FORECASTING SUGARCANE PRODUCTION IN INDIA WITH ARIMA MODEL

[1]B. N. MANDAL

Abstract: Yearly sugarcane production data for the period of 1950-51 to 2002-03 of India were analyzed by time-series methods. Autocorrelation and partial autocorrelation functions were calculated for the data. Appropriate Box-Jenkins autoregressive integrated moving average model was fitted. Validity of the model was tested using standard statistical techniques. The forecasting power of autoregressive integrated moving average model was used to forecast sugarcane production for three leading years.
Key words: ACF = autocorrelation function, **ARIMA** = autoregressive integrated moving average, **ARMA** = autoregressive moving average, **PACF** = partial autocorrelation function, sugarcane.

Introduction:

Autoregressive Integrated Moving Average (ARIMA) model was introduced by Box and Jenkins (hence also known as Box-Jenkins model) in 1960s for forecasting a variable. An effort is made in this paper to develop an ARIMA model for sugarcane production in India and to apply the same in forecasting sugarcane production for the three leading years.

ARIMA method is an extrapolation method for forecasting and, like any other such method, it requires only the historical time series data on the variable under forecasting. Among the extrapolation methods, this is one of the most sophisticated methods, for it incorporates the features of all such methods, does not require the investigator to choose the initial values of any variable and values of various parameters a priori and it is robust to handle any data pattern. As one would expect, this is quite a difficult model to develop and apply as it involves transformation of the variable, identification of the model, estimation through non-linear method, verification of the model and derivation of forecasts. In what follows, we first explain the ARIMA model, then develop the same for sugarcane production using yearly data for India during 1950-1951 to 2002-2003 and finally apply the same to forecast the values of the variable during the future 3 years.

Theoretical Basis of Time-Series Analysis:

A time series is a set of values of a continuous variable Y (Y_1, Y_2, ...,Y_n), ordered according to a discrete index variable t (1, 2, ..., n). The term time-series comes from econometric studies in which the index variable refers to intervals of time measured in a suitable scale. However, it must be clearly stated that this direct reference to time is not required: actually, any different meaning can be attributed to the index variable, provided

[1] PhD scholar, IASRI, New Delhi-12,
mandal_stat@rediffmail.com

that it is able to order the *Y* values. In general, in a given time series the following can be recognized and separated (3):
1) a regular, long-term component of variability, termed trend, that represents the whole evolution pattern of the series;
2) a regular, short-term component whose shape occurs periodically at intervals of *s* lags of the index variable, currently known as seasonality, because this term is also derived by applications in economics;
3) an *AR(p)* autoregressive component of *p* order, which relates each value $Z_t = Y_t$ – (trend and seasonality) to the *p* previous *Z* values, according to the following linear relationship

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + \varepsilon_t \quad (1)$$

where ϕ_i (i= 1, ..., p) are parameters to be estimated and ε_t is a residual term; and

4) a *MA(q)* moving average component of *q* order, which relates each Z_t value to the *q* residuals of the *q* previous *Z* estimates

$$Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... \theta_q \varepsilon_{t-q} \quad (2)$$

where θ_i (i= 1, ..., q) are parameters to be estimated. The theory of time-series analysis has developed a specific language and a set of linear operators. According to Box and Jenkins (1), a highly useful operator in time-series theory is the lag or backward linear operator (B) defined by $BZ_t = Z_{t-1}$
Consider the result of applying the lag operator twice to a series:

$$B(BZ_t) = BZ_{t-1} = Z_{t-2}$$

Such a double application is indicated by B^2, and, in general, for any integer *k*, it can be written

$$B^k Z_t = Z_{t-k}$$

By using the backward operator, Equation [1] can be rewritten as

$$Z_t - \phi_1 Z_{t-1} - \phi_2 Z_{t-2} - ... - \phi_p Z_{t-p} = \varepsilon_t = \phi(B) Z_t \quad (3)$$

where ϕ(B) is the autoregressive operator of *p* order defined by

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$$

Similarly, Equation [2] can be written as

$$Z_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... \theta_q \varepsilon_{t-q} = \theta(B) \varepsilon_t \quad (4)$$

where $\theta(B)$ indicates the moving average operator of q order defined by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$$

