Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms

Alberto Fernández

Universitat Rovira i Virgili, Tarragona, Spain

Sergio Gómez

Universitat Rovira i Virgili, Tarragona, Spain

Abstract: In agglomerative hierarchical clustering, pair-group methods suffer from a problem of non-uniqueness when two or more distances between different clusters coincide during the amalgamation process. The traditional approach for solving this drawback has been to take any arbitrary criterion in order to break ties between distances, which results in different hierarchical classifications depending on the criterion followed. In this article we propose a variable-group algorithm that consists in grouping more than two clusters at the same time when ties occur. We give a tree representation for the results of the algorithm, which we call a *multidendrogram*, as well as a generalization of the Lance and Williams' formula which enables the implementation of the algorithm in a recursive way.

Keywords: Agglomerative methods; Cluster analysis; Hierarchical classification; Lance and Williams' formula; Ties in proximity.

1. Introduction

Clustering methods group individuals into groups of individuals or *clusters*, so that individuals in a cluster are close to one another. In agglomerative hierarchical clustering (Cormack 1971; Sneath and Sokal 1973, Sec. 5.5; Gordon 1999, Chap. 4), one begins with a proximity matrix between

individuals, each one forming a singleton cluster. Then, clusters are themselves grouped into groups of clusters or *superclusters*, the process being repeated until a complete hierarchy is formed. Among the different types of agglomerative methods we find single linkage, complete linkage, unweighted average, weighted average, etc., which differ in the definition of the proximity measure between clusters.

Except for the single linkage case, all the other clustering techniques suffer from a non-uniqueness problem, sometimes called the *ties in proximity* problem, which is caused by ties either occurring in the initial proximity data or arising during the amalgamation process. From the family of agglomerative hierarchical methods, complete linkage is more susceptible than other methods to encounter ties during the clustering process, since it does not originate new proximity values different from the initial ones. With regard to the presence of ties in the original data, they are more frequent when one works with binary variables, or even with integer variables comprising just some few distinct values. But they can also appear using continuous variables, specially if the precision of experimental data is low. Sometimes, on the contrary, the absence of ties might be due to the representation of data with more decimal digits than it should be done. The non-uniqueness problem also depends on the measure used to obtain the proximity values from the initial variables. Moreover, in general, the larger the data set, the more ties arise (MacCuish, Nicolaou, and MacCuish 2001).

The ties in proximity problem is well-known from several studies in different fields, for example in biology (Hart 1983; Backeljau, De Bruyn, De Wolf, Jordaens, Van Dongen, and Winnepenninckx 1996; Arnau, Mars, and Marín 2005), in psychology (Van der Kloot, Spaans, and Heiser 2005), or in chemistry (MacCuish et al. 2001). Nevertheless, this problem is frequently ignored in software packages (Morgan and Ray 1995; Backeljau et al. 1996; Van der Kloot et al. 2005), and those packages which do not ignore it fail to adopt a common standard with respect to ties. Many of them simply break the ties in any arbitrary way, thus producing a single hierarchy. In some cases the analysis is repeated a given number of times with randomized input data order, and then additional criteria can be used for selecting one of the possible solutions (Arnau et al. 2005). In other cases, some requirements are given on the number of individuals and the number of characteristics needed to generate proximity data without ties (Hart 1983; MacCuish et al. 2001). None of these proposals can ensure the complete absence of ties, neither can all their requirements be satisfied always.

Another possibility for dealing with multiple solutions is to use further criteria, like a distortion measure (Cormack 1971, Table 3), and select the best solution among all the possible ones. However, the result of this approach will depend on the distortion measure used, which means that

an additional choice must be made. But this proposal does not ensure the uniqueness of the solution, since several candidate solutions might share the same minimum distortion value. Besides, in ill conditioned problems (those susceptible to the occurrence of too many ties), it is not feasible to perform an exhaustive search for all possible hierarchical classifications, due to its high computational cost. With regard to this, Van der Kloot et al. (2005) analyze two data sets using many random permutations of the input data order, and with additional criteria they evaluate the quality of each solution. They show that the best solutions frequently emerge after many permutations, and they also notice that the goodness of these solutions necessarily depends on the number of permutations used.

An alternative proposal is to seek a hierarchical classification which describes common structure among all the possible solutions, as recommended by Hart (1983). One approach is to prune as little as possible from the classifications being compared to arrive at a common structure such as the maximal common pruned tree (Morgan and Ray 1995). Care must be taken not to prune too much, so this approach can be followed only when the number of alternative solutions is small and they are all known. Furthermore, the maximal common pruned tree need not be uniquely defined and it does not give a complete classification for all the individuals under study.

What we propose in this article is an agglomerative hierarchical clustering algorithm that solves the ties in proximity problem by merging into the same supercluster all the clusters that fall into a tie. In order to do so we must be able to calculate the distance separating any two superclusters, hence we have generalized the definition of distance between clusters to the superclusters case, for the most commonly used agglomerative hierarchical clustering techniques. Additionally, we give the corresponding generalization of Lance and Williams' formula, which enables us to compute these distances in a recursive way. Moreover, we introduce a new tree representation for the results obtained with the agglomerative algorithm: the multidendrogram.

In Section 2 we introduce our proposal of clustering algorithm and the multidendrogram representation for the results. Section 3 gives the corresponding generalization of some hierarchical clustering strategies. In Section 4, Lance and Williams' formula is also generalized consistently with the new proposal. Section 5 shows some results corresponding to data from a real example, and we finish with some conclusions in Section 6.

2. Agglomerative Hierarchical Algorithm

2.1 Pair-Group Approach

Agglomerative hierarchical procedures build a hierarchical classification in a bottom-up way, from a proximity matrix containing dissimilarity

data between individuals of a set $\Omega = \{x_1, x_2, \ldots, x_n\}$ (the same analysis could be done using similarity data). The algorithm has the following steps:

0) Initialize n singleton clusters with one individual in each of them: $\{x_1\}, \{x_2\}, \ldots, \{x_n\}$. Initialize also the distances between clusters, $D(\{x_i\}, \{x_j\})$, with the values of the distances between individuals, $d(x_i, x_j)$:

$$D(\{x_i\}, \{x_j\}) = d(x_i, x_j) \qquad \forall i, j = 1, 2, \ldots, n.$$

1) Find the shortest distance separating two different clusters.

2) Select two clusters X_i and $X_{i'}$ separated by such shortest distance and merge them into a new supercluster $X_i \cup X_{i'}$.

3) Compute the distances $D(X_i \cup X_{i'}, X_j)$ between the new supercluster $X_i \cup X_{i'}$ and each of the other clusters X_j.

4) If all individuals are not in a single cluster yet, then go back to step 1.

