SORBONNE UNIVERSITES

16 mai 2017

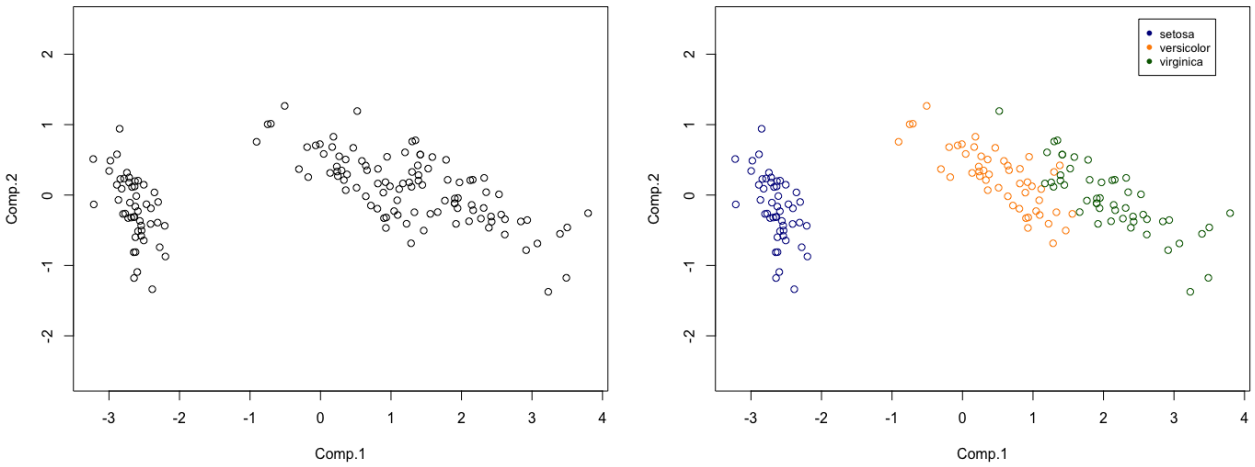
TP 2 - Classification automatique

UV : **SY09**
Branche : **Génie Informatique**
Filière : **Fouille de Données et Décisionnel**
Auteurs : **LU Han - HAMONNAIS Raphaël**

Table des matières

1. Visualisation des données

1.1 Visualisation des données Iris



(a) Données `Iris` dans le premier plan factoriel sans tenir compte de l'espèce

(b) Données `Iris` dans le premier plan factoriel en tenant compte de l'espèce

FIGURE 1.1 – Représentation des données `Iris` dans le premier plan factoriel après ACP

» Affichage dans le premier plan factoriel sans tenir compte de l'espèce

— On observe deux groupes bien distincts (voir Figure 1.1a).

» Affichage dans le premier plan factoriel en tenant compte de l'espèce

— On voit qu'un des deux groupes précédemment observé est en fait constitué de deux espèces différentes (voir Figure 1.1b) ;

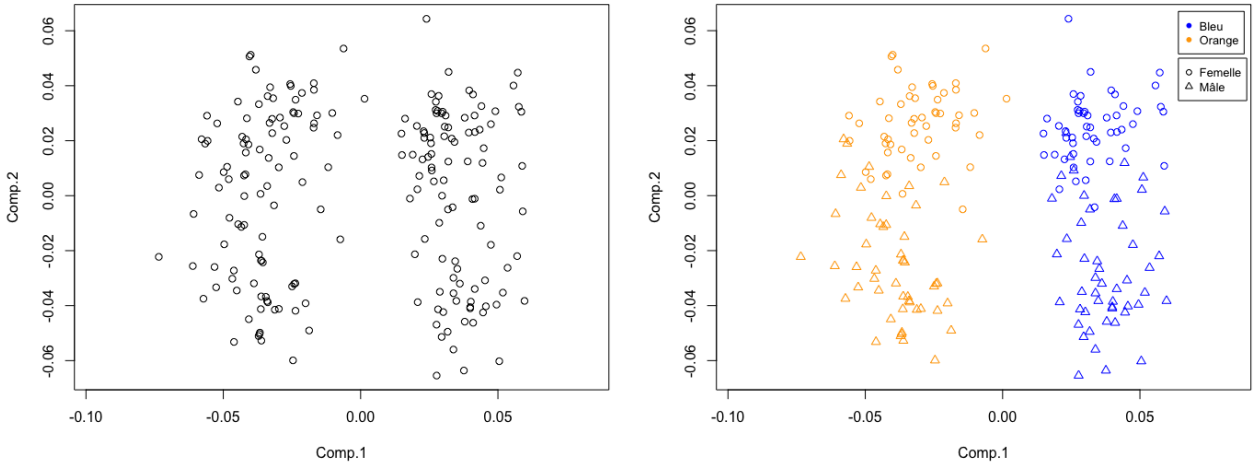
— On obtient donc deux informations précieuses :

— Les méthodes de classification géométriques tendront à nous donner deux classes ;

— On sait que les données contiennent en réalité trois classes bien distinctes quand le facteur discriminant est l'espèce ;

— Il faudra donc faire attention à ce qu'on cherche à obtenir : une nouvelle classification en X classes sans tenir compte de l'espèce, et auquel cas on obtiendra surement deux classes. Ou bien une classification en fonction de l'espèce et alors il faudra spécifier qu'on cherche à obtenir trois classes.

