GEIST Louis

ENSAE 2e année
Stage d'application
Année scolaire 2022 - 2023

Prédiction de la volatilité de rendements financiers avec des variables explicatives de différentes fréquences

KIT - Karlsruher Institut für Technologie
Karlsruhe,
Deutschland

Maître de stage : **Rebekka Buse**
05/06/2023 - 15/09/2023

Remerciements

Je tiens à remercier Prof. Dr. Melanie SCHIENLE, directrice du département ECONSTAT du KIT, pour m'avoir permis de faire mon stage dans son département.

Je remercie ma superviseuse Dr. Rebekka BUSE, chercheuse, pour son écoute et ses aides précieuses.

Je remercie également Lotta RÜTER et Dr. Johannes BRACHER pour les échanges que j'ai eus avec eux, ainsi que l'ensemble de l'équipe.

Je remercie également Prof. Dr. Jean-Michel ZAKOÏAN, responsable du laboratoire Finance-Assurance au CREST, pour m'avoir mis en contact avec Prof. Dr. Melanie SCHIENLE.

Table des matières

Introduction

Ce rapport de stage retrace mon expérience estivale au *Karlsruher Institut für Technologie* (KIT) en Allemagne, où j'ai travaillé dans le département de méthodes statistiques et d'économétrie.

Le sujet était la modélisation et la prédiction en temps réel de la volatilité d'indices des marchés financiers, avec l'utilisation de variables explicatives de différentes fréquences d'actualisation. La prédiction de la volatilité est un enjeu, car la volatilité est une mesure du risque financier. A cette fin, j'ai utilisé le modèle GARCH-MIDAS (*Generalized AutoRegressive Conditional Heteroskedasticity - Mixed Data Sampling*). Le modèle GARCH(1,1) capture déjà de nombreux faits stylisés des séries financières. Le modèle GARCH-MIDAS ajoute la décomposition de la volatilité conditionnelle en produit de deux termes, de fréquence possiblement distincte.

L'objectif final était la construction d'une application R Shiny [2] pour visualiser les prédictions faites par les modèles GARCH-MIDAS. Durant mon stage, mon langage de programmation principal était le langage R. Je me suis aussi servi de GitHub Actions et du langage YAML pour l'automatisation des tâches.

J'ai pu m'appuyer sur un article de référence [3], qui a été la pierre angulaire pour ma démarche de recherche. Ce rapport détaillera les étapes de mon travail, les résultats obtenus et les enseignements tirés de cette expérience enrichissante.

Dans la première partie, nous présentons les données utilisées et quelques statistiques descriptives les concernant. Dans la deuxième partie, nous introduisons le modèle GARCH-MIDAS. Dans la troisième partie, nous présentons les résultats de prédictions ponctuelles de la volatilité journalière, de construction d'intervalles de confiance et comparons la précision des différents modèles.

Chapitre 1

Les données

1.1 Présentation des données

Les données utilisées comportent des indices financiers et des indices macroéconomiques.

1.1.1 Sources et récupération automatique

Les données proviennent de *Refinitiv Eikon* (pour le S&P 500) et de la *Federal Reserve Bank of St. Louis* (pour le reste des données). Toutes les quantités monétaires sont exprimées en dollars américains (USD). Outre le cas de la construction a posteriori de la volatilité dans la partie 3.1.1 avec des données intrajournalières, les indices financiers manipulés sont les valeurs quotidiennes à la fermeture du marché.

L'objectif final du projet étant la prédiction en temps réel, j'ai été amené à utiliser l'API disponible dans le package R *alfred* [10]. Pour planifier et exécuter la récupération automatique des données, j'ai mis en place un script YAML pour GitHub Actions.

1.1.2 Description des indices

Nous décrivons brièvement tous les indices qui sont utilisés dans notre étude. Ces indices économiques et financiers sont considérés pertinents pour la prédiction de la volatilité, voir [4].

Standard & Poor's 500 (S&P 500) (ou NASDAQ-100) Le S&P 500 est un indice boursier qui est composé des 500 plus grandes valorisations d'entreprises cotées aux Etats-Unis (qu'elles soient cotées au NYSE ou au NASDAQ). Le NASDAQ-100 est un autre indice boursier, qui est composé des 100 entreprises non financières étant les plus valorisées au

NASDAQ. On définit par **rendement** la transformation de la différence des logarithmes [1].

Les trois indices suivants sont des mesures quotidiennes des risques financiers.

RVOL22 La *RVOL22* est une mesure quotidienne rétrospective de la volatilité quotidienne de l'indice principal. Elle est définie comme la volatilité moyenne réalisée au cours des 22 jours précédents. En notant r le rendement de l'indice principal (NASDAQ-100), on pose $RVol(22)_t = \sqrt{\frac{1}{22}\sum_{k=0}^{21} r_{t-k}^2}$.

Volatility Index (VIX et VXNCLS) Le VIX est une mesure quotidienne prospective de la volatilité quotidienne du S&P 500. Cette mesure est définie comme la moyenne des volatilités annuelles sur les options d'achats et de ventes sur le S&P 500. Elle traduit ainsi la peur que les investisseurs ont dans le marché financier américain, car l'achat d'options permet à l'investisseur de se prémunir de certaines évolutions du marché. On considérera par la suite une transformation linéaire du VIX, qui correspond à la version convertie à un niveau quotidien : $\frac{VIX}{\sqrt{252}}$.

Le VXNCLS est la même mesure appliquée au NASDAQ-100 au lieu du S&P 500. Lorsque l'indice principal est le NASDAQ-100 et non le S&P 500, VIX désignera implicitement dans ce rapport le VXNCLS.

Volatility risk premium (VRP) La prime de risque associée à la volatilité est approchée par la différence suivante : $\frac{VIX}{\sqrt{252}} - RVol(22)$.

National Financial Conditions Index (NFCI) Le NFCI est un indice hebdomadaire de l'activité financière états-unienne. Il est défini comme la moyenne pondérée de 105 mesures de l'activité financière.

Housing starts (HOUST) Cet indicateur recense le nombre de nouveaux logements privés mis en chantier aux États-Unis. L'indicateur est mensuel. La série brute semble être une réalisation de processus intégrée à l'ordre 1 au sens de Granger. On considère donc sa transformation des différences de logarithmes.

Chicago Fed National Activity Index (NAI) Le NAI est un indice mensuel représentatif de l'activité économique états-unienne et des inflations liées.

1. En notant X_t la série brute et Z_t sa transformation stationnaire, on considère $Z_t = 100 * \log(\frac{X_t}{X_{t-1}})$. On multiplie par 100 pour des raisons de "stabilité numérique". (cf. [3])

Industrial production (IP) L'indice de production industrielle mesure mensuellement la production réelle des industries situées aux États-Unis. La série présentant clairement une tendance haussière, nous en considérons sa transformation des différences de logarithmes.

Indice / Série de données	Date de démarrage
S&P 500 (SPX)	05/01/1971
NASDAQ-100 (NDX)	02/10/1985
VIX	02/01/1990
VXNCLS	02/02/2001
RVOL22	03/02/1971
VRP	02/01/1990
NFCI	04/01/1971
HOUST	01/02/1959
NAI	01/02/1959
IP	01/02/1959

TABLE 1.1 – Disponibilités des cotations quotidiennes de à la fermeture

Dans la table 1.1, VIX, RVOL22 et VRP sont les indices se rapportant à la série du S&P 500. [2]

Les données intrajournalières du S&P 500 et du NASDAQ-100 proviennent de *Refinitiv Eikon*. Nous ne disposons de l'historique de ces données que depuis le 1^{er} juin 2023 pour les deux indices. Cependant, nous disposons des volatilités réalisées quotidiennes, calculées par l'estimateur indiqué en partie 3.1.1, du S&P 500 pour la période 2000 - 2019.

1.2 Statistiques descriptives

De manière usuelle en séries temporelles, nous travaillons sur des séries stationnaires [3]. Il s'agit donc ici de tester leur stationnarité et, dans le cas contraire, d'effectuer la transformation appropriée pour la rendre $I(0)$.

Nous souhaitons réaliser un test ADF (*Augmented Dickey-Fuller* test) sur la série temporelle de l'indice IP, qui teste la stationnarité de la série autour d'une constante. Le retard (paramètre de *lag*) qui permet d'avoir une absence d'autocorrélations des résidus

2. Nous utiliserons aussi les variables équivalentes pour le NASDAQ-100. Elles sont disponibles depuis des dates similaires.

