

DATA MINING II

LES DÉTERMINANTS DU BONHEUR NATIONAL

LAMON Océane
SEZESTRE Émilien

Enseignant : Christophe
Daniel



Table des matières

I.	Introduction	2
i.	Présentation du sujet.....	2
ii.	Problématique et plan.....	3
II.	Présentation des données	3
i.	Présentation de la base de données.....	4
ii.	Description des variables	5
iii.	Traitement des données manquantes.....	7
III.	Statistiques descriptives.....	8
IV.	Analyse en composantes principales	13
i.	Méthodologie.....	13
a)	Présentation de la méthode d'analyse.....	13
b)	Inertie et dimensions retenues.....	14
ii.	Résultats	15
a)	Analyse graphique	16
1.	Description des variables.....	16
2.	Description des individus.....	17
b)	Description des dimensions.....	21
1.	Première dimension	21
2.	Deuxième dimension.....	23
V.	Régression linéaire multiple	24
i.	Méthodologie.....	24
a)	Présentation de la méthode d'analyse.....	24
b)	Variable expliquée et variables explicatives retenues.....	24
ii.	Tests et résultats	28
VI.	Régression sur variables latentes.....	34
i.	PCR.....	34
a)	Présentation de la méthode d'analyse.....	34
b)	Résultats.....	35
ii.	PLSR	38
a)	Présentation de la méthode d'analyse.....	38
b)	Résultats.....	38
VII.	Conclusion.....	44
VIII.	Bibliographie.....	47
IX.	Annexes	48

I. Introduction

Cette partie introductive a pour objectif de présenter le sujet, d'en définir les termes principaux, d'explicitier le contexte dans lequel il a évolué et de poser une problématique liée à la thématique.

i. Présentation du sujet

Dans son ouvrage *Fondements de la métaphysique des mœurs* de 1785, le philosophe allemand Emmanuel Kant définit le bonheur comme suit : "Le pouvoir, la richesse, la considération, même la santé ainsi que le bien-être complet et le contentement de son état, est ce qu'on nomme le bonheur"¹.

Il est ainsi possible de caractériser le bonheur comme étant un état de satisfaction stable et durable, moins éphémère que le plaisir et moins dynamique que la joie. Il repose, toujours selon Kant, sur les expériences de chacun, ce qui fait du bonheur un concept empirique et donc indéterminé. Cela implique qu'il est très difficile de s'accorder sur une définition, et encore plus sur les différents moyens d'atteindre le bonheur.

Malgré cela, le bonheur représente un but à l'existence de chacun, que les individus cherchent à atteindre de façon consciente ou inconsciente. De ce fait, il est universellement recherché par l'ensemble des êtres humains. Cela implique que la notion de bonheur, ou de bien-être, aussi subjective et indescriptible qu'elle puisse être, est au centre des préoccupations et des choix politiques de nos sociétés modernes. En effet, les différentes politiques économiques, bien qu'elles ne soient pas traditionnellement construites sur autre chose que des finalités économiques et de compétitivité des territoires, ont également un impact sur les individus et le bien-être qu'ils ressentent, compte tenu des conséquences de ces politiques économiques sur leur vie quotidienne. Ainsi, en juin 2016, l'OCDE s'était engagée à « redéfinir le discours sur la croissance pour placer le bien-être des personnes au centre des efforts des gouvernements ».

Le bien-être n'est pas un concept nouveau en économie. En effet, le domaine a vocation à imposer un impérialisme économique sur les sciences humaines et sociales, et n'a aucune réticence à s'intéresser à des notions subjectives, qualitatives, relatives à la vie quotidienne des individus, tout en essayant de les expliquer à l'aide d'outils et de méthodes issus de la science économique. Ainsi, on retrouve dès le XVIIIe siècle la notion d'utilitarisme développée par Jérémie Bentham, puis par John Stuart Mill au XIXe siècle, dont l'objectif est de maximiser le bien-être collectif pour atteindre « le plus grand bonheur du plus grand nombre ». La théorie néoclassique a pour sa part introduit « l'économie du bien-être », pour laquelle une situation est considérée comme étant optimale si aucun individu ne peut améliorer son bien-être sans détériorer celui d'un autre (optimum de Pareto). Enfin, cette approche a été formalisée par Arthur Pigou au XXe siècle, notamment dans son ouvrage *L'économie du bien-être* (1920), dans lequel il pose des hypothèses rattachées au bien-être de nature individualiste (l'individu est le seul juge de son bien-être, et le bien-être collectif est exclusivement défini à partir du bien-être de chacun), et où il examine le lien entre satisfaction des individus et variables économiques.

¹ KANT Emmanuel, *Fondements de la métaphysique des mœurs* (1785), Première section, traduit de l'allemand par Victor Delbos, Edition Delagrave, 1977, pp 87-88.

ii. Problématique et plan

Ces éléments soulèvent plusieurs interrogations. D'une part, la difficulté à définir le bonheur peut induire une plus grande difficulté encore à le mesurer de façon convenable. Et si nous étions amenés à analyser un estimateur relativement satisfaisant du bien-être des individus, il n'est pas certain que ce dernier soit à même d'être expliqué par des variables d'origine économique. Puisque le bonheur est subjectif et issu des expériences de chacun, les conséquences de politiques socio-économiques sur les individus pourraient ne pas être généralisées à l'ensemble d'une population comme étant favorables ou défavorables à l'augmentation de leur bien-être.

Cette incertitude sera le point de départ de notre analyse, et elle nous permet de poser les problématiques suivantes :

Quels sont les déterminants du bonheur national ? Le bonheur peut-il être expliqué par des indicateurs et instruments socio-économiques ?

Ce dossier aura vocation à déterminer puis présenter un indicateur du bonheur national considérant les pays du monde entier. En plus de cet indicateur, diverses variables relatives à l'environnement dans lequel évoluent les individus, donc témoignant d'aspects économiques, démographiques, sociétaux et politiques de chaque pays considéré, et susceptibles d'affecter de façon positive ou négative le bonheur, seront introduites. Après avoir présenté la base de données ainsi que les variables qui la composent, le lien entre ces dernières sera dans un premier temps examiné à travers des statistiques descriptives et une méthode d'exploration de données. Dans un second temps, nous mettrons en place des modèles créés à partir de nos variables et susceptibles d'expliquer le bonheur, par le biais de régressions économétriques. Enfin, la conclusion de notre étude nous permettra de revenir sur les principaux résultats obtenus pour apporter une réponse à la problématique, ainsi que d'exposer les limites de notre analyse.

II. Présentation des données

L'objectif de cette seconde partie est de présenter la base de données qui constitue le cœur de notre analyse. Nous évoquerons les sources à partir desquelles elle a été construite, puis nous apporterons une description plus précise à chacune des variables qui la composent. Dans la mesure où notre base comportait des observations manquantes pour certains individus, nous reviendrons également sur le traitement qui a été apporté à ces dernières, dans la mesure où certaines méthodes d'analyse utilisées dans cette étude requièrent d'utiliser des bases de données ne comportant aucune donnée manquante.

i. Présentation de la base de données

La base de données utilisée dans cette étude est un tableau de contingence croisant en ligne des individus, et en colonne des variables. Les individus considérés sont les différents pays du monde, au nombre de 151. La différence entre le nombre de pays présents dans la base de données et le nombre de pays existant dans le monde provient d'un manque de données vis-à-vis de la situation démocratique et politique des pays concernés. Les pays pour lesquels seulement quelques indicateurs étaient manquants ont été maintenus dans la base, et une méthode d'imputation de données manquantes leur a été appliquée.

Les variables de cette étude sont au nombre de vingt-six (26), et concernent l'année 2019 pour les différents indicateurs. Notre jeu de données recense en effet quinze variables quantitatives continues, six variables quantitatives binaires et cinq variables qualitatives.

L'ensemble des variables est listé ci-après :

- Nom du pays
- Code du pays
- Région du monde
- Continent
- Niveau de bonheur
- PIB/habitant
- Taux de chômage
- Taux de fécondité
- Espérance de vie en bonne santé à la naissance
- Support social
- Liberté de faire des choix de vie
- Générosité
- Voix citoyenne et responsabilité
- Stabilité politique et absence de violence
- État de droit
- Efficacité des pouvoirs publics
- Qualité de la réglementation
- Perceptions de la corruption
- Index de démocratie
- Europe de l'Ouest
- Afrique Sub-saharienne
- Asie de l'Est
- Amérique du Nord et ANZ
- Type de démocratie
- Démocratie complète
- Régime totalitaire

Notons que les quatre premières variables permettent de catégoriser chaque pays selon son nom, code et emplacement géographique, la cinquième variable est l'indicateur de mesure du bonheur. Les deux suivantes sont des indicateurs économiques, les deux qui suivent sont des indicateurs démographiques, tandis que les trois suivants ont davantage une dimension sociologique. Les sept suivants sont quant à eux relatifs aux dimensions de la bonne gouvernance, et les quatre qui suivent sont des variables binaires relatives à la géographie mondiale. Enfin, les trois dernières sont relatives au type de démocratie, construites à partir de l'index de démocratie.

Chacune de ces vingt-six variables sera présentée de façon plus précise dans la section suivante.

ii. Description des variables

Variables pays :

- **Country Name** : Nom du pays, en anglais. Variable qualitative.
- **Country Code** : Code de trois lettres majuscules pour identifier chaque pays conformément à la norme internationale ISO 3166-1. Variable qualitative.
- **Region** : Région du monde à laquelle chaque pays est géographiquement rattaché (10 modalités). Variable qualitative.
- **Continent** : Nom du continent auquel chaque pays est géographiquement rattaché. Variable qualitative.

Variable bonheur :

- **Happiness** : Réponse moyenne nationale à la question de l'échelle de Cantril du Gallup World Poll (GWP) invitant les gens à évaluer la qualité de leur vie actuelle sur une échelle de 0 à 10 : « Imaginez une échelle avec des marches numérotées de 0, en bas, à 10, en haut. Supposons que le haut de l'échelle représente la meilleure vie possible pour vous, et le bas de l'échelle la pire vie possible pour vous. Si la marche la plus haute est 10 et la plus basse 0, sur quelle marche vous placeriez-vous en ce moment même ? ». Variable quantitative continue. *Source : World Happiness Report 2020, chapter 2.*

Variables économiques :

- **GDP** : Logarithme du PIB/habitant exprimé en parité de pouvoir d'achat (PPA). Variable quantitative continue. *Source: World Bank's Global Economic Prospects.*
- **Unemployment** : Part de la population active sans travail, mais disponible et à la recherche d'un emploi. Exprimé en pourcentage. Variable quantitative continue. *Source: International Labour Organization, ILOSTAT database.*

Variables démographie :

- **Fertility rate** : Nombre d'enfants qu'une femme mettrait au monde si elle vivait jusqu'à la fin de ses années de procréation, et conformément aux taux de fécondité par âge de l'année spécifiée. Variable quantitative continue. *Source: United Nations Population Division. World Population Prospects: 2019*
- **Healthy life expectancy** : Espérance de vie en bonne santé à la naissance. Exprimé en années. Variable quantitative continue. *Source: World Happiness Report 2020, chapter 2; calculé à partir des données de la World Health Organization (WHO) et des World Bank's World Development Indicators.*

Variables social :

- **Social support** : Moyenne nationale des réponses binaires à la question du Gallup World Poll (GWP) « Si vous étiez en difficulté, auriez-vous des membres de votre famille ou des amis sur lesquels vous pourriez compter pour vous aider quand vous en auriez besoin, ou pas ? ». *Source : World Happiness Report 2020, chapter 2.*
- **Freedom to make life choices** : Moyenne nationale des réponses binaires à la question du Gallup World Poll (GWP) « Etes-vous satisfait ou insatisfait de votre liberté de choisir ce que vous faites de votre vie ? ». *Source : World Happiness Report 2020, chapter 2.*
- **Generosity** : Résidu de la régression de la moyenne nationale des réponses au sondage du Gallup World Poll (GWP) des réponses à la question « Avez-vous donné de l'argent à une association caritative au cours du mois dernier ? » sur le PIB par habitant. Variable quantitative continue. *Source : World Happiness Report 2020, chapter 2.*

Variables gouvernance :

- **Voice and Accountability** : Dans quelle mesure les citoyens participent véritablement au choix de leurs gouvernants, par l'étendue de leur liberté d'expression, leur liberté d'association, ou encore celle des médias. Variable quantitative continue. *Source: Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010).*
- **Pol. Stability and abs. of violence** : Probabilité que le gouvernement puisse être déstabilisé ou renversé, par des moyens inconstitutionnels, par la violence politique, ou par le terrorisme. Variable quantitative continue. *Source: Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010).*
- **Rule of Law** : Dans quelle mesure les citoyens ont confiance en les règles qui régissent la société, et dans quelle mesure ils les respectent ou non. Variable quantitative continue. *Source: Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010).*
- **Government Effectiveness** : Qualité des services publics, fonctionnaires, et degré d'indépendance quant aux pressions politiques. Mais également qualité des politiques publiques (dans leur définition et application) et responsabilité du gouvernement vis-à-vis de ces dernières. Variable quantitative continue. *Source: Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010).*
- **Regulatory Quality** : Capacité du gouvernement dans l'application de politiques et des réglementations adéquates et favorisant le développement économique du secteur privé. Variable quantitative continue. *Source: Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010).*
- **Perceptions of corruption** : Moyenne nationale des réponses à deux questions à réponse fermée du Gallup World Poll (GWP) « La corruption est-elle répandue dans tout le gouvernement ? » ; « La corruption est-elle généralisée au sein des entreprises ? ». *Source: World Happiness Report 2020, chapter 2.*
- **Democracy index** : Indice reposant sur cinq catégories : processus électoral et pluralisme, fonctionnement du gouvernement, participation politique, culture politique, libertés civiles. Variable quantitative continue. *Source: Economist Intelligence Unit, Democracy Index 2020.*

Variables géographie :

- **Western Europe** : Prend la valeur 1 si le pays concerné est localisé en Europe de l'Ouest, 0 sinon. Variable binaire.
- **Sub Saharian Africa** : Prend la valeur 1 si le pays concerné est localisé en Afrique Sub-Saharienne, 0 sinon. Variable binaire.
- **East Asia** : Prend la valeur 1 si le pays concerné est localisé en Asie de l'Est, 0 sinon. Variable binaire.
- **North America and ANZ** : Prend la valeur 1 si le pays concerné est localisé en Amérique du Nord ou ANZ (Australie, Nouvelle-Zélande), 0 sinon. Variable binaire.

Variables démocratie :

- **Democracy** : Catégorise le pays selon le type de démocratie qui le caractérise le mieux (démocratie complète, démocratie imparfaite, régime hybride, régime autoritaire). Variable qualitative. *Source: Economist Intelligence Unit, Democracy Index 2020.*
- **Démocratie complète** : Prend la valeur 1 si le pays concerné est en situation de démocratie complète, 0 sinon. Variable binaire.
- **Régime totalitaire** : Prend la valeur 1 si le pays concerné est en situation de régime totalitaire, 0 sinon. Variable binaire.

iii. Traitement des données manquantes

La base de données utilisée dans cette étude et construite à partir des différentes variables présentées ci-dessus comportait à l'origine des valeurs numériques manquantes, au nombre de 6, pour un taux de complétude concernant les variables quantitatives de l'ordre de 99,81%. Malgré le faible nombre de données manquantes compte tenu du nombre de valeurs totales présentes dans la base, certaines méthodes d'analyse requièrent un jeu de données ne comportant aucune valeur manquante, ce qui nous contraint à traiter les données manquantes que comporte notre base de données.

De ce fait, nous avons procédé à l'utilisation du package *missMDA* sur R pour réaliser de l'imputation, autrement dit remplacer les données manquantes d'un jeu de données par des valeurs plausibles, ces mêmes valeurs étant prédites à partir d'un modèle qui prend en compte à la fois les similarités entre les individus et entre les variables.

Pour procéder à l'imputation de données manquantes, l'algorithme remplace la valeur manquante par la moyenne des valeurs observées pour la variable concernée. Le jeu de données ainsi complété est soumis à une ACP (Analyse en Composantes Principales), et la valeur prédite par l'ACP pour la donnée manquante est utilisée à la place de la moyenne. Une nouvelle ACP est ensuite réalisée, et les étapes précédentes sont répétées jusqu'à convergence de la valeur prédite entre l'ACP précédente et l'ACP qui la suit.

La base de données comportant les données imputées à la suite de cette procédure sera utilisée pour toutes les méthodes d'analyse et de statistiques réalisées dans la suite de notre étude.

III. Statistiques descriptives

CONTINENT	REPARTITION DES PAYS (EFFECTIF)	REPARTITION DES PAYS (%)
AFRIQUE	55	36,43%
AMERIQUE	23	15,23%
ASIE	31	20,53%
EUROPE	40	26,49%
OCEANIE	2	1,32%
MONDE	151	100%

REGION	REPARTITION DES PAYS (EFFECTIF)	REPARTITION DES PAYS (%)
AMERIQUE LATINE ET CARAIBES	21	13,90%
AMERIQUE DU NORD ET ANZ	4	2,65%
AFRIQUE SUB-SAHARIENNE	39	25,83%
ASIE DE L'EST	6	3,97%
ASIE DU SUD	7	4,64%
ASIE DU SUD-EST	9	5,96%
COMMUNAUTE DES ÉTATS INDEPENDANTS	12	7,95%
EUROPE CENTRALE ET EUROPE DE L'EST	17	11,26%
EUROPE DE L'OUEST	20	13,24%
MOYEN ORIENT ET AFRIQUE DU NORD	16	10,60%
MONDE	151	100%

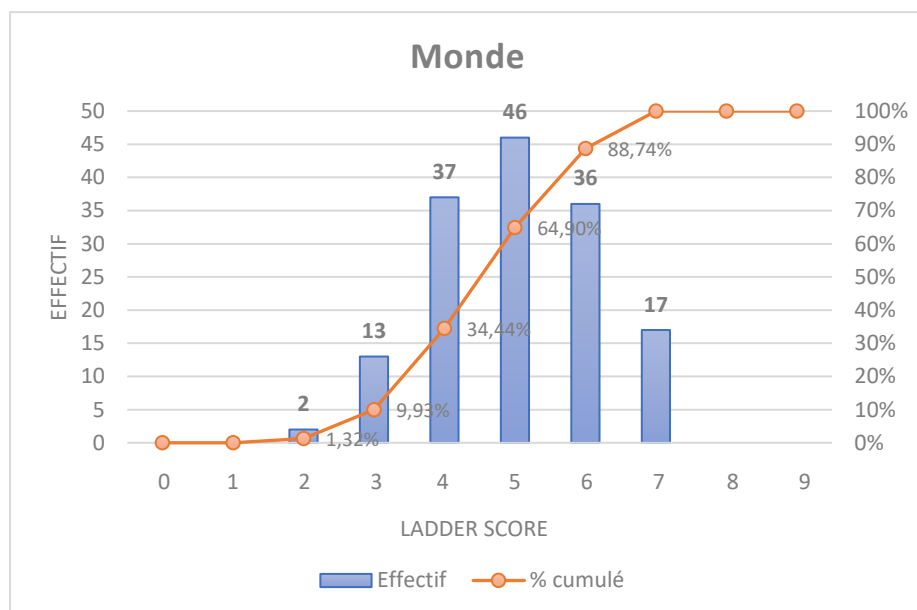
Notre base de données comporte 151 individus qui sont des pays du monde. Les deux tableaux ci-dessus présentent leur répartition géographique, d'abord par continent, puis plus spécifiquement par région du monde.

Ainsi, sur les 151 pays considérés dans cette base de données parmi les 195 reconnus par l'ONU, 55 se situent sur le continent Africain soit plus de 36% du total. Le continent Européen concentre pour sa part plus de 26% des pays de notre jeu de données, quand ce taux n'est que de 1,32% pour l'Océanie avec deux pays seulement : l'Australie et la Nouvelle Zélande. Lorsque l'on considère ces deux pays dans le second tableau, ils sont cette fois-ci recensés sous la catégorie « Amérique du Nord et ANZ » avec les USA et le Canada, ces quatre pays pesant ainsi 2,65% du total des pays de la base de données. La région du monde la plus représentée dans le second tableau est l'Afrique sub-saharienne avec 25,83% des effectifs (39 pays), suivi par l'Amérique latine et les Caraïbes (13,90%) puis les différentes régions d'Europe (Europe de l'Ouest puis Europe centrale et Europe de l'Est) avec respectivement 20 et 17 pays.

HAPPINESS			
MEAN	SD	MIN	MAX
5,479	1,117	2,567	7,809

La variable modélisant le bonheur national est un score compris entre 0 et 10, ce dernier étant évalué par la moyenne des réponses à la question de l'échelle de Cantril qui s'intéresse au bien-être subjectif des individus, affectant ainsi à chaque pays un *Ladder score*, ici sous l'appellation *happiness*. La moyenne mondiale est de 5,479 tandis que l'écart-type témoigne

d'une faible dispersion des différents score nationaux autour de cette moyenne mondiale. Enfin, le score le plus faible est de 2,567, recensé en Afghanistan, tandis que le pays le plus heureux du monde selon cet indicateur serait la Finlande avec un score de 7,809.

