Mémoire statistique

Madeleine Hueber et Justine Biscay

Avril 2024

Abstract

Ce mémoire a pour but d'appliquer les outils vu en MAP 565 Modélisation aléatoire et statistique des processus à des problèmes concrets grâce à des jeux de données réels et récents. Voici les jeux de données et les processus statistiques que nous utilisons au sein du mémoire :

1. Modélisation par série temporelle linéaire des températures au Bangladesh sur la période 1901 - 2023 et analyse de la dépendance avec les précipitations mensuelles grâce aux copules

   Jeu de données : `https://www.kaggle.com/datasets/yakinrubaiat/bangladesh-weather-dataset`

2. Modélisation du prix d'ouverture des actions de Tesla grâce à un processus de Garch.

   Jeu de données : `https://finance.yahoo.com/quote/TSLA/history?period1=1609459200&period2=1709251200`

3. Modélisation du nombre hebdomadaire d'hospitalisations dues au Covid-19 aux États-Unis par un processus de Hawkes.

   Jeu de données : `https://covid.cdc.gov/covid-data-tracker/#trends_weeklyhospitaladmissions_select_00`

1 Températures au Bangladesh - séries linéaires temporelles et copules

1.1 Présentation des données

Le jeu de données considéré présente les températures (en degré Celsius) et précipitations (en mm) moyennées par mois à différentes localisations au Bangladesh, sur la période 1901 - 2023.

Tout d'abord, nous allons procéder à une analyse de la **série temporelle des température mensuelles sur la période 1901 - 2023**, en découpant le jeu de données en un jeu de données d'entraînement pour construire le modèle (janvier 1901 à janvier 2023), et un jeu de donnée test pour le valider ou l'infirmer (février à juin 2023).

Puis on s'intéressera à la **corrélation entre les variables températures et précipitations mensuelles grâce aux copules**. On s'attend à une corrélation linéaire positive entre ces deux variables, car au Bangladesh, des températures plus élevées sont généralement corrélées à des précipitations plus abondantes, notamment pendant la mousson. .

1.2 Séries temporelles

On peut d'abord représenter la série temporelle des températures sur l'ensemble de la période 1901 - 2023.

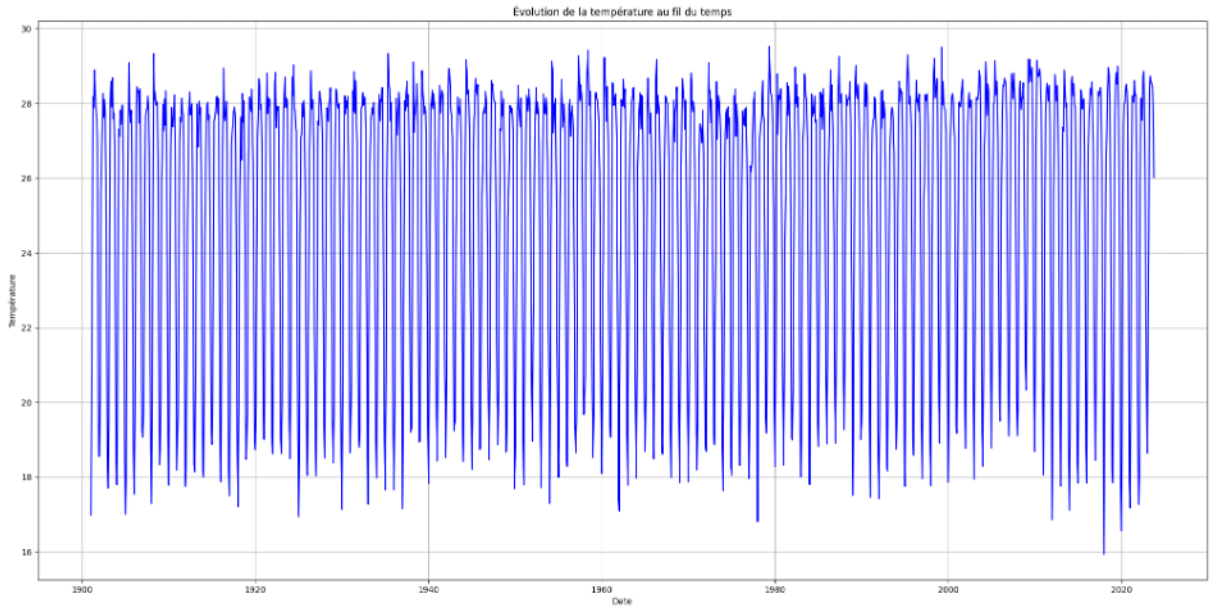


Figure 1: Évolution de la température mensuelle au Bangladesh de 1901 à 2023

On cherche à construire un modèle de prédiction basé sur les séries temporelles linéaires, c'est-à-dire un modèle capablee de prédire les valeurs futures de la température à partir des observations et erreurs passées. Pour cela, nous emploierons la méthode de Box-Jenkins, axée sur l'analyse et l'utilisation de modèles SARIMA (Seasonal AutoRegressive Integrated Moving Average).

Les modèles SARIMA (*Seasonal AutoRegressive Integrated Moving Average*) sont une extension des modèles ARIMA qui ajoutent une structure saisonnière. Un modèle SARIMA est généralement noté SARIMA$(p, d, q)(P, D, Q)_s$ où :

- p, d, q sont les paramètres non saisonniers comme dans ARIMA.
- P, D, Q sont les paramètres saisonniers correspondant aux composants AR saisonnier, intégré saisonnier, et MA saisonnier, respectivement.
- s est la périodicité de la saisonnalité. Par exemple, pour des données mensuelles qui présentent une saisonnalité annuelle, s serait 12.

