The Andhra Agric. J 71 (4): 470-477, 2024
doi :10.61657/aaj.2024.164

An empirical analysis on forecasting area, production and productivity of mustard: A case study in Jalpaiguri District of West Bengal

Suma Deepika and Satyananda Basak
Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyala, West Bengal, India

ABSTRACT

Forecasting area, production, and productivity is vital for optimizing agricultural planning, stabilizing markets, and supporting economics. Accurate predictions help ensure resource allocation, price stability, and improved farmers' incomes. This study aims to forecast the area, production and productivity of mustard in Jalpaiguri district of West Bengal using ARIMA and ARIMAX. Annual data from 1977 to 2022 on mustard cultivation were analyzed alongside weather variables such as rainfall, maximum & minimum temperatures and data on fertilizer consumption. In thes way ARIMA model, which accounts for past values and random shocks, was compared with the ARIMAX model, which incorporates exogenous variables. The findings revealed that the ARIMAX model outperformed the ARIMA model for area, production and productivity in terms of forecasting accuracy, with lower values of RMSE, MAE and MAPE. The study emphasizes the importance of accurate forecasting in agricultural planning, contributing to more efficient resource allocation and price stabilization. The results indicate a positive outlook for mustard cultivation in Jalpaiguri.

Key words: *ARIMAX, ARIMA, Exogenous variables and Forecasting*

Time series forecasting is a method used to predict future values based on previously observed values. It involves analyzing data points recorded at specific time intervals rather than randomly, to identify patterns, trends, and seasonal variations. An important approach to time series forecasting is ARIMA (autoregressive integrated moving average) modeling which aims to describe the autocorrelations in the data. ARIMAX, a special kind of regression model, is the extension of ARIMA model that accounts important exogenous variables in the model development. This article attempts to focus on forecasting the area, production and productivity of mustard in Jalpaiguri district of West Bengal. Mustard is one of the most essential and highly demanded edible oilseed crops grown in West Bengal, accounting for 6.15% of the nation's mustard production MoA & FW, 2nd Advanced Estimate 2023-2024). Within India, West Bengal occupies 5th position in area and production. Forecasting the area, production, and productivity of mustard is important because it ensures efficient allocation of inputs like seeds, water, and fertilizers, aids in predicting supply levels and stabilizing prices. Banakara *et al.* (2022) evaluated weather parameter-based pre-harvest yield forecast models for wheat in Gujarat's Saurashtra region, comparing multiple linear regression (MLR), Time Delay Neural Network (TDNN), and ARIMAX models. Their study found that ARIMAX and TDNN approaches outperformed the conventional MLR technique, providing more accurate and reliable forecasts. The ARIMAX model, in particular, demonstrated consistent performance during both the model training and forecast periods, suggesting its suitability for reliable pre-harvest yield predictions. Dharmaraja *et al.,* (2020) conducted an empirical analysis of crop yield forecasting in India, focusing on Bajra yield in Rajasthan's Alwar district. In this linear regression and time-series models were evaluated by emphasizing the importance of selecting auxiliary variables based on crop growth stages. It was highlighted the ARIMAX model as the best for forecasting of Bajra yield by showcasing its proficiency in incorporating historical data and external environmental factors.

MATERIAL AND METHODS

Data description

Annual data from 1977 to 2022 on area ('000 ha), production ('000 tons) and productivity

(kg/ha) of Mustard in Jalpaiguri is collected from Statistical Abstract, Govt. of West Bengal. for 1977-2014 and Directorate of Agriculture, Govt. of W.B. for 2015-2022. Daily rainfall (mm), maximum temperature (ÚC) and minimum temperature (ÚC) were collected from I.M.D., Pune transformed into annual data by taking average for all weather variables. The weather variables are calculated at different stages of crop growth from the duration of November 15 to February 15(90 days) as vegetative pha reproductive phase and maturity phase respectively. The missing values were substituted using Imputation by Moving Average method (Moritz and Beielstein, 2017). Data on fertilizer Consumption and Price (kg/ha) was collected from Statistical Abstract, Govt. of West Bengal.

