

## 4. CAH – Classification Ascendante Hiérarchique

# Plan

1. Exemples, problématique
2. Complémentarité analyse factorielle / classification
3. Comment réaliser une classification ?
4. Ressemblances entre individus et classes d'individus
5. Agrégation selon l'indice de Ward
6. Le choix du nombre de classes
7. Classification de données qualitatives
8. Classification sur facteurs
9. Interprétation des classes d'une partition

## 4.1 – Problématique, exemples

**Classifier** un ensemble d'objets : répartir en classes un ensemble d'objets décrits par différentes variables ou caractéristiques

L'objectif est d'obtenir :

*Des classes **homogènes** :* les individus d'une même classe partagent de nombreuses caractéristiques (se ressemblent)

*Des classes **séparées** :* les individus de classes différentes ont peu de caractéristiques en commun

**Exemples...**

## Exemples

- **Crédit à la consommation**

Quels sont les différents types de « comportement bancaire » parmi les 66 consommateurs de l'agence ?

- **Enquête Ouest France**

Identifier les différentes « façons » pour les lecteurs de lire le quotidien Ouest France ?

- **Températures mensuelles**

Existe-t-il des villes présentant des profils de températures similaires tout au long de l'année ?

## Exemple traité

- 52 emmentals décrits par 17 descripteurs sensoriels
- Trois types de descripteurs liés au goût, à la texture et au parfum
- Une variable de conformité (binaire)

parfum

texture

goût

Emmental	intensité du parfum	parfum propionique	parfum butyrique	texture ferme	texture souple	texture granuleuse	texture collante	texture fondante	texture caractéristique	intensité du goût	goût acide	goût salé	goût sucré	goût piquant	goût fruité	goût amer	goût caractéristique	Conformité
1	5,1	4	3,7	3,8	4,8	3,7	3,3	3	3,9	5,6	4,8	4,3	3,9	4,1	3,4	3,2	3,6	non
2	4,7	3,9	3,4	5,2	3	4	3,3	4	3,6	5,8	4,3	4,9	4	4,3	4,8	3,2	3,7	oui
3	4,7	4,2	2,9	3,7	3,3	2,8	3,4	4,7	3,9	5,7	4,3	4,7	3,9	4,2	4,9	3	4	oui
4	5,3	4,6	3,9	3	3,6	2,4	3,4	5,4	4,2	5,6	4,2	4,6	4	4,2	4,9	4,1	4	oui
5	4,3	4	2,6	3,9	4,1	3,3	2,9	4,2	3,7	5,1	4	4,3	3,9	4	4,3	3,8	4	oui
6	4,7	4	3,5	4,5	4,6	3,3	3,5	4,2	4,4	4,9	3,6	4,2	3,2	3,5	4	3,2	3,9	oui
7	3,6	3,7	2,6	4,1	4,4	3,1	3,4	4,5	4,1	4,5	3,4	3,5	3,2	3	4,2	2,6	4,3	oui
8	5,4	3,9	4	4,1	3,9	2,5	3,9	4,9	4	5	4,2	4,2	3,1	4,1	3,8	2,9	3,6	oui
9	4	3,8	2,6	4,5	4	4,1	3,3	3,3	4,1	4,5	3,1	3,9	3,3	3,1	3,8	3,2	3,5	oui
10	5	4,2	3,3	4,7	4,3	3,8	3,7	3,5	3,8	4,8	3,6	4	3	3,3	3,3	3,1	3,7	oui
⋮																		
52	3,7	4,1	3	3,7	4,2	4,4	2,7	2,6	3,2	3,8	2,9	3,1	2,9	3	3,4	3,5	3,2	non

## Le tableau de données étudié en classification

- Un tableau « classique » *individus* x *variables*
- En général : des variables de même nature (quantitatives ou qualitatives)

*variables*

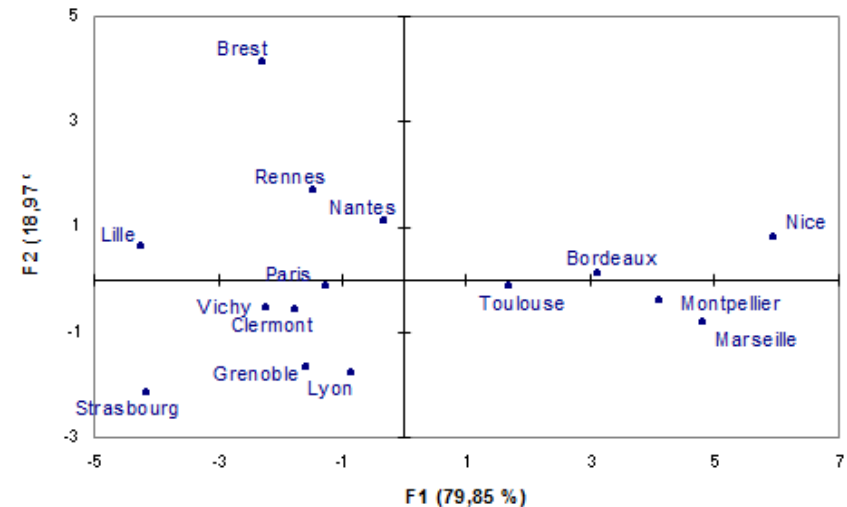
	1	...	$j$	...	$p$
<i>individus</i>	1				
	:				
	$i$		$x_{ij}$		
	:				
	$n$				

$x_{ij}$  : valeur prise par l'individu ( $i$ )  
pour la variable ( $j$ )

## 4.2 – Complémentarité analyse factorielle - classification

Ce que permettent d'obtenir les **analyses factorielles**

- Mise en évidence des principaux **facteurs de variabilités** entre individus
- Des **proximités, regroupements** ou **oppositions** entre individus dans un espace géométrique



## Limites des représentations factorielles

- Souvent : difficultés d'établir des regroupements d'individus (pas de bien nette, **lisibilité**, graphique surchargé)
- Les projections peuvent conduire à des proximités **trompeuses**
- Peut-on se limiter aux proximités observées sur le premier plan ?  
Comment faire la **synthèse de plusieurs axes** simultanément ?
- Comment **caractériser** un regroupement d'individus par quelques **variables** importantes ?



## Intérêt des méthodes de classification

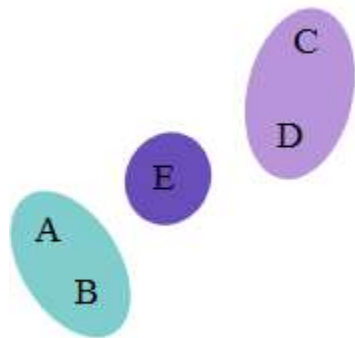
- Réalisent des regroupements d'individus en tenant compte de proximités établies sur **plusieurs dimensions**
- Fournissent une **description synthétique** des classes à partir des variables les plus importantes

## Complémentarité entre analyse factorielle et classification

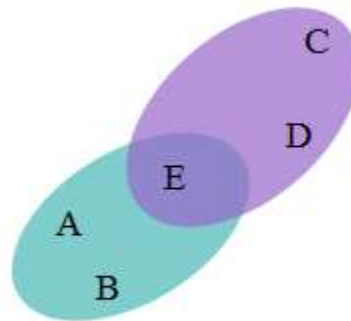
- Caractère **synthétique** de la classification
- Richesse de la représentation **géométrique** issue des méthodes factorielles

## 4.3 – Comment réaliser une classification ?

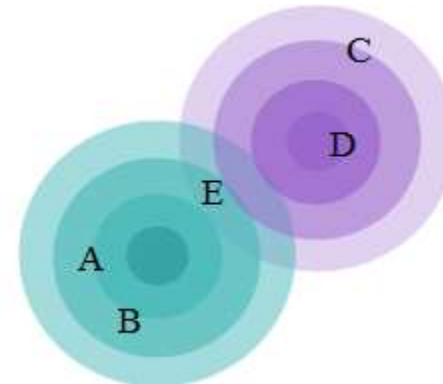
### Les différentes formes de classification



**Partition**



**Recouvrement**



**Classification floue**

- Classification ascendante hiérarchique (CAH)
- Partitionnement direct (ex. *k – means*)

## Peut-on rechercher la partition optimale ?

Peut-on construire toutes les partitions possibles des  $n$  individus en 1, 2, ...,  $k$  classes puis retenir la « meilleure » ?

### Nombres de Bell

Nombre total de partitions possibles de  $n$  objets en  $k$  classes

[d'après Saporta, 1990]

Nombre  $n$  d'individus

Nombre  $k$  de classes

	1	2	3	4	5	6	7	8	9	10	11	12
1	1	1	1	1	1	1	1	1	1	1	1	1
2		1	3	7	15	31	63	127	255	511	1023	4095
3			1	6	25	90	301	966	3025	9330	28501	86526
4				1	10	65	350	1701	7770	34105	145750	611501
5					1	15	140	1050	6951	42525	246730	1379400
6						1	21	266	2646	22827	179487	1323652
7							1	28	462	5880	63987	627393
8								1	36	750	11880	159027
9									1	45	1155	2275
10										1	55	1705
11											1	66
12												1
TOTAL	1	2	5	15	52	203	877	4140	21147	115975	678570	4213597

25



?

## Calcul du nombre total de partitions

$$B_{n+1} = \sum_{k=0}^n \binom{n}{k} B_k$$

$$B_n = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$$

## Conclusion

Il faut se contenter en général de recherche une **partition « sous-optimale »**

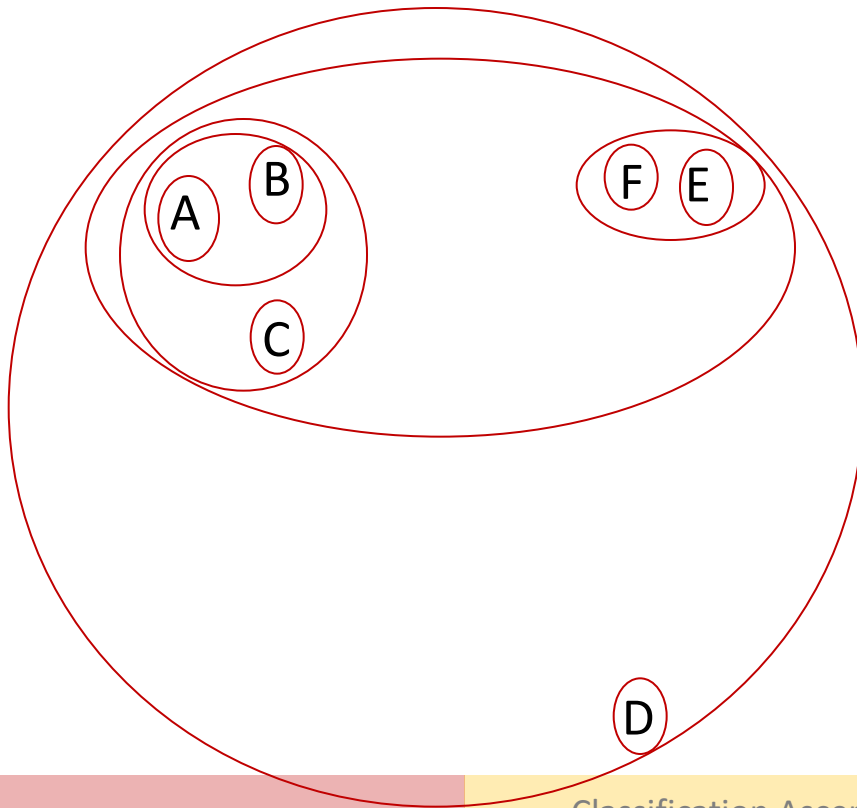
Deux principales familles de méthodes de classification

- ✓ les méthodes produisant des **arbres hiérarchiques** (ex. CAH)
- ✓ les méthodes de **partitionnement direct** (ex. nuées dynamiques, *k – means*)

## L'algorithme de la CAH

La CAH consiste à **agréger de proche en proche** des individus entre eux, puis des classes d'individus entre elles, jusqu'à obtenir une classe englobant l'ensemble de la population

**Illustration** 6 points décrits par leurs coordonnées dans le plan



**Étape 0** : chaque individu = une classe

**Étape 1** : agrégation de E et F

**Étape 2** : agrégation de A et B

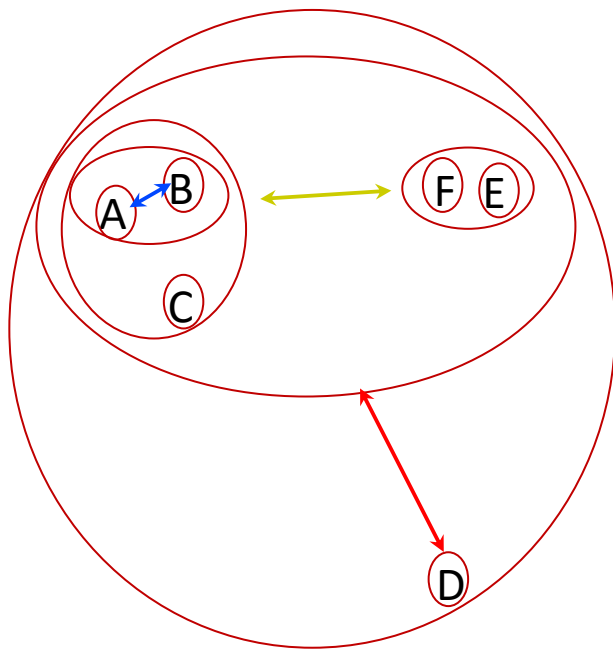
**Étape 3** : agrégation de C et {A, B}

**Étape 4** : agrégation de {A,B,C} et {E,F}

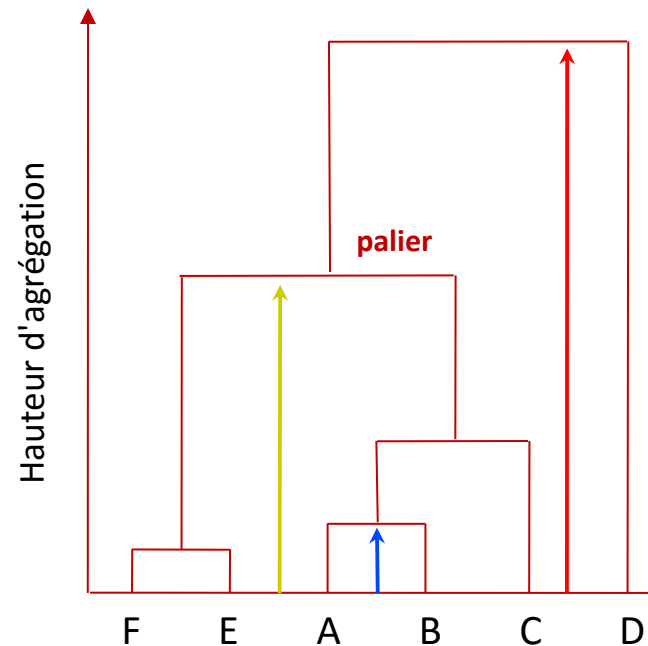
**Étape 5** : agrégation de {A,B,C,E,F} et D

