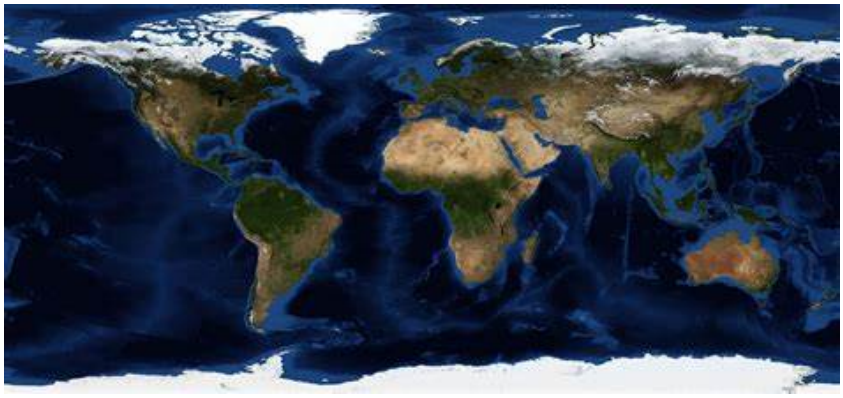


ESILV

Rapport de statistique des valeurs extrêmes, simulation, modélisation et série temporelle

Etude de la mortalité dans le monde causée par différents facteurs

Nicolas COBAN, Cindy DAI, Edmée HOGENMULLER, Léa PAUSE

Encadré par
Gwladys MAO

2023

Table des matières

1 Introduction

Les données de notre base concernent la mortalité causée par 5 types d'évènements : famine, épidémie, conflit et terrorisme, catastrophes naturelles et autres blessures.
Cette mortalité est donnée pour chaque pays, par année.

Il est intéressant d'essayer de comprendre les raisons de variation de cette mortalité, ce qui peut expliquer que certaines années et certains pays ont plus de morts que d'autres.

Nous avons également des caractéristiques des pays sur chaque année : leur PIB, leur population, l'état de la démocratie et le nombre d'évènements par année d'un des 5 types.

Nous avons dans un premier temps étudié les valeurs extrêmes du Nombre de morts pour 100 000 habitants, puis nous avons essayé de modéliser le PIB par habitant, nous avons ensuite effectué des simulations et avons fini par étudier les séries temporelles du nombre d'évènements liés au terrorisme et aux catastrophes naturelles.

2 Présentation de la base et de nos données

2.1 Traitement de la base

Restructuration et jointure :
Nous avions initialement une ligne par type de cause pour chaque année et chaque pays. Nous avons simplifié cette base de données en ne gardant qu'une ligne par pays par année.

Année	Pays	Cause	Mort
2000	France	Famine	190
2000	France	Epidémie	14
2000	France	Catastrophe naturelle	90
2001	France	Famine	60
2001	France	Epidémie	120
2001	France	Catastrophe naturelle	113

→

Année	Pays	Epidémie	Famine	Catastrophe naturelle
2000	France	190	14	90
2001	France	60	120	113

Nous nous sommes aperçus que nous manquions de variables explicatives, et avons donc recherché d'autres bases de données, afin d'avoir plus de variables pour nos modèles.

Nous avons trouvé une base de données qui listait toutes les attaques terroristes et une autre qui recensait les catastrophes naturelles et épidémies.
Nous avons restructuré ces bases de données afin d'obtenir le nombre d'évènements de chaque type par année pour chaque pays.

Date	Pays	Type
01/01/2000	France	Storm
01/06/2000	France	Wildfire
01/11/2000	France	Drought
05/11/2000	France	Tsunami
01/01/2001	France	Storm
01/05/2000	France	Tsunami

→

Année	Pays	Nb_cat
2000	France	4
2001	France	2

Nous avions également une base de données qui décrivait l'état de la démocratie pour chaque pays par année. Cette donnée combine des informations sur la mesure dans laquelle des élections ouvertes, multipartites et compétitives permettent le choix d'un chef exécutif.

Nous avons finalement joint ces différentes bases de données en fonction du pays et de l'année, ce qui nous a donné une base de données avec 18 variables.

Données manquantes :
Nous nous sommes aperçus que nous avions un certain nombre de données manquantes sur la population (151) et sur le PIB (789) parmi les 7181 lignes que nous avions.

Nous avons émis l'hypothèse que d'une année à l'autre, ces données ne variaient pas énormément pour un pays. Nous pouvions donc deviner ces données manquantes en utilisant la technique d'interpolation linéaire.

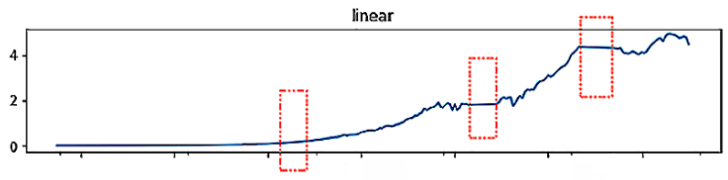


FIGURE 1 – Interpolation

Nous avons cependant été contraints de supprimer certaines lignes, car des pays avaient trop d'années avec des données manquantes, et il n'était donc pas possible de deviner ces données.

Nous avons ensuite normalisé le Nombre de morts pour chaque type et au total, par rapport au nombre d'habitants. Nous avons donc compté ces morts pour 100 000 personnes.
Après ce nettoyage, il nous restait **6854 observations**, pour **38 années** et **185 pays**.

Variables :

Nom	Description
Year	Années, allant de 1980 à 2018
Country	Pays
ISO	Code ISO du pays
Conf_terr	Nombre de morts lié aux conflit et terrorisme
Epi	Nombre de morts lié aux épidémies
Fam	Nombre de morts lié à la famine
Cat_nat	Nombre de morts lié à une catastrophe naturelle
Other	Nombre de morts lié à d'autre blessures
Male.Pop	Nombre d'habitants homme
Female.Pop	Nombre d'habitants femme
Total.Pop	Nombre d'habitants
PCAP	PIB par habitant
Nb_cat	Nombre d'évènements de type catastrophe naturelle
Nb_epi	Nombre d'évènements de type épidémie
Nb_conf	Nombre d'évènements de type conflit et terrorisme
democracy_polity	Etat de la démocratie. Échelle de 21 allant de -10 (monarchie héréditaire) à +10 (démocratie consolidée).
Death_percmill	Nombre de morts total (somme des morts par cause)

TABLE 1 – Description des variables

2.2 Étude univariée et bivariée

Nombre de morts pour 100 0000 habitants :

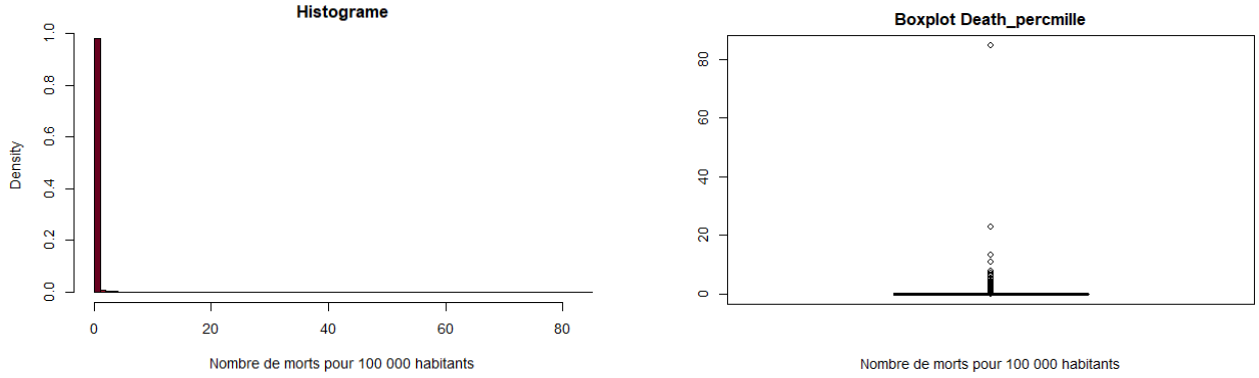


FIGURE 2 – Variable de la mortalité pour 100 000 habitants

Pour notre variable target, nous observons que la plupart des valeurs sont proches de 0, avec la présence de quelques extrêmes, allant jusqu'à 84 morts par an. Cette variable pourra donc être intéressante dans l'étude des extrêmes.

Les années avec le plus de morts sont 1982, 1994, 1983, 1984, 2004 et 2010.

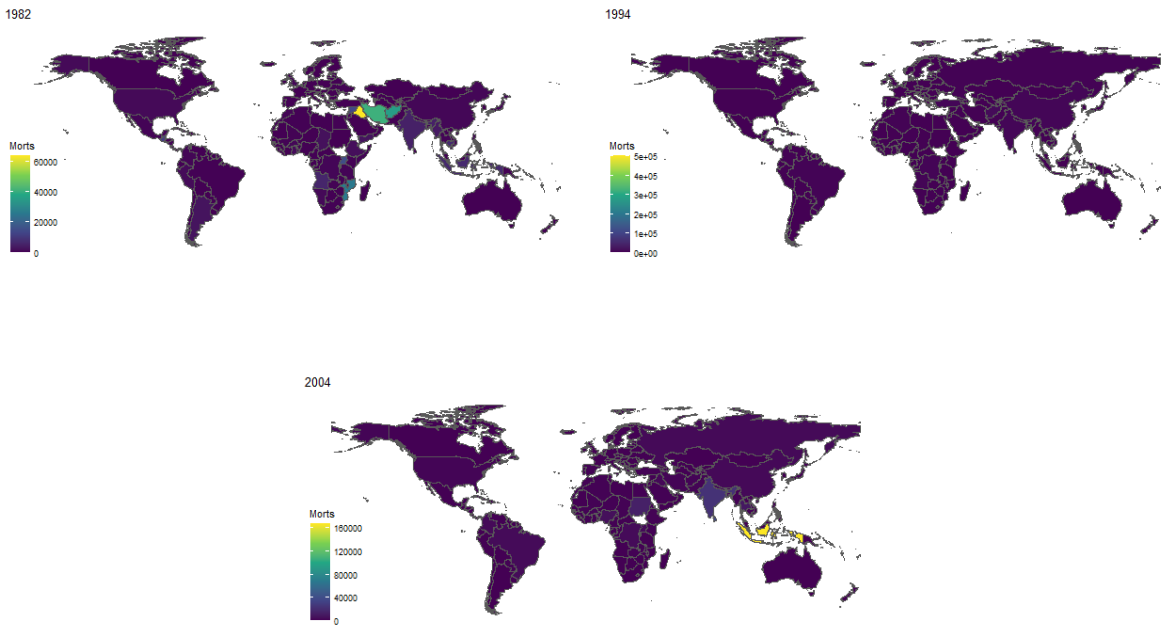
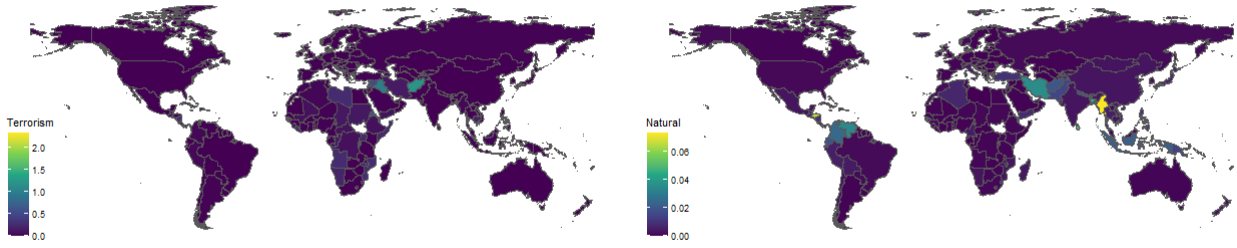


FIGURE 3 – Mortalité dans le monde en fonction de différentes années

Pour chacune de ces années, cette surmortalité est liée à un (ou plusieurs) évènements majeurs, comme :

- 1982 (carte en haut à gauche) : Offensives iraniennes dans la guerre Iran-Irak
- 1994 (carte en haut à droite) : Conflit et génocide au Rwanda
- 2004 (carte en bas) : Un séisme et un tsunami ravagent l'intégralité de l'océan Indien (Thaïlande, Inde, Sri Lanka, etc.)

Nombre de mortss par cause :
Les variables sur le Nombre de mortss par cause ont une répartition semblable à celle du nombre total de morts.



