# 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**

Social media data presents challenges for forecasters since one must convert text into data and deal with issues related to these measures being collected at different frequencies and volumes than traditional financial data. In this paper, we use a deep learning algorithm to measure sentiment within Twitter messages on an hourly basis and introduce a new method to undertake MIDAS that allows for a weaker discounting of historical data that is well-suited for this new data source. To evaluate the performance of approach relative to alternative MIDAS strategies, we conduct an out of sample forecasting exercise for the consumer confidence index with both traditional econometric strategies and machine learning algorithms. Irrespective of the estimator used to conduct forecasts, our results show that (i) including consumer sentiment measures from Twitter greatly improves forecast accuracy, and (ii) there are substantial gains from our proposed MIDAS procedure relative to common alternatives.

**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

# 1 Introduction

Substantial progress has been made in the machine learning literature on quickly converting text to data, generating real time information on social media content. Yet, there remains substantial speculation on whether data created from online social media content provides valuable insights.[1] Two challenges persist that limit the use of this data in both financial and macroeconomic forecasting exercises. First, from the prospective of a practitioner, the potential value of social media content in forecasting stock market performance is likely tied to our understanding of what information it may capture. Without this interpretation, concerns regarding the generalizability of the social media measure may emerge. Second, from an econometric perspective, how one should incorporate this new data which arrives at different frequencies, asynchronously and may exhibit substantial parameter instability due to the time-varying population of social media users in forecasting exercises remains an open question.

In this study, we address these two challenges by exploring the benefits of incorporating an aggregate measure of social media sentiment, the Wall Street Journal-IHS Markit U.S. Sentiment Index (USSI) in forecasting the conference board consumer confidence index (CCI).[2] The CCI is reported regularly in the financial press and is a variable that has been empirically found to have significant impacts on behavior of financial markets. The likely importance of the CCI likely relates to one of the key arguments in behavioral finance which postulates that change in sentiment can profoundly affect people's behavior and decision making. Until 2013, many Wall Street firms willingly paid an extra subscription fee to Thomson Reuters to gain access to monthly consumer confidence data a full two seconds earlier than the rest of its subscribers at 9:54:58 a.m., as opposed to 9:55:00 ex-

[1] A growing body of research makes claims that this data can improve the performance of high-frequency trading algorithms. For example, Mishne and Glance (2006) proposed using Blogger sentiment to predict movie sales; Bollen, Mao, and Zheng (2011) use data from only 19 days and reach the conclusion that Twitter mood predicts the stock market; Karabulut (2013) showed that the stock market activity can also be predicted by measures extracted from Facebook messages.

[2] Briefly, each hour the USSI uses a deep learning algorithm developed in Felbo, Mislove, ogaard, Rahwan, and Lehmann (2017) to analyze a random sample of 10% of all Twitter messages to measure the national real-time mood, as well as subgroups defined by state or gender. Further details are provided in Section 2.

actly. Thus, this new information was clearly valuable and there is strong industry interest in improving forecasts of the CCI.

Data timing presents a serious challenge in using hourly measures of the USSI to forecast the monthly CCI, which is measured at a much lower frequency.[3] To forecast the CCI requires the analyst to convert hourly USSI measure to a monthly aggregate measure. To develop such an aggregate measure Ghysels, Santa-Clara, and Valkanov (2004) propose a data-driven process coined mixed-data sampling (MIDAS) and shows that it outperforms simple averaging. The MIDAS technique computes a weighted average that generally places a larger weight on the most recent observations. MIDAS was not developed for social media sentiment measures such as the hourly USSI that differs from other financial and macroeconomic variables used to forecast CCI by displaying significant asymmetric response to current events that cause large jumps in the sentiment levels that may have an important impact on dynamics of consumer behavior.[4]

In this paper, we propose a new method to assign weights with MIDAS that allows for heterogeneous effects (henceforth, H-MIDAS) of different high frequency observations on the low frequency dependent variable. This flexibility in how weights are constructed reduces concerns from using conventional MIDAS methods that struggle with parameter instability that may reflect jumps, which can be problematic if the frequency mismatch is severe. Further, we prove that the simple averaging estimator introduces asymptotic bias to the coefficient compared with H-MIDAS.

Our empirical application uses both econometric strategies and machine learning algorithms to ascertain whether incorporating an aggregated measure of very high-frequency social media data can create a more lucrative forecast of the CCI.[5] Our main finding is that

[3] Social media data can be collected and analyzed on a second by second basis. At very high frequencies there is substantial temporal volatility in social media data. As such, we focus on the hourly USSI measure that has social media sentiment appear as a highly persistent process with a long memory decay.

[4] The asymmetry arises in part since there are different populations posting Twitter messages during the standard work-day versus late at night.

[5] In a highly cited paper, O'Connor et al. (2010) report that the correlation between Twitter sentiment from the population and the Gallup Poll of consumer confidence is strong and approximately 0.8. This study simply measures sentiment as the ratio positive versus negative messages on a day and then correlates a moving average of these daily measures with a monthly measure of consumer confidence. Our study presents a significant advance by i) using a lower frequency of social media data, ii) measuring sentiment

incorporating social media sentiment can significantly improve forecast accuracy. This result contributes to a rapidly growing empirical literature on the value of social media in financial econometric applications,[6] that we additionally contribute to by providing a new data driven method to aggregate measures of high frequency social media data.

Further, we find that there are also significant improvement in forecasting accuracy once our proposed H-MIDAS procedure is applied to other high frequency financial and macroeconomic variables that are incorporated in the forecasting model. This evidence is suggestive that allowing for more general forms of heterogeneity in the weights used to undertake MIDAS that can vary across explanatory variables is empirically important.

This paper is organized as follows. In the next section, we describe the data used to conduct forecasts as well as how both the consumer confidence index and social media sentiment are measured. Section 3 provides an overview of different strategies including our proposed H-MIDAS that is designed to incorporate high frequency social media data in forecast models for low frequency measures. Section 4 details the out of sample forecasting exercise that evaluates alternative approaches to undertake MIDAS and contrasts econometric estimators with machine learning algorithms. The empirical results are presented and discussed in Section 5. We find that (i) including consumer sentiment measures from Twitter greatly improves forecast accuracy; (ii) there are substantial gains from the new H-MIDAS procedure relative to common alternatives; and (iii) improvements in forecast accuracy from using machine learning approaches relative to econometric strategies. We conclude by discussing the merits and trade-offs researchers face when incorporating social media data in forecasting models and suggesting directions for future research.

from social media, iii) flexibly handle mixed frequencies and iv) considering multivariate relationships with both econometric and machine learning methods rather than reporting a bivariate correlation.

[6] For example, Brown and Cliff (2004) present significant evidence of the importance of sentiment in measuring U.S. stock market returns. Lemmon and Portniaguina (2006) discuss the connection between consumer confidence and asset prices. Stambaugh, Yu, and Yuan (2012) and Stambaugh, Yu, and Yuan (2014) study the predictive power of investor sentiment for anomaly returns. Baele, Bekaert, and Inghelbrecht (2010) investigate sentiment and the time-series relationships between government bond and stock market returns, while Baker and Wurgler (2012) reveal that sentiment connects the cross-section of stock returns with government bonds. Other papers explore how sentiment affects general financing patterns including Chan, Durand, Khuu, and Smales (2017), García (2013), Mclean and Zhao (2014) among others.

# 2 Data Description

In this study, we forecast the Conference Board's Consumer Confidence Index (CCI), arguably the most well-known and followed measure of U. S. consumer confidence. The CCI is considered to be a major predictor of stock market performance since it is hypothesized to approximate the level confidence on future economy. Since 1967, the CCI has been calculated monthly and is the average response to five specific questions contained within a broader survey of consumer attitudes and expectations.[7] Two of the questions focus on the present labor market and the remaining three questions probe respondents about expected changes in business conditions, job availability and respondents' nominal income over the next six months. Since social media data from Twitter is only recently available, we only use data from January 2013 to March 2017.[8]

To forecast the CCI we consider standard predictors including macroeconomic variables, financial variables, and the big data variables. The macroeconomic variables describes the macro-level economic environment that economic theories often postulate would affect one's consumption behavior. Macroeconomic variables are usually reported on a monthly basis, which is the same frequency as CCI. The financial variables measure the overall performance of the financial markets from various perspectives. In finance studies, CCI is considered as a major predictor that approximates the general public confidence on future economy. Our forecasting models consider the inverse or this relationship and financial variables in the current period are used to forecast future CCI values.

For the big data variables, we use Twitter data from 2013-01-01 to 2017-03-22 to calculate consumer sentiment at both daily and the hourly level.[9] We use the identical

[7] The University of Michigan's Consumer Sentiment Index is another well-known study that measures consumer confidence using five slightly different questions. The surveys also differ in the sample size (CCI is much larger) and how the responses are collected (phone vs. mail responses for the CCI). In this study, we follow the practice of each of the four financial forecasters who use this sentiment index as an explanatory variable to forecast the CCI; rather than the converse. This is likely due to the timing of the survey release since the CCI is released on the last Tuesday of each month at 10am EST, whereas preliminary results from the University of Michigan arrives in mid-month.

[8] Expanding the data may lead to challenges from the emergence of bots. That said, our results are robust to smaller time periods within this sampling frame.

[9] Our focus is using the hourly measure since the daily measure is weighted by the volume of tweets average of the hourly measure. We explore the robustness of our results to the daily measure in the Appendix.

Felbo et al. (2017) deep learning algorithm that Janys Analytics uses to construct the Wall Street Journal - IHS U.S. Sentiment Index (USSI) introduced in Zumbrun (2017). In brief, every tweet from a 10% random sample of all Twitter messages within the preceding hour is scored and then these scores are averaged together. These are very large samples to undertake sentiment analysis since in 2005, there was an average of 350,000 tweets sent per minute globally. The number of tweets per hour generally varies between 120,000 to 200,000 tweets per hour in our 10% random sample.

