Statistics and Applications{**ISSN 2454-7395**(online)}
Volume 14, Nos. 1 & 2, 2016 (New Series) pp.63-73

An Application of Multivariate Techniques for Divergence Study in Pearl Millet Germplasm

Urmil Verma[1], Ramesh Kumar[2] and M.S. Dalal[2]

[1]Department of Mathematics, Statistics and Physics,
[2]Bajra Section, Department of Genetics and Plant Breeding, CCS Haryana Agricultural University, Hisar

Received March 20, 2015; Revised December 24, 2015; Accepted: August 01, 2016

Abstract

Selection, which is the basis of every breeding programme operates only on variation which is of genetic nature and a wide range of variability present in any crop always provides the better chances of selecting the desirable types. The emphasis of this work was to study the genetic divergence among the pearl millet germplasm and grouping them into different clusters based on yield and yield attributing traits for the hybridization programme. PC analysis revealed that the first four PCs explained about 73% of the total variation giving an idea of the structure underlying the variables. The most informative variables associated with these PCs were grain yield, earhead weight, ear length, number of effective tillers and 1000-grain weight. Genetic divergence study was performed on the basis of Discriminant analysis using Mahalanobis' D^2-statistics. Based on the relative contributions of different characters; 99 genotypes of pearl millet were grouped into five clusters and plant height, dry fodder yield and grain yield were found the best discriminatory characters for the selection of diverse genotypes.

Key words: Genetic diversity, Principal Component, cluster mean, D^2-value, Inter-cluster distances

1. Introduction

Identification of diverse parents in any crop species is the pre-requisite. Generally, in breeding programmes for crop improvement, a researcher/plant breeder has to deal with several traits of economic importance which are usually inter-related with varying degrees and therefore single trait analysis can often be misleading and hence the correct approach would be multivariate analysis to explain the nature of relationship

Corresponding author: Urmil Verma
E-mail: vermas21@hotmail.com

among different traits. While working on a crop improvement programme, one of the steps to be taken into consideration is; what are the important characters to be considered for selection ? This can be answered by applying the suitable statistical procedures for selection of variables.

Pearl millet, the crop considered for this study is a staple diet for the vast majority of poor farmers and also forms an important fodder crop for livestock population in arid and semiarid regions of the country. Increased emphasis on development of dual purpose (grain cum fodder) pearl millet, is necessary for ensuring high grain yield as well as higher dry fodder yield under rainfed cultivation. The development of an effective breeding programme is dependent upon the existence of genetic variability. Pearl millet is endowed with a rich reservoir of genetic variability for various yield components, adaptation and quality traits. Exploitation of the genetic variability in the available germplasm holds promise for producing high grain and fodder yielding hybrids. The more diverse the parents, the greater are the chances of obtaining new combinations of genes and therefore increasing the probability for crop improvement. Several measures are being used to assess the genetic diversity among plant populations. Of these measures, multivariate analysis {Fisher (1936), Jolliffe (1972), Johnson and Wichern (2006) etc.} provides the most reliable information. Just to cite a few references in this regard ; Joshi *et al.* (1988), Wilson *et al.* (1990), Vidyadhar *et al.* (2004) and Shanmuganathan *et al.* (2006) etc. have worked on genetic divergence in pearl millet at national/international level. Among the multivariate procedures, Mahalanobis (1936) generalized distance (D^2) has been used extensively. Keeping in view the importance of the subject matter, an attempt has been made to perform the multivariate analyses in pearl millet germplasm pertaining to Haryana state for grain yield and its contributing characters.

2. Multivariate statistical techniques for selection of important characters

In a crop improvement programme, measurements are taken on several characters because of their inter-relationships, however, a breeder may be interested in selecting only few important characters in which the improvement is needed. A number of statistical procedures have been proposed from time to time for selection of important characters. There are two basic groups of multivariate techniques – Dependence Methods and Interdependence Methods.

2.1. Analysis of Dependence

If a multivariate technique attempts to explain or predict the dependent variable(s) on the basis of two or more independent variables then the **analysis of dependence is required.** The dependence methods viz., multiple regression analysis, multivariate analysis of variance and canonical correlation analysis have wide applications in this direction.

