Maydica 41 (1996): 307-316

CLUSTER ANALYSIS, AN APPROACH TO SAMPLING VARIABILITY IN MAIZE ACCESSIONS[1]

F. Rincon[2], B. Johnson[2,*], J. Crossa[3], S. Taba[3]

[2] Department of Agronomy, University of Nebraska-Lincoln, 326 Keim Hall, P.O. Box 830915, Lincoln, NE 68583-0915, U.S.A.
[3] International Maize and Wheat Improvement Center (CIMMYT), Lisboa 27, Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

Received August 9, 1996

ABSTRACT - Cluster analysis is frequently used to classify maize (*Zea mays* L.) accessions and can be used by breeders and geneticists to identify subsets of accessions which have potential utility for specific breeding or genetic purposes. Phenograms can be utilized to define subsets of accessions on the basis of dissimilarity coefficients. Phenograms created using cluster analysis depend on the clustering method, and on type and number of attributes used to compute associations among individuals. The objectives of this study were to 1) compare several clustering strategies used for grouping Caribbean maize accessions, 2) define groups having similar characteristics, and 3) obtain a representative subset of the total number of accessions evaluated. Four hierarchical clustering strategies were compared: single linkage, unweighted pair-group method using arithmetic averages (UPGMA), using centroids, and Ward's. Each method was evaluated using two data sets, and varying types and numbers of traits. Average euclidean squared distance was used as the dissimilarity measure. Phenogram agreement was evaluated by the cophenetic correlation coefficients. Cophenetic correlation and inspection of phenograms suggested that in preference to the other strategies, UPGMA can be utilized to group maize accessions using agronomic and morphological data. Number of individuals and number of traits affected computation of dissimilarity measures among accessions. For large data sets, it might be useful to include as many traits as possible to compute the dissimilarity measures. In addition to clustering methods, principal component analysis helped to form groups which had particular characteristics that accounted for phenotypic diversity present in the whole population. Groups were formed on basis of common clusters identified by a consensus analysis. Each group was exposed to a stratified sampling process to define a subset in proportion to their number of accessions. A set of 43 entries (23%) was identified as a selected subset representing the 184 accessions evaluated. The relationships among accessions defined by the phenogram, and the associated race classification indicated that phenetic relationship can be used to group maize accessions, and consequently definine subsets, in proportion to the number of accessions.

KEY WORDS: Clustering strategies; Cophenetic correlations; Maize classification.

INTRODUCTION

Knowledge of relationships and common characteristics of individuals can assist breeders and geneticists with identification of sets or groups of individuals which have potential utility for specific breeding or genetic purposes. Classification is defined as the arraying of organisms into groups (or sets) on basis of relationships (SNEATH and SOKAL, 1973). Classification can be conceptualized as the process by which grouping of individuals is attained, the result of the grouping process, or the resultant hierarchy of classes (STUESSY, 1990). Individuals are often described by numerical measurements, and, in many cases, data involve a mixture of several types of variables. For instance, data can be quantitative, expressed as continuous (plant height) or ordinal (ear quality) measurements. Alternatively data may be qualitative with measurements on a nominal scale (plant color) or binary (presence/absence) scale. Computation of similarities among individuals is affected by type and number of attributes measured. GOWER (1971) described a general coefficient for measuring similarities that can be used under many different circumstances. This coefficient can be based upon various types of characters, e.g., dichotomous, quantitative, and qualitative. ANDERBERG (1973) analyzed the type of variables, their association, and their implication in cluster analysis. He discussed properties of the mean, range, and standard deviation as alternatives for equalizing different types of variables. If measurements are subject to random fluctuations of different

[1] Contribution of the Agric. Res. Div., Univ. of Nebraska-Lincoln. Published as Journal Series Number 11512.

* For correspondence (fax +1 402 472 7904).

magnitudes, recorded on different units or scales, it is often desirable to standardize to a mean of zero and unit variance (EVERITT, 1980; JOHNSON and WICHERN, 1992). Several authors have suggested first computing a principal component analysis, then using the first principal component (PC) scores as input variables for the clustering process (EVERITT, 1980; GOODMAN, 1972; GOODMAN and BIRD, 1977).

Hierarchical cluster analysis has been suggested for classifying entries of germplasm collections based on degree of similarity and dissimilarity (PEETERS and MARTINELLI, 1989; VANHINTUM, 1995). Similarly, a combination of cluster analysis and principal component analysis has been used to classify maize (*Zea mays* L.) accessions (CROSSA *et al.*, 1995). Several studies have reported comparisons of cluster analysis algorithms, including those used for classification of germplasm collections (PEETERS and MARTINELLI, 1989), and classification of maize inbreds (MUMM and DUDLEY, 1994).

The relationships between distances or dissimilarities among all possible pairs of individuals can be determined using correlation and regression analyses, and comparing results obtained with data obtained from an independent source such as co-ancestries based on pedigree records, or genetic analysis (AJMONE-MARSAN *et al.*, 1992; SMITH and SMITH, 1992). However, if there are no reference data for comparison, a particular phenogram must be evaluated by other approaches. Evaluation can be accomplished by using the cophenetic correlation coefficient developed by SOKAL and ROHLF (1962). Several studies have shown that the cophenetic correlation coefficient is appropriate to validate the agreement of phenograms obtained from different clustering algorithms (MOLINA *et al.*, 1992; MUMM *et al.*, 1994).

Difference in magnitude of genotype performance in different environments is commonly characterized by measuring quantitative characters, and determining the magnitude of the genotype-by-environmental interaction. Usually those characters which are effectively used for classification are those less affected by the environment, and not subject to wide variation within the samples considered (STUESSY, 1990). GOODMAN and PATERNIANI (1969) showed that appropriate characters for racial classification are those with higher values of a ratio given by: $r = \sigma^2_r / (\sigma^2_e + \sigma^2_y)$, where σ^2_r, σ^2_e, and σ^2_y are estimates of variance components for collections, environments, and collection-by-environment interaction, respectively. They found that vegetative traits had relatively lower values, and were therefore, considered unstable and strongly affected by the environment.

