Multivariate statistics in R

Hannes PETER
Martin Boutroux

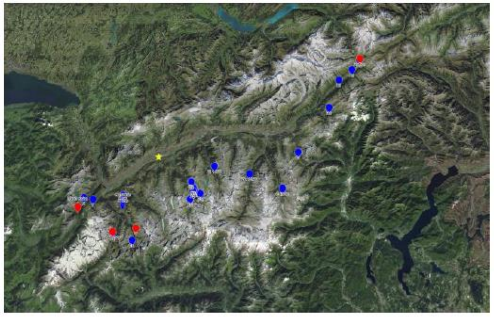
groups

Group 1	Laureen Ahlers, Yaniss Augot, Léa Francomme, Haohua He, Taiqi Lian
Group 2	Aurèle Baretje, Ambre De Herde, Giulia Dohy, Juan Benedetti, Yann Roubaud
Group 3	William Browne, Emma Faval, Gloria Leuenberger, Luca Raviglione, Elie Roth
Group 4	Maëlle Régnier, Katia Todorov, Charlotte Wang
Group 5	Maxwell Bergström, Cécile Bettex, Julie Botzas-Coluni, Agathe Crosnier
Group 6	Chloe Bouchiat, Delaram Mirmohammadi, Désirée Popelka, Luca Soravia

n=26

group 1

Environmental impact of microbial communities in glacial streams



- **Background**:
  - *Hydrurus foetidus* (HF) is a freshwater algae commonly found in glacier-fed stream that has an unstudied associated microbiome
  - Winter samples from 33 sites in Valais, Switzerland
- **Research questions**:
  - Is there a difference between microbial community in water column and algae-associated?
  - Do environmental parameters (topology, geography, water chemistry) affect these microbial communities?
  - Subset analysis: Does distance to glaciers (up-/downstream) affect microbial community composition?

group 2

Multivariate statistics in R - Group 2

Distributions, life-history specialization, and phylogeny of the rain forest vertebrates in the Austalian Wet Tropics

Dataset: https://esapubs.org/archive/ecol/E091/181/default.htm
What do we have?

- Species
- total abundances (circa 200)
- 40 different sites in the studies region constructed by similarities (topography, land use…)
- rainforest specialization
- many species living characteristics (elevation range, humidity range, etc…)

What we will extract:

- site characterization : near the ocean or not ; main land use (forest, agricultural, urban)

Dataset meteo:
https://www.climatechangeinaustralia.gov.au/en/obtain-data/download-datasets/
What do we have?

- temperature (mean, max and min)
- precipitations
- relative humidity

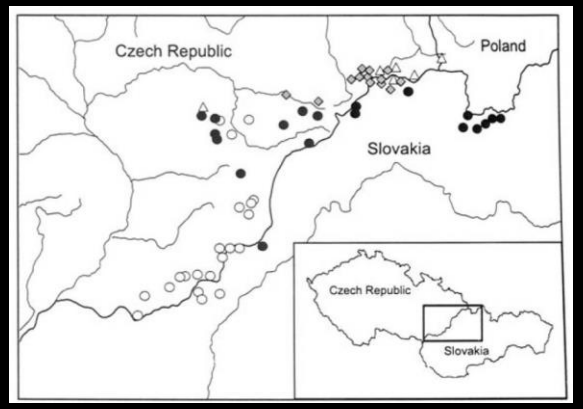
Hypotheses: Climate and topographic parameters mainly explain the spatial distribution of the studied species

Studied zone (in Australia)

group 3

How do the chemical parameters of fen water and slope influence the species diversity of vascular plants and bryophytes in the Western Carpathians?

The data was collected in 2002 from the fens of the Western Carpathian mountain range along the Czechia-Slovakia border. It includes information on bryophyte and vascular plant abundance and richness across 70 locations, as well as 14 chemical parameters of the water and slope measurements, but not the location



group 4

Root fungal communities in organic and conventional agriculture

group 5

Contrasting patterns of functional diversity in coffee root fungal communities associated with organic and conventionally-managed fields

Research question :

What is the effect of conventional vs. organic management on coffee root fungal diversity and community composition?

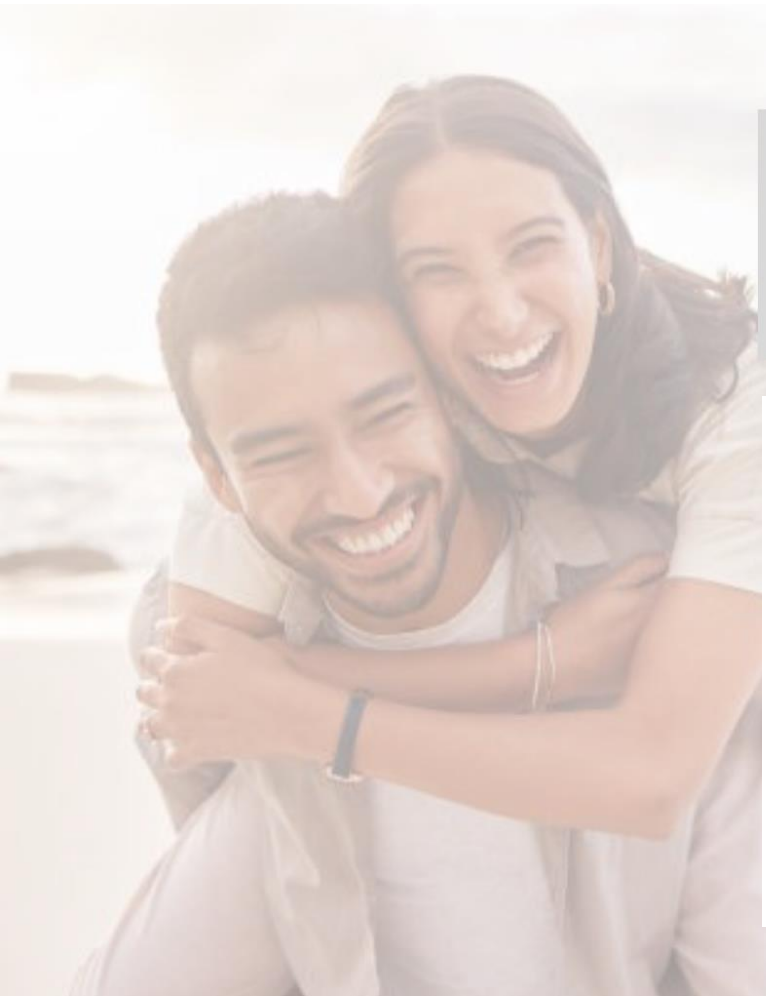
Dataset description : 25 fields

Field characteristics:

- Conventional or organic
- Coordinates
- Density cover, leaf litter depth, elevation, slope
- Coffee density, coffee variety, age of coffee plants, age of coffee field, prior use
- Types of shade tree, types of windbreak tree
- Types of fungicides, herbicides and fertilizers
- Spore richness
- Mean AMF root colonization
- Soil characteristics (pH, heavy metals, N-content)

Clustering and Characterizing Marital Status in Switzerland: A Multivariate Analysis of Unmarried and Married Individuals Over Time

group 6



Research Question

How did the archetypes of married and unmarried individuals in Switzerland change over time?

Hypotheses

H1: Individuals in Switzerland can be clustered into distinct demographic and socioeconomic groups based on factors such as education level, habitat, income, and employment status.

