Section: Numerical classification

# Cluster analysis (hierarchical agglomerative classification)

**Theory** R functions Examples

Cluster analysis is a hierarchical and agglomerative method of numerical classification, creating hierarchically nested groups (clusters) of samples by agglomerating individual samples into larger clusters. Two main choices need to be done: select the **distance metric**, which measures dissimilarity or similarity (in terms of species composition) between samples, and **clustering algorithm**, which searches for boundaries among groups of samples. For types of distances or similarities, consult Ecological resemblance. Overview of most often used clustering algorithms is below. Note that not all distances/similarities can be paired with all clustering algorithms. For example, Ward's method requires distances which are metric (i.e. can be projected into Euclidean space, with which the Ward algorithm is operating), and therefore non-metric distances (like Bray-Curtis) cannot be directly used (but some can be converted into metric; e.g., Bray-Curtis distance by square-root transformation).

Clustering algorithm represents a set of rules which decide to which group (cluster) will be the sample assigned. I like to imagine this problem as an analogy to me coming to the pub full of people (samples) and making a decision to which table (group, cluster) I will sit. Some of the people are my good friends (high similarity. low distance), some I really don't like (low similarity, high distance). Sitting at the table with people I do not like will perhaps ruin the night, so I need to choose carefully. I will call this "where to sit in the pub" analogy below.

Results or hierarchical cluster analysis is often displayed as a dendrogram, which represents the hierarchy on which samples and/or clusters are connected together. The orientation of branches around nodes is arbitrary (it does not matter whether the samples are right or left, Fig. 1).