3. Nous verrons effectivement dans la suite (cf. partie 2.2.2) que le modèle GARCH-MIDAS fait l'hypothèse que la série principale et les variables explicatives soient stationnaires.

est de 12. (Pour tester l'absence d'autocorrélation des résidus, nous réalisons un test de Ljung-Box jusqu'à l'ordre 24.) La p-valeur alors obtenue avec le test ADF en niveau est de 0.7 : nous ne pouvons pas rejeter l'hypothèse H_0 : *présence d'une racine unitaire* à n'importe quel niveau usuel. Nous réalisons alors un test KPSS en niveau pour pouvoir affirmer que la série est non-stationnaire : la p-valeur obtenue est inférieure à 0.01 et nous concluons donc sur la présence d'une racine unitaire.

Nous considérons alors la transformation suivante de la série : $Y_t = \log(X_t) - \log(X_{t-1})$ (où X_t est la série brute). En réalisant les mêmes tests que précédemment : la p-valeur du test ADF est inférieur à 0.01 : nous pouvons rejeter l'hypothèse nulle de présence d'une racine unitaire et concluons donc sur la stationnarité de cette série.

L'indice IP est représenté en figure 1.1. En vue du graphique de la série brute, nous aurions pu directement conclure sur la présence d'une racine unitaire.

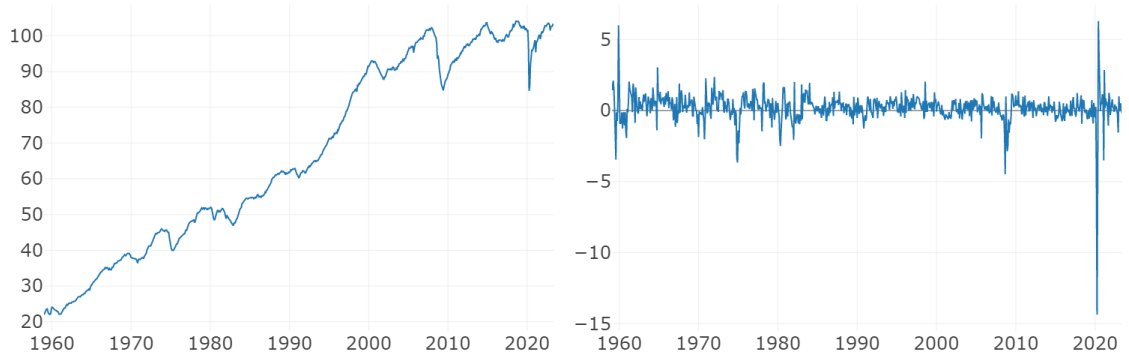


FIGURE 1.1 – Indice de production industrielle (IP) et sa transformation des différences de logarithmes en fonction du temps

L'indice IP est représenté en figure 1.1. En vue du graphique de la série brute, nous aurions pu directement conclure sur la présence d'une racine unitaire.

La représentation de la série brute de l'indice HOUST en figure 1.2 est plus difficile à interpréter quant à la présence d'une racine unitaire.

Le test ADF en niveau rejette la présence d'une racine unitaire et le test KPSS en niveau rejette la stationnarité de la série pour non différenciée.

Pour la série différenciée, le test ADF rejette la présence d'une racine unitaire et le test KPSS ne rejette pas la stationnarité de la série. Pour cette raison, nous utiliserons par la transformation de la série que nous désignerons par *dhoust*.

De la même manière avec un test ADF en niveau, nous avons pu rejeter la présence d'une racine unitaire au niveau de 1% pour les séries VIX, VRP, RVol22 (correspondant aux deux indices, S&P500 et NASDAQ-100), NFCI, NAI et enfin pour les transformations

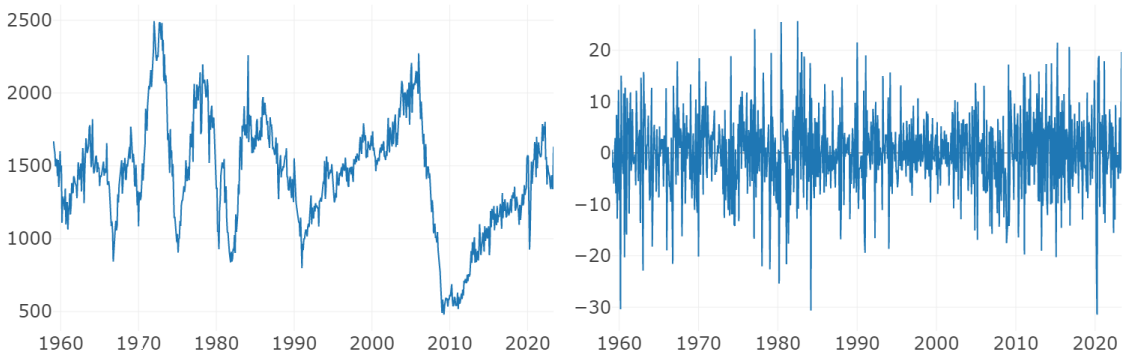


FIGURE 1.2 – Nombre de débuts de constructions de maisons (HOUST) et sa transformation des différences de logarithme en fonction du temps

des différences de logarithmes des séries du S&P 500 et du NASDAQ-100. A présent, nous ne travaillerons plus que sur les transformations de ces deux séries et omettrons de rappeler que nous considérons leur transformation.

Chapitre 2

Modélisation : le choix du modèle GARCH-MIDAS

Dans la première partie de ce chapitre, nous introduisons le modèle GARCH(1,1) puis le modèle GJR-GARCH qui constituent notre point de départ. Dans la seconde partie, nous exposons de ces modèles GARCH : le modèle GARCH-MIDAS, introduit par Engle et al. [6] en 2013. Cette extension permet la prise en compte d'une variable externe (voire deux variables externes) de fréquence d'actualisation éventuellement distincte de la fréquence de l'indice principal.

2.1 Rappels sur les modèles GARCH

2.1.1 Le modèle GARCH(1,1)

Un processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ suit un modèle **GARCH(1,1)** de paramètre $\theta = (\omega, \alpha, \beta)$ s'il existe un bruit blanc fort normalisé η tel que :

$$\forall t \in \mathbb{Z}, \varepsilon_t = \sigma_t \eta_t \text{ avec } \begin{cases} \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \\ \omega > 0,\ \alpha \geq 0,\ \beta \geq 0. \end{cases}$$

Le modèle GARCH(1,1) capture de nombreux faits stylisés des rendements financiers ; par exemple le regroupement de hautes volatilités autour de certaines dates. C'est un modèle phare des séries temporelles financières.

Le modèle GARCH(1,1) sera pris comme modèle de référence dans les comparaisons empiriques. Il s'agit d'un bon modèle de référence, car il est difficile de le battre, comme il a été montré dans [9].

2.1.2 Le modèle GJR-GARCH(1,1)

Un processus $(\varepsilon_t)_{t\in\mathbb{Z}}$ suit un modèle **GJR-GARCH(1,1,1)** de paramètre $\theta = (\omega, \alpha, \beta, \gamma)$ s'il existe un bruit blanc fort normalisé η tel que :

$$\forall t \in \mathbb{Z}, \varepsilon_t = \sigma_t \eta_t \text{ avec } \sigma_t^2 = \omega + (\alpha + \gamma \mathbb{1}_{\eta_{t-1}<0})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2.$$

Les lettres "GJR" signifient Glosten-Jagannathan-Runkle, auteurs de la publication [8]. Ce modèle est un des modèles GARCH qui permet de capter l'asymétrie de la loi des rendements financiers.

