Environment & Ecology 35 (1A) : 224—228, January—March 2017
Website: environmentandecology.com ISSN 0970-0420

# ARIMA Modelling and Forecasting of Cotton Productivity in India

**Manoj Kumar, Rajendra, R. C. Hasija**

Received 28 March 2016 ; Accepted 23 April 2016 ; Published online 20 May 2016

**Abstract** In the present study, autoregressive integrated moving average (ARIMA) methodology has been applied for modelling and forecasting of yearly cotton productivity in India. The order of the best ARIMA model was found to be (0, 1, 1). The first eight year forecast data is used with actual data and it was observed that on an average there is overall 3.66% error in forecast. It was also observed that there is increasing trend in productivity after the year 2013 which continues for subsequent year.

**Keywords** Forecasting, Time series modelling, ARIMA, Cotton productivity.

## Introduction

Cotton is the livelihood for about 60 million India including farming, textile and trade sector. With about 365 Lakhs bales of cotton production in the country for year 2012-it is the second largest in the world next to China and productivity at 500 kg/ha for the past six to seven years which is much lower to leading cotton growing countries like China. Brazil and USA. Over the years India has achieved significantly in quantitative and qualitative cotton production with available technologies. Therefore to know the status of cotton productivity this study has been undertaken. The data of 64 years data related to cotton productivity in India was used to fit the model. The data was taken from the website of Central Institute of Cotton Research, Nagpur for the year 1950-51 to 2013-14. Out of these 64 years data 56 years data was used to develop the model and remaining eight years are used to find out the average forecast error in the developed model.

In this paper, an Autoregressive Integrated Moving Average (ARIMA) model introduced by Box and Jenkins in 1960 was used to forecast cotton productivity for the leadings four years this model is also known as Box-Jenkins Model which is used to forecast a single variable. The characteristic of the ARIMA model is that it assumes non-zero autocorrelation between the successive values of the time series data.

Many attempts have been made in the past to develop forecast models for various commodities. Paul et al. [1] have studied the fluctuations in export price of spice ; Chandran and Pandey [2] have studied the seasonal fluctuation in potato price in Delhi ; Paul and Das [3] have attempted forecasting of inland fish production in India by using ARIMA approach. Paul [4] has also studied the application of stochastic modelling for forecasting of wholesale

M. Kumar*
Assistant Professor CCS HAU, COA, Kaul-Kaithal-136021, Haryana, India

Rajendra
Assistant Professor, CCS HAU COA, Kaul

R. C. Hasija
Professor Statistics, CCS HAU Hisar, India
e-mail: m25424553@gmail.com
*Correspondence

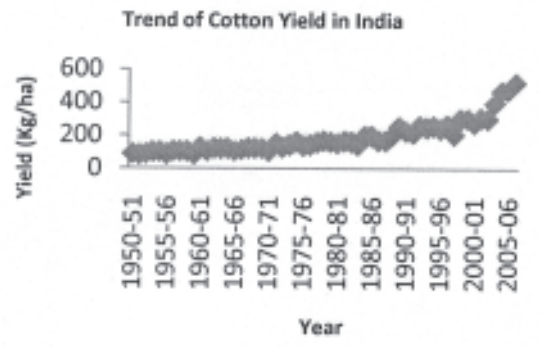


**Fig. 1.** Cotton productivity (bales) in India from 1950 to 2006.

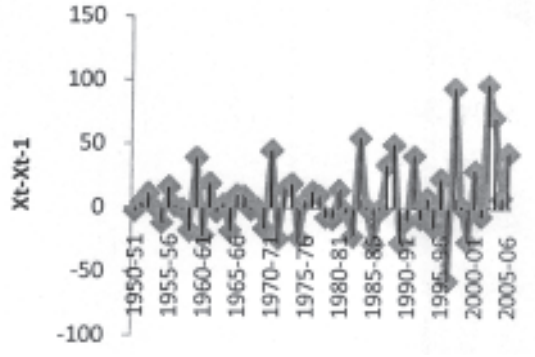


**Fig. 2.** Line plot of differenced cotton productivity in India data of first order (d =1).

price of Rohu in West Bengal, India. Contreras et al. [5] in their study, using ARIMA methodology, provided a method to predict next-day electricity prices both for spot markets and long-term contracts for mainland Spain and Californian markets.

## ARIMA model

A time series is defined as a sequence of data observed over time. ARIMA models are a class of models that have capabilities to represent stationary as well as non-stationary time series and to produce accurate forecasts based on a description of historical data of single variable. It is carried out in three stages, viz. identification, estimation and diagnostic checking. The parameters of 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. A detailed discussion on various aspects of this approach is given in Box et al. [6]. Figure 1 represents the line plot of cotton productivity in India and shows increasing trend.

