Classification ascendante hiérarchique

Classification automatique, typologie, clustering

Ricco RAKOTOMALALA
Université Lumière Lyon 2

PLAN

La classification automatique

La typologie ou le Clustering ou l'Apprentissage non-supervisé

Classification automatique

Typologie, apprentissage non-supervisé, clustering

X (tous quantitatifs)

Pas de Y à prédire

Modele	Prix	Cylindree	Puissance	Poids	Consommation	Groupe
Daihatsu Cuore	11600	846	32	650	5.7	
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8	
Fiat Panda Mambo L	10450	899	29	730	6.1	
VW Polo 1.4 60	17140	1390	44	955	6.5	
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8	
Subaru Vivio 4WD	13730	658	32	740	6.8	
Toyota Corolla	19490	1331	55	1010	7.1	
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4	
Peugeot 306 XS 108	22350	1761	74	1100	9	
Renault Safrane 2.2. V	36600	2165	101	1500	11.7	
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5	
VW Golt 2.0 GTI	31580	1984	85	1155	9.5	
Citroen ZX Volcane	28750	1998	89	1140	8.8	
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3	
Fort Escort 1.4i PT	20300	1390	54	1110	8.6	
Honda Civic Joker 1.4	19900	1396	66	1140	7.7	
Volvo 850 2.5	39800	2435	106	1370	10.8	
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6	
Hyundai Sonata 3000	38990	2972	107	1400	11.7	
Lancia K 3.0 LS	50800	2958	150	1550	11.9	
Mazda Hachtback V	36200	2497	122	1330	10.8	
Mitsubishi Galant	31990	1998	66	1300	7.6	
Opel Omega 2.5i V6	47700	2496	125	1670	11.3	
Peugeot 806 2.0	36950	1998	89	1560	10.8	
Nissan Primera 2.0	26950	1997	92	1240	9.2	
Seat Alhambra 2.0	36400	1984	85	1635	11.6	
Toyota Previa salon	50900	2438	97	1800	12.8	
Volvo 960 Kombi aut	49300	2473	125	1570	12.7	

?

Identifier les catégories (groupes) de voitures « similaires » (c.-à-d. qui se ressemblent)

Objectif : identifier des groupes d'observations ayant des caractéristiques similaires (ex. comportement d'achats de clients, caractère « polluant » de véhicules, etc.)

On veut que :

(1) Les individus dans un même groupe se ressemblent le plus possible

(2) Les individus dans des groupes différents se démarquent le plus possible

Pourquoi ?

- → Identifier des structures sous-jacentes dans les données
- → Résumer des comportements
- → Affecter de nouveaux individus à des catégories
- → Identifier les cas totalement atypiques

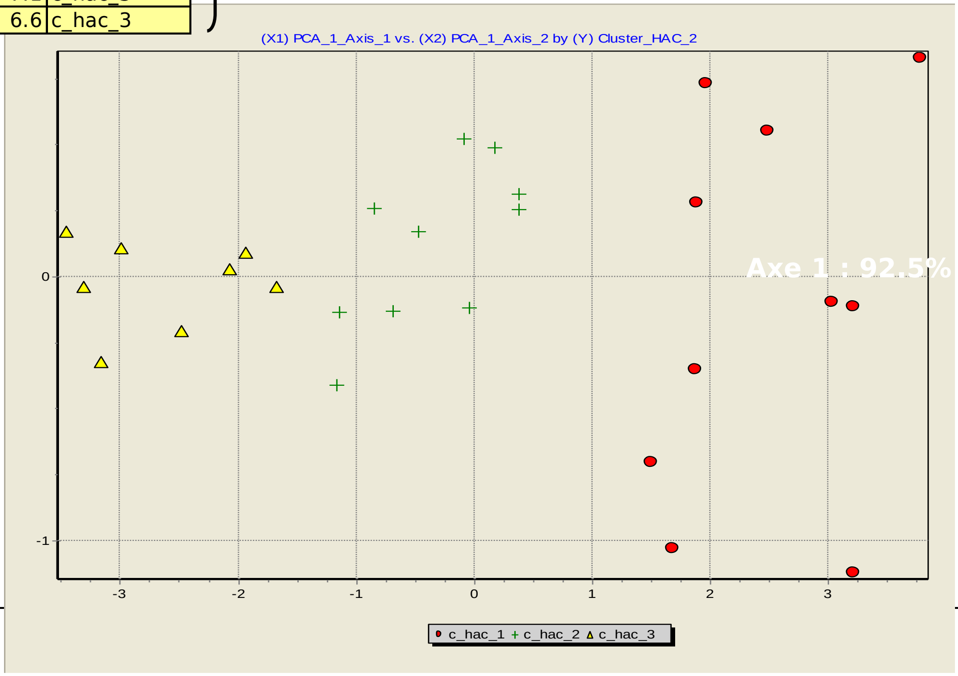
Modele	Prix	Cylindree	Puissance	Poids	Consommation	Groupe
Renault Safrane 2.2. V	36600	2165	101	1500	11.7	c_hac_1
Volvo 850 2.5	39800	2435	106	1370	10.8	c_hac_1
Hyundai Sonata 3000	38990	2972	107	1400	11.7	c_hac_1
Lancia K 3.0 LS	50800	2958	150	1550	11.9	c_hac_1
Mazda Hachtback V	36200	2497	122	1330	10.8	c_hac_1
Opel Omega 2.5i V6	47700	2496	125	1670	11.3	c_hac_1
Peugeot 806 2.0	36950	1998	89	1560	10.8	c_hac_1
Seat Alhambra 2.0	36400	1984	85	1635	11.6	c_hac_1
Toyota Previa salon	50900	2438	97	1800	12.8	c_hac_1
Volvo 960 Kombi aut	49300	2473	125	1570	12.7	c_hac_1
Opel Astra 1.6i 16V	25000	1597	74	1080	7.4	c_hac_2
Peugeot 306 XS 108	22350	1761	74	1100	9	c_hac_2
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5	c_hac_2
VW Golt 2.0 GTI	31580	1984	85	1155	9.5	c_hac_2
Citroen ZX Volcane	28750	1998	89	1140	8.8	c_hac_2
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3	c_hac_2
Fort Escort 1.4i PT	20300	1390	54	1110	8.6	c_hac_2
Honda Civic Joker 1.4	19900	1396	66	1140	7.7	c_hac_2
Mitsubishi Galant	31990	1998	66	1300	7.6	c_hac_2
Nissan Primera 2.0	26950	1997	92	1240	9.2	c_hac_2
Daihatsu Cuore	11600	846	32	650	5.7	c_hac_3
Suzuki Swift 1.0 GLS	12490	993	39	790	5.8	c_hac_3
Fiat Panda Mambo L	10450	899	29	730	6.1	c_hac_3
VW Polo 1.4 60	17140	1390	44	955	6.5	c_hac_3
Opel Corsa 1.2i Eco	14825	1195	33	895	6.8	c_hac_3
Subaru Vivio 4WD	13730	658	32	740	6.8	c_hac_3
Toyota Corolla	19490	1331	55	1010	7.1	c_hac_3
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6	c_hac_3

Exemple des voitures

Segments « classiques » des voitures : petites, moyennes, berlines/monospaces

Sur les 2 premiers axes de l'ACP

Petites voitures — Moyennes — Grosses berlines et/ou monospaces



Classification automatique

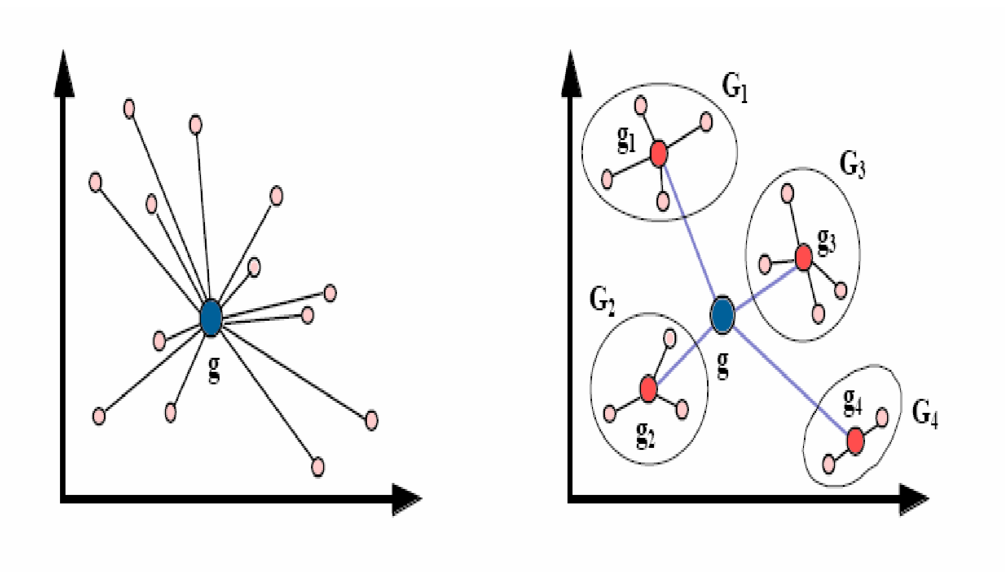
Objectifs

Principe : Constituer des **groupes** (classes, clusters) « naturels » de manière à ce que les individus dans un même groupe se ressemblent, et les individus dans des groupes différents soient dissemblables.

Autres visions :

- Identifier des groupes d'individus ayant un comportement (ou des caractéristiques) homogènes
- Proposer un résumé des données en explicitant ses principales dimensions (oppositions)
- Mettre en évidence les principales structures dans les données (définir des « concepts »)
- Construction d'une taxonomie (classification hiérarchique) d'objets (cf. taxonomie des espèces)

Illustration dans le plan



Points clés dans la constitution des groupes.

