Carolina BRUNO
Clara PAGANI

TP02 – SY09

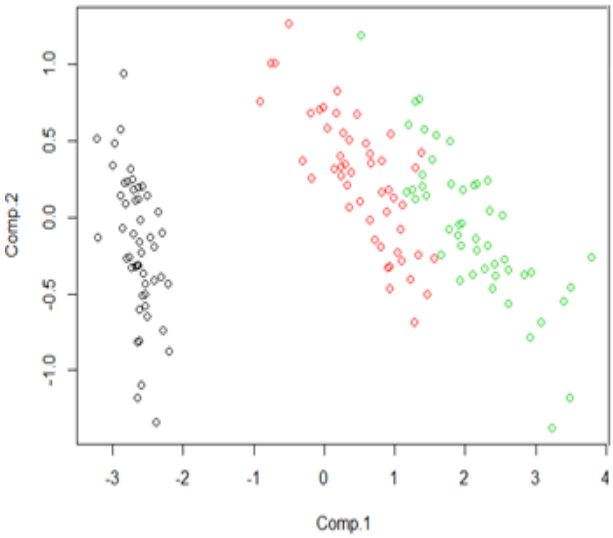
Classification Automatique

Printemps 2013

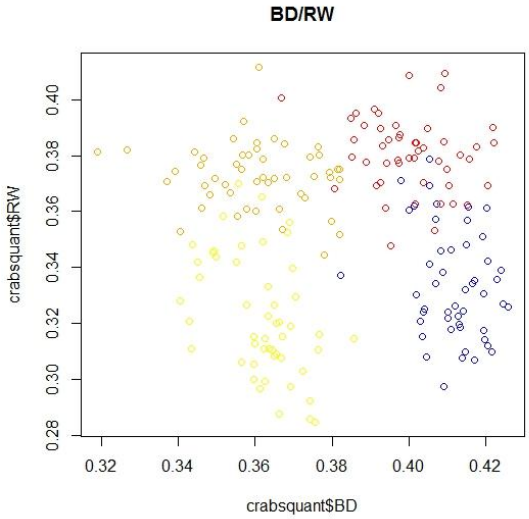
1 Visualisation de données

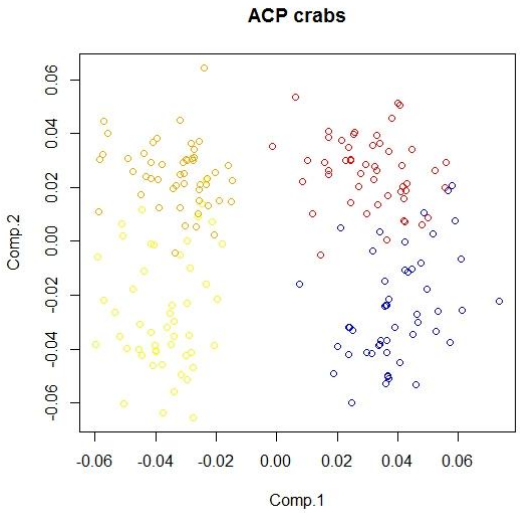
1.1 Visualisation des données Iris

Pour une meilleure compréhension, on va utiliser différentes couleurs pour différentes espèces d'iris : la couleur noir pour Setosa, rouge pour Versicolor et vert pour Virginica.
La représentation de l'ACP dans le premier plan factoriel permet de distinguer les trois espèces de manière satisfaisante. Nous pouvons remarquer que les espèces Versicolor et Virginica sont plus proches que l'espèce Setosa.
Le pourcentage d'inertie expliqué du premier plan factoriel est de 97,7 %. Cette représentation des données est donc satisfaisante.



