
LECTURE 6. CLASSIFICATION AUTOMATIQUE

Prof. Kizungu Vumilia Roger

**UNIKIN (FACAGRO-BIOLOGIE), UNILU (FACAGRO), UEA (FACAGRO), UCB
(FACAGRO), ISS, ISTA (ENVIRONNEMENT), UPN (FACAGRO-MEDVET)**

---GII-GIII-DEA---

Release: 27 décembre 2009

Au vu des résultats de l'ACP, on peut être tenté de classer les accessions suivant les critères de distance entre les individus. En considérant ces critères, si on est rigoureux, chaque individu sera dans sa propre classe. Si l'on est tolérant, alors tous les individus sont dans une même classe. Entre les deux, selon les échelles de rigueur, on a différentes classes qui se forment.

La classification hiérarchique s'opère sur un ensemble de dissimilarités pour n objets qui doivent être classifiés. Chaque objet est assigné à sa propre classe et l'algorithme fonctionne de façon itérative à chaque étape, joignant deux objets de classes similaires. L'algorithme fonctionne jusqu'à ce que l'on ne reste qu'avec une seule classe.

Il existe plusieurs définitions des similarités.

Syntaxe dans R :

```
hclust(d, method = "complete", members=NULL)
```

avec:

d: la structure des dissimilarités telle que donnée par "dist".

method : méthode d'agglomération des objets. Les méthodes utilisées sont :

"ward", "single", "complete", "average", "mcquitty", "median", "centroid"

members: 'NULL' ou un vecteur de dimension d.

Les chapitres sur les analyses factorielles (Analyse en Composantes Principales, Analyse Factorielle des Correspondances, Analyse Factorielle des Correspondances Multiples) permettent de grouper les individus voisins dans des classes. Ce n'est pas leur objectif premier. Ce n'est qu'une conséquence de la méthode. Leur objectif est la représentation des échantillons par des cartes (Tomassone, 1993). Il existe en revanche, des méthodes qui ont pour objectif principal, la représentation par des classes. C'est l'objet de ce chapitre.

L'histoire de cette analyse remonte au temps d'Aristote qui a proposé les règles pour classer le monde animal. Après c'était Buffon qui a proposé la même chose pour le monde végétal et Mendeleiev pour les éléments chimiques. Le monde auquel la méthode peut être appliquée s'est accru. C'est le cas d'un chercheur qui dispose d'un millier d'accessions sur une culture donnée dans une collection (individus) où l'on a mesuré un certains nombres des caractéristiques (variables). Face à une maladie qui se déclare dans la région, il peut souhaiter regrouper les accessions suivant certains critères

sur leurs caractéristiques. Il se posera la question de savoir quel sont les individus qui se ressemblent et qui sont voisins par rapport au critère. Symétriquement, quels sont les individus qui sont dissemblables ?

Une façon de procéder serait la manière séquentielle. On considère en premier lieu une variable qui permet une discrimination en deux groupes. C'est par exemple, pour les variétés de manioc, le goût. Deux classes peuvent se dessiner : manioc tendre et manioc amer (fig. 1.). A ce niveau, on poursuit la classification en manioc à peau rouge et peau brune. La classification va ainsi en s'affinant jusqu'à ce qu'on atteigne une variété particulière.

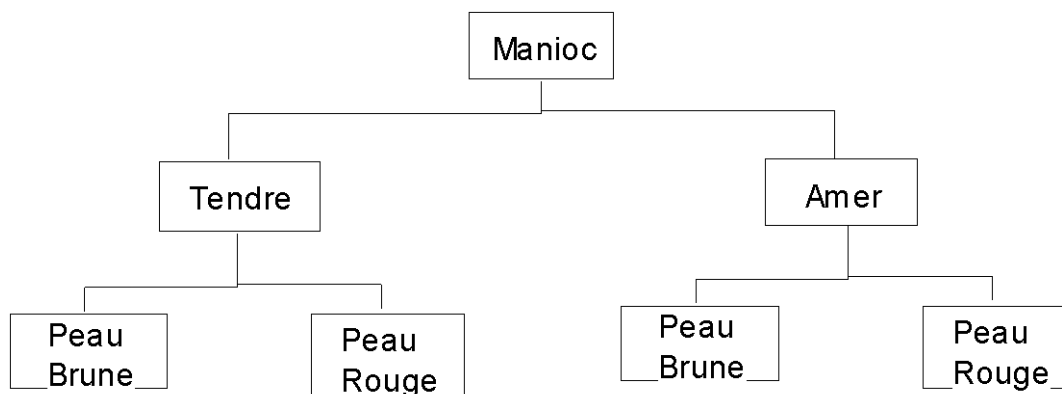


Fig.58 Classification séquentielle

Ce type de classification est dit classification monothétique (Tomassone, 1993). Une autre façon de procéder, celle que nous allons développer dans ce chapitre consiste à former des classes voisine en considérant simultanément un groupe des variables. C'est la classification polythétique.