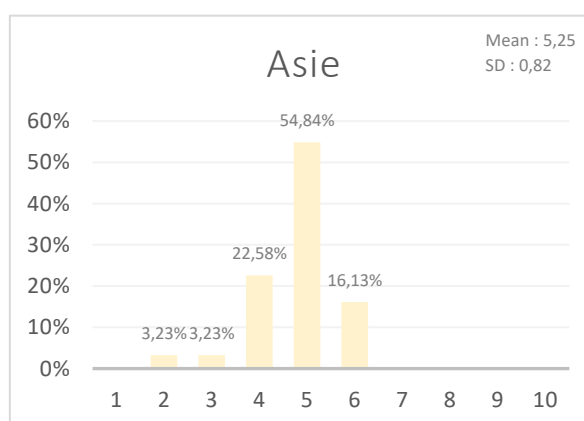
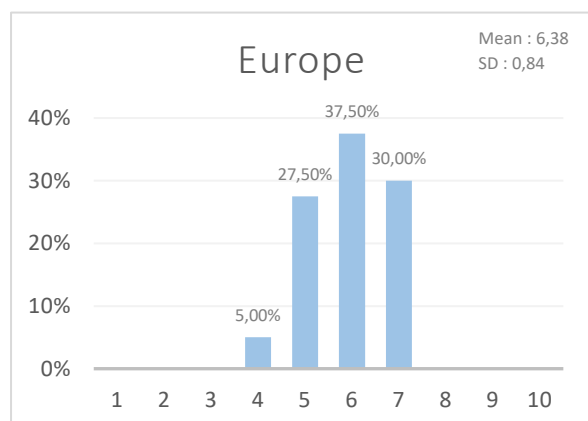
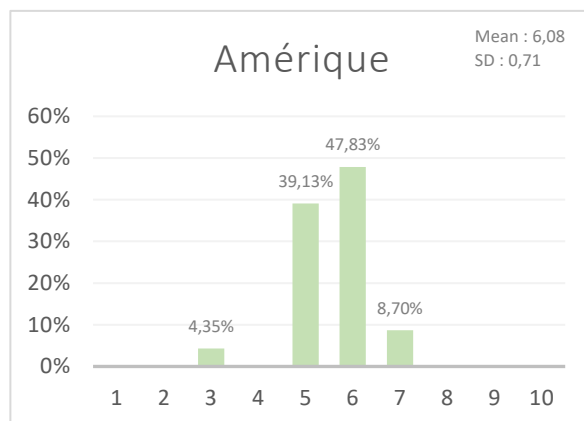
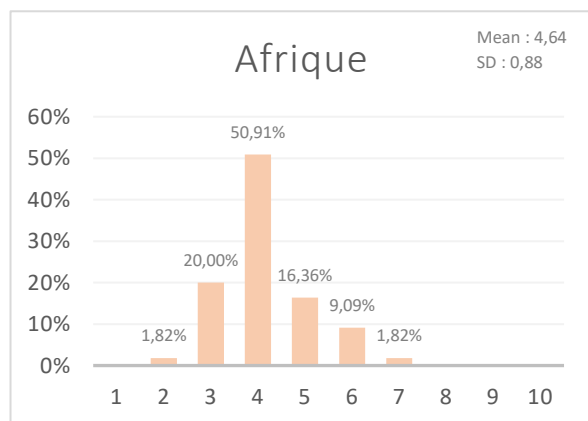


Cet histogramme présente la répartition des scores par pays du monde (effectifs) ainsi que les fréquences cumulées associées.

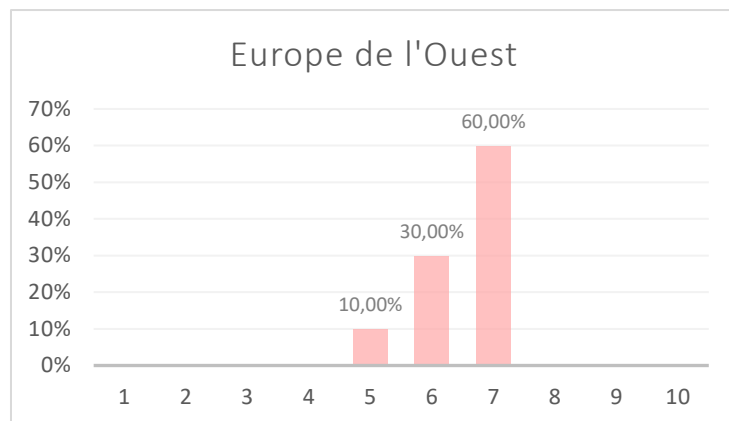
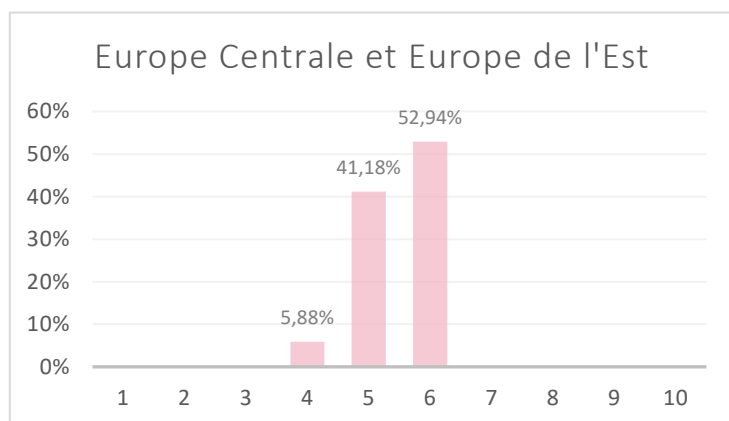
La classe de score la plus représentée est le 5, soit 46 pays ayant un ladder score compris dans l'intervalle [5;6[.

Si l'on considère les fréquences cumulées, l'on peut notamment constater que plus de 90% des pays

de notre jeu de données ont un score supérieur à 3 tandis que seulement 11,86% d'entre eux ont un score supérieur à 7. En effet, seuls 17 pays sur 151 ont un score compris dans l'intervalle [7;8[, 70% d'entre eux, soit 12 pays, étant situés en Europe de l'Ouest (Finlande, Danemark, Suisse, Islande, Norvège, Pays-Bas et Suède qui occupent respectivement les sept premières places en termes de ladder score). D'autres pays comme le Canada, l'Australie ou encore le Costa Rica présentent également des scores très élevés.



Les histogrammes ci-dessus nous permettent notamment de constater une répartition relativement inégale des ladder score en fonction des différents continents. En effet, 72,73% des pays du continent Africain ont un score strictement inférieur à 5, quand ce taux n'est que de 5% pour l'Europe ou encore de 4,35% pour l'Amérique. De la même façon, seuls 1,82% des pays africains et 8,70% des pays du continent Américain ont un score strictement supérieur à 7, contre 30% des pays européens. Aucun pays Asiatique de notre jeu de données n'est doté d'un tel score, le plus élevé étant atteint par l'État insulaire de Taiwan avec une note de 6,45. Enfin, les écarts-type des différentes distributions peuvent suggérer des disparités plus ou moins marquées au sein d'un même continent.



On dénombre dans notre jeu de données 17 pays d'Europe centrale et de l'Est pour 20 pays d'Europe de l'Ouest. Leurs moyennes respectives de score de bonheur sont de 5,88 et de 6,97. Manifestement, les pays d'Europe de l'Ouest semblent en moyenne plus heureux que les autres, encore faut-il déterminer si cette différence est significative ou non.

PAYS	EFFECTIF	HAPPINESS (MEAN)
EUROPE DE L'OUEST	20	6.967
MONDE, HORS EUROPE DE L'OUEST	131	5.252

Le bonheur moyen des pays d'Europe de l'Ouest est supérieur à celui du reste du monde, mais il convient de tester la significativité de cet écart entre les deux moyennes. La réalisation d'un test de Student nous permet de rejeter l'hypothèse nulle (la différence entre les moyennes est égale à zéro) et donc d'accepter l'hypothèse alternative selon laquelle la moyenne du groupe Europe de l'Ouest est significativement supérieure, avec un risque de première espèce de l'ordre de 1%. Les pays d'Europe de l'Ouest sont donc, en moyenne, plus heureux que les pays du reste du monde.

Pour poursuivre nos statistiques descriptives, nous pouvons nous intéresser à une autre variable qualitative : le type de démocratie. En effet, notre variable *Democracy* est construite à partir de l'index de démocratie, variable quantitative présente dans notre jeu de données, et permet de catégoriser chaque pays parmi l'un des quatre types de démocraties suivantes : démocratie complète, démocratie imparfaite, régime hybride et régime autoritaire. Il peut être intéressant de croiser cette variable avec notre variable *Continent* pour étudier le lien entre position géographique et type de démocratie.

Democracy	Africa	America	Continent Asia	Europe	Oceania	Total
Authoritarian	31	2	12	1	0	46
	67.39	4.35	26.09	2.17	0.00	100.00
	56.36	8.70	38.71	2.50	0.00	30.46
	20.53	1.32	7.95	0.66	0.00	30.46
Flawed democracy	7	12	8	20	0	47
	14.89	25.53	17.02	42.55	0.00	100.00
	12.73	52.17	25.81	50.00	0.00	31.13
	4.64	7.95	5.30	13.25	0.00	31.13
Full democracy	1	4	4	14	2	25
	4.00	16.00	16.00	56.00	8.00	100.00
	1.82	17.39	12.90	35.00	100.00	16.56
	0.66	2.65	2.65	9.27	1.32	16.56
Hybrid regime	16	5	7	5	0	33
	48.48	15.15	21.21	15.15	0.00	100.00
	29.09	21.74	22.58	12.50	0.00	21.85
	10.60	3.31	4.64	3.31	0.00	21.85
Total	55	23	31	40	2	151
	36.42	15.23	20.53	26.49	1.32	100.00
	100.00	100.00	100.00	100.00	100.00	100.00
	36.42	15.23	20.53	26.49	1.32	100.00

Tout d'abord, sur les 151 pays de notre jeu de données, 46 d'entre eux sont qualifiés de régimes autoritaires soit plus de 30% du total. Les démocraties complètes sont minoritaires et ne représentent que 16,6% des pays, tandis que les démocraties imparfaites sont bien plus répandues avec 47 pays recensés soit 31% du total.

Nous pouvons également remarquer que si l'on considère le continent africain, seul un pays est considéré comme une démocratie complète : il s'agit de l'île Maurice, qui bien que située en Afrique de l'Est est toutefois une ancienne colonie britannique indépendante depuis 1968. D'autre part, plus de la moitié des pays de ce continent (56,4%) est en situation de régime autoritaire, quand ce taux n'est que 2,5% pour l'Europe. En effet, le seul pays du continent européen étant considéré comme un régime autoritaire est la Biélorussie. De plus, 85% des pays d'Europe sont des démocraties, qu'elles soient complètes ou imparfaites, quand 56% des démocraties complètes et 42,5% des démocraties imparfaites du monde entier se situent sur le continent européen.

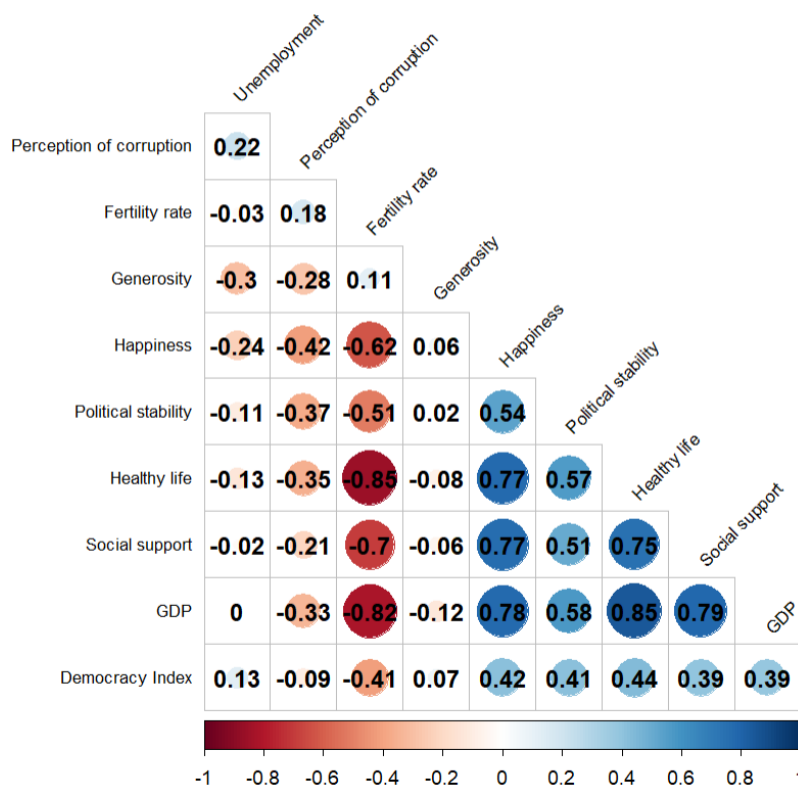
Enfin, les deux pays recensés dans notre jeu de données en Océanie, l'Australie et la Nouvelle-Zélande, sont des démocraties complètes, tandis que l'Asie n'est pas sur-représentée par un type de démocratie spécifique : sur 31 pays, 12 soit 38,7% du total sont des régimes autoritaires, 8 sont des démocraties imparfaites, 7 sont des régimes hybrides et 4 sont des démocraties complètes. Il s'agit de ce fait du seul continent dont l'un des quatre types de démocraties ne représente pas plus de 50% des pays sur son sol : 56,4% des pays Africains sont des régimes hybrides, 100% des pays d'Océanie sont des démocraties complètes, et respectivement 52,2% et 50% des pays d'Amérique et d'Europe sont des démocraties imparfaites.

Il peut à présent être intéressant de voir si le bonheur est significativement différent en fonction des types de démocratie qui régissent les pays du monde. En effet, l'étude de David Dorn et al. « Is it culture or democracy? The impact of democracy and culture on happiness » (2005) met en lumière l'existence d'une relation positive entre démocratie et bonheur, ainsi qu'un effet de la démocratie sur le bonheur plus marqué dans les pays à tradition démocratique établie.

DEMOCRATIE	EFFECTIF	HAPPINESS (MEAN)
DEMOCRATIE COMPLETE	25	6.898
AUTRE	126	5.197

Le bonheur moyen des pays en démocratie complète est supérieur à celui des autres, mais il convient une fois de plus de tester la significativité de cet écart entre les deux moyennes. La réalisation d'un test de Student nous permet de rejeter l'hypothèse nulle et donc d'accepter l'hypothèse alternative selon laquelle le bonheur moyen des pays en démocratie complète est significativement supérieur, avec un risque de première espèce de l'ordre de 1%. Les pays en démocratie complète sont donc, en moyenne, plus heureux que les pays dotés d'un autre type de démocratie.

Pour conclure cette partie, et face à l'impossibilité d'évoquer en détails l'intégralité des variables présentes dans notre jeu de données, nous allons nous pencher plus spécifiquement sur la corrélation entre plusieurs de nos variables, ce qui nous permettra d'introduire des variables sur lesquelles nous n'avons pas nécessairement focalisé notre attention dans le cadre des statistiques descriptives précédentes.



Le graphique ci-dessus est un corrélogramme, autrement dit une représentation graphique de la matrice des corrélations des variables que nous avons sélectionnées. Il permet ainsi de mettre en évidence les variables les plus corrélées entre-elles, ainsi que la nature de cette corrélation (positive ou négative).

Nous pouvons dans un premier temps constater que le niveau de bonheur (*happiness*) est corrélé positivement avec l'espérance de vie en bonne santé à la naissance (*healthy life*), le support social (*social support*) et le logarithme du PIB/hab en PPA (*GDP*), et dans une moindre mesure avec l'index de démocratie (*democracy index*) et la stabilité politique (*political stability*).

Cela signifie que l'une et l'autre variables évoluent dans le même sens : ainsi, le niveau de bonheur et l'espérance de vie en bonne santé à la même naissance tendent à augmenter (diminuer) ensemble, de même que pour le niveau de bonheur et le support social, ou encore le niveau de bonheur et l'index de démocratie.

D'autre part, le niveau de bonheur est négativement corrélé avec la perception de la corruption (perception of corruption), le taux de fécondité et le taux de chômage. Cela signifie que l'une et l'autre variables évoluent dans le sens contraire : ainsi, quand le niveau de bonheur augmente, le taux de fécondité a tendance à baisser (et inversement), de même que lorsque le taux de chômage baisse, le niveau de bonheur a tendance à augmenter (et inversement).

Rappelons enfin que le fait que deux variables soient fortement corrélées entre-elles n'implique pas qu'il y ait une relation de causalité entre l'une et l'autre. Autrement dit, voir deux variables évoluer ensemble, dans des sens contraires ou non, ne signifie pas nécessairement qu'une variable est la cause de l'autre.

Pourtant, nous pouvons notamment remarquer une corrélation négative très élevée entre le taux de fécondité et l'espérance de vie en bonne santé à ma naissance (-0,85), entre le taux de fécondité et le logarithme du PIB/hab en PPA (-0,82) ainsi qu'une corrélation positive très élevée entre le logarithme du PIB/hab en PPA et l'espérance de vie en bonne santé à la naissance (0,85). Ainsi, dans le cadre de la régression linéaire multiple (V.) et de la régression sur variable latentes (VI.), il sera nécessaire de s'assurer que les variables sélectionnées dans les modèles retenus ne soient pas sujet à de la multicollinéarité.

IV. Analyse en composantes principales

i. Méthodologie

a) Présentation de la méthode d'analyse

Notre jeu de données est un tableau à double entrée croisant individus en ligne, ici des pays, et variables quantitatives en colonne. À condition de traiter les variables qualitatives présentes dans notre base de données, l'analyse en composantes principales, ou ACP, sera la méthode adéquate aux données dont nous disposons, et nous permettra de mieux visualiser et résumer ces dernières.

L'Analyse en Composantes Principales (ACP) est une des méthodes d'analyse de données multivariée permettant d'explorer des jeux de données constitués de variables quantitatives. Elle permet de réduire la dimensionnalité et de rendre l'information moins redondante en déformant le moins possible la réalité. De ce fait, il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales tel qu'un maximum d'information soit conservé.

Comme relevé précédemment, et dans le cadre de la réalisation d'une ACP, il est nécessaire de traiter les variables qualitatives se trouvant dans notre base de données. À ce titre, nous pouvons définir les variables que nous considérerons comme actives, c'est-à-dire prises en

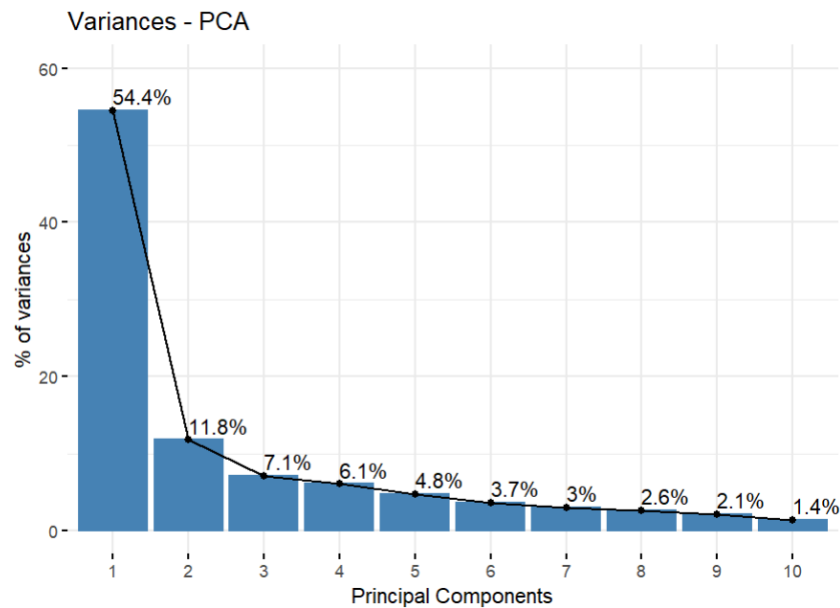
comptes dans l'analyse, ainsi que les variables illustratives, n'influençant pas la construction des composantes principales de l'analyse mais aidant à l'interprétation des dimensions. Le tableau ci-dessous recense, en plus de son nom et de son type, le statut affecté à chaque variable de notre jeu de données.

