A Comparative Statistical Approach for Forecasting Maize Yield in India Using ARIMA Models

Original Research Article

Abstract

This paper investigates the scenario of maize yield in India using several conventionally developed autoregressive integrated moving average (ARIMA) models. The forecasting performances of the generated models were evaluated using akaike information criterion (AIC), root mean square error (RMSE) and mean absolute percentage error (MAPE). The best fitted model was ARIMA(2,1,0) with drift, having AIC value of 894.95, RMSE value of 148.05, and MAPE value of 7.71%. Additionally, a comparative performance assessment of the fitted models were made with the automatically generated model viz., ARIMA(1,1,2) with drift, which was obtained on using $auto.arima$ () function in *R-studio*. Furthermore, the Ljung-Box test was performed for diagnostic checking of residuals of the generated models. The results of the analysis revealed that ARIMA(1,1,2) with drift model was slightly more precise as compared to ARIMA(2,1,0) with drift. The forecast values of maize yield for five consecutive years were obtained with 80% and 95% prediction

intervals using ARIMA(1,1,2) with drift model. The findings of the study revealed that the trend of maize yield is significantly rising over the recent years, which is a good sign for policymakers and scientists regarding development of strategies pertaining to global food trade and nutritional security.

Keywords: ARIMA; time series; stationarity; autocorrelation; residual.

1 Introduction

The agriculture sector holds a prominent position for sustaining global food demand and nutritional security. Food grain crops are essential source of carbohydrates, proteins, fibers, and other vital nutrients, which have enormous health benefits. The demand for food grain crops are increasing at a rapid rate worldwide. To meet the global food demand, efforts are needed for enhancement of crop yield through improved varieties, policy support, subsidies, resource allocation, market development, and farmers' motivation towards cultivation of profitable crops. However, the yield of agricultural crops is influenced by several extraneous factors like climate change, pest attacks, resource scarcity, and land acquisition for constructions and urbanization. Hence, it becomes indispensable to analyze the long-term trend of agricultural produce to boost sustainable agriculture through effective policy formulation regarding inventory management, transportation, pricing and trade of agricultural produce.

Several attempts have been made in the past regarding the development of statistical models for forecasting the scenario of crop yield for various crops. For instance, Choudhury and Jones (20[illegible] applied several forecasting methods for evaluating crop yield estimates in Ghana. In the study, yield forecasts were compared using Simple Exponential Smoothing, Double Exponential Smoothing, Damped-Trend Linear Exponential Smoothing, and ARMA models. Sahu and Mishra (2014) analyzed the instability and trend in area, production and yield of maize in major states of India pertaining to the period from 1950 to 2009. Tripathi et al. (2014) utilized ARIMA models for forecasting area, production, and productivity of rice in Odisha on the basis of historical data concerning the period from 1950-51 to 2008-09. Akossou et al. (2016) conducted spatial and temporal analysis of maize crop yields in various agro-ecological zones of Benin from 1987 to 2007. Ilić et al. (2016) forecasted corn production in Serbia using ARIMA model by considering the period from 1947 to 2014. Cheng-Zhi et al. (2017) projected Chinese maize yield on the basis of ARIMA model. Mohammad et al. (2022) forecasted maize production in Bangladesh using yearly data for the growing seasons 1970-71 to 2019-20, and applying ARIMA and mixed model approach. Some other noteworthy contributions towards statistical modeling and time series analysis of crops have been made by Mesike (2012), Rathod et al. (2017), Ji et al. (2019), Yonar et al. (2021), Kumar et al. (2024), Maheshnath et al. (2024), Prakash et al. (2025), Rana et al. (2025), and Singh and Kumar (2025).

India is ranked among the top producers of maize at global level with production of 37.67 million tons over 11.24 million hectares of area during the year 2023-24 (ESE Division, 2024), Also, as per the exploratory data analysis conducted by Tubiello et al. (2025), it is revealed that India holds significant position in global ranking among the top ten countries with the largest area of maize. Considering the significance of maize crop in Indian economy, an attempt is made in this paper to investigate and forecast the scenario of maize yield in India using autoregressive integrated moving average (ARIMA) models.