The autoregressive and moving average components can be combined in an autoregressive moving average (ARMA) (p, q) model

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

or in lag operator form

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Z_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) \varepsilon_t .$$

Finally,

$$\phi(B) Z_t = \theta(B) \varepsilon_t \qquad (5)$$

In a preliminary analysis of a series it is useful to independently evaluate the long- and short-term periodic components, which are essential to define the regular structure of the series. The trend component can be evaluated by fitting a regular function, a polynomial, or a more complicated general function. The seasonal component can be estimated by a seasonal decomposition procedure, which calculates a seasonal index based on the ratio of the observed values to the moving average. In the final stage of series modeling, however, both the trend and the seasonal component will be integrated in the ARMA (p, q) process (1). For the trend, such an integration is obtained by using the difference linear operator (∇), defined by

$$\nabla Y_t = Y_t - Y_{t-1} = Y_t - B Y_t = (1 - B) Y_t$$

A single application of the ∇ operator corrects the data for a linear increasing trend, whereas its repeated use for d times corrects for a trend that can be fitted by a d-order polynomial. The stationary series Z_t obtained as the dth difference (∇^d) of Y_t,

$$Z_t = \nabla^d Y_t = (1 - B)^d Y_t$$

can be then modeled by an ARMA (p,q) process. The combined use of the ∇ operator and the ARMA (p, q) process results in an ARIMA (p, d, q) model. Furthermore, ARIMA can account for the seasonal component of s lag period, by using both correlations between Z_t and Z_{t-s} values and those between the corresponding residuals ε_t and ε_{t-s} . In mathematical terms, therefore, a seasonal ARIMA model is an ARIMA (p,d,q) model whose residuals ε_t can be further modeled by an ARIMA$(P,D,Q)_s$ structure with linear operators (P,D,Q) being functions of the B^s operator.

The operators of a seasonal ARIMA model, defined as $(p,d,q) \times (P,D,Q)_s$, can be expressed as follows:

AR (p) nonseasonal operator of p order, $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p$;

AR (P) seasonal operator of P order, $\phi(B) = 1 - \phi_1 B^s - ... - \phi_p B^{sP}$;

MA (q) nonseasonal operator of q order, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q$;

MA (Q) seasonal operator of Q order, $\theta(B) = 1 - \theta_1 B^s - \theta_2 B^{2s} - ... - \theta_Q B^{Qs}$; and

difference operator of d order, $\nabla^d = (1 - B)^d$.

The Box-Jenkins methodology for analyzing and modeling time series is characterized by three steps:
1) Model identification, 2) parameter estimation, and 3) model validation.

Model identification defines the (p, d, q) orders of the AR and MA components, both seasonal and nonseasonal. In this step, fundamental analytical tool is the autocorrelation functions.

The autocorrelation function (ACF) and partial ACF (PACF) are very important for the definition of the internal structure of the analyzed series. The ACF ρ (k) at lag k of the Z_t series is the linear correlation coefficient between Z_t and Z_{t-k}, calculated for k =0, 1, 2...

$$\rho_k = \frac{C\,\mathrm{ov}(Z_t, Z_{t-k})}{\sqrt{Var(Z_t)Var(Z_{t-k})}}$$

The PACF is defined as the linear correlation between Z_t and Z_{t-k}, controlling for possible effects of linear relationships among values at intermediate lags. Theoretically, both an AR (p) process and an MA (q) process should be associated with well-defined patterns of ACF and PACF, usually decreasing exponential or alternate in sign or decreasing sinusoidal patterns. A precise correspondence between ARMA (p, q) processes and defined ACF and PACF patterns is more difficult to recognize. When the order of at least one of the two components (AR or MA) is clearly detectable, however, the other can be identified by attempts in the following step of parameter estimation. Finally, the existence of a seasonal component of length s is underlined by the presence of a periodic pattern of period s in the ACF.

