Travail d'analyse et de synthèse

Should Macroeconomic Forecasters Use Daily Financial Data and How?

Léo Renault · Nicolas Annon · Théo Verdelhan · Arthur Le Net

31 janvier 2026

Gestion Quantitative

Master 2 272 – Ingénierie Économique et Financière
Parcours Finance Quantitative
Université Paris-Dauphine–PSL

Table des matières

1 Introduction **1**

2 Contexte, motivation et contribution de l'étude **2**
2.1 Problématique et contexte 2
2.2 Objectifs et contribution de l'étude 3

3 Modèles de régression MIDAS **4**
3.1 Problèmes d'agrégation temporelle 5
3.2 Nowcasting et avances 6

4 Données et pré-traitements **8**
4.1 Sources, fréquences et périmètres d'échantillon 8
4.2 Construction et alignement des séries 8
4.3 Transformations des données journalières 9
4.4 Extraction des facteurs financiers journaliers 9
4.5 Analyse des facteurs PCA 10
4.6 Difficultés rencontrées et limites des données 10

5 Design de réplication et protocole de prévision **12**
5.1 Échantillons, horizons et fenêtre m 12
5.2 Modèles estimés et benchmarks 12
5.3 Procédure pseudo hors-échantillon (expanding/rolling) 13
5.4 Métriques d'évaluation 13

6 Résultats de réplication **15**
6.1 Réplication des résultats du papier : *Long sample* (1988–2008) 15
6.2 Réplication des résultats du papier : *Short sample* (1999–2008) 16
6.3 Analyse hors-échantillon récente (2024–2025) 17
6.4 Illustrations graphiques (période récente) 18

7 Extension : MIDAS à deux paramètres β (lags vs leads) **19**
7.1 Motivation et intuition économique 19
7.2 Spécification du modèle 19
7.3 Protocole d'évaluation (OOS) et configuration 19
7.4 Résultats : performance et comparaison au single-β 20
7.5 Application aux échantillons du papier (Long/Short sample) 20
7.6 Positionnement par rapport aux autres modèles (période récente) 21
7.7 Illustrations graphiques 22
7.8 Discussion et limites 22

8 Conclusion **23**

Préambule

Objet du document. Ce document présente un travail de lecture, d'analyse et de restitution portant sur l'article de recherche intitulé *Should Macroeconomic Forecasters Use Daily Financial Data and How?*. Nous y synthétisons la contribution méthodologique des auteurs, en particulier l'utilisation des approches MIDAS et la manière d'exploiter des données financières quotidiennes pour la prévision macroéconomique.

Elena Andreou
Département d'économie, Université de Chypre, CY 1678 Nicosie, Chypre
elena.andreou@ucy.ac.cy

Eric Ghysels
Département d'économie, Université de Caroline du Nord, Chapel Hill, NC 27599–3305,
et Département de finance, Kenan–Flagler Business School, Chapel Hill, NC 27599
eghysels@unc.edu

Andros Kourtellos
Département d'économie, Université de Chypre, CY 1678 Nicosie, Chypre
andros@ucy.ac.cy

1 Introduction

Ce travail propose une réplication de l'article *Should Macroeconomic Forecasters Use Daily Financial Data and How ?* d'Andreou, Ghysels et Kourtellos (2013). L'étude s'inscrit dans la littérature sur la prévision macroéconomique et interroge l'intérêt d'exploiter des données financières quotidiennes pour anticiper des variables observées à plus basse fréquence, en particulier la croissance trimestrielle du PIB.

L'argument central est que les prix d'actifs incorporent rapidement l'information et les anticipations des agents, ce qui peut fournir un signal utile pour la prévision de l'activité réelle. Le défi consiste toutefois à intégrer correctement cette information haute fréquence sans perdre la dynamique intra-période par une agrégation trop naïve. L'article met en avant deux difficultés principales : d'une part, la gestion des fréquences mixtes (quotidien vs trimestriel) ; d'autre part, la synthèse de l'information contenue dans un grand nombre de séries financières couvrant plusieurs classes d'actifs.

Les auteurs proposent d'y répondre à l'aide de régressions MIDAS (*Mixed Data Sampling*), qui permettent de relier une variable trimestrielle à des observations quotidiennes sans imposer une pondération uniforme. Dans cette approche, un schéma de pondération paramétrique résume l'information de haute fréquence de manière parcimonieuse. L'article discute également des extensions de type nowcasting via l'introduction d'avances (*leads*), afin de mobiliser l'information disponible à l'intérieur du trimestre en cours tout en respectant le calendrier d'information.

Notre réplication poursuit trois objectifs. Premièrement, reproduire les principaux résultats empiriques de l'article, en suivant les périodes d'échantillon et la logique d'évaluation hors-échantillon. Deuxièmement, expliciter les choix opérationnels nécessaires à l'implémentation (construction des facteurs, paramétrisation, protocole pseudo hors-échantillon et métriques). Troisièmement, proposer et tester une extension simple du cadre initial, afin d'évaluer si un léger assouplissement de la structure de pondération peut améliorer les performances dans certaines configurations.

2 Contexte, motivation et contribution de l'étude

2.1 Problématique et contexte

La prévision macroéconomique repose globalement sur des modèles exploitant des variables observées à fréquence mensuelle ou trimestrielle, alors même que les marchés financiers génèrent en continu une information abondante journalière. Cette dissociation soulève une problématique centrale : comment exploiter efficacement l'information contenue dans les données financières quotidiennes pour améliorer la prévision de l'activité économique agrégée, sans introduire de biais liés à l'agrégation temporelle ou à la prolifération des paramètres ? Cette question revêt une importance particulière dans un contexte où les prix d'actifs sont réputés intégrer rapidement les anticipations et les chocs informationnels, comme lors de périodes de forte instabilité macrofinancière depuis l'élection du Président des États-Unis, Donald Trump.

Sur le plan méthodologique, Les solutions classiques pour relier les variables haute fréquence avec les variables basses fréquences consistaient à agréger les données financières de haute fréquence :

(1) L'agrégation temporelle (par exemple moyenne, somme, dernière valeur, etc.) : On transforme la donnée de haute fréquence $x_{t,i}$ (par exemple quotidienne) en une variable de basse fréquence X_t (par exemple trimestrielle) par la moyenne par exemple :

$$X_t = \frac{1}{m} \sum_{i=1}^{m} x_{t,i}, \tag{1}$$

où m désigne le nombre d'observations de haute fréquence contenues dans la période t environ 60 pour un trimestre. Puis, on estime ensuite une régression linéaire classique reliant la variable agrégée X_t à la variable cible de basse fréquence :

$$Y_{t+1} = \alpha + \beta X_t + \varepsilon_{t+1}. \tag{2}$$

Résultat : perte d'informations sur la dynamique intra-période, l'ensemble des données ont le même poids. Rien ne justifie économiquement que les données anciennes soient aussi informatives que les récentes.

(2) La Régression naïve avec toutes les données de haute fréquence : vise à inclure directement les observations de haute fréquence dans la régression reliant la variable cible de basse fréquence à ses prédicteurs. la relation estimée peut s'écrire comme :

$$Y_{t+1} = \alpha + \sum_{i=1}^{m} \beta_i x_{t,i} + \varepsilon_{t+1}, \tag{3}$$

où $x_{t,i}$ désigne la i-ème observation de haute fréquence (par exemple quotidienne) au sein de la période t, et m correspond au nombre total d'observations de haute fréquence contenues dans cette période. Bien que cette spécification exploite pleinement l'information de haute fréquence disponible, elle présente plusieurs limites économétriques majeures.

1. *Explosion du nombre de paramètres.* lorsque m est élevé (par exemple $m \approx 60$ jours de bourse par trimestre), le modèle requiert l'estimation d'un grand nombre de coefficients. Étant donné la taille réduite de l'échantillon en basse fréquence, cela conduit à une sur-paramétrisation sévère.
2. *Multicolinéarité extrême.* si les observations $x_{t,i}$ sont fortement autocorrélées, ce qui induit une forte colinéarité entre les régresseurs et entraîne une variance très élevée des estimateurs des coefficients β_i.
3. *Sur-apprentissage.* le modèle s'ajuste excessivement aux données in-sample, produisant une qualité des performances de prévision moindres hors échantillon.

   Résultat : Inapplicabilité en pratique car le nombre de paramètres explose

L'approche conventionnelle pour prédire une variable observée à un basse fréquence avec des séries financières haute fréquence consiste à utiliser un modèle de régression à retards distribués augmentés, ADL (p_Y^Q, q_X^Q). Les régressions ADL s'appuient sur des variables de haute fréquence agrégées, ce qui revient à imposer un schéma de pondération fixe, uniforme. Ce choix d'agrégation n'est pas neutre économétriquement, car il suppose que l'information intra-période est homogène et indépendante de l'horizon (ne se déprécie pas dans le temps). Or, dans les séries financières, l'information est temporellement hiérarchisée, les données les plus récentes sont plus pertinentes (l'information se diffuse progressivement dans le prix de l'actif). Le problème central réside dans le caractère exogène et non adaptatif des poids imposés par rapport aux données et à l'objectif de prévision.

Par ailleurs, des recherches préalables tentent de résoudre ce problème par le processus de nowcasting. Selon les auteurs, le nowcasting répond à deux difficultés structurelles majeures de la prévision macroéconomique en temps réel :

1. La disponibilité asynchrone au cours du temps d'un large échantillon hétérogène des variables du modèle (problème dit de bord irrégulier ou jagged/ragged edge).
2. L'inadéquation des modèles standards en temps réel : inadaptés pour des mise à jour séquentielles.