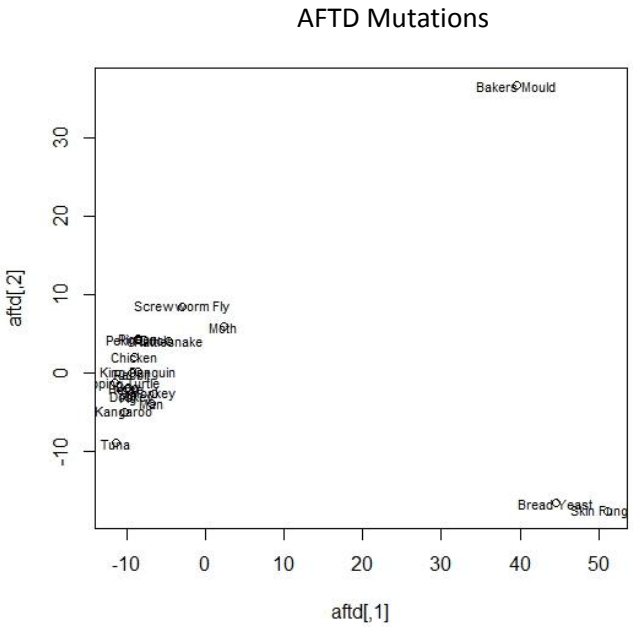
1.2 Visualisation des données Crabs

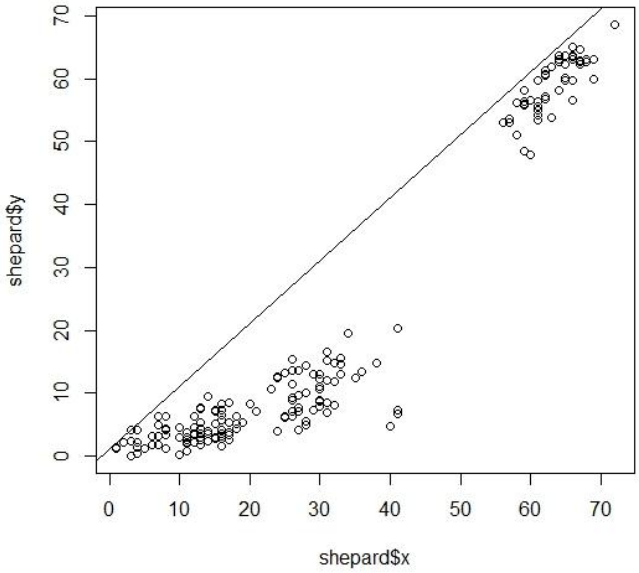




Lorsque l'on compare la représentation de l'ACP des données crabs normalisées dans le premier plan factoriel et la représentation des crabes selon les critères BD et RW, on remarque qu'on distingue les 4 catégories de crabes (rouge : femelle orange /bleu : male orange/ jaune : male blanc/orange : femelle blanche). Ces catégories se distinguent un peu mieux sur le premier plan facoriel de l'ACP. De plus, la représentation de l'ACP permet de déterminer de manière distincte la couleur de chaque crabe : rouge et bleu : crabes oranges, orange et jaune : crabes blancs.

1.3 Visualisation des données Mutations





Les données mutations sont sous la forme d'une matrice de distances. Nous ne pouvons donc pas utiliser les méthodes utilisées dans les jeux de données précédents pour évaluer la répartition des données. Nous allons donc effectuer une AFTD permettant de représenter les différences entre les espèces.

Cette représentation nous permet de distinguer plusieurs groupes d'espèces. Certaines données marquent une grande différence avec les autres et peuvent être considérées comme des données atypiques : Bakers Mould , Bread Yeast et Skin Fing.

