Why is it useful to use multivariate statistical methods for microfacies analysis?

- A microfacies is a multivariate object: each sample is characterized by several variables (texture, allochems…);
- Multivariate statistical methods allow to study changes in several properties simultaneously and to manipulate more variables/samples than we can do.

CLUSTER ANALYSIS

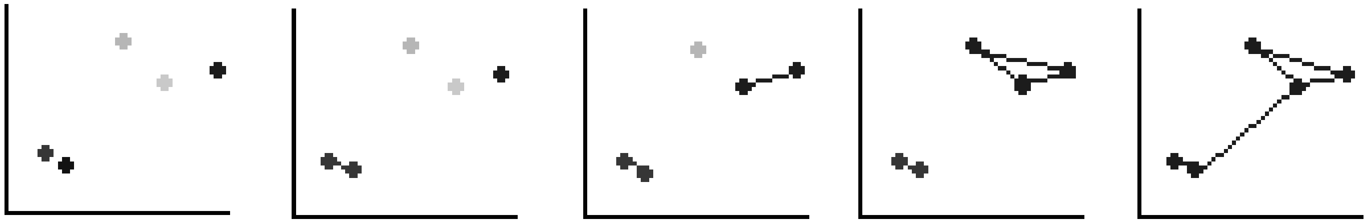
(hierarchical, agglomerative)

Basics

- Grouping of objects (samples) based on similarity or difference of their variables (components) > Q-mode (R-mode = variables);
- Reduces the dimensionality of your (multivariate) data table;
- Matrix of similarity coefficients: numerical similarity between all pairs of objects.

Procedure

1) Select variables (mixing different types is **not** adviced!);
2) Calculate distance/similarity between all samples (= initial ‘clusters’) and store in a distance matrix (= similarity matrix);
3) Select the two most similar initial clusters (samples) in the matrix and fuse them;
4) Calculate the distance between that new cluster and all others (mono-sample). Only the distances involving that cluster will have changed, no need to re-calculate all distances;
5) Repeat 3 until all samples are in one cluster.



Similarity measures

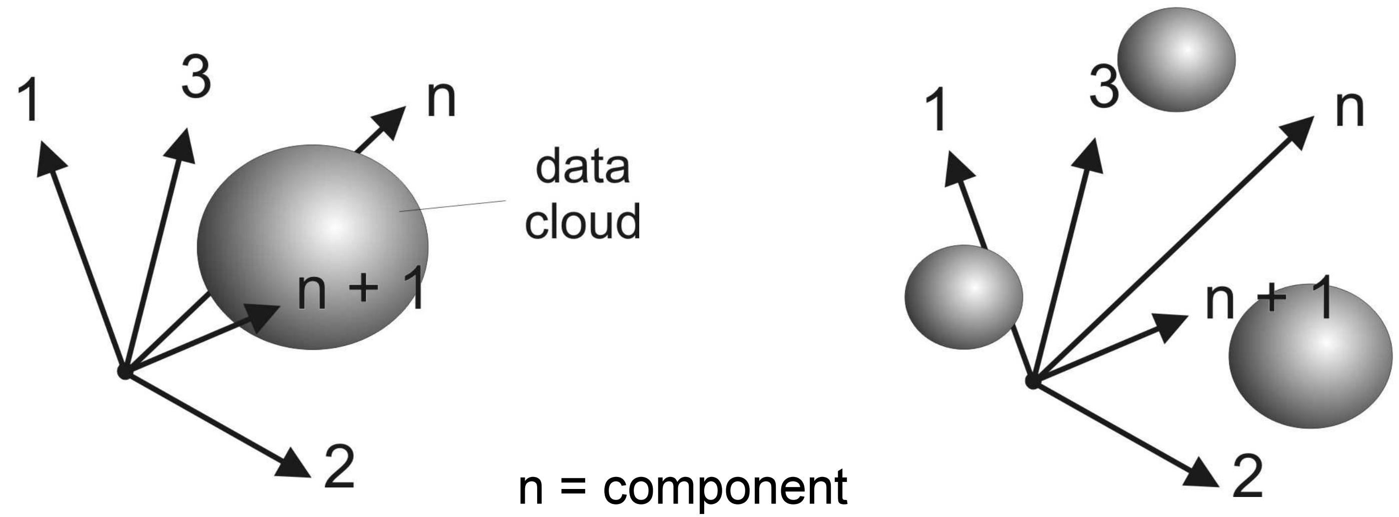
1. Distance coefficients: 2 main types, Euclidian or not (e.g. Manhattan);

2. Correlation similarity coefficient;

3. Association coefficients (only for binary 1-0 data).

1. Distance coefficients

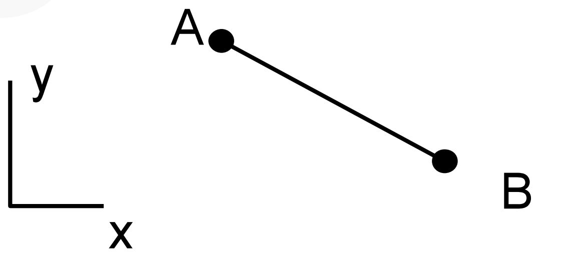
- Data = scatter of points (samples) in a multidimensional space (components of a microfacies) > distance = (dis-)similarity.



n = component

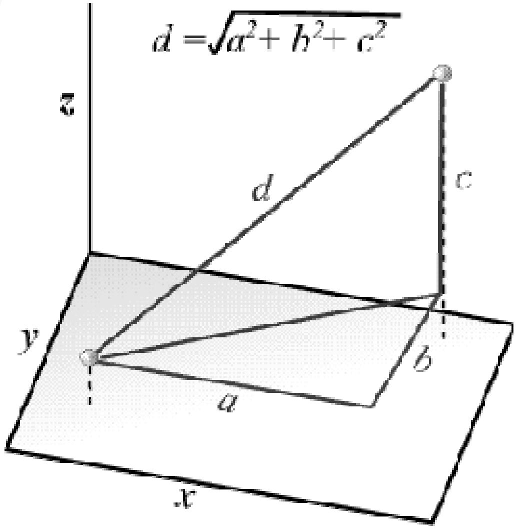
Euclidian = straight line (hypo.)

