Seasonal Autoregressive Integrated Moving Average model for tax revenue forecast in Kenya

Fredrick Kyalo Samuel [1] and Titus Kithanze Kibua [2]

[1]Department of Mathematics and Actuarial Science, Kenyatta University, Nairobi, Kenya
P.O. Box 8747-00300, Nairobi, Kenya
E-mail: kyaloo2004@yahoo.com
[2]Department of Mathematics and Actuarial Science, Kenyatta University, Nairobi, Kenya
P.O. Box 43936-00100, Nairobi, Kenya
E-mail: tkibua@yahoo.com

[1]Corresponding Author

Published: 18 July 2019

Abstract
Modelling and forecasting tax revenue is desirable in an economy for long-term projections and proper fiscal planning. This study sought to establish Seasonal Autoregressive Integrated Moving Average (SARIMA) model for tax revenue forecast in Kenya. SARIMA $(2,0,0)(2,0,0)_{12}$ was identified as an appropriate model based least Akaike Information Criteria (AIC) value. The model passed residual normality test and the model residuals followed a white noise process. The predictive ability tests of Root Mean square (RMSE) and Mean Absolute Error (MAE) revealed that SARIMA $(2,0,0)(2,0,0)_{12}$ was accurate, consistent and appropriate for forecasting tax revenue. The Monte Carlo simulations of tax revenue using SARIMA $(2,0,0)(2,0,0)_{12}$ model produced similar plots for original and simulated time series models. The tax revenue forecasts will exhibit the similar patterns with no significant growth in the next five years. The study recommended application of SARIMA $(2,0,0)(2,0,0)_{12}$ model in tax revenue forecast in Kenya and enhancing tax revenue collections.

Keywords: Tax revenue, SARIMA model and forecasting

Cite this article: Samuel, F.K. & Kibua, T.K. (2019). Seasonal Autoregressive Integrated Moving Average model for tax revenue forecast in Kenya. *European International Journal of Science and Technology*, 8(6), 11-30.

1. Introduction

Tax is a compulsory contribution to the government, paid by individuals and corporate entities, which does not bear any relationship to the benefit received (Hyman, 1987). Tax revenue is a major component and largest source of government revenue in Kenya. Despite this, tax revenue performance has consistently failed to meet the set targets. The total cumulative revenue, including Appropriation in Aid (A-I-A) over the first half of the financial year 2015/16 amounted to Ksh 581.1 billion against a target of Ksh 678 billion, hence below target by Ksh 96.9 billion. Ordinary revenue amounted to Ksh 544.2 billion against a target of Ksh 591.8 billion over the same period. Thus, total revenue fell below the target by 17.0 per cent, while ordinary revenue also fell below target by 8.7 per cent (KIPPRA, 2016). A review of the performance of tax handles also showed that income tax and Valued Added Tax (VAT) were below their target by 10.0 per cent and 9.0 per cent, respectively in the first half of the financial year 2015/16. Excise and import duties were also below target by 6.0 per cent and 5.0 per cent, respectively over the same period (KIPPRA, 2016).

Besides failure of tax revenue to meet set targets, the government expenditures have continually exceeded revenues and maintained consistent growth patterns resulting to continuous increase in the public debt (KNBS, 2016). For instance, government revenue increased by Ksh 676.36 billion compared with an increase of Ksh 787.54 billion in expenditure between financial years 2011/12 and 2015/16. In relation to the Gross Domestic Product (GDP), government revenue averaged 19.88 per cent, while mean expenditure was 30.12 per cent, generating a resource gap of about 10.24 per cent for the period 2011 to 2015 (KIPPRA, 2016).

In order to ensure proper fiscal management, the tax structure must be stable and flexible to ensure revenues can be easily and predictably be forecasted with certainty (Todaro & Smith, 2003). Taxation should aim at raising sufficient revenue to fund public expenditure without too much public sector borrowing (Moyi & Ronge, 2006). In contrast, the total debt more than quadrupled between 2000 and 2014 thereby raising concern on the issue of debt sustainability given the increasing debt levels. A comparison of fiscal indicators across East Africa Community (EAC) countries for year 2015 indicated that Kenya had the highest debt to GDP ratio, estimated at 52.7 per cent compared to Uganda's 35.4 per cent, Tanzania's 40.5 per cent, Rwanda's 34.6 per cent, and Burundi's 38.4 per cent (KIPPRA, 2016).

Kenya has undergone numerous tax reforms aimed at raising tax revenue for funding government operations without excessive government borrowing (IEA, 2012). This is buttressed by increasing need for governments to mobilize their own internal resources to meet public expenditure (IMF, 2011). Despite this, an array of factors including tax evasion, corruption and fraud has been key impediment in meeting tax revenue targets. It's estimated that Ksh. 639 billion is lost annually in tax evasion by multinational corporations, thus significantly hampering economic growth (Lilian, 2015).

Forecasting tax revenue may be challenging owing to seasonality behaviour of the some tax revenue components such as VAT, Imports and Exercise duties. Nevertheless, modelling and forecasting tax revenue is generally desirable in an economy to enable government in long term projections and proper fiscal planning with view of containing public expenditure and escalating public debt. Consequently, appropriate tax revenue forecasting model capable of producing valid and reliable tax revenue forecast estimates is paramount.

The purpose of this study is to establish Seasonal Autoregressive Integrated Moving Average (SARIMA) model for tax revenue forecast in Kenya. SARIMA models are suitable for time series data exhibiting seasonality-periodic fluctuations. The models are capable of describing time series data that

exhibit non-stationary behaviours both within and across seasons, hence appropriate for studies concerning tax revenue that exhibit seasonality fluctuations.

2. Literature Review

Shengwei Wang et al. (2011) investigated the advantages and disadvantages SARIMA and the regression model with seasonal latent variable in forecasting precipitation demand. The study found out that rainfall had a strong autocorrelation of seasonal characteristics in time series. The SARIMA model provided good model fitting degree in decision-making for agricultural irrigation and SARIMA $(2, 0, 2)(1, 1, 1)_{12}$ model was suggested as appropriate prediction model. The study recommended full use of natural rainfall for corresponding areas and save underground water resources to ensure long-term stability of groundwater resources for sustainable development.

Michael (2014) used SARIMA model to determine an adequate forecasting model for the mean temperature of Ashanti Region in Ghana. SARIMA $(2,1,1)(1,1,2)_{12}$ had the least Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) and was considered as adequate model for prediction. The residuals of the model were white noise of passing Ljung-Box at 5%. The forecasted mean temperature values showed similar pattern of previous recordings.

Otu et al. (2014) applied SARIMA models in modelling and forecasting Nigeria's monthly inflation rates for the period November 2003 to October 2013. SARIMA $(1, 1, 1)(0, 0, 1)_{12}$ was identified as appropriate model for forecasting monthly inflation rates since it had least AIC. The forecast results revealed a decreasing pattern of inflation rates in the first quarter of 2014 and turning point at the beginning of the second quarter of 2014, where the rates took an increasing trend till the September.

Kibunja et al. (2014) used SARIMA model to forecast precipitation using a case study of Mt. Kenya region. SARIMA $(1, 0, 1)(1, 0, 0)_{12}$ was identified as the appropriate model since it had the least values of AIC and BIC. The model passed residual normality test and the forecasting evaluation statistics of mean error (ME), mean square error (MSE), root mean square error (RMSE) and mean absolute error (MAE). The study concluded SARIMA model was good for forecasting precipitation in Mt. Kenya region.

Susan et al. (2015) used SARIMA model to forecast inflation rate in Kenya. SARIMA $(0,1,0)(0,0,1)_4$ was identified as the best model since it had least AIC. Autocorrelation function (ACF) and Partial autocorrelation function (PACF) plots for the residuals and squared residuals revealed that they followed a white noise process and were homoscedastic respectively. The predictive ability tests of RMSE and MAE showed that the model was appropriate for forecasting the inflation rate in Kenya. The study recommended that appropriate policies and measures have to be adopted by the government and major stakeholders to ensure that the aim of single digit inflation rate value is achieved in Kenya.

Tadesse et al. (2017) Applied SARIMA model to forecast monthly flows in Waterval River in South Africa where Mean monthly flows from 1960 to 2016 were used for modelling and forecasting. Different SARIMA models were evaluated and SARIMA $(3, 0, 2)(3, 1, 3)_{12}$ model was selected for Waterval River flow forecasting based on the minimum values of AIC. Diagnostic check-up of forecasts was made using white noise and heteroscedasticity tests. Comparison of forecast performance of SARIMA models with that of computational intelligent forecasting techniques was recommended for future study.

Fazidah et al. (2018) forecasted Dengue Hemorrhagic Fever (DHF) cases using a case study in Asahan District of Indonesia. The results demonstrated that the reported DHF cases showed a seasonal variation. The SARIMA $(1,0,0)(0,1,1)_{12}$ model was identified as the best model that could be used to

predict incidence of DHF in Asahan District. Further research was recommended to integrate the forecasting model into the existing disease control program in terms of reducing the disease occurrence.

