FACULTÉ DES SCIENCES ECONOMIQUES

UNIVERSITE DE RENNES 1

La classification

Fabien Chevalier
Jérôme Le Bellac

2012-2013

UNIVERSITÉ DE RENNES 1

Introduction :

- Classification : méthode d'analyse de données

- Objectif : Obtenir une représentation schématique simple d'un tableau de données complexe à partir d'une typologie (segmentation), c'est à dire d'une partition des n individus dans des classes, définies par l'observations de p variables.

UNIVERSITÉ DE RENNES 1

Introduction :

- Méthode : Classifier, c'est regrouper entre eux des objets similaires selon certains critères. Les diverses techniques de classification visent toutes à répartir *n individus, caractérisés par p variables X1, X2, ..., Xp en un* certain nombre m de sous-groupes aussi homogènes que possible, chaque groupe étant bien différencié des autres.

- Deux grandes techniques de classification :
  - Le partitionnement
  - La classification hiérarchique

Introduction :

- Présentation à partir d'un jeu de données

  - 22 régions de France métropolitaine
  - 5 variables (voir si on garde les 10)
    - Densité
    - Criminalité
    - Espérance de vie
    - Pauvreté
    - Enseignement

UNIVERSITÉ DE RENNES 1

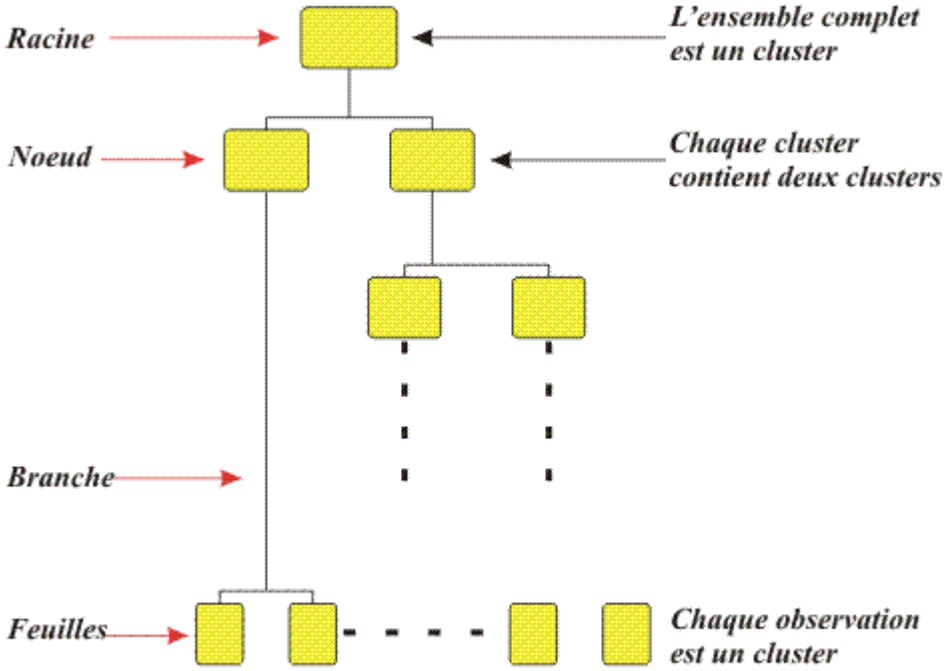
Plan

- CAH
  - Présentation de la méthode
  - Importance du choix de la distance
  - Exemple sur données réelles
  - Limites de la méthode
- Méthode de partitionnement
  - Centres Mobiles
  - Variantes (Présentation et exemple)
  - Limites des méthodes
- Classification mixte
  - Principe de la méthode
  - Exemple
- Validation et sélection de la classification optimale

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Objectif: obtenir une hiérarchie, c'est-à-dire une collection de groupes d'observations.
- Ne pas confondre hiérarchie et typologie. Une typologie est la partition de l'ensemble des données.
- Plusieurs typologies peuvent donc être définies à partir d'une seule hiérarchie.



Hiérarchie

UNIVERSITÉ DE RENNES 1

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- 1ère phase: Initialisation de l'algorithme.
  - Les classes initiales = n singletons individus.
  - Calcul de la matrice des distances des individus 2 à 2
- 2ème phase : Itération des étapes suivantes.
  - Regrouper les 2 éléments (individus ou groupes) les plus proches au sens d'un critère chosi.
  - Mise à jour du tableau des distances en remplaçant les deux éléments regroupés par le nouveau et en recalculant sa distance avec les autres classes.
- Fin de l'itération : agrégation de tous les individus en une seule classe.

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Réflexions pré-algorithme

  - Nécessité de définir une distance entre les individus

  - Définir un critère de regroupement des individus à minimiser aussi appelé stratégie d'agrégation.

  - Stratégie pour définir la meilleure typologie finale.

UNIVERSITÉ DE RENNES 1

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- 4 grandes étapes
  - Préparation des données
  - Choix de l'indice de dissimilarité entre les individus
  - Choix de l'indice d'agrégation
  - Choix de la partition finale

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Choix de l'indice de dissimilarité entre les individus

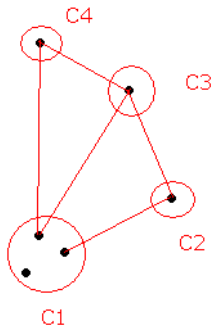
  - Le choix de la mesure de distance entre individus dépend des données étudiées et des objectifs.

  - Exemples :
    - Distance Euclidienne : le type de distance le plus couramment utilisé. Il s'agit d'une distance géométrique dans un espace multidimensionnel. $distance(x,y) = \{\sum_i (x_i - y_i)^2\}^{1/2}$
    - Distance Euclidienne au carré : Permet de "sur-pondérer" les objets atypiques (éloignés), en élevant la distance euclidienne au carré. $distance(x,y) = \sum_i (x_i - y_i)^2$
    - Distance du City-block (Manhattan) : cette distance est simplement la somme des différences entre les dimension. $distance(x,y) = \sum_i |x_i - y_i|$

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Choix de l'indice d'agrégation
  - On regroupe les éléments en minimisant l'indice d'agrégation
  - Plusieurs méthodes encore, mais la méthode la plus connu : Méthode de ward

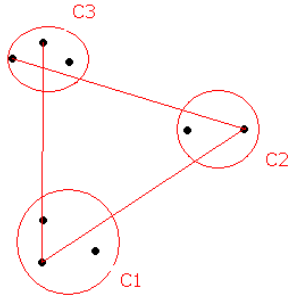


  - autres stratégies :
    - ❖ stratégie du saut minimum ou single linkage :

      On regroupe les 2 éléments présentant la plus **petite** distance entre éléments des deux classes.

    - ❖ stratégie du saut maximum ou du diamètre ou complete linkage :

      On regroupe les 2 éléments présentant la plus **grande** distance entre éléments des deux classes.



