Available online at www.ijpab.com

DOI: http://dx.doi.org/10.18782/2320-7051.6596

ISSN: 2320 – 7051

Int. J. Pure App. Biosci. **6 (3):** 181-190 (2018)

Research Article

International Journal of Pure & Applied Bioscience

OPEN ACCESS

Genetic Diversity by Multivariate Analysis Using R Software

Immad Ahmad Shah[1], Imran Khan[2*], Shakeel A Mir[3], M. S. Pukhta[4], Zahoor A Dar[5], Ajaz Lone[6]

[1,2,3,4]Division of Agricultural Statistics, SKUAST-K, Shalimar, 190025, J&K, India

[5,6] Dry Land Agricultural Research Station, SKUAST-K, Budgam

*Corresponding Author E-mail: immad11w@gmail.com

Received: 15.05.2018 | Revised: 22.06.2018 | Accepted: 28.06.2018

ABSTRACT

The present investigation was conducted to study the genetic divergence pattern using Multivariate analysis techniques viz. Cluster Analysis (CA) and Principal Component Analysis (PCA). Cluster analysis identified and classified the accessions on the basis of the similarity of the characteristics into seven distinct clusters. The highest inter cluster distance was observed between Cluster I and Cluster III and lowest between Cluster II and Cluster V. The Principal Component Analysis revealed two principal components, PC I and PC II, and accounted for nearly 76.92% of the total variation. R software has been used to execute the above mentioned techniques of analysis.

Key words: *Principal Component Analysis, Eigen Values/Vectors, Cluster Analysis, R Software, SAS Software.*

INTRODUCTION

Genetic diversity is the total number of genetic characteristics in the genetic makeup of a species. Genetic diversity serves as a way for populations to adapt to changing environments. In context of the estimation of genetic diversity among the accessions Multivariate techniques have been proved to be important. Several tools are now in hand for studying the variability and relationships between accessions. The genetic divergence analysis estimates the extent of diversity existed among selected accessions[9].

Cluster analysis identifies and classifies objects individuals or variables on the basis of the similarity of the characteristics they possess, so the degree of association will be strong between members of the same cluster and weak between members of different clusters. It seeks to minimize within-group variance and maximize between-group variance. The cluster analysis aims to allocate a set of individuals to a set of mutually exclusive, exhaustive groups such that the individuals within a particular group are similar to one another while the individuals in the different groups are dissimilar. It is also helpful for parental selection in the breeding program and crop modelling. In order to reduce the volume of data and identify a few key or minimum descriptors that effectively account for the majority of the diversity observed, saving time and effort for future characterization efforts the data was subjected to principal component analysis.

Cite this article: Shah, I.A., Khan, I., Mir, S. A., Pukhta, M.S., Dar, Z.A., Lone, A., Genetic Diversity by Multivariate Analysis Using R Software, *Int. J. Pure App. Biosci.* **6(3):** 181-190 (2018). doi: http://dx.doi.org/10.18782/2320-7051.6596

Principal component analysis makes it possible to transform a given set of characteristics (variables), which are mutually correlated, into a new system of characteristics, known as principal components, which are not correlated. The obtained variables may also be used for further analysis, where the assumption of no co-linearity is required. Moreover, the analysis is characterized by the fact that it includes the total variance of variables, explains maximum of variance within a data set, and is a function of primary variables. The Principal Component Analysis shows which of the traits are decisive in genotype differentiation[7]. PCA enables easier understanding of impacts and connections among different traits by finding and explaining them.

R is a language and an environment for statistical computing and graphics for Multivariate data. It is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large and coherent, integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

MATERIAL AND METHODS

The multivariate analysis techniques viz. CA and PCA were used to study the genetic diversity of 55 accessions of Maize (Table 1) maintained by Division of Genetics and Plant Breeding, SKUAST-K, Shalimar. CA was used to group the accessions into various clusters. Hierarchical clustering approach was used using Mahalanobis[8] generalized distance (D^2). The clustering was performed using Euclidean distance and Average Linkage method which best reflects the differences among the accessions[6]. PCA was carried out for trait variability analysis and the Eigen Values and Eigen Vectors were computed which represent the variance and the loadings of the corresponding principal components. The variables were Plant height, Ear Height, Days to 50% Tasseling, Days to 50% Silking, 75% HB, Cob Length, Cob per Plant, Rows per Cob, Grains per Row, Cob Diameter, 100 Seed Weight, Yield per Plant. To perform data analysis R statistical package (Version 3.4.4) downloaded from http://cran.r-project.org was used.