L'équation est donnée par :

$$\nabla^d \nabla_s^D X_t = c + \sum_{i=1}^{p} \phi_i \nabla^d \nabla_s^D X_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \sum_{i=1}^{P} \Phi_i \nabla^d \nabla_s^D X_{t-is} + \sum_{j=1}^{Q} \Theta_j \varepsilon_{t-js} + \varepsilon_t \tag{1}$$

où :

- ∇^d et ∇_s^D sont les opérateurs de différenciation non saisonnière et saisonnière, respectivement.
- c est une constante (le terme d'interception).
- $\phi_1, \ldots, \phi_p$ et $\theta_1, \ldots, \theta_q$ sont les coefficients des termes AR et MA non saisonniers.
- $\Phi_1, \ldots, \Phi_P$ et $\Theta_1, \ldots, \Theta_Q$ sont les coefficients des termes AR et MA saisonniers.
- ϵ_t est le terme d'erreur au temps t, qui est supposé être un bruit blanc.

Nous allons donc sélectionner de manière optimale les ordres et les coefficients du modèle en suivant la méthode de Box-Jenkins.

1.2.1 Vérification de l'hypothèse de stationnarité de la série

Les résultats du test de Dickey-Fuller après différenciation simple puis différenciation par rapport à la saisonnalité annuelle sont les suivants :

Statistique de test	-12.088950455342887
P-value	$2.144782626370246 \times 10^{-22}$
Valeurs Critiques 1%	-3.4349408214067227
Valeurs Critiques 5%	-2.8635675309927153
Valeurs Critiques 10%	-2.5678494453155656

Table 1: Résultats du test statistique de Dickey-Fuller

On fixe ainsi dans le modèle SARIMA les paramètres $d = 1$ et $D = 1$ car on a différencié par rapport à la période annuelle ($S = 12$). La série temporelle stationnaire obtenue après ces différentiations est représentée ci-dessous.

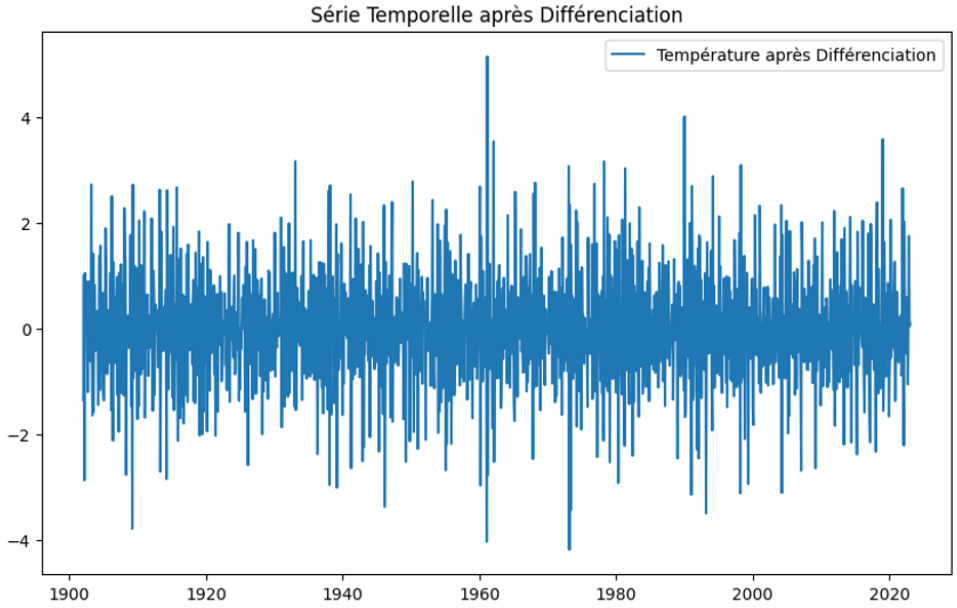


Figure 2: Série temporelle stationnaire obtenue

1.2.2 Estimation des ordres de l'équation SARIMA et des paramètres du modèle

Pour déterminer les ordres les plus appropriés du modèle, on commence par examiner les graphiques d'autocorrélation (ACF) et d'autocorrélation partielle (PACF) des données.

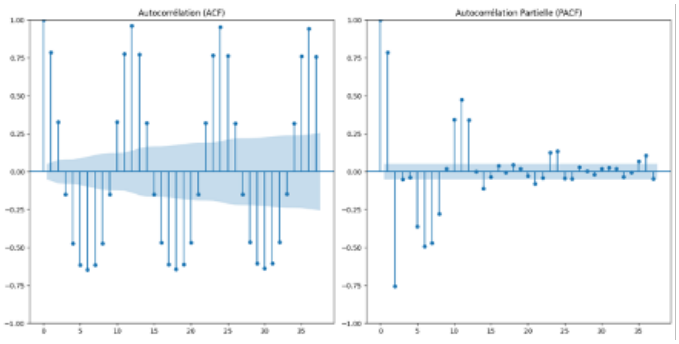


Figure 3: Diagrammes d'autocorrélation et d'autocorrélation partielle

Ces diagrammes présentent des corrélations fortes et supérieures aux seuils sur de nombreux lags et notamment sur des lags réguliers (12), ce qui confirme la nécessité d'utiliser un modèle SARIMA qui va prendre en compte cette saisonnalité.

En utilisant la bibliothèque `pmdarima` de Python, il est possible d'explorer une gamme de configurations de modèles $(p, d, q)(P, D, Q)$ et de les comparer en fonction du critère d'information bayésien (BIC). Le BIC évalue la pertinence du modèle en considérant la vraisemblance, mais avec une pénalité plus forte pour les modèles ayant un excès de paramètres. Cette technique vise à sélectionner les ordres de modèle les plus adéquats.

En appliquant cette approche, le modèle SARIMA $(3, 1, 0)(2, 1, 0)_7$ a été identifié comme ayant le BIC le plus bas, suggérant qu'il offre un bon équilibre entre l'ajustement du modèle et sa simplicité. Ce modèle est alors ajusté aux données pour en estimer les paramètres par la méthode de maximisation de la vraisemblance, ce qui permet d'obtenir les coefficients ϕ, θ, ϕ_{maj} et θ_{maj}.

1.2.3 Vérification du modèle

On réalise réalise plusieurs tests pour vérifier le modèle.

On vérifie avec le test de Ljung-Box si les résidus sont indépendants. Cependant, cette condition n'est pas vérifiée avec le modèle SARIMA considéré (le test donne une p-valeur < 0.05). Les résidus ont une autocorrélation significative qui n'a pas été capturée par le modèle, comme on peut l'observer sur les diagrammes d'ACF et de PACF ci-dessous.