Social media users are not demographically representative of the population and prior research has found they are more likely to reside in urban areas (Mislove et al., 2011) that are wealthier with younger populations (Malik et al., 2015). The Twitter users themselves tend to be younger and more educated than the general population (Greenwood, Perrin, and Duggan, 2016). Yet, for consumer confidence, a predominately younger population may be quite relevant for forecasts, given the standard hump shaped curve of how consumer expenditures vary over the life-cycle.

Measuring sentiment in social media is a challenge in the field of natural language processing. The algorithm we selected to analyze sentiment was trained on 124.6 million tweets containing emojis. The algorithm does not score individual emotion words in a Twitter message, but rather calculates a score based on the probability of each of 64 different emojis capturing the sentiment in the full Twitter message taking the structure of the sentence into consideration. Thus, each emoji has a fixed score and the sentiment of a message is a weighted average of the type of mood being conveyed. Tests of the validity of the Felbo et al. (2017) algorithm with samples drawn from Amazon mechanical turk, have found it to be more accurate than competing sentiment algorithms.[10] The USSI is a national measure and includes both investors and non-investors that has recently used to forecast volatility (Lehrer, Xie, and Zhang, 2019).[11] In total, we have 37,008 observations for the USSI variable at the hourly level as well as 1,543 observations for the USSI

[10] This likely arises since it considers the ordering of all the words in a Twitter message rather than using a binary indicator such as positive or not, to those based on scoring words via emotional valence.

[11] The prior algorithm used by Janys Analytics to measure social media sentiment was used in applications to forecast revenue for the film industry (Lehrer and Xie, 2017, 2018).

at the daily level. Last, we created a monthly USSI variable, denoted as USSI$_a$, by simple averaging of the hourly measures.

Beyond social media data, we also account for macroeconomic and financial variables in our forecast model. These explanatory variables are also collected at different frequencies. Thus, for ease of exposition we use (M), (D), and (H) to indicate whether a specific data series is reported on a monthly basis, daily basis, or hourly basis. The explanatory variables that we control for in our forecasting models are listed and described in Table 1.

While the macroeconomic variables are measured at the same frequency (monthly) as the CCI, both the financial and big data variables are measured at a higher frequency (daily and hourly). In this paper, we focus on alternative conversions for the big data variables and also convert all financial variables from daily to monthly using the conventional MIDAS method that is described in the next section. We consider three alternative measures of the USSI: (i) USSI$_a$ is the monthly basis USSI converted from hourly basis USSI using simple weighted averaging;[12] (ii) we denote the monthly USSI converted from hourly basis using conventional MIDAS as USSI$_h$; and (iii) USSI$_{new}$ is the USSI converted from hourly basis using the newly proposed H-MIDAS method that we introduce in Section 3.1.

Summary statistics for each data series included in the forecasting exercises are presented in Table 2. Note, that prior to including each series in this exercise, we perform the augmented Dickey-Fuller test (ADF) test of the null hypothesis that a unit root is present in each respective time series. The results suggest that the original series of nearly every macroeconomic and financial variable is non-stationary; with the exception of the unemployment rate. To construct a stationary data series for variable $z_t$, we transform the data by calculating the first difference $\Delta z_t \equiv z_t - z_{t-1}$. Applying the ADF-test to $\Delta z_t$ we next confirmed that each transformed data series is stationary. Notice in Table 2, that there is a significant heterogeneity in both the CCI, MCSI and USSI measures. Among the alternative USSI measures we consider in the forecasting exercises, the USSI converted from hourly basis using the newly proposed H-MIDAS method exhibits the lowest variability.

[12] The hourly USSI is accompanied with a hourly volume that measures the total number of tweets involved in estimating the sentiment. The monthly basis USSI$_a$ is a simple weighted average of the hourly USSI using volume as weights.

Table 1: List of Explanatory Variables to Forecast the CCI

| | | |
|---|---|---|
| *Panel A: Macroeconomic Variables at the Monthly Level (M)* | | |
| (1) | MCSI | The University of Michigan Consumer Sentiment Index (MCSI) is a well-established index prepared by University of Michigan's Institute for Social Research to present an alternative measure of consumer confidence. |
| (2) | LEI | The Conference Board Leading Economic Index (LEI) is intended to forecast future economic activity using values of ten key variables that measure the overall performance of the U.S. economy. |
| (3) | UR | The seasonally adjusted U.S. unemployment rate (UR) released by the U.S. Bureau of Labor Statistics. |
| (4) | SR | The seasonally adjusted U.S. personal saving rate (SR) as % of disposable personal income released by the U.S. Bureau of Labor Statistics. |
| (5) | CPI | The seasonally adjusted consumer price index (CPI) for all urban consumers that is released by the U.S. Bureau of Labor Statistics. |
| *Panel B: Daily Financial Variables (D)* | | |
| (6) | SP500 | The Standard & Poor's 500 (SP500) is a stock market index based on the market capitalization of 500 large companies having common stock listed on the NYSE or NASDAQ. |
| (7) | VIX | The monthly basis Chicago board of exchange (CBOE) volatility index (VIX) is a popular measure of the implied volatility of S&P 500 index options. It is colloquially referred to as the fear index or the fear gauge. |
| (8) | USD | The US Dollar Index (USD) is an index (or measure) of the value of the United States dollar relative to a basket of foreign currencies. It is a weighted geometric mean of the dollar's value relative to six other select currencies. |
| (9) | TS | The term spread (TS) is calculated as the difference between the 10-year and 3-year treasury constant maturity rates. |
| *Panel C: Big Data Variable (H)* | | |
| (10) | USSI | The hourly Wall Street Journal - IHS U.S. Sentiment Index (USSI). |

Last, the variability in each of the financial variables appears small, but the range in the data appears quite large relative to the other predictor variables.

Table 2: Summary Statistics

| Variable | Mean | Median | Minimum | Maximum | Std.Dev. |
|---|---|---|---|---|---|
| *Panel A: Dependent Variable* | | | | | |
| $\Delta$CCI[*] | 1.1549 | 1.5000 | -8.8000 | 10.8000 | 5.0883 |
| | | | | | |
| *Panel B: Macroeconomic Variable* | | | | | |
| $\Delta$MCSI | 0.4680 | 0.3500 | -5.2000 | 8.1000 | 3.2773 |
| $\Delta$LEI | 0.3640 | 0.4000 | -0.4000 | 1.3000 | 0.3973 |
| UR[†] | 5.9039 | 5.6000 | 4.6000 | 8.0000 | 1.0505 |
| $\Delta$SR | -0.1040 | 0.1000 | -6.1000 | 0.6000 | 0.8905 |
| $\Delta$CPI | 0.2647 | 0.3610 | -1.3770 | 1.3730 | 0.4860 |
| | | | | | |
| *Panel C: Financial Variable* | | | | | |
| $\Delta$SP500 | 2.1888 | 0.0200 | -27.3900 | 72.9000 | 17.6210 |
| $\Delta$VIX | -0.0976 | 0.0616 | -4.9285 | 2.8389 | 1.1653 |
| $\Delta$USD | 0.0475 | 0.0386 | -0.8440 | 1.0179 | 0.3003 |
| $\Delta$TS | -0.0029 | -0.0040 | -0.0204 | 0.0191 | 0.0089 |
| | | | | | |
| *Panel D: Big Data Variable* | | | | | |
| $\text{USSI}_a$ | 0.1067 | 0.3835 | -6.5576 | 7.5495 | 2.3374 |
| $\text{USSI}_h$ | 0.1745 | 0.7438 | -18.1427 | 12.9880 | 7.0689 |
| $\text{USSI}_{new}$ | 0.2560 | 0.3077 | -10.8029 | 10.2103 | 5.3858 |

[*] The $\Delta$ sign indicates the first-difference of the associated variable.

[†] Parameter UR is stationary, and hence does not require first-difference.

## 3 Data Sampling Techniques

Mixed-frequency problems are ubiquitous in many forecasting exercises for the banking and finance industry. The CCI is not sampled at the same frequency as its potential predictors listed in Table 2. Numerous solutions to this challenge have been proposed beginning with simply averaging the high-frequency data ($\text{USSI}_a$) as in Section 2 to MIDAS techniques initially proposed in Ghysels et al. (2004) and subsequently in Ghysels et al. (2005, 2006 and 2007). Unlike simple averaging which equally weights all the data in the high frequency series, MIDAS uses a pre-determined weighting function with a small number of hyperparameters relative to the sampling rate of the higher-frequency variable. The hyperparameters are estimated (usually as the unique solution with a specific optimization algorithm) and the estimates are then used to compute the MIDAS weighted averaged predictors in the same frequency as the dependent variable.

Formally, if $Y_t$ is a low frequency variable that is sampled at periods denoted by a time index $t$ for $t = 1, ..., n$. Consider a higher frequency (indicated by a superscript $h$ throughout the paper) predictor $\boldsymbol{X}_t^h$ that are sampled $m$ times within the period of $t$:

$$\boldsymbol{X}_t^h \equiv \left[ X_t^h, X_{t-\frac{1}{m}}^h, ..., X_{t-\frac{m-1}{m}}^h \right]^\top . \tag{1}$$

A specific element among the high frequency observations in $\boldsymbol{X}_t^h$ is denoted by $X_{t-\frac{i}{m}}^h$ for $i = 0, ..., m-1$.[13] Denoting $L^{i/m}$ as the lag operator, then $X_{t-\frac{i}{m}}^h$ can be reexpressed as $X_{t-\frac{i}{m}}^h = L^{i/m} X_t^h$ for $i = 0, ..., m-1$.

Since $\boldsymbol{X}_t^h$ on $Y_t$ are measured at different frequencies, data snooping may arise if researchers choose which $X_{t-\frac{i}{m}}^h$ to include as an explanatory variable. Converting the higher-frequency data to match the sampling rate of the lower-frequency data solves the problem of mixed sampling frequencies. The simplest way to to estimate a low frequency $X_t$ that matches the frequency of $Y_t$ is a simple average of the high frequency observations $\boldsymbol{X}_t^h$:

$$\bar{X}_t = \frac{1}{m} \sum_{i=0}^{m-1} L^{i/m} X_t^h.$$

When $Y_t$ and $\bar{X}_t$ are measured in the same time domain, a regression approach is simply

$$Y_t = \alpha + \gamma \bar{X}_t + \epsilon_t = \alpha + \frac{\gamma}{m} \sum_{i=0}^{m-1} L^{i/m} X_t^h + \epsilon_t, \tag{2}$$

where $\alpha$ is the intercept, $\gamma$ is the slope coefficient on the time-averaged $\bar{X}_t$. This approach assumes that each element in $\boldsymbol{X}_t^h$ has an identical effect on explaining $Y_t$, since they share the same coefficient $\gamma$.

