Projet AD

Dellouve Théo
Gazzo Sandro

14 mai 2020

1 Introduction

Notre base de données étaient en anglais. Pour des raisons pratiques, nous garderons les noms des pays et des variables en anglais. Nous traduirons cependant le nom des variables en français.

Notre base de données exprime le pourcentage de chômage, d'emploi et de participation dans différents pays en fonction du genre et de la nationalité pour l'année 2018. Cela nous donne un total de 18 variables quantitatives, qui ne sont pas toutes indépendantes deux à deux d'après la matrice de corrélation.

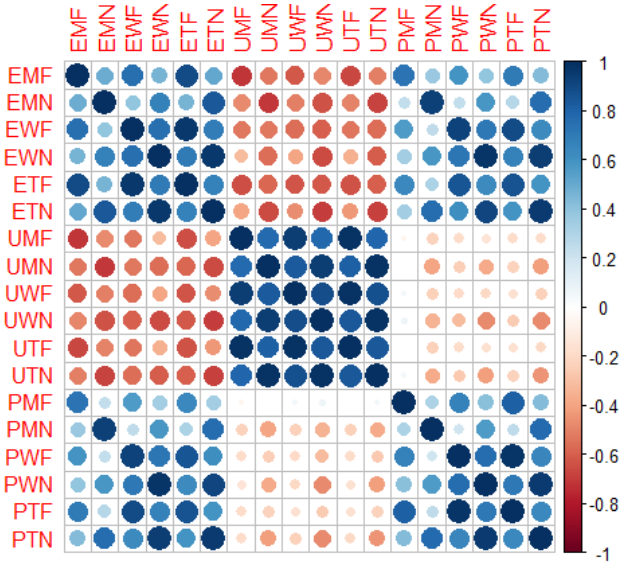

FIGURE 1 – Matrice de corrélation.

Ainsi, il est cohérent de faire une ACP normée.

Pour améliorer la lisibilité lors de l'ACP nous avons utilisé les acronymes de nos variables dont les noms sont assez longs. La première lettre désigne le taux d'emploi, de chômage ou de participation des deux lettres suivantes. La deuxième lettre fait référence au genre, c'est à dire masculin, féminin ou les deux cumulés (Total). Enfin, la dernière lettre correspond soit aux personnes nées dans le pays étudié (natif), soit aux personnes nées ailleurs que dans le pays étudié (étranger).

Voici la signification détaillée des variables :

— EMF : Employment rate, Men, Foreign-born (Taux d'emploi masculin étranger) ;
— EMN : Employment rate, Men, Native (Taux d'emploi masculin natif) ;
— EWF : Employment rate, Women, Foreign-born (Taux d'emploi féminin étranger) ;
— EWN : Employment rate, Women, Native (Taux d'emploi féminin natif) ;
— ETF : Employment rate, Total, Foreign-born (Taux d'emploi total étranger) ;
— ETN : Employment rate, Total, Native (Taux d'emploi total natifs) ;
— UMF : Unemployment rate, Men, Foreign-born (Taux de chômage masculin étranger) ;
— UMN : Unemployment rate, Men, Native (Taux de chômage masculin natif) ;
— UWF : Unemployment rate, Women, Foreign-born (Taux de chômage féminin étranger) ;
— UWN : Unemployment rate, Women, Native (Taux de chômage féminin natif) ;
— UTF : Unemployment rate, Total, Foreign-born (Taux de chômage total étranger) ;
— UTN : Unemployment rate, Total, Native (Taux de chômage total natif) ;
— PMF : Participation rate, Men, Foreign-born (Taux de participation masculin étranger) ;
— PMN : Participation rate, Men, Native (Taux de participation masculin natif) ;
— PWF : Participation rate, Women, Foreign-born (Taux de participation féminin étranger) ;
— PWN : Participation rate, Women, Native (Taux de participation féminin natif) ;
— PTF : Participation rate, Total, Foreign-born (Taux de participation total étranger) ;
— PTN : Participation rate, Total, Native (Taux de participation total natif).