Nunes (2005) et Giannone, Reichlin et Small (2008), entre autres, ont formalisé le nowcasting qui vise à exploiter toute l'information partiellement disponible à un instant donnée. Ce processus met à jour une prévison de la variable basse fréquence (par exemple le PIB trimestriel), avant même que cette variable ne soit publiée officiellement, à mesure que les nouvelles données haute fréquence deviennent disponibles (estimer l'état courant de la variable cible) tout en prenant en compte l'incertitude liées aux données manquantes. L'approche de nowcasting par filtre de Kalman peut être interprétée comme une généralisation probabiliste de l'agrégation temporelle, dans laquelle les poids implicites appliqués aux observations de haute fréquence sont endogènes et déterminés par un modèle espace–état. Mais ils sont lourds à spécifier, sensibles aux erreurs de modélisation lorsque le nombre de séries quotidiennes est élevée. Contrairement aux régressions ADL à agrégation plate, ces poids résultent de la dynamique estimée des facteurs latents et du calendrier des publications.

Andreou, Ghysels et Kourtellos proposent une extension des régressions ADL sans agrégation arbitraire lors de l'intégration des données hautes fréquence, tout en offrant une alternative forme réduite au nowcasting par filtre de Kalman. Leur approche évite la lourdeur paramétrique et la dépendance aux états latents propres aux modèles espace–état mais conserve la flexibilité informationnelle du nowcasting.

2.2 Objectifs et contribution de l'étude

L'objectif principal de l'étude est d'évaluer l'apport informationnel des données financières quotidiennes pour la prévision de la croissance trimestrielle du PIB réel, en proposant une approche de prévision fondée sur des régressions à fréquences mixtes (MIDAS). L'analyse vise à dépasser l'usage des indicateurs financiers agrégés en exploitant directement leur dynamique quotidienne, tout en évitant la complexité inhérente aux modèles structurels à espace d'état. Bien que performants dans certains contextes, ces derniers nécessitent l'estimation d'un grand nombre de paramètres et deviennent difficilement applicables lorsque l'ensemble informationnel inclut des centaines de séries financières quotidiennes. Sur le plan empirique, l'étude essaie de mettre en évidence des gains de prévision robustes et d'identifier les classes d'actifs financiers les plus informatives pour la prévision de l'activité réelle.

3 Modèles de régression MIDAS

Introduit dans les années 2000 par par Ghysels, Santa-Clara et Valkanov, le modèle MIDAS est une régression à fréquences mixtes qui permet de relier une variable dépendante observée à basse fréquence (par exemple le PIB trimestriel) à des variables explicatives observées à haute fréquence (des données financières quotidiennes), sans agréger arbitrairement ces données.

Supposons que nous souhaitons prévoir une variable observée à basse fréquence (ex : trimestrielle) notée $Y_{t+h}^{Q,h}$ en utilisant des séries financières journalières considérées comme prédicteurs utiles. Notons $X_{m-j,t}^{D}$ la j-ième observation quotidienne comptée à rebours au cours du trimestre t. le dernier jour du trimestre correspond à j = 0 (donc l'avant dernier j = 1). Alors $X_{m,t}^{D}$ la j-ième observation quotidienne où m désigne le nombre de retards quotidiens, ou de manière équivalente le nombre de jours de bourse par trimestre, supposé constant par souci de simplicité.

Soit le modèle de régression ADL–MIDAS(p_Y^Q, q_X^D) défini par :

$$Y_{t+h}^{Q,h} = \mu^h + \sum_{j=0}^{p_Y^Q-1} \rho_{j+1}^h Y_{t-j}^Q + \beta^h \sum_{j=0}^{q_X^D-1} \sum_{i=0}^{m-1} w_{i+jm}^h X_{m-i,t-j}^D + u_{t+h}^h. \tag{4}$$

Dans ce modèle, le schéma de pondération w_{i+jm}^h dépend d'un vecteur de faible dimension d'hyperparamètres inconnus θ, ce qui permet d'éviter le problème de prolifération des paramètres.

Le principe fondamental des modèles MIDAS consiste à approximer la projection linéaire complète reliant la variable de basse fréquence à l'ensemble des observations de haute fréquence à l'aide d'une structure de pondération paramétrique de faible dimension. Plutôt que d'estimer un coefficient distinct pour chaque retard (observation) de haute fréquence, les modèles MIDAS imposent que les coefficients suivent une forme fonctionnelle dépendant d'un nombre réduit d'hyperparamètres (prolifération des paramètres), ce qui réduit considérablement la dimension du problème d'estimation.

Andreou, Ghysels et Kourtellos (2013) font le choix d'utiliser un polynôme de retards exponentiels d'Almon. Soit $x_{t,j}$ la j-ème observation de haute fréquence (par exemple quotidienne) disponible au sein de la période t, avec $j = 1, \ldots, m$. La pondération exponentielle d'Almon définit les poids $w_j(\theta)$ associés à chaque observation comme :

$$w_j(\theta) = \frac{\exp\left(\theta j^2\right)}{\sum_{k=1}^{m} \exp\left(\theta k^2\right)}, \tag{5}$$

où $\theta \in R$ est un hyperparamètre à estimer.

Par construction, cette spécification impose deux contraintes essentielles :

1. Les poids sont strictement positifs, i.e. $w_j(\theta) > 0$ pour tout j ;
2. Les poids sont normalisés, c'est-à-dire $\sum_{j=1}^{m} w_j(\theta) = 1$;

Ces contraintes assurent l'identification du coefficient de pente associé (β^h) au prédicteur agrégé et l'interprètent comme un effet marginal global des données de haute fréquence sur la variable de basse fréquence. La fonction exponentielle permet de générer des profils de pondérations linéaires et décroissants, capturant un mécanisme de mémoire décroissante selon lequel les observations les plus récentes sont plus informatives (importances relatives). Ainsi, au lieu d'estimer m coefficients distincts pour chaque observation, La pondération exponentielle d'Almon dans les régressions MIDAS permet de résumer l'information contenue dans les m observations de haute fréquence à l'aide d'un unique hyperparamètre (θ^h) en vue d'obtenir une projection linéaire des données quotidiennes sur la variable trimestrielle. Cet hyperparamètre garantit une forte parcimonie sur les coefficients associés aux retards tout en conservant une flexibilité suffisante. Si $\theta^h < 0$, les poids décroissent avec j, les obersvations récentes reçoivent plus de poids, ce qui se justifie économiquement. Ce schéma de pondérationn est particulièrement utile dans notre contexte compte tenu de la taille réduite de notre échantillon de

données. En imposant un seul hyperparamètre, on réduit la variance des estimateurs (qui peut être élevée avec des données trimestrielles), améliorant la stabilité de l'optimisation et la robustesse hors échantillon en conséquence.

Dans des exercices non rapportés, les auteurs indiquent avoir expérimenté un polynôme de retards exponentiel d'Almon à deux paramètres sans constater d'amélioration des performances de prévision. D'abord, la taille limitée de l'échantillon en basse fréquence rend toute sur-paramétrisation particulièrement coûteuse en termes de variance des estimateurs. Puis, l'hypothèse d'une structure de pondération monotone et décroissante est économiquement cohérente dans un contexte de prévision macroéconomique, où l'information financière la plus récente est supposée contenir d'avantage de signaux sur l'activité future.

Notons tout de même que les auteurs estiment les paramètres $(\mu^h, \rho_1^h, \rho_2^h, \ldots, \rho_{p_Q}^h, \beta^h, \theta^h)$ du modèle de régression MIDAS de l'équation (4) par des moindres carrés non linéaires.

En résumé, contrairement à l'agrégation temporelle uniforme, la pondération exponentielle d'Almon permet aux données de déterminer endogènement la contribution relative de chaque observation intra-période, évitant ainsi une perte d'information et une spécification arbitraire des poids.

Par ailleurs, en comparaison avec une régression naïve, la pondération d'Almon impose une régularisation structurelle forte qui réduit le nombre de paramètres à estimer. Cette contrainte permet de limiter la multicolinéarité induite par l'autocorrélation élevée des données de haute fréquence et de réduire le risque de sur-apprentissage, améliorant ainsi les performances de prévision hors échantillon.

3.1 Problèmes d'agrégation temporelle

L'agrégation paramétrique guidée par les données (MIDAS) peut être reliée à la littérature sur l'agréagtion temporelle et au modèle ADL en considéreant la variable basse fréquence filtrée suivante, déterminée par les paramètres :

$$X_t^Q(\tilde{\theta}) = \sum_{i=0}^{m-1} w_i(\tilde{\theta})\, X_{m-i,t}^D \tag{6}$$

Cette équation (6) montre que l'approche MIDAS reste une forme d'agrégation des données de haute fréquence. Les modèles MIDAS peuvent être interprétés comme une extension flexible des modèles ADL classiques, dans laquelle la structure des pondérations intra-période est déterminée de manière endogène (paramétrée). La forme du schéma de pondération est estimée à partir des données pour mieux refléter la dynamique informationnelle des séries de haute fréquence. Cette approche permet de capturer des profils de mémoire décroissants, plus cohérents avec la dynamique des séries financières, tout en conservant une structure parcimonieuse adaptée aux échantillons de taille limitée.

Nous pouvons alors définir le modèle ADL–MIDAS–M(p_Y^Q, q_X^Q), où le M fait référence au schéma de pondération multiplicatif du modèle, à savoir :

$$Y_{t+h}^{Q,h} = \mu^h + \sum_{k=0}^{p_Y^Q-1} \rho_k^h Y_{t-k}^Q + \sum_{k=0}^{q_X^Q-1} \beta_k^h X_{t-k}^Q(\tilde{\theta}^h) + u_{t+h}^h. \tag{7}$$

Une question se pose sur la manière dont la régression de l'équation (4) se rapporte à l'approche plus tradtionnelle faisant intervenir le filtre de Kalman. Dans ces modèles, la variable macroéconomique d'intérêt est généralement interprétée comme une composante latente, observée de manière imparfaite à travers un ensemble de signaux de haute fréquence. Le filtre de Kalman permet alors de combiner ces signaux de façon optimale, en tenant compte de leur dynamique et de leur précision relative. Bien que le filtre de Kalman soit optimal dans un cadre gaussien, il est sensible aux erreurs de spécification et nécessite l'estimation d'un grand nombre de paramètres, en particulier lorsque le nombre de séries de haute fréquence est élevé.

