Clustering

- *Objectives:*

Discuss the theory and practice of clustering

Illustrate diverse applications of this technique

Disclaimer: *In ecology and systematics, "cluster analysis" usually means agglomerative hierarchical cluster analysis.*

However, there are 100's of different (and diverse) methods:

Some are divisive (break-up groups)
Others place samples in multiple clusters

For overview, see Clarke & Warwick (2001)

Properties of Hierarchical Clustering

Hierarchical clustering (also called hierarchical cluster analysis or HCA) is a classification method which seeks to build a hierarchy of clusters.

It can follow two approaches:

•**Agglomerative** ("bottom up“): each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

•**Divisive** ("top down“): all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Properties of Hierarchical Clustering

Three key properties of hierarchical strategies:

Combinatorial or noncombinatorial

Compatible or incompatible

Space-conserving or space-distorting

Properties I

Combinatorial or not: Can all distances be calculated from original dissimilarity matrix ?

The basic combinatorial equation is

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma |d_{ip}^2 - d_{iq}^2|$$

where values of α_p, α_q, β, and γ determine the type of sorting strategy (Table 11.1). Think of these parameters as weights that define how distances from two groups are fused into a set of new distances for the new group.

Why does it matter: Combinatorial methods are faster and easier to compute (require less memory)

Properties II

Compatible or not: Are the dissimilarities consistently calculated using the same measures

A **compatible** strategy is one in which the dissimilarities calculated later in the analysis are calculated in the same fashion as the initial dissimilarity matrix.

An example of an incompatible strategy would be to choose Sørensen (Bray-Curtis) dissimilarity along with a hierarchical method that calculates the new intergroup dissimilarities as Euclidean distances. Incompatible strategies should be considered experimental at present.

Why does it matter: Compatible approaches are consistent.

TO AVOID INCOMPATIBILITIES – check next table

Summary of properties of linkage methods / distance measures

Linkage Method	Dissimilarity Measure: Euclidean distance (absolute and relative) — Combinatorial compatible?	Euclidean distance — Space contracting, expanding, or conserving?	Sørensen distance (1 - 2w/a+b) — Combinatorial compatible?	Sørensen distance — Space contracting, expanding, or conserving?
Nearest neighbor (single linkage)	yes	contracting	yes	contracting
Farthest neighbor (complete linkage)	yes	expanding	yes	expanding
Median (Gower's method)	yes	contracting	no	unknown
Group average (average linkage)	yes	conserving	yes	conserving
Centroid (weighted group)	yes	contracting	no	contracting
Ward's method (Orloci's method)	yes	conserving	no	unknown
Flexible beta	yes	flexible	yes	flexible
McQuitty's method	yes	contracting	no	unknown

Properties III

Space conserving or not: Are relative distances conserved

The initial dissimilarity matrix can be thought of as defining distances in a space with certain properties conferred by the choice of dissimilarity measure. As groups form, measures of intergroup distances may alter the original properties of the space. If the properties of the original space are preserved, then the strategy is **space-conserving**. With certain strategies the space in the vicinity of a group may become expanded or contracted. Such strategies are **space-distorting**. Chaining is the result of a **space-contracting** strategy.

Why does it matter: Affects the shape of the dendogram

Summary of properties of linkage methods / distance measures

Linkage Method	Dissimilarity Measure: Euclidean distance (absolute and relative) — Combinatorial compatible?	Euclidean distance — Space contracting, expanding, or conserving?	Sørensen distance (1 - 2w/a+b) — Combinatorial compatible?	Sørensen distance — Space contracting, expanding, or conserving?
Nearest neighbor (single linkage)	yes	contracting	yes	contracting
Farthest neighbor (complete linkage)	yes	expanding	yes	expanding
Median (Gower's method)	yes	contracting	no	unknown
Group average (average linkage)	yes	conserving	yes	conserving
Centroid (weighted group)	yes	contracting	no	contracting
Ward's method (Orloci's method)	yes	conserving	no	unknown
Flexible beta	yes	flexible	yes	flexible
McQuitty's method	yes	contracting	no	unknown

Reminder About Distance Measures

What range of values can they take on? Are they metrics?

Name (synonyms)	Domain of x	Range of $d = f(x)$	Comments
Euclidean (Pythagorean)	all	non-negative	metric
Relative Euclidean (Chord distance; standardized Euclidean)	all	$0 \le d \le \sqrt{2}$ for quarter hypersphere; $0 \le d \le 2$ for full hypersphere	Euclidean distance between points on unit hypersphere; metric
Sørensen (Bray & Curtis; Czekanowski)	$x \ge 0$	$0 \le d \le 1$ (or $0 \le x \le 100\%$)	proportion coefficient in city-block space; semimetric
Relative Sørensen (Kulczynski; Quantitative Symmetric)	$x \ge 0$	$0 \le d \le 1$ (or $0 \le x \le 100\%$)	proportion coefficient in city-block space; same as Sørensen but data points relativized by sample unit totals; semimetric
Jaccard	$x \ge 0$	$0 \le d \le 1$ (or $0 \le d \le 100\%$)	proportion coefficient in city-block space; metric

Recommendations

Euclidean / Relative Euclidean Distance Metrics

Linkage Method	Combinatorial compatible?	Space contracting, expanding, or conserving?
Nearest neighbor (single linkage)	yes	contracting
Farthest neighbor (complete linkage)	yes	expanding
Median (Gower's method)	yes	contracting
Group average (average linkage)	yes	conserving
Centroid (weighted group)	yes	contracting
Ward's method (Orloci's method)	yes	conserving
Flexible beta	yes	flexible
McQuitty's method	yes	contracting

All eight linkage methods are compatible

But, only two do not distort the relationships in variable space:

Group Average
Ward's method

Recommendations

Sorensen / Relative Sorensen Distance Semi-Metric

Linkage Method	Combinatorial compatible?	Space contracting, expanding, or conserving?
Nearest neighbor (single linkage)	yes	contracting
Farthest neighbor (complete linkage)	yes	expanding
Median (Gower's method)	no	unknown
Group average (average linkage)	yes	conserving
Centroid (weighted group)	no	contracting
Ward's method (Orloci's method)	no	unknown
Flexible beta	yes	flexible
McQuitty's method	no	unknown