2D



$$d\ A\text{-}B = \sqrt{(xB-xA)^2 + (yA-yB)^2}$$

3D



mD

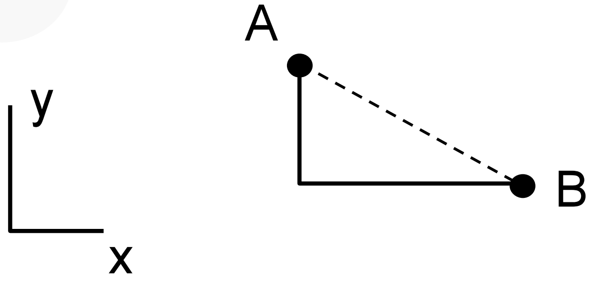
$$d\ A\text{-}B = \sqrt{\sum_{i=1}^{m} (iB-iA)^2}$$

Or to avoid the measure to increase with more variables:

$$d\ A\text{-}B = \sqrt{1/m \sum_{i=1}^{m} (iB-iA)^2}$$

Manhattan = sum

2D



$$d_{A-B} = 1/2 \; | x_B - x_A | + | y_A - y_B |$$

mD

$$d_{A-B} = 1/m \sum_{i=1}^{m} | i_B - i_A |$$

According to some, more robust to outliers.

Remarks:

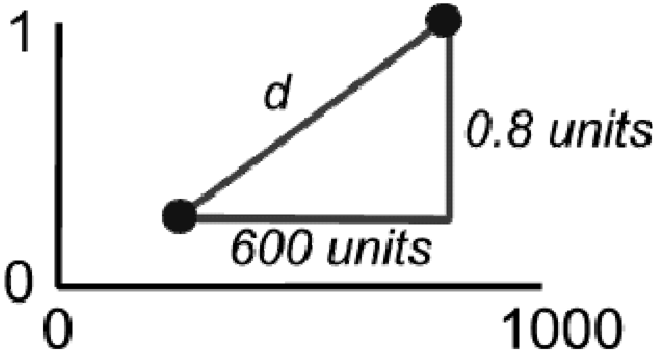
1. Euclidian distance is intuitive but underestimates joint differences, ex. 2 shape characters of an organism should be regarded as due to 2 separate genetic changes, so the real difference between them is the sum of the differences, not the length of the hypothenuse.

↓

So the choice between Euclidian or Manhattan is fct of the independence of variables in the causative process: do 2 differences really mean 2x the difference or just 2 linked consequences of 1 difference?

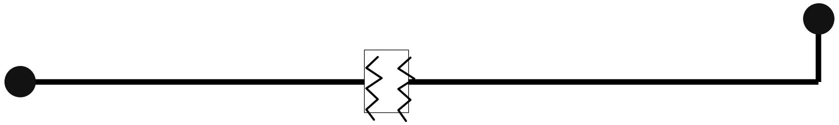
Remarks:

2. Standardisation prior to distance calculation: units / scale.



Euclidian distance = $\sqrt{(600^2 + 0.8^2)}$

= 600,005



ex. force both in 0-1 = 0.6 and 0.8

Even when units are the same, a small variation in one variable might be geologically as important as a large variation in another!

And…

- Distance measures are dependent on the magnitude of the variables, not always desirable…
- Ex.a: 2 fossils may be identical in shape [correlation] but have very different sizes [distances] > in this case we might want to regard similarity in terms of ratios between variable values.
- Ex.b: Two biostratigraphic samples are more similar if the relative proportions of species are similar [correlation] or if abundances (counts) of the species are similar [distances]?

2. Correlation similarity coefficients

- Uses Pearson’s correlation coefficient r but instead of many objects (samples) and 2 variables (components) we have two objects and many variables > scatter plot with axes = samples and data points are variables.

- Standardisation is less important in this case but outliers can affect strongly the results (high or low values in one or two variable).

3. Association coefficients

- For binary data (microfacies, palaeontology);
- A and B are compared on the basis of a contingency matrix:

a to d are number of variables

		sample B	
		present	absent
sample A	present	a	b
	absent	c	d

- There is a large variety of association coefficients calculated on a, b, c and d designed to do well according to various criteria. Here are two common examples:

Jaccard:

$$\boxed{J_{AB} = \frac{a}{a + b + c}}$$

→ Joint absences (d) are not considered as indicative of similarity

Dice-Sorensen:

$$\boxed{D_{AB} = \frac{2a}{2a + b + c}}$$

→ More weight is given to joint-presences

In PAST 1.33

- Various measures are proposed to build the matrix of similarity:
  - Euclidian (robust) and Manhattan;
  - Correlation using r;
  - Dice-Sorensen, Jaccard, Simpson, Raup-Crick for presence/absence;
  - Various for abundances (Bray-Curtis, Cosine, Chord, Morisita, Horn);
  - Hamming for categorical data.

