Received: 26 October 2020 | Revised: 27 March 2021 | Accepted: 30 March 2021

DOI: 10.1002/jae.2859

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

News media versus FRED-MD for macroeconomic forecasting

Jon Ellingsen[1] | **Vegard H. Larsen**[2] | **Leif Anders Thorsrud**[1]

[1]Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School, Oslo, Norway

[2]Norges Bank and Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School, Oslo, Norway

Correspondence
Leif Anders Thorsrud, Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School, Oslo, Norway.
Email: leif.a.thorsrud@bi.no

Summary

Using a unique dataset of 22.5 million news articles from the *Dow Jones Newswires Archive*, we perform an in depth real-time out-of-sample forecasting comparison study with one of the most widely used data sets in the newer forecasting literature, namely the *FRED-MD* dataset. Focusing on US GDP, consumption and investment growth, our results suggest that the news data contains information not captured by the hard economic indicators, and that the news-based data are particularly informative for forecasting consumption developments.

KEYWORDS

Forecasting, Machine Learning, News, Real-time, Text data

1 | INTRODUCTION

During the last decades advances in econometric techniques have substantially improved short-term forecasting performance in economics (Aastveit et al., 2014; Ghysels et al., 2004; Giannone et al., 2008; Stock & Watson, 2002). However, while much research has leveraged the qualities of traditional economic data to construct new and better models, less attention has been given to new and alternative data sources (Varian, 2014).

In this paper, we use a unique corpus of 22.5 million news articles from the *Dow Jones Newswires Archive*to perform an in depth real-time out-of-sample (OOS) macroeconomic forecasting comparison study with what has become the "industry standard" in the newer forecasting literature, namely the *FRED-MD* dataset. This dataset is compiled by McCracken and Ng (2016), contains real-time vintages (from 1999) of over 100 monthly (leading) economic indicators, and builds upon the seminal contribution by Stock and Watson (1989), and the literature that followed, using large datasets for macroeconomic forecasting and monitoring.

Intuitively, what we simply denote as news data has several appealing features compared to traditional (hard) economic statistics. First, news data is available at a high frequency allowing forecasts to be updated without a time-lag, which is often an issue when working with traditional economic data (Giannone et al., 2008). Second, the news covers a broad set of topics and thus provide a narrative about economy-wide developments (Larsen & Thorsrud, 2018). In contrast, traditional high-frequency economic data mostly covers financial markets. These are important markets, but their predictive power for macroeconomic developments have been proven to be unstable (Stock & Watson, 2003). Third, from an informational perspective, one could argue that news data potentially provides a better description of the information agents, at least households, actually have when forming expectations (Larsen et al., 2021). Thus, as expectations translate into outcomes, using news might be beneficial. Likewise, news data might capture stories and developments that are not easily measured

by traditional economic data, for example, politics and uncertainty (Baker et al., 2016), making it a useful supplement for capturing the complexity of expectations (Sims, 2003). In addition, news data are not subject to revisions which often influence forecasting performance negatively when using traditional economic statistics (Croushore & Stark, 2001).[1]

Still, the raw news data is textual, unstructured, and high-dimensional. In economics, the most prevalent way of turning this type of data into quantitative time series has been to use dictionary- or Boolean-based techniques (Bholat et al., 2015). These methods essentially searches through the text and counts specific words. This has been shown to work well when one knows exactly what to search for, but is less suited when the underlying signal might be multifaceted, as here. For this reason, we decompose the text data into something relatively small, dense, and interpretable, using a Latent Dirichlet Allocation (LDA; Blei et al., 2003) topic model.

The LDA is one of the most popular topic models in the Natural Language Processing (NLP) literature, and treats articles as a mixture of topics, and topics as a mixture of words. It automatically classifies text in much the same manner as humans would (Chang et al., 2009) and is also proposed as a valuable tool in recent economic research using text as data, including, for example, business cycle and monetary policy analysis (Hansen & McMahon, 2016; Hansen et al., 2018; Larsen & Thorsrud, 2019; Thorsrud, 2018).[2] Compared to many other NLP methods, and despite being an unsupervised algorithm, the LDA has the attractive feature of delivering interpretable output. Thus, the narrative realism of the approach can be validated since the topics have narrative content.

In total, we extract 80 topics from the corpus. These topics cover a broad set of economic narratives, ranging from politics and trade to finance and health, and are transformed into monthly time series measuring how much the media reports on the different topics across time. For example, if something newsworthy happens in the oil market, the hypothesis is that oil market related topics spike relative to the other topics and that this variation across time can be informative about current and future economic developments.

We focus on nowcasting (Banbura et al., 2011) and short-term predictions of quarterly US GDP, consumption, and investment growth, and leverage the news data's large scope to evaluate more than two decades of OOS performance.

To form predictions, off the shelf, but state-of-the-art, Machine Learning (ML) and econometric forecasting techniques are combined and applied. The unrestricted MIDAS (Foroni et al., 2015; Ghysels et al., 2004) is used to bridge the frequency gap between the quarterly outcome variables and the monthly predictors, while the Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani, 1996), Principal Component Analysis (PCA; Stock & Watson, 1989), and the Random Forest (RF; Breiman, 2001) are used to handle the high dimensionality of the predictive problem and potential non-linearities.

The forecasting horse-race design is simple. First, separate models with either the news or hard economic data are estimated, and then their OOS point forecasting accuracy is evaluated ex post. Next, to mimic a more realistic forecasting process, simple forecast combination schemes and aggregated models, including all the data, are considered. To avoid look-ahead biases, all experiments are conducted using real-time data, and the LDA is only estimated using data from an initial training sample. To facilitate the comparison with the *FRED-MD* data, all the predictors are recorded on a monthly basis, although the news data has the potential benefit of being available on a higher frequency.

We reach three main conclusions: First, relative to the hard economic indicators, the news data has a (significant) lower forecast error variance when predicting consumption developments, but is inferior in terms of predicting investment developments. For GDP, we do not find any statistically significant differences between the two datasets. Likewise, when optimally combining forecasts recursively throughout the evaluation sample, without the benefit of ex post knowing the best data, the models containing news-based predictors consistently obtain a higher weight than the models containing hard economic indicators, at least when predicting GDP and consumption growth.

Second, consistent with the view that news affects economic agent's expectations about the future (Larsen et al., 2021), the news data seems to be more forward-looking than the hard economic indicators. The best performance of the news data relative to the hard economic indicators, for example, is obtained when doing one-quarter ahead consumption predictions. It is also a general pattern that the news-data is more informative in the beginning of any given quarter, when little hard economic information is available, than toward the end of the quarter.

Third, we find that the news-based predictors are more short-lived and sparse relative to the hard-based predictors. Still, the narrative realism of the news-based predictive approach is good. For example, on average across the evaluation sample,

[1] The news data also has a clear benefit over other high-frequency alternative data sources, such as social media or Internet search volume, whose usage might lead to unreliable inference because long time series for such data do not exist (Lazer et al., 2014).

[2] Similar in spirit to the earlier work by Larsen and Thorsrud (2019) and Thorsrud (2018), Bybee et al. (2019) also apply the LDA to describe how news data can provide meaningful signals about economic developments in the United States.

topics related to *Personal finance*, *Health care*, and *Bond market* all receive a high weight when predicting consumption developments.