La méthode de classification automatique vise donc la répartition des objets en classes homogènes disjoints. Les objets à classer sont généralement des individus. Ils peuvent aussi être des variables. Dans le cas d'un tableau individus x variables quantitatives, on y définit une distance qui permet d'évaluer la dissemblance entre deux vecteurs individus. La méthode peut être carrément utilisée sur un tableau des distances entre ces individus. Pour les tableaux à variables qualitatives, les méthodes sont particulières. Le principe est bien simple : on cherche une partition des individus qui optimise un critère qui tend à regrouper deux individus qui sont semblables et à séparer ceux qui sont différents.

Les principales structures de classification

Les problèmes de classification automatique diffèrent selon le type d'information recherché: Selon le cas, il s'agira d'une hiérarchie ou d'une partition. Le souci qui nous anime quand nous effectuons la classification est que les objets qui sont dans une même classe soient plus proches (cohésion interne des classes) que ceux qui sont dans des classes différentes (isolement entre les classes) (Fig.1). On peut ainsi observer un isolement sans cohésion. I (Fig. xxx), un recouvrement (Fig. xxx) ou une hiérarchie (Fig. xxx).

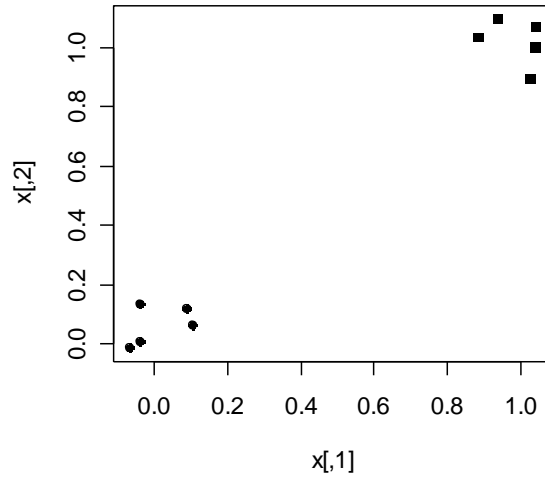


Fig. 59. Cohésion et isolement

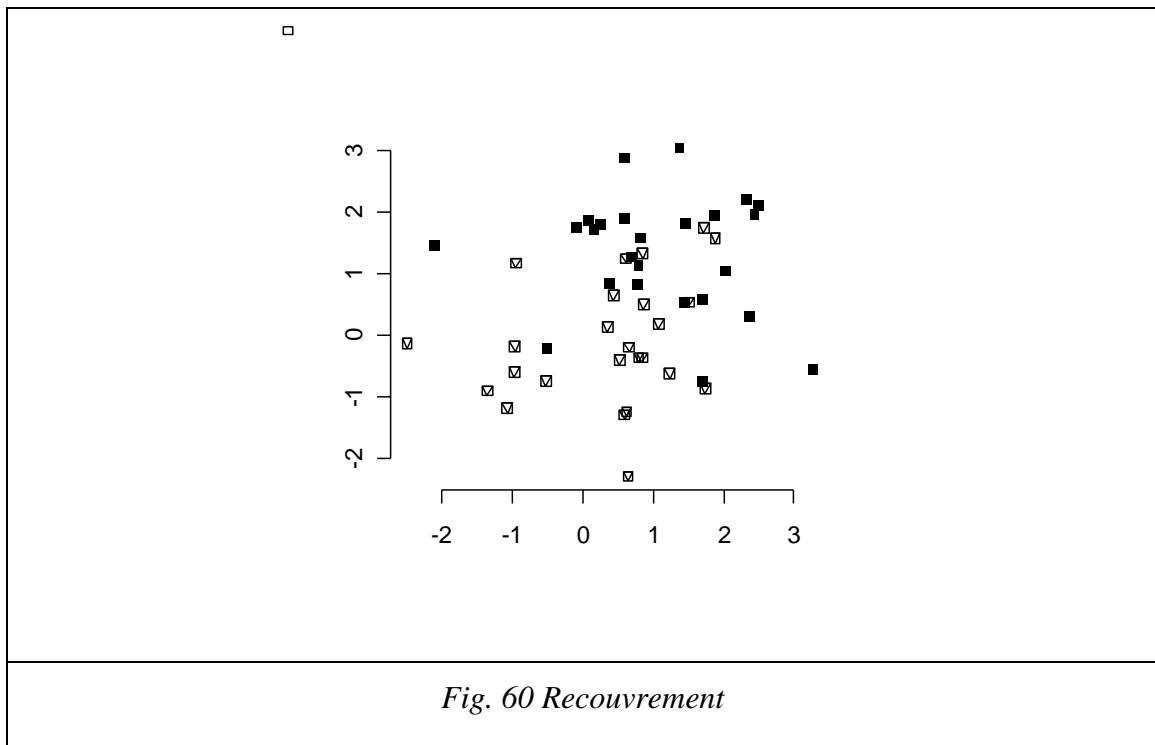


Fig. 60 Recouvrement

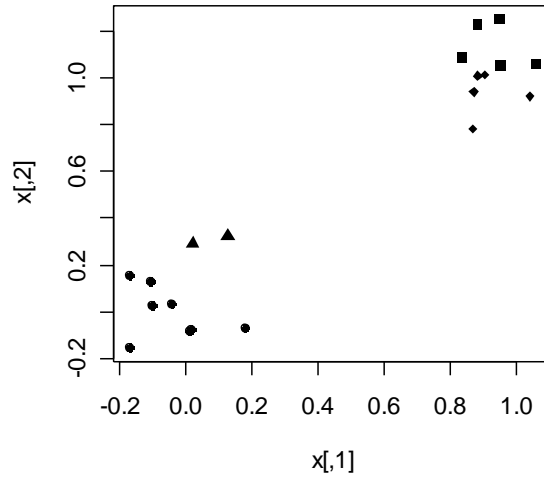


Fig. 61. Hiérarchie

Les partitions

Une partition de l'ensemble des observations Ω est un ensemble de parties non vides $P = (P_1, \dots, P_k)$ d'intersection vides deux à deux et dont la réunion forme Ω avec :

- 1) $\forall j \in \{1, 2, \dots, k\} P_j \neq \Phi$
- 2) $\forall i, j \in \{1, 2, \dots, k\} i \neq j; P_i \cap P_j = \Phi$
- 3) $\bigcup_{j=1}^k P_j = \Omega$

Ainsi avec les sept points de la **figure xxx**, on peut, par exemple, construire une partition en trois classes:

$P = (P_1, P_2, P_3)$ représentée par $P_1 = \{w_7\}$, $P_2 = \{w_5, w_4, w_6\}$ et $P_3 = \{w_1, w_2, w_3\}$.

