SY09 - TP02
Classification automatique

Bertrand Bon - Antoine Hars

May 3, 2013

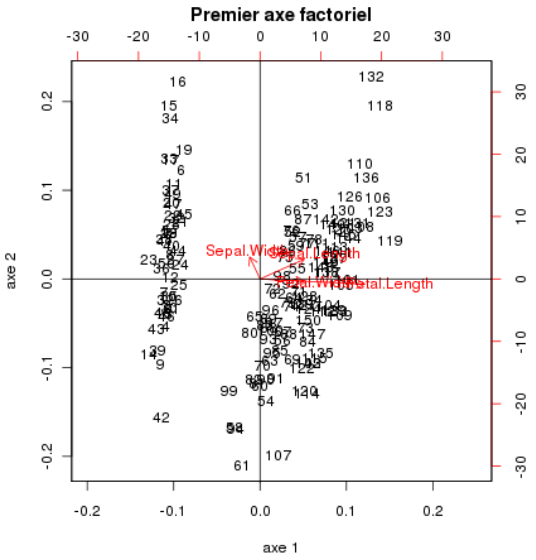
Introduction

Ce TP a pour but de nous permettre de nous familiariser avec différentes méthodes de classification automatique.
Nous avons pu revoir la méthode d'Analyse en Composantes Principales, et découvrir la méthode d'Analyse Factorielle d'un Tableau de Distances, la méthode des centres mobiles, ainsi que les classifications hiérarchiques ascendantes et descendante.

1 Visualisation des données

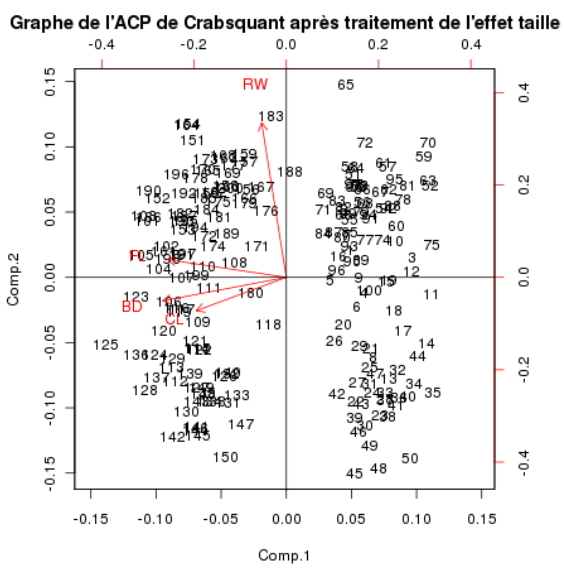
L'objectif de cet exercice est de visualiser les données qui seront étudiées dans la suite de ce TP.
Pour visualiser ces données, nous avons utilisé l'Analyse en Composantes Principales (ACP), ainsi que l'Analyse Factorielle d'un Tableau de Distances (AFTD) qui est équivalente à l'ACP lorsque les données disponibles se présentent sous la forme d'une matrice de dissimilarités.

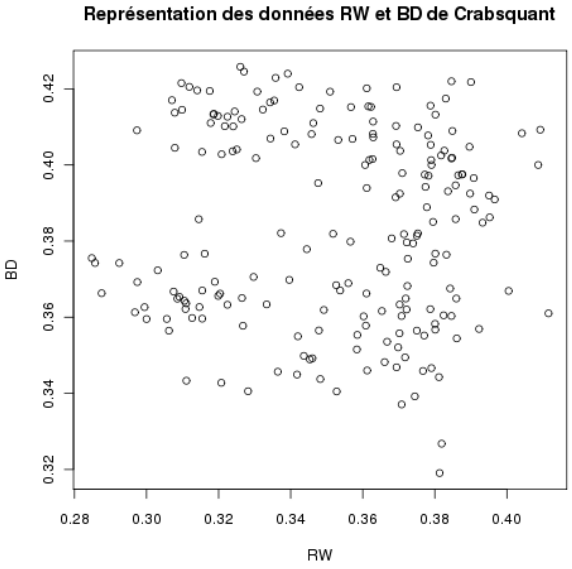
Charger les données *Iris* et sélectionner les variables quantitatives et afficher les données dans le premier plan factoriel.



On constate que les variables *petal.width* et *petal.length* sont fortement corrélées (ce qui semble logique car ces deux variables sont basées sur la taille des pétales).
En revanche, les variables *sepal.width* et *sepal.length* ne sont pas du tout corrélées car nous pouvons observer un angle approchant les 90° entre ces deux variables.

