Classification ascendante hiérarchique (CAH)

François Husson

Laboratoire de mathématiques appliquées - Agrocampus Rennes

husson@agrocampus-ouest.fr

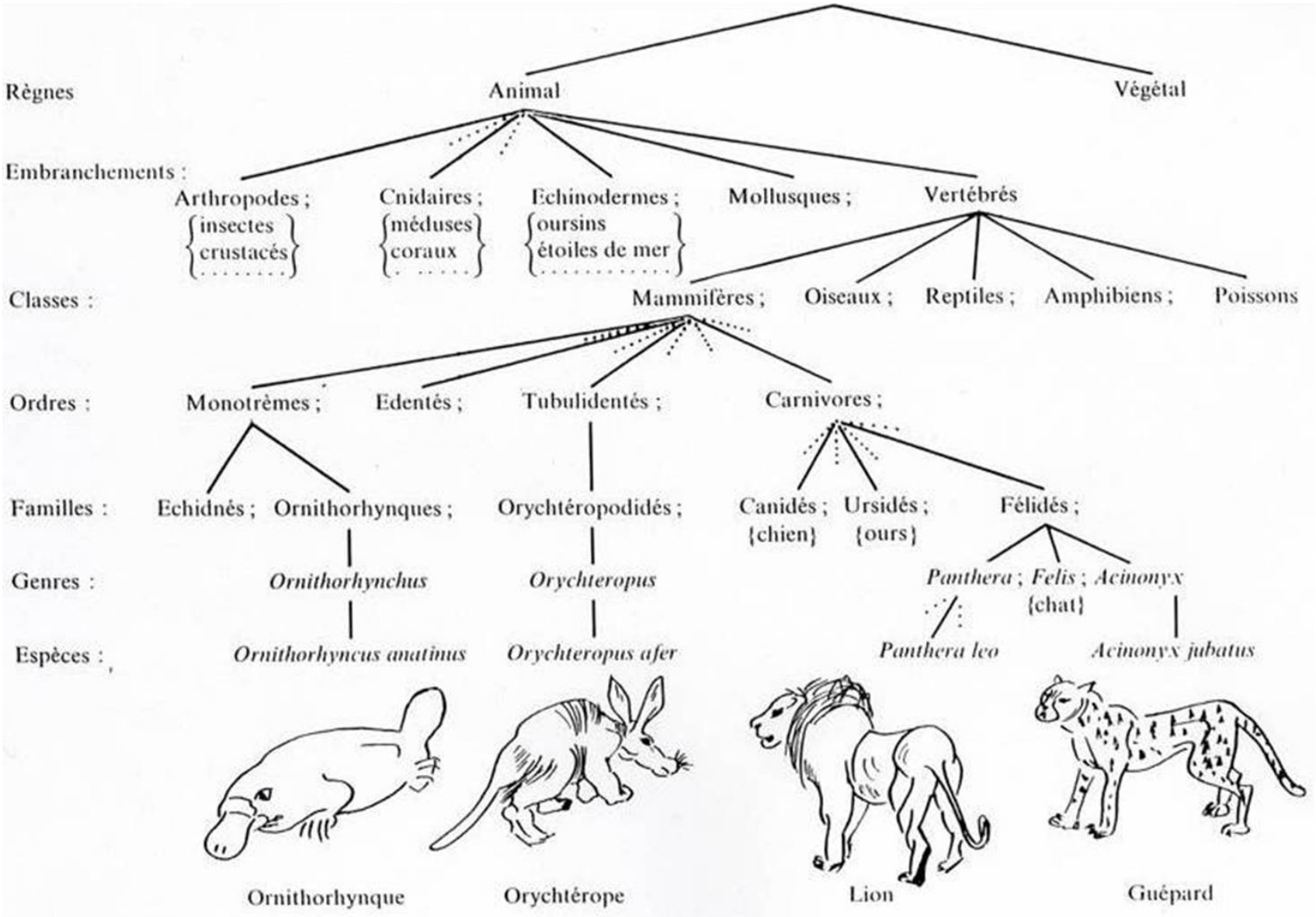
Classification ascendante hiérarchique (CAH)

Classification ascendante hiérarchique (CAH)

Introduction

- Définitions :
  - Classification : action de constituer ou construire des classes
  - Classe : ensemble d'individus (ou d'objets) possédant des traits de caractères communs (groupe, catégorie)
- Exemples
  - de classification : règne animal, disque dur d'un ordinateur, division géographique de la France, etc.
  - de classe : classe sociale, classe politique, etc.
- Deux types de classification :
  - hiérarchique : arbre, CAH
  - méthode de partitionnement : partition

Exemple de hiérarchie : le règne animal

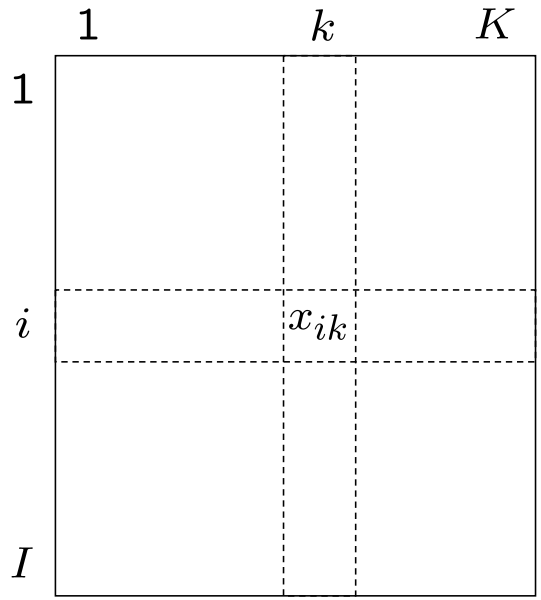


[CL. Nat.], TIB n°2, § 1.1

Classification ascendante hiérarchique (CAH)

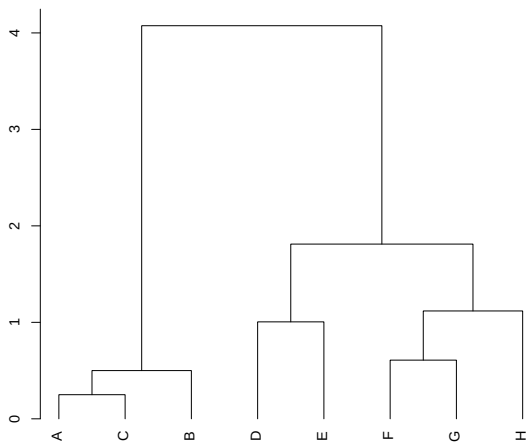
Quelles données pour quels objectifs ?

La classification s'intéresse à des tableaux de données individus $\times$ variables quantitatives



Objectifs : production d'une structure (arborescence) permettant :

- la mise en évidence de liens hiérarchiques entre individus ou groupes d'individus
- la détection d'un nb de classes « naturel » au sein de la population



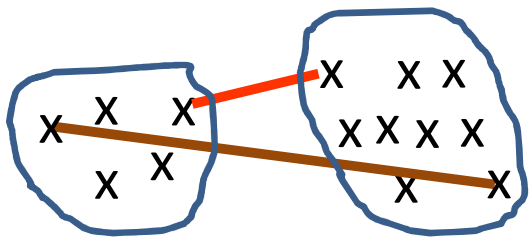
Critères

Ressemblance entre individus :

