the tourists arriving in the country, there is needed a model that can forecast the accommodation demands by the tourists. This will make it possible for the hotel industry players to respond in good time to the anticipated changes in demand over time and also to maximize returns on investments.

Seasonal variations are important in tourism and hospitality demands. The Box-Jenkins models for time series analysis allow the analyst to forecast future values of a series with only the past period's data, without having some related variable's data (Shumway and Stoffer, 2011). The authors therefore focused on the Box-Jenkins models to generate a forecasting model using quarterly data on bed occupancy by tourists visiting Kenya from 1974 to 2011. The SARIMA (1,1,2)(1,1,1)[4] model was found to be suitable for forecasting future quarterly demand on tourist accommodation in Kenya. This model shall therefore be useful to the tourism and related industries in forecasting future demands and maximize their returns on investment.

Keywords: Tourist Accommodation Demand, Kenya Tourism Accommodation, Tourism Accommodation Modeling

1. Introduction

1.1. Background of the study

Tourism is a very important sector in the world economy and contributes significantly to foreign exchange earnings. In 2012, international tourism generated a total of US $ 1.3 trillion globally or US$ 3.5 billion a day on average (United Nations World Tourism Organization, 2013). Tourism is not an independent sector but very much dependent on others such as transport and hotel, among others.

Kenya's migration statistics are collected under the Immigration Act 1967 (ROK, 2012). In 2011, a total of 1,822,885 tourists arrived in Kenya (ROK, 2012). Earnings from tourism in Kenya increased annually from

Kenya Shillings 24.3 billion in 2001 to 73.7 billion in 2010. The number of tourists coming into the country increased from 1,146,102 in the year 2003 to 1,822,885 in the year 2011. In the year 2011, the majority of visitors came from Europe zone (930,527), representing 51% of all arrivals. In quarter 4 of 2011, tourists from Europe had the highest average length of stay (25.7 days) compared to the other zones. The average for tourists from Asia was 8.5 days whereas from Other Africa, it was at 5.5 days.

Figure 1: Administrative Map of Kenya

Source: http://www.nationsonline.org/oneworld/map/kenya_map.htm

The major tourist zones in Kenya are: Nairobi, Beach, Mombasa, Coast Hinterland, Maasailand, Nyanza basin, Western, Central and North (ROK, 2012). These can further be reduced to three: Nairobi, Coastal and Others. An administrative map of Kenya is shown in ***Figure 1.***

Tourism in Kenya relies on many other sectors and industries, one of which is the hotel and accommodation.

The fluctuating number of tourists visiting the country affects the hotel industry, a key player in tourism. There have been variations in the number of visitors from different regions of the world as well as type of visit, length of stay and type of accommodation over time. The hotel industry has not been able to reliably predict the number of visitors requiring accommodation in Kenya.

In order to enable these related industries match the specific accommodation needs for the tourists arriving in the country, there is needed a model that can forecast the accommodation demands by the tourists. This will make it possible for the hotel industry players to respond in good time to the anticipated changes in demand over time and also to maximize returns on investments.

A time series is a set of observations measured sequentially through time (Chatfield, 2000). Consider a time series as a sequence of random variables, x_1, x_2, x_3, … , where the random variable x_1 denotes the value taken by the series at the first time point, the variable x_2 denotes the value for the second time period, x_3 denotes the value for the third time period, and so on. In general, a collection of random variables, $\{x_t\}$, indexed by t is referred to as a stochastic process. The index t, will typically be discrete and vary over the integers $t = 0, \pm1, \pm2, \ldots$ or some subset of the integers. The observed values of a stochastic process are referred to as a realization of the stochastic process. The main objectives of a time series analysis are: description, modeling, forecasting and control.

A time series may consist of the following components: seasonal variations, cyclic variations, trend component and random errors. Trend may be termed as a long term change in the mean level of the time series e.g. population. Seasonal variations are oscillations with a period of one year and are exhibited by many time series such as sales figures and temperature readings. Cyclic variations have a fixed period not equal to one year due to some other physical causes e.g. the daily temperatures of a town in East Africa will always be high at around midday every day. Such variation has a period of 24 hours. In addition some time series exhibit oscillations which are predictable to some extent e.g. economic data are sometimes thought to be affected by business cycles with a period varying between five and seven years. After trend, seasonal and cyclic have been removed from a data set, what remains are a series of residuals which may or may not be random. These fluctuations are due to forces which are beyond human control such as earthquakes, wars, draughts and floods.