Four linkage methods are compatible

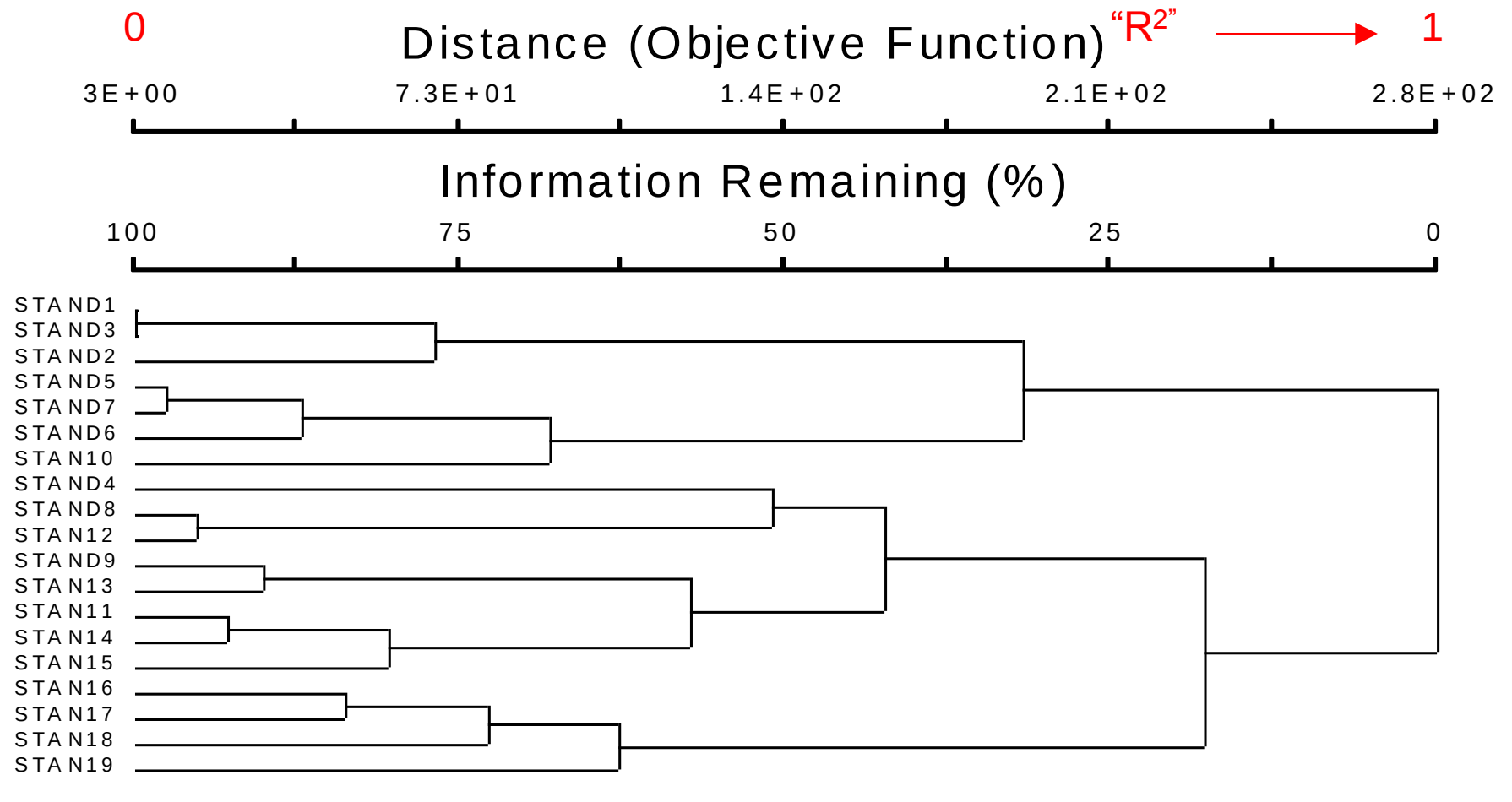
Only one does not distort the relationships in variable space:

Group Average

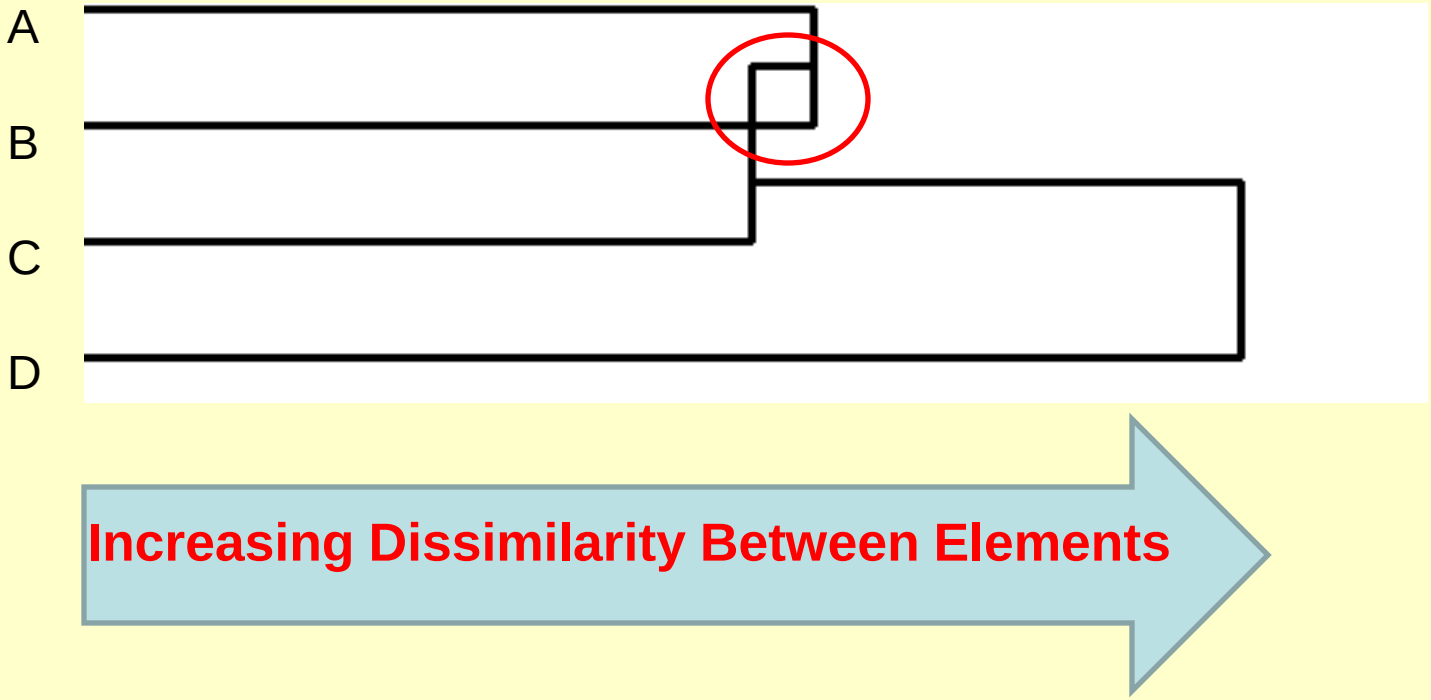
Dendogram Properties I

The objective function rescaled from 0% to 100% of information:

% information remaining = 100 (SSt - E) / SSt



Dendogram Properties II

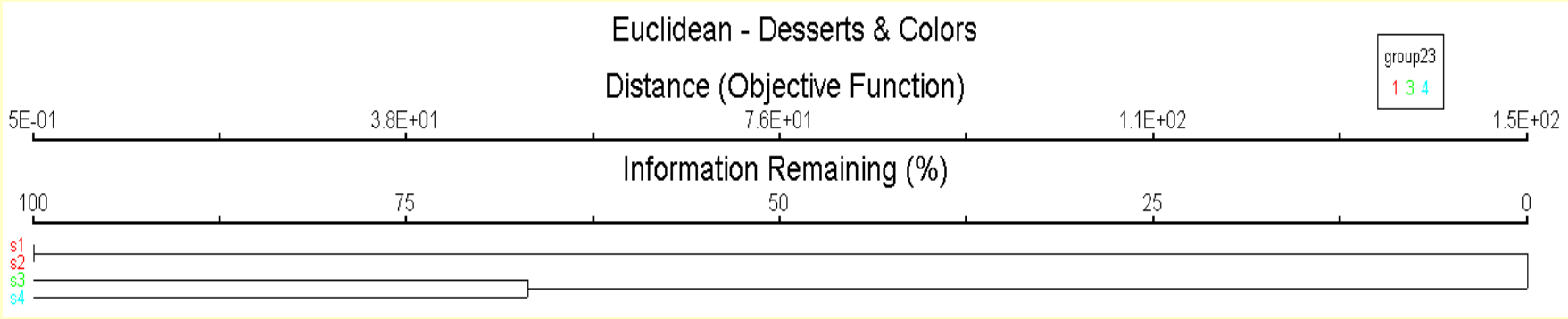


Elements in a dendrogram are always linked according to the “objective function” (more similar elements linked first)

Take Home: Successive Links cannot “decrease” in similarity

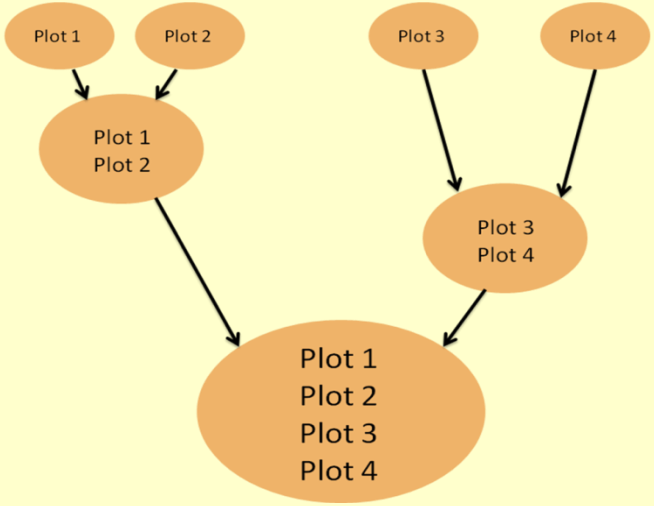
Dendogram Properties III

The samples are labelled by group membership:



Main - Clustering_Example.wk1

4	Stands	
2	Species	
	Q	Q
	sp1	sp2
s1	1	0
s2	1	1
s3	10	0
s4	10	10



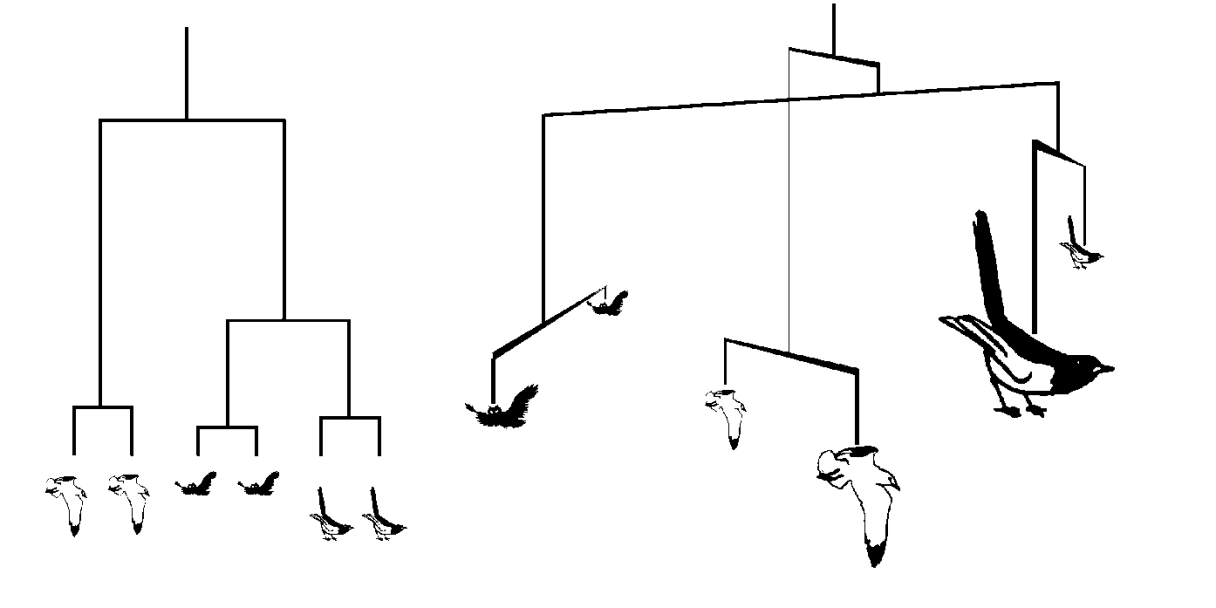
Second - WORK2.WK1

4	Stands	
2	Groups	
	C	C
	group23	group22
s1	1	1
s2	1	1
s3	3	3
s4	4	3

Main Matrix: Data
(Input by user)

Second Matrix: Groups
(Output by computer)

Dendogram Properties IV



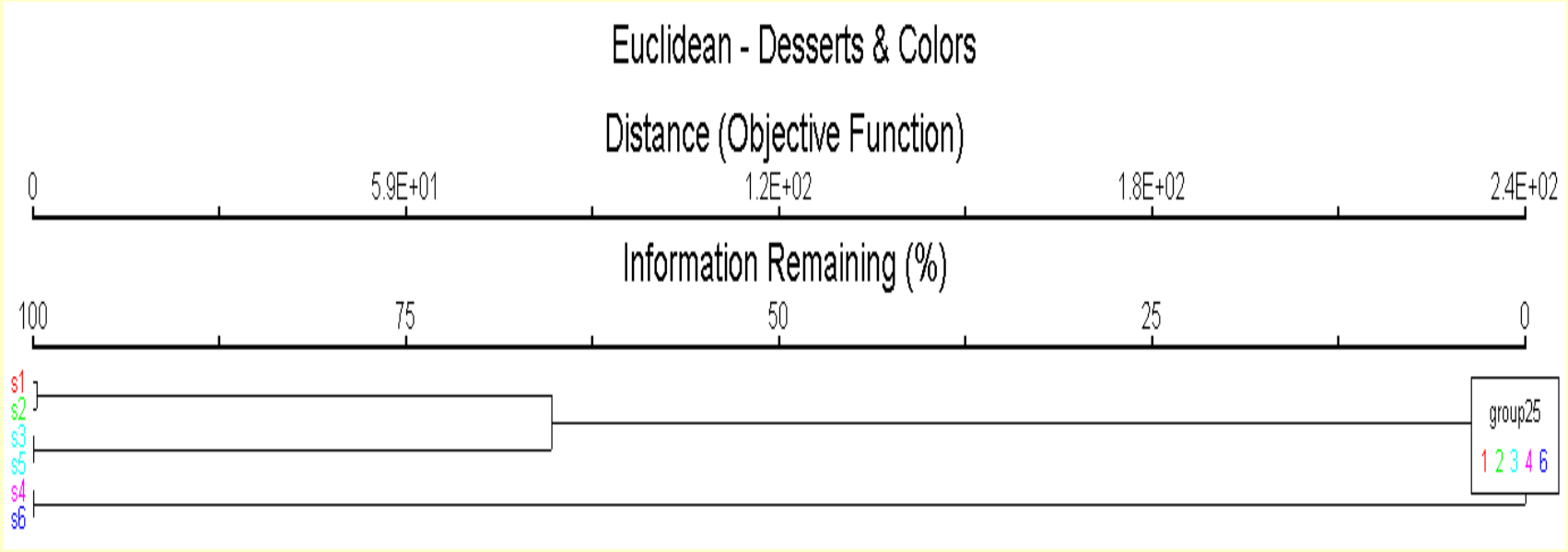
A dendrogram is an inherently nondimensional representation. Imagine the branches as free to pivot, like a child's mobile.

Dendogram Properties V

Multiple sample pairs can be linked on same cycle:

	Sp1	sp2
s1	1	0
s2	1	1
s3	10	0
s4	10	10
s5	10	0
s6	10	10

What happens when two sample pairs are at the same distance?

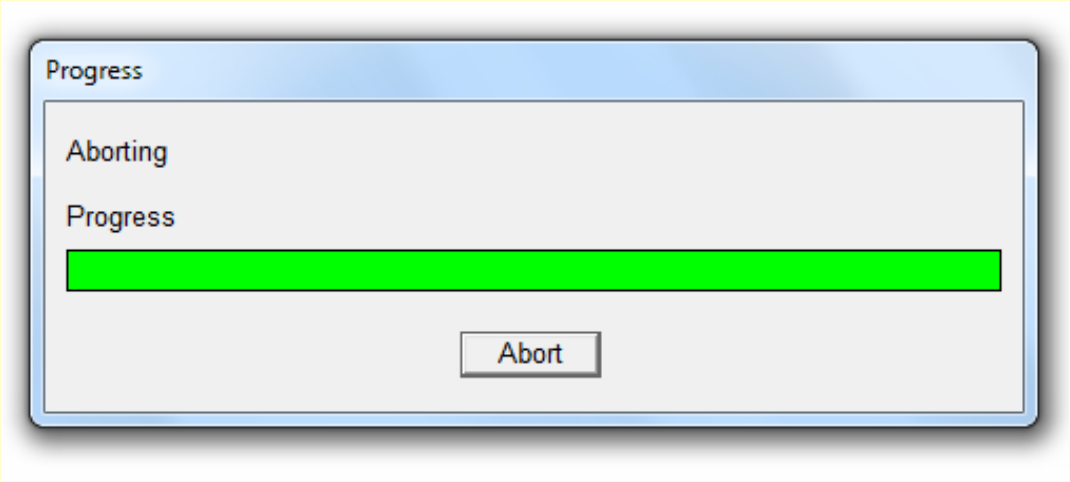


Dendogram Properties VI

Yet, there has to be structure in the data: distances.

