Compte-rendu du TP n°2: Classification automatique

Sarah Richard et Solène Weill

29 avril 2016

1 Introduction

L'objectif de ce second TP est d'utiliser les différentes méthodes de classification vues en cours : la classification ascendante hiérarchique, la classification descendante hiérarchique et la méthode des centres mobiles. Trois jeux de données sont proposés à l'étude : Mutations, Iris et Crabs. Nous commencerons par un exercice de visualisation des données afin de pouvoir appréhender les résultats à analyser.

2 Visualisation des données grâce à l'ACP et l'AFTD

2.1 Question 1

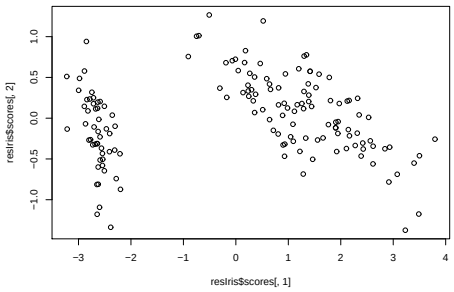

FIGURE 1 – Représentation des données iris dans le premier plan factoriel.

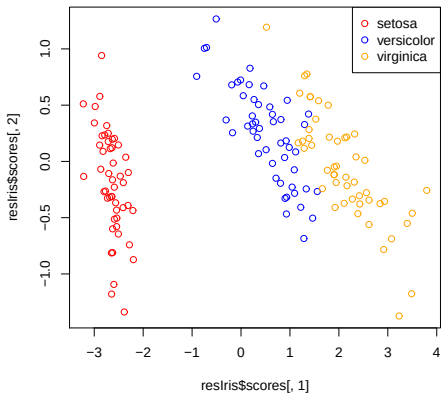

FIGURE 2 – Représentation des données iris par espèce dans le premier plan factoriel.

Afin de visualiser les données iris, nous réalisons tout d'abord une ACP. Nous affichons ensuite les données dans le premier plan factoriel et nous distinguons deux groupes de points.
Nous recommençons cette fois-ci en tenant compte de l'espèce. Pour cela nous attribuons une couleur à chaque espèces. On peut ainsi observer que le second groupe de points à droite est en fait constitué de deux espèces différentes. Les caractères des espèces sont proches, il est donc impossible de les différencier sans les identifier clairement sur le graphe. Lorsque l'on fera une partition des

données en deux clusters, on pourra s'attendre à ne trouver que deux sous-ensembles dans la partition. Un des deux sous ensemble sera composé des espèces Versicolor et Virginica. Lorsque l'on augmentera le nombre de clusters pour les partitions, on peut s'attendre à avoir majoritairement des erreurs de classification sur les deux dernières espèces citées.

2.2 Question 2

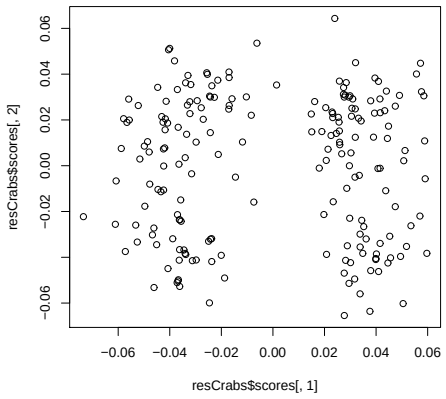

FIGURE 3 – Représentation des données crabs dans le premier plan factoriel.

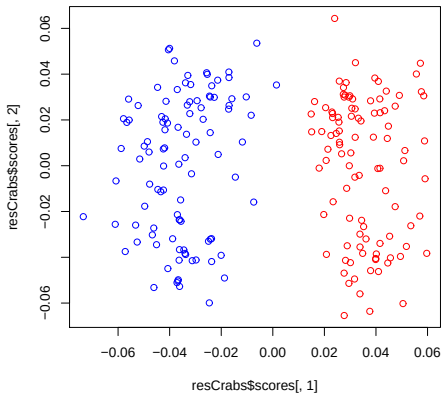

FIGURE 4 – Représentation des données crabs par espèce dans le premier plan factoriel.

