Hierarchical clustering

François Husson

Applied Mathematics Department - Rennes Agrocampus

husson@agrocampus-ouest.fr

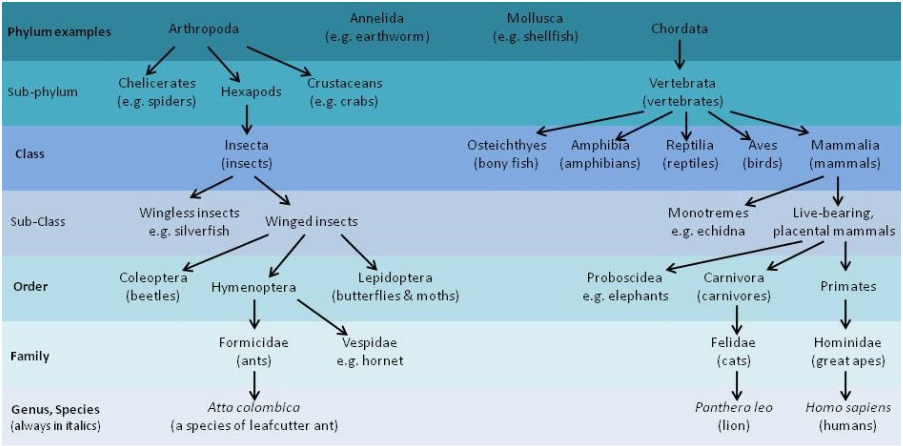
Hierarchical clustering

Hierarchical clustering

Introduction

- Definitions :
  - Clustering is : making or building classes
  - Class : set of individuals (or objects) with similar shared characteristics
- Examples
  - of clustering : animal kingdom, computer hard disk, geographic division of France, etc.
  - of classes : social classes, political classes, etc.
- Two types of clustering :
  - hierarchical : tree
  - partitioning methods

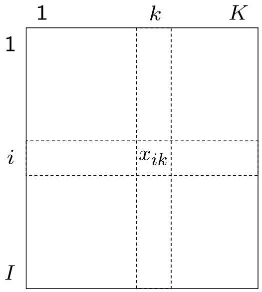
Hierarchical example : the animal kingdom



Hierarchical clustering

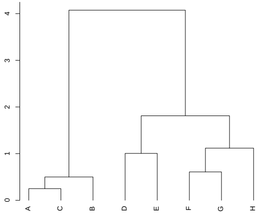
What data ? What goals ?

Clustering is for data tables : rows of individuals, columns of quantitative variables



Goals : build a tree structure that :

- shows hierarchical links between individuals or groups of individuals
- detects a "natural" number of classes in the population



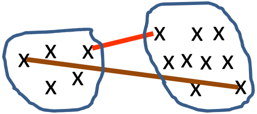
Critères

Measuring similarity of individuals :

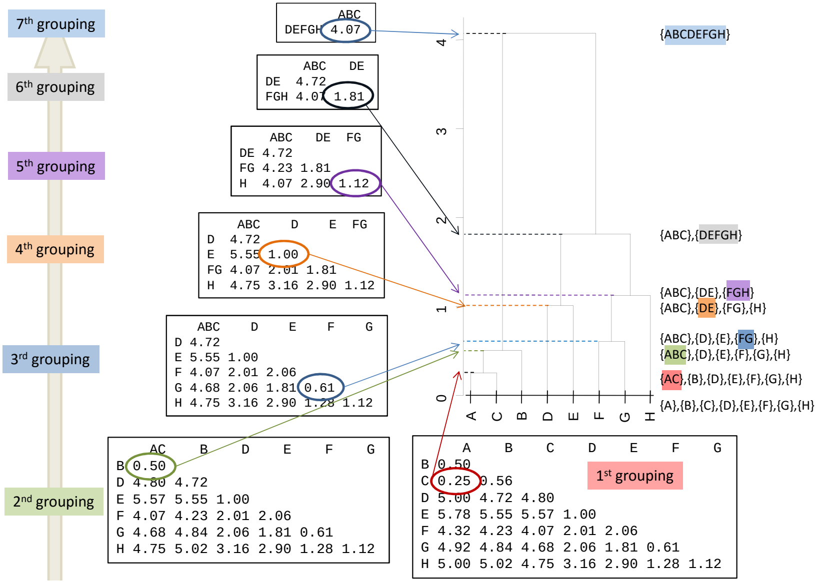
- Euclidean distance
- similarity indices
- etc.

