École Centrale Lyon

UE INF

Analyse de données et reconnnaissance de formes

Rapport

ANALYSE DES DONNÉES SUR LA POPULATION

Élèves :
Hugo PUYBAREAU

Enseignant :
Emmanuelle Delandrea

5 octobre 2023

Table des matières

1 Introduction

L'objectif de ce devoir est de réaliser une analyse de données complète sur le fichier *population.txt*. Les données consistent, pour chaque catégorie de population, en un certain nombre d'heures passées à la réalisation d'une activité. Les lignes du tableau représentent donc les catégories de population, et les colonnes représentent le nombre d'heures passées à la réalisation des différentes activités.

Les consignes sont de combiner une analyse factorielle avec une méthode de classification des données. Les quatres combinaisons suivantes doivent être réalisées : ACP + CAH, ACP + centres mobiles, ACM + CAH, ACM + centres mobiles.

2 Analyse en composantes principales

2.1 ACP

La première chose à noter lorsque l'on réalise l'ACP est qu'il n'y pas de variables illustratives mais uniquement des variables actives.

Nous regardons tout d'abord le diagramme d'inertie suivant :

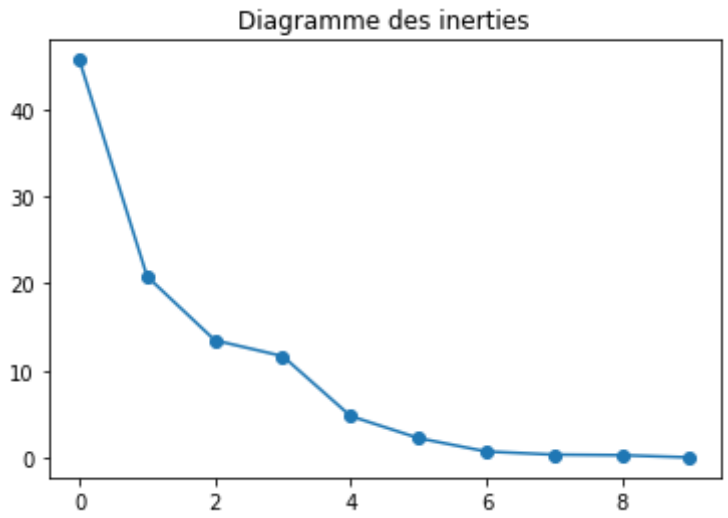

FIGURE 1 – Diagramme des inerties pour l'ACP

On comprend que, contrairement au jeu de données sur les villes, les contributions des axes factorielles 3 et 4 sont importantes. L'observation sur un seul plan factoriel risque de donner des résultats dont il sera difficile de tirer des conclusions intéressantes. Rappelons que l'on étudie un jeu de données où les individus sont tous très différents. Il y en a 28 pour 10 occupations, on s'attend donc à une grande diversité des profils d'activité. Il est normal de ne pas avoir plus de 50% de l'inertie sur un seul axe.

C'est d'ailleurs pour cela que l'on nous fait réaliser une CAH et un Kmeans sur les composantes factorielles. Ces opérations ont pour but de "filtrer le bruit des facteurs non principaux" pour pouvoir obtenir une analyse des données dont il est possible de tirer des conclusions.

En regardant la projection des variables, je suppose que le premier axe est celui qui décrit le sexe de la personne considérée (1 pour Homme, -1 pour Femme). On remarque,

en effet, que les loisirs, télévision et toilette en sont dé-corrélés alors que le travail, les transports, les tâches ménagères, le fait de s'occuper des enfants y sont corrélés. C'est également évident lorsque l'on regarde la projection des individus. Je ne trouve pas d'interprétation évidentes aux autres axes pour l'instant.