RESULTS AND DISCUSSION

ARIMA model

In ARIMA model, time series variable is assumed to be a linear function of past actual values and random shocks. An ARIMA (p, d, q) model is defined by the following equation.

$$\hat{y}_t = \mu + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} \cdots + \varphi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

In general, an ARIMA model is characterized by the notation ARIMA (p, d, q) where p, d and q denote orders of auto regression, differencing and moving average respectively. The underlying principle in the ARIMA model is that the time series under consideration is stationary. A Stationary series is one whose statistical properties such as mean, variance and covariance do not vary with time in other words a stationary series do not have any trend. Time plots will show the series to be roughly horizontal (although some cyclic behavior is possible), with constant variance (Hyndman, Athanasopoulos (2018). Even though in many practical situations non stationary time series arises, it can be transformed to stationary series through appropriate differencing. The Box Jenkins method for finding a good ARIMA model involves three steps: 1) Identification, 2) Estimation and 3) Diagnostic checking.

Model is tentatively selected based on the orders of *p* and *q* through ACF and PACF plots at the identification stage and the best model was selected based on the criterion which is having lowest values of AIC, AICc and BIC and the selected models are evaluated by the accuracy measures such as RMSE, MAE and MAPE. Adequacy of model is tested through diagnostic checking which consists of visualizing ACF and PACF plot of residuals and statistical tests.

$$AIC = -2 \log l(\hat{\theta}) + 2k$$

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

$$BIC = -2 \log l(\hat{\theta}) + k \log n$$

Where, k= Number of parameters and $\hat{\theta}$ is the maximum likelihood estimates of the model parameters.

$$Mean\ Absolute\ Error\ (MAE) = (1/n) * \Sigma |y_i - \hat{y}_i|$$

$$Mean\ Absolute\ Percentage\ Error\ (MAPE) = (1/n) * \Sigma |(y_i - \hat{y}_i)/y_i| * 100$$

$$Root\ Mean\ Square\ Error\ (RMSE) = \sqrt{(\Sigma (y_i - \hat{y}_i)^2 / n)}$$

n = number of observations.

y_i = actual value of the dependent variable for observation i.

$\hat{y}_i$ = predicted value of the dependent variable for observation

ARIMAX model (Autoregressive Integrated Moving average with Exogenous variables)

ARIMAX, a special kind of regression model, is the extension of ARIMA model that accounts significant exogenous variables in the model development, it can be written as

$$y_t = c + \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + \varepsilon_t$$

Where y_t is the dependent variable and β_k are the coefficients of exogenous variables and x_{tk} are the exogenous variables at time t. In this regression process is used in selecting and including significant exogenous variable into the ARIMAX model. Stepwise regression is an automatic and step-by-step iterative construction of a regression model that

