Online Appendix for "Does High Frequency Social Media Data Improve Forecasts of Low Frequency Consumer Confidence Measures?"*

Steven Lehrer†, Tian Xie‡, and Tao Zeng◊

†Queen's University, NYU Shanghai, and NBER, lehrers@queensu.ca
‡Shanghai University of Finance and Economics, xietian001@hotmail.com
◊Zhejiang University, ztzt6512@gmail.com

July 28, 2019

Abstract

This is the online appendix for Lehrer, Xie, and Zeng (2019). Four sections are included, which provide further details on the data collection as well all of the econometric estimators and machine learning algorithms listed in the main text. The appendix also contains all formal proofs of the econometric theory and additional intuition explaining why the H-MIDAS method yields improvements in forecast accuracy.

JEL classification: C53, E27, G17

Keywords: Forecasting, Social Media, Big Data, Machine Learning, MIDAS

*We would like to thank Jun Yu, Yue Qiu, seminar participants at the IFABS Asia 2017 Ningbo Conference, ShanghaiTech University, Wuhan University, and Xiamen University for comments and suggestions that improved this paper. Xie's research is supported by the Natural Science Foundation of China (71701175) and the Chinese Ministry of Education Project of Humanities and Social Sciences (17YJC790174). The usual caveat applies. Correspondence author: Tian Xie, Rm.424, College of Business, Shanghai University of Finance and Economics, Shanghai, China, 200433. Tel: +8615021784437, email: xietian001@hotmail.com

A Asymptotic Properties of the H-MIDAS Estimator

In this section, we analyze the asymptotic properties of the H-MIDAS estimator. We derive its asymptotic variance for inference purpose and demonstrate that not using our H-MIDAS estimator can introduce bias to the estimation of the coefficients under certain circumstances.

Let $\boldsymbol{\tau} = \left[\beta, \boldsymbol{\theta}^\top\right]^\top$, $\boldsymbol{\theta} = \left[\theta_1, \theta_2, \ldots \theta_{p-1}\right]^\top$, then we have $w_j(\boldsymbol{\theta}) = \theta_j$, if $1 \leq j \leq p-1$, and $w_p(\boldsymbol{\theta}) = 1 - \sum_{k=1}^{p-1} \theta_k$, note that the first order derivative of the weights can be expressed as

$$\frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} = \begin{bmatrix} \boldsymbol{I}_{p-1} \\ -\boldsymbol{\iota}_{p-1}^\top \end{bmatrix},$$

where ι is a column vector with all elements to be 1. If we define $g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau}) = \beta X_t^{new}(\boldsymbol{w}) = \beta \sum_{j=1}^{p} w_j(\theta) \bar{X}_t^{(l_j)} = \sum_{j=1}^{p} \beta w_j(\boldsymbol{\theta}) \frac{1}{l_j} \sum_{i=0}^{l_j-1} X_{t-\frac{i}{m}}$, we have the derivatives of $g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})$ as

$$\frac{\partial g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}} = \begin{bmatrix} \frac{\partial g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})}{\partial \beta} \\ \frac{\partial g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})}{\partial \boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} X_t^{new}(\boldsymbol{w}) \\ \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \beta \tilde{\boldsymbol{X}}_t \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^\top \tilde{\boldsymbol{X}}_t \\ \beta \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \tilde{\boldsymbol{X}}_t \end{bmatrix},$$

and the regression problem can be viewed as

$$Y_t = g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau}) + \epsilon_t$$

which can be estimated by the nonlinear least square method. According to Andreou, Ghysels, and Kourtellos (2010), the estimator of $\boldsymbol{\tau}$, which is denoted by $\hat{\boldsymbol{\tau}}$, has asymptotically distribution

$$\sqrt{T}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2 \left[\mathbb{E}\left(\frac{\partial g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \frac{\partial g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}}^\top\right)\right]^{-1}\right).$$

We derive the asymptotic variance of the estimator of β in the following lemma

Lemma 0 *Suppose $X_{t-\frac{i}{m}}^h$ is an AR(1) process*

$$X_{t-\frac{i}{m}}^h = \rho X_{t-\frac{i-1}{m}}^{(h)} + e_{t-\frac{i}{m}},$$

where $|\rho| \in (0,1)$ is the AR coefficient and the error term $e_{t-\frac{i}{m}} \overset{iid}{\sim} (0, \sigma_e^2)$. The asymptotic variance of the estimated coefficient $\hat{\beta}$ is

$$AVar(\hat{\beta}) = \sigma_e^2 \left[\boldsymbol{w}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \boldsymbol{w} - \boldsymbol{w}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}\right.$$

$$\times \left(\frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right)^{-1} \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \boldsymbol{w} \Bigg]^{-1},$$

where the i^{th} element of $\mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right]$ is

$$Cov\left(\bar{X}_t^{(l_i)}, \bar{X}_t^{(l_s)}\right) = \frac{l_i\left(1-\rho^2\right) + \rho\left(1-\rho^{l_i}\right) A_{s,j}}{l_i l_s \left(1-\rho\right)^2 \left(1-\rho^2\right)}$$

where

$$A_{s,j} \equiv \left(1-\rho^{l_s}-\rho^{l_s-l_j}\right)\left(\rho^{l_s-l_j+1}+1\right) - \rho\left(1-\rho^{l_j}\right) + \rho^{l_s} - 2.$$

Proof of Lemma 0 Since $X^h_{t-\frac{i}{m}}$ is an AR(1) process

$$X^h_{t-\frac{i}{m}} = \rho X^h_{t-\frac{i-1}{m}} + e_{t-\frac{i}{m}},$$

with an error term $e_{t-\frac{i}{m}} \overset{iid}{\sim} \left(0, \sigma_e^2\right)$, following the derivatives of $g(\tilde{\boldsymbol{X}}_t, \boldsymbol{\tau})$, the general formula for estimating the asymptotic variance of the estimator of β is

$$\begin{aligned} \text{AVar}\left(\hat{\beta}\right) &= \sigma_e^2 \Bigg[\boldsymbol{w}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \boldsymbol{w} - \boldsymbol{w}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \\ &\quad \times \left(\frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \right)^{-1} \frac{\partial \boldsymbol{w}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}^\top \mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right] \boldsymbol{w} \Bigg]^{-1}. \end{aligned}$$

However, we still need to derive the i^{th} element of $\mathbb{E}\left[\tilde{\boldsymbol{X}}_t \tilde{\boldsymbol{X}}_t^\top\right]$ from the above equation.

