SY09 TP2
Classification automatique

Thibault Subreville & Yuting Chen

P17

L'objectif de ce TP est d'apprendre à manipuler les différentes méthodes de classification automatique. Nous utiliserons dans la suite de ce TP l'Analyse en Composantes Principales (ACP), la méthode d'Analyse Factorielle d'un Tableau de Distances (AFTD), la classification hiérarchique et la méthode des centres mobiles. L'outil d'analyse statistique utilisé est le logiciel R.

1 Exercice 1: Visualisation des données

1.1 Données Iris

Les données Iris sont composées de 150 spécimens de fleurs décrites par 5 variables dont 4 quantitatives (longueur et largeur du pétale et du sépale) et une variable qualitative: l'espèce.
Tout d'abord, on charge les données Iris pour réaliser une ACP avec **iris_acp = princomp(iris[,1:4])**. Nous enlevons la dernière colonne qui correspond à l'espèce car c'est une donnée qualitative. Ensuite, on réalise un **biplot(iris_acp)** pour obtenir la représentation pour le premier plan factoriel. Au niveau des axes, on remarque que **Petal.Width** et **Petal.Length** sont corrélés;

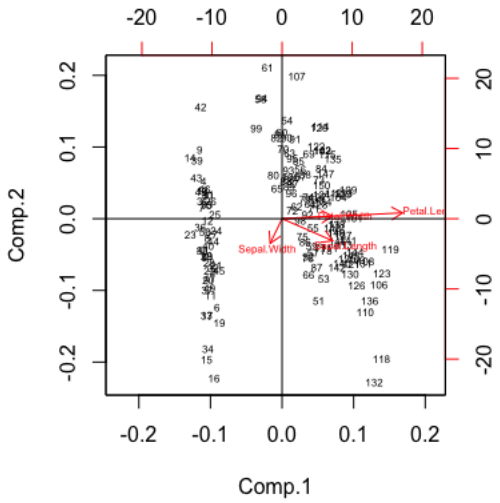

Figure 1: ACP des données Iris dans le premier plan factorielle

cela s'explique par **l'effet de taille** car ces deux variables décrivent la taille d'un pétale d'une fleur. En revanche pour **Sepal.width** et **Sepal.Length** les deux variables ne sont que très faiblement corrélés de par la présence d'un angle presque droit. Un phénomène qui pourrit être

vraisemblablement une propriété unique de ce type de fleur.
Après représentation des données dans le premier plan factoriel, on observe distinctement deux groupes de points, le premier à gauche de l'axe des ordonnées et le deuxième à droite de ce même axe. On constate directement que l'hypothèse de différencier 3 groupes tombent à l'eau

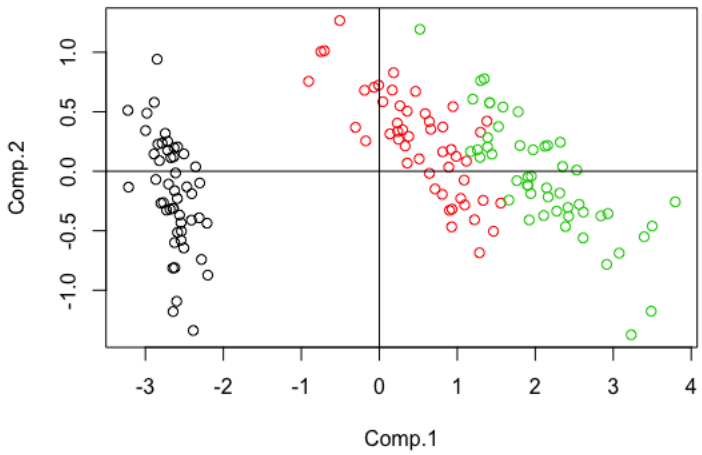

Figure 2: ACP des données Iris colorés en fonction de l'espèce

car nous distinguons 3 espèces (rouge: Versicolor, noir: Setosa , vert:Virginica) au lieu de 2. Si on recherche une partition de données on pourrait s'attendre à n'obtenir une partition de données uniquement composé de 2 groupes stables alors que la réalité est tout autre: on dispose de 3 espèces qui correspondent aux 3 couleurs ci-dessus.

