Université de Paris

Projet TND – SUJET 7

Table des matières

I. Analyse du jeu de données "pays"

A. Informations sur les variables

Le jeu de données "pays" met en avant les parités mais surtout les disparités entre les pays du jeu au travers de ces variables permettant d'observer leurs status économiques, éducatifs ou encore sanitaires. "pays" contient 8 variables :

"pays" : Donnée textuelle listant des noms de pays.

"esp vie F" : Variable quantitative continue. Nombre moyen d'années que vivrait une fille née en 2001 si la mortalité féminine par âge demeurait la même qu'en 2001.

"mort_inf": Variable quantitative continue. (Nombre d'enfants <1 an morts en 2001 / nombre d'enfants nés vivants en 2001).

"activF": Variable quantitative continue. (Nombre de femmes ayant un emploi / nombre de femmes d'âge actif).

"% chom.": Variable quantitative continue. (Nombre de chômeurs / nombre des actifs de plus de 15 ans)*100.

"pnb/hb": Variable quantitative continue. Produit national brut annuel par habitant (exprimé en $).

"% education": Variable quantitative continue. Dépenses d'éducation (de source publique ou privée) en % du Pnb.

"% santé": Variable quantitative continue. Dépenses de santé (de source publique ou privée) en % du Pnb.

Chacune de ces 8 variables contient 27 objets. Aucune valeur manquante ou non disponible (NA) n'est à déplorer.

B. Choix de la méthode d'analyse multidimensionnelle

Nous avons le choix entre des méthodes d'analyse fréquentielles telles que l'ACP (Analyse en Composantes Principales) ou des méthodes de classifications telles que la CAH (Classification hiérarchique ascendente). Le jeu de données "pays" est une matrice 8 * 27. En ce sens, l'ACP et la CAH peuvent convenir pour des données de tailles relativement petites. L'ACP traite des variables quantitatives; ici nous en avons 7 sur 8, donc 1 qualitative que nous pouvons mettre en tant que variable supplémentaire. La CAH traite tout type de données. Nous décidons donc d'appliquer ces deux méthodes.

II. Traitement du jeu de données "pays"

A. Étude des statistiques de base

```
pays = read.table("C:/Users/admin/Documents/Binome7/pays.txt", header = T, sep = "")
# À remplacer avec votre chemin vers le fichier pays.txt
# Nous avons modifié les noms pour rendre la récupération plus facile. (Annexe)
summary(pays)
```

```
    pays              esp_vie_F       mort_inf        activ_F         X._chom.
Length:27          Min.   :64.50   Min.   : 3.400   Min.   :22.90   Min.   : 2.400
Class :character   1st Qu.:72.25   1st Qu.: 4.550   1st Qu.:46.55   1st Qu.: 4.450
Mode  :character   Median :75.00   Median : 5.300   Median :50.90   Median : 7.400
                   Mean   :73.31   Mean   : 5.789   Mean   :50.65   Mean   : 7.652
                   3rd Qu.:75.50   3rd Qu.: 6.050   3rd Qu.:53.55   3rd Qu.: 9.450
                   Max.   :77.50   Max.   :10.400   Max.   :76.20   Max.   :18.100
    pnb.hb       X._education      X._santé
Min.   : 7809   Min.   :2.300   Min.   : 4.200
1st Qu.:12728   1st Qu.:4.650   1st Qu.: 6.250
Median :26756   Median :5.200   Median : 7.900
Mean   :22380   Mean   :5.356   Mean   : 7.541
3rd Qu.:28514   3rd Qu.:5.750   3rd Qu.: 8.450
Max.   :50410   Max.   :8.100   Max.   :10.700
```

Pour chacune des variables, nous avons un résumé des résultats des différents calculs tels que la moyenne ou la médiane. Ceci nous permet d'avoir un premier aperçu du contenu des variables.

