Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.5834

ISSN: 2320 – 7051

Int. J. Pure App. Biosci. **5 (5):** 1538-1546 (2017)

International Journal of Pure & Applied Bioscience

Research Article

OPEN ACCESS

Forecasting of Soybean Yield in India through ARIMA Model

Mahesh Kumar[1*], Rohan Kumar Raman[2] and Subhash Kumar[1]

[1]Dr. Rajendra Prasad Central Agricultural University, Pusa, Samastipur, Bihar-848125, India

[2]ICAR-Central Inland Fisheries Research Institute, Barrackpore, Kolkata-700120, India

*Corresponding Author E-mail: mahesh_smca@yahoo.co.in

Received: 15.09.2017 | Revised: 21.10.2017 | Accepted: 25.10.2017

ABSTRACT

Forecasting of soybean productivity is of immense value and plays an important role in many important decisions. There are several models by the help of which forecasting of production of soybean can be carried out. In this research paper, we have discussed yield of soybean in India for last 40 years. In addition to that, we have forecasted the production of soybean for next 5 years.

Key words: *Soy protein; Productivity; ARIMA; Box-Jenkins modeling*

INTRODUCTION

Soybean (*Glycine max* L.) is the leading cash crop of India. It is grown in area of 11.65 million hectares with production of 8.00 million ton in India[6]. The average productivity is 687 Kg/hectare (Year 2015-16). It is now a complex scientific activity aimed at producing maximum amount of agricultural produce with minimum expenditure in terms of time, space and energy to meet the needs of a growing population and economy. In spite of recent technological advances, the soybean productivity is low. Forecasting of soybean productivity is of immense value and plays an important role in many important decisions. The Univariate Box-Jenkins[3] approach for forecasting is based on the solid foundation of classical probability theory and mathematical statistics. It is a family of models out of which one appropriate model is selected having optimal Univariate forecast. For the purpose, data on yield (t/ha) of soybean has been collected for the period of 46 years i.e. from 1970 to 1915 (Table 1) (Source: Agricultural Statistics at a Glance-2014 and, Oilseed: World Market and Trade, March 2016 issue, published by USDA) for building forecast model and generating short term forecast on soybean productivity.

Cite this article: Kumar, M., Raman, R.K. and Kumar, S., Forecasting of Soybean Yield in India through ARIMA Model, *Int. J. Pure App. Biosci.* **5(5):** 1538-1546 (2017). doi: http://dx.doi.org/10.18782/2320-7051.5834

Table 1: Area, Production and Productivity of soybeans in India: 1970 to 2015

Sl. No	Year	Area (Million hec.)	Production (Million Tonnes)	Yield Kg/hec	Sl. No	Year	Area (Million hec)	Production (Million Tonnes)	Yield Kg/hec
1	1970-71	0.03	0.01	426	24	1994-95	4.32	3.93	911
2	1971-72	0.03	0.01	426	25	1995-96	5.04	5.10	1012
3	1972-73	0.03	0.03	819	26	1996-97	5.44	5.38	989
4	1973-74	0.05	0.04	829	27	1997-98	5.99	6.46	1079
5	1974-75	0.07	0.05	768	28	1998-99	6.49	7.14	1100
6	1975-76	0.09	0.09	975	29	1999-00	6.22	7.08	1138
7	1976-77	0.13	0.12	988	30	2000-01	6.42	5.28	823
8	1977-78	0.20	0.18	940	31	2001-02	6.34	5.96	940
9	1979-80	0.50	0.28	568	32	2002-03	6.11	4.65	762
10	1980-81	0.61	0.44	728	33	2003-04	6.55	7.82	1193
11	1981-82	0.48	0.35	741	34	2004-05	7.57	6.87	908
12	1982-83	0.77	0.49	637	35	2005-06	7.71	8.27	1073
13	1983-84	0.84	0.61	735	36	2006-07	8.33	8.85	1063
14	1984-85	1.24	0.95	768	37	2007-08	8.88	10.97	1235
15	1985-86	1.34	1.02	764	38	2008-09	9.51	9.91	1041
16	1986-87	1.53	0.89	584	39	2009-10	9.73	9.96	1024
17	1987-88	1.54	0.90	582	40	2010-11	9.60	12.74	1327
18	1988-89	1.73	1.55	892	41	2011-12	10.11	12.21	1208
19	1989-90	2.25	1.81	801	42	2012-13	10.84	14.67	1353
20	1990-91	2.56	2.60	1015	43	2013-14	12.20	9.50	779
21	1991-92	3.18	2.49	782	44	2014-15	10.90	8.70	798
22	1992-93	3.79	3.39	894	45	2015-16	11.65	8.00	687
23	1993-94	4.37	4.75	1086					

Methodology for Selecting Model Through ARIMA

This approach automatically selects most reliable forecast model from the family of ARIMA model by going through three iterative stages i.e., Identification stages, Estimation stages and Diagnostic checking stage. This technique provides a parsimonious model that is a model with smallest number of parameters for describing the available data. The secondary data are covering the period from the year 1970 to 2015 for India. Building an ARIMA (p,d,q) model basically consisted of three steps, namely; (a) Identification of the order of the model (b) Estimation of model parameters and (c) Diagnostic checking for adequacy of the fitted model as mentioned above also[3].

