Time Series Model to Forecast Production of Cotton from India: An Application of Arima Model

***Sundar rajan**
***Palanivel**

*Research Scholar, Department of Statistics, Govt Arts College, Udumalpet, Tamilnadu, India
*Assistant Professor, Department of Statistics, Govt Arts College, Coimbatore, Tamilnadu, India

Abstract

This paper attempts forecasting the cotton production of India by fitting of univariate Auto regressive Integrated Moving Average (ARIMA) models The data on cotton production collected during the years from 1951 to 2021 are calculated based on the selected model. The study considered ARIMA model was introduced by Box and Jenkins. This study introduces a forecasting Cotton production in India. Based on ARIMA (p,d,q) and its components ACF, PACF, Normalized BIC, Box-Ljung Q statistics and residuals estimated, ARIMA model (0,1,0) was selected for forecasting model and the analysis revealed that ARIMA (0,1,0) was the best model for forecasting cotton production. Although, eight years forecast with the model shows an increasing trend in production, the forecast value 47.25 million bales (of 170kgs each) in 2021. Based on the chosen model, it could be predicted that the cotton production would increase to 47.25 million bales (of 170kgs each) in 2021 from 35.1 million bales (of 170kgs each) in 2012-13 in India. This study also estimates increase in the production of cotton in future.

Keywords: (Autoregressive Integrated Moving Average) ARIMA, Cotton production, Forecasting, ACF, PACF, Normalized BIC.

Introduction

India is an agricultural country with about 80% of its population dependent on income from agriculture. Cotton is an important cash crop in India and plays a significant role in the national economy and one of the most ancient and every important commercial fibre crop at global importance with a significant role in Indian agriculture industrial development and improving the national economy. An India's total cotton cultivation areas in Gujrat, Maharastra, Andhra Pradesh and Madhya Pradesh are the four main states which contribute 80 percent of the total cotton production in the country.

The cotton Association of India (CAI) has placed the cotton crop for the season 2012-13 at 35.1[1] million bales (1 bales=170kgs) as against 37.3[1] million bales (1 bales=170kgs) in 2011-12. Cotton the "white gold" is a leading commercial crop grown for its valuable fibre. India being one of the oldest countries in the world for domesticated cotton production and manufacture of cotton fabrics has also become the largest grower of cotton.

The major producers of cotton are America, India, China, Egypt, Pakistan, Uzbekistan, Argentina, Australia, Greece, Brazil, Mexico and Turkey. These countries contribute about 85% to the global cotton production. India has the largest acreage (9.4 m.ha) under cotton at global level and has the productivity of 560 kg Lint/ha and

ranks second in production (5.334 m.MT 31.0 m. bales) after China during 2007/2008.

In the present study, ARIMA stochastic modeling is used on the cotton production of India for forecasting purpose.

Objectives of study

The objectives of present research are mentioned below:

1. To suggest appropriate ARIMA model for the generation of forecasting production of Cotton in India and to make eight year forecasts with appropriate prediction interval.
2. To generate forecasts of production of Cotton in India by using appropriate ARIMA models.

(Business Line 11.04.2013)

Material and Methods

This study was based on time series data for Cotton production in India is collected from the Agricultural Statistics at a Glance for the period 1950 to 2011 and Business Line. Box and Jenkins (1970) was frequently used for discovering the pattern and predicting the future values of the time series data. The most popular and widely used forecasting models for uni-variate time series data. Akaike discussed with the stationary time series by an AR(p), p is finite and bounded by the same integer. Moving Average (MA) models were used by Slutzky(1973). Hannan and Quinn (1979) for pure AR models and Hannan (1980) for ARIMA models. A second order determination method could be considered as a variance of Schwarz's Bayesian Criterion (SBC) which gives a consistent estimate of the order of an ARMA model. Hosking (1981) introduced a family of models; called fractionally differenced autoregressive integrated moving average models. In general, ARIMA model is characterized by the notation ARIMA (p,d,q):

- A p^{th}-order autoregressive model: AR(p),which has the general form:

$Y_t=\phi_0+ \phi_1 Y_{t-1} + \phi_2 Y_{t-2} +....+ \phi_p Y_{t-p} +\varepsilon_t$

Where,

Y_t= Response (dependent) variable at time t

$Y_{t-1}, Y_{t-2},...,Y_{t-p}$= Response variable at time

Lags t-1,t-2,....,t-p, respectively

$\phi_0, \phi_1, \phi_2,...,\phi_p$ = Coefficient to be estimated

ε_t = Error term at time t.

