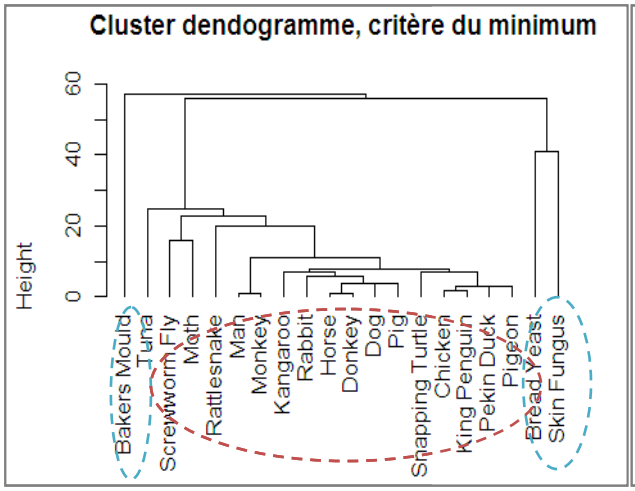
SY09 Printemps 2010 - TP 2
Classification Automatique

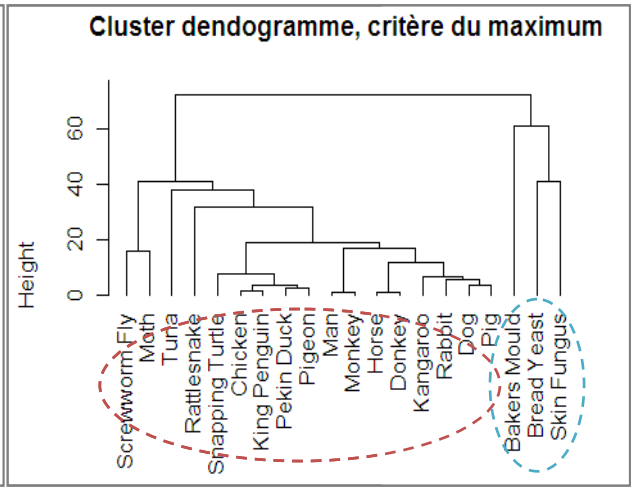
Le principal but de ce TP est d'utiliser la méthode de classification automatique, qui consiste à attribuer une classe ou catégorie à chaque individu à classer, en se basant sur des données statistiques, de manière à comparer ces résultats à ceux obtenus avec les méthodes vu lors des précédents TP.

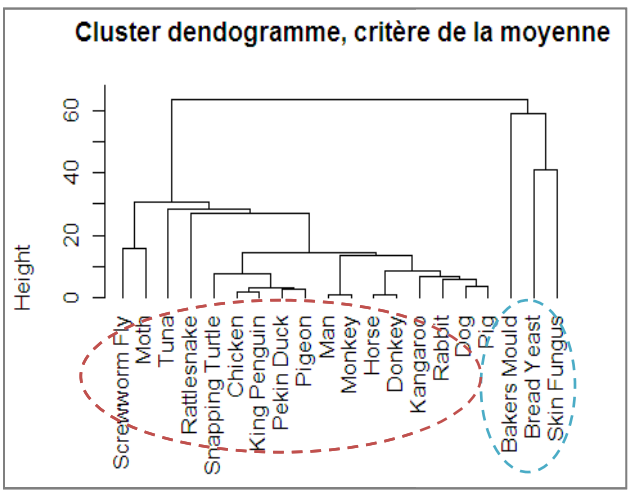
I. Classification hiérarchique

a) Classification hiérarchique ascendante

La classification est réalisée grâce à la fonction hclust. Les différents critères d'agrégation disponibles sont average, single, complete et ward. Average sert à calculer la distance moyenne entre tous les éléments d'une classe et ceux d'une autre classe, single correspond à la plus petite distance entre 2 classes séparant un individu d'une classe et un individu de l'autre, « complete » la distance maximum entre un individu d'un classe et un individu de l'autre.
Le jeu de données « mutation » est un tableau de distance, nous ne pouvons pas appliquer le critère de Ward, car il nécessite un espace euclidien.
Nous testons les différents critères grâce à la représentation en dendogramme. Sur ces graphiques, nous observons la composante height, qui correspond à un vecteur des distances entre les classes aux différents stades de la classification.







