Development of Pre Harvest Forecasting Models Based on Discriminant Function Analysis (DFA) and Principal Component Analysis (PCA) on Yield of Chickpea for Chhattisgarh Plain Zone

Kamal Nayan Pammar[1] Pandey K. K.[2]
[1,2]Department of Agril Statistics
[1,2]Indira Gandhi Agriculture University, Raipur (C.G.), India

Abstract— The present investigation deals with the model have been also developed by Discriminant function analysis and Principal Component Analysis on Chhattisgarh plain Zone. Models fitted with DFA then 2 Discriminant score (ds1 and ds2) and time trend (T) the models found are highly significance and R^2 value 68% for plain zone for Chickpea crop and Models fitted with Principal Components then 3 (PC1, PC2 and PC3) Components and Time trend (T) entered Models are highly significant and R2 value 73% for Chickpea for plain Zone, highly significant at 0.1% level of significance. The present investigation covers under the study of individual effect of weather variables, joint effect of weather variables following the concept of older studies by the different scientists in this direction for many crops.

Keywords: Chickpea, Coefficient of Determination, Discriminant Function Analysis and Principal Component Analysis

I. INTRODUCTION

Forecasting is a technique that uses historical data as inputs to make informed estimates that are predictive in determining the direction of future trends. Forecast of the crop production at suitable stages of crop period before the harvest are vital for rural economy. Forecasts are used by the government & industry to protect life and property. The weather and climatic information plays a major role before and during the cropping season and if provided in advance it can be helpful in stirring the farmer to form and use their own resources in order to gather the benefits. The advance knowledge of weather parameters in a particular region is advantageous in effective planning.

The final estimates based on objective crop-cutting experiments are of limited utility as those become available quite later after the crop harvest. The Statistical techniques employed for forecasting purposes should be able to provide objective, consistent and comprehensible forecasts of crop yield with reasonable precisions well in advance before the harvest. Several studies have been carried out to forecast crop yield using weather parameters etc.

In the current situation of India faces increasing population and industrial development which are adversely distressing the crop yield in India. Keeping in mind early crop yield forecast will help farmer to formulate the cropping pattern, agricultural practices which will results in to the increase yield of the farmers. Therefore main objective of the present study was to develop different approaches for forecasting the rice yield before harvesting with help of weather parameters.. The forecasting of crop yield may be done by using three major objective methods (i) biometrical characteristics (ii) weather variables and (iii) agricultural inputs (Agrawal *et al.* 2001).

II. MATERIAL AND METHODS

The yield data on Chickpea crop for 28 years (1990-91 to 2017-18) for Raipur, Durg, Rajnadgaon and Bilaspur Districts. 20 year (1998-99 to 2017-18) for Dhamtari, Mahasamund, Korba, Kawardha and Janjgir have been procured from the published booklets and official website (http://agridept.cg.gov.in/index.htm) of the Directorate of Agricultural, Govt.of Chhattisgarh to development of the statistical models at zone level.

III. DISCRIMINATE FUNCTION ANALYSIS

The data is divided in to three groups *viz.* Congenial, Average and Adverse on the basis of productivity of chickpea for the given time series data. In this procedure discriminant scores have been computed using the un weighted and weighted averages (weather indices) for the first weather variable (Here, discriminating factors will be only two).

This process is continued up to seventh weather variables, and finally we get two discriminant scores ds_1, and ds_2. Using crop-yield as dependent variable and discriminant scores ds_1, ds_2 and the time trend T as independent variable, the following model is fitted for the development of forecast model.

$$\text{Model : } y = \beta o + \beta_1 ds_1 + \beta_2 ds_2 + \beta_3 T + \varepsilon$$

Where y = crop yield
βo = intercept of the model
$\beta_{i's}$ $(i=1,2,3)$= the regression coefficients
ds_1 and ds_2 are the two discriminant scores
T = trend variable ($T= 1,2,....n$) n is crop-years under consideration
ε is error~ N (0, σ^2).

IV. PRINCIPAL COMPONENT ANALYSIS

PCA is a multivariate technique for data reduction. It is a mathematical function, which does not require user to specify the statistical model or assumption about distribution of original variables. (Anderson *et al.*1984).

