Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23–27 February 2022; pp 115-127

Time Series Analysis of Major Cotton Production States in India using Box–Jenkins Approach

Prema Borkar
Gokhale Institute of Politics and Economics（Deemed University, Pune, India

Received: 19 July 2022; Revised: 25 July 2022; Accepted: 28 July 2022

Abstract

Cotton is one of the most important fibers and cash crops of India and plays a dominant role in the industrial and agricultural economy of the country. It provides the basic raw material (cotton fiber) to the cotton textile industry. In India, there are ten major cotton – growing states which are divided into three zones, viz., north zone, central zone, and south zone. The north zone consists of Punjab, Haryana, and Rajasthan. The central zone includes Madhya Pradesh, Maharashtra, and Gujarat. The South zone comprises Andhra Pradesh, Telangana, Karnataka, and Tamil Nadu.

In this study, the data on cotton production in major cotton-producing states in India were collected from the website of Cotton Corporation of India for the period from 1964-65 to 2021-22 and were used to fit the ARIMA model and to predict future production. The Box Jenkins (1970) ARIMA methodology has been used for forecasting. ARIMA forecasting model is the most popular and widely used forecasting model for time series data. Autocorrelations and partial autocorrelation functions were calculated for the data. Model parameters were estimated using R programming software. The performance of the fitted model was examined by computing various measures of goodness of fit *viz.*, AIC, BIC, and MAPE. Empirical results showed that ARIMA (0,1,0) model was most suitable to forecast the future production of cotton in India. Similarly, the ARIMA model was fitted separately for major cotton-producing states in India. The forecasts from 2022-23 to 2029-30 are calculated based on the selected model. Overall cotton production is expected to be 362.18 million tons by 2029-30. The forecasting power of the autoregressive integrated moving average model was used to forecast cotton production for eight leading years. The results of major cotton-growing states are presented numerically and graphically.

Key words: ACF **-** autocorrelation function; ARIMA **-** autoregressive integrated moving average; Cotton production; PACF **-** partial autocorrelation function; Residual analysis.

1. Introduction

Cotton is often referred to as the "birth place of India" and the industry plays an important role in the country's economy. It is important in both developed and developing countries as a cash crop for millions of farmers, including small and marginal farmers, and as a strategic raw material for the textile industry. Although cotton is farmed in almost 100 nations, only six countries—China, India, the United States, Brazil, Pakistan, and Uzbekistan contribute to nearly 80% of global production (19.84 million tons) (FAO, 2022). With a yearly production

Corresponding Author: Prema Borkar
Email: premaborkar@rediffmail.com

volume of 6.16 million tonnes, India is the world's second largest cotton producer (FAO, 2022). However, for a variety of well-known reasons, there are variances between nations in terms of the fundamental crop/commodity performance metrics, such as area, production, productivity, trade, etc. Cotton production has increased significantly since the introduction of genetically modified crops, especially in the US, China, India, Australia, Argentina, and South Africa.

The cotton sector in India directly supports approximately 5 million farmers spread across 10 states, and it plays an important role in the domestic economy as a strategic industrial raw material for the textile industry. India ranks first in world cotton area cultivation which is about 37 % of the world area under cotton cultivation between 12.0 million hectares to 13.5 million hectares (www.cotcorp.org.in). It is the second largest producer of cotton in the world accounting for about 22% of the world cotton production. Despite the fact that India has the most cotton land, its productivity is among the lowest in the world. Among the main factors cited for India's low cotton productivity are the prevalence of small and marginal holdings, insufficient transfer of production technologies, and insufficient financial resources. The yield per kgs hectare which is presently 469 kgs/ha is still lower against the world average yield of about 787 kgs /ha (www.cotcorp.org.in).

Millions of farmers and those working in industries related to cotton, such as transportation and processing, are employed by the crop. In terms of acres planted to cotton and cotton production, India leads the globe. Currently, ten major cotton-growing states account for the majority of the nation's cotton production. These states can be divided into three regions: The Northern Zone, which includes Punjab, Haryana, and Rajasthan; the Central Zone, which includes Gujarat, Maharashtra and Madhya Pradesh; and the Southern Zone, which includes Andhra Pradesh, Telangana, Karnataka, and Tamil Nadu.

Cotton cultivation, marketing, processing, and exports provide a living for nearly 60 million people today (www.ibef.org). India is also the only country in the world that commercially grows not only the four cultivated cotton species, but also their intra- and inter-specific hybrids. The textile industry, which uses cotton as its primary raw material, contributes about 4% of GDP and is the country's largest foreign exchange earner. As a result, the growth and development of the cotton and cotton-based textile industries is critical to the overall development of the Indian economy. Thus, it becomes important to study the cotton production in major states of India and to forecast cotton production in India. The main objective of this study is to develop an ARIMA model for forecasting the cotton production in major states of India using Box-Jenkins approach.

2. Material and Methods

In this study, the data on cotton production in major cotton−producing states in India were collected from the website of Cotton Corporation of India for the period from 1964-65 to 2021-22 and were used to fit the ARIMA model and to predict future production using Box-Jenkins approach.

2.1. Autoregressive integrated moving average

ARIMA stands for auto-regressive integrated moving average and is defined by three order parameters: (*p, d, q*). The Box-Jenkins technique is another name for the procedure of fitting an ARIMA model. ARIMA model is a technique for prediction the future values or events of the variable. This method is suitable for any time series with any pattern of change. It requires a long time series data for analysis (Biswas *et al*., 2014). When past values are used in the regression equation for the series Y, this is referred to as an auto regressive AR(*p*) component. The auto-regressive parameter *p* provides the model's number of lags. The general model AR(*p*) is represented as:

$$\boldsymbol{Y_t = \mu + \sum_{i=1}^{p} \phi_i Y_{t-i} - i + e_t}$$

where, *p* is the order of the AR model and ϕ_1, ϕ_2, …., ϕ_p are *p* partial autocorrelation parameters for the AR(*p*) model. The AR(*p*) model contains only *p* statistically significant partial autocorrelations. The AR(*p*) model's autocorrelation coefficient approaches zero and is constrained between –1 and 1. The degree of differentiation in the integrated I(*d*) component is represented by *d*. To differentiate a series, just subtract its current and prior values *d* times. When the stationarity assumption is violated, differencing is frequently utilized to stabilize the series. Hossain *et. al*. (2006) forecasted three different varieties of pulse price in Bangladesh using ARIMA model. Mandal (2005) forecasted sugarcane production in India. Assis *et al*. (2010) forecasted cocoa bean prices in Malaysia along with other competing models. Cooray (2006) forecasted Sri Lanka's monthly total production of tea and paddy monthly data from January 1988 to September 2004.

