Journal of Agrometeorology 21 (4) : 462-467 (December 2019)

Bayesian discriminant function analysis based forecasting of crop yield in Kanpur district of Uttar Pradesh

VANDITA KUMARI*, KAUSTAV ADITYA, HUKUM CHANDRA and AMARENDER KUMAR[1]

ICAR-Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi – 110012, India
ICAR-Indian Agricultural Research Institute, Pusa, New Delhi – 110012, India
** Corresponding author's email: vandita.iasri@gmail.com*

ABSTRACT

Discriminant function analysis technique using Bayesian approach has been attempted for wheat forecasting in Kanpur district of Uttar Pradesh, India both qualitatively and quantitatively. Crop yield data and weekly weather data on temperature (maximum and minimum), relative humidity (maximum and minimum), rainfall for 16 weeks of the crop cultivation have been used in the study. These data have been utilized for model fitting and validation. Crop years were divided into two and three groups based on the de-trended yield. Crop yield forecast models have been developed using posterior probabilities calculated through Bayesian approach in stepwise discriminant function analysis along with year as regressors for different weeks. Suitable strategy has been used to solve the problem of number of variables more than number of data points. Performance of the models obtained at different weeks was compared using Adjusted R^2, PRESS (Predicted error sum of square), number of misclassifications. Forecasts were evaluated using RMSE (Root Mean Square Error) and MAPE (Mean absolute percentage error) of forecast. The result shows that the model based on three groups case perform better. The performance of the proposed Bayesian discriminant function analysis technique approach was better as compared to existing discriminant function analysis score based approach both qualitatively and quantitatively.

Key words: Crop yield forecasting, posterior probabilities, discriminant function analysis

India is facing the problem of rapid population growth as well as decrease in arable lands and challenge is to produce more and more to feed the ever growing population. Moreover, there have been changes in the weather pattern in the recent decade which poses risk to the agricultural production. Crop yield forecast is a crucial for timely planning and sound decisions making for the country. Thus, reliable and timely pre-harvest forecasting of crop yield is very important for formulation and implementation of policies related to the crop procurement, distribution and import export decisions etc. Since, agriculture production depends highly on the weather parameters, therefore weather based approach for forecasting crop yield has been useful. Various workers have attempted to develop methodology for weather based models for crop yield forecasting such as weather indices based regression models (Agrawal *et al.*, 1986; 2001; Giri *et al.,* 2017; Gupta *et al.,* 2018; Kumar *et al.,* 2019), neural network (Kumar *et al.*, 2010; Laxmi and Kumar, 2011), ordinal logistic regression (Kumari and Kumar, 2014; Kumari *et al.* 2016), crop simulation models (Singh *et al*., 2017) and reference therein.

Discriminant function analysis is a multivariate technique concerned with separating distinct sets of objects (or sets of observations) and allocating new objects (or observations) to the previously defined groups. This technique has been discussed in several book to name few Anderson (1984), Johnson and Wichern (2006), etc. Discriminant function analysis has been used by several authors for pre-harvest forecasting of crop yield. Rai and Chandrahas (2000) made use of discriminant function analysis of weather variables to develop statistical models for pre-harvest forecasting of rice-yield in Raipur district of Chhattisgarh. Agrawal *et al.* (2012) have developed forecast models using scores for wheat yield using discriminant functions analysis of weakly data on weather variables. Sisodia *et al.* (2014) used discriminant function analysis of meteorological parameters for developing suitable statistical models to forecast wheat yield in Faizabad district of Eastern Uttar Pradesh. Singh *et al.* (2017) developed statistical models to forecast on wheat yield in Sultanpur district of Eastern Utter Pradesh using discriminant function analysis of meteorological parameters. All these works were done using discriminant function

analysis based on scores and fitting the models with discriminant scores treated as regressors. In this paper discriminant function analysis approach was used to generate posterior probabilities through Bayesian approach for forecasting wheat yield for Kanpur district of Uttar Pradesh in India both qualitatively and quantitatively.

