Unfolding the potential of the ARIMA model in forecasting maize production in Tanzania

Joseph Lwaho and Bahati Ilembo
Department of Mathematics and Statistics Studies, Mzumbe University, Morogoro, United Republic of Tanzania

Received 2 July 2023
Revised 27 August 2023
6 September 2023
Accepted 5 October 2023

Abstract

Purpose – This paper was set to develop a model for forecasting maize production in Tanzania using the autoregressive integrated moving average (ARIMA) approach. The aim is to forecast future production of maize for the next 10 years to help identify the population at risk of food insecurity and quantify the anticipated maize shortage.
Design/methodology/approach – Annual historical data on maize production (hg/ha) from 1961 to 2021 obtained from the FAOSTAT database were used. The ARIMA method is a robust framework for forecasting time-series data with non-seasonal components. The model was selected based on the Akaike Information Criteria corrected (AICc) minimum values and maximum log-likelihood. Model adequacy was checked using plots of residuals and the Ljung-Box test.
Findings – The results suggest that ARIMA (1,1,1) is the most suitable model to forecast maize production in Tanzania. The selected model proved efficient in forecasting maize production in the coming years and is recommended for application.
Originality/value – The study used partially processed secondary data to fit for Time series analysis using ARIMA (1,1,1) and hence reliable and conclusive results.

Keywords ARIMA, Time series, Maize production, Forecast
Paper type Research paper

1. Introduction

Zea mays is regarded as the major food crop produced and consumed worldwide (Nyaligwa, Hussein, Laing, Ghebrehiwot, & Amelework, 2017). It can be consumed by humans and livestock and used as raw materials for biofuel production. In Sub-Saharan Africa (SSA), maize is the most crucial primary cereal crop grown in over half of the countries and one of the top two cereals in over three-quarters of these countries (Suleiman & Kurt, 2015; Faostat, 2021). The crop flourishes on soils with pH between 5.0 and 7.0; nevertheless, a moderately acidic environment of pH 6.0 and 7.0 is mostly favorable (Baijukya *et al.,* 2020).

Tanzania has been among the world's top 25 maize-producing countries (Twilumba, Ahmad, & Shausi, 2020). According to Suleiman and Kurt (2015), Tanzania is among the significant maize producers in SSA. It is a famous and prominent staple food (Laudien, Schauberger, Makowski, & Gornott, 2020), and both rural and urban population consumes it (Baijukya *et al.*, 2020).

Food crop production, particularly maize, has a direct effect on the status of food security (Mkonda & He, 2017), which is one of the major areas of concern all over the world due to its contribution to all forms of human survival (Quaye, Yawson, Ayeh, & Yawson, 2012; Ngongi & Urassa, 2014). The significance of food security cannot be overemphasized, and in recent

Business Analyst Journal
Emerald Publishing Limited
e-ISSN: 2754-6721
p-ISSN: 0973-211X
DOI 10.1108/BAJ-07-2023-0055

decades, the developing world has experienced food shortage, which led to food insecurity (Rwanyiziri *et al.,* 2019). It is estimated that about 355 million people in SSA will be under food shortage by 2050 (Rwanyiziri *et al.*, 2019).

In recent years, forecasting food crop production has become more challenging (Liu & Basso, 2020) due to several major drivers, particularly climate extremes like heavy rains, storms and floods. It is clear that major players' efforts to ensure the world has enough food are shattered by several major drivers, particularly climate change (WHO, 2021). Despite the efforts from stakeholders and the government, the country still has not done well regarding crop yield sustainability and food security (Mkonda & He, 2017). Although modelling cereals production has attracted intensive research due to the vitality of major food crops, little is known, especially in maize yield forecasting in response to food insecurity.

Time series methods have been widely used in forecasting future values based on past observations (Enders, 2015). Time series analysis involves studying the variables on which observations are arranged sequentially over time. In most cases, forecasting concentrates on univariate time series models pioneered by Box and Jenkins, including autoregressive (AR), moving average (MA) and autoregressive integrated moving average (ARIMA) models (Box & Jenkins, 1970). This method has been successful in many applications, including economics (Petrevska, 2017; Yildiran & Fettahoğlu, 2017), agriculture (Uwamariya & Ndanguza, 2018; Mgaya, 2019; Bezabih, Wale, Satheesh, Fanta, & Atlabachew, 2023), climate and environment (Fayaz, Meraj, Khader, & Farooq, 2022).

