MAP565 : ANALYSES STATISTIQUES DE JEUX DE DONNÉES

Avril 2023

Etienne Gauthier, Imrane Alioua

TABLE DES MATIÈRES

1 DONNÉES UTILISÉES POUR LE MÉMOIRE

1.1 Données financières du Nasdaq

Dans le but de compléter notre étude, nous avons récupéré les symboles des entreprises du Nasdaq à l'adresse suivante :`https://www.nasdaq.com/market-activity/stocks/screener`

Nous avons utilisé le fichier csv pour obtenir une liste d'entreprises côtées sur le marché. Ensuite, à partir d'un jupyter notebook nous avons extrait des données de Yahoofinance à partir du ticker lié à leur symbole.

L'explication du choix de nos données est faite plus en détail dans la partie correspondante.

1.2 Données météorologiques à New-York

Pour notre étude, nous avons décidé d'étudier un phénomène présentant à la fois une "tendance", une "saisonnalité" et du "bruit", Afin d'étudier le comportement de certaines variables aléatoires, pour en comprendre son passé et prédire son futur. C'est pourquoi nous avons décidé de récupérér des données météorologiques à l'adresse suivante : `https://www.visualcrossing.com/weather/weather-data-services`

Il s'agit de données journalières qui recensent des données météos (température, humidité, etc...) dans la ville de New-York. Nous avons travaillé avec les données de la période Juillet 2020 - Mars 2023 qui ont été enregistrées au format csv.

Voici une documentation des données en questions :`https://www.visualcrossing.com/resources/documentation/weather-data/weather-data-documentation/`

2
DONNÉES DE MARCHÉS FINANCIERS

2.1 Matrices aléatoires

2.1.1 • Motivations

L'un des domaines de la théorie des matrices aléatoires consiste à étudier la distribution des valeurs propres d'une "grande" matrice aléatoire. Sous certaines hypothèses, les valeurs propres trouvées dans une plage théorique prédite sont considérées comme dues à des interactions purement aléatoires dans les données. Pour cette raison, nous pouvons écarter les valeurs propres trouvées dans une telle plage prédite pour essayer de "filtrer" le caractère aléatoire de nos données et ne conserver que les tendances "réelles", celles qui ont un sens physique ou économique.

La matrice aléatoire dont nous allons parler est une matrice carrée de la forme W = X * Xtranspose où X est une matrice N×T et les entrées de X sont des variables normales aléatoires i.i.d.

Dans le cas où T > N et que N est grand, la distribution des valeurs propres d'une telle matrice W (telle que définie ci-dessus) suivra la distribution de Marchenko-Pastur.

Une matrice de corrélation peut être un exemple d'une telle matrice W si la matrice de données X contenant T points de données (correspondant à l'horizon d'étude) de N variables a des entrées normales i.i.d (qui seront des rendements log-gaussiens).

Bien que cela ne soit pas toujours vrai, il est courant de modéliser le logarithme des rendements quotidiens d'un cours boursier comme une distribution normale. Cette approximation fonctionne suffisamment bien pour qu'une grande partie de la théorie moderne de la finance soit basée dessus. Par conséquent, la matrice de corrélation des log-rendements correspond en partie à l'exemple ci-dessus.

En finance, un filtrage des corrélations est réalisée à partir de la méthode suivante :
- On calcule les valeurs propres de la matrice de corrélation.
- Si elles se situent dans la fourchette théorique donnée par la distribution de Marchenko-Pastur, nous fixons ces valeurs propres à 0, puis nous reconstruisons la matrice.

Toute valeur propre trouvée dans cette fourchette est considérée comme une information due au hasard et est donc rejetée.

2.1.2 • Choix de nos données

Étant donné que N et T doivent être grands, nous avons utilisé un ensemble de données d'entreprises côtées sur le Nasdaq. Nous avons besoin à la fois d'un grand nombre d'actifs financiers et d'un grand nombre de données. Nous avons récupéré les symboles des entreprises du Nasdaq à l'adresse suivante :`https://www.nasdaq.com/market-activity/stocks/screener`

Ensuite, nous avons extrait des données de Yahoofinance à partir du ticker lié à leur symbole. Nous avons décidé de sélectionner aléatoirement, N actions pour le tracé des graphes à venir. Ce nombre N doit être assez grand pour dénoter d'une "convergence suffisante" des théorèmes sur les matrices aléatoires, toutefois pour des raisons d'ordre esthétiques (l'affichage des graphes pouvant devenir illisible avec des données trop volumineuses) nous retiendrons une valeur N de l'ordre de 100 pour notre étude. Etant donné les valeurs des actions, nous en déduisons leurs rendements à partir d'opérations sur les dataframes de la librairie pandas. Le lecteur intéréssé pourra retrouver les codes utilisés en annexes.

Après un nettoyage des données qui consistait principalement à vérifier qu'il n'y avait pas de données manquantes, où de phénomènes totalement exogènes (cours de l'action divisé par deux en un jour), nous avons appliqué une transformation logarithmique à nos rendements.

$$f(r_t) = \log(1 + r_t)$$

.