- A q^{th}- order moving average model: MA(q), which has the general form:

$Y_t= \mu+ \varepsilon_t- \theta_1 \varepsilon_{t-1} – \theta_2 \varepsilon_{t-2}-.... - \theta_q \varepsilon_{t-q}$

Where,

Y_t= Response (dependent) variable at time t

μ= Constant mean of the process

$\theta_1 , \theta_2,...., \theta_q$ = Coefficients to be estimated

ε_t = Error term at time t

$\varepsilon_{t-1}, \varepsilon_{t-2},..., \varepsilon_{t-q}$ = Error in previous time

Periods that are incorporated in the response Y_t .

- And the general form of ARIMA model of order (p,d,q) is

$Y_t=\phi_0+ \phi_1 Y_{t-1} + \phi_2 Y_{t-2} +.... + \phi_p Y_{t-p}+ \mu+ \varepsilon_t- \theta_1 \varepsilon_{t-1} – \theta_2 \varepsilon_{t-2}-.... - \theta_q \varepsilon_{t-q}+ \varepsilon_t$

Where Y_t is cotton production, ε_t's are independently and normally distributed with zero mean and constant variance δ^2 for t=1,2,...,n; d is the fraction differenced while interpreting AR and MA and ϕs and θs are coefficients to be estimated.

Stochastic time series ARIMA models were widely used in time series data having the characteristics (Alan Pankratz, 1983) of parsimonious, stationary, invertible, significant estimated coefficients and statistically independent and normally distributed residuals. When a time series is non-stationary, it can often be made stationary by taking first differences of the series i.e., making a new time series of successive differences (Y_t- Y_{t-1}). If first differences do not convert the series to stationary form, then first differences can be created. This is called second-order differencing. A distinction is made between a second-order differences (Y_t- Y_{t-2}).

While Mendelssohn (1981) used Box-Jenkins models to forecast to forecast fishery dynamics, Prajneshu and Venugopalan (1996) discussed various statistical modeling techniques viz., polynomial, ARIMA time series methodology and nonlinear mechanistic growthing approach for describing marine, inland as well as total fish production in India during the period 1950-51 to 1994-95. Tsitsika at al. (2007) also used univariate and multivariate ARIMA models to model and forecast the monthly pelagic production of fish species in the Mediterranean Sea during 1990-2005.Jai Sankar et al. (2010) also used stochastic modeling for cattle in the Tamilnadu state during 1970-2010. Faqir Muhammad, Muhammad Siddique Javed and Mujahid Bashir(1992) also used a forecasting sugarcane production in Pakistan using ARIMA models during 1947-48 to 1988-89.

In general models for time series data can have many forms and represent differenced follows both AR and MA models and is known as Autoregressive integrated moving averages (ARIMA) model. Univariate ARIMA models use only the information contained in the series itself. Thus, models are constructed as linear functions of past values of the series and/or previous random shocks (or error). Forecasts were generated under the assumption that the past history could be translated into predictions for the future. Hence, ARIMA model was used in this study, which required a sufficiently large data set and involved four steps:

1. Model Identification: Orders of AR and MA components were determined.
2. Estimating the parameters: Linear model coefficients were estimated.
3. Diagnostic checking: Certain diagnostic methods were used to test the sutability of the estimated model.
4. Forecasting: The best model chosen was used for forecasting.

The model parameters to be estimated using the Statistical Package for Social Science (SPSS) package and to fit the ARIMA models.

Trend fitting

For evaluating the adequacy of AR, MA and ARIMA processes, various reliability statistics like R^2, Stationary R^2, Root Mean Square Error(RMSE), Mean Absolute Percentage Error(MAPE) and Bayesian Information Criterion(BIC)[as suggested by Schwartz, 1978] have been used. The reliability statistics viz. RMSE, MAPE, BIC and Q statistics have also been used.