2 ACP normée

2.1 Individus supplémentaires

On remarque qu'on obtient des individus pour lesquels la contribution à l'axe 1 ou à l'axe 2 est supérieur a 20%, on choisit donc de mettre les individus "Sweden", "Spain" et "Greece" en individus supplémentaires.

2.2 Choix du nombre d'axes et interprétation

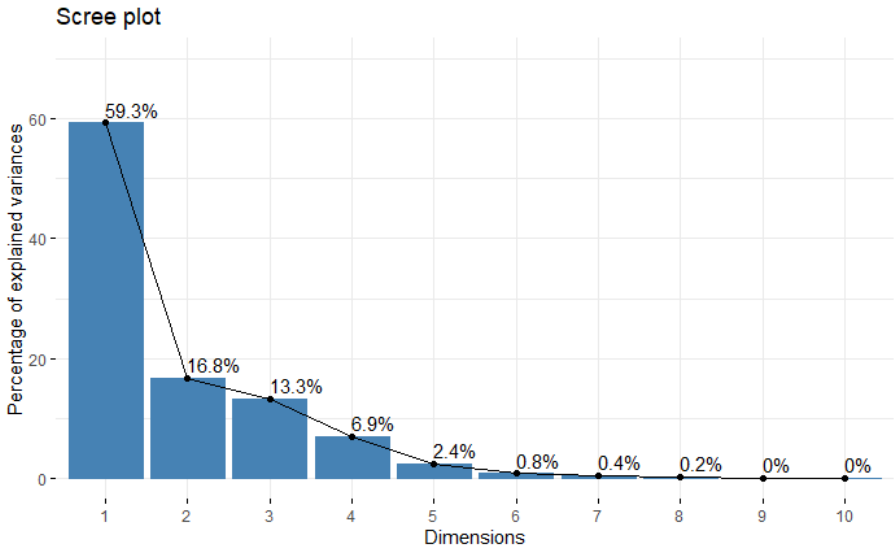

FIGURE 2 – Valeurs propres obtenues et leur pourcentage d'inertie.

— Règle du coude : on conserve 1 axes.

— Kaiser : on conserve les 4 premier axes car les trois première valeurs propres sont supérieur à la moyenne

— Thumb : on conserve 3 variables pour conserver au moins 80% d'inertie.

On va s'intéresser maintenant à l'interprétabilité des axes.

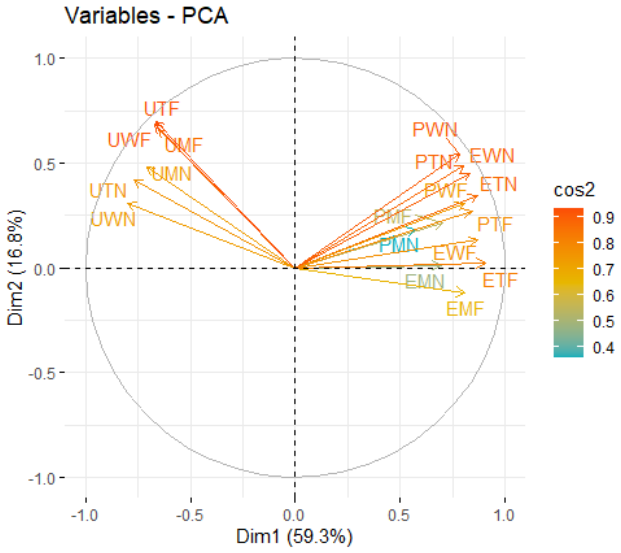

FIGURE 3 – Cercle de corrélation.

a) Premier axe

On commence par regarder le premier axe.

	correlation
ETF	9.064E-01
EWF	8.697E-01
ETN	8.692E-01
PTF	8.431E-01
EWN	8.333E-01
EMF	8.080E-01
PWF	8.074E-01
PTN	8.036E-01
PWN	7.838E-01
PMF	7.058E-01
EMN	6.954E-01
PMN	5.710E-01
UMF	-6.525E-01
UTF	-6.673E-01
UWF	-6.747E-01
UMN	-7.122E-01
UTN	-7.699E-01
UWN	-8.014E-01

FIGURE 4 – Tableau des corrélations de l'axe 1.

