Indian J. Genet., **51** (3): 335–341 (1991)

A COMPARISON OF CLUSTERING PROCEDURES BASED ON MULTIPLE TRAITS IN *GERBERA* AND *DAHLIA*

S. D. WAHI AND K. K. KHER

Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110012

(Received: August 19, 1989; accepted: September 4, 1990)

ABSTRACT

The multivariate methods, namely, principal component analysis (PCA), Tocher's technique, and an iterative method of clustering based on successive reallocation of elements using Mahalanobis D^2 statistic, were compared on the basis of average intracluster distance and a ratio index of homogeneity of clusters utilizing the intra- and intercluster distances. The results from two different populations under study showed a superior performance of the iterative method. The optimum clusters formed by the iterative method were uniformly more homogeneous and unique as compared to the clusters obtained by the other two procedures. In addition to these advantages, the iterative method can be easily computerised and is useful in handling the large number of genotypes generated in breeding programmes, whereas the other two methods need a lot of manual labour in addition to their computer analysis.

Key words: Cluster analysis, D^2 statistic, *Gerbera*, *Dahlia*, component analysis, iterative method.

The summarization of large quantities of data by various multivariate methods, especially by cluster analysis, is increasingly practised in various branches of science. The cluster methods refer to whole subject or restricted class of partitioning or clumping the objects on the basis of similarities/dissimilarities. Several measures of similarity/dissimilarity, such as, Euclidean distances, city-block metric [1], Minkowski metrics [2], D^2 statistic [3, 4] etc. are used as the basis of clustering. An exhaustive review of the various methods of clustering has been given by Cormack [5]. The various methods of clustering have one thing in common, i.e., whether two groups are to be regarded as separate clusters depends on the distances between their means as compared to the mean distances within the clusters.

One of the basic problems faced by breeders is to classify a large number of genotypes into fewer number of homogeneous clusters. For the past 30 years the D^2 statistic of Mahalanobis [4] has been widely used by the breeders as a measure of the distance between

two populations. These distances coupled with principal components are used to form homogeneous clusters of large numbers of genotypes. The most widely used procedures of clustering using D^2 statistic is the Tocher's technique given by Rao [6]. Two of the serious drawbacks in this technique are that the stopping rule for formation of any cluster is arbitrary and the genotypes are often wrongly clustered which increases the average intracluster distance. Moreover, this method of clustering is manual and takes a long time when a large number of objects are to be clustered. An iterative clustering method, which can be computerized and is free from the above drawbacks, can be used effectively. In this paper, attempts have been made to compare the performance of the three different methods of clustering, i.e., principal component analysis, Tocher's method, and computer oriented iterative method with the help of two different sets of data on *Gerbera* and *Dahlia*.

MATERIALS AND METHODS

The data on 31 cultivars of *Gerbera jemesonaides* and 39 cultivars of *Dahlia variabilis*, collected from different sources in the country, maintained at Hessarghatta Farm of the Indian Institute of Horticultural Research, Bangalore were used. Data were collected on nine morphological characters of *Gerbera*: shoots/plant, leaves/plant, leaf length, leaf width, days from flower bud appearence to opening, length of flower stalk, flowers/plant, flower diameter, and longevity of flowers. Similarly, data were collected on seven morphological characters of *Dahlia*: plant height, leaves/plant, branches/plant, flowers/plant, flower diameter, flower stalk length, and flower longevity. The D^2 statistic based on multiple characters among the strains of both sets of data were obtained by the method of pivotal condensation as described by Rao [6]. The principal component analysis was also performed on both the sets.

The computer algorithm used for iterative procedure for clustering the genotypes was used after modification as described below:

1. Identify the two genotypes having maximum dissimilarity and consider them as nuclei of the first two clusters.

2. The remaining genotypes are considered one by one and allotted to the cluster for which the dissimilarity with the nucleus is minimum.

3. Take out the genotypes excluding those forming the nuclei one by one from the cluster to which it was allocated, calculate average dissimilarity with all clusters and reallocate the genotype to a cluster from which its average dissimilarity is minimum. This step is repeated again until the clusters remain stable.

