Comparison of Several Clustering Methods in Grouping Kale Landraces

Guillermo Padilla, Maria Elena Cartea1, and Amando Ordás

ABSTRACT. Four clustering methods were compared for classification of a collection of 148 kale landraces (Brassica oleracea L. acephala group) from northwestern Spain based on morphologic characters: the unweighted pair group method using arithmetic averages (UPGMA) and the Ward method, hierarchical cluster algorithms, and the modified location model (MLM) applied to both the UPGMA and the Ward method (UPGMA-MLM and Ward-MLM, respectively). Comparisons were based on five criteria and on subjective considerations about the structure of each method and the characteristics of the material evaluated. Although the UPGMA-MLM was superior according to the objective criteria, its slight advantage with respect to the Ward-MLM strategy did not overcome the fact that the initial UPGMA cluster generated a classification with little value. The Ward-MLM strategy generated five homogeneous groups with defined morphologic characteristics. Moreover, the Ward-MLM strategy allowed the identification of redundant landraces, which would permit the number of accessions in further critical trials to be reduced.

To preserve and characterize Galician (northwestern Spain) landraces of the genus Brassica L., a germplasm bank has been created at the Misión Biológica de Galicia (MBG), part of the Spanish Council for Scientific Research. To date, 246 Brassica oleracea, 207 Brassica rapa L., and 50 Brassica napus L. accessions have been collected, representing most of the region. The B. oleracea collection includes kale (B. oleracea acephala group), cabbage (B. oleracea capitata group), and Portuguese Tronchuda cabbage (B. oleracea costata group). Kale, closely related to wild B. oleracea (Song et al., 1990), is by far the most well-represented variety of the species (214 accessions), being an important vegetable and forage crop in the traditional farming systems of Galicia. It was, until recently, an important vegetable grown of the region and still is the only crop suitable for marginal and mountainous areas.

A critical aim of genetic resource conservation programs is to identify redundant genotypes in the collection and to select those genotypes that best represent the gene collection with the minimum loss of genetic diversity (Crossa and Franco, 2004). Previous studies have attempted to characterize the MBG kale collection, but these only examined representatives from a small number of accessions and considered traits that could be characterized as continuous variables (Cartea et al., 2003; Picoaga et al., 2003). Several works on Portuguese coles, including “Galician kales” landraces, have been reported based on agronomic traits (Dias et al., 1993), isozyme differences (Dias et al., 1994), and restricted fragment length polymorphism (RFLP) markers (Dias and Monteiro, 1994). These works have shown that kale landraces have an extremely high intravariability and are thus difficult to classify into natural groups. This is not unexpected for an allogamous species in which rudimentary seed production methods are followed (Cartea et al., 2003; Dias and Monteiro, 1994). Intercrossing among landraces can produce highly heterogeneous cultivars that may have intermated some degree (Dias and Monteiro, 1994). Grouping accessions, therefore, is difficult because variation within landraces may be as great as among them.

A characterization trial of a large number of landraces for mixed variables yields a multivariate data set, which can be confusing if the classification method is inappropriate. The choice of the method that best reflects preconceptions produces a highly subjective result. Because each method has some distinctive features and properties, it is better to consider the desired objective at the outset and to select the approach that best meets this objective (Krzanowski, 2000). On the basis of agronomic and morphologic characters, the unweighted pair group method using arithmetic averages (UPGMA) is most commonly used (Mohammadi and Prasanna, 2003). This method has been effective in grouping kale populations (Cartea et al., 2003; Dias et al., 1993). In studies that compare the UPGMA with other hierarchical methods, the former showed a higher estimation of the cophenetic correlation coefficient and demonstrated genotype associations (Mohammadi and Prasanna, 2003; Rincón et al., 1996). However, in clusters based on molecular data, the Ward method (Ward, 1963) has been more suitable because it avoided the “chaining” effects often seen with UPGMA (Dubreuil et al., 1996; Lombard et al., 2000; Mohammadi and Prasanna, 2003). Franco et al. (1997a, 1997b) extended the comparison between cluster methods and statistical models. They concluded that the Normix model (Wolfé, 1970) applied to groups formed by the Ward method was an appropriate strategy for grouping maize (Zea mays L.) accessions into relatively homogeneous groups compared with other methods. Later, Franco et al. (1998) improved the Normix method creating the modified location model (MLM), which allows the use of categorical traits in addition to other improvements.

