Fondements statistiques pour la science des données

TP n°4. Classification ascendante hiérarchique (CAH)

1. CAH d'un petit jeu de données

1. Créer le jeu de données des cinq individus décrits par deux variables (exemple du cours).
2. Représenter le nuage de points associé.
3. À l'aide de Rcmdr, réaliser la CAH de ces individus avec les choix suivants : distance entre individus = city-block ; indice d'agrégation = lien minimum.
4. Résumer la classification hiérarchique obtenue
5. Ajouter les classes au jeu de données puis représenter à nouveau le nuage de points en affichant la classe d'appartenance de chaque individu.

2. Températures mensuelles dans 15 villes de France

On dispose des températures mensuelles de 15 villes de France (fichier `temperature.txt`). L'objectif est d'établir une partition de ces 15 villes en fonction de leur profil de température à l'aide d'une classification ascendante hiérarchique (CAH).

On réalisera la CAH selon les deux démarches suivantes :

(A.) Classification directement sur le tableau des données
(menu de Rcmdr : Statistique → Analyse multivariée).

(B.) Classification sur les coordonnées factorielles
(via l'ACP du tableau de données initial à l'aide de FactoMineR.)

(A.) CAH Directe

1. La réduction des variables est-elle obligatoire ? Centrer réduire les variables *Janvier* à *Décembre* : menu = Données → Gérer les variables → standardiser les variables
2. Réaliser la CAH des données centrées-réduites avec les choix suivants : distance = euclidienne ; méthode de classification = indice d'agrégation de Ward.
3. La commande suivante indique l'ordre dans lequel les différents éléments ont été agrégés :
   ```
   HClust.1$merge
   ```
4. La commande suivante indique à quel niveau ceux-ci ont été agrégés (hauteur des paliers) :
   ```
   HClust.1$height
   ```
5. À partir du menu *Classification*, ajouter les classes au jeu de données en choisissant trois classes, puis résumer la classification hiérarchique.
6. Comparer la partition en trois classes précédente avec celle que l'on obtiendrait à partir des choix suivants : distance = euclidienne, agrégation = liens simples.

(B.) CAH sur coordonnées factorielles

1. Sous FactoMiner, dans le menu ACP, sélectionner les 12 mois comme variables actives, les autres variables comme illustratives et cocher `réduire les variables`. Par défaut, la CAH est réalisée sur 5 coordonnées factorielles.

   Pour conserver toutes les coordonnées factorielles, saisir `Inf` dans la case "Nombre de dimensions". Cliquer ensuite sur le bouton `Réaliser une classification après l'ACP`. Dans la boîte de dialogue : choisir le nombre de classes de façon interactive ; le nombre optimal de classes à choisir entre 2 et 10 ; cocher les sorties graphiques et l'écriture des résultats pour les classes. Appliquer l'ACP.
2. Sur la fenêtre affichant le dendrogramme, choisir une partition en 3 classes
3. Visualiser les deux autres sorties graphiques proposées par la fonction HCPC :

   – le plan des individus (axes 1 et 2) de l'ACP où chaque objet est identifié par une couleur différente selon sa classe d'appartenance dans la partition ;

   – l'arbre hiérarchique en trois dimensions reconstitué sur le premier plan factoriel.
4. Produire le ratio entre l'inertie intra de la partition en (k) classes et celle en $(k-1)$ classes :

   ```
   res.hcpc$call$t$quot
   ```

   Interpréter les valeurs obtenues.
5. Soumettre à nouveau la procédure de construction d'un arbre hiérarchique et choisir une partition en 4 classes.
6. Produire les éléments les plus proches du centre de gravité de chacune des classes (les parangons) ainsi que les individus les plus spécifiques.

   Pour obtenir tous les éléments d'une classe, on peut par exemple demander à lister tous les parangons en ajoutant l'argument `nb.par = Inf` à la fonction `hcpc`.
7. Quelles sont les variables caractéristiques de chacune des classes ?

3. Évaluation sensorielle de 52 emmentals

Lors d'une évaluation sensorielle, 52 emmentals ont été dégustés par un panel d'experts et notés selon 17 caractéristiques sensorielles de goût, de texture et de parfum.

1. Réaliser l'ACP normée (variables actives = les 17 descripteurs sensoriels) puis la CAH à partir de toutes les coordonnées factorielles.
2. Quelle est la caractéristique sensorielle principale des emmentals de la classe 2 ? Comment définir brièvement cette classe ?
3. Quel est l'emmental représentant le mieux la classe 1 (dans le sens où il se rapproche le plus de l'emmental moyen de la classe) ? Idem pour la classe 4.
4. Quel est l'individu le plus spécifique de la classe 2 (dans le sens où il est le plus éloigné des centres des autres classes) ? Idem pour la classe 6.

4. Enquête sur les OGM

Réaliser une typologie des 135 personnes (adultes françaises) ayant répondu à l'enquête sur les organismes génétiquement modifiés (OGM).