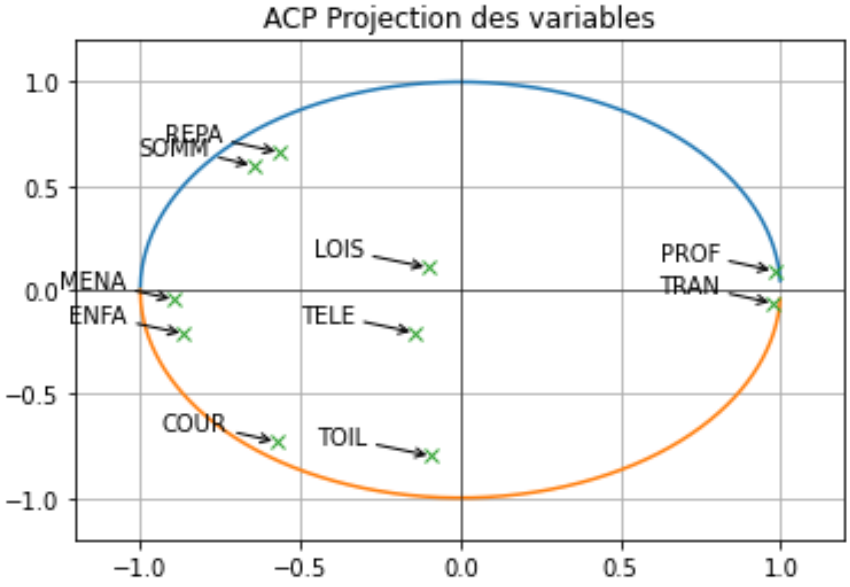

FIGURE 2 – Projection des variables pour l'ACP seule

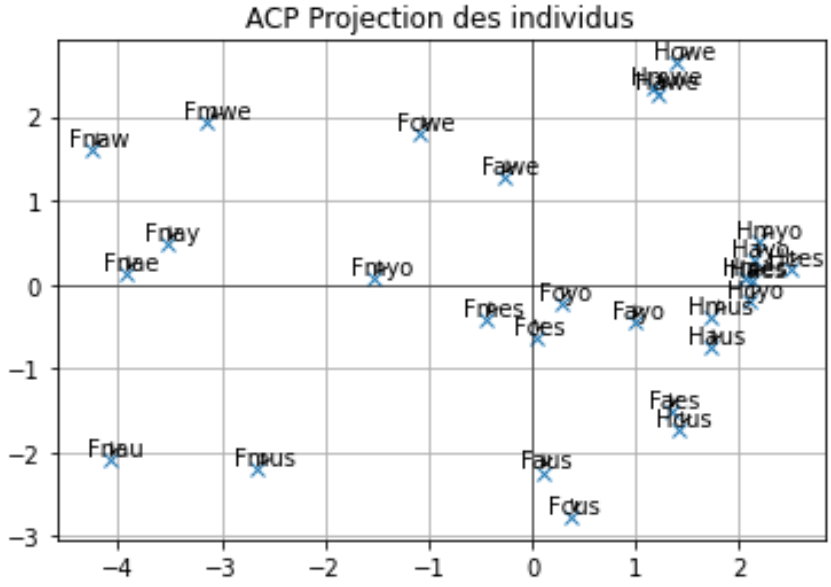

FIGURE 3 – Projection des individus pour l'ACP seule

2.2 ACP et CAH

On vient récupérer les valeurs pour les 2 premiers facteurs afin d'avoir une analyse filtrée des autres facteurs. On n'utilise pas *whiten* car c'est justement la valeur des facteurs non normalisée qui nous intéresse.

J'obtiens le dendogramme suivant :

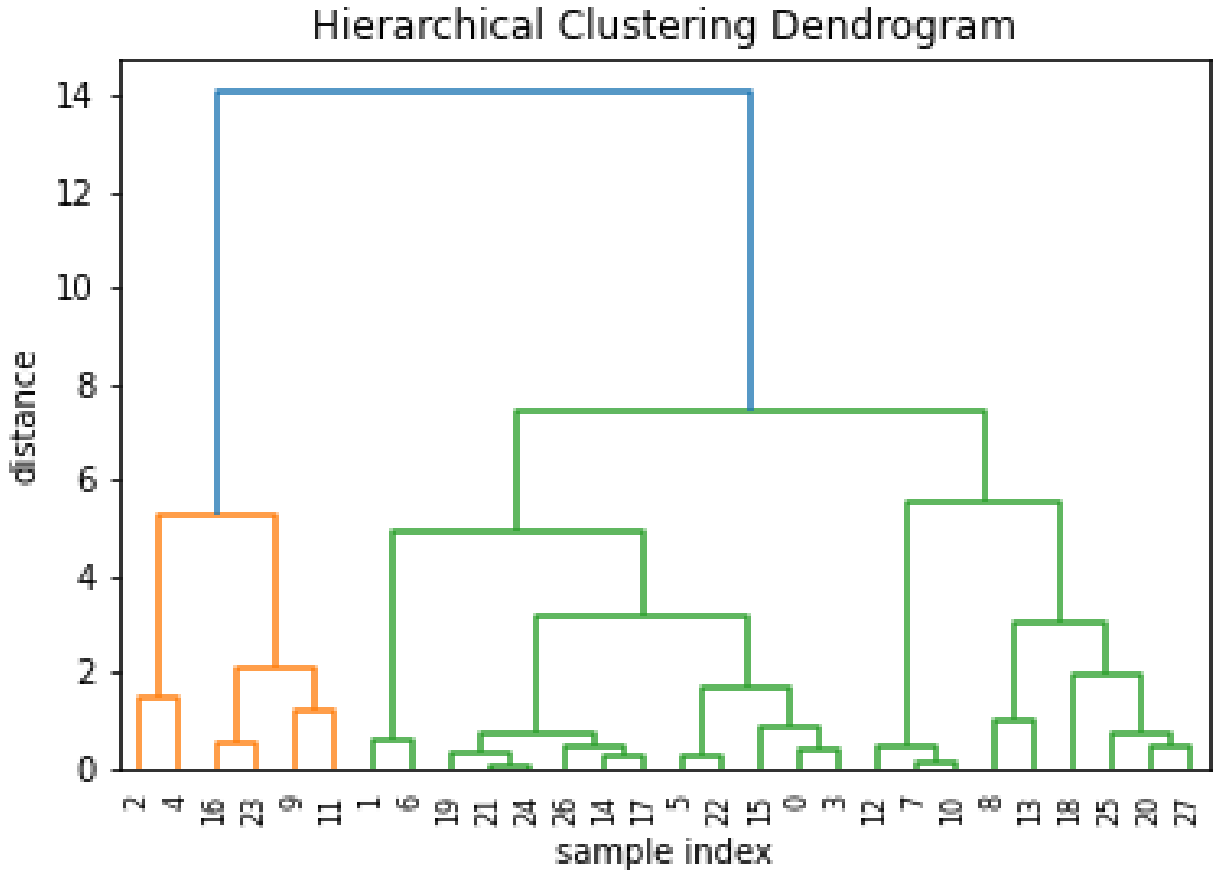

FIGURE 4 – Dendogramme de la CAH pour les deux premiers facteurs

Une des décisions faisant varier grandement l'analyse finale et le nombre de clusters à choisir. Pour la première analyse je fais le choix de prendre 3 clusters. Je paramètre la fonction *fcluster* ainsi :

```
131     k=3
132     clusters = fcluster(Z, k, criterion='maxclust')
```

FIGURE 5 – Premier paramétrage de la fonction *fcluster*

Ce qui revient à réaliser cette coupure :