2.2 Le modèle GARCH-MIDAS

2.2.1 Définition formelle

Un processus $(\varepsilon_{i,t})_{t\in\mathbb{Z},i\in I_t}$ suit un modèle **GARCH-MIDAS** s'il existe un triplet $(\alpha, \beta, \gamma) \in \mathbb{R}^3$ et un bruit blanc $(Z_{i,t})_{t\in\mathbb{Z},i\in I_t}$ tels que, pour tout $t \in \mathbb{Z}$ et pour tout $i \in [1, I_t]$:

$$\frac{\varepsilon_{it}}{\sqrt{\tau_t}} = \sqrt{g_{it}} Z_{i,t}$$

avec :

1) $g_{i,t} = (1 - \alpha - \frac{\gamma}{2} - \beta) + (\alpha + \gamma \mathbb{1}_{\varepsilon_{i-1,t}<0}) \frac{\varepsilon_{i-1,t}^2}{\tau_t} + \beta g_{i-1,t}$,

2) τ_t est une fonction fixe d'un processus explicatif (X) de basse fréquence,

3) $[1, I_t]$, la liste des indices que peut prendre "i" pour la période t, avec $I_t \in \mathbb{N}\backslash\{0\}$

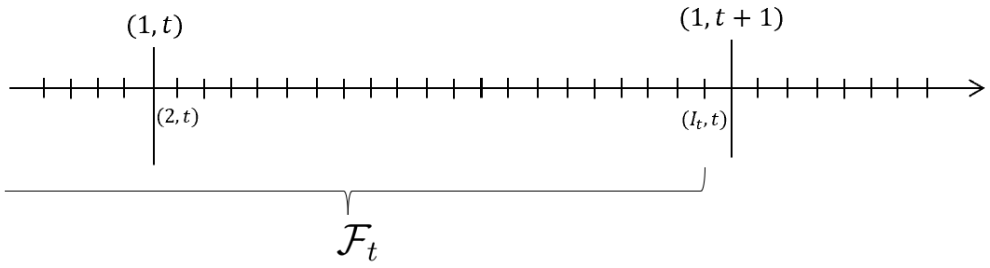


FIGURE 2.1 – Notation des indices temporels (i,t)

Pour faciliter les cohérences des formules, les couples d'indices $(I_t + 1, t)$ et $(1, t + 1)$ sont les mêmes. Par extension, $(0, t)$ et $(I_{t-1}, t - 1)$ sont aussi les mêmes. τ est $\mathcal{F}-$prédictible. Conformément à la figure 2.2, la valeur τ_{t+1} est donc disponible en théorie

à la date (I_t, t). En temps réel, les variables de long terme prennent du temps à être calculées après la fin de la période t. Ils sont donc disponibles au cours de la période $t+1$.

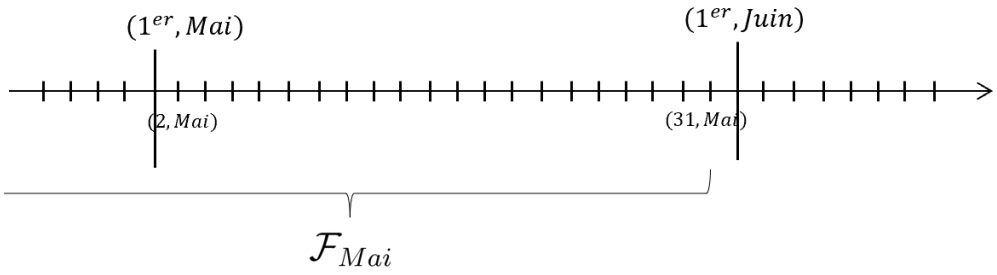


FIGURE 2.2 – Exemple de notation des indices temporels : cas mensuel

On appelle **variance conditionnelle** (ou volatilité) en (i, t) :

$$\sigma_{i,t}^2 = Var(\varepsilon_{i,t}|\mathcal{F}_{i-1,t}).$$

2.2.2 Hypothèses

Nous ferons les quatre hypothèses suivantes :

Hypothèse 2.2.1 (Bruit blanc fort normalisé à kurtosis fini) $(Z_{it})_{t\in\mathbb{Z},i\in I_t}$ *est indépendant et identiquement distribué (iid) avec* $\forall t \in \mathbb{N}, \forall i \in I_t, \mathbb{E}[Z_{it}] = 0, \mathbb{E}[Z_{it}^2] = 1$ *et* $1 < \kappa := \mathbb{E}[Z_{it}^4] < \infty$.

Hypothèse 2.2.2 (Condition de stationnarité) *Les paramètres du modèle satisfont :*

1) $\alpha > 0, \beta \geq 0, \alpha + \beta > 0$

2) $\alpha + \frac{\gamma}{2} + \beta < 1$

3) $(\alpha + \frac{\gamma}{2})^2\kappa + 2(\alpha + \frac{\gamma}{2})\beta + \beta^2 < 1$

Nous admettons que les hypothèses 2.2.1 et 2.2.2 permettent de montrer que $\left(\frac{\epsilon_{it}}{\sqrt{\tau_t}}\right)$ est un processus GJR-GARCH(1,1) stationnaire au second ordre.

Hypothèse 2.2.3 (Structure de la composante de long-terme) *Soit* $f > 0$ *une fonction mesurable et* $(X_t)_{t\in\mathbb{Z}}$ *un processus strictement stationnaire et ergodique, tels que*

$$\forall t \in \mathbb{N}, \tau_t = f((X_{t-k})_{k\in\mathbb{N}^*})$$

avec $\mathbb{E}[|X_t|^q] < \infty$ *et* q *assez grand pour avoir* $\mathbb{E}[\tau_t^2] < \infty$.

Le théorème ergodique permet de montrer avec cette dernière hypothèse que $(\tau_t)_{t\in\mathbb{Z}}$ est strictement stationnaire et ergodique. La finitude du moment d'ordre 2 de ce même processus donne la stationnarité au second ordre.

Hypothèse 2.2.4 (Indépendance) $\forall(j,t) \in \mathbb{Z}^2, \forall i \in I_{t-j}, X_t$ *est indépendant de* $Z_{i,t-j}$.

2.2.3 Interprétation

Tout d'abord, il s'agit d'expliciter la variance conditionnelle. En raison de l'hypothèse 2.2.3, τ est $(\mathcal{F}_t)_{t\in\mathbb{Z}}-$prévisible. Par la définition 2.2.1 du modèle GARCH-MIDAS, g est $(\mathcal{F})_{t\in\mathbb{Z},i\in I_t}-$prévisible. Or, le produit de deux fonctions mesurables est mesurable. Donc le produit $\tau_t g_{it}$ est $\mathcal{F}_{it}-$mesurable et comme Z_t est indépendant de $\mathcal{F}_{it}$ et $\mathbb{E}[Z_{it}^2] = 1$, on obtient : $\forall t \in \mathbb{Z}, \forall i \in [1, I_t]$,

$$\boxed{\sigma_{it}^2 = \tau_t g_{it}}$$

Le but du modèle GARCH-MIDAS est ainsi d'améliorer la modélisation de la volatilité par rapport au modèle classique GARCH(1,1) par deux composantes de fréquence d'actualisation possiblement différente. La composante de fréquence élevée d'actualisation est notée $(g)_{t\in\mathbb{Z},i\in I_t}$ et est aussi appelée composante de court terme. La composante de plus basse fréquence est $(\tau_t)_{t\in\mathbb{Z}}$ et est aussi appelée composante de long terme.

La fréquence élevée correspond dans notre cas à une période de un jour. La basse fréquence dépend de la variable explicative que nous utiliserons. Cela correspondra typiquement à des périodes d'un mois, une semaine ou un jour. Pour ce dernier cas, nous n'avons donc pas deux fréquences d'actualisations distinctes et pouvons alors abandonner l'indice i dans nos notations.

2.2.4 Composante de long terme de la volatilité

La composante de long terme de la volatilité est la variable τ, qui est une transformation de la variable explicative de basse fréquence notée X. Pour la suite du projet, nous utiliserons la fonction f définie de la manière suivante (cf. l'hypothèse 2.2.3) :

$$\forall t \in \mathbb{N}, \tau_t = f((X_{t-k})_{1\leq k\leq K}) = \exp(m + \theta \sum_{k=1}^{K} \varphi_k X_{t-k})$$

où :

— K est le nombre de retards de la variable X utilisés pour définir τ,

— m est un paramètre du modèle à estimer,

— θ est un paramètre du modèle à estimer,

— φ_k est le schéma de poids, qui est une fonction de deux paramètres, w_1 et w_2 qui sont à estimer.

Le schéma de poids est défini de la manière suivante : $\forall k \in [1, K]$

$$\varphi_k = \lambda \left[\left(\frac{k}{K+1} \right)^{w_1 - 1} \left(1 - \frac{k}{K+1} \right)^{w_2 - 1} \right]$$

où λ est choisi de telle sorte que $\sum_{k=1}^{K} \varphi_k = 1$.

Il est souvent préférable de contraindre la pondération à être décroissante en k, ce qui est la cas lorsque $w_1 = 1$. Dans ce cas, on appelle la pondération *beta restricted weighting scheme*. S'il n'y a pas de telle contrainte, on l'appelle *beta unrestricted weighting scheme*.

