Classification automatique

YANQING ZENG GI05
MINH TRI LÊ GI02

27 avril 2017

Introduction

L'objectif de ce TP est d'appliquer des méthodes de visualisation et de classification automatique afin d'explorer des jeux de données selon les similarités ou dissimilarités des individus. Nous utiliserons l'ACP et l'AFTD (CMDS : Classical MultiDimentional Scaling) pour la visualisation ainsi que la classification hiérarchique et la méthode des centres mobiles (K-Means) pour la classification automatique.

1 Visualisation des données

Le but de cette partie est de visualiser les données dans un espace de dimension 2 avec l'ACP ou l'AFTD.

1.

Les données Iris sont des mesures de certaines espèces de fleurs (3 espèces).
On effectue donc l'ACP sur ce jeu de données, d'abord sans et avec coloration pour distinguer les espèces. (Figure [1]).

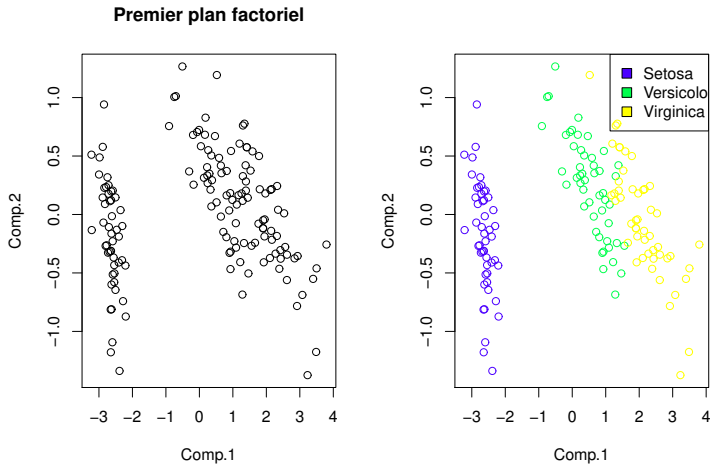

FIGURE 1 – Représentation des données *iris* par ACP

Après l'ACP, on peut facilement distinguer deux groupes sur la figure de gauche (1). Seulement, on se rend compte que la figure de droite (1) montre en fait trois groupes. Le groupe

de données de droite est en effet plus difficilement distinguable en deux groupes. Les espèces Versicolor et Virginica sont assez proches.

Il en résulte que si l'on recherche une partition de donnée et si le nombre K de partitions vaut 2, les Versicolor et Virginica seront dans le même groupe. Avec K=3, les partitions correspondront aux trois espèces.

2.

Comme vu au TP précédent, les données crabs correspondent à des mensurations de deux espèces de crabes, mâle et femelle.

De la même manière, on effectue l'ACP sans et avec distinction de l'espèce et du sexe, on obtient la figure suivante (2).

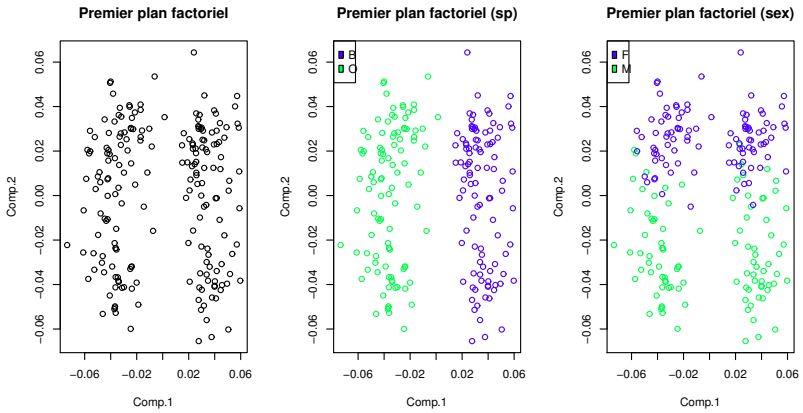

FIGURE 2 – Représentation des données *crabs* par ACP

On distingue 4 groupes sur la figure de gauche (2) qui correspondent aux 4 combinaisons d'espèces et de sexe.

La figure du milieu (2) pour l'espèce montre 2 groupes de données correspondant aux deux espèces de crabes. De même pour la figure de droite (2) pour le sexe, on remarque qu'il y a un léger recouvrement des données au milieu de cette figure.

Ainsi, on constate bien qu'il y a quatre groupes d'individus différents pour ce jeu de données avec l'ACP.

3.

Les données Mutations contiennent les mesures de dissimilarités d'une protéine entre différentes espèces.