1.2 Données Crabs

Les données Crabs sont composés de 6 variables dont 4 quantitatives et 2 qualitatives. Pour leur description complète nous faisons directement référence au "Compte Rendu" dernier.
On a réalisé une ACP sur les données Crabs avec la même méthode qu'à la section précédente. La représentation dans le premier plan factoriel, nous donne les figures suivantes:
Légende: vert: Bleu+Male, rose: Bleu+Femelle, bleu: Orange+Male, rouge: Orange+Femelle

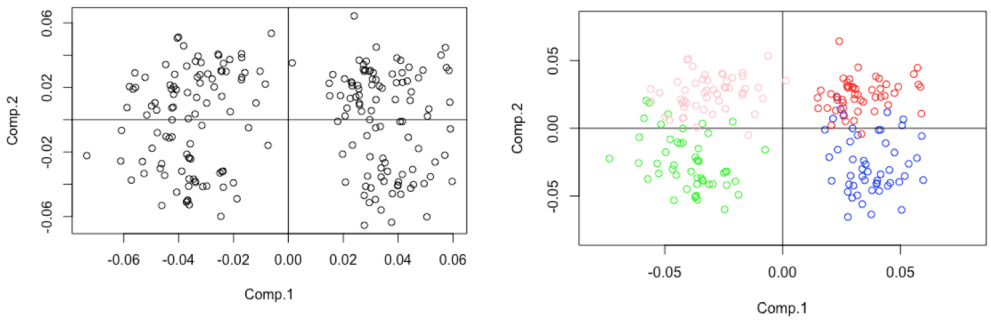

Figure 3: ACP des données Crabs dans le premier plan factoriel

Sans l'aspect des couleurs, on pouvait encore ici distinguer 2 groupes bien séparés sans aucune hésitation; ces deux groupes correspondent aux 2 espèces Bleu et Orange. La distinction du sexe est ici plus difficile à déterminer sans les couleurs.

1.3 Données Mutations

Le jeu de données est une matrice de dissimilarité (**mut**) contenant 20 espèces. Les distances ont été calculées en fonction du nombre d'acides aminés différents entre les espèces par rapport au Cytochrome c, une protéine.
Nous réalisons une AFTD à 2 variables des données de mutations avec la commande: **mut.aftd = cmdscale(mut, k = 2, eig = T)** et la fonction **plot**. Nous obtenons les diagrammes

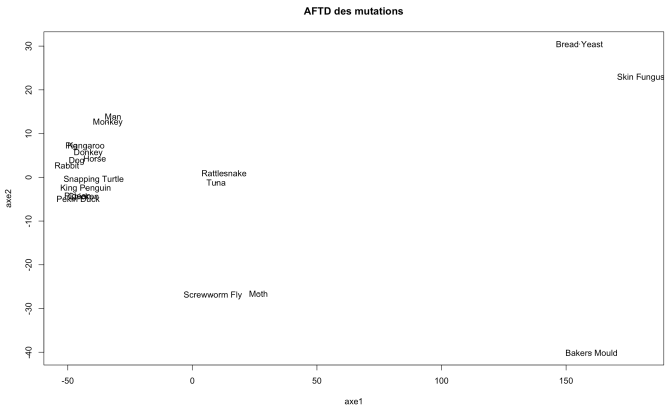

Figure 4: AFTD des mutations entre espèces

de Shepard pour un nombre de variable allant de 2 à 5. Cette compilation de diagrammes de

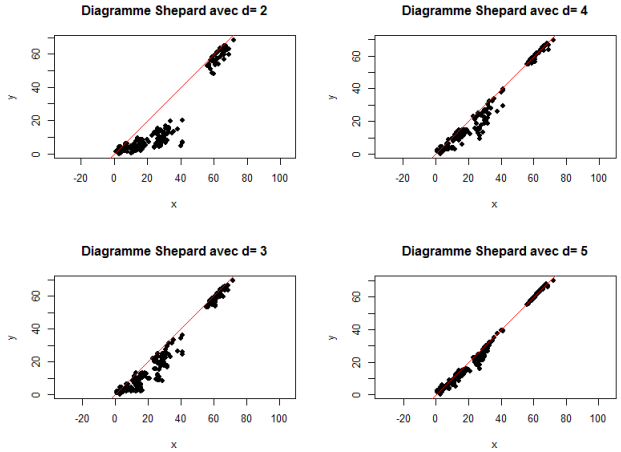

Figure 5: Diagramme de Shepard pour d allant de 2 à 5

Shepard permet de mesurer la différence entre les distances originales et les distances déterminées par l'AFTD; par conséquent la qualité de la représentation. Ce résultat est en accord avec le tableau suivant:

Nombre de variables	2	3	4	5
Pourcentage d'inertie expliqué %	0.70	0.81	0.88	0.93

Qualité de la présentation en fonction du nombre de variables à 10^{-2} *près*

Les premières valeurs propres sont positives donc on constate que les le pourcentage d'inertie croit en fonction du nombre de variables. Sur les 20 valeurs propres, les 6 dernières sont négatives, cela est concordant le fait que nous ne sommes pas dans un espace d'étude euclidien (les valeurs de la matrice de distance étant toutes entières, cela nous avait déjà donné une idée sur les propriétés de l'espace d'étude).

2 Exercice 2: Classification hiérarchique

2.1 Données Mutations

Avec la fonction hclust, il existe plusieurs critères d'agrégations :
— *ward.D*: Avec la méthode de liaison de Ward, la distance entre deux groupes est égale à la somme des écarts moyens quadratiques entre les points et les centres. Le but de la liaison de Ward est de minimiser la somme des carrés à l'intérieur du groupe.
— *single*: Avec la méthode de liaison single (dite du "voisin le plus proche"), la distance entre deux groupes est égale à la distance minimale entre une observation d'un groupe et une observation de l'autre groupe.
— *complete*: Avec la méthode de liaison complète (dite du "voisin le plus éloigné"), la distance entre deux groupes est égale à la distance maximale entre une observation d'un groupe et une observation de l'autre.
— *median*: Avec la méthode de liaison médiane, la distance entre deux groupes est égale à la distance médiane entre une observation d'un groupe et une observation de l'autre.
— *mcquitty*: Avec la méthode de liaison de McQuitty, lorsque deux groupes sont réunis, la distance entre le nouveau groupe et tout autre groupe est calculée comme étant la moyenne des distances entre cet autre groupe et les groupes à réunir prochainement.
— *centroid*: Avec la méthode de liaison du point central, la distance séparant deux groupes est la distance entre les centres ou les moyennes des groupes.
— *average*: Avec la méthode de liaison de la moyenne, la distance entre deux groupes est égale à la distance moyenne entre une observation d'un groupe et une observation de l'autre.

Nous obtenons les hiérarchies suivantes selon la suivante:

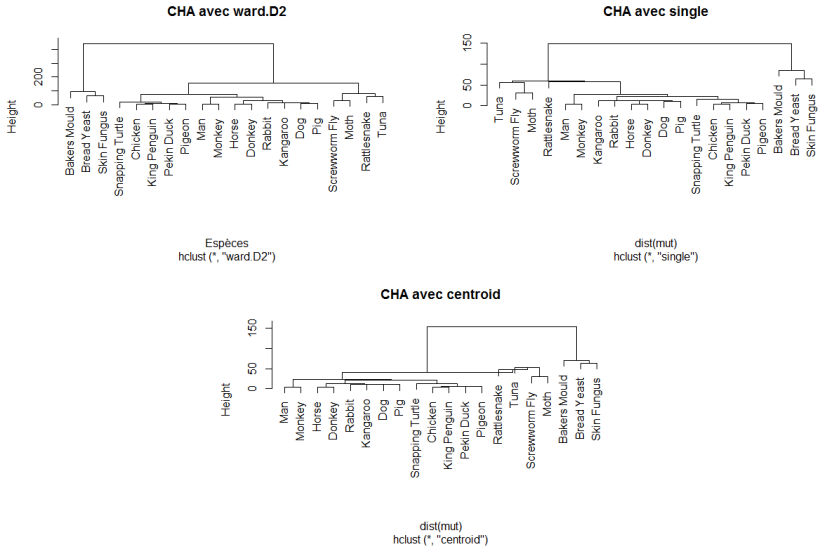

Figure 6: Application des différentes méthodes de classification

Selon les méthodes, elles présentent toutes des similitudes, nous pouvons remarquer que la classification hiérarchique ascendante avec les critères d'agrégation WARD, Complete, Mcquitty,

Average sont similaires. 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 car nous nous basons sur ce critère ici. L'ensemble des représentations selon la méthode de classification est disponible en annexe de ce document.

