Cost Action CA16208 "KNOWLEDGE CONVERSION FOR ENHANCING MANAGEMENT OF EUROPEAN RIPARIAN ECOSYSTEMS AND SERVICES" (CONVERGES) CA16208

Training School on Diversity and development of phytocoenological databases and using of different numerical methods for analysis of vegetation data
21-26 October, 2019, Sofia, Bulgaria

Agglomerative methods – PC-ORD

Agglomerative methods for classification – basic points

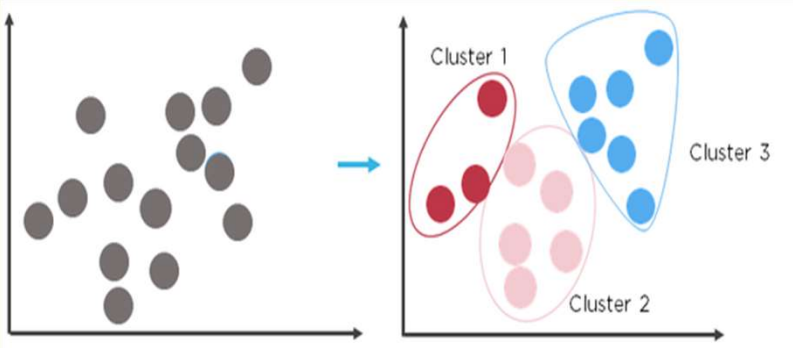
Desislava Sopotlieva

The position of agglomerative methods among numerical methods

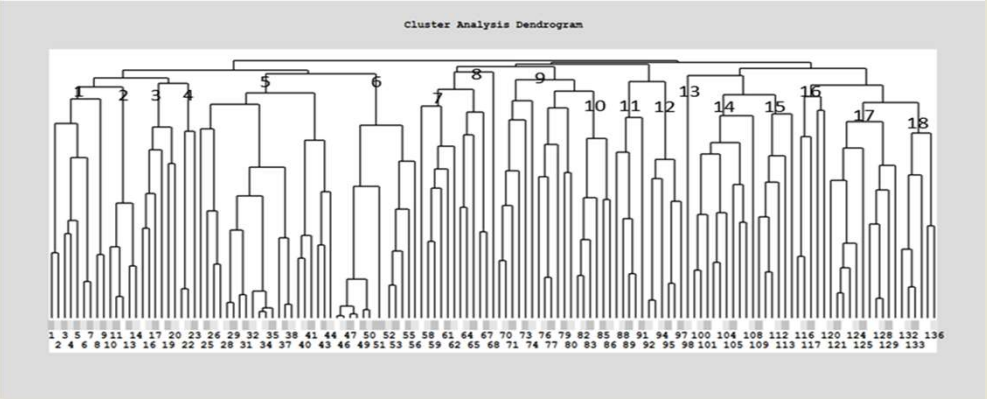
Cluster analysis or clustering

- Classification
- Grouping of objects into groups (clusters) according their similarity
- Not the only one specific kind of algorithm

Hierarchical and non-Hierarchical classification



Non-hierarchical



Hierarchical

The position of agglomerative methods among numerical methods

Supervised and unsupervised classification

- ***Unsupervised classification*** **–** calculated by software; depends of the chosen classification algorithm

- ***Supervised classification*** **–** human(expert)-guided; class definition; assignment criteria; examples of objects belonging to each class

Agglomerative methods are

Unsupervised hierarchical classification approaches

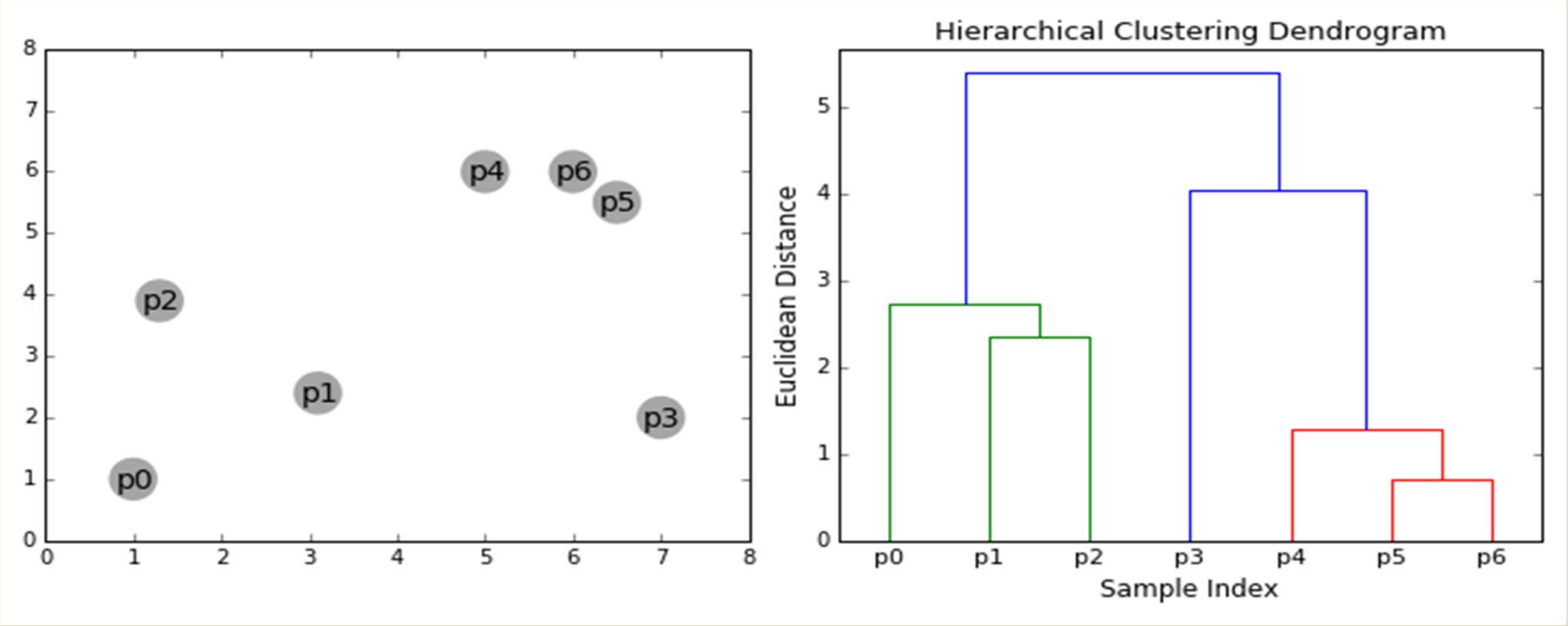
The nature of agglomerative methods

Divisive methods vs. Agglomerative methods

Divisive - „up-bottom" procedure

Agglomerative methods: "bottom- up" procedure

The classification process started with a single object (relevé) and associated to it into one group (cluster) the most similar ones, then joined the groups into bigger and bigger groups up to only one group