On effectue l'AFTD sur avec cette mesure de dissimilarités, ici on choisit la représentation à 2 variables (figure 3).

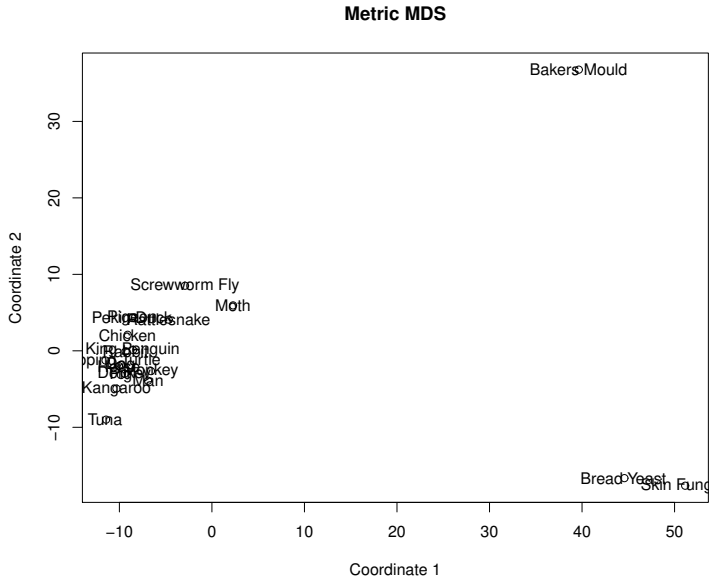

FIGURE 3 – Représentation des données à 2 variables par AFTD

L'AFTD (3) montre une grande concentration des points au milieu à droite et des singularités pour $\{BakersMould\}$ et $\{BreadYeast, SkinFungus\}$.

Pour vérifier la qualité de la représentation, on effectue un diagramme de Shepard qui représente la distance euclidienne en fonction de la dissimilarité. Si la représentation est bonne, les données sont alignées selon la droite $y = x$.

Sur la figure 4, on remarque que plus d augmente, plus les données sont alignées selon la bissectrice et donc la qualité de la représentation est meilleure.
Pour d=2, la qualité de la représentation est donc moins bonne que pour d=5 par exemple.

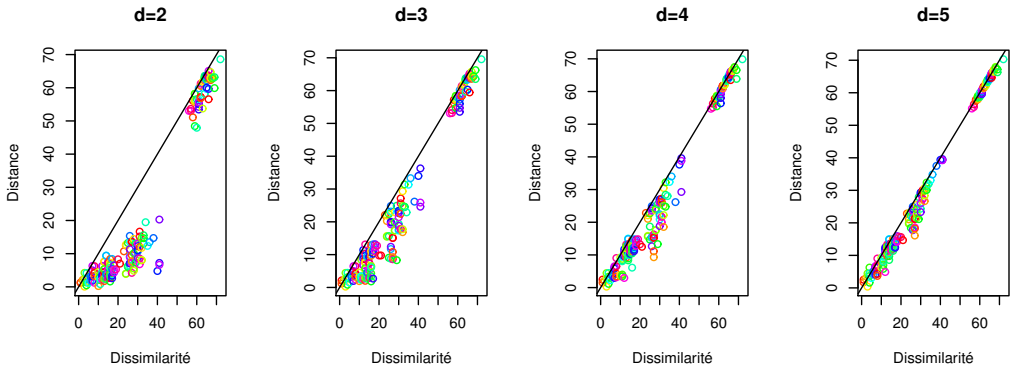

FIGURE 4 – Diagramme de Shepard à d variables

2 Classification hiérarchique

Dans cette partie, nous effectuerons des classifications hiérarchiques ascendantes avec la fonction *hclust* et descendantes avec *diana* sur les données *mutations* et *iris*.
La classification hiérarchique ne nécessite pas de choisir le nombre de classe à l'avance comparé à l'algorithme des K-Means que l'on étudiera après.

1.

Nous avons effectué des classifications hiérarchiques ascendantes avec quatre critères d'agrégations différents.
La classification hiérarchique ascendante (CAH) part des individus, donc du cas particulier, pour arriver au cas général. (5)

Les critères d'agrégations utilisés sont les suivants :

— *complete* : Critère du lien maximum : plus grande distance entre deux classes. $D(A,B) = max\{d(i,i'), i \in A \text{ et } i' \in B\}$

— *single* : Critère du lien minimum : plus petite distance entre deux classes. $D(A,B) = min\{d(i,i'), i \in A \text{ et } i' \in B\}$

— *ward D2* : Critère de l'inertie intra-classe minimisée au maximum. $D(A,B) = \frac{n_A n_B}{n_A + n_B} d^2(g_a, g_b)$ où g_i est le centre de gravité du sous-espace i.