The regressors $X^h_{t-\frac{i}{m}}$ can be expressed as

$$X^h_{t-\frac{i}{m}} = \rho X^h_{t-\frac{i-1}{m}} + e_{t-\frac{i}{m}} = \sum_{j=0}^{m-i-2} \rho^j e_{t-\frac{i+j}{m}} + \rho^{m-i-1} X^h_{t-\frac{m-1}{m}}.$$

We can write $\bar{X}_t^{(l_i)}$ as follows

$$\begin{aligned} \bar{X}_t^{(l_i)} &= \frac{1}{l_i}\sum_{i=0}^{l_i-1} X^h_{t-\frac{i}{m}} = \frac{1}{l_i}\sum_{i=0}^{l_i-1}\left(\sum_{j=0}^{m-i-2}\rho^j e_{t-\frac{i+j}{m}} + \rho^{m-i-1} X^h_{t-\frac{m-1}{m}}\right) \\ &= \frac{1}{l_i}\sum_{i=0}^{l_i-1}\sum_{j=0}^{m-i-2}\rho^j e_{t-\frac{i+j}{m}} + \left(\rho^{m-l_i}\frac{1}{l_i}\sum_{i=0}^{l_i-1}\rho^{l_i-i-1}\right) X^h_{t-\frac{m-1}{m}} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{l_i}\sum_{i=0}^{l_i-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}} + \frac{1}{l_i}\cdot\frac{1-\rho^{l_i}}{1-\rho}\sum_{k=l_i}^{m-2}\rho^{k-l_i+1}e_{t-\frac{k}{m}} + \rho^{m-l_i}\frac{1}{l_i}\cdot\frac{1-\rho^{l_i}}{1-\rho}X^{(h)}_{t-\frac{m-1}{m}} \\
&\equiv A^{(l_i)} \quad + \quad B^{(l_i)} \quad + \quad C^{(l_i)},
\end{aligned}
$$

where we define $A^{(l_i)}$, $B^{(l_i)}$, and $C^{(l_i)}$ accordingly to simplify the complicated polynomial.

In order to compute the covariance $\text{Cov}\left(\bar{X}_t^{(l_i)}, \bar{X}_t^{(l_s)}\right)$, where $l_s > l_i$, we first decompose $\bar{X}_t^{(l_i)}$ and $\bar{X}_t^{(l_s)}$. For the $\bar{X}_t^{(l_i)}$ term, we decompose the middle component, $B^{(l_i)}$, and obtain

$$
\bar{X}_t^{(l_i)} = A^{(l_i)} + \left(\frac{1}{l_i}\cdot\frac{1-\rho^{l_i}}{1-\rho}\sum_{k=l_i}^{l_s-1}\rho^{k-l_i+1}e_{t-\frac{k}{m}} + \frac{1}{l_i}\cdot\frac{1-\rho^{l_i}}{1-\rho}\sum_{k=l_s}^{m-2}\rho^{k-l_i+1}e_{t-\frac{k}{m}}\right) + C^{(l_i)}.
$$

For the $\bar{X}_t^{(l_s)}$ term, we decompose the $A^{(l_s)}$ component as

$$
\bar{X}_t^{(l_s)} = \left(\frac{1}{l_s}\sum_{i=0}^{l_i-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}} + \frac{1}{l_s}\sum_{k=l_i}^{l_s-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}}\right) + B^{(l_s)} + C^{(l_s)}.
$$

Then, it is straightforward to show that

$$
\begin{aligned}
\text{Cov}\left(\bar{X}_t^{(l_i)}, \bar{X}_t^{(l_s)}\right) &= \frac{1}{l_i l_s}\cdot\underbrace{\text{Var}\left(\sum_{i=0}^{l_i-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}}\right)}_{\equiv D} + \frac{1}{l_i l_s}\cdot\underbrace{\frac{1-\rho^{l_i}}{1-\rho}\text{Cov}\left(\sum_{k=l_i}^{l_s-1}\rho^{k-l_i+1}e_{t-\frac{k}{m}}, \sum_{k=l_i}^{l_s-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}}\right)}_{\equiv E} \\
&\quad + \frac{1}{l_i l_s}\cdot\underbrace{\text{Cov}\left(\frac{1-\rho^{l_i}}{1-\rho}\sum_{k=l_s}^{m-2}\rho^{k-l_i+1}e_{t-\frac{k}{m}}, \frac{1-\rho^{l_s}}{1-\rho}\sum_{k=l_s}^{m-2}\rho^{k-l_s+1}e_{t-\frac{k}{m}}\right)}_{\equiv F} \\
&\quad + \frac{1}{l_i l_s}\cdot\underbrace{\frac{\rho^{2m-l_i-l_s}\left(1-\rho^{l_i}\right)\left(1-\rho^{l_s}\right)}{(1-\rho)^2}\text{Var}\left(X^h_{t-\frac{m-1}{m}}\right)}_{\equiv G} \\
&= \frac{1}{l_i l_s}\cdot(D \quad + \quad E \quad + \quad F \quad + \quad G),
\end{aligned}
$$

where D, E, F, and G represent the associated terms.

Since the D term can be explicitly written as

$$
\begin{aligned}
\text{Var}\left(\sum_{i=0}^{l_i-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}}\right) &= \frac{\sigma_e^2}{(1-\rho)^2}\left[\sum_{i=0}^{l_i-1}\left(1-2\rho^{k+1}+\rho^{2(k+1)}\right)\right] \\
&= \frac{\sigma_e^2}{(1-\rho)^2}\left[l_i - 2\frac{\rho\left(1-\rho^{l_i}\right)}{1-\rho} + \frac{\rho^2\left(1-\rho^{2l_i}\right)}{1-\rho^2}\right] \\
&= \frac{l_i\left(1-\rho^2\right) - 2\rho\left(1-\rho^{l_i}\right)(1+\rho) + \rho^2\left(1-\rho^{2l_i}\right)}{(1-\rho)^2(1-\rho^2)}\sigma_e^2
\end{aligned}
$$

$$= \frac{l_i\left(1-\rho^2\right)-2\rho-\rho^2+2\rho^{l_i+1}+2\rho^{l_i+2}-\rho^{2l_i+2}}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2, \quad \text{(A1)}$$

the E term can be expressed as

$$\begin{aligned}
& \frac{1-\rho^{l_i}}{1-\rho}\text{Cov}\left(\sum_{k=l_i}^{l_s-1}\rho^{k-l_i+1}e_{t-\frac{k}{m}}, \sum_{k=l_i}^{l_s-1}\frac{1-\rho^{k+1}}{1-\rho}e_{t-\frac{k}{m}}\right) \\
= & \frac{1}{\left(1-\rho\right)^2}\sum_{k=l_i}^{l_s-1}\left(1-\rho^{k+1}\right)\rho^{k-l_i+1}\sigma_e^2 \\
= & \frac{1}{1-\rho}\left(\rho\frac{1-\rho^{l_s-l_i}}{1-\rho}-\rho^2\frac{1-\rho^{2(l_s-l_i)}}{1-\rho^2}\right)\sigma_e^2 \\
= & \frac{\left(1-\rho^{l_i}\right)\left[\rho\left(1-\rho^{l_s-l_i}\right)\left(1+\rho\right)-\rho^2\left(1-\rho^{2(l_s-l_i)}\right)\right]}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2 \\
= & \frac{\rho-\rho^{l_s-l_i+1}-\rho^{l_s-l_i+2}+\rho^{2(l_s-l_i)+2}-\rho^{l_i+1}+\rho^{l_s+1}+\rho^{l_s+2}-\rho^{2l_s-l_i+2}}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2 \quad \text{(A2)}
\end{aligned}$$

