OPEN SPAT

Erasmus+

INSTITUTO SUPERIOR D AGRONOMIA
Hinc Patriam Sustinet
Universidade de Lisboa

Pattern recognition on spatial data

Finding structures : numerical classification

Pr Yves Brostaux, GxABT, University of Liege (Belgium)

OpenSpat, ULisboa

June 2019

LIÈGE université Gembloux Agro-Bio Tech

Contents

Contents

Introduction

- Introduction
- Example data

Introduction

Objective

- divide n individuals into k groups, based on their similarities on the values of p variables

Introduction

Data

- Raw data : matrix of numerical data
  - n rows $\equiv$ individuals
  - p columns $\equiv$ variables
- Standardized data : for each variable j
  - $x_{ij} = (y_{ij} - \bar{y}_j)/\hat{\sigma}_j$
  - $\bar{x}_j = 0$
  - $\hat{\sigma}_{x_j} = 1$

Introduction

- Introduction
- Example data

Example

Raw data : Cations and anions in mineral waters

HCO_3^-	Ca^{2+}
SO_4^{2-}	Mg^{2+}
Cl^-	Na^+

Example

```
# set working directory
setwd("....")
# source supplementary function
source("hclust.info.R")

# read data file
waters <- read.csv2("ClassEaux.csv", row.names=7)

# select 7 mineral waters
waters7 <- subset(waters, rownames(waters) %in%
 c("G", "L", "O", "S", "T", "c", "o"))
```

Raw data

	HCO3	SO4	CL	CA	MG	NA	NOM
G	357	10	2	78	24	5	EVIAN (F)
L	398	218	15	157	35	8	ONDINE (F)
O	11	65	5	4	1	3	SPA (B)
S	402	306	15	202	36	3	VITTEL (F)
T	64	7	8	10	6	8	VOLVIC (F)
c	1580	112	137	89	104	425	APOLLINARIS (D)
o	251	32	14	81	11	6	SPONTIN (B)

Scale variables

Important Unlike most of PCA tools, clustering methods don't standardize the data internally. It has to be done manually using them for clustering.

```
# scale variables (mean=0 and SD=1)
waters7std <- scale(waters7[1:6])
```

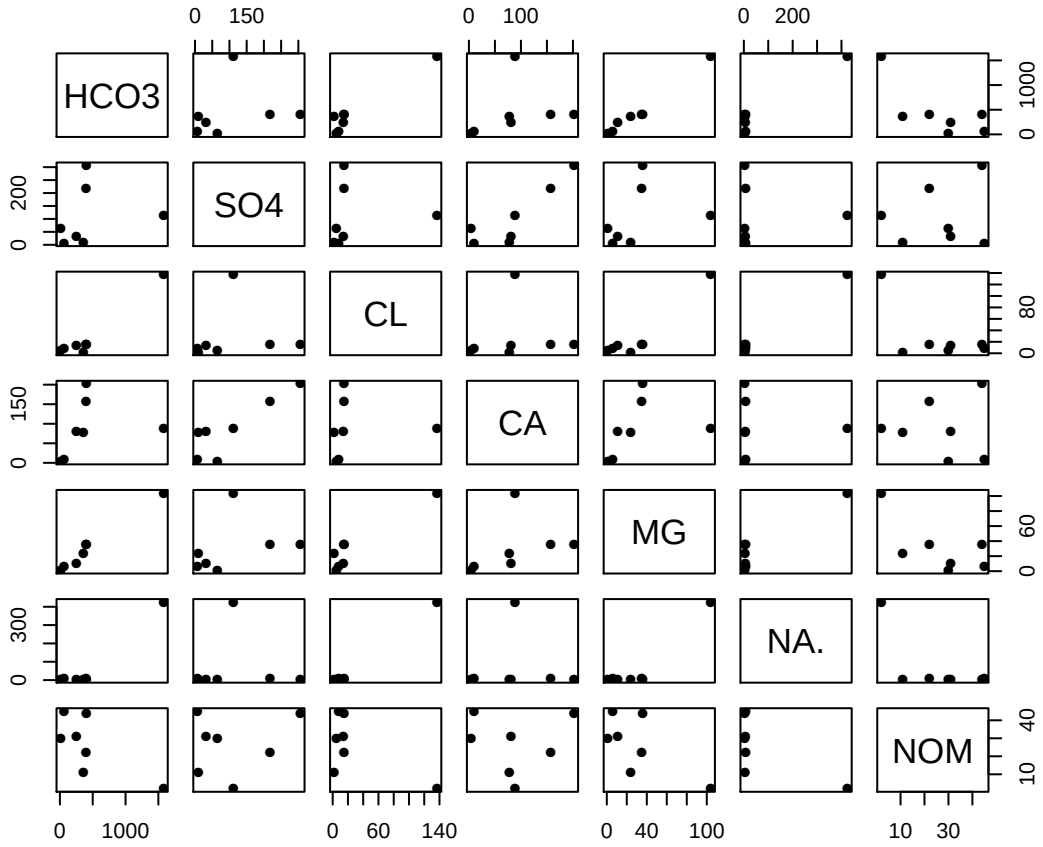
Standardized data

	HCO3	SO4	CL	CA	MG	NA	NOM
G	-0.1527	-0.8480	-0.5379	-0.1490	-0.2001	-0.3811	EVIAN (F)
L	-0.0750	0.9677	-0.2689	0.9493	0.1143	-0.3622	ONDINE (F)
O	-0.8084	-0.3679	-0.4758	-1.1777	-0.8575	-0.3937	SPA (B)
S	-0.0674	1.7358	-0.2689	1.5749	0.1429	-0.3937	VITTEL (F)
T	-0.7079	-0.8741	-0.4137	-1.0943	-0.7146	-0.3622	VOLVIC (F)
c	2.1649	0.0424	2.2549	0.0040	2.0866	2.2676	APOLLINARIS (D)
o	-0.3536	-0.6559	-0.2896	-0.1072	-0.5717	-0.3748	SPONTIN (B)

Matrix plot

```
# Matrix plot
plot(waters7, pch=16)
```

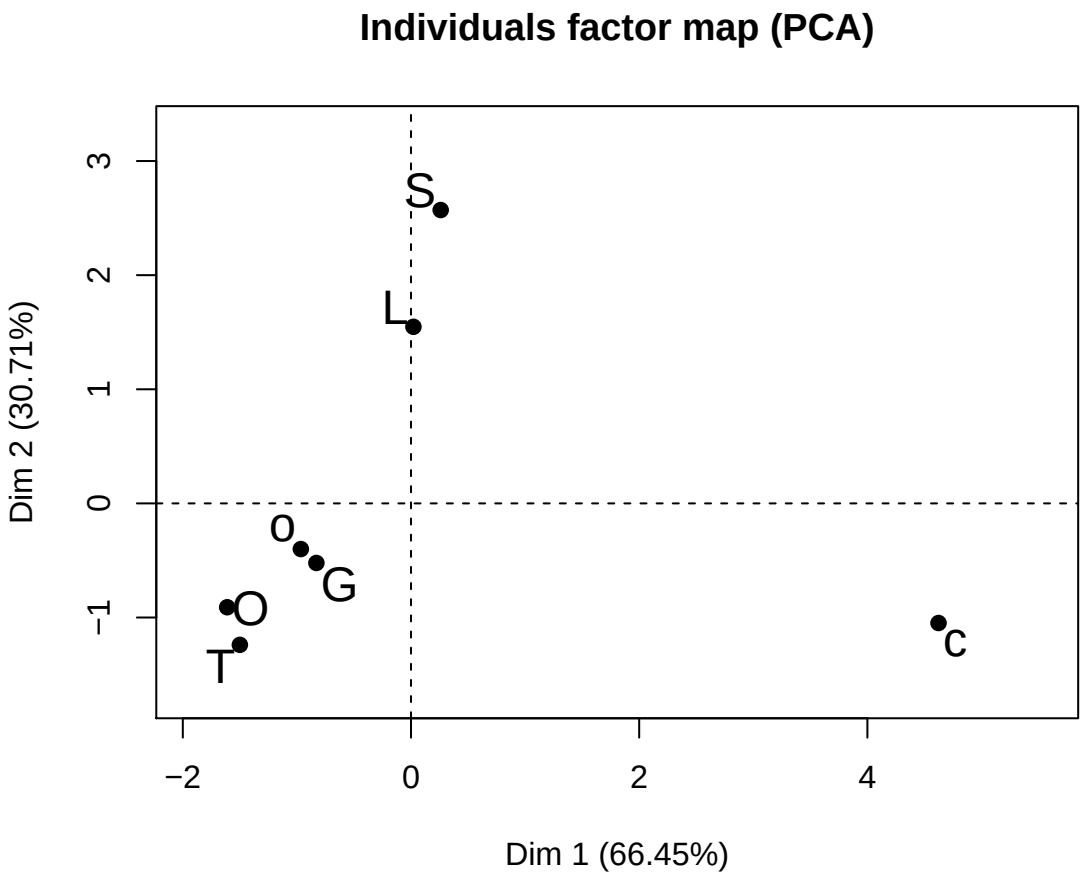
Waters = Matrix Plot



Example - PCA

```
# ACP
library(FactoMineR)
cluster.pca <- PCA(waters7std)
```

Example - PCA


Individuals factor map (PCA)
Dim 2 (30.71%)
Dim 1 (66.45%)
S
L
o
G
O
T
c
3
2
1
0
−1
−2
0
2
4