Following Sneath and Sokal (1973, p. 216), this type of approach is known as a *pair-group* method, in opposition to *variable-group* methods which will be discussed in the next subsection. Depending on the criterion used for the calculation of distances in step 3, we can implement different agglomerative hierarchical methods. In this article we study some of the most commonly used ones, which are: single linkage, complete linkage, unweighted average, weighted average, unweighted centroid, weighted centroid and joint between-within. The problem of non-uniqueness may arise at step 2 of the algorithm, when two or more pairs of clusters are separated by the shortest distance value (i.e., the shortest distance is tied). Every choice for breaking ties may have important consequences, because it changes the collection of clusters and the distances between them, possibly resulting in different hierarchical classifications. It must be noted here that not all tied distances will produce ambiguity: they have to be the shortest ones and they also have to involve a common cluster. On the other hand, ambiguity is not limited to cases with ties in the original proximity values, but ties may arise during the clustering process too.

The use of any hierarchical clustering technique on a finite set Ω with n individuals results in an *n-tree* on Ω, which is defined as a subset T of parts of Ω satisfying the following conditions:

(i) $\Omega \in T$,

(ii) $\emptyset \notin T$,

(iii) $\forall x \in \Omega \quad \{x\} \in T$,

(iv) $\forall X, Y \in T \quad (X \cap Y = \emptyset \quad \vee \quad X \subseteq Y \quad \vee \quad Y \subseteq X)$.

An n-tree gives only the hierarchical structure of a classification, but the use of a hierarchical clustering technique also associates a height h with each of the clusters obtained. All this information is gathered in the definition of a *valued tree* on Ω, which is a pair (T, h) where T is an n-tree on Ω and $h : T \longrightarrow \mathbb{R}$ is a function such that $\forall X, Y \in T$:

(i) $h(X) \geq 0$,

(ii) $h(X) = 0 \quad \Longleftrightarrow \quad |X| = 1$,

(iii) $X \subsetneq Y \quad \Longrightarrow \quad h(X) < h(Y)$,

where $|X|$ denotes the cardinality of X.

For example, suppose that we have a graph with four individuals like that of Figure 1, where the initial distance between any two individuals is the value of the shortest path connecting them. This means, for example, that the initial distance between x_2 and x_4 is equal to 5. Using the unweighted average criterion, we can obtain three different valued trees. The graphical representation of valued trees are the so called *dendrograms*, and Figure 2 shows the three corresponding dendrograms obtained for our toy graph. The first two dendrograms are quite similar, but the third one shows a considerably different hierarchical structure. Hence, if the third dendrogram is the only one obtained by a software package, one could extract from it the wrong conclusion that x_3 is closer to x_4 than it is to x_2.

2.2 Variable-Group Proposal: Multidendrograms

Any decision taken to break ties in the toy graph of Figure 1 would be arbitrary. In fact, the use of an unfortunate rule might lead us to the worst dendrogram of the three. A logical solution to the pair-group criterion problem might be to assign the same importance to all tied distances and, therefore, to use a variable-group criterion. In our example of Figure 1 this means the amalgamation of individuals x_1, x_2 and x_3 in a single cluster at the same time. The immediate consequence is that we have to calculate the distance between the new cluster $\{x_1\} \cup \{x_2\} \cup \{x_3\}$ and the cluster $\{x_4\}$. In the unweighted average case this distance is equal to 5, that is, the arithmetic mean among the values 7, 5 and 3, corresponding respectively to the distances $D(\{x_1\}, \{x_4\})$, $D(\{x_2\}, \{x_4\})$ and $D(\{x_3\}, \{x_4\})$. We must also decide what height should be assigned to the new cluster formed by x_1, x_2 and x_3, which could be any value between the minimum and the maximum distances that separate any two of them. In this case the minimum distance is 2 and corresponds to both of the tied distances $D(\{x_1\}, \{x_2\})$ and $D(\{x_2\}, \{x_3\})$, while the maximum distance is the one separating x_1 from x_3 and it is equal to 4.
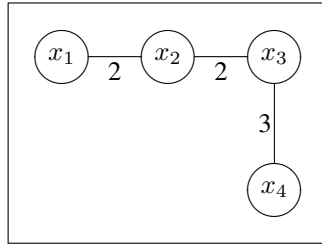
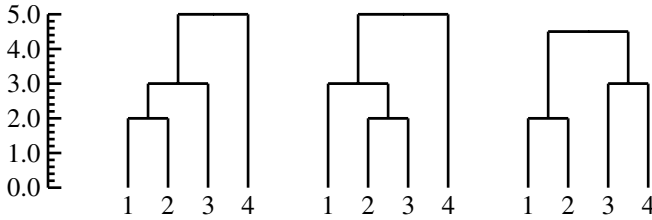Figure 1. Toy Graph with Four Individuals and Shortest Path Distances

Figure 2. Unweighted Average Dendrograms for the Toy Example

Following the variable-group criterion on a finite set Ω with n individuals, we no longer get several valued trees, but we obtain a unique tree which we call a *multivalued tree* on Ω, and we define it as a triplet (T, h_l, h_u) where T is an n-tree on Ω and $h_l, h_u : T \longrightarrow \mathbb{R}$ are two functions such that $\forall X, Y \in T$:

(i) $0 \leq h_l(X) \leq h_u(X)$,

(ii) $h_l(X) = 0 \iff h_u(X) = 0 \iff |X| = 1$,

(iii) $X \subsetneq Y \implies h_l(X) < h_l(Y)$.

A multivalued tree associates with every cluster X in the hierarchical classification two height values, $h_l(X)$ and $h_u(X)$, corresponding respectively to the lower and upper bounds at which member individuals can be merged into cluster X. When $h_l(X)$ and $h_u(X)$ coincide for every cluster X, the multivalued tree is just a valued tree. But, when there is any cluster X for which $h_l(X) < h_u(X)$, it is like having multiple valued trees because every selection of a height $h(X)$ inside the interval $[h_l(X), h_u(X)]$ corresponds to a different valued tree. The length of the interval indicates the degree of heterogeneity inside cluster X. We also introduce here the concept of *multidendrogram* to refer to the graphical representation of a multivalued tree. In Figure 3 we show the corresponding multidendrogram for the toy example.
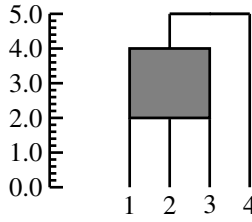
Figure 3. Unweighted Average Multidendrogram for the Toy Example

The shadowed region between heights 2 and 4 refers to the interval between the respective values of h_l and h_u for cluster $\{x_1\} \cup \{x_2\} \cup \{x_3\}$, which in turn also correspond to the minimum and maximum distances separating any two of the constituent clusters $\{x_1\}$, $\{x_2\}$ and $\{x_3\}$.