## Le dendrogramme

Il représente le résultat du processus d'agrégation sous la forme d'un arbre hiérarchique (binaire) ou hiérarchie



dendrogramme ou hiérarchie



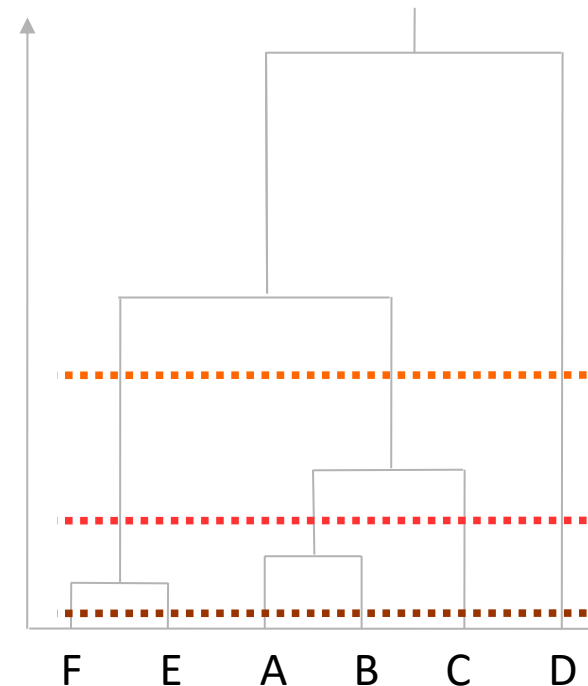
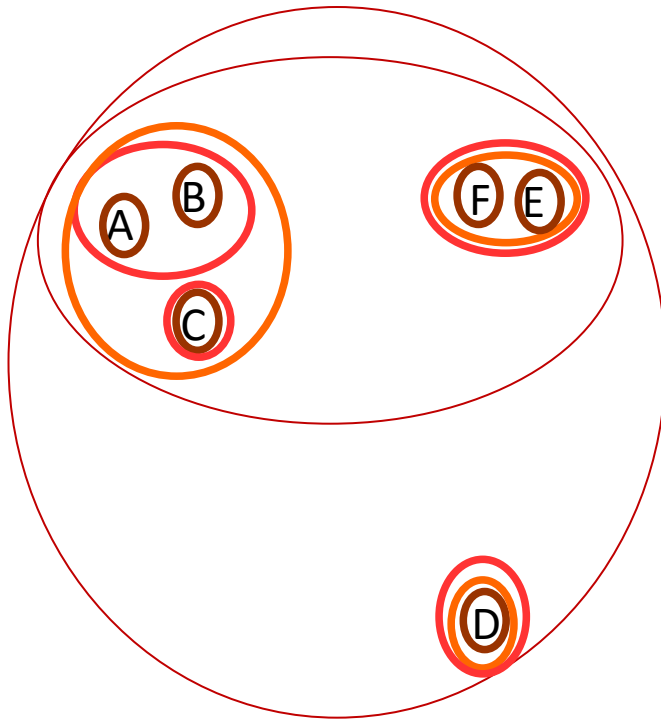
### Hauteur d'agrégation

Traduit le niveau de dissemblance entre les éléments agrégés

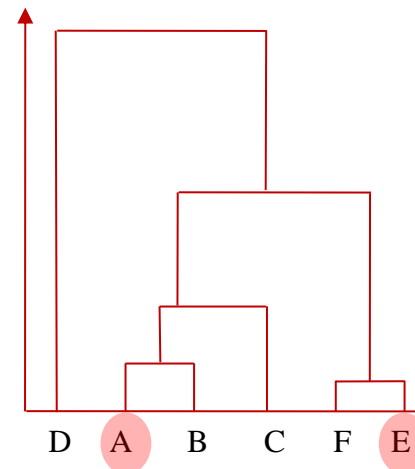
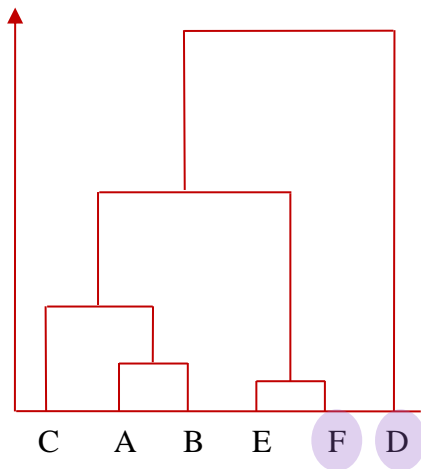
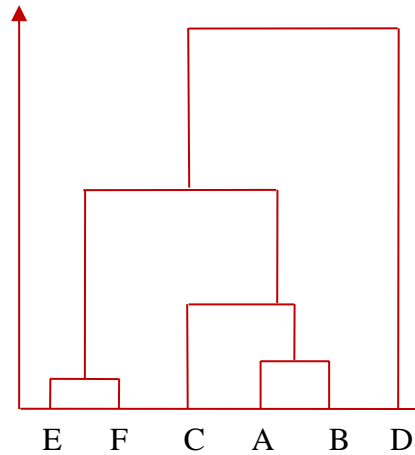
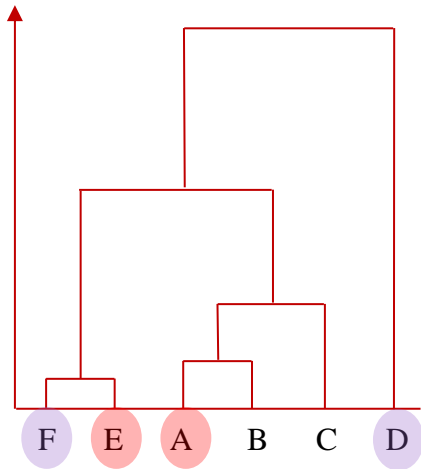
« plus le palier est haut, moins les éléments réunis se ressemblent »

## Hiérarchie et partitions emboîtées

- Chaque « **coupure** » de la hiérarchie définit un **niveau de partition** des objets
- Les partitions successives sont **emboîtées** les unes dans les autres



## Attention aux « proximités » trompeuses !





## 4.4 – Ressemblance entre individus et classes d'individus

### Classification de données quantitatives

- Un tableau de données individus x variables quantitatives
- Centrage- réduction des données si besoin (*mêmes arguments que pour l'ACP*)

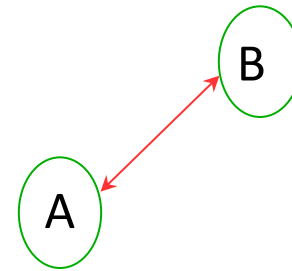
*variables*

		1	...	$j$	...	$p$
<i>individus</i>	1					
	$\vdots$					
	$i$			$x_{ij}$		
	$\vdots$					
	$n$					
	$m$			$\bar{x}_j$		
	$s$			$s_j$		

## Ressemblance entre individus

On utilise un **indice de distance**

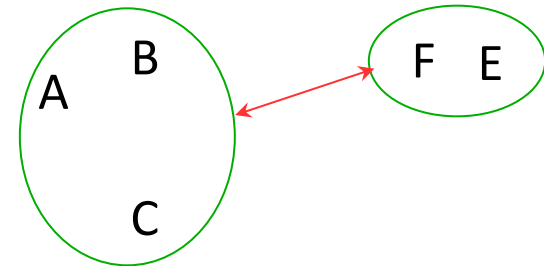
- *distance euclidienne usuelle*
- *distance du Chi2*
- *distance de Manhattan*



## Ressemblance entre classes d'individus

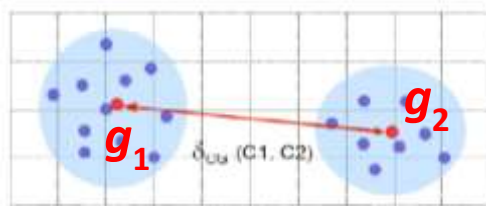
On utilise un **indice d'agrégation**

- *indice du lien minimum*
- *indice du lien maximum*
- *distance moyenne*
- *distance entre centres de gravité*
- *indice de Ward*



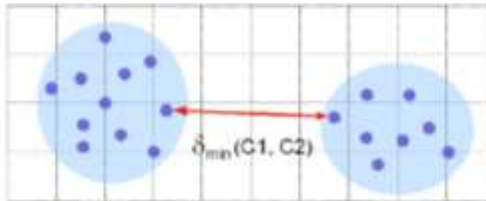
## Quelques indices d'agrégation

- Distance entre centre de gravité



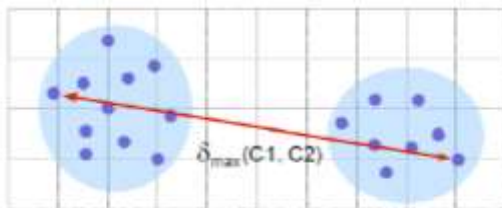
$$\delta_{CG}(C_1, C_2) = d(g_1, g_2)$$

- Indice du lien minimum (*single linkage*)



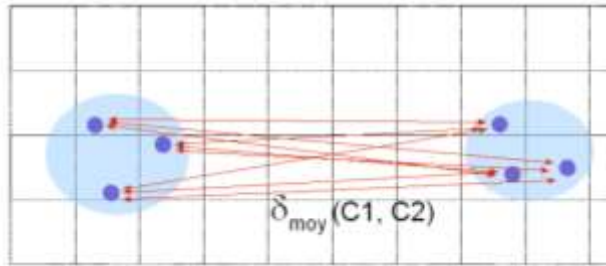
$$\delta_{\min}(C_1, C_2) = \min_{\substack{i \in C_1 \\ j \in C_2}} \{d(i, j)\}$$

- Indice du lien maximum (*complete linkage*)



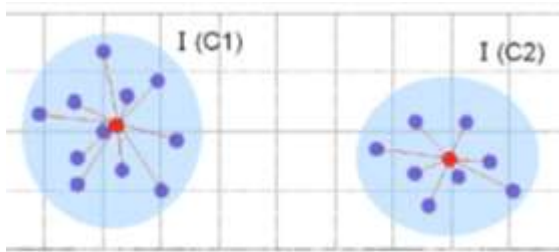
$$\delta_{\max}(C_1, C_2) = \max_{\substack{i \in C_1 \\ j \in C_2}} \{d(i, j)\}$$

- Distance moyenne

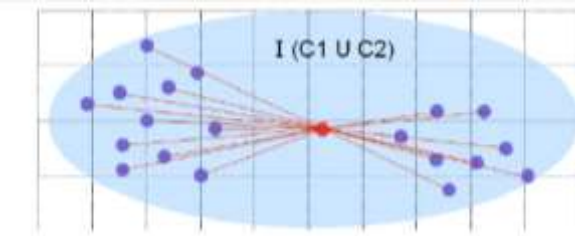


$$\delta_{\text{moy}}(C_1, C_2) = \frac{1}{\text{card}(C_1) \times \text{card}(C_2)} \sum_{i \in C_1, j \in C_2} d(i, j)$$

- Indice de Ward



Deux classes sont d'autant plus **proches** que leur agrégation conduit à une **faible augmentation d'inertie** (*intra classe*)

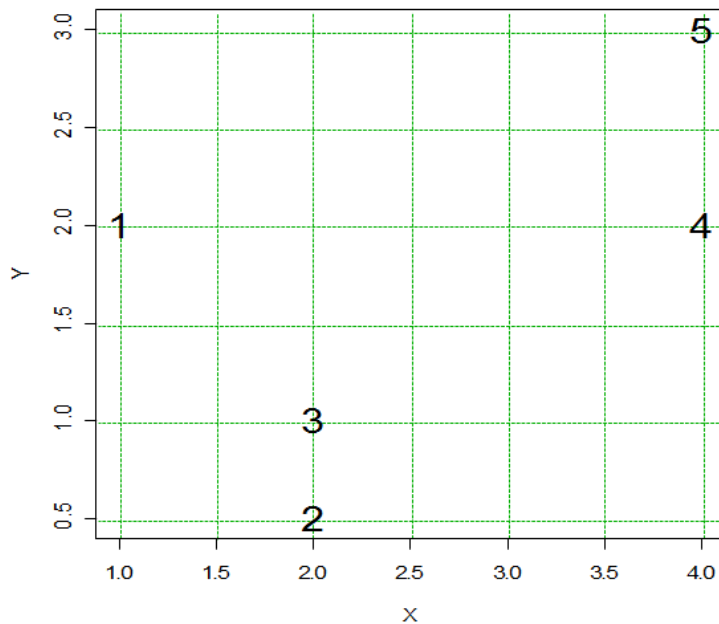


$$\delta_{\text{Ward}}(C_1, C_2) = I(C_1 \cup C_2) - [I(C_1) + I(C_2)]$$

$$\delta_{\text{Ward}}(C_1, C_2) = \frac{m_1 m_2}{m_1 + m_2} d^2(g_1, g_2)$$

## Application

Cinq individus 1, 2, 3, 4 et 5 sont décrits par leurs coordonnées  $(x, y)$  dans le plan



Réaliser la CAH de ces cinq individus sur la base des choix suivants :

- **Indice de distance** : distance du city – block
- **Indice d'agrégation** : lien minimum

