Classification

Exemple : Enquête d'opinion sur les OGM

Pauline Le Badezet

Alexandra Lepage

SOMMAIRE

- Introduction
- Méthodologie
- Méthode de partitionnement
- Classification Ascendante Hiérarchique
- Interprétation des résultats
- Conclusion

I. Introduction

I.1. Contexte de la classification

- Dans quel cas utiliser la classification ?
  - Grand volume de données
  - Besoin de partitionner en sous-ensembles homogènes
  - Traitement et analyse spécifique à chacun d'eux
  - Dans de nombreux domaines : sciences humaines, médecine, marketing…

I. Introduction

I.2. Définition et application

- Définition : Opération statistique qui consiste à regrouper des objets (individus ou variables) en un nombre limité de groupes (classes, segments) qui ont deux propriétés :
  - Homogénéité dans chaque classe et disparité entre les classes
  - Classes non prédéfinies mais découvert au cours de l'opération

I. Introduction

I.2. Définition et application

- Secteur d'application :

  - **Domaine médical** : classification permettant de déterminer des groupes de patients susceptibles d'être soumis à des protocoles thérapeutiques, chaque groupe regroupant tous les patients réagissant identiquement

  - **Domaine commercial** : classification répartissant l'ensemble des magasins d'une enseigne en établissements homogènes d'un point de vue de type de clientèle, CA…

I. Introduction

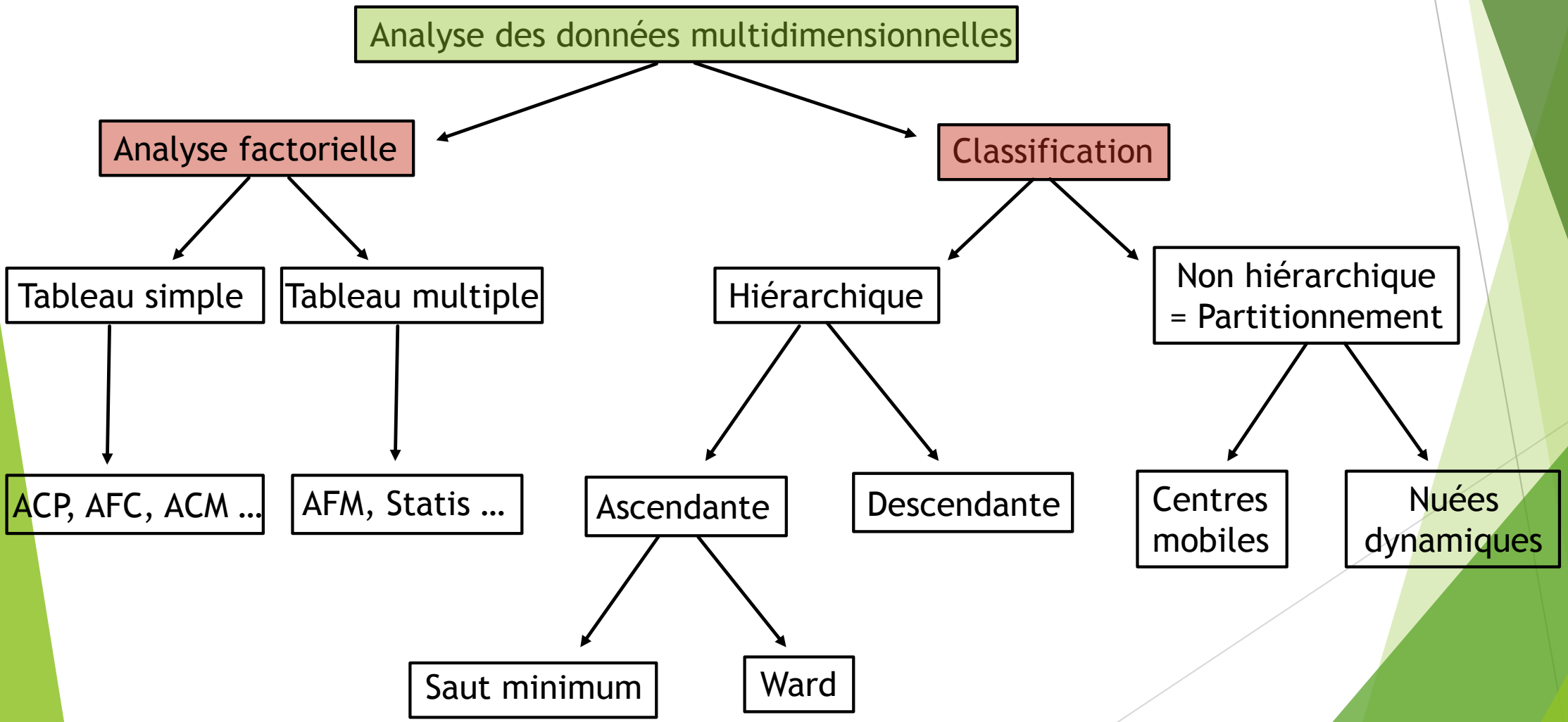
I.2. Définition et application

- **Marketing** :
  - classification appelée plus fréquemment segmentation.
  - Recherche des différents profils de clients constituant la clientèle. Après avoir détecté les classes de la clientèle, l'entreprise peut adapter sa stratégie marketing à chaque profil.
  - Objectif :
    - Identification des prospects les plus susceptibles de devenir clients
    - Identification des clients les plus rentables afin de concentrer sur eux les efforts commerciaux
    - Identification des clients susceptibles de partir à la concurrence
    - Détermination des profils de consommateurs

I. Introduction

I.3. Comparaison avec l'analyse factorielle

- Deux grands types d'analyses de données



I. Introduction

I.3. Comparaison avec l'analyse factorielle

➢ Pourquoi choisir la classification plutôt que l'analyse factorielle ?