B. Études des corrélations entre les variables

La corrélation est un indicateur qui nous permet d'évaluer l'évolution mutuelle entre deux valeurs, c-à-d d'observer si ces valeurs évoluent dans le même sens (augmentation parallèle) ou dans des sens contraires (l'une augmente, l'autre diminue et vice-versa).

```
corPays = cor(pays[c,(2:8)] # Nous excluons la variable qualitative "pays".
```

```
               esp_vie_F    mort_inf      activ_F   X._chom.        pnb.hb  X._education
esp_vie_F     1.000000000 -0.86208619 -0.003531218 -0.4116394  0.6511476638 -0.0628593657
mort_inf     -0.862086189  1.00000000 -0.176850877  0.3782614 -0.6166925865 -0.0862421524
activ_F      -0.003531218 -0.17685088  1.000000000 -0.2425572  0.2178180313  0.5999525892
X._chom.     -0.411639449  0.37826143 -0.242557163  1.0000000 -0.5932732219 -0.2757933732
pnb.hb        0.651147664 -0.61669259  0.217818031 -0.5932732  1.0000000000  0.0006702915
X._education -0.062859366 -0.08624215  0.599952589 -0.2757934  0.0006702915  1.0000000000
X._santé      0.701675316 -0.67458514 -0.009578663 -0.3407958  0.4231796846 -0.0867528687
                 X._santé
esp_vie_F     0.701675316
mort_inf     -0.674585139
activ_F      -0.009578663
X._chom.     -0.340795849
pnb.hb        0.423179685
X._education -0.086752869
X._santé      1.000000000
```

Nous pouvons constater que plusieurs variables sont corrélées. Ainsi par exemple, les variables "X_education" et "activ_F" sont corrélées postivement d'où nous pouvons déduire qu'investir dans l'éducation influence positivement le nombre de femme active.

De même, les variables "esp_vie_F" et "X_santé" sont corrélées positivement d'où nous pouvons également déduire qu'investir dans la santé augmente l'esp. de vie des femmes.

En ce qui concerne des exemples de corrélations négatives (donc évolution de sens inverse), nous pouvons observer que les variables "esp_vie_F" et "mort_inf" sont fortement corrélées négativement, et que donc, plus la mort infantile est basse, et plus l'espérance de vie d'une femme est haute.

C. Étude de l'ACP

L'ACP nous permet d'observer les données que nous avons et qui sont projetés sur des axes virtuels. Ces axes sont issus de combinaisons linaires entre les différentes variables du jeu de sorte à ce que la variance de toutes ces variables soit maximale, afin que l'un des axes n'ait pas d'influence sur l'autre. Ainsi nous représentons ces données dans un cercle des corrélations des variables et dans un plan factoriel des individus:

```
library(FactoMineR)
res.pca = PCA(pays, quali.sup = c(1))
# La variable qualitative "pays" est  en variable supplémentaire. Elle sert de référence.
plot(res.pca, choix = "var")  # Cercle des corrélations
```

Interprétation de l'ACP – Graphe des variables:

Dans un cercle des corrélations, nous nous intéressons qu'aux variables proches de la circonférence du cercle. Les flèches tracées permettent d'indiquer le sens d'évolution d'une variable par rapport aux axes virtuels. En abscisse, nous considérerons l' "Axe 1" et en ordonnée, nous considérerons l' "Axe 2".

Nous constatons notamment que:

- Les variables "X_education" et "activ_F" sont corrélées positivement par rapport à l'Axe 1. Les deux augmentent ou diminuent ensemble.

- Les variables "pnb.hb", "esp_vie_F et X_santé" sont corrélées positivement par rapport à l'Axe 1. Les trois augmentent ou diminuent ensemble.

- Les variables "mort_inf" et "X_chom." sont corrélées postivement par rapport à l'axe par rapport à l'Axe 1. Les deux augmentent ou diminuent ensemble.

