Forecasting Mustard Yield Using Weather Parameters: An Application of the ARIMAX Model

Surbhi Kaushik[1], Nitin bhardwaj[2*], B. Parashar[3], Jitender Kumar[4]
[1,2]Department of Mathematics and Statistics, CCSHAU, Hisar
[3]Department of Mathematics, JSS Academy of Technical Education, Noida
[4] Directorate of Human resources Management, CCSHAU, Hisar
Department of Mathematics and Statistics, CCSHAU, Hisar
Email id: nitinbhardwaj.mdu@hau.ac.in

ABSTRACT

Accurate forecasting of crop yields is essential for effective agricultural planning, policymaking, and ensuring food security. This study focuses on forecasting mustard yield in the Bhiwani district of Haryana using the ARIMAX model, a time-series approach that incorporates weather parameters as exogenous variables. The analysis leverages historical data on mustard yield from 1980 to 2023 and weather parameters including minimum temperature, maximum temperature, and rainfall during the crop-growing season. The results demonstrated that the inclusion of weather variables in the ARIMAX model significantly improved its forecasting accuracy. Statistical validation metrics such as Root Mean Square Error (RMSE = 1.3) and Mean Absolute Error (MAE = 1.27) highlighted the superior performance of the ARIMAX model. Additionally, the Percent Relative Deviation (PRD) analysis showed a close alignment between the forecasted and observed mustard yields, with minimal deviations. The findings of this study underscore the importance of incorporating climatic factors into crop yield forecasting models, particularly for mustard, where weather conditions significantly influence productivity.

Keywords: Time Series, ARIMAX, RMSE, MAE, PRD

INTRODUCTION

Agriculture is the backbone of India's economy, and accurate crop yield forecasting plays a crucial role in ensuring food security, optimizing resource allocation, and guiding agricultural policymaking. Among various crops, mustard holds a significant position in India's oilseed production, contributing to the livelihood of farmers and meeting the country's edible oil demand. Haryana, particularly the Bhiwani district, is one of the major mustard-producing regions in the country. The productivity of mustard is heavily influenced by climatic factors such as temperature and rainfall, which fluctuate considerably during the crop's growing season . This variability underscores the need for accurate forecasting models that can integrate both historical yield data and weather parameters to provide reliable estimates. Traditional forecasting methods, such as the Auto-Regressive Integrated Moving Average (ARIMA) model, have been widely used for time series forecasting in agriculture. However, ARIMA is a univariate model that relies solely on historical crop yield data and cannot account for the influence of external factors, such as weather, on yield variations. This limitation restricts its applicability, especially in the context of climate-sensitive crops like mustard. To address this gap, the Auto-Regressive Integrated Moving Average with Exogenous Variables (ARIMAX) model has emerged as a powerful alternative. The ARIMAX model extends ARIMA by incorporating external variables (exogenous inputs), making it well-suited for agro-climatic studies. In this study, we develop and evaluate ARIMAX models to forecast mustard yield in the Bhiwani district of Haryana, using climatic factors such as minimum temperature, maximum temperature, and rainfall as exogenous variables. These weather parameters are critical during the mustard crop's growth stages, influencing germination, flowering, and seed development. The inclusion of these variables allows the ARIMAX model to capture the complex interactions between climatic conditions and crop yield, enhancing its predictive accuracy compared to ARIMA. Yogarajah *et al.* (2013) used the ARIMAX model to forecast annual paddy production in the Trincomalee district of Sri Lanka, incorporating rainfall as an input variable in the model. In another study, Hamjah (2014) developed ARIMAX time-series models to assess and forecast the impact of temperature and rainfall on the production of Bangladesh's primary spice crops. The ARIMAX(2, 1, 2) model was the best fit for forecasting chilli production, ARIMAX(2, 0, 1) for garlic, and ARIMAX(2, 1, 1) for ginger production. Kour *et al.* (2018) explored the feasibility of predicting rice production in the Kheda district of central Gujarat, India, using the ARIMAX time-series model. The model incorporated meteorological parameters to estimate the combined effects on rice production, demonstrating its potential for improving forecasting accuracy. By demonstrating the superior performance of the ARIMAX model, this research aims to contribute to the growing body of literature on climate-integrated crop forecasting and provide a robust methodological framework for agricultural planners and policymakers. The findings of this study hold practical significance for pre-harvest planning, resource allocation, and market regulation. Moreover, the methodological approach can be extended to other crops and regions, emphasizing the importance of integrating weather parameters into crop yield forecasting models to address the challenges posed by climate variability.