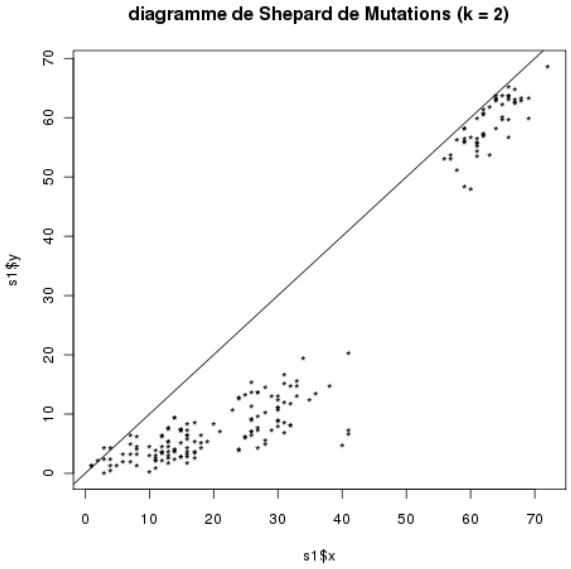
Effectuer l'ACP des données *Crabs*, préalablement traitées de manière à supprimer l'effet taille et comparer à la représentation des Crabes suivant les variables *RW* et *BD*.
Après traitement des données pour supprimer l'effet taille, nous obtenons la représentation suivantes des données :

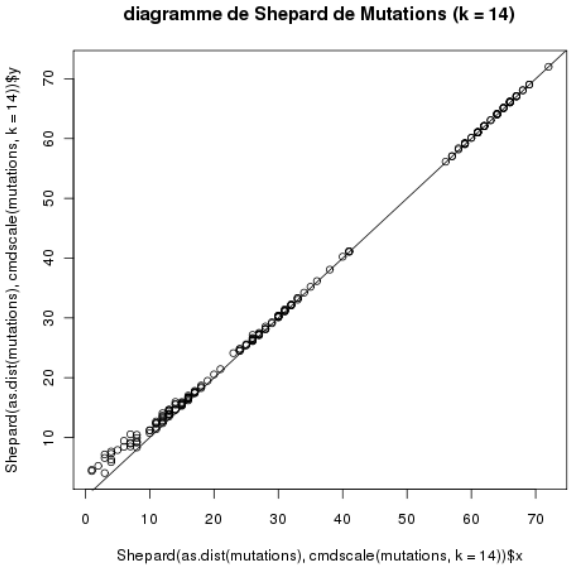




Grâce à ces deux graphiques, nous pouvons dire que les variables RW et BD ne sont pas du tout corrélées car elles forment un angle droit sur la représentation de l'ACP de Crabsquant après traitement de l'effet taille.
De plus, nous pouvons remarquer sur le deuxième graphique que les points sont très éparpillés et donc appuient cette non-corrélation entre les variables RW et BD.

Effectuer l'AFTD des données *Mutations*, puis afficher et analyser la représentation obtenue.
Le traitement de l'AFTD des données Mutations au moyen de diagrammes de Shepard nous donne les représentations suivantes :



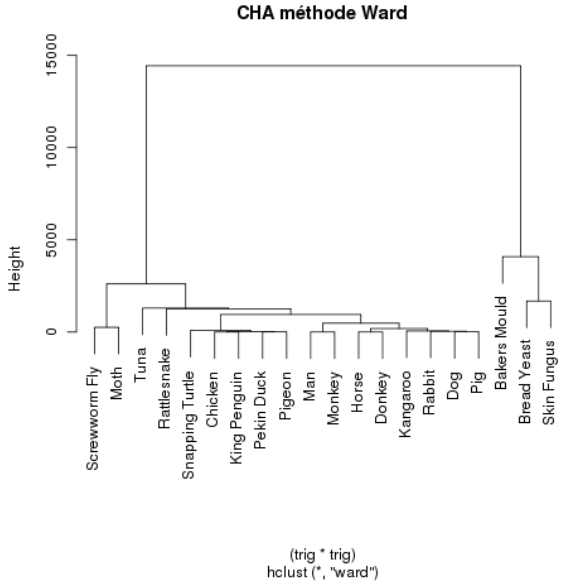


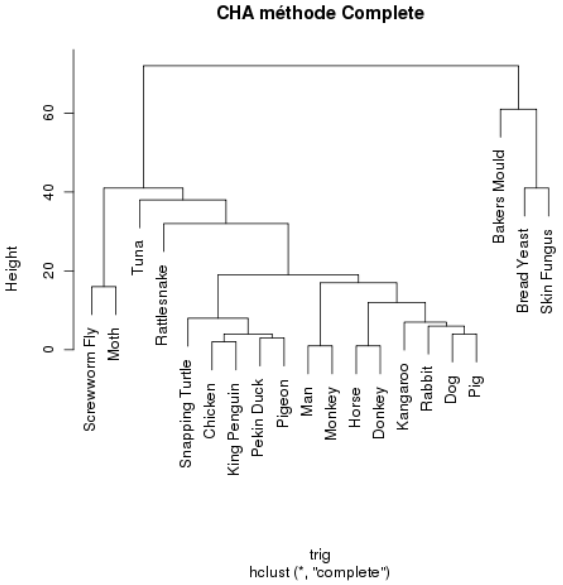
Dans le cas des données Mutations, nous utilisons plutôt une AFTD au lieu d'une ACP puisque les données sont données sour la forme d'une matrice de distances entre espèces.
Le diagramme de Shepard sert principalement à mesurer l'exactitude entre les distances originales et les distances retrouvées par l'AFTD.
Ce diagramme nous permet d'apprécier simplement la qualité des représentations. En effet, plus le nuage de points du diagramme est proche de la droite d'équation y = x, plus la représentation est fidèle.

