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)

Clustering - Objectives

- Objectives and Limitations (James & McCulloch 1990)

Objectives:

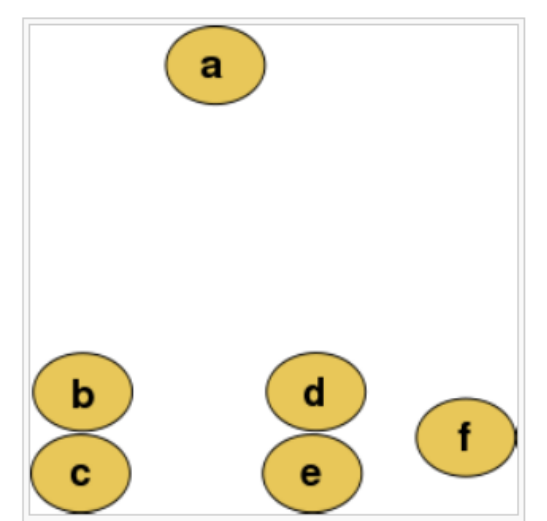
1. To classify groups of objects judged to be similar according to distance or similarity measure
2. To reduce consideration of n objects to g (g less than n) group of objects

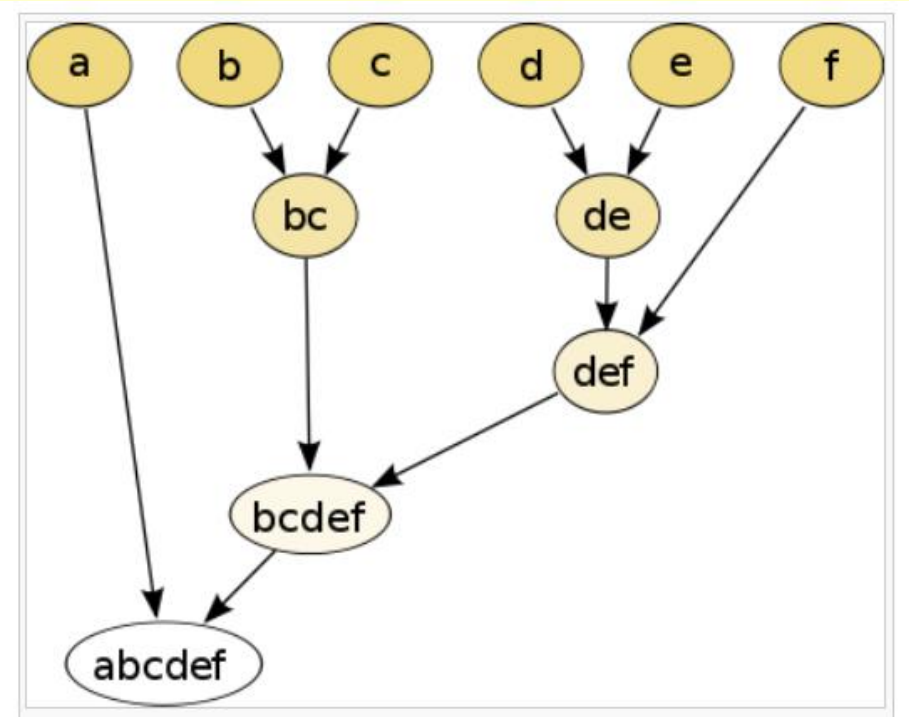
Limitations:

1. Results depend on the distance measure chosen.
2. Results depend on the algorithm chosen for forming clusters

Clustering - Introduction

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





Clustering - Introduction

1. Start with pairwise similarity matrix among objects (individuals, sites, populations, taxa).

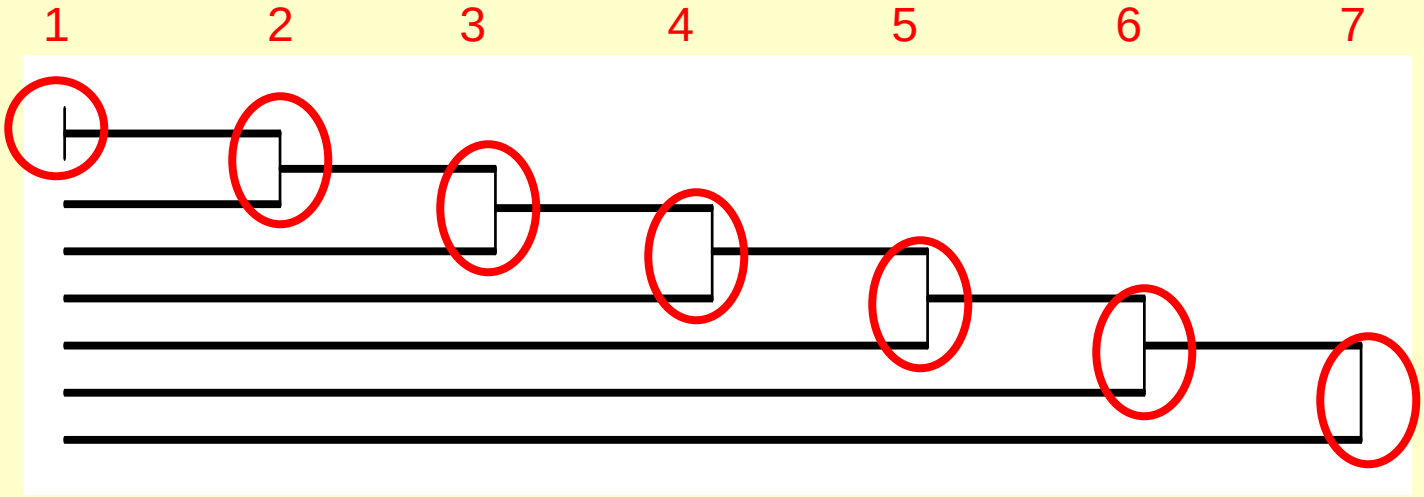
2. Two most similar objects are joined into a group, and the similarities of this group to all other units are calculated.

3. Repeatedly the two closest groups are combined until only a single group remains.

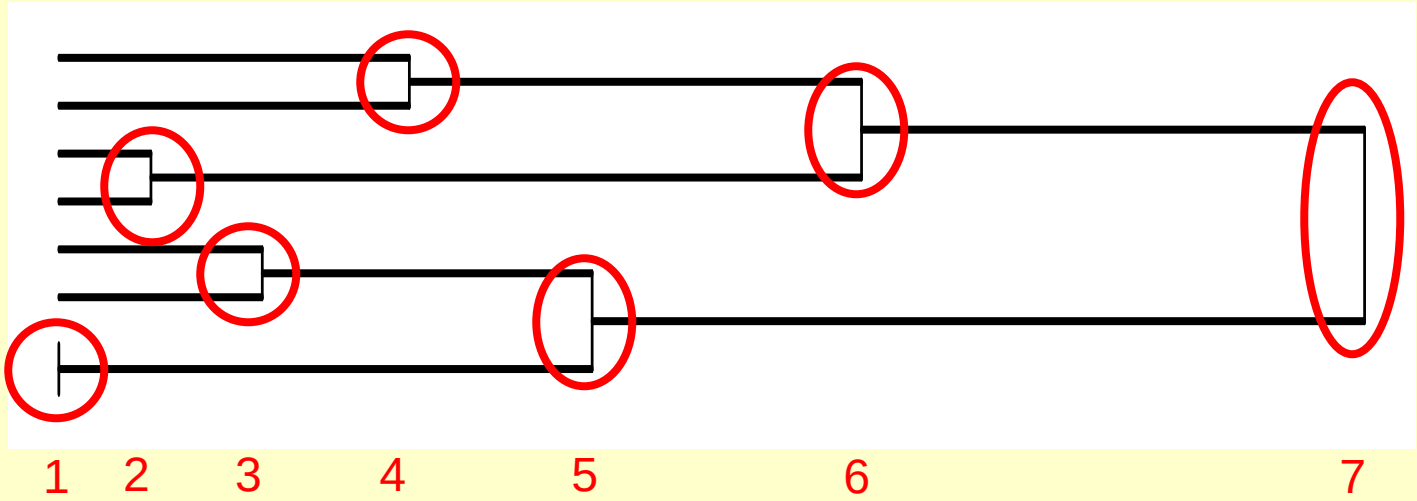
4. Results usually expressed in the form of a dendrogram, a two-dimensional hierarchical tree diagram representing the complex multi-variate relationships among the objects.