The nature of agglomerative methods

'Data Transformation' option – usually used to reduce the weight of higher cover values

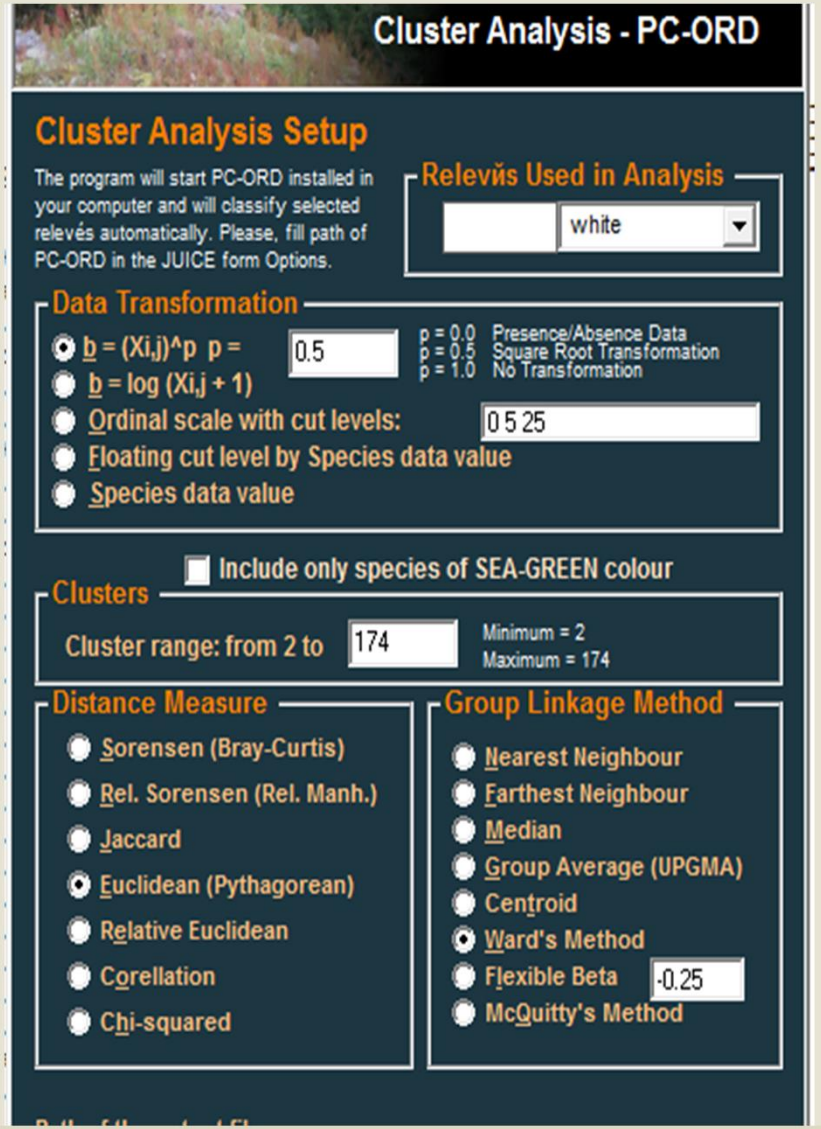
Two main stages of calculation procedure

- **Calculation of similarity/dissimilarity between objects**

→dissimilarity matrix

→Distance Measure

- **Clustering method (Group Linkage Method) – sorting strategy**



The nature of agglomerative methods
Distance Measure's calculation

Jaccard index (Jaccard similarity coefficient)

The Jaccard coefficient measures similarity between certain samples, and is defined as the size of the intersection divided by the size of the union of the sample sets

J (rel1, rel2) = join species(rel1, rel2)/join species(rel1, rel2)+species(rel1)+species(rel2)
Example: rel 1 - 5 species; rel 2 - 6 species; join species(rel1, rel2) – 3 species
→ J = 3/ 3+5+6= 3/14 = 0.214

0≤ J (A, B) ≥1
J (rel1,rel2) = 0 → absolutely different according plant species composition
J (rel1,rel2) = 1 → absolutely the same according plant species composition

Jaccard distance

dJ (A,B) = 1- J (A,B)

Use qualitative data – presence/absence of plant species

The nature of agglomerative methods
Distance Measure's calculation

Sørensen index (Sørensen similarity coefficient)

Very similar to Jaccard index, but give twice weight to the join species

Use qualitative data – presence/absence of plant species

Bray–Curtis dissimilarity

Similar to *Sørensen index* in term of twice weight to the join species, but **use quantitative data – abundance of plant species**

Dimension is between 0 and 1; It is dissimilarity index → 0 – absolute similarity

Sørensen vs. Bray–Curtis dissimilarity

The nature of agglomerative methods
Distance Measure's calculation

Table 6.2. Reasonable and acceptable domains of input data, x, and ranges of distance measures, $d = f(x)$.