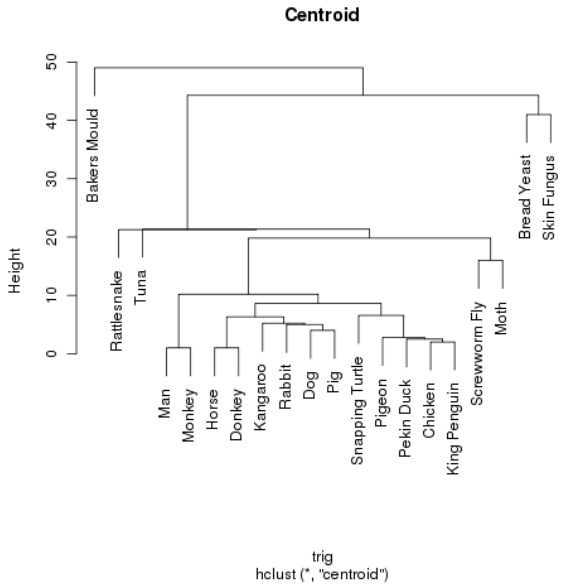
Nous pouvons donc observer que pour la dimension 2, la représentation est moins fidèle que dans le cas de la dimension 14.

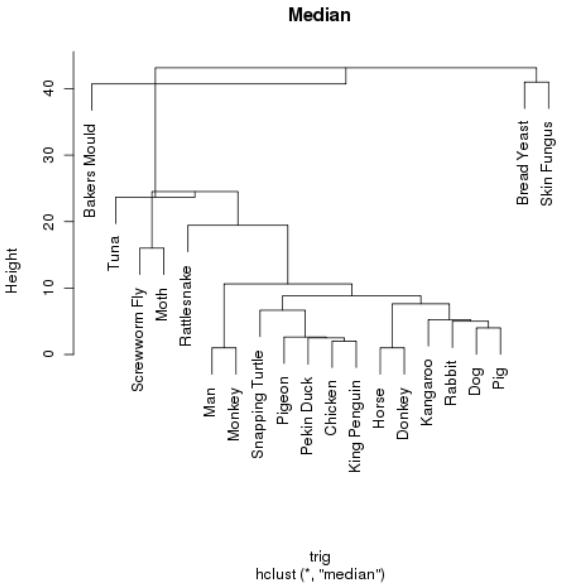
2 Classification hiérarchique

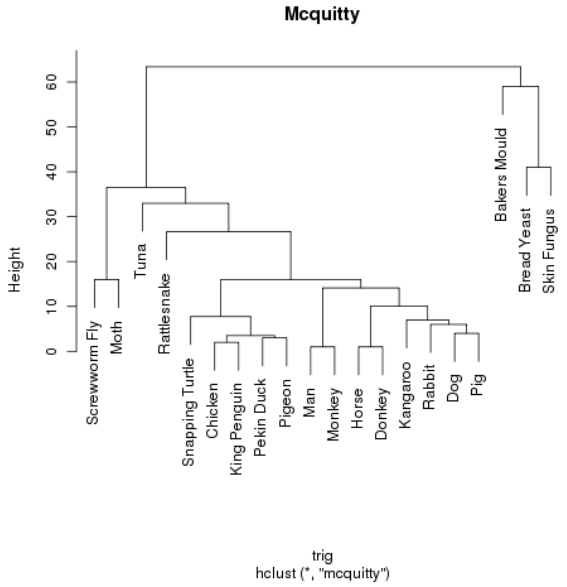
En utilisant la fonction *hclust*, effectuer la classification hiérarchique ascendante (avec les différents critères d'agrégation disponibles) des données de mutations. Commenter et comparer les résultats obtenus.

