Journal of Agriculture, Biology and Applied Statistics
Vol. 3, No. 1, 2024, pp. 11-20

URL : www.arfjournals.com
https://DOI:10.47509/JABAS.2024.v03i01.02

Comparative Study Using Principal Component and Cluster Analysis for Crop Yield Classification

S.R. Krishna Priya[1], N. Naranammal[2] and S. Sujith[3]
[1,2,3] Department of Statistics, PSG College of Arts & Science, Coimbatore

To cite this article

S.R. Krishna Priya, N. Naranammal & S. Sujith (2023). Comparative Study Using Principal Component and Cluster Analysis for Crop Yield Classification. Journal of Agriculture, Biology and Applied Statistics. Vol. 3, No. 1, pp. 11-20. *https://DOI:10.47509/JABAS.2023.v03i01.02*

Abstract: This paper is an attempt to classify different crop yields of Coimbatore district using multivariate techniques. Fifteen different crop yields including food and cash crops from the year 1970 to 2021 have been used for the analysis. A comparative study using Principal Component analysis and Cluster analysis has been carried out for classifying different crop yields. Using both Principal Component analysis and Cluster analysis, four groups have been classified. The Principal Component analysis performed better than Cluster analysis in the classification of crop yields. The four groups obtained from classification are Pulses, Cereal grains, Seasonal cash crops and Annual cash crops.

Keywords: Multivariate Techniques, Crop Yield, Agriculture, Classification, Sustainability.

1. Introduction

Agriculture is the mainstay of life for about half the working population of Tamil Nadu. Tamil Nadu's total area of cropland was used for plantations, amounting over 658 thousand hectares in 2023. In Coimbatore district, agriculture plays a pivotal role in the economy, with crops ranging from staples like Rice, Pulses, and Millets to cash crops such as Cotton, Sugarcane, and Coconut. Paddy cultivation is prevalent in the district, particularly in areas with access to irrigation facilities and its production ranges from 3 to 5 tons per hectare. Sugarcane production ranges from 70 to 100 tons under favourable conditions. Coimbatore is often referred to as the "Manchester of South India" due to its prominent textile industry, of which Cotton plays a vital role and its production ranges from 2 to 4 tons per hectare.

Previously studies have been carried out using Principal Component analysis for classifying yield attributing traits in chilli (Pragya, *et al.* 2020), to understand the Azerbaijan vegetables and fruit sectors (Ibrahim. 2021), irrigating black gram influenced by liquid

Article History Received : 30 December 2023; Revised : 29 January 2024; Accepted : 11 February 2024; Published : 30 March 2024

organic bio stimulants (Ajaykumar, *et al.* 2023), to explore consumers' perception toward quinoa health (Tavagwisa, *et al.* 2020), classifying bean genotypes for agronomic, morphological, and biochemical characteristics (Girgel, 2021), application in agricultural equipment's (Constantin, *et al.* 2011).

Many researchers have used Cluster analysis in assessing the sustainability of organic farms (Maciej, *et al.* 2019), designing strategy of region's agro-food complex (Eugene, *et al.* 2017), commercialisation of farmers in developing rural areas (Moraka, *et al.* 2008), productivity of major crops across different agro-climatic zone (Halagunegowda, *et al.*2015), agriculture and other allied sectors (Sarojamma, *et al.* 2019). Some other researchers have used both Principal Component analysis and Cluster analysis for grouping bread wheat genotypes (Urgaya, *et al.* 2022), for ground water quality assessment (Mehmet, *et al.* 2021), agricultural productivity in crop commodities (Ibrahim and Gubad 2023) and characterization of maize fields (Daniel, *et al.* 2022).

2. Materials and Methods

2.1. Data Description

In this present study, the crop yield data has used for classification. The crops used for the study are Rice, Cholam, Cumbu, Ragi, Maize, Bengal Gram, Red Gram, Green Gram, Black Gram, Horse Gram, Sugarcane, Cotton, Tobacco. The data has been collected from various volumes of Season and Crop report. The crop yield data from year 1970 to 2021 has been used for the Principal Component and Cluster analysis.

