FORECASTING COTTON PRODUCTION IN INDIA USING ARIMA MODEL

M. Gopinath N. Bharath[2] M. Kavithamani[3]

[1] Assistant Professor, Department of Mathematics, Sri Krishna Arts and Science College, Coimbatore

[2]Assistant Professor, Department of Mathematics, Sri Krishna Arts and Science College, Coimbatore

[3]Assistant Professor, Department of Mathematics, Sri Krishna Arts and Science College, Coimbatore

Abstract : Cotton production plays major role in India, it gives the elementary raw material which is cotton fibre to the Cotton Industries to produce the yarn, from that the industries are producing cloths. The cotton seed called 'Binola' is using the some of the industries to produce 'Vanaspathi', also it is useful for milk cattle to get more milk. The main objective of this paper is to forecast the cotton Production in India. The Auto Regressive Integrated Moving Average (ARIMA) models were used to forecast the future value of the production of cotton in all over the India. The model parameters were found using maximum likelihood method, then the software Eviews 9 were used to predict the future value with the help of ARIMA models output. In this study the ARIMA (1,1,0) model were used to find out the smallest value of mean absolute percentage error (MAPE).

Keywords: Forecasting, Prediction, Box-Jenkins, Auto-Regressive Integrated Moving Average (ARIMA), Auto Regressive (AR), Moving Average (MA), Autocorrelation Function (ACF), Partial Autocorrelation Function (PACF).

I. INTRODUCTION

The first witness of cotton utilization was found in India and dates from about 6000 B.C. scientists assume that cotton was initially cultivated in the Indus delta. Cotton is most important cash crop and economy of our country. India alone provides 6 million formers and 40-50 million people are occupied in the cotton business and it's processing. Also India is the first place in production of cotton.

In India there are several cotton growing states which are segregated into three major zones, particularly north zone, south zone and central zone. North zone contains Punjab, Rajasthan and Haryana. South zone consist of Tamil Nadu, Andhra Pradesh, Karnataka and Telangana. Central zone comprises Madhya Pradesh, Gujarat and Maharashtra. Cotton is also produce in the small areas like Uttar Pradesh, Tripura and West Bengal.

The government of India has developed "Technology Mission on Cotton" in the year 2000, the objective of the development is to improve the production of cotton and develop the high yielding. Raised the global demand for fibre should motivate the highest production in the upcoming decades. With a help of these information it is needed to know the industry of cotton cultivation in future with a help of available data sources. Several methods have been used for predicting such agricultural systems. [1] From the various research about ARIMA are explained the modelling and forecasting. [2] Several models of demand forecast include Auto Regressive (AR), Moving Average (MA), Auto Regressive Moving Average (ARMA), and Auto Regressive Integrated Moving Average (ARIMA). [3] Compared with AR, MA and ARMA model, ARIMA model is pliable in the function and more detailed in the quality of simulative forecasting results. [4] The ARIMA analysis, an identified fundamental process is generated based observations. [5] Some earlier research about demand prediction with enduring typical and can be stored. The objective of this research is to choose a suitable ARIMA model in forecasting cotton production.

The article is systematized as follows, Section-1 contains background of the research and problems in the real system. The basic theory and applications have been discussed in section-2. In Section-3, present a basic methodology and solutions to the particular problem. Section-4 gives the analysis and discussion. Finally the conclusions have drawn in section-5.

1. METHODOLOGY

The Box – Jenkins method or ARIMA is used for predicting short terms. For the long term process this output cannot stable. ARIMA can be defined as the assemblage of two autoregressive (AR) model that is integrated with the Moving Average (MA) model. By writing the notation of Autoregressive

Integrated Moving Average is an ARIMA (p, d, q) [6]. P is the degree of process of AR, d is the order of differencing and q is the degree of MA process.