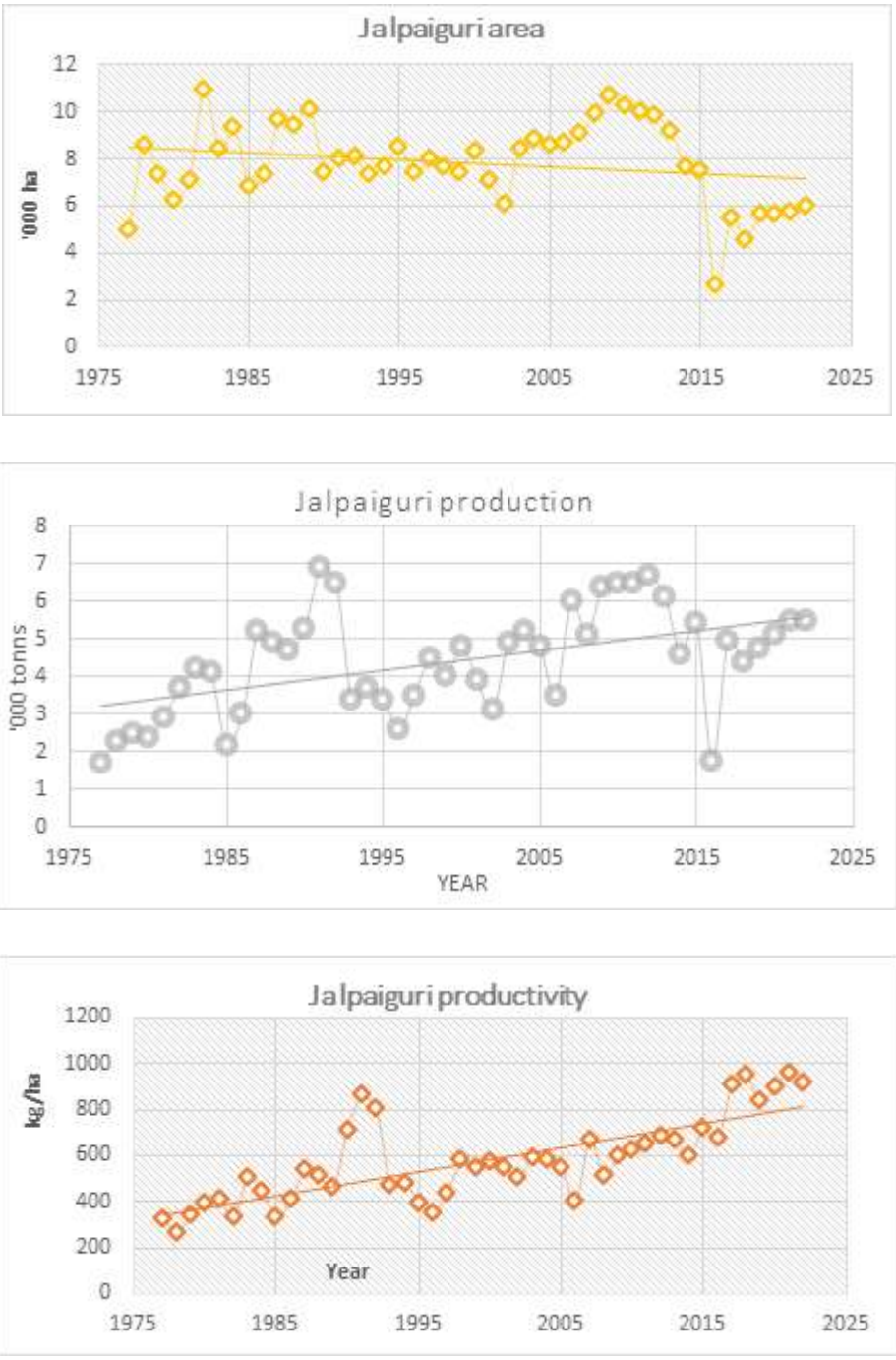

Fig 1: Time plots of area production and productivity

Table 1: ADF test results of Jalpaiguri area, production and productivity

ADF TEST	Augmented Dickey-Fuller	Differencing order	p-value
Area	-2.34	d=0	0.43
	-3.93	d=1	0.02
Production	-2.95	d=0	0.19
	-3.65	d=1	0.03
Productivity	-2.29	d=0	0.45
	-2.65	d=1	0.31
	-6.61	d=2	0.01

The series is found to be stationary at first differencing for area and production whereas for productivity the series becomes stationary at 2nd differencing. The ACF and PACF plots helps in choosing the appropriate values for p and q. The ACF plot is showing significant spike at lag 1 which gives MA order i.e., q=1, for area, production and productivity. PACF graph is showing significant spike at lag 1 for area and production whereas for productivity the significant spike is at lag 3 which gives the AR orders i.e., p=1 and p=3 respectively.