2 Materials and Methods

Secondary time series data on maize yield in India encompassing the period from 1954 to 2023 were obtained from the repository of Economics, Statistics & Evaluation (ESE) Division, Department of Agriculture and Farmers Welfare, India. The analysis is carried out by generating ARIMA models based on the concerned data using *R-studio* software. The performances of models have been evaluated using various model fit statistics criteria viz., akaike information criterion (AIC), root mean square error (RMSE) and mean absolute percentage error (MAPE). Moreover, the diagnostic checking of residuals has been performed using Ljung-Box test.

The steps involved in ARIMA model fitting are elaborated below:

a) ARIMA model specification

The ARIMA model is a generalization of the ARMA (Autoregressive Moving Average) model to handle non-stationary time series. The integrated component 'I' refers to differencing the time series to achieve stationarity

(i.e., constant mean). The ARIMA model is generally written as ARIMA(p, d, q), where p, d, q refer to the order of autoregression, differencing and moving averages components, respectively.

The mathematical form of ARIMA(p, d, q) model is given by

$$y'_t = c + \sum_{i=1}^{p} \emptyset_i \, y'_{t-i} + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j} + \epsilon_t$$

Where

$y'_t \, , y'_{t-i} =$ differenced version of the time series
$c =$ constant term (or drift)
$\emptyset_i =$ autoregressive parameters ; $(i = 1, 2, \dots, p)$
$\theta_j =$ moving average parameters ; $(j = 1, 2, \dots, q)$
$\epsilon_t \, , \epsilon_{t-j} =$ random error terms

b) Model selection

The initial step in ARIMA model fitting is the identification of optimal orders (i.e., p, d, q). The order of differencing 'd' is identified using a statistical test for stationarity. In the present study, Augmented Dickey-Fuller (ADF) test is considered for the same on considering the null hypothesis (H_0) that the series is non-stationary against the alternative hypothesis (H_1) that the series is stationary. The decision regarding the rejection or acceptance of the null hypothesis (H_0) is made on the basis of p-value. If the p-value comes out to be less than 0.05, then the null hypothesis (H_0) is rejected, and the conclusion is made that the series is stationary (i.e., the series has a constant mean and variance).

Furthermore, the orders of autoregressive and moving average components, i.e., 'p' and 'q', are determined by analyzing the partial autocorrelation function (PACF) and autocorrelation function (ACF) of the differenced time series, respectively. The accuracy of the fitted models are measured using akaike information criterion (AIC), root mean square error (RMSE) and mean absolute percentage error (MAPE), which are symbolically mentioned below:

$$AIC = -2log(L) + 2N,$$
i. e., $$AIC = -2log(L) + 2(p + q + k + 1),$$

where L is the likelihood of data, and N denotes the number of model parameters. Also, $k = 1$ if $c \neq 0$, and $k = 0$ if $c = 0$ (Hyndman and Athanasopoulos, 2018).

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100$$

Here, 'n' represents the number of observed values. Also, y_t denotes the actual maize yield at time 't', and $\hat{y}_t$ refers to the predicted maize yield at time 't'.

c) Parameter estimation

After the identification of optimal orders of the ARIMA model, the next step is the estimation of model parameters. The parameters are usually estimated using the method of maximum likelihood.

d) Diagnostic checking of residuals

The model adequacy is inferred on checking the autocorrelation and normality of the residuals. In the analysis, the Ljung-Box test is performed for diagnostic checking of residuals for resembling white noise, i.e., to identify whether the residuals are uncorrelated and identically distributed. Also, for assessing the normality of residuals, the in-built function viz., *checkresiduals* () function in *R-studio* is utilized for various fitted ARIMA models.

The Ljung-Box test statistic is given by

$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{r}_k^2}{(n-k)}$$

where

$n =$ number of observations
$h =$ number of lags being tested
$\hat{r}_k =$ estimated autocorrelation coefficient of the residuals at the k^{th} lag