## 4.5 – Agrégation selon l'indice de Ward

### Notations

- $x_i$  : description de l'individu ( $i$ )
- $g$  : centre de gravité des  $n$  individus
- $g_k$  : centre de gravité des  $n_k$  individus de la classe  $C_k$

### Qualité d'une partition en termes d'inertie

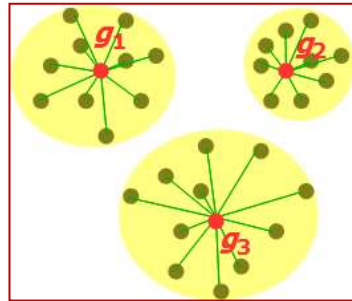
Une partition est de « bonne qualité » si

- les classes sont **homogènes** : l'inertie **intra** – classes est **faible**
- les classes sont **bien séparées** : l'inertie **inter** – classes est **élevée**

## Inerties Intra et Inter

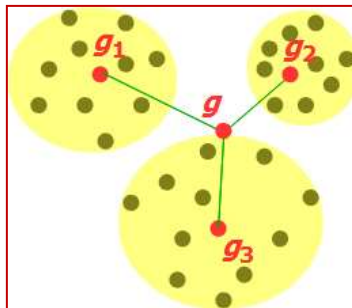
### *Illustration pour une partition en trois classes*

- Inertie **intra** (**W**ithin)



$$I_W =$$

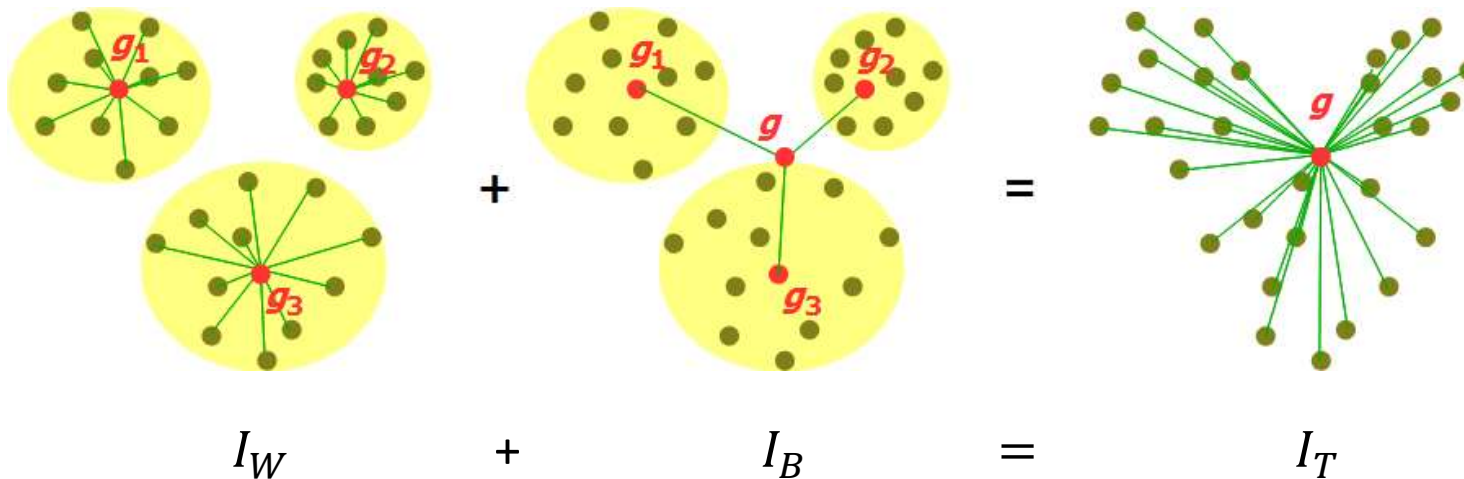
- Inertie **inter** (**B**etween)



$$I_B =$$

## Le théorème de Huygens

En présence d'une partition en  $K$  classes, l'inertie totale se décompose en la somme des inerties Inter et Intra



Où  $I_T =$



## Indicateur de qualité d'une partition

$$\frac{\text{inertie } \mathbf{inter} - \text{classes}}{\text{inertie } \mathbf{totale}}$$

## Propriété de l'indice de Ward

Minimiser l'augmentation  
d'inertie INTRA



Minimiser la perte  
d'inertie INTER

- Partition en  $K$  classes :  $I_T(K) = I_W(K) + I_B(K)$
- Partition en  $K - 1$  classes :  $I_T(K - 1) = I_W(K - 1) + I_B(K - 1)$

## Evolution des inerties et de l'indice au cours de l'algorithme

Nombre de classes	Inertie TOTALE	Inertie INTER	Inertie INTRA	Indice de Ward
$n$				
$n - 1$				
...				
2				
1				

- L'augmentation d'inertie INTRA :
- La somme des indices de Ward :
- Seconde expression de l'indice de Ward :  $\delta_{Ward}(C_1, C_2) = \frac{m_1 m_2}{m_1 + m_2} \times d^2(g_1, g_2)$

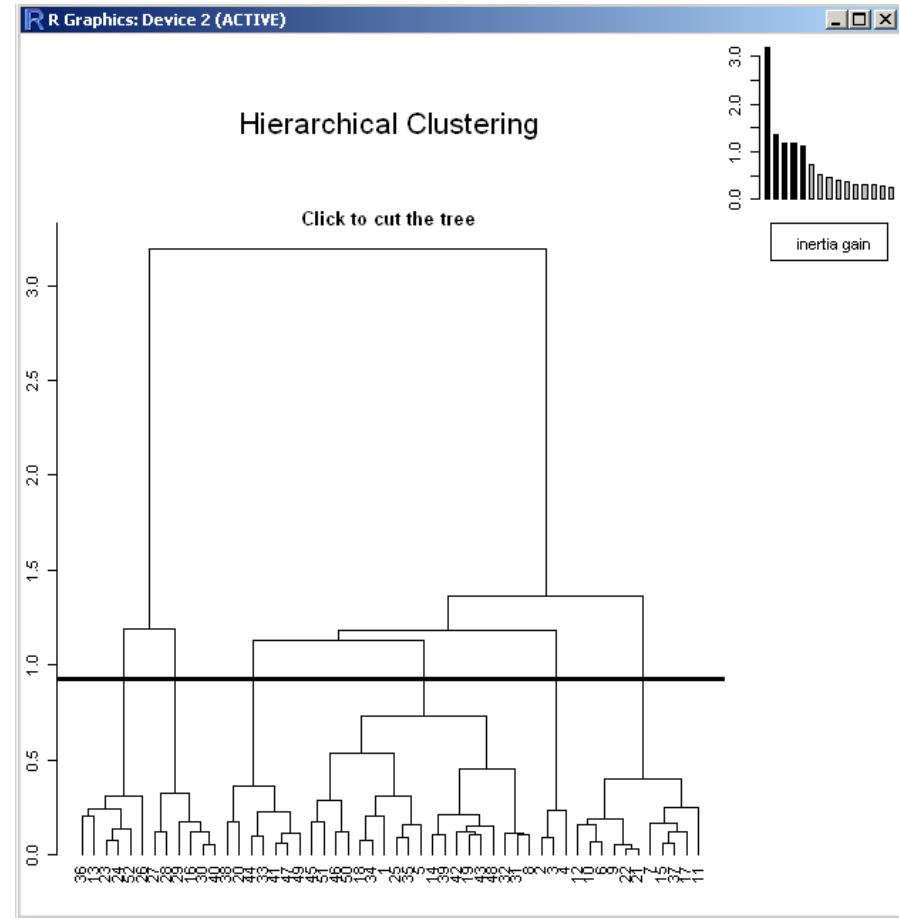
## 4.6 – Le choix du nombre de classes

### Dendrogramme

produit par *FactoMineR*  
pour l'exemple « Emmental »

### Questions

- Combien de classes choisir ?
- Existe-t-il un nombre de classes « naturel » ?

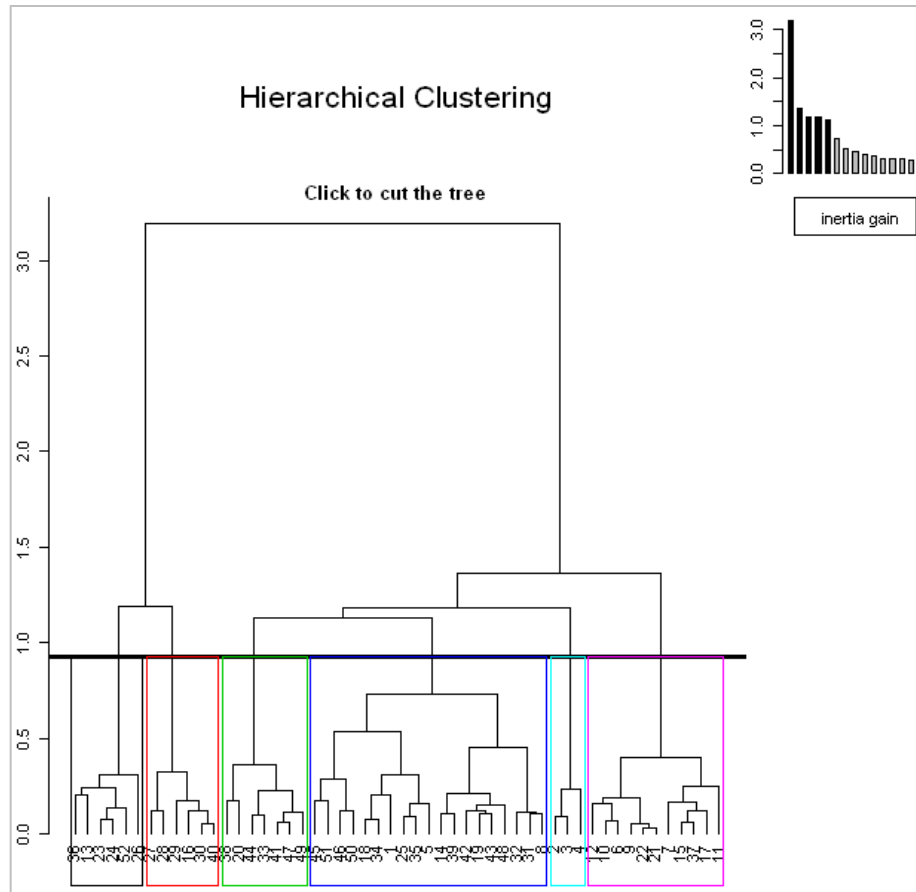


## La première étape de l'algorithme...

Matrice  
de distances  
entre les  
52 emmentals



## Une partition = coupure des branches de l'arbre

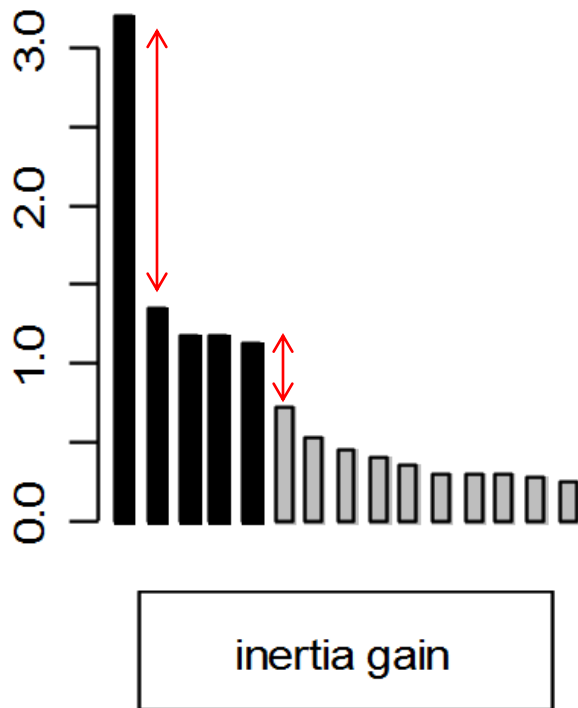


Une bonne partition :

**Les branches coupées  
sont « longues »**

*Les classes sont autant que  
possible distantes les unes  
des autres*

## Choix d'une partition à partir des indices de niveaux



On repère les étapes de l'algorithme où l'**augmentation de l'inertie** INTRA est importante

**Critère maximisé** dans FactoMineR

Le nombre optimal de classes  $k^*$  est tel que

$$k^* = \operatorname{argmin}_k \left\{ \frac{I_W(k)}{I_W(k-1)} \right\}$$

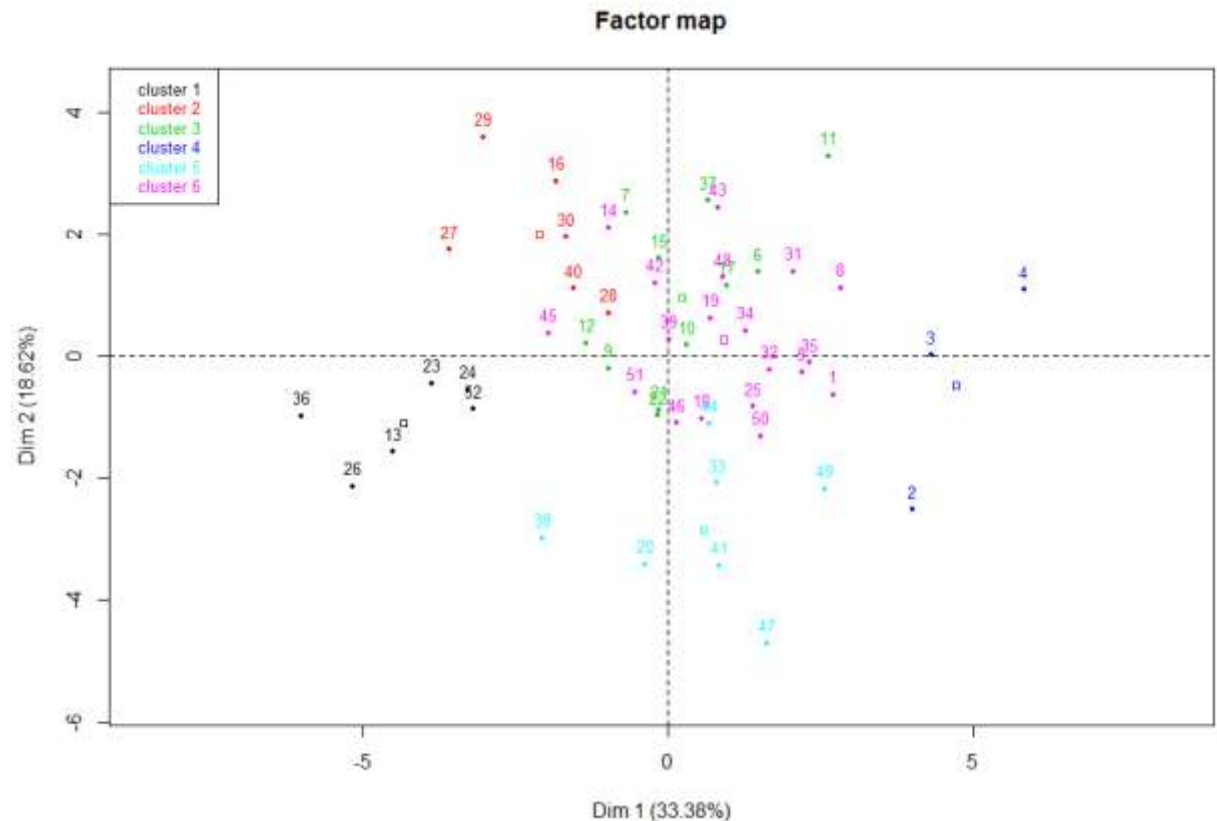
## Représentation de la partition sur un plan factoriel

