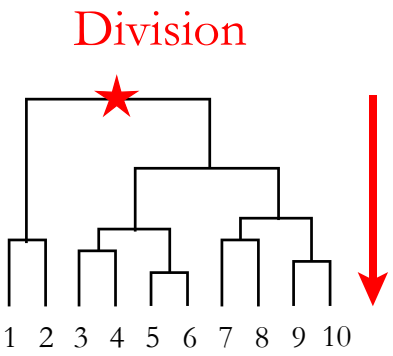
Polythetic Divisive Hierarchical Clustering

- PDHC techniques use the information on all the variables.
- PDHC begins with all entities together in a single cluster and successively divide the entities into a hierarchy of smaller and smaller clusters until, finally, each cluster contains only one entity or some specified number of entities.
- There are only a few different PDHC techniques, perhaps due to the computational burdens of a truly divisive approach (2^{n-1}-1 possible partitions of the samples into two groups).



If n = 100, 6.34×10^{29} possibilities

Polythetic Divisive Hierarchical Clustering

Ordination Space Partitioning

- Ordination is first used to position entities in low dimensional ordination space.
- Ordination scores are partitioned:
  - *Gauch (1982)* – groups delineated by drawing boundaries on ordination plots by hand.
  - *Williams (1976) and Lefkovitch (1976)* – ordination axis (preferably PCO) divided by assigning samples to one group or a second according to their sign on the axis (i.e., dividing the axis in half).
  - *Pielou (1984)* – ordination axis partitioned by breaking it into segments that maximize the ratio of between- to within-segment variance or dissimilarity.

Polythetic Divisive Hierarchical Clustering

Two-Way Indicator Species Analysis (TWINSPAN)

1. RA is used to position entities in low dimensional ordination space.
2. Those "*indicator*" species that characterize the RA axis extremes are emphasized in order to polarize the samples. These species are assigned higher weights – positive or negative depending on which end of the axis they fall on, and the ordination is recomputed using these weights. The weighting effectively polarizes the ordination with respect to the indicator species.
3. The samples are divided into two clusters by breaking the ordination axis near its middle, at a natural discontinuity forced by the polar weightings.

Polythetic Divisive Hierarchical Clustering

Two-Way Indicator Species Analysis (TWINSPAN)

4. The division process is then repeated on the two sample subsets to give four clusters, and so on, until each cluster has no more than a chosen minimum number of members.
5. A corresponding species clustering is produced, and the sample and species hierarchical clusterings are used together to produced an *arranged data matrix*.
6. The resulting sample hierarchy (and species hierarchy) may also be displayed as a *dendrogram*, using the sequences of divisions as integral levels or computing the levels as the average distances between samples in ordination space.

Polythetic Divisive Hierarchical Clustering

Two-Way Indicator Species Analysis (TWINSPAN)

Pseudospecies:

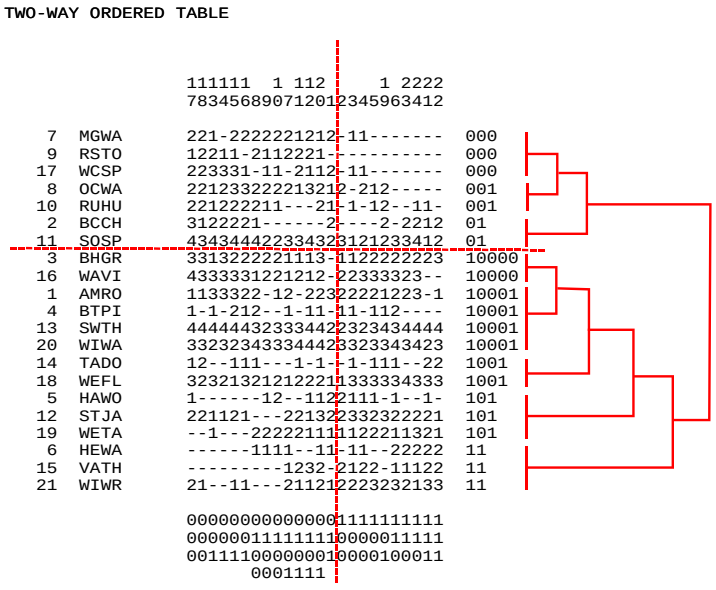
- TWINSPAN approximates quantitative abundance data by creating a variable number of "*psuedospecies*" representing abundance classes. Psuedospecies "*cut levels*" are used to define the ranges of the abundance classes.