1.2 Visualisation des données Crabs



(a) Données `Crabs` dans le premier plan factoriel sans tenir compte de l'espèce ou du sexe

(b) Données `Crabs` dans le premier plan factoriel en tenant compte de l'espèce et du sexe

FIGURE 1.2 – Représentation des données `Crabs` dans le premier plan factoriel après ACP

- Affichage dans le premier plan factoriel sans tenir compte de l'espèce ou du sexe
  - On observe deux groupes bien distincts (voir Figure 1.2a).
- Affichage dans le premier plan factoriel en tenant compte de l'espèce et du sexe (voir Figure 1.2b)
  - On constate que les deux groupes observés précédemment correspondent à l'espèce des crabes ;
  - On voit aussi apparaître deux autres groupes au sein des premiers qui délimitent le sexe ;
  - On va donc chercher à faire une classification à 4 classes ;
  - On note quand même que la délimitation entre les sexes est plus floue que celle entre les espèces.

1.3 Visualisation des données Mutation

Les données `Mutation` représentent par le biais d'une matrice de dissimilarités les liens entre espèces : plus la distance (dissimilarité) est faible, plus les espèces sont proches.

Nous allons effectuer une Analyse Factorielle de Tableau de Distance (AFTD). On rappelle que l'AFTD peut être vue comme un équivalent de l'ACP pour des données se présentant sous la forme d'un tableau $n \times n$ de dissimilarités δ_{ij} entre n individus ($i,\ j \in \{1, ..., n\}$) : elle calcule une représentation multidimensionnelle de ces individus (dont le tableau de dissimilarités ne donne qu'une description implicite) dans un espace euclidien de dimension $p \leqslant n$. Cette représentation est exacte lorsque les dissimilarités sont des distances euclidiennes, ce qui n'est pas toujours le cas.

Après sélection d'un certain nombre de variables, la qualité de la représentation peut être évaluée numériquement par un critère similaire au pourcentage d'inertie de l'ACP, ou graphiquement au moyen d'un diagramme de Shepard : sur ce graphique, la distance $d_{ij} = d(x_i, x_j)$ entre les représentations de x_i et x_j déterminées par l'AFTD est représentée en fonction de la dissimilarité initiale δ_{ij}, pour chaque couple d'individus (x_i, x_j).

1.3.1 Données Mutation dans le premier plan factoriel après AFTD

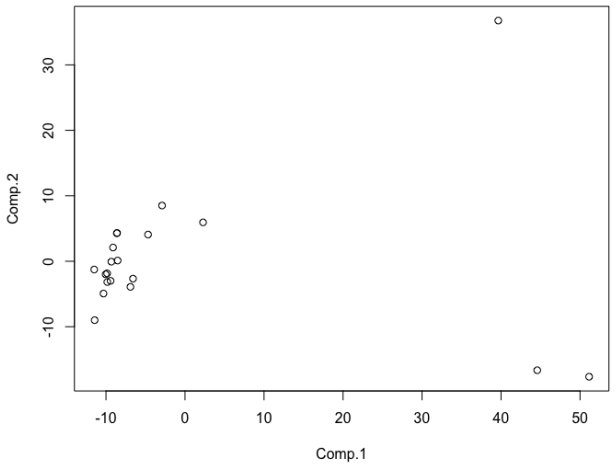


FIGURE 1.3 – Représentation euclidienne des données `Mutation` en deux dimensions par AFTD

Cette représentation a le mérite de nous permettre d'appréhender plus facilement le tableau de dissimilarités : on voit clairement que beaucoup des espèces sont proches les unes des autres. Certaines très proches. On voit aussi que trois d'entre elles sont particulièrement éloignées.

On pourrait former deux classes : la première regroupant l'ensemble des points proches, et la seconde les trois points éloignés. Ou bien trois, si on décide que l'espèce tout en haut à droite du graphique est trop loin pour être intégrée à une classe. Il est aussi tout à fait possible de subdiviser la première classe d'espèces proches les unes des autres en plusieurs classes plus petites.

1.3.2 Analyse de la qualité de la représentation par AFTD

Certaines des valeurs propres (inertie expliquée de la composante principale correspondante) sont négatives. Calculer le pourcentage d'inertie expliquée demande alors de faire un choix : transformer les valeurs propres négatives en leur inverse positif (valeur absolue) ou bien ne tenir compte que des valeurs propres positives. Nous avons ici effectué les calculs avec les deux possibilité afin de comparer les résultats (voir Tableau 1.1 et Tableau 1.2).

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Pourcentage d'inertie expliquée	52.71	16.01	10.94	6.72	4.78
Pourcentage cumulé d'inertie expliquée	52.71	68.72	79.67	86.38	91.16

TABLE 1.1 – Inertie avec valeurs absolues

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Pourcentage d'inertie expliquée	53.43	16.23	11.09	6.81	4.84
Pourcentage cumulé d'inertie expliquée	53.43	69.66	80.75	87.56	92.40

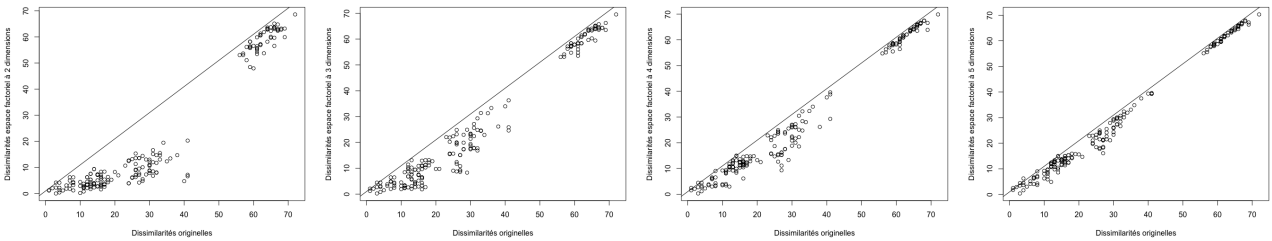
TABLE 1.2 – Inertie avec valeurs positives seulement

On remarque que les pourcentages d'inertie expliquée ne varient pas énormément entre les deux choix de calcul. Cela s'explique par la faible importance des valeurs propres négatives dans ce jeu de données.