Nous pouvons également prendre en compte deux variables explicatives de long terme. En prenant X_1 une variable explicative quotidienne et X_2 une variable explicative de plus faible fréquence d'actualisation, on définit τ de manière analogue :

$$\forall t \in \mathbb{N}, \forall i \in I_t, \ \tau_{i,t} = \exp(m + \theta^{X_1} \sum_{k=1}^{K_1} \varphi_k^{X_1} X_{1;i-k,t} + \theta^{X_2} \sum_{k=1}^{K_2} \varphi_k^{X_2} X_{2;t-k}).$$

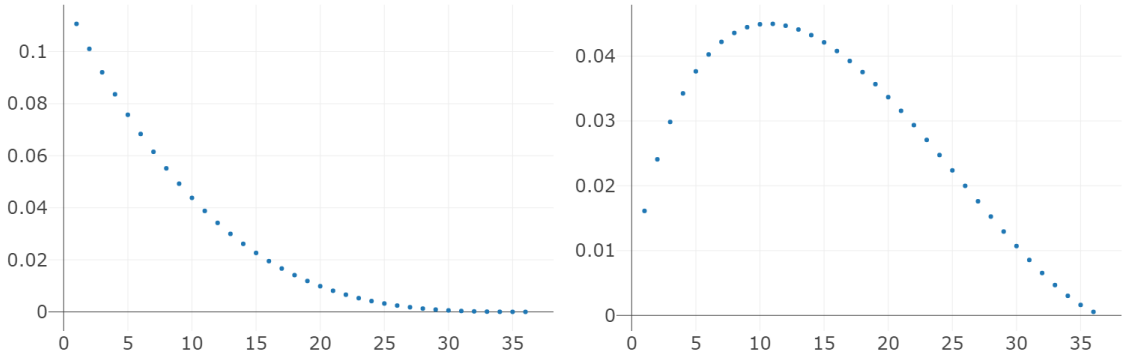


FIGURE 2.3 – Exemples de *beta weighting scheme* : *restricted* à gauche, *unrestricted* à droite – valeurs des poids φ_k en fonction du retard k pour les variables IP (gauche) et HOUST (droite), issus des estimations du 01/09/2023 des modélisations GARCH-MIDAS du S&P500 avec comme variable explicative respectivement IP et HOUST

On constate en figure 2.4 que la variable X, ici l'indice NFCI, est transformée en une variable τ plus lisse.

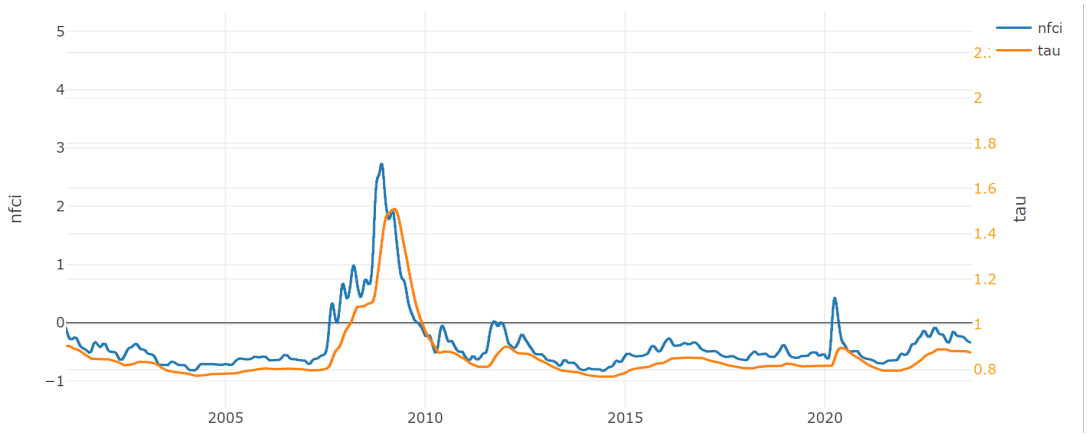


FIGURE 2.4 – Exemple de la composante de long-terme de la volatilité – $X = NFCI$ et sa transformation τ

2.2.5 Prédiction optimale

A) Prédiction optimale pour un modèle GARCH(1,1)

Nous appellerons **prédiction optimale en t à l'horizon k** : $h_{t+k|t} = \mathbb{E}[\sigma^2_{t+k}|\mathcal{F}_t]$. Pour rappel, le modèle est le suivant : $\exists\, \eta\ iid(0,1)$

$$\begin{cases} \sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \\ \epsilon_t = \eta_t\sigma_t. \end{cases}$$

Nous avons donc : $\sigma^2_{t+k} = \omega + \alpha\epsilon^2_{t+k-1} + \beta\sigma^2_{t+k-1}$.
Ainsi : $h_{t+k|t} = \omega + \alpha\mathbb{E}[\eta^2_{t+k-1}\sigma^2_{t+k-1}|\mathcal{F}_t] + \beta h_{t+k-1|t}$.
Par construction, η_{t+k-1} et σ_{t+k-1} sont indépendants. Et le second moment de η_{t+k-1} est égal à 1 (car c'est un bruit blanc normalisé par hypothèse).

Alors $h_{t+k|t} = \omega + (\alpha+\beta)h_{t+k-1|t}$.
Puisque σ est $\mathcal{F}_t-$prévisible, cela signifie que :

$$\boxed{\forall k \geq 1,\quad h_{t+k|t} = \omega\frac{1-(\alpha+\beta)^{k-1}}{1-(\alpha+\beta)} + (\alpha+\beta)^{k-1}\sigma^2_{t+1}}.$$

B) Prédiction optimale pour le modèle GARCH-MIDAS

Nous appellerons **prédiction optimale en t (à l'horizon k pour la période courte, s pour la période longue)** : $h_{k,t+s|t} = \mathbb{E}[\sigma^2_{k,t+s}|\mathcal{F}_t]$.

Hypothèse 2.2.5 *La loi des innovations $\mathcal{L}(Z)$ est telle que* $\mathbb{E}[Z^2\mathbb{1}_{Z>0}] = \mathbb{E}[Z^2\mathbb{1}_{Z<0}] = \frac{1}{2}$.

Composante de court terme g On a $g_{i,t} = (1-\alpha-\frac{\gamma}{2}-\beta)+(\alpha+\gamma\mathbb{1}_{\epsilon_{i-1,t}<0})g_{i-1,t}Z_{i-1,t}^2+\beta g_{i-1t}$.

Donc : $g_{i,t+1|t} = (1-\alpha-\frac{\gamma}{2}-\beta)+\mathbb{E}[(\alpha+\gamma\mathbb{1}_{Z_{i-1,t+1}<0})g_{i-1,t+1}Z_{i-1,t+1}^2|\mathcal{F}_t]+\beta g_{i-1,t+1|t}$.

Par construction du modèle, $Z_{i-1,t+1}$ et $g_{i-1,t+1}$ sont indépendants. Par hypothèse, $\mathbb{E}[Z^2\mathbb{1}_{Z<0}] = \frac{1}{2}$. Par conséquent, en notant $g_{i,t+s|t} = \mathbb{E}[g_{i,t+s}|\mathcal{F}_t]$,

$$g_{i,t+1|t} = (1-\delta) + \delta g_{i-1,t+1|t}$$

avec $\delta := \alpha + \frac{\gamma}{2} + \beta$.

C'est-à-dire : $g_{i,t+1|t} - 1 = \delta(g_{i-1,t+1|t} - 1)$.

Cette relation de récurrence permet de généraliser à la formule suivante : $\forall s > 1$,

$$g_{k,t+s|t} = 1 + \delta^{(I_{t+1}+\ldots+I_{t+s-1}+k-1)}(g_{1,t+1} - 1).$$

Composante de long terme τ Les valeurs futures de la composante de long terme seront alors approchées par la dernière valeur de τ disponible. Ainsi, en date t, pour tout $s > 1$, nous utiliserons τ_{t+1} pour τ_{t+s}. τ_{t+1} est bien disponible en t, car τ est $\mathcal{F}-$prévisible.

Indépendance des deux composantes Nous avons $X \perp\!\!\!\perp Z$ par l'hypothèse 2.2.4.

D'une part, τ est fonction de X, donc $\tau \perp\!\!\!\perp Z$.