Under the assumption of null hypothesis (H_0) that the residuals are independently and identically distributed, the Ljung-Box test statistic (Q) follows a chi-square distribution with 'h' degrees of freedom. Also, the critical region for rejection of null hypothesis (H_0) at 'α' level of significance is given by

$$Q > \chi^2_{(1-\alpha),h}$$

3 Results and Discussion

The secondary time series data on maize yield in India encompassing the period from 1954 to 2023 is summarized in Table 1, and the graphical plot of maize yield is demonstrated in Fig. 1. On using the ADF test, it is observed that the original time series on maize yield is non-stationary with test result: Dickey-Fuller = 1.0752, Lag order = 4, p-value = 0.99. Consequently, the first differencing of the series, with logarithmic transform, is utilized and the ADF test is applied on the differenced series for checking the stationarity. In this case, the first order differenced series becomes stationary (i.e., the differenced series has constant mean and variance) with test result: Dickey-Fuller = -5.7496, Lag order = 4, p-value = 0.01. The plot of first order differenced stationary series on maize yield is demonstrated graphically in Fig. 2.

In order to develop appropriate ARIMA(p,d,q) models, the PACF and ACF plots of the first order differenced series on maize yield are obtained, which are represented in Figs. 3 and 4, respectively. It is observed from the PACF plot of Fig. 3 that the lags 1 and 2 are significantly outside the threshold limits. In a similar manner, from the ACF plot of Fig. 4, it is revealed that the lags 0 and 1 are significantly outside the threshold limits. Hence, the possible orders of autoregressive (AR) component, i.e., 'p' are taken as 1 and 2. Also, the possible orders of moving average (MA) component, i.e., 'q' are taken as 0 and 1. Furthermore, the order of differencing 'd' is taken as 1. On using a combination of these possible orders, several ARIMA models are developed for the analysis of maize yield in India, which are enlisted in Table 2, along with the estimates of model parameters viz., autoregressive parameters, moving average parameters, and the drift parameter. The accuracy of the fitted ARIMA models are measured using model fit statistics criteria viz., akaike information criterion (AIC), root mean square error (RMSE) and mean absolute percentage error (MAPE), and the findings are presented in Table 2. Moreover, on using $auto.arima$ () function in *R-studio*, the automatically generated ARIMA model for maize yield is obtained as ARIMA(1,1,2) with drift, which is elaborated along with model parameters and model fit statistics in Table 2.

From Table 2, it is revealed that the fitted ARIMA models with drift terms are more precise as compared to their counterpart models without drift terms, in terms of achieving least values of AIC and RMSE, along with comparable values of MAPE. Also, among the several fitted models, the best fit model is found to be ARIMA(2,1,0) with drift, having least values for model fit statistics criteria, i.e., AIC value of 894.95, RMSE

value of 148.05, and MAPE value of 7.71%. Furthermore, the performance of automatically generated model viz., ARIMA(1,1,2) with drift, is slightly better as compared to the conventionally developed best fit model, i.e., ARIMA(2,1,0) with drift.

Table 1. Time series data on maize yield (kg/ha) in India during 1954-2023

Sl. No.	Year	Yield (kg/ha)	Sl. No.	Year	Yield (kg/ha)	Sl. No.	Year	Yield (kg/ha)	Sl. No.	Year	Yield (kg/ha)
1	1954	795	21	1974	948	41	1994	1570	61	2014	2632
2	1955	703	22	1975	1203	42	1995	1595	62	2015	2563
3	1956	819	23	1976	1060	43	1996	1720	63	2016	2689
4	1957	772	24	1977	1051	44	1997	1711	64	2017	3065
5	1958	810	25	1978	1076	45	1998	1797	65	2018	3070
6	1959	938	26	1979	979	46	1999	1792	66	2019	3006
7	1960	925	27	1980	1159	47	2000	1822	67	2020	3199
8	1961	956	28	1981	1162	48	2001	2000	68	2021	3387
9	1962	994	29	1982	1145	49	2002	1681	69	2022	3545
10	1963	995	30	1983	1352	50	2003	2041	70	2023	3351
11	1964	1010	31	1984	1456	51	2004	1907			
12	1965	1005	32	1985	1146	52	2005	1938			
13	1966	964	33	1986	1282	53	2006	1912			
14	1967	1123	34	1987	1029	54	2007	2335			
15	1968	997	35	1988	1395	55	2008	2414			
16	1969	968	36	1989	1632	56	2009	2024			
17	1970	1279	37	1990	1518	57	2010	2542			
18	1971	900	38	1991	1376	58	2011	2478			
19	1972	1094	39	1992	1676	59	2012	2566			
20	1973	965	40	1993	1602	60	2013	2676			