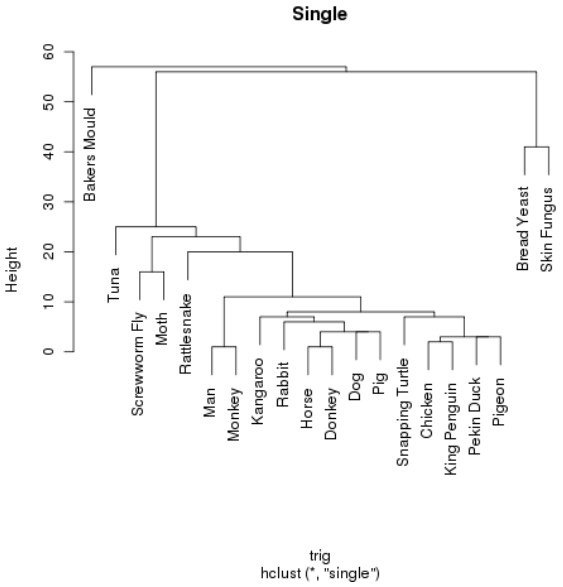






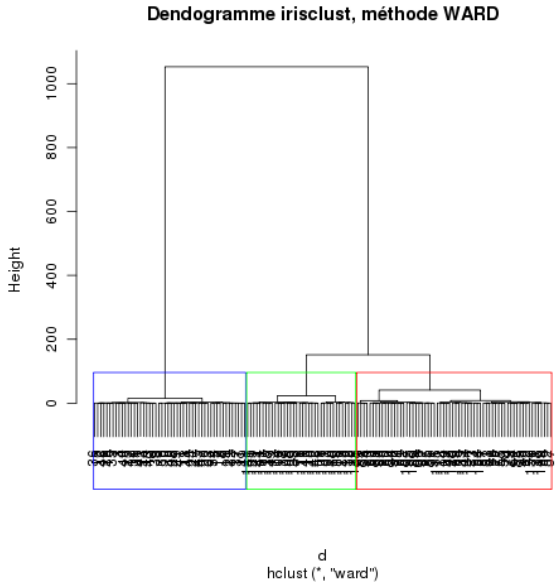






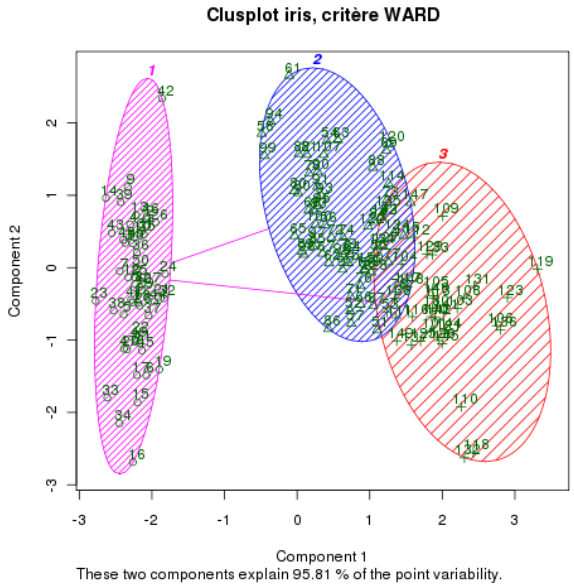
Nous pouvons remarquer que les CAH avec les critères d'aggrégation WARD, Complete, Mcquitty sont similaires.
Les CAH avec les critères d'aggrégation Single, Median diffèrent un peu des 3 premières.
Quant à la CAH avec le critère d'aggrégation Centroid, on peut observer une plus nette différence comparé aux 5 autres graphiques. S'il y en a une à utiliser, il s'agit de la méthode WARD car pour les variables quantitatives, le critère de WARD minimise l'inertie intra-classe, ce qui en fait le plus fiable pour notre cas de figure.

Effectuer la classification hiérarchique ascendante des données *Iris*. Commenter les résultats obtenus.
Pour effectuer cette CAH, nous utilisons le critère d'aggrégation WARD, ce qui nous donne le Dendogramme suivant :



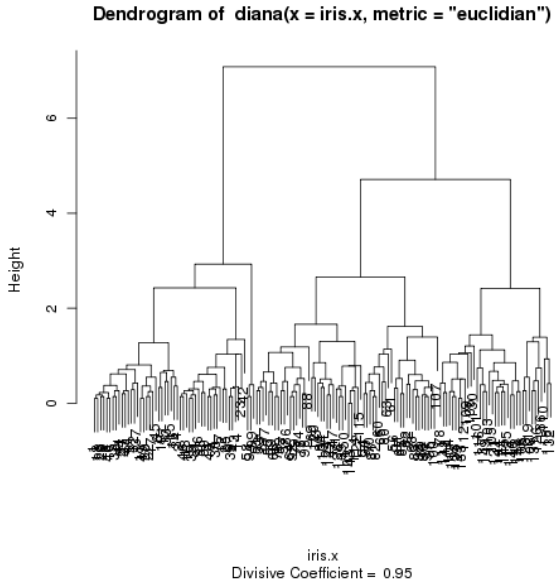
Nous avons mis en évidence les 3 principaux groupes qui ressortent de ce graphique, qui sont les espèces Setosa, Versicolor et Virginica.

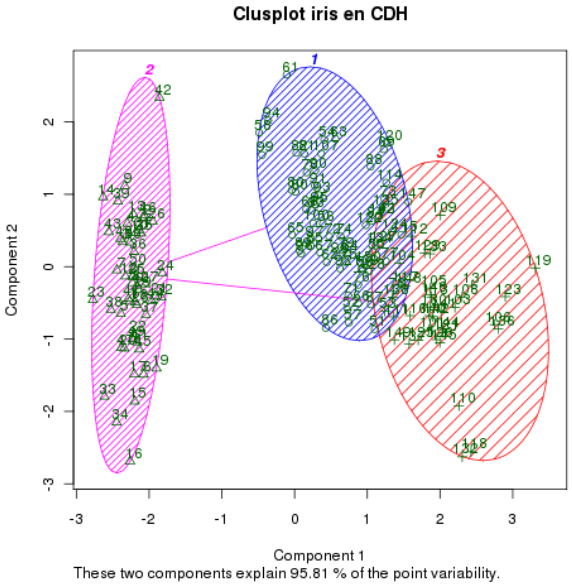
Au moyen d'un Clusplot avec le critère de WARD, nous obtenons la représentation suivante :



Cette représentation nous donne une meilleure vision de la répartition des espèces par rapport au Dendogramme. Nous pouvons observer sur le dernier graphique que les espèces Versicolor et Virginica se chevauchent et donc peuvent être difficiles à différencier alors que pour reconnaître l'espèce Setosa, ce souci ne se pose pas.