3. Methodology

3.1. SARIMA model

3.1.1. Notation

Box and Jenkins (1976) generalized the Autoregressive Integrated Moving Average (ARIMA) model to Seasonal Autoregressive Integrated Moving Average (SARIMA) model to deal with seasonality in time series. SARIMA is the product of seasonal and non-seasonal polynomials and is denoted by SARIMA (p, d, q) x (P, D, Q)$_s$, where (p, d, q) and (P, D, Q) are non-seasonal and seasonal components, respectively with a seasonality's'. SARIMA decomposes time series data into four components:

(1) Non seasonal Autoregressive (AR) process defined by:

$$\theta_p(B) = (1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p) \quad (3.1)$$

where $\theta_p(B)$ is a polynomial of order p; α_i $(i = 1,2,\ldots p)$ are coefficients to be estimated; B is backward shift operator (defined as $B^d x_t = x_{t-d}$); and d is the number of non-seasonal differences.

(2) Seasonal Autoregressive (SAR) process defined by:

$$\varphi_p(B^s) = (1 - a_1 B^s - a_2 B^{2s} - \cdots - a_p B^{Ps}) \quad (3.2)$$

where $\varphi_p(B^s)$ is a polynomial of order P; a_i $(i = 1,2,\ldots P)$ are coefficients to be estimated; B is backward shift operator; and s is the length of season

(3) Non seasonal Moving Average (MA) process defined by:

$$\omega_q(B) = (1 - \beta_1 B - \beta_2 B^2 - \cdots - \beta_q B^q) \quad (3.3)$$

where $\omega_q(B)$ is a polynomial of order q; β_i $(i = 1,2,\ldots q)$ are coefficients to be estimated; B is backward shift operator (defined as $B^d x_t = x_{t-d}$); and d is the number of non-seasonal differences.

(4) Seasonal Moving Average (SMA) process defined by:

$$\vartheta_Q B^s = (1 - b_1 B^s - b_2 B^{2s} - \cdots - b_Q B^{Qs}) \quad (3.4)$$

where $\vartheta_Q B^s$ is a polynomial of order Q; b_i $(i = 1,2,\ldots Q)$ are coefficients to be estimated; and B is backward shift operator; and s is the length of season

The general notation of SARIMA model may thus be written as;

$$\theta_p(B)\varphi_P(B^s)\nabla^d\nabla_s^D x_t = \mu + \omega_q(B)\vartheta_Q(B^s)z_t \quad (3.5)$$

where B is backward shift operator (defined as $B^d x_t = x_{t-d}$); ∇ is non-seasonal differencing operator (defined as $\nabla^d x_t = (1-B)^d x_t = \{\sum_{i=0}\binom{d}{i}(-B)^{d-i}\}x_t$); ∇_s is seasonal differencing operator (defined as $\nabla_s^D x_t = (1-B^s)^D x_t$); $\theta_p(B)$ and $\omega_q(B)$ are polynomials of order p and q, respectively; $\varphi_p(B^s)$ and $\vartheta_Q B^s$ are polynomial of degrees P and Q, respectively; p is the order of non-seasonal auto regression; d is the number of non-seasonal differences; q is the order of non-seasonal moving average; P is the order of seasonal auto regression; D is the number of seasonal

differences; Q is the order of seasonal moving average; s is the length of season; and z_t is the error term.

3.1.2. SARIMA Modelling

Box and Jenkins (1976) proposed four sequential stages for SARIMA modelling and forecasting:

(1) Model Identification

The objective of this stage is to determine if the time series is stationary by having constant mean and variance. That is, $E(x_t) = \mu$ and $\text{var}(x_t) = \sigma^2$. If the model is found to be non-stationary, stationarity is achieved mostly by differencing the series. Consider $x_t = x_1, x_2, \ldots, x_n$ to be non-stationary time series. Through differencing, the stationary time series may be obtained by first order differencing given by $\nabla x_t = x_t - x_{t-1}$. In the second order differencing we use the operator ∇^2 such that $\nabla^2 x_{t+2} = \nabla x_{t+2} - \nabla x_{t+1}$. The number of times that the original series is differenced to achieve stationarity is the order of homogeneity. Stationarity could also be achieved by some mode of transformation like the log transformation. Transformation helps to stabilize the variance in a series where the variation changes with the level.

Once stationarity has been achieved, the type (non-seasonal and seasonal) and order (p, q, P, Q) of model parameters are determined by inspection of ACF and PACF patterns to identify potential SARIMA model which might provide the best fit to the data. Identification criteria of non seasonal (p, q) and seasonal (P,Q) orders of stationary Autoregressive Moving Average (ARMA) model is indicated in **Table 3.1.** Based on the inspection of ACF and PACF, several members of SARIMA model could be identified, whose parameters are estimated using the Maximum Likelihood method.

(2) Parameter Estimation

This stage entails finding the values of the model coefficients which provide the best fit to the data. A range of potential members of SARIMA model are estimated by maximum likelihood methods, and for each, the Akaike Information Criteria (AIC) is calculated as:

$$\text{AIC}(p,q) = -2\,\text{In}(L) + 2k \qquad (3.6)$$

where L is the maximized value of the likelihood function for the estimated model; and k(= p+q+1) is the number of parameters in the statistical model. The model with smallest AIC value is judged as the best model.

(3) Diagnostics Checking

This involves checking whether the residuals follow a white noise process and the estimated parameters are statistically significant. Residuals from the model are examined to ensure that the model is adequate (random) by inspecting the time plot of residuals, ACF and PACF plots of residuals, the residual normal Quantile Quantile (QQ) Plot and Ljung-Box tests.

(4) Forecasting

When a satisfactory SARIMA model has been found to be adequate, we proceed to forecast or predict for a period or several periods ahead. Forecasting is based on the assumption that the past pattern and behaviour of the variable will continue in the future. However, chances of forecast errors are inevitable as the period advances. Suppose $x_1, x_2, \ldots, x_n$ be the observed time series, we estimate the future values such as x_{n+k} made at time $t = n$ for k steps ahead. Let $\hat{X}_{n+k}$ denote the estimate/forecast where k is known as the lead time. We need to look for an estimate in such that the Mean Squared

Error (MSE) of the predictor is minimized. That is, $MSE(\hat{X}_{n+k}) = E(X_{n+k} - \hat{X}_{n+k})^2$ should be minimized. The accuracy and consistency the competing SARIMA models is determined by comparison of some statistics such as mean error (ME), root mean square error (RMSE), mean absolute error (MAE), mean absolute square error (MASE) and mean absolute percentage error (MAPE) where the best forecasting model is considered based on minimum of these statistics.

3.2. Asymptotic properties of estimator

3.2.1. Convergence of an estimator

A sequence $\{x_t\}$ converges to x $(x_t \to x)$, if $|x_t \to x| \to 0$ as $t \to \infty$ (or for every $\varepsilon > 0$, there exists an n where for all $t > n$, $|x_t - x| < \varepsilon$. As the sample size increases, the estimator should converge to the parameter of interest. Consider a likelihood criterion, $L_n(a)$ which we shall use to estimate parameters a_0, $\bar{a}_n$ by maximizing (or minimizing) a criterion where; $\bar{a}_n = \arg max_{a\epsilon\theta} L_n(a)$ and θ is the parameter space we do the maximization (or minimization) over. Typically, the true parameter a should maximize (minimize) the 'limiting' criterion L.

3.2.2. Consistency property of an estimator

Suppose $\bar{a}_n = \arg max_{a\epsilon\theta} L_n(a)$ and $a_0 = \arg max_{a\epsilon\theta} L(a)$ is the unique maximum. If $sup_{a\epsilon\theta}|L_n(a) - L_a| \to^{a.s} 0$ as $n \to \infty$ and $L(a)$ has a unique maximum. Then $\bar{a}_n \to^{a.s} a_0$ as $n \to \infty$.

Proof

We note that by definition we have $L_n(a_0) \leq L_n(\bar{a}_n)$ and $L(\bar{a}_n) \leq L(a_0)$. Using this inequality, we have: $L_n(a_0) - L(a_0) \leq L_n(\bar{a}_n) - L(a_0) \leq L_n(\bar{a}_n) - L(\bar{a}_n)$

Therefore from the above we have:

$|L_n(\bar{a}_T) - L(a_0)| \leq \max\{|L_n(a_0) - L(a_0)|, |L_n(\bar{a}_T) - L(\bar{a}_n)|\} \leq sup_{a\epsilon\theta}|L_n(a) - L_a|$

Hence since we have uniform convergence, we have $L_n(\bar{a}_n) - L(a_0) \to^{a.s} 0$ as $n \to \infty$. Now since $L(a)$ has a unique maximum, we see that $L_n(\bar{a}_n) - L(a_0) \to^{a.s} 0$ implies $\bar{a}_n \to a_0$

3.2.3. Asymptotic normality of an estimator

We derive asymptotic normality of an estimator under the following assumptions:

1. That θ is univariate.
2. The third derivative of contrast function, $L_n(\theta)$ exists, its expectation is bounded and it's variance converges to zero as $n \to \infty$.

