ISSN: 2633-4828 Vol. 6 No.1, January, 2024

International Journal of Applied Engineering & Technology

FORECASTING COTTON GROWTH AREA OF PUNJAB AND HARYANA USING THE BEST FITTED ARIMA MODEL

Pooja Devi, Daljeet Kaur and Madhuchanda Rakshit
Guru Kashi University, Talwandi Sabo, Bhathinda

ABSTRACT

The paper attempts forecasting the Cotton Growth area in Punjab and Haryana using the best fitted Auto-Regressive Integrated Moving Average (ARIMA) model. The time series data on area growth of cotton in Punjab and Haryana for the period of last 10 Years i.e. from 2012-13 to 2021-22 is analyzed for this study. The best models are selected by calculating Normalized BIC; Mean Absolute Percentage Error (MAPE) and maximum values of R_2. The study revealed that ARIMA (1,1,1) and ARIMA (0,1,1) are the best fitted models for forecasting Growth area of cotton in both states. The analysis shows an increasing trend in area of cotton for both the states Punjab and Haryana.

Keywords: BIC, MAPE, R_2, Autoregressive integrated moving average (ARIMA) model, Box and Jenkins, Cotton Growth Area Forecasting.

INTRODUCTION

India is primarily an agricultural country, emerged as a second-fastest growing economy; with the agricultural and peripheral activities contributing significantly to the overall Gross Domestic Product (GDP) with a share of 19.9 percent (India Economic Survey, 2021). Agriculture contributes a major source of alimentation providing direct and passive employment to about 43.21 percent of the total workforce (International Labour Organization, 2019) and constitutes 11 percent of total exports. India alone is home to the second-largest livestock population globally following Brazil; it breeds 15 percent of total world livestock in a comparably small area which also comes under agrarian related activities. It is evident from the integration and dependence on agriculture, the vitality and critical role of assessing and ascertaining the performance and production analysis of an economic parameter as huge as agriculture in India. The nation produces the largest lot of conventional Cotton next to China comprising. India also is a global leader in organic cotton production catering to 70% of total global demands (Mohapatra & Saha, 2019). It is therefore important for country like India, which lacks proper crop-grain storage, maintenance, effective transportation, logistics, to have real-time data of production and distribution (Rajendran, 2003) with an effective boost in infrastructure that has a direct implication on the production of Cotton (Chinnadurai, Sangeetha, Anbarassan & Kavitha, 2019).

Indian textile manufacturing is the second-largest pool of employment generators followed by agriculture; it has a national share of 26 percent of total manufacturing output. Cotton is the main raw material required in textile manufacturing, making its sustainable production an important requirement for the prolongment of a healthy economic profile.

Cotton in a popular term is referred to as "White Gold and King of Fibers" for its importance and worldwide use as a commercial fiber crop with prime economic value. The yarn obtained is spun from the cotton fiber of the plant, further woven or knitted into the desirable kind of fabric. The main reason for its prolonged association with humans is its desirable characteristics; it adds moisture-absorbing characteristics to the fabric, with good drape adaptability and shell life. It's lightweight and comfortable feature keep it on the top of consumer preference, making it among the most demanded textile raw material. The processed product obtained (yarn) from the cotton plant is composed of a natural polymer, making it 100 percent biodegradable. The natural polymer can degrade within four weeks under both aerobic and anaerobic conditions, making its greenhouse gas footprint neutral. The study of Debnath *et al*. (2015) revealed that area, production and yield of cotton in India would increase from 2016-17 to 2020- 21. Similar studies have been conducted by Payyamozhil and Kachi (2017) and Rajan *et al*. (2018).

Information on crop yield beforehand is a crucial aspect in designing the stock distribution framework and entailing commercial activities. Forecasting techniques in agriculture is applied mainly in analyzing the harvest/yield and productivity of the crop. Agronomy forecast has evolved encompassing inter-alia factors like pest concentration, rainfall variability, water availability, diseases, and natural calamity. The forecasting technique involves various disciplines of mathematics and statistics having webbed into different disciplines. The model is appropriated according to the variables and closeness to the real values once the comparison between the predicted value and real values are matched. Studies by Rachana *et al.* (2010) for forecasting pigeon pea production in India by using ARIMA Modeling and Rahman (2010) for forecasting of boro rice production in Bangladesh. Iqbal *et al.* (2005) also use the ARIMA Model for forecasting wheat area and production in Pakistan.

METHODOLOGY

The time series data on growth area of cotton in Punjab and Haryana for the period of last 10 Years i.e. from 2012-13 to 2021-22 is analyzed for this study. Data is collected from Annual Reports issued by Ministry of Textile, Government of India and Central Institute of Cotton Research, Ministry of Agriculture and farm Welfare, Government of India.