the F term can be shown as

$$\begin{aligned}
& \text{Cov}\left(\frac{1-\rho^{l_i}}{1-\rho}\sum_{k=l_s}^{m-2}\rho^{k-l_i+1}e_{t-\frac{k}{m}}, \frac{1-\rho^{l_s}}{1-\rho}\sum_{k=l_s}^{m-2}\rho^{k-l_s+1}e_{t-\frac{k}{m}}\right) \\
= & \frac{\left(1-\rho^{l_i}\right)\left(1-\rho^{l_s}\right)}{\left(1-\rho\right)^2}\sum_{k=l_s}^{m-2}\rho^{2k-l_i-l_s+2}\sigma_e^2 \\
= & \frac{\left(1-\rho^{l_i}-\rho^{l_s}+\rho^{l_i+l_s}\right)\rho^{-l_i+l_s+2}\left(1-\rho^{2(m-l_s-1)}\right)}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2 \\
= & \frac{\left(\rho^{-l_i+l_s+2}-\rho^{l_s+2}-\rho^{2l_s-l_i+2}+\rho^{2l_s+2}\right)\left(1-\rho^{2(m-l_s-1)}\right)}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2 \\
= & \frac{\rho^{-l_i+l_s+2}-\rho^{l_s+2}-\rho^{2l_s-l_i+2}+\rho^{2l_s+2}-\rho^{2m-l_s-l_i}+\rho^{2m-l_s}+\rho^{2m-l_i}-\rho^{2m}}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2, \quad \text{(A3)}
\end{aligned}$$

and the G term is simply

$$\begin{aligned}
\frac{\rho^{2m-l_i-l_s}\left(1-\rho^{l_i}\right)\left(1-\rho^{l_s}\right)}{\left(1-\rho\right)^2}\text{Var}\left(X^h_{t-\frac{m-1}{m}}\right) &= \frac{\rho^{2m-l_i-l_s}\left(1-\rho^{l_i}\right)\left(1-\rho^{l_s}\right)}{\left(1-\rho\right)^2}\frac{\sigma_e^2}{1-\rho^2} \\
&= \frac{\rho^{2m-l_i-l_s}-\rho^{2m-l_s}-\rho^{2m-l_i}+\rho^{2m}}{\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2, \quad \text{(A4)}
\end{aligned}$$

combining the results from Equations (A1), (A2), (A3) and (A4), we have

$$\begin{aligned}
&\operatorname{Cov}\left(\bar{X}_t^{(l_i)}, \bar{X}_t^{(l_s)}\right) \\
=\ & \frac{l_i\left(1-\rho^2\right)-\rho-\rho^2+\rho^{2l_s+2}-\rho^{l_s-l_i+1}+\rho^{2(l_s-l_i)+2}+\rho^{l_s+1}-2\rho^{2l_s-l_i+2}+\rho^{l_i+1}+2\rho^{l_i+2}-\rho^{2l_i+2}}{l_i l_s\left(1-\rho\right)^2\left(1-\rho^2\right)}\sigma_e^2 \\
=\ & \frac{l_i\left(1-\rho^2\right)+\rho\left(1-\rho^{l_i}\right)\left[-1-\rho^{l_s-l_i}-\rho\left(1-\rho^{l_i}\right)+\left(1-\rho^{l_s}-\rho^{l_s-l_i}\right)\rho^{l_s-l_i+1}\right]}{l_i l_s\left(1-\rho\right)^2\left(1-\rho^2\right)} \\
=\ & \frac{l_i\left(1-\rho^2\right)+\rho\left(1-\rho^{l_i}\right)A_{s,j}}{l_i l_s\left(1-\rho\right)^2\left(1-\rho^2\right)},
\end{aligned}$$

where

$$A_{s,j} \equiv \left(1-\rho^{l_s}-\rho^{l_s-l_j}\right)\left(\rho^{l_s-l_j+1}+1\right)-\rho\left(1-\rho^{l_j}\right)+\rho^{l_s}-2.$$

This completes the proof. ∎

Remark 1 Lemma 0 and the following Lemma 1 both assume a dynamic autoregressive data generating process. The high frequency data $X^h_{t-\frac{i}{m}}$ in our exercise is the USSI variable, which quantifies the consumers' hourly sentiment change. Psychologically speaking, people's past sentiment usually affects his/her current sentiment. Therefore, the dynamic data generating process assumption is more reasonable than the conventional i.i.d. assumption in the MIDAS literature.

In line with Andreou et al. (2010), we define the aggregate regressor based on flat weights as X_t^A which is $\bar{X}_t^{(l_p)}$ in our case. Following Andreou et al. (2010), the regression function can be decomposed as the combination of an equal weight component X_t^A and a non-equal weight component X_t^B:

$$\begin{aligned}
Y_t &= \beta X_t^{new}+\epsilon_t = \beta\tilde{\boldsymbol{X}}_t^{\top}\boldsymbol{w}+\epsilon_t = \beta\sum_{j=1}^{p} w_j\bar{X}_t^{(l_j)}+\epsilon_t \\
&= \beta\sum_{j=1}^{p-1} w_j\bar{X}_t^{(l_j)}+\beta\left(w_p-1\right)\bar{X}_t^{(l_p)}+\beta X_t^A+\epsilon_t \\
&= \beta\sum_{j=1}^{p-1} w_j\bar{X}_t^{(l_j)}-\beta\sum_{j=1}^{p-1} w_j\bar{X}_t^{(l_p)}+\beta X_t^A+\epsilon_t \\
&= \beta X_t^B+\beta X_t^A+\epsilon_t, \qquad \text{(A5)}
\end{aligned}$$

where $X_t^B \equiv \sum_{j=1}^{p-1} w_j\left(\bar{X}_t^{(l_j)}-\bar{X}_t^{(l_p)}\right)$. As shown in Lemma 1, omitting X_t^B can introduce bias to the estimation of β.