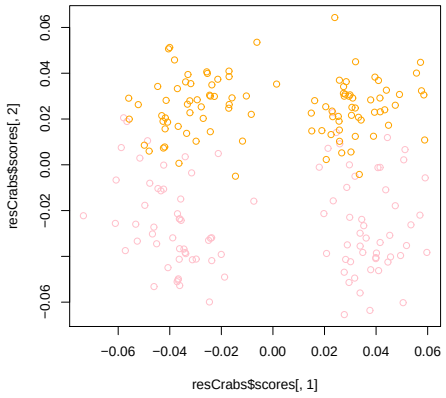

FIGURE 5 – Représentation des données crabs par sexe dans le premier plan factoriel.

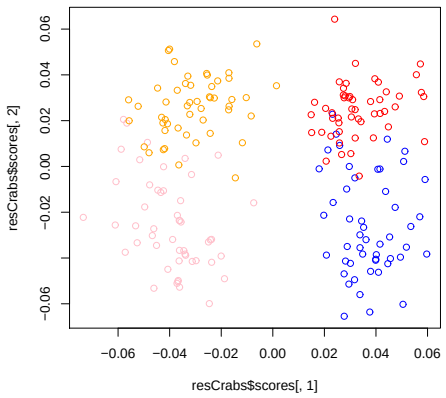

FIGURE 6 – Représentation des données iris par espèce et sexe dans le premier plan factoriel.

Pour étudier les données crabs, on réalise une ACP. Nous ne tenons, tout d'abord, ni compte du sexe ni de l'espèce. On observe deux groupes de points distincts. Nous recommençons ensuite

l'ACP en prenant en compte le sexe puis l'espèce séparément. Lorsque nous tenons compte d'une seule des deux variables qualitatives, nous observons également 2 groupes distincts. Cependant, ces deux groupes de points sont différents en fonction de la représentation par sexe ou par espèce. En réalisant une ACP qui tient à la fois compte du sexe et de l'espèce des crabes, nous observons 4 groupes de points. Un groupe représentant une espèce et un sexe précis.

2.3 Question 3

Le jeu de données ici à l'étude est un fichier CSV. Nous le chargeons en tant que tableau de dissimilarité. C'est pourquoi nous utiliserons l'AFTD pour visualiser les données Mutations. L'AFTD est l'équivalent d'une ACP avec pour données de départ un tableau de dissimilarités.

Lorsque l'on calcule une représentation euclidienne des données avec l'AFTD, on se rend compte que l'on a des valeurs propres négatives. Cela est dû au fait que la matrice de départ n'est pas une matrice de distance euclidienne. Pour parer à ce problème dans le calcul des pourcentages d'inertie, plusieurs solutions s'offrent à nous. On peut soit décider de ne pas prendre en compte les valeurs propres négatives soit prendre leurs valeurs absolues.

Une fois l'AFTD effectuée, nous affichons les résultats dans le premier plan factoriel de l'AFTD. Afin de mieux visualiser les groupes, nous avons affecté à chaque point le nom de l'espèce qu'il représente.

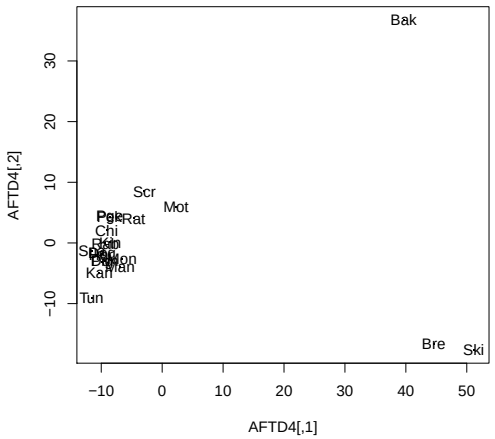

FIGURE 7 – Représentation de la distance des mutations dans le premier plan factoriel de l'AFTD.

La distance entre les différents points représente la dissimilarité de la protéine Cytochrome c entre chaque espèce. Ainsi, nous pouvons observer 3 groupes différents. Le premier à gauche composé des animaux et insectes et le deuxième et troixième regroupant les bactéries.
Afin de s'assurer de la qualité de la représentation, nous utilisons le diagramme de Shepard. Ainsi, plus la représentation est fidèle, plus les points sont alignés sur la bissectrice. On observe que pour K=2 la représentation ne semble pas très fidèle. Cependant, il est possible d'apercevoir 2 classes différentes.