Results and Discussion

Model Identification:

In this stage we use ARIMA model was designed after assessing that transforming the variable under forecasting was stationary series. The stationary series was the set of values that varied over time around a constant mean and constant variance. In this most common method to check the stationary was to explain the data through figure and hence is done in Figure 1.

Figure 1 reveals in this data used were non-stationary. And again, non-stationary in mean was corrected through first differencing of the data. The newly constructed variable Y_t could now be examined for stationary. Since, Y_t was stationary in mean, the next step was to identify the values of p and q. For this, the autocorrelation and partial autocorrelation coefficients (ACF and PACF) of various orders of Y_t were computed and presented in Table 1 and Figure 2.

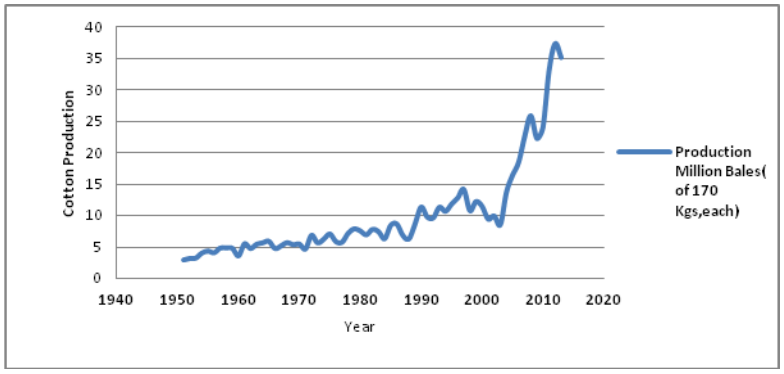


Figure: 1 Time plot of Cotton production in India

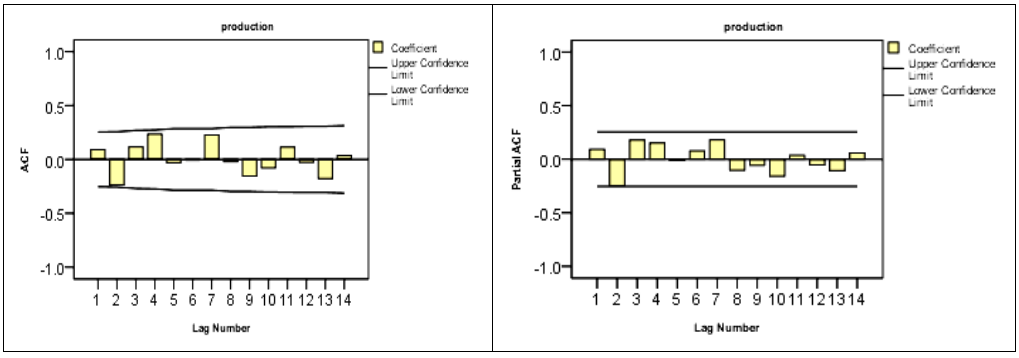


Figure 2. ACF and PACF of differenced data

Table 1.ACF and PACF of Cotton production

Lag	Auto Correlation		Box-Ljung Statistics			Partial Auto Correlation	
	Value	Df	Sig.	Value	Df	Value	Df
1	.093	.127	.562	1	.454	.093	.127
2	-.236	.128	4.247	2	.120	-.247	.127
3	.116	.135	5.149	3	.161	.178	.127
4	.235	.137	8.926	4	.063	.152	.127
5	-.029	.143	8.986	5	.110	-.014	.127
6	-.008	.143	8.991	6	.174	.078	.127
7	.229	.143	12.766	7	.078	.180	.127
8	-.021	.149	12.799	8	.119	-.104	.127
9	-.154	.149	14.577	9	.103	-.058	.127
10	-.079	.151	15.054	10	.130	-.159	.127
11	.114	.152	16.057	11	.139	.036	.127
12	-.027	.153	16.113	12	.186	-.055	.127
13	-.178	.153	18.670	13	.134	-.107	.127
14	.037	.157	18.781	14	.173	.060	.127

The tentative ARIMA models were discussed with values differenced once (d=1) and the model which had the minimum normalized BIC was chosen. The various ARIMA models and the corresponding normalized BIC values are given in Table 2. The value of normalized BIC of the chosen ARIMA was 1.467.

