Classification automatique - clustering

Marie Chavent

Université de Bordeaux

Introduction

Comment définir automatiquement des groupes d'individus ou de variables qui se ressemblent ?

Exemple : données quantitatives décrivant 8 eaux minérales sur 13 variables.

	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline	appréciation.globale
St Yorre	3.4	3.1	2.9	6.4	4.8	2.9
Badoit	3.8	2.6	2.7	4.7	4.5	2.9
Vichy	2.9	2.9	2.1	6.0	5.0	2.8
Quézac	3.9	2.6	3.8	4.7	4.3	3.5
Arvie	3.1	3.2	3.0	5.2	5.0	2.9
Chateauneuf	3.7	2.8	3.0	5.2	4.6	3.3
Salvetat	4.0	2.8	3.0	4.1	4.5	3.4
Perrier	4.4	2.2	4.0	4.9	3.9	2.8

- A partir des distances entre individus : quelle mesure de distance ?
- A partir des liaisons entre les variables : quelle mesure de liaison ?

Dépend de la nature des données : quantitatives, qualitatives ou mixtes.

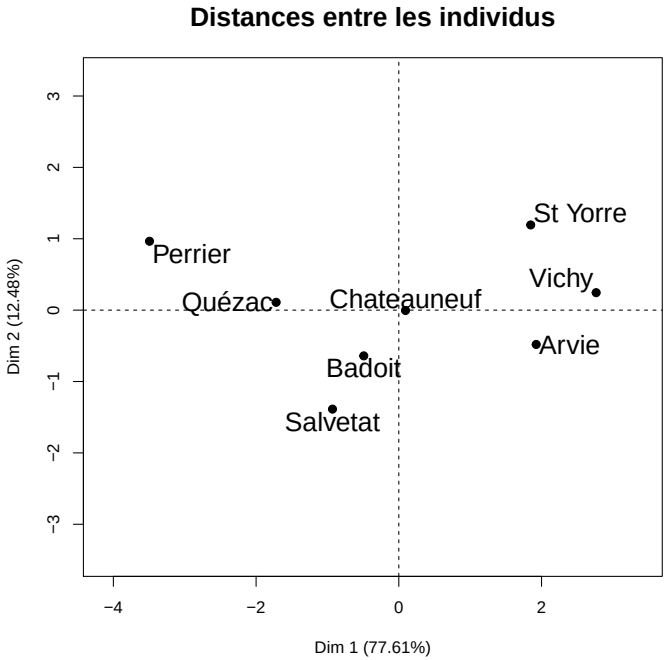
A partir des distances euclidiennes entre individus ?

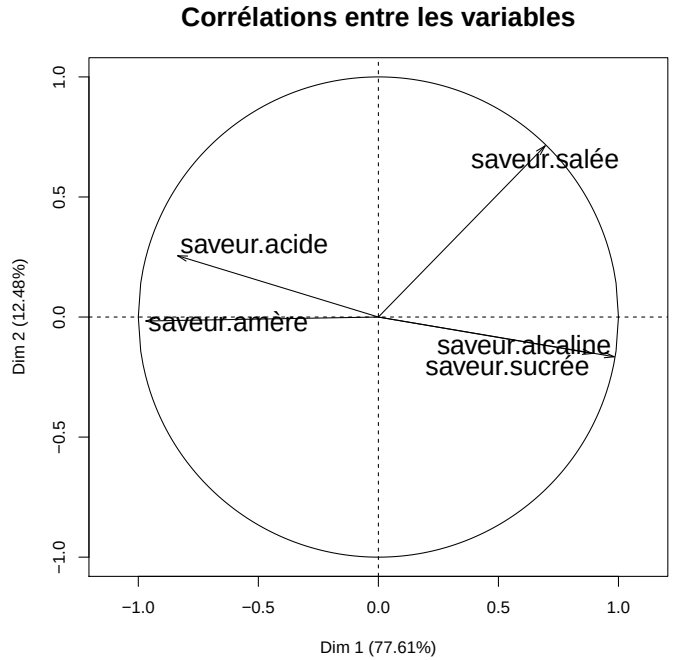
	St Yorre	Badoit	Vichy	Quézac	Arvie	Chateauneuf	Salvetat	Perrier
St Yorre	0.0	4.1	7.9	2.9	3.0	2.9	4.0	8.2
Badoit	4.1	0.0	4.8	5.3	1.8	1.8	1.2	10.6
Vichy	7.9	4.8	0.0	9.7	5.5	5.7	5.4	14.7
Quézac	2.9	5.3	9.7	0.0	4.7	4.3	4.9	6.2
Arvie	3.0	1.8	5.5	4.7	0.0	1.3	1.8	10.1
Chateauneuf	2.9	1.8	5.7	4.3	1.3	0.0	1.6	9.9
Salvetat	4.0	1.2	5.4	4.9	1.8	1.6	0.0	10.3
Perrier	8.2	10.6	14.7	6.2	10.1	9.9	10.3	0.0

A partir des corrélations entre les variables ?

	saveur.amère	saveur.sucrée	saveur.acide	saveur.salée	saveur.alcaline
saveur.amère	1.00	-0.83	0.78	-0.67	-0.96
saveur.sucrée	-0.83	1.00	-0.61	0.49	0.93
saveur.acide	0.78	-0.61	1.00	-0.44	-0.82
saveur.salée	-0.67	0.49	-0.44	1.00	0.56
saveur.alcaline	-0.96	0.93	-0.82	0.56	1.00

A partir d'une analyse en composantes principales (si les données sont quantitatives) ?





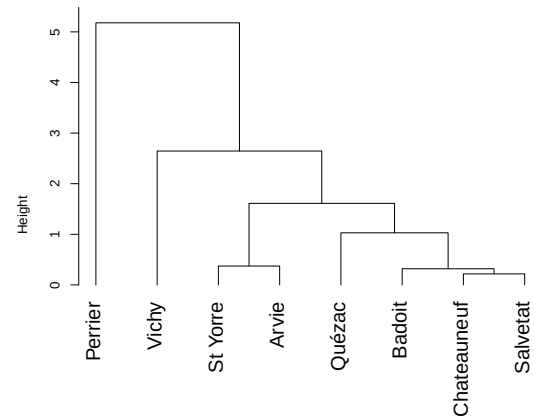
⇒ à partir d'une méthode de classification automatique (clustering).

Exemples de sorties de méthodes de classification automatique - clustering.

Partition en 4 classes des individus.

	P4
St Yorre	1
Badoit	2
Vichy	3
Quézac	2
Arvie	1
Chateauneuf	2
Salvetat	2
Perrier	4

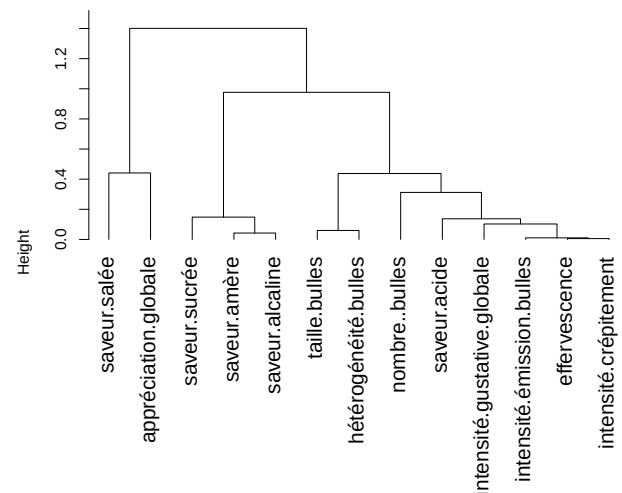
Hiérarchie des individus



Partition en 3 classes des variables.

	P3
saveur.amère	1
saveur.sucrée	1
saveur.acide	2
saveur.salée	3
saveur.alcaline	1
appréciation.globale	3
intensité.émission.bulles	2
nombre..bulles	2
taille.bulles	2
hétérogénéité.bulles	2
effervescence	2
intensité.gustative.globale	2
intensité.crépitement	2

Hiérarchie des variables



Il existe de nombreux algorithmes de classification automatique qui se distinguent par :

- la nature des objets à regrouper : les individus ou les variables,
- la nature des données : quantitatives, qualitatives ou mixtes,
- la nature de la structure de classification : partition ou hiérarchie,
- la nature de l'approche utilisée : approche géométrique (distance, dissimilarité, similarité) ou approche probabiliste (modèles de mélange).

Ici, on s'intéresse à la classification d'individus décrits par des données quantitatives, à l'aide d'approches géométriques utilisant les distances.

Outline

Données et objectifs

On s'intéresse à un tableaux de données numériques où les individus sont en lignes et les variables en colonnes.

	$1 \ldots$	j	$\ldots p$
1			
$\vdots$		$\vdots$	
i	$\ldots$	$x_{ij} \in \mathbb{R}$	$\ldots$
$\vdots$		$\vdots$	
n			

On notera w_i le poids de l'individu i avec en général :

- $w_i = \frac{1}{n}$ pour des observations aléatoires,
- $w_i \neq \frac{1}{n}$ pour des données ajustées, agrégées...

Objectifs.

Production d'une structure de classification (partition ou hiérarchie) permettant de mettre en évidence :

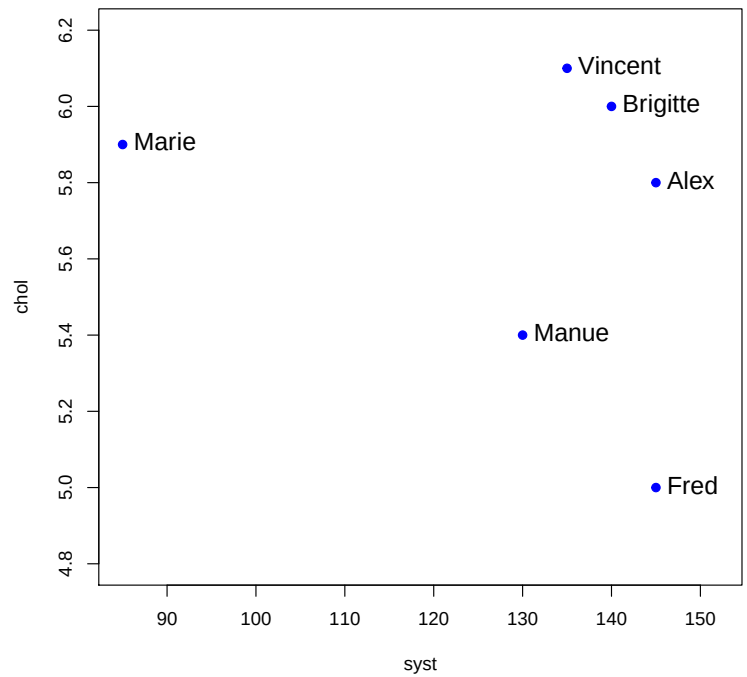
- des groupes d'individus (classes - clusters) : méthodes de partitionnement
- des liens hiérarchiques entre les individus : méthodes de classification hiérarchique.

Une partition en K classes des individus est un ensemble de classes non vides, deux à deux disjointes et dont la réunion est l'ensemble des individus.

On notera $P_K = (C_1, \ldots, C_k, \ldots, C_K)$.