Once a suitable ARIMA (p, d, q) $\times (P,D,Q)_s$ structure is identified, subsequent steps of parameter estimation and model validation must be performed. Parameter estimates are usually obtained by maximum likelihood, which is asymptotically correct for time series. Estimators are usually sufficient, efficient, and consistent for Gaussian distributions and are asymptotically normal and efficient for several non-Gaussian distribution families.

Validation of the goodness of fit of an ARIMA model can be developed according to the following steps:
1) Evaluation of statistical significance of parameters by the usual comparison between the parameter value and the standard deviation of its estimate. For a test statistic that is

valid only asymptotically, a parameter whose value exceeds twice its standard error can be considered significant.

2) Analysis of the ACF of residuals. In this step, residuals (ε_t) are considered as a new time series, and ACF and PACF are estimated to be sure that values at lag $k \geq 0$ are not statistically different from zero.

For prediction purposes, ARIMA models are different from the analytical functions of time: $Z_t = f(t)$, because ARIMA forecasting uses previous values of the series and errors in the previous estimates. Actually, this peculiarity of ARIMA forecasting is valid in the short term because parameters of the model cannot account, in the long term, for changes in the dynamics of the series.

Building ARIMA model for sugarcane production data and Forecasting:

To fit an ARIMA model requires a sufficiently large data set. In this study, we used the data for sugarcane production for the period 1950-51 to 2002-2003. As we have earlier stated that development of ARIMA model for any variable involves three steps: identification, estimation and verification.

Each of these three steps is now explained for sugarcane production.

Year	Sugarcane production (million tonnes)	Year	Sugarcane production (million tonnes)
1950-51	57.05	1977-78	176.97
1951-52	61.63	1978-79	151.66
1952-53	51	1979-80	128.83
1953-54	44.41	1980-81	154.25
1954-55	58.74	1981-82	186.36
1955-56	60.54	1982-83	189.51
1956-57	69.05	1983-84	174.08
1957-58	71.16	1984-85	170.32
1958-59	73.36	1985-86	170.65
1959-60	77.82	1986-87	186.09
1960-61	110	1987-88	196.74
1961-62	103.97	1988-89	203.04
1962-63	91.91	1989-90	225.57
1963-64	104.23	1990-91	241.05
1964-65	121.91	1991-92	254
1965-66	123.99	1992-93	228.03
1966-67	92.83	1993-94	229.66
1967-68	95.5	1994-95	275.54
1968-69	124.68	1995-96	281.1

1969-70	135.02	1996-97	277.56
1970-71	126.37	1997-98	279.54
1971-72	113.57	1998-99	288.72
1972-73	124.87	1999-00	299.32
1973-74	140.81	2000-01	295.96
1974-75	144.29	2001-02	297.21
1975-76	140.6	2002-03	281.58
1976-77	153.01		

Model identification:

ARIMA model is estimated only after transforming the variable under forecasting into a stationary series. The stationary series is the one whose values vary over time only around a constant mean and constant variance. There are several ways to ascertain this. The most common method is to check stationarity through examining the graph or time plot of the data. Fig1 reveals that the data is nonstationary. Non-stationarity in mean is corrected through appropriate differencing of the data. In this case difference of order 1 was sufficient to achieve stationarity in mean.

The newly constructed variable X_t can now be examined for stationarity. The graph of X_t was stationary in mean. The next step is to identify the values of p and q. For this, the autocorrelation and partial autocorrelation coefficients of various orders of X_t are computed (Table 2). The ACF and PACF (fig 2 and 3) shows that the order of p and q can at most be 1. We entertained three tentative ARIMA models and chose that model which has minimum AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). The models and corresponding AIC and BIC values are

ARIMA (*p, d, q)*	AIC	BIC
210	409.759	415.612
211	410.289	418.094
212	412.358	422.114

So the most suitable model is ARIMA (2, 1, 0) as this model has the lowest AIC and BIC values.

