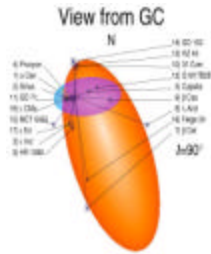


## Cluster Analysis



Introduction & definitions  
The basic model: single linkage clustering  
Cophenetic matrix and ultrametric property  
A diversity of clustering methods  
Hierarchical agglomerative clustering  
Single linkage  
Complete linkage  
Intermediate linkage  
UPGMA & WPGMA  
UPGMC & UPGMC  
Ward's minimum variance  
General agglomerative model  
Flexible clustering  
Information analysis  
Hierarchical divisive clustering  
Monothetic methods  
Polythetic methods  
TWINSPAN  
Partitioning by K-means  
Cluster validation & choice of method  
Exploratory data analysis via graphics

## Cluster Analysis

The collection and organization of objects is almost an innately human trait. It requires the recognition of discontinuous subsets.

Cluster analysis is a process of identification and categorization of subsets of objects that are, more often than not, continuously distributed.

Measures of similarity between objects (Q-mode) or descriptors (R-mode) is the first step of this process.

## Cluster Analysis

Usually the result of clustering in EEB is a *typology* (i.e., a system of types). The primary purpose being to describe the structure of a continuum and identify various *object types*. The "nature" of the types is often immaterial.

Most multivariate statistical software offers some form of cluster analysis. The result of these procedures is usually some form of *dendrogram* (tree) or *skyline plot*.

## Cluster Analysis

BEFORE embarking upon a cluster analysis, you should clearly state what your *goals* are. Moreover, it is important to justify why you believe discontinuities exist in the data and what you hope to gain by identifying and evaluating those discontinuities.

Alternatively, *ordination* stresses the continuous nature of data and is used to emphasize gradients in the environment or community traits.

Cluster analysis and ordination are not mutually exclusive, but one may be more preferable than the other for certain questions.

## Introduction

*Clustering* is an operation of multidimensional analysis which consists of partitioning a set of objects into *subsets*, such that each object or descriptor belongs to one and only one subset for that partition. Thus, subsets of any level, by definition, must consist of *mutually exclusive cells*.

The Polish ecologist Kulczynski (1928) may have been the first EEB to use cluster analysis to group observations. Thus, the method has been used since the early years of EEB.

Most clustering methods proceed from an association matrix.

## Introduction

The choice of clustering methods is just as critical as is the choice of an association measure.

It is important to fully understand the whole range of clustering methods and options in order to correctly evaluate and interpret the structure in your data under different circumstances.

There are two major types of clustering: *descriptive* and *synoptic*. In descriptive clustering one attempts to avoid misclassification at all costs; whereas, in synoptic clustering, the goal is more of a structured conceptual model.

## Single Linkage Clustering

- The Basic Model -

There is an entire classification of clustering methods. For most natural scientists, the simplest and most straightforward clustering method to understand is *single linkage* (also referred to in the literature as *nearest neighbor*).

We will assume at this point that you have carefully chosen a similarity measure and applied it to a primary data matrix to obtain a similarity hemi-matrix.

We will also focus on the classification of objects, but recognize that all of the method can just as easily be applied to descriptors.

## Single Linkage Clustering

- The Basic Model -

The method proceeds in two steps:

First, the association hemi-matrix is rewritten in order of decreasing similarities (or increasing distances, as appropriate), heading the list with the two most similar objects, and proceeding until all pairs are accounted for.

Second, the clusters are formed hierarchically, starting with the most similar objects, and then letting the objects clump in to groups, and then the groups aggregate in to one another, as the similarity criterion is relaxed.

## Single Linkage Clustering

- Example -

Let's return to an earlier example where 38 species of plankton were studied in 5 ponds. The data were recorded on a relative abundance scale from 0 = absent to 5 = abundant.

After computing the similarity coefficient  $S_{20}$  with parameter  $k = 2$ , the symmetric similarity hemi-matrix is used to derive a single linkage clustering.

## Single Linkage Clustering

- Example, Step 1 -

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	-

$S_{20}$	Pairs
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

## Single Linkage Clustering

A *dendrogram* is the most commonly used method of summarizing the hierarchical clustering results (although "skyline plots" are prevalent in SAS [not recommended for classification procedures]).

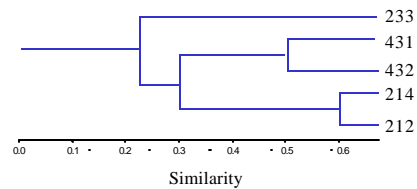
Dendrograms only display the clustering *topology* and object labels, not the links between objects.

Dendrograms are made of *branches* that meet at *nodes* which are drawn at the similarity value where *fusion* of the branches takes place. Note that the branches *furcating* from any node can be switched (swiveled) without ever affecting the information content.

## Single Linkage Clustering

- Example, Step 2 -

Now, arrange ordered similarity data in to a dendrogram:



## Single Linkage Rule

From the previous example, it should be clear that the rule for assigning an object to a cluster, in single linkage clustering, requires an object to display a similarity at least equal to the level of partition with *at least one object already a member of the cluster*.

The assignment rules differ among clustering procedures.

SL clustering forms a *chaining rule* whereby at each level of partition, two objects must be allocated to the same subset if their degree of similarity is equal to or higher than that of the partitioning level considered.

## Single Linkage Clustering

Single linkage clustering provides a fairly accurate picture of the relationships between pairs of objects. But, because of the propensity to exhibit *chaining*, it may be undesirable in many ecological analyses.

The chaining phenomenon is an issue because this means that the presence of an object midway between two compact clusters, or a few intermediates connecting the two clusters, is enough to turn them in to a single cluster.

*Clusters only chain if intermediates* are present, so the chaining of the data sometimes provides some useful insight in to the data.

## Cophenetic Matrix

Any classification or partition can be fully described by a cophenetic matrix. This matrix is used for comparing different classifications of the same subjects.

The cophenetic similarity (or distance) of two objects  $x_1$  and  $x_2$  is defined as the similarity level at which objects  $x_1$  and  $x_2$  become members of the same cluster during the course of clustering.

## Cophenetic Matrix

Consider the single linkage cluster analysis we just performed. the cophenetic similarity matrix (derived from the dendrogram itself) for that data set is:

S	212	214	233	431	432
212	–				
214	0.600	–			
233	0.214	0.214	–		
431	0.214	0.214	0.300	–	
432	0.214	0.214	0.300	0.500	–

## Ultrametric Property

If there are no reversals in the clustering, a classification has the following ultrametric property and the cophenetic matrix is called ultrametric:

for every triplet of objects ( $x_1, x_2, x_3$ ) in the study.

The ultrametric property may also be expressed in terms of similarities:

## A Cornucopia of Clustering Methods

Clustering algorithms have been developed using a wide range of conceptual models for studying all sorts of problems. There are *5 major dichotomies among methods*:



1. Sequential vs. Simultaneous algorithms
2. Agglomerative vs. Divisive methods
3. Monothetic vs. Polythetic methods
4. Hierarchical vs. Non-hierarchical methods
5. Probabilistic vs. Non-probabilistic methods

### Sequential vs. Simultaneous Algorithms

Most clustering methods are *sequential* in that they proceed by applying a *recurrent sequence of operations* to a group of objects (e.g., single linkage clustering).