La partition en  $K$  classes obtenue définit une nouvelle **variable qualitative** à  $K$  modalités

Emmental	intensité du parfum	parfum propionique	parfum butyrique	texture ferme	texture souple	texture granuleuse	texture collante	texture fondante	texture caractéristique	intensité du gout	gout acide	gout salé	gout sucré	gout piquant	gout fruité	gout amer	gout caractéristique	Conformité	Partition 6 classes
1	5,1	4	3,7	3,8	4,8	3,7	3,3	3	3,9	5,6	4,8	4,3	3,9	4,1	3,4	3,2	3,6	non	C4
2	4,7	3,9	3,4	5,2	3	4	3,3	4	3,6	5,8	4,3	4,9	4	4,3	4,8	3,2	3,7	oui	C5
3	4,7	4,2	2,9	3,7	3,3	2,8	3,4	4,7	3,9	5,7	4,3	4,7	3,9	4,2	4,9	3	4	oui	C5
4	5,3	4,6	3,9	3	3,6	2,4	3,4	5,4	4,2	5,6	4,2	4,6	4	4,2	4,9	4,1	4	oui	C5
5	4,3	4	2,6	3,9	4,1	3,3	2,9	4,2	3,7	5,1	4	4,3	3,9	4	4,3	3,8	4	oui	C5
6	4,7	4	3,5	4,5	4,6	3,3	3,5	4,2	4,4	4,9	3,6	4,2	3,2	3,5	4	3,2	3,9	oui	C6
7	3,6	3,7	2,6	4,1	4,4	3,1	3,4	4,5	4,1	4,5	3,4	3,5	3,2	3	4,2	2,6	4,3	oui	C6
8	5,4	3,9	4	4,1	3,9	2,5	3,9	4,9	4	5	4,2	4,2	3,1	4,1	3,8	2,9	3,6	oui	C4
9	4	3,8	2,6	4,5	4	4,1	3,3	3,3	4,1	4,5	3,1	3,9	3,3	3,1	3,8	3,2	3,5	oui	C6
10	5	4,2	3,3	4,7	4,3	3,8	3,7	3,5	3,8	4,8	3,6	4	3	3,3	3,3	3,1	3,7	oui	C6
⋮																			⋮
52	3,7	4,1	3	3,7	4,2	4,4	2,7	2,6	3,2	3,8	2,9	3,1	2,9	3	3,4	3,5	3,2	non	C1

## Identification des 6 classes sur le plan (1,2) de l'ACP

La variable de partition peut être projetée en tant que **variable supplémentaire** dans l'ACP du tableau des données sensorielles



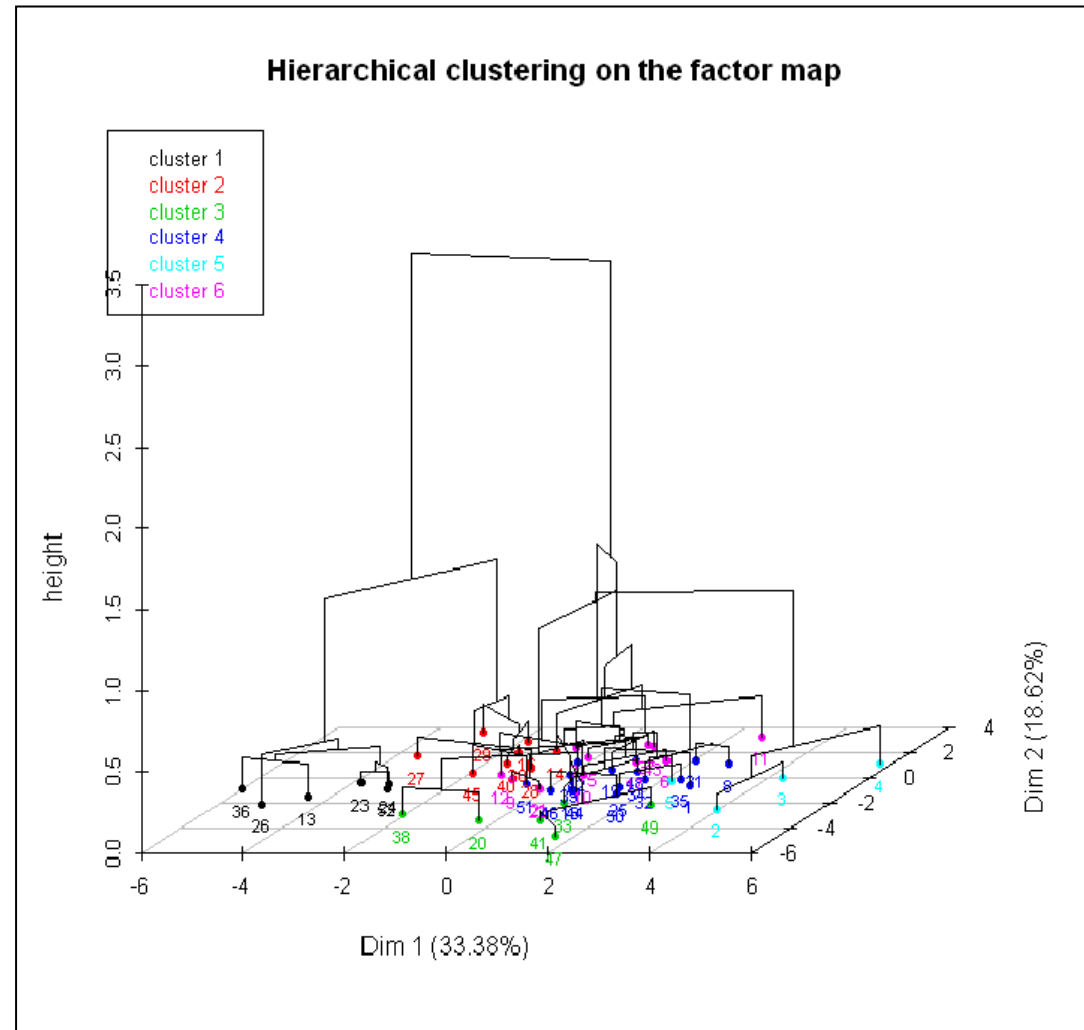


# Représentation simultanée du plan factoriel et du dendrogramme

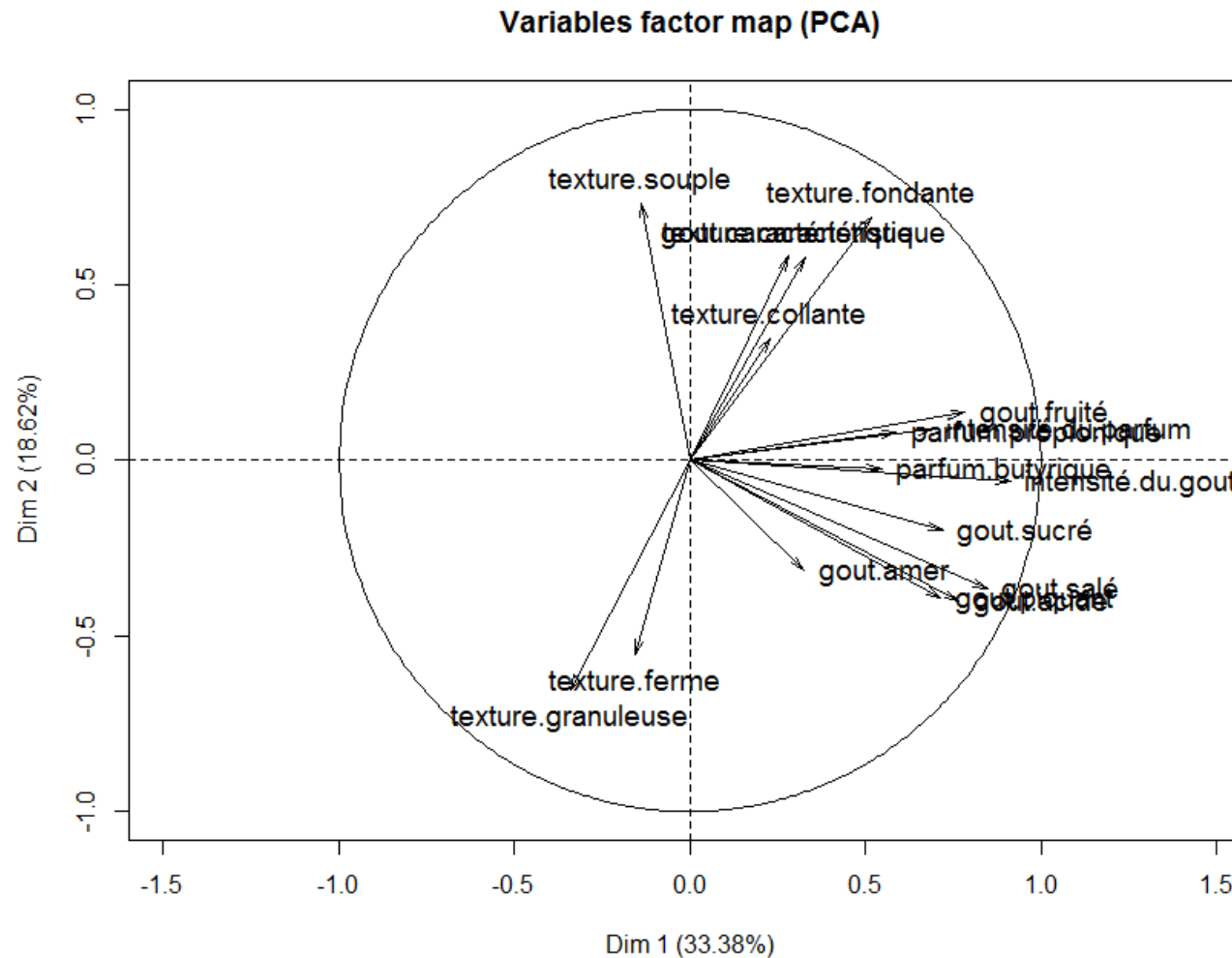
Combine les avantages

- d'une **vision géométrique** des proximités entre individus et entre classes d'individus
- de l'aspect **synthétique** de la classification

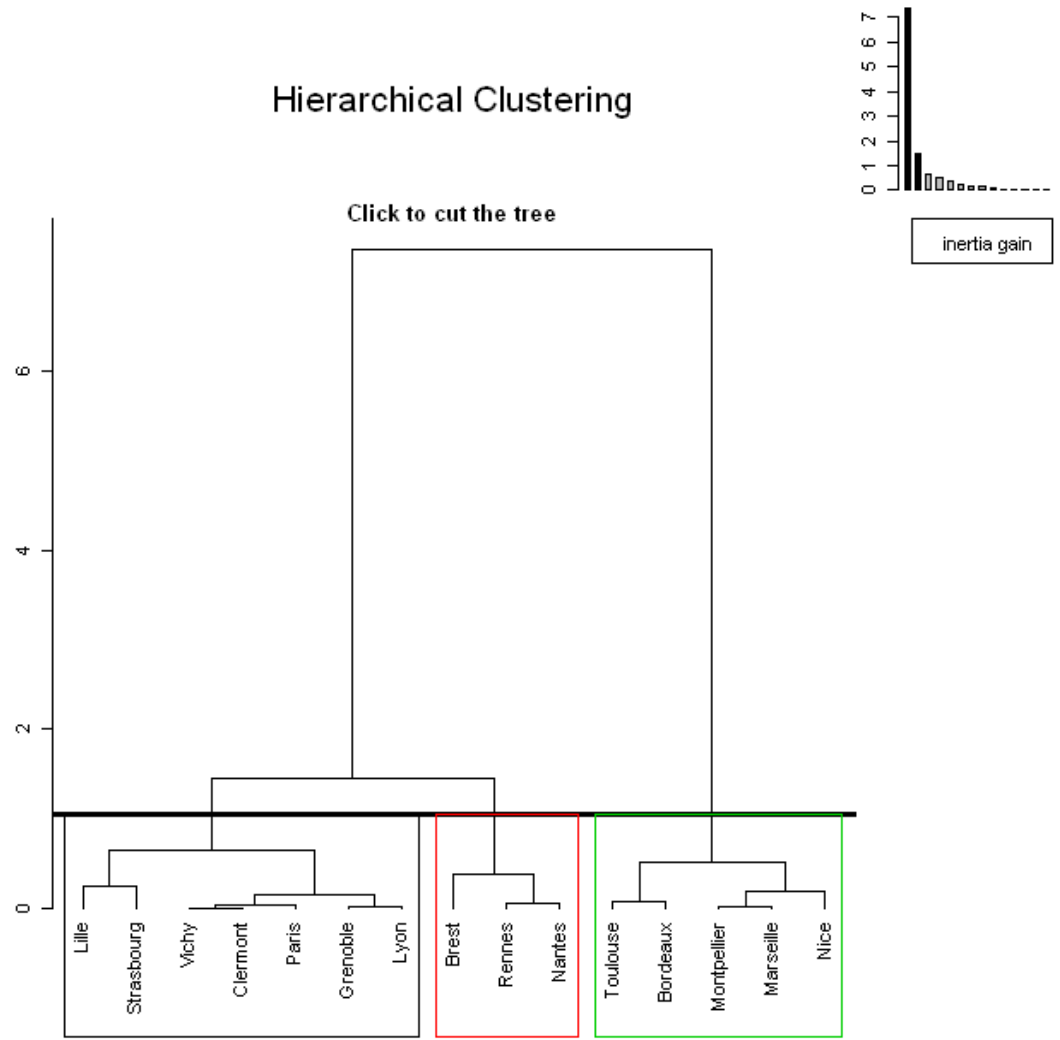
**Question** : comment caractériser chacune des classes obtenues à l'aide des variables ?

