Compte rendu TP 2 SY09

ARTCHOUNIN Daniel / VALLOIS Célestin

27 avril 2016

Résumé

Dans le cadre du deuxième sujet des séances de Travaux Pratiques (TP) de l'Unité de Valeur (UV) SY09 enseignée à l'Université de Technologie de Compiègne (UTC), nous avons principalement mené des classifications automatiques sur plusieurs jeux de données.

Dans un premier temps, nous avons réalisé des analyses descriptives, des `ACP` et des `AFTD` sur les jeux de données : `Iris`, `Crabs` et `Mutations`.

Dans un second temps, nous avons mené des classifications hiérarchiques sur les jeux de données : `Mutations` et `Iris`.

Enfin dans un troisième temps, nous avons utilisé l'algorithme des centres mobiles sur les jeux de données suivants : `Iris`, `Crabs` et `Mutations`.

Le dossier `code_source` associé au présent rapport et contenant le code source `R` écrit afin de répondre aux différentes questions présentes dans le sujet s'organise ainsi :

— `ex1.r` : le script `R` associé à la section 1
— `ex2.r` : le script `R` associé à la section 2
— `ex3.r` : le script `R` associé à la section 3

Nous pourrons noter que nous n'utiliserons pas la fonction `clusplot` durant ce rapport de part le fait que nous ne savons pas exactement son fonctionnement en arrière-plan.

1 Visualisation des données

1.1 Données Iris

Le jeu de données `Iris` contient des mesures en centimètres des variables quantitatives `sepal length`, `sepal width`, `petal length` et `petal width` de 150 iris : 50 `Iris setosa`, 50 `versicolor` et 50 `virginica`.

Tout d'abord, nous avons commencé par représenter les individus de l'échantillon selon l'espèce, tour à tour, en fonction de 2 variables quantitatives parmi les 4 à notre disposition (figure 1). Il semble que les individus des différentes espèces sont distinguables via les 4 variables quantitatives.

Ensuite, nous avons effectué une `ACP` sur les 4 variables quantitatives. Les pourcentages d'inertie expliquée par les sous-espaces principaux sont consultables dans la figure 2. Nous constatons que le pourcentage d'inertie expliquée par le sous espace vectoriel E_1 (le premier plan factoriel) est de $97.77 \geq 80$. Ainsi, cela nous a incités à représenter les données dans le premier plan factoriel.

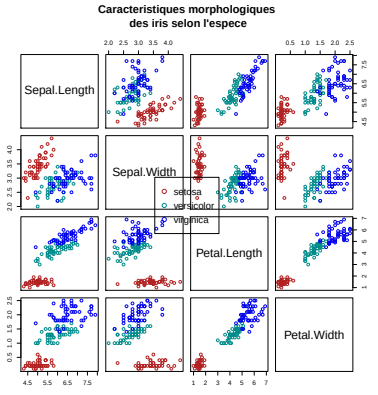

FIGURE 1 – Représentation des individus dans les plans formés de 2 variables quantitatives (Iris)

Nous avons commencé par représenter les individus dans le premier plan factoriel sans tenir compte de l'espèce (à gauche dans la figure 3). On constate que le premier axe factoriel tra-

duit principalement les variables `petal length` et `petal width` tandis que le deuxième axe factoriel semble traduire la variable `sepal width`. Par ailleurs, le jeu de données semble principalement comporter 2 groupes de points.

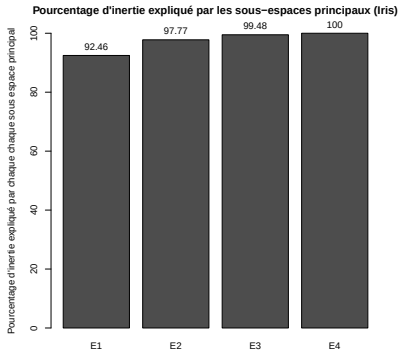

FIGURE 2 – Pourcentage d'inertie expliquée par les sous-espaces principaux (Iris)

Afin de savoir si ces deux groupes de données sont liés aux espèces des individus, nous avons également représenté ces derniers en tenant compte de ce paramètre (à droite dans la figure 3). Ainsi, il s'avère que l'un des deux groupes est constitué des individus de l'espèce `setosa`. Les membres de l'autre groupe ont pour espèce, soit `versicolor`, soit `virginica`.