Proposer une "bonne" et une "mauvaise" partition $P_3 = (C_1, C_2, C_3)$ en 3 classes des 6 individus ci-dessous.

	syst	chol
Brigitte	140	6.0
Marie	85	5.9
Vincent	135	6.1
Alex	145	5.8
Manue	130	5.4
Fred	145	5.0

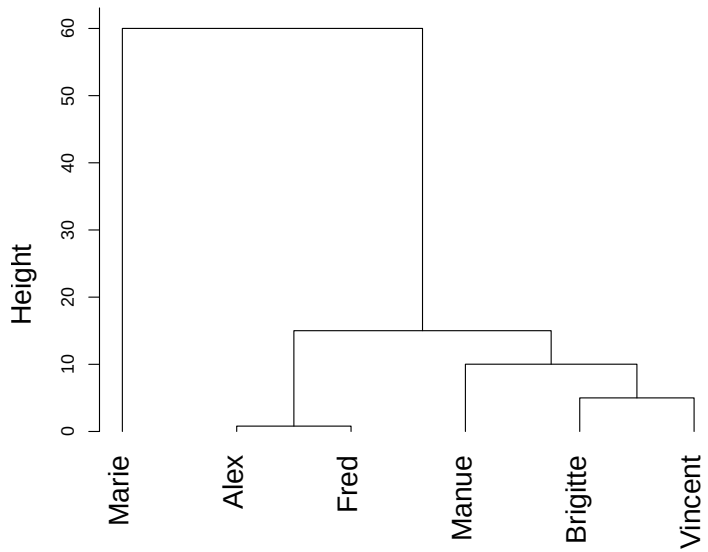


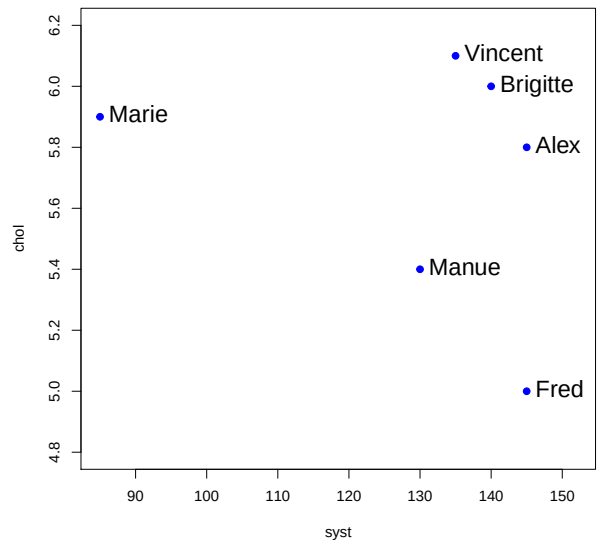
Une hiérarchie H des individus est un ensemble de classes non vides qui vérifie :

- H contient la classe de tous les individus,
- H contient tous les singletons (classes réduites à 1 individu),
- deux classes de H sont soit disjointes soit contenues l'une dans l'autre.

Un indice h est une fonction de H dans $\mathbb{R}^+$ qui donne la hauteur de chaque classe dans l'arbre hiérarchique (dendrogramme).

Quelle est la hiérarchie H du dendrogramme ci-dessous ?





Outline

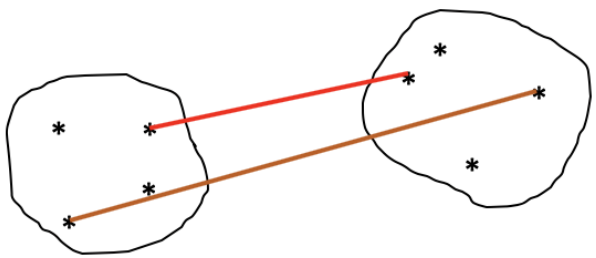
Classification ascendante hiérarchique (CAH)

L'algorithme de classification ascendante hiérarchique.

Idée : partir de la partition en n classes (des singletons) et agréger à chaque étape les deux classes qui se ressemblent le plus jusqu'à obtenir la partition en 1 classe.

$\Rightarrow$ Ressemblance entre individus : distance, dissimilarité, similarité.

$\Rightarrow$ Ressemblance entre classes : 2 exemples.



- Lien minimum (single link) : plus petite distance.
- Lien complet (complete link) : plus grande distance.

Exemple : 8 points de $\mathbb{R}^2$

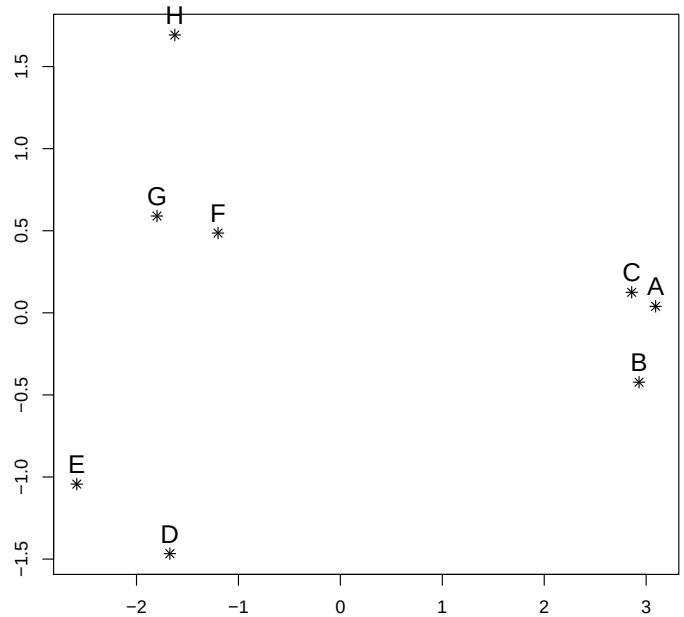
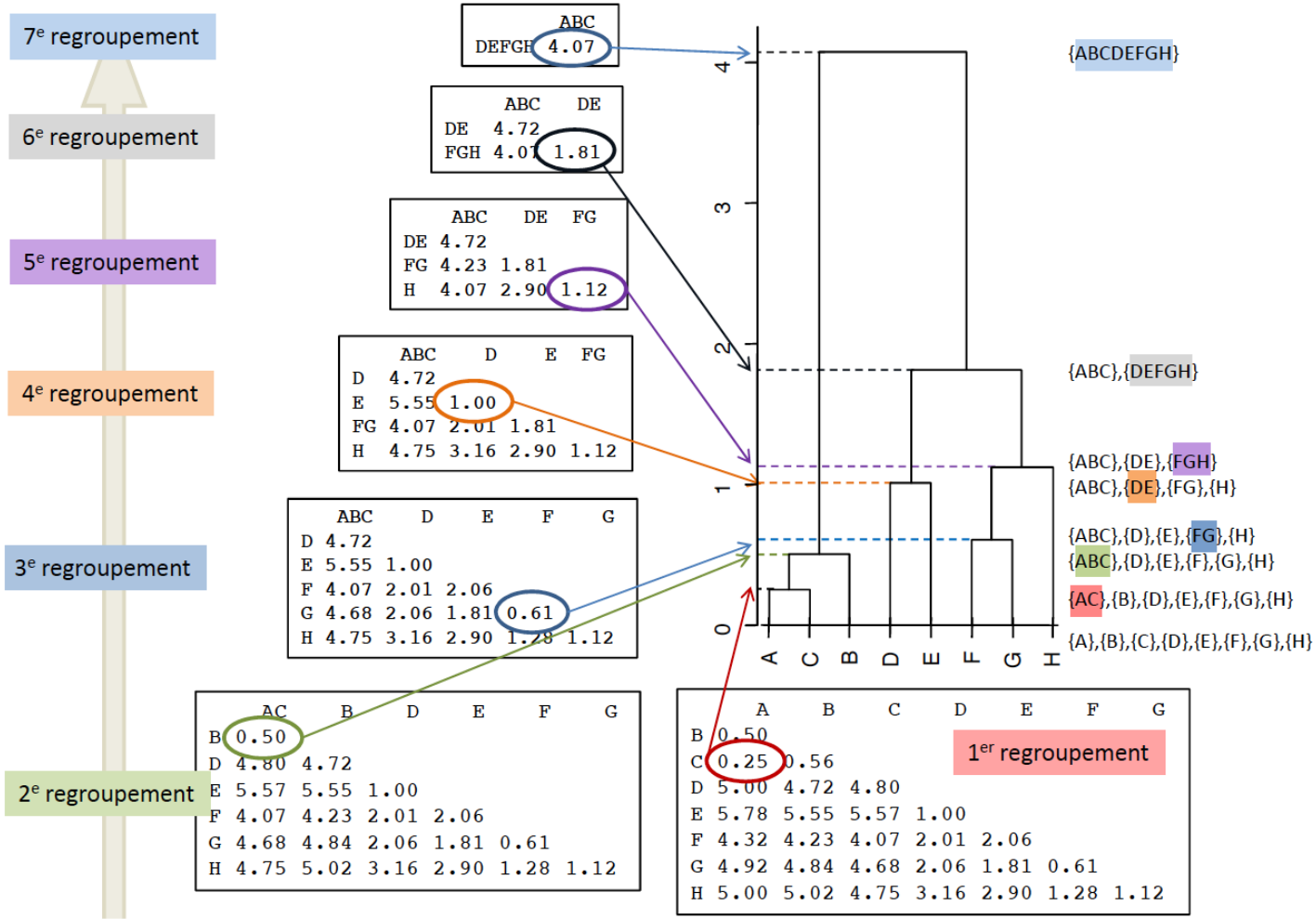


TABLE – Distances Euclidiennes entre les 8 individus

	A	B	C	D	E	F	G	H
A	0.00	0.50	0.25	5.0	5.8	4.32	4.92	5.0
B	0.50	0.00	4.72	4.7	5.5	4.23	4.84	5.0
C	0.25	0.56	0.00	4.8	5.6	4.07	4.68	4.8
D	5.00	4.72	4.80	0.0	1.0	2.01	2.06	3.2
E	5.78	5.55	5.57	1.0	0.0	2.06	1.81	2.9
F	4.32	4.23	4.07	2.0	2.1	0.00	0.61	1.3
G	4.92	4.84	4.68	2.1	1.8	0.61	0.00	1.1
H	5.00	5.02	4.75	3.2	2.9	1.28	1.12	0.0

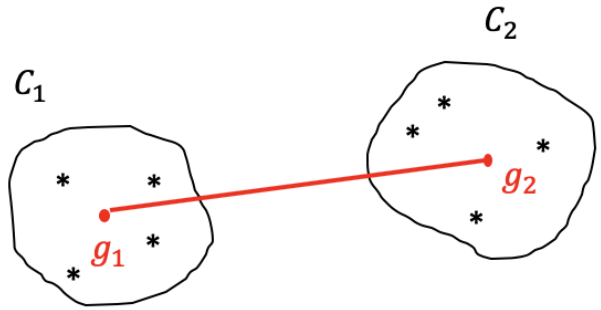
Algorithme du lien minimum appliqué aux 8 points.



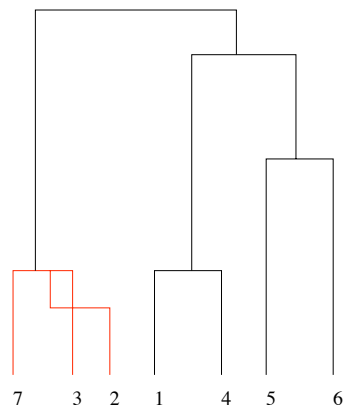
$\Rightarrow$ Indice h (hauteur d'une classe dans le dendrogramme) = lien minimum entre les deux sous-classes.

Autres mesures de ressemblances entre classes.

- Centroid : distance Euclidienne entre les centres de gravités des classes.



Problème : des inversions peuvent être observées dans l'arbre.