Let X_{ij} be the value of j^{th} biometrical character (j= 1, 2,...p) corresponding to i^{th}varieties of experiment (i= 1, 2,...n). The PCA for $x_{ij's}$ will be carried out. Let PC_1, PC_2, ...PC_K be first K (K< P) principal components explaining variability about more than 70 to 90 percent of the total variation in $x_{ij's}$. Using these K principal components as regressor variables and varieties yield (y_i) as regressand, the following linear multiple regression model for pre-harvest forecast of crop yield has been proposed.

$$Y_i = \beta_0 + \beta_1 PC_{1i} + \beta_2 PC_{2i} + ... + \beta_k PC_{ki} + e_i$$
$$,i= 1, 2, ...n.$$

Where Y_i is the crop yield of the i^{th} plot; β_0, β_1, β_2 β_k are model parameters and e_i is error term assumed to follow

independently normal distribution with mean 0 and variance σ^2. The aforesaid model is fitted with the data by least square technique.

A. Measures for Validation and Comparison of the Models

Two procedures have been used for the comparison and the validation of the developed models. These procedures are given bellow.

B. Coefficient of determination (R^2) and (R^2_{adj})

The models were validated on the basis of (R^2) and (R^2_{adj}) which can be computed from the formula given by Drapper and Smith (1988).

Adjusted R^2 is given by the following formula

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

Where $ss_{res}/(n\text{-}p)$ is the residual mean square and $ss_t/(n\text{-}1)$ is the total mean square. The total mean square is constant regardless of how many variables are in the model. On adding a regressed in the model Adjusted R^2 increases only if the addition of the regressed reduces the residual mean square. It also penalizes for adding terms that are not helpful, so it is very important in evaluating and comparing the regression models.

1) Root Mean Square Error (RMSE):- RMSE indicate the magnitude of the average error, but provide no information on the relative size of the average difference between (F) and (O). This method is most popular and important for the validation of the model. The formula for calculating the RMSE is given by Varshneya *et al.* 2010 of forecast is given below:

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

2) Percent Error:- PE is defined as ratio of RMSE to mean observed value expressed as percentage This is another important tool of the model validation. The formula for calculating the Percent Error is given by Varshneya *et al.* 2010 of forecast is given below

$$\text{PE} = \frac{RMSE}{\bar{o}} \times 100$$

C. Percent Deviation:-

The formulae for computation of Percent Deviation of forecast yield from actual yield are given by Mohd. Azfar *et al.* (2015) and Yadav and Sisodia (2015). This measures the deviation (in percentage) of forecast from the actual yield data. The formula for calculating the percent deviation of forecast is given below:

$$Percentage\ deviation = \frac{(\text{Actual yield} - \text{Forecasted yield})}{\text{Actual yield}} \times 100$$

V. Results and Discussion:

Model developed through SRA based on two discriminate scores and time trend T, which is given below:

Y = 532.29 - 39.57 ds1 + 15.06 ds2 + 16.15T

Where;

ds1, and ds2 = Discriminant score 1and Discriminant score 2

T= Time Trend

	Variable	Coefficient	Standard error	P-Value	Significance (R^2)	R^2 (Adjusted)
Zone	Intercept	532.29***	54.98		0.001	0.68***
	ds1	-39.57$^+$	21.83	0.08		
	ds2	15.06***	23.29	0.52		
	T	16.15***	3.48	0.0001		

***P<0.001, **P<0.01,*P<0.05,$^+$P<0.1

Table 1: Model Developed for C.G. Plain Zone on cheakpea through Discriminant Function

	Year	Actual Yield	Predicted Yield	Percent Deviation	R^2	RMSE	PE
Zone	2013	815.1111	904.2837	-10.93	0.68***	139.32	13.77
	2014	1011.111	941.4998	6.88			
	2015	804.2222	1053.512	-30.99			
	2016	1206.444	1071.799	11.16			
	2017	1022.444	1085.579	-6.174			

Table 2: Validation of model:

Table-1 shows on the result of the model. P-value of coefficients indicates that significance of the coefficients.

The table-2 is the table for the validation of model. Minimum percent deviation show for 2015 followed by 2013 and maximum in 2016 followed by 2014. The value of R^2 is highly significant at 1% level of significance. The RMSE and percent error for Zone district is 139.3 and 13.7 respectively.

