Forecasting of Soybean Yield in India through ARIMA Model

Mahesh Kumar1*, Rohan Kumar Raman2 and Subhash Kumar1

1Dr. Rajendra Prasad Central Agricultural University, Pusa, Samastipur, Bihar-848125, India
2ICAR-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. 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 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.
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.

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

\[
\phi(B) \Delta^d Z_t = \Theta(B) a_t
\]

Where,

\[
\Delta^d = (1-B)^d
\]

\[
\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)
\]

\[
\Theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)
\]

\[
Z = Z_t - \mu
\]

\[
d = \text{Stationary time series data}
\]

\[
a_t = \text{Order of differencing}
\]

\[
\alpha_r = \text{Random shock}
\]

\[
p = \text{Order of auto-regression}
\]

\[
q = \text{Order of moving average}
\]

Under identification phase, the first order differencing \((r=1,2, \ldots)\) 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, 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 \(\chi^2\) (Chi-square) of the residual ACF.

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

ARIMA Model: ARIMA 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;

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

<table>
<thead>
<tr>
<th>Sl. No</th>
<th>Year</th>
<th>Area (Million hec.)</th>
<th>Production (Million Tonnes)</th>
<th>Yield Kg/hec</th>
<th>Sl. No</th>
<th>Year</th>
<th>Area (Million hec.)</th>
<th>Production (Million Tonnes)</th>
<th>Yield Kg/hec</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1970-71</td>
<td>0.03</td>
<td>426</td>
<td>24</td>
<td>24</td>
<td>1994-95</td>
<td>4.32</td>
<td>3.93</td>
<td>911</td>
</tr>
<tr>
<td>2</td>
<td>1971-72</td>
<td>0.03</td>
<td>426</td>
<td>25</td>
<td>24</td>
<td>1995-96</td>
<td>5.04</td>
<td>5.10</td>
<td>1012</td>
</tr>
<tr>
<td>3</td>
<td>1972-73</td>
<td>0.03</td>
<td>819</td>
<td>26</td>
<td>26</td>
<td>1996-97</td>
<td>5.44</td>
<td>5.38</td>
<td>989</td>
</tr>
<tr>
<td>4</td>
<td>1973-74</td>
<td>0.05</td>
<td>829</td>
<td>27</td>
<td>27</td>
<td>1997-98</td>
<td>5.99</td>
<td>6.46</td>
<td>1079</td>
</tr>
<tr>
<td>5</td>
<td>1974-75</td>
<td>0.07</td>
<td>768</td>
<td>28</td>
<td>28</td>
<td>1998-99</td>
<td>6.49</td>
<td>7.14</td>
<td>1100</td>
</tr>
<tr>
<td>6</td>
<td>1975-76</td>
<td>0.09</td>
<td>975</td>
<td>29</td>
<td>29</td>
<td>1999-00</td>
<td>6.22</td>
<td>7.08</td>
<td>1138</td>
</tr>
<tr>
<td>7</td>
<td>1976-77</td>
<td>0.13</td>
<td>988</td>
<td>30</td>
<td>30</td>
<td>2000-01</td>
<td>6.42</td>
<td>5.28</td>
<td>823</td>
</tr>
<tr>
<td>8</td>
<td>1977-78</td>
<td>0.20</td>
<td>940</td>
<td>31</td>
<td>31</td>
<td>2001-02</td>
<td>6.34</td>
<td>5.96</td>
<td>940</td>
</tr>
<tr>
<td>9</td>
<td>1979-80</td>