Clustering algorithms

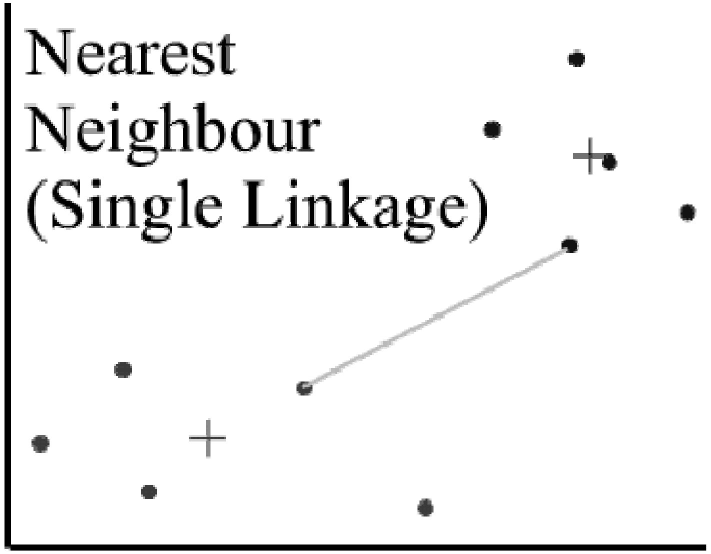
1. Divise methods = find the sparse areas for positioning boundaries between clusters;

2. Density methods = multivariate space is searched for concentrations of points;

3. Linkage methods = nearby points are iteratively linked together.

Common methods (linkage)

- A. Nearest-neighbour = single linkage:

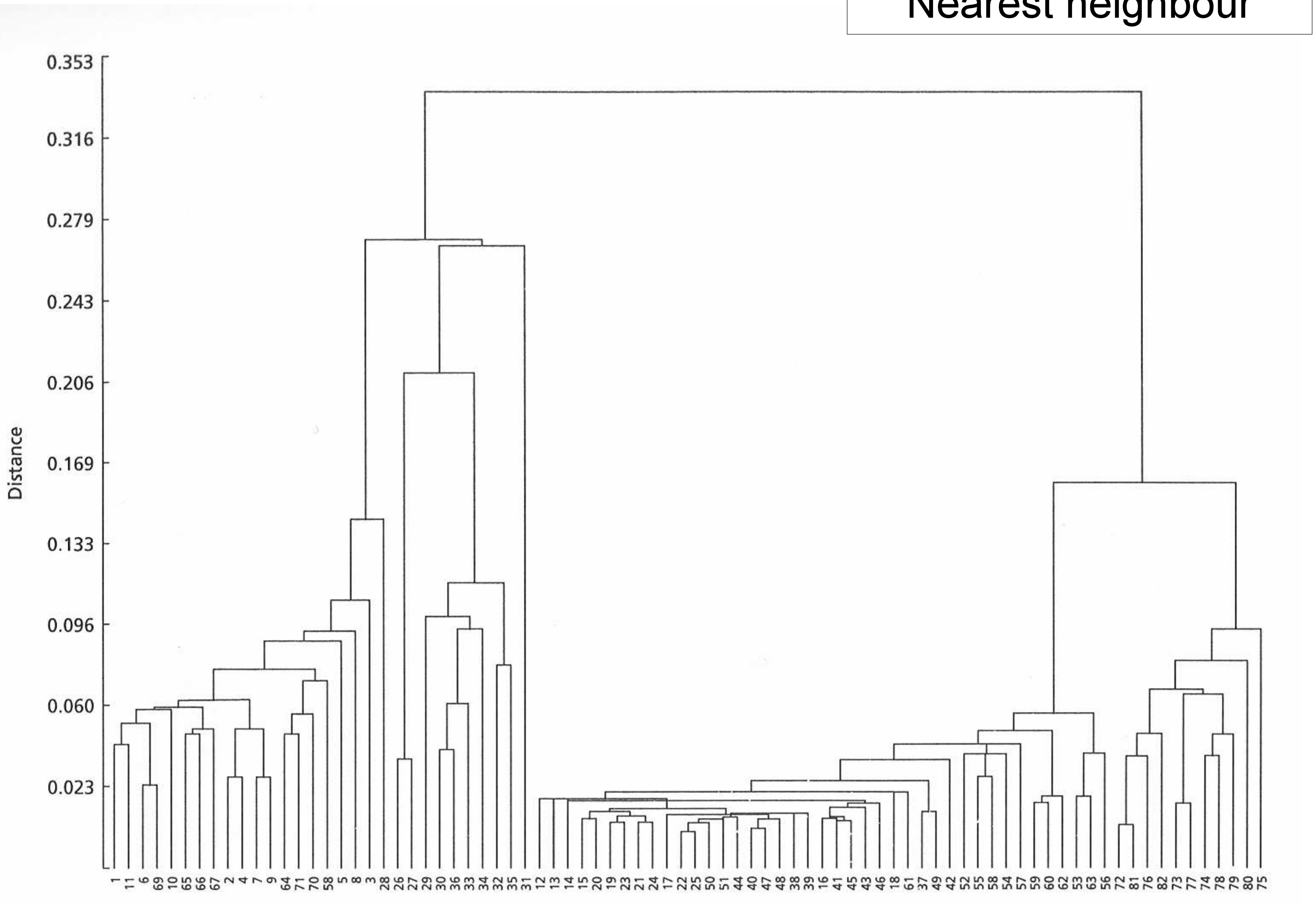


similarity between one point and a new cluster (or 2 clusters) = similarity between that point and the most similar point in the cluster

less than true distance for most points so easy for points to link on to the ends of dispersed, elongated clusters with points at oppsite ends substantially different