ARIMA Model

One of the most important and widely used time series models is the Auto Regressive Integrated Moving Average (ARIMA) Model. The popularity of the model is due to its statistical properties as well as the well known Box-Jenkins methodology in the model building process. Box-Jenkins Model is based on the methodology set by George Box and Gwilym Jenkins, which is based on the early works of Yule (1929) who established the inference of stochasticity in historical data (time-series), the breakthrough was the universal assumption of time series as a stochastic process. It's a combination of autoregressive (AR) and moving average (MA), along with the difference value. It is expressed in the form of order (p,d,q). In an autoregressive integrated moving average (ARIMA) model, the future value of a variable is assumed to be a linear function of several past observations and random errors.

ARIMA is composed of the following components:

AR(p) : p implying the order of the autoregressive part

I(d) : d implying a degree of first differencing

MA(q) : q implying order of the moving average part.

- Autoregressive Model (AR) with pth order has the general form:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots . \varphi_p Y_{t-p} + \epsilon_t$$

……………..1

Where Y_t is the response variable or independent variable at the time $'t'$,

Y_{t-1}, Y_{t-2}, Y_{t-p} are the response variable at lag t - 1, t - 2.., t – p respectively,

$\varphi_0, \varphi_1, \varphi_2, \varphi_p$ are the co-efficient to be estimated,

ϵ_t is the error term at time t

- Moving Average model (MA) with qth order has the general form:

$$Y_t = \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots \theta_q \epsilon_{t-q}$$

……………..2

Where Y_t is the response variable at time t,

μ is the constant mean of the process,

θ_1, θ_2 ... θ_q is the coefficient to be estimated.

ϵ_t is the error term at time t

$\epsilon_{t-1}, \epsilon_{t-2}, \epsilon_{t-q}$ are the error in the previous time.

The value is d is the degree of the difference applied in order to make the time series data stationary.

- Differencing (value of d)

The degree of difference required to transform the non-stationarity of the time series is taken as 'd' which finally used in the ARIMA model building, the value of d is usually 0-2 depending upon the level of differencing needed to ensure constant mean throughout the time series data. The combination of these models AR (p) and MA (q) and differencing (d) is incorporated in the given form:

$$Y_t = \varphi_0 + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \cdots . \varphi_p Y_{t-p} + \mu + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} - \cdots \theta_q \epsilon_{t-q} + \epsilon_t$$

……………..3

The objective of this paper is to develop a Box Jenkins methodology of ARIMA model and forecast the Growth area of cotton in the states Punjab and Haryana and a comparative study is produced.

Box-Ljung Statistic

Ljung – Box Statistics statistic tests whether a group of autocorrelations of a time series are less than zero (G. Ljung, and G. Box, 1978). The test statistic is given as:

$$Q = T(T+2) \sum_{k}^{s} 1 \frac{{r_k}^2}{T-K}$$

……………..4

T: number of observations

s: length of coefficients to test autocorrelation

r_k: Autocorrelation coefficient (for lag k)

The hypothesis of Ljung - Box test are:

H_0: Residual is white noise

H_1: Residual is not white noise

If the sample value of *Q* exceeds the critical value of aχ2 distribution with degrees of freedom, then at least one value of is statistically different from zero at the specified significance level.

Normalized BIC

In statistics, the Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models. It is based, in part, on the likelihood function, and it is closely related to Akaike information criterion (AIC). When fitting models, it is possible to increase the like hood by adding parameters, but doing so may result in over fitting. The BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The penalty term is large in BIC than in AIC. The BIC was developed by Gideon E. Schwarz (G. E.

Schwarz, 1978), who gave a Bayesian argument for adopting it. It is closely related to the Akaike information criterion (AIC). In fact, Akaike was so impressed with Schwarz's Bayesian formalism that he developed his own Bayesian formalism, now often referred to as the ABIC for "a Bayesian Information Criterion" or more casually "Akaike's Bayesian information criterion" (Akaike, H., 1977) .

Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE), also known as mean absolute percentage deviation (MAPD), is a measure of accuracy of a method for constructing fitted time series values in statistics, specifically in trend estimation. It usually expresses accuracy as a percentage, and is defined by the formula:

$$M = \frac{100}{n}\sum_{i=1}^{n}\left[\left|\frac{A_t - f_d}{A_t}\right|\right]$$

……………..5

Where, At is the actual value and $Ft$is the forecast value. The difference between At and is divided by the actual value At again. The absolute value in this calculation is summed for every fitted or forecasted point in time and divided again by the number of fitted points n. multiplying by 100 makes it a percentage error.