In *simultaneous* algorithms, which are infrequent, the solution is obtained in a *single step*. ordination procedures tend to be more of the latter type.

### Agglomerative vs. Divisive Methods

Among the sequential algorithms, *Agglomerative* procedures begin with the discontinuous partition of all objects, i.e., the objects are considered as being separate from one another. They are successively grouped into larger and larger clusters.

If the single group of all objects is used as the starting point of the procedure, and smaller and smaller groups are derived by partitioning, the algorithm is *divisive*.

### Monothetic vs. Polythetic Methods

Divisive clustering methods may be monothetic or polythetic.

*Monothetic* models use a single descriptor as a basis for the partitioning (at each partition, one descriptor is chosen).

*Polythetic* models use several descriptors which, in most cases, are combined in to an association matrix prior to clustering.

### Hierarchical vs. Non-hierarchical Methods

In hierarchical methods, the members of inferior-ranking clusters become members of larger, higher-ranking clusters. Most of the time, hierarchical methods produce non-overlapping clusters. (Single linkage clustering is of this type.)

Non-hierarchical methods are very useful in EEB. They produce a single partition which optimizes within-group homogeneity, instead of a hierarchical series of partitions. This method should be used when the aim is to obtain a direct representation of the relationships among objects instead of a summary of their hierarchy.

### Probabilistic vs. Non-probabilistic Methods

Probabilistic methods include the clustering model of Clifford & Goodall (1967) and the parametric and nonparametric methods for estimating density functions in multivariate space.

The method of C&G is recommended when employing Goodall's similarity coefficient ( $S_{23}$ ), for the clustering of species in to biological associations.

### Hierarchical Agglomerative Clustering

1. Single linkage agglomerative clustering
2. Complete linkage agglomerative clustering
3. Intermediate linkage clustering
4. Unweighted arithmetic average clustering (UPGMA)
5. Weighted arithmetic average clustering (WPGMA)
6. Unweighted centroid clustering (UPGMC)
7. Weighted centroid clustering (WPGMC)
8. Ward's minimum variance method
9. General agglomerative clustering model
10. Flexible clustering
11. Information analysis

## Single Linkage Agglomerative Clustering

In single linkage agglomerative clustering (already covered), two clusters fuse when the two objects closest to each other (one in each cluster) reach the similarity of the reconsidered partition.

As a consequence of chaining, results of SL clustering are fairly sensitive to noise in the data, because noise changes the similarity values and may thus easily modify the order in which objects cluster.

## Single Linkage Clustering

Example: Cockroaches

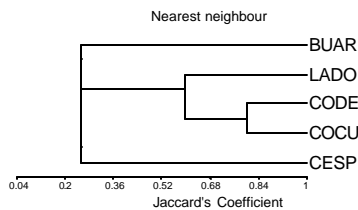
Let's return to an example we investigated previously (when looking at the simple matching similarity coefficient): presence/absence data are available for 5 species of cockroaches in 6 localities.

	BCI	LC	FORT	BOQ	MIR	CORG
CESP	0.000	0.000	0.000	0.000	0.000	1.000
COCU	1.000	1.000	1.000	1.000	1.000	0.000
CODE	1.000	1.000	1.000	0.000	1.000	0.000
BUAR	1.000	0.000	0.000	0.000	0.000	0.000
LADO	1.000	1.000	0.000	0.000	1.000	1.000

## Single Linkage Clustering

Example, Cockroach Data

After generating a similarity hemi-matrix of Jaccard's coefficients, we conduct a SLC or (*nearest neighbor* analysis) using MVSP software:



## Complete Linkage Agglomerative Clustering

Complete linkage agglomeration (also known as *farthest neighbor* analysis) is essentially opposite in approach to the single linkage analysis.

In this method, the fusion of two clusters relies on the *most distant pair of objects* instead of the closest. Thus, an object joins a cluster only when it is linked to all the objects already in the cluster. Two clusters can fuse only when all objects of the first cluster are linked to all objects of the second cluster.

## Complete Linkage Agglomerative Clustering

PROS:

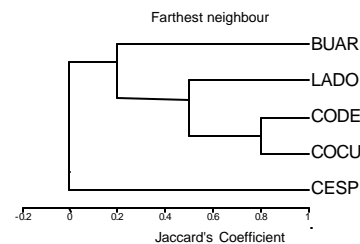
CL clustering produces *maximally linked and spherical clusters* (instead of the chained clusters of SL). This feature is often desirable in EEB when wishing to emphasize discontinuities.

CONS:

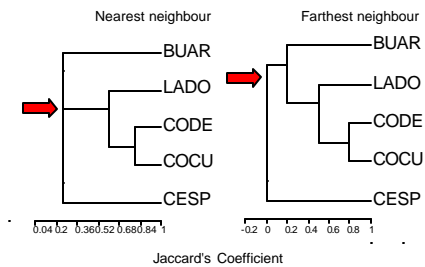
CL clustering relies on a virtually arbitrary rule known as the "*right-hand rule*" in choosing to include objects in a cluster. An example is when two objects or two clusters could be included in a third cluster, while these two objects or clusters have not completed the linkage with each other.

## Complete Linkage Agglomerative Clustering

Example, Cockroach Data, MVSP



### Single vs. Complete Linkage Comparison Cockroach Data, MVSP



### Intermediate Linkage Clustering Methods

Between the chaining problem of SL and the extreme space dilation of CL, the most interesting solution in EEB may be a type of clustering that approximately conserves the metric space A.

If the interest lies in the clusters shown in the dendrogram, and not the actual similarity links between clusters shown by the subgraphs, the average clustering methods of the next four (4) sections may be the most useful since they conserve the metric properties of the reference space.

### Intermediate Linkage Clustering Methods

In intermediate linkage clustering, the fusion criterion of an object or a cluster with another cluster is considered satisfied when a given proportion of the total possible number of similarity links is reached.

For example, if the criterion of connectedness is  $C = 0.5$ , two clusters are only required to share 50% of the possible links to fuse. This criterion has been referred to in the literature as *proportional link linkage*.

When  $C = 1$ , the method is called *average linkage clustering* (next 4 methods).

### Average Clustering

There are four methods of average clustering that conserve the metric properties of reference space.

Since they do not tally the links between clusters, they are not object-linkage methods in the sense of the previous three methodologies. They rely instead on average similarities among:

- objects (UPGMA, WPGMA), or
- centroids (UPGMC, WPGMC) of clusters.

### Unweighted Arithmetic Average Clustering (UPGMA)

"Unweighted arithmetic average clustering", or "unweighted pair-group method using arithmetic averages" (as originally defined by Sneath and Sokal 1973), or "average linkage" by SAS and SYSTAT are all the same UPGMA procedure.

The highest similarity identifies the next cluster to be formed. Following this event, the method computes the arithmetic average of the similarities between a candidate object and each of the cluster members. All objects receive equal weights in the computation. The similarity matrix is updated and reduced in size at each clustering step. Clustering proceeds by agglomeration.