Mathematically, an ARIMA (p,d,q) model is given by-

$$\phi(B)\,\Delta^d \bar{Z}_t = \Theta(B)\, a_t$$

Where,

Δ^d = $(1-B)^d$

$\phi(B)$ = $(1-\phi B-\phi B^2-\ldots\ldots\ldots-\phi B^p)$

$\Theta(B)$ = $(1-\Theta_1 B\ \Theta_2 B^2-\ldots\ldots\ldots\ \Theta_q B^q)$

Z = $Z_t - \mu$

Z_t = *Stationary time series data*

d = *Order of differencing*

a_t = *Random shock*

p = *Order of auto-regression*

q = *Order of moving average*

Under identification phase, the first order differencing (r=1,2, …) of Z_t is done till a stationary time series is achieved. The order p & q is decided on the basis of ACF & PACF and the criteria led down by Box and Jenkins[3], after determining the value of p, d and q. The model parameters are estimated. Diagnostic checking of the fitted model is done through some important statistics such as t-test and χ^2 (Chi-square) of the residual ACF.

Brief descriptions of various models of L-Jung, G.M., Box, G.E.P[5]. ARIMA family are cited here:

ARIMA ModelARIMA model is an algebraic statement telling how the observations on a variable are statistically related to past observation on the same variable. In fact, ARIMA model is a family of models consisting of three kinds of model, which are given below;

a) Autoregressive Model: This can be represented as

$Z_t = C + \phi_1 Z_{t-1} + a_t$ …(1)

Where

C = *μ (1- ϕ₁) = Constant term*

μ = *Constant parameter*

ϕ = *Deterministic coefficient its value determines the relationship between Z_t and Z_{t-1} (Lagged observation)*

a_t = *Random shock having some continuous statistical distribution.*

The term $\phi_1 Z_{t-1}$ is autoregressive term, and the longest lag attached to it is t-1 thus, above is autoregressive model of order 1, denoted as AR (1). The parameters of model (1) are estimated by least square method. Approximate estimates for μ and ϕ_1 can be obtained as Z (mean of the available observation) and r_1 (autocorrelation function) respectively. Similarly, second order autoregressive model denoted as AR (2) can be represented as

$Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t$

In this model, Z_t is linearly related to the past observation Z_{t-1} and Z_{t-2}. The lease square estimate of ϕ_1 and ϕ_2 are approximated by

$\phi_1 = r_1 (1-r_2) / 1-r^2_1$ and $\phi_2 = r_2 - r^2_1) / 1-r^2_1$

Where,

r_1 & r_2 are autocorrelation function for first and second lag respectively.

In general, one can represent autoregressive model of order p denoted as AR (p) as a linear combination of p-past values and a random term i.e.

$Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$

b) Moving Average (MA) Model: A moving average model of order one denoted as MA (1) can be represented as

$Z_t = C - \Theta_1 a_{t-1} + a_t$ …(2)

Where,

C = *μ (1- Θ_1) = constant term*

Θ_1 = *Moving average coefficient determines the statistical relationship between Z_t and a_{t-1} (lagged random shock)*

a_t = *random shock with mean '0' and variance σ^2.*

***Estimation of Parameters of MA Model*:** Estimation of parameters of MA model is more difficult than an AR model because efficient explicit estimators cannot be found. Instead some numerical iteration method is used. For example, to estimate μ and Θ of Equation 2 i.e.

$Z_t = C - \Theta_1 a_{t-1} + a_t$

residual sum of square (RSS) $\sum a^2_t$ in terms of observed Z's and the parameters μ and Θ are obtained and then it is differentiated with respect to μ and Θ to obtain estimated μ and Θ. Unfortunately, the RSS is not a quadratic function of the parameters and so explicit least square estimates cannot be found. An iterative procedure suggested by Box-Jenkins is used in which suitable values of μ and Θ such as μ = Z and Θ given by the solution of Equation 3.

