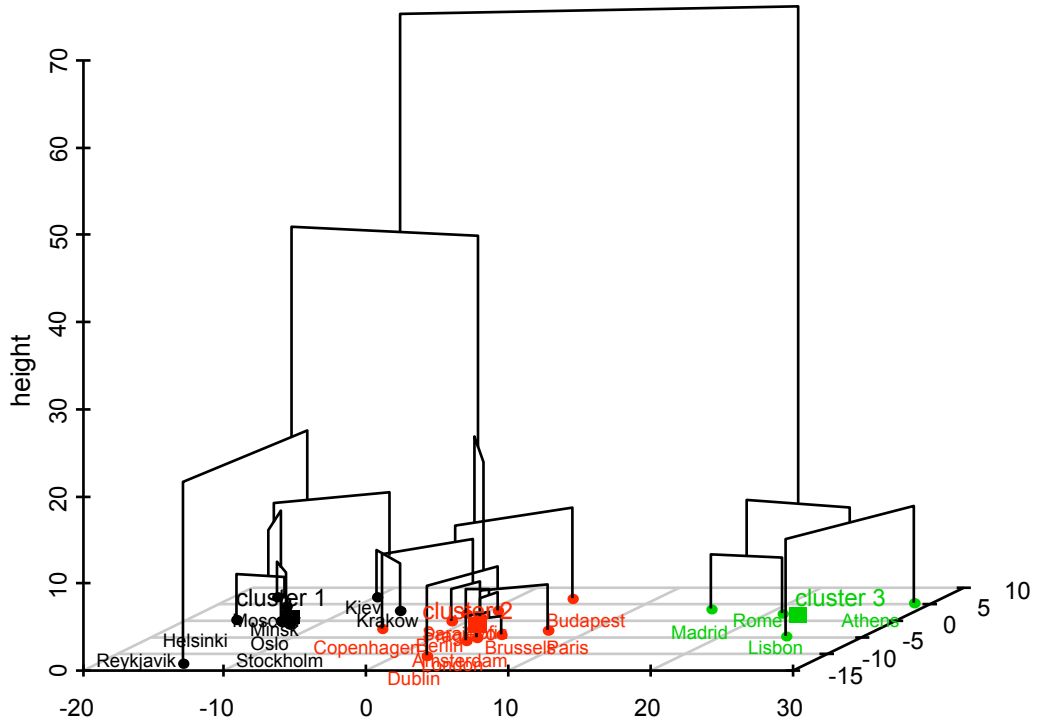
Hierarchical Clustering on Principal Components (HCPC)



LE RAY Guillaume
MOLTO Quentin
Students of AGROCAMPUS OUEST majored in applied statistics

Context

- **R:** A free, opensource software for statistics (1875 packages).
- **FactoMineR:** a R package, developped in Agrocampus-Ouest, dedicated to factorial analysis.
- The aim is to create a complementary tool to this package, **dedicated to clustering**, especially **after a factorial analysis**.
- Wide range of choices and uses, results, and graphical representations.

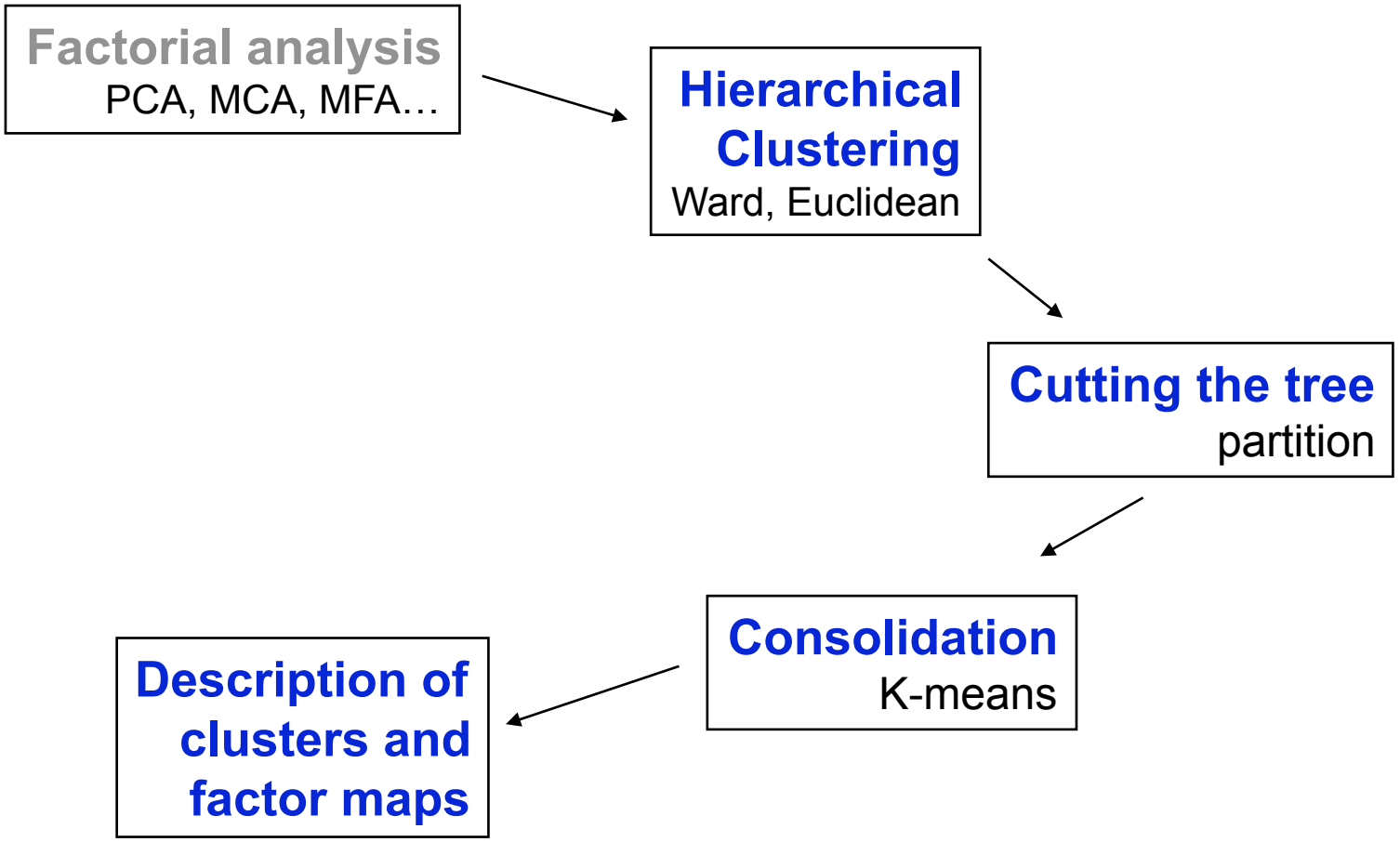
Clustering and factorial analysis

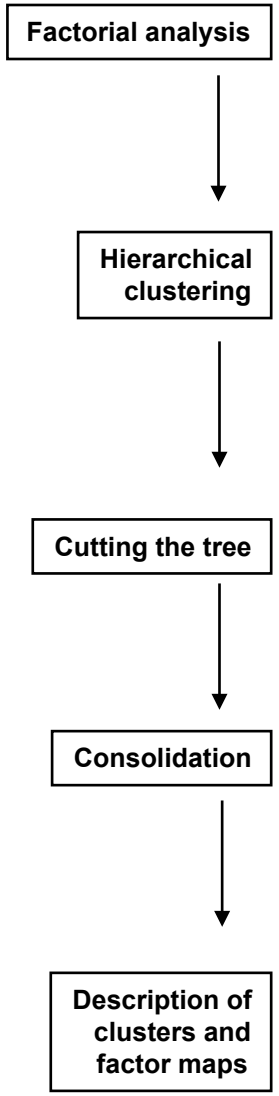
- Factorial analysis and hierarchical clustering are very complementary tools to explore data.

- Removing the last factors of a factorial analysis remove noise and makes the clustering robuster.