Autoregressive model with the ordo of the AR (p) model of ARIMA (p,0,0) is stated as follows [7]:

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \ldots + \theta_p Y_{t-p} + e_t \quad (1)$$

Where :
Y_t = Stationary time series
θ_0 = Constant
θ_p = Parameter of autoregressive model
e_t = Residual time (t)

Moving Average model with the ordo of the MA (q) or ARIMA (0,0, q) is stated as follows:

$$Y_t = \theta_0 — \theta_1 Y_{t-1} — \theta_2 Y_{t-2} — \theta — \theta_q Y_{t-q} + e_t \quad (2)$$

Where :
Y_t = Stationary time series
θ_0 = Constant
θ_q = Coefficient of the model which shows the moving average weights
e_t = Residual tense used

To Check the results available right from ARIMA has precise and decrease the level of error can be utilized with four models-selection criteria comprise root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE) and Theil Inequality Coefficient.

2. PRELIMINARIES

TABLE 1

Criteria	Formula
Root Mean Square Error (RMSE)	$\sqrt{\sum_{t=T+1}^{T+h} (\hat{y} - y_t)^2 / h}$
Mean Absolute Error (MAE)	$\sum_{t=T+1}^{T+h} \lvert \hat{y} - y_t \rvert / h$
Mean Absolute Percentage Error(MAPE)	$100 \sum_{t=T+1}^{T+h} \left\lvert \frac{\hat{y}_t - y_t}{y_t} \right\rvert / h$
Theil Inequality Coefficient	$\frac{\sqrt{\sum_{t=T+1}^{T+h} (\hat{y}_t - y_t)^2 / h}}{\sqrt{\sum_{t=T+1}^{T+h} \hat{y}_t^{\,2} / h} - \sqrt{\sum_{t=T+1}^{T+h} y_t^{\,2} / h}}$

The first estimation decisive factor, RMSE is conserving the units of the assessment variable. This approach is highly sensitive and decreases large errors. However, the ability to evaluate different time series is restricted with this criterion. On the other hand, MAE, the second criterion, establishes the error level for a precise set of forecasts. MAE describes how close forecasts are to the real outcomes. This metric does not reflect on the direction of the forecasts. In addition, these criteria decide the accuracies of continuous variables. The third criteria is Theil Inequality Coefficient (U_1 and U_2), in that order. The former allows different forecasts to be evaluated, which implies that definite values are evaluated with forecasted values. U_1 presents a array of values on a zero-to-one scale. The nearer U_1 is to zero, the more accurate the forecast is. While faced with substitute predictions, the prediction with the smallest value of U_1 is observed as the most excellent and is thus selected. On the other hand, U_2 carry outs relative comparisons rooted in random walk models and prediction models (naïve model). The naïve model may be explained as the actual programmed forecast model applied rooted in an indiscriminate-walk process. While U_2 levels off at unity, the naïve method is measured to be equally useful for predicting. $U_2 < 1$ point outs that the predictive model would work much better than the naïve approach. MAPE, the fourth criterion, allows comparison of different time-series data without major relation or percent error. This metric is significant in examples in which the measured variables are very big. In this research using MAPE since data availability.

This research is based on estimation of parameter, model, and forecasting production of cotton. The data were used for this study with a help of Kaggle data source. Analysis of the performance data contains test and non-test stationary make use of the ADF test, following that analysis model applied Box-Jenkins method and software Eviews 9. Box Jenkins method used for estimation model equations mean. At this stage the data confirmation and justification problem analysis in order to time-series and prediction parameter from cotton production index data so attained the best model to fit the actual conditions.

3. DISCUSSION AND ANALYSIS

The matter of this research was production of cotton in India, data sample for the study is obtained from the kaggle data source, the plot represent the time series plot for the production of cotton for original data.