2.1.3 • Valeurs propres de la matrice de corrélation

Grâce à l'opération corr() de la librairie Numpy, on obtient la matrice de corrélation

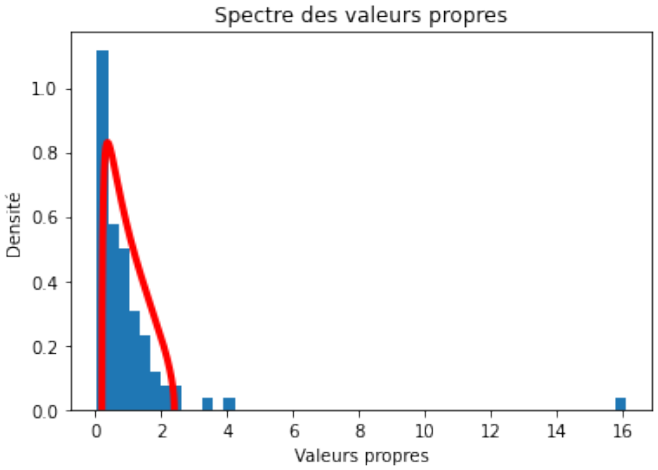

Figure 1 – Valeurs propres de la matrice de corrélation

En rouge, il s'agit de la loi théorique de Marcenko-Pastur. En bleu, l'histogramme des valeurs propres obtenu pour une matrice de taille 169 (jours de marchés) x 81 (actifs du Nasdaq).

On remarque qu'une valeur propre est plus grande que les autres, avec une marge non négligeable. Il s'agit de la valeur propre du marché, qui est parfois appelé "mode marché".

Enfin, nous avons supposé que la matrice de corrélation était stationnaire. Toutefois, les marchés, régis par un monde extrêmement complexe, ne sauraient être stationnaires. Ainsi, l'étude de matrices aléatoires, utilisée dans le management de risque de portefeuille, peut sous-estimer le risque réel du portefeuille optimal déterminer à partir de la matrice de corrélation. En effet, les actifs financiers ont une volatilité qui change au cours du temps, ils réagissent à de nouvelles tendances du marché, et la correlation entre deux actifs dépend elle-même des conditions du marché. Quelquefois, les correlations peuvent connaître des sauts soudains en raison d'informations extérieures (géopolitique, catastrophes naturelles...).

2.2 Copules

2.2.1 • Motivations

Pour de nombreuses applications financières, le problème est de trouver une distribution efficace pour décrire certains phénomènes, par exemple les relations entre les rendements de différents actifs. Très souvent, les distributions des returns sont supposées être des gaussienne multivariées ou une distribution log-normale afin de faciliter les calculs alors que cette hypothèse est bien souvent inapropriée. En effet, la distribution des rendements présente des queues lourdes. Dans le cadre de l'analyse de portefeuille, la variance correspond à la mesure du risque, mais cela implique que le monde possède un comportement gaussien... C'est pourquoi les copules sont utiles, pour décrire la structure de dépendance du modèle. Elles relient les lois marginales à leur distribution multivariée.

En général, une copule est utilisée pour séparer le caractère purement aléatoire d'une variable (actif financier par exemple) des interdépendances entre cette variable et d'autres variables. On peut ainsi décider modéliser chaque variable séparément et donc, disposer d'une mesure des relations entre ces variables. On peut choisir pour chaque actif le type de distribution le plus approprié, sans influencer les interdépendances entre ces variables/actifs.

2.2.2 • Propriétés des marchés financiers

Les données empiriques ont largement prouvé que la distribution multinormale est inadéquate pour modéliser la distribution des rendements des actifs financiers d'un portefeuille...

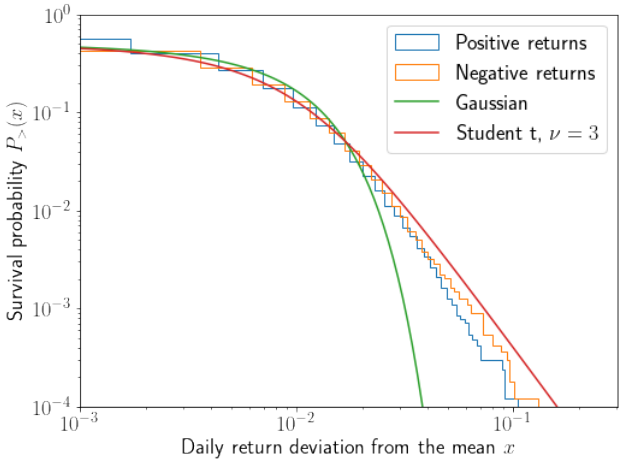

Figure 2 – Queues de distribution du SP500 de 1957 à aujourd'hui

Les distributions marginales empiriques sont asymétriques et à queue large. La distribution normale ne prend pas en compte la possibilité de co-mouvements conjoints extrêmes pour les rendements des actifs financiers.

2.2.3 • Modélisation de la dépendance entre deux actifs financiers

Sur les marchés certains actifs peuvent présenter une corrélation particulière, en fonction du type d'industrie sous-jacent, des croyances des acheteurs, de l'état de santé de l'économie...

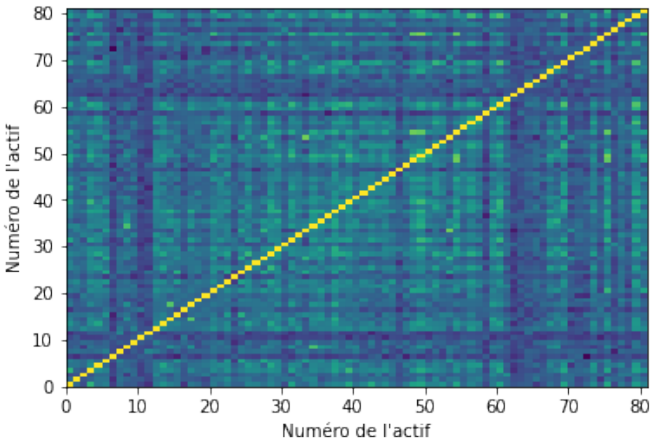

Figure 3 – Matrice de corrélation d'actifs du Nasdaq choisis aléatoirement sur Yahoo Finance

La période de l'étude s'étend pour tous les jours d'ouverture de la bourse entre le 1er janvier 2010 et le 31 décembre 2022.

