Indian Journal of Agricultural Sciences **81** (6): 520–3, June 2011

Modelling and forecasting of pigeonpea (*Cajanus cajan*) production using autoregressive integrated moving average methodology

SARIKA[1], M A IQUEBAL[2] and C CHATTOPADHYAY[3]

Indian Institute of Pulses Research, Kanpur, Uttar Pradesh 208 024

Received: 11 November 2009; Revised accepted: 10 March 2011

ABSTRACT

A study was conducted on modelling and forecasting time-series data of pigeonpea production [*Cajanus cajan* (L.) Millsp.] in India. Box-Jenkins Autoregressive Integrated Moving Average (ARIMA) time-series methodology was considered for modelling and forecasting country's pigeonpea production data (1969–70 to 2007–08). The augmented Dicky Fuller test was applied to test stationarity in data set. Root mean square error, Akaike information criterion and Bayesian information criterion were used to identify the best model. The performance of fitted model was examined using mean absolute error, mean per cent forecast error, root mean square error and Theil's inequality coefficients. ARIMA (2, 1, 0) model performed better among other models of ARIMA family for modelling as well as forecasting purpose. One and two-step ahead forecast value for 2006–07 and 2007–08 for India's pigeonpea production was computed as 2.54 and 2.53 million tonnes with standard errors 0.29 and 0.31, respectively.

Key words: Autoregressive integrated moving average model, Box-Jenkins, Forecasting, Modelling, Pigeonpea production, Time-series data

Pigeonpea [*Cajanus cajan* (L.) Millsp.] is one of the major legume (pulse) crops of the tropics and sub-tropics. Although it ranks sixth in area and production in comparison to other grain legumes such as chickpea, beans and pea, it is used in more diverse ways than the others. The most important products come from seed, and dominant among these products is *dal* made by dehulling the dry seed. It is also known to provide several benefits to the soil, in which it is grown. This crop is outstanding in the depth and lateral spread of its root system, which incidentally enables to tolerate drought. It is widely grown in the Indian subcontinent, which accounts for almost 90% of the world's crop (Nene *et al*. 1990). Maharashtra, Uttar Pradesh, Karnataka, Madhya Pradesh, Gujarat and Tamil Nadu are the major pigeonpea producing states in the country (http://dacnet.nic.in).

Forecasting the crop yield or any agricultural produce is a formidable challenge. Accurate forecasting is important to both government and industry that needs to predict future production of foodgrains. Such kind of exercise would enable the policy-makers to foresee the future requirements of pigeonpea, its import/export thereby supporting them to take appropriate measures in this regard. The forecast would thus help save much of the precious resources of our country, which otherwise might be wasted.

In many scientific or technical application, data is generated in the form of time-series, thus making time-series analysis one of the major tools in research and development. Since its inception, the univariate Box-Jenkins ARIMA approach is widely used throughout the world for different types of agricultural and industrial time-series analysis. The most significant point of this approach is that the explanatory variables in these models are the past values of the same variable. The models are constructed as a linear function of past values of the series and/or previous random shocks (or errors). It can be used when the series is stationary and there is no missing data within the time-series. Forecasts are generated under the assumption that the past history can be translated into predictions for the future. This paper aims to develop model from the observed pigeonpea data applying ARIMA methodology for uses in future forecasts.

MATERIALS AND METHODS

For the present study, India's pigeonpea time series production data from 1969–70 to 2007–08 were collected from different sources (http://www.indiaagristat.com, http://dacnet.nic.in). Data from 1969–70 to 2005–06 were used

[1,2]Scientist (e mail: aijaiswal@gmail.com, [2]e mail: jiqubal @gmail.com) Agricultural Statistics; [3]Head (e mail: chirantan_cha @hotmail.com), Division of Crop Protection

for model development and 2006–07 and 2007–08 for validation.

