MAUSAM, **63**, 3 (July 2012), 455-458

551.509.39 : 633

Use of discriminant function analysis for forecasting crop yield

RANJANA AGRAWAL, CHANDRAHAS and KAUSTAV ADITYA

Indian Agricultural Statistics Research Institute, New Delhi, India

(*Received 2 May 2011, Modified 18 July 2011*)

e-mail : ranjana@iasri.res.in

सार – प्रस्तुत शोधपत्र कानपुर (भारत) में गेहूँ की उपज के पूर्वानुमान मॉडल को विकसित करने के लिए विक्तिकर फलन के उपयोग से संबंधित है। विक्तिकर फलन विश्लेषण रैखिक/द्विघात फलन को विकसित करने की एक पद्धति है जिसके द्वारा विभिन्न वर्गो में अंतर भलीभाँति स्पष्ट किया जा सकता है तथा इस प्रकार इसके द्वारा संभावित उपज का गुणात्मक अनुमान लगाया जा सकता है। इस अध्ययन में विक्तिकर फलन विश्लेषण द्वारा प्राप्त मौसम स्कोरों को समाश्रयण के रूप में लेते हुए बहुसमाश्रयण पद्धति का उपयोग कके उपज के मात्रात्मक अनुमान प्राप्त किए गए। उपज के बंटन के आधार पर 30 वर्षों (1971–2000) के श्रृंखला आँकड़ों को सामान्य के अनुकूल, सामान्य व सामान्य से खराब तीन वर्गों में विभाजित किया गया। इन तीन मॉडलिंग में समाश्रयण के रूप में उपयोग किया गया। साप्ताहिक आँकड़ों के उपयोग के लिए विभिन्न नीतियाँ प्रस्तावित की गई। इन मॉडलों का उपयोग करके वर्ष 2000–01 से 2002–03 तक के (जिनका उपयोग मॉडलों के विकास में नहीं किया गया) उपज के लिए पूर्वानुमान प्राप्त किए गए। इस पद्धति से कटाई से लगभग दो माह पूर्व विश्वसनीय उपज के पूर्वानुमान प्राप्त किए जा सकते हैं।

ABSTRACT. The present paper deals with use of discriminant function analysis for developing wheat yield forecast model for Kanpur (India). Discriminant function analysis is a technique of obtaining linear/Quadratic function which discriminates the best among populations and as such, provides qualitative assessment of the probable yield. In this study, quantitative forecasts of yield have been obtained using multiple regression technique taking regressors as weather scores obtained through discriminant function analysis. Time series data of 30 years (1971-2000) have been divided into three categories: congenial, normal and adverse, based on yield distribution. Taking these three groups as three populations, discriminant function analysis has been carried out. Discriminant scores obtained from this have been used as regressors in the modelling. Various strategies of using weekly weather data have been proposed. The models have been used to forecast yield in the subsequent three years 2000-01 to 2002-03 (which were not included in model development). The approach provided reliable yield forecast about two months before harvest.

Key words – Weather variables, Weather indices, Discriminant function analysis, Crop yield forecast modelling.

1. Introduction

Timely and effective forecasts of the crop yields are vital for an agrarian economy. Crop yield forecasts are important for advance planning, formulation and implementation of policies related to food procurement, distribution, price structure and import-export decisions etc. These are also useful to farmers to decide in advance their future prospects and course of action. Thus reliable and timely pre-harvest forecasts of crop yield are very important. For this purpose, weather based models using different statistical approaches have been tried by the researchers. In the present paper, use of discriminant function analysis has been explored for forecasting crop yield. The methodology has been demonstrated using data of wheat yield in Kanpur district in Uttar Pradesh.

2. Data

District level wheat yield data for thirty three years (1971-2003) have been collected from Directorate of Economics and Statistics, Ministry of Agriculture, New Delhi. Weekly data (1971-2003) on weather variables during the different growth phases of crop were obtained from Central Research Institute for Dry-Land Agriculture (C.R.I.D.A.), Hyderabad. First thirty years' data from 1970-71 to 1999-2000 have been used for model development and data from 2000-01 to 2002-03 were used for validation of the models.

Wheat is generally sown in the month of October when average daily temperature falls around 23-25° C. After sowing of the crop, germination takes 6-7 days or

near about one week after the pre-sowing phase. Crown root initiation occurs after 20-25 days of sowing or in about 3 weeks from germination. Tillering phase lasts up to nearly about 2-3 weeks after crown root initiation. Jointing and Reproductive phase is the peak plant growth stage and starts after the tillering phase or 45-60 days after sowing. The reproductive phase lasts 60-85 days after sowing.

As weather during pre-sowing period is important for establishment of the crop, data starting from two weeks before sowing have been included in model development. Further, as the forecast is required well in advance of harvest, weather data about 2 months before harvesting have been considered. Thus, data on four weather variables *viz*., Maximum and Minimum Temperature, Rainfall and morning Relative Humidity during 16 weeks [40th standard meteorological week (smw) to 52nd smw of a year and 1st smw to 3rd smw of the next year] have been used in the analysis.