2.2. Analysis of Interdependence

The goal of interdependence method is to give meaning to a set of variables or to seek to group things together. No one variable or variable subset is to be predicted from

the others or explained by them. The most common of these methods are principal component analysis (PCA), factor analysis, cluster analysis, discriminant analysis and multidimensional scaling etc. PC analysis reduces the data with large number of correlated variables into a substantially smaller set of new variables, through linear combination of the variables that accounts for most of the variation present in the original variables. The linear combinations so selected are called the principal components. Correlation or co-variance matrices are generally used to estimate principal components. When the variables are measured in different units, scale effects can influence the composition of derived components. In such situations, it becomes desirable to standardize the variables. Therefore, correlation matrix is considered to be better as it does not require standardization.

Let $X_1, X_2,\ldots\ldots X_p$ be the elements of a p-component random vector **X**. Assuming that the mean vector of **X** is **0** and variance-covariance matrix $\sum$ which is a real positive definite matrix. Supposing the non-zero eigen values of $\sum$ are $\lambda_1 > \lambda_2 > \ldots\ldots > \lambda_p$ with the corresponding eigenvectors $\boldsymbol{\gamma_1, \gamma_2, \ldots\ldots, \gamma_p}$. For distinct λ_i's (i = 1, 2, ……p); a (p × p) orthogonal matrix **Γ** can be formed as

$$\boldsymbol{\Gamma} = [\gamma_1, \gamma_2, \ldots . \gamma_p]$$

The **Γ** matrix diagonalizes **Σ** matrix such that

$$\Sigma = \Gamma \Lambda \Gamma',$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) = \Gamma' \Sigma \Gamma$
Next, considering an orthogonal transformation of **X** vector to **Y** vector as

$$Y = \Gamma' X$$

where $\mathbf{Y_1, Y_2, \ldots, Y_p}$ are the p components of **Y** and are called principal components.

The percentage of variation of **X** explained by *i*-th principal component is

$$\frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100$$

Interpretation of principal components is often facilitated by computing the components loadings. PC loadings measure the importance of each variable in accounting for the variability in PC. Thus, PCA may often indicate which variables in a data set are important and which ones may be of little consequence. Some of these low-performance variables might therefore be removed from consideration in order to simplify the overall analysis.

Cluster analysis is also one of the methods of data reduction technique. PCA reduces the number of variables whereas cluster analysis reduces the number of observations. The cluster analysis identifies homogeneous groups or clusters and no assumptions are made concerning the number of groups. There is no unified approach on which actually constitutes a cluster. One of the important aspects of clustering is to study the differences among the formed clusters. One way MANOVA may be performed to check the accuracy of the clustering and to decide about the appropriate number of clusters. As per definition, a cluster constitutes with similar objects and that is decided on the basis of a measure of inter-object similarity. There are two main types of measures used to estimate this relation; distance measures and similarty measures. Hierarchical Cluster analysis is a commonly used method for forming clusters and displaying similarities and dissimilarities between pairs of objects of a set by using one of the two methods; agglomerative or divisive. In agglomerative analysis, clusters are formed by grouping cases into bigger and bigger clusters until all cases are members of a single cluster. Thus, based on the chosen similarity measure, the basic criterion is that the objects in a cluster should be close to each other than to objects in other clusters.

Cluster analysis has similarity with discriminant analysis in respect of classification of observations but the discriminant analysis derives a rule for allocating an object to its known proper population based on some prior information of the group membership of the objects. This technique is used to discriminate between/among various groups of objects when the dependent variable is categorical and independent variables are metric. It is a multivariate technique concerned with separating distinct sets of objects and allocating new objects to the previously defined groups. The steps of discriminant analysis are: i) To develop an equation or a function using variables under consideration for computing a new variable or index that will parsimoniously represent the differences between/among various groups ii) Use of discriminant function to classify the observations into any of the pre-defined groups. Three different methods namely i) Maximum Likelihood Discriminant Rule, ii) Fisher's Linear Discriminant Function and iii) Bayes Discriminant Rule are in common use. However for the present study, the Fisher's linear discriminant function method was used for the purpose.