À l'inverse, l'approche MIDAS ne repose pas sur l'introduction d'une variable latente ni sur la spécification complète d'un modèle d'état. Elle adopte une représentation en forme réduite, dans laquelle la variable de basse fréquence est directement reliée à une agrégation pondérée des données de haute fréquence observées. Les pondérations jouent un rôle analogue à celui du mécanisme de filtrage dans les modèles à espace d'état, en déterminant la contribution relative des différentes observations intra-période à la prévision. C'est une approximation d'un processus de filtrage optimal.

Cette approche présente l'avantage d'être plus simple à estimer, plus robuste dans des échantillons de taille limitée, et particulièrement adaptée à des objectifs de prévision hors échantillon, ce qui motive son utilisation par rapport aux méthodes fondées sur le filtre de Kalman uniquement.

3.2 Nowcasting et avances

Les auteurs du papier proposent une stratégie alternative en forme réduite, reposant sur des régressions MIDAS avec avances. Elles reposent sur l'exploitation des observations de haute fréquence disponibles entre les dates t et $t+1$, avant la publication officielle de la variable macroéconomique de basse fréquence. Imaginons, nous sommes à deux mois à l'intérieur du trimestre $t+1$, c'est-à-dire à la fin novembre 2025, et que notre objectif soit de prévoir l'activité économique trimestrielle à fin décembre 2025. Nous disposons donc de l'équivalent d'au moins 44 jours de bourse (deux mois) de données financières quotidiennes.

Notons $X^D_{m-i,t+1}$ le i-ème jour compté à rebours dans le trimestre $t+1$ et considérons J^D_X avances quotidiennes pour le prédicteur quotidien en termes de multiples de mois (nombre de mois d'avance utilisé), avec $J^D_X = 2. Alors, X^D_{2m/3,t+1}$ correspond à 44 avances, tandis que $X^D_{1,t+1}$ correspond à une avance pour le prédicteur quotidien (3 mois moins 2 mois).

Formellement, On peut spécifier un modèle ADL–MIDAS(p^Q_Y, q^D_X, J^D_X) avec avances de la forme :

$$\begin{aligned} Y^{Q,h}_{t+h} = \mu^h + \sum_{k=0}^{p^Q_Y-1} \rho^h_k Y^Q_{t-k} \\ + \beta^h \left[\sum_{i=(3-J^D_X)m/3}^{m-1} w^{\theta^h}_{i-m} X^D_{m-i,t+1} + \sum_{j=0}^{q^D_X-1} \sum_{i=0}^{m-1} w^{\theta^h}_{i+jm} X^D_{m-i,t-j} \right] + u^h_{t+h}. \end{aligned} \tag{8}$$

Dans cette équation, $Y^{Q,h}_{t+h}$ désigne la variable macroéconomique trimestrielle à prédire à l'horizon h, tandis que le premier terme de somme capture la dynamique autorégressive de la variable de basse fréquence. Le terme central correspond à une agrégation pondérée des données financières quotidiennes, où les pondérations MIDAS w^{θ^h} permettent de résumer l'information de haute fréquence à l'aide d'un nombre réduit de paramètres. Il convient de noter qu'il existe diverses manières d'hyperparamétrer les polynômes MIDAS d'avances et de retards.

$$\beta^h \Bigg[\underbrace{\sum_{i=(3-J^D_X)m/3}^{m-1} w^{\theta^h}_{i-m} X^D_{m-i,t+1}}_{\text{(1) composante avec avances}} + \underbrace{\sum_{j=0}^{q^D_X-1} \sum_{i=0}^{m-1} w^{\theta^h}_{i+jm} X^D_{m-i,t-j}}_{\text{(2) composante sans avances (passé)}} \Bigg]. \tag{9}$$

La première composante de cette agrégation exploite les observations quotidiennes disponibles observées avant la fin du trimestre dans le trimestre $t+1$, c'est la notion d'avance (nowcasting avec le terme $X^D_{m-i,t+1}$). La seconde composante intègre les données quotidiennes des trimestres précédents, assurant la continuité temporelle du prédicteur et donc une forme de mémoire avec le terme $X^D_{m-i,t-j}$.

Les pondérations MIDAS jouent un rôle de filtrage statique, déterminant la contribution relative des observations quotidiennes les plus récentes par rapport aux plus anciennes. Cette approche traite le problème du bord irrégulier sans modéliser explicitement des facteurs latents, ce qui permet une estimation parcimonieuse et robuste.

Il existe deux différences importantes entre le nowcasting (utilisant le filtre de Kalman) et les modèles MIDAS avec avances :

(1) Nature des prévisions : mises à jour intra-période vs prévisions directes multi-horizons. Le *nowcasting* renvoie typiquement à des mises à jour fréquentes des prévisions à l'intérieur de la période courante (par exemple prévisions de croissance du PIB réel du trimestre courant). Les modèles MIDAS avec avances peuvent également jouer ce rôle, mais aussi prédire le PIB réel à horizon futur (à plusieurs trimestres). Surtout, une différence importante est que les régressions MIDAS permettent d'obtenir des prévisions directes à h pas en avant (une régression estimée spécifiquement pour chaque horizon), par opposition aux approches itérées qui reposent sur la dynamique implicite du modèle. Les modèles de nowcasting fondés sur une représentation état–espace estimée par filtre de Kalman et les régressions MIDAS partagent la capacité de produire des prévisions à plusieurs horizons. Toutefois, une différence méthodologique centrale réside dans le fait que modèles état–espace reposent généralement sur des prévisions itérées, tandis que les régressions MIDAS permettent d'estimer des prévisions directes à l'horizon h. Dans une approche itérée, le modèle est estimé pour une prévision à un pas en avant, puis les valeurs prévues sont réinjectées de manière récursive afin d'obtenir une prévision à h pas. Cette procédure exploite la dynamique implicite du modèle estimé, mais toute erreur de spécification est transmise et amplifiée au fil des itérations. En conséquence, la qualité des prévisions tend à se détériorer à mesure que l'horizon de prévision s'allonge. À l'inverse, les régressions MIDAS reposent sur une stratégie de prévision directe. Pour chaque horizon h, une équation spécifique est estimée, reliant directement la variable d'intérêt $Y_{t+h}^{Q,h}$ à l'ensemble de l'information disponible à la date de prévision. Cette approche évite la propagation des erreurs inhérente aux prévisions itérées et confère aux modèles MIDAS une plus grande robustesse face aux erreurs de spécification.

(2) Traitement explicite du *ragged edge* et rôle du calendrier des publications. La seconde différence concerne la nature à bords irréguliers (*jagged/ragged edge*) des bases macroéconomiques en temps réel. Le *nowcasting* par filtre de Kalman traite explicitement ce problème, car le calendrier des publications et la structure des données manquantes joue un rôle important dans la spécification des équations de mesure du modèle espace–état. De plus, le filtre de kalman autorise la présence d'observations manquantes dans les équations de mesure et permet les révisions ex post des données (source d'incertitude supplémentaire). L'approche MIDAS avec avances ne modélise ni les processus de publication, ni la dynamique latente des variables (filtre de Kalman), ni leur loi jointe. Le principe central consiste à reformuler le problème du nowcasting en exploitant directement, dans une régression, toute l'information effectivement disponible à la date de prévision. L'asynchronie des données n'est plus traitée comme un problème d'observations manquantes, mais comme une structure d'avances, permettant d'intégrer des données de haute fréquence appartenant à la période courante ou suivante. les auteurs s'appuient sur l'idée qu'avec l'approche MIDAS l'information nouvelle est rapidement incorporée dans les prix d'actifs. Ainsi, le flux d'information en temps réel est capturé via les variables financières de haute fréquence (nature fondamentalement prospective des prix des actifs financiers), ce qui permet de produire des mises à jour de prévision et dispense de la mise à jour continue des séries macroéconomiques de basse fréquence et de la modélisation explicite du calendrier de publication. Contrairement aux indicateurs macroéconomiques, ces variables financières sont observées sans erreur de mesure significative et ne font pas l'objet de révisions, ce qui renforce la fiabilité empirique des prévisions. Enfin, le filtre de Kalman, dans le contexte du nowcasting, facilite l'étude de l'impact des annonces macroéconomiques, "chocs", sur les prévisions. Les régressions MIDAS avec avances peuvent également traiter le caractère irrégulier des séries et fournissent des outils similaires. En effet, les régressions MIDAS permettent d'estimer des régressions autour des dates d'annonces (données financières antérieures et postérieures aux annonces) et en analyser les variations induites des prévisions.

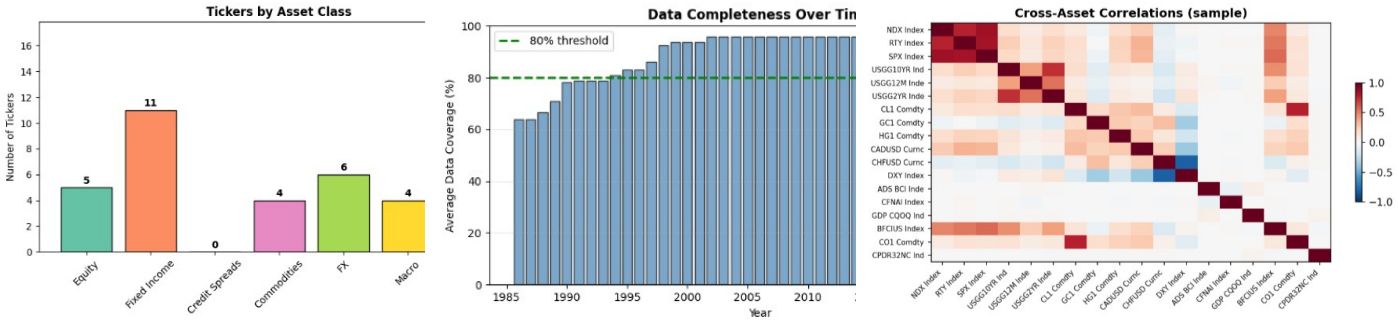
4 Données et pré-traitements

Cette section décrit (i) les séries utilisées, (ii) les transformations appliquées avant l'extraction de facteurs et l'estimation MIDAS, et (iii) les principales difficultés rencontrées lors de la mise en cohérence de données multi-fréquences. Notre protocole suit l'approche de l'article d'Andreou, Ghysels et Kourtellos (2013) (MIDAS + facteurs financiers journaliers), tout en l'adaptant à un univers de données plus restreint que celui du papier.