	Sp1	Sp2
s1	1	10
s2	1	10
s3	1	10
s4	1	10
s5	1	10
s6	1	10

What happens when all sample pairs are at the same distance?



Clustering cannot be performed;

PC-ORD blows up

Clustering – An Example

Cluster step 1:

Calculate all pair-wise dissimilarities – across samples (see data matrix below).

Data Matrix

	Sp1	Sp2
Plot 1	1	0
Plot 2	1	1
Plot 3	10	0
Plot 4	10	10

Squared Euclidean Distance Matrix

	Plot 1	Plot 2	Plot 3	Plot 4
Plot 1	0	1	81	181
Plot 2	1	0	82	162
Plot 3	81	82	0	100
Plot 4	181	162	100	0

Clustering – An Example

Cluster step 2:

Combine group 2 (plot 2) into group 1 (plot 1) at given level of *E* . This fusion produces the least possible increase in the objective function (below).

$$E_t = \sum_{i=1}^{k_t}\sum_{j=1}^{p}(x_{ijt} - \bar{x}_{jt})^2$$

	Sp1	Sp2
Plot 1	1	0
Plot 2	1	1
Mean	1	0.5

$$E_1 = \sum_{i=1}^{2}\sum_{j=1}^{2}(x_{ij1} - \bar{x}_{j1})^2$$

$$= (1-1)^2 + (1-1)^2 + (0-0.5)^2 + (1-0.5)^2$$

$$= 0.5$$

Clustering – An Example

Cluster step 2:

Obtain the coefficients for basic combinatorial equation by applying the coefficients for Ward's method

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma |d_{ip}^2 - d_{iq}^2|$$

To calculate d for group 12 and sample 3:

NOTE:

r = new group (merge 1 & 2)

p = 1 (merged)

q = 2 (merged)

i = unmerged (3 and 4)

$$\alpha_1 = \frac{1+1}{1+2} = \frac{2}{3} \quad \alpha_2 = \frac{1+1}{1+2} = \frac{2}{3}$$

$$\beta = -\frac{1}{3} \quad \gamma = 0$$

	Plot 1	Plot 2	Plot 3	Plot 4
Plot 1	0	1	81	181
Plot 2	1	0	82	162
Plot 3	81	82	0	100
Plot 4	181	162	100	0

So, for sample 3: $d_{3,1+2}^2 = \frac{2}{3}(81) + \frac{2}{3}(82) - \frac{1}{3}(1) = \frac{325}{3} = 108.3$

Clustering – An Example

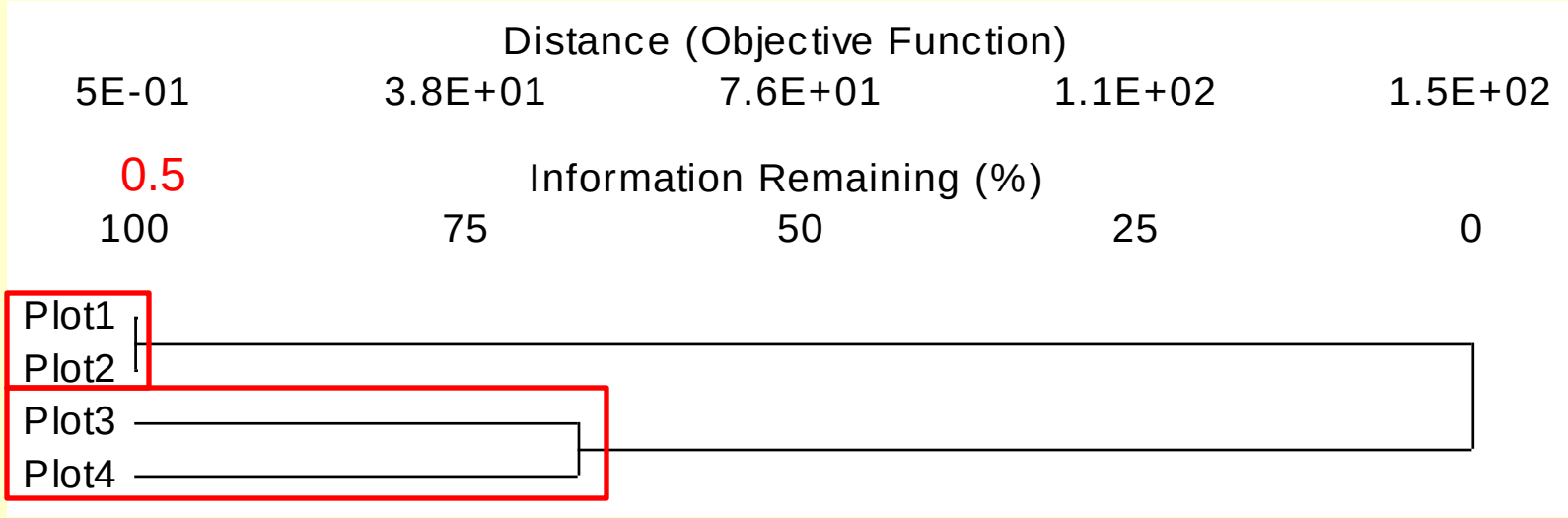
Cluster step 3:

Create new dissimilarity matrix, including the new group (union of plot 1 and plot 2)

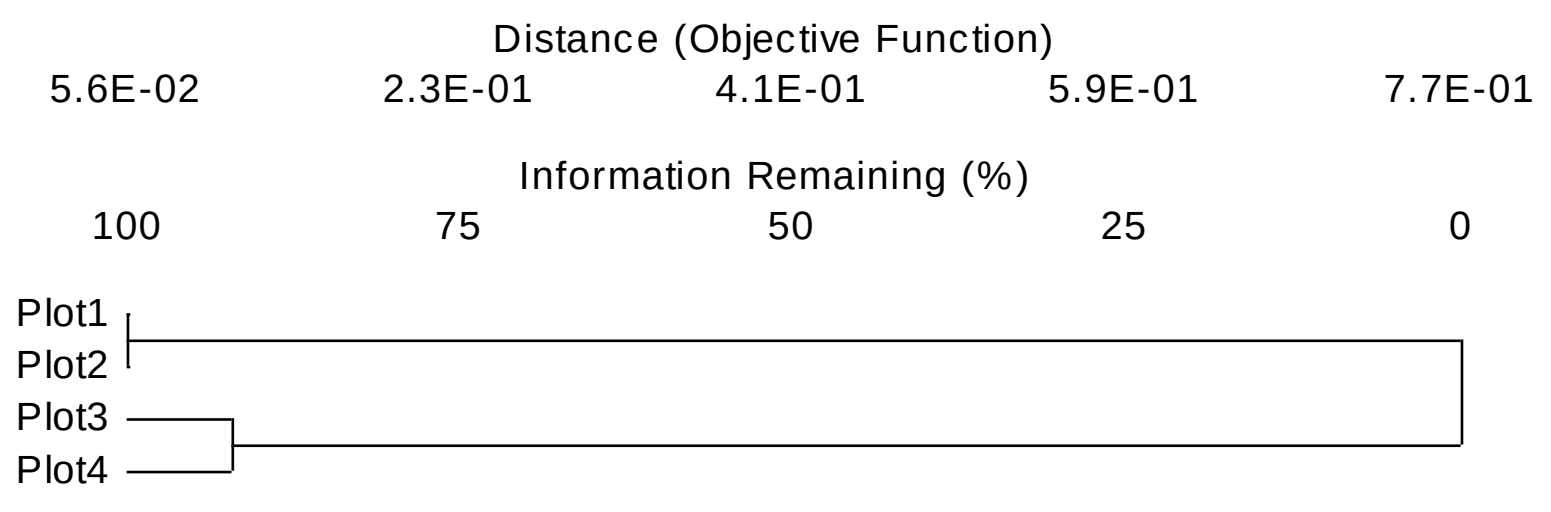
Revised distance matrix after the first fusion.

	Plots 1+2	Plot 3	Plot 4
Plots 1+2	0	108.3	228.3
Plot 3	108.3	0	100
Plot 4	228.3	100	0

Clustering – An Example



Cluster analysis of plots using Ward's method and Euclidean distance.