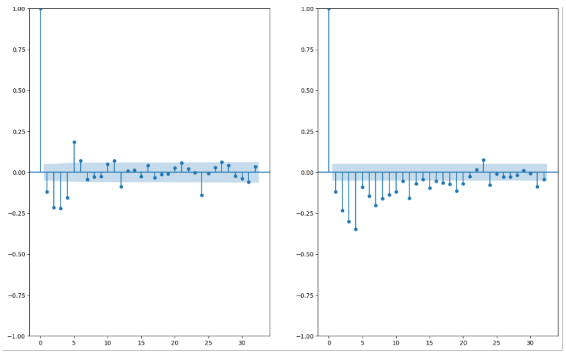


Figure 4: Diagrammes d'autocorrélation et d'autocorrélation partielle

On vérifie avec le test de Jarque-Bera si les résidus suivent une loi gaussienne centrée. Cependant, cette condition n'est pas vérifiée avec le modèle SARIMA considéré (le test donne une p-valeur < 0.05).

On vérifie si les résidus présentent une valeur constante. Cette condition est vérifiée. Cependant, cette condition n'est pas vérifiée avec le modèle SARIMA considéré (le test donne une p-valeur de $0.16 > 0.05$).

Le premier modèle considéré ne satisfait pas deux des conditions nécessaire. On opère alors les 3 tests sur le modèle avec le deuxième meilleur BIC.

1.2.4 Prédiction avec le modèle

On peut utiliser le modèle construit pour, même si on s'attend à des résultats peu probants au regard des hypothèses non vérifiée.

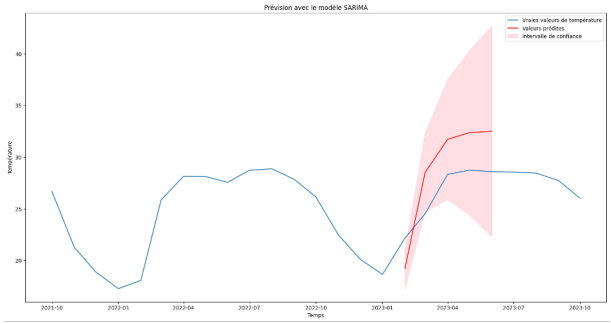


Figure 5: Prédiction avec le modèle SARIMA

1.3 Copules

On commence par s'intéresser à la corrélation entre températures et précipitations mensuelles. On peut d'abord représenter l'évolution mensuelle de ces variables sur la période 2018 - 2023.

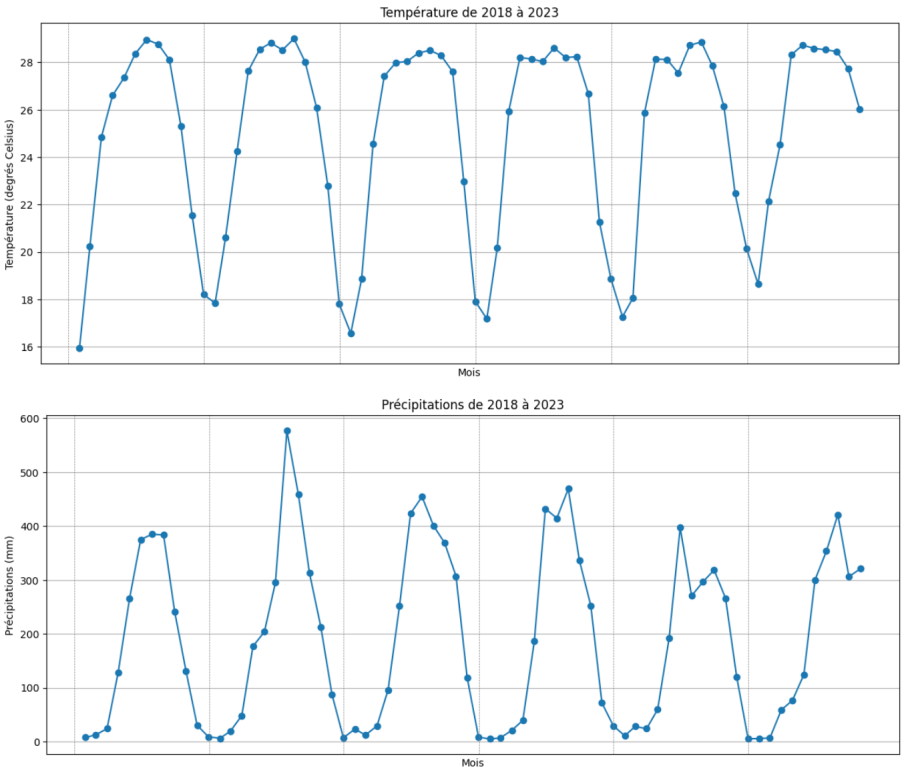


Figure 6: Évolution de la température et des précipitations mensuelles au Bangladesh de 2018 à 2023

On peut noter X la variable température et Y la variable précipitations, et on peut extraire la moyenne et l'écart-type de leur distribution comme présenté en figure 2.

On observe par ailleurs un coefficient de corrélation linéaire de $\rho = 0.76$, ce qui montre une corrélation linéaire positive, comme attendu (mais pas parfaite car l'écart à 1 reste important).