3. Experimental details and Statistical analysis

The experimental material comprising of 99 entries of Pearl millet germplasm was grown in Augmented Design during *kharif* season, 2014 at the research farm of Bajra section, Department of Genetics and Plant Breeding, CCS HAU, Hisar. Each genotype was sown in a plot consisting of two rows of size 3.6m x 0.45m each with plant to plant distance as 12 cm. The quantitative data for yield and yield contributing characters viz., days to 50% flowering, plant height (cm), ear length (cm), ear diameter (cm), effective number of tillers per plant, earhead weight (g/plant), grain yield (g/plant), dry fodder yield (g/plant) and 1000-grain weight (g) were recorded.

The genetic variability parameters including mean, standard deviation and coefficient of variation (CV%) of various traits were obtained (Table 1). Regression analysis across all the genotypes was carried out and earhead weight was found the best predictor towards the grain yield (R^2 = 0.81). PC analysis for investigating the inter-dependence was attempted to simplify the complex and diverse relationships existing among a set of observed variables, by revealing common dimensions that link seemingly unrelated variables. The procedure consists of finding the eigen roots and eigen vectors of the correlation matrix of explanatory variables. One of the most commonly used criteria eigenvalue-one (Kaiser, 1960) was used to observe the percent variance explained by different PC components as shown in Table 2. With the Scree test (Cattell, 1966), we have also plotted the eigenvalues associated with each component and looked for a break between the components with relatively larger eigenvalues and those with smaller eigenvalues as shown below in Scree plot. The components that appeared before the break were assumed to be more meaningful and those appearing after the break were not retained.

Table 1. Descriptive statistics of yield and yield attributing traits

Variables	**Min**	**Max**	**Mean**	**Std. Deviation**	**CV(%)**
Days to 50% flowering	44.0	80.0	60.02	7.16	11.93
Plant height (cm)	107.0	218.0	160.67	26.13	16.26
No.of effective tillers/plant	2.0	6.4	3.04	0.63	20.69
Ear length (cm)	11.6	28.4	19.89	3.02	15.22
Ear diameter (cm)	10.2	35.8	21.23	4.42	20.81
Earhead weight (g/ plant)	10.0	52.0	27.38	8.96	32.72
Grain yield (g/plant)	4.0	35.0	14.53	6.80	46.81
Dry fodder yield (g/plant)	13.0	112.0	51.06	19.66	38.50
1000- grain wt (g)	3.6	13.6	8.23	2.08	25.31

Table 2. Percent variance explained by different PC components

Component (s)	Eigen value	% Variance	% Cumulative Variance
1	2.53	28.13	28.13
2	1.57	17.47	45.61
3	1.39	15.44	61.06
4	1.04	11.63	72.69
5	0.85	9.54	82.24
6	0.63	7.06	89.31
7	0.49	5.51	94.83
8	0.37	4.12	98.96
9	0.09	1.03	100.00

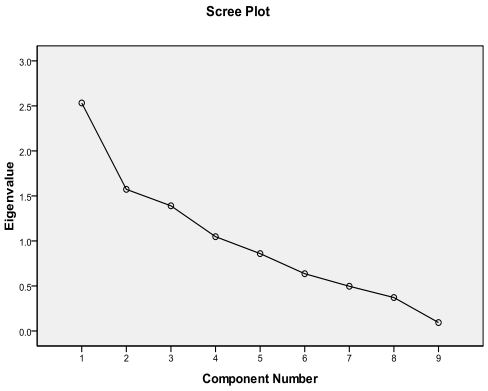

Table 3. Principal Component matrix showing higher loading displaying characters

Characters	Component(s)			
	1	2	3	4
Grain yield (g/ plant)	**.930**	.078	.079	-.103
Earhead weight (g/plant)	**.923**	.128	.087	-.114
Dry fodder yield (g/plant)	.785	.014	-.060	.287
Ear length (cm)	.057	**.822**	-.046	.209
Days to 50% flowering	.156	.614	-.145	-.309
Plant height (cm)	-.013	.610	.593	.127
No.of effective tillers/plant	.172	-.131	**.812**	.054
Ear diameter (cm)	.152	.038	-.636	.530
1000- grain wt (g)	-.013	.037	.015	**.814**

PC analysis revealed that the first four PCs explained about 73% of the total variation giving an idea of the structure underlying the variables. The higher loading displaying variables associated with these PCs in Table 3 were found as grain yield, earhead weight, ear length, number of effective tillers and 1000-grain weight.