METHODOLOGY

ARIMAX stands for Autoregressive Integrated Moving Average with Exogenous Variables. It is a logical extension of traditional ARIMA modelling that incorporates independent variables to improve the explanatory power of the model. Conceptually, it combines regression and ARIMA modelling. When the AR and MA terms in a pure ARIMA model are insufficient to produce a model with an acceptable level of overall explanatory power, it is only to look for additional variable whose influence is on the dependent variable. A univariate process is said to follow ARIMA(p, d, q), if it can be represented as:

$$\phi_p(B)\nabla^d y_t = \theta_q(B)\varepsilon_t$$

Where p is order of Autoregressive, q is order of Moving average and d is differencing order. $\phi_p \ and \ \theta_q$ are the autoregressive and moving average polynomial respectively.

$\nabla^d = (1-B)^d, B(y_t) = y_{t-1}$,
$\{\varepsilon_t\} \sim (0, \sigma^2)$
A constant term may also be added to the right-hand side. if y_t is not mean-adjusted. The ARIMA methodology is followed in three stages, viz. identification, estimation and diagnostic checking. Parameters of the tentatively selected ARIMA model at the identification stage are estimated at the estimation stage, and adequacy of tentatively selected model is tested at the diagnostic checking stage. If the model is found to be inadequate, the three stages are repeated until satisfactory ARIMA model is selected for the time series under consideration. The ARIMA with exogenous variable (ARIMAX) model is a generalization of the ARIMA model, which is capable of incorporating an external input variable (X). Given a historic input vector (x_t), the ARIMAX model assumes the form,

$$\phi_p(B)\nabla^d y_t = \mu + v(B)x_t + \theta_q(B)\varepsilon_t$$

Where $v(B) = \beta_1 x_t + \beta_2 x_{t-1} + \cdots + \beta_M x_{t-m+1}$

The ARIMAX models has been developed using meteorological variables as input series, which are average maximum temperature (Tmx), average minimum temperature (Tmn) and total rainfall (Trf) for 30 fortnights during the entire crop period. Secondarydata on weather parameters was collected on a daily basis for the period spanning from September 15 to February 15, covering the years 1980 to 2023. To facilitate further analysis, this daily data was aggregated into fortnightly intervals by computing averages for maximum and minimum temperatures and calculating totals for rainfall.

For instance, the average maximum and minimum temperatures for the period from September 1 to September 15 represented the first fortnight, while the total rainfall during the same period represented the first fortnightly rainfall value. This approach was repeated for successive fortnights throughout the total growth period of the crop, resulting in 30 weather parameters such as Tmx_1, Tmx_2..., Tmx_{10}, Tmn_1, Tmn_2, …, Tmn_{10} and Trf_1, Trf_2…, Trf_{10}.Significant variables among all are selected using Stepwise Regression method. The study uses historical data on mustard yield (1980–2023) and weather parameters recorded during the crop's growing season. The data were divided into training (1980–2017) and testing (2018–2023) sets to evaluate the forecasting accuracy of the models.

Stepwise Regression Method is a statistical technique used to select a subset of predictor variables for building a regression model. It is an iterative approach that evaluates which variables to include or exclude from the model based on specific criteria, such as R^2, Akaike Information Criterion (AIC), or p-values. Check the residuals of the model to ensure they behave like white noise *i.e.* uncorrelated, constant variance and normally distributed on the basis of Ljung-Box test and Shapiro-Wilk test.