Many methods and techniques for analyzing time series and forecasting are available in various literatures. This research shall focus on the Box-Jenkins models for time series analysis and modeling using the quarterly data on bed occupancy by tourists. The Box-Jenkins models allow the analyst to forecast future values of a series with only the past period's data, without having some related variable's data (Shumway and Stoffer, 2011). Seasonal variations are important in tourism and hospitality demands. The seasonality phenomenon is always contributed to by natural factors such as weather and temperature; institutional factors such as vacations, religious events and natural events among others. Forecasts produced by time series models are always useful in planning, inventory control, policy making and other related activities in the industry.

1.2. Objectives of the study

1.2.1. General objective

The general objective of the study is to model tourist accommodation demand in Kenya using time series analysis.

1.2.2. Specific objectives

i. To build a time series model for tourist accommodation demand in Kenya
ii. To apply the built model for tourist accommodation demand in Kenya to forecast future values

2. Literature review

The primary objective of time series analysis is to develop mathematical models that provide credible descriptions for sample data collected. Time series analysis can be done either in the time domain or frequency domain. Analysis in the time domain involves relating series to each other or to evaluating the effects of treatments or design parameters that arise when time-varying phenomena are subjected to periodic stimuli. However for the analysis in the frequency domain, the nature of the physical or biological phenomena under study are best described by their Fourier components rather than by the difference equations involved in ARIMA or state-space models.

A model of a time series is a mathematical formulation that summarizes the series. This formulation has a model form and several parameters (Rossiter, 2012). The aims of modeling a time series are: (1) to understand the underlying process (es) that produced it; (2) to predict into the future, at missing data points, or into the past and (3) to use the model to simulate similar time series (Rossiter, 2012).

Time series models come in three kinds: moving average (MA), auto-regressive (AR) and auto regressive moving average (ARMA) models (Box et al., 2008). An ARMA model has components of moving average and auto-regressive nature. A seasonal auto regressive moving average (SARIMA) model; SARIMA (2,0,2)(0,1,2) was developed to forecast hotel-motel room nights in New Zealand in 2007 (Andersen et al., 2009). It used the auto regressive moving average processes to explain the seasonal patterns of hotel-motel room nights occupied. It also used the HEGY test in order to check for the presence of seasonal and non-seasonal unit roots in the room-nights series. The challenge with this model was that it did not perform well in the out-of-sample forecast.

An analysis of Kenya's monthly tourism data from 1971 to 1990 was done, but encountered missingness in the period 1984 when there were no monthly records at the Kenya Central Bureau of Statistics (CBS) (Kihoro, 1998). The immediate problem was to fill the missing block of data and that was the main focus of the work and it had used data for fewer years (Kihoro, 1998). A suggestion made was that monthly data for more years need to be incorporated to have a wider view of the behavior of the series over a long period of time.

There are lots of literatures covering studies done on modeling of tourism data globally and in Kenya. However, much has not been done in modeling accommodation demand for tourists in Kenya. The studies conducted so far looked at data available for shorter periods of time. There is therefore a need to model the tourism accommodation demand using monthly data covering a longer period of time.

3. Methodology

3.1. Autocorrelation function

An important guide to the properties of a time series is provided by a series of quantities called sample autocorrelation coefficients or serial coefficients which measure the correlation between observations at different distances apart. These coefficients often provide some insight into the probability model that generated the data. Given n observations $x_1, x_2, ..., x_n$, on a discrete time series ,we can form n-1 pairs of observations namely: $(x_1, x_2),(x_2, x_3),(x_3, x_4),(x_4, x_5),...,(x_{n-1}, x_n)$.

Regarding the first observation in each year as one variable and the second observation as a second variable the correlation between x_t and x_{t+1} is given by:-

$$r_1 = \frac{\sum_{t=1}^{n-1} (x_t - \bar{x}_1)(x_{t+1} - \bar{x}_2)}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \quad (1)$$

Whereas the correlation between observations a distance k apart is given by:-

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2} \quad (2)$$

3.2 Stationary time series

A time series is stationary if it has no systematic change in mean (this means no trend), no systematic change in variance and if strictly there are no periodic variations. Most time series analysis techniques require data to be stationary and therefore non-stationary series always need to be made stationary. Mathematically, stationary time series is defined as follows:

A Time series is said to be strictly stationary if the joint distribution of $x_{t_1} + x_{t_2} + x_{t_3}, ..., x_{t_n}$ is the same as the joint distribution of $x_{t_{1+h}} + x_{t_{2+h}} + x_{t_{3+h}}, ..., x_{t_{n+h}}$, for all $t_1, t_2, t_3, ..., t_n$ and h_i i.e. the probability structure of the process does not change with time. In practice it is always difficult to define stationarity in this manner, therefore there is a less stringent way of defining stationarity; in a weak sense. A time series is said to be of second order stationarity (or weak stationarity) if its mean is constant and its auto covariance function is independent of time but depends only on distance between the variables. No assumptions are made about higher order moments than those of the second order.

3.3. Seasonal Autoregressive Integrated Moving Average (SARIMA) Model

Since seasonality is dominant in tourism demand, the SARIMA $(p,d,q)(P,D,Q)_S$ with D seasonal differences and periods/seasons S is more appropriate.

The general SARIMA model introduced by Box and Jenkins includes: auto regressive and moving average components as well as the required order of differencing. The three types of parameters in the model are: the auto regressive parameter (p), the moving average parameter (q) and the number of differencing required (d). These three are summarized as ARIMA (p,d,q). For example a model that is described as ARIMA (1,1,1) means that it contains 1 auto regressive (p) parameter and 1 moving average (q) parameter after differencing the data once so as to attain stationarity.

The seasonal ARIMA (P,D,Q) parameters can also be identified for the data. These are; seasonal auto regressive (P), seasonal moving average (Q) and seasonal differencing (D) components. If the p, d, and q are as stated above, and the seasonal parameters are 1,1 and 1 for P, D and Q respectively, then an ARIMA (1,1,1)(1,1,1)12 describes a model that includes 1 AR parameter, 1 MA parameter, 1 SAR parameter and 1 SMA parameter. These parameters are generated after differencing the series once at lag 1 and then once at lag 12 to remove seasonality).

The general form for the model is:

$$(1 - \emptyset_1 B)(1 - \propto_1 B^{12})(1 - B)(1 - B^{12})X_t = (1 - \theta_1 B)(1 - \gamma_1 B^{12}) \quad (3)$$

Where:

ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.4, No.10, 2014

$(1 - \emptyset_1 B)$ is the non-seasonal autoregressive component of order 1

$(1 - \propto_1 B^{12})$ is the seasonal autoregressive of order 1

X_t is the current value of the time series examined

B is the backward shift operator (BX_t=X_{t-1}) and ($B^{12}X_t = X_{t-12}$)

$(1 - B)$ is the first order non-seasonal difference

$(1 - B^{12})$ is the seasonal difference of order 1

$(1 - \theta_1 B)$ is the non-seasonal moving average of order 1 and,

$(1 - \gamma_1 B^{12})$ is the seasonal moving average of order 1.

3.4. Data

Quarterly data on tourist accommodation demand were obtained from various Statistical Abstract publications by the Kenya National Bureau of Statistics (KNBS) for the period between 1985 and 2013. Data were collected about: reported visitor arrivals by purpose, reported visitor arrivals by continent of residence, length of stay of departing visitors by purpose and country of residence and hotel bed availability and occupancy by zones in Kenya.

