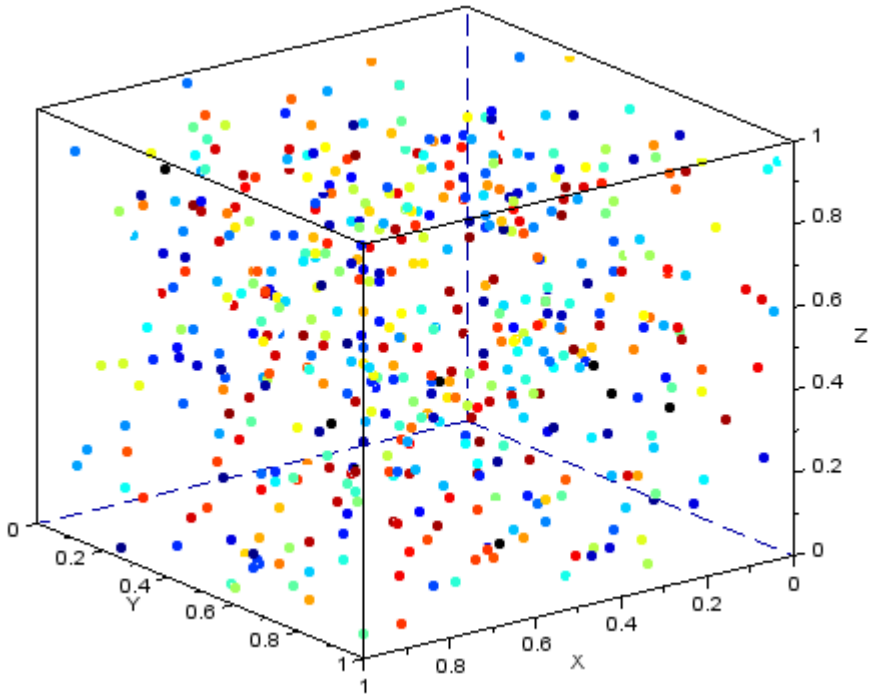
Exploratory data analysis in environmental health

Stéphane Joost & Mayssam Nehme

Hierarchical Ascendant Classification

Hierarchical Ascendant Classification - HAC



Euclidian distances

Values are standardized:

Centered (x - mean) *and*

Reduced (divided by std)

N dimensions (1 per variable)

Hierarchical Ascendant Classification - HAC

- Searching for homogeneous groups of individuals (clusters) in the population:
  - Two individuals belonging to the same group are close to each other (similar behaviors)
  - Two individuals belonging to different groups are far from each other (different behaviors);
- Build a partition of the population into homogeneous clusters (low within-variability) which are different one from the other (high between-variability)

How to calculate the pairwise distances?

Socio-eco variables

Stat. sectors

	V1	V2
A	2	5
B	7	8
C	3	3
D	8	9
E	4	5

$$d_{A*B}=\sqrt{(2-7)^2+(5-8)^2} \rightarrow d_{A*B}=5.83$$

$$d_{A*C}=\sqrt{(2-3)^2+(5-3)^2} \rightarrow d_{A*C}=2.23$$

$$d_{A*D}=\sqrt{(2-8)^2+(5-9)^2} \rightarrow d_{A*D}=7.21$$

Table of Euclidian distances

	A	B	C	D	E
A	0				
B	5.83	0			
C	2.23	6.40	0		
D	7.21	1.41	7.81	0	
E	2.00	4.24	2.23	5.65	0

Grouping the closest points

- The minimum distance is d_{B*D} , B and D are the closest individuals
- B and D are grouped at a level of 1.41

	A	B	C	D	E
A	0				
B	5.83	0			
C	2.23	6.40	0		
D	7.21	1.41	7.81	0	
E	2.00	4.24	2.23	5.65	0

Successive table of distances

- To calculate the distance between group BD and the rest of the individuals, the mean distance is used (aggregation criterion)

$$d_{A*BD} = \frac{d_{AB} + d_{AD}}{2} \rightarrow d_{A*BD} = 6.52$$

	A	BD	C	E
A	0			
BD	6.52	0		
C	2.23	7.05	0	
E	2.00	4.94	2.23	0

- The minimum distance is d_{A*E} , the closest individuals are A and E
- A and E are grouped at a level of 2.00