These homogeneity assumptions may be quite strong in practice. For example, elements of the high frequency variable may have a heterogeneous effect. One could assume that each of the slope coefficients for each element in $\boldsymbol{X}_t^h$ is unique. Extending Model (2) to

[13] In this case, the high frequency observation $X_t^h$ at exact time period of $t$ is included in estimating $Y_t$. In practice, this is possible when the low frequency $Y_t$ is observed after the period $t$, for example, GDP, GNP, etc. For simplicity, we adopt this framework in the remainder of this paper.

allow for heterogeneous effects of the high frequency observations generates

$$Y_t = \alpha + \sum_{i=0}^{m-1} \gamma_i L^{i/m} X_t^h + \epsilon_t, \tag{3}$$

where $\gamma_i$ represents a set of slope coefficients for all high frequency observations $X^h_{t-\frac{i}{m}}$. Estimating $\gamma_i$ can be problematic when $m$ is a relatively large number.[14]

Thus, while the simple averaging model (2) is parsimonious, it discards information related to the timing of innovations to higher-frequency data. In contrast, the heterogeneous weighting model (3) preserves the timing information, although it may require the analyst to estimate a potentially large number of parameters. To reduce the dimensionality of the number of parameters while preserving some timing information, Ghysels et al. (2004) proposed the following MIDAS model:

$$Y_t = \alpha + \gamma \sum_{i=0}^{m-1} \Phi(i;\boldsymbol{\theta}) L^{i/m} X_t^h + \epsilon_t, \tag{4}$$

where the function $\Phi(i;\boldsymbol{\theta})$ is a polynomial that determines the weights for temporal aggregation based on the hyperparameter $\boldsymbol{\theta}$. The weighting function, $\Phi(i;\boldsymbol{\theta})$, is not restricted and can take a variety of functional forms. Researchers should select a $\Phi(i;\boldsymbol{\theta})$ that is both flexible and parsimonious. For example, Ghysels, Santa-Clara, and Valkanov (2005) suggest using an exponential Almon specification:

$$\Phi(i;\theta_1,\theta_2) = \frac{\exp(\theta_1 i + \theta_2 i^2)}{\sum_{j=0}^{m-1} \exp(\theta_1 j + \theta_2 j^2)}.$$

With this weighting function, simple time averaging is obtained when $\theta_1 = \theta_2 = 0$.[15]

A nonlinear least squares (NLS) estimator is used to obtain the unknown coefficients $\boldsymbol{\theta}$

[14] Problems with high-dimensional explanatory variables are a major feature of research involving big data. Estimators such as the LASSO zero out many of the $\gamma_i$ to satisfy a strong sparsity condition. We follow an approach developed in the econometrics literature to develop a parsimonious specification.

[15] Another popular choice among forecasters for the weighting function is the beta formulation:

$$\Phi(i;\theta_1,\theta_2) = \frac{f(\frac{i+1}{m},\theta_1,\theta_2)}{\sum_{j=0}^{m-1} f(\frac{j+1}{m},,\theta_1,\theta_2)}$$

from MIDAS regression. We can reexpress the right-hand-side of equation (4) and define

$$\hat{X}_t \equiv \sum_{i=0}^{m-1} \Phi(i;\hat{\boldsymbol{\theta}}) L^{i/m} X_t^h. \tag{5}$$

Intuitively, this converts the higher frequency variable $X^h_{t-\frac{i}{m}}$ to the same frequency as $Y_t$ with dynamic weights $\Phi(i;\hat{\boldsymbol{\theta}})$; such that $\hat{X}_t$ has better explanatory power on $Y_t$.

Using the conventional MIDAS method presented in equation (4), the hourly USSI is aggregated to a monthly measure using the exponential Almon polynomial as the weight function. Figure 1(a) illustrates the estimated weights for each high frequency group with $m = 650$ observations. For brevity, we only present the first 100 Almon polynomial lags, since weights after the first 20 periods are very close to 0.[16] In the context of this study, the evidence in the graph implies that only hourly measures of the USSI collected on the last day of each month are used to construct the monthly USSI.

The extreme weights in Figure 1(a) arise from the choice of an exponential Almon polynomial as the weight function.[17] The exponential Almon polynomial gives near zero weight to observations collected earlier in the data series based on the belief that more recent observations should have larger impacts on the dependent variable. The exponential Almon performs well in settings where analysts convert monthly data to quarterly, or annual data, which involves either 3 observations averaged to 1 or 12 observations averaged to 1.[18] In our application, however, we need to convert both daily data to monthly data as well as hourly data to monthly data, that is approximately 650 observations averaged to 1 in the latter example.

where $f(x,\theta_1,\theta_2) = \frac{x^{\theta_1-1}(1-x)^{\theta_2-1}\Gamma(\theta_1+\theta_2)}{\Gamma(\theta_1)\Gamma(\theta_2)}$ with $\theta_1$ and $\theta_2$ being hyperparameters governing the shape of the weighting function, and $\Gamma(\theta) = \int_0^\infty e^{-x}x^{\theta-1}dx$ is the standard gamma function. Simple time averaging is nested within and obtained when $\theta_1 = \theta_2 = 1$. In our forecasting exercise, we considered both specifications of the weighting function and the results using the Almon specification strictly dominate the beta specification. As such, we present results using the Almon specification in the main text. The full set of results that utilized the beta specification are available from the corresponding author upon request.

[16] The exponential Almon polynomial only considers approximately 20 most recent observations in both our application and earlier work including Ghysels, Santa-Clara, and Valkanov (2006). The 20 most recent observations roughly corresponds to using data from a single day in a month.

[17] Note, the beta formulation leads to even more extreme weights.

[18] See Ghysels et al. (2004,2005) and Ghysels, Sinko, and Valkanov (2007) for more examples.

Recall, the CCI is constructed from responses to survey questions that are received throughout the month. This transformation places greater weight on most recent events but if survey responses vary across the month and the completions are related to economic conditions, this strategy ignores the potential timing. As such, we next consider a simple modification to the conventional MIDAS procedure to allow for greater heterogeneity.

## 3.1 Heterogenous Mixed Data Sampling (H-MIDAS)

We modify the (conventional) MIDAS method described in Section 3 to a method that uses a step function to allow for heterogeneous effects of different high frequency observations on the low frequency dependent variable. We coin this new method as heterogeneous MIDAS, or H-MIDAS for short.[19]

To demonstrate this H-MIDAS procedure, recall that $\boldsymbol{X}_t^h$ is defined as

$$\boldsymbol{X}_t^h = \left[X_t^h, X_{t-\frac{1}{m}}^h, ..., X_{t-\frac{m-1}{m}}^h\right]^\top.$$

A low frequency $\bar{X}_t^{(l)}$ can be constructed following

$$\bar{X}_t^{(l)} \equiv \frac{1}{l}\sum_{i=0}^{l-1} L^{i/m} X_t^h = \frac{1}{l}\sum_{i=0}^{l-1} X_{t-\frac{i}{m}}^h, \tag{6}$$

where $l$ is a pre-determined number and $l \leq m$. Equation (6) implies that $\bar{X}_t^{(l)}$ is computed by a simple average of the first $l$ observations in $\boldsymbol{X}_t^h$ and ignore the remaining observations.

We consider different values of $l$ and group all $\bar{X}_t^{(l)}$ into $\tilde{\boldsymbol{X}}_t$ such that

$$\tilde{\boldsymbol{X}}_t = \left[\bar{X}_t^{(l_1)}, \bar{X}_t^{(l_2)}, \ldots, \bar{X}_t^{\left(l_p\right)}\right],$$

where we set $l_1 < l_2 < \cdots < l_p$. Consider a weight vector $\boldsymbol{w} = \left[w_1, w_2, \ldots, w_p\right]^\top$ with

[19] Our method is inspired by the heterogeneous autoregression (HAR) of Corsi (2009), who proposed an additive cascade model of volatility components defined over different time periods that leads to a simple AR-type model in the realized volatility with the feature of considering different volatility components realized over different time horizons.

$\sum_{j=1}^{p} w_j = 1$, we can construct regressor $X_t^{new}$ as $X_t^{new} = \tilde{\boldsymbol{X}}_t \boldsymbol{w}$. The regression based on our H-MIDAS estimator can be expressed in the same fashion as the conventional MIDAS estimator of Ghysels et al. (2004) such that

$$Y_t = \beta X_t^{new} + \epsilon_t = \beta \sum_{s=1}^{p} \sum_{j=s}^{p} \frac{w_j}{l_j} \sum_{i=l_{s-1}}^{l_s-1} L^{i/m} X_t^h + \epsilon_t = \beta \sum_{s=1}^{p} \sum_{i=l_{s-1}}^{l_s-1} w_s^* L^{i/m} X_t^h + \epsilon_t. \quad (7)$$

This specification nests the weights considered in conventional MIDAS when $l_0 = 0$ and $w_s^* = \sum_{j=s}^{p} \frac{w_j}{l_j}$. For ease of exposition, we ignore the intercept $\alpha$ in the H-MIDAS regression (7). In empirical practice, one can demean $Y_t$ and $\tilde{\boldsymbol{X}}_t$ when estimating (7).