Let us consider the situation shown in Figure 4, where nine different clusters are to be grouped into superclusters. The clusters to be amalgamated should be those separated by the shortest distance. The picture shows the edges connecting clusters separated by such shortest distance, so we observe that there are six pairs of clusters separated by shortest edges. A pair-group clustering algorithm typically would select any of these pairs, for instance (X_8, X_9), and then it would compute the distance between the new supercluster $X_8 \cup X_9$ and the rest of the clusters X_i, for all $i \in \{1, 2, \ldots, 7\}$. What we propose here is to follow a variable-group criterion and create as many superclusters as groups of clusters connected by shortest edges. In Figure 4, for instance, the nine initial clusters would be grouped into the four following superclusters: X_1, $X_2 \cup X_3$, $X_4 \cup X_5 \cup X_6$ and $X_7 \cup X_8 \cup X_9$. Then, all the pairwise distances between the four superclusters should be computed. In general, we must be able to compute distances $D(X_I, X_J)$ between any two superclusters $X_I = \bigcup_{i \in I} X_i$ and $X_J = \bigcup_{j \in J} X_j$, each one of them made up of several clusters indexed by $I = \{i_1, i_2, \ldots, i_p\}$ and $J = \{j_1, j_2, \ldots, j_q\}$, respectively.

The algorithm that we propose in order to ensure uniqueness in agglomerative hierarchical clustering has the following steps:

0) Initialize n singleton clusters with one individual in each of them: $\{x_1\}$, $\{x_2\}$, ..., $\{x_n\}$. Initialize also the distances between clusters, $D(\{x_i\}, \{x_j\})$, with the values of the distances between individuals, $d(x_i, x_j)$:

$$D(\{x_i\}, \{x_j\}) = d(x_i, x_j) \qquad \forall i, j = 1, 2, \ldots, n.$$

1) Find the shortest distance separating two different clusters, and record it as D_{lower}.
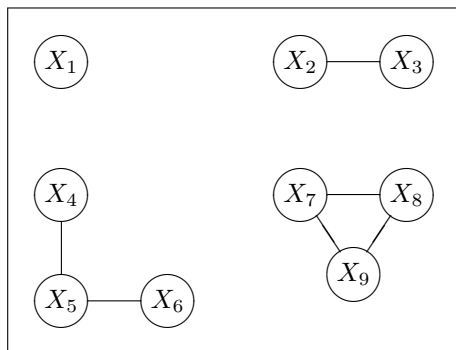
Figure 4. Simultaneous Occurrence of Different Superclusters

2) Select all the groups of clusters separated by shortest distance D_{lower} and merge them into several new superclusters X_I. The result of this step can be some superclusters made up of just one single cluster ($|I| = 1$), as well as some superclusters made up of various clusters ($|I| > 1$). Notice that the latter superclusters all must satisfy the condition $D_{min}(X_I) = D_{lower}$, where

$$D_{min}(X_I) = \min_{i \in I} \min_{\substack{i' \in I \\ i' \neq i}} D(X_i, X_{i'}).$$

3) Update the distances between clusters following the next substeps:

   3.1) Compute the distances $D(X_I, X_J)$ between all superclusters, and record the minimum of them as D_{next} (this will be the shortest distance D_{lower} in the next iteration of the algorithm).

   3.2) For each supercluster X_I made up of various clusters ($|I| > 1$), assign a common amalgamation interval $[D_{lower}, D_{upper}]$ for all its constituent clusters X_i, $i \in I$, where $D_{upper} = D_{max}(X_I)$ and

$$D_{max}(X_I) = \max_{i \in I} \max_{\substack{i' \in I \\ i' \neq i}} D(X_i, X_{i'}).$$

4) If all individuals are not in a single cluster yet, then go back to step 1.

Using the pair-group algorithm, only the centroid methods (weighted and unweighted) may produce *reversals*. Let us remember that a reversal arises in a valued tree when it contains at least two clusters X and Y for which $X \subset Y$ but $h(X) > h(Y)$ (Morgan and Ray 1995). In the case of the variable-group algorithm, reversals may appear in substep 3.2. Although reversals make dendrograms difficult to interpret if they occur during the last stages of the agglomeration process, it can be argued that they are not very

disturbing if they occur during the first stages. Thus, as happens with the centroid methods in the pair-group case, it could be reasonable to use the variable-group algorithm as long as no reversals at all or only unimportant ones were produced.

Sometimes, in substep 3.2 of the variable-group clustering algorithm, it will not be enough to adopt a fusion interval, but it will be necessary to obtain an exact fusion value (e.g., in order to calculate a distortion measure). In these cases, given the lower and upper bounds at which the tied clusters can merge into a supercluster, one possibility is to select the fusion value naturally suggested by the method being applied. For instance, in the case of the toy example and the corresponding multidendrogram shown in Figures 1 and 3, the fusion value would be 2.7 (the unweighted average distance). If the clustering method used was a different one such as single linkage or complete linkage, then the fusion value would be 2 or 4, respectively. Another possibility is to use systematically the shortest distance as the fusion value for the tied clusters. Both criteria allow the recovering of the pair-group result for the single linkage method. The latter criterion, in addition, avoids the appearance of reversals. However, it must be emphasized that the adoption of exact fusion values, without considering the fusion intervals at their whole lengths, means that some valuable information regarding the heterogeneity of the clusters is being lost.

3. Generalization of Agglomerative Hierarchical Methods

In the variable-group clustering algorithm previously proposed we have seen the necessity of agglomerating simultaneously two families of clusters, respectively indexed by $I = \{i_1, i_2, \ldots, i_p\}$ and $J = \{j_1, j_2, \ldots, j_q\}$, into two superclusters $X_I = \bigcup_{i \in I} X_i$ and $X_J = \bigcup_{j \in J} X_j$. In the following subsections we derive, for each of the most commonly used agglomerative hierarchical clustering strategies, the distance between the two superclusters, $D(X_I, X_J)$, in terms of the distances between the respective component clusters, $D(X_i, X_j)$.