Model estimation and verification:

Model parameters were estimated using SPSS package. Results of estimation are reported in table 3. The model verification is concerned with checking the residuals of the model to see if they contain any systematic pattern which still can be removed to improve on the chosen ARIMA. This is done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, the various correlations upto 14 lags were computed and the same along with their significance which is tested by Box-Ljung test are provided in table 4. As the results indicate, none of these correlations is significantly different from zero at a reasonable level. This proves that the selected ARIMA model is an appropriate model. The ACF and PACF of the

residuals (fig 4 and 5) also indicate 'good fit' of the model. So the fitted ARIMA model for the sugarcane data is

$$Z_t = 4.6022 + 1.1209Z_{t-1} - .7630Z_{t-2} + .6421Z_{t-3} + \varepsilon_t \qquad (6)$$

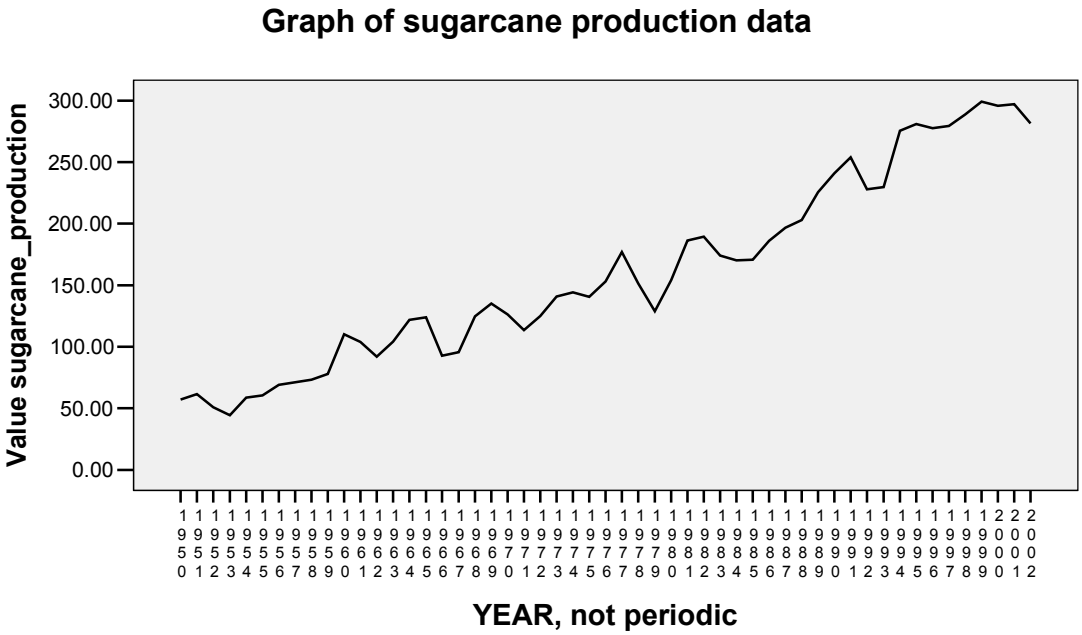