This analysis speaks to a growing literature entertaining text as data in economics (see Gentzkow et al., 2019 for an overview) and a large economic (short-term) forecasting literature. The work most closely related to ours are recent research by Ulbricht et al. (2017), Ardia et al. (2019), and Kalamara et al. (2020). They propose and test (news) text-based (sentiment) indicators for economic forecasting, and focus on predicting developments in industrial production and other macroeconomic variables in Germany, the United States, and the United Kingdom, respectively.

We contribute along several dimensions: First, we contribute by performing the first in depth OOS forecasting comparison experiment with news and the much used *FRED-MD* dataset. Accordingly, all our results are new in the literature and establishes several "stylized facts." These are not only useful for future research on the topic but also relevant for practitioners wanting to improve short-term forecasting performance. We show, for example, that when something abrupt happens and expectations change rapidly, like during and after the Great Recession episode, the value of news seems especially high relative to the hard economic indicators. In contrast, when evaluating the news-based performance to publicly available soft data (Giannone et al., 2009) in the form of the Survey of Professional Forecasters (SPF), we find little evidence suggesting that the news-based predictions constructed here are superior. From a cost benefit perspective, this might make the usage of news topics less attractive. Still, the news-based data is available at a much higher frequency than typical soft data, and although we deliberately have not utilized this property in this article, related research suggests that the news-based signal-to-noise ratio remains solid when increasing the sampling frequency (Larsen & Thorsrud, 2018; Thorsrud, 2018).

By using ML techniques to form predictions our analysis also relates to recent research by Medeiros et al. (2019) and Babii et al. (2020). Whereas Medeiros et al. (2019) use the *FRED-MD* dataset to compare ML models for inflation forecasting, we focus on the (textual) news versus hard economic data dimension when forecasting National Account Statistics. Interestingly, our study complement theirs in terms of documenting that the (news-based) RF method is better than both the LASSO and the PCA across nearly all outcome variables and forecasting horizons. In contrast, Babii et al. (2020) propose a new sparse-group LASSO estimator, and show that it performs favorably compared to other alternatives, especially when combined with using text as data, when nowcasting US GDP growth.

Finally, our analysis casts light on the role of the media in the expectation formation process of economic agents. This has been a relatively unexplored field in (macro)economics, but studies by, for example, Carroll (2003), Nimark and Pitschner (2019), and Larsen et al. (2021), show how the media channel might be important both in practice and in theory. In particular, under the assumption that consumption and investment decisions are mostly done by households and professionals, respectively, our results are consistent with Larsen et al. (2021) who find that news has good predictive power for households' inflation expectations, but much less so for expectations among professional forecasters.

The rest of this article is organized as follows: In Section 2 we describe the data and the LDA. Section 3 describes the models and experiment used for prediction and evaluation, while Section 4 presents the results. Section 5 concludes.

2 | DATA

In the following the news data and how these are transformed into time series objects are presented. We describe the outcome variables, the hard-based economic indicators, and provide simple descriptive statistics comparing the two datasets.

2.1 | News data and topics

The news data consists of news articles from the *Dow Jones Newswires Archive (DJ)* for the period January 1985 to April 2020. The unique feature with this corpus, that is, the text and articles, is its coverage in terms of time span and the broad scope of news reported. In total we have access to roughly 22.5 million news articles and over 1.5 million unique terms. All text is business-focused and written in English and covers a large range of *Dow Jones's* news services, including content from *The Wall Street Journal*. The Dow Jones company is one of the leading international providers of business news, while *The Wall Street Journal* is one of the largest newspapers in the United States in terms of circulation and naturally leaves a large footprint in the US media landscape.

The textual data is high-dimensional and unstructured. This makes statistical computations challenging. Therefore, as is common in the NLP literature, the news corpus is cleaned prior to estimation. We remove stop-words, conduct

stemming, and apply term frequency—inverse document frequency calculations. A more detailed description of these steps is given in Appendix SD.

The cleaned text corpus is decomposed into news topics using a LDA model (Blei et al., 2003). The LDA is an unsupervised model that clusters words into topics, which are distributions over words, while at the same time classifying articles as mixtures of topics. It is one of the most popular topic algorithms in the NLP literature and used here because of its simplicity, because it has proven to classify text in much the same manner as humans would do (Chang et al., 2009), and because it delivers interpretable output. For these reasons it has also been one of the most widely used NLP algorithms in recent economic applications (Dybowski & Adämmer, 2018; Hansen & McMahon, 2016; Hansen et al., 2018; Larsen, 2021; Larsen & Thorsrud, 2017).

From a forecasting perspective, it is worth noting that the LDA shares many features with latent (Gaussian) factor models used with success in conventional economic forecasting applications, but with factors (representing topics) constrained to live in the simplex and fed through a multinomial likelihood at the observation equation. Appendix SB provides a brief description on how the LDA is implemented here, while Blei (2012) provides a nice layman introduction to topic modeling in general and more technical expositions of the LDA approach can be found in, for example, Blei et al. (2003) and Griffiths and Steyvers (2004).

How many topics to extract when estimating the LDA is a choice variable, just as deciding how many factors to use in conventional exploratory factor analysis. We use 80 topics in the main analysis, and discuss how our results are robust to other choices in Section 4.5.

Finally, the output of the LDA topic decomposition is transformed into time series. The LDA produces two outputs; one distribution of topics for each article in the corpus, and one distribution of words for each of the topics. Using the former distributions, each day in the sample is given a topic weight, measuring how much each topic is written about on that particular day. Thus, while the time series will sum to one on any given day, they can vary substantially in terms of their relative weights across time. Our simple hypothesis is that this variation across time can be informative about current and future economic developments. To align the frequency of topic observations to those available for the *FRED-MD*data, these statistics are then aggregated to a monthly frequency using the mean of the daily weights.

To build intuition, Figure 1 illustrates the output from the above steps for six of the 80 topics. A full list of the estimated topics is given in Table SA2. The LDA topic distributions are illustrated using word clouds. A bigger font illustrates a higher probability for the terms. As the LDA estimation procedure does not give the topics any name, labels are subjectively given to each topic based on the most important terms associated with each topic. How much each topic is written about at any given point in time is illustrated in the graphs below each word cloud. Since the time series in the graphs are normalized, they should be read as follows: Progressively more positive (negative) values means the media writes more (less) than on average about this topic.

To help interpretation, one could also interpret each topic as belonging to clusters of higher order abstractions, like, politics, technology, etc. This is illustrated in Figure SA1, where a clustering algorithm has been used to group the topics into broader categories. For example, the *Korea*, *China*, and *Trade* topics are automatically grouped together, making it apparent that these news types are related to trade and East-Asia. As news stories and narratives are not based on only one topic, viewing them as belonging to higher order abstraction like this can be useful.

2.2 | Outcomes and hard economic variables

The outcome variables are monthly real-time vintages of real GDP (*GDP*), real personal consumption expenditures (*Consumption*) and real gross private domestic nonresidential investment (*Investment*), obtained from the *ALFRED* (ArchivaL Federal Reserve Economic Data) real-time database maintained by the *Federal Reserve Bank of St. Louis* (Croushore & Stark, 2001). By institutional convention, the first release of a given quarter is published in the second month of the subsequent quarter and revisions two and three are published in the following months. Prior to estimation, all the outcome variables are transformed to quarterly percentage (log) growth rates.

