École d'ingénieurs du numérique

Analyse de données

Projet final

2018 - 2019

Franck Nguyen
Merlin Rousseau
Guillaume Dupont
Maxence Vercauteren
Thibault Van De Walle

1. Explication de notre approche

Le jeu de donnée que nous avons choisi est : WDR2011 Dataset. Dans ce jeu de donnée issue du group The World Bank on y retrouve de nombreuses variables permettant de décrire les conflits, la sécurité, la politique, le développement ou encore la situation socio-économique de certains pays.

Ainsi, on y retrouve 211 pays avec des variables les décrivant cependant il y a de nombreuses valeurs manquantes et avant de mettre en place notre études nous allons devoir traiter ces valeurs.

L'une des variables notables est la présence de la variable **year** qui couvre la période 1960 à 2009. Sa présence nous indique qu'avec ce jeu de donnée nous pouvons réaliser une étude temporelle.

Cependant, commencer directement avec une analyse temporelle nous paraissait compliqué ce qui nous a poussé à d'abord faire une étude descriptive du jeu de donnée.

Pour ce faire, nous avons donc mis en place une stratégie d'analyse :

- Réduire la dimension de notre jeu donné, cela passe par la catégorisation des variables, un choix subjectif de notre part et la non prise en compte de la variable **year**

- Les pays étant déjà grouper géographiquement grâce à la variable **regionA** nous allons essayer de pousser plus loin le clustering (Par exemple : pays riche/corrompu, pays pauvre/corrompu etc ...)

Concernant les valeurs de la variable **regionA** celles-ci sont :
EAP : Asie de l'Est et du Pacifique
ECA : Europe / Asie Central
LAC : Amérique Latine/ Caraïbe
MNA : Afrique du Nord et du centre-est
NAM : Amérique du Nord
SAR : Région d'Asie du Sud
SSA : Afrique Sud – Saharienne

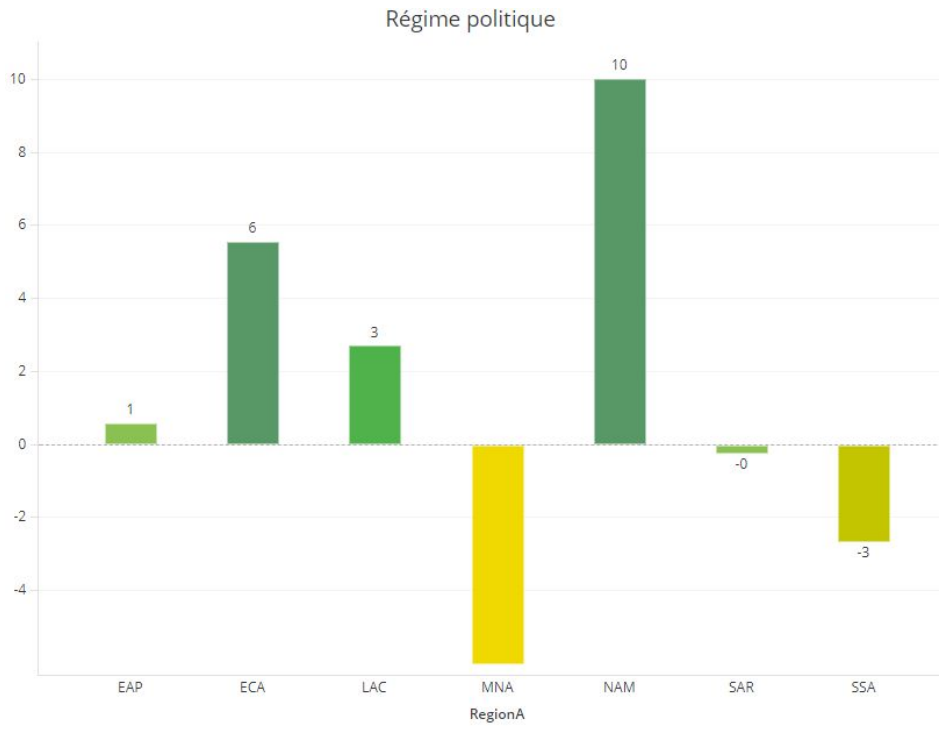
- Pour donner suite à cela, nous allons faire une analyse sur les guerres, les morts, les conséquences extérieurs et intérieurs afin de voir s'il existe des corrélations
- On cherchera ensuite les facteurs potentiels (religion, présence d'ethnies/racisme, langues) expliquant ces phénomènes

Enfin nous finirons par une étude temporelle du jeu donnée qui prendra en compte la variable **year**. On va chercher dans cette partie à déterminer des valeurs inconnues d'une de nos variables à l'aide d'un modèle de régression linéaire et prédire son évolution si c'est possible avec un modèle tel que ARIMA. Pour cela nous devrons réaliser des études sur la corrélation des différentes variables afin de choisir la meilleure. Pour le choix des pays nous prendrons les pays qui ressortiront le plus de notre analyse descriptive.

2. Analyse Descriptive

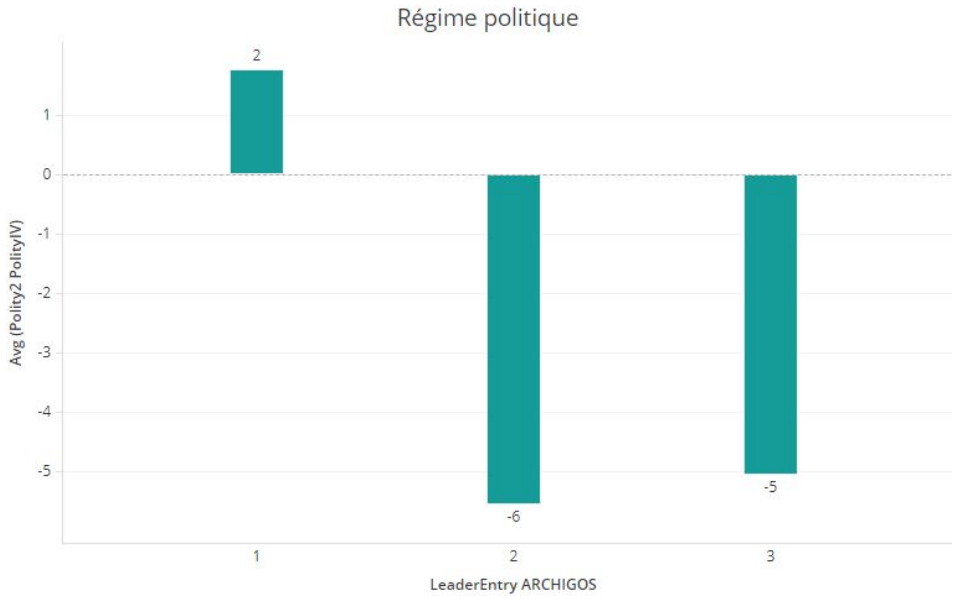
a. Une étude Politique

Pour cette première étude, nous allons nous intéresser aux variables suivantes, ceux-ci ont été choisi pour leurs descriptions et leurs pertinences :



Ici est représentée la moyenne de la variable “Polity2” pour chaque région. Cette variable combinant deux autres variables “Autoc” et “Democ” nous indique sur une échelle de -10 à 10 le type de régime politique mis en place. On voit par exemple que c’est en Amérique du Nord que les régimes politiques sont très démocratiques et peu autoritaires tandis qu’en Afrique la moyenne est à -6 pour le nord et l’est est à -3 pour la partie subsaharienne.

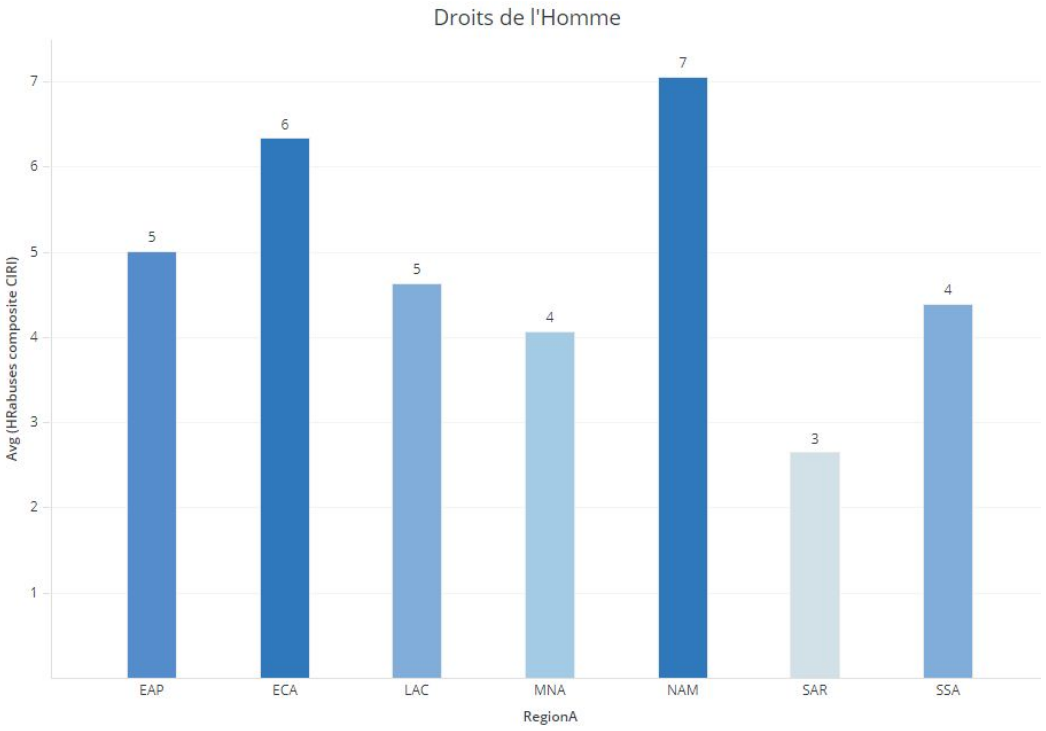
On peut se demander si cette variable est corrélée avec d’autres variables présentes dans notre dataset. Par exemple :



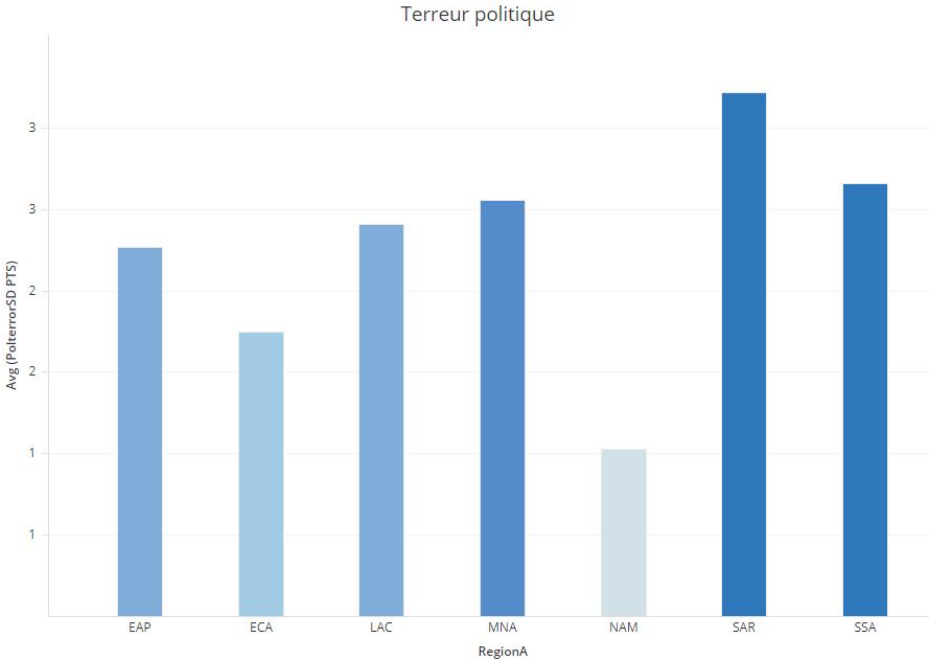
Ici nous voyons que la moyenne du régime politique dépend énormément de la manière dont le leader du pays est arrivé. En effet, les régimes sont plus autoritaires lorsque le dirigeant du pays est arrivé de manière illégale comme un coup d'état (type 2) ou bien qu'il a été placé là par un état étranger (type 3).

