mdendro: An R Package for Extended Agglomerative Hierarchical Clustering

Alberto Fernández
Universitat Rovira i Virgili

Sergio Gómez
Universitat Rovira i Virgili

Abstract

mdendro is an R package that provides a comprehensive collection of linkage methods for agglomerative hierarchical clustering on a matrix of proximity data (distances or similarities), returning a multifurcated dendrogram or multidendrogram. Multidendrograms can group more than two clusters at the same time, solving the nonuniqueness problem that arises when there are ties in the data. This problem causes that different binary dendrograms are possible depending both on the order of the input data and on the criterion used to break ties. Weighted and unweighted versions of the most common linkage methods are included in the package, which also implements two parametric linkage methods. In addition, package **mdendro** provides five descriptive measures to analyze the resulting dendrograms: cophenetic correlation coefficient, space distortion ratio, agglomerative coefficient, chaining coefficient and tree balance.

Keywords: multifurcated dendrogram, parametric linkage, dendrogram descriptor, R.

1. Introduction

Agglomerative hierarchical clustering (AHC) is widely used to classify individuals into a hierarchy of clusters organized in a tree structure called dendrogram (Gordon 1999). There are different types of AHC linkage methods, such as single linkage, complete linkage, average linkage and Ward's method, which only differ in the definition of the distance measure between clusters. All these methods start from a distance matrix between individuals, each one forming a singleton cluster, and gather clusters into groups of clusters, this process being repeated until a complete hierarchy of partitions into clusters is formed.

Except for the single linkage case, all the other AHC linkage methods suffer from a nonuniqueness problem known as the *ties in proximity* problem. This problem arises whenever there are more than two clusters separated by the same minimum distance during the agglomerative process of a pair-group AHC algorithm. This type of algorithm breaks ties choosing any pair of clusters, and proceeds in the same way until a binary dendrogram is obtained. However, different binary dendrograms are possible depending both on the order of the input data and on the criterion used to break ties.

The ties in proximity problem is long known (Hart 1983; Morgan and Ray 1995; Backeljau, De Bruyn, De Wolf, Jordaens, Van Dongen, and Winnepenninckx 1996), even from studies in different fields, such as biology (Arnau, Mars, and Marín 2005), psychology (van der Kloot, Spaans, and Heiser 2005) and chemistry (MacCuish, Nicolaou, and MacCuish 2001).

The extend of the problem in a particular field has been analyzed for microsatellite markers (Segura-Alabart, Serratosa, Gómez, and Fernández 2022). Nevertheless, this problem is ignored by some software packages: the `hclust()` function in the **stats** package and the `agnes()` function in the **cluster** package of R (R Core Team 2021), the `cluster()` and `clustermat()` commands of Stata (StataCorp LLC 2021), the `linkage()` function in the **Statistics and Machine Learning Toolbox** of MATLAB (The MathWorks Inc. 2022), and the `hclust()` function in the **Clustering.jl** package of Julia (Bezanson, Edelman, Karpinski, and Shah 2017).

There are some other statistical packages that just warn against the existence of the nonuniqueness problem in AHC. For instance, the `Hierarchical Cluster Analysis` procedure of SPSS Statistics (IBM Corporation 2021), the `CLUSTER` procedure of SAS (SAS Institute Inc. 2018), the `Agglomerate()` function in the **Hierarchical Clustering Package** of Mathematica (Wolfram Language & System Documentation Center 2020), and the `linkage()` function in the `scipy.cluster.hierarchy` module of the **SciPy** package in Python (Virtanen, Gommers, Oliphant, Haberland, Reddy, Cournapeau, Burovski, Peterson, Weckesser, Bright, van der Walt, Brett, Wilson, Millman, Mayorov, Nelson, Jones, Kern, Larson, Carey, Polat, Feng, Moore, VanderPlas, Laxalde, Perktold, Cimrman, Henriksen, Quintero, Harris, Archibald, Ribeiro, Pedregosa, van Mulbregt, and SciPy 1.0 Contributors 2020).

Software packages that do not ignore the nonuniqueness problem fail to adopt a common standard with respect to ties, and they simply break ties in any arbitrary way. Here we introduce **mdendro**, an R package that implements a variable-group AHC algorithm (Fernández and Gómez 2008) to solve the nonuniqueness problem found in any pair-group AHC algorithm.

Package **mdendro** was designed using state-of-the-art methods based on neighbor chains, and its base code was implemented in C++. The package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=mdendro` and on GitHub at `https://github.com/sergio-gomez/mdendro`. The functionality of the R package **mdendro** makes it very similar and compatible with the main ones currently in use, namely the R functions `hclust()` in package **stats** and `agnes()` in package **cluster** (Maechler, Rousseeuw, Struyf, Hubert, and Hornik 2021). The result is a package **mdendro** that includes and extends the functionality of these reference functions.

The rest of the article is structured as follows. In Section 2 we describe the pair-group and the variable-group AHC algorithms, the latter grouping more than two clusters at the same time when ties occur. Section 3 describes the most common AHC linkage methods: single linkage, complete linkage, average linkage, centroid linkage and Ward's method. Package **mdendro** also includes two parametric linkage methods: β-flexible linkage and versatile linkage. In the same section, five descriptive measures for the resulting dendrograms are included: cophenetic correlation coefficient, space distortion ratio, agglomerative coefficient, chaining coefficient and tree balance. Section 4 compares package **mdendro** with other state-of-the-art packages for AHC. Finally, in Section 5, we give some concluding remarks.

2. Agglomerative hierarchical clustering algorithms

2.1. Pair-group algorithm

AHC algorithms build a hierarchical tree in a bottom-up way, from a matrix of pairwise distances between individuals of a set $\Omega = \{x_1, \ldots, x_n\}$. The pair-group algorithm (Sneath

and Sokal 1973) has the following steps:

0) Initialize n singleton clusters with one individual in each one of them: $X_1 = \{x_1\}, \ldots, X_n = \{x_n\}$. Initialize also the distances between clusters, $D(X_i, X_j)$, with the values of the distances between individuals, $d(x_i, x_j)$:

$$D(X_i, X_j) = d(x_i, x_j)\,, \qquad \forall i, j = 1, \ldots, n\,.$$

1) Find the shortest distance separating two different clusters, D_{shortest}.

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

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

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

The nonuniqueness problem in the pair-group algorithm arises when two or more shortest distances between different clusters are equal during the agglomerative process (Hart 1983). The standard approach consists in choosing only a single pair to break the tie. However, different hierarchical clusterings are possible depending on the criterion used to break ties (usually a pair is just chosen at random), and the user is unaware of this problem.

For example, let us consider the genetic profiles of 51 grapevine cultivars at six microsatellite loci (Almadanim, Baleiras-Couto, Pereira, Carneiro, Fevereiro, Eiras-Dias, Morais-Cecilio, Viegas, and Veloso 2007). The distance between two cultivars is defined as one minus the fraction of shared alleles, and this definition is used to calculate a distance matrix `d`. The main characteristic of this kind of data is that the number of different distances is very small:

```
R> length(unique(d))
```

```
[1] 11
```

As a consequence of these 11 unique values out of 1275 pairwise distances in the matrix, it becomes very easy to find tied distances during the agglomeration process. The reach of the nonuniqueness problem for this example is the existence of 11,160 structurally different binary dendrograms. This number corresponds to the average linkage method and a resolution of 3 decimal digits, and it has been computed using the `Hierarchical_Clustering` tool in **Radatools** (Gómez and Fernández 2021). We can check the diversity of results by just calculating binary dendrograms for random permutations of the data and plotting the broad range of values of their cophenetic correlation coefficients (see Figure 1), what clearly indicates the existence of many structurally different binary dendrograms.

