



# Forecasting Sugarcane Yield of Tamilnadu Using ARIMA Models

K. K. Suresh · S. R. Krishna Priya

Received: 6 May 2010 / Accepted: 22 November 2010 / Published online: 26 May 2011  
© Society for Sugar Research & Promotion 2011

**Abstract** This paper attempts forecasting the sugarcane area, production and productivity of Tamilnadu through fitting of univariate Auto Regressive Integrated Moving Average (ARIMA) models. The data on sugarcane area, production and productivity collected from 1950–2007 has been used for present study. ARIMA (1, 1, 1) model is found suitable for sugarcane area and productivity. ARIMA (2, 1, 2) is found appropriate for modeling sugarcane production. The performances of models are validated by comparing with actual values. Using the models developed, forecast values for sugarcane area, production and productivity are developed for subsequent years.

**Keywords** Sugarcane · Forecast · Modeling · ARIMA model

## Introduction

Sugarcane, a traditional crop of India plays an important role in agricultural and industrial economy of the country. It is cultivated in most of the states and though it covers an insignificant share of about 2 and 20% of gross cropped area of our country and world respectively, its share in our country's economic growth has become significant. Among the sugarcane growing states in our country, Tamilnadu ranks first in per hectare productivity of sugarcane with 113.9/ha of cane yield. Sugarcane is a versatile crop. Because of its diversified uses in different industries, this

crop is considered as “Karpagavirucham” and in modern terminology as “wonder cane” (Mohan et al. 2007).

Sugarcane occupies a significant position among the commercial crops in India. A proper forecast of production of such important commercial crops is very important in an economic system. There is close association between crop productions with prices. An unexpected decrease in production reduces marketable surplus and income of the farmers and leads to price rise. A glut in production can lead to a slump in prices and has adverse effect on farmers' incomes. Impact on price of an essential commodity has a significant role in determining the inflation rate, wages, salaries and various policies in an economy. In case of commercial crops like sugarcane, production level affects raw material cost of user industries and their competitive advantages in the market. In our present study sugarcane area, production and productivity of Tamilnadu have been forecasted using Auto Regressive Integrated Moving Average (ARIMA) models.

Previously attempts were made to forecast sugarcane production and productivity using ARIMA models (Bajpai and Venugopalan 1996; Yaseen et al. 2005). ARIMA models have been used for modeling and forecasting of fish catches (Venugopalan and Srinath 1998; Tsitsika et al. 2007). The forecasting efficiency of ARIMA models were compared with neural network models (Hanson et al. 1999). ARIMA models have been developed to forecast the cultivable area, production and productivity of various crops of Tamilnadu (Balanagammal et al. 2000). Wheat production in Pakistan and Canada were forecasted using ARIMA models (Saeed et al. 2000; Boken 2000). ARIMA models were used to obtain seasonal forecast of paddy in Tamilnadu and food grains in India (Balasubramanian and Dhanavanthan 2002). ARIMA models were compared with structural time series models (Ravichandran and Prajneshu

---

K. K. Suresh · S. R. Krishna Priya (✉)  
Department of Statistics, Bharathiar University,  
Coimbatore, India  
e-mail: sr\_priyamadhan@yahoo.co.in

2001; Prajneshu et al. 2002). ARMA models were used in forecasting of milk, fat and protein yields of Italian Simmental cows (Maccioitta et al. 2000, 2002). Univariate forecasting of state level agricultural production was done using ARIMA models (Indira and Datta 2003). ARIMA models were compared with nonparametric regression approach for forecasting oilseed production in India (Chandran and Prajneshu 2005). Forecasting of irrigated crops like Potato, Mustard and Wheat were forecasted using ARIMA models (Sahu 2006). Milk production in India was forecasted using time-series modeling techniques (Pal et al. 2007). The objective of our present study using ARIMA models to forecast sugarcane area, production and productivity of Tamilnadu.