Another approach to biological classification is phenetics. Phenetics is a system of classification based on numerous precisely determined characters which uses overall similarity to describe relationships (SNEATH and SOKAL, 1973; STUESSY, 1990). In phenetics, as many characters as possible are included without any weighting. When using phenetics, it is desirable to include characters that represent the total life-cycle of the organism. The phenograms obtained illustrate only "phenetic similarity," based entirely on the comparison of character states.

Effective utilization of germplasm collections, usually represented by large numbers of accessions, requires information from previous evaluation and characterization. To access and use genetic diversity in large collections, BROWN (1989a) indicated that diversity of a subset could be efficiently retained if a good sampling procedure was applied to the collection. He recommended retaining 10% of the accessions in the collection, up to about 3000 entries to establish a core. Considering number of accessions in non-overlapping groups, BROWN (1989b) compared the constant and proportional strategies of sampling. Constant strategy biased selection in favor of groups with reduced number of accessions, whereas, proportional strategy biased selection in favor of groups with large number of accessions. To avoid bias in selection, BROWN (1989b) proposed that a representation of the groups should be in proportion to the logarithm of the number of accessions in the group. YONEZAWA *et al.* (1995) compared five stratification strategies and concluded that a proportional strategy should be used in defining subsets, particularly if genetic diversity within groups is unknown, as is the case for many germplasm collections. Sample size would be in proportion to the number of accessions in the groups. Alternatively, when genetic diversity is known, stratification would be in proportion to the range of genetic diversity.

The objectives of this study were to 1) compare several clustering strategies used to define groups of maize accessions based on agronomic and morphological traits, 2) describe the phenetic relationships among accessions used to define similarity groups, and 3) develop a representative subset of the total number of Caribbean accessions included in this study.

MATERIALS AND METHODS

Experimental data

The study included 184 accessions of the Caribbean collection of maize maintained at the International Maize and Wheat Improvement Center (CIMMYT), El Batan, Mexico. The original field

TABLE 1 - *Selected traits used for cluster analysis, and estimates of variance components for 249 maize accessions evaluated in four environments in Mexico.*

Trait	Selected traits			Variance components †			
	Type ‡	Evaluation §	Grouping ¶	σ^2_e	σ^2_g	σ^2_{ge}	r #
Field germination (%)	1	x	x	1.94	3.00	4.06	5.50
Vigor	2		x	0.04	0.07	0.07	0.64
Days to silk (50%)	1	x	x	204.25	38.39	5.61	0.18
Days to pollen (50%)	1	x	x	182.92	33.08	4.76	0.18
Plant height (cm)	1	x	x	526.24	925.98	56.72	1.59
Ear height (cm)	1	x	x	457.12	775.95	52.12	1.52
Leaves above ear	1	x	x	0.12	0.13	0.02	0.93
Forage rating	2	x	x	0.04	0.16	0.08	1.33
Husk cover	2			8.40	0.03	0.02	0.00
Root development	2		x	0.50	0.08	0.08	0.14
Root lodging (%)	1			31.74	29.63	24.80	0.52
Stalk lodging (%)	1			10.64	26.57	9.87	1.30
Senescence	1	x	x	8.48	8.23	6.73	0.54
Ear length (cm)	1	x	x	1.94	1.19	0.30	0.53
Ear diameter (cm)	1	x	x	0.56	0.09	0.01	0.16
Ear rot (%)	2		x	0.46	0.06	0.06	0.12
Ear quality	2		x	0.51	0.08	0.09	0.13
Kernel row number	1		x	0.55	0.50	0.57	0.45
Adaptation	2		x	0.51	0.08	0.12	0.13
Agronomic scale	2		x	0.55	0.09	0.09	0.14
Ears per plant	1		x	0.62	0.01	0.01	0.02
Grain yield (t ha^{-1})	1	x	x	2.16	0.54	0.32	0.22

† Variance components for environments (σ^2_e), accessions (σ^2_g), and interaction (σ^2_{ge}).
‡ Quantitative traits: 1= Continuous; 2= Ordinal (scale 1 to 5).
§ Traits selected for the evaluation of clustering methods.
¶ Traits selected for cluster analysis used to identify similar groups.
# Ratio, $r = \sigma^2_g / (\sigma^2_e + \sigma^2_{ge})$.

evaluation included 250 accessions and six checks planted in four environments. Field evaluations were conducted at two locations in Mexico: Tlaltizapan, Morelos (940 masl), and Poza Rica, Veracruz (60 masl), with two growing seasons per location (dry and wet), identified as A and B, during 1992 and 1994. The four experiments were organized as two replication 16 x 16 lattice square designs and were designated as PR92B, PR94A, TL92B, and TL94A. From the total accessions evaluated, only 184 belonged to the Caribbean collection, and those 184 formed the basis for the present study. Checks included were Oaxaca 244, an accession from Mexico; three CIMMYT populations, Across 8331, Poza Rica 8432, and Across 8443; and two experimental hybrids.

Accessions were characterized by 22 agronomic and morphological traits (Table 1). Because of difficulties in computing dissimilarity measures using a mixture of types of characters, only quantitative traits (continuous and ordinal) were considered in this study. Specific traits used in the analysis displayed significant variability among accessions, and were recorded from all environments. Data on traits were first examined using analyses of variance and frequency distributions. Data were also tested for effective discrimination among accessions by using stepwise discriminant analysis (PROC STEPDISC) of SAS (SAS, 1989). Collectively, the traits covered all stages of plant development, from field germination to harvest, and are described on the descriptors for maize, International Board for Plant Genetic Resources (IBPGR, 1991). Individual and combined analyses of variance, and computation of least-squares means were made using PROC GLM of SAS (SAS, 1989).

Estimates of variance components were obtained using the multiple-trait, derivative-free, restricted maximum-likelihood method (MTDFREML) developed for animal models (Boldman K.G., L.A. Kriese, L.D. Van Vleck, C.P. Van Tassell, S.D. Kachman, 1995 A manual for use of MTDFREML. A set of programs to obtain estimates of variances and covariances. Draft. USDA-ARS). These estimates were used to compute a ratio given by $r = \sigma^2_g / (\sigma^2_e + \sigma^2_{ge})$, where σ^2_g, σ^2_e, and σ^2_{ge} are estimates of variance components for accessions, environments, and accession-by-environment interaction, respectively. These ratios have been used to identify elucidative traits for maize classification (Goodman and Paterniani, 1969; Sánchez *et al.*, 1993).