Pour ce qui est de la qualité de la représentation, on voit très vite qu'avec simplement deux dimensions, on n'obtient qu'environ 69% d'inertie expliquée. C'est relativement peu pour une ACP, mais cela peut être

suffisant pour une représentation simple des données. Tenir compte de trois, quatre ou cinq dimensions nous permet à chaque fois de mieux représenter des données, avec respectivement 80%, 87% et 92% d'inertie cumulée.

On remarque aussi cette augmentation de qualité de représentation des données initiales lorsqu'on regarde les diagrammes de Shepard :



(a) Diagramme de Shepard (2 dimensions)

(b) Diagramme de Shepard (3 dimensions)

(c) Diagramme de Shepard (4 dimensions)

(d) Diagramme de Shepard (5 dimensions)

FIGURE 1.4 – Diagrammes de Shepard pour les dimensions 2 à 5

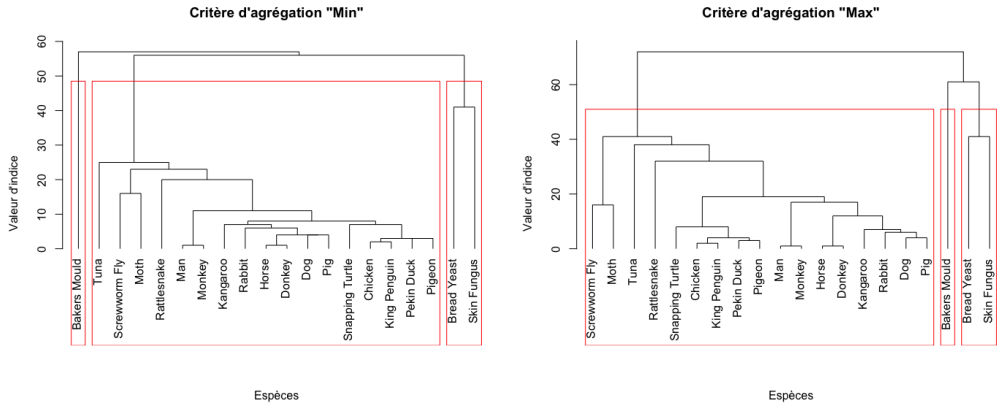
L'axe des abscisses représente les dissimilarités originelles avant AFTD et l'axe des ordonnées les distances euclidiennes entre les observations sur le nouvel espace factoriel obtenu par l'AFTD.

Plus le nombre de dimensions augmente, plus les valeurs des dissimilarités avant et après AFTD s'approchent de la droite $y = x$, c'est à dire une dissimilarité initiale égale à la distance entre les individus représentés dans l'espace euclidien à k dimensions défini par l'AFTD.

2. Classification hiérarchique

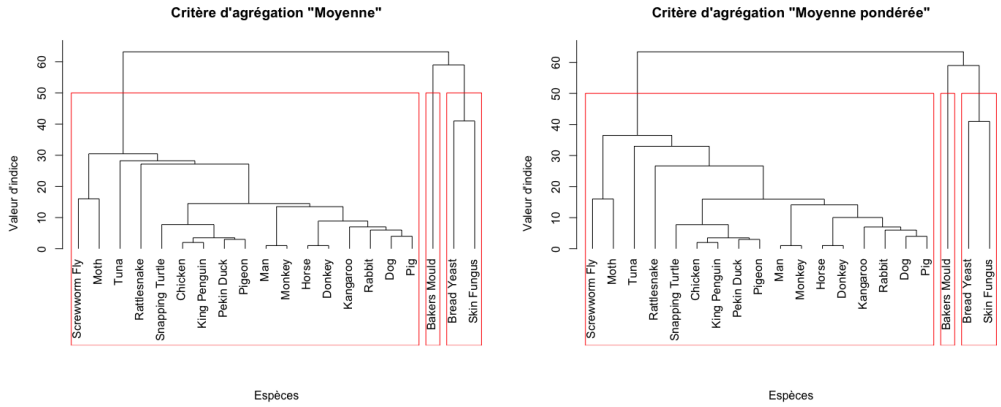
2.1 Classification hiérarchique ascendante sur les données Mutation

Remarque : certains critères d'agrégation sont « pondérés ». Cela signifie que les classes sont considérées comme étant de poids équivalents, quel que soit leur effectif.



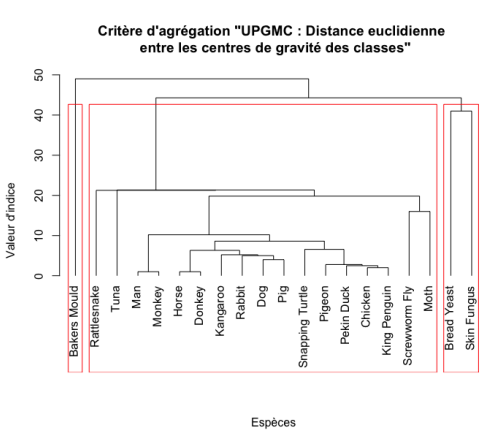
(a) Critère agrégation Min (b) Critère agrégation Max

FIGURE 2.1 – Classification avec les critères d'agrégation Min et Max

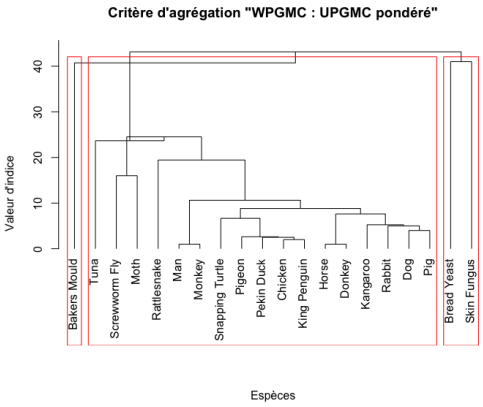


(a) Critère agrégation Moyenne (b) Critère agrégation Moyenne pondérée

FIGURE 2.2 – Classification avec les critères d'agrégation Moyenne et Moyenne pondérée