Etudions maintenant une variable nommée "HRabuses_composite_CIRI". Elle est la composante de 4 variables : Torture_CIRI, Pol_prisoners_CIRI, Extrajudicialkilling_CIRI, Disappearances_CIRI.
Elle mesure donc le respect des droits de l'Homme se basant sur un indicateur de torture, de disparition politique, de prisonniers politiques et de meurtres organisés. Cette variable se situe sur une échelle de 0 à 8, 8 signifiant un respect maximal pour ces droits.

Ainsi nous pouvons mettre cette variable en relation avec le régime politique en vigueur, en effet en Europe et en Amérique, là où les régimes sont plutôt démocratiques, on remarque que les respects pour les droits de l'Homme sont élevés. A l'inverse ils sont bas là où les régimes politiques sont autoritaires (MNA, SAR et SSA) .



On peut également étudier revenir sur la terreur politique calculée par les US pour voir qu'elle est aussi en corrélation avec les régimes politiques et les respects des droits de l'Homme. En effet, la terreur politique est très élevée là où les respects pour les droits de l'Homme sont très faibles (SAR,SSA).



Il paraît aussi intéressant d'afficher la niveau de régime en fonction de la terreur politique. On voit donc ci-dessous que plus la terreur augmente, plus le niveau de régime baisse et donc se rapproche d'un régime autoritaire.

Pour finir cette étude descriptive sur la politique, nous prenons 4 variables. La première ayant déjà été étudié est Polity, ensuite FreePress exprime la liberté de la presse, Corruption mesure la corruption du service public et VHpart exprime le pourcentage de la population participant aux élections.

Tableau des corrélations pour ces 4 valeurs :

	Polity	FreePress	Corruption	VHpart
Polity	1	-0.7321923	0.5352643	0.6173442
FreePress	-0.7321923	1	-0.6460097	-0.4669308
Corruption	0.5352643	-0.6460097	1	0.3809388
VHpart	0.6173442	-0.4669308	0.3809388	1

Ici l'on remarque que la liberté de la presse et le type de régime politique sont plutôt corrélés (-0.73), ce qui signifie que 0.54% de l'information de l'un est contenu dans l'autre. Néanmoins on ne peut pas vraiment retirer d'informations concluantes sur les autres indices de corrélations.

Il peut être intéressant dans le cas de notre étude de réaliser un clustering des pays en fonction de paramètre qui nous semblent pertinent afin de voir si une tendance se dégage. Pour cela nous avons pris en considération 2 variables qui ont un rapport avec la politique. Cependant afin de faciliter le clustering nous avons modifié la dimension de nos tableaux mais ceci est expliqué plus bas.

PolterrorSD_PTS (qui décrit l'intensité de la terreur politique selon les US) est notre 1er paramètre.

Il y a aussi la présence de PolterrorAM_PTS (qui décrit la terreur politique selon Amnesty) mais celle-ci possède plus de variable manquante que celle des US : 6331 contre 5420. Cette variable varie de 1 à 5 (1 indiquant qu'il y a peu de terreur). Pour le clustering et réaliser la réduction de dimension on va utiliser sa moyenne arrondie sur les dernières années. Les valeurs manquantes auront pour valeurs 0 car cela ne change pas la moyenne. Il suffira de faire attention au nombre d'éléments différents de 0.

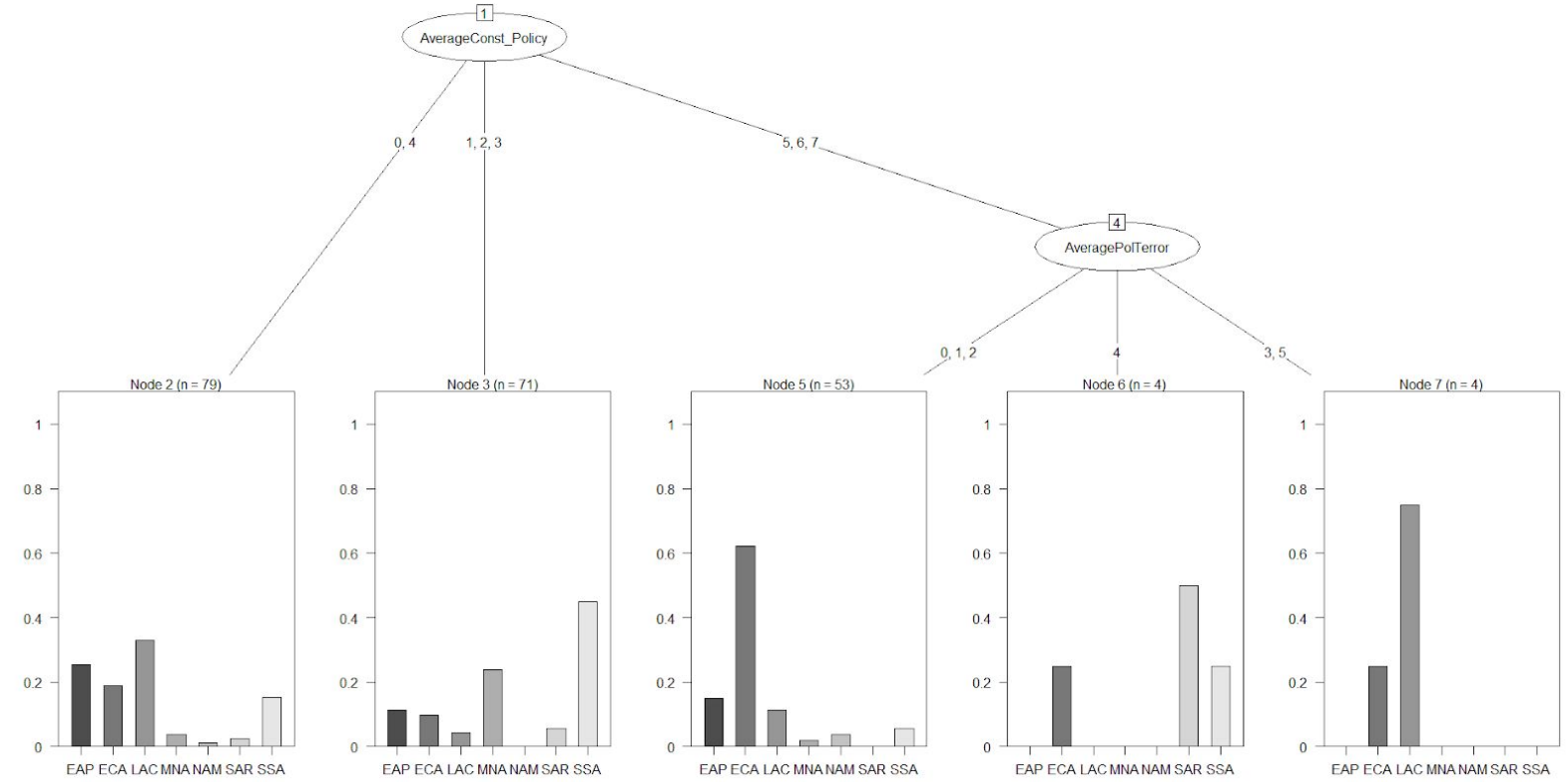
L'idée de notre réduction c'est donc d'avoir un tableau de ce type :

	CountryName	RegionA	AveragePolTerror
1	Afghanistan	SAR	5
2	Albania	ECA	3
3	Algeria	MNA	3
4	American Samoa	EAP	0
5	Andorra	ECA	0
6	Angola	SSA	4
7	Antigua and Barbuda	LAC	0
8	Argentina	LAC	3
9	Armenia	ECA	1
10	Aruba	LAC	0

Xconst_PolityIV (qui prend une valeur indiquant le pouvoir de l'exécutif) est notre 2nd paramètre.

Concernant cette variable elle prend des valeurs entre 1 et 7 (1 indiquant un pouvoir illimité) mais si on regarde les valeurs présentes il y a aussi les valeurs -77, -88, -66. Ces valeurs ont des significations qu'on a pu retrouver sur http://home.bi.no/a0110709/polityiv_manual.pdf à la page 17. Pour le traitement de ces valeurs nous allons leurs assigner la même valeur que les valeurs manquantes c'est-à-dire 0. En effet, ces valeurs indiquent qu'il y a eu une transition, interruption de la politique et donc que l'exécutif du pays n'a eu aucun pouvoir.

En utilisant que ces 2 variables on peut déjà réaliser un arbre de décision nous donnant des informations sur une manière de compléter nos valeurs nulles et quel pays ont des points communs au niveau de la politique:



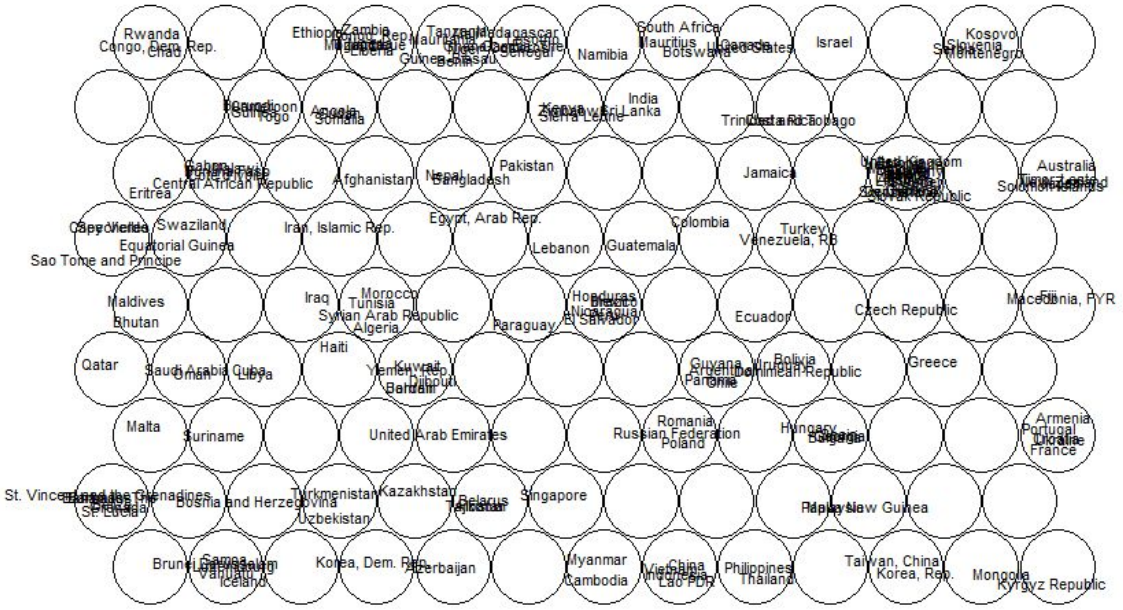
Ici on voit que les pays ayant une constance politique nulle sont groupés avec ceux ayant une constance politique de 4. Les pays où l'exécutif ont un pouvoir important sont dans les régions d'Afrique du Nord et du Sud. Dans les pays où il existe des groupes ayant la même autorité que l'exécutif ceux-ci se divisent par l'intensité de terreur dans le pays. Les pays où on a très peu peur du gouvernement sont dans la région ECA. Mais les pays où la terreur est répandue à la majorité de la population (**AveragePolTerror > 4**) et le pouvoir de l'exécutif bien répartis sont peu nombreux : 8 pays sur 211.