Evaluation of clustering strategies

Four hierarchical clustering methods were compared: 1) single linkage, 2) unweighted pair-group method using arithmetic averages (UPGMA), 3) unweighted pair-group method using centroids (UPGMC), and 4) Ward's minimum-variance method. Anderberg (1973) grouped these algorithms into linkage methods (single linkage and UPGMA), centroid, and error sum of squares methods (Ward's method). The algorithms were also classified as agglomerative techniques by Everitt (1980). Single linkage, UPGMA, and UPGMC methods, referred to as the standard strategies, are described in detail by Lance and Williams (1967) and Sneath

and SOKAL (1973). Features of the Ward's method of clustering are described by WARD (1963).

Assuming that D_{IJ} measures dissimilarity between groups I and J, containing n_i and n_j elements, respectively. These groups fuse to form a new group K with n_k elements ($n_i + n_j$). The association of this K group, with a new group H, is denoted by the combinatorial formula:

$$D_{HK} = \alpha_i D_{HI} + \alpha_j D_{HJ} + \beta D_{IJ} + \gamma | D_{HI} - D_{HJ} |$$

where the parameters α_i, α_j, β and γ determine the differences between clustering strategies (EVERITT, 1980; LANCE and WILLIAMS, 1967). Descriptions of the four hierarchical clustering methods follow.

Single linkage is the simplest of the hierarchical strategies. In this strategy, groups with n_i, and n_j elements are fused using individual entries by merging nearest neighbors (D_{IJ}). Thus, the distance with any other cluster is computed by $D_{(IJ)W} = \min(D_{IW}, D_{JW})$. The parameters used by the combinatorial formula are $\alpha_i = \alpha_j = 0.5$; $\beta = 0$; and $\gamma = -0.5$ (EVERITT, 1980; LANCE and WILLIAMS, 1967).

UPGMA. This strategy computes the average distance to form a cluster (IJ). This process uses the distance of all pairs of individuals in the cluster (n_i, n_j). The distance between the group (IJ) and another cluster H is obtained by:

$$D_{(IJ)H} = \frac{\Sigma_i \Sigma_k D_{ik}}{N_{(K)} N_{(H)}}$$

where D_{ik} is the distance between individual in cluster (IJ) and individual k in cluster H. $N_{(K)}$ and $N_{(H)}$ are the number of items in clusters (IJ) and H, respectively. The parameters used by the combinatorial formula are $\alpha_i = n_i/n_k$; $\alpha_j = n_j/n_k$; and $\beta = \gamma = 0$ (EVERITT, 1980; LANCE and WILLIAMS, 1967).

UPGMC. This strategy merges two clusters with the most similar mean vectors of centroids, starting with the smallest distance. The coefficients used by the combinatorial formula are: $\alpha_i = n_i/n_k$; $\alpha_j = n_j/n_k$; $\beta = -\alpha_i \alpha_j$; and $\gamma = 0$ (EVERITT, 1980; LANCE and WILLIAMS, 1967).

Ward's method. This strategy finds those two clusters whose merger at each stage gives the minimum increase in the total within-group error sum of squares (ESS). The minimum increase in the ESS is proportional to the squared euclidean distance between the centroids of the merged clusters. The ESS is given by:

$$ESS = \sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2$$

where X_i is the score of the individual. The parameters used by the combinatorial formula are $\alpha_i = (n_k+n_i)/(n_k+n_i+n_j)$; $\alpha_j = (n_k+n_j)/(n_k+n_i+n_j)$; $\beta = -n_k/(n_k+n_i+n_j)$; and $\gamma = 0$ (EVERITT, 1980).

Means across environments for two data subsets containing 17 and 62 accessions were generated to test the clustering methodologies. These subsets represent groups of accessions classified as Haitian yellow (HATIYE) and coastal tropical flint (COSTCR), respectively. Number of accessions was selected to compare dissimilarity measures of two contrasting data sets. In addition to the accessions, data from the six checks augmented the two data sets, generating final data sets of 23 and 68 individuals, respectively. The main objective of including the checks was to provide a way to compare the clusters formed in phenograms. Because the checks represented five improved materials and one accession, it was expected that they would constitute different clusters. Each data set was evaluated using different types and numbers of traits: Analysis I, all 22 traits, Analysis II, a selected group of 11 traits, and Analysis III, PC scores. A total of 22 quantitative traits was selected for the first Analysis, and eleven for the second (Table 1). For Analysis III, the first PC scores from PCA obtained by the PRINCOMP procedure of SAS (SAS, 1989) were used as input traits. A PC score was used as an input trait when either the eigenvalue was equal to or greater than one ($\lambda \geq 1$), or when the cumulative variance explained exceeded 80%. Four and five PCs were selected when the number of individuals was 23, and five when the number of individuals was 68.

For Analysis I and II, data were standardized by subtracting the mean of each trait, and dividing by its standard deviation (Z-score) before the computation of association among individuals: $Z_{ij} = (X_{ij} - \bar{X}_j)/S_j$. Phenograms were obtained for all clustering methods, using the average euclidean squared distance as the dissimilarity measure among pairs of individuals. Phenograms were generated by NTSYS-pc, the Numerical Taxonomy and Multivariate Analysis System (ROHLF, 1994) except for Ward's method. Phenograms for Ward's method were obtained by using the procedures CLUSTER and TREE of SAS (SAS, 1989).

Phenograms were checked for agreement using the cophenetic correlation coefficient following the method developed by SOKAL and ROHLF (1962). Cophenetic correlation is an ordinary product-moment correlation between the corresponding elements of two matrices. The distance (dissimilarities) matrix and the cophenetic values (average dissimilarities) were obtained from the phenograms (SOKAL and ROHLF, 1962; SNEATH and SOKAL, 1973). Matrices of cophenetic values were obtained by using NTSYS-pc (ROHLF, 1994) except for the Ward's method. Matrices of cophenetic values for Ward's method were extracted directly from phenograms obtained by SAS. Cophenetic values obtained by using NTSYS-pc and those extracted from phenograms generated by SAS were compared with the UPGMA method. Correlation coefficients of $r_{CC} = 0.99$ were found between matrices for two data sets of 13 and 68 individuals, respectively. The comparison of dissimilarity and cophenetic value matrices for all clustering strategies evaluated was made by using NTSYS-pc.