Similarity between groups of individuals :

- minimum jump or single linkage (smallest distance)
- complete linkage (largest distance)
- Ward criterion



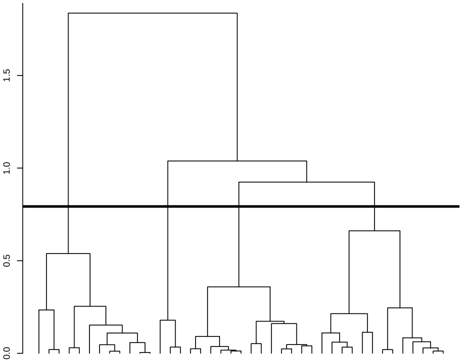
Algorithm


7th grouping
6th grouping
5th grouping
4th grouping
3rd grouping
2nd grouping
1st grouping
ABC
DEFGH 4.07
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABCDEFGH}
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Trees and partitions

Trees always end up . . . cut through !

Choosing a height to cut at gives a partition



Remark : given how it was made, the partition is interesting but not optimal

Partition quality

When is a partition a good one ?

- If individuals placed in the same class are close to each other
- If individuals in different classes are far from each other

Mathematically speaking ?

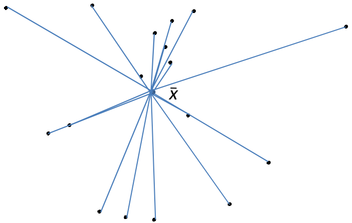
- small within-class variability
- large between-class variability

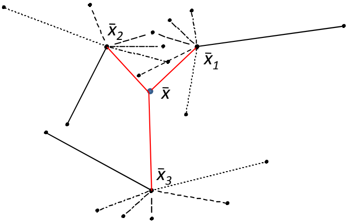
$\Longrightarrow$ Two criteria. Which one to use ?

Partition quality

$\bar{x}_k$ the mean of the x_k, $\bar{x}_{qk}$ the mean of the x_k in class q

$$\underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_k)^2}_{\text{total inertia}} = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_{qk})^2}_{\text{within-class inertia}} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_{qk}-\bar{x}_k)^2}_{\text{between-class inertia}}$$





$\Longrightarrow$ 1 criterion only !

Partition quality

Partition quality is measured by :

$$0 \leq \frac{\text{between-class inertia}}{\text{total inertia}} \leq 1$$

$$\frac{\text{inertia}_{\text{between}}}{\text{inertia}_{\text{total}}} = 0 \Longrightarrow \forall k, \forall q, \bar{x}_{qk} = \bar{x}_k$$

by variable, classes have the same means

Doesn't allow us to classify

$$\frac{\text{inertia}_{\text{between}}}{\text{inertia}_{\text{total}}} = 1 \Longrightarrow \forall k, \forall q, \forall i, x_{iqk} = \bar{x}_{qk}$$

individuals in the same class are identical

Ideal for classifying

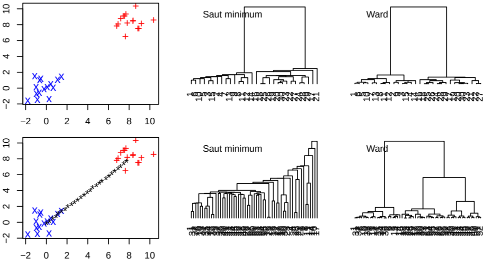
Warning : don't just accept this criteria at face value : it depends on the number of individuals and classes

Ward's method

- Initialize : 1 class = 1 individual $\Longrightarrow$ Between-class inertia = total inertia
- At each step : combine classes a and b that minimize the decrease in between-class inertia

$$\text{Inertia}(a) + \text{Inertia}(b) = \text{Inertia}(a \cup b) - \underbrace{\frac{m_a m_b}{m_a + m_b} d^2(a, b)}_{\text{to minimize}}$$