Analyses factorielles simples et multiples 4^{éme} édition, Escofier,Pagès 2008

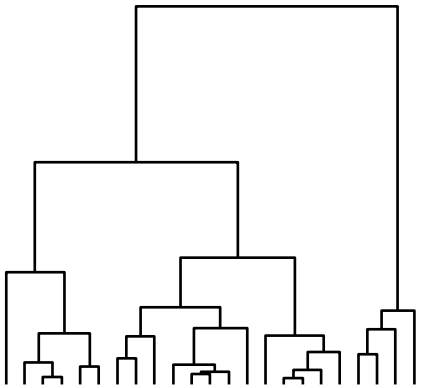
Program structure



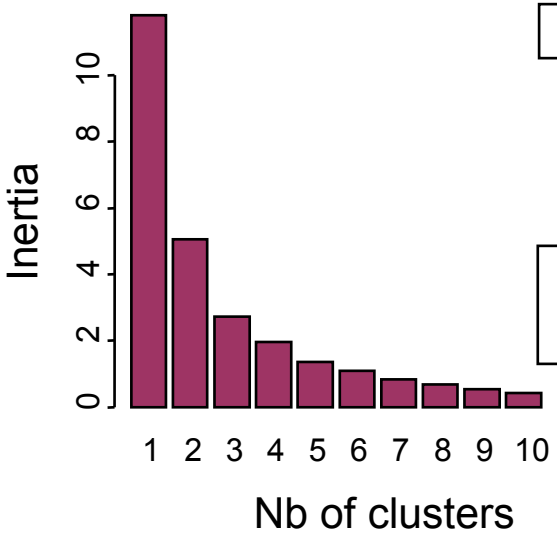


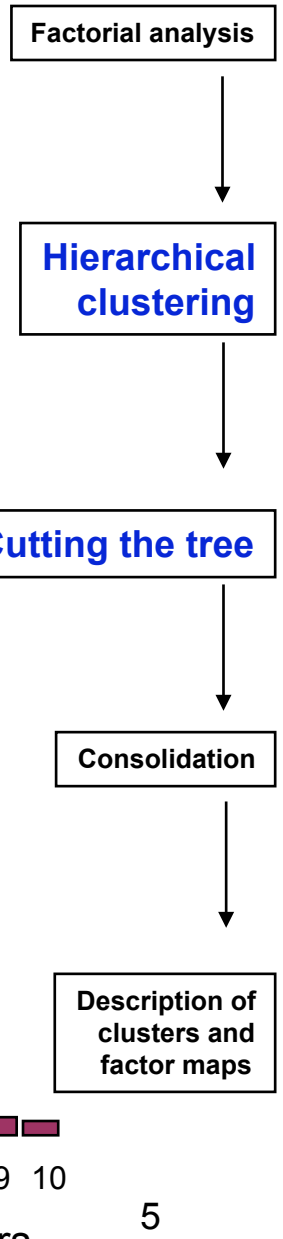
Statistic methods (1)

- **Hierarchical clustering:**
  - Function *agnes*
  - Euclidean distance
  - Ward criterion=$d^2(i,j)x(mi.mj)/(mi+mj)$



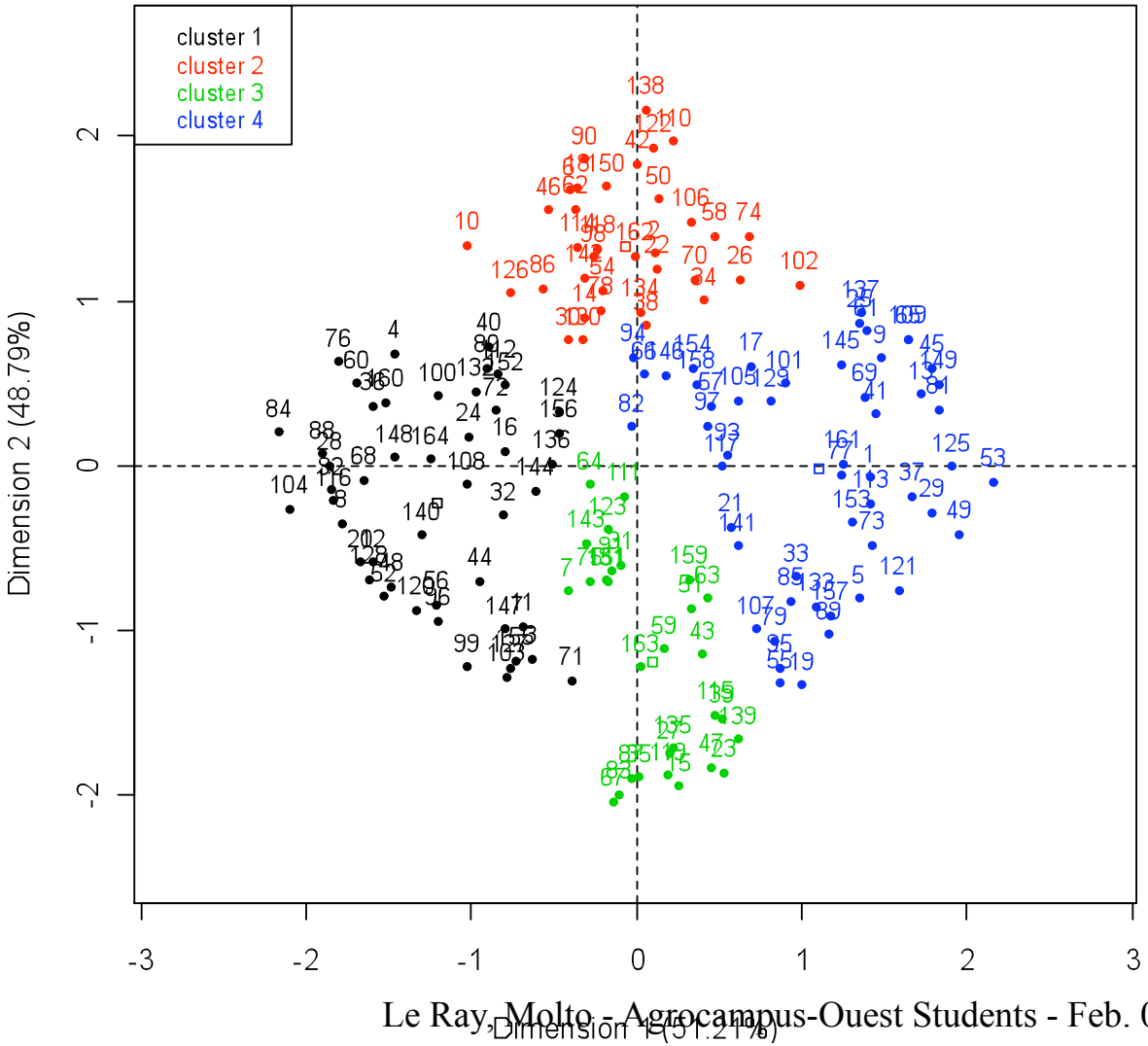
- **Suggested level to cut the tree:**
  - Intra-cluster inertia
  - Partition comparison: $Q=(I_{n+1} - I_n)/I_{n+1}$
  - Max = nb of individuals/2





Statistic methods (2)