Pour node6 : Inde, Afrique du Sud, Sri Lanka, Turquie

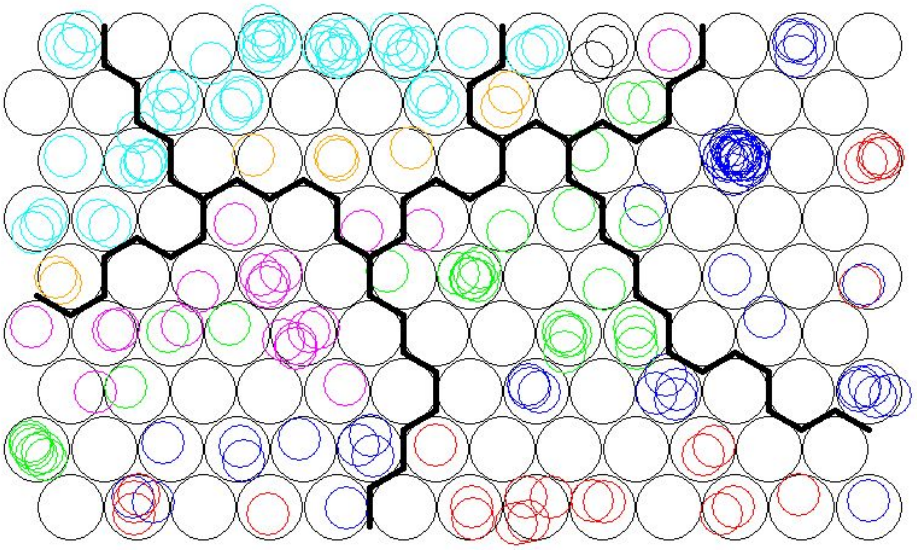
Pour node7 : Colombia, Bolivie, Bulgarie, Venezuela

Nous pouvons aussi réaliser une visualisation des clusters à l'aide de méthode comme SOM. Dans cette visualisations nous avons supprimé tous les pays ayant les 2 variables égalent à 0 car il y a un manque d'information cela n'aurait pas été utile et aurait faussé notre visualisation. Après traitement nous obtenons une image ci-dessous avec les noms de pays réparti sur un hexagone 13 x 9. De nombreux pays se trouve sur la même case et cela n'est pas très visuel mais si on affiche désormais en couleur la région des pays on a une autre figure plus intéressant. Concernant le code couleur :

EAP : Asie de l'Est et du Pacifique **(rouge)**
ECA : Europe / Asie Central **(bleu)**
LAC : Amérique Latine/ Caraïbe **(vert)**
MNA : Afrique du Nord et du centre-est **(violet)**
NAM : Amérique du Nord **(noir)**
SAR : Région d'Asie du Sud **(orange)**
SSA : Afrique Sud – Saharienne **(cyan)**



Mapping plot



Les très noir en gras ont été obtenus à l'aide d'un clustering hiérarchique ascendant coupé en 6. Le but était de voir si les régions ont une influence sur la politique du pays. Il semble en effet d'après cette image que pour les pays d'**Europe/Asie Central** (zone en haut à droite) et pour les pays d'**Afrique** (zone en haut à gauche) ceux-ci sont très peux mélangé avec d'autre régions. Pour les pays d'**Amérique Latine** et d'**Afrique du Nord et du centre** est ceux-ci se mélange avec d'autre pays. Quant aux pays **Asie de l'Est et du Pacifique** ils semblent être regroupés pour la plupart en bas de l'image.

Ce cluster semble plus intéressant que celui obtenu avec un arbre décisionnel car les pays sont mieux répartis en terme de nombre.

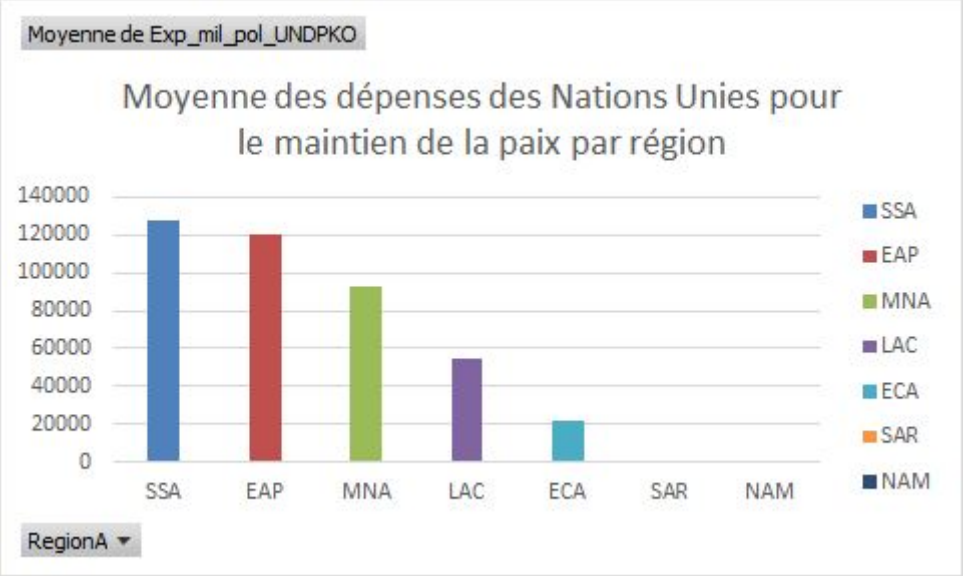
b. Une étude Économique

Le critère économique reste un des critères les plus important dans la société actuelle et pour lequel on peut le plus remarquer les pays qui sont en développement d'un pays qui est en guerre ou qui est déjà développé.

On peut relier le critère politique et le critère économique notamment grâce aux aides financière dépensé par les Etats-Unis sous forme de millions de dollars dans les déploiements militaires de leur armé (plus grande puissance mondiale militaire)

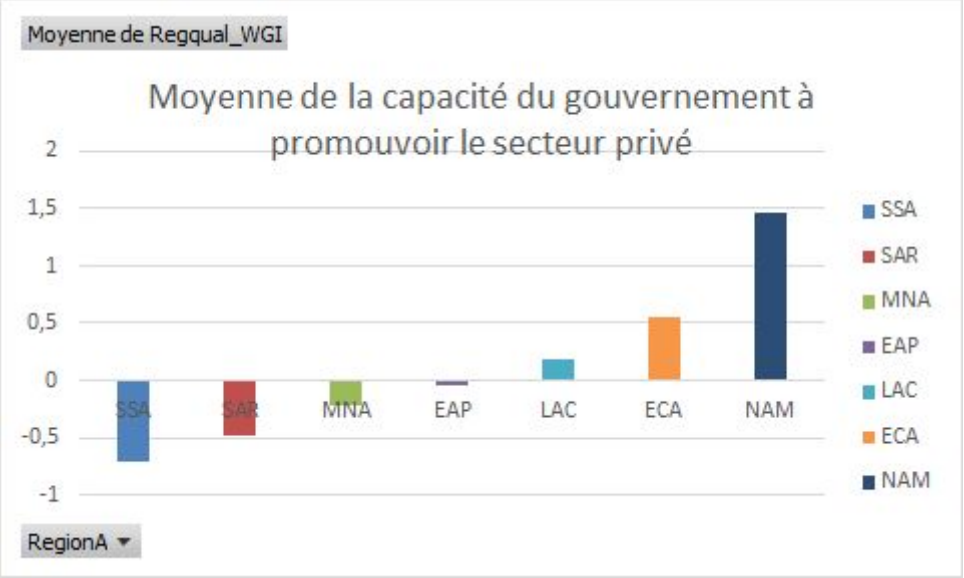
Somme de MilgrantsFr	Étiquettes de colonnes										
Étiquettes de lignes	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	Total général
Afghanistan	0	0	0	0	1,4	163,4	49,3	0	0	0	214,1
Colombia	33,7	53,5	0	0	0	0	0	0	0	0	87,2
Jordan	22,3	25,4	0	0	0	0	0	0	0	0	47,7
Iraq	0	0	0,2	0	1,1	44,6	0	0	0	0	45,9
Georgia	0	0	0	0	24,5	0	0	0	0	0	24,5
Antigua and Barbuda	5,1	13,8	0	0	3,2	0,2	0	0	0	0	22,3
Tunisia	0	4,1	3,3	4,8	5,4	0	0	0	0	0	17,6
Philippines	0	0	0	0	7,6	0	0	4,6	0	0	12,2
Timor-Leste	0	11	0	0	0	0	0	0	0	0	11
Pakistan	0	0	0	0	0	0	0	0	9,9	0	9,9
Indonesia	0	0	0	0	0	0	0	7,3	1,5	0	8,8
Peru	3,9	4,7	0	0	0	0	0	0	0	0	8,6
Bosnia and Herzegovina	6,5	0	0	0	1,1	0,4	0	0	0	0	8
Rwanda	0	0	0	0	0	0	0	7,7	0	0	7,7
Sri Lanka	0	0	0	0	0	0	0	7,3	0	0	7,3
Bolivia	7,3	0	0	0	0	0	0	0	0	0	7,3

On constate que ce critère se rapproche aux déploiements des troupes américaines dans des zones de guerre (L'Afghanistan arrive en tête dû à la volonté des pays du Nord d'éradiquer les menaces terroristes qui s'y trouvait et qui s'y trouve encore). On peut également regarder les dépenses que les Nations Unies ont effectué pour maintenir la paix dans ces régions **:**

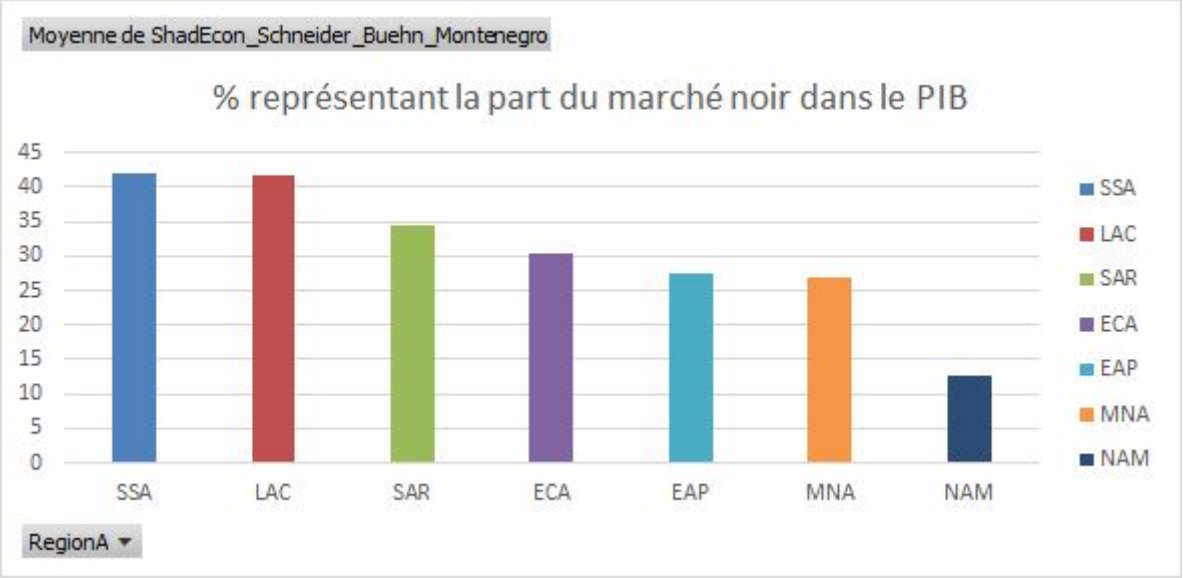


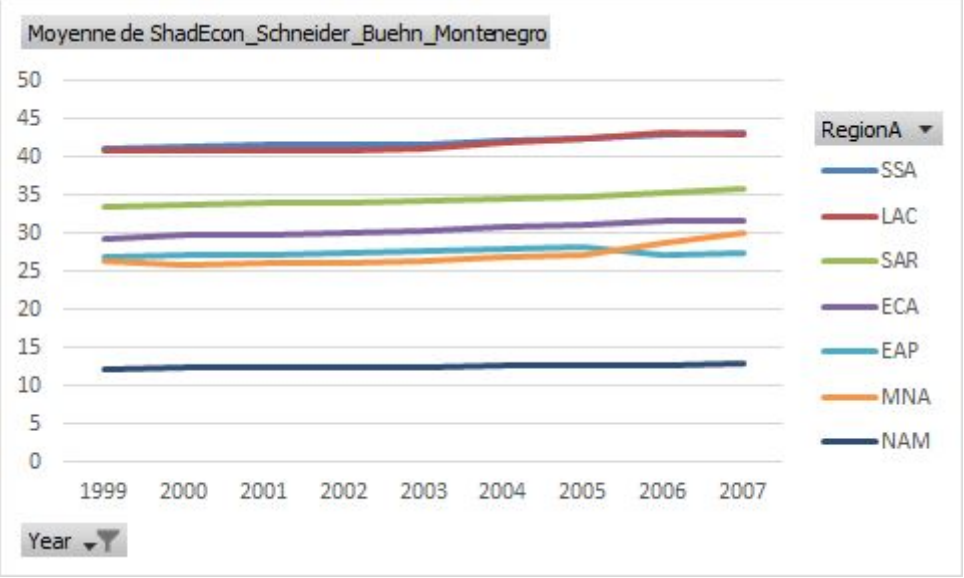
On remarque la force présence des Nations Unies dans des zones où les guerres civiles sont importantes comme en Afrique ou encore en Asie de l'Est. Comme le montre le tableau ci-dessous :

Étiquettes de lignes	Moyenne de Exp_mil_pol_UNDPKO
Sudan	261314,45
Congo, Dem. Rep.	228419,9944
Sierra Leone	203488,625
Somalia	193947,95
Cambodia	192987,175
Cote d'Ivoire	161956,84
Liberia	160167,9636
Lebanon	93167,3
Timor-Leste	88214,45556
Haiti	81165,075
Burundi	77357,8
Kosovo	66756,01
Namibia	63122,5
Croatia	62168,23333
Mozambique	44831,675
Rwanda	21737,26
Central African Republic	19007,53333
Cyprus	18725,52813
Angola	18660,25
Bosnia and Herzegovina	10347,05714
Georgia	3132,55
Guatemala	1462,4
Tajikistan	1176,085714
El Salvador	506,46



On remarque ici que les régions développé comme le nord de l'amérique, ou encore l'Europe favorise le développement du secteur privé. On va voir ensuite la part du marché noir dans les PIB des pays.