Analyse factorielle	Classification
• Manque de lisibilité des plans principaux lorsque le volume de données est important • Seules les premières composantes sont interprétables -> Possible perte d'informations	• Prend en compte toutes les dimensions d'un problème (toutes les variables) donc moins de perte d'information • Simplicité d'interprétation et d'utilisation (description directe des classes) • Représentation graphique plus lisible

II. Méthodologie

II. 1. Les méthodes

- Différentes méthodes de classification :
  - Méthode de partitionnement : deux classes sont toujours disjointes.
    - Centres mobiles, K-means, nuées dynamiques
    - Nombre de classes connu à priori
  - Méthode hiérarchique : deux classes sont disjointes ou l'une contient l'autre.
    - Ascendante
    - Descendante
  - Méthode dite d'analyse floue : deux classes peuvent avoir plusieurs objets en commun (classes « empiétantes » ou « recouvrantes »)
    - Très rare

II. Méthodologie

II. 2. Présentation des données

- Présentation des données : Enquête d'opinion sur les OGM
  - Données issues de l'agrocampus ouest
  - Enquête réalisée sur 135 individus
  - Caractéristiques de l'individu : âge, sexe, CSP, parti politique
  - 15 questions :

    - Concerne
    - Position
    - Manifestation
    - Information active
    - Famine
    - Amélioration agricole
    - Danger
    - Risque écologique
    - Procédé inutile
    - Classement des risques entre :
      - OGM
      - Agriculture intensive
      - Pénurie

II. Méthodologie

II. 2. Présentation des données

Données qualitatives : on réalise dans un premier temps une analyse des correspondances multiples (ACM : proc corresp sous SAS) après avoir discrétiser la variable âge

Avantage de l'ACM : permet de traiter simultanément les variables quantitatives (mises en classe) et qualitatives

→ coordonnées des différents axes factoriels.

- Ici, nous conservons pour la classification 12 axes factoriels
- Etude uniquement sur les individus

II. Méthodologie

II. 3. Inertie

- Inertie totale : moyenne pondérée des carrés des distances des individus au barycentre de la population : **Inertie intraclasse + inertie interclasse**

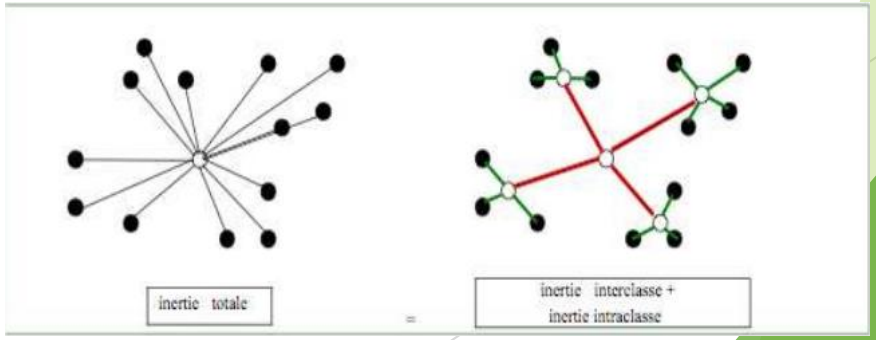
$$I=\sum_{i\in I} p_i\ (xi-\bar{x})^{\,2}$$

- Inertie intraclasse : somme des inerties totales de chaque classe

$$I_A=\sum_{j=1}^{k} I_j$$

- Inertie interclasse : moyenne des carrés des distances des barycentres de chaque classe au barycentre global

$$I_R=\sum_{j\in clases}\left(\sum_{i\in I\,j} p_i\right)\ (\overline{x_j}-\bar{x})^2$$



