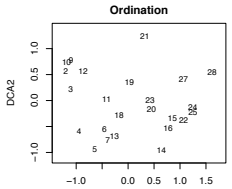
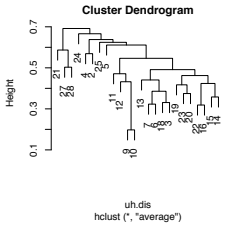
Classification

- Nobody should **want** to make clustering, but they are desperate with multivariate data.
- Reduce data into a few classes and describe these instead of original observations.






```
                    11111111122222222
           2345679012345689012345678
Cal.vul ++5.1..14185352.+.5.+...
Emp.nig 652+665737461676675746 74
Led.pal ..........+......54...42
Vac.myr ......1...1+24.4.76...67
Vac.vit 654566777766757678787776
Vac.uli .4+..6........4.2.4..+31
Dic.sp  .........1+.23..31+.87.1
Dic.fus +2113221114778414185.752
Dic.pol ...+.+111...+1.++5.+4..1
Hyl.spl ................4....+66
Ple.sch 3+4+++537666875774878899
Pol.jun +.+2+1111.+11++4+1.416.+
Pti.cil ++..+.1.3.1.+++1.6+4+1++
Cla.arb 26778823658776866636656 41
Cla.ran 88798956878467857556646 3
Cla.ste 99837199897+1+58+71++.4.
Cla.unc 113152114338531151436 3++
Cla.cor +1+1111111+21+11+++4122+
Cla.cri ++++1+111123311++114++++
Cet.niv .+61+131......++...+....
Ste.sp  .136+3.13.+23+21...4++.+
Cla.def +++11+11.124331232+42+++
```

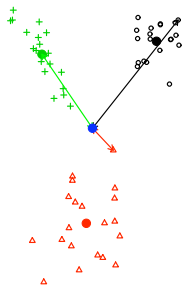

Classification of classification

Formal	$\Longleftrightarrow$	Informal
Hierarchic	$\Longleftrightarrow$	Non-hierarchic
Quantitative	$\Longleftrightarrow$	Qualitative
Divisive	$\Longleftrightarrow$	Agglomerative
Polythetic	$\Longleftrightarrow$	Monothetic

- **Cluster analysis:** Formal, hierarchic, quantitative (usually), agglomerative, polythetic.
- **TWINSPAN:** Formal, hierarchic, semi-quantitative, divisive, polythetic.

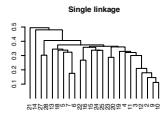
Cluster Analysis

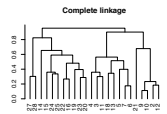
- Agglomerative: Combine two most similar observations, and continue until every point is in the tree.
- Various criteria for similarity between clusters:
  1. Single linkage or distance to the nearest neighbour.
  2. Complete linkage or distance to the furthest neighbour.
  3. Average linkage or distance to the class centroid.

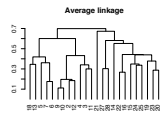


Clustering strategies

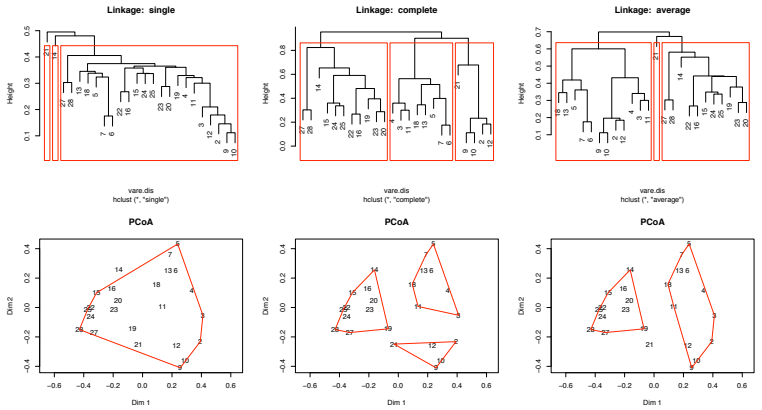
- **Single linkage or nearest neighbour**
  - Finds the *minimum spanning tree*: Shortest tree that connects all points.
  - Finds discontinuities.
  - Chaining: Groups of inequal size.
- **Complete linkage or furthest neighbour**
  - Compact clusters of ±equal size.
  - Makes compact groups even when none exist.
- **Average linkage methods** (e.g. UPGMA)
  - Between single and average linkage.
  - UPGMA minimizes *cophenetic correlation*.





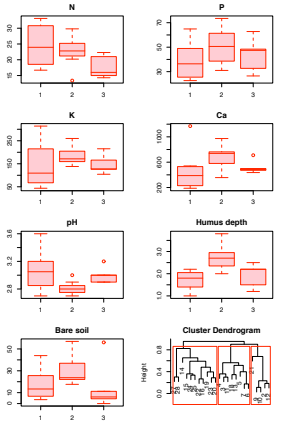


Clustering and space



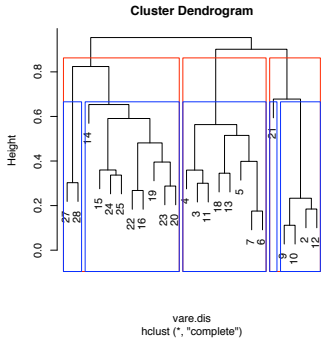
Interpreting clusters

- Predictive clusters are different in their environment.
- Community classification for environmental indication.
- Clustering may detect local peculiarities, whereas (most) ordination methods show the global gradient pattern.