Successive table of distances

- We need to calculate the distance between group AE and the rest of the individuals

	AE	BD	C
AE	0		
BD	5.73	0	
C	2.23	7.05	0

$$d_{AE*C} = \frac{d_{AC} + d_{EC}}{2} \rightarrow d_{AE*C} = 2.23$$

$$d_{AE*BD} = \frac{d_{A*BD} + d_{E*BD}}{2} \rightarrow d_{AE*BD} = 5.73$$

- The minimal distance is d_{C*AE} , the closest individuals are AE and C
- AE and C are grouped at a level of 2.23

Successive table of distances

- We need to calculate the distance between group AEC and the rest of the individuals

	AEC	BD
AEC	0	6,39
BD	6,39	0

$$d_{AEC*BD} = \frac{d_{AE*BD} + d_{C*BD}}{2} \rightarrow d_{AEC*BD} = 6.39$$

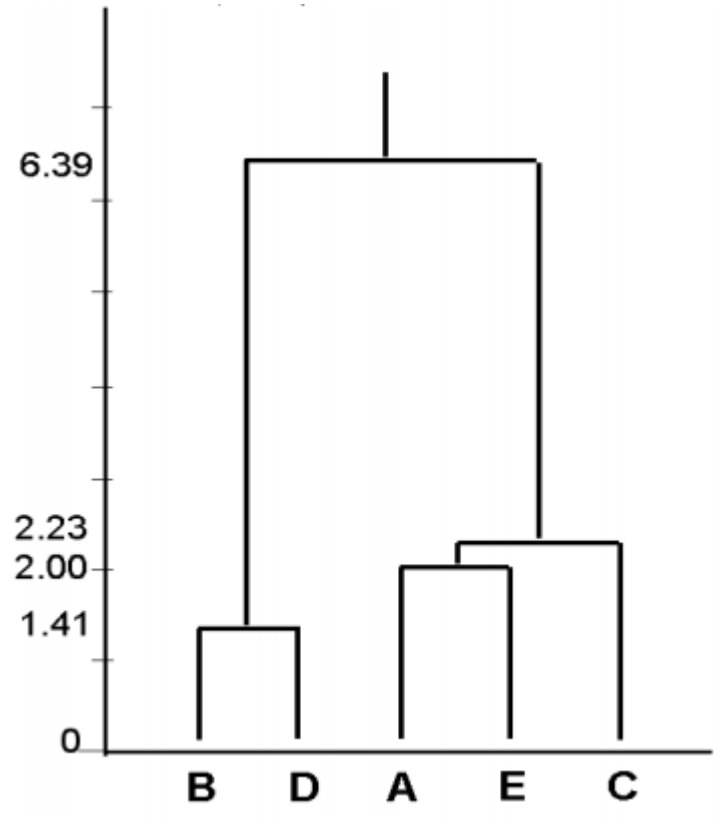
- AEC and BDC are grouped at a level of 6.39