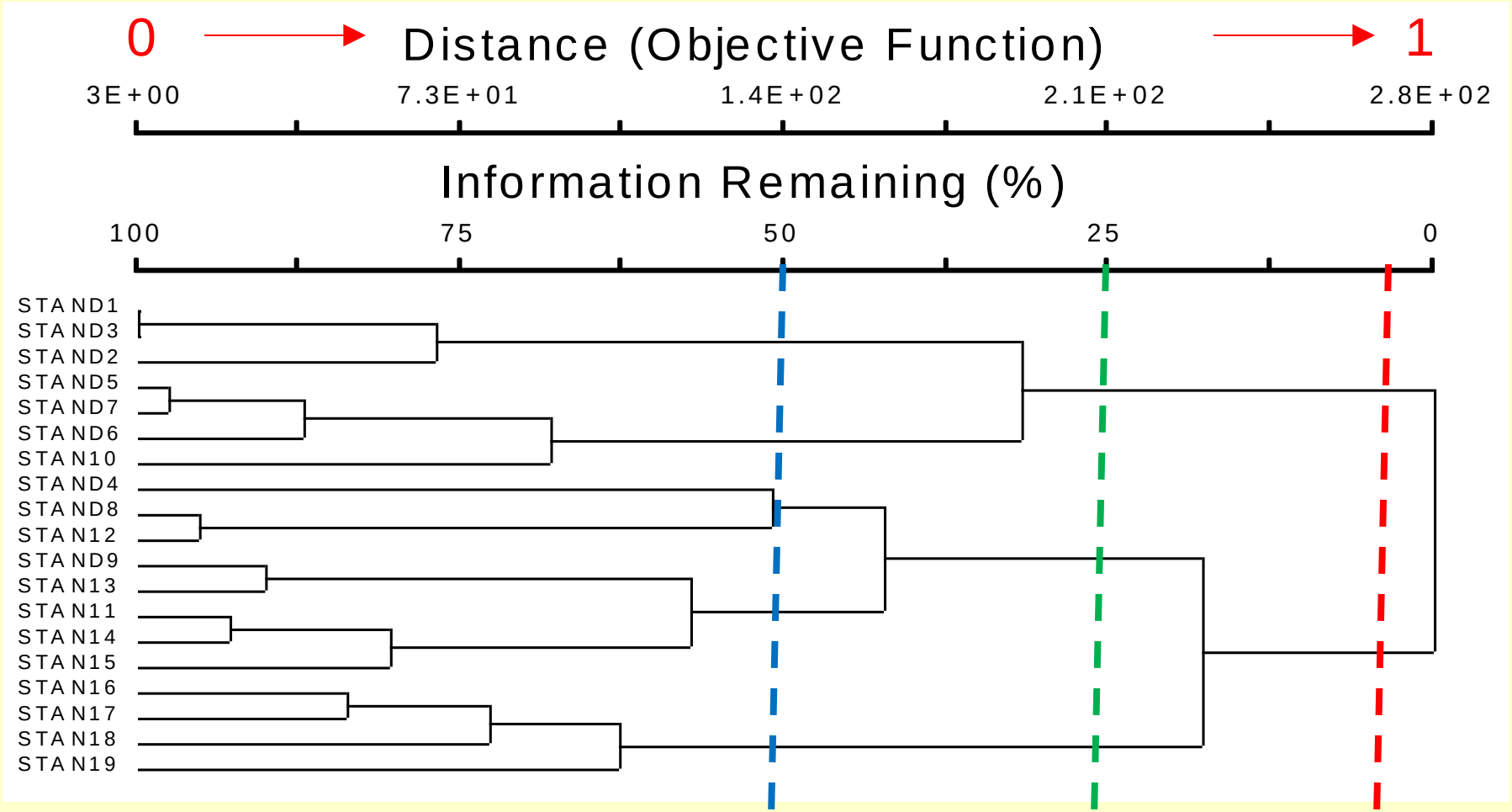


Cluster analysis of plots using Ward's method and Sørensen distance.

Classification - Output

Approach: Objects placed in groups according to “objective” function: similarity measure and a grouping algorithm.

NO information explained R^2 ALL information explained



Summary – The Good / The Bad / The Ugly

- Objects placed in groups according to similarity measure and then a grouping algorithm.

- The reduction in the data comes from forming g groups ($g < n$) out of n objects.

- Most appropriate for categorical rather than continuous data (But used extensively for species data: P / A or abundance).

- Cluster analysis produces clusters whether or not "real" groupings exist, and results depend on both the similarity measure chosen and the algorithm used for clustering.

(BEWARE: Clustering less efficient than principal components or discriminant function analyses, when data cross-correlated).

Summary – What to Report

Distance Measure

Linkage Method

} **Ensure they are compatible**

Show your dendogram (include a graph)

If dendogram re-scaled, explain what method used

If groups defined, explain rule for “prunning” the tree

And show the amount of information retained

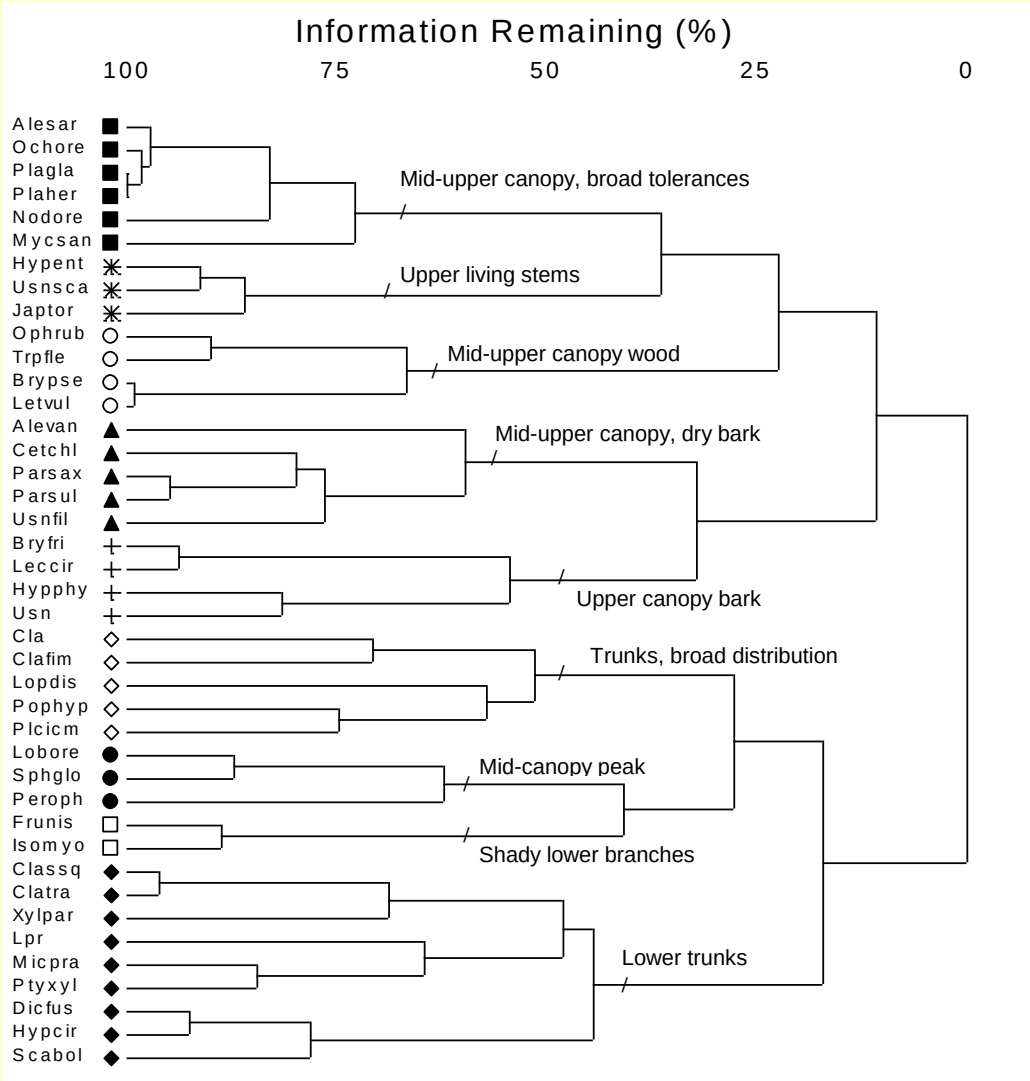
Dendogram Example

Example dendrogram from hierarchical cluster analysis of a species by sample unit matrix.

