Visualisation des prix des actions des entreprises du S&P 500 et du CAC40 et prédiction de la volatilité des rendements quotidiens : une approche automatique de la modélisation GARCH.

Alexandre Brunet

Décembre 2024

Introduction

Dans un environnement économique marqué par des fluctuations constantes, la volatilité des entreprises émerge comme un indicateur stratégique incontournable pour les investisseurs, les analystes financiers et les gestionnaires de risques. Maîtriser cette volatilité, la comprendre et pouvoir la prédire devient essentiel pour prendre des décisions éclairées et anticiper les risques avec précision.

Ce projet vise à offrir une solution innovante et pratique pour analyser et prévoir les prix et rendements des 500 plus grandes entreprises aux Etats-Unis et des 40 plus grandes entreprises françaises cotées en bourse. En développant une application intuitive, l'objectif est de fournir aux professionnels un outil performant pour générer des descriptions détaillées des mouvements boursiers sur une période spécifiée. L'application permet de suivre simultanément jusqu'à quatre entreprises, en observant à la fois l'évolution individuelle de leurs valeurs ainsi que les relations complexes entre leurs prix de clôture. Cette application va au-delà de la simple observation en permettant la prévision dynamique de la volatilité des rendements des actions. À partir d'une date choisie par l'utilisateur et d'un horizon temporel de 2 à 15 jours, elle offre une estimation précise de la volatilité à venir pour les entreprises sélectionnées, en tenant compte de la structure sous-jacente des données.

L'approche automatique intégrée dans l'application privilégie une sélection exhaustive et objective des modèles, minimisant ainsi les biais humains. Elle permet une exploration systématique et approfondie des paramètres, garantissant que les choix effectués reposent sur une analyse complète et rigoureuse des données.

Au cœur de cette modélisation se trouve le modèle $GARCH(p, q)$, développé par Tim Bollerslev dans son ouvrage fondateur *Generalized Autoregressive Conditional Heteroskedasticity* (1986), qui sera le sujet principal de cette documentation.

L'application a été développée en `Python` et déployée sur `Streamlit`. Elle se trouve ici : financevolatility.streamlit.app.

Contents

1 Origine et traitement des données

Pour ce projet, il n'y aucune nécessité que l'utilisateur importe des données. Toutes les données nécessaires, selon les choix de l'utilisateur, sont importées directement depuis le module `yfinance`. Plus précisément, elles proviennent de Yahoo Finance, qui est une plateforme financière en ligne offrant des informations notamment sur les matières premières, les devises, les actions, etc. La récupération peut se faire très facilement grâce à l'API de Yahoo Finance et les télécharge dans un format facilement gérable par `Python`. La chose à faire est de saisir les informations nécessaires. Par exemple, si l'utilisateur souhaite analyser les actions d'Apple entre le 01/01/2023 et le 01/01/2024, le code prend la forme générale comme suit :

```
import yfinance as yf
data = yf.download("AAPL", start="2023-01-01", end="2024-01-01")
```

Dans le cas où il souhaiterait effectuer des prédictions sur la volatilité des actions Apple à partir du 01/01/2024, la date de début sera fixée automatiquement 1 an et 6 mois avant celle qu'il a choisi. Ce choix repose sur deux éléments : d'abord, le modèle GARCH est un modèle de court terme, donc il n'est pas forcément pertinent d'entraîner le modèle sur une grande quantité de données. La seconde raison tient à la lourdeur de l'algorithme qui peut faire saturer le serveur de Streamlit qui possède des ressources limitées, surtout si l'utilisateur sélectionne plusieurs entreprises. C'est également pour cette raison qu'il ne peut sélectionner que 4 entreprises au plus.

Après le téléchargement des données selon les besoins de l'utilisateur, il est fréquent que certaines données manquent pour certains jours de la période sélectionnée. Un modèle ne peut pas correctement s'entraîner avec des dates manquantes. Pour combler ces manques, une interpolation polynomiale quadratique est effectuée. Ce type d'interpolation se situe entre l'interpolation linéaire, qui est simple mais peu réaliste dans le contexte d'un marché boursier, et l'interpolation cubique, qui est très précise pour modéliser des variations complexes mais coûteuse en termes de calcul et susceptible de biaiser les données en raison de son instabilité, notamment la création d'oscillations artificielles.

2 Visualisation et exploration des données financières

Dans le cadre de ce projet, cette partie est conscarée à fournir des sorties graphiques dynamiques, qui s'adaptent en fonction des choix de l'utilisateur (période et entreprises sélectionnées). Par conséquent, aucune analyse des données au sens strict du terme n'est effectuée.

Ces sorties graphiques portent principalement sur les prix de clôture et les rendements quotidiens (performance), avec une attention moindre accordée à la volatilité (risque). Une autre visualisation synthétise les informations sur les prix d'ouverture

et de clôture pour l'ensemble de la période sélectionnée, sous forme de graphiques en chandelles interactifs.

Enfin, les relations entre les prix de clôture des différentes entreprises sont mises en évidence à la fois visuellement, notamment par une droite de régression linéaire robuste aux valeurs aberrantes, et statistiquement. Pour quantifier ces relations, le coefficient de corrélation de Pearson est utilisé. Il est défini comme suit :

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$$

où $\mathrm{Cov}(X,Y)$ représente la covariance entre X et Y, et σ_X, σ_Y sont les écarts-types respectifs de X et Y et $\rho(X,Y)$ le coefficient de corrélation linéaire entre les deux séries contenant les prix de clôture de deux entreprises, notées X et Y.