Lemma

Suppose that the third derivative of the contrast function $L_n(\theta)$ exists for $k = 0,1,2$ $E\left(\frac{\partial^k L_n(\theta)}{\partial\theta^k}\right) = \frac{\partial^k L}{\partial\theta^k}$ and $var\left(\frac{\partial^k L_n(\theta)}{\partial\theta^k}\right) \to 0$ as $n \to \infty$ and $\frac{\partial^3 V(\theta)}{\partial\theta^3}$ is bounded by a random variable z_n which is independent of n where $E(z_n) < \infty$ and $var(z_n) \to 0$. Then we have $(\theta_n - \theta_0) = V(\theta)^{-1}\frac{\partial L_n(\theta)}{\partial\theta}_{\theta=\theta_0} + o_p(I)\frac{\partial L_n(\theta)}{\partial\theta}_{\theta=\theta_0}$, where $V(\theta_0) = \frac{\partial^2 L(\theta)}{\partial\theta^2}_{\theta_0}$

Proof

$$\frac{\partial L_n(\theta)}{\partial\theta\ \ \theta=\theta_0} = \frac{\partial L_n(\theta)}{\partial\theta\ \ \theta=\bar{\theta}_n} - (\bar{\theta}_n - \theta_0)\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n} = -(\bar{\theta}_n - \theta_0)\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n} \quad (3.7)$$

where $\bar{\theta}_n$ lies between θ_0 and $\bar{\theta}_n$. We first study $\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n}$. By using the mean theorem we have:

$\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n} = \frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta_0} + (\bar{\theta}_n - \theta_0)\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n}$ where $\bar{\theta}_n$ lies between θ_0 and $\bar{\theta}_n$.

Since $\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta_0} \to \frac{\partial^2 L(\theta)}{\partial\theta^2\ \ \theta_0} = V(\theta_0)$, under the stated assumptions, we have

$$|\frac{\partial^2 L(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n} - (V(\theta_0)| \leq |\bar{\theta}_n - \theta_0||\frac{\partial^2 L_n(\theta)}{\partial\theta^2\ \ \theta=\bar{\theta}_n}| \leq |\bar{\theta}_n - \theta_0|L_n$$

Therefore, by consistency of the estimator, it is clear that $\frac{\partial^2 L}{\partial\theta^2\ \ \theta=\bar{\theta}_n} \to^{\mathrm{p}} V(\theta_0)$. Substituting this into equation (3.7), we have:

$\frac{\partial \mathrm{L}}{\partial\theta\ \ \theta=\theta_0} = -(\bar{\theta}_n - \theta_0)(V(\theta_0) + \mathrm{o}(\mathrm{I}))$, since $(V(\theta_0)$ is bounded away from zero, we have $[\frac{\partial^2 L}{\partial\theta^2\ \ \theta=\bar{\theta}_n}]^{-1} = (V(\theta_0)^{-1} + \mathrm{o}_p(I)$ hence the proof.

3.3. Monte Carlo Simulation

Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. Monte Carlo simulation involves three steps:

(1) Sampling on input random variables

The purpose of sampling on the input random variables $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is to generate samples that represent distributions of the input variable from their cumulative density functions (cdfs), $F_{x_i}(x_i)(i = 1,2, \ldots, n)$. The samples of the random variables are then used as inputs to the simulation experiments.

(2) Numerical Experimentation

Suppose that N samples of each random variable are generated, then all the samples of random variables constitute N sets of inputs, $x_i = (x_{i1}, x_{i2, \ldots} x_{in}), (i = 1,2, \ldots, N)$, to the output model $\boldsymbol{Y} = g(\boldsymbol{X})$. Solving the problem N times deterministically yields N sample points of the output Y: $y_i = g(x_i), (i = 1,2, \ldots, N)$

(3) Extraction of probabilistic information of output variables

After N samples of output $\boldsymbol{Y}$ have been obtained, statistical analysis is carried out to estimate the characteristics of the output $\boldsymbol{Y}$, such as the mean, variance, reliability, the probability of failure, probability density function (pdf) and cumulative density function (cdf).

4. Results and Discussion

In this study, secondary data on monthly tax revenue collections for July 2000 through to June 2015 which was made up of 192 monthly data series was obtained from the Kenya Revenue Authority (KRA). The analysis was conducted using R version 3.5.1 statistical software.

4.1. Exploratory data analysis

The time plot of monthly tax revenue collections in Kenya from 2000 to 2015 is indicated in **Figure 4.1.** The time plot exhibits a systematic change, therefore giving evidence of trend in the data. The increasing linear trend component indicates that the data in each month of the year are increasing with time. Therefore, the tax revenue data is non-stationary, hence differencing is needed to achieve stationarity. The sample Autocorrelation function (ACF) plot of monthly tax revenue collections in Kenya from 2000 to 2015 is indicated in **Figure 4.2.** The ACF plot dies down in attenuating sine wave pattern implying existence of seasonal component of the time series.

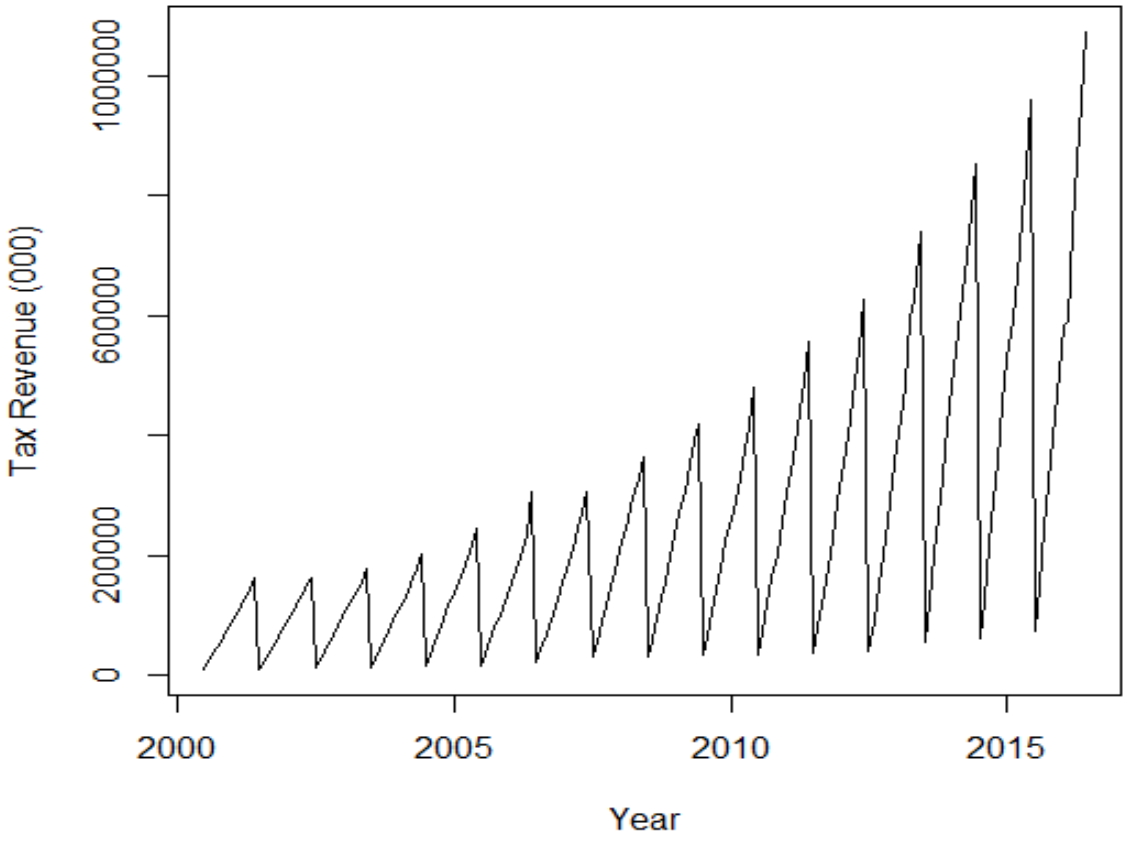

Figure 4.1: Time plot of monthly tax revenue collections in Kenya from 2000 to 2015

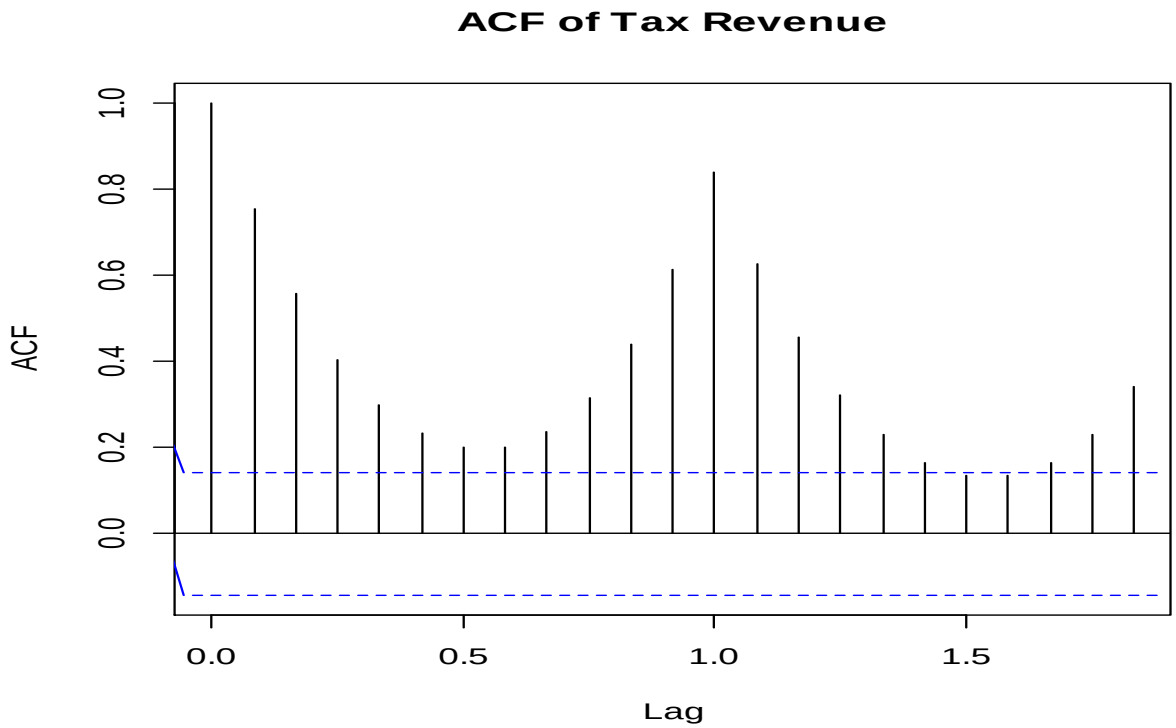

Figure 4.2: Time Sample ACF for the monthly tax revenue collections in Kenya from 2000 to 2015