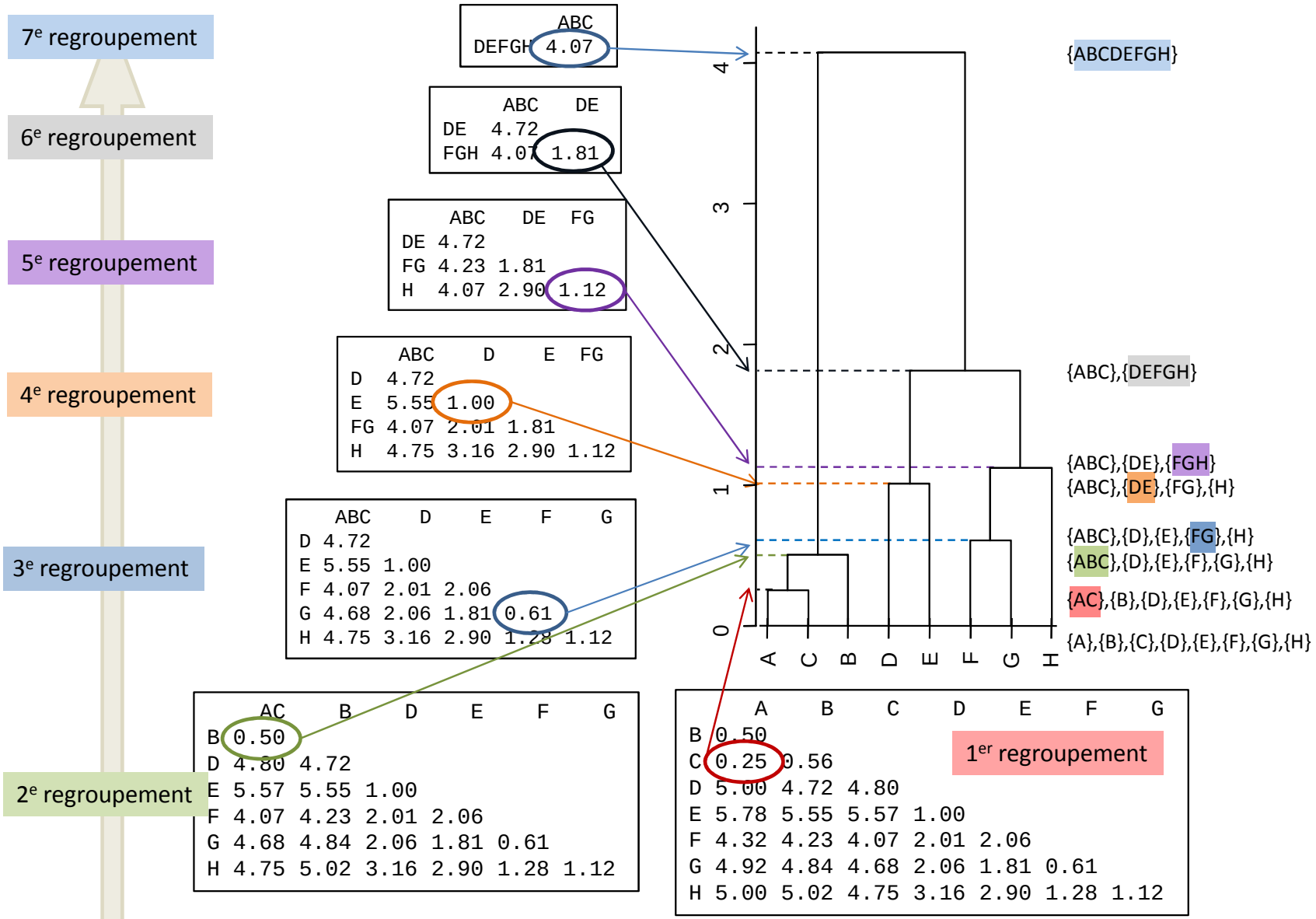
- distance euclidienne
- indice de similarité
- ...

Ressemblance entre groupes d'individus :

- saut minimum ou lien simple (plus petite distance)
- lien complet (plus grande distance)
- critère de Ward



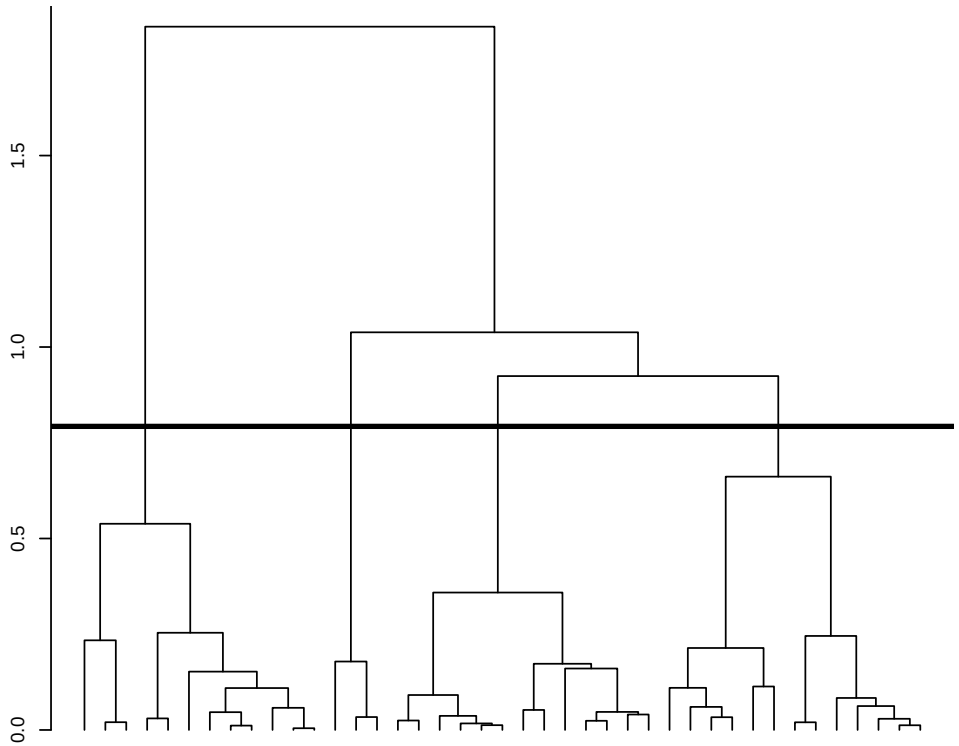
Algorithme


7e regroupement
6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
ABC
DEFGH 4.07
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
1er regroupement
4
3
2
1
0
A C B D E F G H
{ABCDEFGH}
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Arbres et partitions

Les arbres finissent tous ... par être coupés ! ! !

En définissant un niveau de coupure, on construit une partition



Remarque : vu le mode de construction, la partition n'est pas optimale mais est intéressante

Qualité d'une partition

Quand une partition est-elle bonne ?

- Si les individus d'une même classe sont proches
- Si les individus de 2 classes différentes sont éloignés

Et mathématiquement ça se traduit par ?

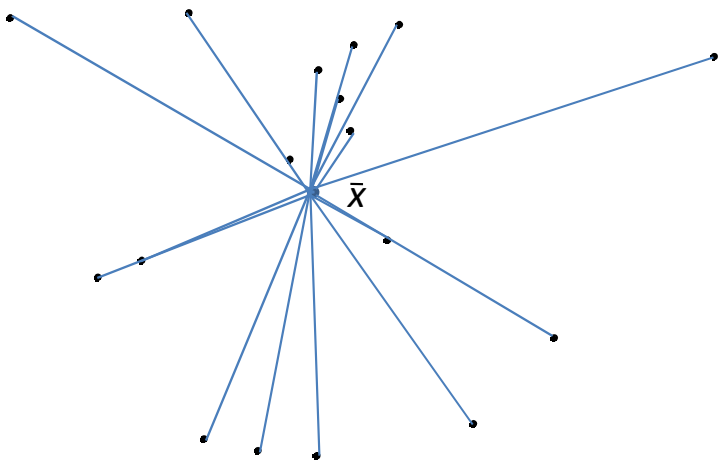
- Variabilité intra-classe petite
- Variabilité inter-classes grande

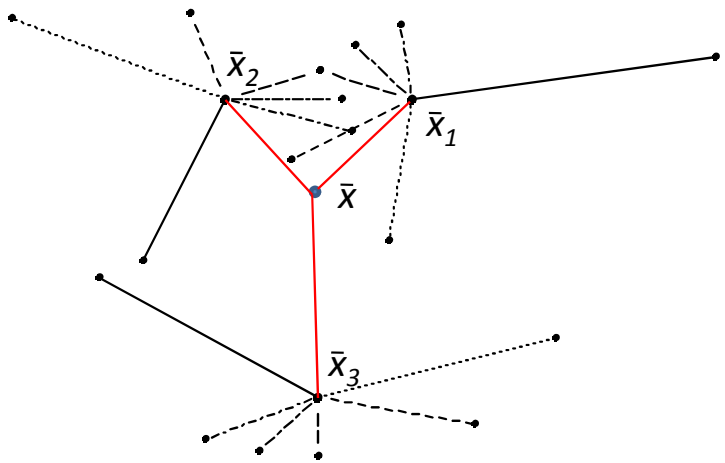
$\Longrightarrow$ Deux critères, lequel choisir ?

