Mémoire de MAP566 - Modélisation aléatoire et Statistiques des processus

Marc BOËLLE et Tess BRETON

Abstract :

Dans ce mémoire, nous menons des études statistiques sur quatre jeux de données différents pour utiliser les notions abordées en MAP565. Le premier jeu de données contient des données sur les cours boursiers d'entreprises du NASDAQ ; nous l'utilisons pour modéliser le cours de l'action Amazon avec des processus de GARCH, puis pour modéliser la dépendance entre les cours des actions Amazon et Google à l'aide de copules. Le deuxième jeu de données contient des données de consommation énergétique en Suisse, que nous analyserons à l'aide de la théorie des séries temporelles. Le troisième jeu de donnés contient des donnés de trafic routier (nombre de véchiules) à Istanbul pour la période de décembre 2023. Nous utilisons un modèle SARIMA pour modéliser leur évolution périodique. Enfin, le dernier jeu de données recense les séismes au Chili sur les dix dernières années ; nous utilisons les processus de Hawkes pour modéliser leurs occurences.

Datasets :

Les quatre jeux de données utilisés contiennent des données de 2023 (même de 2024), et sont accessibles aux liens suivants :

— NASDAQ stock prices
`https://www.kaggle.com/datasets/svaningelgem/nasdaq-daily-stock-prices`

— Consommation énergétique en Suisse
`https://www.swissgrid.ch/en/home/operation/grid-data/generation.html#total-energy-consumpti`

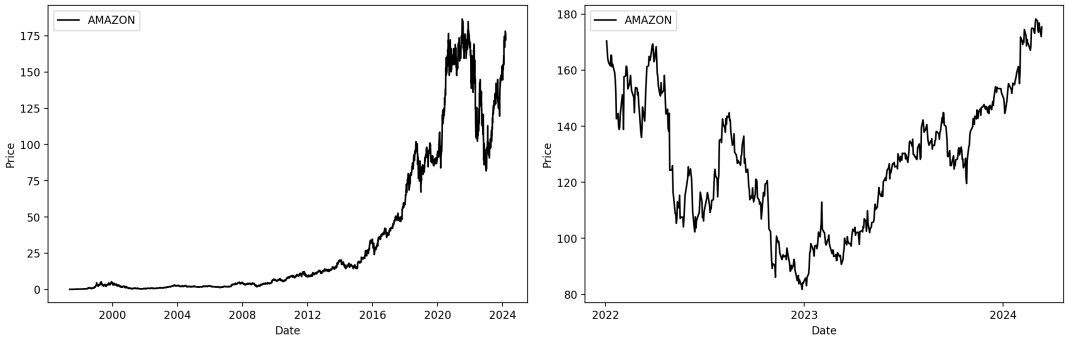
— Traffic routier à Istanbul
`https://data.ibb.gov.tr/en/dataset/hourly-traffic-density-data-set/resource/aa58374d-ef6f-411f-8271-5b63eefe4fde?filters=`

— Séismes au Chili
`https://www.kaggle.com/datasets/nicolasgonzalezmunoz/earthquakes-on-chile`

1 Évolution du cours de l'action Amazon

1.1 Visualisation des données

Notre premier jeu de données contient les cours quotidiens de l'action Amazon à la fermeture de la Bourse, du 15 Mai 1997 jusqu'au 12 Mars 2024 (week-ends et jours féries exclus). Nous visualisons les données sur la Figure 1, et nous choisissons de ne conserver pour notre étude que les données à partir de l'année 2022.



(a) Depuis 1997 (b) Depuis 2022

FIGURE 1 – Évolution du cours de l'action Amazon (AMZN)

Comme souvent lorsqu'on étudie des données financières, nous considérons le rendement journalier ε défini par :

$$\varepsilon_t = \log\left(\frac{P_t}{P_{t-1}}\right) \tag{1}$$

Nous représentons l'évolution de ε et ε^2 sur la Figure 2. Nous remarquons que la série temporelle ε^2 est autocorrélée, ce qui suggère une modélisation par un processus GARCH.

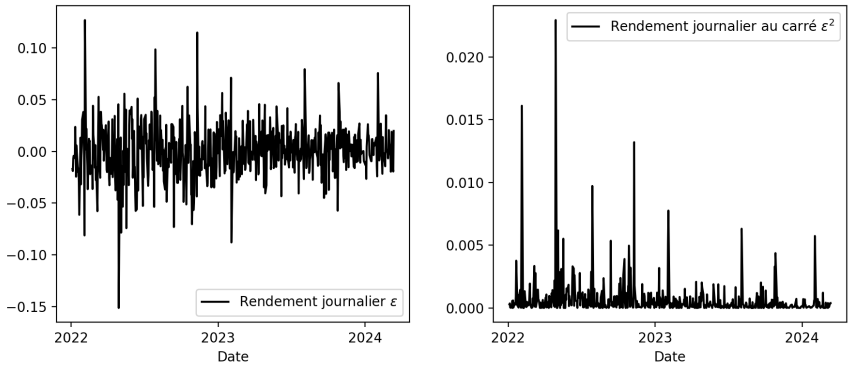


FIGURE 2 – Rendements et rendements au carré de l'action Amazon

1.2 Modélisation par un processus GARCH

1.2.1 Rappels sur les processus GARCH(p,q)

Les processus GARCH (*Generalized Autoregressive Conditional Heteroskedasticity*) sont généralement utilisés pour modéliser la volatilité conditionnelle dans les séries temporelles financières. Ces modèles ont été développés pour capturer la présence de l'hétéroscédasticité conditionnelle, c'est-à-dire la variabilité conditionnelle de la volatilité dans les données financières.