4.1 Sources, fréquences et périmètres d'échantillon

L'ensemble des séries financières et macroéconomiques est extrait de Bloomberg (export Excel), puis restructuré en un panel *date* × *ticker*. Les données financières sont observées à fréquence journalière (jours de bourse), tandis que la variable cible est trimestrielle. Notre base contient 47 tickers journaliers, couvrant la période du 02/01/1986 au 31/12/2025 (14 631 dates), soit une matrice 14 631 × 47 (dates × tickers). Les séries couvrent plusieurs classes d'actifs (taux/souverains, crédit/spreads, actions/volatilité, change, matières premières), en cohérence avec la typologie du papier. La variable à prévoir est la croissance trimestrielle du PIB réel américain (`GDP CQOQ Index`, en %), disponible sur 161 trimestres du 31/03/1986 au 31/12/2025. Sur l'échantillon, la moyenne est 2.71, l'écart-type 4.22, avec un minimum à -28.0 et un maximum à 34.9, reflétant notamment des épisodes de rupture (crise COVID-19).

Les auteurs travaillent sur deux fenêtres (1986–2008 et 1999–2008) et un univers quotidien beaucoup plus large (jusqu'à ∼991 séries quotidiennes). Notre réplication conserve la logique méthodologique (régression MIDAS, extraction de facteurs journaliers, évaluation pseudo hors-échantillon), mais s'appuie sur un univers de tickers plus réduit et une période étendue jusqu'en en fin 2025 (disponibilité des données), ce qui peut modifier certaines propriétés de stabilité (ruptures structurelles, volatilité extrême, etc.).



(a) Répartition des tickers par classe d'actifs.

(b) Complétude moyenne des données (journalières) dans le temps.

(c) Corrélations croisées (extrait du panel).

FIGURE 1 – Composition, couverture et structure de dépendance de la base Bloomberg utilisée dans l'étude.

4.2 Construction et alignement des séries

Les séries sont alignées sur un index temporel commun (`DatetimeIndex`) et pivotées en format large (colonnes = tickers). Les valeurs infinies sont traitées comme manquantes et aucune interpolation n'est effectuée afin d'éviter l'introduction d'information artificielle. La variable PIB trimestrielle est alignée sur les fins de trimestre. Dans notre implémentation, le PIB trimestriel est reconstruit en prenant la dernière observation disponible de `GDP CQOQ Index` au sein de chaque trimestre (fin de période). Les prédicteurs quotidiens sont ensuite utilisés via des blocs de m jours précédant chaque date trimestrielle (voir §4.3).

4.3 Transformations des données journalières

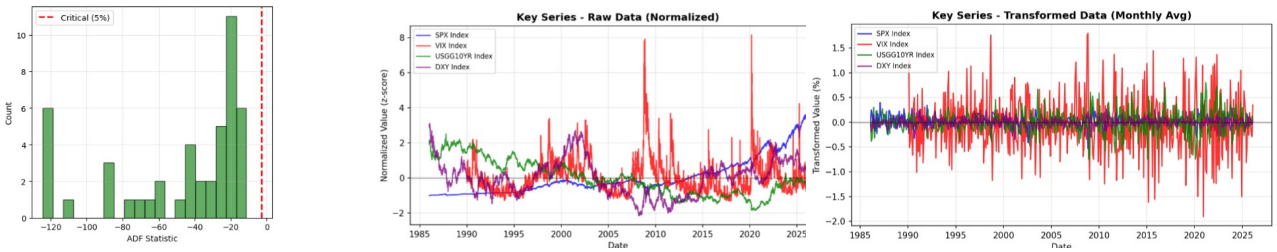
Les séries financières sont transformées en séries stationnaires avant l'extraction factorielle et l'estimation MIDAS, conformément aux pratiques standards en finance empirique et à l'esprit du papier (returns/différences, robustification aux extrêmes). Pour les séries strictement positives (prix/indices), nous utilisons des log-returns $x_t = 100\big(\ln X_t - \ln X_{t-1}\big)$. Pour les séries pouvant être nulles ou négatives (taux, spreads), nous utilisons des premières différences $x_t = 100(X_t - X_{t-1})$. Cette règle évite les problèmes de logarithme et standardise les unités en points de pourcentage.

Data Type	Transformation	Formula	Rationale
Prices/Indices	Log-returns	$r_t = \ln(P_t) - \ln(P_{t-1})$	Prices are I(1), returns are I(0)
Interest Rates	First differences	$\Delta x_t = x_t - x_{t-1}$	Rates can be ≤0, so log impossible
Spreads	First differences	$\Delta x_t = x_t - x_{t-1}$	Already in % but often I(1)
GDP Growth	None (kept in levels)	—	Already a growth rate, stationary

FIGURE 2 – Règles de transformation appliquées aux principales catégories de séries (prix/indices, taux, spreads, PIB).

Après transformation, chaque série est winsorisée aux percentiles 1% et 99% (capage des extrêmes sans suppression d'observations). Ce choix limite l'influence d'épisodes atypiques (crises, erreurs de cotation) tout en préservant la taille d'échantillon.

Deux séries ne sont pas incluses dans l'extraction des facteurs journaliers : (i) la variable cible `GDP CQOQ Index` (fréquence trimestrielle), et (ii) une série macro mensuelle `NFP TCH Index` (fréquence non-daily). Ces séries peuvent être mobilisées séparément comme indicateurs macro (benchmark FAR/FADL), mais ne doivent pas entrer dans une PCA journalière.



(a) Statistiques ADF (seuil 5%). (b) Séries clés (données brutes, normalisées). (c) Séries clés (après transformation).

FIGURE 3 – Diagnostics des transformations : stationnarité (ADF) et illustration sur séries représentatives avant/après traitement.

4.4 Extraction des facteurs financiers journaliers

Nous construisons des facteurs financiers journaliers en appliquant une ACP (PCA) au panel de séries quotidiennes transformées. L'objectif est de résumer l'information commune contenue dans la coupe transversale de séries, tout en évitant une sur-paramétrisation. Avant l'ACP, les séries sont standardisées (centrage-réduction) afin qu'aucune variable ne domine mécaniquement l'extraction des composantes à cause de sa variance.

Soit $X_t \in R^N$ le vecteur des N séries transformées et standardisées au jour t. L'ACP construit des facteurs F_t comme des combinaisons linéaires

$$F_t = W^\top X_t,$$

où les colonnes de W sont choisies pour maximiser la variance expliquée, sous contraintes d'orthogonalité. Dans notre cas, nous retenons $K = 5$ facteurs journaliers $\{DF1, \ldots, DF5\}$.

Concrètement, l'ACP est estimée sur la période 03/01/1998–31/12/2025, ce qui produit une matrice de facteurs de dimension $10\,247 \times 5$. Les facteurs expliquent au total 57.5% de la variance du panel (DF1 : 20.0%, DF2 : 16.5%, DF3 : 8.6%, DF4 : 8.0%, DF5 : 4.4%). Afin d'assurer une extraction stable, seules les séries présentant une couverture minimale (au moins 70% d'observations non manquantes) sont retenues pour la PCA.

Enfin, dans une réplication strictement pseudo hors-échantillon, l'ACP devrait idéalement être recalculée de manière récursive (fenêtre *expanding* ou *rolling*) afin d'éviter toute fuite d'information (*look-ahead bias*). Dans notre pipeline, ce point est traité comme une limite et discuté en §4.6.

4.5 Analyse des facteurs PCA

Afin de vérifier que les facteurs extraits résument bien l'information macro-financière contenue dans notre panel, nous analysons deux éléments : (i) la part de variance expliquée par chaque facteur, et (ii) la corrélation moyenne des facteurs avec les principales classes d'actifs (actions, taux, matières premières, change). Cette étape permet de rapprocher nos facteurs des interprétations usuelles proposées dans l'article de référence, sans préjuger des performances de prévision.

Facteur	**Variance expliquée (%)**	**Corr. avec le PIB (CQOQ)**
DF1	20.0	0.165
DF2	16.5	-0.050
DF3	8.6	-0.034
DF4	8.0	0.067
DF5	4.4	0.186
Cumul (DF1–DF5)	**57.5**	

TABLE 1 – Variance expliquée par les facteurs PCA et corrélation contemporaine avec le PIB.

Le tableau 1 montre que les deux premiers facteurs concentrent une part importante de l'information commune (36.5% de variance expliquée), tandis que les facteurs suivants capturent des dimensions plus spécifiques. La corrélation contemporaine avec le PIB est globalement modérée, ce qui est cohérent avec l'idée que les facteurs financiers ne sont pas des proxies directs de l'activité réelle mais peuvent contenir un signal utile pour la prévision. Dans notre échantillon, DF5 présente la corrélation la plus élevée avec le PIB (0.186).

Classe d'actifs	**DF1**	**DF2**	**DF3**	**DF4**	**DF5**
Equity	0.405	-0.084	0.373	-0.117	-0.024
Fixed Income	0.558	0.252	-0.135	0.187	0.041
Commodities	0.262	-0.247	-0.138	-0.148	0.020
FX	-0.080	-0.339	0.025	0.221	0.016

TABLE 2 – Corrélation moyenne des facteurs avec les séries de chaque classe d'actifs.