2.2 La classification hiérarchique ascendante des données Iris

Nous allons à présent appliquer le même traitement que précédemment aux données Iris, nous utilisons le critère d'agrégation WARD, en ayant pris soin de donner à chaque espèce une couleur pour pouvoir mieux distinguer les clusters.
Setosa: noir, Versicolor: vert, Virginica: rouge

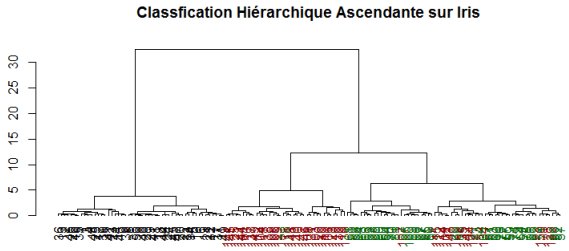

Figure 7: Application des différentes méthodes de classification

En effet, dans le cas du dendogramme issu de l'application de la méthode de Ward, nous voyons qu'à la hauteur 10, il y a bien trois groupes formés. Puis à partir d'une hauteur égale à 15 environ, deux groupes sont réunis et il y a alors les 2 groupes mentionnées plus haut. On remarque qu'on peut mettre en évidence trois principaux groupes, qui sont les trois espèces. Les Setosa qui se distinguent vraiment, et les Virginica et Versicolor se chevauchent.

2.3 La classification hiérarchique descendante des données Iris

Nous effectuons à présent une classification hiérarchique descendante des données Iris.

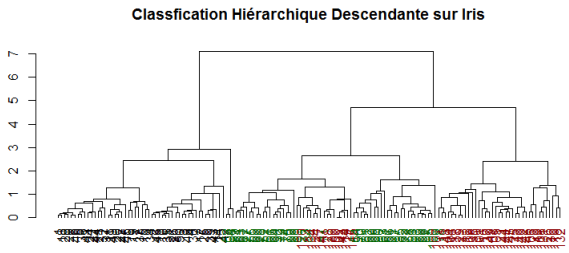

Figure 8: Application des différentes méthodes de classification

On retrouve bien les 3 groupes lorsque l'on se place à une hauteur de 4, dont 2 se rejoignent à une hauteur de 5 environ. En effet, les Setosa ne bougent pratiquement pas, les Virginica et les Versicolor qui se chevauchent encore plus que lors de la classification ascendante. Mais on remarque clairement mois de mélange entre espèces. Nous en déduisons que dans ce cas présent que la méthode ascendante donne de moins bon résulat dans la mesure où il y a moins de mélange d'espèces.

3 Exercice 3: Méthode des centres mobiles

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

3.1 Données Iris

Pour K=2, le cluster différencie l'espèce Setosa des deux autres espèces (Versicolor et Virginica) Nous remarquons que pour K=3, on retrouve le partitionnement par espèces.

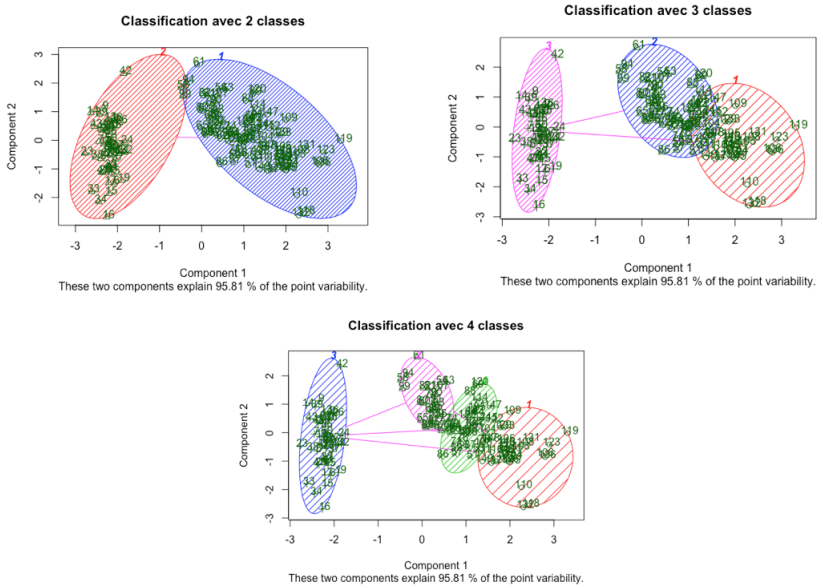

Figure 9: Représentation des classification pour un partitionnement allant de 2 à 4

Néanmoins, s'il on réalise un grand nombre de kmeans (supérieur à 30), nous obtenons des représentons différentes de la précédente telle que:

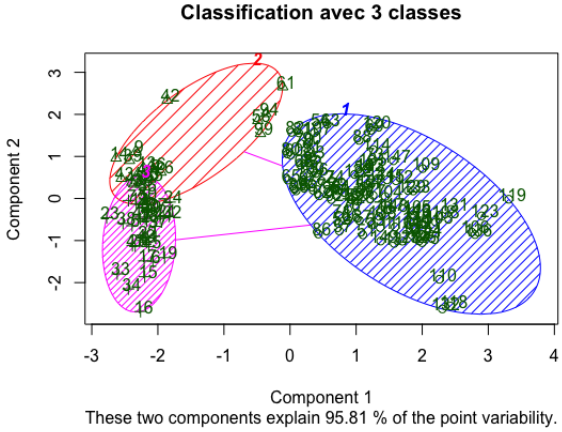

Figure 10: Une autre représentation de la fonction kmeans pour K=3

Cela s'explique par le fait que la fonction kmeans choisi des points au hasard pour réaliser la méthodes des centres mobiles. Le résultat de cette fonction peut donc être variable selon les initialisations. Nous remarquons néanmoins que cette représentation est moins bonne que la

précédente dans la mesure où l'inertie intraclasse est de 142.76 contre 78.86 pour la première représentation. La figure 10 apparait environ dans 20% des kmeans réalisé (sur un échantillon de 369 kmeans réalisé).

Pour K supérieur à 3, la fonction kmeans se voit contrainte de fusionner/scinder deux groupes, dès lors cela ne correspondant plus à aucune espèce connue, sauf pour l'espèce Setosa qui reste une partition à elle seule (partion bleu).

Ensuite, nous réalisons 100 kmeans par partitionnement allant de 2 à 10. Nous extrayons la valeur minimal des 100 kmeans, on obtient le tableau suivant:

Nombre de classes	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Inertie minimale	152.35	78.85	57.23	46.45	39.04	34.29	29.98	27.78	25.85

Table des sommes d'inertie intraclasse minimal en fonction du nombre de partitionnement à 10^{-2} *près*

La représentation graphique de ces données nous donne la graphique suivant:

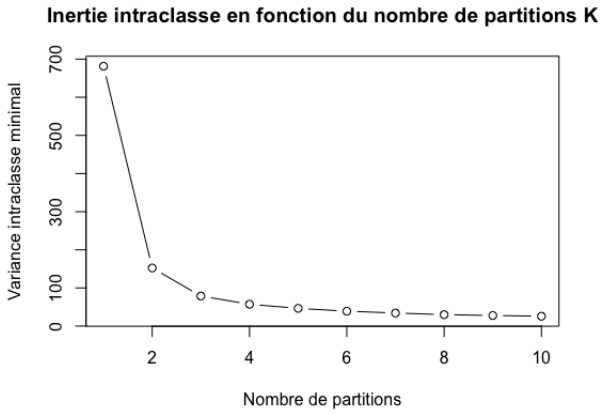

Figure 11: Graphique des sommes d'inertie intraclasse minimal en fonction du nombre de partitionnements

Pour déterminer le nombre optimal de classe pour le partitionnement nous utilisons la méthode du coude (heuristique). Cette méthode nous indique que le nombre optimal est 2 ou 3 classes. Par cohérence au vue du nombre d'espèces présente dans l'échantillon, nous sélectionnons la valeur de 3.

3.2 Données Crabs

Les partitions ne sont pas stables avec K=2, on se trouve dans le cas de la figure de droite avec une partition des Crabs selon le sexe alors que pour la la figure de droite (cf. figure 12), la partition est réalisé en fonction de l'espèce. La somme des inerties intraclasse dans le partitionnement par espèce est de 0.259 alors que pour celui par sexe est de 0.356 à 10^{-3} près.