(*Source: Economics, Statistics & Evaluation Division, DA&FW, India*)

Table 2. Model parameters and model fit statistics of the various ARIMA models for maize yield in India

Model	Model Parameters					Model Fit Statistics		
	Autoregressive Parameters		Moving Average Parameters		Drift			
	$\emptyset_1$	$\emptyset_2$	θ_1	θ_2	c	AIC	RMSE	MAPE
ARIMA(1,1,0)	-0.387	-	-	-	-	907.35	167.12	7.85
ARIMA(1,1,0) with drift	-0.458	-	-	-	38.707	901.65	157.97	8.14
ARIMA(2,1,0)	-0.463	-0.199	-	-	-	906.56	163.68	7.80
ARIMA(2,1,0) with drift	-0.615	-0.345	-	-	39.135	894.95	148.05	7.71
ARIMA(1,1,1)	-0.191	-	-0.249	-	-	907.84	165.27	7.77
ARIMA(1,1,1) with drift	-0.145	-	-0.486	-	38.927	896.26	149.52	7.89
ARIMA(2,1,1)	-0.834	-0.341	0.389	-	-	908.10	163.11	7.76
ARIMA(2,1,1) with drift	-0.411	-0.254	-0.233	-	39.128	896.37	147.40	7.77
ARIMA(1,1,2) with drift	0.9212	-	-1.821	0.919	47.310	887.41	135.29	6.92

The residual diagnostics of the various models for maize yield are performed using Ljung-Box test, and the findings are depicted in Table 3. In the Ljung-Box test, the assumption made under the null hypothesis (H_0) is that the residuals of the generated model have no autocorrelation, which is tested against the alternative hypothesis (H_1) that the residuals are autocorrelated.

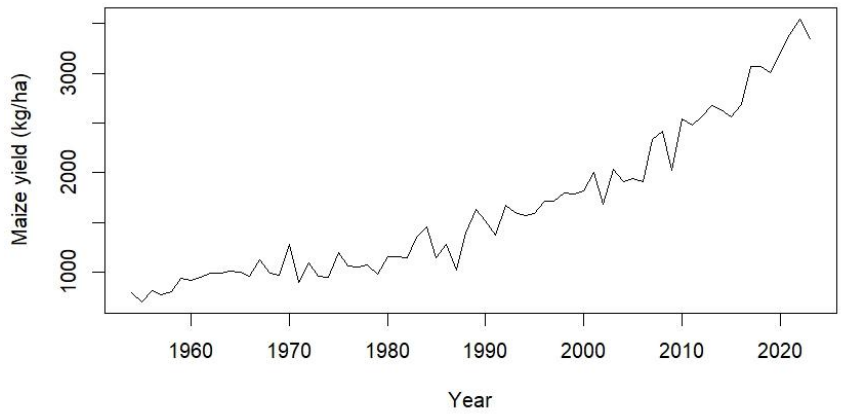

Fig. 1. Plot of maize yield (kg/ha) in India during 1954-2023

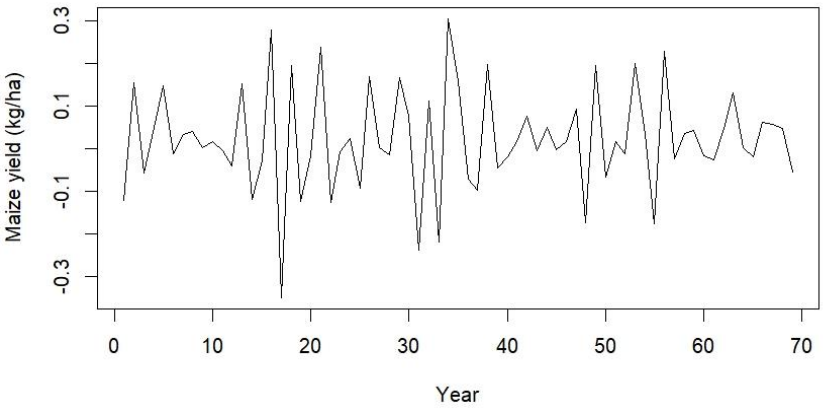

Fig. 2. Plot of first order differenced stationary series on maize yield (kg/ha)