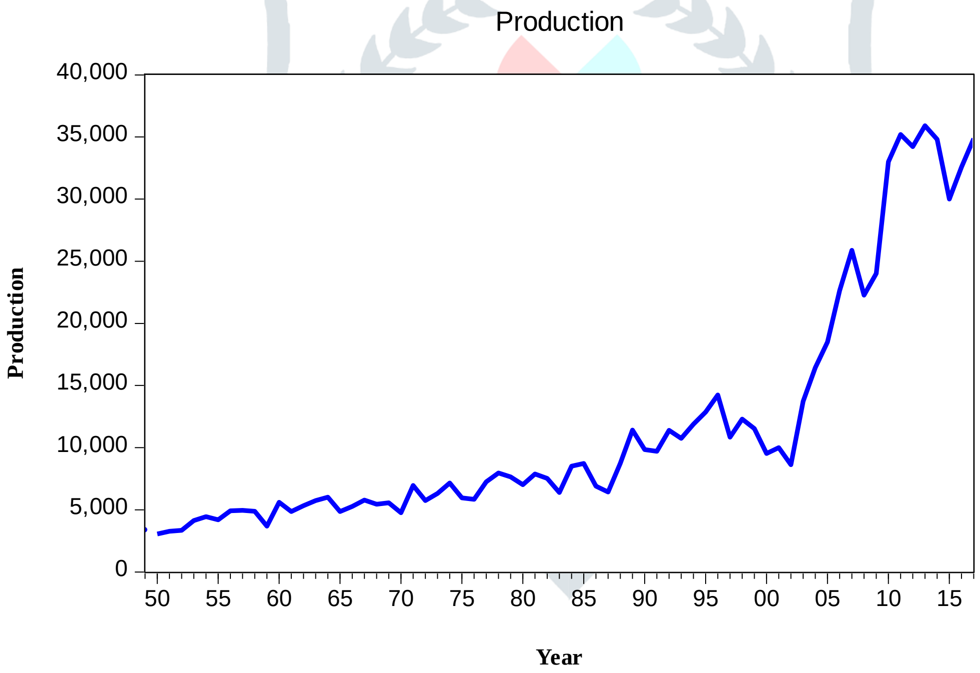


Figure – 1 Plot of original Production data

From the above plots explains that the amount of production is very fluctuating that tends to upward. Based on the above data plot it indicates that the data has not been stationer against the mean and variation of the original data. Particularly it is needed to be done to test the Augmented Dickey-Fuller (ADF) so that known production of cotton data has stationary. The result of the ADF test looks like Table 2.

Table 2 ADF Test

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		0.708896	0.9916
Test critical values:	1% level	-3.530030	

5% level	-2.904848
10% level	-2.589907

*MacKinnon (1996) one-sided p-values.

The value of the t-statistic in output is in output is 0.708896, still smaller than the value in table t Mackinnon at trust level 1%, 5% or 10%. As well as the value of the probability of 0.9916 is still greater than the value of the critique of $\alpha = 0.05 (0.9916 > 0.05)$. The result of the output indicates that the data are not stationary. This data indicates need for differentiation and transformation. So that the data becomes stationary, ADF test done first with differentiation result as in table 3.

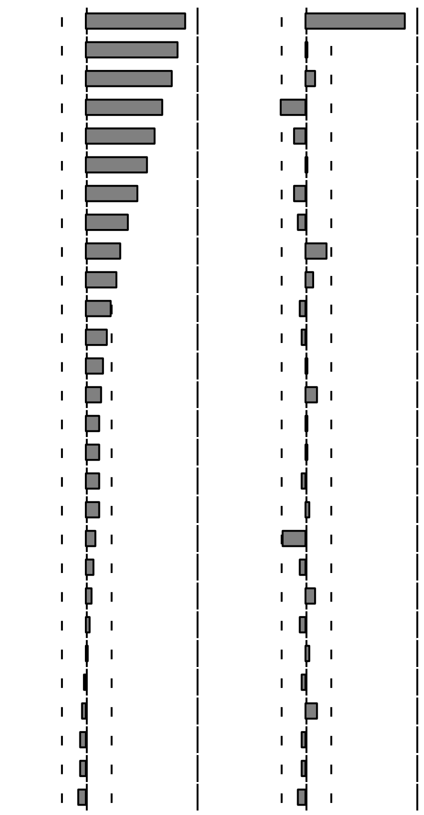
Table 3. ADF Test with First Differences

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.45597	0.0000
Test critical values:	1% level	-3.534868	
	5% level	-2.906923	
	10% level	-2.591006	

*MacKinnon (1996) one-sided p-values.