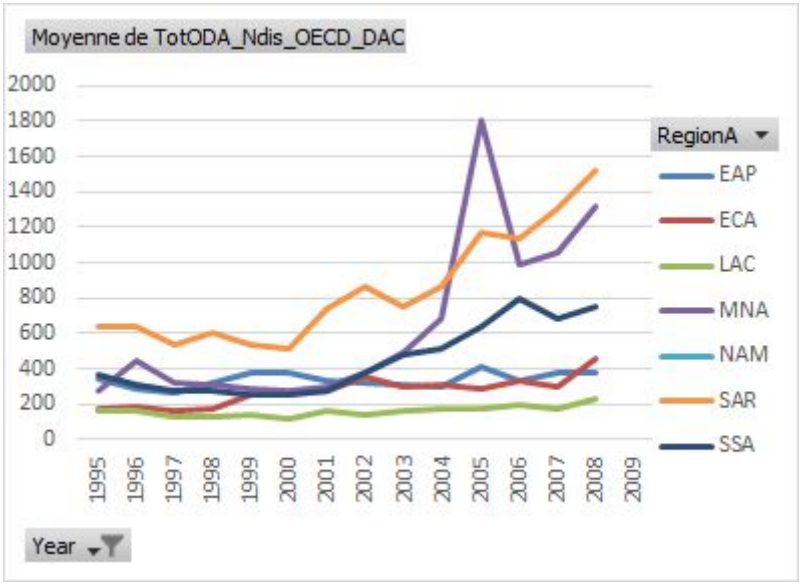


Pour ce qui est du marché noir, on remarque qu'il représente une part très importante dans certaine région du monde. Même en Amérique du Nord : avec environ 13% du PIB du pays. On peut aussi constaté qu'il est en augmentation entre 1999 et 2007 dans toutes les régions du monde excepté en Asie de l'Est. Ce marché représente plus de 50% du PIB de certains pays comme le montre le tableau suivant :

Étiquettes de lignes	Moyenne de ShadEcon_Schneider_Buehn_Montenegro
Georgia	68,81111111
Bolivia	68,12222222
Panama	64,65
Azerbaijan	63,3
Peru	61,81111111
Tanzania	60,2
Nigeria	59,65
Zimbabwe	57,04444444
Ukraine	54,9
Myanmar	54,9
Thailand	54,71111111
Haiti	54,37777778
Guatemala	52,5
Uruguay	51,54444444
Cambodia	51,45555556
Honduras	50,88888889
Zambia	50,81111111
Congo, Rep.	50,12222222

On remarque ici que 2 pays d'Europe de l'Est sont présents dans les pays dont le PIB est en grosse partie influencer par le marché noir, l'Ukraine et la Géorgie (presque 69% ce qui est énorme pour un pays européen).

Ensuite nous pourrons faire une corrélation entre le développement des pays et le total d'argent net employé pour le développement du pays.



On remarque ici les régions en développement comme l'Afrique sub-saharienne. On verra ensuite la catégorie sociétale.

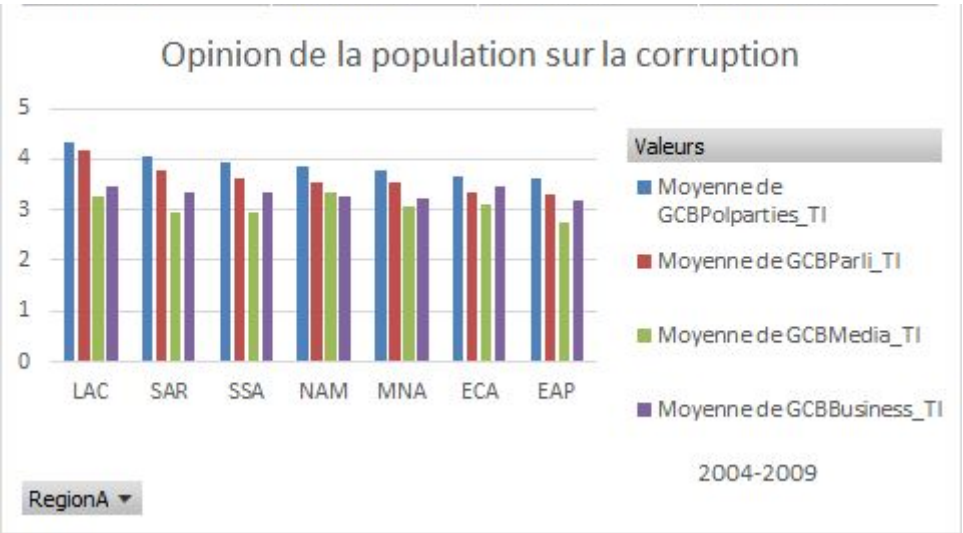
c. Une étude Sociétale

Une catégorie primordiale à prendre en compte en plus de celles de l'économie et de la politique reste la société. En effet, une étude sociétale peut permettre de constater que le réseau d'institutions, de traditions et de relations mis en place a un impact sur la qualité de vie de chacun. Ainsi, cette étude se base non seulement sur des faits mais également sur l'opinion des populations quant aux différentes actions entreprises par leurs pays et les agissements de leurs gouvernements.

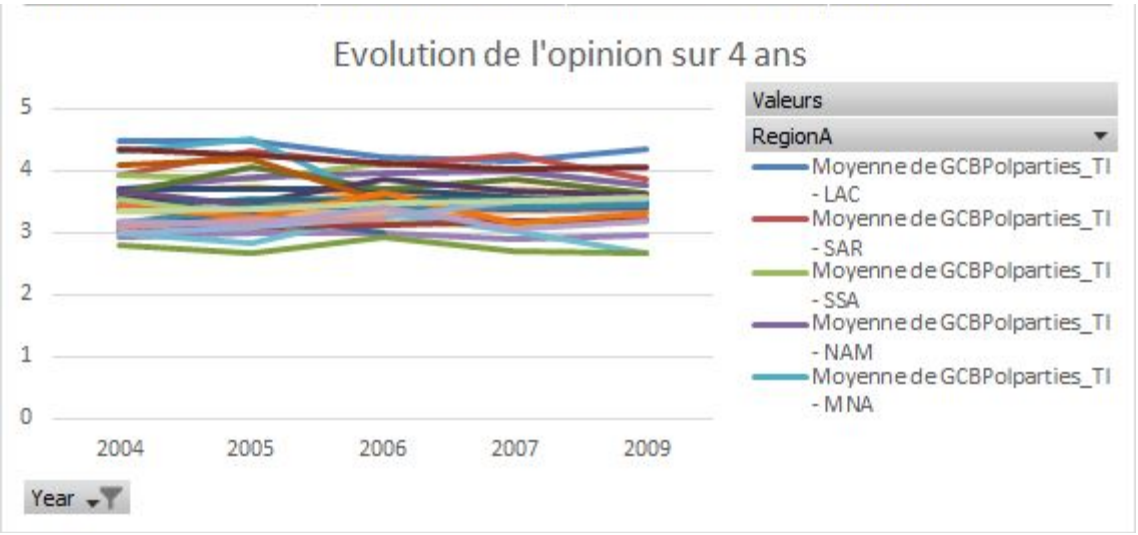
Nous sélectionnons, parmi l'ensemble des données qui nous sont proposées, des critères reliant ces phénomènes de société. Il est donc nécessaire d'analyser et regrouper de multiples données pour en faire une analyse concrète.

On précise que l'ensemble des analyses qui sont réalisées s'effectuent sur une plage annuelle **"contenant"** des données. En effet, dans un souci d'objectivité, il serait impensable d'effectuer une analyse sur 50 ans avec pour seules informations des données regroupées sur 5 ans.

Tout d'abord, sur la question de la question de la corruption, il semble important de connaître l'opinion des populations concernant la potentielle corruption qui agirait dans leurs pays. En effet, un certain degré de méfiance à l'égard des actions du pays pourrait influencer le fonctionnement de ceux-ci:

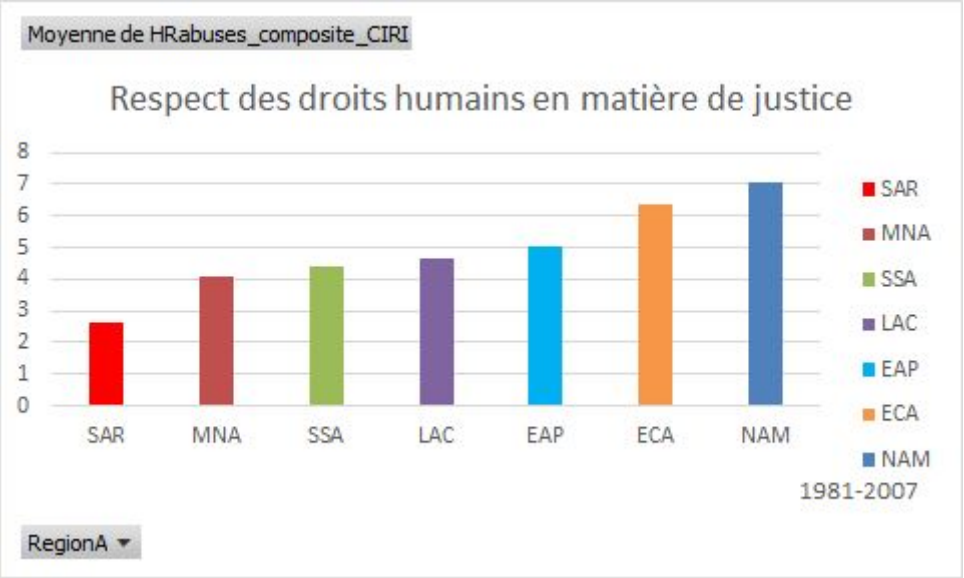


Ce graphique résume la moyenne (sur une échelle de 1 à 4) de l'ensemble des perceptions des populations sur différentes catégories de corruption (respectivement envers les partis politiques, le parlement, la presse et les médias, le secteur des affaires).