In this study, several criteria are used to compare UPGMA, the Ward method, and MLM applied to the groups created by the two hierarchical methods (i.e., UPGMA-MLM and Ward-MLM) as clustering techniques to classify the collection.

Received for publication 3 Nov. 2006. Accepted for publication 14 Mar. 2007. Research supported by the National Plan for Research and Development (AGL2003-01366) and Excmia. Diputación Provincial de Pontevedra. G. Padilla acknowledges a fellowship from the Cabildo Insular de La Palma. We thank Prof. J. Franco for providing statistical programs used in this work.

*Corresponding author. E-mail: ecartea@mbg.csic.es.
of kales from the MBG germplasm bank. All four classifications are based on the same morphologic and agronomic traits. These criteria, based on statistical properties, are contrasted with more subjective considerations about the material studied. After choosing the most appropriate classification method, redundant phenotypes can be identified. This will allow performing further critical tests with less landraces.

The objectives of this study were: 1) to compare four classification methods from a statistical point of view and to test the results with some subjective considerations and 2) to identify redundant genotypes in the kale collection and characterize this germplasm.

Materials and Methods

**Plant material.** One-hundred forty-eight kale populations from Galicia and six commercial controls were grown at the MBG in Pontevedra, Spain (lat. 42°24′ N, long. 8°38′ W) 20 m above sea level in acidic sandy loam soil. Seeds were sown in multistop trays and seedlings were transplanted into the field at the five- to six-leaf stage in Apr. 2003. Populations were evaluated by a Petersen augmented design (Petersen, 1985) with six controls in a randomized complete block design with four replications. Therefore, four blocks were formed, each one having 37 different landraces and a replication of the controls. The experimental plots consisted of two rows with 10 plants in each row. Rows were spaced 0.6 m apart and plants within rows were 0.6 m apart. Before transplanting, the soil was fertilized with 57 kg ha⁻¹ N, P, and K. After transplanting, 30 kg ha⁻¹ N was applied. Weeds were controlled mechanically. Plants were treated with diazinon to combat cabbage root fly (*Delia radicum*).

Data were collected on 18 morphologic and agronomic traits following the International Board for Plant Genetic Resources (IBPGR) *Brassica* and *Raphanus* L. descriptors list (IBPGR, 1990), actually named International Plant Genetic Resources Institute (IPGRI) (Table 1). Raw data were adjusted for the block effect using the GLM procedure of SAS (version 8.0; SAS Institute, Cary, N.C.) (Petersen, 1985; Wolfinger et al., 1997).

**Statistical analysis.** A dissimilarity matrix was created with the Gower general coefficient (Gower, 1971), which combines continuous variables (numeric and ordinal alphanumeric variables) and categorical variables (nominal alphanumeric variables, both binary and multistate). This coefficient includes several distances as special cases (Gower, 1971). These are the Jaccard coefficient for binary attributes, the simple matching coefficient for multistate traits, and the Manhattan distance for continuous traits (Romesburg, 1984).

Two hierarchical cluster algorithms, UPGMA and Ward’s minimum variance method, gave two different classifications. Both were implemented with the MLM method (Franco et al., 1998). All four classifications, UPGMA, Ward, UPGMA-MLM, and Ward-MLM, were derived from the procedures CLUSTER and IML (SAS Institute, 2000) following a set of programs created by Franco et al. (1998). The optimal number of groups generated by the hierarchical methods was determined with the upper tail approach (Mojena, 1977). Crossa and Franco (2004) consider that a fusion value greater than the 95th percentile marks the optimum number of clusters. For the MLM method, the point of the likelihood profile where maximum growth occurs was the optimal group number (Crossa and Franco, 2004). Five criteria were used to compare the four classifications. The influence of categorical variables on the classification was evaluated by 1) the χ² test of independence and by 2) the affinity among groups for categorical traits. The influence of continuous variables was determined by 3) the number of continuous variables effectively used in the formation of groups, 4) the value of det(A) and det(T) ratio and the value of the Hotelling-Lawley trace, and 5) the Mahalanobis distance among clusters. Each criterion is explained as follows:

**Criterion 1:** For each categorical variable, two-way tables of “group by values of the variable” were produced and the observed and the

Table 1. Traits used for a preliminary classification of a collection of kale landraces from northwestern Spain. *

| Traits | Description |
| --- | --- |
| **A. Continuous (numeric and ordinal) traits** | |
| Early vigor | Subjective scale from 1 (very poor) to 5 (excellent) |
| Leaf length | Measurement of the whole leaf (cm) |
| Fresh matter | Average fresh weight of a single leaf (g) |
| Secondary stems | Average number of secondary stems per plant |
| Plant height | Distance from the soil to the top of the plant (cm) |
| Vegetative stem length | Distance from soil to the highest point on vegetative or preflowering apex (cm) |
| Plant diameter | Diameter of the widest point on plant at the beginning of the flowering period (cm) |
| Stem width | Diameter of the widest point on stem at the beginning of the flowering period (mm) |
| Leaves per plant main stem | Average number of leaves per main stem |
| Leaf scars per plant main stem | Average number of leaf scars in 20 cm of the main stem of a plant at medium height |
| Time to flowering | Days from transplanting until 50% of plants have flowered |
| **B. Categorical traits** | |
| Leaf color | Score (3 = light green, 5 = green, 7 = dark green) |
| Leaf bloom | Score (3 = low, 5 = intermediate, 7 = high) |
| Leaf blistering | Score (0 = none, 3 = low, 5 = intermediate, 7 = high) |
| Leaf lobes | 0 (absent), 1 (present) |
| Leaf lamina attitude | 0 (concave), 1 (convex) |
| First flowering habit | Score (1 = terminal, 2 = axillary) |
| Flower color | Score (1 = white, 2 = yellow, 3 = mixed) |

*Eleven traits are continuous and seven are categorical.*
expected values for each cell were calculated (Franco et al., 1997b). The procedure PROC FREQ of SAS was used. 

Criterion 2: The affinity among groups was measured as \(\sum_{i=1}^{S}(p_{ia}p_{ja})^{1/2}\) where \(p_{ia}\) and \(p_{ja}\) are, respectively, the proportion of individuals in the groups \(i\) and \(j\) sharing the \(s\)th level for the multinomial variable that joins all categorical variables (Franco et al., 1998). This criterion is given as the number of pairs of clusters that share any particular \(s\)th level.

Criterion 3: The STEPDISC procedure of SAS was used to perform a stepwise discriminating analysis to select a subset of the continuous variables that could be used in discriminating among the clusters (Franco et al., 1997b).

Criterion 4: \(T\) and \(A\) are the matrices of sum of squares and crossproducts corresponding to the total variability and the variability within groups, respectively. The Hotelling-Lawley trace is the trace of the product \(A^{-1}B\), with \(B\) the matrix of between-group variability. The smallest value of \(det(A)\) and \(det(T)\) ratio and the largest value of the Hotelling-Lawley trace gives the best clustering (Crossa and Franco, 2004). The CANDISC and IML procedures of SAS were used to perform these analyses.

Criterion 5: The best strategy would produce the largest Mahalanobis distance among clusters (Crossa and Franco, 2004; Franco et al., 1997b). This value was obtained with the CANDISC procedure of SAS.

With the results of each clustering, a canonical analysis was performed. The graphic representation of this analysis allows better differentiation among groups when the first two canonical variables are plotted in a diagram. The correlation coefficients between the original traits and the canonical variables are plotted in a diagram. The correlation allows better differentiation among groups when the first two canonical variables are plotted in a diagram. The Hotelling-Lawley trace is the trace of the product \(A^{-1}B\), with \(B\) the matrix of between-group variability. The smallest value of \(det(A)\) and \(det(T)\) ratio and the largest value of the Hotelling-Lawley trace gives the best clustering (Crossa and Franco, 2004). The CANDISC and IML procedures of SAS were used to perform these analyses.

Results and Discussion

Of the 18 traits assessed, 11 were continuous variables and seven were categorical variables (Table 1). All of the categorical variables were considered nominal variables. Five local cultivars out of the 148 evaluated had not flowered when the trial was concluded (15 months after transplanting) and they were not included in the analysis.