	X	Y
Moyenne (m)	25.02	187.37
Écart type (σ)	4.13	162.56

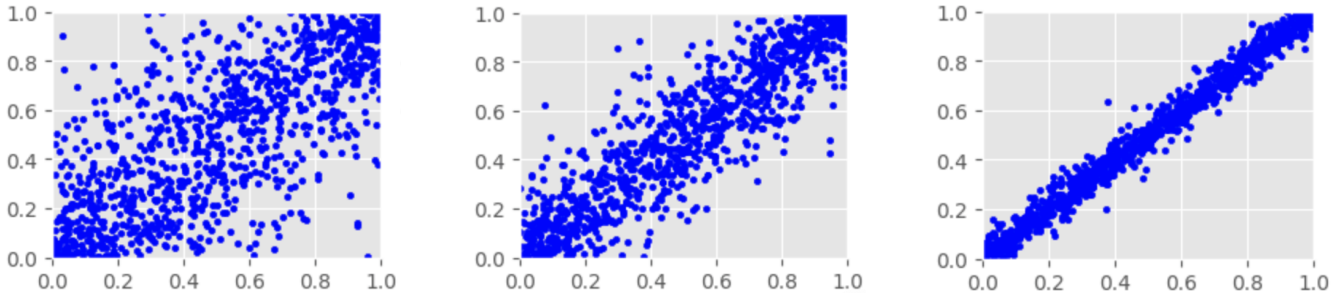
Table 2: Statistiques des températures et des précipitations

1.3.1 Analyse de la dépendance des variables avec les copules

On utilise les copules pour modéliser la dépendance entre les deux variables aléatoires considérées, en séparant la structure de dépendance de leurs distributions marginales. On considère trois types de copules étudiés :

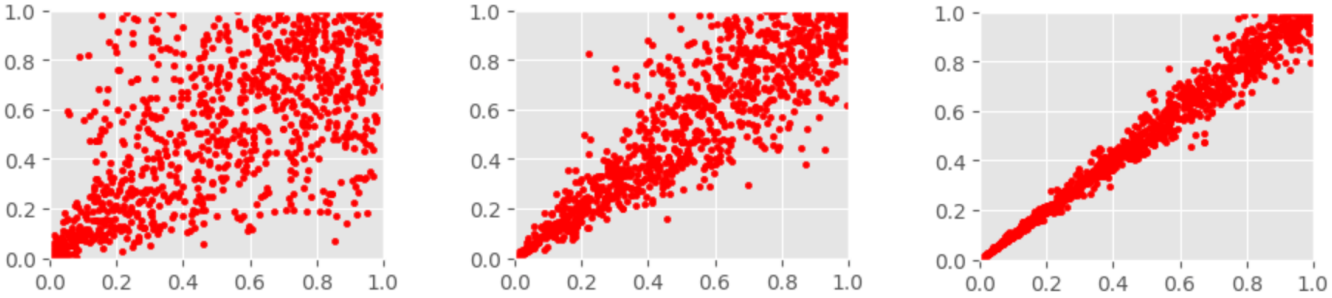
- **Copule de Clayton :** $C(x, y) = (x^{-\theta} + y^{-\theta} - 1)^{-1/\theta}$
- **Copule de Frank :** $C_\theta(x, y) = \frac{1}{\theta} \ln\left(1 + \frac{(e^{-\theta x}-1)(e^{-\theta y}-1)}{e^{-\theta}-1}\right)$
- **Copule de Gumbel :** $C(x, y) = \exp\left(-\left((-\log(x))^{-\theta} + (-\log(y))^{-\theta}\right)^{1/\theta}\right)$

On peut représenter les différentes distributions obtenues à l'aide de ces copules en considérant différents paramètres θ :



(a) $\theta = 5$ (b) $\theta = 12$ (c) $\theta = 43$

Figure 7: Copule de Franck

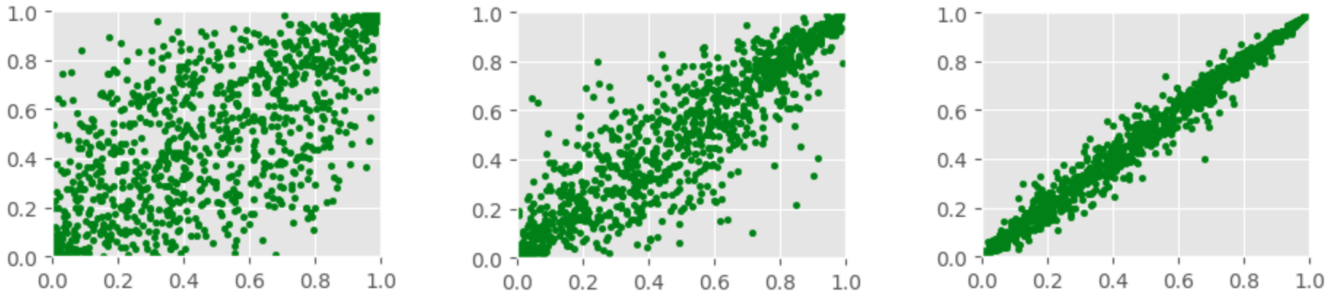


(a) $\theta = 2$ (b) $\theta = 5$ (c) $\theta = 20$

Figure 8: Copule de Clayton

Pour utiliser les variables aléatoires considérées, on doit d'abord les ramener à une répartition dans l'intervalle $[0; 1]$ selon la méthode suivante :

1. On trie dans l'ordre croissant les X_1^t et Y_1^t avec t entre 0 et n.
2. On note le rang $r(X, t)$ et $r(Y, t)$ de chacun de ces X_1^t et Y_1^t.
3. On définit $X_2^t = \frac{r(X,t)}{n}$ et $Y_2^t = \frac{r(Y,t)}{n}$.



(a) $\theta = 2$ (b) $\theta = 4$ (c) $\theta = 11$

Figure 9: Copule de Gumbel

On peut représenter ensuite la distribution obtenue :