Name (synonyms)	Domain of x	Range of $d = f(x)$	Comments
Sørensen (Bray & Curtis; Czekanowski)	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq x \leq 100\%$)	proportion coefficient in city-block space; semimetric
Relative Sørensen (Kulczynski; Quantitative Symmetric)	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq x \leq 100\%$)	proportion coefficient in city-block space; same as Sørensen but data points relativized by sample unit totals; semimetric
Jaccard	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq d \leq 100\%$)	proportion coefficient in city-block space; metric
Euclidean (Pythagorean)	all	non-negative	metric
Relative Euclidean (Chord distance; standardized Euclidean)	all	$0 \leq d \leq \sqrt{2}$ for quarter hypersphere; $0 \leq d \leq 2$ for full hypersphere	Euclidean distance between points on unit hypersphere; metric
Correlation distance	all	$0 \leq d \leq 1$	converted from correlation to distance; proportional to arc distance between points on unit hypersphere; cosine of angle from centroid to points; metric
Chi-square	$x \geq 0$	$d \geq 0$	Euclidean but doubly weighted by variable and sample unit totals; metric
Squared Euclidean	all	$d \geq 0$	metric
Mahalanobis	all	$d \geq 0$	distance between groups weighted by within-group dispersion; metric

after McCune& Grace (2002)

The nature of agglomerative methods
Distance Measure's calculation

XLSTAT by Addinsoft PRODUCTS TRAINING TRIAL SUPPORT ORDER 0 ENGLISH MYXL

XLSTAT proposes several similarities/dissimilarities that are suitable for a particular type of data:

	Similarity	Dissimilarity
Quantitative data	Pearson's coefficient of correlation Spearman's coefficient of rank correlation Kendall's coefficient of rank correlation Inertia Covariance (n) Covariance (n-1) Percent agreement	Euclidean distance Chi-square distance Manhattan distance Pearson's dissimilarity Spearman's dissimilarity Kendall's dissimilarity Percent disagreement
Binary data (0/1)	Jaccards coefficient Dice coefficient Sokal & Sneath coefficient (2) Rogers & Tanimoto coefficient Simple matching coefficient Indice de Sokal & Sneath coefficient (1) Phi coefficient Ochiais coefficient Kulczinskis coefficient Percent agreement	Jaccards coefficient Dice coefficient Sokal & Sneath coefficient (2) Rogers & Tanimoto coefficient Simple matching coefficient Indice de Sokal & Sneath coefficient (1) Phi coefficient Ochiais coefficient Kulczinskis coefficient Percent agreement

https://www.xlstat.com/en/solutions/features/agglomerative-hierarchical-clustering-ahc

The nature of agglomerative methods
Group Linkage Method

Internet sources -main

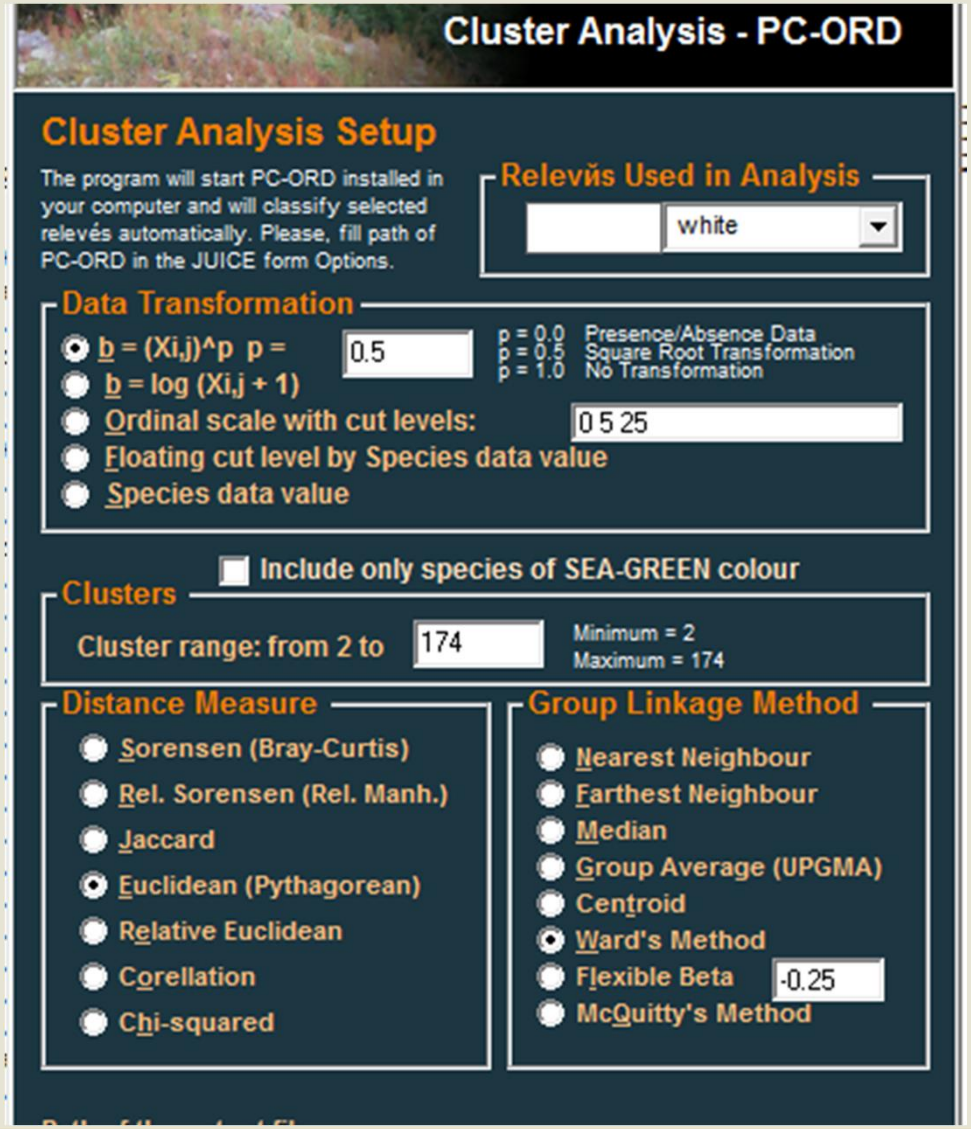
Names
Maximum or complete-linkage clustering
Minimum or single-linkage clustering
Unweighted average linkage clustering (or UPGMA)
Weighted average linkage clustering (or WPGMA)
Centroid linkage clustering, or UPGMC
Minimum energy clustering

Internet sources-others

….