Eight groups was the optimal number determined by the UPGMA and the Ward method using the upper tail approach method (Mojena, 1977). Although the optimal number of clusters obtained was the same, their structure clearly differs. The number of landraces in each group is shown in Table 2. The differences result from inherent properties of each method. The Ward method tends to force data into clusters of equal size (Krzanowski, 2000). Thus, the eight clusters formed by the Ward method are more balanced in size than the clusters obtained with the UPGMA. Homogeneous groups do not mean a better classification per se. In fact, small groups could be an indication of the existence of variants in the collection (Franco et al., 1997b). Gouesnard et al. (1997) classified maize populations by the UPGMA and Ward methods and obtained similar results. UPGMA generated a main cluster and many small clusters, whereas the Ward method built clusters of similar size subdividing the big cluster generated by UPGMA into several groups. With respect to continuous traits, these small groups did not show an exclusive range of values (Fig. 1). Leaf color was the only categorical trait with a level (purple-green), which is exclusively represented in two small groups (data not shown).

| Cluster | UPGMA | Ward | UPGMA-MLM | Ward-MLM |
|---------|-------|------|------------|----------|
| A       | 79    | 19   | 62         | 20       |
| B       | 2     | 6    | 6          | 51       |
| C       | 4     | 18   | 11         | 38       |
| D       | 12    | 12   | 19         | 25       |
| E       | 4     | 15   | 17         | 9        |
| F       | 1     | 12   | 10         | —        |
| G       | 23    | 30   | 10         | —        |
| H       | 18    | 31   | 5          | —        |
| I       | —     | —    | 1          | —        |
| J       | —     | —    | 2          | —        |

Comparing the results of both hierarchical cluster techniques, mean variation among clusters was larger and variability within each cluster was, generally, narrower for each continuous trait with UPGMA than with the Ward method (Fig. 1). Krzanowski (2000) observed the high internal affinity that clusters created by UPGMA exhibit compared with other methods. In this study, the higher internal affinity of the UPGMA compared with the Ward method is the result of the cluster structure of each method. The presence of several small clusters (four or less accessions) in the UPGMA classification allows detecting larger differences among clusters and smaller variations within clusters.

Clusters formed by hierarchical methods were used to carry out the MLM. Using the point of the likelihood profile where the maximum growth was observed to settle the optimal number of groups, the UPGMA-MLM strategy fixed 10 groups and the Ward-MLM strategy fixed five groups. Because these differences were observed for both cluster composition and number, differences in the cluster structure become stronger after applying the MLM. The UPGMA-MLM strategy formed a large cluster with 62 landraces and six small groups containing 10 units or fewer (Table 2). Landraces comprising the smallest groups of the UPGMA and the UPGMA-MLM clusters were not coincident (data not shown), indicating the low consistence of the possible variants. The Ward-MLM strategy reduced the number of groups from eight to five compared with the Ward method (Table 2). The reasons for all these different patterns among classification methods could be attributable to the high variability within populations and the resemblance between accessions along with the relatively low number of characters observed.

With respect to continuous traits, the variation of means among clusters was higher for all traits, except leaf length, when comparing the UPGMA-MLM strategy with UPGMA. Mean variation was also greater for all traits in the Ward-MLM strategy, except for time to flowering, compared with the Ward method (Fig. 1). This indicates that the addition of the MLM strategy enhanced the differences among clusters for continuous traits.

When applying MLM to clusters obtained by hierarchical clustering methods, all categorical traits were reduced to a
multinomial variable (W) with 70 levels, each level being a combination of states of all the categorical variables. Only two levels of W were present in a significant number of landraces (data not shown). The first level, found in 18 landraces, included medium bloomed and lobed dark green leaves with low blade blistering and concave lamina, terminal first flowering habit, and white flowers. The second level, present in 15 landraces, had the same characteristics except for leaf blade and for the first flowering habit. UPGMA-MLM divided the 18 populations with the first level into two clusters, whereas Ward-MLM divided these populations into three clusters. Regarding the second level, the UPGMA-MLM strategy split the 15 populations into four clusters and the Ward-MLM strategy into three clusters. Comparing the use of MLM with the hierarchical clustering methods, UPGMA grouped together all landraces with both levels in just one cluster, whereas the Ward method joined all landraces with one of each level in the same cluster. This means that the MLM reduced the importance of the categorical traits in the clustering process. This is compensated for by a better classification of the continuous traits.

