International Journal of Mathematics and Statistics Studies, 13 (2), 1-20, 2025
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Time Series Modelling of Yearly Cassava Production in Nigeria: A Comparative Study

[1] Uwem Paul Abraham and [2]Emmanuel Wilfred Okereke
[1]Department of Statistics, Akwa Ibom State Polytechnic Ikot Osurua, Ikot Ekpene, Akwa Ibom State

[2]Department of Statistics, Michael Okpara University of Agriculture, Umudike

doi: https://doi.org/10.37745/ijmss.13/vol13n2120 Published May 03, 2025

Citation: Abraham U.P. and Okereke E.W. (2025) Time Series Modelling of Yearly Cassava Production in Nigeria: A Comparative Study, *International Journal of Mathematics and Statistics Studies*, 13 (2), 1-20

Abstract: *Cassava production is an important agricultural activity in Nigeria, as it contributes to the GDP of the polity. Appropriate prediction of cassava production in the nation Nigeria is fundamental to the development of a long-term plan to sustain agricultural productivity and promote food security. This study investigated in detail statistical characteristics of yearly cassava production in Nigeria over the period 1961 to 2022 with a view to choosing a befitting model for the data. The data set was divided into training set and test set. By virtue of ADF test, the training set was found to be nonstationary. The Zivot-Andrew test revealed the presence of a structural break in the data. The break date was found to be 1990. Holt's linear model with multiplicative errors, ARIMA (1,1,2) model and SETAR (2,2,1) model were fitted to the training set following their automatic selection using ets, auto.arima and Selectsetar functions in the forecast and tsDyn packages in R. The out of sample comparison of the three models based on their associated root mean squared errors (RMSEs) and mean absolute percentage errors (MAPEs) provided the evidence of the SETAR mode having the smallest RMSE and MAPE values. Hence, SETAR (2,2,1) model is the best for forecasting annual cassava production in Nigeria among the three models.*

Keywords: Cassava production, food security, Holt's linear model with multiplicative errors, ARIMA model, out of sample comparison, SETAR model

INTRODUCTION

Cassava is important, not only as a food crop but even more so as a major source of income for rural households. Nigeria is currently the largest producer of cassava in the world with an annual production of over 34 million tonnes of tuberous roots. Cassava is largely consumed in many processed forms in Nigeria.

The demand for cassava roots and products is high and fast rising. However, the current food production is far from being able to meet the food needs of the geometrically growing population

in the sub-region (Food and Agricultural Organisation (FAO), 2018). It provides a strong incentive for more economic agents to be involved in the cassava market. According to FAO (2018), cassava is a choice crop for rural development, poverty alleviation, economic growth and ultimately, food security. More importantly, proceeds from cassava like starch, is increasingly demanded in developed world making cassava production a more sustainable source of foreign income and a great contributor to the national economy. Local processing of cassava has created jobs for many rural women and the local fabricators and thus, has significantly stimulated the rural economy in SSA. Similarly, it has also influenced the agricultural input supply market. Therefore, it contributes to capital formation and securing markets for the agro-industry in Nigeria. However, whether or not, the present cassava production (supply) can meet the increasing demand for cassava as food and industrial use remains a serious concern.

The increasing importance of cassava (*Manihot esculenta*) among crops grown in Nigeria is not only connected to its increasing demand as food but also as food security (FAO, 2018). Cassava products are dietary staple food in Nigeria and other countries in SSA. Nigeria is populated with about 200 million people, and 7 in every 10 Nigerians consume, at least, a product of cassava once in a day (Njoku and Muoneke, 2008). These products include: cassava flakes (garri), cassava flour (pupuru and lafun), cassava paste (fufu) which are derived from cassava roots. Ukwuru and Egbonu (2013) elucidated emerging processed products from cassava to include, but not limited to, cassava paste, starch, ethanol, biofuels, flour, paper, adhesives, glucose, cassava chips, pies, cookies, noodles, flakes and cakes. It is a widely acceptable energy food source to over 600 million consumers of cassava across the globe (Hershey et al., 2001; and FAO, 2015).

The Nigeria government has on several occasions introduced policies and initiatives with a view to enhancing cassava production, ensuring sustainable growth, and integrating cassava more effectively into industrial sectors. For example, the Presidential Cassava Initiative (PCI) was launched in cassava-producing countries in West Africa, including Nigeria in 2001. The programme which lasted for six years (2001–2007) aimed at enhancing the productivity and production of cassava by increasing the area cultivated to 5 million ha, with the hope of harvesting 150 million tons of fresh cassava tubers annually, producing 37.5 million tons of processed cassava products for the local and export markets, organizing the export of cassava and processed cassava products as a revenue-generating project and generating about 5 billion dollars annually from exporting value-added cassava products (Sanogo & Adetunji, 2008; Donkor et al., 2017). An extensive review of the presidential interventions on cassava over the period 2002 to 2012 is available in Ohimain (2015). Evidence across states in Nigeria shows that government investments and intervention to enhance cassava production have resulted to increased output and also stimulated the rural economy (Okhankhuele, et al. 2017; Ugbem-Onah and Mbakuuv, 2024).

Yearly cassava production for a reasonable number of years constitutes a time series. Several time series modelling approaches have been developed for modelling time series data across walks of life. In the literature, few studies have been done on modelling of Nigeria's cassava production. The most recent one is the work of Oni and Akanle (2018), which compared the performance of

various exponential smoothing models. Their study recommended the use of the Holt's Exponential Trend model to forecast future cassava production in Nigeria.

Forecasting cassava production using time series models is essential for informed decision-making and the sustainable development of the agricultural sector. Suffice it to say that a time series model can only be used to make reliable forecasts of yearly cassava production data in a particular country if it takes into consideration the necessary properties of the data. Until now, studies on time series modelling of cassava production in Nigeria are limited to ARIMA models and exponential smoothing procedures (Oni and Akanle, 2018; Omoluabi and Ibitoye, 2024).

Policy change is a known factor that is responsible for the presence of structural breaks in time series (Çamalan et al., 2024). With the series of policies made by government of Nigeria to boost cassava production, it is necessary to perform a test for structural break in the series. Again, time series with potential breaks often exhibit nonlinear behavior. The imposition of a linear time series model on a time series with nonlinear features implies model misspecification and inadequacy. As a consequence, this study intends to improve on the existing Nigeria yearly cassava forecasting models by investigating more properties of the concerned series for the purpose of building a suitable model for forecasting the series. Thus, the object of this paper is to determine an adequate model for forecasting yearly cassava production in Nigeria. The remaining component of the series is structured as follows. Section 2 contains a brief and succinct explanation of statistical procedures employed to analyze the cassava production data. Results obtained in the course of analyzing the data are enshrined in Section 3 while the conclusion of the study is found in Section 4.