RESULTS AND DISCUSSION

The data on variability parameters are presented in Table 2. The lowest values of standard deviation were recorded in the case of the Cob/Plant (0.171) followed by the Cob Diameter (0.603). The coefficients of variation were the lowest for Days to 50% Silking (0.027) followed by 75% Husk Browning (0.029). However, the highest coefficient of variation value was recorded for Yield/Plant (0.382) followed by Ear Height (0.217). A wide range of diversity was observed in the experimental material for the majority of the characters studied including morphological traits.

Cluster Analysis: Based on Average Linkage Clustering and the distance measure Euclidean the 55 maize accessions were grouped into seven distinct clusters (Figure 1 & Table 3). Cluster IV consisted of a maximum of 14 accessions (25.45%), Cluster II consisted of 12 accessions (21.81%), Cluster I consisted 10 accessions (18.18%), Cluster V consisted 9 accessions (16.36%) and Cluster III, VI and VII consisted 6, 3, 1 accessions accounting to 10.9, 5.45 and 1.81% respectively. Cluster VI had the highest intra cluster distance (29.65) indicating the high divergence among the accessions of the cluster (Table 4).

While considering the inter cluster distance minimum distance (34.37) was noticed between Cluster II and V (Table 4). Maximum inter cluster distance (184.71) was noticed between Cluster I and III followed by distance 178.96 between (I and V) and distance 177.88 between Cluster I and VI

(Table 4). According to Ghaderi *et al*.[2], increasing parental distance implies a great number of contrasting alleles at the desired loci, and to the extent that these loci recombine in the F2 and F3 generation following a cross of distantly related parents, the greater will be the opportunities for the effective selection for yield factors. Thus, crossing of accessions from these clusters with other clusters may produce higher amount of heterotic expression in the first filial generations (F1's) and wide range of variability in subsequent segregating (F2) populations.

Based on cluster means, Cluster III was important for Cob Length, Cob per Plant, Grain per Row, Cob Diameter and Yield per Plant, Cluster VI for Plant Height & Ear Height and Cluster VII for 50% Tasseling, 50% Silking, Rows per Cob as shown in Table 5. The maximum distances existed between Cluster I and III, I and V and I and VI. From the results it was concluded that accessions in Cluster III are important for Cob Length, Cob per Plant, Grain per Row, Cob Diameter and Yield per Plant, accessions in Cluster VII for 50% Tasseling, 50% Silking, Rows per Cob and accessions in Cluster VI for Plant Height & Ear Height, which could be selected as parents for hybridization programme. Crosses involving parents belonging to more divergent clusters would be expected to manifest maximum heterosis and wide variability in genetic architecture[10]. In the present study, Cluster VI was more divergent than the others. However, the chance of getting segregates with high yield level is quite limited when one of the characters has a very low yield level. The selection of parents should also consider the special advantage of each cluster and each genotype within a cluster depending on specific objective of hybridization[1]. Thus, crosses involving Cluster I and III with any other cluster except Cluster IV and VII are suggested to exhibit high heterosis and could result in segregates with higher Maize yield.

Correlation: The correlation matrix generated for the 12 characters revealed that several characters were found to be highly correlated such as Grain/row and Yield (0.925), 50% Tasseling and 50% Silking (0.938), Cob Length and Grain per Row (0.942), and 100 Seed Weight and Yield (0.895) as shown in Table 6.