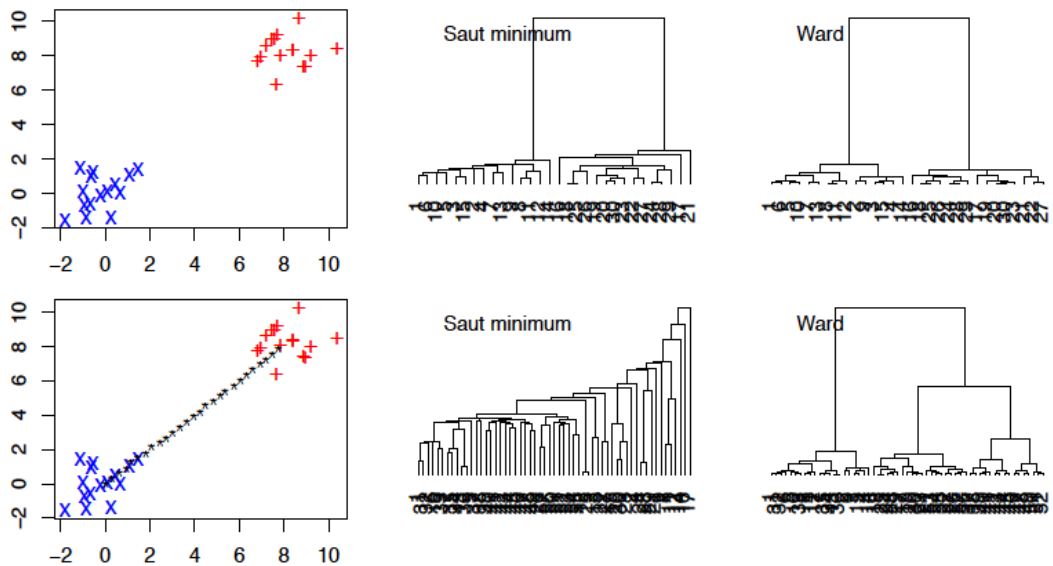
- Ward : distance pondérée entre les centres de gravités.

$$\frac{\mu_1 \mu_2}{\mu_1 + \mu_2} d^2(g_1, g_2)$$

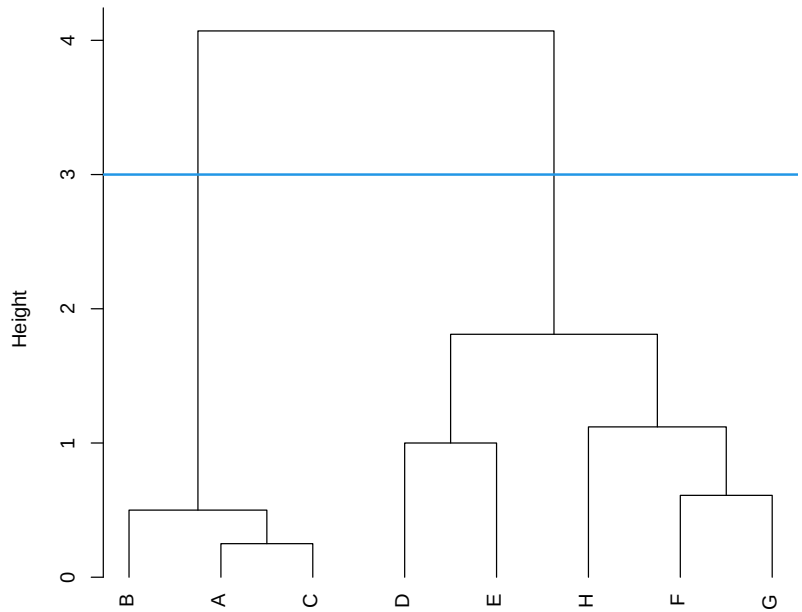
où $\mu_k = \sum_{i \in C_k} w_i$ est le poids de la classe k.

$\Rightarrow$ Pas d'inversion dans l'arbre.

$\Rightarrow$ Favorise l'agrégation de classes de faibles poids et casse l'effet chaine du lien minimum.



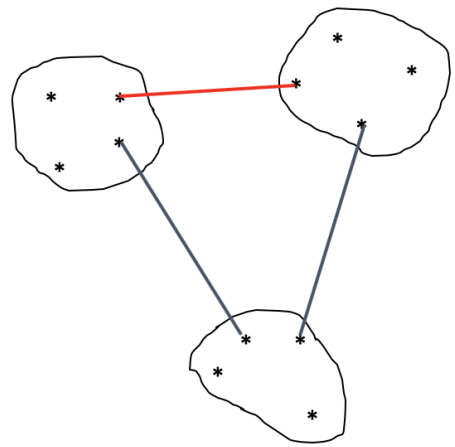
Couper un arbre pour obtenir une partition.



- En définissant un niveau de coupure, on construit une partition.
- Ici $P_2 = (\{A, B, C\}, \{D, E, F, G, H\}$.

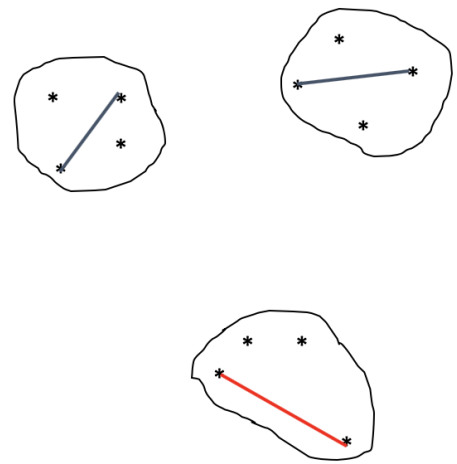
Qualité des partitions.

- L'isolation des classes d'une partition peut se mesurer par le plus petit lien minimum.



Ce critère est maximisé par l'algorithme du lien minimum qui construira des classes isolées mais pas nécessairement cohérentes et qui peuvent être désiquilibrées.

- La cohésion des classes d'une partition peut se mesurer par le plus grand diamètre.

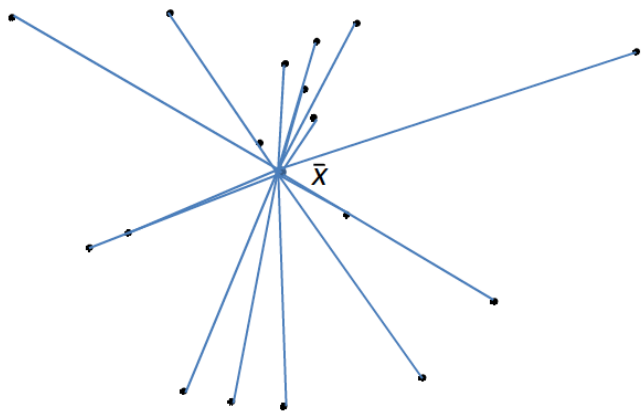


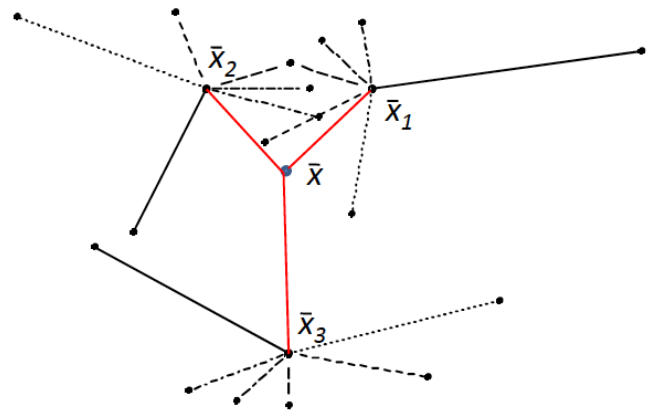
Ce critère est minimisé (approximativement) par l'algorithme du lien maximum qui construira des classes cohérentes mais pas nécessairement isolées et qui sont sensibles aux valeurs abérrantes.

- L'isolation et la cohésion avec un seul critère.

Pour des poids w_i quelconques :

$$\underbrace{\sum_{i=1}^{n} w_i d^2(x_{i.}, \bar{x})}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K} \sum_{i \in C_k} w_i d^2(x_{i.}, \bar{x}_k)}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K} \mu_k d^2(\bar{x}_k, \bar{x})}_{\text{Inertie inter}}$$





Minimiser l'inertie intra (hétérogénéité des classes) $\Leftrightarrow$ Maximiser l'inertie inter (séparation des classes)

La qualité d'une partition est alors mesurée par :

$$0 \leq \frac{\text{Inertie inter}}{\text{Inertie total}} \leq 1$$

Interprétation de ce critère :

- Part de l'inertie totale expliquée par la partition.
- $\frac{\text{Inertie inter}}{\text{Inertie total}} = 0$
  - $\Rightarrow$ Les variables ont les mêmes moyennes dans toutes les classes (moyenne globale).
  - $\Rightarrow$ La partition ne permet pas de classifier.
- $\frac{\text{Inertie inter}}{\text{Inertie total}} = 1$
  - $\Rightarrow$ Les individus d'une même classe sont identiques.
  - $\Rightarrow$ La partition est idéale pour classifier.

Attention. Ce critère ne peut pas être jugé en absolue car il dépends du nombre d'individus et du nombre de classes.

En effet il vaut :

- 1 pour la partition en n classes (1 individu par classe),
- 0 pour la partition en 1 classes (contenant tous les individus).

Il augmente donc avec le nombre de classes. Il permet de comparer des partitions ayant le même nombre de classes.

Ce critère est maximisé (approximativement) par l'algorithme de Ward qui construira des classes isolées, cohérentes et équilibrées.

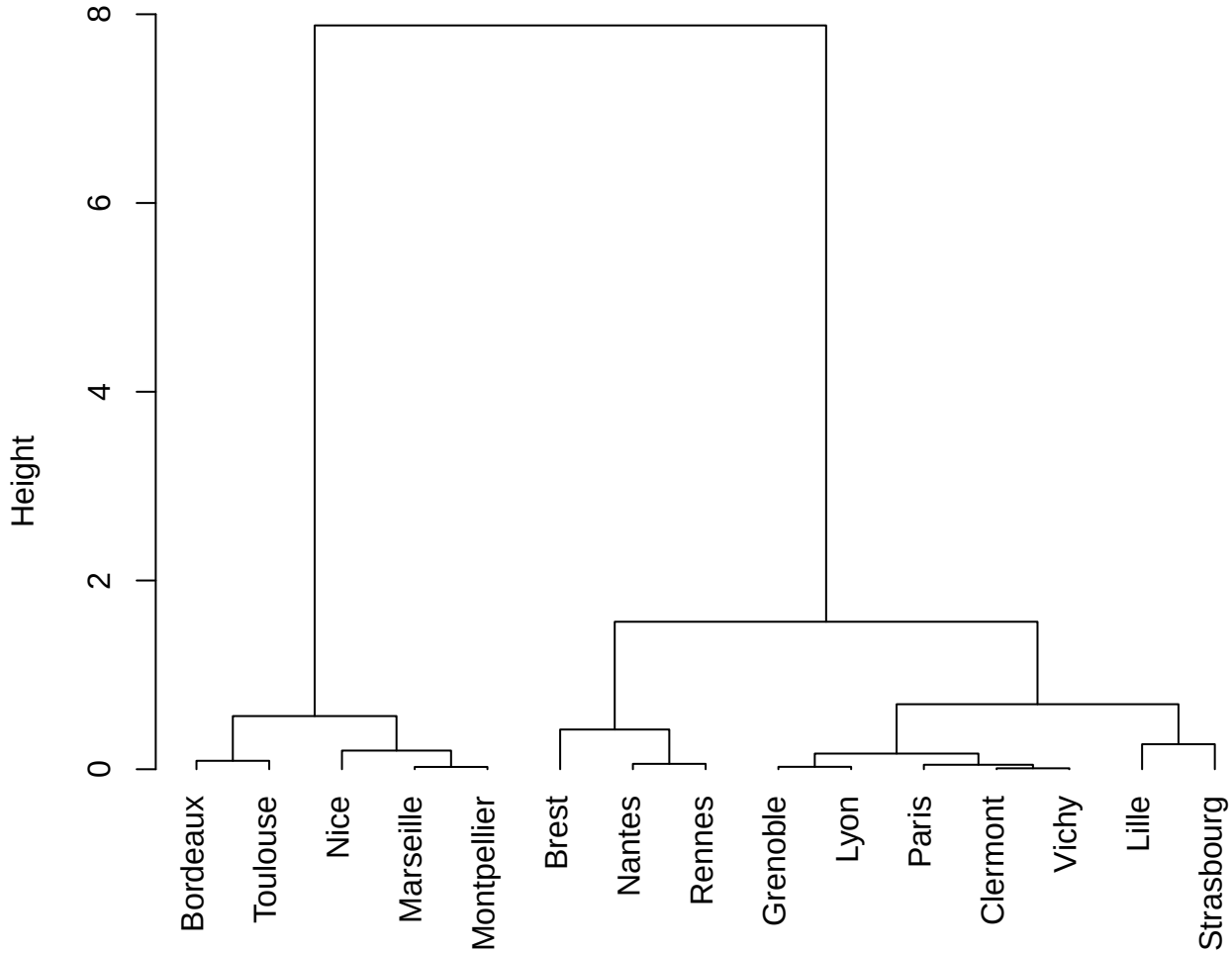
Exemple des données températures.

- 15 individus : villes de France,
- 12 variables : températures mensuelles moyennes (sur 30 ans).

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	5.6	6.6	10.3	12.8	16	19	21	21	19	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	12	14	16	16	15	12.0	9.0	7.0
Clermont	2.6	3.7	7.5	10.3	14	17	19	19	16	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	10.6	14	18	20	20	17	11.4	6.5	2.3
Lille	2.4	2.9	6.0	8.9	12	15	17	17	15	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	10.9	15	18	21	20	17	11.4	6.7	3.1
Marseille	5.5	6.6	10.0	13.0	17	21	23	23	20	15.0	10.2	6.9
Montpellier	5.6	6.7	9.9	12.8	16	20	23	22	19	14.6	10.0	6.5
Nantes	5.0	5.3	8.4	10.8	14	17	19	19	16	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13.3	17	20	23	22	20	16.0	11.5	8.2
Paris	3.4	4.1	7.6	10.7	14	18	19	19	16	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10.1	13	16	18	18	16	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14	17	19	18	15	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	11.6	15	19	21	21	18	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	14	17	19	19	16	11.0	6.6	3.4

Quelles villes ont des profils météorologiques similaires ?

