Mémoire Statistique

Axel Benyamine et Guillaume Février

MAP565 - Modélisation aléatoire et statistique des processus

1 Introduction

Dans ce mémoire, nous nous intéressons à l'étude de 4 jeux de données. Nous étudions chaque jeu de données à l'aide d'un outil lié à la modélisation aléatoire et statistique des processus.

Dataset 1 - Séries temporelles linéaires
Hauteur de la mer à Brest sur l'année 2023 :
https://data.shom.fr/donnees/refmar/3#001=eyJjIjpbLTUwMDM4NS41NTcy
MTYyMjcsNjE3MDc5My44NDMxMzg3NjFdLCJ6Ijo2LCJyIjowLCJsIjpbeyJ
0eXBlIjoiSU5URVJOQUxfTEFZRVIiLCJpZGVudGlmaWVyIjoiRkRDX0dFQ
kNPX1BZUi1QTkdfMzg1N19XTVRTIiwib3BhY2l0eSI6MSwidmlzaWJpbGl0
eSI6dHJ1ZX1dfQ==

Dataset 2 - Copules
Morts de la pneumonie et du COVID-19 aux Etats-Unis depuis avril 2022 :
https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab/about_data

Dataset 3 - Théorie des extrêmes
Retards des avions de ligne américains en janvier 2023 :
https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr$\bar{\text{b}}$0-gvzr

Dataset 4 - Processus de Hawkes
Activité sismique en Europe en 2024
https://earthquake.usgs.gov/earthquakes/search/

2 Hauteur de la mer à Brest sur l'année 2023

2.1 Présentation du dataset

Ce dataset provient des marégraphes de Brest et donne pour chaque minute depuis le 1er janvier 2023 le niveau de la mer. Nous nous restreignons à des données horaires, ce qui facilite l'analyse et nous semble pertinent au vu de l'ordre de grandeur des périodes des marées.

2.2 Tendance et saisonnalités

Nous nous intéressons à la série temporelle sur les 5 premiers mois de l'année 2023.

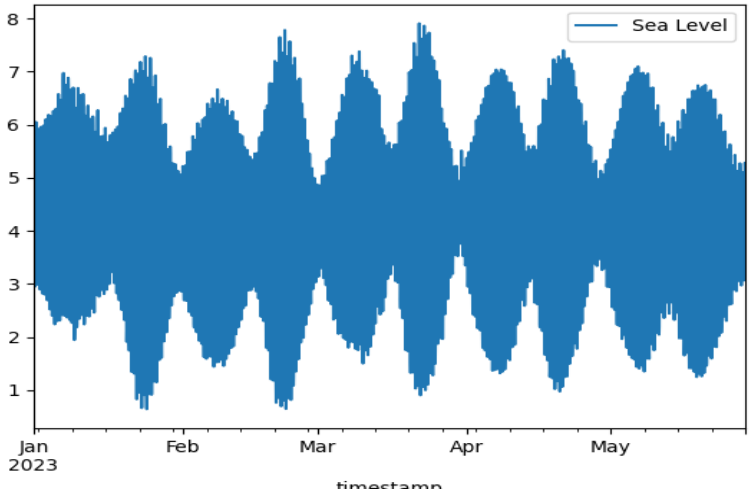

Figure 1: Niveau de la mer à Brest de janvier à mai 2023

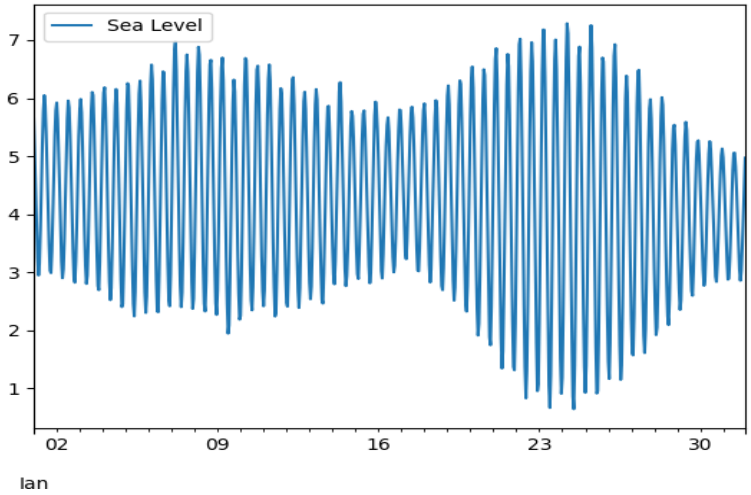

Figure 2: Niveau de la mer en janvier

Nous cherchons d'abord à identifier la tendance globale de la courbe. On effectue une régression linéaire qui donne un coefficient directeur de $1.27 * 10^{-6}$. Cela représente une augmentation d'environ 4,6 mm entre le 1er janvier et le 30 avril. Nous considérons donc que la série temporelle présente une tendance nulle vu l'ordre de grandeur du marnage.

