Pre-harvest forecast models for rapeseed & mustard yield using principal component analysis of weather variables

MOHD. AZFAR, B. V. S. SISODIA, V. N. RAI and MONIKA DEVI
Deptt. of Agril. Statistics, Narendra Deva University of Agril. & Tech. Kumarganj, Faizabad (U. P.) – 224 229
(Received 25 April 2014, Modified 27 October 2014)
e mail : mazfar38@gmail.com

ABSTRACT. The present paper deals with use of principal component analysis of weekly data on weather variables for developing rapeseed & mustard yield forecast model for Faizabad district of U. P. (India). Time series data on rapeseed & mustard yield and weekly data of six weather variables for the crop season for 22 years (1990-91 to 2011-12) have been used to develop weather indices. In all, six models have been developed and have been used to forecast yield for three subsequent years 2009-10 to 2011-12 (which were not included in model development). The model with six weather variables was found to be most appropriate to provide reliable yield forecast about one and half months before the harvest.

Key words – Weather variables, Weather indices, Principal component analysis and forecast model.

1. Introduction

Pre-harvest forecasting of crop yield is an important means of assessing total available food supplies and thereby providing early warning about the emerging food situation in a country. From the time of sowing to harvesting, a crop evolves through different growth stages and can reach its genetically determined yield potential only when all environmental and other input factors remain optimal during each phases of the growing cycle. There are several methods which can be applied to arrive at forecasts of food crop production. These are: (a) monitoring crop conditions on the basis of agro-climatic data; (b) making regular survey to assess area, yield and production of crop; and (c) estimating regression models describing quantitative relationship between selected weather variables and final yields of the crop. Various research workers in the recent past have attempted to develop forecast models for rice and wheat yield. Notably among them are Agrawal et al. (1986); Agrawal et al. (2012); Jain et al. (1980); Sisodia et al. (2014) etc. Jain et al. (1984) have developed forecast model for rice yield using principal component analysis of biometrical characters. The present paper is concerned with formulation of appropriate yield forecasting models using the principal component analysis of data on weather variables for obtaining pre-harvest forecast of rapeseed & mustard crop yield in Faizabad district of Uttar Pradesh.

2. Materials and statistical methodology

2.1. Area and crop covered

The present study is related to Faizabad district (Uttar Pradesh, India) which is situated between 26° 47' N latitude and 82° 12' E longitudes. It lies in the eastern plain zone of Uttar Pradesh. It has an annual rainfall of about 1002 mm and nearly 85% of total precipitation is received from south west monsoon during the months of July to September. However, occasional mild shower occur during winter season. The average minimum temperatures lies 18.6 °C and 31.3 °C. It is liberally sourced by the Saryu (Ghaghar) river and its tributaries. Soils are deep alluvial, medium to medium heavy textured but are easily ploughable. The favourable climate, soil and the availability of ample irrigation facility make growing of rapeseed & mustard a natural choice for the area. Rapeseed & mustard crop is generally cultivated during
the Rabi season because during this period it provides a better environment for the cultivation of this crop.

2.2. Sources and description of data

Time series data of rapeseed & mustard yield of Faizabad district of Uttar Pradesh for 22 years (1990-91 to 2011-12) have been used for development of the models. Data were collected from the Bulletins of Directorate of Agricultural Statistics and Crop Insurance, Govt. of Uttar Pradesh. Weekly weather data for the same period on six weather variables, viz., Minimum Temperature, Maximum Temperature, Relative Humidity at 7 and 14 hours IST, Wind-Velocity and Sun-shine hours have been used in the study. These data were obtained from the Department of Agro-meteorology, N.D. University of Agriculture & Technology Kumarganj, Faizabad, U. P., India.