(a) Critère agrégation distance inter-classe



(b) Critère agrégation distance inter-classe pondérée

FIGURE 2.3 – Classification avec les critères d'agrégation de distance inter-classe (distance entre centres de gravité des classes)

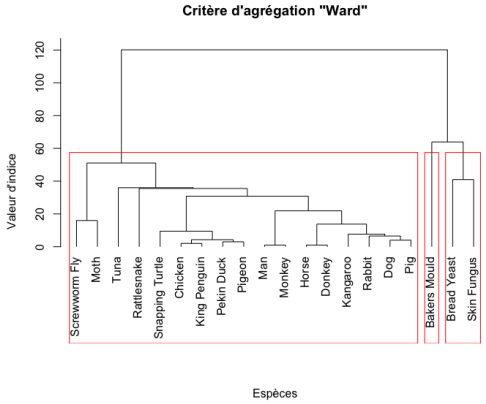


FIGURE 2.4 – Classification avec le critère d'agrégation de Ward

La première chose qu'on remarque est que, lorsqu'on divise en trois classes tel que nous le suggérait la représentation graphique des données après AFTD (cf. sous-section 1.3.1), tous les critères d'agrégation donnent les mêmes classes. A savoir {*Bakers Mould*}, {*Bread Yeast, Skin Fungus*}, et le reste dans la dernière classe.

Il est aussi intéressant de remarquer que deux des critères d'agrégation ne sont pas monotones (cf. Figure 2.3a et Figure 2.3b). On voit en effet que l'indice n'est pas décroissant.

2.2 Classification hiérarchique sur les données Iris

2.2.1 Classification hiérarchique ascendante

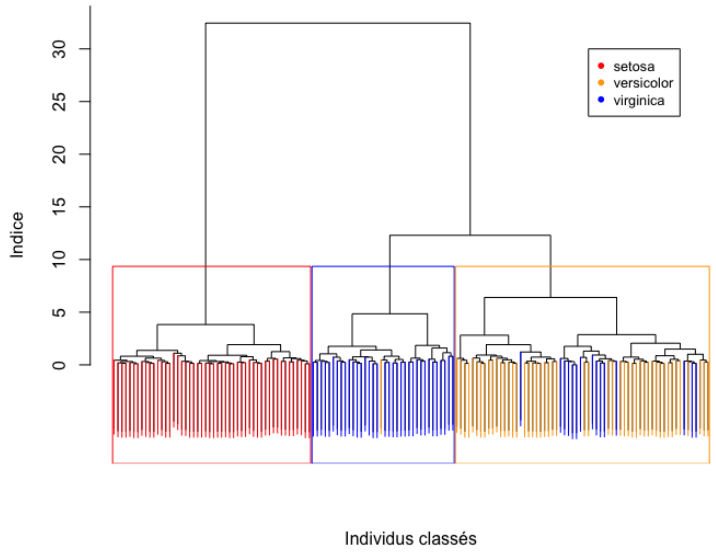


FIGURE 2.5 – Iris : classification hiérarchique ascendante

La classification identifie très bien les individus de l'espèce *setosa*. Par contre, elle tend à confondre certains individus des espèces *versicolor* et *virginica*. Compte tenu de la représentation graphique obtenue après ACP (cf. Figure 1.1b), ce n'est pas étonnant, ces deux espèces étant très proches sur le graphe, contrairement à *setosa* qui est bien distincte graphiquement parlant.

2.2.2 Classification hiérarchique descendante

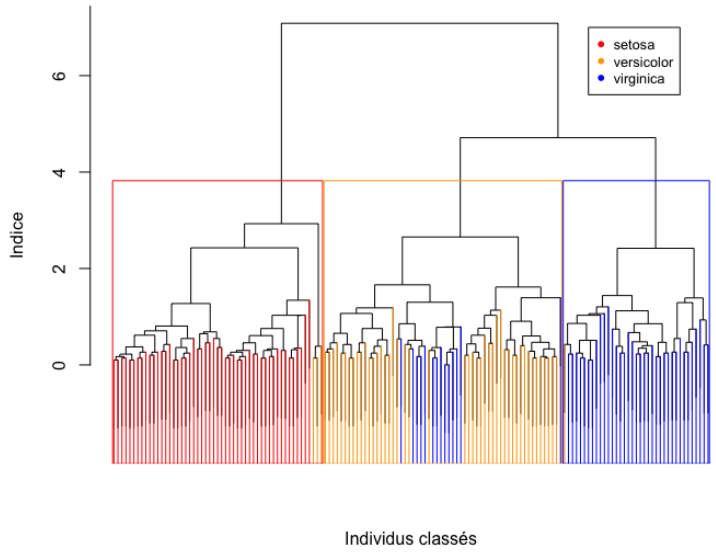


FIGURE 2.6 – Iris : classification hiérarchique descendante

La classification hiérarchique descendante confond elle aussi certains des individus des espèces *versicolor* et *virginica*. Mais elle classe en plus certains individus de l'espèce *versicolor* dans la classe contenant les individus

d'espèce *setosa*. Elle semble à priori moins précise que la classification hiérarchique ascendante.

Afin de savoir laquelle des deux méthodes reste la plus précise, nous utilisons l'indice de Rand corrigé. Le principe est de sélectionner l'ensemble des paires possibles d'individus au sein de chaque partition et comparer leur classement au classement initial qui est connu. La valeur de cet indice est comprise entre 0 et 1 : plus elle se rapproche de 1, plus la classification obtenue est proche de la classification initiale.