4. To increase the number of clusters, the genotype showing highest dissimilarity to the nucleus of its cluster is taken as an additional nucleus. New clusters are formed and optimized by repeating steps 2 and 3.

RESULTS AND DISCUSSION

The first two canonical roots in *Gerbera* and *Dahlia* populations explained 55% and 76% of the total variation, respectively.

The distribution of strains to different clusters formed by the three procedures along with their average intra- and intercluster distances (D^2 value) are given in Tables 1 and 2. The distribution of strains over different clusters by Tocher's method is not as widespread as in the case of PCA and the iterative procedure. The Tocher's method has a tendency of clustering most of the strains in a single cluster. The average intracluster distance was the smallest at all the stages of clustering due to the iterative method (Fig 1, 2). A plot of the average intracluster distance against the number of clusters formed by the three procedures showed a regular fall in the average intracluster distance with an increase in the number of clusters formed by the iterative procedure in both *Gerbera* and *Dahlia* populations. This regular decline in the average intracluster distance in case of iterative method gives the clue to optimum number of clusters as 7 and 4 in the *Gerbera* and *Dahlia* populations, respectively. In spite of low intracluster distance shown by the iterative method, there was no regular trend of intercluster distances. The results of *Gerbera* populations show that the intercluster distance was higher among the clusters formed by the Tocher's method as compared to the PCA and iterative method. But in case of *Dahlia* populations, the intercluster distance was initially higher among the clusters formed by the iterative procedure as compared to the Tocher's method and this trend was reversed with increase in the number of clusters.

Theoretically, the best method of clustering will be the one which gives the most homogeneous clusters (minimum intracluster distance) along with the maximum possible intercluster distance. As none of these three procedures satisfies both the conditions uniformally, as it was decided to form an index of homogeneity of clusters by taking the ratio of intracluster to intercluster distance to compare the efficiency of the three procedures; the lower the value of this index the more homogeneous will be the clusters. The ratio index of the iterative method was uniformally lower as compared to the PCA and Tocher's method, except in the two initial stages of clustering in *Gerbera* (Table 3). The ratio index also exhibited a regular fall with the increase in number of clusters. This further confirms the optimum number of clusters formed in the two populations by this procedure. The results of PCA were also comparable to those of iterative method in the *Dahlia* populations. This was perhaps due to the higher percentage of total variation explained by the first two principal components in *Dahlia*.

Table 1. Composition of clusters along with their intra- and intercluster D^2 values among 31 strains of *Gerbera**