3 Le modèle GARCH : une modélisation forte mais stricte de la volatilité passé

Comme expliqué lors de l'introduction, pour chaque entreprise, un modèle GARCH est utilisé pour effectuer des prévisions sur un certain horizon temporel. Bollerslev (1986) décrit ce modèle comme étant utile pour modéliser les séries finacières où la volatilité n'est pas constante mais évolue au fil du temps de manière dynamique. Ainsi, la la variance conditionnelle est modélisée (ou la volatilité dans le cas financier) à partir des erreurs et de la volatilité passée. Basiquement le modèle GARCH(p,q) se formalise comme suit :

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

Avec

$$\epsilon_t = \sigma_t z_t = y_t - \mu$$
$$\epsilon_t \mid \Omega_{t-i} \sim \mathcal{N}(0, \sigma_t^2)$$

Où :

- y_t est le rendement observé au temps au t,
- μ est la moyenne des rendements, supposée constante,
- ϵ_t correspond au résidu au temps t. L'ensemble des résidus sont *i.i.d.*,
- Ω_{t-i} est l'ensemble de l'information disponible jusqu'à l'instant $t-1$,
- σ_t^2 représente est la variance conditionnelle au temps t,
- z_t est un processus de bruit blanc puissant *i.i.d.* d'espérence nulle et de variance 1,
- ω est une constante strictement positive,

- Les coefficients α_i et β_j mesurent respectivement l'impact des résidus passés et de la volatilité passée.

Dans le cas de projet p et q seront fixés au minimum à 1 pour représenter le modèle le plus simple possible. Au maximum de sa complexité, p et q seront fixés à 10. À partir de ces informations, les hypothèses statistiques du modèle GARCH apparaissent : normalité des résidus, indépendance des résidus du modèle, et stationnarité conditionnelle. Deux autres hypothèses peuvent être ajoutées : l'indépendance des résidus au carré et l'homogénéité conditionnelle. Cette dernière implique que la variance conditionnelle dépend uniquement des informations passées et reste cohérente à travers les différentes sous-populations temporelles, sous réserve des paramètres estimés.

Ces hypothèses garantissent la validité des inférences tirées du modèle et permettent de s'assurer que les dynamiques de volatilité sont correctement représentées par la structure conditionnelle. Cependant, dans le contexte des séries financières, ces hypothèses peuvent s'avérer très restrictives. Si des violations sont détectées, des ajustements peuvent être effectués. Par exemple, en cas d'autocorrélation des résidus (au carré ou non), la spécification du modèle peut être modifiée pour inclure des termes supplémentaires dans la moyenne. Si la normalité des résidus n'est pas respectée, leur distribution a *posteriori* peut être étudiée, et une distribution alternative (comme la distribution t de Student ou GED) peut être utilisée pour modéliser les erreurs. Dans cette optique, ce projet propose une approche itérative et automatique dans le choix du meilleur modèle GARCH tout en tenant compte des violations potentielles des hypothèses statistiques.

4 Sélection du meilleur modèle GARCH

Comme expliqué dans la première section, le choix du modèle GARCH est réalisé par un processus itératif et automatisé. Voici les étapes principales de ce processus, pour l'ensemble des données de chaque entreprise :

1. **Recherche initiale :** Sélection exhaustive du meilleur modèle GARCH selon un critère d'information, sur la base d'apprentissage (2/3 des données les plus anciennes) avec le modèle par défaut :
   - Moyenne constante (ou nulle si le test de Student confirme que la moyenne de la série est égale à 0) ;
   - Distribution gaussienne (ou normale) des erreurs.
2. **Validation des hypothèses :** Vérification du respect des hypothèses du modèle par défaut, notamment l'indépendance et la normalité des résidus.
3. **Révision du modèle :** Ajustement de la moyenne et/ou de la distribution des erreurs en fonction des hypothèses violées.

4. **Nouvelle recherche :** Reprise de la recherche exhaustive du meilleur modèle GARCH, en tenant compte des ajustements.
5. **Application finale :** Utilisation du modèle GARCH retenu à l'issue de ce processus pour des prédiction :
    - Prévisions glissante sur la base de validation (le tiers des données restantes) pour visualiser comment le modèle se généralise sur de nouvelles données ;
    - Prédiction sur l'horizon temporelle choisit par l'utilisateur.

4.1 Une sélection exhaustive par le critère d'information

Pour automatiser la sélection du modèle, le critère d'information est un outil très précieux. Il évalue les caractéristiques d'un modèle, notamment la valeur de la vraisemblance, qui mesure dans quelle mesure les données observées sont compatibles avec un ensemble de paramètres du modèle. Deux critères d'information sont fréquemment utilisés : l'AIC (*Akaike Information Criterion*) et le BIC (*Bayesian Information Criterion*), définis comme suit :

$$\text{AIC} = -2\ln(\hat{L}) + 2k$$

$$\text{BIC} = -2\ln(\hat{L}) + k\ln(n)$$

Avec $\hat{L}$ la valeur de la vraisemblance maximisée du modèle, k le nombre de paramètre estimés dans le modèle et n le nombre d'observation. En général, plus la valeur du critère est faible, mieux le modèle s'ajuste aux observations. Dans le cadre de ce projet, et en raison de la complexité des données financières, l'AIC est privilégié. En effet, le BIC, qui favorise les modèles les plus simples, risque de sélectionner un modèle excessivement simplifié. Un tel modèle pourrait produire des prédictions stationnaires, ce qui n'est pas toujours adapté pour des données financières très volatiles.