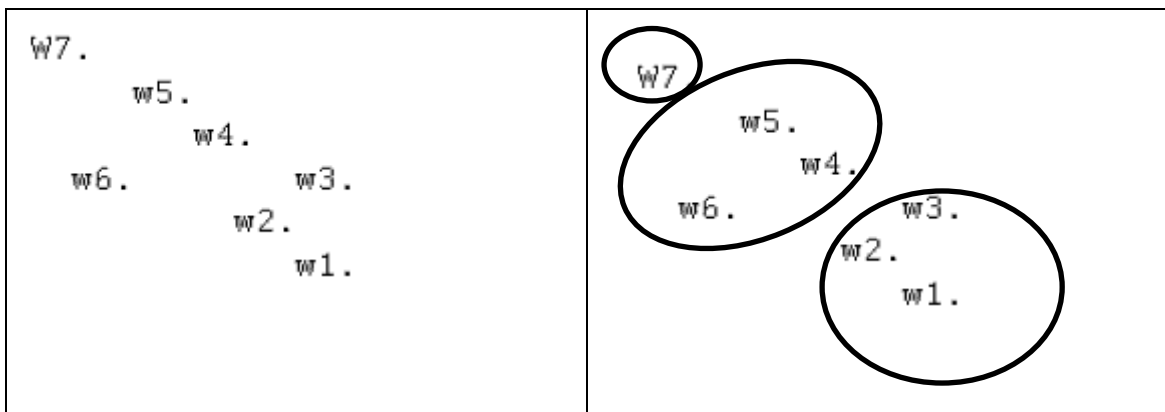


Fig. xxx

Fig. 62

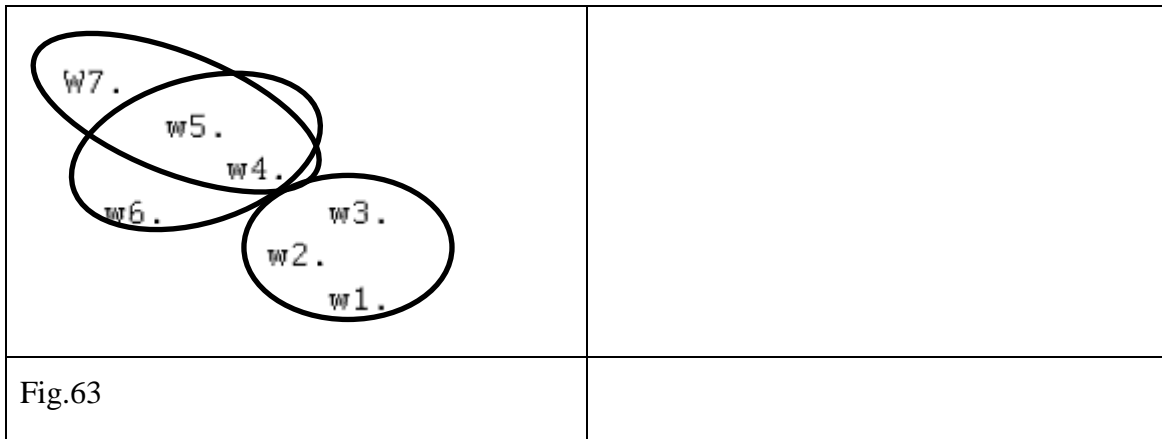
Les recouvrements

Un recouvrement de Ω est un ensemble de parties non vides $P = (P_1, \dots, P_k)$ dont la réunion forme Ω

$$1) \forall j \in \{1, 2, \dots, k\} P_j \neq \Phi$$

$$2) \bigcup_{j=1}^k P_j = \Omega$$

Avec les sept points précédents, on peut aussi construire un recouvrement à trois classes $P = (P_1, P_2, P_3)$: $P_1 = \{w_7, w_5, w_4\}$; $P_2 = \{w_5, w_4, w_6\}$; et $P_3 = \{w_1, w_2, w_3\}$ représenté par la **figure xxx**



Une partition est donc un cas particulier de recouvrement:

Les Hiérarchies

On cherche à représenter Ω par un ensemble de partitions emboîtées. Soit Ω un ensemble fini, H un ensemble de parties (appelées paliers) non vides de Ω . H est une hiérarchie sur Ω si :

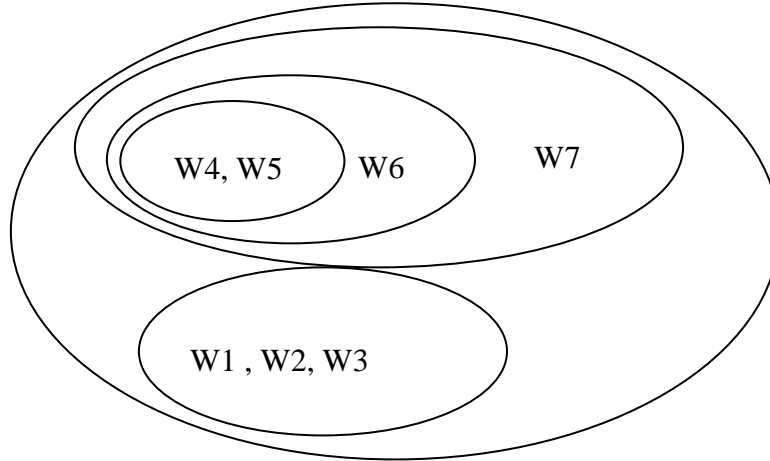
1) $\Omega \in H$ (le palier le plus haut contient tous les individus)

2) $\forall w \in \Omega, \{w\} \in H$ (les points terminaux)

3) $\forall h, h' \in H$ on a $h \cap h' \neq \Phi \Rightarrow h \subset h'$ ou $h' \subset h$