Few studies exist in Tanzania that model and forecast maize production using different methods. For example, Ogutu, Franssen, Supit, Omondi, and Hutjes (2018) forecasted maize production in Tanzania using dynamic ensemble seasonal climate forecasts. Laudien *et al.* (2020) used LASSO regression to forecast maize yields before harvest. Liu and Basso (2020) forecasted maize yields for smallholder farmers in three selected regions in Tanzania by integrating a field-based survey with a process-based mechanistic crop simulation model. In Mkonda and He (2018), the analysis focused on the trend of yields by plotting graphically, and the results were unreliable and inconclusive. This paper attempts to model annual maize production and forecast future production using the ARIMA model for the next ten years. The ARIMA method is chosen due to its proven forecasting reliability and ability to predict sequential series accurately.

2. Materials and methods

2.1 Data and data source

The paper used time series data of maize yield (hg/ha) from 1961 – 2021 obtained from FAO STAT calculated annually. Maize was selected due to its potential as the primary food crop for most households in Tanzania (Laudien *et al.*, 2020). The dataset contained 60 observations, meeting the requirement of the generalization ability of time series analysis. Hyndman and Athanasopoulos (2018) argued that there is no justification for the minimum number of observations for ARIMA modelling, the only theoretical limit being that there should be more observations than the number of parameters in the forecasting model. Time series analysis was done using *forecast, ggplot2 and ur* packages ensembled in R software. The function auto.arima ( ) was used to obtain the appropriate ARIMA model with estimated parameters automatically instead of specifying using autocorrelation function (ACF) and partial autocorrelation function (PACF), as suggested by Hyndman and Athanasopoulos (2018).

2.2 The model formulation

The ARIMA method was considered appropriate to achieve the maize production forecasting objective. ARIMA is a family of time series models introduced by Box and Jenkins (1970) and is a widely used method in modelling and forecasting stationary and no-stationary time series with non-seasonal components. The reason for choosing the ARIMA technique is its ability to generate high predictive accuracy compared to the AR and MA as standalone models for short-run forecasting (Box & Jenkins, 1970). This methodology involves three successive phases: identification, which determines the order of the model required (p, d and q) to capture the data's salient dynamic features. This mainly leads to the use of graphical procedures (plotting the series, the ACF and PACF, etc.), estimation involves estimating the parameters of the different models. It proceeds to a first selection of models (using information criteria), and the diagnostic checking involves determining whether the model(s) specified and estimated is adequate. Notably, one uses residual diagnostics.

2.2.1 Autoregressive (AR). An *AR* model is the one in which Y_t depends only on its own past values $Y_{t-1}, Y_{t-2}, Y_{t-3}$ e.t.c

A typical representation of an *AR* model where it depends on p of its past values called *AR* (p) is as shown below:

$$Y_t = \alpha + \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \psi_3 Y_{t-3} + ... + \psi_p Y_{t-p} + \varepsilon_t \tag{1}$$

$$Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + \psi_3 Y_{t-3} + ... + \psi_p Y_{t-p} + \varepsilon_t \tag{2}$$

the $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$ is predictor variable at lags $t-1, t-2, \ldots, t-p$, $\alpha = \mu(1-\psi_1 - \psi_2 - \psi_3 - \ldots - \psi_p)$, $\psi_1, \psi_2, \psi_3, \ldots, \psi_p$ are the parameters to be measured, and ε_t is the error term which follow white noise process with zero mean and a constant variance, i.e. $\varepsilon_t \sim N(0, \sigma^2)$. This *AR* ($p$) can also be presented as follows using backward shift operator;

$$\left(1 - \psi_1 B - \psi_2 B^2 - \psi_3 B^3 - \psi_p B^p\right) Y_t = \varepsilon_t \tag{3}$$

where B is the backshift operator defined by

$$B^m Y_t = Y_{t-m} \; (m = 0, 1, 2 \ldots p)$$

2.2.2 Moving average (MA). An MA model is the one when Y_t depends only on the error terms which follow a white noise process. The general form is given by:

$$Y_t = \delta + \varepsilon_t - \theta_1 \varepsilon_{t-1} \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - ... - \theta_p \varepsilon_{t-p} \tag{4}$$