Le premier axe est corrélé :

— positivement avec EMF (0.808), EWF (0.87), EMN (0.695), EWN (0.833), ETF (0.906), ETN (0.869), PMN (0.571), PWN (0.784), PMF (0.706), PWF (0.807), PTF (0.843), PTN (0.804).

— négativement avec UMF (-0.652), UWF(-0.675), UMN(-0.712), UWN(-0.801), UTF(-0.667), UTN(-0.77).

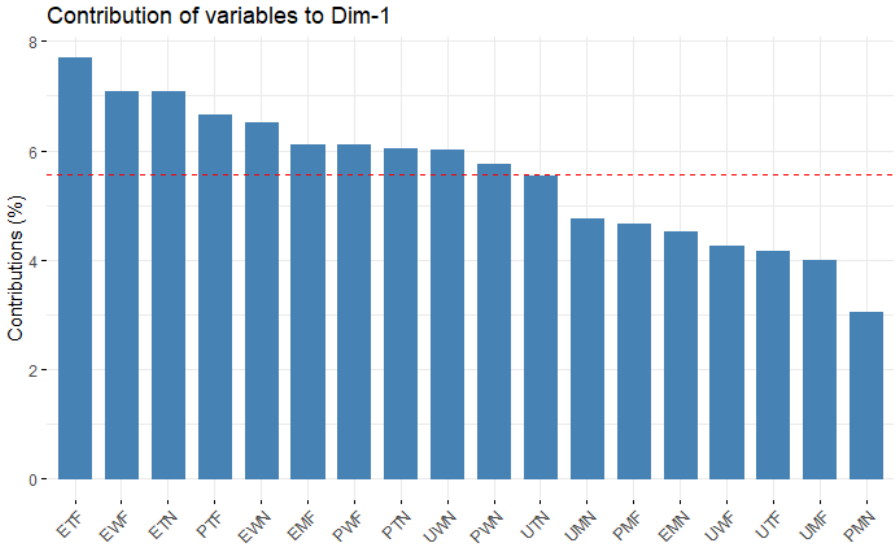

FIGURE 5 – Contribution des variables pour l'axe 1.

La valeur des corrélations obtenues renforce l'observation des contributions, toutes les variables contribuent approximativement de la même façon a cette axe.

Cet axe oppose donc deux groupes. Un premier contenant des pays où le taux de chômage est élevé tel que "Turkey" (caractérisé par le côté négatif de l'axe 1), et un deuxième contenant des pays où le taux d'emploi et le taux de participation est élevé tel que "Iceland","New Zealand" et "United Kingdom" (caractérisés par le côté positif de l'axe 1). Les individus cités sont bien représentés sur l'axe 1. En effet, le $\cos^2$ de ces individus est proche de 1 (respectivement 0.906 pour "Iceland", 0.903 pour "New Zealand" et 0.848 pour "United Kingdom").

On peut observer également ceci via la figure 9.

b) Deuxième axe

On s'intéresse ensuite au deuxième axe.

	correlation
UTF	6.998E-01
UWF	6.822E-01
UMF	6.597E-01
PWN	5.408E-01
PTN	4.860E-01
UMN	4.801E-01
EWN	4.501E-01
UTN	4.174E-01

FIGURE 6 – Tableau des corrélations de l'axe 2.

Le deuxième axe est corrélé positivement avec UTF (0.7), UWF (0.682) et UMF(0.66).