Nous utilisons encore l'ensemble Ω formé des sept points précédents; une hiérarchie associée H associée peut être:

$$H = \left\{ \begin{array}{l} h1 = \{w1\}, h2 = \{w2\}, h3 = \{w3\}, h4 = \{w4\}, h5 = \{w5\}, h6 = \{w6\}, h7 = \{w7\}, \\ h8 = \{w4, w5\}, h9 = \{w1, w2, w3\}, h10 = \{w6, w4, w5\}, \\ h11 = \{w7, w6, w4, w5\}, h12 = \{w7, w6, w4, w5, w1, w2, w3\} \end{array} \right\}$$



Indice sur une hiérarchie

Soit la fonction $i : H \rightarrow \mathfrak{R}^+$ $i : H$ vérifiant

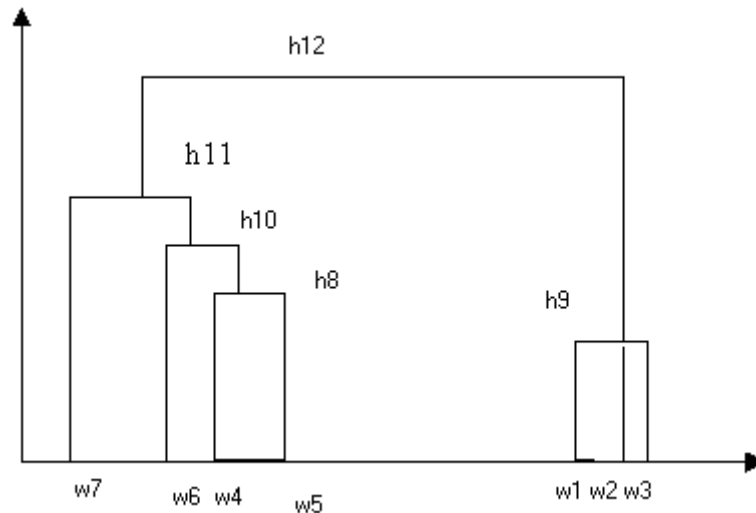
$$h \subset h' \text{ et } h \neq h' \Rightarrow i(h) < i(h')$$

$$\forall x \in \Omega \quad i(\{x\}) = 0$$

(H, i) est une hiérarchie indicée. L'indice donne la hauteur du nœud.

La représentation de l'arbre hiérarchique est dite dendrogramme

Chaque niveau d'une hiérarchie indicée est une partition : une hiérarchie correspond à un ensemble de partitions emboîtées.



Caractère combinatoire de la classification

La question que l'on peut se poser est la suivante : quelle est la meilleure classification de n objets en k groupes ?

Le nombre de toutes les partitions d'un ensemble est le nombre de Bell.

$$S(n, k) = \frac{1}{k!} \sum_{i=0}^k (-1)^{i-1} C_i^k i^n$$

Ce nombre devient vite très grand avec n. Ce qui rend nécessaire les algorithmes pour définir les groupes. Pour avoir une idée, voici comment évolue S (Tableauxxx).

	1	2	3	4	5	6	7	8
1	1							
2	1	1						
3	1	3	1					
4	1	7	6	1				
5	1	15	25	10	1			
6	1	31	90	65	45	1		
7	1	63	301	350	140	1	1	
8	1	127	966	1710	1050	266	8	1

Objectifs de la classification

Rappelons que l'objectif est la constitution des classes homogènes. La notion d'homogénéité est liée à celle de la proximité (similarité ou dissimilarité). D'où la notion de mesure de ressemblance.

Pour mesurer le voisinage ou la ressemblance, on doit définir une distance sur un ensemble Ω de n individus. On la définit comme une application de $\Omega \times \Omega$ dans \mathbb{R}^+ telle que :

1. $\forall i, j \ d(i, j) \geq 0$ et $d(i, j) = 0 \Leftrightarrow i = j$
2. $d(i, j) = d(j, i)$
3. $d(i, j) \leq d(i, k) + d(k, j)$ (inégalité triangulaire)

La mesure ainsi définie est dite mesure euclidienne si elle est engendrée par le produit scalaire suivant :

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2 = \|x_i - x_j\|^2 = (x_i - x_j)'(x_i - x_j)$$

Si l'inégalité triangulaire n'est pas définie, alors on a une *dissimilarité*.