Wards's method

……



The nature of agglomerative methods
Group Linkage Method

- **Single linkage clustering or Minimum or Nearest Neighbour**

  The distance between two clusters is defined as the minimum distance value of the nearest pair between the elements in cluster 1 and the elements in cluster 2

- **Complete-linkage clustering or Maximum or Farthest Neighbour**

  The distance between two clusters is defined as the minimum distance value of farthest pair between the elements in cluster 1 and the elements in cluster 2

- **Average linkage clustering (for expl. Unweighted Pair-Groups Method using Arithmetic averaging or UPGMA)**

  The distance between two clusters is defined as the minimum distance value of average distance between the all elements in cluster 1 and the elements in cluster 2

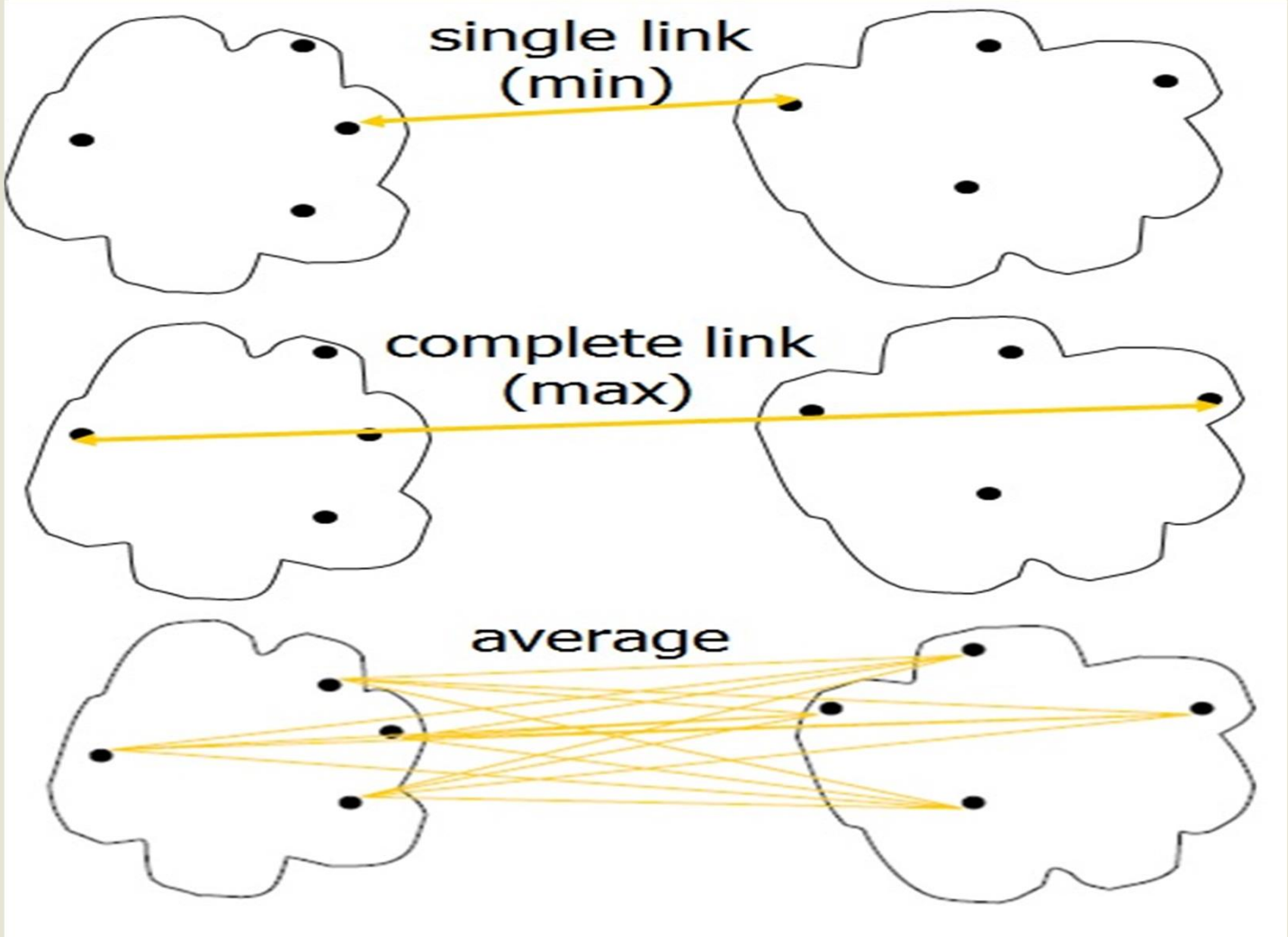
- **Centroid linkage**

  The distance between two clusters is defined as the distance between the centroid for cluster 1 (a mean vector of length p variables) and the centroid for cluster 2.