3.1 Single Linkage

In *single linkage* clustering, also called *nearest neighbor* or *minimum* method, the distance between two clusters X_i and X_j is defined as the distance between the closest pair of individuals, one in each cluster:

$$D(X_i, X_j) = \min_{x \in X_i} \min_{y \in X_j} d(x, y). \tag{1}$$

This means that the distance between two superclusters X_I and X_J can be defined as

$$D(X_I, X_J) = \min_{x \in X_I} \min_{y \in X_J} d(x, y) = \min_{i \in I} \min_{x \in X_i} \min_{j \in J} \min_{y \in X_j} d(x, y). \tag{2}$$

Notice that this formulation generalizes the definition of distance between clusters in the sense that equation (1) is recovered from equation (2) when $|I| = |J| = 1$, that is, when superclusters I and J are both composed of a single cluster. Grouping terms and using the definition in equation (1), we get the equivalent definition:

$$D(X_I, X_J) = \min_{i \in I} \min_{j \in J} D(X_i, X_j). \tag{3}$$

3.2 Complete Linkage

In *complete linkage* clustering, also known as *furthest neighbor* or *maximum* method, cluster distance is defined as the distance between the most remote pair of individuals, one in each cluster:

$$D(X_i, X_j) = \max_{x \in X_i} \max_{y \in X_j} d(x, y). \tag{4}$$

Starting from equation (4) and following the same reasoning as in the single linkage case, we extend the definition of distance to the superclusters case as

$$D(X_I, X_J) = \max_{i \in I} \max_{j \in J} D(X_i, X_j). \tag{5}$$

3.3 Unweighted Average

Unweighted average clustering, also known as *group average* method or *UPGMA* (Unweighted Pair-Group Method using Averages), iteratively forms clusters made up of pairs of previously formed clusters, based on the arithmetic mean distances between their member individuals. It uses an unweighted averaging procedure, that is, when clusters are joined to form a larger cluster, the distance between this new cluster and any other cluster is calculated weighting each individual in those clusters equally, regardless of the structural subdivision of the clusters:

$$D(X_i, X_j) = \frac{1}{|X_i||X_j|} \sum_{x \in X_i} \sum_{y \in X_j} d(x, y). \tag{6}$$

When the variable-group strategy is followed, the UPGMA name of the method should be modified to that of *UVGMA* (Unweighted Variable-Group Method using Averages), and the distance definition between superclusters in this case should be

$$D(X_I, X_J) = \frac{1}{|X_I||X_J|} \sum_{x \in X_I} \sum_{y \in X_J} d(x, y)$$

$$= \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{x \in X_i} \sum_{j \in J} \sum_{y \in X_j} d(x, y).$$

Using equation (6), we get the desired definition in terms of the distances between component clusters:

$$D(X_I, X_J) = \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| D(X_i, X_j). \qquad (7)$$

In this case, $|X_I|$ is the number of individuals in supercluster X_I, that is, $|X_I| = \sum_{i \in I} |X_i|$.

3.4 Weighted Average

In *weighted average* strategy, also called *WVGMA* (Weighted Variable-Group Method using Averages) in substitution of the corresponding pair-group name *WPGMA*, we calculate the distance between two superclusters X_I and X_J by taking the arithmetic mean of the pairwise distances, not between individuals in the original matrix of distances, but between component clusters in the matrix used in the previous iteration of the procedure:

$$D(X_I, X_J) = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} D(X_i, X_j). \qquad (8)$$

This method is related to the unweighted average one in that the former derives from the latter when we consider

$$|X_i| = 1 \quad \forall i \in I \qquad \text{and} \qquad |X_j| = 1 \quad \forall j \in J. \qquad (9)$$

It weights the most recently admitted individuals in a cluster equally to its previous members. The weighting discussed here is with reference to individuals composing a cluster and not to the average distances in Lance and Williams' recursive formula (see next section), in which equal weights apply for weighted clustering and different weights apply for unweighted clustering (Sneath and Sokal 1973, p. 229).

3.5 Unweighted Centroid

The next three clustering techniques assume that individuals can be represented by points in Euclidean space. This method and the next one further assume that the measure of dissimilarity between any pair of individuals is the squared Euclidean distance between the corresponding pair of points. When the dissimilarity between two clusters X_i and X_j is defined to be the squared distance between their centroids, we are performing *unweighted centroid* (or simply *centroid*) clustering, also called *UPGMC* (Unweighted Pair-Group Method using Centroids):

$$D(X_i, X_j) = \|\overline{x}_i - \overline{x}_j\|^2, \qquad (10)$$

where $\overline{x}_i$ and $\overline{x}_j$ are the centroids of the points in clusters X_i and X_j respectively, and $\|\cdot\|$ is the Euclidean norm. Therefore, under the variable-group point of view, the method could be named *UVGMC* and the distance between two superclusters can be generalized to the definition:

$$D(X_I, X_J) = \|\overline{x}_I - \overline{x}_J\|^2. \tag{11}$$

In the Appendix it is proved that this definition can be expressed in terms of equation (10) as

$$\begin{aligned}
D(X_I, X_J) = {} & \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| D(X_i, X_j) \\
& - \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}| D(X_i, X_{i'}) \\
& - \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} |X_j||X_{j'}| D(X_j, X_{j'}).
\end{aligned} \tag{12}$$

3.6 Weighted Centroid

In *weighted centroid* strategy, also called *median* method or *WVGMC* (Weighted Variable-Group Method using Centroids) in substitution of the pair-group name *WPGMC*, we modify the definition of dissimilarity between two clusters given in the unweighted centroid case, assigning each cluster the same weight in calculating the "centroid". Now the center of a supercluster X_I is the average of the centers of the constituent clusters:

$$\overline{x}_I = \frac{1}{|I|} \sum_{i \in I} \overline{x}_i.$$

This clustering method is related to the unweighted centroid one by relation (9), which also related the weighted average strategy to the corresponding unweighted average. So, in this case we define the distance between two superclusters as

$$\begin{aligned}
D(X_I, X_J) = {} & \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} D(X_i, X_j) \\
& - \frac{1}{|I|^2} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} D(X_i, X_{i'}) - \frac{1}{|J|^2} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} D(X_j, X_{j'}).
\end{aligned} \tag{13}$$

3.7 Joint Between-Within

Székely and Rizzo (2005) propose an agglomerative hierarchical clustering method that minimizes a *joint between-within* cluster distance, measuring both heterogeneity between clusters and homogeneity within clusters. This method extends Ward's minimum variance method (Ward 1963) by defining the distance between two clusters X_i and X_j in terms of any power $\alpha \in (0, 2]$ of Euclidean distances between individuals:

$$
D(X_i, X_j) = \frac{|X_i||X_j|}{|X_i| + |X_j|} \left(\frac{2}{|X_i||X_j|} \sum_{x \in X_i} \sum_{y \in X_j} \|x - y\|^\alpha \right.
$$
$$
\left. - \frac{1}{|X_i|^2} \sum_{x \in X_i} \sum_{x' \in X_i} \|x - x'\|^\alpha - \frac{1}{|X_j|^2} \sum_{y \in X_j} \sum_{y' \in X_j} \|y - y'\|^\alpha \right). (14)
$$

When $\alpha = 2$, cluster distances are a weighted squared distance between cluster centers

$$
D(X_i, X_j) = \frac{2|X_i||X_j|}{|X_i| + |X_j|} \|\bar{x}_i - \bar{x}_j\|^2,
$$

equal to twice the cluster distance that is used in Ward's method.

