Versatile Linkage: a Family of Space-Conserving Strategies for Agglomerative Hierarchical Clustering

Alberto Fernández[1] · Sergio Gómez[2]

Abstract

Agglomerative hierarchical clustering can be implemented with several strategies that differ in the way elements of a collection are grouped together to build a hierarchy of clusters. Here we introduce versatile linkage, a new infinite system of agglomerative hierarchical clustering strategies based on generalized means, which go from single linkage to complete linkage, passing through arithmetic average linkage and other clustering methods yet unexplored such as geometric linkage and harmonic linkage. We compare the different clustering strategies in terms of cophenetic correlation, mean absolute error, and also tree balance and space distortion, two new measures proposed to describe hierarchical trees. Unlike the β-flexible clustering system, we show that the versatile linkage family is space-conserving.

Keywords Hierarchical clustering · Versatile linkage · Space distortion · Tree balance · Multidendrogram

1 Introduction

Agglomerative hierarchical clustering constitutes one of the most widely used methods for cluster analysis. Starting with a matrix of dissimilarities between a set of elements, each element is first assigned to its own cluster, and the algorithm sequentially merges the more similar clusters until a complete hierarchy of clusters is obtained (Sneath and Sokal 1973; Gordon 1999). This method requires the definition of the dissimilarity (or distance) between clusters, using only the original distances between their constituent elements. The way these distances are defined leads to distinct strategies of agglomerative hierarchical clustering. To name just two such clustering strategies, single linkage usually leads to an elongate

✉ Alberto Fernández
  alberto.fernandez@urv.cat

  Sergio Gómez
  sergio.gomez@urv.cat

[1] Departament d'Enginyeria Química, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain

[2] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain

growth of clusters, while complete linkage generally leads to tight clusters that join others with difficulty. Average linkage clustering strategies were developed by Sokal and Michener to avoid the extreme cases produced by single linkage and complete linkage (Sokal and Michener 1958). They require the calculation of some kind of average distance between clusters; average linkage, for instance, calculates the arithmetic average of all the distances between members of the clusters.

More than 50 years ago, Lance and Williams introduced a formula for integrating several agglomerative hierarchical clustering strategies into a single system (Lance and Williams 1966). Based on this formula, they proposed β-flexible clustering (Lance and Williams 1967), a generalized clustering procedure that provides an infinite number of hierarchical clustering strategies just varying a parameter β. Similarly, in this work, we introduce *versatile linkage*, a new parameterized family of agglomerative hierarchical clustering strategies that go from single linkage to complete linkage, passing through arithmetic average linkage and other clustering strategies yet unexplored such as geometric linkage and harmonic linkage.

Both β-flexible clustering and versatile linkage are presented here using variable-group methods (Sokal and Michener 1958; Fernández and Gómez 2008) that, unlike pair-group methods, admit any number of new members simultaneously into groups. In the case of pair-group methods the resulting hierarchical tree is called a dendrogram, which is built upon bifurcations, while in the case of variable-group methods the resulting hierarchical tree is called a *multidendrogram* (Fernández and Gómez 2008), which consists of multifurcations, not necessarily binary ones. Here we use the variable-group algorithm introduced in Fernández and Gómez (2008) that solves the non-uniqueness problem, also called the ties in proximity problem, found in pair-group algorithms (Sneath and Sokal 1973; Hart 1983; Day and Edelsbrunner 1984). This problem arises when there are more than two clusters separated by the same minimum distance during the agglomerative process. Pair-group algorithms break ties between distances choosing a pair of clusters, usually at random. However, different output dendrograms are possible depending on the criterion used to break ties. Moreover, very frequently results depend on the order of the elements in the input data file, which is an undesired effect in hierarchical clustering except for the case of contiguity-constrained hierarchical clustering, which is used to obtain a hierarchical clustering that takes into account the ordering on the input elements. The variable-group algorithm used here always gives a uniquely determined solution grouping more than two clusters at the same time when ties occur, and when there are no ties, it gives the same results as the pair-group algorithm.

Section 2 reviews the β-flexible family of hierarchical clustering strategies, while Section 3 introduces the versatile linkage family. Four case studies are used in Section 4 to perform a descriptive analysis of different hierarchical clustering strategies in terms of cophenetic correlation, mean absolute error, and the proposed new measures of space distortion and tree balance. Finally, some concluding remarks are given in Section 5.

2 β-Flexible Clustering

In any procedure implementing an agglomerative hierarchical clustering strategy, given a set of individuals $\Omega = \{x_1, x_2, \ldots, x_n\}$, initially each individual forms a singleton cluster, $\{x_i\}$, and the distances $D(\{x_i\}, \{x_j\})$ between singleton clusters are equal to the dissimilarities between individuals, $d(x_i, x_j)$. During the subsequent iterations of the procedure, the distances $D(X_I, X_J)$ are 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 $I = \{i_1, i_2, \ldots, i_p\}$ and $J = \{j_1, j_2, \ldots, j_q\}$, respectively. Lance and Williams introduced a formula for integrating several agglomerative hierarchical clustering strategies into a single system (Lance and Williams 1966). The variable-group generalization of Lance and Williams' formula, compatible with the fusion of more than two clusters simultaneously, is:

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

$$+ \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} \beta_{ii'} D(X_i, X_{i'}) + \sum_{j \in J} \sum_{\substack{j' \in J \\ j' > j}} \beta_{jj'} D(X_j, X_{j'}), \qquad (1)$$

where the values of the parameters α_{ij}, $\beta_{ii'}$, and $\beta_{jj'}$ determine the nature of the clustering strategy (Fernández and Gómez 2008). This formula is combinatorial (Lance and Williams 1967), i.e., the distance $D(X_I, X_J)$ can be calculated from the distances $D(X_i, X_j)$, $D(X_i, X_{i'})$ and $D(X_j, X_{j'})$ obtained from the previous iteration and it is not necessary to keep the initial distance matrix $d(x_i, x_j)$ during the whole clustering process.