4. Seasonal ARIMA model building

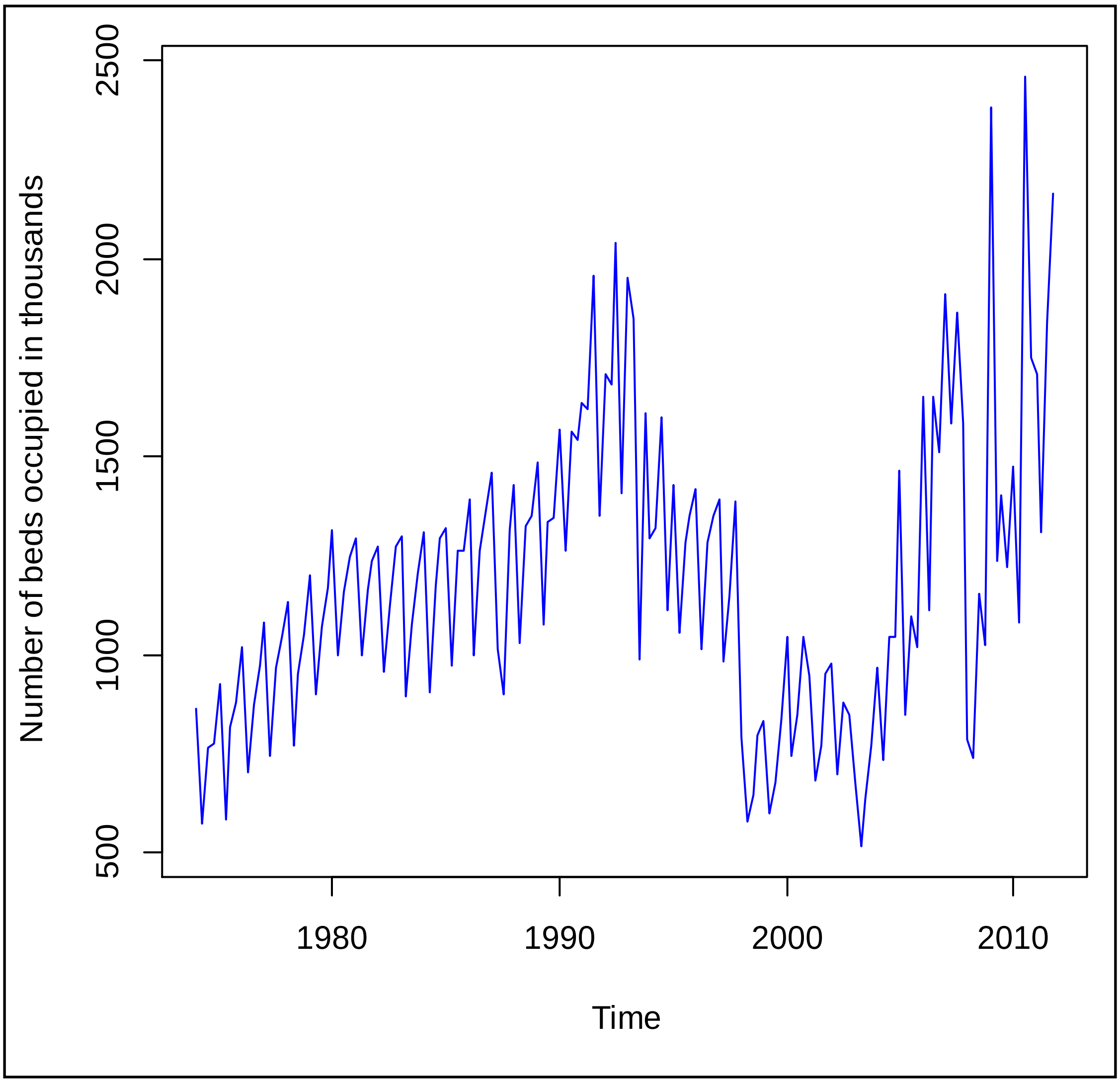
The model parameters were identified from various plots of the series. These parameters were then used to estimate time series models. Diagnostic checks were then conducted on the models to ensure an optimal one is selected. These diagnostic checks include the analysis of residuals from the fitted models. Forecasting of future values was then done from the optimal model.

R statistical software was used to model the seasonal ARIMA for the data on beds occupied. The R software and the required packages were downloaded from the CRAN website (CRAN, 2014) and installed on the computer for use. Various time series packages such as tseries, forecast, and zoo were downloaded and installed and used for the required analyses (McLeod et al., 2011).

4.1. Model identification

In order to generate a time series model for the accommodation demand, quarterly data on the number of beds occupied by visitors were collected for the period 1974 to 2011 and entered in to Ms. Excel. R statistical software was used to import the data set from Ms. Excel and perform time series analyses. The R package *tseries* was used in the analysis.

The time plot of the series of the number of hotel beds occupied by both East African and foreign residents was generated and is shown in ***Figure 2*****Error! Reference source not found.**.



The time plot in ***Figure 2*** shows that the original series is non-stationary and needs differencing. There is an upward trend and quarterly seasonality present as evident in the time plot. With trend and seasonality, first and fourth differences were taken so as to attain stationarity.