Formellement, un processus GARCH(p, q) peut être exprimé comme suit :

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

avec p l'ordre du terme autorégressif et q l'ordre du terme d'erreur conditionnelle [1].

1.2.2 Choix du modèle, entraînement et prédictions

Après avoir testé plusieurs choix de paramètres (p, q), nous choisissons de travailler avec un GARCH$(3, 3)$ (des valeurs plus élevées nous donnaient la même log-vraisemblance opitmale). Nous fixons notre ensemble d'entraînement aux données précédant juin 2023, sur lequel nous ajustons notre GARCH$(3, 3)$ avec le module `arch`. Nous pouvons ensuite simuler plusieurs trajectoires d'évolution des rendements de l'action, que nous visualisons sur la Figure 3 :

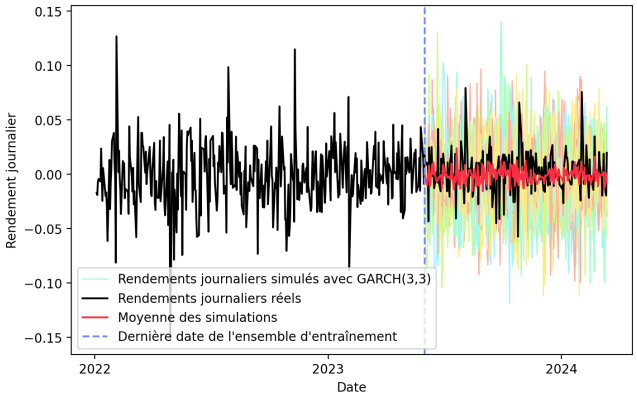


FIGURE 3 – Rendements réels et simulés, avec entraînement jusqu'en juin 2023

En inversant la transformation (1), nous simulons 200 trajectoires de l'évolution du cours de l'action :

1. Pour ne pas alourdir le rapport, les autre notations ne sont pas explicitées ici.

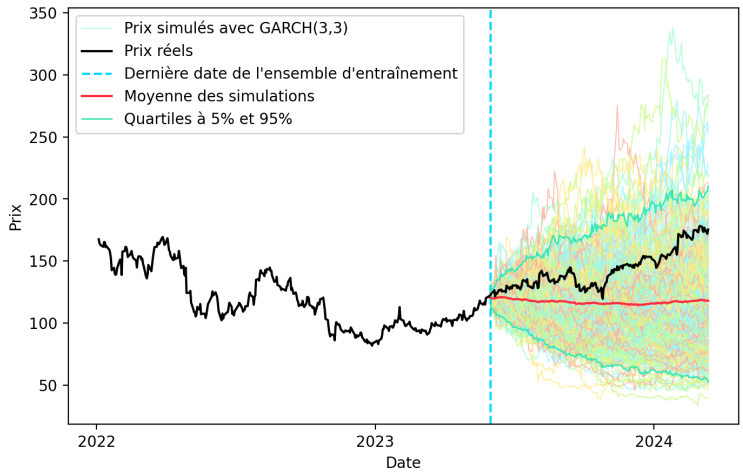


FIGURE 4 – Données réelles et simulées (200 trajectoires), avec entraînement jusqu'en juin 2023

Pour mieux visualiser les trajectoires individuelles, nous avons aussi choisi d'en visualiser seulement quelques unes sur la Figure 5.

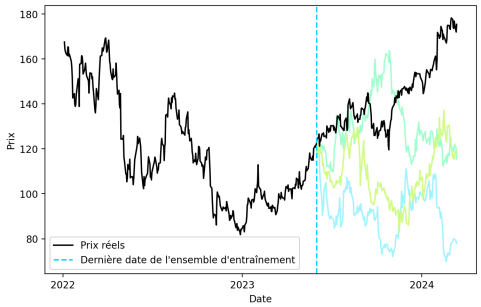


FIGURE 5 – Trois trajectoires simulées par GARCH(3,3)

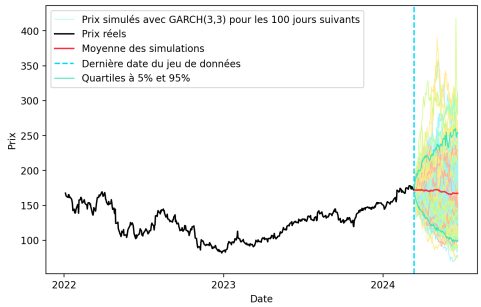


FIGURE 6 – Données réelles et prédictions sur les trois mois suivants

De la même manière, nous avons aussi entraîné notre modèle sur l'ensemble des données disponibles, puis simulé des trajectoires pour des dates auxquelles la valeur réelle n'a pas encore été observée. Nous visualisons les résultats sur la Figure 6, où les simulations sont réalisées sur un horizon de trois mois.