Principal Component Analysis: Principal Component Analysis (PCA) reflects the importance of the largest contributor to the total variation at each axis of differentiation[11]. The Eigen vectors represent the coefficients of the principal components and the Eigen values represent the variances represented by the Principal Component. Eigen values are often used to determine how many Principal Components to retain. Usually Components with Eigen values less than 1 are excluded[5,4]. A scree plot as shown in Figure 2 also indicates the appropriate number of the Principal Components to be retained. The sum of the Eigen values is usually equal to the total number of variables under study or is equal to the trace of the correlation matrix. The first Principal component of the observations is the one whose variance is greatest among all the other Principal Components. Highest eigenvalue was obtained for principal component 1 with an eigenvalue of 6.92 followed by principal component 2 with an eigenvalue of 2.30 indicating that the variance of PC 1 is the largest of all, thus explaining maximum variability in the data (Table 7). Since the Eigen values of these components were greater than 1 hence they were retained (Fig. 2)

The loadings defining the five principal components of these data are given in Table 8. The coefficients are scaled, so that they present correlations between observed variables and derived components. Two Principal Components, PC I & PC II, which were extracted from the original data and having latent roots greater than one, accounting for nearly 76.92% of the total variation. Suggesting these Principal Component scores might be used to summarize the original 12 variables in any further analysis of the data. Out of the total

principal components retained, PC I, PC II with values of 57.69% and 19.23% respectively contributed more to the total variation. According to Chahal and Gosal[1] characters with largest absolute value closer to unity within the first principal component influence the clustering more than those with lower absolute value closer to zero. Therefore, in the present study, differentiation of the accessions into different clusters was because of relatively high contribution of few characters rather than small contribution from each character. Different characters have different loadings on the Principal Components. Most of the characters load on PC1 and only three of the characters were found to load on PC2. Accordingly, the first principal component had high positive component loading from Yield/Plant, Grain/Row, Cob Diameter, Seed Weight & Ear Height and high negative loading from Days to 50% Husk Browning and Days to 50% Tasseling while remaining traits in this PC1 did not contribute rather their effects were distributed in PC2 (Table 8).

The positive and negative loading showed the presence of positive and negative correlation trends between the components and the variables. This is evident from Figure 3 where the correlations of the characters for the Principal Components is depicted based on the relative loadings. Therefore, the above mentioned characters which load high positively or negatively contributed more to the diversity and they were the ones that most differentiated the clusters. The accessions in the PC1 were more likely to be associated with Plant Height, Ear Height, Rows per Cob, Cob per Plant, Grain per Row, 100 Seed Weight, Cob Length, Cob Diameter and yield/plant whereas the accessions with 50% Tasseling, 50% Silking, and 75% Husk Browning were contributing PC2 (Figure 2). In the component pattern diagram (Figures 2 and 3) correlation between variables and the Principal Component depicted and it has been found that Plant height, Ear height, Rows per Cob, Cob per Plant, Grain per Row, 100 Seed Weight, Cob Length, Cob Diameter and Yield per Plant are strongly correlated, with respect to PC1 and 50% Tasseling, 50% Silking, 75% Husk Browning strongly positively correlated with respect to PC2. Similarly 50% Tasseling, 50% Silking were found to be strongly negatively correlated with respect to PC1 and Ear Height, Yield per Plant, Plant Height, Cob diameter, Cob per Plant were found to be strongly negatively correlated with respect to PC2. The component scores and the component pattern of the fifty five Maize accessions and characters under study along the first two Principal Component axis respectively are shown in Figure 4 in the component score plot.

The coordination of the accessions on the entire axis all the axis together (Figure 4) revealed that accessions 3, 13, 14, 31, 37, 52, 18, 14, 18, 21 were found to be most distinct accessions for the characters studied. Usually it is customary to choose one variable from these identified groups. Hence, for the first group Yield/plant is best choice, which had the largest loading from component one and 50% Silking for the second component. The characters which contributed positively to first three Principal Components could be given due consideration while selecting the best accessions without losing yield potential. The present investigation provided considerable information useful in genetic improvement of Maize. Accessions grouped into Cluster I and III showed maximum inter cluster diversity. From cluster mean values, accessions in Cluster III, VI and VII deserve consideration for their direct use as parents in hybridization programs to develop high yielding Maize varieties. There is significant genetic variability among tested accessions which indicates the presence of excellent opportunity to bring about improvement through wide hybridization by crossing accessions in different clusters.