NOM	TYPE	STATUT
COUNTRY NAME	Qualitative nominale	Variable illustrative
COUNTRY CODE	Qualitative nominale	Variable illustrative
REGION	Qualitative nominale	Variable illustrative
CONTINENT	Qualitative nominale	Variable illustrative
DEMOCRACY	Qualitative ordinale	Variable illustrative
HAPPINESS	Quantitative continue	Variable active
GDP	Quantitative continue	Variable active
UNEMPLOYMENT	Quantitative continue	Variable active
FERTILITY RATE	Quantitative continue	Variable active
HEALTHY LIFE	Quantitative continue	Variable active
SOCIAL SUPPORT	Quantitative continue	Variable active
LIFE CHOICES	Quantitative continue	Variable active
GENEROSITY	Quantitative continue	Variable active
REGULATORY QUALITY	Quantitative continue	Variable active
RULE OF LAW	Quantitative continue	Variable active
POL STABILITY	Quantitative continue	Variable active
PERCEPTIONS OF CORRUPTION	Quantitative continue	Variable active
GOVERNMENT EFFECTIVENESS	Quantitative continue	Variable active
VOICE AND ACCOUNTABILITY	Quantitative continue	Variable active
DEMOCRACY INDEX	Quantitative continue	Variable active
WESTERN EUROPE	Quantitative binaire	Variable illustrative
SUB SAHARIAN AFRICA	Quantitative binaire	Variable illustrative
EAST ASIA	Quantitative binaire	Variable illustrative
NORTH AMERICA AND ANZ	Quantitative binaire	Variable illustrative
DEMOCRATIE COMPLETE	Quantitative binaire	Variable illustrative
REGIME TOTALITAIRE	Quantitative binaire	Variable illustrative

Nous avons donc un total de 15 variables actives pour 11 variables illustratives. Nous avons en effet choisi de considérer les variables binaires comme illustratives, et donc de ne construire nos composantes principales à partir des variables quantitatives continues uniquement.

b) Inertie et dimensions retenues

Afin de poursuivre notre analyse, il est nécessaire de déterminer quelle part de l'information est expliquée par chaque dimension, et ainsi de choisir d'en retenir plus ou moins en fonction de ce critère. Le pourcentage d'inertie représente en effet la quantité d'information recueillie par un axe principal. Plus ce pourcentage est élevé, et plus la part de l'information expliquée par l'axe concerné est élevée. Suite à une première réalisation de l'ACP, les inerties correspondantes aux dix premières dimensions sont présentées dans le graphique ci-dessous, tandis que le tableau suivant le complète en associant également à chaque dimension sa valeur propre.



	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	8.16735308	54.4490206	54.44902
comp 2	1.76703834	11.7802556	66.22928
comp 3	1.07166224	7.1444150	73.37369
comp 4	0.90859499	6.0572999	79.43099
comp 5	0.71705130	4.7803420	84.21133
comp 6	0.54902102	3.6601402	87.87147
comp 7	0.45258828	3.0172552	90.88873
comp 8	0.39689125	2.6459417	93.53467

Nous pouvons observer que l'axe 1 explique 54,45% de l'information, tandis que l'axe 2 en explique 11,78%. Cette valeur est de 7,14% pour l'axe 3, et elle décroît davantage pour chaque dimension suivante considérée.

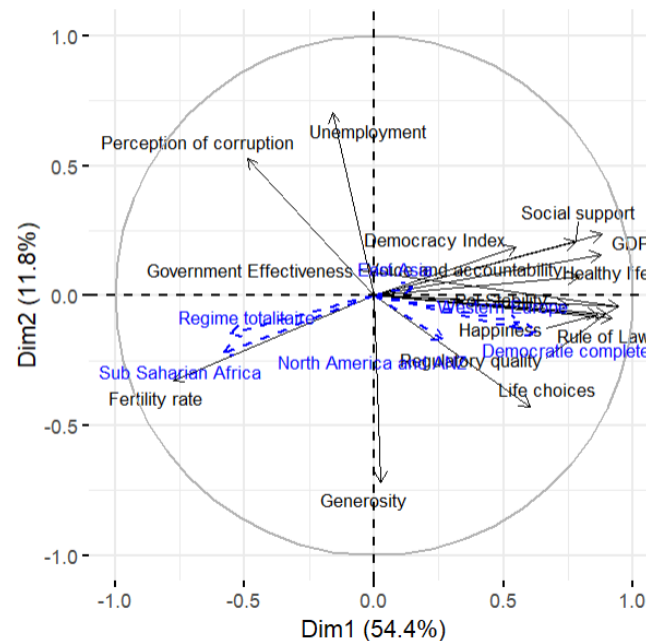
Le tableau nous apporte des informations supplémentaires : les valeurs propres des dimensions 1, 2 et 3 sont supérieures à 1. Selon la règle de Kaiser, ce résultat impliquerait de conserver les trois premiers axes pour notre analyse, et donc d'effectuer l'analyse de trois plans factoriels distincts : Axe 1 et axe 2 ; axe 2 et axe 3 ; axe 1 et axe 3. Nous nous focaliserons cependant sur les deux premières dimensions uniquement, soit le premier plan factoriel, qui explique tout de même à lui seul plus de 66% de l'information.

ii. Résultats

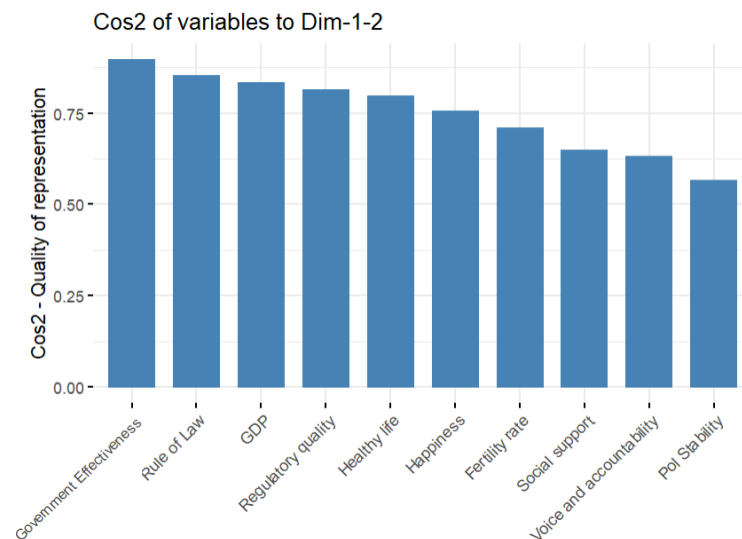
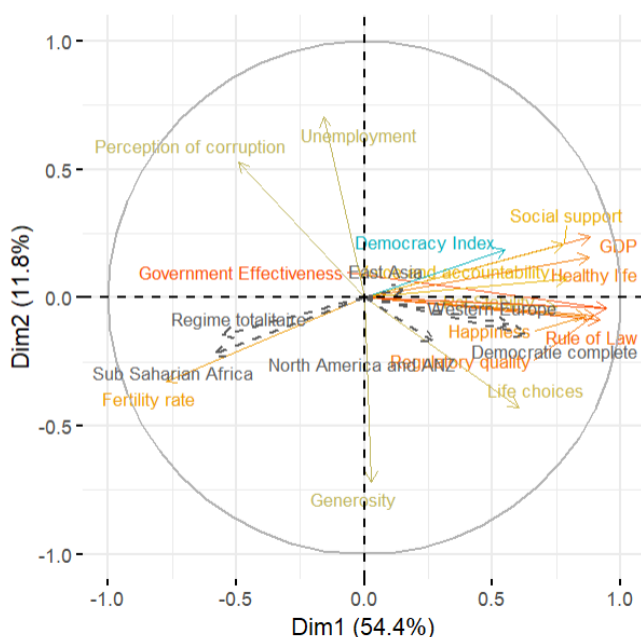
Cette sous-partie a vocation à présenter les principaux résultats issus de l'analyse en composantes principales réalisée sur notre jeu de données. Après avoir procédé à une analyse graphique en mettant notamment en évidence le cercle des corrélations du premier plan factoriel inhérent à nos variables actives, ainsi que les graphiques des individus avec divers habillages les reliant avec les variables du jeu de données, nous utiliserons les éléments précédents ainsi que d'autres mesures pour qualifier de façon précise de chacun de nos deux axes.

a) Analyse graphique

1. Description des variables



Le cercle des corrélations du premier plan factoriel est un cercle de rayon 1. L'axe des abscisses représente le premier axe d'inertie, soit la première dimension, tandis que l'axe des ordonnées représente la seconde dimension. Le pourcentage d'inertie de chaque axe déterminé dans la section précédente lui est également attribué sur ce graphique. Les variables sont représentées par des flèches dont l'origine est le centre du cercle, et plus ou moins proches du bord du cercle. Cette proximité exprime la qualité de représentation du point sur les deux axes considérés. Ainsi, plus une flèche est proche du bord du cercle, et plus la variable associée est bien représentée sur le plan factoriel en question.



Pour une meilleure lisibilité du fait du nombre élevé de variables, nous pouvons mettre en évidence les variables les mieux représentées sur les dimensions 1 et 2 en les colorant en fonction de leur qualité de représentation, ou encore en classant les variables actives par qualité de représentation décroissante. Ces deux idées sont mises en pratique par le biais des deux graphiques ci-dessus, et nous permettent de faire plusieurs observations concernant les variables sur le premier plan factoriel.

Tout d'abord, l'axe 1 est construit sur une opposition entre niveau de bonheur, support social, PIB/habitant, espérance de vie en bonne santé, État de droit, et efficacité du gouvernement d'un côté (coordonnées positives), et perception de la corruption et taux de fécondité (coordonnées négatives). Ainsi, les pays dotés d'un niveau de bonheur et d'un PIB par habitant élevés, dont le gouvernement est jugé efficace et dont les habitants ont une haute espérance de vie en bonne santé à la naissance, se sentent supportés par leur proches et ont confiance en les règles qui régissent la société, sont opposés aux pays où le taux de fécondité est élevé et dont les citoyens ont le sentiment que la corruption est généralisée au sein du gouvernement et/ou des entreprises.

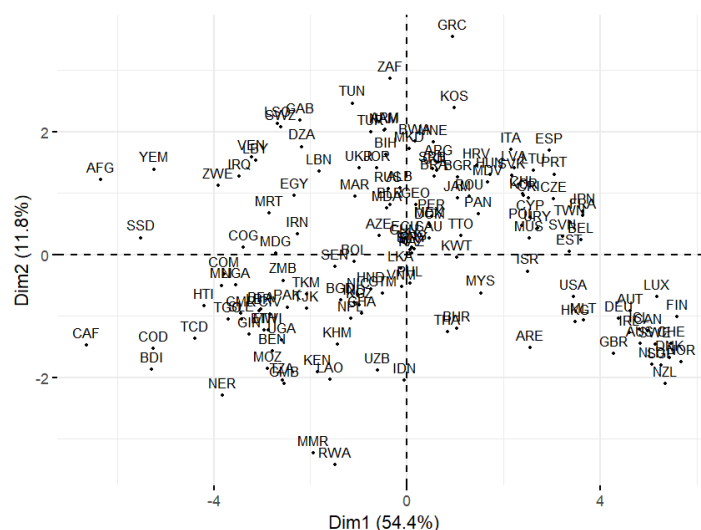
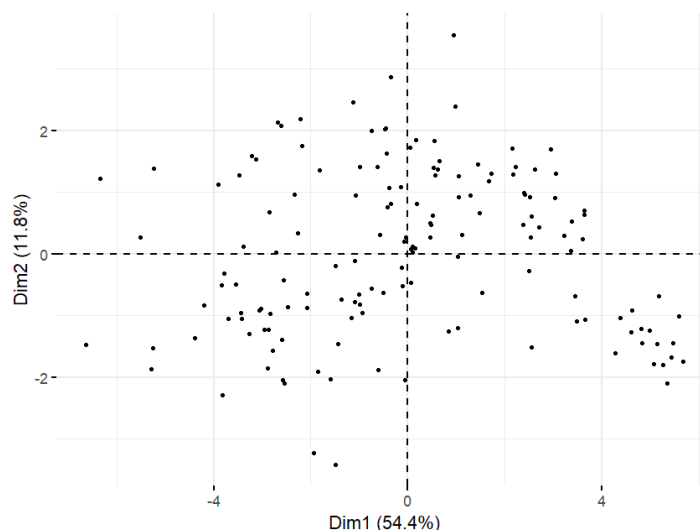
L'axe 2 est pour sa part principalement marqué sur une opposition entre taux de chômage et perception de la corruption d'une part, et générosité d'autre part. Ainsi, les pays dont le taux de chômage est élevé et dont les habitants ressentent que de la corruption émane de leur gouvernement et/ou des entreprises, sont opposés aux pays dans lesquels les citoyens font régulièrement des dons aux associations caritatives.

De plus, les variables les mieux représentées sur le premier plan factoriel sont respectivement l'efficacité du gouvernement, l'État de droit, le logarithme du PIB/hab en PPA, la qualité de la réglementation, l'espérance de vie en bonne santé à la naissance, le niveau de bonheur, et ainsi de suite conformément au diagramme à barres ci-dessus. Nous pouvons de ce fait remarquer que les variables qui sont les mieux représentées sur le premier plan factoriel (axe 1 et axe 2) sont également celles qui sont fortement corrélées positivement avec l'axe 1.

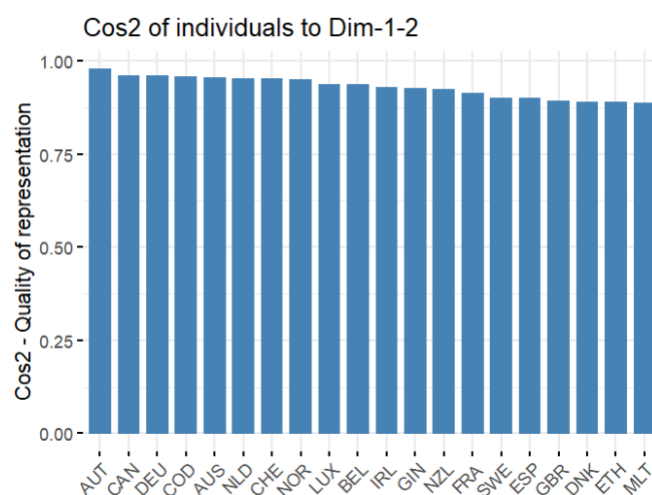
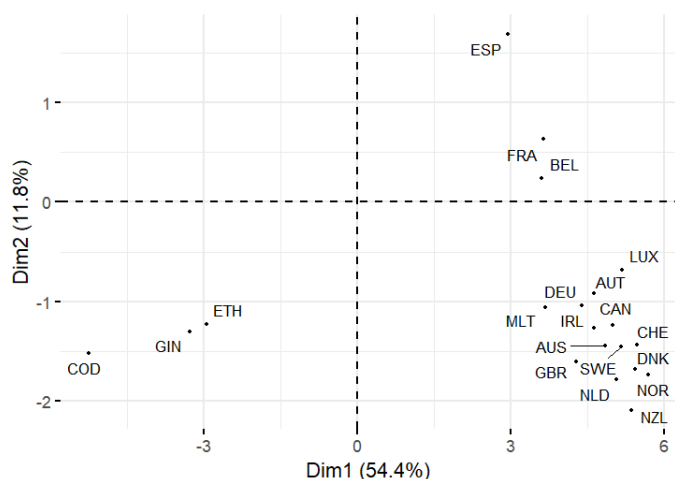
Enfin, si l'on s'intéresse aux variables illustratives, nous pouvons constater que l'Europe de l'Ouest, l'Amérique du Nord et ANZ ainsi que les démocraties complètes sont corrélées positivement à l'axe 1, tandis que l'Afrique Sub-saharienne et les démocraties totalitaires sont corrélées négativement à l'axe 1. Tout cela nous permet de formuler des intuitions quant à la répartition des différents pays sur le premier plan factoriel. En effet, l'opposition marquée sur la dimension 1 laisse penser que les pays développés auront tendance à se situer davantage du côté droit du graphique, du fait de la corrélation positive de l'axe 1 à des variables relatives au bon développement d'un pays. Cependant, l'opposition caractéristique de la dimension 2 est plutôt ambiguë et ne permet pas nécessairement de spécifier la position des pays par niveau de développement ou par position géographique.

2. Description des individus

Le graphique des individus intègre les différents individus, ici les pays, à la représentation du premier plan factoriel. Le premier graphique ci-dessous permet de visualiser de façon précise l'emplacement de tous les pays (chaque point représentant un pays) sans être gêné par un quelconque label, tandis que le second graphique permet de les identifier.

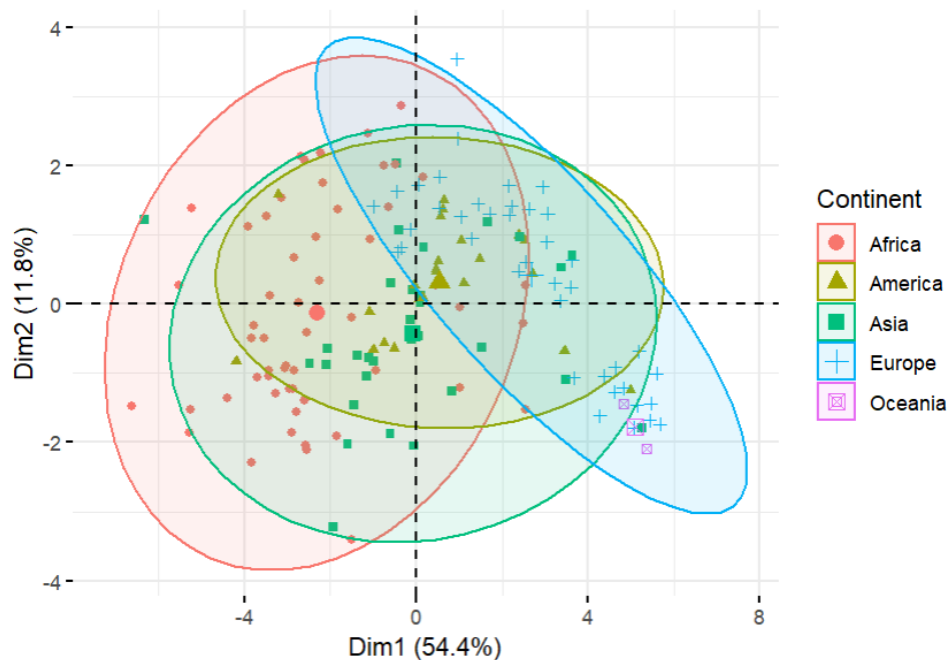


Ces deux représentations nous permettent notamment de constater la présence d'un petit regroupement de quinze pays à droite du graphique, relativement concentrés et surtout détachés du reste des pays, composé entre-autres de la Finlande, du Luxembourg, de la Nouvelle-Zélande, de la Grande-Bretagne, du Danemark ou encore de la Norvège.



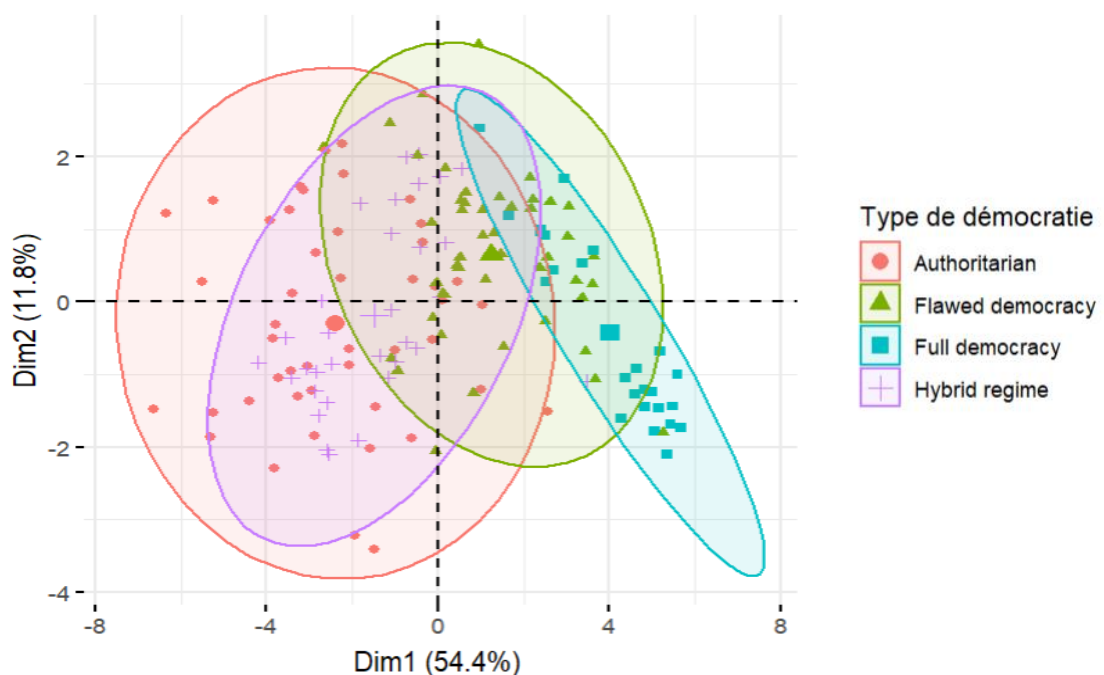
Le graphique ci-dessus est uniquement composé des vingt pays les mieux représentés par le premier plan factoriel. Le diagramme à barres complète cette information en les classant par qualité de représentation décroissante. Ainsi, le pays le mieux représenté par les dimensions 1 et 2 est l'Autriche (AUT), suivi du Canada (CAN), de l'Allemagne (DEU), du Congo (COD), de l'Australie (AUS), et ainsi de suite. Nous retrouvons également la petite concentration de pays dans la partie droite du graphique mise en évidence précédemment.