La classification Ascendante Hiérarchique

A) présentation de l'algorithme

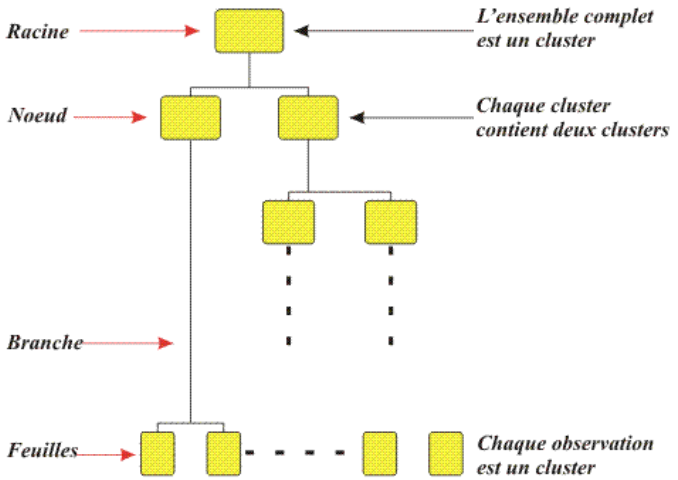
- Méthode de Ward
  - Objectif : gain minimum d'inertie intra-classe à chaque agrégation
  - perte d'inertie interclasse due à cette agrégation
- Calcul : utilise une analyse de la variance approchée afin d'évaluer les distances entre classes.
  - Minimisation de la Somme des Carrés (SC) de tous les couples hypothétiques de classes : agrégation
  - Les indices d'agrégation sont recalculés l'aide de la règle suivante :
    si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par la distance entre les barycentres de la classe M et de J.

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Choix de la partition finale
  - On défini un ensemble de parties, ou classes de l'ensemble des individus tel que :
    - ❖ toute classe soit non vide
    - ❖ deux classes distinctes sont disjointes
    - ❖ tout individu appartient à une classe.

Rappel :



Hiérarchie

La classification Ascendante Hiérarchique

A) présentation de l'algorithme

- Choix de la partition finale
  - Graphique de l'indice de niveau
    - ❖ l'utilisateur doit repérer des sauts extrêmement importants dans les valeurs, en analysant l'histogramme des indices de niveau
    - ❖ Si ces sauts concernent les k derniers nœuds de l'arbre, alors un découpage en (k+1) classes sera pertinent.
  - ⇔ La hauteur d'une branche est proportionnelle à la distance entre 2 classes
    - ❖ On coupe au niveau d'une longue branche
    - ❖ ⇔ coupé avant une forte perte d'inertie dans le cas de la méthode de Ward

La classification Ascendante Hiérarchique

B) Importance du choix des distances

- Calcul de la 1ère matrice de distance

Région	Densité	criminalité
Alsace	0,41	-0,45
Aquitaine	-0,34	-0,07

- Distance Euclidienne

Distance Alsace - Aquitaine

$$\sqrt{(0{,}41-(-0.34))^2+(-0.45-(-0.07))^2}=0{,}84$$

- Distance de Manhattan

Distance Alsace - Aquitaine

$$|0{,}41-(-0{,}34)|+|-0{,}45-(-0{,}07)|=1{,}13$$

La classification Ascendante Hiérarchique

B) Importance du choix des distances

- Calcul de la 1ère matrice de distance
  - 2 matrices des distance totalement différentes

	Alsace	Aquitaine	Auvergne	Basse-Normandie	Bourgogne
Alsace	0	-	-	-	-
Aquitaine	1,12	0	-	-	-
Auvergne	1,48	1,11	0	-	-
Basse-Normandie	1,03	0,72	0,44	0	-
Bourgogne	1,25	0,88	0,23	0,21	0

Distance Manhattan

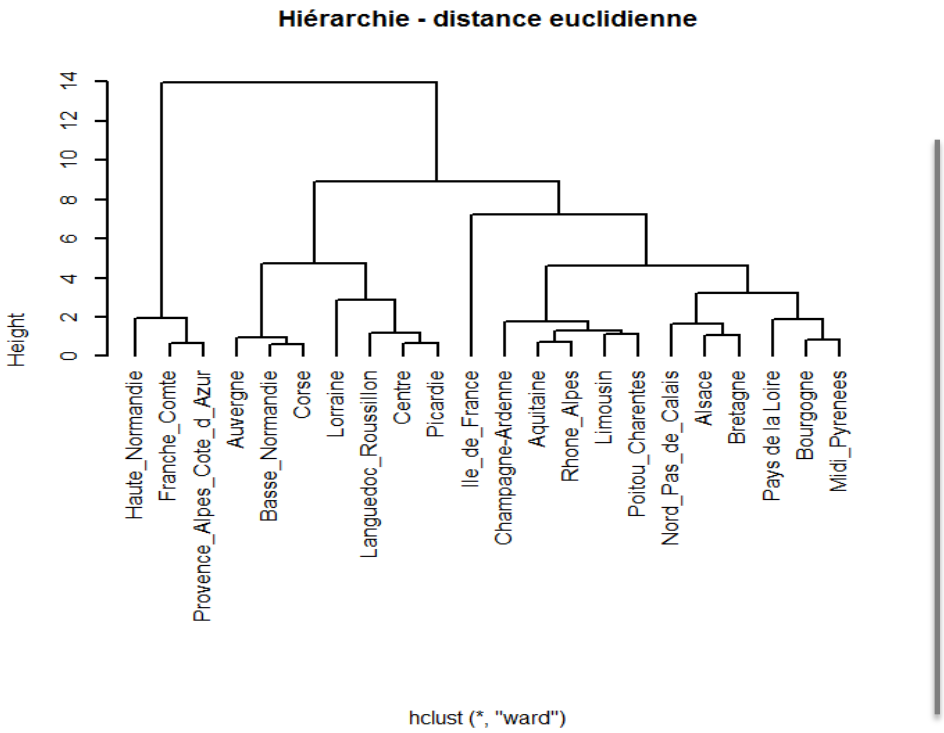
Distance Euclidienne

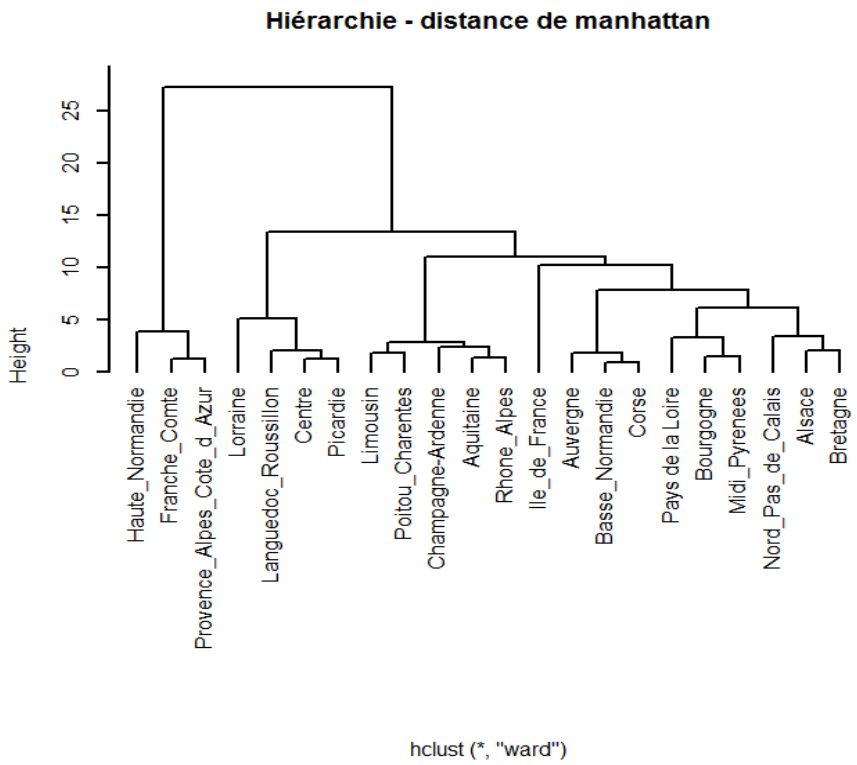
	Alsace	Aquitaine	Auvergne	Basse-Normandie	Bourgogne
Alsace	0	-	-	-	-
Aquitaine	0,84	0	-	-	-
Auvergne	1,07	0,99	0	-	-
Basse-Normandie	0,79	0,7	0,32	0	-
Bourgogne	0,96	0,76	0,22	0,17	0