$Z_t = C + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} - \Theta_1 a_{t-1} \ldots - \Theta_q a_{t-q} + a_t$ …(3)

Then the RSS may be calculated recursively from

$a_t = Z_1 - c + \Theta_1 a_{t-1}$ with $a_0=0$

This procedure then can be repeated for a grid of points in (μ, Θ) plane. We may then by inspection choose that value of (μ,Θ) as estimates which minimized RSS. The lease square estimates are also maximum likelihood estimated conditional on a fixed value of a_0 provided a_t is normally distributed.

c) Autoregressive Moving Average Model (ARMA): The combination of AR (p) and MA (q) models to describe a given series is known as ARMA (p, q) which can be represented as

$Z_t = C + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} - \Theta_1 a_{t-1} \ldots - \Theta_q a_{t-q} + a_t$

The Box-Jenkins Modeling Procedure

Box-Jenkins proposed a practical three stage procedure for finding a good model. A sketch of the broad outline of the Box-Jenkins modeling procedure is summarized schematically in Figure 1.

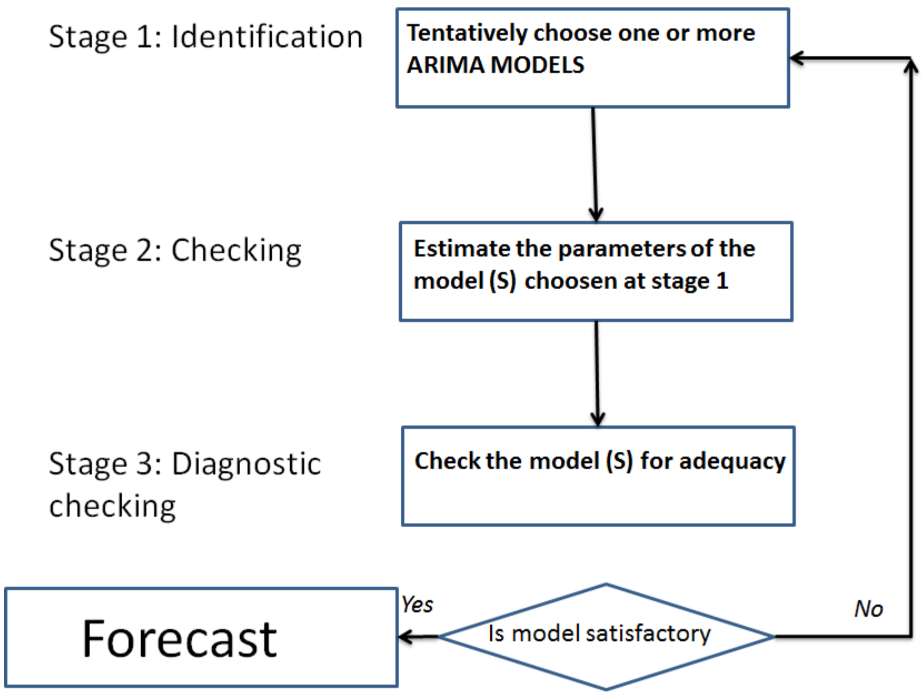

Fig. 1: Stages in the Box-Jenkins iterative approach to model building.

a) Stage 1:- Identification of the order of the model: If $\bar{Z}$ be the mean of a stationary time series such that $Z_t = Z_t - \bar{Z}$ denoting the number of observations by n and the number of computable lags by k the estimated autocorrelation function (ACF) r_k of the observations are separated by k time periods.

b) Stage 2:- Estimation of model parameter: Box-Jenkins time series models written as ARIMA (p, d, q) amalgamate three type of processes namely auto-regressive (AR) or order p; differencing to make a series stationary of degree d and moving average (MA) of order q. At the parameter estimation stage, the aim is to obtain estimates of the tentatively identified ARMA model parameters of Stage-I for given values of p and q. In general, ARIMA coefficients (the ɸ 's and Ɵ's) must be estimated using a nonlinear least square procedure, while several nonlinear least square methods are available, the one most commonly used t estimate ARIMA models is known as "Marquardt's compromise".

c) Stage 3: Diagnostic checking for the adequacy of the model: This is the third stage of model formulation. At this stage, the decision about the statistical adequacy of the model is taken. Most important test of the statistical adequacy at an ARIMA model involves the assumptions that the random shocks (a_t) are independent. Meaning not autocorrelate, since in practice the random shocks cannot be observed, the estimate at residual(a_t) is taken to test the hypothesis about the independent of random shocks. This is mainly performed by the examination of residual ACF, t test for the residual ACF and χ^2 –test based on L-Jung and Box for the residual autocorrelation (L-Jung,G.M., Box, G.E.P[5]. Gupta[4] has discussed about ARIMA model and forecasts on tea production in India. He developed and applied an ARIMA forecasting model for tea production in India. Boran and Bora[2] have discussed about the monthly rainfall around Guwahati using a seasonal ARIMA model. Prajneshu and Venugopalan[8] have studied various statistical modeling techniques viz. polynomial function fitting approach, ARIMA time series methodology and non-linear mechanistic growth modeling approach for describing marine, inland as well as total fish production of the country during the periods 1950-51 to 1994-95.