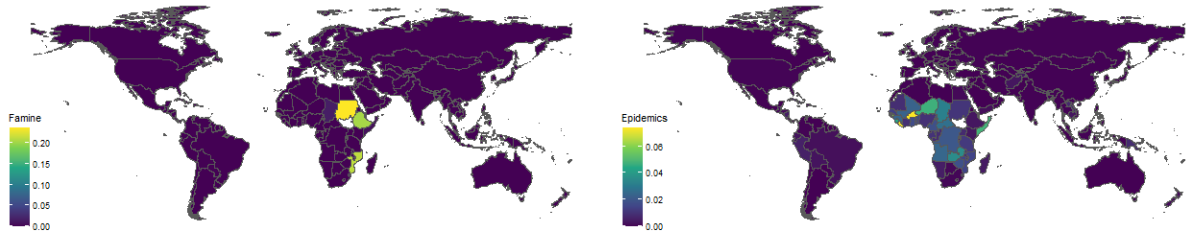


FIGURE 5 – Mortalité dans le monde en fonction des différentes causes

En représentant sur une carte le nombre moyen par pays de morts pour chaque cause, nous pouvons déterminer quels pays sont les plus touchés pour chaque cause :

- Famine : Les pays d'Afrique de l'Est, plus particulièrement en Éthiopie, au Soudan et en Mozambique
- Conflit et terrorisme : Les pays d'Afrique et du Moyen-Orient
- Épidémie : Les pays d'Afrique
- Catastrophe naturelle : Les pays autour de l'Océan Indien et de la mer des Caraïbes

PIB par habitant :

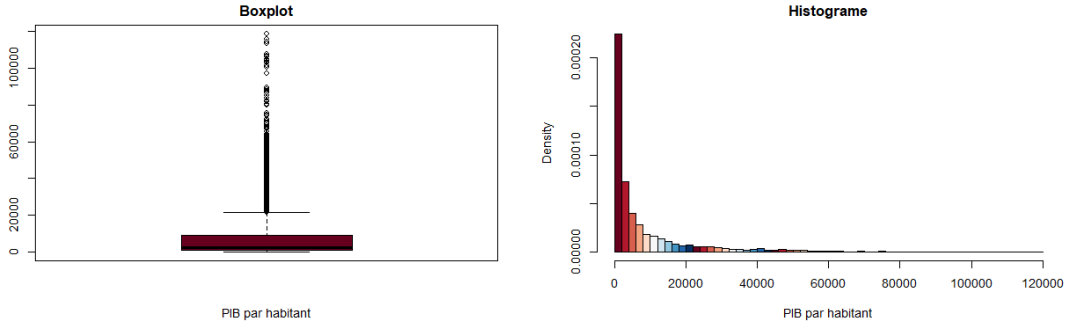


FIGURE 6 – Variable PIB

En étudiant le PIB par habitant, nous observons qu'une majorité des pays ont un PIB en dessous de 10 000$ par habitant, et que le PIB par habitant monte jusqu'à 120 000$ par habitant.

Politique démocratique :

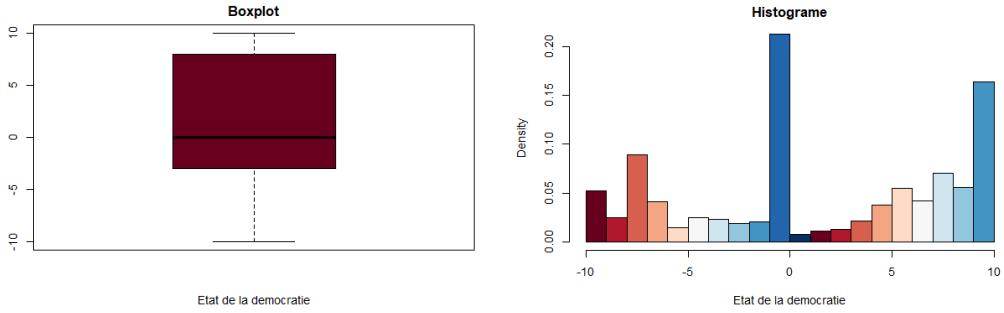


FIGURE 7 – Variable sur la politique démocratique

Concernant le niveau de démocratie, la moitié de nos pays ont un niveau au-dessus de 0, et la moitié en dessous de 0.

En prenant les années 1980 et 2018 (première et dernière année de notre base de donnée), nous constatons que l'état de la démocratie a évolué pour un certain nombre de pays.

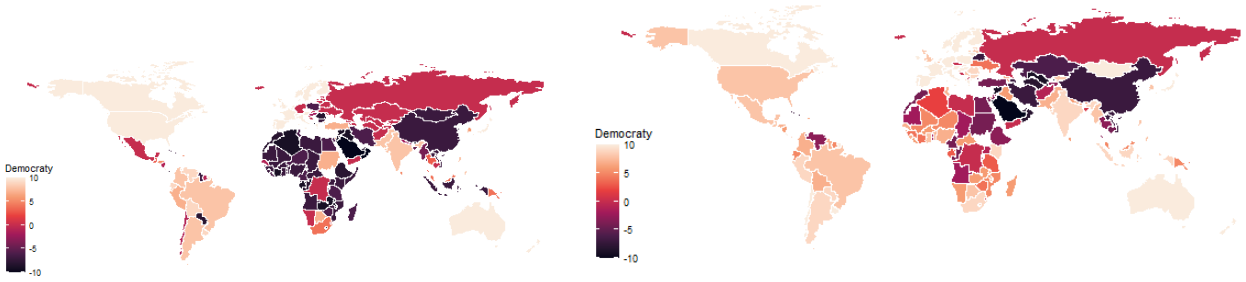


FIGURE 8 – Évolution de l'état de la démocratie dans le monde (à gauche 1980 et à droite 2018)

En effet, la grosse majorité des pays d'Afrique et de l'Europe de l'Est ont vu leur démocratie s'améliorer, tandis que ceux d'Asie de l'Est ont vu leur démocratie se durcir. En Amérique, le Mexique, le Chili et le Paraguay ont vu leur démocratie s'améliorer tandis que le Venezuela a vu sa démocratie se détériorer.

Étude bivariée :

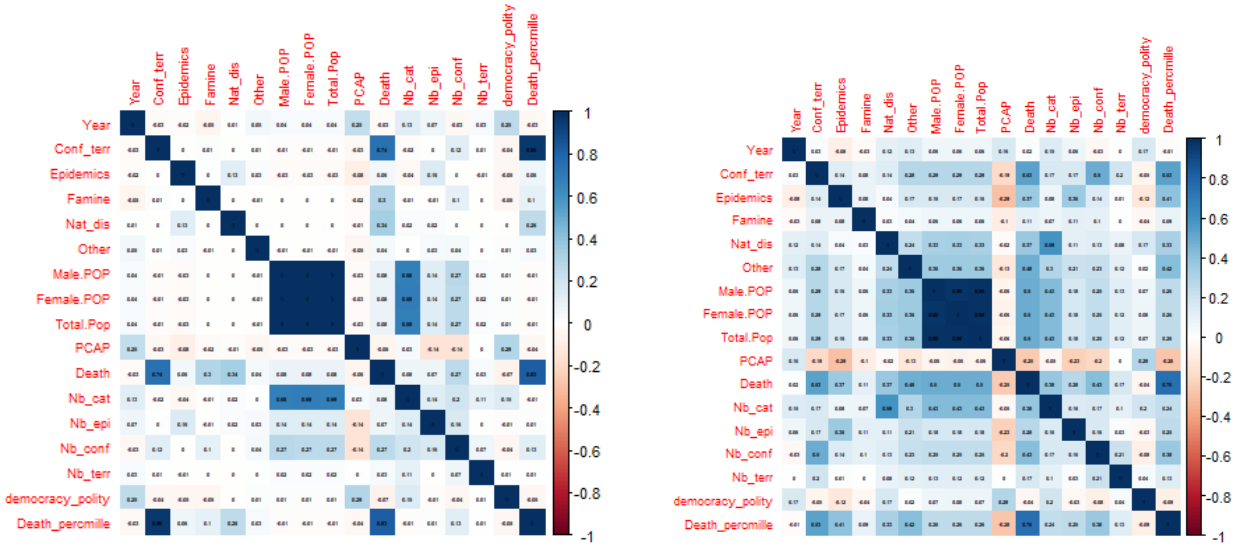


FIGURE 9 – Matrice de corrélation

En analysant les corrélations de Pearson et Kendall, nous observons qu'il n'y a pas de corrélation importante entre les variables autres que celles évidentes (comme entre population des femmes et population totale, ou entre morts pour 100 000 habitants et morts).

3 Statistique des valeurs extrêmes

Notre variable target *death_percmille* comprend une majorité de valeurs inférieures à 1, et quelques très grosses valeurs allant jusqu'à 84. Nous allons étudier les valeurs extrêmes de cette variable.

3.1 Mesure de risque

Le maximum de la variable *death_percmille* est de 84.8 et le minimum de 0. Nous avons un écart-type de 1.13 et une moyenne de 0.35.
Nous avons ici une dispersion très importante de nos données, mais un écart-type plutôt faible comparé à cette dispersion, les valeurs tendent donc à être proches de la moyenne.

La Value-at-Risk de niveau 99% est de 1.8. Nous avons donc 99% de chance qu'il n'y ait pas plus de 1.8 morts pour 100 000 habitants par an.
La TVaR est de 4.74. Ainsi, le nombre moyen de morts au-delà de la VaR est de 4.74.

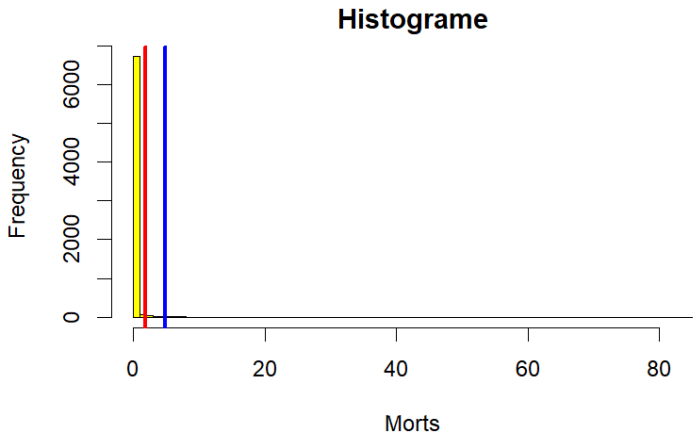


FIGURE 10 – VaR (rouge) et TVaR (bleu)

3.2 Étude du maximum avec GEV

Nous avons ensuite étudié la loi du maximum. Nous avons choisi des blocs annuels, en ne conservant donc qu'une valeur (le maximum) de mort par année. Cela nous donnait donc 38 valeurs.

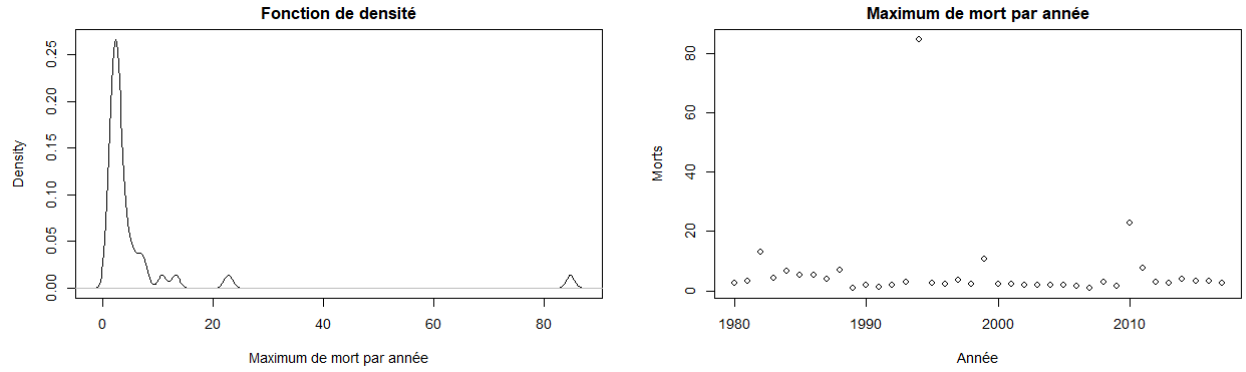


FIGURE 11 – Maximums

À partir du théorème de Fisher-Tippett, nous avons essayé de caractériser les distributions asymptotiques de nos maximas.
En étudiant l'allure de la courbe de distribution, nous observons qu'elle n'est pas bornée, et semble avoir une queue lourde. Nos maximas semblent ainsi suivre une loi de Frechet.

Grâce au maximum de vraisemblance, nous obtenons les paramètres de la loi d'extremum généralisée (GEV) :

- Shape : 1.38
- Scale : 0.65
- Location : 2.29

Notre paramètre shape étant supérieur à 0, nos maximas suivent effectivement une loi de **Frechet**.