The weights $\boldsymbol{w}$ play a crucial role in this procedure. We first estimate $\widehat{\beta \boldsymbol{w}}$ following

$$\widehat{\beta \boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathcal{W}}{\arg\min} \left\| Y_t - \ \tilde{\boldsymbol{X}}_t \cdot \beta \boldsymbol{w} \right\|^2$$

by any appropriate econometric method necessary, where $\mathcal{W}$ is some predetermined weights set. Once $\widehat{\beta \boldsymbol{w}}$ is obtained, we estimate the weight vector $\hat{\boldsymbol{w}}$ by rescaling

$$\hat{\boldsymbol{w}} = \frac{\widehat{\beta \boldsymbol{w}}}{\text{Mean}(\widehat{\beta \boldsymbol{w}})},$$

since the coefficient $\beta$ is a scalar. In this paper, we use OLS to estimate $\widehat{\beta \boldsymbol{w}}$ and then calculate the converted $\hat{X}_t^{new} = \tilde{\boldsymbol{X}}_t \cdot \hat{\boldsymbol{w}}$.

Figure 1: Estimated Weights for USSI Using Various MIDAS Methods

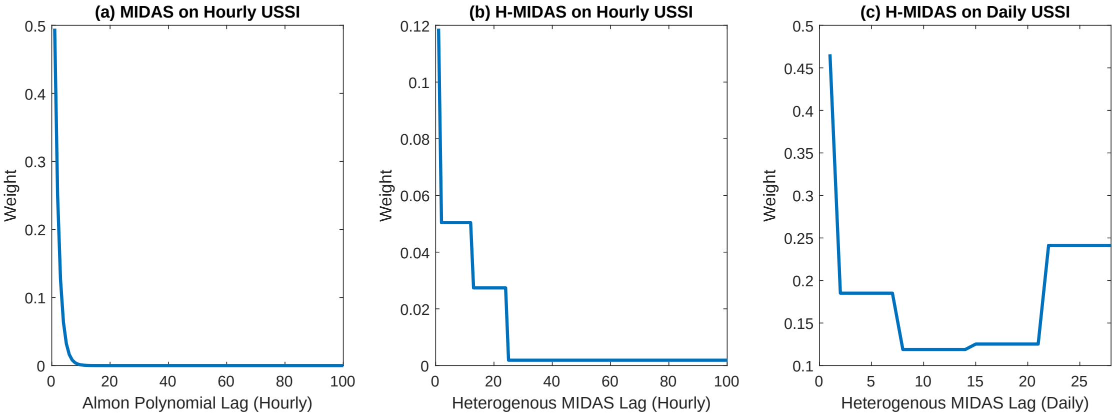


Figure 1(b) illustrates the estimated weights for H-MIDAS when we convert the hourly USSI to monthly using [1, 12, 24, 120, 240, 650] as the lag index to mimic the 1-hour, 1/2-day, 1-day, 1-week, 2-week, and 1-month effects. The estimated weights for H-MIDAS are not as smooth as conventional MIDAS demonstrated in Figure 1(a) and place significantly less weight on the USSI measured in the last few hours. We denote the USSI converted by H-MIDAS as $\text{USSI}_{new}$. Overall, contrasting the first two panels of figure 1 illustrates the benefits that may accrue from relaxing the functional form assumptions embedded in the choice of weighting functions using the conventional MIDAS.

Figure 1(c) displays the estimated weights for H-MIDAS that convert the daily USSI to monthly using $[1, 7, 14, 30]$ as the lag index to mimic the 1-day, 1-week, 2-week, and 1-month effects. Notice that the step function has a very heterogeneous pattern placing larger weight on the most recent and least recent days in the month. Thus, the last panels of Figure 1 illustrate that the H-MIDAS procedure does not restrict the pattern across dates to take a specific shape. The time-varying pattern that is observed in this panel, may arise since we control for the MCSI that may do a tremendous job of capturing similar information as measures of the USSI collected between 8-22 days earlier.

To further understand the properties of the H-MIDAS estimator, we derive the asymptotic properties of the H-MIDAS estimator in the Appendix A. These properties permit us to state the following (dial-down version) lemma:

**Lemma 1** *Let the variable $X^h_{t-\frac{i}{m}}$ follow an AR(1) process. Then, compared to the H-MIDAS method, the simple averaging estimator introduces asymptotic bias to the coefficient $\beta$.*

See Appendix A for an extended statement and a detailed proof. This lemma extends Proposition 4.3 of Andreou, Ghysels, and Kourtellos (2010) that derived conditions under which the simple averaging estimator can introduce asymptotic bias to the coefficient relative to the conventional MIDAS techniques. The above finding can now be applied to a broader set of MIDAS techniques including the H-MIDAS method.

# 4 Forecasting Techniques

Researchers interested in forecasting with social media data are faced with a decision regarding on how to construct aggregate measures from high frequency social media data and also which estimator to apply to the forecasting model. Since time series forecasting can be framed as a supervised learning problem, there is growing evidence (see e.g. Lehrer and Xie, 2018) that standard linear and nonlinear machine learning algorithms display improved performance.[20] To help provide an evidence base to assist future researchers and finance practitioners, we examine the relative prediction efficiency of different estimators with different ways of accounting for social media data using the following experiment.

We contrast a suite of popular approaches from the econometrics literature with those from machine learning. Specifically, the econometric approaches include

(i) OLS using all of the available regressors in a general unrestricted model (GUM);

(ii) Model selection using the Akaike information criterion (AIC) of Akaike (1973);

(iii) Model averaging allowing for model uncertainty where the weights are chosen using the prediction model averaging of Xie (2015) (PMA).

Among machine learning algorithms, we first consider four methods that use algorithms that partition the characteristic space into a series of hyper-cubes. A local constant model is estimated in each partition to approximate the underlying data generation process. The methods considered include

(iv) Regression trees proposed by Breiman, Friedman, and Stone (1984) (RT);

(v) Bootstrap aggregation (BAG) tree technique developed in Breiman (1996);

(vi) Random forest (RF) of Breiman (2001);

[20] For example, Bajari, Nekipelov, Ryan, and Yang (2015) analyzed the advantages of using machine learning methods for demand estimation, Mullainathan and Spiess (2017) provided a up-to-date overview on machine learning methods in economics, while Athey and Imbens (2017) demonstrated how machine learning methods can improve the performance of the standard econometric methods.

(vii) A simple least squares boosting (LSB) tree of RT ensembles (BOOST).

We also consider penalized regression methods from the machine learning literature

(viii) Support vector regression (SVR) machines proposed in by Drucker et al. (1996) using linear and nonlinear kernels

With both the bootstrap aggregation tree and random forest algorithms, we estimate 100 trees in the ensemble and additionally account for an important feature of our data consisting of dependent observations, We use two specific bootstrap methods for time series data in our implementation. Specifically, we consider Kulperger and Prakasa Rao (1989) Markov bootstrap method as well as Künsch (1989) moving block bootstrap (MBB) method. These methods respectively rely on either assuming a specific structural form for a stationary and weakly dependent time series or a weaker restriction that only preserves the dependence structure of the random variable at short lag distances.[21] We consider SVRs with different penalty functions to control which observations are given weight in the objective function of the estimator. We consider both linear (denoted as $\text{SVR}_1$) and two different nonlinear kernels (denoted as $\text{SVR}_2$ for a Gaussian kernel and $\text{SVR}_3$ for a local polynomial kernel). Further details on the implementation and theory underlying each of these estimators is provided in Appendix B.

## 5 Empirical Results

A rolling window exercise that fixes the window length at 36 (3 years) is conducted. For each forecasting strategy, the mean squared forecast error (MSFE) and mean absolute forecast error (MAFE) from a one-step-ahead forecast is computed. To assess how to extract the most value from social media content in forecasting economic outcomes, we consider five alternative methods of including the USSI as a predictor variable:

[21] See Kreiss and Lahiri (2012) for a detailed literature review as well as Appendix B for more details on our implementation.

(i) $\mathcal{M}_0$: data without any USSI variables;

(ii) $\mathcal{M}_a$: data with USSI$_a$ (simple average);

(iii) $\mathcal{M}_m$: data with USSI$_m$ (conventional MIDAS, hourly);

(iv) $\mathcal{M}_{new}$: data with USSI$_{new}$ (H-MIDAS, hourly).

(v) $\mathcal{M}_{all}$: data with all three versions of USSI variables.

Table 3 reports the median MSFE and MAFE from the relative one month ahead prediction efficiency experiment for each of the 10 forecasting methods (columns) described in the preceding section with the above alternative methods of including the USSI across the rows of Table 3. To ease interpretation, we place the lowest MSFE and MAFE in bold, for each row of Table 3. The linear support vector machines for regression demonstrates improved performance relative to the other estimators considered, unless we include three versions of the USSI in the model.

Table 3: One-step-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)* | | | | | | | | | | |
| $\mathcal{M}_0$ | 18.8061 | 17.7036 | 17.9470 | 27.7313 | 18.6351 | 18.8577 | 67.6181 | **13.0906** | 19.5763 | 31.0816 |
| $\mathcal{M}_a$ | 19.4375 | 26.4169 | 20.8762 | 27.7313 | 17.5063 | 17.5503 | 66.9525 | **16.2758** | 19.6363 | 33.2165 |
| $\mathcal{M}_m$ | 17.0214 | 19.6263 | 16.9666 | 12.4811 | 13.2995 | 14.3669 | 35.9071 | **12.1732** | 19.5361 | 28.7186 |
| $\mathcal{M}_{new}$ | 14.1906 | 15.1128 | 13.2115 | 18.7820 | 12.9198 | 13.8692 | 18.5529 | **10.3271**$^{\diamond}$ | 19.6891 | 24.8384 |
| $\mathcal{M}_{all}$ | 17.6537 | 14.8215 | 13.1496 | 13.2403 | **11.3241** | 11.8710 | 18.5307 | 13.1679 | 19.7090 | 32.5857 |
| *Panel B: Mean Absolute Forecast Error (MAFE)* | | | | | | | | | | |
| $\mathcal{M}_0$ | 3.5614 | 3.4078 | 3.3684 | 3.8111 | 3.3645 | 3.3721 | 6.4182 | **3.0174** | 3.7057 | 4.3819 |
| $\mathcal{M}_a$ | 3.5083 | 4.0982 | 3.7362 | 3.8111 | 3.3317 | 3.2610 | 6.6775 | **3.2589** | 3.7403 | 4.7417 |
| $\mathcal{M}_m$ | 3.1777 | 3.4367 | 3.0940 | 2.9261 | 2.7873 | 2.9146 | 4.4979 | **2.6881** | 3.7023 | 4.5727 |
| $\mathcal{M}_{new}$ | 2.7981 | 2.7415 | 2.6811 | 3.0674 | 2.6245 | 2.8316 | 3.4075 | **2.5035**$^{\diamond}$ | 3.7456 | 3.7854 |
| $\mathcal{M}_{all}$ | 3.0771 | 2.6835 | 2.6163 | 3.0374 | **2.5985** | 2.6084 | 3.2112 | 2.7050 | 3.7323 | 4.4751 |

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

There are several findings in Table 3 worth stressing. First, when comparing the results across rows of the Table, irrespective of the estimator, we see that the prediction efficiency increases by more than 25% using MSFE as criterion when we include social media data measured by USSI$_{new}$. This result provides the first piece of evidence demonstrating the importance of using social media data in this forecasting exercise.