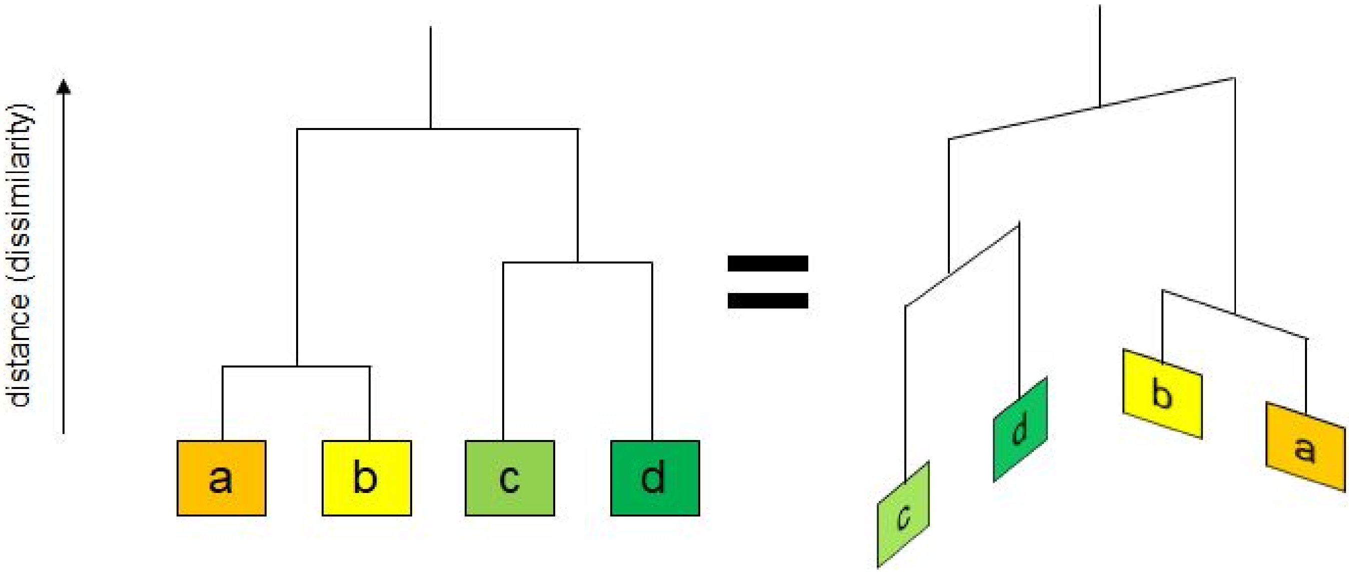
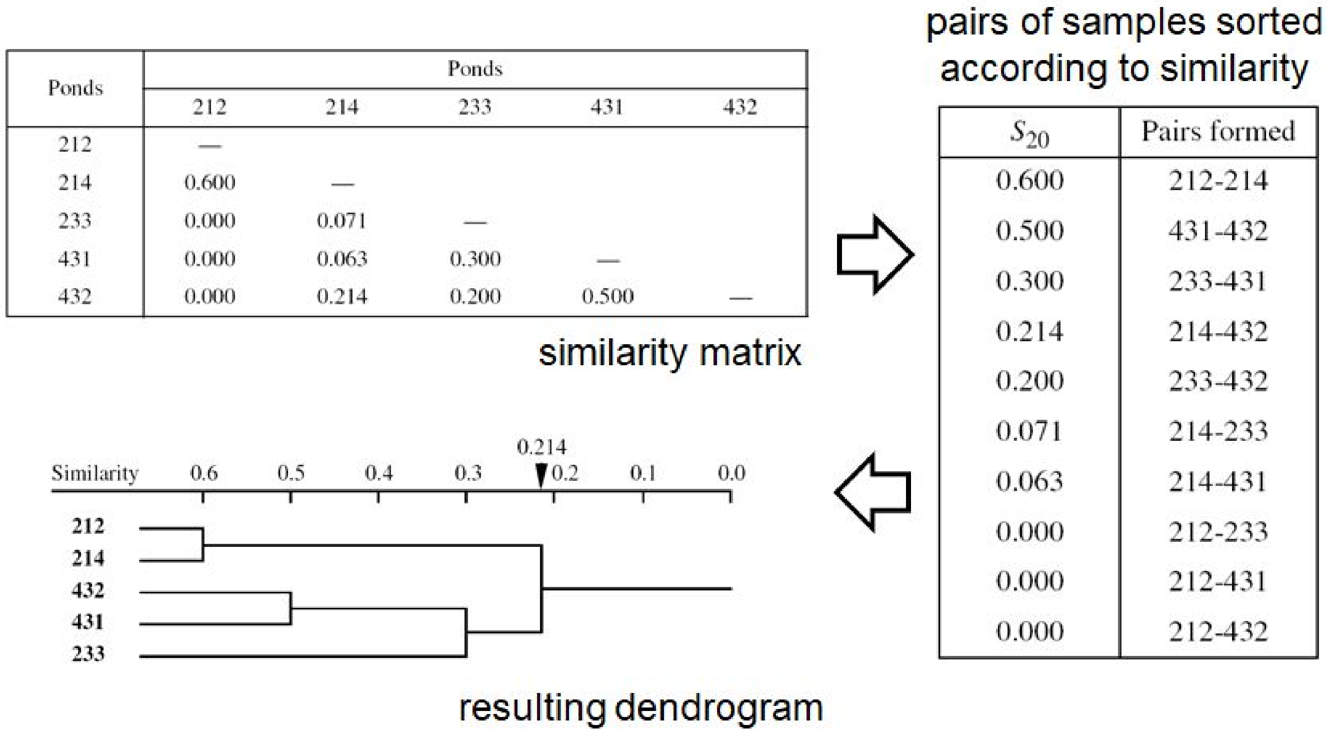


Figure 1: Dendrogram showing a hierarchical relationship between four clusters (a-c), with distance (dissimilarity) measure displayed on the vertical (y) axis. In the case of hierarchical agglomerative cluster analysis, the position of branches around nodes is arbitrary and they can be freely rotated (this is, however, not true e.g. in the case of TWINSPAN). (The rotation dendrogram schema (right) was inspired by Fig. 10.3. from McCune and Grace 2002.)