Qualité d'une partition

$\bar{x}_k$ moyenne de x_k, $\bar{x}_{qk}$ moyenne de x_k dans la classe q

$$\underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_k)^2}_{\text{Inertie totale}} = \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(x_{iqk}-\bar{x}_{qk})^2}_{\text{Inertie intra}} + \underbrace{\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{I}(\bar{x}_{qk}-\bar{x}_k)^2}_{\text{Inertie inter}}$$





$\Longrightarrow$ 1 seul critère !

Qualité d'une partition

La qualité d'une partition est mesurée par :

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

$\frac{\text{Inertie}_{\text{inter}}}{\text{Inertie}_{\text{totale}}} = 0 \Longrightarrow \forall k, \forall q, \bar{x}_{qk} = \bar{x}_k$

par variable, les classes ont mêmes moyennes

Ne permet pas de classifier

$\frac{\text{Inertie}_{\text{inter}}}{\text{Inertie}_{\text{totale}}} = 1 \Longrightarrow \forall k, \forall q, \forall i, x_{iqk} = \bar{x}_{qk}$

les individus d'une même classe sont identiques

Idéal pour classifier

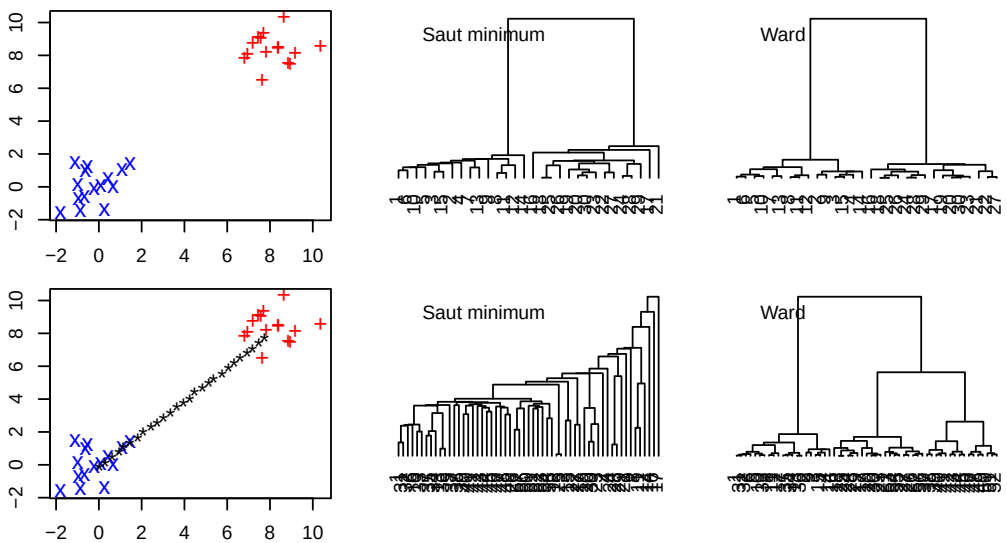
Attention : ce critère ne peut être jugé en absolu car il dépend du nb d'individus et du nb de classes

Méthode de Ward

- Initialisation : 1 classe = 1 individu $\Longrightarrow$ In. inter = In. totale
- A chaque étape : agréger les classes a et b qui minimisent la diminution de l'inertie inter

$$\text{Inertie}(a) + \text{Inertie}(b) = \text{Inertie}(a \cup b) - \underbrace{\frac{m_a m_b}{m_a + m_b} d^2(a,b)}_{\text{à minimiser}}$$

Regroupe les objets de faible poids et évite l'effet de chaîne



Regroupe des classes ayant des centres de gravité proches

Intérêt immédiat pour la classification

Classification ascendante hiérarchique (CAH)

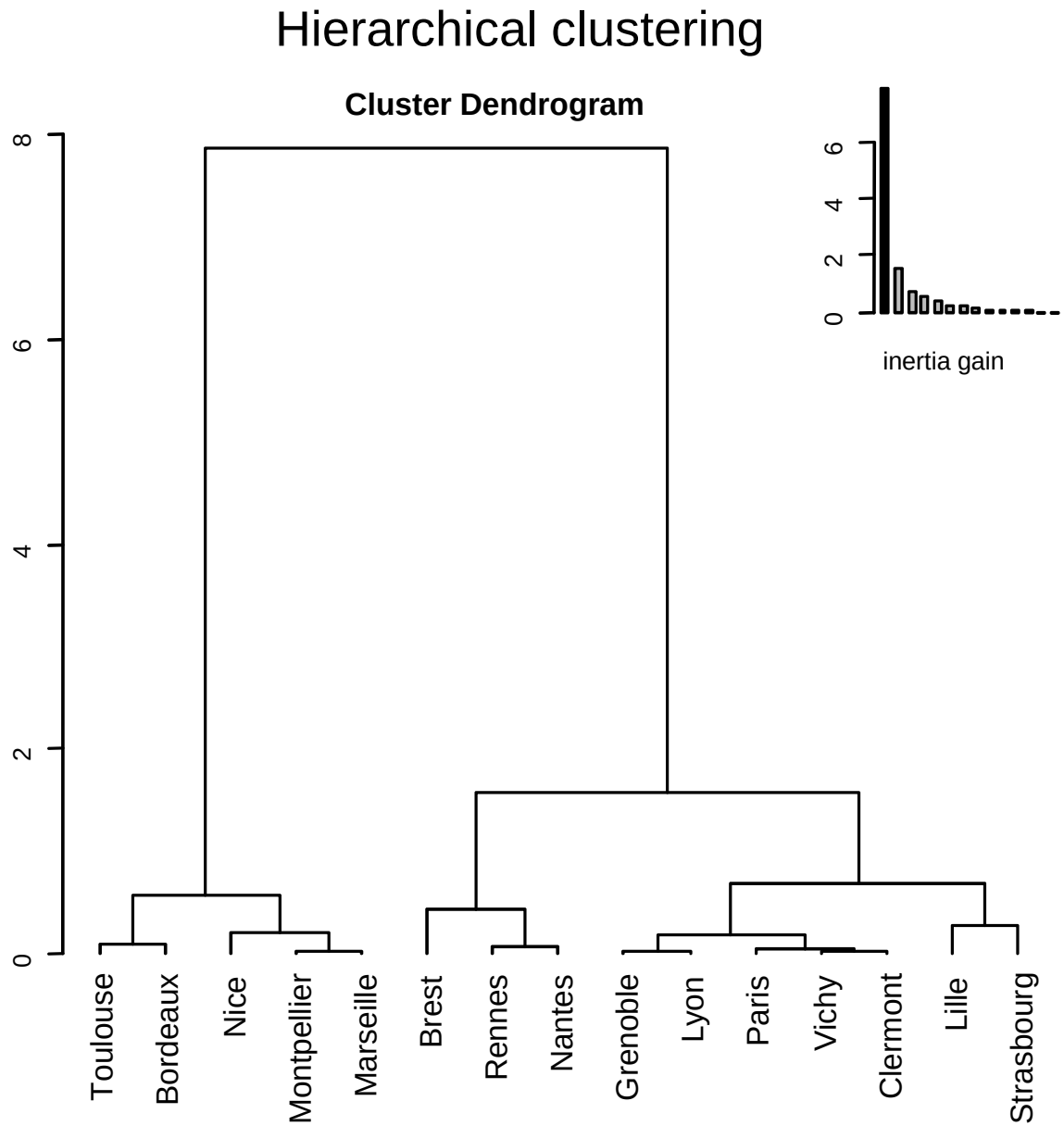
Les données température

- 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	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Quelles villes ont des profils météo similaires ?
Comment caractériser les groupes de villes ?