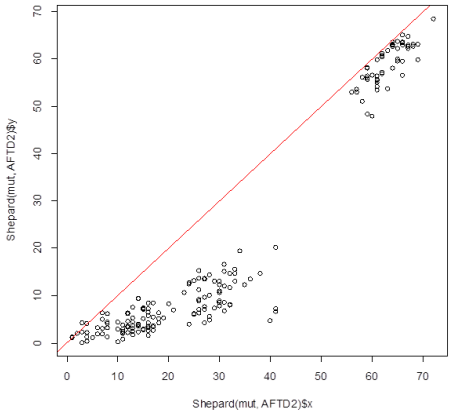

FIGURE 8 – Représentation de Shepard pour K=2.

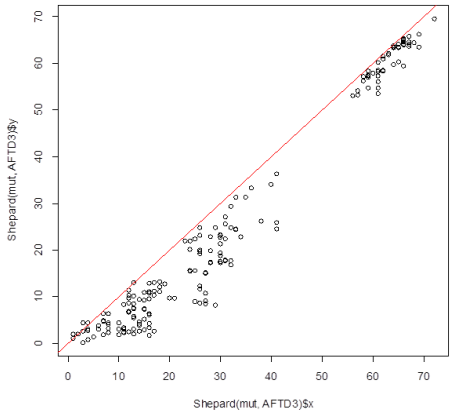

FIGURE 9 – Représentation de Shepard pour K=3.

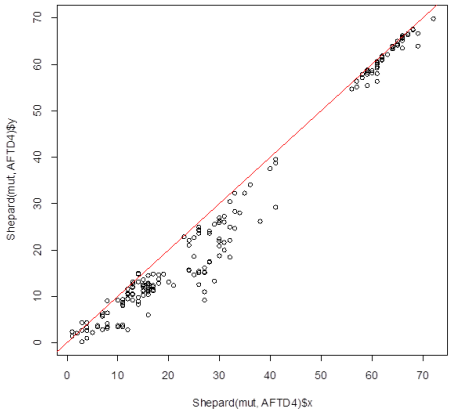

FIGURE 10 – Représentation de Shepard pour K=4.

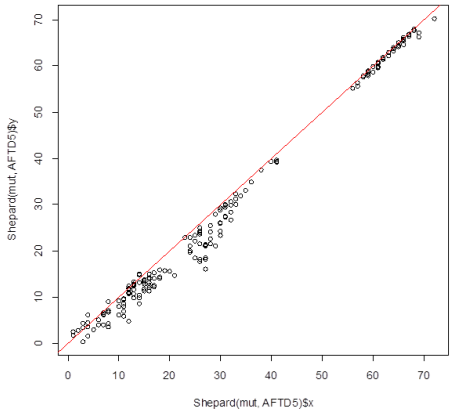

FIGURE 11 – Représentation de Shepard pour K=5.

Plus le nombre de composantes sélectionnées augmente, plus la représentation est fidèle. De plus, on observe maintenant 3 classes distinctes sur le diagramme de Shepard où K=5.

3 Classifications hiérarchiques

3.1 Question 1

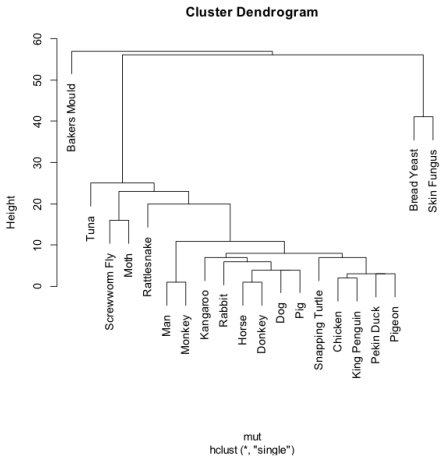

FIGURE 12 – Représentation du dendrogramme des données de mutation avec le seuil minimum comme critère d'agrégation.

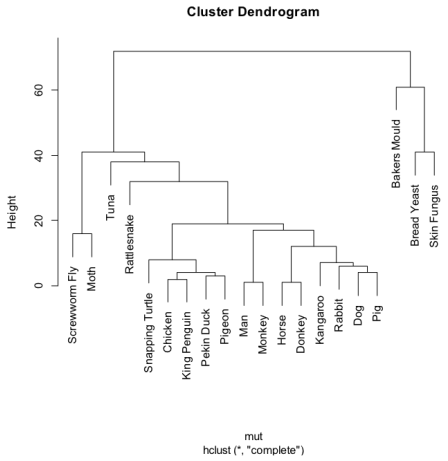

FIGURE 13 – Représentation du dendrogramme des données de mutation avec le seuil maximum comme critère d'agrégation.