Consolidation



- Non optimal partition
- K means with the cluster centers

Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

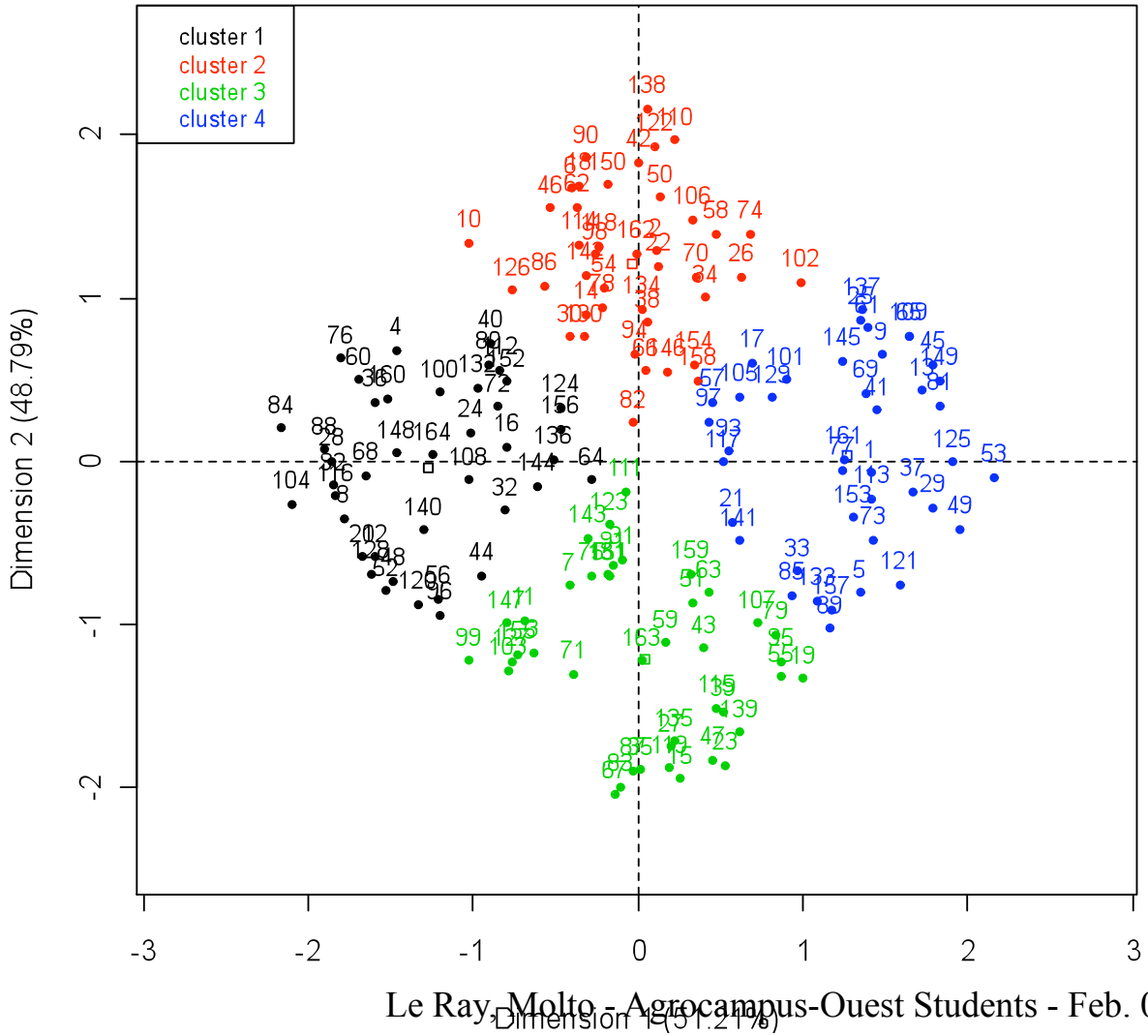
Consolidation

↓

Description of clusters and factor maps

Statistic methods (2)

Consolidation



- Non optimal partition
- K means with the cluster centers

Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

Consolidation

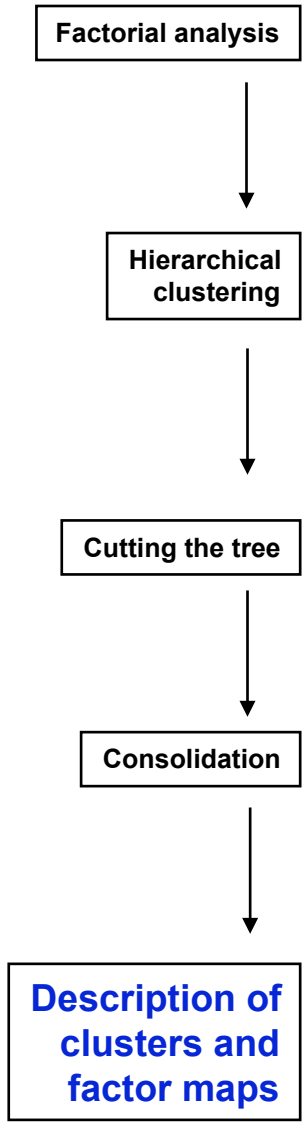
↓

Description of clusters and factor maps

Statistic methods (3)

Clusters description

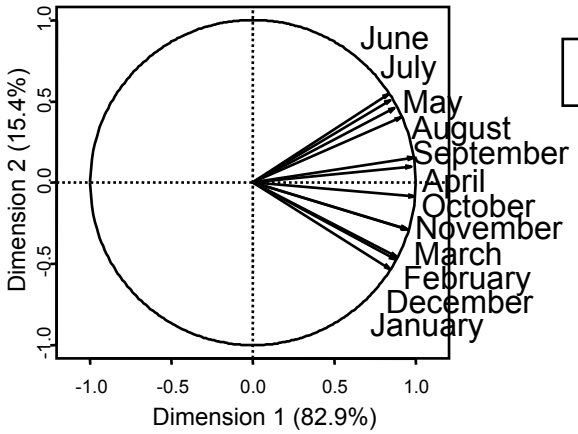
- Description by individuals:
  - Use real individuals to caracterise clusters.
- Description by variables:
  - Give list of typical variable of clusters.
- Description by axes:
  - Like in factorial analysis.