Analysons cette série temporelle visuellement : on remarque tout d'abord une saisonnalité bi-journalière, ce qui est largement confirmé par le graphique d'autocorrélation de la série.

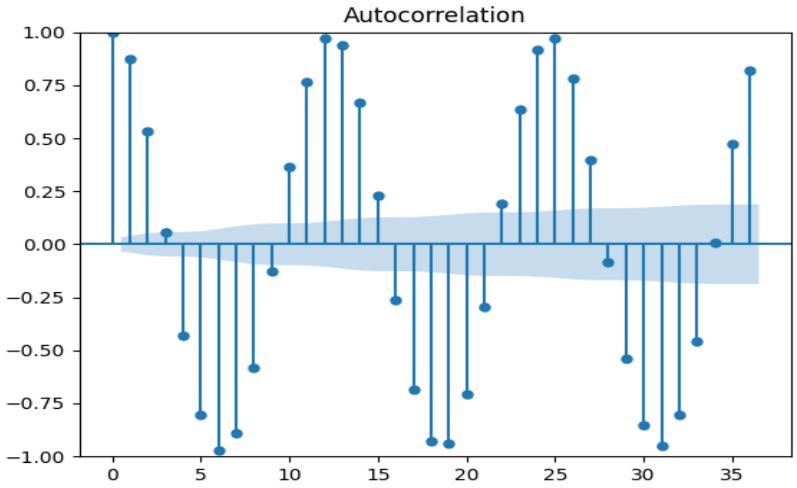

Figure 3: Autocorrélation de la série temporelle

Pour établir précisément la période associée, nous pouvons utiliser une transformation de Fourier. Avant de procéder à l'analyse, nous retranchons à la série sa moyenne pour ne pas avoir de pic en 0.

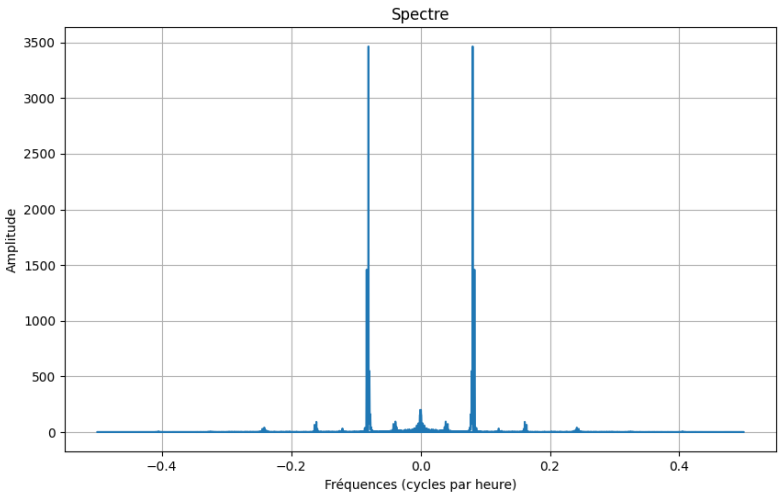

Figure 4: Transformée de Fourier de la série

En considérant le pic dominant, nous obtenons un nombre de cycles par heure d'environ 0.0806, ce qui donne une période d'environ 12.414h. Nous pouvons croiser ce résultat avec la durée théorique des marées qui est de 12.417h : l'estimation est très satisfaisante.

Notre objectif est d'utiliser un modèle SARIMA pour modéliser le processus. On peut remarquer qu'il existe une deuxième saisonnalité dans les données, qui correspond à la rotation de la Lune autour de la Terre. Cela se traduit par une deuxième période de 29.5j, soit 708h, qui rentre aussi en compte. Nous avons donc à présent une double saisonnalité pour le modèle, ce qui n'est pas pris en compte tel quel par SARIMA. Néanmoins, il est possible de contourner ce problème en utilisant des termes de Fourier comme variables exogènes pour le modèle. C'est la solution que nous avons décidé d'implémenter à l'aide de la librairie **pmdarima**.