— *centroid* : Utilise le carré des distances euclidiennes entre les différents centres.

Les classifications hiérarchiques obtenues sont très proches de la représentation de l'AFTD. En effet, les différents groupes d'individus de l'AFTD sont bien hiérarchisés distinctement de la même manière.

On remarque que les différents critères d'agrégations groupent l'espèce Bakers Mould avec ou sans Bread Yeast et Skin Fungus, sachant que ces trois espèces sont bien séparées. Les autres individus sont très proches dans la représentation de l'AFTD et sont donc proches dans la hiérarchie.
Les méthodes complete et ward D2 ont des résultats similaires et groupent Bakers Mould avec Bread Yeast et Skin Fugus.

Les méthodes single et centroid, séparent l'espèce Bakers Mould et sont donc plus proches de la représentation de l'AFTD.

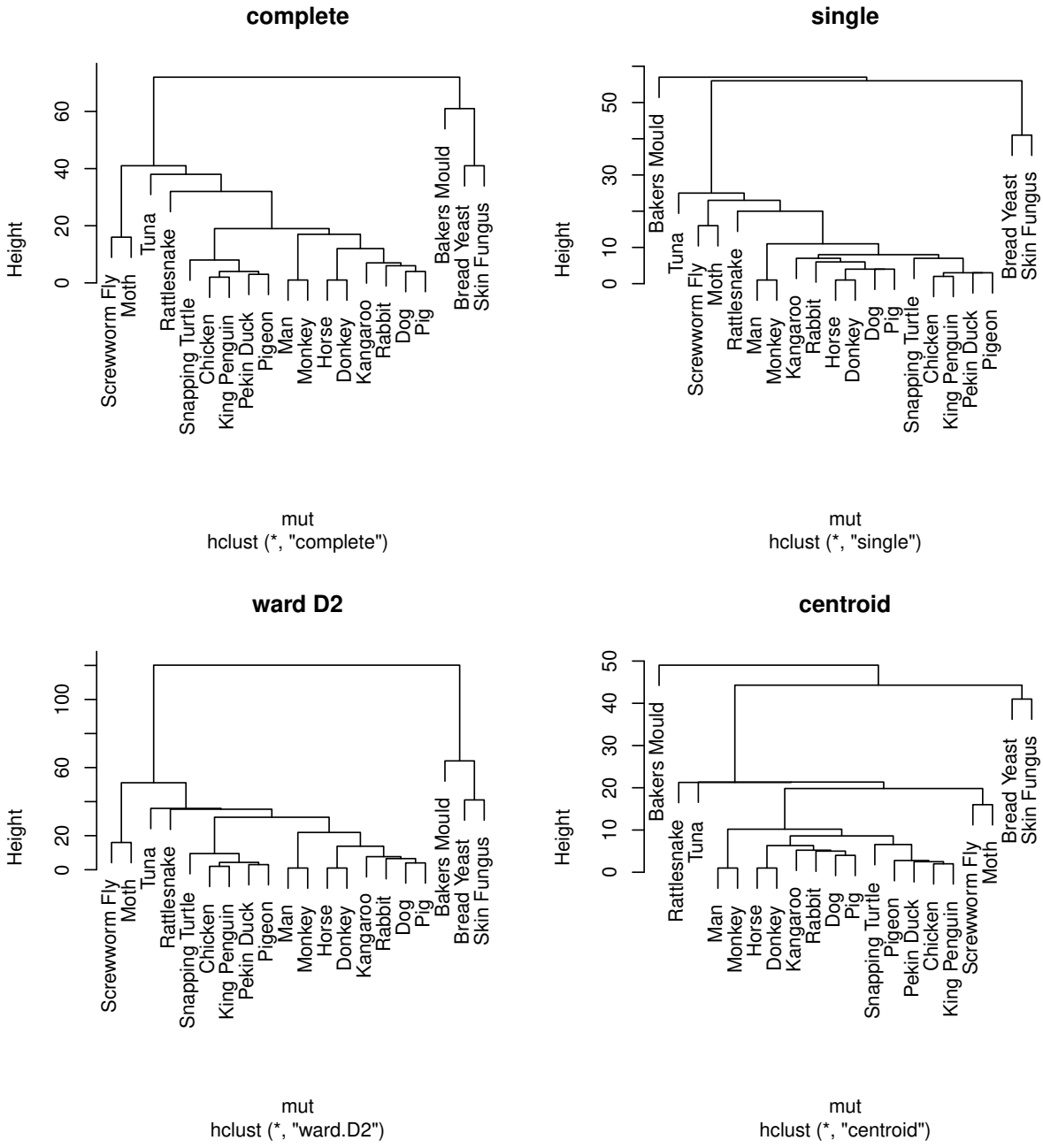

FIGURE 5 – Classification hiérarchique ascendante selon différents critères d'aggregations

2.