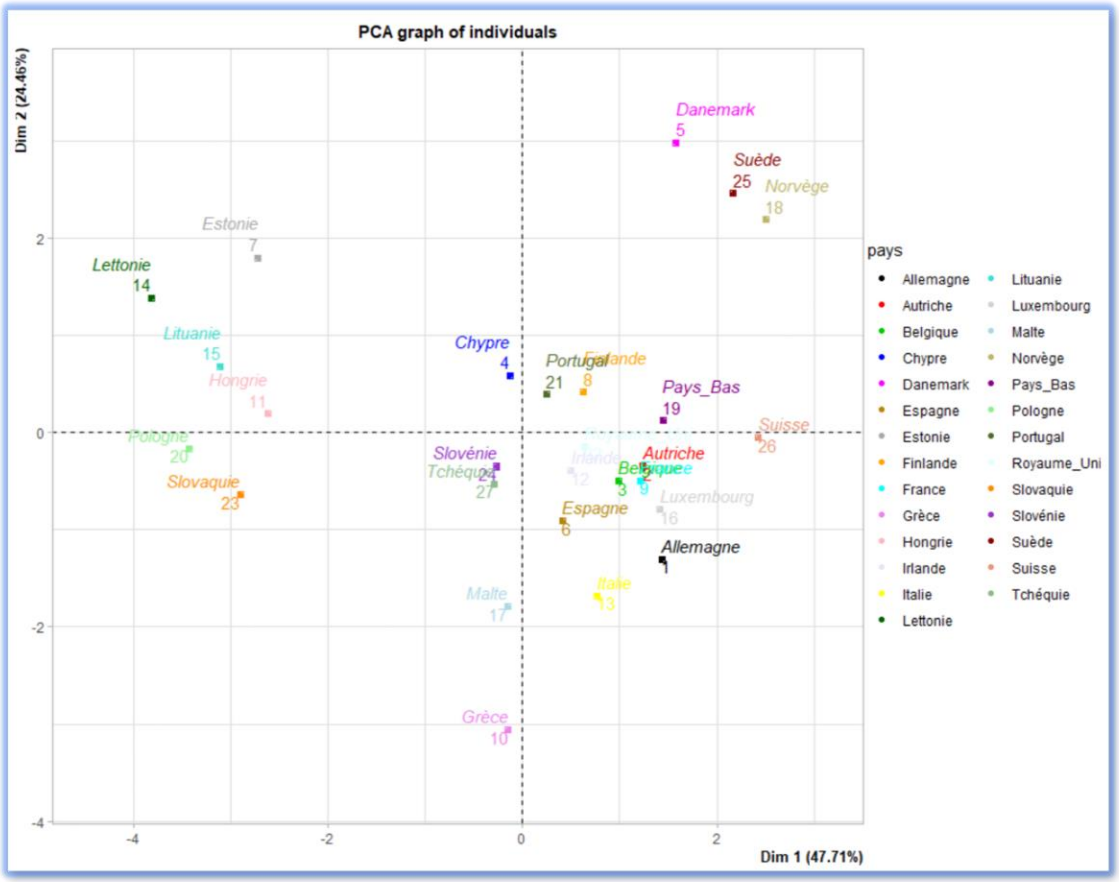
- Les variables "X_santé" et "esp_vie_F" sont corrélées négativement avec "mort_inf" par rapport à l'Axe 2. Lorsque les deux premières variables augmentent, la troisième diminue et vice-versa.

- Les variables "mort_inf" et "X_chom" sont corrélées négativement avec "pnb.hb" par rapport à l'Axe 1. Lorsque les deux premières variables augmentent ou diminuent ensemble, la troisième diminue et vice-versa.

Nous pouvons donc en déduire que plus le PNB est élevé, et plus les investissements en santé et en éducation sont élevés et donc, plus le nombre de femmes actives et l'espérance de vie des femmes évoluent.

Lorsque ces 5 variables augmentent positivement et conjointement alors en parallèle, les morts infantiles et le nombre de chômeurs diminuent. Ainsi à l'inverse, plus les taux de chômages et de morts infantiles sont élevés, et moins les 5 variables précédentes le sont.

```
plot(res.pca, choix = "ind", habillage = 1)
# Plan factoriel en fonction de la variable "pays".
```



Interprétation de l'ACP – Graphe des individus:

Nous constatons notamment que:

- Des pays tels que la Slovaquie ou la Pologne, étant les plus proches des positions des variables "mort_inf" et "X_chom.", en ont une valeur élevée.

- Des pays tels que le Danemark, la Suède ou la Norvège, étant les plus proches des positions des variables "X_education" et "activ_F", en ont une valeur élevée.

- Des pays tels que la Suisse ou l'Autriche, étant les plus proches des positions des variables "pnb.hb", "esp_vie_F et X_santé", en ont une valeur élevée.

Nous pouvons donc en déduire que plus des pays sont proches de la position d'une variable, et plus la valeur affectée à cette variable pour ces pays sera élevée, et vice-versa.