Pour mieux cerner la disposition des individus sur le premier plan factoriel, et avant de procéder à des habillages en fonction des variables actives quantitatives de notre jeu de données, nous pouvons chercher à voir comment sont répartis les pays en fonction de leur proximité géographique et démocratique. Nous allons ainsi, et dans un premier temps, les classer par continents puis par type de démocratie.



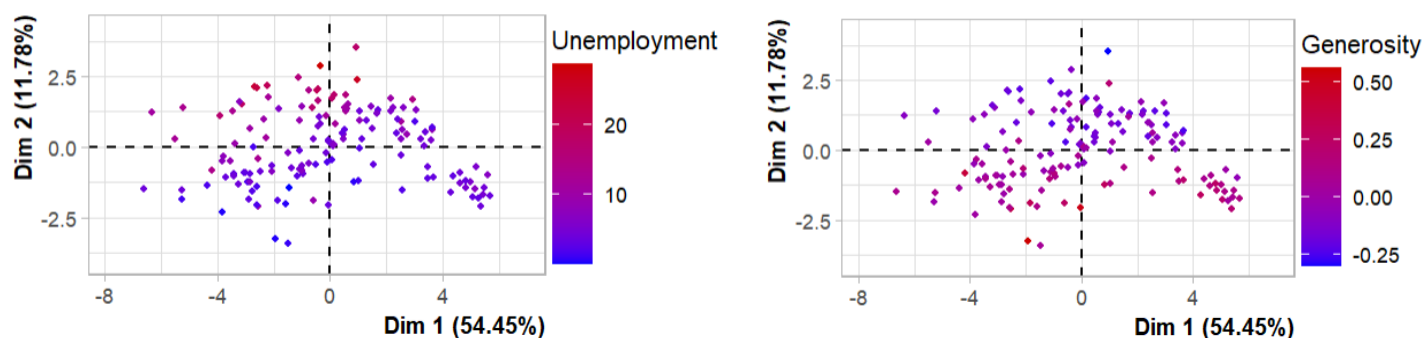
Le graphique ci-dessus consiste en un habillage des individus en fonction de la variable continent. Les pays sont représentés par un symbole correspondant au continent auquel ils appartiennent, l'on notera également la présence d'un point moyen pour chaque continent dont le symbole est plus gros que ceux des individus. Des ellipses de concentration permettent de mieux rendre compte de la possible proximité ou au contraire de la dispersion de pays appartenant au même continent.

Nous pouvons notamment observer que les deux pays d'Océanie et la majorité des pays du continent européen sont situés dans la partie droite du graphique, et peuvent donc être opposés aux pays africains qui sont pour la plupart majoritairement situés dans la partie gauche du graphique. Les pays d'Amérique et d'Asie sont situés de part et d'autre du graphique horizontalement parlant, mais les ellipses de concentration nous permettent de constater que les pays d'Amérique sont relativement plus homogènes que les pays asiatiques. L'on notera également que les points moyens de ces continents sont situés à proximité de l'origine du graphique : il n'est donc pas possible d'établir une stricte corrélation entre l'ensemble des pays constitutifs de ce continent et l'une ou l'autre dimension, ce qui s'explique par une plus ou moins grande hétérogénéité des pays au sein de ces continents.

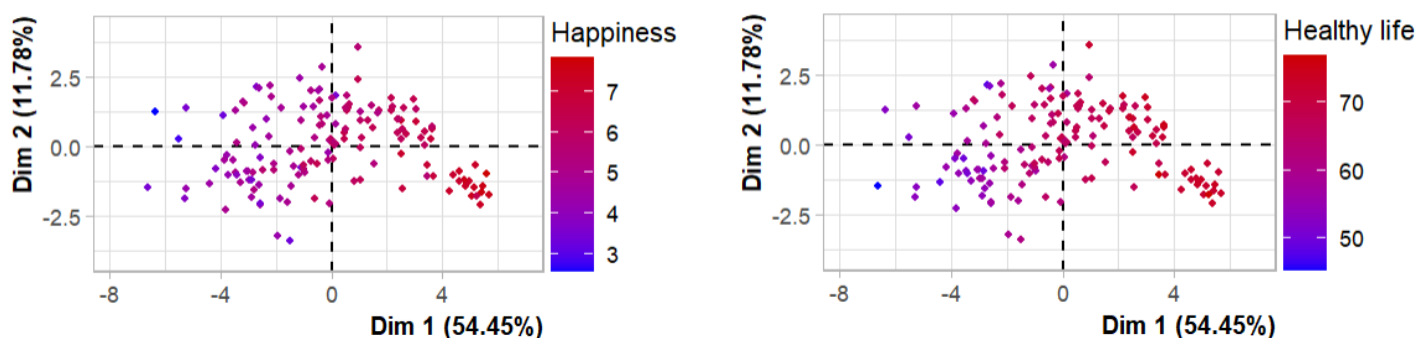


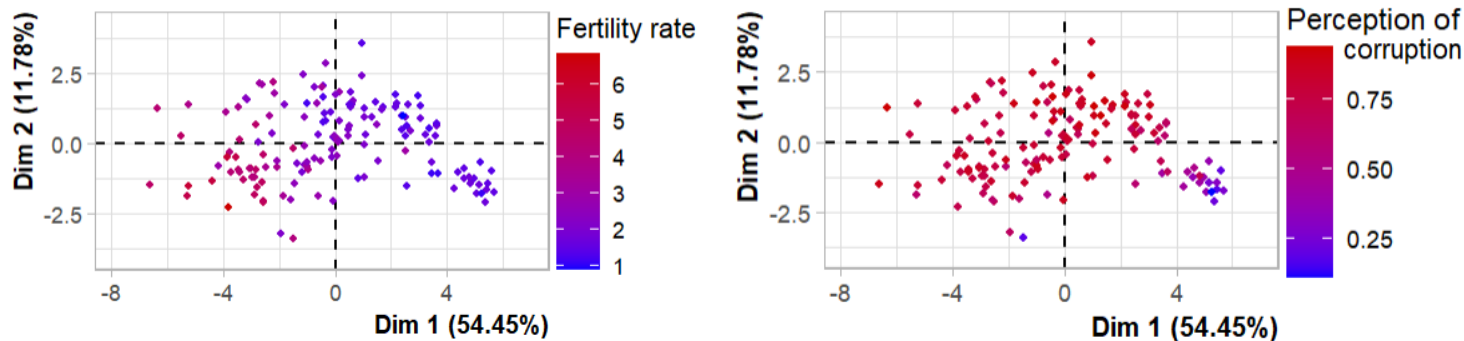
Le graphique ci-dessus concerne cette fois-ci les différents types de démocraties. Dans un premier temps, la forme de l'ellipse de concentration des démocraties complètes n'est pas sans rappeler celle des pays d'Europe évoquée ci-dessus. Rappelons que 56% des démocraties complètes du monde entier se trouvent sur le continent européen, ce qui permet d'expliquer cette similarité. Nous pouvons également constater que l'intégralité des pays considérés comme des démocraties complètes sont situés strictement dans la partie droite du graphique, tandis que la grande majorité des pays en démocratie imparfaite le sont également. Pour leur part, la quasi-totalité des régimes hybrides et régimes totalitaires sont au contraire situés dans la partie gauche du graphique, ce qui les oppose tout naturellement avec les deux autres types de démocraties (complète et imparfaite). Les démocraties complètes semblent être les plus homogènes, tandis que les pays considérés comme des régimes autoritaires ont des profils moins similaires entre eux.

Après avoir catégorisé les pays en fonction de leur emplacement géographique, nous pouvons poursuivre avec un l'habillage de nos individus en fonction de certaines de nos variables actives. Cela nous permettra de distinguer les pays selon d'autres caractéristiques, cela complétant les résultats déjà obtenus lors de l'analyse du cercle des corrélations.



Ces deux graphiques concernent respectivement le taux de chômage et la générosité. Rappelons que la dimension 2 se construit notamment par le biais de l'opposition entre ces deux variables. Nous pouvons en effet constater que plus le taux de chômage d'un pays est élevé, et plus il aura tendance à se trouver dans la partie haute du graphique : le point le plus élevé est la Grèce dont le taux de chômage s'élevait à 17.31% pour l'année 2019, quand les deux points les plus bas sont le Rwanda (0.99%) et Myanmar (0.50%). Si l'on considère à présent la générosité, l'opposition est moins claire et moins marquée même si les pays les plus généreux ont tendance à se trouver dans la partie basse du graphique. Ainsi, et conformément à l'analyse de l'axe 2, ces deux graphiques nous permettent de constater que les pays dont le taux de chômage est faible sont également les pays dont la population est généreuse envers les associations caritatives.





Les quatre graphiques ci-dessus représentent de façon respective le bonheur, l'espérance de vie en bonne santé à la naissance, le taux de fécondité et les perceptions de la corruption. Les caractéristiques de nos individus nous permettent de confirmer l'analyse de variables actives. En effet, les pays dont le taux de fécondité est élevé sont également ceux qui ont l'espérance de vie en bonne santé à la naissance les plus faibles. Ces pays se situent dans la partie gauche du graphique, et ce sont ainsi également ceux qui ont un niveau de bonheur plus faible que les pays qui se situent dans la partie droite du graphique. On retrouve de ce fait l'opposition mise en évidence lors de la construction de la première dimension.

Lorsque l'on considère les perceptions de la corruption, le constat est sans appel : une grande majorité des citoyens du monde considèrent que la corruption est généralisée au sein de leur gouvernement et/ou des entreprises. L'opposition entre les pays à gauche et à droite du graphique est bien moins notable, cela provenant du fait que cette variable n'est en réalité pas très bien représentée sur le premier plan factoriel. Nous pouvons tout de même constater que la petite concentration de pays développés qui se détache à droite du graphique affiche une perception de la corruption très faible par rapport aux autres pays : ces pays ont également un niveau de bonheur parmi les plus élevés, une haute espérance de vie en bonne santé à la naissance ainsi qu'un taux de fécondité très faible.

b) Description des dimensions

Après avoir présenté le premier plan factoriel, constitué des dimensions 1 et 2, nous pouvons davantage nous concentrer sur chacune des dimensions de façon indépendante, pour mettre en évidence les variables puis les individus qui sont les plus contributifs, les mieux représentés, ou encore dont les coordonnées sur l'axe concerné sont les plus élevées/faibles.

Cela passera par la réalisation de tableaux rassemblant chacun de ces critères, pour une lecture plus intuitive, ainsi que de quelques commentaires associés aux résultats. Les tableaux relatifs aux pays classent ces derniers par ordre décroissant de valeur de coordonnée, qualité de représentation et de contribution à la dimension associée, seuls les cinq premiers de chaque catégorie étant présentés.

1. Première dimension

VARIABLE	COORD.	COS ²	CONTRIB.
HAPPINESS	0.865	0.748	9.162
GDP	0.880	0.774	9.480
UNEMPLOYMENT	-0.161	0.026	0.316
FERTILITY RATE	-0.776	0.602	7.366
HEALTHY LIFE	0.877	0.770	9.428
SOCIAL SUPPORT	0.778	0.605	7.414
LIFE CHOICES	0.604	0.365	4.471
GENEROSITY	0.028	0.001	0.009
REGULATORY QUALITY	0.899	0.808	9.889
RULE OF LAW	0.919	0.844	10.34
POL STABILITY	0.750	0.562	6.886
PERCEPTION OF CORRUPTION	-0.490	0.240	2.945
GOVERNMENT EFFECTIVENESS	0.945	0.894	10.94
VOICE AND ACCOUNTABILITY	0.790	0.625	7.652
DEMOCRACY INDEX	0.550	0.302	3.702

TOP	COORD. (-)	COORD. (+)	COS ²	CONTRIB.
1	CENTRAFRIQUE (CAF)	NORVEGE (NOR)	AUTRICHE (AUT)	CENTRAFRIQUE (CAF)
2	AFGHANISTAN (AFG)	FINLANDE (FIN)	BELGIQUE (BEL)	AFGHANISTAN (AFG)
3	SOUDAN DU SUD (SSD)	SUISSE (CHE)	LUXEMBOURG (LUX)	NORVEGE (NOR)
4	BURUNDI (BDI)	DANEMARK (DNK)	ALLEMAGNE (DEU)	FINLANDE (FIN)
5	REP. DEM. CONGO (COD)	NOUV. ZELANDE (NZL)	CANADA (CAN)	SOUDAN DU SUD (SSD)

Concernant la première dimension, qui explique 54,45% de l'information, on remarque aisément que la variable efficacité du gouvernement est à la fois celle possédant la coordonnée positive la plus élevée sur l'axe 1, mais également celle étant la mieux représentée et celle contribuant le plus à la construction de ce dernier. Elle est suivie par les variables État de droit et qualité de la réglementation, dont la contribution et la qualité de représentation sont également très élevées, ces trois variables faisant par ailleurs partie des six dimensions de la bonne gouvernance au sens de la Banque mondiale.

La variable taux de fécondité est celle possédant la coordonnée négative la plus élevée sur l'axe 1 tout en ayant une qualité de représentation et une contribution élevées, ce qui nous permet de retrouver l'opposition précédemment mise en lumière concernant nos individus. En effet, nous avons pu constater que les pays dont le taux de fécondité est élevé sont opposés à ceux disposant d'une bonne gouvernance, d'un score de bonheur élevé ou encore d'une haute espérance de vie en bonne santé à la naissance. De plus, nous remarquons qu'une variable contribue très peu à la construction de la première dimension : il s'agit de la générosité, qui joue pour sa part un grand rôle dans la construction de l'axe 2 comme évoqué lors de l'analyse du cercle des corrélations.

Enfin, nous constatons que les pays possédant les coordonnées négatives les plus élevées sur l'axe 1 sont tous des pays du continent africain, hormis l'Afghanistan qui se situe en Asie du Sud. Les quatre premiers pays dotés des coordonnées positives les plus élevées sont des pays développés d'Europe de l'Ouest, la nouvelle Zélande se situant en Océanie. La Norvège dispose en effet d'un score de bonheur de 7.45 et d'une espérance de vie en bonne santé à la naissance de l'ordre de 72.3 ans, tandis que ces valeurs ne s'élèvent respectivement qu'à 3.48 points et 45.2 ans pour la Centrafrique. Enfin, les quatre pays les mieux représentés se situent en Europe de l'Ouest, quand les plus contributifs à la construction de la première dimension sont à la fois des pays disposant de fortes coordonnées positives et négatives sur l'axe 1: Centrafrique, Afghanistan, Norvège, Finlande et Soudan du Sud.

2. Deuxième dimension

VARIABLE	COORD.	COS ²	CONTRIB.
HAPPINESS	-0.073	0.005	0.299
GDP	0.238	0.057	3.208
UNEMPLOYMENT	0.704	0.495	28.03
FERTILITY RATE	-0.327	0.107	6.056
HEALTHY LIFE	0.159	0.025	1.433
SOCIAL SUPPORT	0.207	0.043	2.416
LIFE CHOICES	-0.428	0.183	10.37
GENEROSITY	-0.720	0.518	29.31
REGULATORY QUALITY	-0.071	0.005	0.284
RULE OF LAW	-0.087	0.007	0.425
POL STABILITY	-0.035	0.001	0.070
PERCEPTION OF CORRUPTION	0.527	0.278	15.73
GOVERNMENT EFFECTIVENESS	-0.043	0.002	0.104
VOICE AND ACCOUNTABILITY	0.073	0.005	0.300
DEMOCRACY INDEX	0.186	0.035	1.961

	COORD. (-)	COORD. (+)	COS ²	CONTRIB.
1	RWANDA (RWA)	GRECE (GRC)	GRECE (GRC)	GRECE (GRC)
2	MYANMAR (MMR)	AFRIQUE DU SUD (ZAF)	SERBIE (SRB)	RWANDA (RWA)
3	NIGER (NER)	TUNISIE (TUN)	MONTENEGRO (MNE)	MYANMAR (MMR)
4	GAMBIE (GMB)	KOSOVO (KOS)	TUNISIE (TUN)	AFRIQUE DU SUD (ZAF)
5	NOUV. ZELANDE (NZL)	GABON (GAB)	ARMENIE (ARM)	TUNISIE (TUN)

La deuxième dimension explique 11,78% de l'information. Les résultats issus du tableau confirment notre analyse graphique du cercle des corrélations, à savoir que le second axe est construit sur une opposition entre taux de chômage, possédant la coordonnée positive la plus élevée, et générosité, avec la coordonnée négative la plus élevée. La variable générosité est celle qui est le mieux représentée sur l'axe 2, également celle qui contribue le plus à sa construction. C'est également le cas pour le taux de chômage, ainsi que la perception de la corruption et les choix de vie dans une moindre mesure.

Enfin, si l'on s'intéresse plus en détails aux individus, nous pouvons remarquer que parmi les cinq pays possédant les coordonnées négatives les plus élevées, nous retrouvons la Nouvelle Zélande qui était auparavant corrélée positivement avec la première dimension. Ce pays possède un taux de chômage de l'ordre de 4.07%, et une générosité évaluée à 0.192 ; la situant ainsi dans le troisième quartile. La Grèce est le pays possédant la coordonnée négative la plus élevée mais est également le pays le mieux représenté et le plus contributif à la construction de la seconde dimension: son taux de chômage est estimé à 17.31%, sa générosité à -0.301 (la valeur minimale mondiale), et son score de choix de vie à 0.54 qui témoigne d'une relative insatisfaction des citoyens quant à la liberté de choisir ce que qu'ils font de leur vie. Enfin, mis à part la Grèce, les pays les plus contributifs à la construction de la seconde dimensions se situent en Asie (Myanmar) ou sur le continent africain (Rwanda, Afrique du Sud et Tunisie).

V. Régression linéaire multiple

Après avoir présenté et résumé notre jeu de données, puis étudié plus en détail le lien entre les différentes variables qui le composent, la suite de notre étude consistera à déterminer un modèle économétrique destiné à déterminer quelles variables ont un effet significatif sur le bonheur national ainsi qu'à interpréter cet effet. Pour ce faire, et après avoir brièvement présenté la méthode d'analyse utilisée, nous allons déterminer les variables qui composeront le modèle, nous soumettrons ce dernier à plusieurs tests et nous interpréterons les résultats inhérents à la régression effectuée sur notre modèle.

i. Méthodologie

a) Présentation de la méthode d'analyse

Le modèle auquel nous allons recourir a vocation à expliquer une variable quantitative, appelée variable expliquée ou dépendante, à l'aide de plusieurs variables exogènes. Cela implique donc de réaliser sur notre jeu de données une régression linéaire multiple. Cette méthode nous permet d'observer l'impact et la significativité de cet impact des variables explicatives sur la variable expliquée.

Notre base de données est composée de vingt-et-une variables quantitatives, dont six sont des variables quantitatives binaires. La mise en place d'un modèle économétrique nécessite d'opérer une sélection parmi ces variables afin de déterminer lesquelles feront partie du modèle.

De ce fait, et après avoir présenté la variable expliquée de notre modèle linéaire multiple, nous reviendrons sur les différentes méthodes de sélection ayant abouti à la spécification du modèle utilisé dans notre étude.

b) Variable expliquée et variables explicatives retenues

Le modèle économétrique que nous allons mettre en place a pour but de mettre en lumière les déterminants du bonheur national. De ce fait, il est nécessaire de déterminer un indicateur pertinent du bonheur dans les différents pays du monde.