1.3 Discussion

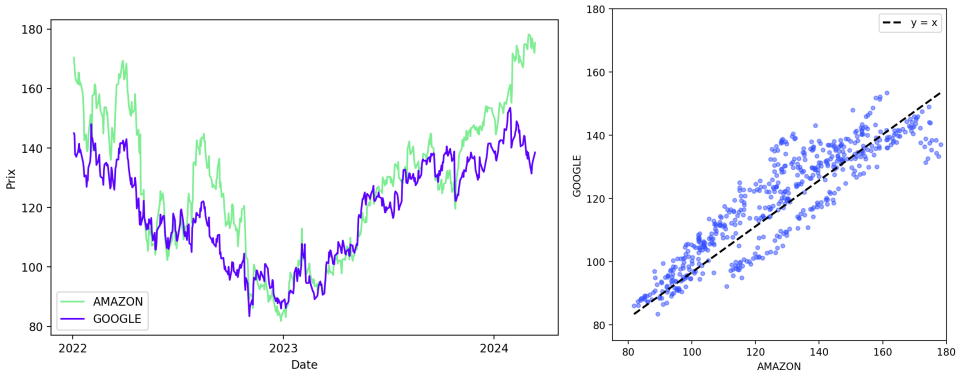
Les trajectoires simulées (voir Figure 5) ont une allure plausible, avec une composante aléatoire similaire à celle observée dans les données réelles. Nous remarquons que la moyenne des trajectoires tend vers une évolution quasi-constante (sur les Figures 4 et 6). Les quar-

tiles à 5%et 95% définissent une zone large, mais assez raisonnable pour considérer que notre modélisation est plutôt correcte.

2 Dépendance des cours des actions Amazon et Google

Pour cette étude, nous utilisons à nouveau le jeu de données NASDAQ Stock Prices, en considérant les données à partir de l'année 2022. Nous noterons Y (resp. X) la variable aléatoire représentant le prix de l'action Google (resp. Amazon).

2.1 Visualisation des données



(a) Cours des actions Google et Amazon (b) Scatter plot Google vs. Amazon

FIGURE 7 – Premières visualisations

Nous voyons sur la Figure 7 que les cours des deux actions ont une allure similaire, avec un scatter plot proche de la droite $y = x$. Avant d'aller plus loin dans l'analyse de la dépendance des deux variables, nous pouvons commencer par déterminer leur coefficient de corrélation linéaire :

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Nous obtenons $\rho(X, Y) = 0.903$, ce qui signifie que les deux cours sont presque linéairement dépendants (ce qui est cohérent avec la Figure 7).

2.2 Modélisation : Copules

Pour étudier la relation de dépendance non linéaire entre les cours des deux actions, nous utilisons trois types de copules : la copule de Clayton, la copule de Frank et la copule de Gumbel. Les copules sont un outil statistique permettant de modéliser la dépendance entre

des variables aléatoires, grâce à la fonction copule qui relie la densité jointe aux densités marginales.

2.2.1 Rappel des définitions

Nous rappelons ci-dessous les déinitions des copules de Clayton, Frank et Gumbel, que nous allons utiliser pour modéliser la dépendance des cours des deux actions :

Copule de Clayton : $C_\theta^{\mathcal{C}}(x,y) = \left(x^{-\theta} + y^{-\theta} - 1\right)^{-1/\theta}$

Copule de Frank : $C_\theta^{\mathcal{F}}(x,y) = \frac{1}{\theta}\ln\left(1 + \frac{(e^{-\theta x}-1)(e^{-\theta y}-1)}{e^{-\theta}-1}\right)$

Copule de Gumbel : $C_\theta^{\mathcal{G}}(x,y) = \exp\left(-\left((-\log x)^{-\theta} + (-\log y)^{-\theta}\right)^{1/\theta}\right)$

2.2.2 Ajustement aux données

À l'aide du module Python `copulalib`, nous obtenons les paramètres optimaux suivants après ajustement aux données :

- Copule de Clayton : $\hat{\theta} = 5.684$
- Copule de Frank : $\hat{\theta} = 13.495$
- Copule de Gumbel : $\hat{\theta} = 3.842$

2.2.3 Visualisation

Nous visualisons les résultats obtenus sur la Figure 8, où nous avons simulé 1000 couples (X_i, Y_i) pour chaque copule. Visuellement, nous remarquons que la copule de Clayton semble mieux capter la structure de dépendance que les deux autres : elle approche très bien la relation entre les deux variables pour les petites valeurs de (X, Y), et les points simulés restent relativement proches des données réelles. De manière analogue, la copule de Gumbel approche bien la dépendance pour les grandes valeurs de (X, Y), mais il y a de gros écarts aux données réelles pour les plus petites valeurs. Pour mieux quantifier les résultats, nous utilisons des coefficients de mesure de dépendance.