Quantifier :

- La proximité entre 2 individus
- La proximité entre 2 groupes
- La proximité entre 1 individu et un groupe (lors de la construction et l'affectation)

- Le degré de compacité d'un groupe
- L'éloignement global entre les groupes (séparabilité)

La classification ascendante hiérarchique

Une technique très populaire… pour de nombreuses raisons

CAH - Algorithme

```
Entrée : tableau de données (X)
Sortie : Indicateur de partition des individus

Calcul du tableau des distances entre individus
Chaque individu constitue un groupe (classe)
REPETER
  Détecter les 2 groupes les plus proches
  Les agréger pour n'en former qu'un seul
JUSQU'À tous les individus forment un seul groupe

Identifier le nombre adéquat de groupes
Procéder au partitionnement
```

Définir **une mesure de distance** entre individus

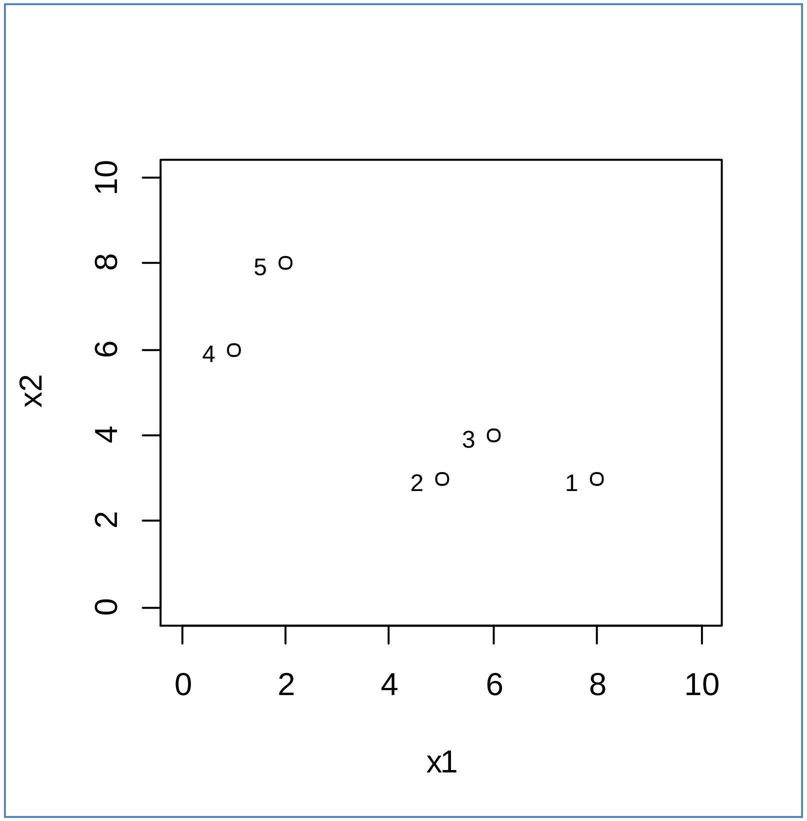
Définir **une stratégie d'agrégation** c.-à-d. une mesure de dissimilarité entre groupes (entre un individu et un groupe)

De quel outil peut-on disposer pour identifier la « bonne » partition ? Dendrogramme.

CAH – Un exemple (1)

Tableau de données

	x1	x2
1	8	3
2	5	3
3	6	4
4	1	6
5	2	8



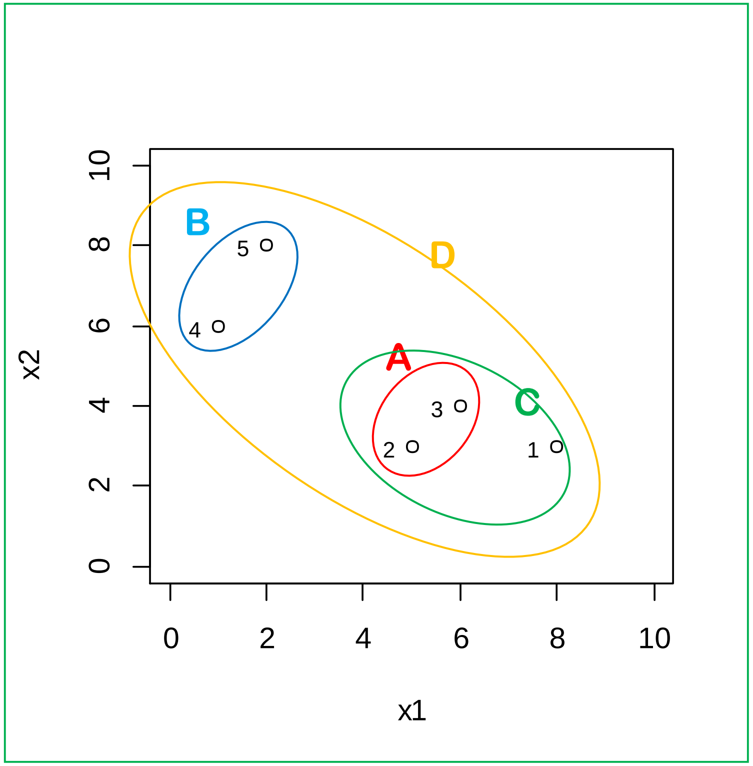
Matrice de distances entre individus

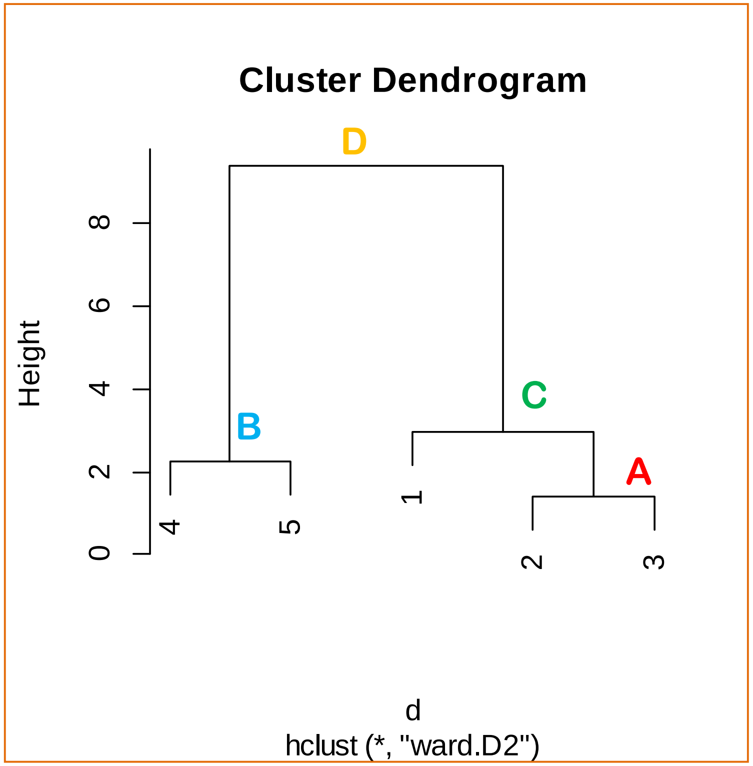
	1	2	3	4
2	3.000000			
3	2.236068	1.414214		
4	7.615773	5.000000	5.385165	
5	7.810250	5.830952	5.656854	2.236068

Distance euclidienne entre individus

$$
\begin{aligned}
d(1,3) &= \sqrt{(8-6)^2 + (3-4)^2} \\
&= \sqrt{4+1} \\
&= 2.236
\end{aligned}
$$

CAH – Un exemple (2)

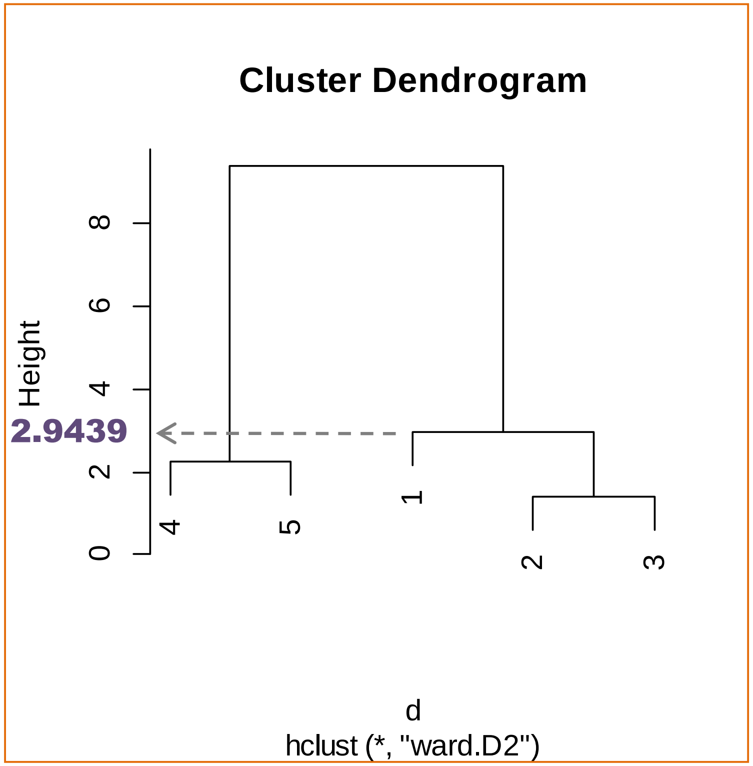




On distingue parfaitement les étapes de l'algorithme

CAH – Un exemple (3) – Niveau d'agrégation

	x1	x2
1	8	3
2	5	3
3	6	4
4	1	6
5	2	8



Distance entre (1) et (2,3)

Coordonnées du groupe (2,3) : centre de classe

$$\left(\frac{5+6}{2}=5.5, \frac{3+4}{2}=3.5\right)$$

Distance de Ward entre (1) et (2,3)

$$D^2 = \frac{n_1 \times n_{23}}{n_1 + n_{23}} \times d^2(1,23) = \frac{1 \times 2}{1+2} \times 6.5 = 4.333$$

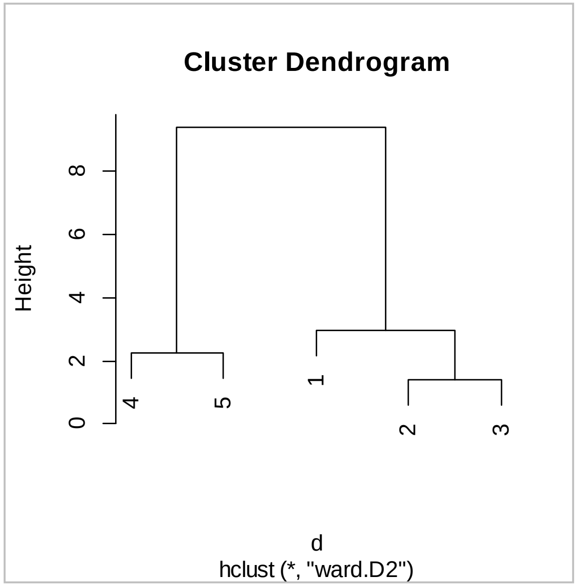
On obtient une hiérarchie indicée. Les niveaux d'agrégation correspondent en général à l'indice de dissimilarité entre les deux parties réunies.

