Print ISSN : 0972-8813
e-ISSN : 2582-2780

[Vol. 22(3) September-December 2024]

Pantnagar Journal of Research

(Formerly International Journal of Basic and Applied Agricultural Research ISSN : 2349-8765)

G.B. Pant University of Agriculture & Technology, Pantnagar

PANTNAGAR JOURNAL OF RESEARCH

Vol. 22(3) September-December 2024

CONTENTS

Application of principal component analysis and discriminant function analysis in developing prediction models to forecast maize yield using weather indices

ANITA YADAV*and A.K. SHUKLA

Department of Mathematics, Statistics and Computer Science, College of Basic Sciences and Humanities, G.B. Pant University of Agriculture and Technology, Pantnagar-263145 (U.S. Nagar, Uttarakhand)
**Corresponding author's email id: anitay225@gmail.com*

ABSTRACT: The importance of precise crop yield forecasts cannot be overstated, as they serve as a critical input for policy formulation and implementation related to crop procurement, distribution, price structure and import/export decisions. Our objective is to develop a reliable forecasting model that can inform evidence-based decision-making for policymakers and stakeholders. A 21-year dataset was split into training (17 years) and testing (4 years) subsets. Weather indices were developed using weekly data, in accordance with the method outlined by Agrawal *et al.* (1983).This study explored three multivariate models for predicting maize yields based on weather variables: SMLR (Stepwise Multiple Linear Regression), PCA (Principal Component Analysis) and DFA (Discriminant Function Analysis). The performance of the model was assessed using two primary metrics: Adjusted R-squared (adj. R^2), which indicates the proportion of variance explained by the model and Root Mean Square Error (RMSE), which measures the average magnitude of prediction errors. On the basis of adj R^2 (89.98 %) and RMSE (218.54 Kg/ha), the DFA-SMLR model performed best for maize yield prediction compared to SMLR and PCA-SMLR models in the studied region.

Key words: Discriminant function analysis, maize yield prediction, model, principal component analysis, stepwise multiple linear regression

Maize (*Zea mays* L.) is one of the most versatile crops, capable of thriving in a variety of agro-climatic conditions. It is planted over nearly 190 million hectares in approximately 165 countries with a broader diversity of soil, temperature, biodiversity and management approaches, accounting for 39% of global grain production. The United States of America (USA) is the world's largest producer of maize, accounting for nearly 30.99% of global production in 2020 and maize is the primary driver of the US economy. In India, maize is farmed all year. It is primarily a Kharif crop, with 85 percent of the region being cultivated during the monsoon season.Yield forecasting is essential for improving decision-making in agriculture by helping farmers, businesses and governments predict crop outcomes, manage risks and optimize resources. Accurate forecasts enable better market planning, fair pricing and targeted policies for food security. It also supports sustainable agriculture by optimizing input use and minimizing waste. Techniques like Principal Component Analysis (PCA), Discriminant Function Analysis (DFA) and Stepwise Multiple Linear Regression (SMLR) offer distinct advantages in data analysis. PCA reduces data complexity by transforming multiple variables into fewer, uncorrelated components, improving visualization and noise reduction. DFA helps in classifying data by maximizing the separation between groups, making it ideal for predictive classification tasks. SMLR enhances model accuracy by automating variable selection, ensuring only the most relevant predictors are included, which simplifies the model and improves its predictive power. These methods collectively assist in creating more accurate, efficient and interpretable yield forecasting models. Various efforts have been made to create crop yield prediction models (Taffler *et al.,* 1982; Aneja *et al.,* 1984; Jain *et al.,* 1985; Chandrahas *et al.,* 1993; Draper *et al.,* 1998; Chauhan *et al.*, 2009; Gosh *et al.*, 2014; Banakara *et al.*, 2018; Diwan *et al.*, 2018; Arvind *et al.*, 2022; Das *et al.*, 2022; Sridhara *et al.*, 2023).

Various researchers have previously attempted to create statistical models based on time series data on crop yield and meteorological variables for agricultural yield predictions prior to harvest.