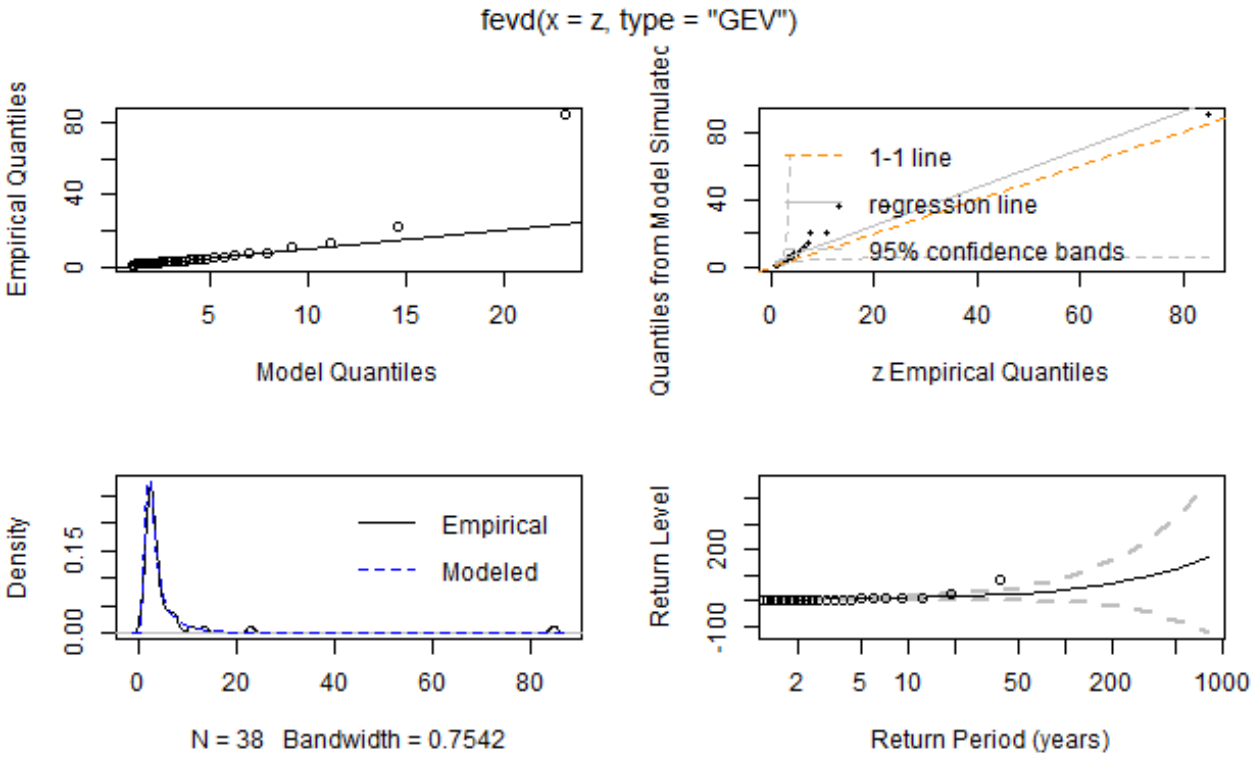


FIGURE 12 – Analyse de la loi de Frechet

En observant nos différents graphiques d'analyse de notre loi de Frechet, il semble que notre loi s'adapte bien avec la plupart de nos données, à l'exception des plus grosses valeurs.
En comparant la densité empirique et la densité modélisée, nous observons qu'elles sont très similaires, hormis pour nos valeurs extrêmes des maximus, que notre modèle n'arrive pas à prédire.

3.3 Étude des excédents/POT

Pour étudier nos excédents, nous avons d'abord recherché l'indice à partir duquel nos données étaient considérées comme excédents.
Nous avons d'abord utilisé la « rule of thumb », en définissant l'indice méthodiquement. Cela nous a donné 3 premiers seuils : 0.08, 1.45 et 0.74.

Nous avons ensuite utilisé une approche graphique pour déterminer nos seuils.

Nous avons étudié le graphique de la durée moyenne, pour chercher le seuil à partir duquel l'espérance de nos excès était linéaire.

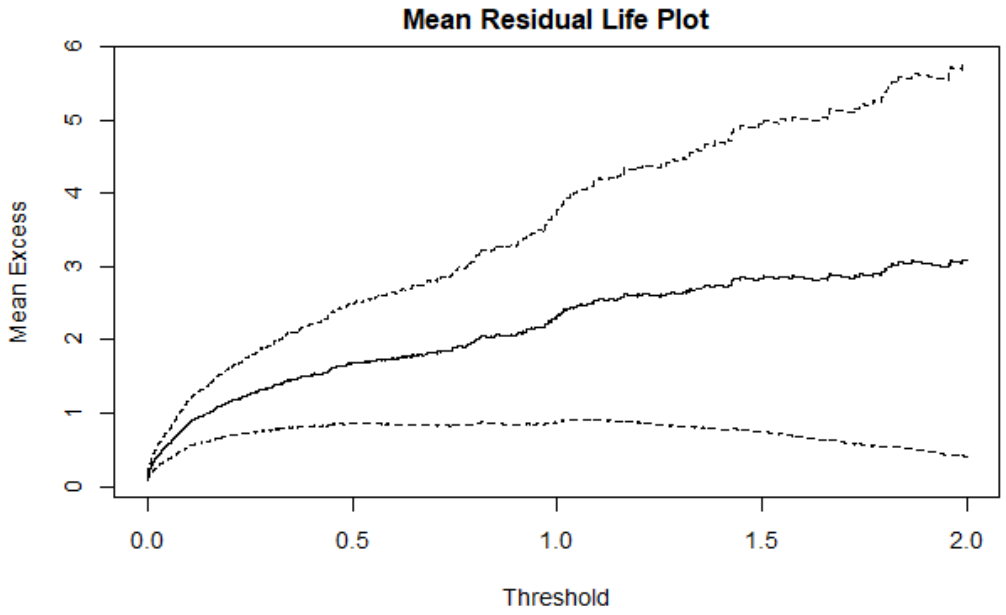


FIGURE 13 – Analyse loi de Frechet

Nous avons déterminé trois seuils possibles : 0.5 ; 1 et 1.5.

Nous avons ensuite étudié la stabilité de notre GPD, avec les graphiques de stabilité de nos paramètres de forme et d'échelle.

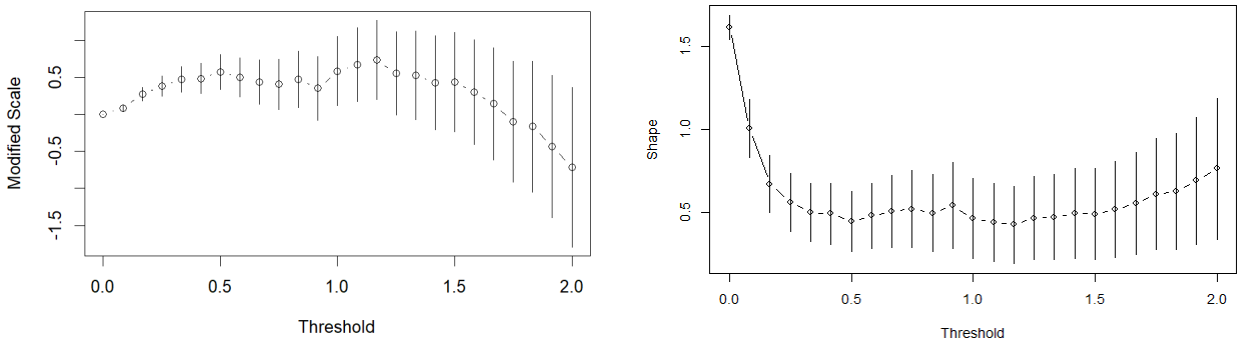


FIGURE 14 – Graphiques de stabilité des paramètres

En étudiant le graphique du paramètre d'échelle, nous observons qu'il est plutôt constant jusqu'à 1.5, avant de devenir décroissant.
Concernant le graphique de forme, il n'est pas du tout constant pour des faibles valeurs (inférieur à 0.5).

Nous avons ensuite étudié l'erreur des paramètres. Nous observons que cette erreur est plus faible pour les seuils de 0.08 et 0.5, et est similaire pour les autres.

Estimation	Seuil	Nombre d'excédents	Shape	Scale	Erreur Shape	Erreur Scale
90^{th}	0.08	686	1.04	0.15	0.01	0.08
$\sqrt{n}$	1.45	81	0.47	1.23	0.2	0.13
$\frac{n^{\frac{2}{3}}}{\log(\log(n))}$	0.74	164	0.51	0.81	0.11	0.11
Graphique 1	0.5	213	0.45	0.45	0.09	0.09
Graphique 2	1	118	0.47	1.05	0.15	0.12
Graphique 3	1.5	80	0.49	1.17	0.19	0.14

Cependant, d'après nos graphiques, un plus grand seuil serait pertinent à choisir. Avoir un seuil plus élevé nous permettra de réduire la variance des excédents et de mieux décrire les valeurs très extrêmes.

Nous décidons donc de choisir un **seuil de 1**.

3.4 Copule et distributions multidimensionnelles des extrêmes

Après avoir dans un premier temps étudié nos valeurs extrêmes de la variable *death_percmille* de façon univariée, nous allons désormais étudier de façon bivariée, en étudiant la distribution jointe de cette variable avec la variable *PCAP*. Nous allons chercher la relation de dépendance de ces variables, en étudiant leur copule.

Dans un premier temps, nous étudions la corrélation de Spearman de nos variables, qui est de 0.46, donc assez faible.

Nous avons ensuite (grâce à un algorithme d'ajustement), sélectionné la copule bivariée la plus adaptée à nos variables, qui est la **Rotated BB8 270 degrees**.
Les paramètres de cette copule (estimés grâce au maximum de vraisemblance) sont :

- par = -2.83
- par2 = -0.72
- tau = -0.27

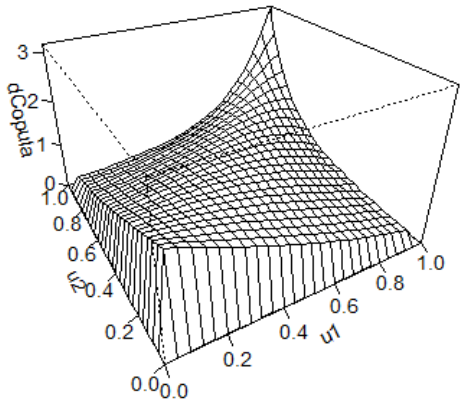


FIGURE 15 – Densité de la copule

Cette copule présente une très forte concentration de points au coin (1,1), et une forte concentration au coin (0,0), et plus faible sur le reste du nuage.
Ainsi, nos valeurs extrêmes semblent dépendantes. Un très fort PIB est lié à une très forte mortalité, tandis qu'un très faible PIB est lié à une très faible mortalité.

Nous allons maintenant construire la copule et en tirer 10 000 échantillons aléatoires.
Nous obtenons ce tracé des échantillons contenus dans le vecteur u :

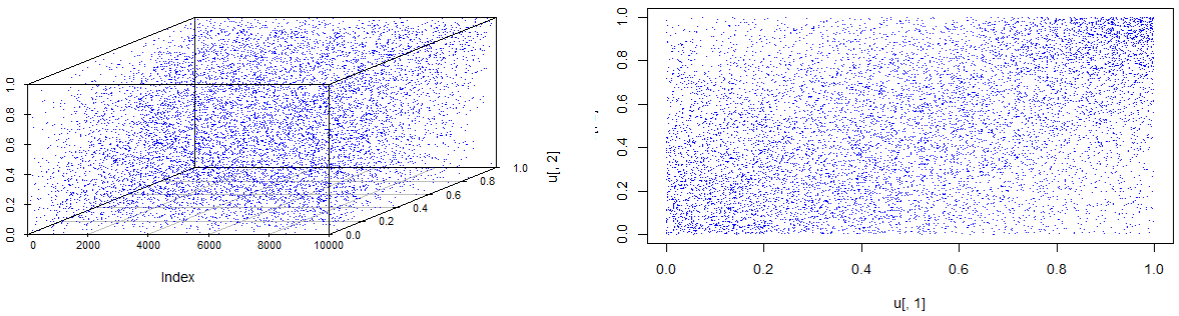


FIGURE 16 – Densité de l'echantillon

L'échantillon aléatoire semble proche du cas de l'indépendance, car les points sont étalés. Cela est cohérent avec la corrélation que nous avions calculée dans une première partie, qui était de 0.46, ce qui n'est pas très élevé.
Nos échantillons ont donc une corrélation similaire à nos données.