Second, the results in Table 3 demonstrate the general improvements in forecasting from a machine learning algorithm relative to an econometric approach presented in any of the first three columns of the Table. Gains from machine learning algorithms arise since variables are added to the forecasting model in a more flexible manner than econometric strategies, since in a tree structure every cut-point in each independent variable is considered allowing for highly nonlinear models with potentially complex interactions. In our application, support vector machines for regressions demonstrate the strongest performance in terms of either MSFE or MAFE for most of the cases, but bagging and random forests also have lower MSFE and MAFE so long as a measure of the USSI is included. Regression trees and boosting do not perform as well (nor the non-linear SVR2 and SVR3) as the SVR1 estimator, which may reflect the small sample size in this forecasting exercise.

Third, the results also suggest the importance of considering model uncertainty when comparing GUM (no model uncertainty) to PMA. The prediction efficiency is improved by 34% when using $\mathcal{M}_{all}$. Interestingly, the model selection (AIC) method in our exercise do not yield better forecasts than GUM, with $\mathcal{M}_a$, $\mathcal{M}_m$, and $\mathcal{M}_{new}$.

To examine if there are more general benefits from using H-MIDAS in place of conventional MIDAS, we next convert the daily financial variables in $\mathcal{M}_s^h$ using H-MIDAS. We replicate the analysis presented in Table 3 where now each of the four financial variables used as predictors is transformed via H-MIDAS. To undertake this transformation, we set the lag index in H-MIDAS as 1 to 22 in a bid to mimic the 1-day to 1-month averages.[22] In other words, the results presented in Table 4 repeats the same forecasting experiment where now every high frequency data is converted by the H-MIDAS procedure.

The rows of in Table 4 continue to explore alternative methods to include the USSI in the forecasting exercise, with input groups, denoted by $\mathcal{M}_s^h$, which is identical to $\mathcal{M}_s$ for $s = 0, a, m, new, all$ with the exception that we in contrast to the conventional MIDAS method used in $\mathcal{M}_{new}$. Exploring each cell of the forecasting results presented in Table 4, we observe improved results indicating improved forecasting performance than those in Table 3. This implies the superiority of our H-MIDAS method over the conventional MIDAS

[22] The choice of optimal lag index is beyond the scope of this paper and leave for future research.

Table 4: One-step-ahead Forecasting Results where All Financial and Macroeconomic Variables Are Transformed by H-MIDAS

| | GUM | AIC | PMA | RT | BAG | RF | BOOST | SVR$_1$ | SVR$_2$ | SVR$_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| *Panel A: Mean Squared Forecast Error (MSFE)* | | | | | | | | | | |
| $\mathcal{M}_0^h$ | 11.1230 | 10.2568 | 11.6820 | 14.6906 | 13.1275 | 12.5885 | 14.4684 | **9.9089** | 19.4999 | 15.6654 |
| $\mathcal{M}_a^h$ | 10.2857 | 10.0643 | 11.1280 | 14.6906 | 12.6923 | 12.4064 | 13.5092 | **9.8430** | 19.8121 | 15.3893 |
| $\mathcal{M}_m^h$ | 11.6393 | 10.2568 | 11.6820 | 16.0860 | 12.7677 | 12.3043 | 16.7293 | **10.7093** | 19.4000 | 23.2848 |
| $\mathcal{M}_{new}^h$ | 8.7328 | 9.3175 | 8.7556 | 13.3832 | 11.8608 | 11.5446 | 12.5212 | **8.2013**$^\diamond$ | 19.6947 | 9.0957 |
| $\mathcal{M}_{all}^h$ | 10.6760 | 9.3175 | 8.7556 | 13.3832 | 11.6948 | 11.4150 | 8.3833 | **8.2310** | 19.7648 | 17.6720 |
| *Panel B: Mean Absolute Forecast Error (MAFE)* | | | | | | | | | | |
| $\mathcal{M}_0^h$ | 2.7011 | 2.6499 | 2.8140 | 3.3228 | 2.9433 | 2.8488 | 3.1374 | **2.5852** | 3.6667 | 3.4733 |
| $\mathcal{M}_a^h$ | 2.5891 | 2.5993 | 2.6907 | 3.3228 | 2.8446 | 2.8743 | 3.2356 | **2.4999** | 3.7491 | 3.1214 |
| $\mathcal{M}_m^h$ | 2.7688 | 2.6499 | 2.8140 | 3.4833 | 2.9018 | 2.8205 | 3.5549 | **2.6738** | 3.6548 | 4.0463 |
| $\mathcal{M}_{new}^h$ | 2.4821 | 2.6057 | 2.4140 | 3.0051 | 2.7590 | 2.7249 | 2.7532 | **2.3717** | 3.7330 | 2.5448 |
| $\mathcal{M}_{all}^h$ | 2.6533 | 2.6057 | 2.4140 | 3.0051 | 2.7336 | 2.6774 | 2.4616 | **2.3535**$^\diamond$ | 3.7337 | 3.4592 |

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

even in the case of converting daily frequency to monthly frequency.

The main result from Table 4 is a clear demonstration of the potentially large benefits from adopting MIDAS methods that allow for more flexible weights and not restrict them to be constant across predictors or to follow a specific functional form. While the H-MIDAS approach was initially developed for social media data, in part since online opinion can shift rapidly in unpredictable directions,[23] our empirical investigation finds that it be beneficial to use with other high frequency variables whose measurements vary significantly within the low frequency period. Moreover, the results in Table 4 reinforce our earlier finding of the importance of using social media data in this forecasting exercise, since the prediction efficiency increases by more than 20% judged by the MSFE. Further, the

[23] As an extreme example of the challenge in incorporating social media data, tweets from U.S. President Donald Trump on economic policies often lead to both large swings in aggregate Twitter sentiment measures and can have large impacts upon intraday volatilities facing futures, equities, and FOREX markets. A related but more concrete and specific example of how aggregate Twitter sentiment moves with financial indicators such as equity prices consider that following the removal of Ivanka Trump's fashion line from their stores, President Trump issued a statement via Twitter:

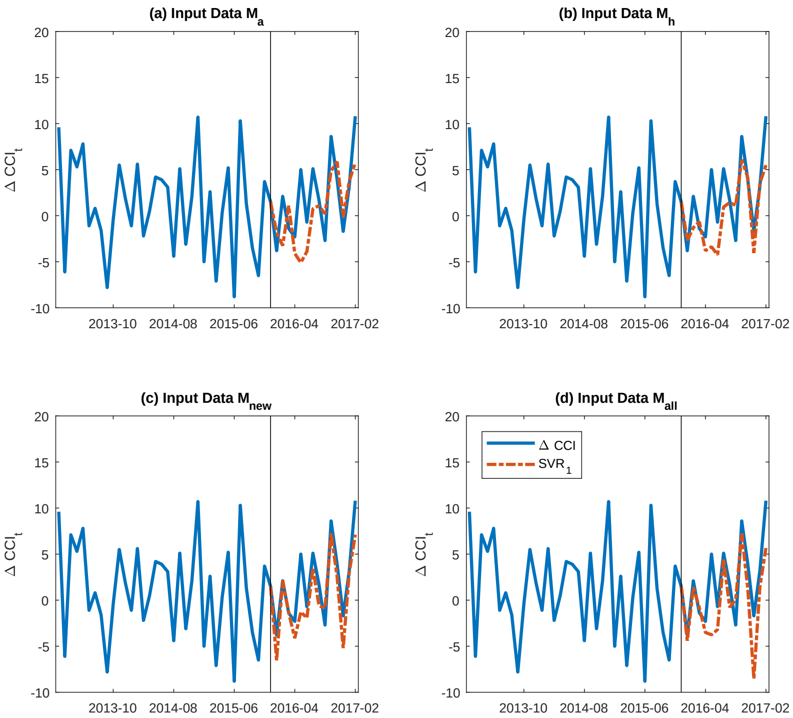
> *My daughter Ivanka has been treated so unfairly by @Nordstom. She is a great person – always pushing me to do the right thing! Terrible!*

The general public response to this Tweet was to disagree with President Trump's stance on Nordstrom so aggregate Twitter sentiment measures rose and the immediate negative effects from the Tweet on Nordstrom stock of a decline of 1% in the minute following the tweet were fleeting since the stock closed the session posting a gain of 4.1%. See http://www.marketwatch.com/story/nordstrom-recovers-from-trumps-terrible-tweet-in-just-4-minutes-2017-02-08 for more details on this episode.

performance of SVR$_1$ continues to dominate other estimators in Table 4.

To provide a visual understanding of why $\mathcal{M}_{new}$ yields the lowest MSFE with SVR$_1$, we present the forecasting results for $\mathcal{M}_a$, $\mathcal{M}_m$, $\mathcal{M}_{new}$, and $\mathcal{M}_{all}$ we account for the USSI in the panels of Figure 2. The solid line represents the actual data and the dashed line represents forecasting results from the SVR$_1$ method. Monthly date ticks are labeled in the horizontal axis. Notice that both $\mathcal{M}_a$ and $\mathcal{M}_m$ struggle with forecasts in August and September 2016. Both $\mathcal{M}_{new}$ and $\mathcal{M}_{all}$ generally tracks the temporal pattern and $\mathcal{M}_{new}$ does experience smaller deviations from the actual line in most months. The results with $\mathcal{M}_{all}$ perform quite well until the US election when they overshoot the negative sentiment associated with Donald Trump's victory relative to sentiment associated with consumer confidence. This result does stress that understanding what twitter sentiment is capturing is important to using it as an explanatory factor in forecasting models.