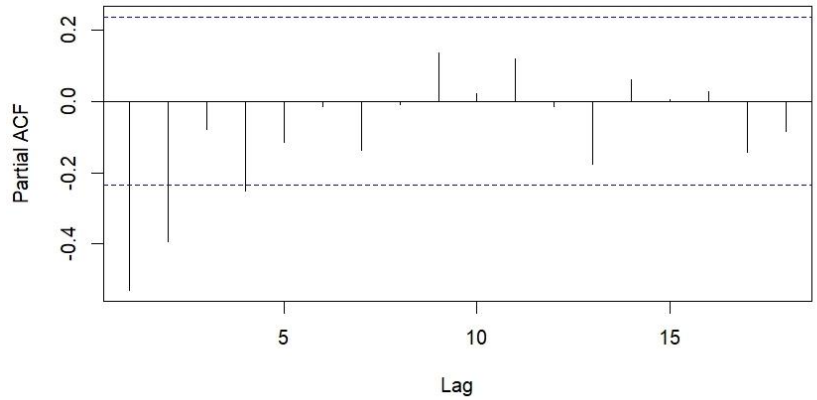

Fig. 3. PACF plot of first order differenced stationary series on maize yield

From Table 3, it is revealed that the Ljung-Box test statistic (Q) achieves p-value greater than 0.05 for residuals of each fitted model, which indicates that the null hypothesis (H_0) is accepted, and hence it can be concluded that the residuals of the various generated models are uncorrelated. Moreover, on using the in-built *checkresiduals* () function in *R-studio* for the concerned models, it is observed that the residuals are normally distributed. As the residuals of the various generated models are uncorrelated and normally distributed, it can be inferred that all the developed models are adequate for forecasting the scenario of maize yield in India.

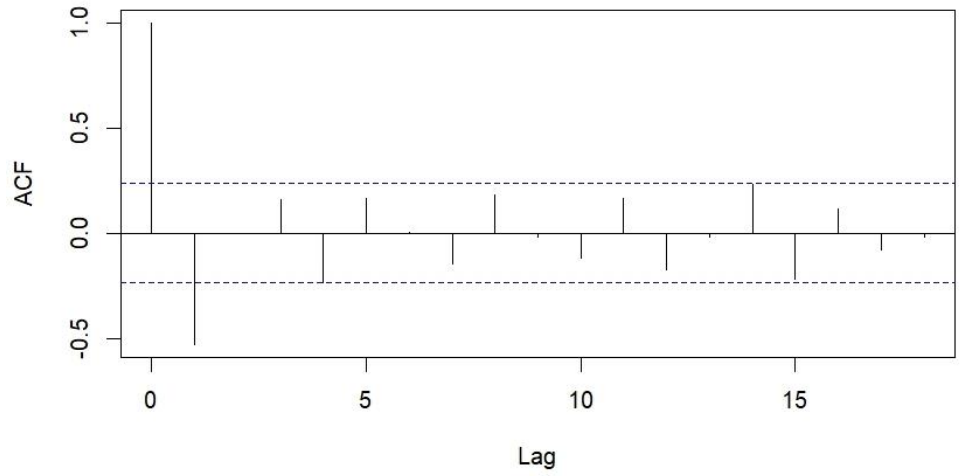

Fig. 4. ACF plot of first order differenced stationary series on maize yield

Table 3. Residual diagnostics of the various ARIMA models for maize yield in India

Model	Residual diagnostics	
	Ljung-Box test statistic (Q)	***p*-value**
ARIMA(1,1,0)	10.348	0.323
ARIMA(1,1,0) with drift	10.764	0.292
ARIMA(2,1,0)	6.215	0.623
ARIMA(2,1,0) with drift	6.716	0.568
ARIMA(1,1,1)	7.733	0.460
ARIMA(1,1,1) with drift	8.034	0.430
ARIMA(2,1,1)	6.699	0.461
ARIMA(2,1,1) with drift	7.313	0.397
ARIMA(1,1,2) with drift	1.969	0.962

Also, as the ARIMA (1,1,2) with drift model is preferable over other fitted models, in terms of achieving least values of AIC, RMSE and MAPE, the forecast values for maize yield in India are obtained with 80% and 95% prediction intervals using the concerned model for five successive years viz., 2024-2028, and the findings are elaborated in Table 4. Furthermore, the plot of observed and forecasted maize yield in India is depicted graphically in Fig. 5.