In the Appendix we derive the following recursive formula for updating cluster distances as a generalization of equation (14):

$$
D(X_I, X_J) = \frac{1}{|X_I| + |X_J|} \sum_{i \in I} \sum_{j \in J} (|X_i| + |X_j|) D(X_i, X_j)
$$
$$
- \frac{|X_J|}{|X_I|(|X_I| + |X_J|)} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|) D(X_i, X_{i'})
$$
$$
- \frac{|X_I|}{|X_J|(|X_I| + |X_J|)} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} (|X_j| + |X_{j'}|) D(X_j, X_{j'}). \quad (15)
$$

4. Generalization of Lance and Williams' Formula

Lance and Williams (1966) put the most commonly used agglomerative hierarchical strategies into a single system, avoiding the necessity of a separate computer program for each of them. Assume three clusters X_i, $X_{i'}$ and X_j, containing $|X_i|$, $|X_{i'}|$ and $|X_j|$ individuals respectively and with distances between them already determined as $D(X_i, X_{i'})$, $D(X_i, X_j)$

and $D(X_{i'}, X_j)$. Further assume that the smallest of all distances still to be considered is $D(X_i, X_{i'})$, so that X_i and $X_{i'}$ are joined to form a new super-cluster $X_i \cup X_{i'}$, with $|X_i| + |X_{i'}|$ individuals. Lance and Williams express $D(X_i \cup X_{i'}, X_j)$ in terms of the distances already defined, all known at the moment of fusion, using the following recurrence relation:

$$D(X_i \cup X_{i'}, X_j) = \alpha_i D(X_i, X_j) + \alpha_{i'} D(X_{i'}, X_j)$$
$$+ \beta D(X_i, X_{i'}) + \gamma |D(X_i, X_j) - D(X_{i'}, X_j)|. \tag{16}$$

With this technique superclusters can always be computed from previous clusters and it is not necessary to return to the original dissimilarity data during the clustering process. The values of the parameters α_i, $\alpha_{i'}$, β and γ determine the nature of the sorting strategy. Table 1 gives the values of the parameters that define the most commonly used agglomerative hierarchical clustering methods.

We next give a generalization of formula (16) compatible with the amalgamation of more than two clusters simultaneously. Suppose that one wants to agglomerate two superclusters X_I and X_J, respectively indexed by $I = \{i_1, i_2, \ldots, i_p\}$ and $J = \{j_1, j_2, \ldots, j_q\}$. We define the distance between them as

$$D(X_I, X_J) = \sum_{i \in I} \sum_{j \in J} \alpha_{ij} D(X_i, X_j)$$

$$+ \sum_{\substack{i \in I \\ i' > i}} \sum_{\substack{i' \in I \\ }} \beta_{ii'} D(X_i, X_{i'}) + \sum_{\substack{j \in J \\ j' > j}} \sum_{\substack{j' \in J \\ }} \beta_{jj'} D(X_j, X_{j'})$$

$$+ \delta \sum_{i \in I} \sum_{j \in J} \gamma_{ij} [D_{max}(X_I, X_J) - D(X_i, X_j)]$$

$$- (1 - \delta) \sum_{i \in I} \sum_{j \in J} \gamma_{ij} [D(X_i, X_j) - D_{min}(X_I, X_J)], \tag{17}$$

where
$$D_{max}(X_I, X_J) = \max_{i \in I} \max_{j \in J} D(X_i, X_j)$$

and
$$D_{min}(X_I, X_J) = \min_{i \in I} \min_{j \in J} D(X_i, X_j).$$

Table 2 shows the values for the parameters α_{ij}, $\beta_{ii'}$, $\beta_{jj'}$, γ_{ij} and δ which determine the clustering method computed by formula (17). They are all gathered from the respective formulae (3), (5), (7), (8), (12), (13) and (15), derived in the previous section.

5. Glamorganshire Soils Example

We show here a real example which has been studied by Morgan and Ray (1995) using the complete linkage method. It is the *Glamorganshire*

Table 1. Parameter Values for the Lance and Williams' Formula

Method	α_i	$\alpha_{i'}$	β	γ
Single linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage	$\frac{1}{2}$	$\frac{1}{2}$	0	$+\frac{1}{2}$
Unweighted average	$\frac{|X_i|}{|X_i|+|X_{i'}|}$	$\frac{|X_{i'}|}{|X_i|+|X_{i'}|}$	0	0
Weighted average	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Unweighted centroid	$\frac{|X_i|}{|X_i|+|X_{i'}|}$	$\frac{|X_{i'}|}{|X_i|+|X_{i'}|}$	$-\frac{|X_i||X_{i'}|}{(|X_i|+|X_{i'}|)^2}$	0
Weighted centroid	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Joint between-within	$\frac{|X_i|+|X_j|}{|X_i|+|X_{i'}|+|X_j|}$	$\frac{|X_{i'}|+|X_j|}{|X_i|+|X_{i'}|+|X_j|}$	$-\frac{|X_j|}{|X_i|+|X_{i'}|+|X_j|}$	0

Table 2. Parameter Values for the Variable-Group Formula

Method	α_{ij}	$\beta_{ii'}$	$\beta_{jj'}$	γ_{ij}	δ
Single linkage	$\frac{1}{|I||J|}$	0	0	$\frac{1}{|I||J|}$	0
Complete linkage	$\frac{1}{|I||J|}$	0	0	$\frac{1}{|I||J|}$	1
Unweighted average	$\frac{|X_i||X_j|}{|X_I||X_J|}$	0	0	0	$-$
Weighted average	$\frac{1}{|I||J|}$	0	0	0	$-$
Unweighted centroid	$\frac{|X_i||X_j|}{|X_I||X_J|}$	$-\frac{|X_i||X_{i'}|}{|X_I|^2}$	$-\frac{|X_j||X_{j'}|}{|X_J|^2}$	0	$-$
Weighted centroid	$\frac{1}{|I||J|}$	$-\frac{1}{|I|^2}$	$-\frac{1}{|J|^2}$	0	$-$
Joint between-within	$\frac{|X_i|+|X_j|}{|X_I|+|X_J|}$	$-\frac{|X_J|}{|X_I|}\frac{|X_i|+|X_{i'}|}{|X_I|+|X_J|}$	$-\frac{|X_I|}{|X_J|}\frac{|X_j|+|X_{j'}|}{|X_I|+|X_J|}$	0	$-$

soils example, formed by similarity data between 23 different soils. A table with the similarities can be found also in Morgan and Ray (1995), where the values are given with an accuracy of three decimal places. In order to work with dissimilarities, first of all we have transformed the similarities $s(x_i, x_j)$ into the corresponding dissimilarities $d(x_i, x_j) = 1 - s(x_i, x_j)$.