2.3. Crop season

Preparation for sowing of rapeseed & mustard starts from the first week of October in Faizabad districts and harvested in the month of April. The entire crop season has been divided broadly into four phases. The first phase includes preparation, sowing, emergence and initial growth stages of the crop which covers about 6 weeks from October 1 (40th SMW) to November 11 (45th SMW). The second phase includes vegetative growth stage which covers about 7 weeks from Nov. 12 (46th SMW) to December 31 (52nd SMW). The third phase includes flowering, reproduction and pod formation stages which covers about 8 weeks from January 1 (1st SMW) to February 25 (8th SMW). The fourth phase includes the ripening and harvesting stage of crop which covers about 7 weeks from February 26 (9th SMW) to April 15 (15th SMW). Therefore, the entire data on weather variables for 28 weeks of the crop production period from 40th SMW to 15th SMW of the subsequent year has been utilized for constructing weighted and unweighted weather indices of weather variables along with their interactions. The weighted indices are weighted average of the weather variables over weeks, weights being the correlation coefficients between the de-trended yield and the weekly data on respective weather variable. The unweighted indices are the simple average of the weather variables over the weeks. Similarly, the unweighted and weighted indices of interactions between the weather variables have been obtained using product of weather variables (taking two at a time). In all 42 indices (21 weighted and 21 unweighted) consisting of 6 weighted weather indices and 15 weighted interaction indices; 6 unweighted and 15 unweighted interaction indices have been obtained. These weather indices and interaction indices have been computed by using the following formula.

\[
Z_{ij} = \frac{\sum_{w=1}^{n} r_{iw}^{j} X_{iw}}{\sum_{w=1}^{n} r_{iw}^{j}} \text{ (A)}
\]

\[
Z_{ii',j} = \frac{\sum_{w=1}^{n} r_{ii'w}^{j} X_{iw} X_{i'w}}{\sum_{w=1}^{n} r_{ii'w}^{j}}, j = 0, 1 \text{ and } i = 1, 2, \ldots, p \text{ (B)}
\]

where, \(Z_{ij}\) is unweighted (for \(j = 0\)) and weighted (for \(j = 1\)) weather indices for \(i^{th}\) weather variable and \(Z_{ii',j}\) is the unweighted (for \(j = 0\)) and weighted (for \(j = 1\)) weather indices for interaction between \(i^{th}\) and \(i'^{th}\) weather variables. \(X_{iw}\) is the value of the \(i^{th}\) weather variable in \(w^{th}\) week, \(r_{ij}/r_{ii'w}\) is correlation coefficient of yield adjusted for trend effect with \(i^{th}\) weather variable/product of \(i^{th}\) and
\(i^{th}\) weather variable in \(w^{th}\) week, \(n\) is the number of weeks considered in developing the indices and \(p\) is number of weather variables. Models are developed using simple regression analysis as given below:

\textbf{Model - 1}

In this model, unweighted weather indices of six weather variables have been used in principal component analysis. The analysis has identified first three components \(PC_1, PC_2 & PC_3\) as most significant ones as per loading and have explained over 79.06 per cent variance of the total variance. Hence, these first three principal components have been used as regressors in the development of forecasting model. The form of model fitted is as follows:

\[ Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \delta T + e \tag{1} \]

where \(Y\) is the crop yield, \(\beta_i^i (i = 0,1,2,3)\) and \(\delta\) are model parameter, \(PC_1, PC_2 & PC_3\) are principal components, \(T\) is the trend variable and \(e\) is error term assumed to follow normal distribution with mean 0 and variance \(\sigma^2\).

\textbf{Model - 2}

In this model, weighted weather indices of six weather variables have been used in principal component analysis. It has identified first principal component as most significant ones as per loading and have explained over 60.83 per cent of the total variance. Hence, only first principal component has been used as regressors in the development of forecasting model. The form of model fitted is as follows:

\[ Y = \beta_0 + \beta_1 PC_1 + \delta T + e \tag{2} \]

where, the notations are described in model-1.

\textbf{Model - 3}

In this model, all 42 weather indices (including interaction indices) of six weather variables have been used in principal component analysis. The first six components were most significant ones and have explained over 94.02 per cent of the total variance. Hence, these have been used as regressors in the model-3.

\textbf{Model - 4}

In this model, weighted and unweighted weather indices of six weather variables have been used in principal component analysis. The first four principal components were most significant ones and explained over 86.26 per cent of the total variance. Hence, these have been used as regressors in the model-4.

\textbf{Model - 5}

Here unweighted and unweighted interaction weather indices of six weather variables have been used in the analysis. The first five principal components were significant and explained over 97.41 per cent variance and were used as regressors in the model-5.