Ainsi nous pouvons distinguer deux catégories: des pays où "il ferait mieux bon vivre" et d'autres, "moins bon vivre". La première catégorie rassemble les pays à droite de l'Axe 2 et la seconde catégorie rassemble les pays à gauche de l'Axe 2. Plus la position d'un pays est éloignée du centre du cercle (et donc plus elle est proche de la circonférence) et plus son appartenance à l'une des deux catégories est prononcée.

D. Étude de la CAH

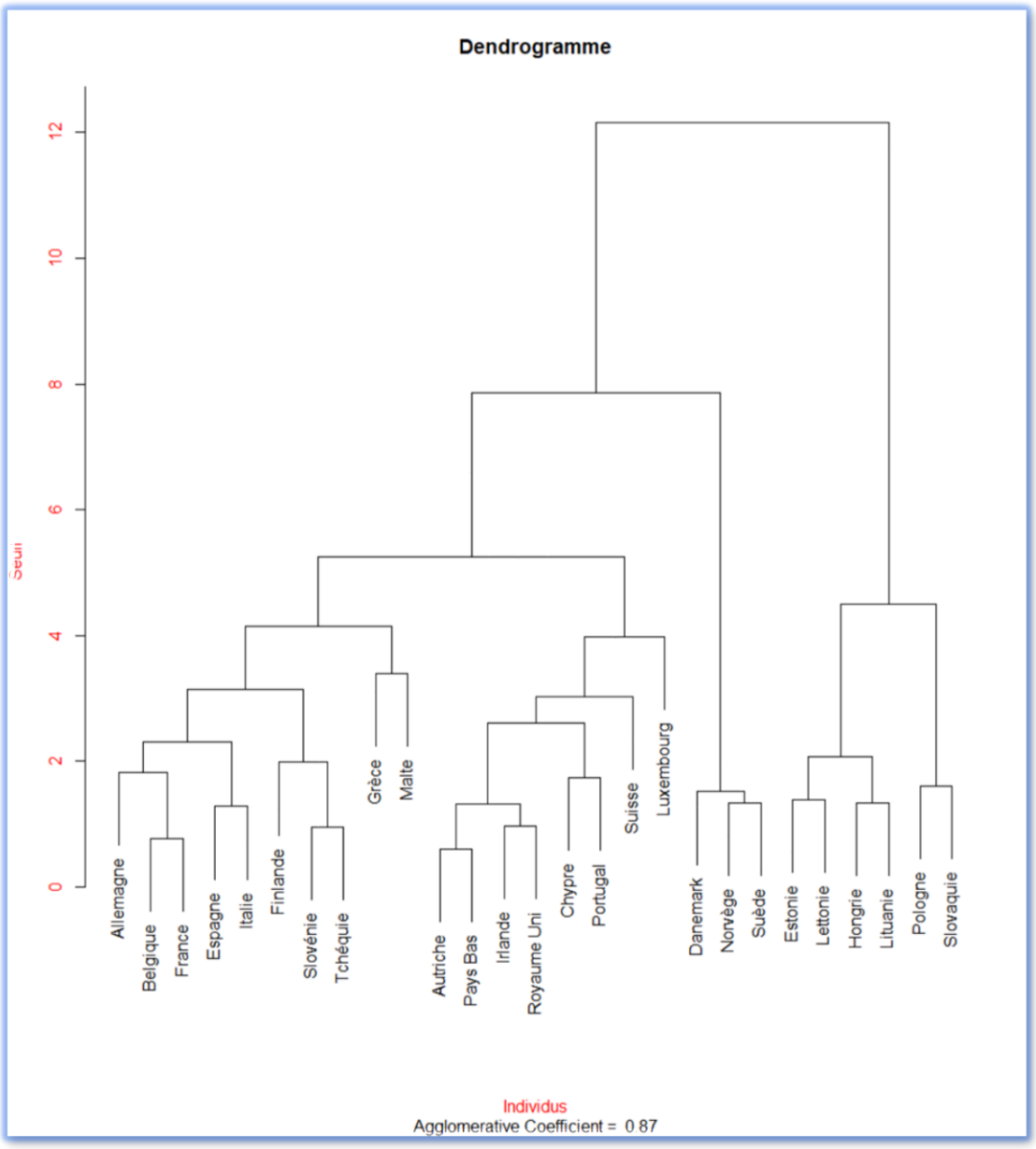
La classification hiérarchique ascendante permet de regrouper itérativement des objets partageant des critères de similarités suivant un critère d'agrégation.