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - ... - \theta_p \varepsilon_{t-p} \tag{5}$$

$$\text{For } \delta = 0$$

$$\varepsilon_t \sim WN\left(0, \sigma^2\right)$$

Using the backshift operator, the above equation (5) can be rewritten as

$$\left(1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 - \theta_p B^p\right) \varepsilon_t = Y_t \tag{6}$$

2.2.3 ARMA model. ARMA (p, q) is defined as the combination of AR(p) and MA(q) for a stationary time series and is given by the following equation:

$$Y_t = \alpha + \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + ... + \psi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_3 \varepsilon_{t-3} - ... - \theta_p \varepsilon_{t-p} \tag{7}$$

$$\text{For } \alpha = 0$$

$$Y_t = \psi_1 Y_{t-1} + \psi_2 Y_{t-2} + ... + \psi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_p \varepsilon_{t-p} \tag{8}$$

Equation (8) can be defined using backshift operator as follows,

$$\psi(B)Y_t = \theta(B)\varepsilon_t \tag{9}$$

Although ARMA (p, q) is useful in modelling the time series process, it is special only if the series is stationary. Most time series data are not stationary, making ARMA unsuitable for those circumstances. Box-Jenkins methodology outlined a solution for the case of non-stationary series where data transformation (differencing) is needed to attain normality. Under this circumstance, the ARIMA (p, d, q) model is appropriate.

2.2.4 ARIMA model. ARIMA model is a combination of AR, i.e. Autoregressive (lagged observations as inputs), whereby I stand for Integrated (differencing to make series stationary) and MA, i.e. moving average (lagged errors as inputs).

According to Box and Jenkins (1970), the ARIMA model is denoted by ARIMA (p, d, q) where p is the order of autoregressive process, d is the order of integration, i.e. the number of differences to make the series stationary and q is the order of MA process. The general form of the ARIMA (p, d, q) is;

$$Y'_t = \alpha + \psi_1 Y'_{t-1} + \psi_2 Y'_{t-2} + ... + \psi_p Y'_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \theta_3\varepsilon_{t-3} - ... - \theta_p\varepsilon_{t-p} \tag{10}$$

$$\text{For } \alpha = 0$$

$$Y'_t = \psi_1 Y'_{t-1} + \psi_2 Y'_{t-2} + ... + \psi_p Y'_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - ... - \theta_p\varepsilon_{t-p} \tag{11}$$

Equation (8) can be defined using backshift operator:

$$\psi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t \tag{12}$$

where Y'_t is differenced response variable, d is the number of difference and $\varepsilon_t \sim WN(0, \sigma^2)$.

The differencing of the response variable can be calculated by using the following relation for non-stationary data:

$$Y'_t = \nabla^d Y_t = (1-B)^d Y_t = \sum_{j=0}^{d} \binom{d}{j} (-1)^j Y_{t-j}$$

Box and Jenkins methodology (step-by-step procedures) were used to achieve the study's objective, which involves model identification, estimation, diagnostic checking and forecasting.

2.3 Model identification

Before ARIMA modelling, the time series data structure was checked for stationarity. Here, the focus was to observe the behavior of the mean and variance of a stochastic process to identify the existence of trends or seasonal patterns. As Montgomery, Jennings, and Kulahci (2015) suggested, checking for stationarity is essential because it brings equilibrium and stability to data. A series is considered stationary if its mean and variance are constant over time, and covariance depends only on lags (Enders, 2015).

2.3.1 Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test was used to determine if the series is stationary. The null hypothesis was that the times series level is stationary (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). Reject null hypothesis for a large value of KPSS test statistic compared to critical values (Müller, 2005). This suggests that data transformation is required, i.e. log-transformation or differencing. The KPSS test statistic is given by

$$KPSS = \frac{1}{T^2 \widehat{\sigma}_T^2} \sum_{t=1}^{T} S_t^2$$

Where $\widehat{\sigma}_T^2$ is a consistent estimator of the long-run variance of the residuals, $S_t = \sum_{s=1}^{t} \widehat{e}_s$ is the sum of the residuals, T is the number of observations (Kwiatkowski *et al.*, 1992).