2.2. Moving average

The error of the model is represented as a linear combination of past error terms by a moving average MA(*q*) component. The number of terms to include in the model is determined by the order.

$$\boldsymbol{Y_t = \mu + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t}$$

The moving average is a linear combination of past forecast errors. The general MA(q) model is represented as:

$$\boldsymbol{Y_t = \mu + \sum_{i=1}^{q} \Theta_t \, e_{t-i} + e_t}$$

where, *q* is the order of the model and Θ_1, $\Theta_{2, \ldots\ldots, \Theta_q}$ are parameters of the model. The e_t, e_{t-1}, ……, e_{t-q} are the white noise error terms. This can be equivalently written in terms of the backshift operator B as:

$$\boldsymbol{Y_t = \mu + \left(\sum_{i=1}^{q} \Theta_i B^i + 1 \right) e_t}$$

The capability of inverting an MA model to obtain an AR model with infinite order is made possible by the invertibility condition of an MA(*q*) process. A non-seasonal ARIMA model is made up of varying, autoregressive, and moving average components and can be expressed as a linear equation

$$\boldsymbol{Y_t = \mu + \sum_{i=1}^{p} \phi_i Y_{di} + \sum_{i=1}^{q} \Theta_t \, e_{t-i} + e_t}$$

where, y_d is *Y* differenced *d* times and μ is a constant mean.

The Box-Jenkins methodology for estimating a time series model consists of four iterative steps: model identification, estimation of model parameters, diagnostic checking, and forecasting. The tentative model parameters are identified first using ACF and PACF, and then the coefficients of the most likely model are determined. The next steps involve forecasting, validating, and checking the model performance by observing the residuals using the Ljung Box test and ACF plot of residuals.

3. Model Identification

Theoretically, ARIMA models are the most general class of models for forecasting a time series that may be made "stationary" by differencing (if necessary), sometimes in conjunction with nonlinear transformations such as logging or deflating (if necessary). A stationary random variable is one whose statistical features remain consistent across time. A stationary series has no trend, constant amplitude variations around its mean, and wiggles in a consistent manner, i.e., its short-term random temporal patterns always look the same statistically. The ARIMA forecasting equation for a stationary time series is a linear (regression-type) equation in which the predictors are dependent variable lags and/or forecast error lags.

To find the best ARIMA model for *Y*, first determine the order of differencing (*d*) required to stationeries the series and eliminate the gross seasonal features, maybe in conjunction with a variance stabilizing operation like logging or deflating. If you stop here and forecast that the differenced series is constant, you've just fitted a random walk or random trend model. However, the stationeries series may still have auto correlated errors, implying that several AR terms (*p* 1) and/or few of MA terms (*q* 1) are required in the forecasting equation. Statistical tests are used to determine if a time series is stationary. To determine if the time series were stationary, Augmented Dickey Fuller (ADF) were used in this study.

3.1. Estimating the parameters

After tentatively identifying the suitable model, next step is to obtain least square estimates of the parameters, such as R^2, Root mean square error (RMSE), Mean absolute percentage error (MAPE), Mean absolute error (MAE) and normalized Bayesian Information Criterion (BIC) to check the accuracy of the model. In this study, three different parameters are considered for the evaluation of the forecasting models *i.e.*, MAPE, AIC and BIC.

The Mean Absolute Percent Error (MAPE)

The mean absolute percent error was used as a measure of accuracy of the models. It is also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses the accuracy as a ratio defined by the formula

$$\mathrm{MAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

where, A_t is the actual value and F_t is the forecast value. Their difference is divided by the actual value A_t.

Low Akaike information criteria (AIC)

AIC is estimated by AIC = $(-2\log L + 2m)$,

where, $m = p + q$ and L is the likelihood function.

The Akaike information criterion (AIC) is an estimator of prediction error and thereby relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

Low Bayesian Information Criteria (BIC)

Bayesian Information Criteria (BIC) is also used and estimated by

$$\text{BIC} = \log \sigma 2 + (m\log n)/n.$$

The Bayesian information criterion (BIC) is a criterion for model selection among a finite set of models; models with lower BIC are generally preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

3.2. Diagnostic checking

After estimating the parameters of a tentatively recognized ARIMA model, diagnostic testing is required to ensure that the model is adequate. Examining the ACF and PACF of residuals may reveal the model's adequacy or insufficiency. If it has random residuals, it means that the model that was tentatively selected is adequate. When an inadequacy is found, the checks should indicate how the model should be updated, followed by more fitting and checking. When all of their ACF were under the limitations, the residuals of ACF and PACF were considered random (Burark and Sharma, 2012).

3.3. Forecasting

Future values of the time series are forecasted. R programming software was used for time series analysis and developing ARIMA models.

4. Results and Discussion

The time series data of production in major cotton growing states were subjected to a stationary check, which demonstrated the non-stationarity of cotton production except Punjab which was found to be stationary. Nonstationary time series data were made stationary using first order differencing and best fit ARIMA models were developed and used to forecast production of cotton in major cotton growing states in India during 2022-23 to 2029-30. The initial values for the orders of the non-seasonal parameters "p" and "q" were used to identify ARIMA models. They were discovered by plotting autocorrelation and partial autocorrelation functions for significant spikes. During the identification stage, one or more models that appear to provide statistically adequate representations of the available data were tentatively chosen.

Initially, all cotton growing state production ACF and PACF plots were plotted and then it was examined. In the plots, the continuous line above and below the x-axis represents the confidence limits. It was observed that the spikes were above the confidence limits. It means

the data was not stationary. No spikes exceeded confidence limits after differencing. This indicated that the series had reached its stationary point. The plot of ACF and PACF with differencing is shown in Figure 1 except Punjab.

