Multivariate statistics in R

Hannes PETER
Martin BOUTROUX
Zhe LIU

groups

Group 1 (n=5)	Marion Calvo Feryel El Phil Zeineb Fourati Emilie Louise Jeanne Germaine Maillard Alena Sergeyevna Vasilyeva
Group 2 (n=4)	Clara Élodie Christine Cornet Juliette Brigitte D De Wasseige Laura Elisabeth Hebert Noemi Montelaghi
Group 3 (n=4)	Aleksei Dukat Anna Dukat Stefan Eckensperger Zhe Liu
Group 4 (n=5)	Manon Julie Bernard Beatrice Bossi Marie-Moea Sylvie Romina Geffard Lemaître Camille Louise Marie Masanet Leila Claudia Mégevand
Group 5	Madeline Deck
Group 6 (n=4)	Myriam Abdelouhabi Samia Benhalima Janine Keller Ana Victoria Sibaja Zepeda
Group 7 (n=5)	Léo Alexandre Dana Sven Henri Pierre Hominal Miyuka Laurenson Lorenzo Prato Raphaëlle Simonet
Group 8 (n=5)	Baptiste Axel Marie Carmier Etienne De Labarriere Frédéric Laurent Lardet Maurus Sebastian Lozza Marc-André Mauron

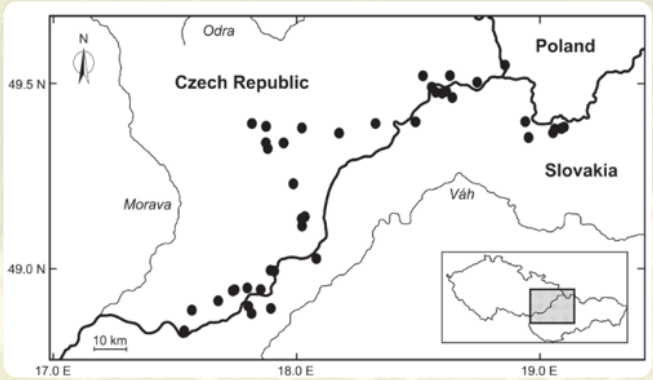
n=32

group 1

group 2

Evaluating Community Composition: A Case Study of Molluscs and their Plant Partnerships in European Wetlands

43 sites of interest
- Located on the border between the **Czech Republic and Slovakia**
- **Wetland** ecosystems



Species of Interest – from soil samples
- **57** molluscs
- **203** vegetation species
  - **43** Bryophytes
  - **160** vascular plants

Environmental Parameters –from water samples
- Ca, Mg, Fe, K, Na, SO4, PO4, N.NO3, N.NH3, Cl, pH, conductivity and redox potential

Research Question:

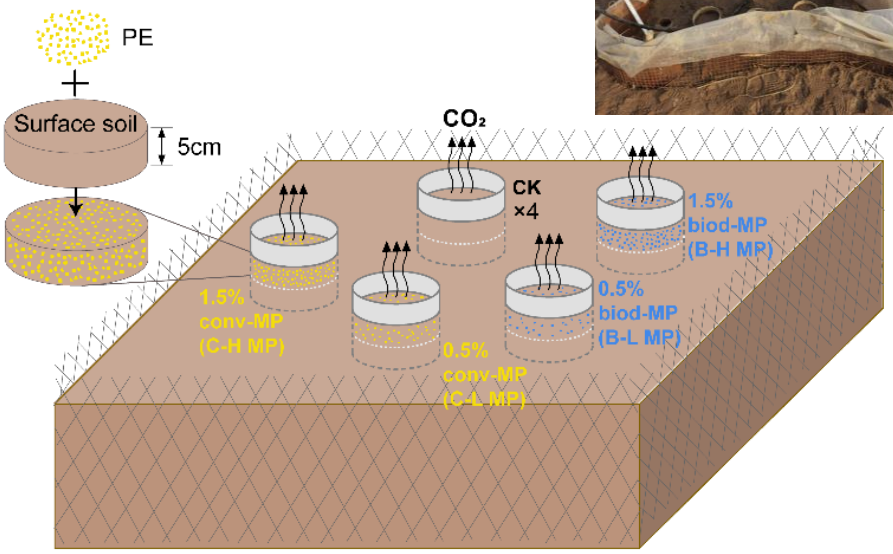
How do the environmental conditions of various wet lands impact the community composition of molluscs, bryophytes and vascular plants?

i.e. based on the clustering of wetlands by environmental condition, how do the mollusc andvegetation communities differ?

group 3

Do Microplastics affect the Soil Microbiome?

- Bio-degradable vs. conventional plastic
- One year duration
  - Measured every season
- Community sequenced
  - 16S and 18S ASVs
- Environmental parameters
  - pH
  - Soil moisture and temperature
  - Soil biomass and soil total C/N/P
  - DOC and DON
  - Gas fluxes



Does the microbial community change after plastic addition?

Is there less of an effect from bio-degradable plastics?

group 4

Development of the microbial community in aged versus new drinking water pipes receiving different organic carbon?

group 5

Multivariate data: 16S rRNA sequences

- Counts of amplicon sequence variants (ASVs)
- Counts of classified taxonomic levels from domain to family

Total number of samples: 144

- 24 pipes = 2 ages of pipe x 4 experimental conditions x 3 biological replicates
- 2 types of sample (biofilm/bulk water) were collected at each pipe over 3 time points
- 24 pipes x 2 types of sample x 3 time points = 144 total samples

Additional environmental data:

- Collected from only bulk water samples (72): pH, dissolved oxygen, temperature, total organic carbon, total cell counts (flow cytometry), intact cell counts (flow cytometry)
- Collected from all samples (144): 16S rRNA gene copies (qpcr)

group 6

Group 6
Myriam Abdelouhabi, Samia Benhalima, Janine Keller, Ana Victoria Sibaja Zepeda

Research Question

What environmental parameters influence the presence of certain algal pigments and toxins, and can we identify a link between them?

Context

The dataset is from the European Multi Lake Survey, conducted in the summer of 2015, which assessed 369 lakes in 27 countries in Europe. The goal of the study was to find out more about the influence of anthropogenic activity on the eutrophication process.

Data Description

Data dimension: 369 x 34

Considered parameters:

- 18 Environmental variables: Location, depth, temperature and nutrients (P and N)
- 16 Analyzed variables :
  - 9 Algal pigments
  - 7 Toxins

Mantzouki, E., Campbell, J., van Loon, E. *et al.* A European Multi Lake Survey dataset of environmental variables, phytoplankton pigments and cyanotoxins. *Sci Data* **5**, 180226 (2018). https://doi.org/10.1038/sdata.2018.226

Sven Hominal
Miyuka Laurenson
Lorenzo Prato
Raphaelle Simonnet
Léo Dana

EPFL

Alpine plants communities in Aravo (F group 7

ENV-513 GROUP 7

To what extent do environmental variables and marmot-induced zoogenic disturbance explain variation in alpine plant community composition ?

Dates, Existing Studies, Authors

Data covers the period 1997-1999
Previous studies on the dataset : Dray et al. (2014), Choler (2005)
Community composition of vascular plants was determined in 75 sites, 5 × 5 m plots.
82 measured species (most abundant) out of 132 recorded species

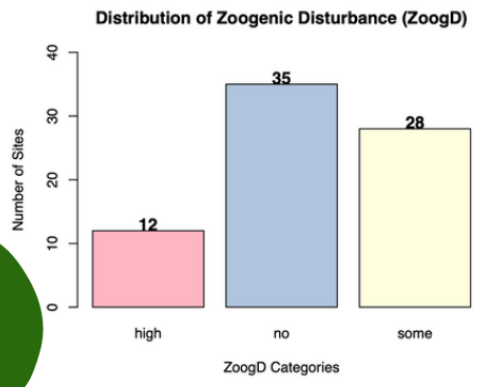
Hypotheses

Environmental variables (snowmelt timing, topographic factors, and physical disturbance) are the primary drivers of alpine plant community composition, explaining significantly more variation than marmot-induced zoogenic disturbance.

The effect of marmot-induced zoogenic disturbance on plant community composition is context-dependent and varies significantly across environmental gradients, with environmental variables and zoogenic disturbance together explaining more variation than either factor alone through complementary mechanisms.

