*Journal of Agrometeorology 23 (1) : 100-105 (March 2021)*

# Rice yield forecasting using agro-meteorological variables: A multivariate approach

**GURMEET NAIN[1], NITIN BHARDWAJ[2], P.K. MUHAMMED JASLAM[2*], CHANDER SHEKHAR DAGAR[3] and ANURAG[3]**

*[1]Division of Agricultural Economics, IARI, Pusa, New Delhi- 110012*

*[2]Department of Mathematics and Statistics, [3]Department of Agricultural Meteorology, CCS HAU, Hisar, Haryana- 125004*

**Corresponding author email: jaslam.stat@hau.ac.in*

The weather variables impact the crop differently throughout the various stages of development. The weather effect on crop yield thus can be determined not only by the magnitude of weather variables but also on the variability of weather over crop season. Crop yield forecasting methods incorporating weather information provide a better prediction of yield accounting the relative effects of each weather component. Regression analysis is the most frequently used statistical technique for investigating and modelling the relationship between variables. Building a multiple regression model is an iterative process. Usually several analyses are required for checking the data quality as well as for improvement in the model structure. The use and interpretation of multiple linear regression models depends on the estimates of individual regression coefficients. However, in some situations the problem of multicollinearity exists when there are near linear dependencies between/among the independent variables. The Principal Component Analysis (PCA) method has been proposed to address the problem of multicollinearity. Using principal component scores (PC) derived from weather variables as predictor variables helps to obtain better estimate the yield. The discriminant analysis is a multivariate technique involving the classification of separate sets of objects (or sets of observations) and assigning of new objects (or observations) to the groups defined previously. Forecasting of crop yield can also be done using discriminant analysis scores based on the weather variables as regressor.

***Keywords -*** Multiple linear regression, principal component analysis, discriminant function analysis, pre-harvest forecast, crop yield and weather variables

The weather variables impact the crop differently throughout the various stages of development. The degree of weather effect on crop yield thus be determined not only by the magnitude of weather variables but also on the variability of weather over crop season. Crop yield forecasting methods incorporating weather information's provides a better prediction of yield accounting the relative effects of each weather components. Regression analysis is the most frequently used statistical technique for investigating and modelling the relationship between variables. Building a regression model is an iterative process. Usually several analyses are required as improvement in the model structure and flaws in the data are discovered. The use and interpretation of multiple linear regression models depends on the estimates of individual regression coefficients. However, in some situations the problem of multicollinearity exists when there are near linear dependencies between/among the independent variables. The Principle component analysis method has been proposed to address when the problem of multicollinearity. Using principle component scores derived from weather variables as predictor variable helps to obtain better estimate the yield. The discriminant analysis is a multivariate technique involving the classification of separate sets of objects (or sets of observations) and the assigning of new objects (or observations) to the groups defined previously. Forecasting of crop yield can also be done using discriminant analysis scores based on the weather variables as regressor.

Rai and Chandrahas (2000) used the discriminant function analysis of weather variables to develop statistical models for pre-harvest forecasting of rice-yield in Raipur district of Chhattisgarh. Bal *et al.* (2004) developed agro-meteorological wheat yield forecasting models using multiple regression technique for the Ludhiana district. Regression models were found to perform better by combining technology trend with weather variables (Mallick *et al.,* 2007). Agrawal *et al.* (2012) have derived prediction models for wheat yield in Kanpur district (U.P.) using discriminant functions study of weekly weather data. To overcome the difficulty of multicollinearity observed among plant biometric

characters Verma *et al.* (2015) developed crop yield models within the framework of principal component analysis (PCA). Goyal (2016) made a comparative study of different multivariate techniques (multiple linear regression, principal component analysis and discriminant function technique) for pre-harvest wheat yield estimation in Hisar (Haryana). Singh and Sharma (2017) developed reliable pre-harvest forecasting models for maize yield in various districts of Himachal Pradesh using composite weather indices. The principal components of the weather parameters spread over the crop growth period were employed to forecast wheat yield in northern zone of Haryana by Goyal and Verma (2018). They inferred that zonal weather models gave the desired predictive accuracy and provided a considerable improvement in the district-level wheat yield prediction.

The present study has been carried out to compare the performance of different multivariate statistical methods such as multiple linear regression, principal component analysis and discriminant function analysis for pre-harvest forecast of rice yield of Karnal district of Haryana.