Based on Eq. 1, Lance and Williams (1967) proposed an infinite system of agglomerative hierarchical clustering strategies defined by the constraint

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

where $-1 \leqslant \beta \leqslant +1$ generates a whole system of hierarchical clustering strategies for the infinite possible values of β. Given a value of β, the value for α_{ij} can be assigned following a weighted approach as in the original β-flexible clustering based on WPGMA (weighted pair-group method using arithmetic mean) and introduced by Lance and Williams (1966), or it can be assigned following an unweighted approach as in the β-flexible clustering based on UPGMA (unweighted pair-group method using arithmetic mean) and introduced by Belbin et al. (1992). The standard WPGMA and UPGMA strategies are obtained from weighted and unweighted β-flexible clustering, respectively, when β is set equal to 0. The difference between weighted and unweighted methods lies 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. In unweighted β-flexible clustering, the value for α_{ij} is determined proportionally to $|X_i||X_j|$:

$$\alpha_{ij} = \frac{|X_i||X_j|}{|X_I||X_J|}(1 - \beta), \qquad (3)$$

where $|X_i|$ and $|X_j|$ are the number of individuals in subclusters X_i and X_j, respectively, 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 $|X_J| = \sum_{j \in J} |X_j|$. In a similar way, the value for $\beta_{ii'}$ is calculated proportionally to $|X_i||X_{i'}|$, and the value for $\beta_{jj'}$ proportionally to $|X_j||X_{j'}|$:

$$\beta_{ii'} = \frac{|X_i||X_{i'}|}{\sigma_I + \sigma_J} \beta, \qquad (4)$$

$$\sigma_I = \sum_{i \in I} \sum_{\substack{i' \in I \\ i' > i}} |X_i||X_{i'}| = \frac{1}{2} \left(|X_I|^2 - \sum_{i \in I} |X_i|^2 \right). \qquad (5)$$

The corresponding values for weighted β-flexible clustering are:

$$\alpha_{ij} = \frac{1}{|I||J|}(1 - \beta), \tag{6}$$

$$\beta_{ii'} = \frac{1}{\sigma_I + \sigma_J}\beta, \tag{7}$$

$$\sigma_I = \frac{|I|\,(|I| - 1)}{2} = \frac{|I|^2 - |I|}{2}, \tag{8}$$

where $|I|$ and $|J|$ are the number of subclusters contained in clusters X_I and X_J, respectively. These formulas derive from the unweighted ones when we take $|X_i| = 1, \forall i \in I$, and $|X_j| = 1, \forall j \in J$.

3 Versatile Linkage

Arithmetic average linkage clustering iteratively forms clusters made up of previously formed subclusters, based on the arithmetic mean distances between their member individuals; for simplicity and to avoid confusion, we will denote it *arithmetic linkage* instead of the standard term *average linkage*. Substituting the arithmetic means by generalized means, also known as power means, this clustering strategy can be extended to any finite power $p \neq 0$:

$$
\begin{aligned}
D_p(X_I, X_J) &= \left(\frac{1}{|X_I||X_J|} \sum_{x \in X_I} \sum_{y \in X_J} [d(x, y)]^p \right)^{1/p} \\
&= \left(\frac{1}{|X_I||X_J|} \sum_{i \in I} \sum_{j \in J} |X_i||X_j| [D_p(X_i, X_j)]^p \right)^{1/p}.
\end{aligned}
\tag{9}
$$

We call this new system of agglomerative hierarchical clustering strategies as *versatile linkage*. As in the case of β-flexible clustering, versatile linkage provides a way of obtaining an infinite number of clustering strategies from a single formula. The second equality in Eq. 9 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 Lance and Williams' recurrence formula given in Eq. 1.

The decision of what power p to use could be taken in agreement with the type of distance employed to measure the initial dissimilarities between individuals. For instance, if the initial dissimilarities were calculated using a generalized distance of order p, then the natural agglomerative clustering strategy would be versatile linkage with the same power p. However, this procedure does not guarantee that the dendrogram obtained is the best according to other criteria, e.g., cophenetic correlation, mean absolute error, space distortion or tree balance (see Section 4). A better 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 dissimilarities between individuals are available, without coordinates for the individuals, as is common in multidimensional scaling problems, or when the dissimilarities have not been calculated using generalized means.