2.3.2 Correlograms. The autocorrelation function (ACF) and partial autocorrelation function (PACF) provided essential information to identify if a time series is stationary. The ACF defines the order of the AR process, while the PACF defines the order of the MA process. These two correlograms provided useful information concerning the stationarity of the time series process. For the time series process to be stationary, the ACFs for the AR process should be characterized by decaying exponentially tails off towards zero while the MA process cuts off to zero after lag q. In terms of PACFs, the AR process cuts to zero at lag p while the MA process decays exponentially tails off towards zero to show that the process is stationary (Wei, 2006).

2.3.3 Autocorrelation function (ACF). The ACF measures the correlation between two observations in a series over the corresponding variable lags, i.e. Y_t and Y_{t+h}, using the following equation:

$$\rho_h = \frac{\sum_{t=1}^{T-h} \left(Y_t - \overline{Y}\right)\left(Y_{t+h} - \overline{Y}\right)}{\sum_{t=1}^{T} \left(Y_t - \overline{Y}\right)^2}$$

Where Y_t is the current observation and Y_{t+h} is the observation h after the current one.

2.3.4 Partial autocorrelation function (PACF). Used to measure the degree of association between Y_{t} and Y_{t-p} when the effects at other time lag Y_{t-1}, X_{t-2} ,. . .., Y_{t-p+1} are removed.

$$\gamma_{hh} = \begin{cases} \rho_1, \quad for\ h = 1 \\ \dfrac{\rho_h - \sum_{j=1}^{h-1} \left(\gamma_{h-1,j} \cdot \rho_{h-j}\right)}{1 - \sum_{j=1}^{h-1} \left(\gamma_{h-1,j} \cdot \rho_j\right)}, for\ h > 1 \end{cases}$$

where $\gamma_{hj} = \gamma_{h-1,j} - \gamma_{hh}\gamma_{h-1,h-j}, j = 1, 2, 3, \ldots, h-1.$

In conditions where the time series is non-stationary, data transformation has to be done by applying appropriate differencing, hence obtaining suitable values (Enders, 2015). The suitable values of p and q will be selected by observing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series data. The appropriate ARIMA models are selected by observing the behavior of ACF spikes and PACF based on the order identified (Hyndman & Athanasopoulos, 2018).

2.4 Model estimation

The parameters α, ψ and θ of the selected ARIMA model with specified values of p, q and d d need to be estimated. Maximum Likelihood Estimation (MLE) Method is proposed to estimate the parameters. The final model for forecasting is selected based on Information Criteria.

2.4.1 Maximum likelihood estimation (MLE). This technique estimates ARIMA model parameters that maximize the likelihood of getting the observed series data. The method is based on providing the log-likelihood, which describes the logarithm of the probability of observed time series data of the estimated model. The function is given by:

$$\log L = -\frac{T}{L}\left(\ln(2\pi) + \ln\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t$$

where T is the time of the time series data, L is likelihood function, σ^2 is the constant variance and ε_t is the error term.

The model with the maximum log-likelihood value is considered for selection for subsequent scrutiny based on Information Criteria.

2.4.2 Information criteria. The ARIMA model subjected to forecasting is selected based on a threshold of Information Criteria. Akaike Information Criterion (AIC), Akaike Information Criterion corrected (AICc) and Bayesian Information Criteria (BIC) are assessed to specify the model with the lowest index (Hyndman & Koehler, 2006). The AIC is suitable for obtaining the appropriate ARIMA model based on the smallest values compared to other competing models. It is given by

$$AIC = -2\ln L + 2p$$

where, L is the likelihood function of time series data, and p is the number of parameters of the fitted model.

Since AIC does not consider the effect of sample size, AICc makes adjustments to allow the criteria to be used in the presence of a small sample. The following equation defines the AICc:

$$AICc = AIC + \frac{2p(p+1)}{n-p-1}$$

The BIC is another information criterion that is useful in deciding the required model. It extends the AIC by penalizing free parameters stronger than AIC as a standalone criterion. The BIC is obtained as follows:

$$BIC = AIC + p(\ln(n) - 2)$$

Although all of the above criteria are provided after model estimation, this study selected AICc as a reference to choose the suitable ARIMA model. The model with the lowest AICc must be considered the best forecasting model.