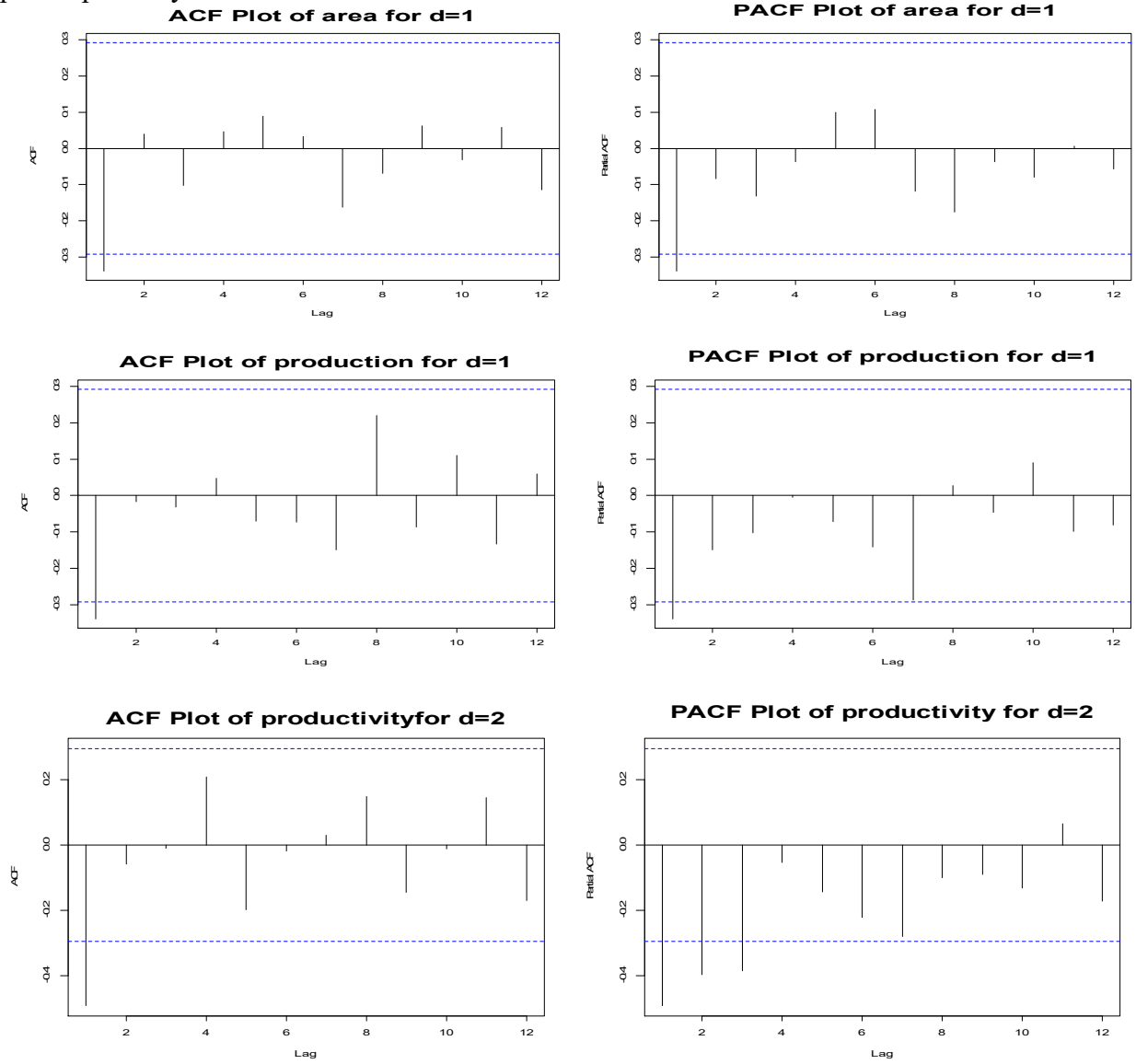

Fig 2: ACF and PACF plots of differenced series

For ARIMAX model, the variables which are found to be significant in stepwise regression. Estimates of the coefficients along with its standard error and significance are shown in table 2

Table 2: Exogenous variables significance of ARIMAX model

	Exogenous variables	Estimate	Std. Error	Z value	p-value
Area	Maize area	0.159	0.061	2.598	0.009
Production	Rainfall (vegetative)	0.957	0.391	2.445	0.014
	phosphorus	0.078	0.038	2.064	0.038
productivity	Rainfall (vegetative)	121.381	40.809	2.974	0.002
	Rainfall (reproductive)	134.514	51.941	2.589	0.009

Based on the lowest AIC, AICc, and BIC values, the best models for area, production and productivity have been identified. The AIC, BIC and AICc values for the best model are shown in table 3. The best fitted models were further evaluated using accuracy measures: RMSE, MAE, and MAPE. A lower value indicates better model performance.