Grouping accessions and development of a representative subset

Adjusted means of 19 traits (Table 1) were standardized to a mean of zero and unit variance to avoid effects due to scaling differences. The standardized values were utilized to compute average euclidean squared distance, which was the dissimilarity (distance) measure between all possible pairs of accessions:

$$E_{ij}^2 = \frac{1}{n}\Sigma_k \left(X_{ik} - X_{jk}\right)^2$$

The inter-individual dissimilarity matrix was then used by the UPGMA clustering method to obtain a phenogram which included the 184 accessions for each environment. To identify common clusters among the 184 accessions, a consensus tree (common part) obtained from the four phenograms was generated by using NTSYS-pc (ROHLF, 1994), given the parameter s=0.5 to the Stinebrickner s-consensus method (STINEBRICKNER, 1984). Each common part or subset represented a group of similar individuals. Principal component analysis was used to detect a pattern within subsets (clusters) identified by the consensus tree (WILLIAMS, 1976). Once the pattern was detected, several clusters were combined to form groups of accessions containing common characteristics. These groups were defined on basis of phenetic dissimilarity computed from the selected traits. Adjusted means across environments for 22 traits were used for PCA by using the PRINCOMP procedure of SAS (SAS, 1989).

To define a representative sample from the total number of

TABLE 2 - *Cophenetic correlation coefficients between the elements of the cophenetic values and distance matrices (r_{CD}) from four clustering methods.*

Clustering method †	Type and number of traits			PC ‡	Average
	No. individuals	Selected 11 traits	All 22 traits		
Single linkage	23	0.82	0.87	0.84	0.84
	68	0.39	0.39	0.37	0.38
UPGMA	23	0.91	0.88	0.88	0.89
	68	0.60	0.61	0.61	0.61
UPGMC	23	0.91	0.90	0.91	0.91
	68	0.59	0.63	0.62	0.61
Ward	23	0.88	0.83	0.83	0.85
	68	0.58	0.55	0.54	0.56
Average (23 individuals)		0.88	0.87	0.87	0.87
Average (68 individuals)		0.54	0.55	0.54	0.54

† UPGMA = Unweighted pair-group method using arithmetic averages.
UPGMC = Unweighted pair-group method using centroids.
Ward = Ward's minimum-variance method.
‡ Principal component scores.

accessions, a stratified random sampling strategy was used. Stratified random sampling results in retention of variability proportional to the number of accessions contained in each group (YONEZAWA *et al.*, 1995). For stratified random sampling, accessions were selected from the phenogram by defining low dissimilarity values, thereby dividing accessions into subgroups, and randomly choosing accessions from each subset (CROSSA *et al.*, 1995). New phenograms were obtained for each group to discriminate individuals within retained accessions. This process continued until the desired number of accessions was included in the final subset.

RESULTS AND DISCUSSION

Evaluation of clustering strategies

The cophenetic correlation coefficients between the cophenetic values and dissimilarities (r_{CD}) was computed to evaluate the different clustering methods (Table 2). In general, cophenetic correlation coefficients decreased as number of individuals changed from 23 (r_{CD} = 0.87) to 68 (r_{CD} = 0.54) for all methods tested. ROHLF and FISHER (1968) reported that when the dissimilarity measures are considered, the magnitude of the cophenetic correlation coefficient decreases if the number of individuals increases about 50, but no changes result over 50. ROHLF and FISHER (1968) stated that for large data sets, the cophenetic correlation coefficients have similar values and are not affected by the number of characters. The average coefficients were very similar for type and number of traits utilized on different clustering methods. The only large difference in magnitude was due to the number of individuals included in the two data sets (Table 2).

When the number of individuals was 23, the selected traits had high cophenetic correlation coefficients for all methods, except single linkage (Table 2). Coefficients obtained from all 22 traits and PC scores had similar or lower values than values as compared to those for the 11 selected traits, with the exception of those obtained using single linkage. The average values show that the UPGMC method resulted in the highest cophenetic correlation coefficient, followed by UPGMA, Ward's, and single linkage (r_{CD} =0.91, r_{CD} =0.89, r_{CD} =0.85, and r_{CD} =0.84, respectively). When the number of individuals was 68, UPGMA and UPGMC had similar performance; high cophenetic values were found when all 22 traits were included, followed by those for PC scores and selected traits. Single linkage and Ward methods had high correlation coefficients when selected traits were used, followed by all 22 traits, and PC scores. On average, UPGMC had the highest cophenetic correlation coefficient, followed by UPGMA, Ward's, and single linkage (r_{CD} =0.61, r_{CD} = 0.61, r_{CD} = 0.56, and r_{CD} =0.38, respectively). This is the same pattern found when the number of individuals was 23. In both scenarios, UPGMA and UPGMC had very similar correlation coefficients. A phenogram is considered a good representation of a matrix of associations if the cophenetic correlation coefficient is r_{CD} =0.85 or higher (STUESSY, 1990). However, low values do not mean

that the phenogram has no utility, only that some distortion may have occurred. It should be noted that there is no statistical test for the correlation coefficient because of the lack of independence of the individual coefficients in the dissimilarity matrices.

Phenograms (not shown) indicated that when the number of individuals included in the analysis was 23, all methods showed acceptable grouping relative to each other, and were not affected by type or number of traits. This conclusion was based on the fact that the clusters formed in the phenograms included a cluster which grouped the six check individuals together, which represent the improved germplasm. These materials originated from different sources and types of germplasm than the Caribbean accessions. The clustering algorithms allocated the checks into distinctive clusters apart from the accessions. Therefore, this check group was utilized as a tag in all phenograms. The same group of entries was identified when the number of individuals was 68. In this instance, when 11 traits were used, the augmented entries were similarly grouped. When PC scores were used, the UPGMA was the only method that grouped the check entries similarly to the previous analysis for 11 traits. Single linkage, UPGMC, and Ward's divided the check entries into different clusters, and separated them from clusters of accessions. When all 22 traits were included, the UPGMA method was consistent and gave similar separation of the check group of entries. The UPGMC and Ward's methods grouped these entries into four different subsets, distributed along the phenogram. The UPGMA method was generally consistent in regard to the allocation of clusters, within the different type and number of traits.