The value of t-statistic in output is -12.45597, already greater than the value in table t McKinnon at trust level 5% and 10%. As well as the value of the probability of 0.0000 is already smaller than the value of the critique of 0.05 (0.0000 < 0.05). Thus the data has been stationary on the differentiation of the first stage (1st difference) and the null hypothesis can be rejected. After that, the next process is to do an analysis of the time series model with ARIMA.

ACF and PACF plot made to identify a suitable data for means of data. Then the result of the correlogram with a first differentiation will show ACF and PACF graph like figure 2.



	AC	PAC	Q-Stat	Prob
1	0.927	0.927	61.928	0.000
2	0.859	-0.007	115.83	0.000
3	0.809	0.097	164.37	0.000
4	0.729	-0.233	204.45	0.000
5	0.640	-0.110	235.83	0.000
6	0.566	-0.001	260.72	0.000
7	0.481	-0.101	279.00	0.000
8	0.388	-0.078	291.09	0.000
9	0.333	0.190	300.17	0.000
10	0.292	0.075	307.23	0.000
11	0.235	-0.059	311.90	0.000
12	0.189	-0.049	314.96	0.000
13	0.159	0.002	317.16	0.000
14	0.140	0.108	318.92	0.000
15	0.128	0.021	320.40	0.000
16	0.126	0.019	321.88	0.000
17	0.120	-0.032	323.23	0.000
18	0.116	0.030	324.51	0.000
19	0.094	-0.219	325.38	0.000
20	0.073	-0.057	325.91	0.000
21	0.064	0.082	326.33	0.000
22	0.036	-0.061	326.47	0.000
23	0.007	0.036	326.47	0.000
24	-0.016	-0.027	326.50	0.000
25	-0.029	0.102	326.59	0.000
26	-0.050	-0.031	326.88	0.000
27	-0.065	-0.047	327.37	0.000
28	-0.078	-0.069	328.09	0.000

Figure 2. ACF and PACE

From the above graph model, it can be predicted that the model of ARIMA is used for proper ARIMA (1,1,0), ARIMA (0,1,1), ARIMA (1,1,1) without constant. Next do the estimation of the value of C, probability, and AIC on each model.

TABLE 4. Models of ARIMA (1,1,0)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	464.2571	273.2514	1.699011	0.0941
AR(1)	0.039248	0.123118	0.318784	0.7509
SIGMASQ	3921066.	423562.0	9.257360	0.0000
R-squared	0.001563	Mean dependent var	463.6176	
Adjusted R-squared	-0.029158	S.D. dependent var	1996.452	
S.E. of regression	2025.349	Akaike info criterion	18.10801	
Sum squared resid	2.67E+08	Schwarz criterion	18.20593	
Log likelihood	-612.6723	Hannan-Quinn criter.	18.14681	
F-statistic	0.050883	Durbin-Watson stat	1.966344	
Prob(F-statistic)	0.950428			
Inverted AR Roots	.04			

TABLE 5. ARIMA (0,1,1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	464.6887	281.5024	1.650745	0.1036
MA(1)	0.084012	0.123587	0.679783	0.4991
SIGMASQ	3914227.	443623.3	8.823314	0.0000
R-squared	0.003305	Mean dependent var	463.6176	
Adjusted R-squared	-0.027363	S.D. dependent var	1996.452	
S.E. of regression	2023.582	Akaike info criterion	18.10635	
Sum squared resid	2.66E+08	Schwarz criterion	18.20426	
Log likelihood	-612.6157	Hannan-Quinn criter.	18.14514	
F-statistic	0.107755	Durbin-Watson stat	2.032064	
Prob(F-statistic)	0.898007			
Inverted MA Roots	-.08			

TABLE 6. Models of ARIMA (1,1,1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	462.9354	284.7213	1.625925	0.1089
AR(1)	-0.728185	0.092646	-7.859869	0.0000
MA(1)	1.000000	378.8569	0.002640	0.9979
SIGMASQ	3386462.	28648652	0.118207	0.9063