Comme expliqué dans la première section, dans le cadre de ce projet, le modèle le plus simple est un GARCH(1,1), et le modèle le plus complexe est un GARCH(9,9). La recherche explore toutes les combinaisons possibles des paramètres p et q, chacun variant dans l'intervalle $[1, 9]$. Cela signifie qu'au total, 81 modèles GARCH sont évalués ($9 \cdot 9 = 81$).

4.2 Les recours en cas de violation des hypothèses du modèle GARCH

4.2.1 Les hypothèses statistiques et méthodes de validation

Le modèle GARCH repose sur des hypothèses statistiques strictes qui doivent être validées pour garantir la pertinence de ses résultats. Cependant, il est fréquent que certaines de ces hypothèses soient violées, nécessitant des ajustements. Les hypothèses principales et leurs méthodes de validation sont décrites ci-dessous.

Parmi ces hypothèses, la **normalité des résidus** spécifie que les résidus du modèle doivent suivre une distribution gaussienne, c'est-à-dire que leur moyenne, conditionnellement aux informations passées, doit être nulle et leur variance doit correspondre à la variance conditionnelle. Cette hypothèse est testée à l'aide du test de Shapiro-Wilk. Si la *p-value* de ce test est inférieure à 0,05, les résidus ne suivent probablement pas une distribution normale, ce qui peut être dû à une asymétrie ou à des queues épaisses. Ces cas seront explorés plus en détail par la suite.

Une autre hypothèse qui garantit la stabilité des coefficients et donc du modèle est la **stationnarité conditionnelle**. Elle se révèle être sûrement très importante car mentionnée de nombreuses fois par Tim Bollerslev (1986). Elle garantit la convergence de la variance conditionnelle et la cohérence des prédictions. Elle se vérifie par la somme des valeurs coefficients, après estimation, qui doit être strictement inférieure à 1 :

$$\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$$

.

Deux autres hypothèses similaires sont **l'absence d'autocorrélation dans les résidus (et leurs carrés)**. Elles sont vérifiées à l'aide du test de Ljung-Box pour k retards, avec $k = 7$ dans ce projet. Si la *p-value* du test est inférieure à 0,05 pour un retard donné, l'autocorrélation est significative. L'hypothèse est validée uniquement si les 7 retards sont indépendants, tant pour les résidus que pour leurs carrés.

La dernière hypothèse est sans doute la plus importante : **l'homogénéité conditionnelle**, ou l'absence d'un « effet arch ». Elle est testée par le test du multiplicateur de Lagrange appliqué aux résidus. Si la *p-value* est inférieure à 0,05, cela suggère une variance non homogène, ce qui remet en question la validité du modèle.

L'étude des résidus au carré permet d'identifier des corrélations résiduelles dans la variance non expliquées par le modèle. Une telle corrélation indique que la volatilité est imparfaitement modélisée, même si l'hétéroscédasticité conditionnelle semble respectée. En combinant cette analyse avec les autres tests, il est possible d'affiner le diagnostic et d'améliorer la précision du modèle.

Face à ces problèmes de non-validation, des recours sont possibles pour au moins diminuer leurs impacts négatifs : changement de moyenne et/ou de distribution d'erreur en réponse à l'autocorrlation des résidus (au carré ou non) et à leur non-normalité.

4.2.2 Ajout de retards supplémentaires pour palier aux violations des hypothèses d'indépendance des résidus

Conernant la moyenne, 4 sont disponibles grâce au module `arch` : `Constant`, `Zero`, `AR` ou `HAR`. Dans le cas où aucune autocorrélation des résidus n'est détectée par, la moyenne peut être `Constant` ou `Zero`. Si la moyenne est `Constant`, alors la moyenne

est estimée automatiquement par la fonction `arch_model` et si elle est fixée à `Zero`, alors aucune estimation n'est et assume que la moyenne des rendements est égale à 0. Dans ce projet, le choix entre l'un ou l'autre repose sur un test de Student qui confirme ou non si la moyenne des rendements est égale à 0. Dans le cas où une autocorrélation des résidus est détectée, la moyenne est `AR` si les résidus au carré sont indépendants, sinon la moyenne est `HAR`.

Ferenstein et Gasowski (2004) ont étudié le modèle AR-GARCH, qui est la résultante de plusieurs contributions au fil du temps dont celle de Tim Bollerslev en 1986. Ce modèle ajoute la composante auto-régressive AR pour capturer les dépendances linéaires dans les données. Considérons X_t comme étant la valeur de la série au temps t, ϕ_k les coefficients auto-régressifs et ε_t le terme d'erreur ayant les mêmes caractéristiques que le terme d'erreur du modèle GARCH de Bollerslev (1986). La composante `AR` se formule donc ainsi :

$$X_t = \sum_{i=1}^{k} \phi_k X_{t-1} + \varepsilon_t$$

Le choix de k, ou plus précisément le nombre de retards, peut se faire grâce aux critères d'information (cf. AIC et BIC). Dans ce projet, ce sera l'AIC qui sera privilégier de nouveau, pour les mêmes raisons expliquées dans la section 4.1 et aussi pour garder une cohérence dans la sélection du modèle. Pour équilibrer le modèle entre complexité (risque de sur-apprentissage) et simpliciter (risque de sous-apprentissage), la valeur de k peut prendre une valeur dans l'intervalle [1,7].