En traçant la série différenciée selon les périodes 12.414 puis 708, on remarque une période résiduelle de 25h ; nous redifférencions donc une troisième fois pour obtenir une série semblant stationnaire :

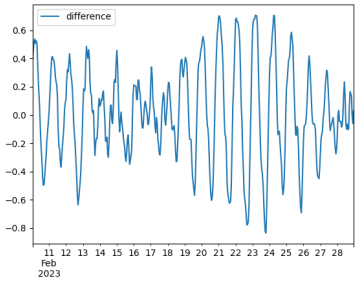

Figure 5: Série différenciée selon les périodes 12.414 puis 708

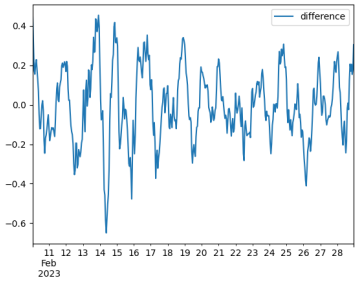

Figure 6: Série différenciée selon les périodes 12.414, 708 puis 25

2.3 Estimation des paramètres de SARIMA

A présent, nous testons la stationnarité de la série temporelle après différentiation selon les trois périodes mentionnées ci-dessus. On peut effectuer un test ADF (Augmented Dickey-Fuller). On obtient une p-valeur de l'ordre de 10^{-21}, ainsi qu'une valeur du test statistique est de -11.74, ce qui est inférieur à la valeur critique -3.43 correspondant à une p-valeur de 0.01. Le processus peut donc être considéré comme stationnaire. Les périodes 12.414 et 708 seront traitées comme variables exogènes, et nous prendrons $s = 25$ comme paramètre du modèle SARIMA, ainsi que $D = 1$ et $d = 0$.

Pour estimer (q, Q) et (p, P), on utilise respectivement les fonctions d'autocorrélation et d'autocorrélation partielle de la série différenciée. Pour chacune, on estime à partir de quel rang l'autocorrélation n'est plus significative et ce rang donne une borne sur les valeurs possibles des 2 paramètres associés. On effectue ensuite un grid-search pour sélectionner le modèle qui présente le meilleur score selon le critère d'information AIC. On choisit finalement $(q, Q, p, P) = (1, 1, 1, 1)$.

2.4 Prédiction des marées

Le modèle est entrainé sur les 4 premiers mois de 2023 avec pour objectif de prédire le niveau de la mer en mai. Nous obtenons une MAE de 0.27 :

