STANDARDISATION OF TECHNIQUES OF CLUSTERING GENOTYPES USING MAHALANOBIS D^2 AND WILKS' Λ CRITERION

By

SURESH K. M.

THESIS

submitted in partial fulfilment of the
requirement for the degree

Master of Science (Agricultural Statistics)

Faculty of Agriculture
Kerala Agricultural University

Department of Statistics
COLLEGE OF VETERINARY AND ANIMAL SCIENCES
Mannuthy - Trichur

1986

To My Loving Parents

DECLARATION

I hereby declare that this thesis entitled "STANDARDISATION OF TECHNIQUES OF CLUSTERING GENOTYPES USING MAHALANOBIS D^2 AND WILKS' Λ CRITERION" is a bonafide record of research work done by me during the course of research and that the thesis has not previously formed the basis for the award to me of any degree, diploma, associateship or other similar title of any University or Society.

Suresh

Mannuthy,
22-07-1986

(SURESH, K.M.)

CERTIFICATE

Certified that this thesis entitled "STANDARDISATION OF TECHNIQUES OF CLUSTERING GENOTYPES USING MAHALANOBIS D^2 STATISTIC AND WILKS' Λ CRITERION" is a record of research work done independently by Mr. SURESH, K.M. under my guidance and supervision and that it has not previously formed the basis for the award of any degree, fellowship or associateship to him.

V K GOPINATHAN UNNITHAN,
CHAIRMAN, Advisory Board,
Associate Professor of
Agricultural Statistics,
College of Horticulture.

VELLANIKKARA,
22-07-1986.

ACKNOWLEDGEMENTS

With immense pleasure I place on record my profound sense of gratitude and personal indebtedness to Sri V.K. Gopinathan Unnithan, Chairman of the advisory committee, Associate Professor of Agricultural Statistics, College of Horticulture under whose meticulous guidance this work has been executed.

I express my sincere thanks to Dr. K.C. George, Professor and Head, Department of Statistics, College of Veterinary and Animal sciences for his valuable guidance and encouragement throughout the study.

I place on record my indebtedness to sri P.V. Prabhakaran, Professor of Statistics, College of Agriculture for his encouragement and help rendered.

Grateful acknowledgement is made to Dr. K.V. Peter, Professor of Olericulture, College of Horticulture and member of the advisory committee for his whole hearted co-operation and valuable suggestions in the preparation of the thesis.

I am thankful to Dr. Rajeevan, P.K., Assistant Professor, Department of Pomology and Floriculture for providing the necessary data for the study and the suggestions offered.

I express my sincere thanks to the Dean College of Veterinary and Animal sciences and Dean college of Horticulture for providing the necessary facilities.

I am grateful to the Kerala Agricultural University for offering financial assistance in the form of fellowship.

The help rendered by Miss Laly John C., Mrs. Malika, V. and Miss Alphy Korath, Junior Assistant Professors of Agricultural Statistics are specially acknowledged.

Lastly, but not leastly I extend my sincere thanks to the staff, department of statistics and to all my friends for their generous help and encouragement.

Suresh

(SURESH K.M.)

CONTENTS

Introduction

INTRODUCTION

Multivariate analysis is very effective to study any object or objects characterised by a number of traits. Any biological phenomenon is manifested through a number of traits and hence could be studied more effectively by multivariate analysis than analysis of several univariate cases.

It is of immense use to biologists and more so to plant breeders to form clusters of a given number of genotypes such that there is more homogeneity among genotypes within clusters than between clusters.

Eventhough a number of clustering algorithms are available, they are not being used for clustering of genotyes. This could perhaps be due to the fact that the theoretical concept of clustering is still vague and in the initial stage. Another reason could be that none of them take within variation of genotypes into consideration.

Mahalanobis D^2 statistic, a measure of distance between two populations taking the within variation also into consideration, is very widely used to cluster genotypes. The procedure now being followed in formation of clusters was suggested by Tocher in a discussion on Rao (1948). One of the drawbacks of this procedure is that the stopping rule for clustering is very arbitrary. Some workers take it as that, if the average D^2 between a genotye and a cluster currently being formed is less

than the maximum among the minimum D^2 values of the genotypes, it is included in the cluster and otherwise not (Singh and Choudhary, 1979). This is very arbitrary and if one of the genotypes is far distant from the rest all the remaining genotypes will form a single cluster. A second disadvantage noticed is that once the clustering is over, often a genotype belonging to a cluster will have atleast on the average a smaller D^2 value with genotypes of a different cluster than the one to which it belongs to.

Wilks Λ criterion was developed to test the differences among a number of populations with respect to a number of characters, where as Mahalanobis D^2 was to test the differences between two populations. Hence a procedure using Λ criterion to form clusters is expected to be much more effective and meaningful than the one using D^2.

Hence the study was taken up with the following objectives.

a). To obtain a critical value for the within cluster distances, similar to the least significant difference, to group a number of genotypes using Mahalanobis D^2.

b). To evolve a method of clustering using Wilks Λ criterion.

c). Comparison of these new methods with the existing widely used methods of Tocher and canonical variates through illustrative examples.

Review of Literature

REVIEW OF LITERATURE

Multivariate analysis which is an extension of univariate analysis gathered momentum when Wishart (1928) derived the distribution of the dispersion matrix of a vector of variables having multivariate normal distribution.

Consequently Hotelling (1931) derived T^2 statistic, the multivariate analog of student's t, to test the difference in two populations with respect to a number of characters.

Wilks (1932) extended analysis of variance in the univariate case to analysis of dispersion in multivariate case. The Λ statistic derived by Wilks is the product of a number of beta variates and hence is very unwieldy to have exact tests of significance. Consequently various workers suggested approximate tests based on Λ statistic.

Bartlett (1947) suggested an approximate test based on Λ as $-m\log_e\Lambda$, which follows a chisquare distribution. Rao (1951, 1973) provided a better approximation. These are useful to test the homogeneity of a group of objects, prior to any attempt on classification.

Early works of classifying objects were initiated due to Karl Pearson's coefficent of racial likeness (C.R.L.) (Tildesley, 1921). If n_{1i} and n_{2i} are the numbers of observations on which the means m_{1i} and m_{2i} of the ith character for the

first and second group are based, and s_i is the standard deviation of the i^{th} character, and 'p' is the number of characters used, the C.R.L. was given as

$$\frac{1}{p} \sum_{i=1}^{p} \frac{n_{1i}\, n_{2i}}{n_{1i}+n_{2i}} \cdot \frac{(m_{1i}-m_{2i})^2}{s_i^2}$$

The C.R.L. does not take the relationship among the different characters into account. All the characters are treated as independent. Also it is greatly affected by slight variation in sample sizes

The concept of generalised distance between populations was developed by Mahalanobis (1936) and the square of the generalised distance between two populations namely Mahalanobis D2 was defined.

Tocher (Rao, 1948) proposed a simple device for obtaining group constellations using D2 values. Rao (1952) opined that the only criterion that could be considered for clustering was that, any two genotypes belonging to the same cluster should atleast on an average have a smaller D^2 than those belonging to two different clusters.

The concept of generalised distance between groups and its use in formation of group constellation have been discussed by Mahalanobis et al. (1949); Rao (1952); Majumdar and Rao (1958).

Rao and Slater (1949) made use of D2 statistic to classify six neurotic groups. Nair (1952); Blackith (1957) found that it worked well in entomological problems relating to the study of desert locusts. Mukherjee (1951) described the applications of D^2 in anthropometric measurments.

Mahalanobis D2 statistic was utilised with great advantage in plant breeding for measuring genetic divergence and forming clusters for hybridisation (Arunachalam, 1981).

A graphical method of deriving group constellations using the canonical variates (canonical analysis) was discussed by Rao (1952).

Recent trends in cluster analysis

Clustering technique which is mainly a multivariate procedure has shown rapid improvement only after 1950's. A number of algorithms have been developed in order to classify a group of objects into smaller groups containing objects of similar character.

Eventhough in a field as diverse as cluster analysis the review and comparative assessment is cumbersome, Cormack (1971) and Orloci (1978) gave detailed accounts of various clustering procedures.

The first step in cluster analysis is the construction of a similarity or dissimilarity matrix between units to be classif-

ied [Mahalanobis et al. 1949 and Majumdar and Rao 1958]. However there is no adequate discussion in the literature on the choices of variables and measures of similarity (Rao, 1952).

The second step in cluster analysis is to build a 'rule' which connects units at various levels of similarity. A number of agglomerative and divisive methods have been developed for this purpose for overlapping and non-overlapping clusters. These various methodologies of cluster analysis attempt to sort a heterogeneous set of previously unpartitioned objects into groups that adequately reflect the original inter-object relationships (Atchely and Bryant, 1975).

Types of classification

A notable advance in the field of cluster analysis could be seen with the advent of electronic computers. It started gathering momentum due to early workers like Zubin (1938); Tyron (1939).

Cluster analysis can be classified as exclusive versus non exclusive according to the appearence of a given element in one or more groups (Williams,1976).

If the clusters are formed by progressive fusion or by progressive division, the procedure is termed hierarchical. In a non hierarchical classification clusters are obtained serially or simultaneously. A serial strategy is one in which a group is formed and removed before the formation of another begins or in the other case groups are formed simultaneously.

Hierarchical classification is further divided into agglomerative versus subdivisive. An agglomerative is one that proceeds by progressive fusion (Sokal and Michener, 1958) and divisive algorithm splits the population into specific number of groups [Edwards and Cavalli Sforza, 1965; Friedman and Rubin, 1967].

Clustering Algorithms

Cox (1957) and Fisher (1958) gave grouping criteria based on a single character (i.e.,univariate case). Cox (1957) considered the case when the variable is normally distributed and Fisher (1958)considered it without the distribution assumption.

Sokal and Michener (1958) evolved a weighted mean pair algorithm for clustering using the correlation between individuals as dissimilarity measures and applied it to an entomological problem. Sokal and Sneath (1963) recommended the above method as the best among the class of a commonly used methods of cluster analysis.

Williams and Lambert (1959) proposed a clustering procedure for data consisting of presence or absence of different traits.

Ward (1963) described an agglomerative sum of square algorithm. Many iterative algorithms which allow re-allocation of a single object at a time were proposed [Forgy 1965; Friedman

and Rubin 1967]. The initial partitions in these cases were randomly chosen, systematically chosen or the partition obtained from another algorithm.

Edwards and Cavalli-Sforza (1965) proposed a clustering technique based on the euclidean distance such that the sum of squares of distances between sets is maximum.

Gower (1967) in an excellent discussion compared the three methods of cluster analysis, those described by Sokal and Michener (1958); Edwards and Cavalli Sforza (1965) and Williams and Lambert (1959) and recommended Sokal and Michener's (1958) weighted mean pair algorithm for a general purpose of classification.

Scott and Symons (1971a,b) proposed an alternative to Edwards and Cavalli Sforza's (1965) algorithm which required the examination of $2^{N-1}-1$ partitions at a time. The algorithm suggested by Scott and Symons limits the consideration to $(2^{V}-2)\ ({}^{N}C_{V})$ partitions.

Friedman and Rubin (1967) proposed three criterion functions to be optimised for clustering which are basically related to Λ criterion developed by Wilks (1932). They were

i) Minimisation of trace W

ii) Maximisation of trace $W^{-1}B$ and

iii)Minimisation of /W/.

They recommended the minimisation of /W/ criterion to be selected in preference to others since it is invariant under non singular transformation. They also pointed out the advantage of using principal components to reduce dimensionality and singularity of /W/ when a large number of variables are considered for clustering. Different authors have discussed the algorithm developed by Friedman and Rubin (1967).