2.2. Variable-group algorithm

Fernández and Gómez (2008) introduced a variable-group algorithm to ensure uniqueness in AHC, which differs from the pair-group algorithm in the following steps:

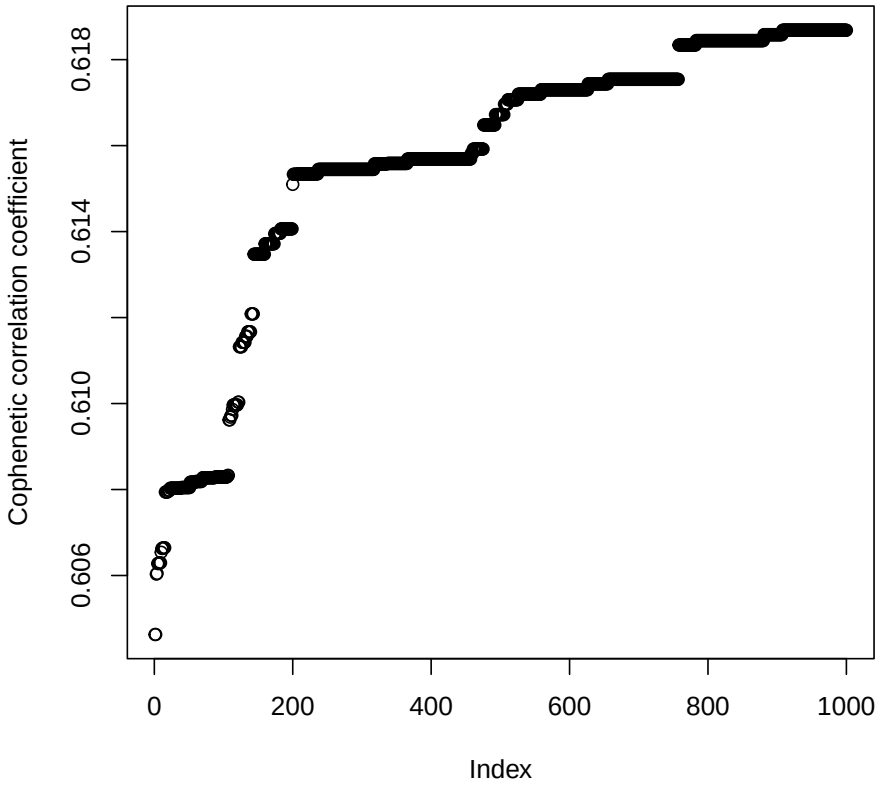

Figure 1: Sorted cophenetic correlation coefficients for the different pair-group dendrograms obtained by using random permutations of the same grapevine cultivars dataset.

2) Select all the groups of clusters separated by the shortest distance D_{shortest}, and merge them into several new clusters $X_I = \bigcup_{i \in I} X_i$, each one made up of several subclusters X_i indexed by i in $I = \{i_1, \ldots, i_p\}$.

3) Compute the distances $D(X_I, X_J)$ between any two clusters $X_I = \bigcup_{i \in I} X_i$ and $X_J = \bigcup_{j \in J} X_j$, each one of them made up of several subclusters X_i and X_j indexed by i in $I = \{i_1, \ldots, i_p\}$ and j in $J = \{j_1, \ldots, j_q\}$, respectively.

When there are tied shortest distances in the agglomerative process, in order to keep track of valuable information regarding the heterogeneity of the clusters that are formed, the `linkage()` function in the **mdendro** package saves for each cluster X_I made up of more than one subcluster ($|I| > 1$) a fusion interval $[D_{\min}(X_I), D_{\max}(X_I)]$, where:

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

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

The variable-group algorithm groups more than two clusters at the same time when ties occur, giving rise to a graphical representation called multidendrogram. Its main properties are:

- When there are no ties, the variable-group algorithm gives the same results as the pair-group one.
- It 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 range of any fusion interval indicates the degree of heterogeneity inside the corresponding cluster.

With the `linkage()` function, you can use both the pair-group algorithm or the variable-group one (see Figure 2):

```
R> par(mfrow = c(2, 1))
R> cars <- round(dist(scale(mtcars)), digits = 1)
R> nodePar <- list(cex = 0, lab.cex = 0.7)
R> lnk1 <- linkage(cars, method = "complete", group = "pair")
R> plot(lnk1, main = "dendrogram", nodePar = nodePar)
R> lnk2 <- linkage(cars, method = "complete", group = "variable")
R> plot(lnk2, col.rng = "pink", main = "multidendrogram", nodePar = nodePar)
```

The identification of ties requires the selection of the number of significant digits in the working dataset. For example, if the original distances are experimentally obtained with a resolution of three decimal digits, two distances that differ in the sixth decimal digit should be considered as equal. If this is not taken into account, ties might be broken just by the numerical imprecision inherent to computer representations of real numbers. In the `linkage()` function, you can control this level of resolution by adjusting its `digits` argument.

3. Linkage methods

3.1. Common linkage methods

During each iteration of the AHC algorithm, the distances $D(X_I, X_J)$ have to be computed between any two clusters $X_I = \bigcup_{i\in I} X_i$ and $X_J = \bigcup_{j\in J} X_j$, each one of them made up of several subclusters X_i and X_j indexed by i in $I = \{i_1, \ldots, i_p\}$ and j in $J = \{j_1, \ldots, j_q\}$, respectively. Lance and Williams (1966) introduced a formula for integrating several AHC linkage methods into a single system, avoiding the need of a separate computer program for each one of them. Similarly, Fernández and Gómez (2008) gave a variable-group generalization of this formula, compatible with the fusion of more than two clusters simultaneously:

$$D(X_I, X_J) = \sum_{i\in I}\sum_{j\in J} \alpha_{ij} D(X_i, X_j) + \sum_{i\in I}\sum_{\substack{i'\in I \\ i'>i}} \beta_{ii'} D(X_i, X_{i'}) + \sum_{j\in J}\sum_{\substack{j'\in J \\ j'>j}} \beta_{jj'} D(X_j, X_{j'})\,. \quad (1)$$

Function `linkage()` in package **mdendro** uses this recurrence relation to compute the distance $D(X_I, X_J)$ from the distances $D(X_i, X_j)$ obtained during the previous iteration, being

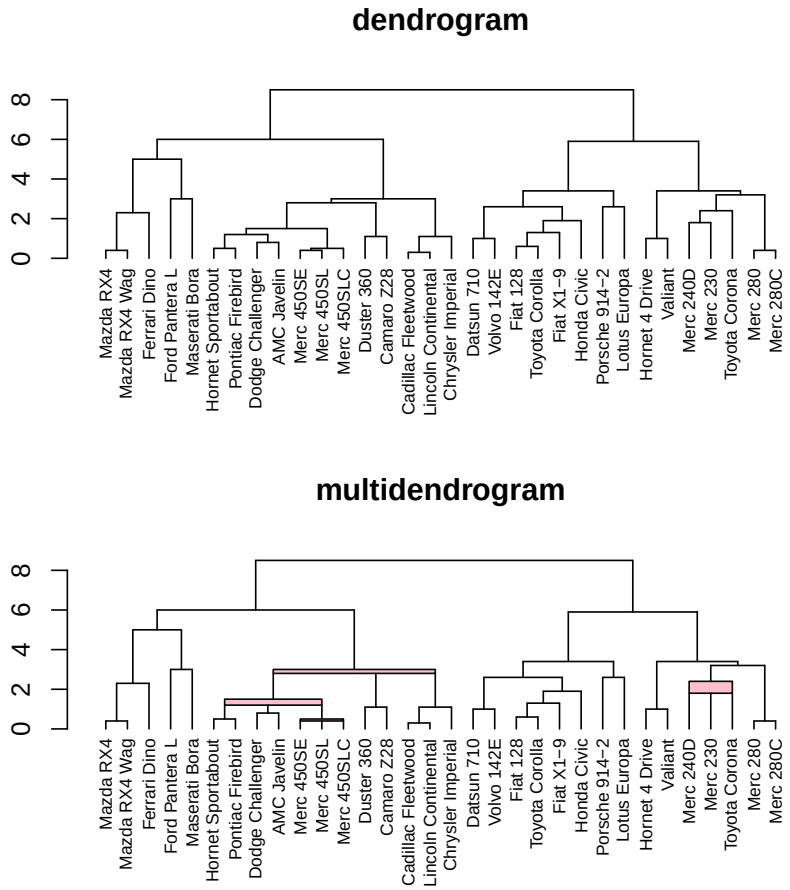

Figure 2: Pair-group dendrogram vs. variable-group multidendrogram. The ranges (rectangles) in the multidendrogram show the heterogeneity of distances within the group, but they are optional in the plots and can be hidden just by setting the `col.rng` argument in the `plot()` function to `NULL`.

unnecessary to look back at the initial distance matrix $d(x_i, x_j)$ at all. The values of the parameters α_{ij}, $\beta_{ii'}$ and $\beta_{jj'}$ determine the nature of the AHC linkage methods (Fernández and Gómez 2008). Some of these methods even have weighted and unweighted forms, which differ in the weights assigned to individuals and clusters during the agglomerative process: weighted methods assign equal weights to clusters, while unweighted methods assign equal weights to individuals. Package **mdendro** implements weighted and unweighted forms of the most commonly used AHC linkage methods, namely:

- `single`: the proximity between clusters equals the minimum distance or the maximum similarity between objects.
- `complete`: the proximity between clusters equals the maximum distance or the minimum similarity between objects.
- `arithmetic`: the proximity between clusters equals the arithmetic mean proximity between objects. Also known as average linkage, UPGMA (unweighted pair-group method

using averages) or WPGMA (weighted pair-group method using averages).