Forecasting accuracy checking

Among the best fitted ARIMA and exponential smoothing technique a best model is used for forecasting based on the accuracy of the testing. The accuracy is checked using two measures namely RMSE and MAPE. A major part of the data used for model fitting is called as training set and a smaller portion (usually 10%) of data used for checking forecasting accuracy is called as testing set.

Maximum values of R-Squared

R-Squared is the percentage of the dependent variable variation that a linear model explains. The maximum value would be 1 but minimum value can be below 0.

RESULTS AND DISCUSSION

ARIMA Model for Cotton Area Growth for Punjab and Haryana

The time-series data for forecasting of Punjab cotton growth area exhibited non- stationary behavior, after its first difference the data indicated stationary. Observing the pattern and similarities between the ACF and PACF, it can be concluded that the model is ARMA (p, q) model. The value of the difference is taken as 1, as the order of differencing (d) is 1 for both Punjab and Haryana.

Tab 1 Autocorrelation and PACF value of Cotton Growth Area Time Series Data

Autocorrelation and Partial Autocorrelation														
	Series: Punjab Area for Cotton Growth							Series: Haryana Area for Cotton						
			Box-Ljung Statistic			PACF	SE			Box-Ljung Statistic			PACF	SE
Lag	ACF	SE[a]	Value	df	Sig.[b]			ACF	SE[a]	Value	df	Sig.[b]		
1	.712	.274	6.761	1	.009	.712	.316	.447	.274	2.663	1	.103	.447	.316
2	.362	.258	8.728	2	.013	-.294	.316	.140	.258	2.956	2	.228	-.075	.316
3	.013	.242	8.731	3	.033	-.248	.316	.101	.242	3.130	3	.372	.084	.316
4	-.198	.224	9.513	4	.049	-.020	.316	-.241	.224	4.294	4	.368	-.394	.316
5	-.308	.204	11.794	5	.038	-.102	.316	-.269	.204	6.026	5	.304	.016	.316
6	-	.183	16.761	6	.010	-.272	.316	-	.183	8.157	6	.227	-.216	.316

	.407							.267					
7	- .373	.158	22.329	7	.002	.060	.316	- .329	.158	12.489	7	.086	-.081	.316
8	- .235	.129	25.633	8	.001	.084	.316	- .081	.129	12.886	8	.116	.086	.316

a. The underlying process assumed is independence (white noise).

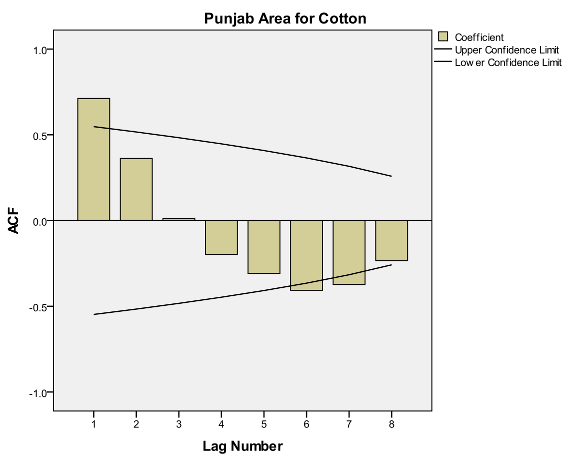
b. Based on the asymptotic chi-square approximation.

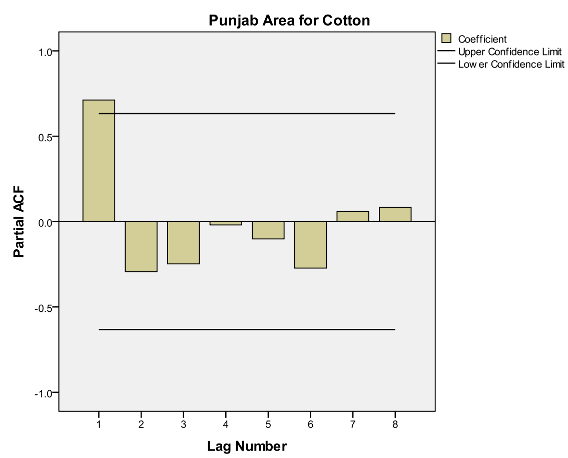
ACF - Autocorrelation Function **PACF-** Partial Autocorrelation Function **df-** Degrees of Freedom **SE-** Std. Error