Monthly real-time economic predictors are obtained from the same source and contains data from the *FRED-MD* dataset defined by McCracken and Ng (2016). Both the outcome variables and the *FRED-MD* variables are collected to span the same time period as the news data, that is, January 1985 to April 2020. The *FRED-MD* dataset is one of the most widely used in the newer forecasting literature and the vintage published in April 2020 contains well over 100 monthly

FIGURE 1 Topic distributions and time series. For each topic, the size of a word in the word cloud reflects the probability of this word occurring in the topic. Each word cloud only contains a subset of all the most important words in the topic distribution. Topic labels are subjectively given. The topic time series are linearly detrended and normalized. January 1985 to April 2020

(leading) economic indicators. This includes output, consumption, and income statistics, labor market data, housing data, money, credit, and interest rates, prices, and stock market indicators. The data is transformed following the transformation scheme used in Medeiros et al. (2019). Table SA1 provides the details. Each monthly real-time vintage contains data that was available by the end of that month, but with potential missing values due to differences in the release calendar across variables. That is, the real-time *FRED-MD* dataset is unbalanced and contains so-called ragged-edges. If a given variable does not exist for an earlier vintage or time period, or has missing data, the series from the first succeeding vintage that contains the variable is used and truncated such that the variable follows the same release pattern as usual. Similarly, the real-time vintage version of the *FRED-MD* dataset only goes back to 1999, and we construct pseudo-real-time vintages for the pre 1999 periods using the same logic as above for variables and vintages with missing data.

2.3 | Descriptive statistics

In terms of simple descriptive statistics, and using the final vintage of the *FRED-MD* data, Figure 2 shows that there are noticeable differences between the news- and hard-based time series data. As a group, the news topic time series tend to be more negatively skewed compared to the hard-based data, and, as seen from the kurtosis plot, the news-based data is by far much more outlier-prone. The news data also tend to be much less auto-correlated than the hard-based data.

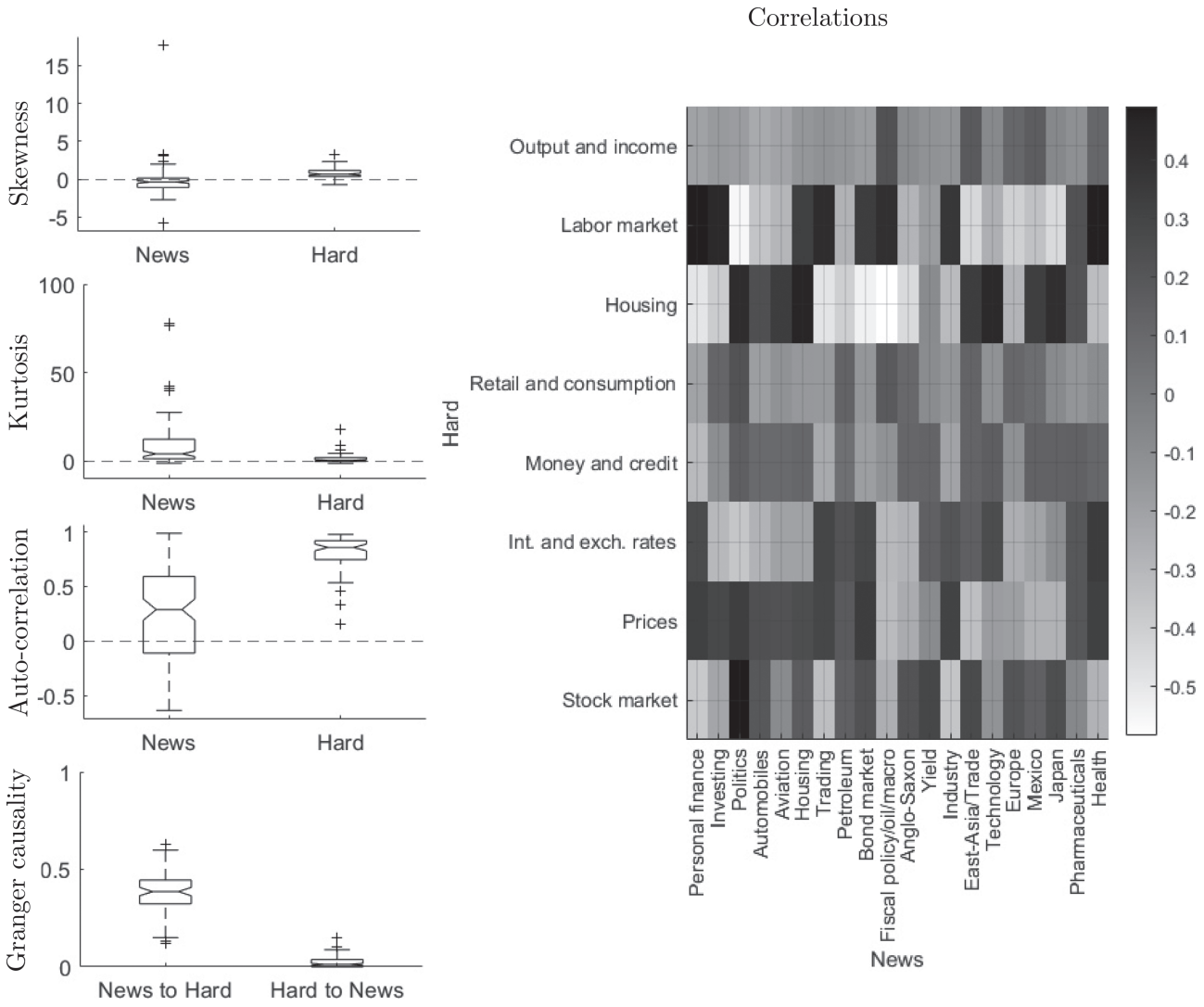


FIGURE 2 Descriptive statistics. The box plots report skewness, kurtosis and the first-order auto-correlation, where the skewness and kurtosis of the normal distribution is defined to be 0. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the remaining data points excluding outliers, which are plotted individually using the + symbol. The correlation image reports the largest (negative/positive) correlation among variables within the 8 and 20 subgroups of the *FRED-MD* and *DJ* datasets, respectively, where the news topic subgroups are constructed using a hierarchical agglomerative clustering algorithm (Figure SA1)

Still, although the datasets differ in terms of simple descriptive statistics, they share some narrative plausible correlation patterns. This is illustrated in the correlation image in Figure 2. For readability, the graph shows the largest (negative/positive) correlation between the news- and hard-based data within the higher order groups they belong to, where the *FRED-MD* group names are given by the structure of the database and the news topic subgroups are those discussed above, see also Table SA1 and Figure SA1. The statistics show that the correlation between news topics and hard-based variables in the *Labor market*, *Housing*, and *Stock market* groups tend to be especially high. Variables in the former group, for example, have a fairly high positive/negative correlation with news topics related to *Personal finance*, *Investing*, *Politics*, and *Health*, whereas hard-based housing and stock market variables are most strongly positively correlated with topics in the *Housing* and *Politics*clusters, respectively. In contrast, the correlation between news topics and hard-based variables related to *Money and credit*, *Retail and consumption*, and *Output and income* tend to be low.

