Numerical classification

The goal of numerical classification is to find discontinuities in community data (which may, in fact, be more or less continuous, Fig. 1) and name them – e.g. to ease the communication, or to see the compositional pattern more clearly. This is done by grouping similar objects (samples, species) into groups that are internally homogeneous while being well distinguishable from the other groups. In a specific case of community data (sample x species matrix), the classification can be either based on samples (resulting groups contain similar samples with similar species composition, e.g. because the communities occur on similar habitats), or on species (each group contains species with similar ecological behaviour).

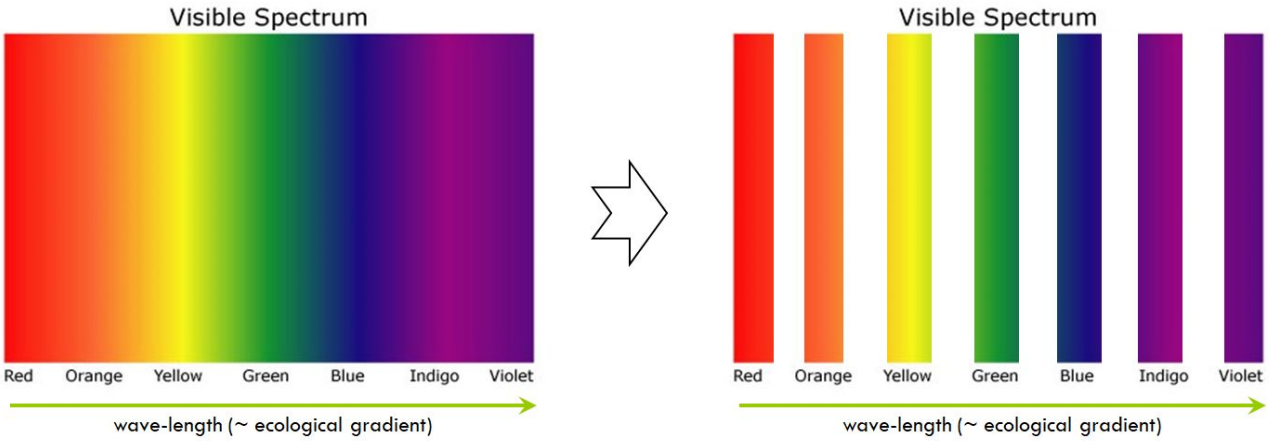

Figure 1: Even if the subject is truly continuous, it may be well worth to classify it into subsets and give these subset names. Example: although the visible spectrum is from principle continuous (depending on the wavelength), it is useful to separate it into relatively homogeneous subsets (colours) and give each subset a name.

Questions to ask before you begin classifying things

What you need the classification for? You want to classify your dataset (*sort books in your personal library*), or you want to create general classification schema, which can be used also by other people (*creating classification/sorting system for public libraries according to which they will sort books*)? In the first case, you may want to opt for unsupervised methods of classification, in the latter case for supervised methods (not discussed here in details).

What criteria you will use to classify things? How to quantify similarity among objects to decide how similar they are (for books this is reflecting the similarity in their content, e.g. group all Serbian horror novels together, or books similar in size). In the case of community samples, this equals to the selection of similarity index or distance measure.

How to decide where are the boundaries between groups? You need to choose rules used to classify objects into groups; in the case of community samples, this equals to the selection of clustering algorithm.

Types of classification methods

Simple “classification” of the numerical classification methods is in Fig. 2. The methods are either **hierarchical** or **non-hierarchical**, depending on whether the resulting groups of samples have a hierarchical relationship (some are more similar than others, which can be displayed by dendrogram) or not. In case of hierarchical methods, two alternatives exist: **divisive** algorithms, which take the whole dataset and cut it into subsets and these subsets into smaller subsets (in top-down direction), and **agglomerative** algorithms, which start from the level of individual samples and merge them together into larger groups (bottom-up direction).