3.1 Particular Cases

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

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

$$D_{\min}(X_I, X_J) = \min_{x \in X_I} \min_{y \in X_J} d(x, y) = \min_{i \in I} \min_{j \in J} D_{\min}(X_i, X_j). \tag{10}$$

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

$$D_{\max}(X_I, X_J) = \max_{x \in X_I} \max_{y \in X_J} d(x, y) = \max_{i \in I} \max_{j \in J} D_{\max}(X_i, X_j). \tag{11}$$

There are also three other particular cases that can be grouped together as *Pythagorean linkages*:

- When $p = +1$, the generalized mean is equal to the arithmetic mean and *arithmetic linkage* (AL), i.e., the standard average linkage or UPGMA, is recovered.
- When $p = -1$, the generalized mean is equal to the harmonic mean and, therefore, *harmonic linkage* (HL) is obtained.
- In the limit when $p \to 0$, the generalized mean tends to the geometric mean. Hence, the distance definition for *geometric linkage* (GL) is:

$$\begin{aligned}
D_{\text{geo}}(X_I, X_J) &= \left(\prod_{x \in X_I} \prod_{y \in X_J} d(x, y) \right)^{1/(|X_I||X_J|)} \\
&= \left(\prod_{i \in I} \prod_{j \in J} [D_{\text{geo}}(X_i, X_j)]^{|X_i||X_j|} \right)^{1/(|X_I||X_J|)}.
\end{aligned} \tag{12}$$

To show the effects of varying the power p in versatile linkage clustering, we have built a small dataset with four individuals: Alice, Bob, Carol, and Dave, which lay on a straight line, separated between them by distances equal to 7, 9, and 12 units, respectively. Table 1 gives the pairwise distances between the four individuals, and Fig. 1 shows some multiden-drograms obtained varying the power p in versatile linkage clustering. Alice and Bob are always grouped together forming the first binary cluster, at a distance equal to 7.00. For values of the exponent $p \in (-\infty, 0)$, the Alice-Bob cluster is joined with Carol's singleton cluster at distances that range between 9.00 and 12.00. More precisely, this distance takes values 9.00 for SL ($p \to -\infty$), 11.52 for HL ($p = -1$) and 12.00 when we approach GL ($p \to 0^-$). For larger values of the exponent, $p > 0$, this distance becomes larger

Table 1 Sample pairwise distances between four individuals

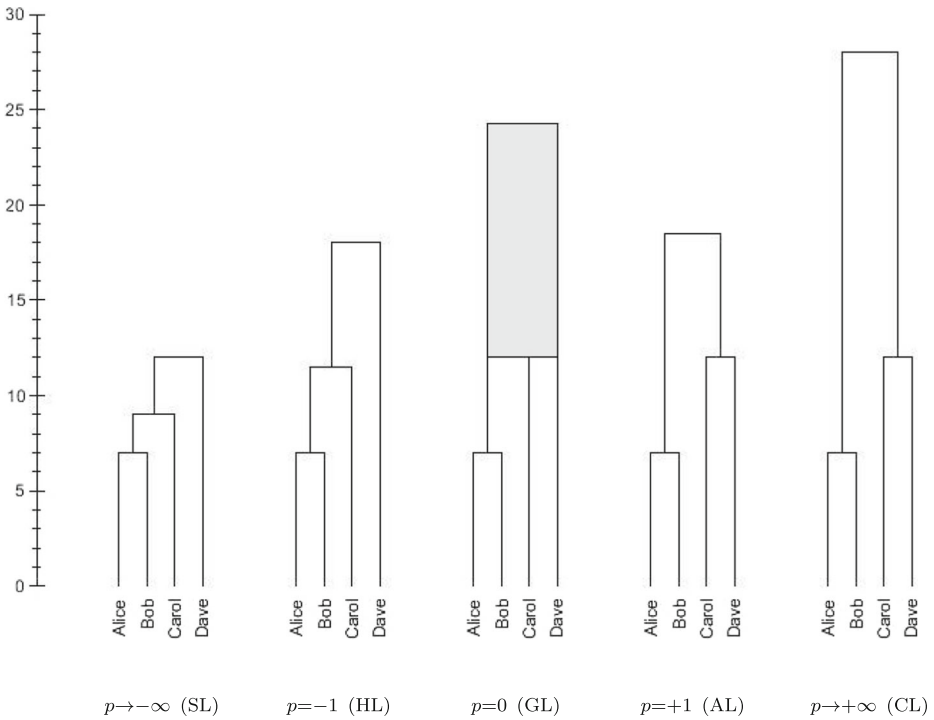|	Alice	Bob	Carol	Dave
Alice	0	7	16	28
Bob		0	9	21
Carol			0	12
Dave				0

Fig. 1 Effects of varying the power p in versatile linkage clustering for the sample distances in Table 1. Computations performed using the MultiDendrograms software (Gómez and Fernández 2018), with the precision parameter equal to 2 significant decimal digits. In MultiDendrograms, to avoid the infinite range of the exponent p, a sigmoidal transformation is performed such that the parameter used is within the range $[-1.0, +1.0]$, with values -1.0, -0.1, 0.0, $+0.1$, and $+1.0$ representing SL, HL, GL, AL, and CL, respectively. When $p = 0$ (GL), the gray band shows the existence of a tie between distances

than 12.00, thus Carol joins instead in a cluster with Dave at their distance 12.00. The remaining cluster for $p \in (-\infty, 0)$, which joins the Alice-Bob-Carol cluster with Dave, happens at heights 12.00 (SL), 18.00 (HL), and 19.18 ($p \to 0^-$), respectively. For the range $p \in (0, +\infty)$, the clusters Alice-Bob and Carol-Dave join at heights 17.06 ($p \to 0^+$), 18.50 (AL), and 28.00 (CL), respectively.

GL ($p = 0$) lays between these two structurally different dendrograms, represented as "(((Alice,Bob),Carol),Dave)" and "((Alice,Bob),(Carol,Dave))." Using pair-group agglomerative clustering methods, we would assign one of these two possible dendrograms to GL, thus breaking the tied pairs (Alice,Bob)-Carol and Carol-Dave (both at distance 12.00) randomly; this is an example of the ties in proximity (non-uniqueness) problem mentioned above. With the variable-group approach (Fernández and Gómez 2008), we join them at once forming the multidendrogram "((Alice,Bob),Carol,Dave)," where the three clusters join at distance 12.00, with a band going up to distance 24.25 to represent the heterogeneity of the new cluster, 24.25 being the distance between the clusters (Alice,Bob) and Dave (see middle multidendrogram in Fig. 1). This simple example shows the ability of versatile linkage to cover structurally different hierarchical clustering structures, including at the same time the traditionally important methods of SL, AL, and CL.