Cluster No.	Tocher's method		Principal component method		Iterative method	
1	12, 24		7, 8, 12, 13, 19, 20, 22, 25, 26, 27		7, 8, 9, 12, 13, 15, 19, 20, 21,	
		103.86		93. 97	22, 23, 25, 26, 27, 28	91.14
	Rest 29 elements	(186.39)	Rest 21 elements	(138.67)	Rest 16 elements	(135.74)
1	7, 12, 13	91.33	12		7, 8, 12, 13, 19, 20, 22, 25, 26,	
2	14, 24	(169.37)	7, 8, 13, 19, 20, 21, 22, 23, 25,		27, 1, 2, 9, 10, 11, 16, 17, 18,	
3	Rest 26 elements		26, 27, 28	88.52	24, 29, 31	
			Res 17 elements	(136.01)	3, 4, 5, 6, 14, 15, 21,	75.88
					23, 28, 30	(131.50)
1	7, 13		12		14, 24	
2	14, 24	91.33	7, 8, 13, 19, 20, 22, 25, 26,		7, 8, 12, 13, 19, 20, 22, 25, 26, 27	
3	12	(168.73)	27 3, 4, 5, 6, 14, 15, 16, 21, 23,		3, 4, 5, 6, 15, 21, 23, 28, 30	
4	Rest 26 elements		24, 28	69.54	1, 2, 9, 10, 11, 16,	65.74
			Rest 10 elements	(132.60)	17, 18, 29, 31	(132.35)
1	7, 13, 27		12		24	
2	6, 14, 16		13	68.48	3, 4, 16, 7, 8, 12,	64.11
3	12	83.71	7, 8, 19, 20, 22, 25,	(131.53)	13, 25, 26, 27	(131.18)
		(152.43)	26, 27			
4	24		3, 4, 5, 6, 14, 15, 16, 21, 23, 24, 28		4, 5, 6, 15, 21, 23, 28, 30	
5	Rest 23 elements		Rest 10 elements		Rest 12 elements	
1	4, 6, 30		12		12	
2	7, 13, 22, 27		13		24	
3	14, 16		7, 8, 19, 20, 22, 25, 26, 27		7, 8, 13, 22, 25, 26, 27	55.13
4	12, 26	73.82	3, 4, 5, 6, 14, 16, 24	62.52	3, 4, 6, 14, 16, 23	(129.52)
5	24	(140.09)	15, 21, 23, 28	(128.33)	5, 15, 19, 20, 21, 28	
6	Rest 19 elements		Rest 10 elements		Rest 12 elements	
1	12		12		12	
2	24		13		24	
3	7, 26	73.31	1, 10, 17, 18	66.08	3, 4, 6, 14, 16	46.19
4	14, 16	(139.11)	15, 21, 23, 28	(123.55)	5, 15, 21, 28	(128.07)
5	4, 6, 30		2, 9, 11, 29, 30, 31		8, 13, 19, 20, 22	
6	13, 22, 27		3, 4, 5, 6, 14, 16, 24		7, 23, 26, 30	
7	Rest 19 elements		7, 8, 19, 20, 22, 25, 26, 27		1, 2, 9, 10, 11, 17, 18, 29, 31	
1	7		12		12	
2	12	63.19	13		24	
3	14	(133.93)	3, 5, 24	58.86	3, 4, 6, 14, 16	41.50
4	24		1, 10, 17, 18	(122.98)	5, 15, 21, 28	(127.25)
5	26, 30		4, 6, 14, 16		8, 19, 20, 22, 25, 27	
6	13, 22, 37		15, 21, 23, 28		1, 2, 9, 10, 11, 17, 18, 29, 31	
7	4, 5, 6, 15, 16, 21, 23		2, 9, 11, 29, 30, 31		7, 13	
8	Rest 15 elements		7, 8, 19, 20, 22, 25, 26, 27		23, 26, 30	
1	6		12		12	
2	12		13		24	
3	14	52.92	27	58.95	4, 6	38.39
4	22	(131.15)	3, 5, 24	(121. 87)	7, 13	(126.47)
5	24		1, 10, 17, 18		3, 14, 16	
6	26, 30		4, 6, 14, 15		23, 26, 30	
7	7, 13, 25, 27		15, 21, 23, 28		5, 15, 21, 28	
8	3, 4, 5, 15, 16, 21, 23, 28		2, 9, 11, 29, 30, 31		8, 19, 20, 22, 25, 27	
9	Rest 12 elements		7, 8, 19, 20, 22, 25, 26		1, 2, 9, 10, 11, 17, 18, 29, 31	

*The figures in parentheses are the intercluster D^2 values.

Table 2. Composition of clusters along with their intra- and intercluster D^2 values among 39 strains of *Dahlia**

