# Classification de variables

## Application à la Base Permanente des Équipements de 2007

*Brigitte GELEIN (Ensai), Olivier SAUTORY* (*Cepe*)

# La classification de variables - pourquoi ?

- ***Description d'un ensemble de variables***
  - Identifier des groupes de variables bien corrélées
  - Regrouper ces variables par un algorithme automatique, et non par un procédé "visuel" = complément de l'ACP
- ***Réduction du nombre de variables***

  Représenter chaque classe par
  - une nouvelle variable synthétique
  - ou par celle des variables analysées qui représente le mieux les classes (sélection de variables).

# La classification de variables - comment ?

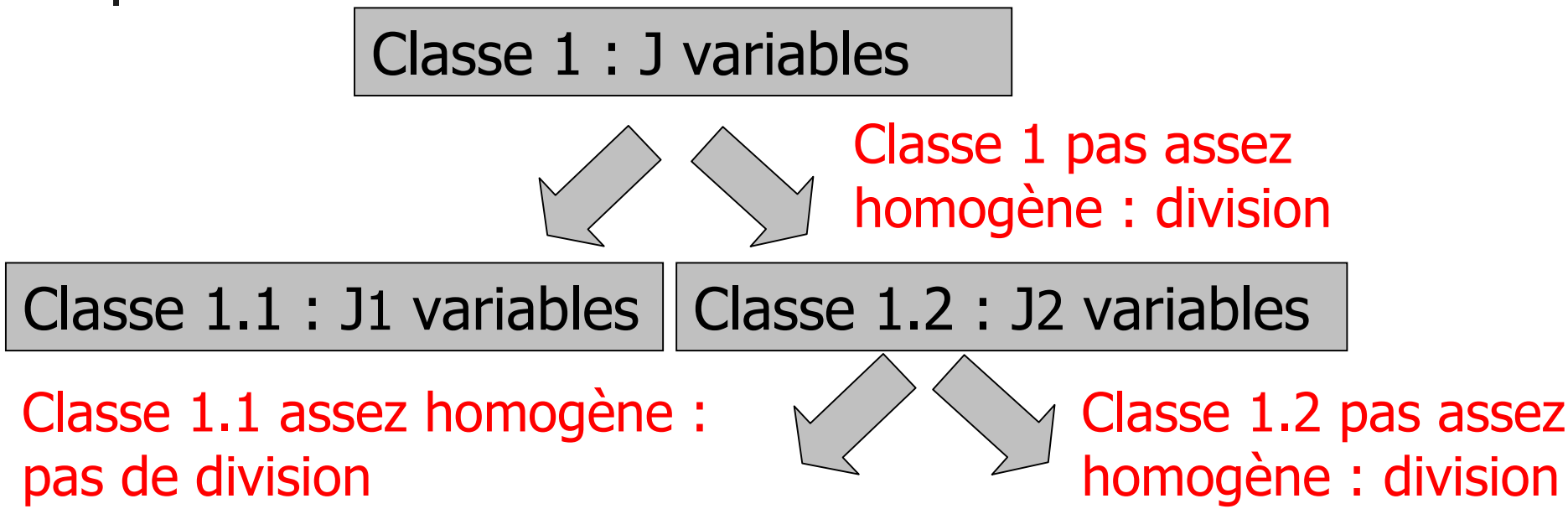
- Il existe des algorithmes de classification :
  - Ascendants
  - Descendants
  - De partitionnement direct
- Toutes ces méthodes sont fondées sur la notion de corrélation
- Dans le cas des algorithmes ascendants, on définit une mesure de dissimilarité, par exemple : $1 - r$ ou $1 - |r|$ ou $1 - r^2$

  en notant r le coefficient de corrélation linéaire.

# Proc VARCLUS de SAS
# La méthode

- **Caractéristiques de la méthode :**
  - Une **méthode descendante**, fondée sur un critère de division d'un groupe de variables en deux classes.
  - Les variables peuvent être numériques ou binaires.
  - Chaque classe est représentée par une combinaison linéaire des variables de la classe, appelée *composante*,
  - On maximise la somme, sur l'ensemble des classes de la partition, des variances de ces composantes.

# Proc VARCLUS de SAS
# La méthode




L'algorithme s'arrête si aucune classe ne peut être divisée => une partition de l'ensemble des variables en un nombre K de classes **disjointes**.