Finally, the box plots in the lower left corner of the figure shows that it is more common that the news-based data Granger causes the hard-based once than vice versa.[3] For example, on average a news topic Granger causes almost 40% of the hard-based variables, whereas a hard-based variable at best Granger causes less than 20% of the news-based data. The results suggest, or at least do not rule out, that news reporting captures economic developments that eventually show up in economic statistics or even affect the outcome of such statistics. Interestingly, although this latter point is not the focus of this article per se, it speaks to a long standing literature in economics analyzing the more structural relationship between news and economic fluctuations. See, for example, Larsen and Thorsrud (2019) and Larsen and Thorsrud (2018) for recent contributions providing evidence consistent with ours, and Wu et al. (2004) for a relatively early example and literature review.

In sum, although the two datasets share interesting correlation patterns, they also clearly differ in terms of simple descriptive statistics and time series patters. The question then becomes whether these differences are useful for forecasting macroeconomic aggregates.

3 | EXPERIMENTAL SETUP

The predictor datasets, *FRED-MD* and *DJ*, are recorded on a monthly frequency, while the outcome variables *GDP*, *Consumption*, and *Investment*, are quarterly. To make use of the high-frequency information captured by the predictors we apply the unrestricted MIDAS technology (Foroni et al., 2015; Ghysels et al., 2004).

Formally, let the quarterly time index be t, and m the number of times the higher sampling frequency (months) appears in the low frequency time unit (quarters). Denote the low frequency outcome variable of interest y_t^L and let a high-frequency predictor be denoted $x_{t-j/m}$, where j represents lags. Then, the unrestricted MIDAS, for a single predictor and forecasting horizon h, has the following form

$$y_{t+h}^L = a_h + \sum_{j=0}^{p} \beta_{j,h} L^{j/m} x_t + \varepsilon_{t+h}^L, \tag{1}$$

where p denotes the number of lags and L is the lag operator.

The MIDAS model is simple, popular, and has proven to produce very good predictions in a wide range of applications (Clements & Galvão, 2008; Foroni et al., 2015; Ghysels et al., 2004). When the set of predictors is low dimensional, estimation can be done by ordinary least squares (OLS). Here, where the set of predictors is large, this is not feasible. For this reason (1) is estimated using three different approaches; LASSO (Tibshirani, 1996); RF (Breiman, 2001); PCA on the predictor set coupled with OLS on a factor augmented version of (1). Individually these methods allow for regularization, potential non-linearities, and dimension reduction. While factor-augmented predictive approaches are well known in the econometrics literature, the usage of the LASSO and the RF methods are more common in ML.

In the interest of preserving space, a description of each estimation method is relegated to Appendix SC. In short, we use fivefold cross validation to tune the amount of regularization in the LASSO, and 500 bootstrap samples and 1/3 of the predictors as the random subset when estimating the RF. For estimating the factors we have explored using the EM algorithm from Stock and Watson (2002) together with the information criterion suggested in Bai and Ng (2002) to determine the numbers of factors, but find that using a fixed number of one factor produces better results.

The OOS forecasting experiment is conducted as follows. For each monthly vintage of the quarterly outcome variables, the predictive models are estimated using vintages of monthly data available at the end of either month one (M1), two (M2), or three (M3) of the quarter. In the benchmark case the models are estimated using either the *DJ* or *FRED-MD* dataset, but, as described later, we also consider a merged dataset and a forecast combination scheme. Next, predictions for the nowcast (H0; $h = 0$, that is, the current quarter), one- (H1; $h = 1$), and two-quarter ahead (H2; $h = 2$) horizons are produced. Since (1) is a direct forecasting equation, separate regressions are estimated for each forecasting horizon. Because of the release calendar of the National Account Statistics this implies that the nowcast will actually be a two-quarter ahead

[3] To handle the high-dimensionality of the problem, the group LASSO (Yuan & Lin, 2006) is used to estimate a Directed Acyclical Graph (DAG), and from that summarize the Granger causality statistics (Lozano et al., 2009; Shojaie & Michailidis, 2010). For each of the predictor variables (news and hard), the Granger causality test is run including three lags of all the predictors, and the amount of regularization is determined by the BIC information criteria. In the summary statistic in Figure 2, two-way predictive relationships, that is, when both the news- and hard-based variable Granger cause each other, are not counted.

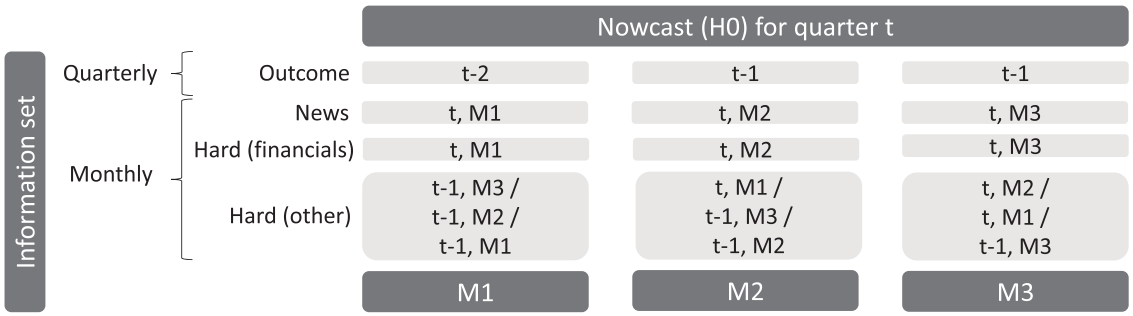


FIGURE 3 The information structure (quarter, month) for a generic quarter *t*. For the *Hard (other)* data category the monthly information structure depends on the release structure ("ragged-edges") in the FRED-MD database

prediction when using M1 data, but a one-quarter ahead prediction using M2 and M3 data. Figure 3 provides an illustration of the assumed information and timing structure. For each new vintage of data, the models are re-estimated using an expanding estimation window. Finally, although including lags of the dependent variable in (1) tend to improve forecasting accuracy, we refrain from this here to focus on the news- versus hard-based data dichotomy (but compare model performance to simple auto-regressive benchmarks later).

In all models we allow for $p = 6$ lags of each predictor, where the time lags are set relative to having a full quarter of monthly information. This ensures that our results across monthly vintages within a quarter reflect differences in available information, and not differences in lag structure, but also highlights the so called ragged-edge problem common to standard real-time forecasting experiments (Banbura et al., 2011). For example, due to lags in the release calendar, standing in M1 of any given quarter means that observations for M2 and M3 are missing for all predictors, and data for M1 (and even M3 in the previous quarter) for some of them. Here we address this by simply filling in the missing observations with the (real-time) mean of the predictors when making the predictions.[4]

Unless otherwise stated, all models are initially estimated using data from 1985Q1 to 1995Q4. The remaining data, 1996Q1 to 2020Q1, is used to recursively re-estimate the models and evaluate the OOS forecasting performance. All data transformations are done in real-time, that is, within each recursion and with the appropriate vintage of data, to avoid look-ahead biases. For the same reason, and because it is computational heavy to re-estimate, the LDA model used to classify the news and construct news topic time series is not updated after 1995Q4. Hence, all the news after 1995Q4 is classified OOS using the topic distributions learned from the 1985Q1 to 1995Q4 sample.[5]

We focus on point forecasting and use root-mean-square errors (RMSEs) to measure average performance over the whole sample and cumulative squared prediction error differences (CSPED) to highlight how forecasts perform relative to each other across time. In the main analysis all predictions are evaluated against the final vintage of data, that is, the release containing data for 2020Q1 and the first COVID-19 economic effects in the United States, but we discuss robustness to this choice in Section 4.5.