La difficulté du choix du critère d'agrégation correspond au fait qu'ils aboutissent à des résultats différents. Nous comparons ces résultats de manière à s'assurer que les variations restent faibles.

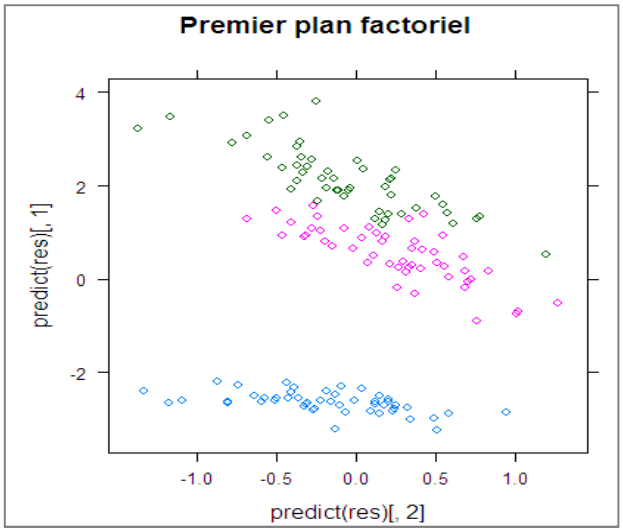
Dans les trois cas, deux groupes se distinguent, Bakers mould, bread yeast, skin fungus sont séparés du reste des donnés.

Les représentations divergent un peu selon la méthode utilisée, même si la répartition en classes finales reste similaire.
Nous ne pouvons donc conclure quant à une représentation idéale pour ce jeu de donnée.

b) Classification hiérarchique des données Iris

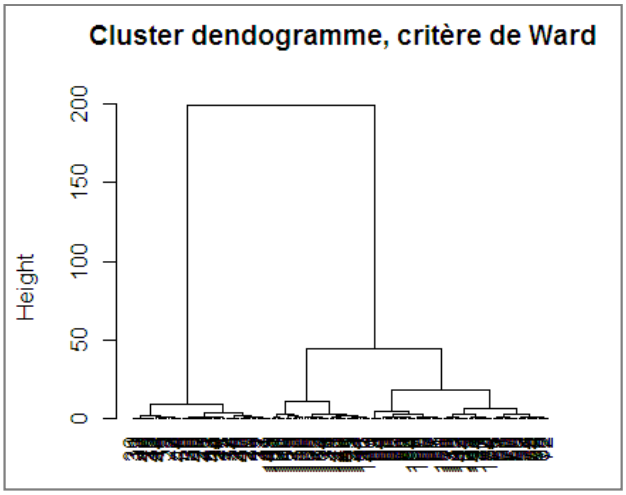
Nous comparons la partition obtenue par classification hiérarchique ascendante, et la représentation plane obtenue par l'analyse en composante principale.
Les analyses factorielles et les techniques de classification sont complémentaires, et on est le plus souvent amené à les utiliser conjointement pour analyser un problème donné. En reliant un module de l'A.C.P à un module de C.A.H, les variables utilisées par le module de C.A.H. seraient les coordonnées des individus sur les premiers axes factoriels.

- **Représentation dans le premier plan factoriel**



La figure représente le premier plan factoriel des donnes Iris.
Nous notons la séparation en **3 classes** suivant les différentes couleurs : Setosa, Versicolor, Virginica.