Group together objects with small weights and avoid chain effects



Group together classes with similar centers of gravity

Direct use for clustering

Hierarchical clustering

Temperature data

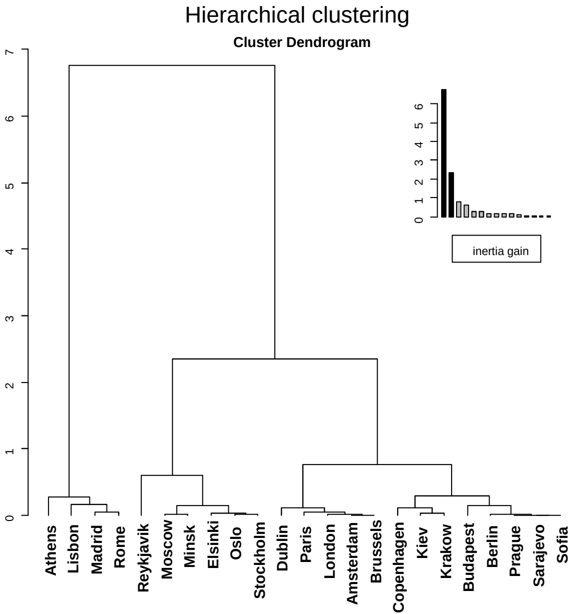
- 23 individuals : European capitals
- 12 variables : mean monthly temperatures over 30 years

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Area
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5	11.4	7.0	4.4	West
Athens	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8	19.2	14.6	11.0	South
Berlin	-0.2	0.1	4.4	8.2	13.8	16.0	18.3	18.0	14.4	10.0	4.2	1.2	West
Brussels	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15.0	11.1	6.7	4.4	West
Budapest	-1.1	0.8	5.5	11.6	17.0	20.2	22.0	21.3	16.9	11.3	5.1	0.7	East
Copenhagen	-0.4	-0.4	1.3	5.8	11.1	15.4	17.1	16.6	13.3	8.8	4.1	1.3	North
Dublin	4.8	5.0	5.9	7.8	10.4	13.3	15.0	14.6	12.7	9.7	6.7	5.4	North
Elsinki	-5.8	-6.2	-2.7	3.1	10.2	14.0	17.2	14.9	9.7	5.2	0.1	-2.3	North
Kiev	-5.9	-5.0	-0.3	7.4	14.3	17.8	19.4	18.5	13.7	7.5	1.2	-3.6	East
Krakow	-3.7	-2.0	1.9	7.9	13.2	16.9	18.4	17.6	13.7	8.6	2.6	-1.7	East
Lisbon	10.5	11.3	12.8	14.5	16.7	19.4	21.5	21.9	20.4	17.4	13.7	11.1	South
London	3.4	4.2	5.5	8.3	11.9	15.1	16.9	16.5	14.0	10.2	6.3	4.4	North
Madrid	5.0	6.6	9.4	12.2	16.0	20.8	24.7	24.3	19.8	13.9	8.7	5.4	South
Minsk	-6.9	-6.2	-1.9	5.4	12.4	15.9	17.4	16.3	11.6	5.8	0.1	-4.2	East
Moscow	-9.3	-7.6	-2.0	6.0	13.0	16.6	18.3	16.7	11.2	5.1	-1.1	-6.0	East
Oslo	-4.3	-3.8	-0.6	4.4	10.3	14.9	16.9	15.4	11.1	5.7	0.5	-2.9	North
Paris	3.7	3.7	7.3	9.7	13.7	16.5	19.0	18.7	16.1	12.5	7.3	5.2	West
Prague	-1.3	0.2	3.6	8.8	14.3	17.6	19.3	18.7	14.9	9.4	3.8	0.3	East
Reykjavik	-0.3	0.1	0.8	2.9	6.5	9.3	11.1	10.6	7.9	4.5	1.7	0.2	North
Rome	7.1	8.2	10.5	13.7	17.8	21.7	24.4	24.1	20.9	16.5	11.7	8.3	South
Sarajevo	-1.4	0.8	4.9	9.3	13.8	17.0	18.9	18.7	15.2	10.5	5.1	0.8	South
Sofia	-1.7	0.2	4.3	9.7	14.3	17.7	20.0	19.5	15.8	10.7	5.0	0.6	East
Stockholm	-3.5	-3.5	-1.3	3.5	9.2	14.6	17.2	16.0	11.7	6.5	1.7	-1.6	North

Which cities have similar weather patterns ?