La classification Ascendante Hiérarchique

B) Importance du choix des distances

- Obtention des hiérarchies
  - Fait avec densite espérance de vie enseigment criminalite pauvrete

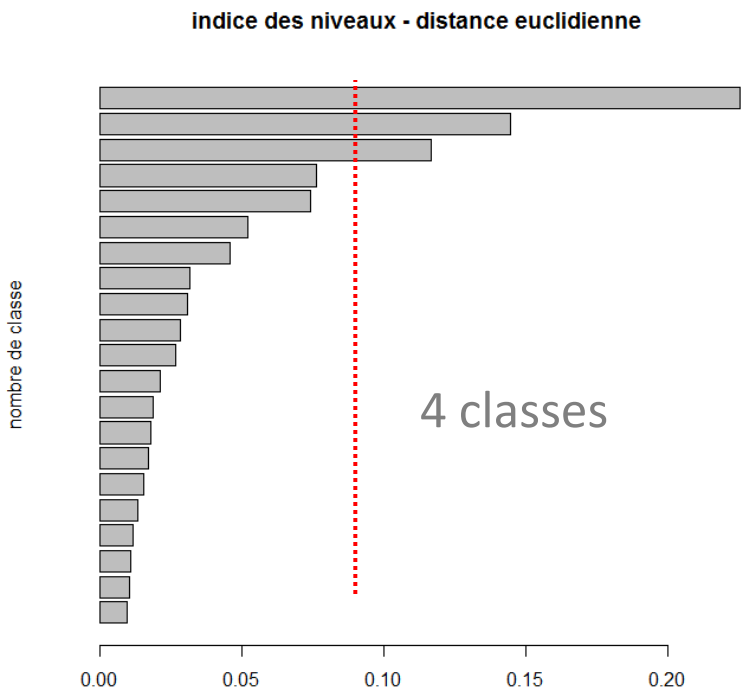




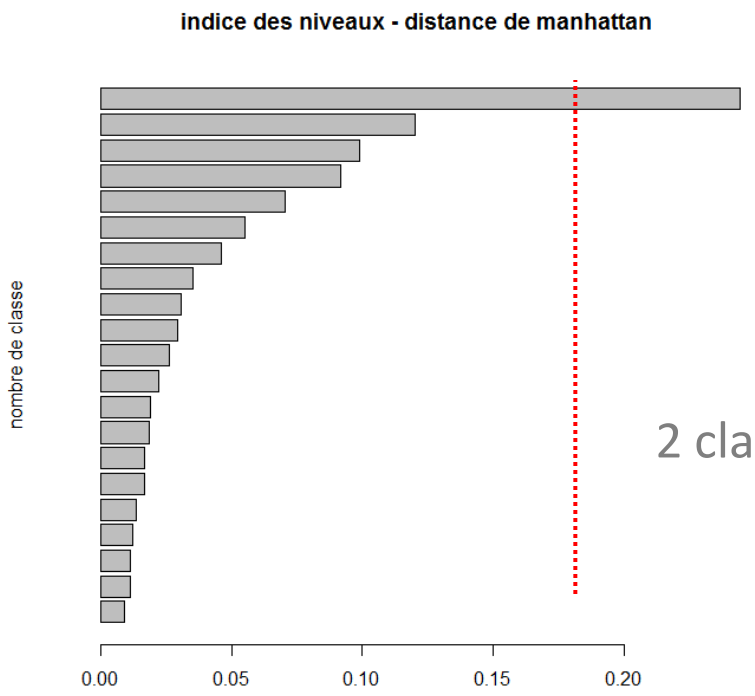
La classification Ascendante Hiérarchique