Lemma 1 (Extended Version) *Suppose $X^h_{t-\frac{i}{m}}$ is an AR(1) process*

$$X^h_{t-\frac{i}{m}} = \rho X^h_{t-\frac{i-1}{m}} + e_{t-\frac{i}{m}},$$

where $|\rho| \in (0,1)$ is the AR coefficient, and consider the H-MIDAS regression model in (A5). Then, the simple averaging estimator that omits the non-equal weight component X^B_t from Model (A5) can introduce the asymptotic bias $ABias\left(\hat{\beta},\beta\right) = \gamma\beta$ to the coefficient β, where

$$\gamma = \sum_{j=1}^{p-1} \frac{w_j}{l_j}\rho\left(\frac{A_{p,j}l_p\left(1-\rho^{l_j}\right)+2l_j\left(1-\rho^{l_p}\right)}{B}\right)$$

is the bias coefficient with

$$A_{p,j} \equiv \left(1-\rho^{l_p}-\rho^{l_p-l_j}\right)\left(\rho^{l_p-l_j+1}+1\right)-\rho\left(1-\rho^{l_j}\right)+\rho^{l_p}-2$$

and

$$B \equiv \frac{l_p\left(1-\rho^2\right)-2\rho+2\rho^{l_p+1}}{\rho}.$$

Proof of Lemma 1 Following the definition of omitted variable bias, we know that

$$\gamma = \frac{\text{Cov}\left(X^A_t, X^B_t\right)}{\text{Var}\left(X^A_t\right)}.$$

We first derive the covariance of X^A_t and X^B_t.

$$\begin{aligned}
&\text{Cov}\left(X^A_t, X^B_t\right)\\
=\ &\text{Cov}\left(\bar{X}^{(l_p)}_t, \sum_{j=1}^{p-1} w_j\left(\bar{X}^{(l_j)}_t - \bar{X}^{(l_p)}_t\right)\right) = \sum_{j=1}^{p-1} w_j \text{Cov}\left(\bar{X}^{(l_p)}_t, \bar{X}^{(l_j)}_t - \bar{X}^{(l_p)}_t\right)\\
=\ &\sum_{j=1}^{p-1} w_j \text{Cov}\left(\bar{X}^{(l_p)}_t, \bar{X}^{(l_j)}_t\right) - \sum_{j=1}^{p-1} w_j \text{Var}\left(\bar{X}^{(l_p)}_t\right)\\
=\ &\sum_{j=1}^{p-1} w_j \text{Cov}\left(\bar{X}^{(l_p)}_t, \bar{X}^{(l_j)}_t\right) - (1-w_p)\,\text{Var}\left(\bar{X}^{(l_p)}_t\right)\\
=\ &\sigma^2_e \sum_{j=1}^{p-1} w_j \left(\frac{l_p\rho\left(1-\rho^{l_i}\right)\left[-1-\rho^{l_p-l_i}-\rho\left(1-\rho^{l_i}\right)+\left(1-\rho^{l_p}-\rho^{l_p-l_i}\right)\rho^{l_p-l_i+1}\right]+2l_i\rho-2l_i\rho^{l_p+1}}{l_i l_p^2\left(1-\rho\right)^2\left(1-\rho^2\right)}\right).
\end{aligned}$$

Following the result in Lemma 0, we have

$$\text{Var}(X^A_t) = \frac{l_p\left(1-\rho^2\right)-2\rho+2\rho^{l_p+1}}{l_p^2\left(1-\rho\right)^2\left(1-\rho^2\right).}$$

Therefore,

$$\begin{aligned}
\gamma &= \frac{\mathrm{Cov}\left(X_t^A, X_t^B\right)}{\mathrm{Var}\left(X_t^A\right)} \\
&= \sum_{j=1}^{p-1} w_j \frac{l_p\rho\left(1-\rho^{l_j}\right)\left[-1-\rho^{l_p-l_j}-\rho\left(1-\rho^{l_j}\right)+\left(1-\rho^{l_p}-\rho^{l_p-l_j}\right)\rho^{l_p-l_j+1}\right]+2l_j\rho-2l_j\rho^{l_p+1}}{l_j l_p^2\left(1-\rho\right)^2\left(1-\rho^2\right)} \\
&\qquad \times \frac{l_p^2\left(1-\rho\right)^2\left(1-\rho^2\right)}{l_p\left(1-\rho^2\right)-2\rho+2\rho^{l_p+1}} \\
&= \sum_{j=1}^{p-1} w_j \left(\frac{l_p\rho\left(1-\rho^{l_j}\right)\left[-1-\rho^{l_p-l_j}-\rho\left(1-\rho^{l_j}\right)+\left(1-\rho^{l_p}-\rho^{l_p-l_j}\right)\rho^{l_p-l_j+1}\right]+2l_j\rho-2l_j\rho^{l_p+1}}{l_j\left[l_p\left(1-\rho^2\right)-2\rho+2\rho^{l_p+1}\right]}\right) \\
&= \sum_{j=1}^{p-1} \frac{w_j}{l_j}\rho\left(\frac{A_{p,j}l_p\left(1-\rho^{l_j}\right)+2l_j\left(1-\rho^{l_p}\right)}{B}\right),
\end{aligned}$$

where

$$A_{p,j} = \left(1-\rho^{l_p}-\rho^{l_p-l_j}\right)\left(\rho^{l_p-l_j+1}+1\right)-\rho\left(1-\rho^{l_j}\right)+\rho^{l_p}-2,$$

and

$$B = \frac{l_p\left(1-\rho^2\right)-2\rho+2\rho^{l_p+1}}{\rho}.$$

This completes the proof. ∎

Remark 2 Lemma 1 states that converting high frequency series to low frequency series using simple averaging is a special case of our H-MIDAS. Moreover, this special case is biased since the non-equal weight component X_t^B is omitted by simple averaging.

B Detailed Description of the Forecasting Techniques

B.1 GUM, AIC, and PMA

Researchers who ignore model uncertainty implicitly assume their selected model is the "true" one that generated the data $(y_t, \boldsymbol{X}_t) : t = 1, ..., n$, where y_t and $\boldsymbol{X}_t = [x_{t1}, x_{t2}, ...]$ are real-valued. We assume the data generating process for an outcome y_t is given as

$$y_t = \mu_t + \epsilon_t, \tag{A6}$$

where $\mu_t = \sum_{j=1}^{\infty} \beta_j x_{tj}$, $\mathbb{E}(\epsilon_t | \boldsymbol{X}_t) = 0$ and $\mathbb{E}(\epsilon_t^2 | \boldsymbol{X}_t) = \sigma^2$.

For researchers who admit ignorance of the true data generating process, they generally select one model from a sequence of linear approximation models $m = 1, 2, ..., M$. An approximation model m using k^h regressors belonging to $\boldsymbol{X}_t$ such that

$$y_t = \sum_{j=1}^{k^h} \beta_j^h x_{tj}^h + \epsilon_t^h \quad \text{for } i = 1, ..., n, \tag{A7}$$

where β_j^h is a coefficient in model m and x_{tj}^h is a regressor in model m. Approximation models can be either nested or non-nested and model averaging approaches first involve solving for the smoothing weights across the set of approximation models based on a specific criterion. We assume that there are K regressors in total among all the potential models. The general unrestricted model (GUM) is like a kitchen sink that consists of all the K regressors we consider in explaining y_t. All potential models are nested within GUM.

Formally, the DGP (A6) and approximation model (A7) can be represented in matrix forms: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ and $\boldsymbol{y} = \boldsymbol{X}^h \boldsymbol{\beta}^h + \boldsymbol{\epsilon}^h$, where $\boldsymbol{y}$ is $n \times 1$, $\boldsymbol{\mu}$ is $n \times 1$, $\boldsymbol{X}^h$ is $n \times k^h$ with the tj^{th} element being x_{tj}^h, $\boldsymbol{\beta}^h$ is $k^h \times 1$ and $\boldsymbol{\epsilon}^h$ is the error term for model m. For an approximation model m, the least squares estimate of $\boldsymbol{\mu}$ from model m can be written as $\hat{\boldsymbol{\mu}}^h = \boldsymbol{P}^h \boldsymbol{y}$, where $\boldsymbol{P}^h$ is a projection matrix.