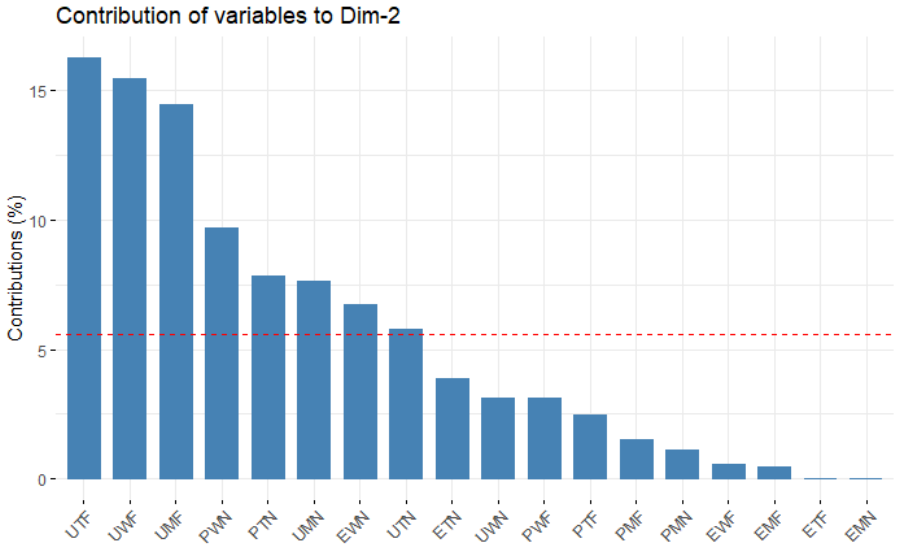

FIGURE 7 – Contribution des variables pour l'axe 2.

Contrairement au cas précédent, on observe 3 variables qui contribuent plus que les autres sur cet axe, les corrélations obtenues soulignent également ce résultat.

Cet axe oppose donc un premier groupe contenant des pays où le taux de chômage des étrangers est plus élevé tel que "Finland" (caractérisé par le côté positif de l'axe 2), et un deuxième groupe où le taux de chômage des étrangers est moins élevé tel que "Hungary", "Poland" et "United States" (caractérisés par le côté négatif de l'axe 2). Les individus cités sont bien représentés sur l'axe 2. En effet, le $\cos^2$ de ces individus est proche de 1 (respectivement 0.790 pour "Hungary", 0.668 pour "Poland", 0.610 pour "Finland" et 0.562 pour "United States").

On retrouve ces observations à la figure 9.

c) Troisième axe

	correlation
PTF	4.499E-01
PMF	4.367E-01
PWF	4.330E-01
ETF	4.178E-01
EWF	4.168E-01
EMF	3.873E-01
PMN	-5.958E-01
EMN	-6.006E-01

FIGURE 8 – Tableau des corrélations de l'axe 3.

Le troisième axe est corrélé négativement avec EMN (-0.601) et PMN (-0.596), mais ces variables sont plus corrélées en valeur absolue avec le premier axe. Il n'est donc pas interprétable.

d) Choix des axes

Au final, on choisit de conserver les deux premiers axes.

Le premier axe contient 59.33% de l'inertie totale, et le deuxième axe contient 16.77% de l'inertie totale. On conserve donc 76.1% de l'inertie totale avec ces deux axes. On est proche de 80% d'inertie conservé en plus de respecter la règle du coude. De plus, le résultat en dimension 2 est facilement interprétable graphiquement.