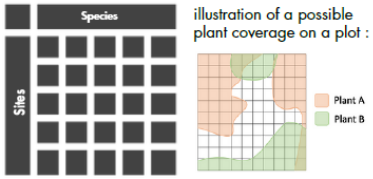
First explorations

- **Community patterns**
- **Environmental distributions**
- **Distribution of Zoogenic Disturbance and correlation with environmental variables and species attributes**
- **Preliminary relationships (linear or non-linear)**



Dataset - Aravo

Species abundancy matrix



6 level scale estimating the cover of each plant species for each 5*5 plot (may overlap on a site)

- 0 → 0 - 5%
- 1 → 5 - 10%
- 2 → 10 - 25%
- 3 → 25 - 50%
- 4 → 50 - 75%
- 5 → > 75%

6 Environmental Variables

- Relative South aspects
- Slope Inclination (°)
- Microtopographic landform index
- Mean Snowmelt date
- Physical disturbance due to solifluction and cryoturbation
- Zoogenic Disturbance due to marmots activity

8 Species Attributes

- Vegetative Height [cm]
- Max. Lateral Spread [cm]
- Leaf elevation angle
- Leaf area
- Maximal leaf's cross-section thickness
- Specific leaf area
- Mass-based leaf nitrogen content
- Seed mass

Preprocessing

Data loading & Quality Check
Numerical variables and categorical variables

Transformation of variables :

- Log, Square root, Hellinger transformations
- Abundance of species by site

Diversity indices :

- Shannon diversity, Simpson diversity

Distance matrices :

- Bray-Curtis (for abundance data)
- Jaccard/Sorensen (for presence/absence)

Next steps

- **Unconstrained Ordination with Principal Component Analysis (PCA)** of environmental variables and zoogenic disturbance due to marmot activities - identify main gradients.
- **Constrained ordination** relating species composition to environmental variables by constraining the ordination axes to be linear combinations of predictor variables.
- **Clustering** of species to identify a trend and similarities
- **Variance Partitioning**: Pure environmental effects vs pure zoogenic effects, test of statistical significance
- **Comparison of methods and results with literature, visualisation and Interpretation of the results**

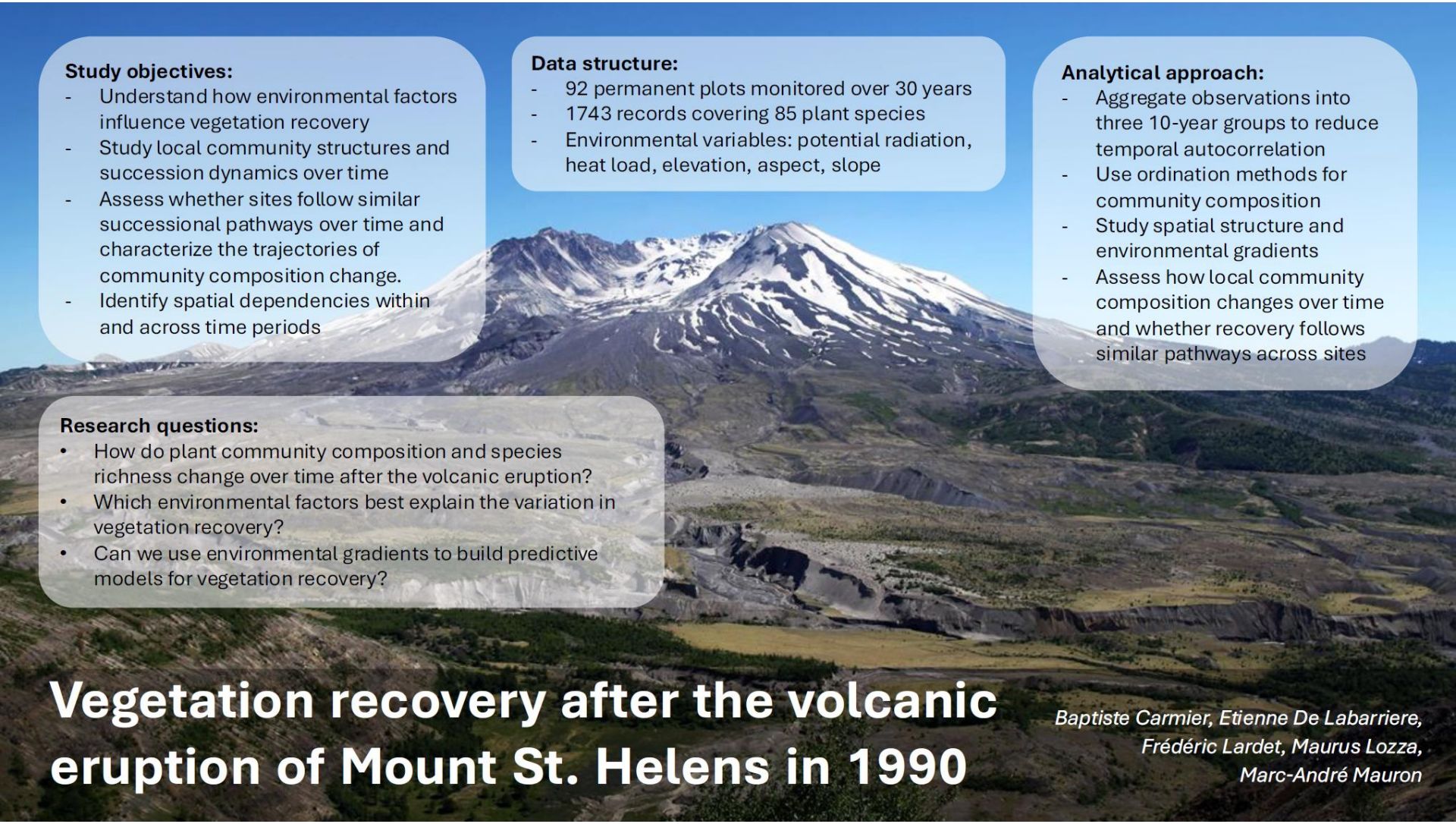
Sources

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.alpvise.co.uk%2Fgalp7g.html&psig=AOvVaw0lt2UPjB0WhJgxaPnM2D&ust=1759391795916000&source=images&cd=vfe&opi=89978449&ved=0CBUQjRxqFwoTCMDI3arDjpADFQAAAAAdAAAAABAJ

5. Dray et al., "Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation," Ecology, vol. 95, no. 1, pp. 14-21, Jan. 2014, doi: https://doi.org/10.1890/13-0196.1.

Philippe Choler, "Consistent Shifts in Alpine Plant Traits along a Mesotopographical Gradient," vol. 37, no. 4, pp. 444-453, Nov. 2005, doi: https://doi.org/10.1657/1523-0430(2005)037[0444:csiapt]2.0.co;2.

group 8



Study objectives:
- Understand how environmental factors influence vegetation recovery
- Study local community structures and succession dynamics over time
- Assess whether sites follow similar successional pathways over time and characterize the trajectories of community composition change.
- Identify spatial dependencies within and across time periods

Data structure:
- 92 permanent plots monitored over 30 years
- 1743 records covering 85 plant species
- Environmental variables: potential radiation, heat load, elevation, aspect, slope

Analytical approach:
- Aggregate observations into three 10-year groups to reduce temporal autocorrelation
- Use ordination methods for community composition
- Study spatial structure and environmental gradients
- Assess how local community composition changes over time and whether recovery follows similar pathways across sites

Research questions:
- How do plant community composition and species richness change over time after the volcanic eruption?
- Which environmental factors best explain the variation in vegetation recovery?
- Can we use environmental gradients to build predictive models for vegetation recovery?