The descriptive statistics given in Table 1 indicated that a considerable diversity exists in the material used. So, the data recorded on all the traits were subjected to cluster analysis and 99 genotypes were grouped into five clusters where each genotype within a cluster was closest to the cluster mean. The Ward's minimum variance method (1963) was used to carry out the agglomerative hierarchical cluster analysis. Thus, the cluster

analysis helped in grouping the genotypes in such a manner that similar types are grouped together while dissimilar ones belong to different groups (Table 7). The mean performance of different clusters calculated for different traits revealed wide range of differences among the clusters.

Further, the discriminant analysis was carried out for the selection of discriminator variables leading to the development of discriminant functions which were then used for classifying the observations. First three canonical discriminant functions were used for the purpose as is mentioned in Tables 4 & 5. Tests for differences between the groups, considering all the variables simultaneously were dealt using Wilk's test statistic given in Table 6. To determine the inter-cluster distances, the data were analysed on the basis of D^2-statistics to measure the genetic divergence among the genotypes and their average inter-cluster distances are shown in Table 8. However, the final cluster means in respect of all characters are given in Table 7. As there are genotypes superior for individual trait belonging to different clusters which indicates that none of the clusters contained genotypes with all the desirable characters. Thus, the genotypes superior for specific characters from different clusters may be selected for further utilization in breeding programme. Based on the relative contributions of different characters; plant height, dry fodder yield and grain yield were found the best discriminatory characters and 99 % of the originally grouped cases were correctly classified as has been depicted in Table 9. Inter and intra-cluster distances from the group centroid are shown in Figure1 while the clustering pattern with name and number of genotypes in each cluster is expressed in Table 10.

Table 4. Percent variance explained by the discriminant functions

Function(s)	Eigen value	% Variance	% Cumulative variance	Canonical correlation
1	4.60	69.0	69.0	0.91
2	1.93	29.0	98.0	0.81
3	0.13	2.0	100.0	0.35

First 3 canonical discriminant functions were used in the analysis

Table 5. Canonical discriminant function coefficients

Characters	Function(s)		
	1	2	3
Plant height (cm)	.080	-.028	.002
Dry fodder yield (g/plant)	.045	.069	-.045
Grain yield (g/plant)	-.004	.046	.187
(Constant)	-15.121	.309	-.681

Table 6. Variables Entered/Removed[a,b,c]

Step	Entered	Wilks' Lambda				Sig.
		Statistic	df1	df2	df3	
1	Plant height (cm)	0.20	1	4	94	.000
2	Dry fodder yield (g/plant)	0.06	2	4	94	.000
3	Grain yield (g/plant)	0.05	3	4	94	.000

At each step, the variable that maximized the Mahalanobis distance between the two closest groups was entered

a. Maximum number of steps was 18

b. Minimum partial F to enter was 3.84

c. Maximum partial F to remove was 2.71

Table 7. Final cluster means

Characters	Cluster(s)				
	1	2	3	4	5
Days to 50% flowering	58.89	60.26	62.79	58.46	61.67
Plant height (cm)	169.98	191.04	151.68	126.58	167.33
No.of effective tillers/plant	3.05	3.27	2.95	2.83	3.33
Ear length (cm)	19.52	21.12	20.14	18.58	21.37
Ear diameter (cm)	20.61	19.80	21.39	22.28	24.90
Earhead weight (g/ plant)	22.67	28.33	34.53	23.50	38.00
Grain yield (g/plant)	10.88	14.69	20.59	11.48	23.37
Dry fodder yield (g/plant)	36.02	54.99	61.27	43.28	102.50
1000- grain wt (g)	8.12	8.22	7.92	8.31	9.47