The order of *p* and *q* were determined based on the ACF and PACF plots for developing the preliminary ARIMA model for major cotton producing states in India. Ten tentative ARIMA models were chosen with different *p, d*, and *q* values that were within a reasonable range. The model's parameters were then precisely estimated using least squares. After fitting the model, accuracy of the model was tested based on diagnostics statistics *i.e*., MAPE, AIC and BIC. The model which had lowest value of these parameters was selected for validation. The fitted ARIMA model is presented in Table 1.

Table 1: Model fit statistics of the fitted ARIMA model

State	Best Fitted ARIMA Model	MAPE	AIC	BIC
Maharashtra	(2,1,2)	8.37	390.36	402.62
Andhra Pradesh and Telangana	(0,1,1)	7.94	397.28	403.41
Gujarat	(0,1,0)	8.53	434.10	436.14
Rajasthan	(0,1,0)	8.34	271.65	273.69
Karnataka	(0,1,2)	9.54	306.04	312.17
Haryana	(1,1,1)	7.32	281.63	287.76
Madhya Pradesh	(2,1,2)	7.34	276.64	286.86
Punjab	(1,0,0)	8.36	309.26	35.44
Tamil Nadu	(2,1,1)	7.45	168.74	176.91
India	(0,1,0)	8.41	509.36	511.40

The auto-correlation function (ACF) and partial auto-correlation function (PACF) of residuals were further examined to see if the selected models contained any systemic pattern that could be removed to improve predictability. The ACF and PACF of these models' residuals for major cotton producing states were plotted. This figure shows that the ACF and PACF of residuals are within the confidence interval and are not significantly different from zero. This indicated that the models were chosen correctly.

To see if the forecast errors are normally distributed with mean zero, plot a histogram of the forecast errors with an overlaid normal curve with mean zero and the same standard deviation as the forecast error distribution. Figure 2 histogram plots demonstrates the histogram of forecast errors of residuals of major cotton production states in India.

The forecast error time plots of major cotton producing states demonstrates that the variation of the forecast errors is almost consistent across time. The forecast error histogram indicates that it is likely that the errors are normally distributed, with mean zero and variance constant. Therefore, it is plausible that the forecast errors are normally distributed with mean

zero and constant variance. Since successive forecast errors do not appear to be connected and the forecast errors appear to be normally distributed with mean zero and constant variance, the ARIMA model (2,1,2), (0,1,1), (0,1,0), (0,1,0), (0,1,2), (1,1,1), (2,1,2), (1,0,0), (2,1,1) and (0,1,0) appears to be appropriate for predicting Maharashtra, Andhra Pradesh and Telangana, Gujarat, Rajasthan, Karnataka, Haryana, Madhya Pradesh, Punjab, Tamil Nadu and India.

Figure 3 shows the actual and forecasted plots of major cotton production states in India. Eight years ahead forecast was done for major cotton production states in India using the fitted ARIMA models *i.e*., ARIMA (2,1,2), (0,1,1), (0,1,0), (0,1,0), (0,1,2), (1,1,1), (2,1,2), (1,0,0), (2,1,2) and (0,1,0) at the 95 per cent confidence interval.

From Figure 3, it is observed that Maharashtra, Karnataka, Haryana and Tamil Nadu is showing a slightly increasing trend. The forecasted values for cotton production in Maharashtra, Karnataka, Haryana and Tamil Nadu in 2029-30 was found to be 95.49, 18.62, 20.88 and 4.31 million tonnes, respectively. While the production of cotton in Andhra Pradesh, Gujarat and Rajasthan will remain constant throughout the study period. The production of cotton in Madhya Pradesh is showing a declining trend. The overall cotton production in India will remain constant throughout the study period i.e., 362.18 million tonnes. The fitted models accurately forecast 91.63 percent for Maharashtra, 92.06 percent for Andhra Pradesh, 91.47 percent for Gujarat, 91.66 percent for Rajasthan, 90.46 percent for Karnataka, 92.68 percent for Haryana, 92.66 percent for Madhya Pradesh, 91.64 percent for Punjab and 92.55 percent for Tamil Nadu, according to the mean absolute percentage error (MAPE).

5. Conclusion

The Box-Jenkins method as an ARIMA model was used to forecast future values based on the historical movement patterns of a variable. In this paper, a model for forecasting of cotton production in major states of India was developed. The forecasted cotton production in Maharashtra, Karnataka, Haryana and Tamil Nadu exhibited a slightly increasing trend and Madhya Pradesh exhibited a declining trend. Whereas, Andhra Pradesh, Gujarat, Rajasthan exhibited no trend *i.e.*, it will remain constant throughout the study period. Based on the forecasting and validation results, it is possible to conclude that the ARIMA model might be used successfully forecast cotton production in major states of India in the coming years. The current study's findings provided direct support for the potential use of accurate forecasts in decision-making, assist the government in formulating policies, production, import, and/or export and cotton production management in India.

References

Assis, K., Amran, A., Remali, Y. and Affendy, H. (2010). A comparison of univariate time series methods for forecasting cocoa bean prices. *Trends in Agricultural Economics*, **3**, 207–215.

Biswas, B., Dhaliwal, L. K., Singh, S. P and Sandhu, S. K. (2014). Forecasting wheat production using ARIMA model in Punjab. *International Journal of Agricultural Sciences*, **10(1)**, 58 -161

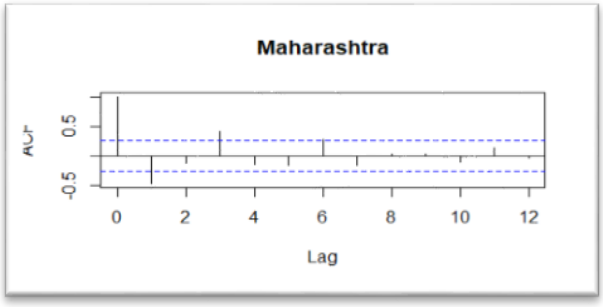
Box, G. E. P. and Jenkins, J. M. (1970). *Time Series Analysis – Forecasting and Control*. Holden-Day Inc., San Francisco.