The cophenetic correlation coefficient and visual inspection of phenograms indicated that the UPGMA method can be used for grouping maize accessions. These results were consistent over the two data sets and were not affected by number or type of traits compared. By analyzing properties of the cophenetic correlation, FARRIS (1969) indicated that pair-grouping, in particular the UPGMA procedure, would maximize cophenetic correlation coefficients. Several studies have shown that the UPGMA cluster algorithm provides consistency in grouping biological material with relationships computed from different types of data (AJMONE-MARSAN *et al.*, 1992; MOLINA *et al.*, 1992; MUMM *et al.*, 1994; ORDÁS *et al.*, 1994). For data sets containing few individuals and few traits, Ward's method can be used as an alternative. However, if the number of individuals is increased considerably, it may be less suitable. This specific method might divide dense clusters in an unacceptable manner (SNEATH and SOKAL, 1973). In this particular case, both the single linkage and Ward's methods gave different solutions when compared to UPGMA, but Ward's method had higher cophenetic correlation coefficients than did the single linkage method (Table 2).

TABLE 3 - *Correlation coefficients between dissimilarity matrices obtained from different traits and two data sets, above diagonal (n=68), below diagonal (n=23).*

	Selected 11 traits	All 22 traits	PC †
Selected 11 traits		0.89	0.89
All 22 traits	0.95		0.99
PC	0.94	0.99	

† Principal component scores.

Correlation coefficients between dissimilarity matrices for different traits and individuals were computed (Table 3). For the two data sets evaluated, correlation coefficients decreased when number of individuals increased from 23 to 68 for all groups of traits. High correlation coefficients were found between dissimilarities obtained from PC scores and all 22 traits. Such values are expected because a few PC scores accounted for most of the variability when all traits were included. There is an apparent tendency that reduction in number of traits had an effect on the correlation coefficient with values of r_{DD} = 0.95 and r_{DD} = 0.89 for 23 and 68 individuals, respectively (Table 3).

Our results suggests that UPGMA can serve as an accurate clustering method to group maize accessions by using agronomic and morphological data. This method had very similar cophenetic correlation coefficients for all types and number of traits. However, if the number of elements is relatively large, it might be useful to include as much information as possible to compute dissimilarity measures. Both the number of individuals and number of traits affected the computation of dissimilarity measures as shown in Table 3. Incorporation of many traits makes computation of associations among accessions more stable, and is advantageous, especially if missing values exist in the data matrix.

Grouping accessions and development of a representative subset

Combined analysis of variance showed a very highly significant accession-by-environment interaction ($P < 0.001$) for all traits listed in Table 1, except ear diameter and number of leaves above the ear ($P <$

0.05). The large interaction indicates a need to consider a procedure suggested by GOODMAN and PATERNIANI (1969). They used variance component estimation to compute a ratio given by: $r = \sigma^2_g / (\sigma^2_e + \sigma^2_{ge})$, as shown in Table 1. They suggested that traits having values of a ratio larger than 3 are appropriate for racial classification. In this study, except for field germination, all traits had values below 3 (Table 1). According to GOODMAN and PATERNIANI's (1969) criteria, we could not identify a group of traits which could accurately be used for classification. Consequently, an alternative approach was used.

To define similar groups of accessions, both the racial classification and country of origin were inspected for possible patterns using PCA. There were no obvious patterns useful to identify groups with similar characteristics. Because identification of groups was necessary for obtaining a good stratified sampling, phenetic relationships among accessions were computed as an alternative approach for defining groups with similar characteristics. In phenetics, as many traits as possible are used to compute associations among individuals. Phenetic relationship can be used by cluster analysis to group individuals according to the dissimilarity coefficient (STUESSY, 1990; VANHINTUM, 1995). Moreover, by increasing the number of traits, the value of similarity coefficient becomes more stable (STUESSY, 1990). For this reason, all traits listed in Table 1 were used to compute dissimilarity measures, excluding only husk cover and root and stalk lodging, which had high coefficients of variation. Consensus analysis of four phenograms (one for each environment) of 184 accessions identified the common subsets, which in total generated 12 clusters. These clusters represented similar groupings of the same number of accessions in all environments.

The scatter plot of the first two PC scores showed a pattern associated with clusters identified by the consensus analysis (not presented). A general approach was suggested by WILLIAMS (1976), who pointed out that PCA can be used as a pattern-finding technique. Considering the pattern associated with the clusters, a set of five groups of accessions was defined, here after referred to as A1, A2, A3, A4, and A5 (Table 4). The partial listing of eigenvalues and eigenvectors of PCA are presented in Table 5. This table contains only information for those traits that contribute the most to define the characteristics of the first two component scores, on basis of the higher coefficients. Collectively, the first two PC scores explained 60% of the total variability expressed by the 22 traits, the third PC score explained only 9%. The scatter plot of the first two PC scores in Fig. 1 shows the dispersion of the five groups. The relative magnitudes of the coefficients (eigenvectors) in Table 5 reflect the relative contribution of each trait to PC scores. Thus, those traits that largely define PC1 included plant height and the negatively associated traits: root development, stalk lodging, and forage rating. The second PC was principally a function of days to anthesis and root lodging, negatively associated with days to senescence and grain yield.

Although racial classification and country of origin did not show a pattern, they can be described in relation to each of the groups established. Group A1 was represented mainly by the races Coastal tropical flint (COSTCR), Cubano amarillo (CUBAAM), and TUSON (Table 4). Group A2 included several races, 50% of

TABLE 4 - *Racial classification associated with assigned groups of 184 maize accessions (SET1).*

	Group					
Race name †	A1	A2	A3	A4	A5	Total
Canilla	4		5	1	1	11
Chandelle	7	1				8
Coastal tropical flint	21	4	9	17	9	60
Cubano amarillo	23	3	2			28
Early Caribbean	1		1		1	3
Haitian yellow		2	15			17
Mezcla		2				2
Puya		3	4			7
St. Croix		1				1
Tuson	13	15	19			47
Total	70	30	55	18	11	184

† Primary race name classification at CIMMYT.