2.2.4 Comparaison des performances

Pour comparer les performances des trois copules, nous utilisons le coefficient de corrélation de Spearman $\rho_{\mathcal{S}}$ et le taux de Kendall $\tau_{\mathcal{K}}$:

$$\rho_{\mathcal{S}}(X,Y) = \rho(F_X(X), F_Y(Y))$$
$$\tau_{\mathcal{K}}(X,Y) = \mathbb{E}\left[\operatorname{sgn}\left((X-\tilde{X})(Y-\tilde{Y})\right)\right]$$

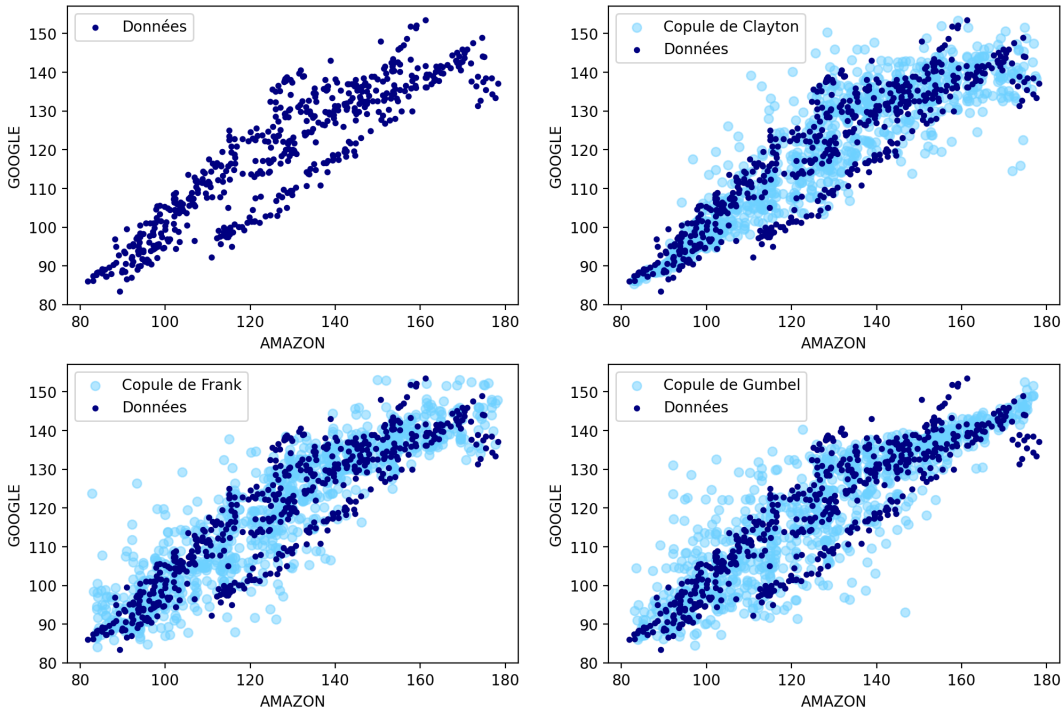


FIGURE 8 – Visualisation des copules ajustées : Simulation de 1000 couples (X, Y)

avec F_X (resp. F_Y) la fonction de répartition de X (resp. Y), et $(\tilde{X}, \tilde{Y})$ une copie i.i.d de (X, Y).

Nous comparons les coefficients obtenus pour les simulations des copules à ceux obtenus sur les données réelles, la modélisation étant considérée meilleure lorsque les coefficients se rapprochent des valeurs calculées sur les données. En estimant ces coefficients sur les données réelles et sur les échantillons de taille 1000 générés précédemment, nous obtenus les résultats présentés Table 1.

	Spearman	Kendall
Données	0.909	0.740
Clayton	0.902	0.739
Frank	0.921	0.750
Gumbel	0.918	0.757

TABLE 1 – $\rho_{\mathcal{S}}$ et $\tau_{\mathcal{K}}$ pour les données réelles et les différentes copules

En conclusion, conformément à ce qui avait été suggéré par la visualisation, la copule de Clayton semble être la plus adaptée pour modéliser la structure de dépendance des actions Amazon et Google.

3 Données Swissgrid - Consommation d'énergie en Suisse

3.1 Visualisation des données

Ce jeu de données recueilli par Swissgrid, la compagnie suisse de distribution d'électricité haute tension, contient la consommation d'énergie (MWh) en Suisse du 1er janvier 2019 au 01 mars 2024, avec un pas de temps horaire. Il contient 45264 points. La figure 9 affiche les données de consommation énergétique.

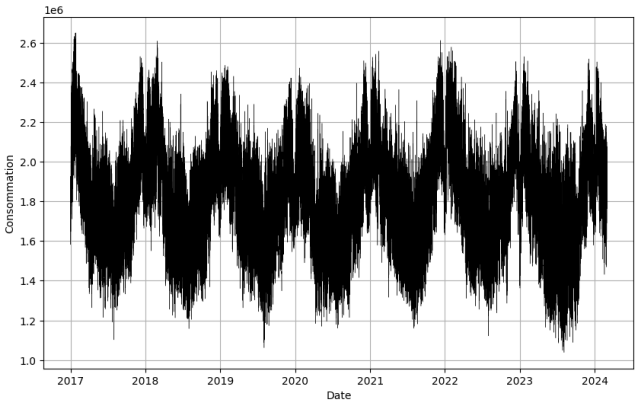


FIGURE 9 – Evolution de la consommation énergétique en Suisse

Au vu de la périodicité de la consommation, on choisit de modéliser ce jeu de données en une série temporelle $x_n = T_n + s_n + \epsilon_n$ où T_n est linéaire, s_n est périodique et ϵ_n est un résidu non-périodique et sans tendance. Dans la suite, nous allons déterminer ces différentes séries.