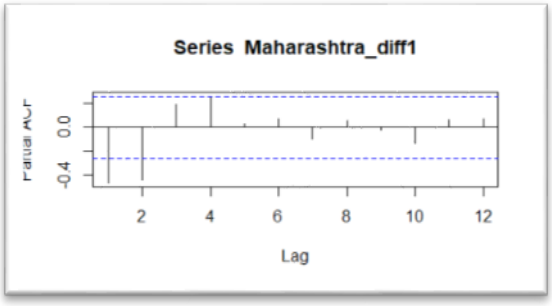
Cooray, T. M. (2006). Statistical analysis and forecasting of main agriculture output of Sri Lanka: rule-based approach. Appeared in 10th International Symposium, **221**: 1–9. Sabaragamuwa University of Sri Lanka

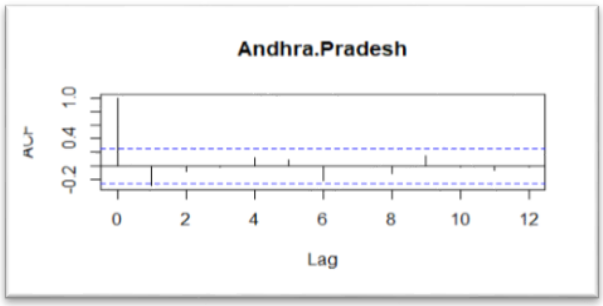
Hossain, M. Z., Samad, Q. A. and Ali, M. Z. (2006). ARIMA model and forecasting with three types of pulse prices in Bangladesh: A case study. *International Journal of Social Economics*, **33**, 344–353.

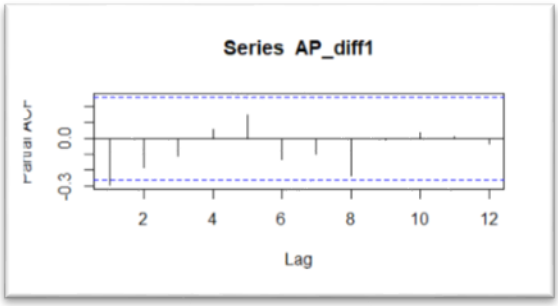
Mandal, B. N. (2005). Forecasting Sugarcane Productions in India with ARIMA Model. *Inter Stat*, October, 2005.