Symbols indicate species groups formed by pruning the dendrogram ("/" are the cut marks).

Each species group is accompanied by an interpretation of the associated habitat

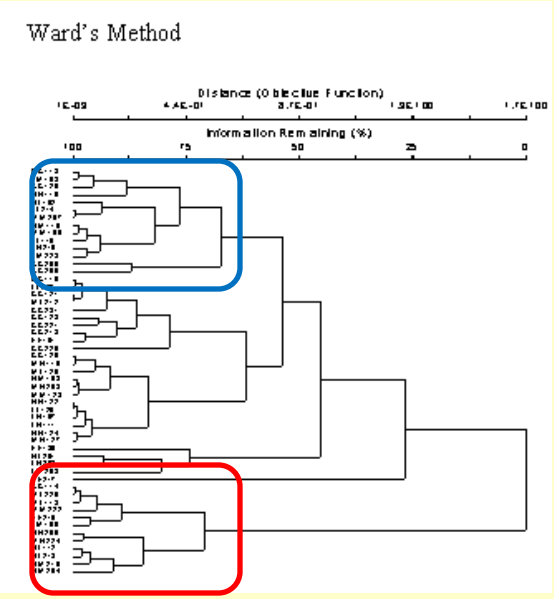
(McCune et al. 2000)

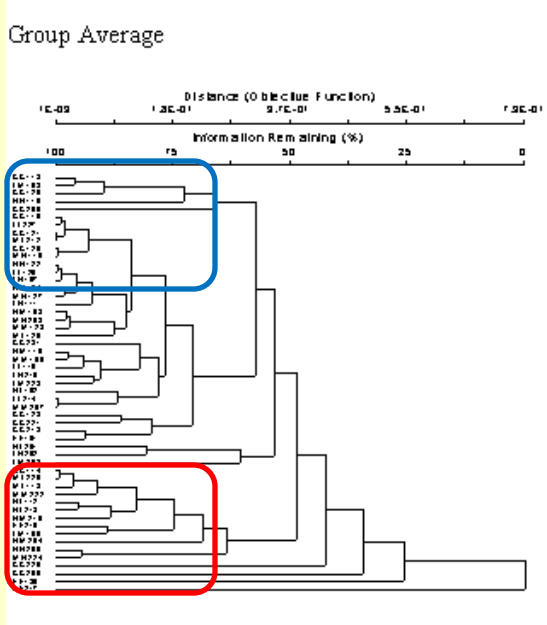


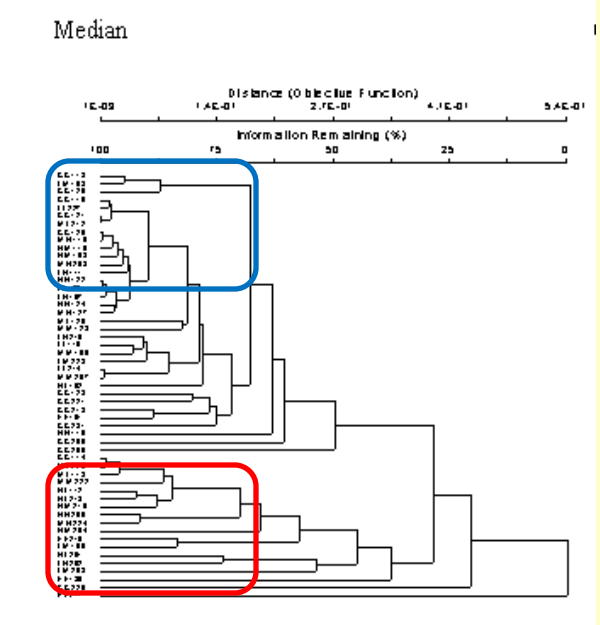
Summary – Recommendations

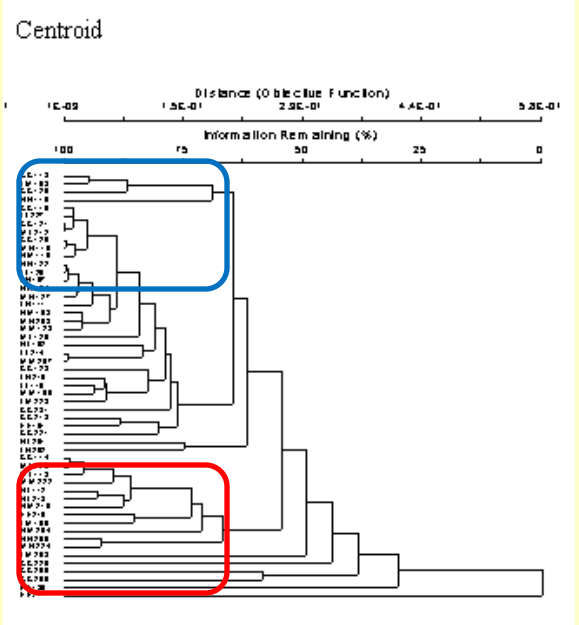
Play with the Output Options:

Especially linkage methods









Summary – Recommendations

Distance Measures for Hierarchical Clustering:

Euclidean (for continuous data)

Sorensen (for categorical data)

Linking Method for Hierarchical Clustering:

Group Linkage Method (for both)

Ward Linkage Method (Euclidean)