Scott and Symons (1971 b) reported that the minimum /W/ criterion had a tendency to form clusters of equal size which was also observed by Marriot (1971) and Everitt (1979). When there was no previous information about any of the populations, minimisation of /W/ was found better by Scott and Symons (1971b).

Marriot (1971) discussed the practical problems of cluster formation. According to him the main advantage of using /W/ criterion was that the variables which are highly correlated in the whole population are not given excessive weight.

An approximate method for deciding the number of groups was also suggested by Marriot (1971). He suggested that the value of g for which $g^2/W/m$ is minimum could be taken as the optimum number of clusters.

Maronna and Jacovkis (1974) in an interesting discussion of the matrices used in cluster analysis suggested minimisation of $[p\sum_{i=1}^{k}(n_i-1)/W_i/(1/n)]$ for clustering, where W_i is the internal scatter matrix of the ith cluster, ni the number of data points in it, k the total number of clusters and p is the number of

variables.

Symons (1981) proposed modifications for /W/ criterion of Friedman and Rubin (1967) to overcome the tendency to have clusters of equal size.

Everitt (1979,1980) provided a comprehensive overview of the practical problems common to users of cluster analysis.

A recent book by Gordon (1981) contains the recent trends of research work in cluster analysis.

Materials and Methods

MATERIALS AND METHODS

Prior to any attempt to form clusters of a number of genotypes based on a set of characters they are to be tested for homogeneity. In case they are homogeneous there is no necessity to form different groups, as they form a single group.

3.1 Analysis of Dispersion

For testing the homogeneity of a given set of genotypes with respect to a number of characters, multivariate analysis of variance (analysis of dispersion) should be carried out as follows.

Assume that there are '$\bar{v}$' genotypes each replicated 'r' times and 'B' is the matrix of corrected sum of squares and sum of products between genotypes and 'W' the matrix of within sum of squares and sum of products. The analysis of dispersion can be presented as follows.

Analysis of dispersion

s.v.	d.f.	s.s.p. matrix
Replications	(r-1)	R
Between genotypes	(v-1)	B
Within genotypes	(r-1)(v-1)	W
Total	(t-1)	$\mathfrak{C}$

Wilks Λ criterion (Wilks, 1932) could be adopted to test the significance of the differences among genotypes. The Λ is given by

$$\Lambda = \frac{/W/}{/W+B/} \qquad \ldots\ldots (3.1)$$

To test the significance of Λ, the approximate F - test suggested by Rao (1951,1973) could be adopted. The procedure is as follows.

$$F = \frac{ms - 2k}{pq} \cdot \frac{1 - \Lambda^{(1/s)}}{\Lambda^{(1/s)}} \qquad \ldots\ldots (3.2)$$

is variance ratio based on pq and ms - 2k d.f. where

$$p = \text{number of variables},$$

$$m = (t-1) - \frac{p+q+1}{2},$$

$$q = v-1 \quad , \quad k = \frac{pq-2}{4},$$

$$s = \frac{(p^2q^2 - 4)^{(1/2)}}{(p^2+q^2 - 5)^{(1/2)}}$$

Once the set of genotypes is found heterogeneous, one has to proceed with formation of different clusters such that those within any group are more alike compared to those belonging to any other.

Before considering the actual grouping, some measure of dissimilarity is to be defined between every pair of genotypes. The measure which is of wide use among plant breeders is

Mahalanobis D^2 statistic (Mahalanobis, 1936). A new measure of dissimilarity, that is, the determinant of the scatter matrix for any pair of genotypes is proposed in the present study.

3.2 Mahalanobis D^2 statistic

The D^2 statistic based on 'p' characteristics between any pair of genotypes was defined by Mahalanobis (1936) as

$$D_p^2 = cd'W^{-1}d \quad \ldots (3.3)$$ where

c = error d.f. , W = matrix of mean error sum of squares and sum of products and $d' = (\bar{X}_{11}-\bar{X}_{12}, \bar{X}_{21}-\bar{X}_{22}, \ldots\ldots, \bar{X}_{p1}-\bar{X}_{p2})$

A simplified, systematic and widely used procedure, as described by Rao (1952) could be made use of, for obtaining the D^2 values.

The first step in this procedure is to transform the original variables (x's) to a set of uncorrelated variables (y's) by the method of pivotal condensation.

Once the new set of uncorrelated variables (y's) are obtained D^2 for i^{th} and j^{th} genotype could be obtained as

$$D_{ij}^2 = \sum_{k=1}^{\phi} (y_{ik} - y_{jk})^2 \quad \ldots (3.4)$$

where y_{ik} is the value of the k^{th} variable for the i^{th} genotype.

3.3 Determinant of the scatter matrix

The determinant of the corrected sum of squares and sum of products of every pair of genotypes is suggested to be used as a measure of dissimilarity between genotypes.

When the number of data points for two genotypes is less than or equal to the number of characters being considered, this matrix often becomes singular. To overcome this difficulty the dimensionality could be reduced by considering a few principal components corresponding to the largest eigen values of the total scatter matrix of all characters taken together. The number of principal components must be less than the number of data points for any pair of genotypes.

3.4 Tocher's method

The method suggested by Tocher (Rao, 1948) is to start with those two genotypes having minimum value of D^2 and find a third genotype which has the smallest average D2 from the first two. The fourth genotype is chosen which has the smallest average D^2 from the first three and so on. If at any stage the increase in average D^2 for a genotype appears to be high compared to the previous one the current cluster is completed without this genotype. A new cluster is tried from the remaining genotypes in a similar way. The procedure is continued until all the genotypes are exhausted.

3.5 Modification over Tocher's method.

Once the grouping by Tocher's method is over an iterative re-location algoritham is suggested to improve the clustering so obtained.

i) Number the genotypes from 1 to V when there are V genotypes

ii) Take out genotype No.1 from the cluster to which it was alloted and calculate the average D2 values between this genotype and each cluster. Allocate this genotype into that cluster where the average D^2 value is found minimum

iii) Repeat (ii) for all the genotypes numbered from 1 to V

iv) With the clustering obtained in step (iii) a second iteration may be started, if necessary. i.e., repeat (ii) and (iii).

The iterations have to be continued till two successive iterations end up with the same configurations of clusters.

3.6 An iterative algorithm for formation of clusters using D^2 values

An iterative re-location algorithm for forming clusters using D2 values is proposed in this study as follows:

1) Identify the two genotypes having maximum D^2 value between them and they are termed as the nuclei of two clusters

ii) Every genotype is considered in turn and allocated to the cluster for which its D^2 value with the nucleus genotype is minimum

iii) To increase the number of clusters by one the maximum D^2 within the above two clusters is found and the genotypes having maximum D^2 will be considered as the nuclei in addition to the nucleus genotype of the remaining clusters. The genotypes may be reassigned as in (ii). In a similar way the number of clusters can be raised to a desired level.

The clustering thus obtained could be further optimised using the iterative re-location algorithm described in section 3.5

3.6.1 Determination of number of clusters

A problem that seeks solution in cluster analysis by mathematical programming is that of deciding on the number of clusters to be formed. A graphical method for determination of optimum number of clusters is suggested herein and is explained below.

A graph of weighted arithmetic mean of the average intracluster D2 values against the number of clusters may be drawn, the weights being the total number of D^2 values in the cluster. The graph will be a decreasing one. The rate of decrease will also be decreasing. The point on the X - axis which is just beyond the maximum curvature could be taken as the optimum number of clusters.

3.7 Method of canonical analysis

It is a graphical method widely used by plant breeders and taxonomists. The steps are as follows.

i) Obtain the eigen vectors corresponding to the largest two eigen roots of the between sum of squares and sum of products matrix of the transformed variables (y's) as in section 3.2

ii) Principal components corresponding to these vectors, say Z_1, Z_2 generated from the means of the transformed uncorrelated variables (y's) are obtained

iii) The Z_1 values are plotted against Z_2 values for getting a graphical representation of the genotypes.

iv) Group the genotypes represented by contiugous points by examining the graph.

3.8 Formation of clusters statistically

By formation of clusters statistically, we mean to form clusters which are maximum nonsignificant sets of genotypes. Every cluster should be such that any addition will make the genotypes in it to be significantly hetrogeneous. A procedure for forming such clusters is to evaluate the least value of D^2 to be significant as follows.

For testing the equlity of two populations with respect to p characters the statistic used is

$$F = \frac{N_1+N_2-p-1}{p} \cdot \frac{N_1 N_2}{N_1+N_2} \cdot \frac{D_p^2}{N_1+N_2-2} \quad \ldots (3.5)$$

If the calculated value of F is greater than F_{α}, the critical value of F_{p,N_1+N_2-p-1} at the α level of significance the two genotypes differ significantly. In other words if

$$D_p^2 > \frac{N_1+N_2-2}{N_1 N_2} \cdot \frac{N_1+N_2}{N_1+N_2-p-1} \cdot p \cdot F \quad \ldots (3.6)$$

the genotypes differ significantly. The R.H.S of the inequality 3.6 is termed as the critical value of D^2.

If the D^2 value between two genotypes is greater than the critical value, the two genotypes could be considered as significantly different and otherwise not. By this method we get overlapping clusters as in the case of comparison of treatments using critical difference after analysis of variance (single variable situation).

3.9 Minimum /W/ clustering

3.9.1 An optimisation technique for clustering

An iterative re-location algoritham suggested by Friedman and Rubin (1967) is proposed to form clusters of genotypes by minimising the determinant of the within cluster scatter matrix.

For a given number of clusters, the iterative procedure starting with some initial solution is as follows.

i) Number the genotypes from 1 to V

ii) Take out genotype No.1 and calculate /W/, the determinant of the within cluster scatter matrix by allocating it to the different clusters in turn and finally allocate to the cluster for which /W/ is minimum

iii) Repeat (ii) for genotype No.2, No.3, ... up to No. V

iv) With the clustering obtained in step (iii) a second iteration may be started, if necessary. i.e., repeat steps (ii) and (iii).

The iteration has to be continued till two successive iterations end up with the same configurations of clusters.

3.9.2 Formation of Initial clusters

a). The clusters obtained in section 3.6 could be used as the initial clusters for optimisation.

b). A procedure which is exactly the same in 3.6 could be adopted for forming initial clusters with the determinant of the pairwise scatter matrix as the measure of dissimilarity in the place of D2 value.

3.9.3 Determination of the number of clusters

A graphical procedure which is exactly same as in 3.6 using minimum /W/ instead of weighted arithmetic mean could be adopted for deciding the number of clusters to be formed

A method suggested by Marriot (1971) was also tried to determine the number of clusters to be formed. The method is to select that value of 'g' for which $g^2/w/_m$ is minimum, where g is the number of clusters and $/w/_m$ is the minimum of /w/ obtained for g.

3.10 Illustration

Two sets of secondary data were used to illustrate the methods described. First set of data was taken from Rajeevan (1985) on 24 accessions of Musa (AAB) group. Sixteen characters which showed significant differences among the accessions were selected.

3.9.2 Formation of Initial clusters

a). The clusters obtained in section 3.6 could be used as the initial clusters for optimisation.

b). A procedure which is exactly the same in 3.6 could be adopted for forming initial clusters with the determinant of the pairwise scatter matrix as the measure of dissimilarity in the place of D2 value.

3.9.3 Determination of the number of clusters

A graphical procedure which is exactly same as in 3.6 using minimum /W/ instead of weighted arithmetic mean could be adopted for deciding the number of clusters to be formed

A method suggested by Marriot (1971) was also tried to determine the number of clusters to be formed. The method is to select that value of 'g' for which $g^2/w/_m$ is minimum, where g is the number of clusters and $/w/_m$ is the minimum of /w/ obtained for g.