3.2 Prévision sur l'année 2024

Nous allons déterminer les différentes séries en deux temps :

- Identification de la composante linéaire T_n par régression linéaire
- Détermination de la composante saisonnière s_n et du résidu ϵ_n

3.2.1 Identification de la tendance linéaire

Pour identifier la tendance linéaire T_n, nous choisissons d'effectuer une régression linéaire. Nous obtenons :

$$T_n = -1.01n + 1.60 \times 10^6 kWh$$

avec n en heures. On note ainsi une tendance à la diminution de consommation de -1.01 kWh/h, soit -8.89 MWh/an. En supprimant la tendance linéaire à la consommation, nous obtenons un signal centré, avec les tendances périodiques inchangées.

3.2.2 Identification de la tendance saisonnière

Expliquer pourquoi la consommation est périodique : annuelle La consommation d'énergie congtient plusieurs composantes périodiques, notamment **annuelle**, **hebdomadaire** et **journalière**. Pour identifier ces tendances, nous effectuons une décomposition de Fourier du résidu ainsi obtenu avec le module **numpy.fft**. Nous choisissons les fréquences dont l'amplitude est la plus forte dans la décomposition de Fourier. Ces fréquences nous apportent des renseignements sur la périodicité du signal. Les contributions les plus importantes sont visible dans le tableau 2.

On reconnaît notamment la périodicité annuelle, d'amplitude 2.39 GWh, puis la périodicité journalière, bi-journalière et enfin hebdomadaire. D'autres périodicités d'amplitude moindre sont aussi présentes, mais moins interprétables.

Fréquence (jour^{-}1)	Amplitude (GWh)	Saisonnalité
0.00265	0.239	annuelle (période 377.2 jours)
1.0	0.199	journalière
2.0	0.105	bi-journalière
0.143	0.103	hebdomadaire (période 7.01 jours)

TABLE 2 – Détail des contributions périodiques les plus importantes

Nous pouvons alors reconstruire la consommation à partir de la décomposition de Fourier, en supposant la même périodicité de la consommation future. A l'ordre 1, seule la périodicité annuelle est prise en compte. A l'ordre 4, on prend aussi en compte la périodicité journalière, hebdomadaire et bi-journalière. Sur la figure 10, on affiche la prédiction de consommation à l'échelle de la semaine. A titre de comparaison, on affiche aussi la prédiction à un ordre bien plus élevé en considérant les 50 fréquences d'amplitude les plus élevées. On observe que la prédiction est améliorée mais le gros de la périodicité est saisi déjà à l'ordre 4.

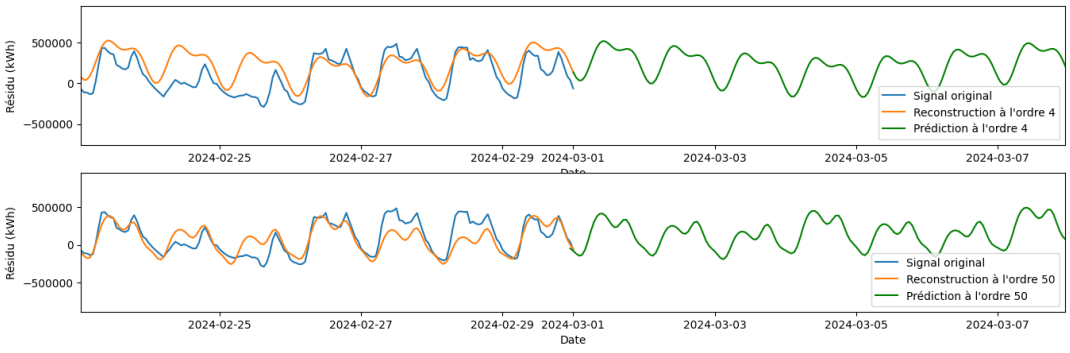


FIGURE 10 – Zoom sur la reconstruction du résidu à l'échelle hebdomadaire

La prédiction finale est obtenue en combinant la régression linéaire à la décomposition périodique à l'ordre 4. Elle est affichée figure 11.