METHODOLOGY

This section deals with the source of the data analyzed in this paper as well as the statistical methods employed in the analysis of the data. The data set used in this study is the yearly cassava production in Nigeria obtained from Food and Agriculture Organization Statistical Database and it covers a period of sixty-one years (1961 – 2022). Statistical tests, namely the augumented Dickey-Fuller (ADF) test, BDS test and Zivot-Andrew test, which were employed in the analysis of the data are discussed in this section. The detail of each of the Holt linear trend model with additive error, ARIMA and SETAR models is also provided below.

The ADF Test

This test is widely utilized to test for stationarity of a time series. In the test, the null hypothesis H_0: the series is not stationary (i.e the series has a unit root) and the alternative hypothesis H_1: the series is stationary (i.e it does not have a unit root). The test equation has the form

$$\Delta x_t = \alpha + \lambda t + \delta x_{t-1} + \sum_{i=1}^{p} \gamma \Delta x_{t-i} + \varepsilon_t , \qquad (2.1)$$

where x_t is the time series. $\Delta x_t = x_t - x_{t-1}$ is the first difference of the time series, α is a constant (drift), λ represents a time trend (optional depending on the test), δ is the coefficient of the lagged value of the series, ε_t is the error term and p is the number of lags included in the model.

At α level of significance, we reject the null hypothesis and conclude that the series is stationary if the associated p-value is greater than α . Otherwise, we infer that the series is not stationary. α

Zivot-Andrews Test

The Zivot-Andrews test (Zivot and Andrews, 1992) is useful in detecting a structural break in a univariate time series model where the break data is not known a priori. It involves testing for presence of a break in the intercept and or trend of the data. The test can be performed using any of Model I, Model II and Model III. The three models are defined in Equations (2.2), (2.3) and (2.4) respectively.

$$x_t = \alpha + \beta t + \gamma_1 D_t + \epsilon_t \quad (2.2)$$

$$x_t = \alpha + \beta t + \gamma_2 D_t t + \epsilon_t \quad (2.3)$$

$$x_t = \alpha + \beta t + \gamma_3 D_t + \delta D_t t + \epsilon_t \quad (2.4)$$

In the three equations above, x_t=the value of the time series at time t, α is the intercept, β is the slope (trend) of the time series, t is the time variable (trend term), D_t is a dummy variable that takes the value 0 before the break and 1 after the break, γ_1 is the coefficient for the structural break (the change in the intercept after the break), γ_2 is the coefficient for the structural break in the trend, $D_t t$ is the interaction term that allows for a change in the slope of the trend after the break, γ_3 is the coefficient for the structural break in the intercept (level shift), δ is the coefficient for the structural break in the trend (slope shift) and ϵ_t.

Holt's Linear Trend Method with Multiplicative Errors

In this double exponential smoothing method, the error term is defined as

$$e_t = \frac{x_t - (l_{t-1} + b_{t-1})}{(l_{t-1} + b_{t-1})}.$$

The innovation state space model corresponding to the Holt's linear method with multiplicative errors is given as

$$X_t = (l_{t-1} + b_{t-1})(1 + e_t);$$

$$l_t = (l_{t-1} + b_{t-1})(1 + \alpha e_t);$$

$$b_t = b_{t-1} + \beta (l_{t-1} + b_{t-1}) e_t,$$

where X_t is the actual observation at time t, l_t is an estimate of the level of the series at time t, b_t an estimate of the trend of the series at time t, α is the smoothing parameter for the level, $0 < \alpha < 1$, γ is the smoothing parameter for the trend, $0 < \gamma < 1$ and $\beta = \alpha\gamma$. Additionally, $e_t \sim NID(0, \sigma^2)$.

The tsDyn package in R can be used to fit the model using the ets() function.

Autoregressive Integrated Moving Average Model Building

A nonstationary time series is said to follow an autoregressive integrated moving average (ARIMA) model of order p, d and q if its dth difference follows an autoregressive moving average (ARMA) model of order p and q. An ARIMA(p,d,q) process is defined by Wei (2000) as

$$\phi_{\mathrm{p}}(\mathrm{B})(1-\mathrm{B})^{\mathrm{d}}\,\mathrm{X}_{\mathrm{t}} = \mu + \theta_q(\mathrm{B})\mathrm{e}_{\mathrm{t}},$$

where B is the backshift operator, μ is a constant, $\phi_{\mathrm{p}}(\mathrm{B}) = 1-\phi_1\mathrm{B}-\phi_2\mathrm{B}^2-\ldots--\phi_{\mathrm{p}}\mathrm{B}^{\mathrm{p}}$ is the autoregressive characteristic polynomial in B of degree p, $\theta_{\mathrm{q}}(\mathrm{B}) = 1-\theta_1\mathrm{B}-\theta_2\mathrm{B}^2-\ldots--\theta_{\mathrm{q}}\mathrm{B}^{\mathrm{q}}$ is the moving average characteristic polynomial in B of degree q. The two characteristic polynomials are not expected to have common factors.

The autoregressive integrated moving average (ARIMA) model. Three stages are critical to autoregressive integrated moving average (ARIMA) model building. They include model identification, estimation, diagnostic checking and forecasting.

Model Identification

An easy and reliable way of identifying an ARIMA process is concerned with the examination of the ACF and PACF of the dth difference of a time series. Mathematically, the sample autocorrelation function at lag k is given by

$$r_k = \frac{\sum_{\mathrm{t}=1}^{\mathrm{n}-\mathrm{k}}\left(x_{\mathrm{t}+k}-\bar{x}\right)\left(x_{\mathrm{t}}-\bar{x}\right)}{\sum_{\mathrm{t}=1}^{\mathrm{n}}\left(x_{\mathrm{t}}-\bar{x}\right)^2}. \tag{2.5}$$