Among all the model selection methods, the most widely used of these is probably the Akaike information criterion, or AIC by Akaike (1973). There are many versions of AIC, the one we considered is the following

$$\text{AIC}^h = n \log(\text{SSR}^h) + 2k^h, \tag{A8}$$

where SSR^h is the sum of squared residuals from approximation model m.[1] We choose the model with the lowest AIC score.

On the other hand, model averaging simply assume that there is no one specific model that dominates all others. Therefore, it is better to take a weighted average of all the

[1] A more precise description of this version AIC is $n \log(\text{SSR}^h) + 2k^h + C$ with C being a constant term irrelevant to m. Of course, the term C can be conveniently ignored, since only differences in AIC are meaningful for model selection purpose.

potential models. Let $\boldsymbol{w} = \left[w^{(1)}, ..., w^{(M)}\right]^\top$ be a weight vector in the unit simplex in $\mathbb{R}^M$,

$$\boldsymbol{H_M} \equiv \left\{ \boldsymbol{w} \in [0,1]^M : \sum_{m=1}^{M} w^h = 1 \right\},$$

which is a continuous set. We define the model average estimator of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^h \hat{\boldsymbol{\mu}}^h = \sum_{m=1}^{M} w^h \boldsymbol{P}^h \boldsymbol{y}. \tag{A9}$$

By defining the weighted average projection matrix $\boldsymbol{P}(\boldsymbol{w})$ as $\boldsymbol{P}(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^h \boldsymbol{P}^h$, equation (A9) can be simplified to $\boldsymbol{\mu}(\boldsymbol{w}) = \boldsymbol{P}(\boldsymbol{w})\boldsymbol{y}$. Thus, the effective number of parameters to be solved is defined as $k(\boldsymbol{w}) \equiv \sum_{m=1}^{M} w^h k^h$. Note that $k(\boldsymbol{w})$ is not necessarily an integer and is a weighted sum of the k^h.

The prediction model averaging (PMA) estimator of Xie (2015) can be understood as the model averaging analog of the prediction criterion of Amemiya (1980). Following Xie (2015), the vector of empirical weight $\hat{\boldsymbol{w}}$ is the solution to

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \boldsymbol{H_M}}{\arg\min}\, \mathrm{PMA}_n(\boldsymbol{w}) = \underset{\boldsymbol{w} \in \boldsymbol{H_M}}{\arg\min}\, \left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\right)^\top \left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{w})\right) \left(\frac{n + k(\boldsymbol{w})}{n - k(\boldsymbol{w})}\right), \tag{A10}$$

where $\boldsymbol{\mu}(\boldsymbol{w})$ and $k(\boldsymbol{w})$ are defined above. The PMA estimator is asymptotically optimal in the sense of achieving the lowest possible mean square error.

B.2 Tree-based Algorithms

This section consists of four machine learning techniques. The building block is called the regression tree (RT) proposed by Breiman, Friedman, and Stone (1984). Note that the full name of the method is Classification and Regression Trees (CART), in which Classification mostly deals with the categorical response of non-numeric symbols and texts and Regression Trees concentrate purely on quantitative responses variables. Given the numerical nature of our data set, we only consider the second part of CART.

Consider the sample of $\{y_t, \boldsymbol{X}_t\}_{t=1}^n$ as defined in Section B.1. A simple regression will yield a sum of squared residuals, SSR_0. Suppose we can split the original sample into two sub-samples such that $n = n_1 + n_2$. The RT method finds the best split of a sample to minimize the SSR from the two sub-samples. That is, the variable and splitting point are chosen to reduce the residual sum of squares (SSR) as much as possible after the split as compared to before the split. This results in partitioning the data into groups that are as different as possible. We can continue splitting the subsamples until we reach a pre-determined stopping rule. To combat concerns related to overfitting, the tree can be pruned using a cost-complexity criterion. This criterion takes into account the amount of squared error explained by each sub-tree plus a penalty chosen by cross-validation for the

number of terminal nodes in the sub-tree in an attempt to trade-off tree size and over-fitting.

Forecasts from RT involve calculating the average of the outcome for the individuals whose covariates land them in a specific terminal node calculated. Put simply, a local constant model is estimated in each terminal node of the tree to generate a forecast. Lehrer and Xie (2018) argue that in the presence of heteroskedastic data, splits made in the tree are biased to be in regions of high heteroskedasticity at the expense of regions of low heteroskedasticity. They additionally advocate using model averaging in place of the local constant model in each terminal leave, an approach we did not consider since the sample size in our application is quite small.

In general, an RT outperforms conventional regressions as it yields smaller SSR values since it can allow for more general nonlinearities in the covariates. If the data are stationary and ergodic, the RT method also demonstrates better forecasting accuracy. Intuitively, for cross-sectional data, the RT method performs better because it removes heterogeneity problems by splitting the sample into clusters with heterogenous features; for time series data, a good split should coincide with jumps and structure breaks, and therefore, it fits the data to the model better.

We also consider the bootstrap aggregation (BAG) technique developed in Breiman (1996). Unlike the RT method, the BAG method involves a training process where the level of training is predetermined. The BAG algorithm is summarized as below:

(i) Take a random sample with replacement from the data.

(ii) Construct a regression tree.

(iii) Use the regression tree to make forecast, denoted by $\hat{\boldsymbol{y}}$.

(iv) Repeat steps (i) to (iii), $b = 1, ..., B$ times and obtain $\hat{\boldsymbol{y}}^b$ for each b.

(v) Take a simple average of the B forecasts $\hat{\boldsymbol{y}}_{\text{BAG}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\boldsymbol{y}}^b$ and consider the averaged value $\hat{\boldsymbol{y}}_{\text{BAG}}$ as the final forecast.

For most of the part, the more bootstrap samples in the training process, the better the forecast accuracy. However, more bootstrap samples means longer computational time. A balance needs to be found between accuracy and time costs and constraints.

The above algorithm is usually executed for cross-sectional data. When the data is time series (dependent observations), we need to replace step (i) with specific bootstrap methods for time series based on different assumptions. A straightforward way is to bootstrap the residuals instead of observations, in which the residual is estimated using an optimal predictor of the $\boldsymbol{X}_t$'s. For observations following a stationary Markov chain with finite state space, Kulperger and Prakasa Rao (1989) initiated the Markov bootstrap method. If we are not willing to assume a specific structural form for a (stationary and weakly dependent) time series, we can use the moving block bootstrap (MBB) method formulated by Künsch

(1989). Instead of performing single-data resampling, Künsch (1989) advocated the idea of resampling blocks of observations at a time. By retaining the neighboring observations together within each block, the dependence structure of the random variable at short lag distances is preserved. See Kreiss and Lahiri (2012) for a detailed literature review.