## Materials and Methods

The Data on sugarcane area (000' ha), production (000' tonnes) and productivity (tonnes/ha) for a period of 57 years from (1950–1951) to (2007–2008) has been collected from various volumes of 'Cooperative Sugar' (CSJ 1980, 2007) and 'Indian Sugar' (ISJ 1985, 2009) journals.

The data for a period of 55 years (1950–2006) was used in model building. The remaining 2 years data (2007–2008) was used for validation of the model.

### Description of the Model

In general, an ARIMA model is characterized by the notation ARIMA (p, d, q) where p, d, q denote orders of auto-regression, integration (differencing) and moving average respectively. In ARIMA, time series is a liner function of past actual values and random shocks. A stationary ARIMA (p, q) process is defined by the equation

$$Y_t = \Phi_0 + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \dots + \Phi_p Y_{t-p} + \varepsilon_t \omega_1 \varepsilon_{t-1} - \omega_2 \varepsilon_{t-2} \dots \omega_q \varepsilon_{t-q} \quad (1)$$

where,  $Y_t$  is the response (dependant) variable at time  $t$ .  $Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}$  is the response (dependant) variable at time lags  $t-1, t-2, \dots, t-p$  respectively; these  $Y$ 's are independent variables.  $\Phi_1, \Phi_2, \dots, \Phi_p$  is the coefficients to be estimated.  $\varepsilon_t$  is the error term at time  $t$  that represents the effects of variables not explained by the model; the assumptions about the error term are the same as those for the standard regression model.  $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q}$  is the error term that represents the effect of variables not explained by the model. The assumptions about the error term are the same as those for the standard regression model.  $\omega_1, \omega_2, \dots, \omega_q$  is the coefficients to be estimated.

## ARIMA Model Building

### Identification

The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for the stationary series. There are two kinds of, viz., stationarity in 'mean' and stationarity in 'variance'. Visual examination of graph of the data and structure of autocorrelation, and partial correlation coefficients helps to check the presence of stationarity. Another way of checking for stationarity is to fit a first order autoregressive model for the raw data and test whether the coefficient ' $\Phi_1$ ' is less than one. If the model is found to be non-stationary, stationarity is achieved by differencing the series.

If ' $X_t$ ' denotes the original series, the non-seasonal difference of first order is

$$Y_t = X_t - X_{t-1} \quad (2)$$

The next step in the identification process is to find the initial values for the orders of non-seasonal parameters, p and q. They are obtained by looking for significant autocorrelation and partial autocorrelation coefficients. There are no strict rules in choosing the initial values. Though sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients, still they are used as initial values while the final models are achieved after going through the stages repeatedly.

### Estimation

At the identification stage, one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then precise estimates of parameters of the model are obtained by least squares. Standard computer packages like SAS, SPSS etc. are available for finding the estimates of relevant parameters using iterative procedures.

### Diagnostics

Different models are obtained for various combinations of Auto Regressive and Moving Average individually and collectively. The best model is selected based on the following diagnostics:

- Low Akaike Information Criteria (AIC)
- Insignificance of auto correlations for residuals (Q-tests)
- Significance of the parameters

(a) *Low AIC*: AIC is given by  $AIC = (-2 \log L + 2m)$  where  $m = p + q$  and  $L$  is the likelihood function. Since

**Table 1** Final estimates of parameters for sugarcane area

Type	Estimate	S.E
AR1	−0.2517	0.2142
MA1	−0.745	0.14999
CONSTANT	0.0047	0.00496

**Table 2** Final estimates of parameters for sugarcane production

Type	Estimate	S.E
AR1	0.486	0.206
MA1	0.48	5.962
MA2	0.519	3.159
CONSTANT	0.591	0.039

**Table 3** Final estimates of parameters for sugarcane productivity

Type	Estimate	S.E
AR1	0.74	0.148
MA1	0.998	2.689
CONSTANT	0.613	0.151

−2 log L is approximately equal to  $\{n(1 + \log 2\pi) + n \log \sigma^2\}$  where  $\sigma^2$  is the model MSE, AIC is written as  $AIC = \{n(1 + \log 2\pi) + n \log \sigma^2 + 2m\}$  and because first term in this equation is a constant, it is omitted while comparing between models. As an alternative to AIC, sometimes SBC is also used which is given by  $SBC = \log \sigma^2 + (m \log n)/n$ .