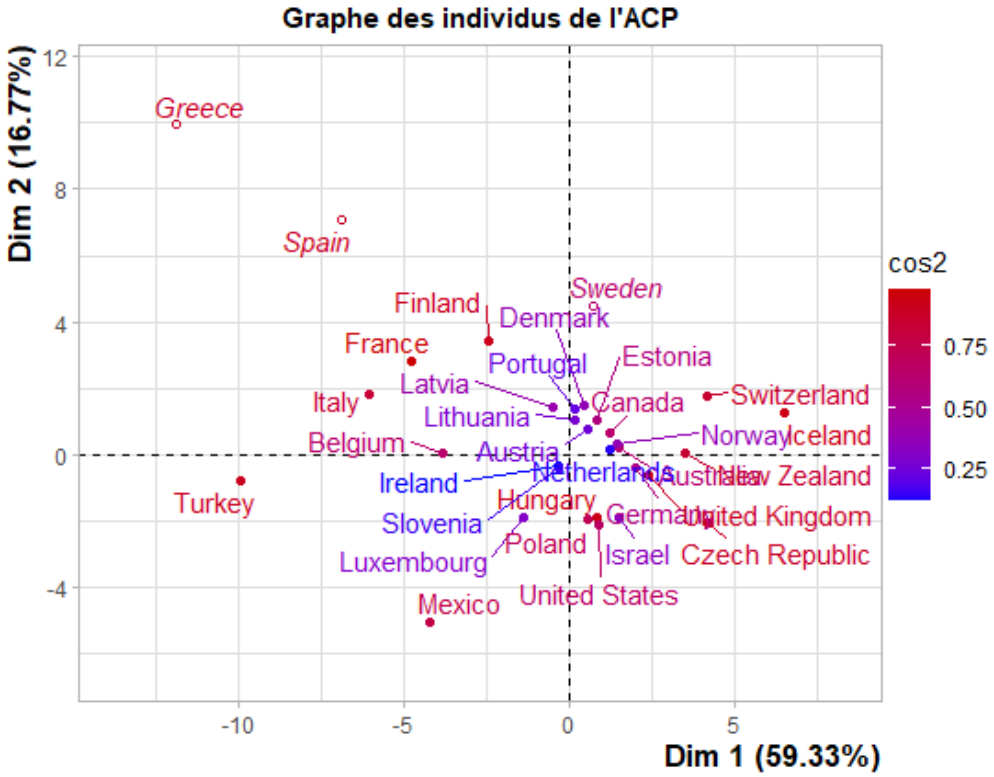

FIGURE 9 – Graphique des individus après ACP.

Les individus supplémentaires sont représentés par un rond vide tandis que les autres individus sont représentés par un rond plein. Ce graphique représente également la valeur cos2 par un dégradé de couleur allant du bleu au rouge. Ainsi, un pays coloré en bleu n'est pas bien représenté dans le plan tandis qu'un pays coloré en rouge est bien représenté.

3 Classification

On effectue ensuite une classification des individus, dont on comparera les résultats avec ceux de l'ACP. Pour commencer, on cherche a savoir le nombre de groupe optimal dans la classification. On choisi le critère *k-means*. On fera donc une classification *k-means* sur les résultats de l'ACP normée effectuée précédemment. On observe le graphique ci-dessous.