Contents

Distance

- Between observations
- Between groups

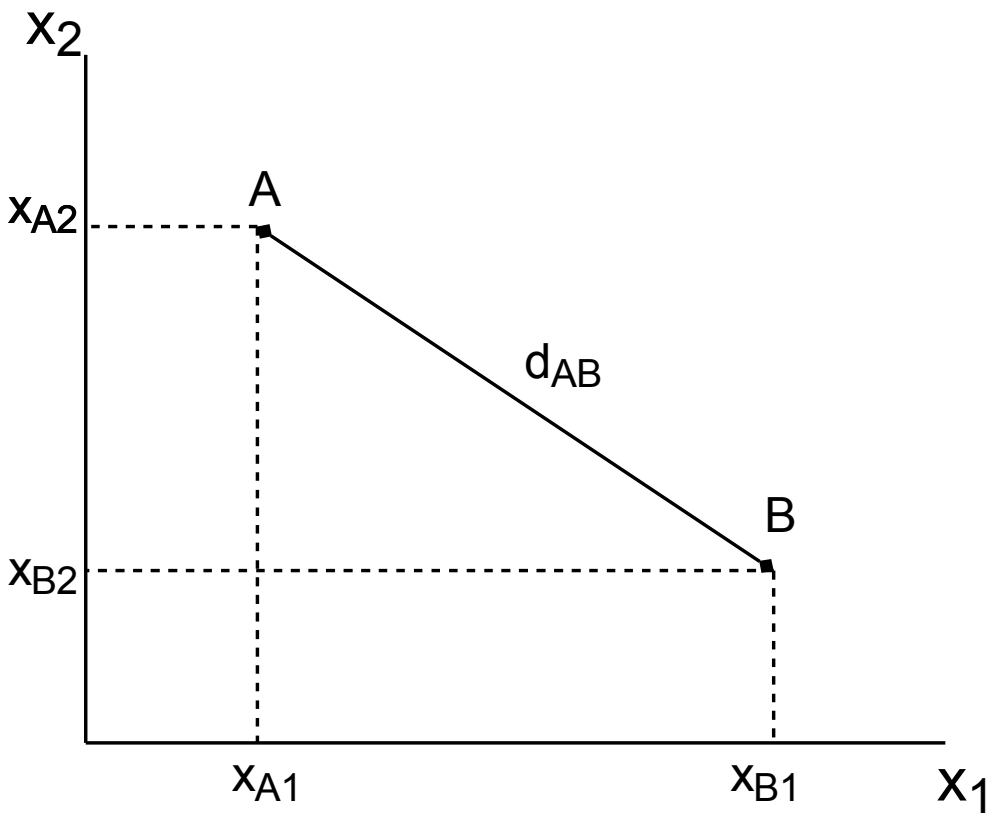
Distance space

Clustering methods heavily rely on (dis)similarity measures.

Euclidian distance is one of the most often used dissimilarity measures. But this distance is calculated into the **data space**, not the **coordinates space**.

Euclidian distance

$$
\begin{aligned}
d^2_{ii'} &= (x_{i1} - x_{i'1})^2 \\
&+ \dots \\
&+ (x_{ip} - x_{i'p})^2 \\
&= \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2
\end{aligned}
$$



Distance matrix

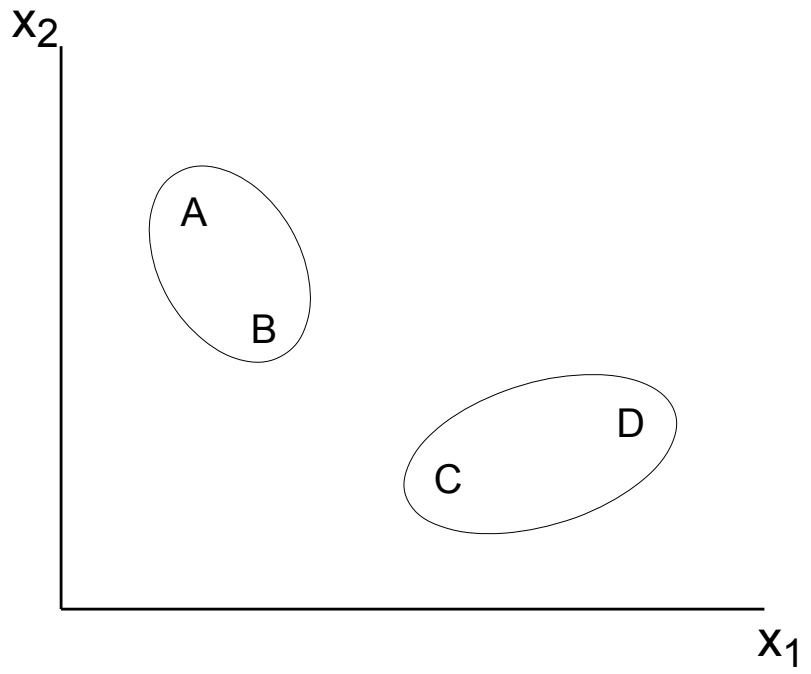
```
# Euclidean distance matrix
waters7.d <- dist(waters7std)
```

Distance

- Between observations
- Between groups

Distances between two clusters

- Nearest neighbour distance : d_{BC}
- Furthest neighbour distance : d_{AD}
- Average distance : $d^2 = (d^2_{AC} + d^2_{AD} + d^2_{BC} + d^2_{BD})/4$
- etc.



Example

Euclidian distances between waters

Distance	G	L	O	S	T	c	o
G	0	2.16	1.47	3.14	1.22	5.12	0.53
L	2.16	0	2.80	0.99	2.95	4.89	2.07
O	1.47	2.80	0	3.69	0.55	5.80	1.25
S	3.14	0.99	3.69	0	3.89	5.25	3.02
T	1.22	2.95	0.55	3.89	0	5.67	1.09
c	5.12	4.89	5.80	5.25	5.67	0	5.23
o	0.53	2.07	1.25	3.02	1.09	5.23	0

Example

Euclidian distances between clusters

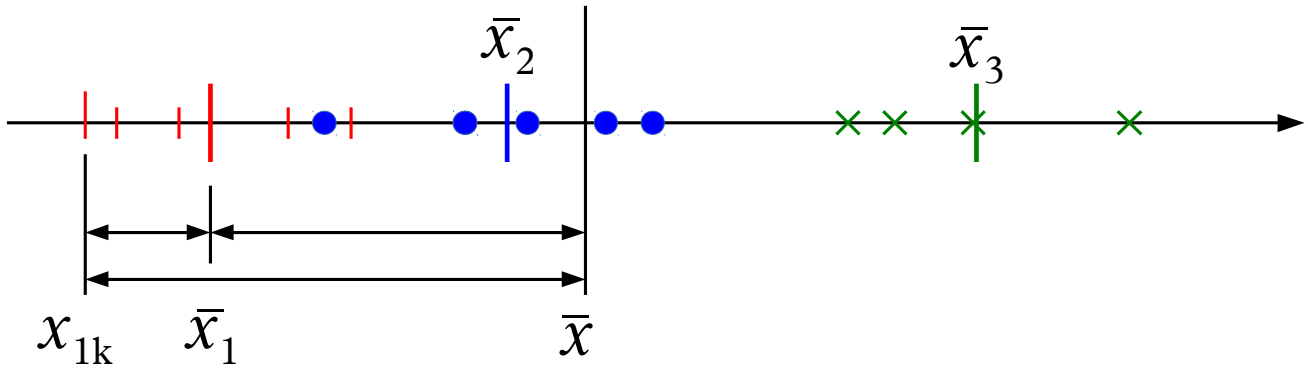
Distance	L	S
G	2.16	3.14
O	2.80	3.69
T	2.95	3.89
o	2.07	3.02

Nearest neighbour distance :
$2.07 = d_{oL}$

Furthest neighbour distance :
$3.89 = d_{TS}$

Average distance :
3.02

Ward's distance



$$\text{TOTAL SS} = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

$$\text{WITHIN SS} = \sum_{i=1}^{g} \left[\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \right]$$

$$\text{BETWEEN SS} = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})^2$$

Ward's distance

$$
\begin{aligned}
(\text{TOTAL})_1 &= (\text{BETWEEN})_1 + (\text{WITHIN})_1 \text{ for variable 1} \\
&\dots \\
(\text{TOTAL})_j &= (\text{BETWEEN})_j + (\text{WITHIN})_j \text{ for variable } j \\
&\dots \\
(\text{TOTAL})_p &= (\text{BETWEEN})_p + (\text{WITHIN})_p \text{ for variable } p
\end{aligned}
$$

$$
\sum_{j=1}^{p}(\text{TOTAL})j = \sum_{j=1}^{p}(\text{BETWEEN})_j + \sum_{j=1}^{p}(\text{WITHIN})_j
$$

$$
R^2 = \frac{\sum_{j=1}^{p}(\text{BETWEEN})_j}{\sum_{j=1}^{p}(\text{TOTAL})_j}
$$

Example

Analysis of variance

4 clusters : (G, o), (O, T), (L, S), (c)

Sources	df	HCO3	SO4	Cl	Ca	Mg	Na	Total
Between	3	5.97	5.56	5.97	5.80	5.92	6.00	35.22
Within	3	0.03	0.44	0.03	0.20	0.08	0.00	0.78
Total	6	6.00	6.00	6.00	6.00	6.00	6.00	36.00

R-squared $= 35.22 \ / \ 36 = 0.978$

LIÈGE université Gembloux Agro-Bio Tech

Example

Analysis of variance

3 clusters : (G, o, O, T), (L, S), (c)