MATERIALS AND METHODS

Description of data

The study was conducted on Kanpur district of Uttar Pradesh (U.P.) of India which is situated between 26°03' N latitude and 80°04' E longitude. It lies in the central plain zone of U.P. It is liberally sourced by the Ganges and Yamuna Rivers and their tributaries. The favorable climate, soil and the availability of ample irrigation facilities make growing of wheat a natural choice for the area. Wheat crop is generally cultivated during the Rabi season because during this period it provides a better environment for the cultivation of the wheat crop. It is generally sown towards the end of October when average daily temperature falls around 23-25°C.

For this study, time series data on yield (q ha^{-1}) of wheat crop for Kanpur district for 39 years (1971-72 to 2009-10) have been collected from Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare, Govt. of India, New Delhi. Weekly weather data for 39 years (1971-72 to 2009-10) of Kanpur district during the different growth phases (Pre-Sowing and germination phase, crown root development phase, tillering phase, jointing and reproductive phase) of wheat crop have been obtained from Central Research Institute for Dryland Agriculture (CRIDA), Hyderabad, India. Since, weather during pre-sowing period is important for establishment of the crop and also the forecast is required well in advance of harvest, weather data starting from two weeks before sowing i.e. first week of October to about two months before harvesting i.e. 15-21 January of the next year have been considered. A data on five weather variables viz. maximum temperature, minimum temperature, morning relative humidity, evening relative humidity and rainfall for 16 weeks of the crop cultivation i.e. for 40th standard meteorological week (SMW) to 52nd SMW of a year and 1st SMW to 3rd SMW of subsequent year has been used in the study. Data from 1971-72 to 2006-07 have been utilized for model fitting and subsequent three years (2007-09) were used for the validation of the model.

Methodology

In the proposed methodology, crop years have been divided into two and three groups on the basis of crop yield adjusted for trend effect. A linear regression has been fitted between yield (dependent variable) and year (independent variable) using data from 1971-72 to 2006-07. The equation thus obtained was utilized for getting residuals (detrended yield). These residuals were then used for the group formation. The years were divided into two groups on the basis of residuals, the residuals with negative values were considered as bad year (0) and with positive values were considered as good year (1). Crop years were grouped into three, where residuals (after fitting linear regression between yield and year) have been arranged into ascending order and were divided into three equal groups namely adverse year (0), normal year (1) and congenial year (2). Using weather variables in these two/three groups, posterior probabilities were obtained by discriminant function analysis. These probabilities along with year as regressors were used for development of forecast model. The models were developed using stepwise regression procedure.

Therefore, use of five variables in 16 weeks as such makes number of explanatory variable 80 for discriminant function analysis. The strategy to solve the problem of number of variables more than number of data points is as follows. At first week (40th SMW), the weather variables corresponding to the pre-defined groups have been used to compute posterior probabilities by stepwise discriminant function analysis. At the second week (41st SMW), the weather variables of this week along with probabilities computed at the first week have been used to compute posterior probabilities using stepwise discriminant function analysis. These steps have been repeated up to last week (3rd SMW). The models were developed at different weeks starting from 52nd SMW using stepwise regression procedure along with Bayesian posterior probabilities obtained at time of forecast where year is used as regressors and yield as the regressand. These models were then validated by forecasting yield of subsequent years. The forms of various forecast models are given below.

Modelling with two groups

Posterior probabilities are calculated based on Bayesian theory. The posterior probability, $P_1 = P(G_1/\mathbf{x}_m)$, that an event is good year (G_1) given the value of weather variable $\mathbf{x}_m$ is given by

$$P_1 = P(G_1/\mathbf{x}_m) = \frac{P(G_1)P(\mathbf{x}_m/G_1)}{P(\mathbf{x}_m/G_1)P(G_1) + P(\mathbf{x}_m/G_0)P(G_0)}$$

$$= \frac{P(\mathbf{x}_m \cap G_1)}{P(\mathbf{x}_m)}$$

$$= \frac{P(G_1)P(\mathbf{x}_m/G_1)}{P(\mathbf{x}_m/G_1)P(G_1)+P(\mathbf{x}_m/G_0)P(G_0)}$$

Forecast model

Models were fitted using stepwise regression at different weeks starting from 52nd SMW taking posterior probability of first group at the week of forecast along with year as regressors. The model fitted was

$$Yield = a + b_1P_1 + b_2T + \varepsilon$$

where, a is intercept of the model

b_i's are the regression coefficients

P_1 is the posterior probability of good year (Y=1) obtained using weather variables

T is year

ε is error ~$N(0,\sigma^2)$.