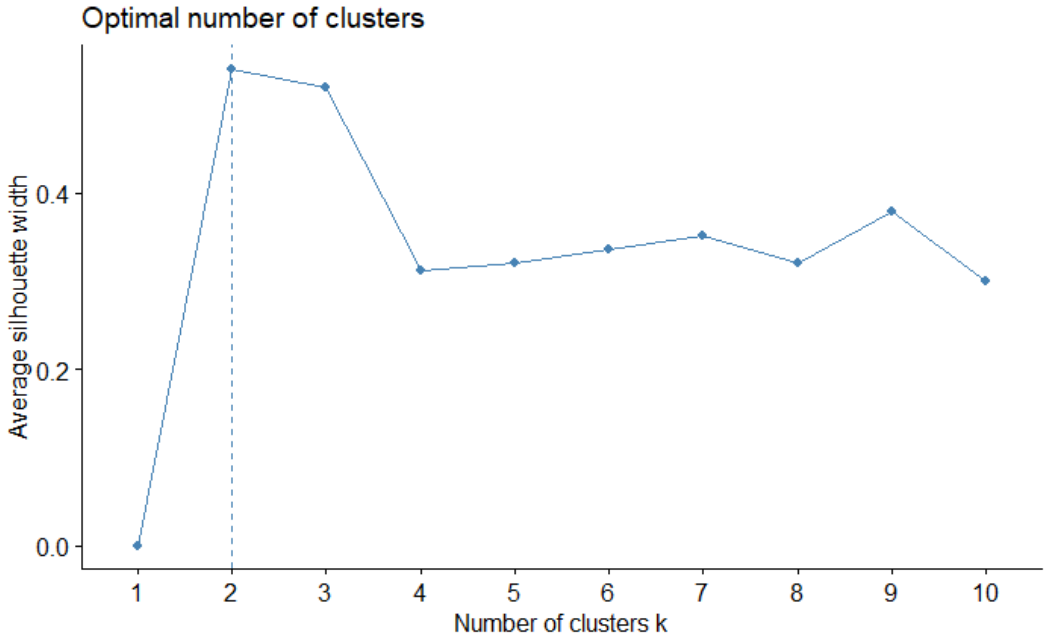

FIGURE 10 – Nombre de classes optimal pour une classification de *k-means*.

On observe que le nombre de groupes optimal d'après ce critère est 2. On trace ensuite le dendrogramme pour avoir une idée de la répartition des pays dans ces groupes.

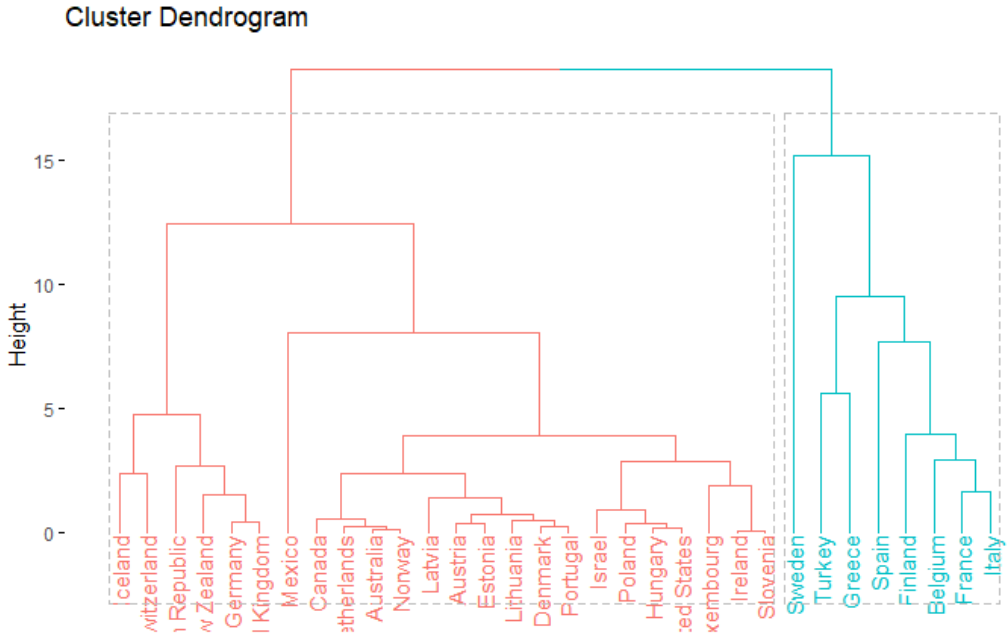

FIGURE 11 – Dendrogramme des individus, répartis en 2 classes.

Pour avoir une meilleure visualisation, on regarde le graphique représentant les groupes obtenus avec les pays contenus dans chacun d'eux en faisant apparaître les centres de gravité.

