Agriculture for Sustainable Development 2(1):82-86, 2014/Mini review ***ISSN 2347-5358 (Print)/ 2349-2228 (Online)***

Agriculture for Sustainable Development

http://www.sustainableagriculture.in

FORECASTING FOR SOYBEAN PRODUCTION IN INDIA USING SEASON TIME SERIES MODEL

PREMA BORKAR*

Gokhale Institute of Politics and Economics, BMCC Road, Deccan Gymkhana, Pune, Maharashtra- 411 004, India

Received: 14.05.2014 **Revised accepted: 30.05.2014**

ABSTRACT

The paper describes an empirical study of modeling and forecasting time series data of soybean production in India. Yearly soybean production data for the period of 1970-1971 to 2011-2012 of India were analyzed by time-series methods. Autocorrelation and partial autocorrelation functions were calculated for the data. The Box Jenkins ARIMA (autoregressive integrated moving average) methodology has been used for forecasting. The diagnostic checking has shown that ARIMA (1, 1, 1) is appropriate. The forecasts from 2012-2013 to 2024-2025 are calculated based on the selected model. The forecasting power of autoregressive integrated moving average model was used to forecast soybean production for thirteen leading years. These forecasts would be helpful for the policy makers to foresee ahead of time the future requirements of soybean seed, import and/or export and adopt appropriate measures in this regard.

Keywords:
ACF, ARIMA, PACF, Soybean

INTRODUCTION

The soybean (*Glycine max*) is known as the "Golden Bean" of the 20th century. Though Soybean is a legume crop, it is classed as an oilseed rather than a pulse. It is grown in tropical, subtropical, and temperate climates. Soybean is the world's most cultivated oilseed. The soybean is often called the miracle crop. It has emerged as one of the important commercial crop in many countries. Due to its worldwide popularity, the international trade of Soybean is spread globally. Soybeans now account for about 60 % of the total global oilseed production of 390-425 million tons with cottonseed, the closest competitor (Commodity Profile-ICEX). Although, a native of China soybean for all practical reason is an American crop today, USA is the major producer of soybean and ranks first in production. Its share in the world production is almost 35 %. Brazil, Argentina and China rank second, third and fourth position in terms of production respectively. India occupies fifth place.

Forecasts have traditionally been made using structural econometric models. Concentration have been given on the univariate time series models known as auto regressing integrated moving average (ARIMA) models, which are primarily due to work of Box and Jenkins (1970). These models have been extensively used in practice for forecasting economic time series, inventory and sales modeling (Brown, 1959 and Holt *et al.*, 1960) and are generalization of the exponentially weighted moving average process. Several methods for identifying special cases of ARIMA models have been suggested by Box- Jenkins and others. Makridakis *et al.* (1982), and Meese and Geweke (1982) have discussed the methods of identifying univariate models. Among others Jenkins and Watts (1968), Yule (1926, 1927), Bartlett (1964), Quenouille (1949), Ljune and Bos (1978) and Pindyck and Tubinfeld (1981) have also emphasized the use of ARIMA models.

In this study, these models were applied to forecast the production of soybean crop in India. This would enable to predict expected soybean production for the years from 2013 onward. Such an exercise would enable the policy makers to foresee ahead of time the future requirements for soybean seed, import and/or export of soybean thereby enabling them to take appropriate measures in this regard. The forecasts would thus help save much of the precious resources of our country which otherwise would have been wasted.

MATERIALS AND METHODS

Respective time series data for this study were collected from various Government Publications of India. Box and Jenkins (1976) linear time series model was applied. Auto Regressive Integrated Moving Average (ARIMA) is the most general class of models for forecasting a time series. Different series appearing in the forecasting equations are called "Auto-Regressive" process. Appearance of lags of the forecast errors

*Corresponding author email:premaborkar@rediffmail.com

in the model is called "moving average" process. The ARIMA model is denoted by ARIMA (p,d,q),

Where,

"P" stands for the order of the auto regressive process,

"d" is the order of the data stationary and

"q" is the order of the moving average process.