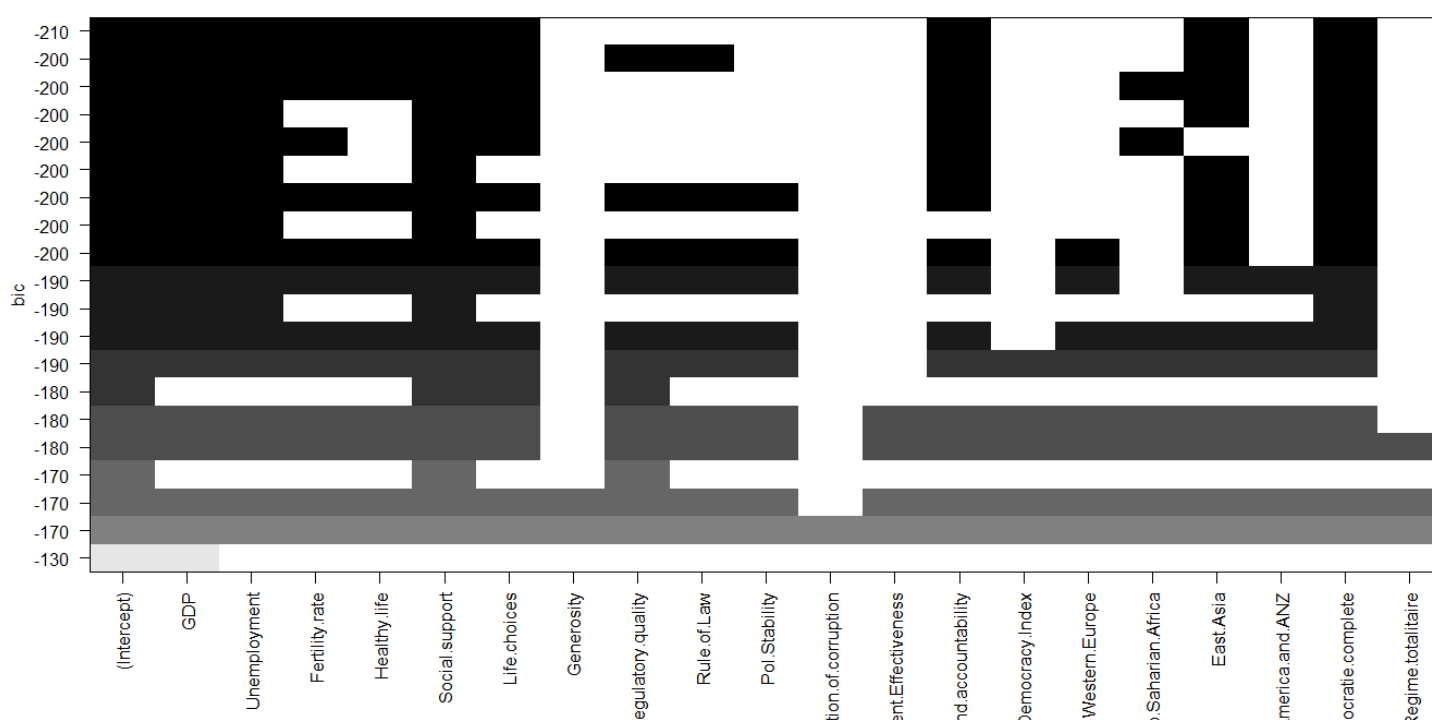
Cet indicateur apparaît dans notre jeu de données comme étant la variable *happiness*. Cette dernière correspond à la réponse moyenne par pays à la question de l'échelle de Cantril, qui invite les gens à évaluer la qualité de leur vie actuelle sur une échelle de 0 à 10. Cette question est extraite d'un sondage sur la mesure du bien-être provenant du Gallup World Poll (GWP), et dont les données ont été recueillies via le *World Happiness Report 2020* (chapitre 2). Ladite question se présente sous la forme suivante : « Imaginez une échelle avec des marches numérotées de 0, en bas, à 10, en haut. Supposons que le haut de l'échelle représente la meilleure vie possible pour vous, et le bas de l'échelle la pire vie possible pour vous. Si la marche la plus haute est 10 et la plus basse 0, sur quelle marche vous placeriez-vous en ce moment même ? ».

Intéressons-nous à présent au choix des variables exogènes qui composeront notre modèle. Les méthodes de sélection de variables explicatives sont très utiles lorsqu'un jeu de données est composé d'un grand nombre de variables et que l'on veut savoir, de prime abord, quelles seront les plus significatives pour ainsi déterminer le modèle qui sera le plus adéquat et performant.

La méthode des sous-ensembles, ou méthode exhaustive, permet d'exprimer sous la forme d'un graphique la contribution de chaque variable pour un critère donné. Nous utiliserons ici le critère du BIC et le critère du R^2 , communément utilisés dans le cadre de cette méthode.

- Sélection via la minimisation du BIC

Le BIC, ou Critère d'information de Bayes, est un critère permettant de mesurer la qualité d'ajustement d'un modèle. Plus la valeur associée à ce critère est faible, et plus le modèle sera bien ajusté.



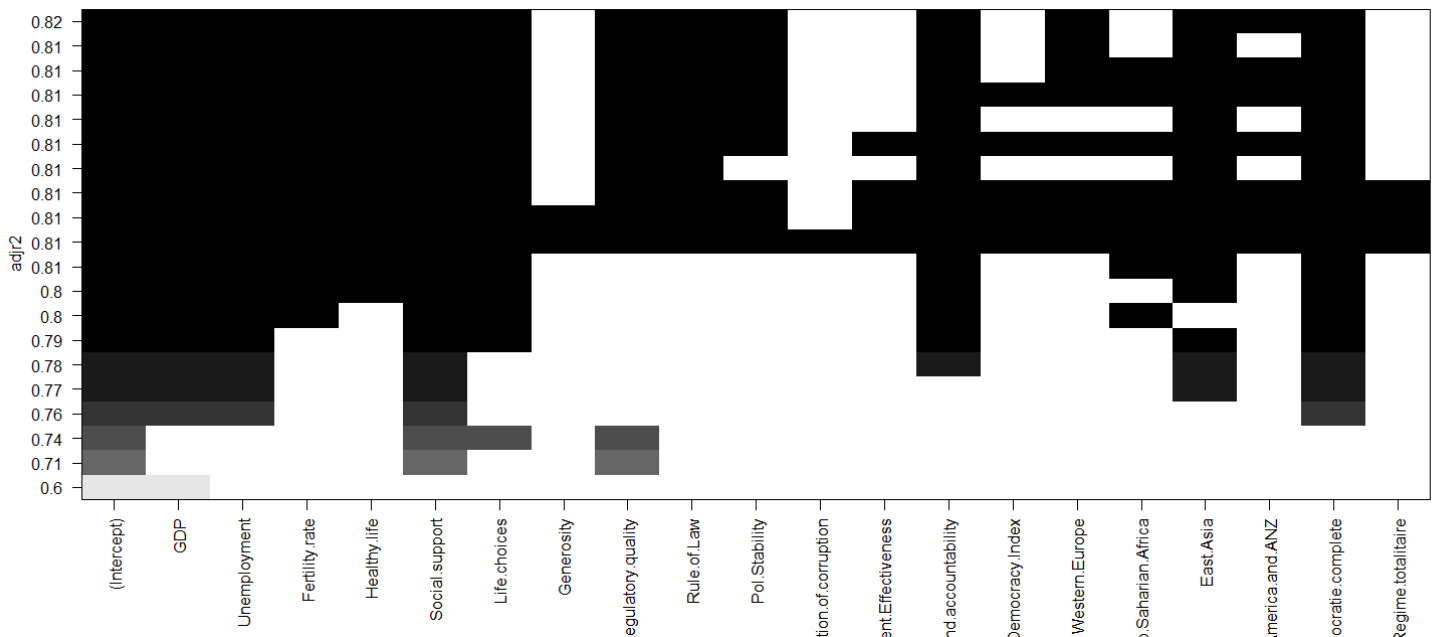
Le graphique ci-dessus est composé de l'ensemble des variables quantitatives de notre base de données, en abscisse, et du critère du BIC en ordonnée. Ainsi, si nous utilisons uniquement la variable *GDP* (en plus de la constante) pour expliquer le niveau de bonheur, nous obtiendrions un BIC de -130. Si nous utilisons les variables *Social support*, *Life choices* et *Regulatory quality*, le BIC serait de -180.

Ainsi, pour savoir quelle est la combinaison de variables qui minimise le BIC, nous devons nous intéresser à la valeur la plus faible sur l'axe des ordonnées, à savoir -210 soit la ligne la plus haute. De ce fait, le modèle le mieux ajusté selon ce critère est composé des variables *GDP*, *Unemployment*, *Fertility rate*, *Healthy life*, *Social support*, *Life choices*, *Voice and accountability*, *East Asia*, *Democratie complete*. Cela mènerait à la spécification suivante :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \varepsilon \end{aligned}$$

- Sélection via la maximisation du R^2

Le R^2 est un indicateur qui mesure le pourcentage de la variance expliquée par un modèle. Il est ainsi compris entre 0 et 1, sachant que plus le R^2 est élevé (proche de 1) et plus notre modèle expliquera une grande part de l'information.



Le graphique ci-dessus est similaire à celui présenté pour le critère BIC, ainsi composé des variables quantitatives de notre jeu de données en abscisse et du R^2 en ordonnée. De la même façon, si nous n'utilisons que la variable GDP pour expliquer le niveau de bonheur, 60% de la variance serait expliquée par ce modèle.

Ainsi pour maximiser le R^2 , nous devrions conserver le modèle suivant, qui permettrait d'expliquer 82% de la variance:

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \beta_{10} \text{Regulatory quality} + \beta_{11} \text{Rule of Law} \\ & + \beta_{12} \text{Pol Stability} + \beta_{13} \text{Western Europe} + \beta_{14} \text{North America and ANZ} + \varepsilon \end{aligned}$$

Nous pouvons constater que le modèle retenu en considérant la maximisation du R^2 comporte quatorze variables, parmi lesquelles l'intégralité des neuf variables retenues avec le critère BIC. Avant de procéder au choix des variables, nous allons cette fois-ci recourir à des méthodes dites « pas à pas ». Ces méthodes sont toutes basées sur l'AIC, ou Critère d'information d'Akaike, qui est une mesure de la qualité d'information du modèle. Plus la valeur de ce critère est faible, et plus le modèle aura une forte qualité explicative.

- Méthode ascendante (forward)

Dans le cadre de cette méthode, l'algorithme part d'un modèle vide uniquement composé de la constante. À chaque pas, une variable est ajoutée au modèle. C'est celle qui permet de réduire au mieux le critère AIC du modèle obtenu. La procédure s'arrête lorsque toutes les

variables sont introduites ou lorsque l'AIC ne décroît plus. Suite à cette méthode, nous obtenons un AIC de -208.06, avec le modèle suivant :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \beta_{10} \text{Regulatory quality} + \beta_{11} \text{Rule of Law} \\ & + \beta_{12} \text{Pol Stability} + \varepsilon \end{aligned}$$

- Méthode ascendante (backward)

Dans le cadre de cette méthode, l'algorithme part d'un modèle composé de toutes les variables du jeu de données. À chaque étape, la variable dont l'élimination conduit à l'AIC le plus faible est supprimée. La procédure s'arrête lorsque l'AIC ne décroît plus.

Cette méthode permet d'obtenir le modèle suivant :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \beta_{10} \text{Regulatory quality} + \beta_{11} \text{Rule of Law} + \varepsilon \end{aligned}$$

- Méthode bidirectionnelle (stepwise)

La méthode bidirectionnelle débute pour sa part comme la méthode ascendante, c'est-à-dire avec un modèle vide en rajoutant une à une les variables. Cependant, l'algorithme introduit une étape d'élimination de variable après chaque étape de sélection, afin de retirer du modèle d'éventuelles variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

Ainsi, nous obtenons le modèle suivant :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \beta_{10} \text{Regulatory quality} + \beta_{11} \text{Rule of Law} + \varepsilon \end{aligned}$$

Modèle retenu :

Après avoir analysé les différentes méthodes de sélection de variables explicatives, nous pouvons constater que si ces dernières ne s'accordent pas toutes sur un modèle spécifique, leurs résultats apparaissent pour le moins similaires et concordants, voire identiques dans le cas de la méthode ascendante et bidirectionnelle. La méthode de maximisation du R^2 tend toutefois à conserver davantage de variables que les autres méthodes de sélection, quand les méthodes se basant sur l'AIC nous conduisent à sélectionner deux à trois variables supplémentaires par rapport au critère BIC.

Ainsi, pour faciliter la lecture de nos résultats ainsi que la sélection des variables pour notre modèle de régression linéaire multiple, le tableau ci-dessous associe chaque variable quantitative aux différentes méthodes de sélection, en fonction de si elle a été retenue pour faire partie du modèle ou non.

	BIC	R ²	Ascendante	Descendante	Bidirectionnelle
GDP					
Unemployment					
Fertility Rate					
Healthy life					
Social Support					
Life Choices					
Generosity					
Regulatory					
Quality					
Rule of Law					
Pol Stability					
Perception of corruption					
Government Effectiveness					
Voice and accountability					
Democracy					
Index					
Western Europe					
Sub Saharian Africa					
East Asia					
North America and ANZ					
Régime totalitaire					
Démocratie complète					

Vis-à-vis des résultats ci-dessus, nous avons décidé de conserver onze variables conformément aux méthodes descendante et bidirectionnelle (critère AIC), ce qui nous amène à la spécification ci-dessous :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Démocratie complète} + \beta_{10} \text{Regulatory quality} + \beta_{11} \text{Rule of Law} + \varepsilon \end{aligned}$$

ii. Tests et résultats

Avant de procéder à l'estimation des coefficients du modèle pour en interpréter les résultats, il convient de confronter ce dernier à plusieurs tests.

Cela nous permettra notamment d'identifier les valeurs atypiques, les problèmes de multicolinéarité, ainsi que de vérifier diverses hypothèses sur les résidus (tests de normalité, d'homoscédasticité et d'autocorrélation des résidus) qui assurent une bonne fiabilité du modèle, ceci étant susceptible d'aboutir à des modifications au sein de notre jeu de données ou du modèle précédemment retenu.

- Test de Rosner

Ce test a pour objectif de détecter si certaines valeurs sont atypiques et pourraient ainsi fausser notre analyse. Ainsi, un ou plusieurs individus pourraient, via des valeurs extrêmes au sein des variables, biaiser les résultats.

Le test de Rosner nous permet d'obtenir un tableau comportant une valeur critique et une valeur test. Si la valeur test est supérieure à la valeur critique, cela veut dire que l'individu est atypique et qu'il doit donc être retiré de la base de données avant d'effectuer la régression. Nous avons ainsi réalisé ce test sur les variables quantitatives contenu du modèle retenu spécifié ci-dessus.

	NUMERO	VALEUR TEST	VALEUR CRITIQUE
GDP	1	2.327204	3.519104
UNEMPLOYMENT	1	3.995965	3.519104
	2	3.758669	3.517009
	3	3.525480	3.514898
	4	3.350329	3.512770
FERTILITY RATE	1	3.344805	3.519104
HEALTHY LIFE	1	2.718603	3.519104
SOCIAL SUPPORT	1	4.000643	3.519104
	2	2.960926	3.517009
LIFE CHOICES	1	3.284110	3.519104
VOICE AND ACCOUNTABILITY	1	2.175401	3.519104
REGULATORY QUALITY	1	2.452995	3.519104
RULE OF LAW	1	2.222959	3.519104

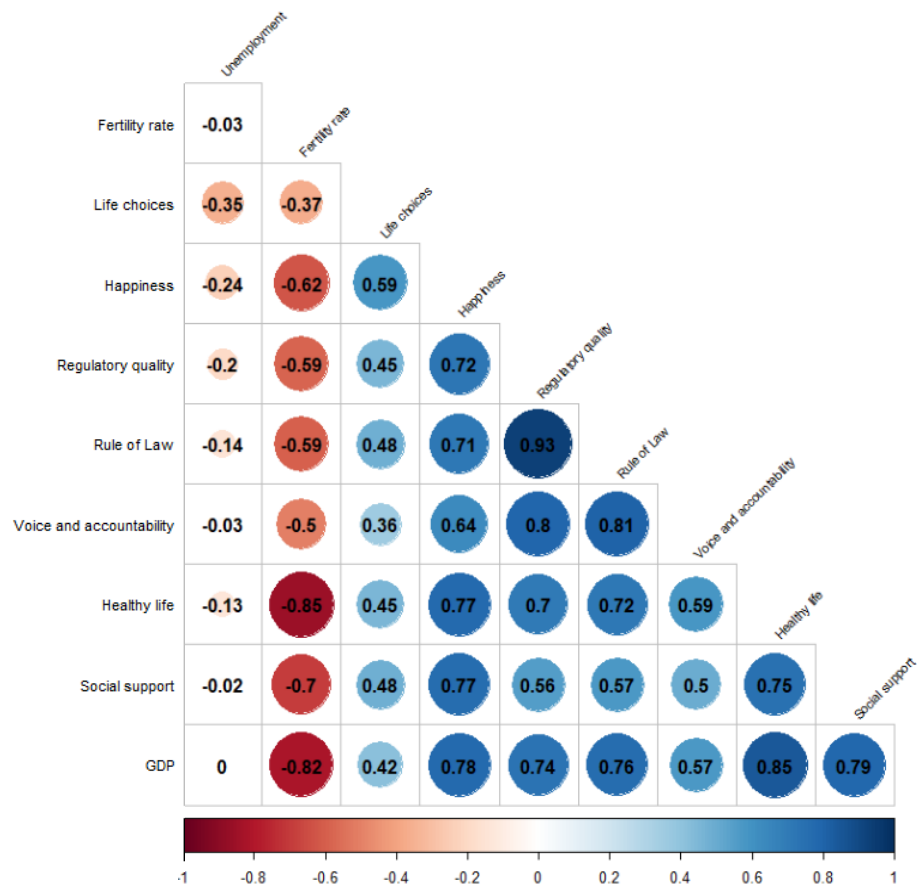
L'analyse de ce tableau révèle la présence de quatre valeurs atypiques dans notre base de données. Concernant le chômage, les pays concernés sont : le Kosovo avec un taux de chômage de 26%, le Lesotho (23,86%) et l'Afrique du Sud (28,47%). Concernant le soutien social, le seul pays concerné est la Centrafrique avec un soutien social de 0,3194. Après suppression de ces individus, le nouveau test de Rosner effectué ne décèle plus aucune valeur atypique dans notre jeu de données.

- Test VIF

Le test VIF, ou Variance inflation factor, est une vérification ex-post pour s'assurer que les corrélations entre les variables explicatives du modèle n'entraînent pas une variance trop forte. Autrement dit, il s'agit d'un test de multicollinéarité.

Une trop forte corrélation positive ou négative entre deux variables a en effet tendance à fortement biaiser les résultats en augmentant fortement le R^2 . Ainsi, si deux variables sont trop fortement corrélées, on ne les supprime pas mais on les estime séparément. Généralement, on ne commence à s'en préoccuper que lorsque la corrélation dépasse |0.7|, et si elle est supérieure à |0.8| on ne préférera pas mettre les deux variables simultanément dans notre modèle.

Le corrélogramme ci-dessous représente graphiquement et plus intuitivement les résultats issus du calcul de la matrice de corrélation des variables du modèle précédemment retenu.



Nous pouvons ainsi constater une trop forte corrélation positive entre

- *Rule of Law* et *Regulatory quality* : 0,93
- *Rule of Law* et *Voice and accountability* : 0,81
- *Regulatory quality* et *Voice and accountability* : 0,8
- *GDP* et *Healthy life* : 0,85

Ainsi qu'une trop forte corrélation négative entre

- *Fertility rate* et *GDP* : -0,82
- *Healthy life* et *Fertility rate* : -0,87

VARIABLE	VALEUR VIF
GDP	7.151829
UNEMPLOYMENT	1.474107
FERTILITY RATE	4.920902
HEALTHY LIFE	6.817978
SOCIAL SUPPORT	3.322355
LIFE CHOICES	1.866644
VOICE AND ACCOUNTABILITY	3.633448
REGULATORY QUALITY	8.881831
RULE OF LAW	11.201346
EAST ASIA	1.157496
DEMOCRATIE COMPLETE	1.839115

L'analyse de la matrice de corrélation ci-dessus est ainsi couplée avec la réalisation d'un test VIF. La lecture du tableau est la suivante: si la valeur du test VIF associée à une variable dépasse 10, nous pouvons considérer qu'elle pose un problème de colinéarité. C'est ici le cas pour la variable *Rule of Law*, que nous avons ainsi décidé de supprimer du modèle initialement retenu. Une fois la variable retirée, le test VIF est réalisé à nouveau et nous n'obtenons que des variables dont le VIF est inférieur à 10.

- Test de Ramsey

Le test de Ramsey est un test de spécification. Il s'agit ici de tester la possible omission d'une ou plusieurs variables pertinentes en introduisant une variable fictive dans le modèle. Le test consiste ainsi à vérifier la significativité de l'effet de la variable fictive introduite sur la variable expliquée. Si la variable fictive n'est pas significative, alors la spécification du modèle est complète, autrement dit il est composé de toutes les variables pertinentes intervenant dans l'explication de la variable endogène.

$$\begin{cases} H_0 : \text{Le modèle est correctement spécifié} \\ H_1 : \text{Le modèle est mal spécifié} \end{cases}$$

La réalisation du test de Ramsey sur notre modèle nous permet d'accepter l'hypothèse nulle et donc de conclure que notre modèle est correctement spécifié.