Remarque : Curieusement, le logiciel R (3.3.1) affiche

$$\text{Height} = \sqrt{2 \times D^2} = 2.9439$$

CAH – Un exemple (4) – Détails sous R

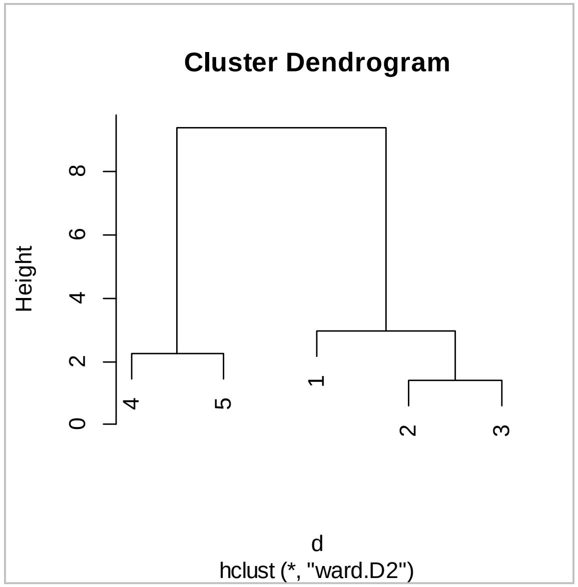
```
#vecteurs de données
x1 <- c(8,5,6,1,2)
x2 <- c(3,3,4,6,8)

#plot
plot(x1,x2,xlim=c(0,10),ylim=c(0,10))
text(x1-0.5,x2,1:5,cex=0.75)

#distance entre individus
X <- data.frame(x1,x2)
d <- dist(X)
print(d)

#CAH
cah <- hclust(d,method="ward.D2")
plot(cah)

#hauteurs d'agrégation
print(cah$height)
```




```
> #hauteurs d'agrégation
> print(cah$height)
[1] 1.414214 2.236068 2.943920 9.398581
```

CAH – Distance ultramétrique

Peut-il y avoir des inversions
dans le dendrogramme ?



A toute hiérarchie indicée H correspond une distance entre éléments de H : d(A, B), qui est le niveau d'agrégation de A et B

Elle possède une propriété supplémentaire **(propriété ultramétrique)**

$$d(A,B) \leq \max\{d(A,C), d(B,C)\}$$

Matrice de distances entre individus

	1	2	3	4
2	3.000000			
3	2.236068	1.414214		
4	7.615773	5.000000	5.385165	
5	7.810250	5.830952	5.656854	2.236068

Ex. d(2,3) ≤ max {d(2,1), d(3,1)}

CAH – Distance entre individus
(il y en a d'autres…)

Propriétés d'une distance

- Symétrie : d(a,b) = d(b,a)
- Séparation : d(a,b) = 0 ⇔ a = b
- Inégalité triangulaire : d(a,c) ≤ d(a,b) + d(b,c)

Distance euclidienne

$$d^2(a,b) = \sum_{j=1}^{p} \left(x_j(a) - x_j(b)\right)^2$$

Distance euclidienne pondérée par l'inverse de la variance

$$d^2(a,b) = \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \left(x_j(a) - x_j(b)\right)^2$$

Permet d'éliminer les problèmes de différences d'échelles entre les variables. Peut être obtenue en appliquant la distance euclidienne sur données réduites.

Distance cosinus

$$d(a,b) = 1 - \cos(a,b) = 1 - \frac{\langle a,b \rangle}{\|a\| \times \|b\|}$$

$$= 1 - \frac{\sum_{j=1}^{p} x_j(a) \times x_j(b)}{\sqrt{\sum_j x_j^2(a)} \times \sqrt{\sum_j x_j^2(b)}}$$

Populaire en text mining lorsque les vecteurs individus comportent de nombreuses valeurs nulles (parce que les textes sont de longueurs différentes).

CAH – Distance entre groupes

(il y en a d'autres…)

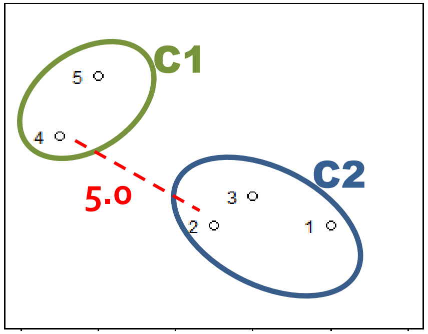
Matrice de distances

	1	2	3	4
2	3.000000			
3	2.236068	1.414214		
4	7.615773	5.000000	5.385165	
5	7.810250	5.830952	5.656854	2.236068

La distance entre deux groupes est comptabilisée à partir des deux éléments qui sont le plus proches. Attention, effet de « chaîne » des groupes.

Saut minimum (single linkage)

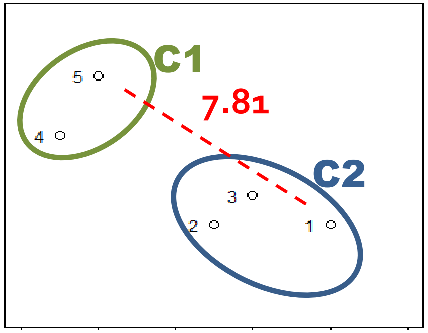
$$d(C1,C2)=\min_{a\in C1, b\in C2} d(a,b)$$



La distance entre deux groupes est comptabilisée à partir des deux éléments qui sont les plus éloignés. Groupes compacts mais problème avec points atypiques.

Saut maximum (complete linkage)

$$d(C1,C2)=\max_{a\in C1, b\in C2} d(a,b)$$

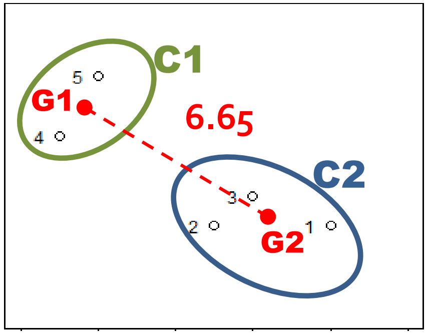


La distance entre deux groupes est comptabilisée à partir de la distance (pondérée) entre les barycentres. Privilégié souvent dans les logiciels.

Distance de Ward

$$d^2(C1,C2)=\frac{n_1\times n_2}{n_1+n_2}d^2(G1,G2)$$

➔ *Ce critère maximise l'inertie inter-classes (à voir plus loin) quand on utilise la distance euclidienne*



CAH – Distance entre groupes

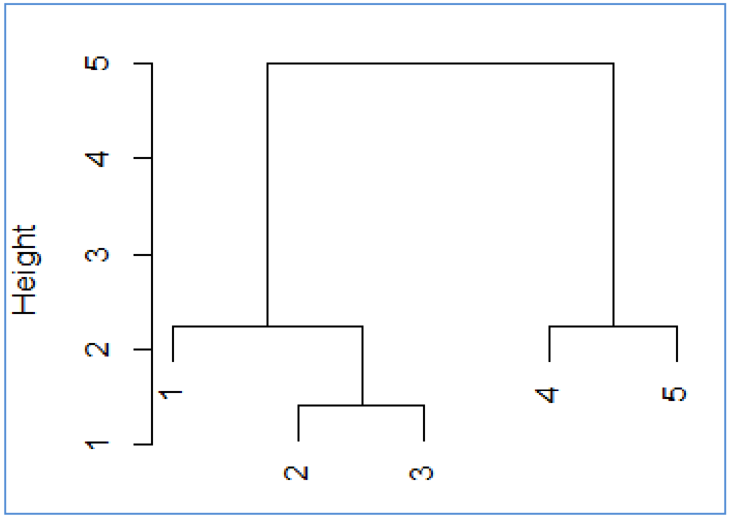
Exemple

Matrice de distances

	1	2	3	4
2	3.000000			
3	2.236068	1.414214		
4	7.615773	5.000000	5.385165	
5	7.810250	5.830952	5.656854	2.236068

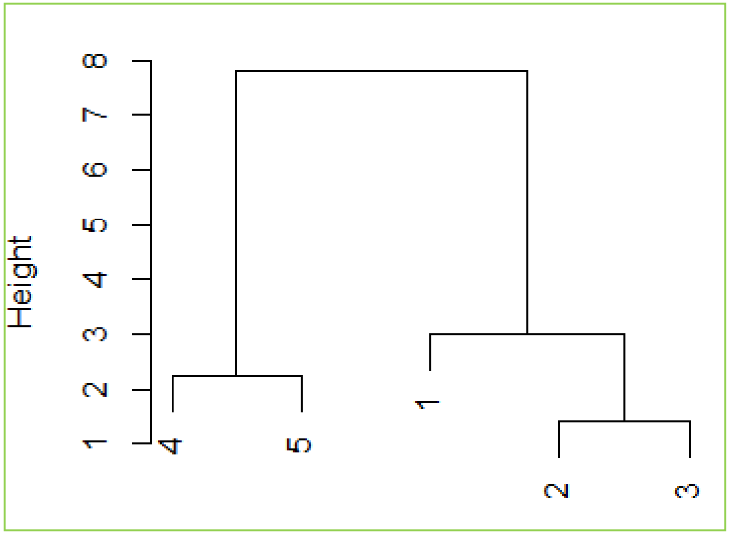
```
#CAH - single linkage
cah <- hclust(d,method="single")
plot(cah)

#hauteurs d'agrégation
print(cah$height)
```

5.000000
2.236068
2.236068
1.414214




```
#CAH - complete linkage
cah <- hclust(d,method="complete")
plot(cah)

#hauteurs d'agrégation
print(cah$height)
```

7.810250
3.000000
2.236068
1.414214



Détection du nombre de classes

La CAH fournit une hiérarchie de partitions imbriquées, et autant de scénarios de solutions

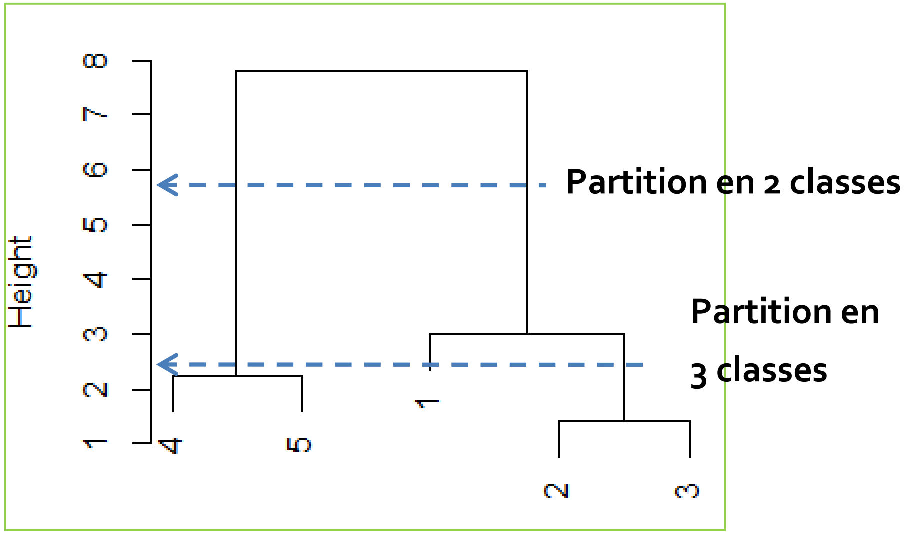
Identification du « bon » nombre de classes

Identifier le bon nombre de classes est un problème « ouvert » en classification automatique