RESULTS AND DISCUSSION

Soybean *(Glycine max L.)* productivity of India is forecasted through fitting of well-known Box Jenkins Univariate Auto Regressive Integrated Moving Average (ARIMA) model. The data on soybean productivity in India from the year 1970 to 2010 were utilized to build an ARIMA model and validated through five year productivity data from 2011 to 2015. Akaike Information Criterion (AIC) and Baysian Information Criteria (BIC) were selected for best model selection criteria. ARIMA (1, 1, 0) model found best suitable model for soybean productivity in India based on AIC and BIC criteria (Table 2).

Table 2: Best seven ARIMA models.

ARIMA MODEL	AIC	BIC
(1,1,0)	597.75	601.08
(1,1,1)	599.4	608.49
(2,1,0)	601.3	608.53
(2,0,0)	611.35	618.66
(1,0,0)	612.89	618.37
(2,0,1)	612.71	621.85
(0,0,1)	620.45	625.93

Using developed ARIMA (1, 1, 0) model soybean productivity in India was forecasted for five year ahead i.e., year 2016 to 2020.The results showed almost equal trend as from 2016 to 2020 i.e. 734.62, 714.14, 722.95, 719.19 and 720.80 kg/hec (Table. 3).

Table 3: Forecast of productivity of soybean for five next years

Point Forecast	Productivity (kg/hec)	Low 80	High 80	Low 95	High 95
2016	734.6207	508.5289	960.7126	388.8431	1080.398
2017	714.1907	453.8389	974.5425	316.0170	1112.364
2018	722.9555	411.6280	1034.2822	246.8224	1199.089
2019	719.1953	372.3681	1066.0224	188.7689	1249.622
2020	720.8085	338.6383	1102.9786	136.3297	1305.287

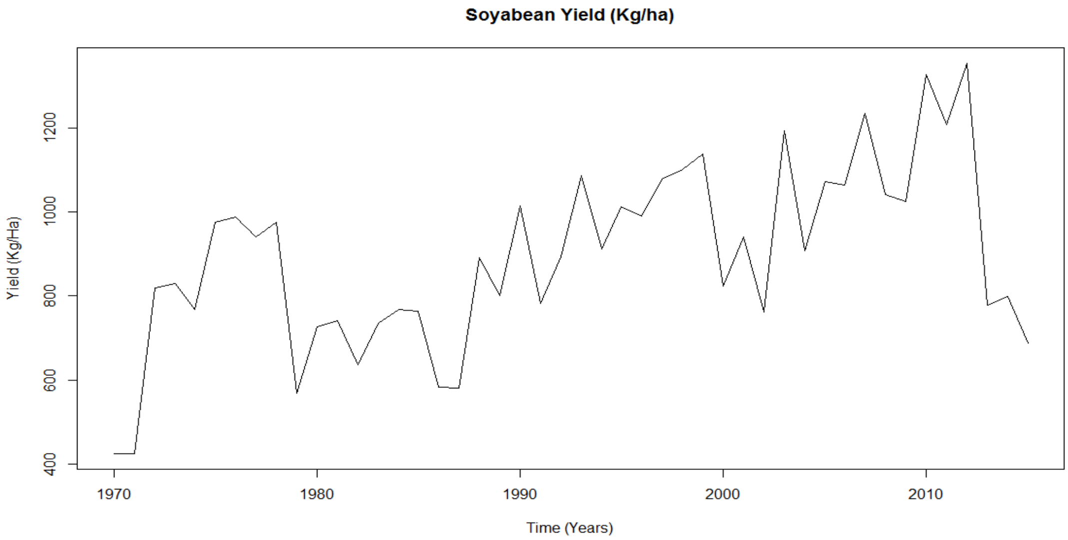

Fig. 2: Plot of soybean productivity (kg/ha) with time from the period 1970-2015

Fig. 3: and Figure 4 show Auto correlation function (ACF) and Partial autocorrelation function (PACF) respectively, in the data