Figure 2: Forecasting Performance of SVR$_1$ Using Various Input Data



The panels in Figure 3 conduct the same graphical evidence of the forecasting performance of the 10 different estimators considered for the $\mathcal{M}_{new}$. The three econometric approaches (GUM, AIC and PMA) as well as boosting tend to forecast too low values for the

CCI in most periods. Both random forests and $\text{SVR}_2$ appear to do a poor job at capturing the monthly fluctuations in the CCI. The similar performance of random forests relative to regression tree is striking since the latter should capture more heterogeneity by averaging across trees. Among potential empirical strategies, support vector machines for regression with linearity appears to perform well overall, as well as exhibit the closest forecasts in most every month. $\text{SVR}_1$ ranks highest in forecast accuracy among the 10 estimators 28.57% of the time; and ranks second and third highest 35.71% and 14.29% of the time. In summary, our results suggest that not only does social media data matter for forecasts, but so does how it is aggregated.

In Appendix C, we repeat the above exercises in Section 5 using a daily USSI in place of the hourly USSI data. The daily USSI is a simple weighted average of the hourly USSI, where the weights reflect the hourly volume of Tweets divided by the total volume per day. Similarly to investigate robustness, Appendix D presents results of forecasting CCI two months ahead and unsurprisingly forecast accuracy declines with dynamic forecasts since they involve more than one step ahead. Yet, the analysis in both of these exercises demonstrate the robustness of our results that find (i) incorporating USSI in forecasting the CCI is empirically important, and (ii) the superior performance with the H-MIDAS estimator.

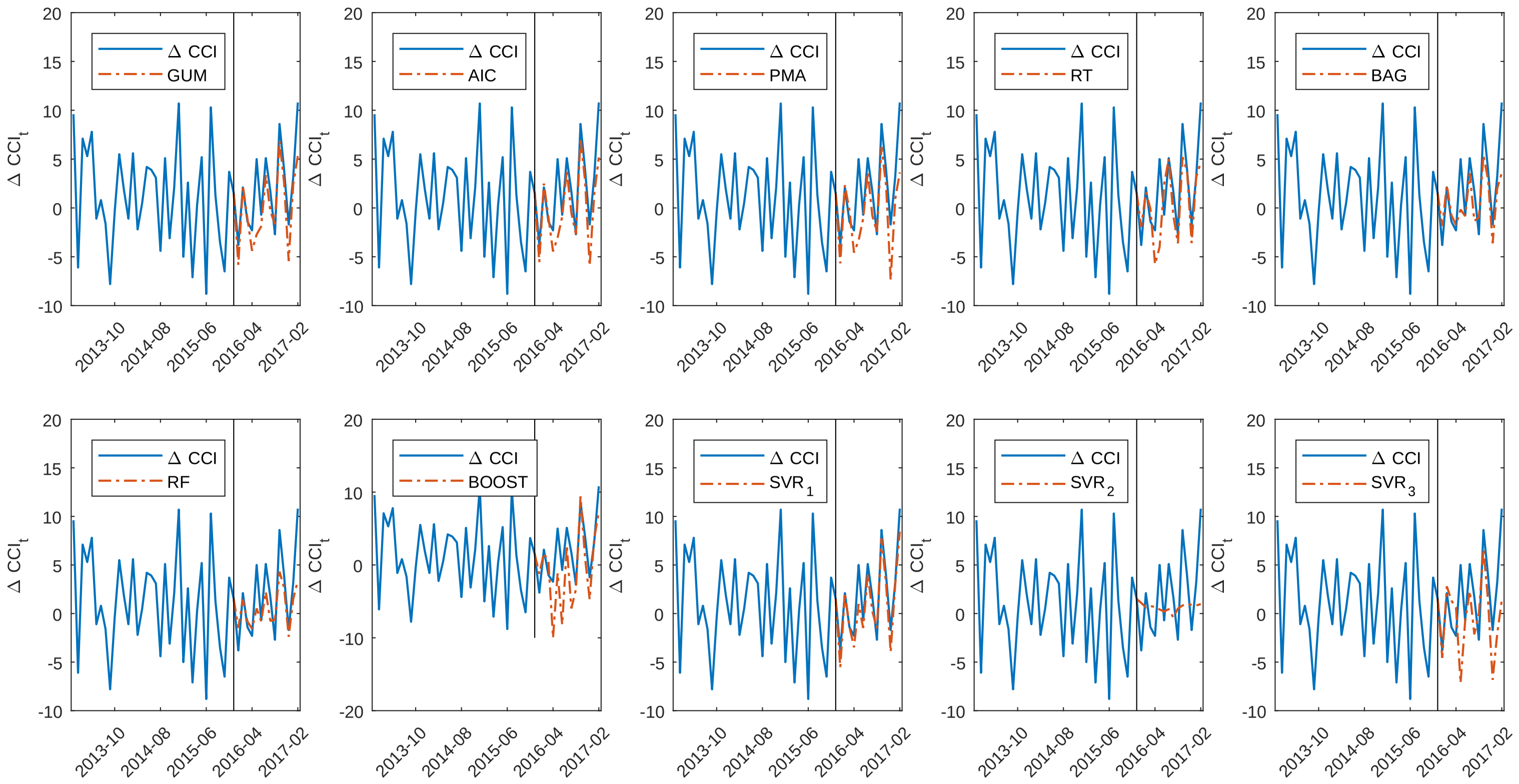
## 5.1 Additional Evidence of Benefits from Including Social Media Data

To further illustrate the benefits of including an appropriately transformed USSI measure as an explanatory variable when forecasting $\Delta\text{CCI}_t$, consider OLS estimates of the GUM specification

$$\begin{aligned}\Delta\text{CCI}_t &= \beta_0 + \beta_1\Delta\text{MCSI}_{t-1} + \beta_2\Delta\text{LEI}_{t-1} + \beta_3\text{UR}_{t-1} + \beta_4\Delta\text{SR}_{t-1} + \beta_5\Delta\text{CPI}_{t-1} \\ &\quad +\beta_6\Delta\text{SP500}_{t-1} + \beta_7\Delta\text{VIX}_{t-1} + \beta_8\Delta\text{USD}_{t-1} + \beta_9\Delta\text{TS}_{t-1} + \beta_{10}\text{USSI}_{s,t-1} + \epsilon_t. \qquad (8)\end{aligned}$$

Table 5 compares OLS estimates across nested specifications that either impose restrictions on some of the coefficients (i.e. all financial variables equal 0, etc.) or utilize different

Figure 3: Forecast Performance of the 10 Estimators Applying $\mathcal{M}_{new}$ as Input Data



aggregations of the USSI. Specifically, the subscript $s = a, m$, or $new$ that respectively represent the USSI converted by simple averaging, conventional MIDAS, and our proposed H-MIDAS method. Panels A to C of Table 5 present the estimated coefficient and associated standard error (in parenthesis) for each variable with variable names list on the first column. Panel D reports the centered $R^2$ and adjusted $R^2$ for each model.

The first two columns of Table 5 exclude the USSI. None of the macroeconomic and financial variables are statistically significant, with the sole exception of USD. Yet, an $F$-test of Model (1) is unable to reject the joint insignificance of all macroeconomic variables at the 10% level. This likely arises since for forecasting to be valid we must use a one-month lag of the macroeconomic variables. In contrast, the set of financial variables (transformed via conventional MIDAS) in Model (2) are jointly significant with a $p$-value 0.0008.[24]

Models (3) to (5) consider the sole inclusion of a single alternative USSI measure. In each specification, the respective USSI enters in a statistically significant manner but there are large differences across the columns in the magnitude of the effect. By comparing the associated $R_c^2$ and $\bar{R}^2$ values, we notice that the regression model containing the USSI created by H-MIDAS explains the most variation in the data.[25] GUM estimates with alternative USSI measures are presented in columns (6) to (8) of Table 5. Surprisingly, given the large marginal effect in Model (3), USSI$_a$ variable is statistically insignificant when one also conditions on macroeconomic and financial variables. This result is suggestive of high degrees of collinearity between the simple averaging USSI$_a$ and subsets of the macroeconomic and financial variables. The estimates in Models (7) and (8) demonstrate that there is unique variation in USSI$_m$ and USSI$_{new}$ and each of them enter in a statistically significant manner. Further, the coefficients in Models (7) and (8) do not differ markedly from those in Models (4) to (5), which increases our confidence that this is explaining

[24] The conference board releases CCI on the last Tuesday of each month at 10am. Since the macroeconomic variables are on the same frequency of CCI, we use one-month lags to avoid simultaneity and have a valid forecasting model. This information is reported approximately one month before the CCI is release. The financial variables, on the other hand, contain information up to one day before the release and such information can be preserved by the conventional MIDAS method to a higher degree than H-MIDAS which would give larger weight to more distant observations in the series. The difference in timing likely explains why financial variables have better forecasting performance than the macroeconomic variables.

[25] It should also be noted that Model (5) yields the highest $\bar{R}^2$ values among all 12 models.