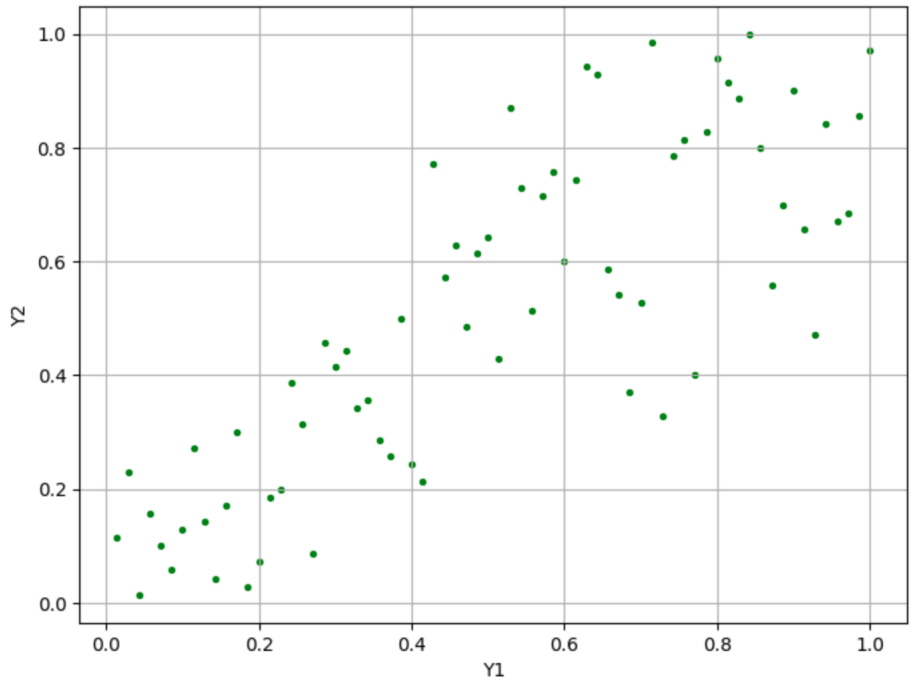


Figure 10: Représentation de (X_2^t, Y_2^t)

En comparant avec les distributions des copules ci-dessus, on s'attend à ce que le **choix de Gumbel soit optimal**, ce qui est cohérent avec le fait que la copule de Gumbel modélise au mieux des scénarios où les événements extrêmes positifs dans une variable sont accompagnés d'événements extrêmes positifs dans une autre. La copule de Gumbel apparaît donc adaptée à notre cas car les valeurs extrêmes de températures sont positivement liées à de fortes précipitations au Bangladesh.

Pour maintenant évaluer précisemment quel copule et quel paramètre offrent la meilleure simulation de la distribution de (X_2, Y_2), **nous utilisons le test d'ajustement du** χ^2. L'approche consiste à découper le pavé $[0,1] \times [0,1]$ en k domaines et à comparer les fréquences théoriques et observées de (X_2, Y_2).

Nous divisons le pavé $[0,1] \times [0,1]$ en $k = 8$ domaines, pour ne pas avoir de domaines vides. La matrice d'observation est obtenue en comptant les occurrences de (X_2, Y_2) dans chacun de ces domaines.

Pour chaque copule et chaque paramètre, nous calculons la matrice théorique $E = nF$, où F est la matrice des fréquences théoriques et n est le nombre d'observations. Ensuite, nous utilisons la formule du χ^2 pour évaluer l'adéquation entre les observations réelles et théoriques.

On obtient les résultats suivants qui indiquent le meilleur paramètres θ à choisir pour chacun des copules :

Copule	θ	χ^2
Frank	10.8	0.33
Clayton	5.4	0.9
Gumbel	3.9	0.23

Table 3: Comparaison des copules avec leurs paramètres θ et χ^2 correspondants

Selon le critère de minimisation du χ^2, on choisit la copule de Gumbel avec $\theta = 3.9$.

On a considéré ici 7 degrés de liberté, donc la table de loi du χ^2 nous permet d'affirmer que la valeur du $\chi^2 = 0.4$ obtenue ne rejette pas l'hypothèse selon laquelle la loi jointe de (X_2, Y_2) a une fonction de répartition donnée par la copule de Gumbel avec $\theta = 3.9$.

2 Processus de GARCH

2.1 Présentation des données

Le jeu de données utilisé pour cette analyse concerne les prix d'ouverture quotidiens des actions de Tesla en bourse, couvrant la période allant du 1er janvier 2021 au 1er mars 2024. Les données sont collectées sur une base journalière, à l'exception des week-ends et des jours fériés où les marchés financiers sont fermés. Le jeu de données initial est sous forme de tableau, chaque ligne représentant une observation quotidienne comprenant la date, le prix d'ouverture, de clôture, le maximum et le minimum atteint ce jour-là, ainsi que le volume des échanges. Pour notre analyse, nous nous intéressons uniquement au prix d'ouverture journalier.

	Date	Open	High	Low	Close	Adj Close	Volume
Date							
2021-01-04	2021-01-04	239.820007	248.163330	239.063339	243.256668	243.256668	145914600
2021-01-05	2021-01-05	241.220001	246.946671	239.733337	245.036667	245.036667	96735600
2021-01-06	2021-01-06	252.830002	258.000000	249.699997	251.993332	251.993332	134100000
2021-01-07	2021-01-07	259.209991	272.329987	258.399994	272.013336	272.013336	154496700
2021-01-08	2021-01-08	285.333344	294.829987	279.463318	293.339996	293.339996	225166500

Figure 11: Extrait du jeu de données

Figure 12: Prix d'ouverture journalier des actions de TESLA

Ces données étant journalières il est intéressant de regarder le au rendement journalier des actions défini par $\epsilon_t = log(\frac{p_t}{p_{t-1}})$ avec p_t le prix d'ouverture à la date t.

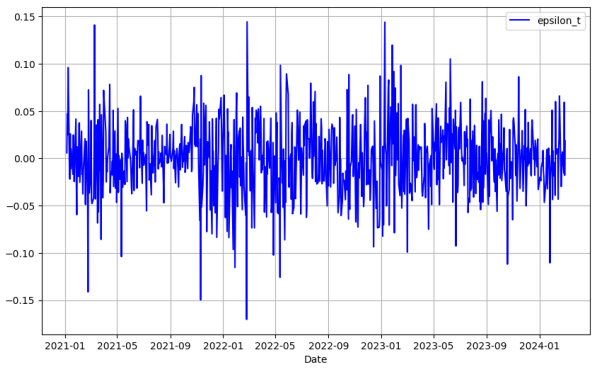


Figure 13: Tracé de ϵ_t

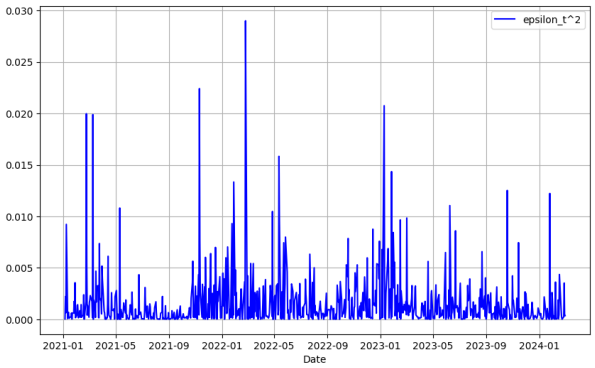


Figure 14: Tracé de ϵ_t^2

Il semble que ϵ_t^2 soit auto-corrélé : de grandes valeurs de ϵ_t^2 sont suivies par de grandes valeurs de ϵ_t^2, et vice versa. De plus, ϵ_t ne semble pas suivre une loi normale, comme le montre le "Quantile-Quantile plot" ci-dessous et le calcul de la kurtosis : $\frac{E[(\epsilon_t - E[\epsilon_t])^4]]}{Var(\epsilon_t)^2} = 4.28$ (> 3 pour une loi normale).