H2: Age and sex will significantly influence the composition of clusters, with younger individuals and women more likely to cluster around certain lifestyle or educational characteristics compared to older individuals and men.

H3: Urban vs. rural habitat (or other geographic distinctions) will play a significant role in forming distinct clusters, with urban residents exhibiting different socioeconomic characteristics compared to rural residents.

H4: Over time, the characteristics of unmarried and married individuals have evolved, resulting in changes to the composition and characteristics of the identified clusters.

Recap

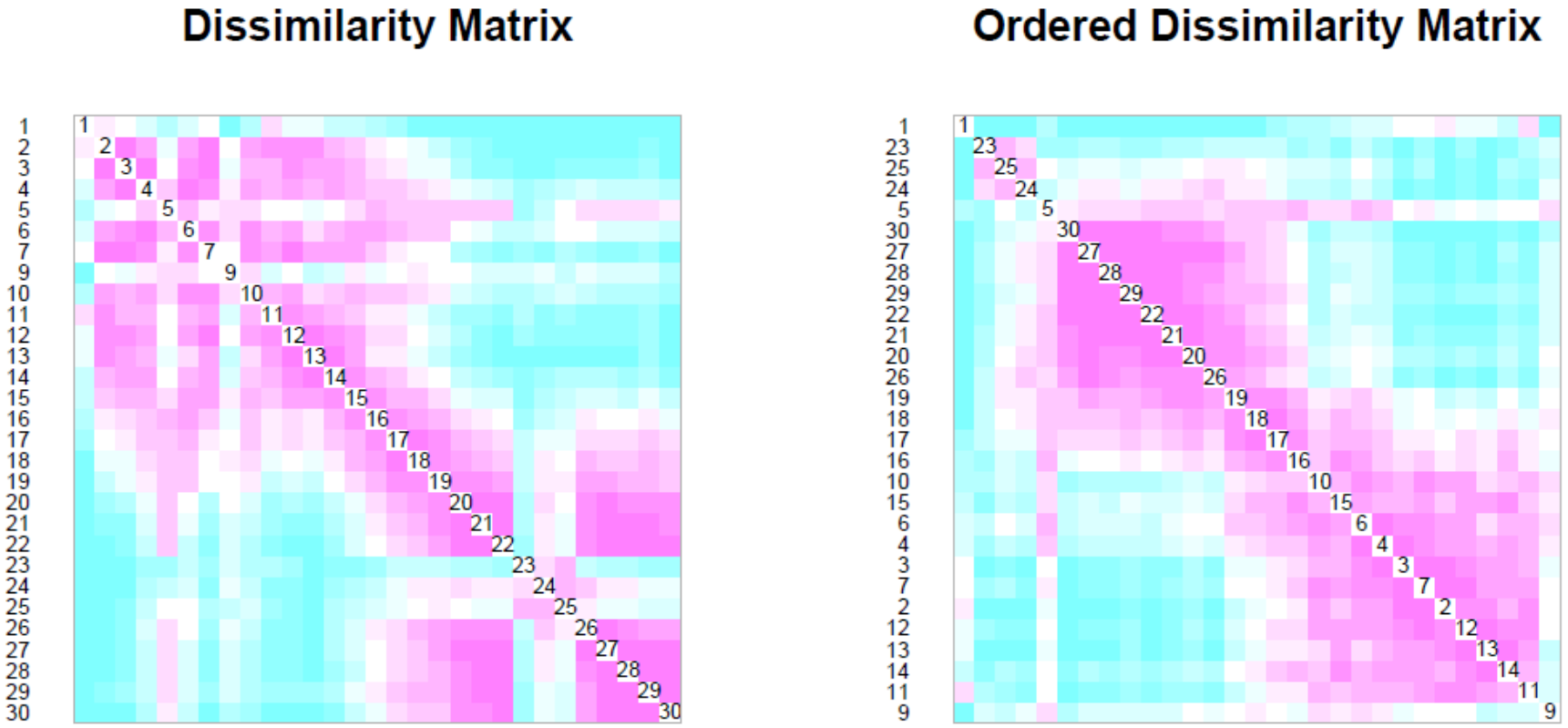
First session

- data exploration
- summary statistics
- visualization

Second session

- transformations
- resemblance
  - dis/similarity, distance

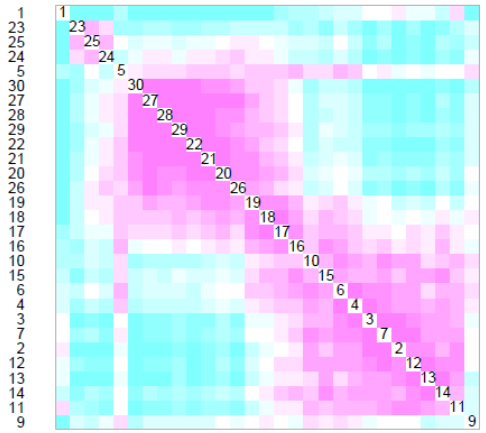
reminder: Bray-Curtis dissimilarity



Classification

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

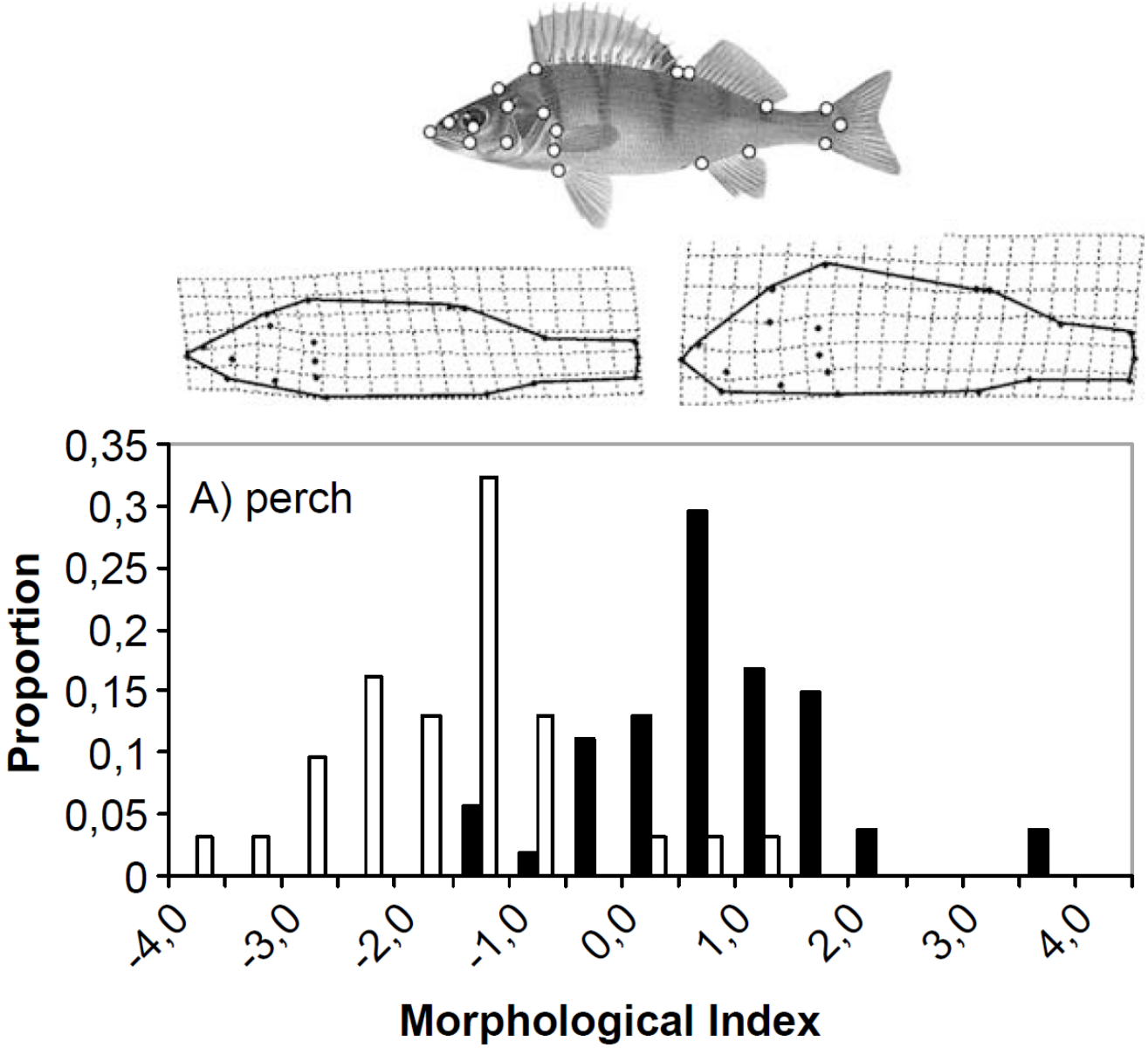
- name them (e.g. to ease communication)
- explore patterns and structure of dataset
- identify groups, types (typology)