The original data present a tied value for pairs of soils (3,15) and (3,20), which is responsible for two different dendrograms using the complete linkage strategy. We show them in Figures 5 and 6. Morgan and Ray (1995) explain that the 23 soils have been categorized into eight "great soil groups" by a surveyor. Focusing on soils 1, 2, 6, 12 and 13, which are the only members of the brown earths soil group, we see that the dendrogram in Figure 5 does not place them in the same cluster until they join soils from five other soil groups, forming the cluster (1, 2, 3, 20, 12, 13, 15, 5, 6, 8, 14, 18). From this point of view, the dendrogram in Figure 6 is better, since the corresponding cluster loses soils 8, 14 and 18, each representing a different soil group. So, in this case, we have two possible solution dendrograms and the probability of obtaining the "good" one is, hence, 50%.

On the other hand, in Figure 7 we can see the multidendrogram corresponding to the *Glamorganshire soils* data. The existence of a tie comprising
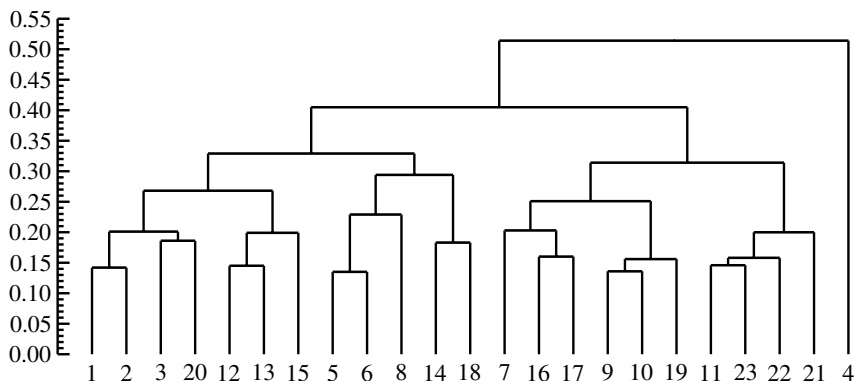
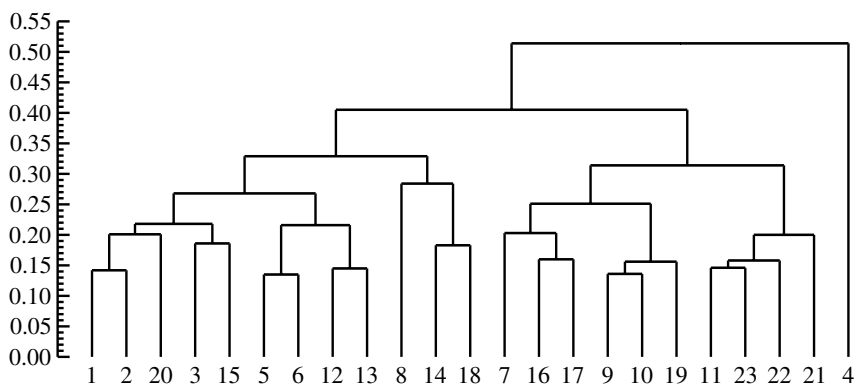Figure 5. First Complete Linkage Dendrogram for the Soils Data

Figure 6. Second Complete Linkage Dendrogram for the Soils Data

soils 3, 15 and 20 is clear from this tree representation. Besides, the multi-dendrogram gives us the good classification, that is, the one with soils 8, 14 and 18 out of the brown earths soil group. Except for the internal structure of the cluster (1, 2, 3, 15, 20), the rest of the multidendrogram hierarchy coincides with that of the dendrogram shown in Figure 6.

Finally, notice that the incidence of ties depends on the accuracy with which proximity values are available. In this example, if dissimilarities had been measured to four decimal places, then the tie causing the non-unique complete linkage dendrogram might have disappeared. On the contrary, the probability of ties is higher if lower accuracy data are used. For instance, when we consider the same soils data but with an accuracy of only two decimal places, we obtain the multidendrogram shown in Figure 8, where three different ties can be observed.
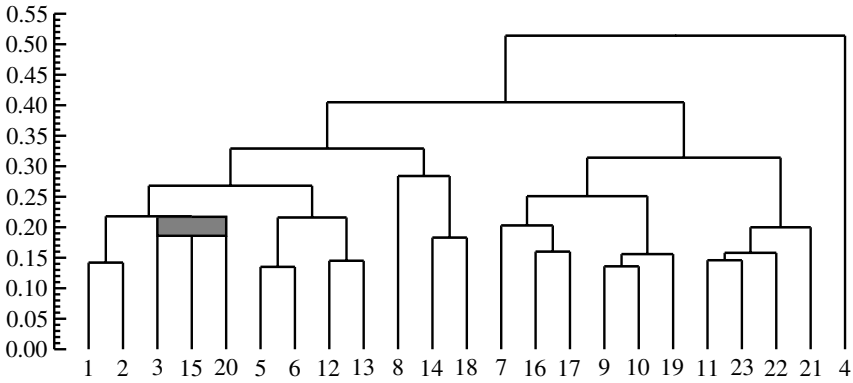
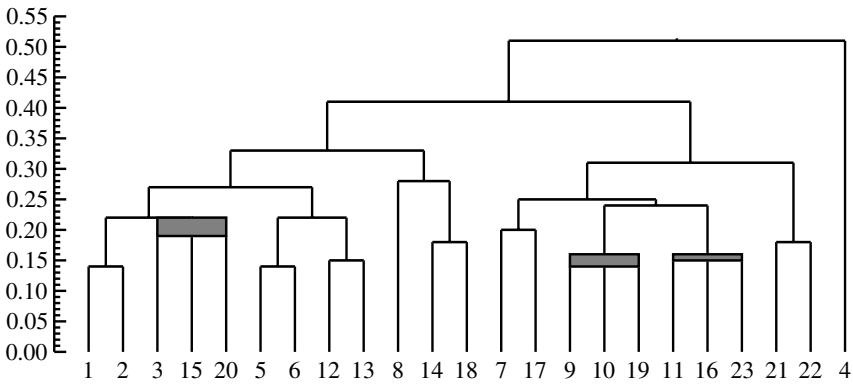Figure 7. Complete Linkage Multidendrogram for the Soils Data

Figure 8. Complete Linkage Multidendrogram for the Soils Data with an Accuracy of Two Decimal Places

6. Conclusions

The non-uniqueness problem in agglomerative hierarchical clustering generates several hierarchical classifications from a unique set of tied proximity data. In such cases, selecting a unique classification can be misleading. This problem has traditionally been dealt with distinct criteria, which mostly consist of the selection of one out of various resulting hierarchies. In this article we have proposed a variable-group algorithm for agglomerative hierarchical clustering that solves the ties in proximity problem. The output of this algorithm is a uniquely determined type of valued tree, which we call a multivalued tree, while graphically we represent it with a multidendrogram.