Table 3. AIC, AICc and BIC values for ARIMA and ARIMAX models

Variables	Model	AIC	AICc	BIC
Area	ARIMA (0,1,1)	122.88	123.24	126.04
	ARIMAX (0,1,1)	123.47	126.37	132.97
Production	ARIMA (1,1,1)	108.29	115.21	122.54
	ARIMAX (0,1,1)	106.48	110.48	117.56
productivity	ARIMA (0,2,1)	439.05	439.42	442.16
	ARIMAX (2,2,1)	438.76	447.93	454.31

The performances of the best fitted models of ARIMA and ARIMAX are validated by accuracy measures such as RMSE, MAE and MAPE as shown in table 4.

Table 4: Performance validation of ARIMA and ARIMAX models

Model	Accuracy measures		
Area	RMSE	MAE	MAPE
ARIMA (0,1,1)	1.629	0.808	25.562
ARIMAX (0,1,1)	**1.133**	**0.561**	**17.485**
Production	RMSE	MAE	MAPE
ARIMA (1,1,1)	1.443	0.957	32.932
ARIMAX (0,1,1)	**1.083**	**0.716**	**24.794**
Productivity	RMSE	MAE	MAPE
ARIMA (0,2,1)	94.489	68.285	8.141
ARIMAX (2,2,1)	**79.93**	**68.251**	**7.928**

From the above table ARIMAX model is showing lowest error values for all the variables. The adequacy of the model is tested through diagnostic checking i.e., visualizing ACF and PACF plots of residual and statistical test for independence and normality. Fig 3 represents the ACF and PACF plot of residuals of best fitted models and statistical test results were shown in table 5 indicating that the fitted models are independent and following normal distribution.

Table 5: Ljung-Box test results

Test	Ljung-Box test		
ARIMA	Q*statistic	df	p-value
Area	6.619	6	0.357
Production	2.722	5	0.742
Productivity	10.222	6	0.115
ARIMAX	Q*statistic	df	p-value
Area	5.394	6	0.198
Production	5.394	6	0.494
Productivity	4.729	4	0.316

ARIMAX produced stabilized results that can be used for reliable forecasts. The forecasted values indicate a steady increase in mustard cultivation area, production, and productivity in West Bengal over the next five years. The area forecasts range from 6.51 to 6.80 thousand hectares, production from 3.83 to 4.96 thousand tons, and productivity from 918.49 to 951.69 kg/ha by 2027. Despite some variability, the overall trend is positive, highlighting a favorable outlook for mustard farming in the region. The forecasted values with 80% and 95% confidence intervals are presented in figure 4.