2.2. Principal Component Analysis

Principal Component analysis is commonly used for dimensionality reduction in data analysis and machine learning. Its primary objective is identifying patterns and structure in high-dimensional data by transforming it into a lower-dimensional space while preserving as much of the original information as possible.

2.2.1. Covariance Matrix

The covariance matrix Σ of the standardized data is computed

$$\Sigma = (X - \bar{X})^T (X - \bar{X})$$

Where X is the standardized data matrix, $\bar{X}$ is the mean vector of the standardized data, and 'n' is the number of samples.

2.2.2. Eigenvalue Decomposition

After obtaining the covariance matrix, the next step is to compute its eigen vectors and eigen values. The eigenvalue is given by

$$\Sigma v = \lambda v$$

Where 'v' is the eigen vector, λ is the corresponding eigen value.

2.2.3. Selection of Principal Components

After computing the eigenvalues and eigen vectors, the Principal Components are selected based on the eigen values. Typically, the eigenvectors corresponding to the 'k' largest eigenvalues are chosen as the Principal Components.

2.2.4. Projection onto Principal Component

Finally, the original data is projected onto the subspace spanned by the selected Principal Components. If V represents the matrix whose columns are the selected eigenvectors (Principal Components), the projection of the standardized data X onto the Principal Components is given by:

$$\text{Projected Data } Y = XV$$

2.3. Cluster Analysis

Cluster analysis involves grouping similar objects or data points into clusters to reveal underlying patterns or structures. There are several methods and algorithms for Cluster analysis, each with its own equations or algorithms. Here are some common ones,

2.3.1. Hierarchical Clustering

There are different linkage methods in hierarchical clustering like single, complete, average, etc. The equations for distance computation vary based on the chosen linkage method. For example, in complete linkage, the distance between two clusters is the maximum distance between any single data point in the first cluster and any single data point in the second cluster.

Single Linkage: The distance between two clusters is defined as the minimum distance between any single data point in the first cluster and any single data point in the second cluster

Complete Linkage: The distance between two clusters is defined as the maximum distance between any single data point in the first cluster and any single data point in the second cluster.

Average Linkage: The distance between two clusters is defined as the average distance between all pairs of points in the two clusters.

3. Results and Discussion

3.1. Summary Statistics

Table 1, summary statistics shows the overall minimum and maximum crop yields and its mean & standard deviation.

Table 1: Summary Statistics of Crop Yield

Crops	*Min*	*Max*	*Mean*	*SD*
Sornavari	2131	5095	3512.77	743.694
Samba	1802	4567	3197.21	638.111
Navarai	2000	4838	3330.70	709.947
Cholam	284	1316	610.71	213.611
Cumbu	592	5670	1633.71	823.203
Ragi	821	3805	2126.13	657.566
Maize	560	8347	2616.85	2439.165
Bengal Gram	375	926	703.02	107.299
Red Gram	192	1273	659.95	249.128
Green Gram	161	745	412.47	153.050
Black Gram	225	887	478.24	164.722
Horse Gram	128	956	401.46	201.736
Sugarcane (tonnes)	87	139	106.33	11.465
Cotton	101	835	397.93	137.154
Tobacco	951	2399	1507.78	215.361