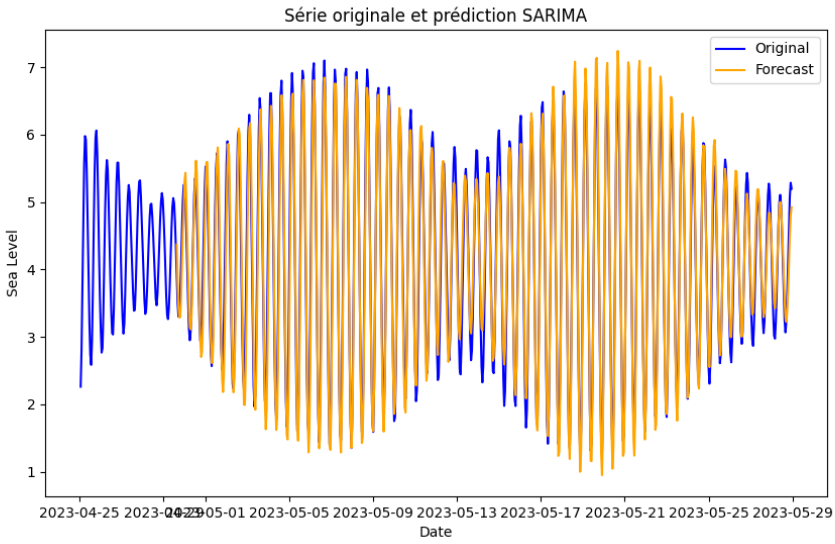

Figure 7: Prédiction du niveau de la mer

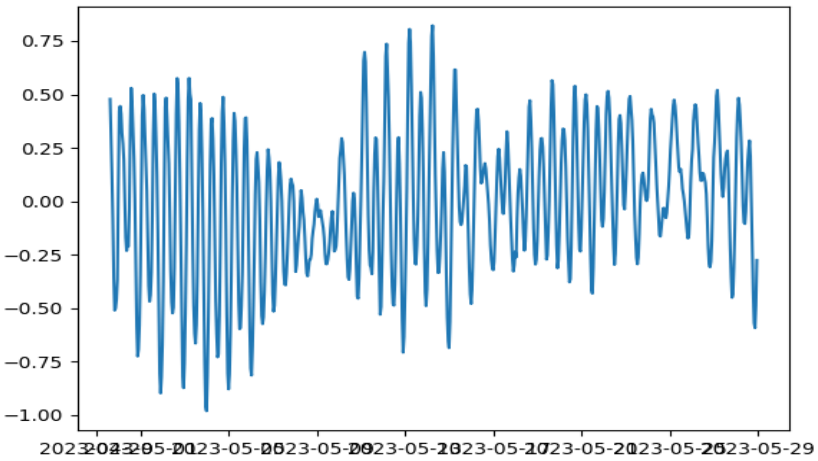

Figure 8: Terme résiduel sur mai

Le modèle est proche de la courbe réelle du niveau de la mer. En revanche, le terme résiduel ne semble pas être un bruit blanc. Il présente une pseudo-période

qui correspond à la période de 12.414h des marées. Le modèle n'est donc pas entièrement satisfaisant, même si la MAE est assez faible, environ 5-10% du marnage. On peut supposer que cette pseudo-périodicité vient de la grande régularité du processus initial. En effet, ce dernier est peu perturbé par rapport à un signal construit uniquement avec des sinusoïdes : une légère erreur dans l'intensité de la prédiction conduit à un résidu relativement régulier. Un signal d'entrée davantage bruité/plus irrégulier présenterait probablement un terme résiduel plus proche d'un bruit blanc.

3 Morts de la pneumonie et du COVID-19 aux Etats-Unis depuis avril 2022

3.1 Présentation du dataset

Ce dataset provient du CDC, une agence de santé américaine. Il recense en particulier les statistiques du nombre de morts du COVID-19, de la pneumonie et de la grippe par état et par semaine. Nous nous intéressons plus spécifiquement au lien entre le nombre de morts du COVID-19 et de la pneumonie à l'échelle nationale depuis avril 2022 (après la fin du pic de l'épidémie de COVID). Nous allons tenter de caractériser la dépendance entre ces deux variables à l'aide d'une copule.

3.2 Analyse

Nous commençons par afficher les courbes correspondant aux deux maladies :

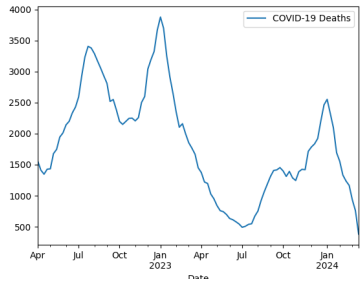

Figure 9: Morts du COVID en fonction du temps

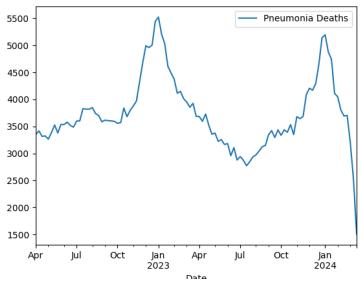

Figure 10: Morts de la pneumonie en fonction du temps

Visuellement, elles semblent avoir une corrélation relativement importante, avec des évolutions qui sont assez similaires (sauf le premier pic du COVID). Nous calculons donc la corrélation linéaire entre les deux variables : elle vaut 0.75, ce qui confirme la corrélation positive observée. Le τ de Kendall vaut 0.59 et la corrélation de Spearman, plus robuste que celle de Pearson, est égale à

0.77. Le nombre de morts du COVID-19 et celui de la pneumonie aux Etats-Unis ne sont donc pas indépendants : les copules vont nous permettre d'estimer la dépendance entre ces deux variables. Pour cela, nous avons utilisé la librairie **copulas**.