## Types of clustering alogirhtms

### Single linkage (nearest neighbour)

In the single linkage algorithm, samples join the group in which is the sample the most similar to them. In the “where to sit in the pub” analogy, this is like joining the table where sits my best friend (highest similarity, lowest distance), without considering the other people sitting with him/her. The example of how this algorithm works is on Fig. 2 (from Legendre & Legendre 2012) and example of resulting dendrogram on Fig. 3 (left panel).



| Ponds | Ponds | | | | |
|---|---|---|---|---|---|
| | 212 | 214 | 233 | 431 | 432 |
| 212 | — | | | | |
| 214 | 0.600 | — | | | |
| 233 | 0.000 | 0.071 | — | | |
| 431 | 0.000 | 0.063 | 0.300 | — | |
| 432 | 0.000 | 0.214 | 0.200 | 0.500 | — |

| $S_{20}$ | Pairs formed |
|---|---|
| 0.600 | 212-214 |
| 0.500 | 431-432 |
| 0.300 | 233-431 |
| 0.214 | 214-432 |
| 0.200 | 233-432 |
| 0.071 | 214-233 |
| 0.063 | 214-431 |
| 0.000 | 212-233 |
| 0.000 | 212-431 |
| 0.000 | 212-432 |

Figure 2: Single linkage algorithm of cluster analysis. First, the pairwise matrix of similarities is calculated, and then pairs of samples are sorted according to their compositional similarity. The most similar pair will form the first cluster (merging at S20 = 0.600), then the second most similar pair will form the second cluster (merging at S20 = 0.500). Sample 233 is the most similar to the sample 431, so it will join the second cluster (at S20 = 0.300). Finally, merging at S20 = 0.214 will connect all samples together. Example from Legendre & Legendre (2012); similarity coefficient S20, i.e. partial similarities, was used in this example (coding follows Legendre & Legendre 2012).

### Complete linkage (furthest neighbour)

The sample joins the group in which the furthest sample is the most similar. In the “where to sit in the pub” analogy, I will join the table in which the person I don’t like the most is still not that bad (each of other tables has someone I don’t like even more). Example of dendrogram is on Fig. 3 (central panel).

### Average linkage (group average)

The sample joins the group to which it has the lowest average distance. In the “where to sit in the pub” analogy, I will take table by table and average how much I like people sitting at that table; then, I join the table with the highest average (in average I like them most). The average linkage includes a range of methods in between single and complete linkage, and in ecology, these are the most useful (e.g. UPGMA, unweighted pair group method with arithmetic mean). Example dendrogram is on Fig. 3 (right panel).

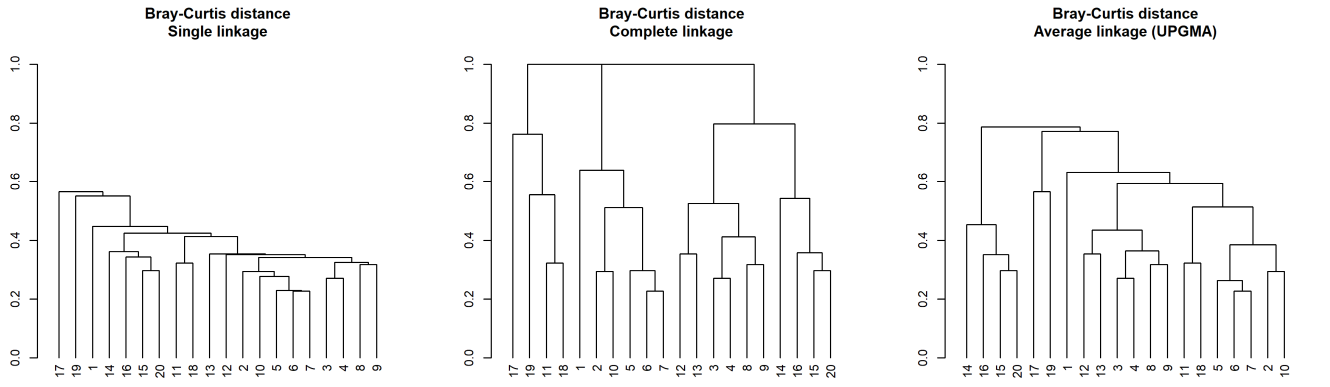
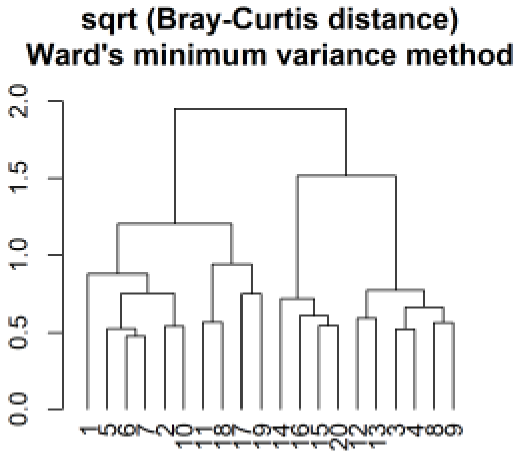