From the table 1, the rice Sornavari's minimum yield is 2131 hectares in 1980 and maximum yield is 5095 hectares in 1994. Samba's minimum yield is 1802 hectare in 2017 and maximum yield in 2014 is 4567 hectares. Navarai's minimum yield is 2000 hectares in 1973 and maximum yield is 4838 hectares in 2014. Cholam's minimum yield is 284 hectares and maximum yield is 1316 hectares in 2015. Cumbu's minimum yield is 592 hectares in 1972 and maximum yield is 5670 hectares in 2009. Ragi's minimum yield is 821 hectares in 1971 and maximum yield is 805 hectares in 2021. Maize's minimum yield is 560 hectares in 1975 and maximum yield is 8374 hectares in 2015. Bengal Gram's minimum yield is 375 hectares in 1975 and maximum yield is 926 hectares in 2017 to 2022. Red Gram's minimum yield is 192 hectares in 2002 and maximum yield is 1273 hectares in 2020. Green Gram's minimum yield is 161 hectares in 1988 and maximum yield is 745 hectares in 2014. Black Gram's minimum yield is 225 hectares in 1976 and maximum yield is 887 hectares in 2015. Horse Gram's minimum yield is 128 hectares in 1975 and maximum yield is 956 hectares in 2015. Sugarcane's minimum yield is 87 hectares in 2017 and maximum yield is 139 hectares in 2006. Cotton's minimum yield is 101 hectares in 2003 and maximum yield is 835 hectares in 1985. Tobacco's minimum yield is 951 hectares in 1973 and maximum yield is 2399 hectares in 1985. Some crop yields are minimum in the year of 1971 to 1975. Some crops yields are maximum in the year of 2014 to 2015.

3.2. Results of Principal Component Analysis

In the present study Principal Component analysis has been used for classification and the results are presented below.

3.2.1. Checking the Adequacy of Data

Kaiser-Meyer-Olkin (KMO) test has been used to measure the relevance of the data for Principal Component analysis. The test measures the sampling adequacy of each variable in the model. The value of KMO test ranges from 0 to 1. KMO test and Bartlett's sphericity test have been conducted for adequacy checking of the data and the results are presented in table 2.

Table 2: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.803
Bartlett's Test of Sphericity	Approx. Chi-Square	499.057
	df	105
	Sig.	.000

From this table 2, KMO measure is 0.803 which indicates that the value is acceptable and satisfactory for principal component analysis. From Bartlett's test of Sphericity, the associated probability is 0.000(<0.05) which makes correlation matrix as identity matrix. The results from the tests indicate that the data is fit for factor analysis.

3.2.3. Total Variance Explained

Total variance explained refers to the proportion of the total variance in the dataset that is accounted by the principal component retained in the analysis. The result of total variance explained in presented in table 3.

Table 3: Total Variance Explained

Component	*Rotation Sums of Squared Loadings*		
	Total	*% of Variance*	*Cumulative %*
1	4.623	30.818	30.818
2	3.839	25.592	56.410
3	1.406	9.374	65.785
4	1.327	8.849	74.634

From the table 3, it is clear that 74.634 % of the original data has been retained to the components using Principal Component analysis.

3.2.4. Scree Plot

Scree plot shows the classification of crop yield into components using graphical representation and it is presented in figure 1.

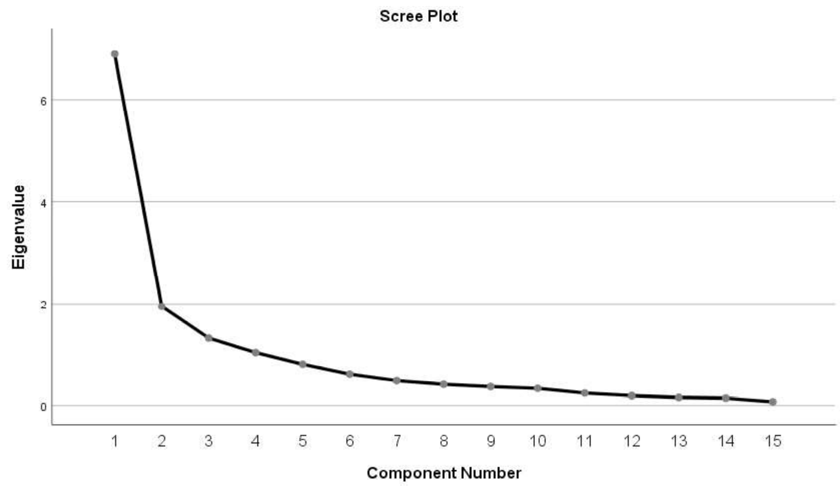

Figure 1: Scree Plot

From the figure 1, eigen values of first four components are greater than 1 which indicates that four components are obtained from 15 crops.