According to criterion 1, leaf lobe was the only categorical trait with no significance in any method, indicating that this variable has no influence in the classification (Table 3). Because all methods had the same number of significant categorical traits, this criterion did not discriminate among methods. Moreover, most of the distributions tested had fewer than five observations, so $\chi^2$ tests may not be valid. Hierarchical
methods formed clusters with less affinity for the categorical traits (criterion 2) with only one pair of clusters sharing a W multinomial variable level for each method (Table 3). The suggestion that categorical traits lose influence in the classification in favor of continuous traits when applying MLM is supported by this result. Overlapping of W levels among clusters occurred in both UPGMA-MLM and Ward-MLM strategies with nine and eight pairs of clusters, respectively (Table 3).

The highest number of relevant continuous traits (criterion 3) for use in discriminating among the groups occurred in UPGMA-MLM, with all 11 continuous variables selected, followed by Ward-MLM with 10 (Table 3). Six continuous traits had a significant influence on the classification in all methods: leaf length, fresh matter, secondary stems, leaves per plant, leaf scars per plant, and time to flowering. The UPGMA-MLM strategy showed the smallest value for det(A) and det(T) ratios along with the largest value for the Hotelling-Lawley trace (criterion 4), suggesting greater differences among the means of the continuous traits for the formed groups and, therefore, better clustering. However, UPGMA-MLM, with 10 clusters of unequal sizes, presented values not very different from those obtained with Ward-MLM with five clusters. Compared with both hierarchical methods, Ward-MLM showed better values according to criterion 4 despite forming less clusters (Table 3). Both UPGMA and Ward showed an average Mahalanobis distance shorter than their respective MLM implementations (criterion 5). Similar results were also obtained by Franco et al. (1997a, 1997b) when classifying maize accessions. UPGMA-MLM had the largest average Mahalanobis distance between clusters (35.7), although the main contribution to its value came from clusters with one and two populations, respectively. When these clusters were removed, the average Mahalanobis distance between clusters was 20.9, very close to the value obtained with Ward-MLM (18.8). It is noteworthy that Ward-MLM had the largest

Fig. 1. Continued
minimum Mahalanobis distance between clusters of all four classification methods.

Plots of the first two canonical variables given by the canonical analyses are shown in Figure 2. Ward-MLM is the method that best distinguished the different groups, displaying clusters that were more compact with less overlap. The first canonical variable was highly and positively correlated with number of leaves and leaf scars per plant in all classification methods except for UPGMA-MLM, in which correlation of the first canonical variable for these two traits was highly negative (Table 4). Therefore, these were the most relevant variables for the formation of groups (Franco et al., 1997a). Correlation coefficients of the second canonical variable with the original continuous traits changed across classifications. Time to flowering had the highest negative correlation with this second canonical variable in UPGMA and the highest positive correlation in the Ward method. Number of secondary stems per plant was the most correlated trait with the second canonical variable in the UPGMA-MLM and Ward-MLM strategies. In the Ward method and the Ward-MLM strategy, the two first canonical variables explained more than 70% of the variation between groups, whereas in UPGMA and UPGMA-MLM, the total variation explained 65% and 51%, respectively (Table 4). Therefore, classifications based on Ward and Ward-MLM are more appropriate to visualize relationships among groups and the variables associated with the axes when plotting the first two canonical variables.

MLM improves both hierarchical methods for all criteria but criterion 2 related to categorical traits. Criteria 3, 4, and 5 showed that UPGMA-MLM was more effective than Ward-MLM, but the slight advantage obtained with respect to Ward-MLM did not overcome more subjective considerations. The structure of the classification created by the hierarchical methods is related to their features and properties. UPGMA tends to create clusters with high internal affinity (Krzanowski, 2000). In this study, because of lack of isolation in the multiplication of the landraces, there is a mixture of morphologic characteristics across accessions. Because of this, UPGMA grouped 55% of the accessions into one cluster because UPGMA is more prone to chaining effects than the Ward method (Dubreuil et al., 1996; Lombard et al., 2000; Mohammadi and Prasanna, 2003). Another consequence of the higher internal affinity of the clusters generated by UPGMA is that the small groups must represent relevant variants with a possible valuable genetic diversity. However, phenotypic properties of accessions belonging to small groups did not have this specific genetic value as they split off when MLM was implemented. Moreover, previous studies concluding that UPGMA was a better classification technique than the Ward method were based on the capability of UPGMA for recovering groups (Gouesnard et al., 1997; Peters and Martinelli, 1989; Rincón