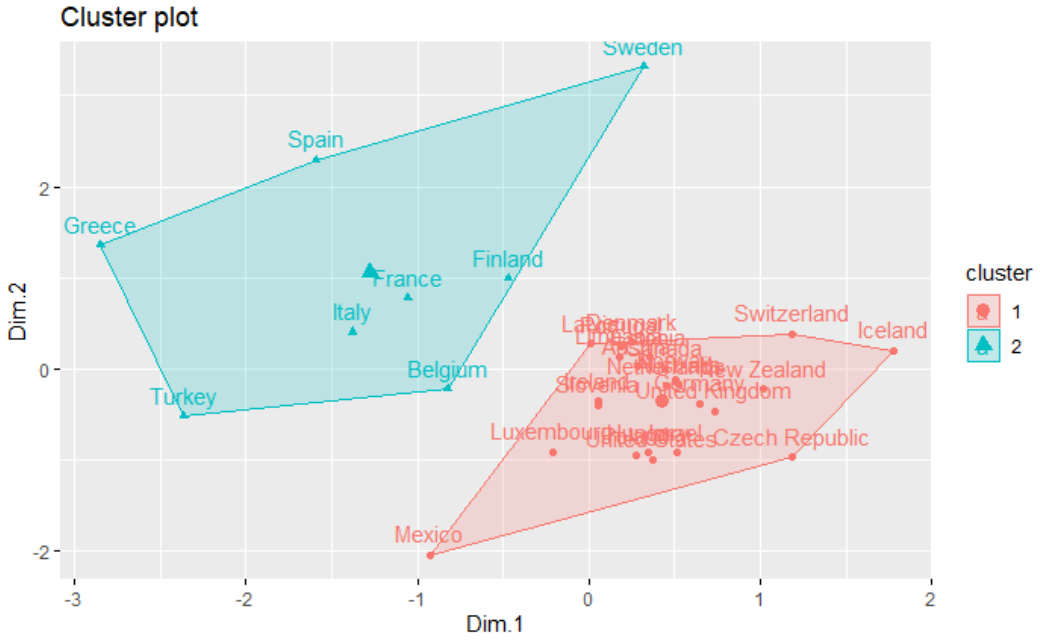

FIGURE 12 – Les 2 classes obtenues.

On observe ensuite le graphe de la silhouette.

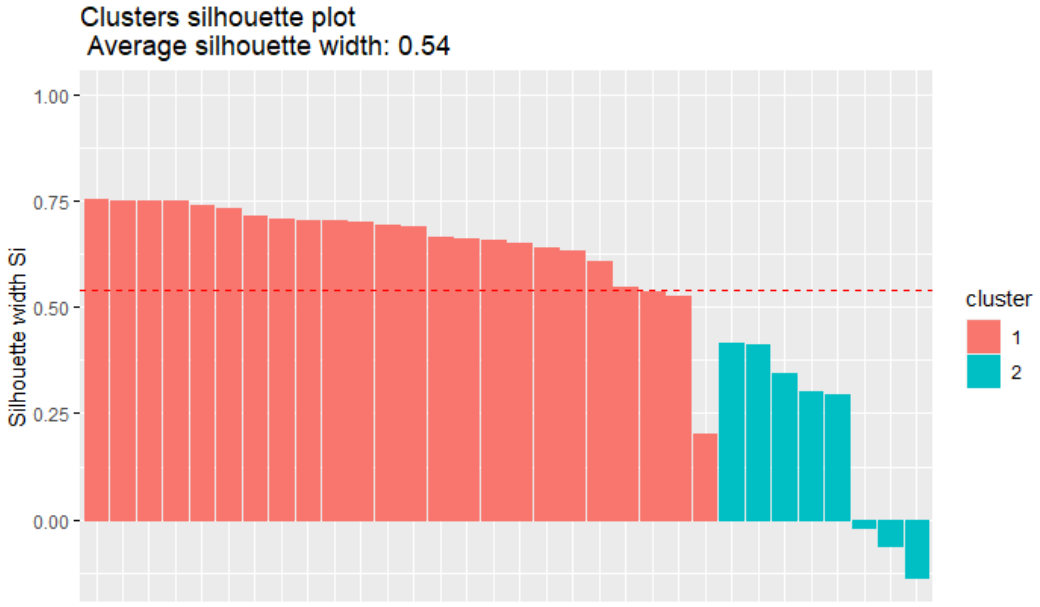

FIGURE 13 – Silhouette de la classification.

```
##      Australia        Austria         Canada Czech Republic        Denmark
##              1              1              1              1              1
##        Germany        Hungary        Iceland        Ireland     Luxembourg
##              1              1              1              1              1
##         Mexico    Netherlands    New Zealand         Norway         Poland
##              1              1              1              1              1
##       Portugal    Switzerland United Kingdom  United States        Estonia
##              1              1              1              1              1
##         Israel       Slovenia         Latvia      Lithuania        Belgium
##              1              1              1              1              2
##        Finland         France          Italy         Turkey         Sweden
##              2              2              2              2              2
##          Spain         Greece
##              2              2
```