Valeurs obtenues :

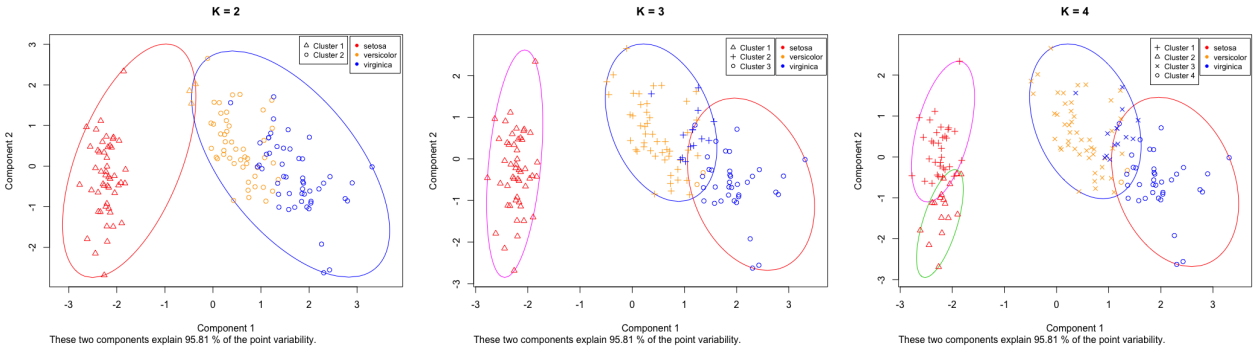
- Classification ascendante : 0.73
- Classification descendante : 0.69

On en déduit que la classification ascendante est meilleure pour ce jeu de données, bien que de peu.

3. Méthode des centres mobiles (K-Means)

3.1 Données Iris

3.1.1 Partition en $K \in 2, 3, 4$



(a) Centres mobiles avec K = 2 (b) Centres mobiles avec K = 3 (c) Centres mobiles avec K = 4

FIGURE 3.1 – Iris : classification à l'aide de l'algorithme des centres mobiles en 2, 3 et 4 classes

Classifications obtenues sans recherche d'un critère de sélection minimal :

- La séparation en deux classes détecte bien les deux groupements de points qui sont visibles (voir Figure 1.1a) ;
- La classification en trois classes semble elle aussi avoir du sens car elle identifie les trois différentes espèces (voir Figure 1.1b) ;
- La classification en 4 classes quant à elle divise la première en deux sous-classes différentes.

3.1.2 Étude de la stabilité du résultat

Afin d'étudier la stabilité du résultat, nous avons effectué 100 classifications consécutives. On s'aperçoit que l'algorithme donne toujours les deux mêmes classifications, présentées dans les tableaux Tableau 3.1 et Tableau 3.2

	Inertie intra classe	Fréquence d'apparition
Classification n° 1	0.53	76%
Classification n° 2	0.95	24%

TABLE 3.1 – Différentes inerties intra-classe obtenues

	1	2	3
setosa	0	50	0
versicolor	48	0	2
virginica	14	0	36

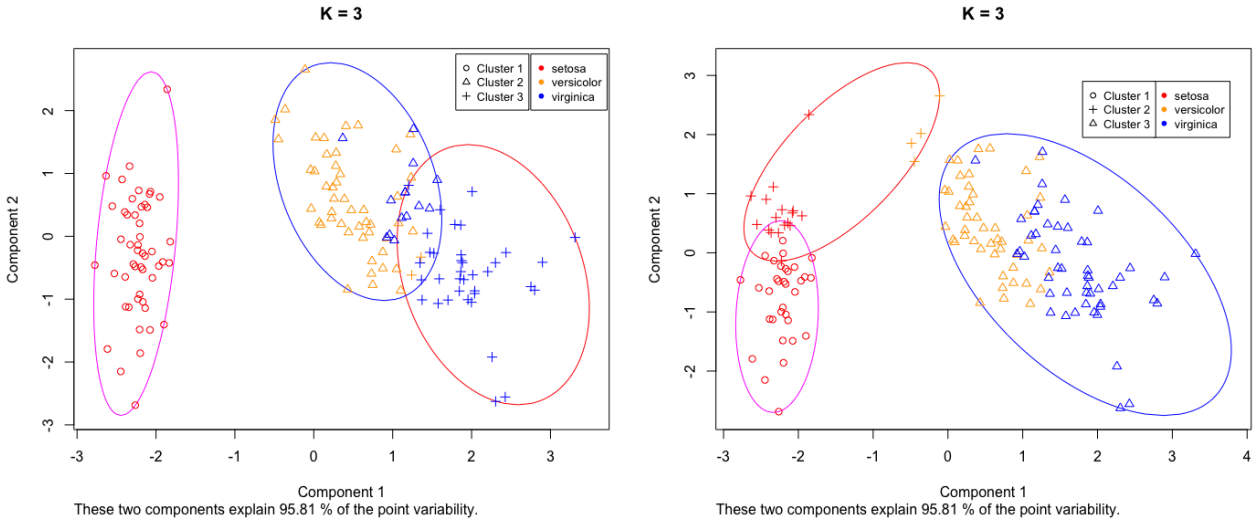
(a) Classification n° 1

	1	2	3
setosa	33	17	0
versicolor	0	4	46
virginica	0	0	50

(b) Classification n° 2

TABLE 3.2 – Différentes classifications en fonction de l'inertie intra-classe

La classification numéro 1 est la meilleure dans le sens où elle minimise la valeur de l'inertie intra-classe de la partition. On peut d'ailleurs s'en assurer en comparant les représentations graphiques des deux classifications obtenues (voir Figure 3.2) :



(a) Classification n°1

(b) Classification n°2

FIGURE 3.2 – Représentation des différentes classifications en fonction de l'inertie intra-classe

Même sans la connaissance préalable des données, la classification numéro deux semble bien moins pertinente que la première.