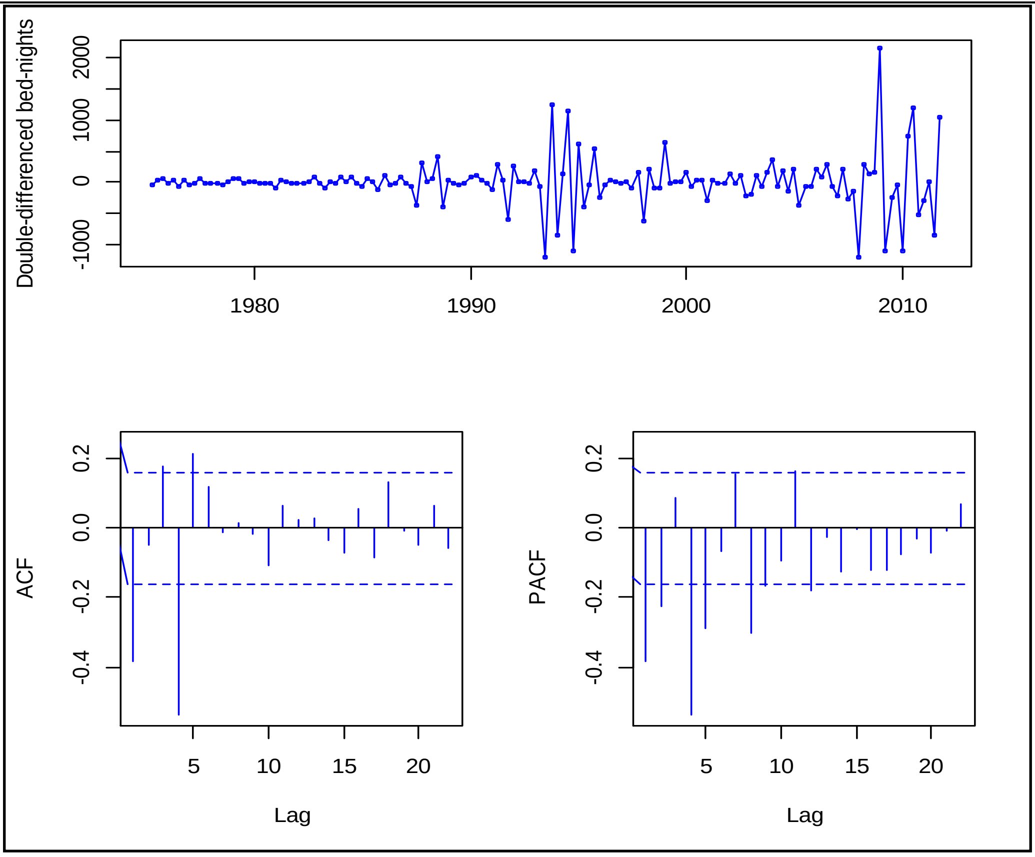


Figure 3: Double-differenced series of beds occupied

Figure 3 shows the plot of the double- differenced series, which is stationary. The double differenced series showed significant spikes at lag 2 in the ACF, suggesting a non-seasonal MA(2) component, and a significant spike at lag 4 in the ACF suggesting a seasonal MA(1) component. This means that q=2 and Q=1. Similarly, the significant spikes at lags 1 and 2 suggest non-seasonal AR(1) and AR(2), and the significant spikes at lags 4, 8 and 12 suggest a seasonal AR(1). This means that p=1 or p=2 and P=1. Since non-seasonal differentiation was done once before stationarity, the component d equals1 and since the seasonal differentiation was done once, then D equals 1. This resulted into two possible seasonal ARIMA models: SARIMA(1,1,2)(1,1,1)[4] and SARIMA(2,1,2)(1,1,1,)[4].

4.2. Model estimation

The two identified models were then estimated and the results are as in Table 1Error! Reference source not found. **and**

Table 2. They show the two models with their MA and AR coefficients and the AIC, AICc and BIC values.

Table 1: Fitted SARIMA (1,1,2)(1,1,1)[4] model

	ar1	ma1	ma2	sar1	sma1
Coefficients	-0.9885	0.3628	-0.5418	-0.1197	-1
s.e.	0.0507	0.0838	0.0829	0.0916	0.0963

sigma^2 estimated as 55428: log likelihood=-1018.53

AIC=2049.06 AICc=2049.66 BIC=2067

Table 2: Fitted SARIMA (2,1,2)(1,1,1)[4] model

	ar1	ar2	ma1	ma2	sar1	sma1
Coefficients	0.7478	-0.1635	-1.3285	0.5204	-0.1088	-0.9256
s.e.	0.2863	0.1961	0.2686	0.229	0.1305	0.1057

sigma^2 estimated as 57927: log likelihood=-1019.11

AIC=2052.23 AICc=2053.03 BIC=2073.16

The coefficients and their standard errors are displayed in Table 1 **and**

Table 2. The AIC, AIC_C and BIC are also indicated in the two tables.