Le tableau 2 suggère que DF1 est un facteur « global » : il est positivement corrélé avec les actions, les taux et, dans une moindre mesure, les matières premières. Cela correspond à une composante commune de marché (*risk-on/risk-off*) qui traverse plusieurs classes d'actifs. DF2 apparaît davantage lié au change, avec une corrélation moyenne négative marquée sur la classe FX (-0.339), et une composante liée aux taux (corrélation positive sur Fixed Income). Les facteurs DF3 et DF4 capturent des dimensions plus spécifiques (par exemple une composante equity additionnelle pour DF3 ou une composante taux/change pour DF4), tandis que DF5 présente des corrélations faibles par classe mais reste celui qui est le plus corrélé au PIB dans notre échantillon.

4.6 Difficultés rencontrées et limites des données

La constitution d'un jeu de données exploitable pour MIDAS soulève plusieurs difficultés pratiques. D'abord, le caractère multi-fréquence du problème impose des choix d'alignement entre les données

quotidiennes (jours de bourse) et la cible trimestrielle : il faut fixer une date d'ancrage (fin de trimestre), définir une taille de fenêtre m et décider comment traiter les jours manquants. Dans l'article, m est supposé constant pour simplifier l'exposé ; en pratique, il s'agit plutôt d'un ordre de grandeur (par exemple $m = 63$ jours $\approx$ un trimestre).

Ensuite, les séries n'ont pas toutes la même profondeur historique. Certaines commencent plus tard, d'autres comportent davantage de valeurs manquantes. Cela réduit l'échantillon effectif dès que l'on utilise des fenêtres longues (par exemple $m = 189$ ou $m = 252$) et rend plus délicate la comparaison des performances si tous les modèles ne produisent pas des prévisions sur exactement les mêmes dates.

Un autre point important concerne la présence d'épisodes extrêmes. Les crises (2008–2009, COVID-19) entraînent des mouvements très atypiques dans la variable cible et peuvent provoquer des sur-réactions lorsque les pondérations deviennent très concentrées. La winsorisation permet de limiter l'impact des valeurs extrêmes sur les prédicteurs, mais elle ne "corrige" pas les ruptures observées sur le PIB lui-même.

Par ailleurs, le travail de *mapping* des tickers Bloomberg n'est pas toujours direct. Certaines séries exigent des ajustements (renommages, proxys, séries substituts) pour s'approcher au mieux des catégories utilisées dans l'article. Cela introduit une part d'incertitude dans la correspondance (*matching*) entre variables, surtout lorsque la série exacte n'est pas disponible sur la même période ou sous le même code.

Enfin, l'implémentation des pondérations MIDAS (polynôme exponentiel de type Almon) demande une attention particulière à la convention temporelle et à l'alignement de l'information disponible. En particulier, l'introduction d'avances (*leads*) doit respecter une logique strictement temps réel : seules les observations effectivement disponibles à la date d'information doivent être utilisées pour prévoir le trimestre cible. Ces points sont vérifiés dans notre pipeline et discutés dans la section des résultats.

5 Design de réplication et protocole de prévision

Cette section décrit le cadre empirique de la réplication : définition des échantillons et des horizons, spécifications estimées, procédure pseudo hors-échantillon, et critères d'évaluation. L'objectif est de reproduire au plus près l'esprit de l'article, tout en tenant compte des contraintes liées à notre univers de données (nombre de séries plus réduit et période étendue).

5.1 Échantillons, horizons et fenêtre m

La variable cible est la croissance trimestrielle du PIB réel (`GDP CQOQ Index`). Les prédicteurs financiers sont observés à fréquence journalière et intégrés dans des régressions à fréquences mixtes via des blocs de m jours de bourse. Dans la suite, m désigne le nombre d'observations quotidiennes utilisées pour construire la composante haute fréquence au sein d'un trimestre.

Conformément à l'ordre de grandeur retenu dans la littérature MIDAS, nous fixons par défaut $m = 63$ jours (environ un trimestre de bourse). Nous testons également des fenêtres plus longues (par exemple $m = 126$, $m = 189$, $m = 252$) afin d'évaluer la sensibilité des performances au choix de la mémoire intra-trimestrielle. Intuitivement, des fenêtres plus longues augmentent la quantité d'information disponible, mais peuvent aussi introduire du bruit et réduire l'échantillon effectif lorsque certaines séries sont plus courtes.

Les prévisions sont réalisées à différents horizons h (en trimestres) selon les configurations. L'horizon court $h = 1$ correspond à une prévision à un trimestre, tandis que des horizons plus longs (par exemple $h = 4$) permettent d'évaluer la capacité des facteurs financiers à porter un signal plus structurel. Lorsque plusieurs horizons sont étudiés, nous estimons une équation distincte pour chaque h (prévision directe).

5.2 Modèles estimés et benchmarks

Nous comparons plusieurs familles de modèles, organisées autour d'un benchmark autorégressif et de variantes MIDAS.

— Benchmark univarié (AR) : Le point de comparaison principal est un modèle autorégressif trimestriel, typiquement un AR(1) :

$$y_{t+h} = \alpha + \rho\, y_t + u_{t+h},$$

où y_t désigne le PIB trimestriel (croissance) et h l'horizon de prévision. Ce benchmark capture la persistance de la variable cible sans information financière.

— Modèle ADL–MIDAS : Nous estimons des régressions MIDAS augmentées de retards de la variable cible, de la forme :

$$y_{t+h} = \alpha^h + \sum_{j=0}^{p_y-1} \rho_{j+1}^h\, y_{t-j} + \beta^h \sum_{i=0}^{m-1} w_i(\theta^h)\, x_{t,i} + u_{t+h},$$

où $x_{t,i}$ est l'observation quotidienne (transformée) associée au i-ème retard dans la fenêtre, et $w_i(\theta^h)$ est une fonction de pondération paramétrique de faible dimension (polynôme exponentiel d'Almon). Dans notre réplication, les prédicteurs x peuvent être des séries individuelles ou des facteurs journaliers issus d'une ACP (DF1–DF5).

— MIDAS avec avances (leads) : Pour le nowcasting intra-trimestre, nous considérons une extension avec avances, qui incorpore les observations quotidiennes disponibles au cours du trimestre courant (par exemple 2 mois d'informations disponibles avant la fin du trimestre). Cette variante permet d'étudier l'apport de l'information financière la plus récente dans une logique de mise à jour en temps réel.

— Combinaison de prévisions : Lorsque plusieurs modèles/facteurs sont disponibles, nous considérons également une combinaison de prévisions (pondération en fonction de la performance passée), afin de réduire la dépendance à un facteur particulier et d'améliorer la robustesse hors-échantillon.

Le choix du nombre de retards trimestriels p_y peut être fixé a priori (souvent faible) ou sélectionné sur un critère d'information (AIC/BIC) dans une grille restreinte, afin d'éviter la sur-paramétrisation.

Combinaison de prévisions (MSFE pondérée) : Lorsque plusieurs prédicteurs/modèles sont disponibles, nous construisons une prévision combinée à l'horizon h comme une moyenne pondérée des M prévisions individuelles :

$$\widehat{Y}^{Q,h}_{c,t+h|t} = \sum_{i=1}^{M} \omega^h_{i,t}\, \widehat{Y}^{Q,h}_{i,t+h|t},$$

où $\omega^h_{i,t}$ désigne le poids (variable dans le temps) attribué au modèle i.

Construction des poids : Les poids sont déterminés à partir d'une MSFE actualisée (*discounted MSFE*) :

$$\omega^h_{i,t} = \frac{\left(\lambda^{-1}_{i,t}\right)^{\kappa}}{\sum_{j=1}^{M}\left(\lambda^{-1}_{j,t}\right)^{\kappa}}, \qquad \lambda_{i,t} = \sum_{\tau=T_0}^{t-h} \delta^{\,t-h-\tau}\left(Y^Q_{\tau+h} - \widehat{Y}^{Q,h}_{i,\tau+h|\tau}\right)^2.$$

Dans cette construction, $\delta \in (0,1)$ est le facteur d'actualisation (dans nos tests, $\delta = 0.9$) : plus δ est élevé, plus les erreurs récentes pèsent dans $\lambda_{i,t}$. Le paramètre $\kappa > 0$ (ici $\kappa = 2$) contrôle la concentration des poids : plus κ est grand, plus la combinaison favorise les modèles ayant les plus faibles erreurs passées.

5.3 Procédure pseudo hors-échantillon (expanding/rolling)

Les performances sont évaluées dans un cadre pseudo hors-échantillon. Une date de départ hors-échantillon T_0 est fixée : les modèles sont estimés sur un échantillon initial (*in-sample*), puis les prévisions sont générées de manière séquentielle au fur et à mesure que de nouvelles observations deviennent disponibles.

Nous utilisons principalement une fenêtre *expanding* (récursive) : à chaque date t, l'échantillon d'estimation comprend toutes les observations disponibles de l'origine jusqu'à t. Les paramètres sont ré-estimés à chaque itération, puis une prévision $\widehat{y}_{t+h|t}$ est produite. Cette procédure se rapproche d'un exercice de prévision en temps réel, tout en restant reproductible.

Dans certaines analyses de sensibilité, une fenêtre *rolling* (taille fixe) peut également être utilisée pour tester la stabilité des coefficients lorsque des ruptures structurelles sont susceptibles d'affecter la relation entre facteurs financiers et activité réelle. Dans ce cas, seuls les L derniers trimestres sont conservés à chaque itération.