Table 1: List of the Maize Accessions

Accessions	Code	Accessions	Code
KDM-381B	1	KDM-916A	29
KDM-930A	2	KDM-382A	30
KDM-899A	3	KDM-926B	31
KDM-381A	4	KDM-909A	32
KDM-913A	5	KDM-362A	33
KDM-340A	6	KDM-347	34
KDM-456A	7	CM-129	35
KDM-911A	8	CM-135	36
KDM-895A	9	CM-128	37
KDM-332A	10	CM-502	38
KDM-914A	11	CML-414	39
KDM-892A	12	CML-72	40
KDM-921A	13	CML-491	41
KDM-924A	14	CML-139	42
KDM-940A	15	CML-334	43
KDM-917A	16	HK1040-4	44
KDM-323A	17	HKI-586	45
KDM-356A	18	KDM-1016	46
KDM-940B	19	KDM-961	47
KDM-912A	20	KDM-439	48
KDM-362B	21	KDM-969	49
KDM-361A	22	KDM-1138	50
KDM-918A	23	KDM-404	51
KDM-935A	24	KDM-3001	52
KDM-932A	25	KDM-895	53
KDM-445A	26	KDM-915	54
KDM-343A	27	KDM-716	55
KDM-332B	28		

Table 2: Variability of the morphological traits in fifty five accessions of maize

Character	Range		Mean	Standard Deviation	CV (%)
	Minimum	Maximum			
Plant Height	100	270	193.127	41.892	0.217
Ear Height	42	122	92.491	20.082	0.217
Daysto50%Tasselling	67	76	71.600	2.122	0.030
Daysto50%Silking	71	79	75.327	2.001	0.027
75% Husk Browning	126	142	134.073	3.920	0.029
Cob Length	7.8	18.8	14.593	3.026	0.207
Cob/Plant	1.2	2	1.511	0.171	0.113
Rows/Cob	8	14.5	11.627	1.667	0.143
Grains/Row	14	35	27.540	5.859	0.213
Cob Diameter	2.8	5.12	4.174	0.603	0.144
100 Seed Weight	16.6	31.5	25.891	3.912	0.151
Yield/Plant	26	196	131.164	50.131	0.382

Table 3: Total Clusters and the accessions falling in each one of them

Cluster No.	Cluster Size	Accessions
1	10	P1, P2, P3, P4, P5, P6, P7, P8, P9, P10
2	12	P11, P14, P15, P20, P24, P33, P40, P44, P49, P50, P51, P54
3	6	P12, P21, P27, P30, P39, P52
4	14	P13, P22, P23, P28, P29, P31, P34, P35, P36, P46, P47, P48, P53, P55
5	9	P16, P18, P19, P25, P38, P41, P42, P43, P45
6	3	P17, P32, P37
7	1	P36

Table 4: Inter and Intra Cluster Distance.

Inter-Cluster distance							
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Cluster 1	0						
Cluster 2	150.34	0					
Cluster 3	184.71	50.27	0				
Cluster 4	135.27	39.61	54.15	0			
Cluster 5	178.96	34.37	45.15	61.12	0		
Cluster 6	177.87	50.87	77.73	83.26	43.03	0	
Cluster 7	121.34	96.91	104.73	71.45	119.45	139.98	0
Intra-Cluster distance							
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**	**Cluster 7**
Distance	23.93	20.23	19.85	26.92	21.88	29.65	0

Table 5: Cluster means

	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Cluster6**	**Cluster7**
Plant Height	124.2	213.58	208.83	183.07	234.78	256.67	118
Ear Height	56.5	102.33	101.5	88.14	110.78	114.67	110
Daysto50%Tasselling	71.1	71.83	70	72.36	71.78	71	73
Daysto50%Silking	74.9	75.67	73.5	76.21	75.33	74.33	77
75% Husk Browning	131.6	136.5	128.33	135.71	135.78	131	135
Cob Length	9	15.27	17.75	15.85	15.49	15.17	16.1
Cob/Plant	1.33	1.51	1.67	1.56	1.54	1.57	1.3
Rows/Cob	8.64	12.23	12.62	12.33	11.96	12.53	12.8
Grains/Row	15.67	29.72	31.87	30.26	30.01	28.83	30
Cob Diameter	3.06	4.25	4.67	4.49	4.31	4.62	4.4
100 Seed Weight	18.63	26.94	28.53	26.46	29.1	27.6	28
Yield/Plant	33.2	141.83	188.67	147.43	160.56	133.33	139