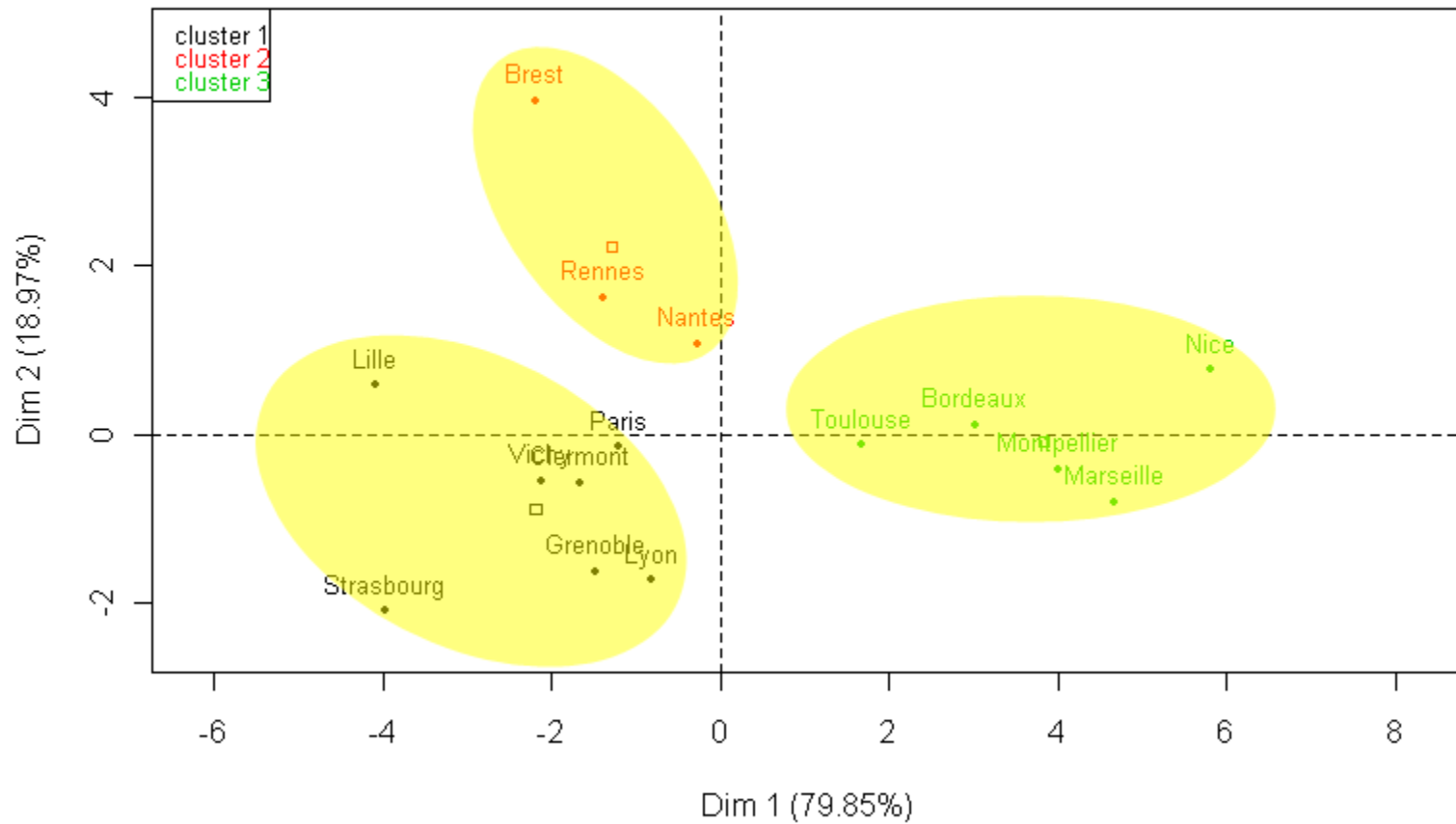


## Le cercle des corrélations – Axes 1 et 2

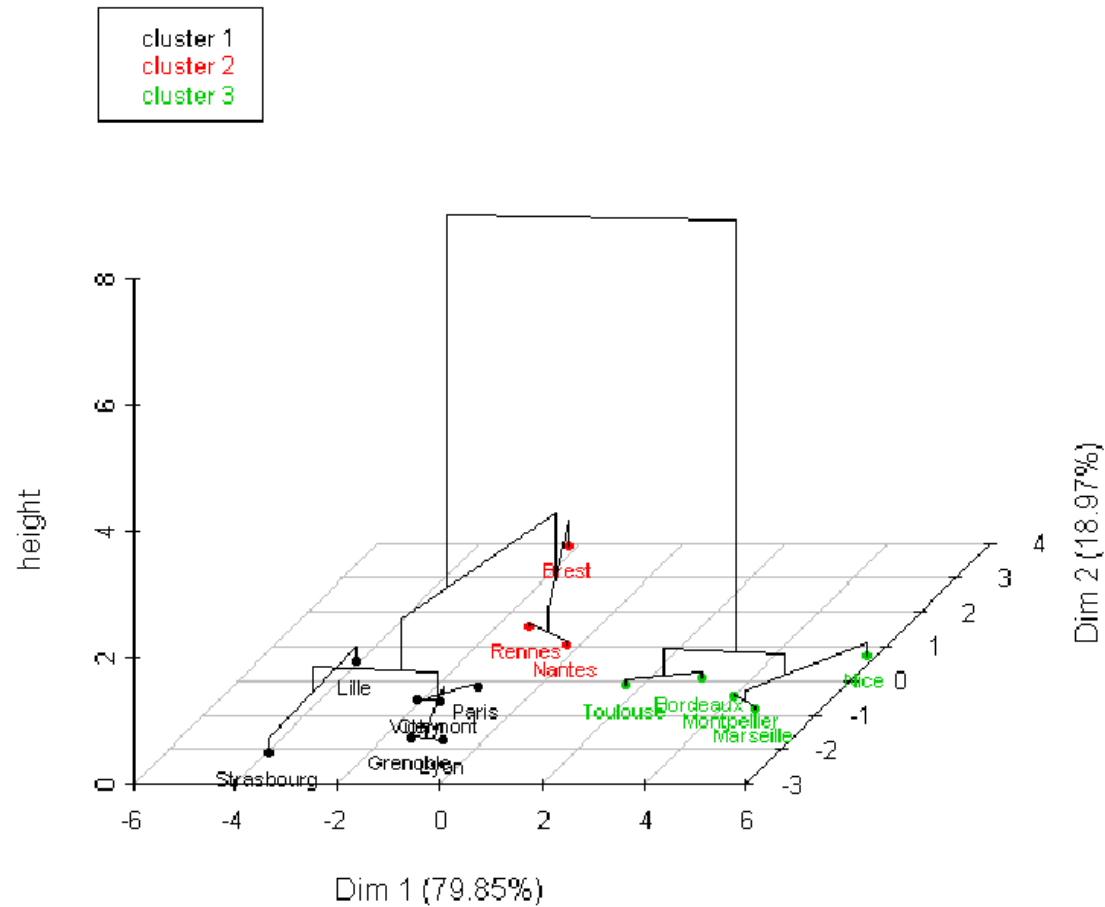


# Exemple « Températures »



**Factor map**

## Hierarchical clustering on the factor map



## 4.7 – Classification de données qualitatives

### Mesures de ressemblance entre individus

- **Variables binaires** de type présence/absence  
Utilisation de *coefficients d'association* ou *indices de similarité*
- **Variables qualitatives quelconques**  
Utilisation de la distance du khi2  
(cohérent avec les choix de l'ACM)



- Jaccard (1908):  $\frac{n_{JK}}{n_{JK} + n_{jK} + n_{Jk}}$
- Sokal et Michener (1958):  $\frac{n_{JK} + n_{jk}}{n}$
- Dice (1915):  $\frac{2n_{JK}}{2n_{JK} + n_{jK} + n_{Jk}}$
- Ochiai (1957):  $\frac{n_{JK}}{\sqrt{(n_{JK} + n_{jK})(n_{JK} + n_{Jk})}}$
- Russel et Rao (1940):  $\frac{n_{JK}}{n}$
- Kulczynski (1927):  $\frac{n_{JK}}{n_{jK} + n_{Jk}}$

### Indices d'agrégation

- Plusieurs choix possibles
- Indice de Ward compatible avec distances quadratiques

## 4.8 – Classification sur facteurs

### Principe

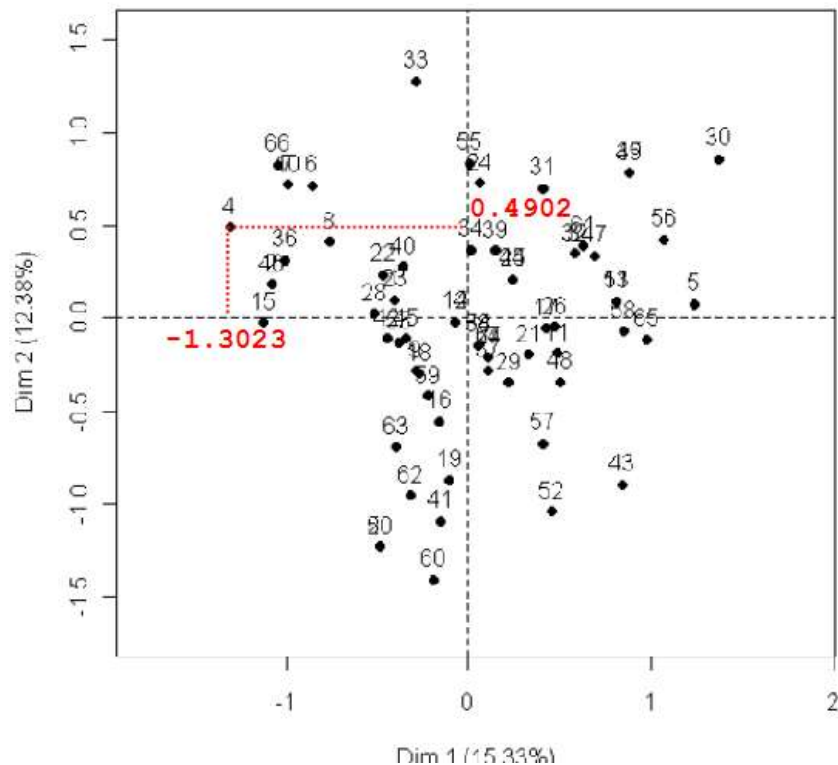
- On remplace les données des variables initiales par les coordonnées des individus sur les différentes dimensions (axes, facteurs) de l'analyse factorielle

*On parle aussi de classification sur composantes principales*

- C'est la stratégie programmée dans FactoMineR (initialement : SPAD)
- Applicable aux coordonnées issues de n'importe quelle analyse factorielle (ACP, AFC, ACM)

## Etape (1) : analyse factorielle

### Illustration : données « Crédit »



```
> res$ind$coord
```

	Dim 1	Dim 2
1	-0.4079	0.0958
2	-1.0737	0.1815
3	0.5780	0.3464
<b>4</b>	<b>-1.3023</b>	<b>0.4902</b>
5	1.2388	0.0722
6	-0.8545	0.7146
7	-0.9888	0.7172
8	-0.7601	0.4085
9	-0.2855	-0.2845
10	-0.9888	0.7172
11	0.4869	-0.1902
12	-0.0726	-0.0187
13	0.8116	0.0899
14	-0.0726	-0.0187
15	-1.1212	-0.0258



## Etape (2) : Construction du tableau des coordonnées

Cient	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12	Dim 13
1	-0.41	0.10	-0.32	-0.18	-0.46	0.23	-0.72	0.40	-0.13	-0.36	-0.15	-0.29	-0.34
2	-1.07	0.18	0.02	0.39	-0.24	0.56	0.10	-0.44	0.50	0.05	-0.26	-0.46	-0.05
3	0.58	0.35	-0.44	0.17	0.24	-0.49	-0.15	-0.57	-0.11	-0.43	0.24	-0.11	-0.04
4	-1.30	0.49	0.59	0.04	-0.14	-0.22	0.06	-0.01	-0.39	-0.19	-0.23	-0.15	0.22
5	1.24	0.07	0.60	-0.14	-1.29	0.05	0.66	0.18	0.13	0.03	-0.23	0.17	0.21
6	-0.85	0.71	0.16	-0.33	0.07	-0.40	-0.38	-0.07	0.00	-0.50	-0.30	0.49	0.46
7	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
8	-0.76	0.41	-0.08	0.81	-0.21	0.31	0.40	-0.42	0.41	0.22	-0.20	-0.16	-0.49
9	-0.29	-0.28	-1.00	0.44	-0.20	0.01	0.09	0.24	0.34	0.79	0.16	0.25	-0.07
10	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
⋮							⋮						⋮



La même information !