Vegetation recovery after the volcanic eruption of Mount St. Helens in 1990

Baptiste Carmier, Etienne De Labarriere, Frédéric Lardet, Maurus Lozza, Marc-André Mauron

Recap

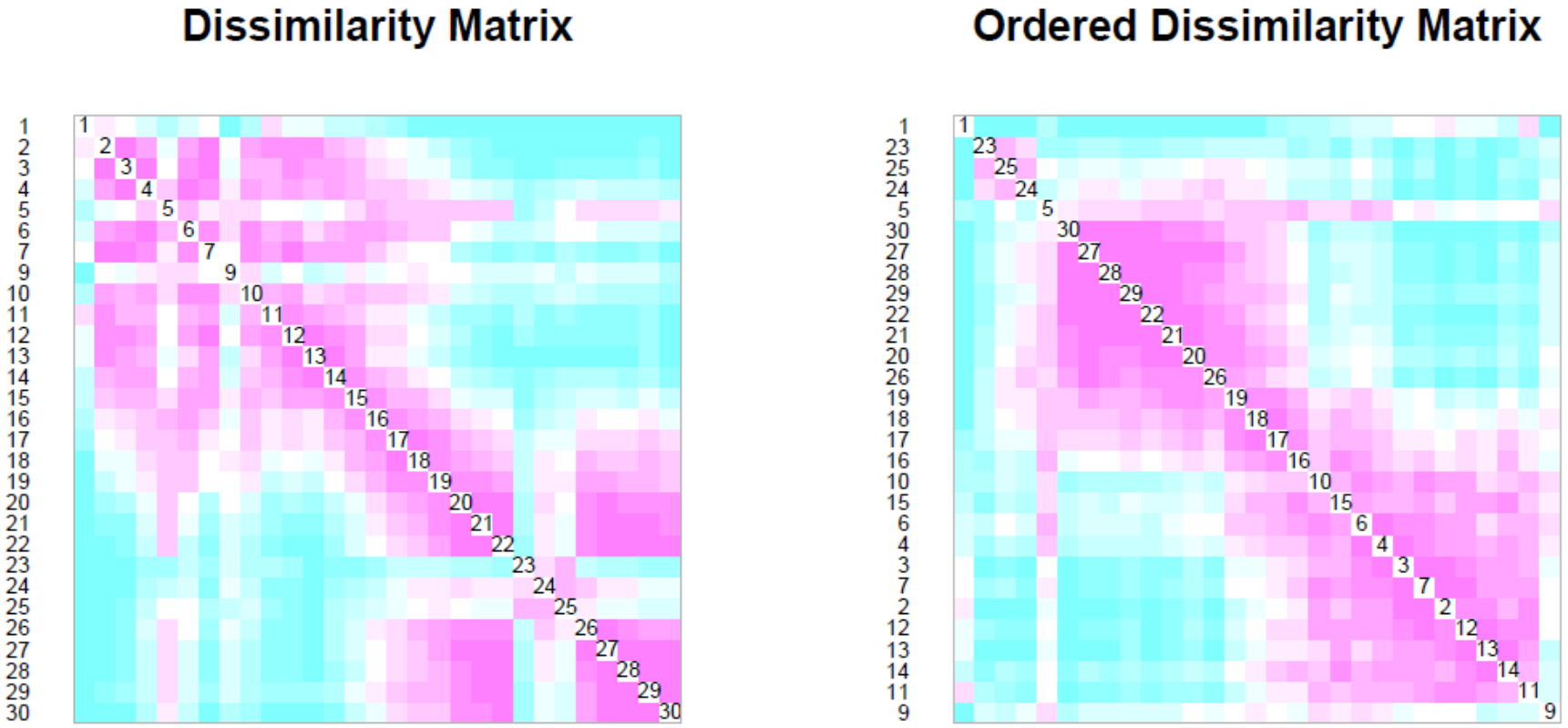
First session

- data exploration
- summary statistics
- visualization

Second session

- transformations
- resemblance
  - dis/similarity, distance

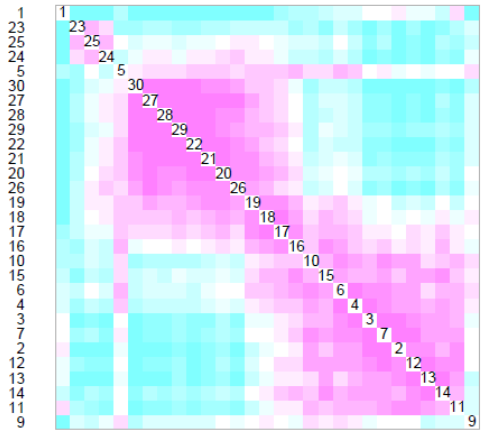
reminder: Bray-Curtis dissimilarity



Classification

Aim: to find discontinuities (breaks/gaps) in data and to group similar objects in order to…

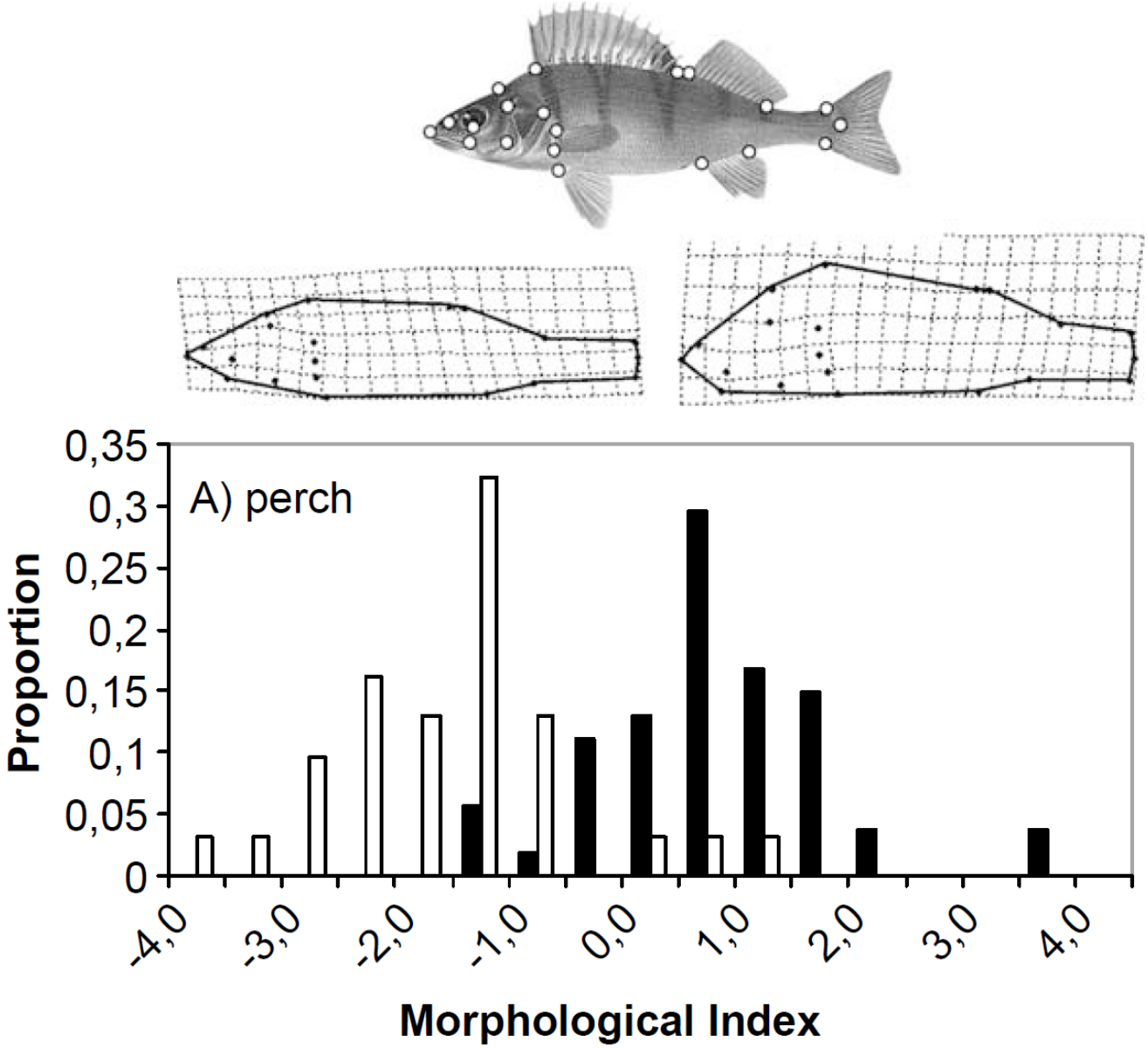
- name them (e.g. to ease communication)
- explore patterns and structure of dataset
- identify groups, types (typology)



Groups/clusters should be internally homogeneous and clearly distinguishable from the other groups.

Multivariate groups are often fuzzy (multiple gradients, continuous variation), and therefore these methods might not be the best ones (alternative: ordinations)

example: morphometerics



Classification

Unsupervised

search for main gradients and homogeneous groups in the data.

- ➢ No a priori knowledge/assumptions
- ➢ Results depend mainly structure of the dataset.

- distance/similarity metric, choice of clustering method
- assignment of samples into groups may change even with slight changes of the dataset (e.g. by adding more samples)
- examples of unsupervised methods are **cluster analysis, TWINSPAN**

Supervised

use external criteria to classify the dataset

- ➢ you supply information/rules about how to classify

- assignment of samples to groups remain the same despite changes in the structure of the dataset

- examples are **classification and regression trees** (CART), **random forest classifier**, artificial neural networks (ANN), etc.