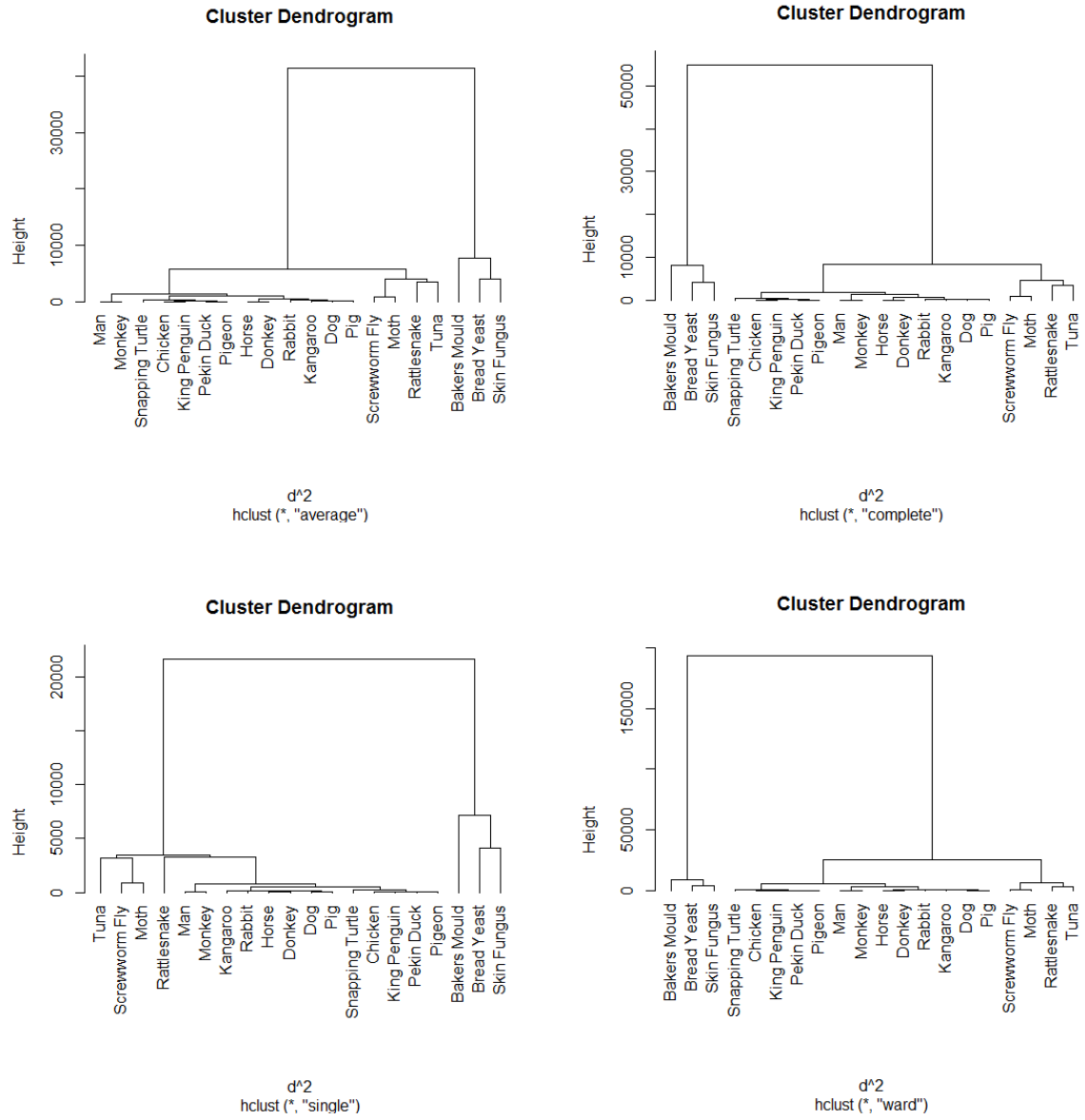
La qualité des données de l'AFTD peut être évaluée à travers un diagramme de Sherpard. La représentation de l'AFTD est d'autant plus exacte que points sont distribués autour de la droite x=y. En observant, la répartition des points autour de la droite x= y sur le graph de Sherpard, nous en déduisons que le rendement obtenu avec l'AFTD est assez bon.

2 Classification hiérarchique

Dans cette partie, nous étudierons les différentes méthodes de classification hiérarchique. La méthode de classification hiérarchique consiste à fusionner les deux classes les plus proches selon un critère, à chaque étape de la construction de la hiérarchie. L'ordre de regroupement des classes permet de définir un indice, on parle de hiérarchie indicée.