Figure 1: Time plot of sugarcane production data

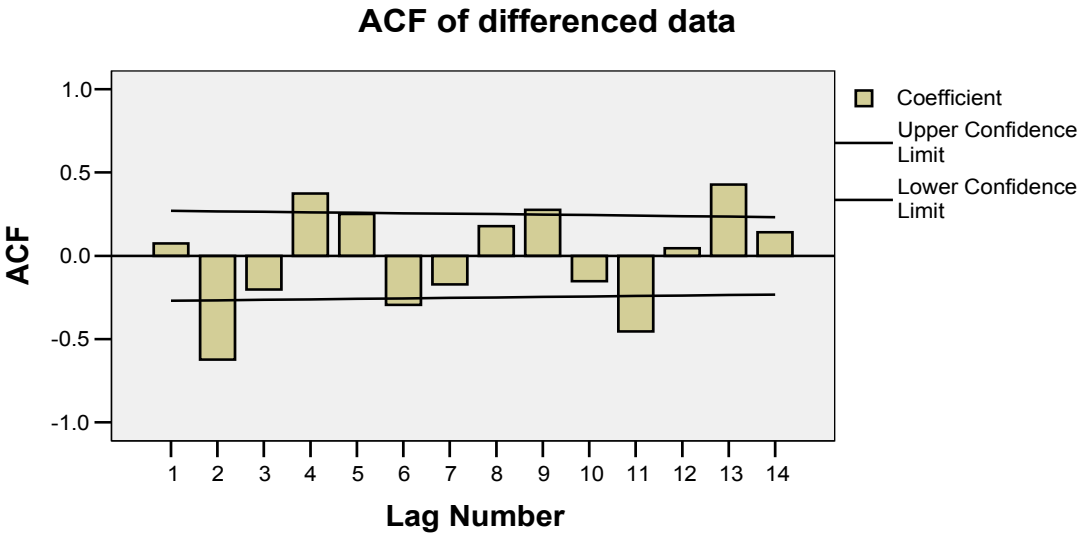

Figure 2: ACF of differenced data

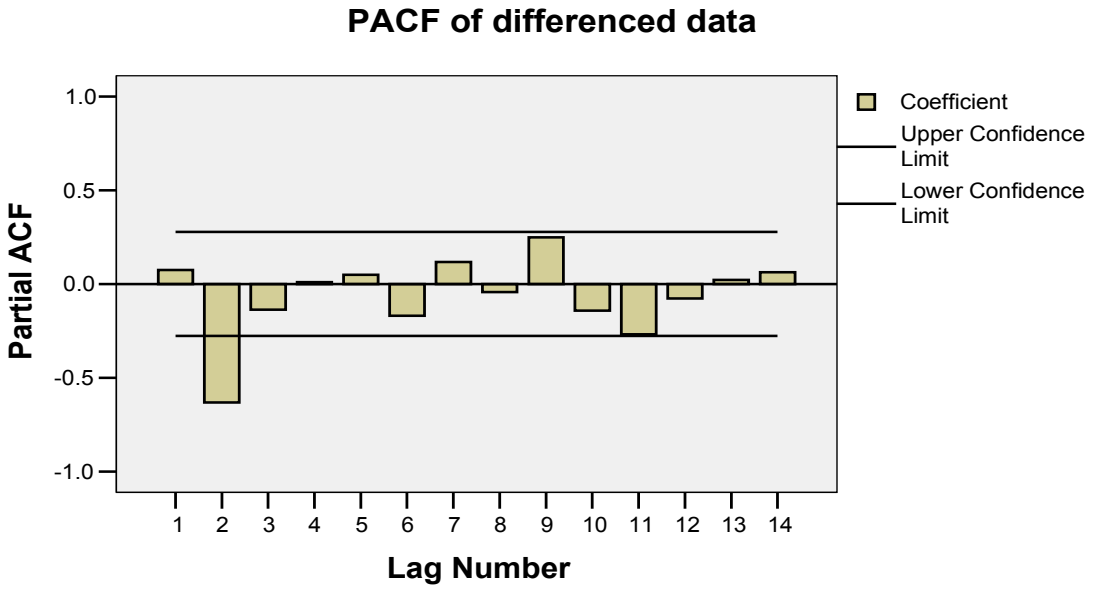

Figure 3: PACF of differenced sugarcane data

Lag	Autocorrelation	Std.Error	Lag	Partial Autocorrelation	Std.Error
1	0.074	0.135	1	0.074	0.139
2	-0.623	0.133	2	-0.632	0.139
3	-0.203	0.132	3	-0.137	0.139
4	0.373	0.131	4	0.010	0.139
5	0.252	0.129	5	0.049	0.139
6	-0.295	0.128	6	-0.169	0.139
7	-0.172	0.127	7	0.117	0.139
8	0.178	0.125	8	-0.043	0.139
9	0.276	0.124	9	0.249	0.139
10	-0.153	0.122	10	-0.143	0.139
11	-0.454	0.121	11	-0.268	0.139
12	0.045	0.119	12	-0.077	0.139
13	0.427	0.118	13	0.022	0.139
14	0.141	0.116	14	0.062	0.139

Table 2: Autocorrelations and partial autocorrelations

		Estimates	Std Error	t	Approx Sig
Non-Seasonal Lags	AR1	0.1209	0.1093	1.1055	0.2743
	AR2	-0.6421	0.1087	-5.9069	0.0000
Constant		4.6022	1.1076	4.1551	0.0001
Number of Residuals	52				
Number of Parameters	2				