Ici, nous avons choisi comme critère d'agrégation la méthode de Ward qui figure parmi les plus utilisés et les plus efficaces car nos données ne sont pas étirées et ne varient pas dans des valeurs extrêmes. Les objets sont donc regroupés deux par deux et ainsi de suite, jusqu'à ce qu'il n'y ait plus qu'un seul groupe contenant tous les objets.

La classification hiérarchique ascendante nous permet donc d'avoir une meilleure vue d'ensemble sur les ressemblances et les différentes entre les données d'un jeu de données.

```
rownames(pays) = c(
                   "Allemagne","Autriche","Belgique","Chypre","Danemark",
                   "Espagne","Estonie","Finlande","France","Grèce","Hongrie",
                   "Irlande,"Italie","Lettonie","Lituanie","Luxembourg","Malte",
                   "Norvège", "Pays Bas","Pologne","Portugal","Royaume
                   Uni","Slovaquie", "Slovénie","Suède","Suisse","Tchéquie"
                   )
```

```
library(cluster)
cah=agnes(scale(pays[, 2:8]), method="ward")
plot(cah, xlab="Individus", ylab="Seuil", col.axis = "red", col.lab = "red", main="Dendrogramme")
```



Nous constatons, d'une part, les objets qui partagent le plus de critères de similarités tels que les objets "Belgique" et "France" ou encore les objets "Pologne" et "Slovaquie".

D'autre part, nous pouvons distinguer 4 groupes (clusters) aux seuils ~ 4.1, ~ 3.9, ~ 5.9 et ~ 5. Ainsi, chacun des objets contenu dans l'un des clusters ont une variance intra-cluster minimale (ressemblance avec les autres objets du cluster) pour une variance inter-cluster maximale (différence avec les objets des autres clusters).

Voici l'appartenance simplifiée des objets à l'un des 4 clusters:

```
classification = as.hclust(cah)
plot(rev(classification$height), type = "h", ylab = "hauteurs")
classes = cutree(cah, k=4); classes
pays.classes = cbind.data.frame(pays, as.factor(classes)); pays.classes
```

	as.factor(classes)
Allemagne	1
Autriche	2
Belgique	1
Chypre	2
Danemark	3
Espagne	1
Estonie	4
Finlande	1
France	1
Grèce	1
Hongrie	4
Irlande	2
Italie	1
Lettonie	4
Lituanie	4
Luxembourg	2
Malte	1
Norvège	3
Pays Bas	2
Pologne	4
Portugal	2
Royaume Uni	2
Slovaquie	4
Slovénie	1
Suède	3
Suisse	2
Tchéquie	1

Nous avons donc les groupes suivants:

1: Allemagne, Belgique, Espagne, Finlande, France, Grèce, Italie, Malte, Slovénie, Tchéquie.
2: Autriche, Chypre, Irlande, Luxembourg, Pays Bas, Portugal, Royaume Uni, Suisse.
3: Danemark, Norvège, Suède.
4: Estonie, Hongrie, Lettonie, Lituanie, Pologne, Slovaquie.

Conformément aux observations précédemments faites, nous pourrions labéliser ces clusters comme suit:

1: Rassemble des pays aux qualités "Moyennes à élevées".
2: Rassemble des pays aux qualités "Élevées".
3: Rassemble des pays aux qualités "Très élevées".
4: Rassemble des pays aux qualités "Médiocres à moyennes".

III. Conclusion sur le jeu de données "pays"

Pour conclure, le jeu de données "pays" contient 8 variables et 27 objets dont 1 variable de type "données textuelles" qui liste les noms des pays.

Les différentes études (dont notamment l'ACP et la CAH) menées au travers du projet ont pu mettre en évidence des corrélations positives ou négatives entre ces variables, ce qui nous permet de mieux comprendre les influences de certains phénomènes sur d'autres phénomènes. De même, les représentations graphiques telles que le cercle des corrélations et le plan factoriel de ces phénomènes nous permettent également d'avoir une meilleure vue d'ensemble, et, encore une fois, d'en dégager des informations quant à leur influence mutuelle ou non.

