Ecole d'Economie Toulouse-TSE et
Institut de Mathématiques de Toulouse
Deuxième année du Diplôme du Magistère
d'Economiste Statisticien
Etude de cas (numéro 1)

Analyse d'une métrique pour évaluer le bien être mentale des habitants

Année universitaire 2022-2023

Pierre Luc Alibert
Bastien CANNARD
Raphael PINAULT

Mots clés : ACP, World Hapiness Report, AFC, K-means

1 Cadre de travail

Nous allons présenter les conditions dans lesquelles le travail s'est effectué. Cela regroupe le cadre de travail, les outils employés, les sources d'information disponible, la problématique et la manière dont nous avons réparti le travail.

1.1 Condition et motivations du choix du sujet

Nous avons cherché à trouver une métrique qui permettrait d'évaluer le bien-être psychique de la population. Cette métrique pourrait être utilisée comme indicateur dans les décisions du gouvernement. Lors des différents cours et lecture des articles nous avons pu remarquer que le bien-être mental n'était pas toujours pris en compte dans les prises de décisions. Nous voulons voir les différentes solutions disponibles dans la littérature. Le "Worl happiness Report" (WHR par la suite) est un rapport annuel qui cherche à répondre à la même problématique. En reprenant leur base de données disponible en ligne, nous souhaitons appliquer des méthodes d'analyse de données plus approfondies. Ainsi, nous pouvons voir cet exercice comme une étude complémentaire au rapport "Worl happiness Report" de 2021.
Les motivations collectives de travailler sur ce sujet peuvent donc être résumé comme suit:

1. Tout d'abord, l'envie de traveiller sur ce sujet a été guidé par les sujets d'actualité d'aujourd'hui et les crises bouleversant le monde dans lequel on vit. Nous pensons qu'il est important d'essayer de comprendre comment un peuple réagit de manière générale aux enjeux de crise qu'il traverse et comment en sort-il plus fort?

2. Puis, le souhait de travailler sur ce sujet a été renforcé quand nous nous sommes aperçus que l'on pouvait utiliser beaucoup de méthodes de clustering et d'analyses statistiques sur la base de données que l'on avait trouvé. La base en elle-même nous a convaincu de travailler sur ce sujet par sa qualité de mise en oeuvre ainsi que par le nombre suffisant d'observations (observations au sens de pays). En effet, il y a beaucoup de données sur beaucoup d'années et sur des pays très différents du monde,renseignées dans cette base de données.C'est ainsi pour cela que le sujet a convenu à tout le monde et que nous avons travaillé sur cette problématique.

La base de données a été fournie par Gallup, un partenaire du WHR. La partie expliquant le sens des variables et l'histoire du WHR est basée des explications données dans le rapport du World Happiness Report de 2022 ; nous pouvons retrouver des similitudes. Enfin, la partie expliquant

le procédé pour sélectionner les ménages pour établir le sondage reprend des informations du site de Gallup, de ce fait, nous pouvons également retrouver les similitudes. Nous pouvons considérer ces parties comme une traduction de la partie présente dans le rapport. Aucune autre partie de cette étude n'a été publié dans un autre document. Nous déclarons sur l'honneur que ce document, n'a pas été présenté, pour une autre EDC ou mémoire. Notre rapport apporte des résultats d'analyses avec des outils d'exploration plus sophistiqués tel que l'analyse par composante principale ou bien le clustering avec la méthode k-means.

1.2 *World happiness report*

L'assemblée générale des nations unies a déclaré en 2011 qu'il était nécessaire d'apporter une plus grande importance au bien-être et au bonheur lorsqu'on abordait les questions sociales et économiques. Suite à cette annonce, en 2012, le premier rapport du "Worl happiness Report" (WHR par la suite) a vu le jour. WHR est devenu un rapport annuel. Ce dernier cherche à mesurer le niveau de bonheur de chaque population avec les derniers outils de statistiques et d'apprentissage automatisée. Ce rapport est le résultat d'une collaboration entre trois grands institues : le "Earth Institute" à l'université de Columbia, le centre de recherche en performance économique au "London School of Economics" (LSE) et l'institue canadien des recherches avancées (CIFAR).

1.3 Plan de travail

La recherche de la base de données et de la problématique a été effectuée en équipe lors des réunions audiovisuelles. Nous avons distribué les tâches de la manière suivante :

- Raphael s'est occupé des analyses par composante principale sur chaque année ;
- Pierre-Luc s'est occupé de la partie sur le clustering et l'afc ;
- Bastien s'est occupé des autres parties.

Nous étions tous les trois dans des villes différentes et avec des contraintes horaires différentes. Cela a considérablement impacté notre manière de travailler. Nous avons fixé la problématique et les tâches lors des rendez-vous audiovisuelle que nous avons eu. Par la suite, nous avons avancé chacun de son côté. Chacun partageait leurs résultats lorsqu'il avait avancé dans sa partie. Les horaires de stage ont été très contraignants pour l'avancement du projet.

1.4 Outils de travail

Les données sont issues du WHR. Cette base de données est une compilation de plusieurs bases de données, mais dont la plus grande partie provient du "Gallup World Poll".

Le sondage est réalisé dans 160 pays chaque année. Le nombre de personnes interrogé varie entre 500 et 2000 personnes pour les plus grands pays tels que la Russie ou la Chine. Chaque année, les mêmes, 100 questions sont posées de la même manière dans le même ordre. L'entretien se fait par téléphone si la proportion de la population possédant un téléphone est supérieure à 80 %. Dans ce cas, un algorithme de sélection aléatoire est utilisé pour sélectionner les candidats. Dans les pays en développement, les entretiens se font principalement en personne.

Dans un premier temps Gallup partitionne l'ensemble des ménages selon la géographie et le nombre d'habitant des villes. Les ménages sont sélectionnés de manière aléatoire suivant une probabilité proportionnelle au nombre d'habitants. La demande d'interview peut être relancée jusqu'à trois fois à différentes périodes de la journée.

Enfin, un poids peut être appliqué dans les réponses pour faire correspondre leur échantillon à la démographique nationale. [1]

Nous avons utilisé le logiciel R studio pour tous les traitements de la base de données ainsi que pour effectuer les statistiques descriptives. L'ACP a été faite à partir du package "FactoMineR" et "factoextra" disponible sur R studio. Nous avons consulté le rappor du WHR pour obtenir toutes les informations concernant la base de données.

1.5 Les difficultés et comment on les a surmonté

Nous avons voulu trouvé une variable qui puisse indiquer plus en détail l'implication de chaque pays dans la crise mondiale de 2008. Ainsi, une comparaison plus fine aurait pu être effectuée entre le bonheur des habitants des pays pas trop impactée par la crise et ceux vivant dans un pays fortement impacté. Nous avons trouvé dans la littérature les variables qui auraient pu être de bons indicateurs pour mesurer l'impact de la crise mondiale dans le pays telle que le *bond spread*, ou bien la variation du taux de change. Nous n'avons pas réussi à trouver une base de données fournissant cette information pour tous les pays de l'échantillon.

Certaines valeurs n'étaient pas présentes dans la base de données. Nous avons fait le choix de supprimer les observations avec ces valeurs manquantes. Nous nous sommes assurés qu'après-avoir fait le tri, les pays restants soient une bonne représentation des pays dans le monde.

[1] La description de la procédure est issue du site *https://www.gallup.com*

Contents

2 Intoduction

Nous cherchons à trouver une métrique permettant d'évaluer l'état psychique des différentes populations. Ainsi, à travers cette métrique, nous pourrons essayer d'évaluer l'impact des chocs politiques et économiques sur l'état psychique de la population. Dans ce rapport, nous proposons d'utiliser des statistiques descriptives simples ainsi que quelques méthodes habituelles d'analyse de données telles que l'analyse par composantes principales ou bien la méthode de clustering k-means. Pour cela, nous utilisons une base de données issue du World Happinesse Report. Ces données de panel apportent une idée sur l'état d'esprit et le niveau de bien-être psychique général des habitants dans tous les pays du monde sur les vingt dernières années. Une idée générale du bien-être psychique peut être obtenue à travers les neuf variables présentes dans la base de données dont nous allons présenter dans la partie suivante. Pour des raisons de simplifications, nous nous restreignons aux données des années 2007, 2008 et 2009. Nous avons sélectionné l'année 2008 car, il y a eu une crise économique et financière à l'échelle planétaire. Ensuite, nous sélectionnons l'année antécédente et précédente pour pouvoir faire des comparaisons et déterminer les évolutions. Cela nous permettra de voir l'impact que la crise a eu sur le psychisme des gens.

3 Présentation des données

Nous allons présenter les différentes variables présentes dans la base de données qui correspondent aux questions posées lors de l'interview. Les réponses correspondent à la moyenne des réponses données sur trois années : l'année en cours lors de l'interview et les deux précédentes années [2].

- Les participants au questionnaire doivent noter leur **qualité de vie** sur une échelle de 0 à 10 ; 0 correspondant à la moins bonne qualité possible. Cette note est présentée dans la variable "qualité de le vie". Cette note est connue sous le nom de l'échelle de Cantril.
- La variable **affect positive** représente la proportion des personnes interrogées qui ont connu un moment de rigolade, de joie ou qui ont acquis une nouvelle connaissance intéressante ces derniers jours.
- La variable **affect négative** représente la proportion des personnes interrogées qui ont connu un moment de tristesse, d'anxiété ou bien de colère durant ces derniers jours.
- La variable **perspective de l'entre aide sociale** représente la proportion des personnes interrogées qui pense avoir un proche qui puisse les aider en cas de soucis.
- La variable **perspective de libre-arbitre** représente la proportion des personnes satisfait de leur liberté de choix dans la vie.
- La variable **générosité** correspond au résidu de la régression de l'indicatrice indiquant si la personne a fait un don récemment sur le logarithme du PIB par habitant.
- La variable **perception de la corruption dans le gouvernement** correspond à la proportion de personne qui pense que la corruption est une pratique courante dans le gouvernement de leur pays.
- La variable indiquant l'espérance de vie à la naissance est issue des bases de données de l'organisation mondiale de la santé.
- Le PIB par habitant est mesuré en parité de pouvoir d'achat et ajusté en dollars constants de 2017.

[2] Toutes ces descriptions sont celles données dans le World Happiness Report de 2022

4 Statistiques descriptives

Analysons nos variables quantitatives et qualitatives à l'aide de statistiques descriptives (on s'intéresse uniquement à décrire et à résumer les données étudiées, on ne fait pas de test d'hypothèses ni d'estimations). Pour cela, nous allons utiliser le logiciel R, logiciel libre destiné à la manipulation de données et aux constructions de graphiques.
La base de données est ici triée sur 3 années : 2007,2008 et 2009, correspondant aux années nous intéressant puisque l'on veut étudier l'impact de la crise de 2008 dans notre base de données.Nous avons choisi les pays avec des données présentes sur les trois années. Nous avons 219 observations soit 219 pays en tout mais ramené à l'année, cela fait 73 pays différents. L'échantillon de pays est plutôt hétérogène, nous avons des pays développés, en voie de développement mais également des pays pauvres: on peut donc dire que notre échantillon est représentatif et ne faussera pas les analyses.