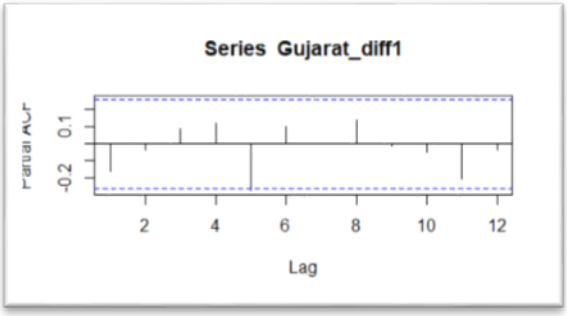
Appendix

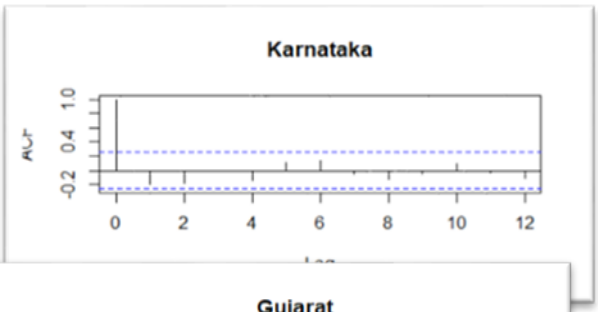


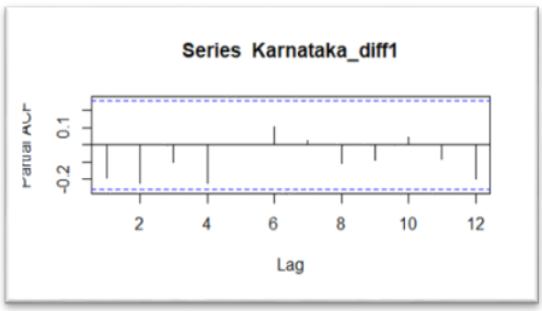


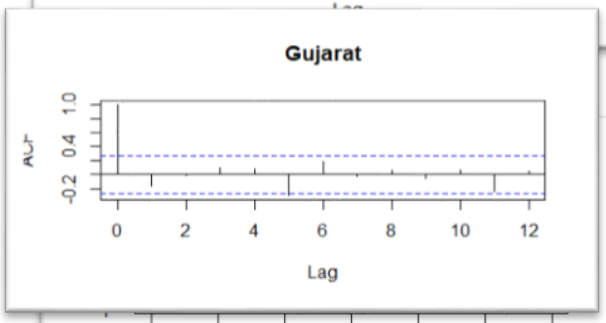


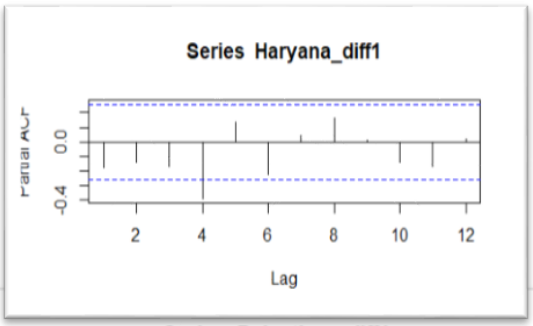


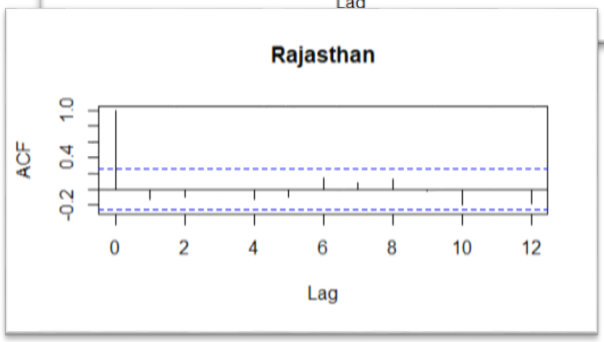


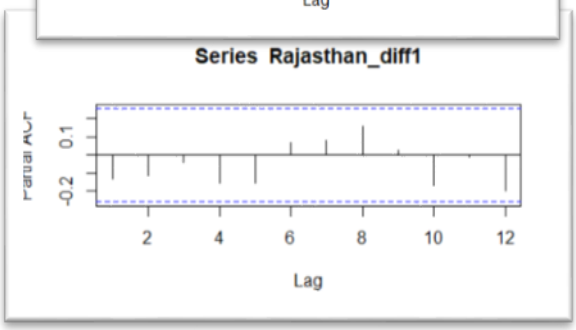

Karnataka
ACF
Lag

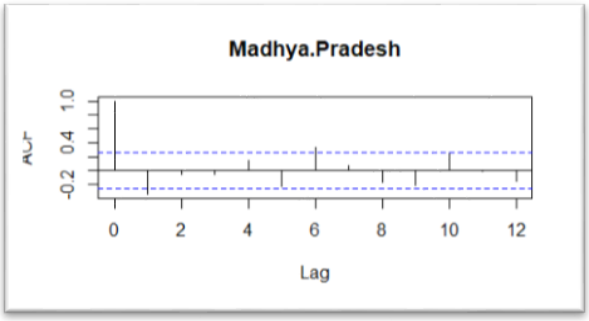

Series Karnataka_diff1
Partial ACF
Lag