Random forest (RF) by Breiman (2001) is a modification of bagging that builds a large collection of de-correlated trees, and then averages them. Similar to BAG, RF also constructs B new trees with (conventional or MBB) bootstrap samples from the original data set. But for RF, as each tree is constructed, we take a random sample (without replacement) of q predictors out of the total $K\,(q < K)$ predictors before each node is split. Such process is repeated for each node. In our application, $\frac{q}{K}$ is set at its default value of $\frac{1}{3}$, and the results are robust to other choices for how many variables to consider to split at each node. Note that if $q = K$, RF is equivalent to BAG. Eventually, we end up with B trees like BAG and the final RF forecast is calculated as the simple average of forecasts from each tree.

The RT method can respond to highly local feature of the data, since it capitalizes on very flexible fitting procedures. An alternative method to accommodate highly local features of the data is to give the observations responsible for the local variation more weight in the fitting process. If a fitting function fits those observations poorly, we reapply that function with extra weight given to the observations poorly fitted. For a large number of trials, we assign relatively more weights to the poorly fitted observations, hence, combine the outputs of many weak fitting functions to produce a powerful committee, as described in Hastie, Tibshirani, and Friedman (2009, Chapter 10).

The procedure we just described is called boosting (BOOST). Although they assemble similarities, the boosting method is fundamentally different from the RF method. Boosting works with the full training sample and all of the predictors. Within each iteration, the poorly fitted observations are given more relative weight, which eventually forces the (poor) fitting functions to evolve in boosting. Moreover, the final output values are a weighted average over a large set of earlier fitting results instead of simple average as in the RF method. In general, since boosting builds trees in a sequential manner, the size of the trees are much shorter to be computationally efficient.

Many of the boosting methods are designed for classification issues, for example, the most popular boosting algorithm `AdaBoost.M1` by Freund and Schapire (1997). In this paper, we consider a simple least squares boosting (LSB) that fits RT ensembles. In line with Hastie et al. (2009, Chapter 8), at every step, the LSB method applies a new learning tree to the difference between the observed response and the aggregated prediction of all trees grown previously. The LSB method fits to minimize MSE.

B.3 Support Vector Machine for Regression

In machine learning, support vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The theory behind SVM is developed in Vapnik (1996). The classic SVM was designed

for classification and a version of SVM for regression, later known as support vector regression (SVR), was proposed in by Drucker, Burges, Kaufman, Smola, and Vapnik (1996). The goal of SVR is to find a function $f(\boldsymbol{X}_t)$ that deviates from y_t by a value no greater than a predetermined e for each observations $\boldsymbol{X}_t$, and at the same time is as flat as possible.

In this paper, we first consider the SVR for the linear regression model (SVR_1). Following Hastie et al. (2009, Chapter 12),

$$y_t = f(\boldsymbol{X}_t) + \epsilon_t = \boldsymbol{X}_t\boldsymbol{\beta} + \epsilon_t = \beta_0 + \tilde{\boldsymbol{X}}_t\boldsymbol{\beta}_1 + \epsilon_t,$$

where $\boldsymbol{X}_t = [1, \tilde{\boldsymbol{X}}_t]$ and $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_1^\top]^\top$. We estimate $\boldsymbol{\beta}$ through the minimization of

$$H(\boldsymbol{\beta}) = \sum_{t=1}^{n} V_e\big(y_t - f(\boldsymbol{X}_t)\big) + \frac{\lambda}{2}\|\boldsymbol{\beta}_1\|^2, \tag{A11}$$

where the loss function

$$V_e(r) = \begin{cases} 0 & \text{if } |r| < e \\ |r| - e & \text{otherwise} \end{cases}$$

is called an e-insensitive error measure that ignores errors of size less than e. As a part of the loss function V_e, the parameter e is usually predetermined. On the other hand, λ is a more traditional regularization parameter, that can be estimated by cross-validation.

Let $\hat{\boldsymbol{\beta}} = \big[\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1^\top\big]^\top$ be the minimizers of function (A11), the solution function can be shown to have the form

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= \sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t)\tilde{\boldsymbol{X}}_t^\top, \\ \hat{f}(\boldsymbol{X}) &= \sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t)\boldsymbol{X}\boldsymbol{X}_t^\top + \hat{\beta}_0\boldsymbol{\iota}_n, \end{aligned}$$

where $\boldsymbol{\iota}_n$ is an $n \times 1$ vector of ones and the parameters $\hat{\boldsymbol{\alpha}}_t$ and $\hat{\boldsymbol{\alpha}}_t^*$ are the nonnegative multiplier of the following Lagrangian equation

$$\min_{\hat{\boldsymbol{\alpha}}_t, \hat{\boldsymbol{\alpha}}_t^*} \; e\sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* + \hat{\boldsymbol{\alpha}}_t) - \sum_{t=1}^{n} y_t(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t) + \frac{1}{2}\sum_{t=1}^{n}\sum_{t'=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t)(\hat{\boldsymbol{\alpha}}_{t'}^* - \hat{\boldsymbol{\alpha}}_{t'})\boldsymbol{X}_t\boldsymbol{X}_{t'}^\top$$

subject to the constraints $0 \le \hat{\boldsymbol{\alpha}}_t^*, \hat{\boldsymbol{\alpha}}_t \le 1/\lambda, \sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t) = 0, \hat{\boldsymbol{\alpha}}_t\hat{\boldsymbol{\alpha}}_t^* = 0$ for all $t = 1, ..., n$. We usually called the non-zero values of $\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t$ for $t = 1, ..., n$ the support vector.

We now extend the above SVR framework from linear regression to nonlinear regression. We approximate the nonlinear regression function $f(\boldsymbol{X}_t)$ in terms of a set of basis function $\{h_m(\tilde{\boldsymbol{X}}_t)\}$ for $m = 1, ..., M$:

$$y_t = f(\boldsymbol{X}_t) + \epsilon_t = \beta_0 + \sum_{m=1}^{M}\beta_m h_m(\tilde{\boldsymbol{X}}_t) + \epsilon_t$$

and we estimate the coefficients $\boldsymbol{\beta} = \left[\beta_0, \beta_1, \ldots, \beta_M\right]^\top$ through the minimization of

$$H(\boldsymbol{\beta}) = \sum_{t=1}^{n} V_\epsilon\big(y_t - f(\boldsymbol{X}_t)\big) + \frac{\lambda}{2}\sum_{m=1}^{M}\beta_m^2. \tag{A12}$$