L'utilisation de copules permet de séparer les comportements marginaux des dépendances entre variables. Ainsi, avant de les estimer, il faut d'abord appliquer à chaque variable sa fonction de répartition empirique pour se ramener à une loi uniforme sur $[0, 1]$. Si on note $(X_t)_{1 \leq t \leq T}$ la série associée au COVID, on prend $U_t = \frac{rang(X_t)}{T}$, ce que l'on fait à l'aide de la fonction **stats.rankdata**. On obtient le graphique suivant avec la variable corrigée du COVID-19 en abscisse et celle de la pneumonie en ordonnée :

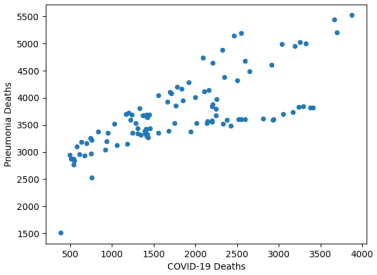

Figure 11: Variables non corrigées

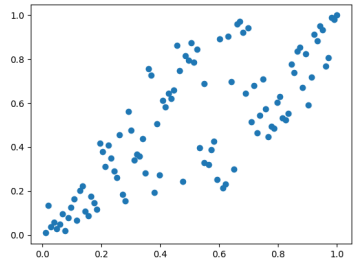

Figure 12: Variables corrigées

A présent, on peut estimer la dépendance avec des copules. Nous avons retenu les trois vues lors du cours :

- **Gaussienne :** $C_P(u) = \Phi_P(\Phi(u_1), ..., \Phi(u_d))$ où Φ est la fonction de répartion de $\mathcal{N}(0, 1)$ et Φ_P celle de $\mathcal{N}(0, P)$
- **Gumbel :** $C_\theta(u_1, u_2) = \exp(-((-\log(u_1))^\theta + (-\log(u_2))^\theta)^{\frac{1}{\theta}})$ avec $1 \leq \theta$
- **Clayton :** $C_\theta(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}}$ avec $0 < \theta$

Après estimation, nous obtenons les paramètres suivants :

- **Gaussienne :** $P = \begin{pmatrix} 1 & 0.86 \\ 0.86 & 1 \end{pmatrix}$
- **Gumbel :** $\theta = 2.43$
- **Clayton :** $\theta = 2.87$

Nous pouvons générer des données synthétiques à l'aide des copules trouvées ci-dessus, et les comparer visuellement à notre échantillon de données pour voir si la dépendance semble effectivement être similaire.

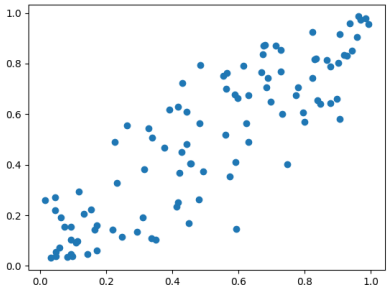

Figure 13: Copule Gaussienne

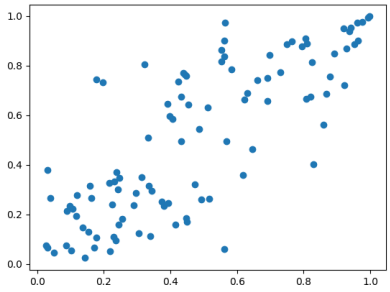

Figure 14: Copule de Gumbel

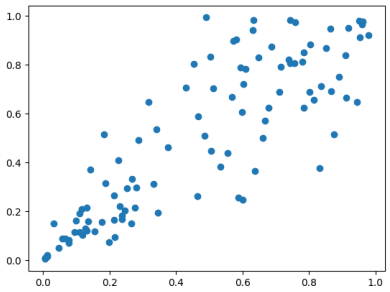

Figure 15: Copule de Clayton

On remarque que la copule de Clayton génère les données les plus proches de notre échantillon : c'est elle qui semble le mieux modéliser la dépendance entre le nombre de morts du COVID-19 et la pneumonie.

4 Retards d'avions américains en janvier 2023

4.1 Présentation du dataset

Ce dataset provient du Bureau of Transportation Statistics, un organisme sous tutelle du Département des Transports des États-Unis. Il contient les informations principales sur chacun des vols de janvier 2023 reliant deux villes américaines (villes de départ et d'arrivée, heures de départ et d'arrivée prévues et réelles) et nous indique également si le vol a été annulé ou détourné. Dans la suite, nous nous restreignons aux vols non détournés, non annulés, à retard positif ou nul.

4.2 Modélisation des retards d'avions

Initialement, nous avons tenté de modéliser la distribution des retards d'avions avec une loi exponentielle puis avec une loi gamma, mais les résultats furent insatisfaisants.