Si l'on recherche une partition de données avec 2 classes, on peut s'attendre à voir que les individus d'une classe ont pour espèce `setosa` tandis que les individus de l'autre classe sont, soit d'espèce `versicolor`, soit d'espèce `virginica`.

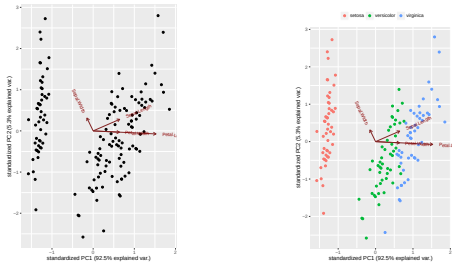

FIGURE 3 – Représentation des individus dans le premier plan factoriel sans tenir compte (à gauche) / en tenant compte (à droite) de l'espèce (Iris)

1.2 Données Crabs

Le jeu de données `Crabs` contient des mesures de 4 variables quantitatives `FL2`, `RW2`, `CL2` et `BD2` de 200 crabes : 50 mâles bleu `M/B`, 50 mâles orange 50 `M/O`, 50 femelles bleu `F/B` et 50 femelles orange `F/O`.

Tout d'abord, nous avons commencé par représenter les individus de l'échantillon selon leurs couleurs et leurs sexes, tour à tour, en fonction de 2 variables quantitatives parmi les 4 à notre disposition (figure 4). Il semble que les individus de différentes couleurs ou de différents sexes sont distinguables via les 4 variables quantitatives.

Ensuite, nous avons effectué une `ACP` sur les 4 variables quantitatives. Les pourcentages d'inertie expliquée par les sous-espaces principaux sont consultables dans la figure 5. Nous constatons que l'inertie expliquée par le sous espace vectoriel E_1 (le premier plan factoriel) est de $92.35 \geq 80$. Ainsi, cela nous a incités à représenter les données dans le premier plan factoriel.

Nous avons commencé par représenter les individus dans le premier plan factoriel sans tenir compte de la couleur et du sexe (à gauche dans la figure 6). On observe que le premier axe factoriel traduit principalement les variables `FL2`, `BD2` et `CL2` tandis que le deuxième axe factoriel semble traduire la variable `RW2`. Graphiquement, le jeu de données semble principalement comporter 2 groupes de points.

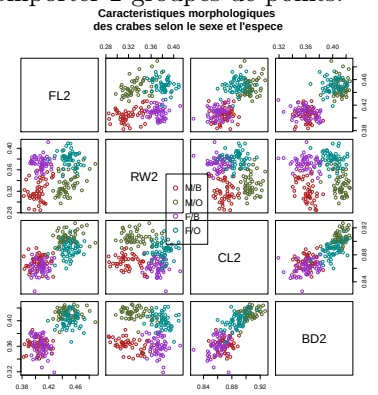

FIGURE 4 – Représentation des individus dans les plans formés de 2 variables quantitatives (Crabs)

Afin de savoir si ces deux groupes de points sont liés aux couleurs ou aux sexes des indivi-

dus, nous avons également représenté ces derniers en tenant compte de ces paramètres (à droite dans la figure 6). Il s'avère que l'un des deux groupes est constitué des individus `F/O` et `M/O`. Les membres de l'autre groupe sont, soit des `F/B`, soit des `M/B`. Ainsi, un groupe contient des individus `O` tandis que l'autre contient des individus `B`.

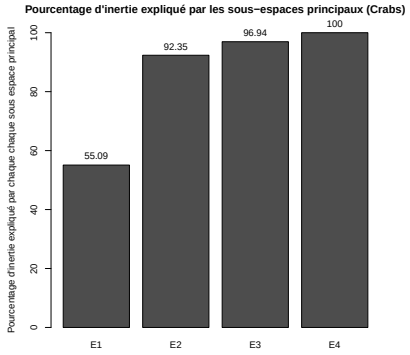

FIGURE 5 – Pourcentage d'inertie expliquée par les sous-espaces principaux (Crabs)

Si l'on recherche une partition de données avec 2 classes, on peut s'attendre à voir que les individus d'une classe sont principalement des `O` tandis que les individus de l'autre classe sont majoritairement des `B`.