4.1 Année 2007: l'année avant la crise

Ci-dessous la répartition des pays selon leurs régions respectives. Cette répartition est donc prise sur l'année 2007 mais c'est exactement la même sur les années 2008 et 2009 puisque l'on a toujours les mêmes pays représentés.

```
Central and Eastern Europe Commonwealth of Independent States                East Asia
                         4                                  10                        3
Latin America and Caribbean                 Middle East and North Africa North America and ANZ
                        17                                   6                        2
                South Asia                      Southeast Asia           Sub-Saharan Africa
                         5                                   7                       13
            Western Europe
                         6
```

Figure 1: Répartition des pays selon leurs régions respectives

On peut ainsi voir que c'est l'Amérique du Sud (modalité **Latin America and Caribbean**) qui est le continent le plus représenté parmi les pays présents dans notre base de données.

A présent, toujours sur l'année d'avant crise, voici le tableau résumant les principales statistiques descriptives de nos variables quantitatives.
On constate que pour la variable **Life Ladder**, le maximum est de **7.834**, ce qui correspond à la note de qualité de vie attribuée dans le pays du Danemark. Parmi tous les pays étudiés, c'est le Danemark avec la note de qualité de vie la plus élevée sur l'année d'avant crise.

Notons également la présence de valeurs non renseignées, manquantes (valeurs **NA's**). Nous avons décidé de remplacer ces valeurs par la médiane de chaque variable, jugeant que la médiane était l'outil le plus représentatif des valeurs prises par chaque variable. Nous avons fait de même pour les autres années.

```
 Life.ladder       log.GDP.per.capita Social.support   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices  Generosity
Min.   :3.280   Min.   : 6.881     Min.   :0.4790   Min.   :42.86                    Min.   :0.2950                Min.   :-0.284000
1st Qu.:4.698   1st Qu.: 8.364     1st Qu.:0.7285   1st Qu.:59.12                    1st Qu.:0.6188                1st Qu.:-0.095000
Median :5.252   Median : 9.181     Median :0.8360   Median :63.50                    Median :0.6840                Median :-0.013000
Mean   :5.449   Mean   : 9.190     Mean   :0.8054   Mean   :61.73                    Mean   :0.6894                Mean   : 0.003616
3rd Qu.:6.138   3rd Qu.: 9.903     3rd Qu.:0.8770   3rd Qu.:66.47                    3rd Qu.:0.7973                3rd Qu.: 0.132000
Max.   :7.834   Max.   :11.212     Max.   :0.9700   Max.   :73.90                    Max.   :0.9320                Max.   : 0.391000
                                   NA's   :2        NA's   :1                        NA's   :1
Perceptions.of.corruption
Min.   :0.0640
1st Qu.:0.7495
Median :0.8230
Mean   :0.7890
3rd Qu.:0.8885
Max.   :0.9680
NA's   :3
```

Figure 2: Statistiques descriptives des variables quantitatives sur l'année 2007

Enfin, nous pouvons voir ci-dessous la boîte à moustaches de notre variable **Life Ladder** correspondant à une note de la qualité de vie.

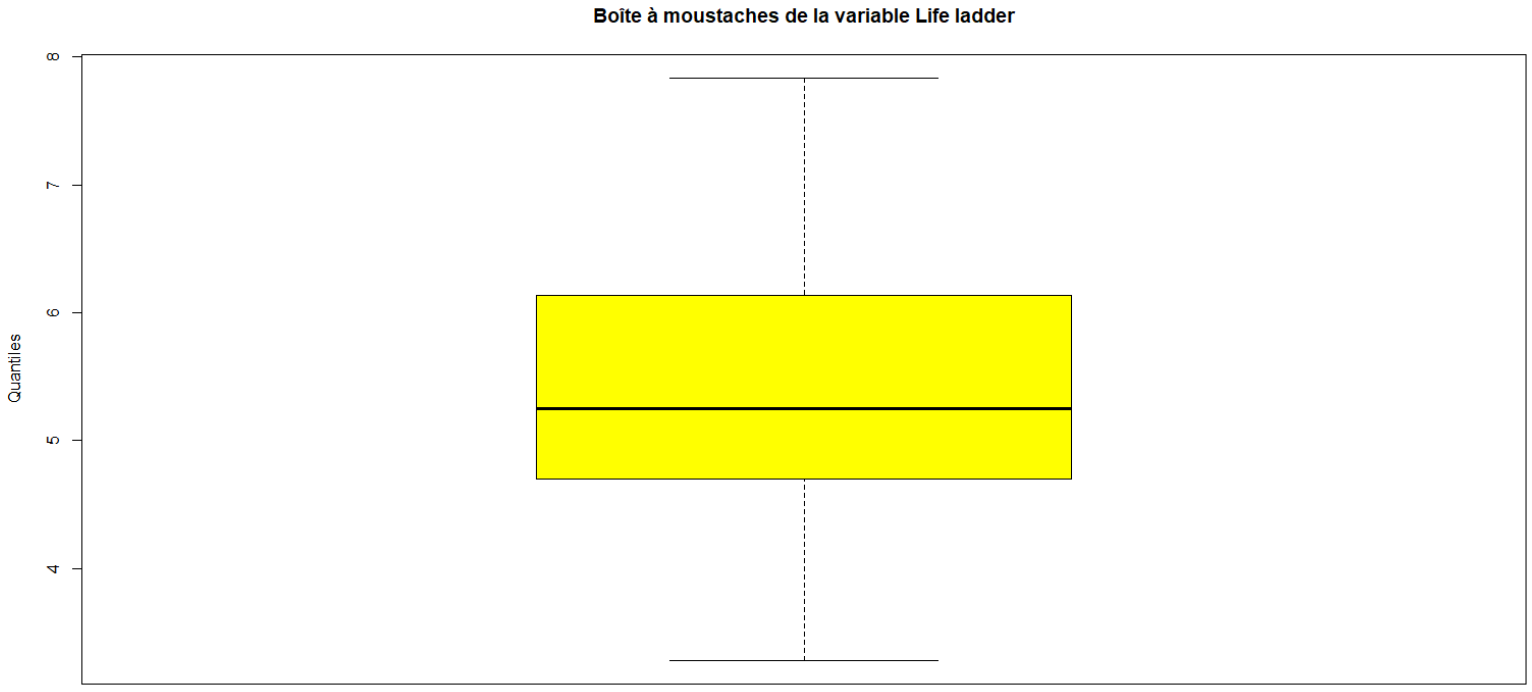


Figure 3: Boîte à moustaches de la variable Life Ladder

Nous pouvons dire que sur l'année 2007, la boîte à moustaches de la variable Life ladder est symétrique, les valeurs de notes de qualité de vie se répartissent de la même façon de part et d'autre de la valeur **5.252** (la valeur de la médiane étant représentée par une barre dans la

boîte à moustaches).On peut dire également que 50% des pays ont des notes comprises entre 4.75 et 6 pour la variable Life ladder.

4.2 Année 2008: l'année de la crise

À présent, étudions les statistiques descriptives basiques de nos variables sur l'année de la crise des subprimes, l'année 2008. Ci-dessous le tableau des différentes statistiques pour nos variables quantitatives.

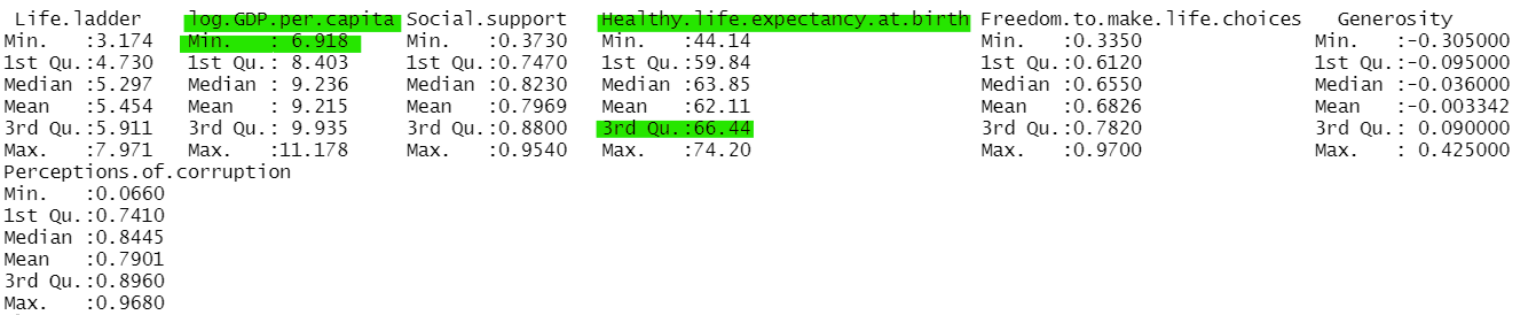


Figure 4: Statistiques descriptives des variables quantitatives sur l'année 2008

Pour la variable **Log GDP per capita**, le minimum est de **6.918**. Ce logarithme du PIB par habitant-là concerne le Niger qui donc parmi tous les pays étudiés est celui qui a le plus faible logarithme du PIB par habitant et donc le plus faible PIB par habitant.
La variable **Healthy life expectancy at birth** a une valeur du troisième quartile égale à **66.44** : 25% des pays de notre base de données ont une espérance de vie à la naissance supérieure à 66 ans. Cela renforce ici le fait que notre échantillon de pays ne contient pas que des pays développés.

À présent, ci-dessous l'histogramme de la variable **Life ladder**, correspondant aux notes de qualité de vie de l'année 2008 des 73 pays étudiés.

Figure 5: Histogramme de la variable Life ladder

Comme titre à cet histogramme, on pourrait donner : **Répartition des pays selon leur note de qualité de vie**. Par ailleurs l'histogramme met en évidence une classe quadri modale, il s'agit de [4;6] : la majorité des pays ont note de qualité de vie comprise entre 4 et 6.

Enfin, analysons la corrélation entre les variables **Life ladder** et **Log GDP per capita** sur l'année 2008. Le coefficient de corrélation est égal à **0.79**.Nous avons donc une relation linéaire positive entre ces deux variables. La valeur de ce coefficient permet de dire que la relation semble être **significative** entre les variables Life ladder et Log GDP per capita.

Le nuage de points ci-dessous vient confirmer le résultat du coefficient de corrélation entre ces deux variables.