On peut le définir comme un paramètre (à fixer) de l'algorithme (ex. K-Means)

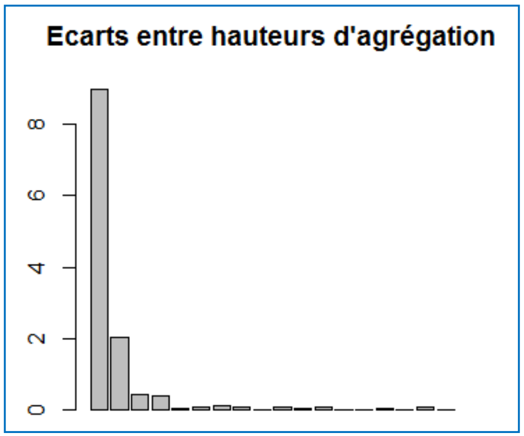
On peut aussi tester différentes solutions et utiliser des mesures insensibles au nombre de classes pour trouver la meilleure configuration (ex. indice silhouette)

La situation est différente avec la CAH. Le dendrogramme décrit un ensemble de partitions imbriquées cohérentes, qui sont autant de solutions potentielles

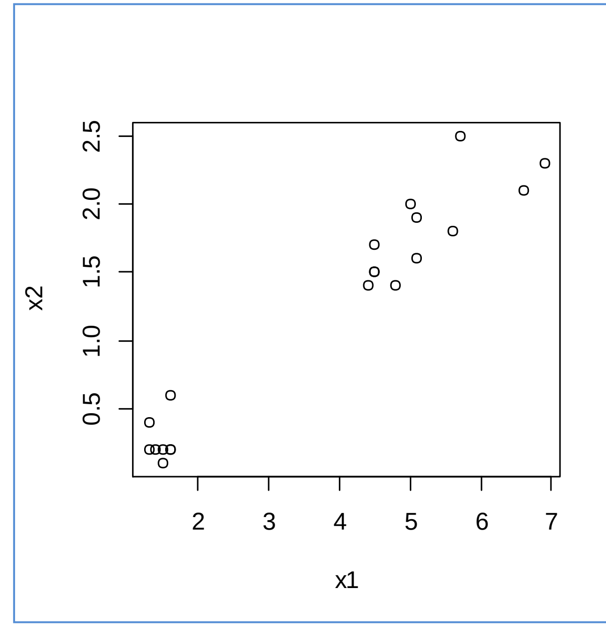


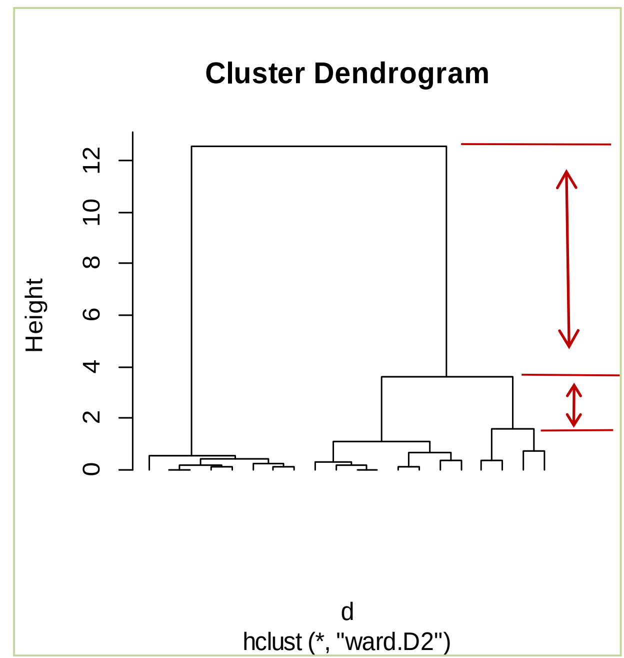
Ecarts entre paliers d’agrégations

Principe : Des fortes différences entre deux niveaux d’agrégation successifs indique une modification « significative » de la structure des données lorsqu’on a procédé au regroupement.



Une solution en 2 groupes est possible, une solution en 3 groupes est envisageable également.





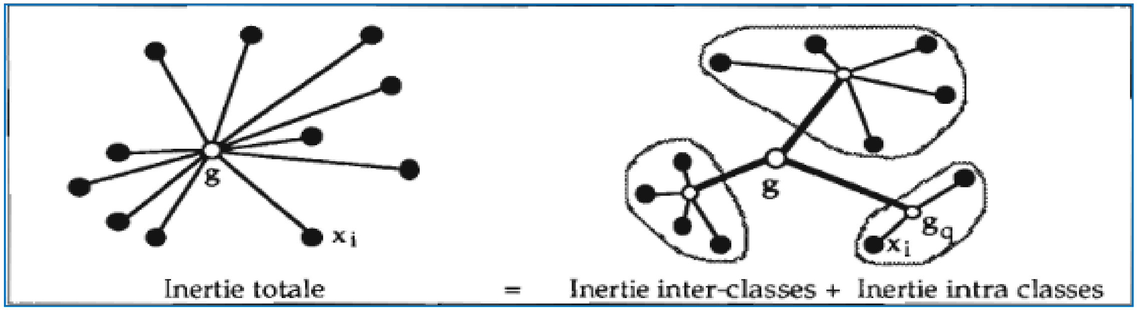
Remarque : La solution en 2 groupes apparaît toujours comme « évidente » dans le dendrogramme. Il faut savoir aller au-delà.

Inertie (1)

L'inertie est un indicateur de dispersion. Elle généralise la variance au cas multidimensionnel. G représente le barycentre global.

$$\sum_{\omega} d^2(X(\omega), G)$$

Relation de Huygens : suite à une partition des observations, décomposition de l'inertie totale en inertie inter-classes (expliquée par l'appartenance aux groupes) et inertie intra-classes (résiduelle, intrinsèque aux groupes).



$$\sum_{\omega} d^2(X(\omega), G) = \sum_{g} n_g \times d^2(G_g, G) + \sum_{g} \sum_{\omega \in g} d^2(X(\omega), G_k)$$

T. Dispersion totale.

B. Dispersion des centres de groupes autour du centre global.

W. Dispersion à l'intérieur des groupes.

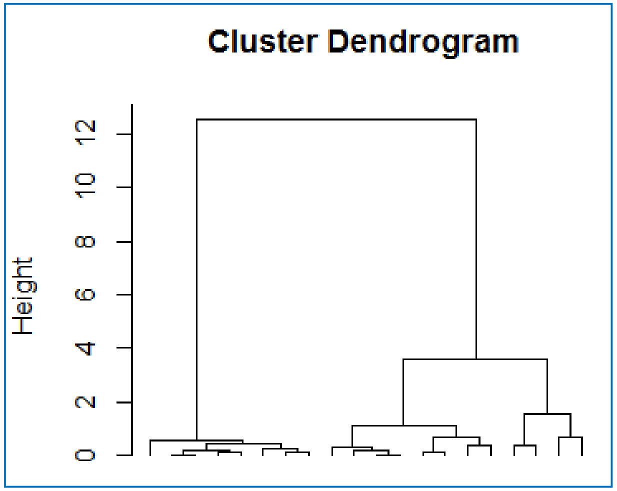
Part d'inertie expliquée par la partition :

$$R^2 = \frac{B}{T}$$

$R^2 = 0$, il y a un seul groupe.
$R^2 = 1$, partition parfaite. Souvent partition triviale : 1 individu = 1 groupe.

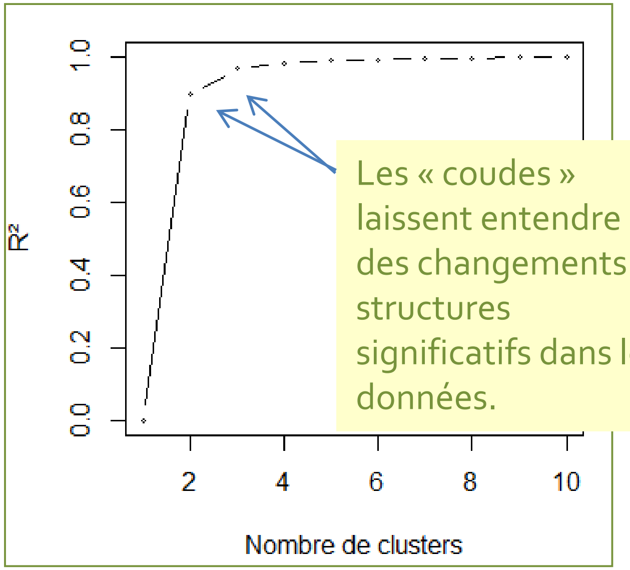
Inertie (2) – Critère de Ward

$$\Delta = \frac{n_1 \times n_2}{n_1 + n_2} d^2(G1, G2)$$



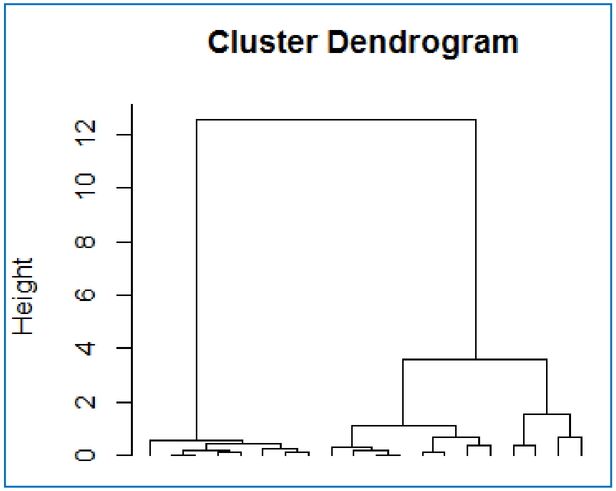
Chaque agrégation entraîne une diminution de l'inertie inter-classes. **On fusionnera donc les deux groupes entraîne la plus petite valeur de Δ**. Ils sont les plus proches au sens du critère de Ward. Leur fusion entraîne également la plus petite perte d'inertie.

On peut élaborer un graphique qui met en relation la part d'inertie expliquée (R^2) en fonction du nombre de groupes.

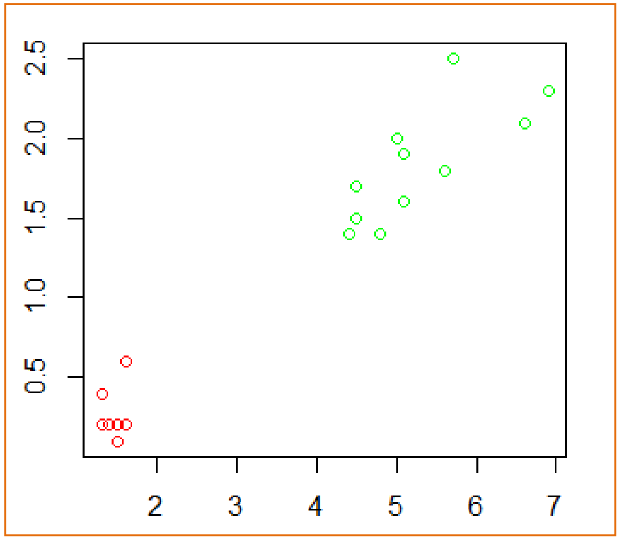


Les « coudes » laissent entendre des changements structures significatifs dans les données.