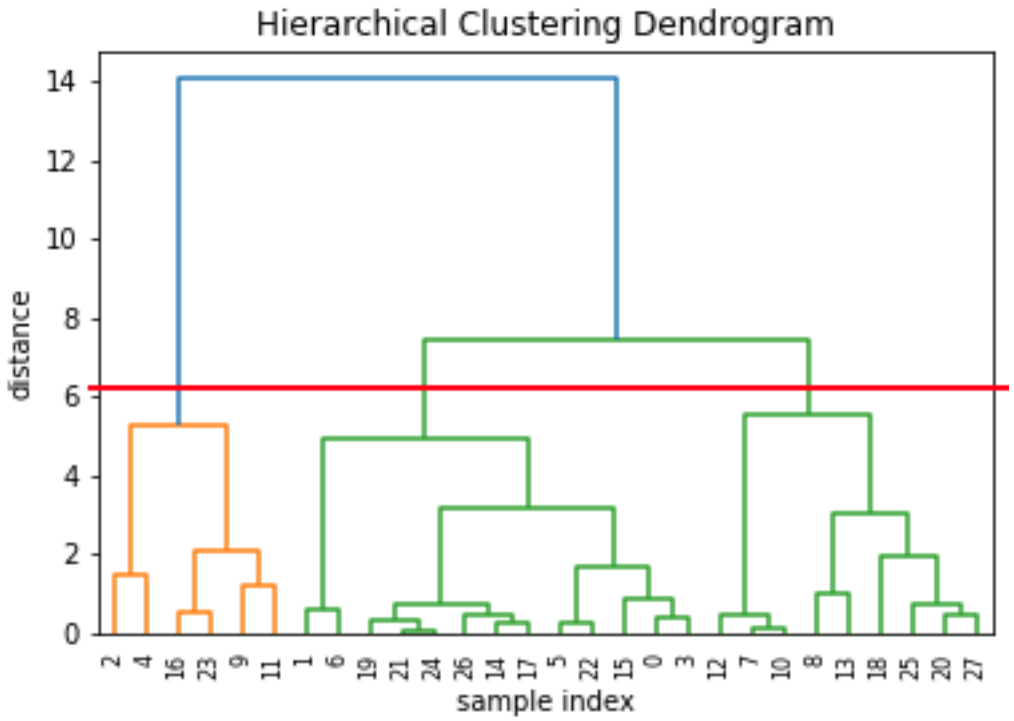

FIGURE 6 – Coupure pour le premier paramétrage

On obtient alors les clusters suivants :

— Cluster 1 en rouge : ['Fnau', 'Fmus', 'Fnaw', 'Fmwe', 'Fnay', 'Fnae']
— Cluster 2 en vert : ['Haus', 'Faus', 'Hmus', 'Hcus', 'Fcus', 'Hayo', 'Fayo', 'Hmyo', 'Hcyo', 'Haes', 'Faes', 'Hmes', 'Hces']
— Cluster 3 en bleu : ['Hawe', 'Fawe', 'Hmwe', 'Hcwe', 'Fcwe', 'Fmyo', 'Fcyo', 'Fmes', 'Fces']

On remarque les hommes et les femmes ne sont pas séparés dans tous les groupes. Cela remet en question l'analyse que j'avais faite sur la nature du premier axe.

On obtient ce graphique ci pour l'analyse des individus :

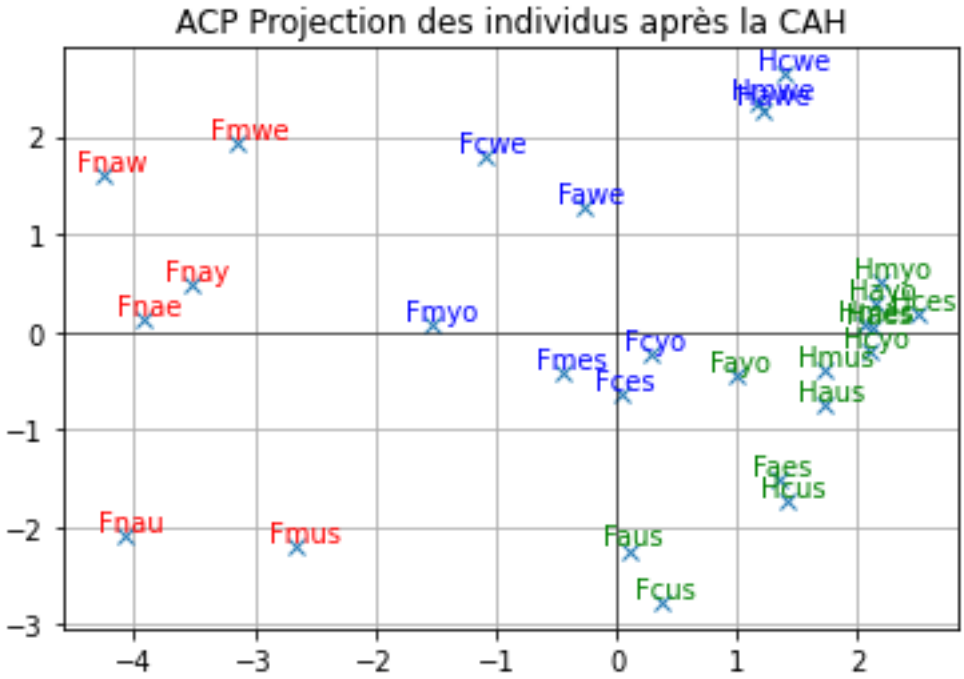

FIGURE 7 – Projection des individus avec les 3 clusters représentés

Plusieurs résultats sont intéressants. Par exemple, un homme célibataire des pays de l'est (HCWE) a une répartition de son temps plus proche de celle d'une femme mariée de Yougoslavie FMYO (intuitivement plutôt remplie de tâches ménagères et de peu d'activités professionnelles) que d'un Homme célibataire de Yougoslavie HCYO (où pour le coup on observe des activités professionnelles dominantes devant les tâches ménagères et familiales comme on s'y attendrait pour HCWE).