(**k-means clustering**, can either be supervised or unsupervised)

General overview of unsupervised clustering

Selection of a resemblance criteria

- (Dis)similarity or distance between objects

Partition (non-hierarchical) clustering

- split objects into groups (e.g. TWINSPAN - Two Way INdicator SPecies ANalysis)
- number of groups can be set initially (k-means)

Hierarchical clustering

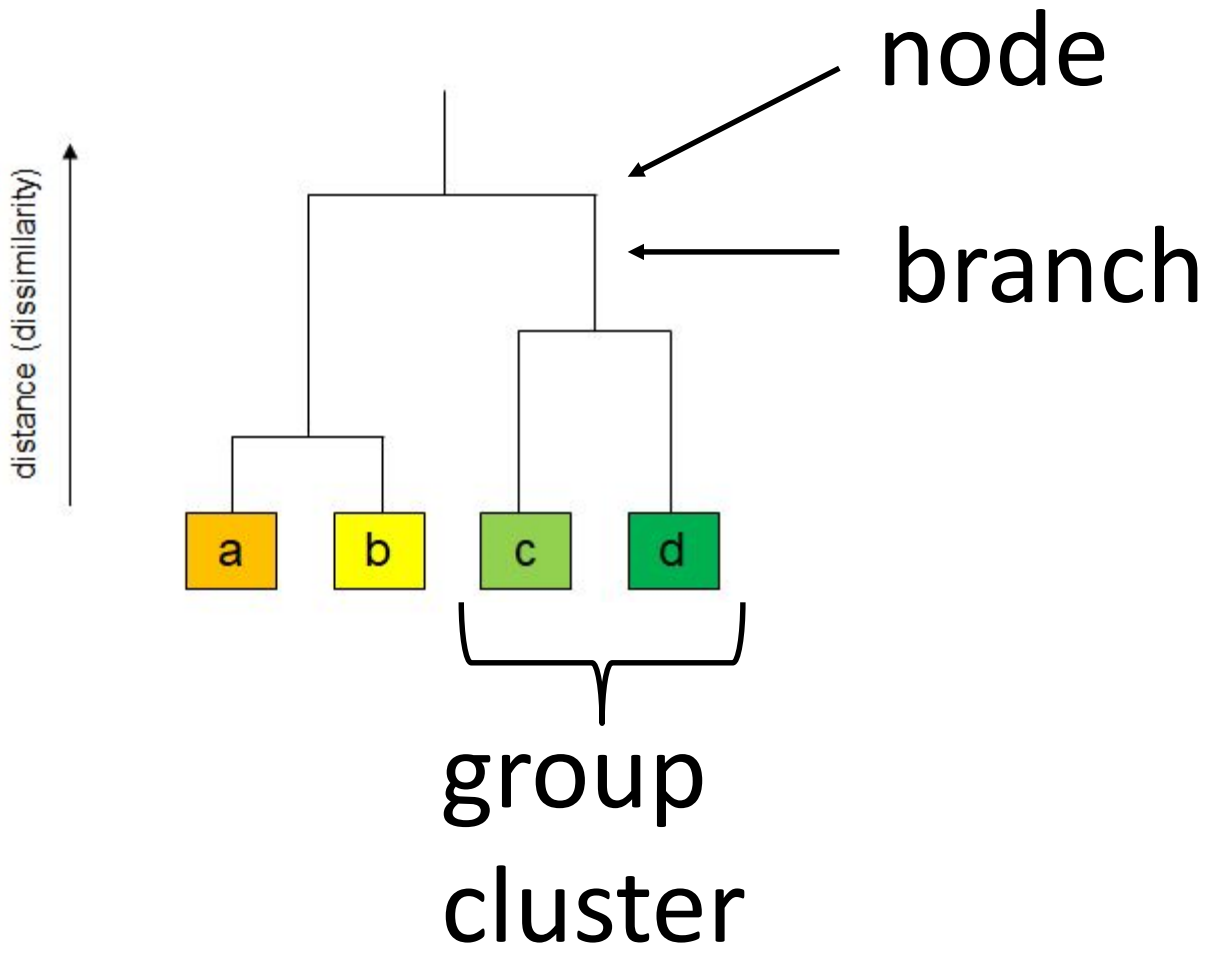
- maintain hierarchy of similarity within group (groups can cluster inside other groups)
- e.g. cluster analysis

=> dendrograms

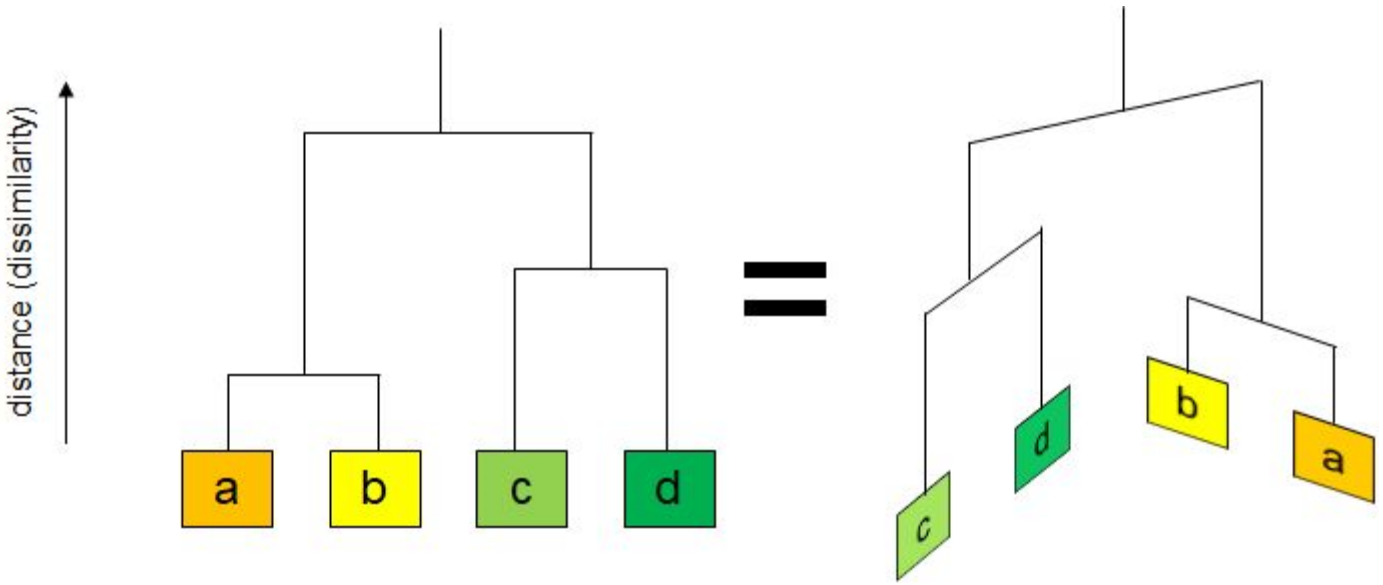
Selection of a grouping criteria

- Are two objects (or descriptors) sufficiently similar to be assigned to the same group ?
- Most methods consider mutually exclusive groups (*binary membership*).