Finalement, ceci nous permet de regrouper ces observations en différents groupes afin de mieux analyser les ressemblances et les différences que constituent un jeu de données. Ceci nous permet de conclure que l'objectif de ce jeu de données est de comparer les prestations et la qualité de vie proposées par les pays étudiés.

IV. Annexe

Code source des programmes utilisés pour ce projet + contenu du fichier pays.txt

pays esp_vie_F mort_inf activ_F %_chom. pnb/hb %_education %_santé
Allemagne 74.8 4.4 48.8 8.2 26768 4.3 10.6
Autriche 75.4 4.8 49 4.1 29075 4.9 8
Belgique 75.1 5 42.3 7.3 27952 5.8 8.8
Chypre 75.3 5.6 50.9 3.8 12724 5.8 6
Danemark 74.5 5.3 73.8 4.5 30096 8.1 8.4
Espagne 75.6 3.9 40.3 11.4 22538 5.6 7.7
Estonie 64.9 8.4 52.2 6.8 10201 6.8 5.5
Finlande 74.6 3.8 56.8 9.1 27215 5.6 6.6
France 75.5 4.6 47.8 8.7 27560 5.6 9.4
Grèce 75.4 6.1 37.6 10.3 17670 2.3 9.2
Hongrie 68.4 9.2 45.6 8.4 12733 5.2 5.7
Irlande 73 5.9 47.5 4.4 32549 4.5 7.2
Italie 76.7 4.5 36 9.1 26946 4.6 8
Lettonie 64.5 10.4 49.7 8.5 7809 6.2 4.8
Lituanie 65.9 8.6 54.6 10.9 8359 5.2 5.7
Luxembourg 75.2 5.8 42.5 2.4 50410 4 6.1
Malte 76.2 6 22.9 7.4 9875 5.7 8.9
Norvège 76.2 3.8 69.2 3.9 37070 7.4 8.5
Pays_Bas 75.5 5.1 52.9 2.7 29614 5.2 8.2
Pologne 70.3 8.1 49.5 18.1 9852 5.1 4.2
Portugal 72.4 5.5 54.1 5 18500 5.5 8.2
Royaume_Uni 75 5.6 53 5.1 26756 4.7 7.3
Slovaquie 69.7 8.6 52.9 17.4 12314 4.3 6.4
Slovénie 72.3 4.9 51.3 11.3 17762 5.2 8.2
Suède 77.5 3.4 76.2 4.9 26849 7.3 7.9
Suisse 77.2 4.9 58.8 3.1 30058 5.1 10.7

Tchéquie 72.2 4.1 51.3 9.8 15011 4.6 7.4

A. Code des statistiques de base

```
pays = read.table("C:/Users/admin/Documents/Binome7/pays.txt", header = T, sep = "")
# À remplacer avec votre chemin vers le fichier pays.txt
summary(pays)
```

B. Code des corrélations entre les variables

```
corPays = cor(pays[c,(2:8)] # Nous excluons la variable qualitative "pays".
```

C. Code de l'ACP

```
library(FactoMineR)
res.pca = PCA(pays, quali.sup = c(1))
# La variable qualitative "pays" est à mettre en variable supplémentaire. Elle sert de référence.
plot(res.pca, choix = "var")
# Cercle des corrélations
plot(res.pca, choix = "ind", habillage = 1)
# Plan factoriel en fonction de la variable "pays".
```

D. Code de la CAH

```
rownames(pays) = c(
                    "Allemagne","Autriche","Belgique","Chypre","Danemark",
                    "Espagne","Estonie","Finlande","France","Grèce","Hongrie",
                    "Irlande,"Italie","Lettonie","Lituanie","Luxembourg","Malte",
                    "Norvège", "Pays Bas","Pologne","Portugal","Royaume
                    Uni","Slovaquie", "Slovénie","Suède","Suisse","Tchéquie"
                    )

library(cluster)
cah=agnes(scale(pays[, 2:8]), method="ward")
plot(cah, xlab="Individus", ylab="Seuil", col.axis = "red", col.lab = "red", main="Dendrogramme")

classification = as.hclust(cah)
plot(rev(classification$height), type = "h", ylab = "hauteurs")
classes = cutree(cah, k=4); classes
pays.classes = cbind.data.frame(pays, as.factor(classes)); pays.classes
```