Autoregressive integrated moving average (ARIMA) methodology

In agricultural research, data are usually collected over time. Each observation of the observed data series, y_t was considered as a realization of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t, t \in T\}$, where $T = \{0, \pm 1, \pm 2 \pm, \ldots\}$. Standard time-series approach was applied to develop an ideal model, which adequately represented the set of realizations and also their statistical relationships in a satisfactory manner. There are number of approaches available for forecasting time-series. In our study, we applied Box-Jenkins ARIMA modelling (Kumar 1990, Hossain *et al*. 2006, Koutroumanidis *et al*. 2009), which is one of the most widely used time-series prediction methods. This method uses a systematic procedure to select an appropriate model from a rich family of ARIMA models. Such models amalgamate three types of processes, viz autoregressive (AR) of order *p*, differencing of degree *d* to make the series stationary and moving average (MA) of order *q*, and is written as ARIMA (*p, d, q*). In general, its mathematical form is represented as follows:

$$\phi_p(B)(1-B)^d Y_t = c + \theta_q(B)\varepsilon_t \qquad \ldots (1)$$

where, ϕ_p (B) and θ_q (B) are polynomials in *B* of degrees *p* and *q* respectively, *c* = constant; *B* = a backshift operator; *d* = order of difference operator; *p* = order of nonseasonal AR operator; and *q* = order of nonseasonal MA operator.

The conditions of stationarity and invertibility of the data under study were met only if all the roots of the characteristic equations $\phi_p(B)= 0$, $\theta_q(B)=0$ lied outside the unit circle.

Choice of the most appropriate values for *p, d* and *q* is major problem in ARIMA modeling technique. In our study, this problem is partially resolved by performing prediction through the following stages:

Model identification: Orders of AR and MA components were determined.

Model estimation: Linear model coefficients were estimated.

Model validation: Certain diagnostic methods were used to test the suitability of the estimated model.

Forecasting: The best model chosen was used for forecasting.

ARIMA methodology may be precisely visualized from Fig 1.

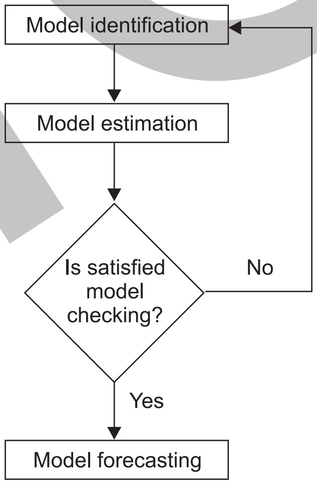

Fig 1 Stages of building ARIMA model

Testing for stationarity and estimation of parameters

Preliminary, but very important step was considered at the identification stage. This was to check whether or not the time-series under study met the condition of stationarity, since univariate ARIMA models are only applicable to stationary series (time-series with no systematic change in mean and variance). In order to test the stationarity (Pankratz 1983), autocorrelation function (ACF) of difference series up to 20 lags were computed. The series, in general is considered to be stationary if ACF for first and higher differences drop abruptly to zero, which is a heuristic approach. In our study, more statistically sound technique, viz augmented Dickey Fuller (ADF) test (Dickey and Fuller 1979) was applied to the data as such for testing the stationarity. Eviews ver. 3.0 software package was used to calculate ADF test statistic.

After identifying appropriate models in the identification stage, precise estimates of the parameters for chosen models were derived. For estimation of parameters Principles of Least Squares technique was used. Briefly, this estimation process produces new coefficients from some given initial values of coefficients in order to minimize residual sum of square. In the present investigation, SPSS, ver 16.0 software package was used. Total number of iterations considered was 100 with convergence criterion 0.001%.

Residual analysis and evaluation of forecast error

Diagnostic stage statistically determines adequacy of the fitted model. It is also necessary to ascertain whether or not the assumption of independence of the white noise residuals is met. If a model is an adequate representation of a time-series, it should capture all the correlation in the series, and the white noise residuals should be independent of each other. Thus, any significant autocorrelation shown in the estimated white noise residuals at the ACF and/or partial autocorrelation function (PACF) indicates model inadequacy and suggests the model modification. With this concept, the residual analysis in our study was carried out through autocorrelation function, partial autocorrelation function and Box-Ljung test (Box *et al.* 1994). To test the randomness of errors, residual analysis was also carried out using run test (Gujarati 2003).

There are many statistics available in literature for evaluating forecast error of any model. We often do not compute all the statistics because one of them is the function of other. In our study, comparison of the forecasting performance was done using statistics such as mean absolute errors (MAE), root mean square error (RMSE), mean per cent forecast error (MPFE) and Theil's inequality coefficients (TIC).