The solution of (A12) has the form $\hat{f}(\boldsymbol{X}) = \sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t)K(\boldsymbol{X}, \boldsymbol{X}_t) + \hat{\beta}_0\boldsymbol{\iota}_n$ with $\hat{\boldsymbol{\alpha}}_t^*$ and $\hat{\boldsymbol{\alpha}}_t$ being the nonnegative multiplier of the following Lagrangian equation

$$\min_{\hat{\boldsymbol{\alpha}}_t, \hat{\boldsymbol{\alpha}}_t^*} \; e\sum_{t=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* + \hat{\boldsymbol{\alpha}}_t) - \sum_{t=1}^{n} y_t(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t) + \frac{1}{2}\sum_{t=1}^{n}\sum_{t'=1}^{n}(\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t)(\hat{\boldsymbol{\alpha}}_{t'}^* - \hat{\boldsymbol{\alpha}}_{t'})K(\boldsymbol{X}_t, \boldsymbol{X}_{t'})$$

similar to the linear SVR case. In the nonlinear SVR case, a kernel function $K(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \sum_{m=1}^{M} h_m(\boldsymbol{X}_t)h_m(\boldsymbol{X}_{t'})$ is used to replace the inner product of the predictors $\boldsymbol{X}_t\boldsymbol{X}_{t'}^\top$ as in the SVR_1 case. In our paper, we consider the following kernel functions

$$K(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \exp\left(-\|\boldsymbol{X}_t - \boldsymbol{X}_{t'}^\top\|^2\right), \tag{A13}$$

$$K(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) = \left(1 + \boldsymbol{X}_t\boldsymbol{X}_{t'}^\top\right)^p \quad \text{with } p \in \{2, 3, \ldots\}, \tag{A14}$$

in which, we label the associated SVR model as SVR_2 and SVR_3, respectively.

C Comparing Daily USSI with Hourly USSI

In this section, we repeat our main empirical results presented in Section 5 where we use the daily USSI instead of hourly USSI. The daily USSI is a simple weighted average of the hourly USSI, where the weights are the hourly volume of tweets used in the construction of the hourly USSI. The results presented below demonstrate that the main results are robust. We continue to find that it important to incorporate the USSI in forecasting CCI, the superiority of the proposed H-MIDAS estimator relative to other strategies to include the CCI and the general improved forecast accuracy of machine learning strategies relative to econometric estimators.

Table A1: Summary of Statistics of Daily USSI Variables

Variable	Mean	Median	Minimum	Maximum	Std.Dev.
USSI_d	0.3595	0.9323	-18.6684	11.9494	6.3654
USSI'_{new}	0.2962	0.2118	-15.0848	9.1623	5.6586

Figure A1: Estimated Weights for Daily USSI with Specific Lag Index

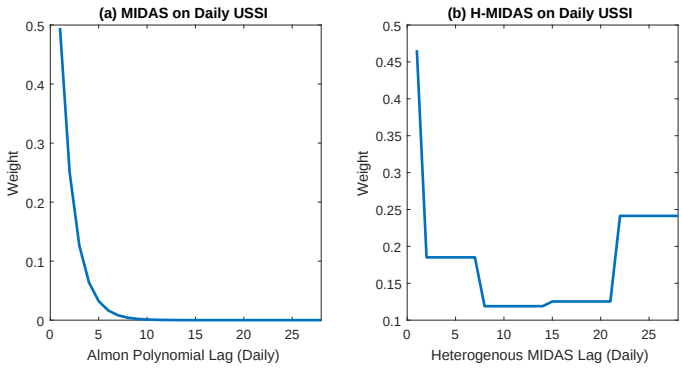

Figure A2: Forecasting Performance of SVR_1 Using Daily USSI as Input Data

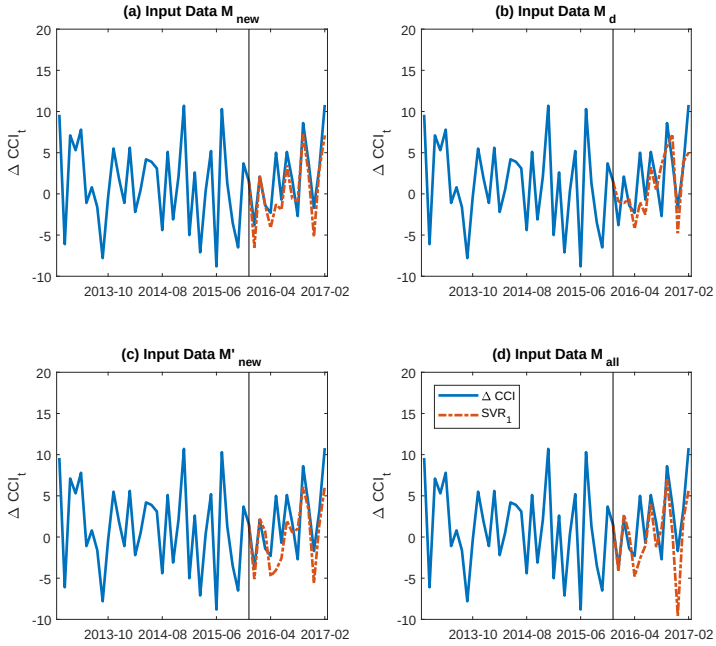

Table A2: Estimation Results with Daily USSI

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Macroeconomic Variable								
MCSI	-	-	-	0.0820	-0.0084	0.0175	0.0853	0.0680
	-	-	-	(0.1992)	(0.1905)	(0.1902)	(0.1997)	(0.2044)
LEI	-	-	-	0.6444	0.3579	0.5687	0.7309	0.5203
	-	-	-	(1.7782)	(1.6852)	(1.6810)	(1.6833)	(1.7329)
UR	-	-	-	-0.2761	-0.2469	-0.3109	-0.3113	-0.2838
	-	-	-	(0.6782)	(0.6440)	(0.6412)	(0.6396)	(0.6534)
SR	-	-	-	-0.7294	-0.5438	-0.4946	-0.3173	-0.3075
	-	-	-	(0.7965)	(0.7639)	(0.7593)	(0.7745)	(0.7953)
CPI	-	-	-	-0.4856	-1.0217	-0.6788	-0.7249	-0.8132
	-	-	-	(1.4254)	(1.3319)	(1.3497)	(1.3470)	(1.3830)
Panel B: Financial Variable								
SP500	-	-	-	-0.0184	0.0019	-0.0050	0.0122	0.0099
	-	-	-	(0.0816)	(0.0778)	(0.0774)	(0.0787)	(0.0809)
VIX	-	-	-	-0.0938	0.1681	0.0738	0.1045	0.2080
	-	-	-	(1.2296)	(1.1706)	(1.1643)	(1.1617)	(1.1906)
USD	-	-	-	3.6418	3.9614*	3.4278	2.8672	2.8250
	-	-	-	(2.4772)	(2.3218)	(2.3431)	(2.3928)	(2.4391)
TS	-	-	-	100.8794	97.7550	96.9602	72.5226	93.5743
	-	-	-	(75.7207)	(72.1231)	(71.5883)	(74.8274)	(80.6411)
Panel C: Big Data Variable								
$USSI_a$	-	-	-	-	-	-	-	-0.1867
	-	-	-	-	-	-	-	(0.3308)
$USSI_h$	-	-	-	-	-	-	-	0.0994
	-	-	-	-	-	-	-	(0.1602)
$USSI_{new}$	0.5479◇	-	-	-	-	-	0.2407	0.2759
	(0.1067)	-	-	-	-	-	(0.2203)	(0.2301)
$USSI_d$	-	0.4262◇	-	0.3740◇	-	0.1683	0.1595	0.0849
	-	(0.0940)	-	(0.1077)	-	(0.1336)	(0.1335)	(0.1982)
$USSI'_{new}$	-	-	0.5167◇	-	0.4675◇	0.3481†	0.1750	0.1686
	-	-	(0.1021)	-	(0.1124)	(0.1464)	(0.2154)	(0.2198)
Panel D: Goodness of Fit								
R^2_c	0.3546	0.2997	0.3480	0.3738	0.4322	0.4549	0.4720	0.4821
$\bar{R}^2$	(0.3412)	(0.2851)	(0.3345)	(0.2132)	(0.2865)	(0.2971)	(0.3007)	(0.2750)

* 10% level of significance.
† 5% level of significance.
◇ 1% level of significance.