Algorithme de Ward appiqué aux données températures standardisées.



Interprétation des hauteurs de l'arbre ?

Hauteur d'une classe dans l'arbre de Ward.

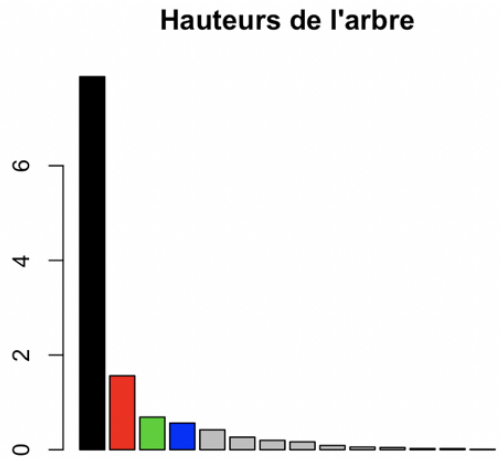
La hauteur de l'agrégation de deux classes A et B est :

$$\underbrace{\frac{\mu_A \mu_B}{\mu_A + \mu_B} d^2(g_A, g_B)}_{\text{Mesure de Ward}} = \underbrace{I(A \cup B) - I(A) - I(B)}_{\text{perte d'inertie expliquée}} = \underbrace{\mu_1 d^2(g_1, g) + \mu_2 d^2(g_2, g)}_{\text{inertie inter}}$$

Pour les données températures :

Pertes d'inertie lors du passage de

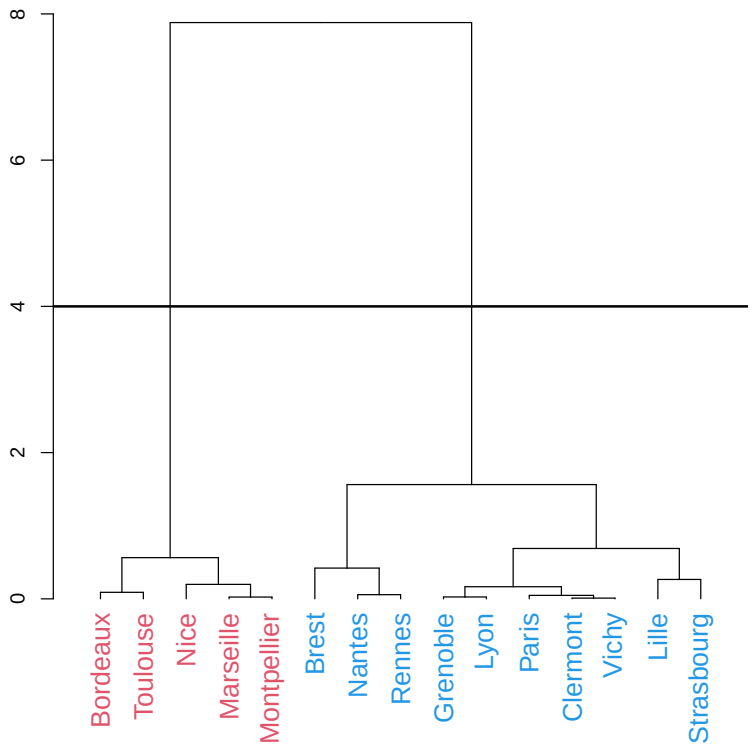
15 classes à 14 classes : 0.01
14 classes à 13 classes : 0.02
13 classes à 12 classes : 0.03
12 classes à 11 classes : 0.05
11 classes à 10 classes : 0.06
10 classes à 9 classes : 0.09
9 classes à 8 classes : 0.17
8 classes à 7 classes : 0.19
7 classes à 6 classes : 0.26
6 classes à 5 classes : 0.42
5 classes à 4 classes : 0.56
4 classes à 3 classes : 0.69
3 classes à 2 classes : 1.56
2 classes à 1 classes : 7.88



Grosse perte si on passe de 2 classes à 1 seule donc on préfère en garder 2.

Somme des pertes d'inertie = 12 (inertie totale).

Couper l'arbre pour obtenir une partition.

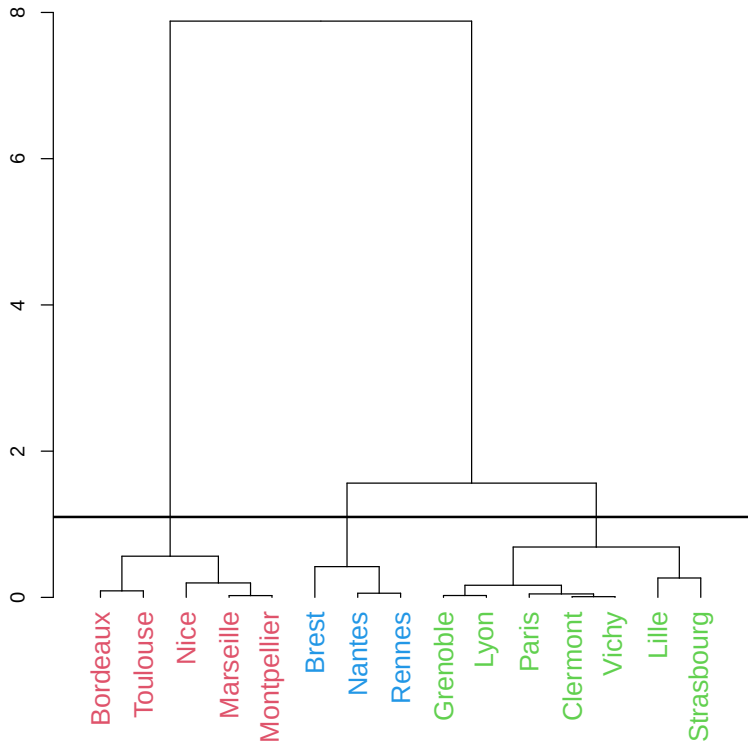


Découpage en 2 classes :

$$\frac{\text{Inertie inter}}{\text{Inertie total}} = \frac{7.88}{12} = 66\%$$

$\Rightarrow$ 66% d'inertie expliquée par la partition en 2 classes.

Séparer les villes froides en deux groupes.



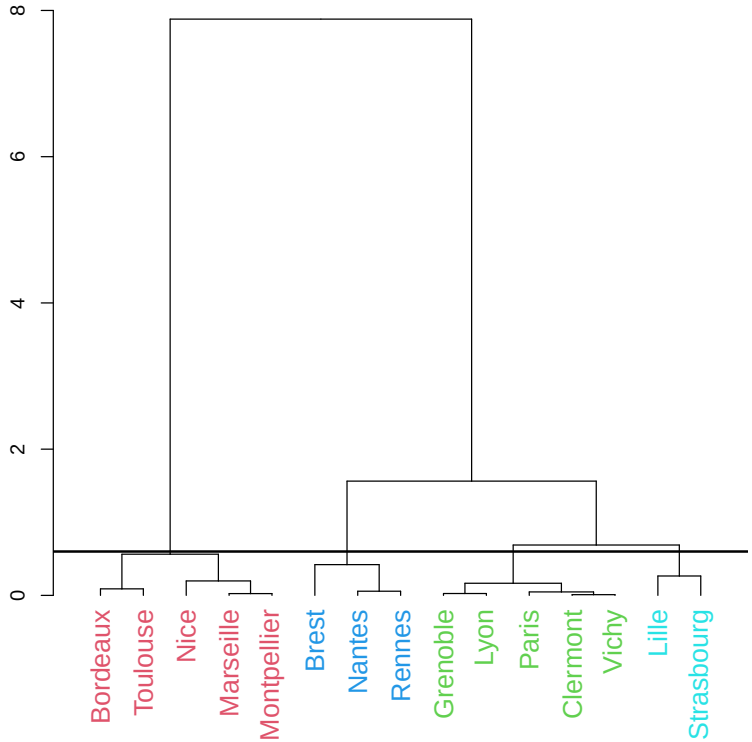
Découpage en 3 classes :

$$\frac{\text{Perte inertie}}{\text{Inertie total}} = \frac{1.56}{12} = 13\%$$

Gain de 13% d'inertie en passant de 2 à 3 classes (perte en passant de 3 à 2).

$\Rightarrow$ 66% + 13% = 79 % d'inertie expliquée par la partition en 3 classes.

Séparer les villes froides de l'Est en deux groupes.



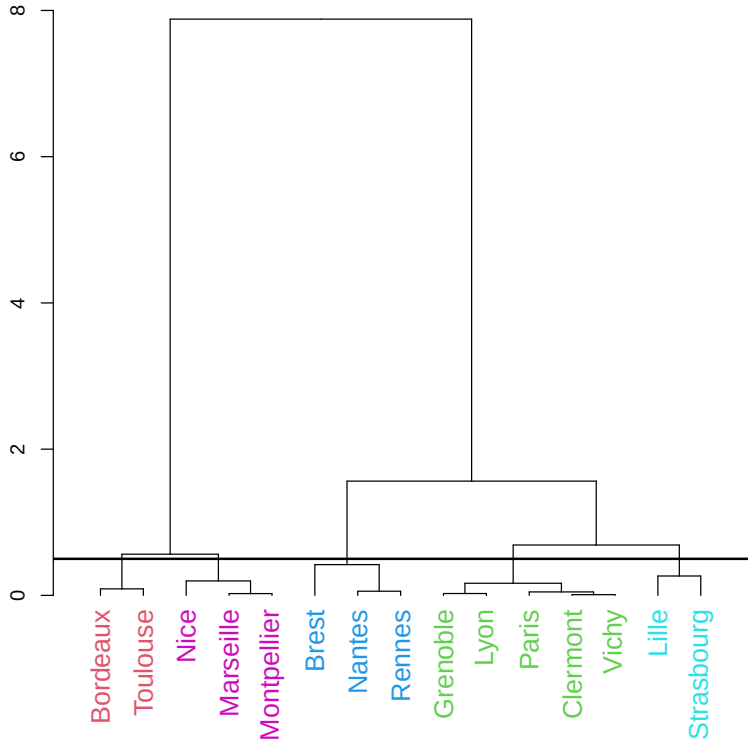
Découpage en 4 classes :

$$\frac{\text{Perte inertie}}{\text{Inertie total}} = \frac{0.69}{12} = 5.75\%$$

Gain de 5.7% d'inertie en passant de 3 à 4 classes.

$\Rightarrow$ 79% + 5.7% = 84.7 % d'inertie expliquée par la partition en 4 classes.

Séparer les villes chaudes en deux groupes.



Découpage en 5 classes :

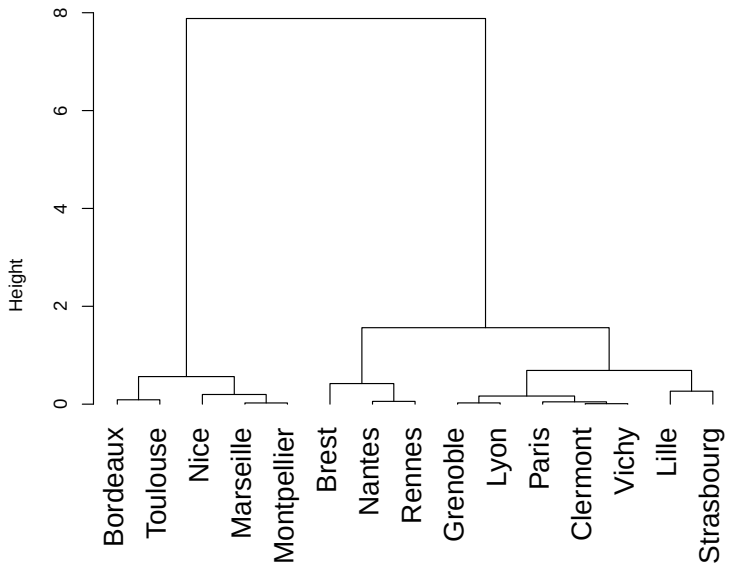
$$\frac{\text{Perte inertie}}{\text{Inertie total}} = \frac{0.56}{12} = 4.7\%$$

Gain de 4.7% d'inertie en passant de 4 à 5 classes (perte en passant de 4 à 3).