The sample partial autocorrelation coefficient at lag k can be computed using the recursive procedure proposed by Durbin (1960). The method depends on the following equations:

$$\hat{\phi}_{\mathrm{k}+1,\mathrm{k}+1} = \frac{r_{\mathrm{k}+1}-\sum_{\mathrm{j}=1}^{\mathrm{k}}\hat{\phi}_{\mathrm{kj}}r_{\mathrm{k}+1-\mathrm{j}}}{1-\sum_{\mathrm{j}=1}^{\mathrm{k}}\hat{\phi}_{\mathrm{kj}}r_{\mathrm{j}}} \tag{2.6}$$

and

$$\hat{\phi}_{\mathrm{k}+1,\mathrm{j}} = \hat{\phi}_{k\mathrm{j}} - \hat{\phi}_{\mathrm{k}+1,\mathrm{k}+1}\hat{\phi}_{\mathrm{k},\mathrm{k}+1-\mathrm{j}},\ \mathrm{j}=1,2,\ldots,\mathrm{k}.$$

(2.7)

If the ACF and PACF of the differenced series behaves like those of an ARMA process, then the original series has an ARIMA data generating process. The structure of the ACF and PACF of an ARMA(p,q) process is well-known and document in the literature. According to Abraham and Ledolter (1983), the ACF of the model decays exponentially, starting from the (q+1)th autocorrelation coefficient and its PACF decays geometrically, starting with the (p+1)th partial autocorrelation coefficient. In situations where the sample ACF and PACF do not mimic those of a known ARMA process, the Akaike information can be used to select the best among several ARMA models can be fitted to the data. Here, the model with the minimum AIC value is chosen.

It is worthy of note that the forecast package in R can be employed in the automatic selection of the best ARIMA model for a time series data set.

Model Estimation

Once the order of an ARIMA (p,d,q) model has been determined, it is expedient to estimate the parameters of the model. To obtain the conditional maximum likelihood estimators of the parameter of the model, let

$$W_t = (1-B)^d X_t. \tag{2.8}$$

Consider the general ARMA (p,q) model

$$\dot{W}_t = \phi_1 \dot{W}_{t-1} + \phi_2 \dot{W}_{t-2} + \ldots + \phi_p \dot{W}_{t-p} + e_t - \theta_1 e_{t-1} \ldots - \theta_q e_{t-q}, \tag{2.9}$$

where $\dot{W}_t = W_t - \mu$ and $\{e_t\}$ is the white noise process. The joint density of $\mathbf{e} = (e_1, .., e_n)'$ is

$$P(\mathbf{e} \mid \boldsymbol{\varphi}, \mu, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\pi\sigma^2} \sum_{t=1}^{n} e_t^{\,2} \right),$$

where $\boldsymbol{\varphi} = (\phi_1, .., \phi_p)'$ and $\boldsymbol{\theta} = (\theta_1, .., \theta_q)'$. From Equation (2.9),

$$e_t = \dot{W}_t + \theta_1 e_{t-1} \ldots + \theta_q e_{t-q} - \phi_1 \dot{W}_{t-1} - \phi_2 \dot{W}_{t-2} - \ldots - \phi_p \dot{W}_{t-p}. \tag{2.10}$$

Let $\mathbf{W} = (W_1, ..., W_n)'$. Suppose that the initial conditions $\mathbf{W}_* = (W_{1-p}, ..., W_{-1}, W_0)'$ and $\mathbf{e}_* = (e_{1-q}, ..., e_{-1}, e_0)'$ are known. The conditional log-likelihood function becomes

$$\ln L_* (\boldsymbol{\varphi}, \mu, \boldsymbol{\theta}, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{\mathbf{S}_* (\boldsymbol{\varphi}, \mu, \boldsymbol{\theta})}{2\sigma^2}, \tag{2.11}$$

where

$$\mathbf{S}_* (\boldsymbol{\varphi}, \mu, \boldsymbol{\theta}) = \sum_{t=1}^{n} e_t^{\,2} (\boldsymbol{\varphi}, \mu, \boldsymbol{\theta} \mid \mathbf{W}_*, \mathbf{e}_*, \mathbf{W}) \tag{2.12}$$

is the conditional sum of sum squares function.

The quantities $\hat{\boldsymbol{\varphi}}$, $\hat{\mu}$ and $\hat{\theta}$, which minimize the conditional log-likelihood function are called the conditional maximum likelihood estimators. These estimators are the same as the conditional least squares estimators obtained by minimizing $\mathbf{S}_* (\boldsymbol{\varphi}, \mu, \boldsymbol{\theta})$, which does not depend on σ^2. Initial conditions for $\mathbf{W}_*$ and $\mathbf{e}_*$ are obtained by replacing unknown W_t with $\overline{W}$ and e_t with its expectation 0. Also, we assume that $e_p = e_{p-1} \ldots = e_{p+1-q} = 0$ and calculate e_t for $t \geq (p+1)$ using Equation (2.10).

Consequently, the conditional sum of squares in Equation (2.12) becomes

$$\mathbf{S}_*\left(\boldsymbol{\varphi},\mu,\boldsymbol{\theta}\right)=\sum_{t=1}^{n}e_t^{\,2}\left(\boldsymbol{\varphi},\mu,\boldsymbol{\theta}\,|\,\mathbf{W}\right). \tag{2.13}$$

After determining the estimators $\hat{\boldsymbol{\varphi}}$, $\hat{\mu}$ and $\hat{\theta}$, the estimator of σ^2 is calculated from

$$\hat{\sigma}^2=\frac{\mathbf{S}_*\left(\boldsymbol{\varphi},\mu,\boldsymbol{\theta}\right)}{df}.$$

If Equation (2.13) used to obtain the conditional sum of squares, $df=n-\left(2p+q+1\right)$.

Diagnostic Checks

Once the parameters of a tentatively entertained ARIMA model have been estimated, it is worthwhile to investigate the adequacy of the fitted model. Several time series model adequacy checking tools are available in the literature. The correlogram and partial correlogram of the residuals from the fitted model are adopted among the tools. If the fitted model is adequate, the associated residuals should behave like white noise. The need to fit a nonlinear time series model to the data under consideration will be substantiated through the application of the BDS test.