Regression models have been utilized by Fisher (1924), Hendricks and Scholl (1943), Agrawal *et al.* (1980, 1983, 1986 and 2001), Jain and Singh (1980) and others. Moreover, Rai and Chandrahas (2000), along with Agrawal *et al.* (2012), aimed to develop statistical models for predicting agricultural yield by applying discriminant function analysis to weather indices and weekly weather variable data. The present study aims to develop suitable statistical models for maize yield forecasting in the Udham Singh Nagar district of Uttarakhand by applying Principal Component Analysis (PCA - It is a linear technique that reduces the dimensionality of data while retaining most of its original information) and Discriminant Function Analysis (DFA– It is a statistical technique that finds linear combinations of features to separate two or more groups of objects or events) to weather indices derived from weather variables. By applying DFA and PCA to various weather variables, our objective was to develop robust statistical models for predicting yields in this region, thereby enabling data-driven decision-making for agricultural planning and management.

MATERIALS AND METHODS

Description of the study area

Udham Singh Nagar is a district located in the state of Uttarakhand, India. It is situated in the Tarai region, which is characterized by flat terrain and relatively lower elevations compared to the hilly areas of Uttarakhand.It is situated between 78° 45' E and 80° 08' E longitude and spans between 28° 53' N and 29° 23' N latitude. The climate ranges from sub-tropical to sub-humid and has three distinct seasons, summer, monsoon and winter. The soils are shallow, with a texture ranging from sandy to loamy. They are calcareous and moderately fertile.

Description of the Data

The maize crop yield data for the period 2001-2021 were obtained from the Directorate of Economics and Statistics, Ministry of Agriculture and Farmers Welfare website, whereas the weather data were gathered from the Meteorological Observatory at the Department of Agrometeorology, College of Agriculture, G.B.Pant University of Agriculture and

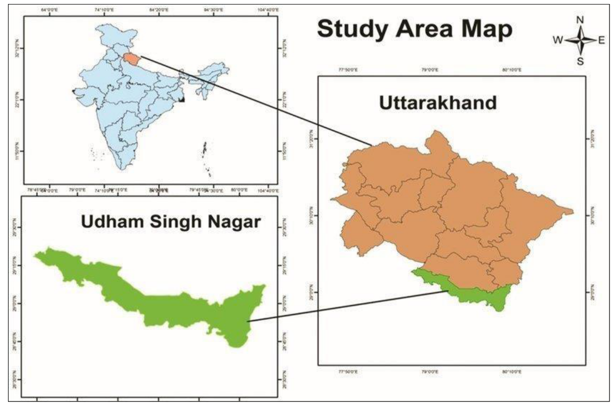

Fig. 1: Study location

Technology, Pantnagar, Uttarakhand. The study considered eight weather parameters: Maximum temperature, Minimum temperature, Relative Humidity I, Relative Humidity II, Total rainfall, Sunshine hours, Wind velocity and Evapotranspiration.

Software used

The data were analyzed using software such as SPSS and MS Excel.

Statistical procedure

A 21-year (2001-2021) dataset was divided into two subsets, with approximately 80% of data used for model training and the remaining 20% reserved for model testing. Weekly weather data were used to generate weather indices, following the methodology outlined by Agrawal *et al.* (1983). In this study three multivariate weather based models including stepwise multiple linear regression (SMLR), principal component analysis (PCA) and discriminant function analysis (DFA) were investigated. Johnson *et al.* (2001).