Table 6: Correlation between characters of 55 Accessions

Correlation Matrix												
	PlHt	**ErHt**	**Tsl**	**Sil**	**HB**	**CobLn**	**Cobpt**	**Rowcob**	**GrnRow**	**Cobdia**	**Sdwt**	**YPlnt**
PlHt	1.0000											
ErHt	0.8919	1.0000										
Tsl	-0.0819	-0.0679	1.0000									
Sil	-0.1444	-0.1106	0.9387	1.0000								
HB	0.1233	0.2235	0.4243	0.4692	1.0000							
CobLn	0.6497	0.7524	0.0114	0.0044	0.1494	1.0000						
Cobpt	0.4810	0.3674	-0.0644	-0.1300	-0.1174	0.3977	1.0000					
Rowcob	0.6287	0.7313	-0.0063	-0.0188	0.1895	0.7328	0.2976	1.0000				
GrnRow	0.7255	0.8123	0.0444	0.0350	0.2373	0.9422	0.4662	0.8021	1.0000			
Cobdia	0.6908	0.7631	-0.0770	-0.1022	0.1334	0.8452	0.4867	0.8594	0.8603	1.0000		
Sdwt	0.7819	0.8609	0.0401	0.0226	0.1846	0.8130	0.3490	0.7542	0.8587	0.7973	1.0000	
YPlnt	0.7473	0.7961	0.0079	-0.0249	0.0993	0.8827	0.5951	0.8141	0.9253	0.8816	0.8947	1.0000

Table 7: Eigenvalues for the PC's

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.922	4.615	0.5769	0.5769
2	2.307	1.405	0.1923	0.7692
3	0.901	0.284	0.0751	0.8443
4	0.616	0.120	0.0514	0.8957
5	0.495	-	0.0413	0.9370

Table 8: Eigen vectors representing the loadings on various PC's

Eigenvectors					
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**
Plant Height	0.319796	-.067038	-.027088	0.410889	-.511479
Ear Height	0.342512	-.021247	-.201203	0.238518	-.360651
50% Tasselling	-.007019	0.613101	0.281505	-.099333	-.218515
50% Silking	-.018869	0.627857	0.209387	-.141535	-.123449
75% Husk Browning	0.072156	0.445834	-.447179	0.581407	0.477140
Cob Length	0.344279	0.028048	-.006805	-.268538	0.149110
Cob/Plnt	0.200405	-.133102	0.755574	0.444714	0.278124
Row/Cob	0.325949	0.026441	-.169869	-.283235	0.199447
Grain/Row	0.362261	0.057522	-.001560	-.108771	0.144094
CoB Diameter	0.350457	-.039399	-.001455	-.164908	0.273434
Seed Weight	0.349616	0.045812	-.101304	-.074992	-.278682
Yield/Plant	0.365496	-.005338	0.160204	-.107666	0.071483