Cient	Marche	Apport	Impaye	Assurance	Endettement
1	Renovation	pas_Apport	Imp_0	AID	End_1
2	Renovation	pas_Apport	Imp_0	Sans Assur	End_2
3	Voiture	Apport	Imp_0	AID	End_3
4	Renovation	pas_Apport	Imp_0	Senior	End_2
5	Scooter	Apport	Imp_3 et +	AID	End_4
6	Renovation	pas_Apport	Imp_0	Senior	End_3
7	Renovation	Apport	Imp_0	Senior	End_2
8	Renovation	Apport	Imp_0	Sans Assur	End_2
9	Mobil Ameub	Apport	Imp_0	Sans Assur	End_1
10	Renovation	Apport	Imp_0	Senior	End_2
⋮					

Les variables **qualitatives** initiales  
sont remplacées par de nouvelles  
variables **quantitatives** :  
*les Facteurs de l'ACM*

## Etape (3) : CAH sur coordonnées factorielles

- Le tableau des coordonnées factorielles est soumis à une CAH avec les choix *Distance euclidienne + Indice d'agrégation de Ward*
- Possibilité de réaliser la CAH sur un sous-ensemble des coordonnées factorielles

Client	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12	Dim 13
1	-0.41	0.10	-0.32	-0.18	-0.46	0.23	-0.72	0.40	-0.13	-0.36	-0.15	-0.29	-0.34
2	-1.07	0.18	0.02	0.39	-0.24	0.56	0.10	-0.44	0.50	0.05	-0.26	-0.46	-0.05
3	0.58	0.35	-0.44	0.17	0.24	-0.49	-0.15	-0.57	-0.11	-0.43	0.24	-0.11	-0.04
4	-1.30	0.49	0.59	0.04	-0.14	-0.22	0.06	-0.01	-0.39	-0.19	-0.23	-0.15	0.22
5	1.24	0.07	0.60	-0.14	-1.29	0.05	0.66	0.18	0.13	0.03	-0.23	0.17	0.21
6	-0.85	0.71	0.16	-0.33	0.07	-0.40	-0.38	-0.07	0.00	-0.50	-0.30	0.49	0.46
7	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
8	-0.76	0.41	-0.08	0.81	-0.21	0.31	0.40	-0.42	0.41	0.22	-0.20	-0.16	-0.49
9	-0.29	-0.28	-1.00	0.44	-0.20	0.01	0.09	0.24	0.34	0.79	0.16	0.25	-0.07
10	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
⋮						⋮							⋮

## Intérêt de la classification sur facteurs

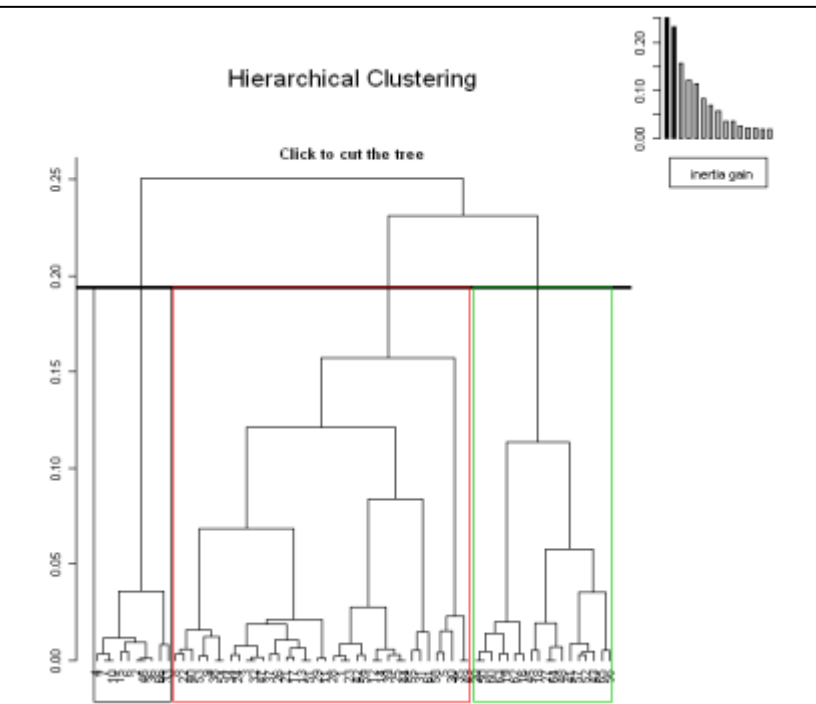
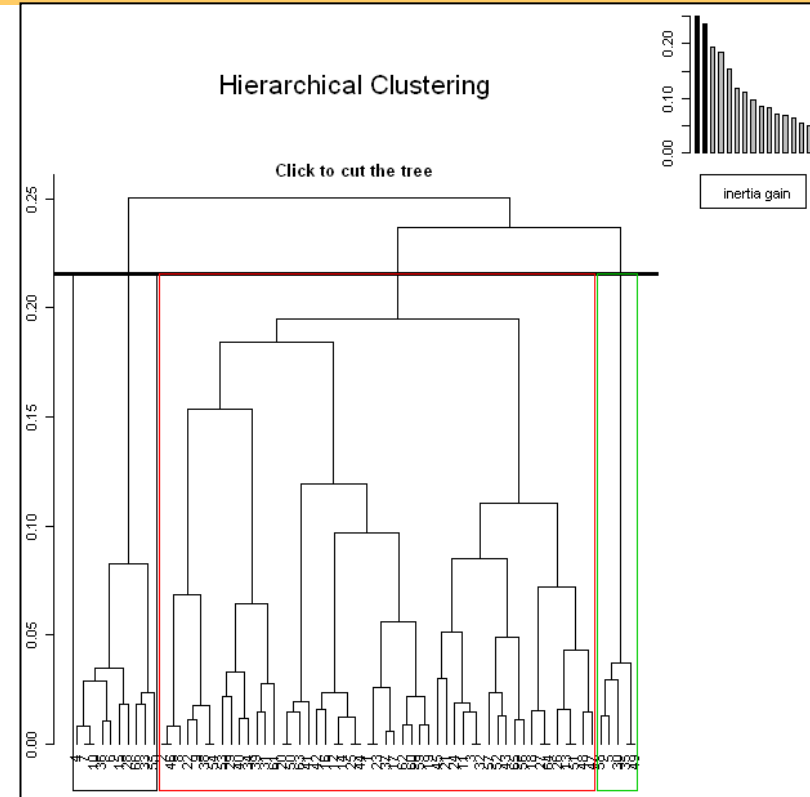
- On élimine les **derniers axes**, souvent porteurs de « **bruit** », de fluctuations aléatoires
- La classification est « **lissée** », les classes souvent plus homogènes
- Peut être mise en œuvre également pour des variables quantitatives

Difficulté parfois de choisir le nombre d'axes à conserver...

Client	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12	Dim 13
1	-0.41	0.10	-0.32	-0.18	-0.46	0.23	-0.72	0.40	-0.13	-0.36	-0.15	-0.29	-0.34
2	-1.07	0.18	0.02	0.39	-0.24	0.56	0.10	-0.44	0.50	0.05	-0.26	-0.46	-0.05
3	0.58	0.35	-0.44	0.17	0.24	-0.49	-0.15	-0.57	-0.11	-0.43	0.24	-0.11	-0.04
4	-1.30	0.49	0.59	0.04	-0.14	-0.22	0.06	-0.01	-0.39	-0.19	-0.23	-0.15	0.22
5	1.24	0.07	0.60	-0.14	-1.29	0.05	0.66	0.18	0.13	0.03	-0.23	0.17	0.21
6	-0.85	0.71	0.16	-0.33	0.07	-0.40	-0.38	-0.07	0.00	-0.50	-0.30	0.49	0.46
7	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
8	-0.76	0.41	-0.08	0.81	-0.21	0.31	0.40	-0.42	0.41	0.22	-0.20	-0.16	-0.49
9	-0.29	-0.28	-1.00	0.44	-0.20	0.01	0.09	0.24	0.34	0.79	0.16	0.25	-0.07
10	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
⋮						⋮							⋮

## classification sur 13 facteurs

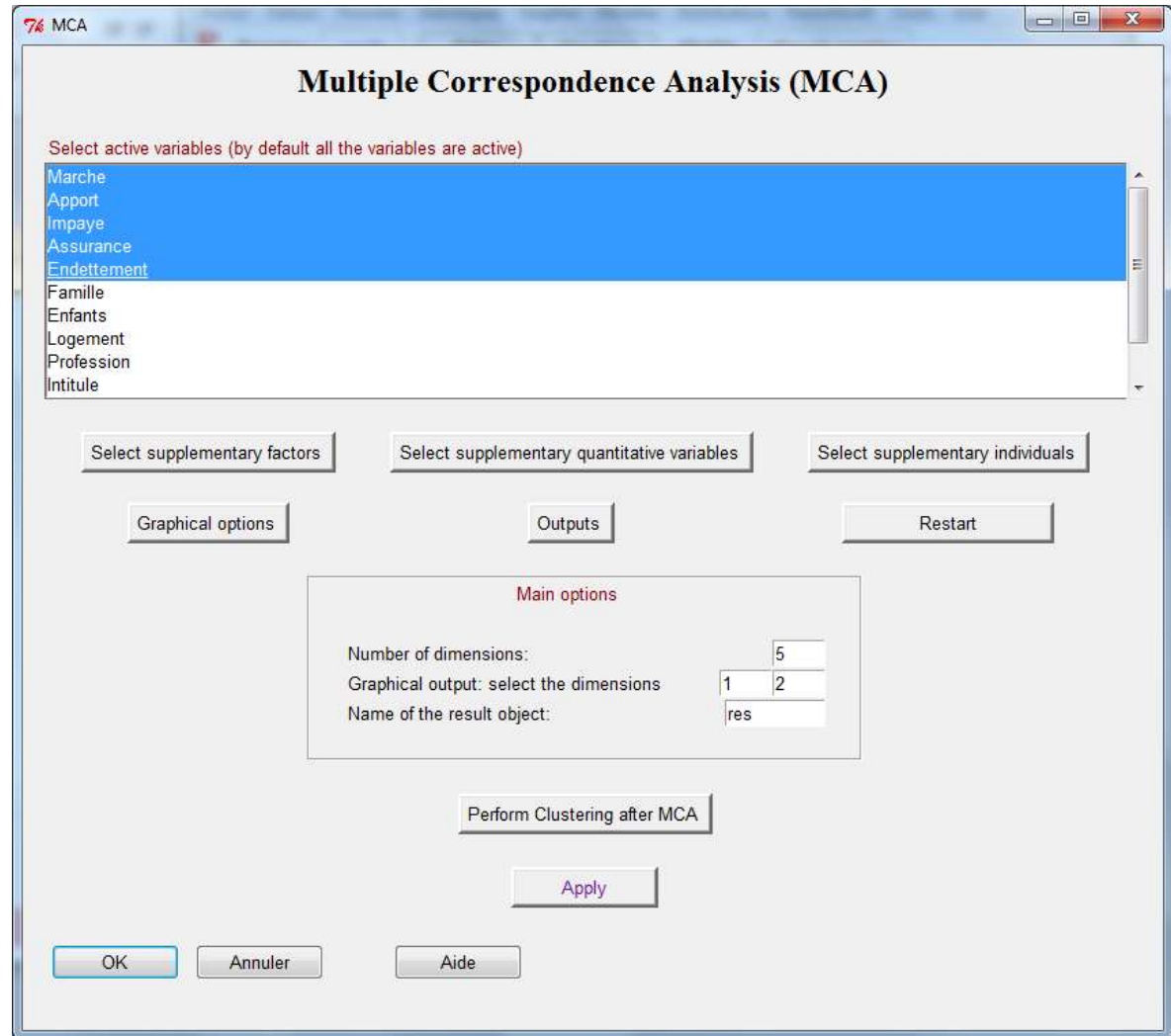
Client	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8	Dim 9	Dim 10	Dim 11	Dim 12	Dim 13
1	-0.41	0.10	-0.32	-0.18	-0.46	0.23	-0.72	0.40	-0.13	-0.36	-0.15	-0.29	-0.34
2	-1.07	0.18	0.02	0.39	-0.24	0.56	0.10	-0.44	0.50	0.05	-0.26	-0.46	-0.05
3	0.58	0.35	-0.44	0.17	0.24	-0.49	-0.15	-0.57	-0.11	-0.43	0.24	-0.11	-0.04
4	-1.30	0.49	0.59	0.04	-0.14	-0.22	0.06	-0.01	-0.39	-0.19	-0.23	-0.15	0.22
5	1.24	0.07	0.60	-0.14	-1.29	0.05	0.66	0.18	0.13	0.03	-0.23	0.17	0.21
6	-0.85	0.71	0.16	-0.33	0.07	-0.40	-0.38	-0.07	0.00	-0.50	-0.30	0.49	0.46
7	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
8	-0.76	0.41	-0.08	0.81	-0.21	0.31	0.40	-0.42	0.41	0.22	-0.20	-0.16	-0.49
9	-0.29	-0.28	-1.00	0.44	-0.20	0.01	0.09	0.24	0.34	0.79	0.16	0.25	-0.07
10	-0.99	0.72	0.49	0.46	-0.11	-0.47	0.35	0.01	-0.47	-0.02	-0.18	0.15	-0.21
...													



Client	Dim 1	Dim 2	Dim 3	Dim 4	Dim 5
1	-0.41	0.10	-0.32	-0.18	-0.46
2	-1.07	0.18	0.02	0.39	-0.24
3	0.58	0.35	-0.44	0.17	0.24
4	-1.30	0.49	0.59	0.04	-0.14
5	1.24	0.07	0.60	-0.14	-1.29
6	-0.85	0.71	0.16	-0.33	0.07
7	-0.99	0.72	0.49	0.46	-0.11
8	-0.76	0.41	-0.08	0.81	-0.21
9	-0.29	-0.28	-1.00	0.44	-0.20
10	-0.99	0.72	0.49	0.46	-0.11
...					