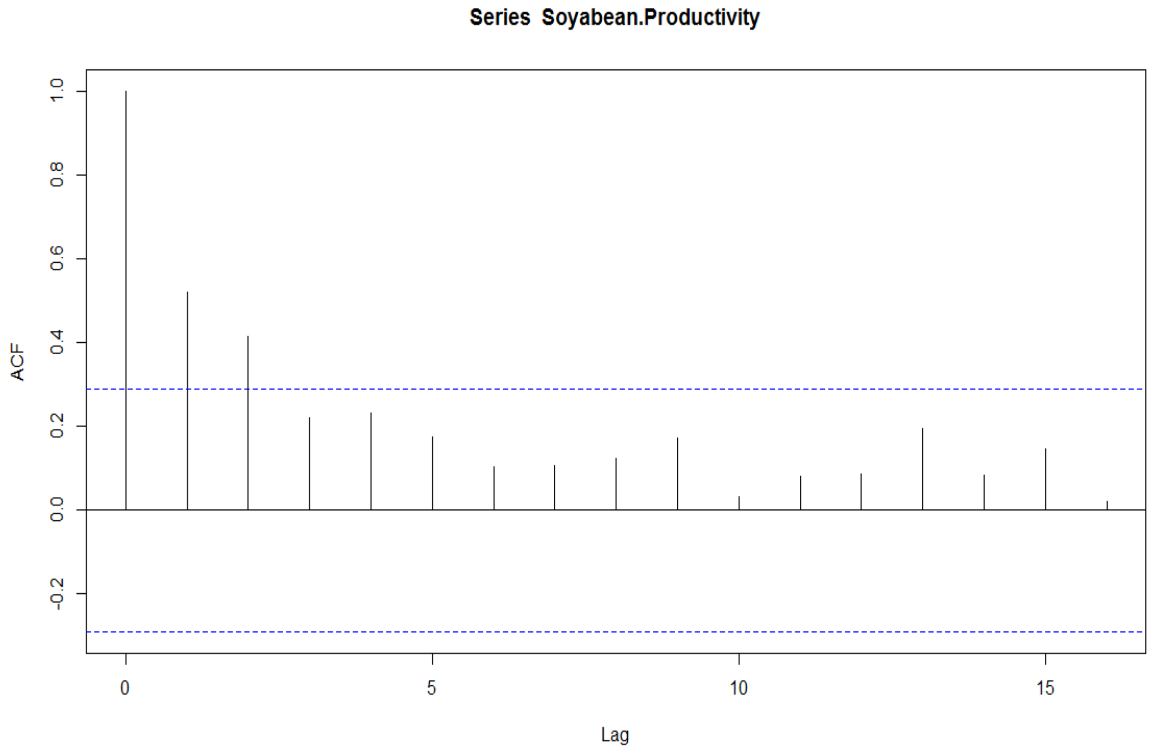

Fig. 3: ACF of soybean yield

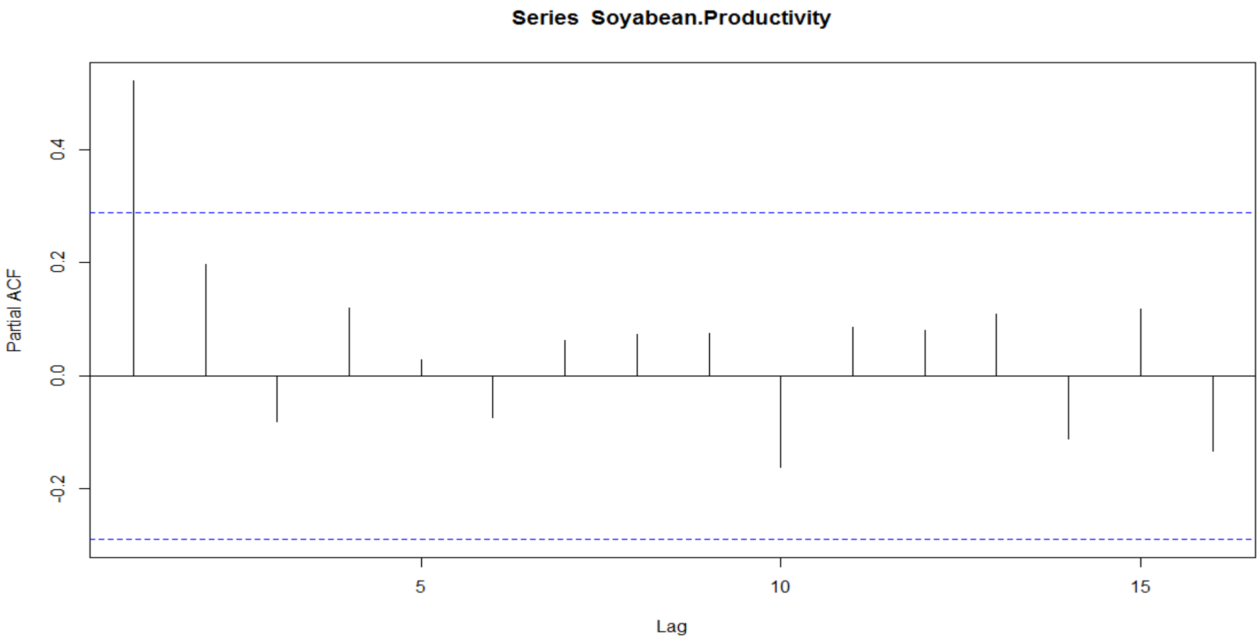

Fig. 4: PACF of soybean yield

Figure2 shows the original catch trend of yield. The data looks non-stationary in nature. Hence to make it stationary first order differentiate (d=1) is presented Figure 5.

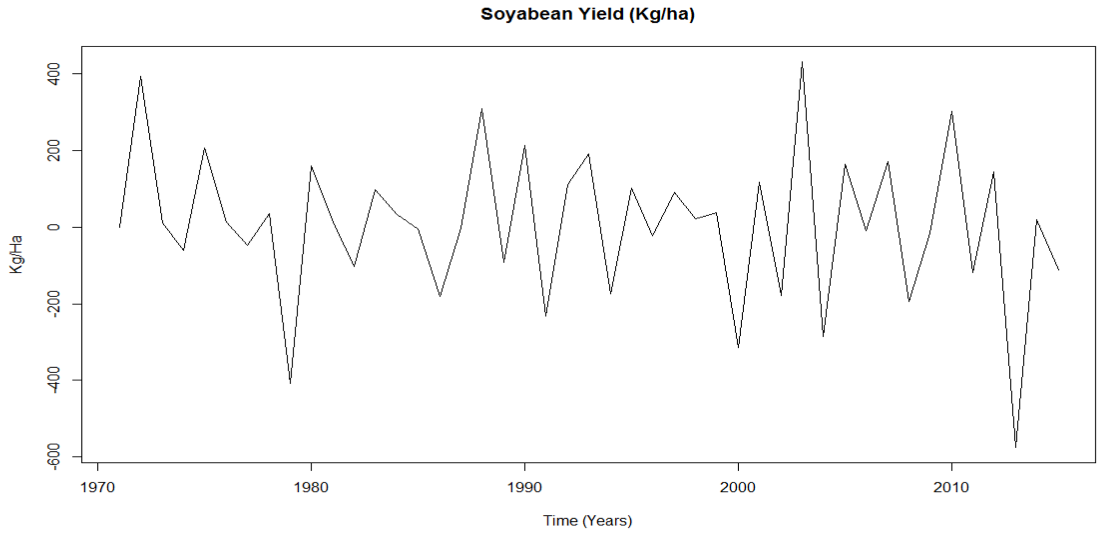

Fig. 5: Soybean yield after first order differentiation

The Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) are helpful to decide the model identification and model order. To confirm stationarity of the data, augmented Dickey–Fuller test (ADF) test was performed. The Dickey-Fuller statistics value = -10.567, p-value = 0.01 suggest to go for alternative hypothesis that is stationary. As we confirm the stationarity of the data, next step is to estimate the Auto Regressive Integrated Moving Average (ARIMA) model parameter estimation after differencing the data ACF (Figure6) and PACF (Figure7).