Summary

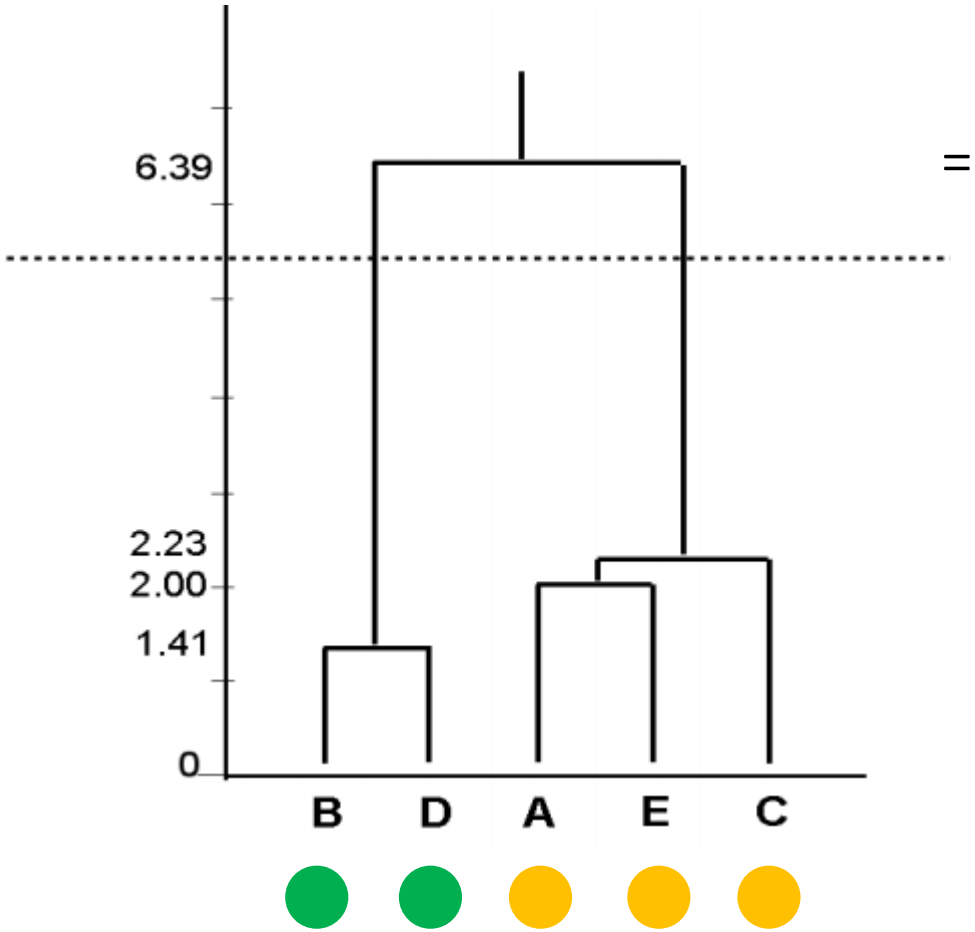
- B and D are grouped at a level of 1.41
- A and E are grouped at a level of 2.00
- AE and C are grouped at a level of 2.23
- AEC and BD are grouped at a level of 6.39