How to characterize groups of cities ?

Temperature data : hierarchical tree



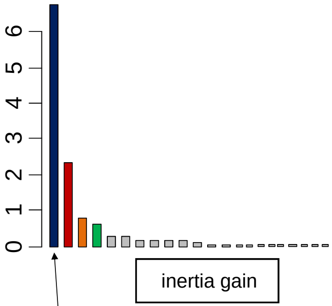
Temperature data

Loss in between-inertia when going from

23 clusters to 22 clusters: 0.01
22 clusters to 21 clusters: 0.01
21 clusters to 20 clusters: 0.01

....

9 clusters to 8 clusters: 0.15
8 clusters to 7 clusters: 0.16
7 clusters to 6 clusters: 0.27
6 clusters to 5 clusters: 0.29
5 clusters to 4 clusters: 0.60
4 clusters to 3 clusters: 0.76
3 clusters to 2 clusters: 2.36
2 clusters to 1 clusters: 6.76



Important loss when going from 2 clusters to a unique cluster thus we prefer to keep 2 custers

Sum of losses of inertia = 12

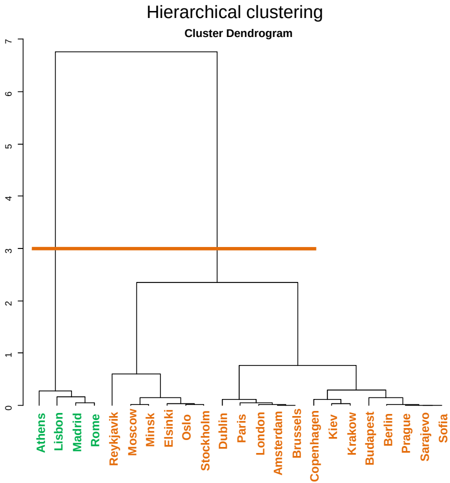
Using the tree to build a partition

Should we make 2 groups ? 3 ? 4 ?

Cut into 2 groups :

$$\frac{\text{between-class inertia}}{\text{total inertia}} = \frac{6.76}{12}$$

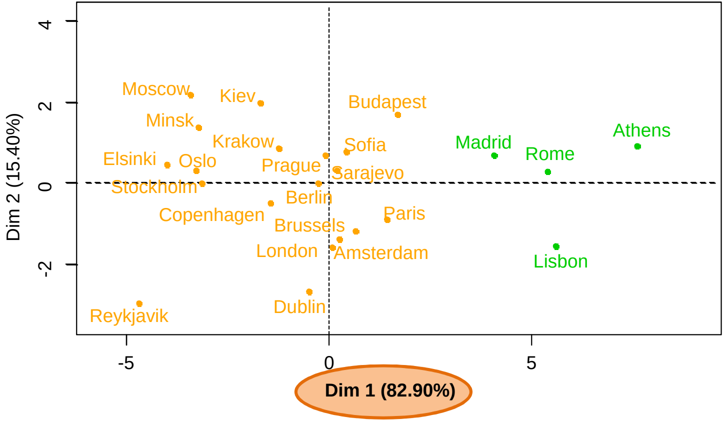
What can we compare this percentage with ?



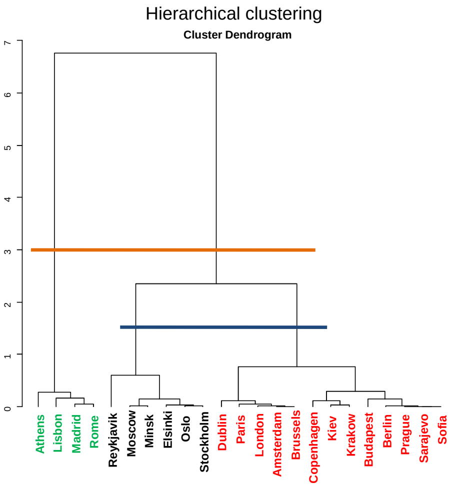
Using the tree to build a partition

66 % of the information is contained in this 2-class cut

What can we compare this percentage with ?