B) Importance du choix des distances

- Choix du nombres d'axes



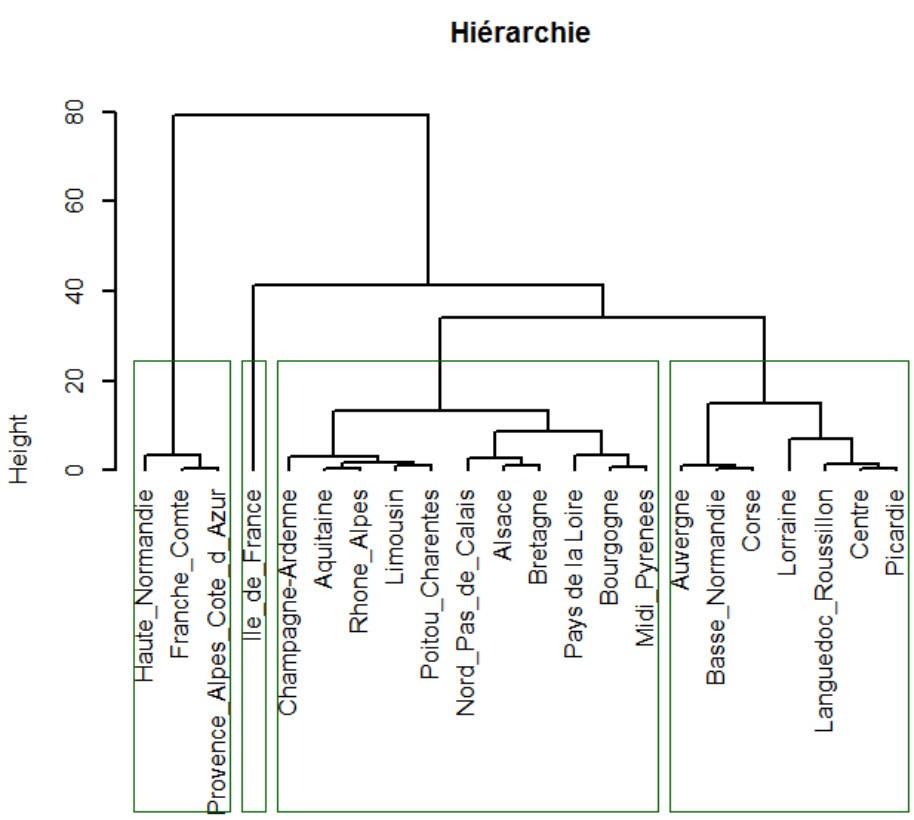
4 classes

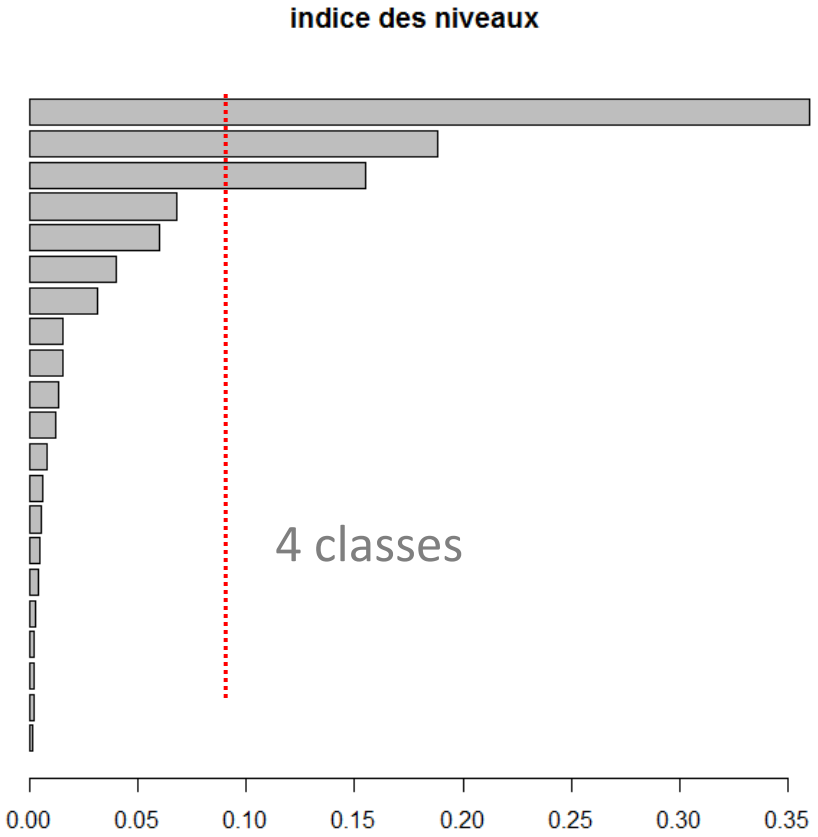


2 classes

La classification Ascendante Hiérarchique

C) application sur données réelles

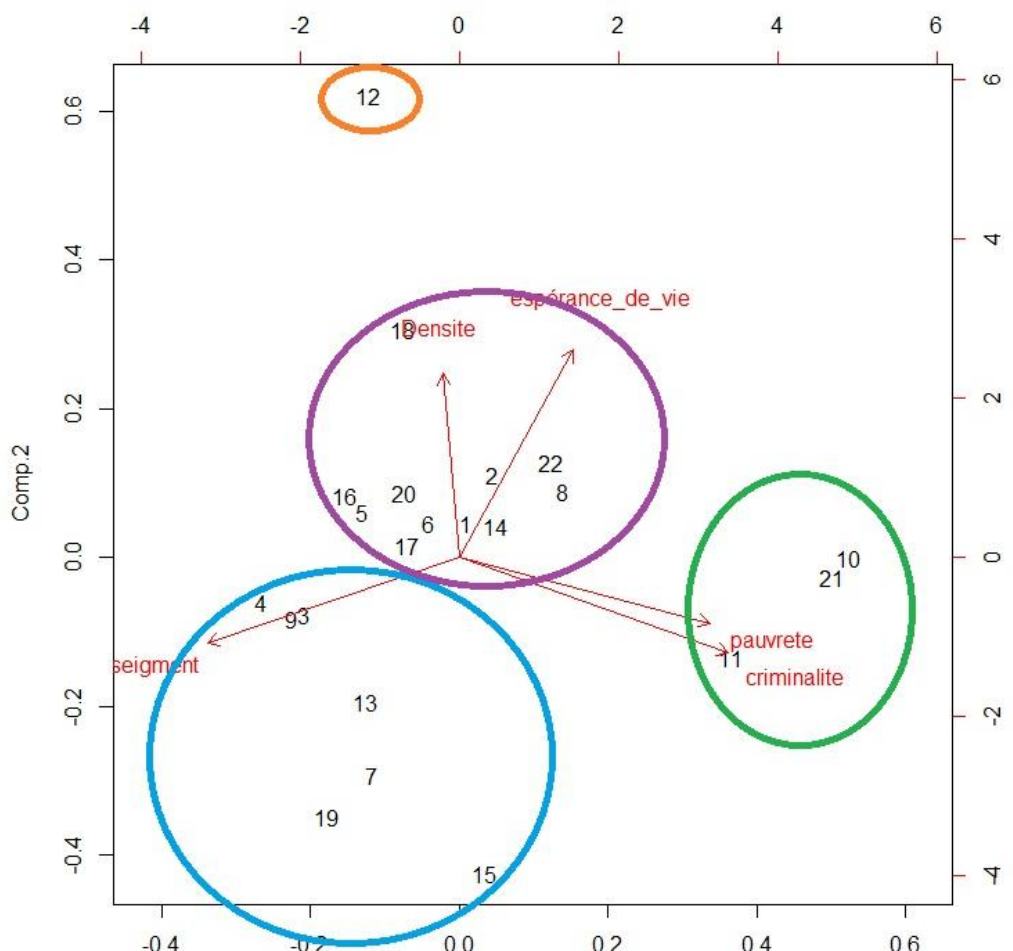




La classification Ascendante Hiérarchique

C) application sur données réelles

1 Alsace
2 Aquitaine
3 Auvergne
4 Basse Normandie
5 Bourgogne
6 Bretagne
7 Centre
8 Champagne-Ardenne
9 Corse
10 Franche Comte
11 Haute_Normandie
12 Ile de France
13 Languedoc Roussillon
14 Limousin
15 Lorraine
16 Midi Pyrenees
17 Nord Pas de Calais
18 Pays de la Loire
19 Picardie
20 Poitou Charentes
21 Provence Alpes Cote d'Azur
22 Rhone Alpes



La classification Ascendante Hiérarchique

D) Limites

- Résultats différents en fonction de la paramétrisation
  - Distances différentes
  - Choix d'agrégation différents
  - Lourdeur des calculs dès qu'on a un nombre de données important

- les regroupements sont définitifs, ce qui ne permet pas d'optimisation postérieure au *clustering*

UNIVERSITÉ DE RENNES 1

Méthode partitionnement

- La structure classificatoire recherchée est la *partition.*

- *Objectif :*

  Trouver, parmi l'ensemble fini de toutes les partitions possibles, une partition qui optimise un critère défini a priori.

- Problème :

  En pratique approche irréalisable, car pour N objets et K classes on a:

  $k^N / K!$ partition possibles.

Méthode partitionnement

- Logique des methodes de partitionnement
  - Une approche typique des methodes de partionnement est l'utilisation de methodes iteratives.
  - Produir une classification par partionnement revient à produir plusieurs classes non vide (leur nombre étant souvent défine à l'avance).
- Critère d'optimisation
  - l'algorithme a pour objectif de minimiser ce critère U defini a priori.

Méthode partitionnement

A) Centres Mobiles

- Critère d'optimisation
  - Différentes approches :
    - ❖ Approche géométrique : une distance.
    - ❖ Approche probabiliste : une vraissemblance.
    - ❖ Approche prototype : une fonction D quelconque qui depend du type de données dont on dispose.

- Approche retenu ici : approche géométrique

$$U = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d^2(\mathbf{x}, \mathbf{m}_i) = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} \| \mathbf{x} - \mathbf{m}_i \|^2$$

Mesure l'homogénéité de chaque classe.

UNIVERSITÉ DE RENNES 1

Méthode partitionnement

A) Centres Mobiles

- Algorithme
  - Etape 1 :

    On choisit aléatoirement k individus comme centres initiaux des classes.
  - Etape 2 :

    On attribue chaque objet à la classe la plus proche, ce qui définit k classes
  - Etape 3 :

    Connaissant les membres de chaque classe on recalcule les centres d'inertie de chaque classe.
  - Etape 4 :

    On redistribue les objets dans la classe qui leur est la plus proche en tenant des nouveaux centre de classe calculés à l'étape précédente.
  - Etape 5 :

    On retourne à l'étape 3 jusqu'à ce qu'il y ai convergence, c'est-à-dire jusqu'à ce qu'il n'y ai plus aucun individu à changer de classe.

Méthode partitionnement

B) Variantes

- K-Means
  - Principe

    le barycentre de chaque groupe est recalculé à chaque nouvel individu introduit dans le groupe, au lieu d'attendre l'affectation de tous les individus.

