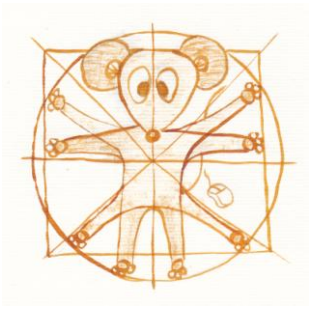

M1 EDSB
Data Mining

TD 4 : CAH et k-means

Robert Sabatier
Christelle Reynès

Mise en œuvre de classifications non supervisées sur le jeu de données « Villes »

On se propose d'explorer les données « Villes », déjà utilisées dans le TP2 sur l'ACP.

Calcul de la distance initiale et de l'arbre pour différentes CAH sur les observations.

1. Calculer la distance euclidienne entre les observations (attention à la transformation initiale des données)
2. Donner les différents arbres en utilisant les CAH avec les méthodes d'agrégations du saut minimum, maximum et Ward. Commenter les sorties et choisir le niveau de coupure.
3. En considérant les composantes choisies dans l'ACP centrée réduite du TD n°1, réaliser la CAH avec le critère de Ward et comparer l'arbre obtenu avec le même critère sur le jeu de données complet. En particulier, on comparera ces regroupements sur le premier plan de l'ACP.

Classification simultanée des observations et des variables

4. On se propose de réaliser sur le même tableau des données, une classification simultanée (mais indépendante) des variables et des observations. On choisira avec soin la mesure de dissimilarité entre les variables ainsi que l'algorithme de classification. La fonction *heatmap* (de R), va permettre de représenter simultanément les deux CAH.

Retour sur la classification des observations avec les k-means

5. Réaliser une classification non supervisée à l'aide des k-means, pour les données précédentes, en choisissant 4 classes. Recommencer l'opération plusieurs fois et comparer les classes obtenues.
6. Rechercher le nombre de classes raisonnables (en quel sens) et dessiner sur le plan 1-2 de l'ACP centrée et réduite les groupes obtenus. Commenter.
7. Que peut-on dire des groupes obtenus dans cette partie par rapport à ceux donnés par coupure de l'arbre de la CAH sur critère de Ward ?