Residual df	49
Adjusted Residual Sum of Squares	7166.782
Residual Sum of Squares	7168.028
Residual Variance	143.286
Model Std. Error	11.97021
Log-Likelihood	-201.879
Akaike's Information Criterion (AIC)	409.7587
Schwarz's Bayesian Criterion (BIC)	415.6125

Table 3: Estimates of the fitted ARIMA model

Lag	Autocorrelation	Std.Error	Box-Ljung Statistic Value	df	Sig.
1	-0.0898	0.1348	0.4438	1.0000	0.5053
2	-0.0156	0.1334	0.4575	2.0000	0.7955
3	-0.0524	0.1321	0.6149	3.0000	0.8930
4	-0.1268	0.1307	1.5559	4.0000	0.8167
5	0.1815	0.1294	3.5250	5.0000	0.6196
6	-0.1441	0.1280	4.7930	6.0000	0.5706
7	0.1967	0.1266	7.2080	7.0000	0.4075
8	-0.0612	0.1252	7.4474	8.0000	0.4892
9	-0.0335	0.1237	7.5208	9.0000	0.5831
10	-0.0292	0.1223	7.5780	10.0000	0.6700
11	-0.2872	0.1208	13.2276	11.0000	0.2787
12	0.0616	0.1194	13.4940	12.0000	0.3342
13	0.0388	0.1179	13.6022	13.0000	0.4024
14	0.2038	0.1163	16.6725	14.0000	0.2741

Lag	Partial Autocorrelation	Std.Error
1	-0.0898	0.1387
2	-0.0239	0.1387
3	-0.0565	0.1387
4	-0.1389	0.1387
5	0.1583	0.1387
6	-0.1309	0.1387
7	0.1831	0.1387
8	-0.0518	0.1387
9	0.0062	0.1387
10	-0.0870	0.1387
11	-0.2323	0.1387