5.4 Métriques d'évaluation

La qualité prédictive est mesurée à l'aide de l'erreur de prévision $e_{t+h} = y_{t+h} - \widehat{y}_{t+h|t}$ et de deux métriques standards :

— **RMSFE** (*Root Mean Squared Forecast Error*) :

$$\text{RMSFE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e^2_{t+h}},$$

— **MAE** (*Mean Absolute Error*) :

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |e_{t+h}|.$$

Lorsque nous reportons des gains de performance, nous les exprimons relativement au benchmark AR(1). Par exemple, un gain en RMSFE peut être résumé sous forme de pourcentage :

$$\text{Gain}(\%) = 100 \times \left(1 - \frac{\text{RMSFE}_{\text{modèle}}}{\text{RMSFE}_{\text{AR}}}\right).$$

Un gain positif indique que le modèle améliore la performance par rapport à AR(1).

Enfin, afin de garantir une comparaison équitable entre modèles, les métriques sont idéalement calculées sur un ensemble commun de dates de prévision (même période hors-échantillon). Cette précaution est importante lorsque certains modèles produisent moins de prévisions en raison de données manquantes ou de fenêtres m plus longues.

6 Résultats de réplication

Cette section présente (i) la réplication des principaux tableaux empiriques de l'article (Tables 1, 3 et 5) sur l'échantillon *Long sample* et (ii) une application de la méthodologie sur une période récente (2024–2025). Les tableaux de réplication reportent des ratios de RMSFE relativement au benchmark Random Walk (RW) : une valeur inférieure à 1 indique une amélioration par rapport à RW. Pour la période récente, nous reportons des RMSFE en niveau ainsi que des RMSFE relatifs à RW.

6.1 Réplication des résultats du papier : *Long sample* (1988–2008)

Table 1 – Modèles sans avances (no leads) : : RW est reporté en RMSFE absolu (2.69 à $h = 1$ et 3.18 à $h = 4$) et tous les autres modèles sont exprimés en ratios à RW.

Modèle	**Long** $h = 1$	**Long** $h = 4$
Univariate models		
RW (absolute)	2.69	3.18
AR	1.01	0.91
Models with macro data		
FAR (CFNAI)	0.91	0.90
Models with financial data (5 DF)		
ADL (5 DF)	1.09	1.12
ADL-MIDAS (5 DF) `ADL-MIDAS(`$J_X^D = 0$`)`	1.11	1.11
Models with macro and financial data (CFNAI, 5 DF)		
FADL (CFNAI, 5 DF)	0.96	1.12
FADL-MIDAS (CFNAI, 5 DF) `FADL-MIDAS(`$J_X^D = 0$`)`	1.07	0.86

TABLE 3 – Table 1 (réplication) – RMSFE sans avances (Long sample) : RW est reporté en RMSFE absolu ; les autres valeurs sont des ratios à RW (valeurs < 1 : amélioration vs RW).

Table 3 – Modèles avec avances (leads) : $J_X^D = 2$ correspond à deux mois de leads quotidiens ; $J_M = 1$ correspond à un mois de lead pour l'indicateur macro mensuel.

Modèle	**Long** $h = 1$	**Long** $h = 4$
Models with leads in daily financial data		
ADL-MIDAS($J_X^D = 2$)	0.97	0.87
FADL-MIDAS($J_X^D = 2$)	0.77	0.73
Models with leads in monthly macro and daily financial data		
FADL-MIDAS($J_M = 1, J_X^D = 2$)	0.93	0.81
Models with leads in monthly macro data		
FAR($J_M = 1$)	0.87	0.73
FADL($J_M = 1$)	0.90	0.88
FADL-MIDAS($J_M = 1, J_X^D = 0$)	0.97	0.82

TABLE 4 – Table 3 (réplication) – RMSFE avec avances (Long sample). Valeurs reportées en ratios à RW.

Table 5 – Cas ADS (indice macro quotidien) : La Table 5 : ADS est un indicateur macro quotidien ; les valeurs sont des ratios RMSFE vs RW.

Modèle	$h=1$	$h=4$
ADL-MIDAS($J^D_{ADS}=2$)	0.57	0.48
FADL-MIDAS($J_M=1, J^D_{ADS}=2$)	0.60	0.52

TABLE 5 – Table 5 (réplication) – Comparaisons avec ADS. Valeurs reportées en ratios à RW.

6.2 Réplication des résultats du papier : *Short sample* (1999–2008)

Le *Short sample* correspond à l'exercice sur une fenêtre plus récente et plus courte, ce qui permet de tester la robustesse des résultats lorsque l'échantillon d'estimation est plus limité. Comme pour le *Long sample*, RW est reporté en RMSFE absolu et les autres valeurs sont des ratios à RW (valeurs < 1 : amélioration vs RW).

Modèle	**Short** $h=1$	**Short** $h=4$
Univariate models		
RW (absolute)	3.46	4.66
AR	1.13	1.00
Models with macro data		
FAR (CFNAI)	0.94	0.98
Models with financial data (5 DF)		
ADL (5 DF)	1.20	1.14
ADL-MIDAS (5 DF) `ADL-MIDAS(`$J^D_X=0$`)`	1.24	1.13
Models with macro and financial data (CFNAI, 5 DF)		
FADL (CFNAI, 5 DF)	1.00	1.14
FADL-MIDAS (CFNAI, 5 DF) `FADL-MIDAS(`$J^D_X=0$`)`	1.02	1.00

TABLE 6 – Table 1 (réplication) – RMSFE sans avances (*Short sample*). RW est reporté en RMSFE absolu ; les autres valeurs sont des ratios à RW (valeurs < 1 : amélioration vs RW).

Modèle	**Short** $h=1$	**Short** $h=4$
Models with leads in daily financial data		
ADL-MIDAS($J^D_X=2$)	0.94	0.89
FADL-MIDAS($J^D_X=2$)	0.70	0.62
Models with leads in monthly macro and daily financial data		
FADL-MIDAS($J_M=1, J^D_X=2$)	0.86	0.82
Models with leads in monthly macro data		
FAR($J_M=1$)	0.84	0.72
FADL($J_M=1$)	0.92	0.86
FADL-MIDAS($J_M=1, J^D_X=0$)	0.92	0.86

TABLE 7 – Table 3 (réplication) – RMSFE avec avances (*Short sample*). Valeurs reportées en ratios à RW.

Modèle	$h=1$	$h=4$
ADL-MIDAS($J^D_{ADS}=2$)	0.56	0.42
FADL-MIDAS($J_M=1, J^D_{ADS}=2$)	0.60	0.46

TABLE 8 – Table 5 (réplication) – Comparaisons avec ADS (*Short sample*). Valeurs reportées en ratios à RW.

Sur le Short sample, les modèles basés uniquement sur les facteurs financiers (ADL, ADL-MIDAS avec 5 DF) ne battent pas RW dans la spécification sans avances. En revanche, l'introduction d'avances améliore nettement les performances, en particulier pour FADL-MIDAS($J^D_X=2$), et l'indicateur ADS reste très performant, surtout à horizon long ($h=4$).

6.3 Analyse hors-échantillon récente (2024–2025)

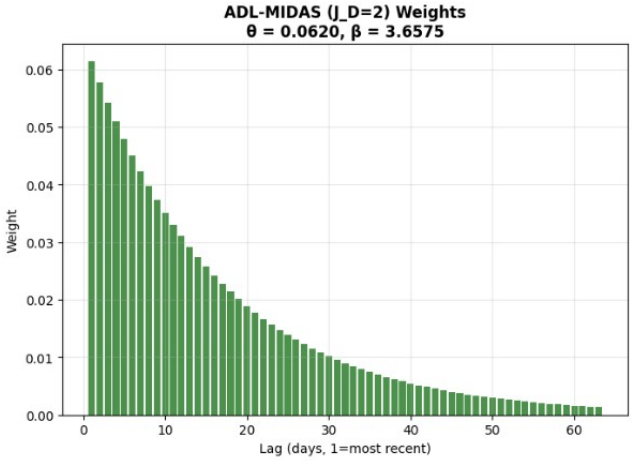
L'échantillon d'entraînement couvre 2020Q1–2023Q4 et l'évaluation hors-échantillon porte sur 2024Q1–2025Q4. Dans la pratique, certaines spécifications ne génèrent pas de prévision pour le tout premier trimestre, car le nowcast est construit à une date d'information (fin du 2^{e} mois du trimestre). Les RMSFE reportées ci-dessous correspondent donc aux prévisions effectivement produites sur la fenêtre récente.

Le tableau 9 présente le classement des modèles sur l'échantillon récent : quatre modèles battent RW (ADL(flat), FAR(CFNAI), AR, FADL($J_M=1$)), tandis que les variantes MIDAS testées sous-performent sur cette période.

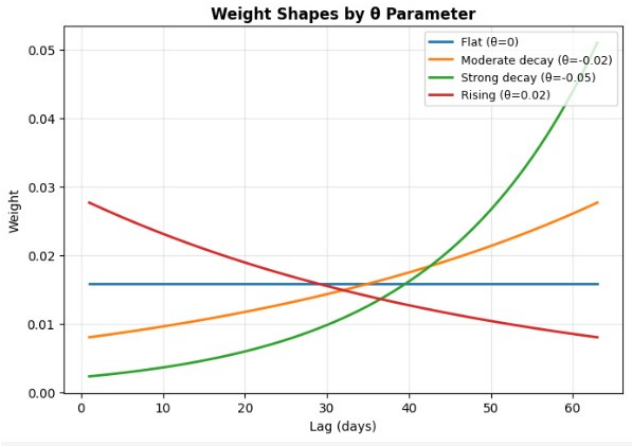
Rang	**Modèle**	**RMSFE**	**Rel. à RW**	**vs RW**
1	ADL(flat)	0.9511	0.449	+55.1%
2	FAR(CFNAI)	1.8016	0.851	+14.9%
3	AR	1.8844	0.890	+11.0%
4	FADL($J_M=1$)	2.1136	0.998	+0.2%
5	RW	2.1172	1.000	Baseline
6	FAR($J_M=1$)	2.2609	1.068	-6.8%
7	ADL-MIDAS($J_D=2$)	3.2400	1.530	-53.0%
8	FADL-MIDAS($J_M=1, J_D=2$)	3.6527	1.725	-72.5%
9	FADL-MIDAS	3.7317	1.763	-76.3%

TABLE 9 – Période récente (OOS 2024–2025, $h=1$) – Classement des modèles.