Avec la classification ascendante hiérarchique (Ward D2) sur Iris (6), on obtient 3 groupes hiérarchiques distincts, ce qui correspond bien aux trois espèces d'Iris. De plus, on remarque que la sous hiérarchie à droite groupe les Versicolor et Virginica ensemble, qui sont proches (d'après l'ACP).

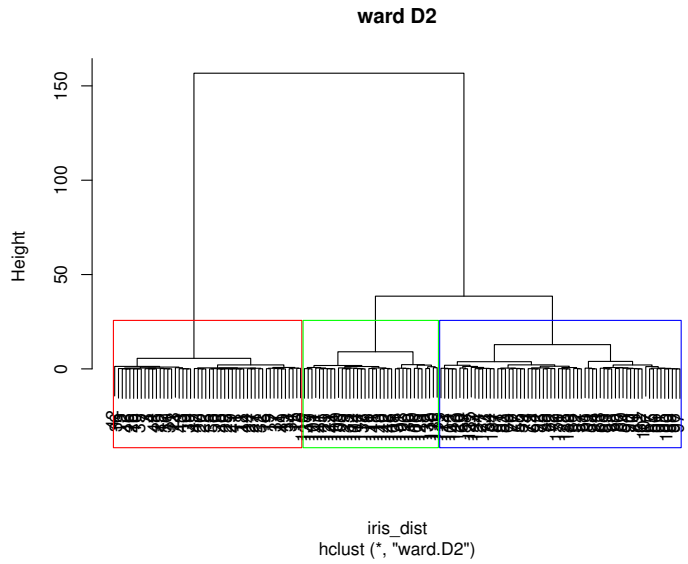

FIGURE 6 – Classification hiérarchique ascendante

3.

Avec la classification hiérarchique descendante, on part du cas général pour différencier des individus.

La classification est (7) donc adéquate, car on distingue bien les trois espèces d'Iris. De plus, on remarque que la sous-hiérarchie à droite groupe les Versicolor et Virginica ensemble, qui sont proches (d'après l'ACP).

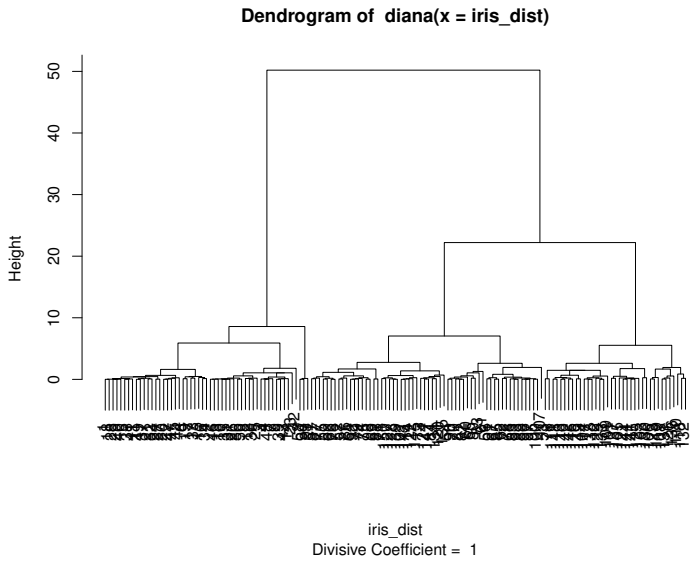

FIGURE 7 – Classification hiérarchique descendante

3 Méthodes des centres mobiles

Dans cette dernière partie, nous appliquerons l'algorithme des K-Means sur les jeux de données précédents pour tenter de trouver le nombre de classes optimales pour classifier les individus.

3.1 Données Iris

1.