3.10 Illustration

Two sets of secondary data were used to illustrate the methods described. First set of data was taken from Rajeevan (1985) on 24 accessions of Musa (AAB) group. Sixteen characters which showed significant differences among the accessions were selected.

The second set of data has been taken from Singh and Choudhary (1979) on 8 varieties of barley. Observations on four characters were used.

All the analyses were carried out in HCL workhorse, level 2 computer available in the Department of Statistics, College of Horticulture, Vellanikkara.

Results and Discussion

RESULTS AND DISCUSSION

Results obtained through the application of various procedures described in chapter 3 are presented below. The merits and demerits of the different methods are also discussed.

Analysis of dispersion for both sets of data was performed to examine the homogeneity of the genotypes. In both the cases the Λ was found to be significant and is presented in Table 4.1. The genotypes were found heterogeneous.

Table 4.1. Wilk's Λ criterion

Data Set	$\lvert W \rvert$	$\lvert W+B \rvert = \lvert T \rvert$	$\Lambda = \frac{\lvert W \rvert}{\lvert T \rvert}$	F	d.f.(Nr.) pq	d.f.(Dr.) ms-2k
I	0.1095×10^{30}	0.1828×10^{34}	5.9927×10^{-5}	1.4503	368	488
II	0.8890×10^{5}	0.8550×10^{8}	0.0010	15.759	28	77

4.1 Measures of dissimilarity

a) D^2 values.

D^2 values for every pair of genotypes were found out for both sets of data and are presented in appendix B.

b) Determinant of the pairwise scatter matrix.

The number of data points (6) for any pair of genotypes was less than the number of characters (16) in the case of data set I. Hence the dimensionality was reduced by taking 5 principal components. The determinants of the pairwise scatter matrices are given in appendix C.

The determinant of the pairwise scatter matrix was proposed as a measure of dissimilarity for forming initial clusters for Friedman and Rubin's (1967) algorithm because its resemblence to /W/.

4.2. Tocher's method

Tocher in his method of clustering suggested that, addition of genotypes to a cluster might be stopped when a 'sudden increase' in average D2 value was exhibited. This 'sudden increase' is subjective to a great extent. Singh and Choudhary (1979) pointed out that this increase could become to the extent of maximum among the minimum D^2 values attached to every genotype. There is no sound basis in fixing such an arbitrary value for deciding a cluster. On the other hand, if there is a genotype which is far distant from the rest, all the genotypes except the distant one will fall into a single cluster.

Even though the cluster formation by Tocher's method is widely used by plant breeders and taxonomists, the clusters

formed by this method often suffer from the defect that the distance of a genotype from another one within the same cluster may be much larger than that from one in another cluster. Often a genotype included in a cluster by this method has smaller average D2 with those in a different cluster than the one to which it belongs to.

4.3 An improvement on clustering by Tocher's method

Due to these various disadvantages of Tocher's method, a refinement over it as described in 3.5 is proposed in this study. Clusters formed by Tocher's method and modified Tocher's method for both sets of data are presented in Table 4.2.

Table 4.2 Clusters formed by Tocher's method and by improved clustering

Data set	Sl.No. of clusters	Sl. No. of the genotypes in the cluster by	
		Tocher's method	Improved clustering
(1)	(2)	(3)	(4)
I	1	3 6 9 11 12 15 16 19 20 21 22 23 24	6 11 12 15 20 21 22 23 24
	2	1 5	1 3 5 9 16 19
	3	2 8 10 13 14 17 18	2 8 10 13 14 17 18
	4	7	7
	5	4	4
II	1	4 7	4 6 7
	2	1 5 6 8	5 8
	3	2	1 2
	4	3	3

The intra- and inter-cluster D^2 values are presented in Tables 4.3 and 4.4

Table 4.3 Average intra & inter cluster D^2 values for data I

Tocher's method				
1134.88	4250.59	9668.11	15698.20	69457.33
	41.72	24274.31	3892.91	40684.86
		721.33	47333.26	127363.30
			0.00	19548.81
				0.00
Improved clustering				
597.59	3073.51	7335.03	18422.71	75481.24
	773.04	18036.47	7676.35	50830.64
		721.33	47333.26	127363.30
			0.00	19548.81
				0.00

Table 4.4 Average intra & inter cluster D^2 values for data II

Tocher's method			
42.47	27.76	48.79	24.15
	23.78	58.80	56.88
		0.00	104.81
			0.00
Improved clustering Technique			
15.08	32.37	39.73	28.47
	15.62	47.67	61.19
		17.70	84.83
			0.00

In the first set of data, the improved clustering re-assigned genotypes 3,9,16, and 19 to cluster II from cluster I. The question is whether such a reassignment is appreciable. For verifying the effectiveness of the improved clustering technique the weighted arithmetic mean of the average intra cluster D^2 values which is nothing but the simple arithmetic mean of all the intra cluster D^2 values was calculated, the weights being the total number of D^2 values in a cluster. They are given in Table 4.5.

Table 4.5 Weighted arithmetic mean of Intra cluster D^2

Data set	Tocher's method	Modified Tocher's method
I	1036.41	670.23
II	20.99	15.71

From the table, it may be noted that the weighted arithmetic mean of intra-cluster D^2 values decreased considerably in both cases. This establishes the superiority of the modified method over the Tocher's method.

Since D^2 value is a measure of variability between two genotypes, the arithmetic mean of D^2 values between every pair of genotypes coming together in a cluster is essentially a measure of within cluster variability. The basic principle to be

GRAPH OF WEIGHTED ARITHMETIC MEAN OF AVERAGE INTRACLUSTER D^2 VALUES AGAINST NUMBER OF CLUSTERS.

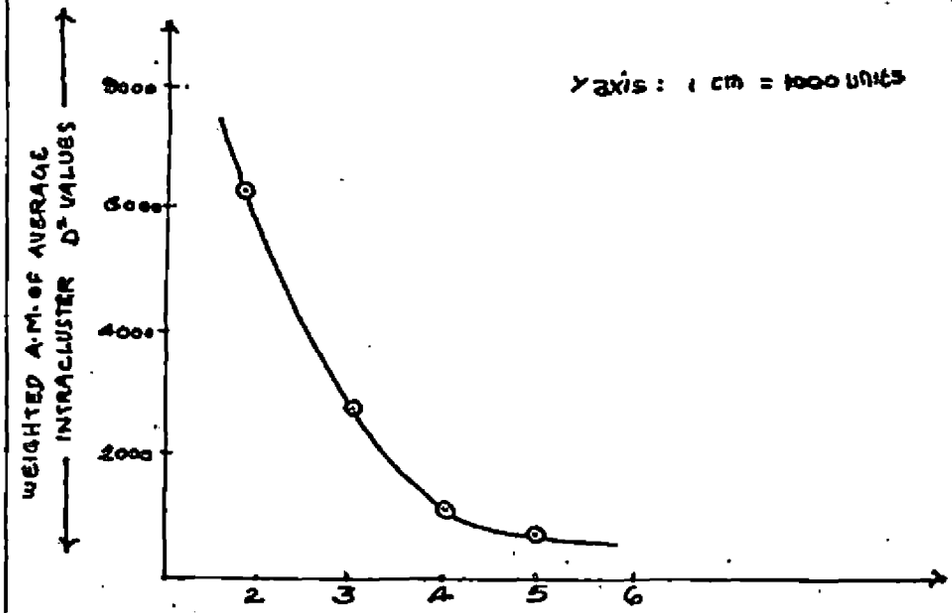


FIG. 4.1 (Dataset I)

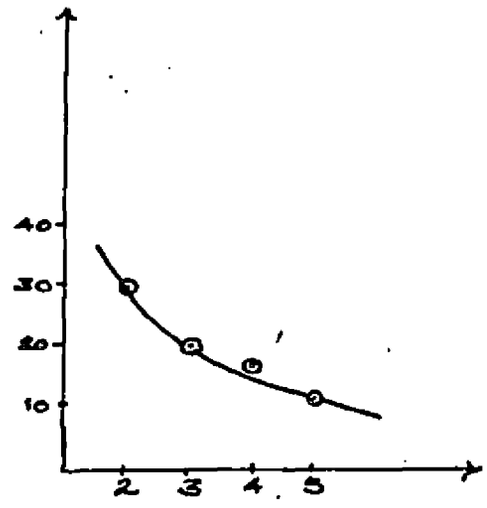


FIG. 4.2 (Dataset II)

followed in the formation of clusters must be that the within cluster variability must be minimum and the between clusters variability the maximum. Hence the average intra-cluster D^2 value is a very logical statistic for comparison of the efficiencies of different clustering making use of D^2 values.

4.4 Formation of clusters by the iterative algorithm using D^2 values

Clusters were formed using the iterative algorithm described in 3.6. For deciding the number of clusters to be formed a graphical method as in 3.6.1 (see Fig 4.1 and 4.2) were adopted and the optimum number of clusters were found as four in both the cases of data sets I and II. The clusters obtained are given in Table 4.6 and Table 4.7

The algorithm suggested here may be used in preference to that by Tocher's method. The disadvantages of Tocher's method mentioned in 4.2 are rectified in the present case.

Table 4.6 Clusters obtained by the iterative algorithm using D^2 from data set I

	Sl.No. of clusters	Sl.No. of genotypes	Weighted A.N.of intra cluster D^2	No.of iterations
		Two clusters		
Initial	1	1 2 3 5 6 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24	6264.71	
	2	4 7		
Final	1	1 2 3 5 6 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24	6264.71	1
	2	4 7		
		Three clusters		
Initial	1	4	2774.90	
	2	1 3 5 6 7 9 11 12 16 19 20 21 23 24		
	3	2 8 10 13 14 15 17 18 22		
Final	1	4	2774.90	1
	2	1 3 5 6 7 9 11 12 16 19 20 21 23 24		
	3	2 8 10 13 14 15 17 18 22		
		Four clusters		
Initial	1	4	1093.10	
	2	1 5 7		
	3	3 6 9 11 12 15 16 19 20 21 22 23 24		
	4	2 8 10 13 14 17 18		
Final	1	4	1093.10	1
	2	1 5 7		
	3	3 6 9 11 12 15 16 19 20 21 22 23 24		
	4	2 8 10 13 14 17 18		
		Five clusters		
Initial	1	4	837.43	
	2	7		
	3	2 6 11 12 15 20 21 22		
	4	1 3 5 9 16 19 23 24		
	5	8 10 13 14 17 18		
Final	1	4	670.23	2
	2	7		
	3	6 11 12 15 20 21 22 23 24		
	4	1 3 5 9 16 19		
	5	2 8 10 13 14 17 18		

Table 4.7 Clusters obtained by the iterative algorithm using D^2 values for data II

	Sl.No. of clusters	Sl.No. of genotypes	Weighted A.M.of intra cluster D^2	No.of iterations
		Two clusters		
Initial	1	1 2		
	2	3 4 5 6 7 8	30.04	
Final	1	1 2		
	2	3 4 5 6 7 8	30.04	1
		Three clusters		
Initial	1	1 2		
	2	3 4	21.52	
	3	5 6 7 8		
Final	1	1 2		
	2	3 4 7	19.69	2
	3	5 6 8		
		Four clusters		
Initial	1	1 2		
	2	3 4		
	3	5	17.60	
	4	6 7 8		
Final	1	1 2		
	2	3 4 7		
	3	5	16.49	3
	4	6 8		
		Five clusters		
Initial	1	1 2		
	2	3		
	3	4 7	11.37	1
	4	5		
	5	6 8		
Final	1	1 2		
	2	3		
	3	4 7	11.37	1
	4	5		
	5	6 8		

4.5. Method of canonical analysis

The two canonical vectors corresponding to the first two eigen values were found out as described in 3.7. and are given in appendix D. A graphical representation of the genotypes based on the canonical variates for data sets I and II are given in Fig. 4.3 and 4.4 respectively.