3.2 Weighted Versatile Linkage

Weighted clustering was introduced by Sokal and Michener (1958) in an attempt to give merging branches in a hierarchical tree equal weight regardless of the number of individuals carried on each branch. Such a procedure weights the individuals unequally, contrasting with unweighted clustering that gives equal weight to each individual in the clusters.

In weighted versatile linkage strategies, the distance between two clusters X_I and X_J is calculated by taking the generalized mean of the pairwise distances, not between individuals in the initial distance matrix, but between component subclusters in the matrix used during the previous iteration of the procedure, thus Eq. 9 being replaced by:

$$D_p(X_I, X_J) = \left(\frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} \left[D_p(X_i, X_j) \right]^p \right)^{1/p}. \tag{13}$$

3.3 Absence of Inversions

Versatile linkage strategies are monotonic, that is, they do not produce inversions. An inversion or reversal appears in a hierarchy when the hierarchy contains two clusters X and Y for which $X \subset Y$ but the height of cluster X is higher than the height of cluster Y (Murtagh 1985; Morgan and Ray 1995). Inversions make hierarchies difficult to interpret, specially if they occur during the last stages of the agglomeration process.

The monotonicity of versatile linkage strategies is explained by the Pythagorean means inequality,

$$\min \leqslant \text{HM} \leqslant \text{GM} \leqslant \text{AM} \leqslant \max, \tag{14}$$

where HM stands for the harmonic mean, GM for the geometric mean, and AM for the arithmetic mean. In the general case given by Eqs. 9 and 13, the generalized mean inequality holds:

$$D_p(X_I, X_J) \leqslant D_q(X_I, X_J), \qquad \forall p < q, \tag{15}$$

and $D_p(X_I, X_J) = D_q(X_I, X_J)$ if, and only if, the initial distances $d(x, y)$ are equal $\forall x \in X_I$ and $\forall y \in X_J$. Supposing that at a certain step of the clustering procedure the minimum distance between any two subclusters still to be merged is equal to δ, then the distance $D(X_i, X_j)$ between any two subclusters to be included in different clusters, $X_i \subseteq X_I$ and $X_j \subseteq X_J$, will be necessarily greater than δ, otherwise subclusters X_i and X_j would be merged into the same cluster. In particular, $D_{\min}(X_I, X_J) > \delta$. Therefore, taking into account the generalized mean inequality in Eq. 15, and given that in the limit when $p \to -\infty$ we have $D_p(X_I, X_J) = D_{\min}(X_I, X_J)$, we can conclude that $D_p(X_I, X_J) > \delta$, $\forall p$, which proves the absence of inversions of versatile linkage strategies.

4 Descriptive Analysis of Hierarchical Trees

We have selected four case studies, drawn from the UCI Machine Learning Repository (Lichman 2013), for a descriptive analysis of several agglomerative hierarchical clustering strategies. Table 2 summarizes the main characteristics of these datasets. The values of the variables in these datasets show different orders of magnitude; therefore, all the variables have been scaled first, and then, the corresponding dissimilarity matrices have been built using the Euclidean distance between all pairs of individuals.

Table 2 Characteristics of the selected datasets

Dataset	Instances	Features
Breast tissue (Jossinet 1996)	106	9
Iris (Fisher 1936)	150	4
Wine (Aeberhard et al. 1992)	178	13
Parkinsons (Little et al. 2009)	195	22

For the comparison of the hierarchical clustering strategies, we have chosen the following methods: β-flexible with $\beta = +0.9$, to avoid the completely flat hierarchical trees obtained with $\beta = +1$; versatile linkage with $p \to -\infty$, i.e., SL; centroid method; versatile linkage with $p = -1$, i.e., HL; versatile linkage with $p \to 0$, i.e., GL; versatile linkage with $p = +1$, which is the same as β-flexible with $\beta = 0$, i.e., AL; versatile linkage with $p \to +\infty$, i.e., CL; Ward's minimum variance method (Ward 1963); and β-flexible with $\beta = -1$. This selection includes five variants of versatile linkage, three of them equivalent to traditional methods (SL, AL, and CL) and the other two introduced in this work (HL and GL), and three variants of β-flexible clustering, one of them equivalent to AL.

Weighted and unweighted versions of the hierarchical clustering strategies have been used. Although weighting has no effect on SL and CL, we have included both of them for visual convenience in all the figures depicted next. The software used to run these experiments is MultiDendrograms (Gómez and Fernández 2018), which from version 5.0 implements all the hierarchical clustering strategies analyzed here and it also computes the necessary descriptive measures.

4.1 Cophenetic Correlation

The cophenetic correlation coefficient (CCC) measures the similarity between the distances in the initial matrix and the distances in the final ultrametric matrix obtained as result of a hierarchical clustering procedure (Sokal and Rohlf 1962). The ultrametric distance between two individuals is represented in a dendrogram by the height at which those two individuals are first joined. The CCC is calculated as the Pearson correlation coefficient between both matrices of distances; thus, the closer to 1, the largest their similarity.