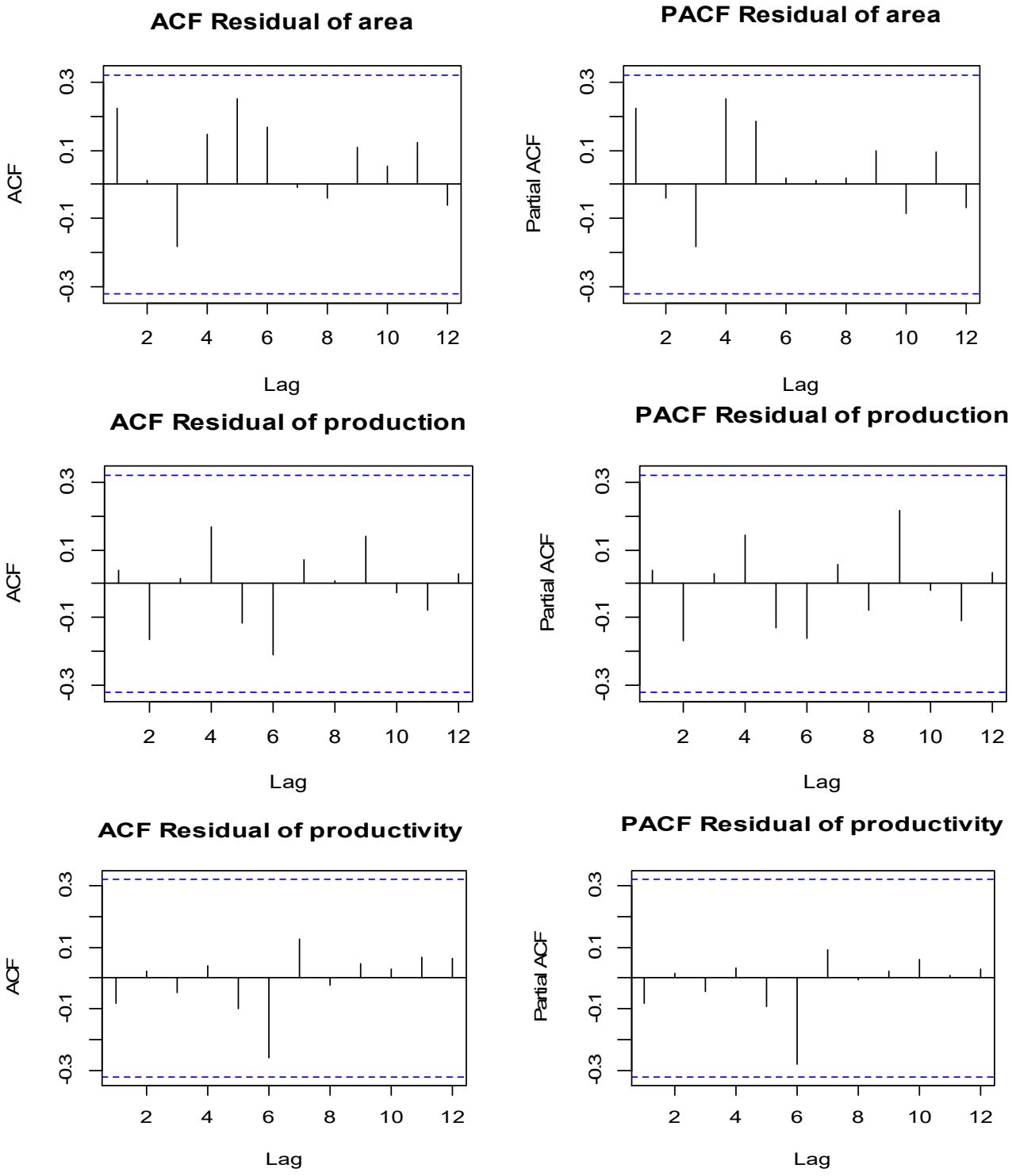

Fig 3: ACF and PACF plot of residuals of area, production and productivity of mustard