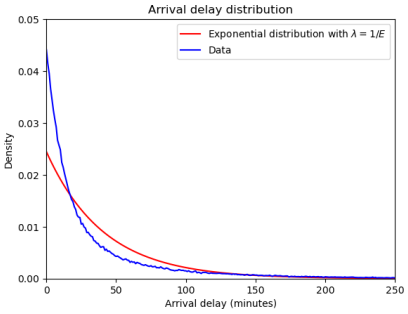

Figure 16: Distributions : dataset et loi Exponentielle

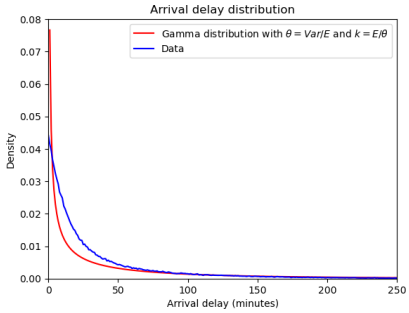

Figure 17: Distributions : dataset et loi Gamma

Finalement, nous avons réussi à modéliser la loi des retards d'avions comme une loi lognormale. Nous avons retenu ici les estimateurs des moments

$$\begin{cases} \hat{\sigma} = \sqrt{\log(1 + Var_{emp}/\mathbb{E}_{emp}^2)} \\ \hat{\mu} = \log(\mathbb{E}_{emp}) - \hat{\sigma}^2/2 \end{cases}$$

bien plus précis que les estimateurs de maximum de vraissemblance ($\hat{\sigma} = Var_{emp}(\log(X))$ et $\hat{\mu} = \mathbb{E}_{emp}(\log(X))$). L'approximation n'est pas tout à fait fidèle pour les très petites (≤ 5) valeurs de X mais épouse très bien la courbe sur tout le reste du support.

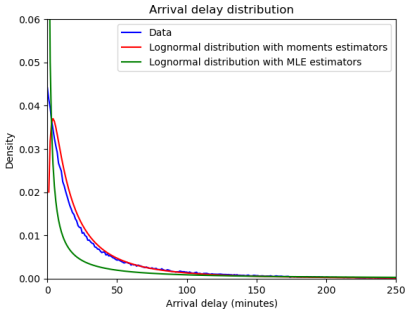

Figure 18: Distributions : dataset et lois Log-normales

4.3 Modélisation des lois max indépendantes

Pour poursuivre notre analyse, nous avons mélangé le dataset et l'avons réparti en 400 listes de longueurs égales afin de modéliser 400 lois max indépendantes. Ensuite, nous avons normalisé chaque loi max avec les paramètres c_n et d_n associés aux lois Log-normales (lois dans le domaine d'attraction de Gumbel) :

$$\begin{cases} d_n = \exp[\mu + \sigma(\sqrt{2\ln(n)} - \frac{\ln(4\pi)+\ln\ln n}{2\sqrt{2\ln(n)}})] \\ c_n = \frac{\sigma d_n}{\sqrt{2\ln(n)}} \end{cases}$$

de sorte que $\frac{M_n - d_n}{c_n}$ converge en loi vers la loi de $Gumbel(0,1)$.

4.4 Comparaison avec la loi limite de Gumbel

Pour comparer $\frac{M_n - d_n}{c_n}$ et sa loi limite de Gumbel, nous nous sommes appuyés sur des comparaisons des distributions et des fonctions de répartition.

D'abord nous avons comparé la distributions empirique de $\frac{M_n - d_n}{c_n}$ en $n = taille\_du\_dataset/400 = 529$ à la distribution de la loi de Gumbel.

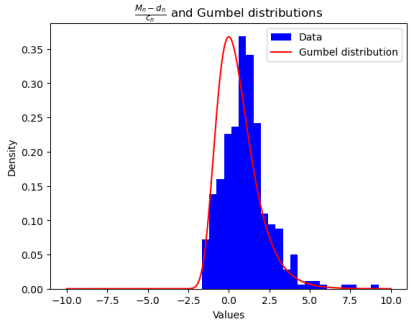

Figure 19: Distributions : $\frac{M_n - d_n}{c_n}$ en $n = 529$ pour 400 échantillons et Gumbel(0,1)

Comme le montre la figure précédente, la distribution empirique présente un léger décalage vers la droite par rapport à la distribution de Gumbel, ainsi que quelques valeurs prises dans la queue de distribution. Cependant, la forme et la valeur max des distributions sont semblables, ce qui justifie la modélisation.