Site	Species A	Psuedospecies A1	A2	A3	A4	A5
1	15	1	1	1	0	0
2	0	1	0	0	0	0
3	5	1	1	0	0	0
4	33	1	1	1	1	0
5	82	1	1	1	1	1

Cut Levels:
0 5 10 20 40

Polythetic Divisive Hierarchical Clustering

Two-Way Indicator Species Analysis (TWINSPAN)



Polythetic Divisive Hierarchical Clustering

Two-Way Indicator Species Analysis (TWINSPAN)

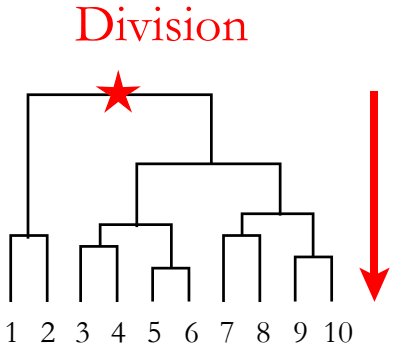
Limitations:

- Solution is based on a Correspondence Analysis (or Reciprocal Averaging) ordination and so it may suffer from the same weaknesses as CA(RA).
- TWINSPAN most amenable to data with a strong and unidimensional underlying gradient.

Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering

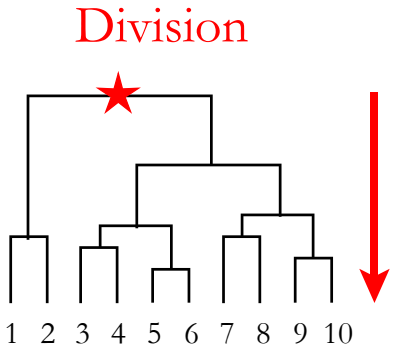
- DIANA...divisive hierarchical *polythetic* clustering algorithm suitable for any dissimilarity matrix
- MONA...divisive hierarchical *monothetic* clustering algorithm designed for binary variables (e.g., presence/absence data).



Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering (DIANA)

- The *diana* algorithm constructs a hierarchy of clusterings, starting with one large cluster containing all *n* observations. Clusters are divided until each cluster contains only a single observation.
- At each stage, the cluster with the largest diameter is selected. (The diameter of a cluster is the largest dissimilarity between any two of its observations.)



Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering (DIANA)

- To divide the selected cluster, the algorithm first looks for its most disparate observation (i.e., which has the largest *average* dissimilarity to the other observations of the selected cluster). This observation initiates the '*splinter group*'.

	Sites						
Sites	1	2	3	4	5	6	*mean*
1	0	1.4	9.7	15.9	15.1	13.7	*11.16*
2	1.4	0	9.3	15.2	14.4	12.7	*10.6*
3	9.7	9.3	0	10.9	10	13.8	*10.74*
4	15.9	15.2	10.9	0	2.2	8.2	*10.48*
5	15.1	14.4	10	2.2	0	8.3	*10*
6	13.7	12.7	13.8	8.2	8.3	0	*11.34*

Split off site #6

Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering (DIANA)

- In subsequent steps, the algorithm reassigns observations that are closer to the '*splinter group*' than to the '*old party*'. The result is a division of the selected cluster into two new clusters.

	Clusters	
Sites	1-5	6
1	10.5	13.7
2	10.1	12.7
3	10.0	13.8
4	11.1	8.2
5	10.4	8.3

Clusters: (1,2,3) and (4,5,6)

- Distance between clusters is given by the *maximum* distance between entities in the two clusters.

Sites	1	2	3	4	5	6
1	0	1.4	9.7	15.9	15.1	13.7
2	1.4	0	9.3	15.2	14.4	12.7
3	9.7	9.3	0	10.9	10	13.8
4	15.9	15.2	10.9	0	2.2	8.2
5	15.1	14.4	10	2.2	0	8.3
6	13.7	12.7	13.8	8.2	8.3	0

Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering (DIANA)

	Sites			
Sites	1	2	3	*mean*
1		1.4	9.7	*5.55*
2	1.4		9.3	*5.35*
3	9.7	9.3		*9.5*

Split off site #3

	Clusters	
Sites	1-2	3
1	1.4	9.7
2	1.4	9.3

Clusters: (1,2) and (3)

	Sites			
Sites	4	5	6	*mean*
4		2.2	8.2	*5.2*
5	2.2		8.3	*5.25*
6	8.2	8.3		*8.25*

Split off site #6

	Clusters	
Sites	4-5	6
4	2.2	8.2
5	2.2	8.3

Clusters: (4,5) and (6)