## MATERIALS AND METHODS

In the present investigation, the study region was Karnal district which is located between 29°9'50" to 29°50' North and longitude 76°31'15" to 77°12'45" East. It comes under Haryana's Eastern Plain Zone. It receives approximately 766 mm of rain per year. The soil type is deep alluvial, medium to medium heavy textured soil which is easy to plough. The suitable climate, soil and availability of irrigation facilities make *kharif* rice and wheat cultivation natural choice for the area.

Time series data for *kharif* rice yield of Karnal district of Haryana for 22 years (1996-2017) were collected from the Directorate of Agricultural Statistics and Crop Insurance, Govt. of Haryana. Data on weather variables (weekly) from the year 1996 to 2017 were obtained from the Department of Agricultural Meteorology, Chaudhary Charan Singh Haryana Agricultural University, Hisar, Haryana. The meteorological parameters of the Karnal district of Haryana through the various rice growth phases, up to the first 20 weeks of the crop growing, corresponding to the period from 22nd to 41st Standard Meteorological Week (SMW) of the study years were collected. Five weather variables considered for the study were minimum temperature, maximum temperature, relative humidity, rainfall and hours of sunshine. The data for the first 20-years from 1996 to 2016 were used for model development and the data of the remaining two-years 2016-17 and 2017-18 were used to validate the model.

The influence of weather parameters on crop production not only depends on the immensity but also on their pattern of distribution. Throughout the crop growth period. To study the influence of the weather variables on the yield of rice crop different statistical models were used in the present study

***Multiple linear regression***

Let Y denotes the dependent variable (rice yield) that is linearly related to k independent (or explanatory) variables $x_1, x_{2,\ldots,} x_k$ (weekly weather variables and time trend). The multiple linear regression model is given by -

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon \quad \ldots\ldots\ldots (1)$$

where, $^{2}_{1}, ^{2}_{2}, \ldots\ldots, ^{2}_{k}$ are the regression coefficients associated with dependent variable and weekly weather data on minimum temperature, maximum temperature, relative humidity, rainfall and the sun shine hour as independent variables along with the time trend and e is the random error component, reflecting the difference between the observed and fitted linear relationship. Stepwise regression model was employed as variable selection procedure to obtain the final form of multiple regression model.

***Principal component analysis***

Principal Component Analysis (PCA) is a data reduction technique. It is a multivariate method that needs no user to define the statistical model or hypothesis of the distribution of original variables. In the analysis original set of correlated variables converts in to a new set of uncorrelated variables, Let $x_{ij}$ be the value of $j^{th}$ weather variable (j= 1, 2,...p) corresponding to $i^{th}$ year yield ( i= 1, 2,...n). The PCA for $x_{ij}$'s will be worked out. Let $PC_1, PC_2, \ldots PC_K$ be first K (K< P) principal components with eigen value greater than one. Using these K principal components as independent variables along with time trend and the yield ($Y_i$) as dependent variable, the following multiple linear multiple regression model for the pre-harvest forecast of crop yield has been proposed:

$$Y_i = \beta_0 + \beta_1 PC_{1i} + \beta_2 PC_{2i} + \ldots + \beta_k PC_{ki} + \beta_l T + \varepsilon \quad \ldots (2)$$

where, $Y_i$ is the crop yield of the $i^{th}$ year, $\beta_0, \beta_1, \ldots, \beta_k$ are model coefficients for PC scores (i = 1, 2, ...n) derived from weekly weather data on minimum temperature, maximum temperature, relative humidity, rainfall and the sunshine

hour along with time trend as independent variables, and $\varepsilon$ is error term assumed to follow independently normal distribution with mean 0 and variance.

***Discriminant function analysis***

Discriminant function analysis is a multivariate technique studied in numerous literatures, Anderson (1984), Hair *et al*. (1995), Sharma (1996), Johnson and Wichern (2006). The discriminant analysis is used to identify a suitable function that best discriminates in one of the previously defined groups between sets of observations. The observations are classified on the basis of p variables into non-overlapping k groups. The procedure distinguishes linear functions where the factors coefficients are obtained to maximize the dissimilarity between the groups relative to the dissimilarity within the groups. The highest amount of discriminating functions achieved is uniform to the lowest level of (number of supervised groups -1) and (number of weather variables). These consequences are utilized to figure out the discriminant scores for the classification of observations into different groups. Because the yield obtained over the years is not the same, it can affect many factors, mostly weather parameters due to its uneven distribution throughout the year.