Nous observons cependant une concentration au niveau des maximas des 2 axes, ce qui montre une relation pour nos variables extrêmes, et confirme ce que l'on avait observé sur la densité de la copule.

4 Modélisation

Nous avons d'abord essayé de construire des modèles avec notre variable target, *death_ percmille*, mais la répartition de nos valeurs (un grand nombre proche de 0, et des extrêmes) rendait la lisibilité graphique très mauvaise.

Nous avons donc choisi une variable avec des données mieux réparties, et mieux corrélées avec les autres variables, le PIB par habitant, nommée *PCAP*.

Pour sélectionner les variables explicatives de nos modèles, nous avons étudié les corrélations de Pearson (pour les modèles linéaires) et de Kendall (pour les modèles polynomiales et non paramétriques).

Pour tous nos modèles, nous avons séparé notre population en 2 jeux de données : un **training** (80% de la population) et un **testing** (20%).

4.1 Modèle paramétrique, modèles mixtes

Variables sélectionnées pour chaque modèle :

— Régression linéaire univariée : *democracy_ polity*
— Régression linéaire multivariée : *democracy_ polity*, *Year*, *Nb_ cat* et *Nb_ conf.*
— Régression polynomiale : *Total.Pop*

Nous avons détecté nos outliers grâce à la **méthode de Cook**, que nous avons ensuite supprimés.

Nos modèles paramétriques ne donnant pas de bons résultats, nous avons utilisé des modèles mixtes, en groupant les données en fonctions du pays.
Nous avions donc 185 groupes différents.

Voici les AIC de nos modèles paramétriques mixtes et non mixtes.

	Régression linéaire univariée	Régression linéaire multivariée	Régression polynomiale
Modèle simple	119106.7	118825.4	119519.7
Modèle mixte	113558.2	111828.5	114057.1

Nous observons que pour chaque type de modèle, le modèle mixte a un AIC plus faible que le modèle non mixte.
Le meilleur modèle est celui de la régression linéaire multivariée en modèle mixte, tandis que le moins bon est le polynomial non mixte.

Nous allons nous concentrer sur l'analyse du modèle mixte univarié ici, mais l'analyse des autres modèles est disponible dans l'annexe.

Voici le modèle univarié ajusté à chacun des groupes :

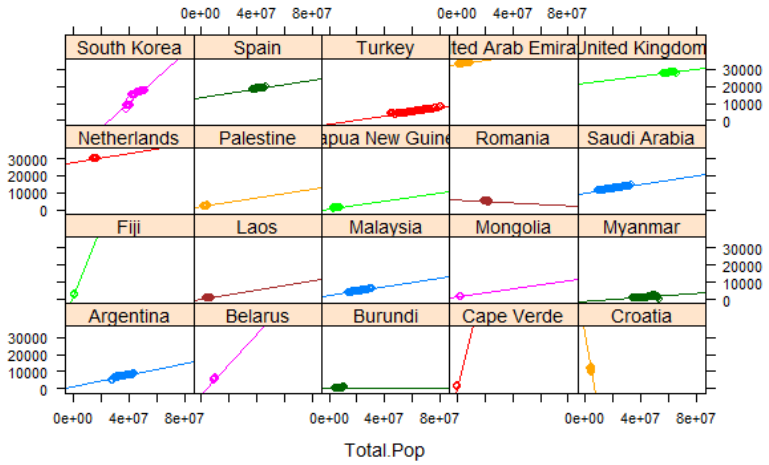


FIGURE 17 – Modèle mixte univarié

Notre modèle a un R^2 ajusté de 0.69, ce qui est bien, mais en observant le modèle, nous remarquons que nous n'avons pas un modèle linéaire.

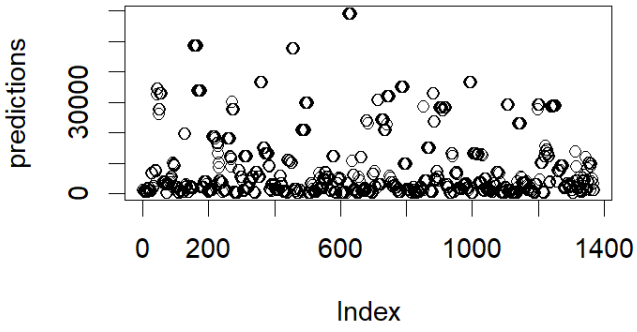


FIGURE 18 – Modèle mixte univarié

De plus, nous n'avons pas de droite $x = y$ qui apparait dans le graphique des valeurs prédites en fonction de celles observées.
Enfin, nous observons que nos résidus ne suivent pas une loi centrée réduite.

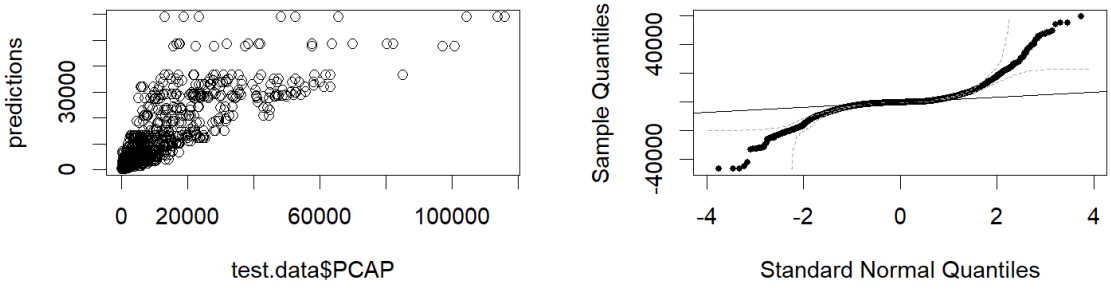


FIGURE 19 – Résidus

Notre modèle n'est donc pas bon.
Nous pouvons cependant observer dans notre graphique 2 un triangle inférieur vide, notre modèle ne prédit pas de valeur faible pour celle qui sont réellement grandes.

4.2 Modèles non paramétriques

Nos modèles paramétriques n'étant pas adaptés à nos données, nous avons essayé de construire des modèles de régression non paramétriques (Kernel, Smoothing Splines et LOESS).

Pour les modèles non paramétriques, nous avons pris comme variable explicative le nombre d'habitant, qui était la variable continue avec la plus forte corrélation de Spearman avec *PCAP*.

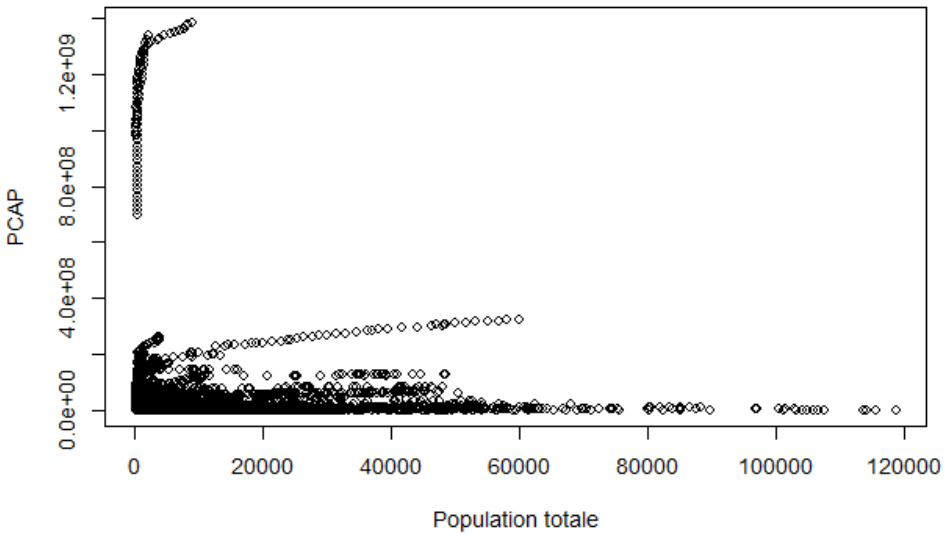


FIGURE 20 – PIB en fonction de la population

En représentant notre variable *PCAP* en fonction de la variable *Total.Pop*, nous nous sommes aperçus que nous avions des pseudoréplications. Cela est dû au fait que nous avions plusieurs données par pays (une par année), ce qui créait de la dépendance dans nos données.

Nous avons donc décidé de prendre la moyenne des données par pays sur les 38 années, en ne gardant donc qu'une seule ligne par pays, et non plus 38.

4.2.1 Kernel

Nous avons d'abord construit un modèle de Kernel gaussien.
Nous avons testé plusieurs modèles, en faisant varier le paramètre de lissage : 0.3, 0.5, 0.75 et 1.

En comparant les modèles avec leur AIC, nous avons pu sélectionner le meilleur modèle, qui est celui avec un paramètre de lissage de 0.5

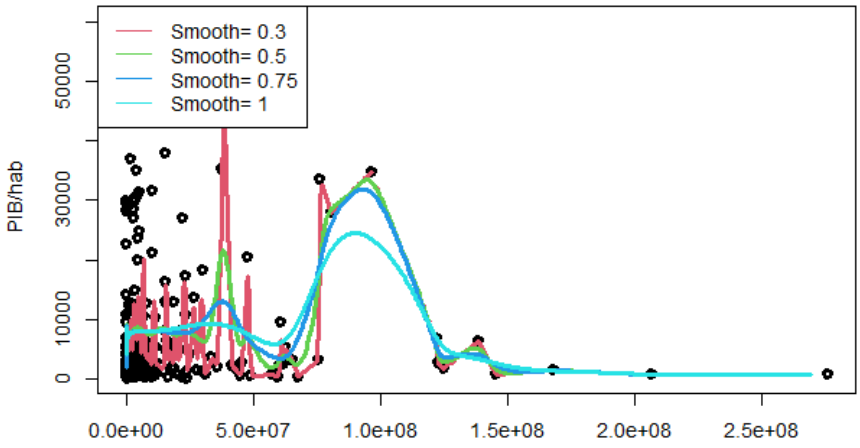


FIGURE 21 – Modèle Kernel

4.2.2 Smoothing Spline

Nous avons ensuite testé le modèle de régression de Smoothing Spline.
Nous avons testé plusieurs paramètres de lissage : 0.3, 0.5, 0.75 et 1.

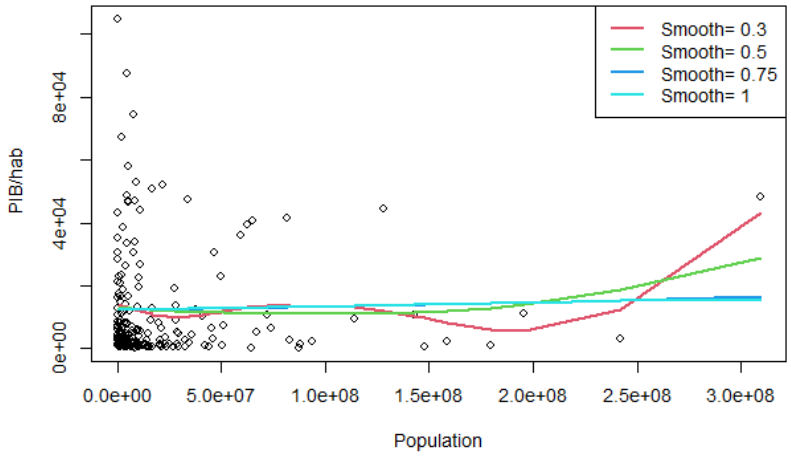


FIGURE 22 – Modèle Smooting Spline

En comparant les modèles avec leur AIC, nous avons pu sélectionner le meilleur modèle, qui est celui avec un paramètre de lissage de 0.3.

4.2.3 LOESS

Notre dernier modèle non paramétrique était le LOESS. Nous avons également fait varier le paramètre de lissage pour notre modèle : 0.3, 0.5, 0.75 et 1.

En comparant les modèles avec leur AIC, nous avons pu sélectionner le meilleur modèle, qui est celui avec un paramètre de lissage de 0.3.