6.4 Illustrations graphiques (période récente)



(a) Poids MIDAS estimés (ADL–MIDAS, $J_D = 2$).



(b) Formes des pondérations selon θ.

FIGURE 4 – Diagnostics sur la forme des pondérations MIDAS (période récente).

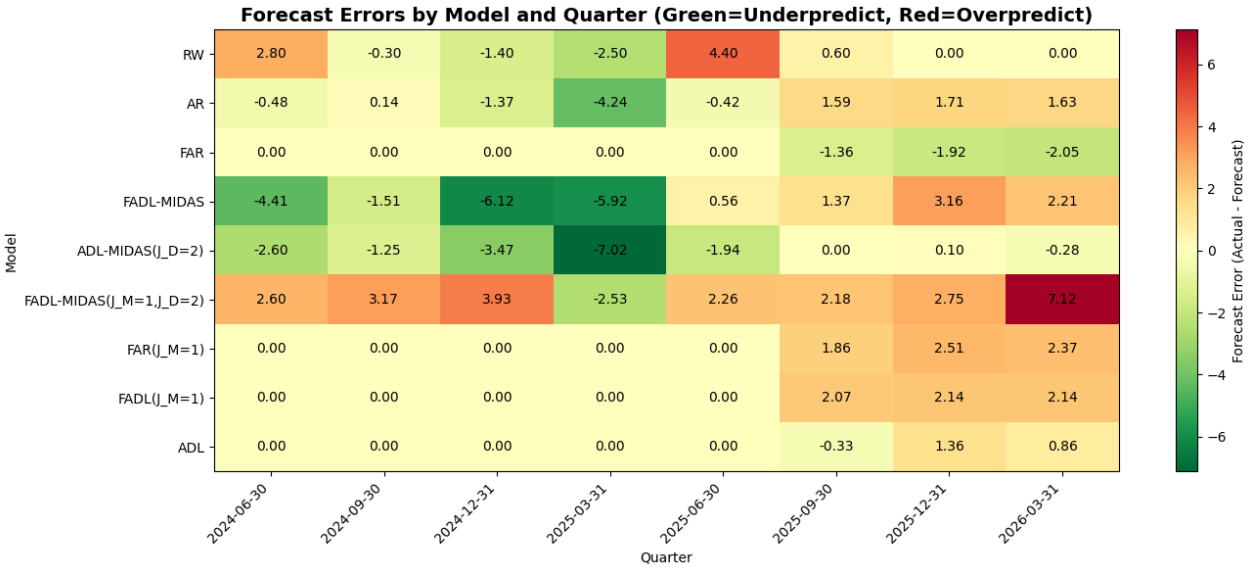


FIGURE 5 – Erreurs de prévision par modèle et par trimestre (2024–2025).

7 Extension : MIDAS à deux paramètres β (lags vs leads)

Cette section propose une extension simple du modèle MIDAS avec avances. Dans la spécification standard (Andreou, Ghysels et Kourtellos, 2013), un *seul* paramètre β détermine la forme des pondérations appliquées à l'ensemble du bloc quotidien utilisé, c'est-à-dire à la fois aux données passées (lags) et aux données de nowcast (leads). Notre extension consiste à dissocier ce paramètre en deux : β_{lag} pour pondérer le bloc des lags, et β_{lead} pour pondérer le bloc des leads. L'objectif est de relâcher une contrainte potentiellement forte, tout en restant dans une approche parcimonieuse (on n'ajoute qu'un paramètre).

7.1 Motivation et intuition économique

Dans notre réplication, le nowcasting est construit à une date fixée à la fin du deuxième mois du trimestre cible. Concrètement, cela revient à utiliser environ deux mois de données quotidiennes "en avance", soit environ 42–44 jours de bourse. Cette convention respecte le calendrier d'information : la prévision pour le trimestre t est construite uniquement à partir des observations disponibles à la date d'information, sans utiliser de données postérieures (pas de *look-ahead bias*).

L'intuition derrière l'extension est la suivante. Si l'on impose un *unique* β pour l'ensemble du bloc (lags + leads), on force la même dynamique de pondération pour l'information passée et pour l'information intra-trimestre. Or les leads correspondent au début du trimestre cible : ce sont les données les plus récentes au moment où la prévision est faite, mais elles peuvent se retrouver relativement sous-pondérées si la forme des poids est principalement dictée par le bloc lag. Autoriser $\beta_{\text{lag}} \neq \beta_{\text{lead}}$ permet donc de laisser au modèle la possibilité de traiter différemment l'information historique et l'information de nowcast, ce qui nous paraît plus cohérent économiquement dans un exercice de nowcasting.

7.2 Spécification du modèle

Nous séparons la composante quotidienne en deux agrégats MIDAS :
— un agrégat lag construit sur $m = 63$ jours (environ un trimestre de bourse) ;
— un agrégat lead construit sur $m_L \approx 42$ jours (environ deux mois) au sein du trimestre cible.
À horizon $h = 1$, la régression s'écrit :

$$y_{t+1} = \alpha + \rho\, y_t + \beta_{\text{lag}} \sum_{k=1}^{m} B(k;\theta_{\text{lag}})\, x_{t-k} + \beta_{\text{lead}} \sum_{j=1}^{m_L} B(j;\theta_{\text{lead}})\, x_{t+j} + \varepsilon_{t+1}, \tag{10}$$

où $B(\cdot;\theta)$ est une pondération exponentielle de type Almon (normalisée) :

$$B(k;\theta) = \frac{\exp(\theta k)}{\sum_{\ell} \exp(\theta \ell)}. \tag{11}$$

Le modèle single-β est un cas particulier lorsque $\beta_{\text{lag}} = \beta_{\text{lead}}$ (et, implicitement, lorsque l'on impose la même dynamique de pondération aux deux blocs).

7.3 Protocole d'évaluation (OOS) et configuration

L'exercice est conduit en pseudo hors-échantillon (OOS) avec estimation récursive. La fenêtre de données utilisée pour cette extension couvre la période 2015–2025, et l'évaluation hors-échantillon démarre en 2024 :Q1. Le modèle est estimé à chaque date de prévision sur l'information disponible jusqu'à la date d'information (fin du 2$^{\text{e}}$ mois du trimestre cible), puis une prévision de y_{t+1} est produite. Sur cette configuration, nous obtenons 8 prévisions OOS.

Les paramètres retenus sont : $h = 1$, $p_y = 1$, $m = 63$, et 2 mois de leads.

7.4 Résultats : performance et comparaison au single-β

Sur la période d'évaluation (OOS 2024–2025), le modèle two-β obtient une RMSFE de 2.7267, soit un ratio de 1.288 relativement au Random Walk (dégradation de 28.8%). En moyenne sur les prévisions, les paramètres estimés sont :

$$\theta_{\text{lag}} = 0.0266 \qquad \theta_{\text{lead}} = 0.0176 \qquad \beta_{\text{lag}} = 2.5252, \qquad \beta_{\text{lead}} = 3.6708.$$

Ces estimations suggèrent des dynamiques de pondération différentes entre passé et nowcast, ce qui est précisément l'objectif de l'extension.

La comparaison au modèle single-β avec 2 mois de leads est néanmoins instructive : bien que le two-β ne batte pas RW sur cette fenêtre récente, il améliore le modèle MIDAS standard (single-β). Le tableau 10 montre une baisse de RMSFE de 3.2400 à 2.7267, soit une amélioration relative de 15.8% par rapport au single-β.

Modèle	**RMSFE**	**RMSFE / RW**	**vs RW**
Two-β MIDAS ($J_D = 2$)	2.7267	1.288	-28.8%
Single-β MIDAS ($J_D = 2$)	3.2400	1.530	-53.0%

TABLE 10 – Comparaison two-β vs single-β (OOS 2024–2025, leads = 2 mois).

Le tableau 11 reporte les prévisions two-β et les erreurs associées sur les 8 trimestres hors-échantillon en Détail par trimestre.

Trimestre	**Réalisé**	**Prévu (two-β)**	**Erreur**	$\boldsymbol{\beta}_{\text{lag}}$	$\boldsymbol{\beta}_{\text{lead}}$
2024-Q2	3.600	5.595	-1.995	0.0145	0.5085
2024-Q3	3.300	0.947	2.353	0.0149	0.5205
2024-Q4	1.900	-0.251	2.151	0.0142	0.5358
2025-Q1	-0.600	4.171	-4.771	0.0130	0.5295
2025-Q2	3.800	0.932	2.868	0.0124	0.5295
2025-Q3	4.400	4.296	0.104	0.0130	0.5251
2025-Q4	4.400	1.734	2.666	0.0130	0.5248
				RMSFE : 2.7267	

TABLE 11 – Prévisions hors-échantillon du modèle two-β (OOS 2024–2025, $h = 1$). Erreur = Réalisé – Prévu.

7.5 Application aux échantillons du papier (Long/Short sample)

En complément de l'exercice sur la période récente, nous appliquons la même extension two-β aux deux échantillons de référence de l'article (*Long sample* et *Short sample*). L'objectif est de vérifier si le gain observé par rapport à la spécification single-β se retrouve sur les périodes du papier. Comme précédemment, les performances sont reportées en RMSFE (niveau), en RMSFE relatif à RW et en gain/perte par rapport à RW.