R-squared	0.137692	Mean dependent var	463.6176
Adjusted R-squared	0.097271	S.D. dependent var	1996.452
S.E. of regression	1896.870	Akaike info criterion	18.02698
Sum squared resid	2.30E+08	Schwarz criterion	18.15754
Log likelihood	-608.9175	Hannan-Quinn criter.	18.07872
F-statistic	3.406461	Durbin-Watson stat	2.161237
Prob(F-statistic)	0.022766		

Inverted AR Roots	-.73
Inverted MA Roots	-1.00

To determine the best model is to compare to the four models that is looking for a model with a value of AIC and Schwarz criterion to the smallest. From the results above, it is well known that the best model is the ARIMA (1,1,0) without constant. Next is doing a diagnostic check to perform a test of normality residue. The results can be seen in Figure 3.

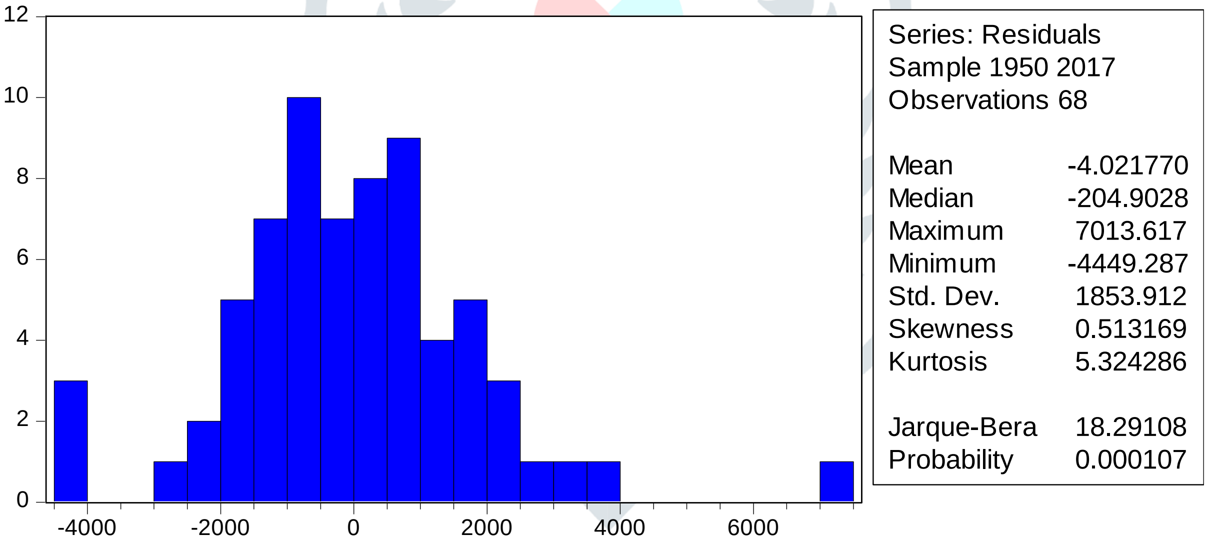


Figure 3. The Results of the Diagnostic Check

Based on the above histogram and descriptive statistics the data is normal and has been stationary against the variation. This implies that these data have relative stable fluctuations from time to time. To prove that data are already normal can use assumptions auto correlation test and assumption hetroscedasticity test.