  - Avantage

    la convergence est parfois possible en une seule itération => plus grande rapidité.

Méthode partitionnement

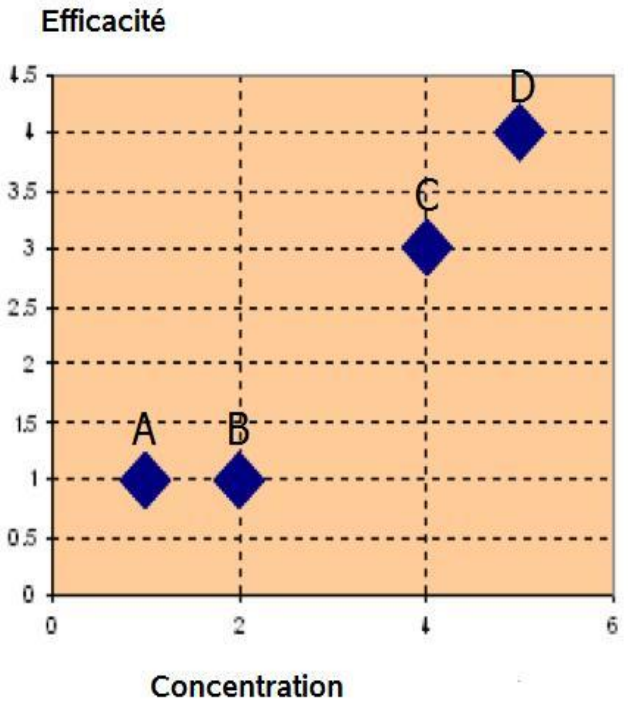
B) Variantes

- K-Means
  - Exemple

4 types de medicaments avec chacun deux modalités

La concentration et l'efficacité, on veut créer deux classes => K=2.

Médicament	Concentration	Efficacité
A	1	1
B	2	1
C	4	3
D	5	4



Méthode partitionnement

B) Variantes

- K-Means
  - Exemple

Etape 1 : On désigne aléatoirement A et B comme centre de classes.

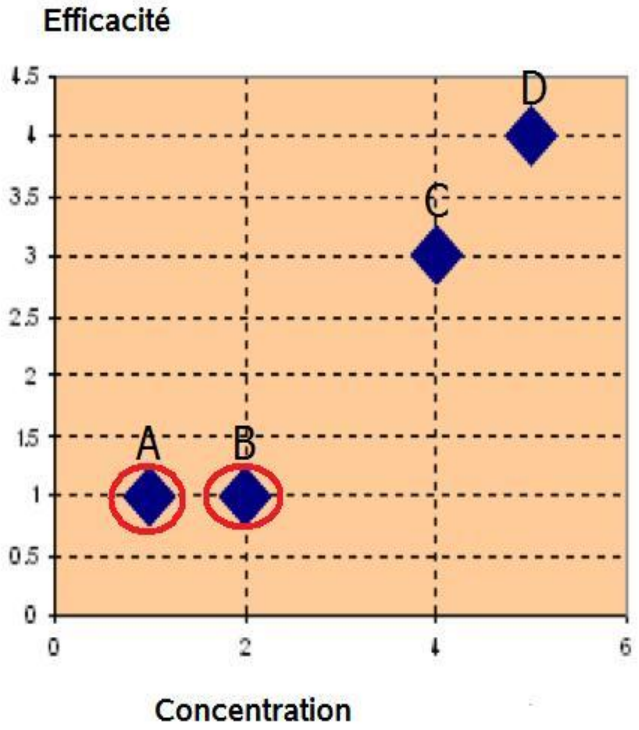
C1 = A *C2* = B

Etape 2 : On assigne chaque point à une des classes.

On commence par D :

$$d(D,c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D,c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$



Méthode partitionnement

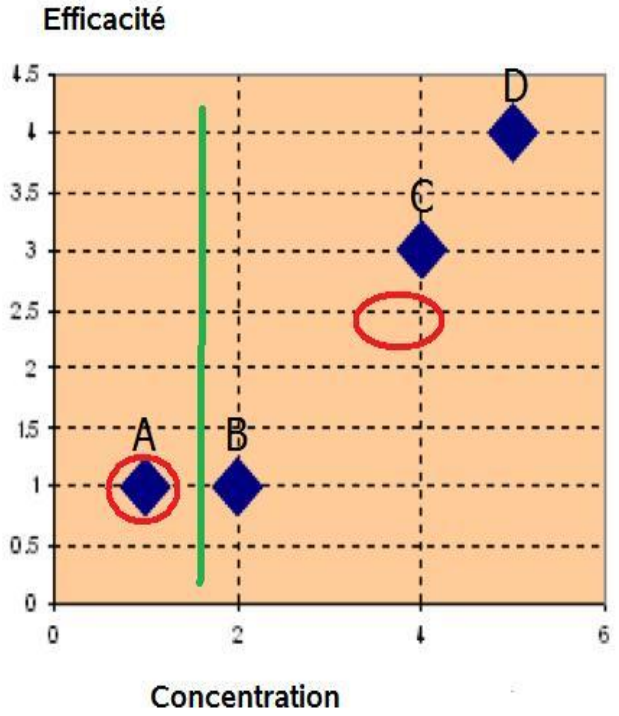
B) Variantes

- K-Means
  - Exemple

Etape 3 : Calcul les nouveaux centres de classe compte tenu de la nouvelle classification.

$$c_1 = (1,\ 1)$$

$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right) = (11/3,\ 8/3) = (3.67,\ 2.67)$$



=> *C1* = (1 , 1) et *C2* = (3.67 , 2.67)

Méthode partitionnement

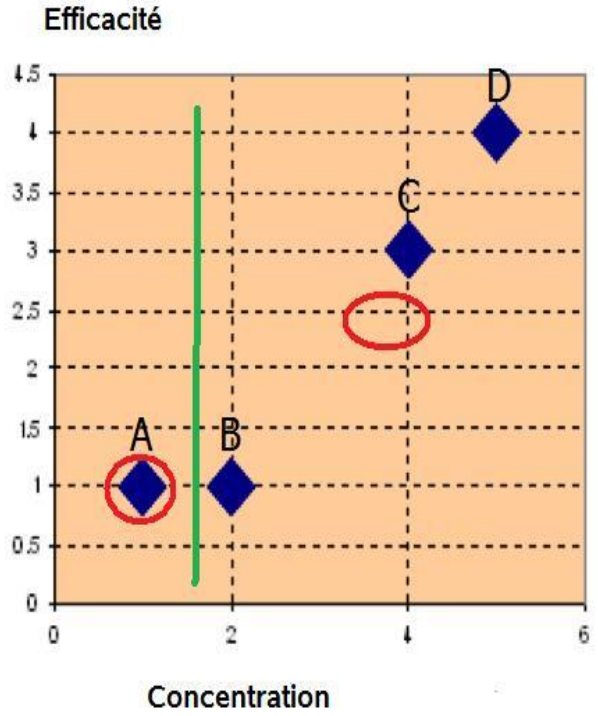
B) Variantes

- K-Means
  - Exemple

Nous voilà à nouveau à l'étape 1. On commence la deuxième itération de l'algorithme.

On réasigne chaque médicament à une classe en calculant la distance les séparant des nouveaux centres de classe .

On repart à l'étape 2.