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

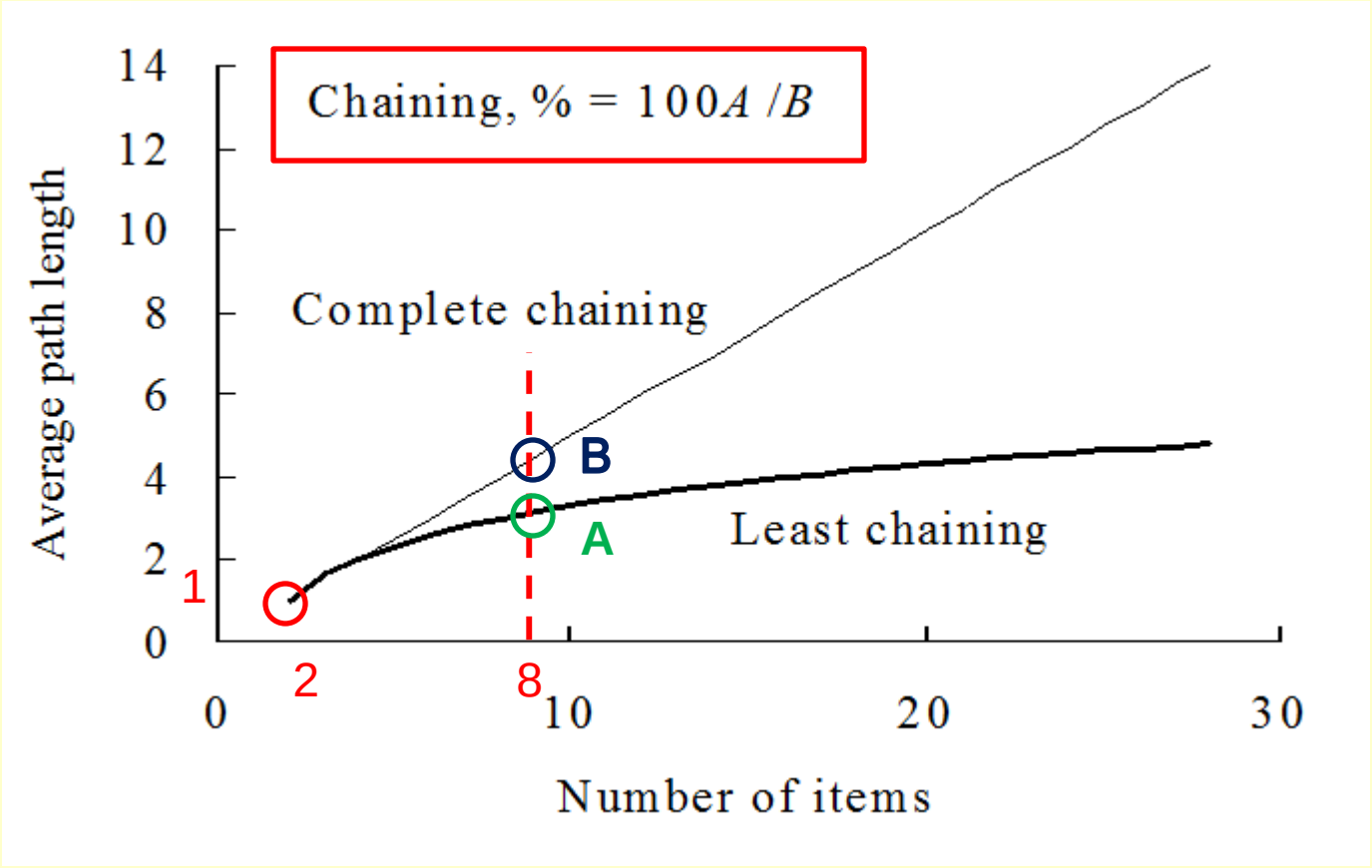


Complete chaining (1 group)



No chaining (2 groups)

Clustering - Introduction



Example with 2 items?

- Two paths
- One node in each
- Avg. Path Length = 1

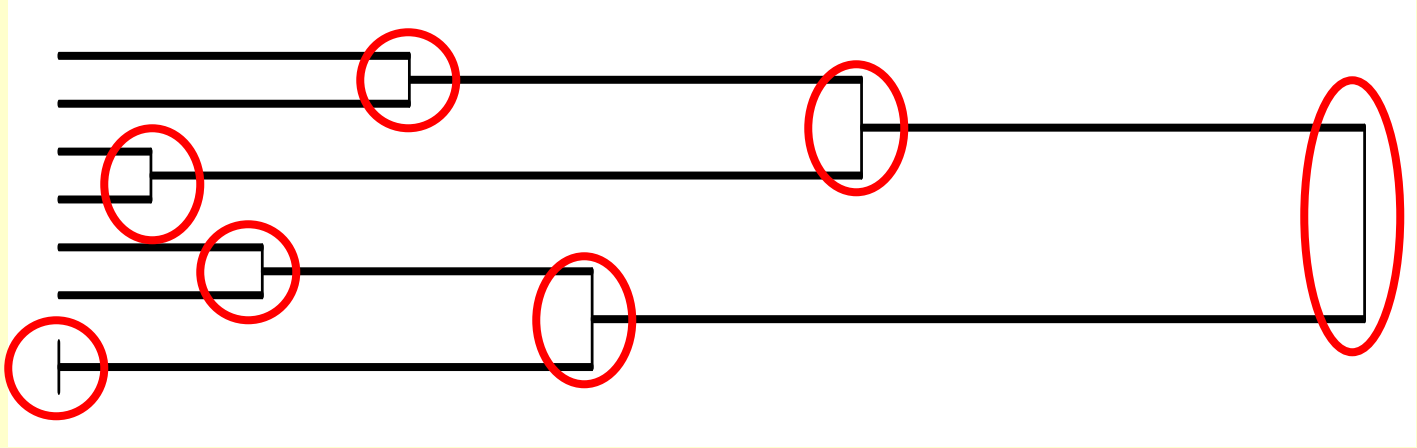
Average path length used to measure percent chaining in cluster analysis. Path length is the number of nodes between tip of a branch and trunk.

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

A) No chaining:

1 2 3 4 5 6 7



Number of paths = 8

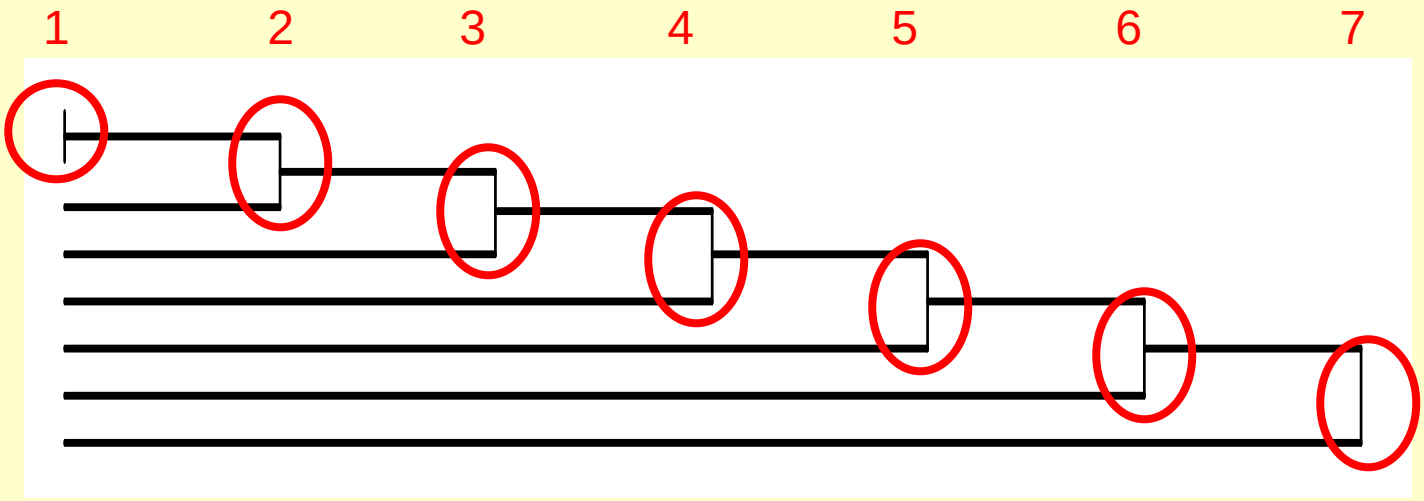
Sum of nodes = 3 3 3 3 3 3 3 3 = 24

Avg. path length = 24 / 8 = 3.00

Clustering - Introduction

Two ways to sort eight samples (multiple species) into groups

B) Complete chaining:

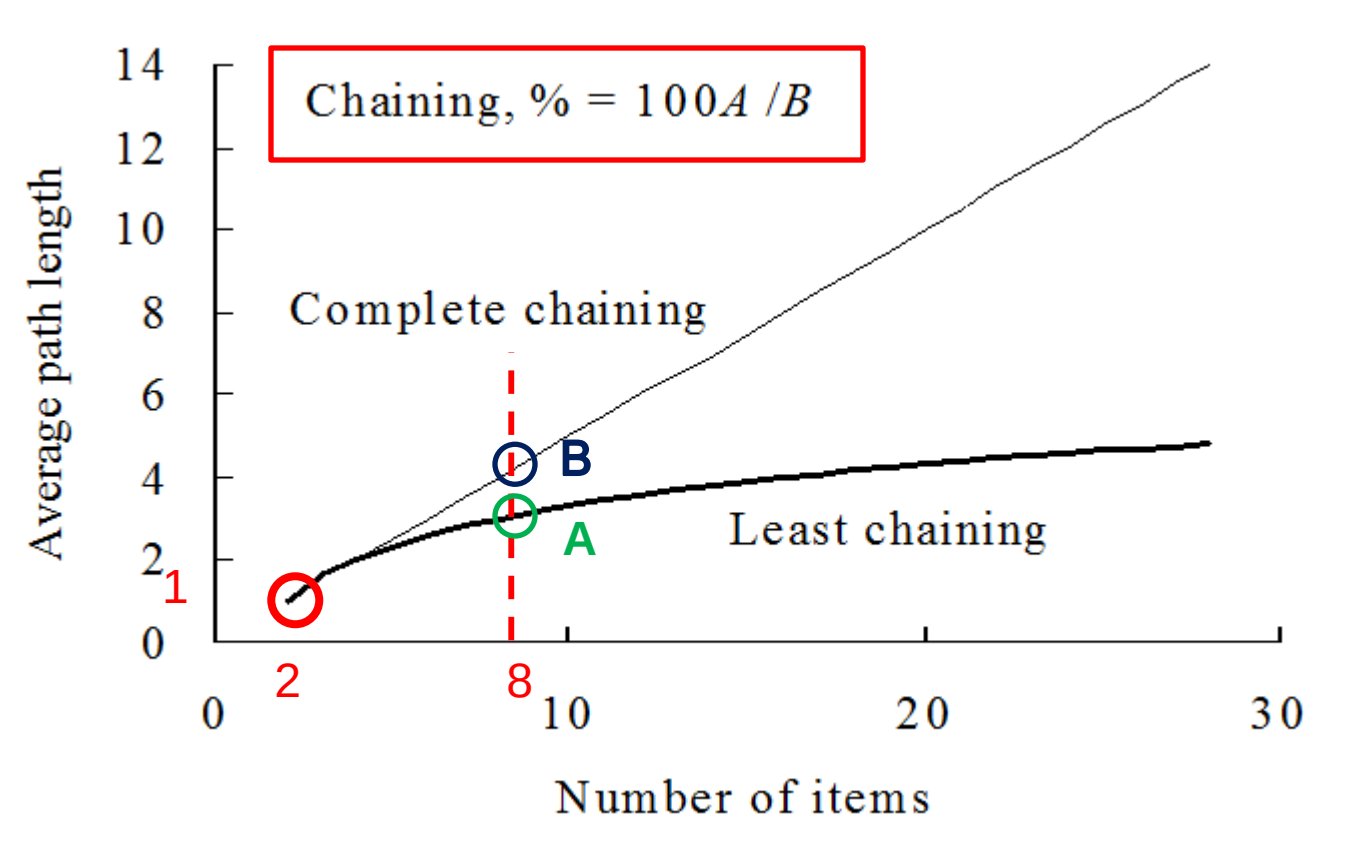


Number of paths = 8

Sum of nodes = 7 7 6 5 4 3 2 1 = 35

Avg. path length = 35 / 8 = 4.375

Clustering - Introduction



NOTE: Chaining can be calculated for any given clustering pattern

Chaining (A) = 100 * (A / complete-chain) = 100 * (3 / 4.375) = 68.57 %

Chaining (B) = 100 * (B / complete-chain) = 100 * (4.375 / 4.375) = 100%

Clustering – How it Works

A dissimilarity matrix of order $n \times n$ (n = number of entities) is calculated and each of the elements is squared. The algorithm then performs n-1 loops (clustering cycles) in which the following steps are done:

1. The smallest element (d_{pq}^2) in dissimilarity matrix sought (groups associated with this element are S_p and S_q).

2. The objective function E_n (the amount of information lost by linking to cycle n) is incremented according to the rule.

3. Group S_p is replaced by $S_p \cup S_q$. Groups S_q and S_q are inactive; their elements assigned to new group $S_p \cup S_q$.

4. The pair-wise distances between the new group ($S_p \cup S_q$) and all other groups are calculated.

Clustering – How it Works

The **objective function (E)** is the sum of the error sum of squares from each centroid to the items in that group.

Where :

t indexes the T clusters

E_t is the error sum of squares for cluster t

$$E = \sum_{t=1}^{T} E_t$$

And each E_t is found by:

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

x_{ijt} is the value of the:

jth variable for the

ith point of cluster t

(which contains k_t points)

$\bar{x}$ is the mean of the jth variable for cluster t.

Clustering – How it Works

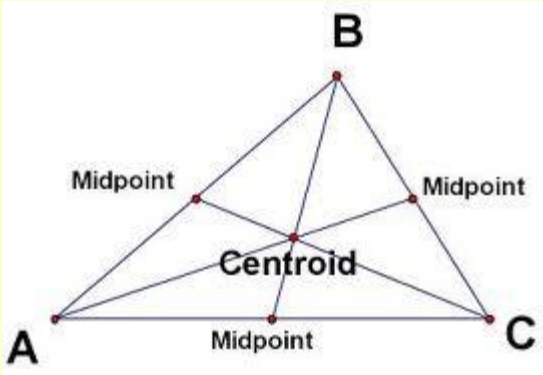
Calculate E for each cluster, separately and sum up (Note: there are T clusters):

$$E = \sum_{t=1}^{T} E_t$$

Calculate E by summing the deviations between all points and centroid, for all variables:

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

What is $\bar{X}$?



A cluster of 3 points, plotted in 2 dimensions

Clustering – How it Works

We need a rule to progressively combine the elements, as we go through the cycles and the groups become larger.

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 (See Table in next slide).

There Are Different Linkage Methods

Use different coefficients in the basic combinatorial equation.

