Original Research Article

Prediction of maize crop yield using principal component analysis of weather parameters

Abstract

The use of principal component analysis in the development of statistical models for crop yield forecasting has been demonstrated. Maize crop yield data for a period of 21 years (2001-2021) were drawn from the Dacnet website and the weather data were collectedfrom the Meteorological Observatory, Department of Agrometeorology, College ofAgriculture,G.B.PantUniversityofAgricultureandTechnologyPantnagar,Uttarakhand. Maximum temperature, Minimum temperature,Relative Humidity A.M, Relative Humidity P.M, Total rainfall, Sunshine hours, Wind velocity andEvapotranspirationweretheweatherparametersconsideredforthestudy.Out of the 21-year data, 17-year data were used for training the model while remaining 4 years data were used for testing the model. Weekly data on weather variables was used to create weather indices (Agrawal *et al*., 1983). This work involves developing forecasting models using principal component analysis (PCA) and multiple linear regression (MLR). Applied to extract principal components (PCs) from the correlation matrix of predictors. Five models (Model 1 to Model 5) are developed using MLR, with varying numbers of PCs as regressors which also include time trend and maize yield as dependent variable. The model performance was measured using Adjusted R-squared (adj R^2) and Root Mean Square Error (RMSE) as goodness of fit criteria. On the basis of adj R^2 and RMSE, model 1 which includes all the calculated weather indices, was found to be best suited model with high adj R^2 (74.18 %) and least RMSE (276.36). Hence, this model can be used to forecast maize yield for the studied region.

Keywords: MaizeYield,PredictionModel,Principal Component Analysis, Weather Parameters

Introduction

Maize (Zea mays L.) is a highly adaptable and resilient crop, capable of thriving in a wide range of agro-climatic conditions. Renowned globally as the "queen of cereals," maize boasts the

highest genetic yield potential among all grains. Its extensive cultivation spans approximately 190 million hectares across 165 countries, encompassing diverse soil types, temperature regimes, biodiversity, and management practices. This accounts for a significant 39% of global grain production. The United States leads the world in maize production, contributing nearly 30.99% of global output in 2020, and plays a vital role in driving the US economy. In contrast, India cultivates maize throughout the year, showcasing its adaptability to various environments.

Maize is a multifaceted crop that extends its utility beyond serving as a primary food source for humans and livestock. Its biochemical constituents, including starch, lipids, and proteins, render it a valuable raw material for various industrial applications. Maize-derived ingredients are integral to the production of numerous products across diverse sectors, such as: Polysaccharides (starch) for paper, textiles, and adhesive industries, Triglycerides (oil) for biofuel, pharmaceutical, and cosmetic applications, Proteins for food, feed, and pharmaceutical industries, Fermentation products (alcoholic beverages, food sweeteners), Cellulose derivatives for film, packaging, and paper products And Additional specialized applications in pharmaceuticals, cosmetics, and other industries.

Numerous studies have previously explored the development of statistical models utilizing time series data on crop yield and meteorological variables to predict agricultural yields**Fisher (1924), Hendricks and Scholl (1943), Agrawal *et al.* (1980, 83, 86, and 2001), Jain and Singh (1980),** and others have all used regression models. **Rai and Chandrahas (2000), as well as Agrawal *et al.* (2012),** sought to create statistical models to forecast agricultural yield using discriminant function analysis of weather indices and weekly data on weather variables. The results obtained from applying discriminant function analysis have been fairly positive.

In this study, we employed principal component analysis (PCA) to analyze weather indices and develop statistical models for predicting maize yields in the Udham Singh Nagar district of Uttarakhand. By applying PCA to weather variables, we aimed to identify the most significant factors influencing maize yield and create reliable forecasting models for this region.

Materials and Methodology

Description of the study area

The study area for developing Yield Prediction models encompasses the Udham Singh Nagar district in Uttarakhand state, specifically focusing on kharif maize yield data (kg/ha) and corresponding weather data. Located in the Kumaon Division's Terai region, this district spans across 78° 45' E to 80° 08' E longitude and 28° 53' N to 29° 23' N latitude. The region experiences a sub-tropical to sub-humid climate, characterized by three distinct seasons: summer, monsoon, and winter. The soil profile is predominantly shallow, with a texture ranging from sandy to loamy.