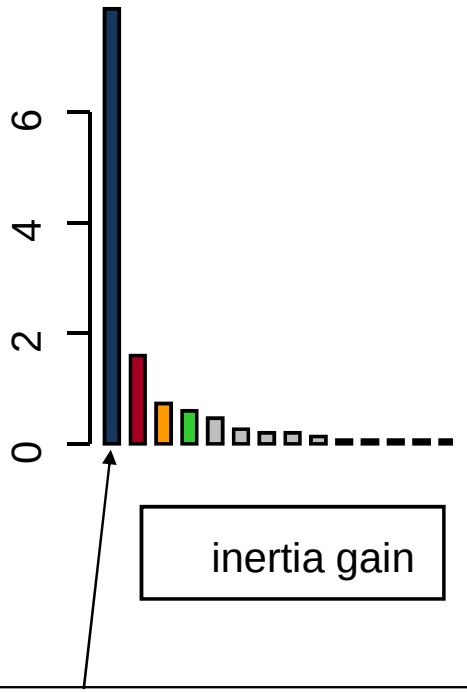
Les données température : l'arbre hiérarchique



Les données température

Pertes d'inertie inter
lors du passage de

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



Somme des pertes d'inertie $= 12$

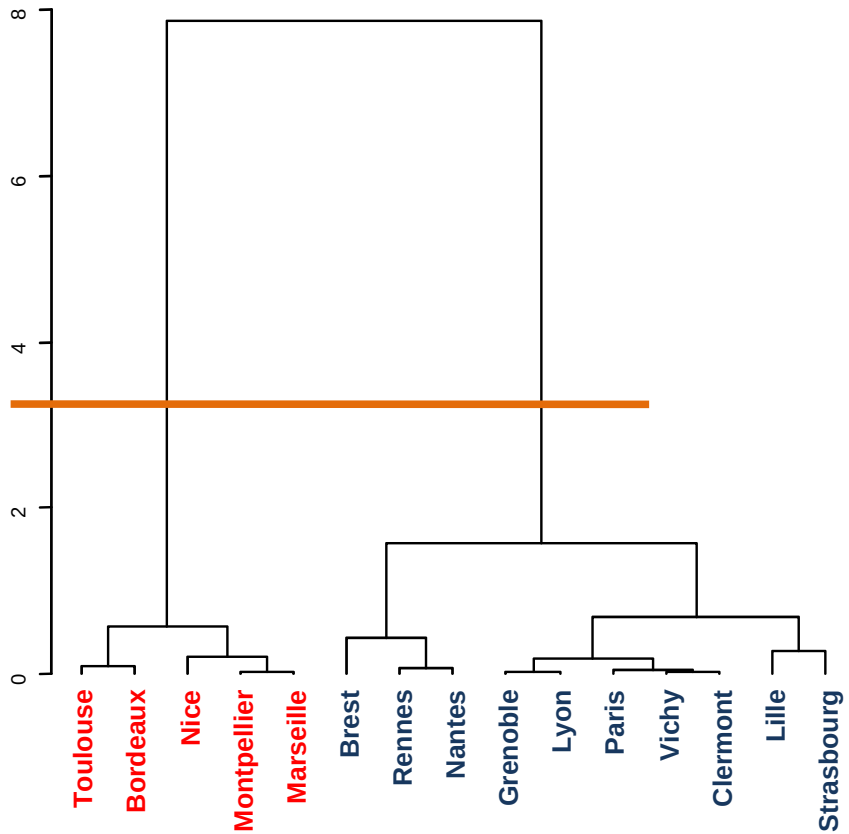
Utilisation de l'arbre pour construire une partition

Doit-on faire 2 groupes ? 3 groupes ? 4 ?

Découpage en 2 groupes :

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

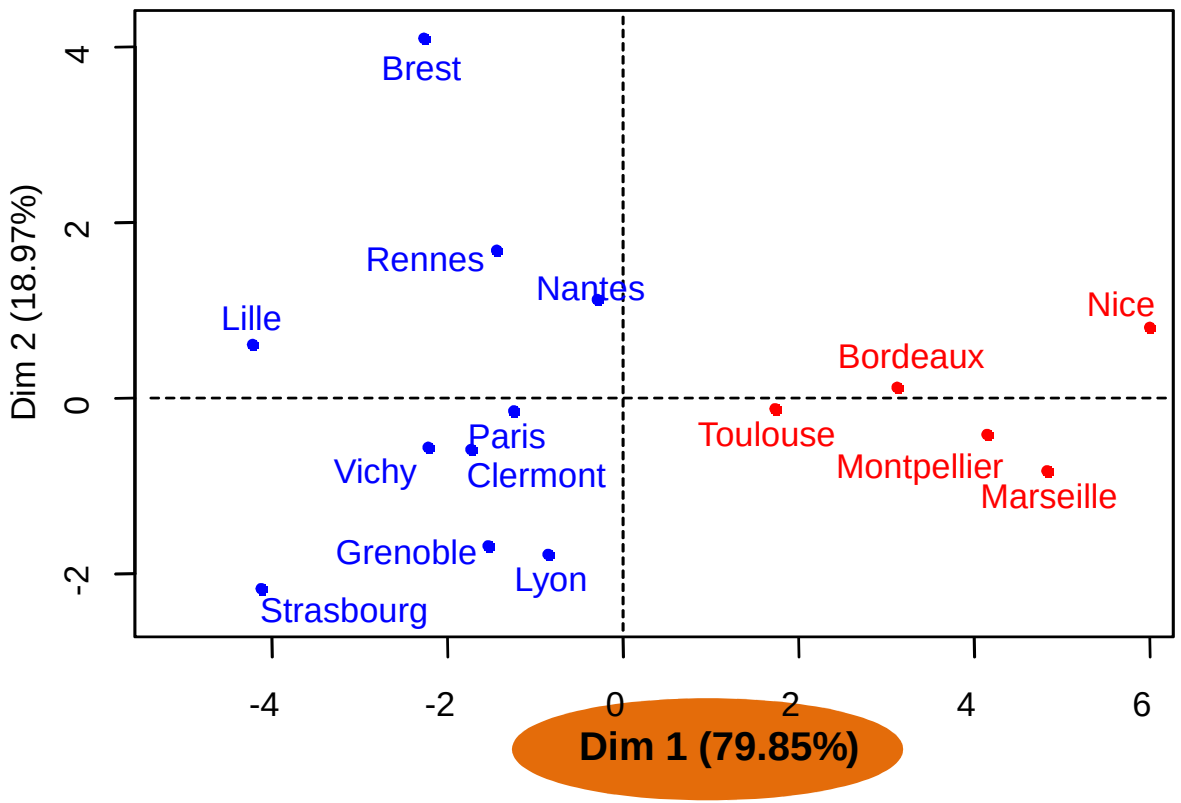
A quoi comparer ce pourcentage ?