- A dendrogram is used to visualize these results

Dendrogram



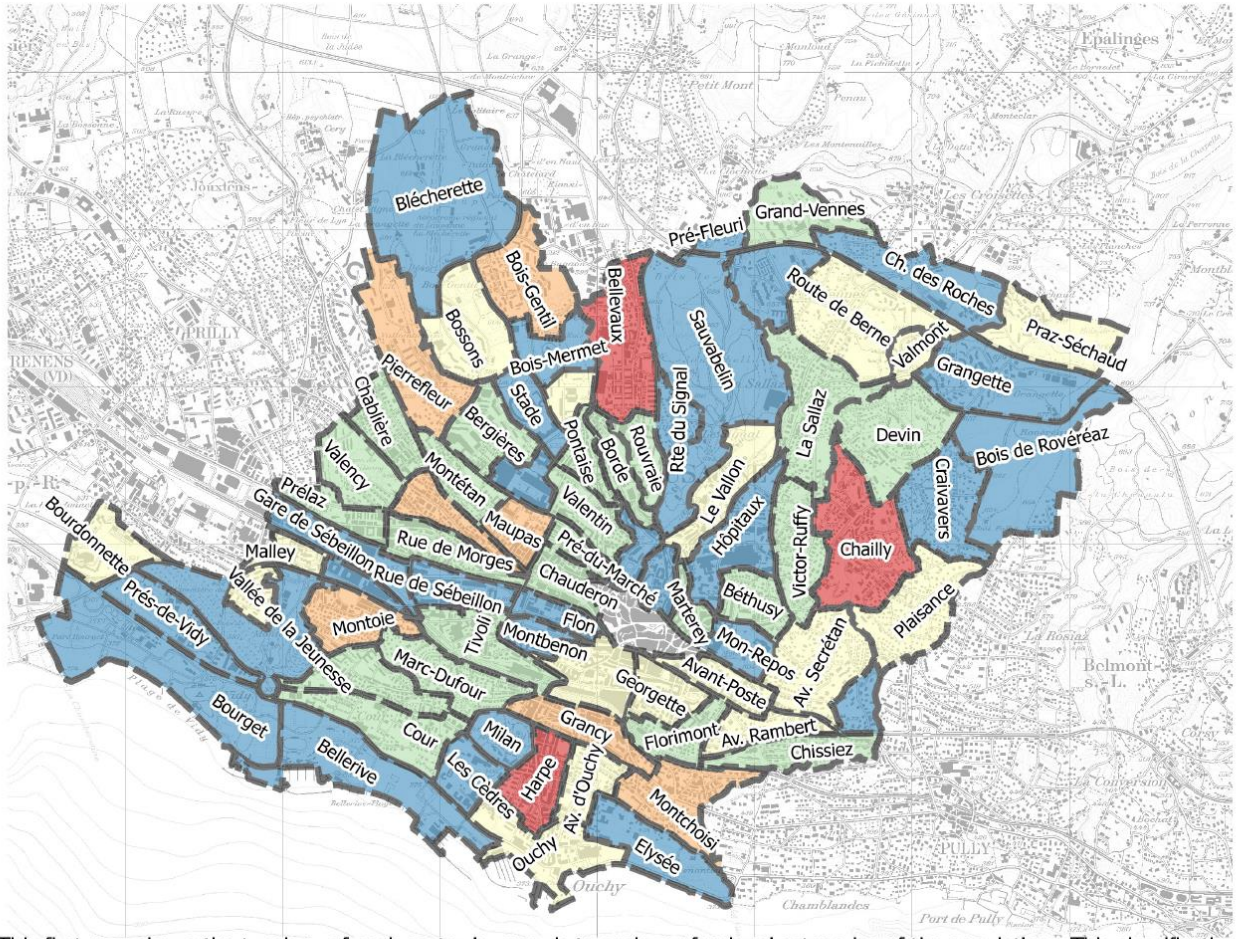
Choice of a number of classes



= number of branches cut

Assignment 1, First Map : Typology of Lausanne sectors, classification depending on socio-professional category
Mathieu Plourde, Civil Engeneering, Master 4, 20/03/2016

Nord

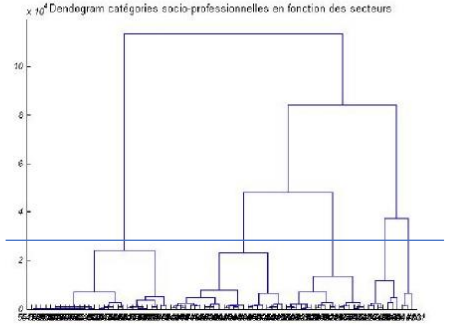


Typology of Lausanne sectors

- Group 1
- Group 2
- Group 3
- Group 4
- Group 5

Données : http://www.scris-lausanne.vd.ch/, population active selon les différentes catégories socio-professionnelles, quartiers statistiques Lausannois

0 1 2 km



This first map shows the typology of each sector in regards to socio-professional categories of the population. This classification was obtained using the Ward's minimum variance method. The grouping shows homogeneous groups of individuals in the population. We observed a concentric pattern in the city of Lausanne : light green, darker green, orange and then blue outside the center. The first group, colored in red, seems arbitrary distributed amongst the sectors.
We can also noticed clusters of sectors in the same group. It would be interesting to compare this map with real estate prices to see if there are similar patterns.


Thank you for your attention!