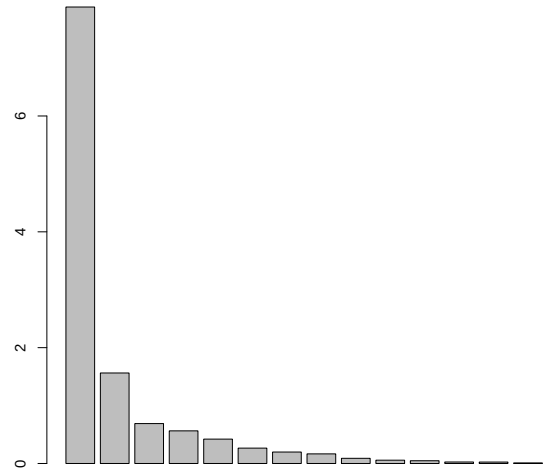
$\Rightarrow$ Gain proche de celui du passage de 3 à 4 classes.

Détermination du nombre de classes.

A partir de l'arbre :



A partir du diagramme des hauteurs :



A partir de l'interprétabilité des classes.

Propriétés de l'algorithme de Ward

- La partition construite à chaque étape maximise l'inertie inter parmi les partitions résultant de l'agrégation de deux classes de la partition précédente.
- La somme des hauteurs du dendrogramme de Ward est l'inertie totale.
- La somme des $K-1$ plus grandes hauteurs est l'inertie inter de la partition en K classes de l'abre.
- La complexité de l'algorithme : quadratique par rapport au nombre d'individus.

$\Rightarrow$ Problème pour les jeux de données ayant un très grand nombre d'individus.

Outline

Algorithme de partitionnement des K-means

Une bonne partition en K classes possède des classes

- homogènes : les individus dans une même classe se ressemblent,
- séparées : les individus de deux classes différentes ne se ressemblent pas.

On a vu que (pour des données quantitatives) :

$$\underbrace{\sum_{i=1}^{n} w_i d^2(x_{i.}, \bar{x})}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K} \sum_{i \in C_k} w_i d^2(x_{i.}, \bar{x}_k)}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K} \mu_k d^2(\bar{x}_k, \bar{x})}_{\text{Inertie inter}}$$

On veut donc :

- minimiser l'inertie intra (l'hétérogénéité des classes),
- maximiser l'inertie inter (la séparation des classes).

$\Rightarrow$ Maximiser la part de l'inertie totale expliquée par la partition (Inertie inter/Inertie total)

Partition optimale.

- Choisir parmi toutes les partitions en K classes celle de plus grande inertie inter.
- Problème : nombre de partitions en K classes des n individus $\sim \frac{K^n}{K!}$.

$\Rightarrow$ Enumération complète impossible.

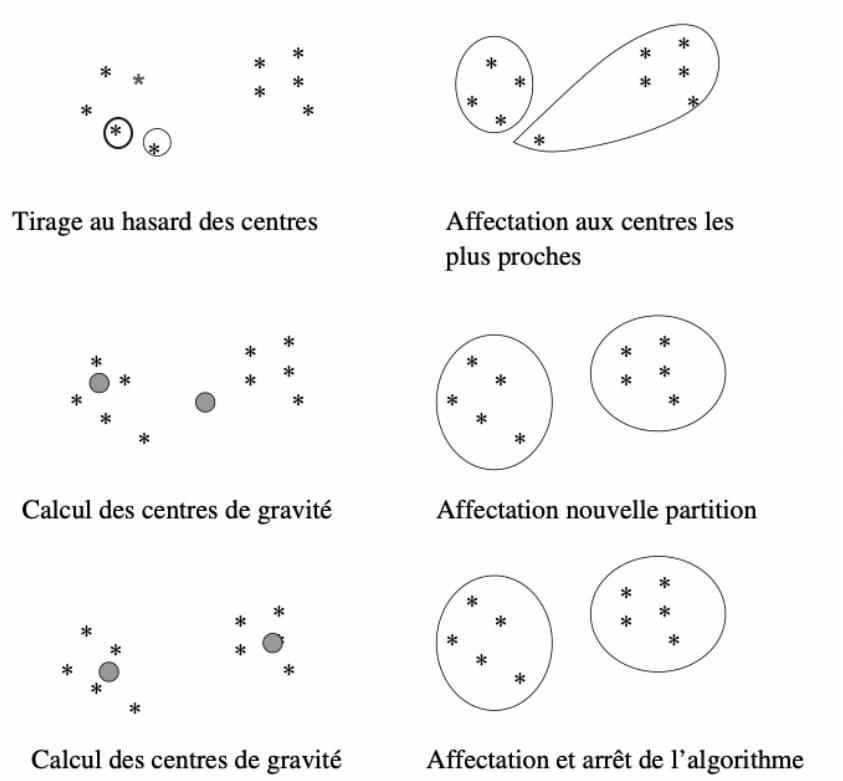
Partition localement optimale.

- Heuristique du type :
  - On part d'une solution réalisable c'est à dire d'une partition P_K^0,
  - A l'étape m+1, on cherche une partition $P_K^{m+1} = g(P_K^m)$ telle que inertie_inter$(P_K^{m+1}) \geq$ inertie_inter(P_K^m).
  - Arrêt lorsqu'aucun individu ne change de classe entre deux itérations.

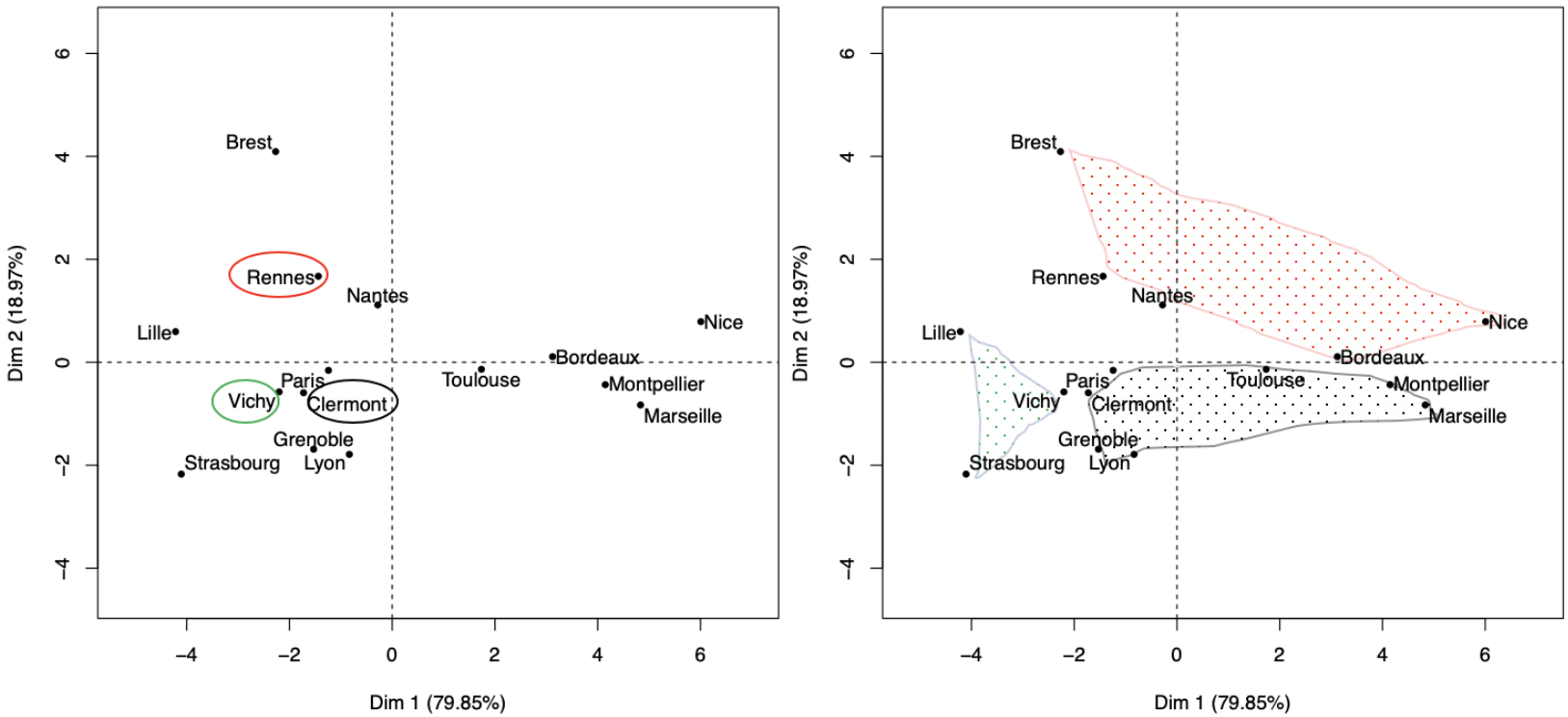
$\Rightarrow$ Méthode de partitionnement des K-means.

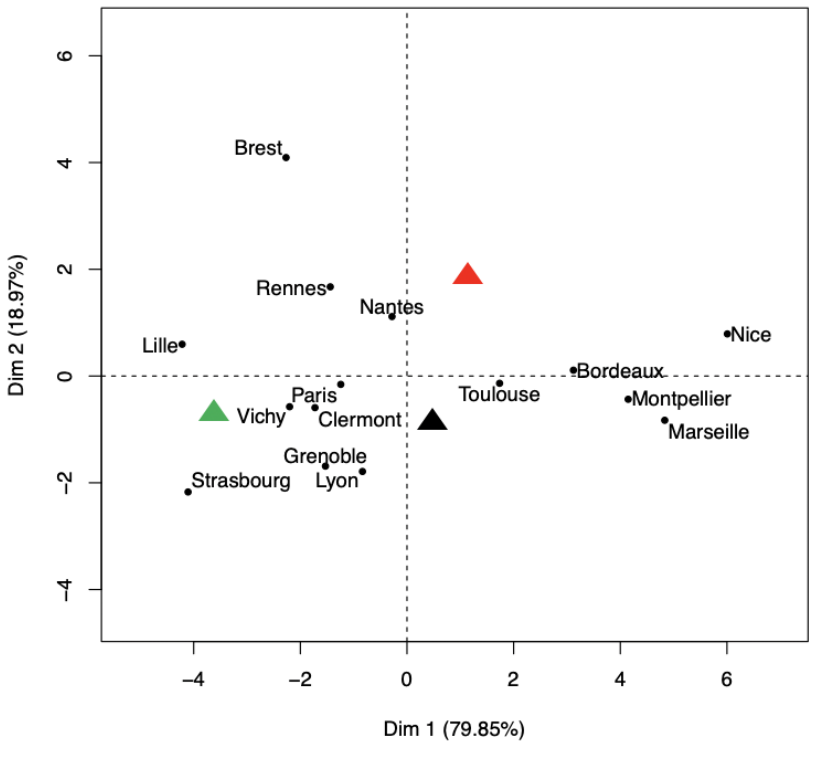
La méthode des K-means.