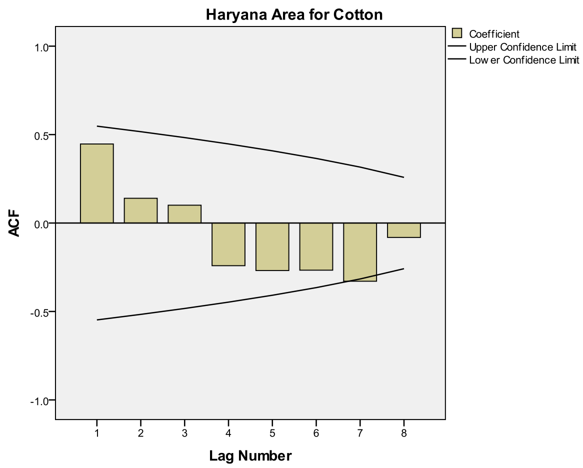
The ACF pattern (Tab 1) shows properties of fair exponential decay (i.e. quantity proportionally changes over time) and damped sine wave pattern (i.e. smooth periodic oscillation) that makes it statistically significant. The consequent step is to find the order of Autoregressive "AR" for the value of "p" and the order of Moving Average "MA" for the value of "q", in order to find these values, examination of Correlogram containing Partial autocorrelation (PACF) and Autocorrelation (ACF) were conducted on the sampled time series data. Accordingly, Correlogram obtained through the stationarity revealed that autocorrelation function falls after the first lag, which gives the value of Auto Regression (AR), p as 1 and similarly from the PACF pattern, the value of Moving Average (MA) was selected as 1. The model that deemed fit for the forecasting process was ARIMA (1,1,1) for Punjab and ARIMA (0,1,1) for Haryana. The model is further investigated and verified for its goodness of fit quality by calculating the Forecasting Criterion.

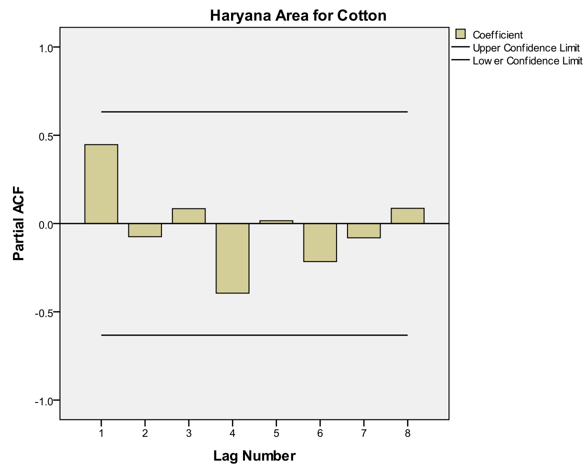
The PACF pattern (Tab 1) properties shows that for model (1,1,1) PACF values lies between -0.712 to -0.294 and for model (0,1,1) is between 0.447 to -0.394 which is verified for its goodness of fit quality.

Fig 1 ACF and PACF Graph for Cotton Growth Area Time Series Data









Forecasting Criterion for Cotton Growth Area

The calculated value of the Forecasting criterion is tabulated below:

Tab 2: Forecasting Criterion Value

AREA	R-squared	RMSE	MAPE	MAE	Normalized BIC
Punjab	0.783	0.385	7.058	0.221	-1.179
Haryana	0.241	0.640	7.875	0.486	-0.404
MAPE- Minimum Absolute Percentage Error **Normalized BIC**- Bayesian Information Criterion **RMSE**- Root Mean Square Error **MAE**- Minimum Absolute Error $\mathbf{R_2}$ – Coefficient of Determination					

The most important indicators of a good forecasting model are its lower Normalized BIC (Bayesian Information Criterion). The BIC value for the model (1,1,1) is -1.179 and for the model (0,1,1) is -0.404.

Tab 3: Model Fit Statistics Value

Model Statistics							
Model	**Number of Predictors**	**Model Fit statistics**		**Ljung-Box Q(18)**			**Number of Outliers**
		Stationary R-squared	**R-squared**	**Statistics**	**DF**	**Sig.**	
Punjab Area for Cotton-Model_1	0	0.232	0.783	.	0	.	0
Haryana Area for Cotton-Model_1	0	0.316	0.241	.	0	.	0

The R-squared value for model (1,1,1) is 0.783 and for model (0,1,1) is 0.241, while Stationary R-squared value for model (1,1,1) is 0.232 and for model (0,1,1) is 0.316 which is a good value for a model. Degree of Freedom for both the model is 0.