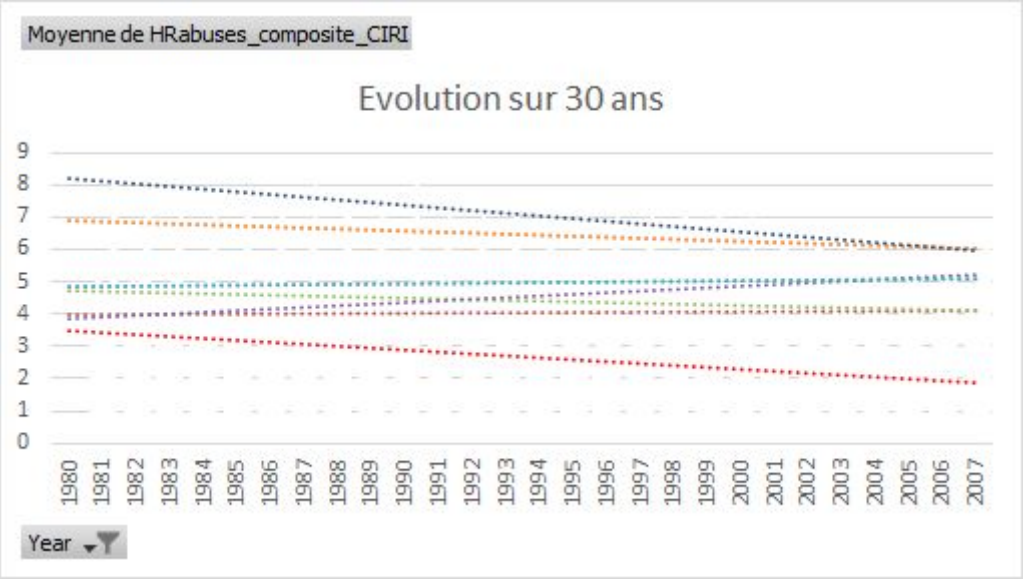


Malgré une domination de la région latino-américaine, on remarque que l'écart est assez faible entre les régions et stagne sur cette période 4 ans sans connaître de réels changements. De plus, moins de la moitié des pays de notre jeu de données est pris en compte dans cette analyse ce qui peut nous faire douter quant à la fiabilité de celle-ci.

De plus, sur la question des droits humains en matière de justice:



Ce graphique vient à résumer la présence et les égards des gouvernements sur les droits humains en matière de justice en fonction de 4 critères: les indicateurs de torture, d'exécutions extrajudiciaires, d'emprisonnements politiques et de disparitions. Une note élevée (sur une échelle de 1 à 8) définit une considération similaire des gouvernements à l'égard de ces 4 critères.



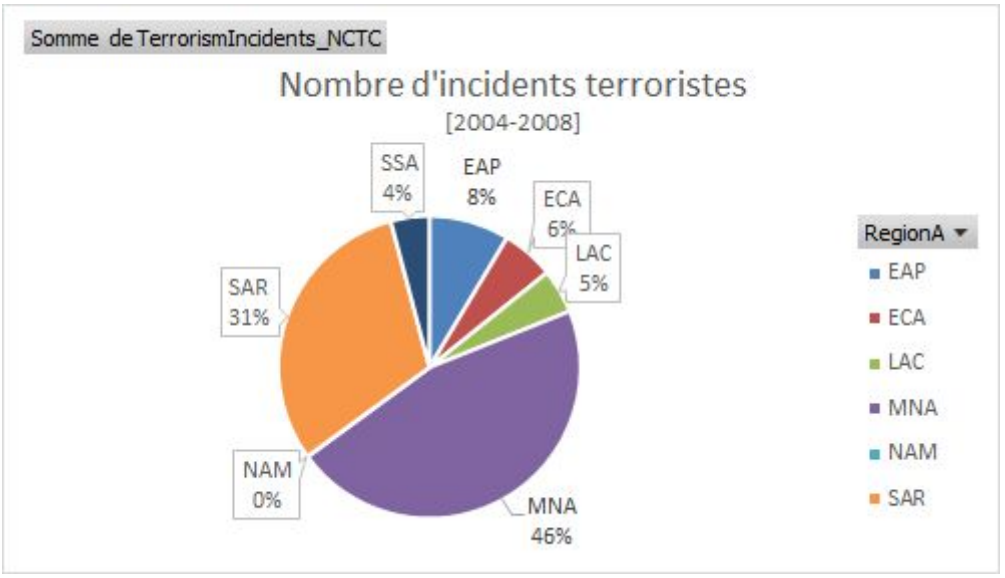
Sur une période longue de 30 ans, la moyenne n'est pas le seul critère d'analyse à prendre en compte car il y a également l'évolution de ces considérations avec leur tendance sur cette plage annuelle.

On remarque clairement la non-action de ces gouvernements en Asie du Sud (moyenne de 2.65) avec une tendance qui s'aggrave progressivement dans cette région ainsi qu'en Afrique Subsaharienne.

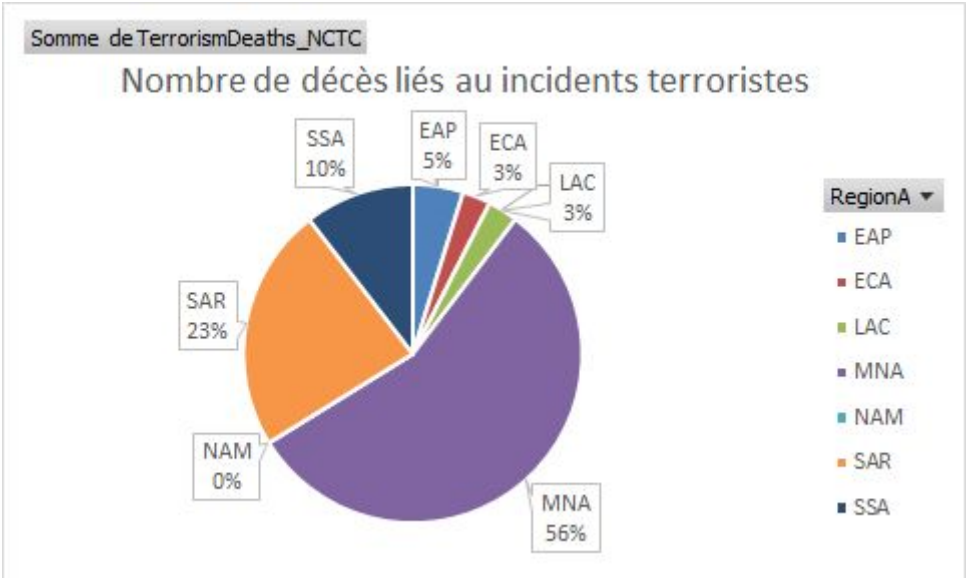
Ce sont notamment les pays comme l'**Inde** et l'**Afghanistan** qui affectent majoritairement la région asiatique puis on retrouve l'essor de certains pays comme la Corée du Nord, l'Iraq et la Colombie.

Étiquettes de lignes	Moyenne de HRabuses_composite_CIRI
India	0,851851852
Korea, Dem. Rep.	0,761904762
Afghanistan	0,761904762
Colombia	0,76
Iraq	0,32

La question du terrorisme n'est pas seulement un facteur politique mais également un repère sociétale surtout quand sa présence dans certaines régions ne relève d'une situation exceptionnelle mais vient à s'installer progressivement.



Ce graphique indique la répartition des incidents terroristes par région entre 2004 et 2008. La prédominance de l'Afrique Centrale et du Nord est à souligner avec plus de 25000 incidents.

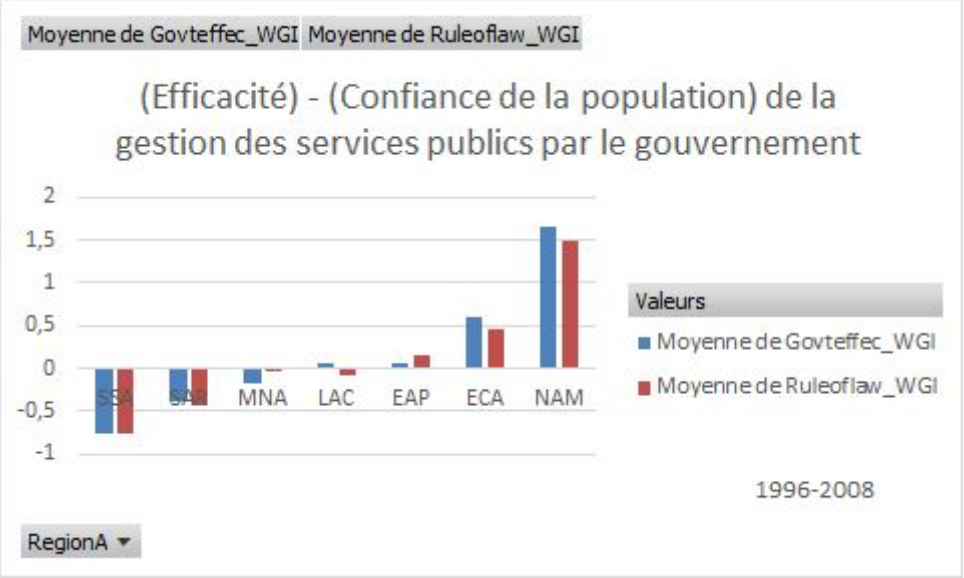


Ainsi, la logique des décès suite à ces incidents est respectée avec plus de 45 000 morts pour la région “SAR” soit plus de la moitié de l’ensemble des décès répertoriés dans le monde à travers ces incidents.

Étiquettes de lignes	Somme de TerrorismIncidents_NCTC
West Bank and Gaza	1295
Israel	2771
Pakistan	3686
Nepal	3836
Afghanistan	3951
India	4462
Iraq	20462
Total général	**40463**

On remarque que l’Irak concentre la majorité des incidents terroristes et résume la situation de ces nombreux pays aux nombres élevés; ils se retrouvent dans une instabilité politique majeure liée à des guerres et affrontements.

La gestion et l’efficacité des services publics est également un repère précieux dans l’étude sociétale que nous menons.