Le partitionnement via l'espèce possède la variance intraclasse la plus faible, on peut donc

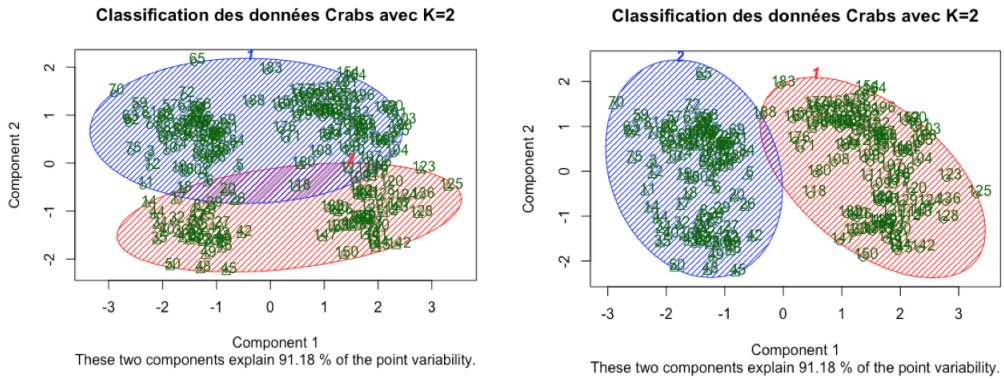

Figure 12: Résulats des classifications des données Crabs avec K=2

en tirer comme conclusion que c'est un meilleur partitionnement que par rapport au sexe. Ce résultat n'est guère étonnant car cela correspond aux deux groupes que l'on avait distingué dans la première partie de ce document. Dans le cas où K=4, le partitionnement en 4 classes corre-

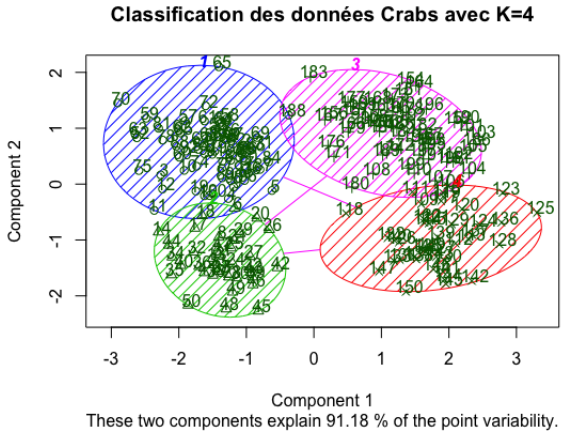

Figure 13: Résulats des classifications des données Crabs avec K=4

spond aux différents crabes en fonction de l'espèce et du sexe. Néanmoins, cette représentation n'est pas absolu car on remarque que si l'on réalise de nombreux kmeans, on trouve d'autres partitionnements possibles (avec une inertie intraclasse supérieur à plus de 60% par rapport à la représentation de la figure ci-dessus). La somme des variance d'inertie intraclasse pour ce partitionnement est de 0.106 à 10^{-3} près. La figure de moins bon partionnement est disponible en annexe de ce document est possède une probalilité d'apparition d'environ 20% (statistique déduite sur une échantillon de 300 kmeans).

3.3 Données Mutations

Après avoir réalisé une AFTD sur les données de **Mutations**, nous réalisons une centaine de kmeans pour pouvoir distinguer les différents partitionnement renvoyés par la fonction kmeans

avec K=3. En résultat, nous obtenons 2 partitionnements différents. La figure de droite ayant

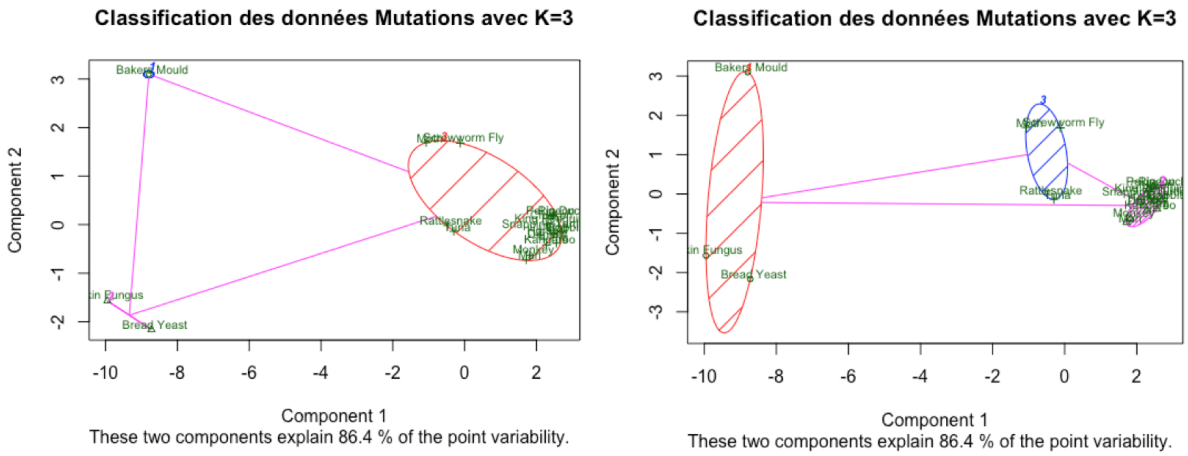

Figure 14: Résultats des classifications des données Mutations avec K=3

une variance intraclasse de l'ordre de 60% de celle de la figure de gauche. On en déduite que le partionnnement de droite est de ce fait de meilleur qualité suivant ce critère.