TABLE 5 - *Partial listing of eigenvalues and eigenvectors from first three principal components (PC) from analysis of 184 accessions*

Traits	PC1	PC2	PC3
Days to anthesis	0.24	0.31	–0.03
Plant height	0.30	0.10	–0.01
Root development †	–0.30	0.01	0.06
Root lodging	–0.04	0.26	0.10
Stalk lodging	–0.24	0.04	0.09
Forage rating †	–0.30	–0.13	0.02
Days to senescence	–0.05	–0.41	–0.13
Grain yield	0.20	–0.31	0.04
Eigenvalues	9.26	4.61	1.96
Proportion of variance explained	0.40	0.20	0.09

† Scored 1 to 5 (1=abundant, 5=poor).

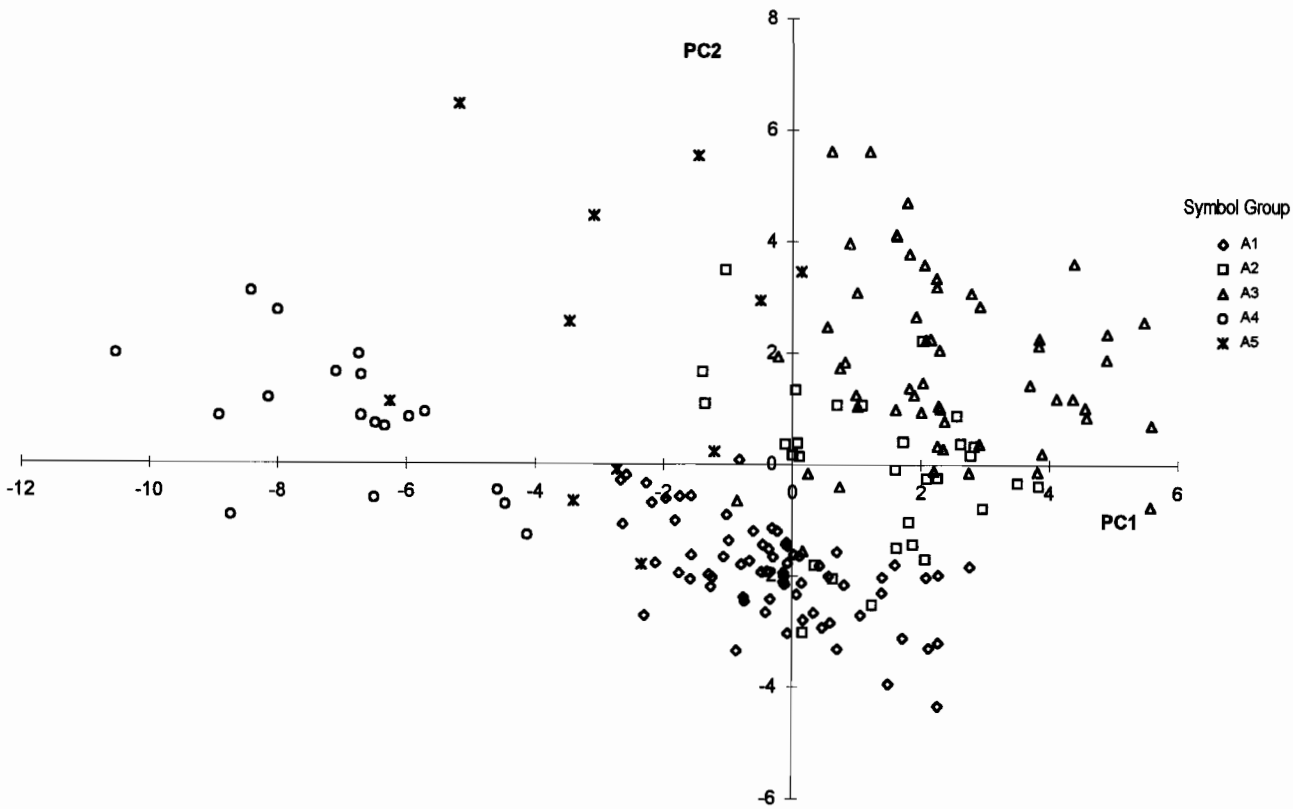

FIGURE 1 - Scatter plot of the first two PC scores showing the dispersion of the assigned groups of 184 Caribbean maize accessions.

which are TUSON. Table 4 shows that Haitian yellow (HAITYE) and TUSON were the races with the most accessions in group A3. Most accessions in group A4 and A5 were COSTCR. Grouping the accessions either by race or country of origin alone is a complex matter. In the West Indies, Brown (1960) identified seven distinctive racial groups, including Cuban flint, Haitian yellow, Coastal tropical flint, Chandelle, Early Caribbean, Tuson and St. Croix. There are about 19 racial groups identified in the Caribbean countries (primary race name) according to passport information available in the CIMMYT germplasm bank (personal communication). The increase in number of racial groups is due in part to introduction of new races, as well as germplasm exchange and interracial hybridization (GOODMAN and BROWN, 1988).

The previous discussion attempted to explain some of the difficulties of working with the Caribbean material, particularly in regard to defining similar groups. Results of cluster analysis and complementary findings by PCA suggested that multivariate techniques were useful in describing complex data sets. Principal component analysis provides interpretation of the biological phenomena based on analysis of a few components, and can provide a two-dimensional graphical representation. Interpretations can be based upon PC scores and magnitude of coefficients, the eigenvectors. The biological interpretations of Fig. 1 are substantiated in the means reported in Table 6. Thus, from Fig. 1, it can be seen that accessions in Group A1 are characterized as early, with low root lodging, high yield performance, and intermediate plant height. Group A2 has average yield performance and inter-

TABLE 6 - *Phenotypic means of six traits for five groups of accessions averaged over four environments.*

Group	Days to anthesis	Plant height	Root lodging	Stalk lodging	Days to senesce	Grain yield
	d	cm	%	%	d	t ha^{-1}
A1	75	254	10	12	48	4.38
A2	81	280	14	13	44	4.16
A3	86	290	16	11	42	3.77
A4	70	187	17	27	44	2.50
A5	79	241	17	17	42	2.92