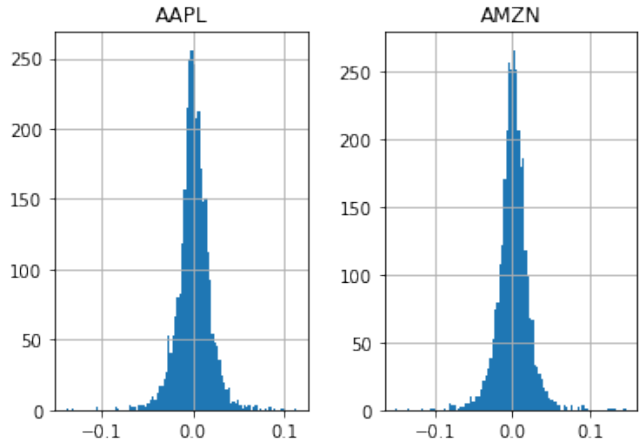

Figure 4 – Histogramme des log rendements pour les actions AAPL et AMZN, bins=100

Nous remarquons des valeurs centrales plus ”piquées” et des queues de distribution plus épaisses que la courbe de la fonction de densité gaussienne. Ceci s'explique en ce que la distribution empirique présente des évènements rares avec une décroissance plus lente en loi de puissance alors que la loi normale a une décroissance exponentielle.

Le choix de la copule représente la première difficulté dans la mise en pratique de la modélisation de la dépendance. Le graphique suivant permet d'appréhender la forme des dépendances qui existent entre l'indice AAPL et l'indice AMZN. Outre le sens (l'orientation de la pseudo-ellipse) et l'intensité des dépendances, ce graphique nous donne un premier renseignement sur la dépendance entre les queues de distribution.

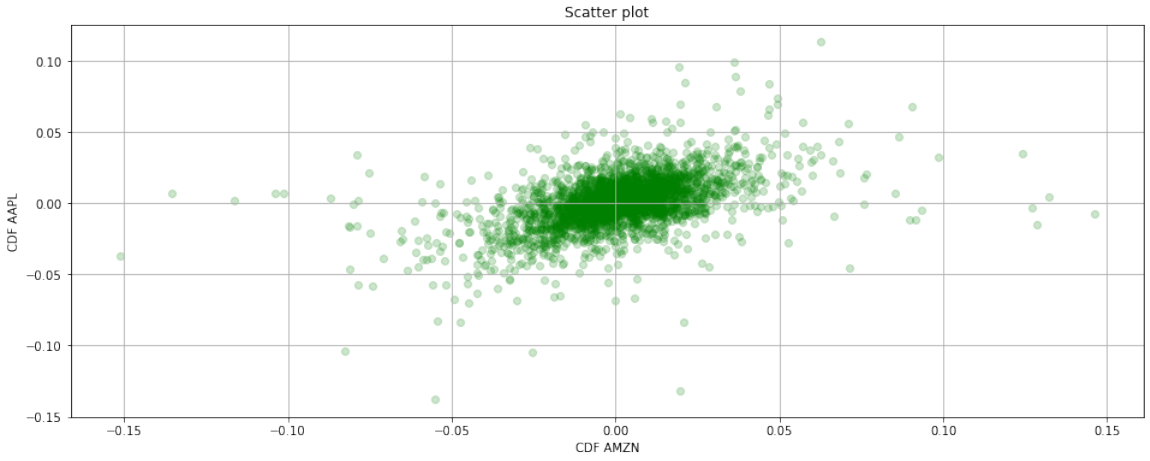

Figure 5 – Scatter plot des log rendements de 2010 à 2022

On remarque une dépendance "positive" entre les deux variables, en effet lorsque les rendements d'une action sont positifs, il en va souvent de même pour l'autre action, de même pour des rendements négatifs. Cela est confirmé par les coefficients de corrélation de Kendall, Pearson et Spearman de 0.35, 0.48, 0.48 respectivement après arrondi au centième.

2.3 Etudes des extrêmes

La partie suivante de notre travail porte sur la théorie des valeurs extrêmes. Cette théorie cherche à analyser les distributions possibles d'événements extrêmes. Il s'agit d'événements qui peuvent se produire, mais qui sont rarement arrivés auparavant (ou même jamais arrivés).

2.3.1 • Choix des actifs

Tout d'abord, nous identifierons l'actif de notre portefeuille dont les queues de distribution pour les rendements sont lourdes. Pour ce faire, nous calculons la kurtosis de la distribution des rendements quotidiens pour certains actifs du Nasdaq.

	Kurtosis
NMCO	1.897084
SMFG	2.992363
MSGE	3.839213
INDB	5.237862
SCYX	6.514629
RVNC	16.959367
JAN	26.076229
AHG	200.858401

Figure 6 – Exemple de Kurtosis d'actifs du Nasdaq entre janvier 2020 et mars 2023

Le kurtosis est défini comme la mesure des valeurs extrêmes présentes dans la queue d'une distribution. Plus le chiffre est élevé, plus la queue est lourde. Nous avons décidé de conserver les actifs 'JAN' et 'RVNC'. Puis, nous vérifions que l'actif 'JAN' a bien une queue de distribution plus lourde.