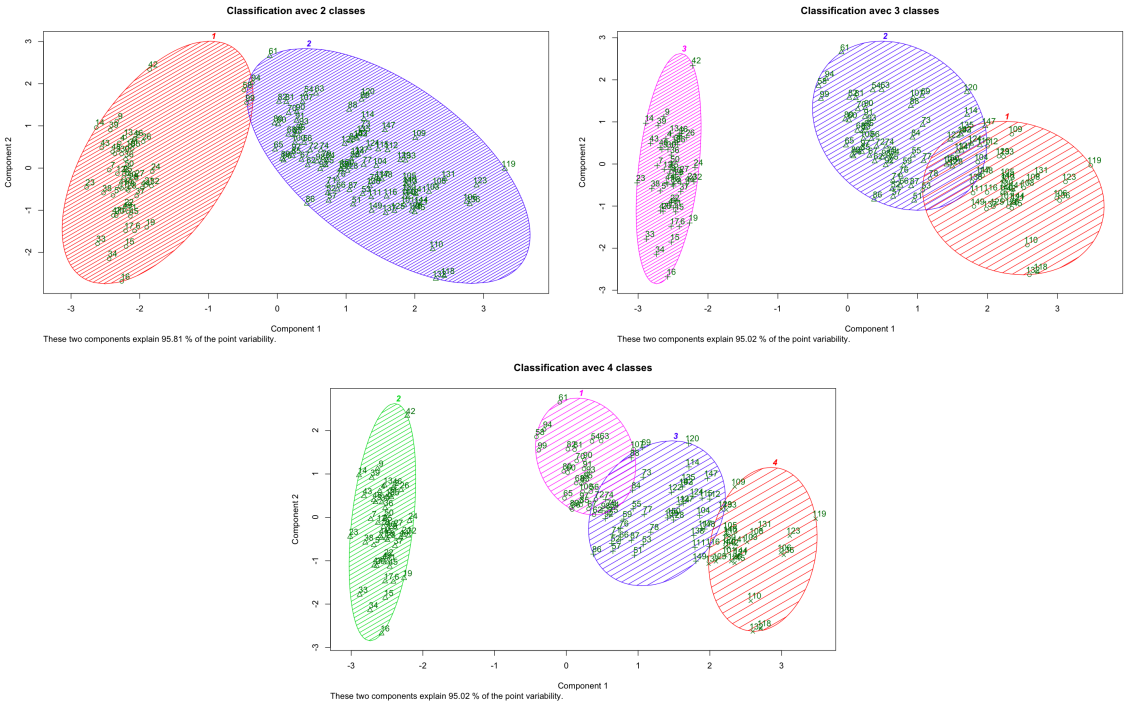

FIGURE 8 – Classification des données iris pour $k \in [2;4]$

Nous avons appliqué l'algorithme des k-means effectué avec $K \in [2;4]$ (figure 8).
Sachant que le jeu de données a 3 espèces, la partition obtenue pour K=3 classifie bien les individus dans la classe correspondante à l'espèce. Cela vérifie la partition réelle *setosa*, *versicolor*, *virginica*.
Pour K=2, k-means fonctionne bien et combine les catégories *versicolor* et *virginica* car ces espèces sont proches, comme nous l'avons vu précédemment.
Pour K=4, la classification partitionne les deux espèces de droite (*versicolor*, *virginica*) en 3 catégories.

Il apparaît donc très important de choisir une valeur de K optimale pour l'algorithme des K-Means pour bien interpréter les données et aussi car nous ne sommes pas censés connaître le nombre de classes.

2.

Après avoir effectué plusieurs répétitions de l'algorithme des k-means, on observe que les inerties intra-classes diffèrent et donc les classifications changent aussi.

On trouve deux partitions différentes (9). La classification à gauche est une classification optimale plus fréquente que l'autre.
L'inertie intra-classe de la classification optimale vaut *78.85*, et l'autre vaut *142.75*.

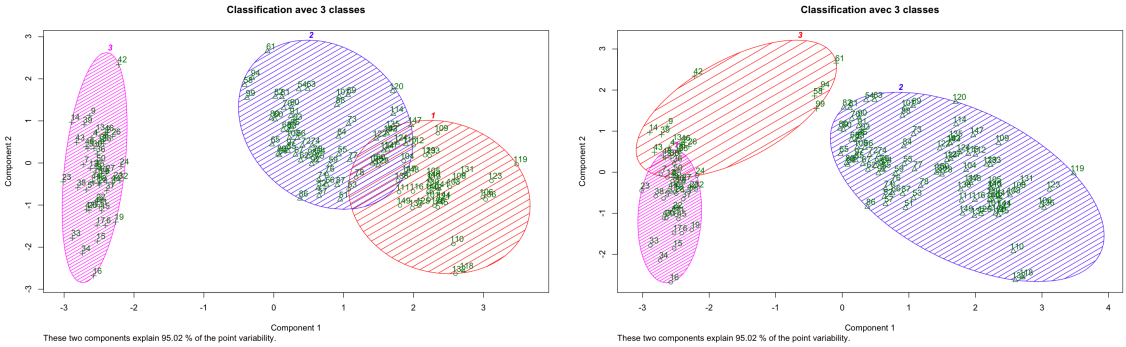

FIGURE 9 – Classification des données iris pour k=3

Cette variation dans le résultat est due à la sélection aléatoire des points initiaux des centres au début de l'algorithme, qui nous donne un minimum local et pas forcément un minimum global.