D'autre part, g se réécrit sans τ : $g_{i,t} = (1-\delta)+(\alpha+\gamma\mathbb{1}_{\sqrt{\tau_t g_{i-1,t}}Z_{i-1,t}<0})g_{i-1,t}Z_{i-1,t}^2+\beta g_{i-1,t}$. Nous constatons que g_{it} est ainsi fonction de $Z_{i-1,t}$ et $g_{i-1,t}$. En raisonnant par récurrence, g_{it} est fonction de $\{Z_{i-j,t}|j \in \mathbb{N}^*\}$.

Finalement, le lemme des coalitions permet de conclure que pour tout $t \in \mathbb{Z}, i \in I_t$,

$$\tau_t \perp\!\!\!\perp g_{i,t}.$$

Conclusion L'indépendance montrée ci-dessus permet d'écrire l'espérance du produit comme le produit des espérances et d'obtenir la formule :

$$h_{k,t+s|t} = \mathbb{E}[\sigma_{k,t+s}^2|\mathcal{F}_t] = \mathbb{E}[\tau_{t+s}|\mathcal{F}_t]\mathbb{E}[g_{k,t|t+s}|\mathcal{F}_t].$$

Notre estimateur de la variance future est donc

$$\forall s \in \mathbb{N}^*, \boxed{h_{k,t+s|t} = \tau_{t+1}\big(1 + \delta^{(I_{t+1}+\ldots+I_{t+s-1}+k-1)}(g_{1,t+1} - 1)\big)}$$

Et en généralisant pour des prédictions dont la date de départ de prédiction est au sein d'une longue période :

$$\forall k \in \mathbb{N}^*,\ h_{l+k,t|l,t} = \tau_t\big(1 + \delta^{k-1}(g_{l+1,t} - 1)\big)$$

2.3 Estimation des modèles

2.3.1 Choix des paramètres

	Période d'actualisation	**K**	**Pondération *beta***
VIX	Journalière	3	restreinte
RVol22	Journalière	264	restreinte
VRP	Journalière	3	restreinte
NFCI	Hebdomadaire	52	restreinte
HOUST	Mensuelle	36	non restreinte
IP	Mensuelle	36	restreinte
NAI	Mensuelle	36	restreinte

TABLE 2.1 – Variables explicatives : caractéristiques et paramètres optés pour les modèles GARCH-MIDAS (K est le retard – une pondération *beta restreinte* est décroissante en le retard, une pondération *non restreinte* est croissante puis décroissante, cf. partie 2.2.4).

Le choix de ce paramétrage est repris du papier [3].

2.3.2 Quasi-maximum de vraisemblance

L'estimation du modèle est réalisée par la méthode du quasi-maximum de vraisemblance, que nous expliquons ici. Pour approximer la vraisemblance inconnue du fait de la non paramétrisation du bruit blanc, on remplace la densité inconnue du bruit blanc (processus noté (Z)) par la densité d'une loi normale centrée réduite. Pour l'estimation des paramètres, on maximise alors cette nouvelle quantité, qu'on appelle quasi-vraisemblance.

L'estimation des modèles GARCH-MIDAS a été réalisée à partir du package mfGARCH (cf. [11]). L'estimation du modèle GARCH(1,1) a été réalisée avec le package ruGARCH (cf. [7]).

Chapitre 3

Résultats empiriques

3.1 Prédictions ponctuelles et intervalles de confiance

3.1.1 Évaluer une prédiction de volatilité

On s'intéresse dans ce projet à la prédiction de la volatilité quotidienne. Or, la volatilité est une quantité théorique à laquelle on n'a pas accès puisque nous ne simulons pas nos données. Il n'existe pas de réalisation de la volatilité qu'on pourrait observer, comme il existe des réalisations des rendements par exemple. Pour évaluer notre modèle, nous devons alors trouver un estimateur de la volatilité qui pourrait être considéré comme la vraie valeur de la volatilité.

L'estimation de la volatilité quotidienne est faite en utilisant des données de fréquence plus élevée que quotidienne, en l'occurrence les cotations intrajournalières à 5 minutes d'intervalle de l'indice. En effet, [1] montre que l'estimateur ci-dessous de la volatilité en date t est bien convergent.

$$\hat{\sigma}_t^2 := \sum_{j \in \Pi} \left(\log \left(\frac{p_{j,t}}{p_{j-1,t}} \right) \right)^2$$

où p est la valeur de l'indice considéré et Π est un maillage régulier de période 5 minutes durant l'ouverture des marchés.

3.1.2 Les prédictions ponctuelles

Nous nommons **date d'origine** la date à partir de laquelle les prédictions sont réalisées. Autrement dit, la date d'origine est la dernière date de la période d'entraînement, qui correspond à toutes les données utilisées pour l'entraînement du modèle et la prédiction de la volatilité à un horizon quelconque. Elle est représentée en gris sur la figure 3.1 et les

figures qui suivent. La variable *real volatility* est la variable à prédire, elle a été définie en partie 3.1.1.

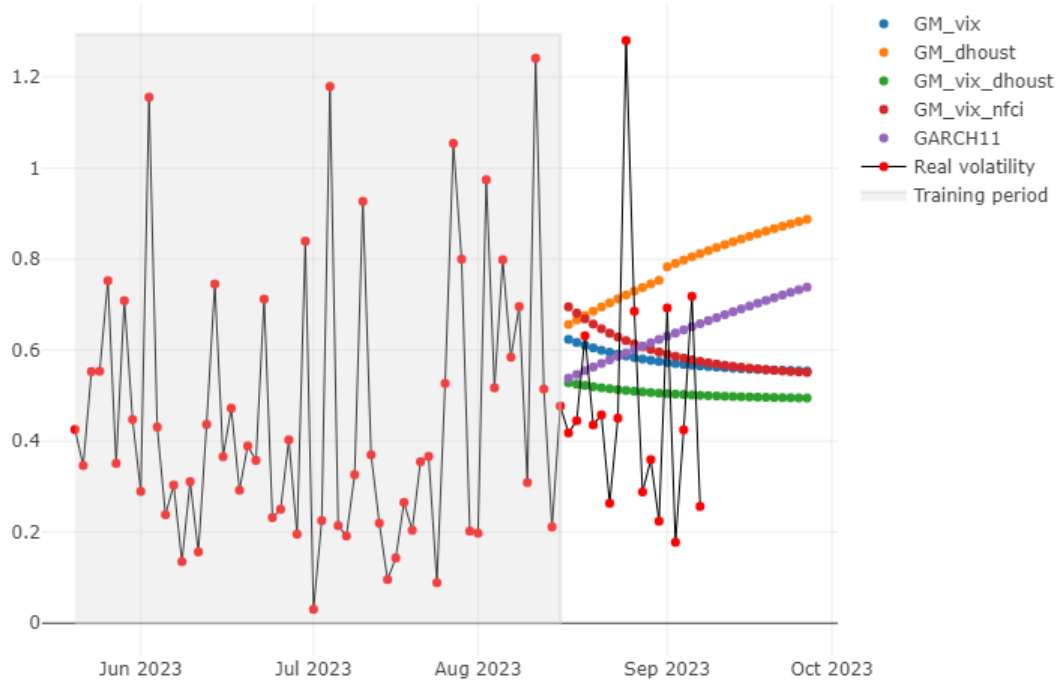


FIGURE 3.1 – Prédiction de la volatilité du S&P500 de date d'origine le 15/08/2023 de l'horizon 1 à 30 pour 5 modèles différents. Dans la légende à droite, "GM" signifie GARCH-MIDAS et les variables indiquées en suffixe du nom du modèle sont les variables explicatives utilisées pour ce modèle. "GARCH11" désigne le modèle GARCH(1,1).

On remarque en figure 3.1 une augmentation significative des prédictions de volatilité quotidienne en fonction de l'horizon pour le modèle GARCH-MIDAS utilisant l'indice HOUST. Cela a lieu au premier septembre et est du au changement de mois, car l'indice change alors de valeur. En effet, nous avons pu voir en partie 2.2.5 que la formule de la prédiction change au changement de mois.