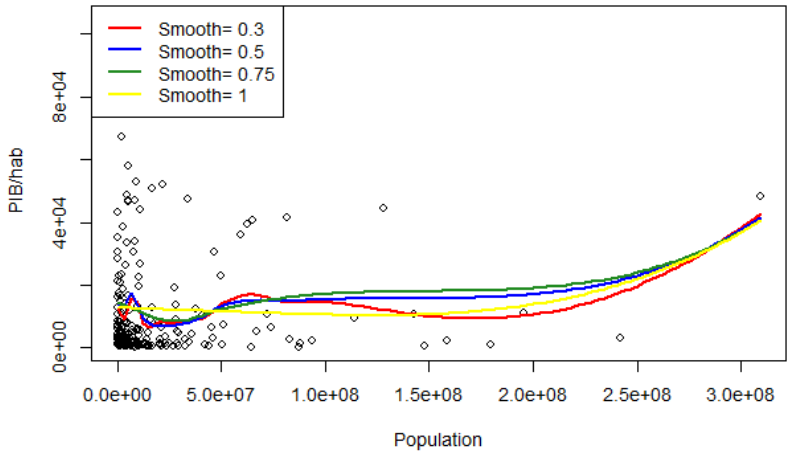


FIGURE 23 – Modèle LOESS

4.2.4 Comparaison de nos modèles

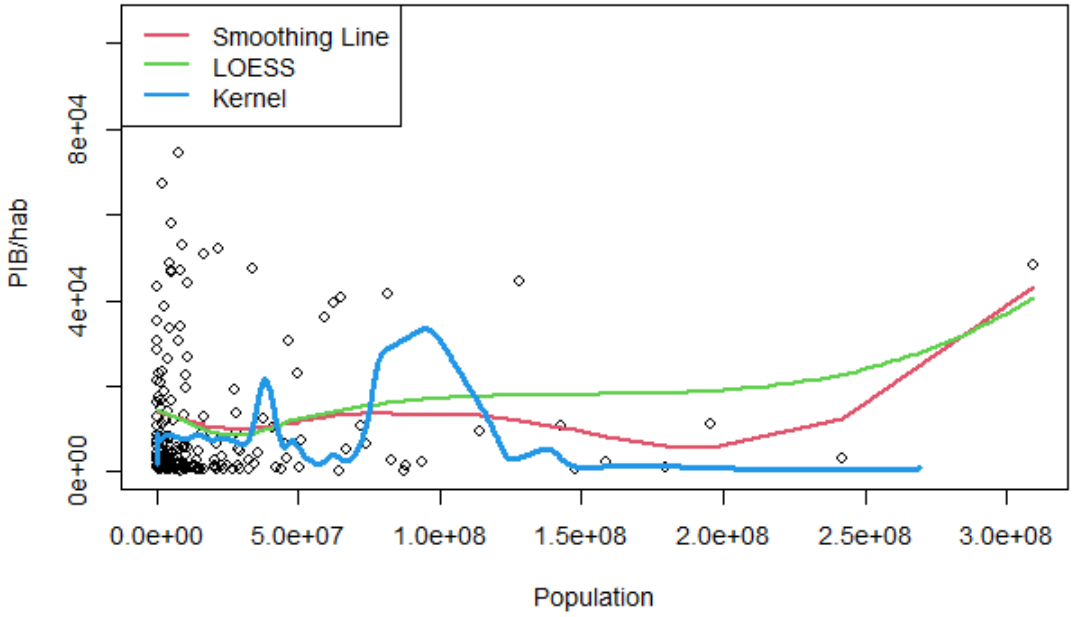


FIGURE 24 – Comparaison des modèles

Ainsi, nous observons que nos modèles sont influencés par les grandes valeurs de x, car moins nombreuses. Ils expliquent donc mieux les grandes valeurs que les faibles valeurs de x, trop nombreuses et dispersées selon l'axe y.

En testant l'AIC, nous observons que le meilleur modèle est celui du Smoothing Spline. Ce modèle a cependant un R^2 ajusté de 0.21 et une p-value de 0.6494. Notre modèle n'explique donc pas tout et n'est pas significatif.

4.3 GAM

Pour ce modèle, nous avons sélectionné 3 variables explicatives : *Total.Pop*, *Nb_epi* et *democraty_polity*.

Nous vérifions tout d'abord grâce à un test de $\tilde{\chi}^2$ s'il existe un lien entre les variables, pour pouvoir effectuer un GAM. Aucune de notre variable n'était liée.

Ensuite, nous avons représenté notre variable target *PCAP* en fonction des variables explicatives, pour chercher si une fonction de lien apparaissait. Ce n'était pas le cas.

Nous laissons donc notre GAM trouver pour nous les fonctions de lien pour chacune des 3 variables.

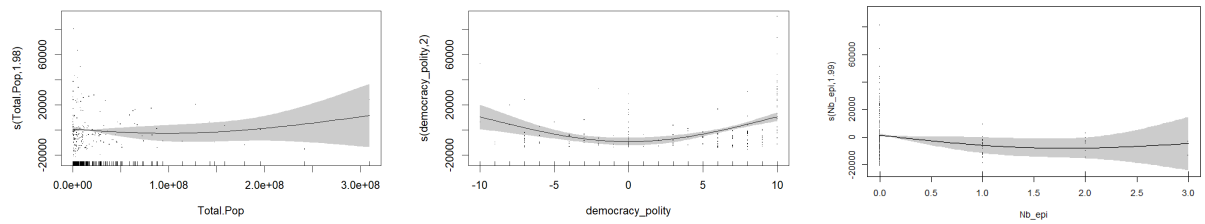


FIGURE 25 – Fonction de lien de chaque variable

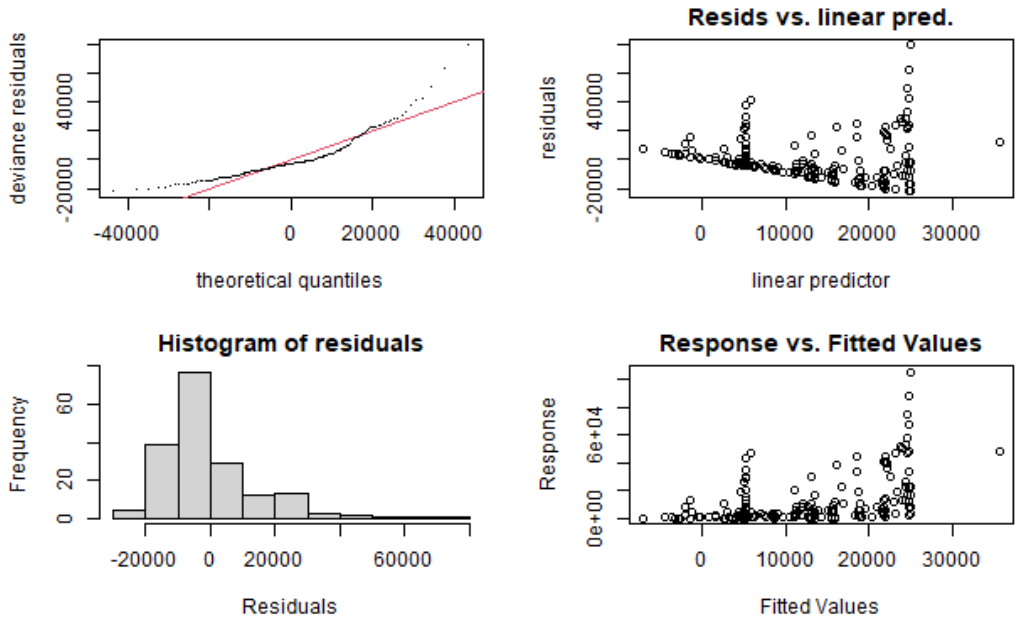


FIGURE 26 – Résidus du modèle GAM

En étudiant les résidus de notre modèle, nous observons qu'ils ne suivent pas une loi normale. De plus, sur le graphique des données estimées en fonction des données empiriques, aucune droite x = y n'apparaît.
Notre modèle n'est donc pas bon.

5 Simulation

5.1 Monte Carlo

Nous avons essayé de simuler les valeurs de *death_percmille* avec Monte-Carlo. Pour cela, nous avons d'abord simulé le Nombre de mortss par risque, puis avons additionné ces morts.

Nous avons effectué 1000 simulations de chacune de ces variables, pour obtenir 10000 valeurs de morts par cause.
Nous avons ensuite calculé la moyenne de ces simulations, que nous avons comparée à la moyenne des morts.

	Famine	Epidemie	Conflit et terrorisme	Catastrophe naturelle	Autre	**Total**
Estimé	5.1×10^{-8}	0.0025	0.0017	0.0019	0.024	0.316
Observé	9.3×10^{-8}	0.0018	0.0019	0.0021	0.029	0.353

Pour chaque cause, la moyenne des estimations est proche de la moyenne des données observés. La moyenne des données estimée est légèrement inférieure à celle observée, mais reste très proche.

5.2 Bootstrap

Nous avons implémenté Bootstrap sur le modèle de régression linéaire multivariée. Ne connaissant pas la loi de nos valeurs, nous avons effectué un Bootstrap non paramétrique, afin d'estimer les paramètres de nos lois.
Nous avons effectué 200 réplications.

Parameter	Coefficient	SE	95% CI	t(6849)	p
(Intercept)	-4.35e+05	28243.03	[-4.91e+05, -3.80e+05]	-15.41	< .001
democracy polity	493.96	24.30	[446.32, 541.60]	20.33	< .001
Nb cat	3.93	47.73	[-89.63, 97.49]	0.08	0.934
Nb conf	-2145.60	197.62	[-2533.00, -1758.20]	-10.86	< .001
Year	221.69	14.14	[193.97, 249.40]	15.68	< .001

FIGURE 27 – Paramètres sans Bootstrap

Parameter	Coefficient	95% CI	p
(Intercept)	-4.43e+05	[-4.88e+05, -3.86e+05]	< .001
democracy polity	497.49	[460.13, 561.25]	< .001
Nb cat	-0.72	[-135.28, 120.59]	0.980
Nb conf	-2158.46	[-2467.83, -1919.86]	< .001
Year	225.42	[197.03, 248.05]	< .001

FIGURE 28 – Paramètres avec Bootstrap

Nous remarquons que nos coefficients sont semblables, hormis celui du nombre de catastrophe, qui passe de 3.93 à -0.72. Les intervalles de confiance sont également similaires hormis pour celui du nombre de catastrophe qui est bien plus grand pour le Bootstrap que pour le normal. Enfin, les p-values sont similaires, et inférieures à 0.05, hormis pour le nombre de catastrophe.

6 Séries temporelles

6.1 Visualisation et description

Ayant des données qui évoluaient au fil du temps, nous nous sommes dits qu'il pouvait être intéressant d'étudier nos données sous forme de Times Series, afin d'analyser leur comportement pour essayer de prévoir leur comportement futur.

Cependant, notre variable *death_percmille* n'avait que des données annuelles.
Nous avons donc pris la base de données qui liste les attaques terroristes, et celle qui liste les catastrophes naturelles.

À partir de cela, nous avons obtenu des données mensuelles sur le nombre d'évènements par mois, par pays.

Nous nous sommes concentrés sur un pays, l'Inde, entre 1980 et 2018, et sur le nombre d'évènements, modélisés par la variable *count*.

Nombre d'attaques terroristes en Inde

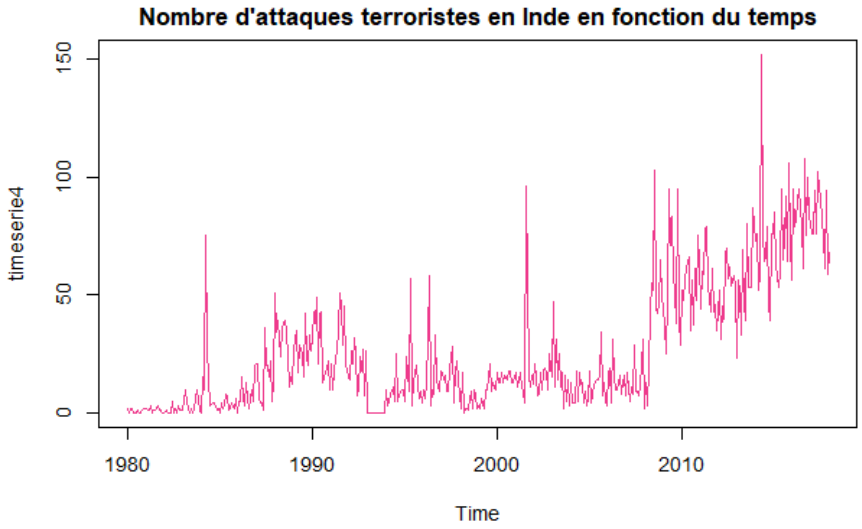


FIGURE 29 – Série temporelle des attaques terroristes

La série temporelle était non stationnaire, sans saisonnalité ni tendance (cf. annexe).
Cela est logique, car les attaques terroristes sont des évènements aléatoires et non prévisibles.

Nous n'avons donc pas tenté de construire un modèle dessus.

Nous pouvons cependant noter qu'il semble y avoir une augmentation du nombre d'attaques terroristes depuis 2008.

Nombre de catastrophes naturelles en Inde

Nous nous sommes concentrés sur les catastrophes naturelles en Inde, qui devrait être plus prédictible.