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

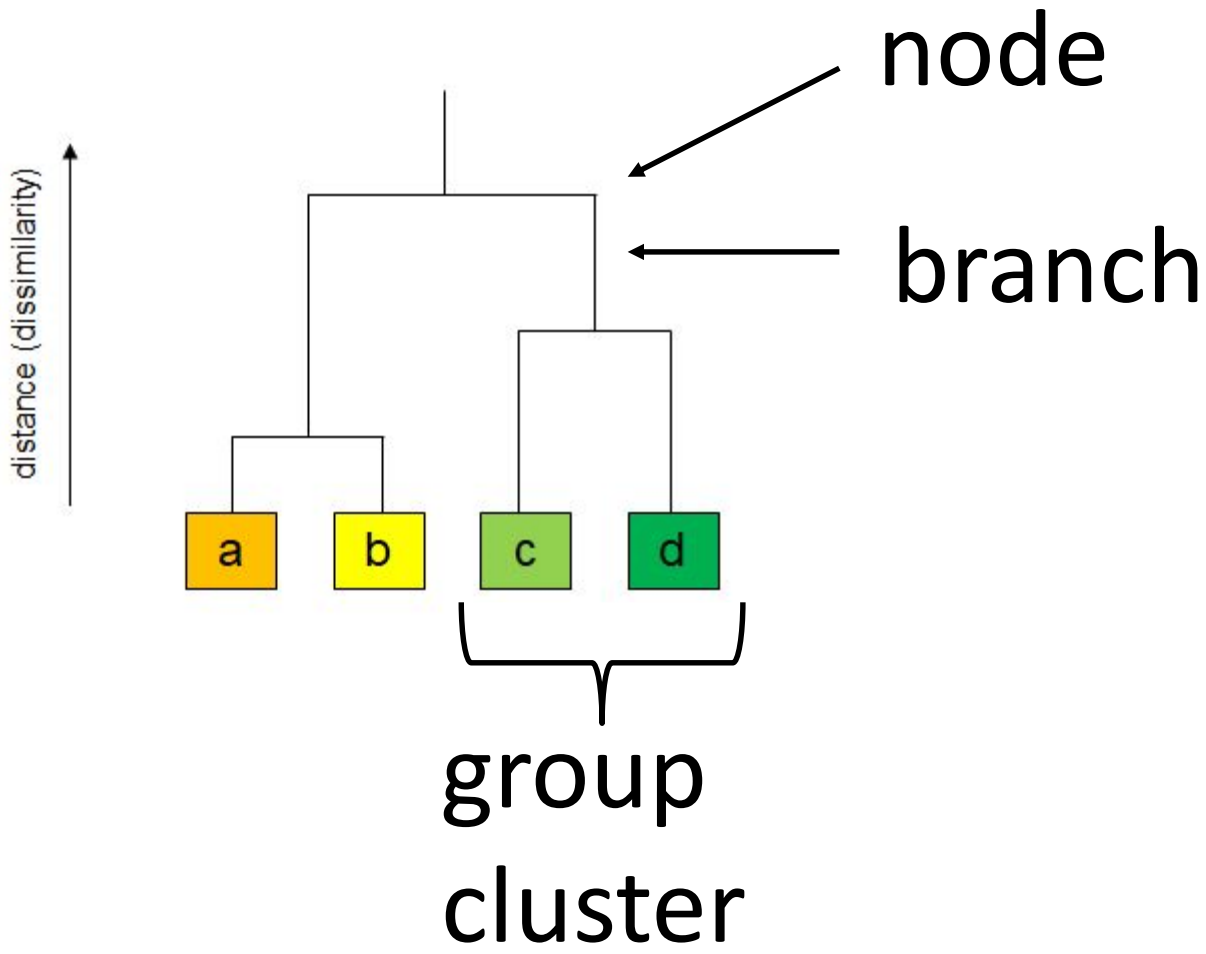
- maintain hierarchy of similarity within group (groups can cluster inside other groups)
- e.g. cluster analysis