4.2. Establishing SARIMA model

4.2.1. Model identification

Differencing of the monthly tax revenue data was applied to achieve the stationarity as indicated in **Figure 4.3.** The first differencing removed the nonstationarity contained in the original series, eliminated the linear trend component as well as the sinusoidal curves. However, differencing failed to remove non-stationarity on variance since the variation in the plot is increasing as we move towards the right of the graph. Hence, the log transformation was applied on the original time series prior to differencing to achieve stationarity on variance as indicated in **Figure 4.4.** Decomposition of log-transformed monthly tax revenue indicated in **Figure 4.5** confirms evidence of trend, seasonality and random effects in the time series data where the seasonal component in the data is also very strong. **Figure 4.6** shows the first differencing of log-transformed tax revenue data which achieves stationary on both mean and variance.

The sample ACF and PACF of seasonally differenced tax revenue displayed in **Figure 4.7** and **Figure 4.8** respectively, were used to identify the type (non-seasonal and seasonal) and number of model parameters (p, q, P, Q) required in the model through inspection of the patterns to identify potential SARIMA model which might provide the best fit to the data while putting into consideration characteristics of ACF and PACF highlighted in **Table 3.1.** The sample ACF is decreasing exponentially and has positive spikes with values that are decreasing in absolute magnitude at lags 0 and 1 which implies evidence of seasonality with very strong short term autocorrelations of past tax revenue collections. The sample PACF cuts off at lag 2 hence two non-seasonal AR parameters and two seasonal AR parameters are required in the SARIMA model.

Based on the identification plots given in **Figure 4.7** and **Figure 4.8**, the following models were suggested:

(a) SARIMA $(2,0,0)(2,0,0)_{12}$
(b) SARIMA $(2,0,0)(2,1,0)_{12}$
(c) SARIMA $(2,1,0)(2,0,0)_{12}$
(d) SARIMA $(2,1,0)(2,1,0)_{12}$

Table 3.1: ***Characteristics of ACF and PACF for stationary Autoregressive Moving Average (p,q) process***

		AR (p)	**MA (q)**	**ARMA (p.q)**
Non-seasonal ARMA (p,q)	**ACF**	tails off at lag k (k=1,2,3...)	cuts off after lag q	Tails off
	PACF	cuts off after lag p	Tails off at lags k (k=1,2,3...)	Tails off
		AR $(P)_s$	**MA $(Q)_s$**	**ARMA $(P.Q)_s$**
Pure Seasonal ARMA (P,Q)	**ACF**	tails off at lag ks (k=1,2,3...)	cuts off after lag Qs	Tails off at ks
	PACF	cuts off after lag Ps	Tails off at lags ks (k=1,2,3...)	Tails off at ks

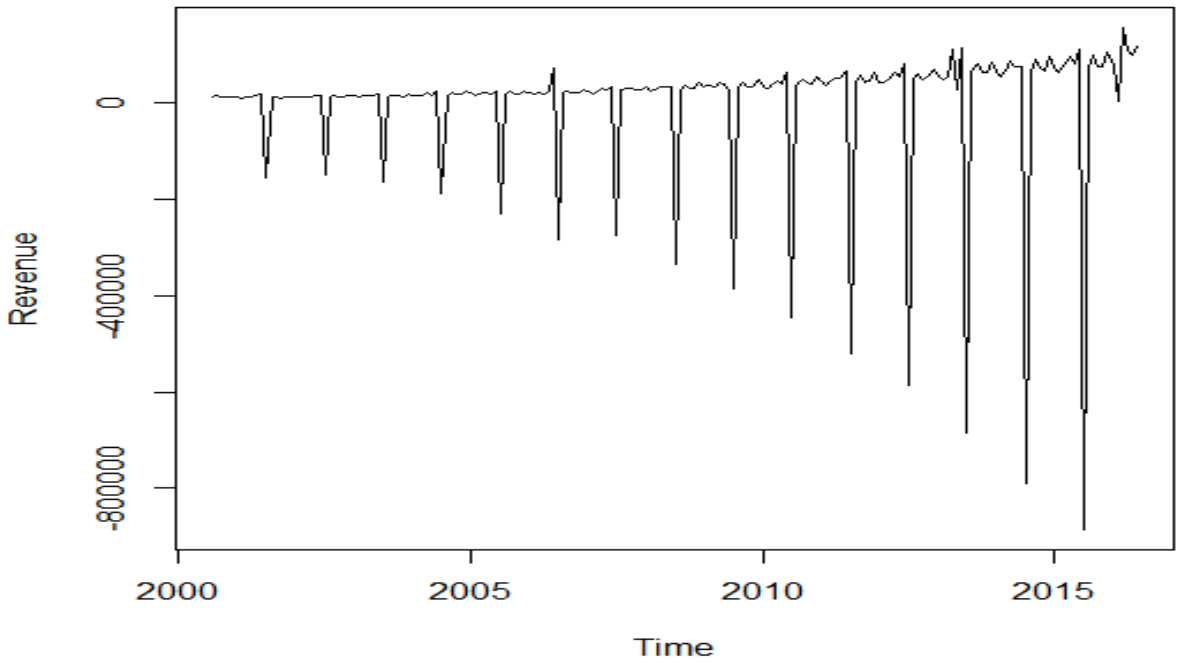

Figure 4.3: First differencing of the monthly tax revenue in Kenya

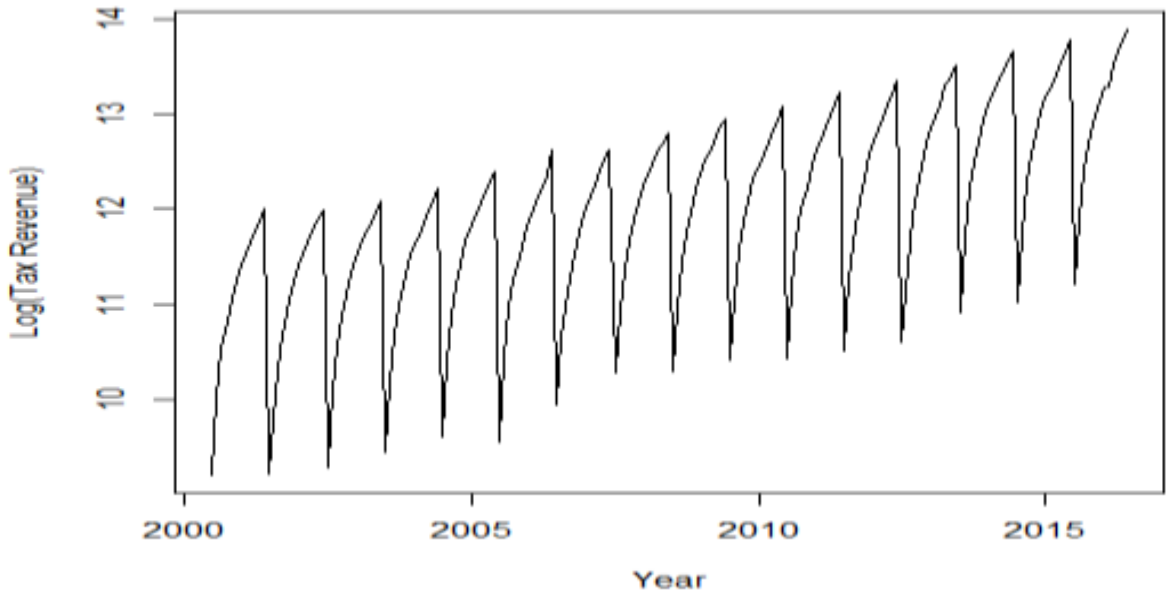

Figure 4.4: Log transformation of monthly tax revenue collections in Kenya from 2000 to 2015