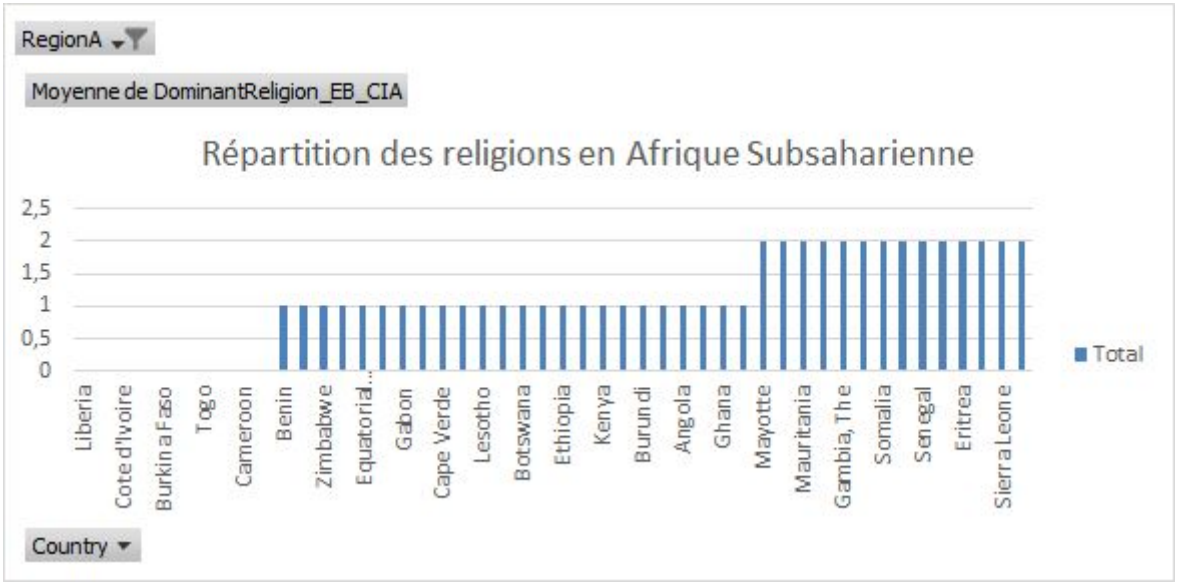


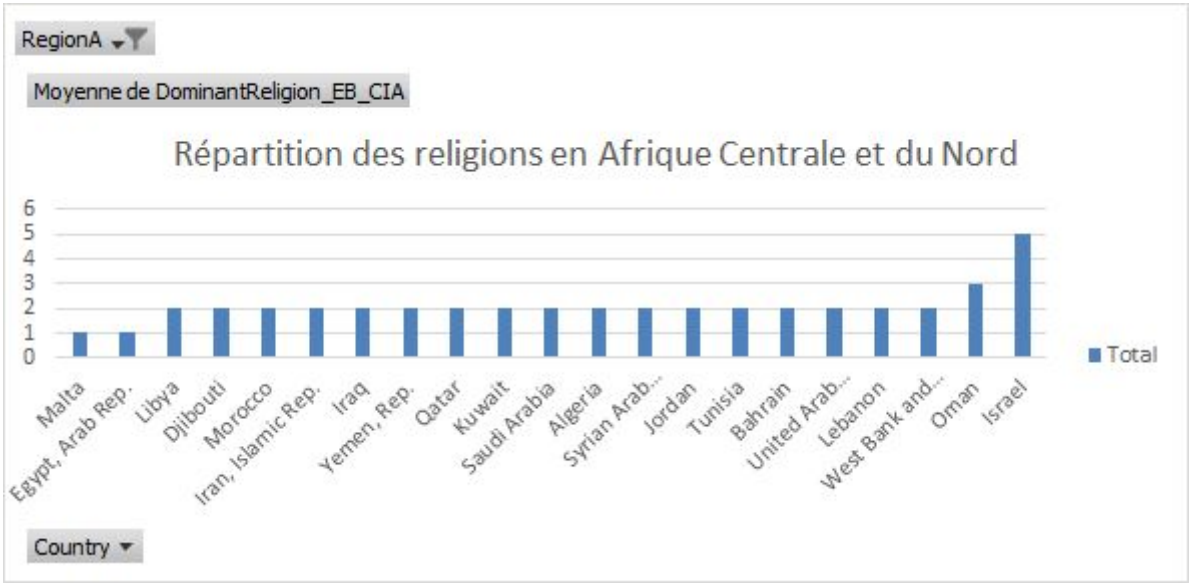
Ce graphique compare l’efficacité et la perception de la population à l’égard de la gestion des services publics par le gouvernement (sur une échelle allant -2.5 à 2.5). On remarque notamment que les appréciations de la population sont en raccord avec les valeurs accordées à l’efficacité du gouvernement. De plus, on dénote le fort écart entre les régions européennes et d’Amérique du Nord avec celles d’Afrique Subsaharienne et d’Asie du Sud.

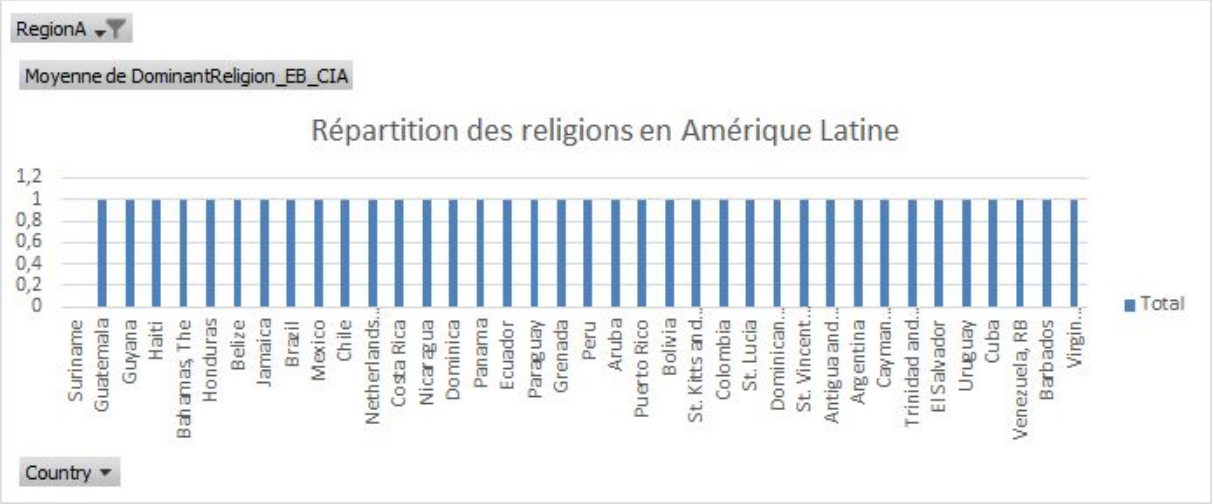
Les raisons de ces évaluations sont majoritairement liées à la gestion du gouvernement dans des pays en crise en raisons de conflits ou de pauvreté.

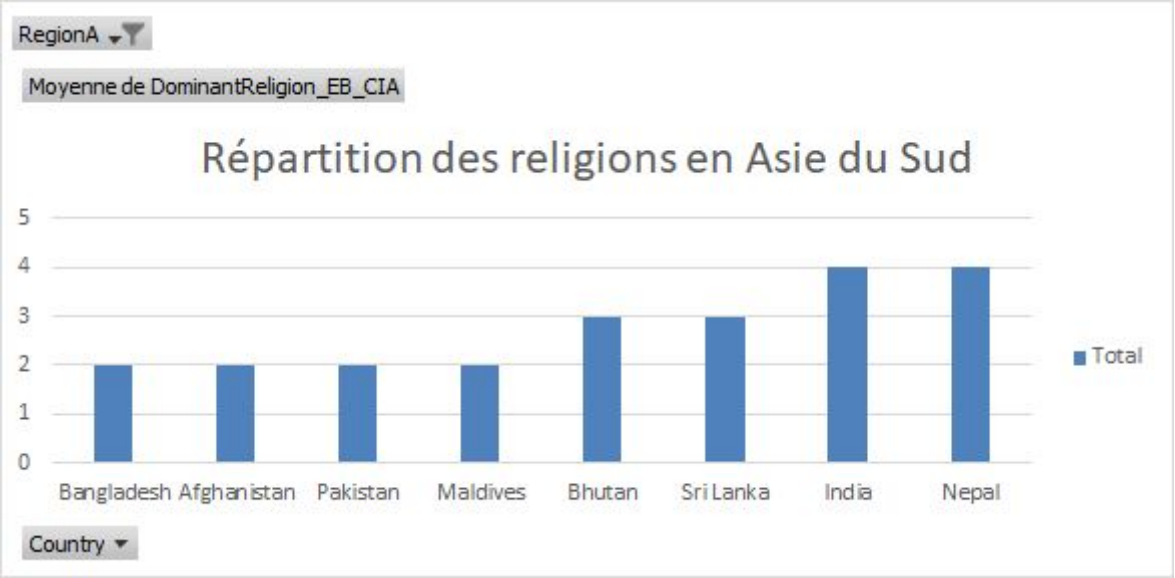
Enfin, notre jeu de données nous donne une variable religion “DominantReligion_EB_CIA” indiquant la domination d’une religion si une population y adhère à plus de 50%. Chaque religion se rapportera donc à un chiffre pour une meilleure analyse. Nous décidons d’analyser les régions qui connaissent de fortes inégalités sociales pour observer si l’on retrouve une religion prédominante.

On définit ces religions de la manière suivante: 0= Pas de religion dominante; 1= Christianisme; 2= Islam; 3= Boudhisme; 4= Hindouisme; 5= Judaisme; 6= Shintoism.









Ces graphiques nous permettent de conclure qu'il n'y a pas de religion dominante concernant l'ensemble de ces régions, la religion ne va donc pas influer sur l'ensemble des crises sociétales que ces pays connaissent.

3. Modèle prédictif

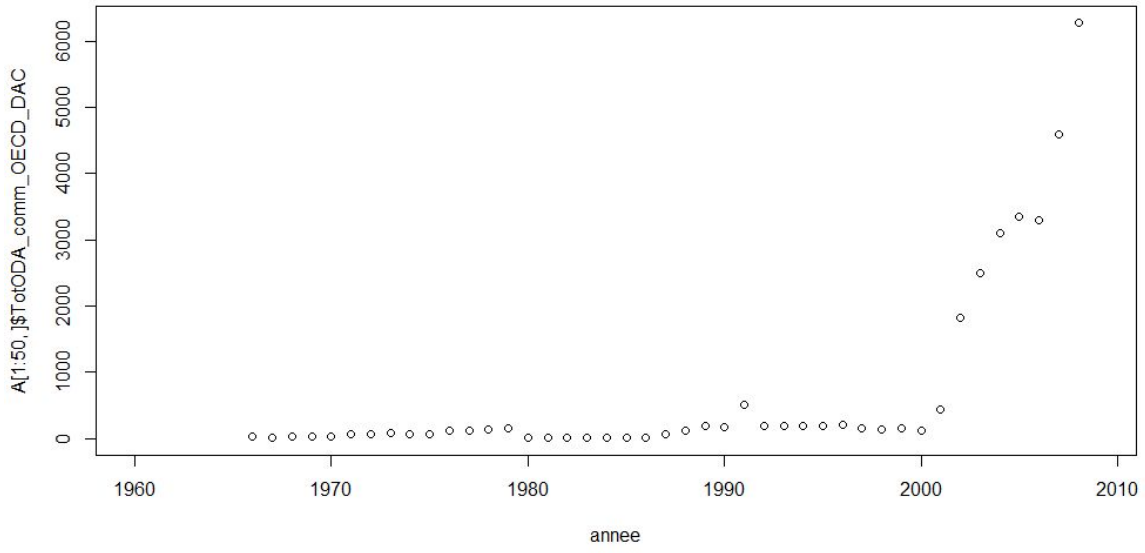
Suite à notre étude descriptive, nous avons choisi de prendre l'Afghanistan comme pays pour notre étude temporelle. Le but va être de trouver un modèle qui correspond au mieux à la variable.

TotODA_comm_OECD_DAC (représente l'aide totale engagé au développement fournie à un pays)

Le choix de ce pays n'est pas sans raison. En effet, c'est un pays ou la stabilité politique est faible et ou la constance des entitées politique aussi. De plus, d'après notre clustering il représente bien la région 'SAR'. Nous voulions réaliser cette étude sur une variable économique car la variable de temps est dans ce cas-là très intéressante. Cependant, dans le dataset la plupart des variables représentant un montant en dollars sont des valeurs provenant de l'OEDC et donnent des informations sur les aides fournies aux pays. Il en vient donc que pour des pays développés il y a de nombreuses valeurs manquantes voir aucune valeurs.

a) Avec régression linéaire

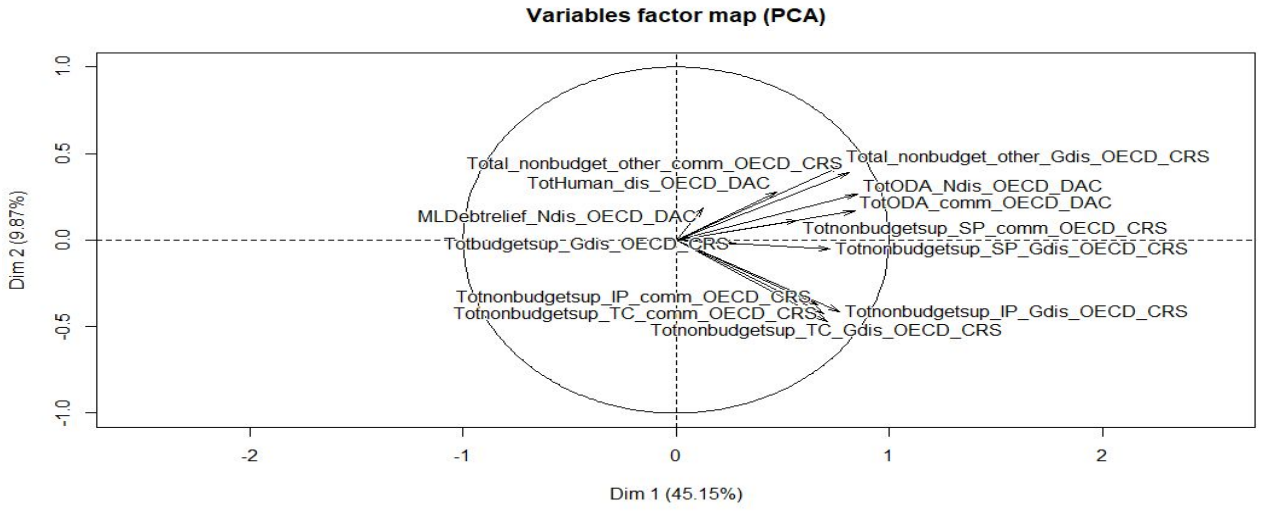
L'un des moyens que nous avons choisi pour prédire l'évolution de l'aide fourni au pays est la régression linéaire. Pour ce faire nous devons trouver d'autres variables très corrélés avec celle choisi. Avant de chercher ces variables nous avons ci-dessous l'évolution de l'aide au développement en millions de dollars sur les 50 dernière années.