- `centroid`: the distance between clusters equals the square of the Euclidean distance between the centroids of each cluster. Also known as UPGMC (unweighted pair-group method using centroids) or WPGMC (weighted pair-group method using centroids). This method is available only for distance data.

- `ward`: the distance between clusters is a weighted squared Euclidean distance between the centroids of each cluster. This method is available only for distance data.

In Figure 3, we can see the differences between these AHC linkage methods on the `UScitiesD` dataset, a matrix of distances between a few US cities:

```
R> par(mfrow = c(2, 3))
R> methods <- c("single", "complete", "arithmetic", "centroid", "ward")
R> for (m in methods) {
+    lnk <- linkage(UScitiesD, method = m)
+    plot(lnk, cex = 0.6, main = m)
+  }
```

3.2. Descriptive measures

The result of the `linkage()` function is an object of class '`linkage`' that describes the resulting dendrogram obtained. In particular, this object contains the following calculated descriptors:

- `cor`: Cophenetic correlation coefficient (Sokal and Rohlf 1962), defined as the Pearson correlation coefficient between the output cophenetic proximity data and the input proximity data. It is a measure of how faithfully the dendrogram preserves the pairwise proximity between objects.

- `sdr`: Space distortion ratio (Fernández and Gómez 2020), calculated as the difference between the maximum and minimum cophenetic proximity data, divided by the difference between the maximum and minimum initial proximity data. Space dilation occurs when the space distortion ratio is greater than 1.

- `ac`: Agglomerative coefficient (Rousseeuw 1986), a number between 0 and 1 measuring the strength of the clustering structure obtained.

- `cc`: Chaining coefficient (Williams, Lambert, and Lance 1966), a number between 0 and 1 measuring the tendency for clusters to grow by the addition of clusters much smaller rather than by fusion with other clusters of comparable size.

- `tb`: Tree balance (Fernández and Gómez 2020), a number between 0 and 1 measuring the equality in the number of leaves in the branches concerned at each fusion in the hierarchical tree.

For instance, when we use the `linkage()` function to calculate the complete linkage of the `UScitiesD` dataset, we obtain the following summary for the resulting dendrogram:

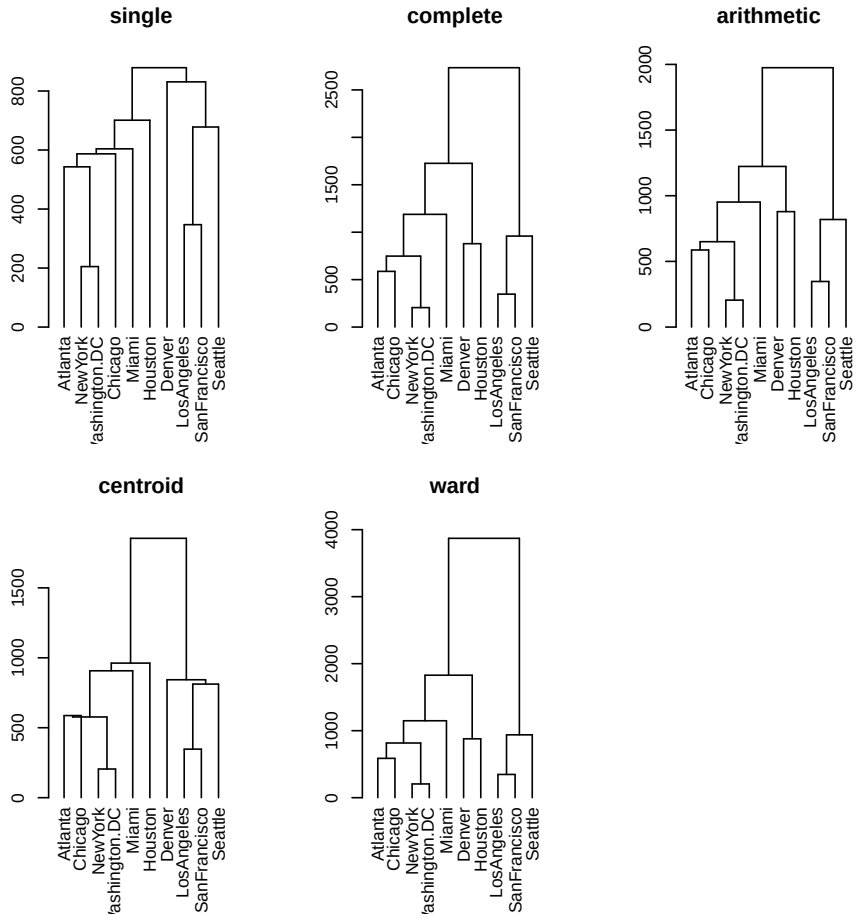

Figure 3: Common linkage methods on the `UScitiesD` dataset.

```
R> lnk <- linkage(UScitiesD, method = "complete")
R> summary(lnk)

Call:
linkage(prox = UScitiesD,
        type.prox = "distance",
        digits = 0,
        method = "complete",
        group = "variable")

Binary dendrogram: TRUE

Descriptive measures:
      cor       sdr        ac        cc        tb
0.8077859 1.0000000 0.7738478 0.3055556 0.9316262
```

While multidendrograms are unique, users may obtain structurally different pair-group dendrograms by just reordering the data. As a consequence, descriptors are invariant to per-

mutations for multidendrograms, but not for pair-group dendrograms. Let us calculate a variable-group multidendrogram and a pair-group dendrogram for the same data:

```
R> cars <- round(dist(scale(mtcars)), digits = 1)
R> lnk1 <- linkage(cars, method = "complete", group = "variable")
R> lnk2 <- linkage(cars, method = "complete", group = "pair")
```

Now, if we apply a random permutation to data:

```
R> set.seed(1234)
R> ord <- sample(attr(cars, "Size"))
R> carsp <- as.dist(as.matrix(cars)[ord, ord])
R> lnk1p <- linkage(carsp, method = "complete", group = "variable")
R> lnk2p <- linkage(carsp, method = "complete", group = "pair")
```

We can check that the original and the permuted cophenetic correlation coefficients are identical for variable-group multidendrograms:

```
R> c(lnk1$cor, lnk1p$cor)

[1] 0.7782257 0.7782257
```

And they are different for pair-group dendrograms:

```
R> c(lnk2$cor, lnk2p$cor)

[1] 0.7780010 0.7776569
```

3.3. Parametric linkage methods

Two of the AHC linkage methods available in package **mdendro**, `flexible` and `versatile`, depend on a parameter that takes values in $[-1, +1]$ for `flexible` linkage, and in $(-\text{Inf}, +\text{Inf})$ for `versatile` linkage. In function `linkage()`, the desired value for the parameter is passed through the `par.method` argument. Here come some examples on the `UScitiesD` dataset (see Figure 4):

```
R> par(mfrow = c(2, 3))
R> vals <- c(-0.8, 0.0, 0.8)
R> for (v in vals) {
+    lnk <- linkage(UScitiesD, method = "flexible", par.method = v)
+    plot(lnk, cex = 0.6, main = sprintf("flexible (%.1f)", v))
+  }
R> vals <- c(-10.0, 0.0, 10.0)
R> for (v in vals) {
+    lnk <- linkage(UScitiesD, method = "versatile", par.method = v)
+    plot(lnk, cex = 0.6, main = sprintf("versatile (%.1f)", v))
+  }
```