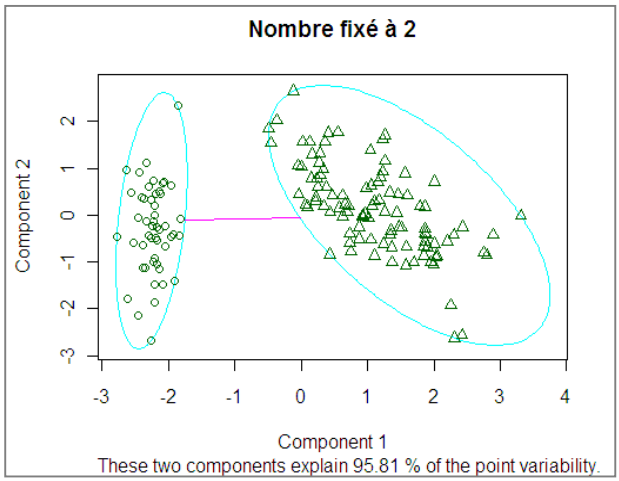
- **Classification hiérarchique ascendante**

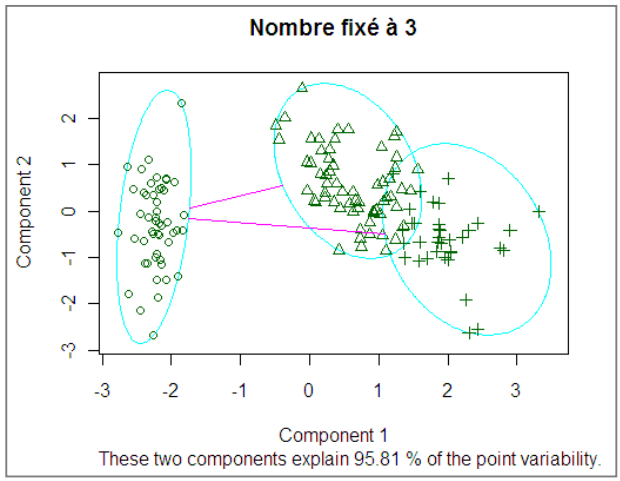


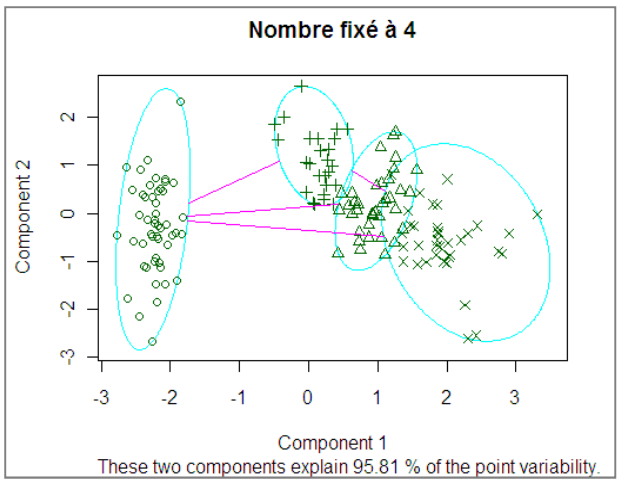
Comme le jeu de données Iris constitue un espace euclidien, nous utilisons le critère d'agrégation de **Ward** (en effet, nous observons que la représentation des autres critères d'agrégation ne permet pas d'obtenir une meilleure classification et une représentation claire).
A la vu de cette représentation, nous pourrions conclure à une classification à **4 classes**.

Ensuite nous visualisons la composition des partitions utilisant la fonction cutree. Cela permet d'imposer un **niveau de coupe** dans les dendogrammes.

La fonction clusplot nous permet ensuite de représenter une partition sur une projection (`cut <- cutree(h,3) ; clusplot(donnees$num,cut))`.