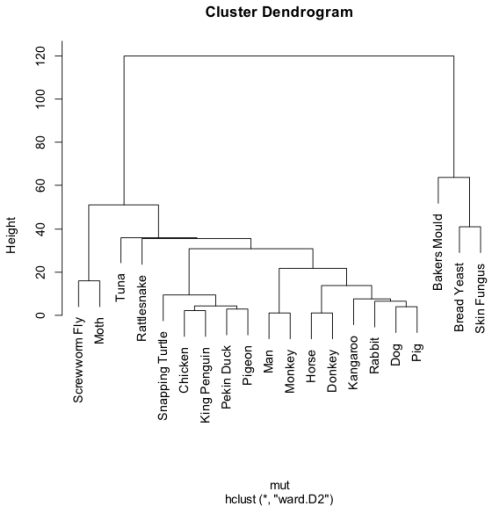

FIGURE 14 – Représentation du dendrogramme des données de mutation avec le critère de Ward comme critère d'agrégation.

Un dendrogramme permet la représentation des liaisons entre les différentes classes mais aussi de leur niveau de proximité grâce à la hauteur des branches. On observe sur nos trois dendrogrammes l'ensemble des espèces représentant les classes minimales. En fonction du critère d'agrégation utilisé, le niveau de proximité entre les classes change. Avec la méthode du seuil maximum et de Ward, on observe distinctement 2 classes. L'une composée de l'ensemble des animaux et insectes et l'autre

composée des bactéries. La classe composée des animaux et insectes peut ensuite être à nouveau subdivisée en plusieurs classes différentes. Comme par exemple, une classe comprenant les insectes et l'autre l'ensemble des animaux. Avec la méthode du seuil minimum, on observe 3 classes distinctes. Une comportant l'ensemble des animaux et insectes et les deux autres comportant les bactéries. Ce dernier dendrogramme peut être mis en relation avec la représentation du premier plan de l'AFTD où l'on retrouve les mêmes classes.

3.2 Question 2

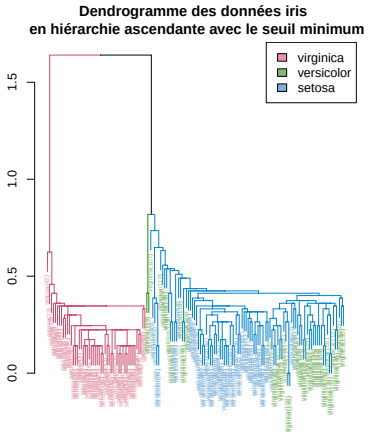

FIGURE 15 – Représentation du dendrogramme des données iris avec le seuil minimum comme critère d'agrégation.

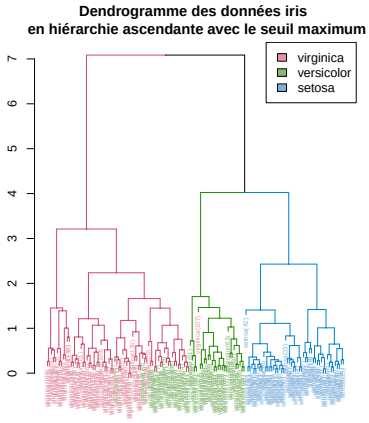

FIGURE 16 – Représentation du dendrogramme des données iris avec le seuil maximum comme critère d'agrégation.

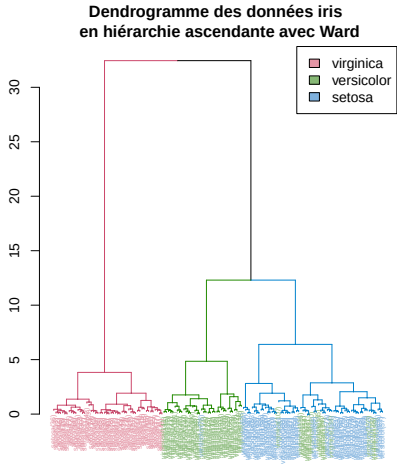

FIGURE 17 – Représentation du dendrogramme des données iris avec le critère de Ward comme critère d'agrégation.