Il est intéressant de regarder ce qu'il se passe si l'on change les paramètres de *fcluster*.

Pour les nouveaux paramètres je prends :

```
131     k=4
132     clusters = fcluster(Z, k, criterion='distance')
133     print(clusters)
```

FIGURE 8 – Second paramétrage de *fcluster*

Cette fois ci, le nombre de cluster est décidé par la distance et pas l'inverse. Cette méthode me permet de fixer la coupure parfaitement ou je veux de sorte à optimiser le gain d'inertie produit par le nombre de clusters choisis. Le risque est cependant d'avoir des clusters qui se ressemblent tellement qu'il ne sera pas possible de tirer les bonnes conclusions. La coupure est la suivante :

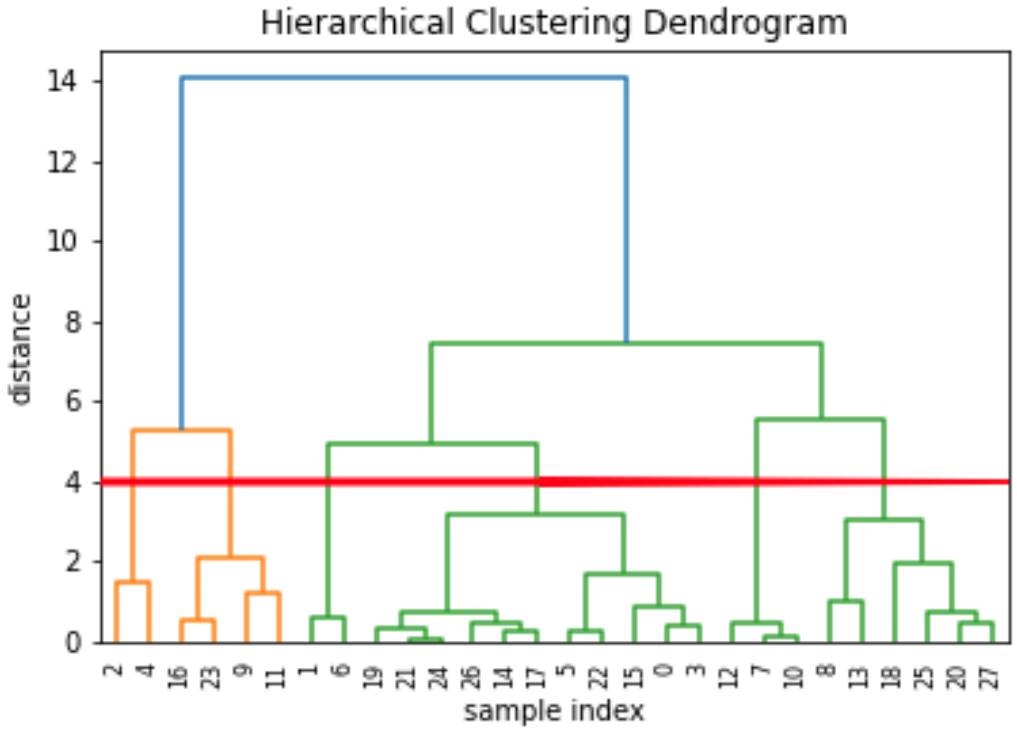

FIGURE 9 – Coupure pour le deuxième paramétrage

J'obtiens les clusters suivants :
— Cluster 1 (Rouge) : ['Fnau', 'Fmus']
— Cluster 2 (Vert) : ['Fnaw', 'Fmwe', 'Fnay', 'Fnae']
— Cluster 3 (Bleu) : ['Faus', 'Fcus']
— Cluster 4 (Noir) : ['Haus', 'Hmus', 'Hcus', 'Hayo', 'Fayo', 'Hmyo', 'Hcyo', 'Haes', 'Faes', 'Hmes', 'Hces']
— Cluster 5 (Violet) :['Hawe', 'Hmwe', 'Hcwe']
— Cluster 6 (Orange) : ['Fawe', 'Fcwe', 'Fmyo', 'Fcyo', 'Fmes', 'Fces']
On obtient cette figure ci pour la projection des individus :

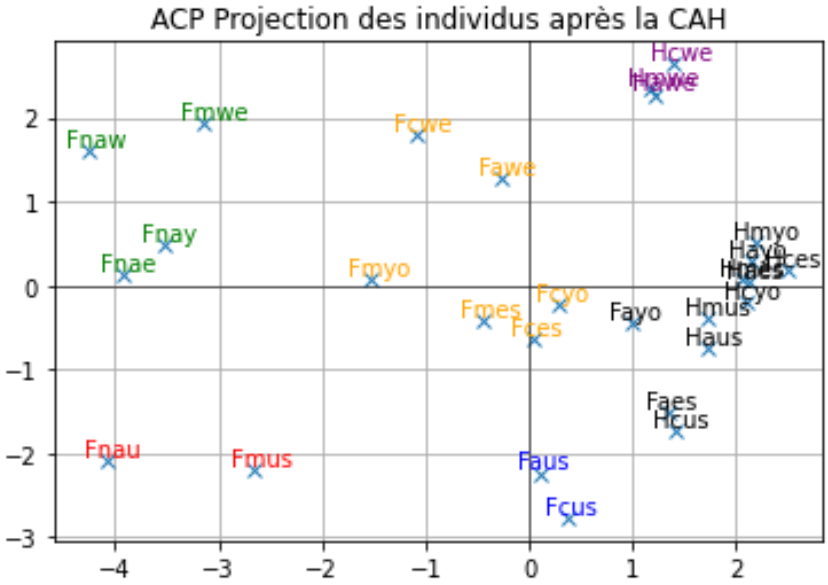

FIGURE 10 – Projection des individus avec les 6 clusters représentés

Cette fois ci, on voit que les hommes sont réunis dans seulement 2 clusters alors qu'on trouve des femmes dans les 6. Je vois cette observation comme un moyen de dire que la répartition des heures des femmes varie beaucoup plus en fonction de leur situation matrimoniale, géographique et professionnelle alors que celle des hommes est globalement toujours la même.