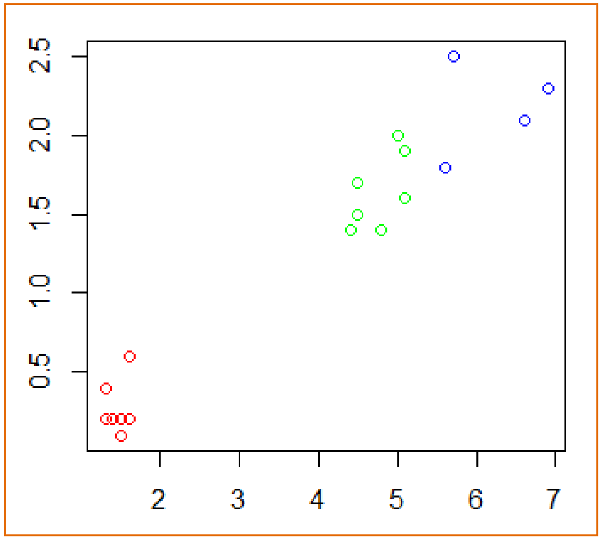
Nombre de classes – Intuition, interprétation



Au final, les techniques de visualisation et l'interprétation des groupes donnent des indications précieuses quant à l'identification du nombre de groupes. Nous avons des **scénarios** de solutions. Il faut aussi tenir compte du cahier des charges de l'étude.



Partition en deux groupes.



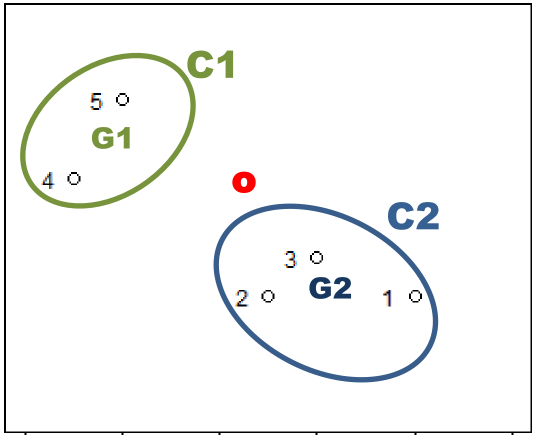
Partition en trois groupes.

Classement d'un nouvel individu

Affectation d'un individu supplémentaire à un des groupes

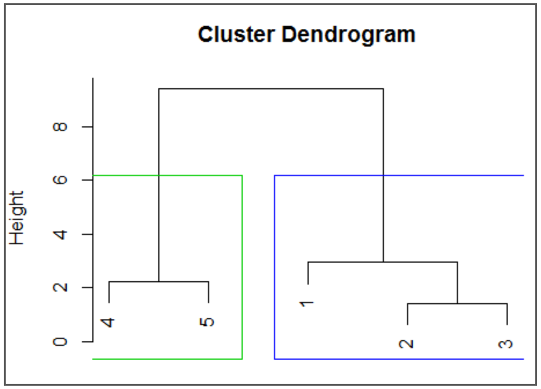
Classement d'un individu supplémentaire

La démarche doit être cohérente avec la distance et la stratégie d'agrégation utilisée.



« Single linkage » : « o » serait associé à C2 à cause du point n°3 (vs. n°5 de C1)

« Complete linkage » : « o » serait associé à C1 à cause du point n°4 (vs. n°1 de C2)



« Ward » : Il faut minimiser la quantité…

$$\Delta_o = \frac{1 \times n_c}{1 + n_c} d^2(o, G)$$

… ce qui correspond *approximativement* à une distance aux centres de classes

Logiciels

Classification des données voitures

Données

Modele	Prix	Cylindree	Puissance	Poids	Consommation
Daihatsu Cuore	11600	846	32	650	5,7
Suzuki Swift 1.0 GLS	12490	993	39	790	5,8
Fiat Panda Mambo L	10450	899	29	730	6,1
VW Polo 1.4 60	17140	1390	44	955	6,5
Opel Corsa 1.2i Eco	14825	1195	33	895	6,8
Subaru Vivio 4WD	13730	658	32	740	6,8
Toyota Corolla	19490	1331	55	1010	7,1
Opel Astra 1.6i 16V	25000	1597	74	1080	7,4
Peugeot 306 XS 108	22350	1761	74	1100	9
Renault Safrane 2.2. V	36600	2165	101	1500	11,7
Seat Ibiza 2.0 GTI	22500	1983	85	1075	9,5
VW Golt 2.0 GTI	31580	1984	85	1155	9,5
Citroen ZX Volcane	28750	1998	89	1140	8,8
Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9,3
Fort Escort 1.4i PT	20300	1390	54	1110	8,6
Honda Civic Joker 1.4	19900	1396	66	1140	7,7
Volvo 850 2.5	39800	2435	106	1370	10,8
Ford Fiesta 1.2 Zetec	19740	1242	55	940	6,6
Hyundai Sonata 3000	38990	2972	107	1400	11,7
Lancia K 3.0 LS	50800	2958	150	1550	11,9
Mazda Hachtback V	36200	2497	122	1330	10,8
Mitsubishi Galant	31990	1998	66	1300	7,6
Opel Omega 2.5i V6	47700	2496	125	1670	11,3
Peugeot 806 2.0	36950	1998	89	1560	10,8
Nissan Primera 2.0	26950	1997	92	1240	9,2
Seat Alhambra 2.0	36400	1984	85	1635	11,6
Toyota Previa salon	50900	2438	97	1800	12,8
Volvo 960 Kombi aut	49300	2473	125	1570	12,7

28 observations

5 variables actives, toutes quantitatives

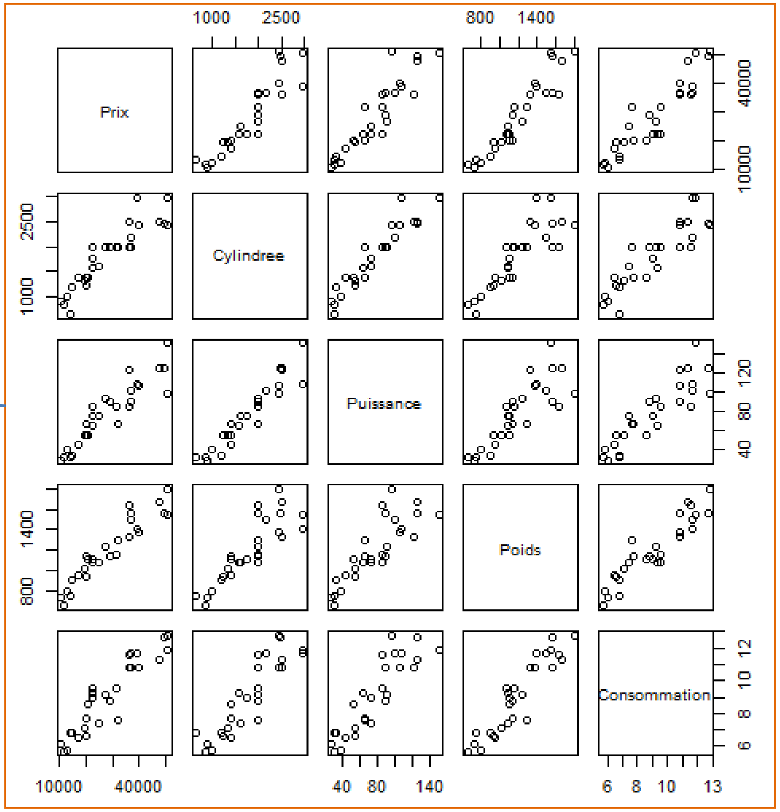
L'objectif est d'identifier des groupes naturels de véhicules, et de comprendre la nature de ces groupes (*Remarque : l'interprétation fait l'objet d'un autre support*).

Logiciel R – Chargement et préparation des données

Les variables ne sont clairement pas sur les mêmes échelles

```
     Prix           Cylindree        Puissance          Poids          Consommation
 Min.   :10450   Min.   : 658   Min.   : 29.00   Min.   : 650.0   Min.   : 5.700
 1st Qu.:19678   1st Qu.:1375   1st Qu.: 54.75   1st Qu.: 996.2   1st Qu.: 7.025
 Median :25975   Median :1984   Median : 79.50   Median :1140.0   Median : 9.100
 Mean   :28394   Mean   :1809   Mean   : 77.71   Mean   :1197.0   Mean   : 9.075
 3rd Qu.:36688   3rd Qu.:2232   3rd Qu.: 98.00   3rd Qu.:1425.0   3rd Qu.:10.925
 Max.   :50900   Max.   :2972   Max.   :150.00   Max.   :1800.0   Max.   :12.800
```

```
#charger les données
autos <- read.table("voitures_cah.txt",header=T,sep="\t",dec=".",row.names=1)

#vérification des données
print(summary(autos))

#graphiques
pairs(autos)

#centrage et surtout réduction
#pour éviter que les variables à forte variance
#« tirent » les résultats à eux
autos.cr <- scale(autos,center=T,scale=T)

#matrice des distances euclidiennes
#sur données centrées et réduites
d <- dist(autos.cr)
```