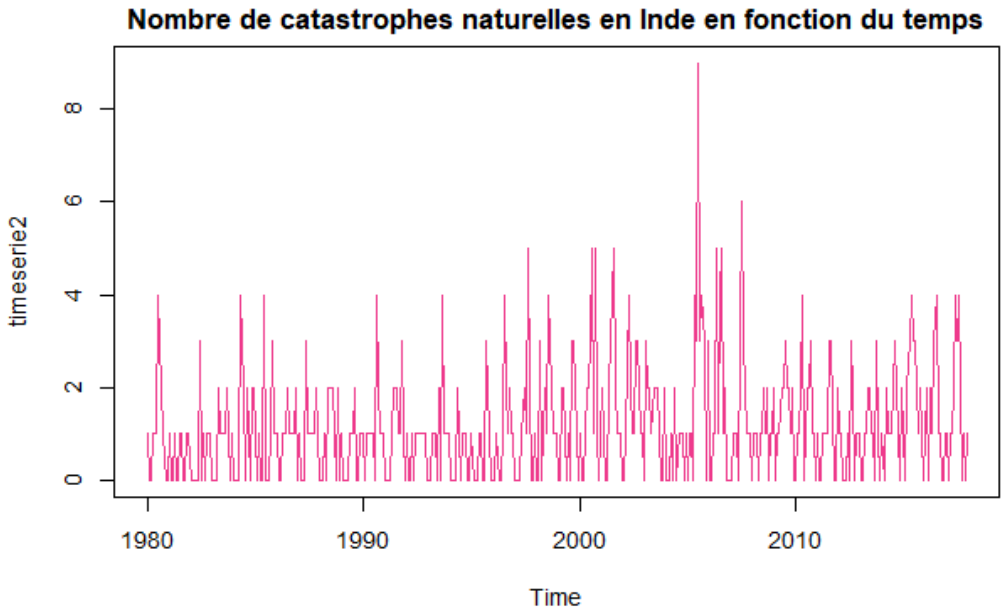


FIGURE 30 – Série temporelle des catastrophes naturelles

En représentant le nombre de catastrophes naturelles en fonction des mois, nous observons qu'il semble y avoir plus de catastrophes naturelles dans les mois de mai à septembre que le reste de l'année.

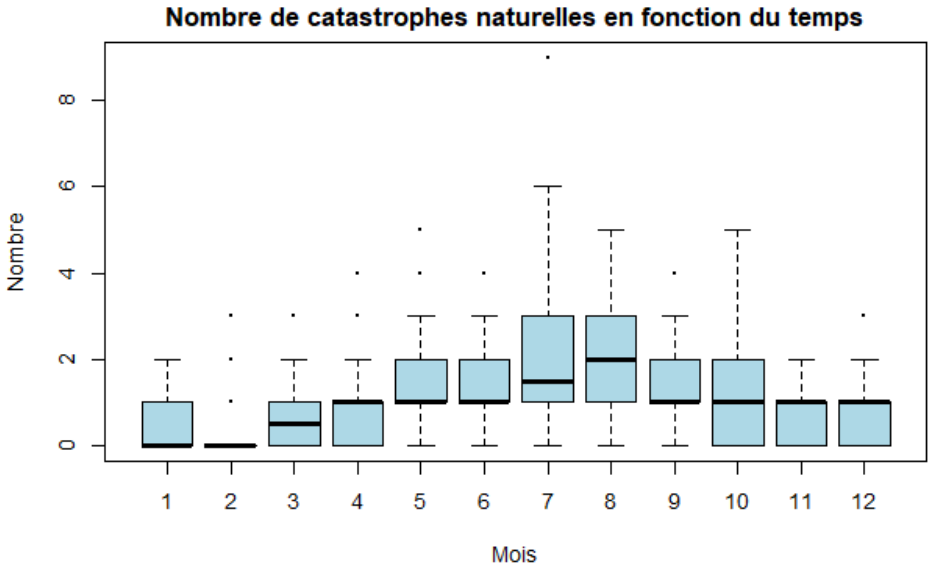


FIGURE 31 – Catastrophes en fonction du mois

Nous avons testé la stationnarité avec le **test de Dickey-Fuller** en prenant comme hypothèse nulle *la série est stationnaire.*
Nous obtenons une p-value de **0.01**, notre série n'est donc pas stationnaire.

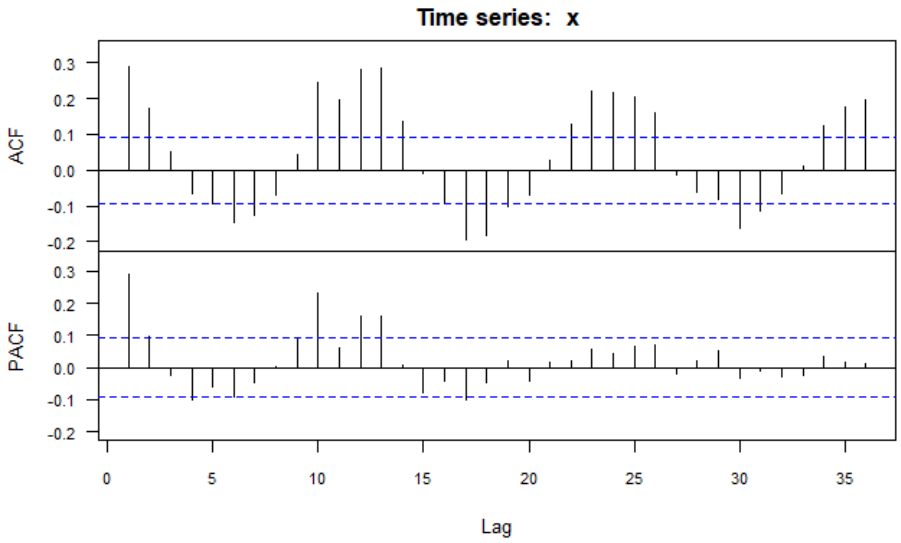


FIGURE 32 – Autocorrélogrammes

La fonction d'autocorrélation nous confirme que notre série n'est pas stationnaire, nous observons en effet une périodicité dans la fonction d'autocorrélation, indiquant une saisonnalité.

Cette saisonnalité est annuelle, car nous avons une autocorrélation forte au lag 12.

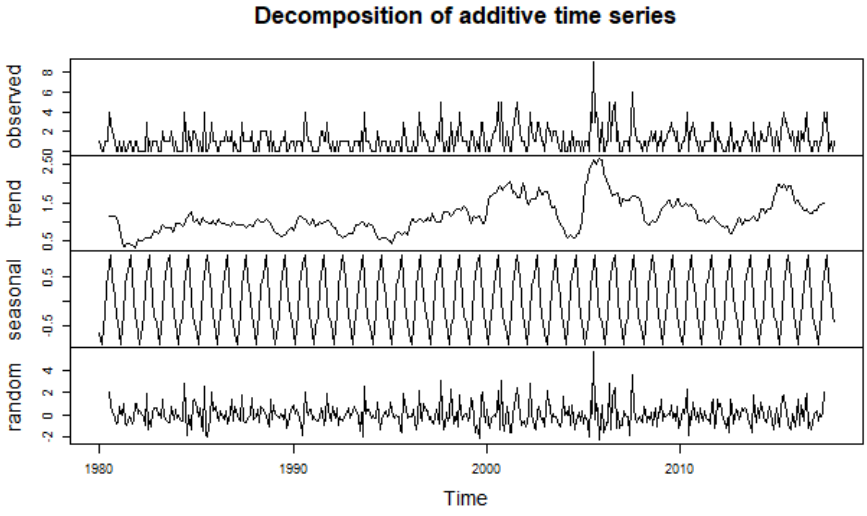


FIGURE 33 – Décomposition de la série temporelle

La décomposition de notre série nous confirme que nous avons une saisonnalité dans nos données.

6.2 Stationarité

Afin de pouvoir implémenter des modèles, nous allons chercher à obtenir une série stationnaire, c'est-à-dire à enlever la saisonnalité et la tendance.

Nous avons d'abord enlevé la saisonnalité de notre série temporelle (en la soustrayant), puis nous avons différencié ce résultat.

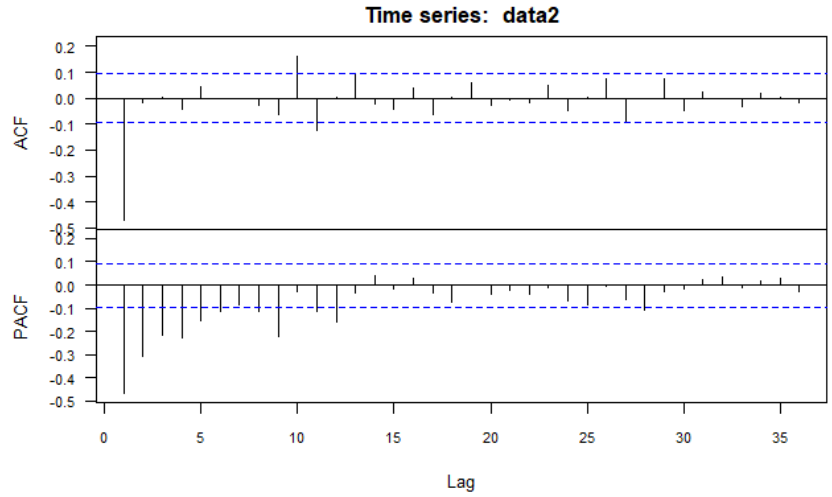


FIGURE 34 – Autocorrélogrammes

Nous observons sur nos autocorrélogrammes que notre modèle est (quasi) stationnaire.

6.3 Modélisation

Pour nos modèles, nous avons séparé nos Times series en 2 parties : une partie **train**, comprenant les données allant de 1980 à 2017, et une partie **test**, comprenant les données allant de 2017 à 2018.

6.3.1 Simple Forecasting methods

Nous avons commencé par faire des modèles simples, utilisant les dernières observations pour prédire.
Nous avons utilisé 3 méthodes :

- Average method : On utilise la moyenne des anciennes données.
- Naive method : On utilise la valeur de la dernière observation.
- Seasonal naive method : On utilise la dernière valeur de la même saison.

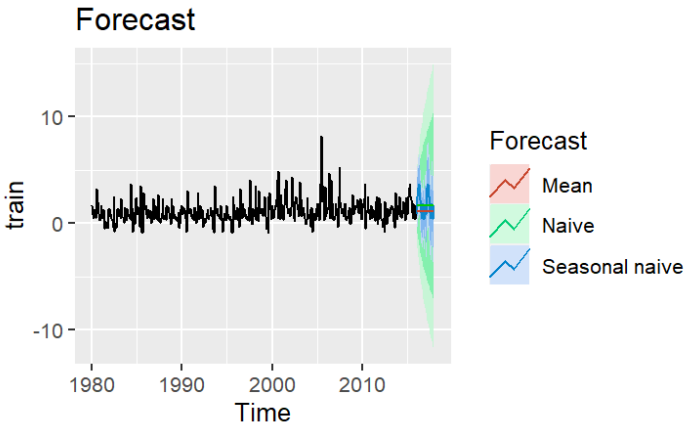


FIGURE 35 – Modèle forecasting simple

Visuellement, nous observons que les modèles de mean et naive ne sont pas bons, car trop simple. Nous allons nous concentrer sur le troisième modèle.

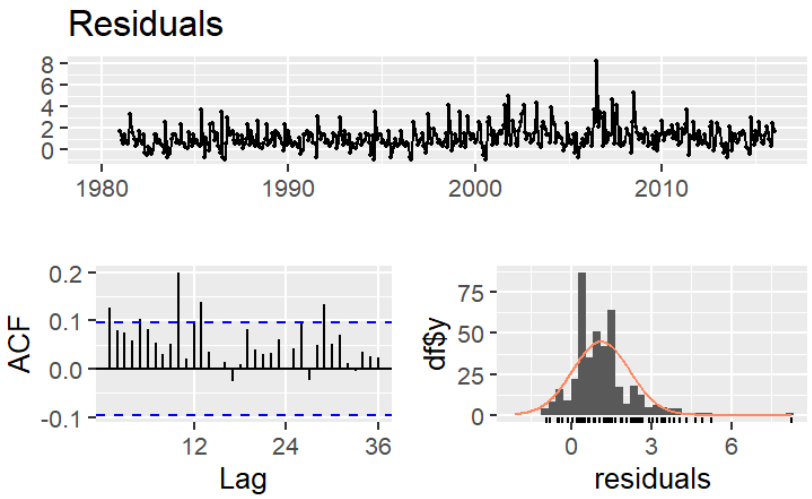


FIGURE 36 – Résidus du modèle

Les résidus semblent à première vue suivre une loi centrée réduite. Leur variance n'est cependant pas constante (nous observons une légère tendance) et l'autocorrélogramme montre des autocorrélations significatives.
De plus, en effectuant le **test de Ljung-Box**, nous obtenons une p-value de 4.4×10^{-16}, il est donc peu probable que les résidus forment un bruit blanc.