Future Forecasting Growth Area for Punjab and Haryana:

Tab 4: Cotton Growth Area Forecasting by using the best fitted model

Year	Punjab Cotton Growth Area Forecasting (2022-23 to 2031-32) in Lakh Hact		Haryana Cotton Growth Area Forecasting (2022-23 to 2031-32) in Lakh Hact.	
	LCL	UCL	LCL	UCL
2022-23	3.32	4.22	6.1	8.8

2023-24	3.09	4.9	6.27	8.98
2024-25	2.94	5.27	6.45	9.16
2025-26	2.77	5.54	6.63	9.33
2026-27	2.61	5.75	6.81	9.51
2027-28	2.44	5.93	6.99	9.69
2028-29	2.28	6.07	7.17	9.86
2029-30	2.12	6.19	7.35	10.04
2030-31	1.95	6.3	7.53	10.22
2031-32	1.79	6.39	7.71	10.39
UCL- Upper Confidence Limit **LCL-** Lower Confidence Limit				

In the study, ARIMA (1,1,1) and ARIMA (0,1,1) models were developed for forecasting area of Cotton Growth for Punjab and Haryana. From the forecasts available by using the best fitted models it can be find out that the cotton area will increase in the next ten years. Model (1,1,1) states that UCL will increases for next Ten Years while LCL goes decline. But for Model (0,1,1) both LCL and UCL goes in upward direction.

REFERENCES

1. Debnath, M.K., Kartic Bera and Mishra, P. Forecasting Area, Production and Yield of Cotton in India Using ARIMA Model, Research and Reviews: Journal of Space & Technology, 2(1): 16-20 (2013).
2. Government of India, (2022). Economic Survey of India. Retrieved from https://www.ibef.org/download/Key_Highlights_of_Economic_Survey_2021-22.pdf .
3. Rajan SM, Palanivel M. Application of regression models for area, production and productivity growth trends of cotton crop in India. Int. J Stat. Distributions & Applications. 2018;4(1):1-5.
4. Rachana W, Suvarna M, Sonal G. Use of ARIMA Model for Forecasting Pigeon Pea Production in India, International Review of Business and Finance. 2010;2(1):97-102.
5. Rahman NMF. Forecasting of boro rice production in Bangladesh: An ARIMA Approach. J Bangladesh Agril. Univ. 2010;8(1):103-112.
6. Iqbal N, Bakhsh K, Maqbool A, Ahamad AS. Use of the ARIMA model for forecasting wheat area and production in Pakistan. J Agric. Soc. Sci. 2005;1(2):120-12.
7. Prabakaran, K. and Sivapragasam, C. Forecasting Areas and Production of Rice in India Using ARIMA Model, International Journal of Farm Sciences, 4(1): 99-106 (2014).
8. Sudar Rajan and Palanivel. Time Series Model to Forecast Production of Cotton in India: An Application of ARIMA Model. AE International Journal of Multi Disciplinary Research, 2(1): 1-9 (2014).
9. Borkar, Prema and Tayade, P.M. Forecasting of Cotton Production in India Using ARIMA Model, International Journal of Research in Economics and Social Sciences, 6(5): 1-7 (2016).
10. Payyamozhil, S., Kachi Mohideen, A. Forecasting of Cotton Production in India Using ARIMA Model. Asia Pacific Journal of Research- A peer reviewed International Journal, XLVIII (I): 70-74 (2017).
11. Wali, V.B., Beeraladinni, D. and Lokesh, H., Forecasting of Area and Production of Cotton in India: An Application of ARIMA Model, Int. J. Pure App. Biosci. 5(5): 341-347 (2017). DOI: http://dx.doi.org/10.18782/2320-7051.5409.
12. Mohapatra, L., & Saha, G. (2019). Cotton Farming in India: Alternative Perspectives and Paradigms. In Transition Strategies for Sustainable Community Systems (pp. 195-213). Springer, Cham. markets in India. International Journal of Agricultural Sciences, 15(1), 141-147.

13. Rajendran, S. (2003). Grain storage: Perspectives and problems. Handbook of postharvest technology: cereals, fruits, vegetables, tea, and spices. Marcel Dekker, USA, 183-192.

14. Chinnadurai, M., Sangeetha, R., Anbarassan, A., & Kavitha, B. (2019). Price integration analysis of major cotton domestic markets in India. International Journal of Agricultural Science, 15(1), 141-147.

15. Yule, G. U. (1929). On Introduction to the Theory of Statistics. London: Chas. Griffin and Co., Ltd, 424.

16. Ljung, G., and Box, G. E. P. (1978): On a Measure of lack of fit in Time Series Models. Biometrika 65:553-564

17. Schwarz, G. E. (1978): Estimating the dimension of a Model. Annals of Statistics. 6(2): 461-464.