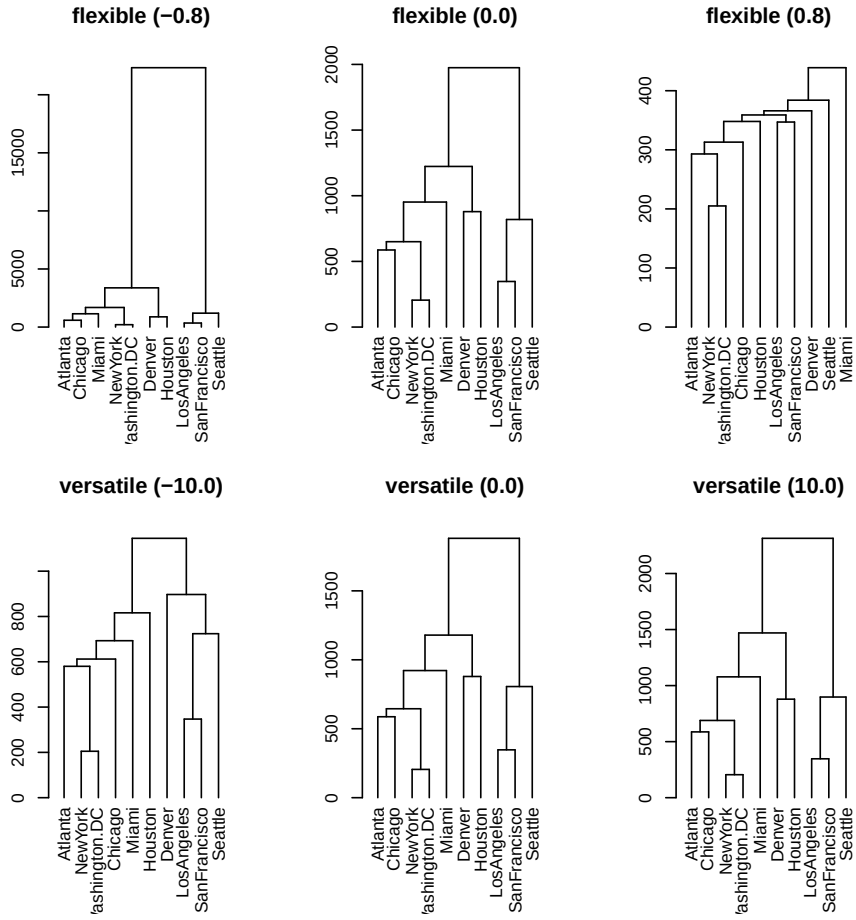

Figure 4: Parametric linkage methods on the `UScitiesD` dataset.

β-flexible linkage

Based on Equation 1, Lance and Williams (1967) proposed an infinite system of AHC strategies defined by the following constraint:

$$\underbrace{\sum_{i\in I}\sum_{j\in J}\alpha_{ij}}_{\alpha}+\underbrace{\sum_{i\in I}\sum_{\substack{i'\in I\\ i'>i}}\beta_{ii'}+\sum_{j\in J}\sum_{\substack{j'\in J\\ j'>j}}\beta_{jj'}}_{\beta}=1\,, \tag{2}$$

where $-1 \leq \beta \leq +1$. Given a value of β, the value for α_{ij} can be assigned following a weighted approach as in the original β-flexible clustering method based on WPGMA and introduced by Lance and Williams (1966), or it can be assigned following an unweighted approach as in the β-flexible clustering method based on UPGMA and introduced by Belbin, Faith, and Milligan (1992). Further details can be consulted in Fernández and Gómez (2020). When β is set equal to 0, `flexible` linkage is equivalent to `arithmetic` linkage.

It is interesting to know how the descriptive measures depend on the parameter of the parametric linkage methods. Package **mdendro** provides the function `descplot()` for this task. For example, using the `flexible` linkage method on the `UScitiesD` dataset (see Figure 5):

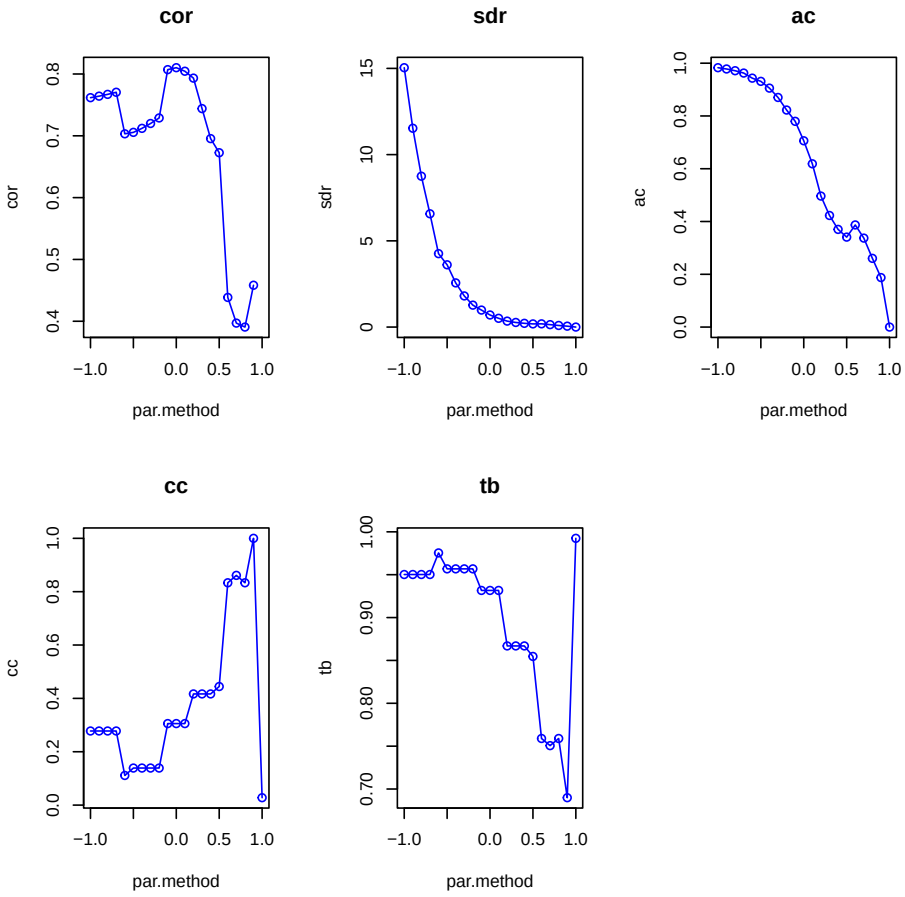

Figure 5: Descriptive measures obtained with the `flexible` linkage method on the `UScitiesD` dataset.

```
R> par(mfrow = c(2, 3))
R> measures <- c("cor", "sdr", "ac", "cc", "tb")
R> vals <- seq(from = -1, to = +1, by = 0.1)
R> for (m in measures)
+    descplot(UScitiesD, method = "flexible",
+             measure = m, par.method = vals,
+             type = "o",  main = m, col = "blue")
```

Versatile linkage

Package **mdendro** also implements another parametric linkage method named versatile linkage (Fernández and Gómez 2020). Substituting the arithmetic means by generalized means, also known as power means, we can extend arithmetic linkage to any finite power $p \neq 0$:

$$D_p(X_I, X_J) = \left(\frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| [D_p(X_i, X_j)]^p \right)^{1/p}, \tag{3}$$

where $|X_i|$ and $|X_j|$ are the number of individuals in subclusters X_i and X_j, and $|X_I|$ and $|X_J|$ are the number of individuals in clusters X_I and X_J, i.e., $|X_I| = \sum_{i \in I} |X_i|$ and $|X_J| = \sum_{j \in J} |X_j|$. Equation 3 shows that versatile linkage can be calculated using a combinatorial formula from the distances $D_p(X_i, X_j)$ obtained during the previous iteration, in the same way as the recurrence formula given in Equation 1.