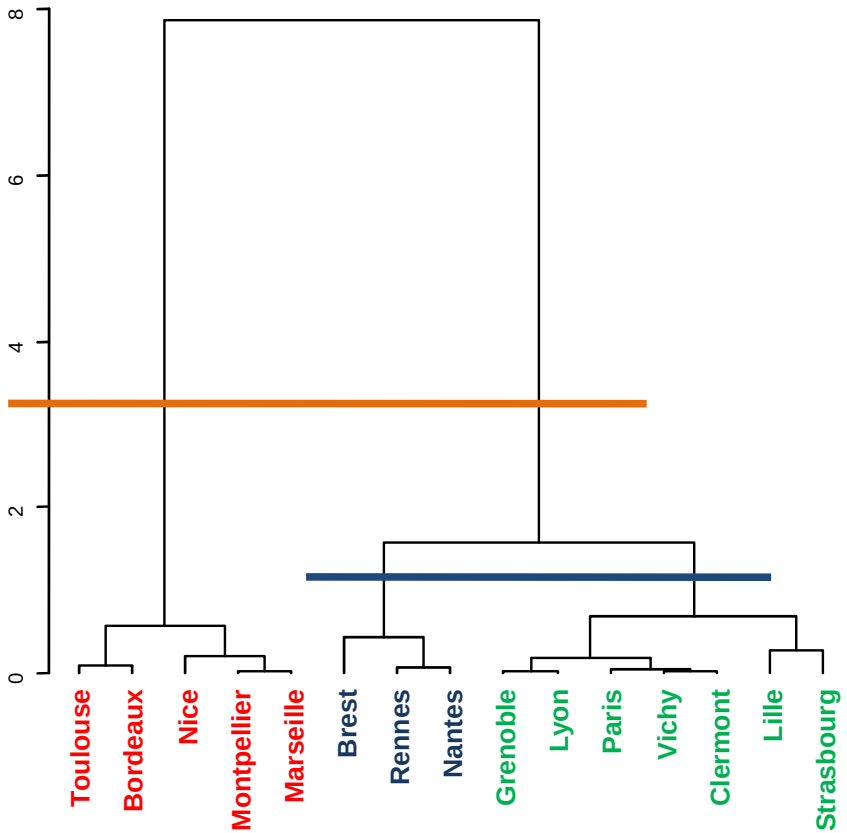
Utilisation de l'arbre pour construire une partition

66 % de l'information résumée avec ce découpage en 2 classes

A quoi comparer ce pourcentage ?



Utilisation de l'arbre pour construire une partition

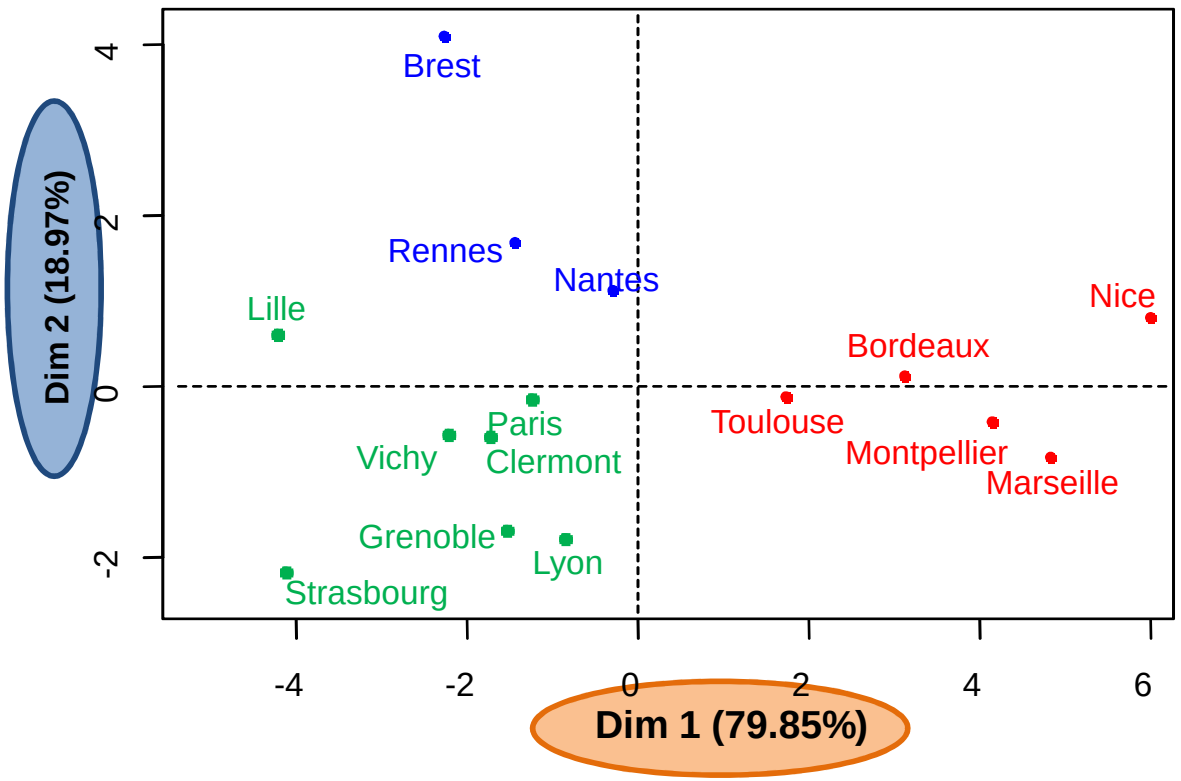


Séparer villes froides en 2 groupes :

$$\frac{\text{Inertie inter}}{\text{Inertie totale}} = \frac{1.56}{12} = 13\%$$

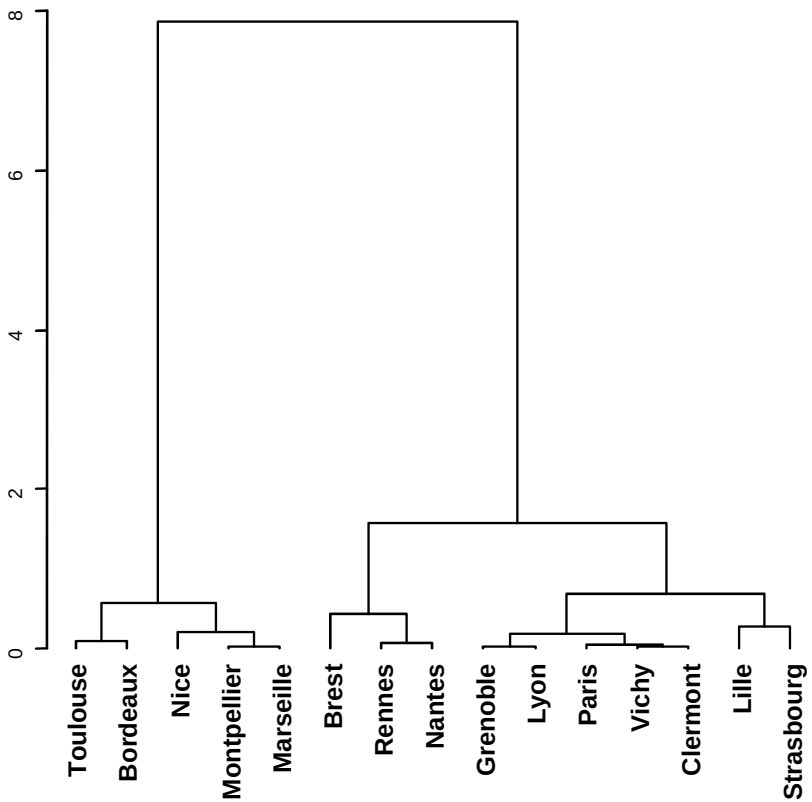
Utilisation de l'arbre pour construire une partition

Passage de 15 villes à 3 classes : 66 % + 13 % = 79 % de la variabilité des données

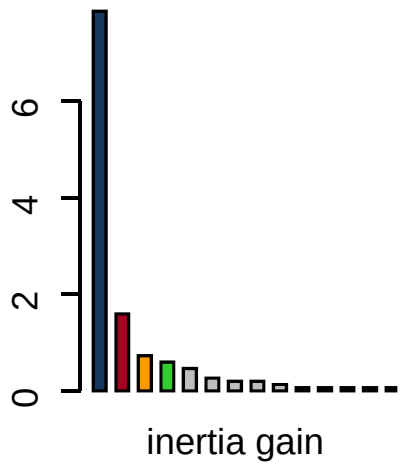


Détermination d'un nombre de classes

- A partir de l'arbre
- Dépend de l'usage (enquête, . . .)