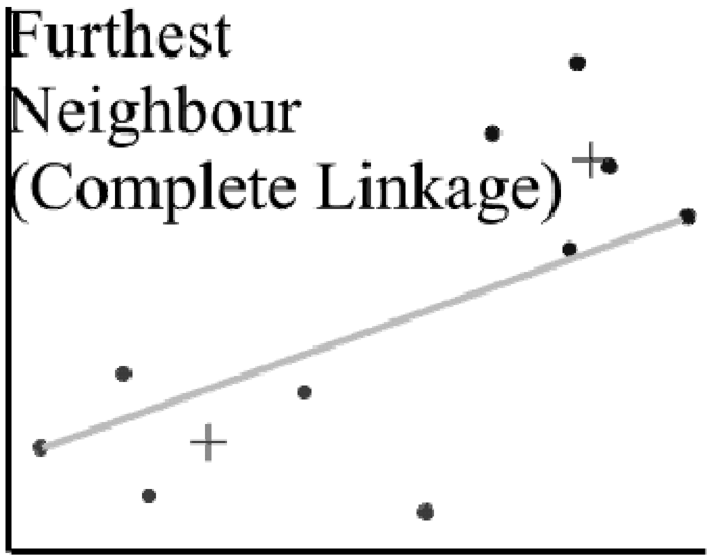
(has been widely used in numerical taxonomy)

Nearest neighbour



Common methods (linkage)

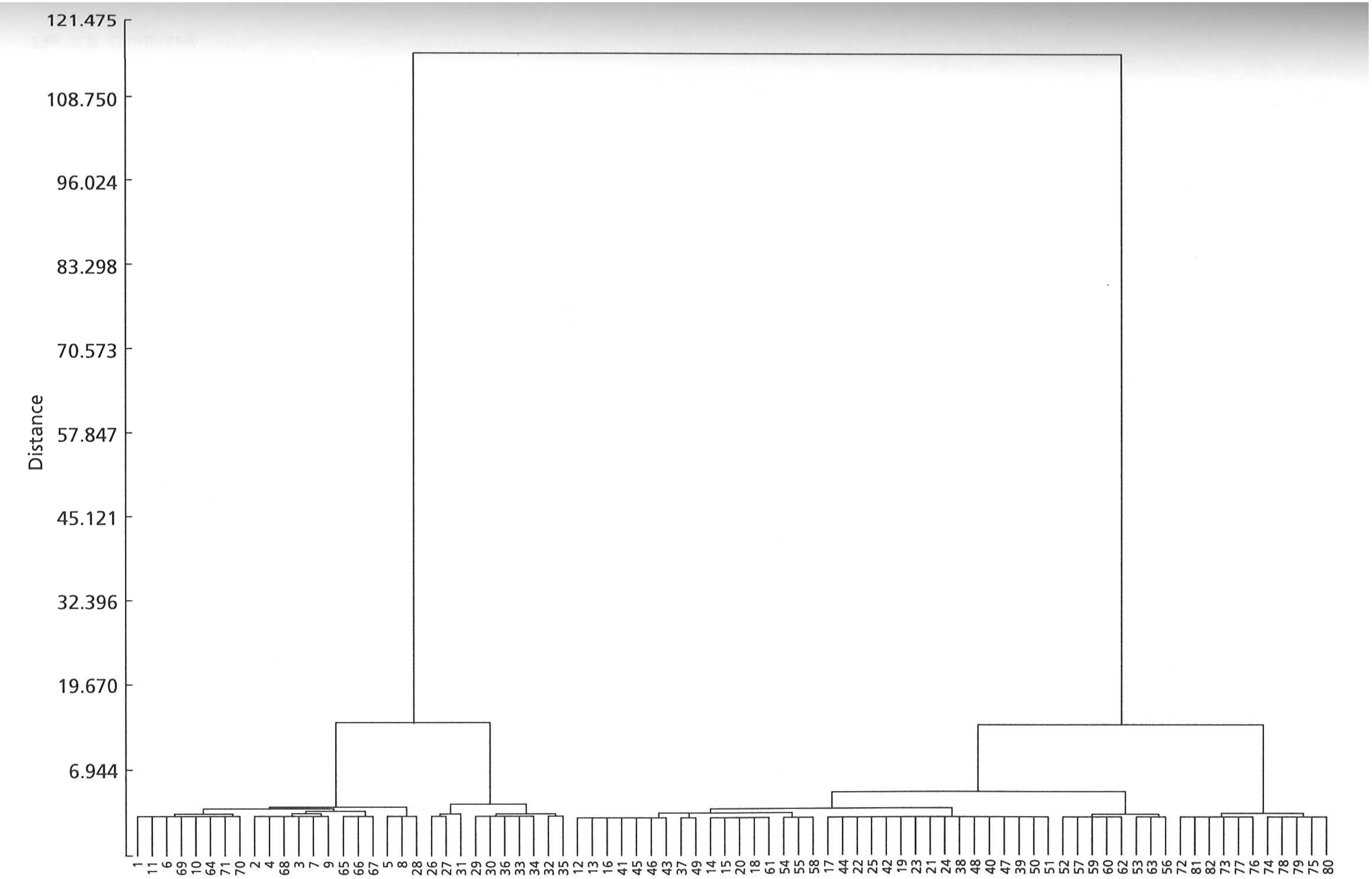
B. Furthest-neighbour = complete linkage:



similarity between one point and a new cluster (or 2 clusters) = weakest of all candidate pairwise similarities, greatest distance

apparent interclusters distances maximised, tends to produce very tight clusters of similar cases, sometimes breaking up ‘too far’