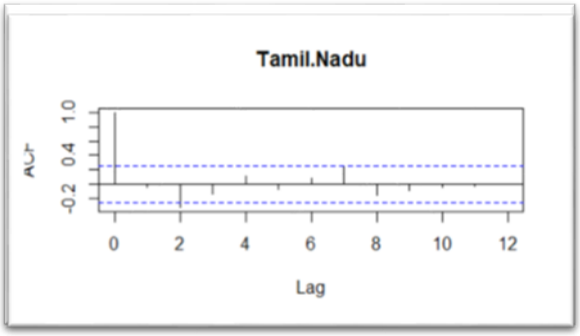

Gujarat
ACF
Lag

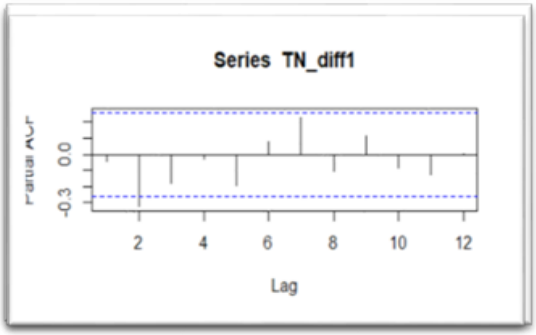

Series Haryana_diff1
Partial ACF
Lag


Rajasthan
ACF
Lag


Series Rajasthan_diff1
Partial ACF
Lag


Madhya.Pradesh
ACF
Lag


Tamil.Nadu
ACF
Lag


Series TN_diff1
Partial ACF
Lag

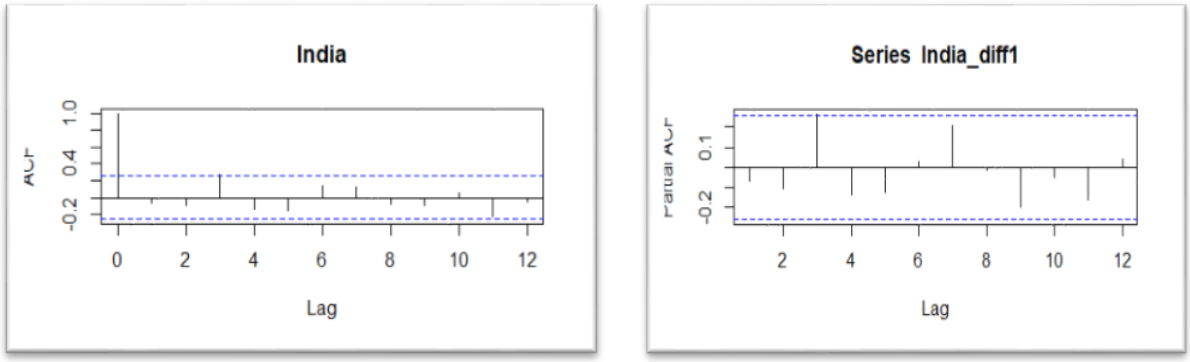
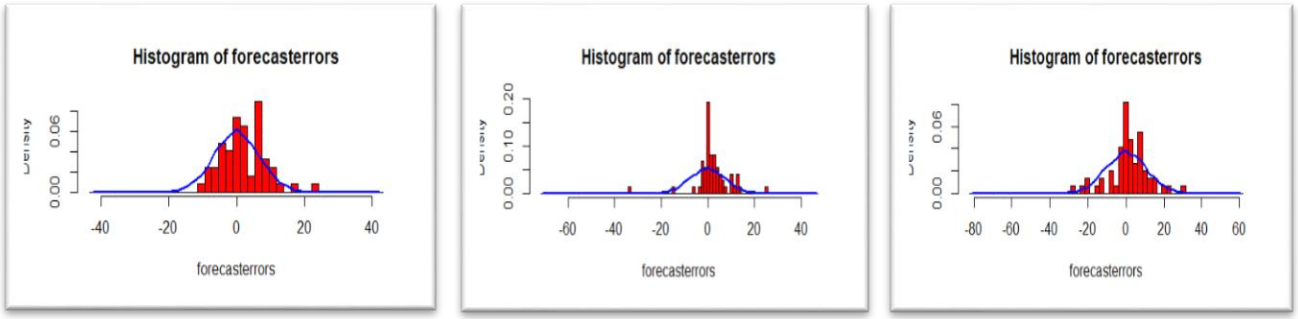
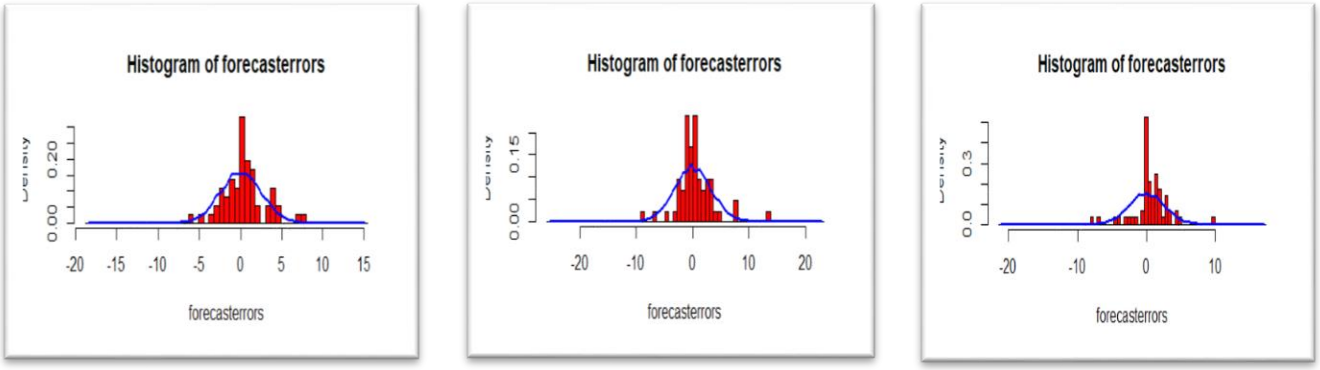