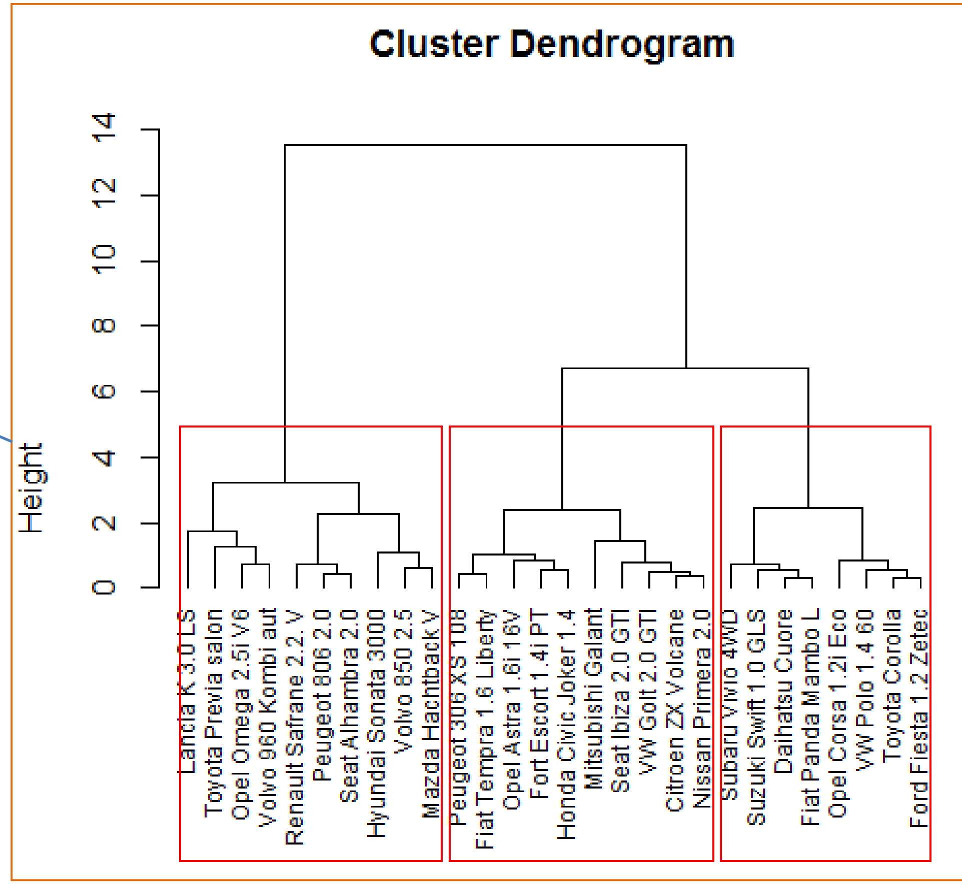


Les variables sont fortement corrélées entre elles. On distingue un peu des groupes déjà.

Logiciel R

La fonction hclust() de « stats »

```
#cah - critère de ward
cah <- hclust(d,method="ward.D2")
plot(cah,hang=-1,cex=0.75)

#mise en évidence des 3 groupes
rect.hclust(cah,k=3)

#découpage en 3 groupes
p <- cutree(cah,k=3)
print(p)
```



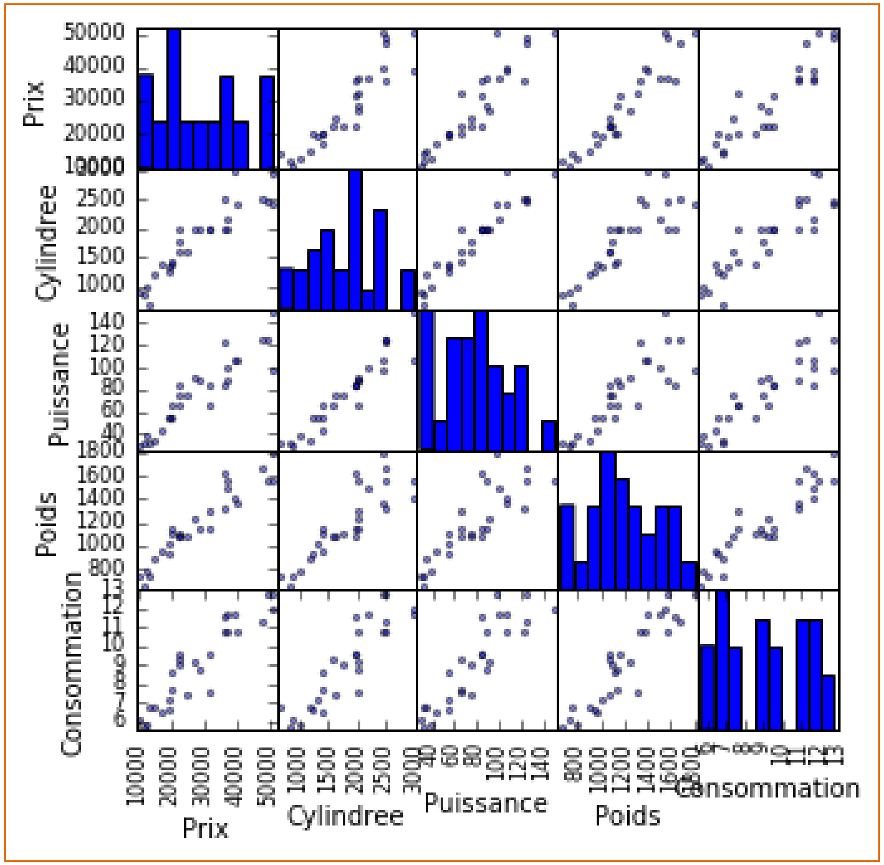

```
        Daihatsu Cuore     Suzuki Swift 1.0 GLS      Fiat Panda Mambo L
                     1                        1                       1
        VW Polo 1.4 60      Opel Corsa 1.2i Eco        Subaru Vivio 4WD
                     1                        1                       1
        Toyota Corolla      Opel Astra 1.6i 16V      Peugeot 306 XS 108
                     1                        2                       2
Renault Safrane 2.2. V        Seat Ibiza 2.0 GTI        VW Golt 2.0 GTI
                     3                        2                       2
    Citroen ZX Volcane Fiat Tempra 1.6 Liberty      Fort Escort 1.4i PT
                     2                        2                       2
  Honda Civic Joker 1.4           Volvo 850 2.5   Ford Fiesta 1.2 Zetec
                     2                        3                       1
   Hyundai Sonata 3000          Lancia K 3.0 LS       Mazda Hachtback V
                     3                        3                       3
     Mitsubishi Galant       Opel Omega 2.5i V6         Peugeot 806 2.0
                     2                        3                       3
    Nissan Primera 2.0        Seat Alhambra 2.0     Toyota Previa salon
                     2                        3                       3
   Volvo 960 Kombi aut
                     3
```

Un indicateur du groupe d'appartenance permet de réaliser tous les calculs en aval. Notamment, ceux utiles à l'interprétation des groupes.

Python

Manipulation des données

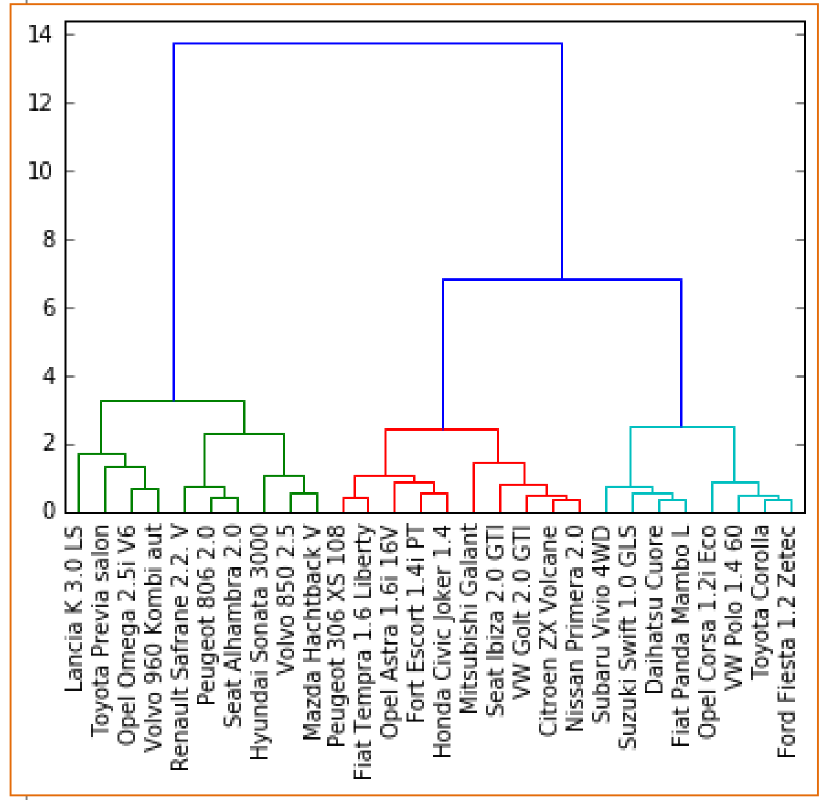
```
#modification du dossier par défaut
import os
os.chdir("...")

#chargement des données
import pandas
autos = pandas.read_table("voitures_cah.txt",sep="\t",header=0,index_col=0)

#stat. descriptives
print(autos.describe())

#graphique, croisement deux à deux
from pandas.tools.plotting import scatter_matrix
scatter_matrix(autos,figsize=(5,5))

#centrage-réduction des variables
from sklearn import preprocessing
autos_cr = preprocessing.scale(autos)
```



Même chose que sous R, on dispose en sus de la distribution unidimensionnelle de chaque variable.

Python - Package SciPy

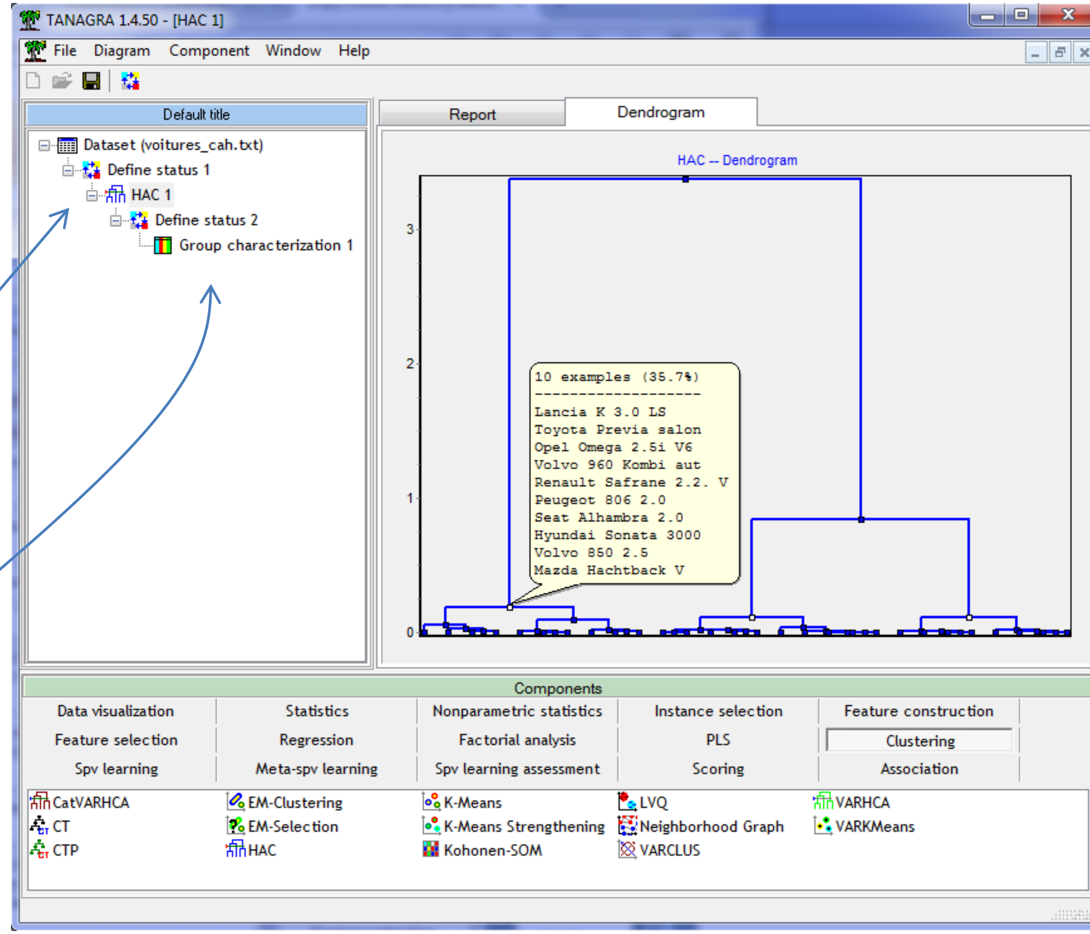
```
#librairie pour la CAH
import matplotlib.pyplot as plt
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster

#construction de la typologie
Z = linkage(autos_cr,method='ward',metric='euclidean')

#affichage du dendrogramme
plt.title("CAH")
dendrogram(Z,labels=autos.index,orientation='top',color_threshold=0,leaf_rotation=90)
plt.show()

#et matérialisation des 3 classes (hauteur de coupure height = 5)
plt.title('CAH avec matérialisation des 3 classes')
dendrogram(Z,labels=autos.index,orientation='top',color_threshold=5,leaf_rotation=90)
plt.show()

#découpage à la hauteur = 5 ==> 3 identifiants de groupes obtenus
groupes_cah = fcluster(Z,t=5,criterion='distance')
print(groupes_cah)

#index triés des groupes
import numpy as np
idg = np.argsort(groupes_cah)

#affichage des observations et leurs groupes
print(pandas.DataFrame(autos.index[idg],groupes_cah[idg]))
```