Table 8. Distances between final cluster means

Cluster	2	3	4	5
1	29.24	34.97	44.06	**69.63**
2		40.89	**65.90**	**54.93**
3			34.36	44.52
4				**74.47**
5				

Table 9. Classification results[a]

		Cluster Number	Predicted group membership					Total
			1	2	3	4	5	
Original	Count	1	27	0	0	0	0	27
		2	0	23	0	0	0	23
		3	1	0	18	0	0	19
		4	0	0	0	24	0	24
		5	0	0	0	0	6	6
	%	1	100.0	.0	.0	.0	.0	100.0
		2	.0	100.0	.0	.0	.0	100.0
		3	5.3	.0	94.7	.0	.0	100.0
		4	.0	.0	.0	100.0	.0	100.0
		5	.0	.0	.0	.0	100.0	100.0

a. 99 % of original grouped cases correctly classified

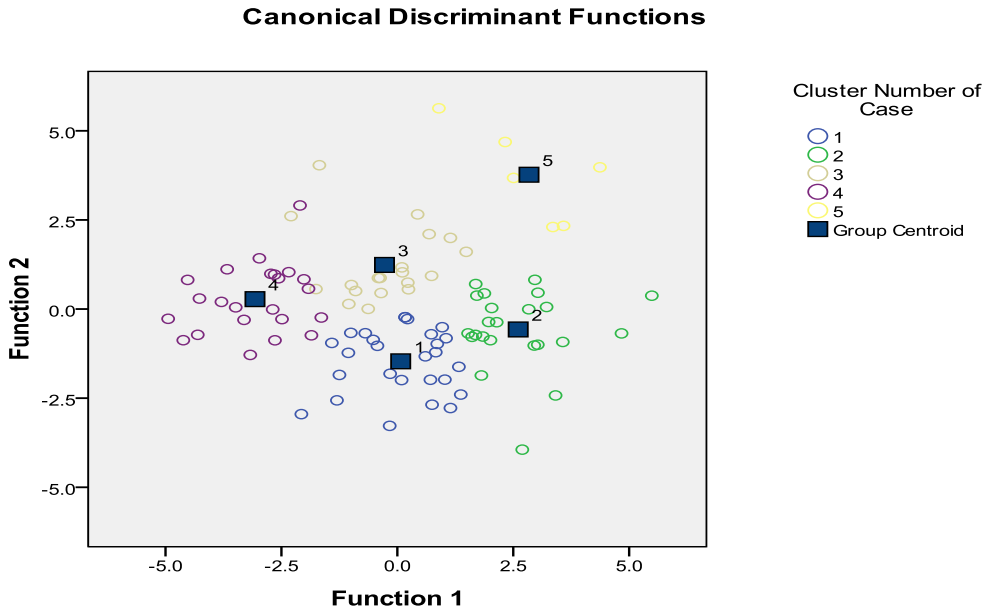

Figure 1. Inter and intra-cluster distances

Table 10. Clustering pattern and number of genotypes in each cluster of 99 genotypes of pearl millet