Model under discriminant function approach using Posterior Probabilities – three groups

The posterior probability, $P(G_1/\mathbf{x}_i)$, that an event belongs to normal year (G_1) given the value of weather variable $\mathbf{x}_m$ is given by

$$P_1 = P(G_1/\mathbf{x}_m) = \frac{P(G_1 \text{ occurs and observe } \mathbf{x}_m)}{P(\text{observe } \mathbf{x}_m)}$$

$$= \frac{P(\mathbf{x}_m \cap G_1)}{P(\mathbf{x}_m)}$$

$$= \frac{P(G_1)P(\mathbf{x}_m/G_1)}{P(\mathbf{x}_m/G_1)P(G_1)+P(\mathbf{x}_m/G_2)P(G_2)+P(\mathbf{x}_m/G_0)P(G_0)}$$

$P_2$is posterior probability, $P(G_2/\mathbf{x}_m)$, that an event belongs to congenial year (G_2) given the value of weather variables x_m is given by

$$P_2 = P(G_2/\mathbf{x}_m) = \frac{P(G_2)P(\mathbf{x}_m/G_2)}{P(\mathbf{x}_m/G_1)P(G_1) + P(\mathbf{x}_m/G_2)P(G_2) + P(\mathbf{x}_m/G_0)P(G_0)}$$

Other symbols are same as defined earlier.

Forecast model

Models were fitted using stepwise regression at different weeks starting from 52nd SMW taking posterior probabilities at the week of forecast along with year as regressors. The model fitted here was

$$Yield = a + b_1P_1 + b_2P_2 + b_3T + \varepsilon$$

where, P_1 and P_2 are the posterior probabilities of Y=1 (normal year) and Y=2 (congenial year) respectively and other symbols are same as defined earlier.

Following the technique proposed by Agarwal *et al.* (2012), forecast model based on discriminant scores has been developed for both two and three groups' cases. Performance of proposed methodology using posterior probabilities based on Bayesian approach and score based approach (Agarwal *et al.* (2012)) in two and three groups has been compared qualitatively on the basis of number of misclassifications. They have been compared quantitatively on the basis of Adj R^2, PRESS of the fitted forecast model, RMSE and MAPE of the forecast. The SAS software has been used for the statistical analysis.

RESULTS AND DISCUSSION

Forecast model with two group

Yield data have been classified into two groups namely good year (1) and bad year (0) on the basis of crop yield adjusted for the year effect taking negative residuals as bad years and positive residuals as good years. Out of thirty six years, the numbers of good crop years are twenty and bad are sixteen. Weather variables in two groups were used to obtain posterior probability of good year through stepwise discriminant function analysis. Regression models have been fitted using stepwise regression by taking yield as the dependent variable and the probability of good year (Y=1) and year as the regressors for different weeks of forecast starting from 52nd SMW. Wheat yield forecast models along with Adj R^2, PRESS and number of misclassifications for different weeks are given in Table 1. The results also show that Adj R^2 varied from a minimum 0.92 in 52nd SMW to a maximum of 0.93 in 3rd SMW. Therefore, on the basis of Adj R^2 3rd week has best model fit as compared to others.

Forecast model with three groups

Yield data have been classified into three groups namely congenial, normal or adverse on the basis of detrended yield and each group consisting of twelve years. Using weather variables in these groups, posterior probabilities have been obtained using Bayesian approach in stepwise discriminant function analysis. Regression models have been fitted using stepwise regression by taking yield as the dependent variable and the probabilities and year as the regressors. Wheat yield forecast models for different weeks are given in Table 2. The results also show that Adj. R^2 varied from a minimum in 3rd SMW to a maximum of in 52nd SMW. Therefore, on the basis of Adj. R^2 52nd week has best model fit as compared to others.