Table 3. Criteria for comparing clusters formed by each method (UPGMA, Ward, UPGMA-MLM, and Ward-MLM methods) used for classifying a collection of kale landraces from northwestern Spain.

| Criterion | UPGMA | Ward | UPGMA-MLM | Ward-MLM |
|-----------|-------|------|-----------|----------|
| 1         |       |      |           |          |
| LC        | **    | **   | **        | **       |
| LBLO      |      | **   | **        | **       |
| LBLI      |      | **   |           |          |
| LL        |      |      | NS        | NS       |
| LLA       |      |      | **        | **       |
| FFH       |      |      | **        | **       |
| FC        |      |      | **        | **       |
| 2         | 1     | 1    | 9         | 8        |
| Min.      | 8     | 6    | 11        | 10       |
| Max.      | 35.6  | 15.1 | 97.9      | 25.8     |
| Mean      | 16.5  | 5.7  | 35.7      | 18.8     |

*UPGMA = unweighted pair group method using arithmetic averages; Ward = Ward’s minimum variance method; UPGMA-MLM = modified location model applied to groups formed by UPGMA; Ward-MLM = modified location model applied to groups formed by Ward’s method.

**Criterion 1 = \( \chi^2 \) test of independence; criterion 2 = affinity among groups for categorical traits given as the number of clusters that share any particular level of the multinomial variable that joins all categorical variables; criterion 3 = number of continuous variables effectively used in the formation of groups; criterion 4 = value of \( \text{det}(A) \) and \( \text{det}(T) \) ratio and value of the Hotelling-Lawley trace; criterion 5 = minimum, maximum, and mean values of the Mahalanobis distance among clusters.

*LC = leaf color; LBLO = leaf bloom; LBLI = leaf blistering; LL = leaf lobes; LLA = lamina leaf attitude; FFH = first flowering habit; FC = flower color.

**\( |A|/|T| \) = \( \text{det}(A) \) and \( \text{det}(T) \) ratio; H-L trace = Hotelling-Lawley trace.

***Min. and Max. = minimum and maximum values, respectively.

Nonsignificant or significant of qualitative variables at \( P \leq 0.05 \) or 0.01, respectively.

Fig. 2. Plots of the first two canonical variables (C1 and C2) on the clusters of kale landraces from northwestern Spain formed by the unweighted pair group method using arithmetic averages [UPGMA (A–H)] and Ward’s minimum variance method [Ward (A–H)] hierarchical methods, and the modified location model applied to groups formed by UPGMA [UPGMA-MLM (A–J)] and modified location model applied to groups formed by Ward’s method [Ward-MLM (A–E)] strategies for a graphic representation of the internal variability within each cluster.
et al., 1996), but in this case, accessions are not divided into races or any other taxonomic division, which could be recovered. Besides this consideration, the Ward method correlates better with the MLM than the UPGMA and also complements the MLM cluster structure (Crossa and Franco, 2004; Franco et al., 1998). Another advantage of the Ward-MLM in

### Table 4. Correlation coefficients of continuous variables with the first two canonical variables (Can1 and Can2) on the clusters of kale landraces from northwestern Spain formed by the UPGMA and Ward hierarchical methods and UPGMA-MLM and Ward-MLM strategies.