12	-0.0695	0.1387
13	0.0737	0.1387
14	0.1628	0.1387

Table 4: Autocorrelation and partial autocorrelations of residuals

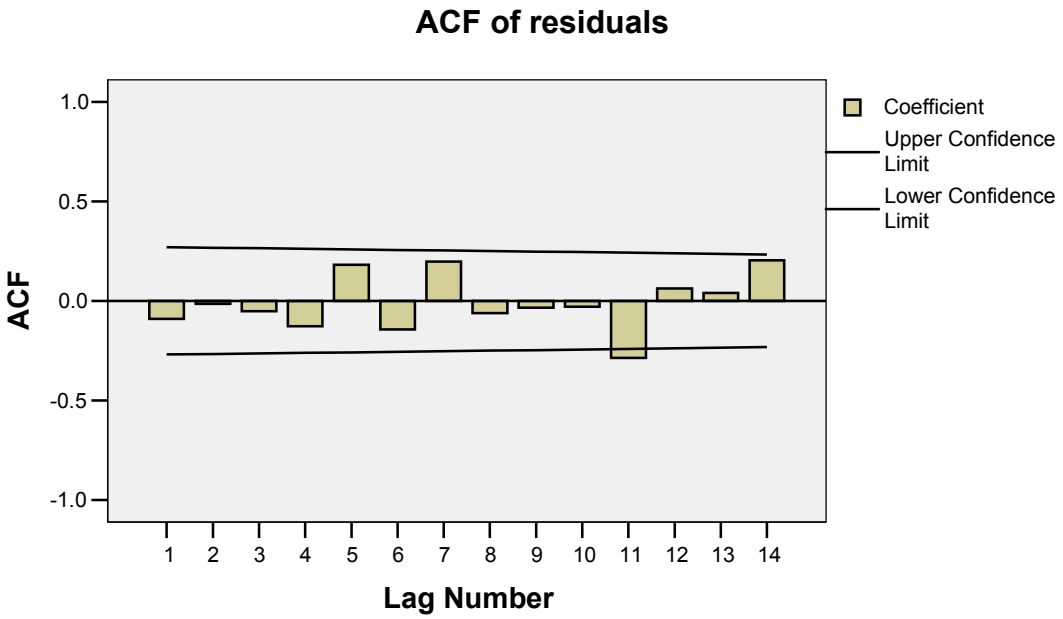

Figure 4: ACF of residuals of fitted ARIMA model

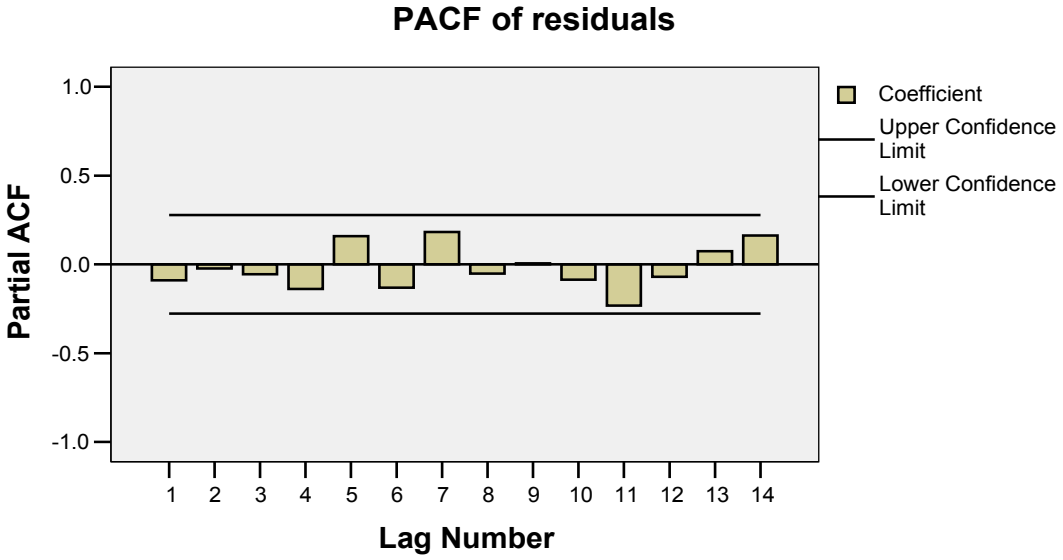

Figure 5: PACF of residuals of fitted ARIMA model

Forecasting with ARIMA model:

ARIMA models are developed basically to forecast the corresponding variable. There are two kinds of forecasts: sample period forecasts and post-sample period forecasts. The former are used to develop confidence in the model and the latter to generate genuine forecasts for use in planning and other purposes. The ARIMA model can be used to yield both these kinds of forecasts.

Sample period forecasts:

The sample period forecasts are obtained simply by plugging the actual values of the explanatory variables in the estimated equation (6).The explanatory variables here are the lagged values of Z_t and the estimated lagged errors. The so obtained values for $\widehat{Z}_t$ together with the actual values of Z_t are shown in table 5.

Year	Actual production (Million tonnes)	Estimated production (Million tonnes)	Residual	Lower CL	Upper CL
1950-51	57.05	.	.	.	.
1951-52	61.63	61.652	-0.022	30.110	93.195
1952-53	51	66.231	-15.231	34.784	97.677
1953-54	44.41	53.775	-9.365	29.483	78.067
1954-55	58.74	57.440	1.300	33.148	81.732
1955-56	60.54	71.705	-11.165	47.413	95.997
1956-57	69.05	58.557	10.493	34.265	82.849
1957-58	71.16	75.924	-4.764	51.632	100.216
1958-59	73.36	72.952	0.408	48.660	97.244
1959-60	77.82	79.272	-1.452	54.980	103.564
1960-61	110	83.948	26.052	59.655	108.240
1961-62	103.97	118.027	-14.057	93.735	142.319
1962-63	91.91	89.579	2.331	65.287	113.871
1963-64	104.23	101.325	2.905	77.033	125.617
1964-65	121.91	120.464	1.446	96.172	144.756
1965-66	123.99	123.137	0.853	98.845	147.429
1966-67	92.83	119.890	-27.060	95.598	144.182
1967-68	95.5	94.729	0.771	70.437	119.021
1968-69	124.68	122.832	1.848	98.540	147.124
1969-70	135.02	133.494	1.526	109.202	157.786
1970-71	126.37	124.534	1.836	100.242	148.826
1971-72	113.57	125.686	-12.116	101.394	149.978
1972-73	124.87	124.578	0.292	100.286	148.870
1973-74	140.81	141.456	-0.646	117.164	165.748
1974-75	144.29	142.482	1.808	118.190	166.774
1975-76	140.6	141.476	-0.876	117.184	165.769
1976-77	153.01	144.920	8.090	120.628	169.213
1977-78	176.97	163.881	13.089	139.588	188.173
1978-79	151.66	178.899	-27.239	154.606	203.191
1979-80	128.83	140.217	-11.387	115.924	164.509
1980-81	154.25	149.323	4.927	125.031	173.616