Échantillon	**Modèle**	**RMSFE**	**Rel. to RW**	**vs RW**
Long sample	Two-β MIDAS ($J_D = 2$)	2.3966	1.132	-13.2%
Long sample	Single-β MIDAS ($J_D = 2$)	3.2400	1.530	-53.0%
Short sample	Two-β MIDAS ($J_D = 2$)	2.8268	1.335	-33.5%
Short sample	Single-β MIDAS ($J_D = 2$)	3.2400	1.530	-53.0%

TABLE 12 – Extension two-β vs single-β sur les échantillons *Long* et *Short* du papier (leads = 2 mois). Les valeurs reportées sont la RMSFE en niveau, la RMSFE relative à RW et le gain/perte vs RW.

Lecture des résultats. Sur les deux échantillons, l'extension two-β améliore systématiquement la version single-β : la RMSFE passe de 3.2400 à 2.3966 sur le *Long sample* (amélioration relative de 26.0%)

et de 3.2400 à 2.8268 sur le *Short sample* (amélioration relative de 12.8%). En revanche, les deux spécifications restent au-dessus de RW (RMSFE/RW > 1) : l'extension corrige donc une partie du problème de pondération (lags vs leads) sans suffire à rendre la stratégie globalement supérieure au benchmark RW sur ces configurations.

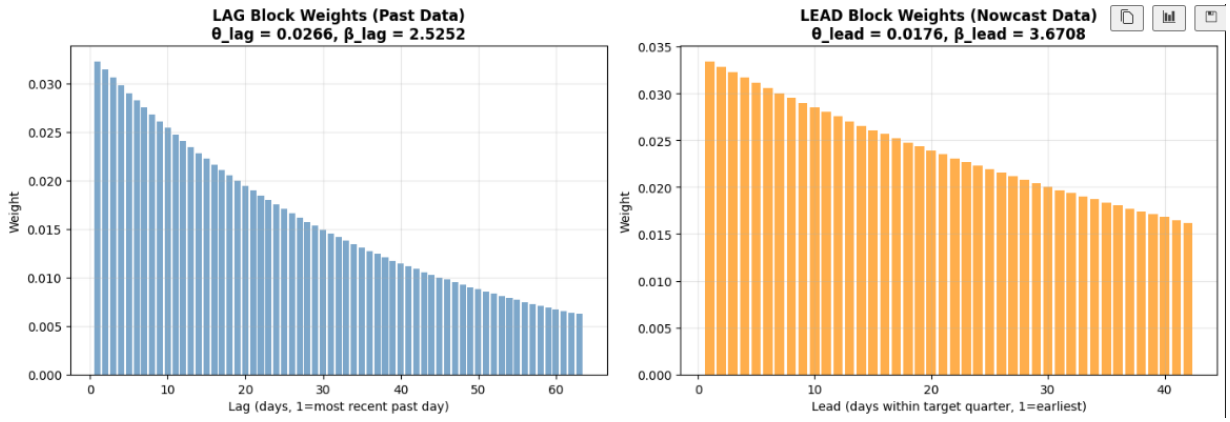


FIGURE 6 – Extension two-β : pondérations estimées sur le bloc des lags (passé) et sur le bloc des leads (nowcast).

La Figure 6 illustre l'apport principal de l'extension : au lieu d'imposer une seule forme de poids sur l'ensemble du bloc (passé + nowcast), le modèle estime deux profils distincts. Le bloc *lag* (à gauche) décroît de manière régulière sur $m = 63$ jours : l'information la plus récente du passé reçoit plus de poids, ce qui correspond à une mémoire décroissante classique. Le bloc *lead* (à droite), construit sur les jours disponibles dans le trimestre cible ($m_L \approx 42$), présente une dynamique propre : les poids ne sont pas contraints à suivre la même décroissance que le passé. Concrètement, cela permet au modèle d'ajuster séparément la façon dont il exploite (i) les retards historiques et (ii) l'information intra-trimestre disponible à la date d'information, ce qui répond directement à la limite de la spécification single-β.

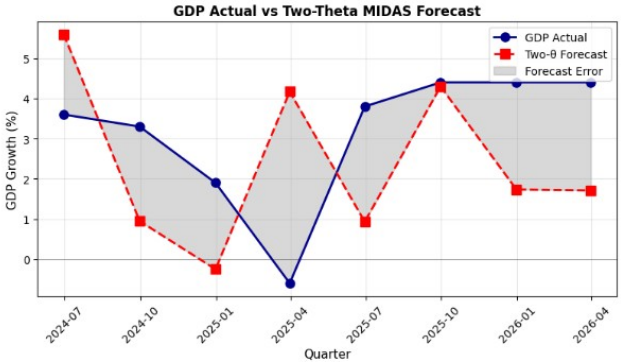
7.6 Positionnement par rapport aux autres modèles (période récente)

Pour situer cette extension dans l'ensemble des modèles testés sur 2024–2025, le tableau 13 reprend le classement complet. L'extension two-β améliore le single-β MIDAS (rang 7 contre rang 8), mais ne surperforme pas les benchmarks simples sur cette fenêtre récente. Ce résultat est cohérent avec l'analyse précedente : sur une période courte et relativement stable, des modèles parcimonieux (ADL(flat), FAR(CFNAI), AR) peuvent dominer des spécifications plus riches, et les modèles MIDAS peuvent souffrir d'une instabilité de la relation entre facteurs et activité réelle.

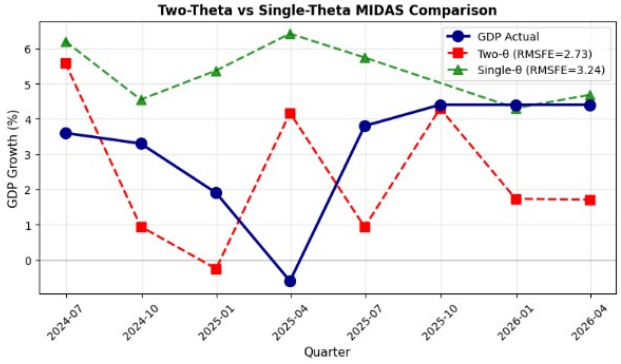
Rang	Modèle	RMSFE	RMSFE / RW	vs RW
1	ADL(flat)	0.9511	0.449	+55.1%
2	FAR(CFNAI)	1.8016	0.851	+14.9%
3	AR	1.8844	0.890	+11.0%
4	FADL($J_M = 1$)	2.1136	0.998	+0.2%
5	RW	2.1172	1.000	Baseline
6	FAR($J_M = 1$)	2.2609	1.068	-6.8%
7	$\star$MIDAS-2β($J_D = 2$)	2.7267	1.288	-28.8%
8	ADL–MIDAS($J_D = 2$)	3.2400	1.530	-53.0%
9	FADL–MIDAS($J_M = 1, J_D = 2$)	3.6527	1.725	-72.5%
10	FADL–MIDAS	3.7317	1.763	-76.3%

TABLE 13 – Classement des modèles sur la période récente (OOS 2024–2025, $h = 1$), incluant l'extension two-β. $\star$ indique notre contribution.

7.7 Illustrations graphiques



(a) Trajectoires : réalisé vs prévision.



(b) Two-β vs single-β.

FIGURE 7 – Visualisations complémentaires du modèle two-β (OOS 2024–2025).

7.8 Discussion et limites

Cette extension augmente la flexibilité du modèle en distinguant explicitement l'information passée et l'information de nowcast, sans introduire d'états latents ni complexifier excessivement l'estimation. Sur la période récente, elle améliore le MIDAS standard avec leads, ce qui suggère que la contrainte single-β peut être trop restrictive.

En revanche, l'extension ne bat pas RW sur 2024–2025. Deux raisons principales semblent plausibles. D'abord, l'échantillon hors-échantillon est très court (8 trimestres), ce qui rend la RMSFE sensible à quelques erreurs importantes. Ensuite, la période post-COVID peut correspondre à un régime où la relation entre facteurs financiers et PIB trimestriel est moins stable, ce qui pénalise les modèles plus structurés.

8 Conclusion

Cette réplication confirme l'intérêt du cadre MIDAS proposé par Andreou, Ghysels et Kourtellos (2013) pour intégrer des données quotidiennes dans la prévision de variables macroéconomiques trimestrielles, en particulier lorsque l'on respecte strictement le calendrier d'information et que l'on évalue les modèles hors-échantillon. Sur le *Long sample*, les tableaux reproduits (Tables 1, 3 et 5) mettent en évidence que les gains ne proviennent pas mécaniquement de l'usage de données haute fréquence : ils dépendent fortement de la spécification retenue, de l'introduction d'avances et du choix des indicateurs. Le cas de l'indice quotidien ADS illustre notamment que certains signaux macroéconomiques quotidiens peuvent être très informatifs.

Sur la période récente (2024–2025), les résultats sont plus contrastés : des modèles simples et parcimonieux (ADL plat, FAR/AR) dominent les variantes MIDAS testées. Cette observation suggère que la relation entre facteurs financiers et croissance du PIB peut être moins stable en régime post-COVID et que, sur une fenêtre courte, le classement des modèles est très sensible à quelques erreurs de prévision. Elle rappelle aussi que la performance prédictive dépend du contexte macro-financier et de la taille effective de l'échantillon hors-échantillon.

Enfin, notre extension two-β (pondérations distinctes pour lags et leads) va dans le sens d'un assouplissement naturel de la contrainte single-β. Sur la fenêtre récente, elle améliore le modèle MIDAS standard avec leads, sans toutefois dépasser le benchmark Random Walk. Ce résultat est cohérent avec l'idée que l'extension corrige une limitation structurelle, mais qu'elle ne suffit pas à elle seule à rendre le modèle robuste dans tous les régimes. Des prolongements immédiats seraient (i) de tester l'extension sur des fenêtres plus longues et sur plusieurs horizons, (ii) d'intégrer une sélection systématique de p_y comme dans le protocole principal, et (iii) d'étendre le two-β au cadre FADL–MIDAS afin de combiner explicitement information macro et financière dans un modèle unique.