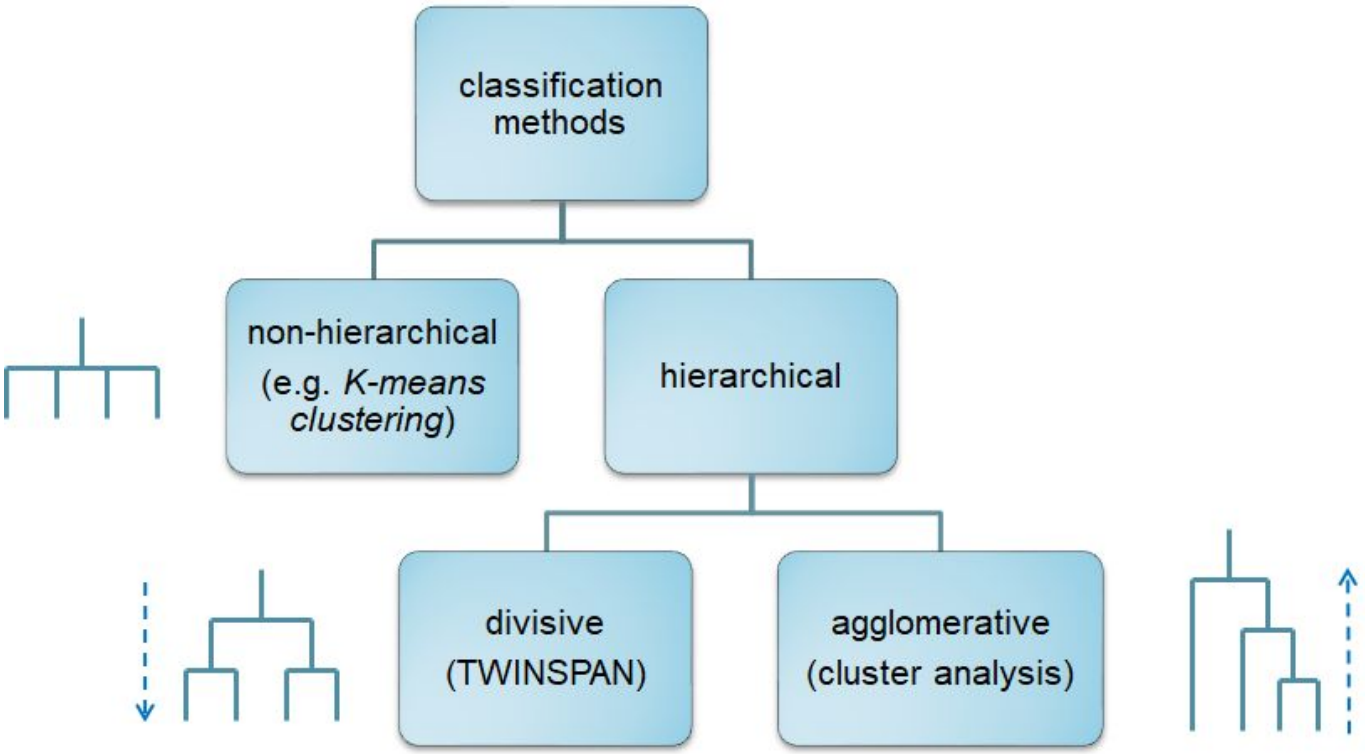

Figure 2: Classification of classification methods. The dendrograms beside the panels indicate whether the clusters (groups) are hierarchically or non-hierarchically related. The dashed arrow in the case of hierarchical methods indicates the direction of clustering – from higher to lower hierarchy in case of divisive algorithms (top-down direction) and from lower to higher hierarchies in the case of agglomerative algorithms (bottom-up direction).

Is the numerical classification producing “objective” results?

Some researchers argue that methods of numerical classification are creating an objective classification of objects that “really exist” (in contrast to a subjective classification which exists only because some other researchers believe in it). However, in reality, all classifications are from principle subjective, because researchers have done a number of subjective decisions before arriving at results (as is true for practically all analytical tasks). Example of such decisions in the case of cluster analysis is on the schema Fig. 3.

But although the numerical classification is not “objective”, it is “formalized” - based on clearly defined methods and reproducible (given that the data and chosen method is identical). In contrast,

the classification which is not formalized is based on criteria which are hard to describe (e.g. feelings, opinions, experience); such classification is difficult (or even impossible) to reproduce by the others, even if the data are the same.