Development of statistical forecast models

This approach is based on the method outlined by Agrawal *et al.* (1986) for developing forecasts using weather indices. In this procedure, the entire 15-week dataset is utilized to construct both weighted and unweighted weather indices for various weather variables, including their interactions. The analysis resulted in 72 indices, categorized into two groups: 36 weighted indices and 36 unweighted indices. The weighted indices include 8 weather indices and 28 interaction indices, while the unweighted indices consist of 8 weather indices and 28 interaction indices. A forecast model has been developed using these 72 indices and the trend variable (T) as predictor variables, with yield as the response variable

Stepwise multiple linear regression (SMLR)

Multiple Linear Regression technique is one of the simplest methods used todevelop yield prediction models. The multiple linear regression equation is typically represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_n X_n + \epsilon \quad ..... \quad (1)$$

In this equation, Y is the dependent variable, X_1, X_2, X_3, ..., X_n are the independent variables, β_0 is the intercept, β_1, β_2, β_3, ..., β_n are the coefficients and ϵ represents the error term.

When used stepwise, the Multiple Linear Regression technique selects the most significant explanatory variables from a large set of independent variables. Theextremely correlated independent variable is incorporated initially to the model. The process is continued by identifying additional significant independent parameter step by step.

Principal component analysis- Stepwise multiple regression model (PCA-SMLR)

The Principal Component Analysis (PCA) is a dimensionality reduction technique that aims to capture the most important features of a dataset while minimizing the loss of information. By transforming the original variables into a new set of uncorrelated variables called principal components, PCA reduces the dimensionality of the data while preserving most of the data's variability. PCA is performed to minimize the risk of overfitting, which can arise from the high dimensionality and intercorrelations among the independent variables. Only the principal components (PCs) with eigenvalues greater than 1 were considered, as recommended by Brejda *et al.* (2000).

In PCA-SMLR technique, PCA score were employed as input for the analysis (Arvind *et al.*, 2022). To address the issue of multicollinearity among weather variables, Principal Component (PC) scores were utilized as regressors for Stepwise Multiple Regression Model (SLMR) in order to construct yield model.Principal Component Analysis (PCA) is used to decompose the original data matrix X into two matrices, P and T, such that $X = TP^t$. The matrix P is known as the "loadings matrix" or "principal component loadings matrix, " and it contains the coefficients or weights that define the principal components. The matrix T, referred to as the "score matrix" or "principal component scores matrix, " is an orthogonal matrix that holds the coordinates of the original data points in the new principal component space. The superscript t indicates the transpose of a matrix.

Discriminant function analysis- Stepwise multiple regression model (DFA-SMLR)

Discriminant analysis is a multivariate statistical method that aims to distinguish between distinct groups of objects or observations and subsequently classify new objects or observations into one of the previously established categories. The technique of discriminant function analysis is used to identify an appropriate function that discriminates best between set of observations from two or more groups and classifying the function observations into one of the previously defined groups.

Consider that observations are classified into k (=3) non-overlapping groups on the basis p (=8) variables. The maximum number of discriminant functions that can be derived is equal to the minimum of (k-1) and p, as stated by Sharma (1996).

For modeling yield using weather variables through discriminant function analysis, the crop years were categorized into three groups based on their trend-adjusted crop yields. The meteorological data for these three groups were then used to develop linear discriminant functions and discriminant scores were calculated for each year. The forecast models are constructed by using the discriminant scores and trend variable as predictors and crop yield as the dependent variable. The models are developed using stepwise regression procedure. In this study, we have three groups of crop years (adverse, normal and congenial) and eight meteorological variables. To classify a crop year into one of these groups, only two discriminant functions are required. We define the groups based on adjusted crop yields, represented by 'y' (average) and's' (range) over 'n' years. The groups are: - Adverse: Yields ≤ y-s , Normal: Yields between y-s and y+sand Congenial: Yields ≥ y+s.

Testing the Performance of the Model

The performance of the developed models was thoroughly evaluated using five key metrics: Coefficient of Determination (R^2), Adjusted Coefficient of Determination (Adj.R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These metrics provided a comprehensive assessment of the models' strengths and weaknesses.

The evaluation criteria were clear: higher R^2 and Adj.R^2 values (closer to 1) and lower RMSE values (closer to 0) indicated better performance.