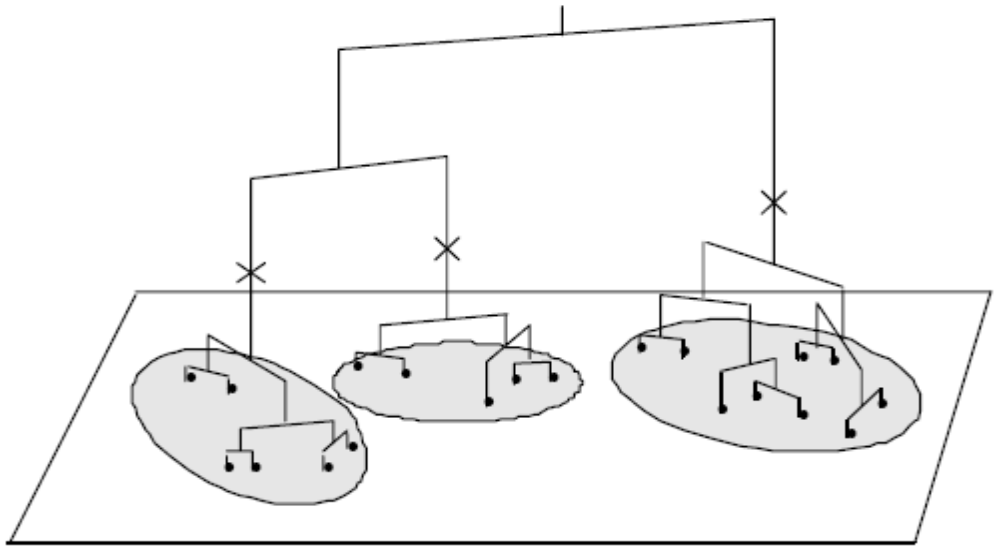
dendrograms



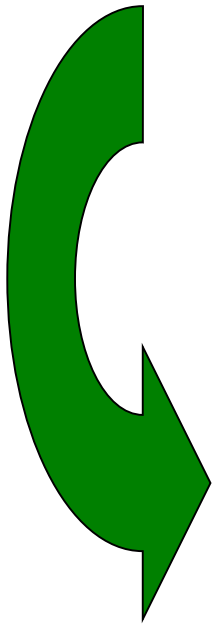
dendrograms

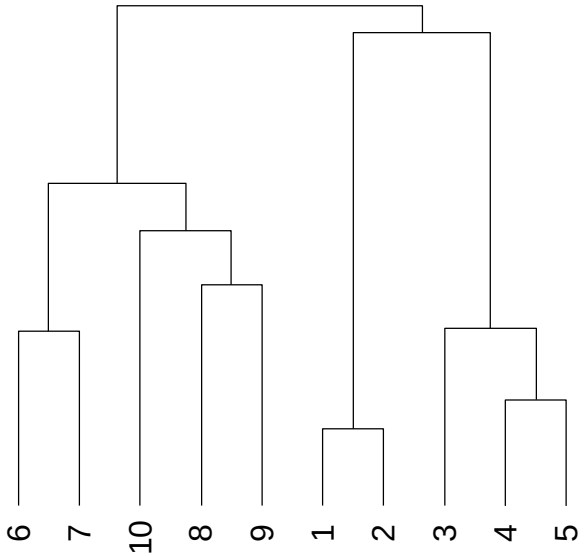


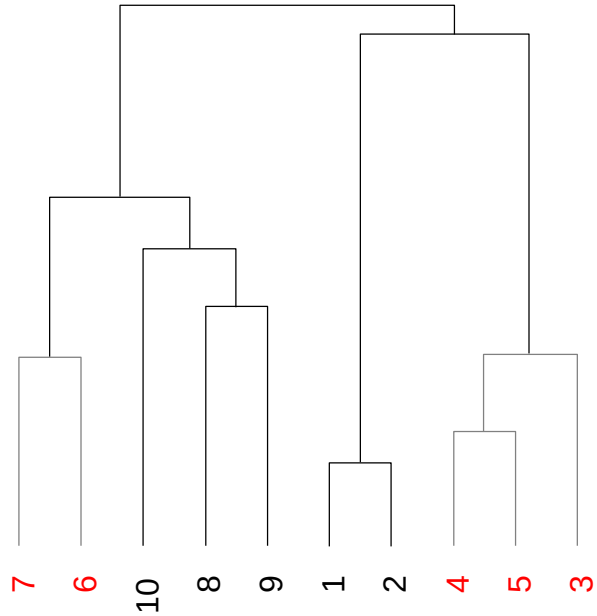
dendrograms



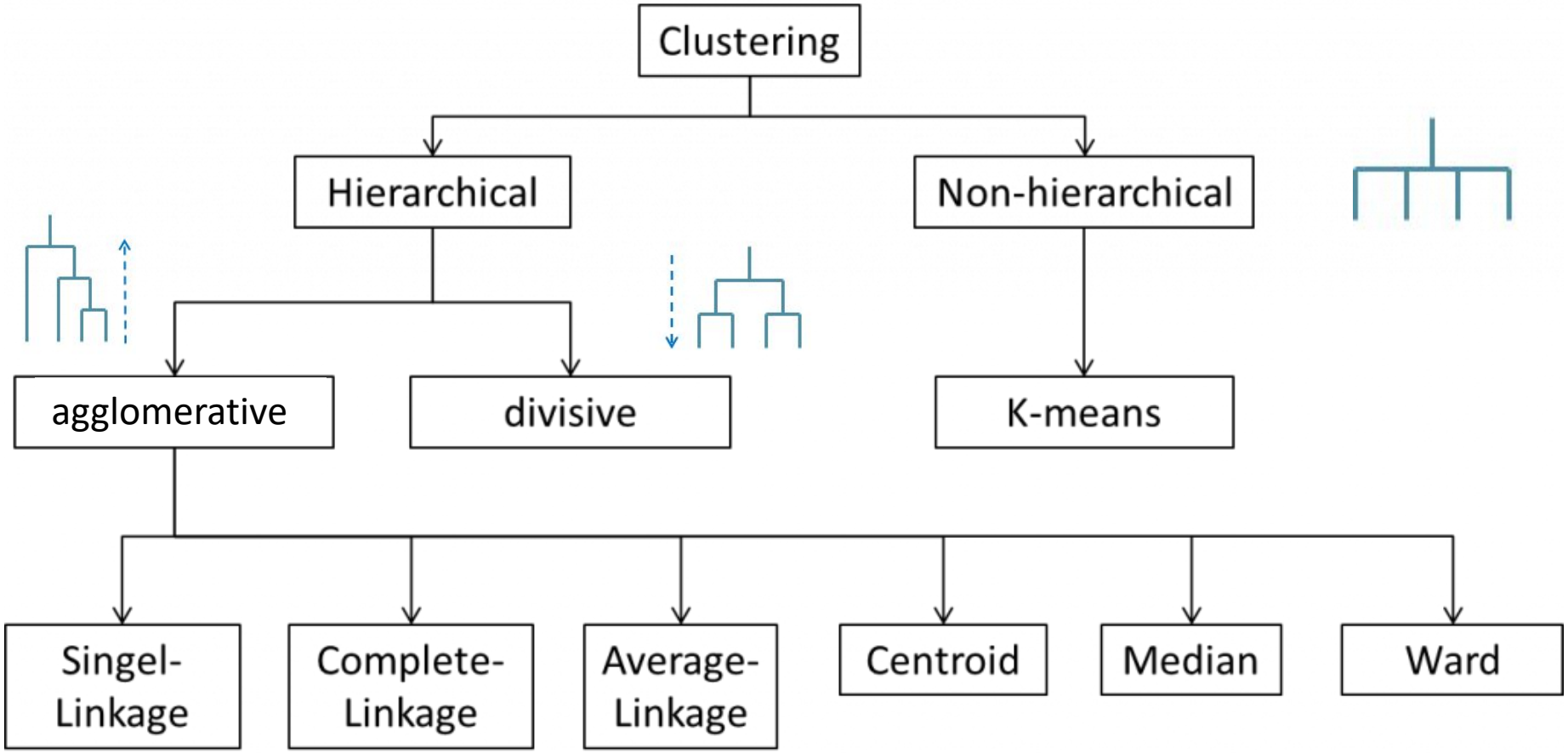
The order of tips on dendrograms can not be used for the interpretation of resemblance!





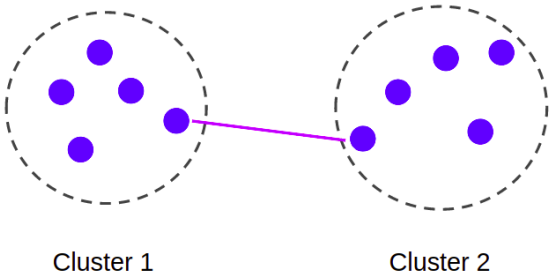


classification of classification methods



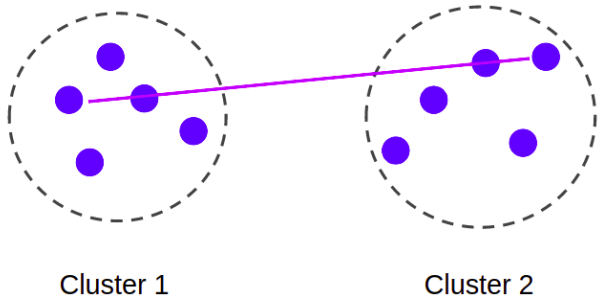

single-linkage
he table with my
best friend
average-linkage
the table with the
highest average
“friendliness”
complete-linkage
the table at with the
person I don’t like the
most is still not that bad

Simple linkage



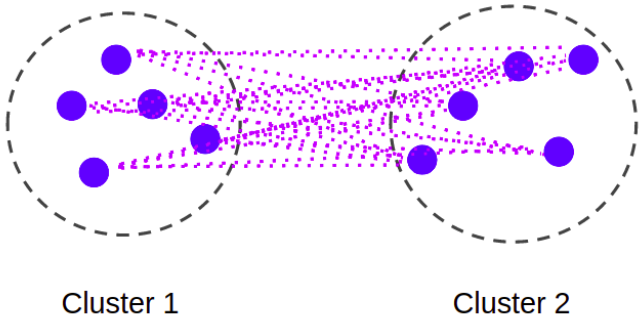
Distance between clusters is defined by the distance between their closest members.