After a careful examination of the graph, five clusters were formed in both cases and the configuration of the clusters are given in Table 4.8.

Table 4.8 Clusters obtained by the method of canonical analysis

Sl. No. of clusters	Sl. No. of genotypes in the cluster	
	data I	data II
1	1 3 5 6 9 11 12 15 16 19 20 21 22 23 24	4 7 6
2	2 8 10 13 14 17 18	5 8
3	15	1
4	4	2
5	7	3

Since the two principal components representing the

Fig. 4·3

REPRESENTATION OF THE 24 GENOTYPES ON THE BASIS OF CANONICAL ANALYSIS.

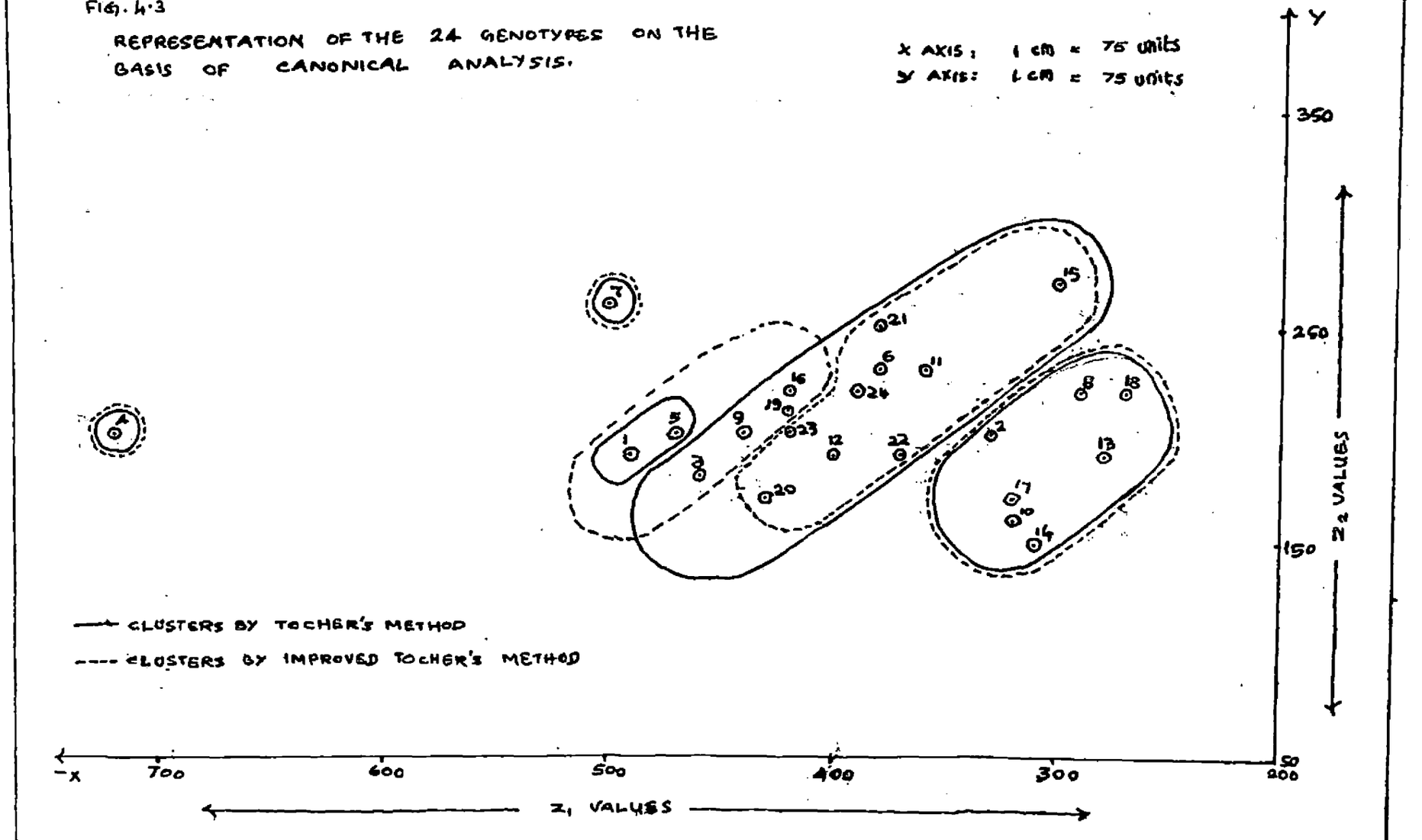


Fig. 4·3

REPRESENTATION OF THE 24 GENOTYPES ON THE BASIS OF CANONICAL ANALYSIS.

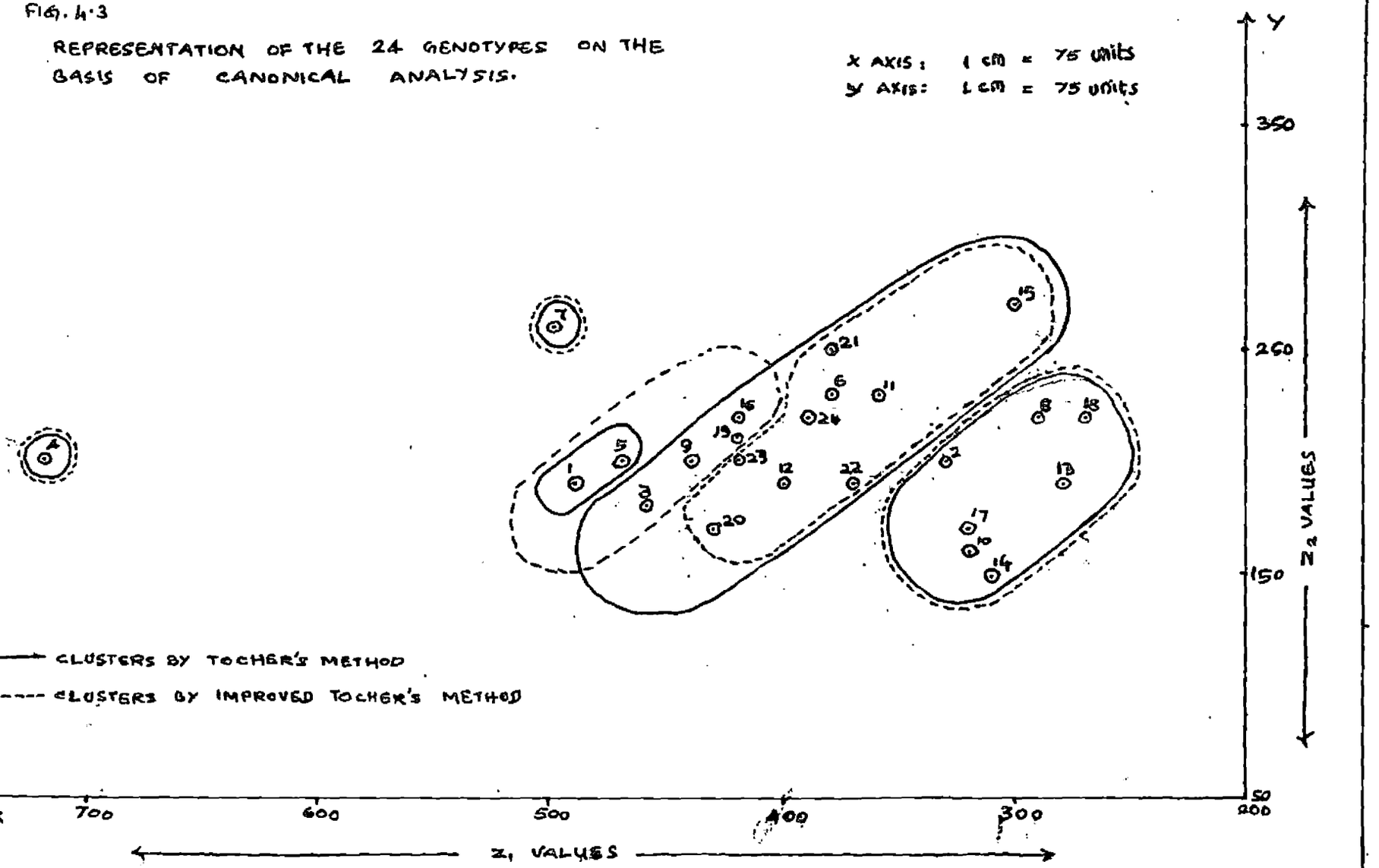
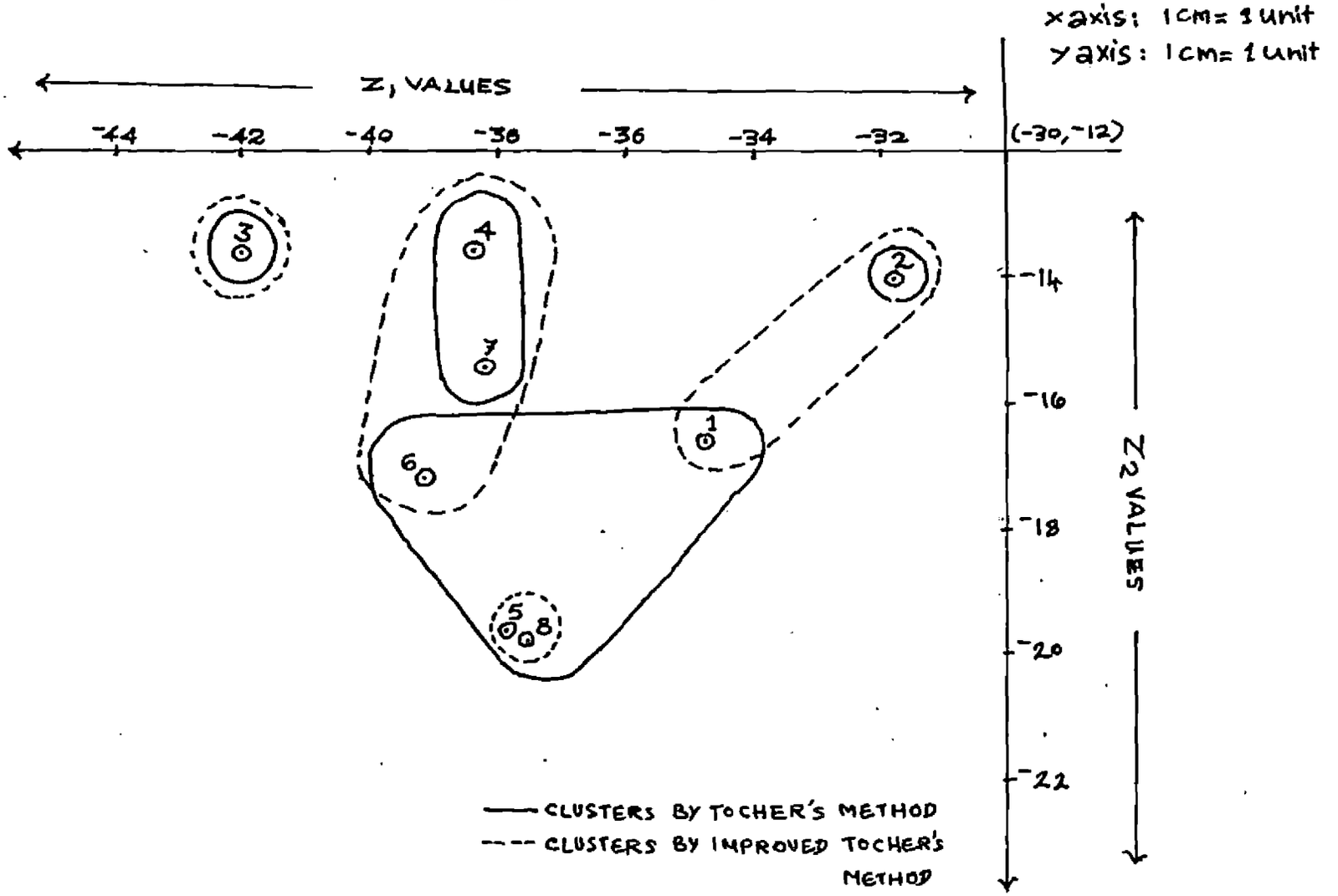


FIG. 4.4 REPRESENTATION OF THE 8 GENOTYPES ON THE BASIS OF CANONICAL ANALYSIS.



genotypes in the graph explained more than 99 percent of the total variation in the first case and 80 percent variation in the case of second set of data, the clustering obtained by the method of canonical analysis can be relied upon for all practical purposes. A comparison of the cluster configurations obtained by this method and the two methods using the D^2 values would be worthwhile. Though there is no complete agreement between the clusterings by the different methods, that by the canonical analysis is more in agreement with the proposed improved method than Tocher's. In otherwords the improved clustering procedure proposed in 3.5 was found more efficient than the Tocher's method in both sets of data.