In order to apply discriminating function analysis on weather variables so as to develop the yield model, three groups have been made namely congenial, normal and adverse by partitioning the crop years based on trend-adjusted crop yields. In these three groups, data on climatic indicators have been utilized to obtain linear discriminant functions and for each year the scores of discriminant functions were obtained. In the development of forecast models together with year as regressors and crop yield as regress and, the discriminant scores have been utilized. The number of groups in the present study is three, therefore only two discriminatory functions are sufficient to discriminate between the three groups over a crop year. Three crop year groups obtained were, adverse, normal and congenial:

Adverse $< \bar{y} - s$

Normal $\bar{y} - s$ to $\bar{y} + s$

Congenial $> \bar{y} + s$

where, $\bar{y}$ is the mean yield and *s* denotes the standard deviation.

Using weekly weather data as such in developing the model poses a problem as there would be a huge increase in the number of independent variables in the regression model. In order to solve this problem, weather indices were built from the weekly weather parameters following the Agrawal *et al*. (2012) procedure.

$$Z_{ij} = \frac{\sum_{w=1}^{n} r_{iw}^{j} x_{iw}}{\sum_{w=1}^{n} r_{iw}^{j}} \quad \text{and} \quad Z_{ii',j} = \frac{\sum_{w=1}^{n} r_{ii',w}^{j} x_{iw} X_{i'w}}{\sum_{w=1}^{n} r_{ii',w}^{j}} \quad \ldots\ldots\ldots\ldots\ldots (3)$$

where,

$Z_{ij}$ = Unweighted (j=0) and weighted (j=1) weather indices for i$^{th}$ variable.

$Z_{ii'j}$ = Unweighted and weighted weather indices for interaction between $i^{th} and\ i'^{th}$ weather variable in $w^{th}$ week

$x_{iw}$ = Value of $i^{th}$ variable in $w^{th}$ week.

$r_{iw}$ = Correlation coefficient of adjusted yield for trend with $i^{th}$ weather variable

$r_{ii'w}$ = Correlation coefficient of adjusted yield for trend with product of $i^{th}$ and $i'^{th}$ weather variable in $w^{th}$ week.

*N* = Number of weeks selected for developing indices.

The weather parameters of twenty weeks from the 22$^{nd}$ to 41$^{st}$ SMW of each year were used to develop thirty weighted and un-weighted weather indices along with their interactions. The 30 indices were composed of 5 weighted weather indices and 10 weighted interaction indices, and 5 unweighted weather indices and 10 unweighted interaction indices. Using regression analysis, 3 models are tried. The first 20-year data from 1996 to 2016 were used to model the yield and the remaining two-year 2016-17 and 2017-18 yield data was used to validate the models.

***Model 3.1***

Model 3.1 is Agrawal *et al*. (2012) fourth model . Using the first week data (22$^{nd}$ SMW) on five weather variables, two discriminating functions and there from two sets of discriminating scores were obtained. Two sets of discriminating scores obtained from data of the first week and five weather variables of the second week (23$^{th}$ SMW) were used as discriminating variables, so there were seven discriminating variables in all, and on the basis of these seven discriminating variables the discriminating analysis

was performed and two sets of discriminating scores were obtained. This process was repeated until the last week until the forecast time (41th SMW or 20th week) and finally two sets of discriminating scores were obtained ($ds_1$ and $ds_2$). On the basis of these two sets of scores obtained during the 20th week, the predictive model taking yield as the dependent variable and the discriminating scores and the trend as the regressor variables were equipped. The equation form is given below.