Selon Ferenstein et Gasowski (2004), l'ajout de la composante `AR` permet une précision accrue en cas de distribution à queues lourdes et en cas d'autocorrélation des résidus, mais modélise la variance conditionnelle de manière symétrique, ignorant les effets de levier.

Corsi *et al.* (2005) propose un modèle HAR-GARCH avec la moyenne `HAR` comme une fonction linéaire des composantes journalières, hebdomadaires et mensuelles. Clements et Preve (2021) ont repris cette proposition, l'ont noté RV_t et l'expriment ainsi :

$$RV_t = \beta_0 + \beta_1 RV_{t-1}^d + \beta_2 RV_{t-1}^w + \beta_3 RV_{t-1}^m + u_t$$

Avec :

- Le vecteur β contenant les paramètres estimés par la méthode des moindres carrés,
- $RV_{t-1}^d = RV_{t-1}$,
- $RV_{t-1}^w = \frac{1}{5}\sum_{i=1}^{5} RV_{t-1}$,
- $RV_{t-1}^m = \frac{1}{22}\sum_{i=1}^{22} RV_{t-1}$,
- u_t étant le terme d'erreur gaussien et homoscédastique.

La solution au problème de minimisation par la méthode des moindre carré se formule ainsi pour le vecteur β, sachant les observations $RV_1, ..., RV_n$:

$$\min_{\beta_0,\beta_1,\beta_2,\beta_3} \sum_{t=23}^{n} \left(RV_t - \beta_0 - \beta_1 RV_{t-1}^d - \beta_2 RV_{t-1}^w - \beta_3 RV_{t-1}^m \right)^2$$

Selon Clements et Preve (2021), le modèle HAR-GARCH intègre des composantes temporelles multiples (journalières, hebdomadaires, mensuelles), ce qui lui permet de capturer efficacement les dynamiques de mémoire à long terme. Toutefois, il a été démontré que les innovations des modèles HAR classiques ne suivent pas une distribution gaussienne et présentent une volatilité variable non prise en compte. L'utilisation d'une distribution flexible au modèle GARCH, permette de mieux modéliser ces propriétés.

4.2.3 Changement de la distribution d'erreur : une alternative à la loi gaussienne

Le choix de la distribution d'erreur se fait à partir de l'hypothèse de la normalité des résidus dont la fonction de distribution $f_g(\varepsilon_t)$ s'exprime ainsi :

$$f_g(\varepsilon_t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{\varepsilon_t^2}{2\sigma_t^2} \right)$$

où σ_t^2 la variance conditionnelle et l'espérance de 0.

En effet, si cette hypothèse est violée, cela signifie que les résidus présentent une asymétrie et/ou des queues lourdes. La distribution normale, notée `normal` en Python, est alors inadaptée et donc peut sous-estimer les risques extrêmes.

Pour évaluer la pertinence de la loi normale, des mesures comme la skewness (asymétrie) et le kurtose (aplatissement) sont utilisées. Dans une distribution gaussienne, la skewness et le kurtose sont nulles ou proche de zéro, traduisant une symétrie parfaite. Lorsque ces valeurs diffèrent significativement de ces références, une autre distribution devient nécessaire. Les caractéristiques des distributions ayant la skewness et le kurtose sont résumés dans les figures 1 et 2.