2.3 ACP et Centres mobiles

Nous procédons à la même analyse en remplaçant la CAH par les kmeans.

Je remarque tout d'abord que si l'on ne spécifie pas le nombre d'itérations de la méthode des kmeans, je me retrouve souvent avec un cluster vide. Cela est sûrement dû au fait que les centroids sont placés aléatoirement sur le plan. Même en utilisant un grand nombre d'itérations et les méthodes d'initialisations *'points'* et *'random'* j'obtiens des clusters toujours différents.

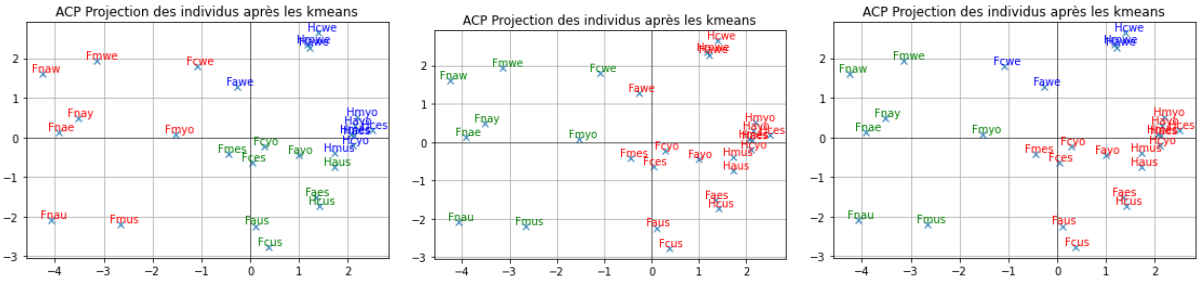

FIGURE 11 – Variation des clusters pour la méthode des kmeans sans initialiser avec une matrice de points

Je décide donc d'initialiser les centroids avec une matrice de points. Je prends 3 clusters et les centres sont les points 'Faes', 'Fnae' et 'Hmwe'. Car ce sont des points qui sont plutôt éloignés ... L'algorithme donne alors toujours le même résultat, ce qui est normal.

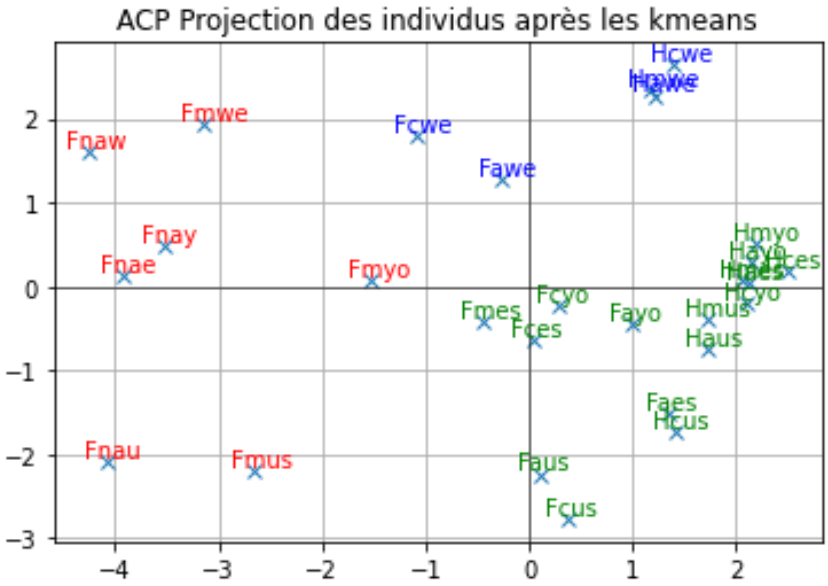

FIGURE 12 – Résultat pour l'initialisation décrite ci-dessus

Il est alors intéressant de comparer avec le résultat de la CAH et les résultats pour des initialisations aléatoires. On remarque que les clusters diffèrent toujours au milieu du graphique. Les points se trouvant aux extrémités sont, eux, globalement toujours dans les mêmes clusters. J'en conclus qu'il est difficile de donner des conclusions sur les points proches du centre du graphique (ce que l'on savait déjà). Les points à exploiter pour avoir des résultats fiables sont ceux qui appartiennent toujours aux mêmes clusters.

3 Analyse des correspondances multiples

3.1 ACM

La première question à se poser et celle du nombre d'intervalles que l'on veut créer. Il est possible de transformer le nombre d'heures en un système de Oui/Non en fixant à deux le nombre d'intervalles. Le risque est alors de perdre de la richesse d'information. Il est également possible, et c'est ce que j'ai fait, de créer 4 intervalles à chaque fois. Ces intervalles représentent une répartition absente, légère, moyenne et élevée de l'occupation. Les variables sont donc les suivantes : ['PROF1', 'PROF2', 'PROF3', 'PROF4', 'TRAN1', 'TRAN2', 'TRAN3', 'TRAN4', 'MENA1', 'MENA2', 'MENA3', 'MENA4', 'ENFA1', 'ENFA2', 'ENFA3', 'ENFA4', 'COUR1', 'COUR2', 'COUR3', 'COUR4', 'TOIL1', 'TOIL2', 'TOIL3', 'TOIL4', 'REPA1', 'REPA2', 'REPA3', 'REPA4', 'SOMM1', 'SOMM2', 'SOMM3', 'SOMM4', 'TELE1', 'TELE2', 'TELE3', 'TELE4', 'LOIS1', 'LOIS2', 'LOIS3', 'LOIS4'] où '4' représente un grand nombre d'heures et '1' un faible nombre d'heures.