Sources	df	HCO3	SO4	Cl	Ca	Mg	Na	Total
Between	2	5.72	5.54	5.97	4.78	5.76	6.00	33.77
Within	4	0.28	0.46	0.03	1.22	0.24	0.00	2.23
Total	6	6.00	6.00	6.00	6.00	6.00	6.00	36.00

R-squared $= 33.77 \ / \ 36 = 0.938$

$\Delta R^2 = 0.978 - 0.938 = 0.04$

LIÈGE université Gembloux Agro-Bio Tech

Contents

Hierarchical clustering techniques

- number of clusters not fixed
- **divisive methods** : from 1 to n clusters
- **agglomerative methods** : from n to 1 cluster
- results presented in the form of a dendrogram

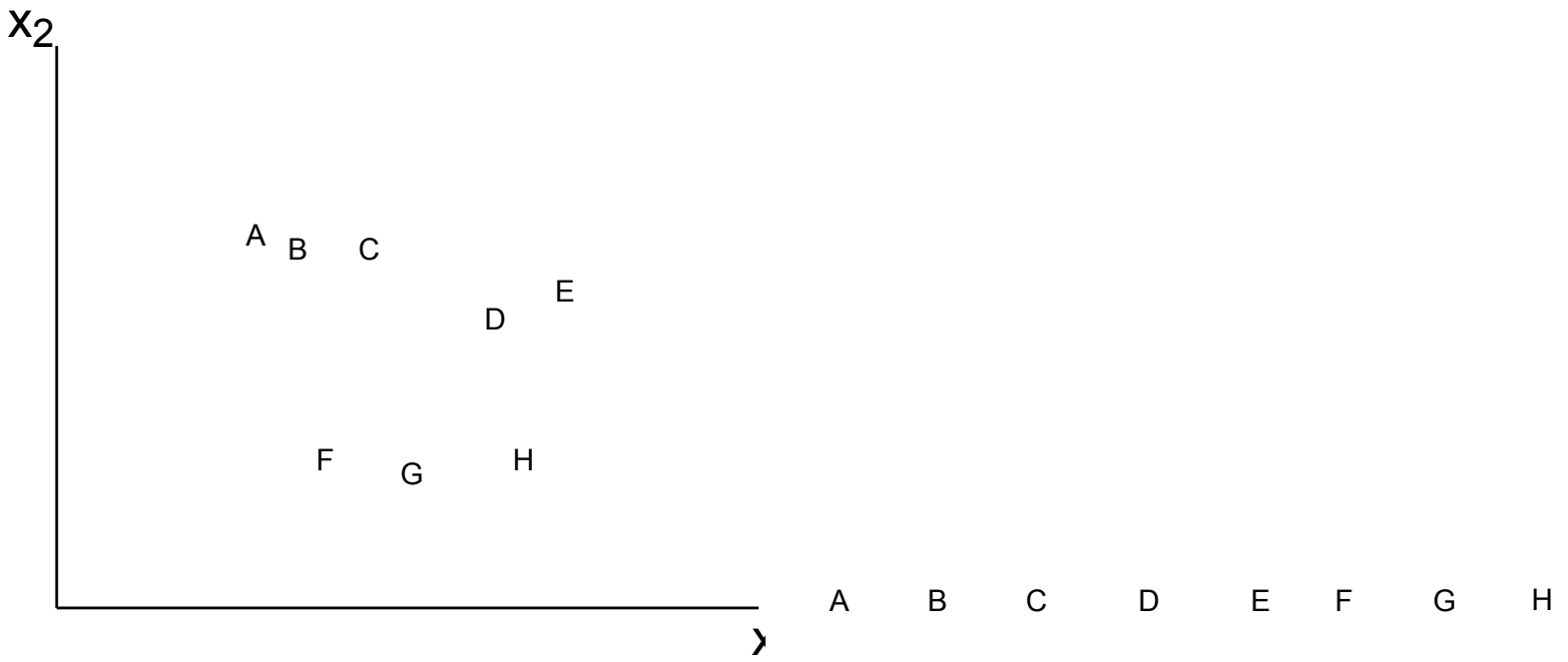
Hierarchical Clustering techniques

- Agglomerative methods
- Example

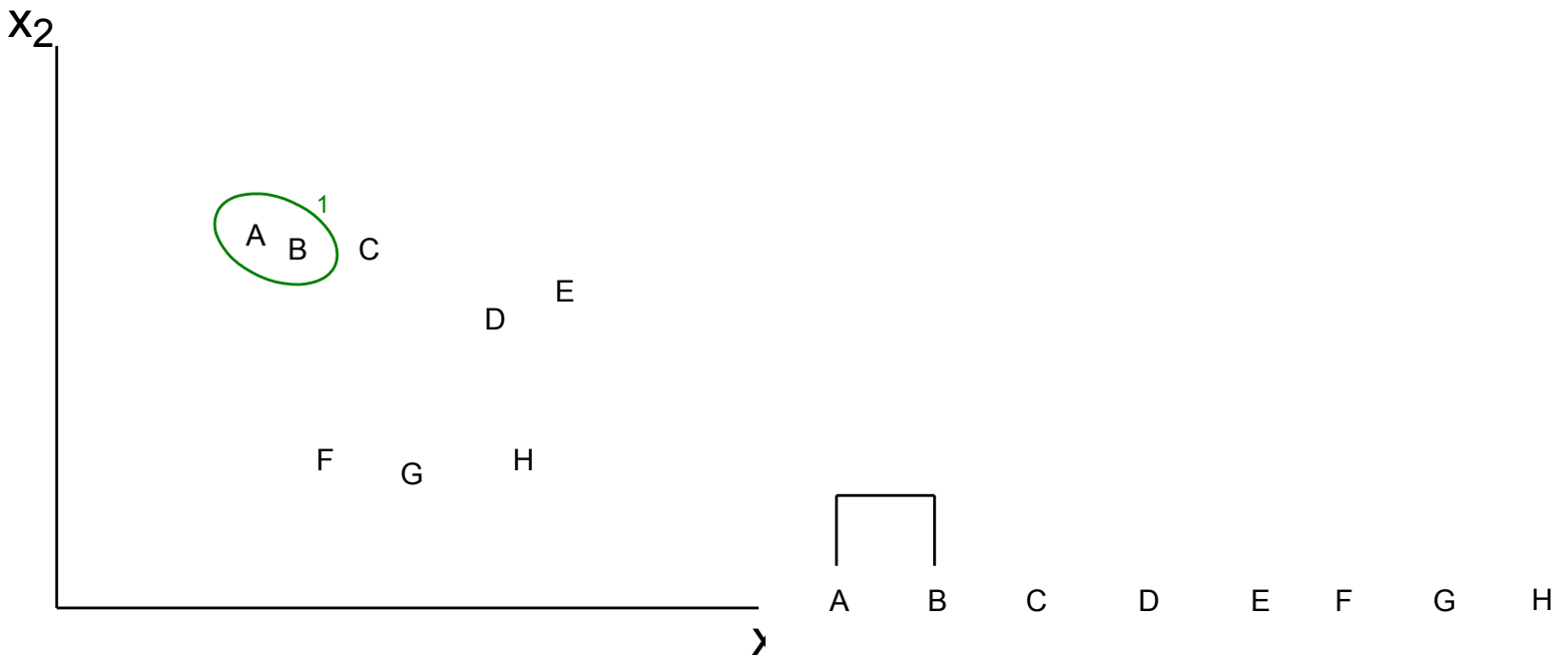
Agglomerative methods

- **WARD'S method**
  fusion of the two groups for which ΔR^2 is minimum
- **Single linkage**
  fusion of the two groups for which the nearest neighbour distance is minimum
- **Complete linkage**
  fusion of the two groups for which the furthest neighbour distance is minimum
- **Average linkage**
  fusion of the two groups for which the average distance is minimum