\textbf{Model - 6}

In this model, weighted and weighted interaction weather indices of six weather variables have been used. The principal component analysis has identified first five as most significant ones and explained over 93.81 per cent of the total variance. Hence, these first five principal components have been used as regressors here.

The models 3, 4, 5 & 6 have almost similar form depending on the number of PC’s identified as significant.

\textbf{2.5. Comparison and validation of forecast models}

Different procedures have been used for the comparison and the validation of the models developed. These procedures are given below:

\textbf{2.5.1. \(R^2\) (Coefficient of Determination)}

\(R^2\) is given by the following formula

\[ R^2 = 1 - \frac{ss_{res}}{ss_t} \]

where, \(ss_{res}\) and \(ss_t\) are the residual sum of square and the total sum of square respectively.

\(R_{adj}^2\) is computed as:

\[ R_{adj}^2 = 1 - \frac{ss_{res} / (n - p)}{ss_t / (n - 1)} \]
TABLE 1
Forecast Models of Rapeseed & Mustard Yield

<table>
<thead>
<tr>
<th>Model</th>
<th>Forecast regression models</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y = 4.934 + 0.843^{*} PC_1 - 0.173 PC_2 - 0.570 PC_3 + 0.273 T^{**}$</td>
</tr>
<tr>
<td>2</td>
<td>$Y = 5.224 - 1.482 PC_1 + 0.235 T$</td>
</tr>
<tr>
<td>3</td>
<td>$Y = 5.570 - 1.268 PC_1 + 0.245 PC_2 - 0.424 PC_3 + 0.518^{*} PC_4 - 0.05 PC_5 + 0.327 PC_6 + 0.208 T^{**}$</td>
</tr>
<tr>
<td>4</td>
<td>$Y = 5.205 - 1.303 PC_1 - 0.00003269 PC_2 - 0.547^{*} PC_3 + 0.240 T$</td>
</tr>
<tr>
<td>5</td>
<td>$Y = 4.731 - 0.952^{*} PC_1 + 0.128 PC_2 - 0.464 PC_3 + 0.255 PC_4 - 0.543 PC_5 + 0.283 T^{**}$</td>
</tr>
<tr>
<td>6</td>
<td>$Y = 5.354 - 1.496 PC_1 + 0.114 PC_2 - 0.182 PC_3 + 0.160 PC_4 + 0.265 PC_5 + 0.225 T$</td>
</tr>
</tbody>
</table>

Significant at *P < 0.05, **P < 0.01

TABLE 2
Actual & Forecasts Yield of Rapeseed & Mustard (Q/ha)

<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Actual yield (Q/ha)</th>
<th>Predicted yield (Q/ha)</th>
<th>Percent deviation</th>
<th>Percent standard error</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2009-10</td>
<td>7.79</td>
<td>10.47</td>
<td>34.42</td>
<td>8.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.94</td>
<td>4.55</td>
<td>8.63</td>
<td></td>
<td>63.1</td>
<td>52.6</td>
</tr>
<tr>
<td>1</td>
<td>2011-12</td>
<td>6.81</td>
<td>9.40</td>
<td>38.00</td>
<td>11.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2009-10</td>
<td>7.79</td>
<td>9.41</td>
<td>20.74</td>
<td>4.46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.27</td>
<td>10.99</td>
<td>5.02</td>
<td>86.1</td>
<td>84.4</td>
<td>1.29</td>
</tr>
<tr>
<td>2</td>
<td>2011-12</td>
<td>6.81</td>
<td>7.85</td>
<td>15.22</td>
<td>6.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2009-10</td>
<td>7.79</td>
<td>9.58</td>
<td>22.95</td>
<td>5.68</td>
<td></td>
<td>89.7</td>
<td>83.1</td>
</tr>
<tr>
<td>3</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.82</td>
<td>5.66</td>
<td>3.96</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2011-12</td>
<td>6.81</td>
<td>7.90</td>
<td>15.97</td>
<td>4.70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2009-10</td>
<td>7.79</td>
<td>10.16</td>
<td>30.49</td>
<td>6.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.85</td>
<td>5.34</td>
<td>6.11</td>
<td>84.6</td>
<td>78.7</td>
<td>1.62</td>
</tr>
<tr>
<td>4</td>
<td>2011-12</td>
<td>6.81</td>
<td>8.18</td>
<td>20.19</td>
<td>9.62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2009-10</td>
<td>7.79</td>
<td>9.14</td>
<td>17.29</td>
<td>12.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.17</td>
<td>11.89</td>
<td>10.67</td>
<td>68.2</td>
<td>52.3</td>
<td>1.42</td>
</tr>
<tr>
<td>5</td>
<td>2011-12</td>
<td>6.81</td>
<td>8.44</td>
<td>23.98</td>
<td>15.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2009-10</td>
<td>7.79</td>
<td>9.74</td>
<td>24.99</td>
<td>6.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2010-11</td>
<td>10.41</td>
<td>9.56</td>
<td>8.12</td>
<td>6.39</td>
<td>90.3</td>
<td>85.4</td>
<td>1.29</td>
</tr>
<tr>
<td>6</td>
<td>2011-12</td>
<td>6.81</td>
<td>7.51</td>
<td>10.13</td>
<td>8.90</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
where, \( n \) and \( p \) are the number of observations and number of regressor variables, respectively.