4 | RESULTS

The results are presented in five parts. In Section 4.1 we present our main predictive results, highlighting the differences in predictive performance between the *DJ* and *FRED-MD* datasets when evaluated ex post. Next, in Section 4.2, we take a more ex ante perspective and evaluate predictive performance when models and data are chosen in real-time without ex post knowledge of the best data and models. Section 4.3 provides a more in depth analysis of the predictor attributes and the narrative realism of the results, while Section 4.4 asks how good the predictions actually are by comparing predictive performance with the SPF. Finally, Section 4.5 shows how our results are robust along a number of dimensions related to modeling choices.

[4] While more sophisticated methods can be used, see, for example, Baffigi et al. (2004), Giannone et al. (2008), Kuzin et al. (2011), and Thorsrud (2018), this comes at the cost of increased computational complexity.

[5] Using news topics estimated on the whole sample, we show in Section 4.5 that the issue related to look-ahead biases likely is not very empirically important.

4.1 | The value of news

Figure 4 summarizes our main results. We highlight four points. First, the left column of the figure reports a scatter plot of the RMSE of each model, forecasting horizon and month, highlighting the overall performance of the news- relative to the hard-based approaches. The news-based scores are on the x axis. Thus, a scatter point above the 45 degree line indicates a better news- than hard-based performance for this model (at a given forecasting horizon and month). For *Consumption* the observations seem to be fairly evenly spread out. For *Investment* and *GDP* the majority of observations are below the 45° line. However, it is a general pattern across all outcome variables that the news-based predictions are more sensitive to method used to produce the predictions. That is, the variance in model performance is larger for the news-based predictions than it is for the hard-based predictions.

Second, focusing on the best (ex post) performing news- and hard-based models at each forecasting horizon and month, column two in Figure 4 reports the Diebold-Mariano test statistics. Here, the color shadings indicate 99%, 95% and 90% confidence bands, while the point estimates are illustrated with a black dot. The best performing news- and hard-based model names are reported to the left and right of the confidence boxes, respectively.

As seen in the figure, news is superior relative to hard economic data in terms of predicting *Consumption*, inferior in terms of predicting *Investment*, and on-par in terms of predicting overall *GDP*. That is, in seven out of nine cases the news-based predictions are the best performing predictors for *Consumption*. The news-based predictions are also significantly different (at the 90% confidence level) from their hard-based counterparts in most of these cases. In contrast, for *Investment* all of the best performing predictions are made using hard-based predictors. The results for overall *GDP* ends up somewhat in between these two extremes, although the best news-based predictions tend to have the lowest RMSE.

In terms of models, we also observe from column two in Figure 4 that the news-based predictors work best together with the RF method, which is the best performing news-based model in 85% of the cases (when looking across all variables, horizons, and months). For comparison, the RF method is the best performing model in less than 50% of the cases when using the hard-based data. Thus, allowing for potential non-linearities in the predictive relationships tend to add more value when using news-based predictors than when using the hard-based data. We explore this topic in greater detail in Section 4.3.

Third, looking at the CSPED plots, where the (ex post) best performing news- and hard-based models are compared over time, and only results for H0 and M1 are reported for visual clarity, one observes that the news-based predictions have a tendency to improve relative to the hard-based predictions during, and after, economic turmoil. For the *Consumption* and *GDP* predictions this is particularly evident around the Great Recession (GR) period, but also somewhat visible during the 2001 recession for *Consumption*.[6] Still, the good overall (relative) performance of the news-based predictions are not driven solely by recessions periods. For example, already in the time period prior to the GR, the news-based *Consumption* predictions had lower RMSE than the predictions based on hard economic data. For the *Investment* predictions, however, this picture is almost the opposite, showing that the hard-based predictions improved a lot upon the news-based predictions both well before and after the GR episode.

Fourth, zooming in on the nowcasting evaluation (H0), our results replicate the well known pattern documented in the earlier nowcasting literature (Aastveit et al., 2014; Banbura et al., 2011; Giannone et al., 2008), namely, that predictive performance improves as more hard-based information becomes available throughout the quarter. In particular, in the graph we compute the improvement in forecasting accuracy in M2 and M3 relative to M1 for the best performing (ex post) news- or hard-based models. For the *Consumption* nowcasts produced using the LASSO and hard-based data, for example, the improvement in RMSE from M1 to M3 is roughly 20%. For the news-based predictions this common finding does not hold, and we find very modest improvement in RMSE throughout the quarter. Together with the finding that the news-based (*Consumption*) predictions are relatively better at H1 and H2 than at H0, see column two in Figure 4, this suggest that the news-based dataset is more forward looking than the hard economic indicators, and thus performs better when either less hard economic information about the current quarter is available or at longer forecasting horizons.[7]

[6] Although our sample only contains the very beginning of the latest Covid-19 episode, we see signs of the same pattern also here, but only if using news topics estimated on a richer sample of data. See Section 4.5.

[7] The finding that news have better relative performance for longer forecasting horizons is also found by Ardia et al. (2019) when analyzing US industrial production.