Forecasting

Forecasting is an important objective of time series analysis. In forecasting, we usually wish to obtain an optimum forecast that has little or no error, leading to the minimum mean squared error forecast. The minimum mean squared error forecasts $\hat{X}_n(l)$ of X_{n+l} at forecast origin n is the conditional expectation

$$\hat{X}_n(l)=E\left(X_{n+l}\,|X_n,X_{n-1},...\right).$$

For the ARIMA(p,d,q) model,

$$\Psi\left(B\right)=\phi\left(B\right)\left(1-B\right)^d=\left(1-\Psi B-...-\Psi_{p+q}B^{p+q}\right).$$

Hence, the general ARIMA(p,d,q) model can be written as the difference equation:

$$\left(1-\Psi B-...-\Psi_{p+q}B^{p+q}\right)X_t=\left(1-\theta B-...-\theta_q B^q\right)e_t.$$

Given that $t=n+l$, we have

$$X_{n+l}=\Psi_1X_{n+l-1}+\Psi_2X_{n+l-2}+...+\Psi_{p+q}X_{n+l-p+q}+e_{n+l}-\theta_1e_{n+l-1}-...-\theta_q e_{n+l-q}.$$

Taking expectations at time origin n leads to

$$\hat{X}_n\left(l\right)=\Psi_1\hat{X}_n\left(l-1\right)+\Psi_2\hat{X}_n\left(l-2\right)+...+\Psi_{p+q}\hat{X}_n\left(l-p-q\right)+\hat{e}_n\left(l\right)-\theta_1\hat{e}_n\left(l-1\right)-...-\theta_q\hat{e}_n\left(l-q\right),$$

where

$$\hat{X}_n(j) = E\left(X_{n+j} | X_n, X_{n-1}, \ldots\right), j \geq 1,$$

$$\hat{X}_n(j) = X_{n+j}, j \leq 0,$$

$$\hat{e}_n(j) = 0, j \geq 1$$

and

$$\hat{e}_n(j) = e_{n+j}, j \leq 0.$$

The SETAR Model

A two-regime self-exciting threshold autoregressive model, denoted by SETAR(2, p_1, p_2) model, can be written in the form:

$$X_t = \begin{cases} \alpha_0^{(1)} + \sum_{i=0}^{p_1} \alpha_i^{(1)} X_{t-i} + e_t^{(1)}, \text{if } X_{t-d} \leq r, \\ \alpha_0^{(2)} + \sum_{i=0}^{p_2} \alpha_i^{(2)} X_{t-i} + e_t^{(2)}, \text{if } X_{t-d} > r, \end{cases} \quad (2.14)$$

where p_1 and p_2 are the orders of the AR models in regime 1 and regime 2 respectively, $e_t^{(1)}$ and $e_t^{(2)}$ are white noise processes, d is the delay parameter and r is the threshold parameter. When the threshold value and threshold variable are fixed, the parameters of the SETAR model can be estimated via the conditional least squares method (Iquebal et al., 2013). The tsDyn package in R, introduced by Antonio et al. (2009) is useful in fitting SETAR models.

The appropriate number of regimes in a SETAR model is determined based on the likelihood ratio test of Hansen (1999). Let SETAR(m) denote a SETAR model with m regimes. Then SETAR(j) $\subset$ SETAR(k), for $j \leq k$. Again, the Hansen's test statistic for testing the null hypothesis of SETAR(j) against SETAR(k), for $j \leq k$ is

$$F_{jk} = n\frac{SSE_j - SSE_k}{SSE_k},$$

where SSE_m is the sum of squared residuals in fitting a SETAR(m) model by least squares method (Magadia, 2016). The test statistic has a non-standard distribution. However, its distribution has been approximated through a bootstrapping procedure (Hansen, 1999). Notably, the null hypothesis has to be rejected whenever the given level of significance is greater than the corresponding p-value.

Often, the threshold value is unknown and needs to be estimated alongside the other parameters in the model. A widely used procedure for finding the unknown threshold value, which is being referred to as the threshold grid search process (TGSP), was introduced by Chan (1993) and subsequently discussed in Enders (2004). In this approach, the threshold value is considered to be

one of the elements of the series itself. Hence, each element of the series is regarded as a potential threshold value and we fit a model to each value X_t. Only the middle 70 or 80% of the series is tested for the threshold so as to have a satisfactory amount of observations on each regime when estimating the threshold and the other parameters in the TAR model.
Thereafter, we specify the regimes and estimate a model in each regime by least squares regression. For each potential threshold, the residual sum of squares is computed.
The threshold value that is associated with the model with the least sum of squared residuals is taken to be the optimal threshold value.

Evaluation Criteria

The process of choosing the best model for forecasting yearly cassava production data in Nigeria begins with the selection of the best model from each of the exponential smoothing class of models, the class of ARIMA models and the class of SETAR models using the Akaike information criteria (AIC). Mathematically,

$$\text{AIC} = -2l + 2\text{k},$$

where l and k refer to the maximized log-likelihood function and number of parameters being estimated respectively.
Suppose that a time series is split into a training set and test set. Let T and H represent the numbers of observations belonging to the training set and test set respectively. Then the out of sample comparison of the fitted models can be done using the following:

(i) Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{h=1}^{H}\left(X_{T+h} - \hat{X}_T(h)\right)^2}{H}}. \qquad (2.15)$$

(ii) Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100}{H}\sum_{h=1}^{H}\left|\frac{X_{T+h} - \hat{X}_T(h)}{X_{T+h}}\right|. \qquad (2.16)$$

Data Analysis

Empirical results based on the methods discussed in Chapter 2 and data on yearly cassava production in Nigeria are presented in this chapter. It is noteworthy that the data are divided into two parts, namely the train data, which comprises observations spanning the period 1961 to 2009 and the test data containing the remaining observations. Figure 3.1 is the time plot of the data. The train data are plotted in blue while the test are plotted in red. The graph reveals the possibility of the series being nonstationary and possessing a structural break in the trained series.

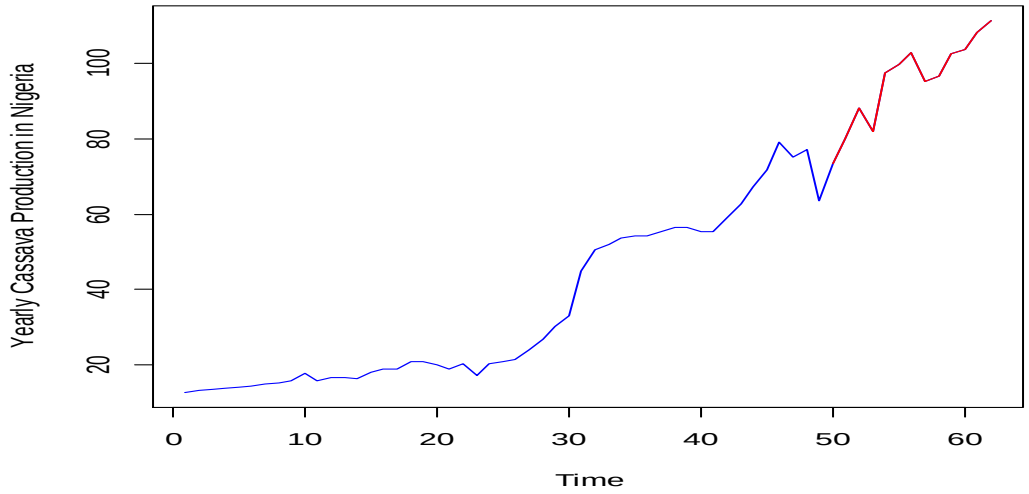

Figure 3.1: Time Plot of Yearly Cassava Production in Nigeria

ADF Unit Test Result

Table 3.1 contains the ADF unit test results on the trained data. In the table, the p-value of 0.6582 exceeds 0.05. Hence, we do not reject the null hypothesis of nonstationarity of the trained data at 5% level of significance.