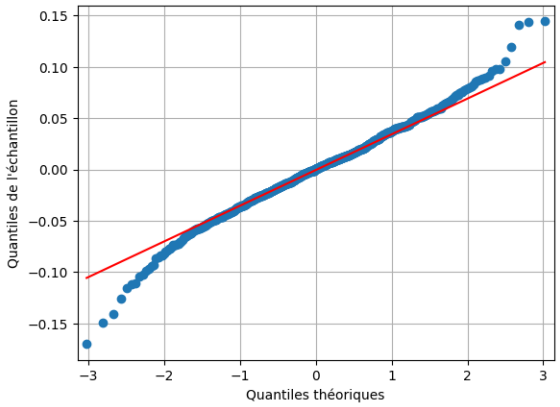


Figure 15: Normal Q-Q plot

Ces éléments nous montrent qu'il ne s'agit pas ici d'un cas de volatilité constante ce qui nous pousse à modéliser $\epsilon(t)$ par un processus de GARCH avec volatilité aléatoire.

2.2 Modélisation

Nous souhaitons donc modéliser les rendements journaliers ϵ_t par un processus de GARCH définit comme suit :

$$\epsilon_t = \sigma_t \eta_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

$$\omega \leq 0,\ \alpha_i, \beta_i \leq 0\ \forall i \text{ et } \eta_t \sim N(0,1)$$

Pour modéliser ce processus on utilise le package `arch`. On choisit tout d'abord de modéliser un GARCH(1,1) en calibrant notre modèle sur nos données jusqu'en janvier 2023 on obtient les valeurs suivantes pour les paramètres :

$$\omega = 0.17\ \alpha_1 = 0.08\ \beta_1 = 0.92$$

Cela nous permets ensuite de réaliser des simulations pour prédire ϵ_t entre janvier 2023 et mars 2024. En se ramenant au prix d'ouverture quotidien on obtient le graphique suivant :

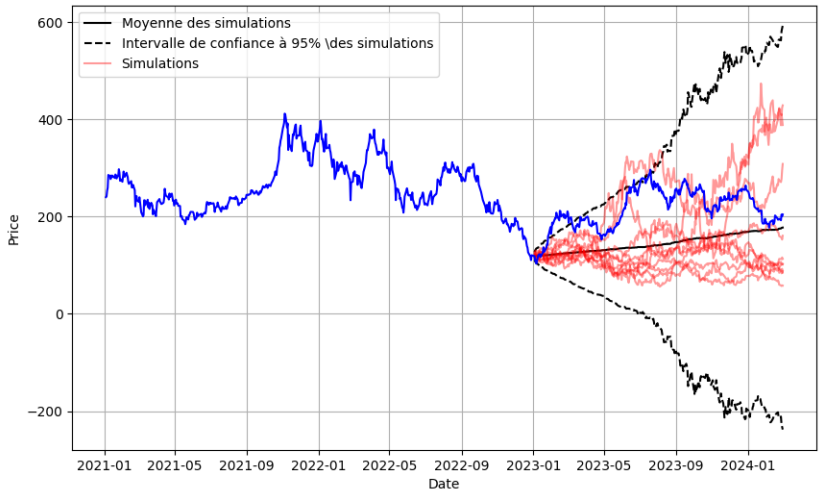


Figure 16: Tracé des données réelles et simulées pour le prix d'ouverture journalier des actions de TESLA

Ici, seules 10 simulations sont présentées en rouge, mais la moyenne des simulations et l'intervalle de confiance à 95% sont basés sur 1000 simulations. La moyenne et l'intervalle de confiance permettent ainsi de prédire les scénarios futurs.

2.3 Remarques et critiques

Sur la figure précédente, nous remarquons que si l'intervalle de confiance à 95% contient bien la courbe réelle, il a tendance à s'agrandir avec le temps et à s'éloigner des valeurs réelles. Notre modèle semble donc plus efficace pour des horizons à court terme.

De plus nous nous sommes demandés si le modèle Garch(1,1) n'était pas limité et si il ne faudrait pas augmenter nos valeurs de p et q. Nous avons donc réalisé plusieurs tests mais en augmentant p et/ou q nous nous sommes rendues compte que au plus 2 valeurs de α_i n'étaient pas nulles (de même pour β_i).

Nous nous sommes donc intéressées à un modèle GARCH(p=2,q=2). Comme le montre les courbes ci-dessous, le résultat est légèrement amélioré puisque l'intervalle à 95% se trouve plus proche des données réelles. On a donc une meilleure prédiction avec ce modèle.