***First iteration***

$$ds_1^1 = a_1W_{11} + a_2W_{12} + a_3W_{13} + a_4W_{14} + .a_5W_{51}$$

$$ds_2^1 = a_1^*W_{11} + a_2^*W_{12} + a_3^*W_{13} + a_4^*W_{14} + .a_5^*W_{51}$$

Second iteration

$$ds_1^2 = a_1W_{21} + a_2W_{22} + a_3W_{23} + a_4W_{24} + .a_5W_{25} + ds_1^1 + ds_2^1$$

$$ds_2^2 = a_1^*W_{21} + a_2^*W_{22} + a_3^*W_{23} + a_4^*W_{24} + .a_5^*W_{25} + ds_1^1 + ds_2^1$$

Third iteration

$$ds_1^3 = a_1W_{31} + a_2W_{32} + a_3W_{33} + a_4W_{34} + .a_5W_{35,} + ds_1^2 + ds_2^2$$

$$ds_2^3 = a_1^*W_{31} + a_2^*W_{32} + a_3^*W_{33} + a_4^*W_{34} + .a_5^*W_{35} + ds_1^2 + ds_2^2$$

Fifth iteration

$$ds_1^5 = a_1W_{51} + a_2W_{52} + a_3W_{53} + a_4W_{54} + .a_5W_{55} + ds_1^4 + ds_2^4$$

$$ds_2^5 = a_1^*W_{51} + a_2^*W_{52} + a_3^*W_{53} + a_4^*W_{54} + .a_5^*W_{55} + ds_1^4 + ds_2^4$$

20st iteration

$$ds_1^{20} = a_1W_{1,20} + a_2W_{2.20} + a_3W_{3,20} + a_4W_{4,20} + .a_5W_{5,20} + ds_1^{19} + ds_2^{19}$$

$$ds_2^{20} = a_1^*W_{1,20} + a_2^*W_{2.20} + a_3^*W_{3,20} + a_4^*W_{4,20} + .a_5^*W_{5,20} + ds_1^{19} + ds_2^{19}$$

And finally developed the model with $ds_1$, $ds_2$ and T is,

$$Y = \beta_0 + \beta_1 ds_1^{20} + \beta_2 ds_2^{20} + \beta_3 T \quad \ldots\ldots\ldots\ldots\ldots (4)$$

where, $W_{ij}$ is the i[th] weather variable for $j^{th}$ week.

***Model-3.2***

In this model, for the first weather variable, discriminating function analysis was performed using the unweighted and weighted averages (weather indices) (there will be only two discriminating factors). Using the two sets of discriminating scores obtained on the basis of the first weather variable and unweighted and weighted averages (weather indices) for the second weather variable, discriminating function analysis was carried out further Here, there will be four discriminating factors. This process has been continued up to fifth weather variables, and finally we have two sets of discriminating scores $ds_1$ and $ds_2$, and the forecast model is developed as follows.

***First iteration***

$$ds_1^1 = a_1Z_{11} + a_2Z_{10}$$

$$ds_2^1 = a_1^*Z_{11} + a_2^*Z_{10}$$

Second iteration

$$ds_1^2 = a_1Z_{21} + a_2Z_{20} + ds_1^1 + ds_2^1$$

Third iteration

$$ds_1^3 = a_1Z_{31} + a_2Z_{30} + ds_1^2 + ds_2^2$$
$$ds_2^3 = a_1^*Z_{31} + a_2^*Z_{30} + ds_1^2 + ds_2^2$$

Fifth iteration

$$ds_1^5 = a_1Z_{51} + a_2Z_{50} + ds_1^4 + ds_2^4$$
$$ds_2^5 = a_1^*Z_{51} + a_2^*Z_{50} + ds_1^4 + ds_2^4$$

And finally developed a model with,and T is,

$$Y = \beta_0 + \beta_1 ds_1^5 + \beta_2 ds_2^5 + \beta_3 T \quad \ldots\ldots\ldots\ldots\ldots (5)$$

where, are the averages of weighted and un-weighted weather indices

***Model 3.3***

In this method the weighted and unweighted indices of weather variables and their interactions were considered. Two discriminating scores were obtained from a total of 30 indices determined using above methods. The regression model was designed to take yield as the predictor variable and discriminating scores and the trend as the independent variables. The model fitted here is

$$Y = \beta o + \beta_1 ds_1 + \beta_2 ds_2 + \beta_3 T \quad \ldots\ldots\ldots (6)$$