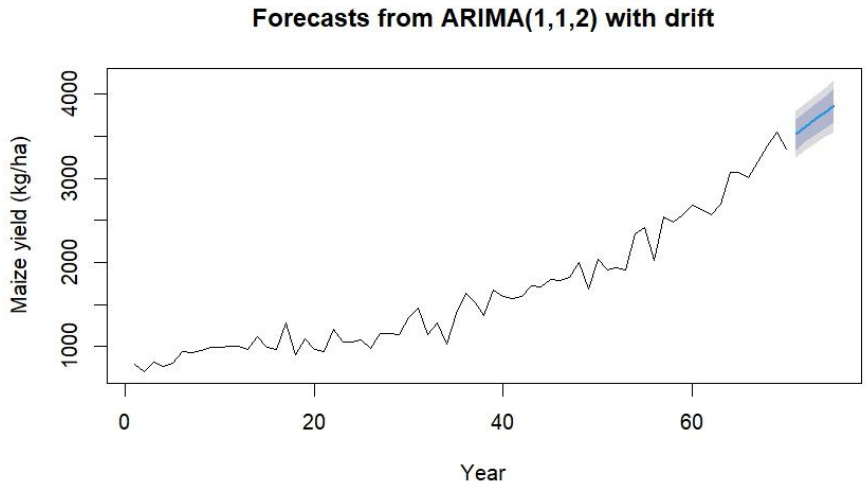

Fig. 5. Plot of observed and forecasted maize yield (kg/ha) in India

Table 4. Forecast values for maize yield in India using ARIMA (1,1,2) with drift model

Year	Forecasted Yield (kg/ha)	Prediction Intervals			
		80%		95%	
		LCL	UCL	LCL	UCL
2024	3520.25	3340.29	3700.20	3245.02	3795.47
2025	3609.43	3428.56	3790.30	3332.81	3886.04
2026	3695.31	3511.17	3879.45	3413.69	3976.93
2027	3778.15	3587.41	3968.89	3486.44	4069.86
2028	3858.19	3657.11	4059.28	3550.66	4165.73

(Note: LCL= Lower Control Limit, and UCL= Upper Control Limit)

The Fig. 5 reveals that the forecast values are significantly rising for the consecutive years, i.e., 2024-2028, which indicates that there will be a noteworthy rise in maize yield for upcoming years in India.

4 Conclusion

Maize holds a prominent position, after rice and wheat, as a highly nutritious cereal crop which is widely consumed in processed form, for instance, cornflakes, cornflour, snacks, baby's food such as cerelac, and much more. Maize is gaining huge familiarity at national as well as at global level. In view of the given fact, the present study was carried out for exploring and forecasting the scenario of maize yield in India. A comparative assessment of conventional and automated generated ARIMA models was made using well-known model fit statistics criteria viz., AIC, RMSE and MAPE. In addition, the Ljung-Box test was used for diagnostic checking of residuals of the generated models.

The results of the analysis revealed that all the generated models achieved MAPE values below 9%, with least value of 6.92% for ARIMA(1,1,2) with drift. Also, the findings of residual diagnostics exhibited that the residuals of the generated models were white noise, i.e., residuals were uncorrelated and normally distributed, which indicated that all the generated models were adequate for forecasting the scenario of maize yield in India. Furthermore, among the various conventional models, the best fitted model was found to be ARIMA(2,1,0) with drift. Moreover, the precision of automated model viz., ARIMA(1,1,2) with drift was slightly better as compared to ARIMA(2,1,0) with drift. Hence, the forecast values for maize yield in India were obtained using ARIMA(1,1,2) with drift for five successive years viz., 2024-2028, with prediction intervals of 80% and 95%.

The findings of the study revealed a noteworthy rise in maize yield of India for the subsequent years 2024 to 2028. Hence, the present study offers significant insights towards the observed and forecasted trend of maize yield in India. The results of the analysis can be effectively used by the scientists and policymakers regarding formulation of strategies pertaining to global food trade and nutritional security.