4.3. Model checking and forecasting

The AIC, AICc and BIC of the SARIMA(1,1,2)(1,1,1)[4] model were 2049.06, 2049.66 and 2067 respectively, while those for the SARIMA(2,1,2)(1,1,1)[4] model were 2052.23, 2053.03 and 2073.16 respectively. We therefore chose the first model; SARIMA (1,1,2)(1,1,1)[4], since it had the smaller AIC, AICc and BIC values. The spikes in the residual display were insignificant and show that the residuals are white noise. A Ljung-Box test was results showed that the residuals were random. The p-value of this test was 0.6309; indicating that the null hypothesis of random residuals is not rejected.

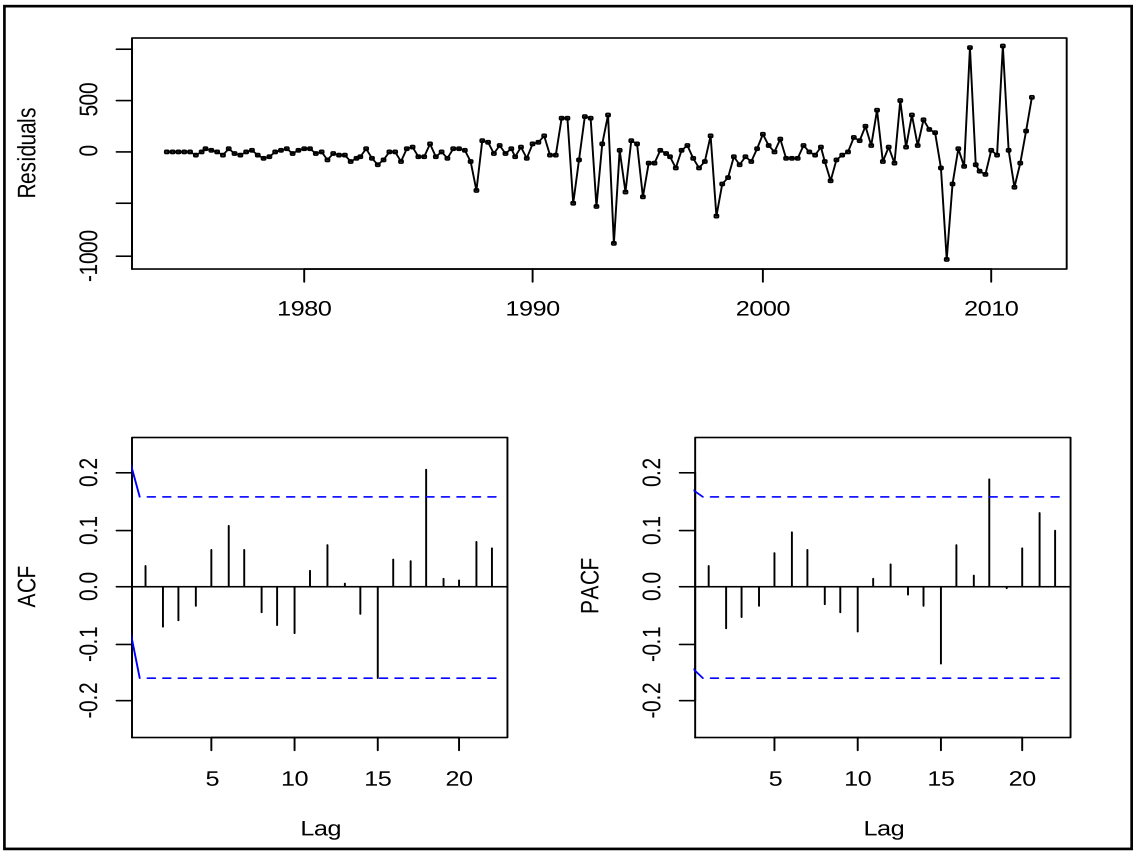


Figure 4: Plot of residuals, ACF and PACF

Error! Reference source not found. displays a plot of the residuals, the ACF and PACF of the residuals. The SARIMA (1,1,2)(1,1,1)[4] model is adequate and can be used for forecasting future values of bed-nights.