Cependant, un tel saut n'est pas visible pour le modèle GARCH-MIDAS utilisant ensemble les indices VIX et HOUST. Nous souhaitons voir si la variable VIX est donc prépondérante face au HOUST dans ce modèle. Pour cela, nous nous intéressons aux paramètres θ et plus particulièrement aux ratios [1] suivants : $\zeta_{VIX} = \frac{\theta_{VIX}^{(GM\_VIX\_HOUST)}}{\theta_{VIX}^{(GM\_VIX)}}$ et $\zeta_{HOUST} = \frac{\theta_{HOUST}^{(GM\_VIX\_HOUST)}}{\theta_{HOUST}^{(GM\_HOUST)}}$ où l'indice correspond à la variable explicative rattachée au θ et l'exposant correspond au modèle GARCH-MIDAS duquel est issu le θ.

Pour les modèles entraînés avec les données jusqu'au 15/08/2023, les ratios valent :

1. Nous ne pouvons pas comparer directement les valeurs des deux paramètres θ du modèle GARCH-MIDAS utilisant le VIX et le HOUST, car ces deux variables n'ont pas le même ordre de grandeur.

$\zeta_{VIX} = 0.986$ et $\zeta_{HOUST} = 0.029$. Ainsi, l'influence du VIX sur τ est environ la même que le modèle soit entraîné avec uniquement le VIX ou avec le VIX et le HOUST. Au contraire, l'influence de HOUST diminue beaucoup (divisé par 30). C'est pourquoi aucun saut de la prédiction de la volatilité n'est visible au changement de mois.

3.1.3 Intervalles de confiance

Construction des intervalles de confiance Soit $h \in \mathbb{N}\backslash\{0\}$ l'horizon de prédiction. Notons X la valeur à prédire et $\hat{X}$ notre prédiction correspondante à l'horizon h. On suppose connu les valeurs de X indicé de 1 à N.

On souhaite calculer l'intervalle de confiance de notre prédiction $\hat{X}_\nu$ à l'horizon h en t, avec $\nu > N$ pour être hors échantillon (*out of sample*).

Algorithme 1 : Intervalle de confiance pour une prédiction à horizon h

Entrées : $(X_t)_{t\in[1,N]}$ valeurs cibles,
$(\hat{X}_t)_{t\in[1,N]}$ prédictions,
$\hat{X}_\nu$ pour $\nu > N$, qui est la prédiction pour laquelle on veut l'intervalle de confiance.

Sorties : q_- et q_+, les limites inférieure et supérieure de l'intervalle de confiance de $\hat{X}_\nu$ au niveau α

pour *i dans [1,N]* **faire**
  $\gamma_i \leftarrow \frac{X_i}{\hat{X}_i}$
fin
Trier γ dans l'ordre croissant;
Calculer $n_- \leftarrow \lfloor \frac{1-\alpha}{2} N \rfloor$;
Calculer $n_+ \leftarrow \lceil (1 - \frac{1-\alpha}{2}) N \rceil$;
Calculer $q_- \leftarrow \gamma_{n_-} \hat{X}_\nu$;
Calculer $q_+ \leftarrow \gamma_{n_+} \hat{X}_\nu$;
Retourner q_- et q_+

Pour utiliser cet algorithme, il faut que toutes les valeurs à prédire et nos prédictions soient de même signe. En l'occurrence, toutes ces valeurs sont bien positives, puisque ce sont des volatilités.

Par ailleurs, afin d'obtenir les quantiles q_- et q_+, on remarque que l'on arrondit à l'entier respectivement inférieur et supérieur pour calculer respectivement n_- et n_+ . Il est préférable, en particulier lorsque N n'est pas grand devant 100, de prendre α et N tels que les arrondis ne soient pas nécessaires. Dans ce qui suit, nous avons $\alpha = 0.9$ et $N = 60$.

Représentation des intervalles de confiance Nous constatons en figure 3.2 que la taille des intervalles de confiance n'augmente pas avec l'horizon. Cela est potentiellement dû

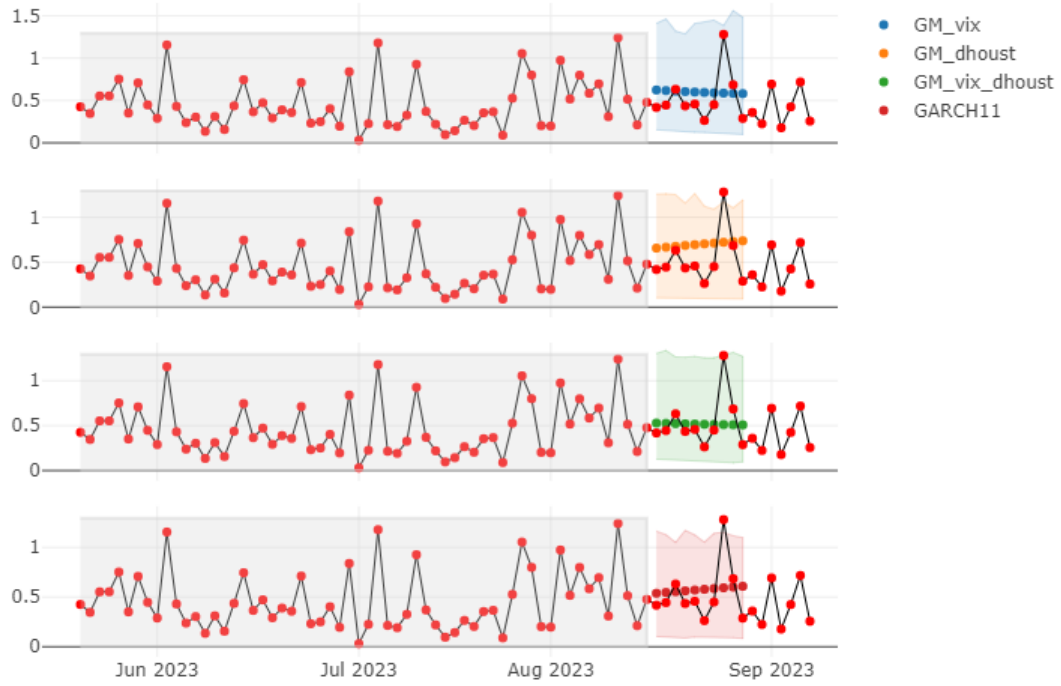


FIGURE 3.2 – Intervalles de confiance au niveau 90% des prédictions de l'horizon 1 à 10 de la volatilité sur S&P500 avec comme date d'origine le 15/08/2023.

aux deux points qui suivent : la série que nous essayons de prédire est stationnaire et les prédictions que nous faisons avec nos modèles sont toujours assez similaires. Ainsi, il n'est pas particulièrement plus dur de prédire la volatilité à 1 jour qu'à 10 jours.

3.2 Comparaison des modèles

3.2.1 Perte Qlike

Nous optons pour la fonction de perte QLIKE définie de la manière suivante : pour σ^2 la variance à une certaine date et h sa prédiction, on a :

$$QLIKE(\sigma^2, h) = \log\left(\frac{h}{\sigma^2}\right) + \frac{\sigma^2}{h} - 1$$

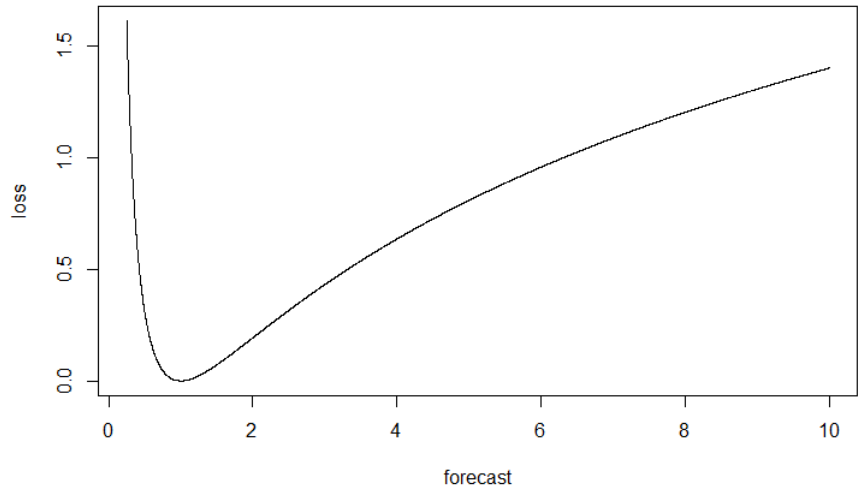


FIGURE 3.3 – Perte QLIKE pour $\sigma^2 = 1$

En raisonnant à σ^2 fixé, la fonction $\mathbb{R}_+^* \longrightarrow \mathbb{R},\ h \mapsto QLIKE(\sigma^2, h)$ est clairement asymétrique. Son minimum est bien atteint en σ^2. Le choix de la perte QLIKE permet de ne pas trop pénaliser les *outliers* qui surestiment la variance.