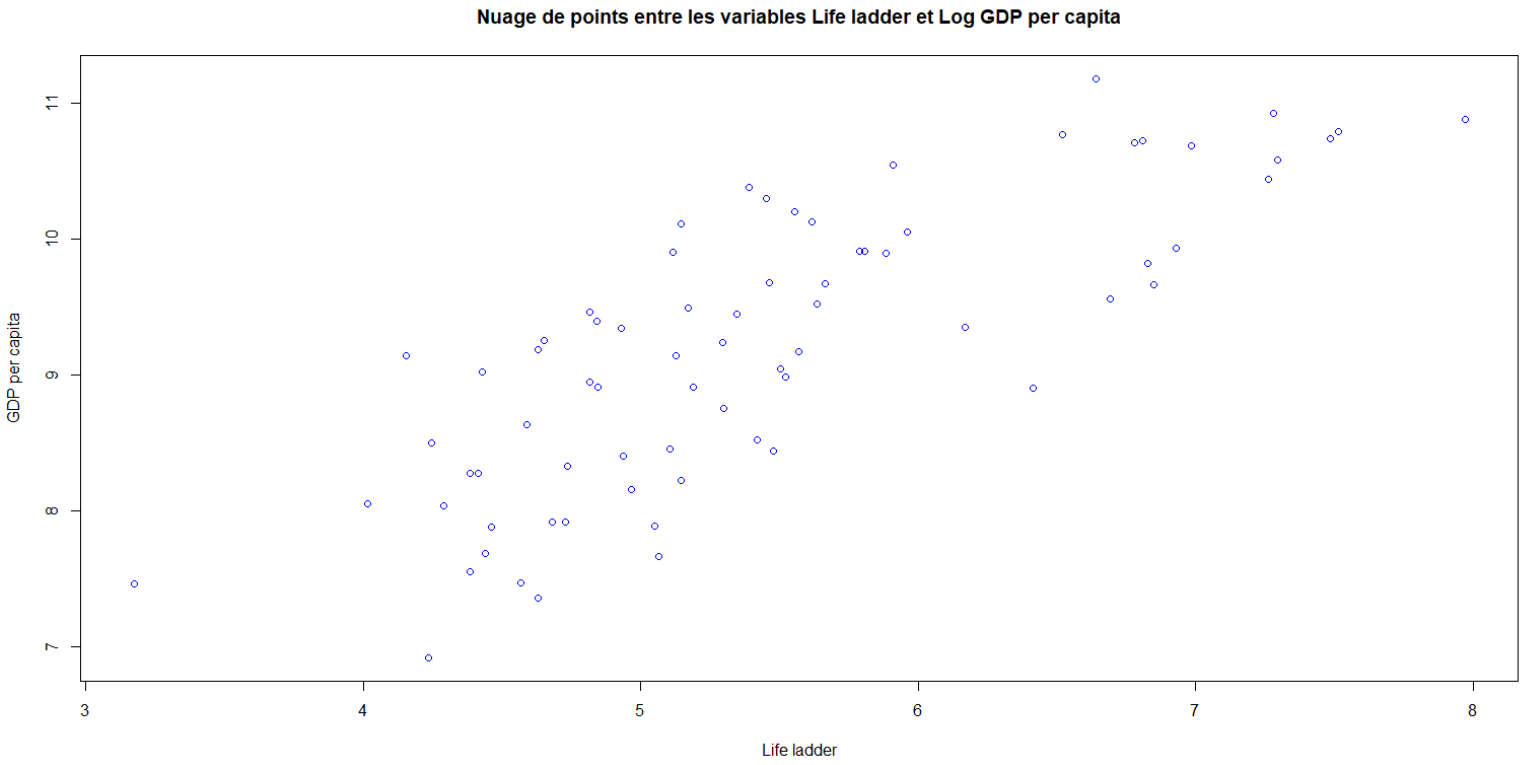


Figure 6: Nuage de points des variables Life ladder et Log GDP per capita

4.3 Année 2009: l'année d'après crise

Maintenant, intéressons nous à l'année d'après crise : l'année 2009 et regardons si les effets sont déjà observables sur les statistiques de cette année-là.

Ci-dessous le tableau de statistiques descriptives de nos variables quantitatives sur l'année d'après crise.

Life.ladder	log.GDP.per.capita	Social.support	Healthy.life.expectancy.at.birth	Freedom.to.make.life.choices	Generosity	Perceptions.of.corruption
Min. :3.408	Min. : 6.898	Min. :0.5220	Min. :44.62	Min. :0.388	Min. :-0.303000	Min. :0.035
1st Qu.:4.669	1st Qu.: 8.407	1st Qu.:0.7790	1st Qu.:59.94	1st Qu.:0.607	1st Qu.:-0.108000	1st Qu.:0.754
Median :5.385	Median : 9.213	Median :0.8320	Median :64.19	Median :0.701	Median :-0.037000	Median :0.839
Mean :5.502	Mean : 9.199	Mean :0.8214	Mean :62.46	Mean :0.689	Mean :-0.008836	Mean :0.787
3rd Qu.:6.199	3rd Qu.: 9.930	3rd Qu.:0.8990	3rd Qu.:66.62	3rd Qu.:0.779	3rd Qu.: 0.077000	3rd Qu.:0.904
Max. :7.683	Max. :11.149	Max. :0.9640	Max. :74.50	Max. :0.949	Max. : 0.525000	Max. :0.979

Figure 7: Statistiques descriptives des variables quantitatives sur l'année 2009

On peut ainsi voir que les écarts dans les statistiques sont très faibles entre les années 2008 et 2009 pour nos variables. On constate une infime baisse de la médiane de la variable **Log GDP per capita** passant de **9.236** en 2008 à **9.213** en 2009. Autrement dit, l'année d'après la crise de 2008, dans notre échantillon de pays, au moins 50% des pays ont un logarithme PIB par

habitant inférieur ou égal à **9.213** tandis qu'au moins 50% des pays ont un logarithme PIB par habitant supérieur ou égal à **9.213**.

Nous pouvons également voir ci-dessous la boîte à moustaches de la variable **Life Ladder** correspondant toujours à la note de la qualité de vie mais maintenant sur l'année d'après crise.

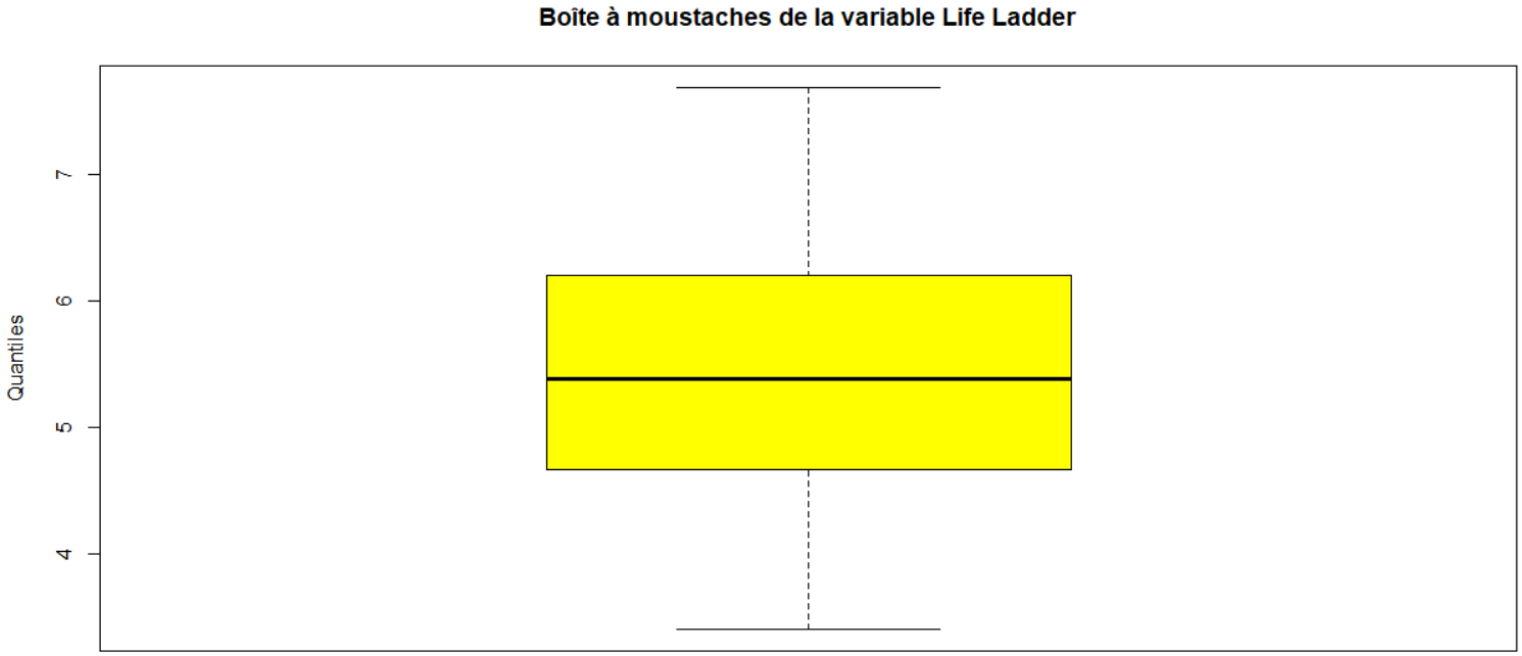


Figure 8: Boîte à moustaches de la variable Life Ladder

On constate également que la boîte à moustaches de cette variable est très similaire à celle de l'année 2007 (l'année d'avant crise). On peut remarquer, même si la médiane est bien centrée, une très légère asymétrie vers le haut : les notes de qualité de vie des observations se répartissent quasiment de la même façon de part et d'autre de la valeur **5.385** avec une plus forte dispersion des 50% des plus grandes valeurs. On peut aussi constater que la plupart des pays ont une note de qualité de vie comprise entre 4.75 et 6.25 (voir bord inférieur et supérieur du rectangle).