The general form of the ARIMA (p,d,q) can be written as described by Judge, *et al*. (1988).

$$\Delta^d y_t = \delta + \theta_1 \Delta^d y_{t-1} + \theta_2 \Delta^d y_{t-2} + \text{----- } + \theta_p y_{t-p} + e_{t-1} \alpha\, e_{t-1} - \alpha_2\, e_{t-2}\, \alpha_q e_{t-2} \quad (1)$$

Where,

Δ^d denotes differencing of order d,i.e., $\Delta y_t = y_t - y_{t-1}$,

$\Delta_2 y_t = \Delta y_t - \Delta_{t-1}$ and so forth,

Y_{t-1} ----- y_{t-p} are past observations (lags),

δ, θ_1 -------- θ_p are parameters (constant and coefficient) to be estimated similar to regression coefficients of the Auto Regressive process (AR) of order "p" denoted by AR (p) and is written as

$$Y = \delta + \theta 1\, y_{t-1} + \theta 2\, y_{t-2} + \text{---------} + \theta p\, y_{t-p} + et \quad (2)$$

Where,

e_t is forecast error, assumed to be independently distributed across time with mean θ and variance $\theta_2 e$, e_{t-1}, e_{t-2} ------ e_{t-q} are past forecast errors,

α_1, ---------- α_q are moving average (MA) coefficient that needs to be estimated.

While MA model of order q (i.e.) MA (q) can be written as

$$Y_t = e_t - \alpha_1 \alpha_{t-1} - \alpha_2 e_{t-2} \text{ ---------- } \alpha_q e_{t-q} \quad (3)$$

The major problem in ARIMA modeling technique is to choose the most appropriate values for the p, d, and q. This problem can be partially resolved by looking at the Auto correlation function (ACF) and partial Auto Correlation Functions (PACF) for the series (Pindyk and Rubinfeld, 1981). The degree of the homogeneity, (d) i.e. the number of time series to be differenced to yield a stationary series was determined on the basis where the ACF approached zero.

After determining "d" a stationary series $\Delta d\, y_t$ its auto correlation function and partial autocorrelation were examined to determined values of p and q, next step was to "estimate" the model. The model was estimated using computer package "SPSS".

Diagnostic checks were applied to the so obtained results. The first diagnostic check was to draw a time series plot of residuals. When the plot made a rectangular scatter around a zero horizontal level with no trend, the applied model was declared as proper. Identification of normality served as the second diagnostic check. For this purpose, normal scores were plotted against residuals and it was declared in case of a straight line. Secondly, a histogram of the residuals was plotted. Finding out the fitness of good served as the third check. Residuals were plotted against corresponding fitted values: Model was declared a good fit when the plot showed no pattern.

Using the results of ARIMA (p,q,d), forecasts from 2013 up to 2025 were made. These projections were based on the following assumptions.

- Absence of random shocks in the economy, internal or external.
- Agricultural price structure and polices will remain unchanged.
- Consumer preferences will remain the same.

RESULTS AND DISCUSSION

ARIMA model for Soybean Production data in India

To fit an ARIMA model requires a sufficiently large data set. In this study, we used the data for Soybean production for the period 1970-1971 to 2011-2012. As we have earlier stated that development of ARIMA model for any variable involves four steps: identification, estimation, diagnostic checking and forecasting. Each of these four steps is now explained for soybean production. The time plot of the soybean production data is presented in Figure-1.

The time plot presented (Figure-1) indicated that the given series 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 Xt are computed (Table-1). The ACF and PACF (Figure-2) 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 (Figure-2). The models and corresponding AIC and BIC values are

ARIMA (p, d, q)	AIC	BIC
1 0 0	115.88	119.35
1 1 1	97.16	102.30
1 0 1	115.72	120.94

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

Model parameters were estimated using SPSS package. Results of estimation are reported in table 2. 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.