=> dendrograms

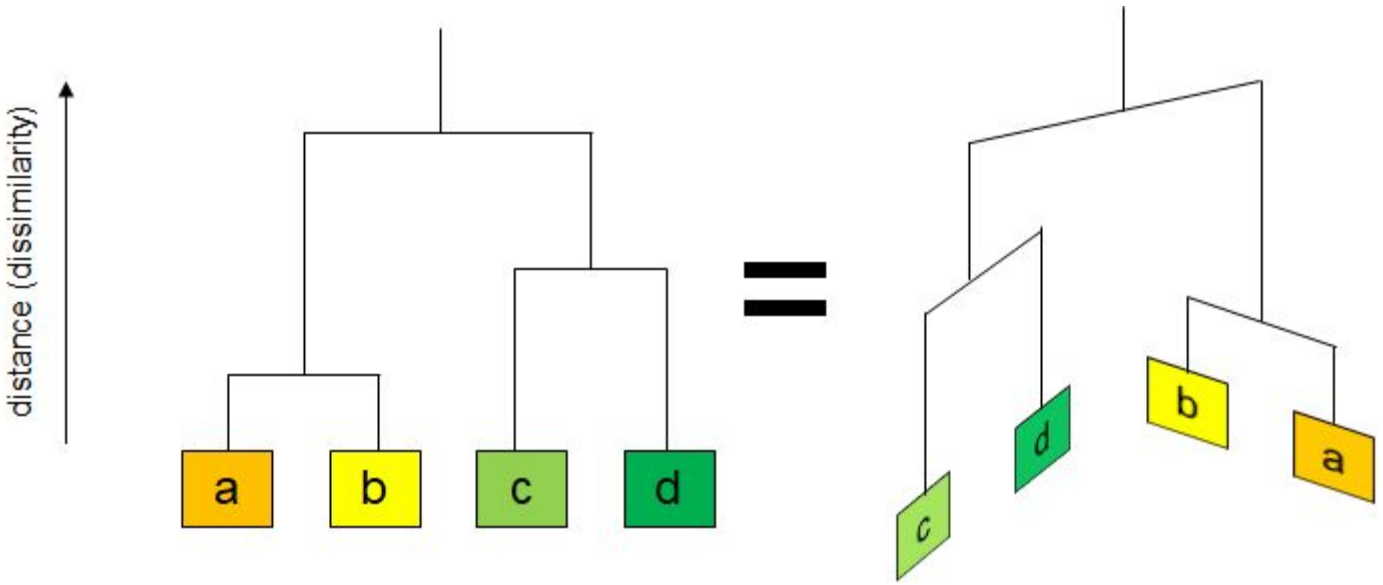
Selection of a grouping criteria

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

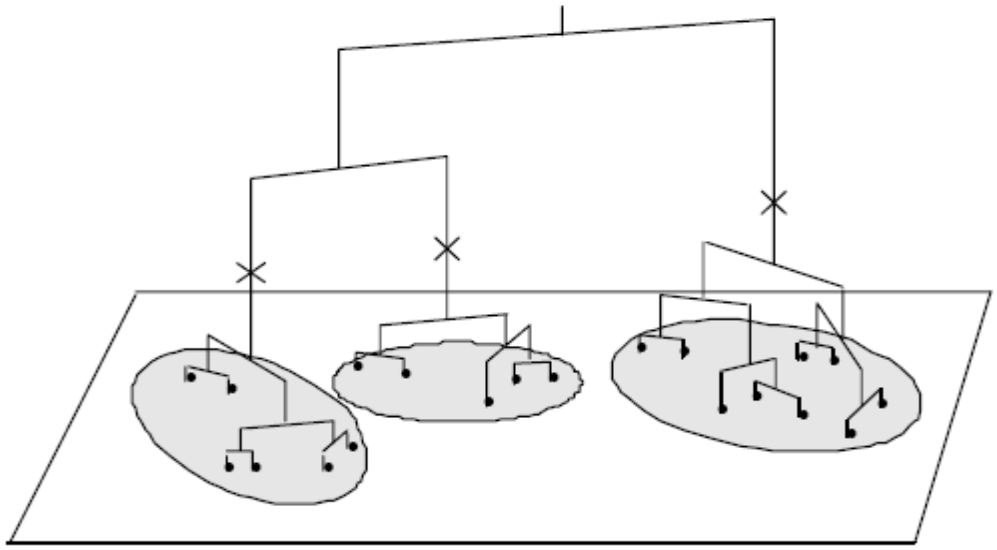
dendrograms



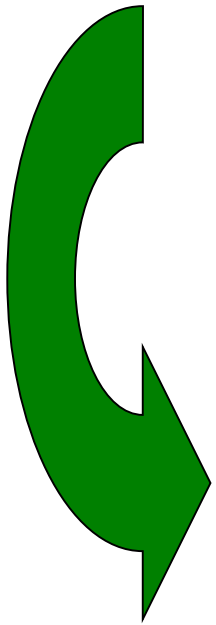
dendrograms

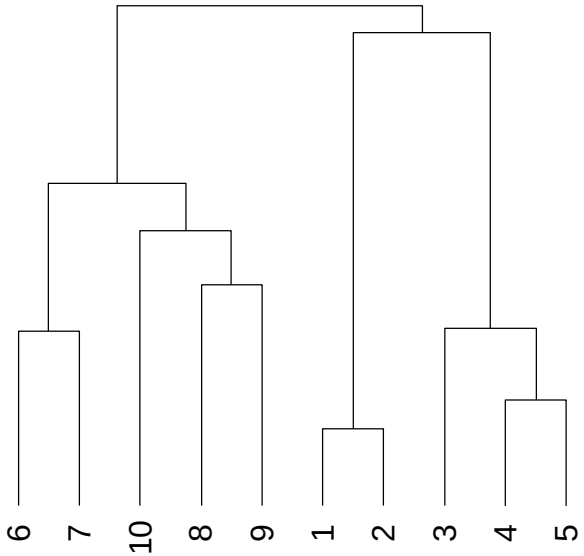


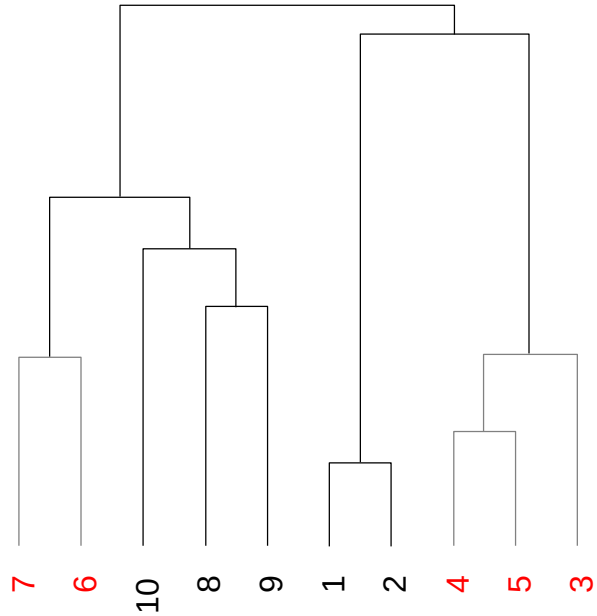
dendrograms



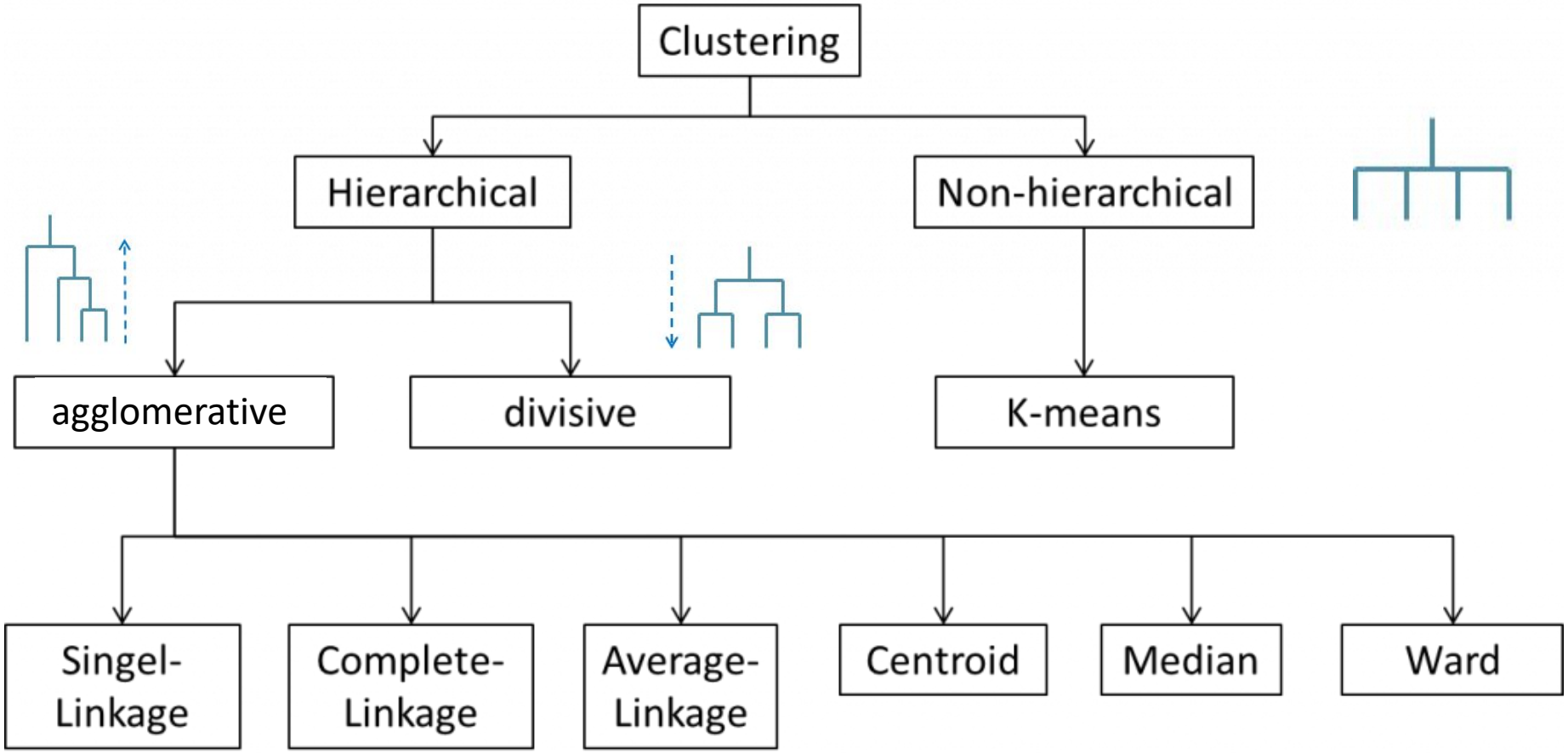
The order of tips on dendrograms can not be used for the interpretation of resemblance!