Table 1: Wheat yield (q ha^{-1}) forecast models for different weeks

Week of forecast (SMW)	Forecast regression equation	Adj. R^2
52	Yield = -1143.51 + 4.10P + 0.59T (60.86) (0.65) (0.03)	0.92**
1	Yield = -1150.76+ 4.23P + 0.59T (57.76) (0.61) (0.03)	0.93**
2	Yield = -1163.51 + 0.65P + 0.59 T (60.04) (0.10) (0.03)	0.92**
3	Yield = -1153.58 + 0.66 P + 0.59T (58.34) (0.09) (0.03)	0.93**

Figures in brackets denote Standard Errors; ** Significant at 1% level of significance

Table 3: Adj. R^2 for discriminant function analysis using posterior probabilities and scores

Week of forecast (SMW)	Two groups		Three groups	
	Probability	Scores	Probability	Scores
52	0.92**	0.92**	0.93**	0.88**
1	0.93**	0.92**	0.93**	0.88**
2	0.92**	0.92**	0.93**	0.88**
3	0.93**	0.92**	0.93**	0.88**

**Significant at p = 0.01

Forecasting of crop yield using Bayesian discriminant function analysis approach using posterior probabilities and discriminant scores in two and three groups cases have been compared. Results of comparison for these approaches are given in Table 3 to Table 8. In these tables results related to the proposed approach is denoted as 'Probability' where as the approach given by Agarwal *et al*. (2012) is denoted as 'Scores'. It can be observed that Adj. R^2 is at par under probability and score based discriminant function analysis for two groups case (Table 3). The posterior probability based discriminant function analysis has best model fit for 3rd week in two group cases. In three groups case Adj. R^2 is more under posterior probability based discriminant function analysis as compared to scores based. Thus, probability based discriminant function analysis has best model fit for 52nd week in three group cases.

PRESS under posterior probability based discriminant function approach was less than score based approach except 52nd SMW in two groups' case. The posterior probability based discriminant function analysis has best model fit for 3rd week in two group cases. In three groups case performance of

Table 2: Wheat yield (q ha^{-1}) forecast models for different weeks

Week of forecast (SMW)	Forecast regression equation	Adj. R^2
52	Yield =-1177.03 + 3.33 P_1 + 5.59 P_2 + 0.60T (56.32) (0.89) (0.76) (0.03)	0.93**
1	Yield =-1179.95 + 3.03 P_1 + 5.36 P_2 + 0.60T (57.08) (0.83) (0.75) (0.03)	0.93**
2	Yield = -1182.27+2.83 P_1 +5.26P_2 + 0.61T (58.25) (0.83) (0.76) (0.03)	0.93**
3	Yield = -1184.03 + 2.74P_1 + 5.21P_2 + 0.61 T (59.01) (0.84) (0.77) (0.03)	0.93**

Table 4: PRESS for discriminant function

Week of forecast (SMW)	Two groups		Three groups	
	Probability	Scores	Probability	Scores
52	144.7	135.9	124.5	222.0
1	130.6	138.6	128.5	219.6
2	139.9	150.9	134.4	219.6
3	132.7	150.9	138.4	219.6

probability based discriminant function analysis was much less as compared to score based approach for all the weeks. Bayesian discriminant function analysis based model has best fit for 52nd week in three group cases (Table 4).

For qualitative comparison, the number of misclassifications in fitted forecast model has been found to be same for posterior probability and score based approach in two groups case. Whereas, in three groups case number of misclassifications has been found to be less for discriminant function analysis using probabilities (Table 5). RMSE of forecast under posterior probability based discriminant function analysis is less than score based discriminant function analysis in two groups case (Table 6). Whereas in three groups case RMSE of forecast is less for score based discriminant function analysis. Further, RMSE of forecast was minimum for 52nd SMW under probability based discriminant function analysis two groups case.

MAPE of forecast under discriminant function analysis using posterior probability was less as compared to discriminant function analysis using score in both two and three groups cases (Table 7). Further MAPE was minimum for 52nd SMW under probability based discriminant function analysis approach. Number of misclassifications was same for

Table 5: Misclassifications for discriminant function

Week of forecast (SMW)	Two groups		Three groups	
	Probability	Scores	Probability	Scores
52	4	4	6	13
1	2	2	7	12
2	2	2	7	12
3	2	2	7	12

Table 6: RMSE of forecast for discriminant function

Week of forecast (SMW)	Two groups		Three groups	
	Probability	Scores	Probability	Scores
52	4.15	5.00	4.44	4.38
1	4.29	4.81	4.46	4.41
2	4.24	4.74	4.50	4.41
3	4.21	4.74	4.53	4.41