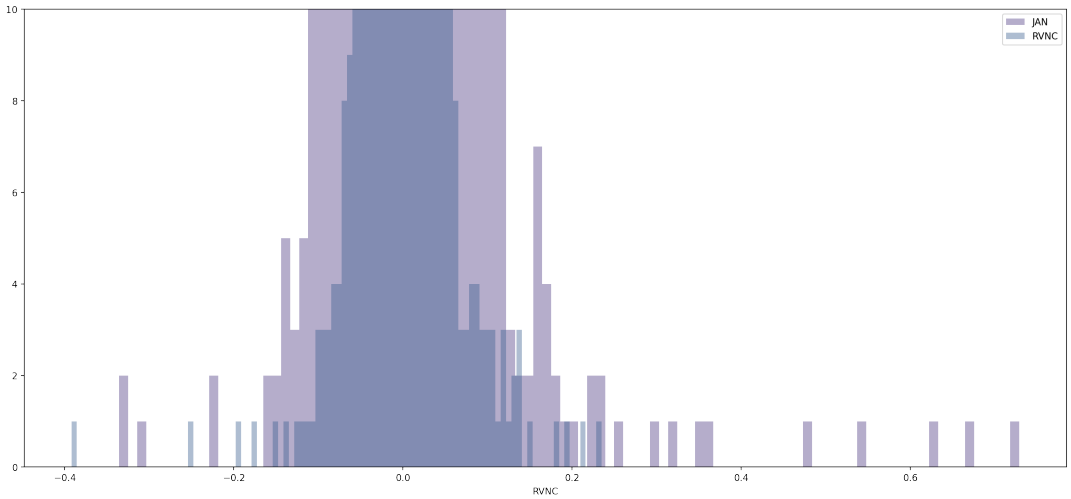

Figure 7 – Histogrammes des actifs JAN et RVNC entre janvier 2018 et décembre 2022

Dans la suite, on considère les log-rendements entre janvier 2018 et décembre 2022 des actifs JAN et RVNC.

2.3.2 • Choix du seuil

Nous suivrons la méthode des peak-over-threshold. Cette méthode répondra à la question de savoir de combien un seuil peut être dépassé quand la perte dépasse un certain seuil.

Le choix du seuil est très important. Il doit y avoir un compromis entre le seuil et la quantité de données. Cela signifie que si l'on choisit un seuil élevé, il y aura moins de données, puisque moins d'observations dépassent le seuil. Comme il y a moins de données disponibles dans ce cas, la variance et l'incertitude augmentent. Au contraire, si nous choisissons un seuil plus bas, nous obtiendrons plus de données, ce qui entraînera une diminution de la variance. Cependant, l'approximation de $F_u(x)$ par la distribution généralisée de Pareto n'est précise que dans les queues. Elle crée donc un biais. En bref, il existe un compromis entre le biais et la variance.

Notre première tentative pour définir le seuil consiste à tracer le paramètre ξ en fonction de différents seuils, avec son intervalle de confiance et le nombre de dépassements :

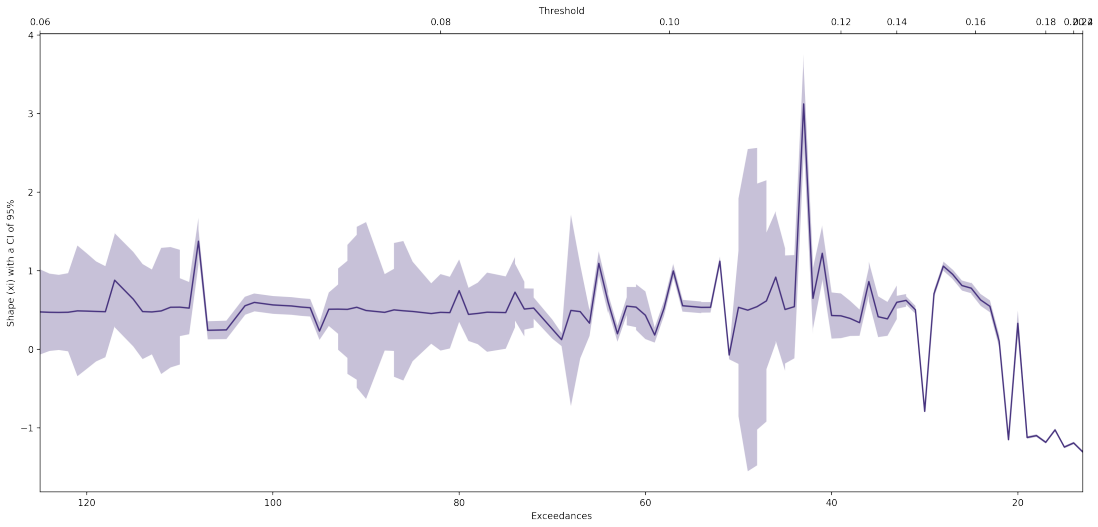

Figure 8 – Graphes des excès de 'JAN' entre janvier 2018 et décembre 2022

Nous travaillons avec le quantile à 0.96 pour déterminer notre seuil. Nous pouvons maintenant trouver les dépassements de ce seuil, ce qui nous permet d'estimer les paramètres de la distribution GPD.

	Valeur
Location	0.108511
Scale	0.059191
Xi	0.498145

Figure 9 – Valeurs de la loi GPD entre 2018 et 2022 pour 'JAN'

La valeur de ξ est d'environ 0.49 et indique que, si l'on effectue un test avec l'hypothèse nulle que ξ est égale à zéro, celle-ci ne serait pas rejetée. Cependant, si nous négligeons cela un instant, nous observons un paramètre positif, ce qui signifie que la queue de distribution est épaisse.