A. Model Developed for Zone on cheakpea through Principle Component Analysis:-

The multiple regression equation fitted on two principal components and time trend T. Which is given below is:-

Y = 390.30 -724.03 PC1–313.03 PC2 -269.73 PC3+ 16.24 T

	Coefficient	Standard error	P-Value	Significance (R^2)	R^2 (Adjusted)
Intercept	390.30	208.1744		0.0007	0.73***
PC1	724.08	1053.442	0.49		
PC2	-313.03	173.8238	0.08		
PC3	-269.73	106.9219	0.01		
T	16.24	3.502146	0.0001		

***P<0.001, **P<0.01,*P<0.05,$^+$P<0.1

Table 3: Model Developed for Zone on cheakpea through Principle Component

	Year	Actual Yield	Predicted Yield	Percent Deviation	R^2	RMSE	PE
Zone	2013	815.11	925.677	-13.56	0.73^{***}	132.89	13.35
	2014	1011.11	1000.946	1.00			
	2015	804.2222	997.5723	-24.04			
	2016	1206.444	1010.876	16.21			
	2017	1022.444	1041.006	-1.81			

$^{***}P<0.001$, $^{**}P<0.01$,$^{*}P<0.05$,$^{+}P<0.1$

Table 4: Validation of Model

Total 28 principal components generated for Zone District. Out of 28 only two components cover more than 80% variation of the data. These two components and Time trend T has been used to fitted the model. Table -3 represents that both the variables and time trend T are significant. Table-4 shows that percent deviation is ranging from -1.81to 16.26. RMSE, PE and R^2 is 132.89, 13.35 and 0.73 respectively.

The result of R^2, Root Mean Square Error (RMSE), Percent Error (PE) and Percent Deviation indicates that model is very good for the pre harvest forecasting.

REFERENCES

[1] Anderson, T.W. (1984) an Introduction to Multivariate Statistical Analysis, *John Wiley & Sons*, New York
[2] AzfarMohd.Sisodia.B.V.S. Rai.V.N& Devi Monika.2015.Pre-harvest forecasting models for rapeseed and mustard yield using principle component analysis of weather variable.*Mausam*, 761-766
[3] Agrawal, R. Jain, R.C. and Jha, M.P (1986) Models for studying rice crop weather relationship. *Mausam*, 37 (1), 67-7
[4] Draper N.R. and Smith H. (1988) Applied Regression Analysis. 3rd edition, *John Wiley & Sons Inc*
[5] Agrawal, R; Jain, R.C. and Mehta, S.C. (2001) Yield forecast based on weather variables and agricultural input on agrioclimatic zone basis. *Ind. J. Agric. Sci.*,71 (7), 487-90
[6] Pandey K.K RaiV.N.SisodiaB.V.S.Bharti A.K. Gairola K.C (2013) Pre -Harvest Forecast Models Based On Weather Variable And Weather Indices For Eastern U.P. *Advances in Bioresearch*, 118- 122
[7] Pandey K. K. Rai V. N. Sisodia B. V. S. And Singh S. K. 2015. Effect of weather variables on rice crop in eastern uttarpradesh, india .*Plant archives*, 575-579
[8] Kumar Neeraj, Pisal R. R, shukla S. P and Pandey K. K (2016) Crop yield forecasting of paddy and sugarcane through modified hendrick and scholl technique for south Gujarat. *Mausam*, 405-410
[9] Pandey K.K, D Maurya, Gupta G and Mishra S.V (2016).Yield forecasting models based on weather parameters for eastern u.p. *Vegetos, An International Journal of Plant Research.*
[10] Pandey K.K., Rai V.N, Sisodia B.V.S, Pandey R.and Pandey V. (2016) Development of forecasting model on weather parameters for eastern U.P.. *Environ. Life sci.rel.sci*, 1370-1372.
[11] Pandey K.K. Chandrakar C.K, Singh S.K. Maurya. D and Gupta G (2017) Identification of most important weeks on minimum temperature for wheat crop, *International Journal of Current Microbiology and Applied Sciences*, 788-794.
[12] Wei, B., Varshney, A., Patwari, N., Hu, W., Voigt, T. and Chou, C.T., 2015, April. drti: Directional radio tomographic imaging. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks* (pp. 166-177).
[13] Abu-Ellail, F.F., Hussein, E.M. and El-Bakry, A., 2020. Integrated selection criteria in sugarcane breeding programs using discriminant function analysis. *Bulletin of the National Research Centre*, *44*(1), pp.1-14.
[14] Smith, A.M. and Blackshaw, R.E., 2003. Weed–crop discrimination using remote sensing: a detached leaf experiment. *Weed Technology*, *17*(4), pp.811-820.