- A partir du diagramme des indices de niveau
- Critère ultime : interprétabilité des classes

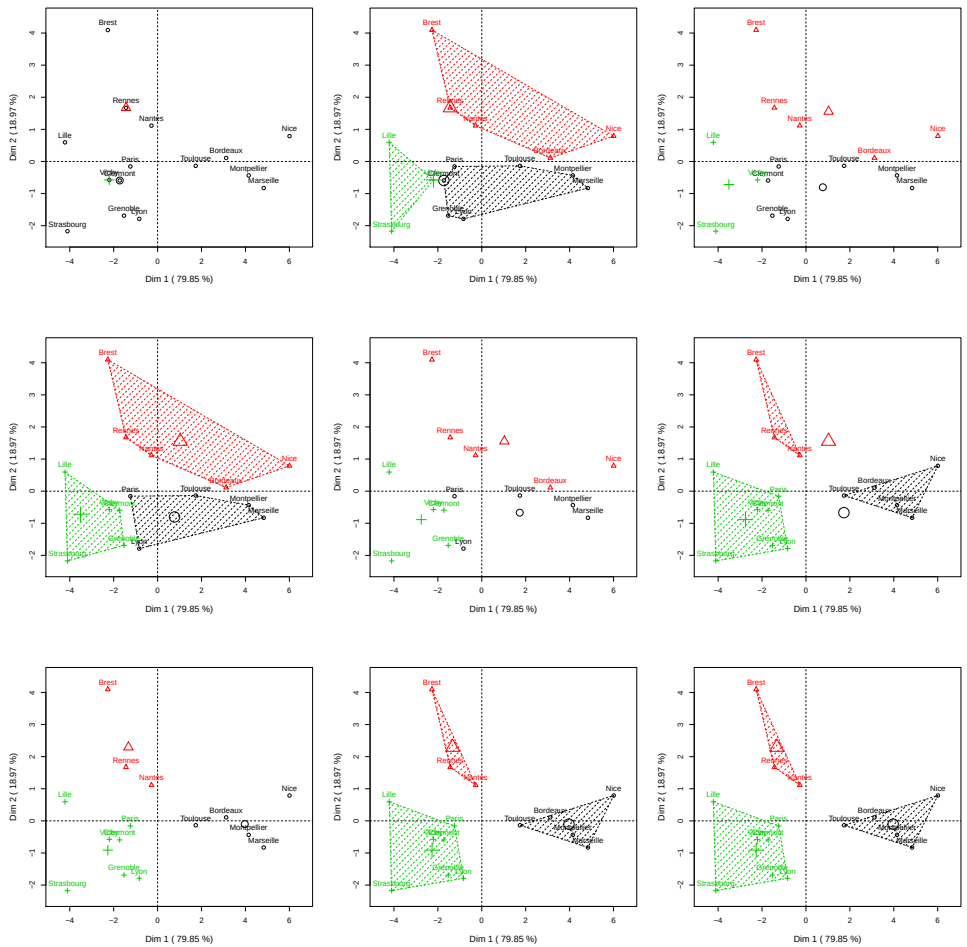


Classification ascendante hiérarchique (CAH)

Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Classification ascendante hiérarchique (CAH)

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

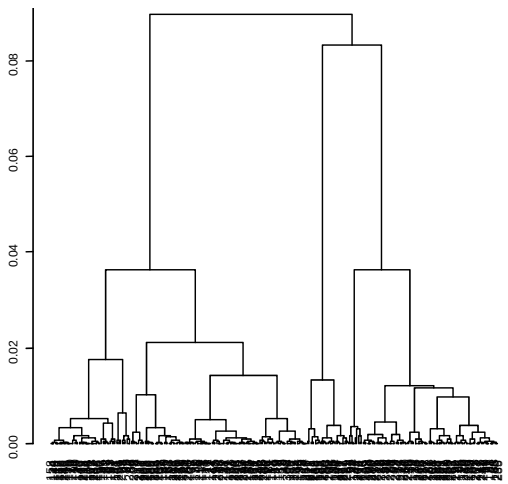
Algorithme de consolidation :

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

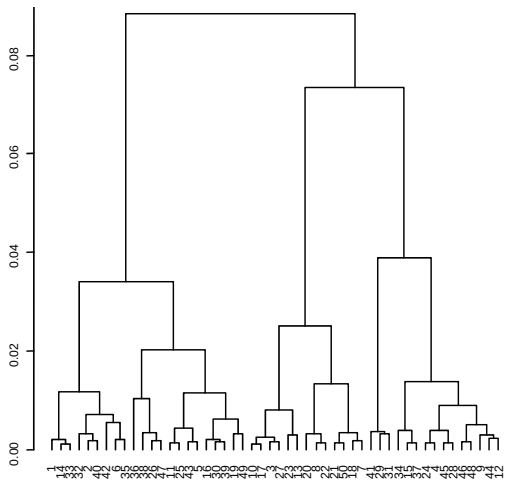
$\Longrightarrow$ amélioration de la partition (souvent non décisive)
Avantage : consolidation de la partition
Inconvénient : perte de l'info de hiérarchie

CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions $\Longrightarrow$ on se ramène au cas classique
- 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)
  - Obtention du « haut » de l'arbre de la CAH



Arbre sur données brutes



Arbre à partir de classes

CAH sur données qualitatives

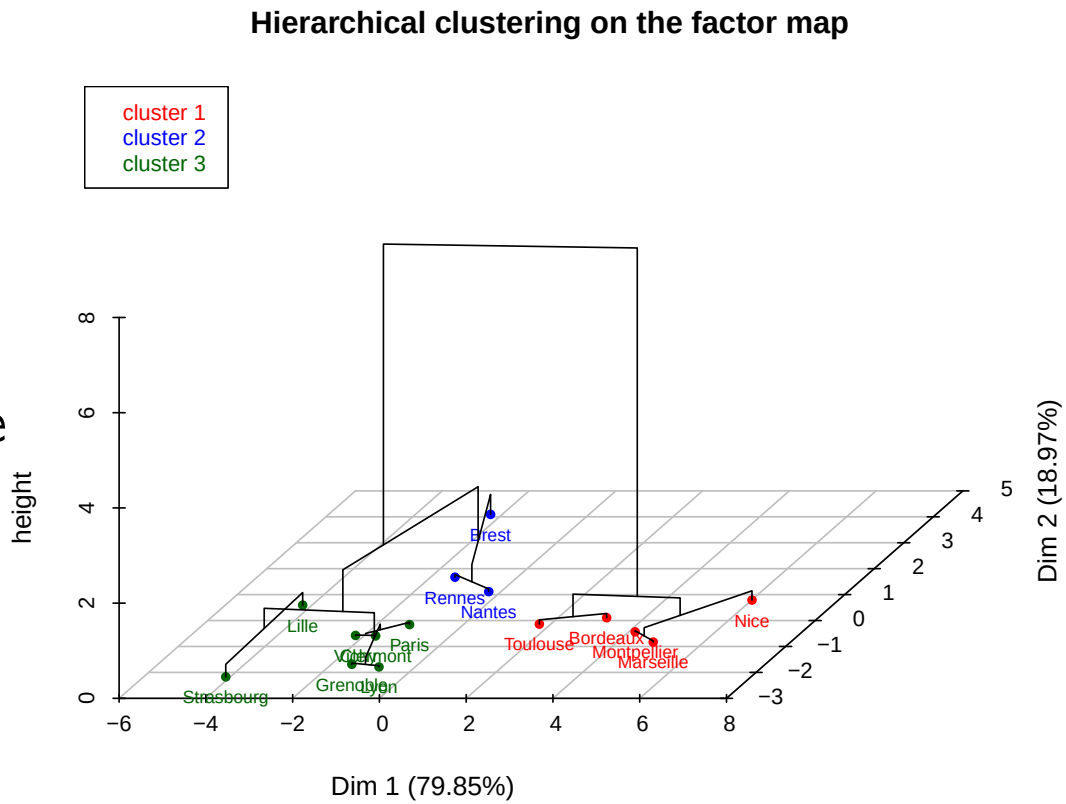
Deux stratégies pour faire une classification sur données qualitatives :

- Se ramener à des variables quantitatives
  - Faire une ACM et ne conserver que les premières dimensions
  - Faire la CAH à partir des composantes principales de l'ACM
- Utiliser des mesures adaptées aux données qualitatives : indice de similarité, indice de Jaccard, etc.