3.2.5. Component Plot in Rotated Space

The component plot in rotated space shows the components of crop yield classification in rotated space diagrammatically. The component plot is presented in figure 2.

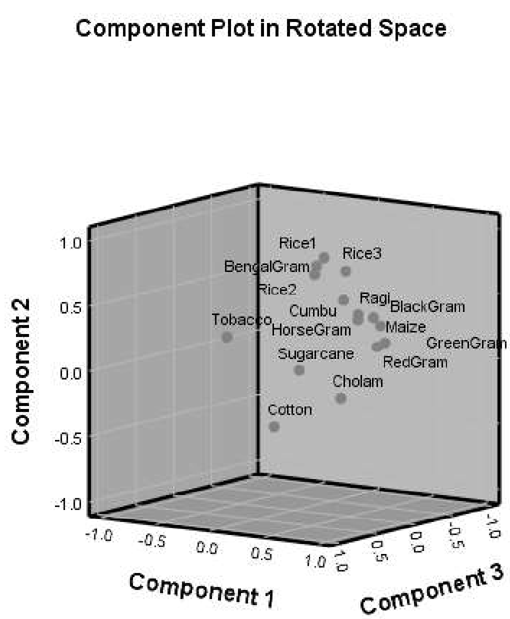

Figure 2: Component Plot in Rotated Space

Figure 2 shows the component plot in rotated space. It shows the visible two components, while the other component can be seen by rotating the figure.

3.2.6. Rotated Component Matrix

The table 4 shows the components of crop yields using Principal Component analysis. The crops belong to the components having the higher loadings.

Table 4: Rotated Component Matrix

Crops	*Rotated Component Matrix*			
	Component			
	1	*2*	*3*	*4*
Cholam	0.769	-0.106	0.488	0.002
Ragi	0.691	0.480	0.146	-0.105
Maize	0.807	0.386	0.020	0.120
Red Gram	0.825	0.236	0.091	0.102
Green Gram	0.756	0.232	-0.114	-0.023
Black Gram	0.699	0.429	-0.043	0.317
Horse Gram	0.722	0.447	0.194	-0.075
Sornavari	0.205	0.831	-0.098	-0.237
Samba	0.210	0.724	0.032	0.168
Navarai	0.494	0.776	0.028	-0.070
Cumbu	0.386	0.523	-0.097	0.243
Bengal Gram	0.254	0.799	0.069	0.073
Cotton	0.289	-0.351	0.680	-0.241
Tobacco	-0.083	0.321	0.771	0.319
Sugarcane	0.088	-0.025	0.062	0.934

The rotated component matrix shows that the crop yields are grouped into 4 components. Pulses are classified under 1st component, Cereal grains are classified under 2nd component, Seasonal cash crops such as Cotton and Tobacco are classified under 3rd component and Sugarcane is the annual cash crop and it classified under 4th component. Here the 1st and 2nd components are food crops while 3rd and 4th components are cash crops.

3.3. Results of Cluster Analysis

In the present study, Cluster analysis has been used to group the crop yield and the results are presented below.

3.3.1. Cluster Table

The table 5 shows the four clusters classified from fifteen crop yield using Cluster analysis.