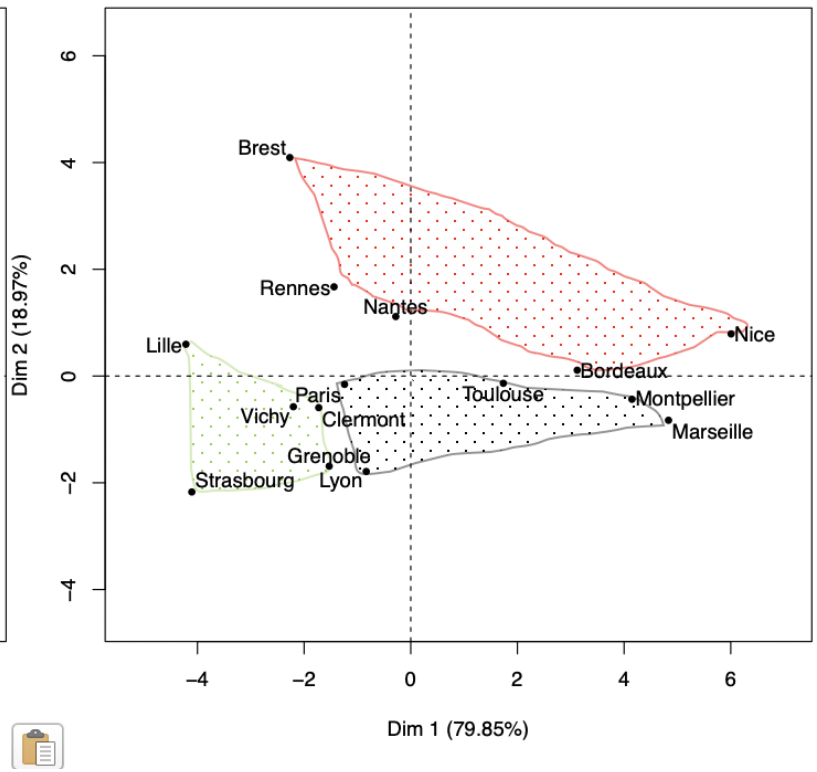
- Tirage aléatoire de K centres de classes (K individus parmi n).
- Répéter jusqu'à la convergence :
  - une étape d'affectation : chaque individu est affecté à la classe dont centre de gravité est le plus proche.
  - une étape de représentation : les centres de gravité des classes sont calculés.

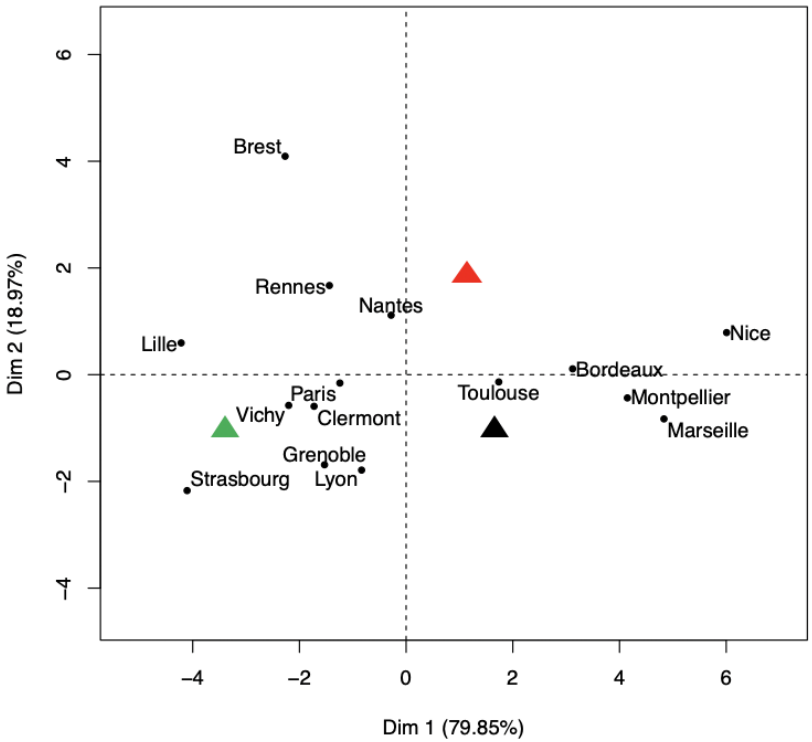


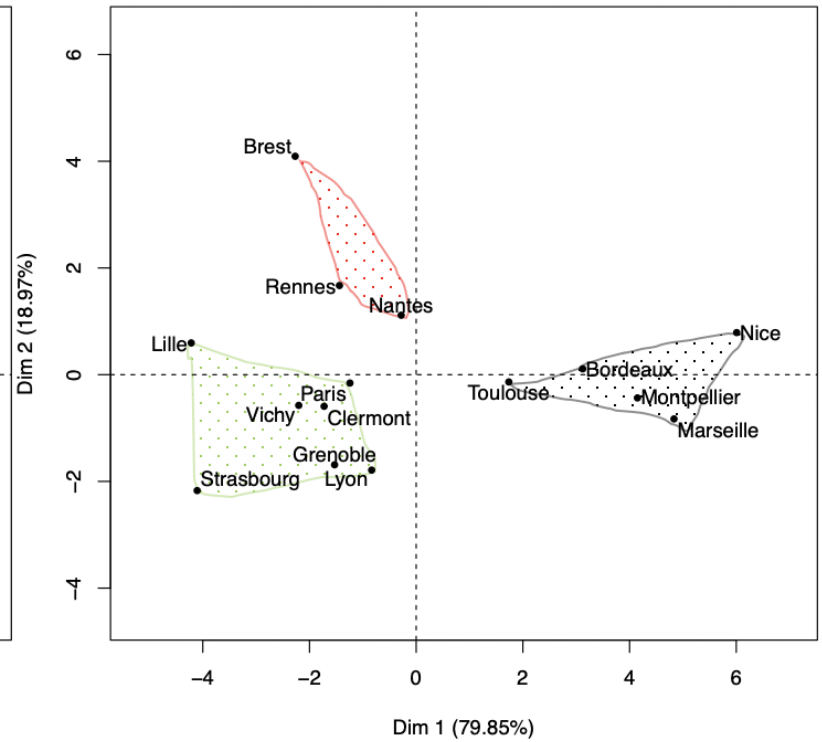
Exemple des données températures.

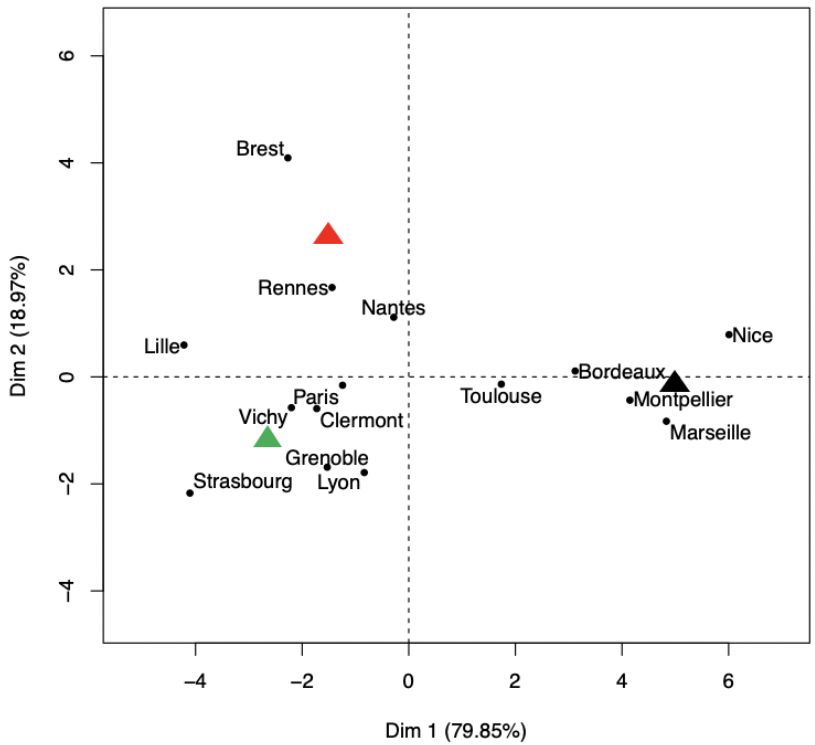


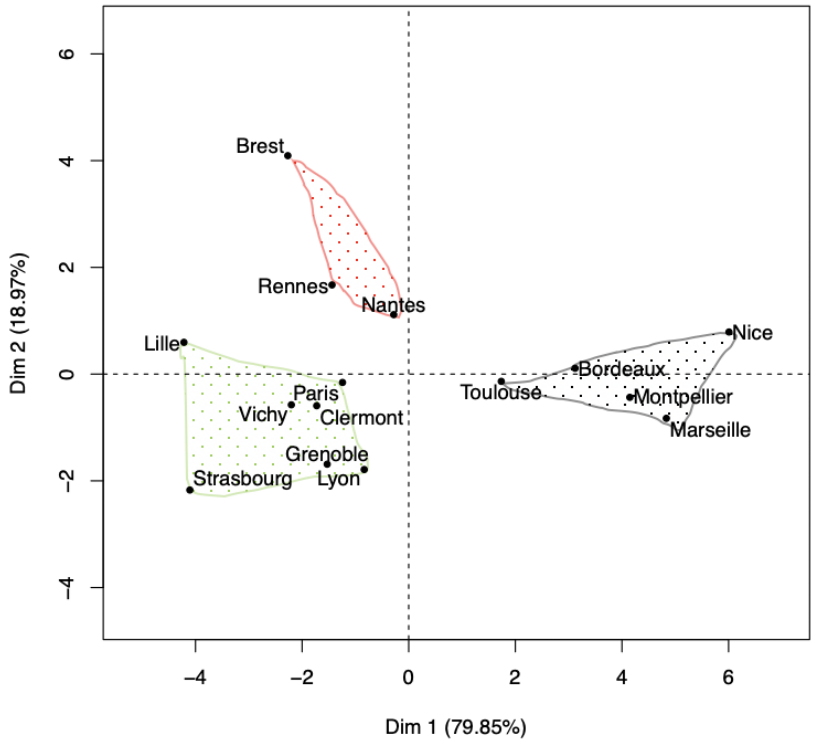

Brest
Rennes
Nantes
Lille
Nice
Bordeaux
Paris
Toulouse
Montpellier
Vichy
Clermont
Marseille
Grenoble
Strasbourg
Lyon
Dim 2 (18.97%)
Dim 1 (79.85%)
6
4
2
0
−2
−4
−4
−2
0
2
4
6


Brest
Rennes
Nantes
Lille
Nice
Bordeaux
Paris
Toulouse
Montpellier
Vichy
Clermont
Marseille
Grenoble
Strasbourg
Lyon
Dim 2 (18.97%)
Dim 1 (79.85%)
6
4
2
0
−2
−4
−4
−2
0
2
4
6


Brest
Rennes
Nantes
Lille
Nice
Bordeaux
Paris
Toulouse
Montpellier
Vichy
Clermont
Marseille
Grenoble
Strasbourg
Lyon
Dim 2 (18.97%)
Dim 1 (79.85%)


Brest
Rennes
Nantes
Lille
Nice
Bordeaux
Toulouse
Montpellier
Paris
Vichy
Clermont
Marseille
Grenoble
Strasbourg
Lyon
Dim 2 (18.97%)
Dim 1 (79.85%)





Cette partition en 3 classes est-elle différente de celle issue de la CAH de Ward ?

Propriétés de l'algorithme des K-means

- L'algorithme converge vers une partition réalisant un maximum local de l'inertie inter-classe (et donc du pourcentage d'inertie expliquée).
- La partition finale dépend de la partition initiale : si on relance l'algorithme avec d'autres centres initiaux, la partition finale peut être différente.

$\Rightarrow$ En pratique,

  - on lance N fois l'algorithme avec des initialisations aléatoires différentes.
  - on retient parmi les N partitions finales, celle ayant le pourcentage d'inertie expliquée le plus grand.

- La complexité de l'algorithme : linéaire par rapport au nombre d'individus.

$\Rightarrow$ Capable de traiter des données avec un très grand nombre d'individus.

Outline

Pourquoi faut-il parfois standardiser les données ?

Le jeu de donnée indique la quantité de protéines consommée dans 9 types d'aliments dans 25 (anciens) pays européens : 25 individus (les 10 premiers ci-dessous) et 9 variables quantitatives.

	Red.Meat	White.Meat	Eggs	Milk	Fish	Cereals	Starchy.Foods	Nuts	Fruite.veg.
Alban	10.1	1.4	0.5	8.9	0.2	42	0.6	5.5	1.7
Aust	8.9	14.0	4.3	19.9	2.1	28	3.6	1.3	4.3
Belg	13.5	9.3	4.1	17.5	4.5	27	5.7	2.1	4.0
Bulg	7.8	6.0	1.6	8.3	1.2	57	1.1	3.7	4.2
Czech	9.7	11.4	2.8	12.5	2.0	34	5.0	1.1	4.0
Den	10.6	10.8	3.7	25.0	9.9	22	4.8	0.7	2.4
E_Ger	8.4	11.6	3.7	11.1	5.4	25	6.5	0.8	3.6
Finl	9.5	4.9	2.7	33.7	5.8	26	5.1	1.0	1.4
Fr	18.0	9.9	3.3	19.5	5.7	28	4.8	2.4	6.5
Greece	10.2	3.0	2.8	17.6	5.9	42	2.2	7.8	6.5

On applique les K-means à ces données pour illustrer deux aspects méthodologiques :

- Pourquoi faut-il parfois standardiser les données ?
- Comment interpréter les classes à l'aide d'une ACP ?