In the analysis shown in Fig. 2, the CCC is higher for Pythagorean linkages (i.e., HL, GL and AL), and also the unweighted clustering strategies generally perform better than the weighted ones, corroborating the empirical observation already stated by Sneath and Sokal (1973). In the case of the almost flat hierarchical trees obtained with β-flexible clustering when $\beta = +1$, the CCC is very close to 0.

4.2 Mean Absolute Error

The CCC is a bounded measure that does not take into account how different the magnitudes of the distances in the initial matrix are from the distances in the final ultrametric matrix. For this reason, in Fig. 3 we show the normalized mean absolute error (MAE), which takes into account this type of differences. Note that in the case of the Iris dataset, Ward's method and β-flexible clustering with $\beta = -1$ showed a very good CCC in Fig. 2, while their MAE observed in Fig. 3 are the worst ones. As a matter of fact, β-flexible clustering with $\beta = -1$ yields results orders of magnitude worse than all the other methods, for the four datasets shown in Fig. 3. The best results are obtained again with Pythagorean linkages, and also unweighted clustering strategies are slightly better than the weighted ones.
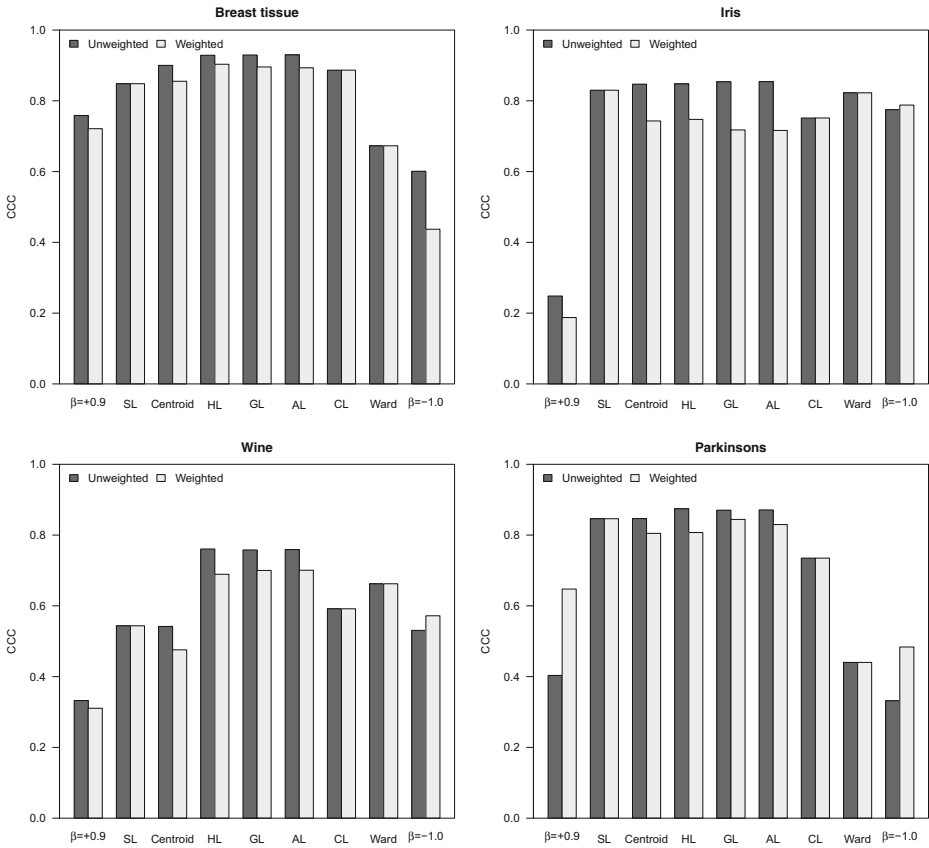
Fig. 2 Cophenetic correlation coefficient (CCC). Weighted and unweighted versions of the clustering strategies are compared

4.3 Space Distortion

For any agglomerative hierarchical clustering strategy, the initial distances between individuals may be regarded as defining a space with known properties (Lance and Williams 1967). When clusters begin to form, if the new distances between clusters are kept within the limits of the same space, then the original model remains unchanged and the clustering strategy is referred to as space-conserving. Otherwise, the clustering strategy is referred to as space-distorting. According to the formalization of the concept of space distortion (Dubien and Warde 1979), a clustering strategy is said to be space-conserving if

$$\min_{i \in I} \min_{j \in J} D(X_i, X_j) \leqslant D(X_I, X_J) \leqslant \max_{i \in I} \max_{j \in J} D(X_i, X_j). \tag{16}$$