Ensuite, nous avons examiné l'évolution de la fonction de répartition empirique $F(\frac{M_n - d_n}{c_n} < k)$ pour 9 valeurs de k réparties uniformément sur le support de la distribution empirique, en comparant ces valeurs avec les valeurs limite pour la loi de Gumbel.

Nous observons ainsi que le calcul des fonctions de répartitions empiriques est certes imparfait, mais propose tout de même une approximation raisonnable avec une erreur relative inférieure à 30% (excepté pour $k = -1.65$ pour lequel

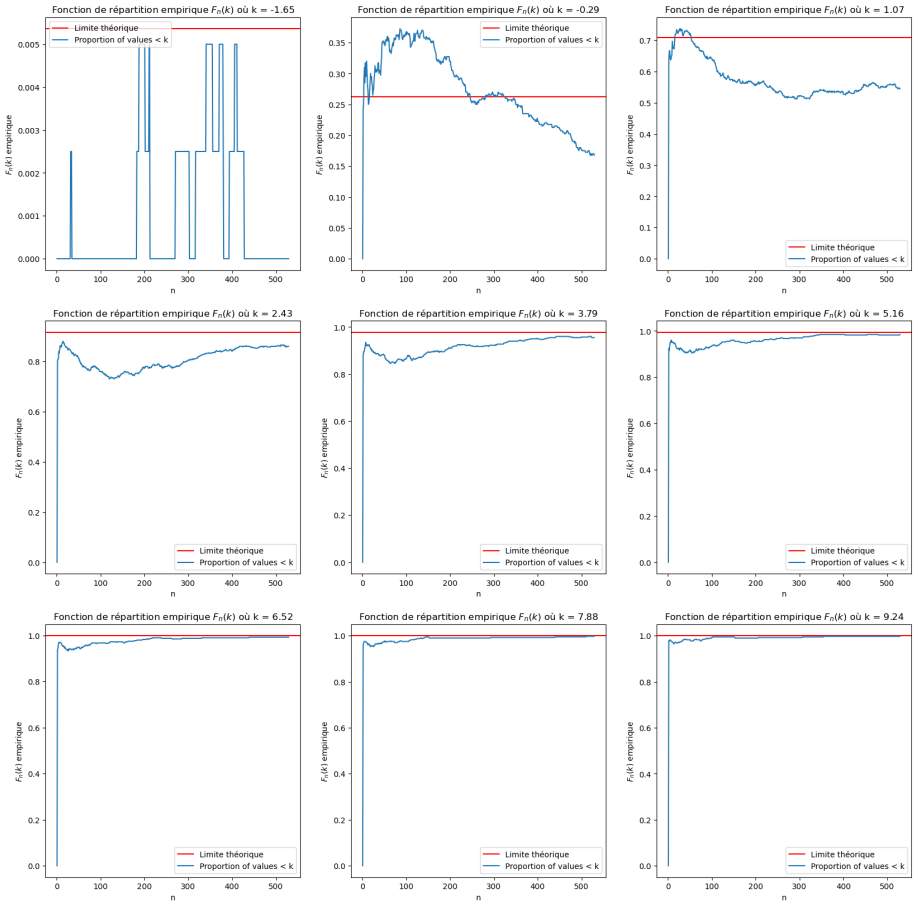

Figure 20: Distributions : Fonction de répartition empirique $F_n(k)$ pour 9 valeurs de k

les valeurs très faibles considérées impliquent une oscillation de la fonction empirique).

5 Activité sismique en Europe en 2024

5.1 Présentation du dataset

Ce dataset provient de l'agence U.S. Geological Survey. Nous avons choisi de nous intéresser aux tremblements de terre récents ayant eu lieu en Europe (de magnitude supérieure à 2.5) et d'essayer de modéliser cette activité sismique. Les séismes en induisent souvent d'autres : ce sont les répliques. Cela leur confère une nature auto-excitante qui nous a motivés à tenter de modéliser les

données à l'aide d'un processus de Hawkes.

5.2 Analyse de la distribution des événements

Les figures suivantes présentent l'évolution, en fonction du temps, du nombre d'événements sismiques et de leur intensité (magnitude). La figure 23 présente la répartition géographique des événements sismiques, donnée d'importance si l'on veut modéliser les différentes corrélations entre les événements sismiques.