It may be noted that in both cases, distances between genotypes in the graph are not in conformity with the corresponding D2 values. For example the genotype which is nearer to 23 with respect to D^2 value is 24, while in the graph it is 19 in the data set I. This sort of situations arise in many cases. Perhaps this could be due to the the effects of departures from the assumptions of multivariate normal distribution and common dispersion matrix, on D2 values and the principal components.

4.6. Formation of clusters statistically

A drawback noted in the procedures discussed is that they do not have any statistical meaning. Though the homogeneity of the genotypes is tested prior to any attempt to form clusters, the homogeneity of the genotypes that are classified into a cluster are usually not tested. Often the genotypes grouped into a cluster differ significantly which could be verified from Appendix B. Hence the procedure in 3.8 was suggested in the study.

The critical values for D_p^2 using equation (3.6) were obtained as

D^2 = 31.33 for the first set of data

= 6.84 for the second set of data

A careful examination of D^2 values of the first set of data in appendix B reveals that only one pair of genotypes viz., 23 and 24 can be grouped together as homogeneous. In other words there are 22 clusters each having a single genotype except one which includes 23 and 24. Similarly in the case of second set of data genotypes 4 and 7 do not differ significantly and could be grouped into a single cluster leaving every other genotypes to form clusters having only one genotype.

There would be situations when we may have a set of overlapping clusters just like the overlapping groups of treatments arrived at using critical diference or multiple range

test after analysis of variance. This should not worry us in adopting the procedure as it is the pattern of natural variation. Hence the procedure could be successfully used in such situations.

In situations as in the two examples we have considered where there is no effective clustering, this procedure may not be admissible. It will be quite useful in cases where this procedure results in the formation of clusters effective to some extent.

4.7. An optimisation technique for clustering

The methods using D^2 and canonical analysis, require the following assumptions to be satisfied by the observations.

i) The variables should follow multivariate normal distribution

ii) The dispersion matrices are homogeneous for the different genotypes.

These assumptions might be violated very often particularly when there are a large number of genotypes to be clustered. Hence a procedure that does not make use of such assumptions would be worthwhile. So the procedure in 3.9 was proposed.

Friedman and Rubin (1967) suggested three criterion functions to be optimised for getting a clustering. They were

i) Minimisation of trace W

ii) Maximisation of trace $W^{-1}B$ and

iii) Minimisation of /W/.

Of these, the minimum /W/ criteria was selected to be the criterion functions for our purpose because of its close relationship with Wilk's Λ .

Wilk's Λ (Wilks, 1932) is the ratio of /W/, the determinant of the within scatter matrix to /W+B/, the determinant of the total scatter (Between + Within) matrix, where 'B' denotes the sum of square and sum of product matrix between clusters. Since /W+B/ remains same for any clustering, minimisation of /W/ amounts to minimisation of Λ . Smaller values of Λ will be the critical region when it is used as a test criterion. In other words the more distant are the groups, the smaller will be the values of Λ. Hence /W/ was chosen as the objective function to be minimised to arrive at the best clustering.

The solution to the above mentioned programming problem is not that straight forward. The iterative procedure [3.9.1] suggested by Friedman and Rubin (1967) leads us, starting from some initial solution, to a local optimum solution. There is no feasible procedure to arrive at a global optimum solution. Hence, perhaps, the alternative is to use an iterative procedure (as described herein) to arrive at a local minimum solution starting from different initial solutions and choose the solution which is optimum among the local optima.

Initial clusters were formed as described in 3.9.2 using determinants of pairwise scatter matrices as well as the D^2 values for number of clusters ranging from 2 to 6 for data set I and 2 to 5 for data set II. Iterative solutions starting from the two different initial solutions were obtained for both data sets and for varying numbers of clusters. Initial and final solutions along with the corresponding /W/ values and $g^2/W/_m$ values are provided in Tables 4.9. to 4.12.

Table 4.9. Initial and final solutions using determinants of pairwise scatter matrices for data I

	Sl.No. of clusters	Sl. No. of genotypes in the cluster	mini./W/ obtained X 10^{-32}	$g^2/W/_m$ X 10^{-32}	No. of Iterations
	(1)	(2)	(3)	(4)	(5)
		Two clusters			
Initial	1	1 2 3 4 5 6 7 8 9 10 11 12			
	2	13 14 15 16 17 18 19 20 21 22 23 24	14.1761		
Final	1	2 3 5 8 10 12 13 16 17 20 22 23	6.9898	27.9592	3
	2	1 4 6 7 9 11 14 15 18 19 21 24			
		Three clusters			
Initial	1	1 2 3 4 5 6 7 8 9 10 11 12			
	2	13 18 21 22	8.7103		
	3	14 15 16 17 19 20 23 24			
Final	1	1 2 3 4 5 6 9 13 16 19 20 23 24	2.4135	21.7213	3
	2	7 10 12 14 17			
	3	8 11 15 18 21 22			

Table 4.9. Contd.....

(1)		(2)	(3)	(4)	(5)
		Four clusters			
	1	1 2 3 4 5 6 7 9 10			
Initial	2	13 18 21 22	6.3382		
	3	8 11 12 17 19			
	4	14 15 16 20 23 24			
	1	7 11 15 18 21			
	2	8 10 12 17 22			
Final	3	2 3 5 13 16 20	0.8920	14.2728	3
	4	1 4 6 9 14 19 23 24			
		Five clusters			
	1	2 3 4 5 6 7 8 9 11			
	2	13 18 21			
Initial	3	12 17 19 22	3.5585		
	4	14 15 16 20 23 24			
	5	1 10			
	1	1 4 6 9 14 19 23 24			
	2	7 11 15 18			
Final	3	8 10 12 17 22	0.3809	9.5224	3
	4	2 3 5 13 16 20			
	5	21			
		six clusters			
	1	2 3 4 5 6 7 8 9 11			
	2	15 18 21 22			
Initial	3	12 17 19	2.5725		
	4	1 10			
	5	13			
	6	14 16 20 23 24			
	1	1 2 3 4 5 6 9 19 20 23 24			
	2	7 11 15 18			
Final	3	10 12 14 17	0.2590	9.3246	5
	4	13 16			
	5	21			
	6	8 22			

Table 4.9. contd.....

	(1)	(2)	(3)	(4)	(5)
		Seven clusters			
	1	2 3 5 6 8 9 10			
	2	1 16 17 19 20			
	3	15 18 21 22			
Initial	4	13	1.6639		
	5	14 23 24			
	6	4 7			
	7	11 12			
	1	2 3 5 20			
	2	1 4 24			
	3	7 11 15 18	0.1657	8.1220	4
Final	4	13 16			
	5	6 9 14 19 23			
	6	8 10 12 17 22			
	7	21			
		Eight clusters			
	1	2 3 5 6 8 9 10			
	2	24			
	3	4 7			
Initial	4	13	1.0631		
	5	14 16 17 20 22 23			
	6	15 18 21			
	7	11 12			
	8	1 19			
	1	2 3 5 20			
	2	1 7 24			
	3	4 6 9 14 19 23	0.0921	5.8979	4
	4	13 16			
Final	5	10 12 17			
	6	11 15 18			
	7	8 22			
	8	21			

Table 4.9. concl.

Table 4.10. Initial and final solutions from D2 values for data I

	Sl.No. of clusters (1)	Sl. No. of genotypes in the cluster (2)	mini./W/ obtained X 10^{-32} (3)	$g^2/W/_m$ X 10^{-32} (4)	No. of Iterations (5)
		Two clusters			
Initial	1	1 2 3 5 6 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24	12.7710		
	2	4 7			
Final	1	1 4 7 8 11 14 15 18 19 21 22	6.7550	27.0199	4
	2	2 3 5 6 9 10 12 13 16 17 20 23 24			
		Three clusters			
Initial	1	4	7.8700		
	2	2 8 10 13 14 15 17 18 22			
	3	1 3 5 6 7 9 11 12 16 1920 21 23 24			
Final	1	7 11 15 18 21	1.9063	17.1570	4
	2	8 10 12 14 17 22			
	3	1 2 3 4 5 6 9 13 16 19 20 23 24			
		Four clusters			
Initial	1	4	6.6955		
	2	2 8 10 13 14 17 18			
	3	1 5 7			
	4	3 6 9 11 12 15 16 19 20 21 22 23 24			
Final	1	1 7 11 14 18	0.9267	14.8270	4
	2	8 10 12 17 22			
	3	2 3 4 5 6 9 13 15 16 19 20 23 24			
	4	21			

Table 4.10. contd.....

(1)	(2)	(3)	(4)	(5)
	Five clusters			
	1 4			
	2 2 8 10 13 14 17 18			
Initial	3 1 3 5 9 16 19	3.7141		
	4 6 11 12 15 20 21 22 23 24			
	5 7			
	1 1 6 7 14 19 24			
	2 8 10 12 17 22			
Final	3 11 15 18 21	0.5917	14.7922	4
	4 2 3 4 5 9 20 23			
	5 13 16			
	Six clusters			
	1 4			
	2 2 8 10 13 14 17 18			
	3 3 9 16 19 23 24			
Initial	4 6 11 12 15 20 21 22	2.5588		
	5 1 5			
	6 7			
	1 1 4 24			
	2 8 10 12 17 22			
	3 2 3 5 13 16 20			
Final	4 7 11 15 18	0.2513	9.0484	4
	5 6 9 14 19 23			
	6 21			
	Seven clusters			
	1 4			
	2 2 8			
	3 1 5			
Initial	4 3 9 16 19 23 24	1.6622		
	5 6 11 12 15 20 21 22			
	6 10 13 14 17 18			
	7 7			
	1 1 4 24			
	2 8 10 12 17 22			
	3 2 3 5 13 16 20			
Final	4 11 15 18	0.1662	8.1453	4
	5 7			
	6 6 9 14 19 23			
	7 21			

Table 4.10. contd....

(1)		(2)	(3)	(4)	(5)
		Eight clusters			
	1	4			
	2	2 8			
	3	1 5			
	4	6 11 12 20 21			
Initial	5	3 9 16 19 23 24	0.9598		
	6	10 13 14 17 18			
	7	15 22			
	8	7			
	1	1 4 24			
	2	10 12 17			
	3	2 3 5 13 16 20			
	4	11 15 18			
Final	5	8 22	0.0937	6.0004	3
	6	4 6 9 19 23			
	7	7			
	8	21			

Table 4.10. concl.