2.5 Diagnostic checking

After choosing a relevant ARIMA model and estimating the corresponding parameters, model adequacy was checked by observing if the fitted model residuals were normally distributed. One of the methods of investigating whether the distribution of residuals from the fitted model is random is using the correlograms. The ACF and PACF can be used to verify if the time series displays a white noise innovation by considering that the model's residuals should not display significant lags.

2.5.1 The Ljung-Box test. The Ljung-Box test was also used to check if the residuals followed a white noise process. The residuals were analyzed using Ljung-Box Statistic to check if the autocorrelation of the time series is significantly different from zero (Ljung & Box, 1978) and is calculated as:

$$the\,Q(k) = n(n+2)\sum_{j=1}^{k}\frac{r_j^2}{(n-j)}$$

where n is the sample size, r_j is the sample autocorrelation at lag j and k is the lag order.

The null hypothesis under this test is that the residuals are white noise, and the hypothesis is rejected if $Q(k)$ is larger than $(1-\alpha)$ quantile.

2.6 Forecasting

At this stage, the idea is to use the selected ARIMA model to forecast future maize production using the past series. Forecasting should be based on the final selected model derived from the diagnostic stage. The plotted graph of the forecast and actual series will inform if the forecast is good or not before a conclusion is made on the ability of the ARIMA model to forecast future values of time series. To measure the accuracy of the prediction model, Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), Mean Absolute Scaled Error (MASE) and Mean Absolute Error (MAE) are suitable (Hyndman & Koehler, 2006). The essence here is to determine the magnitude of errors and bias, and the qualified model should register as minimum errors as possible.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \widehat{Y}_t\right|$$

$$the\,RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right| * 100$$

$$MASE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|Y_t - \widehat{Y}_t\right|}{\frac{1}{n-1}\sum_{t=2}^{n}\left|Y_t - Y_{t-1}\right|}$$

Where n is the sample size, Y_t is the response variable at time t, Y_{t-1} is the response variable at time $t-1$, and $\widehat{Y}_t$ is the forecast value.

3. Results

3.1 Time series plot and trend analysis

The original and annual maize production data are plotted to assess the trend. The plotted graph in Figure 1 shows the volatility of the time series data. Maize production was relatively constant from 1960s to 1970s before it increased rapidly for the next ten years. The production decreased abruptly through the midpoint between 2000 and 2010 before increasing slightly from 2010 onwards. The time series under investigation is characterized by fluctuation behavior, which is a feature of many time series.

3.2 Stationarity test of time series

Table 1 shows the results of the KPSS Unit Root Test.

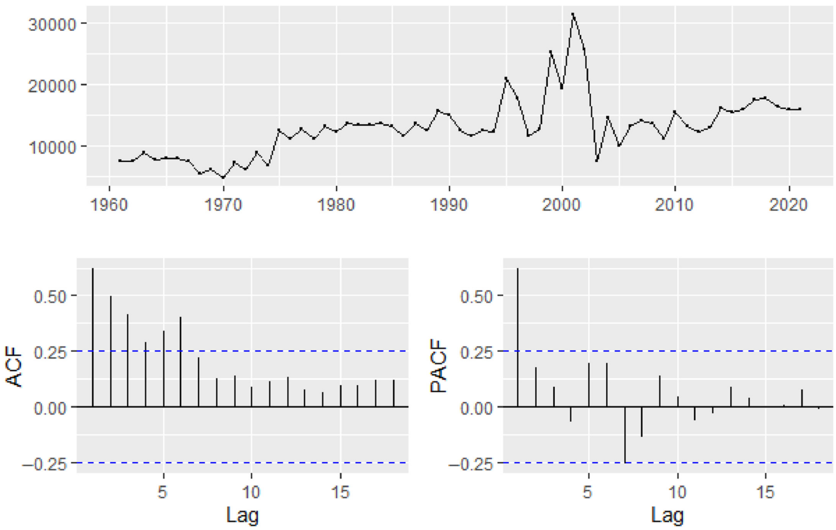


Figure 1. Time plot and ACF and PACF plots for maize production (hg/ha) data

Source(s): Created by authors

Table 1. The results of KPSS unit root test

Test is of type: mu with 3 lags
Value of test-statistic is: 0.9219
Critical value for a significance level of

	10pct	5pct	2.5pct	1pct
Critical values	0.347	0.463	0.574	0.739

Source(s): Created by authors

Table 1 shows that, the test statistic is bigger than the 1% critical value, indicating that the null hypothesis is rejected and that, the data are not stationary. The KPSS result concurs with the ACF and PACF correlograms shown in Figure 1.