Table 5: OLS Estimates of Models to Explain $\Delta \text{CCI}_t$

| Variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Panel A: Macroeconomic Variable* | | | | | | | | | | | | |
| MCSI | -0.0100 | 0.0676 | - | - | - | 0.1037 | 0.0434 | 0.1525 | 0.0593 | 0.1214 | 0.1500 | 0.1138 |
| | (0.2242) | (0.2250) | - | - | - | (0.2228) | (0.2043) | (0.1922) | (0.2077) | (0.1886) | (0.1950) | (0.1912) |
| LEI | -0.1923 | -0.0628 | - | - | - | 0.3375 | 0.0524 | 0.6466 | 0.2013 | 0.5648 | 0.6131 | 0.4790 |
| | (1.9373) | (1.9958) | - | - | - | (1.9824) | (1.8109) | (1.7035) | (1.8429) | (1.6647) | (1.7329) | (1.6905) |
| UR | -0.0792 | 0.0147 | - | - | - | -0.1302 | -0.2114 | -0.2058 | -0.2521 | -0.2871 | -0.1927 | -0.2591 |
| | (0.7700) | (0.7604) | - | - | - | (0.7546) | (0.6937) | (0.6479) | (0.7027) | (0.6347) | (0.6593) | (0.6436) |
| SR | -0.9563 | -1.2259 | - | - | - | -0.9294 | -0.7168 | -0.2898 | -0.6370 | -0.1976 | -0.3036 | -0.2266 |
| | (0.8595) | (0.8852) | - | - | - | (0.8932) | (0.8197) | (0.7862) | (0.8372) | (0.7698) | (0.7990) | (0.7798) |
| CPI | -1.1407 | -1.5440 | - | - | - | -1.1120 | -0.7204 | -1.1787 | -0.6099 | -0.7925 | -1.2257 | -0.8877 |
| | (1.5492) | (1.5732) | - | - | - | (1.5748) | (1.4517) | (1.3387) | (1.4754) | (1.3274) | (1.3757) | (1.3546) |
| *Panel B: Financial Variable* | | | | | | | | | | | | |
| SP500 | - | -0.0099 | - | - | - | -0.0133 | -0.0352 | 0.0352 | -0.0347 | 0.0117 | 0.0368 | 0.0146 |
| | - | (0.0922) | - | - | - | (0.0908) | (0.0840) | (0.0791) | (0.0847) | (0.0785) | (0.0805) | (0.0795) |
| VIX | - | 0.0473 | - | - | - | -0.1274 | 0.0504 | 0.2004 | -0.0184 | 0.1705 | 0.2275 | 0.2351 |
| | - | (1.3884) | - | - | - | (1.3716) | (1.2596) | (1.1794) | (1.2752) | (1.1522) | (1.2019) | (1.1712) |
| USD | - | 5.9640** | - | - | - | 5.2260* | 4.0757 | 3.0869 | 3.9216 | 2.6223 | 3.1064 | 2.6468 |
| | - | (2.6948) | - | - | - | (2.6973) | (2.5195) | (2.3951) | (2.5534) | (2.3556) | (2.4271) | (2.3799) |
| TS | - | 109.0637 | - | - | - | 82.0912 | 150.3913* | 47.9592 | 136.8333* | 83.7340 | 49.8604 | 90.1537 |
| | - | (85.5075) | - | - | - | (86.0450) | (78.7092) | (74.1418) | (82.5336) | (75.4351) | (75.6703) | (77.3005) |
| *Panel C: Big Data Variable* | | | | | | | | | | | | |
| $\text{USSI}_a$ | - | - | 0.7002** | - | - | 0.4988 | - | - | 0.1958 | - | -0.0656 | -0.1598 |
| | - | - | (0.2888) | - | - | (0.3303) | - | - | (0.3271) | - | (0.3278) | (0.3240) |
| $\text{USSI}_h$ | - | - | - | 0.3277*** | - | - | 0.3072*** | - | 0.2851*** | 0.1722* | - | 0.1808* |
| | - | - | - | (0.0866) | - | - | (0.0991) | - | (0.1065) | (0.1016) | - | (0.1041) |
| $\text{USSI}_{new}$ | - | - | - | - | 0.5479*** | - | - | 0.5105*** | - | 0.4051*** | 0.5242*** | 0.4334*** |
| | - | - | - | - | (0.1067) | - | - | (0.1257) | - | (0.1376) | (0.1447) | (0.1504) |
| *Panel D: Goodness of Fit* | | | | | | | | | | | | |
| $R^2_c$ | 0.0394 | 0.1802 | 0.1091 | 0.2297 | 0.3546 | 0.2255 | 0.3422 | 0.4238 | 0.3483 | 0.4643 | 0.4244 | 0.4678 |
| $\tilde{R}^2$ | -0.0698 | -0.0043 | 0.0905 | 0.2137 | 0.3412 | 0.0269 | 0.1735 | 0.2761 | 0.1597 | 0.3093 | 0.2578 | 0.2952 |

Each model contains different explanatory variables and "-" means the corresponding variables are not included in the specification.

* 10% level of significance.

** 5% level of significance.

*** 1% level of significance.

variation in the CCI that was not captured by traditional variables. Finally, the lack of gains when moving from allowing model uncertainty (comparing GUM to PMA columns) in Table 2 may arise from the absence of multiple significant regressors when the USSI$_{new}$ is not included as a regressor.

Last, Models (9) to (12) explore if there is additional value from including multiple USSI measures. Contrasting the estimates across these four columns suggests that there is unique explanatory power in USSI$_{new}$ relative to the other metrics. USSI$_{new}$ always enters in a statistically significant manner.[26]

Overall, the results in Table 5 reinforce the importance of incorporating big data variables on forecasting CCI. The big data series contains information up to one hour prior to the release of the CCI. The series generate significant explanatory power on CCI as the values of $R_c^2$ increase sharply when big data variables are included. Most importantly, the results in Models (7) and (8) demonstrate the necessity of converting higher frequency data to low frequency with sophisticated econometric techniques like MIDAS. When comparing the performance of models that either include the USSI$_m$, USSI$_{new}$, or the simple USSI$_a$ variables, we find that valuable information contained in specific increment of the higher frequency interval can be diluted by simple averaging.

Further, the improved forecast accuracy observed in Table 4 relative to Table 3 across all metrics and estimators points out that allowing for more flexible weights can capture the unsystematic manner by which time-varying conditions underlying these financial and social media measures truly impact consumer confidence. This provides additional intuition for the potential wider applicability of the H-MIDAS estimator since it is not restricted by a functional form assumption. In summary, extracting useful information lurking in the higher frequency data is challenging, but by imposing weaker assumptions when aggregating high frequency data can lead to large rewards in forecast accuracy.

[26] The analysis also indicates a high degree of correlation between USSI$_m$ and USSI$_d$.

# 6 Conclusion

Petabytes of new text data are created every second on social media and it remains an open question if measures extracted from anonymized social media data can help improve our ability to predict future values of the economic indicators ahead of the release of statistical data. However, an additional challenge may arise since social media data differs sharply from other macroeconomic and financial time series in manners beyond simply being text. To incorporate the high frequency part of social media data we propose a new MIDAS strategy that allows for greater heterogeneity in the weights across time, thereby allowing for a more gradual depreciation relative to the common implementation of the mixed data sampling approach. Using both forecasting models from the econometrics and machine learning literature, we provide evidence that incorporating sentiment measures from Twitter greatly improves forecast accuracy of the CCI.

Further, we find major improvements from using our proposed H-MIDAS strategy over other approaches to collapse high frequency data to a single measure. While developed for social media data, our forecasting exercise also shows that there are substantial benefits to using the H-MIDAS on financial variables. An additional advantage of the H-MIDAS estimator is that it allows for a unrestricted step-function to choose weights on elements within a series of the high frequency predictor variables, thereby not imposing arbitrary functional form assumptions that are implicitly embedded with conventional MIDAS strategies. We believe this method can offer substantial benefits in other forecasting exercises within the banking and finance industry.

For practitioners, the evidence in this study suggests that exploiting social media data may provide individuals and firms across numerous industries including banking and finance an advantage to enhance their forecasting capabilities. That said, future research needs to consider developing new tools that may help forecasters gain further advantage as well as investigate forecasting financial measures that are measured at a higher frequency level. On the former, one could consider as an alternative to using a statistical approach to weight Twitter sentiment across periods as in H-MIDAS, it may be interesting to examine how weights derived from either Twitter volume or from the historical timing of survey responses perform in forecasts of consumer confidence. That is, if 8% of surveys used to construct the CCI are historically mailed by survey respondents 11 days prior to the release of CCI, we could assign a weight of 8% to Twitter sentiment measured 11 days prior.

Further, future researchers could consider treat social sentiment as multidimensional rather than a single sentiment score. For example, one could measure mood from subset of tweets based on subgroups characterized by age or occupation or even whether the Twitter message has a positive or negative orientation. By unpacking the USSI in to its components, one could understand what type of emotions conveyed in individual tweets is associated with consumer confidence. In summary, this paper illustrates how in the big data era, many new innovations in the forecasters' toolbox will need to emerge to extract the full potential of these data new sources to improve forecasts of variable of interest to the banking and finance industry.

## References

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

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

ATHEY, S. AND G. W. IMBENS (2017): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31, 3–32.

BAELE, L., G. BEKAERT, AND K. INGHELBRECHT (2010): "The Determinants of Stock and Bond Return Comovements," *The Review of Financial Studies*, 23, 2374–2428.

BAJARI, P., D. NEKIPELOV, S. P. RYAN, AND M. YANG (2015): "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105, 481–85.

BAKER, M. AND J. WURGLER (2012): "Comovement and Predictability Relationships Between Bonds and the Cross-section of Stocks," *The Review of Asset Pricing Studies*, 2, 57–87.

BOLLEN, J., H. MAO, AND X. ZHENG (2011): "Twitter Mood Predicts the Stock Market," *Journal of Computational Science*, 2, 1–8.

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.

BROWN, G. W. AND M. T. CLIFF (2004): "Investor sentiment and the near-term stock market," *Journal of Empirical Finance*, 11, 1 – 27.

CHAN, F., R. B. DURAND, J. KHUU, AND L. A. SMALES (2017): "The Validity of Investor Sentiment Proxies," *International Review of Finance*, 17, 473–477.

CORSI, F. (2009): "A Simple Approximate Long-Memory Model of Realized Volatility," *Journal of Financial Econometrics*, 7, 174–196.

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.

FELBO, B., A. MISLOVE, A. S. OGAARD, I. RAHWAN, AND S. LEHMANN (2017): "Using Millions of Emoji Occurrences to Learn Any-domain Representations for Detecting Sentiment, Emotion and Sarcasm," *Machine Learning*, Accepted at EMNLP 2017.

