AN APPROACH TO MULTIVARIATE AVERAGE LINKAGE CLUSTERING UTILIZING R SOFTWARE

Immad A. Shah
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India
S A Mir
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India
T A Raja
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India
Imran Khan
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India
Nageena Nazir
Sher-e- Kashmir University of Agricultural Sciences and Technology, J&K, India

Abstract: R software was used to hierarchical cluster the Maize genotypes and the distance measure used was Euclidean. Average Linkage was used to group the genotypes on the basis of the similarity. Cluster analysis using the R software grouped the 55 genotypes into distinct clusters. It was found that when the dendrogram for average linkage was cut at a distance of 4, it revealed two distinct clusters for the 55 genotypes. It clearly classified the genotypes, with cluster one containing the individual plants and cluster two containing crosses. Level plot was also obtained which indicated at least two distinct groups with large inter cluster distance. The various commands for R analysis are also mentioned in the paper.

Keywords: R Software, Clustering, Genetic divergence, Maize Genotypes, Multivariate Analysis.

Introduction: Clustering is a distribution free non parametric technique that is no assumptions are made concerning the number of groups and grouping is done on the basis of similarities or distance i.e. dissimilarities (Chatfield and Collins 1990; Jhonson and Wichern 1996). Multivariate analysis is the analysis of observations on several correlated random variables, for a number of individuals. Such analysis becomes necessary when one deals with several variables simultaneously. Multivariate analysis by means of Mahalanobis D^2 Statistics and Euclidean distances is widely used for quantifying the degree of genetic divergence among the population. The genetic divergence analysis estimates the extent of diversity existed among selected genotypes (Mondal, 2003).The cluster analysis is 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. Precise information on the nature and degree of genetic diversity helps the plant breeder in choosing the diverse parents for purposeful hybridization (Samsuddin, 1985).

Agglomerative hierarchical clustering techniques differ primarily in how they measure the distances between or similarity of two clusters (where a cluster may, at times, consist of only a single individual). Two simple intergroup measures are:

$$d_{AB} = min(d_{ij}),$$
$$d_{AB} = max(d_{ij}),$$

Where, d_{AB} is the distance between two clusters A and B, and d_{ij} is the distance between individuals i and j. This is the Euclidean distance. The first intergroup dissimilarity measure above is the basis of single linkage clustering, the second that of complete linkage clustering. A further possibility for measuring intercluster distance or dissimilarity is:

$$d_{AB} = \frac{1}{nAnB} \sum\sum d_{ij}$$

Where, n_A and n_B are the number of individuals in clusters A and B. This measure is the basis of a commonly used procedure known as group average clustering. Average linkage measures the average distance between the two clusters.

Materials and Methods: Data collected for the study constitutes the data base for the present investigation. The material for this study is composed of 55 genotypes of maize. These lines were maintained at SKUAST-K Shalimar and KD station by Department of Genetics and Plant Breeding, SKUAST K Shalimar. The data was subjected to different types of cluster analysis. The experiment was laid out in a RCBD consisting of two replications each containing 55 genotypes considered as treatment. The data generated from the experiment on maize conducted by DARS, Budgam, SKUAST-K, has been used for this study. Data comprised of 55 genotypes of maize, out of which 10 genotypes were individual plants and 45 were crosses and 12 characters were recorded for each of the genotypes. The characters recorded 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. Genetic diversity was studied using Mahalanobis (1936) generalized distance (D^2) extended by Rao (1952).

Result and Discussion: We start our analysis by computing the dissimilarity matrix containing the Euclidean distance of the plant characters on all 55 genotypes. Agglomerative hierarchical cluster analysis that produces partitions by a series of successive fusions of the *n* individuals into groups have been carried out. The resulting 55×55 matrix can be inspected by the level plot (Figure 3) obtained by the function ***levelplot().***Correlation between the characters was observed. The function for obtaining the correlation is : ***cor()***. Several characters were found to be highly correlated such as grain/row and yield (0.925), 50% Tasseling and 50% silking (0.939), cob length and grain per row (0.942), and 100 seed weight and Yield (0.895) as shown in Table 1 and Figure 1.

Table 1: Correlation between the Traits

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

Figure 1: Correlation Matrix Plot

To begin with the euclidean distances are calculated using the ***dist()*** function.

> ***cluster = read.table("clipboard",header=T)***
> ***dist (cluster[,-1])***
> ***distance =dist(cluster[,-1])***

Here we have omitted the 1st column of the data matrix which specifies the genotypes of maize. The function ***hclust()*** performs the average linkage clustering based on the distance matrix of the data. The corresponding plot method draws a dendrogram.

> ***hier=hclust(distance,method="average")***

Here the output is named as ***hier*** and to plot the dendrogram the ***plot*** function is used

> ***plot(hier)***.

The output plot is shown in the figure 2 below:

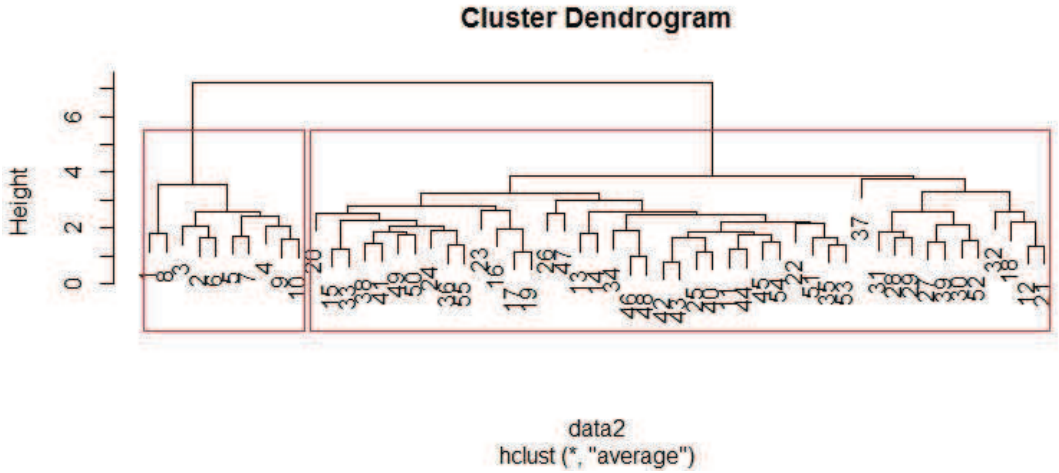

Figure 2: Average Linkage Clustering

Here cluster analysis using Average linkage classified our data set into two types or groups viz. Type 1 and Type 2. Type 1 containing the individual plants and Type 2 containing the crosses. The dendrogram suggests that the cluster 1 contains 10 genotypes which are the individual plants and Cluster 2 contains all the rest of the genotypes which are crosses. The level plot obtained for the Euclidean distances is shown in Fig 3. The diagonal elements of the dissimilarity matrix are dark coloured indicating zero distance while as the pale values indicate greater distance. The level plot indicates that there are at least two distinct groups with large inter cluster distance whereas much larger distance can be observed for the other cells.

Many algorithms have been proposed for cluster analysis but here only average linkage clustering have been used to group the 55 Maize genotypes. Groups are then formed by the process of agglomeration or division. A dendrogram for the average linkage clusters results in the Fig. 2 and Table 2. Also, the level plot for the Euclidean distances is shown in Fig. 3.

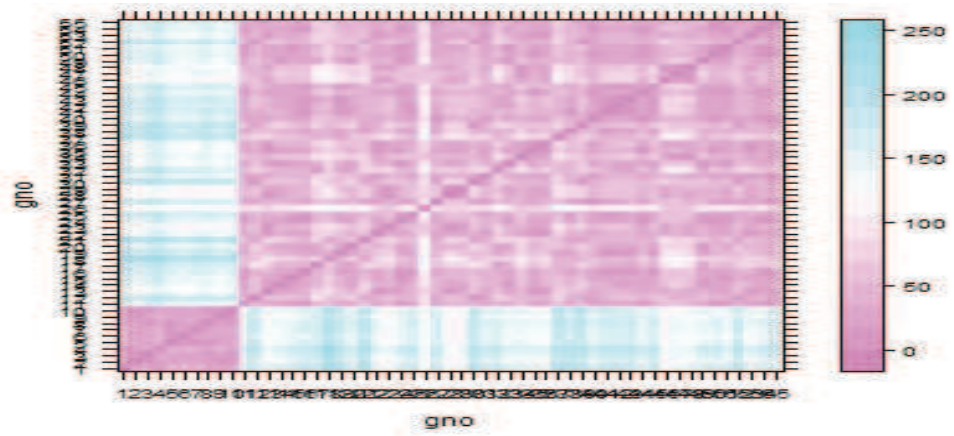

Figure 3: Level Plot

It displays the all distances of the Euclidean matrix of order 55× 55. The x axis represents the total number of observations while as the y axis represents the distances. When the cluster is cut a distance of 4 it specifies two distinct clusters as shown in the dendrogram. The cluster 1 contains individual genotypes (P1, P2, P3, P4, P5, P6, P7, P8, P9 and P10) while as the cluster 2 contains crosses as shown in Fig. 2 and Table 2. 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 (Chahal and Gosal, 2002).

Table 2: Clusters Specifying the Genotypes (Average Linkage Basis)

Cluster	**Genotypes**
1	**P1,P2,P3,P4,P5,P6,P7,P8,P9,P10**
2	**P11,P12,P13,P14, P15,P16,P17,P18, P19,P20,P21,P22, P23,P24,P25,P26 P27,P28,P29,P30, P31,P32,P33,P34 P35,P36,P37,P38, P39,P40,P41,P42, P43,P44,P45,P46, P47,P48,P49,P50, P51,P52,P53,P54,P55**

Table 2 indicates that the genotypes in their respective clusters are homogenous amongst themselves and heterogeneous in between.

Conclusion: Clustering provides considerable useful information about genetic diversity. A significant genetic diversity is found among the genotypes. The above 55 genotypes are classified into 2 distinct clusters on average linkage basis. Thus, the cluster analysis (a multivariate technique) can be utilised as a data classifying technique and can be utilised in diversifying huge gene banks.

References:

1. Chatfield, C. & Collins, A.J. 1990. Introduction to Multivariate Analysis. London: Chapman and Hall, pp. 50, 195, 205, 212-230.
2. Chahal, G.S, Gosal, S.S (2002). Principles and Procedures of Plant Breeding: Biotechnology and Conventional Approaches. Narosa Publishing House, New Delhi.
3. Johnson, R.A. & Wichern, D.W. 1996. Applied Multivariate Statistical Analysis. Englewood Cliffs, NJ: Prentice –Hall, pp. 532-578
4. Mahalanobis PC (1936). On the generalized distance in statistics. *Proc. Nation. Acad. Sci*. (India) 2:49-55.
5. Mondal, M.A.A. 2003. Improvement of potato (Solanum tuberosum L.) through hybridization and in vitro culture technique. A Ph.D Thesis. Rajshahi University, Rajshahi, Bangladesh.
6. Rao, C.R. 1952. Advanced Statistical Methods in Biometrics Research John Wiley and Sons, New York, pp. 357-369.
7. Samsuddin, A.K. 1985. Genetic diversity in relation to heterosis and combining analysis in spring wheat. *Theoretical Appl. Genet*. 70:306308.

***