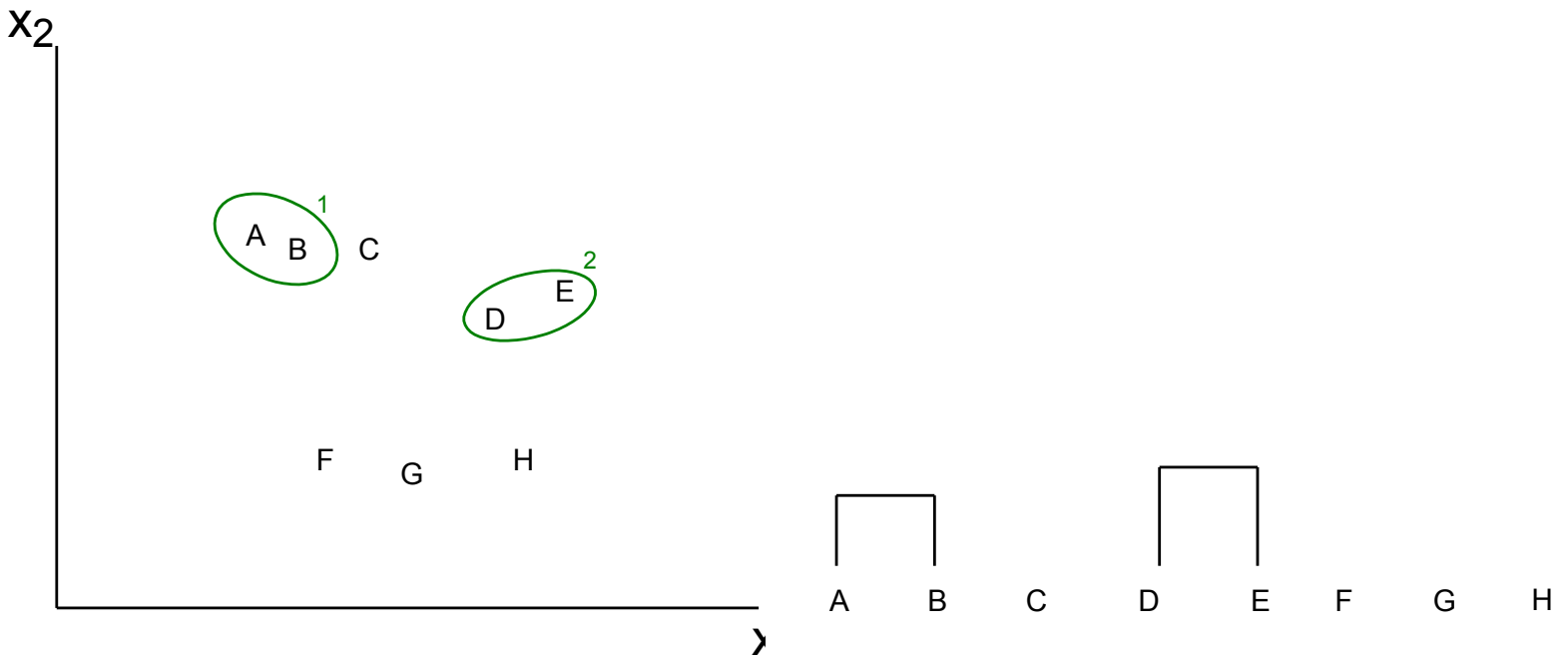
Agglomerative methods



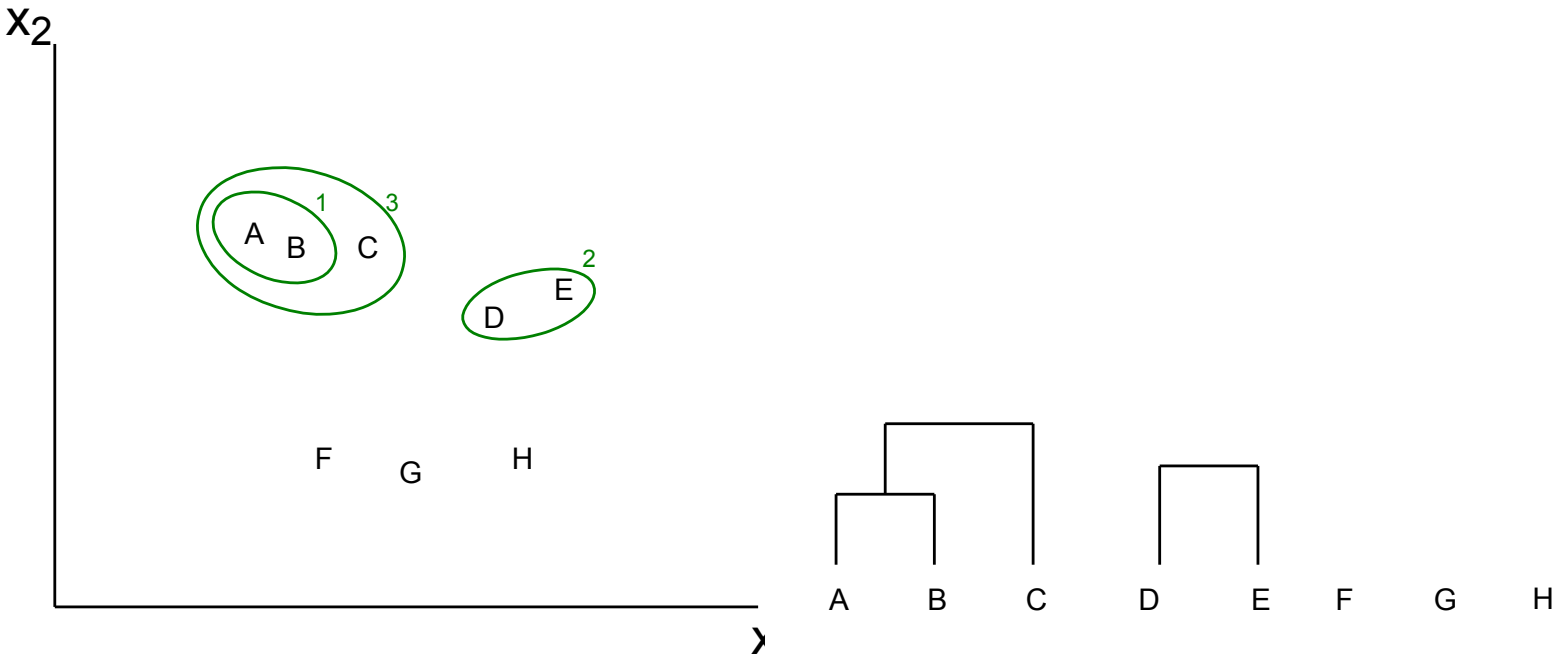
Agglomerative methods



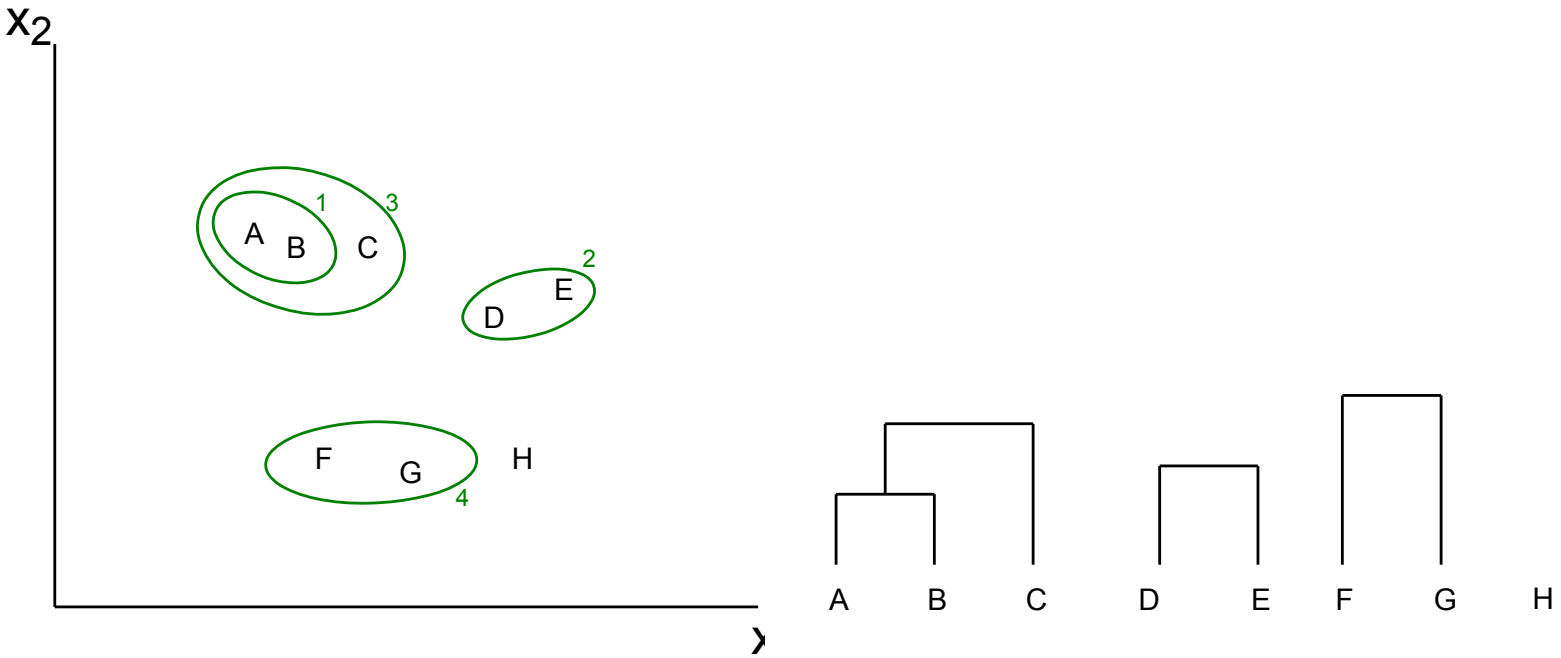
Agglomerative methods



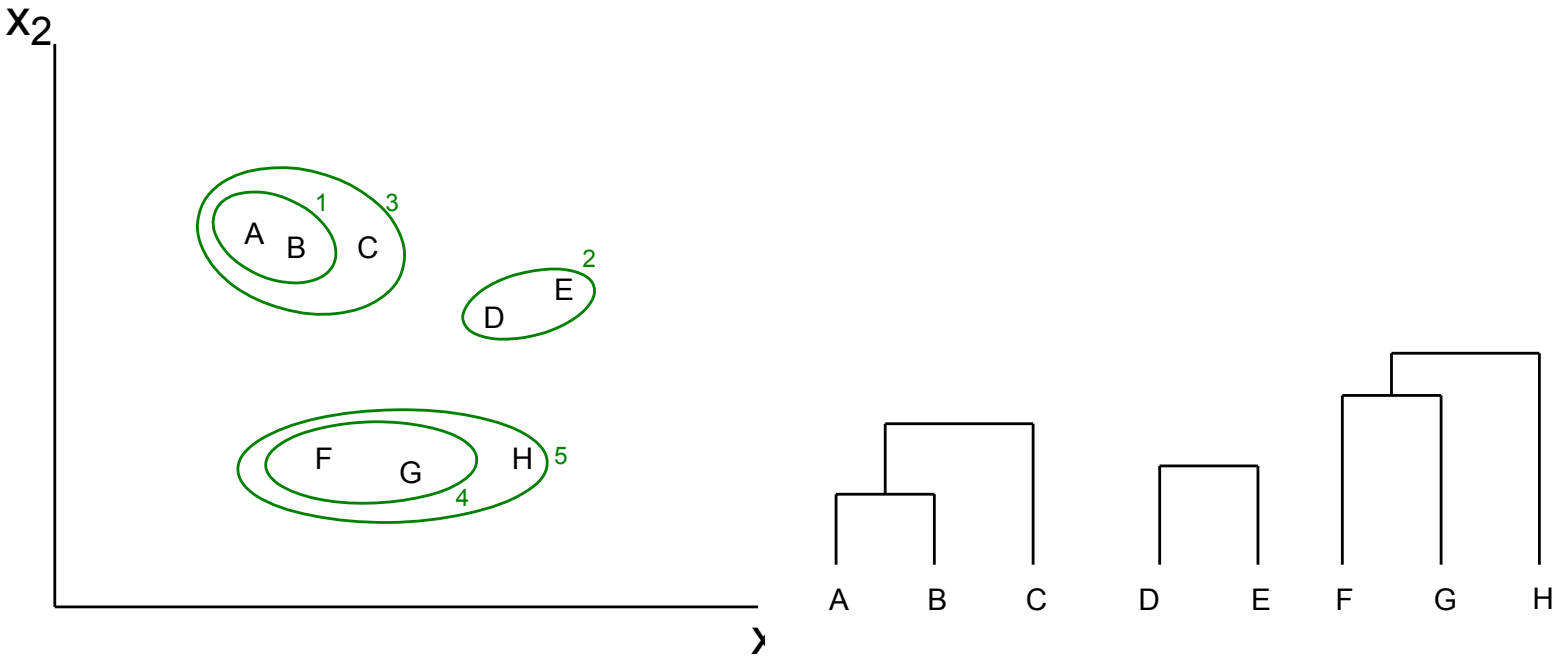
Agglomerative methods



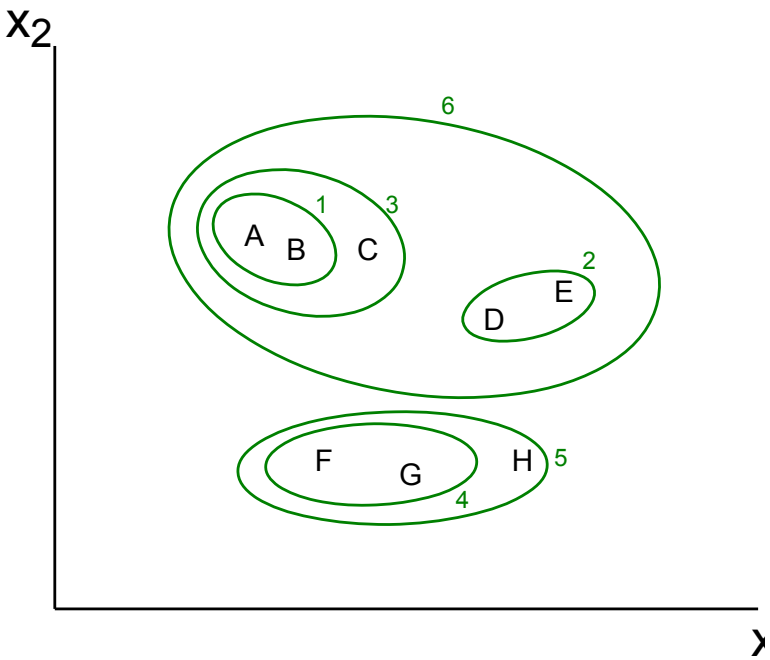
Agglomerative methods

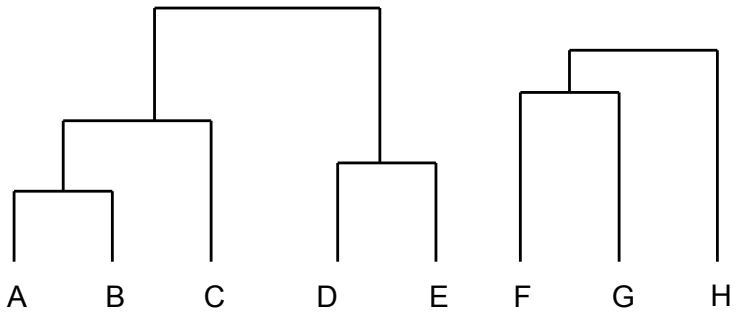


Agglomerative methods

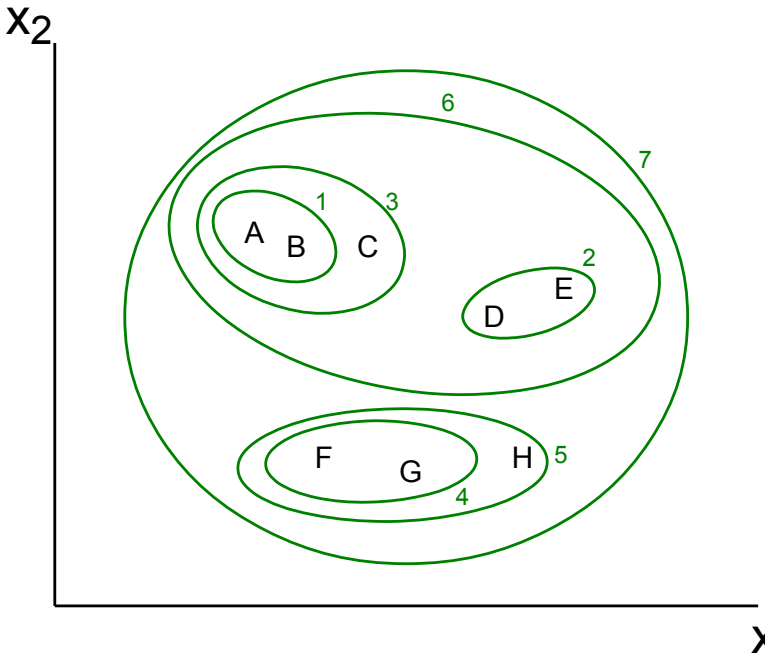


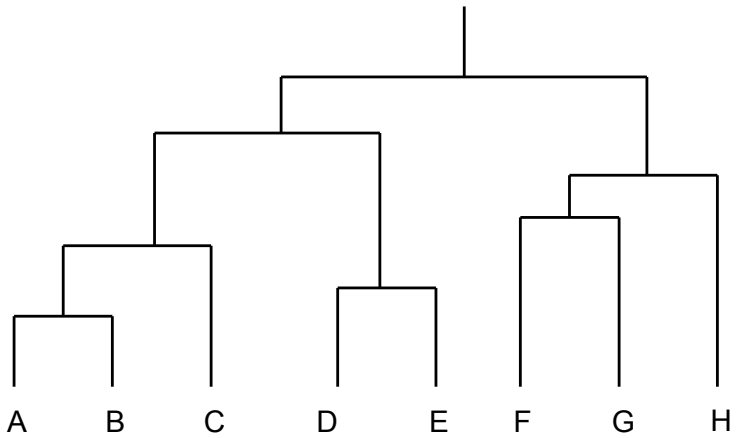
Agglomerative methods





Agglomerative methods





Hierarchical Clustering techniques