Figure 1: Cluster Plot and a Silhouette Plot

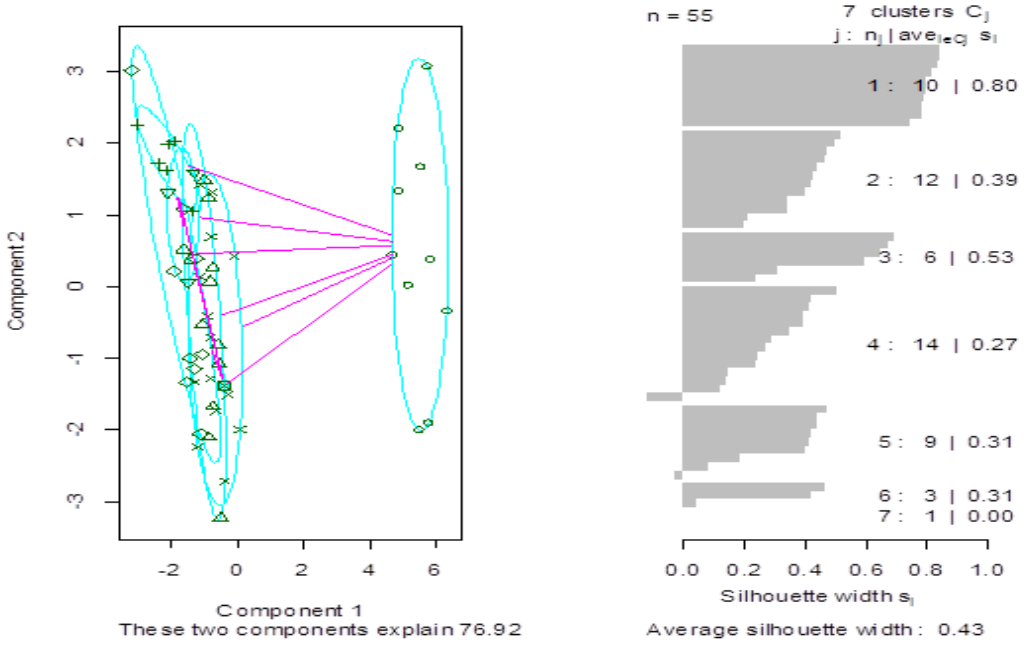


Figure 2: Scree Plot

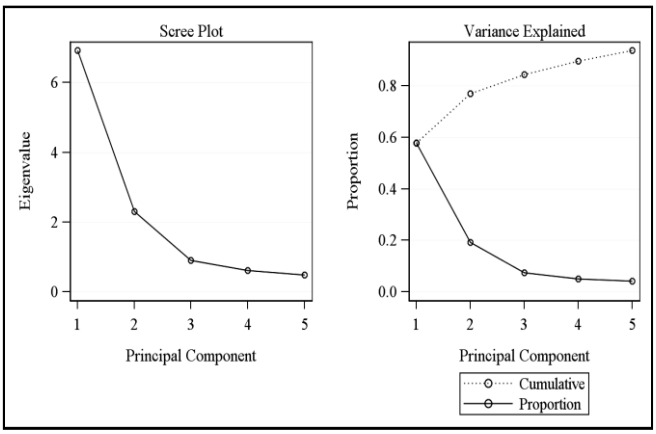