Table 4.11. Initial and final solutions using determinants of pairwise scatter matrices for data II

Sl.No. of clusters (1)		Sl. No. of genotypes in the cluster (2)	mini./W/ obtained X 10^{-7} (3)	$g^2/W/_m$ X 10^{-7} (4)	No. of Iterations (5)
		Two clusters			
Initial	1	1 2 5 8	1.5547		
	2	3 4 6 7			
Final	1	1 2 5 8			
	2	3 4 6 7	1.5547	6.2187	1
		Three clusters			
	1	1 2			
Initial	2	4 6 7	0.8751		
	3	3 5 8			
	1	1 2			
Final	2	3 4 6 7			
	3	5 8	0.4845	4.3609	3

Table 4.11. contd....

(1)		(2)	(3)	(4)	(5)
		Four clusters			
Initial	1	1 2	0.2249		
	2	3 4 7			
	3	5 8			
	4	6			
Final	1	1 2	0.1234	1.9753	2
	2	4 6 7			
	3	5 8			
	4	3			
		Five clusters			
Initial	1	1 2	0.0604		
	2	4 7			
	3	5 8			
	4	6			
	5	3			
Final	1	1 2	0.0604		1
	2	4 7			
	3	5 8			
	4	6			
	5	3			

Table 4.11. concl.

Table 4.12. Initial and final solutions from D^2 values obtained for data II

Sl.No. of clusters (1)	Sl. No. of genotypes in the cluster (2)	mini./W/ obtained X 10^{-7} (3)	$g^2/W/_m$ X 10^{-7} (4)	No. of Iterations (5)
	Two clusters			
Initial 1	1 2	2.2330		
2	3 4 5 6 7 8			
Final 1	1 2	2.2330	8.9321	1
2	3 4 5 6 7 8			

Table 4.12. contd....

(1)		(2)	(3)	(4)	(5)
		Three clusters			
	1	1 2			
Initial	2	3 4 7	0.5554		
	3	5 6 8			
	1	1 2			
Final	2	5 8	0.4845	4.3609	2
	3	3 4 6 7			
		Four clusters			
	1	1 2			
	2	3 4 7			
Initial	3	5	0.2118		
	4	6 8			
	1	1 2			
	2	3 4 7			
Final	3	5	0.2118	3.3890	1
	4	6 8			
		Five clusters			
	1	1 2			
	2	4 7			
Initial	3	6 8	0.0609		
	4	3			
	5	5			
	1	2			
	2	1 4 7			
Final	3	6 8	0.0584	1.4610	2
	4	3			
	5	5			

Table 4.12. concl.

GRAPH OF |W| AGAINST NUMBER OF CLUSTERS

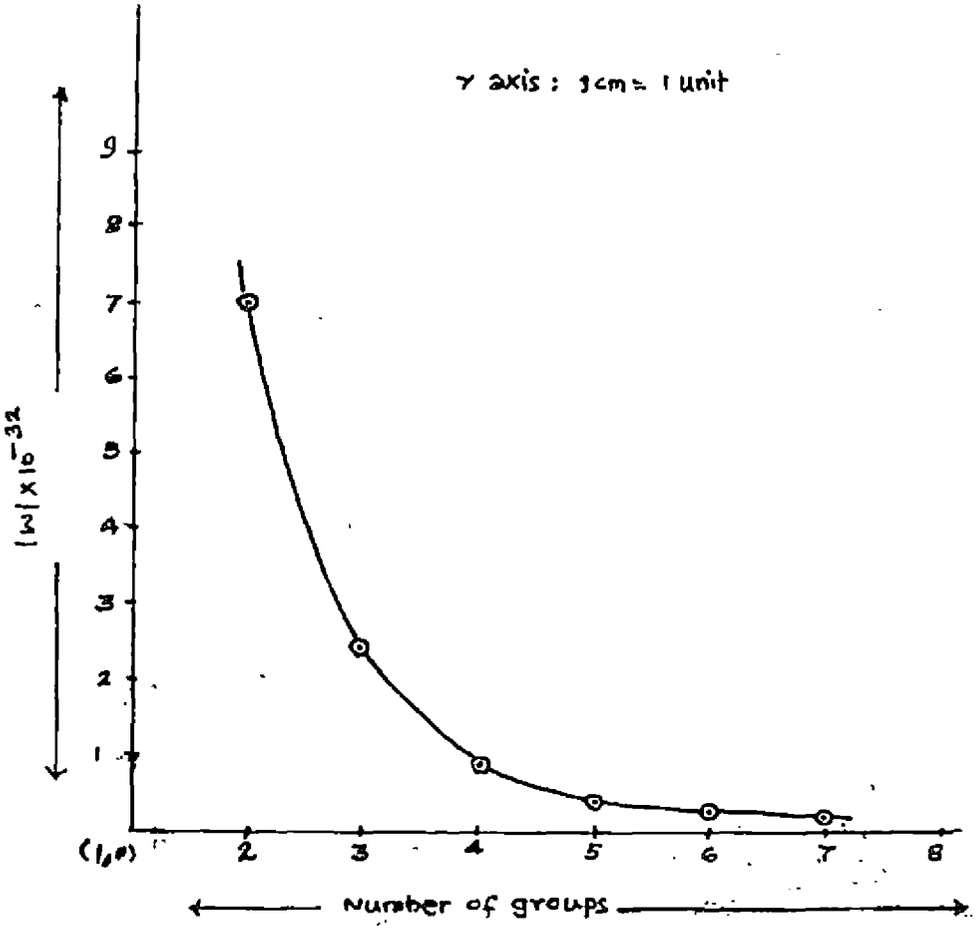


Fig. 4.5 (Data set I)

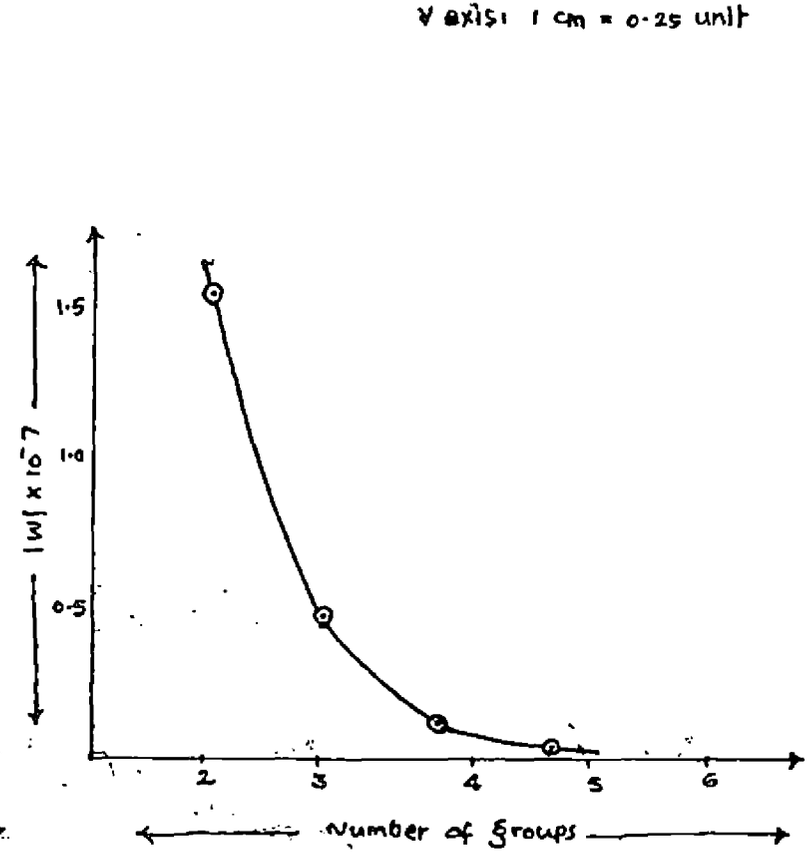


Fig. 4.6 (Data set II)

A close examination of Tables 4.9 and 4.10. reveals that initial solutions arrived neither from D^2 values nor determinants of pairwise scatter matrices can be said to be better than the others, as a general rule.

The cluster configurations obtained by minimising /W/ in the case of two data sets did not show any tendency to have clusters of equal number of genotypes in clusters as pointed out by Scott and Symons (1971 b), Marriot (1971) and Everitt (1979)

A graph of minimum /W/ against g the number of clusters was drawn for each set of data and are given in figures 4.5. and 4.6 Locating the point just beyond the maximum curvature the number of clusters to which the genotypes could be partitioned was determined as four in both sets of data.

The criterion suggested by Marriot (1971) was also tried to find the optimum number of clusters in both cases. The value of $g^2/W/_m$ was found to be decreasing even when the number of clusters equals 8 in the case of first set of data and 5 in the case of data set II, suggesting the non suitability of the technique.

Summary

SUMMARY

Tocher's method of clustering genotypes is very widely used by plant breeders. The following two major drawbacks of this method were pointed out in this study.

i) Stopping rule for formation of any cluster is arbitrary.
ii) Often a genotype belonging to a cluster have on an average, a smaller D^2 value with genotypes of a different cluster than the one to which it belongs to.

A modification of the cluster configuration arrived at by Tocher's method which is an iterative re-location algorithm, that finally re-allocates each genotype to that cluster for which its average D^2 value is least was suggested.

The clusterings obtained by the above two methods were compared with those obtained by canonical analysis method. The modified method was found more in agreement with canonical analysis method.

A new method of clustering using Mahalanobis D^2 values

A new computer oriented iterative algorithm for formation of clusters which does not have the drawbacks mentioned for Tocher's method was suggested as follows:

i) Identify the two genotypes having maximum D^2 value between them and they are termed as the nuclei of two clusters.

ii) Every genotype is considered in turn and allocated to the cluster for which its D^2 value with the nucleus genotype is minimum

iii) To increase the number of clusters by one the maximum D^2 within the above two clusters is found and the genotypes having maximum D^2 is considered as the nuclei in addition to the nucleus genotype of the remaining clusters. The genotypes are re-assigned as in (ii). In a similar way the number of clusters can be raised to a desired level.

The clustering thus obtained may further be optimised using the iterative algorithm as in the modified Tochers method. To decide the number of clusters which reveals the natural pattern of grouping, a graph is drawn with weighted arithmetic mean of average intra cluster D^2 values against the number of clusters. The point just beyond the maximum curvature was taken as the optimum number of clusters to be formed.

Formation of clusters statistically.

The critical value of D^2 was defined as that value beyond which the genotypes attached could be considered significantly different.

A procedure for formation of clusters statistically using the critical value of D^2 was proposed. This is to form maximum nonsignificant subsets of genotypes and the clusters obtained may or may not be overlapping.

A new measure of dissimilarity.

A new measure of dissimilarity, viz., the determinant of pairwise scatter matrix was proposed here in. This does not require any assumption on distribution of the population.

Minimum /W/ criterion for clustering

The iterative algorithm of Friedman and Rubin (1967) is recommended to get the clustering by minimising /W/, the determinant of the within cluster sum of squares and sum of products matrix.

The clustering arrived at using the new iterative procedure for clustering using the determinant of the pairwise scatter matrix and that obtained by D^2 values were considerd as the initial solution for optimisation. It was observed that none of these initial solutions can be considered in prefernce to the other.

The graphical method suggested for the new iterative procedure for clustering using Mahalanobis D^2 with /W/ instead of average intra cluster D^2 values, was used and recommended to determine the number of clusters.

The different methods were illustrated in two sets of data.