Enfin, nous avons jugé intéressant d'utiliser les résultats des statistiques descriptives de nos variables quantitatives selon les régions(du monde) étudiées.
Voici ci-dessous un exemple de tableau récapitulant les statistiques descriptives selon deux régions du monde: l'Amérique du Nord (représenté par **North America and ANZ**) et l'Afrique du Nord et Moyen-Orient (représenté par **Middle East and North Africa**).

```
df_2009[, "Region"]: Middle East and North Africa
  Life.ladder     log.GDP.per.capita Social.support   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices   Generosity
 Min.   :4.470   Min.   : 8.329     Min.   :0.7380   Min.   :59.94                    Min.   :0.4560                Min.   :-0.22700
 1st Qu.:5.103   1st Qu.: 9.246     1st Qu.:0.7468   1st Qu.:62.61                    1st Qu.:0.4993                1st Qu.:-0.10750
 Median :5.606   Median : 9.592     Median :0.8270   Median :64.13                    Median :0.6020                Median :-0.09250
 Mean   :5.708   Mean   : 9.637     Mean   :0.8323   Mean   :64.48                    Mean   :0.5897                Mean   :-0.07083
 3rd Qu.:6.111   3rd Qu.:10.278     3rd Qu.:0.9155   3rd Qu.:64.40                    3rd Qu.:0.6320                3rd Qu.:-0.07750
 Max.   :7.353   Max.   :10.673     Max.   :0.9370   Max.   :72.08                    Max.   :0.7710                Max.   : 0.17200
 Perceptions.of.corruption
 Min.   :0.4450
 1st Qu.:0.7535
 Median :0.7990
 Mean   :0.7597
 3rd Qu.:0.8400
 Max.   :0.9230
------------------------------------------------------------------------------------------------------------
df_2009[, "Region"]: North America and ANZ
  Life.ladder     log.GDP.per.capita Social.support   Healthy.life.expectancy.at.birth Freedom.to.make.life.choices   Generosity
 Min.   :7.158   Min.   :10.70      Min.   :0.9120   Min.   :68.54                    Min.   :0.831                 Min.   :0.2010
 1st Qu.:7.240   1st Qu.:10.74      1st Qu.:0.9197   1st Qu.:69.41                    1st Qu.:0.852                 1st Qu.:0.2122
 Median :7.323   Median :10.79      Median :0.9275   Median :70.28                    Median :0.873                 Median :0.2235
 Mean   :7.323   Mean   :10.79      Mean   :0.9275   Mean   :70.28                    Mean   :0.873                 Mean   :0.2235
 3rd Qu.:7.405   3rd Qu.:10.84      3rd Qu.:0.9353   3rd Qu.:71.15                    3rd Qu.:0.894                 3rd Qu.:0.2347
 Max.   :7.488   Max.   :10.89      Max.   :0.9430   Max.   :72.02                    Max.   :0.915                 Max.   :0.2460
 Perceptions.of.corruption
 Min.   :0.413
 1st Qu.:0.476
 Median :0.539
 Mean   :0.539
 3rd Qu.:0.602
 Max.   :0.665
```

Figure 9: Exemple de statistiques descriptives selon deux régions étudiées

Nous pouvons voir quelques différences statistiques selon ces deux régions. En effet, la valeur de la médiane de la variable **Life ladder** est plus élevée en Amérique du Nord qu'en Afrique du Nord et Moyen-Orient. C'est la même chose pour la variable **Log GDP per capita**, en prenant comme exemple la moyenne. On voit que la moyenne du logarithme PIB par habitant est plus élevée (valeur de **10.79**) en Amérique du Nord qu'en Afrique du Nord et Moyen-Orient (valeur égale à **9.637**).

On peut ainsi donc conclure que dans notre base de données, au niveau des statistiques descriptives "classiques" réalisées sur nos variables, il n'y a pas de véritable changement entre les années d'avant, pendant et d'après crise de 2008. De plus, les indicateurs(moyenne, médiane...) des variables ne semblent pas s'inverser après la crise de 2008 pour les régions. La crise de 2008 étant partie des États-Unis, on aurait pu s'attendre à ce qu'en 2009, certaines régions passent devant les États-Unis (en termes de valeurs de certaines variables comme **Log GDP per capita** ou **Life ladder** par exemple). Cela n'est pas le cas, ces derniers gardant des statistiques supérieures à la plupart des autres régions du monde.

5 ACP

Dans cette partie, nous cherchons à comprendre comment les variables issues de cette base de données interagissent entre elles. Pour cela, nous appliquons une analyse par composante principales (ACP). A partir d'une base de données initiale, nous construisons de nouvelles variables, appelées composante principale, sous forme de combinaison linéaire des variables initiales. Chaque composante principale est construite de sorte à récupérer le plus d'information possible. L'information de la base de données est mesurée sous la forme de l'inertie. Les composantes principales sont orthogonales entre elles. Nous effectuons une ACP pour chacune des trois années sélectionnées pour l'étude : 2007, 2008 et 2009. Il est intéressant d'observer comment les composantes auront été modifiées au cours de ces trois ans. Ces analyses permettront de mettre en relief le sens et la direction des différentes corrélations entre les variables, mais ne permettent pas d'établir les liens causaux.

Pour les trois années, les trois premières composantes suffisent pour récupérer au moins 70 % de l'inertie de la base de données initiale. Dans la suite, nous allons essayer d'interpréter ces composantes ainsi que de voir comment elles évoluent à travers ces trois ans.

5.1 Année 2007

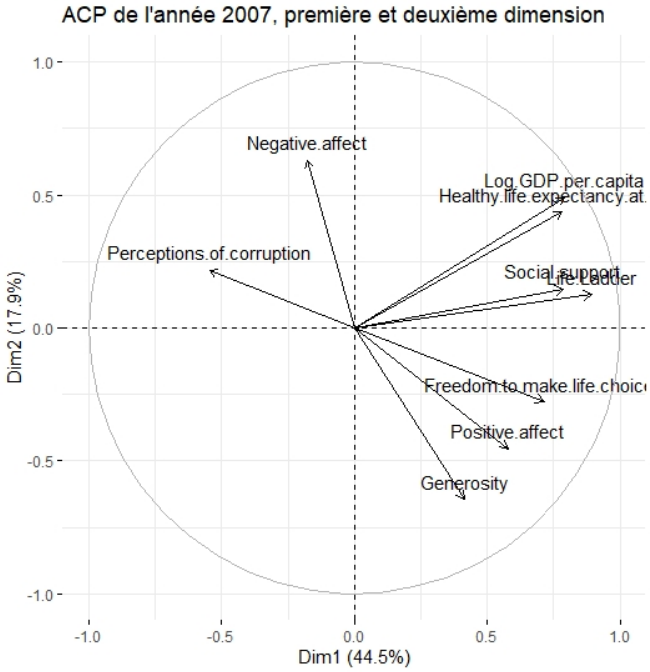
Dans la première composante, nous retrouvons les variables suivantes corrélées positivement : la note de la qualité de la vie (20 %) [3], le logarithme du PIB par habitant (16 %), l'entre aide sociale (15 %), l'espérance de vie à la naissance (15 %), la perception du libre-arbitre (13 %) et l'affect positif (8 %). Celles-ci sont opposées à la variable indiquant la perception de la corruption (8 %). Ainsi, dans cette base de données, nous pouvons dire que les habitants avec une vision positive de la vie et vivant dans un milieu offrant un plus grand niveau de soin médicaux et psychique ont une plus grande confiance dans le gouvernement.

Dans la seconde composante, nous retrouvons les variables indiquant le logarithme du PIB par habitant (15 %), le niveau d'affect négative (25 %) et l'espérance de vie à la naissance (12 %) opposées aux variables indiquant le niveau de générosité (26 %) et d'affect positive (13 %). Les habitants des pays avec un PIB par habitant et un niveau de soins médicaux et un niveau d'affect négatif élevé sont caractérisés par un faible niveau de générosité.

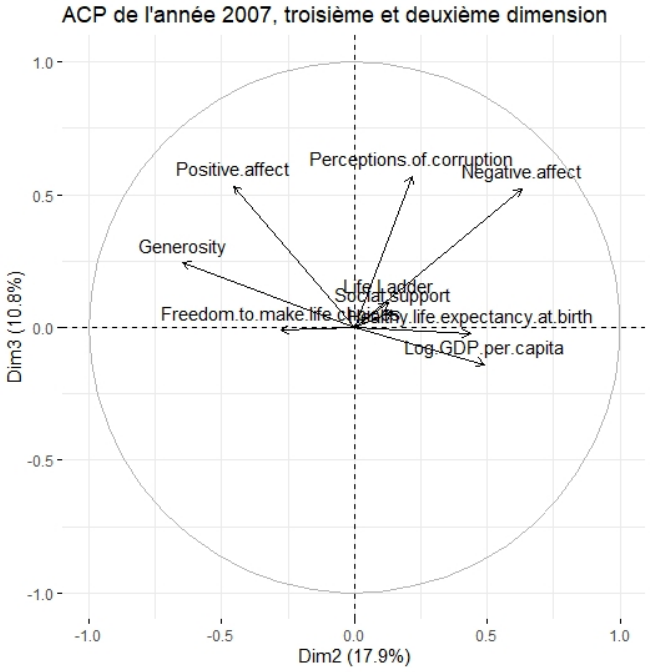
Enfin, dans la troisième composante, nous retrouvons les variables indiquant l'affect négatif (28 %) et positif (29 %) corrélées positivement avec la variable évaluant le niveau de perception

[3] Nous mettons entre parenthèses la part de contribution de la variable dans la création de la composante

de la corruption (33 %).



(a) Première et deuxième composante principale pour l'année 2007



(b) Troisième et deuxième composante principale pour l'année 2007

Figure 10: ACP de l'année 2007 sur 63 pays

5.2 Année 2008

Dans la suite, nous cherchons à comprendre comment la place des variables au sein des composantes a évolué lors de la crise. Dans la première composante, nous retrouvons les mêmes variables. Cependant, les niveaux de contribution ont varié : la note de la qualité de la vie, le logarithme du PIB par habitant ont augmenté de 1 %. Inversement l'entre aide sociale et la perception du libre-arbitre a diminué de 1 %. La contribution de l'espérance de vie à la naissance dans la création de cette composante a augmenté de 2 %. L'affect positif et la variable indiquant la perception de la corruption n'ont pas connu d'évolution.

La deuxième composante a vu une baisse de 2% de la contribution des variables indiquant le logarithme du PIB par habitant, et l'espérance de vie à la naissance. Ainsi qu'une diminution de 6% de la contribution de la variable indiquant l'impression du libre-arbitre. La contribution du niveau d'affect négative a baissé de 7 %. Le niveau de contribution de l'affect positive et le niveau de générosité a augmenté de 5 %.

La troisième composante est entièrement différente de ce dont nous avons vu en 2007. Les variables de la générosité (19 %) et de l'entre aide sociale (14 %) sont opposées à la variable de l'affect négative (47 %). Au cours de cette année, certains pays ont été marqué par une augmentation de l'affect négative accompagnée par une diminution de la générosité et de l'entre aide sociale.