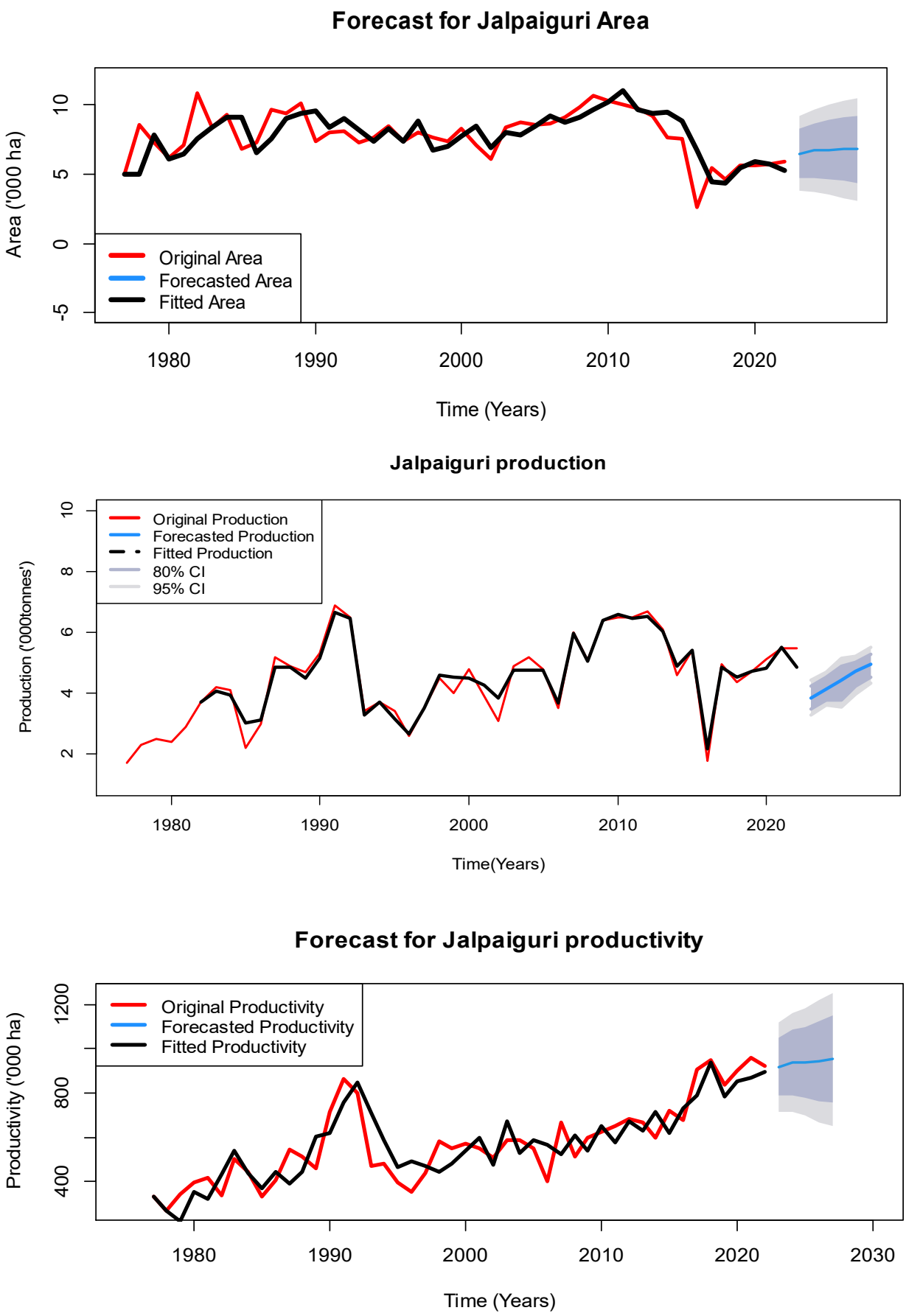

Fig 4: Forecasting of area, production and productivity

involves the selection of independent variables to be used in a final model. once the significant variables are included in the model, the subsequent procedure is same as ARIMA model.

Before modeling the time series data, it is important to check whether data is stationary or not. The time plots of area, production and productivity in fig. 1 below clearly shows that the data is not stationary (actually, it shows an increasing trend in production and productivity while decreasing trend in area) and the Augmented Dickey-Fuller (ADF) test result in table 1 also implies the same, which tests the null hypothesis data is not stationary. Appropriate differencing makes the series stationary.

This study has compared 2 models i.e., ARIMA and ARIMAX in terms of modeling and forecasting of area, production and productivity of mustard in Jalpaiguri district of West Bengal. ARIMAX model has outperformed ARIMA model for forecasting area, production and productivity. ARIMAX model include external regressors that can capture additional information not accounted for by ARIMA models, improving forecast accuracy. This study's findings contribute to the broader field of agricultural forecasting by demonstrating the effectiveness of ARIMAX model in capturing complex dynamics and improving forecast accuracy. The implications are significant for agricultural policy and planning, offering a pathway to more informed and strategic decisions that can enhance the productivity and sustainability of mustard cultivation in the region.

LITERATURE CITED

Banakara K B, Sharma N, Sahoo S, Dubey S K and Chowdary V M 2022. Evaluation of weather parameter-based pre-harvest yield forecast models for wheat crop: a case study in Saurashtra region of Gujarat. *Environmental Monitoring and Assessment*, *195*(1), 51.

Dharmaraja S, Jain V, Anjoy P and Chandra H 2020. Empirical analysis for crop yield forecasting in India. *Agricultural Research*, *9*(1), 132-138.

Ministry of Agriculture and Farmaer's Welfare, Government of India, Advanced Estimatimates. 2023-24.

Moritz S and Bartz-Beielstein T 2017. imputeTS: Time Series Missing Value Imputation in *R*. *R Journal* 9.1; 9(1):207–218. doi: http://dx.doi.org/10.32614/RJ-2017-009

Hyndman R J and Athanasopoulos G 2018. *Forecasting: principles and practice***,** 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2).

Received on 27.11.2024 and Accepted on 17.12.2024