2.3.3 • Résultats

Après avoir fit nos données à une loi généralisée de Pareto, nous cherchons à comparer les résultats obtenus par la GPD avec deux autres méthodes faisant appel à un fit de loi de Student et une méthode utilisant les quantiles historiques.
Etant donné que la crise Covid et la guerre entre l'Ukraine et la Russie ont eu un effet sur les marchés financiers depuis 2020, on réalise une comparaison des différentes méthodes avec et sans les données postérieures au 1er janvier 2020.

	Value-at-Risk	Expected Shortfall
GPD (incl. 2020-2022)	0.695398	1.258388
GPD (excl. 2020-2022)	0.495628	0.711182
Student's t4 (incl. 2020-2022)	0.199135	0.504201
Student's t4 (excl. 2020-2022)	0.221117	0.523041
Historical Simulation (incl. 2020-2022)	0.098077	0.275194
Historical Simulation (excl. 2020-2022)	0.145113	0.297631

Figure 10 – VaR et Expected Shortfall de l'actif 'JAN'

L'Expected Shortfall est exceptionnellement grand pour l'actif JAN, en effet, sa valorisation a déjà connu quelques pics avant 2020.

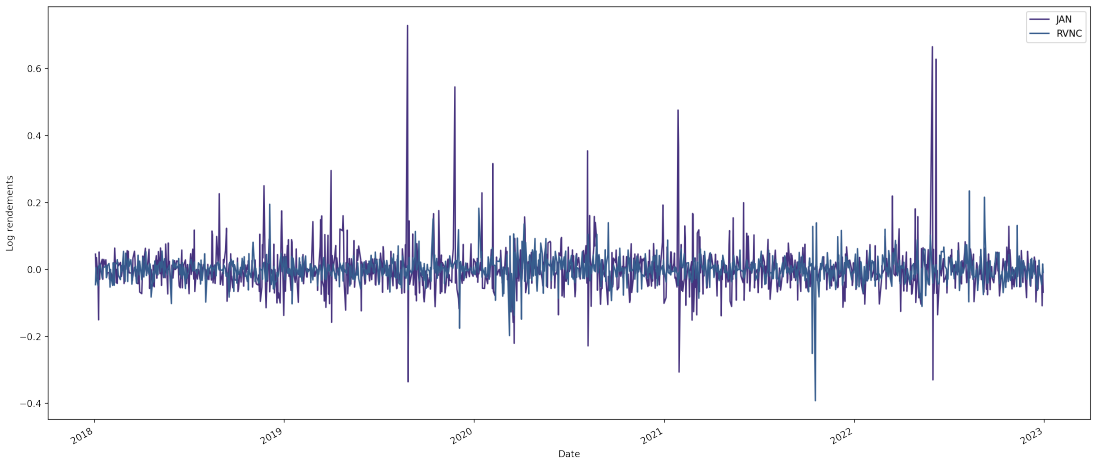

Figure 11 – Log returns des actifs 'JAN' et 'RVNC'

La première chose à noter est que seul le modèle GPD était préparé à une situation aussi extrême que la volatilité de ces dernières années. Le premier janvier 2020, l'ES attendu selon la distribution historique "n'est que" de 0.29 ! Un an plus tard, en février 2021, nous constatons que des pics de 0.45 sont assez fréquents. Les modèle GPD et Student s'attendaient encore à

de telles pertes, mais le mois de mars 2022 dépasse les estimations de la loi de Student.

Il convient de noter que nous comparons la VaR et l'ES du premier jour de 2020 à ceux du premier jour de 2023. Une prochaine étape de travail pourrait consister à examiner une estimation glissante, d'autant plus que nous constatons que le modèle GPD est capable de s'ajuster très bien sur cette courte période ; l'ES passe de 0,71 à 1.26 ! Il serait intéressant de voir ce que cela donne au fil du temps.

2.4 Modélisation de la volatilité avec GARCH

L'année 2022 a connu un marché en baisse, difficile pour les investisseurs. Le SP 500 a perdu près de 30 % de valeur par rapport à son plus haut niveau. La volatilité est au coeur de nombreux sujets de discussion.

Nous concentrons cette étude sur les données de l'actif 'AAPL' entre janvier 2020 et mars 2023, récupérées sur Yahoo Finance.

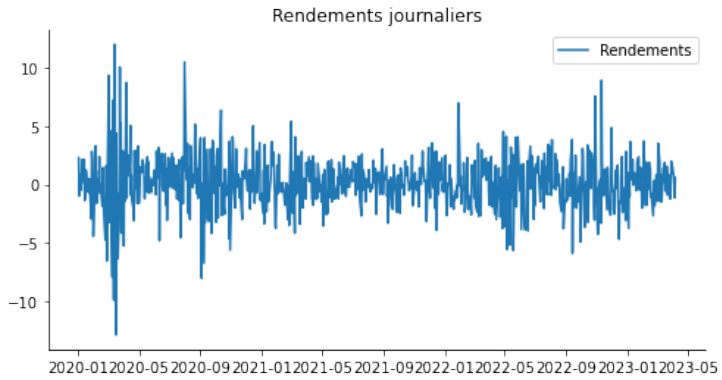

Figure 12 – Rendements de l'actif AAPL entre janvier 2020 et mars 2023

Nous allons calculer la volatilié sur 3 échelles de temps : quotidienne, mensuelle, annuelle.

— Volatilité quotidienne : pour l'obtenir, nous calculons l'écart-type des rendements quotidiens.
— Volatilité mensuelle : nous supposons qu'il y a 21 jours de bourse dans le mois et nous multiplions donc la volatilité quotidienne par la racine carrée de 21.
— Volatilité annuelle : nous supposons qu'il y a 252 jours de bourse dans une année civile, et nous multiplions la volatilité quotidienne par la racine carrée de 252.