In addition we have generalized the definition of distance between clusters for the most commonly used agglomerative hierarchical methods,

in order to be able to compute them using the variable-group algorithm. We have also given the corresponding generalization of Lance and Williams' formula, which enables us to get agglomerative hierarchical classifications in a recursive way. Finally, we have showed the possible usefulness of our proposal with some results obtained using data from a real example.

Gathering up the main advantages of our new proposal, we can state the following points:

- When there are no ties, the variable-group algorithm gives the same result as the pair-group one.
- The new algorithm always gives a uniquely determined solution.
- In the multidendrogram representation for the results one can explicitly observe the occurrence of ties during the agglomerative process. Furthermore, the height of any fusion interval indicates the degree of heterogeneity inside the corresponding cluster.
- When ties exist, the variable-group algorithm is computationally more efficient than obtaining all the possible solutions following out the various ties with the pair-group alternative.
- The new proposal can be also computed in a recursive way using a generalization of Lance and Williams' formula.

Although ties need not be present in the initial proximity data, they may arise during the agglomeration process. For this reason and given that the results of the variable-group algorithm coincide with those of the pair-group algorithm when there are not any ties, we recommend to use directly the variable-group option. With a single action one knows whether ties exist or not, and additionally the subsequent solution is obtained.

Appendix: Proofs

A.1 Proof for the Unweighted Centroid Method

Given a cluster X_i, its centroid is

$$\overline{x}_i = \frac{1}{|X_i|} \sum_{x \in X_i} x,$$

and the centroid of a supercluster X_I can be expressed in terms of its constituent centroids by the equation:

$$\overline{x}_I = \frac{1}{|X_I|} \sum_{i \in I} |X_i| \overline{x}_i. \tag{18}$$

Now, given two superclusters X_I and X_J, the distance between them defined in equation (11) is

$$D(X_I, X_J) = \|\overline{x}_I - \overline{x}_J\|^2 = \|\overline{x}_I\|^2 + \|\overline{x}_J\|^2 - 2\langle \overline{x}_I, \overline{x}_J \rangle,$$

where $\langle \cdot, \cdot \rangle$ stands for the inner product. If we substitute each centroid by its definition (18), we obtain:

$$D(X_I, X_J) = \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{i' \in I} |X_i||X_{i'}|\langle \overline{x}_i, \overline{x}_{i'} \rangle$$

$$+ \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{j' \in J} |X_j||X_{j'}|\langle \overline{x}_j, \overline{x}_{j'} \rangle$$

$$- \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|2\langle \overline{x}_i, \overline{x}_j \rangle.$$

Now, since

$$2\langle \overline{x}_i, \overline{x}_j \rangle = \|\overline{x}_i\|^2 + \|\overline{x}_j\|^2 - \|\overline{x}_i - \overline{x}_j\|^2,$$

we have that

$$D(X_I, X_J) = \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{i' \in I} |X_i||X_{i'}|\langle \overline{x}_i, \overline{x}_{i'} \rangle$$

$$+ \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{j' \in J} |X_j||X_{j'}|\langle \overline{x}_j, \overline{x}_{j'} \rangle$$

$$- \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\|\overline{x}_i\|^2 - \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\|\overline{x}_j\|^2$$

$$+ \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\|\overline{x}_i - \overline{x}_j\|^2.$$

But this can be rewritten as

$$D(X_I, X_J) = \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\|\overline{x}_i - \overline{x}_j\|^2$$

$$- \frac{1}{|X_I|} \sum_{i \in I} |X_i|\|\overline{x}_i\|^2 + \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{i' \in I} |X_i||X_{i'}|\langle \overline{x}_i, \overline{x}_{i'} \rangle$$

$$- \frac{1}{|X_J|} \sum_{j \in J} |X_j|\|\overline{x}_j\|^2 + \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{j' \in J} |X_j||X_{j'}|\langle \overline{x}_j, \overline{x}_{j'} \rangle,$$

and, grouping terms,

$$D(X_I, X_J) = \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\|\overline{x}_i - \overline{x}_j\|^2$$

$$- \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{i' \in I} |X_i||X_{i'}| \left(\|\overline{x}_i\|^2 - \langle \overline{x}_i, \overline{x}_{i'} \rangle \right)$$

$$- \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{j' \in J} |X_j||X_{j'}| \left(\|\overline{x}_j\|^2 - \langle \overline{x}_j, \overline{x}_{j'} \rangle \right).$$

The second and third terms can be simplified a little more, thanks to the equality

$$\sum_{i \in I} \sum_{i' \in I} |X_i||X_{i'}| \left(\|\overline{x}_i\|^2 - \langle \overline{x}_i, \overline{x}_{i'} \rangle \right) =$$

$$= \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}| \left(\|\overline{x}_i\|^2 + \|\overline{x}_{i'}\|^2 - 2\langle \overline{x}_i, \overline{x}_{i'} \rangle \right).$$

With this simplification, we have that

$$D(X_I, X_J) = \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| \|\overline{x}_i - \overline{x}_j\|^2$$

$$- \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}| \|\overline{x}_i - \overline{x}_{i'}\|^2$$

$$- \frac{1}{|X_J|^2} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} |X_j||X_{j'}| \|\overline{x}_j - \overline{x}_{j'}\|^2,$$

and, recalling the definition of distance between two clusters given in equation (10), we finally obtain the desired form of equation (12).

A.2 Proof for the Joint Between-Within Method

We give here a proof based on that of Székely and Rizzo (2005) for their agglomerative hierarchical formulation. Using the following constants:

$$\theta_{ij} = \frac{1}{|X_i||X_j|} \sum_{x \in X_i} \sum_{y \in X_j} \|x - y\|^{\alpha},$$

$$\theta_{ii} = \frac{1}{|X_i|^2} \sum_{x \in X_i} \sum_{x' \in X_i} \|x - x'\|^{\alpha}, \qquad (19)$$

the definition (14) of distance between two clusters X_i and X_j is

$$D(X_i, X_j) = \frac{|X_i||X_j|}{|X_i| + |X_j|} (2\theta_{ij} - \theta_{ii} - \theta_{jj}).$$