Using the tree to build a partition

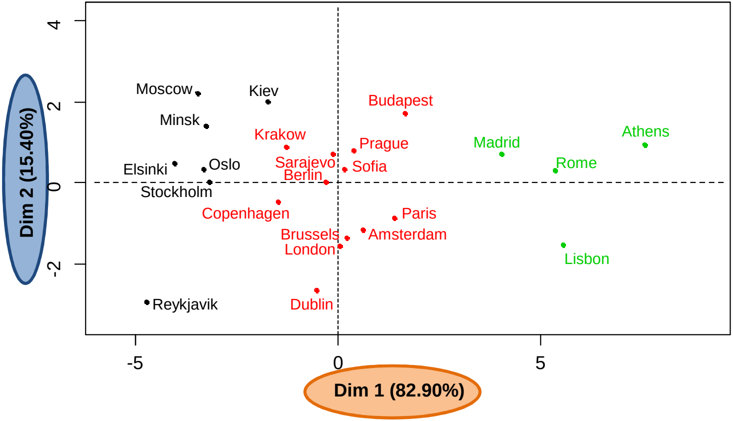


Separate cold cities into 2 groups :

$$\frac{\text{between-class inertia}}{\text{total inertia}} = \frac{2.36}{12} = 20\%$$

Using the tree to build a partition

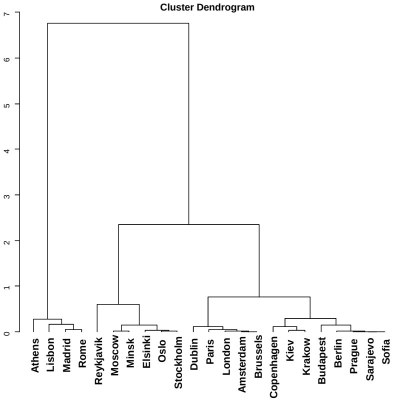
The move from 23 cities to 3 classes : 56 % + 20 % = 76 % of the variability in the data

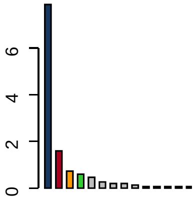


Determining the number of classes

- Starting from the tree
- Depends on the use (survey, etc.)
- Plot with the bars
- Ultimate criterion : interpretability of the classes

Hierarchical clustering



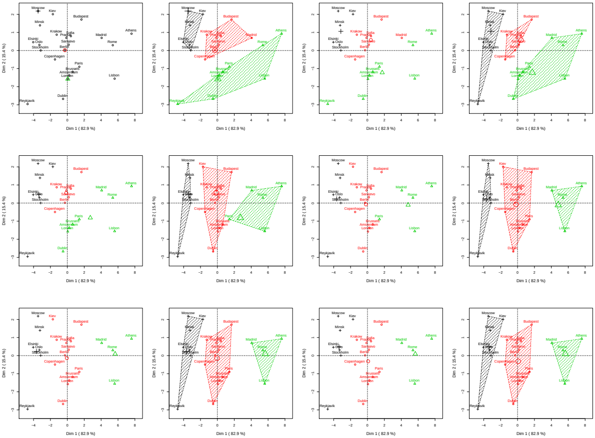


Hierarchical clustering

Partitioning algorithm : K-means

Algorithm for aggregating around moving centers (K-means)

- Choose randomly Q centers of gravity
- Assign the points to the closest center
- Calculate anew the Q centers of gravity



Hierarchical clustering

Robustifying a partition obtained using hierarchical clustering

The partition obtained by hierarchical clustering is not optimal and can be improved or made robust using K-means

Algorithm :

- use the obtained hierarchical partition to initialize K-means
- run a few iterations of K-means

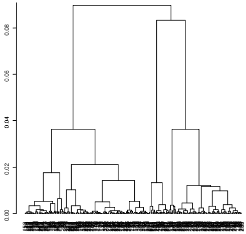
$\Longrightarrow$ potentially improved partition
Advantage : more robust partition
Disadvantage : loss of hierarchical structure

Hierarchical clustering in high dimension

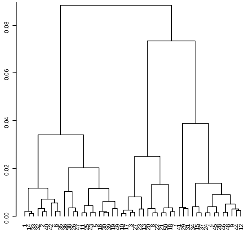
- If many variables : do PCA and keep only first axes $\Longrightarrow$ takes us to classical case
- If many individuals, hierarchical algorithm is too long
  - Use K-means to partition into around 100 classes
  - Build tree using these classes (weighted by the number of individuals in each class)
  - Gives us the “top” of the tree