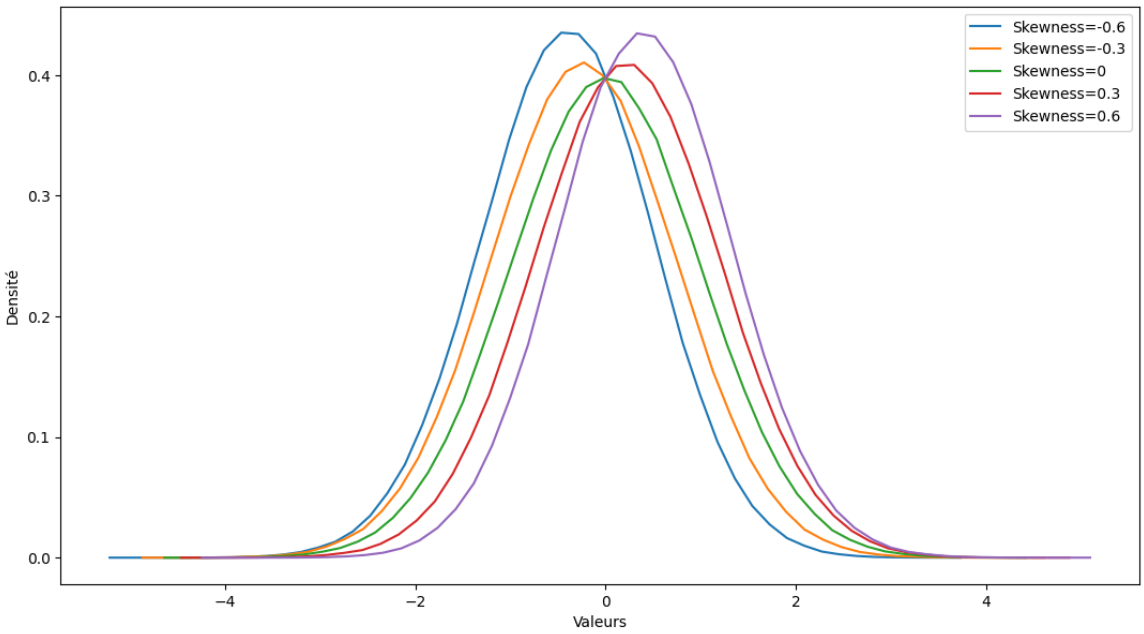

Figure 1: Distributions avec skewness modifiée

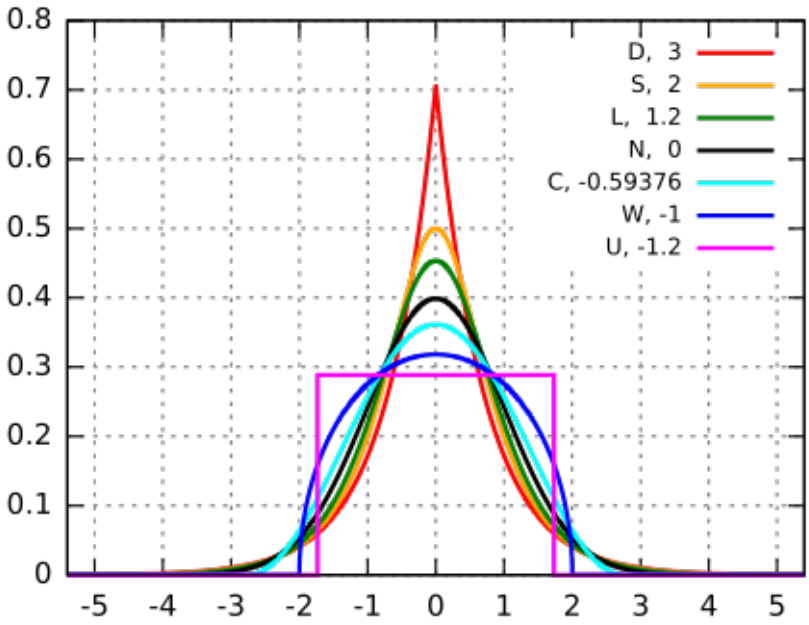

Figure 2: Distributions avec kurtose modifiée

Ampadu *et al.* (2024) ont comparé les performances de plusieurs distributions, notamment la gaussienne (`normal` en Python), la *Generalized Error Distribution* (`GED`), la t-Student (`t`), la skewed t-Student (`skewt`) et la *Skewed Generalized Error Distribution* (SGED). Bien que la SGED ait montré des performances supérieures dans leurs études, elle n'est malheureusement pas disponible dans le module `arch`. Nous nous concentrons donc sur les alternatives disponibles. Lorsque les résidus présentent principalement des queues épaisses mais peu ou pas d'asymétrie, la distribution t-Student constitue un choix approprié. Sa fonction de densité est donnée par :

$$f_{st}(\varepsilon_t) = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\sqrt{\pi(\eta-2)}\Gamma\left(\frac{\eta}{2}\right)} \left(1 + \frac{\varepsilon_t^2}{\eta-2}\right)^{-\frac{\eta+1}{2}}$$

où $\eta > 2$ est le degré de liberté et $\Gamma(\cdot)$ la fonction gamma. Toutefois, à mesure que ν augmente, la t-Student converge vers la loi normale, réduisant ainsi sa capacité à capturer les queues épaisses.

Pour des résidus présentant des queues peu épaisses ou symétrie, la GED peut être utilisée. Sa flexibilité repose sur le paramètre η, qui contrôle l'épaisseur des queues :

$$f_{ged}(\varepsilon_t) = \frac{\eta}{2^{1+1/\eta}\Gamma(1/\eta)\lambda} \exp\left(-\frac{|\varepsilon_t|^\eta}{\lambda^\eta}\right)$$

où $\lambda = \sqrt{\frac{2^{-2/\eta}\Gamma(1/\eta)}{\Gamma(3/\eta)}}$. Pour $\eta = 2$, la GED se réduit à la loi gaussienne, mais elle capture efficacement des queues épaisses lorsque $\eta < 2$. Enfin, si les résidus présentent à la fois une forte asymétrie et des queues épaisses, la distribution skewed t-Student constitue une option robuste, mais dont l'estimation est complexe et nécessite du temps supplémentaire pour son implémentation. Sa fonction de densité est exprimée comme suit :

$$f_{skt}(\varepsilon_t) = \begin{cases} \frac{bc}{\left(1+\frac{1}{\eta-2}\left(\frac{a+b\varepsilon_t}{1-\lambda}\right)^2\right)^{\frac{\eta+1}{2}}}, & \text{si } \varepsilon_t < -\frac{a}{b} \\ \frac{bc}{\left(1+\frac{1}{\eta-2}\left(\frac{a+b\varepsilon_t}{1+\lambda}\right)^2\right)^{\frac{\eta+1}{2}}}, & \text{si } \varepsilon_t \geq -\frac{a}{b} \end{cases}$$

où les paramètres asymétriques sont donnés par : $a = 4\lambda c\left(\frac{\eta-2}{\eta-1}\right)$, $b^2 = 1+3\lambda^2-a^2$, $c = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\sqrt{\pi(\eta-2)}\Gamma\left(\frac{\eta}{2}\right)}$ et où η représente le degré de liberté.