### Average Clustering

Example: Plankton Data

Let's return to the plankton data set to develop the 4 methods of average clustering. Recall there are 38 species of plankton 5 ponds, 0/1 data were collected,  $S_{20}$  coefficient was derived with  $k = 2$ :

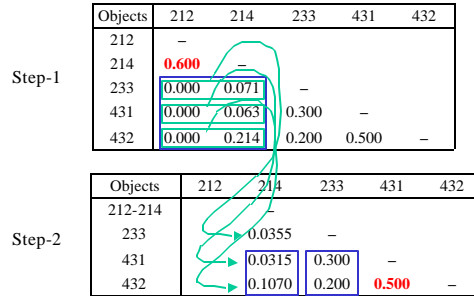
Ponds	Ponds				
	212	214	233	431	432
212	-				
214	0.600	-			
233	0.000	0.071	-		
431	0.000	0.063	0.300	-	
432	0.000	0.214	0.200	0.500	-

## Unweighted Arithmetic Average Clustering UPGMA Example, Plankton Data Set

At step 1, the highest similarity value is identified  $S(212,214) = 0.600$ ; hence the first two objects fuse at 0.600.

The similarity of these two objects with each of the remaining objects in the study must be averaged (values in the inner box of step 1); this results in a reduction of the size of the similarity matrix, and produces the matrix shown in step 2.

## UPGMA Example, Plankton Data Set



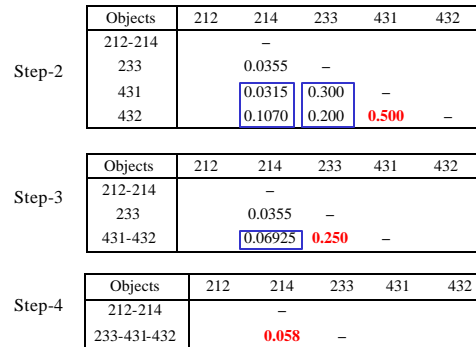
## Unweighted Arithmetic Average Clustering UPGMA Example, Plankton Data Set

At step 2, the highest similarity value remaining is identified ( $S = 0.500$ ); it indicates that 431 and 432 fuse at 0.500.

Again, this similarity value is obtained by averaging the boxed values; this produces a new reduced similarity matrix for the next step.

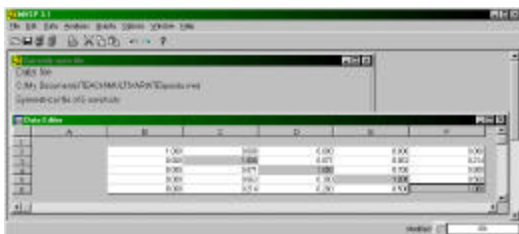
At step 3, the largest similarity is 0.250; it leads to the fusion of the already-formed group (431, 432) with object 233 at a level of 0.250. Because there is 1 object in group 233 and two in group (431,432), the fused similarity is calculated as  $[(0.0355*1) + (0.06925 * 2)]/3$ .

## UPGMA Example, Plankton Data Set



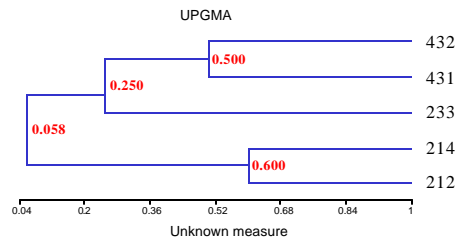
## Unweighted Arithmetic Average Clustering UPGMA Example, Plankton Data Set > MVSP

First, calculate similarity coefficients in Excel, then build a symmetrical data matrix...



## Unweighted Arithmetic Average Clustering UPGMA Example, Plankton Data Set > MVSP

Select UPGMA ( $S$  will be indicated as "pre-calculated").  
The resulting dendrogram:



### Weighted Arithmetic Average Clustering (WPGMA)

It often occurs in EEB that groups of objects, representing different regions or groups, are of unequal sample size. Eliminating objects to equalize the clusters would mean discarding valuable information.

Unfortunately, the presence of a large group of objects, which are more similar *a priori* because of their common origin, may greatly distort a UPGMA.

WPGMA down-weights the largest group by giving equal weights to the two branches of the dendrogram that are about to fuse.

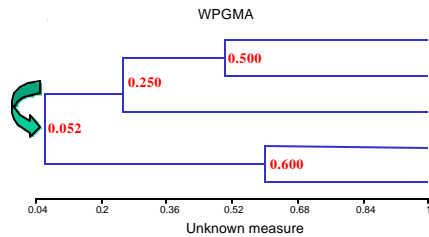
### Weighted Arithmetic Average Clustering (WPGMA)

In the specific case of the plankton example, the only step that changes is step-4 where the last fusion value is calculated.

It is calculated as:  $(0.0355 + 0.06925)/2 = 0.052375$ .

Objects	212	214	233	431	432
212-214		-			
233-431-432			0.05237	-	

### Weighted Arithmetic Average Clustering WPGMA Example, Plankton Data Set > MVSP



### Unweighted Centroid Clustering (UPGMC)

Unweighted centroid clustering, or unweighted pair-group centroid method (UPGMC; Sneath & Sokal 1973) is based on a simple geometric approach.

The centroid of a cluster of objects may be imagined as the type-object of the cluster, whether that object actually exists or is only a mathematical construct.

In A-space, the co-ordinates of the centroid of a cluster are computed by averaging the coordinates of the objects in the group.

### Unweighted Centroid Clustering (UPGMC)

UPGMC proceeds to the fusion of objects or clusters presenting the highest similarity, as in the previous methods. At each step, the members of a cluster are replaced by their *common centroid* (i.e., "mean point").

The centroid is considered to represent a new object for the remainder of the clustering procedure; in the next step, one looks again for the pair of objects with the greatest similarity, on which the procedure of fusion is repeated.

### Unweighted Centroid Clustering (UPGMC)

Gower (1967) proposed the following formula for centroid clustering, where the similarity of the centroid ( $\mathbf{h}_i$ ) of the objects of clusters  $\mathbf{h}$  and  $\mathbf{i}$  with a third object or cluster  $\mathbf{g}$  is computed from the similarities  $S(\mathbf{h}, \mathbf{g})$ ,  $S(\mathbf{i}, \mathbf{g})$ ,  $S(\mathbf{h}, \mathbf{i})$ :

$$S(\mathbf{h}_i, \mathbf{g}) = \frac{w_h}{n_h + n_i} S(\mathbf{h}, \mathbf{g}) + \frac{w_i}{n_h + n_i} S(\mathbf{i}, \mathbf{g}) + \frac{w_h w_i}{n_h n_i} [1 - S(\mathbf{h}, \mathbf{i})]$$

where the  $w$ 's are weights given to the clusters.  $\mathbf{g}$ ,  $\mathbf{h}$ , &  $\mathbf{i}$  can represent points or clusters. The number of objects  $n_h$  and  $n_i$  are often used as weights  $w_h$  and  $w_i$ .