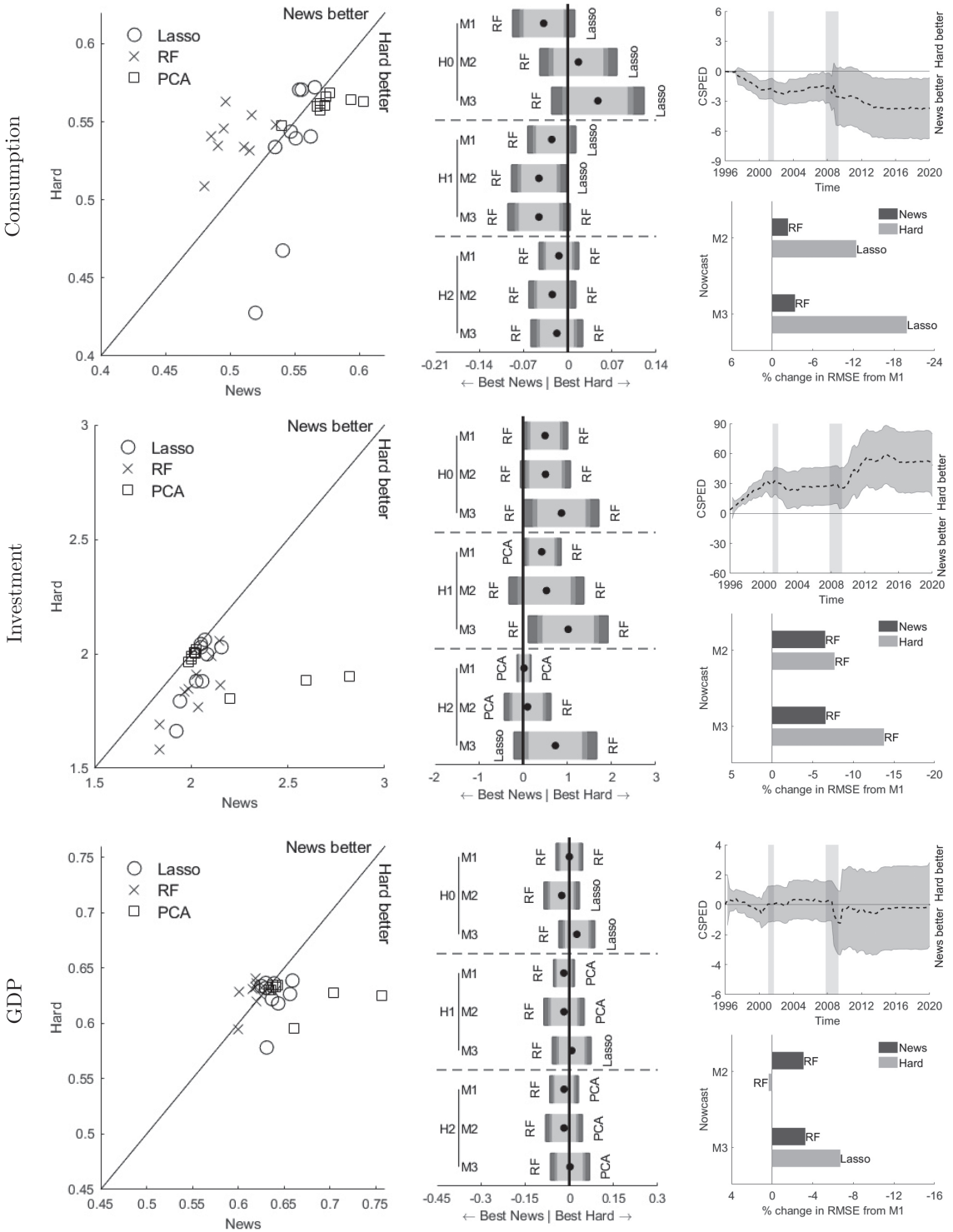


FIGURE 4 Root-mean-square errors (RMSEs), cumulative squared prediction error differences (CSPED) and nowcasting. The evaluation sample is 1996Q1-2020Q1. In columns two and three of the figure the best performing news- and hard-based models, evaluated ex post, are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold & Mariano, 1995). Color shadings illustrate 99%, 95%, and 90% confidence bands. In the CSPED graphs, an upward slope means that the hard economic data outperforms the news data, while the gray band is the equivalent of 90% two-sided levels, based on the Diebold–Mariano test statistic. The nowcast plots display the % improvement in RMSE throughout the quarter for the best performing models, evaluated ex post

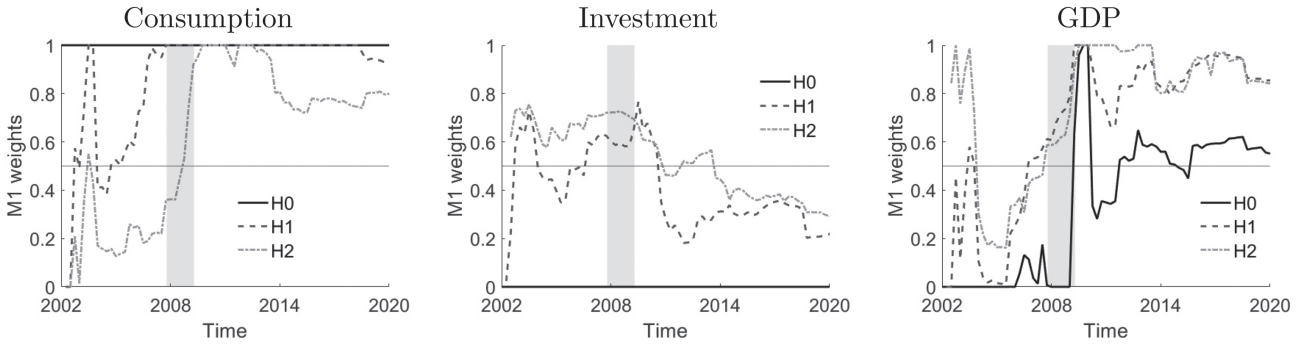


FIGURE 5 Optimal combination and weights. The evaluation sample is 2002Q1-2020Q1, and the weights attached to the news-based models are summed

4.2 | Variable and model combinations

In real-time forecasters do not have the benefit of knowing the ex post best dataset or model. To mimic a more realistic forecasting process, and to ensure that the results from the previous section are not driven by ex post selection, we apply a recursive OOS variable and forecast combination scheme.

In terms of variable combination, the *DJ* and *FRED-MD* datasets are merged into one big panel. Then, the OOS experiment is re-estimated using the same methods as before, but now only using the combined dataset. Going forward, these grand models (GM) are denoted GM-LASSO, GM-RF, and GM-PCA.

In terms of forecast combination, we follow a large point forecast combination literature (see Timmermann, 2006 for an overview) and consider simple linear combinations of the six individual forecasts analyzed in the previous section, that is, the news- and hard-based LASSO, RF, and PCA predictions. More formally, standing at a given forecasting origin t, a combined prediction is constructed as

$$\hat{y}_{t+h} = \sum_{i=1}^{N} w^o_{ith} \hat{y}_{i,t+h}, \tag{2}$$

where $\hat{y}_{i,t+h}$ is the predictions from one of the $N = 6$ ensemble members, and w^o_{ith} is a horizon specific model weight. The weights used here are optimal in the sense that they solve

$$\boldsymbol{w}^o_{th} = argmin_w \sum_{r=1}^{t-h} (y_{r+h} - \boldsymbol{w}\hat{y}_{r+h})^2, \tag{3}$$

which is estimated using OLS under the restriction that the weights are positive and sum to unity.[8]

Twenty-four observations are used to estimate the initial weights. OOS predictions are recursively constructed and updated using an expanding estimation window. Accordingly, both the variable and forecast combination schemes are evaluated over the sample 2002Q1 to 2020Q1.

The two first columns in Figure SD1 report the same type of statistics as in Figure 4, but now comparing the optimal combination to the (best) hard-based models. The qualitative conclusions strengthen those from the ex post OOS analysis in the previous section. That is, the *DJ* dataset contains complementary information to that in the *FRED-MD* dataset when predicting *Consumption* in particular. For *Investment*, the combined predictions have lower RMSE than many of the individual models based on hard data (column one in Figure SD1), but the best performing models still tend to be hard-based only (column two in Figure SD1). As seen from Figure SD2, a similar conclusion is obtained when evaluating the GMs. Accordingly, combining forecasts or combining variables is not an important issue in the experiment conducted here. On the margin, however, the optimal combination scheme performs slightly better in terms of RMSE than the variable combination approach.

To further highlight the news- versus hard-based predictor dichotomy, Figure 5 illustrates how the optimal weights attached to the news-based models vary through time. In the interest of readability and preserving space, the weights

[8] Formally, the weights are optimal in population only to the extent that the joint distribution of outcomes and predictions is Gaussian. Apart from simplifying the interpretation, the restrictions rule out that the combined forecast lies outside the range of the individual forecasts and reduces serial correlation in the combined forecast errors (Timmermann, 2006).