Table 2. BIC values of ARIMA (p,d,q)

0,1,0	**1.467**
0,1,1	1.549
0,1,2	1.539
1,1,0	1.539
1,1,1	1.621
1,1,2	1.550
2,1,0	1.509
2,1,1	1.588
2,1,2	1.557

Model Estimation

The second step was the estimation of model parameters were estimated using SPSS package and the results of estimation were presented in Table 3 and 4. R^2value was 0.93. Hence, the most suitable model for Cotton production was ARIMA (0,1,0), as this model had the lowest normalized BIC value, good R^2 and better model fit statics (RMASE and MAPE). In this justified that the selection of ARIMA(0,1,0) is the best model to represent the data generating process very precisely.

Table 3. Estimated ARIMA model of Cotton production

	Estimate	**SE**	**t**	**Sig.**
Constant	-56.244	27.408	-2.052	.045

Table 4. Estimated ARIMA model fit statistics

Fit Statistics	**Mean**
Stationary R-squared	.067
R-squared	.93
RMSE	1.948
MAPE	14.187
Normalized BIC	1.467

Diagnostic checking

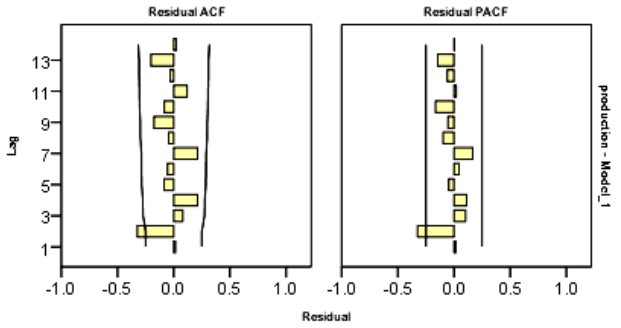
In this model proof that verification was concerned with checking the residuals of the model to see if they contained any systematic pattern which still could be removed to improve the chosen ARIMA, which has been done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, various autocorrelations up to 14 lags were computed and the same along with their significance tested by Box-Ljung statistic are provided in Table 5. As the results indicate, none of these autocorrelations was significantly different from zero at any reasonable level. This proved that the selected ARIMA model was an appropriate model for forecasting Cotton production in India.

Table 5. Residual of ACF and PACF of Cotton production

Lag	ACF		PACF	
	Mean	SE	Mean	SE
1	.015	.127	.015	.127
2	-.325	.127	-.325	.127
3	.082	.140	.104	.127
4	.213	.141	.114	.127
5	-.086	.146	-.047	.127
6	-.055	.146	.047	.127
7	.213	.147	.168	.127
8	-.048	.152	-.097	.127
9	-.176	.152	-.054	.127
10	-.084	.155	-.165	.127
11	.120	.156	.021	.127
12	-.032	.157	-.059	.127
13	-.203	.157	-.144	.127
14	.022	.162	.007	.127

The ACF and PACF of the residuals are given in a Figure 3, which also indicated the 'good fit' of the model.

Figure: 3 ACF and PACF plot of residuals



Forecasting

Based on the model fitted, forecasted cotton production (in million bales of 170kgs each) for the year 2014 through 2021 respectively were 36.52,37.97,39.44,40.95,42.48,44.04,45.63and 47.25 million bales (in 170kgs each) (Table 6). Figure 4 shows the actual and predicted value of cotton production in India.

