MAUSAM, **65**, 3 (July 2014), 357-360

631.153.028 : 551.506.1

Application of principal component analysis in developing statistical models to forecast crop yield using weather variables

R. R. YADAV, B. V. S. SISODIA and SUNIL KUMAR

Narendra Deva University of Agriculture and Technology, Narendra Nagar, Faizabad - 224 229 (U.P.), India

(*Received 4 March 2013, Modified 14 May 2013*)

e mail : bvssisodia@gmail.com; rajaramy22@gmail.com

सार – प्रमुख अवयव विश्लेषण विधि का प्रयोग कर फसल की पैदावार के पूर्वानुमान हेतु सांख्यिकीय निदर्शों के विकास का प्रयत्न किया गया है। उत्तर प्रदेश के फैजाबाद जनपद में गेहूँ की उत्पादकता के समय श्रृंखला आँकड़ों एवं साप्ताहिक परिवर्तिता जैसे न्यूनतम एवं अधिकतम तापमान,सापेक्ष आर्द्रता वायु-गति एवं धूप के घंटे के वर्ष 1990 से 2010 के आँकड़ों का प्रयोग इस अध्ययन में किया गया है। इन मौसम परिवर्तिताओं के साप्ताहिक आँकड़ों के आधार पर विभिन्न सूचकांक तैयार किए गए हैं (अग्रवाल आदि, 1983)। सूचकांक के विभिन्न समूहों का प्रयोग करते हुए प्रमुख अवयव विश्लेषण किया गया तथा तत्पश्चात चार निदर्शों का विकास किया गया। निदर्शों के विकास में प्रमुख अवयवों एवं समय प्रवृत्ति को रिग्रेसर के रूप में प्रयोग किया गया है। गेहूँ की उत्पादकता का पूर्वानुमान फसल की कटाई के दो माह पूर्व लगाने में R^2 adj, पूर्वानुमान का प्रतिशत विचलन, RMSE (%) एवं PSE के आधार पर निदर्श-। एवं 3 को सबसे अधिक उपयुक्त पाया गया है ।

ABSTRACT. Application of principal component analysis in developing statistical models for forecasting crop yield has been demonstrated. The time series data on wheat yield and weekly weather variables, *viz*., Minimum and maximum temperature, Relative Humidity, Wind- Velocity and Sun-Shine hours pertaining to the period 1990 to 2010 in Faizabad district of Uttar Pradesh have been used in this study. Weather indices have been constructed using weekly data on weather variables (Agrawal *et al.*, 1983). Four models have been developed using principal component analysis as regressor variables including time trend and wheat yield as regressand. The model 1 and 3 have been found to be most appropriate on the basis of R^2adj, percent deviation of forecast, RMSE (%) and PSE for the forecast of wheat yield two months before the harvest of the crop.

Key words – Principal components, Weather indices, forecast models.

1. Introduction

Pre harvest forecast of the crop production at suitable stages of crop period before the harvest is vital for advance policy formulation in regards to crop procurement, distribution, price structure and import/ export decisions etc. These are useful to farmers to decide in advance their future prospects and possible course of action. Thus, reliable and timely pre-harvest forecasting of crop yield is very important. Various research workers have made efforts in the past to develop statistical models based on time series data on crop-yield and weather variables for pre-harvest forecasting of crop yield. Notably among them are Fisher (1924), Hendricks and Scholl (1943), Agrawal *et al*. (1980, 83, 86 & 2001), Jain and Singh (1980) etc. using regression models. Application of discriminant function analysis of weather indices and weekly data of weather variables for development of statistical models to forecast crop yield has also been attempted by Rai and Chandrahas (2000), Agrawal *et al*. (2012). The results obtained by application of discriminant function analysis have been quite encouraging. Forecast models based on principal components of biometrical characters have also been developed by Aneja and Chandrahas (1984), Chandrahas and Prem Narain (1993), Jain *et al*. (1985) etc. In the present paper, an attempt has been made to develop suitable statistical models for forecasting of pre-harvest wheat yield in Faizabad district of Uttar Pradesh using principal component analysis of weather indices of weather variables.