References

REFERENCES

Arunachalam, V. (1981). Genetic distance in plant breeding. Indian J. Genet. 41(2) : 226-236

Atchley, W.R. and Bryant, E.H. (1975) Multivariate statistical methods. Vol.I. Dowden, Hutchinson and Ross, Inc. Pennsylvania.

*Bartlett, M.S. (1947). Multivariate analysis, J. Roy. Stat. Soc. Suppl. 9:76

Blackith, R.E. (1957). Polymorphism in some australian locusts and grasshoppers. Biometrics. 13 : 183

Cormack, R.M. (1971). A review of classification. J. Roy. stat. soc.(A). 134(3) : 321-367

Cox, D.R. (1957) Note on grouping. J. Am. Statist. Assoc. 52 : 543-547

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1965). A method for cluster analysis. Biometrics. 21 : 362-375

Everitt, B.S. (1979). Unresolved problems in cluster analysis. Biometrics. 35 : 169-181

Everitt, B.S. (1980) Cluster analysis. Heinemann Educational Books, London, 2nd Ed.

Fisher, W.D. (1958). On grouping for maximum homogeneity. J. Am. statist. Assoc. 53 : 789-798

Forgy, E.W. (1965). Cluster analysis of multivariate data:efficiency versus interpretability of classifications. Biometrics. 21 : 768-769

Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data. J. Am. Statist. Assoc. 62 : 1159-1178

Gordon, A.D. (1981). Classification. Chapman and Hall, London

Gower, J.C. (1967). A comparison of some methods of cluster analysis. Biometrics. 23(4) : 623-628

Hotelling, H. (1931). The generalistaion of students ratio. Ann. Math. Stat. 2(3) : 360-378

Mahalanobis, P.C. (1936). On the generalised distance in statistics. Proc. Natl. Inst. Sci. India. 2(1) : 49-55

Mahalanobis, P.C., Majumdar, D.N. and Rao, C.R. (1949). Antropometric survey of the united provinces, 1941 : A statistical study. Sankhya. 9 : 90-324

Majumdar, D.N. and Rao, C.R. (1958). Bengal antropometric survey,1945: A statistical study. Sankhya. 19 : 201-408

Maronna, R. and Jacovkis, P.M. (1974). Multivariate clustering procedures with variable metrics. Biometrics. 30 : 499-505

Marriot, F.H.C. (1971). Practical problems in a method of cluster analysis. Biometrics. 27 : 501-514

Mukherjee, R.K. (1951) A study of differences in physical development of socio economic strata. Sankhya. 11 : 47

Nair, K.R. (1952) Use of measurments of more than one biometric character for discriminating between phases of six eye striped specimens of desert locusts. Indian J. Entomology. 14 : 126-136

Orloci, L. (1978). Multivariate analysis in Vegetation research. DR. W. Junk B.V. publishers, Boston.

Rajeevan, P.K. (1985). Intra clonal variations and nutritional studies in banana cv. 'Palayankodan' Unpublished Ph.D thesis submitted to Kerala Agricultural University.

Rao, C.R. (1948). The utilisation of multiple measurments in problems of biological classification. J. Roy. Soc. Suppl. 10 : 159-203

*Rao, C.R. (1951). An asymptotic expansion of the distribution of Wilks criterion. Bull. Inst. Inter. Statist. XXXIII(2) : 177-180

Rao, C.R. (1952). Advanced statistical methods in biometric research. John Wiley and sons, New york.

Rao, C.R. (1973). Linear statistical inference and its applications. John Wiley and sons, 2nd Ed., New York

*Rao, C.R. and Slater, P. (1949) Multivariate analysis applied to difference between neurotic groups. British J. Psy. Statistics. Section. 2 : 17

Scott, A.J. and Symons, M.J. (1971 a). On the Edwards and Cavalli Sforza method of cluster analysis. Biometrics. 27 : 217-219

Scott, A.J. and Symons, M.J. (1971 b). Clustering method based on likelihood ratio criteria. Biometrics. 27 : 387-397

Singh, R.K. and Choudhary, B.D. (1979). Biometrical methods in quantitative genetic analysis. Kalyani Publishers, New Delhi.

Sokal, R.R. and Michener, C.D. (1958). A statistical methods for evaluating systemic relationships. Univ. Kansas. Sci. Bull. 38 : 1409-1438

*Sokal, R.R. and Sneath, P.H.A. (1963) Numerical Taxonomy. W.H. Freeman, San Francisco.

Symons, M.J. (1981) Clustering criteria and multivariate normal mixtures. Biometrics 37(1) : 35-43

*Tildesley, M.L. (1921). A first study of Burmese skull. Biometrika. 13 : 176

*Tyron (1939) Cluster analysis. Ann. Arbor : Edward Brothers.

Ward, J.H. (1963). Hierarchical grouping to optimise an objective function. J. Am. Statist. Assoc. 58 : 236-244

Williams, W.T. (1976) Pattern analysis in agricultural science. CSIRO, Amsterdam

*Wilks, S.S. (1932). Certain generalisations in the analysis of variance. Biometrika. 24 : 471-494

Williams, W.T. and Lambert, J.M. (1959). Multivariate methods in plant ecology I. Association analysis in plant communities. J. Ecol. 47 : 83-101

*Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. Biometrika. 20 : 32-58

*Zubin, J. (1938). A technique for measuring likemindedness. J. Abnormal Social Psychol. 33 : 508-516

---

* Original not consulted.

Appendices

Appendix A

Mean values of 16 characters for the 24 genotypes (Data set I)

Sl. No. of genotypes	Girth of pseudo stem at shooting (cm.) (1)	Area of third leaf at shooting (m^2) (2)	Interval of leaf production (days) (3)	Total number of leaves (4)	Stomatal density of upper leaf (per mm^2) (5)	Weight of the bunch (Kg.) (6)	Number of hands (7)	Weight of hands (Kg.) (8)	Mean weight of a hand (Kg.) (9)	Number of fingers (10)	Average weight of a finger (gm.) (11)	Length of the bunch (cm.) (12)	Weight of the ripe finger (gm.) (13)	Weight of the pulp (gm.) (14)	Length of the pedicel (cm.) (15)	Girth of the finger (cm.) (16)
1	63.94	1.239	14.100	31.333	135.597	13.753	11.167	11.023	.990	159.167	70.733	45.833	73.333	56.667	2.933	11.067
2	63.223	1.198	12.433	30.500	105.660	11.383	10.333	9.520	.917	161.000	58.867	41.500	59.000	49.000	2.133	10.267
3	59.333	1.121	11.900	31.833	128.547	10.950	10.167	9.317	.907	155.833	59.233	41.333	57.333	45.000	2.767	10.800
4	62.000	1.192	13.733	31.500	181.210	13.500	10.833	10.877	1.003	165.833	65.433	46.500	60.667	47.000	2.633	10.400
5	59.387	1.180	14.600	30.833	134.353	11.750	10.167	9.770	.957	162.333	59.867	41.667	58.000	46.000	2.767	10.333
6	66.833	1.334	13.500	31.833	118.593	12.650	11.167	10.110	.907	176.000	57.433	46.333	67.667	52.667	2.433	10.300
7	64.943	1.340	15.100	30.833	149.253	14.457	12.000	11.750	.973	190.600	61.833	42.000	70.333	55.333	3.000	10.900
8	63.333	1.152	14.333	29.500	101.180	12.600	11.000	10.487	.953	169.500	61.833	44.083	63.333	52.000	9.400	10.700
9	62.443	1.119	12.667	31.667	127.290	11.883	10.500	9.460	.897	163.367	57.667	44.667	63.667	49.667	2.533	10.833
10	62.887	1.137	16.600	29.167	98.277	10.217	10.000	8.513	.850	142.667	59.633	38.890	59.000	47.333	2.167	10.267
11	65.110	1.264	13.667	29.667	117.350	14.867	11.167	12.387	1.113	177.167	70.067	45.667	75.333	59.333	3.000	11.800
12	61.667	1.104	17.200	30.000	117.350	9.783	10.333	7.887	.767	151.333	52.367	39.057	57.667	47.333	2.333	10.267
13	60.167	1.291	16.067	31.500	97.033	12.917	10.500	10.900	1.037	160.500	67.700	46.500	86.667	65.667	2.900	11.367
14	66.833	1.168	14.133	31.000	95.373	11.917	10.167	9.547	.937	146.500	65.100	43.777	63.333	49.000	2.400	11.100
15	64.000	1.462	15.233	29.667	114.447	15.167	12.000	12.850	1.073	195.333	65.500	50.667	68.667	54.333	2.567	10.100
16	58.390	1.226	12.567	30.667	125.643	12.950	10.333	10.303	1.003	168.000	61.733	45.667	73.667	57.333	3.267	11.167
17	64.220	1.084	15.400	29.833	99.520	10.667	10.167	8.793	.863	150.500	58.200	40.250	61.333	47.667	2.200	10.533
18	65.220	1.354	15.400	29.500	99.520	16.017	11.000	12.800	1.167	173.833	73.767	47.500	86.333	67.000	3.267	11.400
19	66.667	1.317	12.233	31.167	125.643	14.417	11.167	11.873	1.063	174.167	68.100	47.000	60.667	48.667	2.133	10.467
20	61.000	1.209	15.567	30.667	121.497	11.417	10.000	9.210	.913	151.167	60.567	40.667	72.000	54.667	2.600	10.833
21	67.557	1.304	13.133	30.000	120.253	16.333	11.667	13.643	1.170	179.167	76.600	50.333	99.000	82.667	3.367	11.833
22	65.053	1.056	14.467	27.500	111.960	11.150	10.500	9.170	.870	155.667	58.567	46.083	48.333	39.667	1.967	9.867
23	62.837	1.231	14.067	31.667	122.740	10.717	10.333	8.717	.843	165.167	52.800	43.750	54.333	41.667	2.567	10.067
24	65.110	1.194	12.900	31.500	122.490	13.567	11.333	11.170	.987	176.500	64.500	45.000	63.333	48.333	2.767	10.433

Mean values of 4 characters for 8 genotypes (Data II)

Sl. No. of genotypes	Number of ears per plant (1)	Ear length (cm.) (2)	100 grain weight (gm.) (3)	Grain yield per plant (gm.) (4)
1	41.900	20.300	3.900	85.675
2	43.800	19.750	3.650	98.250
3	37.300	18.725	4.600	74.575
4	41.150	20.300	4.300	91.650
5	32.500	20.250	4.100	54.125
6	52.750	19.725	4.375	100.375
7	43.900	20.225	4.275	91.000
8	46.750	20.025	4.150	82.025

Appendix B

D^2 values obtained for data I (24 X 24 matrix)

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00	16233.12	794.30	39976.93	41.72	5611.45	3692.71	20262.00	1285.47	26421.30	6235.84	6607.22
2		0.00	10120.79	107051.56	15401.12	2783.08	35309.98	403.95	8479.31	1512.88	2370.48	2593.69
3			0.00	51405.79	670.33	2367.63	7779.89	13487.47	93.08	18728.60	2802.55	3325.52
4				0.00	41392.80	75480.25	19548.81	116899.13	55336.42	131224.03	77751.66	78632.45
5					0.00	5160.88	4093.11	19246.45	1113.50	25239.15	5725.54	5983.98
6						0.00	18364.75	4713.57	1577.09	7873.48	43.81	327.21
7							0.00	40947.03	9301.24	49520.03	19433.54	19800.82
8								0.00	11614.94	593.18	4119.28	4042.68
9									0.00	16421.22	1951.03	2402.19
10										0.00	7088.64	6732.89
11											0.00	222.83
12												0.00

Contd...