3.

Après avoir effectué 100 répétitions de l'algorithme des K-means, pour $K \in [1;10]$, on trace l'inertie intra-classe minimale en fonction de K sur les 100 répétitions (figure 10).

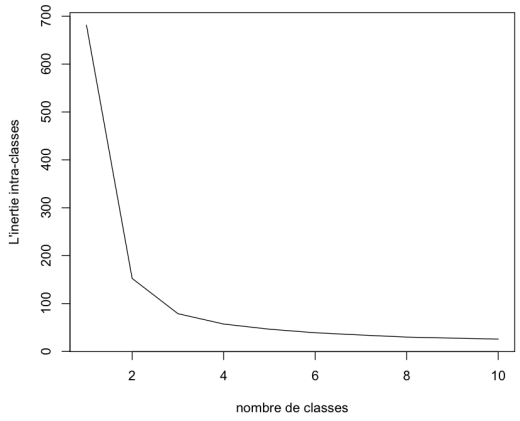

FIGURE 10 – Inertie minimale en fonction de K

D'après la méthode du coude, on trouve qu'une partition à K=2 ou K=3 classes est possible. On propose donc de classifier ce jeu de données par 3 classes, car celle-ci possède une classification avec une inertie intra-classe assez faible.
Pour un nombre de classes égal ou plus grand que 3, l'inertie est un peu plus faible, mais pas de beaucoup donc ces partitions ne sont pas intéressantes.

4.

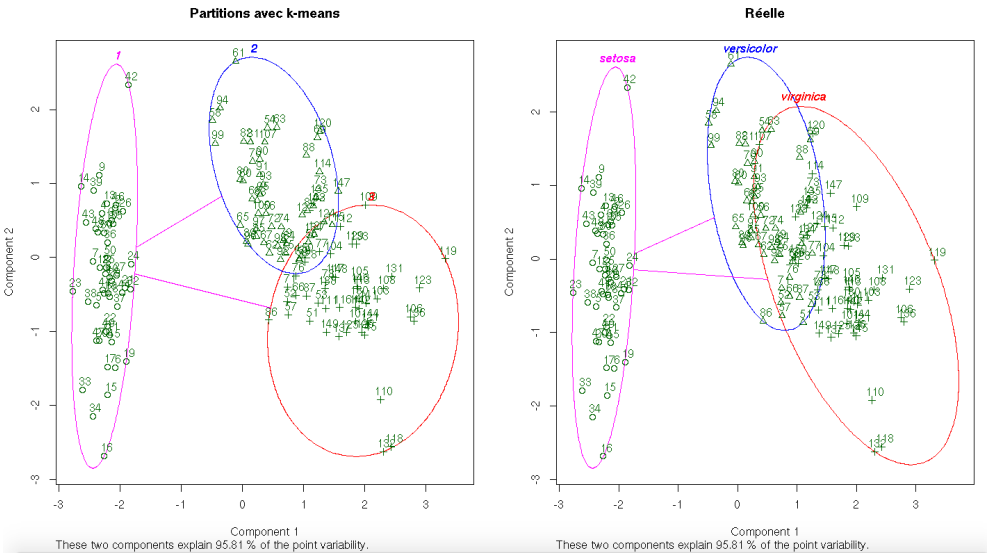

FIGURE 11 – Comparaison du résultat de k-means et la partition réelle

On peut observer que la partition obtenue par les centres mobiles et la partition réelle sont similaires sur la figure 11, à condition que l'on choisisse bien la valeur de K.
La groupe 1 obtenu par k-means est bien identifié par l'espèce *setosa*, mais les groupes 2 et 3 se recouvrent un peu.
Le résultat est donc satisfaisant.

3.2 Données Crabs

1.