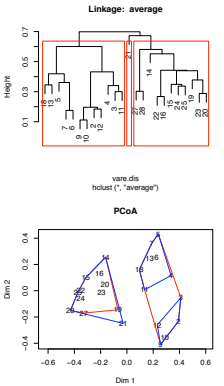
Number of clusters

- As many fusion levels as there are observations: Hierarchic clustering can be cut at any level.
- The scientist usually want to use classes: One level.
- Various optimality criteria doomed to fail: If they are good, they can be made clustering criteria, and then they are just an alternative clustering.



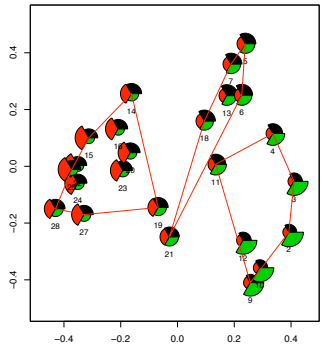
Optimizing classification: K–means clustering

- Agglomerative clustering has a burden of history: Once formed classes cannot be broken although that would be sensible at the chosen level.
- K–means clustering: Iterative procedure for non-hierarchic classification.
- If started with chosen hierarchic clustering, will optimize.
- Best suited with centroid linking, since thinks in that way.



Fuzzy clustering

- Each observation is given a probability profile of class membership.
- Corresponding crisp classification: Class of highest membership probability.
- Non-hierarchic, flat classification.
- Iterative procedure.



TWINSPAN: Two-Way Indicator Species Analysis

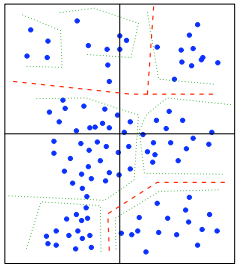
- TWINSPAN is not a method but a program: A bag of tricks.
- Gradient chopping for CA: Ideal if this is the criterion.
- Uses binary data: Trick is to divide each species into a series of 'pseudospecies' by abundance cuts.

Algorithm

1. Get a CA axis on pseudospecies data.
2. Select pseudospecies at the ends of the axis as indicators.
3. Repeat ordination with these pseudospecies: Polarizes the axis.
4. Chop data in two parts in the middle of the axis.
5. Repeat steps for both parts.

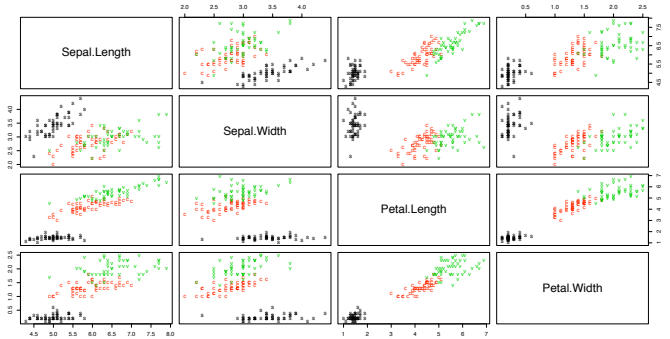
Criteria for good classes

1. Divide environment into equal parts.
2. Compact clusters.
3. Groups of equal size.
4. Discontinuous groups.



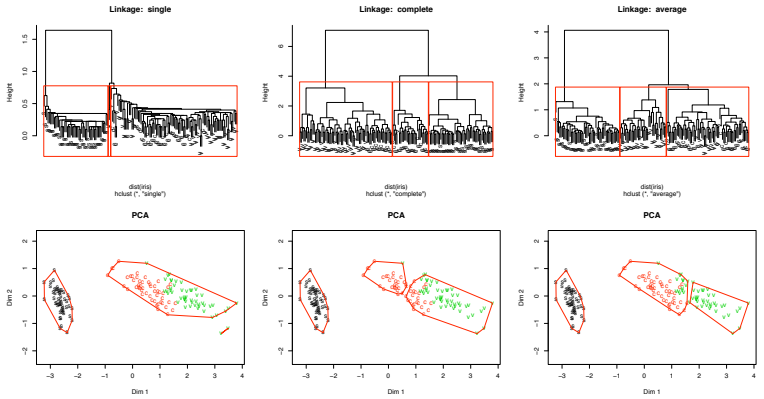
These criteria often in conflict, and cannot be satisfied simultaneously.

Example: A real class structure...



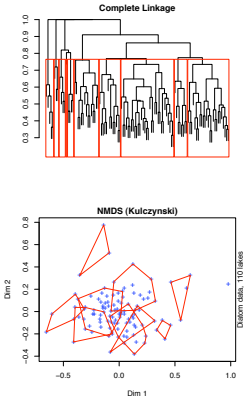
Data on *Iris setosa* (s), *I. versicolor* (c) and *I. virginica* (v).

. . . And all methods fail



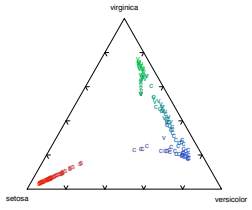
Classification and ordination

- Formerly classification and 'continuum' theoretical ordination were seen as opposites: Only two alternative ways of simplification of multivariate data.
- If classes distinct in ordination, results (or methods!) are consistent.
- Inconsistent results:
  - Either or both results bad.
  - Different criteria.
  - Too few dimensions in ordination.



The choice of clustering method

- Some opt for single linkage: Finds distinct clusters, but prone to chaining and sensitive to sampling pattern.
- Most opt for average linkage methods: Chops environment more evenly.
- All dependent on dissimilarity measure: Should be ecologically meaningful.
- Small changes in data can cause huge visual change in clustering: Classification may be optimized for the chosen level.
- TWINSPAN too unstable and tricky: Better avoided.



Fuzzy clustering may fail as well, but at least shows the uncertainty.