| Traits                  | UPGMA Can1 | UPGMA Can2 | Ward Can1 | Ward Can2 | UPGMA-MLM Can1 | UPGMA-MLM Can2 | Ward-MLM Can1 | Ward-MLM Can2 |
|-------------------------|------------|------------|-----------|-----------|----------------|----------------|---------------|---------------|
| Early vigor             | 0.02       | 0.42       | 0.36      | 0.04      | -0.17          | 0.57           | 0.06          | 0.21          |
| Leaf length             | -0.41      | 0.08       | -0.60     | 0.20      | 0.44           | -0.05          | -0.35         | 0.26          |
| Fresh matter            | -0.19      | 0.41       | -0.14     | -0.57     | 0.15           | -0.06          | -0.34         | -0.13         |
| Secondary stems         | -0.20      | -0.01      | -0.16     | -0.02     | 0.08           | 0.83           | -0.07         | 0.34          |
| Plant height            | -0.12      | 0.47       | 0.65      | 0.05      | -0.16          | 0.36           | 0.00          | -0.20         |
| Vegetative stem length  | -0.38      | 0.30       | -0.74     | 0.26      | 0.21           | 0.42           | -0.27         | 0.21          |
| Plant diameter          | 0.05       | -0.02      | 0.56      | 0.52      | 0.00           | 0.32           | 0.35          | 0.30          |
| Stem width              | 0.25       | -0.09      | 0.66      | 0.60      | -0.09          | 0.02           | 0.33          | 0.23          |
| Leaves per plant        | 0.79       | -0.17      | 0.87      | 0.49      | -0.62          | -0.08          | 0.84          | 0.11          |
| Leaf scars per plant    | 0.79       | 0.06       | 0.95      | 0.20      | -0.61          | -0.30          | 0.81          | 0.04          |
| Time to flowering       | 0.01       | -0.67      | -0.32     | 0.90      | 0.40           | -0.42          | -0.04         | 0.31          |

Explained variation between groups:

| UPGMA | Ward | UPGMA-MLM | Ward-MLM |
|-------|------|-----------|----------|
| 0.40  | 0.58 | 0.31      | 0.42     |
| 0.25  | 0.19 | 0.20      | 0.29     |

*UPGMA = unweighted pair group method using arithmetic averages; Ward = Ward’s minimum variance method; UPGMA-MLM = modified location model applied to groups formed by UPGMA; Ward-MLM = modified location model applied to groups formed by Ward’s method.

Fig. 3. Phenogram of 143 kale landraces from northwestern Spain obtained with the Ward minimum variance method. Dissimilarity matrix was based in the Gower general coefficient.
comparison with UPGMA-MLM is that its canonical analysis allows clearer differentiation of the clusters and the first two canonical variables used in the graphic representation explained much of the variation between groups (71% for Ward-MLM and 51% for UPGMA-MLM). Therefore, because of the nature of the material and the differences in the cluster structure, Ward-MLM was selected for the classification of the collection.

The classification obtained by Ward-MLM assigned each landrace to the cluster where it had the maximum probability of membership (data not shown). The most correlated continuous traits with the first two canonical variables, along with the categorical trait levels with the highest frequencies, defined the characteristics of each cluster (A–E). Cluster A included landraces with high vegetative stem length, high number of secondary stems, and high fresh weight but a low number of leaves and leaf scars per plant. Their leaf blades had no blistering and the leaf lamina attitude was concave. Cluster A landraces had axillary first flowering habit and flowers were mostly white. Cluster B was characterized by a low number of secondary stems, white flowers, and terminal first flowering habit. Most cultivars in cluster C had a low fresh weight and a high number of secondary stems with mixed flower color. Cluster D grouped cultivars with green leaves, low bloom content, and convex lamina. Cluster D also included taller plants and earlier flowering with cultivars having yellow flowers. Finally, cluster E included the shortest landraces as well as those with the highest number of scars and leaves per plant and flowers that were yellow or mixed color.

Ward-MLM also identified redundant genotypes in the collection. The Ward method set the initial groups that were used by MLM. Although the phenogram illustrating the classification obtained by the Ward method does not show the final cluster, the landraces that remain in the same group when applying MLM can be showed (Fig. 3). Within certain distance in the phenogram, cultivars share the same level of the multinominal variable W (i.e., the same categorical traits). Cultivars, thereby grouped and belonging to the same cluster after performing MLM, could be redundant. In this case, 47 cultivars (grouped in 16 pairs), one group of three cultivars, and three groups of four cultivars could be reduced to 20 cultivars if one unit is selected from each group (Fig. 3). The collection, therefore, can be reduced to 121 accessions, which is adequate for further critical trials.
The Ward-MLM strategy was the best classification method to create distinct groups in a collection of kale landraces from northwestern Spain. Redundant genotypes could be identified, resulting in better management and rationalizing the genetic diversity of the collection.