4.2.4 Choix de la moyenne et de la distribution d'erreur : résumé sous Python

Concernant les intervalles d'acceptabilité de skewness et du kurtose sont respectivement $]-0,3;0,3[$ et $]-0,6;1,1[$. Ces choix se font sur la base des observations

simulées (figure 1) et réelles (figure 2).

Ce qui suit résume les choix possibles pour les moyennes et les distributions d'erreur dans le cadre de violations d'hypothèses. L'extrait de code Python présenté est directement dérivé de celui utilisé pour l'application, mais a été légèrement simplifié afin de rendre la logique plus accessible.

```
## Détermination de la moyenne

if autocorr_resid == 1:  # Autocorrélation des résidus
    if autocorr_resid_squared == 0: # Indépendance des résidus au carré
        mean = 'AR'
    elif autocorr_resid_squared == 1: # Autocorrélation des résidus au carré
        mean = 'HAR'
else:  # Pas d'autocorrélation des résidus
    if p_value_ttest >= 0.05:  # Test de moyenne nulle (test de Student)
        mean = 'Zero'
    else:
        mean = 'Constant'

## Détermination de la distribution d'erreur

if normal_pvalue > 0.01: # Seuil fixé à 1% : tolérance plus élevée
    dist = 'normal'
else:
    if (kurtosis >= 1.1 or kurtosis <= -0.6) and abs(skewness) >= 0.3:
        dist='skewt' # Queues épaisses et asymétrie élevée
    elif (kurtosis >= 1.1 or kurtosis <= -0.6) and abs(skewness) < 0.3:
        dist = 't' # Queues épaisses sans asymétrie significative
    elif (kurtosis <= 1.1 and kurtosis >= -0.6) and abs(skewness) >= 0.3:
        dist = 'skewt' # Asymétrie élevée sans queues épaisses marquées
    else:
        dist='ged' # Faibles queues épaisses et asymétrie faible
```

5 Résultats : évaluation des modèles et prédictions

L'évaluation des modèles et des prévisions de volatilité est au cœur de cette application, permettant d'analyser de manière dynamique les rendements et la volatilité des entreprises cotées en bourse. Pour cela, deux types de prévisions sont réalisées : les prévisions glissantes et les prévisions futures sur un horizon spécifique.

5.1 Prévisions glissantes

Les prévisions glissantes (ou rolling predictions) sont une méthode clé dans l'application pour prévoir la volatilité à court terme. L'approche consiste à ajuster le modèle GARCH aux données d'entraînement et à choisir les hyperparamètres optimaux (p, q, moyenne et distribution) sur la base de ces données. Une fois que les hyperparamètres ont été déterminés, le modèle est ensuite entraîné sur une petite portion de la base de test, puis utilisé pour prédire la volatilité pour la période suivante.

Cette méthode est répétée pour chaque instant suivant, en ajustant les prévisions à chaque nouvelle donnée de la série temporelle. L'intérêt principal de cette approche est de pouvoir prédire la volatilité en prenant en compte les évolutions récentes du marché, tout en appliquant un modèle ajusté à la structure des données actuelles. En d'autres termes, chaque prédiction repose sur une fenêtre glissante d'historique, garantissant ainsi que le modèle s'adapte continuellement aux nouvelles informations disponibles.

Le graphique associé à ces prévisions présente deux courbes : la première, en bleu, représente les rendements réels sur la période de test, et la seconde, en rouge, illustre la volatilité prédite par le modèle GARCH. Cette comparaison visuelle permet de juger de la précision des prévisions par rapport aux données réelles, comme le montre la figure 2. Le modèle est jugé de bonne qualité lorsque un pic de la volatilité est observé après une forte variation dans les rendements réels.