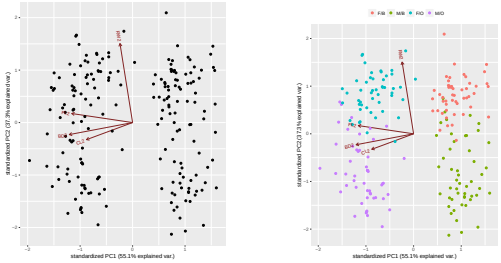

FIGURE 6 – Représentation des individus dans le premier plan factoriel sans tenir (à gauche) / en tenant (à droite) compte de l'espèce et du sexe (Crabs)

1.3 Données Mutations

Le jeu de données `Mutations` est un tableau de dissimilarités Δ sur $n = 20$ espèces. Les dissimilarités entre deux individus de l'échantillon ont été calculées en se basant sur les différentes positions des acides aminés de la protéine `Cytochrome C`.

Pour commencer, nous avons calculé une représentation euclidienne des données en $d = 2$ variables par Analyse Factorielle d'un Tableau de Dissimilarités (`AFTD`). Les résultats obtenus sont consultables dans la figure 7.

Dans ce graphique (figure 7), on peut remarquer que l'homme est proche du singe. De même, le pigeon est proche du canard. De plus, le cheval est proche du chien. Ces résultats semblent plutôt cohérents.

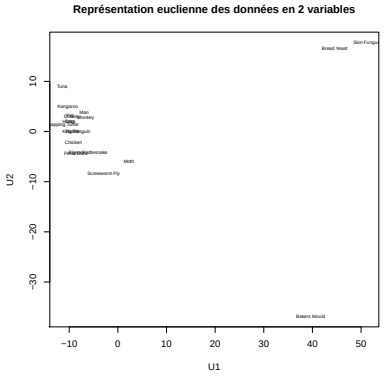

FIGURE 7 – Représentation euclidienne des données en $d = 2$ par AFTD sans ajout de constante préalable

Toutefois, on constate que les valeurs propres ordonnées de la matrice $\frac{1}{n}W = -\frac{1}{2n}Q_n\Delta^2 Q_n$, où $Q_n = I_n - \frac{1}{n}U_n$, ne sont pas toutes positives ou nulles. Effectivement, la dernière valeur propre $\lambda_{20} = -72.75$ représente en valeur absolue plus de $\frac{1}{10}$ de la cinquième valeur propre ($\lambda_5 = 675.04$). Nous jugeons que cela n'est pas négligeable et peut, par conséquent, avoir un impact sur la qualité de la représentation. Par conséquent, nous avons décidé de ne plus appliquer directement l'`AFTD` sur le jeu de données `Mutations`.

Ainsi, nous avons choisi d'ajouter une constante que l'on nommera c^* à la dissimilarité initiale Δ en vue de la transformer en une distance, avant d'appliquer l'`AFTD`.

Les pourcentages d'inertie expliquée par les sous-espaces principaux sont consultables dans la figure 8. Nous constatons que l'inertie expliquée par le sous espace vectoriel E_1 (le premier plan factoriel) est de $91.1 \geq 80$. Ainsi, la représentation euclidienne des données dans un espace de dimension de $d = 2$ semble raisonnable.

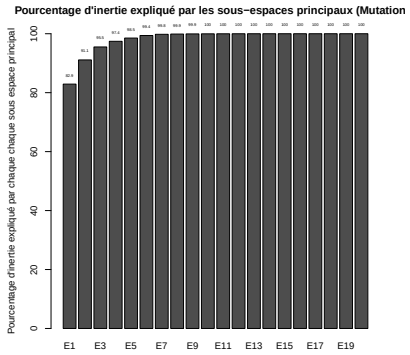

FIGURE 8 – Pourcentage d'inertie expliquée par les sous-espaces principaux (Mutations)

Les résultats obtenus sont consultables dans la figure 9. A l'instar de la figure 7, le graphique semble cohérent. Effectivement, il est très semblable à celui obtenu dans la figure 7.

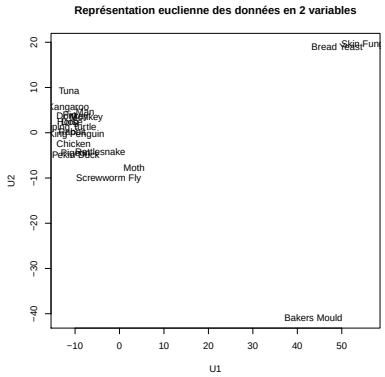

FIGURE 9 – Représentation euclidienne des données en $d = 2$ par AFTD avec ajout de constante préalable