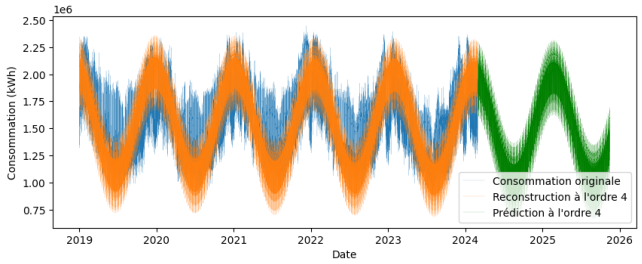


FIGURE 11 – Reconstruction finale de la prédiction de consommation

3.2.3 Analyse

La tendance globale de la consommation est stable au fil des ans, avec un coefficient directeur de -8.89 MWh/an, assez négligeable devant la moyenne de consommation à 1.58 GWh. L'augmentation du nombre d'appareils électriques est mise en concurrence avec les mesures de sobriété énergétiques visant à économiser l'énergie.
Avec l'analyse de Fourier en fréquence, nous gagnons en compréhension sur la périodicité de la consommation. Logiquement, la consommation varie beaucoup annuellement, au rythme des saisons, car l'hiver est plus froid et la consommation en chauffage augmente. Les tendances journalière et hebdomadaire est également importantes. On découvre aussi une forte tendance bi-journalière, avec deux pics de consommation entre 7 et 8h et vers 18h, et des heures de creux entre ces périodes. En poussant plus loin les ordres pris en compte, on peut obtenir un signal plus précis.
L'affichage des résidus en figure 12 montre l'amplitude de la consommation non prédite par notre reconstruction de série temporelle. A ordre 4, on observe encore des traces de périodicité dans le résidu mais à un ordre plus élevé, on obtient bien un résidu sans composante périodique, ce qui conforte notre approche. Dans ce résidu, on observe également des pics de consommation plus élevés en hiver qu'en été, atteignant rarement 0.5 GWh.

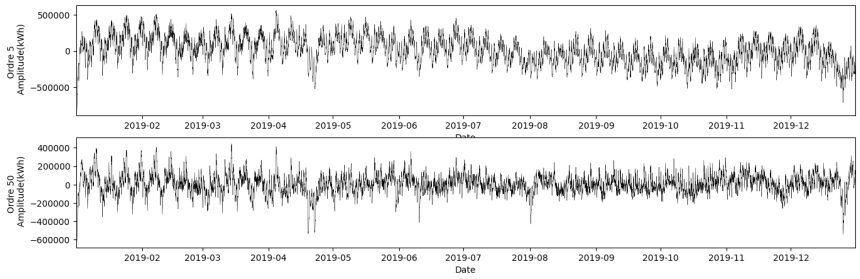


FIGURE 12 – Résidu sans tendance et sans composante périodique

4 Trafic routier à Istanbul

4.1 Visualisation des données

Le trafic routier est souvent étudié car sa prédiction permet d'adapter le nombre d'agents de prévention, de prévenir le risque d'accidents ou de fluidifier le trafic. Dans cette étude, nous nous intéressons au nombre de véhicules à Istanbul en décembre 2023, à la longitude 28.9874267578125 et latitude 41.0696411132813.

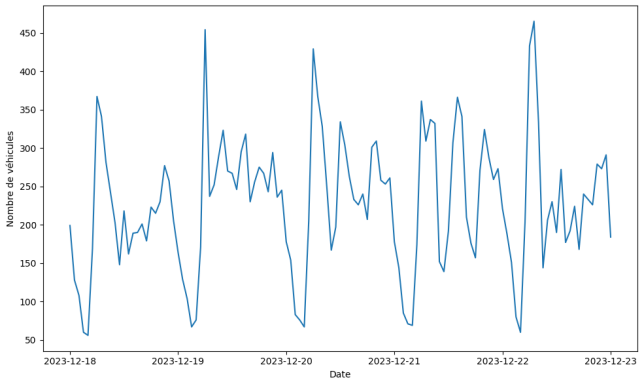


FIGURE 13 – Nombre de véhicules entre le 18 et 23 décembre 2023

Le trafic est visible figure 13. On observe une périodicité journalière importante, avec des pics le matin et des creux au milieu de la nuit. On va donc tenter de prendre en compte cette périodicité avec un modèle SARIMA.

4.2 Prédiction par modèle SARIMA

L'ARIMA (AutoRegressive Integrated Moving Average) est un modèle de série temporelle qui combine les composantes autorégressive (AR), intégrée (I) et moyenne mobile (MA). Le SARIMA (Seasonal ARIMA) étend l'ARIMA en ajoutant également des termes saisonniers pour modéliser la périodicité.
Dans un modèle SARIMA$(p, d, q)(P, D, Q, S)$, p correspond au nombre de termes autorégressifs, d est la fréquence de différenciation nécessaire pour passer d'une série non stationnaire à stationnaire, q est l'ordre de moyenne mobile. P, D et Q sont leurs correspondants saisonniers, et s correspond à la période de la saisonnalité. Certains de ces paramètres peuvent être identifiés empiriquement à partir de la fonction d'autocorrélation (ACF) et de la fonction d'autocorrélation partielle (PACF), affichées figure 14.