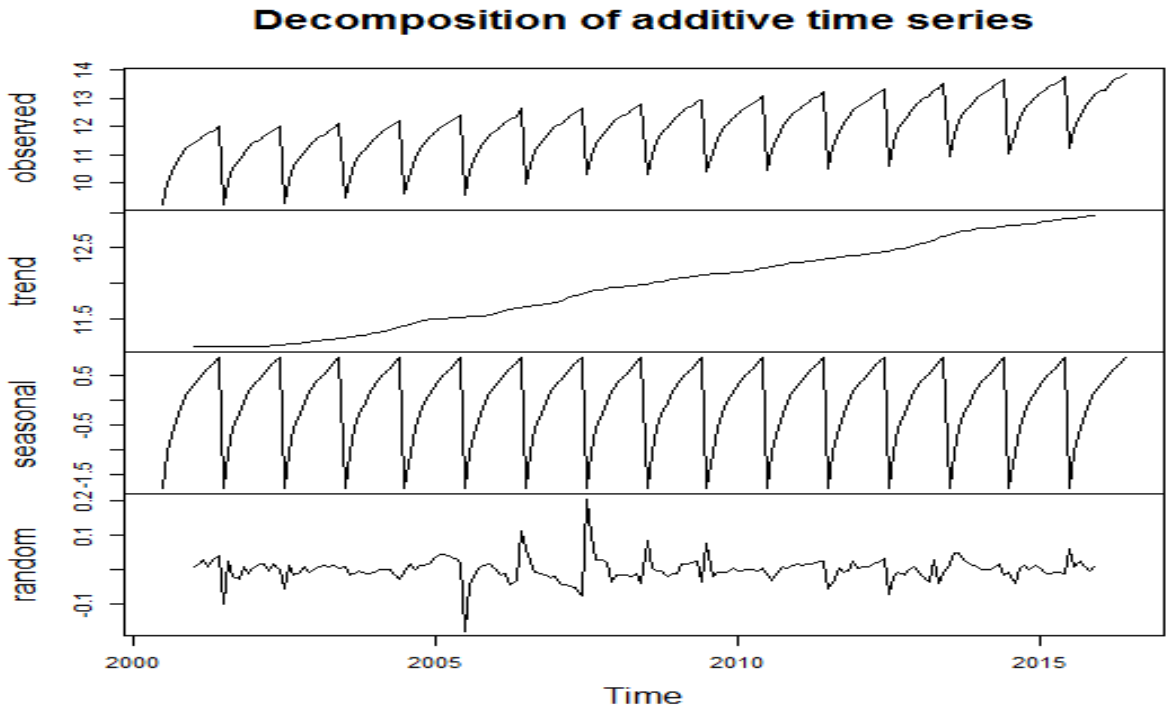

Figure 4.5: Decomposition tax revenue collections in Kenya

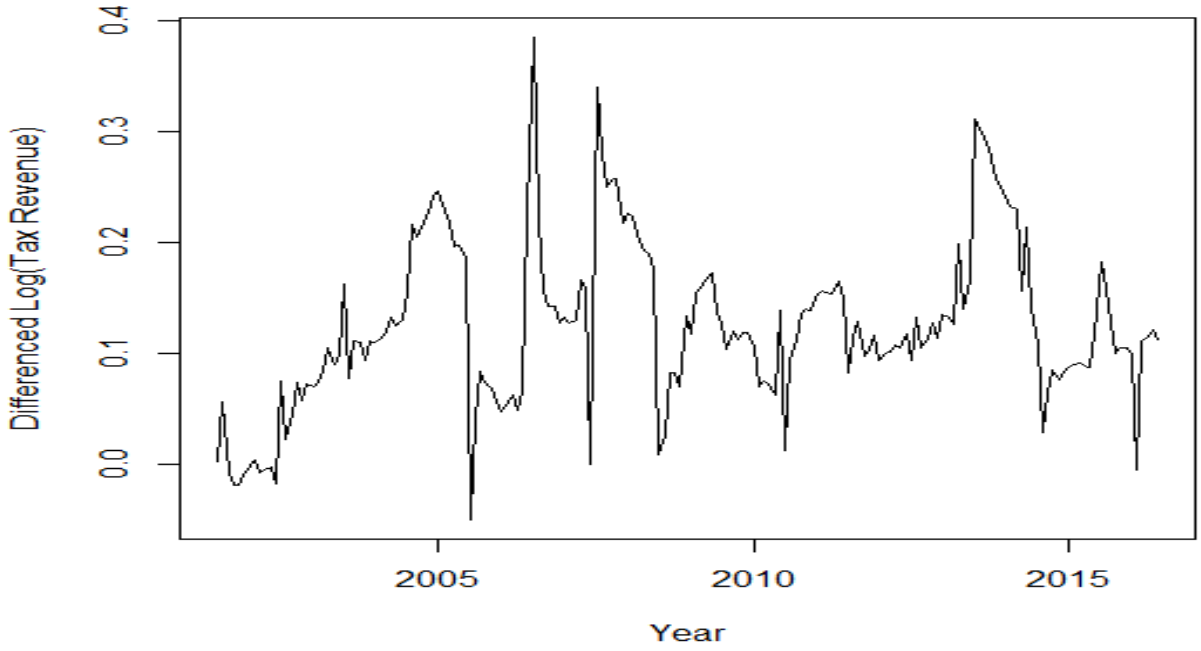

Figure 4.6: First differencing of the log transformed tax revenue

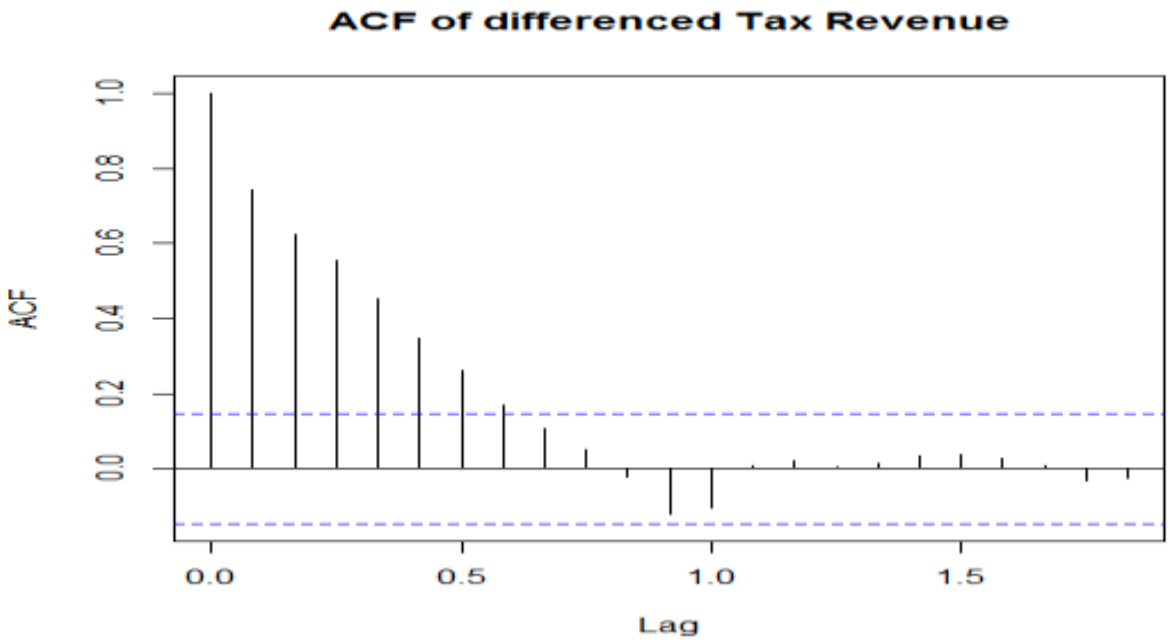

Figure 4.7: ACF of first differencing of tax revenue in Kenya

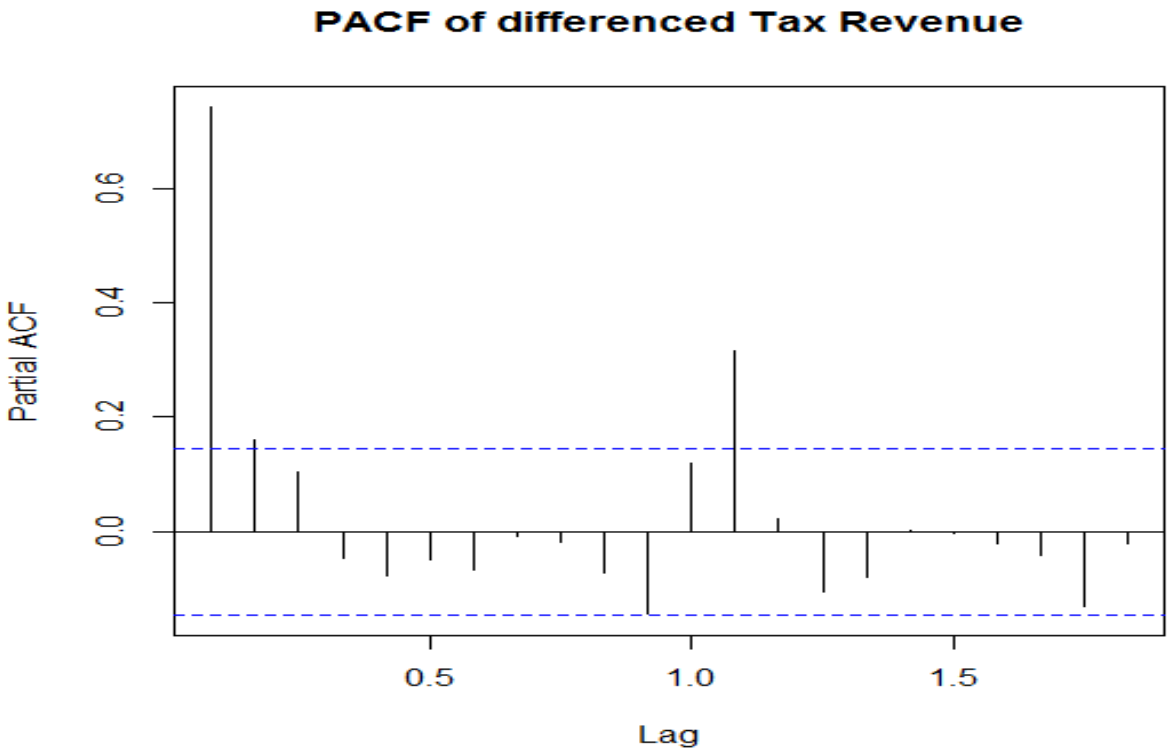

Figure 4.8: PACF of first differencing of tax revenue in Kenya

4.2.2. Parameter Estimation

The Akaike Information Criteria (AIC) for four suggested models are indicated in **Table 4.1.** Based on the diagnostics of AIC values, SARIMA $(2,0,0)(2,0,0)_{12}$ model with a drift (mean) had the minimum AIC value compared to other models hence better model for forecasting monthly tax revenue collections in Kenya. Using the general notation of SARIMA model in equation (3.5), SARIMA $(2,0,0)(2,0,0)_{12}$ model with a mean is thus given by:

$$\theta_p(B)\varphi_p(B^s)x_t = \mu + z_t \qquad (4.1)$$

where $\theta_p(B)$ is the non-seasonal AR part with two parameters (ar1=0.5949 and ar2 =0.2247); $\varphi_p(B^s)$ is the seasonal AR part with two parameters (sar1= -0.2846 and sar2 = -0.3114); B is the backshift operator; μ is mean (0.1272); z_t is error term; and s = 12 indicates monthly series data.