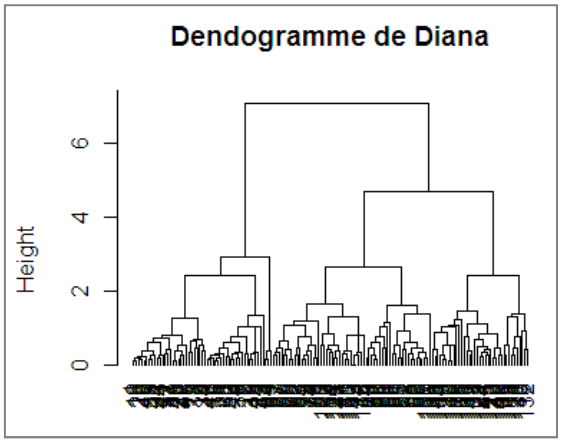




Pour un nombre de partition fixé à 2 : 2 groupes sont formés distinctement.
Pour un nombre de partition fixé à 3 : **3 groupes** sont également bien formés et bien **distincts**.
Cependant, pour un nombre de partition fixé à 4, la distinction entre les groupes et beaucoup moins significatives que leur des deux précédentes représentations.
Le dendogramme permettrait alors une classification en 3 groupes distincts, se qui correspondrait alors aux 3 classes (setosa, versicolor, virginica).

En gardant un nombre de partition fixé à 3, la partition obtenue par classification hiérarchique ascendante, et la représentation plane obtenue par l'analyse en composante principale, convergent vers le même résultat : soit une classification en **3 classes**.

- **Classification hiérarchique descendante**



En comparaison avec la C.A.H, nous pouvons plus facilement conclure une classification débouchant en 3 classes.

Dans ce premier exercice, la classification automatique se rapprochant le plus des résultats obtenus lors des précédents méthodes serait la classification hiérarchique descendante, où nous obtenons une classification en 3 classes distinctes plus facilement.

II. Les 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.

a) Données Iris

- **Partition en K = {2, 3, 4}**

Nous réalisons une classification des données en 2,3 et 4 classes, à l'aide de la fonction kmeans. Nous obtenons les tableaux de contingence suivants:

	K=2		
	setosa	versicolor	virginica
1	0	47	50
2	50	3	0

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

	K=4		
	setosa	versicolor	virginica
1	0	0	27
2	50	0	0
3	0	23	22
4	0	27	1

A partir de la partition en 2 classes, nous obtenons deux classes bien distinctes: une classe contenant toute l'espèce setosa et l'autre contenant versicolor et virginica. Pour une partition en 3 classes, nous obtenons 3 classes avec à peu près la même cardinalité (1/3 chacune) dont la première contenant l'ensemble des éléments setosa et la dernière contenant la majorité des éléments versicolor. Pour k=4, l'espèce setosa est toujours contenue par une classe, ensuite la moitié des éléments de l'espèce versicolor et de virginica sont dans une même classe. Le reste des éléments se retrouvant danss des classes distinctes.

Nous pouvons en conclure que l'espèce setosa est plus éloignée des autres espèces que ne le sont versicolor et virginica l'une de l'autre.

- **Etude de la stabilité du résultat de la partition**

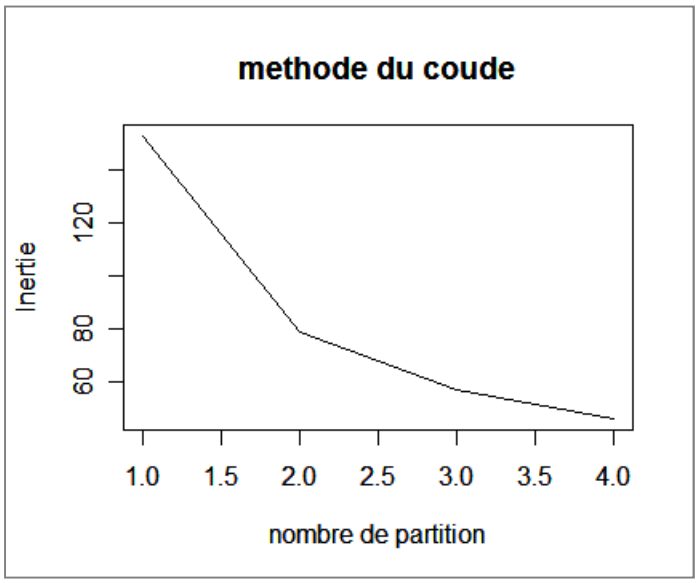
On étudie la stabilité du résultat de partitionnement de kmeans en effectuant 50 fois des classifications en K=3 classes de jeux de donnée.

On remarque que les valeurs obtenues peuvent changer. Les 2 valeurs obtenues pour la partition sont 142,75 et 78,85. La valeur d'inertie intra-classes permettant une plus grande stabilité est la valeur minimum (78,85).