II. Méthodologie

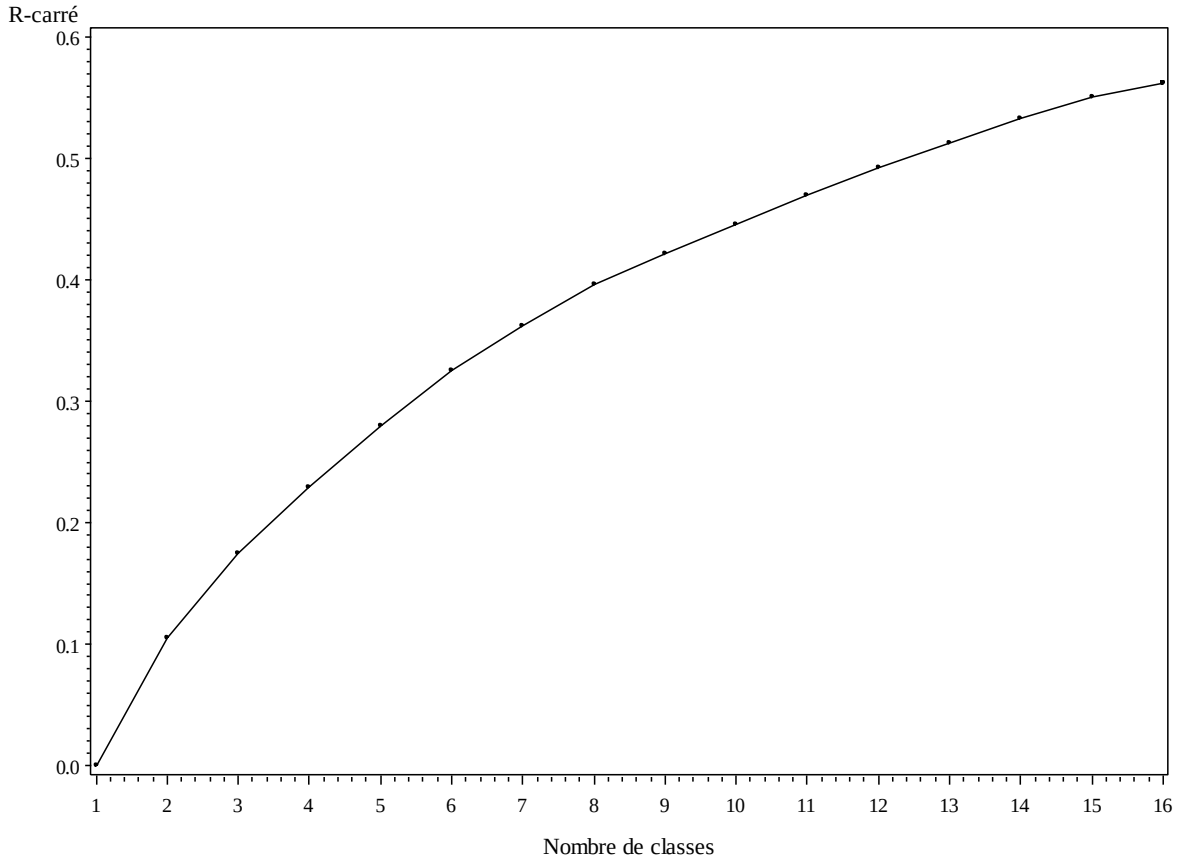
II. 4. Choix du nombre de classes

- Choix du nombre optimal de classe :
  - Trop de classes : segmentation peu utilisable
  - Pas assez de classes : moins de précisions dans les résultats

- Méthodes de partitionnement :
  - Nombre des classes fixées à priori

- Méthodes de classifications hiérarchiques :
  - Représentation graphique du nombre de classes en fonction du R^2
  - Cubic clustering criterion
  - Représentation graphique du nombre de classes en fonction du R^2 semi-partiel

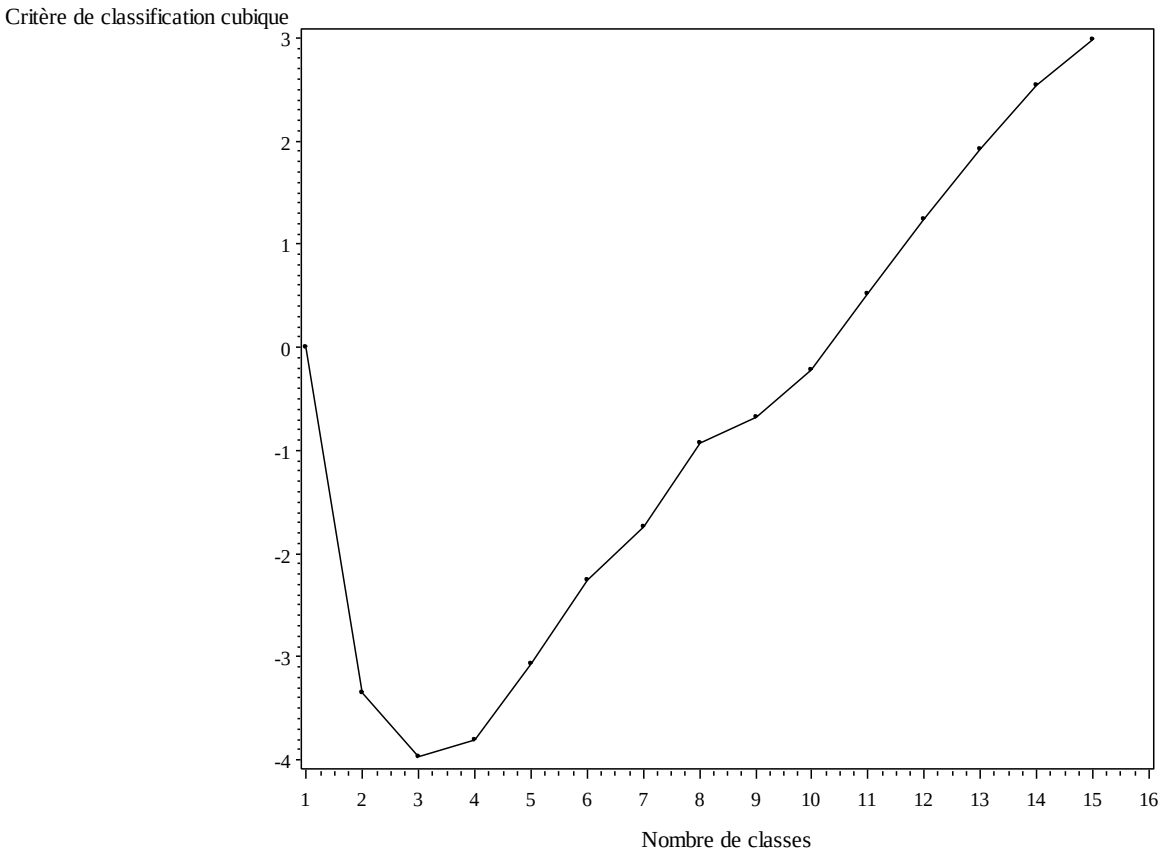
Représentation graphique du nombre de classes en fonction du R^2