RESULTS AND DISCUSSION

Yield prediction models of maize crop

Maize yield prediction models were developed utilizing data on maize yield (in kg/ha) and corresponding weather datafor a period of 21years. For training the model 17 years data, from 2001 to 2017 were used and for testing the model 4 years data, from 2018 to 2021 were used. The training set enables the model to learn and define the relationship between the independent variables and the dependent variable.A testing data set is used to assess the prediction accuracy of the developed models, providing an evaluation of their performance. In this study three multivariate weather based models including stepwise multiple linear regression (SMLR), principal component analysis (PCA) and discriminant function analysis (DFA) were investigated.

Table 2: Comparison between actual and forecast yield of different years using SMLR

Model	Year	Actual yield (kg/ha)	Forecasted yield(kg/ha)	Error (%)
SMLR	2018	1324.50	1301.06	1.77
	2019	1338.24	1256.37	6.12
	2020	2174.60	2202.21	-1.27
	2021	2688.20	2749.38	-2.27

Table 3:Comparison between actual and forecast yield of different years using PCA-SMLR

Model	Year	Actual yield (kg/ha)	Forecasted yield(kg/ha)	Error (%)
PCA-SMLR	2018	1324.50	1295.23	2.21
	2019	1338.24	1293.80	3.32
	2020	2174.60	2088.03	3.98
	2021	2688.20	2703.65	-0.57

Table 4:Comparison between actual and forecast yield of different years using DFA-SMLR

Model	Year	Actual yield (kg/ha)	Forecasted yield(kg/ha)	Error (%)
DFA-SMLR	2018	1324.50	1300.65	1.80
	2019	1338.24	1298.32	2.98
	2020	2174.60	2097.33	3.55
	2021	2688.20	2596.21	3.42

The performance metrics for the SMLR model, PCA-SMLR model and DFA-SMLR model are summarized in Table 1, which displays the statistical indices for each model. In SMLR model Z_{31}, Z_{351}, Z_{141}, $Z_{380}$and Z_{240} are weighted and unweighted weather indices. In PCA-SMLR model P_1, P_2, P_3, P_4, P_5 and P_6 are first, second, third, fourth, fifth and sixth principal components respectively. In DFA-SMLR model ds_1 and ds_2 is discriminant score of function 1 and function 2 respectively.

Table 1 reveals that the performance of model DFA-SMLR is found to be excellent with lower

Table1: Yield Prediction Models developed using SMLR and PCA-SMLR

Model	Equation	R^2 (%)	Adj. R^2 (%)	RMSE (Kg/ha)
SMLR	$Y=321.02-204.32Z_{31}-102.67Z_{240}-87.15Z_{141}-398.24Z_{380}-64.97Z_{351}-104.22T$	86.32	71.02	345.12
PCA-SMLR	$Y=432.12-120.56P_1-52.32P_2+107.65P_3-25.47P_4-54.94P_5+142.86P_6-63.69T$	93.56	87.32	264.20
DFA-SMLR	$Y = 218.02 + 9.35ds_1 + 21.85ds_2 – 6.01T$	95.13	89.98	218.54

RMSE (218.54) and high Adj.R^2 (89.98%). On the other hand, model SMLR performs poorly in comparison with model PCA-SMLR and DFA-SMLR. The model DFA-SMLR is the most appropriate one for the forecast of the maize yield during kharif season in U. S. Nagar district of Uttarakhand.

Error analysis is carried out to evaluate the performance of the models is presented in Table 2, 3 & 4 and a comparison of Actual rice yield with the Forecasted yield is also presented in Table 2, 3 & 4.

Table 2, 3 & 4 reveals that the error percentage of DFA-SMLR model varied from -1.80% to 3.55% and it is the least range of error percentage recorded as compared to SMLR and PCA-SMLR model. Overall predictive performance of model DFA-SMLR found to be excellent. This finding is similar to the finding of Garde *et al.* (2020), who observed that the discriminant function analysis yielded highly accurate predictions.