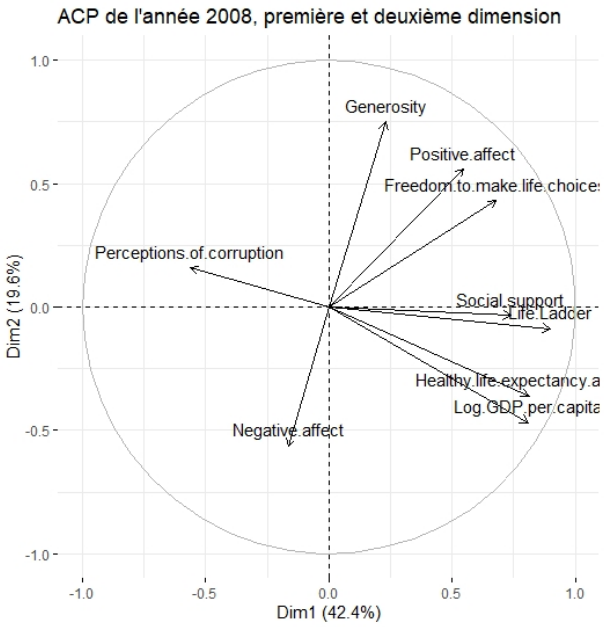


Figure 11: ACP de l'année 2008 sur 63 pays

5.3 Année 2009

Durant cette année, nous retrouvons une composition de la première composante similaire aux deux précédentes années. Dans la seconde composante, seule la variation de la contribution de l'affect positive et négative et de la perception de libre-arbitre est notable. L'affect négatif a baissé de 9 %. L'affect positif a augmenté de 5 %. Alors que la contribution de la variable indiquant la sensation du libre-arbitre a augmenté de 7 %. La composition de la dernière composante sélectionnée est sensiblement similaire à celle de 2007.

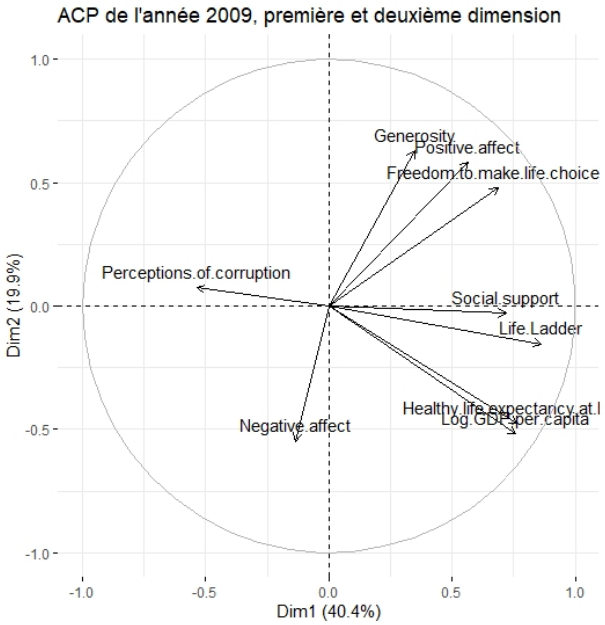


Figure 12: ACP de l'année 2009 sur 63 pays

6 Clustering

Dans la perspective de différences culturelles entre les régions du monde, et en particulier de différences d'appréciation d'une crise telle que celle de 2008, on se propose d'effectuer du clustering pour chacune des trois années, 2007, 2008 et 2009.

Nous utiliserons la méhode des k-Means.

D'abord un point sur les régions auxquelles appartiennent chacun des 73 pays de notre étude, d'après les données fournies par le *World Happiness Report*.

La région "Latin America and Caribbean" comprend les 17 pays suivants : "Argentina", "Bolivia", "Brazil", "Chile", "Colombia", "Costa Rica", "Dominican Republic", "Ecuador", "El Salvador", "Guatemala", "Honduras", "Mexico", "Nicaragua", "Panama", "Paraguay", "Peru", "Uruguay".

La région "Commonwealth of Independent States" comprend les 10 pays suivants : "Armenia", "Azerbaijan", "Belarus", "Georgia", "Kazakhstan", "Kyrgyzstan", "Moldova", "Russia", "Tajikistan", "Ukraine".

La région "South Asia" comprend les 5 pays suivants : "Bangladesh", "India", "Nepal", "Pakistan", "Sri Lanka".

La région "Southeast Asia" comprend les 7 pays suivants : "Cambodia", "Indonesia", "Malaysia", "Philippines", "Singapore", "Thailand", "Vietnam".

La région "Sub-Saharan Africa" comprend les 13 pays suivants : "Cameroon", "Chad", "Ghana", "Kenya", "Mauritania", "Niger", "Nigeria", "Senegal", "South Africa", "Tanzania", "Uganda", "Zambia", "Zimbabwe".

La région "North America and ANZ" comprend les 2 pays suivants : "Canada" et "United States".

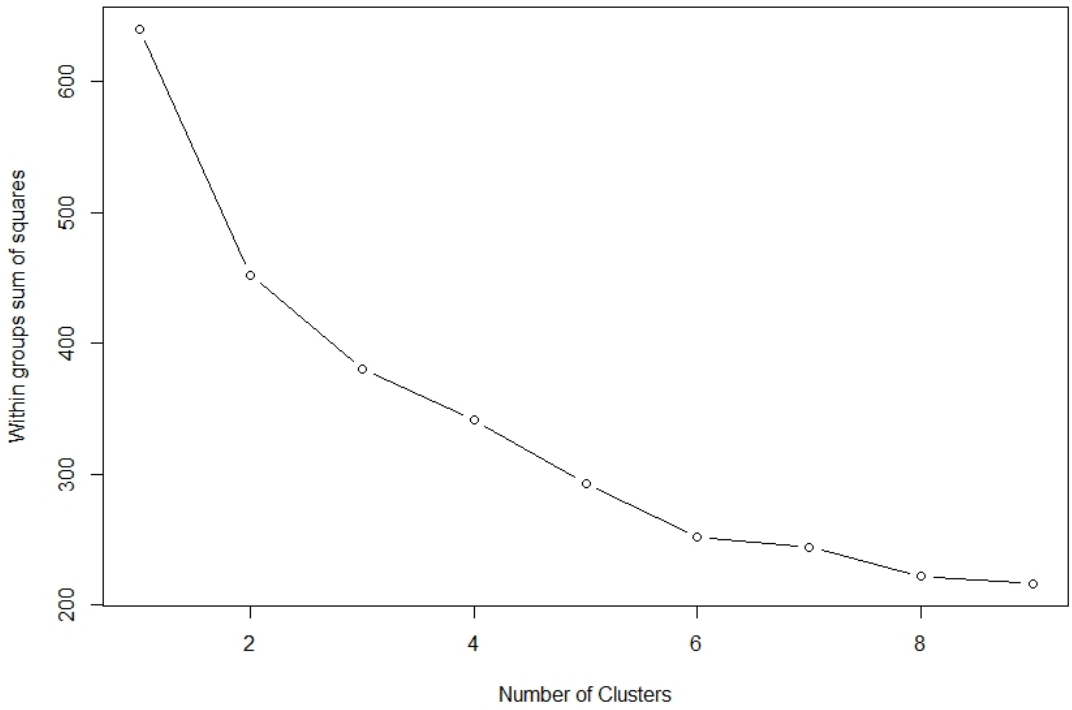
La région "East Asia" comprend les 3 pays suivants : "China", "Japan", "South Korea".

La région "Western Europe" comprend les 6 pays suivants : "Denmark", "Germany", "Italy", "Spain", "Sweden", "United Kingdom".

La région "Middle East and North Africa" comprend les 6 pays suivants : "Egypt", "Israel", "Jordan", "Palestinian Territories", "Saudi Arabia", "Turkey".

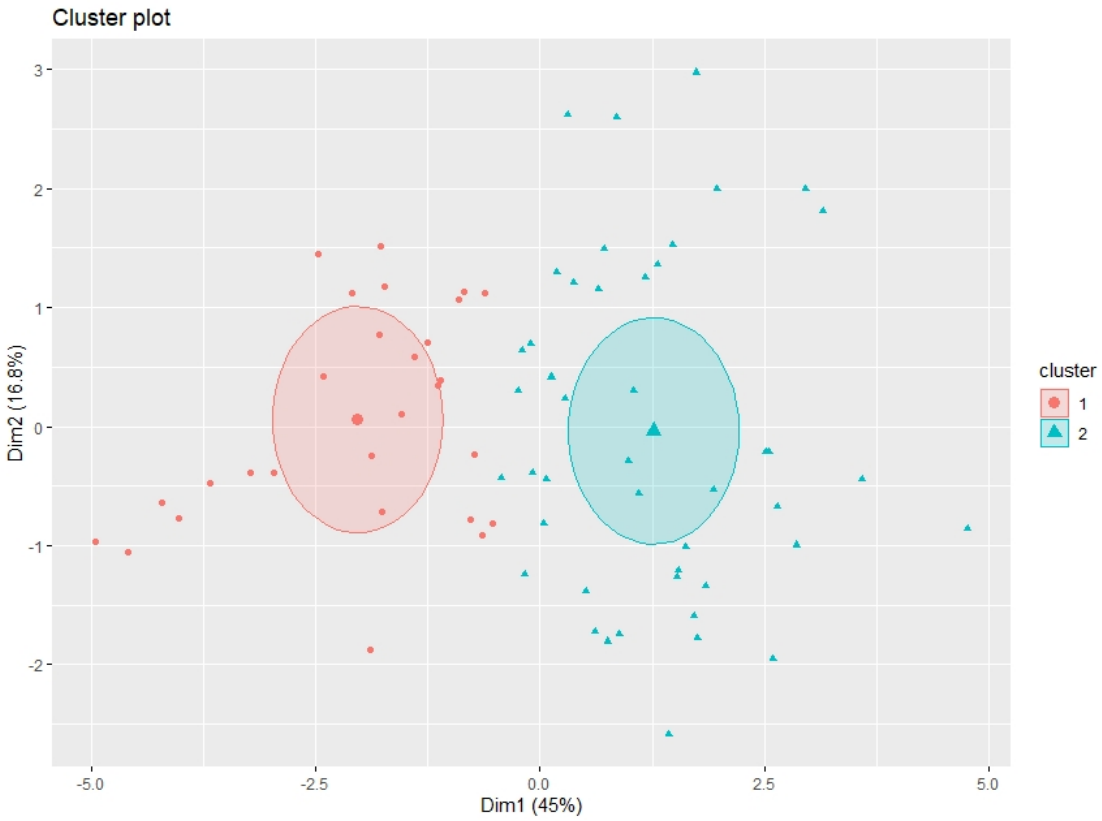
La région "Central and Eastern Europe" comprend les 4 pays suivants : "Estonia", "Kosovo", "Latvia", "Lithuania".