- R^2 = proportion de l'inertie expliqué par les classes
- Doit se rapprocher de 1
- Ici, on choisit 3 classes

$$R^2 = \frac{Inertie\ inter}{Inertie\ totale}$$

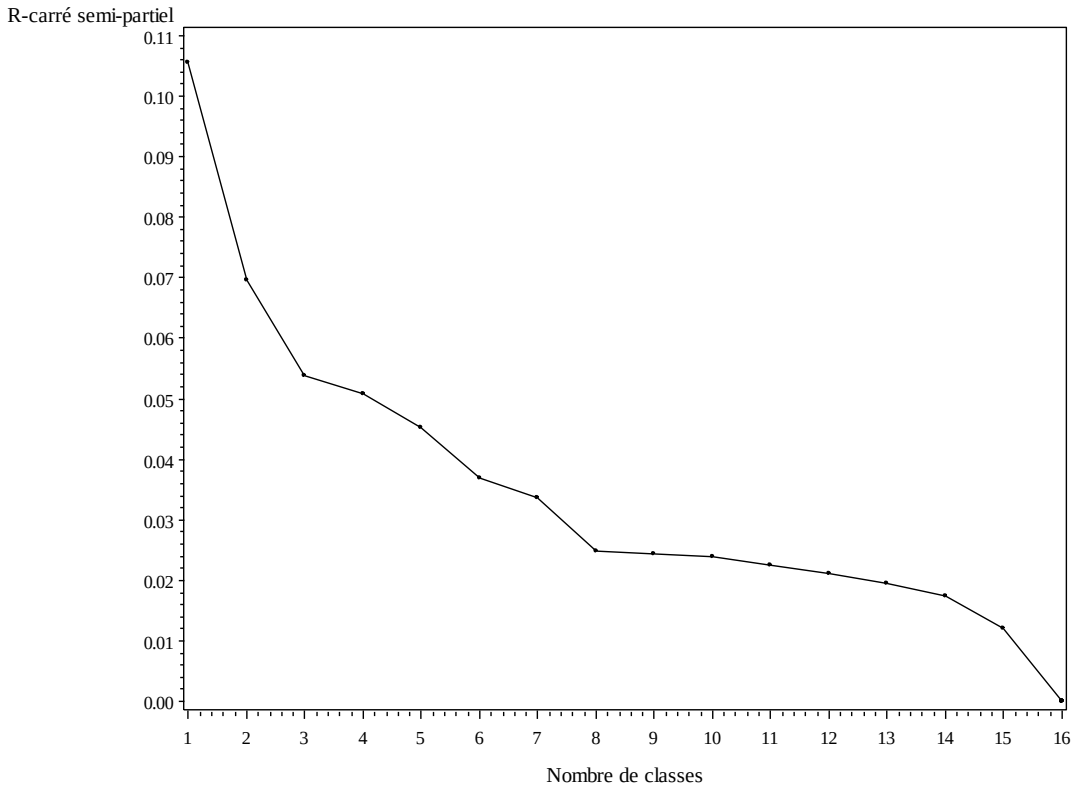
CCC : Critère de classification cubique



- Le CCC indique si la classification est éloignée de celle qui aurait été obtenue par une distribution uniforme du nuage des individus
- CCC>2 : bonne classification
- 0<CCC<2 : classification à vérifier
- CCC<0 : individus hors normes ou petites classes

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

Représentation graphique du nombre de classes en fonction du R^2 semi-partiel



$$SPRSQ = \frac{\Delta\, Inertie\ inter}{Inertie\ totale}$$

- Uniquement pour la CAH
- R^2 semi-partiel mesure la perte d'inertie inter-classe provoquée en regroupant deux classes (baisse du R^2)