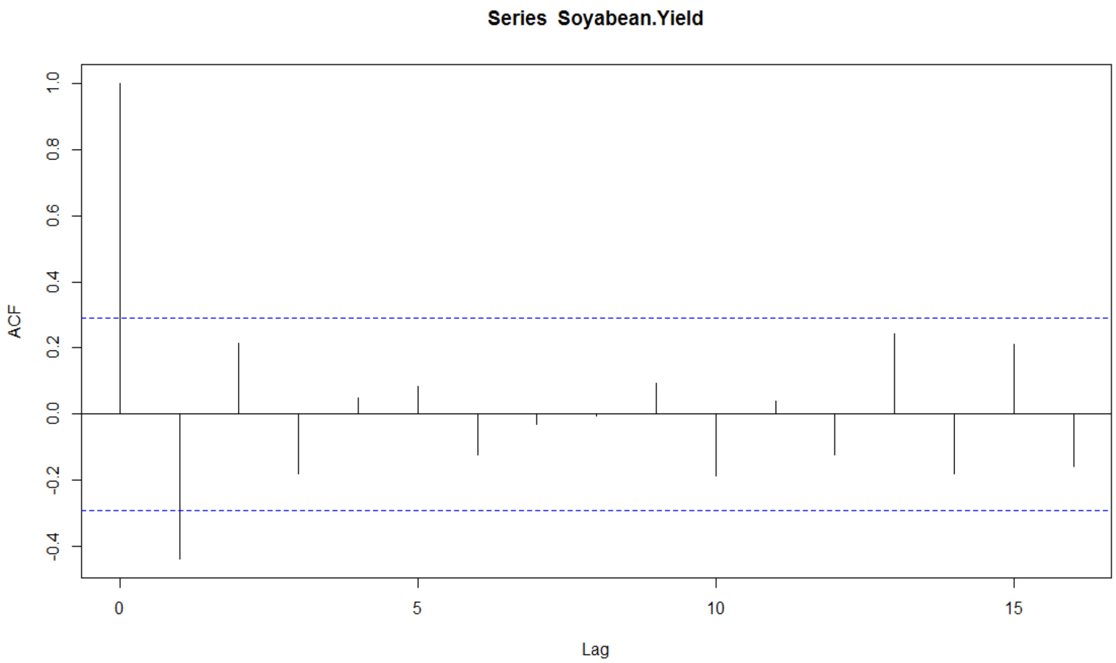

Fig. 6: ACF of soybean yield after first order differentiation

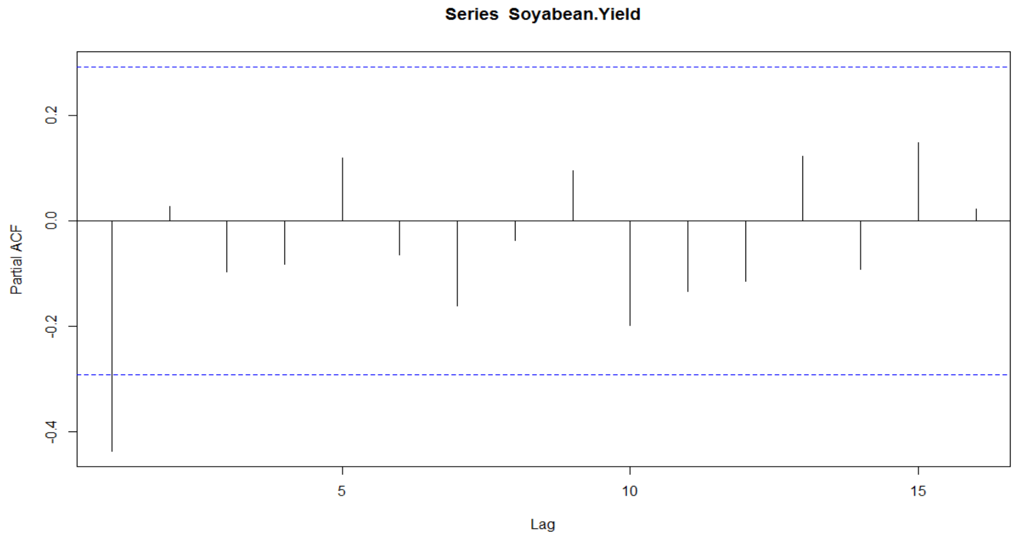

Fig. 7: PACF of soya bean yield after first order differentiation

Yield trend of soybean for 5 next years i.e.2016,2017,2018, 2019 & 2020 is shown in Figure 8.