### Unweighted Centroid Clustering (UPGMC)

Step-1

Objects	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	-

Step-2

Objects	212	214	233	431	432
212-214	-				
233		0.1355	-		
431		0.1315	0.300	-	
432		0.2070	0.200	0.500	-

### Unweighted Centroid Clustering (UPGMC)

Step-2

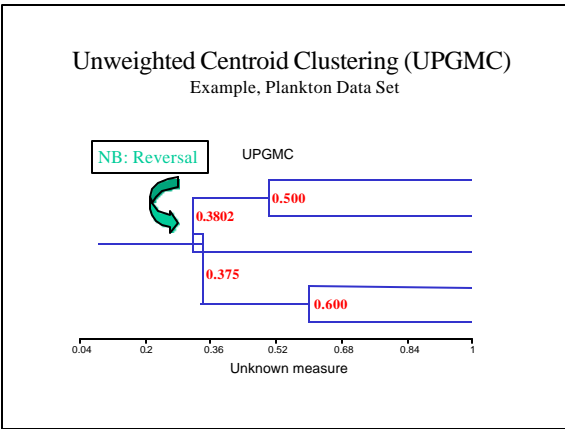
Objects	212	214	233	431	432
212-214	-				
233		0.1355	-		
431		0.1315	0.300	-	
432		0.2070	0.200	0.500	-

Step-3

Objects	212	214	233	431	432
212-214	-				
233		0.1315	-		
431-432		0.29425	0.375	-	

Step-4

Objects	212	214	233	431	432
212-214	-				
233-431-432		0.3802	-		



### Unweighted Centroid Clustering (UPGMC)

Final Comments

As seen in the previous dendrogram, UPGMC may lead to reversals. Many authors feel uncomfortable about reversals since they violate the ultrametric property.

Unweighted centroid clustering may be used with any similarity coefficient, but the Gower's formula presented here is the only one that retains its geometric properties for similarities corresponding to metric distances (cf.  $S_{19}$ )

### Weighted Centroid Clustering (WPGMC)

Weighted centroid clustering was first proposed by Gower (1967) and referred to as the weighted pair-group centroid method (WPGMC) by Sneath and Sokal (1973). It plays the same role with respect to UPGMC as WPGMA plays with respect to UPGMA.

Certain centroids may be biased because they are over-represented in the data and WPGMC adjusts for this. The problem is corrected by giving equal weights to two clusters on the verge of fusing, independently of the number of objects in each cluster.

### Weighted Centroid Clustering (WPGMC)

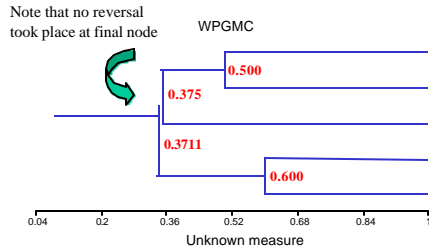
To achieve the weighting, we re-write the previous UPGMC equation to a WPGMC form:

$$S(\mathbf{h}_i \mathbf{g}) = \frac{T_i}{T} S(\mathbf{h}_i \mathbf{g}) + \frac{S(\mathbf{g}_i \mathbf{g})}{T} + \frac{T - T_i}{T} S(\mathbf{h}_i \mathbf{h})$$

In the pond example, the last fusion is calculated as:

$$S_{r(233-431-432)(212-214)} = \frac{T}{T} (0.1355 + 0.29425) + \frac{T - T}{T} (1 - 0.375) = 0.371125$$

### Weighted Centroid Clustering (WPGMC) Example, Plankton Data Set



### Weighted Centroid Clustering (WPGMC) Final Comments

Note that WPGMC did not produce a reversal at the last node, although, WPGMC is capable of producing reversals (in general).

Note also that in R-mode, weighted centroid clustering does not make sense if the measure of association is Pearson's  $r$ . Correlations are cosine transformations of the angles between descriptors and these cannot be combined using the WPGMC equation.

### Ward's Minimum Variance Method (Ward's)

Ward's (1963) minimum variance method is related to the centroid methods in that it too leads to a geometric representation in which cluster centroids play an important role. To form clusters, the method minimizes an objective function which is, in this case, the same "squared error" criterion as that used in the MANOVA.

At the beginning of the procedure, each object is in a cluster of its own, so that the distance of an object to its cluster centroid is 0; hence, the sum of all these distances is also 0. As clusters form, the centroids move away from actual object coordinates and the sums of the squared distances from the objects to centroids increases.

### Ward's Minimum Variance Method (Ward's)

Ward's method finds the pair of objects or clusters whose fusion increases as little as possible the sum, of the squared distances between objects and cluster centroids. The distance of object  $x_i$  to the centroid  $m$  of its cluster is computed using the Euclidean distance formula:

$$\sqrt{\sum_{j=1}^p [y_{ij} - m_j]^2}$$

The sum of squared distances of all objects in cluster  $k$  to their common centroid is called "error" in ANOVA, hence  $e_k$ :

$$e_k^2 = \sum_{i=1}^m \sum_{j=1}^p [y_{ij}^{(k)} - m_j^{(k)}]^2$$

where  $y_{ij}^{(k)}$  is the value of the descriptor  $y_j$  for an object  $i$  member of group  $(k)$  and  $m_j^{(k)}$  is the mean value of the descriptor  $j$  over all members of group  $k$ .

### Ward's Minimum Variance Method (Ward's)

Alternatively, the within-cluster sums of squared errors  $e_k^2$  can be computed as the mean of the squared distances among cluster members:

$$e_k^2 = \left[ \sum_{i=1}^{n_k} \sum_{j=1}^{n_k} D_{ij}^2 \right] / m^k$$

where the  $D_{ij}^2$  are the squared distances among objects in cluster  $k$  and  $n_k$  is the number of objects in that cluster.

### Ward's Minimum Variance Method (Ward's)

The sum of squared errors  $E^2_k$ , for all  $K$  clusters corresponding to a given partition, is the criterion to be minimized at each step:

$$E^2_k = \sum_{k=1}^K e_k^2$$

At each clustering step, two objects or clusters  $h$  and  $i$  are merged into a new cluster  $hi$ , as in previous sections. Since changes occurred only in the groups  $h$ ,  $i$ , and  $hi$ , the change in the overall sum of squared errors,  $\Delta E^2_{hi}$ , may be computed from the changes as:

$$\Delta E^2_{hi} = e_h^2 + e_i^2 - e_{hi}^2$$

### Ward's Minimum Variance Method (Ward's)

It can be shown that this change depends only on the distance between the centroids of clusters **h** and **i** and on their numbers of objects  $n_h$  and  $n_i$ :

$$\Delta E_{hi}^2 = \frac{n_h n_i}{n_h + n_i} \sum_{j=1}^p [m_j^{(h)} - m_j^{(i)}]^2$$

Another way is to calculate the fusion distances between **hi** and **g** for an agglomeration table:

$$D^2(\mathbf{hi}, \mathbf{g}) = \frac{n_h + n_g}{n_h + n_i + n_g} D^2(\mathbf{h}, \mathbf{g}) + \frac{n_i + n_g}{n_h + n_i + n_g} D^2(\mathbf{i}, \mathbf{g}) - \frac{n_g}{n_h + n_i + n_g} D^2(\mathbf{h}, \mathbf{i})$$