	Coefficient			
Linkage method	α_p	α_q	β	γ
Nearest neighbor	0.5	0.5	0	-0.5
Farthest neighbor	0.5	0.5	0	0.5
Median	0.5	0.5	-0.25	0
Group average	n_p / n_r	n_q / n_r	0	0
Centroid	n_p / n_r	n_q / n_r	$- \alpha_p \alpha_p$	0
Ward's method	$\frac{n_i + n_p}{n_i + n_r}$	$\frac{n_i + n_q}{n_i + n_r}$	$\frac{- n_i}{n_i + n_r}$	0
Flexible beta	$(1 - \beta)/2$	$(1 - \beta)/2$	β	0
McQuitty's method	0.5	0.5	0	0

n_p = number of elements in S_p $\quad$ n_q = number of elements in S_q

n_r = number of elements in $S_r = S_p \cup S_q$

n_i = number of elements in S_i $\quad$ $i = 1, n$ $\quad$ except $i \neq p$ and $i \neq q$

Defining Groups (Clusters)

Clusters are defined using two sets of instructions:
distance measures & linkage methods ("sorting strategies")

We consider eight linkage methods:

Nearest neighbor	Farthest neighbor
Median	Group average
Centroid	Ward's method
Flexible beta	McQuitty's method

We consider two generic classes of distance measures:

Euclidean	(absolute, relative)
Proportional	(Sorensen, Relative Sorensen, Jaccard)

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)		Dissimilarity Measure: Sørensen distance (1 - 2w/a+b)	
	Combinatorial compatible?	Space contracting, expanding, or conserving?	Combinatorial compatible?	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

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

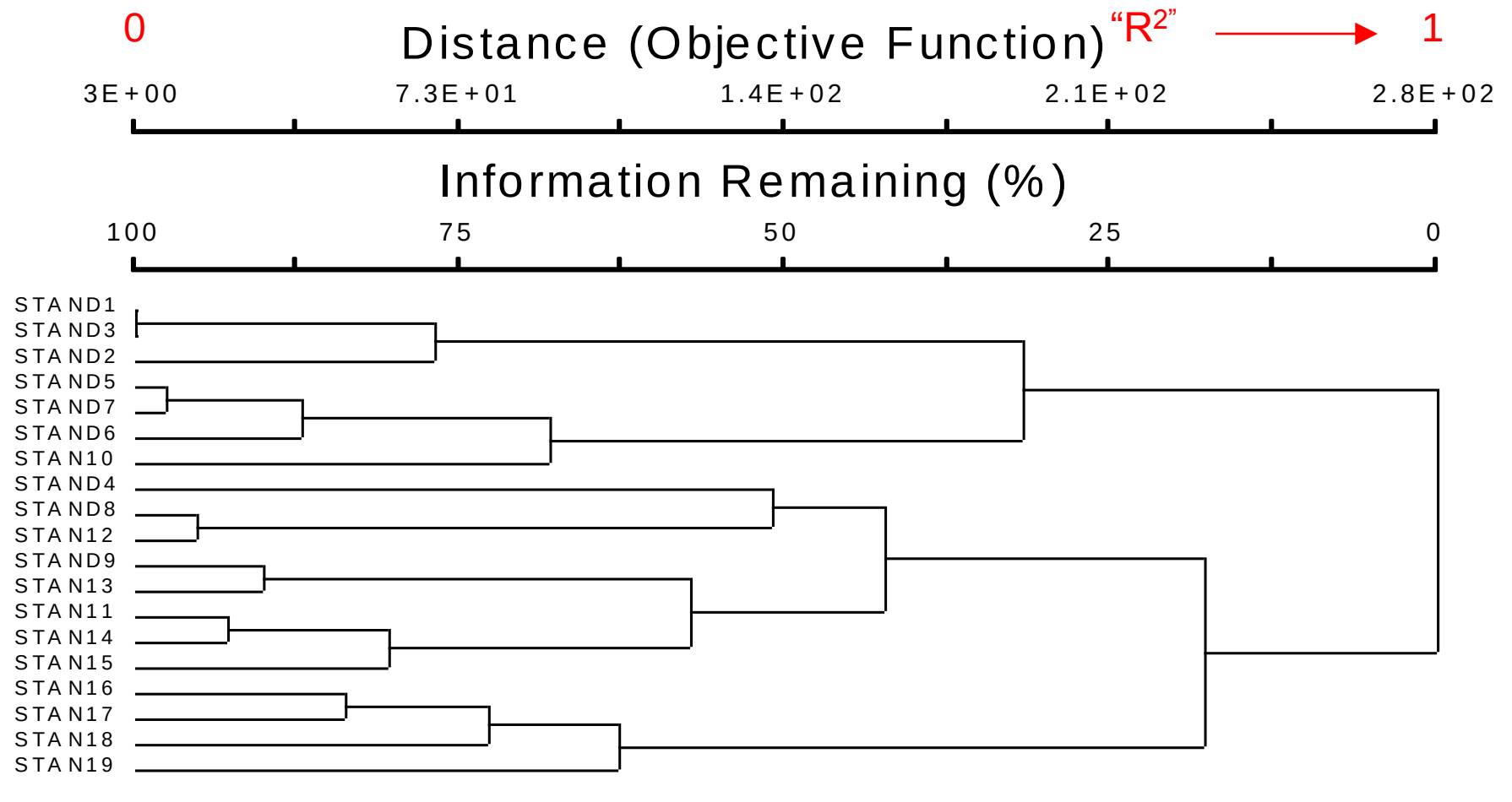
Only one does not distort the relationships in variable space:

Group Average

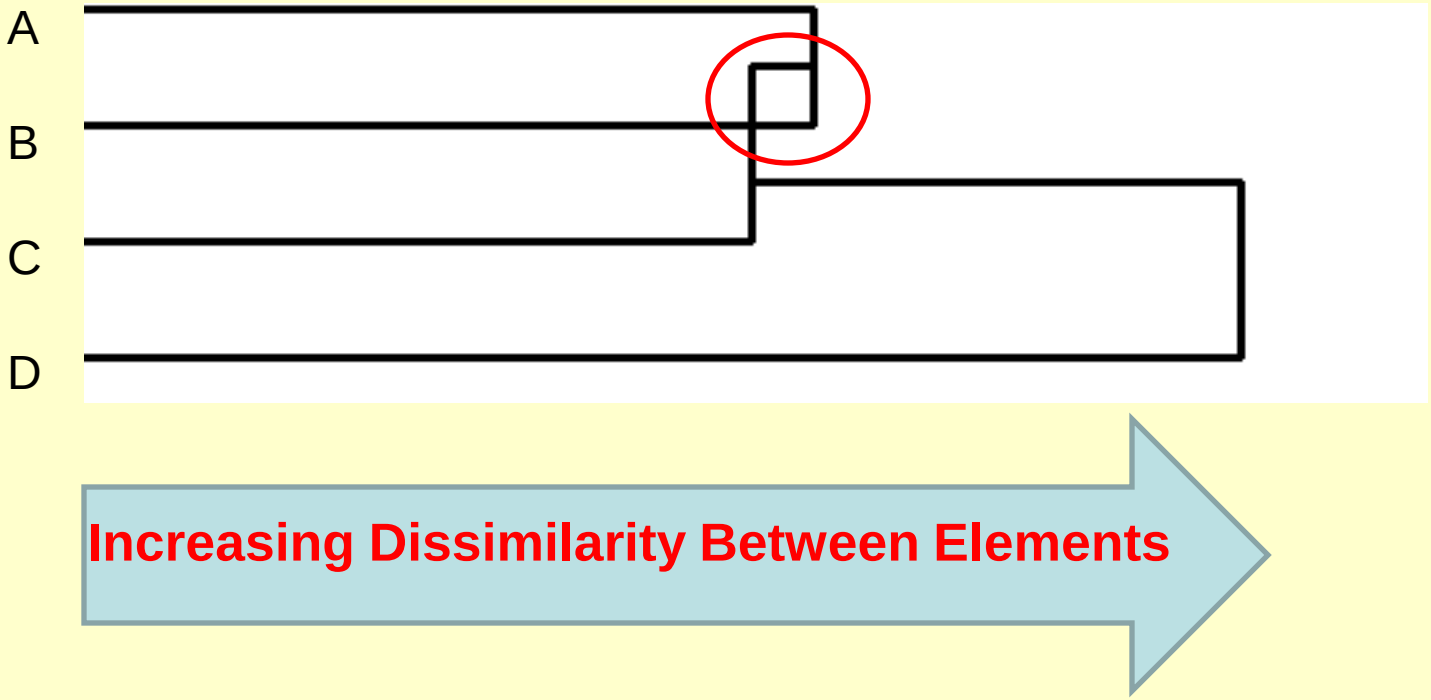
Dendogram Properties I

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

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



Dendogram Properties II

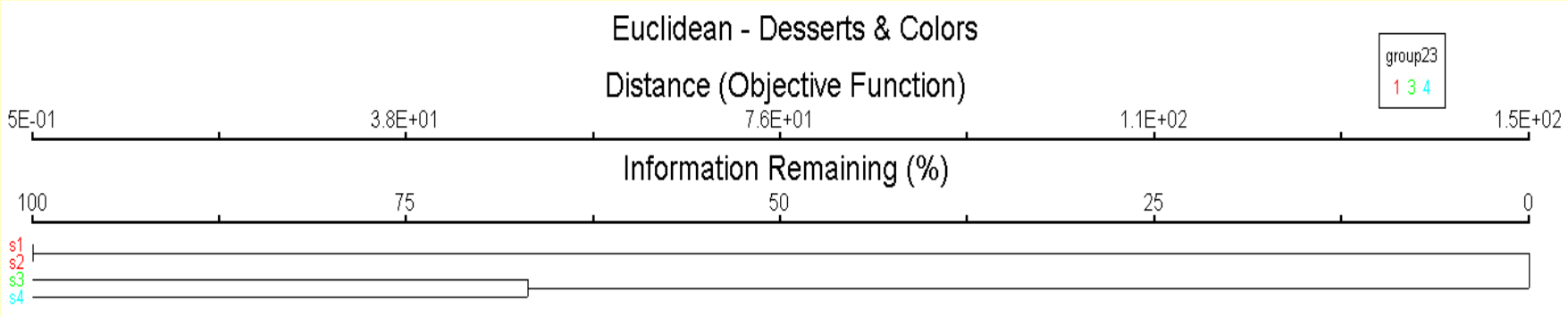


Elements in a dendrogram are always linked according to the “objective function” (more similar elements linked first)

Take Home: Successive Links cannot “decrease” in similarity

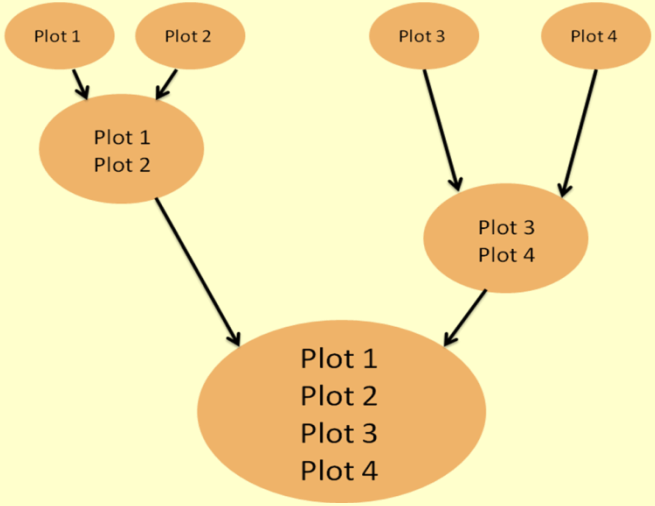
Dendogram Properties III

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



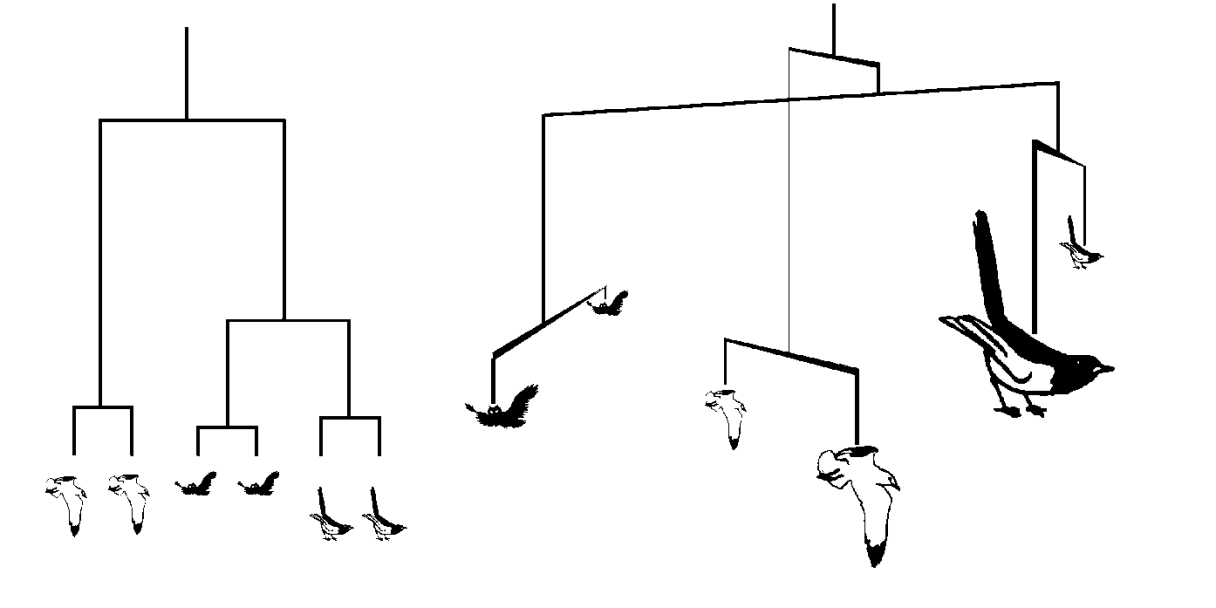
Second - WORK2.WK1

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



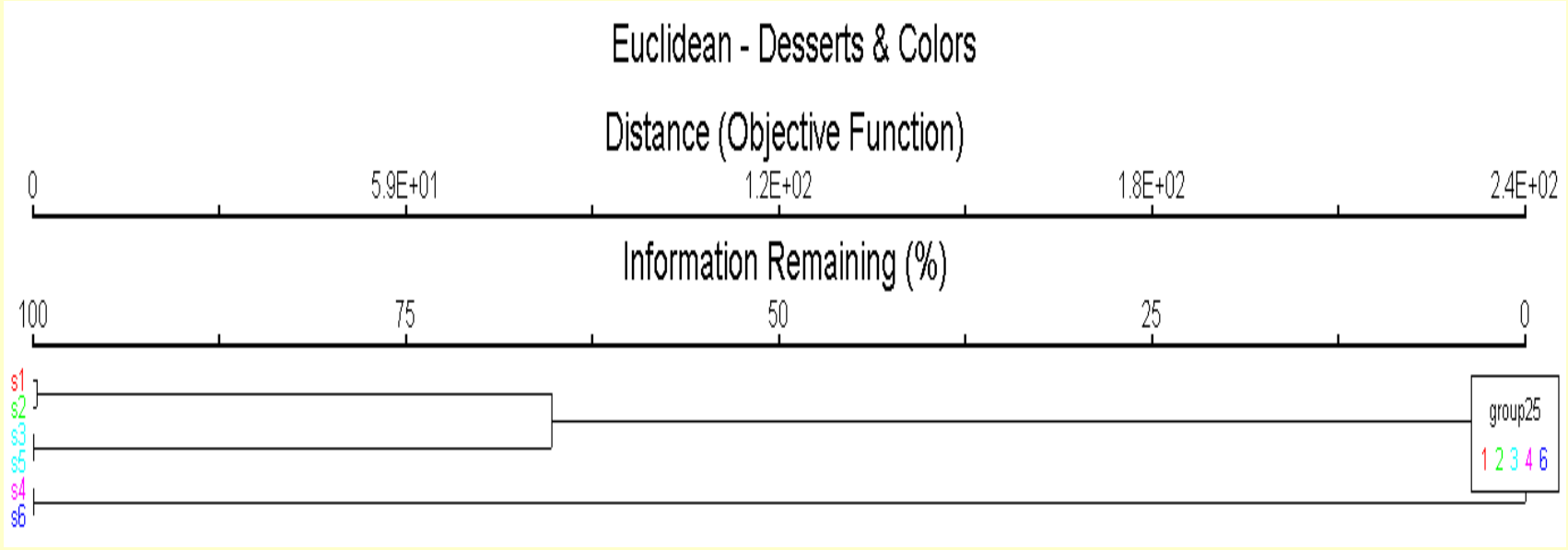
A dendrogram is an inherently nondimensional representation. Imagine the branches as free to pivot, like a child's mobile.