Hierarchical clustering in high dimension

- If many variables : do PCA and keep only first axes $\Longrightarrow$ takes us to classical case
- If many individuals, hierarchical algorithm is too long
  - Use K-means to partition into around 100 classes
  - Build tree using these classes (weighted by the number of individuals in each class)
  - Gives us the “top” of the tree



Tree from original data



Tree using classes

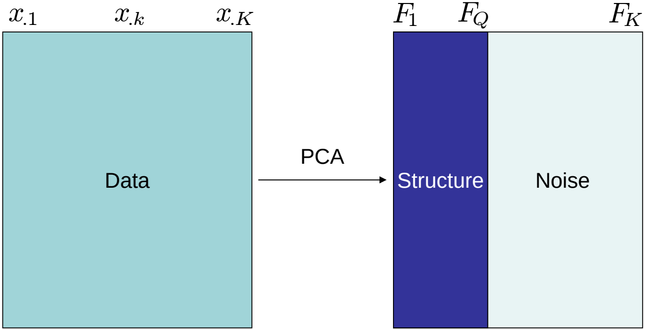
Hierarchical clustering on qualitative data

Two strategies :

- Transform them to quantitative data
  - Do MCA and keep only the first dimensions
  - Do hierarchical clustering using the principal axes of the MCA
- Use measures/indices suitable for qualitative variables : similarity indices, Jaccard index, etc.

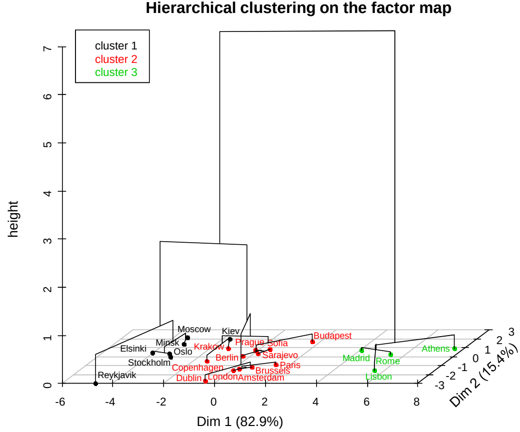
Doing factor analysis followed by clustering

- Qualitative data : MCA outputs quantitative principal components
- Factor analysis eliminates the last components, which are just noise $\Longrightarrow$ more stable clustering



Doing factor analysis followed by clustering

- Representation of the tree and classes on two factor axes $\Longrightarrow$ FA gives continuous information, the tree gives discontinuous information. The tree hints at information hidden in further axes

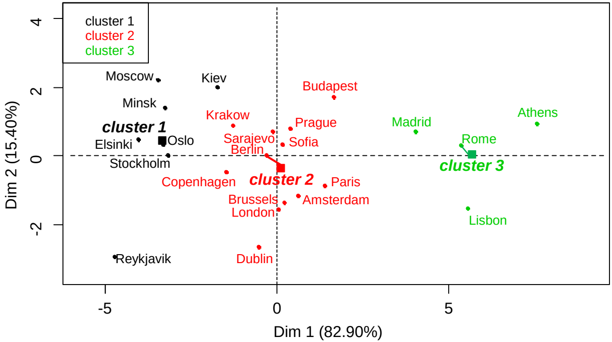


Hierarchical clustering

The class make-up : using “model individuals”

Model individuals : the ones closest to each class center

Cluster 1:	Oslo	Helsinki	Stockholm	Minsk	Moscow
	0.339	0.884	0.9224	0.9654	1.7664
Cluster 2:	Berlin	Sarajevo	Brussels	Prague	Amsterdam
	0.5764	0.7164	1.038	1.0556	1.124
Cluster 3:	Rome	Lisbon	Madrid	Athens	
	0.360	1.737	1.835	2.167	



Characterizing/describing classes

- Goals :
  - Find the variables which are most important for the partition
  - Characterize a class (or group of individuals) in terms of quantitative variables
  - Sort the variables that best describe the classes

- Questions :
  - Which variables best characterize the partition
  - How can we characterize individuals in the 1st class ?
  - Which variables describe them best ?