- Agglomerative methods
- Example

Example: complete linkage clustering I

Distance	L	O	S	T	c	o
G	2.16	1.47	3.14	1.22	5.12	0.53
L	0	2.80	0.99	2.95	4.89	2.07
O		0	3.69	0.55	5.80	1.25
S			0	3.89	5.25	3.02
T				0	5.67	1.09
c					0	5.23

Example: complete linkage clustering II

Distance	L	O	S	T	c
(G,o)	2.16	1.47	3.14	1.22	5.23
L	0	2.80	0.99	2.95	4.89
O		0	3.69	0.55	5.80
S			0	3.89	5.25
T				0	5.67

Distance	L	(O,T)	S	c
(G,o)	2.16	1.47	3.14	5.23
L	0	2.95	0.99	4.89
(O,T)		0	3.89	5.80
S			0	5.25

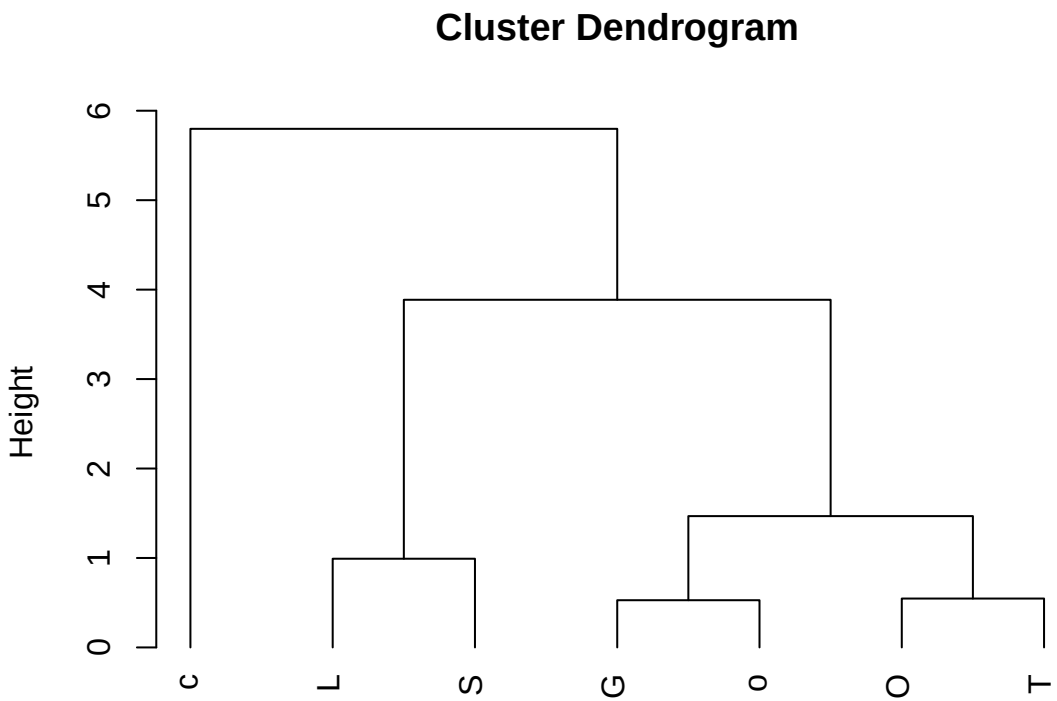
Example: complete linkage clustering III