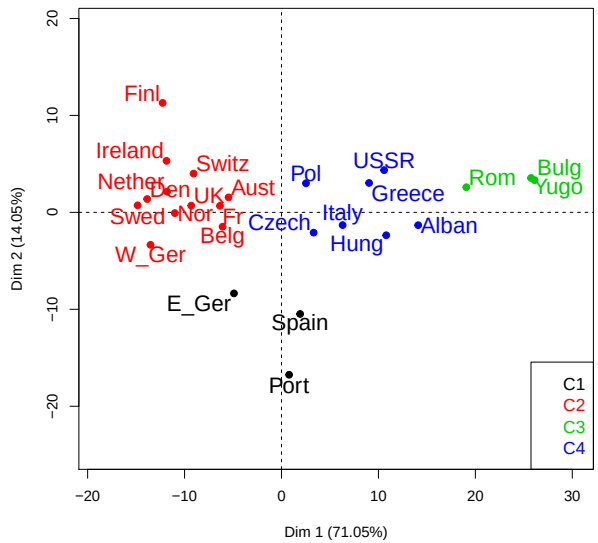
Données brutes : partition en 4 classes des K-means après $N = 5$ initialisations.

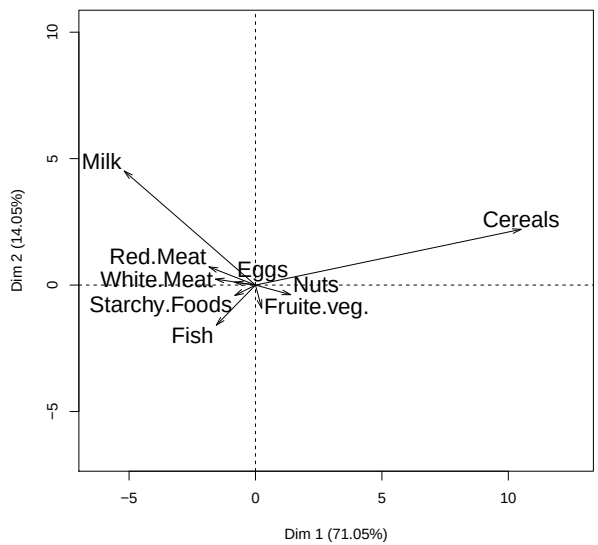
	P4
Alban	4
Aust	2
Belg	2
Bulg	3
Czech	4
Den	2
E_Ger	1
Finl	2
Fr	2
Greece	4
Hung	4
Ireland	2
Italy	4
Nether	2
Nor	2
Pol	4
Port	1
Rom	3
Spain	1
Swed	2
Switz	2
UK	2
USSR	4
W_Ger	2
Yugo	3

	écart-type
Red.Meat	3.4
White.Meat	3.7
Eggs	1.1
Milk	7.1
Fish	3.4
Cereals	11.0
Starchy.Foods	1.6
Nuts	2.0
Fruite.veg.	1.8

Interprétation via l'ACP

non normée.

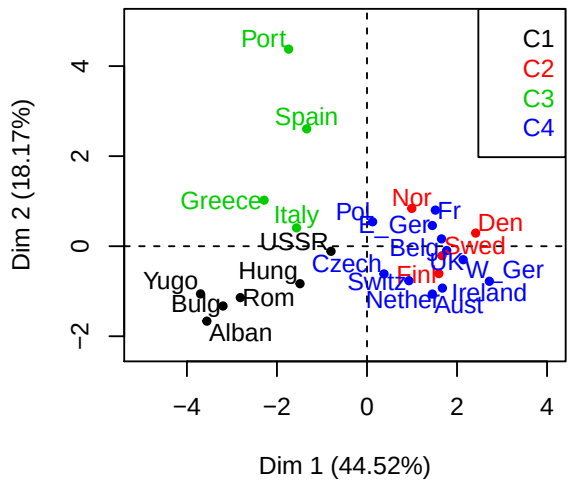


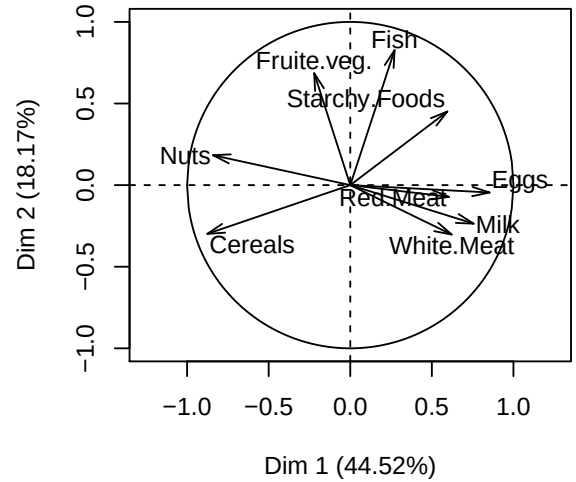


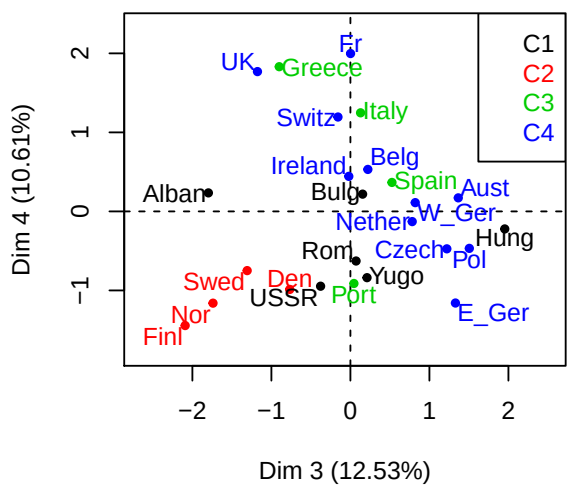
Données standardisées : partition en 4 classes des K-means après $N = 5$ initialisations.

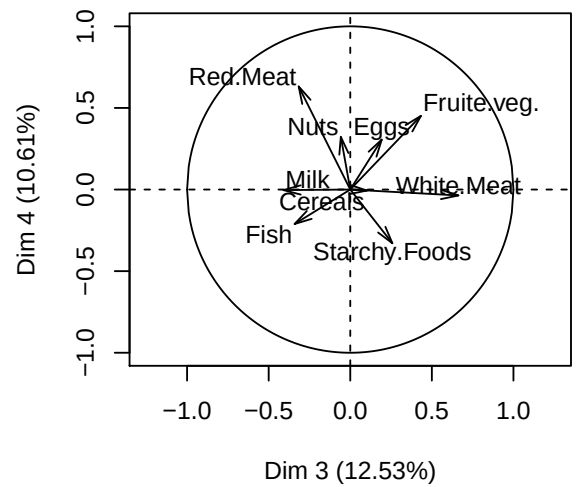
	écart-type
Red.Meat	1
White.Meat	1
Eggs	1
Milk	1
Fish	1
Cereals	1
Starchy.Foods	1
Nuts	1
Fruite.veg.	1

Interprétation via l'ACP normée.









Consolidation d'une partition obtenue par CAH.

La partition obtenue par CAH n'est pas optimale et peut être améliorée, consolidée, par les K-means.

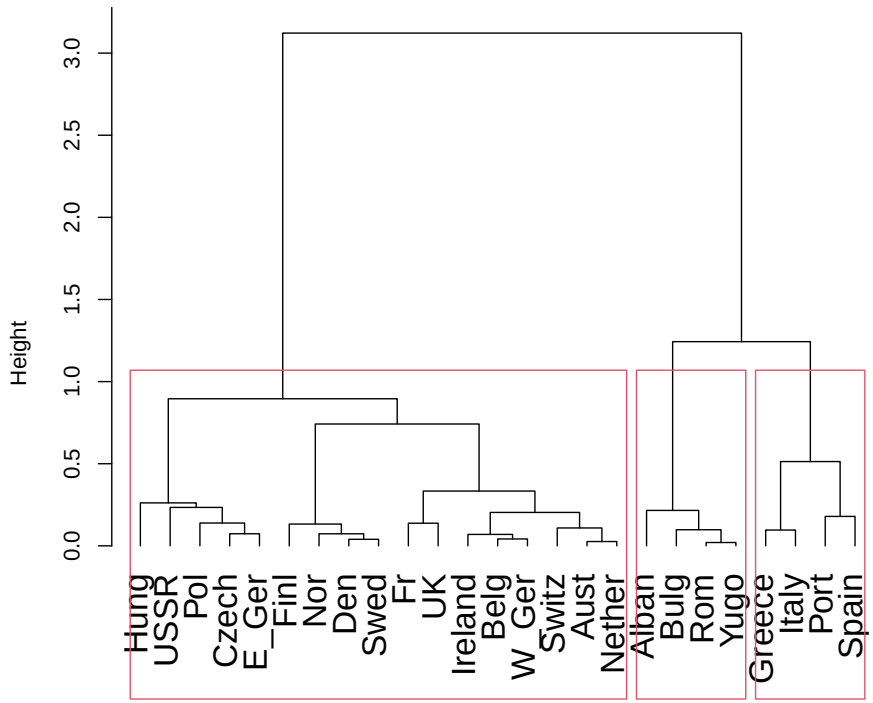
Algorithme de consolidation :

- la partition obtenue par CAH est utilisée comme initialisation de l'algorithme de partitionnement,
- quelques étapes de K-means sont itérées

$\Rightarrow$ amélioration de la partition (souvent non décisive).

Avantage : consolidation de la partition.
Inconvénient : perte de l'info de hiérarchie

Exemple des données protéines standardisées.



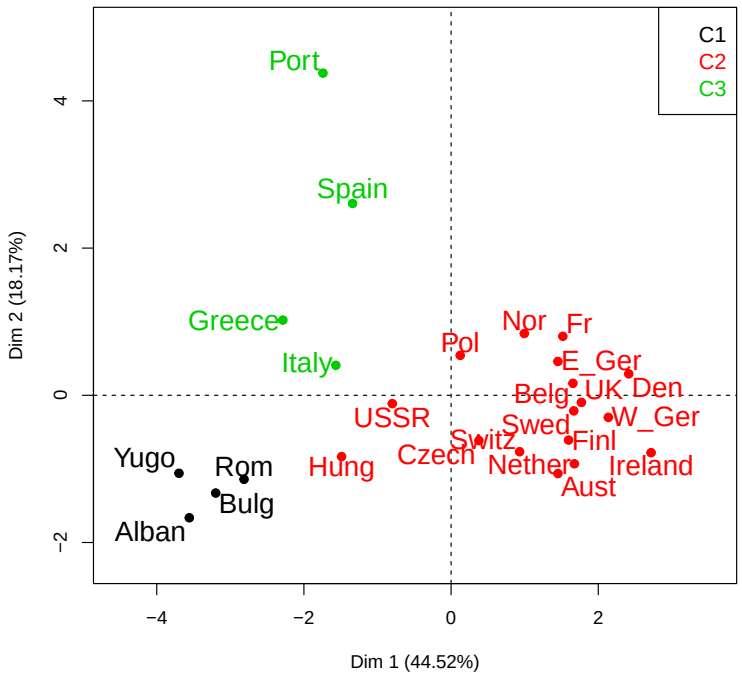
Le dendrogramme suggère de couper l'arbre en 3 ou 5 classes.