### Ward's Clustering

Example, Plankton Data Set

NB: squared distances used and computed as  $D = (1 - S)^2$

Step-1

Objects	212	214	233	431	432
212	-				
214	<b>0.1600</b>	-			
233	1.000	0.8630	-		
431	1.000	0.8779	0.4900	-	
432	1.000	0.6178	0.6400	0.2500	-

Step-2

Objects	212	214	233	431	432
212-214	-				
233		1.1887	-		
431		1.1986	0.4900	-	
432		1.0252	0.6400	<b>0.2500</b>	-

### Ward's Clustering

Step-2

Objects	212	214	233	431	432
212-214	-				
233		0.1187	-		
431		1.1987	0.4900	-	
432		1.0252	0.6400	<b>0.2500</b>	-

Step-3

Objects	212	214	233	431	432
212-214	-				
233		0.18869	-		
431-432		0.54288	<b>0.6700</b>	-	

Step-4

Objects	212	214	233	431	432
212-214	-				
233-431-432		<b>1.67952</b>	-		

### General Agglomerative Clustering

Lance & Williams (1967) proposed a general model that encompasses all the agglomerative clustering methods previously presented, except using an intermediate linkage.

The general model allows one to select an agglomerative clustering model by choosing the values of four parameters called  $\alpha_h$ ,  $\alpha_i$ ,  $\beta$ , and  $\gamma$ , which determine the clustering strategy.

### General Agglomerative Clustering

For similarities, the general model is:

$$D_{hi} = \frac{\alpha_h D_h + \alpha_i D_i + \beta D_{hg} + \gamma D_{ig}}{\alpha_h + \alpha_i + \beta + \gamma}$$

When using distances, the equation becomes:

$$D_{hi} = \frac{\alpha_h D_h + \alpha_i D_i + \beta D_{hg} + \gamma D_{ig}}{\alpha_h + \alpha_i + \beta + \gamma}$$

Clustering proceeds in the same way for all combinatorial agglomerative methods. As the similarity decreases, a new cluster is obtained by the fusion of the two most similar objects and/or clusters. The matrix is thus reduced by one row and one column at a time.

### Flexible Clustering

Lance & Williams (1967) proposed to vary the parameter  $\beta$  between -1 and +1 to obtain a series of intermediate solutions between single linkage chaining and the space dilation of complete linkage. This method is also referred to in the literature as *Beta-flexible Clustering*.

Lance & Williams have shown that, if the other parameters are constrained as follows:

$$\alpha_h = \alpha_i = \beta = \gamma = 1$$

then the resulting clustering will be ultrametric.

## Flexible Clustering

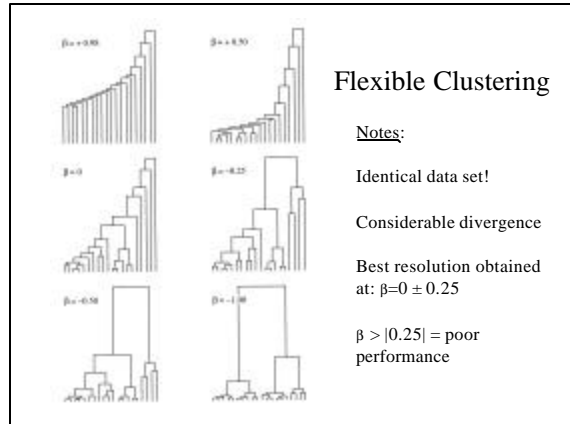
When  $\beta$  is close to 1, strong chaining is obtained.

When  $\beta$  is close to -1, strong space dilation is observed.

The space properties are conserved for small values of  $\beta$  around 0 ( $\pm 0.25$ ).

Like weighted centroid clustering, flexible clustering is compatible with all association measures except Pearson's  $r$ .

The following figure shows the effect of varying  $\beta$  in the clustering of 20 objects:



## Information Analysis

The Q-mode clustering method called information analysis was developed for EEB purposes by Lance & Williams (1966).

It does NOT proceed by the normal steps of similarity calculation followed by clustering. It is a direct method of clustering based on information measures.

Information analysis is a type of unweighted centroid clustering, specifically adapted to species data.

## Shannon's Formula

Despite its widespread use as a measure of species diversity, Shannon's formula was originally designed to measure the diversity of information in a frequency or probability distribution:

$$H = - \sum_{j=1}^p p_j \log p_j$$

In practice, this formula is usually applied to presence-absence data. The information measure is NOT applicable to raw abundance data here because too many different states would be created among species.

## Information Analysis

Example, Pond Data

Sp. j	Ponds					$p_j$	$(1-p_j)$
	212	214	233	431	432		
1	1	1	0	0	0	0.4	0.6
2	0	0	1	1	0	0.4	0.6
3	0	1	1	0	1	0.6	0.4
4	0	0	1	1	1	0.6	0.4
5	1	1	0	0	0	0.4	0.6
6	0	1	0	1	1	0.6	0.4
7	0	0	0	1	1	0.4	0.6
8	1	1	0	0	0	0.4	0.6

## Information Analysis

Example, Pond Data

The entropy of each species presence-absence descriptor  $j$  is calculated on the basis of the probabilities of presence  $p_j$  and absence  $(1-p_j)$  of species  $j$ , which appear in the right-hand side of the table.

The probability of presence is estimated as the number of ponds where species  $j$  is present, divided by the total number of ponds in the cluster under consideration. The probability of absence is estimated likewise.

The entropy of species  $j$  is therefore:

$$H_j = - [p_j \log p_j + (1-p_j) \log (1-p_j)]$$

## Information Analysis

Example, Pond Data

The base of the logarithms is unimportant as long as it is consistent throughout the calculations. We will use natural logs for the present example.

Thus, for the first species,  $H(1)$  would be:

$$H(1) = -[0.4 \ln(0.4) + 0.6 \ln(0.6)] = 0.673$$

The information of the conditional probability table can be calculated by summing the entropies per species (considering all species have the same weight).

## Information Analysis

Example, Pond Data

Since the measure of the total information in the group must also take in to account the number of objects in the cluster, the formula is defined as:

$$I = -n \sum_{j=1}^p [p_j \log p_j + (1-p_j) \log(1-p_j)] \text{ for } 0 < p_j < 1$$

Where  $p$  is the number of species represented in the group of  $n$  objects (ponds). For the group of 5 ponds,

$$I = -5 [8 (-0.673)] = 26.920$$

$I$  is zero when all ponds in a group contain the same species.

## Information Analysis

Example, Pond Data

At each clustering step, three series of values are considered:

- the total information  $I$  in each group; 0 at beginning,
- the value of  $I$  for all possible combinations of groups taken two at a time, and
- the increase of information  $\Delta I$  resulting from each possible fusion.

All of these values can be placed in a matrix, initially of dimension  $n \times n$  which decreases as clustering proceeds.

For the data set here, values of information in each group (a) are placed on the diagonal, values (b) in the lower hemi-matrix, and values (c) in the upper hemi-matrix.