3.1.3 Détermination du nombre de classes optimal

La méthode du coude consiste à étudier la décroissance du critère (ici l'inertie intra-classe minimale) en fonction du nombre K de classes et choisir K avant le premier saut significatif.

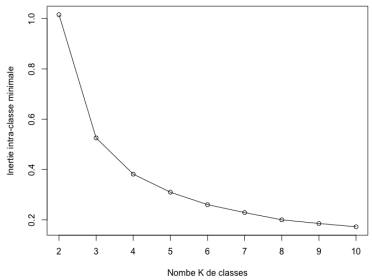


FIGURE 3.3 – Variation de l'inertie minimale en fonction du nombre K de classes

D'après la courbe de variation, il est évident que le passage de 3 à 2 classes présente une augmentation importante du critère. Remarquons aussi que la pente de la droite entre les points K=3 et K=4 est plus importante que lorsque K est supérieur à 4. On pourrait donc choisir 3 ou 4 comme étant le nombre optimal de classes, bien que 3 semble plus indiqué. La bonne solution est de tester avec les deux et de choisir ensuite en fonction des connaissances préalables des données que l'on a.

3.1.4 Comparaison de la classification obtenue avec la classification connue des trois espèces

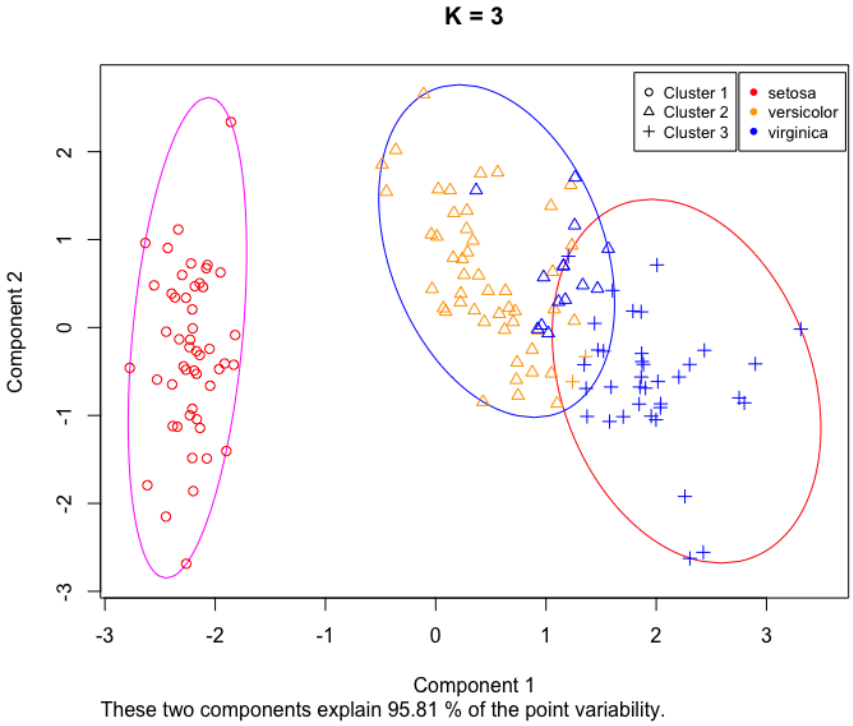


FIGURE 3.4 – Comparaison de la classification K-Means avec la classification initiale connue

On remarque que la classification par les K-Means fait sens mais qu'elle commet quelques erreurs à jonction des espèces *versicolor* et *viginica*, comme le faisait la classification hiérarchique.

3.2 Données Crabs

3.2.1 Classifications en 2 classes

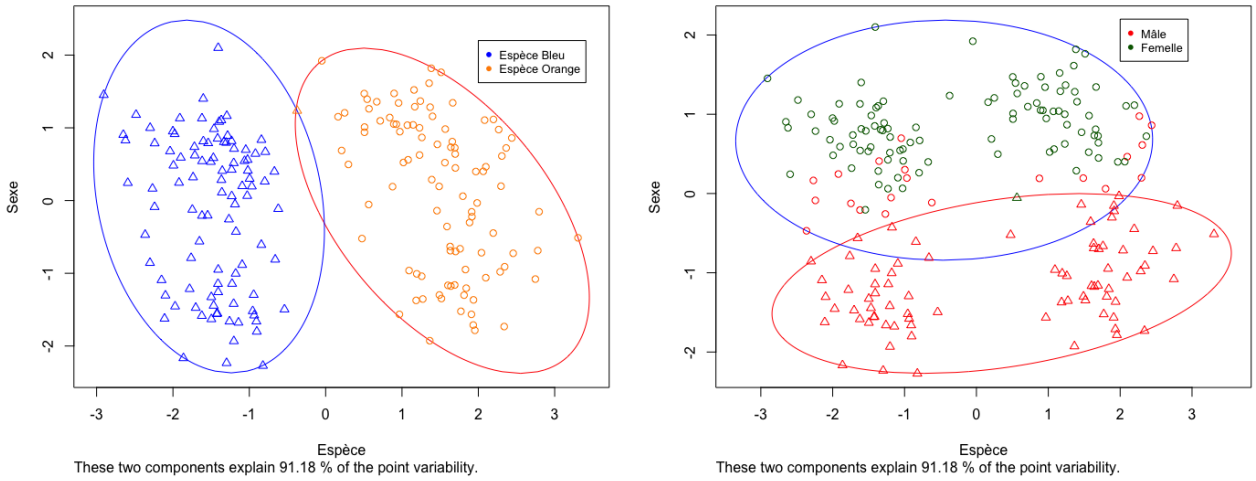
La classification en 2 classes n'est pas stable :

Classification n°	Inertie intra-classe	Pourcentage d'apparition
1	0.25878	73.80
2	0.35607	23.10
3	0.35754	1.00
4	0.358	1.20
5	0.3648	0.90

TABLE 3.3 – Les 5 classifications obtenues sur 1000 essais

Les deux premières classifications sont majoritaires. On remarquera que les classifications numéros 2 à 5 ont toutes une inertie intra-classe très proche les unes des autres : ce sont des « variations » d'une même classification.