- Tests sur les résidus

Pour conclure nos tests et avant de passer à l'interprétation des coefficients issus de la régression de notre modèle, il convient de tester trois des principales hypothèses des MCO concernant les résidus. Lesdits tests sont explicités ci-après :

Test d'autocorrélation des résidus, avec $\begin{cases} H_0 : \text{Les résidus ne sont pas autocorrélés} \\ H_1 : \text{Les résidus sont autocorrélés} \end{cases}$

L'autocorrélation des résidus est vérifiée à l'aide du test de Durbin-Watson. La réalisation de ce test nous conduit à accepter H_0 et donc à conclure que les résidus ne sont pas autocorrélés.

Test sur la normalité des résidus, avec $\begin{cases} H_0 : \text{Les résidus suivent une loi normale} \\ H_1 : \text{Les résidus ne suivent pas une loi normale} \end{cases}$

La normalité des résidus est vérifiée à l'aide du test de Shapiro. La réalisation de ce test ne nous permet pas d'accepter l'hypothèse H_0 , et malgré diverses tentatives de transformation appliquées à notre modèle (transformation logarithmique, transformation en racine carrée et transformation de Box-Cox) ne conduisant pas non plus à valider cette hypothèse, nous décidons finalement de poursuivre l'analyse tel quel.

Test d'hétéroscédasticité, avec $\begin{cases} H_0 : \text{Présence d'homoscédasticité} \\ H_1 : \text{Présence d'hétéroscédasticité} \end{cases}$

L'homoscédasticité, autrement dit le fait que la variance des résidus soit constante, est vérifiée à l'aide du test de Breuch-Pagan. La réalisation de ce test ne nous permet pas d'accepter l'hypothèse H_0 , néanmoins une transformation de Box-Cox nous conduit à rejeter l'hypothèse alternative. Cependant, cette transformation a pour contrariété de remettre en question la bonne spécification du modèle (test de Ramsey), nous décidons ainsi une fois de plus de poursuivre notre étude sans apporter une quelconque modification à notre modèle.

Résultats de la régression :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Life choices} + \beta_7 \text{Voice and accountability} + \beta_8 \text{East Asia} \\ & + \beta_9 \text{Democratie complete} + \beta_{10} \text{Regulatory quality} + \varepsilon \end{aligned}$$

VARIABLE	COEFFICIENT
(INTERCEPT)	-3.20114***
GDP	0.30125***
UNEMPLOYMENT	-0.04011***
FERTILITY RATE	0.19527***
HEALTHY LIFE	0.04320***
SOCIAL SUPPORT	2.66074***
LIFE CHOICES	0.87008*
VOICE AND ACCOUNTABILITY	0.14783*
EAST ASIA	-0.61778***
DEMOCRATIE COMPLETE	0.38571***
REGULATORY QUALITY	0.03040
*** p < 0.01 ; ** p < 0.05 ; * p < 0.1	
N = 147 ; R ² ADJ. = 0.8065 ; P-VALUE = 2.2E-16	

Le tableau ci-dessus a été réalisé suite à la régression du modèle linéaire multiple spécifié précédemment. Il recense les valeurs des différents coefficients, leur significativité (symbolisée par une, plusieurs, ou l'absence d'étoiles), ainsi que le nombre d'observations, la p-value du modèle ainsi que le R² ajusté.

Avant de passer à l'interprétation des coefficients, nous pouvons remarquer qu'une des variables de notre modèle, à savoir celle concernant la qualité de la réglementation, ne se révèle pas significative. Cela implique de réaliser une seconde spécification du modèle en retirant la variable *Regulatory quality*. Une fois cela fait, nous nous rendons à présent compte que la variable *Life choices* n'est à présent plus significative non plus. Après l'avoir également retiré du modèle, nous nous retrouvons avec la spécification suivante :

$$\begin{aligned} \text{Happiness} = & \beta_0 + \beta_1 \text{GDP} + \beta_2 \text{Unemployment} + \beta_3 \text{Fertility rate} + \beta_4 \text{Healthy Life} \\ & + \beta_5 \text{Social support} + \beta_6 \text{Voice and accountability} + \beta_7 \text{East Asia} \\ & + \beta_8 \text{Democratie complete} + \varepsilon \end{aligned}$$

VARIABLE	COEFFICIENT
(INTERCEPT)	-2.984622***
GDP	0.319052***
UNEMPLOYMENT	-0.050113***
FERTILITY RATE	0.200649***
HEALTHY LIFE	0.046232***
SOCIAL SUPPORT	2.856148***
VOICE AND ACCOUNTABILITY	0.166688***
EAST ASIA	-0.703935***
DEMOCRATIE COMPLETE	0.421927***
*** p < 0.01 ; ** p < 0.05 ; * p < 0.1	
N = 147 ; R ² ADJ. = 0.8038 ; P-VALUE = 2.2E-16	

$$\begin{aligned} \text{Happiness} = & - 2.985 + 0.319 \cdot \text{GDP} - 0.050 \cdot \text{Unemployment} + 0.201 \cdot \text{Fertility rate} \\ & + 0.046 \cdot \text{Healthy Life} + 2.856 \cdot \text{Social support} + 0.167 \cdot \text{Voice and accountability} \\ & - 0.704 \cdot \text{East Asia} + 0.422 \cdot \text{Democratie complete} + \varepsilon \end{aligned}$$

Nous remarquons qu'à présent, toutes les variables de notre modèle sont significatives (test de Student) au seuil de 1%. La réalisation du test de Fisher nous permet ainsi de rejeter l'hypothèse alternative selon laquelle il existe au moins un coefficient non nul et donc de conclure que notre modèle est globalement significatif. De plus, notre modèle nouvellement spécifié respecte à présent l'hypothèse d'homoscédasticité des résidus, bien qu'il ne respecte toujours pas l'hypothèse de normalité des résidus.

Nous pouvons également constater que notre coefficient de détermination s'élève à 0.804, autrement dit plus de 80% de la variance est expliquée par notre modèle. Ce coefficient est toutefois inférieur à celui des modèles plébiscités par les différentes méthodes de sélection, dans la mesure où les variables qui ont été retirées pour des problèmes de multicollinéarité ou de significativité apportaient tout de même de l'information supplémentaire à notre modèle.

Concernant l'interprétation des coefficients, nous pouvons remarquer qu'une augmentation d'1% du PIB par habitant va entraîner une hausse du score de bonheur de 0,00319 points, toutes choses égales par ailleurs. De plus, vivre dans un pays considéré comme une démocratie complète provoque une augmentation du score de bonheur à hauteur de 0,422 points, TCEPA. De la même façon, si l'espérance de vie en bonne santé à la naissance augmente d'un an, le score de bonheur augmentera de 0,046 points, toutes choses égales par ailleurs. Nous remarquons également que si le taux de chômage augmente de 1%, le score de bonheur baissera de 0,05 points TCEPA, et que vivre dans un pays d'Asie de l'Est provoque une baisse du score de bonheur à hauteur de -0,704 points, toutes choses égales par ailleurs.

Une revue de la littérature nous a permis de constater une certaine conformité de nos résultats vis-à-vis de la littérature déjà existante à ce sujet. En effet, à un point dans le temps au sein d'une société, les plus riches sont en moyenne plus heureux que les plus pauvres, ce qui est favorable à l'hypothèse selon laquelle le revenu influence positivement le bonheur (Easterlin, 1974), et qu'il n'est fait état d'aucune preuve indiquant l'existence d'un point de satiété au-delà duquel le bonheur n'augmente plus consécutivement à une hausse de revenus (Stevenson et Wolfers, 2008). De plus, le bonheur est négativement impacté par le chômage, et l'effet non-pécuniaire est davantage important que l'effet qui découle de la perte de revenu associé (Winkelmann et Winkelmann, 1998). La démocratie influence positivement le bonheur, son effet étant plus marqué dans les pays à tradition démocratique établie (Dorn et al., 2005). Enfin, concernant la santé, la satisfaction de vie des personnes présentant des invalidités est en moyenne plus faible que celle des personnes n'en ayant déclaré aucune, et plus la maladie

entrave le quotidien des individus, et plus leur satisfaction à l'égard de la vie diminue (Mehnert et al., 1990).

Ainsi, dans l'ensemble, nos résultats concordent avec la littérature, et nous retrouvons pour la quasi-totalité des variables des coefficients aux signes attendus quant à leur effet sur le bonheur. Une exception subsiste cependant : il s'agit du taux de fécondité, qui dans les analyses précédentes était corrélé négativement avec le bonheur, sachant que la première dimension de notre analyse en composantes principales était notamment construite sur l'opposition entre ces deux variables. Les résultats de notre régression montrent cependant qu'un enfant par femme supplémentaire augmente le niveau de bonheur de 0,201 points, toutes choses égales par ailleurs. Si nous examinons la littérature déjà existante, certaines études montrent que de façon globale, le niveau de bonheur décroît quand le nombre d'enfants augmente tout en soulignant l'importance de facteurs contextuels et individuels dans cette relation (Margolis et Myrskylä, 2011), tandis que d'autres suggèrent que l'effet des enfants sur le bonheur est positif et croissant avec le nombre d'enfants, toujours en considérant un effet contingent à des caractéristiques individuelles (Angeles, 2010).

VI. Régression sur variables latentes

Lorsque l'on est en présence d'un trop grand nombre de variables explicatives, ou que ces dernières sont sujettes à de la dépendance linéaire, les estimateurs issus de la méthode des moindres carrés ordinaires d'une régression linéaire multiple peuvent ne plus être aussi performants que dans le cadre d'un modèle composé d'un plus faible nombre de paramètres. Pour pallier ce problème, il est possible de réduire le nombre de coefficients associés aux variables explicatives en constituant des composantes, ces dernières étant construites comme des combinaisons linéaires des variables indépendantes de notre jeu de données.

Dans le cadre de cette étude, nous nous intéresserons plus particulièrement à la régression en composantes principales (PCR) et de la régression des moindres carrés partiels (PLSR), deux méthodes de régression sur variables latentes auxquelles nous aurons recours dans la suite de notre analyse.

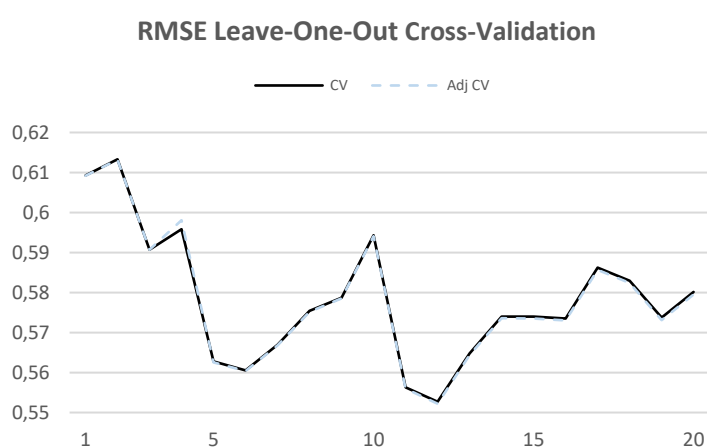
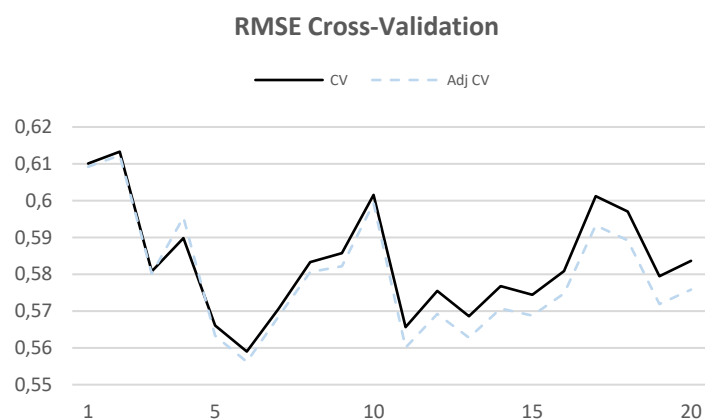
i. PCR

a) Présentation de la méthode d'analyse

La régression en composantes principales est une première alternative à la méthode de régression linéaire multiple (MCO), dont la matrice des variables explicatives est remplacée par une matrice X dont les colonnes sont des variables « artificielles » ainsi obtenues par combinaison linéaire des variables d'origine.

C'est sur cette matrice nouvellement créée par la réalisation d'une analyse en composantes principales (ACP) sur la matrice des variables explicatives, que sera appliquée la méthode de régression linéaire multiple. La méthode PCR cherche plus particulièrement à déterminer les composantes en maximisant la variance de la matrice X .

b) Résultats



Pour réaliser une analyse en composantes principales, il faut en premier lieu déterminer le nombre de composantes qui seront retenues dans le modèle final.

Pour ce faire, nous avons décidé de séparer notre base de données en deux parties de façon randomisée, une partie composée de 80% des données et destinée à construire le modèle prédictif (training set), tandis que les 20% restants auront vocation à évaluer le modèle (testing set).

Le nombre optimal de composantes principales à retenir est sélectionné de sorte à minimiser la RMSE (root-mean-square error). Nous avons comparé deux techniques de validation croisée destinées à trouver le modèle le plus performant, *CV* (k-fold cross-validation) et *LOOCV* (leave-one-out cross-validation) qui est un cas particulier de la validation croisée à k-blocs où à chaque itération l'apprentissage se fait sur n-1 observations et la validation sur l'unique observation restante.

Les deux graphiques ci-dessus nous permettent de constater que la méthode CV (avec k=10 blocs) permet de minimiser le RMSE avec 6 composantes pour une valeur de 0.5590, tandis que la valeur minimale de RMSE via la méthode LOOCV est de 0.5527, atteinte avec 12 composantes soit deux fois plus qu'avec la méthode CV. Pour la suite de notre étude, nous avons choisi d'utiliser un modèle avec méthode de validation croisée CV en retenant 6 composantes.

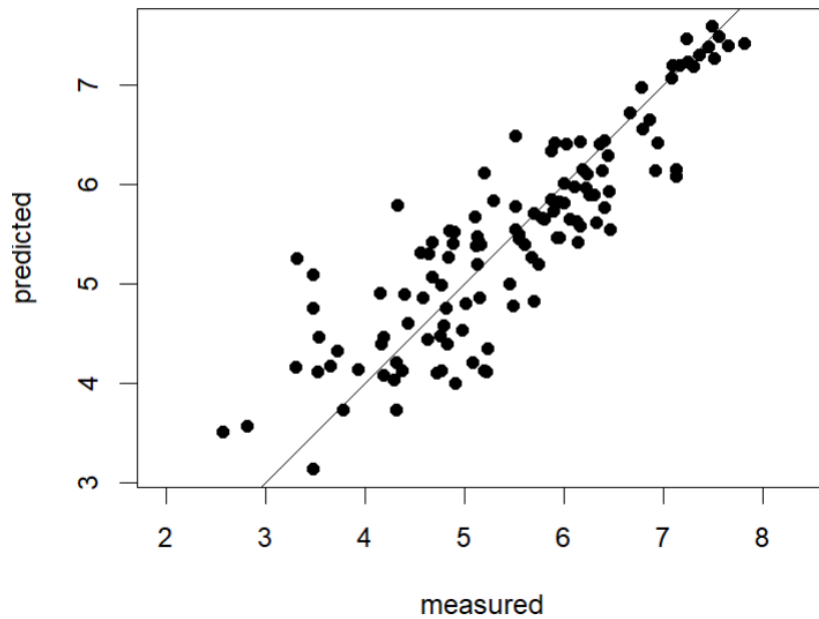
TRAINING: % variance explained

COMP.	1	2	3	4	5	6	...	20
x	45.76	56.16	63.27	69.02	74.69	79.75		100.0
HAPPINESS	72.92	73.01	75.45	76.02	78.31	78.69		84.82

Le tableau ci-dessus présente le montant de la variance cumulée des composantes, soit le vecteur *X*, ainsi que la variance expliquée pour la variable dépendante *Happiness*. Dans ce modèle, nous pourrions expliquer près de 85% de la variance de notre variable expliquée si nous utilisons l'intégralité des 20 composantes.

La méthode CV nous conduit à retenir 6 composantes pour notre modèle, ce dernier étant ainsi capable d'expliquer 78,69% de la variance du niveau de bonheur.

Predicted vs Observed : 6 comps



Le graphique ci-dessus montre les prédictions issues de la validation croisée avec 6 composantes en ordonnée, contre les valeurs réellement observées pour le niveau de bonheur en abscisse. Plus le point est proche de la ligne cible et plus la prédiction est proche de la valeur mesurée, ce qui nous permet ainsi de constater que la grande majorité de nos points suivent la ligne cible de façon relativement correcte.

Nous pouvons cependant remarquer que plus les scores de bonheur observés sont faibles et moins les prédictions associées sont de bonne qualité. En effet, l'on constate plusieurs valeurs relativement voire largement surestimées lors de leur prédiction (notamment un score d'environ 5,2 prédit au lieu d'environ 3,2 réellement observé).

VARIABLE	COEFFICIENT
(INTERCEPT)	2.25150
GDP	0.11555***
UNEMPLOYMENT	-0.0767
FERTILITY RATE	-0.1147***
HEALTHY LIFE	0.13023***
SOCIAL SUPPORT	0.13958***
LIFE CHOICES	0.15749***
GENEROSITY	0.08752*
REGULATORY QUALITY	0.06020**
RULE OF LAW	0.06348***
POL STABILITY	0.04378
PERCEPTION OF CORRUPTION	-0.0859**
GOVERNMENT EFFECTIVENESS	0.08166***
VOICE AND ACCOUNTABILITY	0.02327
DEMOCRACY INDEX	0.01972
WESTERN EUROPE	0.09128***
SUB SAHARIAN AFRICA	-0.0155***
EAST ASIA	-0.0861
NORTH AMERICA AND ANZ	0.02253
DEMOCRATIE COMPLETE	0.04331
REGIME TOTALITAIRE	0.00002

*** p < 0.01 ; ** p < 0.05 ; * p < 0.1

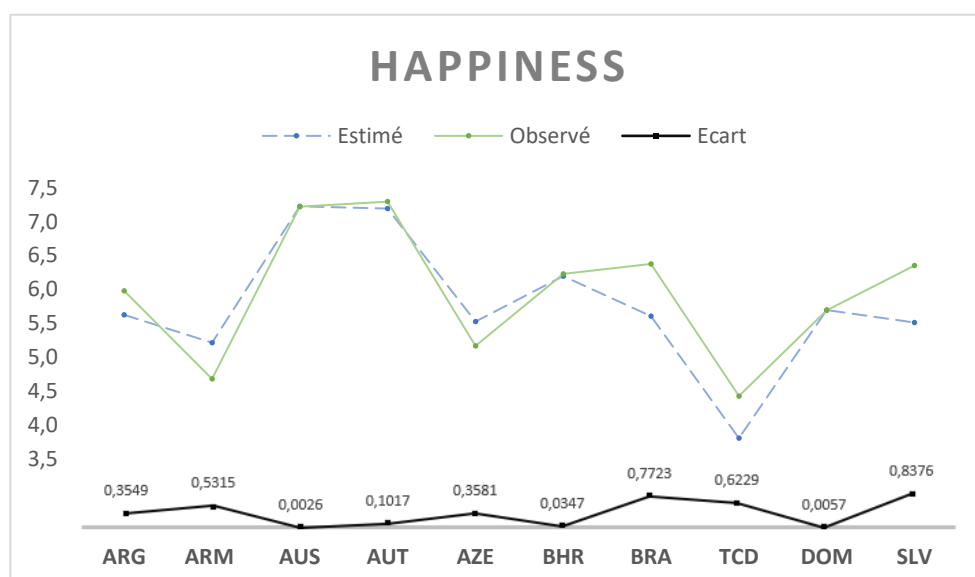
Le tableau ci-dessus recense les coefficients des différentes variables suite à la PCR réalisée avec 6 composantes ainsi que leur significativité. Les signes des coefficients sont dans l'ensemble conformes aux signes attendus ainsi que similaires à ceux obtenus par le biais d'une régression linéaire multiple.

Par ailleurs, des variables comme *Government Effectiveness*, *Western Europe* et *Sub Saharian Africa* qui n'étaient pas retenues à l'issue des différentes méthodes de sélection du modèle linéaire multiple sont cette fois-ci significatives au seuil de 1%. Au contraire, la variable *Unemployment* plébiscitée par l'intégralité des méthodes de sélection auxquelles nous avons eu recours n'est à présent plus significative.

Parmi les variables ayant un effet positif fort sur le niveau de bonheur, on retrouve respectivement les choix de vie, le support social, l'espérance de vie en bonne santé à la naissance et le fait d'être un pays d'Europe. Le taux de fécondité a pour sa part un effet négatif sur le bonheur alors que le coefficient associé à cette variable était positif dans le cadre de la régression linéaire multiple, tandis que la perception de la corruption et le fait d'être un pays d'Afrique Sub-saharienne ont également un effet négatif sur le bonheur.