3.2.2 Erreurs moyennes

Nous ne disposons des volatilités à prédire (cf. partie 3.1.1) pour la période 2000 - 2019 que pour le S&P500. Nous ne pouvons donc pas réaliser le tableau suivant pour le NASDAQ-100.

Tableau des erreurs moyennes cumulatives Le terme *cumulatif* signifie que pour l'horizon h, l'erreur est définie telle que :

$$QLIKE\big(\sum_{i=1}^{h} X_i, \sum_{i=1}^{h} \widehat{X}_i\big)$$

et non telle que $QLIKE\big(X_h, \widehat{X}_h\big)$.

Horizon	1	2	5	10	22	44	66
GM_dhoust	0.29	0.26	0.36	0.42	0.40	0.33	0.29
GM_ip	0.32	0.28	0.35	0.39	0.36	0.29	0.23
GM_nai	0.29	0.26	0.34	0.39	0.36	<u>0.29</u>	<u>0.23</u>
GM_nfci	0.26	0.22	0.31	<u>0.37</u>	<u>0.35</u>	0.30	0.27
GM_Rvol22	0.27	0.23	0.32	0.38	0.36	0.29	0.24
GM_vix	<u>0.20</u>	<u>0.16</u>	<u>0.28</u>	0.40	0.46	0.45	0.44
GM_vrp	0.31	0.27	0.35	0.40	0.38	0.31	0.26
GM_vix_dhoust	0.23	0.21	0.34	0.43	0.47	0.45	0.44
GM_vix_ip	0.23	0.21	0.33	0.43	0.46	0.45	0.42
GM_vix_nai	0.23	0.21	0.32	0.41	0.42	0.39	0.37
GM_vix_nfci	0.22	0.20	0.33	0.43	0.47	0.46	0.44
GARCH(1,1)	0.43	0.42	0.49	0.48	0.43	0.36	0.29

TABLE 3.1 – Erreur moyenne cumulative des prédictions de la volatilité du S&P500 – La période d'entraînement est 1991 - 2014. Le nombre de prédictions faites pour chaque horizon et chaque modèle est 250."GM" signifie GARCH-MIDAS. Pour chaque horizon, la valeur soulignée est l'<u>erreur minimale</u>.

Bien que nous affichons là le résultat d'une année en particulier, il est intéressant de constater que le meilleur modèle pour de petits horizons est le GARCH-MIDAS avec une variable explicative quotidienne ; pour des horizons à moyen terme, le meilleur modèle s'avère être celui avec une variable explicative de fréquence hebdomadaire ; et pour des grands horizons, le meilleur modèle s'avère être celui avec une variable explicative de fréquence mensuelle.

Nous remarquons également que le meilleur modèle dépend de l'horizon que nous considérons. Par ailleurs, le modèle GARCH(1,1) parvient à battre des modèles GARCH-MIDAS pour certains horizons.

Nous avons réalisé pour différentes périodes de tests ce tableau. Le meilleur modèle (au sens de l'erreur qlike cumulative) change pour les grands horizons (un à trois mois). Cependant, le meilleur modèle pour de petits horizons (un jour à deux semaines) est très souvent le modèle GARCH-MIDAS avec la variable explicative VIX.

Tests de significativité Les écarts des erreurs moyennes entre les différents modèles sont éventuellement non significatifs. Il serait ainsi intéressant de tester si, à un horizon donné, un modèle bat significativement un autre modèle. Nous aurions pu utiliser à cette fin le test de Diebold-Mariano [5].

3.3 L'interface graphique

La finalité du stage était la construction d'une interface graphique sur RShiny présentant toutes mes représentations graphiques autour de la prédiction de la volatilité avec les modèles GARCH-MIDAS. Le lien pour y accéder est le suivant :
https://louisgeist.shinyapps.io/real_time_volatility_prediction/.

Dans l'espace gauche (visible sur la figure 3.4) , l'utilisateur choisit les options suivantes pour le premier graphique :

- l'indice dont on veut prédire la volatilité : S&P500 ou NASDAQ-100,
- la dernière date de disponibilité des données : du 01/05/2023 à hier (les modèles de la veille sont calculés et téléversés tous les jours à 8h30),
- l'horizon de prédiction : $[1 : 80]$,
- les modèles utilisés pour la prédiction (GARCH(1,1) ou tous les modèles GARCH-MIDAS qui utilisent soit une seule variable explicative, soit le VIX et une variable explicative non quotidienne),
- l'affichage ou non des intervalles de confiances,
- la superposition des modèles sur un même graphique (cf. figure 3.1) ou alors la création de plusieurs graphiques (cf. figure **??**).

Le deuxième graphique présente la ou les variables explicatives brutes d'un modèle GARCH-MIDAS et leur transformation τ après entraînement du modèle. L'utilisateur peut choisir parmi tous les modèles GARCH-MIDAS déjà présentés ci-dessus et parmi les dates d'entraînement du 01/05/2023 à hier.

Le troisième graphique est simplement la représentation graphique de la transformation stationnaire du S&P500 ou du NASDAQ-100.