Table 3: Five-year forecast of bed-nights by SARIMA(1,1,2)(1,1,1)[4]

Quarter	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2012Q1	2153.0	1847.7	2458.3	1686.1	2619.9
2012Q2	1692.2	1365.9	2018.5	1193.2	2191.2
2012Q3	2023.6	1669.7	2377.5	1482.4	2564.9
2012Q4	1858.3	1486.0	2230.6	1288.9	2427.6
2013Q1	2131.5	1743.8	2519.3	1538.6	2724.5
2013Q2	1683.2	1281.5	2084.9	1068.9	2297.6

2013Q3	2032.9	1612.7	2453.2	1390.2	2675.6
2013Q4	1931.2	1498.0	2364.4	1268.7	2593.8
2014Q1	2166.0	1712.3	2619.8	1472.1	2859.9
2014Q2	1720.8	1254.0	2187.6	1006.9	2434.8
2014Q3	2063.8	1579.8	2547.8	1323.6	2804.0
2014Q4	1959.0	1462.6	2455.3	1199.9	2718.1
2015Q1	2193.9	1679.7	2708.2	1407.4	2980.4
2015Q2	1752.8	1226.3	2279.2	947.7	2557.9
2015Q3	2092.2	1549.9	2634.5	1262.8	2921.5
2015Q4	1992.0	1438.2	2545.9	1145.0	2839.1
2016Q1	2222.7	1652.0	2793.4	1349.9	3095.5
2016Q2	1785.3	1203.0	2367.6	894.8	2675.8
2016Q3	2120.9	1523.8	2718.1	1207.7	3034.2
2016Q4	2024.4	1416.1	2632.6	1094.1	2954.6

Table 3 shows the five-year forecasts for the number of tourists accommodated in Kenya, from quarter 1 of 2012 to quarter 4 of 2016.

5. Discussion and conclusion

The results from the parameter identification tests using the ACF and PACF of the original series and the differenced series suggest that d=1, D=1, q=2, Q=1, p=1 or 2 and P=1. Two SARIMA models are estimated; SARIMA(1,1,2)(1,1,1)[4] and SARIMA(2,1,2)(1,1,1)[4]. The process of identification, estimation and diagnostic checks resulted in the final selection of SARIMA (1,1,2)(1,1,1)[4] as the optimal model. The selected model can therefore be used for the generation of forecasts of tourist accommodation demand in Kenya. A five year quarterly forecast was fitted and shown in

Table 3. The tourism and related industries shall therefore benefit from this SARIMA model and its ability to provide reliable forecasts.

A similar study should be conducted with an objective of looking at monthly tourist accommodation data to generate a model for monthly forecasting of tourist accommodation demand in Kenya.

References

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). Time Series Analysis: Forecasting and Control. John Wiley and Sons Inc., Hoboken, NJ, fourth edition.

Chatfield, C. (2000). Time-Series Forecasting. Chapman and Hall CRC, New York.

CRAN (2014). The comprehensive r archive network.

Kihoro, J. M. (1998). Analysis of seasonal time series with missing observations. Master's thesis, Nairobi.

Lim, C. & Chan, F. (2009). Forecasting tourist accommodation demand in New Zealand. In Andersen, R. S. A., Braddock, R. D., and H., N. L. T. (Eds.), Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation. Paper presented at The 18th World IMACS / MODSIM Congress, Cairns, Australia, 13-17 July (pp. 1251-1257).

McLeod, A. I., Yu, H., and Mahdi, E. (2011). Time series analysis with r. Handbook of Statistics, 30.

Phillips, P. C. B. and Perron, P. (1990). Testing for a unit root in time series regressions. Biometrika, 75(2):336–346.

ROK (2012). Statistical Abstract 2012. Kenya National Bureau of Statistics, Nairobi.

Rossiter, D. G. (2012). Introduction to the R Project for Statistical Computing for use at ITC. University of Twente, Enschede.

Shumway, R. H. and Stoffer, D. S. (2011). Time Series Analysis and Its Applications, With R examples. Springer, New York

United Nations World Tourism Organization, (2013). International tourism receipts grew by 4% in 2012. [online] Available at: http://media.unwto.org/en/press-release/2013-05-15/international-tourism-receipts-grew-4-2012 [Accessed 6 Aug. 2014].