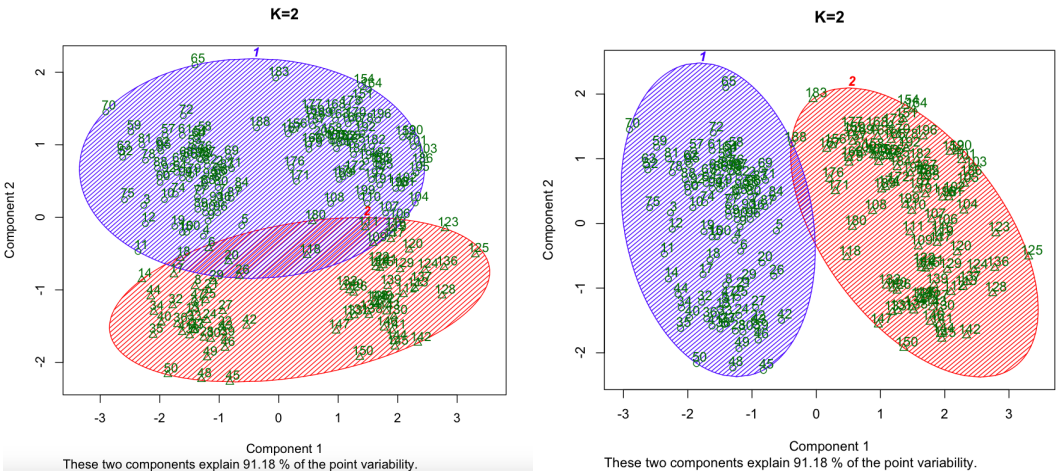

FIGURE 12 – Classification de crabs par k-means avec 2 classes

Les résultats obtenus ne sont pas toujours les mêmes. On trouve 2 différentes partitions. Autrement dit, il existe 2 valeurs minimales locales pour l'inertie intra-classes.
D'après la figure de l'exercice précédent (figure 2), on en déduit que la classification à gauche correspond à la classification du sexe et la classification à droite correspond à l'espèce.

2.

On applique maintenant l'algorithme des K-Means avec K=4

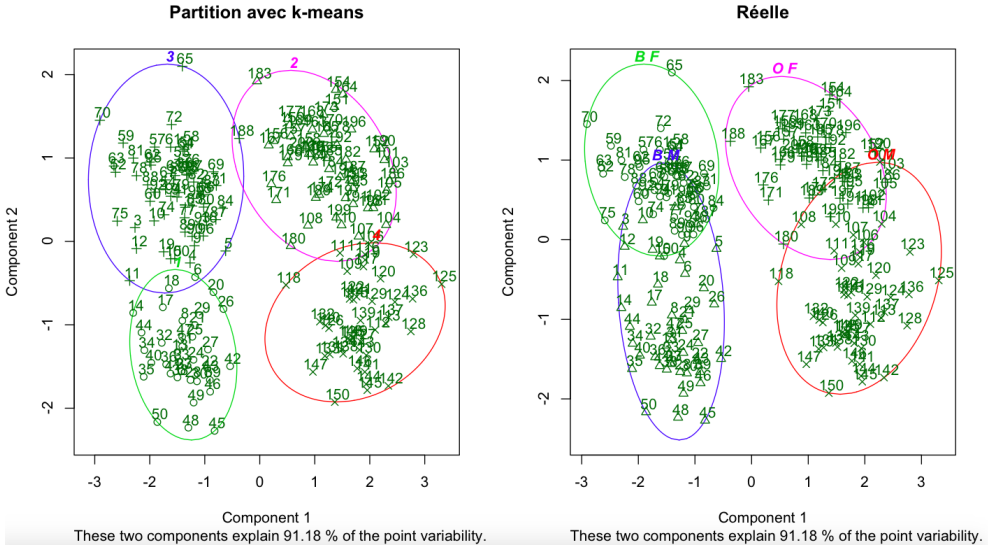

FIGURE 13 – La classification par k-means et la partition réelle

L'étude des 2 graphiques 13 nous montre que la partition obtenue par les k-means correspond bien à la partition réelle si l'on choisit la bonne valeur pour K. Il y a des erreurs de classification due à la similarité de certains individus et du chevauchement dans la classification réelle. Globalement, la classification est similaire à la partition réelle.

3.3 Données Mutations

1.

On applique l'algorithme des K-Means sur le jeu de données Mutations en cherchant si différentes classifications existent pour K=3.