Méthode partitionnement

B) Variantes

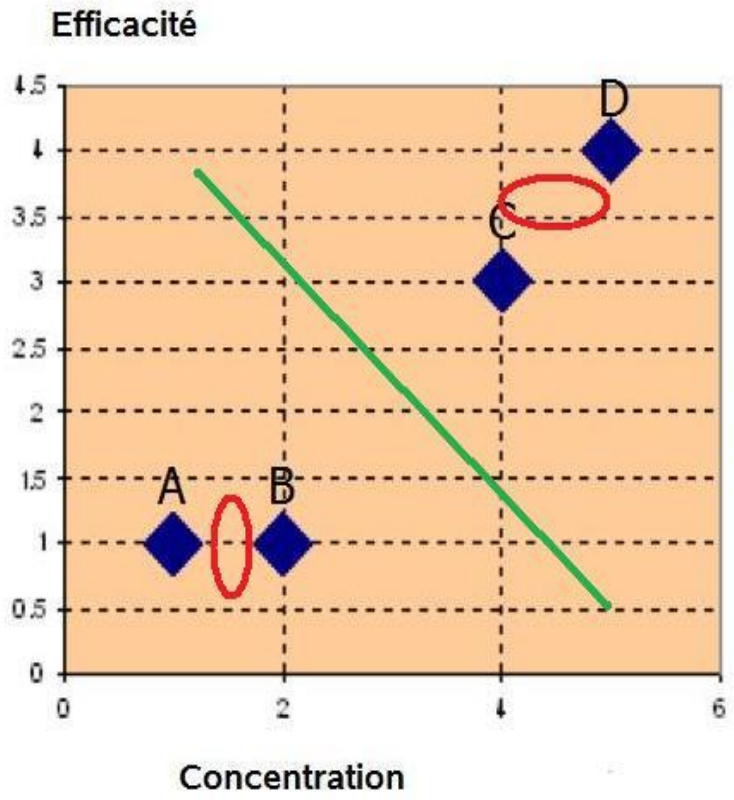
- K-Means
  - Exemple

On répète les étapes jusqu'à convergence.

Connaissant les membres de chaque classe, on recalcule leur centres de classe pour chacun de leur nouveau membre.

$$c_1 = \left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1\frac{1}{2}, 1)$$

$$c_2 = \left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4\frac{1}{2}, 3\frac{1}{2})$$



Méthode partitionnement

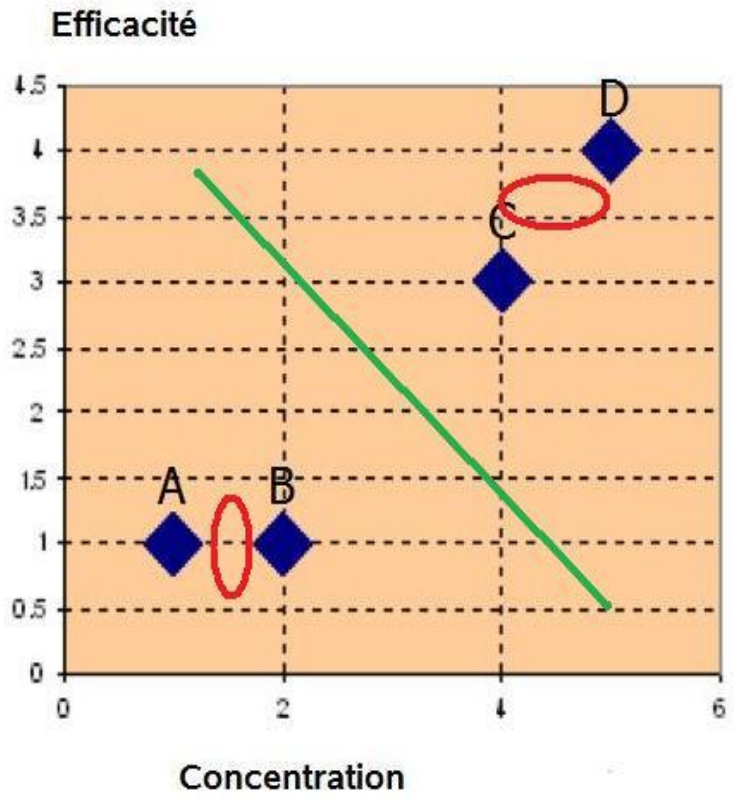
B) Variantes

- K-Means
  - Exemple

Le résultat final est donc:

➢ Classe1 = {A , B} avec comme centre de classe *c1* = (1.5 , 1).

➢ Classe2 = {C , D} avec comme centre de classe *c2* = (4.5 , 3.5).



Méthode partitionnement

B) Variantes

- K-Means
  - Application à nos données

A partir des observations de la classification Ascendante Hiérarchique, on fixe le nombre de classe K = 4.

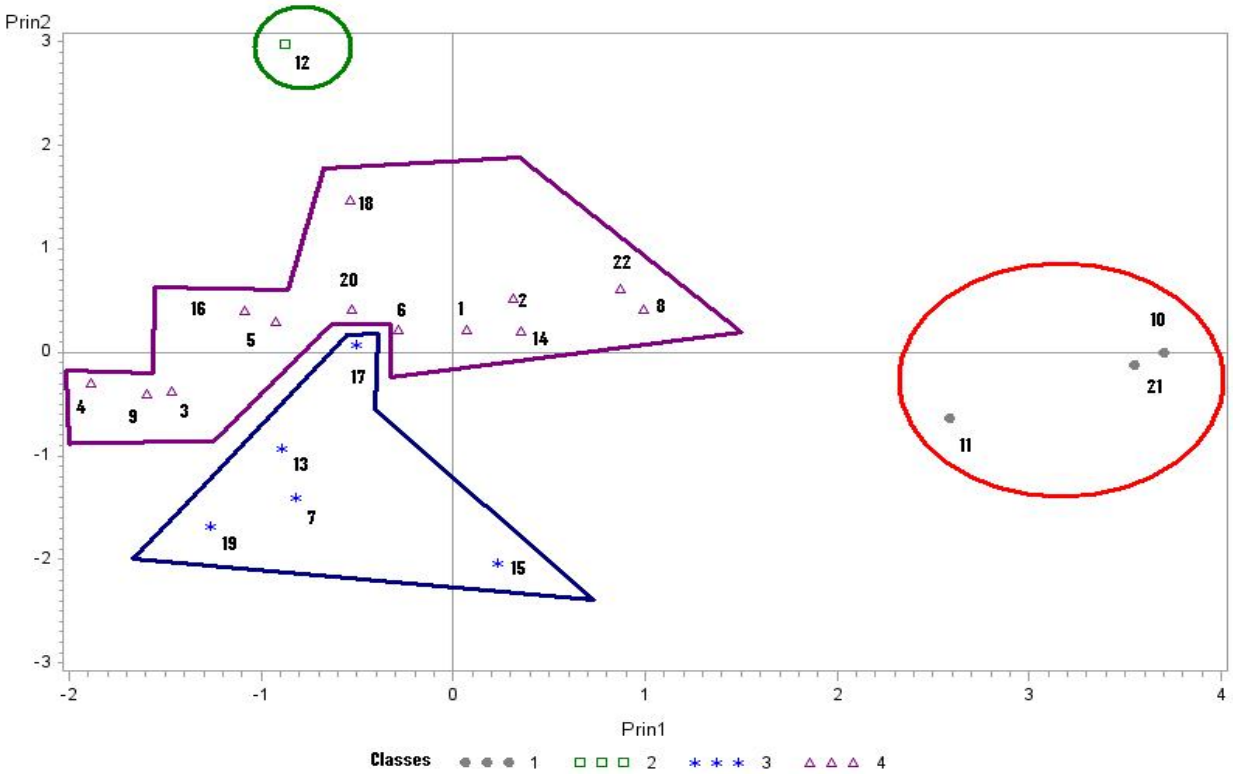
Les centres d'inertie initiaux tirés aux hazard sont :