Versatile linkage provides a way of obtaining an infinite number of AHC strategies from a single formula, just changing the value of the power p. The decision of what power p to use can be taken in agreement with the type of distance employed to measure the initial distances between individuals. For instance, if the initial distances were calculated using a generalized distance of order p, then the natural AHC strategy would be versatile linkage with the same power p. However, this procedure does not guarantee that the dendrogram obtained is the best one according to other criteria, e.g., cophenetic correlation coefficient, space distortion ratio or tree balance (see Section 3.2). Another possible approach consists in scanning the whole range of parameters p, calculate the preferred descriptors of the corresponding dendrograms, and decide if it is better to substitute the natural parameter p by another one. This is especially important when only the distances between individuals are available, without coordinates for the individuals, as is common in multidimensional scaling problems, or when the distances have not been calculated using generalized means.

As in the case of `flexible` linkage, the parameter p of `versatile` linkage is introduced using the `par.method` argument of the function `linkage()`. Here, it is also interesting to know how the descriptors depend on the parameter of this method (see Figure 6):

```
R> par(mfrow = c(2, 3))
R> measures <- c("cor", "sdr", "ac", "cc", "tb")
R> vals <- c(-Inf, (-20:+20), +Inf)
R> for (m in measures)
+    descplot(UScitiesD, method = "versatile",
+             measure = m, par.method = vals,
+             type = "o",  main = m, col = "blue")
```

Particular cases. The generalized mean contains several well-known particular cases, depending on the value of the power p. Some of them reduce `versatile` linkage to the most commonly used methods, while others emerge naturally as deserving special attention:

- In the limit when $p \to -\infty$, `versatile` linkage becomes `single` linkage:

$$D_{\min}(X_I, X_J) = \min_{i \in I} \min_{j \in J} D_{\min}(X_i, X_j)\,. \tag{4}$$

- In the limit when $p \to +\infty$, `versatile` linkage becomes `complete` linkage:

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

There are also three other particular cases that can be grouped together as *Pythagorean linkages*, which show the convenience to rename average linkage as `arithmetic` linkage, to emphasize the existence of different types of averages:

- When $p = +1$, the generalized mean is equal to the arithmetic mean and `arithmetic` linkage is recovered.

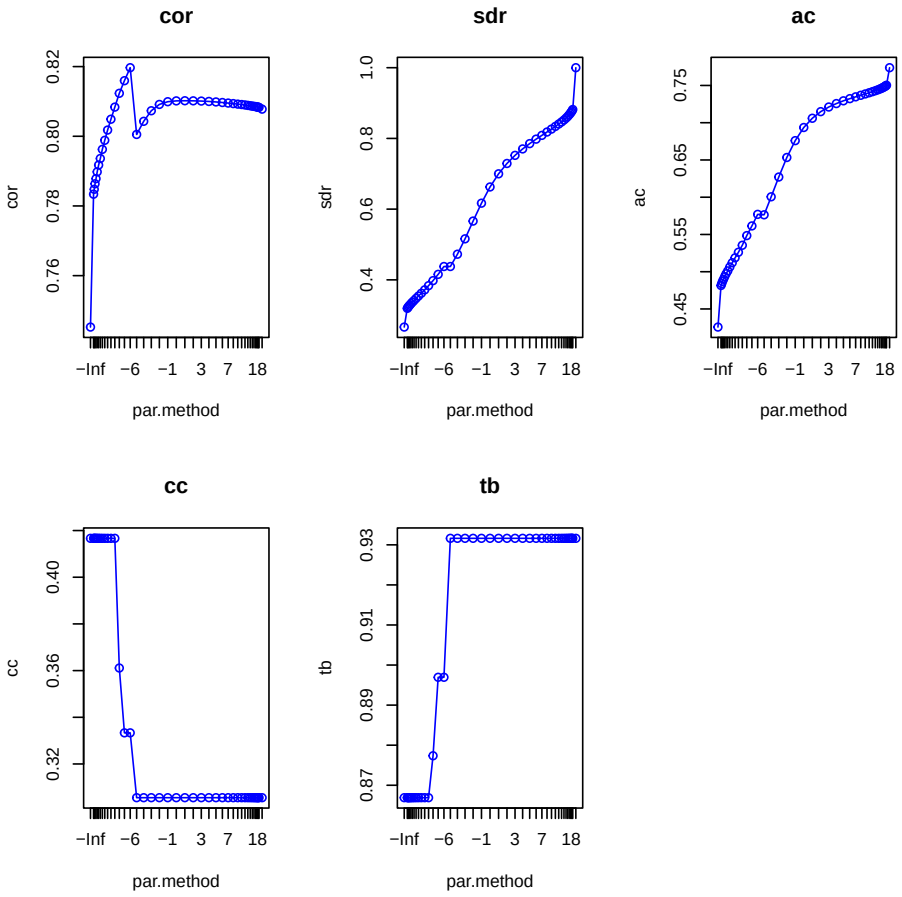

Figure 6: Descriptive measures obtained with the `versatile` linkage method on the `UScitiesD` dataset.

- When $p = -1$, the generalized mean is equal to the harmonic mean and `harmonic` linkage is obtained.

$$D_{\text{har}}(X_I, X_J) = \left(\frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| [D_{\text{har}}(X_i, X_j)]^{-1} \right)^{-1}. \tag{6}$$

- In the limit when $p \to 0$, the generalized mean tends to the geometric mean and `geometric` linkage is obtained:

$$D_{\text{geo}}(X_I, X_J) = \left(\prod_{i \in I} \prod_{j \in J} [D_{\text{geo}}(X_i, X_j)]^{|X_i||X_j|} \right)^{1/(|X_I||X_J|)}. \tag{7}$$

The correspondence between `versatile` linkage and the above mentioned linkage methods is summarized in Table 1.

Let us show a small example in which we plot different dendrograms as we increase the versatile linkage parameter, indicating the corresponding named methods (see Figure 7):

	versatile (par.method)
complete	+Inf
arithmetic	+1
geometric	0
harmonic	-1
single	-Inf

Table 1: Correspondence between `versatile` linkage and other linkage methods.

```
R> d = as.dist(matrix(c( 0,  7, 16, 12,
+                         7,  0,  9, 19,
+                        16,  9,  0, 12,
+                        12, 19, 12, 0), nrow = 4))
R> par(mfrow = c(2, 3))
R> vals <- c(-Inf, -1, 0, 1, Inf)
R> names <- c("single", "harmonic", "geometric", "arithmetic", "complete")
R> titles <- sprintf("versatile (%.1f) = %s", vals, names)
R> for (i in 1:length(vals)) {
+    lnk <- linkage(d, method = "versatile", par.method = vals[i], digits = 2)
+    plot(lnk, ylim = c(0, 20), cex = 0.6, main = titles[i])
+  }
```

4. Comparison with other packages

Except for the cases containing tied distances, the equivalences in Table 2 hold between function `linkage()` in package **mdendro**, function `hclust()` in package **stats** and function `agnes()` in package **cluster**.

For comparison, we can construct the same AHC using the functions `linkage()`, `hclust()` and `agnes()`, where the default plots just show some differences in aesthetics (see Figure 8):

```
R> lnk <- mdendro::linkage(UScitiesD, method = "complete")
R> hcl <- stats::hclust(UScitiesD, method = "complete")
R> agn <- cluster::agnes(UScitiesD, method = "complete")
R> par(mar = c(5, 3, 4, 0), mfrow = c(1, 3))
R> plot(lnk)
R> plot(hcl)
R> plot(agn, which.plots = 2)
```

To enhance usability and interoperability, class '`linkage`' includes the method `as.dendrogram()` for class conversion. In the previous example, converting to class '`dendrogram`' the objects returned by the functions `linkage()`, `hclust()` and `agnes()`, we can see that all three dendrograms are structurally equivalent.