	Volatilité quotidienne %	Volatilité mensuelle %	Volatilité annuelle %
Apple	2.27	10.41	36.06

Figure 13 – Volatilité de l'actif AAPL

La volatilité annuelle de Apple est de 36 %, ce qui est significatif mais relativement faible. Pour l'entreprise Moderna on trouve une volatilité annuelle de 83 %. Cela montre que les investisseurs ont du mal à jauger l'entreprise car sa valorisation "fondamentale" repose principalement sur les vaccins Covid-19. Tandis que Apple est une valeur refuge, présentant une certaine solidité même en temps de crise.

Désormais, on entraine notre modèle à partir de la librairie arch.

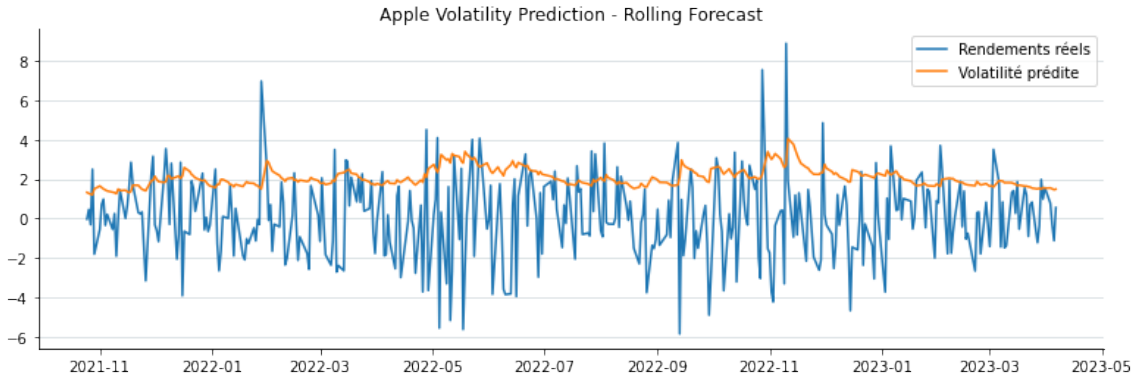

Figure 14 – Volatilité prédite de l'actif AAPL à un horizon de 5 jours

Nous pouvons voir sur les courbes ci-dessus que notre volatilité prédite est généralement cohérente avec les rendements quotidiens.

Cela signifie que notre modèle GARCH fonctionne bien dans cette situation. Les rendements journaliers sont élevés dans les zones où l'on s'attend à ce que la volatilité soit élevée.

3
SÉRIES TEMPORELLES LINÉAIRES APPLIQUÉES À DES DONNÉES MÉTÉOROLOGIQUES

3.1 Librairies utilisées

Dans le cadre de cette partie, nous étudions un phénomène saisonnier en language Python. Nous avons essentiellement utilisé les librairies pandas et statsmodels afin d'exécuter nos calculs et prédictions

3.2 Etude et prévision de la température moyenne

On s'intéresse à l'analyse et la prévision de données de séries temporelles stationnaires (où la moyenne et la variance des données ne changent pas au fil du temps) qui présentent une tendance et une saisonnalité.

3.2.1 • Description

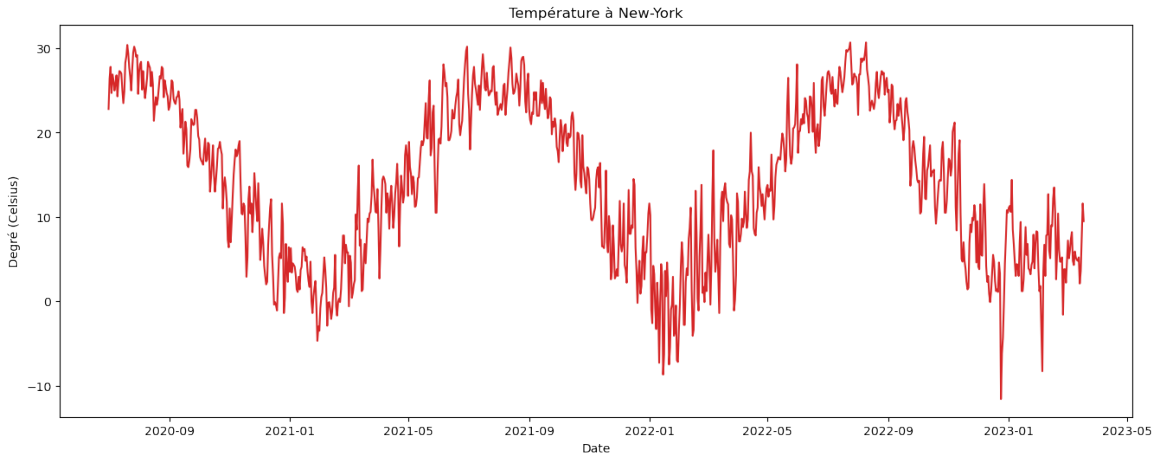

Figure 15 – Données de température à New York

Certains patterns distincts apparaissent lorsque nous représentons les données sur un graphique. La série temporelle présente des caractéristiques saisonnières : les températures sont toujours basses au début de l'année et élevées au milieu de l'année.

Cependant, le graphique ci-dessus est un peu bruyant, car il contient toutes les températures quotidiennes. Toutefois, en examinant attentivement les valeurs journalières, nous pouvons constater que la température ne change que "peu" d'un jour à l'autre.

Nous pouvons également visualiser nos données à l'aide d'une méthode de décomposition des séries temporelles qui nous permet de décomposer nos séries temporelles en trois composantes distinctes : la tendance, la saisonnalité et le bruit.