3. Statistical methodology

Discriminant function analysis is a multivariate technique discussed in many books, to mention a few, Anderson (1984), Hair *et al*. (1995), Sharma (1996), Johnson and Wichern (2006), etc. This technique is used for identifying appropriate functions that discriminate best between set of observations from two or more groups and classifying future observations into one of the previously defined groups.

Suppose observations are to be classified into k groups on the basis of p variables. The technique involves identifying linear/quadratic function(s) where the coefficients are determined in such a way that the variation between the groups gets maximized relative to the variation within the groups. The maximum number of discriminant functions that can be obtained is equal to minimum of k-1 and p. These functions are used to calculate discriminant scores, which are used to classify the observations into different groups.

Rai and Chandrahas (2000) developed forecast models for rice in Raipur district using discriminant function technique and provided reliable yield forecast about two months before harvest. This paper applies the technique used by them alongwith a few modifications.

To apply discriminant function analysis for modelling yield using weather variables, crop years have been divided into three groups namely congenial, normal and adverse on the basis of crop yield adjusted for trend effect. Data on weather variables in these three groups were used to develop linear discriminant functions and the discriminant scores were obtained for each year. These scores were used alongwith year as regressors in developing the forecast models. In the present study the number of groups is three and number of weather variables is four, therefore only two discriminant functions are sufficient for discriminating a crop year into either of the three groups.

However using weather variables as such in developing the model poses a problem. Weather variables affect the crop differently during different phases of development. Thus extent of weather influence on crop yield depends not only on the magnitude of weather variables but also on the distribution pattern of weather over the crop season which, as such, calls for the necessity of dividing the whole crop season into finer intervals. But, doing so will increase number of variables in the model and in turn a large number of parameters will have to be evaluated from the data and sufficient number of observations may not be available for precise estimation of these parameters. This gives rise to the problem of number of variables under study more than number of observations. To solve this problem suitable strategies have been suggested and following four models were proposed:

3.1. *Model-1*

Total time starting from two weeks before sowing up to the time of forecast (*i.e*., 16 weeks starting from 40th smw) has been divided into four phases where each phase consists of different number of weeks. For each phase and each weather variable simple average of the weather data in the different weeks within the phase was obtained. This way for each phase four average weather variables were obtained. Taking these four average weather variables, phase wise discriminant function analysis was carried out and entire data on weather variables were converted to two discriminant scores for each phase in each year. Thus, in all eight scores were obtained for each year. Using these eight discriminant scores and time trend as regressors, model was fitted using stepwise regression technique. The form of the model is as follows:

$$\text{Yield} = \alpha + \sum_{l=1}^{2} \sum_{m=1}^{4} \beta_{lm} ds_{lm} + \beta_9 T + \varepsilon$$

where α = intercept of the model, $\beta_{lm}{}'s \left(l = 1, 2; m = 1, 2, ..., 4\right)$ and β_9 are the regression coefficients, ds_{lm} is the l^{th} discriminant score in m^{th} phase, T is the trend variable (year) and ε is error $\sim N\left(0, \sigma^2\right)$.

This model is same as proposed by (Rai & Chandrahas 2000). This model is based on significant discriminant scores in some phases selected by stepwise regression and as such suffers from the drawback that it is not based on complete data. Further, it gives equal weightage to the weather variables in all the weeks within the phases as it uses phase-wise simple averages.

3.2. *Model-2*

In this model, for each weather variable, an index was developed as weightage total of weather data over all the weeks upto the time of forecast, weights being the correlation coefficients between detrended yield and respective weather variable in different weeks. Thus the entire data of four weather variables in sixteen weeks were converted into four weather indices. Discriminant function analysis was carried out on these four weather indices and two discriminant functions and therefrom two discriminant scores were obtained for each year. These two scores and trend were utilized in developing forecast model using stepwise regression technique. The form of the model is as follows:

$$\text{Yield} = \alpha + \beta_1 ds_1 + \beta_2 ds_2 + \beta_3 T + \varepsilon$$

where α = intercept of the model, $\beta_i's\ (i = 1, 2, 3)$ = the regression coefficients, ds_1 ds_2 are the two discriminant scores, T and ε are as defined in model-1.

This model utilises complete data over all 16 weeks and also considers relative importance of weather variables in different weeks as against model-1 in which equal importance in different weeks was assigned.

3.3. *Model-3*

In model-1, for each phase discriminant scores based on averages of weather variables were computed. In the proposed model, for each phase and each weather variable, weather index as in model-2 has been computed. In all, sixteen weather indices were obtained. Discriminant function analysis was carried out using these sixteen indices which provided two discriminant functions giving two scores for each year. These two scores alongwith the trend variable as the regressors were utilized for developing forecast model. Like model-2, this model also considered complete data and relative importance of weather variables in different weeks. The form of the model is as follows:

$$\text{Yield} = \alpha + \beta_1 ds_1 + \beta_2 ds_2 + \beta_3 T + \varepsilon$$

where symbols α, $\beta_i's$, ds_1, and ds_2 are as defined earlier.