De la même manière que la section dernière si on trace le graphe des inerties intraclasse minimal et qu'on applique la méthode du coude, on remarque que la valeur optimal du nombre de classes est de 2 :

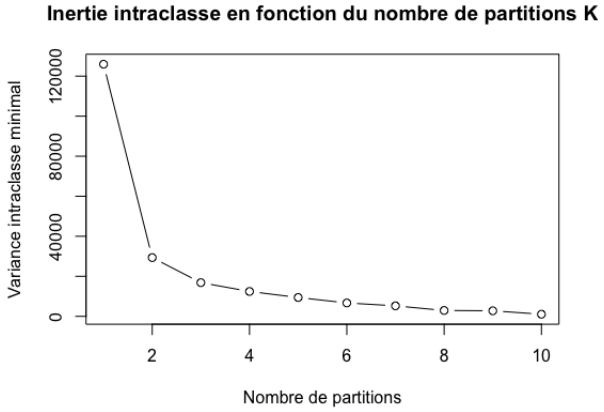

Figure 15: Variance intraclasse en fonction du nombre de partitionnements

Pour K=2, on observe une stabilité de la partition, la figure y est disponible en annexe.

4 Conclusion du TP

Ce TP nous permis de voir différents méthodes pour la classification automatique ascendante et descendantes ainsi que la méthode des centres mobiles. Nous avons pu remarquer que la méthode des centres mobiles est avant tout une heuristique et ne donne pas directement la meilleur partition possible. Ce TP nous a aussi permis d'analyser et d'interpréter des résultats dans le but de les améliorer.