6.1 Année 2007



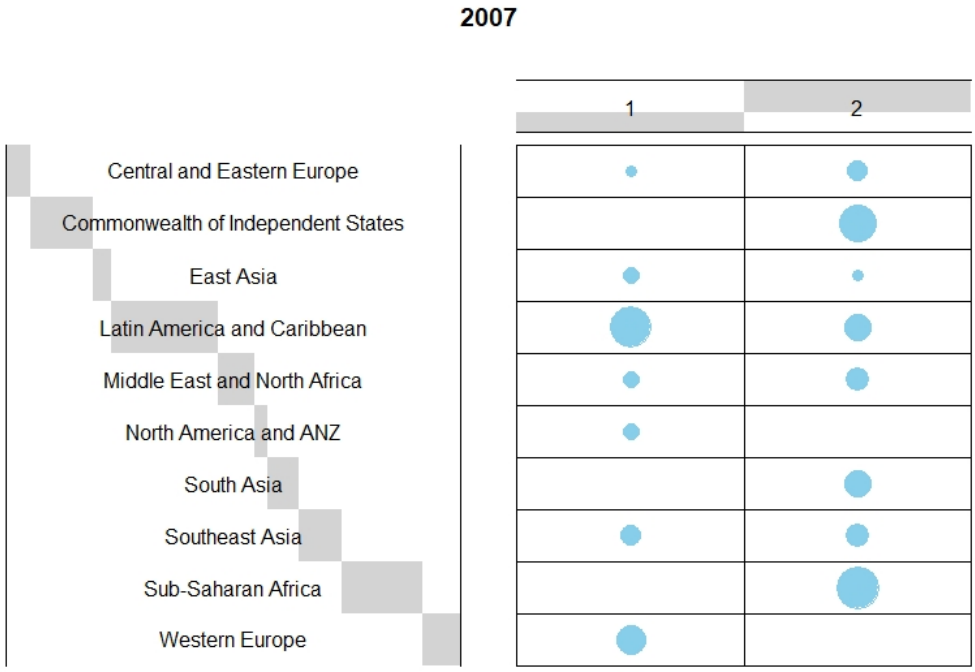
D'après la courbe des variances intra-clusters ainsi que la règle "du coude", on retient deux clusters, bien que cela ne soit pas totalement clair. On obtient des clusters assez équilibrés : respectivement 28 et 45 observations. Le premier cluster est associé à des valeurs positives pour toutes les variables hormis Perceptions of corruption et Negative affect, et inversement pour le deuxième cluster.

Voici les deux clusters dans le plan des deux principales composantes :



Les deux clusters semblent séparés en fonction de la valeur des observations sur le premier axe.

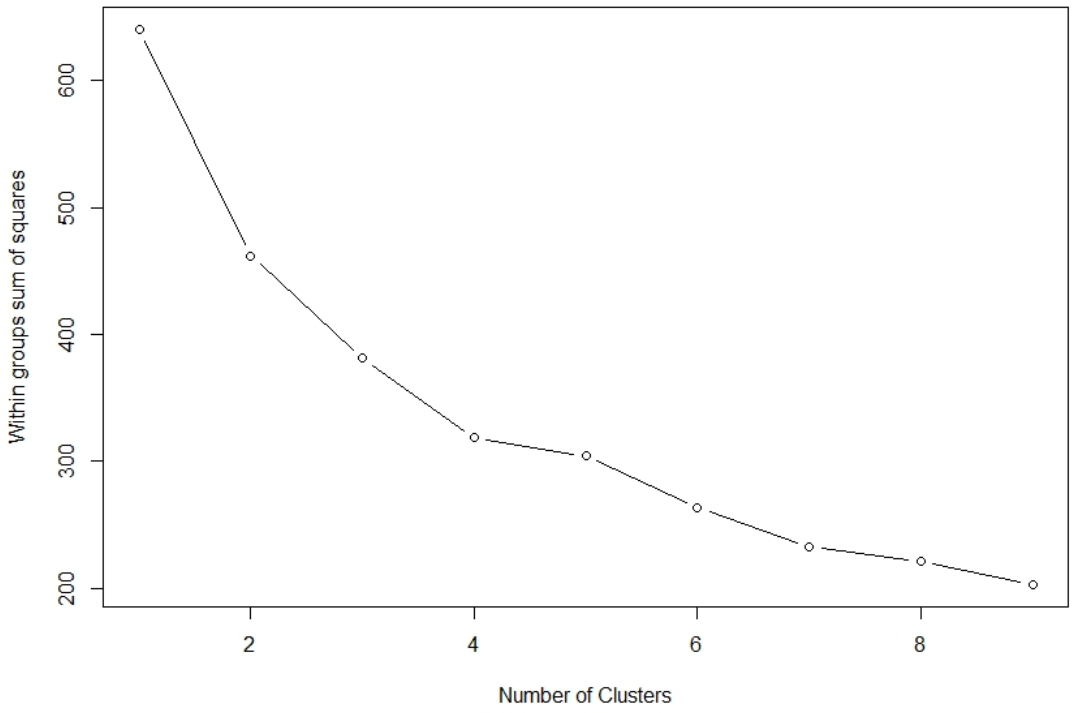
Regardons la distribution des régions en fonction du cluster :



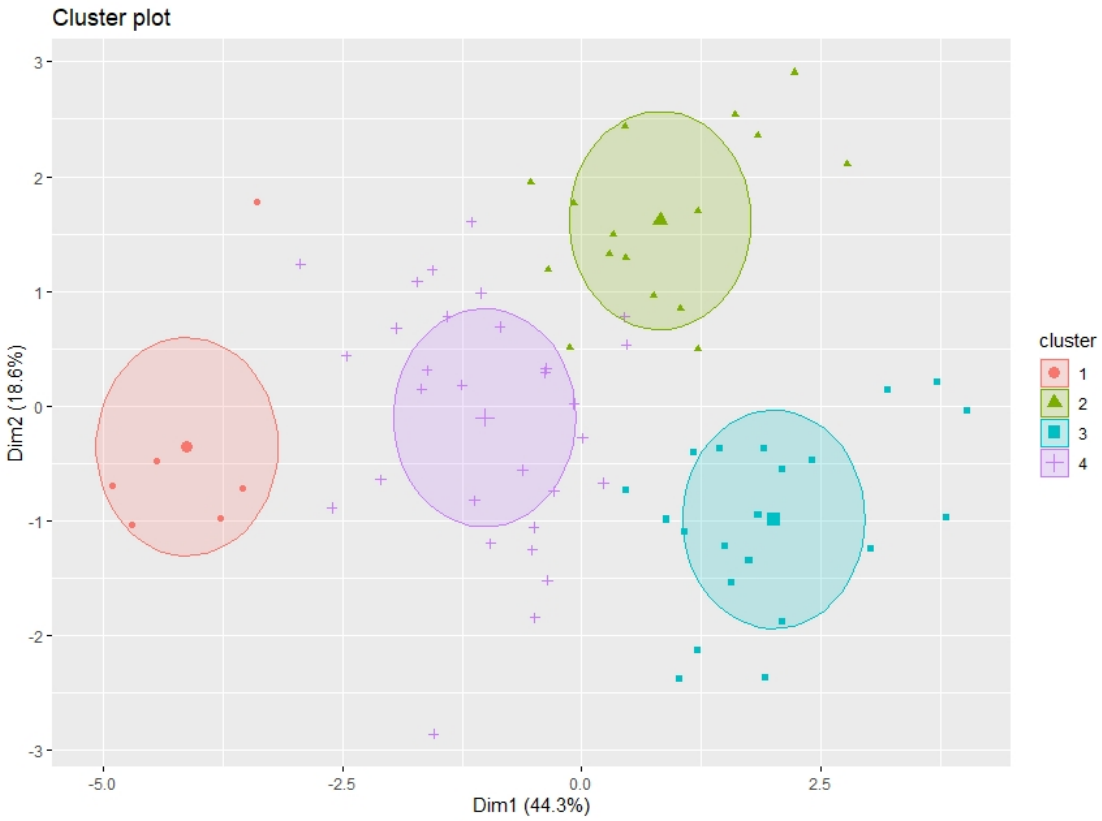
Certaines régions sont totalement liées à un cluster : North America and ANZ et Western Europe avec le cluster 1 ; Commonwealth of Independent States, South Asia et Sub-Saharan Africa avec le cluster 2.

6.2 Année 2008

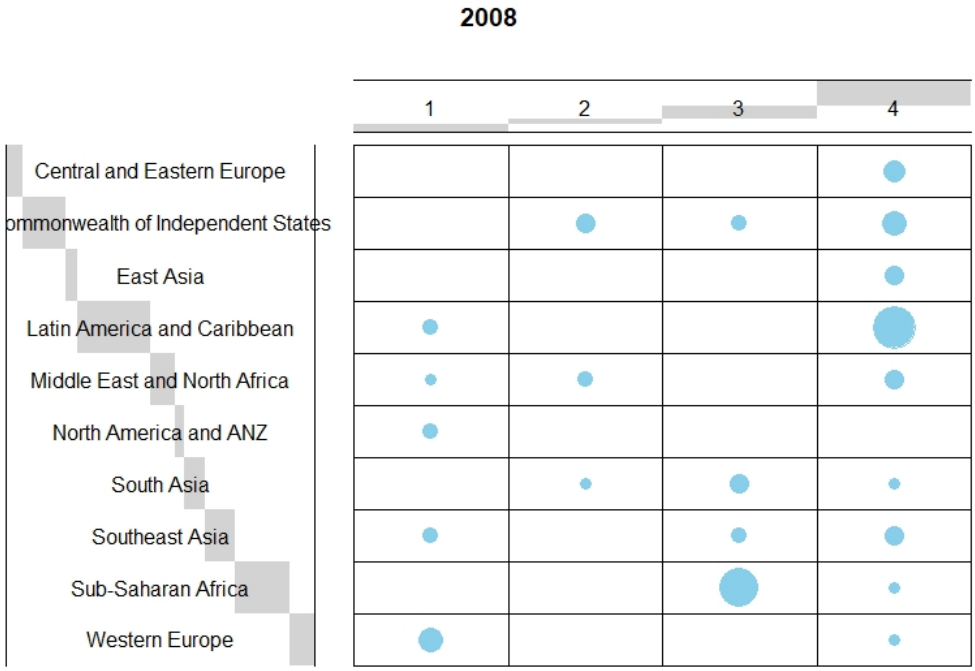
6.2.1 k-Means



D'après la courbe des variances intra-clusters ainsi que la règle "du coude", on retient quatre clusters. On obtient des clusters moyennement équilibrés : respectivement 6, 16, 21 et 30 observations. Le premier cluster est associé à des valeurs positives pour Life ladder, log GDP per capita, Social support, Healthy life expectancy at birth, Freedom to make life choices, Generosity et Positive Affect. Il est associé à des valeurs négatives pour Perceptions of corruption et pour Negative affect. Le deuxième cluster est associé à des valeurs positives pour Negative affect. Le troisième cluster est associé à des valeurs positives pour Perceptions of corruption. Il est associé à des valeurs négatives pour Life ladder, log GDP per capita, Social support et Healthy life expectancy at birth. Voici les quatre clusters dans les deux premières dimensions de l'ACP :