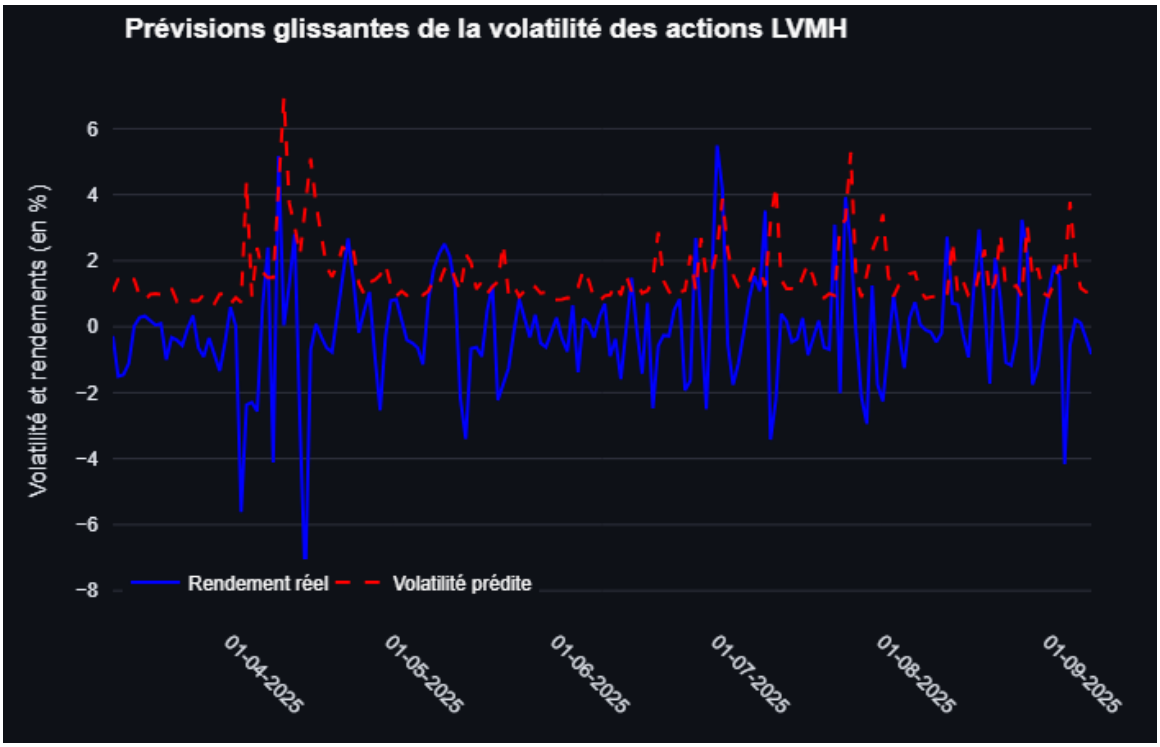

Figure 3: Prévisions glissantes de la volatilité des actions LVMH.

5.2 Prédictions sur le court terme

Une autre approche utilisée dans cette application est la prévision de la volatilité pour un horizon temporel spécifique, allant de 2 à 15 jours. Contrairement aux prévisions glissantes qui sont effectuées à chaque étape du temps, cette méthode prédit la volatilité future pour une période donnée, en ajustant le modèle GARCH sur les données passées. Le graphique met alors en évidence la volatilité estimée pour l'horizon choisi. Ce graphique permet à l'utilisateur d'obtenir une estimation visuelle de la volatilité à venir, en s'appuyant sur les modèles GARCH ajustés aux séries temporelles passées. Ainsi le résultat final se présente comme dans la figure 3.

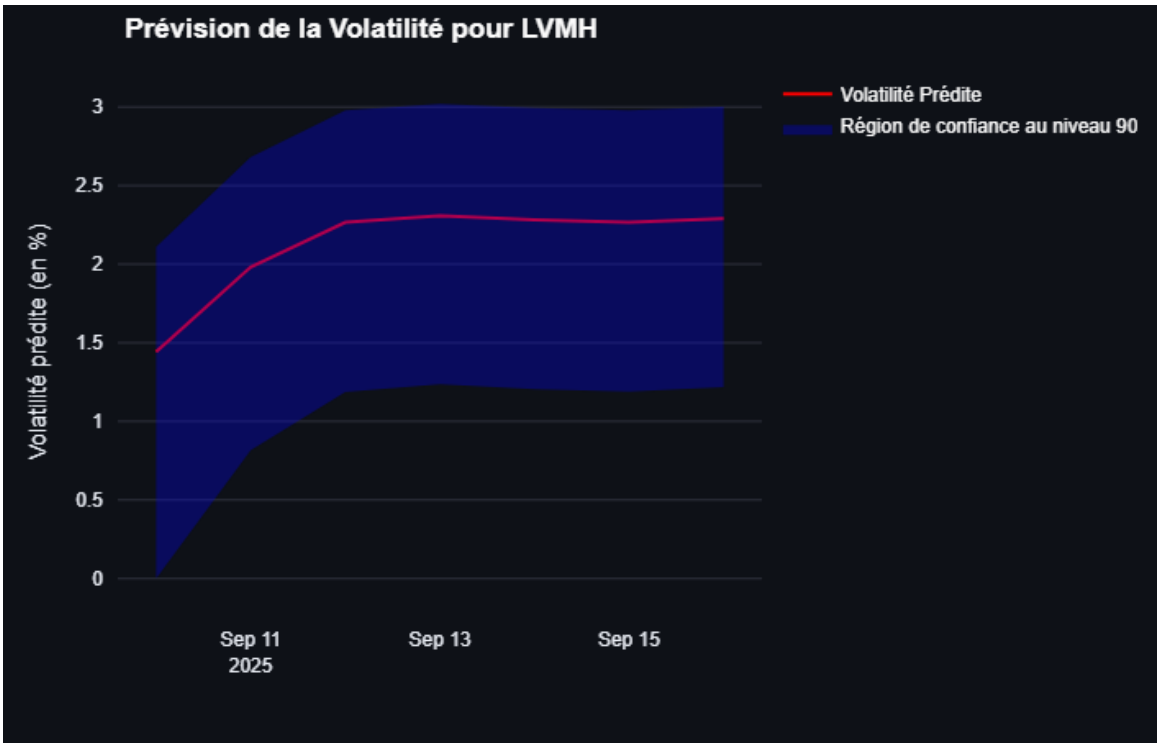

Figure 4: Prédiction de la volatilité des actions LVMH pour les 7 prochains jours.