1981-82	186.36	178.983	7.377	154.691	203.275
1982-83	189.51	180.920	8.590	156.628	205.212
1983-84	174.08	176.273	-2.193	151.981	200.566
1984-85	170.32	177.193	-6.873	152.901	201.485
1985-86	170.65	186.774	-16.124	162.482	211.067
1986-87	186.09	180.105	5.985	155.813	204.397
1987-88	196.74	194.745	1.995	170.453	219.038
1988-89	203.04	195.114	7.926	170.822	219.406
1989-90	225.57	203.964	21.606	179.672	228.256
1990-91	241.05	231.249	9.801	206.957	255.541
1991-92	254	235.455	18.545	211.163	259.747
1992-93	228.03	252.626	-24.596	228.334	276.919
1993-94	229.66	223.577	6.083	199.284	247.869
1994-95	275.54	253.534	22.006	229.242	277.826
1995-96	281.1	287.040	-5.940	262.748	311.332
1996-97	277.56	259.313	18.247	235.021	283.605
1997-98	279.54	280.563	-1.023	256.271	304.855
1998-99	288.72	289.053	-0.333	264.761	313.346
1999-00	299.32	295.559	3.761	271.267	319.851
2000-01	295.96	301.708	-5.748	277.416	326.000
2001-02	297.21	295.748	1.462	271.456	320.041
2002-03	281.58	306.520	-24.940	282.227	330.812
2003-04	.	285.889	.	261.597	310.181
2004-05	.	303.447	.	266.607	340.288
2005-06	.	309.804	.	270.755	348.853

Table 5: Actual and estimated values of sugarcane production and 95% confidence limit (CL)

To judge the forecasting ability of the fitted ARIMA model, important measures of the sample period forecasts' accuracy were computed. The Mean Absolute Percentage Error (MAPE) for sugarcane production turns out to be 6.264337. This measure indicates that the forecasting inaccuracy is low.

Post sample forecasts:

The principal objective of developing an ARIMA model for a variable is to generate post sample period forecasts for that variable. This is done through using equation (6). The forecasts for sugarcane production during 2003 to 2005 are given in lower part of table 5.

Conclusions:

ARIMA model offers a good technique for predicting the magnitude of any variable. Its strength lies in the fact that the method is suitable for any time series with any pattern of change and it does not require the forecaster to choose a priori the value of any parameter. Its limitations include its requirement of a long time series. Often it is called a

'Black Box' model. Like any other method, this technique also does not guarantee perfect forecasts. Nevertheless, it can be successfully used for forecasting long time series data.

In our study the developed model for sugarcane production was found to be ARIMA (2, 1, 0). From the forecast available by using the developed model, it can be seen that forecasted production for the year 2003-04 is lower than 2002-03 but in later years the production increases. The validity of the forecasted values can be checked when the data for the lead periods become available. The model can be used by researchers for forecasting of sugarcane production in India. However, it should be updated from time to time with incorporation of current data.

References:

1) Box, G.E.P., and G. M. Jenkins. 1970. Time series analysis: forecasting and control. Holden Day, San Francisco, CA.

2) Brockwell, P.J., and Davis, R. A. 1996. Introduction to time series and forecasting. Springer.

3) Kendall, M. G., and A. Stuart. 1966. The advanced theory of statistics. Vol. 3. Design and Analysis and Time-Series. Charles Griffin & Co. Ltd., London, United Kingdom.