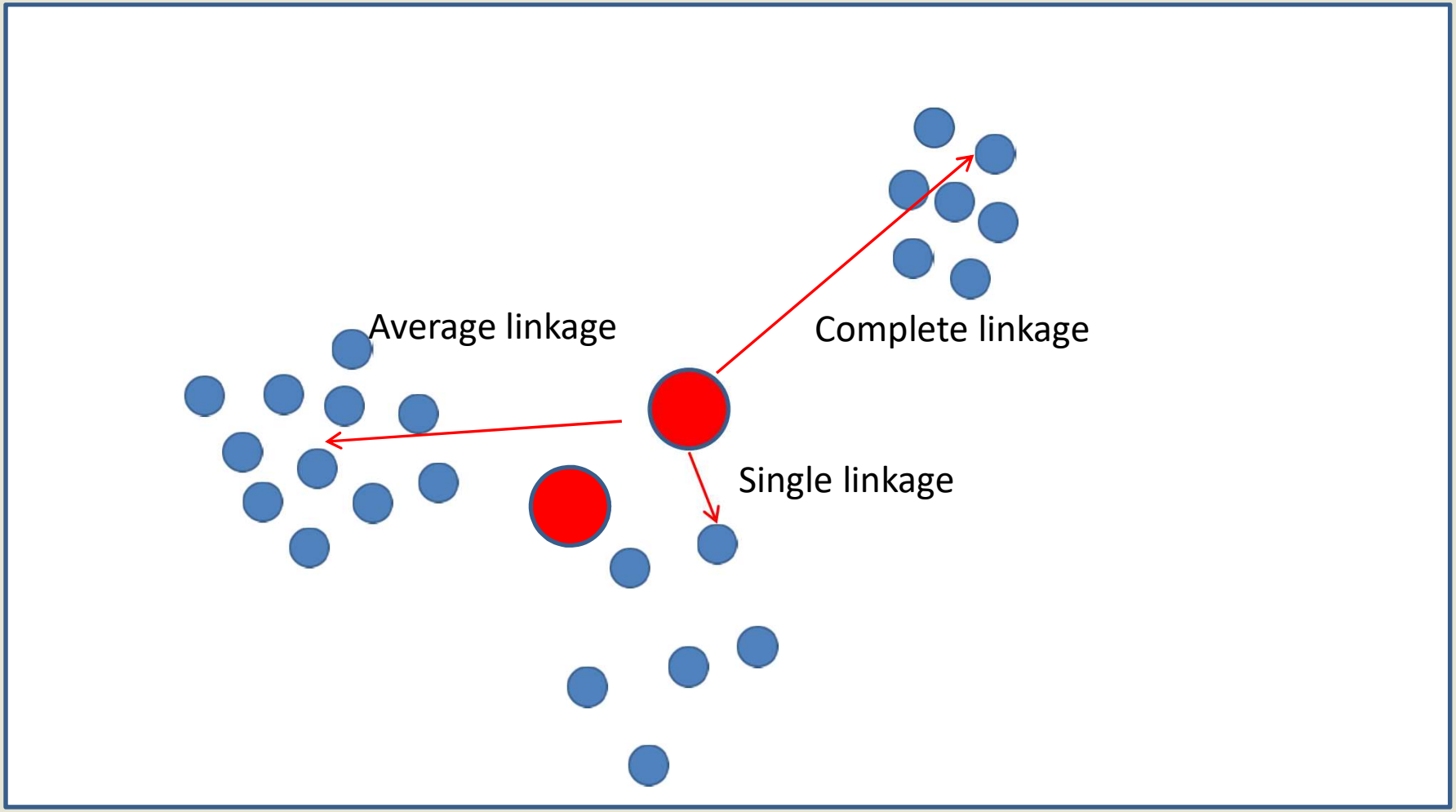
- **Ward's (minimum variance) method**

  It minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged

The nature of agglomerative methods
Group Linkage Method

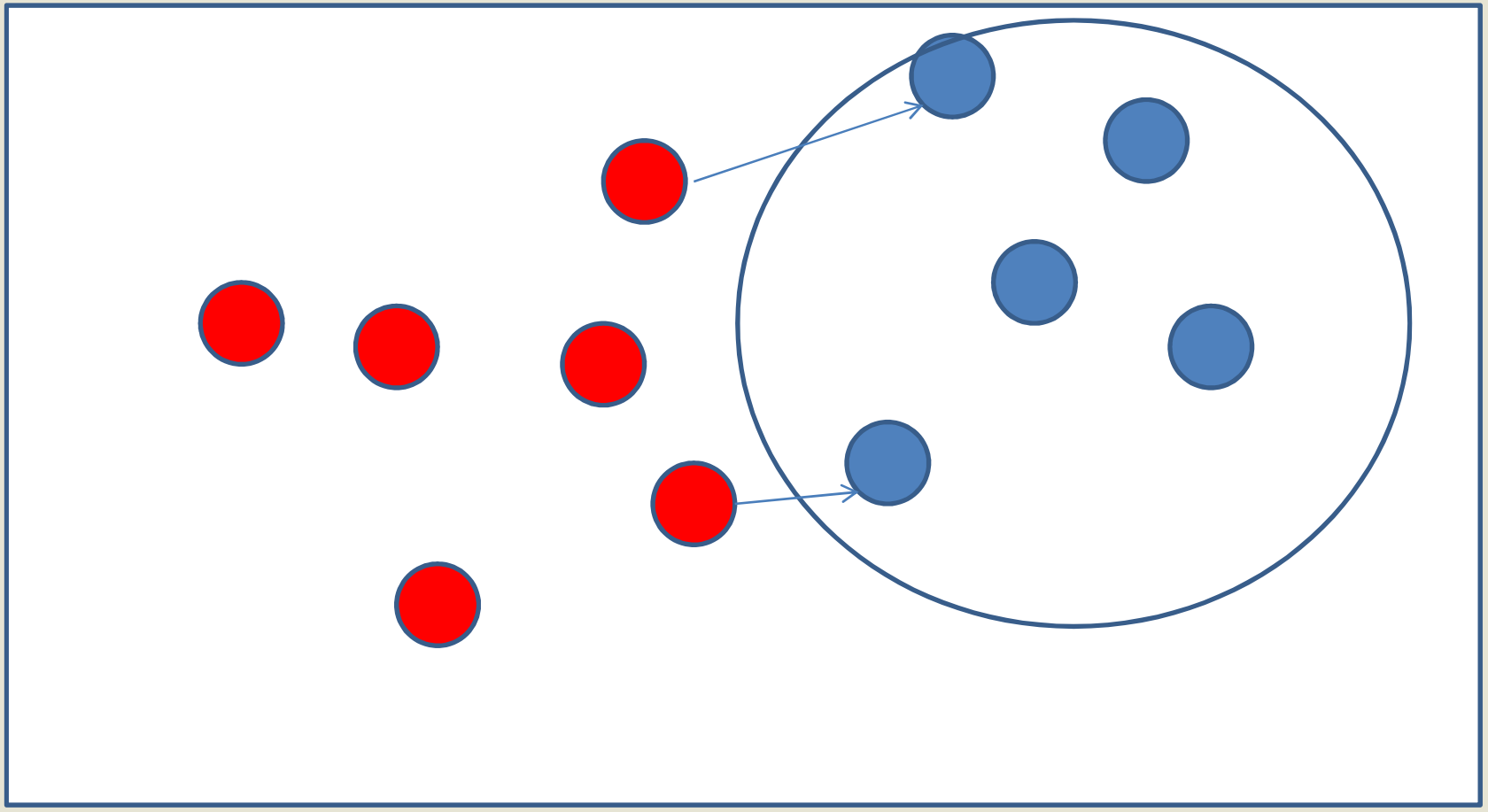


The nature of agglomerative methods
Group Linkage Method



The nature of agglomerative methods
Group Linkage Method

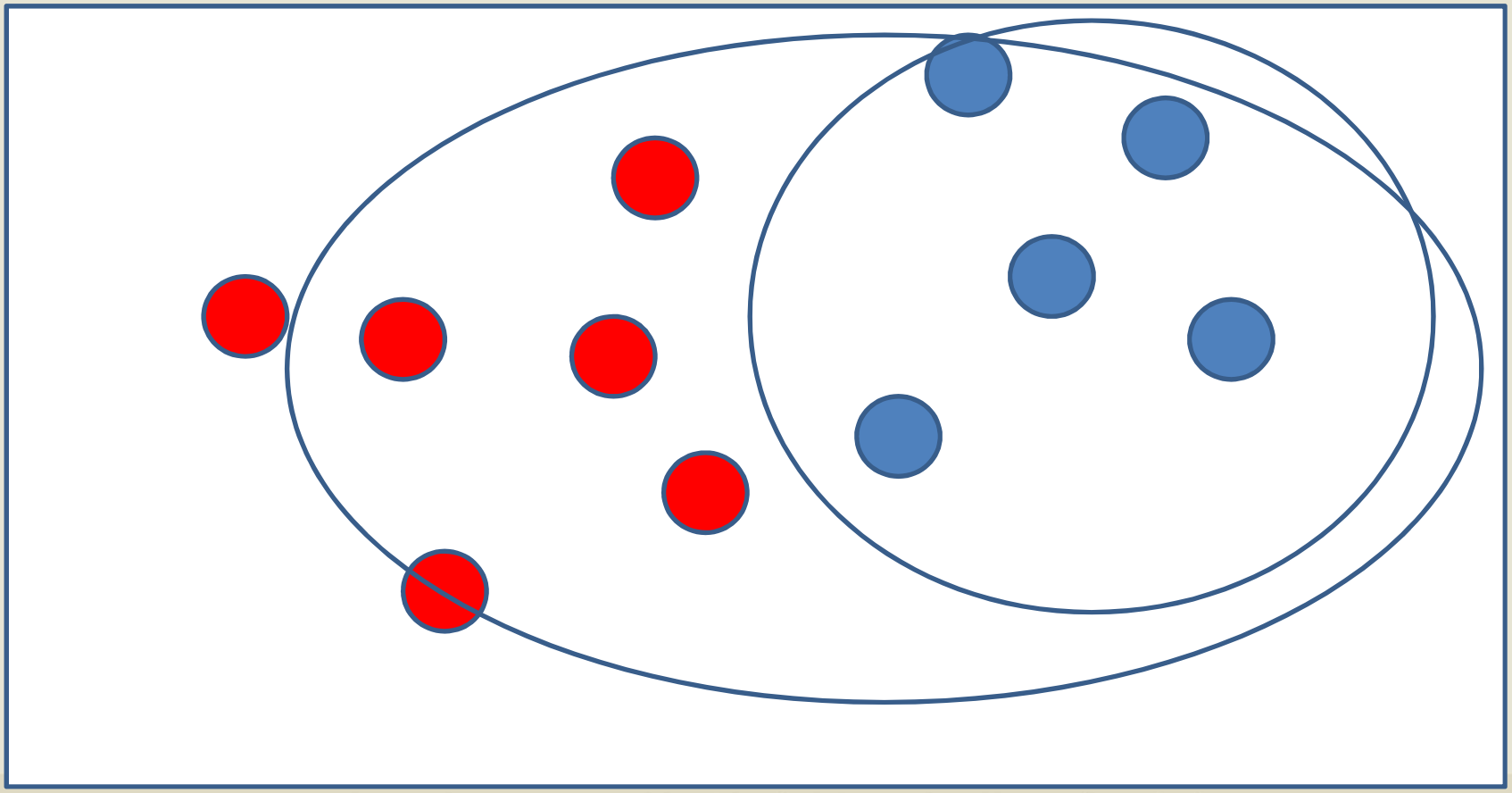
Single linkage clustering or Minimum or Nearest Neighbour



The nature of agglomerative methods
Group Linkage Method

Single linkage clustering or Minimum or Nearest Neighbour