## Information Analysis

Example, Pond Data

Ponds	Ponds				
	212	214	233	431	432
212	0	2.733	8.318	9.704	9.704
214	2.733	0	8.318	9.704	6.931
233	8.318	8.318	0	4.159	4.159
431	9.704	9.704	4.159	0	2.773
432	9.704	6.931	4.159	2.773	0

The  $\Delta I$  for two groups is found by subtracting the corresponding values  $I$ , on the diagonal, from the value  $I$  of their combination in the lower hemi-matrix. Values on the diagonal are all 0 in the first matrix only, thus values in the upper & lower hemi-matrices are identical (not the case in subsequent matrices).

## Information Analysis

Example, Pond Data

The first fusion is identified by the *lowest*  $\Delta I$  value found in the upper hemi-matrix. This value is 2.773 for pairs (212,214) and (431,432), which therefore fuse. A new matrix of  $I$  values is computed:

Groups	Groups		
	212-214	233	431-432
212-214	2.773	10.594	15.588
233	13.367	0	4.865
431-432	21.134	7.638	2.773

Groups	Groups		
	212-214	233	431-432
212-214	2.773	10.594	15.588
233	13.367	0	4.865
431-432	21.134	7.638	2.773

## Information Analysis

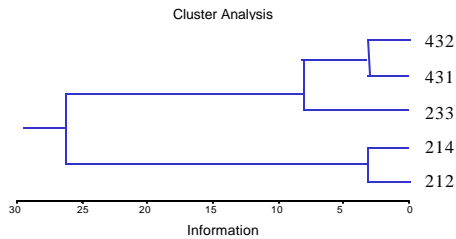
Example, Pond Data

The  $\Delta I$  values in the upper triangle now differ from the  $I$ 's in the lower triangle since there are  $I$  values on the diagonal. The  $\Delta I$  corresponding to the group (212, 214, 431, 432), for example, is calculated as  $21.134 - 2.773 - 2.773 = 15.588$ .

The lowest value of  $\Delta I$  for (233, 431, 432) fuses at the next step (7.638). The matrix is reduce to  $2 \times 2$  and the last fusion occurs at  $I = 26.920$ .

## Information Analysis

Example, Pond Data



## Hierarchical Divisive Clustering

Contrary to agglomerative methods, hierarchical divisive techniques use the whole set of objects as the starting point. They divide it into two or several subgroups, after which they consider each subgroup and divide it again, until the criterion chosen to end the divisive procedure is met.

Options include:  
 Monothetic methods  
 Polythetic methods  
 Twinspan

## Monothetic Methods

The clustering methods that use one descriptor at a time are less than ideal. The best known of these analyses is *association analysis* (Williams & Lambert 1959).

Association analysis may actually be applied to any binary data table, not just species.

The problem is to identify, at each step of the procedure, which descriptor is the most strongly associated with all of the others. In other words, individual species are used as divisors.

## Association Analysis

Chi-square values are computed for all pairs of descriptors using the usual formula:

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

The formula may include Yates' correction for small sample sizes, as in similarity coefficient  $S_{25}$ . The  $\chi^2$  values relative to each descriptor  $k$  are summed:

$$\sum_j \chi_{jk}^2 \text{ for } j \neq k$$

## Association Analysis

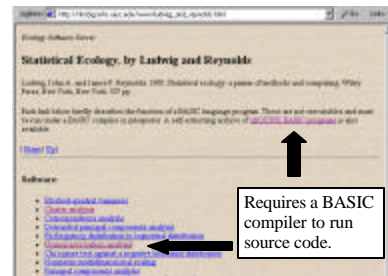
The largest sum identified the descriptor that is most closely related to all of the others. The first partition is made along the states of this descriptor (present/absent).

The original association analysis is an excellent tool for understanding the essence of how cluster analysis works, and while not widely used anymore, still has some useful applications and is prevalent in the older literature.

There is an excellent discussion and worked example of the methodology in Causton (1988). The only available computer version that I am aware of is by Ludwig and Reynolds (1988).

## Association Analysis

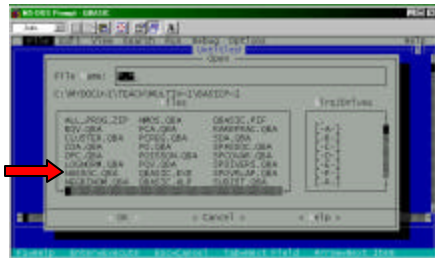
[http://nhsbig.inhs.uiuc.edu/www/ludwig\\_and\\_reynolds.html](http://nhsbig.inhs.uiuc.edu/www/ludwig_and_reynolds.html)



Requires a BASIC compiler to run source code.

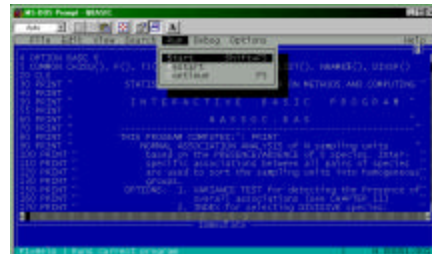
### Association Analysis

You need to copy 2 files to your windows directory called QBASIC.EXE and QBASIC.HLP (Windows CD under \other\olddos\). Open NAASOC.QBA to get QB code.



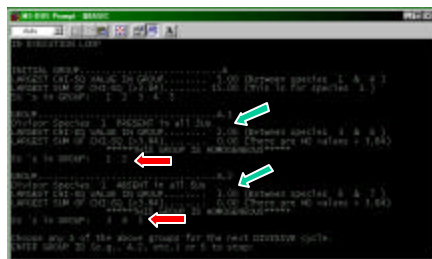
### Association Analysis

QB code comes up for program. If you are familiar with BASIC code, you can alter the programming or add comments for specific applications. Otherwise, Run, Start to launch program; analyze pond data (5 sites, 8 spp., 0/1).



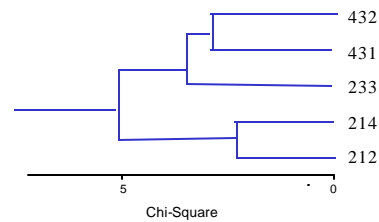
### Association Analysis

Answer queries as appropriate. Output in tabular form. Dendrogram constructed manually. Results show Ponds 212&214 in one group and 233&431&432 in another.



### Association Analysis

After completing all divisions, the chi-square values from the printout can be used to construct a dendrogram. Note the similarity of overall result with WRT previous algorithms.



### Polythetic Methods

There is no satisfactory algorithm for the hierarchical division of objects based upon the entire set of descriptors.

One form is the dissimilarity analysis of Mcnaughton-Smith (1964) that firsts looks for the object that is the most different from everything else and then removes it. One by one, each object is removed. Objects are removed up to the point where the distance between clusters can no longer be increased.

Unfortunately, dissimilarity is known to produce some very odd results, particularly when many small clusters are present.

### TWINSpan

Two Way INdicator SPecies ANalysis

Despite the name, TWINSpan is not just designed to identify indicator species (name came from its original development in vegetation science). This is a very popular procedure that classifies the objects by hierarchical division and constructs an ordered two-way table.