Table 3.1: ADF test result on the trained data

Test Statistic	p-value
-1.7901	0.6582

Zivot-Andrew Test Result

Zivot-Andrew test results pertaining to the train data are presented in Table 3.2. It can be deduced from the table that the coefficient for the lagged dependent variable (y.l1), trend and dummy variable for the structural break are all significant at 5% significance level. However, the coefficient for the difference in the lagged dependent variable (y.dl1) is not significant at 5% significance level. On balance, the significant du variable indicates that there is significant evidence of a structural break in the time series data. The potential breakpoint, pointed out by test corresponds to the 30^{th} observation on the time series. That is there is a breakpoint in 1990.

The significant trend, suggests there is a deterministic trend in the data. This implies that the data are not stationary around a single mean.

Table 3.2: Zivot-Andrew test result on the trained data

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	3.08241	1.19835	2.572	0.013731
y.l1	0.66722	0.07903	8.442	1.36e-10
trend	0.21886	0.08516	2.570	0.013807
y.dl1	0.22154	0.17551	1.262	0.213812
du	8.73970	2.39283	3.652	0.000715

Figure 3.2 depict the time plot for the train data with the breakpoint.

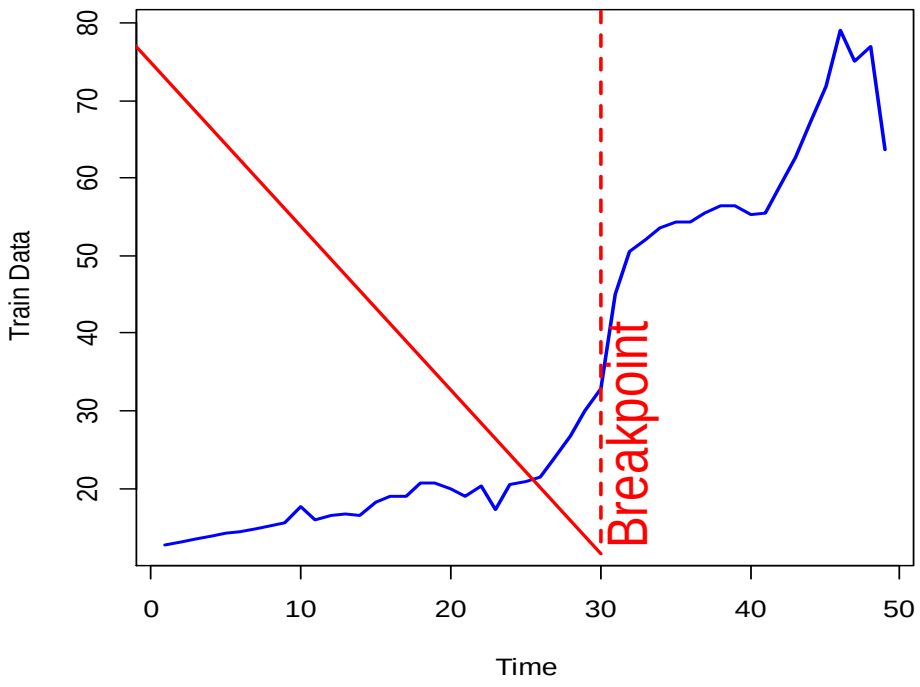

Figure 3.2: Time plot for the train data with breakpoint

Model Fitting and Comparison

Here, three models, which are the Holt's linear model with multiplicative errors, ARIMA (1,1,2) model and SETAR (2,2,1) model are fitted to the training set. In-sample and out-of-sample comparisons of the models are also considered.

Fitting Holt's Linear Model with Multiplicative Errors to the Data

Here, we use the ets() function in forecast package to determine the best fitting exponential smoothing method for the data. Accordingly, the package identifies ets(M, A,N) , which is the Holt's linear model with multiplicative errors as the best model for the data using minimum AIC AICc and BIC values. Estimates of the parameters of the model and associated results are presented in Table 3.3.

Table 3.3: Result based on the Holt's linear model with multiplicative errors fitted to the data

Smoothing parameters	Initial states	sigma	AIC	AICc	BIC
alpha = 0.6203	l = 12.4756	0.0792	277.2813	278.6767	286.7404
beta = 0.6203	b = 0.2881				

The ACF and PACF of residuals from the Holt's linear model plotted in Figure 3.3 possess the properties of white noise.

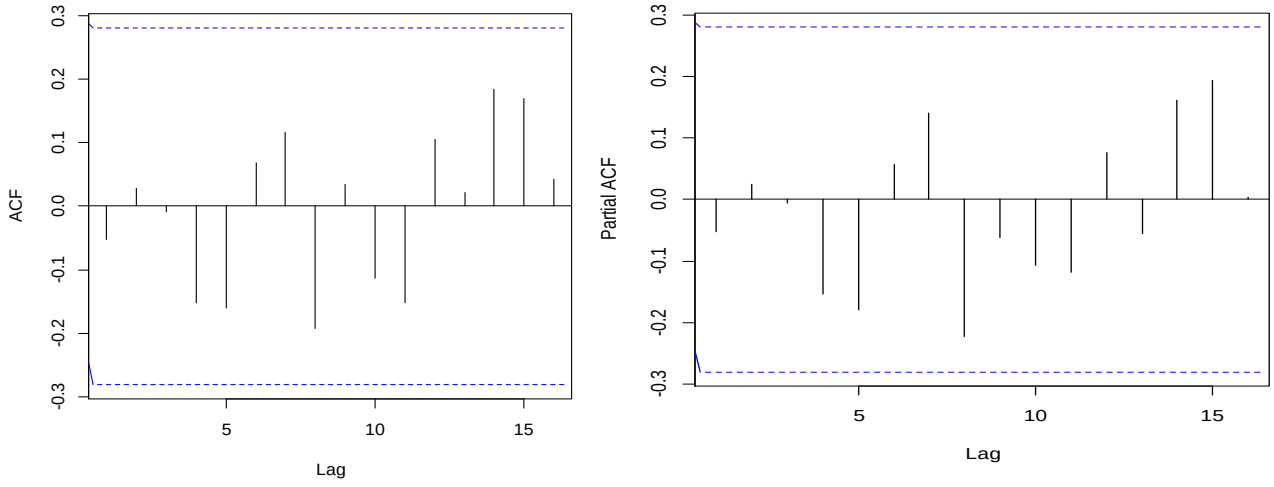

Figure 3.3: Residual ACF plot (Left Panel) and PACF plot (Right Panel) based on Holt's linear method with multiplicative errors