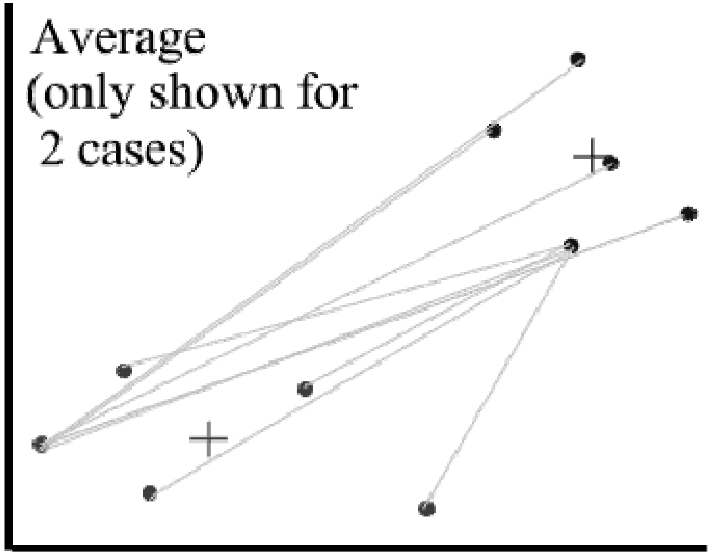
Furthest neighbour



Common methods (linkage)

C. Average linkage:

similarity between one point and a new cluster (or 2 clusters) = average (many different ways)



- Most common: Unweighted Pair-Groups Method Average (UPGMA) = average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form the new cluster.

Common methods (linkage)

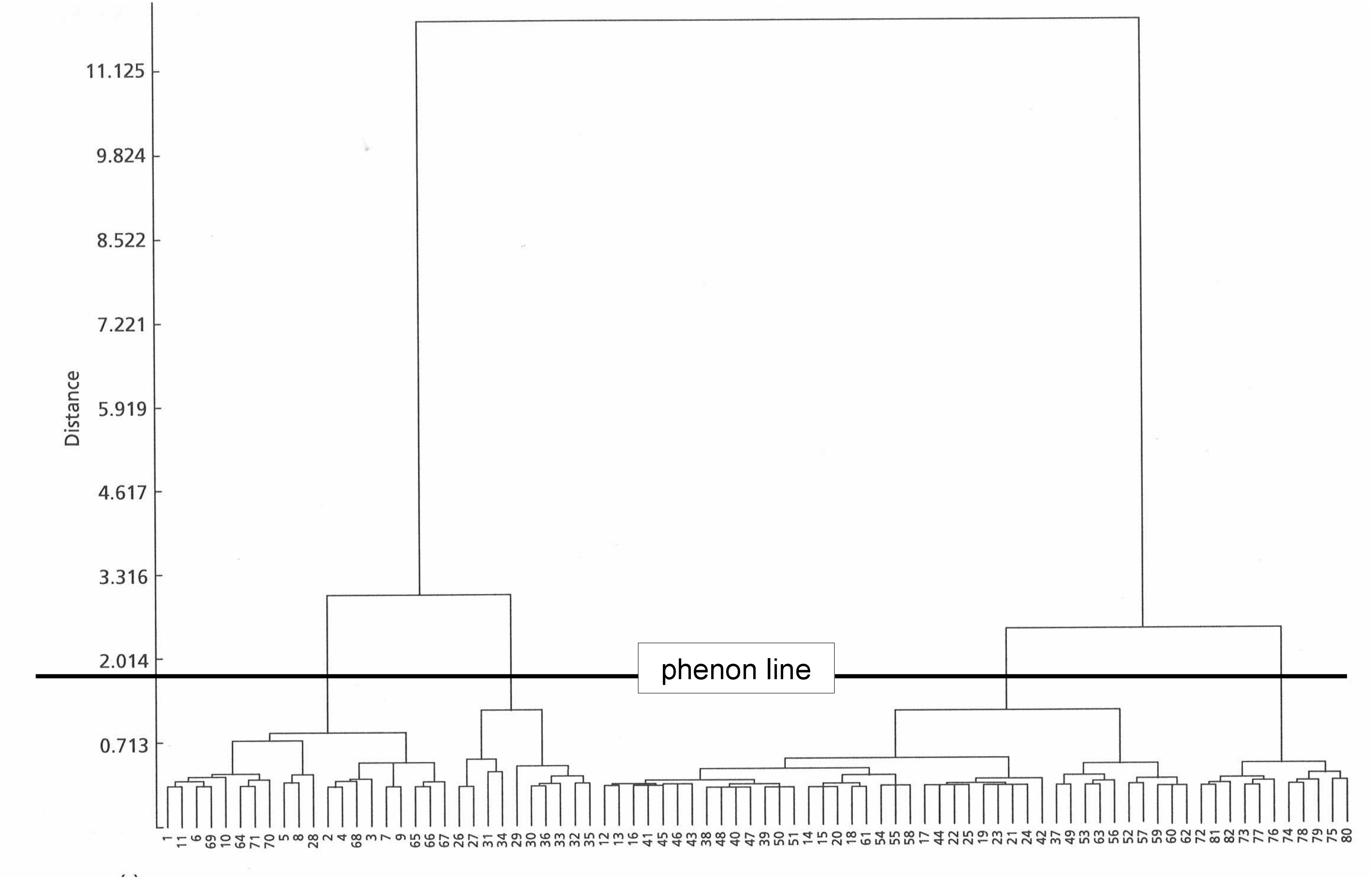
- D. Ward’s method:

Linkage such that there is the least increase in the sum of squared deviations from the cluster means in order to control the increase in variance of clusters during linkage. The criterion for fusion is that it should produce the smallest possible increase in the error sum of squares.