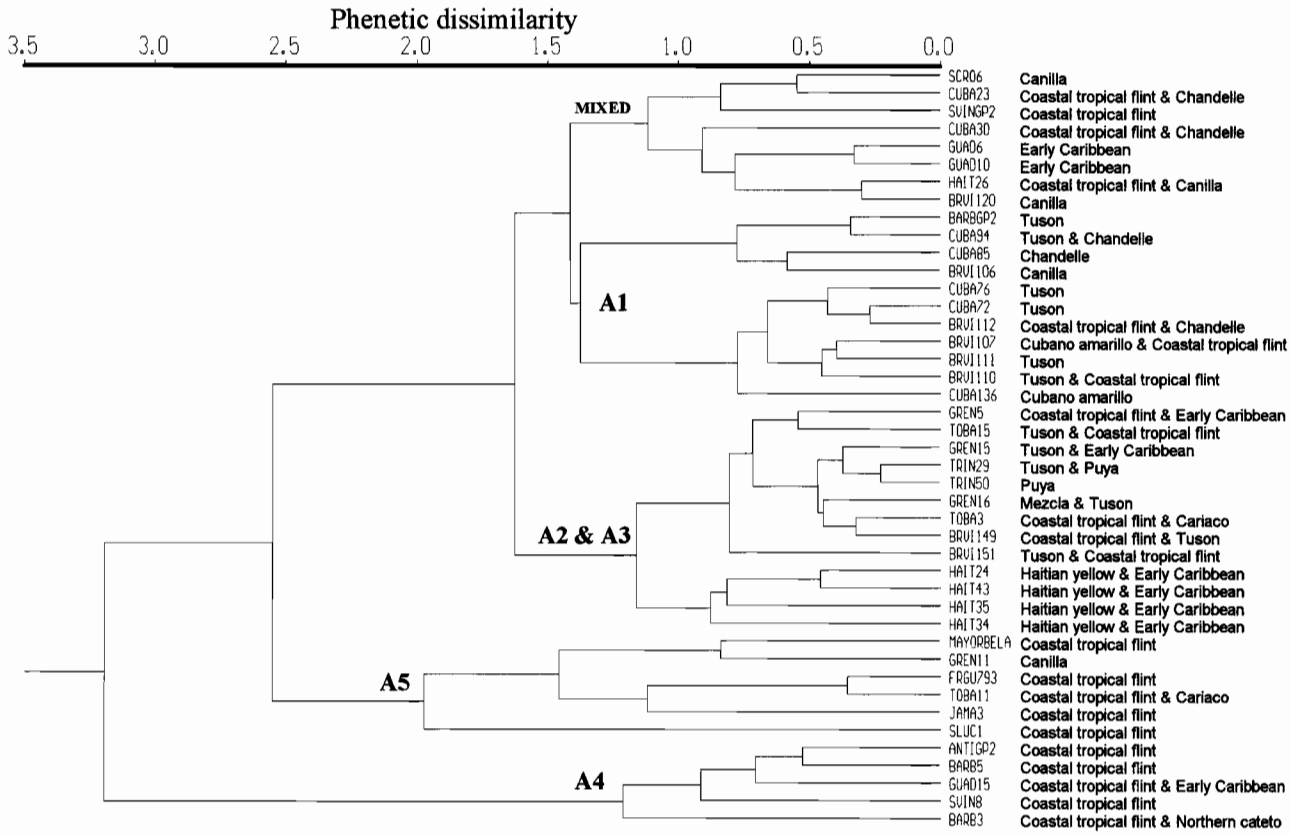

FIGURE 2 - Phenogram of relationships among 43 accessions obtained by the UPGMA cluster method representing a subset (23%) of 184 Caribbean maize accessions.

mediate maturity, but a little taller plants than Group A1. Group A3 is later and taller material with relatively low yield and considerable root lodging. Group A4 has an intermediate maturity and yield performance with lower plant height, poor root development, and a high percentage of stalk lodging. Group A5 represents a sample of those accessions that showed different performance across the environments. The dispersion in Fig. 1 indicates that the accessions can be characterized as a reflection of the variability of the groups. It is suggested that these groups can be used as source of variability for a stratified sampling process to define a subset in proportion to the number of accessions in each group as suggested by YONEZAWA *et al.* (1995) and CROSSA *et al.* (1995).

A selected subset (23%) from 184 accessions evaluated is presented in a phenogram (Fig. 2). Except for Group A5, all groups were exposed to a sampling process in proportion of their contribution to the variability. The phenogram (Fig. 2) shows the racial classification and the associated five groups. The different clusters in the phenogram are described considering the relationships among races given by GOODMAN and BROWN (1988). Six of 11 entries of Group A1 are classified as TUSON, and two as COSTCR. These two races are closely related, and are unrelated to the other three entries, CUBAAM, CANILLA, and CHANDELLE. With the exclusion of the race PUYA, Groups A2 and A3 are composed of three races that are related (HAITYE, TUSON, and COSTCR). Groups A4 and A5 are all COSTCR, except for the race CANILLA from Group A5. A "mixed" Group is identified in the phenogram, which is composed of three races: COSTCR, Early Caribbean, and CANILLA, all unrelated to each other. In general, the phenogram shows the groups separated as indicated in Fig. 1. The associated race name indicates that phenetic relationships described the accessions in a manner fairly congruent with their racial classification. The association of clusters and the groups in phenogram (Fig. 2) indicated that variability in the representative subset reflected the variability of all accessions evaluated. It must be empha-

sized that the phenogram was obtained based on phenetic dissimilarity computed from evaluation data. The racial classification is assigned when the accession is collected, or during the evaluation. Moreover, introgression of genes among racial groups makes any classification more difficult.

Considering these results, computation of phenetic relationships seems to be an appropriate method for defining a subset, particularly when agronomic and morphological information are utilized to characterize accessions at different environments. SNEATH and SOKAL (1973) pointed out that interpretation of phenograms produced by different clustering methods, differences in estimating relationships, and measures of characters in some parts of the organism or life stage, are among the main problems in phenetic classification. Utilization of many unweighted traits to compute dissimilarity relationships, followed by an appropriate clustering strategy, in part, addresses these problems. Therefore, the selected subset represents a sample of the variability attributable to phenetic dissimilarity among accessions.