## Model identification

First stage of ARIMA model building is to identify whether the variable, which is being forecasted, is stationary in time series or not. By stationary we mean, the values of variable over time varies around a constant mean and variance. Figure 1 shows that the data is not stationary, it shows an increasing trend. Therefore to build ATIMA model first make the series stationary. For this first difference the time series 'd' times to obtain a stationary series in order to have an

**Table 1**. ACF and PACF coefficients for lag 1 to 20.

| Lag | ACFY | ACFD.Y | PACFY | PACFd.Y | Lag | ACFY | ACFd.Y | PACFY | PACFd.Y |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.843 | – 0.219 | 0.843 | – 0.219 | 11 | 0.326 | 0.147 | 0.076 | 0.166 |
| 2 | 0.706 | – 0.052 | – 0.016 | – 0.105 | 12 | 0.274 | –0.100 | – 0.085 | – 0.020 |
| 3 | 0.615 | 0.041 | 0.085 | 0.006 | 13 | 0.237 | –0.138 | – 0.018 | – 0.115 |
| 4 | 0.582 | 0.015 | 0.159 | 0.022 | 14 | 0.205 | 0.196 | 0.015 | 0.142 |
| 5 | 0.540 | 0.149 | – 0.011 | 0.173 | 15 | 0.177 | 0.088 | – 0.059 | 0.173 |
| 6 | 0.497 | – 0.043 | 0.023 | 0.038 | 16 | 0.128 | – 0.049 | – 0.086 | – 0.013 |
| 7 | 0.441 | 0.055 | – 0.035 | 0.082 | 17 | 0.064 | – 0.012 | – 0.047 | – 0.025 |
| 8 | 0.377 | – 0.115 | – 0.062 | – 0.107 | 18 | 0.033 | 0.011 | 0.014 | 0.001 |
| 9 | 0.382 | – 0.038 | 0.201 | – 0.106 | 19 | 0.019 | – 0.020 | 0.063 | – 0.044 |
| 10 | 0.346 | 0.048 | – 0.144 | – 0.040 | 20 | 0.004 | 0.004 | – 0.079 | – 0.045 |

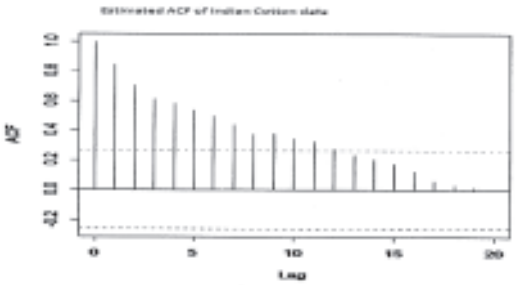

**Fig. 3.** Autocorrelations (ACF) of first differenced series by lag.

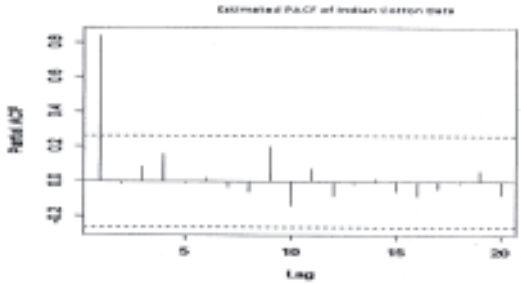

**Fig. 4.** Partial autocorrelations (PACF) of first differenced series by lag.

ARIMA (p, d, q) model with d as the order of differencing used. Caution to be taken in differencing as over differencing will tend to increase in the standard deviation, rather than a reduction. The best idea is to start with differencing with lowest order (of first order, d=1) and test the data for unit root problems. So we obtained a time series of first order differencing and Figure 2 is the line plot of the first order differenced cotton productivity in India. It can easily be inferred from the above graph (Figure 2) that the time series appears to be stationary both in its mean and variance. But before moving further, we will first test the differenced time series data for stationary (unit root problem) using augmented Dickey-Fuller test. The ADF test result, as obtained upon application, is shown below: Dickey-Fuller = – 9.5793, Lag order = 0, *p*-value = 0.01 ; Dickey-Fuller = – 3. 5375 Lag order = 3, *p* - value = 0.04652.

It fail to accept the $H_0$ and hence can conclude that the alternative hypothesis is true i.e. the series is stationary in its mean and variance. Thus, there is no need for further differencing the time series and we adopt d = 1 for our ARIMA (p, d, q) model. This test enables us to go further in steps for ARIMA model development i.e. to find suitable values of *p* in AR and q in MA in our model. For that, we need to examine the correlogram and partial correlogram of the stationary (first order differenced) time series.