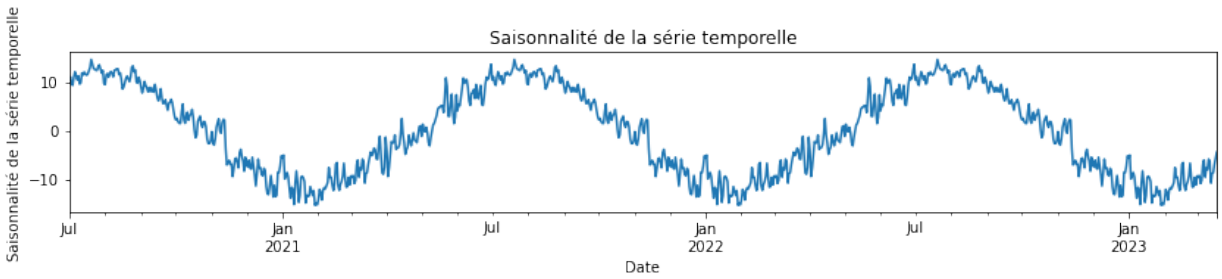

Figure 16 – Saisonnalité de la température à New York

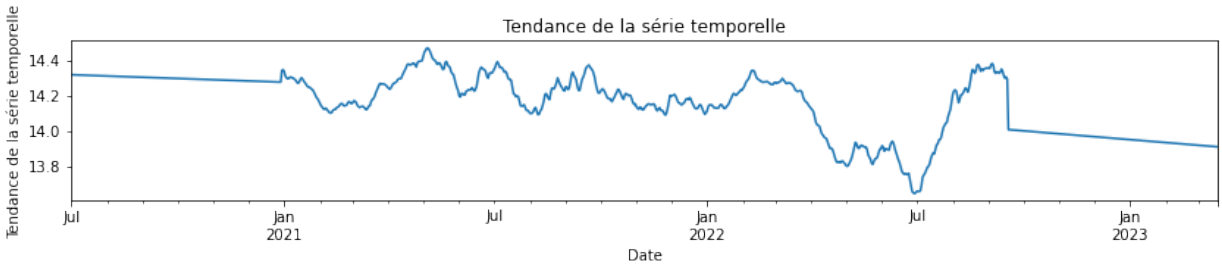

Figure 17 – Tendance de la température à New York

On remarque que la tendance varie peu, ce qui était prévisible. En effet, sur des échelles de temps aussi courte, on ne s'attendait pas à observer des phénomènes d'échelle de temps charactéristiques plus importantes tels que le réchauffement climatique par exemple.

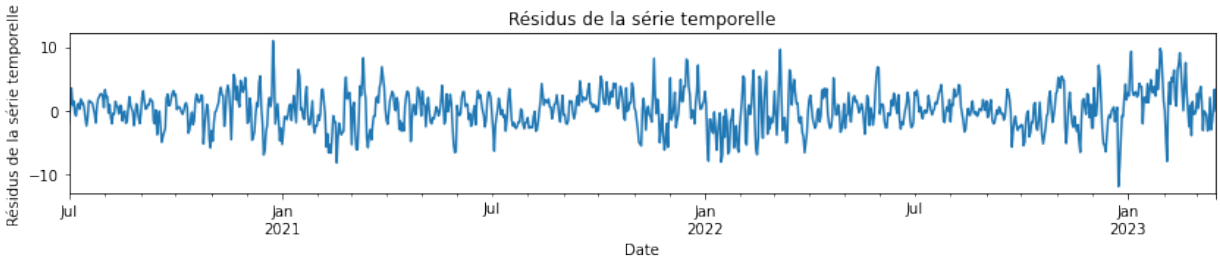

Figure 18 – Résidus de la température à New York

3.2.2 • Utilisation de SARIMA

Nous allons appliquer l'une des méthodes les plus couramment utilisées pour la prévision des séries temporelles, connue sous le nom de SARIMA (Seasonal Autoregressive Integrated Moving Average). Les modèles SARIMA sont désignés par la notation SARIMA(p,d,q)(P,D,Q,s). Ces trois paramètres tiennent compte de la saisonnalité, de la tendance et du bruit dans les données.

Nous utiliserons une "recherche en grille" pour explorer itérativement différentes combinaisons de paramètres. Pour chaque combinaison de paramètres, nous ajustons un nouveau modèle saisonnier SARIMA avec la fonction SARIMAX() du module statsmodels et évaluons sa qualité globale.

Comme critère de choix des paramètres on choisit l'AIC qui mesure l'adéquation d'un modèle aux données tout en tenant compte de la complexité globale du modèle. Un modèle qui s'ajuste très bien aux données tout en utilisant de nombreuses caractéristiques se verra attribuer un score AIC plus élevé qu'un modèle qui utilise moins de caractéristiques pour obtenir la même qualité d'ajustement. Par conséquent, nous cherchons à trouver le modèle qui produit la valeur AIC la plus faible.

La sortie de notre code suggère que "SARIMAX(1, 1, 1)x(1, 1, 1, 12)12" donne la valeur AIC la plus faible de 4844.91. Nous devrions donc considérer qu'il s'agit de l'option optimale parmi tous les modèles que nous avons envisagés.

On fit le modèle avec les paramètres détérminés plus haut. Puis on affiche certains opérateurs statistiques afin de détecter s'il y a des comportements anormaux dans notre modèle.