Description of the data

A 21-year dataset (2001-2021) of maize crop yield was obtained from the Dacnet website, while weather data for the same period was sourced from the Meteorological

Observatory, Department of Agrometeorology, College of Agriculture, G.B. Pant University of Agriculture and Technology, Pantnagar, Uttarakhand. The study considered eight key weather parameters:

- Maximum temperature
- Minimum temperature
- Morning relative humidity (A.M.)
- Afternoon relative humidity (P.M.)
- Total rainfall
- Sunshine hours
- Wind velocity
- Evapotranspiration

These parameters were analyzed to investigate their impact on maize crop yield.

Software used

Data analysis is conducted utilizing statistical software packages, including SPSS and MS-EXCEL, to examine and interpret the data.

Development of weather indices using correlation coefficient as weight

This is based on the method given by **Agrawal et al., (1986)** for developing forecast using weather indices. In this procedure, the entire 15 weeks data have been utilized for constructing weighted and un-weighted weather indices of weather variables along with their interactions. In all, 72 indices (36 weighted and 36 unweighted) consisting of 8 weighted weather indices and 28 weighted interaction indices; 8 un-weighted indices and 28 un-weighted interaction indices have been obtained. These weather indices and interaction indices have been computed by using the following formula:

$$Z_{ij} = \sum_{w=1}^{n} r_{iw}^{j} X_{iw} \text{ and } Z_{ii',j} = \sum_{w=1}^{n} r_{ii'w}^{j} X_{iw} X_{i'w}$$

where,

j = 0, 1 (where, '0' represents unweighted indices and '1' represents weighted indices)

n = Number of weeks considered in developing the indices

r_{iw} = Correlation coefficient between de-trend crop yield and i^{th} weather variable in w^{th} week

$r_{ii'w}$ = Correlation coefficient between de-trend crop yield and the product of i and i'^{th} weather variable in w^{th} week

X_{iw} and $X_{i'w}$ are the i and i'^{th} weather variable in w^{th} week respectively

Statistical procedure

Out of the 21-year data, 17-year data were used for training the model while remaining 4 years data were used for testing the model. Weekly data on weather variables was used to create weather indices **(Agrawal *et al*., 1983)**. In this study five models were created with principal component analysis as independent variables which also include time trend and maize yield as dependent variable.

Principal component method has been used for extraction of factors which consists of finding the eigen values and eigen vectors. The most frequently used convention is to retain the components whose eigen values are greater than one. **Kaiser (1958)** also suggested the dropping of components having eigen roots less than 1. The principal component scores can be used as new regressors in multiple regression analysis for selecting the suitable yield models.

The primary objective of this research is to develop statistical models for maize yield prediction using Principal Component Analysis (PCA), a widely used technique in multivariate analysis, as described in standard texts such as **Johnsonand Wichern (2001)**.

By extracting Principal Components (PCs) from high-dimensional data, researchers can identify the most informative features, reduce noise and redundancy, improve model performance & stability and facilitate model interpretation & visualization.

In this work, using PCs for prediction enables the models to focus on the most important patterns in the data, improve prediction accuracy and enhance model generalizability.

Model 1:

In this approach, all 72 indices were subjected to principal component analysis (PCA), and the first twelve principal components, which collectively explained 91.36 % of the total variance, were selected as regressors for developing the forecasting model. The resulting model takes the form:

$$Y= \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots + \beta_{12} PC_{12} + \beta_{13} T + e$$

where Y is un-trended crop yield, $\beta_i's$ (i=0,1,2,….,13) are the model parameters, PC_1, PC_2, …., PC_{12} are first twelve principal components, T is the trend variable and e is error term assumed to follow normal distribution with mean 0 and variance σ^2.

Model 2:

In this approach, 16 weather indices (8 weighted and 8 unweighted) derived from 8 weather variables were analyzed using principal component analysis (PCA). The PCA results identified the first six principal components as the most significant, explaining 87.95% of the total variance. These six components were then used as regressors to develop a forecasting model, which takes the form:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots + \beta_6 PC_6 + \beta_7 T + e$$

where the notations represent the same variables as defined in Model 1.