Distance	(L,S)	(O,T)	c
(G,o)	3.14	1.47	5.23
(L,S)	0	3.89	5.25
(O,T)		0	5.80

Distance	(L,S)	c
(G,o,O,T))	3.89	5.80
(L,S)	0	5.25

Distance	c
(G,o,O,T,L,S))	5.80

Example - waters



eaux7.d
hclust (*, "complete")

Classification

```
# hierarchical clustering (default to complete link)
waters7.hc <- hclust(waters7.d)

# dendrogram
plot(waters7.hc, hang=-1)
```

Contents

LIÈGE université Gembloux Agro-Bio Tech

Steps in clustering

1. Data collection
2. Measurement of proximity
3. Choice of a clustering algorithm
4. Number of clusters
5. Interpretation of the results

Steps in clustering

- Data collection
- Measurement of proximity
- Choice of a clustering algorithm
- Number of clusters
- Interpretation of the results

Steps in clustering

- Data collection
- Measurement of proximity
- Choice of a clustering algorithm
- Number of clusters
- Interpretation of the results

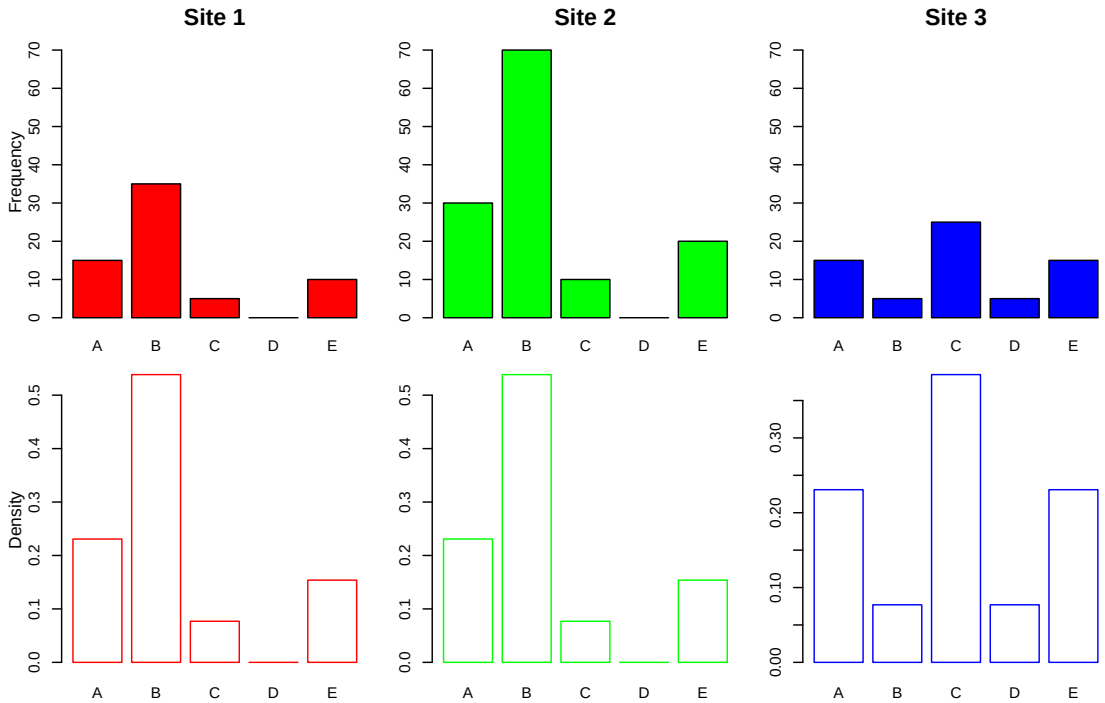
Measurement of proximity

- Euclidean distance
- other distances
  - χ^2
  - correlation
  - similarity/dissimilarity coefficients
  - ...
- transformation of coefficients :

$$d_{ii'} = 1 - s_{ii'} \quad d_{ii'} = 1 - r_{ii'} \quad d_{ii'} = 1 - r_{ii'}^2$$

χ^2 distance

- for (relative) frequencies distribution comparison (ecology)
- linked to χ^2 independance test



Correlation

- to cluster **variables** instead of **individuals**
- importance of the choice of the transformation (relation between anticorrelated variables)

$$d_{ii'} = 1 - r_{ii'} \quad d_{ii'} = 1 - r_{ii'}^2$$

Similarity coefficients for binary data

		individual i		
		1	0	
individual i'	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	a+b+c+d

a co-presences

d co-abscences

b, c mismatches

$$s_{ii'} = \frac{a+d}{a+b+c+d}$$ SOKAL & MICHENER

$$s_{ii'} = \frac{a}{a+b+c+d}$$ RUSSEL & RAO

$$s_{ii'} = \frac{a}{a+b+c}$$ JACCARD

LIÈGE université Gembloux Agro-Bio Tech

Steps in clustering

- Data collection
- Measurement of proximity
- Choice of a clustering algorithm
- Number of clusters
- Interpretation of the results

Hierarchical clustering

- **Distances between observations**
  - cf. Measurement of proximity
- **Distances between groups**
  - WARD'S method
  - single linkage
  - complete linkage
  - average linkage
  - centroïd linkage
  - median linkage

Non hierarchical clustering

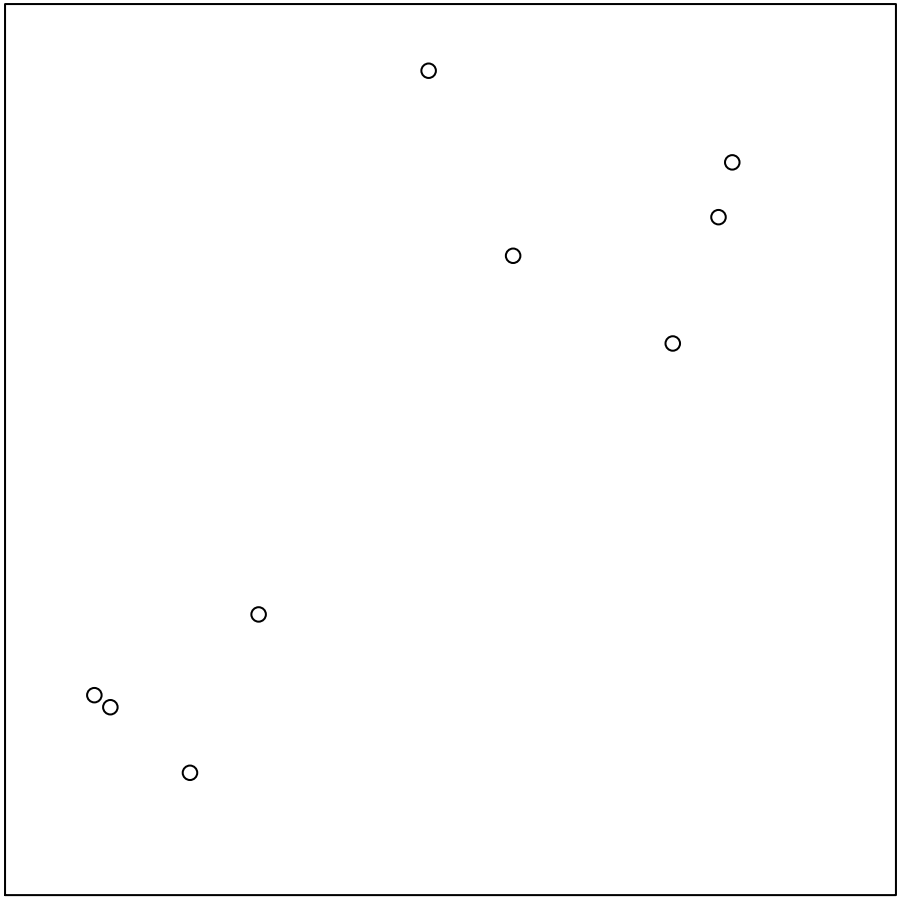
- fixed number of clusters
- methods based on the SS decompositon : $T = H + E$ (minimum of the trace of $E^{-1}H$)
- relocation methods (k-means)