A présent, nous allons pouvoir tester les capacités prédictives de notre modèle, ce qui nous permettra également de comparer les différentes méthodes de régressions sur variables latentes, à savoir la PCR et la PLS. Pour ce faire, nous allons utiliser le modèle construit par le biais d'un entraînement sur 80% des individus de notre jeu de données afin de le tester sur les 20% restants, à savoir 28 pays.

PAYS	HAPPINESS (ESTIMATE)	HAPPINESS (OBSERVED)	ECART
ARG	5,6198	5,9747	0,3549
ARM	5,2082	4,6768	0,5314
AUS	7,2254	7,2228	0,0026
AUT	7,1925	7,2942	0,1017
AZE	5,5229	5,1648	0,3581
BHR	6,1926	6,2273	0,0347
BRA	5,6033	6,3756	0,7723
TCD	3,7998	4,4227	0,6229
DOM	5,6949	5,6892	0,0057
SLV	5,5107	6,3483	0,8376
SOMME DES ECARTS (28 PAYS)			14,19274



Le tableau ci-dessus comporte les dix premiers pays issus de notre groupe *test*. Cet échantillon n'a pas été pris en compte lors de la création du modèle prédictif, dans la mesure où son rôle est de tester la capacité de notre modèle à prédire le niveau de bonheur national en ayant un point de comparaison existant, autrement dit les valeurs observées pour certains pays du monde qui pourront être comparées avec les valeurs estimées.

Pour une lecture plus intuitive des résultats, nous pouvons également utiliser le graphique issu des données du tableau. Nous pouvons notamment constater que pour les dix premiers pays de l'échantillon *test*, la prédiction la plus qualitative a été réalisée sur l'Australie avec un écart de 0,0026 (7,2228 prédit contre 7,2254 observé). Les plus grands écarts de prédiction concernent le Salvador avec une sous-estimation de 0,838 points environ, puis le Brésil avec une sous-estimation de 0,772 points environ.

Enfin, concernant le groupe *test* dans son intégralité, la somme des écarts entre valeur prédite et valeur observée est d'environ 14.193, soit un écart moyen de l'ordre de 0,5069 points. Ce résultat pourra être comparé à l'issue de la mise en place d'un modèle inhérent à la méthode PLSR (régression sur moindres carrés partiels) qui sera construit et testé à partir des mêmes groupes que le modèle de la méthode PCR.

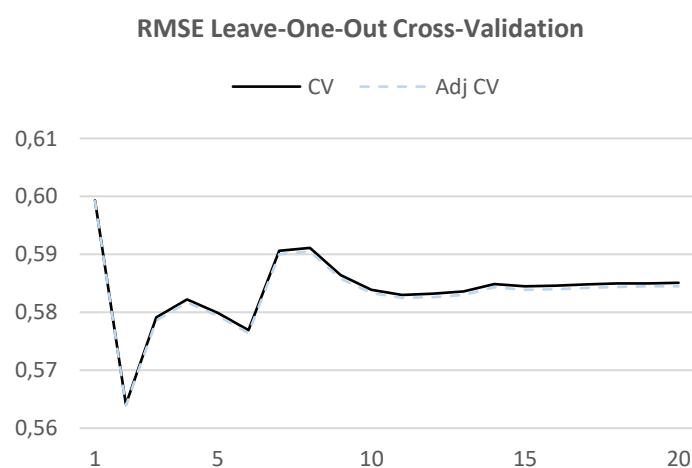
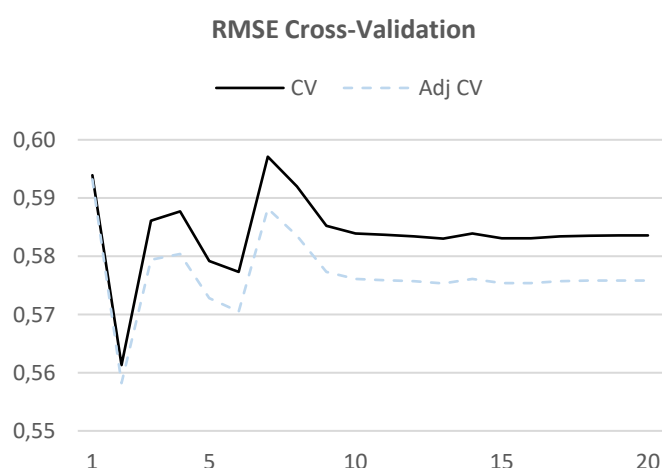
ii. PLSR

a) Présentation de la méthode d'analyse

La régression des moindres carrés partiels est une autre alternative à la méthode de régression linéaire multiple. Là où les dimensions de la méthode PCR sont déterminées en maximisant la variance de la matrice X , la méthode de régression des moindres carrés partiels cherche les axes en maximisant la covariance entre la matrice X et la variable expliquée.

Ainsi, la spécificité de cette méthode est que contrairement à la régression en composantes principales, la PLS considère et utilise la variable dépendante pour la construction des dimensions.

b) Résultats



Pour déterminer le nombre de composantes à retenir dans le modèle final, la méthode est la même que pour la régression en composantes principales, autrement dit via la minimisation du RMSE.

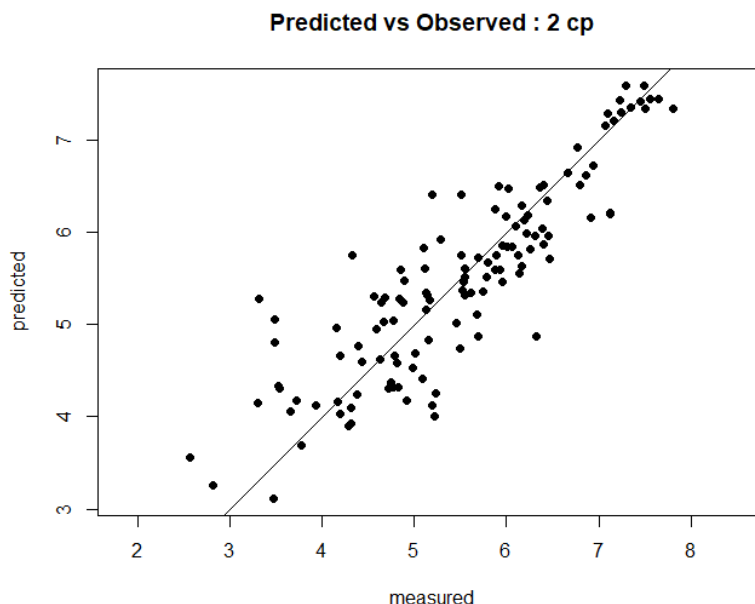
Les deux graphiques ci-dessus nous permettent de constater que les méthodes de validation croisée LOO et CV nous amènent à retenir le même nombre de composantes, soit 2. Nous remarquons toutefois que c'est le modèle avec validation croisée LOO qui permet de minimiser le RMSE à nombre égal de composantes (0.5616 pour la méthode CV contre 0.5591 pour la méthode LOO). Pour la suite de notre étude, nous utiliserons donc un modèle avec validation croisée LOO et 2 composantes.

Nous pouvons ainsi constater que par rapport à la PLS, la PCR a besoin de davantage de composantes pour obtenir un niveau similaire d'erreur de prédiction. En effet, bien que le modèle retenu dans le cadre de la PCR affiche un RMSE plus faible (0.5590), il n'est obtenu qu'avec l'utilisation de six composantes contre deux pour la méthode PLS.

TRAINING: % variance explained

COMP.	1	2	...	20
X	45.69	51.98		100.0
HAPPINESS	74.77	81.20		84.82

Ce tableau, dont la lecture est identique à celui issu de la méthode PCR, permet de mesurer la variance cumulée des composantes ainsi que la variance expliquée pour la variable expliquée. Le modèle à deux composantes retenu nous permet d'expliquer 81.20% de la variance du niveau de bonheur, quand notre modèle PCR à 6 composantes n'en expliquait que 78,69%.

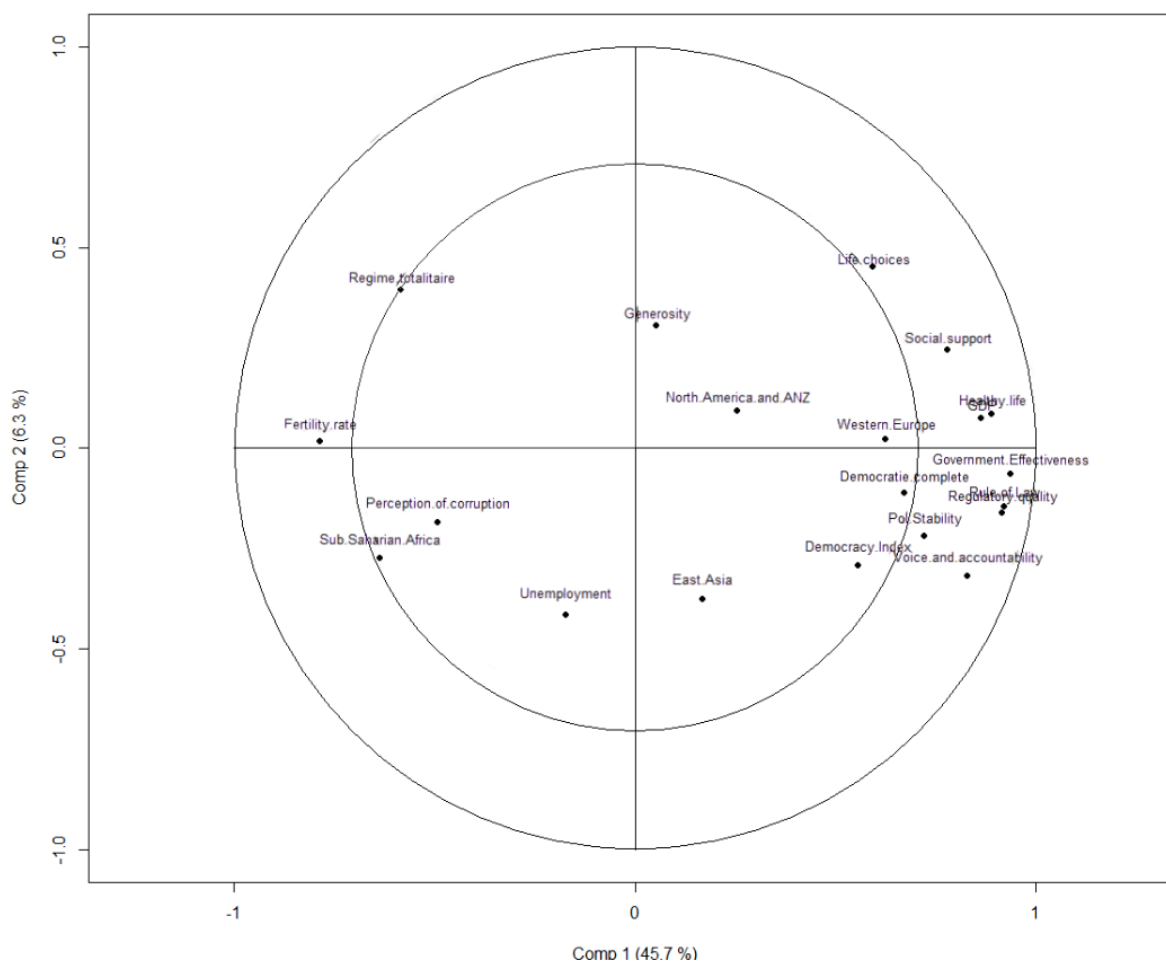


À l'instar de la PCR, nous pouvons également constater que les prédictions issues de notre modèle PLS sont relativement qualitatives. Nous remarquons de la même façon que le nuage de points semble former un « entonnoir » qui se resserre vers le haut, ce qui signifie que les prédictions de niveaux de bonheur élevés sont plus proches des valeurs réellement observées que celles de faible niveaux de bonheur.

- **Loadings**

VARIABLE	COMPOSANTE 1	COMPOSANTE 2
GDP	0,287	
UNEMPLOYMENT		-0,398
FERTILITY RATE	-0,261	
HEALTHY LIFE	0,295	
SOCIAL SUPPORT	0,259	0,235
LIFE CHOICES	0,197	0,432
GENEROSITY		0,292
REGULATORY QUALITY	0,305	-0,155
RULE OF LAW	0,306	-0,140
POL STABILITY	0,240	-0,211
PERCEPTION OF CORRUPTION	-0,164	-0,178
GOVERNMENT EFFECTIVENESS	0,311	
VOICE AND ACCOUNTABILITY	0,275	-0,304
DEMOCRACY INDEX	0,185	-0,281
WESTERN EUROPE	0,207	
SUB SAHARIAN AFRICA	-0,211	-0,263
EAST ASIA		-0,360
NORTH AMERICA		
DEMOCRATIE COMPLETE	0,223	-0,107
REGIME TOTALITAIRE	-0,194	0,378

Le tableau ci-dessus est le tableau des *loadings*, ou chargements. Ces derniers reflètent les corrélations entre les composantes et les variables explicatives, correspondant ainsi aux coordonnées des variables sur les axes factoriels.



Pour interpréter le tableau des *loadings*, nous pouvons le coupler avec la représentation graphique des variables sur les deux premières dimensions. Cela nous permet de mettre en évidence de façon plus intuitive les variables les mieux représentées sur chaque axe, ainsi que les éventuelles oppositions qui caractérisent chaque composante.

L'interprétation des coefficients du tableau des *loadings* se fait par le biais d'une lecture horizontale. De ce fait, nous pouvons par exemple lire : *Regulatory quality* = 0,305 c_1 -0,155 c_2 . Cela signifie que la variable qualité de la réglementation est corrélée positivement à l'axe 1 et négativement à l'axe 2. De plus, nous remarquons que cette variable est mieux expliquée par la composante 1, ce qui est aussi visible graphiquement dans la mesure où plus un point est éloigné de l'origine et proche de l'axe d'une composante (ordonnée ou abscisse), et mieux il sera représenté sur ce dernier.

Nous pouvons ainsi remarquer que la composante 1 est caractérisée par l'opposition entre des variables telles qu'efficacité du gouvernement, qualité de la réglementation, espérance de vie en bonne santé à la naissance et logarithme du PIB/hab en PPA -de façon respective les quatre variables les mieux représentées sur la première composante- et des variables comme taux de fécondité, Afrique Sub-saharienne, régime totalitaire et perception de la corruption. La première composante oppose donc des pays développés (d'Europe notamment) disposant d'une bonne gouvernance au sens de la Banque mondiale, d'une espérance de vie en bonne santé à la naissance et d'un PIB élevés, à des pays dont le taux de fécondité et la perception de la corruption par les citoyens sont élevés, associés à la région Sub-saharienne du continent africain et/ou aux régimes totalitaires.

Concernant la seconde composante, l'on constate une opposition marquée entre générosité, liberté de faire des choix de vie et régimes totalitaires d'une part, et Asie de l'Est, taux de chômage et voix et responsabilité de l'autre. Nous remarquons que vis-à-vis de la première composante, cette opposition semble plus difficile à interpréter voire même contradictoire compte tenu des résultats obtenus précédemment (liberté de faire des choix de vie et régimes totalitaires qui sont corrélés positivement sur cet axe, et opposés à l'Afrique Sub-Saharienne notamment)

- **Weights**

VARIABLE	COMPOSANTE 1	COMPOSANTE 2
GDP	0,304	0,168
UNEMPLOYMENT		-0,334
FERTILITY RATE	-0,244	0,163
HEALTHY LIFE	0,301	
SOCIAL SUPPORT	0,295	0,358
LIFE CHOICES	0,235	0,379
GENEROSITY		0,185
REGULATORY QUALITY	0,292	-0,126
RULE OF LAW	0,288	-0,179
POL STABILITY	0,207	-0,32
PERCEPTION OF CORRUPTION	-0,164	
GOVERNMENT EFFECTIVENESS	0,303	
VOICE AND ACCOUNTABILITY	0,259	-0,163
DEMOCRACY INDEX	0,161	-0,233
WESTERN EUROPE	0,211	
SUB SAHARIAN AFRICA	-0,232	-0,205
EAST ASIA		-0,376
NORTH AMERICA		
DEMOCRATIE COMPLETE	0,223	
REGIME TOTALITAIRE	-0,162	0,319

Le tableau ci-dessus est le tableau des *weights*, ou poids. Les poids correspondent aux coordonnées le long des axes, ces dernières étant proportionnelles aux contributions des variables aux composantes. Ainsi, la lecture se fait de façon verticale, de sorte que la première composante s'interprète de la manière suivante :

$$c_1 = 0,304 * GDP - 0,244 * Fertility Rate + 0,301 * Healthy Life + 0,295 * Social Support + 0,235 * Life choice + 0,292 * Regulatory Quality + 0,288 * Rule of Law + \dots$$

Nous pouvons donc constater que le premier axe est le mieux expliqué positivement par la variable GDP, et le mieux expliqué négativement par le taux de fécondité. Les cinq variables les plus contributrices à la construction de la première composante sont respectivement le logarithme du PIB/hab en PPA, l'efficacité du gouvernement, l'espérance de vie en bonne santé à la naissance, le support social et la qualité de la réglementation, qui sont toutes des variables aux coordonnées positives sur le premier axe. Concernant la seconde composante, elle est la mieux expliquée positivement par la variable choix de vie et négativement par la variable Asie de l'Est.

VARIABLE	COEFFICIENT
(INTERCEPT)	1.155617
GDP	0.156727 ***
UNEMPLOYMENT	-0.126324 **
FERTILITY RATE	-0.042122
HEALTHY LIFE	0.123285 ***
SOCIAL SUPPORT	0.206618 ***
LIFE CHOICES	0.190867 ***
GENEROSITY	0.064716
REGULATORY QUALITY	0.069775 ***
RULE OF LAW	0.053517 **
POL STABILITY	-0.015147
PERCEPTION OF CORRUPTION	-0.059536
GOVERNMENT EFFECTIVENESS	0.087212 ***
VOICE AND ACCOUNTABILITY	0.047520 *
DEMOCRACY INDEX	-0.007384
WESTERN EUROPE	0.085324 ***
SUB SAHARIAN AFRICA	-0.140887 ***
EAST ASIA	-0.098745 **
NORTH AMERICA AND ANZ	0.043884 *
DEMOCRATIE COMPLETE	0.078715 **
REGIME TOTALITAIRE	0.031110

*** p < 0.01 ; ** p < 0.05 ; * p < 0.1

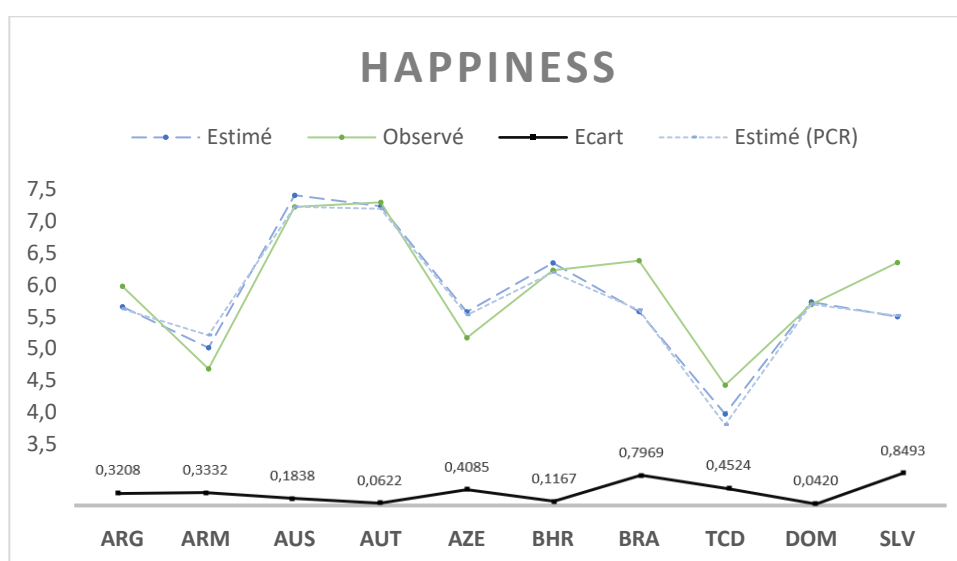
Le tableau ci-dessus comporte les coefficients des variables de notre modèle suite à la PLSR réalisée avec 2 composantes ainsi que leur significativité. De même que pour la PCR, les signes des coefficients sont dans l'ensemble conformes aux signes attendus et similaires à ceux obtenus par le biais d'une régression linéaire multiple.