Cluster No.	Tocher's method		Principal components method		Iterative method	
1	16, 30, 32, 2	62. 74	14, 22, 23, 27, 34	55.81	14, 22, 23, 27, 34	55.81
2	Rest 35 genotypes	(133.45)	Rest 34 genotypes	(144.24)	Rest 34 genotypes	(144.24)
1	35	58.34	2, 13, 16, 30, 32	33.79	2, 13, 16, 28, 30, 32	32.98
2	2, 13, 16, 30, 32	(123.15)	14, 22, 23, 27, 34	(133.30)	14, 22, 23, 27, 34	(127.56)
3	Rest 33 genotypes		Rest 29 genotypes		Rest 28 genotypes	
1	15, 35		27		15, 35	
2	2, 13, 16, 30, 32	26.22	2, 13, 16, 30, 32	33.55	2, 13, 16, 30, 32	26.22
3	14, 22, 23, 27, 34	(126.38)	14, 22, 23, 34	(132.38)	14, 22, 23, 27, 34	(126.38)
4	Rest 27 genotypes		Rest 29 genotypes		Rest 27 genotypes	
1	15, 35		27		15, 35	
2	16, 27		2, 13, 16, 30, 32		2, 13, 30, 32	
3	2, 13, 30, 32	25.54	14, 22, 23, 34	30.70	22, 23, 34	25.69
4	14, 22, 23, 24	(125.17)	28, 36, 39	(115.42)	14, 16, 27	(124. 76)
5	Rest 27 genotypes		Rest 26 genotypes		Rest 27 genotypes	
1	27		27		14, 27	
2	15, 35		15, 35		15, 35	
3	16, 30	25.21	14, 22, 23, 34	21.77	22, 23, 34	21.30
4	2, 13, 32	(124.70)	2, 13, 16, 30, 32	(112.23)	2, 13, 28, 36, 39	(112.14)
5	14, 22, 23, 24		28, 36, 39		16, 30, 32	
6	Rest 27 genotypes		Rest 24 genotypes		Rest 24 genotypes	
1	32		27		13, 32	
2	15, 35		14, 22		14, 27	
3	14, 27		23, 34		15, 35	
4	2, 13, 28	21.74	15, 35	21.58	22, 23, 34	20.88
5	16, 30, 36	(115.41)	28, 36, 39	(111.55)	2, 28, 39	(111.20)
6	22, 23, 34		2, 13, 16, 30, 32		16, 30, 36	
7	Rest 25 genotypes		Rest 24 genotypes		Rest 24 genotypes	
1	32		27		32	
2	36		14, 22		14, 27	
3	15, 35		15, 35		15, 35	
4	14, 27	21.63	23, 34	18.63	16, 30	18.80
5	16, 30	(115.05)	28, 36, 39	(95.08)	2, 13	(194.75)
6	2, 13, 28		2, 13, 16, 30, 32		22, 23, 34	
7	22, 23, 34		3, 6, 17, 19, 25, 33		20, 26, 28, 29, 36, 39	
8	Rest 25 genotypes		Rest 18 genotypes		Rest 21 genotypes	
1	22		27		32	
2	32		32		14, 27	
3	15, 35		14, 22		15, 35	
4	14, 27	19.26	15, 35	17.92	16, 30	15.45
5	16, 30	(107.55)	23, 34	(94.76)	2, 13	(90.10)
6	2, 13		28, 36, 39		22, 23, 34	
7	28, 29, 36, 39		2, 13, 16, 30		3, 6, 17, 19, 25, 33	
8	23, 34		3, 6, 17, 19, 25, 33		20, 26, 28, 29, 36, 39	
9	Rest 23 genotypes		Rest 18 genotypes		Rest 15 genotypes	

*The figures in parentheses are the intercluster D^2 values.

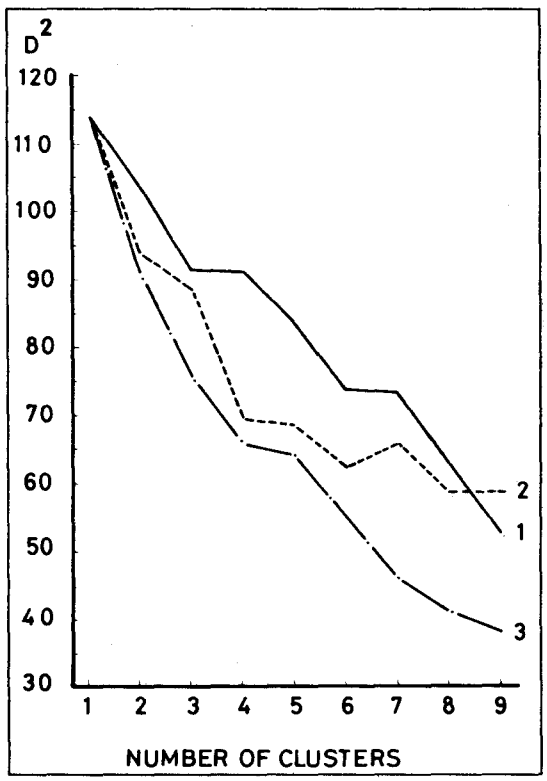

Fig. 1. Average intercluster D^2 values of clusters formed by three different procedures in *Gerbera*:

(1) Tocher's method; (2) principal component analysis; and (3) iterative method.