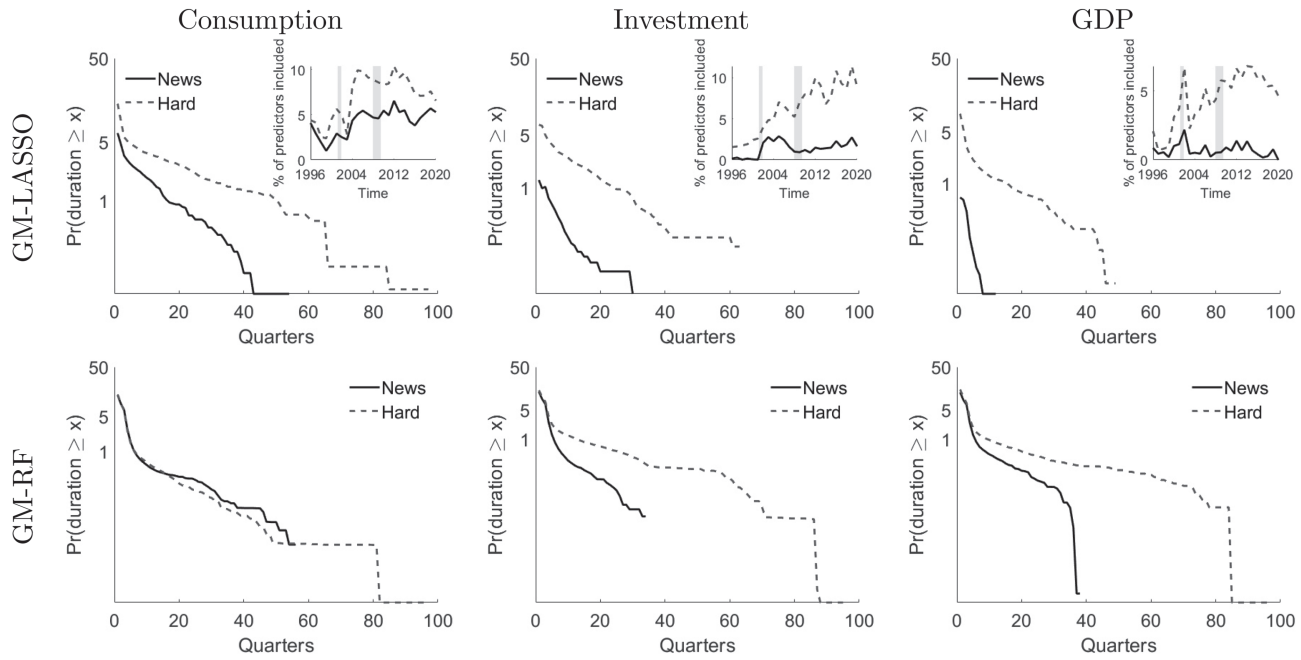


FIGURE 6 Dynamic sparsity and predictor importance. The first row displays the average duration of the predictors and the sparsity (aggregated into average yearly observations) implied by the grand models Least Absolute Shrinkage and Selection Operator (GM-LASSO). The duration is computed as the probability that a predictor is used by the LASSO when making forecasts in more than x consecutive quarters. The second row shows the average duration of the predictors using the GM-Random Forest (RF). This is computed as the probability that a predictor stays in the same decile in terms of ranking by predictor importance in more than x consecutive quarters. In all graphs the mean across forecasting horizons and months are reported

attached to the news-based models are summed and only results for M1 are reported. Apart from some volatility in the beginning, when relatively few observations are available for constructing the weights, the news-based predictions get a substantial weight in terms of predicting *Consumption* and *GDP*. For example, standing in month one of any given quarter, the weight attached to the news-based predictions is above 70% and 50% for a bigger part of the sample irrespective of the forecasting horizon. Moreover, even for *Investment* the news-based predictions receive a weight above 20% for H1 and H2 when standing in M1. Thus, although the hard-based *Investment* predictions were superior in the ex post analysis in the previous section, news adds value in the more realistic forecast combination scheme conducted here.

4.3 | Predictor attributes and narrative realism

There are noticeable differences between the news- and hard-based predictor attributes and how they operate within the individual models. This is illustrated in Figure 6, where recursively estimated in-sample statistics from the GM-RF and GM-LASSO models are reported. For the GM-RF model the importance of each predictor is calculated at each forecasting vintage in the sample.[9] The plot shows the probability that a predictor stays in the same decile in terms of ranking by predictor importance in more than x consecutive quarters. For the GM-LASSO model the degree of sparsity at each forecasting vintage is computed, that is, the fraction of predictors selected, in addition to how likely it is that a predictor is selected for more than x consecutive quarters once it has first been selected as a predictor. All statistics are aggregated across forecasting horizons and months.[10]

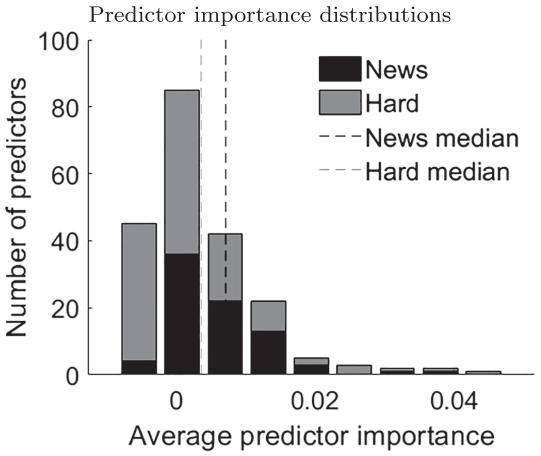
The big picture is clear: The news-based predictors are more short-lived and sparse relative to the hard-based predictors. Using the GM-LASSO, for example, there is roughly 1% probability that a news-based *Consumption* predictor will be in the

[9] For a given predictor, the predictor importance measures the increase in prediction error when the values of that predictor are permuted across the out-of-bag observations. The measure is computed for all the individual trees and then averaged over the entire ensemble and normalized by the standard deviation of the whole ensemble of trees.

[10] We have confirmed that the same qualitative conclusions also hold when looking at each forecasting horizon and month separately. These additional results can be obtained on request.

Consumption and top 10 predictors

Rnk	Type	Group	Name
1	Hard	Housing	New Private Housing Permits, Midwest
2	News	Personal finance	Topic0 Personal finance
3	Hard	Retail and cons.	Real personal consumption expenditures
4	News	Bond market	Topic49 Bond market
5	Hard	Housing	Housing Starts, Midwest
6	Hard	Labor market	Civilians Unemployed - 15 Weeks & Over
7	Hard	Labor market	All Employees: Construction
8	Hard	Labor market	All Employees: Financial Activities
9	News	Politics	Topic39 Negotiation
10	News	Health	Topic19 Health care