Par ailleurs, afin de prendre conscience de la qualité de la représentation, nous avons tracé le diagramme de `Shepard` à partir de la représentation euclidienne des données dans un espace à $d = 2$ dimensions obtenue via l'`AFTD`. L'utilisation de ce diagramme est cohérente puisque l'`AFTD` vise à minimiser le critère $\sum_{i\in\Omega}\sum_{i'\in\Omega}(\delta^2_{ii'} - d^2_{ii'})$, où $d_{ii'}$ représente la distance entre les individus i et i' calculée à partir de la représentation euclidienne X obtenue via l'`AFTD` et où $\delta_{ii'}$ représente la dissimilarité initiale entre les individus i et i'. Le diagramme de `Shepard` permet de représenter la dissimilarité au sein de chaque couple de deux individus en fonction de leur distance. Ainsi, plus la représentation est bonne, plus le nuage de points est proche de la bissectrice.

Le diagramme de `Shepard` obtenu est consultable à gauche dans la figure 10. Il est évidemment difficile d'interpréter directement ce graphique. Afin de pouvoir le faire de manière pertinente, nous avons décidé de réaliser un diagramme de `Shepard` (à droite dans la figure 10) à partir d'une représentation euclidienne dans un espace de $d = 5$ dimensions obtenue via l'`AFTD` avec ajout d'une constante c^* au préalable.

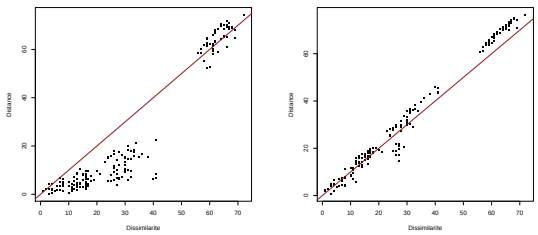

FIGURE 10 – Diagramme de Shepard associé à la représentation euclidienne des données en d=2 (à gauche) / en d=5 (à droite) (Crabs)

Évidemment, nous constatons que le nuage de points à gauche dans la figure 10 est plus proche de la bissectrice que celui à droite dans la figure 10 : cela est logique puisque on a perdu moins d'informations à droite dans la figure 10 qu'à gauche dans la figure 10. Toutefois, la représentation dans un espace à dimensions 5 est difficilement graphiquement exploitable. Par ailleurs, on peut également ajouter que la baisse de la qualité de la représentation est "raisonnable" compte tenu du gain apporté par la représentation dans un plan. Ainsi, de ce fait, nous continuerons à utiliser la représentation dans un espace à $d = 2$ dimensions dans ce rapport.

2 Classification hiérarchique

2.1 Données Mutations

Dans cette sous-section 2.1, nous allons étudier le jeu de données `Mutations` à travers une Classification Ascendante Hiérarchique (CAH). Pour ce faire, nous allons utiliser le tableau de dissimilarités Δ lui étant associé.

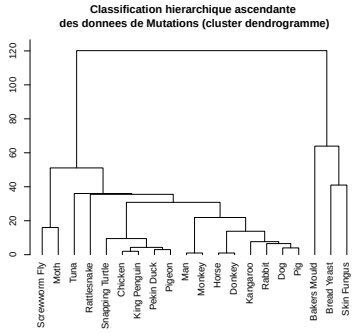

FIGURE 11 – Dendrogramme de la CAH avec la méthode de Ward (données Mutations)

La méthode consiste à commencer par considérer une partition constitués de classes qui sont des singletons. A chaque étape, on fusionne des classes suivant un critère d'agrégation jusqu'à l'obtention d'une seule classe constitué de l'ensemble des individus. Les résultats seront représentés graphiquement dans les figures 11, 12, 13, 14, 15 et 16 via des dendrogrammes symbolisant les hiérarchies indicées obtenues avec les différentes méthodes utilisées.

Suite aux résultats trouvés, on peut observer que les CAH avec les critères d'agrégations de `McQuitty` (figure 13), `Complete` (figure 12) et `Ward` (figure 11) permettent d'obtenir des résultats "très" similaires. Celles avec les critères `Single` (figure 14) et `Median` (figure 15) donnent des résultats "légèrement" différents (notamment pour les classes `Bakers Mould` et `Tuna`).