2.1 Classification hiérarchique ascendante des données de mutations

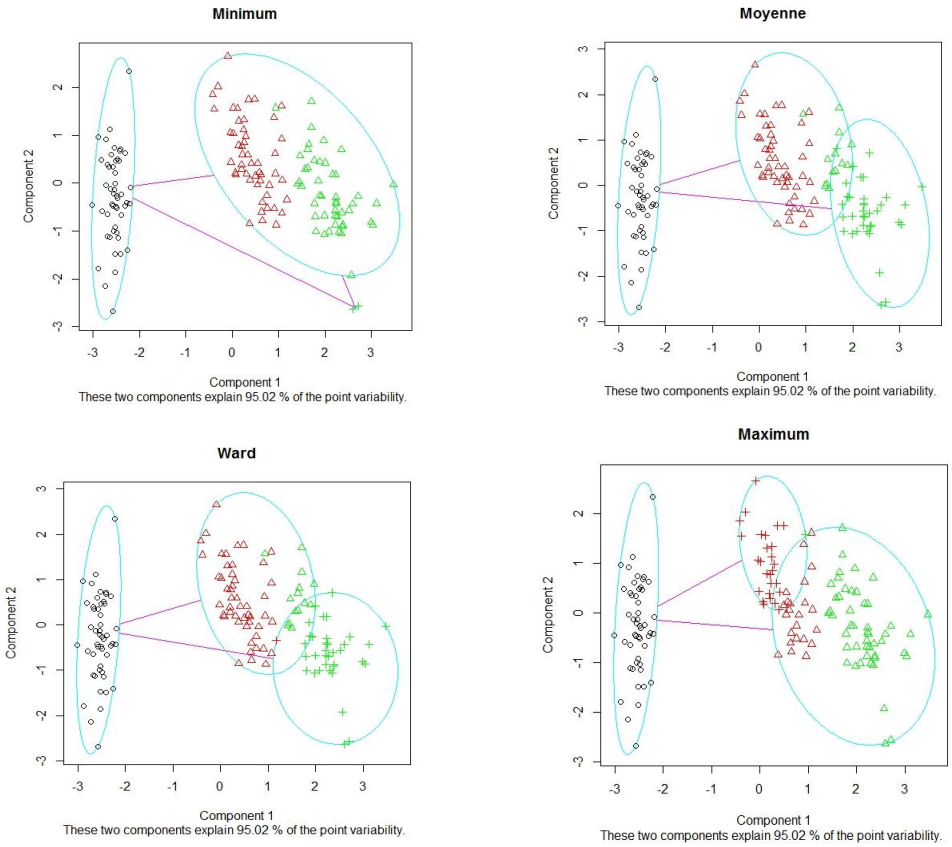
Nous avons réalisé la classification hiérarchique ascendante des données mutations en utilisant 4 critères d'agrégation différents : *ward, moyen, maximum, et minimum*. Les classifications hiérarchiques obtenues sont représentées ci-dessous.



Une classification distincte est obtenue pour chaque critère, mais nous pouvons observer qu'ils sont très similaires et qu'il n'existe pas une grande différence entre chaque classification. Nous pouvons tout de même distinguer une légère différence du critère *minimum* avec les 3 autres critères. Dans toutes les classifications, nous avons bien les mêmes groupes distingués lors de l'AFTD : deux classes : une classe pour Bakers Mould, Bread Yeast ,Skin Fungus et une pour tous les autres.

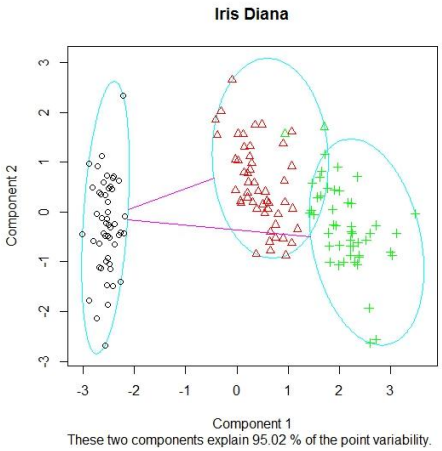
2.2 Classification hiérarchique ascendante des données Iris

De même que précédemment, nous réalisons la classification hiérarchique ascendante des données Iris selon les 4 mêmes critères. Nous choisissons de représenter ces classifications dans le premier plan factoriel :



Premièrement, nous remarquons que nous obtenons des résultats en accord avec l'ACP réalisée dans la première partie. En effet, toutes les méthodes permettent de très bien distinguer l'espèce Sétosa (en noir). Les deux autres espèces Versicolor et Virginica (respectivement en rouge et en vert) apparaissent comme étant plus proches et sont plus difficilement distinguables : les méthodes *maximum* et *minimum* ne permettent pas de les distinguer. Nous notons par ailleurs que leurs taux d'erreur sont logiquement élevés : 32% pour le critère *minimum* et 16% pour le critère *maximum*. Les méthodes *moyenne* et *ward* distinguent bien plus nettement les trois espèces. Nous pouvons en déduire que ces deux derniers critères sont plus aptes à distinguer des classes proches. Le taux d'erreur de ces deux méthodes est d'ailleurs plus faible que celui des deux autres critères : 9.3 % pour le critère *moyenne* et 10.7% pour le critère *ward*.