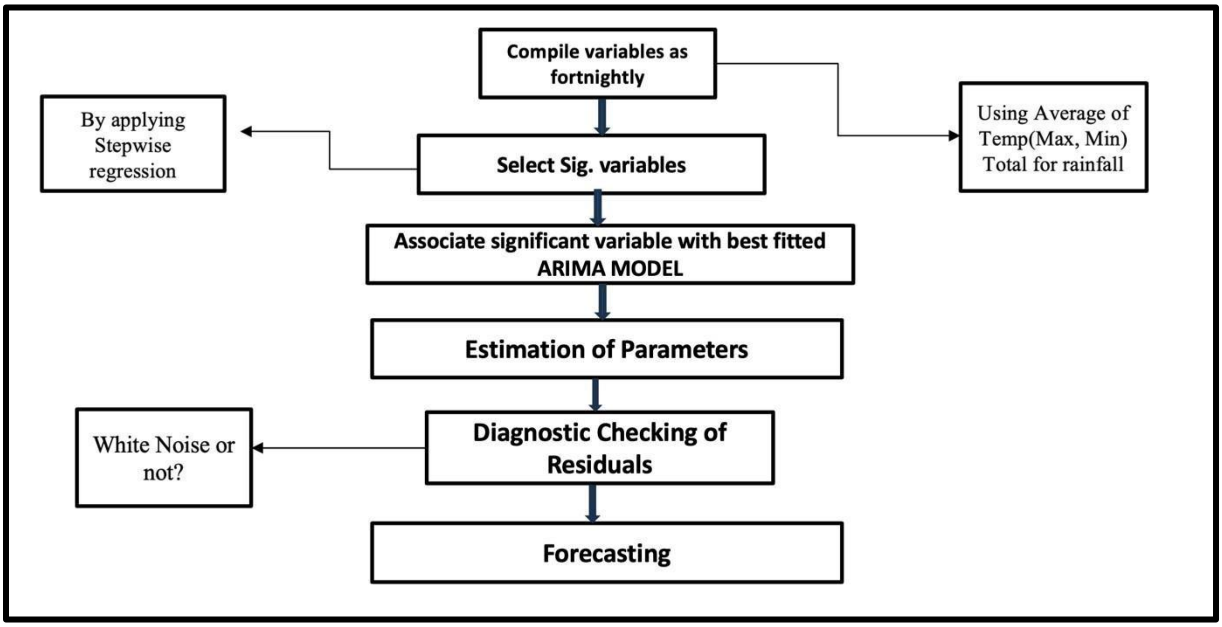
Compile variables as fortnightly
By applying Stepwise regression
Select Sig. variables
Using Average of Temp(Max, Min) Total for rainfall
Associate significant variable with best fitted ARIMA MODEL
Estimation of Parameters
White Noise or not?
Diagnostic Checking of Residuals
Forecasting

Figure 1: Flow Chart of forecasting using ARIMAX Model

RESULTS AND DISCUSSION

To analyse the mustard yield in Bhiwani from 1980 to 2023, we began by plotting the time series data to observe its overall trend and patterns. The visual inspection of the time series plot revealed a non-stationary behaviour, characterized by a clear trend over time. This suggested that the series would require further statistical validation for stationarity.

To confirm the non-stationarity of the series, we applied the Augmented Dickey-Fuller (ADF) test, a widely used statistical test to determine the presence of a unit root in time series data. The results of the ADF test, as presented in Table 1, indicated that the null hypothesis of a unit root could not be rejected, confirming that the time series is non-stationary. Since stationarity is a prerequisite for ARIMA modelling, the series was subjected to differencing. Differencing was applied iteratively, and at the first level of differencing (d=1), the series achieved stationarity. This was evident from the ADF test results conducted on the differenced series, which rejected the null hypothesis of a unit root. Additionally, the absence of a trend and a more consistent variance in the differenced series further supported the stationarity assumption.