TWINSPAN
Two-Way Indicator Species Analysis

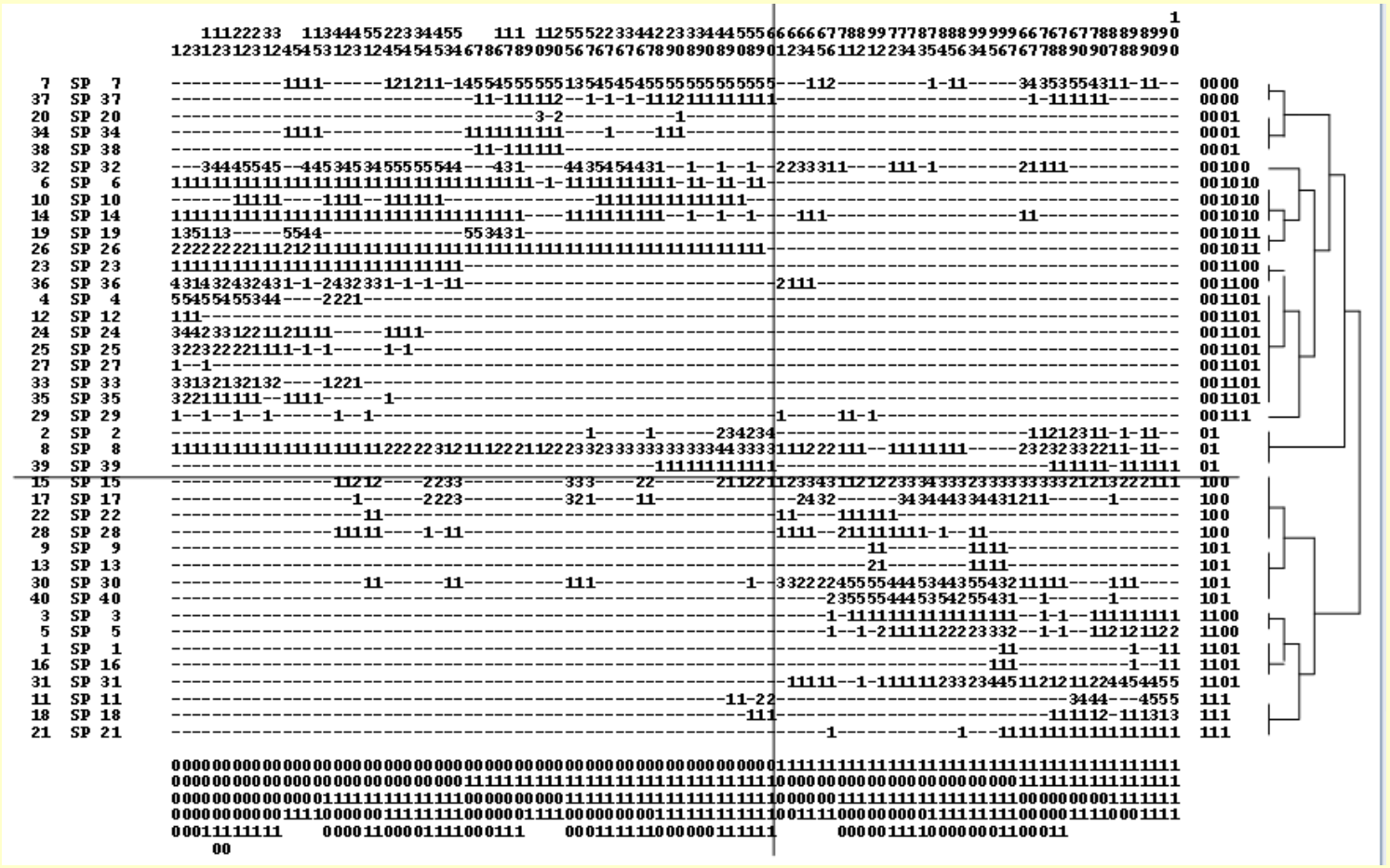
- *Objectives:*

Present this method

Discuss its limitations

Disclaimer: *This approach has "fallen out of favor" because of its limitations and interpretation problems.*

The two-way ordered table from TWINSPAN



TWINSPAN – Pros and Cons

Pros	Cons
Conceptual appeal of two-way ordered table (samples and species at once)	Two-way table effectively displays only 1-D pattern Performs poorly with large heterogeneous data sets Underlying method requires chi-square distance "Pseudospecies" needed to make method semi-quantitative Algorithm complex and difficult to communicate

NOTE: What does it mean to have a pseudo-species"?

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
Chi-square	$x \geq 0$	$d \geq 0$	Euclidean but doubly weighted by variable and sample unit totals; metric

Using Two-Way Dendogram in PC-Ord

Purpose of two-way clustering (known as biclustering) is to graphically illustrate relationship between cluster analyses and your individual data points.

The resulting graph shows graphically how groups of rows and columns relate to each other.

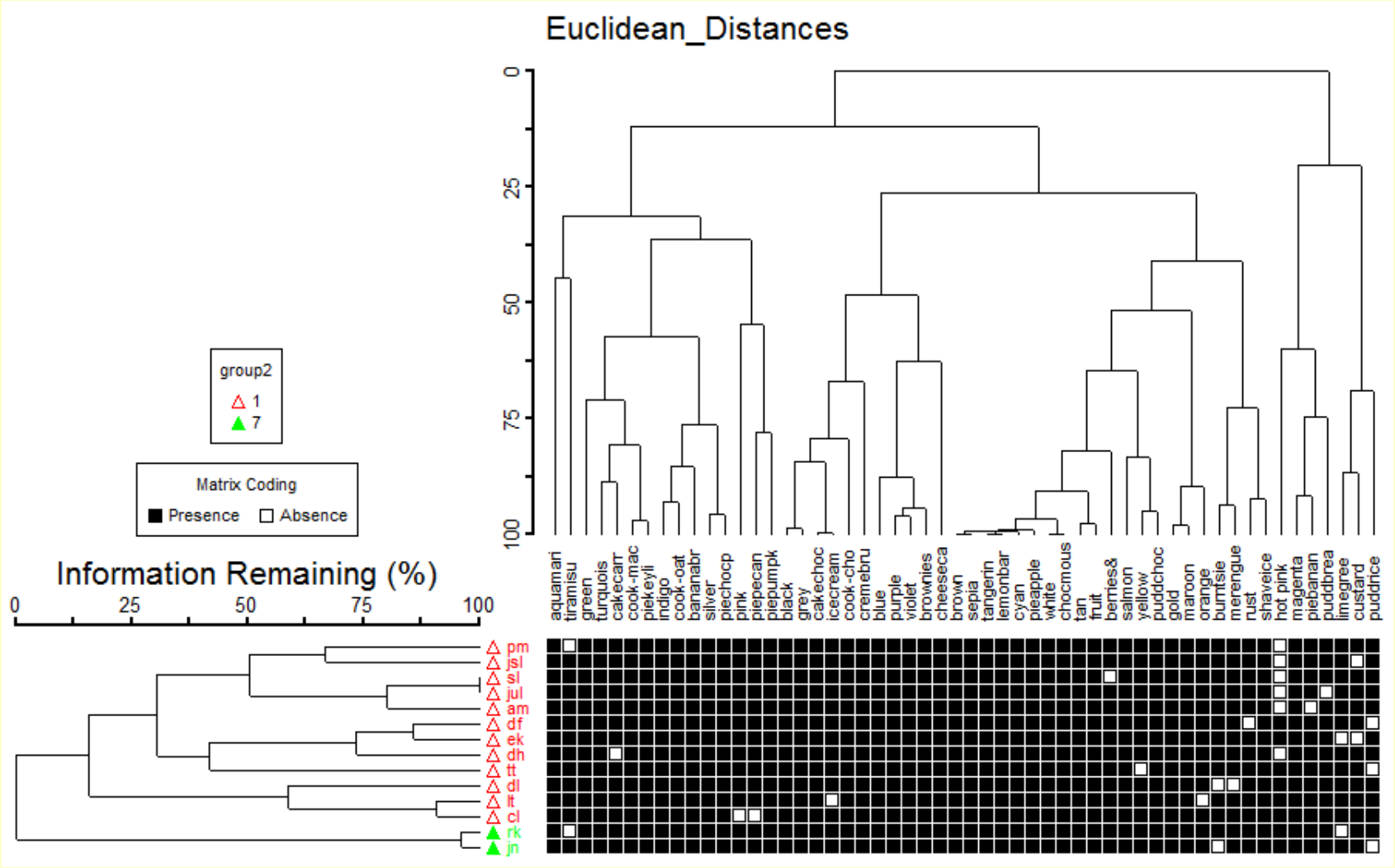
A dendrogram from two-way cluster analyses has 3 elements:

a dendrogram for the rows

a dendrogram for columns

and a graphical representation of the main matrix, re-ordered according to the order in the dendrograms.

Using Two-Way Dendogram in PC-Ord



Use Two-Way Dendogram in PC-Ord

Setup for two-way clustering very similar to regular clustering.

One important difference is that the user can use column relativization. Applied ONLY to clustering of columns.

The clustering of rows is performed first, using the matrix in whatever condition it is showing in main matrix window.

Unless you have good reason for another relativization, relativize by column (species) maximum for two-way clustering of community data sets.

However: still difficult to interpret what groups of species mean

Thus, recommend only using 1-d clustering for samples

Homework #2 – Readings

- Critically reading of journal articles

ELSEVIER

Fisheries Research 31 (1997) 147–158

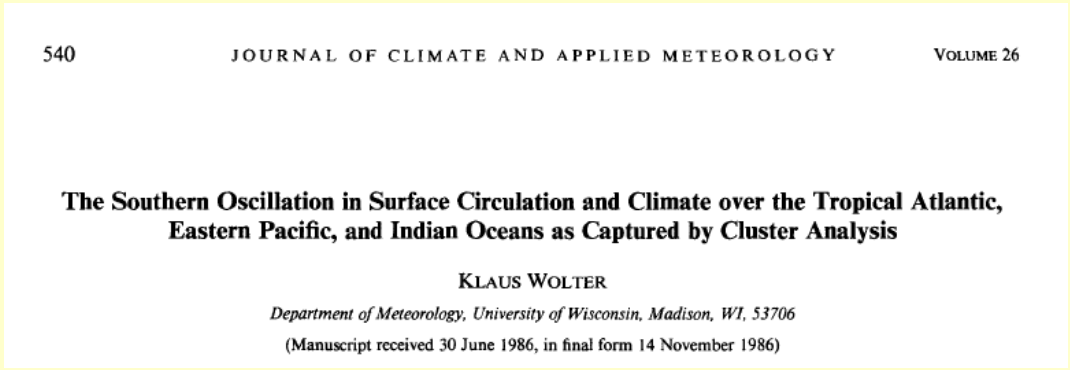
FISHERIES RESEARCH

Cluster analysis of longline sets and fishing strategies within the Hawaii-based fishery