Disclaimer (Artificial intelligence)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

Competing Interests

Authors have declared that they have no known competing financial interests or non-financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Akossou, A.Y.J., Attakpa, E.Y., Fonton, N.H., Sinsin, B. and Bosma, R.H. (2016). Spatial and temporal analysis of maize (Zea mays) crop yields in Benin from 1987 to 2007. *Agricultural and Forest Meteorology*, *220*, 177-189.

Cheng-Zhi, C., Fang, W. and Ying, L. (2017). Chinese maize yield projected on ARIMA model basis. *International Journal of Agricultural and Statistical Sciences*, *13*(2), 403-408.

Choudhury, A. and Jones, J. (2014). Crop yield prediction using time series models. *Journal of Economics and Economic Education Research*, *15*(3), 53-67.

ESE Division (2024). Agricultural statistics at a glance. DA&FW, Government of India.

Hyndman, R.J. and Athanasopoulos, G. (2018). Forecasting: Principles and Practice. OTexts.

Ilić, I., Jovanović, S. and Janković-Milić, V. (2016). Forecasting corn production in Serbia using ARIMA model. *EkonomikaPoljoprivrede*, *63*(4), 1141-1156.

Ji, Y., Zhou, G., Wang, L., Wang, S. and Li, Z. (2019). Identifying climate risk causing maize (Zea mays L.) yield fluctuation by time-series data. *Natural Hazards*, *96*, 1213-1222.

Kumar, M., Singh, G., Singh, S. and Mishra, A. (2024). Performance of the major pulses crop in India: Growth and instability. *Asian Journal of Research in Crop Science*, 9(4), 348-357.

Maheshnath, M., Kumari, R.V., Suhasini, K., Reddy, D.S. and Meena, A. (2024). Forecasting maize production in Telangana state using arima model. *Archives of Current Research International*, 24(6), 223-229.

Mesike C.S. (2012). Short term forecasting of Nigerian natural rubber exports. *Wudpecker Journal of Agricultural Research*, 1(10), 396-400.

Mohammad, N., Islam, M.A., Rahman, M.M., Ahmed, I. and Mahboob, M.G. (2022). Forecasting of maize production in Bangladesh using time series data. *The Bangladesh Journal of Agricultural Economics*, *43*(2), 18-32.

Prakash, G., Kumar, M.,Rana, S.K., and Gowda K.E., S. (2025). A statistical approach for assessment of growth rate and instability of wheat in selected states of India. *Journal of Modern Applied Statistical Methods*, 24(1), 76-89.

Rana, S.K., Kumar, M., Supriya, Mishra, P. and Tiwari, A. (2025). Comparative analysis of instability and growth of pearl millet and maize in India. *Environment and Ecology*, 43(2), 462-467.

Rathod, S., Singh, K.N., Arya, P., Ray, M., Mukherjee, A., Sinha, K., Kumar, P. and Shekhawat, R.S. (2017). Forecasting maize yield using ARIMA-Genetic Algorithm approach. *Outlook on Agriculture*, *46*(4), 265-271.

Sahu, P.K. and Mishra, P. (2014). Instability in production scenario of maize in India and forecasting using ARIMA model. *International Journal of Agricultural and Statistical Sciences*, *10*(2), 425-435.

Singh, G. and Kumar, M. (2025). A statistical approach for analysis of trend pattern of pigeon pea in India. *Journal of Agriculture and Ecology Research International*, 26(1), 1-12.

Tripathi, R., Nayak, A.K., Raja, R., Shahid, M., Kumar, A., Mohanty, S., Panda, B.B., Lal, B. and Gautam, P. (2014). Forecasting rice productivity and production of Odisha, India, using autoregressive integrated moving average models. *Advances in Agriculture*, *2014*(1), 1-9.

Tubiello, F.N., Fabi, C., Conchedda, G., Casse, L. and Bottini, G. (2025). Comparison of WorldCereal with FAOSTAT data: an exploratory analysis. *FAO Statistics Working Paper Series*, No. 25-46. Rome, FAO.

Yonar, A., Yonar, H., Mishra, P., Kumari, B., Abotaleb, M. and Badr, A. (2021). Modeling and forecasting of wheat of South Asian region countries and role in food security. *Advances in Computational Intelligence*, *1*, 1-8.