On the contrary, a clustering strategy is space-contracting if the left inequality, delimited by SL, is not satisfied; and a clustering strategy is space-dilating if the right inequality, delimited by CL, is not satisfied. For space-contracting clustering strategies, as clusters grow in size, they move closer to other clusters. This effect is called chaining and it refers to the successive addition of elements to an ever expanding single cluster (Lance and Williams
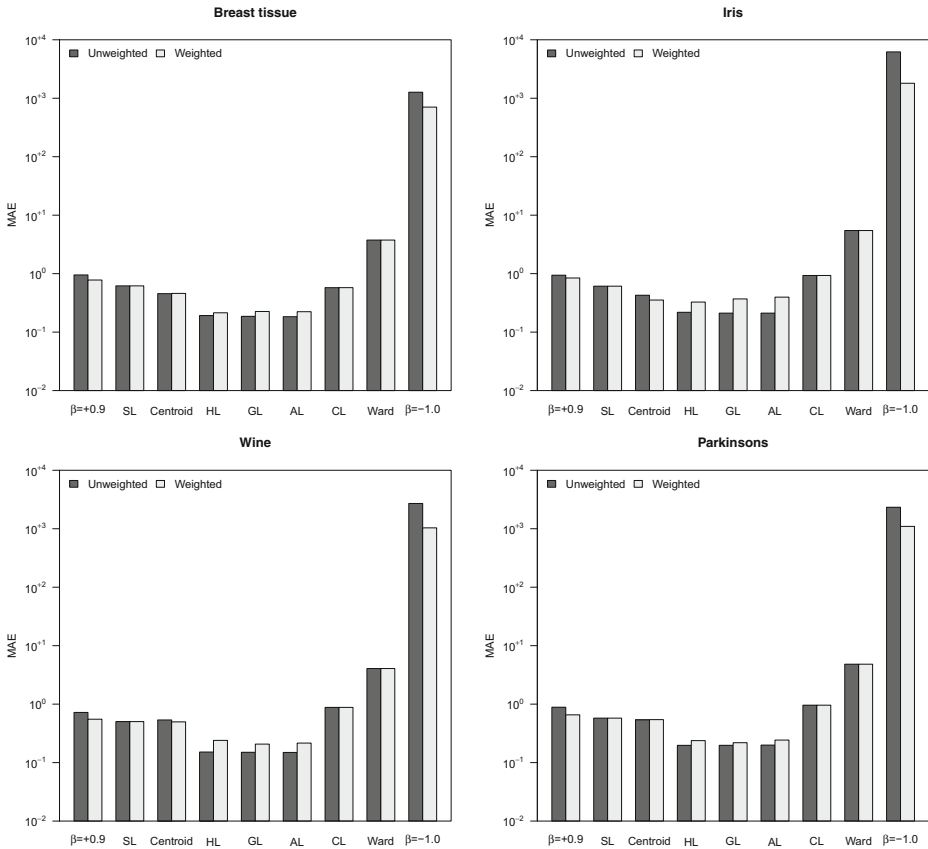
Fig. 3 Normalized mean absolute error (MAE), in logarithmic scale. Weighted and unweighted versions of the clustering strategies are compared

1967). Space-dilating clustering strategies produce the opposite effect, i.e., clusters moving further away from other clusters as they grow in size.

To numerically assess space distortion, we propose a *space distortion ratio* (SDR) measure, calculated as the quotient between the range of final ultrametric distances, $u(x_i, x_j)$, and the range of initial distances, $d(x_i, x_j)$:

$$\text{SDR}(u, d) = \frac{\max u(x_i, x_j) - \min u(x_i, x_j)}{\max d(x_i, x_j) - \min d(x_i, x_j)}. \tag{17}$$

The SDR is equal to 1 for CL; thus, this value separates space-conserving hierarchical trees from space-dilating ones. Figure 4 shows the SDR values corresponding to our four case studies. The outstanding differences between initial distances and ultrametric distances in the case of Ward's method and β-flexible clustering with $\beta = -1$, already observed in Fig. 3, allow the classification of both hierarchical clustering methods as space-dilating. With regard to weighting, it cannot be stated that neither weighted nor unweighted clustering strategies produce more space distortion: it depends on the particular dataset.

In Fig. 4 it can also be observed the increasing space distortion when β decreases in β-flexible clustering, or when the power p increases in versatile linkage clustering. Both
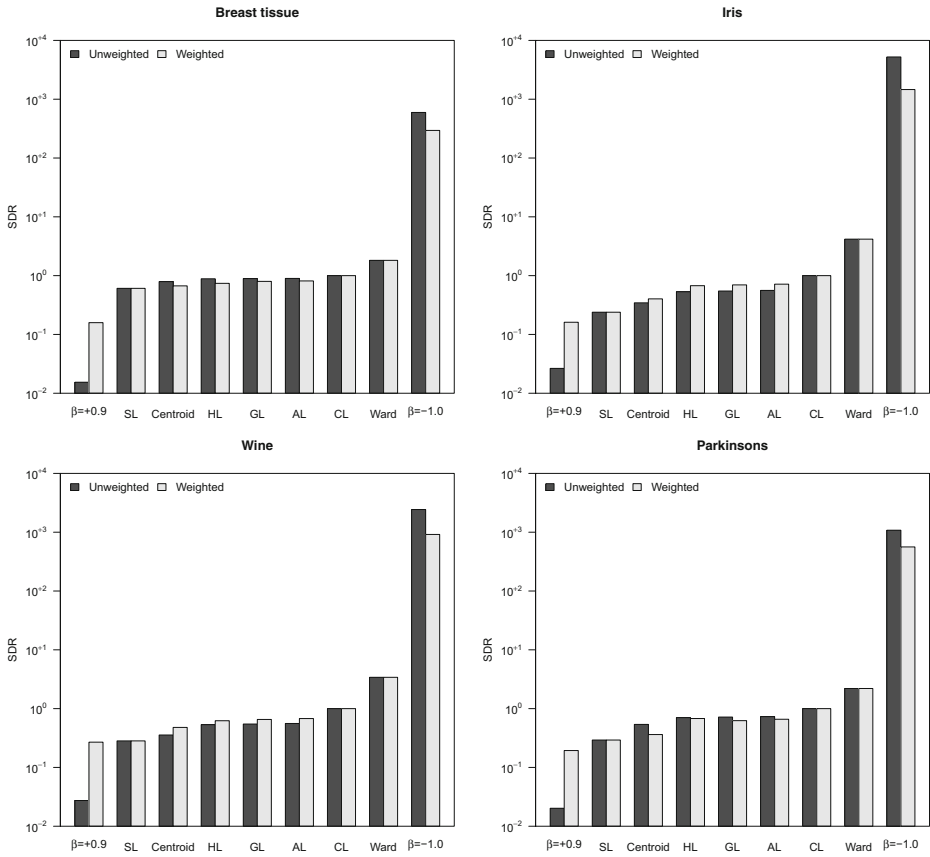
Fig. 4 Space distortion ratio (SDR), in logarithmic scale. Weighted and unweighted versions of the clustering strategies are compared