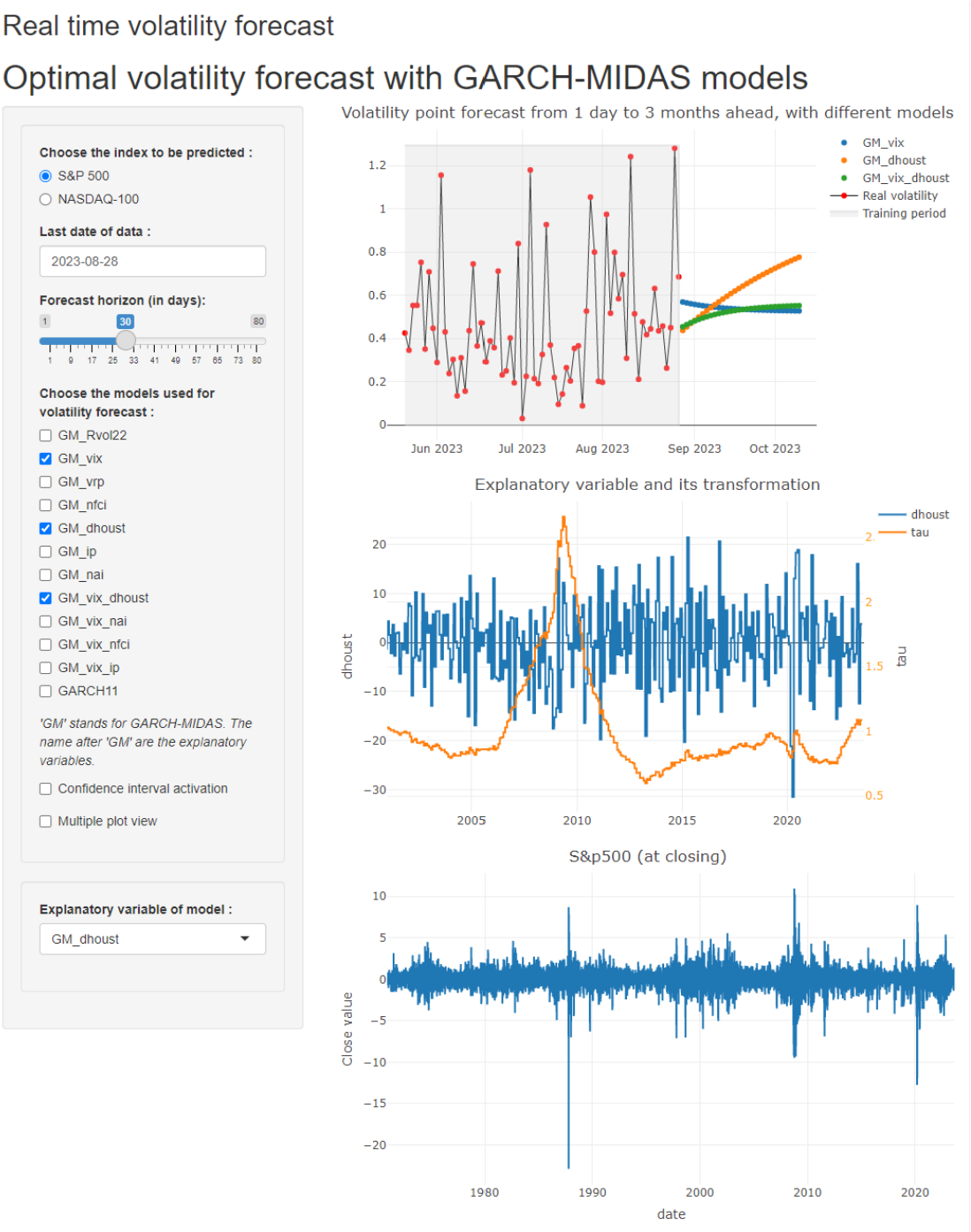
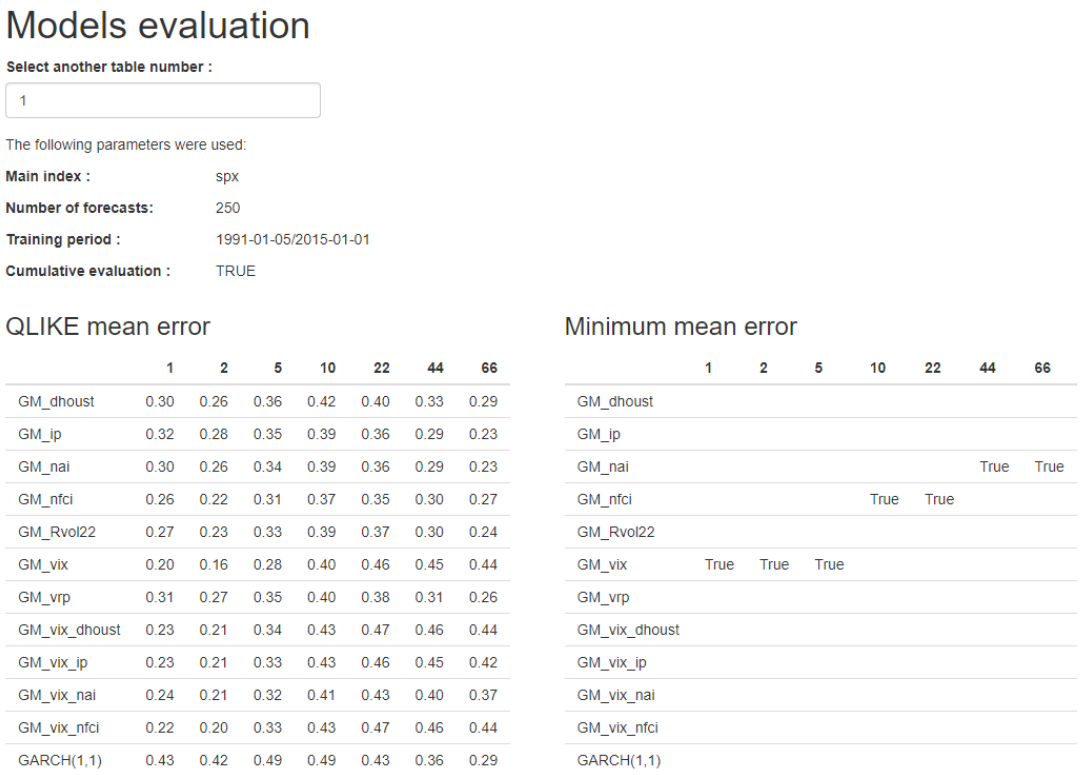


FIGURE 3.4 – RShiny app – Première partie

La figure 3.5 montre l'affichage des tableaux d'erreurs moyennes cumulatives. L'utilisateur peut sélectionner un autre numéro de tableau, affichant un tableau avec une autre période d'entraînement ou de test.



Models evaluation

Select another table number :

1

The following parameters were used:

Main index : spx

Number of forecasts: 250

Training period : 1991-01-05/2015-01-01

Cumulative evaluation : TRUE

QLIKE mean error

	1	2	5	10	22	44	66
GM_dhoust	0.30	0.26	0.36	0.42	0.40	0.33	0.29
GM_ip	0.32	0.28	0.35	0.39	0.36	0.29	0.23
GM_nai	0.30	0.26	0.34	0.39	0.36	0.29	0.23
GM_nfci	0.26	0.22	0.31	0.37	0.35	0.30	0.27
GM_Rvol22	0.27	0.23	0.33	0.39	0.37	0.30	0.24
GM_vix	0.20	0.16	0.28	0.40	0.46	0.45	0.44
GM_vrp	0.31	0.27	0.35	0.40	0.38	0.31	0.26
GM_vix_dhoust	0.23	0.21	0.34	0.43	0.47	0.46	0.44
GM_vix_ip	0.23	0.21	0.33	0.43	0.46	0.45	0.42
GM_vix_nai	0.24	0.21	0.32	0.41	0.43	0.40	0.37
GM_vix_nfci	0.22	0.20	0.33	0.43	0.47	0.46	0.44
GARCH(1,1)	0.43	0.42	0.49	0.49	0.43	0.36	0.29

Minimum mean error

	1	2	5	10	22	44	66
GM_dhoust							
GM_ip							
GM_nai						True	True
GM_nfci				True	True		
GM_Rvol22							
GM_vix	True	True	True				
GM_vrp							
GM_vix_dhoust							
GM_vix_ip							
GM_vix_nai							
GM_vix_nfci							
GARCH(1,1)							

FIGURE 3.5 – RShiny – Deuxième partie

Conclusion

Ainsi, l'utilisation de variables explicatives dans un modèle GARCH-MIDAS permet d'accroître la précision des prédictions de la volatilité au sens de l'erreur QLIKE par rapport à un modèle GARCH(1,1) classique.

Cependant, la précision d'un modèle GARCH-MIDAS dépend non-seulement de la ou des variables explicatives choisis, mais aussi de l'horizon de prédiction. Par exemple, pour certains horizons, tous les modèles GARCH-MIDAS ne battent pas le modèle GARCH(1,1) sur certaines périodes de tests

Le GARCH-MIDAS qui utilise le VIX, mesure prospective de la volatilité quotidienne, reste un très bon modèle parmi les modèles testés pour prédire la volatilité quotidienne du S&P 500 à n'importe quel horizon, et en particulier pour de petits horizons.

Bibliographie

[1] Ole E. Barndorff-Nielsen et Neil Shephard. “Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models”. In : *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64.2 (2002), p. 253-280. issn : 13697412, 14679868. url : http://www.jstor.org/stable/3088799 (visité le 24/07/2023).

[2] Winston Chang et al. *shiny : Web Application Framework for R*. R package version 1.7.4. 2022. url : https://CRAN.R-project.org/package=shiny.

[3] C. Conrad et O. Kleen. “Two are better than one : Volatility forecasting using multiplicative component GARCH-MIDAS models”. In : *J Appl Econ.* 35 (2020), p. 19-45. url : https://doi.org/10.1002/jae.2742.

[4] Christian Conrad et Karin Loch. “Anticipating Long-Term Stock Market Volatility”. In : *Journal of Applied Econometrics* 30.7 (2015), p. 1090-1114. url : https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2404.

[5] Francis X Diebold et Robert S Mariano. “Comparing Predictive Accuracy”. In : *Journal of Business & Economic Statistics* 20.1 (2002), p. 134-144. url : https://doi.org/10.1198/073500102753410444.

[6] R. F. Engle, E. Ghysels et B. Sohn. “Stock market volatility and macroeconomic fundamentals”. In : *Review of Economics and Statistics* 95 (2013), p. 776-797.

[7] Alexios Ghalanos. *rugarch : Univariate GARCH models*. R package version 1.4-9. 2022.

[8] LAWRENCE R. GLOSTEN, RAVI JAGANNATHAN et DAVID E. RUNKLE. “On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks”. In : *The Journal of Finance* 48.5 (1993), p. 1779-1801. url : https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1993.tb05128.x.

[9] Peter R. HANSEN et Asger LUNDE. "A forecast comparison of volatility models : does anything beat a GARCH(1,1) ?" In : *Journal of Applied Econometrics* 20.7 (2005), p. 873-889. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.800.

[10] Onno KLEEN. *alfred : Downloading Time Series from ALFRED Database for Various Vintage*. R package version 0.2.1. URL : https://github.com/onnokleen/alfred/.

[11] Onno KLEEN. *mfGARCH : Mixed-Frequency GARCH Models*. R package version 0.2.1. 2021. URL : https://github.com/onnokleen/mfGARCH/.