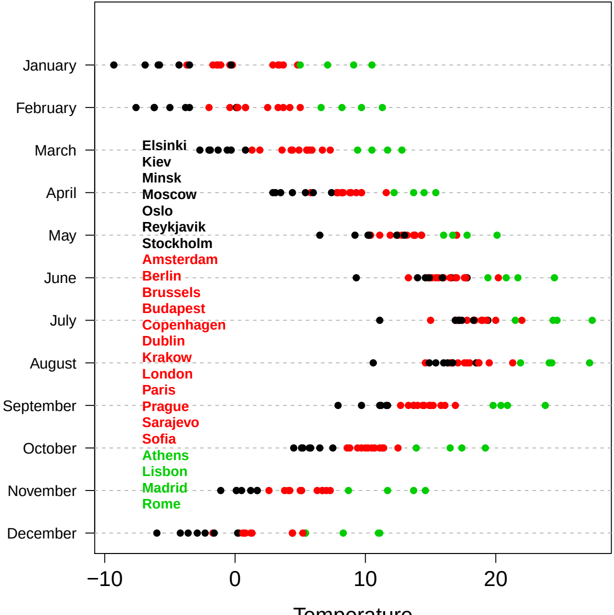
Characterizing/describing classes

Which variables best represent the partition ?

- For each quantitative variable :
  - build an analysis of variance model between the quantitative variable and the class variable
  - do a Fisher test to detect class effect
- Sort the variables by increasing *p*-value

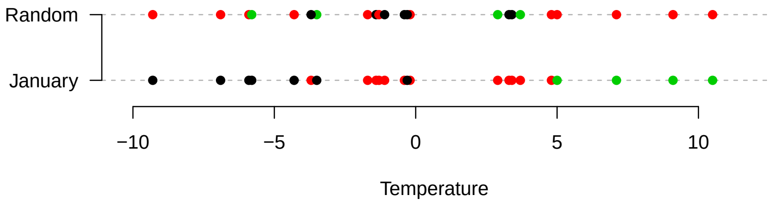
```
              Eta2   P-value
October   0.8990 1.108e-10
March     0.8865 3.556e-10
November  0.8707 1.301e-09
September 0.8560 3.842e-09
April     0.8353 1.466e-08
February  0.8246 2.754e-08
December  0.7730 3.631e-07
January   0.7477 1.047e-06
August    0.7160 3.415e-06
July      0.6309 4.690e-05
May       0.5860 1.479e-04
June      0.5753 1.911e-04
```

Characterizing classes using quantitative variables



Characterizing classes using quantitative variables

1st idea : if the values of X for class q seem to be randomly drawn from all the values of X, then X doesn't characterize class q.



2nd idea : the more a random draw appears unlikely, the more X characterizes class q.

Characterizing classes using quantitative variables

Idea : use as reference a random draw of n_q values from N

What values can $\bar{x}_q$ take ? (i.e., what is the distribution of $\bar{X}_q$?)

$$\mathbb{E}(\bar{X}_q) = \bar{x} \qquad \mathbb{V}(\bar{X}_q) = \frac{s^2}{n_q} \left(\frac{N - n_q}{N - 1} \right)$$

$$\mathcal{L}(\bar{X}_q) = \mathcal{N} \quad \text{because } \bar{X}_q \text{ is a mean}$$

$$\Longrightarrow \quad \text{Test statistic} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{n_q} \left(\frac{N-n_q}{N-1} \right)}} \sim \mathcal{N}(0, 1)$$

- If $|\text{test statistic}| \geq 1.96$ then X characterizes class q
- and the more the test statistic is large, the better X characterizes class q.

Idea : rank the variables by decreasing $|\text{test statistic}|$

Characterizing classes using quantitative variables

$quanti$‘1‘

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
July	-1.99	16.80	18.90	2.450	3.33	0.046100
June	-2.06	14.70	16.80	2.520	3.07	0.039600
August	-2.48	15.50	18.30	2.260	3.53	0.013100
May	-2.55	10.80	13.30	2.430	2.96	0.010800
September	-3.14	11.00	14.70	1.670	3.68	0.001710
January	-3.26	-5.14	0.17	2.630	5.07	0.001130
December	-3.27	-2.91	1.84	1.830	4.52	0.001080
November	-3.36	0.60	5.08	0.940	4.14	0.000781
April	-3.39	4.67	8.38	1.550	3.40	0.000706
February	-3.44	-4.60	0.96	2.340	5.01	0.000577
October	-3.45	5.76	10.10	0.919	3.87	0.000553
March	-3.68	-1.14	4.06	1.100	4.39	0.000238