International Journal of Mathematics and Statistics Studies, 13 (2), 1-20, 2025
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

Fitting ARIMA(1,1,2) Model to the Data

Following the result of the ADF in Section 3.1, we proceed to determine the best fitting ARIMA model to the train data using the auto.arima function in the forecast package. The result of the ARIMA model fitting is in Table 3.4. In accordance with the ARIMA model selection process, ARIMA(1,1,2) model is the most suitable ARIMA model for the data.

Table 3.4: Estimates of ARIMA(1,1,2) model fitted to the data

Parameter	Estimate	Standard Error	$-l$	$\hat{\sigma}^2$	AIC	AICc	BIC
ϕ	0.4490	. 0.2199	120.22	9.165	248.45	249.38	255.93
θ_1	-0.2160	0.1749					
θ_2	0.5487	0.1568					

The ACF and PACF of the residuals from the fitted ARIMA(1,1,2) model are graphed in Figure 3.4. It can be deduced that the residual ACF and PACF behave like those of a white noise process.

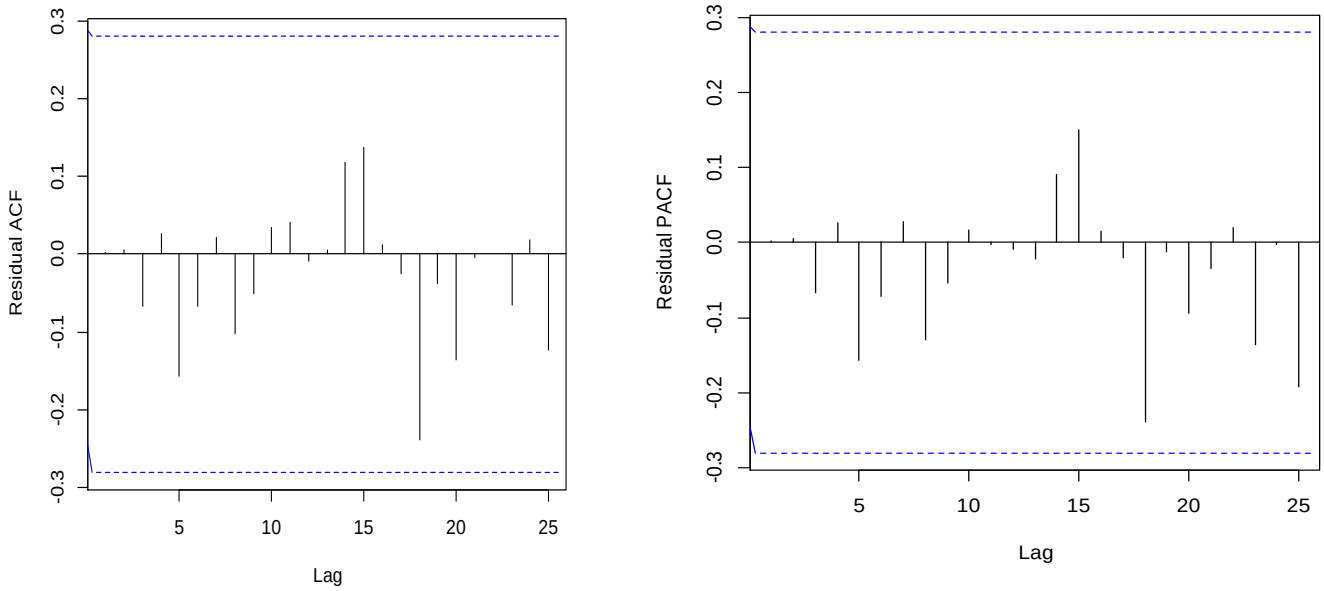

Figure 3.4: Residual ACF plot (Left Panel) and PACF plot (Right Panel) based on ARIMA(1,1,2) Model

The SETAR Model Based Results

First, we apply Hansen's test to the train data to justify the need to consider a SETAR model in this study. The test is also useful in choosing the number of regimes to be put into consideration in the entertained SETAR model. Table 3.5 comprises Hansen's test results dealing with testing the null hypothesis of linearity against each of 2- regime and 3-regime SETAR models. The null hypothesis is rejected in each case at 5% level of significance because the concerned p-values are all less than 0.05.

Table 3.5: Result on test of linearity against SETAR (2) and SETAR(3)

Hypotheses	Test Statistic	p-value
1vs2	45.04176	0.0175
1vs3	157.87644	0.0000

For the purpose of choosing the appropriate number of regimes between 2 and 3, we consider the result in Table 3.6. From the result, preference should be given to a 2-regime SETAR model. This is because at 5% level of significance, as the associated p-value is greater than 0.05.

Table 3.6: Result on test of SETAR (2) against SETAR(3)

Hypotheses	Test Statistic	p-value
2vs3	57.01115	0.08

Since the threshold value (th) is unknown, a grid search for the best thresh value based on TGSP algorithm is performed and the result of the test is contained in Table 3.7. In the table, mL is the minimum autoregressive order for low regime and mH maximum autoregressive order for high regime. On the basis of minimum pooled-AIC value, the best threshold value for the data under consideration is th=20.90. This threshold value corresponds to thDelay =0, mL=2 and mH=1.