where, Y = crop yield

$\beta_0$ = intercept of the model

$\beta_{1's}$ = the regression coefficients

$ds_1$, $ds_2$ are the two discriminant scores and $T$ is the time trend

**Table 1:** Yield forecast models for rice crop of Karnal district of Haryana

| Models | Forecasting regression models | $R^2$ (%) | $R^2_{adj}$ (%) | SE |
|---|---|---|---|---|
| 1 | $Y = 3268.75 - 99.63sh_{19} + 38.11T$ | 58.2 | 53.7 | 256.29 |
| 2 | $Y = 2440.48 + 160.45pc_{11} - 121.01pc_{16} - 120.33pc_{15} + 44.47T$ | 79.1 | 74.1 | 191.71 |
| 3.1 | $Y = 2694.403 + 1.419ds_1^{20} - 18.487ds_2^{20} + 22.401T$ | 85.8 | 83.1 | 161.01 |
| 3.2 | $Y = 2463.541 + 148.067ds_1^{5} - 94.559ds_2^{5} + 44.023T$ | 84.2 | 81.3 | 169.40 |
| 3.3 | $Y = 2544.320 - 50.079ds_1 + 131.826ds_2 + 36.230T$ | 78.2 | 74.1 | 199.19 |

**Table 2:** Variance (%) explained by different principal components

| Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| % of Variance | 13.553 | 11.791 | 10.884 | 7.608 | 6.708 | 6.279 | 5.800 | 5.157 | 4.598 | 4.196 |
| Cumulative % | 13.553 | 25.344 | 36.228 | 43.835 | 50.544 | 56.823 | 62.623 | 67.780 | 72.378 | 76.575 |
| Component | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| % of Variance | 3.407 | 3.096 | 2.964 | 2.669 | 2.369 | 1.946 | 1.885 | 1.684 | 1.433 | 1.165 |
| Cumulative % | 79.982 | 83.078 | 86.043 | 88.711 | 91.080 | 93.026 | 94.911 | 96.595 | 98.027 | 99.192 |

**Table 3:** Percent relative deviation of statistical models for 2016-17 & 2017-18

| Year | | Model 1 | Model 2 | Model 3.1 | Model 3.2 | Model 3.3 |
|---|---|---|---|---|---|---|
| 2016-2017 | Actual yield (kg ha$^{-1}$) | 3068 | 3068 | 3068 | 3068 | 3068 |
| | Estimated yield (kg ha$^{-1}$) | 3172 | 3252 | 3103 | 3390 | 3203 |
| | PED | 3.29 | 5.66 | 1.15 | 10.51 | 4.39 |
| 2017-2018 | Actual yield (kg ha$^{-1}$) | 3171 | 3171 | 3171 | 3171 | 3171 |
| | Estimated yield (kg ha$^{-1}$) | 3397 | 3356 | 3138 | 3242 | 3236 |
| | PED | 6.65 | 5.52 | 1.04 | 2.25 | 2.04 |

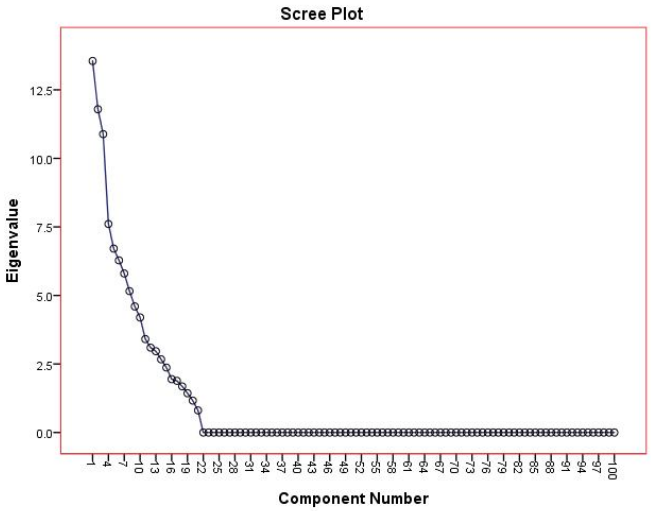


**Fig. 1:** Scree plot of principal component analysis of weekly weather variables

## RESULT AND DISCUSSION

### *Yield forecast models*

The final model equation derived from stepwise Multiple Regression Analysis, Principal Component Analysis and Discriminant Function Analysis are given in Table 1. The adding variable cutoff probability of 0.05 and removal cutoff 0.10 are assigned in stepwise regression for these models.