(b) *Insignificance of auto correlations for residuals (Q-tests)*: After tentative model is fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity  $Q$  known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and is computed as follows:

$$Q = n \sum r^2(j) \tag{3}$$

where summation extends from 1 to  $k$  with  $k$  as the maximum lag considered,  $n$  is the number of observations in the series,  $r(j)$  is the estimated autocorrelation at lag  $j$ :  $k$  is a positive integer and is usually around 20.  $Q$  follows Chi-square with  $(k - m_1)$  degrees of freedom where  $m_1$  is the number of parameters estimated in the model. A modified  $Q$  statistic is the Ljung-box statistic which is given by

**Table 4** Performance of ARIMA (1, 1, 1) model for sugarcane area

Year	Observed area ('000 ha)	Predicted area ('000 ha)	% of deviation (±)
2006	0.391	0.372	4.836
2007	0.352	0.397	−12.78

**Table 5** Performance of ARIMA (2, 1, 2) model for sugarcane production

Year	Observed area ('000 tonnes)	Predicted area ('000 tonnes)	% of deviation (±)
2006	41.124	38.647	6.02
2007	38.346	35.474	7.49

**Table 6** Performance of ARIMA (1, 1, 1) model for sugarcane productivity

Year	Observed area (tones/ha)	Predicted area (tones/ha)	% of deviation (±)
2006	105.1	106.352	−1.191
2007	108.9	107.734	1.070

$$Q = n(n + 2) \sum r^2(j)/(n - j) \tag{4}$$

The  $Q$  statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and  $Q$  should be small (the probability value should be large). A significant value indicates that the chosen model does not fit well.

### Results and Discussion

The stationary check of time series revealed that the time series data on sugarcane area, production and productivity was not stationary. It was made stationary by using the first order differencing technique. For different values of  $p$  and  $q$  (0, 1 or 2), various ARIMA models were fitted and appropriate model was chosen corresponding to minimum value of the selection criterion i.e. Akaike Information Criteria (AIC) and Schwarz-Bayesian Information Criteria (SBC). In this way, ARIMA (1, 1, 1) model was found to be appropriate for sugarcane area and productivity. ARIMA (2, 1, 2) was suitable for sugarcane production. The estimates of parameters along with their standard errors have been presented in Tables 1, 2 and 3 for sugarcane area, production and productivity. After model fitting, next step is diagnostic checking of the fitted model. ACF and PACF were plotted for residuals of the fitted model. For the present study ACF and PACF were lying within the limits

**Table 7** Forecast values for the future

	Years			
	2009	2010	2011	2012
Area (million ha)	0.334	0.333	0.338	0.342
Production (million tonnes)	33.238	35.244	36.312	36.823
Productivity (tonnes/ha)	108.917	110.876	111.720	112.504

for sugarcane area, production and productivity which shows that ARIMA model fitted well.

The fitted models were validated by comparing the actual values with predicted values. The observed and predicted values for sugarcane area, production and productivity along with percentage of deviation has been presented in Tables 3, 4 and 5.

The results of Table 4 indicate that the predicted values of sugarcane area are slightly higher than the actual values. From Table 6 it could be seen that the predicted values are much closer to the observed values for sugarcane productivity.

## Conclusions

The drought of 2009 has brought home the critical need for a short-term forecasting model for agriculture sector at sub-national level, since good and bad agricultural years are not synchronous across states.

The ARIMA models developed could successfully used for forecasting sugarcane area, production and productivity of Tamilnadu for subsequent years. The forecast values for the year (2010–2012) for sugarcane area, production and productivity are presented in Table 7.

**Acknowledgments** The authors would like to acknowledge Sugarcane Breeding Institute, Coimbatore, for their support rendered in conducting this study and making possible to bring out this article.