Cluster	Densite	esperance_vie	enseigment	criminalite	pauvrete
1	0.064956164	0.704186851	-1.838586521	2.454107345	1.864566842
2	4.320667747	0.160604019	0.014522800	-1.274488888	-0.043362020
3	-0.229485969	-2.557310142	-0.304978807	0.870286825	0.338223753
4	-0.311674100	0.024708311	1.931532443	-0.765757453	-0.806533564

Méthode partitionnement

B) Variantes

- K-Means
  - Résultat après 5 itérations



1 Alsace
2 Aquitaine
3 Auvergne
4 Basse Normandie
5 Bourgogne
6 Bretagne
7 Centre
8 Champagne-Ardenne
9 Corse
10 Franche Comte
11 Haute_Normandie
12 Ile de France
13 Languedoc Roussillon
14 Limousin
15 Lorraine
16 Midi Pyrenees
17 Nord Pas de Calais
18 Pays de la Loire
19 Picardie
20 Poitou Charentes
21 Provence Alpes Cote d'Azur
22 Rhone Alpes

Méthode partitionnement

B) Variantes

- Nuée dynamique
  - Principe

    chaque classe n'est plus représentée par son barycentre (éventuellement extérieur a la population), mais par un sous-ensemble de la classe, appelé noyau.

    Le noyau est formés des formes fortes. C'est un petit groupe d'observation qu'on retrouve systématiquement dans chaque classe quelque soit le centres d'inertie initiaux.

  - Avantage

    s'il est bien compose (des individus les plus centraux, par exemple), sera plus représentatif de la classe que son barycentre.

UNIVERSITÉ DE RENNES 1

Méthode partitionnement

C) Limites des méthodes

- Obliger de fixer a priori le nombre de classe.
- Dépendance au choix des centres ou noyaux initiaux.
- Manque de flexibilité
  - Bien adaptée à des données numériques , mais moins flexible que la classification Ascendante Hiérarchique pour des données plus “originales”.

Classification mixte

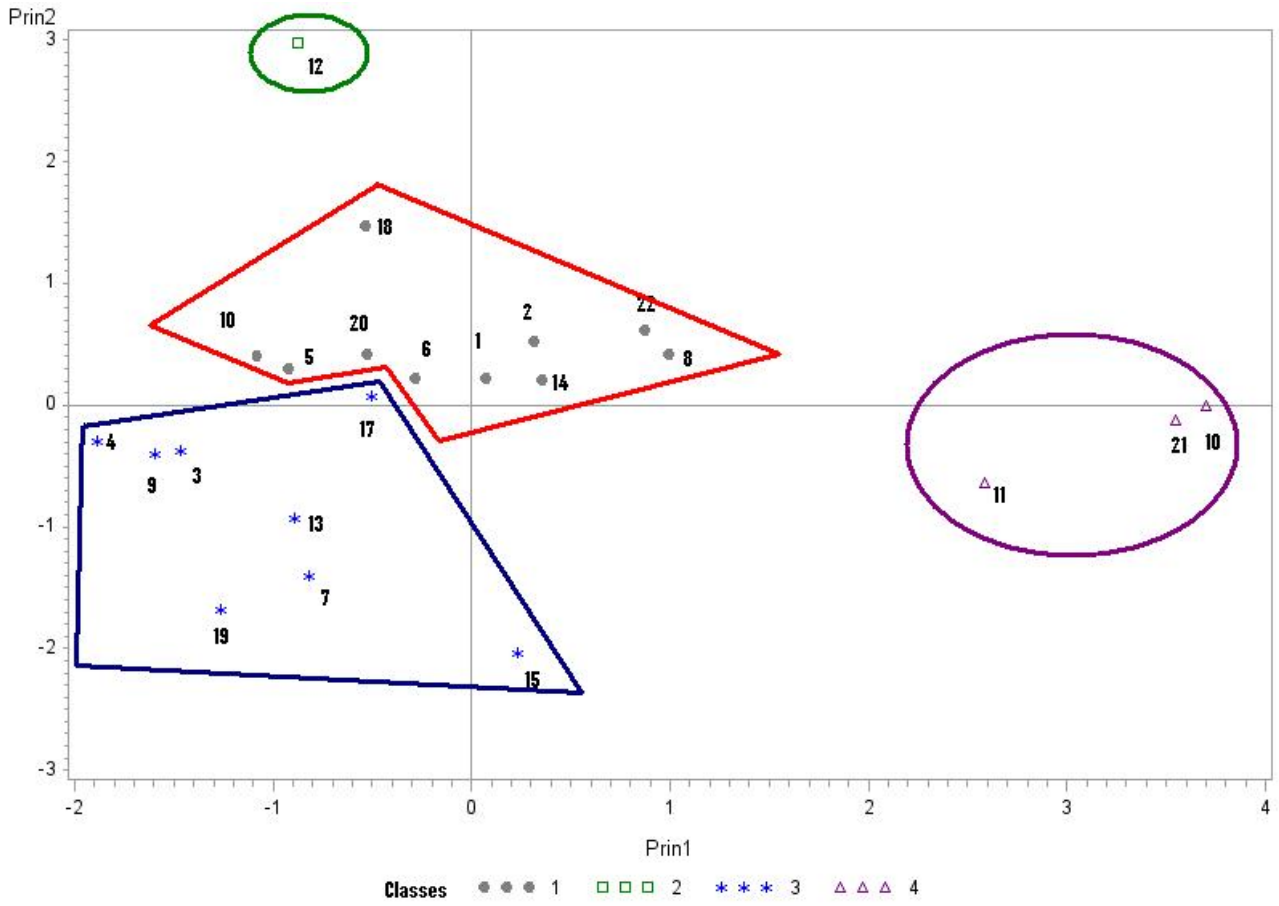
A) présentation

- Objectifs:
  - Combiner les avantages des 2 types de méthodes vues et permettre d'en annuler les inconvénients
- Principe :
  - Réaliseation d'une CAH
    - définie le nombres de classes optimales
    - Donne les barycentres des classes
  - On lance les centres mobiles à partir des barycentres des K classes
    - Obtention d'un optimum local
- Avantage :
  - On ne part de centres de classes définis au hasard
  - On autorise quelques réaffectations individuelles

UNIVERSITÉ DE RENNES 1

Classification mixte

B) application



1 Alsace
2 Aquitaine
3 Auvergne
4 Basse Normandie
5 Bourgogne
6 Bretagne
7 Centre
8 Champagne-Ardenne
9 Corse
10 Franche Comte
11 Haute_Normandie
12 Ile de France
13 Languedoc Roussillon
14 Limousin
15 Lorraine
16 Midi Pyrenees
17 Nord Pas de Calais
18 Pays de la Loire
19 Picardie
20 Poitou Charentes
21 Provence Alpes Cote d'Azur
22 Rhone Alpes

Validation et sélection

A) Validation

- Mesure de la qualité
  - R^2 :proportion de la variance expliquée par les classes

$$R^2 = \frac{I_r}{I} \qquad 0 < R^2 < 1$$

  - Pseudo F = mesure la séparation entre toutes les classes

$$F = \frac{R^2/k - 1}{1 - R^2/n - k}$$

n=observations

k=classes

Validation et sélection

A) Validation

- Mesure de la qualité
  - Cubic clustering criterion (CCC)
  - H0 = Les données sont issues d'une distribution uniforme (pas de classes)

$$CCC = ln\left[\frac{1 - E(R^2)}{1 - R^2}\right] * K$$

K est une constante (voir Sarle (1983))