Lorsqu'on représente graphiquement ces classifications, on s'aperçoit qu'elles mettent en évidence une des catégories de l'individu : son espèce (Figure 3.5a) ou son sexe (Figure 3.5b).



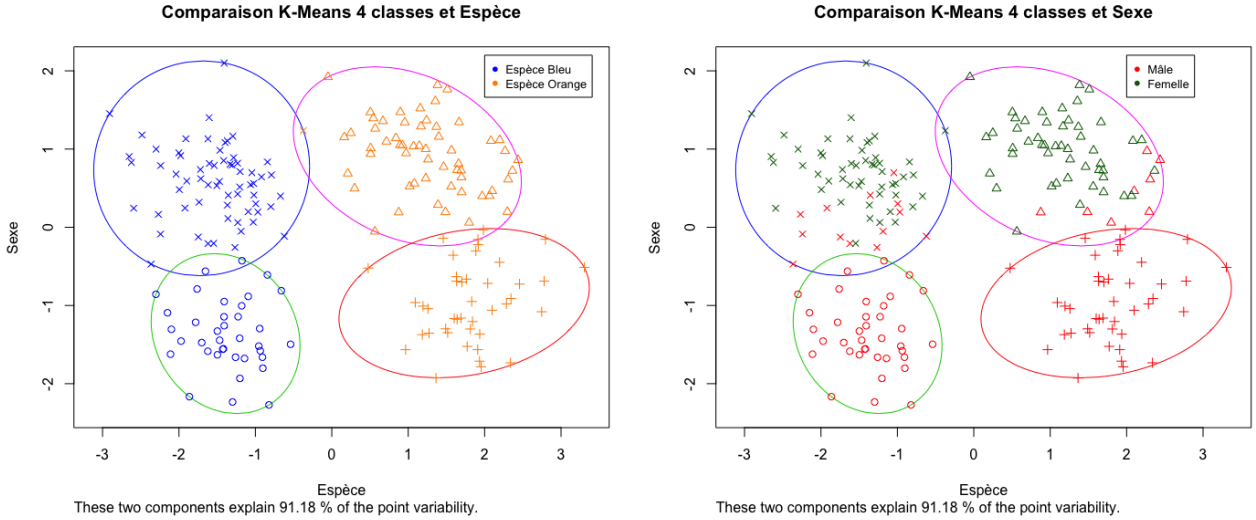
(a) Classification n°1

(b) Classification n°2

FIGURE 3.5 – Deux principales classifications des données `crabs` en deux classes

3.2.2 Classification en 4 classes

On choisit la meilleure classification pour $K = 4$, c'est à dire celle qui minimise l'inertie intra-classe.



(a) Comparaison avec l'espèce

(b) Comparaison avec le sexe

FIGURE 3.6 – Comparaison entre la meilleure classification en 4 classes (obtenue via la méthode des centre mobiles) et les classes « Sexe » et « Espèce » connues

On en conclut que la classification avec la méthode des centres mobiles est efficace : elle identifie bien les 4 groupes différents d'individus en fonction de leur espèce et de leur sexe. Notons tout de même que la différenciation sur le critère de l'espèce est plus efficace que celle sur le critère du sexe : seulement une erreur tandis qu'il y en a bien plus pour la différenciation du sexe des individus.

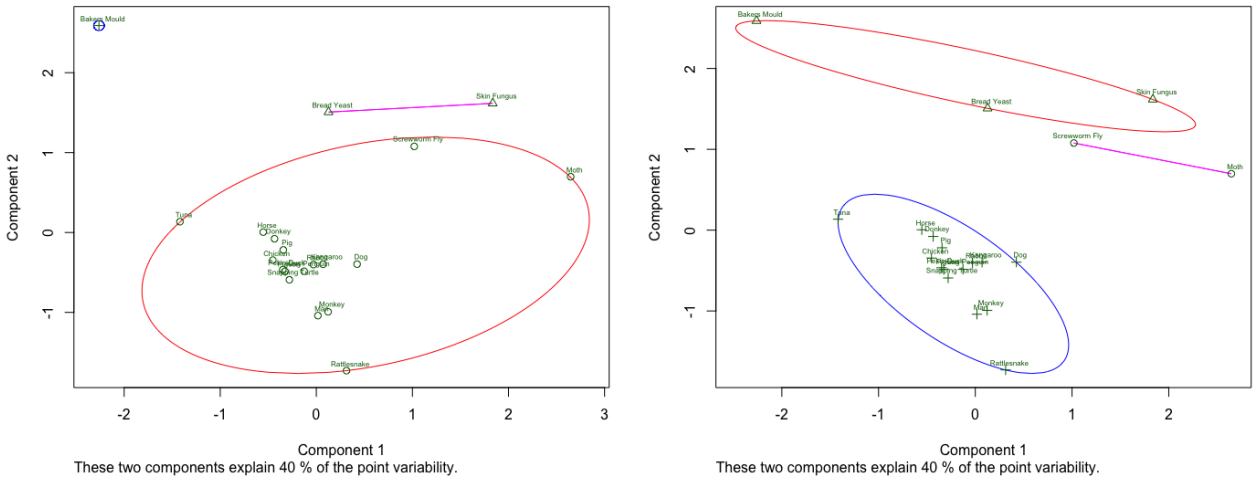
3.3 Données Mutations

La classification en 3 classes n'est pas stable :