Figure 3: Component Pattern Profiles & Component Scores Matrix

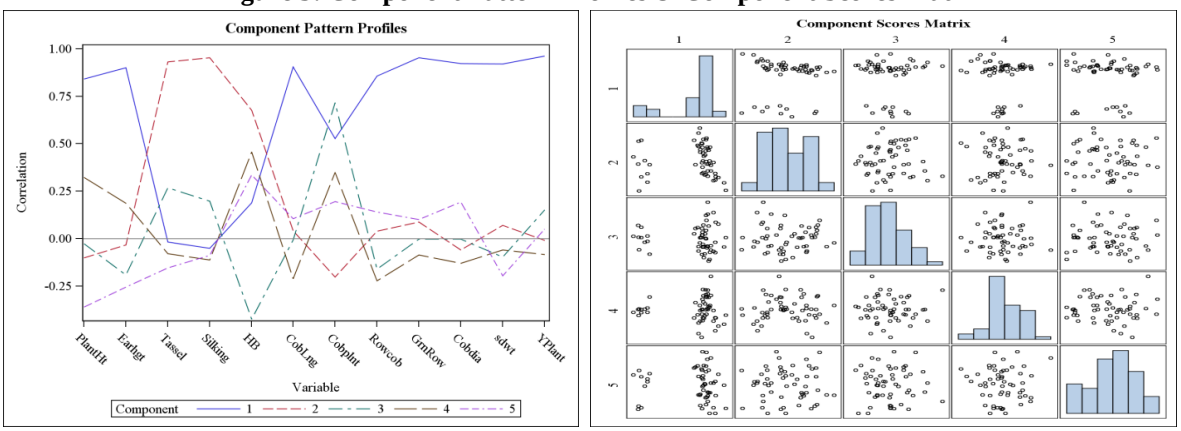
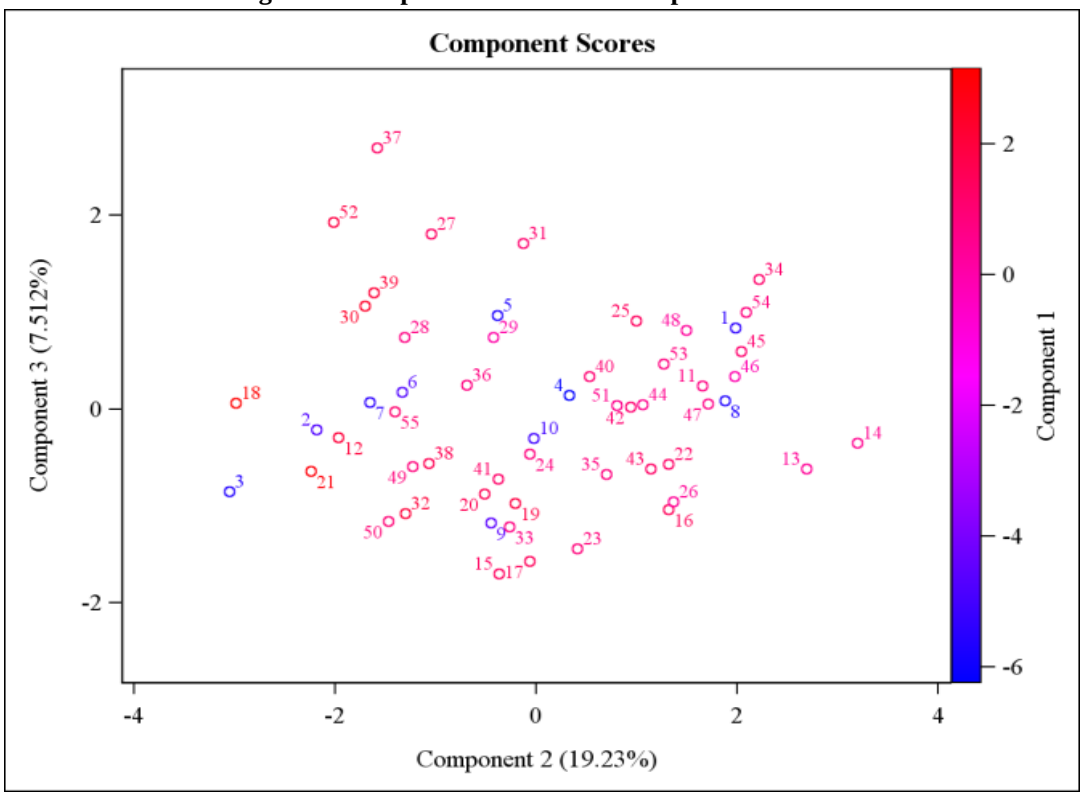
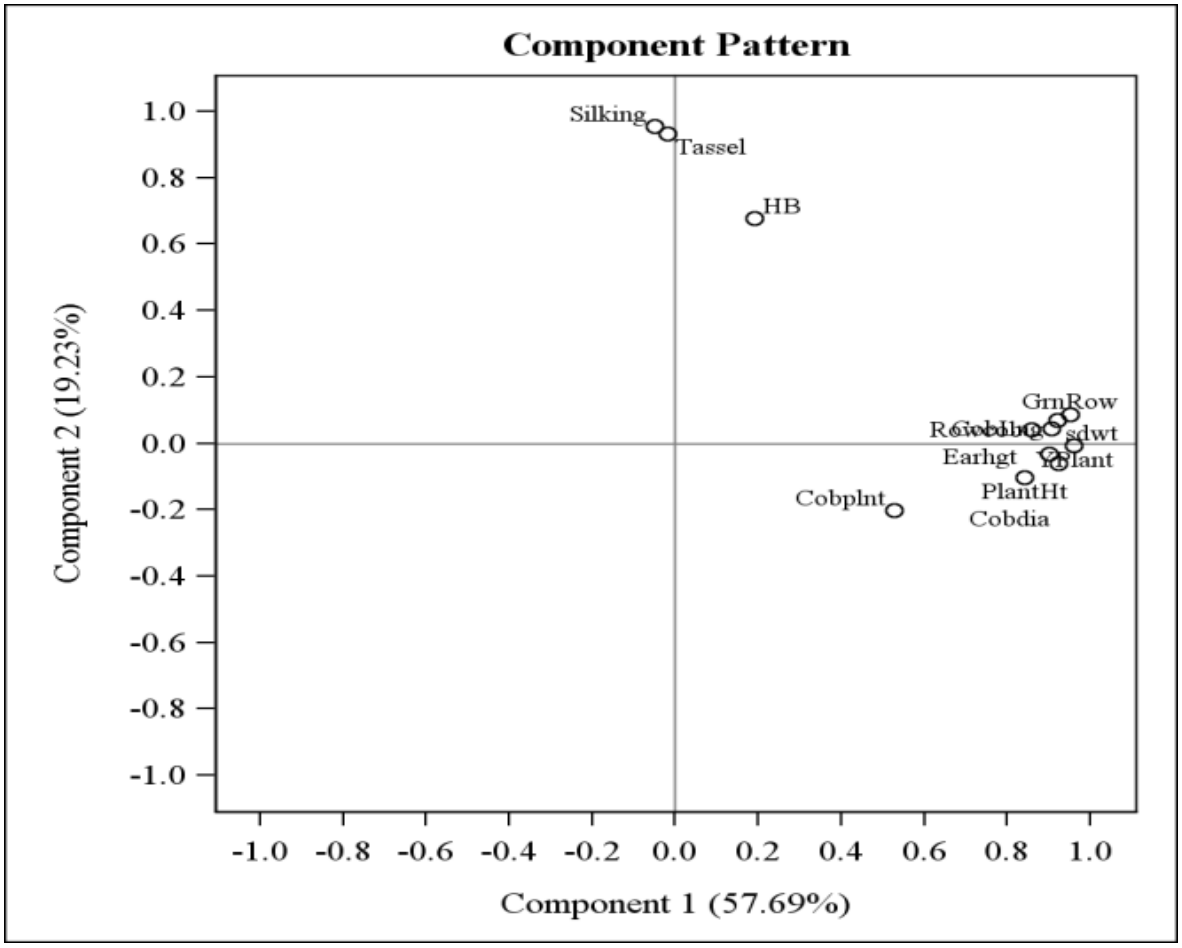