Effectuer la classification hiérarchique descendante des données *Iris*, au moyen de la de la fonction *diana*. Comparer aux résultats obtenus au moyen de la CAH.
La représentation de la classification hiérarchique descendante des données iris nous donne les graphiques suivants :





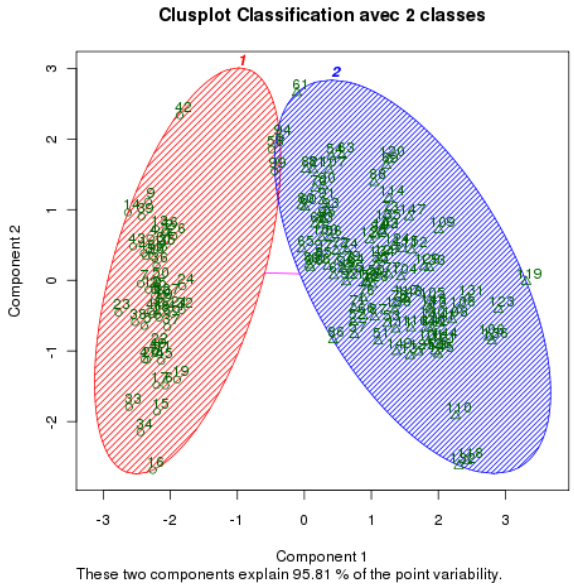
Dans notre cas, nous n'avons pas observé de différences flagrantes entre les classifications hiérarchiques ascendante et descendante.

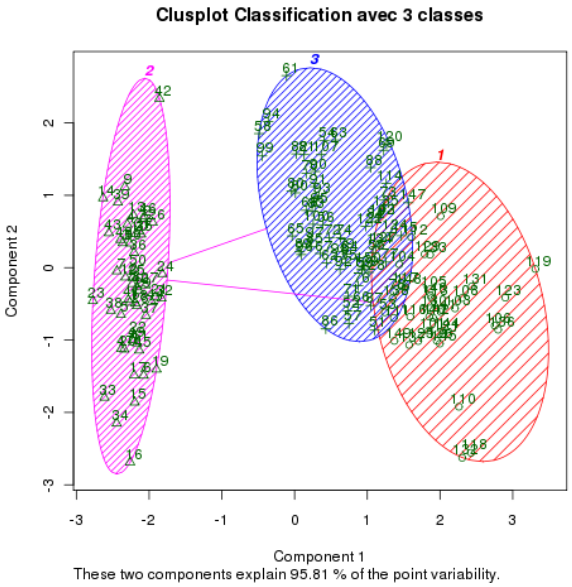
3 Centres mobiles

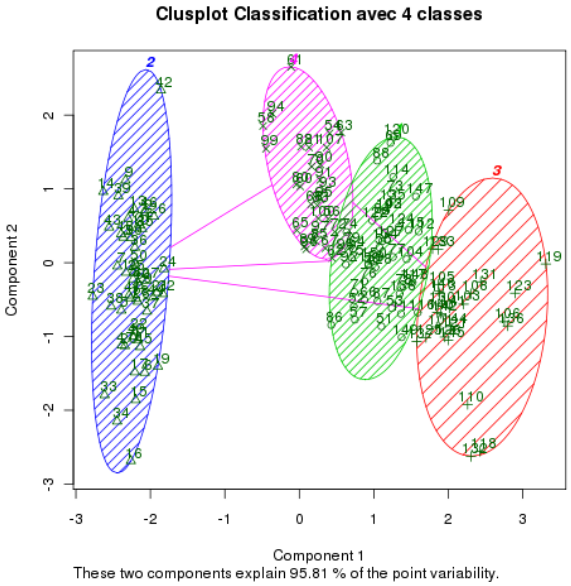
Le but de cet exercice est de tester les performances de l'algorithme des centres mobiles sur deux jeux de données réelles : **Iris** et **Crabs**.

3.1 Données Iris

Tenter une partition en K in {2,3,4} classes avec la fonction *kmeans*. Visualiser et commenter.
En application la fonction *kmeans* sur le jeu de données **iris**, nous obtenons les représentations suivantes des partitions demandées :