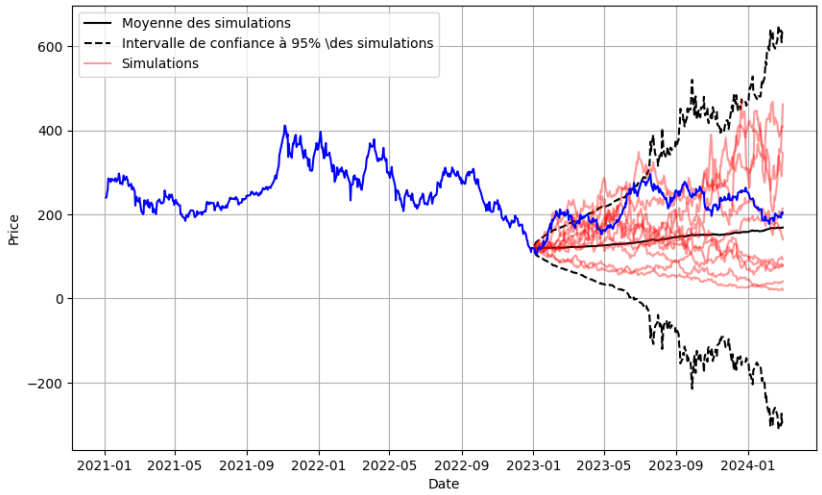


Figure 17: Tracé des données réelles et simulées avec p=q=2

3 Processus de Hawkes

3.1 Présentation des données

Le jeu de données utilisé concerne le nombre hebdomadaire de nouveaux patients admis à l'hôpital pour Covid-19 aux États-Unis entre le 8 août 2020 et le 23 mars 2024. Le jeu de données initial est présenté sous forme de tableau où chaque ligne représente une observation hebdomadaire comprenant la date, le lieu (États-Unis), et le nombre de nouveaux patients admis à l'hôpital pour cette semaine.

Date	Geography	Weekly COVID-19 Hospital Admissions
2020-08-08	United States	31623
2020-08-15	United States	29869
2020-08-22	United States	29424
2020-08-29	United States	28194
2020-09-05	United States	26462

Figure 18: Extrait du jeu de données

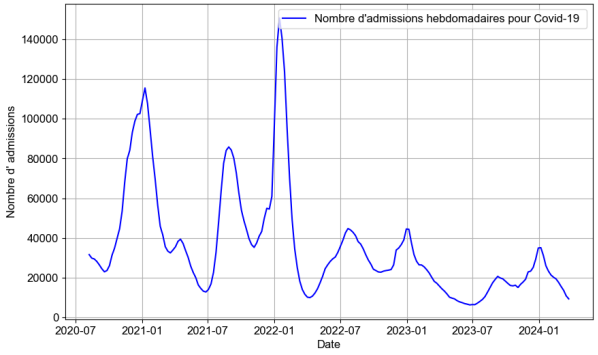


Figure 19: Tracé du nombre hebdomadaire de nouveaux hospitalisés pour Covid-19 aux Etats-Unis

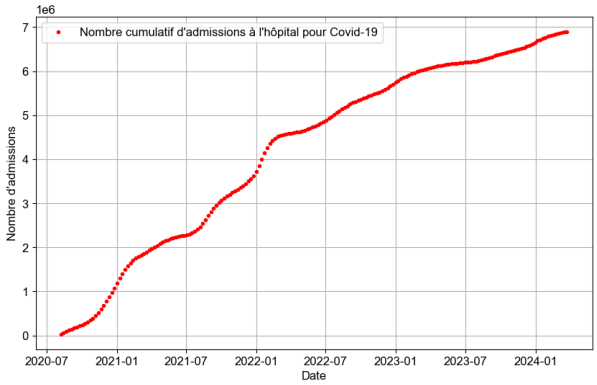


Figure 20: Tracé du nombre cumulatif d'admis

Dans une pandémie, la probabilité d'infection future est influencée par le nombre actuel de cas, en particulier pour des maladies très contagieuses comme le COVID-19. Cette auto-excitation signifie que plus il y a d'hospitalisations pour le COVID-19, plus il est probable que de nouvelles hospitalisations surviennent par la suite. C'est pourquoi nous avons choisi d'utiliser un processus de Hawkes, qui capture cette dynamique d'auto-excitation pour modéliser nos données.

Pour notre étude nous nous intéressons à un processus de Hawkes univarié dans le cas où la fonction d'excitation est définie comme suit : $g(t) = \mu exp(\beta t)$.

Dans ce cas, l'intensité conditionnelle du processus de Hawkes est définie comme suit :

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

Cette intensité représente le nombre attendu d'événements qui se produisent à l'intervalle de temps t, conditionnellement au passé. Ce qui se rapproche de notre donnée de départ : le nombre hebdomadaire d'hospitalisés.

Ou t_i représentent les temps d'hospitalisation. Comme nos données sont fournies en nombre de personnes hospitalisées par semaine, on fait l'hypothèse que ces hospitalisations sont uniformément réparties sur la semaine pour obtenir les temps de chaque hospitalisation. De plus, comme le nombre d'hospitalisés est très important, on divise le nombre hebdomadaire de personnes hospitalisées par un facteur 1000 pour des calculs plus efficaces.

3.2 Modélisation

Pour modéliser notre processus de Hawkes, on utilise le package `Hawkes`. Dans un premier temps, on calibre le modèle sur tous nos temps pour évaluer les résultats.

On obtient alors les valeurs des paramètres suivants :

$$\mu = 1.95 \ \alpha = 0.95 \ \beta = 1.14$$

En traçant $\lambda(t)$ et en comparant avec nos données initiales on retrouve bien les mêmes courbes à un facteur 1000 près.