Figure 4: Component Scores and Component Pattern.





CONCLUSION

The data when subjected to Multivariate Analysis using CA, firstly sorted the accessions into groups, or clusters, such that the degree of association was strong between members of the same cluster and weak between members of different clusters. Secondly, PCA resulted in the reduction in the data from twelve parameters to just two principal components. The obtained principal components may also be used for further analysis. Moreover, the analysis is characterized by the fact that it includes the total variance of variables, explains maximum of variance within a data set, and is a function of primary variables. Thus, the Genetic divergence analysis estimated the extent of diversity that existed among selected accessions. Such precise information on the nature and degree of genetic diversity helps the plant breeder in choosing the diverse parents for purposeful hybridization.

Acknowledgement

The first author gratefully acknowledges the Division of Genetics & Plant Breeding, SKUAST-K, Shalimar for providing the data on the Maize genotypes.

REFERENCES

1. Chahal, G. S., Gosal, S. S., Principles and Procedures of Plant Breeding: Biotechnology and Conventional Approaches. *Narosa Publishing House*, New Delhi (2002).
2. Ghader, A., Adams, M. W., Nassib, A. M., Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and faba bean. *Crop Science*. **24:** 37-42 (1984).
3. Shah Immad, A., *et al*. "*International Journal of Chemical Studies.*" *Multivariate Clustering Utilizing R Software Analytics*, **6(1):** pp. 971–974., doi: 10.22271/chemi (2017).
4. Jolliffe, I. T., Discarding variables in a principal component analysis I: Artificial data. *Applied Statistics*, **21:** 160-173 (1972).
5. Kaiser, H. F., The varimax rotation for analytic rotation in factor analysis. *Psychometrika*, **23:** 187-200 (1958).
6. Kendall, M., *Multivariate Analysis* (Second Edition). Charles Griffin and Co London (1980).
7. Kovacic, Z., Multivariate analysis. Faculty of Economics. *University of Belgrade*. (In Serbian). P. 293 (1994).
8. Mahalanobis, P. C., on the generalized distance in statistics. *Proc. Nation. Acad. Sci*. (India) **2:** 49-55 (1936).
9. Mondal, M. A. A., Improvement of potato (Solanum tuberosum L.) through hybridization and in vitro culture technique. A Ph.D *Thesis. Rajshahi University, Rajshahi*, Bangladesh (2003).
10. Singh, R. K., Choudhary, B. D., Diametrical Methods in Quantitative Genetic Analysis. *Kalyani Publishers*, New Delhi, P. 318 (1985).
11. Sharma, J. R., Statistical and biometrical techniques in Plant Breeding. *New Age International*, New Delhi (1998).