Model 3:

In this approach, 36 unweighted weather indices (8 primary indices and 28 interaction terms) were analyzed. The first 8 principal components, which explained approximately 85.87% of the total variance, were selected as regressors for the forecasting model. The resulting model takes the form:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots + \beta_8 PC_8 + \beta_9 T + e$$

where the notations represent the same variables as defined in Model 1.

Model 4:

This approach utilizes 36 weighted weather indices (8 primary indices and 28 interaction terms). The first 8 principal components, explaining 78.96% of the total variance, are chosen as regressors for the forecasting model, which is represented by the following equation:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots + \beta_8 PC_8 + \beta_9 T + e$$

where the notations represent the same variables as defined in Model 1.

Model 5:

This approach utilizes 28 weighted weather interaction indices and 28unweighted weather interaction indices. The first 10 principal components, explaining 92.76% of the total variance, are chosen as regressors for the forecasting model, which is represented by the following equation:

$$Y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \ldots\ldots + \beta_{10} PC_{10} + \beta_{11} T + e$$

where the notations represent the same variables as defined in Model 1.

All the aforesaid models have been fitted with the data pertaining to the years 2001 to 2017 and the data pertaining to the year 2018 to 2021 were used for validation of the forecast models.

Testing the Performance of the Model

Finally the performance of the developed models was evaluated on the basis of 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).

R^2 &Adj.R^2 towards 1 and RMSE towards 0 indicate better performance of the developed models. Also lesser the MAE and MAPE values, better fit the model is.

Results and Discussion

Yield prediction models were developed using maize yield data(kg/ha) and weather data of maize for a period of 21years. For training the model 17 years data, from 2001 to 2017 were used and for testing the model4yearsdata,from2018to2021wereused.Training set establishes the relationship between predictors and dependent variable.Testingdataset determinetheprediction accuracyof developed models. In this study five models were created with principal component analysis as independent variables which also include time trend and maize yield as dependent variable.

Table1: Maize yield forecast models

Model	Forecast equation	R^2 (%)	Adj. R^2 (%)
1.	Yield = 421.12-130.56*PC_1-152.42PC_2+97.65*PC_3-89.47PC_4-102.94*PC_5+14.86PC_6+132.56*PC_7-45.32PC_8+10.65PC_9+205.47*PC_{10}-104.94*PC_{11}+142.86PC_{12}+63.69*T	89.31	79.78
2.	Yield = 542.08+12.78**PC_1-103.30PC_2-83.07PC_3-67.11PC_4-167.47PC_5+14.73*PC_6-47.12**T	62.56	57.32
3.	Yield = 375.21-10.46PC_1-19.88**PC_2+57.22PC_3-12.39PC_4-32.08PC_5+107.11PC_6-98.72*PC_7-37.08PC_8-16.44*T	73.12	65.39
4.	Yield = 194.47-57.18PC_1+82.23PC_2-127.41PC_3-55.12PC_4-38.46PC_5+102.75PC_6+100.87PC_7-67.19PC_8+81.21T	84.95	73.30
5.	Yield = 263.43+40.24PC_1+97.10PC_2+195.44PC_3-61.24PC_4-107.98PC_5+59.05PC_6-55.39.48**PC_7+125.67PC_8-31.28PC_9-71.51PC_{10}-57.97T	72.68	65.87

Note: *Significant at P< 0.05, **Significant at P< 0.01

Forecast models are presented in Table 1 along with their values of Adj. R^2. In model 1, first, third, fifth, seventh, eleventh principal components including time trend (T) have shown significant effect on maize crop yield. First principal component including time trend (T) have shown significant effect in model 2. In model 3, second and seventh principal components including time trend (T) have shown significant effect while none of the principal components have shown significant effects on maize crop yield in model 4. Only seventh principal component has shown significant effect in model 5.

The value of Adj. R^2 has been found to be maximum of about 79.78 percent in model 1 followed by about 73.30 percent in model 4. Using these forecast models the forecast values of maize crop yield for the years 2018, 2019, 2020 and 2021 were obtained and the results are presented in Table 2.