On observe pour les 3 représentations deux groupes facilement distinguables. L'un comprenant l'espèce virginica et l'autre composé des espèces versicolor et setosa. On en déduit que ces deux

dernières ont beaucoup de caractéristiques communes. De plus, on peut remarquer qu'avec la méthode du seuil minimum, il est impossible de retrouver 3 espèces distinctes car l'iris versicolor et setosa sont trop entremêlés et similaires pour pouvoir les séparer. Avec les 2 autres méthodes, les 3 espèces sont identifiables. Cependant, en utilisant les couleurs pour identifier les espèces, on peut se rendre compte que certains spécimens ne se trouvent pas dans la bonne classe. Cela prouve ainsi que ces deux espèces ont beaucoup de caractéristiques communes.

3.3 Question 3

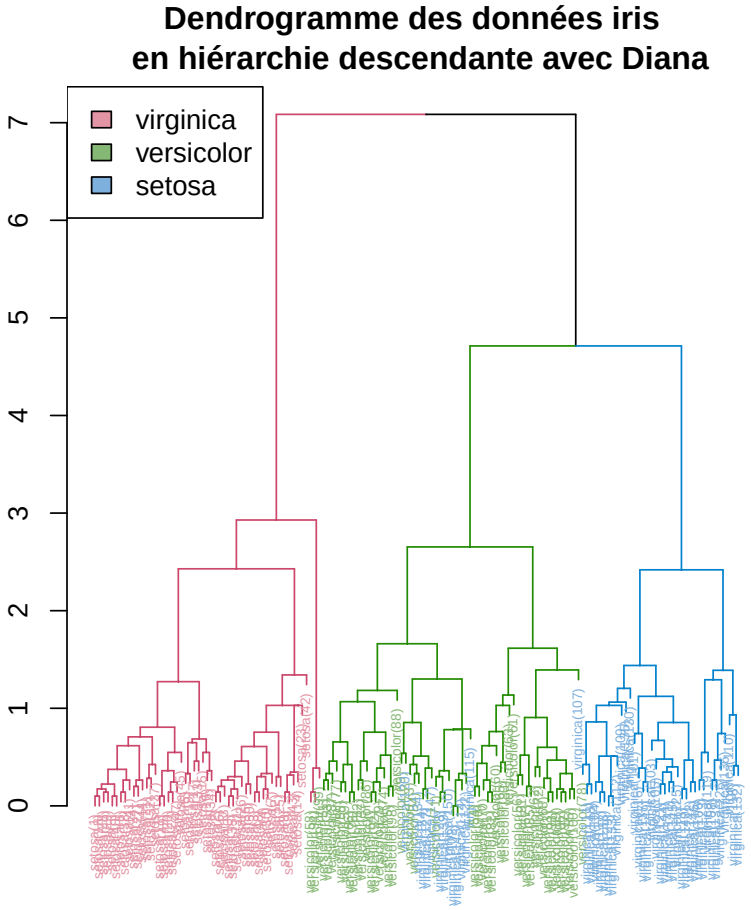

FIGURE 18 – Représentation du dendrogramme des données iris en hiérarchie descendante.

En hiérarchie descendante, on observe à nouveau 2 classes distinctes composée de l'iris virginica d'un côté et de l'iris setosa et versicolor de l'autre. Il est, de plus, possible de subdiviser cette deuxième classe en deux classes distinctes. Cependant, en identifiant les espèces par leur couleur, on remarque encore une fois que ces deux dernières espèces ne sont pas bien réparties en deux classes différentes mais sont intriquées les unes avec les autres. Les résultats de la CAH et de la CDH sont similaires, nos hypothèses semblent donc confirmées.

4 Méthode des centres mobiles

Dans cette partie, nous appliquerons la fonction kmeans, permettant de réaliser l'algorithme des centres mobiles, sur divers jeux de données.

4.1 Données Iris

4.1.1 Question 1

Nous réalisons premièrement des partitions en K= 2,3,4. Lorsque K est égal à 2, nous pouvons voir que deux des espèces d'iris ont été rassemblées en un seul cluster. Lorsque K=3, nous avons 3 clusters, représentant de façon assez fidèle la réalité de la répartition des individus dans le premier plan factoriel. Lorsque K=4, un nouveau cluster est ajouté, partitionant les individus réels d'une espèce en deux sous classes.