REFERENCES

AJMONE-MARSAN P., C. LIVINI, M.M. MESSMER, A.E. MELCHINGER, M. MOTTO, 1992 Cluster analysis of RFLP data from related maize inbred lines of the BSSS and LSC heterotic groups and comparison with pedigree data. Euphytica **60:** 139-148.

ANDERBERG M.R., 1973 Cluster analysis for applications. Academic Press, New York.

BROWN A.H.D., 1989a The case for core collections. pp. 136-156. *In:* A.H.D. Brown, O.H. Frankel, D.R. Marshall, J.T. Williams (eds.) The use of plant genetic resources. Cambridge University Press, Cambridge, UK.

BROWN A.H.D., 1989b Core collections: a practical approach to genetic resources management. Genome **31:** 818-824.

BROWN W.L., 1960 Races of maize in the West Indies. National Academy of Sciences, National Research Council Publication no. 792. Washington, D.C.

CROSSA J., I.H. DELACY, S. TABA, 1995 The use of multivariate methods in developing a core collection. pp. 77-92. *In:* T. Hodgkin, A.H.D. Brown, Th.J.L. VanHintum, E.A.V. Morales (eds.) Core Collections of Plant Genetic Resources. John Wiley & Sons, Chichester, UK.

EVERITT B., 1980 Cluster analysis. Second edition. Halsted Press, New York.

FARRIS J.S., 1969 On the cophenetic correlation coefficient. Syst. Zool. **18:** 279-285.

GOODMAN M.M., 1972 Distance analysis in biology. Syst. Zool. **21:** 174-186.

GOODMAN M.M., R.M. BIRD, 1977 The races of maize. IV. Tentative grouping of 219 Latin American races. Econ. Bot. **31:** 204-221.

GOODMAN M.M., W.L. BROWN, 1988 Races of corn. pp. 33-79. *In:* G.F. Sprague, J.W. Dudley (eds.) Corn and corn improvement. Third edition. ASA, CSSA, and SSSA, Madison, WI.

GOODMAN M.M., E. PATERNIANI, 1969 The Races of Maize. III. Choices of appropriate characters for racial classification. Econ. Bot. **23:** 265-273.

GOWER J.C., 1971 A general coefficient of similarity and some of its properties. Biometrics **27:** 857-871.

IBPGR, 1991 Descriptors for maize. International Maize and Wheat Improvement Center, El Batan, Mexico, and International Board for Plant Genetic Resources, Rome.

JOHNSON R.A., D.W. WICHERN, 1992 Applied multivariate statistical analysis. Third edition. Prentice-Hall, Inc., New Jersey.

LANCE G.N., W.T. WILLIAMS, 1967 A general theory of classificatory sorting strategies: 1. Hierarchical systems. Comp. J. **9:** 373-380.

MOLINA F.I., P. SHEN, S.C. JONG, K. ORIKONO, 1992 Molecular evidence supports the separation of *Lentinula* edodes from *Lentinus* and related genera. Can. J. Bot. **70:** 2446-2452.

MUMM R.H., J.W. DUDLEY, 1994 A classification of 148 U.S. maize inbreds: I. Cluster analysis based on RFLPs. Crop Sci. **34:** 842-851.

MUMM R.H., L.J. HUMBERT, J.W. DUDLEY, 1994 A classification of 148 U.S. maize inbreds: II. Validation of cluster analysis based on RFLPs. Crop Sci. **34:** 852-865.

ORDÁS A., R.A. MALVAR, A.M. DE RON, 1994 Relationships among American and Spanish populations of maize. Euphytica **79:** 149-161.

PEETERS J.P., J.A. MARTINELLI, 1989 Hierarchical cluster analysis as a tool to manage variation in germplasm collections. Theor. Appl. Genet. **78:** 42-48.

ROHLF F.J., 1994 NTSYS-pc. Numerical taxonomy and multivariate analysis system. Version 1.80. Department of Ecology and Evolution, State University of New York, Stony Brook, New York.

ROHLF F.J., D.R. FISHER, 1968 Tests for hierarchical structure in random data sets. Syst. Zool. **17:** 407-412.

SÁNCHEZ J.J., M.M. GOODMAN, J.O. RAWLINGS, 1993 Appropriate characters for racial classification in maize. Econ. Bot. **47:** 44-59.

SAS INSTITUTE INC., 1989 SAS/STAT User's guide, version 6, fourth edition, volume 1 and 2. SAS Institute, Inc., Cary, NC.

SMITH O.S., J.S.C. SMITH, 1992 Measurement of genetic diversity among maize hybrids; a comparison of isozymic, RFLP, and heterosis data. Maydica **37:** 53-60.

SNEATH P.H.A., R.R. SOKAL, 1973 Numerical taxonomy. The principles and practice of numerical classification. W.H. Freeman, San Francisco.

SOKAL R.R., F.J. ROHLF, 1962 The comparison of dendrograms by objective methods. Taxon **11:** 33-40.

STINEBRICKNER R., 1984 s-Consensus trees and indices. Bull. Math. Biol. **46:** 923-935.

STUESSY T.F., 1990 Plant taxonomy. The systematic evaluation of comparative data. Columbia University Press, New York.

VANHINTUM TH.J.L., 1995 Hierarchical approaches to the analysis of genetic diversity in crop plants. pp. 23-34. *In:* T. Hodgkin, A.H.D. Brown, Th.J.L. VanHintum, E.A.V. Morales (eds.) Core Collections of Plant Genetic Resources. John Wiley & Sons, Chichester, UK.

WARD JR. J.H., 1963 Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. **58:** 236-244.

WILLIAMS W.T., 1976 The meaning of pattern. pp. 124-129. *In:* W.T. Williams (ed.) Pattern analysis in agricultural science. Elsevier. Amsterdam, Netherlands.

YONEZAWA K., T. NOMURA, H. MORISHIMA, 1995 Sampling strategies for use in stratified germplasm collections. pp. 35-53. *In:* T. Hodgkin, A.H.D. Brown, Th.J.L. VanHintum, E.A.V. Morales (eds.) Core Collections of Plant Genetic Resources. John Wiley & Sons, Chichester, UK.