L'explication est la suivante : initialement, la méthode des kmeans prend au hasard un nombre de points du jeu de données égal au nombre de classes désiré. L'algorithme est ensuite appliqué à ces centres initiaux. Ce mécanisme engendre le constat suivant : d'une exécution à l'autre, les points de départ de la méthode ne sont pas les mêmes. La fonction kmeans n'est donc pas stable, le résultat dépend donc des premiers centres mobiles choisis.

- **Détermination du nombre de classes optimal**

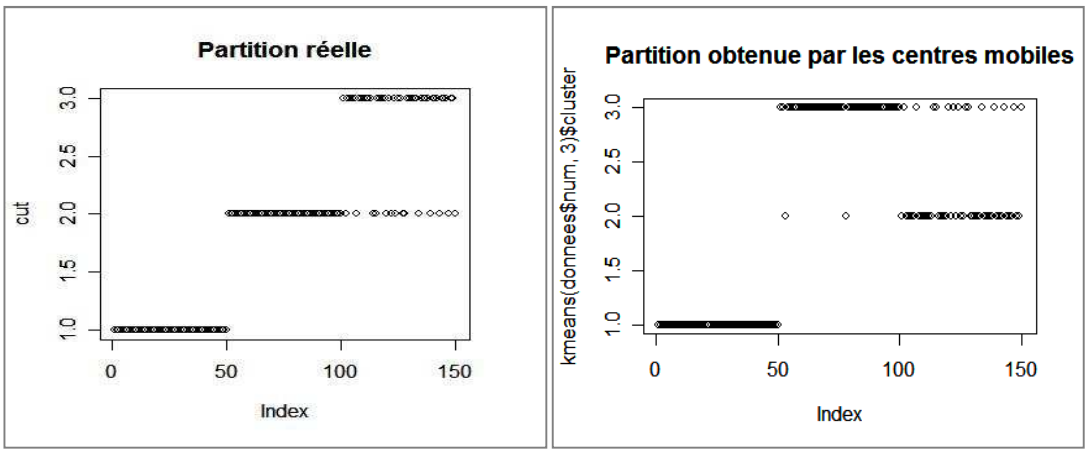
La méthode du coude nous aide à proposer un nombre de classes plausible pour la méthode des kmeans.



On voit clairement ici le ”coude” pour un partitionnement en 3 classes. Dans la mesure où nous voulons l'information sur l'espèce, un partitionnement en 3 classes est largement suffisant puisque au-delà de 3 classes, le partitionnement ne réduit plus de manière significative l'inertie intra-classe.

- **Comparaison entre les 2 méthodes**

On compare la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes:



Les résultats sont comparés avec la réalité. 100% des Setosa sont reconnus par la partition des centres mobiles, 96% des Versicolors sont identifiés, et seulement 72% des Virginia sont dans la bonne classe. Ce qui donne un pourcentage global de réussite de 89,3%. Ceci prouve que cette méthode n'est pas la plus optimale pour étudier ce jeu de données, en comparaison avec la partition réelle.

b) Données Crabs

Nous comparons les résultats de la partition obtenus par les centres mobiles avec la partition réelle des iris en trois groupes. Voici le résultat avec F = femelle et M = mâle et B = bleu et O = orange:

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

Partition obtenue avec les centres mobiles

	F&B	F&O	M&B	M&O
1	50	1	16	0
2	0	0	34	0
3	0	49	0	12
4	0	0	0	38

Partition réelle

Nous remarquons que les résultats sont un peu différents mais que la répartition dans les classes est quasiment la même dans les 2 cas. Seulement quelques spécimens (4 au maximum) différent par rapport à la partition réelle. Pour ces données, la méthode des centres mobiles est aussi efficace qu'une partition réelle.

III. Conclusion

Ainsi, en conclusion de ce TP, nous pouvons dire qu'il n'y a pas de méthode optimale dans tous les cas, et que parfois même une partition réelle s'avère plus efficace qu'une méthode de classification automatique.