Complete linkage



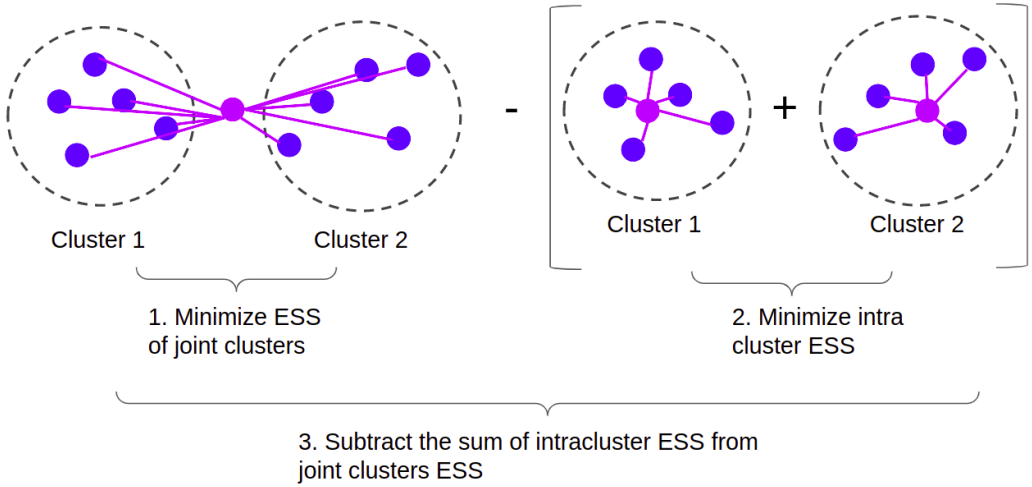
Distance between clusters is defined by the distance between their furthest members.

Average linkage



The percentage of the number of points of each cluster is calculated with respect to the number of points of the two clusters if they were merged.

Ward linkage



Specifies the distance between two clusters, computes the sum of squares error (ESS), and successively chooses the next clusters based on the smaller ESS.

EPFL

Single linkage algorithm

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	—

similarity matrix

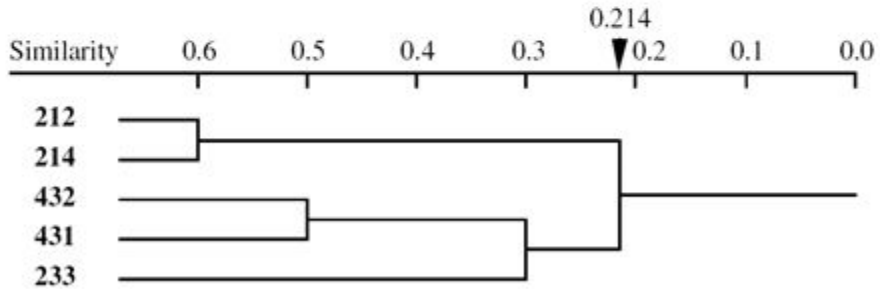
Single linkage algorithm

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	—

similarity matrix

pairs of samples sorted according to similarity

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

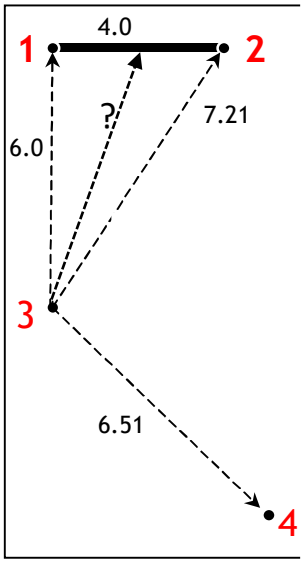


resulting dendrogram

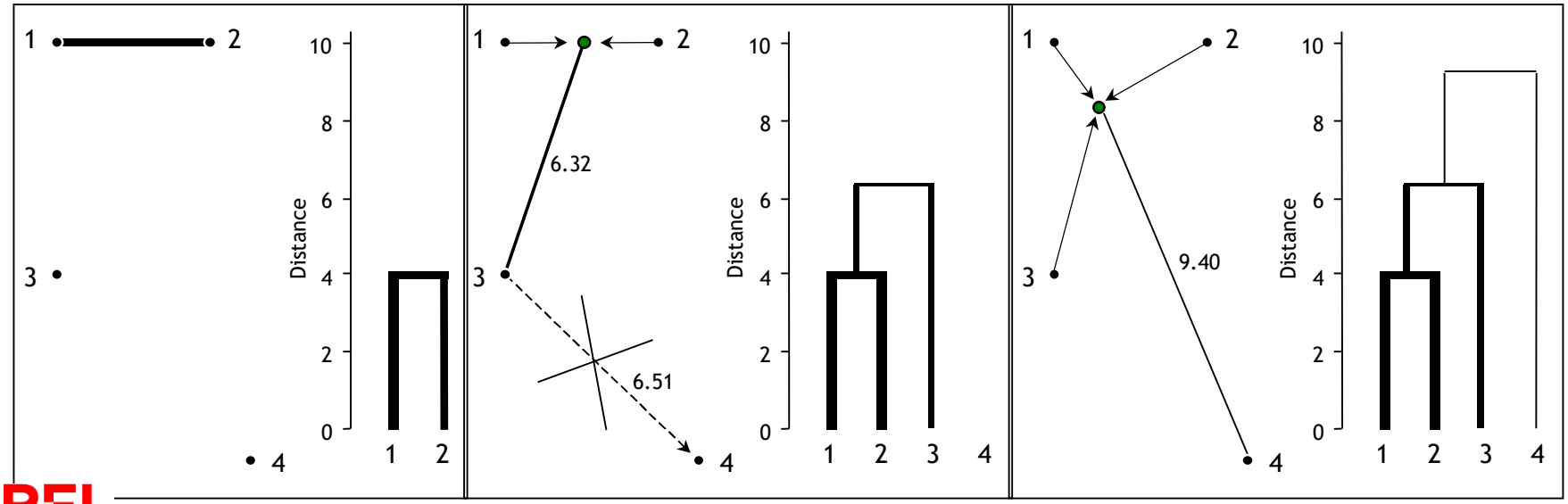
pair group methods (PGM)

	Arithmetic averages of distances or dissimilarities	Centroids of groups
Without weighing	*UPGMA (average)* Grouping by mean association	*UPGMC (centroid)* Grouping by centroids
With weighing	*WPGMA* Grouping by proportional weights	*WPGMC* Grouping by median