From SARIMA $(2,0,0)(2,0,0)_{12}$ model, the order of p (= 2) means that the current time series (x_t) is reliant on its preceding data x_{t-1} and x_{t-2}. The order P (=2) means that x_t is reliant on its preceding years' data of x_{t-12} and x_{t-24}. The orders q,Q (=0) means x_t is not reliant on its preceding random shocks, z_t. The model was stationary as depicted by non-seasonal (d=0) and seasonal (D=0) differencing components.

Using backshift operator, $B^d x_t = x_{t-d}$, model (4.1) can thus be written as:

$$(1 - \alpha_1 B - \alpha_2 B^2)(1 - a_1 B^s - a_2 B^{2s})x_t = \mu + z_t$$

$$(1 - a_1 B^s - a_2 B^{2s} - \alpha_1 B + \alpha_1 a_1 B^{s+1} + \alpha_1 a_2 B^{2s+1} - \alpha_2 B^2 + \alpha_2 a_1 B^{s+2} + \alpha_2 a_2 B^{2s+2})x_t = \mu + z_t$$

$$x_t = \mu + a_1 x_{t-s} + a_2 x_{t-2s} + \alpha_1 x_{t-1} - \alpha_1 a_1 x_{t-s-1} - \alpha_1 a_2 x_{t-2s-1} + \alpha_2 x_{t-2} - \alpha_2 a_1 x_{t-s-2} - \alpha_2 a_2 x_{t-2s-2} + z_t \quad (4.2)$$

The model parameters for SARIMA $(2,0,0)(2,0,0)_{12}$ shown in **Table 4.1** are as follows:

α_1 =ar1 = 0.5949 (Nonseasonal AR parameter); α_2 =ar2 = 0.2247 (Nonseasonal AR parameter); a_1 = sar1 = -0.2846 (Seasonal AR parameter); a_2 = sar2 = -0.3114 (Seasonal AR parameter); μ = 0.1272 (Mean/Constant in the model); and s = 12 (The length of seasonal cycle). Thus, replacing model parameters in equation (4.2), we obtain the fitted model for SARIMA $(2,0,0)(2,0,0)_{12}$ as follows:

$$x_t = 0.1272 + 0.5949x_{t-1} + 0.2247x_{t-2} + 0.0639x_{t-10} + 0.1693x_{t-11} - 0.2846x_{t-12} + 0.0700x_{t-22} + 0.1853x_{t-23} - 0.3114x_{t-24} + z_t \quad (4.3)$$

Table 4.1: ***AIC for suggested models***

SARIMA Model	**Coefficient**	**Parameter estimates**	**AIC**
SARIMA $(2,0,0)(2,0,0)_{12}$	ar1	0.5949	-586.10
	ar2	0.2247	
	sar1	-0.2846	
	sar2	-0.3114	
	Mean	0.1272	
SARIMA $(2,0,0)(2,1,0)_{12}$	ar1	0.5825	-451.05
	ar2	0.1868	
	sar1	-0.6698	
	sar2	-0.4451	
SARIMA $(2,1,0)(2,0,0)_{12}$	ar1	-0.3778	-579.17
	ar2	-0.1822	
	sar1	-0.3097	
	sar2	-0.3100	
SARIMA $(2,1,0)(2,1,0)_{12}$	ar1	-0.3620	-439.8
	ar2	-0.1972	
	sar1	-0.6918	
	sar2	-0.4518	

4.2.3. Model Diagnostics

The model diagnostic checking was used to check appropriateness of the model. The Inverse AR roots of the SARIMA $(2,0,0)(2,0,0)_{12}$ model lie inside the circle as indicated in **Figure 4.9** hence the model is stationary.

The autocorrelation test for residuals was performed to check whether the residuals followed a white noise process as indicated by **Figure 4.10.** The standardized residuals plot in the first panel of the plot shows no obvious trend, have zero mean and constant variance since they are concentrated around -2 to 2. The diagnostics of residuals plot of the ACF in the second panel of the plot shows no evidence of autocorrelation in the residuals, hence residuals assume mean of zero and constant variance hence

uncorrelated. Since there are no spikes outside the insignificant zone for ACF plot, we can further conclude that residuals are random. The diagnostics of the time plot of the Ljung- Box statistics plot in the third panel indicate that all p-values exceed 5% for all lag orders hence not significant at any positive lag, thus residuals follow a white noise process. We therefore conclude that suggested SARIMA $(2,0,0)(2,0,0)_{12}$ satisfies all model assumptions and provides adequate fit to the data.