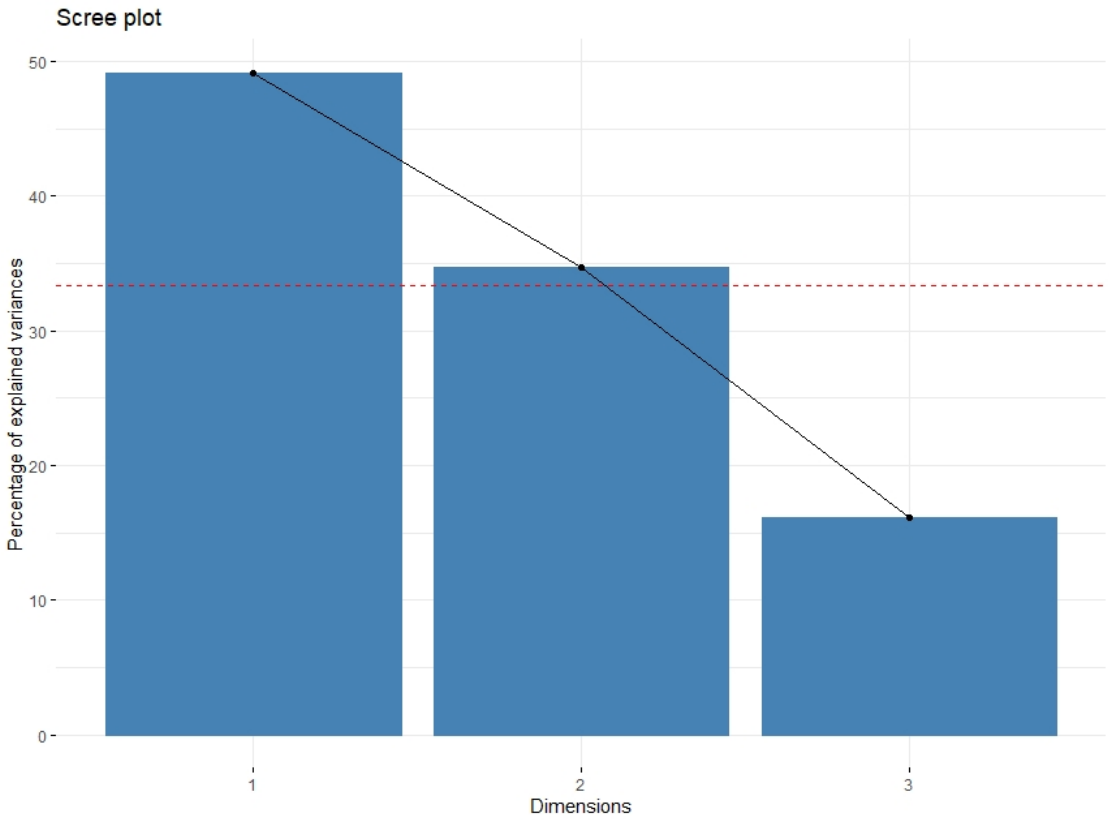
Regardons la distribution des régions en fonction du cluster :



6.2.2 AFC

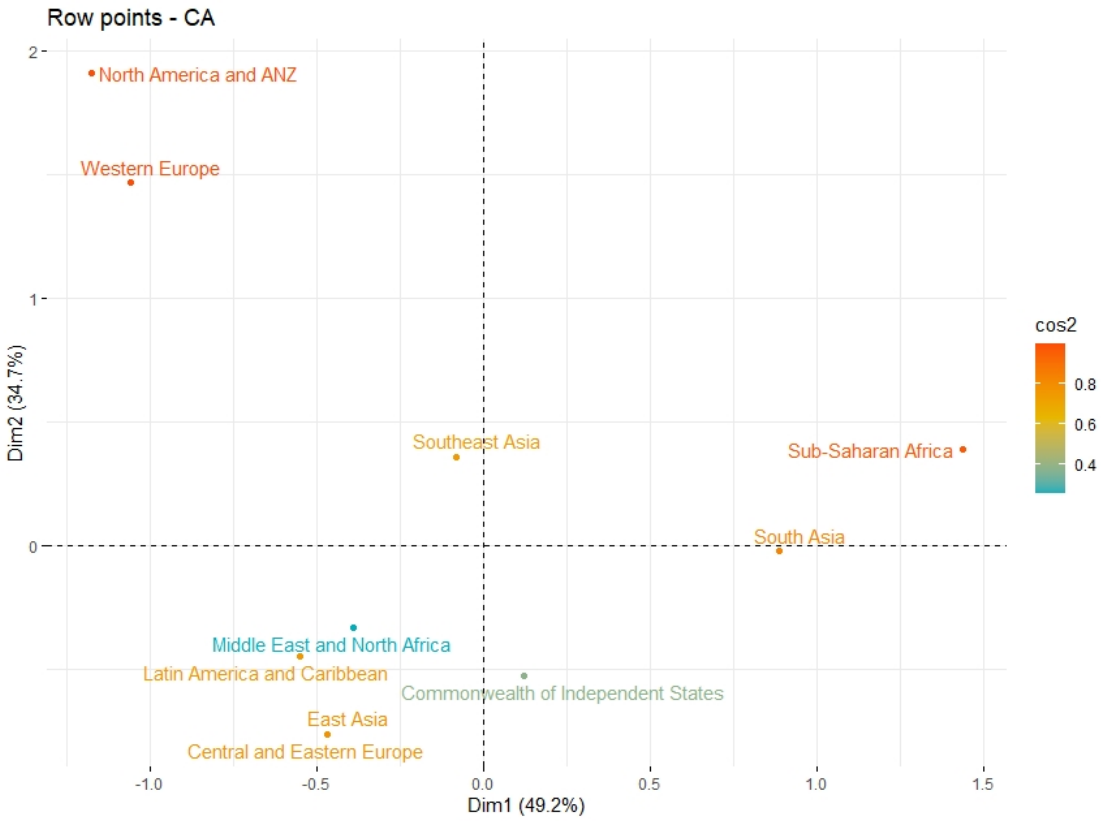
Essayons de faire une AFC pour voir s'il y a des correspondances entre les régions et les clusters auxquels sont assignés les pays.

Le test du khi-2 nous donne une p-valeur égale à 5.925631e-10. On peut donc dire qu'il y a une relation statistiquement significative entre les lignes et les colonnes.



Le scree plot ci-dessus nous montre que l'on doit retenir deux dimensions (les pointillés rouges indiquent la variance expliquée minimale pour qu'un axe ait plus d'information que si les données étaient totalement aléatoires).

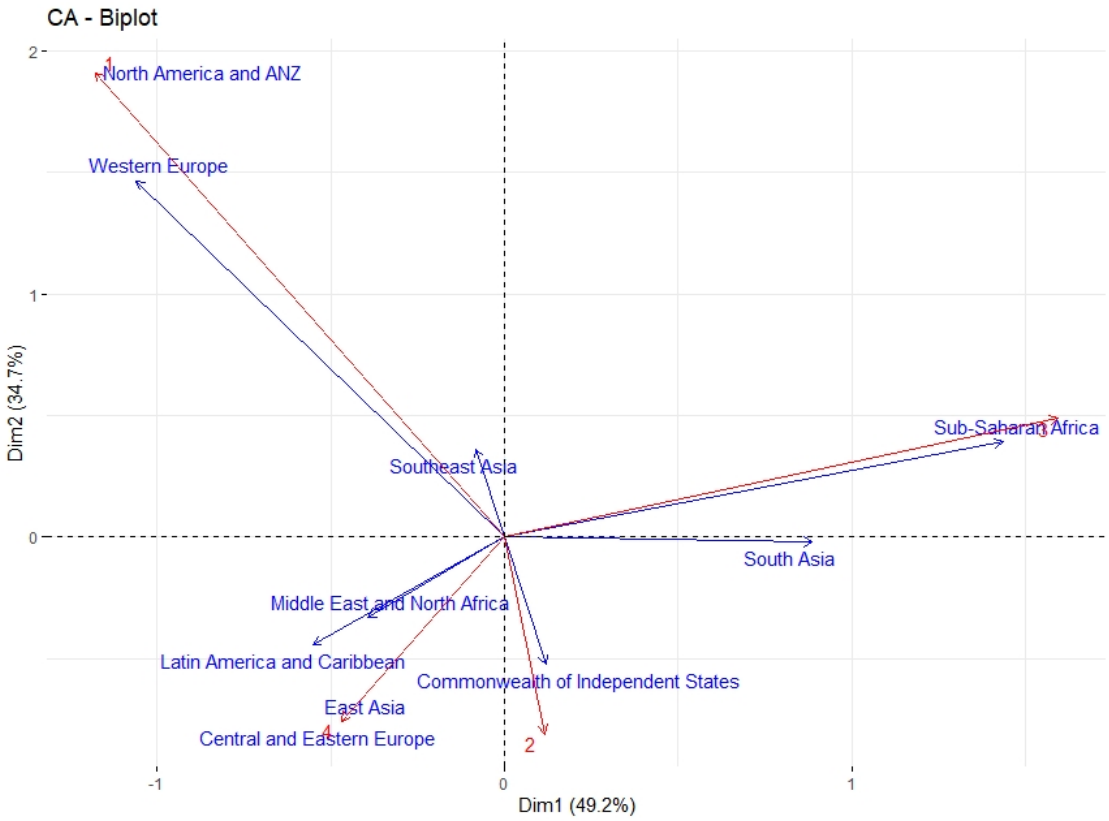
Voici le graphe des lignes dans le plan des deux premières dimensions :



On remarque que :

- Western Europe et North America and ANZ sont corrélés
- de même pour South Asia et Sub-Saharan Africa
- de même pour Latin America and Caribbean, Central and Eastern Europe et East Asia

Ci-dessous nous avons le biplot asymmétrique des lignes et des colonnes :