III. Méthode de partitionnement

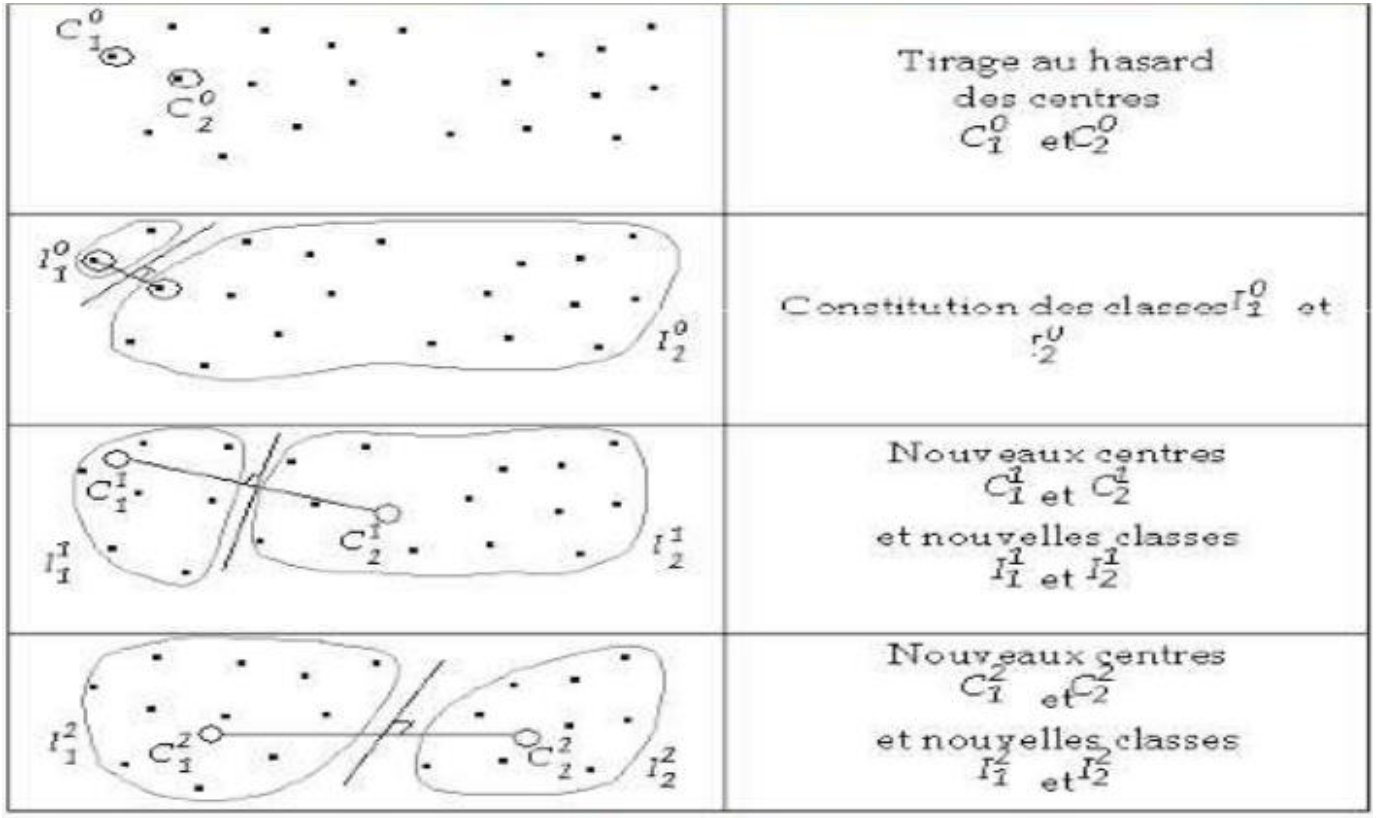
- On part d'une partition arbitraire en K classes que l'on améliore itérativement jusqu'à convergence du critère choisi

- **Méthode des centres mobiles**

  - Etape 1 : on choisit k individus comme centres initiaux des classes
  - Etape 2 : on calcule les distances entre chaque individu et chaque centre et on affecte chaque individu au centre le plus proche → k classes
  - Etape 3 : on remplace les k centres par les barycentres des k classes
  - Etape 4 : on regarde si les centres sont restés suffisamment stables

III. Méthode de partitionnement

- **Méthode des centres mobiles**



III. Méthode de partitionnement

- **K-means**
  - Le barycentre de chaque groupe est recalculé à chaque nouvel individu introduit dans le groupe plutôt que d'attendre l'affectation de tous les individus avant de recalculer les barycentres.

    → Convergence plus rapide

    → Ordre des individus non neutre

- **Nuées dynamiques**
  - Ce n'est plus un seul point qui représente un centre de classe mais un noyau de points

    → Permet de corriger l'influence d'éventuelle valeur extrême

III. Méthode de partitionnement

- Avantages et inconvénients

Avantages	Inconvénients
• Complexité linéaire • Applicable à de grands volumes de données • Permet de détecter les individus isolés ou hors norme • Amélioration continue de la qualité des classes	• Partition finale dépend des choix initiaux : pas d'optimum global • Nombre de classes k fixé • Ne détectent bien que les formes sphériques

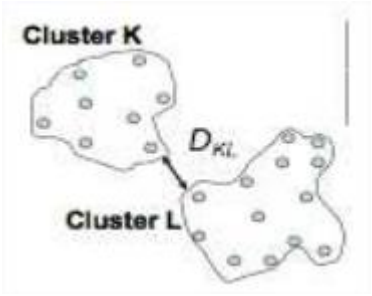
- Procédure sous SAS : fastclus

IV. Classification Ascendante Hiérarchique

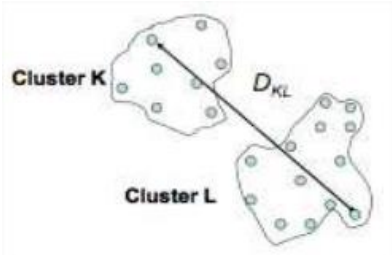
- Algorithme :
  - Etape 1 : Les classes initiales sont les objets
  - Etape 2 : On calcule les distances entre classes
  - Etape 3 : Les deux classes les plus proches sont fusionnées et remplacées par une seule
  - Etape 4 : On reprend à l'étape 2 jusqu'à n'avoir qu'une seule classe

IV. Classification Ascendante Hiérarchique

- Principales distances utilisées
  - Stratégie du saut minimum (single linkage)



  - Stratégie du saut maximum (complete linkage)



  - Méthode de Ward : on affecte les individus qui font le moins varier l'inertie interclasse

$\rightarrow$ d(A,B) = $\frac{d(a,b)^2}{\frac{1}{n_A}+\frac{1}{n_B}}$ avec A et B deux classe de barycentres a et b