Figure 3: Comparison of dendrograms calculated by single linkage (left), complete linkage (centre) and average linkage (UPGMA, right) clustering algorithm, using the Bray-Curtis distance calculated on Dune dataset.

### Ward’s minimum variance method

Figure 4: Dendrogram of Ward's algorithm with Bray-Curtis distances (square rooted) calculated on the Dune dataset.



Ward’s clustering method is based on minimizing increases in the total error sum of squares, which is defined as the sum of squared distances of each sample to the centroid of its group. Ward’s method searches for the solution, which merges two groups of samples (clusters) with the lowest overall increase in the error sum of squares. In the beginning, each sample is forming an independent cluster, and the error sum of squares is equal to zero. In each step, the method merges a sample with the sample, sample with cluster or cluster with the cluster, and the error sum of squares is increasing. Since the distances of samples to the centroids are calculated in Euclidean space, only metric (Euclidean) distances can be used in this method; the use of non-metric distances (e.g. Bray-Curtis) is possible, but these need to be either converted into metric ones (e.g. by square-rooting in the case of Bray-Curtis dissimilarity, Fig. 4), or one has to acknowledge that the results may not follow the underlying principles of minimizing within-group sum of squares.

The total error sum of squares, used in Ward's and K-means clustering algorithms (Fig. 5): for each cluster *k*, it calculates the sum of square distances of individual samples ($y_{ij}^{(k)}$) to the centroid of the cluster ($m_j^{(k)}$; left formula), or, alternatively, as the mean squared distance among all pairs of samples ($D_{hi}^2$) within the cluster (right formula; both calculations yield the same result). These within-cluster error sums of squares are summed across all *K* clusters (formula at the bottom).