2. Materials and methodology

2.1. *Area and crop covered*

The study has been conducted for Faizabad district of Eastern Uttar Pradesh, India, which is situated at 26° 47' N latitude / 82° 12' E longitudes. It lies in the Eastern plain zone of Uttar Pradesh with an annual rainfall of 1002 mm and is sourced by the Saryu (Ghaghra) River and its tributaries. Soils are deep alluvial, medium to medium heavy textured but are easily ploughable. The favorable climate, soil and the availability of ample

irrigation facilities make growing of rice and wheat a natural choice for the area. The objective is to develop pre-harvest forecast model for wheat yield.

2.2. *Data*

Time series data on yield for wheat crop of Faizabad district of Uttar Pradesh for 20 years (1991-2010) have been collected from the bulletins of Directorate of Agricultural Statistics and Crop Insurance, Govt. of Uttar-Pradesh. Weekly weather data for the period (1991-2010) on the weather variables of Faizabad district of Uttar Pradesh during the different growth phases of wheat crop have been obtained from the Department of Agrometeorology, N. D. University of Agriculture & Technology Kumarganj, Faizabad. Preparation for sowing of wheat starts from the end of October in Faizabad districts and its harvesting starts from the first week of April. Therefore, the data have been collected for 15 weeks of the crop production which included 44th Standard Meteorological Week (SMW) that starts from 29th October to 52nd SMW of a year and 1st SMW to 6th SMW of the next year which falls during the first week of February. That means the forecasting has to be made before two months of the harvest of the crop. The data on five weather variables, *viz*., Minimum Temperature, Maximum Temperature, Relative Humidity, Wind-Velocity and Sun-shine hours have been used in the study.

2.3. *Statistical methodology*

The primary objective of this study is to develop suitable statistical models for forecasting pre-harvest yield of wheat crop using principal component analysis. Method of principal component analysis is available in many standard books on multivariate analysis, *viz*., Johnson and Wichern (2001) etc.

2.3.1. *Development of the forecast model*

The entire 15 weeks data from 44th SMW to 52nd SMW of a year and 1st SMW to 6th SMW 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 respective weather variable. The un-weighted 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 30 indices (15 weighted and 15 unweighted) consisting of 5 weighted weather indices and 10 weighted interaction indices; 5 unweighted weather indices and 10 unweighted interaction indices have been obtained. These weather indices and interaction indices have been computed by using the following formula.

$$Z_{ij} = \frac{\sum_{w=1}^{n} r_{iw}^{j} X_{iw}}{\sum_{w=1}^{n} r_{iw}^{j}} \qquad Z_{ii',j} = \frac{\sum_{w=1}^{n} r_{ii',w}^{j} X_{iw} x_{i'w}}{\sum_{w=1}^{n} r_{ii'w}^{j}}$$

$j = 0, 1$ and $i = 1, 2, \ldots, p$

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 un-weighted (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, $r_{iw}/r_{ii'w}$ is correlation coefficient between yield adjusted for trend effect and value of 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 used. Models are developed using simple regression analysis

Model 1: In this procedure, all 30 indices have been used in principal component analysis and first six principal components have been used as regressors in the development of forecasting model because these principal components have explained 90.44 percent of total variance. The form of the model fitted is as follows:

$$y = \beta_0 + \beta_1 pc_1 + \beta_2 pc_2 + \beta_3 pc_3 + \beta_4 pc_4 + \beta_5 pc_5 + \beta_6 pc_6 + \beta_7 T + e$$

where, y is un-trended crop yield, β_i's (i = 0, 1, 2, …7) are model parameters; pc_1, pc_2, …, pc_6 are first six principal components, T is the trend variable and e is error term assumed to follow independently N (0, σ^2).

Model 2: In this procedure, five weighted and five unweighted weather indices of five weather variables have been used. The principal component analysis has identified first three principal components as most significant ones as per loading and have explained over 75 per cent 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:

$$y = \beta_0 + \beta_1 pc_1 + \beta_2 pc_2 + \beta_3 pc_3 + \beta_4 T + e$$

where, the notations stand as usual as described in model 1.