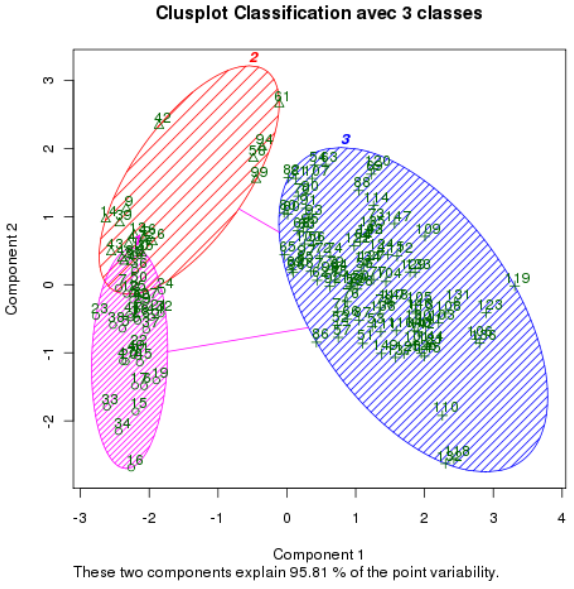


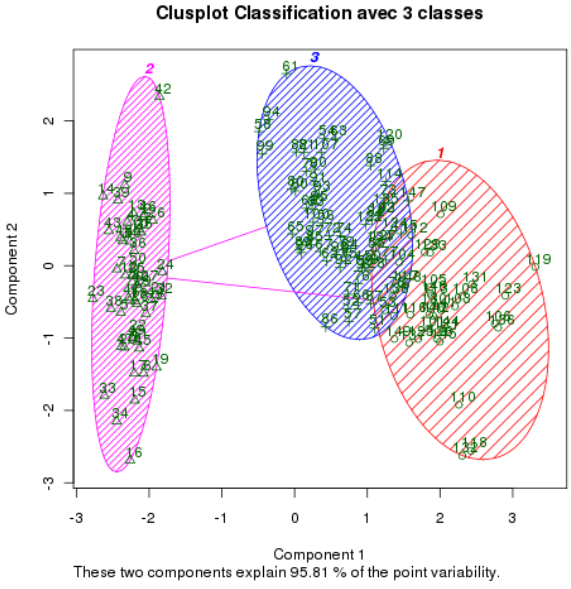




Du fait que le jeu de données *iris* regroupe 3 espèces, nous remarquons que la fonction *kmeans* regroupe les espèces Versicolor et Virginica pour un nombre de classes égal à 2.
Pour un nombre de classes égal à 4, la fonction *kmeans* se voit contrainte d'inventer un nouveau groupe ne correspondant à aucune espèce.
Et dans le cas où nous avons 3 classes, nous retrouvons la même forme de représentation que celle étudiée précédemment dans le tp.

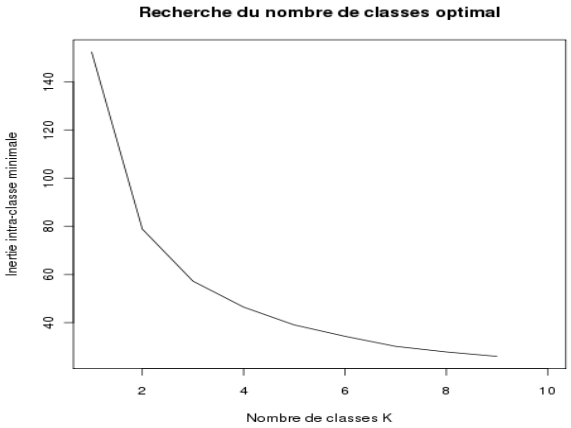
Étude de la stabilité du résultat de la partition. Effectuer plusieurs classifications en $K = 3$ classes du jeu de données. Observer ces résultats, en termes de classification et d'inertie intra-classes. Les résultats ont-ils changé ?
Nous avons pu remarquer que les résultats différaient parfois en terme de classification et d'inertie intra-classes comme nous pouvons le voir sur les représentations suivantes :





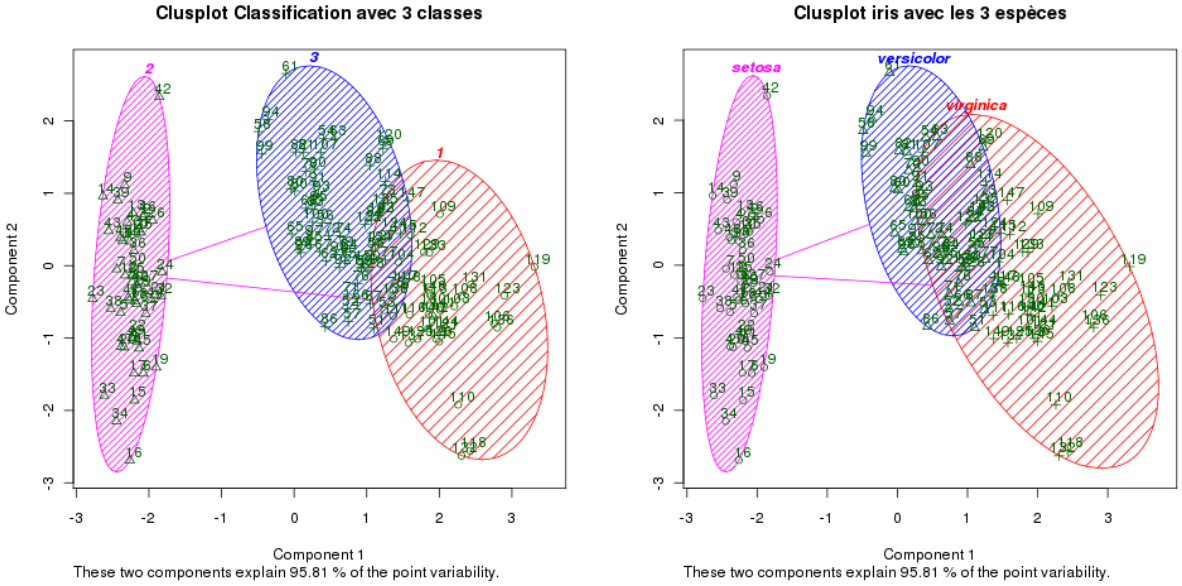
Cette différence de résultats pour l'exécution d'une même instruction provient de l'algorithme de *kmeans* en lui-même.
Cette variation dans les résultats est dûe à la sélection aléatoire des centres au début de l'exécution du kmeans. Pour palier à ce souci aléatoire, nous pouvons indiquer la valeur de départ à la fonction *kmeans* avec l'option *nstart = 20* qui indique à la fonction le nombre de points à choisir pour avoir le premier centre en début d'exécution de la fonction.