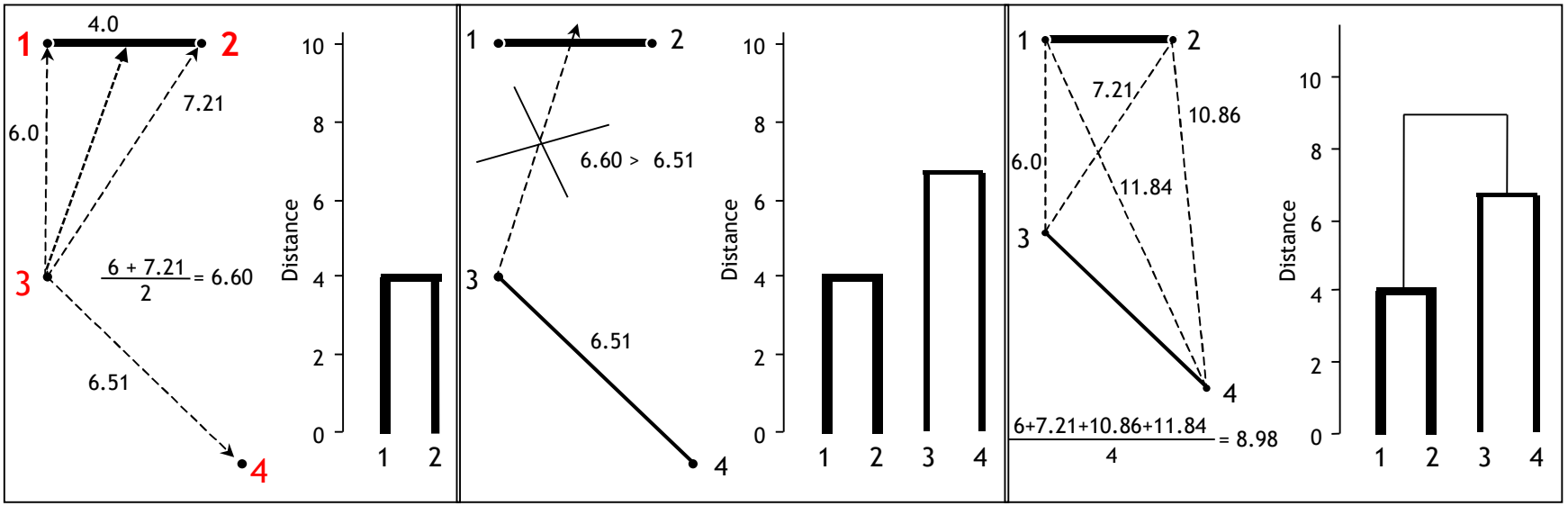
UPGMA (unweighted pair group method with arithmetic mean)



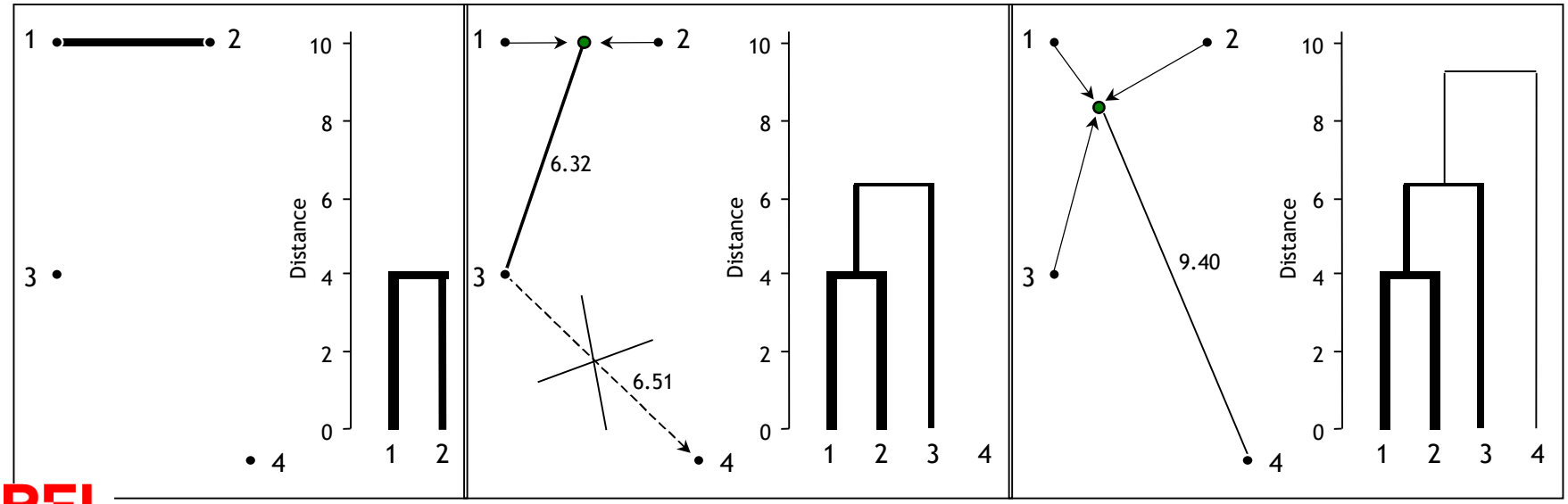
UPGMC (unweighted pair group method with centroids)



UPGMA (unweighted pair group method with arithmetic mean)



UPGMC (unweighted pair group method with centroids)



EPFL

Ward agglomerative clustering

- Minimizes the variance within groups
- Robust method
- Tends to produce dendrograms with compact groups of equal size

TWINSPAN (Two Way INdicator SPecies ANalysis)

TWINSPAN is a **divisive** method:

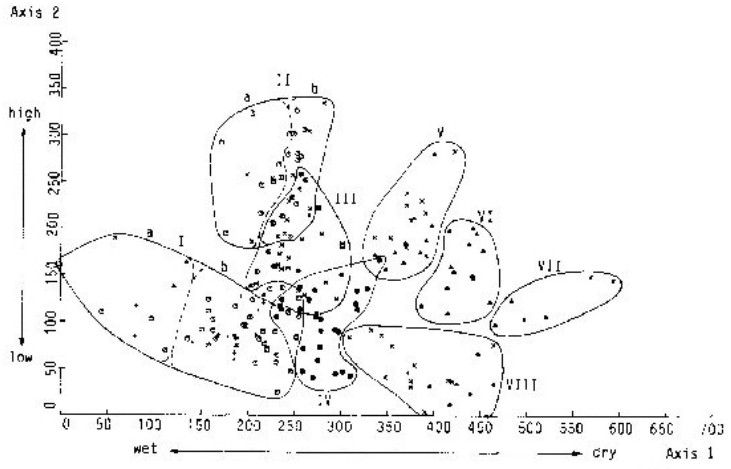
1. Samples are ordinated

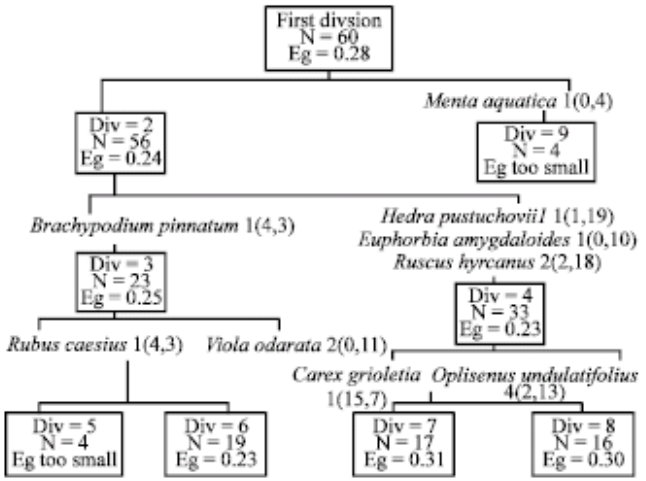
2. A crude dichotomy is formed: the ordination centroid is used as a dividing line between two groups (negative and positive)

3. The dichotomy is refined by a process comparable to iterative character weighting

4. Dichotomies are ordered so that similar clusters are near each other

5. Stopping criteria
   - Number of samples per group
   - Number of divisions





implemented in R: *twinspanR*

Comparison of methods

Criteria for «good» classification (ease of interpretation):

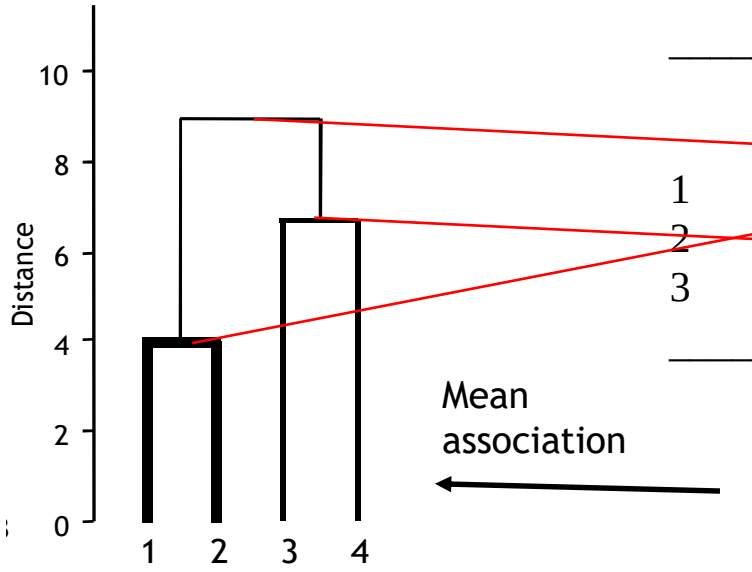
- Compact Groups
  - Minimal intra-group variance
  - Elements grouped at low distance level
- Groups of comparable sizes
  - Roughly the same number of elements in each group
  - No or very few groups with only one element
- Distinctly separated groups
  - Maximal inter-group variance