2.3 Classification hiérarchique descendante des données Iris.

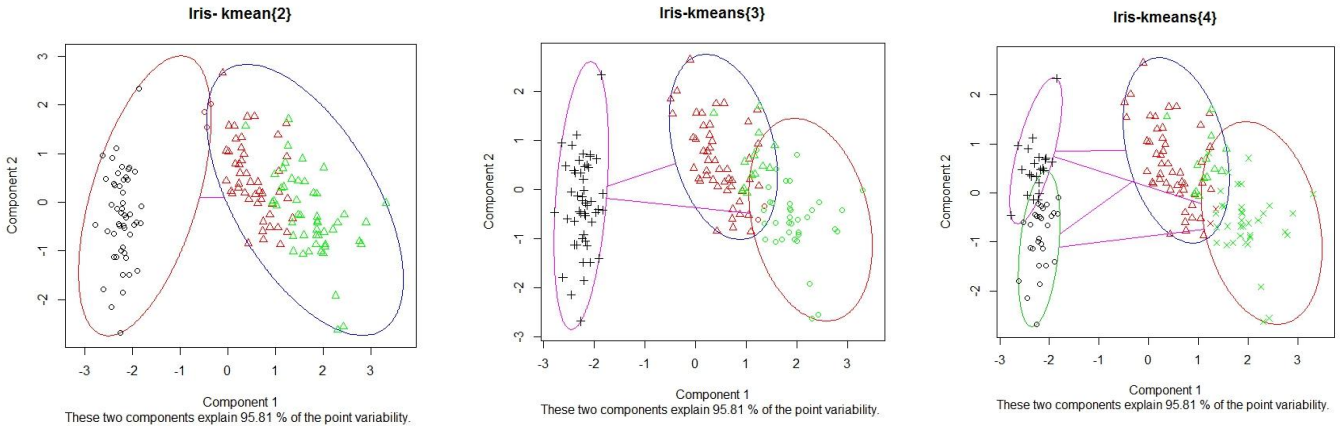


Avec cette méthode, on distingue tout comme avec la classification ascendante, trois classes d'espèces bien définies. Pour la classification hiérarchique descendante donnée d'iris, 3 classes bien définies sont obtenues aussi comme dans l'exercice antérieur.

Cette méthode divise l'ensemble Ω en classes, puis on recommence sur chacune de ces classes et ainsi de suite jusqu'à ce que les classes soient réduites à des singletons. C'est une bonne méthode dans ce cas, nous avons un taux d'erreur de 1.3% seulement. Cependant si l'on avait eu un grand nombre de données, le temps d'exécution aurait été beaucoup plus long que pour les autres méthodes.

3 Centres Mobiles

3.1 Données Iris



En utilisant la fonction kmeans(), nous application l'algorithme de centres mobiles sur les données Iris en choisissant des partitions de 2, 3 et 4 classes.

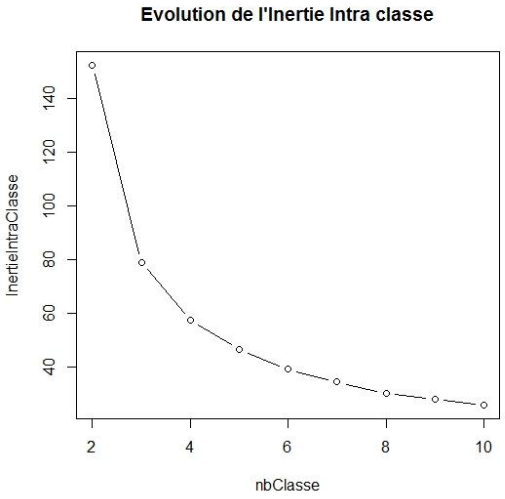
Nous remarquons que dans la partition en 2 classes, la première classe correspond à l'espèce Setosa et la seconde aux espèces Versicolor et Virginica.

Sur la partition en trois classes, nous observons les trois classes correspondants aux trois espèces observées précédemment lors de l'ACP par exemple.

La partition en 4 classes fait apparaitre une nouvelle classe scindant la classe Setosa en deux ce qui n'a pas vraiment de sens.

Si maintenant, nous étudions la stabilité de l'algorithme des centres mobiles en l'appliquant plusieurs fois d'affilé sur le même jeu de données avec la même valeur de partition (ici 3), nous observons que nous n'obtenons pas le même résultat à chaque fois. La classification et l'inertie intra classe varie d'un résultat à l'autre. Ceci est du au caractère itératif de l'algorithme et à sa sélection aléatoire des centres initiaux.