parameters, β and p, work as cluster intensity coefficients in their respective clustering systems. In the case of versatile linkage, the increasing space distortion when the power p increases is explained by the generalized mean inequality in Eq. 15. Therefore, taking also into account that, according to Eq. 16, space-conserving clustering strategies are lower bounded by SL ($p \rightarrow -\infty$) and upper bounded by CL ($p \rightarrow +\infty$), we can state that versatile linkage defines an infinite system of space-conserving strategies for agglomerative hierarchical clustering.

4.4 Tree Balance

We use the concept of entropy from information theory, more concretely Shannon's entropy (Shannon 1948), to introduce a new measure to assess the degree of homogeneity in size of the clusters in a hierarchical tree. Given a cluster X_I, we define its entropy as

$$H_I = -\sum_{i \in I} p_i \log_{|I|}(p_i), \tag{18}$$

where $p_i = \frac{|X_i|}{|X_I|}$ is the proportion of individuals in cluster X_I that are also members of sub-cluster X_i. Next, we define the *tree balance*, H, of a hierarchical tree as the average entropy of all its internal clusters. The maximum tree balance is equal to 1 and it is obtained, for instance, for a completely flat hierarchical tree with a single cluster containing the N individuals in the collection. Another example of hierarchical trees with maximum tree balance are the regular m-way trees obtained when applying the Baire-based divisive hierarchical clustering algorithm on a collection of sequences with uniformly distributed prefixes (Bradley 2010; Contreras and Murtagh 2012). On the contrary, the minimum tree balance, $H_{\min}$, corresponds to a binary tree where individuals are chained one at a time:

$$H_{\min} = \frac{1}{N-1}\left[\log_2(N) + \sum_{n=2}^{N-1}\frac{1}{n+1}\log_2(n)\right]. \tag{19}$$

Now, we can define the *normalized tree balance* (NTB) as

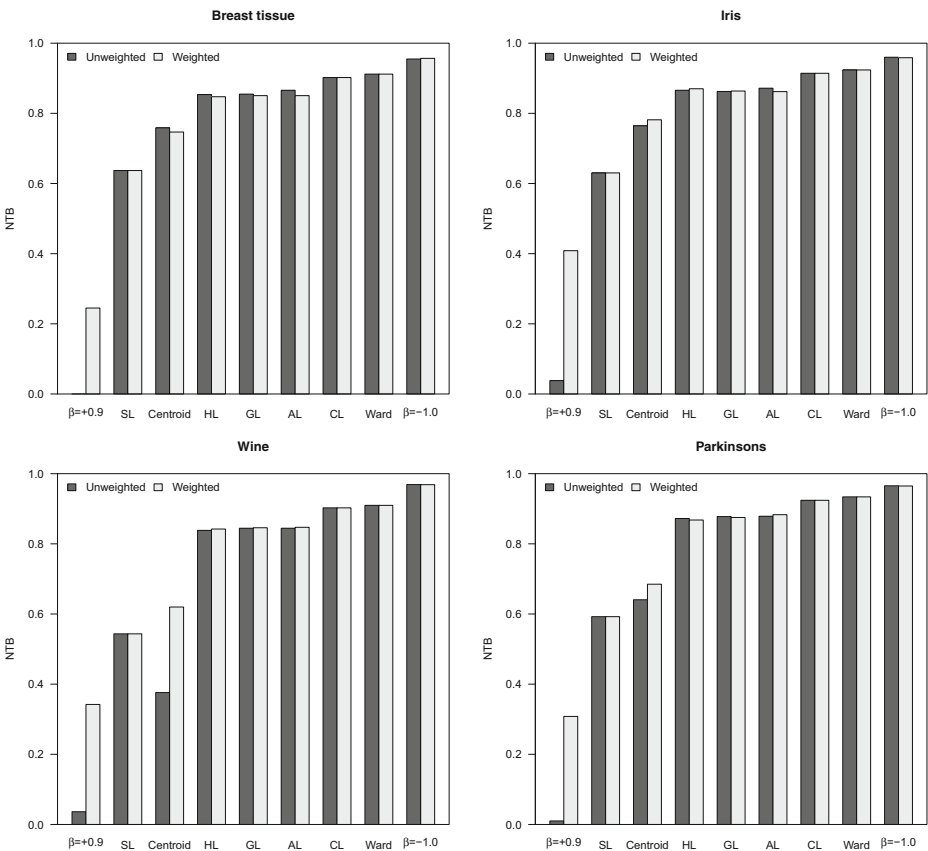$$\text{NTB} = \frac{H - H_{\min}}{1 - H_{\min}}, \tag{20}$$

Fig. 5 Normalized tree balance (NTB). Weighted and unweighted versions of the clustering strategies are compared

which becomes a measure with values between 0 and 1. Figure 5 shows the NTB values obtained for our case studies. Similarly to space distortion, tree balance increases when β decreases in β-flexible clustering, or when the power p increases in versatile linkage clustering. In the case of the almost flat hierarchical trees obtained with β-flexible clustering when $\beta = +1$, the NTB is very close to 1. Finally, according to the values observed in Fig. 5, it cannot be stated that neither weighted nor unweighted clustering strategies produce hierarchical trees that are more balanced.