TABLE 1

Wheat yield forecast models

Model	Forecast regression equation	R^2(%)	R^2_{adj}(%)
1.	Yield = 22.57 - 0.134pc_1 + 0.846*pc_2 - 0.595pc_3 - 0.510pc_4 - 0.806*pc_5 - 0.105pc_6+ 0.294T (1.204) (0.643) (0.297) (0.287) (0.268) (0.264) (0.262) (0.116)	88.6	79.8
2.	Yield = 22.69 - 0.203pc_1 + 0.574pc_2 – 1.063*pc_3 + 0.272T (1.400) (0.626) (0.509) (0.389) (0.137)	75.2	66.9
3.	Yield = 22.03 - 0.903*pc_1 + 0.458pc_2 - 0.256pc_3 - 0.983**pc_4 + 0.348**T (0.811) (0.306) (0.293) (0.373) (0.279) (0.082)	85.1	78.3
4.	Yield = 23.33 - 0.483pc_1 + 0.602pc_2 - 0.725pc_3 - 0.560pc_4 + 0.221T (1.296) (0.732) (0.337) (0.351) (0.343) (0.121)	77.3	67.0

Note: Figures in brackets denote standard error of regression coefficients.*P < 0.05,**P < 0.01

Model 3: In this procedure, 5 unweighted weather indices and 10 unweigthed interactions have been used. First four principal components have been taken as regressors in the forecasting model as these have explained about 95.36 percent of total variance. The form of the model fitted is as follows:

$$y = \beta_0 + \beta_1 pc_1 + \beta_2 pc_2 + \beta_3 pc_3 + \beta_4 pc_4 + \beta_5 T + e$$

where, the notations stand as usual as described in model 1.

Model 4: In this procedure, 5 weighted weather indices and 10 weighted interactions have been used. First four principal components have been used as regressors in forecasting model because these have explained about 85 percent of total variance. The form of the model fitted is as follows:

$$y = \beta_0 + \beta_1 pc_1 + \beta_2 pc_2 + \beta_3 pc_3 + \beta_4 pc_4 + \beta_5 T + e$$

where, the notations stand as described in model 1.

All the aforesaid models have been fitted with the data pertaining to the years 1990-91 to 2006-07 and the data pertaining to the year 2007-08 to 2009-10 were used for validation of the forecast models.

Comparison and validation of forecast models

Different procedures have been used in the present study for the comparison and the validation of the models developed. These procedures are given bellow:

(*i*) R_{adj}^2*:* The models were compared on the basis of adjusted coefficient of determination (R_{adj2}) is as follows:

$$R^2_{adj} = 1 - \frac{ss_{res}/(n-p)}{ss_t/(n-1)}$$

where, $ss_{res}/(n - p)$ is the residual mean square and $ss_t/(n - 1)$ is the total mean square.

(*ii*) The percent deviation of forecast from actual have been computed by the following formula:

$$\text{Percentage deviation} = \frac{\text{Actual yield} - \text{Forecast yield}}{\text{Actual yield}} \times 100$$

(*iii*) *Root Mean Square Error (RMSE)*

It is also a measure for comparing two models. The formula of RMSE is given bellow:

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

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.

(*iv*) *Percent standard error of the forecast*

Let $\hat{y}_f$ be forecast value of crop yield and X_0 be the column vector of values of P independent variables at which y is forecasted then variance of $\hat{y}_f$ is given by (Draper and Smith, 1998).

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

where, $(X'X)$ is the matrix of the sum of square and cross products of regressors matrix X (independent variables) and $\hat{\sigma}^2$ is the estimated residual variance of the model. Therefore, the percent standard error (cv) of forecast is given by

$$\text{Percent S. E.} = \frac{\sqrt{V(\hat{y}_f)}}{\text{forecast value}} \times 100$$

TABLE 2

Actual & forecasts of wheat yield (Q/ha)