Table 7: MAPE of forecast for discriminant function

Week of forecast (SMW)	Two groups		Three groups	
	Probability	Scores	Probability	Scores
52	10.69	13.80	10.50	13.06
1	11.09	13.57	10.57	13.23
2	10.99	13.37	10.87	13.23
3	10.95	13.37	11.06	13.23

Table 8: Misclassifications in forecasted years

Week of forecast (SMW)	Two Groups		Three Groups	
	Probability	Scores	Probability	Scores
52	1	1	1	2
1	1	1	1	2
2	1	1	1	2
3	1	1	1	2

Table 9 : Observed and forecasted yield (q ha^{-1}) for different years at 52nd SMW

Week of forecast (SMW)	Year	Observed yield	Forecasted yield
52	2007-08	30.08	32.79
	2008-09	33.56	33.63
	2009-10	32.31	39.51

posterior probability based discriminant function analysis in two as well as three groups cases. When compared to score based discriminant function analysis, the probability based approach has less number of misclassification in all the weeks in three groups cases (Table 8). The values of the observed and forecasted yield of different years for 52nd week using posterior probability based discriminant function analysis are given in Table 9.

CONCLUSION

This study reveals that the performance of Bayesian discriminant function analysis approach is better than score based discriminant function analysis for forecasting crop yield both qualitatively and quantitatively in most of the cases. Thus, it can be concluded that the performance of posterior probability based discriminant function analysis is better than score based discriminant function analysis for forecasting crop yield. Three groups classification with posterior probability based discriminant function analysis gave better result quantitatively on the basis of fitted forecast model. Optimum time of forecast has been obtained as 52nd SMW (11 weeks after sowing). Thus model based on discriminant function analysis using posterior probability with three groups' classification at 52nd week can be recommended for forecasting wheat yield in Kanpur district of Uttar Pradesh.

REFERENCES

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

Anderson, T.W. (1984). An introduction to applied Multivariate Statistical Analysis, John Wiley & Sons Inc. New York.

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

Giri, A.K., Bhan, M. and Agrawal, K.K. (2017). District-wise wheat and rice yield predictions using meteorological variables in eastern Madhya Pradesh. *J. Agrometeorol.*, 19(4):366-368.

Gupta S., Singh A., Kumar A., Shahi, U.P., Sinha, N.K. and Roy, S. (2018). Yield forecasting of wheat and mustard

for western Uttar Pradesh using statistical model. *J. Agrometeorol.*, 20(1):66-68.

Kumar, S., Attri, S.D. and Singh, K.K. (2019). Comparison of Lasso and stepwise regression technique for wheat yield prediction. *J. Agrometeorol.*, 21(2):188-192.

Kumari, V. and Kumar A. (2014). Forecasting of wheat (*Triticum aestivum*) yield using ordinal logistic regression. *Indian J. Agric. Sci.*, 84(6):691-694.

Kumari, V., Agrawal, R. and Kumar A. (2016). Use of ordinal logistic regression in crop yield forecasting. *Mausam*, 67(4):913-918.

Laxmi, R. R. and Kumar, A. (2011). Weather based forecasting model for crops yield using neural network approach. *Stat. Appl.*, 9**(**1-2):55-69.

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

Singh, H.G., Rai, V.N, Sisodia, B.V.S., Nand, V. and Singh, H.Y. (2017). Forecasting of pre-harvest crop yield using discriminant function analysis of meteorological parameters. *Res. Environ. Life Sci.*, 10(1):11-14.

Singh P.K., Singh, K.K., Singh, P., Balasubramanian, R., Baxla, A.K, Kumar, B., Gupta, A., Rathore, L.S. and Kalra, N. (2017). Forecasting of wheat yield in various agro-climatic regions of Bihar by using CERES-Wheat model. *J. Agrometeorol.*, 19(4):346-349.

Sisodia, B.V.S., Yadav, R. R., Kumar S. and Sharma, M.K. (2014). Forecasting of pre-harvest crop yield using discriminant function analysis of meteorological parameters. *J. Agrometeorol.*, 16(1):121-125.

Received : November 2018 ; Accepted : November 2019