5 Conclusions

Agglomerative hierarchical clustering methods have been continually evolving since their origins back in the 1950s, and historically they have been deployed in very diverse application domains, such as geosciences, biosciences, ecology, chemistry, text mining, and information retrieval, among others (Murtagh and Contreras 2017a). Nowadays, with the advent of the big data revolution, hierarchical clustering methods have had to address the new challenges brought by more recent application domains that require the hierarchical clustering of thousands of observations (Murtagh and Contreras 2017b).

In this work, we have introduced versatile linkage, an infinite family of agglomerative hierarchical clustering strategies based on the definition of generalized mean. We have shown that the versatile linkage family contains as particular cases not only the traditionally important strategies of single linkage, complete linkage, and arithmetic linkage, but also two new clustering strategies such as geometric linkage and harmonic linkage. In addition, we have given both weighted and unweighted versions of these hierarchical clustering strategies, and we have proved the monotonicity of versatile linkage strategies, which guarantees the absence of inversions in the hierarchy. Although we have built versatile linkage upon the multidendrograms variable-group methods to ensure the uniqueness of the clustering, it may also be used with the common pair-group approach just by breaking ties randomly.

We have shown that any descriptive analysis of hierarchical trees in terms of cophenetic correlation should be complemented with the use of other measures capable of describing the space distortion that different hierarchical clustering strategies cause. Under this point of view, we have shown that it is helpful to use other measures such as the mean absolute error or the space distortion ratio. The latter, in addition, provides a way to describe numerically the increase in space distortion observed all along a system of hierarchical clustering strategies such as versatile linkage.

Space distortion is inversely proportional to clustering intensity: space-contracting clustering strategies drive systems to cluster very weakly and produce a chaining effect, while space-dilating clustering strategies drive systems to cluster with high intensity and produce very compact clusters. These differences are described by the normalized tree balance measure introduced here, which is based on Shannon's entropy. Tree balance and space distortion are two new descriptive measures meant to be helpful to analyze and understand any hierarchical tree.

The β-flexible clustering also integrates an infinite number of agglomerative hierarchical clustering strategies into a single system, driven by a parameter β that works as a cluster intensity coefficient. However, to the best of our knowledge, no one has rigorously defined yet a range of values of β for which the corresponding β-flexible clustering strategies can be regarded as space-conserving. Unlike the β-flexible clustering system, we have shown that the versatile linkage family is space-conserving.

References

Aeberhard, S., Coomans, D., De Vel, O. (1992). Comparison of classifiers in high dimensional settings. Dept. Math. Statist., James Cook Univ., North Queensland, Australia, Tech. Rep. no. 92-02.

Belbin, L., Faith, D.P., Milligan, G.W. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research*, *27*(3), 417–433.

Bradley, P.E. (2010). Mumford dendrograms. *The Computer Journal*, *53*(4), 393–404.

Contreras, P., & Murtagh, F. (2012). Fast, linear time hierarchical clustering using the Baire metric. *Journal of Classification*, *29*(2), 118–143.

Day, W.H.E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*(1), 7–24.

Dubien, J.L., & Warde, W.D. (1979). A mathematical comparison of the members of an infinite family of agglomerative clustering algorithms. *Canadian Journal of Statistics*, *7*, 29–38.

Fernández, A., & Gómez, S. (2008). Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms. *Journal of Classification*, *25*(1), 43–65.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Gómez, S., & Fernández, A. (2018). MultiDendrograms: a hierarchical clustering tool (Version 5.0). http://deim.urv.cat/~sergio.gomez/multidendrograms.php.

Gordon, A.D. (1999). *Classification*, 2nd edn. Boca Raton: Chapman & Hall/CRC.

Hart, G. (1983). The occurrence of multiple UPGMA phenograms. In J. Felsenstein (Ed.) *Numerical taxonomy* (pp. 254–258). Berlin: Springer.

Jossinet, J. (1996). Variability of impedivity in normal and pathological breast tissue. *Medical and Biological Engineering and Computing*, *34*(5), 346–350.

Lance, G.N., & Williams, W.T. (1966). A generalized sorting strategy for computer classifications. *Nature*, *212*, 218.

Lance, G.N., & Williams, W.T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal*, *9*(4), 373–380.

Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., Ramig, L.O. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, *56*(4), 1015–1022.

Morgan, B.J.T., & Ray, A.P.G. (1995). Non-uniqueness and inversions in cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *44*(1), 117–134.

Murtagh, F. (1985). Multidimensional clustering algorithms. In *Compstat lectures*. Vienna: Physica-Verlag.

Murtagh, F., & Contreras, P. (2017a). Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(6), e1219.

Murtagh, F., & Contreras, P. (2017b). Clustering through high dimensional data scaling: applications and implementations. *Archives of Data Science, Series A*, *2*(1), 1–16.

Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Sneath, P.H.A., & Sokal, R.R. (1973). *Numerical taxonomy: the principles and practice of numerical classification*. San Francisco: W. H. Freeman and Company.

Sokal, R.R., & Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *The University of Kansas Science Bulletin*, *38*, 1409–1438.

Sokal, R.R., & Rohlf, F.J. (1962). The comparison of dendrograms by objective methods. *Taxon*, *11*(2), 33–40.

Ward, J.H. Jr. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.