Good looking and well-proportioned so became *de facto* standard…

Works only with euclidian distance.

Ward’s method



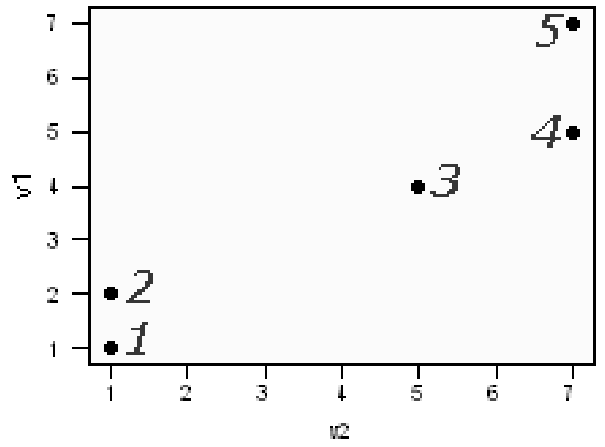
Dendrogram

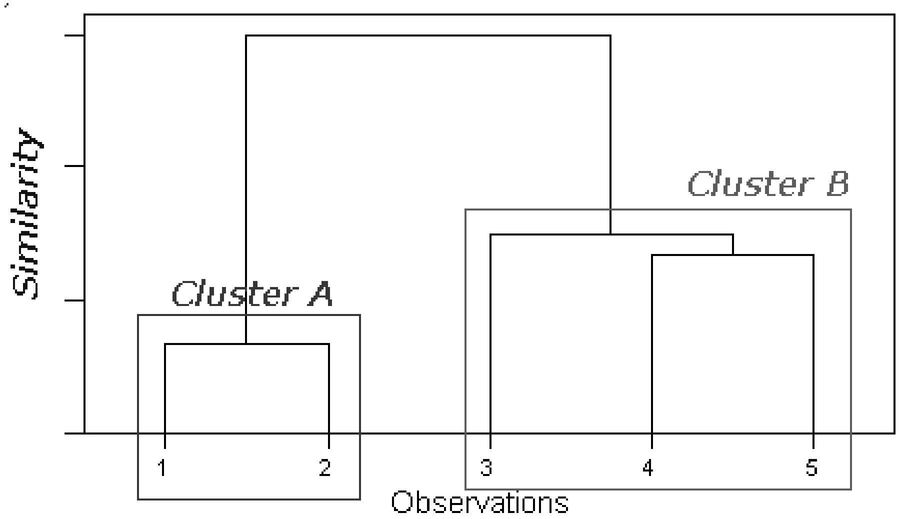
- Result of the analysis = ordered series of linkages between clusters, each at a specific magnitude of similarity. Best represented graphically by a dendrogram;

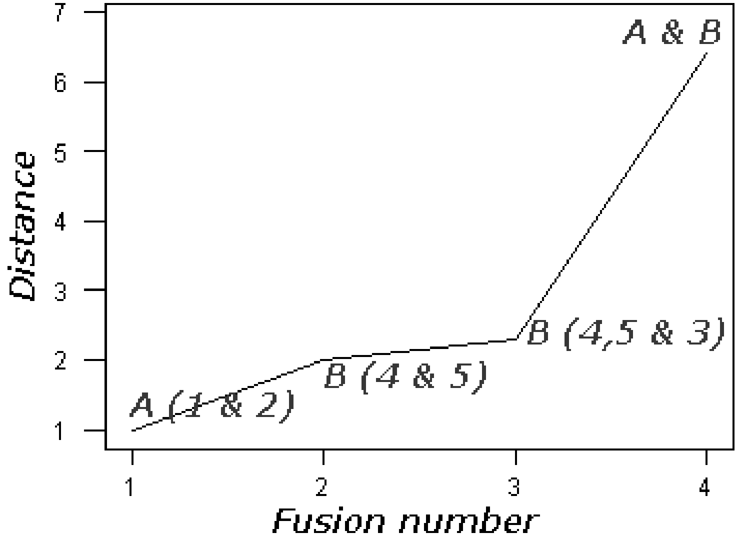
- The phenon line cuts the structure at a chosen level to isolate meaningful clusters. Indeed all clusters will be linked ultimately by the method;

- Where to draw that line is based on: pragmatic requirements, preconceptions (if the number of categories is not itself under investigation) and ‘natural’ divisions if they exist (gaps, jumps).