Table 3.7: Results of the grid search for 1 threshold

thDelay	mL	mH	th	pooled-AIC
0	2	1	20.90	214.3567
0	2	2	20.90	215.5412
0	2	1	30.09	215.8577
0	2	1	21.42	215.9941
0	1	2	30.09	216.8385
0	2	2	30.09	217.3704
0	2	2	21.42	217.3970
0	1	1	30.09	217.8067
0	2	1	19.02	217.8227
0	2	2	19.02	218.0120

Having established statistically the need to fit SETAR(2,2,1) model to the training data, the results relating to the fitted SETAR model are summarized in Table 3.8. Also, the autoregressive model in the low regime (Regime 1) is of order 2 while that in the high regime (Regime 2) is of order. The sub model in each regime is estimated using the data falling in that regime. The residual variance, AIC and MAPE associated the estimated model are residuals variance = 8.246, AIC = 113 and MAPE = 5.054%.

Interestingly, the SETAR model is stationary as the autoregressive model in each regime is stationary.

Table 3.8: Results Based on SETAR(2,1,1) model fitted to the training set

Regime 1				
Parameter	Estimate	Std.Err	t-value	Pr(>|t|)
$\alpha_0^{(1)}$	1.83670	4.36227	0.4210	0.6758
$\alpha_1^{(1)}$	0.49097	0.52853	0.9289	0.3580
$\alpha_2^{(1)}$	0.43091	0.52264	0.8245	0.4141
Regime 2				
$\alpha_0^{(2)}$	8.04097	2.16144	3.7202	0.0005612
$\alpha_1^{(2)}$	0.88264	0.03910	22.5741	< 2.2e-16

Residual ACF and PACF plots for the estimated SETAR model are graphed in Figure 3.5. With respect to Figure 3.5, the residuals from the fitted SETAR model are white noise. This is because none of the autocorrelation coefficients are significantly different from zero.

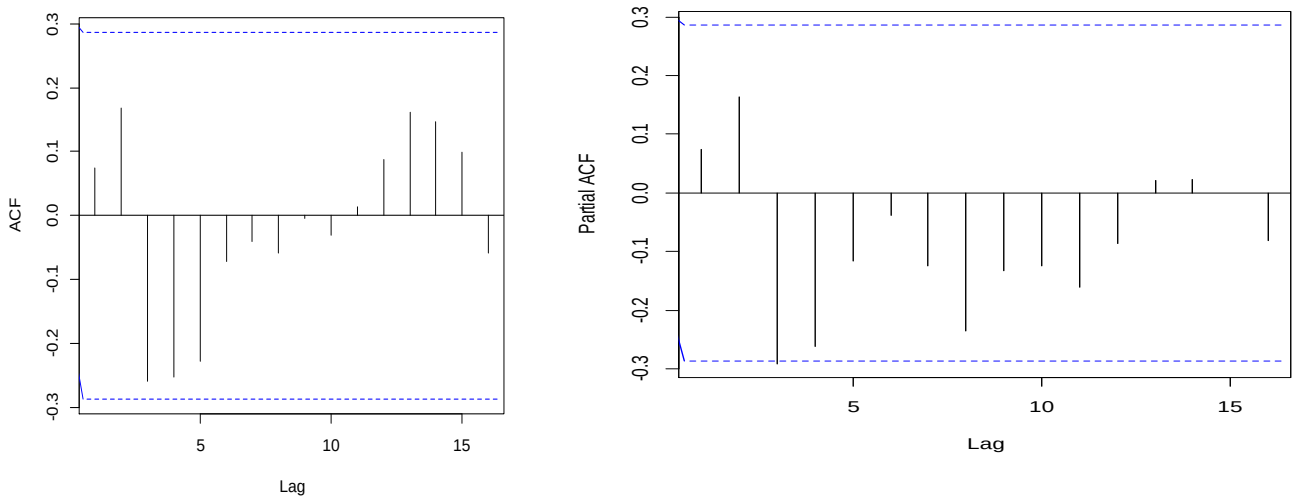

Figure 3.5: SETAR(2,2,1) model residual ACF and PACF plots

In-Sample and Out of Sample Comparison of the Fitted Models

Having fitted several models to the training set, it becomes expedient to compare the performance of the fitted models with a view to determining the best among them. For effective in sample and out of sample comparison of the models, RMSE and MAPE are given due consideration. These accuracy measures are computed and recorded in Table 3.9. Certainly, the ARIMA(1,1,2) outperforms the other two models based on the in sample comparison. However, in terms of the out of sample comparison, the SETAR(2,2,1) model is the best model following the out of sample forecast accuracy measures, namely, RMSE and MAPE.

Table 3.9: Accuracy Measures for the Fitted Models

Comparison	Accuracy Measure	Holt's Linear Model With Multiplicative Errors	ARIMA(1,1,2) Model	SETAR(2,2,1) Model
In Sample	RMSE	3.3008	2.9013	8.246
	MAPE	5.0354	5.0261	5.054
Out of Sample	RMSE	101.0595	49.6280	30.8899
	MAPE	90.9269	48.6416	29.7093

Figure 3.6 contains the time plot of the original data in blue and that of the forecast values in red. The forecast values from the plot indicates reduced future annual cassava production in Nigeria.

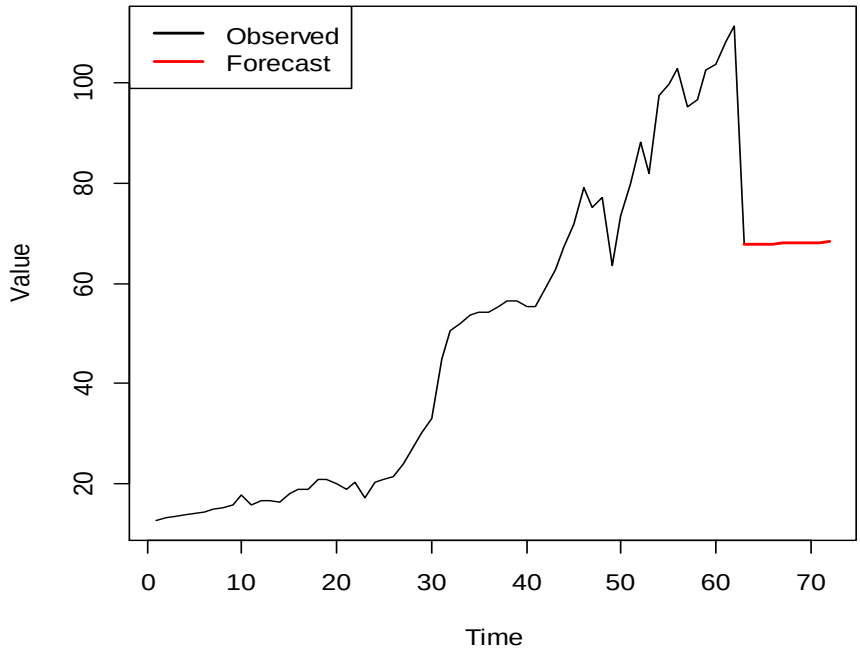

Figure 3.6: Time plot of the original series and forecast values

The forecast values for 2023 through 2032 are available in Table 3.10. The forecast value increases as time increases. However, there are all smaller than the last few actual values of the series, indicating reduced production.