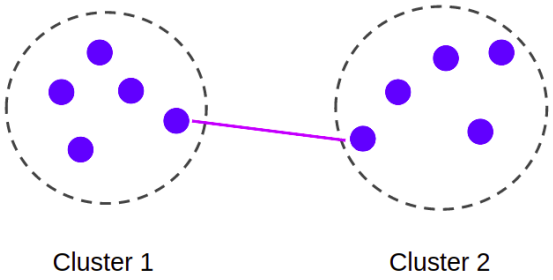


classification of classification methods



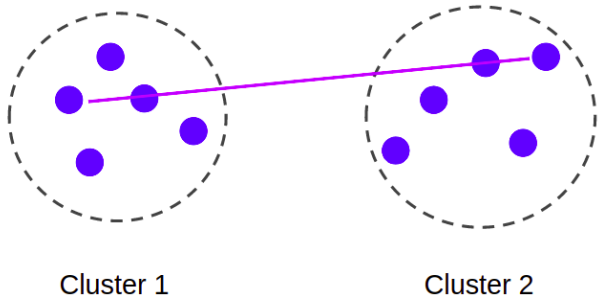

single-linkage
he table with my
best friend
average-linkage
the table with the
highest average
“friendliness”
complete-linkage
the table at with the
person I don’t like the
most is still not that bad

Simple linkage



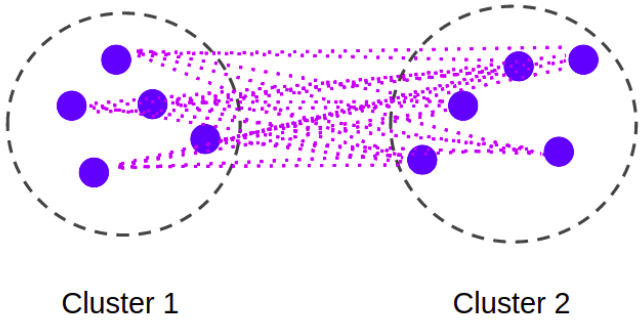
Distance between clusters is defined by the distance between their closest members.

Complete linkage



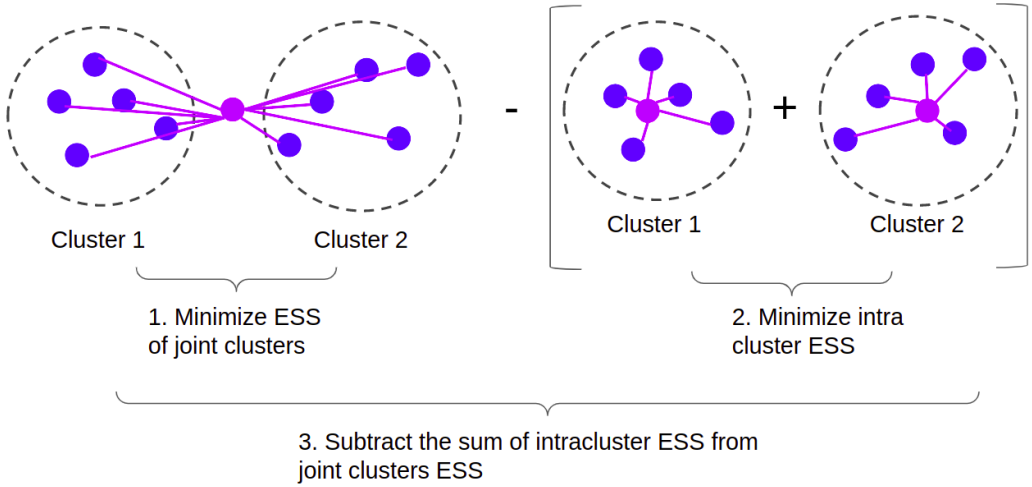
Distance between clusters is defined by the distance between their furthest members.