3.4. *Model-4*

In this model, two discriminant functions were developed using four weather variables data for the first week (40th smw). These discriminant functions were used to compute two scores for each year. Taking data on four weather variables in the second week (41st smw) and two discriminant scores computed from the first week (*i.e.* 6 discriminator variables) discriminant function analysis was carried out which provided two scores for each year based on data upto second week (41st smw). The process was repeated for the successive weeks data till the time of forecast (16th week *i.e.*, 3rd smw) and finally two discriminant scores based on the entire data were obtained for each year. The data on these two discriminant scores alongwith trend were used to develop the model through stepwise regression technique. The form of the model is

$$\text{Yield} = \alpha + \beta_1 ds_1 + \beta_2 ds_2 + \beta_3 T + \varepsilon$$

where symbols are as defined earlier.

This model was also based on complete data and relative importance of weather variables in different weeks.

3.5. *Comparison and validation of models*

The four models were compared on the basis of adjusted coefficient of determination $\left(R_{adj}^{\ 2}\right)$ which is as follows:

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

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

From the fitted models, wheat yield forecasts for the years 2000-01 to 2002-03 were obtained and forecasts were compared on the basis of Root Mean Square Error (RMSE).

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n} (\text{O}_i - \text{E}_i)^2\right]^{1/2}$$

where O_i and the E_i are the observed and forecast values of crop yield respectively and n is the number of years for which forecasting has been done.

TABLE 1

Wheat yield forecast models

Model	Forecast regression equation	R^2	R^2_{adj}
1.	Yield = 12.86 - 0.66 ds_{11} + 0.97 ds_{12} - 0.80 ds_{23} + 0.65 T (0.90) (0.32) (0.38) (0.41) (0.05)	0.91	0.90
2.	Yield = 11.85 - 1.46 ds_1 + 0.71 T (0.71) (0.29) (0.04)	0.92	0.92
3.	Yield = 11.66 + 1.51 ds_2 + 0.72 T (0.66) (0.26) (0.04)	0.93	0.93
4.	Yield = 12.03 + 0.59 ds_1 + 0.64 ds_2 + 0.70 T (0.57) (0.12) (0.11) (0.03)	0.95	0.95

Note : Figures in brackets denote Standard Error of regression coefficients

TABLE 2

Actual & forecasts of wheat yield (Q/ha)

Crop year	Actual yield	Forecast using			
		Model -1	Model -2	Model -3	Model -4
2000-01	32.14	35.37	35.32	34.61	30.69
		(10.05)	(9.89)	(7.68)	(4.51)
2001-02	29.64	33.58	31.59	28.79	29.31
		(13.29)	(6.58)	(2.86)	(1.11)
2002-03	33.36	36.73	34.57	33.35	35.22
		(10.01)	(3.71)	(0.03)	(5.57)
	RMSE	3.25	2.25	1.51	1.37

Note : Figures in brackets denote % deviation of forecast

4. Results and conclusions

The forecast models obtained under the four strategies (as discussed under section 3) alongwith adjusted coefficient of determination $\left(R_{adj}^2\right)$ are presented in Table 1. In all the models, trend variable T was found significant. Apart from trend T other significant variables were found as discriminant scores ds_{11}, ds_{12}, ds_{23} in model-1, ds_1 in model-2, ds_2 in model-3, and ds_1 & ds_2 in model-4. Adjusted coefficient of determination $\left(R_{adj}^2\right)$ varied between 0.90 to 0.95 in different models, the maximum (0.95) being in model-4. RMSE was computed on the basis of yield forecasts for the years 2000-01 to 2002-03. The results (Table 2) revealed that the per cent deviation of forecast varied from 10.01 to 13.29 in model-1, 3.71 to 9.89 in model-2, 0.03 to 7.68 in model-3 and 1.11 to 5.57 in model-4 over the three years. The RMSE varied from a minimum of 1.37 in model-4 to a maximum of 3.25 in model-1. Thus it is concluded that model-4 is the most suitable model among the models considered for forecasting wheat yield for Kanpur district of Uttar Pradesh. The model provides reliable forecast around two months before harvest.

References

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

Hair, Joseph F., Anderson, Rolph, E., Tatham, Ronald L., Black, William C., 1995, "Multivariate Data Analysis with Readings", Prentice Hall, Inc. New Jersey.

Johnson, Richard A. and Wichern, Dean W., 2006, "Applied Multivariate Statistical Analysis", *Pearson Education*.

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

Sharma, S., 1996, "Applied Multivariate Techniques", *John Wiley & Sons*, New York.