On parle d'une similarité ou d'une ressemblance quand on a une application s telle que :

1. $\forall i, j \ s(i, j) \geq 0$
2. $s(i, j) = s(j, i)$
3. $s(i, i) = s(j, j) \geq s(i, j) \ \forall i, j$

L'information est d'autant moins riche que i et j sont plus ressemblants que k et l .

Si r_{ij} est le rang de la ressemblance entre les objets i et j parmi les $n(n-1)/2$ possibles alors :

$$\Pr\{\chi_p^2 < \frac{1}{2}d^2(i, i)\} \cong \frac{2r_{ij} - 1}{n(n-1)}$$

Quand les variables sont quantitatives, on utilise la distance euclidienne calculée sur les variables réduites (Voir ACP).

Quand les variables sont qualitatives, on utilise la métrique du χ^2 (Voir AFC)

Quand les données sont qualitative de nature binaire, il existe un grand nombre de mesure de similarité fonction de 4 nombres suivants :

- a : nombre des caractéristiques communes,
- b : nombre des caractéristiques possédées par i et non pas par j
- c : nombre des caractéristiques possédées par j et non pas par i

- d : nombre des caractéristiques possédées ni par i , ni par j .

On transpose souvent par des coefficients comme :

Jaccard :	$\frac{a}{a+b+c}$	Si on choisit ce coefficient dans le cas de la présence ou absence, ce coefficient fait intervenir les symptômes réellement répertoriés.
Russel et Rao :	$\frac{a}{a+b+c+d}$	Tandis que le coefficient de Russel fait intervenir même les cas non répertoriés.

Notons que ces coefficients sont compris entre 0 et 1. On passe alors d'une similarité à une dissimilarité en prenant le complément à 1.

Une fois que l'on fait un choix sur la mesure de la dissemblance entre les objets, il suffit de calculer cette mesure sur les $n(n-1)/2$ couples pour obtenir une matrice symétrique $D_{n \times n}$. Et c'est à partir de cette matrice que se fait la classification.

Classification ascendante hiérarchique

Classification par partition

Si l'on admet que les n objets à classer sont des points d'un espace euclidien, faire une classification en g classes consiste à répartir les n points en g groupes.

Inertie inter-classe et inertie intra-classe

L'inertie d'un nuage de n points est la moyenne des carrés de la distance de ces points à leur centre de gravité g . Si $d(i,g)$ est la distance du point i au centre de gravité, on peut réécrire :

$$I_T = \frac{1}{n} \sum_{i=1}^n d^2(i,g)$$

Pour un ensemble des points donnés, cette inertie est constante.

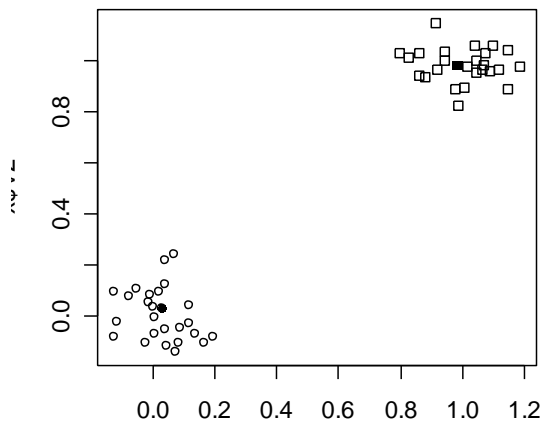
Supposons que l'on ait g classes ; la k ème classe C_k de centre de gravité g_k est formée de n_k objets ($k=1, \dots, g$).

Son inertie vaut :

$$I_k = \frac{1}{n_k} \sum_{i \in C_k} d^2(i, g_k)$$

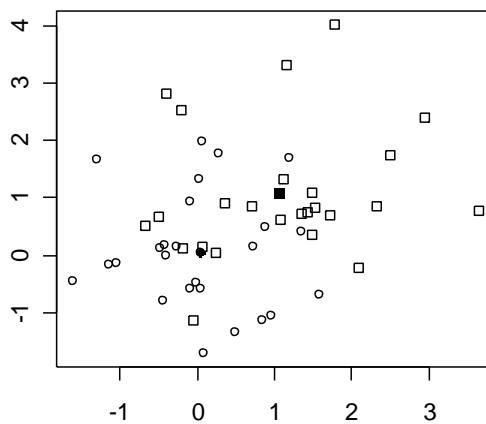
Elle est d'autant plus homogène que ses éléments sont proches de son centre figxx. Ce qui est vrai si l'inertie est faible.