# Proc VARCLUS de SAS
# La méthode

- **Les partitions peuvent être :**
  - emboîtées (option hierarchy) => arbre
  - ou non (option par défaut)

- **La méthode met en œuvre des analyses en composantes principales (ACP) sur des groupes de variables :**
  - réduites (option par défaut)
  - ou non (option covariance)

# Proc VARCLUS de SAS
# Caractéristiques d'une classe

- **Une classe composée des variables $Y_j$ est représentée par une composante notée C = combinaison linéaire des variables :**
  - C est la 1ère composante principale des variables $Y_j$ (option par défaut),
  - C est la moyenne arithmétique des $Y_j$ (option centroid).

# Proc VARCLUS de SAS
# Division d'une classe

- **Une classe est divisée en deux si elle n'est pas suffisamment homogène :**

  si la *composante* représentant la classe ne "résume" pas elle seule l'ensemble des variables de la classe, au sens d'un certain critère.

  => Une deuxième *composante* est nécessaire pour représenter les variables de la classe, qui doit donc être divisée en deux.

# Proc VARCLUS de SAS
# Division d'une classe

- ***<u>Premier critère possible pour décider de diviser ou non une classe :</u>***
  - diviser la classe en deux si sa *2ème valeur propre* est supérieure à un certain seuil λ

=> critère λ (λ =1 par défaut – Kaiser )

# Proc VARCLUS de SAS
# Division d'une classe

- ***Deuxième critère possible pour décider de diviser ou non une classe :***
  - diviser la classe en deux si le ratio :

**variance expliquée par la *composante* C**
**variance de la classe**

est inférieur à un certain seuil *p*

Où la variance de la classe est égale à la somme des variances des variables

=> critère p (p=75% par défaut)

# Proc VARCLUS de SAS
# Division d'une classe

- ***Le choix de la classe à diviser dépend du critère de division utilisé :***
  - avec le ***critère λ***, on sélectionne la classe ayant la plus forte deuxième valeur propre $\lambda_2$.
    Si $\lambda_2 \leq \lambda$, l'algorithme s'arrête, sinon, la classe est divisée
  - avec le ***critère p*** : on sélectionne la classe ayant la plus petite part de variance expliquée par sa *composante*.
    Si **Var Composante / Var Classe > *p*** l'algorithme s'arrête, sinon, la classe est divisée.

# Proc VARCLUS de SAS
# Division d'une classe

- ***Initialisation du processus de division de la classe :***
  - On réalise une analyse en composantes principales (ACP) sur les variables de la classe à diviser.
  - On effectue sur les deux premières composantes principales une ***rotation orthoblique*** **:**
    **=>** obtenir deux nouvelles composantes plus facilement interprétables en fonction des variables initiales, car mieux corrélées (en valeur absolue) avec certaines de ces variables.

# Proc VARCLUS de SAS
# Affectation des variables

***L'affectation des variables dans les classes :***

- 1ère phase : *NCS* (*nearest component sorting*)

1. Chaque variable est affectée à la classe dont la *composante* est la plus corrélée avec la variable (au sens du carré du coefficient de corrélation linéaire $r^2$).
2. On calcule la *composante* de chacune des nouvelles classes ainsi constituées, et on réaffecte chaque variable à la classe dont la *composante* est la plus corrélée avec la variable (au sens du $r^2$).
3. Le processus est itéré jusqu'à ce que la composition des classes ne varie plus.

# Proc VARCLUS de SAS
# Affectation des variables

- 2ème phase : *Search*
  - À l'issue de la 1ère phase, on teste chaque variable pour voir si l'affectation de cette variable à une autre classe augmente la *variance expliquée par la partition*.
  - Si c'est le cas, on change donc la variable de classe, la *composante* de chacune des deux classes concernées par le transfert est recalculée avant le test de la variable suivante.
- Avec l'option **hierarchy**, ces deux phases d'affectation ne concernent que les variables des nouvelles classes

# Application à la Base Permanente des Équipements

- L'accès aux équipements et aux services constitue une problématique importante pour les acteurs locaux : **l'attractivité d'un territoire**.
  - On peut décrire les communes avec des **variables binaires** de présence-absence des différents équipements.
  - On peut décrire les communes avec des **variables quantitatives** représentant le temps nécessaire pour accéder aux équipements les plus proches.