3.3 Stationarity transformation

Since the results above conclude that the time series data is not stationary, differencing and natural log transformations were used for transforming the non-stationary data. This is in agreement with (Enders, 2015), that when time series data is not stationery performing differencing will make the data stationary, and therefore further analysis can be carried out.

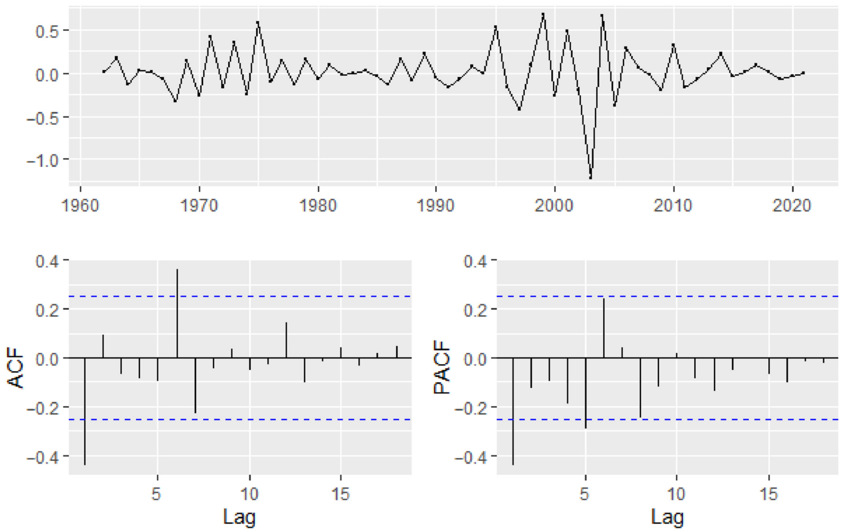
Results in Table 2 show that the test statistic is insignificant, meaning the differenced data are stationary. This can be supported by the time series plot, as shown in Figure 2.

Table 2. KPSS unit root test after first differencing

Test is of type: mu with 3 lags
Value of test-statistic is: 0.0368
Critical value for a significance level of

	10pct	5pct	2.5pct	1pct
Critical values	0.347	0.463	0.574	0.739

Source(s): Created by authors



Source(s): Created by authors

Figure 2. Time plot and ACF and PACF plots for differenced maize production data

3.4 Model selection

Since the series tends to be stationary after transformation, the next step is identifying the order of the model's AR (p) and MA (q) to be estimated. To achieve this, ACF and PACF plots are going to be used. Based on the ACF plot in Figure 2, the order of AR is 1 because of the appearance of one significant spike compared to others. On the other hand, the MA order appears to be 1 due to one significant spike, as the PACF shown in Figure 2. Thus, the proposed model is the combination of AR (1) and MA (1), which resulted to the ARIMA (1, 1, 1) model since the order of differencing for our data is 1.

3.5 Model estimation

ARIMA (1,1,1) model was selected because of the lowest Akaike Information Criterion Corrected (AICc) of 1163.14 and the largest log-likelihood of −578.36 among other models, and it was considered the best model for forecasting. The parameters were estimated by using the maximum likelihood estimation method. The estimated ARIMA (1,1,1) model is given as:

$$MP = 0.3413Y_{t-1} + \varepsilon_t - 0.8272\varepsilon_{t-1}$$

The established ARIMA (1, 1, 1) model is then scrutinized for suitability in forecasting maize production. The ACF plot of the residuals from the ARIMA (1,1,1) model showed that all autocorrelations were within the threshold limits and were white noise. The Liung–Box test statistic was 12.568, and the p-value was 0.1276, which is greater than 0.05 suggesting that the residuals are white noise and, thus, the model is declared fit for forecasting.