6 Discussion

L'objectif principal de ce projet est d'appliquer un modèle GARCH pour prédire la volatilité des entreprises, en ajustant dynamiquement la moyenne et la distribution en cas de violations d'hypothèses. Cette approche entièrement automatisée présente l'avantage de permettre une sélection rapide et efficace des modèles tout en tenant compte de la structure des données, notamment les rendements réels et les résidus du modèle. Elle exploite un maximum d'informations disponibles pour établir une recherche exhaustive des ordre p et q, notamment grâce à l'utilisation de critères d'information.

Cependant, cette automatisation présente certaines limites. En premier lieu, le choix des seuils (seuil des p-value et différence de 0,3 avec les kurtoses et skewness "standards") pour détecter les violations d'hypothèses reste arbitraire et pourrait influencer les décisions de modélisation. De plus, l'absence d'une évaluation visuelle des modèles (cf. ACF et PACF), qui est impossible dans une approche totalement automatisée mais pourtant fondamentale, peut introduire des biais. Par exemple, une moyenne pourrait être privilégiée à tort au détriment d'une autre, entraînant ainsi des erreurs dans la prévision.

Indépendamment de l'approche automatique adoptée, le modèle GARCH lui-même peut s'avérer inadapté pour capturer toute la complexité des séries financières. En effet, bien que l'ajustement de la moyenne et de la distribution d'erreur permette d'obtenir des prédictions relativement fiables à un horizon de 3 ou 4 jours, cela ne garantit pas que le modèle soit suffisant pour modéliser toutes les dynamiques sous-jacentes. Cette limitation est fréquente dans le cadre des données financières, qui présentent souvent des comportements non linéaires et des chocs difficiles à modéliser.

La littérature propose de nombreux modèles dérivés du GARCH, tels que le modèle EGARCH (Exponential GARCH), conçu pour traiter des séries non-symétriques. Ce modèle offre une meilleure modélisation des chocs en supprimant l'hypothèse de symétrie de la variance conditionnelle, comme le mentionnent notamment Ferenstein et Gasowski (2004).

Dans le cadre de ce projet, les variantes plus avancées du modèle GARCH n'ont pas été explorées en raison de contraintes techniques et computationnelles. La recherche des paramètres optimaux pour le modèle GARCH de base étant déjà exigeante, l'intégration de modèles plus complexes aurait considérablement alourdi le processus.

7 Conclusion

Ce projet a permis de développer une solution simple et automatisée permettant de prédire la volatilité des entreprises cotées en bourse, en utilisant le modèle GARCH, qui reste un modèle de base pour étudier et prévoir la volatilité des séries temporelles. L'application développée vise à offrir un outil simple et rapide pour les professionnels, leur permettant de prendre en compte la structure des données et d'effectuer une sélection de modèles de manière objective, en minimisant les biais humains.

Cependant, bien que cette approche automatique offre de nombreux avantages, elle présente également certaines limites. Le choix des seuils pour la violation des hypothèses, par exemple, peut influencer les résultats, tout comme l'absence d'évaluation visuelle des modèles. De plus, bien que le modèle GARCH soit efficace dans de nombreux cas, il peut ne pas suffire à capturer toutes les dynamiques complexes des séries financières, comme l'indiquent les travaux sur des modèles dérivés tels que l'EGARCH.

Dans les perspectives futures, il serait intéressant d'enrichir cette approche en explorant d'autres modèles et en affinant la prise en compte des critères visuels. L'application pourrait également être étendue à un plus grand nombre d'entreprises, en prenant en compte des facteurs supplémentaires qui pourraient élargir le champ d'application de ce projet.

En somme, bien que ce projet apporte une première étape vers une analyse plus accessible de la volatilité des entreprises, il reste des pistes d'amélioration qui, avec le temps, permettront de rendre l'outil plus précis et plus adapté aux besoins du marché.

References

[1] Ampadu, Samuel, Eric T. Mensah, Eric N. Aidoo, Alexander Boateng, et Daniel Maposa. « A comparative study of error distributions in the GARCH model through a Monte Carlo simulation approach ». *Scientific African*, 23 (1 mars 2024): e01988. https://doi.org/10.1016/j.sciaf.2023.e01988.

[2] Bollerslev, Tim. « Generalized autoregressive conditional heteroskedasticity ». *Journal of Econometrics*, 31, n°3 (1 avril 1986): 307-27. https://doi.org/10.1016/0304-4076(86)90063-1.

[3] Clements, Adam, et Daniel P. A. Preve. « A Practical Guide to harnessing the HAR volatility model ». *Journal of Banking & Finance*, 133 (1 décembre 2021): 106285. https://doi.org/10.1016/j.jbankfin.2021.106285.

[4] Corsi, Fulvio, Stefan Mittnik, Christian Pigorsch, et Uta Pigorsch. « The Volatility of Realized Volatility ». *Econometric Reviews*, 2005. https://doi.org/10.1080/07474930701853616.

[5] Ferenstein, Elzbieta, et Mirosław Gasowski. « Modelling Stock Returns with AR-GARCH Processes ». *ResearchGate*, 2004. https://www.researchgate.net/publication/39430560_Modelling_stock_returns_with_AR-GARCH_processes.

[6] « Kurtosis ». In *Wikipédia*, 12 décembre 2023. https://fr.wikipedia.org/w/index.php?title=Kurtosis&oldid=210477241.