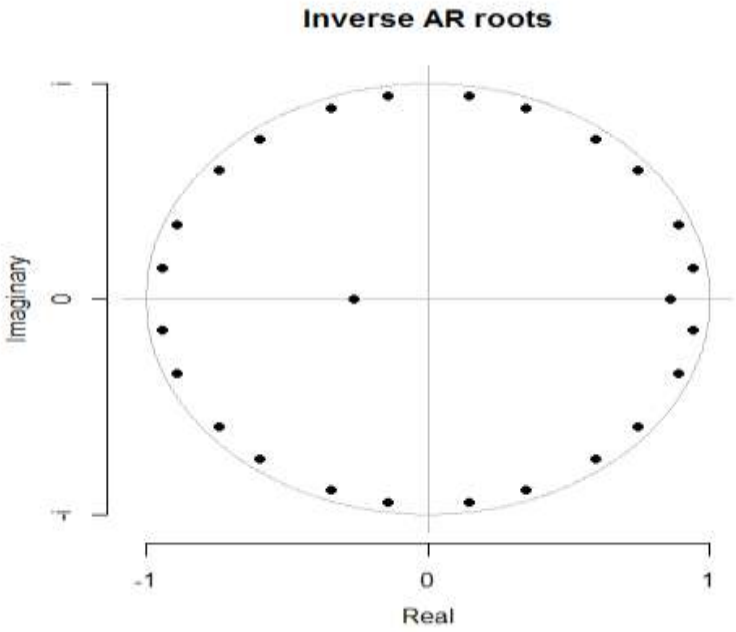

Figure 4.9: Roots of SARIMA $(2,0,0)(2,0,0)_{12}$ Model

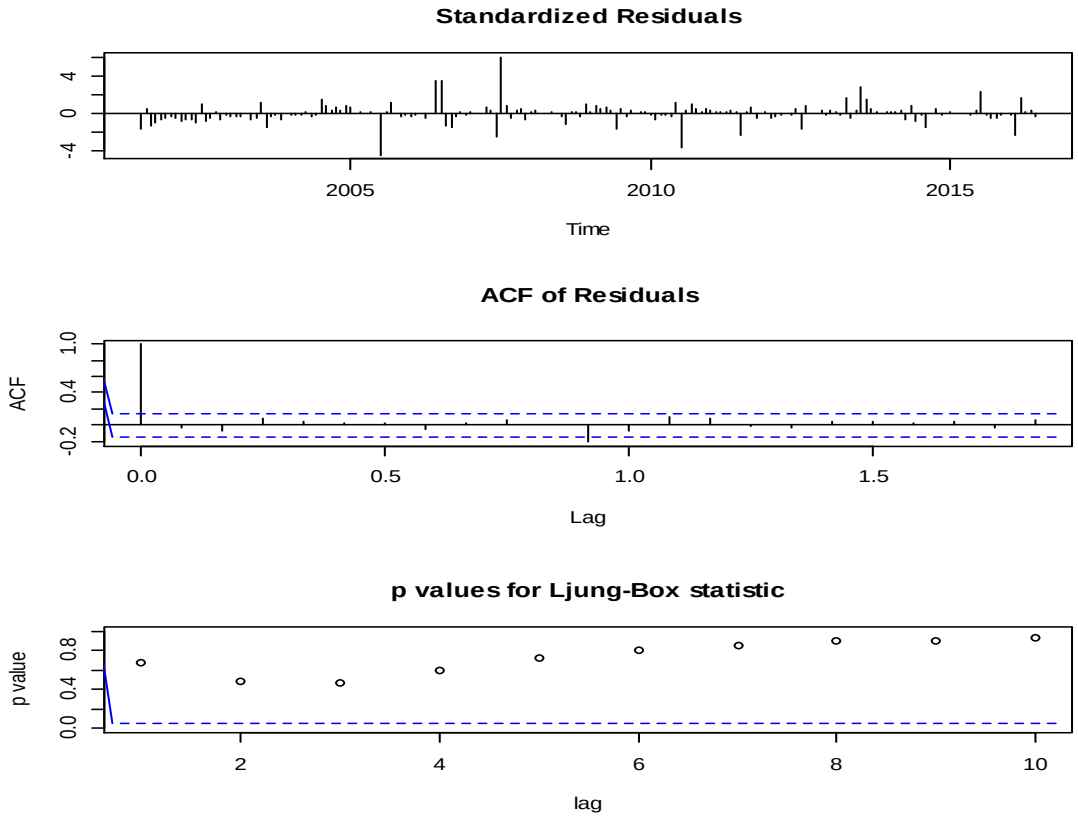

Figure 4.10: Plots of SARIMA $(2,0,0)(2,0,0)_{12}$ model residuals

4.3. Asymptotic properties of SARIMA $(2,0,0)(2,0,0)_{12}$ model

4.3.1.Consistency

The accuracy and consistency of the four suggested models was evaluated based on performance statistics of forecasting errors, namely the root mean square error (RMSE) and mean absolute error (MAE) as indicated in **Table 4.2**. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction while RMSE measures differences between values predicted by the model and values observed. Both measures compare forecasts for the same series across different models, the smaller the error the better for forecasting ability of that model. SARIMA $(2,0,0)(2,0,0)_{12}$ model performed well compared to other models with RMSE=0.04544746 and MAE=0.02665831 hence was identified as better model for forecasting monthly tax revenue in Kenya.

Table 4.2: ***Performance Statistics of suggested models***

SARIMA Model	RMSE	MAE
SARIMA $(2,0,0)(2,0,0)_{12}$	0.04544746	0.02665831
SARIMA $(2,0,0)(2,1,0)_{12}$	0.05772624	0.03324039
SARIMA $(2,1,0)(2,0,0)_{12}$	0.04611399	0.02636878
SARIMA $(2,1,0)(2,1,0)_{12}$	0.05772624	0.03178921

4.3.2.Normality

The histogram plot of the model residuals and normal Q-Q plot of standardized residuals are illustrated in **Figure 4.11** and **Figure 4.12**, respectively. The distribution of the residuals follows a normal distribution as illustrated in the histogram. The majority of the residuals as illustrated in the normal Q-Q plot are located on the straight line except some few outliers. This indicates that residuals are normal. Therefore, SARIMA $(2,0,0)(2,0,0)_{12}$ model satisfies all model assumptions hence good for forecasting.

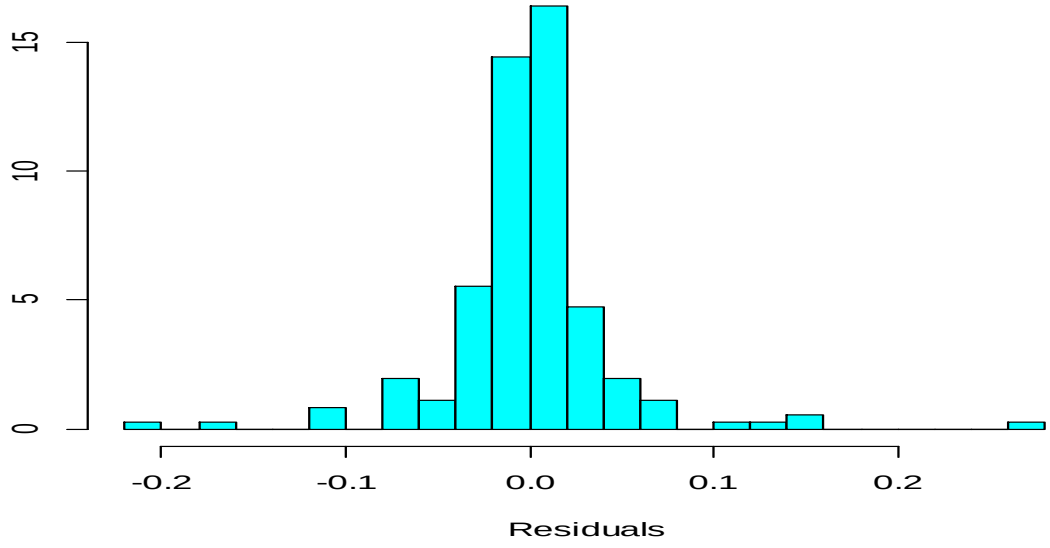

Figure 4.11: Histogram plot of residuals

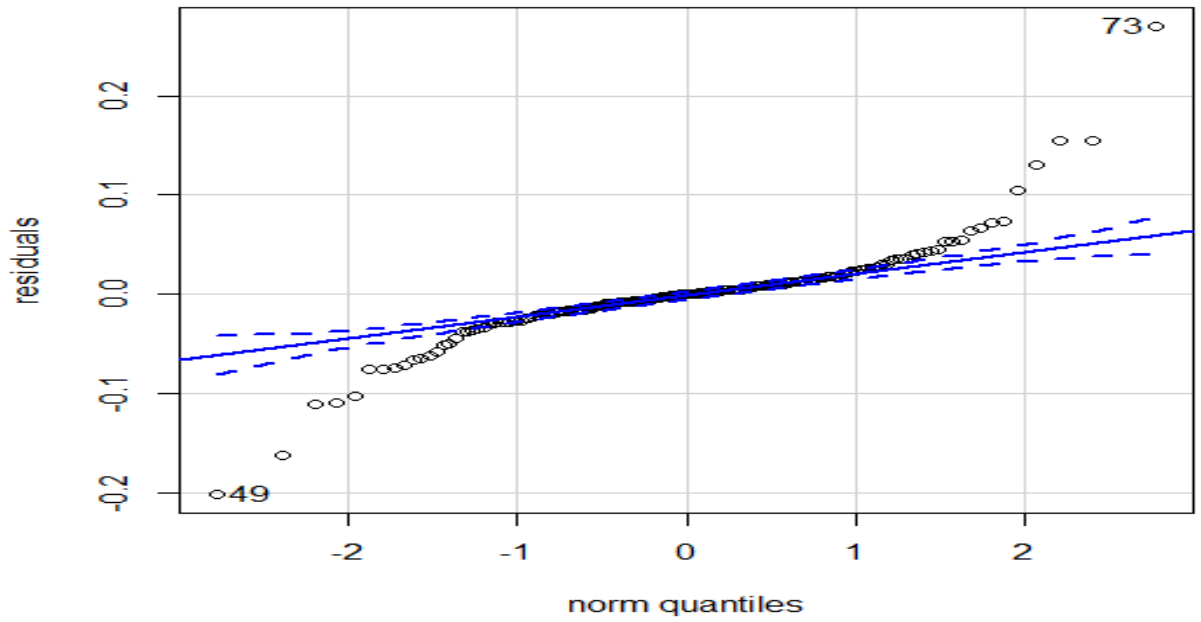

Figure 4.12: Normal Q-Q Plot of standardized residuals

4.4. Simulation of SARIMA $(2,0,0)(2,0,0)_{12}$ model

Monte Carlo simulation was performed on the SARIMA $(2,0,0)(2,0,0)_{12}$ model as illustrated in **Figure 4.13.** The plots of stationary original time series model (continuous line) and simulated time series model (dotted line) produced similar patterns and moved in the same direction implying SARIMA $(2,0,0)(2,0,0)_{12}$ model is appropriate for tax revenue forecasting.

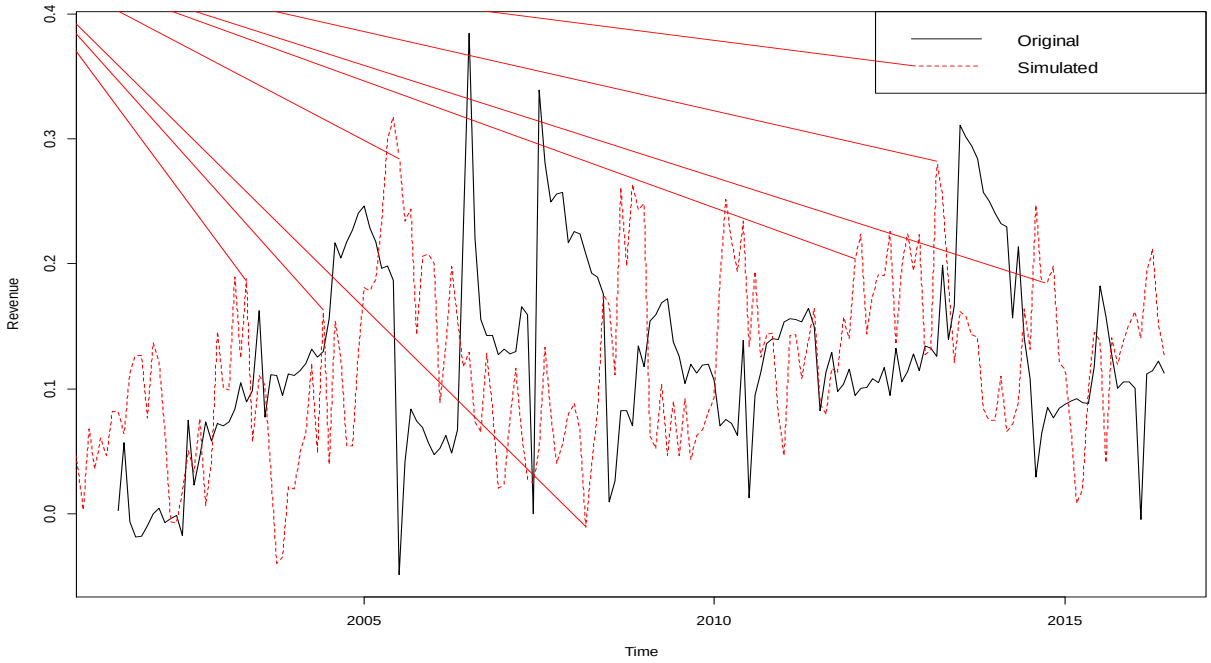

Figure 4.13: Original and Simulated models

4.5. Forecasting using SARIMA $(2,0,0)(2,0,0)_{12}$ model

The SARIMA $(2,0,0)(2,0,0)_{12}$ model was used to forecast monthly tax revenue collections in Kenya for the next five years (July 2016 – June 2021) and results are illustrated in **Figure 4.14.** The light blue line (continuous line within the shaded region) shows the forecast estimates provided by the SARIMA $(2,0,0)(2,0,0)_{12}$ model at 80% confidence limits (shaded in darker blue) and 95% confidence limits (shaded in lighter blue). Based on the forecasts, it can be deduced that monthly tax revenue collections in Kenya will exhibit the same patterns with no significant growth in the next 5 years.