L'algorithme est déterministe, nous avons exactement les mêmes résultats que sous R.

Tanagra

L'outil HAC peut réduire automatiquement ou non les variables ; le nombre de groupes peut être détecté (basé sur les écarts de niveaux d'agrégation, en ignorant la solution à 2 classes) ; seule méthode de Ward est disponible ; possibilité de classement des individus supplémentaires.



L'outil de caractérisation des classes guide l'interprétation.

Cluster_HAC_1=c_hac_1			
Examples			[35.7 %] 10
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Consommation	4.40	11.61 (0.73)	9.08 (2.23)
Prix	4.37	42364.00 (6452.11)	28393.75 (12386.57)
Poids	4.28	1538.50 (144.95)	1196.96 (308.99)
Puissance	3.96	110.70 (19.80)	77.71 (32.26)
Cylindree	3.93	2441.60 (339.57)	1809.07 (623.66)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_2			
Examples			[35.7 %] 10
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Cylindree	-0.25	1768.40 (257.56)	1809.07 (623.66)
Puissance	-0.33	75.00 (12.43)	77.71 (32.26)
Poids	-0.69	1142.00 (74.32)	1196.96 (308.99)
Consommation	-0.72	8.66 (0.81)	9.08 (2.23)
Prix	-1.00	25192.00 (4432.02)	28393.75 (12386.57)
Discrete attributes : [Recall] Accuracy			

Cluster_HAC_1=c_hac_3			
Examples			[28.6 %] 8
Att - Desc	Test value	Group	Overral
Continuous attributes : Mean (StdDev)			
Prix	-3.57	14933.13 (3533.06)	28393.75 (12386.57)
Poids	-3.81	838.75 (128.64)	1196.96 (308.99)
Puissance	-3.86	39.88 (10.45)	77.71 (32.26)
Cylindree	-3.90	1069.25 (259.34)	1809.07 (623.66)
Consommation	-3.90	6.43 (0.51)	9.08 (2.23)
Discrete attributes : [Recall] Accuracy			

Tandem Analysis

Classification Ascendante Hiérarchique sur Composantes Principales

Tandem analysis – Principe et intérêt

Principe

Réaliser une analyse factorielle sur les données (ex. ACP)

Lancer la CAH à partir des premiers axes factoriels « pertinents »

Il ne faut plus réduire les axes dans ce cas : variance axe = son importance

Quel intérêt ?

1. La distance euclidienne considère implicitement que les variables ne sont pas liées, ce qui est faux. En utilisant les axes qui sont par définition orthogonaux deux à deux, la distance euclidienne devient parfaitement adaptée

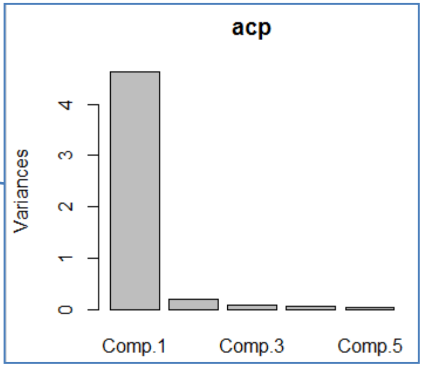
$$d^2(a,b) = \sum_{j=1}^{p} \left(x_j(a) - x_j(b)\right)^2$$

2. On procède à un nettoyage des données en ne considérant que les premiers axes porteurs d'information, et en laissant de côté les derniers axes correspondant à des fluctuations d'échantillonnage (du bruit)
3. L'approche permet d' appliquer la CAH même lorsque les variables actives ne sont pas toutes quantitatives (ACM si toutes qualitatives, AFDM si mélange quanti-quali)

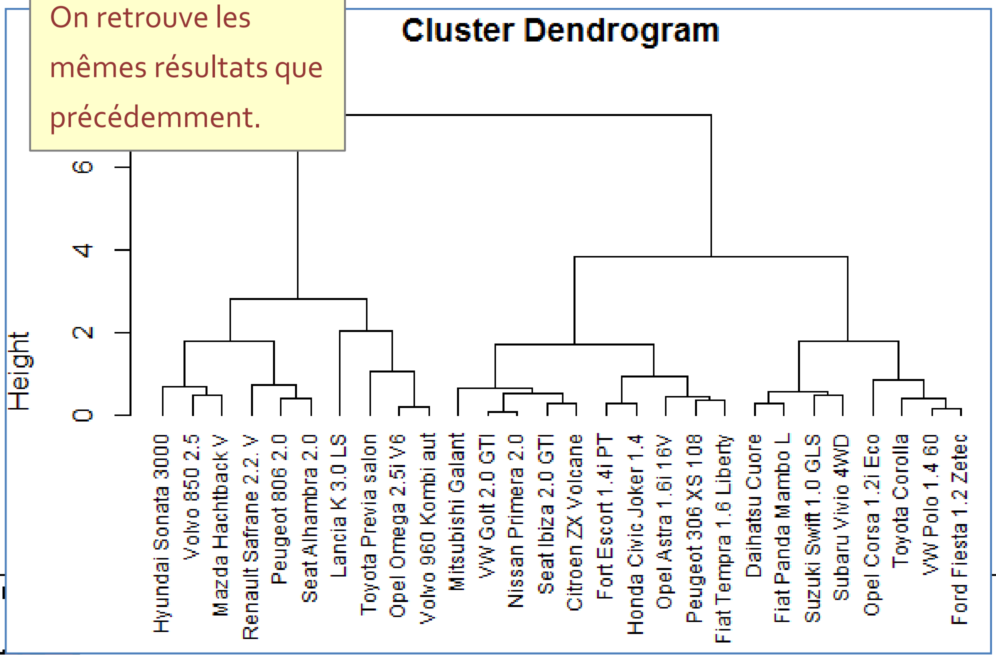
Tandem analysis – Un exemple

```
#acp normée
acp <- princomp(autos,cor=T,scores=T)
screeplot(acp)
#distance sur les 2 premiers axes
dacp <- dist(acp$scores[,1:2])
#cah
cah.acp <- hclust(dacp)
plot(cah.acp,hang=-1,cex=0.75)
#découpage en 3 groupes
p.acp <- cutree(cah.acp,k=3)
#matérialisation dans le plan factoriel
plot(acp$scores[,1],acp$scores[,2],type="n",xlim=c(-4.5,4.5),ylim=c(-4.5,4.5),xlab="92.56 %",ylab="4.10 %")
text(acp$scores[,1],acp$scores[,2],labels=rownames(autos),col=c("red","green","blue")[p.acp],cex=0.5)
```

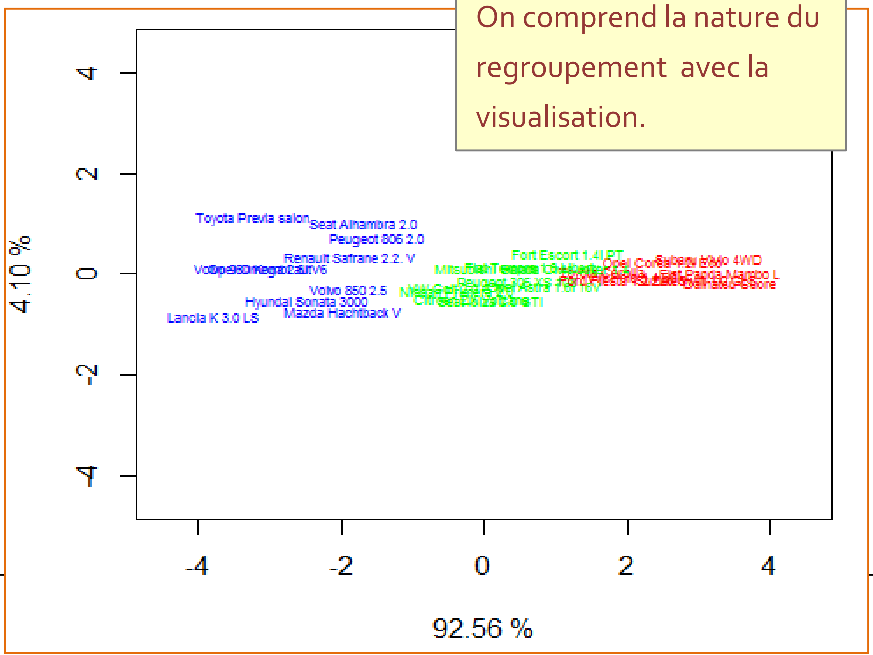
On conserve malgré tout 2 facteurs pour la visualisation.



On retrouve les mêmes résultats que précédemment.