Deux groupes: m=0, sd=0.1 et m=1, sd=0.



Si l'inertie est forte, alors on est dans le cas des recouvrements fig yy.

Deux groupes: m=0, sd=1.1 et m=1, sd=1.



Une mesure globale de l'homogénéité des classes appelée inertie intra-classes est donc la moyenne des inerties des g classes, chacune pondérée par son importance relative n_k/n .

$$I_w = \sum_{k=1}^g \frac{n_k}{n} I_k = \frac{1}{n} \sum_{k=1}^g \sum_{i \in C_k} d^2(i, g_k)$$

L'inertie du nuage des g centres de gravité, ou inertie inter-classes est :

$$I_B = \sum_{k=1}^g \frac{n_k}{n} d^2(g_k, g)$$

Le théorème de König Huyghes nous permet d'écrire que :

$$I_T = I_B + I_w$$

Il est naturel de choisir comme critère de classification une partition qui rende l'inertie intra-classe minimale, ou ce qui est équivalente à l'inertie inter-classe maximale. On peut prendre comme critère global de la qualité de classification :

$$C_q^2 = \frac{I_B}{I_T}$$

Si le nombre g des classes n'est pas fixé, la meilleure partition est naturellement celle qui conduit à $g=n$, puisque l'inertie intra-classe est alors nulle. Cette démarche n'est valable que pour un nombre fixé de classes.

Si la matrice des données est centrée réduite, on montre que la somme des carrés des n distances au centre de gravité g était lié à la trace de la matrice de covariance qui vaut p , le nombre de variables.

$$I_T = (n-1)p/n \text{ (le montrer)}$$

Classification non hiérarchique ou regroupement autour des centres mobiles

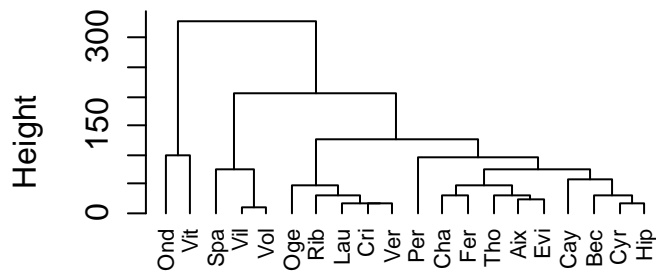
On peut choisir une partition initiale et essayer pas à pas de l'améliorer. Cette méthode s'appelle méthode des centres mobiles. La partition initiale peut être choisie au vu des résultats d'une ACP (Tomassone(1993) ou beaucoup plus simplement, en prenant au hasard g points $(c_1^0, c_2^0, \dots, c_g^0)$ qui sont des centres provisoires de g classes $C_1^0, C_2^0, \dots, C_g^0$. On affecte ensuite chaque point à la classe C_k^0 si c_k^0 est le centre le plus proche du point. On peut alors déterminer le centre de gravité de ces classes $g_1^0, g_2^0, \dots, g_g^0$.

L'étape suivante consiste à recommencer l'opération précédente d'affectation de chaque point à la classe C_k^1 si g_k^1 est le centre de gravité le plus proche du point. On obtient de nouvelles classes $C_1^1, C_2^1, \dots, C_g^1$ et donc de nouveaux centres $g_1^1, g_2^1, \dots, g_g^1$.

Le processus s'arrête quand le contenu de chaque classe n'est plus modifié. On atteint un minimum de l'inertie intra-classe.

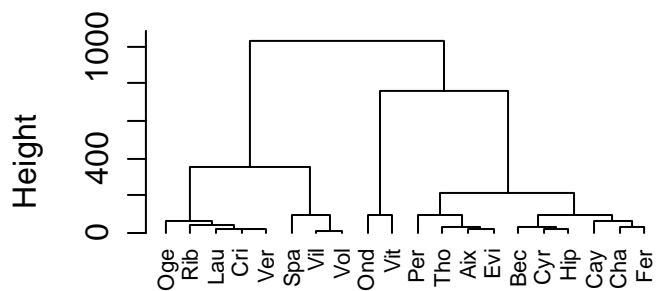
Il est conseillé de recommencer plusieurs fois à partir de plusieurs partitions initiales différentes. Si on a toujours la même chose, le résultat est alors acceptable (Tomassone (1993)).