On cherche à déterminer le nombre de classes optimal.
Afin de rechercher le nombre de classes optimal pour le jeu de données iris, nous effectuons 100 classifications avec un nombre de classes compris entre 2 et 10, ce qui nous donne 9 échantillons contenant chacun 100 valeurs d'inertie intra-classe.
Ensuite pour chaque valeur de classe, nous calculons l'inertie intra-classe minimale et nous représentons sa variation en fonction du nombre de classes :



Pour déterminer le nombre de classe optimal à partir de cet représentation des variations de l'inertie intra-classe minimale, nous utilisons la méthode du coude qui nous indique que le nombre de classes optimal est égal à soit 2, soit 3, mais nous avons préférez prendre la valeur 3 qui est plus cohérente vis-à-vis du nombre d'espèces disponibles dans le jeu de données *iris*.

Comparer les résultats de la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes.
Nous avons comparé les partitions obtenues par les centres mobiles et réelles des iris :



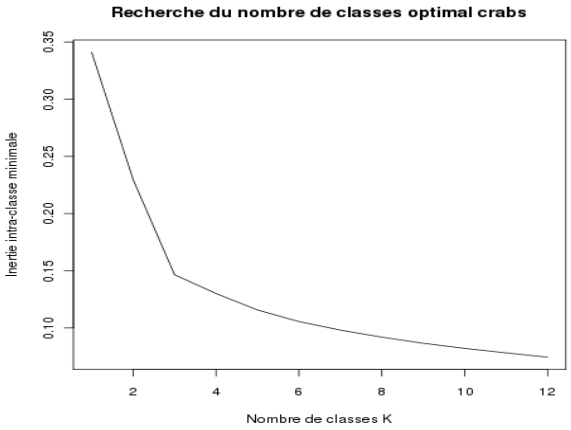
Nous avons pu observer que les 2 représentations sont similaires, l'espèce Setosa est bien identifiée dans les 2 cas tandis que des valeurs des espèces Versicolor et Virginica se recoupent.

3.2 Données Crabs

Effectuer la classification des données *Crabs* au moyen de l'algorithme des centres mobiles. Comparer à la partition réelle des crabes suivant l'espèce et le sexe.

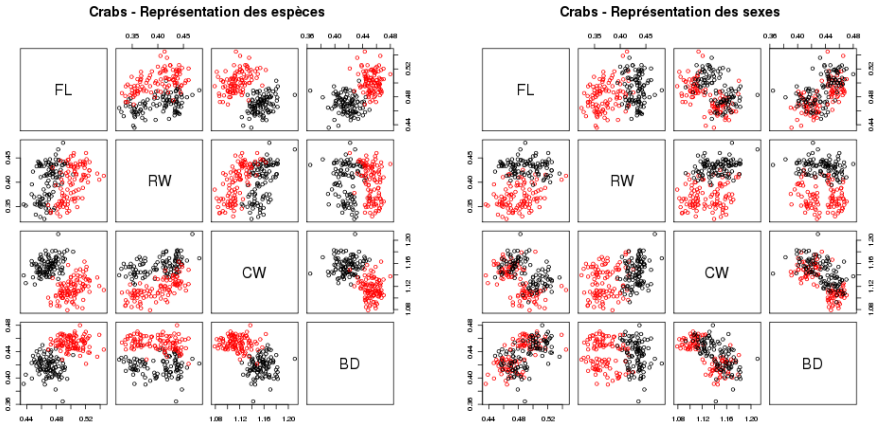
Dans un premier temps, nous récupérons puis traitons les données pour supprimer la forte corrélation présente dans le jeu de données crabs entre les variables.