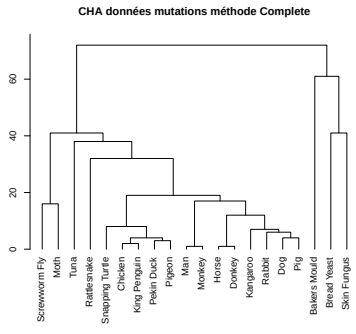

FIGURE 12 – Dendrogramme avec le critère d'agrégation Complete (Mutations)

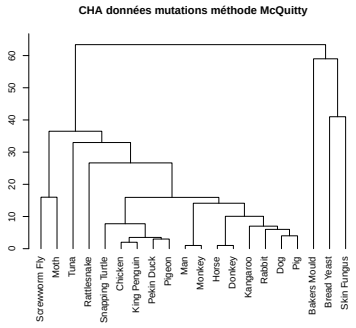

FIGURE 13 – Dendrogramme avec le critère d'agrégation McQuitty (Mutations)

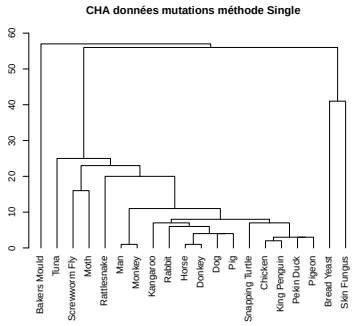

FIGURE 14 – Dendrogramme avec le critère d'agrégation Single (Mutations)

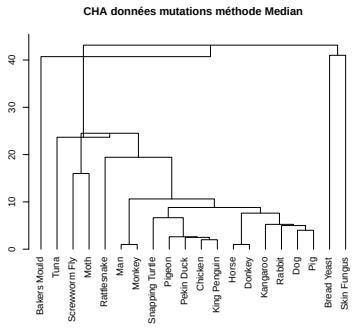

FIGURE 15 – Dendrogramme avec le critère d'agrégation Median (Mutations)

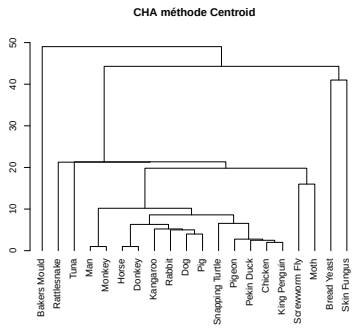

FIGURE 16 – Dendrogramme avec le critère d'agrégation Centroid (Mutations)

On notera la présence d'inversions dans le dendrogramme associé au critère `Median` (figure 15). Finalement, la méthode `Centroid` (figure 16) diffère des autres dendrogrammes sur de "nombreux" points.

En comparant les différents dendrogrammes avec les résultats obtenus précédemment (sous-section 1.3), le critère de `Ward` semble être le plus adapté à nos données. On retrouve d'ailleurs nos 3 espèces très éloignées des autres : `Bakers Mould`, `Bread Yeast` et `Skin Fungus`.

On pourra noter qu'à chaque étape de la méthode de `Ward`, on cherche à fusionner deux classes de sorte à augmenter le moins possible le critère d'inertie intra-classe. Ainsi, cette propriété de cette méthode nous incite à continuer de l'utiliser dans la suite de notre étude.

2.2 Données Iris

On effectue la CAH sur les données `Iris` avec le critère d'agrégation de `Ward` suite aux précédentes remarques.

A l'instar de la sous-section 1.1, sur le dendrogramme obtenu (figure 17), 3 grandes classes correspondant aux 3 espèces du jeu de données `Iris` se distinguent : `Setosa`, `Versicolor` et `Virginica`.

Contrairement à la représentation dans le premier plan factoriel (figure 3 dans la sous-section 1.1), les espèces `Versicolor` et `Virginica` sont "facilement" distinguables. De plus, on peut observer que les individus des deux espèces (dans les rectangles vert et bleu de la figure 17) sont issus d'un même sous-arbre (branche droite de la racine) : cela semble confirmer le fait que les individus des deux classes précédemment citées sont plus "proches" les uns des autres qu'avec ceux de la troisième classe au sens du critère d'inertie intra-classe.