<td>0.50</td>
<td>568</td>
<td>32</td>
<td>32</td>
<td>2002-03</td>
<td>6.11</td>
<td>4.65</td>
<td>762</td>
</tr>
<tr>
<td>10</td>
<td>1980-81</td>
<td>0.61</td>
<td>728</td>
<td>33</td>
<td>33</td>
<td>2003-04</td>
<td>6.55</td>
<td>7.82</td>
<td>1193</td>
</tr>
<tr>
<td>11</td>
<td>1981-82</td>
<td>0.48</td>
<td>741</td>
<td>34</td>
<td>34</td>
<td>2004-05</td>
<td>7.57</td>
<td>6.87</td>
<td>908</td>
</tr>
<tr>
<td>12</td>
<td>1982-83</td>
<td>0.77</td>
<td>637</td>
<td>35</td>
<td>35</td>
<td>2005-06</td>
<td>7.71</td>
<td>8.27</td>
<td>1073</td>
</tr>
<tr>
<td>13</td>
<td>1983-84</td>
<td>0.84</td>
<td>735</td>
<td>36</td>
<td>36</td>
<td>2006-07</td>
<td>8.33</td>
<td>8.85</td>
<td>1063</td>
</tr>
<tr>
<td>14</td>
<td>1984-85</td>
<td>1.24</td>
<td>768</td>
<td>37</td>
<td>37</td>
<td>2007-08</td>
<td>8.98</td>
<td>10.87</td>
<td>1255</td>
</tr>
<tr>
<td>15</td>
<td>1985-86</td>
<td>1.34</td>
<td>764</td>
<td>38</td>
<td>38</td>
<td>2008-09</td>
<td>9.51</td>
<td>9.91</td>
<td>1041</td>
</tr>
<tr>
<td>16</td>
<td>1986-87</td>
<td>1.53</td>
<td>584</td>
<td>39</td>
<td>39</td>
<td>2009-10</td>
<td>9.73</td>
<td>9.96</td>
<td>1024</td>
</tr>
<tr>
<td>17</td>
<td>1987-88</td>
<td>1.54</td>
<td>582</td>
<td>40</td>
<td>40</td>
<td>2010-11</td>
<td>9.60</td>
<td>12.74</td>
<td>1327</td>
</tr>
<tr>
<td>18</td>
<td>1988-89</td>
<td>1.73</td>
<td>892</td>
<td>41</td>
<td>41</td>
<td>2011-12</td>
<td>10.11</td>
<td>12.21</td>
<td>1208</td>
</tr>
<tr>
<td>19</td>
<td>1989-90</td>
<td>2.25</td>
<td>801</td>
<td>42</td>
<td>42</td>
<td>2012-13</td>
<td>10.84</td>
<td>14.67</td>
<td>1353</td>
</tr>
<tr>
<td>20</td>
<td>1990-91</td>
<td>2.56</td>
<td>1015</td>
<td>43</td>
<td>43</td>
<td>2013-14</td>
<td>12.20</td>
<td>9.50</td>
<td>779</td>
</tr>
<tr>
<td>21</td>
<td>1991-92</td>
<td>3.18</td>
<td>782</td>
<td>44</td>
<td>44</td>
<td>2014-15</td>
<td>10.90</td>
<td>8.70</td>
<td>798</td>
</tr>
<tr>
<td>22</td>
<td>1992-93</td>
<td>3.79</td>
<td>894</td>
<td>45</td>
<td>45</td>
<td>2015-16</td>
<td>11.65</td>
<td>8.00</td>
<td>687</td>
</tr>
<tr>
<td>23</td>
<td>1993-94</td>
<td>4.37</td>
<td>475</td>
<td>1086</td>
<td>1086</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
a) **Autoregressive Model:** This can be represented as
\[ Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \alpha_t \]  
Where
\[ c = \mu (1 - \phi_1) = \text{Constant term} \]
\[ \mu = \text{Constant parameter} \]
\[ \phi = \text{Deterministic coefficient its value determines the relationship between } Z_t \text{ and } Z_{t-1} \text{ (Lagged observation)} \]
\[ \alpha_t = \text{Random shock having some continuous statistical distribution.} \]

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 \( \mu \) and \( \phi \) 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
\[ Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \alpha_t \]