Ce que l'on peut remarquer est l'évolution drastique à partir des années 2000. Si l'on va sur le site de world bank on peut récupérer l'évolution du PIB de l'Afghanistan. Les 2 courbes se ressemblent beaucoup et il est facile de conclure que l'aide obtenu par les autres pays aide à l'évolution du PIB. Si l'on observe les valeurs on peut remarquer que l'aide obtenue représente près de 50% du PIB du pays. Dans la suite les valeurs que l'on obtiendra pour l'aide donnée seront comparé au PIB du pays afin de voir si il y a une certaine évolution où si ce pourcentage reste le même.

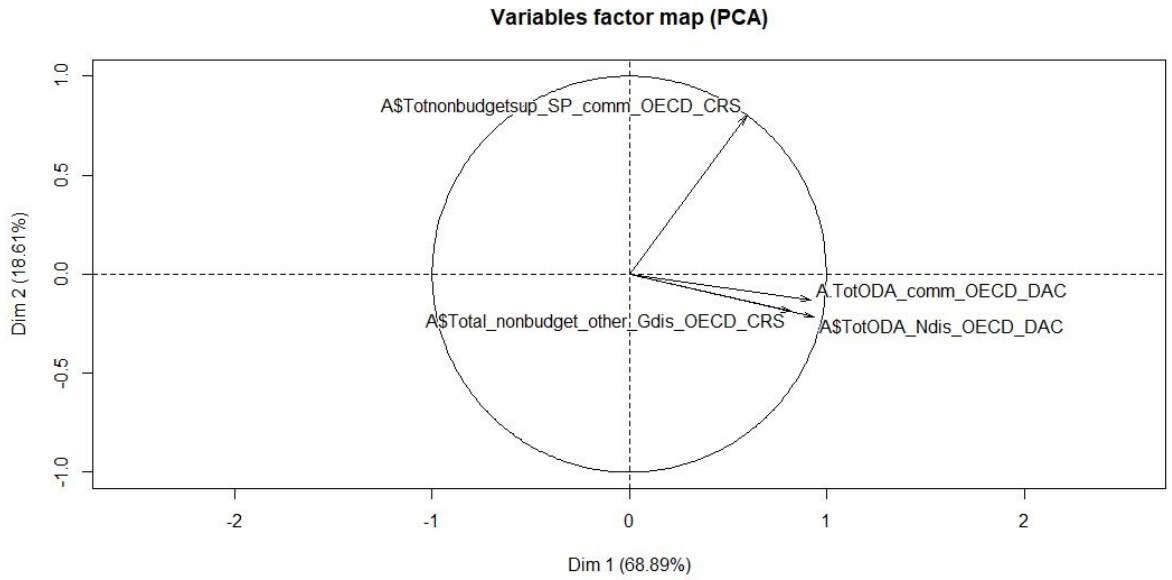
Un point sur lequel on a pas insisté précédemment est le manque de valeur sur la période précédent 1666. Nous supposons que durant cette période il n'y a pas eu d'aide qui leur ont été fourni. Nous ne traiterons pas ces valeurs et les supprimons dans notre étude.

Pour réaliser la régression linéaire nous avons déjà énoncé le fait que nous devions trouver des variables très corrélés. Nous avons donc réalisé une Analyse en Composante Principale afin de déterminer quel variables nous pourrions prendre. Avec le logiciel R nous avons pris en compte l'ensemble des variables provenant de OEDC-CRS et OEDC-DAC. Nous obtenons alors cette figure :

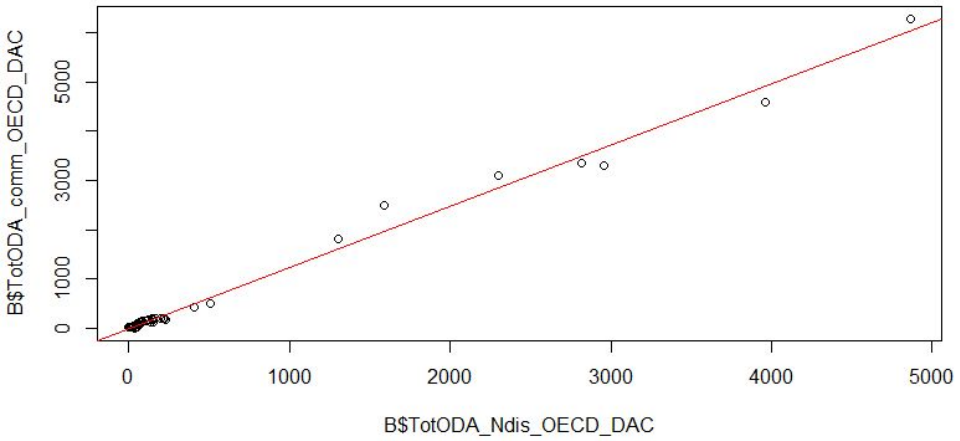


Ce que l'on peut voir c'est que notre variable **TotODA_comm_OECD_DAC** semble très corrélé à plusieurs variables :
TotODA_Ndis_OECD_DAC (qui représente le total d'argent net donné au pays pour le développement)
Total_nonbudget_other_Gdis_OECD_CRS (Total de l'appui non budgétaire en brut pour les autres domaines)
Totnonbudgetsup_SP_comm_OECD_CRS (Total de l'appui non budgétaire aux programmes sectoriels engagé dans le pays)

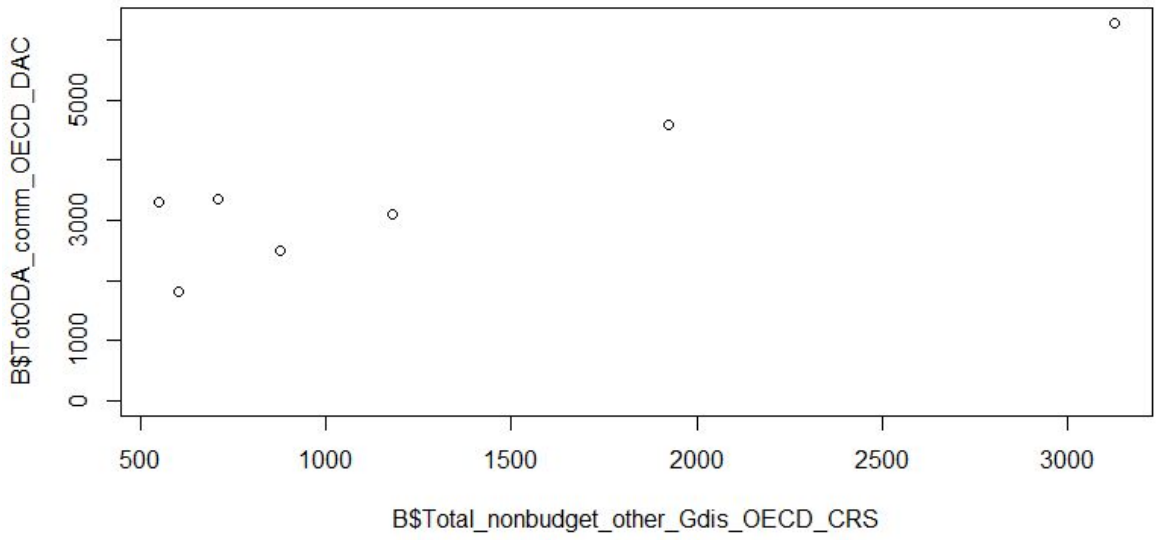
Si on refait une ACP avec seulement ces 3 variables on obtient :



On voit que l'une des variables est très éloigné de celle qui nous intéresse **TotODA_comm_OECD_DAC** on a donc pris les 2 variables restantes. Puis, afin de choisir la meilleur pour notre régression linéaire nous allons réaliser la régression et allons regarde le coefficient de corrélation R^2. Avant cela regardons ce que donne les courbes et regardons ce que nous donne la régression celle-ci sera représenté en rouge :
Avec **TotODA_Ndis_OECD_DAC**



Avec **Total_nonbudget_other_Gdis_OECD_CRS**



On remarque qu'il y a très peu de points seulement 7 il apparaît donc évident que nous n'allons pas prendre cette variable en compte au vu du nombre de valeurs manquante. Malgré tout les points semblent bien alignés et cette variable aurait pu être intéressant pour la régression.

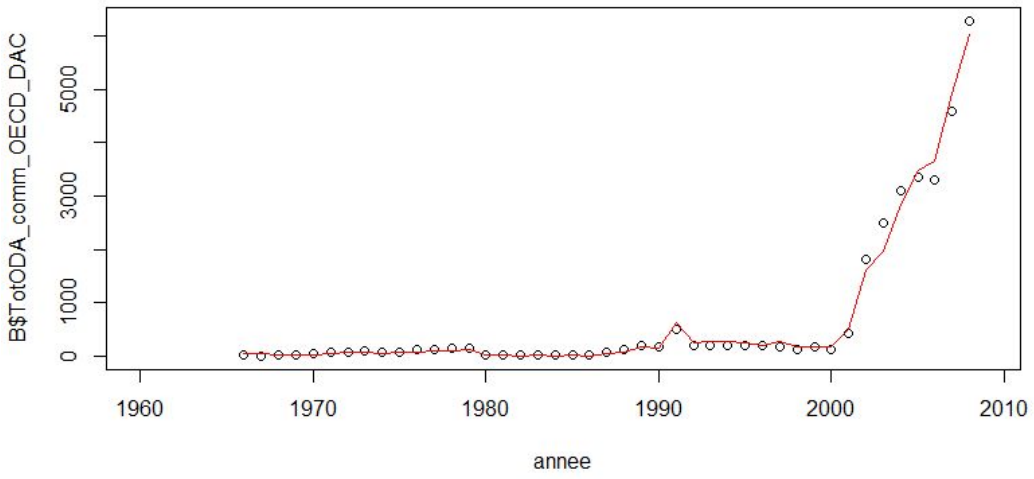
Retournant au budget net donnée par les pays à l'afghanistan. Si on utilise la méthode summary() sur notre précédente régression on obtient ce résultat :

```
> summary(mod)

Call:
lm(formula = test[1:50, ]$A.TotODA_comm_OECD_DAC ~ test[1:50,
    ]$`A$TotODA_Ndis_OECD_DAC`)

Residuals:
    Min      1Q  Median      3Q     Max
-357.49  -45.78    7.93   21.54  538.48

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -8.76222   23.35902  -0.375     0.71
test[1:50, ]$`A$TotODA_Ndis_OECD_DAC`  1.24114    0.01879  66.047   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.3 on 41 degrees of freedom
  (7 observations deleted due to missingness)
Multiple R-squared:  0.9907,    Adjusted R-squared:  0.9905
F-statistic:  4362 on 1 and 41 DF,  p-value: < 2.2e-16
```

La valeur de R^2 est très proche de 1, 0.99 nous allons donc garder ce modèle pour nos prédictions. Concernant les valeurs des coefficients nous les avons avec l'image ci-dessous :

```
Call:
lm(formula = B$TotODA_comm_OECD_DAC ~ B$TotODA_Ndis_OECD_DAC)

Coefficients:
           (Intercept)  B$TotODA_Ndis_OECD_DAC
                -8.762                   1.241
```

Si nous traçons la courbe prédite par R nous obtenons la courbe ci-dessous en rouge :



La courbe suit plutôt bien notre courbe initial. Cette régression est un bon modèle. Si on regarde le tableau de **TotODA_Ndis_OECD_DAC** on voit que il y a des informations entre 1960 et 1966. Nous pouvons donc remplacer nos valeurs manquantes grâce au coefficient de régression et de la constante trouvé. On a alors le tableau suivant :

Année	**TotODA_Ndis_OECD_DAC**	**TotODA_comm_OECD_DAC (new value)**
1960	17.2	13.0
1961	34.7	34.3
1962	16.9	12.2
1963	36.7	36.8
1964	46.2	48.6
1965	53.9	58.1