K-means clustering

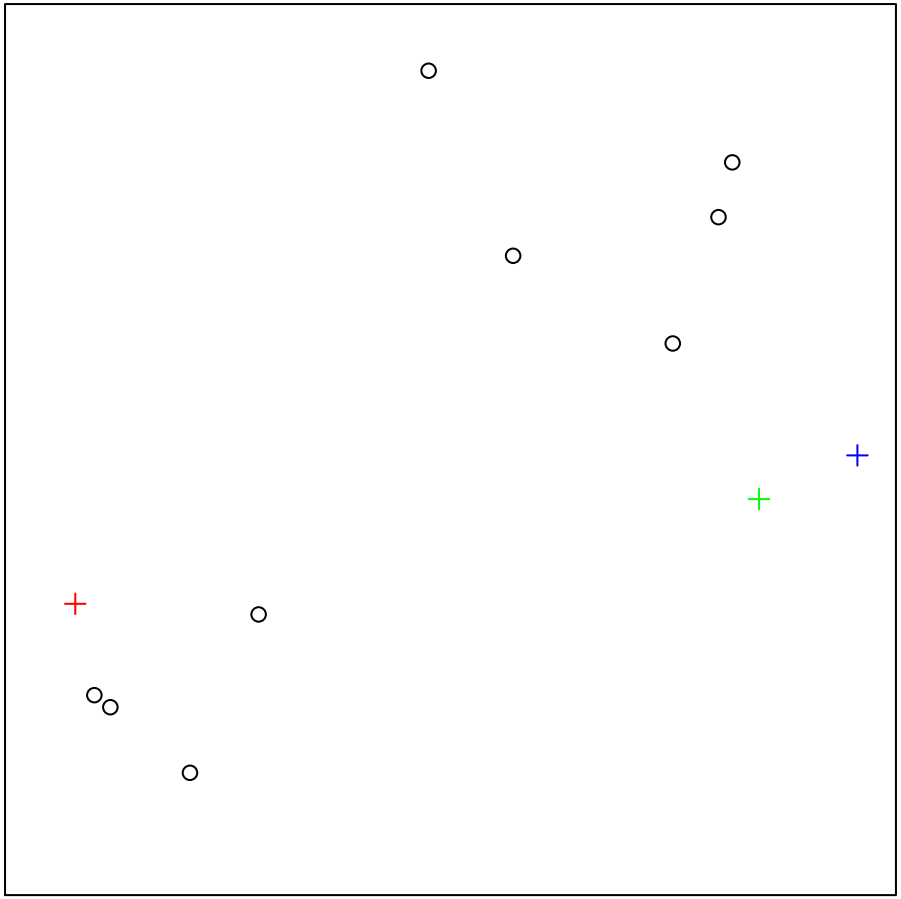
Algorithm

- generate k random centroids
- repeat until no more modifications
  - calculate distance between each points and each centroids
  - assign points to the nearest centroid group
  - recalculate centroids coordinates of the new groups