The cophenetic or ultrametric matrix is readily available as component `coph` of the returned '`linkage`' object, and coincides with those obtained using the functions`hclust()` and `agnes()`:

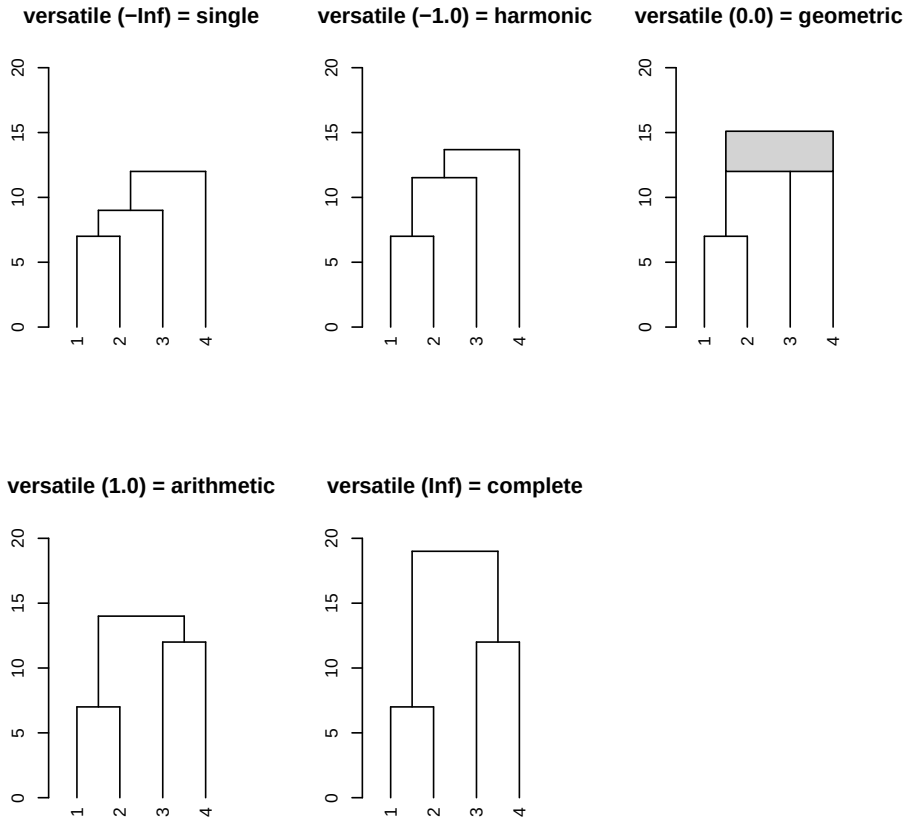

Figure 7: Example of different dendrograms obtained as we increase the versatile linkage parameter.

```
R> hcl.coph <- cophenetic(hcl)
R> all(lnk$coph == hcl.coph)

[1] TRUE

R> agn.coph <- cophenetic(agn)
R> all(lnk$coph == agn.coph)

[1] TRUE
```

The coincidence also applies to the cophenetic correlation coefficient and the agglomerative coefficient, with the advantage that function `linkage()` has both of them already calculated:

```
R> hcl.cor <- cor(UScitiesD, hcl.coph)
R> all.equal(lnk$cor, hcl.cor)

[1] TRUE
```

`linkage()`	`hclust()`	`agnes()`
`single`	`single`	`single`
`complete`	`complete`	`complete`
`arithmetic`, U	`average`	`average`
`arithmetic`, W	`mcquitty`	`weighted`
`geometric`, U/W	—	—
`harmonic`, U/W	—	—
`versatile`, U/W, p	—	—
—	`ward`	—
`ward`	`ward.D2`	`ward`
`centroid`, U	`centroid`	—
`centroid`, W	`median`	—
`flexible`, U, β	—	`gaverage`, β
—	—	`gaverage`, α_1, α_2, β, γ
`flexible`, W, β	—	`flexible`, $(1-\beta)/2$
—	—	`flexible`, α_1, α_2, β, γ

Table 2: Equivalences between functions `linkage()`, `hclust()` and `agnes()`. When relevant, weighted (W) or unweighted (U) versions of the linkage methods and the values for `par.method` are indicated.

```
R> all.equal(lnk$ac, agn$ac)

[1] TRUE
```

Plots including ranges are only available if you directly use the `plot.linkage()` function from package **mdendro**. Anyway, you may still take advantage of other dendrogram plotting packages, such as **dendextend** (Galili 2015) and **ape** (Paradis and Schliep 2019) (see Figure 9):

```
R> par(mar = c(5, 2, 4, 0), mfrow = c(1, 2))
R> cars <- round(dist(scale(mtcars)), digits = 1)
R> lnk <- linkage(cars, method = "complete")
R> lnk.dend <- as.dendrogram(lnk)
R> plot(dendextend::set(lnk.dend, "branches_k_color", k = 4),
+       main = "dendextend package",
+       nodePar = list(cex = 0.4, lab.cex = 0.5))
R> lnk.hcl <- as.hclust(lnk)
R> pal4 <- c("red", "forestgreen", "purple", "orange")
R> clu4 <- cutree(lnk.hcl, 4)
R> plot(ape::as.phylo(lnk.hcl),
+       type = "fan",
+       main = "ape package",
+       tip.color = pal4[clu4],
+       cex = 0.5)
```

And users can also use the function `linkage()` to plot heatmaps containing multidendrograms (see Figure 10):

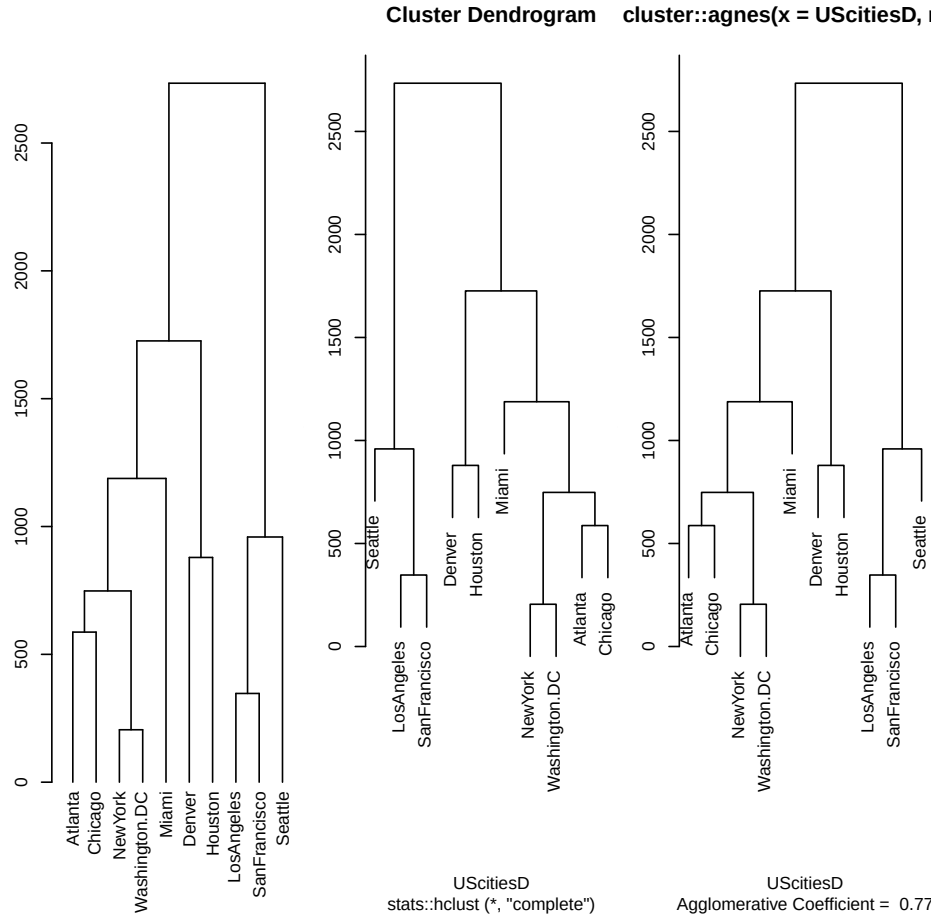

Figure 8: Comparison of `complete` linkage on the `UScitiesD` dataset, using the functions `linkage()`, `hclust()` and `agnes()`.

```
R> heatmap(scale(mtcars), hclustfun = linkage)
```

In addition, it is possible to work directly with similarity data without having to convert them to distances, provided they are in the range [0.0, 1.0]. A typical example would be a matrix of nonnegative correlations (see Figure 11):

```
R> sim <- as.dist(Harman23.cor$cov)
R> lnk <- linkage(sim, type.prox = "sim")
R> plot(lnk)
```

5. Summary and discussion

mdendro is a simple yet powerful R package to make hierarchical clusterings of data. It implements a variable-group algorithm for AHC that solves the nonuniqueness problem found in pair-group algorithms. This problem consists in obtaining different hierarchical clusterings from the same matrix of pairwise distances, when two or more shortest distances between