Table A3: Daily USSI Forecasting Results Measured by MSFE and MAFE

	GUM	AIC	PMA	RT	BAG	RF	BOOST	SVR_1	SVR_2	SVR_3
Panel A: Mean Squared Forecast Error (MSFE)										
$\mathcal{M}_{new}$	14.1906	15.1128	13.2115	18.7820	12.5265	13.6034	18.5529	**10.3271**◇	19.6891	24.8384
$\mathcal{M}_d$	16.9378	17.5787	16.9245	26.5476	14.7155	15.8912	28.1784	**12.1432**	19.7022	27.4054
$\mathcal{M}'_{new}$	19.6599	17.7067	18.2117	35.3171	14.7824	15.0189	26.7917	**11.9981**	19.7905	39.4239
$\mathcal{M}'_{all}$	18.4680	16.5564	15.9801	26.2613	**12.8549**	12.9826	20.5764	13.9349	19.7615	39.0403
Panel B: Mean Absolute Forecast Error (MAFE)										
$\mathcal{M}_{new}$	2.7981	2.7415	2.6811	3.0674	2.5939	2.7793	3.4075	**2.5035**◇	3.7456	3.7854
$\mathcal{M}_d$	3.3534	3.4959	3.4380	3.9928	**2.8991**	2.9925	3.9965	2.9837	3.7207	4.5058
$\mathcal{M}'_{new}$	3.4512	3.0836	3.0908	4.2269	2.7442	2.7923	4.0550	**2.7254**	3.7539	4.7528
$\mathcal{M}'_{all}$	3.1581	2.7500	2.7642	4.1651	**2.6184**	2.6592	3.8997	2.8705	3.7325	4.5498

Note: numbers with ◇ indicate the best performing methods in each panel.

D More Empirical Results

This section presents the tables associated with a variety of robustness checks that are referenced in the main text. For example, we present a variant of Table 3 where we replicate the results of Table 3 for a 2-period-ahead forecasting exercise, These results are presented in Table A4. Similar to the results in the main text, SVR_1 under $\mathcal{M}_{new}$ has the best forecasting accuracy (indicated by the $\diamond$ symbol) in both panels.

Table A4: Two-period-ahead Forecasting Results Measured by MSFE and MAFE

	GUM	AIC	PMA	RT	BAG	RF	BOOST	SVR_1	SVR_2	SVR_3
Panel A: Mean Squared Forecast Error (MSFE)										
$\mathcal{M}_0$	18.2308	18.3155	19.0210	17.1771	15.2140	**16.1259**	43.0629	26.1228	20.2316	73.9860
$\mathcal{M}_a$	20.2511	18.3089	18.3411	16.1720	16.0738	16.6856	46.2792	**15.0084**	20.3537	45.3306
$\mathcal{M}_m$	24.5094	21.8483	24.9359	25.5146	**17.4539**	16.0889	46.8264	27.6693	20.3768	40.2297
$\mathcal{M}_{new}$	15.3268	14.1511	14.5746	35.9327	17.8323	17.5119	30.8213	**13.1977**$^\diamond$	20.5168	46.1630
$\mathcal{M}_{all}$	26.6746	**15.5066**	16.9475	42.3883	20.2647	18.2060	45.2701	18.5380	20.7128	31.9512
Panel B: Mean Absolute Forecast Error (MAFE)										
$\mathcal{M}_0$	3.8519	3.6987	3.8723	3.6061	3.4920	**3.4551**	5.0305	4.5402	3.7657	6.2458
$\mathcal{M}_a$	4.0248	3.6965	3.8063	**3.3751**	3.5468	3.4833	5.6841	3.5076	3.7593	5.7095
$\mathcal{M}_m$	4.4515	4.2879	4.5581	4.3620	3.5388	**3.3646**	5.8107	4.7262	3.7756	5.4154
$\mathcal{M}_{new}$	3.6007	3.3624	3.4939	4.2865	3.7569	3.6214	5.0424	**3.2378**$^\diamond$	3.7881	5.4488
$\mathcal{M}_{all}$	4.6207	**3.5139**	3.7173	4.9082	3.8264	3.6267	5.3703	3.7880	3.7978	4.8966

Note: numbers with $\diamond$ indicate the best performing methods in each panel.

References

AKAIKE, H. (1973): "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory*, 267–281.

AMEMIYA, T. (1980): "Selection of Regressors," *International Economic Review*, 21, 331–354.

ANDREOU, E., E. GHYSELS, AND A. KOURTELLOS (2010): "Regression models with mixed sampling frequencies," *Journal of Econometrics*, 158, 246–261.

BREIMAN, L. (1996): "Bagging Predictors," *Machine Learning*, 26, 123–140.

——— (2001): "Random Forests," *Machine Learning*, 45, 5–32.

BREIMAN, L., J. FRIEDMAN, AND C. J. STONE (1984): *Classification and Regression Trees*, Chapman and Hall/CRC.

DRUCKER, H., C. J. C. BURGES, L. KAUFMAN, A. J. SMOLA, AND V. VAPNIK (1996): "Support Vector Regression Machines," in *Advances in Neural Information Processing Systems 9*, ed. by M. C. Mozer, M. I. Jordan, and T. Petsche, MIT Press, 155–161.

FREUND, Y. AND R. E. SCHAPIRE (1997): "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119 – 139.

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2009): *The Elements of Statistical Learning:*, Springer Series in Statistics, New York, NY, USA: Springer New York Inc.

KREISS, J.-P. AND S. N. LAHIRI (2012): "Bootstrap Methods for Time Series," in *Time Series Analysis: Methods and Applications, Volume 30*, ed. by T. S. Rao, S. S. Rao, and C. Rao, North Holland, chap. 1, 3–26.

KULPERGER, R. J. AND B. L. S. PRAKASA RAO (1989): "Bootstrapping a Finite State Markov Chain," *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 51, 178–191.

KÜNSCH, H. R. (1989): "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.

VAPNIK, V. N. (1996): *The Nature of Statistical Learning Theory*, New York, NY, USA: Springer-Verlag New York, Inc.

XIE, T. (2015): "Prediction Model Averaging Estimator," *Economics Letters*, 131, 5–8.