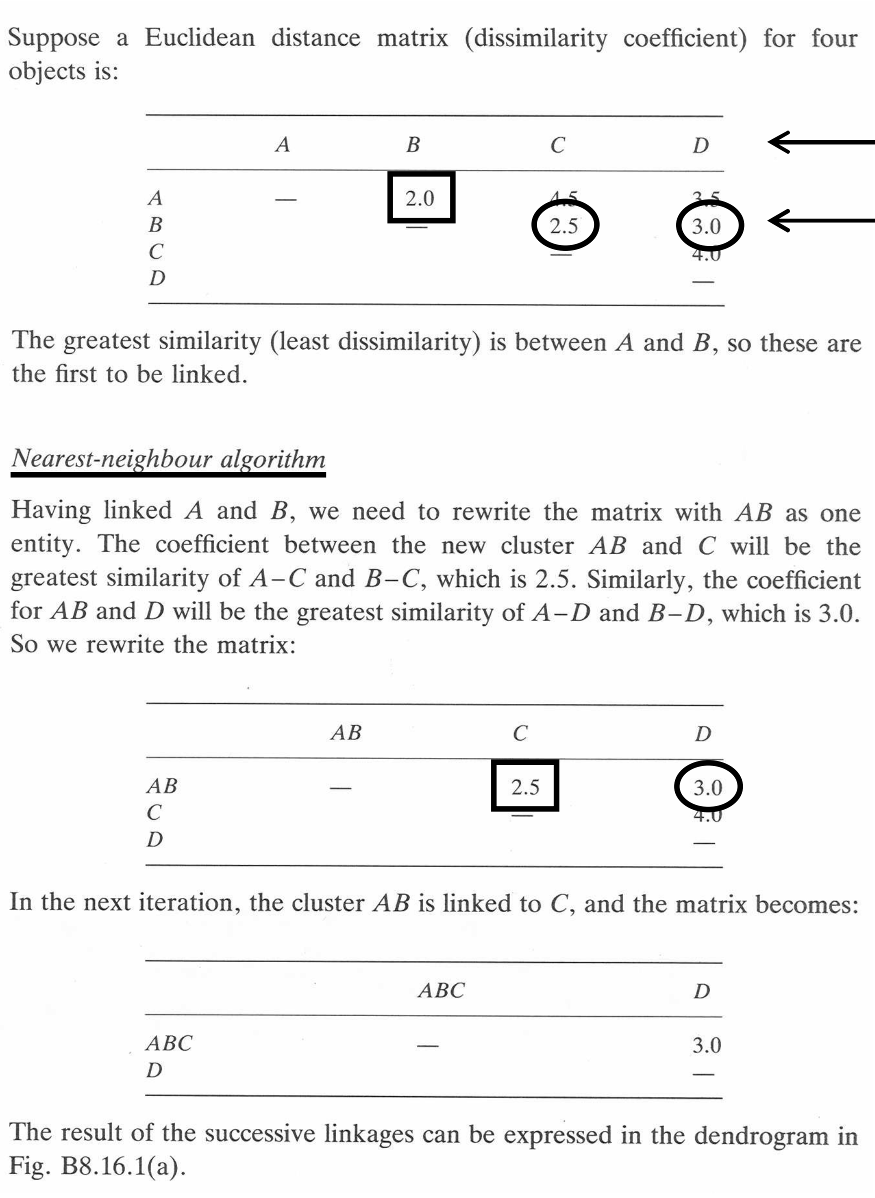
‘natural division’







Example



Suppose a Euclidean distance matrix (dissimilarity coefficient) for four objects is:

	A	*B*	*C*	*D*
A	—	2.0	4.5	3.5
B		—	2.5	3.0
C			—	4.0
D				—

The greatest similarity (least dissimilarity) is between *A* and *B*, so these are the first to be linked.

Nearest-neighbour algorithm

Having linked *A* and *B*, we need to rewrite the matrix with *AB* as one entity. The coefficient between the new cluster *AB* and *C* will be the greatest similarity of *A–C* and *B–C*, which is 2.5. Similarly, the coefficient for *AB* and *D* will be the greatest similarity of *A–D* and *B–D*, which is 3.0. So we rewrite the matrix:

	AB	*C*	*D*
AB	—	2.5	3.0
C		—	4.0
D			—

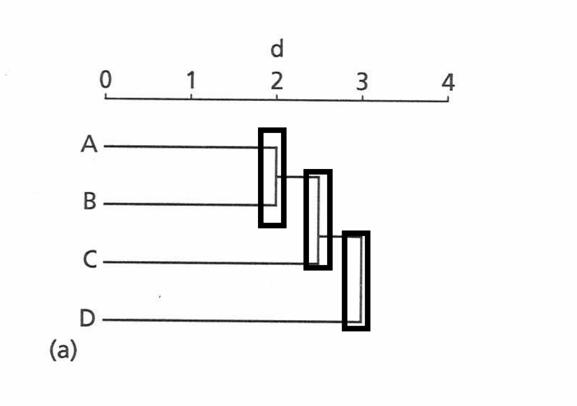
In the next iteration, the cluster *AB* is linked to *C*, and the matrix becomes:

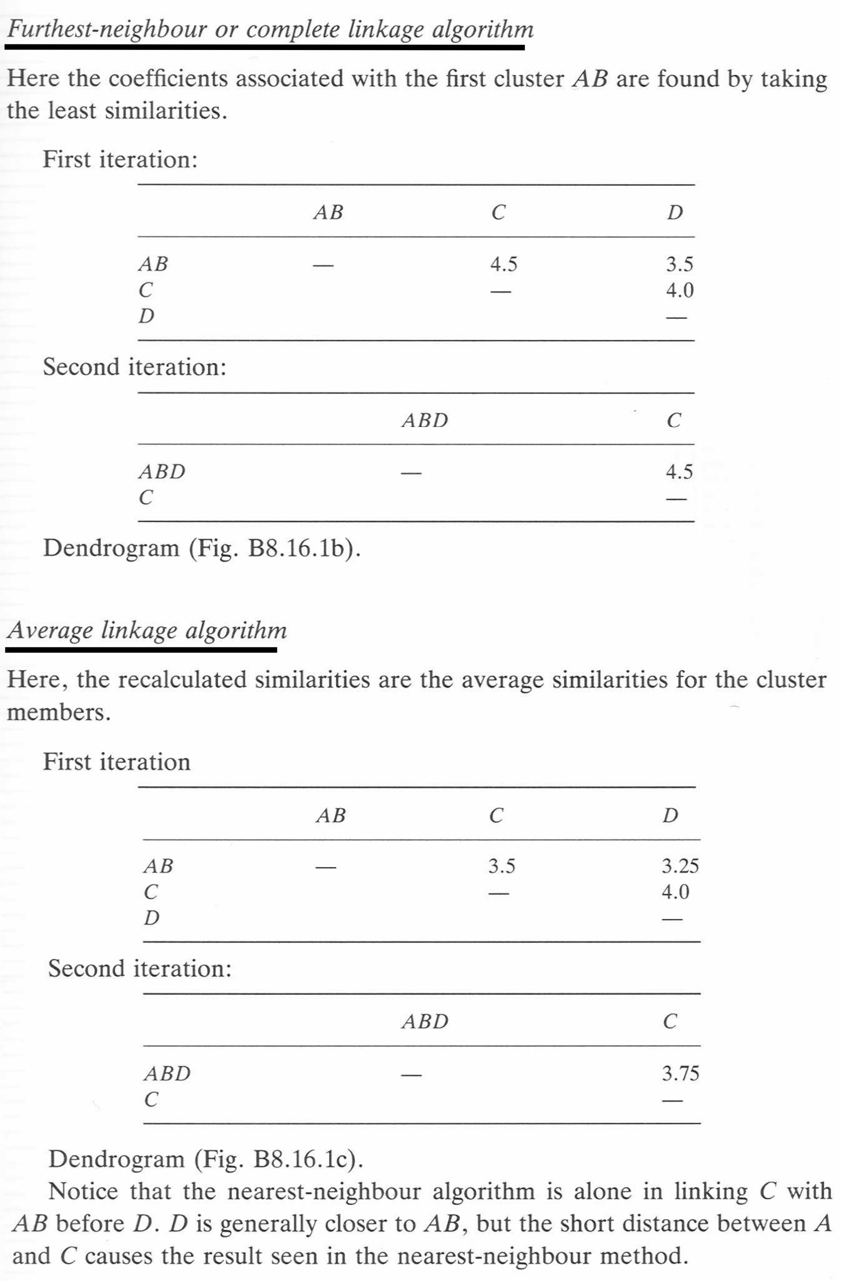
	ABC	*D*
ABC	—	3.0
D		—

The result of the successive linkages can be expressed in the dendrogram in Fig. B8.16.1(a).

samples

euclidian distance in 2 variables space





Furthest-neighbour or complete linkage algorithm

Here the coefficients associated with the first cluster *AB* are found by taking the least similarities.

First iteration:

	AB	*C*	*D*
AB	—	4.5	3.5
C		—	4.0
D			—

Second iteration:

	ABD	*C*
ABD	—	4.5
C		—

Dendrogram (Fig. B8.16.1b).

Average linkage algorithm

Here, the recalculated similarities are the average similarities for the cluster members.

First iteration

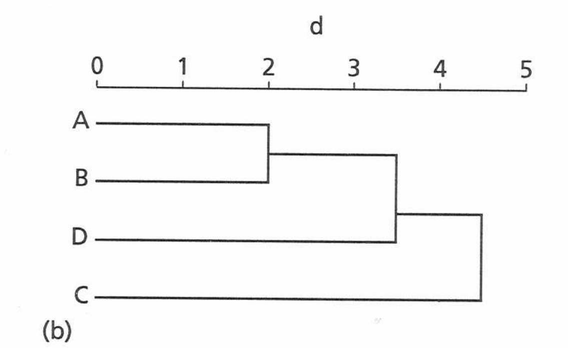
	AB	*C*	*D*
AB	—	3.5	3.25
C		—	4.0
D			—

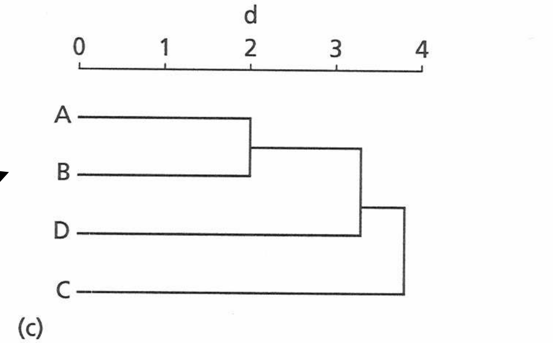
Second iteration:

	ABD	*C*
ABD	—	3.75
C		—

Dendrogram (Fig. B8.16.1c).

Notice that the nearest-neighbour algorithm is alone in linking *C* with *AB* before *D*. *D* is generally closer to *AB*, but the short distance between *A* and *C* causes the result seen in the nearest-neighbour method.





How good is cluster analysis?

- Objective classification but with (most often) subjective choices at many levels; Same data > very different (valid) results;
- New observations will modify the clusters, sometimes strongly > instabilty;
- No available test for difference from random population;
- « profound conclusions should not be based on such uncertain foundations » Swan & Sandilands (1995).
- Test various clustering methods on your data and see if results are comparable!! Remove isolated outliers prior to analysis.
- Average linkage seems to offer the best stability for clusters.

References

- PAST: http://folk.uio.no/ohammer/past/

- Good websites:

  - http://149.170.199.144/multivar/ca.htm
  - http://www.statsoft.com/textbook/stcluan.html
  - http://www2.chass.ncsu.edu/garson/pa765/cluster.htm

Very good reference for data analysis in geology:

Swan, A.R.H. & Sandilands, M. 1995. Introduction to geological data analysis. Blackwell Science.