Polythetic Divisive Hierarchical Clustering

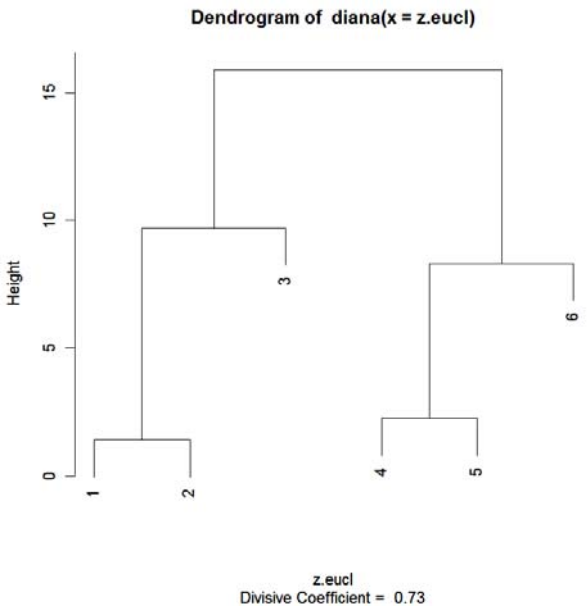
Divisive Hierarchical Clustering (DIANA)

Sites	1	2
1		1.4
2	1.4	

Split 1 and 2

Sites	4	5
4		2.2
5	2.2	

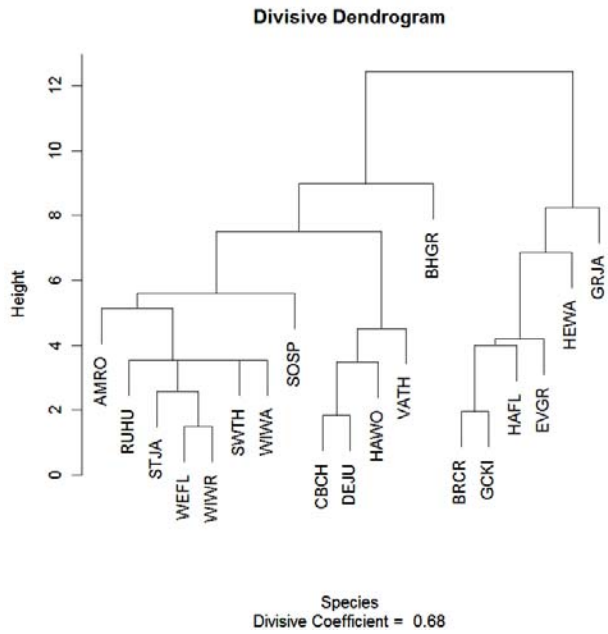
Split 4 and 5



Polythetic Divisive Hierarchical Clustering

Divisive Hierarchical Clustering (DIANA)

- Divisions are based on *average* distances (similar to 'average-linkage'), but cophenetic distance is based on *maximum* distances between entities in the two subclusters (similar to 'complete linkage').

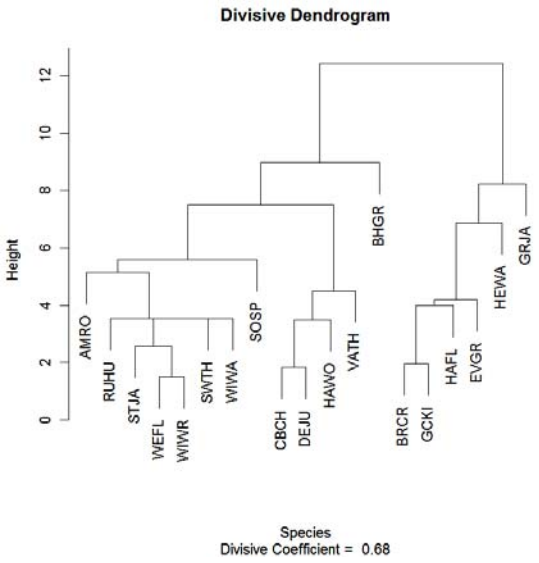


Polythetic Divisive Hierarchical Clustering

Evaluating the Cluster Solution

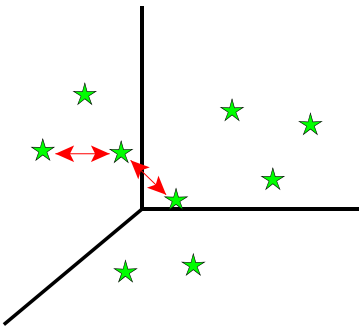
Divisive Coefficient

- *Divisive coefficient* (cluster library) is a measure of the clustering structure of the dataset.
- For each observation i, denote by *d*(i) the diameter of the last cluster to which it belongs (before being split off as a single observation), divided by the diameter of the whole dataset. The *DC* is the average of all 1 - *d*(i).