Nous obtenons alors le diagramme d'inertie suivant :

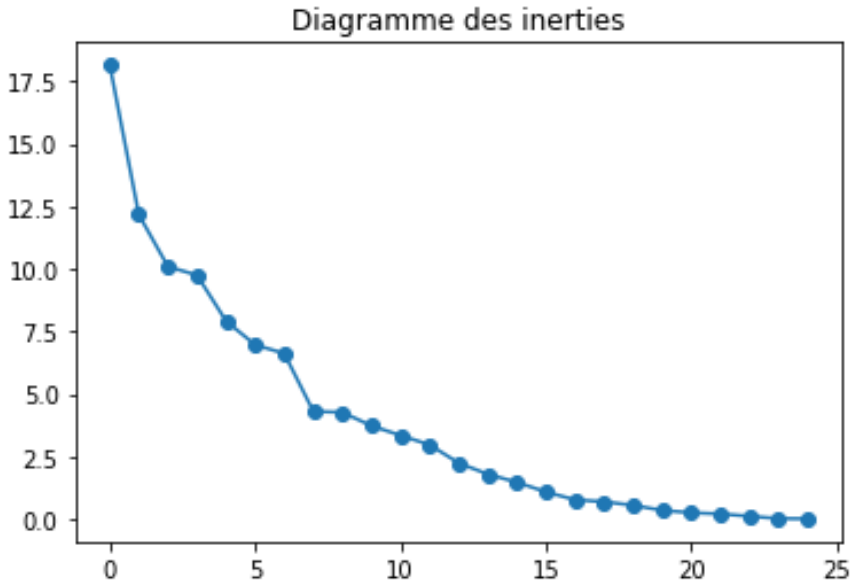

FIGURE 13 – Diagramme d'inertie pour l'ACM

Les deux premiers facteurs sont à 18,14 et 12,164 ce qui est plutôt satisfaisant.

Il est alors possible de comparer les résultats obtenus pour l'ACP et l'ACM. La première chose à remarquer est qu'à nouveau l'axe principal semble être fortement corrélé au sexe de l'individu pour la répartitions des modalités.

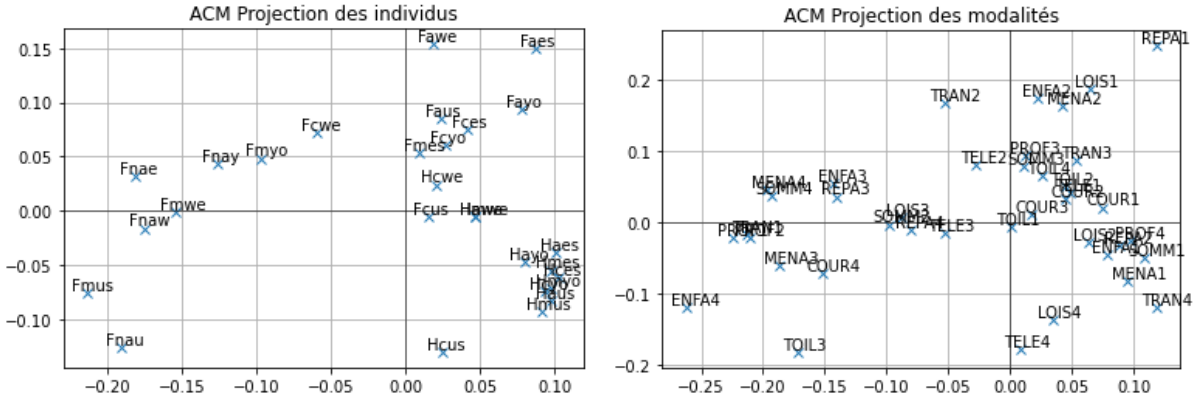

FIGURE 14 – Projection des individus et des modalités pour l'ACM seule

On voit bien que les hommes sont présents uniquement sur la partie droite du graphique et que les modalités présentes à cet endroit sont des 1 pour les tâches ménagères et familiale et des 4 pour les tâches professionnelles.

Les femmes, quant à elles, sont à nouveau très dispersées. On remarque que les REPA, ENFA, MENA sont très élevés dans les zones où on ne trouve que des femmes. Cela confirme l'analyse qui avait été faite pour l'ACP. La valeur ajoutée de la présente analyse est qu'elle permet d'ajouter du sens au deuxième axe factoriel. En effet, on observe que selon le deuxième axe, les valeurs de TELE et LOIS sont très éloignées et sont dé-corrélées du premier axe. Je considère que le deuxième axe décrit la 'non-récréativité' des modalités.

Plus une modalité est placée haut sur cette axe, plus elle est peu amusante. TELE4, LOIS4, MENA1 sont placés en bas alors que LOIS1, MENA3 sont placés en haut. Aussi, les modalités que l'on observe au centre du graphique sont dé-corrélées des deux axes et donc du sexe des individus ainsi que de l'amusement qu'elles procurent. Ce sont des modalités que tous les individus pratiquent sans en avoir le choix (toilettes et courses globalement).

3.2 ACM et CAH

Le seule différence avec l'AFC est qu'il faut faire les manipulations de CAH pour les individus et aussi pour les modalités.