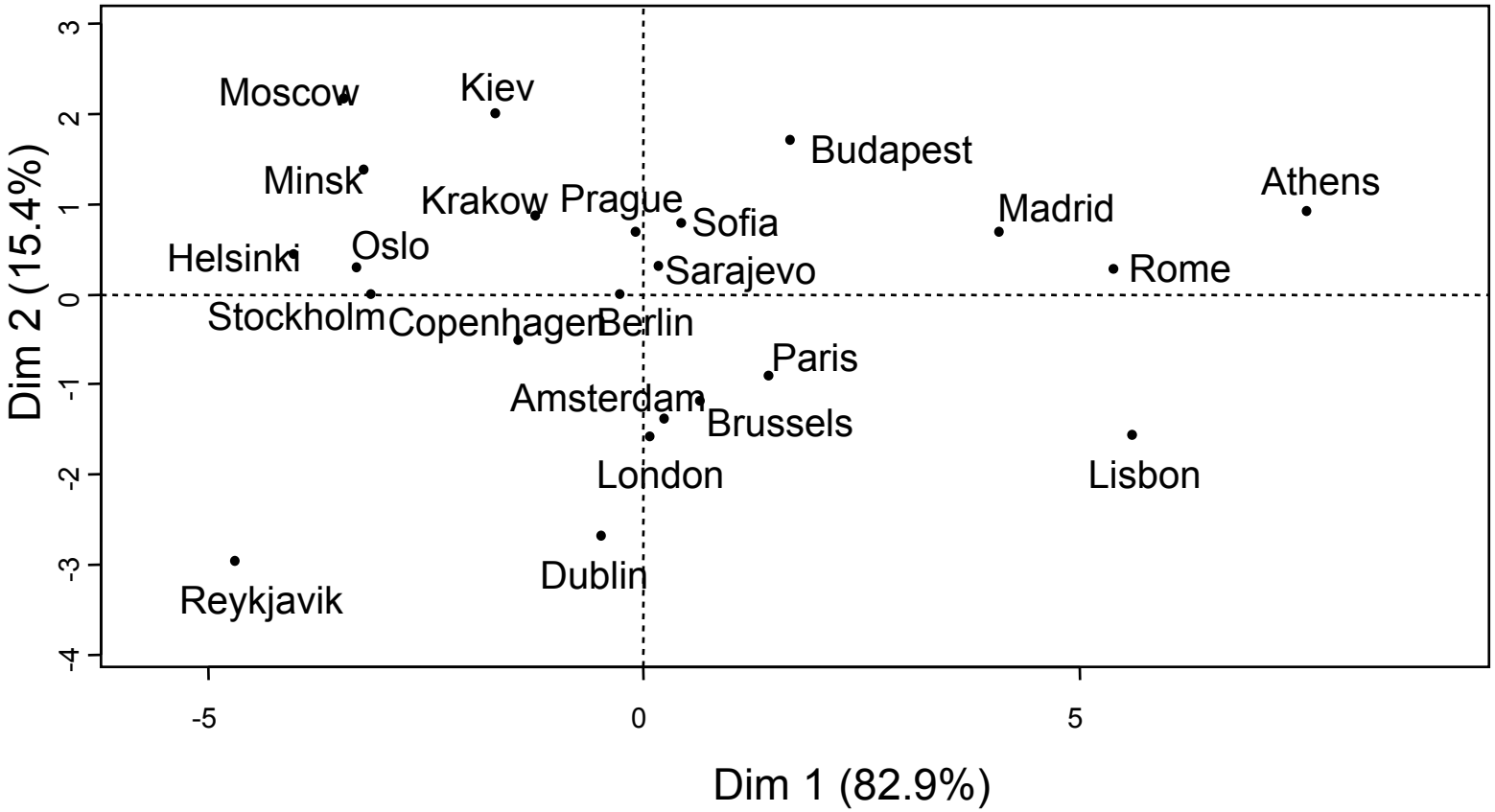


Dataset presentation

row.names	January	February	March	April	May	June	July	August	September	October	November	December
Amsterdam	2.9	2.5	5.7	8.2	12.5	14.8	17.1	17.1	14.5	11.4	7	4.4
Athens	9.1	9.7	11.7	15.4	20.1	24.5	27.4	27.2	23.8	19.2	14.6	11
Berlin	-0.2	0.1	4.4	8.2	13.8	16	18.3	18	14.4	10	4.2	1.2
Brussels	3.3	3.3	6.7	8.9	12.8	15.6	17.8	17.8	15	11.1	6.7	4.4
Budapest	-1.1	0.8	5.5	11.6	17	20.2	22	21.3	16.9	11.3	5.1	0.7
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
Dublin	4.8	5	5.9	7.8	10.4	13.3	15	14.6	12.7	9.7	6.7	5.4
Helsinki	-5.8	-6.2	-2.7	3.1	10.2	14	17.2	14.9	9.7	5.2	0.1	-2.3
Kiev	-5.9	-5	-0.3	7.4	14.3	17.8	19.4	18.5	13.7	7.5	1.2	-3.6
Krakow	-3.7	-2	1.9	7.9	13.2	16.9	18.4	17.6	13.7	8.6	2.6	-1.7
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
London	3.4	4.2	5.5	8.3	11.9	15.1	16.9	16.5	14	10.2	6.3	4.4
Madrid	5	6.6	9.4	12.2	16	20.8	24.7	24.3	19.8	13.9	8.7	5.4
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
Moscow	-9.3	-7.6	-2	6	13	16.6	18.3	16.7	11.2	5.1	-1.1	-6
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
Paris	3.7	3.7	7.3	9.7	13.7	16.5	19	18.7	16.1	12.5	7.3	5.2
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
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
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
Sarajevo	-1.4	0.8	4.9	9.3	13.8	17	18.9	18.7	15.2	10.5	5.1	0.8
Sofia	-1.7	0.2	4.3	9.7	14.3	17.7	20	19.5	15.8	10.7	5	0.6
Stockholm	-3.5	-3.5	-1.3	3.5	9.2	14.6	17.2	16	11.7	6.5	1.7	-1.6

Factorial Analysis





Factorial analysis

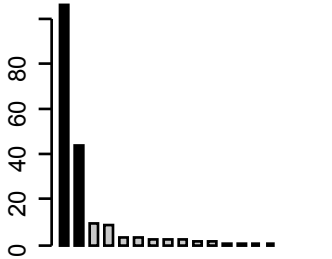
Hierarchical clustering

Cutting the tree

Consolidation

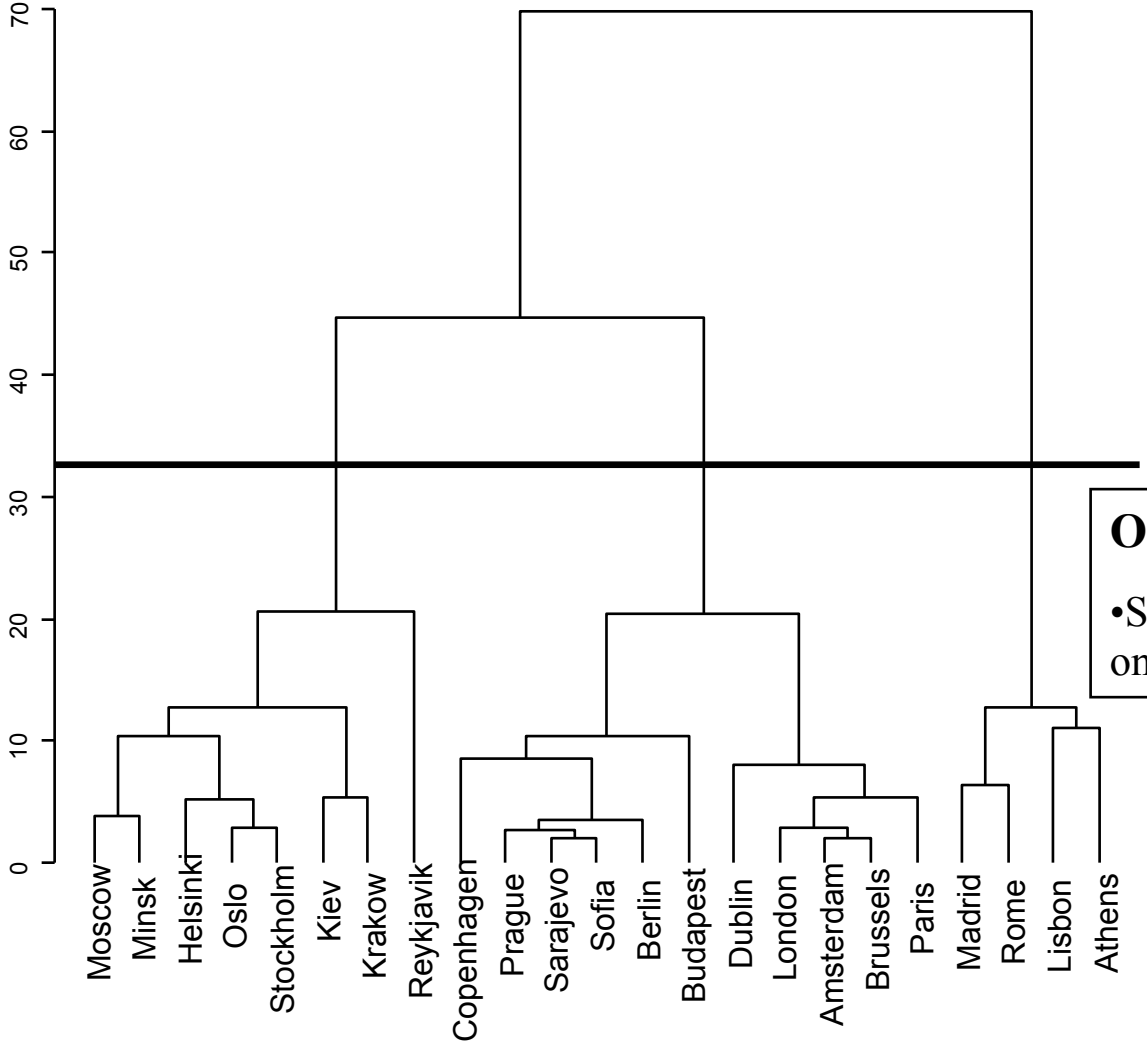
Description of clusters and factor maps

Hierarchical Clustering



inertia gain

Click to cut the tree



suggested level of cutting.

Option:

•Sort the individuals as on the first component.

Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

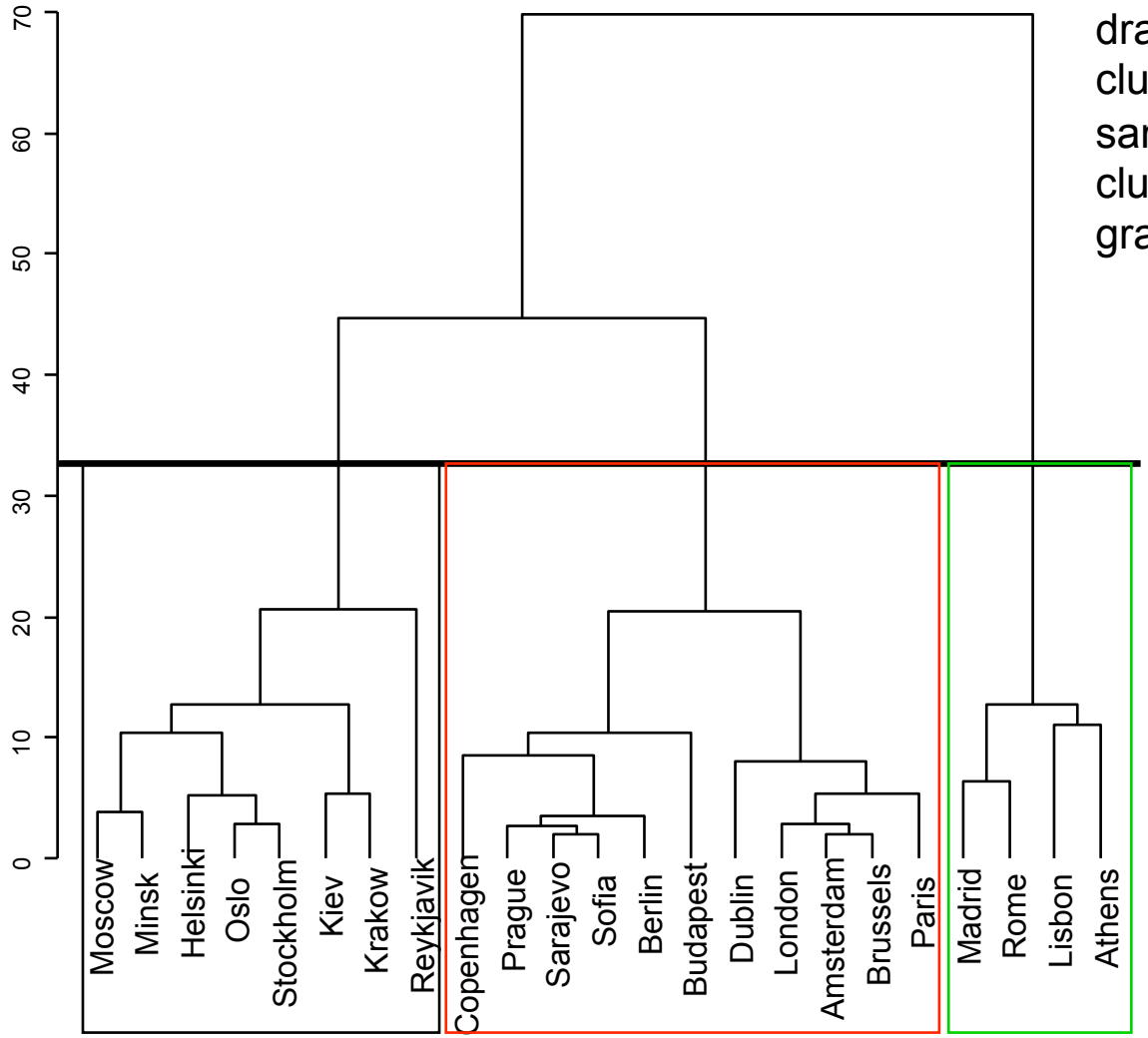
↓

Consolidation

↓

Description of clusters and factor maps

Hierarchical Clustering



Colored rectangles are drawn around the clusters. We keep the same color for each cluster in the next graphs (function *rect*).

Options:

- cut automatically the tree at the suggested level,
- Cut at level with a choosen number of clusters.

Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

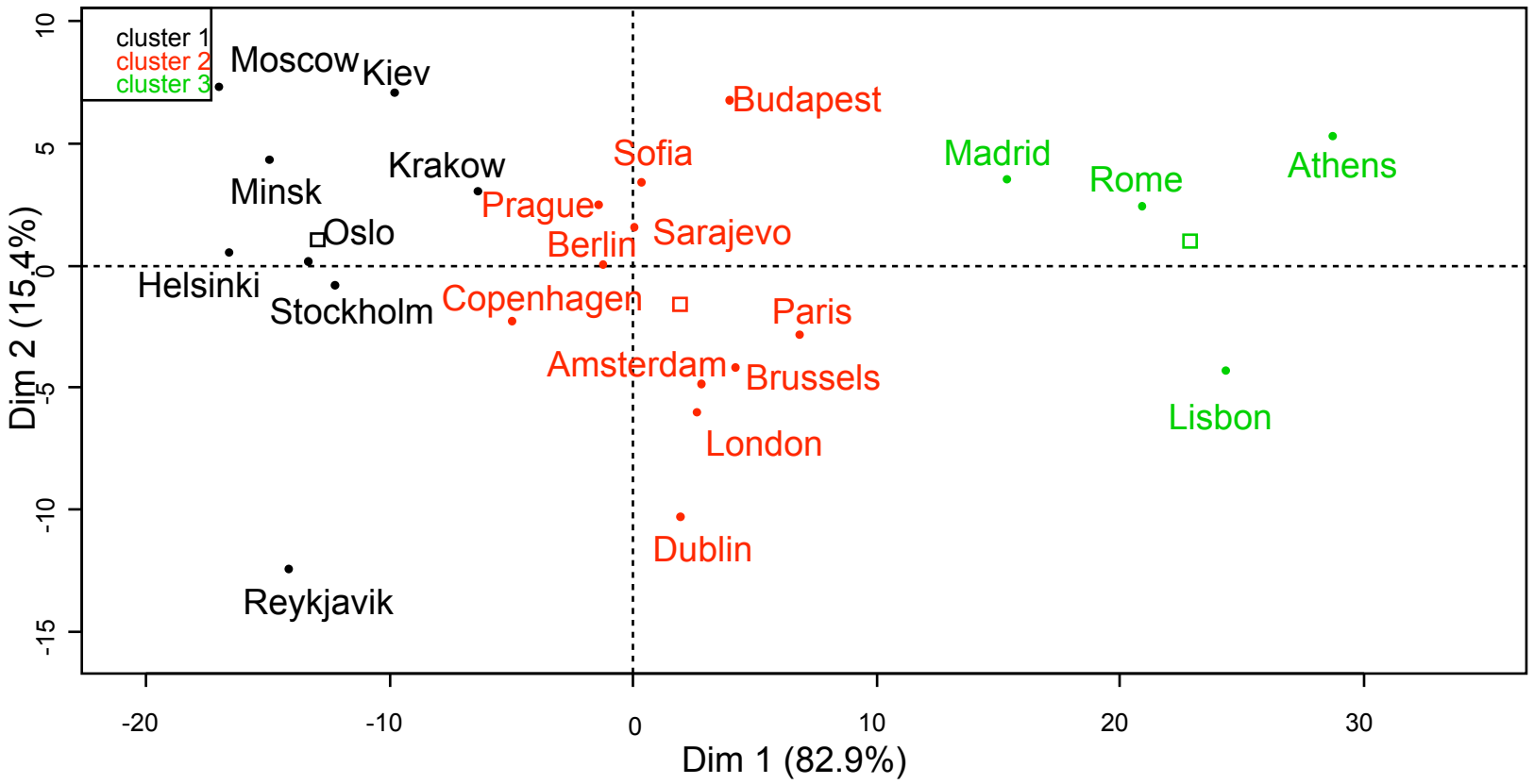
↓

Consolidation

↓

Description of clusters and factor maps

Factor map and clusters



Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

Consolidation

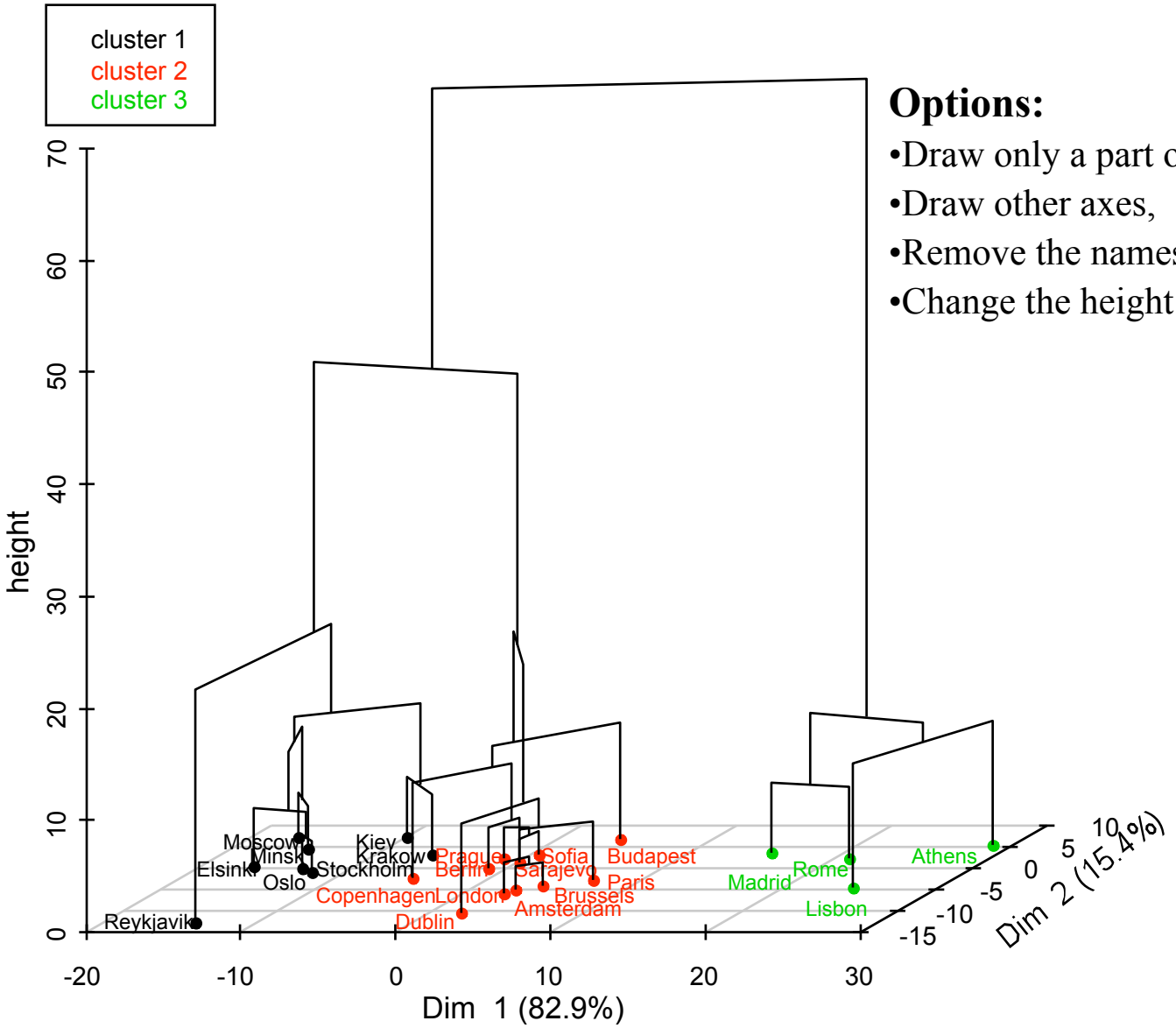
↓

Description of clusters and factor maps

Options:

- Draw other axes,
- Remove the names, the centers.

Factor map, clusters, and tree



Options:

- Draw only a part of the tree,
- Draw other axes,
- Remove the names
- Change the height

Factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

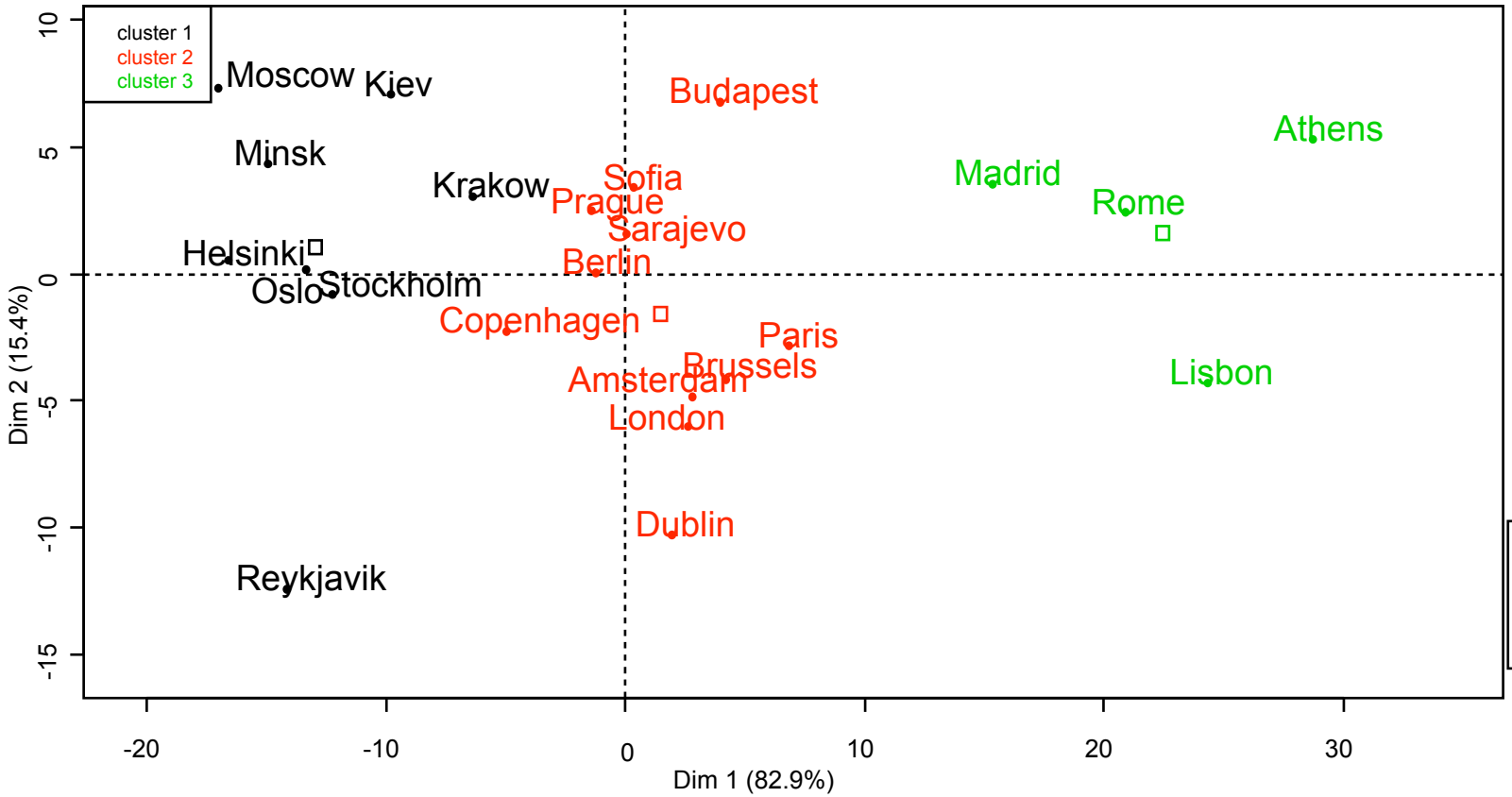
Consolidation

↓

Description of clusters and factor maps

Cluster description (1)
By individuals

Option: the number of individuals for each cluster (here 2)



factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

Consolidation

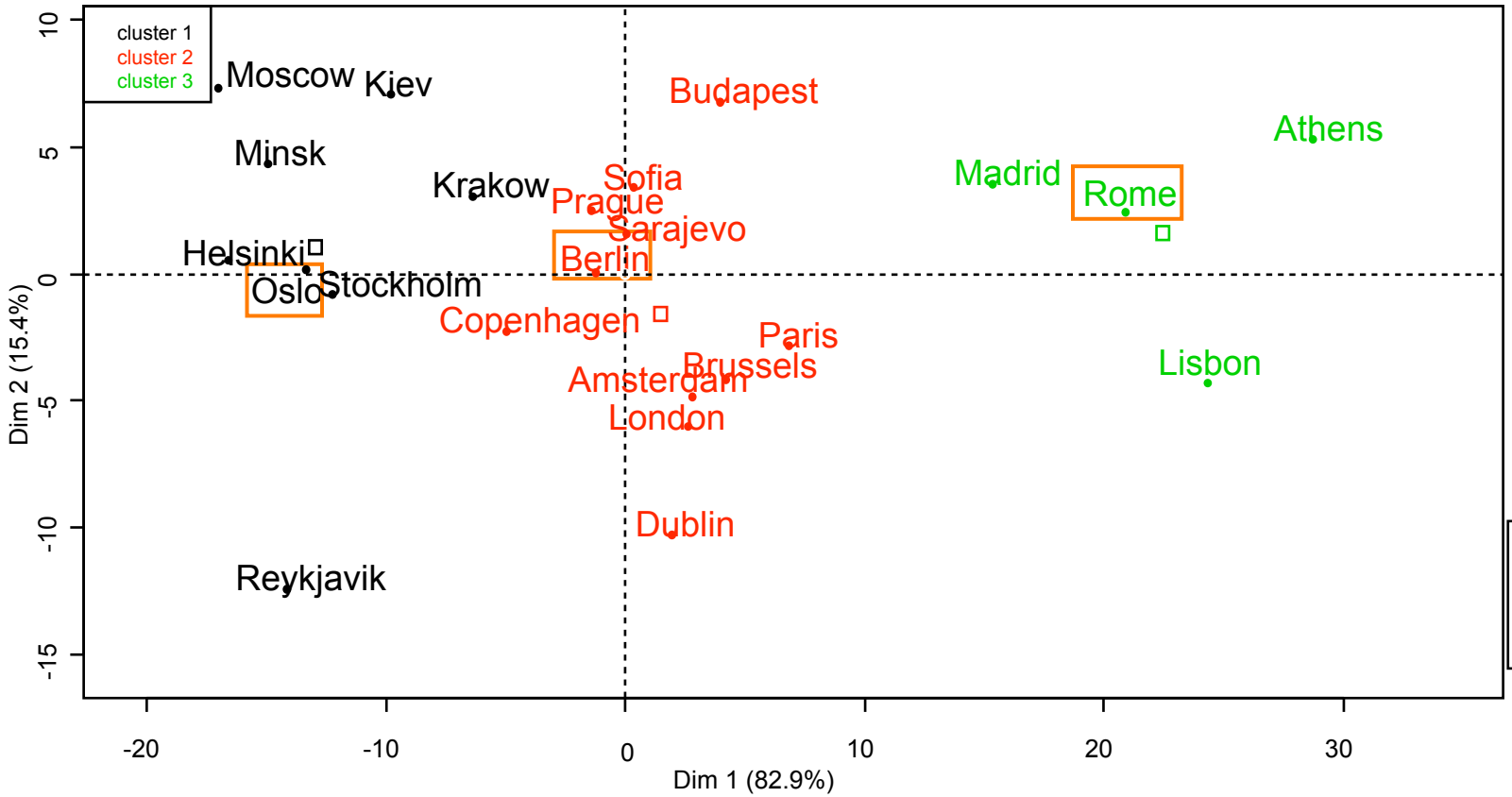
↓

Description of clusters and factor maps

```
> t$desc.ind$para
cluster: 1
      Oslo  Stockholm
  1.533073   3.448958
--------------------
cluster: 2
   Berlin  Sarajevo
 3.180714  3.606640
--------------------
cluster: 3
    Rome    Lisbon
1.612452  6.653195
```

Cluster description (1)
By individuals

Option: the number of individuals for each cluster (here 2)



factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

Consolidation

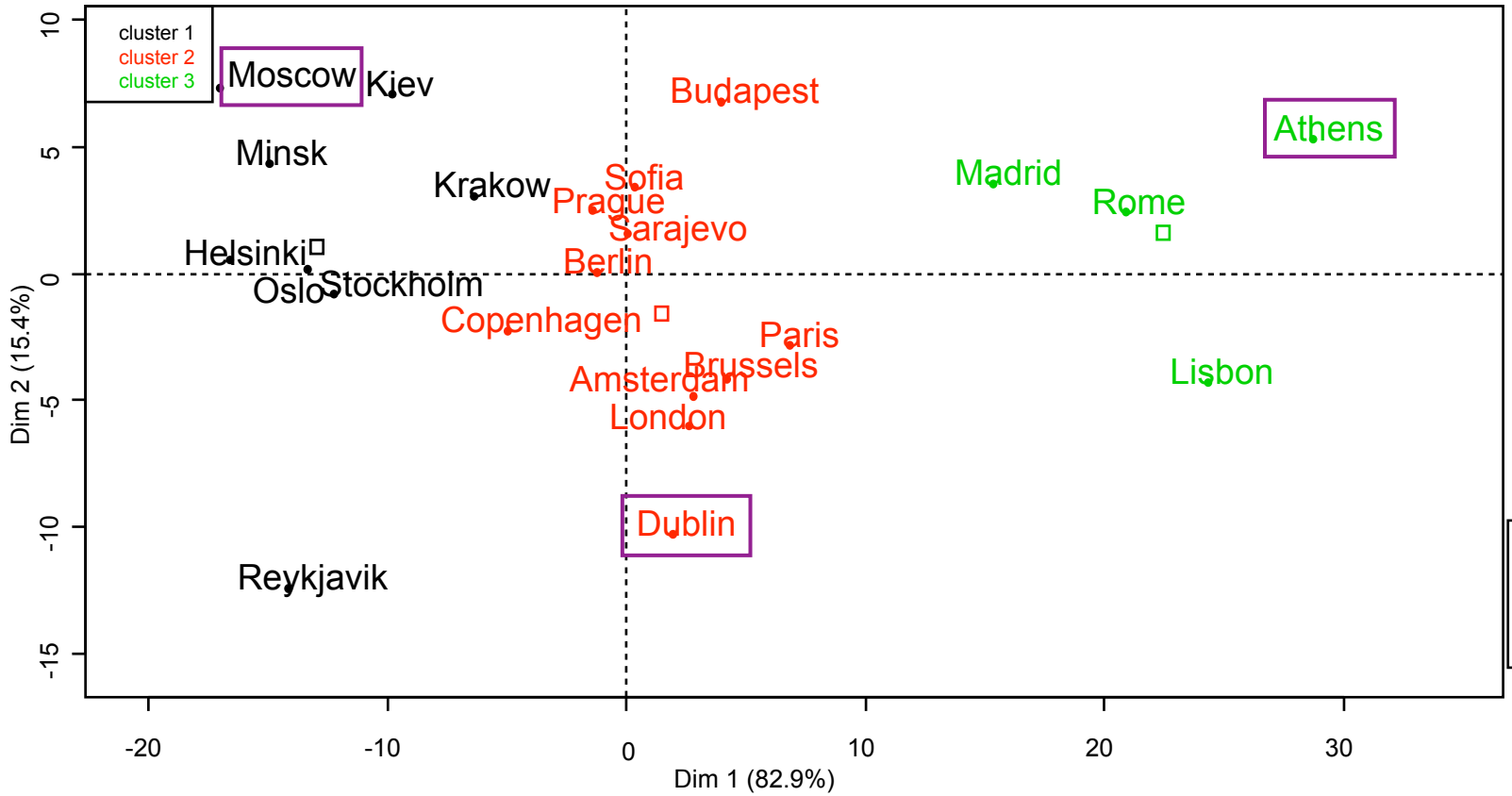
↓

Description of clusters and factor maps

```
> t$desc.ind$dist
cluster: 1
     Moscow Reykjavik
   20.41224   19.02204
----------------------
cluster: 2
 Dublin  Brussels
18.90669 18.10691
----------------------
cluster: 3
 Athens   Lisbon
28.38450 23.14866
```

Cluster description (2)
By individuals

Option: the number of individuals for each cluster (here 2)



factorial analysis

↓

Hierarchical clustering

↓

Cutting the tree

↓

Consolidation

↓

Description of clusters and factor maps

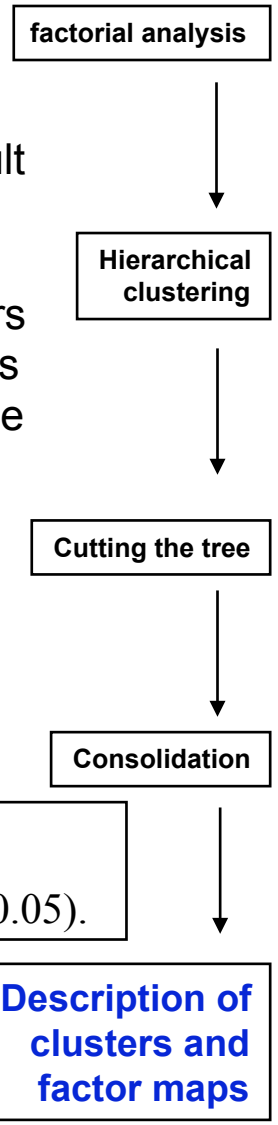
Cluster description (3)
By variables

```
$desc.var$var
$desc.var$var$quanti
$desc.var$var$quanti$`1`
             v.test Mean in category Overall mean sd in category Overall sd
June      -1.969659          15.0000    16.7652174       2.461707   3.069855
July      -1.994828          16.9875    18.9260870       2.357667   3.328822
May       -2.473376          11.1375    13.2739130       2.402050   2.958733
August    -2.481388          15.7500    18.3043478       2.225421   3.526111
September -3.148060          11.3250    14.7086957       1.800521   3.681790
April     -3.332580           5.0750     8.3782609       1.799826   3.395260
January   -3.472699          -4.9625     0.1739130       2.501468   5.066447
December  -3.494357          -2.7625     1.8434783       1.758506   4.515079
October   -3.498795           6.1125    10.0652174       1.274203   3.869795
November  -3.501940           0.8500     5.0782609       1.100000   4.135841
February  -3.578172          -4.2750     0.9565217       2.348803   5.008152
March     -3.761097          -0.7625     4.0608696       1.442166   4.392854

$desc.var$var$quanti$`2`
NULL

$desc.var$var$quanti$`3`
             v.test Mean in category Overall mean sd in category Overall sd
September 3.808964          21.225    14.7086957       1.536839   3.681790
October   3.717610          16.750    10.0652174       1.911151   3.869795
August    3.705134          24.375    18.3043478       1.883315   3.526111
November  3.692832          12.175     5.0782609       2.264260   4.135841
July      3.603579          24.500    18.9260870       2.089258   3.328822
April     3.531688          13.950     8.3782609       1.175798   3.395260
March     3.448552          11.100     4.0608696       1.274755   4.392854
February  3.434969           8.950     0.9565217       1.744276   5.008152
June      3.389407          21.600    16.7652174       1.864135   3.069855
December  3.387321           8.950     1.8434783       2.337199   4.515079
January   3.292484           7.925     0.1739130       2.076505   5.066447
May       3.183059          17.650    13.2739130       1.553222   2.958733
```