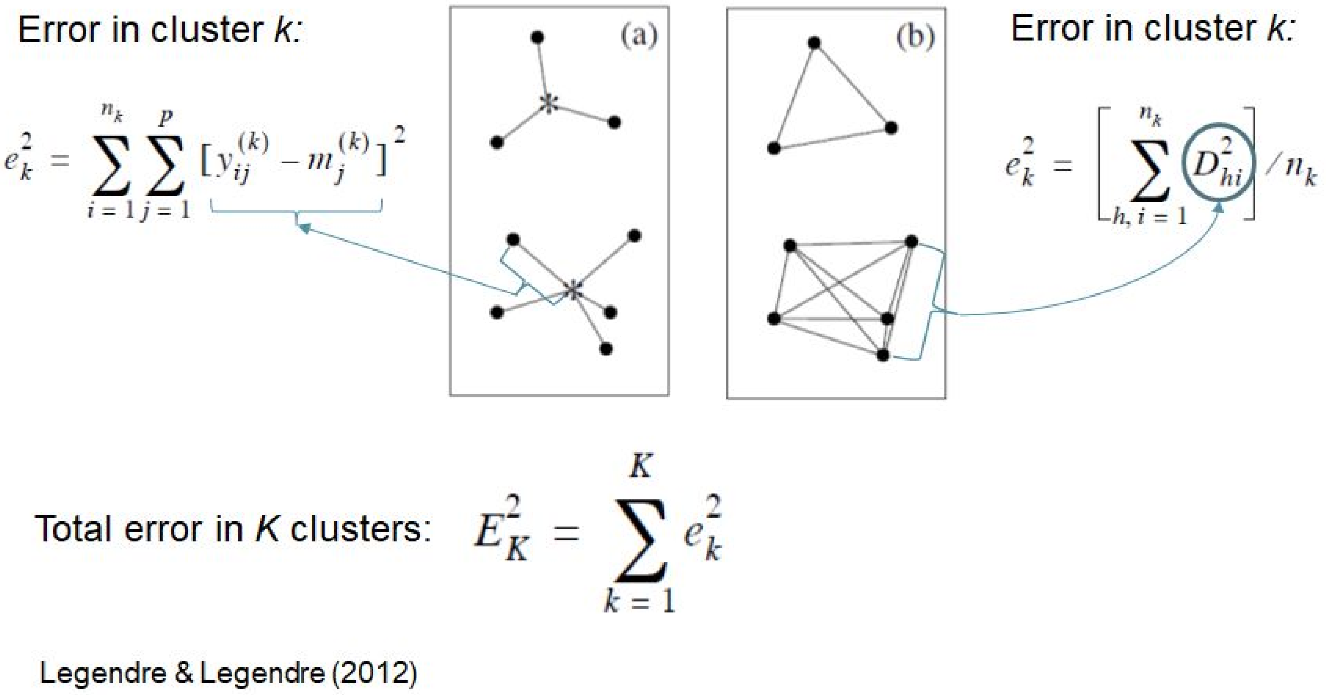

Error in cluster *k*:
$e_k^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{p} [y_{ij}^{(k)} - m_j^{(k)}]^2$
(a)
(b)
Error in cluster *k*:
$e_k^2 = \left[ \sum_{h,i=1}^{n_k} D_{hi}^2 \right] / n_k$
Total error in *K* clusters:
$E_K^2 = \sum_{k=1}^{K} e_k^2$
Legendre & Legendre (2012)

Figure 5: Two ways of calculating total error sum of squares (TOSS). Schema and equations from Legendre & Legendre (2012).

### Flexible beta linkage

Flexible clustering algorithm, which allows influencing the level of chaining by modifying the parameter β (β ≥ -1 and β < 1). With β = -1 the chaining is the lowest, with beta approaching +1 the chaining is highest (similar to simple linkage method, Fig. 6). Optimal representation of distances among samples is for β = -0.25 (results similar to Ward's algorithm). The method can be used with both metric and non-metric coefficients of similarity/dissimilarity.