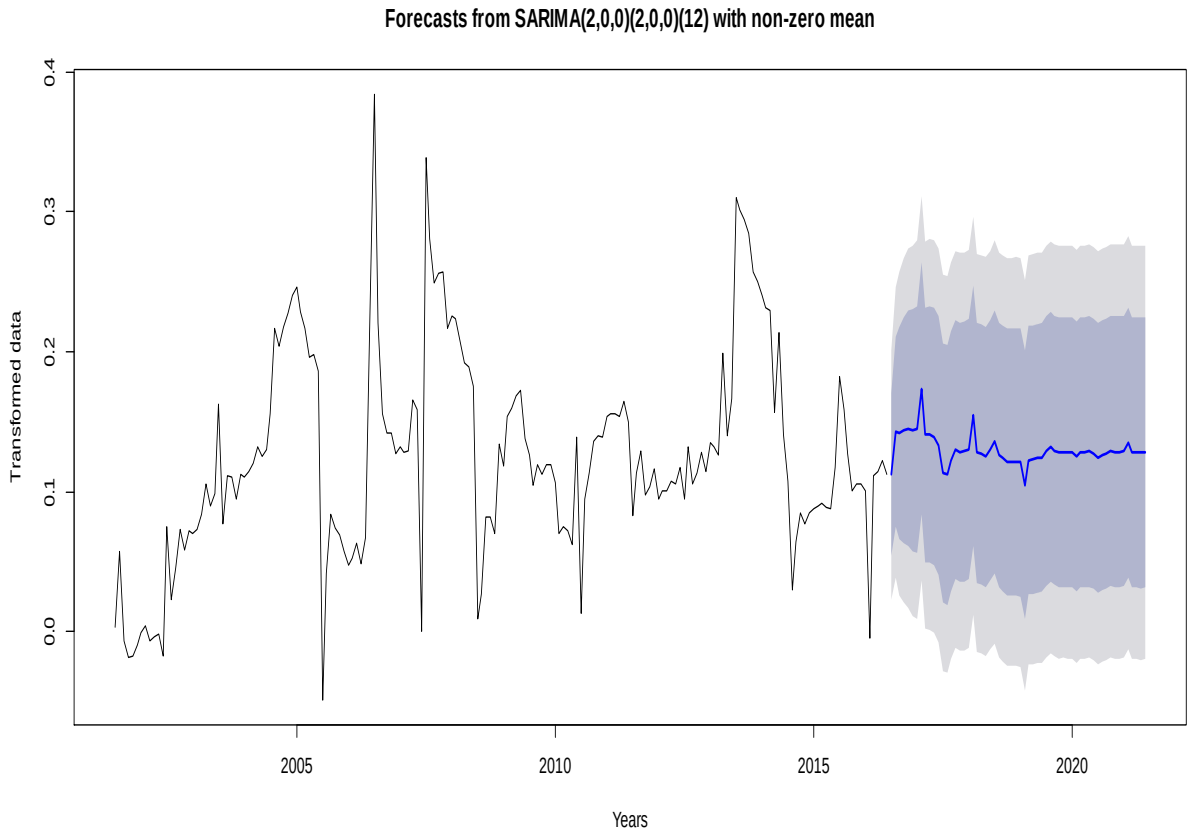

Figure 4.14: Plot of tax revenue, forecasts and confidence intervals

5. Conclusion and Recommendations

The study sought to establish SARIMA model that can be used in tax revenue forecast in Kenya. Several SARIMA models were suggested based on the inspection ACF and PACF plots of stationary time series. SARIMA $(2,0,0)(2,0,0)_{12}$ was considered as appropriate model for forecasting monthly tax revenue in Kenya based on the minimum AIC value The diagnostics of the model residuals were found to follow a white noise process with zero mean and constant variance, hence uncorrelated.

The accuracy and consistency of SARIMA $(2,0,0)(2,0,0)_{12}$ model in tax revenue forecast was ascertained through evaluation of the performance statistics of forecasting errors, namely RMSE and MAE. The normality test was performed through inspection of histogram and normal Q-Q plots of standardized residuals which reviewed that model residuals where normally distributed hence the model was adequate and fit for forecasting.

The Monte Carlo simulation of tax revenue using SARIMA $(2,0,0)(2,0,0)_{12}$ model produced was similar plots for both stationary original time series model and simulated time series model hence SARIMA $(2,0,0)(2,0,0)_{12}$ model was appropriate for tax revenue forecasting.
The tax forecasts for the next five years using SARIMA $(2,0,0)(2,0,0)_{12}$ model reviewed that monthly tax revenue collections in Kenya will exhibit the same patterns with no significant growth.
Based on the findings, the study recommends application of SARIMA $(2,0,0)(2,0,0)_{12}$ model in tax revenue forecast in Kenya and enhancing tax revenue collections. Further study on modelling tax revenue in presence of other macro-economic variables such as inflation, public expenditure, public debt and exchange rates should be explored.

References

[1] Box, G. E. P., and Jenkins, G. M. (1976). Time Series Analysis; Forecasting and Control, Holden-Day Inc. U.S.A.

[2] Fazidah A. S., Tri M., & Saprin, S.(2018). Forecasting dengue hemorrhagic fever cases using ARIMA model: a case study in Asahan district. *4th International Conference on Operational Research (InteriOR). IOP Publishing*

[3] Hyman, D. N. (1987). Public Finance: A contemporary Application of Theory of Policy (2nd ed.). Chicago: The Dryden Press.

[4] Institute of Economic Affairs. (2012).Tax Incentives and Exemption Regime in Kenya: *Is It Working*? Issue No. 30

[5] International Monetary Fund (2011). Revenue Mobilization in Developing in Developing Countries. [Online] Available: https://www.imf.org/external/np/pp/eng/2011/030811.pdf

[6] Kenya Institute for Public Policy Research and Analysis (2016). Kenya Economic Report 2016: Fiscal Decentralization in support of Devolution. Nairobi: KIPPRA

[7] Kenya National Bureau of Statistics (2016). Economic Survey. Nairobi: KNBS

[8] Kibunja H., Kihoro J., Orwa G., and Yodah W.(2014), "Forecasting Precipitation Using SARIMA Model: A Case Study of Mt. Kenya Region", International Institute for Science, Technology and Education, 4(11), 50-58

[9] Lilian, O. (2015). Kenya loses over Sh600bn every year in tax evasion. [Online] Available: https://www.nation.co.ke/business/Kenya-loses-over-Sh600bn-every-year-in-tax-evasion/996-2724818-5jswbtz/index.html

[10] Michael (2014). Using SARIMA to Forecast Monthly Mean Surface Air Temperature in the Ashanti Region of Ghana. *International Journal of Statistics and Applications 2014, 4(6): 292-299*

[11] Moyi, E., & Ronge, E., (2006). Taxation and Tax Modernization in Kenya: A Diagnosis of Performance and Options for Further Reforms, *Institute of Economic Affairs*

[12] Otu, A. O., Osuji G, A., Opara, J., Mbachu, H. I., & Iheagwara, A. I. (2014) "Application of Sarima Models in Modelling and Forecasting Nigeria's Inflation Rates." *American Journal of Applied Mathematics and Statistics 2, no. 1 (2014): 16-28. doi: 10.12691/ajams-2-1-4.*

[13] Shengwei, W., Juan F. & Gang L. (2011). Application of seasonal time series model in the precipitation forecast. Elsevier Ltd.

[14] Susan, W. G., Anthony, G. W., & John, M. K (2015). Forecasting Inflation Rate in Kenya Using SARIMA Model. *American Journal of Theoretical and Applied Statistics. Vol. 4, No. 1, 2015, pp. 15-18. doi: 10.11648/j.ajtas.20150401.13*

[15] Tadesse K.B., Dinka M.O. (2017). Application of SARIMA model to forecasting monthly flows in Waterval River, South Africa. Journal of Water and Land Development. No. 35 p. 229–236. DOI: 10.1515/jwld-2017-0088.

[16] Todaro, P. M. & Smith, C. S. (2003). *Economic development*. Eighth Edition. Addison Wesley