Enchaînement analyse factorielle - classification

- Données qualitatives : ACM renvoie des composantes principales qui sont quantitatives
- L'analyse factorielle élimine les dernières composantes qui ne contiennent que du bruit $\Longrightarrow$ classification plus stable

- Représentation de l'arbre et des classes sur un plan factoriel $\Longrightarrow$ vision continue avec AF, discontinue avec CAH ; vision de l'information sur d'autres axes avec CAH

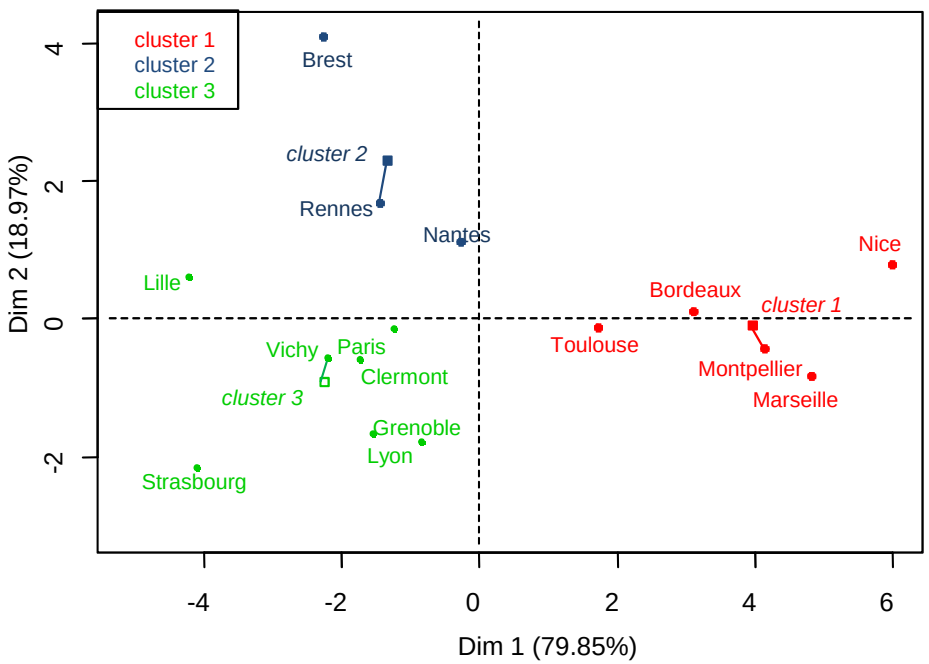


Classification ascendante hiérarchique (CAH)

Constitution des classes - Édition des parangons

Parangon : individu le plus proche du centre d'une classe

```
classe 1 : Montpellier  Bordeaux  Marseille    Nice  Toulouse
             0.419        1.141      1.193   2.242     2.256
classe 2 :     Rennes    Nantes      Brest
             0.641        1.586      2.045
classe 3 :     Vichy  Clermont   Grenoble   Paris      Lyon
             0.428        0.669      1.184   1.339     1.680
```



Caractérisation des classes

- Objectifs :
  - Trouver les variables les plus caractérisantes pour la partition
  - Caractériser une classe (ou un groupe d'individus) par des variables quantitatives
  - Trier les variables qui caractérisent les classes

- Questions :
  - Quelles variables caractérisent le mieux la partition ?
  - Comment caractériser les individus de la classe 1 ?
  - Quelles variables les caractérisent le mieux ?

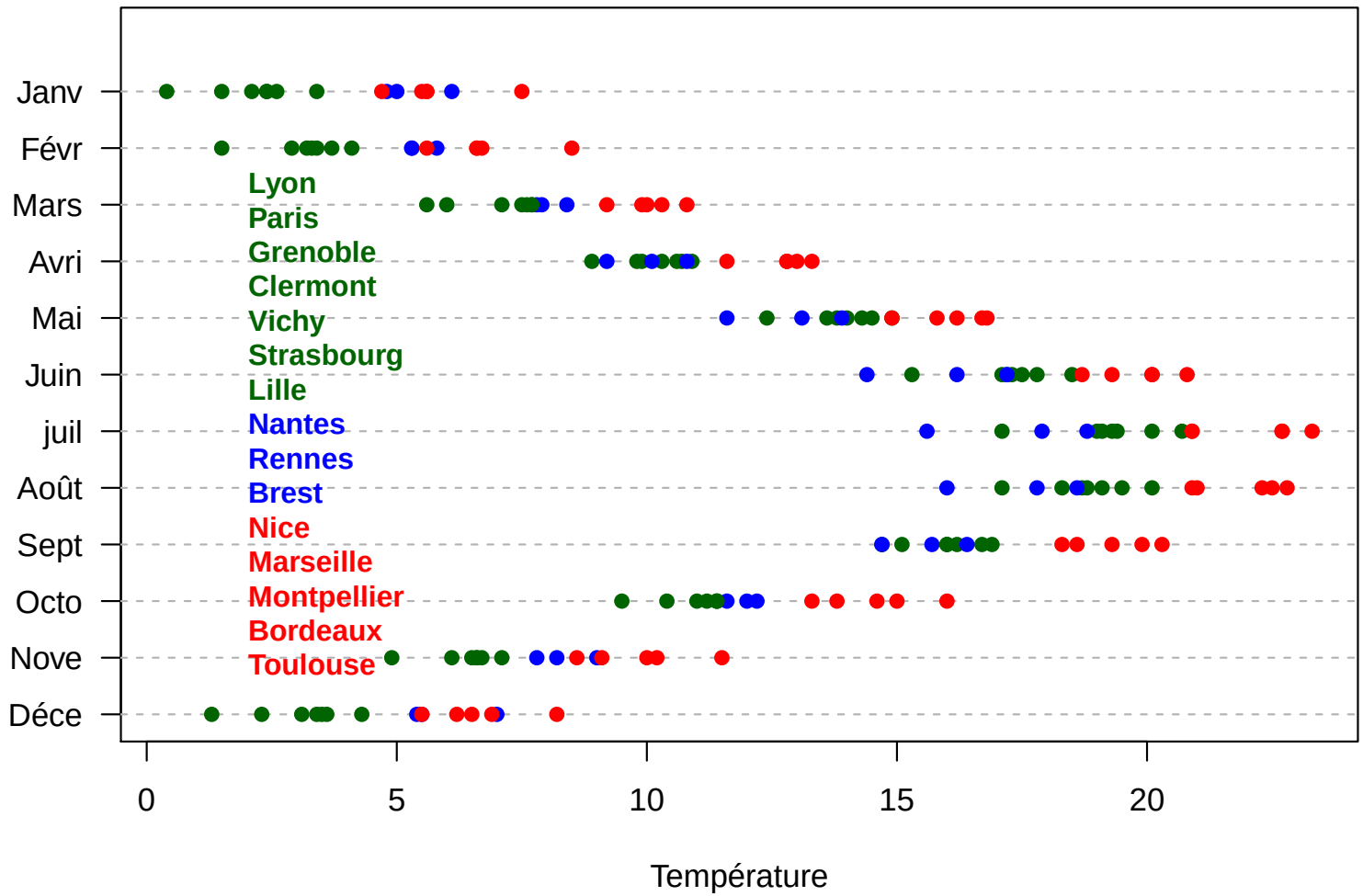
Caractérisation des classes

Quelles variables caractérisent le mieux la partition ?