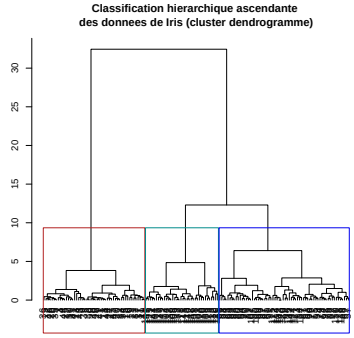

FIGURE 17 – Dendrogramme du CAH avec la méthode de Ward (données Iris)

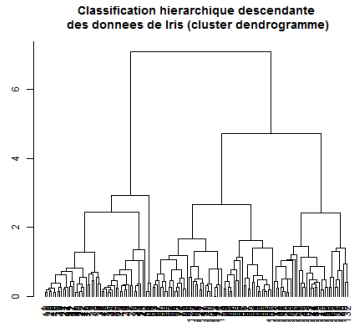

FIGURE 18 – Dendrogramme du CDH avec la méthode de Ward (données Iris)

Enfin, on réalise une Classification Descendante Hiérarchique (CDH) sur le jeu de données `Iris`. A titre de rappel, à contrario de la CAH, au départ, tous les individus sont dans une seule et même classe. Puis, on divise cette classe en 2 classes. Ensuite, on réitère le processus sur chacune des classes jusqu'à ce que chacune d'entre elles contiennent un et un seul individu. Afin d'y parvenir, nous avons utilisé la fonction `Diana` de la bibliothèque `cluster` (figure 18).

On peut remarquer que le résultat (figure 18) est très proche de celui obtenu avec la

CAH (figure 17). Il n'y a pas de différences "flagrantes" et nous pouvons à nouveau observer une partition de 2 classes au sein du dendrogramme. Cette partition contient une classe constituée des individus des deux espèces difficilement différentiables (branche de droite de la racine dans la figure 18) et d'une classe constituée des individus de la troisième espèce (branche de gauche de la racine dans la figure 18).

3 Méthode des centres mobiles

3.1 Données Iris

Tout d'abord, nous avons commencé par trouver des partitions en $K \in \{2; 3; 4\}$ classes via l'algorithme des centres mobiles. Les résultats sont respectivement consultables à gauche dans la figure 19, à droite dans la figure 19 et à gauche dans la figure 20.

La partition P_{11} en $K = 2$ classes (à gauche dans la figure 19) est proche de celle prédite dans la sous-section 1.1 : les individus d'une classe ont pour espèce `setosa` tandis que les individus de l'autre classe sont majoritairement, soit d'espèce `versicolor`, soit d'espèce `virginica`.

On constate que la partition P_{12} en $K = 3$ classes (à gauche dans la figure 20) obtenue via la méthode des centres mobiles est proche de la partition réelle (figure 1).

La partition P_{13} en $K = 4$ classes (à droite dans la figure 19) s'interprète ainsi : les individus d'une classe ont majoritairement pour espèce `setosa`, les individus d'une deuxième classe sont majoritairement d'espèce `versicolor`, les individus d'une troisième classe ont majoritairement pour espèce `virginica` et la quatrième classe est majoritairement constituée d'individus `versicolor` et `virginica`.

Afin d'étudier la stabilité du résultat de la partition, nous avons effectué plusieurs classifications des données en $K = 3$ classes.

A certaines reprises, nous obtenions la partition P_{14} consultable à droite dans la figure 20. Cette partition s'interprète ainsi : une classe est majoritairement constituée d'individus `versicolor` et `virginica` tandis que les deux autres se sont majoritairement formés à partir des individus `setosa` restants. La somme des inerties des classes par rapport à leur centre de la partition P_{14} est proche de 0.95 tandis que la même somme de la partition P_{12} est proche de 0.53. Ainsi, comme prévu, la partition P_{12} est meilleure que la partition P_{14} au sens du critère mentionné ci-dessus.

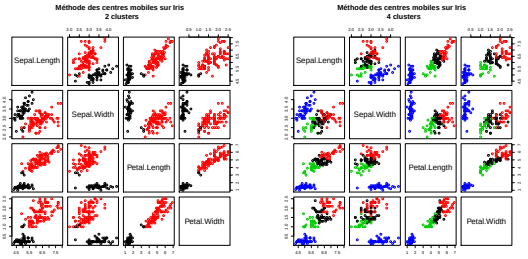

FIGURE 19 – Partition en 2 classes (à gauche) / en 4 classes (à droite) (Iris)

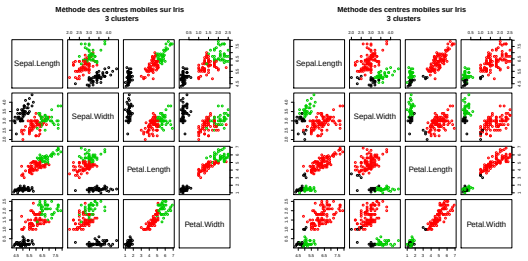

FIGURE 20 – Partition en 3 classes : première configuration (à gauche) / deuxième configuration (à droite) (Iris)