	13	14	15	16	17	18	19	20	21	22	23	24
1	30035.33	30594.03	8575.83	1913.88	24612.15	25548.24	1711.84	4039.36	5012.61	10051.53	3029.62	2978.57
2	2368.76	2316.34	1340.34	7159.07	1011.05	1213.46	7545.77	4277.03	3351.25	831.00	5267.52	5331.41
3	21725.87	22046.70	4514.24	315.78	17124.78	17920.08	214.97	1560.01	2021.12	5535.68	869.05	789.30
4	139209.55	140492.44	85505.70	59047.04	127266.88	129399.47	57824.91	69279.95	73088.18	90009.06	64956.83	64627.03
5	28857.86	29423.13	7929.00	1681.75	23511.55	24454.90	1499.94	3596.45	4638.58	9347.64	2670.38	2649.98
6	9801.68	10044.62	396.54	1059.39	6821.15	7304.27	1230.62	241.60	69.26	733.15	414.90	448.54
7	54582.51	55386.92	23339.50	10861.79	47132.97	48453.41	10329.36	15294.09	17265.63	25731.75	13342.26	13232.08
8	1381.98	1240.31	2581.33	10117.21	361.12	550.85	10464.47	6415.08	5538.03	1838.48	7684.91	7799.66
9	19180.73	19551.19	3407.61	79.41	14919.77	15636.87	68.88	932.92	1278.64	4332.34	428.85	392.17
10	274.90	315.78	4910.48	14557.05	61.21	83.31	15160.35	9852.76	8815.99	3929.28	11634.96	11883.02
11	9004.33	9234.66	225.34	1358.95	6116.52	6580.57	1540.98	320.07	150.32	488.79	598.19	634.10
12	8770.15	9189.45	232.11	1789.00	5933.54	6434.13	2024.09	358.77	526.12	523.83	825.80	994.77
13	0.00	157.33	6587.23	17090.02	369.66	225.76	17923.60	12097.78	10717.95	5525.22	14070.60	14355.81
14		0.00	6849.38	17510.34	376.32	294.34	18136.00	12561.06	11054.15	5645.11	14402.07	14578.27
15			0.00	2609.94	4148.47	4546.40	2871.29	881.04	662.91	88.62	1446.19	1554.61
16				0.00	13138.88	13781.69	110.83	557.71	791.77	3441.70	218.18	208.90
17					0.00	35.49	13713.93	8782.88	7685.48	3239.53	10412.78	10615.51
18						0.00	14434.90	9335.53	8129.98	3632.26	11043.06	11254.90
19							0.00	756.43	1021.07	3670.85	282.94	203.21
20								0.00	213.24	1435.66	129.25	234.04
21									0.00	1100.22	348.43	370.68
22										0.00	2078.71	2162.99
23											0.00	30.69
24												0.00

D^2 values for data II (8 X 8) matrix

	1	2	3	4	5	6	7	8
1	0.000	17.704	64.856	24.082	24.329	30.325	14.737	20.481
2		0.000	104.806	50.074	78.105	71.648	47.517	67.752
3			0.000	23.082	59.423	37.118	25.217	62.957
4				0.000	44.991	28.555	4.247	45.929
5					0.000	39.775	28.767	15.618
6						0.000	12.447	12.178
7							0.000	22.597
8								0.000

Appendix C

Pairwise determinant values obtained for data I

	1	2	3	4	5	6	7	8	9	10	11	12
1	.0000E+01	.1921E+19	.2503E+21	.6390E+21	.3924E+20	.4417E+19	.5726E+21	.9004E+19	.6316E+20	.1789E+18	.3068E+19	.9393E+21
2		.0000E+01	.2183E+14	.3128E+18	.1734E+17	.1035E+15	.3565E+18	.8550E+15	.6367E+15	.1091E+16	.7672E+15	.1784E+18
3			.0000E+01	.1518E+19	.1466E+19	.1734E+19	.1550E+19	.9447E+17	.4788E+18	.1885E+19	.5666E+18	.4895E+22
4				.0000E+01	.4965E+19	.5825E+16	.3968E+19	.7857E+17	.2054E+20	.3733E+19	.8203E+17	.2031E+21
5					.0000E+01	.7536E+18	.1585E+19	.2104E+17	.1919E+18	.9034E+17	.1928E+15	.3654E+19
6						.0000E+01	.3981E+19	.6734E+18	.3807E+17	.6450E+16	.3980E+18	.4895E+21
7							.0000E+01	.1420E+19	.2993E+18	.2527E+20	.8405E+18	.1404E+22
8								.0000E+01	.1099E+18	.6034E+18	.1285E+16	.3209E+19
9									.0000E+01	.1294E+18	.1917E+17	.1618E+21
10										.0000E+01	.6087E+15	.8588E+21
11											.0000E+01	.1018E+17
12												.0000E+01

Contd ...

	13	14	15	16	17	18	19	20	21	22	23	24
1	.1224E+23	.2437E+20	.1135E+21	.4235E+20	.2227E+19	.7515E+20	.1082E+17	.5540E+18	.1507E+22	.5840E+19	.5868E+19	.1449E+23
2	.1142E+22	.2048E+19	.9318E+19	.1297E+20	.3944E+19	.1074E+21	.1544E+19	.5886E+19	.1232E+20	.1346E+19	.8262E+18	.7715E+21
3	.7129E+23	.1915E+20	.4086E+20	.1927E+22	.3638E+21	.5035E+22	.2447E+21	.4771E+20	.1063E+23	.9789E+17	.1024E+21	.2928E+23
4	.1737E+24	.1609E+20	.4877E+23	.1704E+22	.1142E+22	.1749E+24	.6332E+20	.2030E+22	.8331E+22	.6960E+19	.4620E+21	.2502E+24
5	.6457E+22	.8891E+20	.1554E+21	.7176E+20	.1189E+21	.1202E+23	.3591E+20	.4823E+20	.1075E+23	.1471E+19	.2240E+19	.7609E+22
6	.3352E+23	.5408E+20	.8252E+22	.8421E+20	.5471E+19	.2376E+23	.7115E+19	.8255E+19	.8469E+20	.8760E+18	.6552E+19	.4031E+23
7	.8097E+22	.1451E+21	.1189E+23	.4248E+22	.2459E+21	.9912E+22	.3821E+18	.9749E+21	.9109E+23	.1309E+21	.1268E+21	.3113E+22
8	.1145E+21	.4199E+20	.4094E+20	.3321E+20	.3829E+17	.5538E+22	.2493E+18	.1241E+20	.5169E+22	.8909E+19	.1804E+19	.7300E+21
9	.2757E+23	.1692E+21	.2530E+22	.4828E+20	.2846E+20	.1027E+22	.2208E+20	.1287E+20	.2726E+22	.2804E+20	.3368E+18	.1043E+23
10	.1685E+23	.2448E+21	.8395E+20	.1387E+22	.6096E+21	.9410E+20	.2173E+21	.2629E+21	.9906E+20	.5434E+18	.4881E+20	.1269E+23
11	.8986E+21	.9063E+16	.5759E+19	.3368E+20	.1068E+19	.4360E+21	.1659E+18	.1777E+19	.1938E+21	.4156E+17	.2429E+19	.1195E+22
12	.4543E+23	.1522E+21	.1372E+23	.3388E+21	.3093E+19	.3293E+24	.1224E+20	.1977E+22	.1156E+24	.9249E+19	.2356E+20	.8059E+22
13	.0000E+01	.3681E+20	.3142E+21	.2301E+21	.6983E+21	.2761E+20	.2267E+21	.1265E+22	.7227E+20	.7939E+19	.8358E+20	.1160E+21
14		.0000E+01	.3395E+20	.1618E+17	.8210E+14	.1180E+21	.5641E+17	.2718E+17	.5130E+20	.1736E+15	.2916E+15	.8318E+18
15			.0000E+01	.4387E+20	.1555E+20	.9168E+15	.5750E+19	.1337E+20	.1111E+17	.1300E+19	.3184E+19	.1151E+20
16				.0000E+01	.1028E+17	.5345E+21	.4301E+17	.6131E+18	.1873E+20	.3469E+18	.5482E+15	.5319E+20
17					.0000E+01	.1501E+20	.8600E+15	.1502E+17	.2454E+21	.3996E+17	.5892E+15	.7383E+20
18						.0000E+01	.4715E+20	.1195E+21	.9282E+17	.8908E+17	.7288E+19	.4141E+21
19							.0000E+01	.1716E+14	.4745E+20	.1913E+16	.9654E+15	.6936E+20
20								.0000E+01	.5724E+21	.1348E+18	.4379E+17	.4018E+20
21									.0000E+01	.1868E+19	.6587E+19	.3857E+22
22										.0000E+01	.1247E+14	.2505E+20
23											.0000E+01	.2631E+18
24												.0000E+01

Pairwise determinant values for data II (8 X 8 matrix)

	1	2	3	4	5	6	7	8
1	0.000	1146.202	8391.336	3363.443	1636.960	6264.366	2099.664	3744.833
2		0.000	7475.054	4729.512	2604.030	8440.415	1940.173	5524.935
3			0.000	2254.306	1684.364	4724.824	349.789	2048.417
4				0.000	1677.954	1919.894	120.374	6124.912
5					0.000	8426.336	377.527	290.896
6						0.000	1252.997	3342.383
7							0.000	1535.694
8								0.000

Appendix D

The first two canonical values and the corresponding canonical vectors of the between scatter matrix of transformed mean values

	eigen values		percentage of variance
	1	2	explained
Data I	28391.40	2502.42	99.67
Data II	255.452	180.315	80.75

The canonical vectors corresponding to the first two eigen values

Data set I		Data set II	
(1)	(2)	(1)	(2)
0.00221	0.00874	0.06872	-0.02558
0.00013	0.00717	0.17807	-0.18546
0.00847	-0.00683	-0.96944	0.11912
-0.00977	0.00619	0.15410	0.97507
-0.13187	0.11793		
-0.02296	0.04335		
0.07683	-0.08038		
-0.05747	0.05250		
-0.08944	0.05192		
0.65352	-0.00003		
0.67972	-0.02939		
-0.10941	0.06981		
0.20317	-0.14803		
0.12999	0.17233		
-0.06839	-0.69435		
-0.02783	-0.65774		

STANDARDISATION OF TECHNIQUES OF CLUSTERING GENOTYPES USING MAHALANOBIS D^2 AND WILKS' Λ CRITERION

By

SURESH K. M.

ABSTRACT OF THE THESIS

submitted in partial fulfilment of the
requirement for the degree

Master of Science (Agricultural Statistics)

Faculty of Agriculture
Kerala Agricultural University

Department of Statistics
COLLEGE OF VETERINARY AND ANIMAL SCIENCES
Mannuthy - Trichur

1986

ABSTRACT

Two major drawbacks of Tocher's method of clustering genotypes using Mahalanobis D^2 were pointed out and an improvement over Tocher's method was suggested. The cluster configuration obtained by these two methods were compared with those obtained by canonical analysis method.

A new computer oriented iterative algorithm for clustering using Mahalanobis D^2 values was proposed.

A procedure for formation of clusters statistically, using Mahalanobis D^2 was suggested to form maximum nonsignificant subsets of genotypes.

A new measure of dissimilarity which does not require any assumption on distribution of the population, viz., the determinant of the pairwise scatter matrix was proposed in the study.

Minimum /W/ criterion of Friedman and Rubin (1967) was also used for clustering. The clustering obtained by the new iterative algorithm using either Mahalanobis D^2 or determinant of pairwise scatter matrix or both could be used as the initial solution for it.

A graphical method for determining the optimum number of clusters was suggested.

The different methods were illustrated in two sets of data.