García, D. (2013): "Sentiment during Recessions," *The Journal of Finance*, 68, 1267–1300.

Ghysels, E., P. Santa-Clara, and R. Valkanov (2004): "The MIDAS Touch: Mixed Data Sampling Regression Models," CIRANO Working Papers 2004s-20, CIRANO.

——— (2005): "There is a risk-return trade-off after all," *Journal of Financial Economics*, 76, 509 – 548.

——— (2006): "Predicting volatility: getting the most out of return data sampled at different frequencies," *Journal of Econometrics*, 131, 59 – 95.

Ghysels, E., A. Sinko, and R. Valkanov (2007): "MIDAS Regressions: Further Results and New Directions," *Econometric Reviews*, 26, 53–90.

Greenwood, S., A. Perrin, and M. Duggan (2016): "Social Media Update 2016," *Pew Research Center*.

Karabulut, Y. (2013): "Can Facebook Predict Stock Market Activity?" *Working Paper*.

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.

Lehrer, S. F. and T. Xie (2017): "Box Office Buzz: Does Social Media Data Steal the Show from Model Uncertainty When Forecasting for Hollywood?" *The Review of Economics and Statistics*, 99, 749–755.

——— (2018): "The Bigger Picture: Combining Econometrics with Analytics Improve Forecasts of Movie Success," *NBER Working Paper: w24755*.

Lehrer, S. F., T. Xie, and X. Zhang (2019): "Twits versus Tweets: Does Adding Social Media Wisdom Trump Admitting Ignorance when Forecasting the CBOE VIX?" *Working Paper*.

Lemmon, M. and E. Portniaguina (2006): "Consumer Confidence and Asset Prices: Some Empirical Evidence," *The Review of Financial Studies*, 19, 1499–1529.

Malik, M., H. Lamba, C. Nakos, and J. Pfeffer (2015): "Population Bias in Geotagged Tweets," *International AAAI Conference on Web and Social Media*.

Mclean, R. D. and M. Zhao (2014): "The Business Cycle, Investor Sentiment, and Costly External Finance," *The Journal of Finance*, 69, 1377–1409.

MISHNE, G. AND N. GLANCE (2006): "Predicting Movie Sales from Blogger Sentiment," *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*.

MISLOVE, A., S. L. JØRGENSEN, Y.-Y. AHN, J.-P. ONNELA, AND J. N. ROSENQUIST (2011): "Understanding the Demographics of Twitter Users," *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 554–557.

MULLAINATHAN, S. AND J. SPIESS (2017): "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87–106.

O'CONNOR, B., R. BALASUBRAMANYAN, B. R. ROUTLEDGE, AND N. A. SMITH (2010): "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series." in *ICWSM*, ed. by W. W. Cohen and S. Gosling, The AAAI Press.

STAMBAUGH, R. F., J. YU, AND Y. YUAN (2012): "The short of it: Investor sentiment and anomalies," *Journal of Financial Economics*, 104, 288 – 302, special Issue on Investor Sentiment.

——— (2014): "The long of it: Odds that investor sentiment spuriously predicts anomaly returns," *Journal of Financial Economics*, 114, 613 – 619.

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

ZUMBRUN, J. (2017): "A New Index Tracks Our National Mood One Tweet at a Time," *The Wall Stree Journal (online)*, `https://blogs.wsj.com/economics/2017/05/08/a-new-index-tracks-our-national-mood-one-tweet-at-a-time/`.

# 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 $w_j\left(\boldsymbol{\theta}\right) = \theta_j$, if $1 \leq j \leq p-1$, and $w_p\left(\boldsymbol{\theta}\right) = 1 - \sum_{k=1}^{p-1} \theta_k$, note that the first order derivative of the weights can be expressed as

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

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

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

and the regression problem can be viewed as

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

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

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

We derive the asymptotic variance of the estimator of $\beta$ 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 $e_{t-\frac{i}{m}} \overset{iid}{\sim} \left(0, \sigma_e^2\right)$. The asymptotic variance of the estimated coefficient $\hat{\beta}$ is*

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

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

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

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

*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 $\beta$ 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 $\mathrm{Cov}\left(\bar{X}_t^{(l_i)}, \bar{X}_t^{(l_s)}\right)$, 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

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

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

Since the $D$ term can be explicitly written as

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

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

the $E$ term can be expressed as

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

the $F$ term can be shown as

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

and the $G$ term is simply

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

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 $\beta$.

**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 $\beta$, 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 $X^B_t$.

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

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,

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

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 $\beta_j^h$ 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 $\text{SSR}^h$ 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, $\mathrm{SSR}_0$. 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, $\lambda$ is a more traditional regularization parameter, that can be estimated by cross-validation.

Let $\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_0,\, \hat{\boldsymbol{\beta}}_1^\top\right]^\top$ 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 $t = 1, \ldots, n$. We usually called the non-zero values of $\hat{\boldsymbol{\alpha}}_t^* - \hat{\boldsymbol{\alpha}}_t$ for $t = 1, \ldots, n$ 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 $m = 1, \ldots, M$:

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

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

$$\begin{aligned} K(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) &= \exp\left(-\|\boldsymbol{X}_t - \boldsymbol{X}_{t'}^\top\|^2\right), & \text{(A13)} \\ K(\boldsymbol{X}_t, \boldsymbol{X}_{t'}) &= \left(1 + \boldsymbol{X}_t\boldsymbol{X}_{t'}^\top\right)^p \quad \text{with } p \in \{2, 3, \dots\}, & \text{(A14)} \end{aligned}$$

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

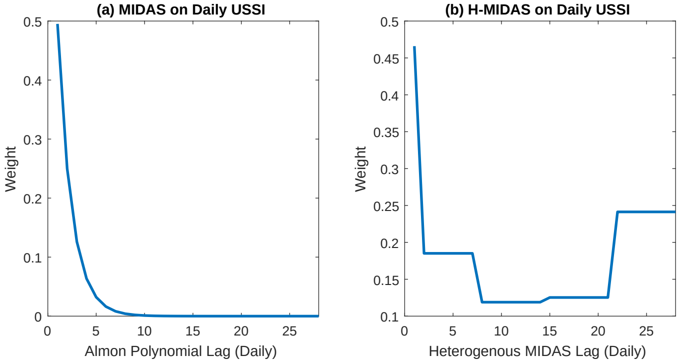
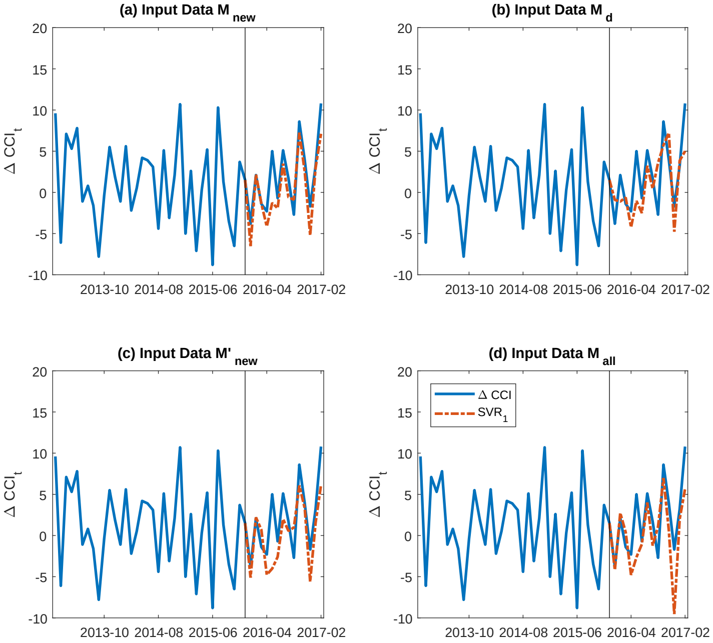


Figure A2: Forecasting Performance of $\mathrm{SVR}_1$ Using Daily USSI as Input Data



## 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* | | | | | | | | |
| $\text{USSI}_a$ | - | - | - | - | - | - | - | -0.1867 |
| | - | - | - | - | - | - | - | (0.3308) |
| $\text{USSI}_h$ | - | - | - | - | - | - | - | 0.0994 |
| | - | - | - | - | - | - | - | (0.1602) |
| $\text{USSI}_{new}$ | 0.5479$^{\diamond}$ | - | - | - | - | - | 0.2407 | 0.2759 |
| | (0.1067) | - | - | - | - | - | (0.2203) | (0.2301) |
| $\text{USSI}_d$ | - | 0.4262$^{\diamond}$ | - | 0.3740$^{\diamond}$ | - | 0.1683 | 0.1595 | 0.0849 |
| | - | (0.0940) | - | (0.1077) | - | (0.1336) | (0.1335) | (0.1982) |
| $\text{USSI}'_{new}$ | - | - | 0.5167$^{\diamond}$ | - | 0.4675$^{\diamond}$ | 0.3481$^{\dagger}$ | 0.1750 | 0.1686 |
| | - | - | (0.1021) | - | (0.1124) | (0.1464) | (0.2154) | (0.2198) |
| *Panel D: Goodness of Fit* | | | | | | | | |
| $R_c^2$ | 0.3546 | 0.2997 | 0.3480 | 0.3738 | 0.4322 | 0.4549 | 0.4720 | 0.4821 |
| $\tilde{R}^2$ | (0.3412) | (0.2851) | (0.3345) | (0.2132) | (0.2865) | (0.2971) | (0.3007) | (0.2750) |

$^{*}$ 10% level of significance.
$^{\dagger}$ 5% level of significance.
$^{\diamond}$ 1% level of significance.

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

| | GUM | AIC | PMA | RT | BAG | RF | BOOST | $\text{SVR}_1$ | $\text{SVR}_2$ | $\text{SVR}_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| *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**$^{\diamond}$ | 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**$^{\diamond}$ | 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 $\diamond$ 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, $\text{SVR}_1$ 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 | $\text{SVR}_1$ | $\text{SVR}_2$ | $\text{SVR}_3$ |
|---|---|---|---|---|---|---|---|---|---|---|
| *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.