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?

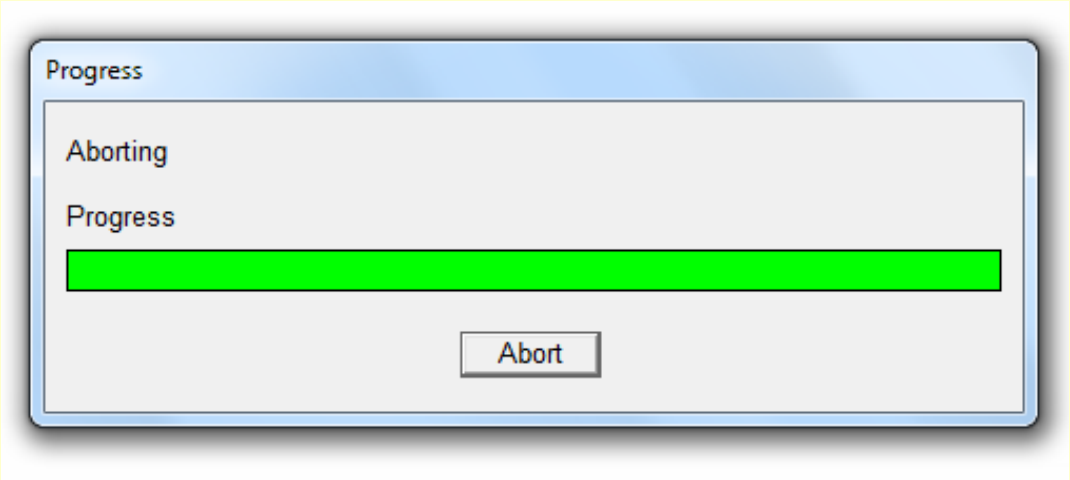


Dendogram Properties VI

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

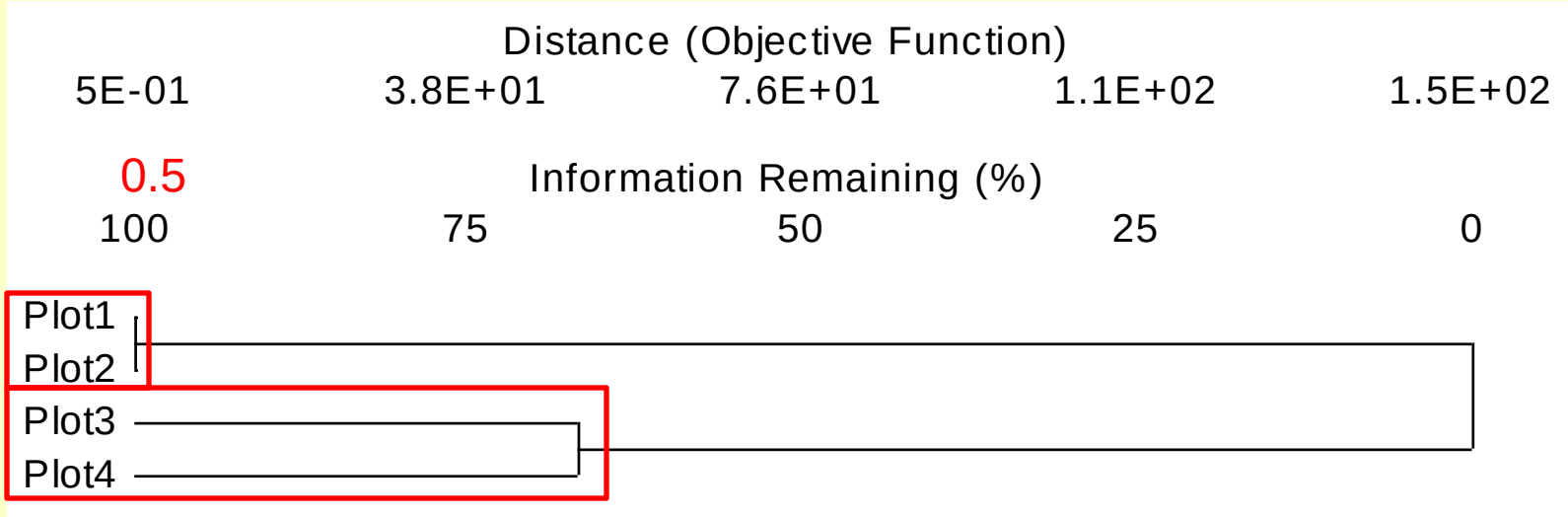
Cluster step 3:

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

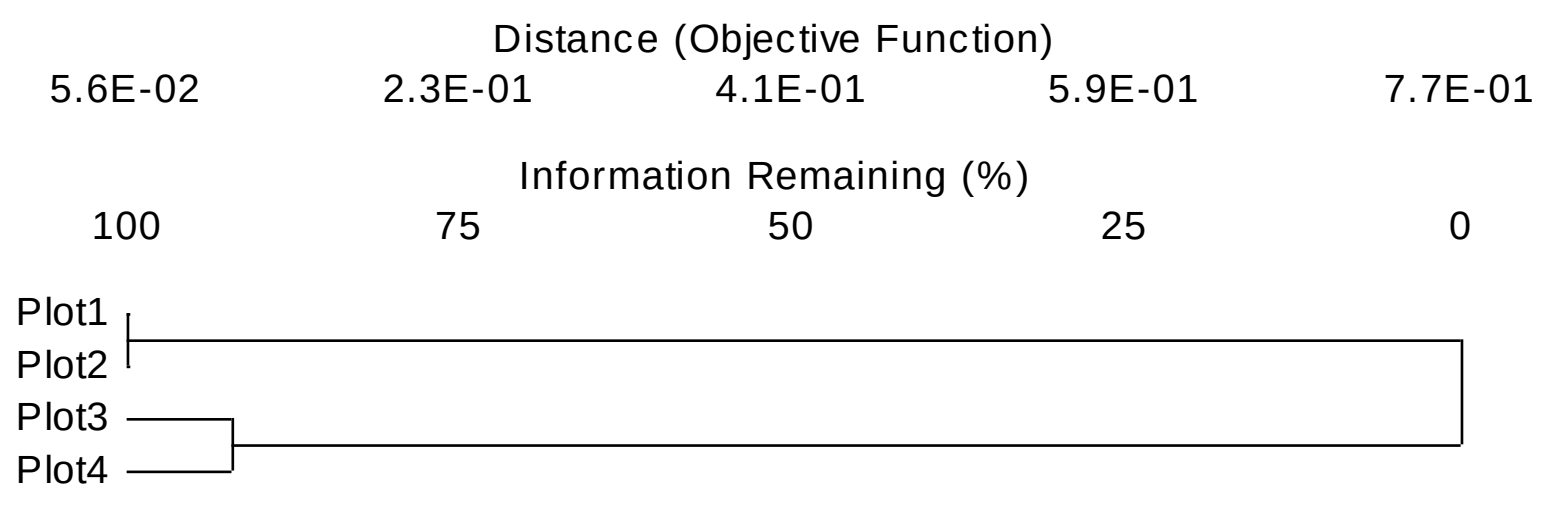
Revised distance matrix after the first fusion.

	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.