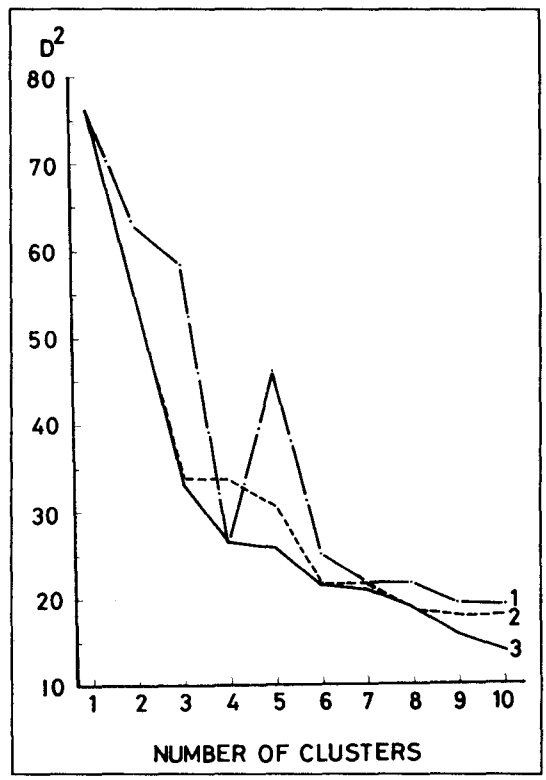

Fig. 2. Average intracluster D^2 values of clusters formed by three different procedures in *Dahlia*:

(1) Tocher's Method; (2) principal component analysis; and (3) iterative method.

The results of the present study indicate that the iterative method of clustering has the following advantages over the PCA and Tocher's method: 1) the clusters formed by the iterative method are more homogeneous as compared to the other two methods, 2) the iterative method provides the unique clusters at all the stages of clustering, whereas arbitrary decisions are required to be taken in the other two methods, 3) the iterative method has the flexibility of computerization and the final clusters can be obtained on the computer, whereas in the other two procedures manual work is also required; and 4) the iterative method also provides the optimum number of clusters, whereas the decision in the other two methods is arbitrary.

Table 3. Ratio of average intra- and intercluster D^2 values of the clusters formed by the three procedures in *Gerbera* and *Dahlia* populations

Cluster No.	*Gerbera*			*Dahlia*		
	Tocher's method	P.C.A.	Iterative method	Tocher's method	P.C.A.	Iterative method
2	0.56	0.68	0.67	0.47	0.39	0.39
3	0.54	0.65	0.58	0.47	0.25	0.26
4	0.54	0.52	0.50	0.21	0.25	0.21
5	0.55	0.52	0.49	0.20	0.27	0.20
6	0.53	0.49	0.43	0.20	0.19	0.19
7	0.53	0.53	0.36	0.19	0.19	0.19
8	0.47	0.48	0.33	0.19	0.20	0.20
9	0.40	0.48	0.30	0.18	0.19	0.17
10	—	—	—	0.18	0.19	0.16

REFERENCES

1. R. L. Johnson and D. D. Wall. 1969. Cluster analysis of semantio differential data. Educ. Psycolal. Measure., **29**: 769–780.

2. A. J. Boyce. 1969. Mapping diversity: a comparative study of some numerical methods. *In*: Numerical Taxonomy (ed. A. J. Code). Academic Press, New York: 30.

3. P. C. Mahalanobis. 1930. On tests of measures of divergence. J. Proc. Asiat. Soc. Beng., **36**: 541–588.

4. P. C. Mahalanobis. 1936. On the generalised distances in statistics. Proc. Natl. Inst. Sci. India., **2**: 49–55.

5. R. M. Cormack. 1971. A review of classification (with discussion). J. R. Stat. Soc. A., **134**: 321–367.

6. C. R. Rao. 1952. Advanced Statistical Methods in Biometerical Research. John Wiley and Sons, New York, USA: 390.