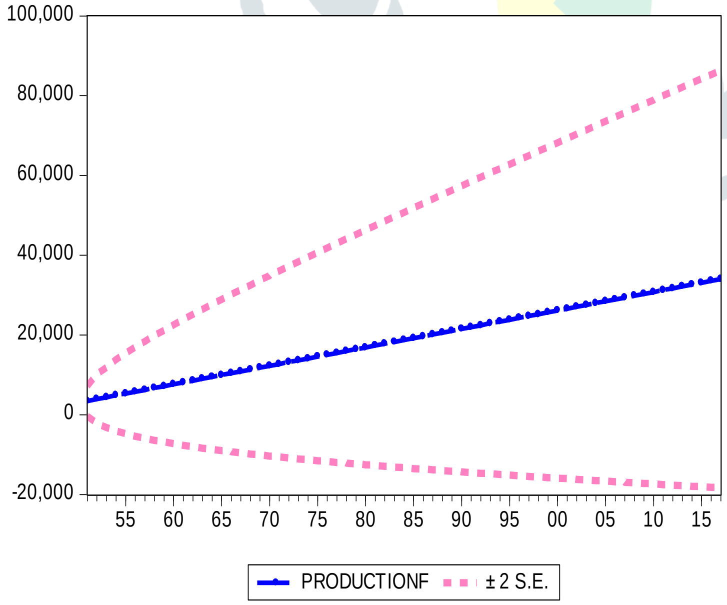
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.083	-0.083	0.4864	
		2	-0.135	-0.143	1.7982	
		3	0.129	0.107	3.0097	0.083
		4	0.124	0.130	4.1594	0.125
		5	-0.240	-0.198	8.5233	0.036
		6	0.092	0.082	9.1794	0.057
		7	0.304	0.265	16.398	0.006
		8	-0.151	-0.079	18.198	0.006
		9	-0.131	-0.082	19.591	0.007
		10	-0.073	-0.250	20.026	0.010
		11	0.088	0.064	20.680	0.014
		12	-0.096	0.058	21.461	0.018
		13	-0.073	-0.151	21.920	0.025
		14	0.065	-0.041	22.289	0.034
		15	0.166	0.235	24.778	0.025
		16	-0.062	0.140	25.125	0.033
		17	-0.018	0.048	25.154	0.048
		18	0.200	0.030	28.955	0.024
		19	-0.104	-0.123	30.014	0.026
		20	-0.164	-0.075	32.683	0.018
		21	0.094	-0.031	33.578	0.021
		22	0.147	-0.024	35.803	0.016
		23	-0.107	0.009	37.020	0.017
		24	-0.108	-0.140	38.285	0.017
		25	0.103	0.059	39.469	0.018
		26	-0.056	0.116	39.831	0.022
		27	-0.066	0.103	40.333	0.027
		28	0.034	0.003	40.468	0.035

FIGURE 4. Correlation Assumption

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.095	0.095	0.6355	0.425
		2	0.226	0.220	4.3354	0.114
		3	0.012	-0.027	4.3454	0.227
		4	0.138	0.095	5.7719	0.217
		5	0.274	0.277	11.437	0.043
		6	0.139	0.061	12.919	0.044
		7	0.234	0.135	17.199	0.016
		8	-0.002	-0.054	17.200	0.028
		9	0.093	-0.020	17.890	0.036
		10	0.004	-0.065	17.891	0.057
		11	0.156	0.068	19.912	0.047
		12	0.103	0.019	20.821	0.053
		13	0.223	0.187	25.135	0.022
		14	-0.052	-0.135	25.376	0.031
		15	-0.047	-0.114	25.574	0.043
		16	-0.008	-0.021	25.580	0.060
		17	0.017	-0.031	25.608	0.082
		18	0.060	-0.076	25.951	0.101
		19	-0.019	0.012	25.985	0.131
		20	0.008	-0.001	25.991	0.166
		21	-0.015	0.083	26.012	0.206
		22	0.006	0.019	26.016	0.251
		23	-0.051	-0.064	26.289	0.287
		24	-0.006	-0.041	26.293	0.338
		25	-0.030	-0.024	26.395	0.387
		26	-0.008	-0.037	26.402	0.441
		27	-0.039	0.037	26.579	0.487
		28	-0.042	0.019	26.789	0.530

FIGURE 5. Test Assumption Hetroscedasticity

After that it can be determined the sales forecast for the short period of time. The results of the Prediction have shown in Figure 6. From the result MSE is 8678.668; MAE is 7324.376; MAPE is 84.72279.