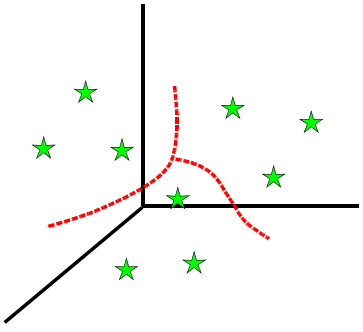


Relationship Between PAHC and PDHC

- PAHC techniques begin by examining small distances between similar samples, yet these small distances are likely to be a reflection of noise than anything else.

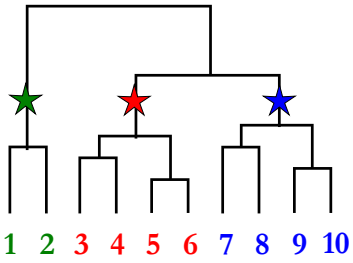


- PDHC techniques begin by examining overall, major gradients in the data; all the available information is used to make the critical topmost divisions.



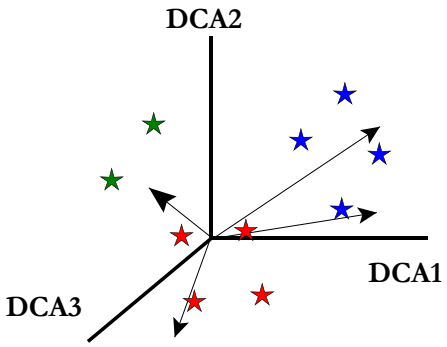
Relationship Between PAHC and PDHC

- PAHC techniques rely on subsequent analysis to describe ecological differences among derived clusters.



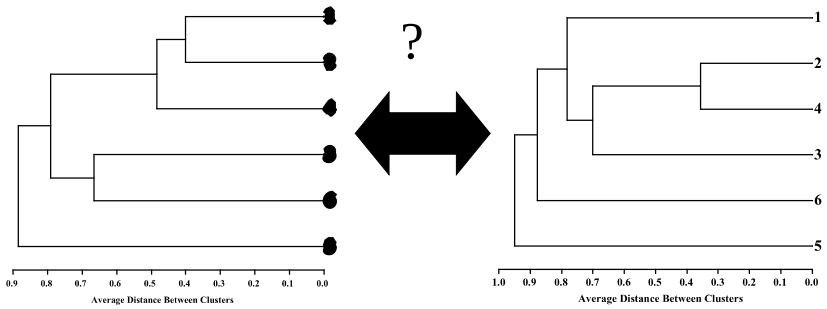
  - Group means
  - ANOVA
  - Discriminant Analysis

- PDHC techniques (e.g., OSP, TWINSPAN) often hybridize ordination and cluster analysis; which combines the power of clustering for summarization with the effectiveness of ordination in revealing directions of relationship.



Evaluating the Stability of the Cluster Solution

1. Compare solutions among alternative clustering algorithms
   - Try PAHC and PDHC where appropriate.
   - Try different resemblance/fusion strategies.



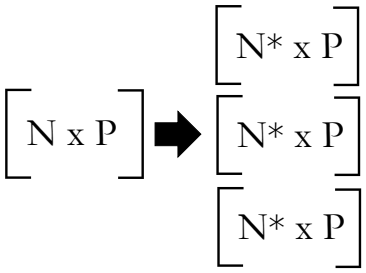
Evaluating the Stability of the Cluster Solution

2. Data Resampling

- Resample data set via *bootstrapping* and perform cluster analysis on each bootstrap sample separately.
  - Bootstrap cluster membership should agree with original membership when the data is clearly structured (e.g., Jaccard similarities of the original clusters to the most similar clusters in the resampled data, after dropping any entities not in the bootstrap sample).

Rule of thumb:

$JD_i < 0.5$ *'dissolved' (unstable)*
$JD_i > 0.75$ *'recovered' (stable)*



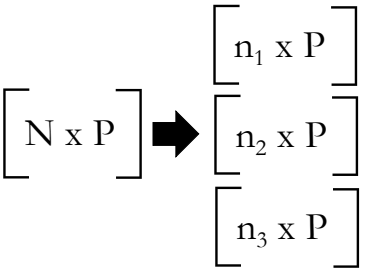
Evaluating the Stability of the Cluster Solution

3. Data Subsetting

- Randomly subset data set and perform cluster analysis on each subset sample separately.
  - Subset cluster membership should agree with original membership when the data is clearly structured (e.g., Jaccard similarities of the original clusters to the most similar clusters in the subset, after dropping any entities not in the subset sample).