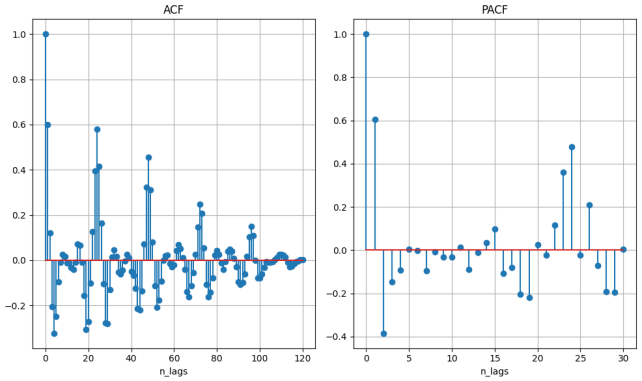


FIGURE 14 – Fonction d'ACF et de PACF du nombre de véhicules au cours du temps

D'abord, nous pouvons effectivement identifier une période $s = 24h$ grâce à la visualisation de la fonction d'autocorrélation, avec des corrélations fortes tous les 24 pas. Le nombre p peut être choisi comme le premier *lag* où la fonction PACF vaut environ 0, ici $p = 5$. En différenciant la série, on trouve les valeurs de d et D optimales permettant de rendre la série stationnaire. Nous utilisons ensuite le critère d'information d'Akaike (AIC) pour déterminer les paramètres restants.
Nous effectuons l'entraînement du modèle sur 3 jours de données (18 au 20 décembre 2023), et nous tentons de prédire les 2 jours suivants. Les résultats obtenus sont présentés figure 15. Nous avons comparé le résultat empirique à un modèle obtenu par recherche automatique des meilleurs paramnètres grâce à la fonction *auto-arima* de **pmdarima**, qui cherche à optimiser l'AIC.

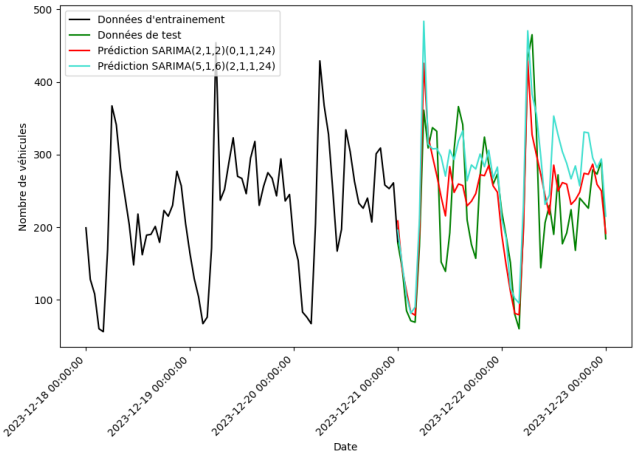


FIGURE 15 – Prédiction du nombre de véhicules par 2 modèles SARIMA. En bleu, la prédiction du SARIMA avec paramètres empiriques. En rouge, la prédiction du SARIMA par recherche automatique des meilleurs paramètres

Nous pouvons constater que la tendance périodique est bien saisie par les deux modèles SARIMA, avec une bonne reproduction des pics le matin et des creux. Pour les périodes entre ces pics, la prédiction est plus complexe et la prédiction a tendance à s'éloigner de la réalité. Le modèle automatique obtient une Mean Average Error (MAE) de 42 contre 51 (en unité nombre de véhicules) pour le modèle SARIMA avec paramètres déterminés empiriquement, il est donc légèrement plus performant.

5 Séismes au Chili

5.1 Visualisation et preprocessing des données

Le Chili est couramment frappé par des séismes, souvent suivis de plusieurs répliques. L'objet de notre étude est de comprendre la structure temporelle de ces événements, en inférant leur fréquence mais aussi le nombre de répliques générées, ainsi que le délai précédant leur déclenchement.

Pour cela, nous disposons d'un jeu de données recensant les séismes au Chili depuis 2012 et jusqu'en 2024. On distingue sur la figure 16 des paliers vers 2014 et 2019, pour lesquels apparaissent successivement des points de magnitude plus faible. Cela correspond vraisemblablement à la prise en compte de séismes moins importants dans le jeu de données, grâce à de nouveaux systèmes de mesure.

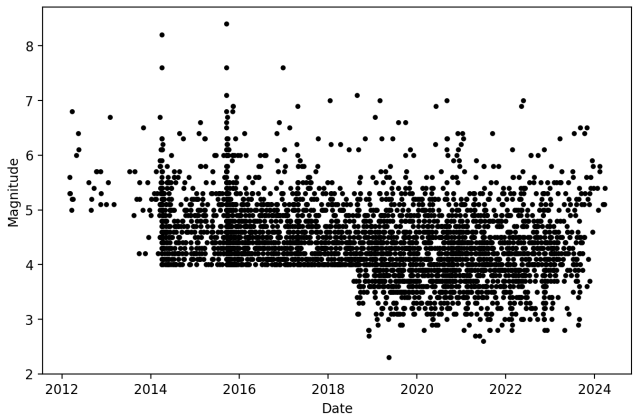


FIGURE 16 – Magnitude des séismes enregistrés au cours du temps

Pour mener une analyse correcte compte tenu de la collecte des données, nous avons choisi de ne considérer que les séismes de magnitude supérieure à 6. On obtient alors la figure 17, pour laquelle nous avons choisi une représentation en escalier afin de mieux visualiser l'écart temporel entre les événements [2]. Les durées sont exprimées en nombre de jours.