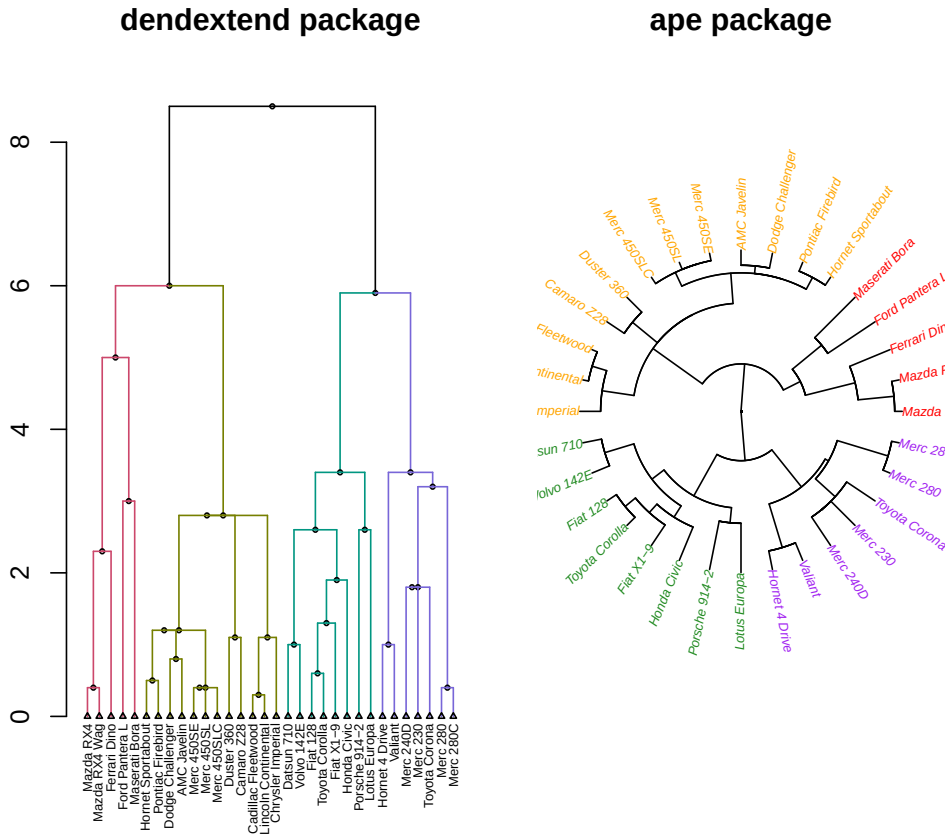

Figure 9: Converting objects of class '`linkage`' using the function `as.dendrogram()`, one can take advantage of other dendrogram plotting packages, such as **dendextend** and **ape**.

different clusters are equal during the agglomeration process. In such cases, selecting a unique clustering can be misleading. Software packages that do not ignore this problem fail to adopt a common standard with respect to ties, and many of them simply break ties in any arbitrary way.

Package **mdendro** computes dendrograms grouping more than two clusters at the same time when ties occur. It includes and extends the functionality of other reference packages in several ways:

- Native handling of both distance and similarity matrices.
- Calculation of variable-group multifurcated dendrograms, which solve the nonuniqueness problem of AHC when there are tied distances.
- Implementation of the most common AHC linkage methods: single linkage, complete linkage, average linkage, centroid linkage and Ward's method.
- Implementation of two parametric linkage methods: β-flexible linkage and versatile linkage. The latter leads naturally to the definition of two new linkage methods: harmonic

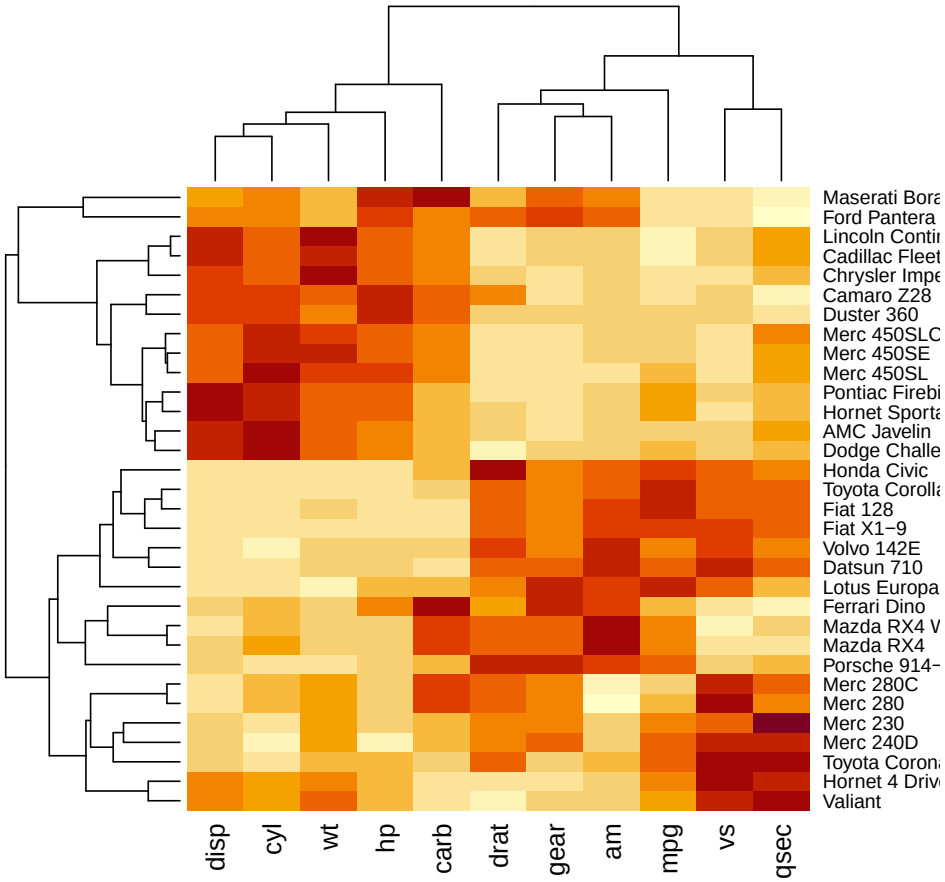

Figure 10: Example of heatmap constructed using the function `linkage()`.

linkage and geometric linkage.

- Implementation of both weighted and unweighted forms for the previous linkage methods.
- Calculation of the cophenetic (or ultrametric) matrix.
- Calculation of five descriptive measures for the resulting dendrogram: cophenetic correlation coefficient, space distortion ratio, agglomerative coefficient, chaining coefficient and tree balance.
- Plots of the descriptive measures for the parametric linkage methods.

Although ties need not be present in the initial proximity data, they may arise during the agglomerative 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 no ties, we recommend to directly use package **mdendro**. With a single action one knows whether ties exist or not, and additionally the subsequent hierarchical clustering is obtained.

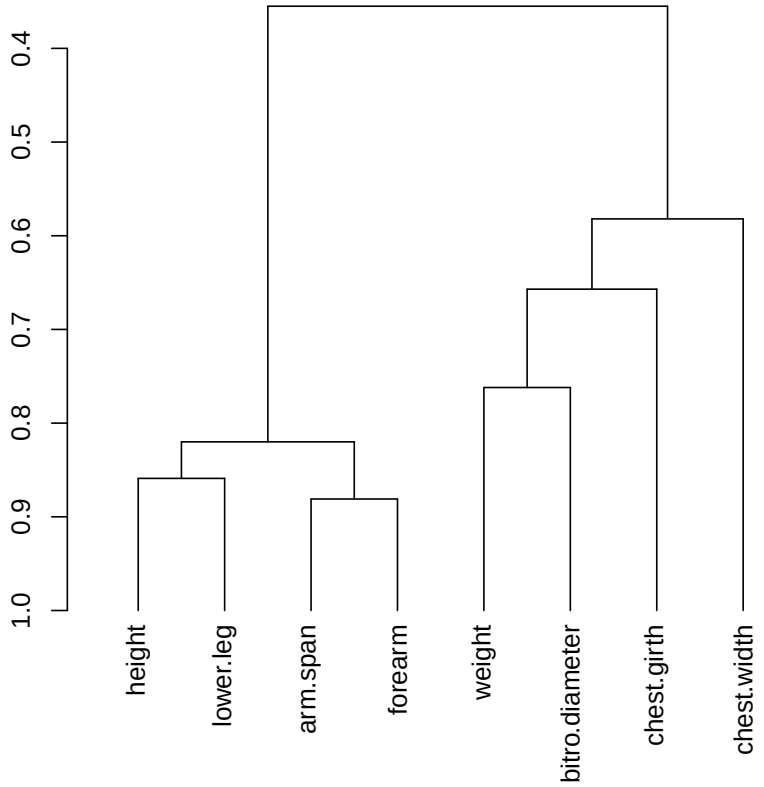

Figure 11: Example of a dendrogram constructed from a matrix of nonnegative correlations, i.e., directly using similarities instead of distances.