## classification sur 5 facteurs

## Mise en œuvre dans FactoMineR



The screenshot shows the 'Multiple Correspondence Analysis (MCA)' dialog box in the FactoMineR software. The window title is '76 MCA'. The main title is 'Multiple Correspondence Analysis (MCA)'. Below the title, it says 'Select active variables (by default all the variables are active)'. A list of variables is shown: Marche, Apport, Impaye, Assurance, Endettement, Famille, Enfants, Logement, Profession, and Intitule. The first five variables are highlighted in blue. Below the list, there are three buttons: 'Select supplementary factors', 'Select supplementary quantitative variables', and 'Select supplementary individuals'. Below these are three more buttons: 'Graphical options', 'Outputs', and 'Restart'. In the center, there is a 'Main options' section with three fields: 'Number of dimensions:' with a value of 5, 'Graphical output: select the dimensions' with values 1 and 2, and 'Name of the result object:' with the value 'res'. Below this section is a button 'Perform Clustering after MCA'. At the bottom, there is an 'Apply' button. At the very bottom, there are three buttons: 'OK', 'Annuler', and 'Aide'.

76 MCA

### Multiple Correspondence Analysis (MCA)

Select active variables (by default all the variables are active)

- Marche
- Apport
- Impaye
- Assurance
- Endettement
- Famille
- Enfants
- Logement
- Profession
- Intitule

Select supplementary factors    Select supplementary quantitative variables    Select supplementary individuals

Graphical options    Outputs    Restart

**Main options**

Number of dimensions: 5

Graphical output: select the dimensions 1 2

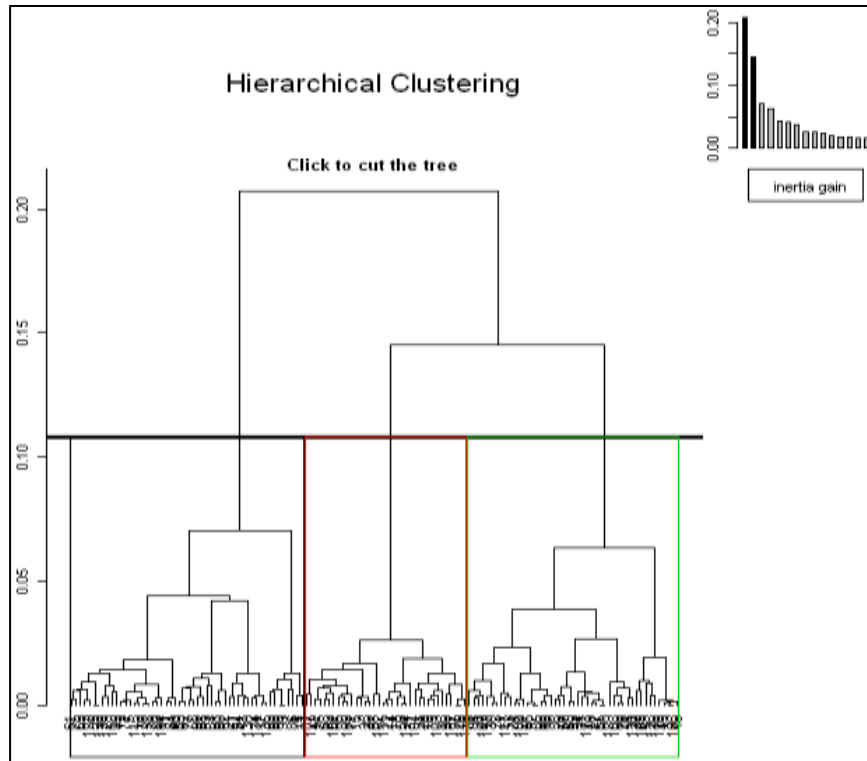
Name of the result object: res

Perform Clustering after MCA

Apply

OK    Annuler    Aide

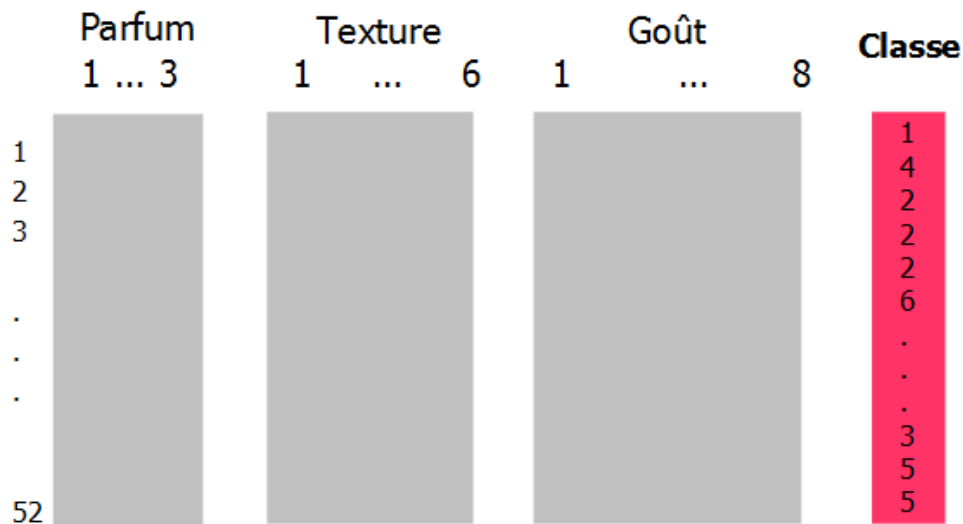
## 4.9 – La description des classes



- Quelles sont les variables les plus importantes pour caractériser la **partition** en général ?
- Peut-on décrire les individus de chaque **classe** par quelques variables caractéristiques ?
- Existe-t-il des **individus caractéristiques** ou typiques au sein d'une classe ?

## Exemple 1 : « Emmental »

Variables quantitatives



La CAH a permis de mettre en évidence 6 classes d'emmentals

## Exemple 2 : « OGM »

## Variables qualitatives

							Classe
Concerné	Position Culture	Position AI H	Position AI A	Manif	Media Actif	Info Active	
Pas du Tout	Plutôt Défavorable	Pas Favorable du Tout	Pas Favorable du Tout	Non	Oui	Non	2
Moyen	Favorable	Plutôt Défavorable	Favorable	Non	Oui	Non	2
Un Peu	Favorable	Plutôt Défavorable	Plutôt Défavorable	Non	Non	Non	2
Un Peu	Favorable	Favorable	Très Favorable	Non	Oui	Non	1
Moyen	Plutôt Défavorable	Favorable	Favorable	Non	Non	Non	3
Beaucoup	Favorable	Favorable	Très Favorable	Non	Non	Oui	2
Moyen	Favorable	Favorable	Très Favorable	Non	Non	Non	3
Pas du Tout	Favorable	Favorable	Favorable	Non	Oui	Non	
Moyen	Plutôt Défavorable	Plutôt Défavorable	Plutôt Défavorable	Non	Non	Oui	
Moyen	Pas Favorable du Tout	Pas Favorable du Tout	Pas Favorable du Tout	Non	Non	Oui	3
Un Peu	Favorable	Favorable	Très Favorable	Non	Oui	Non	1
Moyen	Favorable	Favorable	Favorable	Non	Oui	Non	2
Moyen	Favorable	Favorable	Favorable	Non	Non	Oui	3
Moyen	Plutôt Défavorable	Plutôt Défavorable	Plutôt Défavorable	Non	Oui	Non	3
Un Peu	Plutôt Défavorable	Plutôt Défavorable	Plutôt Défavorable	Non	Non	Non	
Moyen	Plutôt Défavorable	Plutôt Défavorable	Plutôt Défavorable	Non	Non	Oui	1
Moyen	Favorable	Plutôt Défavorable	Favorable	Non	Non	Non	1

La CAH a permis de mettre en évidence 3 classes d'enquêtés



## 4.9.1 Caractérisation de la partition par les variables

### Question

Quelles sont les variables qui ont « contribué » le plus à la création des classes de la partition ?

### Méthodologie statistique

Étude de la liaison entre la variable de partition (qualitative) et chaque variable (qualitative ou quantitative) :

- *Si qualitative : liaison entre deux variables qualitatives*  
Tableau de contingence, **test du Chi2**
- *Si quantitative : liaison entre une variable qualitative et une quantitative*  
**Analyse de la variance** à un facteur

## Description de la partition par les variables qualitatives

- *Méthodologie statistique : le test du khi2*

Pour chaque variable, on teste l'hypothèse

$H_0$  : **Indépendance** entre la partition et la variable

- *Exemple*

Lien entre la partition et la question

« Quelle est votre position face à la culture des OGM ? »

Classe	PC_Favorable	PC_Pas Favorable du Tout	PC_Plutôt Défavorable
1	42	0	4
2	4	1	47
3	2	32	3

$p - value = 4,46 \text{ e-}41$

## Tableau de synthèse

```

$test.chi2
              p.value df
Position.Culture 4.464795e-41 4
Position.Al.A    4.943024e-36 6
Position.Al.H    1.202362e-33 4
Danger           9.559578e-12 2
Manif            2.314111e-07 2
Grds.Parents     4.492277e-07 2
Menace           3.346178e-06 2
Famine           3.937522e-06 2
Procédé.Inutile  2.553569e-05 2
Concerné         2.870236e-05 6
Parti.Politique  4.777906e-04 8
Risque.Eco       1.788315e-03 2
Produits.Phytosanitaires 4.014243e-03 2
Info.Active      7.627213e-03 2
CSP              8.994752e-03 18
Relation         1.325999e-02 2
Futur.Progrès    1.504807e-02 2
  
```

*Sortie FactoMineR*

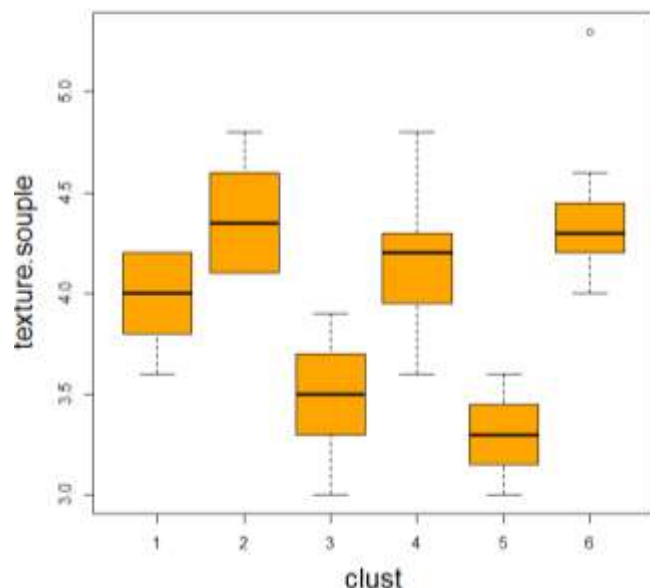
## Description de la partition par les variables quantitatives

- Méthodologie statistique : le test de Fisher*

Pour chaque variable, on teste l'hypothèse

$H_0$  : **égalité des moyennes** de la variable dans les  $K$  classes :  $\mu_1 = \mu_2 = \dots = \mu_K$

Test de Fisher de l'analyse de la variance à un facteur



Une variable fortement liée à la partition  
« Texture souple »

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
clust	5	6.135	1.2270	12.29	<b>1.36e-07 ***</b>
Residuals	46	4.592	0.0998		

## Tableau de synthèse

```
$quanti.var
      Eta2      P-value
intensité.du.gout    0.7701139 1.276654e-13
gout.salé            0.7186823 1.203066e-11
gout.fruité         0.6506794 1.520774e-09
texture.granuleuse   0.6440460 2.310547e-09
texture.souple      0.5719232 1.363082e-07
texture.fondante     0.5703218 1.479529e-07
gout.sucré           0.5301961 1.042116e-06
gout.caractéristique 0.4972990 4.524208e-06
texture.caractéristique 0.4822602 8.543592e-06
gout.acide           0.4774031 1.044420e-05
gout.piquant         0.4752752 1.139734e-05
texture.collante     0.3703465 5.380891e-04
parfum.propionique   0.3577279 8.108694e-04
intensité.du.parfum  0.3320218 1.811232e-03
texture.ferme        0.3309750 1.869837e-03
parfum.butyrique     0.2758158 9.117640e-03
```

*Sortie FactoMineR*

## 4.9.2 Caractérisation d'une classe par les variables

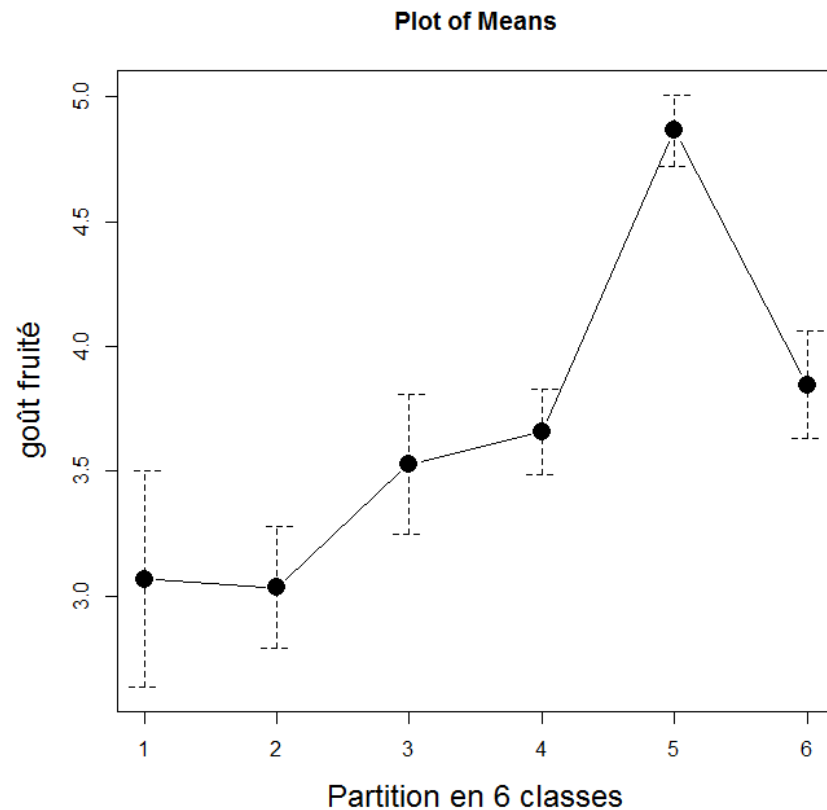
### Méthodologie

Toutes les variables du tableau de données, active ou supplémentaire, quantitative ou qualitative, sont analysées à tour de rôle dans chaque classe