Les résultats obtenus concernant les coefficients coïncident avec l'analyse précédemment opérée sur les *weights* et les *loadings*. En effet, le support social qui était une variable très bien représentée et très contributive sur les deux composantes est également significative dans notre modèle avec un fort effet positif sur le niveau de bonheur. En plus du support social, des variables telles que la liberté des choix de vie, le logarithme du PIB/hab en PPA, l'espérance de vie en bonne santé, l'efficacité du gouvernement ou encore le fait d'être un pays situé sur le continent européen sont toutes significatives au seuil de 1% et ont un fort effet positif sur le niveau de bonheur national.

D'autre part, le taux de chômage et le fait d'être un pays d'Asie de l'Est ou d'Afrique Sub-saharienne sont des variables significatives au seuil de 5% et ont un effet négatif sur le niveau de bonheur.

Enfin, si l'interprétation des coefficients de la régression ne constitue pas un point de comparaison exploitable entre la PLSR et la PCR, le test des capacités prédictives du modèle nous permettra de comparer les deux méthodes, pour déterminer laquelle des deux aura effectué les prédictions les plus proches de la réalité sur le groupe *test*, après avoir été entraînées toutes les deux sur le même *training set*.

PAYS	HAPPINESS (ESTIMATE)	HAPPINESS (OBSERVED)	ECART
ARG	5,6538	5,9747	0,3208
ARM	5,0100	4,6768	0,3332
AUS	7,4066	7,2228	0,1838
AUT	7,2329	7,2942	0,0622
AZE	5,5733	5,1648	0,4085
BHR	6,3440	6,2273	0,1167
BRA	5,5787	6,3756	0,7969
TCD	3,9703	4,4227	0,4524
DOM	5,7312	5,6892	0,0420
SLV	5,4990	6,3483	0,8493
SOMME DES ECARTS (28 PAYS)			14,33750



Ce tableau et le graphique associé nous permettent de constater que pour les dix premiers pays de l'échantillon test, la prédiction la plus qualitative a été réalisée sur la République Dominicaine avec un écart de 0,0420 (5,7312 prédit contre 5,6892 observé). Tout comme pour la PCR, plus grands écarts de prédiction concernent le Salvador avec une sous-estimation de 0,849 points environ, puis le Brésil avec une sous-estimation de 0,797 points environ.

Enfin, concernant le groupe test dans son intégralité, la somme des écarts entre valeur prédite et valeur observée est d'environ 14.3375, soit un écart moyen de l'ordre de 0,5120 points. Nous pouvons ainsi remarquer que bien que le modèle PLS permettait d'expliquer davantage de variance du niveau bonheur, et affichait un RMSE à peine plus élevé avec bien moins de

composantes incluses, il n'en demeure pas moins que le modèle PCR a été le plus performant concernant les prédictions réalisées sur le groupe test.

METHODE	PCR	PLSR
VALIDATION CROISEE	LOO	CV
NOMBRE DE COMPOSANTES	6	2
RMSE	0.5590	0.5591
VARIANCE EXPLIQUEE DE LA VAR. DEPENDANTE	78.69%	81.20%
MOYENNE DES ECARTS DE LA PREDICTION	0,5069	0,5120

Enfin, le tableau ci-dessus a vocation à rappeler les différences relevées entre les méthodes PCR et PLSR. Ce dernier nous permet de constater que la méthode PLSR nous amène à obtenir des résultats relativement proches voire même meilleurs que ceux issus de la PCR, et ce en n'utilisant que deux composantes au lieu de six pour la régression en composantes principales. Le score plus élevé en termes de variance expliquée du niveau de bonheur avec si peu de composantes n'est pas anodin dans la mesure où les composantes issues de la PLSR sont construites en utilisant la variable expliquée, contrairement à la méthode PCR où ce n'est pas le cas. Néanmoins, comme relevé précédemment, la méthode de régression en composantes principales avec six composantes retenues dans le modèle est celle qui a obtenu l'écart minimum lors des prédictions sur le groupe test avec 0.0051 points d'écarts par rapport à la PLSR.

VII. Conclusion

Arrivés au terme de notre étude, cette conclusion nous permettra de revenir sur les différentes méthodes d'analyses employées, d'en rappeler les principaux résultats obtenus, et d'exposer les principales limites de notre analyse.

Dans un premier temps, le but de notre étude était de rechercher les principaux déterminants du bonheur national. Pour ce faire, nous avons construit une base de données regroupant un indicateur de mesure du niveau de bonheur (Cantril ladder score) ainsi que des variables témoignant d'aspects économiques, démographiques, sociétaux et politiques des 151 pays considérés, susceptibles d'exercer une influence avérée sur le bonheur national.

Nous avons ensuite procédé à la réalisation de statistiques descriptives sur notre base de données dans le but de résumer cette dernière. Cependant, bien que les statistiques descriptives constituent une première vue d'ensemble sur notre jeu de données et nous permettent d'extraire certaines informations sur nos individus ou les corrélations entre les variables, force est de constater que cette méthode n'est pas suffisante pour synthétiser une base de données aussi conséquente.

Pour palier ce problème, nous avons décidé de recourir à la méthode d'analyse en composantes principales (ACP) qui nous permet de réduire la dimensionnalité d'un jeu de données, autrement dit obtenir un résumé le plus pertinent possible des données initiales tel qu'un maximum d'information soit conservé. Nous avons donc analysé les deux premières composantes soit le premier plan factoriel qui explique à lui seul plus de 66% de l'information. Cela nous a permis de mettre en évidence le fait que dans la construction du premier axe

factoriel, le niveau de bonheur était associé au PIB par habitant, à l'efficacité du gouvernement, à l'espérance de vie en bonne santé à la naissance, au support social ou encore à la confiance envers les institutions et les lois, à l'Europe de l'Ouest et aux démocraties complètes, et que toutes ces variables étaient opposées au taux de fécondité, aux perceptions de la corruption, à l'Afrique Sub-saharienne et aux régimes totalitaires.

L'analyse en composantes principales, au même titre que les statistiques descriptives, ne permettent cependant pas de déterminer lesquelles de ces variables ont un effet significatif sur le bonheur, qu'il soit positif ou négatif, ni de construire un modèle économétrique destiné à expliciter les déterminants du bonheur national. Nous avons donc choisi dans un premier temps de construire un modèle de régression linéaire multiple, dans lequel les variables retenues avaient préalablement été sélectionnées par le biais de méthodes d'algorithmes de sélection, puis de comparer les résultats obtenus avec la littérature déjà existante. Nous avons ainsi pu mettre en place un modèle contenant huit variables, à savoir le taux de fécondité, espérance de vie en bonne santé, support social, voix et responsabilité et démocratie complète ayant un effet positif et significatif sur le niveau de bonheur, et le taux de chômage et le fait d'être un pays situé en Asie de l'Est ayant un effet négatif significatif sur le bonheur national. Cela nous permet de constater que bien que le bonheur soit un état d'esprit et un ressenti tout à fait personnel et subjectif, aussi complexe à définir qu'à mesurer, force est de constater que des indicateurs économiques, démographiques, sociétaux et politiques sont à même d'expliquer le niveau de bonheur national.

Pour compléter cette analyse, et palier aux problèmes inhérents à un grand jeu de données (nombre trop élevé de variables explicatives et/ou dépendance linéaires entre ces dernières qui peuvent nuire à la performance des coefficients du modèle), nous avons décidé d'appliquer à notre base de données des méthodes de régression sur variables latentes, plus précisément la régression sur composantes principales (PCR) et la régression des moindres carrés partiels (PLS). La réalisation successive de ces deux méthodes et la mesure de leur performance (à travers la prédiction du niveau de bonheur national pour un groupe test à l'issue d'un entraînement sur un même échantillon d'individus du jeu de données) nous a permis de constater que la méthode PLS permettait d'obtenir des résultats similaires à ceux obtenus à l'issue de la PCR en requérant cependant moins de composantes. La méthode PCR est néanmoins celle ayant permis de minimiser l'écart moyen entre niveau de bonheur prédit par le modèle et niveau de bonheur réellement observé (0,5069 en moyenne en valeur absolue).

Enfin, pour conclure cette étude, il est nécessaire de mettre en évidence les limites inhérentes à certaines méthodologies employées ou aux résultats obtenus à l'issue des différentes analyses. Nous pouvons dans un premier temps rappeler que, même si sans réelle gravité, notre modèle linéaire multiple ne satisfaisait pas l'hypothèse de normalité des résidus, bien que nous puissions nous passer de cette hypothèse dans la mesure où notre base de données comporte un nombre suffisamment élevé d'observations.

Une autre limite à notre modèle de régression linéaire multiple réside dans l'utilisation d'algorithmes de sélection : bien que ces derniers soient utiles pour des jeux de données relativement denses, il est plus rigoureux de se fier à la littérature existante pour sélectionner les variables à inclure au modèle.

Enfin, dans le cadre de l'interprétation des coefficients de la régression linéaire multiple, le signe du coefficient associé au taux de fécondité semble discutable. En effet, l'analyse des corrélations entre les variables ainsi que l'analyse en composantes principales suggéraient une relation d'ordre négative entre niveau de bonheur et taux de fécondité : coefficient de corrélation de -0.62, et nette opposition entre ces deux variables dans la construction du

premier axe factoriel où le taux de fécondité était entre-autres associé aux pays d'Afrique qui sont des pays à niveau de bonheur faible (72,73% des pays de cette région du monde ayant un niveau de bonheur strictement inférieur à 5, contre 5% seulement pour l'Europe notamment). Néanmoins, le coefficient associé au taux de fécondité dans notre modèle de régression linéaire suppose que toutes choses égales par ailleurs, un enfant par femme supplémentaire en moyenne induirait une augmentation de 0.2 points du bonheur national. Si ce résultat n'est pas en adéquation avec les observations réalisées sur les individus de notre base de données, et notamment la répartition des niveaux de bonheur et taux de fécondité par continents, nous pourrions trouver une simple explication à cet effet positif dans le fait que l'arrivée d'un enfant provoque une hausse du bonheur pour la famille concernée. Ainsi, même si les pays développés et majoritairement à haut niveau de bonheur sont caractérisées par une transition démographique achevée et donc un faible taux de fécondité, il n'en demeure pas moins que l'accueil d'un nouveau-né au sein d'une famille est source d'une augmentation du bonheur.

VIII. Bibliographie

« Democracy Index 2020 In sickness and in health », *The Economist Intelligence Unit*, 2021, pp. 8-13.

AGGARWAL Sangeet (2020). What is Cross-Validation ? [en ligne]. Towards Data Science. Disponible sur Internet : What is Cross-Validation?. Also, what are LOOCV and k-Fold... | by Sangeet Aggarwal | Towards Data Science (consulté le 22/05/2022).

Angeles, Luis. Children and Life Satisfaction. *J Happiness Stud*, volume 11, 2010, pp. 523–538.

Dorn, D., Fischer, J.A., Kirchgässner, G. et al. « Is It Culture or Democracy? The Impact of Democracy and Culture on Happiness », *Social Indicators Research*, 82(3), 2007, pp. 505-526.

Easterlin, Richard. *Does Economic Growth Improve the Human Lot ?*, 1974.

KASSAMBARA Alboukadel (2018), Principal Component and Partial Least Squares Regression Essentials [en ligne]. STHDA. Disponible sur internet: Principal Component and Partial Least Squares Regression Essentials - Articles - STHDA (consulté le 22/05/2022).

Margolis Rachel, Myrskylä Mikko, A Global Perspective on Happiness and Fertility. *Population and Development Review*, volume 37, 2011, p. 29-56.

MCNAMARA Amelia, CROUSER Jordan (2016), Principal Components Régressions. Disponible sur Internet : Lab 11 - PCR and PLS Regression in R (smith.edu) (consulté le 22/05/2022).

Mehnert, T., Krauss, H. H., Nadler, R., & Boyd, M. « Correlates of life satisfaction in those with disabling conditions », *Rehabilitation Psychology*, 35(1), 1990, pp. 3–17.

MEVIK Bjørn-Helge, WEHRENS Ron (2021), Introduction to the pls Package. Cran R project. Disponible sur Internet : pls-manual.pdf (r-project.org) (consulté le 22/05/2022).

PREDAL, Christian, et Vincent Vandewalle (2018), « TP R: PCR, PLS Ridge et Lasso », Université de Lille 1: pp 1-8 Principal Component Regression in R [en ligne]. Educational Research Techniques. Disponible sur internet: Principal Component Regression in R | educational research techniques (consulté le 22/05/2022).

ROUVIERE Laurent, Régression en grande dimension. Disponible sur internet : Régression en grande dimension (lrouviere.github.io) (consulté le 22/05/2022)

Stephenson Betsey, et Justin Wolfers. « Economic growth and subjective well-being: Reassessing the Easterlin Paradox », *Brookings Papers on Economic Activity*, Vol. 39, 2008, pp. 1–87.

Winkelmann, Liliana, et Rainer Winkelmann. « Why Are the Unemployed So Unhappy ? Evidence from Panel Data », *Economica*, Vol. 65, 1998, pp. 1-15.

Zach, Principal Components Regression in R (Step-by-Step), Statology. Disponible sur Internet: Principal Components Regression in R (Step-by-Step) (statology.org) (consulté le 22/05/2022)

IX. Annexes

#PCR – CV

VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
CV	1.158	0.6101	0.6133	0.5808	0.5898	0.5661	0.5590	0.5706
adjCV	1.158	0.6092	0.6123	0.5802	0.5955	0.5633	0.5562	0.5685
	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
CV	0.5833	0.5857	0.6016	0.5656	0.5755	0.5686	0.5768	0.5744
adjCV	0.5806	0.5822	0.5993	0.5601	0.5692	0.5628	0.5707	0.5688
	16 comps	17 comps	18 comps	19 comps	20 comps			
CV	0.5809	0.6012	0.5970	0.5795	0.5836			
adjCV	0.5748	0.5933	0.5892	0.5719	0.5758			

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
X	45.76	56.16	63.27	69.02	74.69	79.75	84.49	87.90
Happiness	72.92	73.01	75.45	76.02	78.31	78.69	78.78	78.87
	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps	
X	90.45	92.55	94.18	95.5	96.73	97.73	98.38	
Happiness	79.32	79.39	82.33	82.8	83.00	83.10	83.21	
	16 comps	17 comps	18 comps	19 comps	20 comps			
X	98.91	99.37	99.68	99.87	100.00			
Happiness	83.53	83.79	84.08	84.82	84.82			

#PCR-LOOCV

VALIDATION: RMSEP

Cross-validated using 123 leave-one-out segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
CV	1.158	0.6093	0.6133	0.5907	0.5958	0.5628	0.5605	0.5669
adjCV	1.158	0.6092	0.6132	0.5907	0.5981	0.5625	0.5604	0.5668
	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
CV	0.5754	0.5787	0.5943	0.5563	0.5527	0.5648	0.5740	0.5740
adjCV	0.5752	0.5785	0.5941	0.5560	0.5522	0.5644	0.5736	0.5735
	16 comps	17 comps	18 comps	19 comps	20 comps			
CV	0.5735	0.5862	0.5829	0.5737	0.5801			
adjCV	0.5730	0.5857	0.5824	0.5731	0.5795			

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
X	45.76	56.16	63.27	69.02	74.69	79.75	84.49	87.90
Happiness	72.92	73.01	75.45	76.02	78.31	78.69	78.78	78.87

	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
X	90.45	92.55	94.18	95.5	96.73	97.73	98.38
Happiness	79.32	79.39	82.33	82.8	83.00	83.10	83.21

	16 comps	17 comps	18 comps	19 comps	20 comps
X	98.91	99.37	99.68	99.87	100.00
Happiness	83.53	83.79	84.08	84.82	84.82

#PLS-CV

VALIDATION: RMSEP

Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
CV	1.158	0.6101	0.6133	0.5808	0.5898	0.5661	0.5590	0.5706
adjCV	1.158	0.6092	0.6123	0.5802	0.5955	0.5633	0.5562	0.5685

	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
CV	0.5833	0.5857	0.6016	0.5656	0.5755	0.5686	0.5768	0.5744
adjCV	0.5806	0.5822	0.5993	0.5601	0.5692	0.5628	0.5707	0.5688

	16 comps	17 comps	18 comps	19 comps	20 comps
CV	0.5809	0.6012	0.5970	0.5795	0.5836
adjCV	0.5748	0.5933	0.5892	0.5719	0.5758

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
X	45.76	56.16	63.27	69.02	74.69	79.75	84.49	87.90
Happiness	72.92	73.01	75.45	76.02	78.31	78.69	78.78	78.87

	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
X	90.45	92.55	94.18	95.5	96.73	97.73	98.38
Happiness	79.32	79.39	82.33	82.8	83.00	83.10	83.21

	16 comps	17 comps	18 comps	19 comps	20 comps
X	98.91	99.37	99.68	99.87	100.00
Happiness	83.53	83.79	84.08	84.82	84.82

#PLS-LOOCV

VALIDATION: RMSEP

Cross-validated using 123 leave-one-out segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
CV	1.158	0.6093	0.6133	0.5907	0.5958	0.5628	0.5605	0.5669
adjCV	1.158	0.6092	0.6132	0.5907	0.5981	0.5625	0.5604	0.5668

	8 comps	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
CV	0.5754	0.5787	0.5943	0.5563	0.5527	0.5648	0.5740	0.5740
adjCV	0.5752	0.5785	0.5941	0.5560	0.5522	0.5644	0.5736	0.5735

	16 comps	17 comps	18 comps	19 comps	20 comps
CV	0.5735	0.5862	0.5829	0.5737	0.5801

adjCV 0.5730 0.5857 0.5824 0.5731 0.5795

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps
X	45.76	56.16	63.27	69.02	74.69	79.75	84.49	87.90
Happiness	72.92	73.01	75.45	76.02	78.31	78.69	78.78	78.87

	9 comps	10 comps	11 comps	12 comps	13 comps	14 comps	15 comps
X	90.45	92.55	94.18	95.5	96.73	97.73	98.38
Happiness	79.32	79.39	82.33	82.8	83.00	83.10	83.21

	16 comps	17 comps	18 comps	19 comps	20 comps
X	98.91	99.37	99.68	99.87	100.00
Happiness	83.53	83.79	84.08	84.82	84.82

#PCR : ESTIMATION

PAYS	HAPPINESS (ESTIMATE)	HAPPINESS (OBSERVED)	ECART
ARG	5,619841	5,9747	0,354859
ARM	5,20825	4,6768	0,53145
AUS	7,225401	7,2228	0,002601
AUT	7,192479	7,2942	0,101721
AZE	5,522912	5,1648	0,358112
BHR	6,192623	6,2273	0,034677
BRA	5,603286	6,3756	0,772314
TCD	3,799839	4,4227	0,622861
DOM	5,694868	5,6892	0,005668
SLV	5,510662	6,3483	0,837638
GTM	5,510452	6,3989	0,888448
GIN	4,225124	4,9493	0,724176
IND	5,364369	3,5733	1,791069
LBR	4,288478	4,5579	0,269422
MYS	6,125724	5,3843	0,741424
MLI	3,969472	4,7293	0,759828
MUS	5,970258	6,1013	0,131042
MAR	5,034133	5,0948	0,060667
MMR	5,708952	4,308	1,400952
NAM	4,715389	4,5711	0,144289
PHL	5,63364	6,006	0,37236
SGP	7,261076	6,3771	0,883976
SVK	5,948145	6,2806	0,332455
TJK	5,176386	5,5557	0,379314
TTO	5,892355	6,1919	0,299545
VEN	4,778011	5,0532	0,275189
VNM	5,776807	5,3535	0,423307
ZMB	4,452781	3,7594	0,693381

#PLS : ESTIMATION

PAYS	HAPPINESS (ESTIMATE)	HAPPINESS (OBSERVED)	ECART
ARG	5,653851	5,9747	0,320849
ARM	5,010041	4,6768	0,333241
AUS	7,406564	7,2228	0,183764
AUT	7,231992	7,2942	0,062208
AZE	5,573289	5,1648	0,408489
BHR	6,343981	6,2273	0,116681
BRA	5,578734	6,3756	0,796866
TCO	3,970256	4,4227	0,452444
DOM	5,731222	5,6892	0,042022
SLV	5,498966	6,3483	0,849334
GTM	5,593477	6,3989	0,805423
GIN	4,29073	4,949300	0,658569812
IND	5,254542	3,573300	1,681242115
LBR	4,291219	4,557900	0,266680952
MYS	6,127266	5,384300	0,742965768
MLI	4,25135	5,384300	1,132950232
MUS	6,058756	6,101300	0,042543763
MAR	4,879105	5,094800	0,215694995
MMR	5,735993	4,308000	1,427992912
NAM	4,663287	4,571100	0,092186765
PHL	5,741106	6,006000	0,264894042
SGP	7,132003	6,377100	0,754903009
SVK	5,959697	6,280600	0,320903071
TJK	5,23332	5,555700	0,322379825
TTO	5,959025	6,191900	0,232874776
VEN	4,874157	5,053200	0,179042768
VNM	5,862969	5,353500	0,509469111
ZMB	4,42009	3,299200	1,120889942