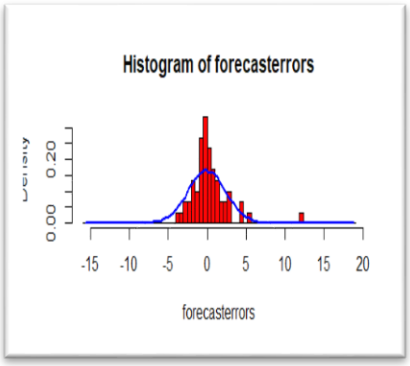
Figure 1: Auto-correlation function (ACF) and Partial auto-correlation function (PACF) of fitted ARIMA models for major cotton producing states



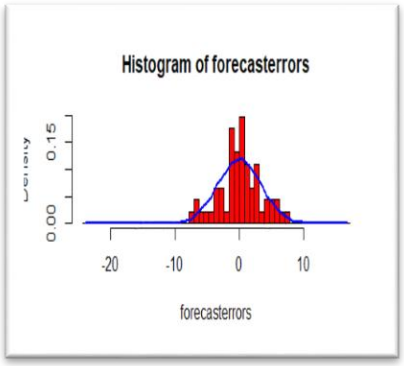
Maharashtra Andhra Pradesh Gujarat



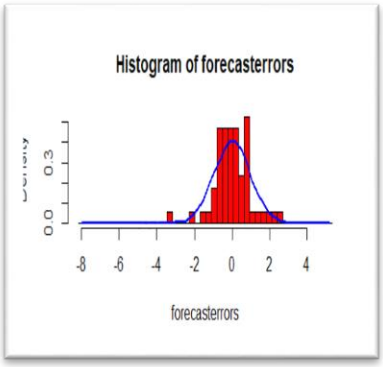
Rajasthan Karnataka Haryana



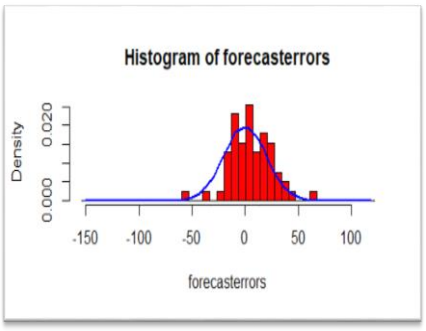
Madhya Pradesh



Punjab

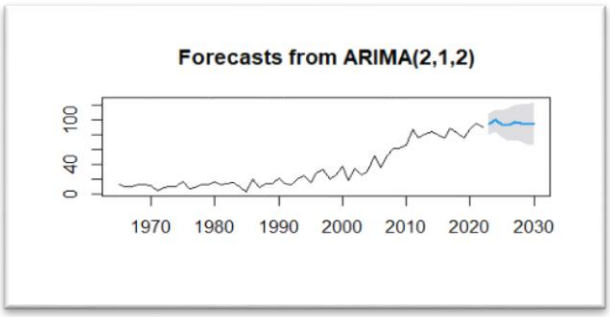


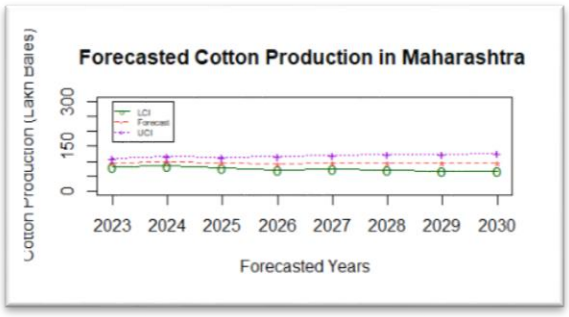
Tamil Nadu

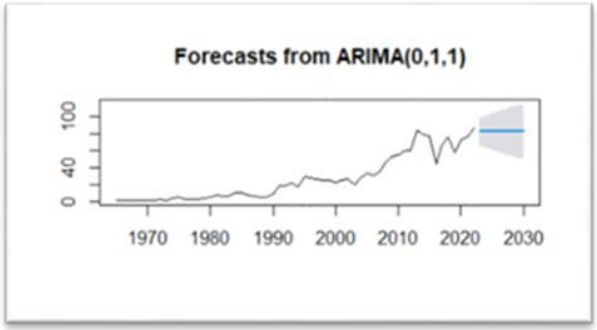


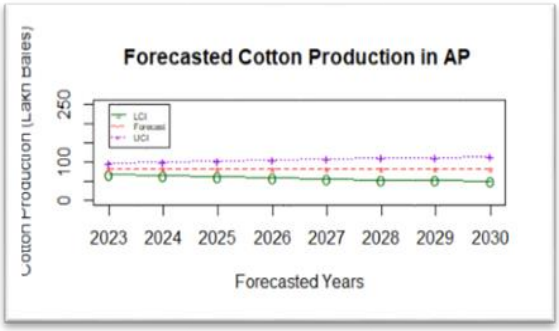
India

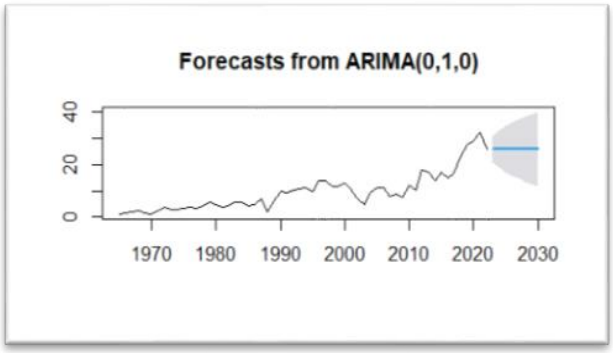
Figure 2: Histogram of forecast errors of residuals of major cotton production states in India

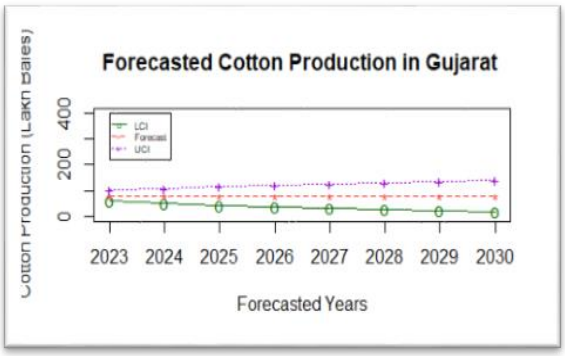


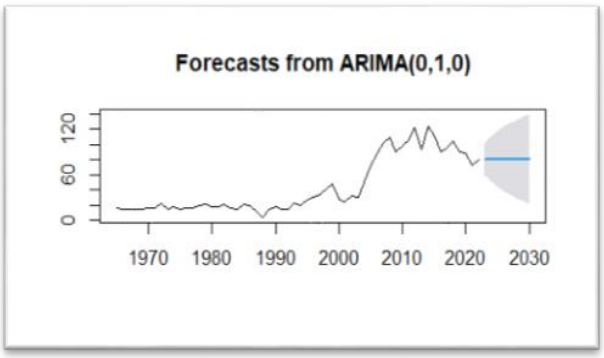


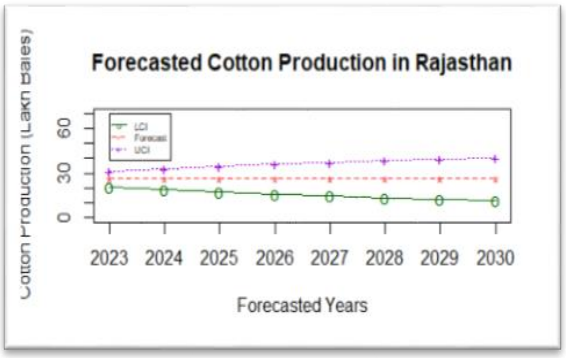


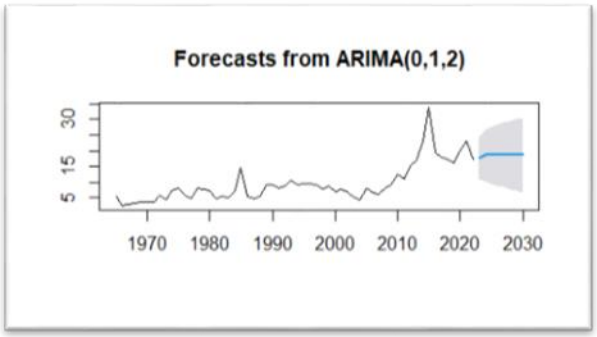


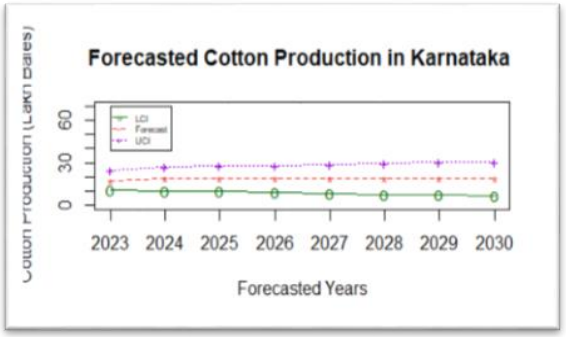

Forecasts from ARIMA(0,1,0)
0
20
40
1970
1980
1990
2000
2010
2020
2030

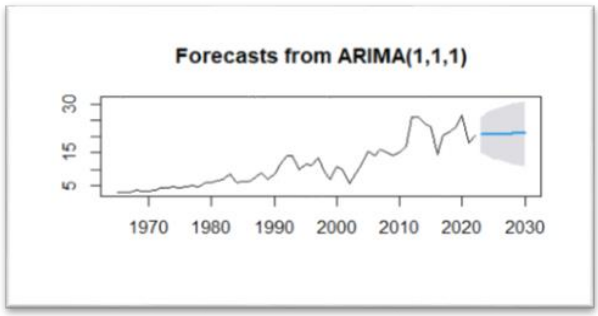

Forecasted Cotton Production in Gujarat
Cotton Production (Lakh Bales)
0
200
400
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