Table 5: Cluster Table of Different Crop Yields

S. No.	*Crops*	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*
1	Sornavari	✓			
2	Samba	✓			
3	Navarai	✓			
4	Cholam		✓		
5	Bengal Gram		✓		
6	Red Gram		✓		
7	Green Gram		✓		
8	Black Gram		✓		
9	Horse Gram		✓		
10	Sugarcane		✓		
11	Cotton		✓		
12	Cumbu			✓	
13	Ragi			✓	
14	Tobacco			✓	
15	Maize				✓

From the table 5, three types of rice are clustered in 1st cluster. The crops Cholam, Bengal gram, Red gram, Green gram, Black gram, Horse gram, Sugarcane and Cotton are clustered in 2nd cluster. Cumbu, Ragi and Tobacco are clustered in 3rd cluster and Maize is clustered in 4th cluster.

3.3.2. Dendrogram of Crop Yield Classification

Figure 3 shows the classification of crop yield using Cluster analysis in dendrogram.

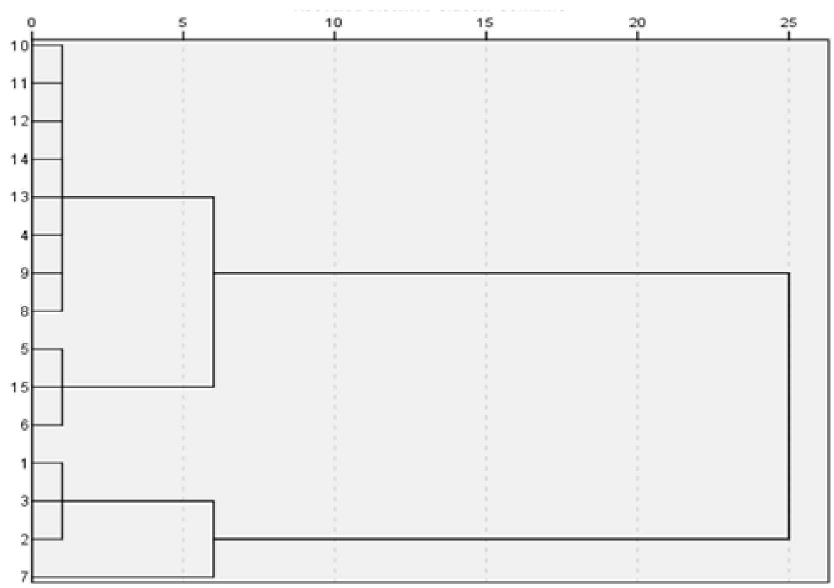

Figure 3: Dendrogram of Crop Yield Classification

From the figure 3, fifteen crops are clustered into four clusters. The crops are 1-Sornavari, 2-Samba, 3-Navarai, 4-Cholam, 5-Cumbu, 6-Ragi, 7-Maize, 8-Bengal Gram, 9-Red Gram, 10-Green Gram, 11-Balck Gram, 12-Horse Gram, 13-Sugarcane, 14-Cotton and 15-Tobacco.

The rice verities such as Sornavari, Samba and Navarai are grouped in cluster-1. Cholam, Bengal gram, Red gram, Green gram, Black gram, Horse gram, Cotton and Sugarcane are grouped in cluster-2. Cumbu, Ragi and Tobacco are grouped in cluster-3. Maize is the only crop to grouped in cluster-4.

4. Conclusion

The crop yield classification plays a crucial role in enhancing agricultural productivity, profitability, and sustainability by providing actionable insights and decision support to farmers, policymakers, researchers, and stakeholders across the agricultural value chain. In the present study, Principal Component analysis and Cluster analysis have been used for classifying the crop yields. The Principal Component analysis classifies the crop yields in four components like Pulses, Cereal grains, Seasonal cash crops and Annual cash crops. The Cluster analysis has clustered the crop yields in four clusters. The 1st and 2nd components are food crops then 3rd and 4th components are cash crops in classification of crop yields using Principal Component analysis. Principal Component analysis performed better in classification than the Cluster analysis. It classified food crops and cash crops in separate components.