RESULTS AND DISCUSSION

Box-Jenkins ARIMA methodology resolved the problem of deciding appropriate values for *p*, *d* and *q* partially by following the steps described earlier. The preliminary step for fitting ARIMA model started with the stationarity test. The computed ADF test statistic 'tau' (–0.009891) was found to be greater than critical values (–2.6280, –1.9504, –1.6206) at 1%, 5% and 10% significant level respectively, leading to the acceptance of null hypothesis (here H_0: Data set is non-stationary). Hence, the production data series was non-stationary. After taking the first difference, ADF test statistic 'tau' (–9.008) came to be smaller than critical values (–2.6280, –1.9504, –1.6206) at 1, 5 and 10% significant level respectively, which made the series stationary. Hence, the value of *d* was assumed to be 1.

The following step was to choose the most appropriate values for *p* and *q*. This problem was partially overcome by looking at ACF and PACF for the series. Coefficients of selected model were estimated. On the basis of minimum RMSE, AIC and BIC criteria, the ARIMA (2, 1, 0) model was selected. Graph of fitted ARIMA model along with data points and error series indicated that model fits well to the country's pigeonpea production data (Fig 2). The RMSE, AIC and BIC values were computed as 0.28, 14.57 and 19.31, respectively. Parameter estimates of the model with standard errors and significant values are reported in Table 1. The fitted ARIMA (2, 1, 0) model represented as:

$$\hat{Y}_t - Y_{t\text{-}1} = 0.021 - 0.629(Y_{t-1} - Y_{t-2}) - 0.476(Y_{t\text{-}2} - Y_{t-3}) \quad \ldots (2)$$

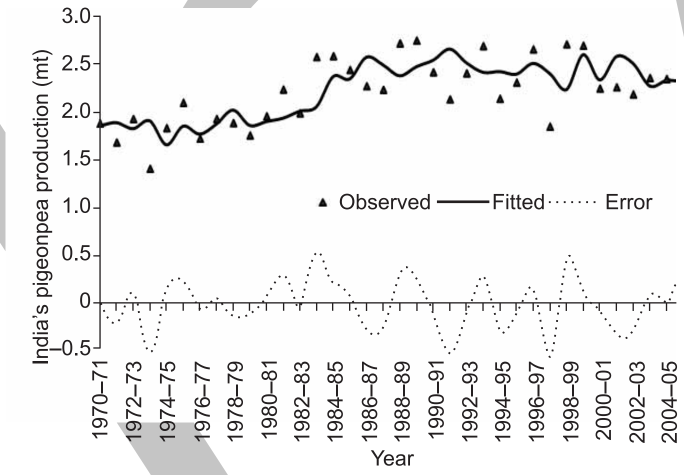

Fig 2 Fitted ARIMA (2, 1, 0) model along with actual data and error series

Table 1 Parameter estimates along with standard errors and significant values

Parameter	Parameter estimate	Standard error	$\|t\|$ -value	Aprox. sig
c (constant)	0.021	0.023	0.919	0.365
ϕ_1	–0.629	0.153	4.111	0.000
ϕ_2	–0.476	0.153	3.114	0.004

Table 2 Residual analysis along with calculated Box-Ljung statistics

Lag	Autocorrelation	Standard error	Box-Ljung statistics		
			Value	Df	Sig[b]
1	–0.056	0.160	0.123	1	0.726
2	–0.160	0.158	1.155	2	0.561
3	–0.152	0.160	2.115	3	0.549
4	–0.115	0.153	2.681	4	0.613
5	–0.056	0.151	4.634	5	0.462
6	0.064	0.148	4.823	6	0.567
7	0.020	0.146	4.842	7	0.679
8	–0.165	0.143	6.176	8	0.627
9	0.023	0.140	6.203	9	0.719
10	0.019	0.138	6.221	10	0.796
11	–0.066	0.135	6.461	11	0.841
12	–0.129	0.132	7.410	12	0.829
13	–0.040	0.130	7.504	13	0.874
14	–0.095	0.127	8.070	14	0.886
15	0.142	0.124	9.388	15	0.856
16	–0.037	0.121	9.483	16	0.892

[a]The underlying process assumed is independent (white noise)
[b]Based on the asymptotic chi-square approximation

The Box-Ljung statistic reported insignificant values which were consistent with the hypothesis that residuals are random (Table 2). The residual ACF and PACF showed no significant values (Fig 3). Coupled with the results from residual ACF and PACF plots, it was concluded that the assumption of independence of error terms was not violated.

The 95% confidence interval for runs was obtained as (13.17, 24.72). Since the number of runs (21) computed for run test was in this interval, the null hypothesis that the residuals are random was accepted at 5% level of significance.