Table 3.10: Forecast values of yearly cassava production in Nigeria for 2023 to 2032

Year	Forecast
2023	67.6686
2024	67.7678
2025	67.8553
2026	67.9326
2027	68.0009
2028	68.0611
2029	68.1142
2030	68.1611
2031	68.2025
2032	68.2391

CONCLUSION

This study aimed to provide an adequate model for forecasting yearly cassava production in Nigeria. Having partitioned the series into the training set and test set, properties of the training set, such as nonstationarity and structural change were investigated through the ADF test and Zivot-Andrew test respectively. The tests revealed that the series is nonstationary and possesses a structural break in 1990.

Furthermore, three time series forecasting models, Holt's linear model with multiplicative errors, ARIMA(1,1,2) model and SETAR(2,2,1) model were fitted to the series. The forecasting performance of the three models were compared based on the RMSE and MAPE. On the basis minimum RMSE and MAPE values, SETAR(2,2,1) model is certainly the best for forecasting the series among the three models under consideration. Forecasts of the values of the series for the period 2023 to 2032 are small compared to the values of the series in the few previous years, detailing reduction in cassava production in the future if the prevailing circumstances persist.

REFERENCES

1. Abraham, **B** and Ledolter, J. (1983). Statistical Methods for Forecasting. John Wiley & Sons, Inc., Hoboken, New Jersey.
2. Antonio, F. D. N, Aznarte, J. L and Stigler, M. (2009). tsDyn: Time Series Analysis Based On Dynamical Systems Theory Package.
3. Çamalan, Ö, Hasdemir, E., Omay, T and Can, M. (2024). Comparison of the Performance of Structural Break Tests in Stationary and Nonstationary Series: A New Bootstrap Algorithm. Computational Economics, https://doi.org/10.1007/s10614-024-10651-z.
4. Chan, K. S. (, 1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressisve model. The Annuals of Statistics, 21:520–533.
5. Donkor, E, Onakuse, S, Bogue, J and Carmenado and Ignacio, d L R. (2017). The impact of the presidential cassava initiative on cassava productivity in Nigeria: Implication for sustainable food supply and food security. Cogent Food & Agriculture, 3: 1368857.
6. Durbin, J. (1960). The fitting of time series models. Review of the Institute of International Statistics, 28: 233-244.
7. Enders, W. (2004). Applied Econometric Time Series. John Wiley and Sons, Third Edition.
8. FAO (2015). Report on Regional Conference on Cassava in the Caribbean and Latin American, 10–12 February 2014. Food and Agriculture Organization of the United Nations, Rome.
9. FAO (2018). Food Outlook - Biannual Report on Global Food Markets – November 2018. Rome; p. 104. http://www.fao.org/3/ca2320en/CA2320EN.pdf License: CC BY-NC-SA 3.0 IGO.
10. Hansen, B. (1999). Testing for Linearity. Journal of Economic Surveys: Volume 13, Issue 5, p. 503-674.
11. Hershey C., Henry G., Best R., Kawano K and Howeler R.H. (2001). Iglesias C. Validation Forum on the Global Cassava Development Strategy, Proceedings. vol. 3. Food and

Agriculture Organization of the United Nations (FAO); International Fund for Agricultural Development (IFAD); Rome, IT: 2001. Cassava in Asia: Expanding the Competitive Edge in Diversified Markets. A Review of Cassava in Asia with Country Case Studies on Thailand and Vietnam; pp. 1-62.

12. Iquebal, M. A, Ghosh, H and Prajneshu. (2013). Fitting of SETAR Three-regime nonlinear time series model to Indian lac production data through genetic algorithm. Indian Journal of Agricultural Sciences 83 (12): 1406-8.
13. Magadia, J. C. (2016). Value-at-Risk Estimates from a SETAR Model. The Philippine Statistician, 65(1): 15-26.
14. Njoku, D.N and Muoneke, C.O. (2008). Effect of cowpea planting density on growth, yield and productivity of component crops in cowpea/cassava intercropping system. Agro-Science, 7(2): 106-113.
15. Ohimain, E. I. (2015). A Decade (2002-2012) of Presidential Intervention on Cassava in Nigeria; the Successes and Challenges. Asian Journal of Agricultural Extension, Economics & Sociology, 6(4): 185-193.
16. Okhankhuele, O. T, Opafunso, Z. O, Akinrinola, O. O and Ojo, O. J. (2017). Evaluation of Presidential Cassava Transformation Initiative on marketing of cassava products, produced by Micro-Scale Cassava Processing Enterprises in Southwest Nigeria. CARD International Journal of Management Studies, Business & Entrepreneurship Research, 2(3): 222-243.
17. Omoluabi, J. E and Ibitoye, S. J. (2024). Cassava production and agricultural growth in Nigeria: Analysis of effects and forecast. GSC Advanced Research and Reviews, 2024, 21(01), 037-046.
18. Oni, O. V and Akanle, Y. O. (2018). Comparison of exponential smoothing models for forecasting cassava production. International Journal of Scientific research in mathematical and Statistical Sciences, 5(1): 65-68.
19. Sanogo, D and Adetunji, O. (2008). Presidential intiatives on cassava in Africa: case studies of Ghana and Nigeria, p. 73. Malawi: IITANEPAD.
20. Ugbem-Onah, C. E and Mbakuuv, P. A. (2024). IFAD/Value Chain Development Programme and Cassava Production In Logo Local Government Area of Benue State. African Scholars Multidisciplinary Journal (ASMJ), 8: 188 – 197.
21. Ukwuru, M. U and Egbeonu, S. E. (2013). Recent development in cassava–based products research. Academic Journal of Food Research, 1910: 1-13.
22. Wei, W. W. S. (2006). Time series analysis: Univariate and multivariate methods. Pearson Addison Wesley, London.
23. Zivot, E and Andrews, D. W. K. (1992). Further evidence on the great crash, the oil-price shock and unit—root hypothesis. Journal of Business and Economic Statistics, 10: 251-270.