5 annexe

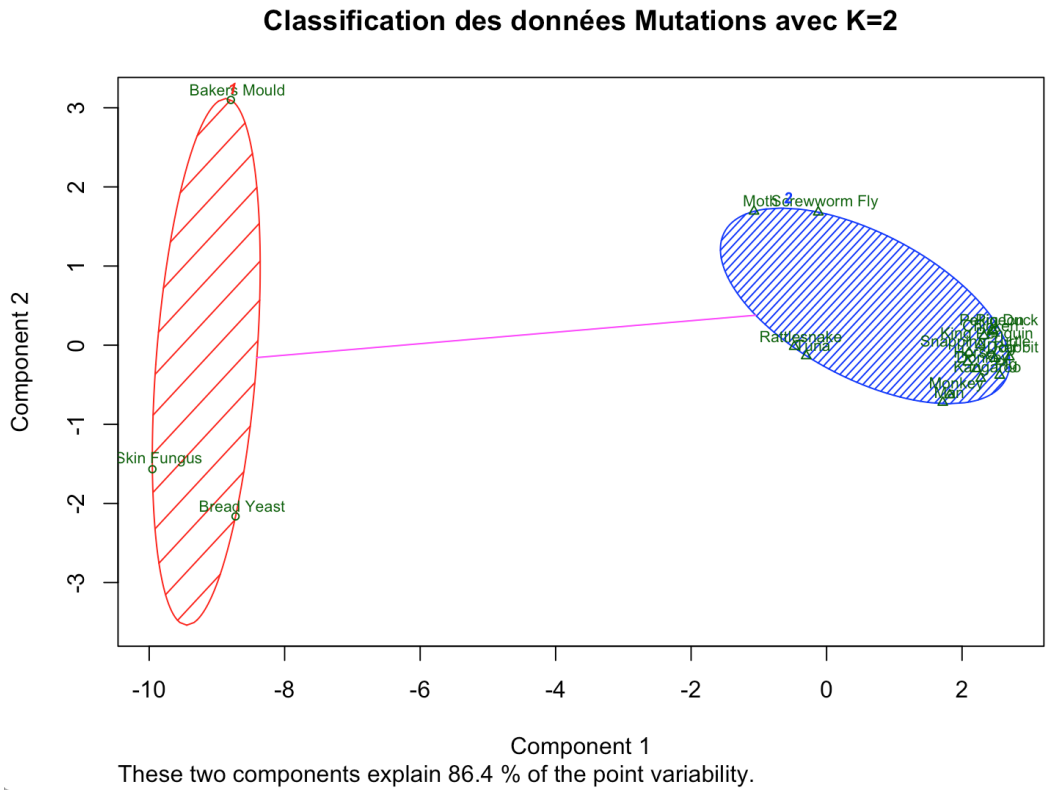

Figure 16: Classification (kmeans) des données Mutations avec K=2

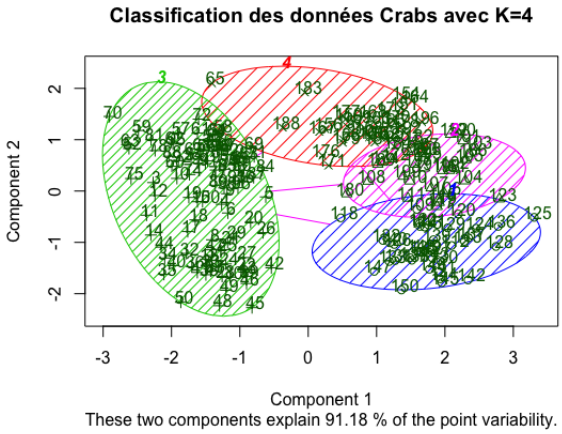

Figure 17: Autre résulat des classifications des données Crabs avec K=4

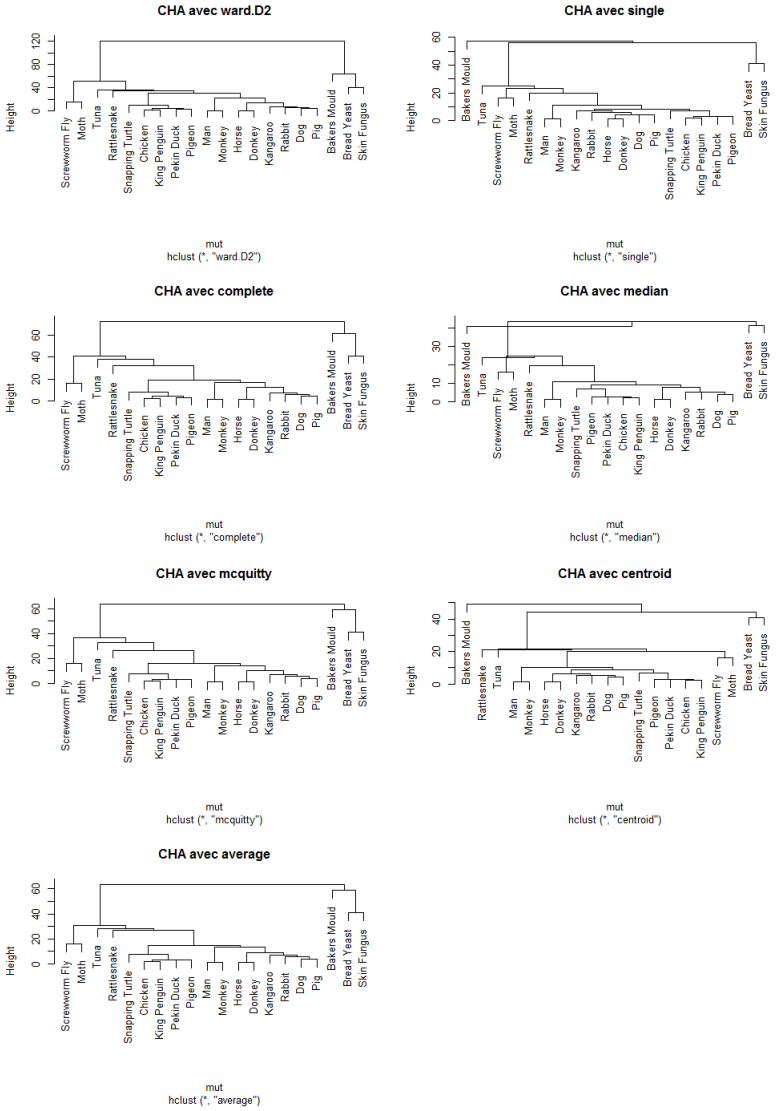

Figure 18: Application des différentes méthodes de classification