Characterizing classes using quantitative variables

```
$`2`
NULL

$`3`
          v.test   Mean in  Overall     sd in  Overall   p.value
                  category     mean  category       sd
September   3.81     21.20    14.70      1.54     3.68  0.000140
October     3.72     16.80    10.10      1.91     3.87  0.000201
August      3.71     24.40    18.30      1.88     3.53  0.000211
November    3.69     12.20     5.08      2.26     4.14  0.000222
July        3.60     24.50    18.90      2.09     3.33  0.000314
April       3.53     14.00     8.38      1.18     3.40  0.000413
March       3.45     11.10     4.06      1.27     4.39  0.000564
February    3.43      8.95     0.96      1.74     5.01  0.000593
June        3.39     21.60    16.80      1.86     3.07  0.000700
December    3.39      8.95     1.84      2.34     4.52  0.000706
January     3.29      7.92     0.17      2.08     5.07  0.000993
May         3.18     17.60    13.30      1.55     2.96  0.001460
```

Characterizing classes using qualitative variables

Which variables best characterize the partition ?

- For each qualitative variable, do a χ^2 test between it and the class variable
- Sort the variables by increasing *p*-value

```
$test.chi2
           p.value df
Area   0.001195843  6
```

Characterizing classes using qualitative variables

Does the *South* category characterize the 3rd class ?

	Cluster 3	Other cluster	Total
South	$n_{mc} = 4$	1	$n_m = 5$
Not south	0	18	18
Total	$n_c = 4$	19	$n = 23$

Test : $H_0 : \frac{n_{mc}}{n_c} = \frac{n_m}{n}$ versus H_1 : m abnormally overrepresented in c

Under $H_0 : \mathcal{L}(N_{mc}) = \mathcal{H}(n_c, \frac{n_m}{n}, n) \qquad P_{H_0}(N_{mc} \geq n_{mc})$

Cluster 3

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Area=South	80	100	21.74	0.000564	3.448

$\frac{4}{5} \times 100 = 80$; $\frac{4}{4} \times 100 = 100$; $\frac{5}{23} \times 100 = 21.74$; $P_{\mathcal{H}(4, \frac{5}{23}, 23)}[N_{mc} \geq 4] = 0.000564$

$\Longrightarrow H_0$ rejected, *South* is overrepresented in the 3rd class

Sort the categories in terms of p-values

Characterizing classes using factor axes

These are also quantitative variables

```
$`1`
      v.test   Mean in   Overall     sd in   Overall    p.value
              category      mean  category        sd
Dim.1  -3.32     -3.37         0      0.85      3.15   0.000908

$`2`
      v.test   Mean in   Overall     sd in   Overall    p.value
              category      mean  category        sd
Dim.3  -2.41     -0.18         0      0.22      0.36   0.015776

$`3`
      v.test   Mean in   Overall     sd in   Overall    p.value
              category      mean  category        sd
Dim.1   3.86      5.66         0      1.26      3.15   0.000112
```

Conclusions

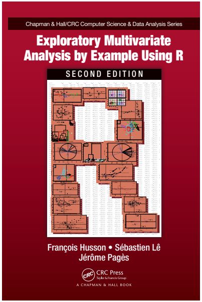
- Clustering can be done on tables of individuals vs quantitative variables
  $\Rightarrow$ MCA transforms qualitative variables into quantitative ones

- hierarchical clustering gives a hierarchical tree $\Rightarrow$ number of classes

- K-means can be used to make classes more robust

- Characterize classes by active and supplementary variables, quantitative or qualitative

More



Husson F., Lê S. & Pagès J. (2017)
Exploratory Multivariate Analysis by Example Using R
2nd edition, 230 p., CRC/Press.

The FactoMineR package for performing clustering :
http://factominer.free.fr/index_fr.html

Movies on Youtube :

- a Youtube channel: youtube.com/HussonFrancois
- a playlist with 11 movies in English
- a playlist with 17 movies in French