6.3.2 Holt Winter

Notre série temporelle contenant une saisonnalité, nous allons donc utiliser la méthode de Holt Winter pour construire notre modèle. Notre saisonnalité étant constante, nous utiliserons la méthode additive.

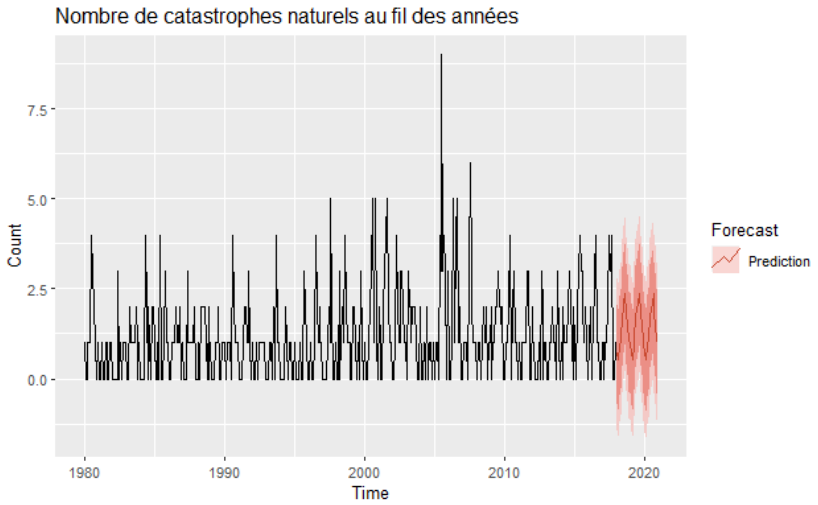


FIGURE 37 – Modèle Holt Winter

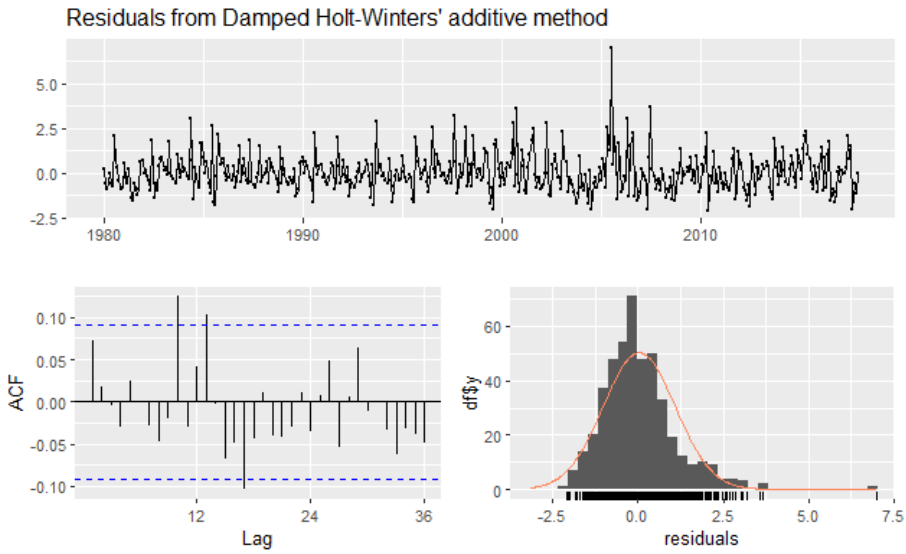


FIGURE 38 – Résidus du modèle

Ici, nos résidus semblent suivre une loi normale centrée réduite, et leur variance est constante (même s'il y a quelques pics).
Cependant, notre autocorrélogramme montre toujours des autocorrélations significatives.

6.3.3 SARIMA

Pour notre modèle SARIMA, nous avons utilisé la série stationnaire que nous avions trouvée précédemment.

Grâce à l'ACF et le PCAF, nous avons pu déterminer nos paramètres :

- L'ordre de la partie autorégressive (AR) : 7
- L'ordre de la moyenne mobile (MA) : 6

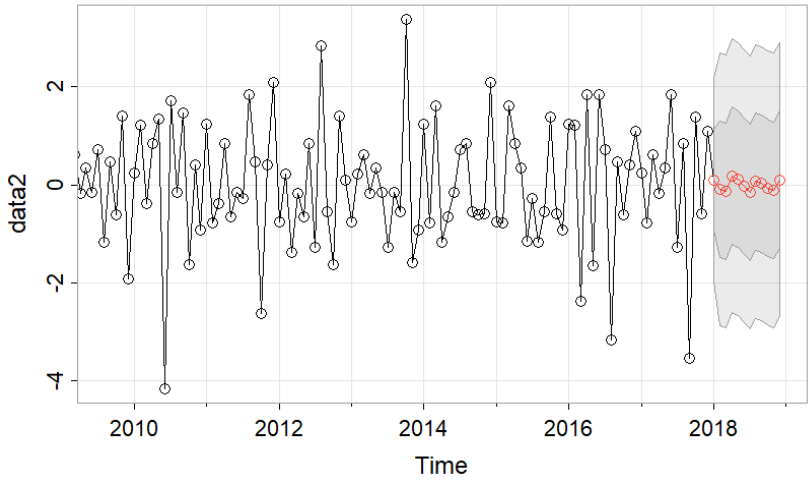


FIGURE 39 – Résidus du modèle

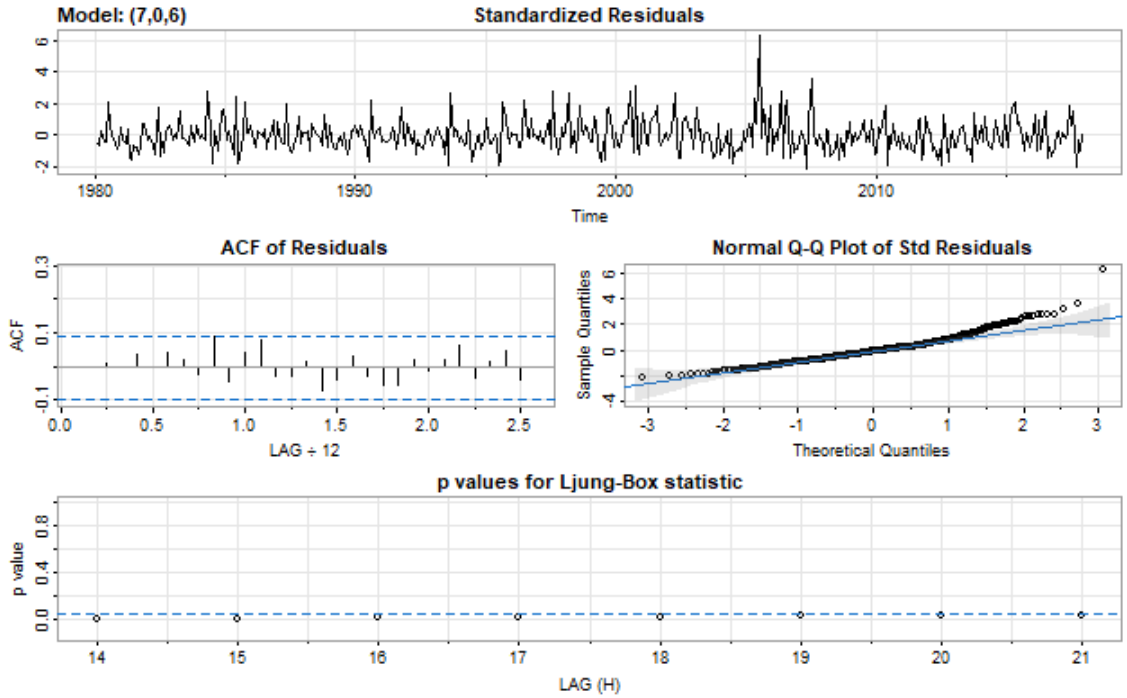


FIGURE 40 – Résidus du modèle

Ici, nos résidus suivent bien une loi normale et sont indépendants.
De plus, notre modèle a bien des résidus avec une série stationnaire, qui n'ont pas d'autocorrélation pour chaque lag. Enfin, les tests de Ljung Box sont tous sous la ligne bleu.

Notre modèle est donc bon.

7 Conclusion

Cette étude s'est décomposée en plusieurs étapes :

Dans un premier temps, nous avons appliqué la théorie des valeurs extrêmes à notre variable *death_ percmille*, en calculant les mesures de risques, en étudiant les maximums (avec la loi d'extremum généralisé) et les excès (avec la loi de Pareto), et en observant la dépendance des extrêmes avec les copules.

Nous avons ensuite essayé de construire des modèles paramétriques et non paramétriques sur notre variable *PCAP*.
Pour les modèles paramétriques, nous avons utilisé des régressions linéaires simple et multivarié, des régressions polynomiales, en modèle simple ou mixte. Pour les modèles non paramétriques, nous avons utilisé Kernel, Smoothing Spline, LOESS et les GAM.
Hélas, nous n'avons pas obtenu de bons résultats, aussi bien pour les modèles paramétriques que pour les non paramétrique. Cela est probablement dû à un manque de variables explicatives sur nos pays, qui pourraient mieux expliquer le PIB.

Dans un troisième temps, nous avons effectué des simulations, de Monte Carlo et de Bootstrap.

Nous avons fini notre étude par étudier les Série Temporelles du nombre d'attaque terroriste et du nombre de catastrophe naturelle en Inde.
La première série temporelle ne nous a pas données de résultat concluant, cela était dû au caractère très aléatoire des attaques terroristes.
Cependant, la deuxième série temporelle était beaucoup plus intéressante à étudier, et nous avons réussi à construire un bon modèle SARIMA.

8 Annexe

8.1 Analyse modèle paramétrique

Régression linéaire univariée

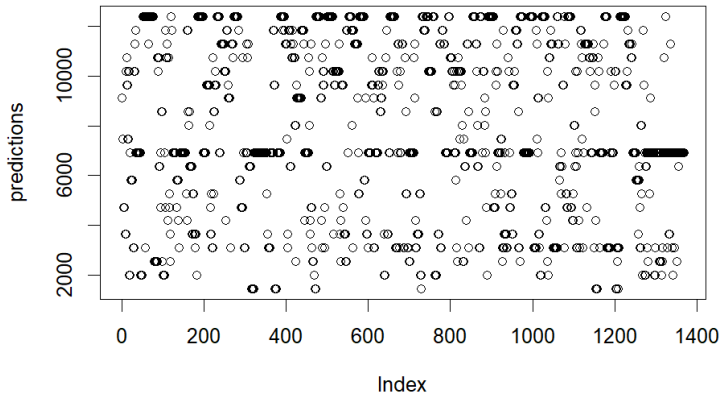
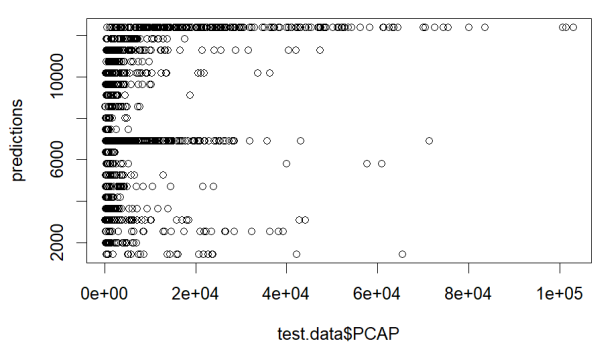


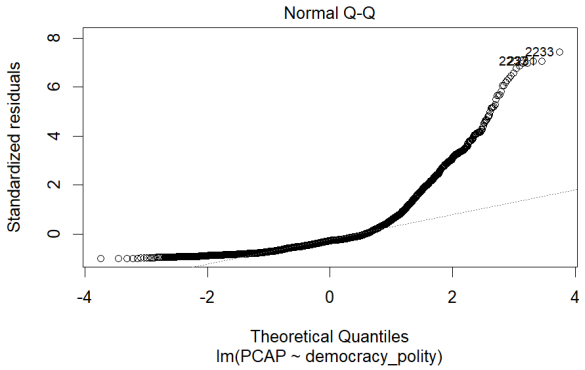
FIGURE 41 – Modèle univarié

Notre modèle n'a pas une apparence linéaire du tout.

Bien que sa p-value soit inférieur à 0.05, son R^2 est de 0.09, ce qui est très faible.
Notre modèle n'explique pas beaucoup la variance des données.