Ensuite nous nous resservons du travail effectué en amont pour déterminer le nombre de classes optimal pour ce jeu de données :



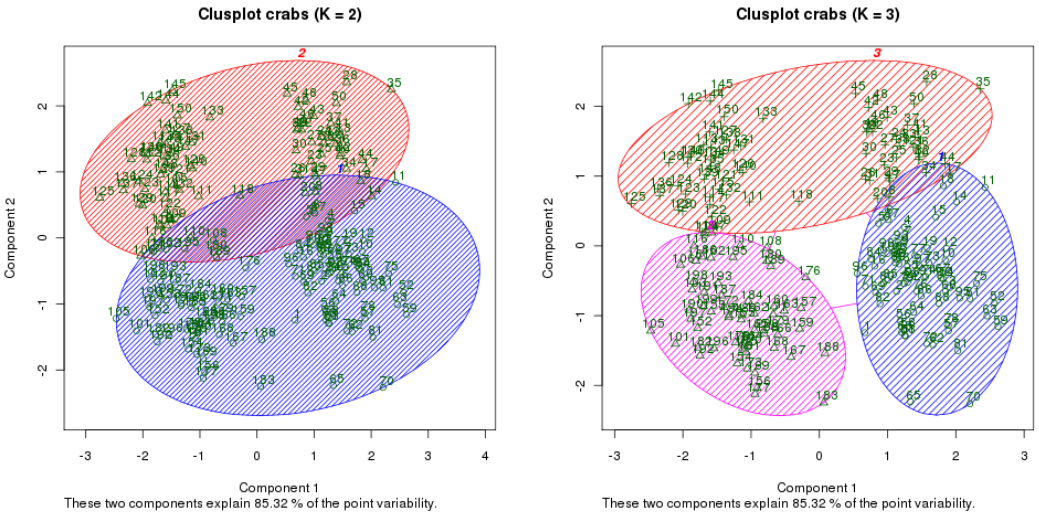
Nous remarquons sur ce graphique que la méthode du coude relève que le nombre de classes optimal est égal à 3 et la présence de coudes pour les valeurs 2 et 4 n'est pas assez prononcée.

Cela signifie peut être qu'une erreur de manipulation sur les données a été effectuée mais nous avons remarqué que sur les représentations des sexes et des espèces, les valeurs, pour les espèces, étaient bien séparées mais que dans certains cas pour les sexes, les valeurs étaient mélangées :

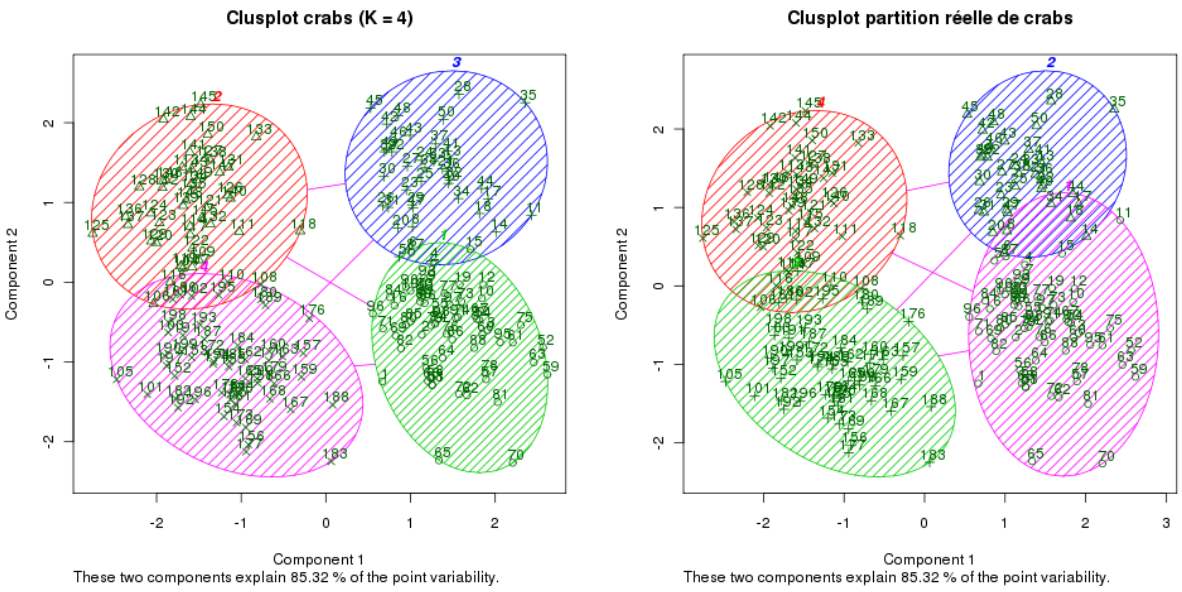


Cela pourrait expliquer le fait que le nombre de classes optimal soit égal à 3 au lieu de 4.

Les représentations de la classification du données de jeu Crabs pour K = 2 et K = 3 sont les suivantes :



Mais pour comparer la classification du données de jeu Crabs à la partition réelle, nous avons tout de même préféré nous baser sur un nombre de classes optimal égal à 4 :



L'étude de ces 2 graphiques nous montre que la classification du jeu de données crabs semble juste même si nous pouvons remarquer que les cercles ont tendance à se chevaucher 2 à 2 dans les deux partitions.

Conclusion

Les exercices de ce TP nous ont donc permis de découvrir plusieurs techniques de classification automatique telles que les classifications hiérarchiques ascendante et descendante et ainsi que la méthode des centres mobiles.
Ce TP nous a surtout permis de connaître les éléments importants pour l'utilisation de ces outils de classification automatique.