# Application à la Base Permanente des Équipements

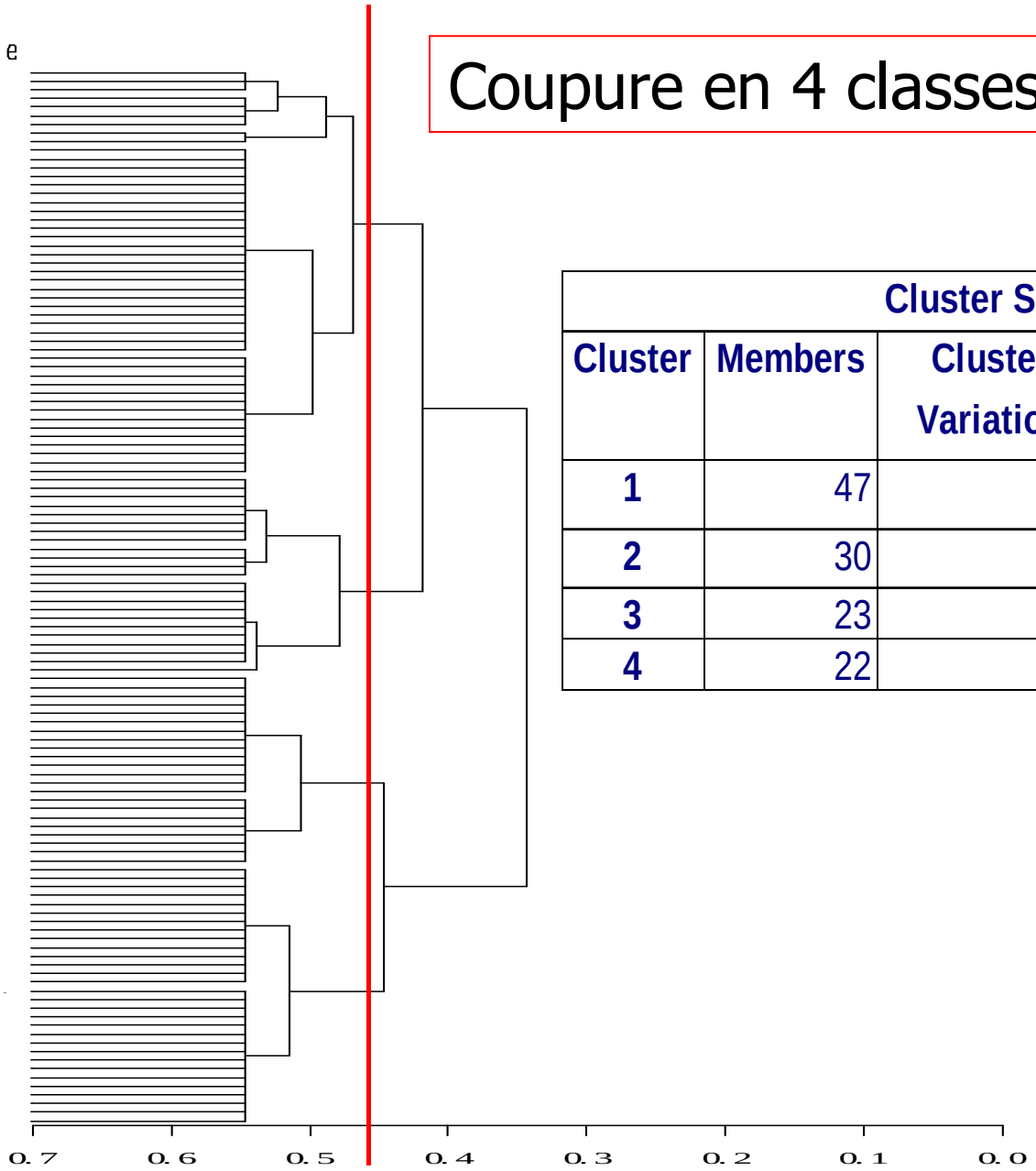
- Construite à l'aide de la BPE de 2007, la table utilisée comporte :
  - en ligne les communes (36 000 environ) traitées comme unités statistiques
  - en colonne les équipements (plus de 120 catégories d'équipements) traités comme des **variables qualitatives binaires** (présence de l'équipement modalité = 1 ou absence modalité = 0)
- Recours à la classification de variables pour
  - réduire le nombre de descripteurs de nos communes,
  - repérer des grands groupes d'équipements.