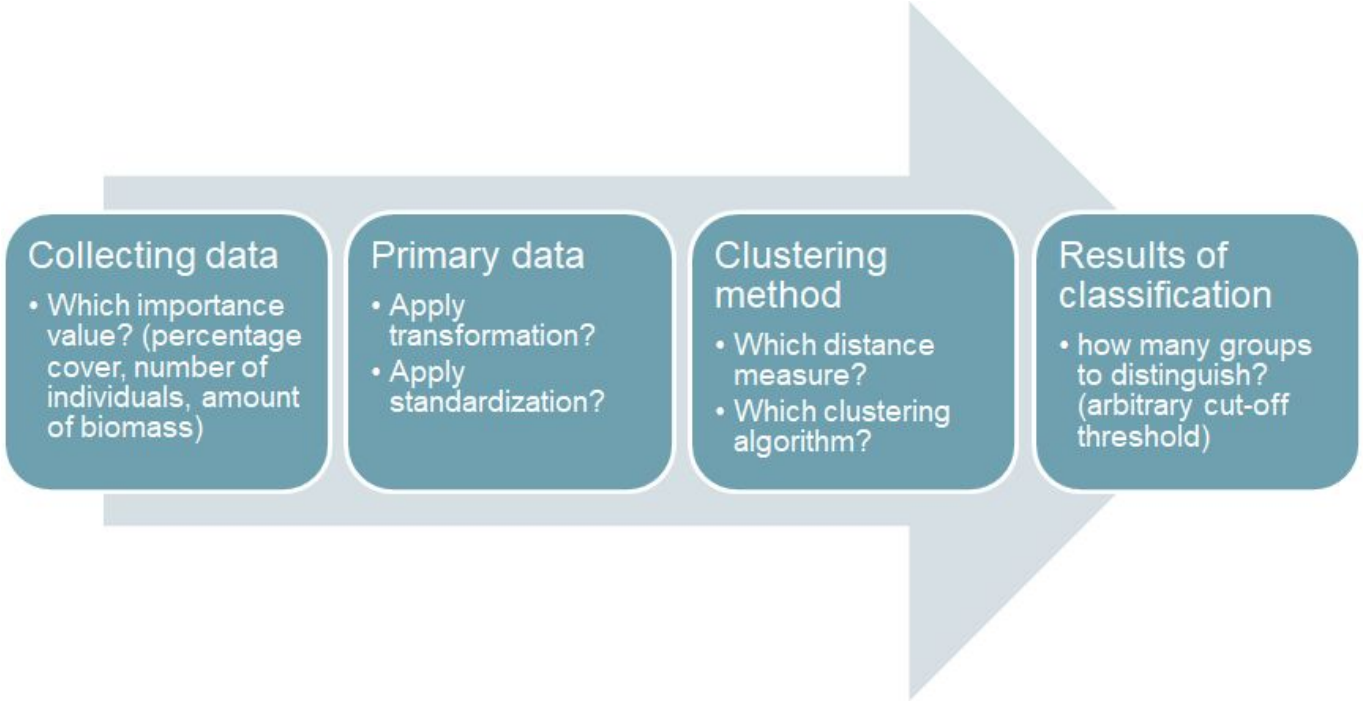

Figure 3: Example of subjective decisions done during the cluster analysis.

Unsupervised vs supervised classification

We can use classification methods in two alternative modes: unsupervised and supervised. **Unsupervised** methods search for main gradients in the species composition, main discontinuities or homogeneous groups of samples, and returns the result that is dependent only on the chosen method and internal structure of the dataset. In contrast, **supervised** classification methods use external criteria to classify the dataset – you can supply them with information about how to process the classification, and it will apply it on the existing dataset. In the case of unsupervised classification, one is able to modify the results by subjective choices (like clustering algorithm, distance metric, cut-off threshold for forming the groups), but the main results are dependent on the internal structure of the dataset and the assignment of samples into groups may change even with slight changes of the dataset (e.g. by adding more samples). In contrast, supervised methods are simply reproducing the classification criteria supplied externally, and assignment of the sample to the group will remain the same despite changes in the structure of the dataset.

Examples of unsupervised methods are TWINSPAN or cluster analysis, supervised methods (not discussed in detail on this website) include artificial neural networks (ANN), classification and regression trees (CART), random forests, COCKTAIL (logical formulas, designed for veg. data). Some methods, like K-means clustering, can run in either unsupervised or supervised mode – in the unsupervised mode the method first searches for the centroids of the predefined number of groups and assigns individual samples to these groups, while in the supervised mode the centroids are defined by user and the method just assigns the samples into these predefined groups.