Cluster Dendrogram



dist(ceaux)
hclust (*, "average")

Cluster Dendrogram



dist(ceaux)
hclust (*, "ward")

Cette analyse ressort 4 groupes

$$I_T = (n-1)p/n = (20-1)*6/20 = 5.7$$

Le choix du nombre des classes est une affaire de bon sens. Il est facilité par l'ACP. A priori, on voit apparaître 4 groupes. En pratique, il faut essayer avec ce nombre et l'encadrer, c'est à dire essayer avec 3 et 5 groupes.

Pour trois groupes :

```
> mobile=kmeans(ceaux,3,20)
> mobile
K-means clustering with 3 clusters of sizes 10, 2, 8
Cluster means:
  HCO3   SO4    Cl    Ca    Mg    Na
1 304.30 16.20 14.200 86.600 12.900 10.50
2 400.00 262.00 15.000 179.500 35.500  5.50
3 145.75  20.25 12.625  40.625  4.625 10.75
Clustering vector:
Aix Bec Cay Cha Cri Cyr Evi Fer Hip Lau Oge Ond Per Rib Spa Tho Ver Vil Vit Vol
```

```

 1  1  1  1  3  1  1  1  1  3  3  2  1  3  3  1  3  3  2  3
Within cluster sum of squares by cluster:
[1] 23651.10 4905.50 33654.13
Available components:
[1] "cluster" "centers" "withinss" "size" >
cl=cbind(row.names(eaux),mobile$cluster)
> cl
      [,1] [,2]
[1,] "Aix" "2"
[2,] "Bec" "2"
[3,] "Cay" "2"
[4,] "Cha" "2"
[5,] "Cri" "1"
[6,] "Cyr" "2"
[7,] "Evi" "2"
[8,] "Fer" "2"
[9,] "Hip" "2"
[10,] "Lau" "1"
[11,] "Oge" "1"
[12,] "Ond" "3"
[13,] "Per" "2"
[14,] "Rib" "1"
[15,] "Spa" "1"
[16,] "Tho" "2"
[17,] "Ver" "1"
[18,] "Vil" "1"
[19,] "Vit" "3"
[20,] "Vol" "1"

```

Nuée dynamique

C'est une variante de la méthode des moyennes mobiles. On remplace le centre de gravité par un ensemble de q points qui constituent le noyau de la classe : ils représentent mieux la classe qu'un seul centre de gravité.

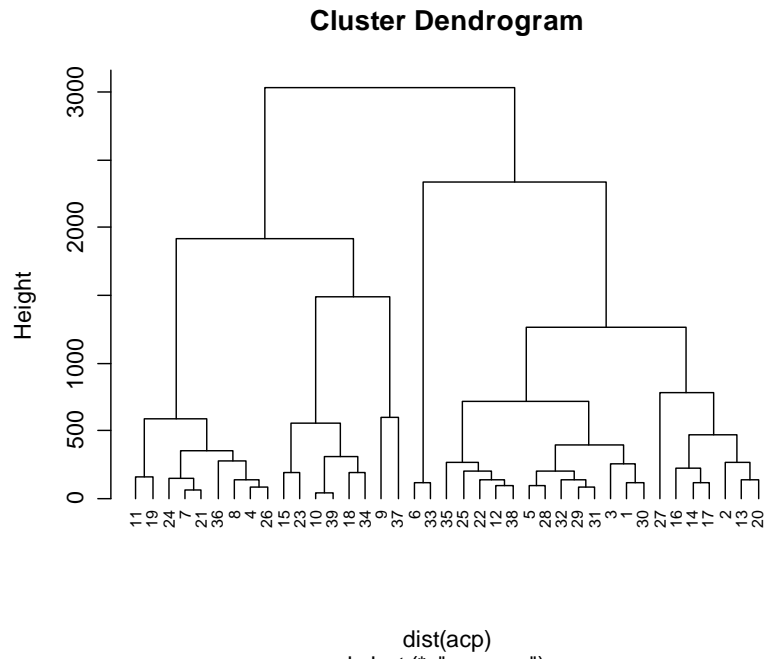
Exemple

Considérons l'exemple traité ci-haut. On veut connaître les classes qui se forment de manière hiérarchique.

```

> hc1 <- hclust(dist(acp), method = "ave")
> plot(hc1)
> plot(hc1, hang=-1)
> plot(hc1, hang=-1, cex=0.5)
> plot(hc1, hang=-1, cex=0.7)
>

```



Commenter en comparant ces résultats avec ceux de l'ACP.