A tendency to produce “straggly” clusters; clusters are rather heterogeneous

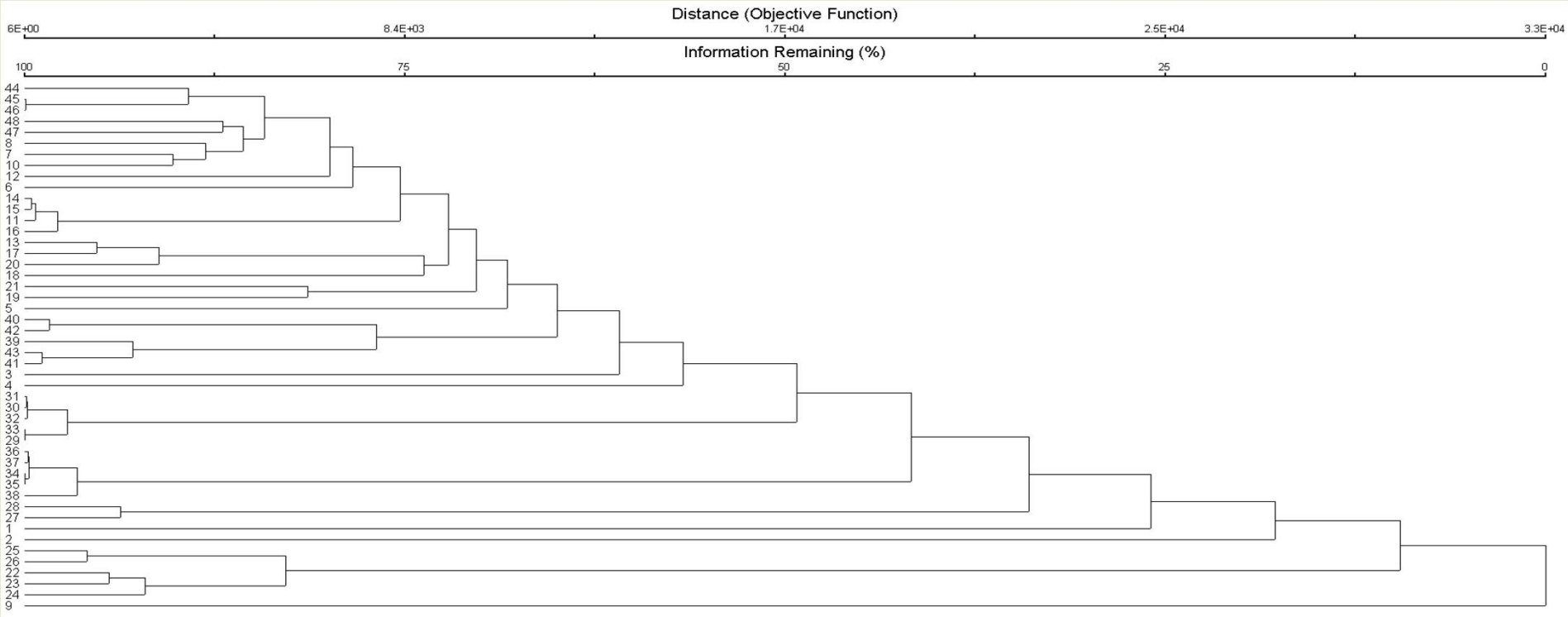


The nature of agglomerative methods
Group Linkage Method

Single linkage clustering or Minimum or Nearest Neighbour

Single linkage is rather prone to chaining (also known as space-contracting), which is the tendency for newly formed clusters to move closer to individual observations, so observations end up joining other clusters rather than another individual observation. This lack of clear separation in the clusters might indicate that there isn't too much variation.