### Critère d'intérêt d'une variable

Une variable est jugée caractéristique d'une classe si les individus de cette classe possèdent des **valeurs remarquables** pour cette variable par rapport aux individus de la population en général

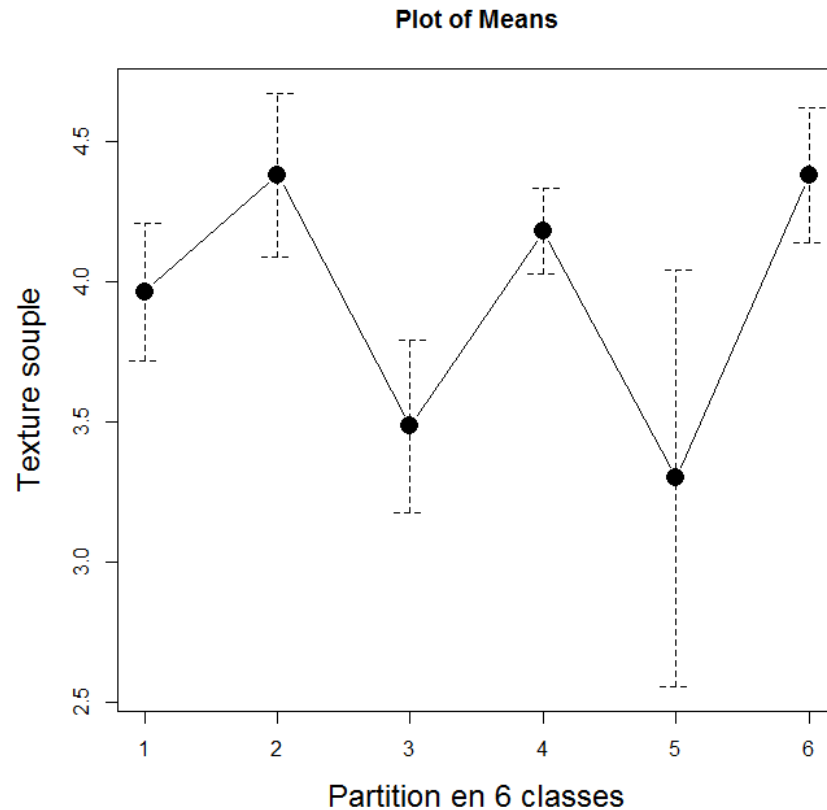
## Description d'une classe par les variables quantitatives



### *Goût fruité*

est caractéristique des emmentals de la classe 5

Sa moyenne dans la classe est significativement supérieure à la moyenne générale



### *Texture souple*

est caractéristique des emmentals  
des classes 3 et 5

Sa moyenne dans la classe est  
significativement inférieure à la  
moyenne générale



- *Méthodologie statistique*

Pour chaque variable  $X$ , on teste l'hypothèse :

$H_0$  : les individus de la classe  $k$  ont été tirés au hasard dans la population

*Statistique de test*

Sous  $H_0$ , la moyenne de  $X$  dans la classe ( $\bar{X}_k$ ) est peu différente de la moyenne de  $X$  dans la population ( $\bar{X}$ )

Plus précisément : sous l'hypothèse de normalité de  $X$ ,

$$\bar{X}_k \approx \mathcal{N}\left(\bar{X}, \frac{s}{\sqrt{n_k}} \sqrt{\frac{n-n_k}{n-1}}\right) \quad \text{ou} \quad V - \text{Test} = \frac{\bar{X}_k - \bar{X}}{s_{\bar{X}_k}} \approx \mathcal{N}(0,1)$$

## Description des classes 2 et 5

\$quanti\$`2`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
gout.amer	-2.058989	3.066667	3.448077	0.3299832	0.4777714	0.039495306
parfum.butyrique	-2.136781	2.700000	3.176923	0.4396969	0.5756653	0.032615779
gout.acide	-2.148858	3.083333	3.571154	0.2967416	0.5855098	0.031645675
intensité.du.gout	-2.408617	4.150000	4.659615	0.2565801	0.5457034	0.016013105
gout.salé	-2.470901	3.416667	3.873077	0.2477678	0.4764110	0.013477301
parfum.propionique	-2.644705	3.183333	3.663462	0.2192158	0.4682327	0.008176211
gout.fruité	-2.849798	3.033333	3.609615	0.2134375	0.5215582	0.004374695
gout.sucré	-3.084028	2.516667	3.075000	0.2671870	0.4669356	0.002042181
texture.granuleuse	-3.187580	2.383333	3.430769	0.2733537	0.8475164	0.001434686

\$quanti\$`5`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
gout.fruité	4.258905	4.866667	3.609615	0.04714045	0.5215582	2.054311e-05
gout.sucré	3.374374	3.966667	3.075000	0.04714045	0.4669356	7.398375e-04
intensité.du.gout	3.368876	5.700000	4.659615	0.08164966	0.5457034	7.547550e-04
gout.salé	3.190757	4.733333	3.873077	0.12472191	0.4764110	1.419003e-03
texture.fondante	2.473516	4.700000	3.776923	0.57154761	0.6594331	1.337907e-02
parfum.propionique	2.150617	4.233333	3.663462	0.28674418	0.4682327	3.150643e-02
gout.acide	2.099030	4.266667	3.571154	0.04714045	0.5855098	3.581426e-02
texture.souple	-3.030172	3.300000	4.078846	0.24494897	0.4541842	2.444142e-03

## Description d'une classe par les variables qualitatives

La fréquence d'une modalité au sein d'une classe est-elle  
*sur* ou *sous représentée* par rapport à sa fréquence dans la population ?

\$category\$`1`	Cla/Mod	Mod/Cla	Global	p.value	v.test
Position.Culture=Position.Culture_Favorable	87.500000	91.304348	35.555556	1.856756e-23	9.980422
Position.Al.H=Position.Al.H_Favorable	97.368421	80.434783	28.148148	3.676870e-23	9.912405
Position.Al.A=Position.Al.A_Favorable	84.090909	80.434783	32.592593	2.069184e-17	8.489842
Danger=Danger_Non	79.487179	67.391304	28.888889	5.692246e-12	6.887147
Grds.Parents=Grds.Parents_Non	63.265306	67.391304	36.296296	1.920186e-07	5.206903
Menace=Menace_Non	58.333333	60.869565	35.555556	2.626576e-05	4.203639
Famine=Famine_Oui	50.000000	73.913043	50.370370	1.441665e-04	3.800909
Parti.Politique=UMP	55.000000	47.826087	29.629630	1.996694e-03	3.090724
Position.Al.A=Position.Al.A_Très Favorable	87.500000	15.217391	5.925926	4.537076e-03	2.838186
Manif=Manif_Non	37.704918	100.000000	90.370370	6.427002e-03	2.725162
...					
Manif=Manif_Oui	0.000000	0.000000	9.629630	6.427002e-03	-2.725162
Famine=Famine_Non	17.910448	26.086957	49.629630	1.441665e-04	-3.800909
Menace=Menace_Oui	20.689655	39.130435	64.444444	2.626576e-05	-4.203639
Grds.Parents=Grds.Parents_Oui	17.441860	32.608696	63.703704	1.920186e-07	-5.206903
Position.Culture=Position.Culture_Pas Favorable du Tout	0.000000	0.000000	24.444444	1.659400e-07	-5.233934
Position.Al.A=Position.Al.A_Plutôt Défavorable	2.564103	2.173913	28.888889	1.447494e-07	-5.259116
Position.Culture=Position.Culture_Plutôt Défavorable	7.407407	8.695652	40.000000	4.642692e-08	-5.464479
Position.Al.A=Position.Al.A_Pas Favorable du Tout	2.272727	2.173913	32.592593	6.241264e-09	-5.810161
Danger=Danger_Oui	15.625000	32.608696	71.111111	5.692246e-12	-6.887147
Position.Al.H=Position.Al.H_Pas Favorable du Tout	0.000000	0.000000	37.037037	1.694777e-12	-7.057539

# Caractérisation d'une classe par les individus

## Individus moyens ou parangons

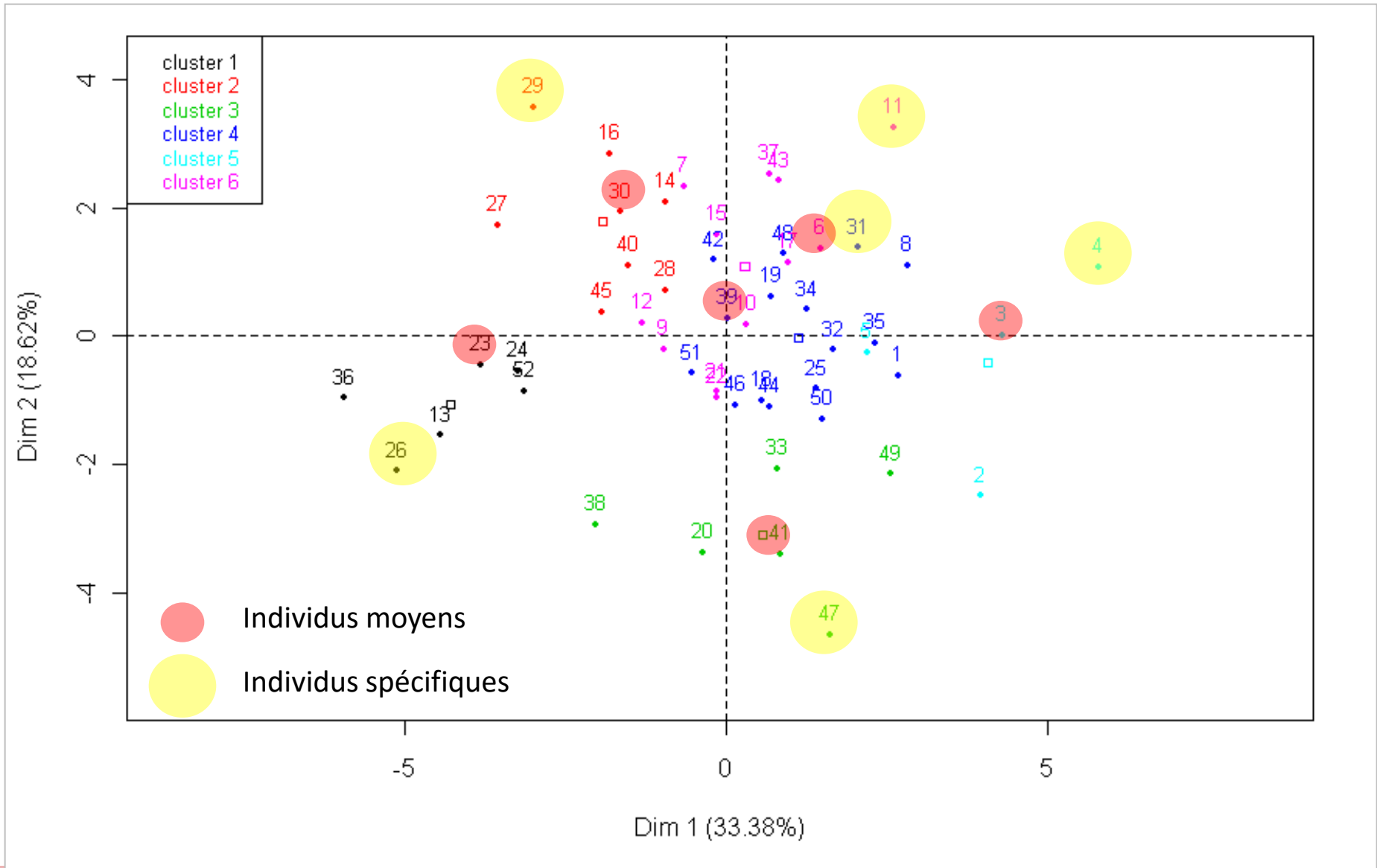
```

cluster: 1
      23      24      52      36      13
2.061208 2.478712 2.576674 3.146382 3.223483
-----
cluster: 2
      30      40      27      16      29
1.750817 2.157566 2.729179 2.740139 2.943888
-----
cluster: 3
      41      33      47      49      20
1.717848 1.896040 2.415650 2.784670 2.969118
-----
cluster: 4
      39      34      19      43      25
2.103091 2.409888 2.481580 2.659681 2.813457
-----
cluster: 5
      3      2      4
1.423763 2.657145 2.872238
-----
cluster: 6
      6      15      9      21      17
1.746877 1.993348 2.223134 2.296515 2.429671
  
```

## Individus spécifiques (éloignés des autres centres de gravité)

```

cluster: 1
      26      36      13      52      23
6.325038 6.313429 5.907366 4.799391 4.603200
-----
cluster: 2
      29      27      28      16      30
5.948479 5.025242 4.034969 3.688640 3.537056
-----
cluster: 3
      47      20      38      41      49
5.349783 5.104888 4.352613 3.948951 3.555841
-----
cluster: 4
      31      51      1      50      32
5.297346 4.955671 4.633673 4.626562 4.448787
-----
cluster: 5
      4      3      2
5.687283 5.098952 4.579356
-----
cluster: 6
      11      17      12      15      37
4.876756 4.777940 4.373645 4.264669 4.104429
  
```



PCA

### Principal Components Analysis (PCA)

Select active variables (by default all the variables are active)

- texture.fondante
- texture.caractéristique
- intensité.du.gout
- gout.acide
- gout.salé
- gout.sucré
- gout.piquant
- gout.fruité
- gout.amer
- gout.caractéristique

Select supplementary factors    Select supplementary variables    Select supplementary individuals

Graphical options    Outputs    Restart

Main options

Name of the result object: res

Number of dimensions: 5

Scale the variables: ☒

Graphical output: select the dimensions 1 2

Perform Clustering after PCA

Apply

OK    Annuler    Aide

Paramétrage de la classification  
sur facteurs  
(fonction HCPC de R)

HCPC options

### Hierarchical Clustering on Principal Components

Select options for the HCPC

Clustering is performed on the first 5 dimensions of the PCA.  
(Change your choice in the main options to change this number)

Choice of the number of clusters:    interactive ☒    automatic ☐

The optimal number of clusters is chosen between: 3 10

Consolidate clusters ☐

Print graphs ☒

Print results for clusters ☒

OK    Cancel