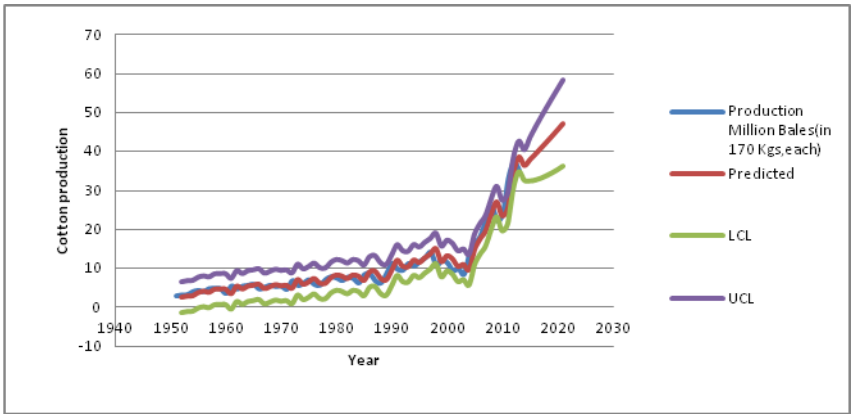


Fig 4. Actual and estimate of Cotton production

Table 6. Forecast for the Production of Cotton in India (in million bales of 170Kgs. each)

Year	Predicted	LCL	UCL
2014	36.52	32.62	40.42
2015	37.97	32.46	43.48
2016	39.44	32.69	46.19
2017	40.95	33.15	48.74
2018	42.48	33.77	51.19
2019	44.04	34.5	53.59
2020	45.63	35.32	55.94
2021	47.25	36.23	58.27

Note: LCL- Lower Confidence Level and **UCL**-Upper Confidence Level

Conclusion

In the following are the conclusions of this study:

- ARIMA is an appropriate model to forecast the production of cotton in India.
- ARIMA (0, 1, and 0) is the most appropriate model to forecast production of cotton in India.
- The forecast indicates that in the year 2021 the production of cotton in India will be 47.25 million bales (in170kgs each), which is 35.1million bales (in170kgs.each) more than the production this year.
- That is, using time series data from 1951 to 2013 on cotton production, this study provides evidence on future cotton production in India, which can be considered for future policy making and formulating strategies for augmenting and sustaining cotton production in India.

Reference:

1. Akaike H. 1970. Statistical Predictor Identification. Annals of Institute of Statistical Mathematics 22: 203-270.
2. Alan Pankratz. 1983. Forecasting with Univariate Box-Jenkins models-concepts and cases. John Wiley, New York, Page 81.
3. Agricultural Statistics at a Glance at 2012, Cotton production in India-Current Scenario-(2012-13) and cotton production 37.3 million bales data (2011-12) from Business Line 11.04.2013.
4. Box G E P and Jenkins J M. 1970. Time series Analysis-Forecasting and Control. Holden-Day Inc., San Francisco.
5. Faqir Muhammad, Muhammad Siddique Javed and Mujahid Bashir(1992).Forecasting sugarcane production in Pakistan using ARIMA models.Pakistan Journal of Agriculture Science.,Vol.9, No.1,1992.
6. Hannan E J and Quinn B G. 1979. The determination of the order of an autoregression. Journal of Royal Statistical Society B(41):190-195.
7. Hannan E J. 1980. The estimation of the order of an ARMA process. Annuals of Statistics 8:1071-1081.
8. Harris.E, Abdul-Aziz, A.R, Avuglah. R.K. 2012. Modeling Annual Coffee Production in Ghana Using ARIMA Time Series Model. International Journal of Business and Social Research, Vol-2, No-7.
9. Hosking J R M.1981. Fractional differencing. Biometrika 68(1):165-176.

10. Jai Sankar, R. Prabakaran, K. Senthamarai Kannan, and S. Suresh, "Stochastic Modeling for Cattle Production Forecasting." Journal of Modern Mathematics and Statistics 4(2): 53-57, 2010.

11. Khadi B M (University of Agricultural Sciences, Dharward, Karnataka. India). Biotech Cotton: Issues for consideration.

12. Mendelssohn R. 1981. Using Box-Jenkins models to forecast fishery dynamics: Identification, Model estimation and checking. Fishery Bulletin 78(4): 887-896.
13. Slutzky E. 1973. The summation of random causes as the source of cyclic processes. Econometrica 5:105-146.
14. Tsitsika E V,Maravelias C D and Haralabous J. 2007. Modeling and forecasting pelagic fish production using univariate and multivariate ARIMA models. Fisheries Science 73:979-988.
15. Wankhade, R. 2, 97-102. Use of the ARIMA model for forecasting pigeonpea production in India. International Review of Business.