## Correlogram and partial correlogram

Figure 3 represents the plot of correlogram (auto-correlation function, ACF) for lags 1 to 20 of the first order differenced time series of the cotton productivity in India. The correlogram infers that the auto-correlation at lag 1 to 11 exceed the significance limits this is due to autocorrelation in the data and rest all coefficients between lag 12 to lag 20 are well within the limits. Figure 4 represents the partial correlogram (partial auto-correlation function, PACF) for lags 1 to 20 of the differenced time series.The partial correlogram, Figure 4, shows that all the PACFs lag 1 to 20 are within the significant limits (Table 1 to 3), represents the ACF and PACF coefficients for lag 1 20 of that original and first order differenced time series.

The auto.arima () function is used in R software to find the appropriate ARIMA model. The appropri-

**Table 2.** AIC and BIC values of fitted ARIMA models.

| ARIMA Model | Coefficients: Coefficient | Coefficients: Intercept | Sigma^2 | Log likelihood | AIC | BIC | AICc |
|---|---|---|---|---|---|---|---|
| (0, 1, 1) | –0.2505 (0.1353) | 7.0374 (2.8311) | 774.4 | –261.01 | 528.01 | 534.03 | 528.48 |

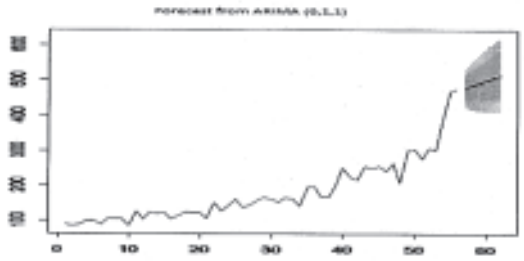

**Fig. 5.** Forecasts from ARIMA (0, 1, 1).

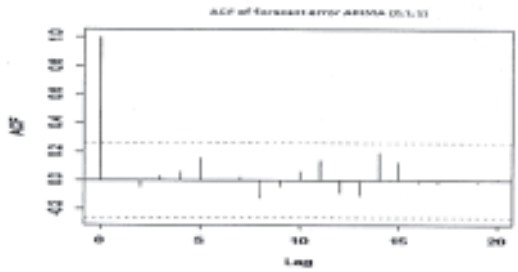

**Fig 6 (a).** Estimated ACF of residuals (Forecast Errors)–ARIMA (0, 1, 1).

ate model was found to be ARIMA (0, 1, 1), which has one parameter.

Forecasting using selected ARIMA model

The above selected model ARIMA (0, 1, 1), fitting to our time series data, means that we are fitting ARIMA (0, 1) model of first order difference to our time series. Also, ARIMA (0, 1) model, which has one parameters in it, can be rewritten an MR model of order 1, or MR (1) model, since $p$ is zero in MA. Therefore, this model can be expressed as: $Y_t = \mu + Y_{t-1} - \theta\varepsilon_{t-1}$, where $Y_t$ is the stationary time series we are studying, $\mu$ is the mean of time series. Here we shall fit the chosen ARIMA (0, 1, 1) model to forecast for the future values of our time series. Table 4 shows the forecast for the next 4 years with 80%, 95% and 99.5% (low and high) prediction intervals. In the picture 5 the two shaded zones of forecast represent the 80% and 95% (lower and upper side) projection of prediction intervals. To investigate further whether there are any correlations between successive forecast errors, we will plot the correlogram (ACF) and partial correlogram (PACF) of the forecast errors. Pictures 6 (a) and 6 (b) represents ACF and PACF of the forecast errors. Similarly ACFs, all the PACFs or partial autocorrelation coefficients of residuals of fitted ARIMA for lag 1 to lag 20 are within the significant limits. This means ACF and PACF concluded that there is non-zero autocorrelations in the forecast residuals (or standard errors) at lag 1 to 20 in the fitted ARIMA (0, 1, 1) model. The Box-Ljung test results are given below: X-squared = 11.259 df = 20, $p$-value = 0.9392 since $p$-value is greater than .05 this means that there is no autocorrelation.