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

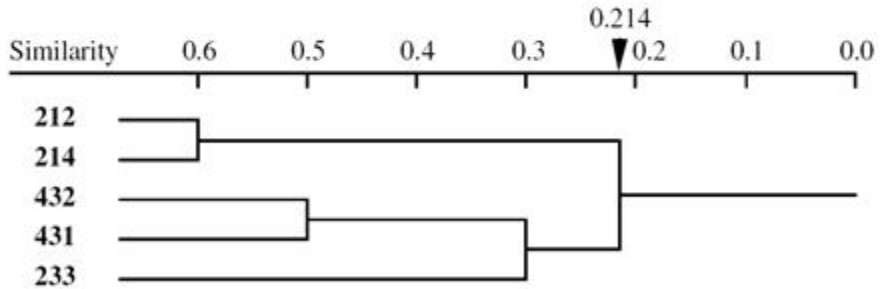
Single linkage algorithm

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

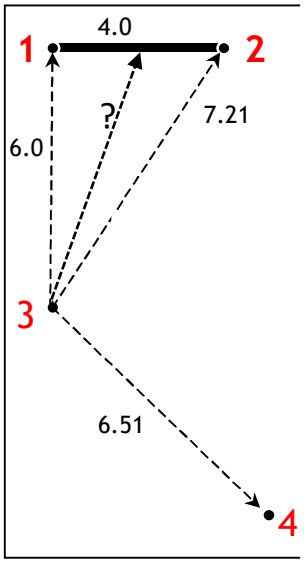


resulting dendrogram

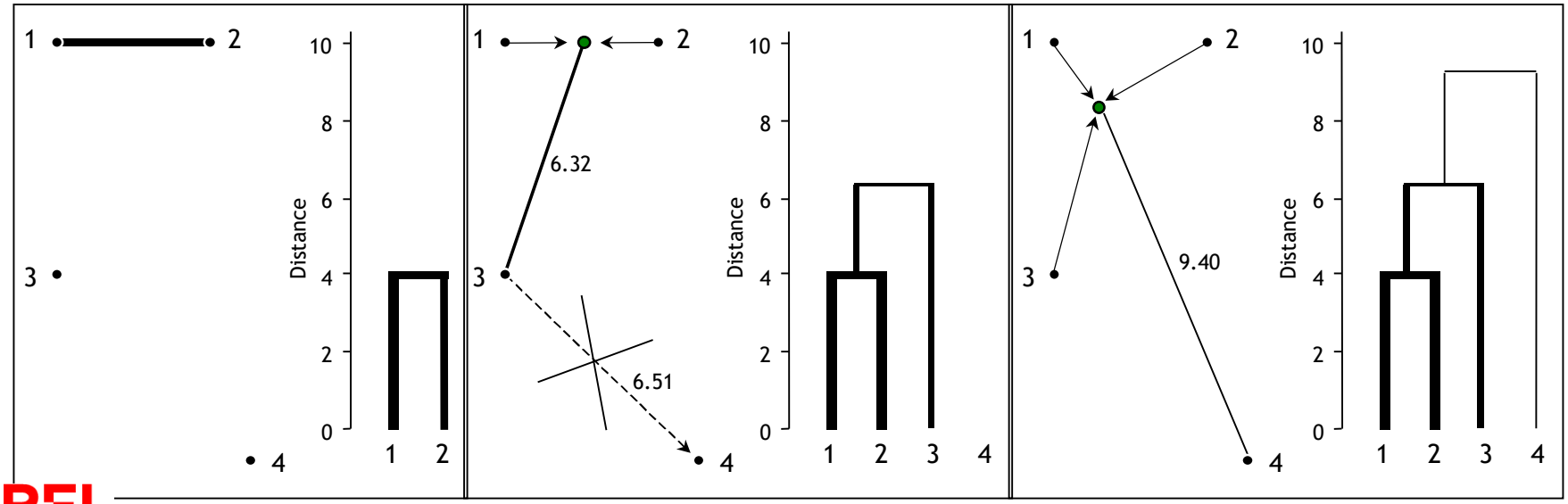
pair group methods (PGM)

	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

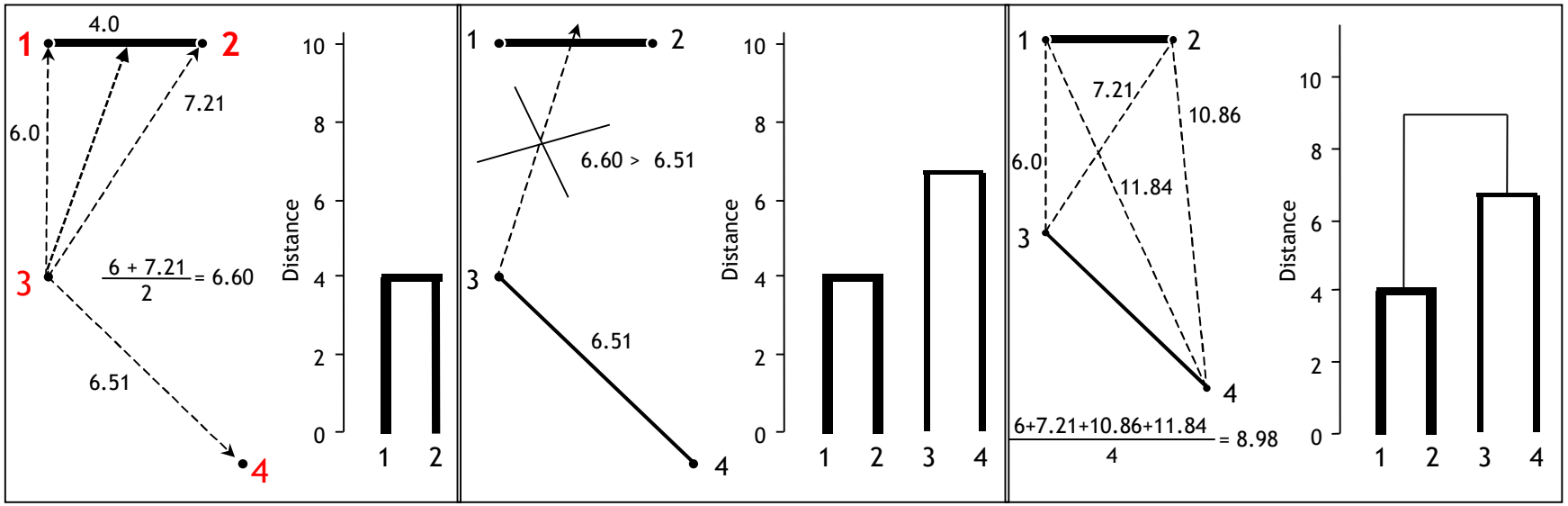
UPGMA (unweighted pair group method with arithmetic mean)



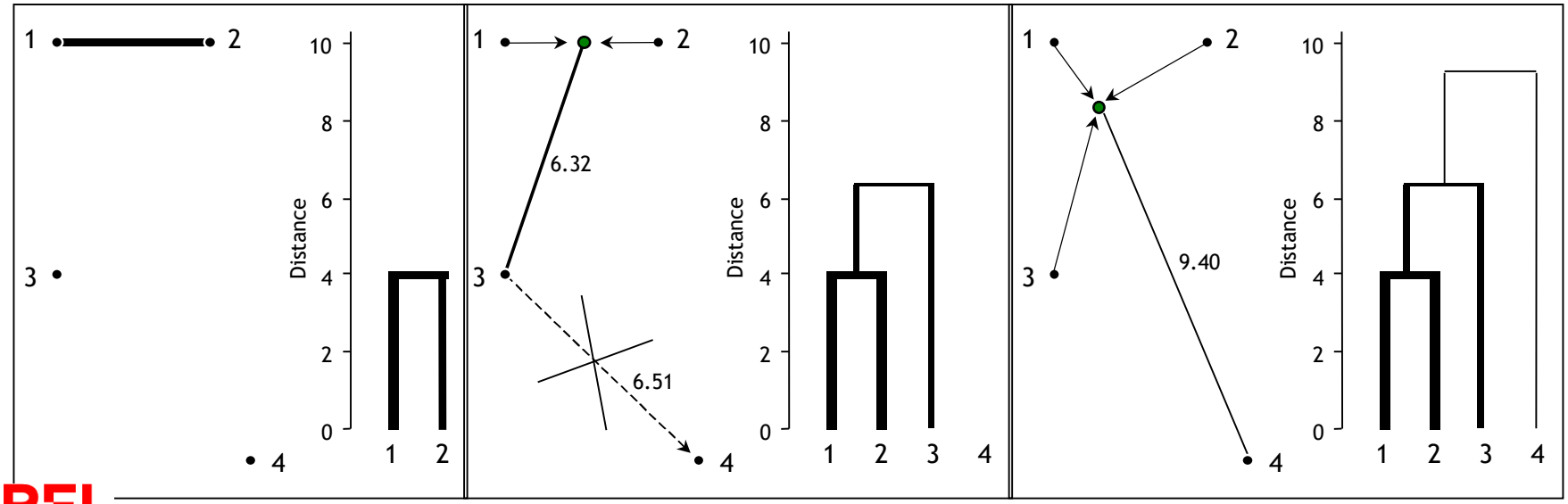
UPGMC (unweighted pair group method with centroids)



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:

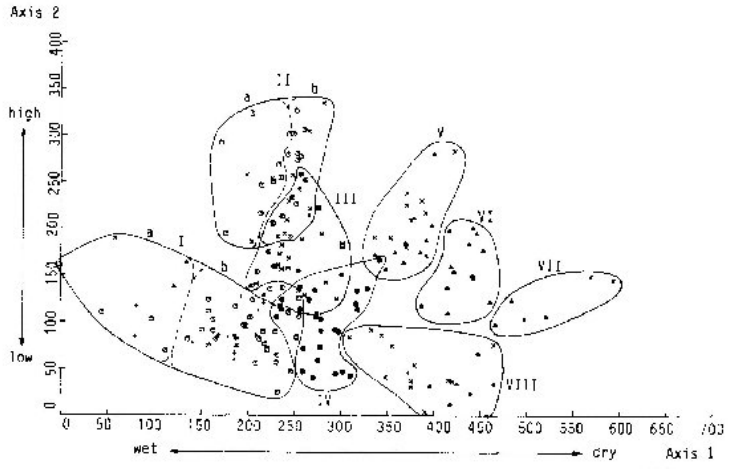
1. Samples are ordinated

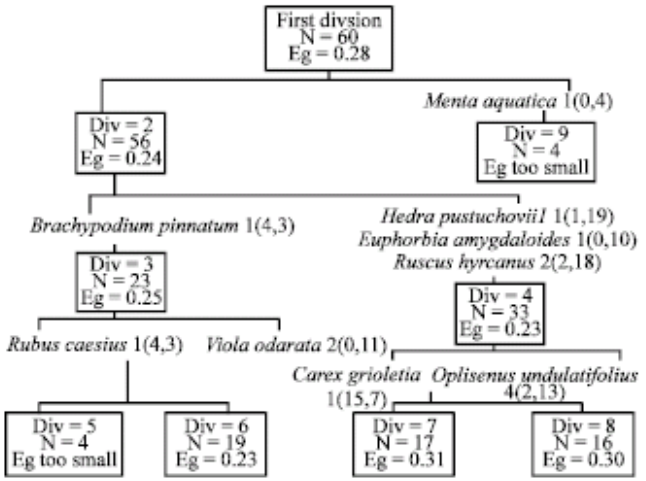
2. A crude dichotomy is formed: the ordination centroid is used as a dividing line between two groups (negative and positive)

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):

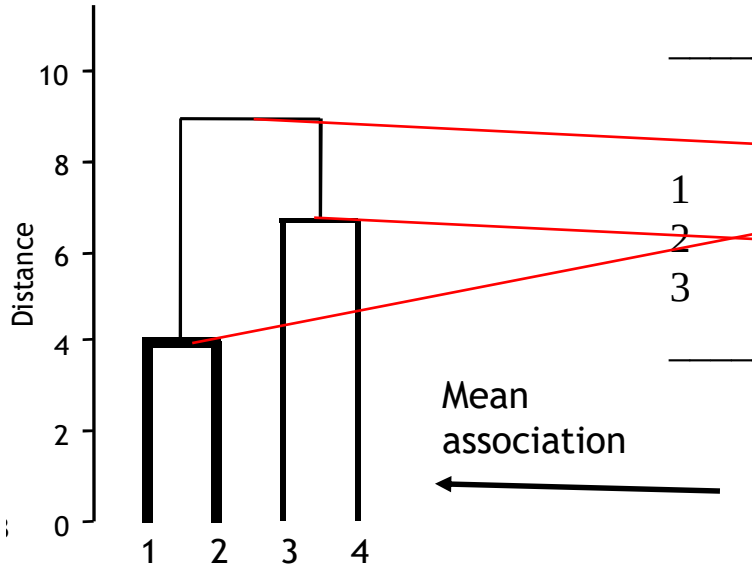
- Compact Groups
  - Minimal intra-group variance
  - Elements grouped at low distance level
- Groups of comparable sizes
  - Roughly the same number of elements in each group
  - No or very few groups with only one element
- Distinctly separated groups
  - Maximal inter-group variance