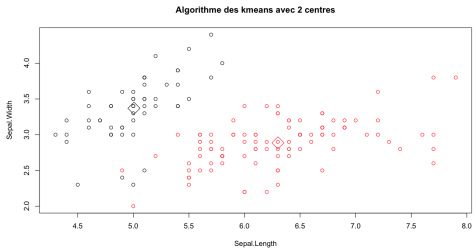

FIGURE 19 – Représentation d'un exemple de répartition des individus dans les clusters avec la fonction kmeans pour K=2 classes.

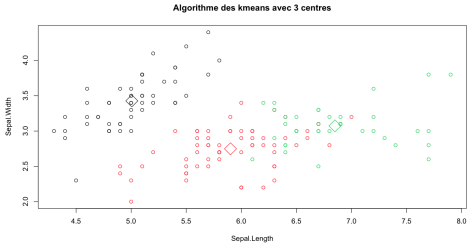

FIGURE 20 – Représentation d'un exemple de répartition des individus dans les clusters avec la fonction kmeans pour K=3 classes.

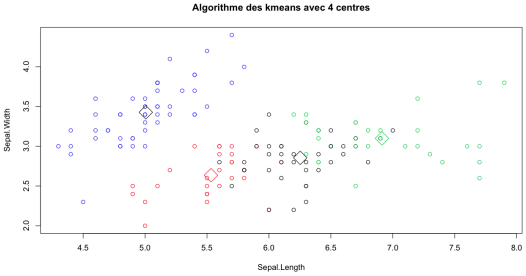

FIGURE 21 – Représentation d'un exemple de répartition des individus dans les clusters avec la fonction kmeans pour K=4 classes.

4.1.2 Question 2

Nous avons répété l'algorithme du kmeans avec K = 3 classes dix fois, ce qui nous permet de tester la stabilité de l'algorithme.
Nous observons 2 répartitions différentes. Avec la méthode des centres mobiles, K centres sont choisis aléatoirement au début de l'algorithme. En fonction des K premiers centres choisis, l'appartenance d'un individu à une classe peut varier. La seconde répartition est cependant majoritairement présente car c'est celle qui minimise la somme des inerties intra-classe.

$$\begin{pmatrix} & cluster1 & cluster2 & cluster3 \\ setosa & 17 & 33 & 0 \\ versicolor & 4 & 0 & 46 \\ virginica & 0 & 0 & 50 \end{pmatrix}$$

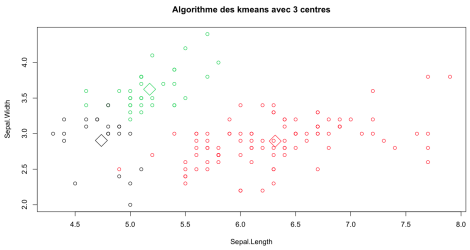

FIGURE 22 – Première répartition des individus dans les clusters observés .

$$\begin{pmatrix} & cluster1 & cluster2 & cluster3 \\ setosa & 0 & 50 & 0 \\ versicolor & 48 & 0 & 2 \\ virginica & 14 & 0 & 36 \end{pmatrix}$$

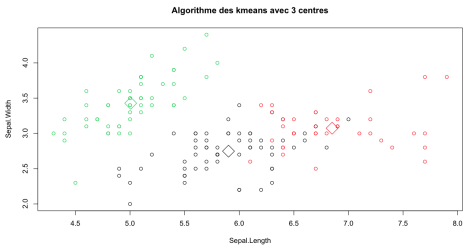

FIGURE 23 – Seconde répartition des individus dans les clusters observés .

4.1.3 Question 3

a Afin de calculer et stocker les valeurs d'inertie intra-classe pour les données iris, nous avons créé une fonction (cf annexes) qui permet de calculer la somme des inerties intra-classe 100 fois pour chaque K, nombre de classe. Le résultat est stocké dans une matrice.

b Une fois la matrice obtenue, nous avons sélectionné le minimum des inerties de chaque colonne grâce à la fonction apply(). Nous avons ainsi tracé le graphique suivant.

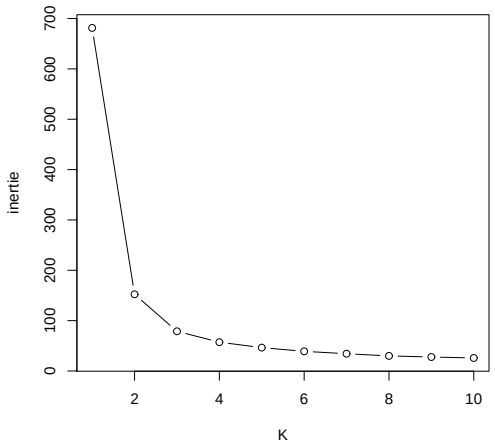

FIGURE 24 – Représentation de la variation de l'inertie minimale en fonction de K.