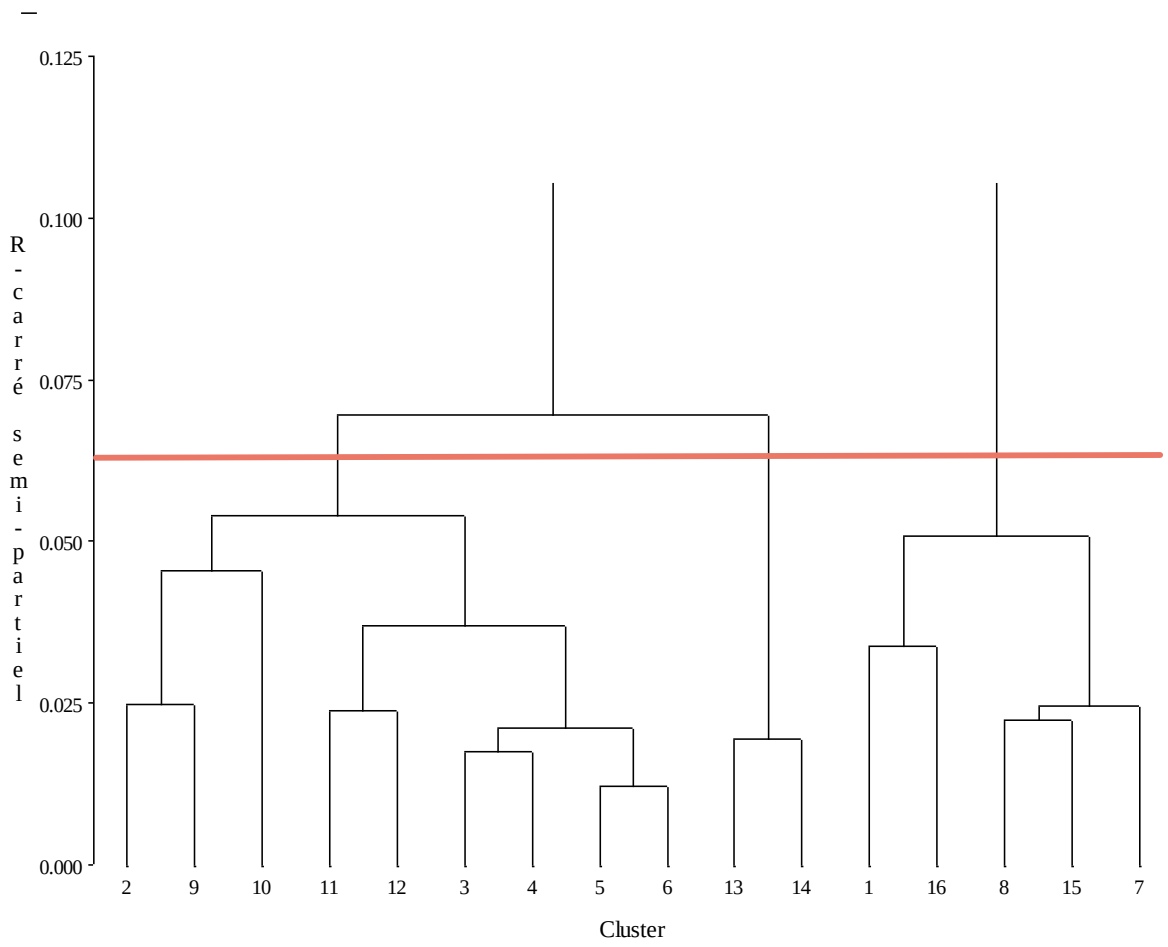
- Procédure SAS : Proc cluster (par défaut, distance de Ward)

IV. Classification Ascendante Hiérarchique

- Dendrogramme :
  - Arbre binaire présentant les agrégations successives, jusqu‘à réunion en une classe unique. La hauteur d'une branche est proportionnelle à la distance entre les objets regroupés. Pour la distance de Ward, la distance est simplement la perte de variance interclasses.
  - Plus l’arbre est coupé bas, plus la classification est fine
  - Une hauteur de coupe est pertinente si elle se trouve entre deux nœuds dont les distances sont relativement élevées.

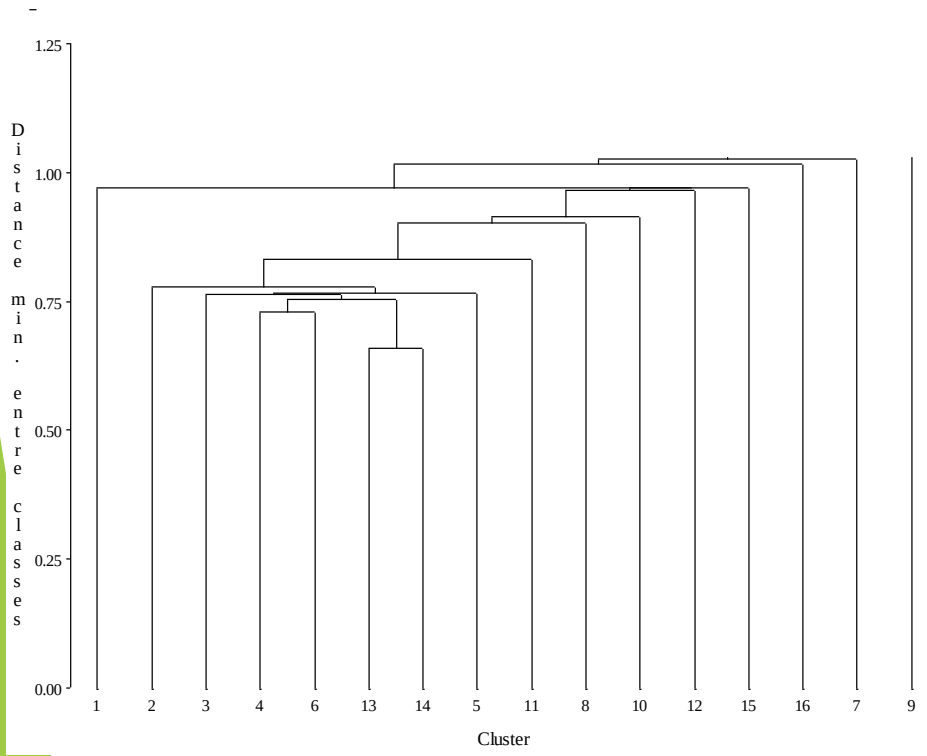
IV. Classification Ascendante Hiérarchique