Table2: Actual and Forecasted yield of maize crop

Year	Actualyield(kg/ha)	Forecasted yield(kg/ha)

		Model 1	Model 2	Model 3	Model 4	Model 5
2018	1324.50	1438.12	1301.06	1258.02	1523.87	1102.98
2019	1338.24	1293.80	1256.37	1136.95	1420.00	1304.52
2020	2174.60	2088.03	2202.21	2312.02	2032.55	2278.03
2021	2688.20	2703.65	2749.38	2798.63	2576.31	2931.85
RMSE(Kg/ha)		236.17	576.34	476.01	264.20	398.25

The evaluation metrics in Tables 1 & 2 and Figs. 1 & 2 indicate that Model 1 outperforms the other models, with the lowest RMSE (236.17) and highest adjusted R-squared value (79.78%). Model 4 also shows promising results, with an adjusted R-squared value of 73.30%. In contrast, Model 2 performs poorly compared to the other models. These findings suggest that Model 1 and Model 4 are the most suitable options for forecasting maize crop yield during the kharif season in Udham Singh Nagar district of Uttarakhand. This is consistent with previous research by **Yadav et al. (2014)**, who developed models using principal component analysis on weather variables for wheat yield forecasting.

The values of Adj. R^2 for the models have not been found to be so high in comparison to the models developed by an application of discriminant function analysis **(Agrawal *et al.*, 2012)** but taking into account the RMSE of the models, model 1 has relatively performed well and can be recommended for the forecast of the maize crop yield in Udham Singh Nagar district of Uttarakhand.

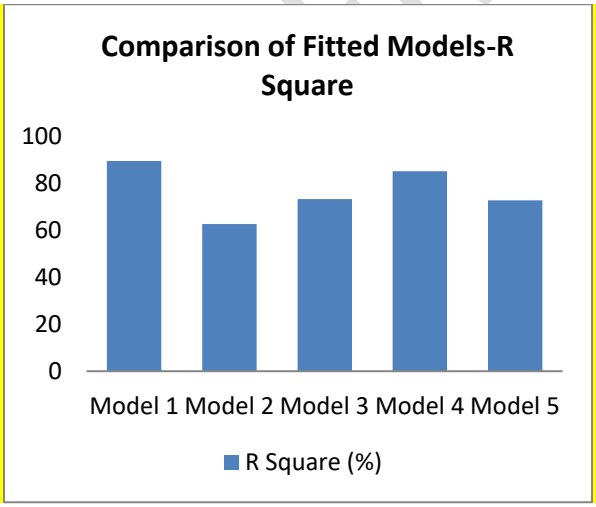

Fig.1: Comparison of developed models in terms of R^2

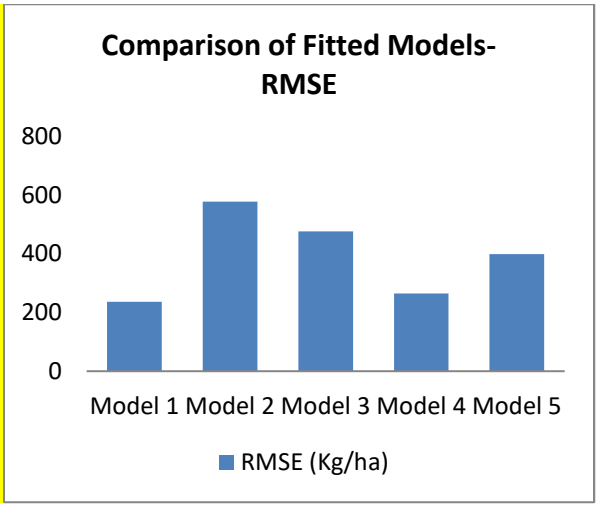

Fig.2: Comparison of developed models in terms of RMSE

Figs. 1 and 2 show the performance of the developed statistical models in terms of R^2 and RMSE.