Among the proposed models, **ARIMA(2,1,1)** was identified as the most suitable model for the mustard yield data in Bhiwani. This model demonstrated the lowest AIC and BIC values, indicating better goodness-of-fit compared to other competing models. The significance of the model parameters was also tested to ensure the reliability of the selected model.

Table 1: Results of ADF test

	Order(d)	**ADF value**	**p - value**	**Interpretation**
BHIWANI	d=0	-2.96	0.19	Non-Stationary
	d=1	-4.55	0.01	Stationary

Once the series was rendered stationary, the next step involved identifying the appropriate ARIMA model. By analysing the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of the differenced series (d=1) in figure 2, identified significant lag values for the Auto-Regressive (AR) and Moving Average (MA) components. Based on these observations and with the objective of minimizing the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), we proposed several tentative ARIMA models.

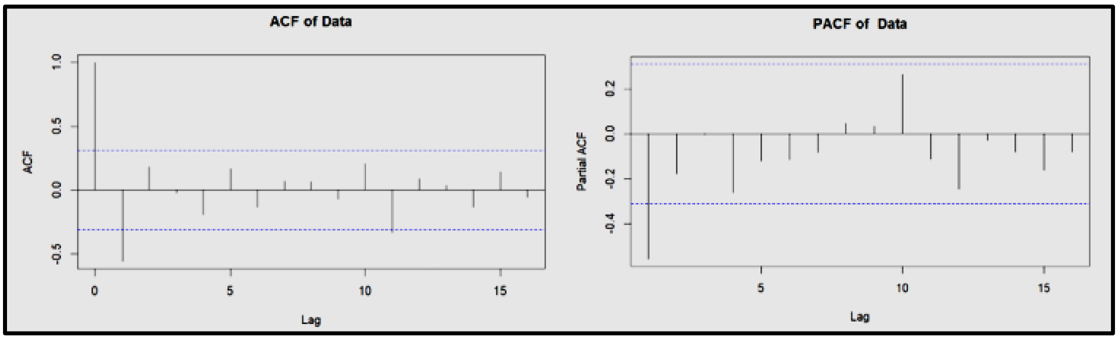

Figure 2 : ACF and PACF plots at d=1

ARIMA(2,1,1) incorporating the exogenous variables with Tmx_3, Tmn_9, Trf_3 is selected as significant model for Bhiwani using stepwise regression method. The parameter estimation for ARIMAX Model is shown in table 2. In table 3, Residuals of the model is tested to ensure they behave like white noise *i.e.* uncorrelated, constant variance and normally distributed on the basis of Ljung-Box test and Shapiro-Wilk test.

Table 2 : Parameter Estimation of ARIMAX Model

District	Parameters	Estimate	St. Error	p-value
Bhiwani ARIMA(2,1,1) with Tmx_3, Tmn_9, Trf_3	AR (1)	0.34*	1.12	0.01
	AR (2)	0.14*	0.41	0.03
	MA (1)	-0.45*	1.26	0.00
	Tmx_3	-9.41*	5.83	<0.01
	Tmn_9	11.49*	3.59	<0.01
	Trf_3	-3.31*	1.24	0.02

Table 3 : Diagnostic Checking for Residual of ARIMAX Model

District	Model	Ljung-Box test		Shapiro-Wilk test	
		Q-value	p-value	W-value	p-value
Bhiwani	ARIMAX (2,1,1) with Tmx_3, Tmn_9, Trf_3	5.17	0.34	0.89	0.25

The p-value is greater than 0.05, indicating that both the Ljung-Box test and the Shapiro-Wilk test are non-significant. This implies that the residuals of the fitted model are uncorrelated and follow a normal distribution. These results confirm that the assumptions of white noise and normality are satisfied, further validating the reliability of the models. Figures 3 represent a comparison between the observed and predicted mustard yields for the districts of Bhiwani using ARIMAX Model. Table 4 represents forecasted yield for 2018-19 to 2022-23.