Forecast: PRODUCTIONF	
Actual: PRODUCTION	
Forecast sample: 1947 2017	
Adjusted sample: 1951 2017	
Included observations: 67	
Root Mean Squared Error	8678.668
Mean Absolute Error	7324.376
Mean Abs. Percent Error	84.72279
Theil Inequality Coefficient	0.241652
Bias Proportion	0.624812
Variance Proportion	0.003100
Covariance Proportion	0.372087

FIGURE 6. The Result of the Forecast

From the above diagram shows the forecasted value of the cotton production in India. Obtained ARIMA models that enable the following:

$$\Delta Y_t = \theta_0 + \theta_1 \Delta Y_{t-1} + e_{t-1}$$
$$\Delta Y_t = C + 0.039248 \Delta Y_t + e_{t-1}$$
$$\Delta Y_t = 464.2571 + 0.039248 + 84.7$$
$$\Delta Y_t = \pm 548$$

From the above result the production of cotton for each year is ± 548

4. CONCLUSION

As a conclusion, the ARIMA model works better forecast when compare with other models. Hence, this ARIMA model is used to predict the future value of cotton production. This method helps us to improve the future production.

Our main objective of this analysis is to find out the appropriate ARIMA model for the production yearly data for cotton in India. On the other hand we were interested in forecasting future production value of this cotton using this model. From the above analysis we can conclude that the future cotton production will be ± 548.

REFERENCES

1. Carina Intan Permatasaria, Wahyudi Sutopob, and Muh. Hisjamc, Sales Forecasting Newspaper with ARIMA: A Case Study, The 3rd International Conference on Industrial, Mechanical, Electrical, and Chemical Engineering, AIP Conf. Proc. 1931, 030017-1–030017-10; https://doi.org/10.1063/1.5024076 Published by AIP Publishing. 978-0-7354-1623-9/$30.00
2. Singh, A. P., Manoj Kumar Gaur, Dinesh Kumar Kasdekar, and Sharad Agrawal, A Study of Time Series Model for Forecasting of Boot in Shoe Industry, *International Journal of Hybrid Information Technology*, vol. 8, no. 8, pp. 143 – 152, 2015.
3. Cardoso, G. and F. Gomide, Newspaper Demand Prediction and Replacement Model based on Fuzzy Clustering and Rules, *International Journal of Information Sciences*, pp. 4799 – 4809, 2007.
4. Spil and Kijl, *A Business Model for the E-Newspaper from a Customer Perspective*, Master thesis: MSc in Business administration, University of Twente, 2014.
5. Lertuthai, M., Manisara Baramichai, and Ungul Laptaned, Development of the Adaptive Forecasting Model for Retail Commodities by Using Leading Indicator: Retailing Rule Based Forecasting Model (RRBF), *Proceedings of the World Congress on Engineering and Computer Science*, vol. 2, 2009.
6. Incesu G., Baris Asikgil, and Mujgan Tez, Sales Forecasting System for Newspaper Distribution Companies in Turkey, *Pak.j.sts.oper.res.* vol. 8, no. 3, pp. 685 – 699, 2012.
7. Rojas, Juan Pedro Sepúlveda, Felipe Rojas, Héctor Valdés-González, dan Mario San Martín, Forecasting Models Selection Mechanism for Supply Chain Demand Estimation, *Procedia Computer Science*, pp. 1060 – 1068, 2015.
8. Adhikari, R. and R.K. Agrawal, *An Introductory Study on Time Series Modeling and Forecasting*, English, Lambert Academic Publishing (LAP), 2013.
9. Raicharoen, T., C. Lursinsap, P. Sanguanbhoki, Application of Critical Support Vector Machine to Time Series Prediction, *Proceedings of the 2003 International Symposium, ISCAS '03*, vol. 5, pp. 741 – 744, 2003.
10. Box, G.E.P. and Jenkins, G.M., *Time Series Analysis: Forecasting and Control*, USA, Revised ed. Holden-Day, 1976.
11. Jansson, Jan Owen, Car Demand Modelling and Forecasting, *Journal of Transport Economics and Policy*, pp. 125 – 140, 1989.
12. Udom, P. and N. Phumchusri, A Comparison Study between Time Series Model and ARIMA Model for Sales Forecasting of Distributor in Plastic Industry, *IOSR Journal of Engineering*, vol. 4, no. 2, pp. 32 – 38, 2014.