2. Nous nous intéressons seulement à l'intervalle de temps entre deux séismes, la magnitude n'intervient

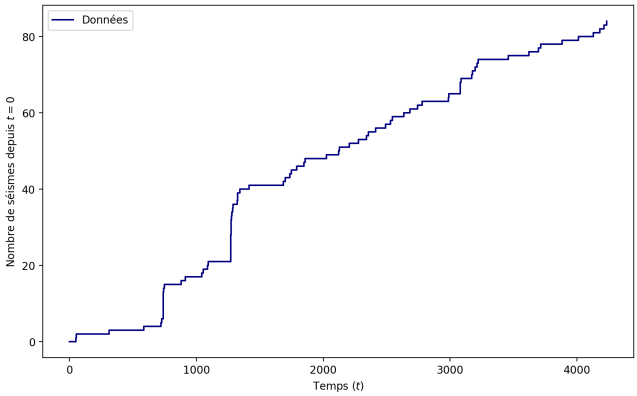


FIGURE 17 – Évolution du nombre de séismes de magnitude supérieure à 6 depuis $t = 0$

5.2 Modélisation : Processus de Hawkes

5.2.1 Définition

Un processus de Hawkes $(N_t)_{t\geq 0}$ est un processus ponctuel auto-excitateur, dont l'intensité au temps t, notée λ_t, est donnée par

$$\lambda_t = \mu + \int_0^t \phi(t-s)dN_s,$$

où μ est un nombre réel positif et ϕ un noyau de régression.

Les processus de Hawkes ont été introduits dans le but de modéliser les tremblements de terre et leurs répliques. Ils sont donc couramment utilisés en sismologie car ils permettent de modéliser l'auto-excitation, et les séismes sont connus pour déclencher des répliques. On voit en effet sur la figure 17 que les séismes arrivent souvent par paquets. Pour modéliser les occurences des séismes, nous allons donc utiliser un processus de Hawkes univarié, dont il reste à choisir le noyau.

Choix du kernel : Nous choisissons de travailler avec un noyau exponentiel, de la forme $\phi(t) = \alpha e^{-\theta t}$, où α et θ seront des paramètres à estimer. Nous obtenons donc l'intensité

$$\lambda_t = \mu + \alpha \int_0^t e^{-\theta(t-s)}dN_s,$$

avec les interprétations suivantes pour les différents paramètres :

— μ est le taux latent d'apparition d'un séisme (de manière exogène),

— α peut être interprété comme le nombre moyen de répliques déclenchées par chaque séisme,

pas dans la suite de l'étude.

— θ est le taux de la loi exponentielle définissant le délai entre un séisme et une réplique ; en moyenne, ce délai vaut donc $1/\theta$.

5.2.2 Estimation et interprétation des paramètres

Partant de la définition mathématique précédente, nous ajustons le modèle aux données pour maximiser la vraisemblance (à l'aide du module `hawkeslib`). En exprimant les durées en nombre de jours, on obtient les valeurs optimales suivantes pour les paramètres du modèle :

$$\hat{\mu} = 0.0129,\ \hat{\alpha} = 0.3636,\ \hat{\theta} = 0.3303$$

Interprétation : Nous concluons que les séismes de magnitude supérieure à 6 se produisent de manière exogène environ une fois tous les $1/\hat{\mu} \sim 78$ jours au Chili, soit presque tous les trois mois (ce qui semble cohérent avec la figure 17). Chaque secousse principale entraîne en moyenne $\hat{\alpha} = 0,36$ répliques, qui surviennent en moyenne avec un délai de $1/\hat{\theta} = 3$ jours.

5.2.3 Simulations

Pour finir, nous utilisons le modèle et les paramètres optimaux précédents pour simuler plusieurs trajectoires. Sur la figure 18, nous avons visualisé les tracés pour 100 trajectoires, ainsi que la moyenne du nombre de séismes à chaque instant sur l'ensemble des trajectoires.

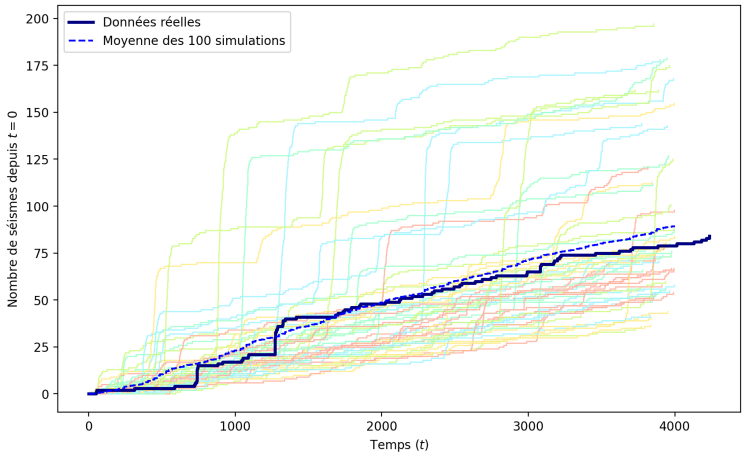


FIGURE 18 – Données réelles et $n = 100$ simulations

Nous pouvons voir que la moyenne capture bien la tendance globale du jeu de données, et que l'allure des simulations est plausible. Il y a quelques emballements exponentiels (ce qui s'explique par la définition du modèle), mais la plupart des trajectoires restent du même ordre de grandeur que les données réelles.