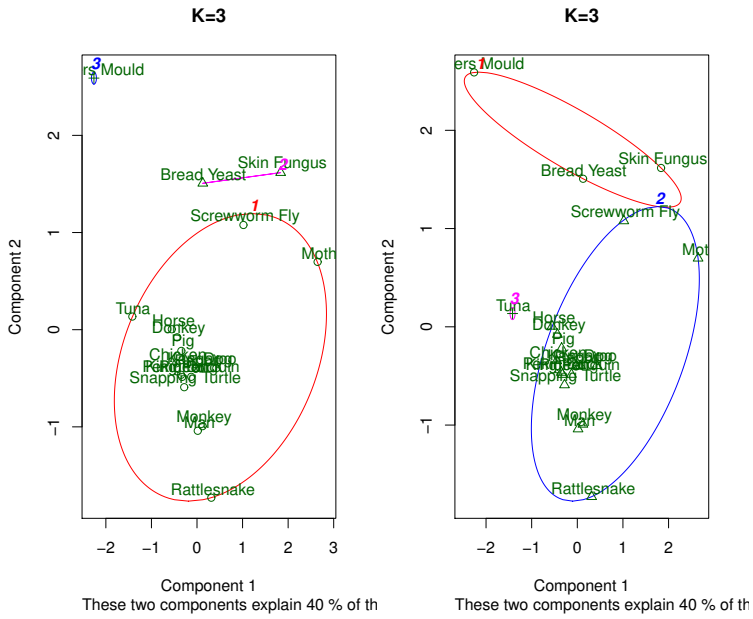

FIGURE 14 – Classifications à 3 classes avec K-Means

Nous avons donc trouvé deux classifications différentes (14).

2.

Pour étudier la stabilité des classifications trouvées, on calcule la variance pour chaque K variant de 2 à 5 (table 1). La variance la plus faible correspond à la partition K la plus stable.

On représente d'abord l'inertie intra-classe en fonction de K de la même manière qu'en [3.1].

La figure 15 montre clairement que la partition optimale par la méthode du coude est pour K=2.

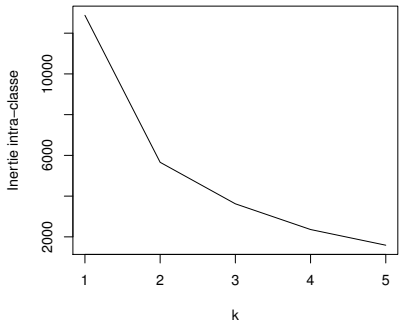

FIGURE 15 – Inertie intra-classe en fonction de K

En calculant la variance des inerties intra-classe pour chaque K, on remarque qu'elle est nulle pour K=2, donc cette partition est parfaitement stable.

On choisit donc d'appliquer un K-Means avec K=2 sur les données mutations (16).

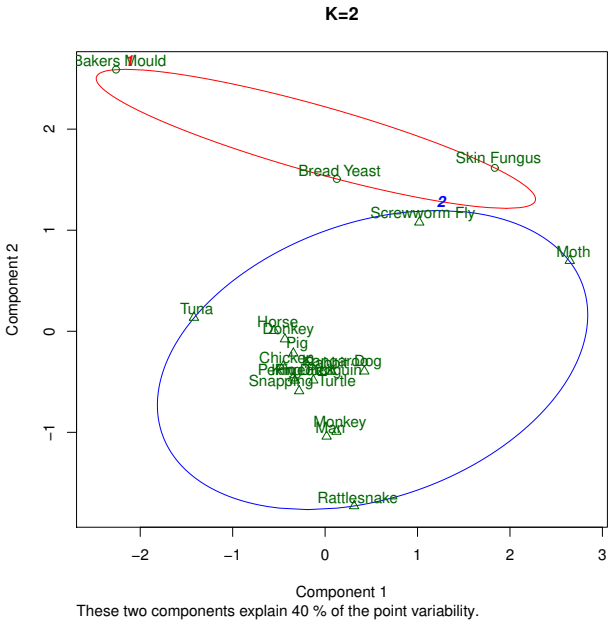

FIGURE 16 – Classifications à 2 classes avec K-Means

Conclusion

Ce TP nous a permis d'appliquer différents algorithmes de classification automatique afin de chercher à regrouper des données ensemble. Nous avons cherché à trouver le nombre de classifications optimales, qui représente le mieux les données.

La visualisation des données nous a aussi été utile pour avoir une première intuition sur nos

données et leurs groupements possibles. De plus, cela nous a aidés à choisir un nombre de partitions finales.

La classification automatique permet donc d'aborder les problèmes de l'apprentissage non supervisé en identifiant des patterns dans les données.

Références

[1] Hierarchical Clustering
[2] K-Means Clustering
[3] Quick-R: Cluster analysis

Annexes

TABLE 1 – Variance des inerties intra-classe pour différents k

k	k=2	k=3	k=4	k=5
Variance	0	365494	537365	468997