Often difficult to satisfy these criteria simultaneously

statistics

cophenetic matrix (distances)

Symmetric matrix of the distance in the dendrogram



	2	3	4
1	4.0	8.98	8.98
2		8.98	8.98
3			6.51

Mean association

Original matrix of distances or dissimilarities

Objects	2	3	4
1	4.00	6.00	11.84
2		7.21	10.86
3			6.51

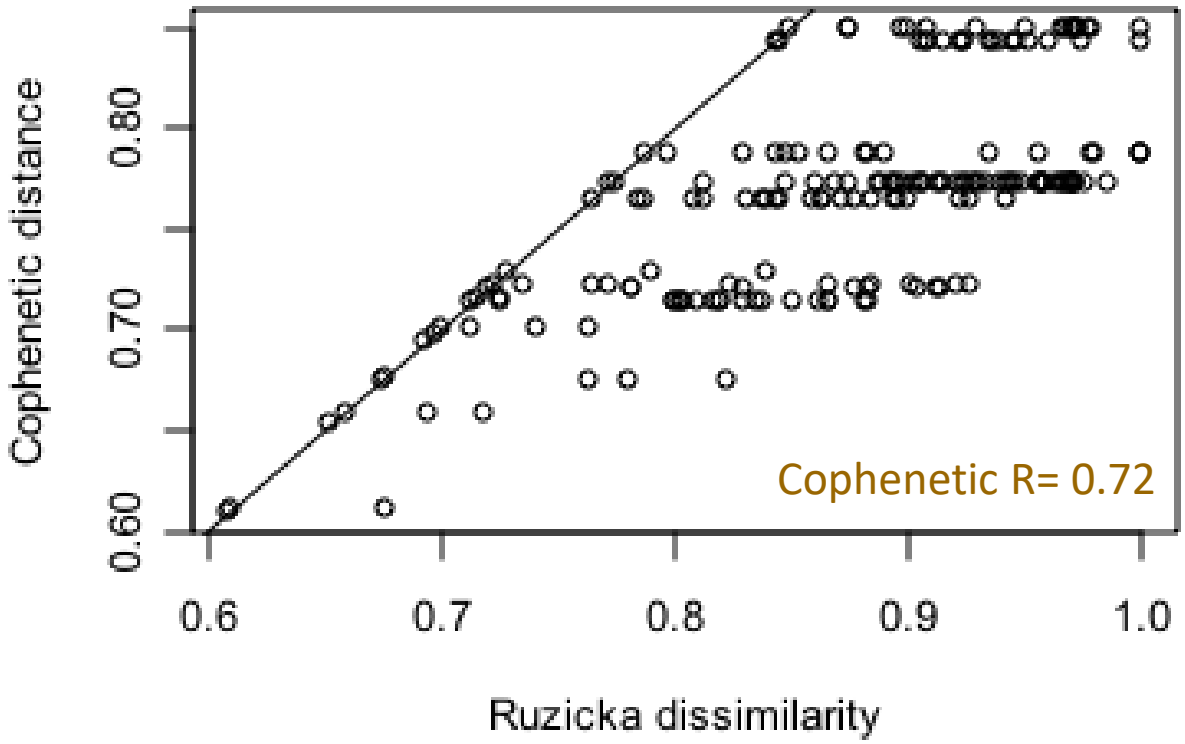
cophenetic correlation

- Correlation between the cophenetic distances and the original dissimilarities

example R= 0.79, indicating that 79% of the variance of the original association matrix is reproduced in the dendrogram

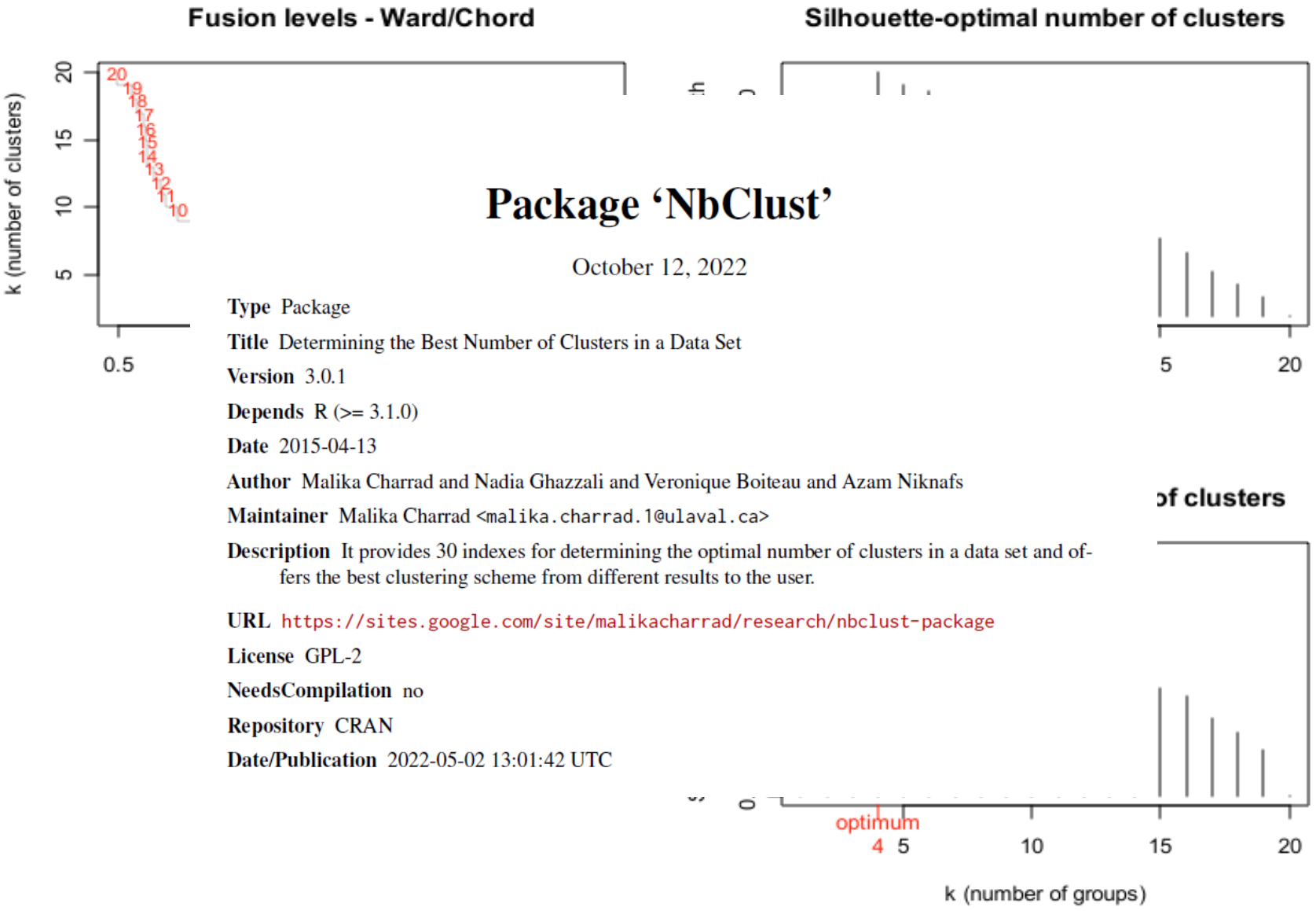
Shepard Diagrams

Comparison of the cophenetic distance matrix and the original dissimilarity matrix of each object.



narrow scatter around a 1:1 line indicates a good representation while large scatter or a nonlinear pattern indicates a lack of representativity.

Choice of the optimal number of groups



Package 'NbClust'

October 12, 2022

Type Package

Title Determining the Best Number of Clusters in a Data Set

Version 3.0.1

Depends R (>= 3.1.0)

Date 2015-04-13

Author Malika Charrad and Nadia Ghazzali and Veronique Boiteau and Azam Niknafs

Maintainer Malika Charrad <malika.charrad.1@ulaval.ca>

Description It provides 30 indexes for determining the optimal number of clusters in a data set and offers the best clustering scheme from different results to the user.

URL https://sites.google.com/site/malikacharrad/research/nbclust-package

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2022-05-02 13:01:42 UTC