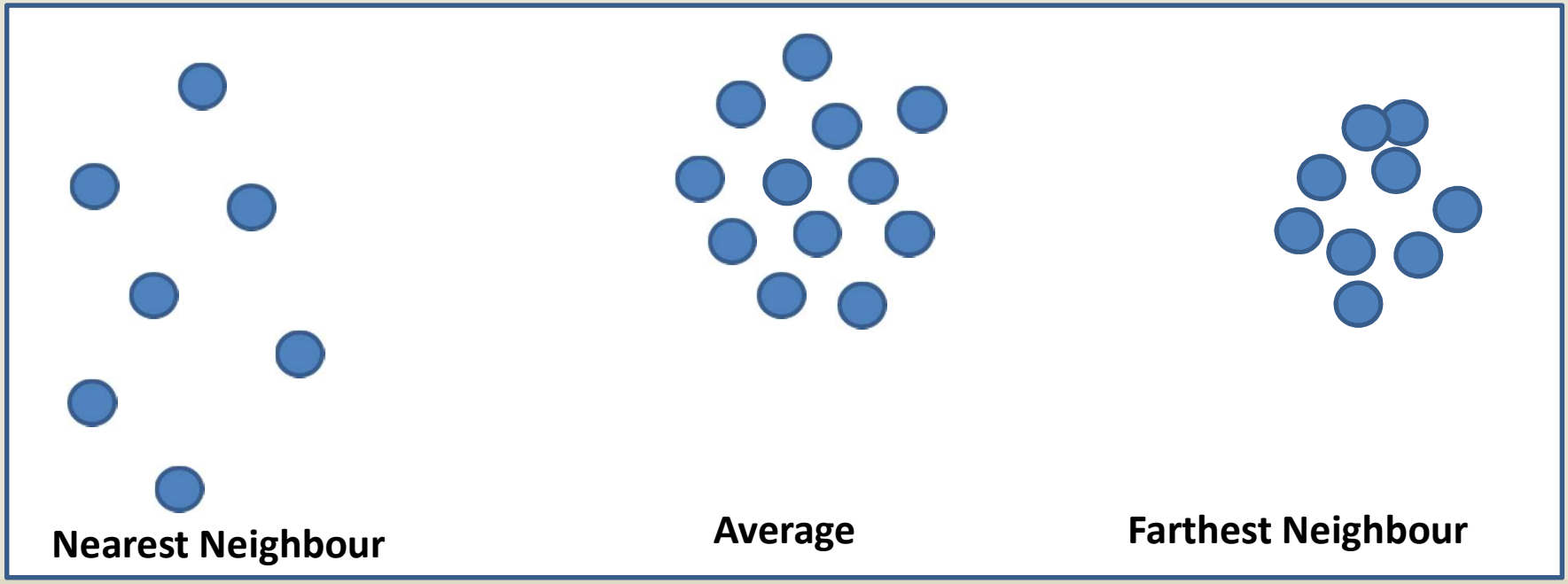
It is not common in Phytosociology, but rather popular in Taxonomy



Training data set of Riparian forest (Euclidean distance, Nearest neighbour, no data transformation)

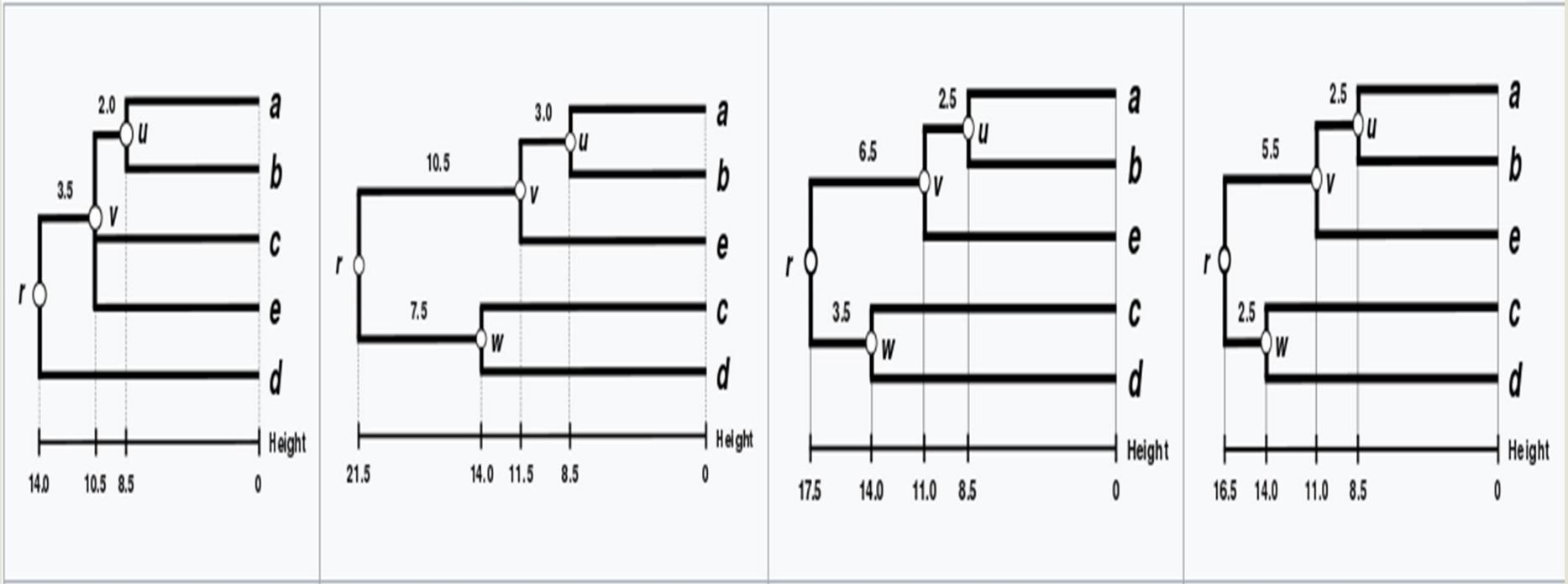
The nature of agglomerative methods
Group Linkage Method

Single linkage clustering /Nearest Neighbour - produce “straggly” clusters; clusters are rather heterogeneous
Complete-linkage clustering /Farthest Neighbour - produce more compacted cluster
Average linkage methods (for example UPGMA) - the individual objects (relevés) are relatively evenly spaced apart each other in the two-dimensional ordination space



The nature of agglomerative methods
Group Linkage Method

Comparison of dendrograms obtained under different clustering methods from the same distance matrix.