Conclusion

The Application of Principal Component Analysis (PCA) on weather variables significantly enhance crop yield prediction accuracy since PCA converts the set of correlated variables into non-correlated components. Our study demonstrates that, the PCA model including all the weighted and unweighted indices performs better as compare to other models. Consequently, the model 1 which utilizes the principal components of all the weighted and unweighted indices as regressors can be used to forecast the maize yield during kharif season for Udham Singh Nagar district of Uttarakhand.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

References

1. 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.
2. Agrawal, R., Jain, R. C. and Jha, M. P., 1983, "Joint effects of weather variables on rice yields", Mausam, 34, 2, 189-194.
3. Agrawal, R., Jain, R. C. and Jha, M. P., 1986, "Models for studying rice crop weather relationship", Mausam, 37, 1, 67-70.
4. 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.
5. Agrawal, R., Chandrahas and Aditya, K., 2012, "Use of discriminant function analysis for forecasting crop yield", Mausam, 63, 3, 455-458.
6. Aneja, K. G. and Chandrahas, 1984, "Preharvest crop yield forecast based on plant biometrical characters", Silver jubilee souvenir, IASRI, New Delhi.
7. Angon, P. B., Tahjib-Ul-Arif, M., Samin, S. I., Habiba, U., Hossain, M. A., & Brestic, M. (2022). How do plants respond to combined drought and salinity stress?—a systematic review. *Plants*, *11*(21), 2884.

8. Angon, P. B., Anjum, N., Akter, M. M., KC, S., Suma, R. P., & Jannat, S. (2023). An overview of the impact of tillage and cropping systems on soil health in agricultural practices. *Advances in Agriculture*, *2023*(1), 8861216.
9. Aravind, 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
10. Bhatnagar, R., Kumar, P., & Singh, D. (2019). Wheat yield prediction using principal component analysis and random forest algorithm. Journal of Intelligent Information Systems, 55(1), 147-162. doi: 10.1007/s10844-018-0514-5
11. 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.
12. Draper, N. R. and Smith, H., 1998, "Applied Regression Analysis", 3rd edition, John Wiley & Sons Inc.
13. 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.
14. 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., 74.
15. Jain, R. C. and Singh, D., 1980, "Forecasting rainfall over Puerto Rico. Annual Report, Department of Meteorology", The Florida State University.
16. Jain, R. C., Jha, M. P. and Agrawal, Ranjana, 1985, "Use of growth indices in yield forecast", Biometrical Journal, 27, 4, 435-439.
17. Jiang, H., Li, Z., & Zhang, J. (2019). Maize yield prediction using principal component analysis and multi-temporal satellite data. Remote Sensing, 11(2), 151. doi: 10.3390/rs11020151
18. Johnson, R. A. and Wichern, D. W., 2001, "Applied Multivariate Statistical Analysis", 3rd edition, Prentice-Hall of India.
19. Kumar, P., Singh, D., & Kumar, A. (2017). Sugarcane yield prediction using principal component analysis and artificial neural networks. Journal of Sugar Research, 4(2), 1-12. doi: 10.5958/2454-2664.2017.00002.1
20. Li, M., Li, Z., & Huang, L. (2019). Maize yield prediction using principal component analysis and neural networks. Computers and Electronics in Agriculture, 157, 271-278. doi: 10.1016/j.compag.2019.02.017
21. Patil, P. D., Kumar, P., & Singh, D. (2018). Crop yield prediction using principal component analysis and support vector machines. Journal of Intelligent Information Systems, 51(2), 247-263. doi: 10.1007/s10844-017-0454-4
22. Rai, T. and Chandrahas, 2000, "Use of discriminant function of weather parameters for developing forecast model of rice crop", Publication of IASRI, New Delhi.
23. Singh, A., Kumar, P., & Singh, D. (2020). Comparative study of machine learning algorithms for maize yield prediction using principal component analysis. Journal of Intelligent Information Systems, 57(2), 257-273. doi: 10.1007/s10844-019-00574-4

24. Yadav, R. R., Sisodia, B. V. S. and Kumar, S. Application of principal component analysis in developing statistical models to forecast crop yield using weather variables. Mausam 2014; 65 (3): 357-360.
25. Yadav, S. K., Kumar, P., & Singh, D. (2014). Principal component analysis for wheat yield forecasting using weather variables. Journal of Intelligent Information Systems, 43(2), 257-271. doi: 10.1007/s10844-013-0264-4
26. Zhang, Y., Li, Z., & Chen, L. (2018). Rice yield prediction using principal component analysis and machine learning algorithms. Journal of Integrative Agriculture, 17(5), 931-941. doi: 10.1016/S2095-3119(18)62023-5