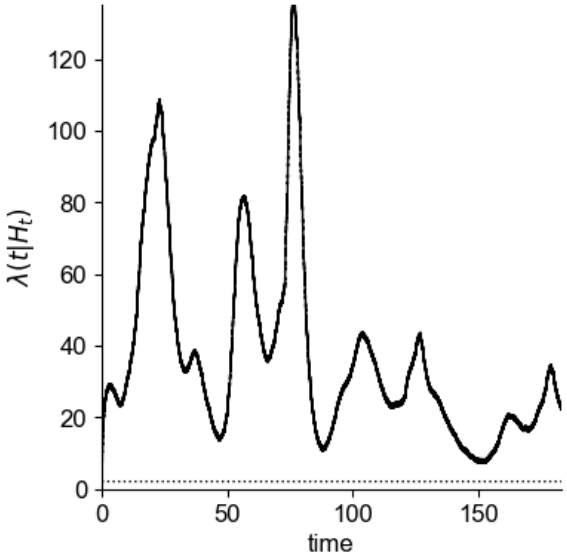


Figure 21: Tracé de $\lambda(t)$ avec notre modèle

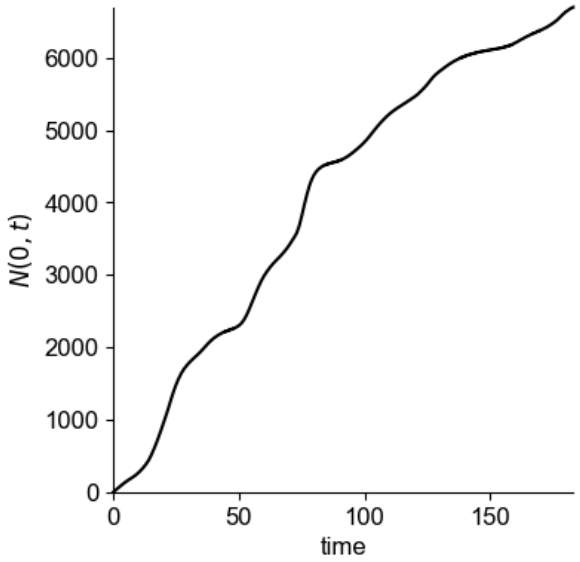


Figure 22: Tracé de $N(t)$ avec notre modèle

On cherche maintenant à voir si le modèle est capable de prédire le future nombre d'hospitalisés pour les semaines à venir. Pour cela on calibre notre modèle sur les 150 premières semaines et on obtient les paramètres suivants :

$$\mu = 2.17 \ \alpha = 0.94 \ \beta = 1.22$$

On réalise ensuite 100 simulations sur les 40 semaines suivantes et on obtient le graph ci-dessous, où la courbe bleue représente le nombre cumulatif réel d'hospitalisés et les courbes rouges représentent les simulations de notre modèle. Ces simulations fournissent ainsi un intervalle de confiance pour le futur nombre d'hospitalisés. On remarque que les valeurs réelles se trouvent bien dans cet intervalle de confiance.

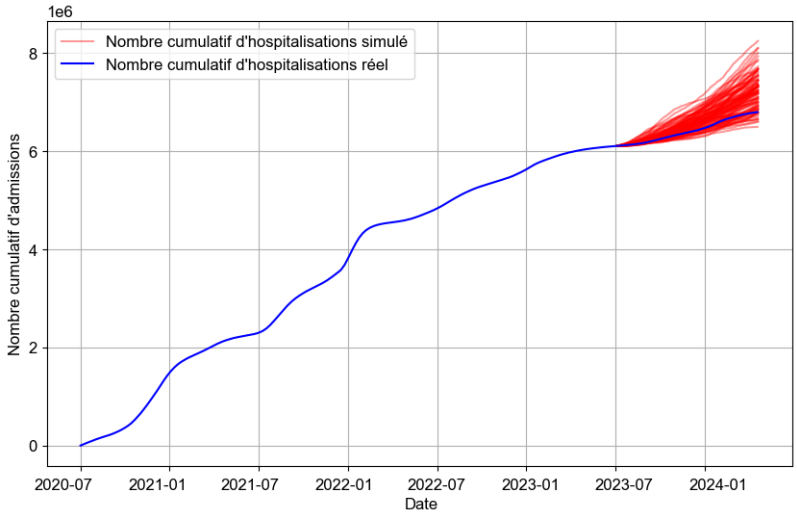


Figure 23: Tracé des simulations

3.3 Remarques et critiques

Le modèle permet ainsi de fournir une première approximation du nombre d'hospitalisations pour les prochaines semaines. Cependant, dans le cas précédent, on constate que l'intervalle de confiance devient très large lorsqu'on tente de prédire sur un horizon trop lointain. En effet, le modèle ne fonctionne plus bien pour des horizons trop étendus.

Si on calibre notre modèle sur les 100 premières semaines d'une part et les 60 première de l'autre on obtient les courbes suivantes :

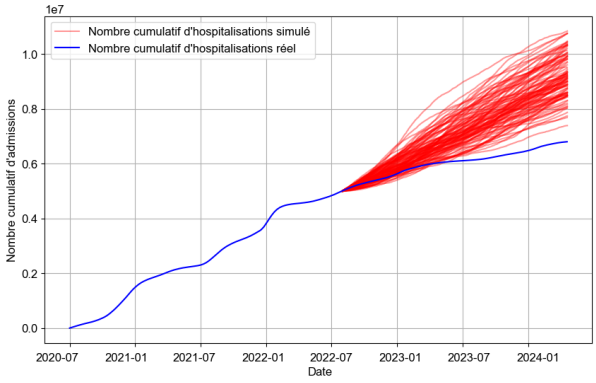


Figure 24: Simulation avec une prédiction sur 90 semaines

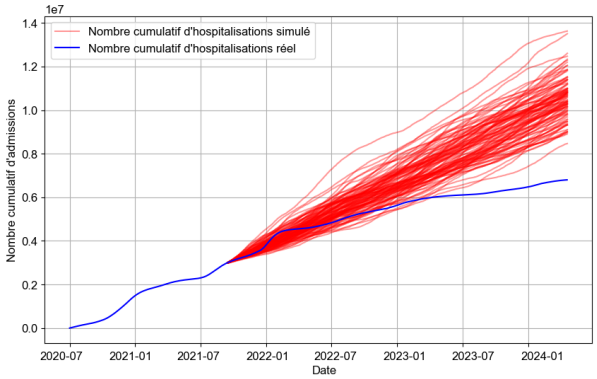


Figure 25: Simulation avec une prédiction sur 130 semaines

On voit que plus on cherche à prédire longtemps en avance plus le modèle a tendance à surévaluer le nombre d'hospitalisés.

Cela est probablement dû au fait que ce modèle univarié est une représentation simplifiée des dynamiques complexes impliquées dans la propagation du Covid-19. La dynamique de la pandémie est influencée par divers facteurs externes et évolutifs, tels que les mesures de santé publique, les comportements individuels, les variations de la virulence du virus et l'évolution des variants, ainsi que d'autres variables socio-économiques et environnementales.

Par conséquent, lorsque nous tentons de prédire à long terme, le modèle Hawkes peut ne pas saisir pleinement ces facteurs externes en constante évolution, ce qui conduit à une surestimation du nombre d'hospitalisations.