Classification n°	Inertie intra-classe	Pourcentage d'apparition
1	181.092	30.90
2	219.91	18.00
3	245.904	24.60
4	248.75	11.30
5	254.623	8.80
6	261.861	6.40

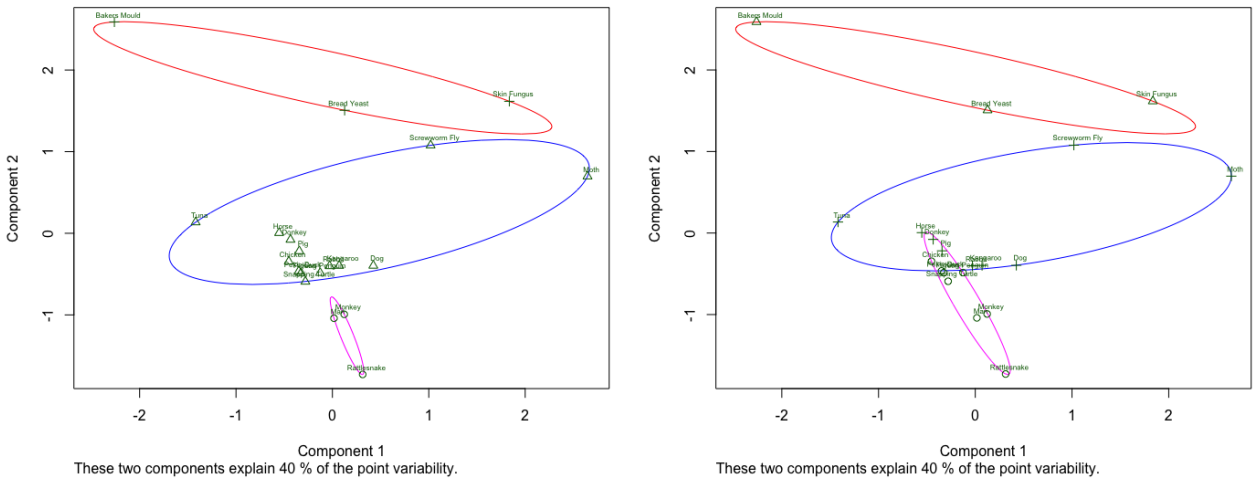
TABLE 3.4 – Classifications obtenues sur les données « Mutations » pour 1000 essais

Voici la représentation graphique de chacune des six classifications possibles avec la méthode des K-means :



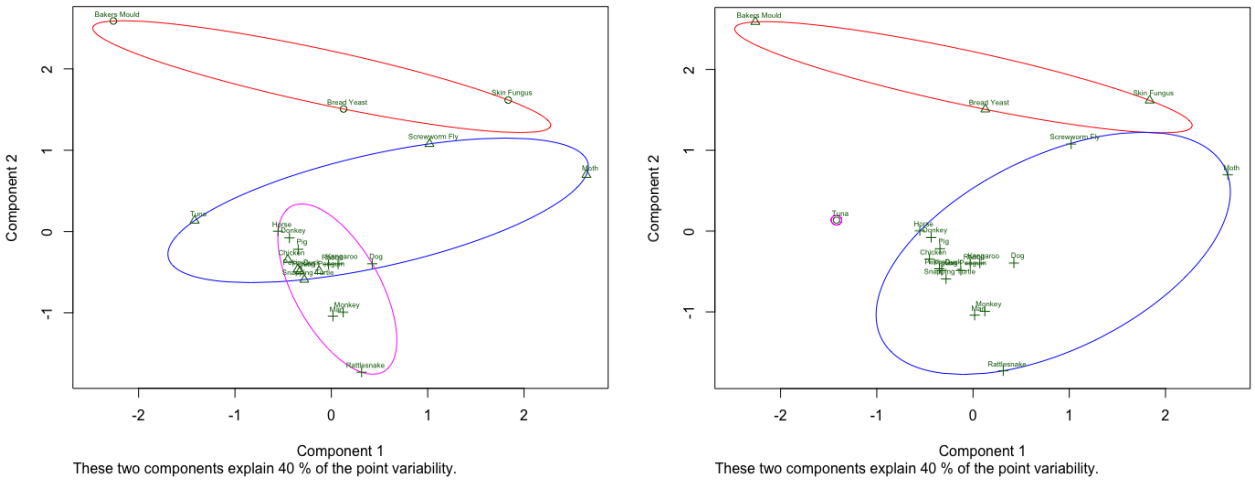
(a) Classification n° 1 (b) Classification n° 2

FIGURE 3.7 – Classifications 1 et 2



(a) Classification n° 3 (b) Classification n° 4

FIGURE 3.8 – Classifications 3 et 4



(a) Classification n° 5

(b) Classification n° 6

FIGURE 3.9 – Classifications 5 et 6

Lorsqu'on observe les représentations graphiques de ces classifications, on remarque plusieurs choses :

- La classification n°1 fait graphiquement sens et rejoint les résultats de la classification hiérarchique (voir section 2.1) ;
- La classification numéro 2 fait elle aussi sens : en regroupe certains des individus les plus éloignés et on recentre la classe des éléments proches les uns des autres ;
- La classification n°3, bien que d'inertie intra-classe plus importante que la n°2, apparaît plus souvent. Cette classification pourrait faire sens avec notre connaissance préalable des données : regrouper *Man* et *Monkey* est logique. Notons tout de même qu'il y a aussi le *Rattlesnake* avec, ce qui fait moins sens ;
- Les autres classifications sont visiblement mauvaises : chevauchements de classes, et singleton n'ayant graphiquement pas lieu d'être.