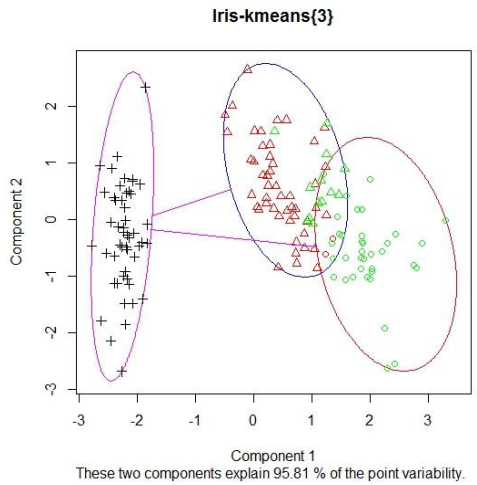
Afin de déterminer le nombre de classe optimale, on se propose d'effectuer 9 séries de 100 classifications, la première série sera 100 classifications à 2 classes, la seconde 100 classifications à 3 classes, et ainsi de suite jusqu'à 10 classes. Pour chacune des ces séries, nous calculons l'inertie intra –classe minimale. Nous obtenons ainsi le graph ci-dessous :

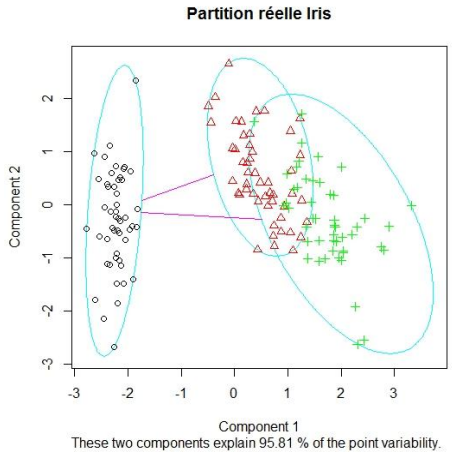


En observant, l'allure de cette courbe, nous observons que l'inertie intra classe diminue fortement de k = 2 à k=3. Nous pouvons donc penser que le nombre de classe optimale est supérieur à 2.

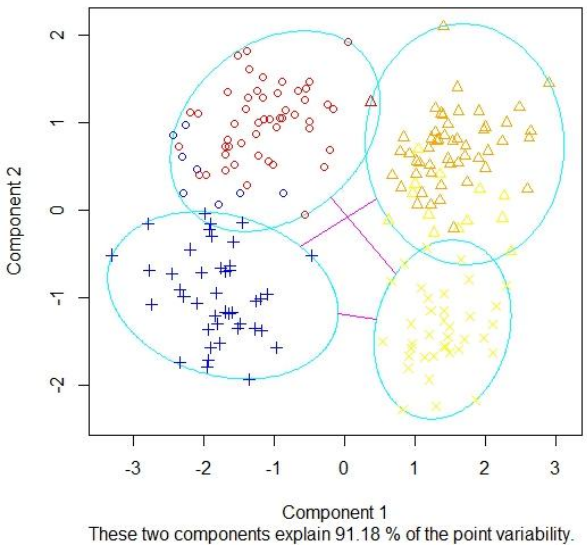
Nous observons que l'inertie intra classe diminue mais de manière moins significative de 4 à 10 classes. Nous choisissons k = 3 classes comme optimale. Nous pourrions prendre un k optimale supérieure à 3 mais cela ne serait pas gagnant sachant la gêne entrainée par un nombre de classes trop importants.

Nous remarquons que la partition réelle des espèces dans chacune des classes est fidèle à la partition effectuée par le kmeans. Nous remarquons cependant que la distinction entre les individus de l'espèce Versicolor et Virginica est toujours moins certaine que celle de l'espèce Setosa. Cependant c'est aussi le cas dans la partition réelle des données.





3.2 Données Crabs



Nous avons appliqué le kmeans à nos données crabs normalisées. La légende est la suivante : rouge : femelle orange /bleu : male orange/ jaune : male blanc/orange : femelle blanche.

Nous remarquons que les espèces se distinguent très bien. Les sexes sont plus difficiles à distinguer. Ceci est en accord avec les observations effectuées dans la partie 1.