FIGURE 14 – Ordres des pays (dans la figure 13).

On remarque qu'il y a 3 pays en négatif dans la classe 2. Ces 3 individus seraient donc dans le mauvais groupe.. Ces 3 pays correspondent à "Sweden", "Spain" et "Greece", autrement dit, les 3 individus supplémentaires, ce qui peut expliquer cette valeur négative.

On regarde ensuite la position des pays de chaque groupe dans le graphiques des individus obtenu avec l'ACP pour pouvoir comparer les résultats.

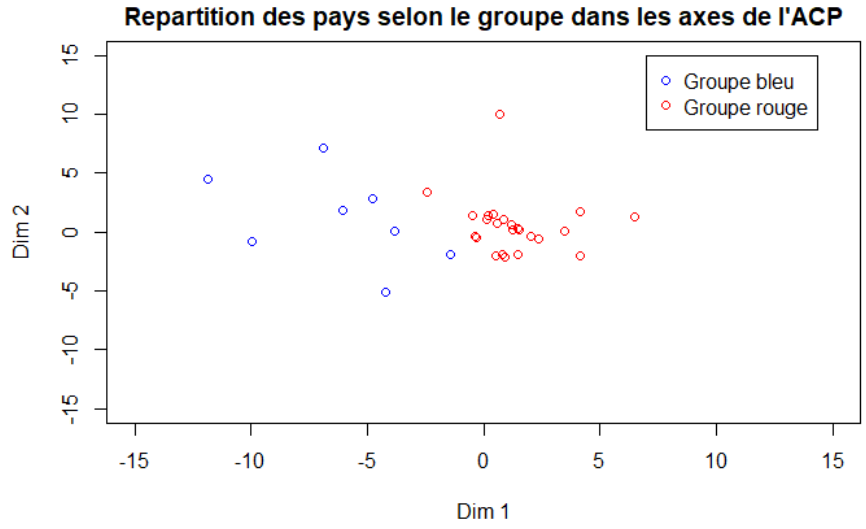

FIGURE 15 – Graphique des individus après ACP et d'après leur classe (voir figure 9).

Les pays dans le groupes bleue sont dans la partie gauche du graphique des individus. En effet, on trouve "Greece", "Spain" et "France" du côté négatif de l'axe 1 et du coté positif de l'axe 2. On trouve "Turkey", "Italy" et "Belgium" du coté négatif de l'axe 1 et proche de l'origine de l'axe 2. Enfin, on trouve "Mexico" et "Luxembourg" du coté négatif de l'axe 1 e du coté négatif de l'axe 2.

Le groupe rouge contient les pays au centre du graphique et les pays sur la partie droite. Il contient donc les pays moyen ainsi que les pays ayant un coordonné positif sur l'axe 1. Par exemple on peut citer "Iceland".

Ainsi, la classification donne deux groupes distinct opposant les pays avec un taux de chômage élevé (groupe bleue) et les pays ayant un taux d'emploi et de participation élevé (groupe rouge).

En comparant aux informations obtenues par l'ACP, on remarque que la classification représente surtout l'axe 1 qui représente a lui seul plus de 50% de l'inertie totale, on perd l'information donnée par l'axe 2 concernant le taux de chômage des étrangers.

On remarque également que dans l'ACP et dans la classification, les informations sur le genre sont perdues.

Dans notre cas, il vaut donc mieux accompagné la classification avec une ACP pour conserver le maximum d'informations.