(a) Prédiction vs valeurs observées



(b) Résidus

L'étude du graphique des notes prédites en fonction des observées, ainsi que celui des résidus, nous confirme que notre modèle n'est pas bon.

En effet, aucune droite x = y ne se dessine, et nos résidus ne suivent pas une loi normale centrée.

Notre modèle n'est donc pas bon.

Régression linéaire bivariée

Encore une fois, notre modèle n'a pas une apparence linéaire du tout.

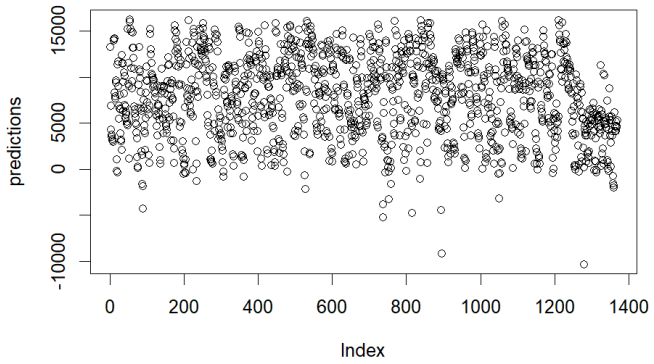
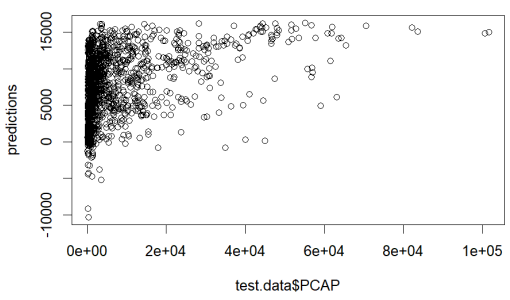
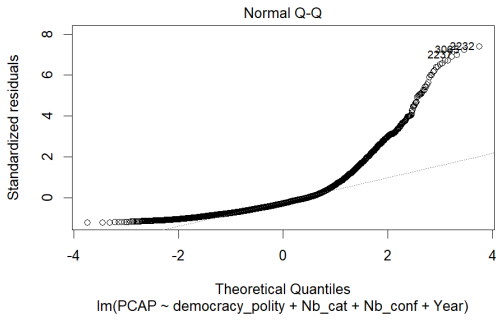


FIGURE 43 – Modèle univarié

La p-value est toujours inférieur à 0.05,mais ici le R^2 est meilleur, il vaut 0.19.
Cela reste très faible, notre modèle n'explique pas beaucoup la variance des données.



(a) Prédiction vs valeurs observées



(b) Résidus

L'étude du graphique des notes prédites en fonction des observées, ainsi que celui des résidus, nous confirme que notre modèle n'est pas bon.
En effet, aucune droite x = y ne se dessine, et nos résidus ne suivent pas une loi normale centrée.

Régression polynomial

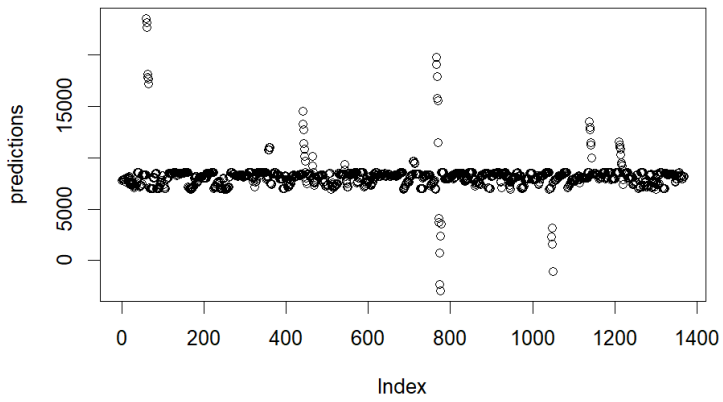
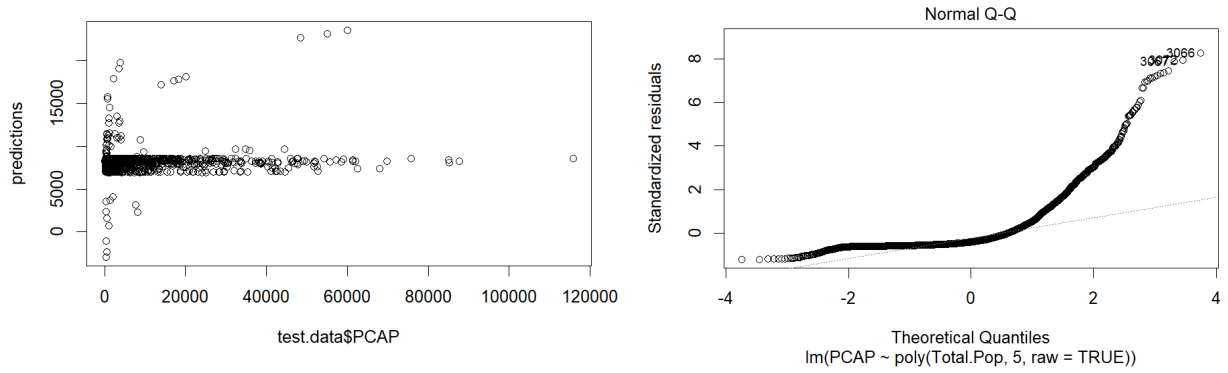


FIGURE 45 – Modèle univarié

Notre modèle n'a pas l'apparence d'une fonction polynomiale. Une grande majorité des données sont alignées sur l'axe y = 0.
De plus le R^2 est de 0.012.

Notre modèle n'explique pas grand chose.



(a) Prédiction vs valeurs observées

(b) Résidus

L'étude du graphique des notes prédites en fonction des observées, ainsi que celui des résidus, nous confirme que notre modèle n'est pas bon.
En effet, aucune droite x = y ne se dessine, et nos résidus ne suivent pas une loi normale centrée.

8.2 Série temporelle sur le nombre d'attaques terroristes

Sur le boxplot du nombre d'attaques en fonction du mois, nous n'observons pas de mois avec plus d'attaques que les autres.

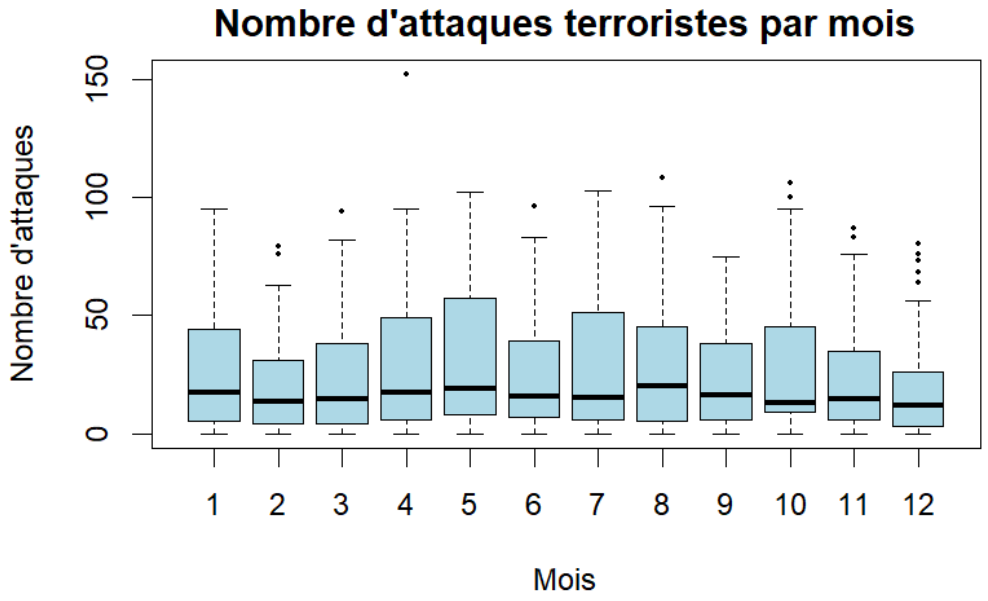


FIGURE 47 – Modèle univarié

L'autocorrélogramme montre que la série n'est pas stationnaire.

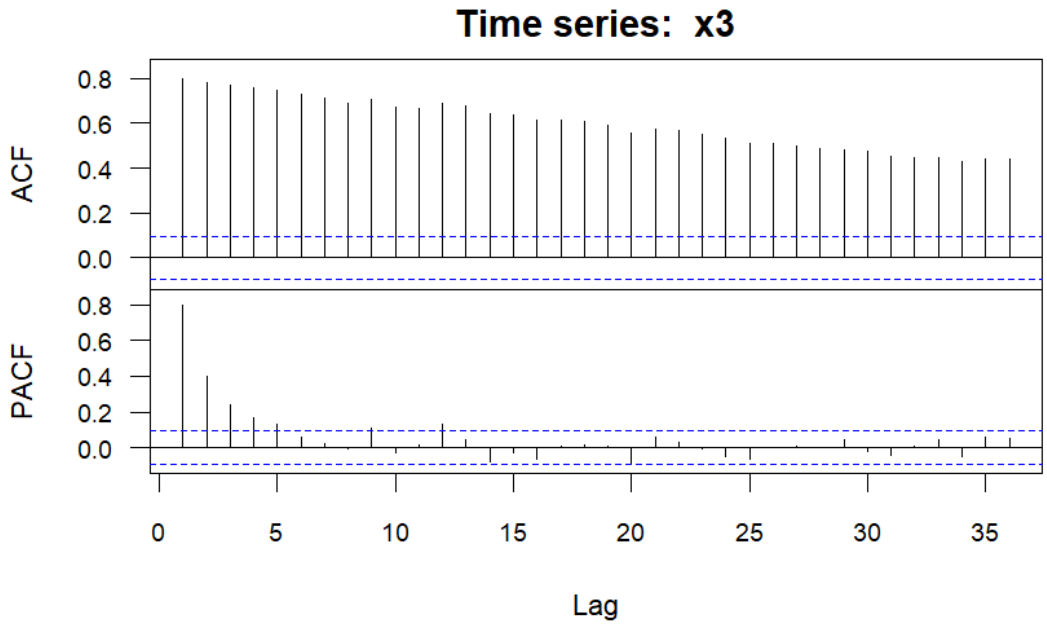


FIGURE 48 – Modèle univarié

La décomposition nous confirme que la série n'est pas stationnaire, en effet la variance du bruit n'est pas constante. On observe qu'il n'y a pas de tendance constante le long de la série temporelle, mais qu'il y à une très forte augmentation d'attaques depuis 2008.

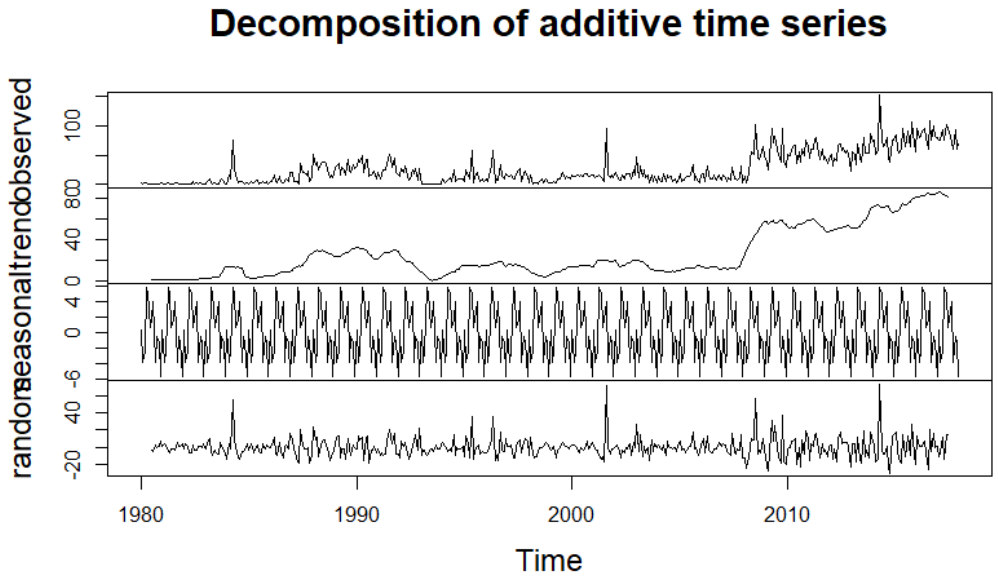


FIGURE 49 – Modèle univarié

Notre série semble donc difficilement exploitable pour construire des modèles, car n'est pas stationnaire, ne contient ni saisonalité, ni tendance apparente.