3.6 Forecasting maize production

Based on the ARIMA (1, 1, 1) model, the maize production forecast and its 95% confidence interval for the next ten (10) years are provided in Table 3. In addition, Figure 3 shows the trend of forecasted maize production of actual and forecasts, which suggests that the

Table 3.
Ten years forecast of maize production (hg/ha)

Year	Forecasts	95% confidence interval Lower	Upper
2022	15910.95	8541.473	23280.43
2023	15881.24	7595.026	24167.45
2024	15871.10	7196.617	24545.58
2025	15867.64	6930.918	24804.36
2026	15866.46	6707.220	25025.69
2027	15866.05	6499.859	25232.25
2028	15865.92	6300.619	25431.21
2029	15865.87	6106.655	25625.08
2030	15865.85	5916.876	25814.83
2031	15865.85	5730.785	26000.91

Source(s): Created by authors

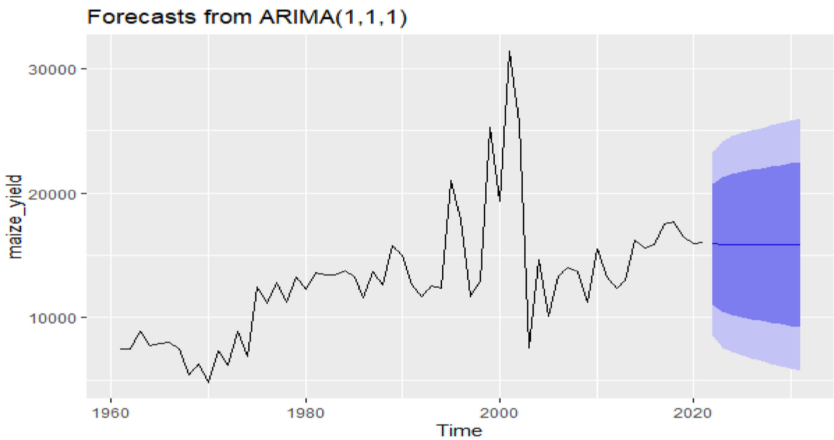


Figure 3.
Forecasts for the maize production in hg/ha

Source(s): Created by authors

production has stagnated for a long time. That means the production of maize for the ten (10) years ahead is decreasing slightly with a nearly constant movement with no sign of bouncing back soon.

4. Conclusion

This paper used time series analysis to forecast maize production. Specifically, the examination used the ARIMA model to forecast the future values based on the observed series. ARIMA (1, 1, 1) model was used as it is declared fit to be used in forecasting for such data points, which were collected from 1961 to 2021. Results indicate that maize production for the next ten years (2022 - 2031) is decreasing slightly with a nearly constant movement and no sign of returning soon. The study concludes that this model performs well in forecasting maize production in Tanzania in the short run. As limitations, the present study used data on the production of maize from FAOSTAT for the period 1961 to 2021. Hence, findings may not be necessarily the same when other sources of data are used. Also, the study used maize as a single variable in the ARIMA model; the use of any additional variable would have influenced the results. Lastly, the data used are measured in hectograms per hectare and not kilograms per hectare. Further research may use more than one variable in maize forecasting, for example, the amount of rainfall recorded in mmHg for a specified period of time.

5. Policy implication

Findings from this study will enable policymakers in Tanzania and government officials to make a well-informed decision to improve maize production, which showed a slightly declining trend for the next ten years. An informed decision can include improving the delivery of new farming technologies, promoting and increasing the use of fertilizer and other complementary practices to achieve yield potential and closing gaps in technology adoption and productivity among males and females in maize production. These, together with other well-thought-out interventions, will address the declining trend of maize production.

References

Baijukya, F. P., Sabula, L., Mruma, S., Mzee, F., Mtoka, E., Masigo, J., Ndunguru, A., & Swai, E. (2020). *Maize production manual for smallholder farmers in Tanzania*. Ibadan: IITA.

Bezabih, G., Wale, M., Satheesh, N., Fanta, S. W., & Atlabachew, M. (2023). Forecasting cereal crops production using time series analysis in Ethiopia. *Journal of the Saudi Society of Agricultural Sciences*. doi: 10.1016/j.jssas.2023.07.001.

Box, G., & Jenkins, G. (1970). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Enders, W. (2015). *Applied econometric time series* (Fourth edition). New York: University of Alabama.

Faostat, F. A.O. (2021). Rome: The Food and Agriculture Organization of the United Nations.

Fayaz, M., Meraj, G., Khader, S. A., & Farooq, M. (2022). ARIMA and SPSS statistics-based assessment of landslide occurrence in western Himalayas. *Environmental Challenges*, *9*, 100624.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. Melbourne: OTexts.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?. *Journal of Econometrics*, *54*(1-3), 159–178.