- Pour chaque variable quantitative :
  - construire le modèle d'analyse de variance entre la variable quantitative expliquée par la variable de classe
  - faire le test de Fisher de l'effet de la classe
- Trier les variables par probabilité critique croissante

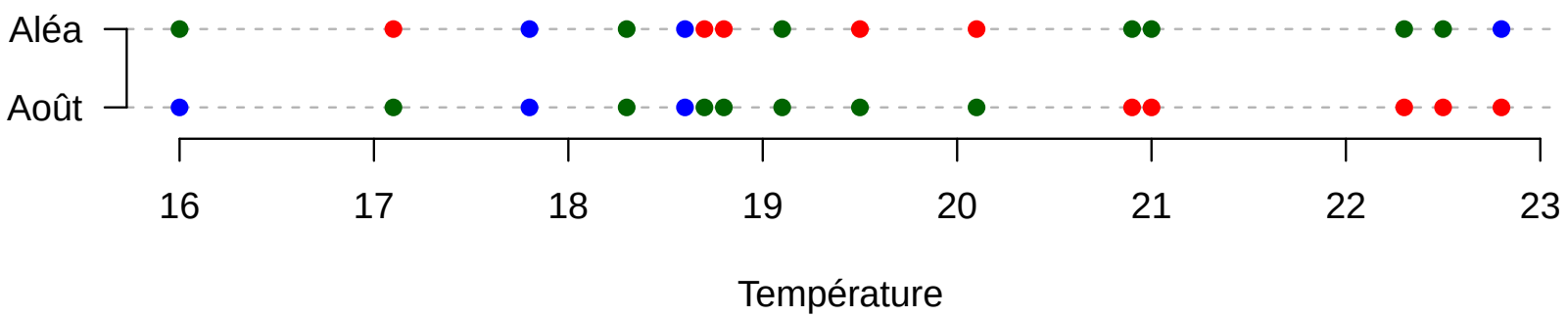
```
       Eta2   P-value
Octo 0.8362 1.930e-05
Sept 0.8301 2.407e-05
Févr 0.8227 3.103e-05
Mars 0.8126 4.326e-05
Janv 0.8118 4.444e-05
Nove 0.8083 4.963e-05
Avri 0.7929 7.890e-05
Déce 0.7871 9.316e-05
Août 0.7864 9.503e-05
Juin 0.7241 4.409e-04
Mai  0.7164 5.205e-04
juil 0.7156 5.287e-04
```

Caractérisation d'une classe par les variables quantitatives



Caractérisation d'une classe par les variables quantitatives

Idée 1 : si les valeurs de X pour la classe q semblent tirées au hasard parmi les valeurs de X, alors X ne caractérise pas la classe q



Idée 2 : plus l'hypothèse d'un tirage au hasard est douteuse, plus X caractérise la classe q

Caractérisation d'une classe par les variables quantitatives

Idée : référence du tirage au hasard de n_q valeurs parmi N

Quelles valeurs peut prendre $\bar{x}_q$? (*i.e.* quelle est la loi de $\bar{X}_q$?)

$$\mathbb{E}(\bar{X}_q) = \bar{x} \qquad \mathbb{V}(\bar{X}_q) = \frac{s^2}{n_q}\left(\frac{N-n_q}{N-1}\right)$$

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

$$\Longrightarrow \quad \text{Valeur-test} = \frac{\bar{x}_q - \bar{x}}{\sqrt{\frac{s^2}{n_q}\left(\frac{N-n_q}{N-1}\right)}} \sim \mathcal{N}(0,1)$$

- Si $|\text{Valeur-test}| \geq 1.96$ alors X caractérise la classe q
- X caractérise d'autant mieux la classe q que V-test grande

Idée : classer les variables par $|\text{Valeur-test}|$ décroissante

Caractérisation d'une classe par les variables quantitatives

$quanti$`1`

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Sept	3.40	19.30	17.00	0.755	1.79	0.000678
Moye	3.39	13.80	11.80	0.742	1.55	0.000705
Avri	3.33	12.70	11.00	0.580	1.37	0.000871
Octo	3.32	14.50	12.30	0.941	1.77	0.000893
Mars	3.24	10.00	8.23	0.524	1.48	0.001210
Août	3.18	21.90	19.60	0.792	1.94	0.001490
Juin	3.00	19.80	17.80	0.727	1.73	0.002670
Mai	3.00	16.10	14.40	0.691	1.45	0.002720
Nove	2.97	9.88	7.93	0.999	1.74	0.003020
juil	2.92	22.10	19.80	1.000	2.06	0.003550
Févr	2.88	6.80	4.83	0.940	1.81	0.003940
Déce	2.54	6.66	4.85	0.896	1.89	0.011200
Janv	2.46	5.78	3.97	0.924	1.94	0.013700

Caractérisation d'une classe par les variables quantitatives

$‘2‘

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Mai	-2.02	12.90	14.40	0.953	1.45	0.04380
Août	-2.02	17.50	19.60	1.090	1.94	0.04330
Juin	-2.05	15.90	17.80	1.160	1.73	0.04020
juil	-2.18	17.40	19.80	1.350	2.06	0.02900
Long	-2.88	-2.34	2.58	1.380	3.21	0.00404
Ampl	-2.95	12.40	15.90	1.560	2.25	0.00316

$‘3‘

	v.test	Mean in category	Overall mean	sd in category	Overall sd	p.value
Sept	-2.05	15.90	17.00	0.738	1.79	0.040700
Avri	-2.11	10.20	11.00	0.637	1.37	0.035100
Moye	-2.60	10.70	11.80	0.620	1.55	0.009220
Octo	-2.81	10.90	12.30	0.661	1.77	0.004940
Mars	-2.85	7.03	8.23	0.807	1.48	0.004310
Nove	-3.15	6.36	7.93	0.654	1.74	0.001620
Févr	-3.25	3.16	4.83	0.763	1.81	0.001150
Déce	-3.28	3.07	4.85	0.911	1.89	0.001020
Janv	-3.36	2.11	3.97	0.876	1.94	0.000793

Caractérisation des classes par les variables qualitatives

Quelles variables caractérisent le mieux la partition ?

- Pour chaque variable qualitative, construire un test du χ^2 entre la variable et la variable de classe
- Trier les variables par probabilité critique croissante

```
$test.chi2
           p.value df
Région 0.001700272  6
```

Caractérisation d'une classe par les variables qualitatives

La modalité *Nord-Est* caractérise-t-elle la classe 3 ?

	Classe 3	Autre classe	Total
Nord-Est	$n_{mc} = 3$	0	$n_m = 3$
Pas NE	4	8	12
Total	$n_c = 7$	8	$n = 15$

Test : H_0 : $\frac{n_{mc}}{n_c} = \frac{n_m}{n}$ contre H_1 : m anormalement élevée dans c

Sous H_0 : $\mathcal{L}(N_{mc}) = \mathcal{H}(n_c, \frac{n_m}{n}, n)$ $\quad P_{H_0}(N_{mc} \geq n_{mc})$

Classe 3

	Cla/Mod	Mod/Cla	Global	p.value	v.test
Région=NE	100.00	42.86	20.00	0.077	1.769

$\frac{3}{3} \times 100 = 100$; $\frac{3}{7} \times 100 = 42.86$; $\frac{3}{15} \times 100 = 20$; $P_{\mathcal{H}(7, \frac{3}{15}, 15)}[N_{mc} \geq 3] = 0.077$

$\Longrightarrow$ H_0 acceptée, *Nord-Est* n'est pas sur-représenté dans la classe 3

Tri des modalités en fonction des probabilités critiques

Caractérisation d'une classe par les axes

Les axes factoriels sont aussi des variables quantitatives

```
$`1`
      v.test   Mean in   Overall     sd in  Overall   p.value
                category      mean  category       sd
Dim.1   3.39      3.97         0      1.46      3.1  0.000693

$`2`
      v.test   Mean in   Overall     sd in  Overall   p.value
                category      mean  category       sd
Dim.2   2.84      2.29         0      1.29     1.51   0.00447

$`3`
      v.test   Mean in   Overall     sd in  Overall   p.value
                category      mean  category       sd
Dim.2  -2.11    -0.911         0     0.927     1.51    0.0346
Dim.1  -2.56    -2.270         0     1.260     3.10    0.0104
```

Conclusion

- La classification s'applique à des tableaux individus $\times$ variables quantitatives
  $\Rightarrow$ L'ACM transforme des variables qualitatives en variables quantitatives

- CAH donne un arbre hiérarchique $\Rightarrow$ nombre de classes

- K-means consolide les classes

- Caractérisation des classes par des variables actives et supplémentaires, quantitatives et qualitatives