Xi He [a,*], Keith A. Bigelow [a], Christofer H. Boggs [b]

[a] *Pelagic Fisheries Research Program, Joint Institute for Marine and Atmospheric Research, School of Ocean and Earth Science and Technology, University of Hawaii, 2570 Dole Street, Honolulu, HI 96822, USA*

[b] *Honolulu Laboratory, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2570 Dole Street, Honolulu, HI 96822, USA*

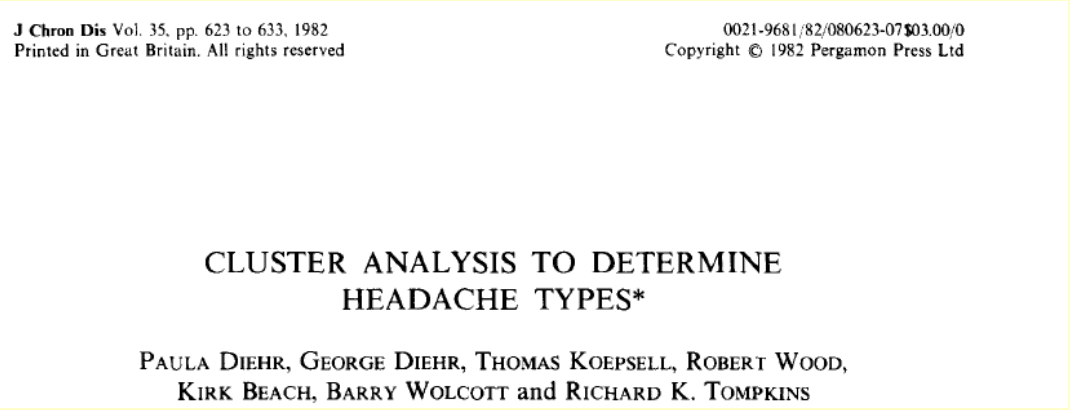


The Southern Oscillation in Surface Circulation and Climate over the Tropical Atlantic, Eastern Pacific, and Indian Oceans as Captured by Cluster Analysis

KLAUS WOLTER

Department of Meteorology, University of Wisconsin, Madison, WI, 53706

(Manuscript received 30 June 1986, in final form 14 November 1986)



J Chron Dis Vol. 35, pp. 623 to 633, 1982
Printed in Great Britain.

0021-9681/82/080623-07$03.00/0

CLUSTER ANALYSIS TO DETERMINE HEADACHE TYPES*

PAULA DIEHR, GEORGE DIEHR, THOMAS KOEPSELL, ROBERT WOOD, KIRK BEACH, BARRY WOLCOTT and RICHARD K. TOMPKINS

Clustering Readings (He et al. 1997)

ELSEVIER

Fisheries Research 31 (1997) 147–158

FISHERIES RESEARCH

Cluster analysis of longline sets and fishing strategies within the Hawaii-based fishery

Xi He [a,*], Keith A. Bigelow [a], Christofer H. Boggs [b]

[a] *Pelagic Fisheries Research Program, Joint Institute for Marine and Atmospheric Research, School of Ocean and Earth Science and Technology, University of Hawaii, 2570 Dole Street, Honolulu, HI 96822, USA*

[b] *Honolulu Laboratory, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 2570 Dole Street, Honolulu, HI 96822, USA*

Accepted 16 September 1996

1. Rationale & Objective: What was the purpose of this paper?

Clustering Readings (He et al. 1997)

2. Methods:

What data went into their primary matrix?

What data went into their secondary matrix?

3. Results:

How many clusters did they detect?

What other fishing strategies were operating in the fishery, in addition to the "tuna" and "swordfish" sets?

Clustering Readings (He et al. 1997)

Methods:

What data went into primary matrix? *Species catches*

What data went into secondary matrix? *Gear / Practices*

Table 2
Mean percentages of catches for eight species and three species groups within five clusters of sets from the Hawaii-based longline fishery (1991–1994)

Species group	Cluster				
	1	2	3	4	5
Albacore	2.6	9.4	1.4	6.0	4.5
Bigeye tuna	36.1	24.0	15.4	5.0	7.8
Blue marlin	2.5	3.8	2.2	1.0	2.7
Blue shark	7.4	9.3	11.1	45.1	0.3
Mahimahi	3.9	9.1	43.3	5.2	12.4
Other billfishes	2.9	4.9	3.2	0.6	2.0
Other fishes	5.3	17.9	4.0	1.2	2.2
Other sharks	3.3	2.8	0.9	1.0	0.8
Striped marlin	5.4	12.6	6.3	1.8	4.2
Swordfish	15.2	1.0	9.0	31.1	59.2
Yellowfin tuna	15.0	4.9	3.0	1.9	3.9

Table 3
Characteristics of fishing operations for five clusters of sets from the Hawaii-based longline fishery (1991–1994). When two numbers are listed within a column they are the mean and standard deviation (in parentheses). Set duration is the interval between the start of deployment and the start of retrieval, not the entire duration of the fishing operation

	Cluster				
	1	2	3	4	5
Primary catch	Tuna–Swordfish	Tuna	Mahimahi–Tuna	Shark–Swordfish	Swordfish
No. of sets	5790	14573	4077	15630	6688
Vessel length (m)	65 (10.7)	60 (10.1)	69 (11.0)	75 (9.8)	71 (9.3)
Hooks per set	1051 (343)	1345 (331)	975 (279)	847 (283)	852 (176)
Lightsticks per set	174 (210)	21 (96)	221 (243)	477 (283)	359 (227)
Duration of set (h)	11.4 (2.9)	9.7 (2.9)	11.8 (2.4)	12.4 (1.6)	12.4 (2.2)
% time of fishing					
Day	37.0	84.6	24.5	3.9	6.5
Day–night	3.0	2.7	2.2	1.2	2.7
Night	60.0	12.7	73.3	94.9	90.8
% sets by lunar phase					
New	23.5	32.5	20.6	25.3	23.7
1st and 3rd quarter	33.6	34.1	32.0	35.7	35.7
Full	42.9	33.4	47.4	39.0	40.6
% sets within MHI EEZ	58.9	68.6	48.6	9.0	24.4

Clustering Readings (He et al. 1997)

Results:

How many clusters did they detect? *5 clusters*

What other fishing strategies were operating in the fishery, in addition to the "tuna" and "swordfish" sets?

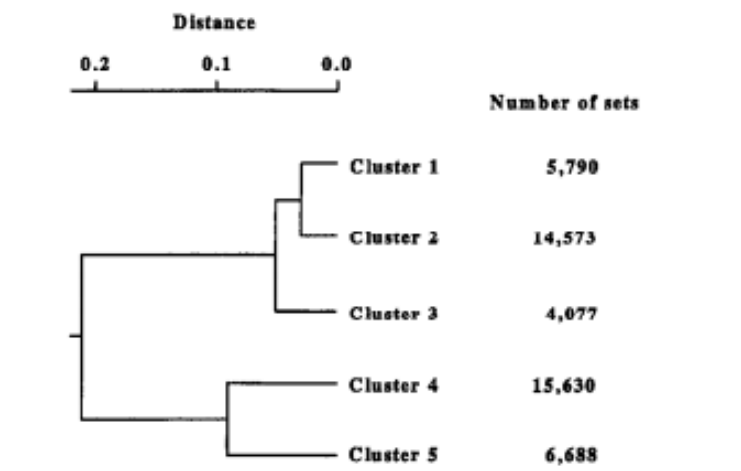


Fig. 1. Dendrogram of five clusters of longline sets from the Hawaii-based longline fishery showing Euclidian distance between clusters.

Clusters 1 and 3 reflect 'mixed' fishing strategies because they caught substantial proportions of both tuna and swordfish