On constate que l'inertie diminue lorsque le nombre de classes augmente. Cela est normal car la distance entre les individus et le centre de classe diminue. L'inertie nulle est atteinte lorsque le nombre de classes est égal au nombre d'individus. Afin d'interpréter ce graphique, nous utilisons la méthode du coude. Celle-ci stipule que le nombre de classes optimal est obtenu lorsque l'inertie chute brusquement. Ce nombre correspond alors à la classe où la baisse abrupte des inerties se

termine. Entre l'inertie pour K=1 et K=2 on observe un décrochage, cependant l'inertie pour K=2 n'est pas proche de l'inertie pour K=1. Ainsi, l'inertie ne diminue pas encore doucement de façon continue. Nous choisissons donc K=3, valeur pour laquelle l'inertie a encore diminué tout en restant proche de l'inertie pour K=2.

4.1.4 Question 4

Le jeu de données iris présente 3 espèces chacune représentée par 50 individus. Lorsque l'on réalise la partition avec les kmeans, on se rend compte que les individus des différentes espèces peuvent être mélangés dans les clusters. On peut remarquer que, dans la partition qui revient le plus souvent avec la méthode des centres mobiles pour K=3, l'espèce setosa semble être distingable de façon explicite. En revanche, les deux autres espèces sont souvent entremêlées car leurs caractéristiques sont plus proches. Cependant, le taux d'erreur reste faible.

4.2 Données Crabs

4.2.1 Question 1

Nous réalisons 100 classifications en K=2 des données crabs. Pour visualiser la répartition des individus au sein des deux groupes, nous avons décidé de tenir compte du sexe et de l'espèce de chaque crabe. Nous affichons donc le nombre de crabes de l'espèce X et de sexe Y dans chaque sous-ensemble de la partition. Nous obtenons deux types de résultats : l'algorithme de classification rassemble les individus soit selon leur espèce (dans la majorité des cas), soit selon leur sexe. Nous obtenons ainsi 3 partitions différentes. On observe 5 centres mobiles différents mais 4 de ces centres mobiles forment seulement 2 partitions différentes. En effet, c'est juste le numéro de cluster affecté au centre mobile qui est modifié mais la partition reste la même.

Inertie du 1er centre	0.126809103539227	0.131969157861268	0.150537520424091	0.169341596514762	0.205527667560138
Répartition en fonction du 1er centre	40	48	5	1	6

Inertie du 2eme centre	0.126809103539227	0.131969157861268	0.150537520424091	0.188195803387884	0.205527667560138
Répartition en fonction du 2eme centre	48	40	6	1	5

4.2.2 Question 2

Pour cette question, nous avons procédé de façon similaire. Nous avons réalisé la classification 100 fois en ne regardant la répartition qu'en fonction de l'espèce, 100 fois en ne regardant la répartition qu'en fonction du sexe et 100 fois en fonction des deux. En fonction des espèces, il est facile de reconnaître les espèces qui sont bien séparées et les erreurs sont faibles. Dans l'exemple ci dessous seulement un crabe a été mis dans le mauvais cluster. On peut même supposer le sexe des individus de certains clusters. En fonction du sexe, les clusters comprennent des individus féminin et masculin mélangés en plus grands nombres. Les clusters regroupent ainsi plus fidèlement les espèces entre elles que les individus par sexe. En séparant par espèce et sexe, on observe que certains individus B et masculins sont proches des individus féminins ce qui induit des erreurs.

	1	2	3	4
B.F	0	50	0	0
O.F	49	1	0	0
B.M	0	13	37	0
O.M	8	0	0	42

Exemple d'une répartition dans les clusters des crabs en fonction de l'espèce et du sexe

Nous avons également cherché le nombre optimal de classes grâce à la méthode du coude comme précédemment. Le nombre optimal de classes est 4 comme on peut le voir sur le graphique ci-dessous. Cela confirme les résultat obtenus en regardant la répartition des individus dans les clusters.