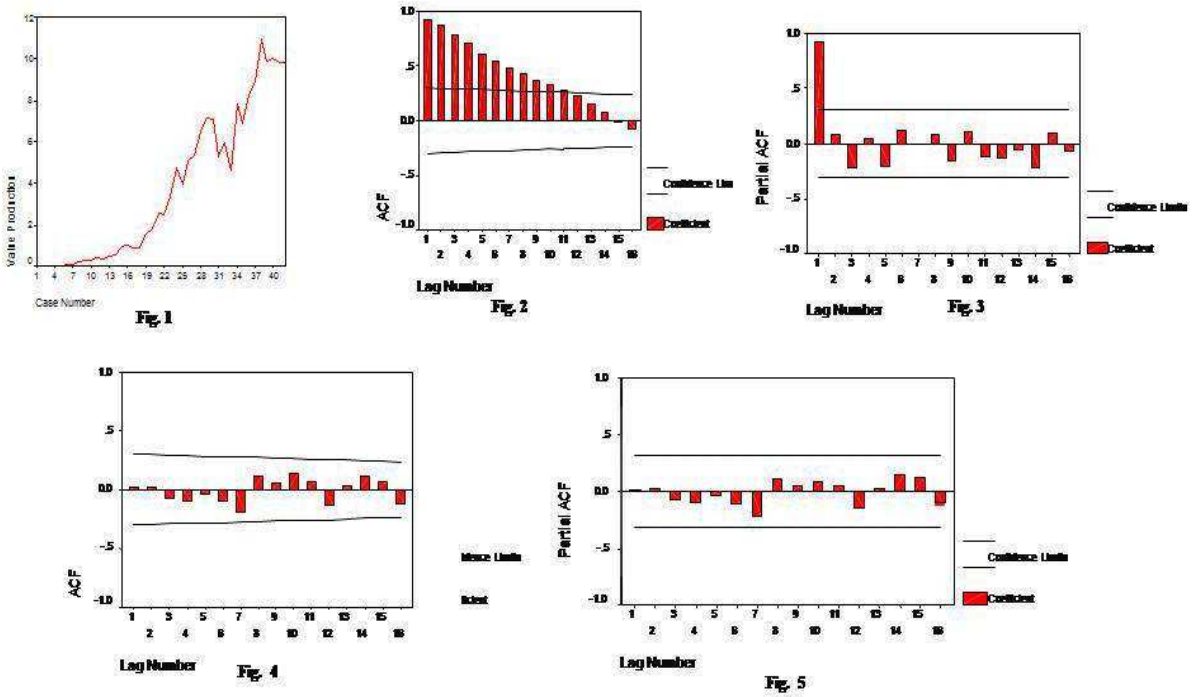

Figure 1-5: Shows time plot of soybean production data; ACF of differenced data; PACF of differenced data; ACF of residuals of fitted ARIMA model; PACF of residuals of fitted ARIMA model, respectively

Table 1: Autocorrelations and partial autocorrelations

Lag	Autocorrelation	Std.error	Lag	Partial Autocorrelation	Std.error
1	0.923	0.149	1	0.923	0.154
2	0.865	0.147	2	0.088	0.154
3	0.777	0.145	3	-0.215	0.154
4	0.709	0.143	4	0.038	0.154
5	0.609	0.141	5	-0.207	0.154
6	0.544	0.140	6	0.128	0.154
7	0.472	0.138	7	0.003	0.154
8	0.431	0.136	8	0.084	0.154
9	0.361	0.134	9	-0.156	0.154
10	0.330	0.132	10	0.106	0.154
11	0.273	0.130	11	-0.119	0.154
12	0.220	0.127	12	-0.137	0.154
13	0.147	0.125	13	-0.054	0.154
14	0.069	0.123	14	-0.224	0.154
15	-0.006	0.121	15	0.093	0.154
16	-0.074	0.119	16	-0.070	0.154

Table 2: Estimates of the fitted ARIMA model

		Estimates	Std Error	t	Approx sig
Non- Seasonal lag	AR1	-0.8871998	0.14239325	-6.230	0.0000
	MA1	-0.6342917	0.24798666	-2.557	0.01464
Constant		0.23689333	0.10290754	2.3020	0.02690
Number of Residuals		41			
Number of Parameters		2			
Residual df		38			
Adjusted Residual Sum of Squares		22.131553			
Residual Sum of Squares		24.109698			
Model Std. Error		0.75937641			
Log-Likelihood		-45.580591			
Akaike's Information Criteria (AIC)		1124.84			
Schwarz's Bayesian Criterion (BIC)		97.16			

Table 3: Autocorrelations and partial autocorrelations of residuals

Lag	Auto correlation	Std.error	Box- Ljung	df	Sig.	Lag	Partial Autocorrelation	Std.error
1	0.012	0.151	0.006	1.000	0.937	1	0.012	0.156
2	0.023	0.149	0.031	2.000	0.985	2	0.023	0.156
3	-0.074	0.147	0.285	3.000	0.963	3	-0.075	0.156
4	-0.092	0.145	0.687	4.000	0.953	4	-0.091	0.156
5	-0.042	0.143	0.774	5.000	0.979	5	-0.037	0.156
6	-0.104	0.141	1.323	6.000	0.970	6	-0.107	0.156
7	-0.199	0.139	3.375	7.000	0.848	7	-0.216	0.156
8	0.117	0.137	4.107	8.000	0.847	8	0.110	0.156
9	0.058	0.135	4.295	9.000	0.891	9	0.045	0.156
10	0.137	0.133	5.364	10.000	0.866	10	0.084	0.156
11	0.069	0.130	5.647	11.000	0.896	11	0.047	0.156
12	-0.136	0.128	6.775	12.000	0.872	12	-0.148	0.156
13	0.033	0.126	6.842	13.000	0.910	13	0.022	0.156
14	0.106	0.124	7.581	14.000	0.910	14	0.141	0.156
15	0.059	0.121	7.818	15.000	0.931	15	0.126	0.156
16	-0.126	0.119	8.944	16.000	0.916	16	-0.134	0.156

This is done through examining the autocorrelations and partial autocorrelations of the residuals of various orders. For this purpose, the various correlations up to 16 lags were computed and the same along with their significance which is tested by Box-Ljung test are provided in table-3. 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 (Figure-4) also indicate 'good fit' of the model.