Cette différence de résultats est facilement explicable : effectivement, l'algorithme des centres mobiles propose une partition visant à minimiser le critère mentionné ci-dessus. Toutefois, suivant les points de départ choisis, les résultats sont différents : l'algorithme converge donc vers un optimum local qui peut-être global. Ainsi, à une reprise, les points de départ choisis ont fait que l'algorithme a convergé vers la partition P_{12}. A une autre reprise, d'autres points de départ ont fait que l'algorithme a convergé vers la partition P_{14}.

Ensuite, on a cherché à déterminer le nombre de classes K optimal. Pour ce faire, on a effectué $N = 100$ classifications en prenant $K = 2$ classes, puis, à nouveau $N = 100$ classifications en $K = 3$ classes, $K = 4$ classes, $\cdots$,

jusqu'à $K = 10$ classes. Ainsi, on a constitué neuf échantillons iid que l'on notera $I_2, \cdots, I_{10}$ contenant chacun $N = 100$ sommes d'inerties intra-classe. Pour chaque valeur de $K, K \in \{2; \cdots; 10\}$, on a calculé la somme minimale d'inertie intra-classe $\hat{I}_K = \min_{i=1,\cdots,100}(I_{Ki})$. Puis, on a représenté la variation de somme minimale d'inerties $\hat{I}_K, K \in \{2; \cdots; 10\}$ en fonction du nombre de classes K dans la figure 21. En utilisant la méthode des coudes, nous proposons 2 comme nombre de classes.

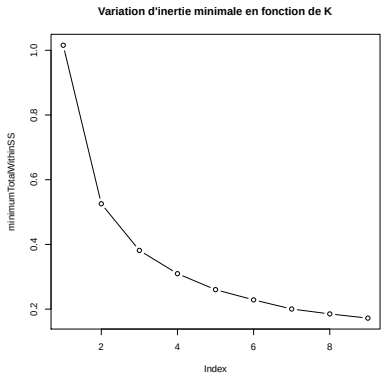

FIGURE 21 – Variation de la somme minimale d'inerties intra-classe en fonction du nombre de classes (Iris)

Le nombre de classes $K = 2$ proposé avec la méthode des coudes est celui prédit graphiquement dans la sous-section 1.1. Par ailleurs, si on calcule la somme des inerties intra-classe de la partition réelle, on constate qu'elle est proche de 0.60. Évidemment, l'inertie de la partition à deux classes obtenue à l'aide de l'algorithme des centres mobiles est plus élevée, proche de 1.02. Cependant, on constate que la même inertie pour la partition à trois éléments obtenue avec la même méthode est proche de 0.53 : elle est inférieure à celle de la partition réelle. En cherchant à minimiser le critère, l'algorithme s'est permis de déplacer certains individus vers des classes différentes de leurs classes d'origine.

3.2 Données Crabs

Afin d'étudier la stabilité de la partition à 2 classes, nous avons effectué plusieurs classifications des données en $K = 2$ classes. A certaines reprises, nous obtenions la partition P_{21} consultable à gauche dans la figure 22. Cette partition s'interprète ainsi : une classe est majoritairement constituée d'individus `O` tandis que l'autre est majoritairement formée à partir des individus `B` restants. D'autres fois, nous obtenions la partition P_{22} consultable à droite dans la figure 22. Cette partition s'interprète ainsi : une classe est majoritairement constituée d'individus `F` tandis que l'autre est majoritairement formée à partir des individus `M` restants.

La somme des inerties intra-classe de la partition P_{21} est proche de $1.29 * 10^{-3}$ tandis que la même somme de la partition P_{22} est proche de $1.78 * 10^{-3}$. Ainsi, comme prévu, la partition P_{21} est meilleure que la partition P_{22} au sens du critère mentionné ci-dessus. On notera également que la partition P_{21} en 2 classes (à gauche dans la figure 22) est proche de celle prédite dans la sous-section 1.2.