TWINSpan produces a tabular matrix arrangement that approximates a Braun-Blanquet phytosociological analysis table.

## TWINSpan

The technique is based on the concept that a group of samples which constitute a community type will have a corresponding group of species that characterize that type (*indicator species*).

Since *reciprocal averaging* (RA ordination procedure to be covered soon) arranges species and samples in a way that best expresses that relationship, RA is the basis of TWINSpan.

TWINSpan is largely based on presence-absence data, but quantitative data are incorporated by considering different abundance levels of the same species to be different species. This results in what is commonly referred to as *pseudo-species*, and, depending on how you set it up, a single species may "become" 5 pseudo-species.

## TWINSpan

This technique is computationally pretty complex & we can not easily do a worked example here. However, it relies on basic principles that we have already discussed. An excellent treatment of the method can be found in Kent and Coker (1992).

Basically, TWINSpan begins with an RA ordination, which is then divided at its centroid. Each sample is then classified into one of two groups, and a discriminant function analysis is used to refine the classification; some samples will switch groups, depending on the maximum likelihood determined by the DFA.

This new dichotomy is used to look for preferential (indicator) species.

## TWINSpan

Species are then scored according to their degree of preference for one side or the other, highly preferential species are those that are at least 3 times more common on one side. These species scores are once again applied back to the samples, using a weighting algorithm, and borderline cases may once again switch position. This is now the first division.

The process is hierarchical, and so each of the new groups then undergoes the same process, until either a certain number of divisions have been reached or a group is too small to subdivide further. Once all the samples have thus been classified, the species are classified according to their overall fidelity to the groups, and a sorted table is produced.

## TWINSpan

While this is the most popular of the hierarchical divisive clustering techniques, there are some problems. The original TWINSpan code was buggy and produced poor results. There is a new version contained within PC-ORD (v.4) which is very reliable and produces excellent results.

One computational problem that has considerable biological consequence has to do with the way early cuts are defined—these make all the difference, and can be affected too strongly by only a few species. Nonetheless, this procedure has good utility in your repertoire of multivariate methods.

## TWINSpan

-Example-

Consider a data set that comes from Gauch (1982) that contains 14 spp. of forest trees sampled in 10 forests. The data are octave transforms (0-10 scale) based upon relative abundances.

	QUMA	QUVE	CAOV	FRSE	QUAL	JUNI	QURU	JUCI	ULAM	TIAM	ULRU	CACO	OSVI	ACSA
1	9	8	6	3	5	2	3	2	2		4			
2	8	9	6	5	4		4		2					
3	3	8	2	6	9			5	4		2			
4	5	7	7	6	9		6		5		2			
5	6		6	7	3	9	2	6	2	5				
6			2	4	7	5	8		7	7	5		5	
7	5			5	4	6	7		5	6	8	6	7	4
8				6	4	6	2		6	8	4	4	8	
9				4		3	4		2	7	8		6	8
10				1	2		3	2	5	6	7	3	5	9

## TWINSpan



<http://www.ptinet.net/~mjm/pcordwin.htm>

## TWINSPAN PC-ORD

## TWINSPAN PC-ORD

Note that a dendrogram can be drawn by hand using the columns of numbers at the right.

In the first column, the 0/1s reflect the first bifurcation. Each additional column reflects subsequent bifurcations.

### Partitioning by K-means

Partitioning consists in finding a single partition of a set of objects.

In other words, given  $n$  objects in a  $p$ -dimensional space, determine a partition of objects into  $K$  groups, or clusters, such that the objects within each cluster are more similar to each other than to objects in other clusters.

The number of  $K$  groups is determined by the user.

### Partitioning by K-means

The difficulty here is in deciding what “*more similar*” means. In other words, what criterion determines the level of similarity?

A *global criterion* would be, for instance, to represent each cluster by a type-object on a priori grounds and assign each object to the nearest type.

A *local criterion* uses the local structure of the data to delineate clusters; groups are identified by identifying *high density regions* in the data (most common approach).

### Partitioning by K-means

This algorithm was originally proposed by MacQueen (1967) and popularized by Lance and Williams (1967).

The objective function that the partition to discover should minimize is the same as in Ward’s classification method: the total sum of squares ( $E^2_k$ , or TESS). A problem encountered by the algorithm is that the solution depends, to some extent, on the initial position of the centroids.

Generically, this issue of the final answer depending upon initial conditions is referred to as the “*local minimum*” problem in algorithms.

### Partitioning by K-means

This is an excellent method, particularly for very large data sets, but has some limitations:

First, it can only be used with Euclidean distances. Many types of EEB data (e.g., species data in particular) are not amenable to analysis with Euclidean distance.

Second, it requires continuous data with no outliers. Some categorical data can be included, but this usually creates problems.

## Partitioning by K-means

[www.ncss.com](http://www.ncss.com)

One of the best places I have found to run K-means analysis is in NCSS

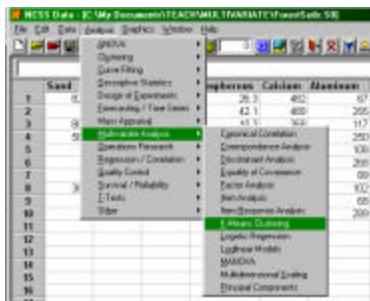


## Partitioning by K-means

Example: Forest Soils, 10 stands, 6 variables

	Sand	Clay	Nitrate	Phosphorous	Calcium	Aluminum
1	82.3	8.2	3.5	25.3	482	67
2	47	27	5.3	42.1	489	255
3	80.5	10.2	3.6	10.3	368	117
4	56.3	25.1	5.2	16.9	600	250
5	80	8.5	3.8	15.6	389	108
6	59	20.9	5.6	32.5	350	280
7	76	6.9	3.1	19.5	369	89
8	38.5	25	4.2	15.3	577	102
9	75	6.3	3.8	25.3	369	68
10	58	26.3	5.3	29.3	509	209

## Partitioning by K-means



## Partitioning by K-means

Iteration	No. of Clusters	Percent of Variables	Bar Chart
1	2	34.81	818
2	2	34.81	818
3	2	34.49	818
4	3	33.14	818
5	3	33.14	818
6	4	35.11	818
7	4	35.11	818
8	4	35.11	818
9	4	35.11	818

Cluster Summary	Cluster 1	Cluster 2
Sand	35.78	51.75
Clay	1.82	3.48
Nitrate	1.82	3.11

Helps determine the optimal no. of clusters.

Helps determine if you have chosen enough starting configurations. This suggests 3 was inadequate, go back & select 5 & re-run.

## Partitioning by K-means

Variable	DF1	DF2	Sum of Squares	Mean Square	F-Ratio	F-Ratio	F-Ratio
Sand	1	8	822.5	822.5	89.78	10.00000	0.00000
Clay	1	8	888.24	2.000	200.06	10.00000	0.00000
Nitrate	1	8	8.884	0.198	5.09	10.00000	0.00000
Phosphorous	1	8	188.501	88.576	3.10	10.00000	0.00000
Calcium	1	8	34425.6	30350.0	5.00	10.00000	0.00000
Aluminum	1	8	61122.9	30451.5	16.79	10.00000	0.00000

This report summarizes the results of performing a one-way ANOVA on each variable, using the currently defined clusters as the factor. This report helps you investigate the importance of each variable in the clustering process.