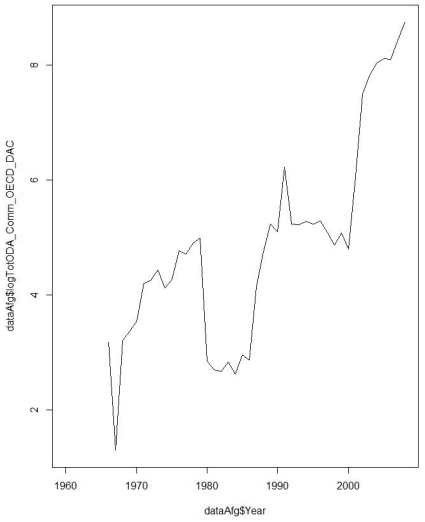
Globalement, l'ordre des points est plutôt bien respecté. On ne trouve pas de valeurs aberrantes. En effet, si l'on compare au PIB du pays qu'on a trouvé sur World Bank ces valeurs ne dépasse jamais le PIB et reste pour la plupart dans les 50% du PIB du pays. Malgré tout nous ne pouvons déterminer la valeur de l'année en 2009 car il nous manque cette valeur aussi dans le tableau de **TotODA_Ndis_OECD_DAC.**

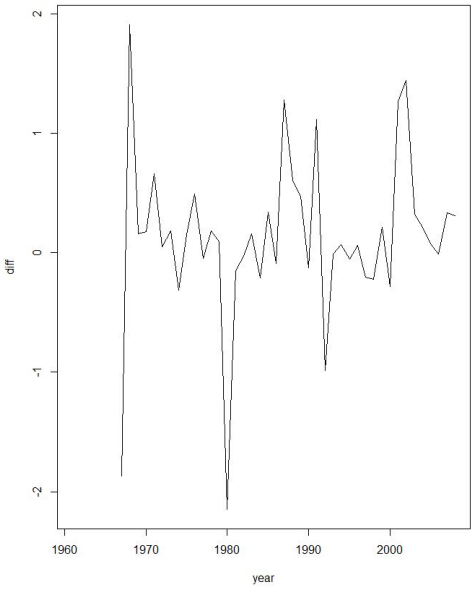
Pour conclure, nous n'avons pas réalisé le test sur les autres pays de la région 'SAR' mais il apparaît qu'un modèle de régression linéaire prenant en compte comme paramètre **TotODA_Ndis_OECD_DAC** semble pertinent

Notre prochaine étape est de voir si avec le modèle ARIMA, qui est plus complexe, les résultats que nous avons peuvent nous permettre de prédire l'évolution de l'argent engagé pour le développement.

b) Avec le modèle arima

En étudiant l'évolution de l'aide totale allouée à l'Afghanistan, on se rend compte que l'on ne peut appliquer le modèle arima car la série n'est clairement pas stationnaire. On applique donc le logarithme pour que la courbe ait un profil plus linéaire, ce qui nous donne la courbe à droite.

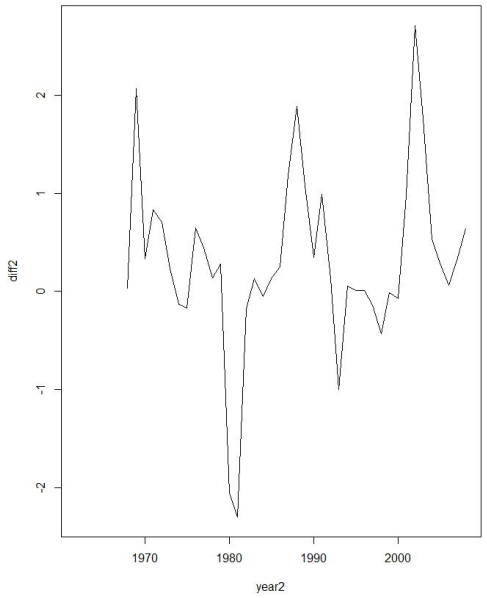





```
        Box-Pierce test

data:  diff
X-squared = 0.56957, df = 1, p-value = 0.4504
```

```
> t.test(diff)

        One Sample t-test

data:  diff
t = 1.2033, df = 41, p-value = 0.2357
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.08988688  0.35493165
sample estimates:
mean of x
0.1325224
```

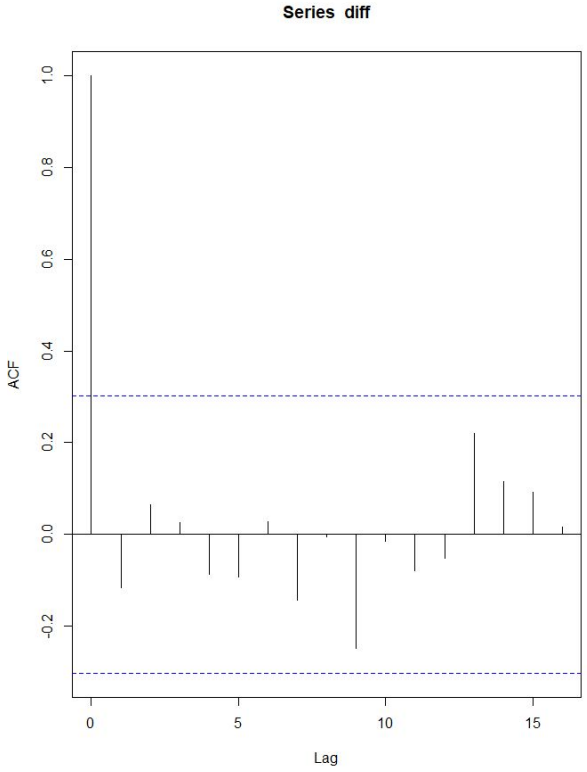
On va dérive ensuite la série temporelle ce qui nous donne les résultats ci-dessus. La p-value du test de Box-pierce indique qu'on ne peut pas dire que la série est dépendante du temps. De même, le test de student montre qu'il est possible que la moyenne soit égale à 0. On va quand même essayer de dériver une seconde fois pour voir si l'on obtient pas de meilleurs résultats.

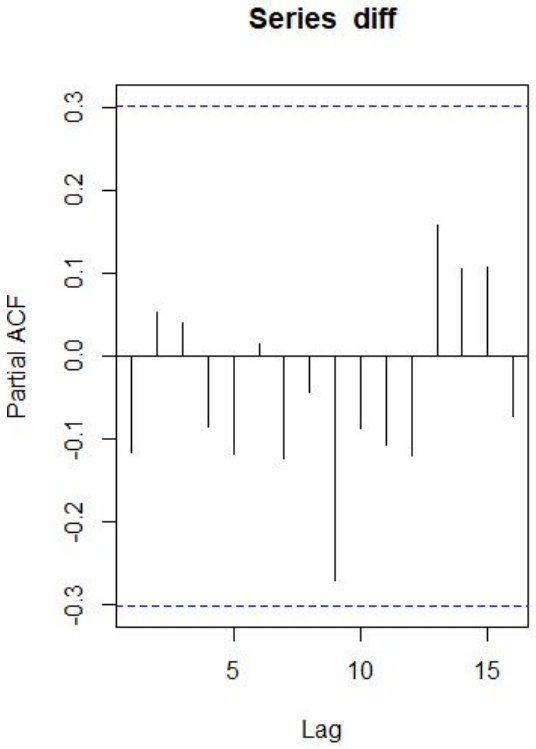



```
        Box-Pierce test

data:  diff2
X-squared = 10.752, df = 1, p-value = 0.001042
```

```
        One Sample t-test

data:  diff2
t = 2.1892, df = 40, p-value = 0.03448
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.02376867 0.59527295
sample estimates:
mean of x
0.3095208
```

Pour la dérivée seconde, la p-value du test de box pierce montre que l'on rejette l'hypothèse nulle. Donc la série est dépendante du temps. De plus, d'après le test de student, la moyenne n'est clairement pas égale à 0 et 0 n'est même pas dans l'intervalle de confiance de la moyenne. On en conclut dont que cette dérivation était inutile. Nous continuerons donc notre étude sur la série diff(log(TotODA_comm_OECD_DAC)).





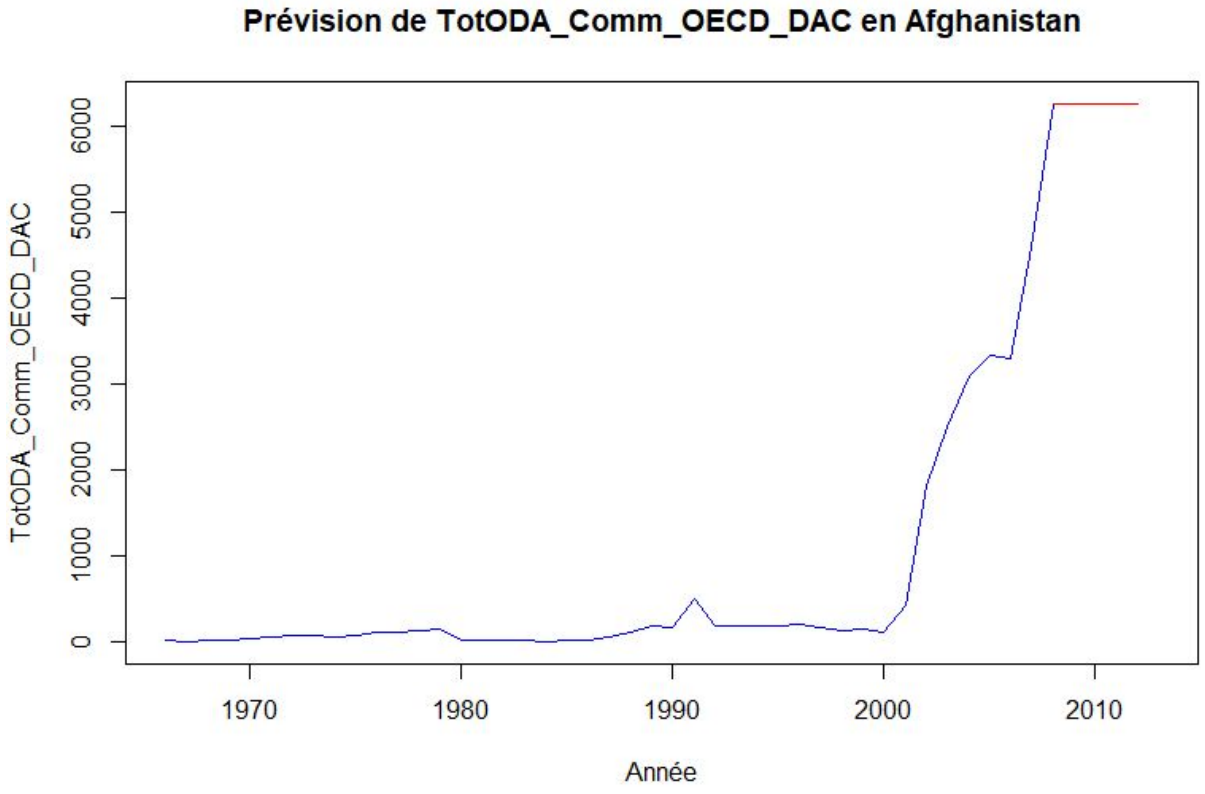
Grâce à l'autocorrélogramme et à l'autocorrélogramme partiel, on se rend compte qu'il n'y a aucun pic qui dépasse la limite des pointillés. On va donc utiliser le modèle arima(0,1,0) sur la série log(TotODA_Comm_OECD_DAC). La fonction auto.arima confirme mes observations.

On va maintenant étudier les résidus du modèle arima(0,1,0) pour vérifier qu'ils soient bien stationnaires.

```
        Box-Pierce test

data:  residuals(arima)
X-squared = 0.46506, df = 1, p-value = 0.4953
```

```
        Shapiro-Wilk normality test

data:  residuals(arima)
W = 0.85968, p-value = 9.233e-05
```

Le test de Box-Pierce indique que les résidus sont plus susceptibles d'être indépendants. Cependant, la faible valeur du test de shapiro montre que les résidus ne sont pas issus d'une population normalement distribuée, ce qui est assez gênant. Les tests réalisés avec d'autres paramètres potentiellement intéressants (p=0 ou 9, q=0 ou 9) se sont révélés infructueux.



En effectuant une prévision à 5 ans, on se rend compte que la courbe reste stable. On aurait pu le prédire plus tôt sachant que les coefficients p et q du modèle sont à 0. Les prédictions avec les modèles arima(0,1,9), arima(9,1,0) et arima(9,1,9) donnent des valeurs décroissantes ce qui semble encore moins pertinent.