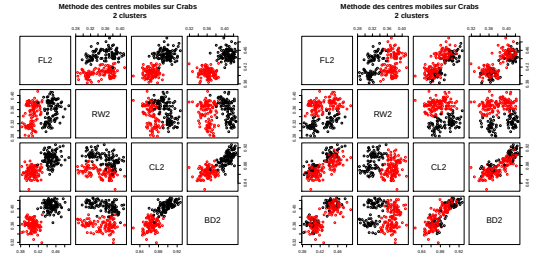

FIGURE 22 – Partition en 2 classes : première configuration (à gauche) / deuxième configuration (à droite) (Crabs)

Ensuite, on a également tenté d'effectuer une classification P_{23} en $K = 4$ classes des données `Crabs`. Le résultat obtenu est consultable dans la figure 23.

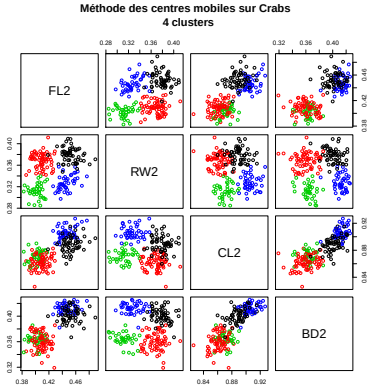

FIGURE 23 – Partition en 4 classes (Crabs)

On constate que la partition P_{23} en 4 classes (figure 23) obtenues via la méthode des centres

mobiles est proche de la partition réelle (figure 4). Effectivement, la partition s'organise ainsi : chacune des 4 classes est majoritairement respectivement constituée de `F/O`, `M/O`, `F/B` et de `F/B`.

Par ailleurs, si on calcule la somme des inerties intra-classe de la partition réelle, on constate qu'elle est proche de $6.24 * 10^{-4}$. Cependant, on constate que la même inertie pour la partition en 4 classes obtenue avec l'algorithme des centres mobiles est proche de $5.30 * 10^{-4}$: elle est donc inférieure à celle de la partition réelle. En cherchant à minimiser le critère, l'algorithme s'est à nouveau autoriser à déplacer certains individus vers des classes différentes de leur classe d'origine.

3.3 Données Mutations

Tout d'abord, nous avons tenté de calculer la représentation des données `Mutations` dans un espace de dimension $d = 5$.

Ensuite, sur cette représentation, on a effectué 10 classifications en $K = 3$ classes via l'algorithme des centres mobiles. Sur ces 10 classifications, nous avons obtenu 4 partitions différentes ayant les sommes d'inertie intra-classe suivantes : 245.90, 181.09, 248.75 et 261.86. Le résultat obtenu pour la partition d'inertie 181.09 dans le premier plan factoriel de l'`AFTD` est consultable dans la figure 24.

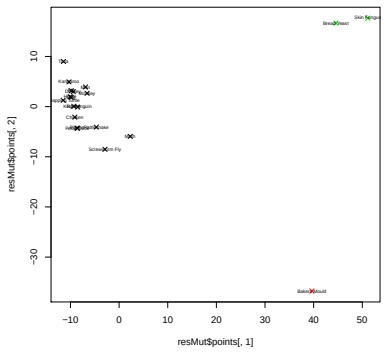

FIGURE 24 – Partition en 3 classes (Mutations)

La partition obtenue (figure 24) dans le premier plan factoriel de l'`AFTD` s'avère être particulièrement intuitive.

Au premier abord, avec ses 4 partitions différentes sur 10 classifications, la classification en $K = 3$ classes semble peu stable compte tenu de la "dissimilarité importante" entre deux individus appartenant à deux classes différentes parmi les trois classes obtenues (figure 24). Toutefois, ce phénomène peut probablement se justifier par le fait que la représentation dans le premier plan factoriel peut masquer ou accentuer certaines dissimilarités entre deux individus (visibles dans d'autres dimensions). Par ailleurs, il est également possible que le fait que la matrice de dissimilarités, avec laquelle on travaille, ne soit pas une matrice de distances, impacte négativement la qualité de la représentation des données obtenues par l'`AFTD` provoquant ainsi une baisse de la stabilité de l'algorithme.

4 Conclusion

En conclusion, au cours des trois séances de travaux dirigés et de la rédaction du présent rapport, nous avons pu à nouveau prendre conscience de la puissance et des multiples possibilités qui s'offrent à nous en terme de traitement statistique de données avec `R`. Ainsi, nous avons notamment appris à effectuer une Analyse Factorielle d'un Tableau de Distances (AFTD) en `R`. Nous avons également eu l'occasion de réaliser des classifications hiérarchiques sur différents jeux de données. Enfin, cela nous a offert la possibilité de tester les performances de la méthode des centres mobiles.