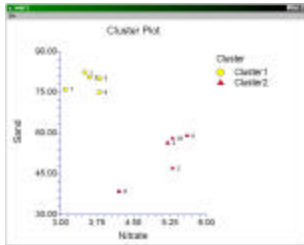
**DANGER:** recognize that this does NOT account for inter-variable correlation. DFA might be a better approach to reduce the number of variables.

## Partitioning by K-means

Cluster 1	Cluster 2	Cluster 3
1	1	1.0000
2	1	1.0000
3	1	0.5296
4	1	2.0734
5	1	0.9253
6	2	4.3208
7	2	4.1303
8	2	3.4756
9	2	4.9359
10	2	4.9359

This section displays which objects belong to which groups based upon the variables and the number of clusters specified. A desirable result is produced when:  $|Dist1 - Dist2|$  is "large". This example produces a good result.

## Partitioning by K-means



A series of bivariate plots, comparing all variables is useful to spot outliers, anomalies, & various other data problems. If the analysis was successful, clusters should be fairly distinct in these plots.

## Cluster Validation

At this point you should be thinking, after all of these algorithms, which one is producing “real” clusters? Alternatively, which clusters of a given technique are largely artifacts of that particular algorithm?

It is important to validate cluster analyses. One needs to show that the clustering structure is unique; i.e., that it departs from what might be expected from unstructured data.

**BAD NEWS:** validation procedures are generally not available in most stats software packages.

## Cluster Validation

Validation can be carried out in non-statistical ways, as well as statistical ways. The latter certainly provides stronger inference (several nice review papers have been written on how to do this; e.g., Milligan 1996). Commonly used non-statistical methods include:

- (1) Plot the clusters onto an ordination plot and look for separation of clusters.
- (2) Compare the results of several clustering algorithms.

## Summary of Clustering Methods

Pros & Cons

Hierarchical agglomeration: linkage clustering

### Single Linkage

Computation simple; contraction of space (chaining); combinatorial method. Good complement to ordination.

### Complete Linkage

Dense nuclei of objects; space expansion; many objects cluster at low similarity; arbitrary rules to resolve conflicts; combinatorial method; Increases contrast among clusters.

### Intermediate Linkage

IL: Preservation of reference space A; non-combinatorial; not included in the L&W “general model”. Preferable to SL & CL in most cases.

## Summary of Clustering Methods

Pros & Cons

Hierarchical agglomeration: average clustering

UPGMA: Fusion of clusters when the similarity reaches the mean cluster similarity. For objects obtained by simple or random sampling.

WPGMA: Same, but adjustment for group sizes. Preferable when sampling was other than simple or random.

UPGMC: Fusion of clusters with closest centroids; may produce reversals. For objects obtained by simple or random sampling.

WPGMC: Same, with adjustment for group sizes; may produce reversals; Preferable when sampling was other than simple or random.

## Summary of Clustering Methods

Pros & Cons

Hierarchical agglomeration:  
flexible clustering & information analysis

Flexible clustering: algorithm permits contraction, conservation, or dilation of A-space; pairwise relationships between objects are lost; combinatorial method. All are implemented with the same simple straightforward algorithm.

Information analysis: minimal chaining; only for Q-mode clustering with presence-absence data. Use is unclear, similarities reflect double absences as well as double presences; not recommended.

## Summary of Clustering Methods

Pros & Cons  
Hierarchical division

**Monothetic**: division of the objects following the states of the "best" descriptor; useful only to split data sets in to large clusters.

**Polythetic**: computationally intense, small data sets only (except if super-computer available).

**TWINSpan**: dichotomized ordination analysis; gives clear ordered table classifying sites and species; ecological justification of some steps questionable (e.g., *pseudospecies*).

## Summary of Clustering Methods

Pros & Cons  
K-means clustering

**K-means clustering**: minimizes within-group sum of squares; different rules may suggest different optimal numbers of clusters; danger of incorrect separation of members of minor clusters near the beginning of clustering; produces a partition of  $K$  groups as determined by the user.

## Clustering Methods

Final Comments

Most clustering methods remain very sensitive to outliers. These should be removed prior to any analysis.

The question of which similarity or distance measure to use remains largely unanswered, but there are useful guidelines.

Comparative studies seem to favor the use of average linkage in terms of its ability to recover known clusters.

With most of the clustering techniques, there are a large number of parameters to be set by the user and care must be exercised.

## Exploratory Data Analysis

Another approach to the search for classification or typologies of objects is through the use of graphical methods.

Graphical methods provide an extremely flexible medium for explaining, interpreting, and analyzing data by means of points, lines, areas, faces, or other geometric forms.

Collectively, these methods are often referred to as *icon plots*. They permit the graphical representation of  $p$ -dimensions in alternative forms.

## Exploratory Data Analysis

Graphical Methods, Example

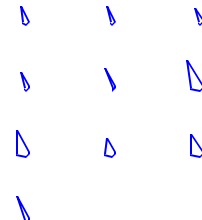
Let's return to the forest soils example ( $n = 10, p = 6$ ) that we used to develop the previous K-means procedure and explore some of the graphical EDA procedures.

I will use SYSTAT as it has the largest number of options in this category.

We will look at 5 graphical options for the exploration of multivariate data: Star, Fourier Blob, Chernoff Faces, Weather Vane, and Sun plots.

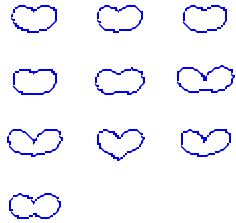
## Star Plots

Star icons are profile icons in polar coordinates; the distance of each point from the center of the icon shows the value of the corresponding variable. Separate icons are drawn for each case ( $n = 10$ ).



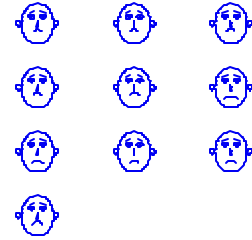
### Fourier Blobs

Polar coordinate Fourier waveforms. Each case in the data set is shown by a blob, and cases with similar values across all variables will have similar shapes.



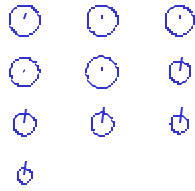
### Chernoff Faces

Chernoff faces represent many variables by assigning each variable to a distinct facial feature (head, nose, mouth, eyes, eyebrows, and ears). Cases with similar values for particular variables will have similar corresponding facial features. Max  $p = 20$ .



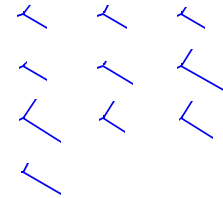
### Weather Vane Plots

Icons represent three variables. The first determines the radius of the central circle, the second determines the length of the vane, and the third determines its direction.



### Sun Plots

Sun plots are similar to star plots. However, the order of the variables is determined by the first principal component, which makes them easier to interpret.



# The End