Cluster 1 (27)	Cluster 2 (23)	Cluster 3 (19)	Cluster 4 (24)	Cluster 5 (6)
H77/833-2-202	110250	PT-1-10-1047	RAJ 3	PT-1-10-1002
HTP 94/54	HPT-2-12-7	TCH 26-1	G73-107	HPT-10-144
SGP-10-120	HPT-2-12-10	(96111B x 4025-3-2-B)-1-6-1	H 1305	110041
HTP 93/37 [HTBC-208-1-B-2-1-1x B- line Bulk]-28	IP 4846 IP 7846	EMRT-11-109 (4B x 6B/1)	SGP-10-111 TPT-A2-1-11-155	IP 17846 CPBL-11-113
HBL-0538	IP 11522	EMRT-11-104	A5 R-10-119	99 HS-23
HMP 802	IP 18132	HBL 927	AR5-07114	
HPT-1-12-35	IP 21202	HBL 1110	HBL 11	
HPT-1-12-90	IP 22278	CPBL-11-107	TPT-A2-1-11-109	
HPT-1-12-84	WG 36-7	CPBL-11-112	HPT-1-12-120	
HPT-2-12-6	HBL-0535	CPBL-11-118	HPT-1-12-189	
HPT-2-12-15	HTP 93/54	TPBL-11-118	HPT-2-12-11	
HPT-2-12-48	MIR-97238	DPHBL-11-118	HPT-2-12-54	
HPT-2-12-69	HTP 93/227	HBL-0529	HPT-2-12-59	
IP 4839	JBV 3 S1-133-1-2-2-1-B	99 ABL-9	HPT-2-12-61	
IP 22271	Stay Green PT-216 0p	ICMP451 S1-10-1	HPT-2-12-62	
HTP 91/42	HTP 07-45	LPR-10-114	HTP 94/53	
203/1-2-1	HTP 0816	LPR-10-123	HBL34	
HTP09/60 (S1-ICMV 94474)	HTP09/113	G 73-107	HBL 1103	
HTP 0819	BBNH-601-7-1		CPBL-11-114	
R-112	1219/2		ISK 48	
MRC HS-178-1-3-1-3-B-B-B-B	MRC HS-109-1-1-1-B-2-B-1-B		HMS16B	
HFI-10-144			(ICMB 01888 x ICMB 01222)-15-2-B-2-3)	
HFI-10-140	HFIT-1-129		AC 04/29 -05K-2 M.T.	
Sweat 09-8				
HMP 808 (ICMV 98107 -07-1)				
HTP 91/42				

The statistical exercise showed that the sufficient variability exists in the material under study and cluster/discriminant analysis clearly helped in differentiating genotypes into major groups for various traits and to be used further for breeding purpose. Summarizing the overall results, the maximum inter-cluster distance was observed between clusters 4 and 5 (74.47) followed by clusters 1 and 5 (69.63), clusters 2 and 4 (65.90), clusters 2 and 5 (54.93) etc. while the lowest inter-cluster distance was observed between clusters 1 and 2 (29.24) followed by clusters 3 and 4 (34.36), clusters 1 and 3 (34.97), clusters 2 and 3 (40.89) etc. The genotypes from the clusters showing higher inter-cluster difference could be utilized in the hybridization programme as crossing between diverse parents is likely to produce wide genetic variability among the progenies of the segregating generations.

Acknowledgement

The reviewers' remarks are highly appreciated and thankfully acknowledged for the improvement of the research documentation.

References

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problem. *Ann. Eugen*, London, **7**, 179-88.

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

Jolliffe, I.T. (1972). Discarding variables in a principal component analysis. I: Artificial data, *Applied Statistics*, **21**, 160-73.

Joshi, R.P., Chauhan, G.S. and Yadav, H.S. (1988). Genetic divergence in Pearl millet. *Experimental Genetics*, **4**, 16–28.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis, Educational and Psychological Measurement, **20**, 141-51.

Mahalanobis, P.C. (1936). On the generalised distance in statistics. *Journal of Genetics*, **41**, 159-93.

Shanmuganathan, M., Gopalan, A. and Mohanraj, K. (2006). Genetic variability and multivariate analysis in pearl millet germplasm for dual purpose. *The Journal of Agricultural Sciences*, **2(1)**, 73-80.

Vidyadhar, B., Chand, P., Devi, S.L.I., Reddy, M.V.S. and Ramachandraiah, D. (2004).Variability and correlation analysis for the yield attributes in pearl millet germplasm. In: 3rd National seminar on millet research and development – Future policy options in India. (Eds). Y.K.Sharma and I.S. Khairwal, Jodhupur, India, pp. 12.

Ward, J.H., Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, **58**, 236-44.

Wilson, J.P., Burton, G.W., Zongo, J.D. and Dicko, I.O. (1990). Diversity among pearl millet land races collected in Central Burkina Faso. *Crop Science*, **30**, 40-43.