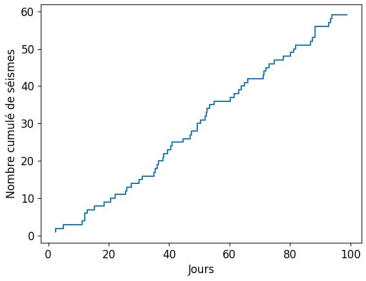

Figure 21: Evolution temporelle du nombre cumulé d'événements sismiques

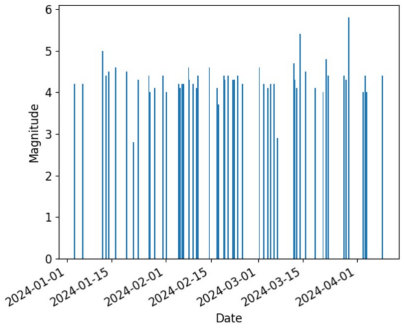

Figure 22: Evolution temporelle de la magnitude des séismes

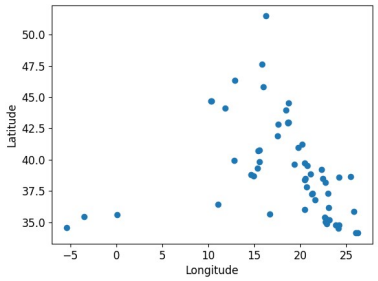

Figure 23: Répartition géographique des séismes

5.3 Estimation des paramètres du processus de Hawkes

Nous avons retenu un modèle à noyau exponentiel. Afin d'estimer les paramètres de la loi, nous avons utilisé la bibliothèque python **Hawkes**.
Les paramètres à estimer pour obtenir un tel processus de Hawkes sont : μ (background intensity), α (self-excitation) et β (decay), de sorte que :

$$\lambda(t) = \mu + \sum_{t_i<t} \alpha \exp(-\beta(t - t_i))$$

L'estimation nous donne :

$$\begin{cases} \hat{\mu} = 0.509 \\ \hat{\alpha} = 0.175 \\ \hat{\beta} = 0.038 \end{cases}$$

On applique un test de Kolmogorov-Smirnov pour juger de la pertinence du modèle estimé vis à vis de nos données. Nous obtenons une p-valeur supérieure à 0.9 ce qui est extrêmement satisfaisant. En effet, la courbe reste très nettement dans l'intervalle de confiance à 95%.

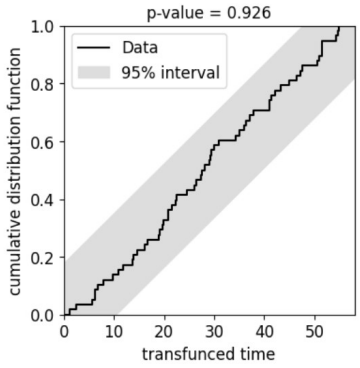

Figure 24: Test de Kolmogorov-Smirnov sur le modèle retenu

5.4 Simulations

Avec l'aide du modèle retenu, nous avons simulé plusieurs scénarios pour les comparer à nos données.

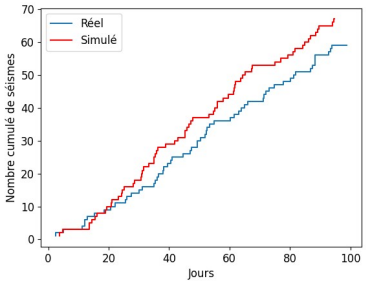

Figure 25: Comparaison entre un scénario simulé et la réalité

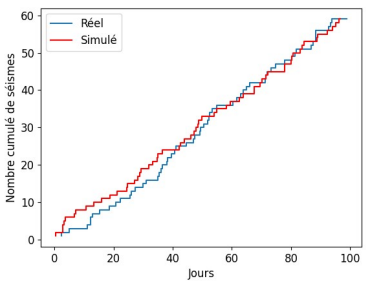

Figure 26: Comparaison entre un scénario simulé et la réalité

Les résultats sont satisfaisants et prédisent des scénarios très proches de la réalité, tant dans leur tendance que dans les valeurs prises pour chaque jour.

En observant ces résultats, il semble donc tout à fait envisageable de tenter de modéliser l'activité sismique de certains territoires par des processus de Hawkes.