J'obtiens alors les dendogrammes suivants :

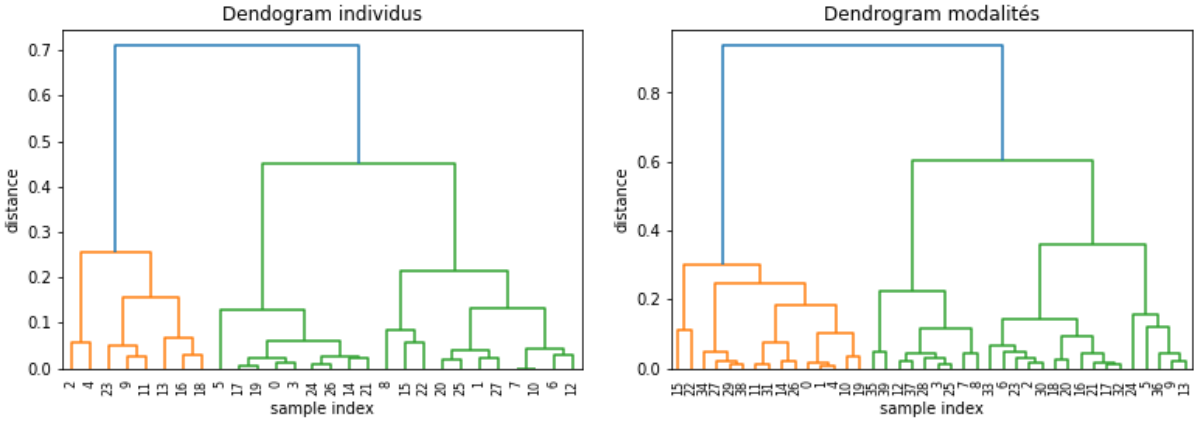

FIGURE 15 – Dendogrammes pour les individus et les modalités

Je décide alors de fixer le nombre de clusters à 3 car on voit que la plus grande perte d'inertie se fait au moment de cette séparation.

J'obtiens alors les clusters suivants pour les individus :

— Cluster 1 (Rouge) : ['Fnau', 'Fmus', 'Fnaw', 'Fmwe', 'Fcwe', 'Fnay', 'Fmyo', 'Fnae']
— Cluster 2 (Vert) : ['Haus', 'Hmus', 'Hcus', 'Hayo', 'Hmyo', 'Hcyo', 'Haes', 'Hmes', 'Hces']
— Cluster 3 (Bleu) : ['Faus', 'Fcus', 'Hawe', 'Fawe', 'Hmwe', 'Hcwe', 'Fayo', 'Fcyo', 'Faes', 'Fmes', 'Fces']

J'obtiens ceux-ci pour les modalités :

— Cluster 1 (Orange) : ['PROF1', 'PROF2', 'TRAN1', 'MENA3', 'MENA4', 'ENFA3', 'ENFA4', 'COUR4', 'TOIL3', 'REPA3', 'REPA4', 'SOMM2', 'SOMM4', 'TELE3', 'LOIS3']
— Cluster 2 (Violet) : ['PROF4', 'TRAN4', 'MENA1', 'ENFA1', 'REPA2', 'SOMM1', 'TELE4', 'LOIS2', 'LOIS4']
— Cluster 3 (Cyan) : ['PROF3', 'TRAN2', 'TRAN3', 'MENA2', 'ENFA2', 'COUR1', 'COUR2', 'COUR3', 'TOIL1', 'TOIL2', 'TOIL4', 'REPA1', 'SOMM3', 'TELE1', 'TELE2', 'LOIS1']

Il est intéressant d'analyser les clusters qui ont été formés pour les individus. On voit que le cluster 2 est bien composé uniquement d'hommes ce qui veut dire que leur contributions aux axes factoriels étaient identiques. Cela confirme les hypothèses faites sur la signification du premier axe factoriel à mon sens.

D'autres parts, je trouve qu'il est encore compliqué d'apporter une conclusion sur les deux clusters de femmes qui semblent apparaître à chaque fois. La seule chose à dire est la suivante : dans les pays où la condition de la femme n'a pas encore grandement évoluée, on retrouve les 3 types de femmes dans le même cluster. Dans les endroits où la condition de la femme est en progrès (ie USA, et pays de l'ouest). Les répartitions des femmes dans les clusters diffèrent en fonction de leur situation matrimoniale. On voit que dans les pays de l'ouest, là où la situation des hommes et des femmes est la plus égalitaire selon moi, toutes les modalités appartiennent quasiment au cluster 3 alors que dans le reste des cas, les hommes sont dans le cluster 1. Après avoir fait quelques recherches, j'ai trouvé qu'en effet l'Europe de l'ouest est un des secteurs où la situation des femmes diffère le moins de celle des hommes.

J'ai sélectionné ce passage d'un rapport de l'ONU qui me semble bien illustrer mes propos : "Seulement 50% des femmes en âge de travailler sont actives, comparé à 77% des hommes, précise par ailleurs le rapport, ajoutant que les femmes ont tendance à se cantonner à des emplois faiblement rémunérés et gagnent en moyenne entre 70 et 90% de ce que gagnent les hommes.

En outre, **les femmes passent en moyenne trois heures de plus par jour que les hommes à accomplir des tâches ménagères et à prendre soin des membres de leur famille dans les pays en développement (et deux heures de plus par jour que les hommes dans les pays développés).**". L'article est disponible en cliquant [ici (Rapport ONU)]

Pour les projections des individus et des modalités, j'obtiens les graphiques suivants :