On regarde la partition en 3 classes :

$$\frac{\text{Inertie inter}}{\text{Inertie total}} = \frac{3.1 + 1.2}{9} = 0.485$$

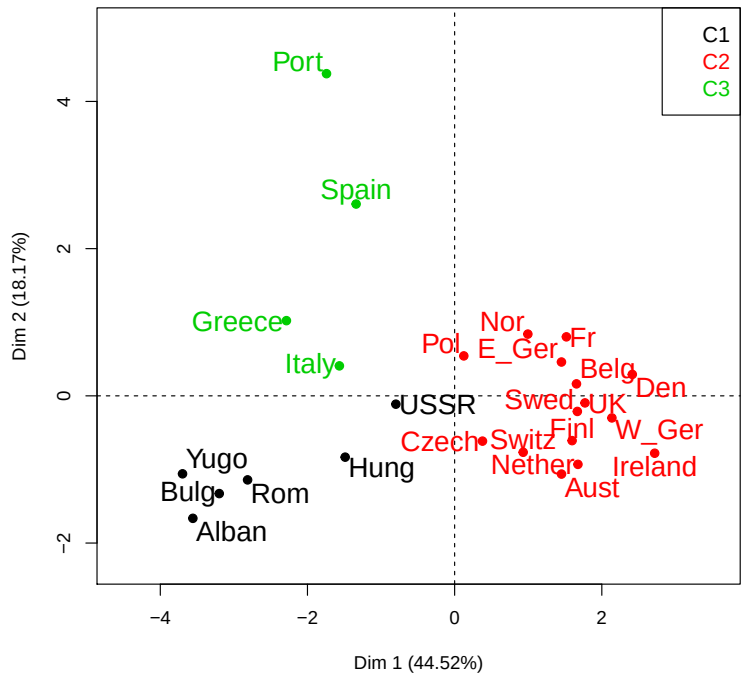
La partition en 3 classes de Ward explique 48.5% de l'inertie.

Avant consolidation.



- 48.5% de l'inertie expliquée par la partition.

Après consolidation



- 50.9% de l'inertie expliquée par la partition.

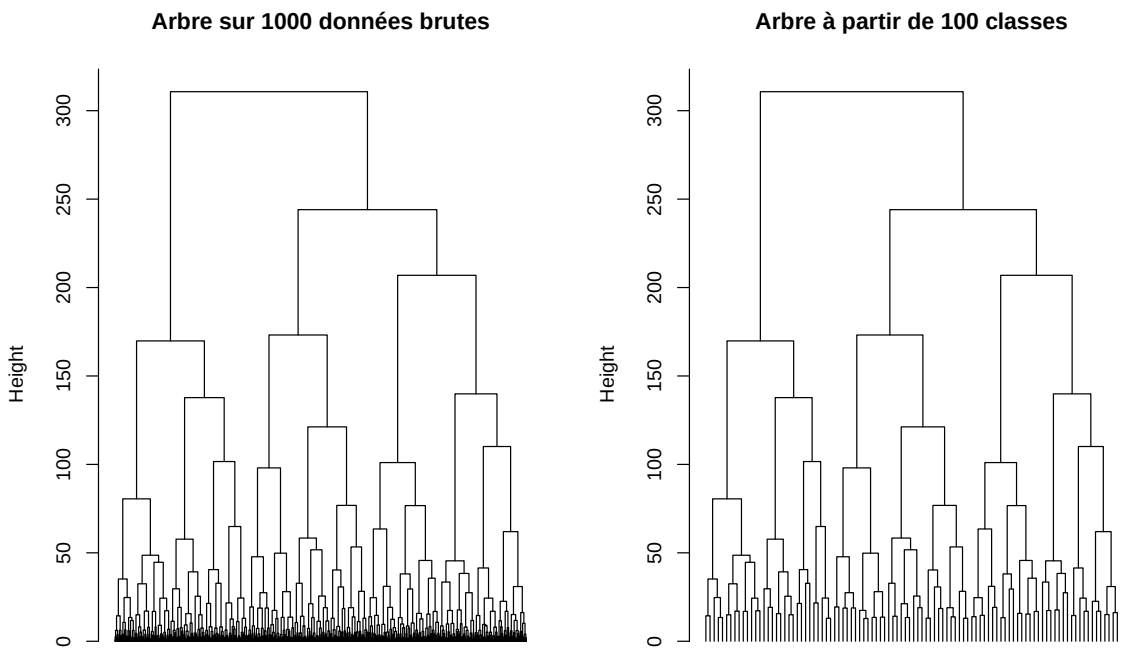
$\Rightarrow$ légère amélioration (2 individus ont changés de classe).

CAH avec beaucoup d'individus

Si beaucoup d'individus : algorithme de CAH trop long

- Faire une partition (par K-means) en une centaine de classes
- Construire la CAH à partir des classes (utiliser l'effectif des classes dans le calcul)

$\Rightarrow$ Obtention du « haut » de l'arbre de la CAH.



CAH ou K-means avec beaucoup de variables

Si beaucoup de variables : réduire la dimension.

- Faire une ACP et conserver les q premières composantes principales.
  - Si on conserve toutes les composantes principales de l'ACP normée (resp. non normée), on trouve la même classification qu'avec les données standardisées (resp. brutes).
- Faire un clustering de variables en q classes et conserver les q variables synthétiques (1ères CP des classes).

$\Rightarrow$ Difficulté du choix de q.

CAH ou K-means sur données qualitatives ou mixtes

- Se ramener à des variables quantitatives :
  - faire une ACM (ou une ACP mixte) et conserver toutes les composantes principales (ou les q les premières) ,
  - faire un clustering de variables (qui gère les données qualitatives et mixtes) et conserver les q variables synthétiques des classes.

$\Rightarrow$ Difficulté du choix de q.

- Utiliser des mesures adaptées aux données qualitatives ou mixte : indice de similarité, indice de Jaccard, etc. puis appliquer un algorithme de CAH sur cette matrice de similarité (dissimilarité, distance).

$\Rightarrow$ Que signifie Ward si les distances ne sont pas Euclidienne ?

Outline

Interprétation des classes d'individus

On peut interpréter les classes d'une partition à partir :

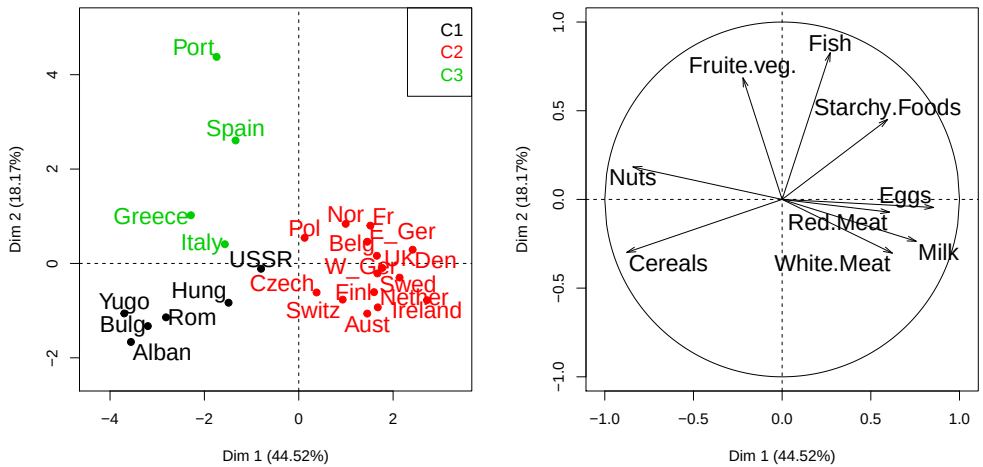
- des variables actives : variables utilisées dans le processus de clustering,
- de variables illustratives : utilisées uniquement pour la description des classes.

Ces variables peuvent être quantitatives ou qualitatives.

En pratique, on caractérisera souvent les classes (ou des groupes d'invididus) par :

- les modalités des variables qualitatives : une modalité est-elle plus fréquente dans la classe, la classe contient-elle tous les individus possédant cette modalité, etc ?
- les variables quantitatives : la moyenne dans la classe est-elle différent de la moyenne chez tous les individus, etc ?

Exemple de la partition en 3 classes après consolidation des données protéines.



Questions :

- Quelles variables caractérisent le mieux la partition ?
- Comment caractériser les villes d'une classe en particulier ?

Quelles variables caractérisent le mieux la partition ?

- Pour chaque variable quantitative :
  - calculer le rapport de corrélation (η^2) entre la partition (variable qualitative à K modalités) et les variables quantitatives.
  - faire le test de Fisher de l'effet de la partition sur la variable quantitative (modèle d'analyse de la variance).
  - Trier les variables par probabilité critique (p-valeur) croissante.

	Eta2	P-value
Nuts	0.79	3.0e-08
Cereals	0.75	2.1e-07
Eggs	0.58	8.1e-05
Fruite.veg.	0.54	1.7e-04
Milk	0.53	2.3e-04
White.Meat	0.43	1.9e-03
Fish	0.40	4.0e-03
Red.Meat	0.33	1.3e-02

- Pour chaque variable qualitative : idem avec un test du χ^2 d'indépendance entre la partition et la variable qualitative.

TABLE – variable qualitative

	zone
Alban	east
Aust	west
Belg	west
Bulg	east
Czech	east
Den	north
E_Ger	east
Finl	north
Fr	west
Greece	south
Hung	east
Ireland	west
Italy	south
Nether	west
Nor	north
Pol	east
Port	south
Rom	east
Spain	south
Swed	north
Switz	west
UK	west
USSR	east
W_Ger	west
Yugo	east

TABLE – Test du Chi2

	p.value	df
zone	1e-06	6

Quelles variables quantitatives caractérisent une classe en particulier ?

Quelle variable quantitative X caractérise le mieux la classe C_k ?

Hypothèse testée : tirage au hasard de n_k valeurs de X parmi n.

Quelles valeurs peut prendre $\bar{x}_k$? (i.e. quelle est la loi de $\bar{X}_k$ sous H_0 ?) :

$$\mathbb{E}(\bar{X}_k) = \bar{x} \qquad \mathbb{V}(\bar{X}_k) = \frac{s^2}{n_k} \frac{(n - n_k)}{n - 1}$$

$$\mathcal{L}(\bar{X}_k) = \mathcal{N} \quad \text{car } \bar{X}_k \text{ est une moyenne}$$

$\Rightarrow$ Sous H_0 :

$$\text{valeur-test} = \frac{\bar{x}_k - \bar{x}}{\sqrt{\frac{s^2}{n_k} \frac{(n-n_k)}{n-1}}} \underset{approx}{\sim} \mathcal{N}(0, 1)$$

- Si |valeur-test| ≥ 1.96 alors la variable X caractérise la classe C_k (H_0 rejetée avec un risque inférieur à 5%)
- X caractérise d'autant mieux la classe C_k que |valeur-test| grande

$\Rightarrow$ trier les variables par valeur-test décroissante

TABLE – Caractérisation de la classe 1

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Cereals	4.0	48.0	32.2	6.52	10.8	0.00005
Nuts	2.5	4.8	3.1	0.92	1.9	0.01269
Starchy.Foods	-2.1	3.0	4.3	1.91	1.6	0.03262
Red.Meat	-2.2	7.2	9.8	2.07	3.3	0.02641
Milk	-2.5	10.7	17.1	2.78	7.0	0.01102
Fish	-2.7	1.1	4.3	0.94	3.3	0.00757
Eggs	-3.3	1.6	2.9	0.74	1.1	0.00106

Quelles modalités des variables qualitatives caractérisent une classe en particulier ?

Quelle modalité m de la variables X caractérise le mieux la classe C_k ?

	C1	C2	C3	Total
east	$n_{mk} = 6$	3	0	$n_m = 9$
north	0	4	0	4
south	0	0	4	4
west	0	8	0	8
Total	$n_k = 6$	15	4	$n = 25$

Hypothèse testée : $\frac{n_{mk}}{n_k} = \frac{n_m}{n}$.

Sous H_0 :

$$\text{valeur-test} = \frac{\frac{n_{mk}}{n_k} - \frac{n_k}{n}}{\sqrt{\frac{(n-n_k)}{n-1}\frac{s^2}{n_k}}} \underset{approx}{\sim} \mathcal{N}(0,1)$$

avec $s^2 = \frac{n_m}{n}(1 - \frac{n_m}{n})$.

TABLE – Caractérisation de la classe 1

	Cla/Mod	Mod/Cla	Global	p.value	v.test
zone=east	67	100	36	0.00047	3.5

$Cla/Mod = \frac{6}{9} \simeq 0.67$, $Mod/Cla = \frac{6}{6} = 1$, $Global = \frac{9}{25} = 0.36$.

$\Rightarrow$ Modalité `east` caractérise la classe (sur-représentée).

TABLE – Caractérisation de la classe 2

	Cla/Mod	Mod/Cla	Global	p.value	v.test
zone=west	100	53	32	0.0059	2.8
zone=south	0	0	16	0.0166	-2.4

TABLE – Caractérisation de la classe 3

	Cla/Mod	Mod/Cla	Global	p.value	v.test
zone=south	100	100	16	7.9e-05	3.9