Single-linkage clustering. | Complete-linkage clustering. | Average linkage clustering: WPGMA. | Average linkage clustering: UPGMA.

https://en.wikipedia.org/wiki/Complete-linkage_clustering

Agglomerative methods
Where can we do this?

- PC-ORD for Windows – Software
- SYN- TAX 2000 Software
- R-script
- ….
- ….
- etc.

Agglomerative methods
Advantages and disadvantages

Advantages

- As the algorithm is computerized and many software packages exist **– the agglomerative methods are relatively easy to use and are especially suitable for large data sets** (it is also the truth for divisive methods)

- **It is repeatable →** important when you published your results

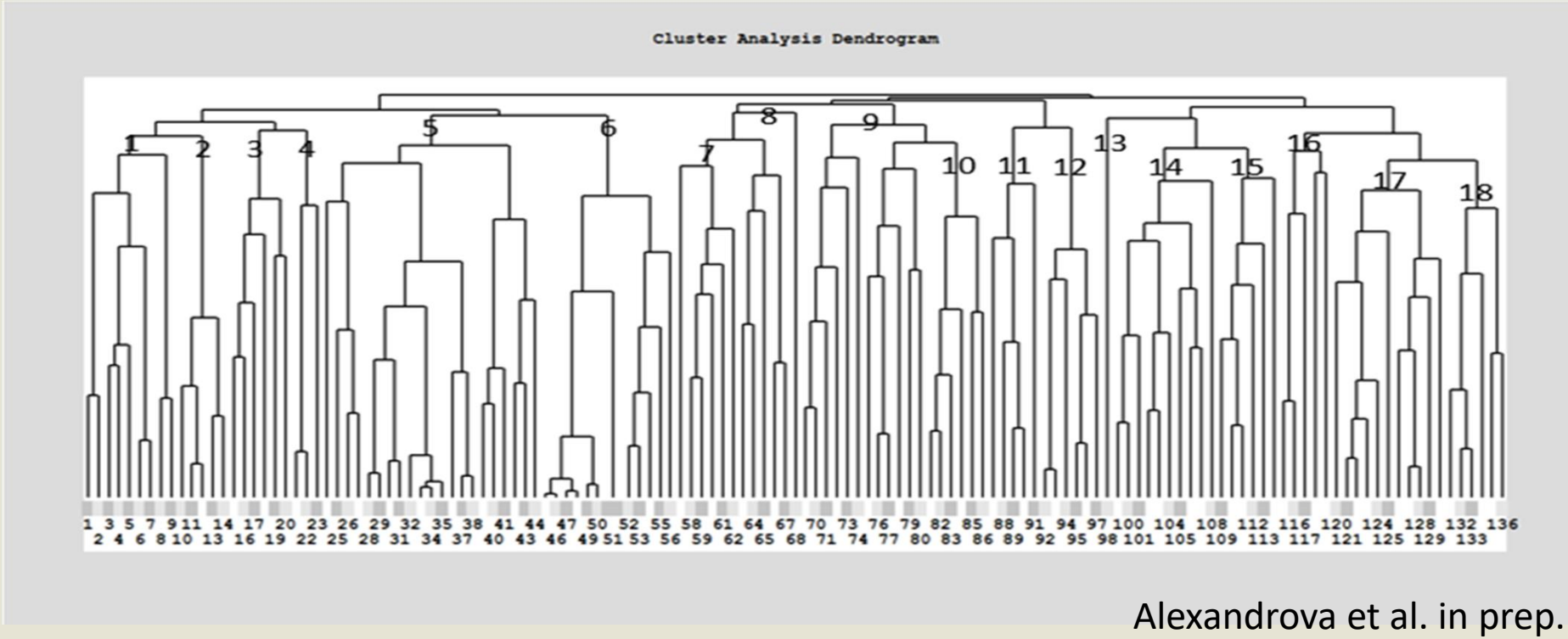
- **Use the difference between the objects to be grouped →** A coefficient of difference can be chosen that is appropriate for the nature of the data type (qualitative or quantitative) and the topic of research (e.g. asked question(s))

- **Different grouping approaches are used →** The one that is the most appropriate for the research topic can be selected

Agglomerative methods
Advantages and disadvantages

Advantages

- **The results are presented as a dendrogram showing progressive (hierarchical) grouping of data →** It is then possible to get an idea of the appropriate number of classes in which the data can be grouped



Alexandrova et al. in prep.

Agglomerative methods
Advantages and disadvantages

Advantages

- **The possibility of an odd number of clusters →** After Modified TWINSPAN (Roleček et al. 2009) it is no longer an advantage

Disadvantages

- Due to the complicated calculations (especially for some linkage methods at each step of linkage procedure) **→ the agglomerative methods are time consuming** (in comparison of other ones e.g. divisive, k-means, etc.)
- **Strong dependence of data set →** if you add only one relevé to the your initial data set, you could obtain different classification results
- **Difficulties to determined the correct number of clusters by the dendrogram →** others procedures to solve this problem exist (for example OPTIMCLASS)
- **The order of the data has an impact on the final results →** the starting relevé impacts the results
- **Rather sensitive procedure to outliers**

Agglomerative methods
Final comments

Does the best algorithm exist? **NO!**

Which distance measure and/or clustering procedure I should choose?
How many clusters/groups I really I have?
How I should interpret the resulting dendrograms?

It is really depend of the initial data and/or study question(s)
And… of your ecological and vegetation knowledge

Kent, M. & Cooker, P. 1992. Vegetation description and analysis: A practical approach. New York, John Wiley and Sons.

“ The best classification is the best ecologically interpreted classification”

Cited by my memory