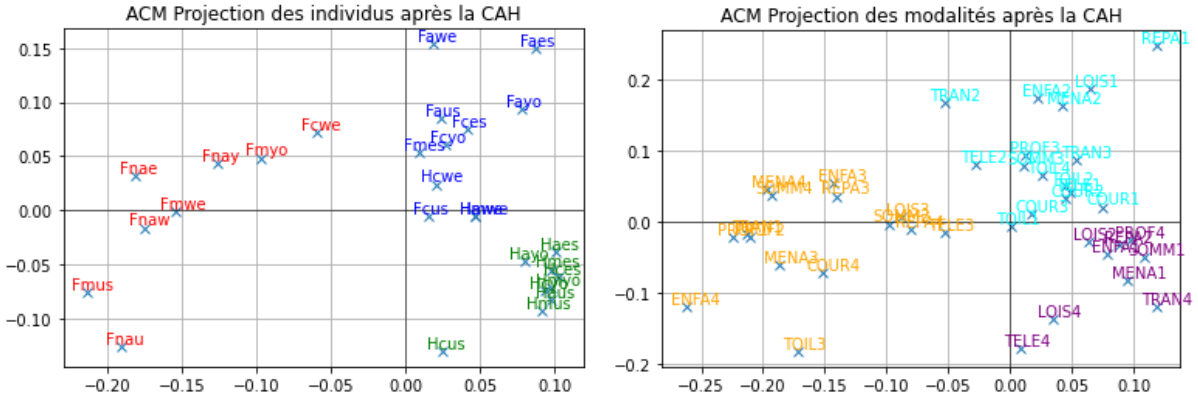

FIGURE 16 – Projections des individus et des modalités après la CAH

On peut alors apporter quelques nouvelles idées.

On voit que les hommes sont, en moyenne, plus enclin à pratiquer des activités du cluster 2, c'est à dire des activités récréatives. Les femmes du cluster 3 (celui où la condition des femmes a évolué) sont moins enclin à faire des choses amusantes que les femmes du cluster 1.

3.3 ACM et Centres mobiles

Comme avec l'ACP, on initialise les centroids avec des points particuliers pour essayer d'avoir des résultats qui ressemblent à ceux obtenus avec la CAH.

Je décide de prendre [TRAN2, MENA2, MENA3] pour les premiers centroids des modalités et [Fnaw, Fces, Hmes] pour les centroids des individus.

J'obtiens les clusters suivants pour les individus :

— Cluster 1 (Rouge) : ['Faus', 'Fcus', 'Fawe', 'Hcwe', 'Fcwe', 'Fayo', 'Fcyo', 'Faes', 'Fmes', 'Fces']
— Cluster 2 (Vert) : ['Fnau', 'Fmus', 'Fnaw', 'Fmwe', 'Fnay', 'Fmyo', 'Fnae']
— Cluster 3 (Bleu) : ['Haus', 'Hmus', 'Hcus', 'Hawe', 'Hmwe', 'Hayo', 'Hmyo', 'Hcyo', 'Haes', 'Hmes', 'Hces']

J'obtiens ceux-ci pour les modalités :

— Cluster 1 (Orange) : ['PROF3', 'TRAN2', 'TRAN3', 'MENA2', 'ENFA2', 'TOIL2', 'TOIL4', 'REPA1', 'SOMM3', 'TELE2', 'LOIS1']
— Cluster 2 (Violet) : ['PROF1', 'PROF2', 'TRAN1', 'MENA3', 'MENA4', 'ENFA3', 'ENFA4', 'COUR4', 'TOIL3', 'REPA3', 'REPA4', 'SOMM2', 'SOMM4', 'LOIS3']
— Cluster 3 (Cyan) : ['PROF4', 'TRAN4', 'MENA1', 'ENFA1', 'COUR1', 'COUR2', 'COUR3', 'TOIL1', 'REPA2', 'SOMM1', 'TELE1', 'TELE3', 'TELE4', 'LOIS2', 'LOIS4']

On retrouve des clusters différents de ce qu'on avait avec la CAH puisque ce coup-ci, il a un cluster qui réunit tous les hommes et deux clusters de femmes. Il est possible d'apporter la même analyse qu'avec la CAH pour les différences entre les deux clusters de femmes.

Pour les projections des individus et des modalités, j'obtiens les graphiques suivants :

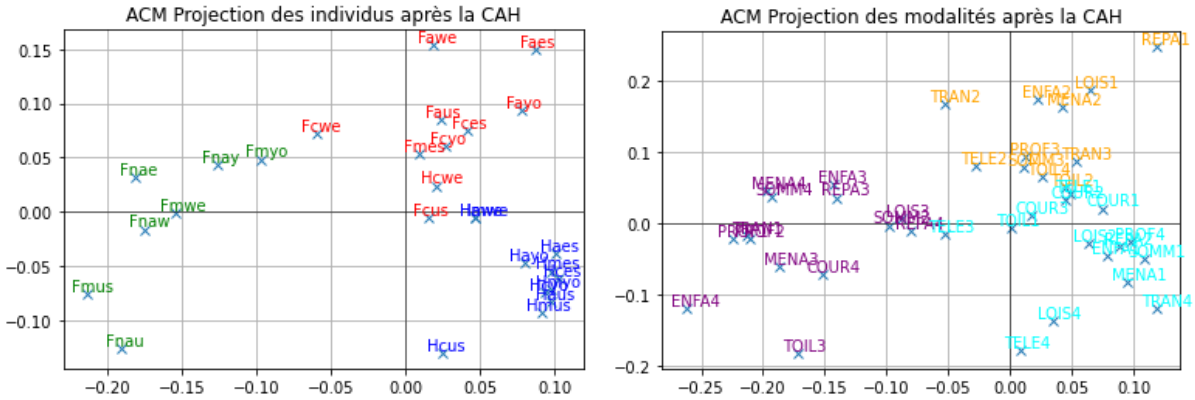

FIGURE 17 – Projections des individus et des modalités après la méthode des kmeans

Les changements de clusters ont lieu proche du centre du repère, il serait idiot d'essayer d'apporter des conclusions là dessus. A mon sens, c'est uniquement dû au fait que les poids des facteurs sont faibles pour les modalités et les individus du centre. C'est pour cela que leurs appartenances aux clusters diffèrent en fonction des méthodes.

4 Conclusion

Pour conclure, l'apport de la CAH sur l'ACP et l'ACM est à mon sens très utile. C'est en analysant les CAH que j'ai compris le plus de choses vis à vis de ce jeux de donnnées. Je pense que l'analyse la plus riche est celle pour la CAM + CAH.