**Table 3.** Four years forecasting for cotton productivity in India.

| Year | Actual | Forecast | Low 80 | %age Forecast | High 80 | %age Forecast | Low 95 | %age Fore cast | High 95 | %age Forecast |
|---|---|---|---|---|---|---|---|---|---|---|
| 2006-07 | 521 | 479.4 | 443.74 | -7.44 | 515.06 | 7.44 | 424.86 | 11.38 | 533.94 | -11.38 |
| 2007-08 | 553 | 486.44 | 441.87 | -9.16 | 531.00 | 9.16 | 418.28 | 14.01 | 554.60 | -14.01 |
| 2008-09 | 524 | 493.48 | 441.51 | -10.53 | 545.44 | 10.53 | 414.00 | 16.11 | 572.95 | -16.10 |
| 2009-10 | 502 | 500.52 | 442.07 | -11.68 | 558.95 | 11.67 | 411.14 | 17.86 | 589.89 | -17.86 |
| 2010-11 | 517 | 507.55 | 443.29 | -12.66 | 571.81 | 12.66 | 409.27 | 19.36 | 605.83 | -19.36 |
| 2011-12 | 493 | 514.59 | 444.99 | -13.53 | 584.19 | 13.53 | 408.14 | 20.69 | 621.03 | -20.68 |
| 2012-13 | 518 | 521.62 | 447.07 | -14.29 | 596.18 | 14.29 | 407.60 | 21.86 | 635.65 | -21.86 |
| 2013-14 | 565 | 528.66 | 449.46 | -14.98 | 607.86 | 14.98 | 407.53 | 22.91 | 649.79 | -22.91 |
| 2014-15 | | 535.7 | 452.11 | -15.60 | 619.29 | 15.60 | 407.86 | 23.86 | 663.79 | -23.91 |
| 2015-16 | | 542.74 | 454.98 | -16.17 | 630.50 | 16.17 | 408.52 | 24.73 | 676.45 | -24.64 |
| 2016-17 | | 549.77 | 458.03 | -16.69 | 641.50 | 16.69 | 409.47 | 25.52 | 690.08 | -25.52 |
| 2017-18 | | 556.81 | 461.26 | -17.16 | 652.36 | 17.16 | 410.67 | 26.25 | 702.95 | -26.25 |

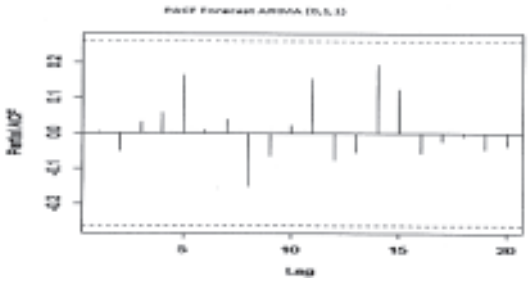

**Fig. 6 (b).** Estimated PACF of residuals–ARIMA (0, 1, 1).

## Conclusion

In this study, the ARIMA (0, 1, 1) was the best candidate model selected for making predictions for up to 4 years for the productivity of cotton in India using a 56 years' time series data. It is concluded that the selected ARIMA (0, 1, 1) seem to provide an adequate predictive model for the cotton yield in India. The ARIMA (0, 1, 1) model predicted a decrease in the productivity for the year 2006 to 2009, then a increase in year 2010 and in subsequent year upto 2017, overall an increase in production (Table 5). The prediction for 2015 is resulted approximately 542 kg per hectare (± 16% at confidence interval 80%, ± 24% at confidence interval 95% and for 2017, the prediction is approximately 556 kg per hectare (± 17% at confidence interval 80%, ± 26% at confidence interval 95%). Although, like any other predictive models in forecasting, ARIMA also has limitations on accuracy of predictions yet it is used more widely for forecasting the future successive values in the time series. It was used for the reasons of its capabilities to make predictions using a time series data with any kind of pattern and with autocorrelations between the successive values in the time series.

### References


1. Paul RK, Prajneshu, Ghosh H (2009) GARCH nonlinear time series analysis for modelling and forecasting of India's volatile spices export data. J Ind Soc Agric Stat 63 : 123—131.
2. Chandran KP, Pandey NK (2007) Potato price forecasting using seasonal ARIMA approach. Potato J 34 : 137—138.
3. Paul RK, Das MK (2010) Statistical modelling of inland fish production in India. J Inland Fish Soc India 42 : 1—7.
4. Paul RK (2010) Stochastic modeling of wholesale price of rohu in West Bengal, India. Interstat 11 : 1—9.
5. Contreras J, Espinola R, Nogales FJ, Conejo AJ (2003) ARIMA models to predict nextday electricity prices. IEEE Tran Power Systems. 18 : 1014—1020.
6. Box GEP, Jenkins GM, Reinsel GC (2007) Time-series analysis: Forecasting and control. Pearson Education, India.