Consider now the superclusters X_I and X_J formed by merging clusters X_i, for all $i \in I$, and X_j, for all $j \in J$. Define the corresponding constants:

$$\theta_{IJ} = \frac{1}{|X_I||X_J|} \sum_{x \in X_I} \sum_{y \in X_J} \|x - y\|^{\alpha}$$

$$= \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} \sum_{x \in X_i} \sum_{y \in X_j} \|x - y\|^{\alpha},$$

$$\theta_{II} = \frac{1}{|X_I|^2} \sum_{x \in X_I} \sum_{x' \in X_I} \|x - x'\|^\alpha$$

$$= \frac{1}{|X_I|^2} \sum_{i \in I} \sum_{i' \in I} \sum_{x \in X_i} \sum_{x' \in X_{i'}} \|x - x'\|^\alpha$$

$$= \frac{1}{|X_I|^2} \sum_{i \in I} \left(\sum_{x \in X_i} \sum_{x' \in X_i} \|x - x'\|^\alpha + 2 \sum_{\substack{i' \in I \\ i' > i}} \sum_{x \in X_i} \sum_{x' \in X_{i'}} \|x - x'\|^\alpha \right),$$

so that in terms of the original constants (19) we have:

$$\theta_{IJ} = \frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\theta_{ij},$$

$$\theta_{II} = \frac{1}{|X_I|^2} \sum_{i \in I} \left(|X_i|^2 \theta_{ii} + 2 \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|\theta_{ii'} \right).$$

Therefore, the distance between superclusters X_I and X_J is given by

$$D(X_I, X_J) = \frac{|X_I||X_J|}{|X_I| + |X_J|} (2\theta_{IJ} - \theta_{II} - \theta_{JJ})$$

$$= \frac{|X_I||X_J|}{|X_I| + |X_J|} \left[\frac{2}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j|\theta_{ij} \right.$$

$$- \frac{1}{|X_I|^2} \sum_{i \in I} \left(|X_i|^2 \theta_{ii} + 2 \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|\theta_{ii'} \right)$$

$$\left. - \frac{1}{|X_J|^2} \sum_{j \in J} \left(|X_j|^2 \theta_{jj} + 2 \sum_{\substack{j' \in J \\ j' > j}} |X_j||X_{j'}|\theta_{jj'} \right) \right].$$

Simplify

$$\sum_{i \in I} \left(|X_i|^2 \theta_{ii} + 2 \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|\theta_{ii'} \right) =$$

$$= \sum_{i \in I} \left[|X_i|^2 \theta_{ii} + \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|(2\theta_{ii'} - \theta_{ii} - \theta_{i'i'} + \theta_{ii} + \theta_{i'i'}) \right]$$

$$= \sum_{i \in I} \left[|X_i|^2 \theta_{ii} + \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|(\theta_{ii} + \theta_{i'i'}) \right.$$

$$\left. + \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|)D(X_i, X_{i'}) \right]$$

$$= |X_I| \sum_{i \in I} |X_i| \theta_{ii} + \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|) D(X_i, X_{i'}),$$

where in last equality we have used the equivalence

$$\sum_{i \in I} \left[|X_i|^2 \theta_{ii} + \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}|(\theta_{ii} + \theta_{i'i'}) \right] = |X_I| \sum_{i \in I} |X_i| \theta_{ii}.$$

Hence,

$$(|X_I| + |X_J|) D(X_I, X_J) = 2 \sum_{i \in I} \sum_{j \in J} |X_i||X_j| \theta_{ij}$$

$$- |X_J| \sum_{i \in I} |X_i| \theta_{ii} - \frac{|X_J|}{|X_I|} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|) D(X_i, X_{i'})$$

$$- |X_I| \sum_{j \in J} |X_j| \theta_{jj} - \frac{|X_I|}{|X_J|} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} (|X_j| + |X_{j'}|) D(X_j, X_{j'}),$$

or, equivalently,

$$(|X_I| + |X_J|) D(X_I, X_J) = \sum_{i \in I} \sum_{j \in J} |X_i||X_j| 2\theta_{ij}$$

$$- \sum_{i \in I} |X_i| \theta_{ii} \sum_{j \in J} |X_j| - \sum_{i \in I} |X_i| \sum_{j \in J} |X_j| \theta_{jj}$$

$$- \frac{|X_J|}{|X_I|} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|) D(X_i, X_{i'})$$

$$- \frac{|X_I|}{|X_J|} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} (|X_j| + |X_{j'}|) D(X_j, X_{j'}),$$

which is also the same as

$$(|X_I| + |X_J|) D(X_I, X_J) = \sum_{i \in I} \sum_{j \in J} (|X_i| + |X_j|) D(X_i, X_j)$$

$$- \frac{|X_J|}{|X_I|} \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} (|X_i| + |X_{i'}|) D(X_i, X_{i'})$$

$$- \frac{|X_I|}{|X_J|} \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} (|X_j| + |X_{j'}|) D(X_j, X_{j'}).$$

And this is exactly the desired formulation given in equation (15).

References

ARNAU, V., MARS, S., and MARÍN, I. (2005), "Iterative Cluster Analysis of Protein Interaction Data," *Bioinformatics, 21(3)*, 364–378.

BACKELJAU, T., DE BRUYN, L., DE WOLF, H., JORDAENS, K., VAN DONGEN, S., and WINNEPENNINCKX, B. (1996), "Multiple UPGMA and Neighbor-Joining Trees and the Performance of Some Computer Packages," *Molecular Biology and Evolution, 13(2)*, 309–313.

CORMACK, R.M. (1971), "A Review of Classification" (with discussion), *Journal of the Royal Statistical Society, Ser. A, 134*, 321–367.

GORDON, A.D. (1999), *Classification* (2nd ed.), London/Boca Raton, FL:Chapman & Hall/CRC.

HART, G. (1983), "The Occurrence of Multiple UPGMA Phenograms," in *Numerical Taxonomy*, ed. J. Felsenstein, Berlin Heidelberg: Springer-Verlag, pp. 254–258.

LANCE, G.N., and WILLIAMS, W.T. (1966), "A Generalized Sorting Strategy for Computer Classifications," *Nature, 212*, 218.

MACCUISH, J., NICOLAOU, C., and MACCUISH, N.E. (2001), "Ties in Proximity and Clustering Compounds," *Journal of Chemical Information and Computer Sciences, 41*, 134–146.

MORGAN, B.J.T., and RAY, A.P.G. (1995), "Non-uniqueness and Inversions in Cluster Analysis," *Applied Statistics, 44(1)*, 117–134.

SNEATH, P.H.A., and SOKAL, R.R. (1973), *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, San Francisco: W. H. Freeman and Company.

SZÉKELY, G.J., and RIZZO, M.L. (2005), "Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method," *Journal of Classification, 22*, 151–183.

VAN DER KLOOT, W.A., SPAANS, A.M.J., and HEISER, W.J. (2005), "Instability of Hierarchical Cluster Analysis Due to Input Order of the Data: The PermuCLUSTER Solution," *Psychological Methods, 10(4)*, 468–476.

WARD, J.H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association, 58*, 236–244.