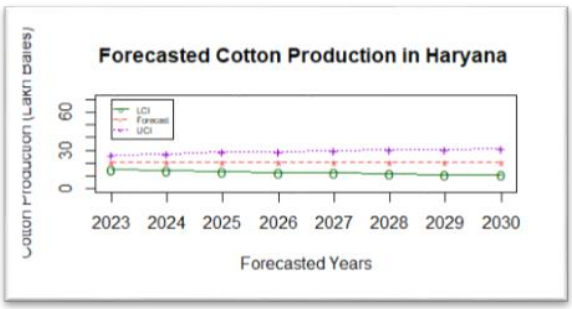

Forecasts from ARIMA(0,1,0)
0
60
120
1970
1980
1990
2000
2010
2020
2030

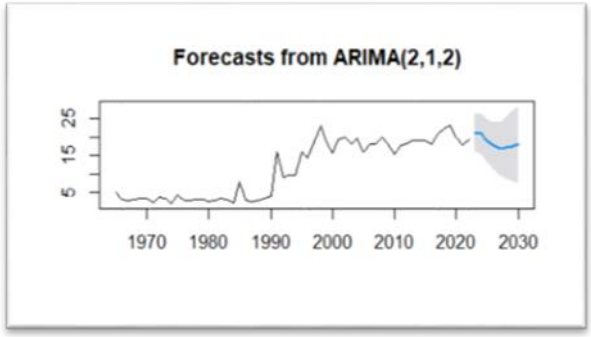

Forecasted Cotton Production in Rajasthan
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

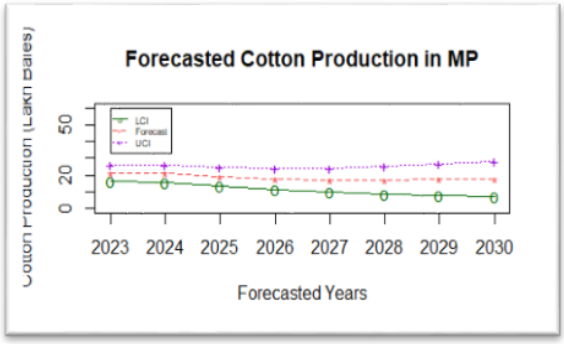

Forecasts from ARIMA(0,1,2)
5
15
30
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in Karnataka
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years


Forecasts from ARIMA(1,1,1)
5
15
30
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in Haryana
Cotton Production (Lakh Bales)
0
30
60
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years


Forecasts from ARIMA(2,1,2)
5
15
25
1970
1980
1990
2000
2010
2020
2030


Forecasted Cotton Production in MP
Cotton Production (Lakh Bales)
0
20
50
LCI
Forecast
UCI
2023
2024
2025
2026
2027
2028
2029
2030
Forecasted Years

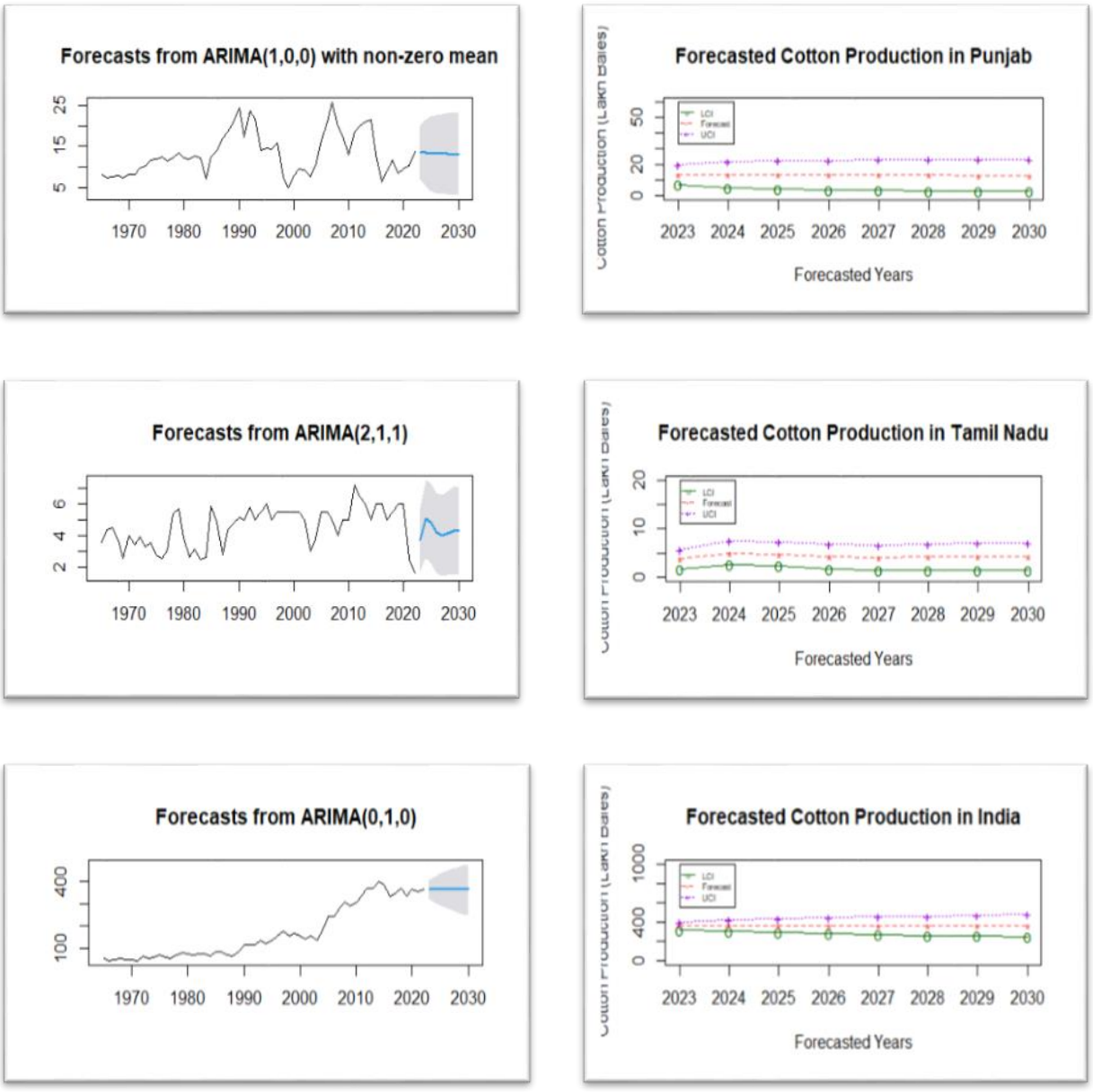


Figure 3: Observed and forecasted plots of major cotton production states in India

Disclaimer: The views expressed in this paper are of the author only and that the Gokhale Institute is not responsible for it.