In the multiple regression model, the time trend (T) entered in to the equation initially. In the second stage, sun shine hour at 19$^{th}$ week ($sh_{19}$) also entered in to the regression equation.

In principal component analysis (PCA) study, first 20 eigen values of the correlation matrix of weather variables suggested 20 factor solution. Eigen values and percent variance explained by different PCs are presented in Table 2 and the Scree plot is given in Fig. 1. The final form of the regression model via step-wise regression method showed

the principal components, pc11, pc13 and pc15 and time trend of yield (T) as the dependent variables for forecasting of wheat yield.

In discriminant function analysis the regression equations for the three models (3.1, 3.2 and 3.3) developed based on discriminate scores (ds) and time trend (*T*) values are presented in Table 1. The individual t-test results showed that probability values of calculated test statistic of intercept, coefficient of $ds_2^{20}(\beta_2)$ and coefficient of $T(\beta_3)$ were highly significant ($P < 0.01$), whereas coefficient of was non-significant in model 3.1. However, in model 3.2 and 3.3 all the coefficients respect to *ds* and intercepts were statistically significant at 1% level.

***Comparison of Model accuracy***

It is evident from the Table 1 that the coefficient of determination ($R^2$) for the simple multiple regression model was the lowest (0.582). The PCA based multiple regression model showed $R^2$ of 0.791whereas the $R^2$ values of discriminant function analysis-based models were 0.858, 0.842 and 0.782 for models 3.1, 3.2 and 3.3 respectively. The standard error was also found to be the lowest (161.01) in discriminant function analysis-based model 3.1.

The model predicted yields along with observed yields and per cent relative deviations for the validating years 2016-17 and 2017-18 are given in Table 3.

It is evident from the results that percent deviation of forecast yield from actual yield varied from 1.04 to 10.51 from all five models validated using the data of 2016-17 and 2017-18. Among all these models the discriminant function analysis-based model, 3.1 showed the lowest percent relative deviation in both the years of validation study.

## CONCLUSION

The results indicate the preference of using prediction equations-based discriminant scores as predictor variable over using principal component scores or weather parameters as such as predictor variables. Among the models based on the discriminant score, the model 3.1 suggested by Agrawal *et al*. (2012) comes out to be the best model on account of high values of adjusted $R^2$ and the lower value of standard error followed by model 3.2. Percent relative deviation also suggest model 3.1 as the most suitable model to forecast wheat yield in Karnal district of Haryana. Therefore, from the suggested model 3.1, a reliable prediction of wheat yield can be achieved about one and a half months before harvest.

## REFERENCES

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 applied multivariate statistical analysis, John Wiley & Sons Inc. New York.

Bal, S.K., Mukherjee, J., Mallick, K. and Hundal, S.S. (2004). Wheat yield forecasting models for Ludhiana district of Punjab state. *J. Agrometeorol.,* 6 (Sp. Issue): 161-165.

Goyal, M. (2016). Use of Different Multivariate Techniques for Pre-Harvest Wheat Yield Estimation in Hisar (Haryana). *Int. J. Comput. Math.,* 12(5): 6-11.

Goyal, M. and Verma, U. (2018). Wheat yield prediction using weather based statistical model in northern zone of Haryana. *Int. J. Humanit. Soc. Sci.,* 7(4): 47-50.

Hair, J. F., Anderson, R. E., Tatham, R. L. and Black, W. C. (1995). Multivariate Data Analysis with Readings, Prentice Hall, Inc. New Jersey.

Johnson, R. A. and Wichern, D. W. (2006). Applied Multivariate Statistical Analysis. Pearson Education.

Mallick, K., Mukherjee, J., Bal, S.K., Bhalla, S.S. and Hundal, S.S. (2007). Real time rice yield forecasting over Central Punjab region using crop weather regression model. *J. Agrometeorol.,* 9(2): 158-166.

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.

Singh, M. and Sharma, S. (2017). Forecasting the maize yield in Himachal Pradesh using climatic variables. *J. Agrometeorol*., 19 (2): 167-169.

Verma, U, Aneja, D.R and Hooda, B.K. (2015). Principal component technique for pre harvest estimation of cotton yield based on plant biometrical characters. *J. Cotton Res. Dev*., 29(2): 339-343.

*Received : February 2020 ; Accepted : December 2020*