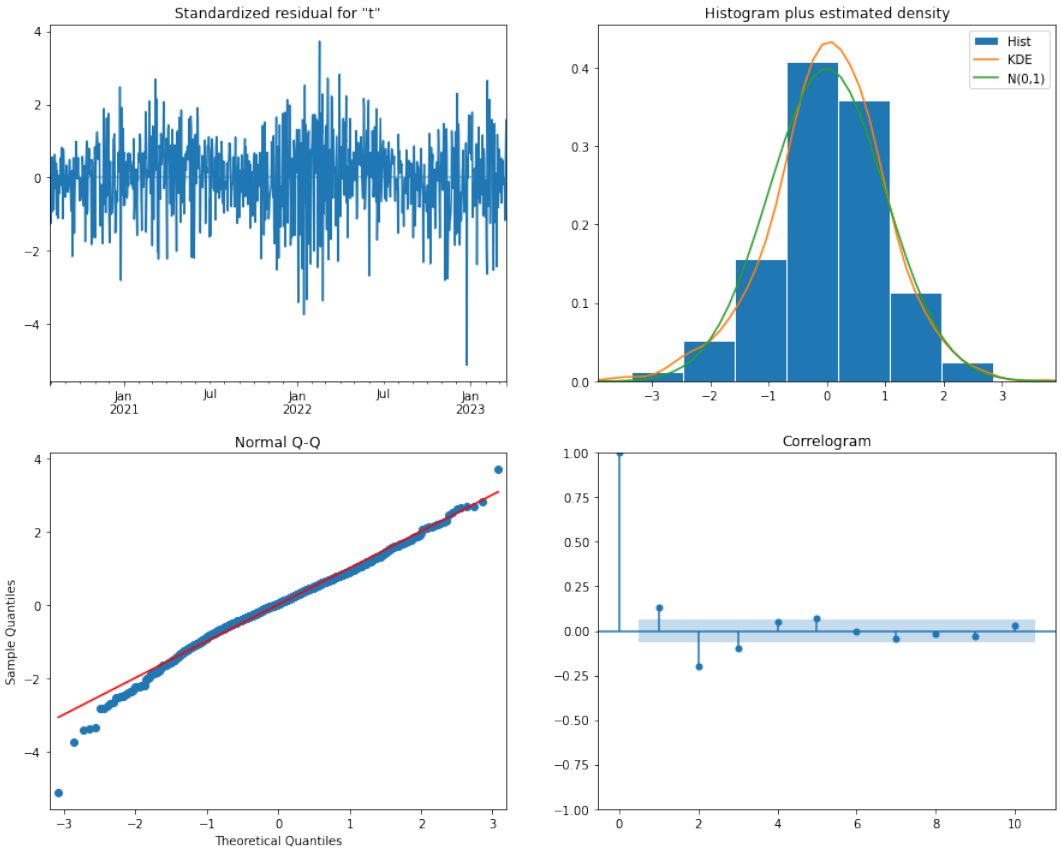

Figure 19 – Analyse statistique des résidus

Dans ce cas, nos diagnostics suggèrent que les résidus du modèle sont distribués suivant une loi normale sur la base de ce qui suit :

- Dans le graphique en haut à droite, nous voyons que la ligne rouge du KDE suit de près la ligne N(0,1) (où N(0,1) est la notation standard pour une distribution normale avec une moyenne de 0 et un écart type de 1). C'est une bonne indication que les résidus suivent bien une loi normale.
- Le graphique quantile - quantile en bas à gauche montre que la distribution ordonnée des résidus (points bleus) suit la tendance linéaire des échantillons prélevés dans une distribution normale standard N(0,1). Là encore, il s'agit d'une indication forte que les résidus sont distribués selon une loi normale.
- Les résidus dans le temps (graphique en haut à gauche) ne présentent pas de saisonnalité claire et semblent être un bruit blanc. On peut confirmer cela par le graphique d'autocorrélation (le corrélogramme) en bas à droite, qui montre que les résidus de la série temporelle ont une faible corrélation avec les versions retardées (shift à gauche) d'eux-mêmes.

Ces observations nous amènent à conclure que notre modèle s'ajuste de façon satisfaisante et qu'il pourrait nous aider à comprendre les données de notre série temporelle tout comme prévoir les valeurs futures.

3.2.3 • Prédiction

On réalise une prédiction non dynamique, c'est à dire qu'on utilise tout l'historique de données pour prédire la température du jour suivant.

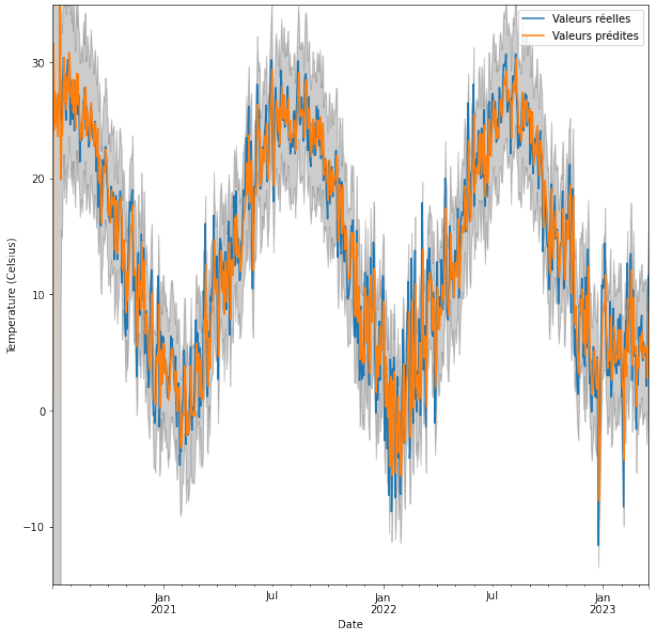

Figure 20 – Prédiction à un horizon de 1 jour

Les prédictions affichent également une saisonnalité de 365 jours.