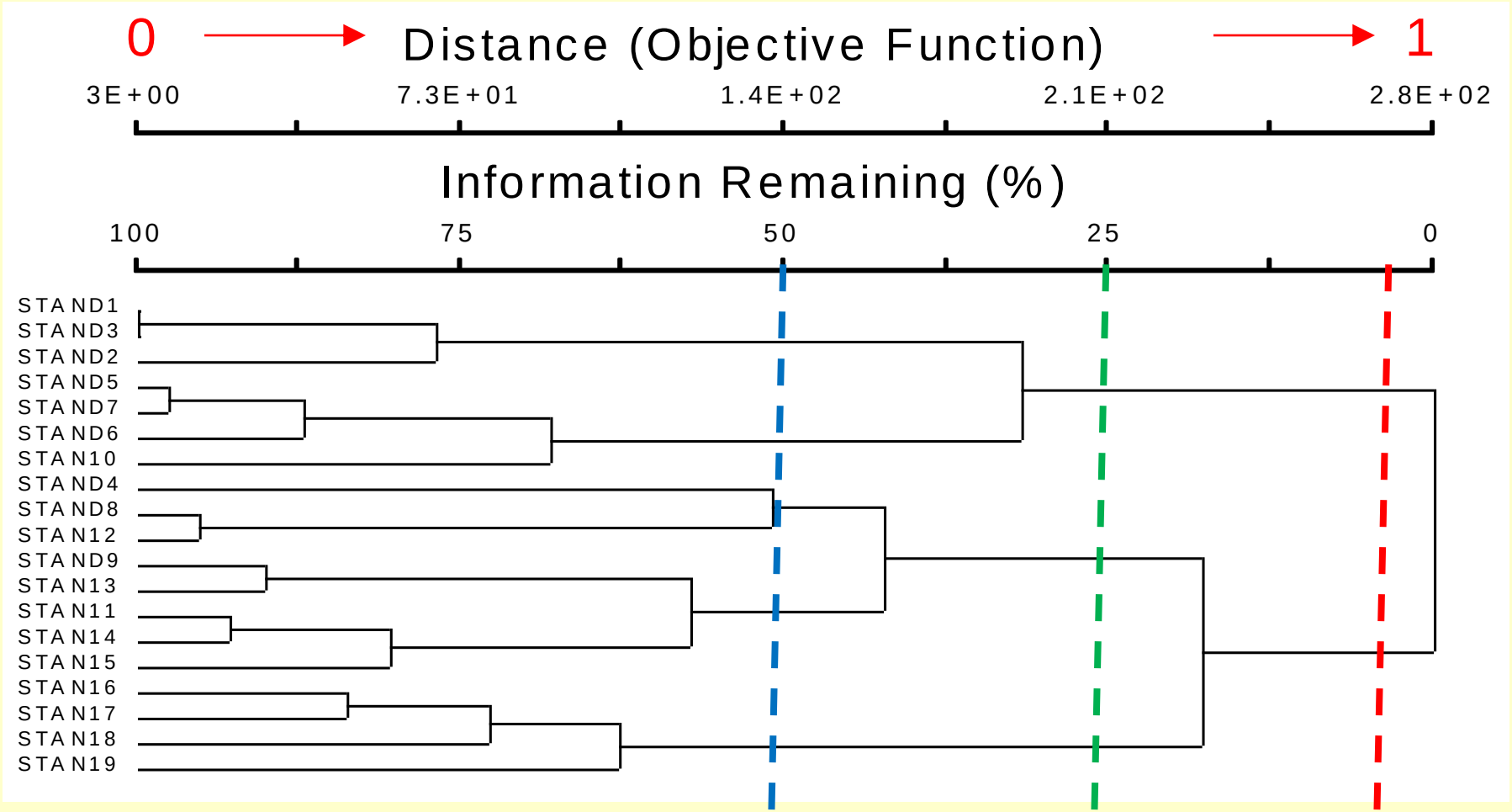


Cluster analysis of plots using Ward's method and Sørensen distance.

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 a similarity measure and 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 (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.

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

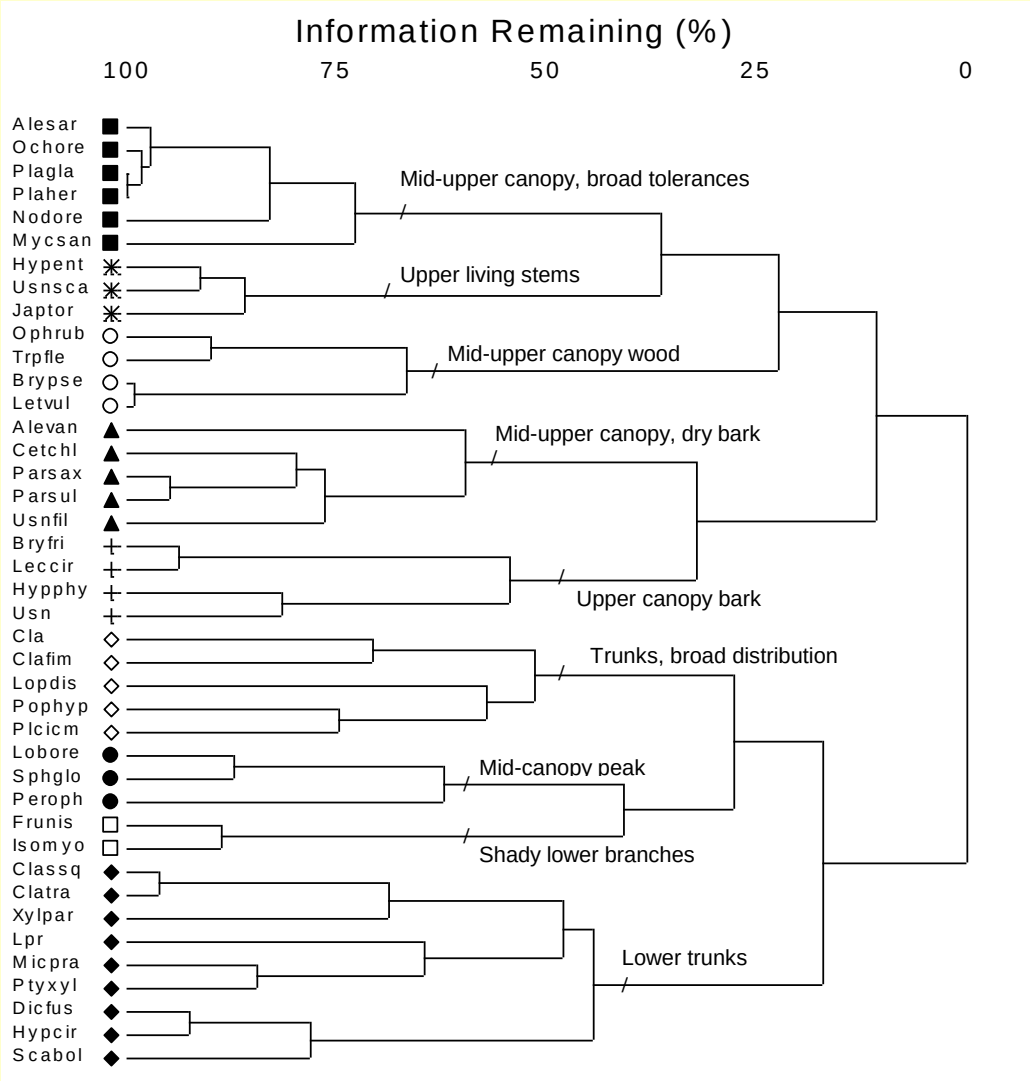
Dendogram Example

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

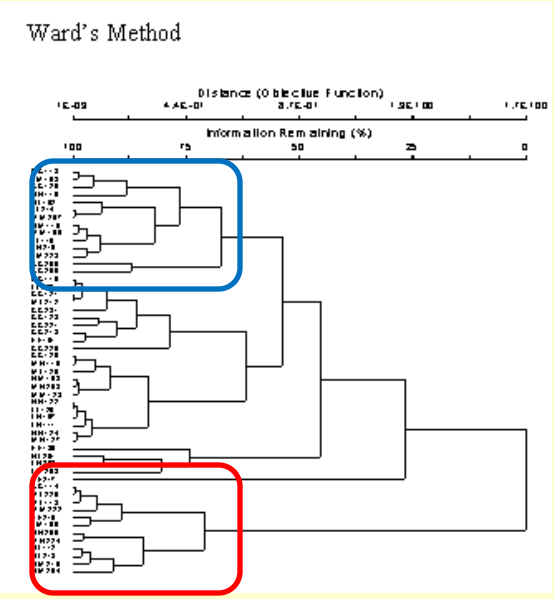
(McCune et al. 2000)