CCC > 2 : bonne classification

0< CCC<2 : classification peut être OK mais à vérifier

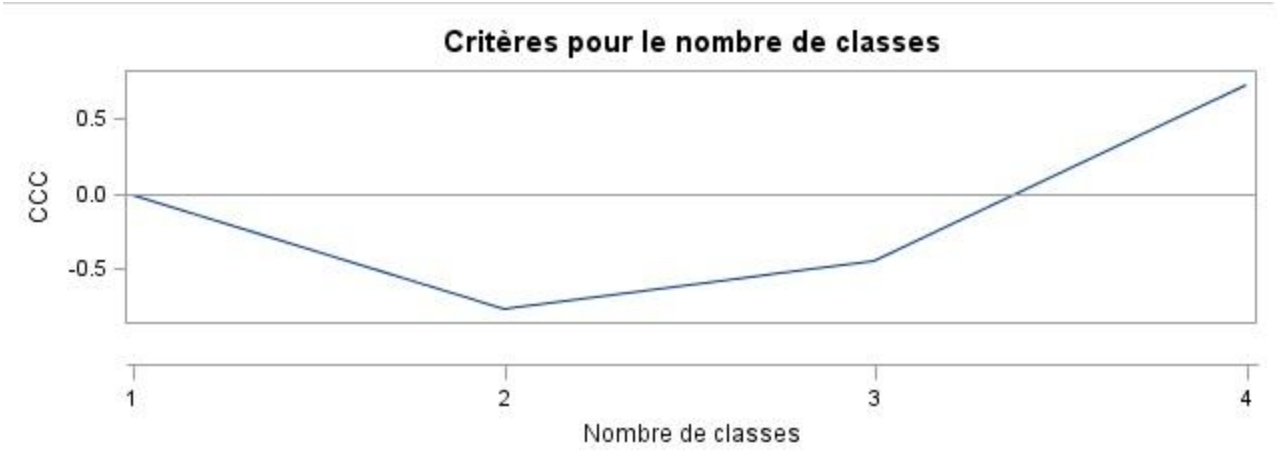
CCC< 0 : présence d'outliers gênant (surtout si CCC < -30)

- On trace CCC versus le nombre de classes. Un creux pour k classes suivi
- d'un pic pour k+1 classes indique une bonne classification en k+1 classes
- (surtout si on a une croissance ou décroissance douce à partir de k+2 classes)

Validation et sélection

A) Validation

- Mesure de la qualité
  - Cubic clustering criterion (CCC)
- On trace le CCC en fonction du nombre de classes.

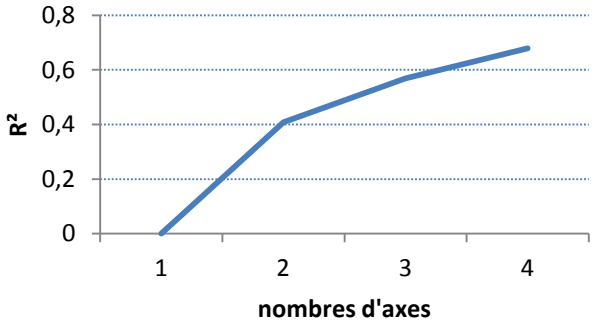


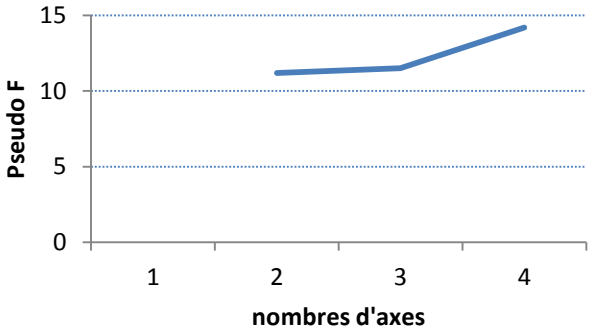
Un creux pour k classes suivi d'un pic pour k+1 classes indique une bonne classification en k+1 classes (surtout si on a une croissance ou décroissance douce à partir de k+2 classes)

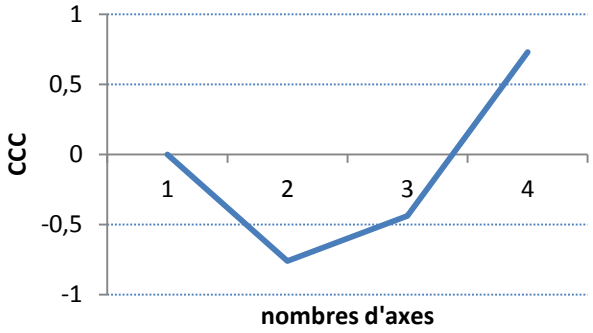
Validation et sélection

B) Sélection sur les exemples

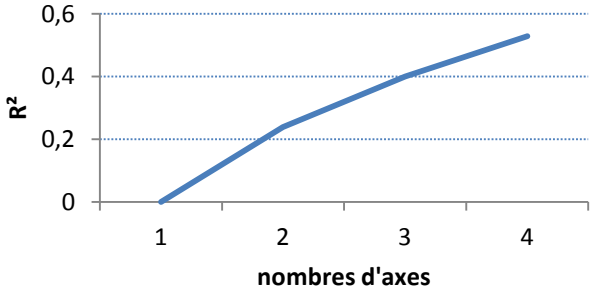
CAH

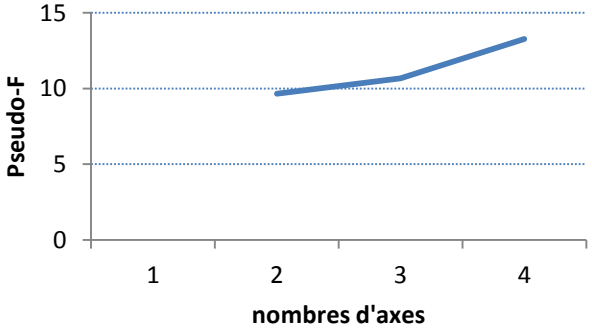


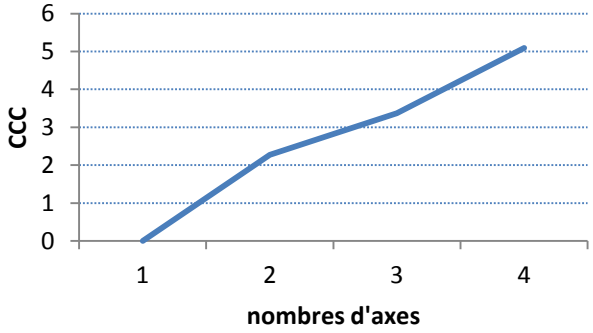




K-Means







UNIVERSITÉ DE RENNES 1

Validation et sélection

B) Sélection sur les exemples

- On compare nos modèle avec 4 classes.

- Comparaison des statistiques

	Pseudo-F	CCC
K-Means	13,6	5,3
Mixte	14,71	5,97

- Mixte est meilleure => pseudo-F le plus grand
  => CCC plus grand

UNIVERSITÉ DE RENNES 1

Conclusion

- Multitude de technique de classification
  - Attention au distance
  - Bien réfléchir à la démarche
  - Ne pas oublier de valider sa classification

- Ouverture
  - Technique de mélange
  - Ouverture de la classification au données multimédias (classification de texts par exemple)