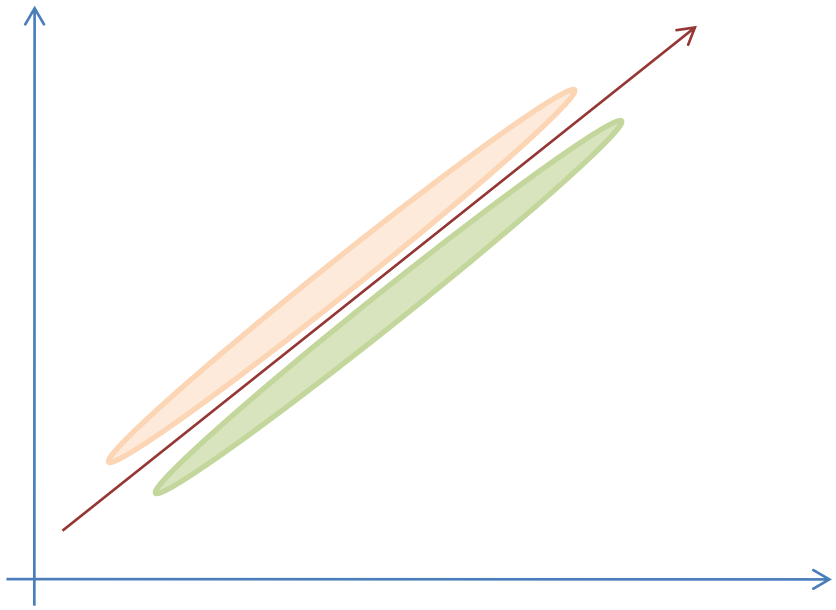


On comprend la nature du regroupement avec la visualisation.



« Tandem Analysis » n'est pas la panacée

Attention, ne retenir que les axes « significatifs » peut masquer la structuration des données en groupes.



Visuellement, les deux groupes sont évidents.

Le premier axe factoriel porte 97% de l'information, personne n'aurait l'idée de retenir le second axe.

Projetés sur le premier axe, les individus des deux groupes sont indiscernables.

➔ Faire des graphiques encore et toujours pour vérifier ce que nous propose le calcul !!!

Classification mixte

Traitement des grands ensembles de données

Classification mixte - Principe

Problème

La CAH nécessite le calcul des distances entre individus pris deux à deux. Il nécessite également l'accès à cette matrice à chaque agrégation. Infaisable sur des grands ensembles de données (en nombre d'observations).

Démarche

Réaliser un pré-regroupement (ex. en 50 classes) à l'aide de méthodes adaptées (ex. k-means, cartes de Kohonen), démarrer la CAH à partir des pre-clusters.

Intérêt

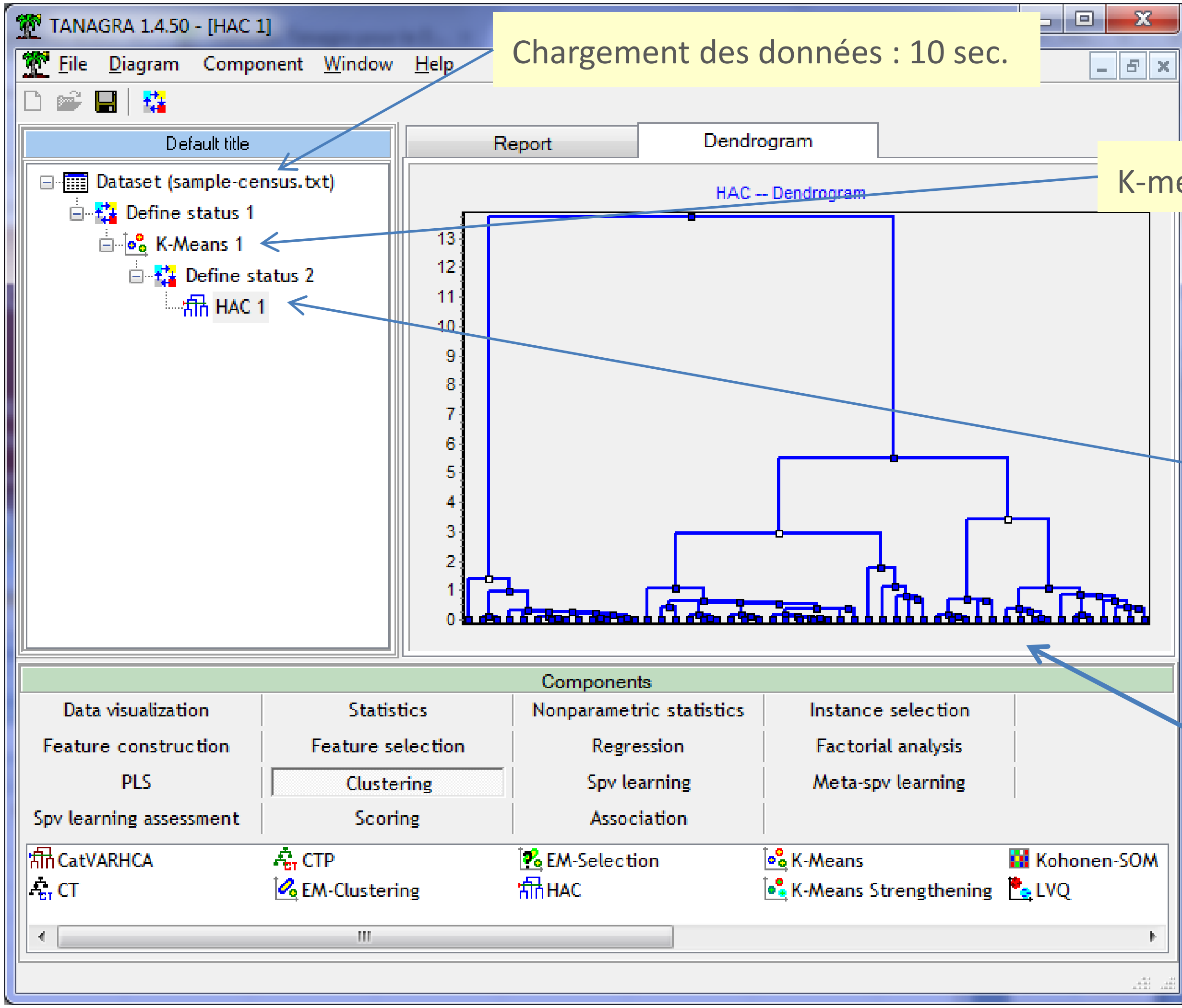
Pouvoir traiter des très grandes bases, tout en bénéficiant des avantages de la CAH (hiérarchie de partitions imbriquées, dendrogramme pour la compréhension et l'identification des classes).

Classification mixte – Un exemple

Core 2 Duo 9400 – 2.53 Ghz – Windows 7 – 4 Go RAM

500.000 observations, 68 variables

Lancer une CAH directement dessus est insensé.



Chargement des données : 10 sec.

K-means (50 groupes) : 6 mn 40 sec.

CAH à partir des 50 groupes : 5 sec.

Les solutions alternatives sont : regroupements en 2, 3 ou 5 groupes.

Voir détails dans « [Traitement de gros volumes – CAH Mixte](#) », oct. 2008.

Contient du code R pour la réalisation de la même analyse sous R.

Bilan

Principe :

• Calculer la dissimilarité entre les individus

• Agglomérations successives en fusionnant en priorité les groupes les plus proches (cf. stratégies d'agrégation : saut minimum, méthode de WARD, etc.)

• Hauteur = Distance entre groupes

Avantages

• Hiérarchie de partition (taxonomie)

• Indications sur la proximité entre groupes (choix du nombre de groupes → très difficile, il n'y a pas de solution « optimale »)

• Propose des solutions alternatives (que l'on peut interpréter ou approfondir)

Inconvénients

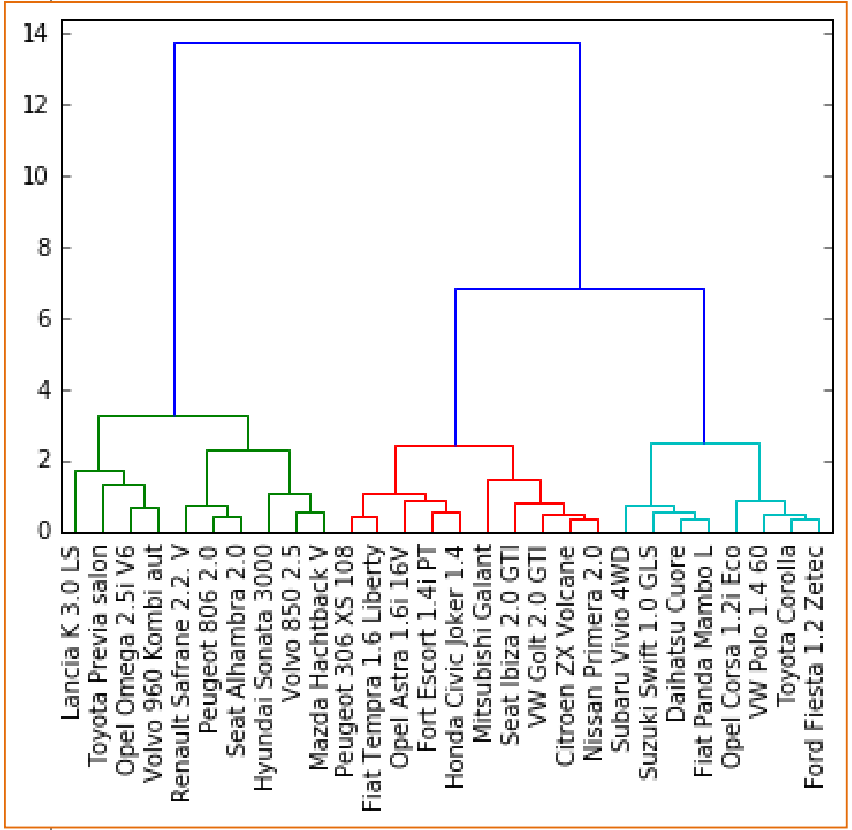
• Mise en œuvre sur des grandes bases (cf. stratégies mixtes)

Problèmes récurrents de la classification

• Détection du « bon » nombre de groupes

• Interprétation des groupes (avec ou non des variables illustratives)

• Classement d'un nouvel individu



La représentation en dendrogramme des alternatives de solutions est très séduisante.

Bibliographie

Ouvrages

Chandon J.L., Pinson S., « Analyse typologique – Théorie et applications », Masson, 1981.

Diday E., Lemaire J., Pouget J., Testu F., « Eléments d'analyse de données », Dunod, 1982.

L. Lebart, A. Morineau, M. Piron – « Statistique exploratoire multidimensionnelle », DUNOD, 2004.

Saporta G, « Probabilités, analyse des données et statistique », Technip, 2011.

Tutoriels

« Classification automatique sous R », octobre 2015.

« Classification automatique sous Python », mars 2016.

« Classification automatique sur données mixtes », novembre 2013.

« Classification automatique – Déploiement de modèles », octobre 2008.

« Traitement de gros volumes – CAH Mixte », octobre 2008.

« La complémentarité CAH et ACP », mars 2008.