References

Ajaykumar, R., Harishankar, K., Chandrasekaran, P., Navinkumar, C., Sekar, S., Sabarinathan, C. and Reddy, B.K.K.R. (2023). Principal Component Analysis (PCA) and Character Interrelationship of Irrigated Black gram [Vigna mungo (L.) Hepper] Influenced by Liquid Organic Bio stimulants in Western Zone of Tamil Nadu. *Legume Research.* 46(3): 346-362.

Balcha, U., Mekbib, F. and Dagnachew. (2022). Cluster and Principal Component Analysis among Bread Wheat (Triticum Aestivum L) Genotypes in Mid Rift Valley of Oromia, Ethiopia. *Advances in Crop Science and Technology.* 10(8):525.

Girgel, U. (2021). Principal Component Analysis (PCA) of Bean Genotypes (Phaseolus vulgaris L.) Concerning Agronomic, Morphological and Biochemical Characteristics. *Applied Ecology and Environmental Research.* 19(3): 1999-2011.

Halagundegowda, G.R., Surendra, H.S., Kumar H.M.P. and Nagaraja, M.S. (2015). Cluster Analysis on Productivity of Major Crops Across Different Agro-Climatic Zones of Karnataka, Progressive Research. 10:2479-2480.

Makhura, M.T., Goode, F.M. and Coetzee, G.K. (1998). A Cluster Analysis of Commercialisation of Farmers in Developing Rural Areas of South Africa. *Development Southern Africa.* 15(3):429-448.

Markos, D., Mammo, G. and Worku, W. (2022). Principal Component and Cluster Analyses based Characterization of Maize Fields in Souther Central Rift Valley of Ethiopia. *Open Agriculture.* 7:504-519.

Muziri, T., Chaibva, P., Chofamba, A., Madanzi, T., Mangeru, P., Mudada, N., Manhokwe, S., Mugari, A., Matsvange, D., Murewi, C.T.F., Mwadzingeni, L., and Mugandani, R. (2020). Using Principal Component Analysis to Explore Consumers' Perception Toward Quinoa Health and Nutritional Claims in Gweru, Zimbabwe. Food Science & Nutrition.

Niftiyev, I. (2021). Understanding Principal Component Analysis (PCA) in the Azerbaijan Economy: Case Studies of Vegetable and Fruit Sectors, Leibenz-Information Centre for Economics. SSRN, Rochester, NY.

Niftiyev, I. and Ibadoghlu, G. (2023). Longitudinal Principal Component and Cluster Analysis of Azerbaijan's Agriculture Productivity in Crop Commodities. *Commodities*. 2(2): 147-167.

Sarojamma, B., Geetha, K. and Reddy, B.H.M. (2019). Cluster Analysis for Agriculture and Other Allied Sectors. *Think India Journal.* 22(10):4186-4189.

Singh, P., Jain, P.K. and Tiwari, A. (2021). Principal Component Analysis for Yield Attributing Traits in Chilli (Capsicum annum) Genotypes. *Chemical Science Review and Letters*. 9(33): 87-91.

Sporysz, M., Szczuka, M., Tabora, S., Molenda, K. and Kubon, M. (2019). The Use of Cluster Analysis in Assessing the Sustainability of Organic Farms. *Agricultural Engineering.* 23(4): 69-76.

Stovba, E., Stovba, A., Abdrashitova, A. and Baygildina, A. (2017). Use of Methods of Cluster Analysis in Designing Strategy of Region's Agro-Food Complex. *Advances in Economics, Business and Management Research*. 38: 648-652.

Tarcolea, C., Paris, A.S. and Voicu, P. (2011). Principal Component Analysis Applied to Agriculture Equipments. *Journal of Agricultural Machinery Science*. 7(3): 305-308.

Tasan, M., Demir, Y. and Tasan, S. (2021). Groundwater Quality Assessment Using Principal Component Analysis and Hierarchical Cluster Analysis in Alacam, Turkey. *Water Supply.* 22(3): 3431.