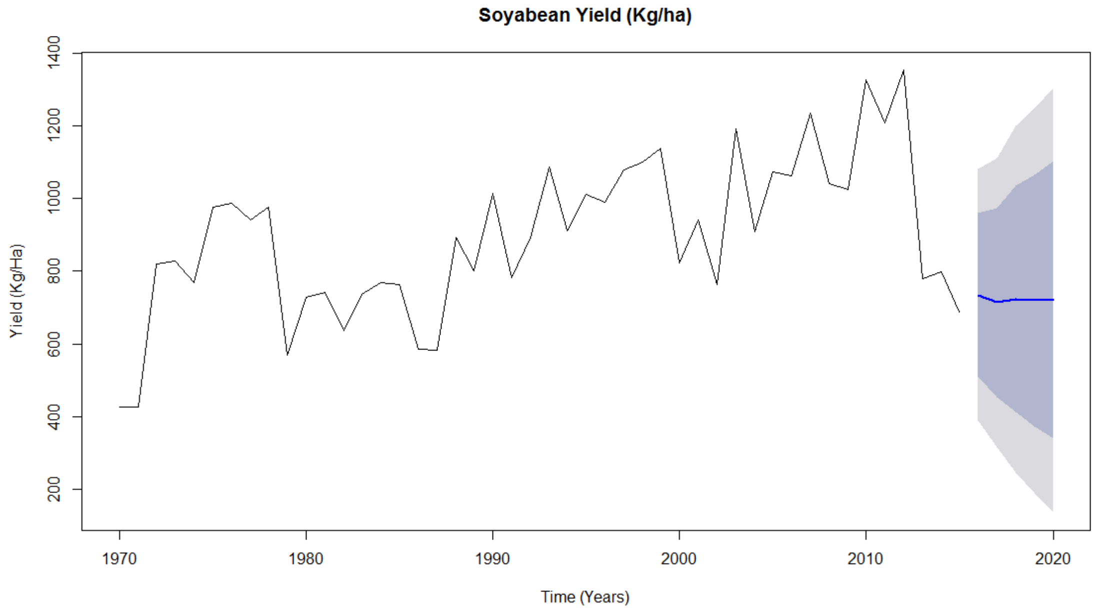

Fig. 8: Yield trend of soya bean for next 5 years i.e. from 2016 to 2020

The University of Illinois, USAID and the foundation of an American missionary, Robert W. Nave, played a key role in the commercial development of soy products and in setting up initial processing facilities. Mr Nave founded the Soy Production and Research Association (SPRA) Trikha et.al[7] (1979) as a joint venture of the Nave Technical Institute, already established by him in Bareilly and Pantnagar University. There is scope to increase area in Madhya Pradesh, Maharashtra, Rajasthan, Tamil Nadu, Andhra Pradesh and Karnataka. Some more area can be brought under soybean in North East and Bihar also.

During 1970-1971, the regional spread of soybean cultivation covered 7700 hectares in Madhya Pradesh, 5900 hectares in Uttar Pradesh, and 18000 hectares in Maharashtra. Soon the crop started spreading based on comparative advantage. These three state together account for more than 96 percent of the area under cultivation as well as the production of soybeans in the country.

According to a study conducted in Madhya Pradesh, the cropped area allocated to soybean cultivation was found to be 60.12 percent on small-size holdings, 40.31 percent on medium-size holdings and 27.27 percent on large-size holdings for 1984-1985[1].

CONCLUSION

India normally produces only a little over 3% of world's soybean which is estimated to be 320 million tons this year. In 2015, production fell to only 7.4 million tons (SOPA estimates), due to erratic monsoon. Using developed ARIMA (1, 1, 0) model soybean productivity in India was forecasted for next five years. The results showed almost equal trend as from 2016 to 2020 i.e. 734.62, 714.14, 722.95, 719.19 and 720.80 kg/hec. India is the only country which does not grow genetically modified (GM) soybean. There is scope to increase cultivation area in Madhya Pradesh, Maharashtra, Rajasthan, Tamil Nadu, Andhra Pradesh and Karnataka. Some more area can be brought under soybean cultivation in North East and Bihar also. Recognizing the importance of soybean cultivation, in 1987 the ICAR established the National Research Centre for Soybean (NRCS) at Indore in the State of Madhya Pradesh to support soybean production systems research with basic technology and breeding material.

REFERENCES

1. Bapna, S.L., Seetharaman, S.P. and Pichholiya, K.R., Soybean system in India. Oxford & IBH Publishing Co., Pvt. Ltd., New Delhi. (1992).
2. Boran, D.K. and Bora, P.K., Predicting the monthly rainfall around Guwahati using a seasonal ARIMA model. *J. Indian Society of Agricultural Statistics*, **47:** 278-287 (1995).
3. Box, G.E.P. and Jenkins, G.M., Time series analysis: Forecasting and Control. 2nd Ed Holden-Day, San Francisco (1976).
4. Gupta, G.S., ARIMA Model forecasting on tea production in India. *The Indian Economic journal,* **41(2):** 88-110 (1993).
5. L-Jung, G.M. and Box, G.E.P., On a measure of lack of fit in time series models.*Biometrika*, **65:** 297-303 (1978).
6. Pathak, D.N. "Soyabean cultivation in India breaking the productivity"(http://www.ofievents.com/india/contentimages/advertising/D.N._Pathak_OFI_presentation.pdf).
7. Trikha, R.N. and Nave, R.W., SPRA's activities on soybean 1978-79, Soya Production and Research Association, Bareilly, (1979).
8. Venugopalan, R. and Prajneshu, Trend analysis in all India marine Products exports using statistical modeling techniques. *Indian J. of Fish,* **43(2):** 107-113 (1996).