Laudien, R., Schauberger, B., Makowski, D., & Gornott, C. (2020). Robustly forecasting maize yields in Tanzania based on climatic predictors. *Scientific Reports*, *10*(1), 19650.

Liu, L., & Basso, B. (2020). Linking field survey with crop modeling to forecast maize yield in smallholder farmers' fields in Tanzania. *Food Security*, *12*(3), 537–548.

Ljung, G. M., & Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Mgaya, J. F. (2019). Application of ARIMA models in forecasting livestock products consumption in Tanzania. *Cogent Food and Agriculture*, *5*(1), 1607430.

Mkonda, M. Y., & He, X. (2017). Yields of the major food crops: Implications to food security and policy in Tanzania's semi-arid agro-ecological zone. *Sustainability*, *9*(8), 1490.

Mkonda, M. Y., & He, X. (2018). Agricultural history nexus food security and policy framework in Tanzania. *Agriculture and Food Security*, *7*(1), 1–11.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. New Jersey: John Wiley & Sons.

Müller, U. K. (2005). Size and power of tests of stationarity in highly autocorrelated time series. *Journal of Econometrics*, *128*(2), 195–213.

Ngongi, A. M., & Urassa, K. (2014). Farm households food production and households' food security status: A case of Kahama district, Tanzania. *Tanzania Journal of Agricultural Sciences*, *13*(2), 40–58.

Nyaligwa, L., Hussein, S., Laing, M., Ghebrehiwot, H., & Amelework, B. A. (2017). Key maize production constraints and farmers' preferred traits in the mid-altitude maize agroecologies of northern Tanzania. *South African Journal of Plant and Soil*, *34*(1), 47–53.

Quaye, W., Yawson, R. M., Ayeh, E. S., & Yawson, I. (2012). Climate change and food security: The role of biotechnology. *African Journal of Food, Agriculture, Nutrition and Development*, *12*(5), 6354–6364.

Ogutu, G., Franssen, W. H., Supit, I., Omondi, P., & Hutjes, R. W. (2018). Probabilistic maize yield prediction over East Africa using dynamic ensemble seasonal climate forecasts. *Agricultural and Forest Meteorology*, *250*(251), 243–261.

Petrevska, B. (2017). Predicting tourism demand by ARIMA models. *Economic Research-Ekonomska Istraživanja*, *30*(1), 939–950.

Rwanyiziri, G., Uwiragiye, A., Tuyishimire, J., Mugabowindekwe, M., Mutabazi, A., Hategekimana, S., & Mugisha, J. (2019). Assessing the impact of climate change and variability on wetland maize production and the implication on food security in the highlands and central plateaus of Rwanda. *Ghana Journal of Geography*, *11*(2), 77–102.

Suleiman, R. A., & Kurt, R. A. (2015). Current maize production, postharvest losses and the risk of mycotoxins contamination in Tanzania. In *2015 ASABE Annual International Meeting* (pp. 26–29). New Orleans, LA: American Society of Agricultural and Biological Engineers.

Twilumba, J. K., Ahmad, A. K., & Shausi, G. L. (2020). Factors influencing use of improved postharvest storage technologies among small scale maize farmers: A case of Kilolo district, Tanzania. *Tanzania Journal of Agricultural Sciences*, *19*(1), 11–21.

Uwamariya, D., & Ndanguza, D. (2018). Bayesian inference approach in modeling and forecasting maize production in Rwanda. *African Journal of Applied Statistics*, *5*(2), 503–517.

Wei, W. W. (2006). *Time series analysis: Univariate and multivariate. Methods*. Boston, MA: Pearson Addison Wesley.

World Health Organization. (2021). *The State of Food Security and Nutrition in the World 2021: Transforming food systems for food security, improved nutrition and affordable healthy diets for all*, *2021*. Food and Agriculture Org. doi: 10.4060/cb4474en.

Yildiran, C. U., & Fettahoğlu, A. (2017). Forecasting USDTRY rate by ARIMA method. *Cogent Economics and Finance*, *5*(1), 1335968.

Corresponding author
Bahati Ilembo can be contacted at: ilembobahati@gmail.com