In this model, \( Z \) is linearly related to the past observation \( Z_{t-1} \) and \( Z_{t-2} \). The least square estimate of \( \phi_1 \) and \( \phi_2 \) are approximated by
\[ \hat{\phi}_1 = \frac{r_1 (1-r_2)}{1-r_1^2} \quad \text{and} \quad \hat{\phi}_2 = \frac{r_2 - r_1^2}{1-r_1^2} \]

Where,
\[ r_1 \text{ & } r_2 \text{ are autocorrelation function for first and second lag respectively.} \]

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.
\[ Z_t = C + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + \alpha_t \]

b) **Moving Average (MA) Model:** A moving average model of order one denoted as MA (1) can be represented as
\[ Z_t = C - \Theta_1 \alpha_{t-1} + \alpha_t \]  
\[ \ldots (2) \]

Where,
\[ C = \mu (1 - \Theta_1) = \text{constant term} \]

\( \Theta_1 \) = Moving average coefficient determines the statistical relationship between \( Z \) and \( \alpha_{t-1} \) (lagged random shock)
\[ \alpha_t = \text{random shock with mean '0' and variance } \sigma^2. \]

**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 \( \mu \) and \( \Theta \) of Equation 2 i.e.
\[ Z_t = C - \Theta_1 \alpha_{t-1} + \alpha_t \]

residual sum of square (RSS) \( \sum a_i^2 \) in terms of observed \( Z \)'s and the parameters \( \mu \) and \( \Theta \) are obtained and then it is differentiated with respect to \( \mu \) and \( \Theta \) to obtain estimated \( \mu \) and \( \Theta \). 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 \( \mu \) and \( \Theta \) such as \( \mu = Z \) and \( \Theta \) given by the solution of Equation 3
\[ Z_t = C + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \Theta_1 \alpha_{t-1} \ldots + \Theta_q \alpha_{t-q} \ldots (3) \]

Then the RSS may be calculated recursively from
\[ a_t = Z_t - C + \Theta_1 \alpha_{t-1} \quad \text{with } a_0 = 0 \]

This procedure then can be repeated for a grid of points in (\( \mu, \Theta \)) plane. We may then by inspection choose that value of (\( \mu, \Theta \)) as estimates which minimized RSS. The least square estimates are also maximum likelihood estimated conditional on a fixed value of \( a_0 \) provided \( a_i \) 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
\[ Z_t = C + \phi_1 Z_{t-1} + \ldots + \phi_p Z_{t-p} + \Theta_1 \alpha_{t-1} \ldots + \Theta_q \alpha_{t-q} + \alpha_t \]

**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 \( \phi \)'s and \( \Theta \)’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 \( \chi^2 \) -test based on L-Jung and Box for the residual autocorrelation (L-Jung,G.M., Box, G.E.P). Gupta\(^1\) 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\(^3\) 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.

<table>
<thead>
<tr>
<th>ARIMA MODEL</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1,0)</td>
<td>597.75</td>
<td>601.08</td>
</tr>
<tr>
<td>(1,1,1)</td>
<td>599.4</td>
<td>608.49</td>
</tr>
<tr>
<td>(2,1,0)</td>
<td>601.3</td>
<td>608.53</td>
</tr>
<tr>
<td>(2,0,0)</td>
<td>611.35</td>
<td>618.66</td>
</tr>
<tr>
<td>(1,0,0)</td>
<td>612.89</td>
<td>618.37</td>
</tr>
<tr>
<td>(2,0,1)</td>
<td>612.71</td>
<td>621.85</td>
</tr>
<tr>
<td>(0,0,1)</td>
<td>620.45</td>
<td>625.93</td>
</tr>
</tbody>
</table>

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

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

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
Figure 2 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.

![Soybean Yield (Kg/ha)](image)

**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 (Figure 6) and PACF (Figure 7).

![ACF of soybean yield after first order differentiation](image)

**Fig. 6: ACF of soybean 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.

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. (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.

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