Goodness of fit measures, viz mean absolute error (MAE),

Table 3 Various measures of goodness of fit

Measures	Calculated value
Mean absolute error	0.22
Mean per cent forecast error	–0.01
Root mean square error	0.27
Theil's inequality coefficients	0.06

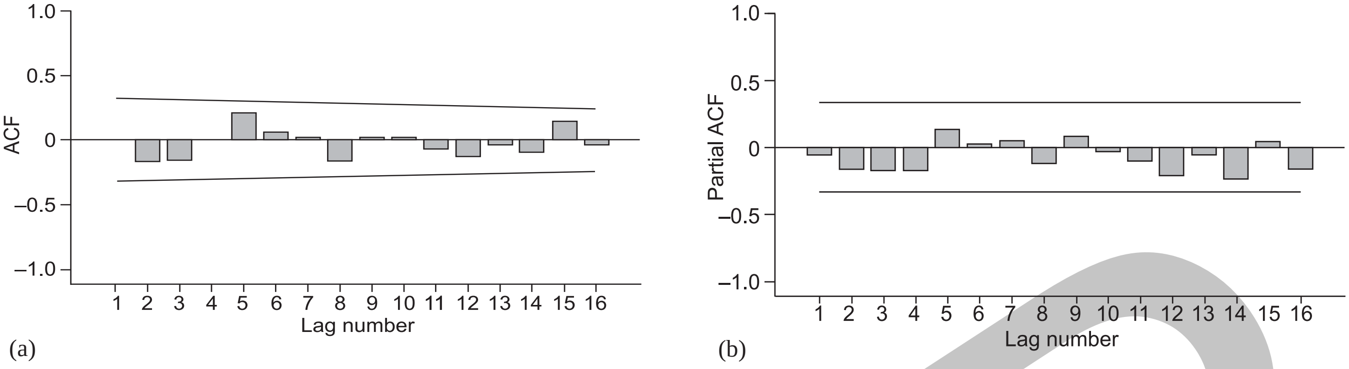

Fig 3 Residual (*a*) Autocorrelation (ACF) and (*b*) Partial Autocorrelation (PACF) plots for ARIMA (2, 1, 0)

mean per cent forecast error (MPFE), root mean square error (RMSE) and Theil's inequality coefficients (TIC) were computed (Table 3) which indicated that the ARIMA (2, 1, 0) model provided a good fit to data taken under study.

After an appropriate time-series model was decided, its unknown parameters estimated and it was established that the model fitted well; forecasting future values of the series were taken ahead. Forecast value obtained for Y_{t+1} was used further to obtain forecast for Y_{t+2} and then these two forecasts might be used to generate forecast for Y_{t+3}. For our study, we considered one and two-step ahead forecast values for 2006–07 and 2007–08 for India's pigeonpea production. These values were computed as 2.54 and 2.53 million tonnes, respectively with standard errors 0.29 and 0.31, which are quite close to their actual values (2.31 and 2.90 million tonnes). 95% confidence interval (upper and lower) for the 2006–07 and 2007–08 was obtained as (1.96, 3.12) and (1.90, 3.15), respectively. The process might be continued to obtain forecast to any point further. Since uncertainty increases as prediction is made further from the data we have, the standard errors associated with predictions increases. Thus, it is advisable to use ARIMA methodology for short-term forecast.

REFERENCES

Agricultural Statistics at a Glance. 2008. http://dacnet.nic.in

Box G E P, Jenkins G M and Reinsel G C. 1994. Time Aeries Analysis: Forecasting and Control, 3rd edn. pp 21–83. Prentice Hall, USA.

Dickey D A and Fuller W A. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Jounal of American Statistical Association* **74**: 427–31.

Gujarati D N. 2003. *Basic Econometrics*, 4th edn. pp 465–7. McGraw-Hill Companies Inc., New York.

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** (4): 344–53.

Koutroumanidis T, Ioannou K and Arabatzis G. 2009. Predicting fuelwood prices in Greece with the use of ARIMA models, artificial neural networks and hybrid ARIMA-ANN model. Energy Policy **37**: 3627–34.

Kumar Kuldeep. 1990. Some recent developments in time series analysis. *Singapore Journal of Statistics* **1**: 45–73.

Nene Y L, Hall S D and Sheila V K. 1990. *The Pigeonpea*, pp 1–14. CAB International, UK.

Pankratz A.1983. *Forecasting with Univariate Box-Jenkins Models–Concepts and Cases*, pp 119–54. John Willey, New York.