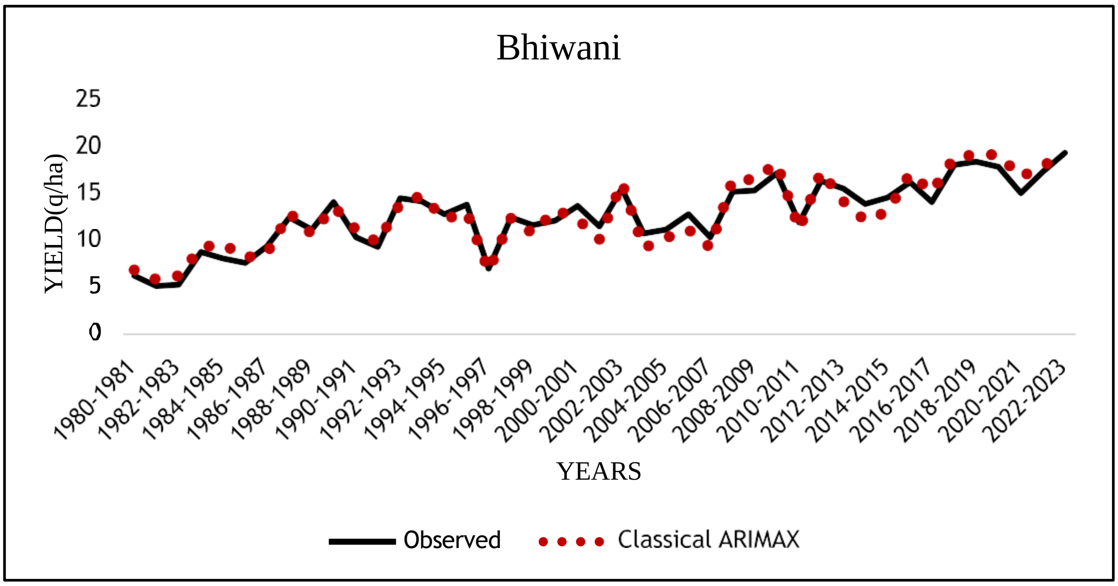

Figure 3: Plot of Observed and predicted yield of Bhiwani using ARIMAX(2,1,1) with Tmx_3, Tmn_9, Trf_3 as exogenous variable.

`

Table 4: Forecasted and Observed Mustard yield using ARIMAX Model Bhiwani

Year	Observed Yield (q/ha)	ARIMAX	
		Forecasted Yield(q/ha)	PRD (%)
2018-19	18.47	19.28	4.29
2019-20	17.959	18.13	0.94
2020-21	15.14	17.01	11.63
2021-22	17.37	18.25	4.94
2022-23	19.43	18.14	6.86

The Percent Relative Deviation (PRD) analysis indicates that the forecasted mustard yield aligns closely with the observed yield, demonstrating minimal deviation between the two. This close agreement highlights the accuracy and reliability of the ARIMAX model for forecasting mustard yield. Further validation of the model's performance was conducted using statistical error metrics, namely the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE). The calculated RMSE and MAE values for the ARIMAX model were 1.3 and 1.27, respectively. These low error values underscore the superior predictive capability of the ARIMAX model, as it effectively captures the underlying dynamics of mustard yield influenced by exogenous weather parameters.

