CC2 - Clustering, hclust et kmeans

Etienne JEAN

L'étude consiste à effectuer des regroupements des poches du jeu de données qui présentent des caractéristiques semblables, afin de structurer les données. Cette approche non supervisée sera décomposées en deux parties : une étape de *classification hiérarchique ascendante,* et une étape de détermination des clusters par la *méthode des k-means.*

Classification hiérarchique ascendante

Le but de cette méthode est de visualiser, mettre en évidence des groupes dans le jeu de données. le terme *ascendante* fait référence au fait que l'on part des individus pour former des groupes de plus en plus larges, et non l'inverse. Les données sont d'abord normalisées afin de pouvoir calculer des distances entre les individus (distance euclidienne), puis on représente ces distances dans un dendrogramme (réalisé par la méthode de Ward comme critère d'aggrégation) (fig. 1).

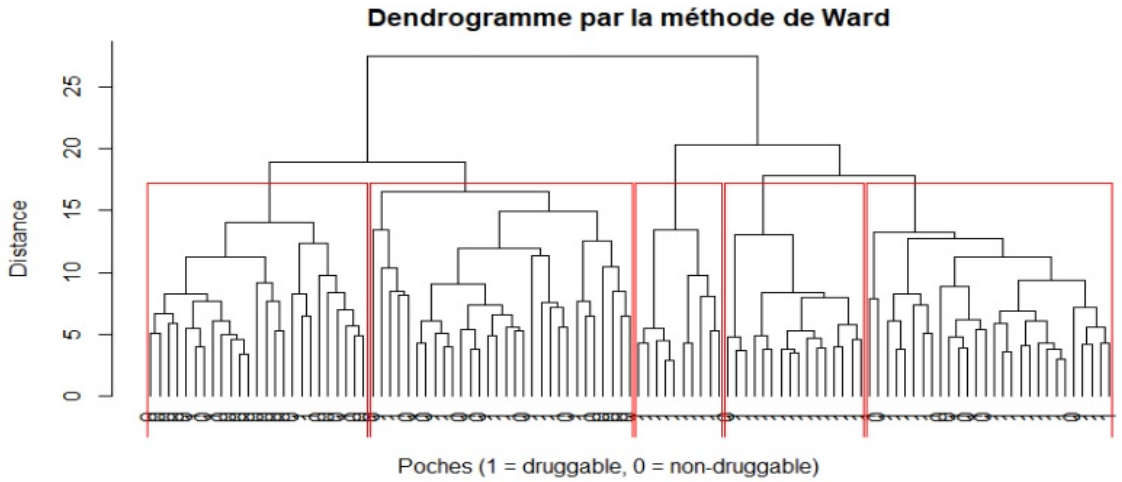

figure 1 : dendrogramme représentant les distances entre les poches.

Le problème est de savoir où effectuer la coupe pour définir les clusters. Le dendrogramme permet de visualiser s'il existe une coupe «naturelle», là où une branche serait plus longue que les autres, ce qui indiquerait une distance élevée entre les groupes. Nous avons ici effectué 5 clusters «à l'oeil», mais il n'apparait pas de coup évidente sur cet arbre.

Détermination du nombre de cluster idéal par la méthode des k-means

La méthode des k-means permet de partitionner le jeu de données en minimisant l'inertie intra-classe (distance entre les individus d'un même groupe). L'algorithme de partitionnement est simple et converge rapidement, mais il demande cependant de connaitre à l'avance le nombre de clusters. Il s'agit donc d'effectuer l'algorithme pour des nombres différents de clusters, et de regarder pour quel nombre de cluster on obtient une décroissance importante de l'inertie intra-classe (fig. 2).

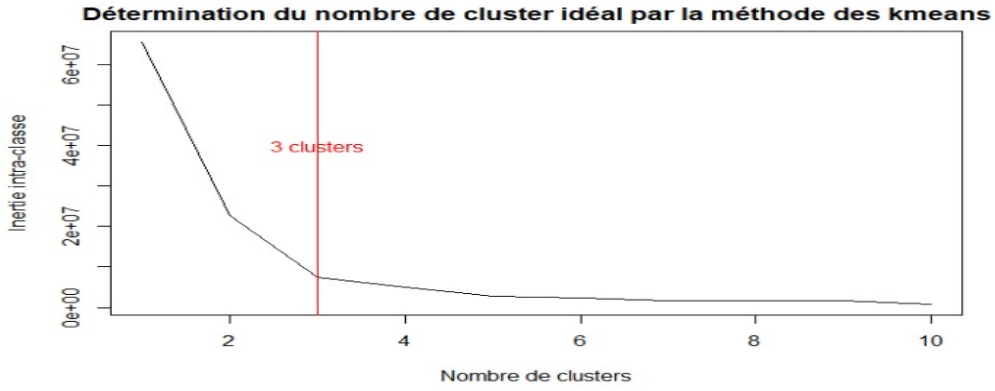

figure 2 : variation de l'inertie inter-classe en fonction du nombre de clusters.

On observe qu'il y a une décroissance assez significative de l'inertie intra-classe pour 3 clusters, et pas d'autre décroissance ne semble satifsfaisante au-delà de 3. Le meilleur partitionnement serait donc obtenu grâce à la méthode des k-means, pour 3 clusters.