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

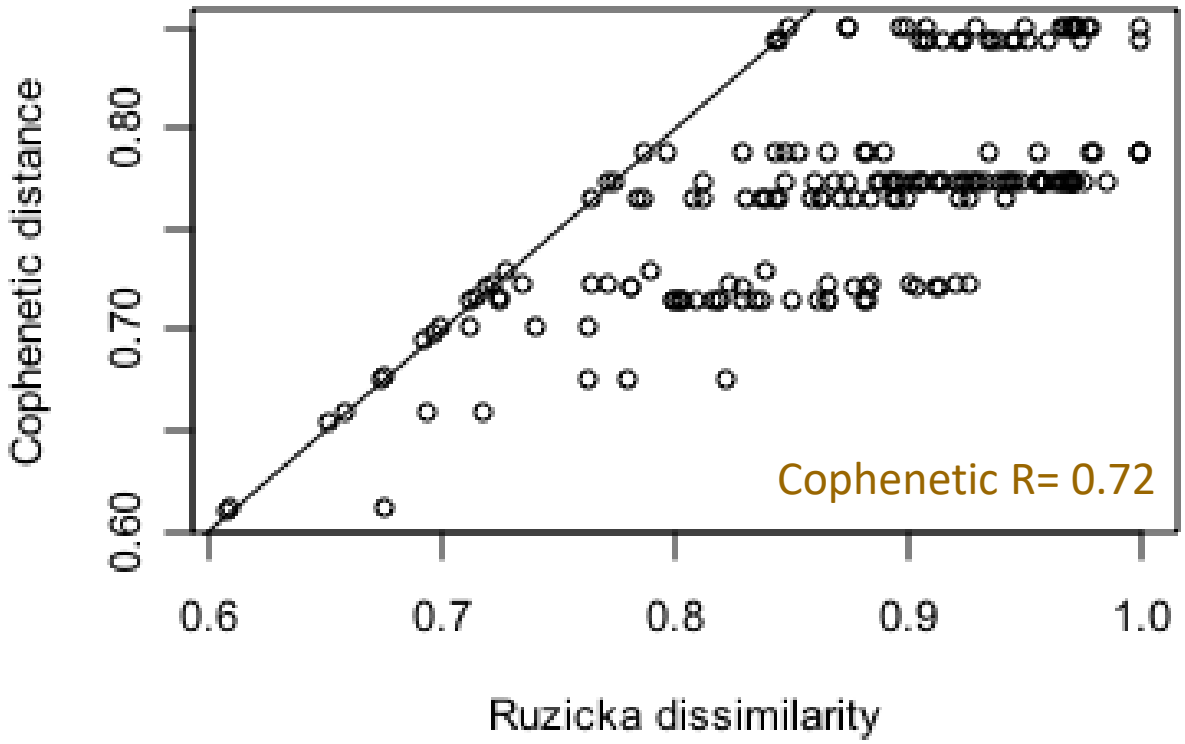
cophenetic correlation

- 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

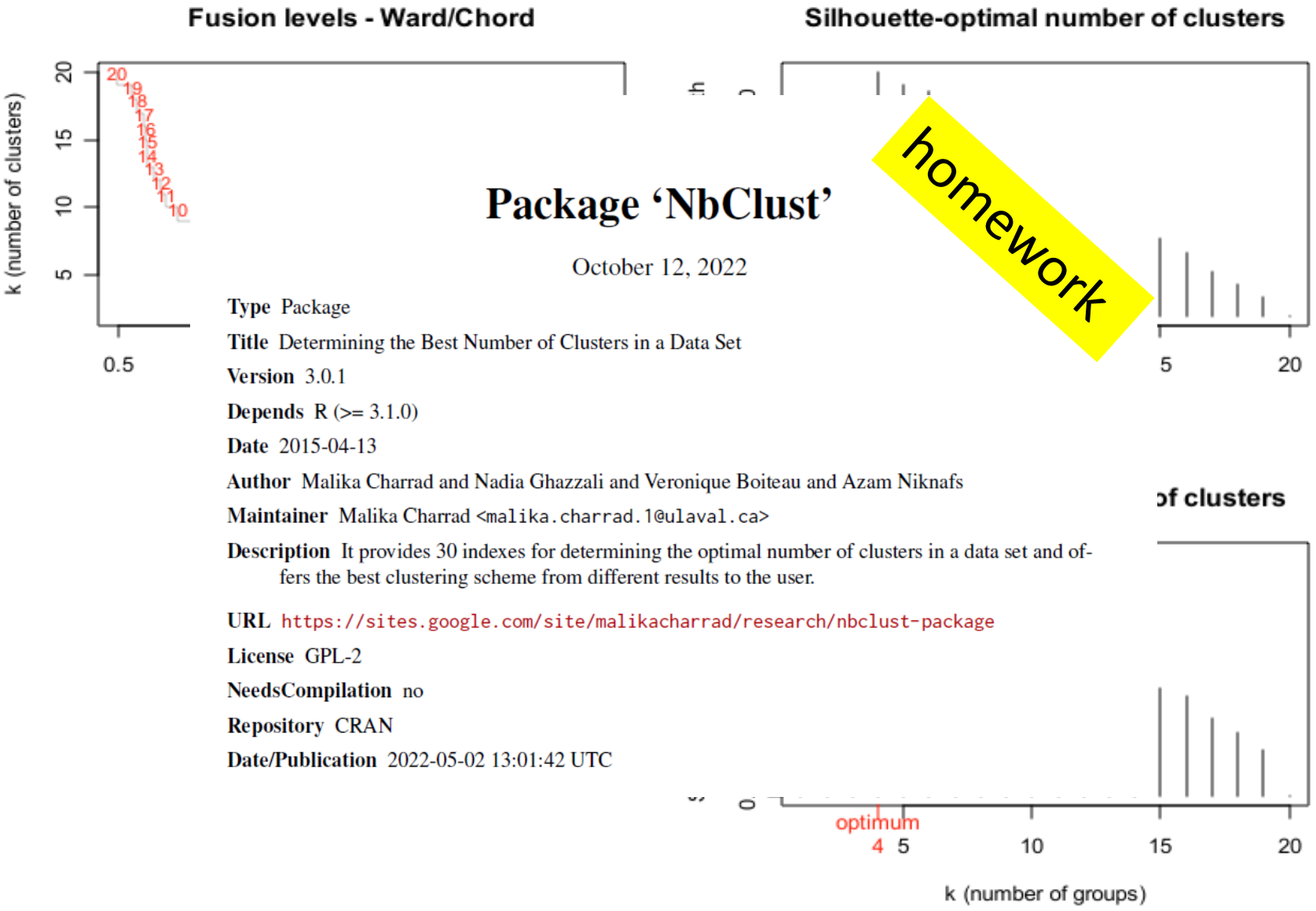
Shepard Diagrams

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

Paper for next week

Deep-Sea Research II 109 (2014) 293–299

Contents lists available at ScienceDirect

Deep-Sea Research II

journal homepage: www.elsevier.com/locate/dsr2

Regular article

Connecting subsistence harvest and marine ecology: A cluster analysis of communities by fishing and hunting patterns

Martin Renner*, Henry P. Huntington

Tern Again Consulting, 388 E Bayview Ave, Homer, AK 99603, USA

ARTICLE INFO

Available online 19 March 2014

Keywords:
Alaska
Subsistence harvest
Community ecology
Fisheries
Cluster analysis

ABSTRACT

Alaska Native subsistence hunters and fishers are engaged in environmental sampling, influenced by harvest technology and cultural preferences as well as biogeographical factors. We compared subsistence harvest patterns in 35 communities along the Bering, Chukchi, and Beaufort coasts of Alaska to identify affinities and groupings, and to compare those results with previous ecological analyses done for the same region. We used hierarchical cluster analysis to reveal spatial patterns in subsistence harvest records of coastal Alaska Native villages from the southern Bering Sea to the Beaufort Sea. Three main clusters were identified, correlating strongly with geography. The main division separates coastal villages of western Alaska from arctic villages along the northern Chukchi and Beaufort Seas and on islands of the Bering Sea. K-means groupings corroborate this result, with some differences. The second node splits the arctic villages, along the Chukchi, Beaufort and northern Bering Seas, where marine mammals dominate the harvest, from those on islands of the Bering Sea, characterized by seabird and seal harvests. These patterns closely resemble eco-regions proposed on biological grounds. Biogeography thus appears to be a significant factor in groupings by harvest characteristics, suggesting that subsistence harvests are a viable form of ecosystem sampling.

1. Introduction

Subsistence hunting and fishing account for a large proportion of the food produced and consumed in rural Alaska (ADF&G, 2012). The types of fish, marine mammals, seabirds, invertebrates, and plants that are harvested reflect cultural preferences, access, harvest technology, and of course the underlying ecology of the surrounding land, freshwaters, and sea (e.g., Wolfe, 2004). In effect, we attempt to use subsistence harvests as a means of sampling the local environment, acknowledging that hunting and fishing practices depend on more than just the presence of potential prey species and that the existence of a potential prey species does not necessarily mean it will be harvested. Analyzing and comparing community-level harvests offered the prospect of insights into biogeographical patterns. By comparing our analysis to previous analyses done on the available fish and seabird fauna, we also hoped to be able to assess the degree to which cultural or other factors further influence subsistence harvest patterns, by identifying any anomalies that could not be explained by biogeographical factors.

The biogeographical contribution is implied in various regional characterizations of subsistence hunting in Alaska, for example showing that marine mammals are the largest category by weight of harvest in the region of the state designated as Arctic by the Alaska Department of Fish and Game (ADF&G), whereas fish occupy the top spot in all other regions of the state (Huntington et al., 1998). The reason for this difference is readily apparent. Bowhead whales (*Balaena mysticetus*), for example, occur and are harvested only in the Arctic, and Pacific walrus (*Odobenus rosmarus divergens*) are only taken in small numbers in southwestern Alaska in contrast to harvests of several hundred animals per year in several arctic communities (ADF&G, N.D.). A more detailed look at regional and community harvest patterns, however, has not previously been undertaken.

Although aspects of the ecology of subsistence harvests and similar local uses of plants and animals have been examined in depth in many parts of the world (e.g., Smith and Winterhalder, 1981; Hurtado and Hill, 1987; Smith, 1991; Lauer and Aswani, 2008), we are unaware of any studies that have looked at regional patterns. Here we use cluster analysis of the harvest patterns of Alaska Native communities to identify patterns across communities and regions.

We use subsistence harvest data from 35 Aleut, Yup'ik, St. Lawrence Island Yupik, and Iñupiaq communities on the coasts and islands of the Bering, Chukchi, and Beaufort Seas. First we

* Corresponding author. Tel.: +1 907 226 4672.
E-mail addresses: auklet@bigfoot.com (M. Renner), hph@alaska.net (H.P. Huntington).

http://dx.doi.org/10.1016/j.dsr2.2014.03.005
0967-0645/

EPFL