## References

Bajpai, P.K., and R. Venugopalan. 1996. Forecasting sugarcane production by time series modeling. *Indian Journal of Sugarcane Technology* 11(1): 61–65.

Balanagammal, D., C.R. Ranganathan, and K. Sundaresan. 2000. Forecasting of agricultural scenario in Tamilnadu—A time series analysis. *Journal of Indian Society of Agricultural Statistics* 53(3): 273–286.

Balasubramanian, P., and P. Dhanavanthan. 2002. Seasonal modeling and forecasting of crop production. *Statistics and Applications* 4(2): 107–118.

Boken, V.K. 2000. Forecasting spring wheat yield using time series analysis: A case study for the Canadian prairies. *Agricultural Journal* 92(6): 1047–1053.

Chandran, K.P., and Prajneshu. 2005. Nonparametric regression with jump points methodology for describing country's oilseed yield data. *Journal of Indian Society of Agricultural Statistics* 59(2): 126–130.

CSJ. *Cooperative Sugar Journal—a monthly journal* (various volumes from 1980–2007) (Published by National Federation of Cooperative Sugar Factories Ltd. New Delhi).

Hanson, J.V., J.B. Macdonald, and R.D. Nelson. 1999. Time series prediction with genetic-algorithm designed neural networks: An empirical comparison with modern statistical models. *Computational Intelligence* 15(3): 171–184.

Indira, R., and A. Datta. 2003. Univariate forecasting of state-level agricultural production. *Economic and Political Weekly* 38: 1800–1803.

ISJ. *Indian Sugar Journal* (various volumes from 1985–2009) (Published by Indian Sugar Mills Association, New Delhi).

Maccioitta, N.P.P., A. Cappio-Borlino, and G. Pulina. 2000. Time series autoregressive integrated moving average modeling of test-day milk yields of dairy ewes. *Journal of Dairy Science* 83: 1094–1103.

Maccioitta, N.P.P., D. Vicario, G. Pulina, and A. Cappio-Borlino. 2002. Test day and lactation yield predictions in Italian simmental cows by ARMA methods. *Journal of Dairy Science* 85: 3107–3114.

Mohan, S., K. Rajendran, D. Sivam, and B. Saliha. 2007. Sugar—The wonder cane. *Co-operative Sugar* 38(10): 21–24.

Pal, S., V. Ramasubramanian, and S.C. Mehta. 2007. Statistical models for forecasting milk production in India. *Journal of Indian Society of Agricultural Statistics* 61(2): 80–83.

Prajneshu, S. Ravichandran, and S. Wadhwa. 2002. Structural time series models for describing cyclical fluctuations. *Journal of Indian Society of Agricultural Statistics* 55: 70–78.

Ravichandran, S., and Prajneshu. 2001. State space modeling versus ARIMA time-series modeling. *Journal of Indian Society of Agricultural Statistics* 54(1): 43–51.

Saeed, N., A. Saeed, M. Zakria, and T.M. Bajwa. 2000. Forecasting of wheat production in Pakistan using ARIMA models. *International Journal of Agricultural Biology* 2(4): 352–353.

Sahu, P.K. 2006. Forecasting yield behavior of potato, mustard, rice, and wheat under irrigation. *Journal of Vegetable Science* 12(1): 81–99.

Tsitsika, E.V., C.D. Maravelias, and J. Haralabous. 2007. Modeling and forecasting pelagic fish production using univariate and multivariate ARIMA models. *Fisheries Science* 73: 979–988.

Venugopalan, R., and M. Srinath. 1998. Modeling and forecasting fish catches: Comparison of regression, univariate and multivariate time series methods. *Indian Journal of Fisheries* 45(3): 227–237.

Yaseen, M., M. Zakria, Islam-ud-din-Shahzad, M. Imran Khan, and M. Aslam Javed. 2005. Modeling and forecasting the sugarcane yield of Pakistan. *International Journal of Agricultural Biology* 7(2): 180–183.