- Dendrogramme : ⇨ 3 classes

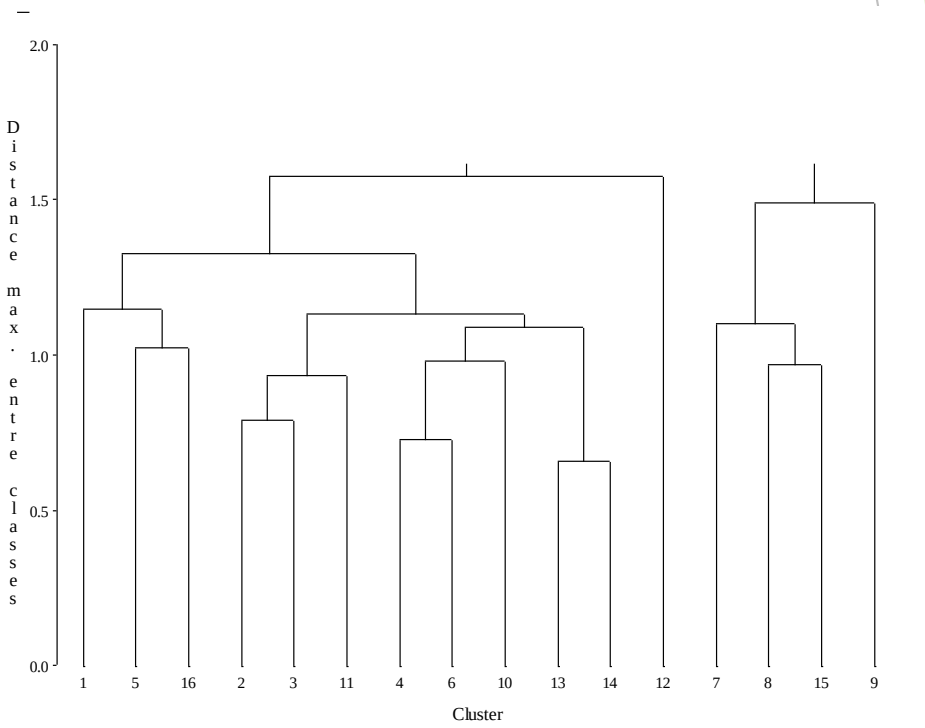


IV. Classification Ascendante Hiérarchique

- Single linkage



- Complete linkage



IV. Classification Ascendante Hiérarchique

- Avantages et inconvénients

Avantages	Inconvénients
• Pas de dépendance au choix de centres initiaux • Pas de fixation à priori du nombre de classes • Détecte des classes de forme diverse • Permet de classer des individu, des variables ou des centres de classes	• Complexité algorithmique (non linéaire) • A chaque étape, le critère de partitionnement n'est pas global mais dépend des classes déjà obtenues