# Application à la Base Permanente des Équipements

- **<u>Choix d'une structure hiérarchique</u>** car :
  - en termes de variance expliquée totale les performance de l'algorithme non hiérarchique étaient très proches de celles de l'algorithme hiérarchique.
  - la création de gammes d'équipements s'apparente à celle d'une nomenclature. Or, dans le domaine des nomenclatures, on peut souhaiter disposer de plusieurs niveaux de détails. Les différents niveaux sont alors imbriqués les uns dans les autres de façon hiérarchique.
- Toutes les autres options de **VARCLUS** par défaut

# Application à la Base Permanente des Équipements

- L'algorithme de classification descendante hiérarchique s'arrête de lui-même à 13 classes d'équipements.
  - **Dans une logique de réduction de dimensions**, on retiendrait ces 13 classes : chaque classe étant bien représentée par une seule variable synthétique (la première composante).
  - **Dans l'optique de description**, on souhaite un résumé un peu plus synthétique de la réalité avec un niveau de division moindre. **On accepte donc que les classes de variables ne soient pas forcément unidimensionnelles => 4 classes**




# Coupure en 4 classes

| Cluster Summary for 4 Clusters | | | | | |
|---|---|---|---|---|---|
| Cluster | Members | Cluster Variation | Variation Explained | Proportion Explained | Second Eigenvalue |
| 1 | 47 | 47 | 22.12561 | 0.4708 | 1.5126 |
| 2 | 30 | 30 | 16.02305 | 0.5341 | 1.1133 |
| 3 | 23 | 23 | 8.238803 | 0.3582 | 1.2973 |
| 4 | 22 | 22 | 10.838 | 0.4926 | 1.1814 |

Proportion de variance expliquée

# Application à la Base Permanente des Équipements

- **Gamme « proximité »**
  - Boulangerie
  - Bureau de poste
  - Banque caisse d'épargne
  - Coiffure
  - Boucherie charcuterie
  - Médecin omnipraticien
  - Chirurgien dentiste
  - Masseur kiné.
  - Pharmacie
  - ...
- **Gamme « intermédiaire »**
  - Ecole de conduite
  - Vétérinaire
  - Blanchisserie teinturerie
  - Supermarché
  - Magasin de chaussures
  - Collège
  - Orthophoniste
  - Pédicure-podologue
  - Laboratoire d'analyses médicales
  - ...

# Application à la Base Permanente des Équipements

- **Gamme « supérieure »**

  Spécialistes en :
  - Cardiologie
  - Dermatologie - vénérologie
  - Gastro-entérologie
  - Ophtalmologie
  - Oto-rhino-laryngologie
  - Radio Imagerie médicale
  - Pédiatrie

  Lycée d'enseignement :
  - général et/ou technologique
  - professionnel
  - ...

- **Gamme « métropolitaine »**
  - Formation commerce
  - UFR
  - Institut universitaire
  - Autre enseignement supérieur
  - Résidence universitaire
  - Restaurant universitaire
  - ...

# Application à la Base Permanente des Équipements

| Population totale 2006 | Classe Proximité | Classe Intermédiaire | Classe Supérieure | Classe Métropolitaine |
|---|---|---|---|---|
| Pearson Correlation Coefficients | 0.35 | 0.51 | 0.72 | 0.80 |
| RV Coefficients | 0.12 | 0.27 | 0.52 | 0.62 |

- Le niveau commune peut bien sûr faire l'objet d'une discussion.
- On pourrait également exclure de l'analyse certains équipements trop atypiques et les traiter en éléments supplémentaires.

# Application à la Base Permanente des Équipements

L'algorithme divisif de la procédure **varclus** pour**:**

- s'affranchir des limites en capacités de calcul inhérentes aux algorithmes ascendants.
- avoir une aide à la décision quant au choix du nombre de classes

**Un grand MERCI à**

- Suzanne Faudon (PSAR Analyse Territoriale)
- Christophe Barret (PSAR Analyse Territoriale)
- Cyrille Van Puymbroeck (PSAR Synthèses locales)