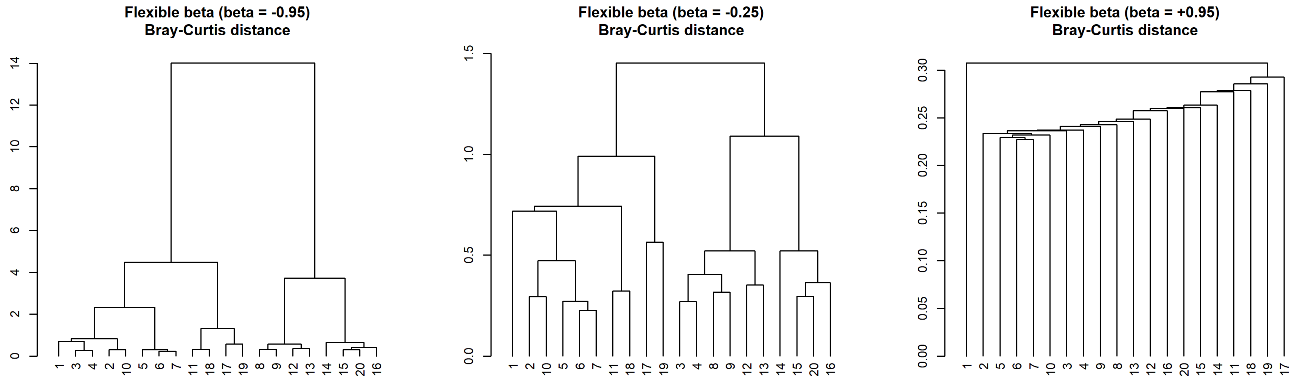


Figure 6: Flexible beta linkage algorithm applied on Bray-Curtis distance calculated from Dune

dataset; three alternative beta parameters result in quite different outcomes: beta = -0.95 (left), beta = -0.25 (middle) and beta = +0.95 (right).

## Effect of transformation applied on species composition data

In most clustering algorithms based on quantitative similarity/dissimilarity coefficient, the transformation of species composition data (e.g. logarithmic) can considerably influence resulting dendrogram (like on Fig. 7 in the case of single linkage algorithm applied on Bray-Curtis distances; Danube dataset). This does not apply if the original species composition data are presence-absence (in that case, the transformation does not change the relative distribution of the values).

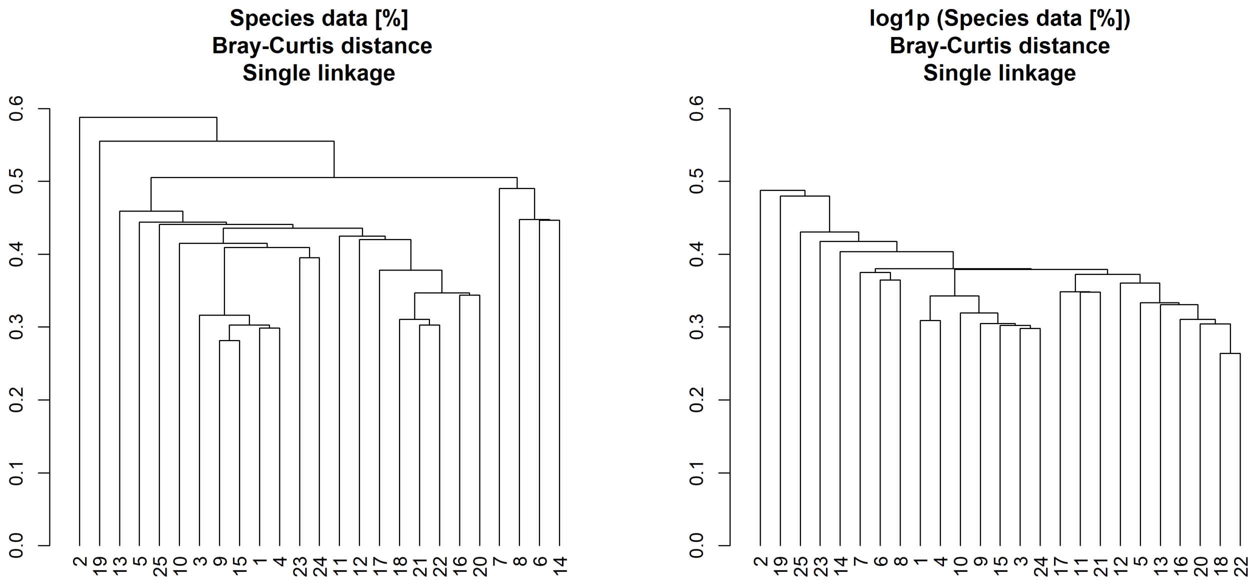


Figure 7: Effect of data transformation (log+1 in this case) on the results of classification (single linkage algorithm applied on Bray-Curtis distances from Danube dataset).