Computational details

The results in this paper were obtained using R 4.3.1 with the **mdendro** 2.1.0 package. R itself and all packages used are available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/.

Acknowledgments

This work was supported by Ministerio de Ciencia e Innovación (PID2021-124139NB-C22, PID2021-128005NB-C21, RED2022-134890-T and TED2021-129851B-I00), Generalitat de Catalunya (2021SGR-633) and Universitat Rovira i Virgili (2021PFR-URV-100 and 2022PFR-URV-56).

References

Almadanim M, Baleiras-Couto M, Pereira H, Carneiro L, Fevereiro P, Eiras-Dias J, Morais-Cecilio L, Viegas W, Veloso M (2007). "Genetic Diversity of the Grapevine (Vitis Vinifera L.) Cultivars Most Utilized for Wine Production in Portugal." *Vitis*, **46**(3), 116. doi:10.5073/vitis.2007.46.116-119.

Arnau V, Mars S, Marín I (2005). "Iterative Cluster Analysis of Protein Interaction Data." *Bioinformatics*, **21**(3), 364–378. doi:10.1093/bioinformatics/bti021.

Backeljau T, De Bruyn L, De Wolf H, Jordaens K, Van Dongen S, Winnepennincks B (1996). "Multiple UPGMA and Neighbor-Joining Trees and the Performance of Some Computer Packages." *Molecular Biology and Evolution*, **13**(2), 309–313. doi:10.1093/oxfordjournals.molbev.a025590.

Belbin L, Faith D, Milligan G (1992). "A Comparison of Two Approaches to Beta-Flexible Clustering." *Multivariate Behavioral Research*, **27**(3), 417–433. doi:10.1207/s15327906mbr2703_6.

Bezanson J, Edelman A, Karpinski S, Shah V (2017). "Julia: A Fresh Approach to Numerical Computing." *SIAM Review*, **59**(1), 65–98. doi:10.1137/141000671.

Fernández A, Gómez S (2008). "Solving Non-Uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms." *Journal of Classification*, **25**(1), 43–65. doi:10.1007/s00357-008-9004-x.

Fernández A, Gómez S (2020). "Versatile Linkage: A Family of Space-Conserving Strategies for Agglomerative Hierarchical Clustering." *Journal of Classification*, **37**(3), 584–597. doi:10.1007/s00357-019-09339-z.

Galili T (2015). "**dendextend**: An R Package for Visualizing, Adjusting, and Comparing Trees of Hierarchical Clustering." *Bioinformatics*, **31**(22), 3718–3720. doi:10.1093/bioinformatics/btv428.

Gómez S, Fernández A (2021). "Radatools 5.2: communities detection in complex networks and other tools." URL https://deim.urv.cat/~sergio.gomez/radatools.php.

Gordon A (1999). *Classification*. 2nd edition. Chapman & Hall/CRC.

Hart G (1983). "The Occurrence of Multiple UPGMA Phenograms." In J Felsenstein (ed.), *Numerical Taxonomy*, pp. 254–258. Springer Berlin Heidelberg. doi:10.1007/978-3-642-69024-2_30.

IBM Corporation (2021). *IBM SPSS Statistics Base 28*. IBM Corporation, Armonk, NY, USA. URL https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Base.pdf.

Lance G, Williams W (1966). "A Generalized Sorting Strategy for Computer Classifications." *Nature*, **212**, 218. doi:10.1038/212218a0.

Lance G, Williams W (1967). "A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems." *The Computer Journal*, **9**(4), 373–380. doi:10.1093/comjnl/9.4.373.

MacCuish J, Nicolaou C, MacCuish N (2001). "Ties in Proximity and Clustering Compounds." *Journal of Chemical Information and Computer Sciences*, **41**(1), 134–146. doi:10.1021/ci000069q.

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2021). ***cluster**: Cluster Analysis Basics and Extensions*. R package version 2.1.2, URL https://CRAN.R-project.org/package=cluster.

Morgan B, Ray A (1995). "Non-Uniqueness and Inversions in Cluster Analysis." *Applied Statistics*, **44**(1), 117–134. doi:10.2307/2986199.

Paradis E, Schliep K (2019). "**ape 5.0**: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics*, **35**(3), 526–528. doi:10.1093/bioinformatics/bty633.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rousseeuw P (1986). "A Visual Display for Hierarchical Classification." In E Diday, Y Escoufier, L Lebart, J Pagés, Y Schektman, R Tomassone (eds.), *Data Analysis and Informatics, IV*, pp. 743–748. North-Holland, Amsterdam.

SAS Institute Inc (2018). *SAS/STAT 15.1 User's Guide*. SAS Institute Inc., Cary, NC, USA. URL http://documentation.sas.com/api/collections/pgmsascdc/9.4_3.4/docsets/statug/content/statug.pdf.

Segura-Alabart N, Serratosa F, Gómez S, Fernández A (2022). "Nonunique UPGMA clusterings of microsatellite markers." *Briefings in Bioinformatics*, **23**(5), bbac312. doi:10.1093/bib/bbac312.

Sneath P, Sokal R (1973). *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W.H. Freeman and Company.

Sokal R, Rohlf F (1962). "The Comparison of Dendrograms by Objective Methods." *Taxon*, **11**(2), 33–40. doi:10.2307/1217208.

StataCorp LLC (2021). *Stata 17*. StataCorp LLC, College Station, TX, USA. URL http://www.stata.com.

The MathWorks Inc (2022). *MATLAB — Statistics and Machine Learning Toolbox (R2022a)*. The MathWorks Inc., Natick, MA, USA. URL http://www.mathworks.com/help/stats/linkage.html.

van der Kloot W, Spaans A, Heiser W (2005). "Instability of Hierarchical Cluster Analysis Due to Input Order of the Data: The **PermuCLUSTER** Solution." *Psychological Methods*, **10**(4), 468–476. doi:10.1037/1082-989X.10.4.468.

Virtanen P, Gommers R, Oliphant T, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt S, Brett M, Wilson J, Millman K, Mayorov N, Nelson A, Jones E, Kern R, Larson E, Carey C, Polat İ, Feng Y, Moore E, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero E, Harris

C, Archibald A, Ribeiro A, Pedregosa F, van Mulbregt P, SciPy 10 Contributors (2020). "**SciPy** 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, **17**, 261–272. doi:10.1038/s41592-019-0686-2.

Williams W, Lambert J, Lance G (1966). "Multivariate Methods in Plant Ecology: V. Similarity Analyses and Information-Analysis." *Journal of Ecology*, **54**(2), 427–445. doi:10.2307/2257960.

Wolfram Language & System Documentation Center (2020). *Mathematica 12.1 — Hierarchical Clustering Package Tutorial*. Wolfram Research Inc., Champaign, IL, USA. URL https://reference.wolfram.com/language/HierarchicalClustering/tutorial/HierarchicalClustering.html.

Affiliation:

Alberto Fernández
Departament d'Enginyeria Química
Universitat Rovira i Virgili
Av. Països Catalans 26
43007 Tarragona, Spain
E-mail: alberto.fernandez@urv.cat

Sergio Gómez
Departament d'Enginyeria Informàtica i Matemàtiques
Universitat Rovira i Virgili
Av. Països Catalans 26
43007 Tarragona, Spain
E-mail: sergio.gomez@urv.cat
URL: https://deim.urv.cat/~sergio.gomez/