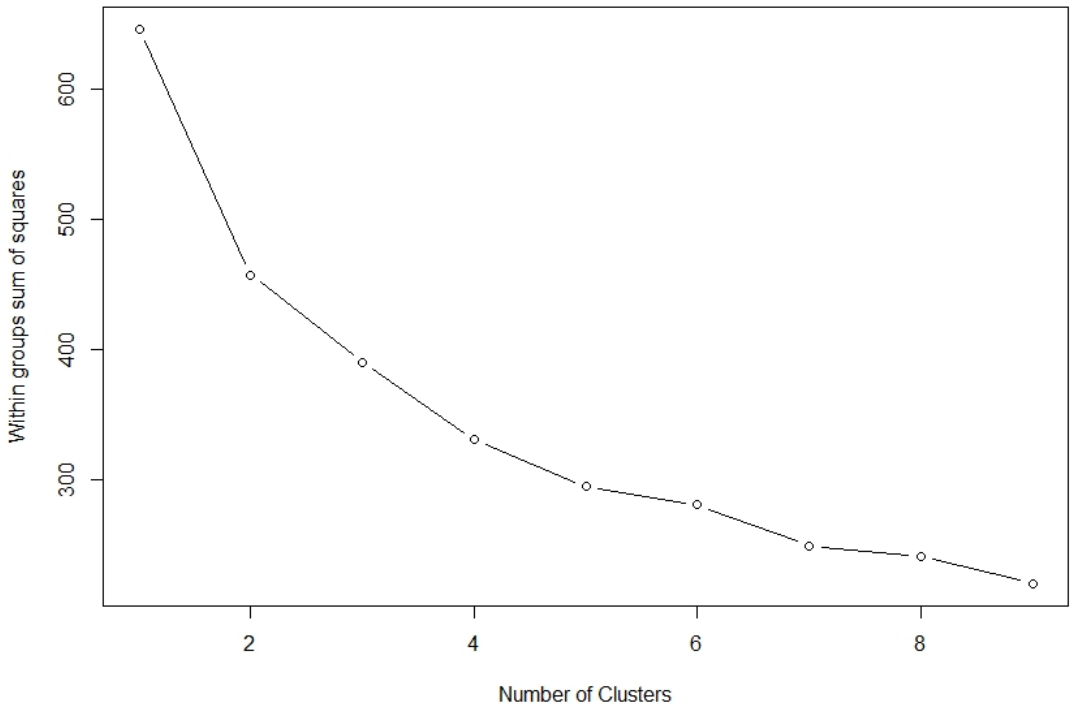


Ce graphe nous permet véritablement d'analyser les correspondances entre lignes et colonnes : il faut, pour cela, projeter orthogonalement les vecteurs lignes sur les vecteurs colonnes ; plus le projeté est proche de la flèche colonne, plus la corrélation est importante.

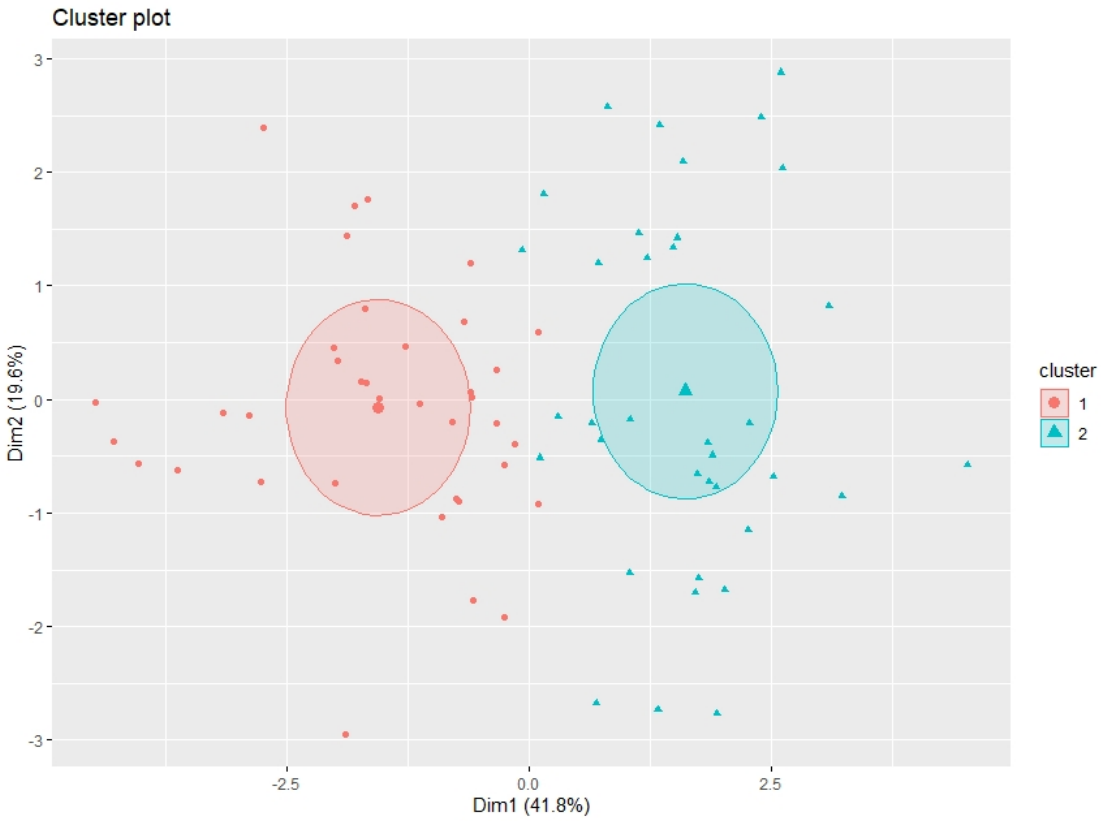
Ainsi, on peut dire que :

- North America and ANZ et Western Europe sont fortement corrélés au cluster 1
- Commonwealth of Independent States est corrélé au cluster 2
- Sub-Saharan Africa est fortement corrélé au cluster 3, South Asia assez corrélé
- Latin America and Caribbean est fortement corrélé au cluster 4, Middle East and North Africa assez corrélé

6.3 Année 2009

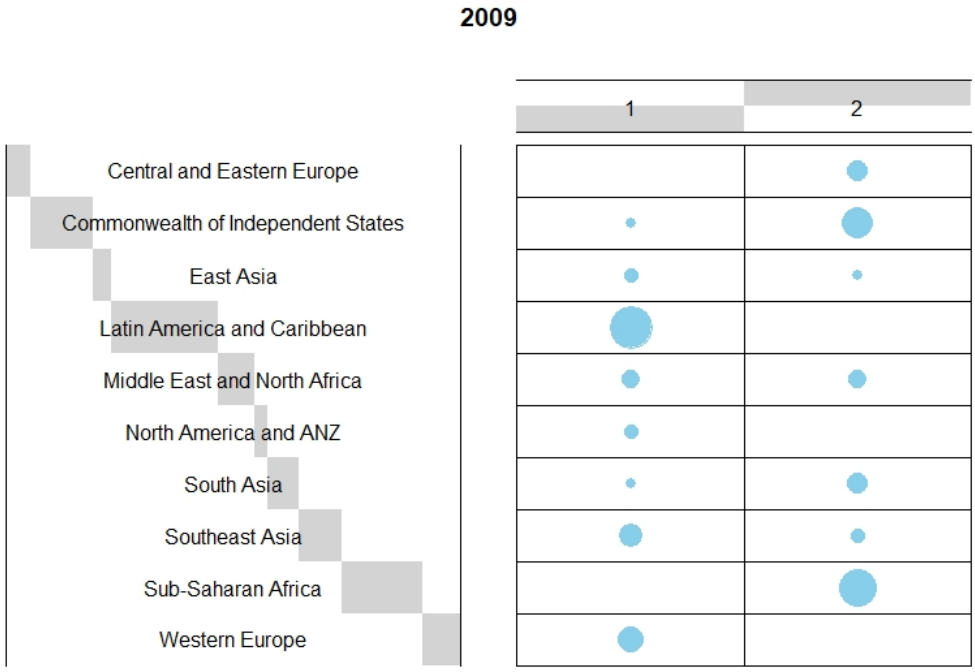


D'après la courbe des variances intra-cluster ainsi que la règle "du coude", on ne retient que deux clusters. On obtient des clusters parfaitement équilibrés : respectivement 37 et 36 observations. Le premier cluster est associé à des valeurs positives pour toutes les variables hormis Perceptions of corruption, et inversement pour le deuxième cluster. Voici les deux clusters dans les deux premières dimensions de l'ACP :



De la même manière que pour l'année 2007, les deux clusters se partagent les observations en fonction des coordonnées de celles-ci sur l'axe 1.

Voici la distribution des régions en fonction du cluster :



Latin America and Caribbean, North America and ANZ et Western Europe sont totalement liés au cluster 1 ; Central and Eastern Europe et Subsaharan Africa sont totalement liés au deuxième cluster.

L'utilisation du clustering nous a permis de distinguer des permanences dans les différences entre régions du monde, qu'il faut toutefois nuancer du fait de la dispersion des pays au sein d'une même région, en particulier pour Latin America and Caribbean, qui regroupe beaucoup de pays.

De manière générale, et c'est ce qui mériterait d'être étudié plus précisément, la richesse qui semble être un facteur déterminant : en 2007 et en 2009, les clusters se séparent en fonction de la première dimension, sur laquelle interviennent fortement la qualité de vie et le logarithme du PIB par habitant.

Une autre perspective pourrait être celle d'étudier les années de survenue de crise. En effet, on a remarqué qu'en 2008, les pays avaient des valeurs plus diffuses dans les différentes variables, nous conduisant à établir quatre clusters. On peut donc émettre l'hypothèse que le choc qu'est la survenue d'une crise est apprécié différemment d'un pays à l'autre ou d'une région à l'autre. Une question importante serait d'ailleurs de distinguer l'effet direct de la crise et son incidence du fait qu'elle bouleverse les valeurs des individus.

7 Conclusion et les limites

Cette étude a pour but de trouver une métrique évaluant l'état général du psychisme des habitants dans les différents pays. Cette métrique peut être utilisée pour évaluer l'impact des chocs économiques et financiers sur le psychisme des gens. À travers les différentes ACP, nous retrouvons une modification de la composition des composantes lors de l'année 2008 correspondant à l'année de la crise financière mondiale. Il est possible d'utiliser la composante principale comme métrique.

Cependant, ces analyses ne permettent pas de trouver les liens causaux entre les variables de la base de données. Par ailleurs, les questions posées ne permettent pas de projeter entièrement l'état psychique des habitants. On notera également que les réponses sont des moyennes glissantes sur les trois dernières années lors de l'interrogation. Cela fausse l'état psychique des habitants pour l'année en question. Nous avons retenu uniquement les pays où nous disposons une réponse pour toutes les questions sur les trois ans en question. Cela a supprimé entre 37 et 40 pays selon l'année.

Il peut être plus intéressant de faire une même analyse en utilisant les méthodes conventionnelles pour évaluer les politiques publiques. Pour cela, il nous aurait fallu avoir une variable discriminatoire qui puisse indiquer l'impact de la crise de 2008 sur le pays.