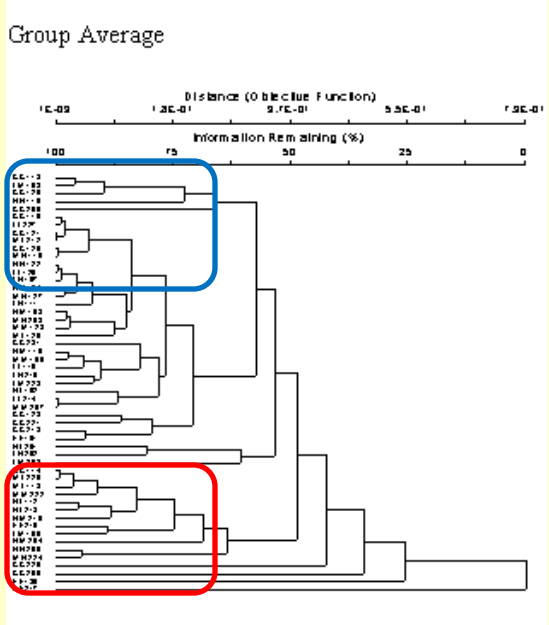


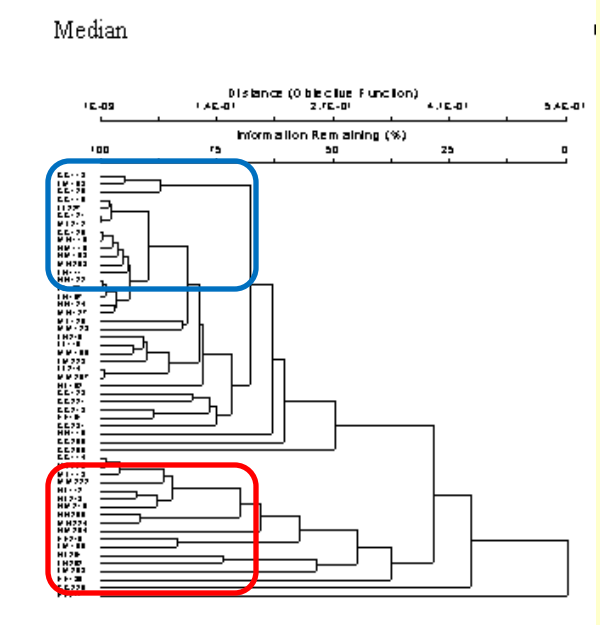
Summary – Recommendations

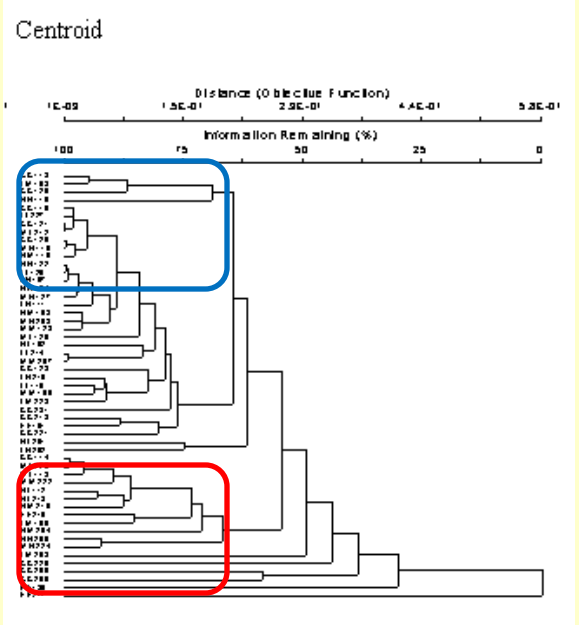
Play with the Output Options:

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)

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

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.*

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

The two-way ordered table from TWINSPAN

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.

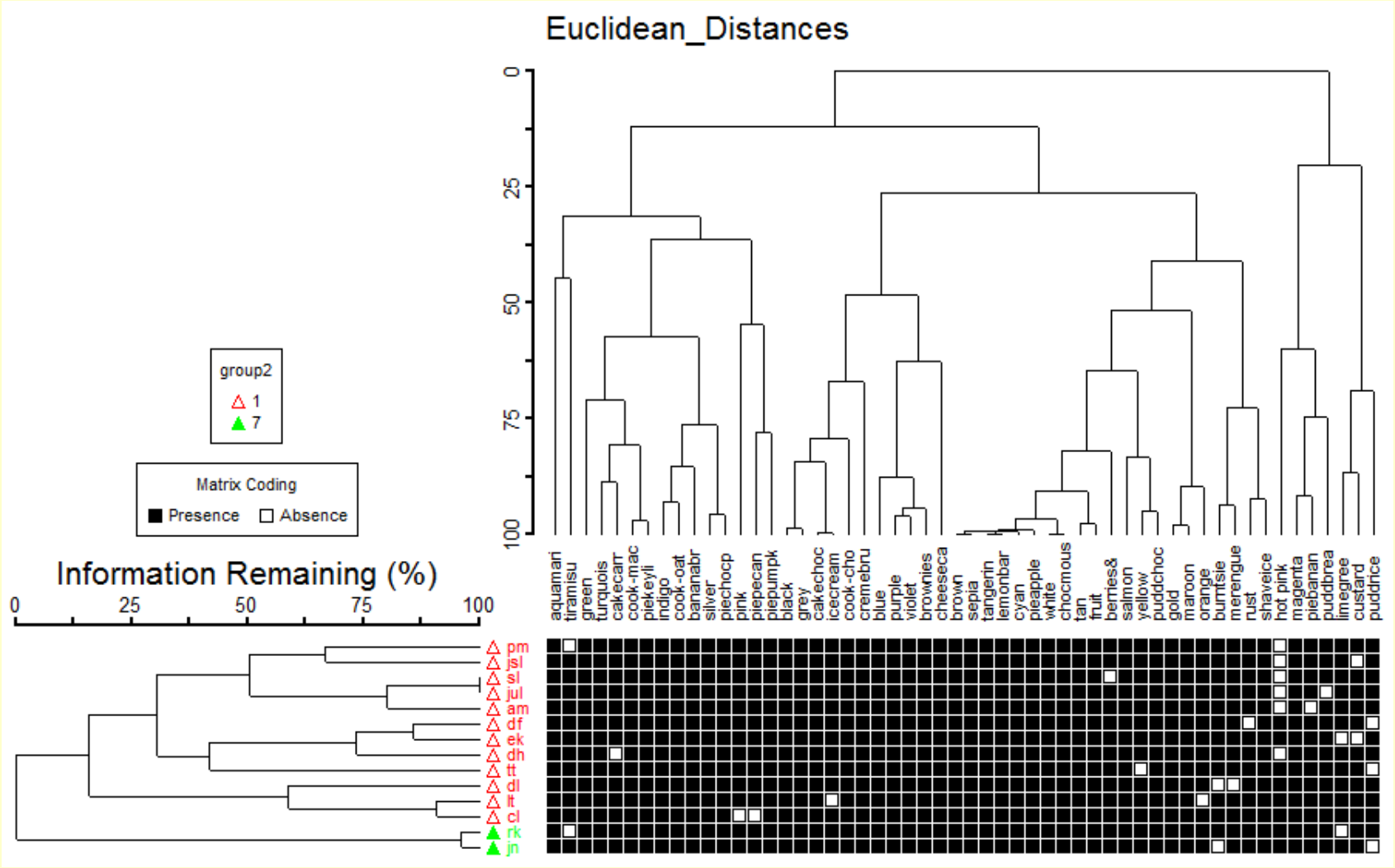
A dendrogram from two-way cluster analyses has 3 elements:

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.

The two-way dendogram from TWINSPAN



However: Difficult to interpret what groups of species mean

Recommend only using 1-d clustering for samples