**Literature Cited**

Cartea, M.E., A. Picoaga, P. Soengas, and A. Ordás. 2003. Morphological characterization of kale populations from northwestern Spain. Euphytica 129:25–32.

Crossa, J. and J. Franco. 2004. Statistical methods for classifying genotypes. Euphytica 137:19–37.

Dias, J.S. and A.A. Monteiro. 1994. Taxonomy of Portuguese Tronchuda cabbage and Galega kale landraces using morphological characters, nuclear RFLPs, and isozyme analysis: A review. Euphytica 79:115–126.

Dias, J.S., A.A. Monteiro, and S. Kresovich. 1994. Genetic diversity and taxonomy of Portuguese Tronchuda cabbage and Galega kale landraces using isozyme analysis. Euphytica 75:221–230.

Dias, J.S., A.A. Monteiro, and M.B. Lima. 1993. Numerical taxonomy of Portuguese tronchuda cabbage and Galega kale landraces using morphological characters. Euphytica 69:51–68.

Dubreuil, P., P. Dufour, E. Krecji, M. Causse, D. de Vienne, A. Gallais, and A. Charcosset. 1996. Organization of RFLP diversity among inbred lines of maize representing the most significant heterotic groups. Crop Sci. 36:790–799.

Franco, J., J. Crossa, J. Diaz, S. Taba, J. Villaseñor, and S.A. Eberhart. 1997a. A sequential clustering strategy for classifying gene bank accessions. Crop Sci. 37:1656–1662.

Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1997b. Classifying Mexican maize accessions using hierarchical and density search methods. Crop Sci. 37:972–980.

Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A. Eberhart. 1998. Classifying genetic resources by categorical and continuous variables. Crop Sci. 38:1688–1696.

Gouesnard, B., J. Dallard, A. Panouillac, and A. Boyat. 1997. Classification of French maize populations based on morphological traits. Agronomie 17:491–498.

Gower, J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27:857–874.

International Board for Plant Genetic Resources (IBPGR). 1990. Descriptors for *Brassica* and *Raphanus*. International Board for Plant Genetic Resources, Rome.

Krzanowski, W.J. 2000. Principles of multivariate analysis. Oxford University Press, New York.

Lombard, V., C.P. Baril, P. Dubreuil, F. Blouet, and D. Zhang. 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Sci. 40:1459–1469.

Mohammadi, S.A. and B.M. Prasanna. 2003. Analysis of genetic diversity in crop plants. Salient statistical tools and considerations. Crop Sci. 43:1235–1248.

Mojena, R. 1977. Hierarchical grouping methods and stooping rules: An evaluation. Computer J. 20:359–363.

Peters, J.P. and J.A. Martinelli. 1989. Hierarchical cluster analysis as a tool to manage variation in germplasm collection. Theor. Appl. Genet. 78:42–48.

Petersen, R.G. 1985. Augmented designs for preliminary yield trials. RACHIS 4:27–32.

Picoaga, A., M.E. Cartea, P. Soengas, L. Monetti, and A. Ordás. 2003. Resistance of kale populations to lepidopterous pests in northwestern Spain. J. Econ. Entomol. 96:143–147.

Rincón, F., B. Johnson, J. Crossa, and S. Traba. 1996. Cluster analysis, an approach to sampling variability in maize accessions. Maydica 41:307–316.

Romesburg, H.C. 1984. Cluster analysis for researchers. Wadsworth, Belmont, Calif.

SAS Institute, 2000. SAS Online Doc, ver. 8. SAS Institute, Inc., Cary, N.C.

Song, K.M., T.C. Osborn, and P.H. Williams. 1990. *Brassica* taxonomy based on nuclear restriction fragment length polymorphisms (RFLPs), 3- Genome relationships in *Brassica* related genera and the origin of *B. oleracea* and *B. rapa* (syn. *campestris*). Theor. Appl. Genet. 79:497–506.

Ward, J. 1963. Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assn. 58:236–244.

Wolfe, J.H. 1970. Pattern clustering by multivariate mixture analysis. Multivariate Behav. Res. 5:329–350.

Wolfinger, R.D., W.T. Federer, and O. Cordero-Brana. 1997. Recovering information in augmented designs, using SAS PROC GLM and PROC MIXED. Agron. J. 89:856–859.