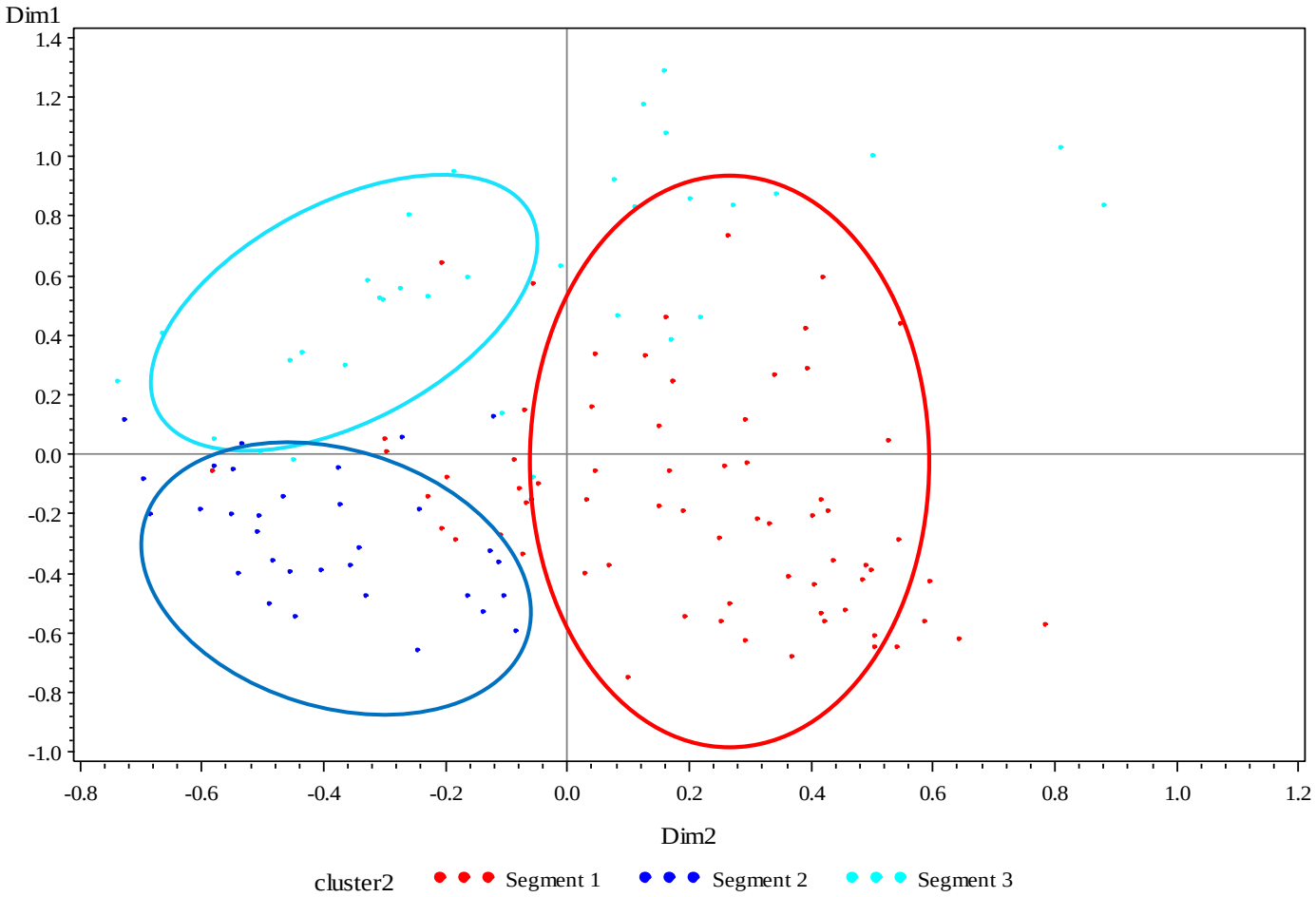
V. Méthode mixte

- Combine les points forts de :
  - La CAH : précision et absence d'à priori
  - Le partitionnement : rapidité

- Etape 1 : Première classification sur les n observations avec centres mobiles ou k-means

- Etape 2 : Classification ascendante Hiérarchique sur les centres de ces classes

V. Interprétation des résultats



Segment	Fréquence	Pourcentage	Fréquence cumulée	Pctage. cumulé
Segment 1	69	51.11	69	51.11
Segment 2	33	24.44	102	75.56
Segment 3	33	24.44	135	100.00

V. Interprétation des résultats

	Parti Politique					
	Centre	Extrême gauche	PPS	UMP	Verts	Total
Segment 1	14 10.37 20.29 43.75	4 2.96 5.80 44.44	19 14.07 27.54 40.43	30 22.22 43.48 75.00	2 1.48 2.90 28.57	69 51.11
Segment 2	13 9.63 39.39 40.63	2 1.48 6.06 22.22	12 8.89 36.36 25.53	6 4.44 18.18 15.00	0 0.00 0.00 0.00	33 24.44
Segment 3	5 3.70 15.15 15.63	3 2.22 9.09 33.33	16 11.85 48.48 34.04	4 2.96 12.12 10.00	5 3.70 15.15 71.43	33 24.44
Total	32 23.70	9 6.67	47 34.81	40 29.63	7 5.19	135 100.00

CSP	Segment 1	Segment 2	Segment 3	Total
Autre	7	0	2	9
Cadre	13	0	4	17
Commerçant	2	0	1	3
Etudiant	20	5	8	33
Fonction Publique	3	0	6	9
INSFA	2	28	6	36
Inactif	4	0	0	4
Liberal	2	0	1	3
Ouvrier	0	0	1	1
Retraite	13	0	1	14
Technicien	3	0	3	6
Total	69	33	33	135

V. Interprétation des résultats

	Opinion sur les OGM				
	Pas Favorable du Tout	Plutôt Défavorable	Favorable	Très Favorable	Total
Segment 1	13 9.63 18.84 39.39	21 15.56 30.43 38.89	**32** 23.70 46.38 71.11	3 2.22 4.35 100.00	69 51.11
Segment 2	1 0.74 3.03 3.03	**23** 17.04 69.70 42.59	9 6.67 27.27 20.00	0 0.00 0.00 0.00	33 24.44
Segment 3	**19** 14.07 57.58 57.58	**10** 7.41 30.30 18.52	4 2.96 12.12 8.89	0 0.00 0.00 0.00	33 24.44
Total	33 24.44	54 40.00	45 33.33	3 2.22	135 100.00

	Concerner				
	Pas du Tout	Un Peu	Moyen	Beaucoup	Total
Segment 1	15 11.11 21.74 100.00	14 10.37 20.29 45.16	**24** 17.78 34.78 45.28	**16** 11.85 23.19 44.44	69 51.11
Segment 2	0 0.00 0.00 0.00	**10** 7.41 30.30 32.26	**18** 13.33 54.55 33.96	5 3.70 15.15 13.89	33 24.44
Segment 3	0 0.00 0.00 0.00	7 5.19 21.21 22.58	**11** 8.15 33.33 20.75	**15** 11.11 45.45 41.67	33 24.44
Total	15 11.11	31 22.96	53 39.26	36 26.67	135 100.00

VI. Conclusion

- Deux types de classification
  - Partitionnement
  - Hiérarchique
- Importance du nombre de classes
- Minimisation de l'inertie intra-classe et maximisation de l'inertie inter-classe
- Importance du choix des distances pour la CAH