551.506.1 : 631.153.028

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

सार – प्रस्तुत शोध पत्र उत्तर प्रदेश के फैजाबाद जिले के लाही और सरसों की उपज के लिए पूर्वानुमान मॉडल के विकास हेतु मौसम परिवर्तिताओं के साप्ताहिक आँकड़ों के प्रमुख घटक विश्लेषण के उपयोग के बारे में है। लाही और सरसों की उपज और छः मौसम परिवर्तिताओं में 22 साल (1990-91 से 2011-12) के समय श्रृंखला आँकड़ों का उपयोग मौसम सूचकाँक को विकसित करने के लिए किया गया है। कुल छः मॉडल विकसित किए गए हैं और विकसित मॉडलों का उपयोग कर तीन वर्षों अर्थात 2009-10 से 2011-12 के लिए उपज के पूर्वानुमान निकाले गए। छः मौसम परिवर्तिताओं वाला मॉडल फसल की कटाई के लगभग डेढ माह पूर्व उपज का विश्वसनीय पूर्वानुमान देने में सबसे उपयुक्त पाया गया।

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 (Ghaghara) 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 weekly data on weather variables have been collected for 28 weeks of the crop production which included 40th SMW that starts from 1st October to 52nd SMW of a year and 1st SMW to 15th SMW of the subsequent year which ends by the second week of April.

2.4. *Statistical methodology*

2.4.1. *Principal component analysis*

Principal component analysis primarily deals with explaining the variance and covariance structure through linear combinations of original variables. The objectives are (1) data reduction (2) interpretation.

The basic theory of principal component analysis is available in many standard books on multivariate analysis (Anderson, 1984; Johnson & Wichern, 2001 etc.). So, the theoretical concept of this technique is not presented here. Let PC_1, PC_2,..., PC_k be first k (k< p) principal components explaining variability up to about 90 percent. Using these k principal components as regressors in the regression models instead of original p variables and crop yield (y) as regressand, the forecasting models have been developed. This technique reduces the number of regressors to be used in the model and hence even for small set of observations (n) the forecasting model can be developed with reasonable precision.

2.4.2. *Development of the forecast model*

The entire 21 weeks data from 40th SMW to 8th week of the next year have 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.

$$Z_{ij} = \sum_{w=1}^{n} r_{iw}^{j} X_{iw} \Bigg/ \sum_{w=1}^{n} r_{iw}^{j} \qquad \text{(A)}$$

$$Z_{ii',j} = \sum_{w=1}^{n} r_{ii'w}^{j} X_{iw} X_{i'w} \Bigg/ \sum_{w=1}^{n} r_{ii'w}^{j}, \; j = 0, 1 \text{ and } i = 1, 2, \ldots, p \qquad \text{(B)}$$

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. $\boldsymbol{X_{iw}}$ is the value of the i^{th} weather variable in w^{th} week, $r_{iw}/r_{ii'w}$ is correlation coefficient of yield adjusted for trend effect with i^{th} weather variable/product of i^{th} and

$i^{\text{'th}}$ 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:

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:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \beta_3 PC_3 + \delta T + e \qquad (1)$$

where Y is the crop yield, $\beta_i^{'s} (i = 0,1,2,3)$ and δ 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 σ^2.

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:

$$Y = \beta_0 + \beta_1 PC_1 + \delta T + e \qquad (2)$$

where, the notations are described in model-1.

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.

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.

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.

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.

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:

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.

Adjusted R^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

Model	Forecast regression models
1	$Y = 4.934 + 0.843^{*} PC_1 - 0.173 PC_2 - 0.570 PC_3 + 0.273 T^{**}$
2	$Y = 5.224 - 1.482 PC_1 + 0.235 T$
3	$Y = 5.570 - 1.268 PC_1 + 0.245 PC_2 - 0.424 PC_3 + 0.518^{*} PC_4 - 0.05 PC_5 + 0.327 PC_6 + 0.208 T^{**}$
4	$Y = 5.205 - 1.303 PC_1 - 0.00003269 PC_2 - 0.547^{*} PC_3 + 0.190 PC_4 + 0.240 T$
5	$Y = 4.731 - 0.952^{*} PC_1 + 0.128 PC_2 - 0.464 PC_3 + 0.255 PC_4 - 0.543 PC_5 + 0.283 T^{**}$
6	$Y = 5.354 - 1.496 PC_1 + 0.114 PC_2 - 0.182 PC_3 + 0.160 PC_4 + 0.265 PC_5 + 0.225 T$

Significant at $^{*}P < 0.05$, $^{**}P < 0.01$

TABLE 2

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

Model	Year	Actual yield (Q/ha)	Predicted yield (Q/ha)	Percent deviation	Percent standard error	R^2	Adjusted R^2	RMSE
1	2009-10	7.79	10.47	34.42	8.17	63.1	52.6	2.17
	2010-11	10.41	9.94	4.55	8.63			
	2011-12	6.81	9.40	38.00	11.61			
2	2009-10	7.79	9.41	20.74	4.46	86.1	84.4	1.29
	2010-11	10.41	9.27	10.99	5.02			
	2011-12	6.81	7.85	15.22	6.59			
3	2009-10	7.79	9.58	22.95	5.68	89.7	83.1	1.26
	2010-11	10.41	9.82	5.66	3.96			
	2011-12	6.81	7.90	15.97	4.70			
4	2009-10	7.79	10.16	30.49	6.74	84.6	78.7	1.62
	2010-11	10.41	9.85	5.34	6.11			
	2011-12	6.81	8.18	20.19	9.62			
5	2009-10	7.79	9.14	17.29	12.21	68.2	52.3	1.42
	2010-11	10.41	9.17	11.89	10.67			
	2011-12	6.81	8.44	23.98	15.59			
6	2009-10	7.79	9.74	24.99	6.72	90.3	85.4	1.29
	2010-11	10.41	9.56	8.12	6.39			
	2011-12	6.81	7.51	10.13	8.90			

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.

$$percentage\ deviation = \frac{(actual\ yield - forecasted\ yield)}{actual\ yield} \times 100$$

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

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

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 $\hat{\sigma}^2$ is the estimated residual variance.

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

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

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:

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

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 (2nd phase) and flowering/ reproduction stage (3rd 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

Agrawal, R., Jain, R. C. and Jha, M. P., 1986, "Models for studying rice crop weather relationship", *Mausam*, **37**, 1, 67-70.

Agrawal, R., Chandrahas and Aditya, K., 2012, "Use of discriminant function analysis for forecasting crop yield", *Mausam*, **36**, 3, 455-458.

Anderson, T. W., 1984, "An Introduction to Multivariate Statistical Analysis", John Wiley & Sons Inc., New York.

Draper, N. R. and Smith, H., 1998, "Applied Regression Analysis", 3rd edition, John Wiley & Sons Inc., New York.

Johnson, R. A. and Wichern, D. W., 2001, "Applied Multivariate Statistical Analysis Third Edition", Prentice Hall of India Private Limited, New Delhi.

Jain, R. C., Agrawal, Ranjana and Jha, M. P., 1980, "Effect of climatic variables on rice yield and its forecast", *Mausam*, **31**, 4, 591-96.

Jain, R. C., Sridharan, H. and Agrawal, R., 1984, "Principal component technique for forecasting of sorghum yield", *Indian J. Agric. Sci*., **54**, 6, 467-470.

Sisodia, B. V. S., Yadav, R. R., Kumar, Sunil and Sharma, M. K., 2014, "Forecasting of Pre-harvest crop yield using discriminant function analysis of meteorological parameters*", Journal of Agro-meteorology*, **16**, 1, 121-125.