Rule of thumb:

$JD_i < 0.5$ *'dissolved' (unstable)*
$JD_i > 0.75$ *'recovered' (stable)*



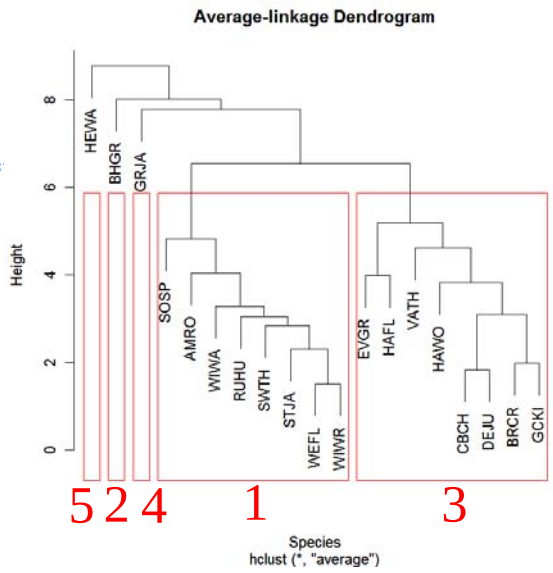
Evaluating the Stability of the Cluster Solution

3. Data Resampling & Subsetting

```
* Cluster stability assessment *
Cluster method:  hclust
Full clustering results are given as parameter result
of the clusterboot object, which also provides further s
of the resampling results.
Number of resampling runs:  100

Number of clusters found in data:  5

 Clusterwise Jaccard bootstrap mean:
[1] 0.8647738 0.5900000 0.7256259 0.6200000 0.6295000
dissolved:
[1] 11 42 20 38 42
recovered:
[1] 77 58 41 62 58
 Clusterwise Jaccard subsetting mean:
[1] 0.8316667 0.4400000 0.6707143 0.4800000 0.4683333
dissolved:
[1]  7 56 32 52 54
recovered:
[1] 60 44 22 48 46
```



Evaluating the Stability of the Cluster Solution

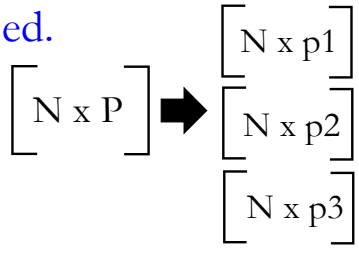
4. Variable Splitting

- Repeat cluster analysis using subsets of variables and compare the results.
  - Variables might be divided into logical subsets based on measurement scale or some other natural division or random subsets of variables could be created.
  - Deletion of a small number of variables from the analysis should not, in most cases, alter greatly the clusters found, if the clusters are "real" and not mere artifacts of the particular technique used.

Rule of thumb:

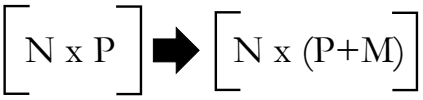
$JD_i < 0.5$ *'dissolved' (unstable)*

$JD_i > 0.75$ *'recovered' (stable)*



Evaluating the Stability of the Cluster Solution

5. Variable Adding
   - Compare the original cluster solution with the solution using additional variables of interest which were NOT included in the original analysis.
     - Perhaps variables gathered during subsequent research.
     - If differences between clusters persist with respect to these new variables then this is some evidence that a "useful" solution has been obtained, in the sense that by stating that a particular entity belongs to a particular cluster we convey information on variables other than those used to produce the cluster.



Limitations of Cluster Analysis

- Some clustering techniques (especially PDHC) are sensitive to the presence of *outliers*.
  - Careful screening for outliers in these cases.
- Most clustering procedures are biased towards finding clusters of a particular shape and are usually biased specifically towards finding spherical clusters.
  - Compare results among several different clustering techniques. General agreement in cluster solutions probably indicates a pronounced structure to the data. Disagreement probably indicates less clear structure to the data (e.g., nonspherical or fuzzy clusters, or no "real" clusters at all), and perhaps that each procedures bias is governing the particular cluster solution found.

Limitations of Cluster Analysis

- Often difficult to judge, from the results of clustering, the "*realness*" of the clusters or the number of clusters suitable for the representation of the data matrix.
  - Ultimately, the validiy of clusters can only be judged qualitatively, by subjective evaluation and by interpretability.
- Because cluster analysis involves choosing among a vast array of different procedures and alternative measures (e.g., choice of resemblance measure and fusion strategy), its application is *more an art than a science* and it can easily be abused.
  - Be aware of this fact and, if possible, replicate the analysis under varying conditions and employ different clustering strategies to ensure reliable results.

Complementary Use of Ordination and Cluster Analysis