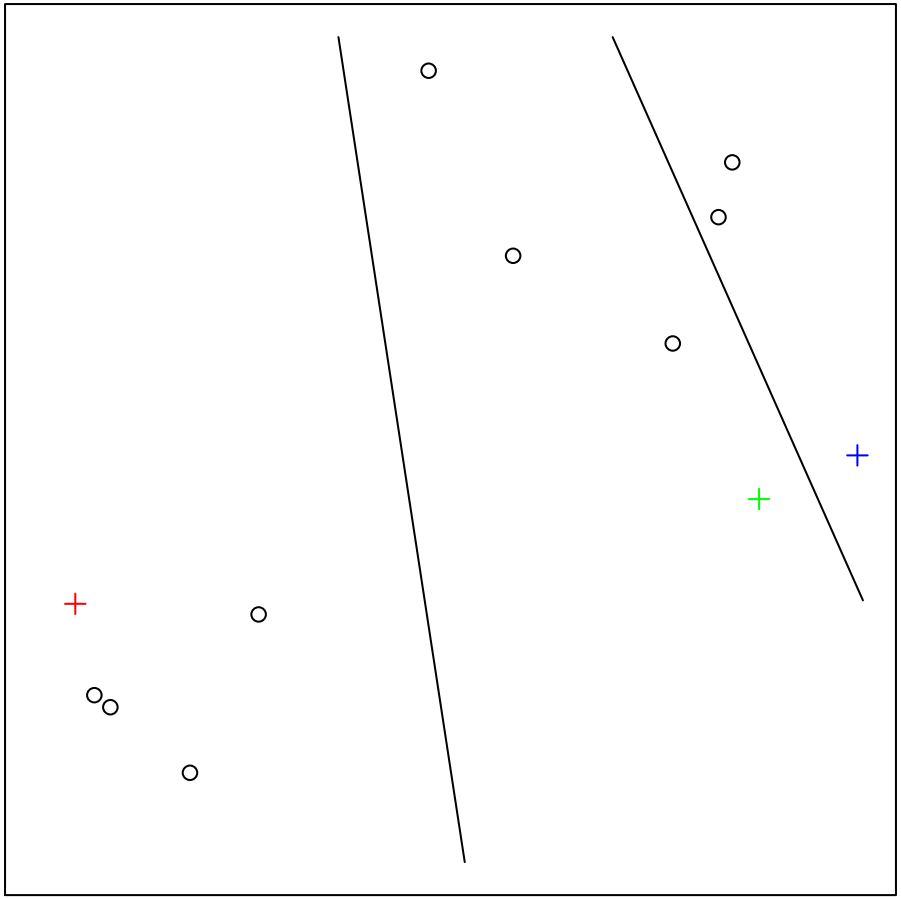
Example - K-means



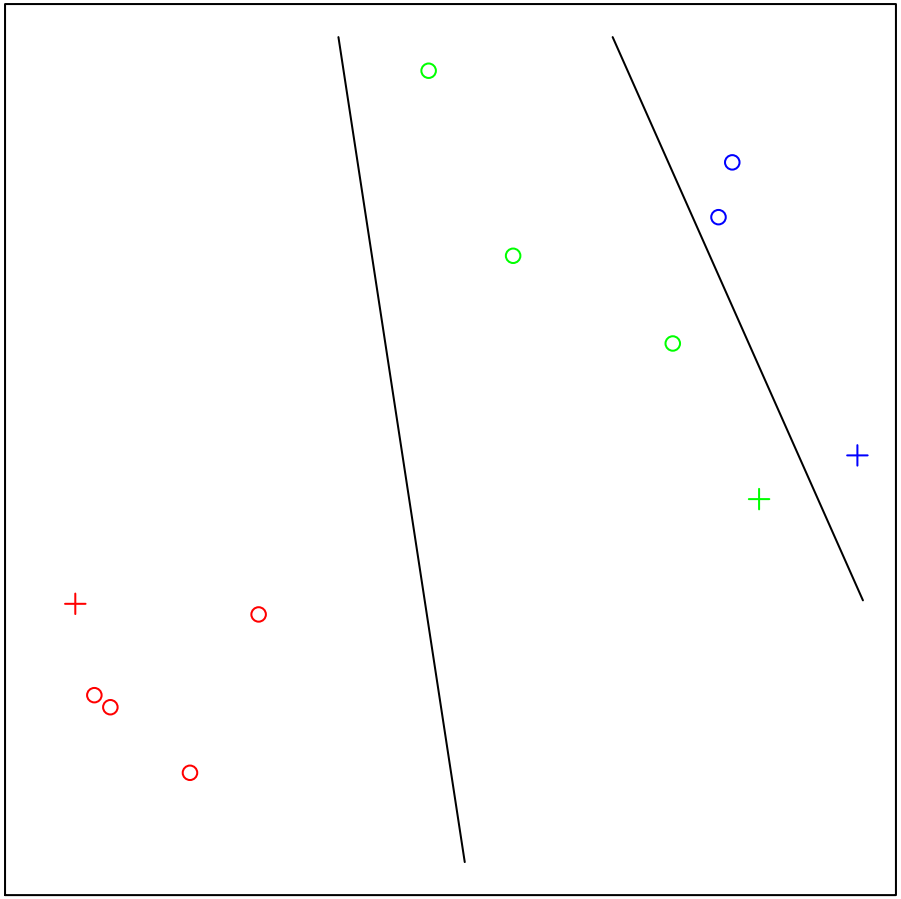
Example - K-means



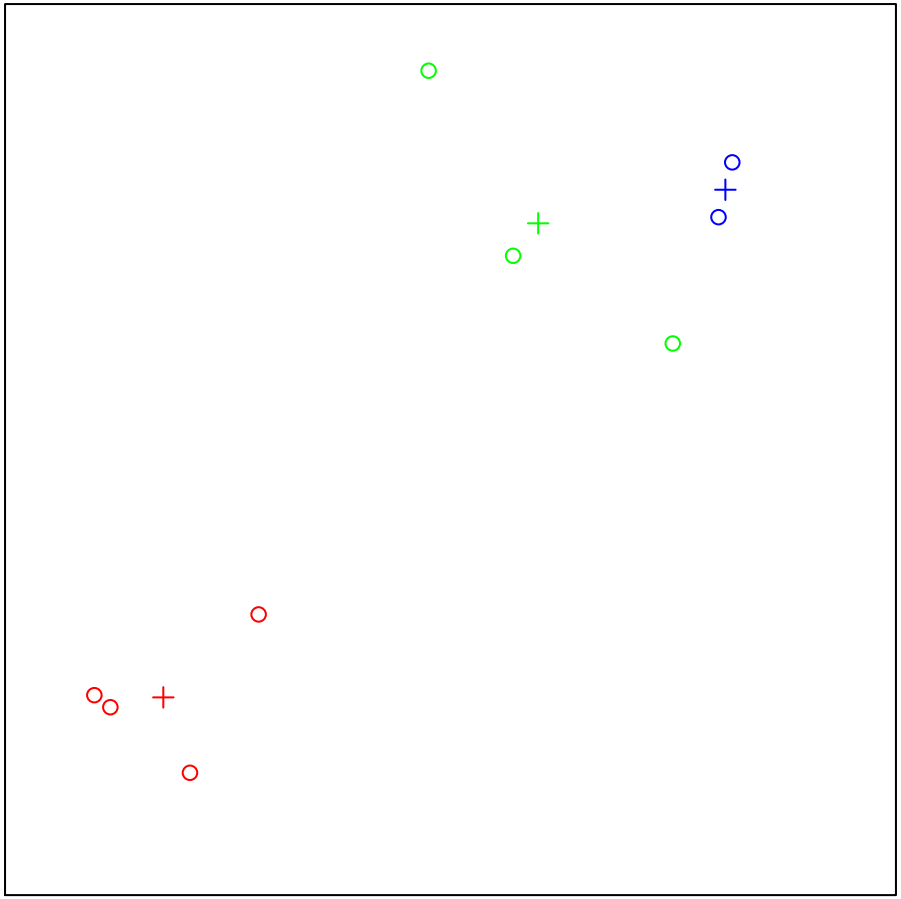
Example - K-means



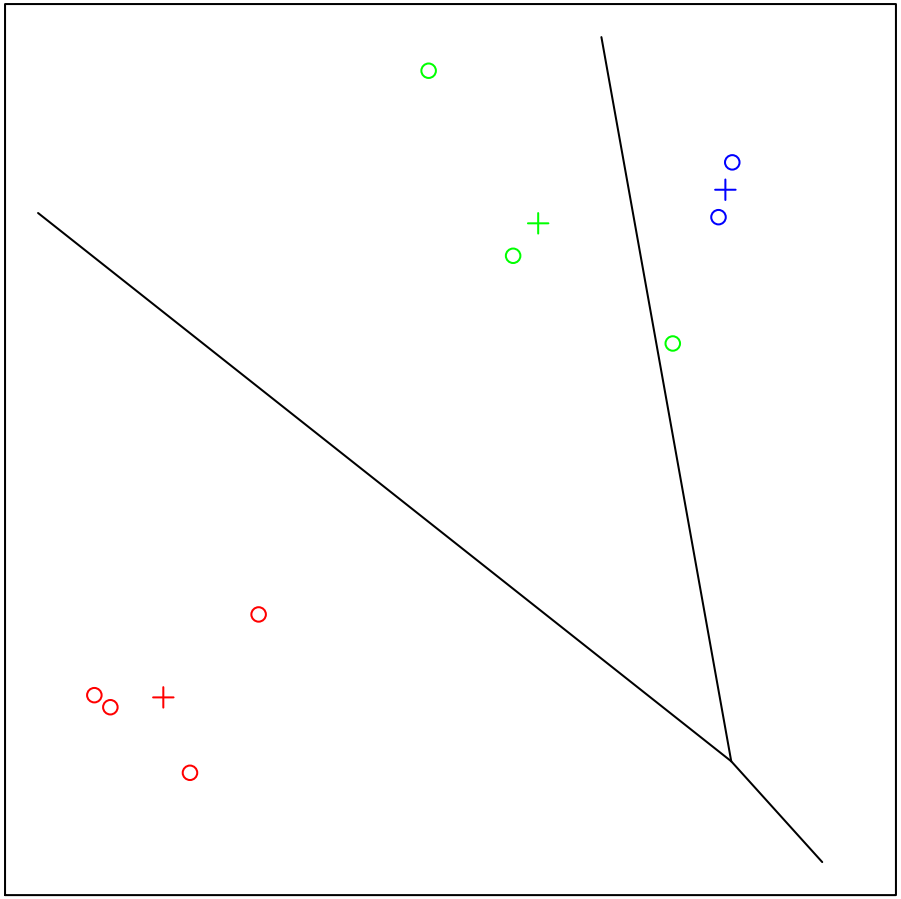
Example - K-means



Example - K-means



Example - K-means



```
waters7.km <- kmeans (waters7std, centers=3)

waters7.km
```

K-means clustering

Pros/cons

- Simple, quick and easy
- Fixed number of clusters
- Sensitive to initial conditions
  - multiple runs with random initial centroids
- Sensitive to outliers
- works best with linearly separable clusters

Combination of methods

- k-means $\rightarrow$ hierarchical clustering
  - reduce the number of initial observations
  - used when initials steps are not important to speed up process
- hierarchical clustering $\rightarrow$ k-means
  - test robustness of partition
  - fix initial centroids as centers of the hierarchical clusters

```
# calculate groups centroïds
init <- aggregate(.~groups, data=as.data.frame(data.std), FUN=mean)
# generate kmeans partition based on previous centroïds
kmean.res <- kmeans(data.std, centers=init[-1])
# compare groupings
table(groups, kmean.res$cluster)
```

Steps in clustering

- Data collection
- Measurement of proximity
- Choice of a clustering algorithm
- Number of clusters
- Interpretation of the results

Number of clusters

- plot of R^2 as a function of the number of clusters
- pseudo $F = \frac{tr(H)/(g-1)}{tr(E)/(n-g)}$
- etc.

More generally, look for discontinuities in the clustering process (sudden raise of the height of merging)

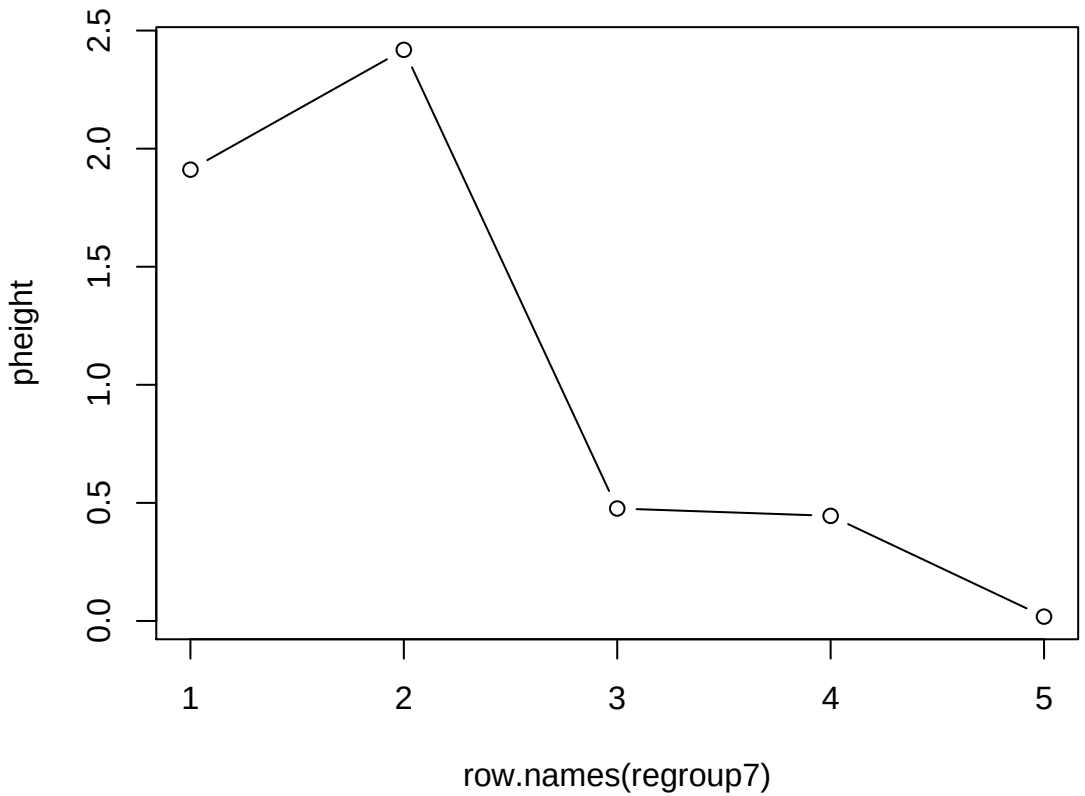
Successive merging steps

```
# informations on successive merging steps

regroup7 <- hclust.info(waters7.hc)

plot(height~row.names(regroup7), data=regroup7,
type="b")

plot(height~row.names(regroup7), data=regroup7,
type="b")
```

Example - waters

	height	pheight
1	5.798	1.911
2	3.886	2.418
3	1.468	0.476
4	0.992	0.445
5	0.546	0.019

Example - waters



Steps in clustering

- Data collection
- Measurement of proximity
- Choice of a clustering algorithm
- Number of clusters
- Interpretation of the results

Interpretation of the results

- univariate descriptive techniques
  - mean and standard deviation of each variable for each group
- multivariate descriptive techniques
  - distances between centroids of groups
  - principal components analysis

Group partition

```
# Group partition
gr3 <- cutree(waters7.hc, 3)
sort(gr3)
# Frequencies
table(gr3)
# Means
aggregate(.~gr3, data=waters7[1:6], FUN=mean)
# Standard diviations
aggregate(.~gr3, data=waters7[1:6], FUN=sd)
```

Example - waters

3 groups: $(G,O,T,o)_1$, $(L,S)_2$, $(c)_3$

Group means

Groups	HCO3	SO4	Cl	Ca	Mg	Na
1	170.75	28.5	7.25	43.25	10.5	5.5
2	400	262	15	179.5	35.5	5.5
3	1580	112	137	89	104	425
Total	437.57	107.14	28.00	88.71	31.00	65.43

LIÈGE université Gembloux Agro-Bio Tech

ACP

```
# ACP
library(FactoMineR)
cluster.pca <- PCA(waters7std)

# generate colors vector according to groups
col.gr <- rainbow(length(unique(gr3)))[gr3]

# plot groups
plot(cluster.pca, col.ind=col.gr, cex=1.5)
```

Example - waters