### 2.5.2. Percent deviation

This measures the deviation (in percentage) of forecast from the actual yield data.

\[
\text{percentage deviation} = \left( \frac{\text{actual yield} - \text{forecasted yield}}{\text{actual yield}} \right) \times 100
\]

### 2.5.3. Percent Standard Error of the forecast

Let \( \hat{y}_f \) be forecast value of crop yield and \( X_0 \) be the vector of selected value of regressor variables for which the yield is forecasted.

The variance of \( \hat{y}_f \) as given in Draper and Smith (1998) is obtained as

\[
V(\hat{y}_f) = \sigma^2 X_0 (X'X)^{-1} X_0
\]

where, \( X'X \) is the dispersion matrix of the sum of square and cross products of regressors variables used for the fitting the model and \( \sigma^2 \) is the estimated residual variance.

The percent standard error (PSE) of forecast yield \( \hat{y}_f \) is given by

\[
\text{PSE} = \frac{\sqrt{V(\hat{y}_f)}}{\text{Forecast yield}} \times 100
\]

In fact, the PSE is the coefficient of variation (C.V.) of the forecast yield.

### 2.5.4. Root Mean Square Error (RMSE)

It is also a measure of comparing two models and is given below:

\[
\text{RMSE} = \left( \frac{1}{n} \sum_{i=1}^{n} (O_i - E_i)^2 \right)^{\frac{1}{2}}
\]

\( O_i \) and the \( E_i \) are the observed and forecasted value of the crop yield respectively and \( n \) is the number of years for which forecasting has been done.

### 3. Results and conclusion

The forecast models developed under the six strategies are given in the Table 1. In only three models, the time trend \( T \) has been found to be significant. Based on these, the forecast yields for the seasons 2009-10, 2010-11 and 2011-12 have been computed (Table 2). It is evident from these results, model-3 is the most appropriate one followed by the models-2 and 6 for the pre-harvest forecast of the rapeseed & mustard yield one and half months before the harvest of crops in Faizabad district of Uttar Pradesh.

It may be seen from the Table 2 that the actual yield during the year 2010-11 was substantially high as against the years 2009-10 and 2011-12. This might be because of the following factors.

(i) Good rain (31.7mm) during third week of October (42 SMW) in the year 2010-11, whereas there was no rain in the years 2009-10 and 2011-12.

(ii) There was heavy rain of about 59.88 mm during first week of January 2012 (flowering etc. stage) followed by some rains in the two subsequent weeks. Similarly, there were some rain of about 6 mm during second week of January 2010. Rain might have affected pollination during 2009-10 and 2011-12.

(iii) The ranges of minimum temperature during vegetative growth stage (2\textsuperscript{nd} phase) and flowering/reproduction stage (3\textsuperscript{rd} phase) were almost similar during the years of forecast.

Acknowledgement

The authors are highly grateful and thankful to the reviewer of the paper for his fruitful and constructive comments and suggestions which have improved the manuscript of the paper in its present form.

### References