CONCLUSION

Predicting crop yield offers significant benefits in agriculture, including improved resource management, optimized practices and better financial planning. Accurate yield forecasts help farmers adjust irrigation, fertilization and pest control, leading to higher productivity. It also supports supply chain management by forecasting crop availability and demand, reducing waste. Yield prediction aids in adapting to climate change by enabling farmers to select resilient crop varieties and adjust practices. In the future, technologies like AI, machine learning and IoT sensors will improve the accuracy and timeliness of predictions. These advancements will be crucial for enhancing food security, crop management and sustainability in agriculture. In this study, three multivariate weather-based models—stepwise multiple linear regression (SMLR), principal component analysis (PCA) and discriminant function analysis (DFA)—were examined. The performance of these models was evaluated using Adjusted R-squared (adj R^2) and Root Mean Square Error (RMSE). Based on an adj R^2 of 89.98% and an RMSE of 218.54 kg/ha, the DFA-SMLR model outperformed the SMLR and PCA-SMLR models in predicting maize yield in the studied region.

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.

Arvind, K., Vashisth, A., Krishanan, P. and Das, B. (2022). Wheat yield prediction based on weather parameters using multiple linear, neural network and penalised regression models. *J. Agrometeorol.*, 24(1): 18-25. https: //doi.org/10.54386/jam.v24i1.1002

Banakara, K. B., Garde, Y. A., Pisal, R. R. and Bhatt, B. K. (2018). Pre-Harvest Forecasting of Rice Yield forEffective Crop Planning Decisionin Surat District of South Gujarat, India. *Int. J. Curr. Microbiol. App. Sci.*, 7(6): 3410-3422.

Banakara, K.B., Popat, R.C., Amaresh and Pandya, H.R.(2018).Pre-harvest forecast of Kharif Rice using weather parameters in Bharuch district of Gujarat state. *Int.J. Chem. Stud.*, 6(6): 1111-1116.

Brejda, J. J., Moorman, T. B., Karlen, D. L. and Dao, T. H. (2000). Identification of Regional Soil Quality Factors and Indicators, I. Central and Southern High Plain. *Soil Sci. Soc. Am. J.*, 64(6): 2115-2124.

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.

Chauhan, V. S., Shekh, A. M., Dixit, S. K., Mishra, A. P. and Kumar, S. (2009).Yield prediction model of rice in Bulsar district of Gujarat. *J. Agrometeorol*., 11(2): 162-168.

Das, B., Murgaonkar, D., Navyashree, S. and Kumar, P. (2022). Novel combinationartificial neural network models could not outperform individual models forweather-basedcashewyield prediction.*Int. J.Biometeorol*., 66: 1-12.

Diwan, U. K., Puranik, H. V., Das, G. K. and Chaudhary, J. L. (2018). Yield prediction of wheat at pre harvest stage using regression based statistical model for 8 district of Chhattisgarh, India*. Int. J. Curr. Microbiol. Appl. Sci*., 7(1): 2180-2183.

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.

Garde, Y. A., Thorat, V. S., Pisal, R. R. and Shinde, V.T. (2020). Pre-Harvest Forecasting of Kharif Rice Yield Using Weather Parameters for Strategic Decision Making inAgriculture. *Int. J. Environ. Clim.,* 10(12): 162-170.

Ghosh, K., Balasubramanian, R., Bandopadhyay, S., Chattopadhyay, N., Singh, K. K. and Rathore, L. S. (2014). Development of crop yield forecast modelsunder FASAL-a case study of kharif rice in West Bengal. *J. Agrometeorol.*, 16(1): 1-8.

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., 74p.

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.

Sharma, S. (1996). Applied Multivariate Techniques. John Wiley & Sons, New York.

Sridhara, S., Manoj, K. N., Gopakkali, P., Kashyap, G. R., Das, B., Singh, K. K. and Srivastava, A. K. (2023). Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India. *Int. J. of Biometeorology*, 67(1): 165-180.

Taffler, R. J. (1982). Forecasting company failure in the UK using discriminant analysis and financial ratio data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 145(3): 342-358.

Received: November 23, 2024
Accepted: December 22, 2024