CONCLUSION

The ARIMAX model has proven to be a reliable and effective tool for forecasting mustard yield in the Bhiwani district of Haryana. By integrating weather parameters such as minimum temperature, maximum temperature, and rainfall as exogenous variables, the ARIMAX model enhances predictive accuracy compared to traditional univariate models like ARIMA. The statistical validation of the model, with RMSE and MAE values of 1.3 and 1.27, respectively, underscores its superior performance in capturing the relationship between climatic factors and crop yield. The Percent Relative Deviation analysis further highlights the ARIMAX model's robustness, as it closely aligns forecasted yields with observed values, demonstrating minimal deviations. This integration of weather data enables the

ARIMAX model to account for climatic variability, which is critical in agricultural forecasting. The ARIMAX model offers significant practical implications for the agricultural sector. It can serve as a valuable decision-support tool for policymakers in framing strategies for production, pricing, and resource allocation. Additionally, it provides farmers with reliable pre-harvest yield predictions, enabling better planning for input management and harvesting.

References

1. Ahmer, S. A., Singh, P. K., Ruliana, R., Pandey, A. K., & Gupta, S. (2023). Comparison of ARIMA, Sutte ARIMA, Holt-Winters and NNAR models to predict food grain in India. *Forecasting 5*(1), 138152.
2. Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21,* 243-47.
3. Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis: Forecasting and Control. *Holden Day, San Francisco*. Dharmaraja, S., Jain, V., Anjoy, P., & Chandra, H. (2020). Empirical analysis for crop yield forecasting in India, *Agricultural Research, 9*(1), 132-138.
4. Goyal, M., & Verma, U. (2015). Development of weather-spectral models for pre-harvest wheat yield prediction on agro-climatic zone basis in Haryana (India). *International Journal of Agricultural and Statistical Sciences, 11*(1), 73-79.
5. Goyal, M., Goyal, S. K., Agarwal, S., & Kumar, N. (2021). Forecasting of Indian agricultural export using ARIMA model. *Journal of Community Mobilization and Sustainable Development, 16*(3), 655658.
6. Hamjah, M. A. (2014). Temperature and rainfall effects on spice crops production and forecasting the production in Bangladesh: An application of Box-Jenkins ARIMAX model. *Mathematical Theory and Modeling,* **4**(10): 149- 159.
7. Kour, S., Shitap, M. S., Pradhan, U. K., Paul, R. K., Arya, P. and Kumar A. (2018). Forecasting of rice yield based on weather parameters in Kheda district of Gujarat. *Indian Journal of Agricultural Science,* **14**(2): 611-615
8. Kumar, N. P., Muhammed, J. P. K., & Aniket, C. (2022). Modelling and forecasting of area, production and productivity of tomatoes in Haryana and India.*Indian Journal of Extension Education*, *58*(2), 205208.
9. Kumar, S., Mishra, H. S., Sharma, A. K., & Kumar, S. (2001). Effect of weather variables on the yield of early, timely and late sown wheat in the Tarai region. *Journal Agricultural Physics, 1*(1), 58-62.
10. Ljung, G. M., & Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika, 65*(2), 297- 303.
11. Marquardt, D. W. (1963). An algorithm for least-squares estimation of non-linear parameters, *Journal of Society for Industrial and Applied Mathematics, 11*(2), 431-441.
12. Sanjeev and Verma, U. (2016). ARIMA versus ARIMAX modelling for sugarcane yield prediction. *International Agricultural Statistics of Science in Haryana, 12*(2), 327-334.
13. Schwarz, G. (1978). Estimating the dimension of a model, *The Annals of Statistics, 6*(2), 461-64.
14. Sharma, P. K., Dwivedi, S., Ali, L., & Arora, R. K. (2018). Forecasting maize production in India using ARIMA Model. *Agro Economist - An International Journal, 5*(1), 1-6.
15. Sreekumar, J., & Sivakumar (2022). Forecasting of Sweet Potato (Ipomoea batatas L.) Prices in India. *Indian Journal of Extension Education, 58*(2), 15-20.
16. Yogarajah, B., Elankumaran, C. and Vigneswaran, R. (2013). Application of ARIMAX model for forecasting paddy production in Trincomalee district in Sri Lanka. *Proceedings of third international symposium*, 21-25.