Year	Actual yield	Forecasted yield			
		Model-1	Model-2	Model-3	Model-4
2007-08	30.19	27.55	26.48	28.14	26.22
		(8.74)	(12.28)	(6.79)	(13.15)
2008-09	30.82	27.60	26.56	27.45	27.01
		(10.44)	(13.82)	(10.93)	(12.36)
2009-10	28.32	28.48	27.38	27.67	27.74
		(0.25)	(3.31)	(2.29)	(2.04)
	RMSE (%)	3.378	3.755	3.378	3.032
2007-08	PSE (CV)	3.29	3.77	2.82	4.68
2008-09	PSE (CV)	4.33	5.18	4.57	4.52
2009-10	PSE (CV)	3.19	3.55	3.08	4.08

Note : Figures in brackets denote % deviation of forecast,
C V : Coefficient of variation

3. Results and discussion

Forecast models are presented in Table 1 along with their values of R^2_{adj}. In model 1, second and fifth principal components have shown significant effect on wheat yield. Only third principal component has shown significant effect in model 2. In model 3, first and fourth principal components including time trend (*T*) have shown significant effect while none of the principal components have shown significant effects on wheat yield in model 4. The value of R^2_{adj} has been found to be maximum of about 80 percent in model 1. Followed by about 78 per cent in model 3. Using these forecast models the forecast values of wheat yield for the years 2007-08, 2008-09 and 2009-10 were obtained and the results are presented in Table 2. The perusal of Tables (1 and 2) reveal that the model 1 is the most appropriate one followed by the model 3 for the pre-harvest forecast of the wheat yield two months before the harvest of the crops in Faizabad district of U. P. The values of R^2_{adj} for the models have not been found to be so high in comparison to the models developed by an application of discriminant function analysis (Agrawal *et al*., 2012) but taking into account the percent deviation of forecast, RMSE (%) and PSE of the model 1 and 3, these two models have relatively performed well and can be recommended for the forecast of the wheat yield two months before the harvest of the crop.

Acknowledgement

The authors are very much thankful to the referee for his valuable comments and suggestions which have improved the earlier version of the paper.

References

Agrawal, R., Jain, R. C., Jha, M. P. and Singh, D., 1980, "Forecasting of rice yield using climatic variables", *Ind. J. Agric. Sci*., **50**, 9, 680-684.

Agrawal, R., Jain, R. C. and Jha, M. P., 1983, "Joint effects of weather variables on rice yields", *Mausam*, **34**, 2, 189-194.

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

Agrawal, R., Jain, R. C. and Mehta, S. C., 2001, "Yield forecast based on weather variables and agricultural input on agro-climatic zone basis", *Ind. J. Agric. Sci*., **71**, 7, 487-490.

Agrawal, R., Chandrahas and Aditya, K., 2012, "Use of discriminant function analysis for forecasting crop yield", *Mausam*, **63**, 3, 455-458.

Aneja, K. G. and Chandrahas, 1984, "Preharvest crop yield forecast based on plant biometrical characters", *Silver jubilee souvenir*, IASRI, New Delhi.

Chandrahas and Narain, Prem, 1993, "Pilot studies on preharvest forecasting of apple yield on the basis of data on biometrical characters, weather factors and crop inputs in Shimla District (H. P.) during 1984-86", IASRI, New Delhi.

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

Fisher, R. A., 1924, "The influence of rainfall on the yield of wheat at Rothamsted", *Roy. Soc. (London)*, *Phil. Trans. Ser. B*., **213**, 89-142.

Hendricks, W. A. and Scholl, G. C., 1943, "Technique in measuring joint relationship: The joint effects of temperature and precipitation on crop yield", *N. Carolina Agric. Exp. Stat. Tech. Bull*., **74**.

Jain, R. C. and Singh, D., 1980, "Forecasting rainfall over Puerto Rico. Annual Report, Department of Meteorology", The Florida State University.

Jain, R. C., Jha, M. P. and Agrawal, Ranjana, 1985, "Use of growth indices in yield forecast", *Biometrical Journal*, **27**, 4, 435-439.

Johnson, R. A. and Wichern, D. W., 2001, "Applied Multivariate Statistical Analysis", 3rd edition, Prentice-Hall of India.

Rai, T. and Chandrahas, 2000, "Use of discriminant function of weather parameters for developing forecast model of rice crop", Publication of IASRI, New Delhi.