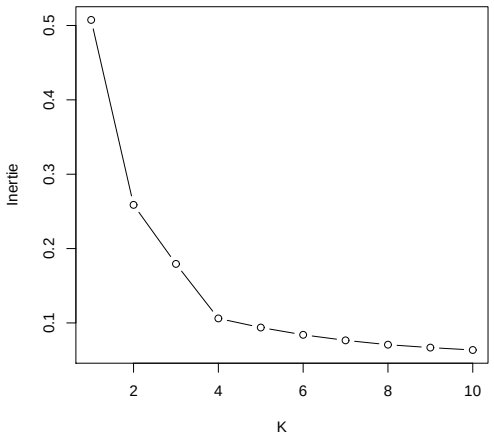

FIGURE 25 – Représentation de la variation de l'inertie minimale en fonction de K.

4.3 Données Mutation

4.3.1 Question 1

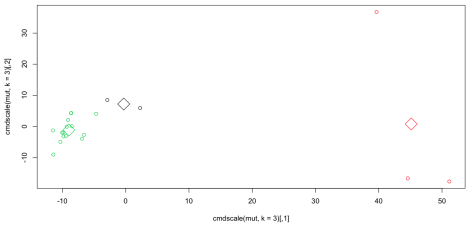

FIGURE 26 – Représentation de la répartition des individus dans le premier plan factoriel avec la méthode des centres mobiles. Inertie intra-classe totale moyenne.

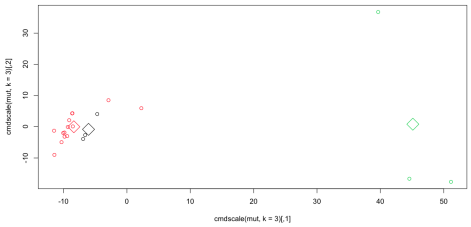

FIGURE 27 – Représentation de la répartition des individus dans le premier plan factoriel avec la méthode des centres mobiles. Inertie intra-classe totale maximum.

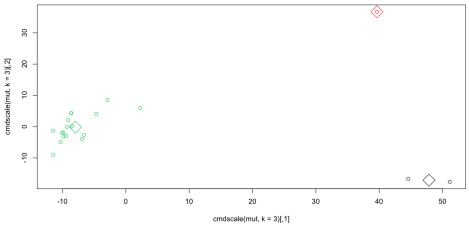

FIGURE 28 – Représentation de la répartition des individus dans le premier plan factoriel avec la méthode des centres mobiles. Inertie intra-classe totale minimum.

On obtient 3 groupes, dont 2 plutôt rapprochés lorsque l'inertie totale intra classe est moyenne et maximum. On peut supposer que cela correspond aux animaux et aux bactéries. Si l'on compare le graphique de la répartition minimisant l'inertie intra classe avec la représentation des données dans le premier plan factorielle après AFTD, on peut voir que les graphiques sont similaires. Les trois groupes correspondes donc : à gauche aux animaux et insectes, à droite à deux groupes bactéries.

4.3.2 Question 2

Nous effectuons un calcul des différentes inerties répété 100 fois pour K=3 classes.

Répartition des individus dans les clusters	Inerties	Fréquence
1,2,17	2046.391	58
3,3,14	3491.282	18
2,3,15	3025.057	24

Fréquence d'apparition des inerties et répartition correspondantes des individus dans les clusters

On peut ainsi remarquer que la partition n'est pas stable car l' inertie totale intra classe qui apparait le plus souvent n'apparait seulement que 58% du temps. Les deux autres inerties totales apparaissent 18% et 24% ce qui fait presque 1 fois sur 4 pour la répartition maximisant l'inertie intra classe et 1 fois sur 5 pour l'autre .

5 Conclusion

Ce TP nous a permis de mettre en pratique différentes méthodes de classification automatique. On a pu apprécier l'importance du choix du critère d'agrégation en en testant plusieurs et l'importance de choisir la bonne métrique. Il nous a aussi permis d'identifier le nombre de classe par différentes méthodes, même s'il n'est pas toujours évident de décider du nombre optimal de classes. Enfin, l'utilisation de différentes méthodes nous a permis de mettre leurs résultats en perspective et de les comparer.

6 Annexe

6.1 Fonction de calcul des inerties totales intra-classe pour K allant de 2 à 10. Les inerties sont calculées 100 fois pour chaque K.

```
classopt<- function(n, j) {
res<-matrix(0,100,9);
for(k in 2:10){
        for (i in 1:n) {
                class<-kmeans(j,k);
                Inertie<-class$tot.withinss
                res[i, k-1]<-Inertie;
        }
}
res
}
```