The last stage in the modeling process is forecasting. ARIMA models are developed basically to forecast the corresponding variable (Figure-5). 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. The residuals calculated during the estimation process, are considered as the one step ahead forecast errors. The forecasts are obtained for the subsequent agriculture year from 2012-13 to 2024-2025.

Table 4: Forecasts for Soybean Production (2012-13 to 2024-2025) (Million tonnes)

Years	Forecasted Production	Lower limit	Upper limit
2012-2013	9.83283	8.27685	11.38882
2013-2014	10.26851	8.30411	12.23292
2014-2015	10.32904	7.81106	12.84703
2015-2016	10.72241	7.88908	13.55573
2016-2017	10.82048	7.56599	14.07497
2017-2018	11.18053	7.64235	14.71872
2018-2019	11.30816	7.41164	15.20467
2019-2020	11.64199	7.47787	15.80612
2020-2021	11.79288	7.30757	16.27819
2021-2022	12.10608	7.36304	16.84911
2022-2023	12.27527	7.23547	17.31507
2023-2024	12.57223	7.28152	17.86293
2024-2025	12.75583	7.18548	18.32619

CONCLUSION

In our study, the developed model for soybean production was found to be ARIMA (1,1,1). The forecasts of soybean production, lower control limits (LCL) and upper control limits (UCL) are presented in Table-4. 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 soybean production in India. However, it should be updated from time to time with incorporation of current data.

This paper discloses the production of soybean from 1970-1971 to 2011-2012 and also shows the future movement. To formulate future development plan for soybean production, it is essential to know the previous condition and also see the future trend. In this study, forecasting is done by using some sophisticated statistical tools so that the government and policy makers can easily realize about the future development of soybean production and could take initiatives to improve the production.

REFERENCES

Bartlett, M.S. 1964. On The Theoretical Specification of Sampling Properties of Auto correlated Time Series. *Journal of Royal Statistic Society*. **8**:27–41.

Box, G.E.P. and Jenkins, G.M. 1976. *Time Series Analysis: Forecasting and Control*, San Francisco: Holden-Day.

Box, G.E.P. and Jenkins, G. M.. 1970. Time series analysis: forecasting and Control. San Francisco: Holden-Day.

Brown, R.G. 1959. Statistical Forecasting For Inventory Control. New York, McGraw-Hill.

Holt, C.C., Modigliani, F., Muth, J.F. and. Simon, H.A. 1960. Planning, Production, Inventores, and Work Force. Prentice Hall, Englewood Cliffs, NJ, USA.

Jenkins, G. M. and Watts, D.G. 1968. Spectral Analysis and its Application, Day, San Francisco, California, USA.

Ljunge, G.M. and Box, G.E.P. 1978. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, ***65:***67–72.

Makridakis, S., Anderson, A., Filds, R., Hibon, M., Lewandowski, R., Newton, J., Parzen E. and Winkler, R. 1982. The Accuracy of Extrapolation (time series) methods: Results of a Forecasting Competition. *Journal of Forecasting Competition.* **1:**111–53.

Meese, R. and J. Geweke, 1982. A Comparison of Autoregressive Univariate Forecasting Procedures for Macroeconomic Time Series. Thesis, University of California, Berkeley, CA, USA.

Prindycke, R.S. and Rubinfeld, D.L. 1981. Econometric Models and Economic Forecasts**.** 2nd Ed. New York, McGraw-Hill.

Quenouille, M.H. 1949. Approximate Tests of Correlation in Time- Series. *Journal of Royal Statistic Society.* B**11**:68–84.

Yule, G.U. 1926. Why Do We Sometimes Get Nonsence-corrleations Between Times Series. A study in Sampling and the Nature of Series. *Journal of Royal Statistic Society.* **89**:1–69.

Yule, G.U. 1927. On a method of Investigation Periodicities in Disturbed Series,With Specia; Reference To Wolfer's Sunspot Number. *Philosophical Transactions of the Royal Society A.* **226:**267–98.