This is the result of a catdes, it describes the different clusters by the variables (the mean in the category, the v.test…)



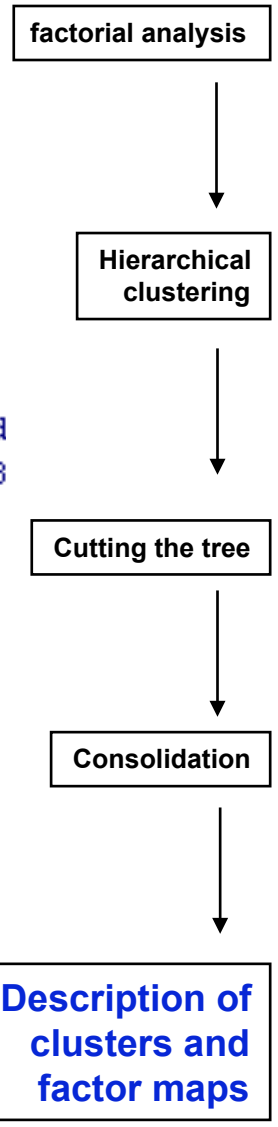
Option:
the p.value (here 0.05).

Cluster description (3)
By axes

```
$axe
$axe$quanti
$axe$quanti$`1`
         v.test Mean in category Overall mean sd in category Overall sd
Dim.1 -3.539546        -13.05878 8.285944e-16       3.334186   12.63763

$axe$quanti$`2`
         v.test Mean in category  Overall mean sd in category Overall sd
Dim.3 -2.083096       -0.5968921 -8.634209e-16      0.8911895   1.286778

$axe$quanti$`3`
        v.test Mean in category Overall mean sd in category Overall sd
Dim.1 3.804897         22.34313 8.285944e-16       4.891736   12.63763
```



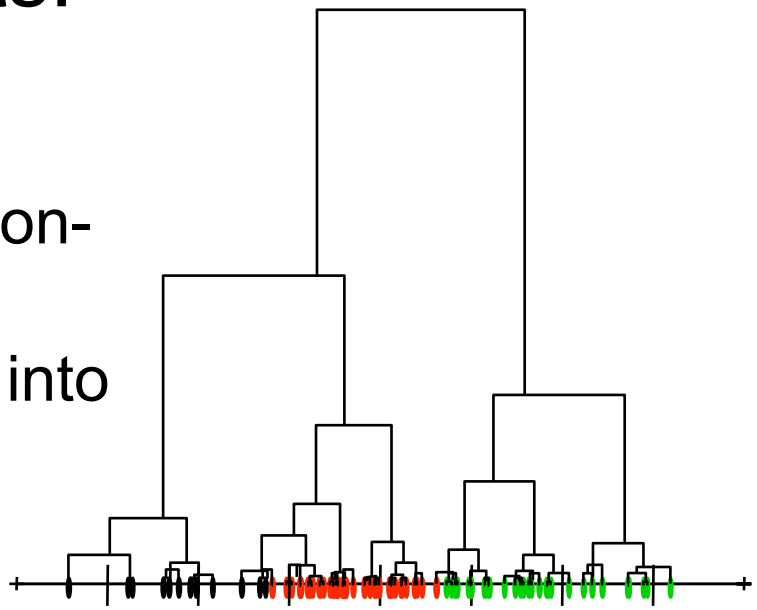
This is the result of a catdes, it describes the different clusters by the axes (the mean in the category, the v.test…)

Option:
the p.value (here 0.05).

Conclusion

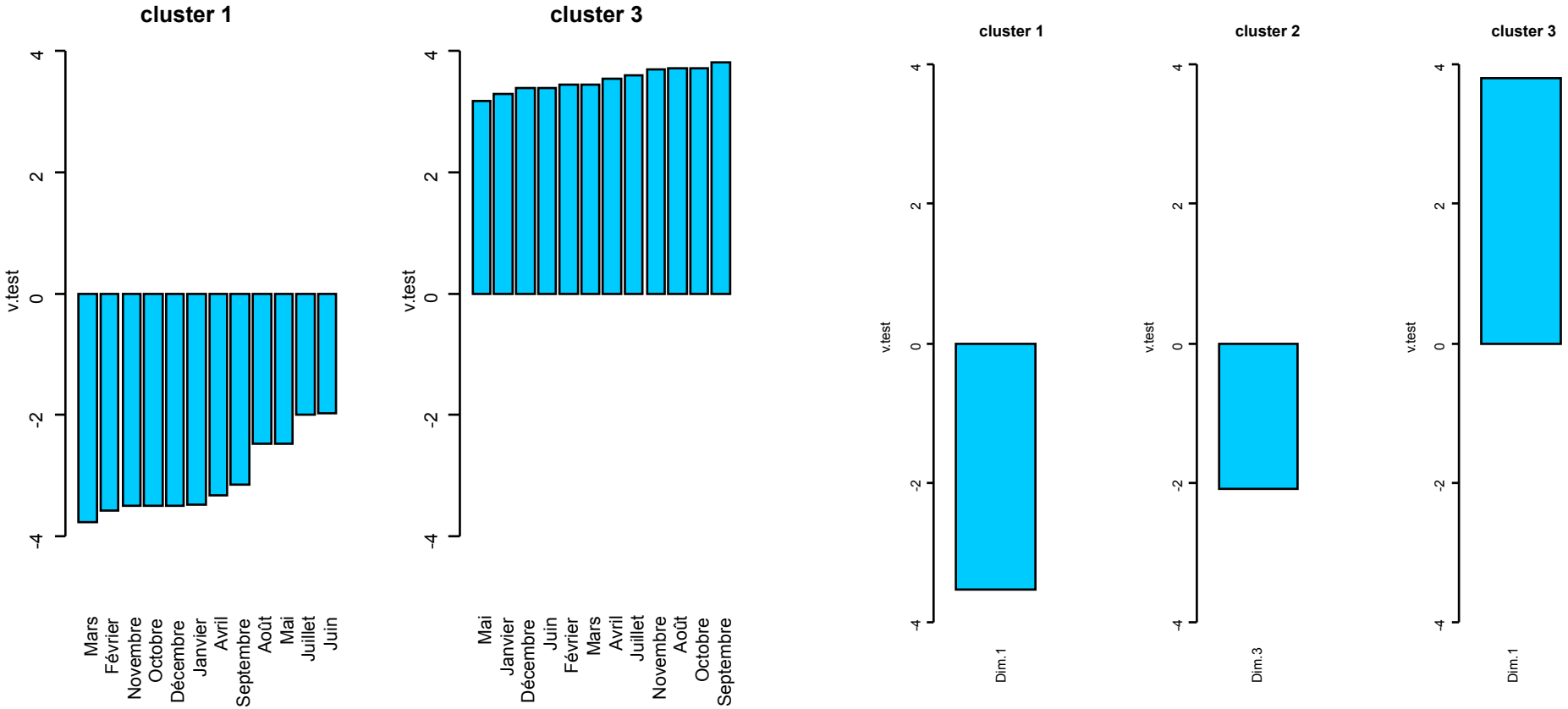
This function was presented with a PCA, but it also acepts:

- MCA and MFA results,
- directly a quantitative dataset (non-scaled PCA),
- a continuous variables to divide into modalities.



A normal distribution divided in 3 clusters

Function plot.catdes



It is a graphical representation of the desc.var results

Option:

- show only the quantitative, qualitative variables or all