Tree map of relative news-based importance

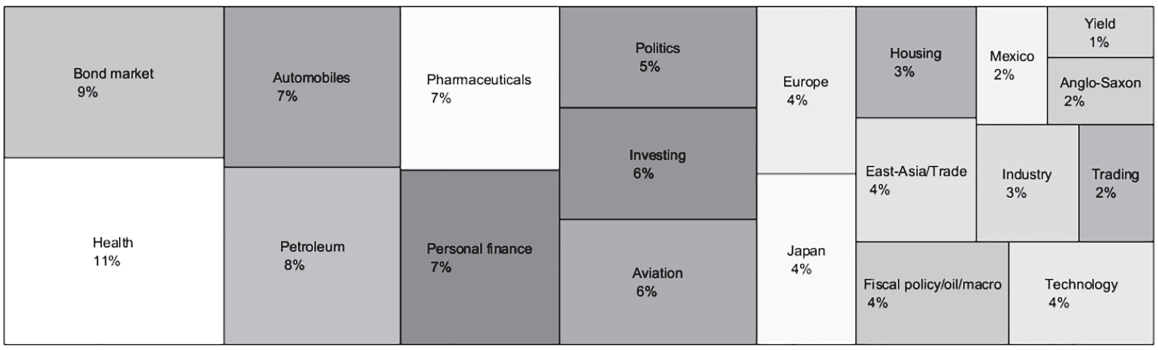


FIGURE 7 Grand models Random Forest (GM-RF) and predictor importance for *Consumption*. The table reports the top 10 most important predictors on average across the sample, while the histogram reports the empirical distribution of the average predictor importance statistics for the news- and hard-based datasets as a whole. In the tree map figure the news-based predictors are categorized into 20 groups using a hierarchical agglomerative clustering algorithm (see Section 2.1 and Figure SA1). The graph then illustrates the average importance of predictors within each group, where the size of the rectangles represent the group's relative weight

selected variable set for up to 15 consecutive quarters, while the comparable probability for the hard-based predictors is more than three times as large. Likewise, the degree of sparsity is high, particularly for the news-based predictors, where only roughly 5% of them are selected on average. Qualitatively, the same conclusions hold for *Investment* and *GDP*, and when looking at the GM-RF duration and predictor importance statistics. The only exception is for *Consumption* and the GM-RF statistic, where the news- and hard-based data behave similarly, although some of the hard-based predictors have longer duration.

While there might be many explanations for these patters, one reason might be that the news media foremost report on newsworthy events and stories. Thus, the news-topic time series becomes more like economic shock series with substantial spikes at specific time periods, as also illustrated in Figure 1 and discussed in Section 2.3. Relatedly, and as pointed out by Larsen and Thorsrud (2018), the particular topic composition of a given story at a given point in time, might very well be unique, but the topics that the narrative constitute are potentially shared by many other stories at different time periods and with different weighting. Thus, how the topics operate together to form narratives change and evolve over time to a much larger extent than it does for hard economic variables. Or, in other words, industrial production measures industrial production regardless of time, whereas a topic's contribution to time dependent narratives is time dependent. This makes it natural that the news-based data is more short-lived and sparse relative to the hard-based predictors.

Figure 7 reports the most influential predictors when using the GM-RF model for *Consumption*predictions. Again, to focus on the overall picture, only averages across time, forecasting horizons, and months are reported, while results for *Investment* and *GDP* are reported in Figure SD3.

Among the most influential hard-based variables are series related to housing, consumption expenditures, and employment conditions. Still, news topic time series related to personal finance, the bond market, and health are all among the 10 most influential series. From a *Consumption* prediction perspective this makes narrative sense. Health care, for example, is not only an important component of most Americans' expenses, but has also been shown to be particularly important

in households expectations formation process (Larsen et al., 2021). Moreover, in line with the sparsity statistics discussed above, the upper right histogram in Figure 7 shows that the predictor importance statistic is skewed to the right for both types of data, but more so for the news-based predictors than the hard-based ones.

However, while *Personal finance* and *Bond market* are the most important news topics for *Consumption*, the tree map in the lower row in Figure 7 shows that news topics related to health, petroleum, and automobiles are (roughly) equally important as a group. In particular, using the hierarchical agglomerative clustering algorithm discussed in Section 2.1, and illustrated in Figure SA1, to group the individual topics into higher order abstractions highlights that many news topic groups are relatively important for describing *Consumption*. At the same time, the figure also shows that some groups are relatively unimportant. For example, news narratives related to *Mexico*, *Anglo-Saxon*, and *Yield* receive a small weight in the US consumption context.

4.4 | How good are the predictions? A soft-based comparison

The set of models used in the preceding sections are commonly used when working with high-dimensional data. Still, more accurate predictions could potentially be constructed using more tailored modeling approaches. Despite this, it is of practical interest to evaluate how good the predictions actually are. To do so we continue to focus on the data dimension, and compare predictions from the best performing news- and hard-based models to those from simple auto-regressive and constant growth rate benchmarks as well as predictions made by the SPF.

TABLE 1 Relative RMSE scores

		H0			H1			H2		
		M1	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
Consumption	AR	0.90***	0.90**	0.87**	0.90**	0.91***	0.87***	0.96**	0.90***	0.89**
	RW	0.82*	0.86***	0.47	0.82	0.87***	0.47	0.89*	0.84*	0.49
	SPF	1.23*	1.19	1.17	1.16	1.13	1.10	1.16**	1.11	1.09
Investment	AR	0.93	0.92	0.87	0.93	0.97	0.93	1.00	1.02	1.03
	RW	0.75	0.77	0.68*	0.75	0.81	0.72	0.73*	0.79	0.73
	SPF	1.53***	1.42***	1.42***	1.46***	1.43***	1.44***	1.39***	1.38**	1.38**
GDP	AR	0.95	0.95*	0.91	0.94**	0.97*	0.94	0.96	0.96	0.97
	RW	0.78**	0.84**	0.60	0.77**	0.86**	0.62	0.77**	0.84*	0.61
	SPF	1.36**	1.31*	1.32**	1.25	1.25*	1.26*	1.12	1.11	1.13

Note: The best news-based models are compared to an auto-regressive model (AR), a constant growth rate model (RW), and the Survey of Professional Forecasters (SPF). The lag order in the AR is chosen (in real-time) using the BIC. The evaluation sample is 1996Q1-2020Q1. A value less than 1 indicates that the best news-based model has the lowest root-mean-square error (RMSE). Significant differences in forecasting performance are calculated using the Diebold-Mariano test (Diebold & Mariano, 1995).

* 10% significance level.

** 5% significance level.

*** 1% significance level.

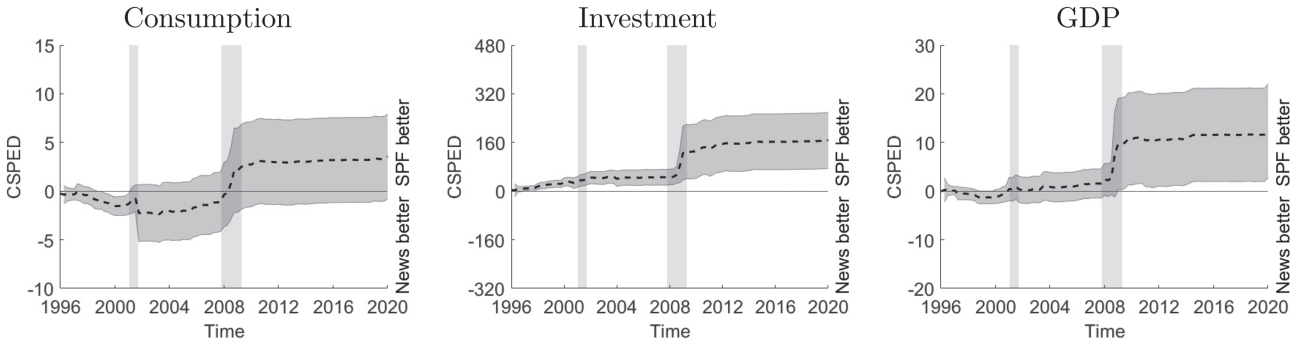


FIGURE 8 Survey of Professional Forecasters (SPF) and ex post best news-based forecasts. The graphs compare nowcasting performance (H0), and to align informational available at the time of prediction between the SPF and the model-based forecasts, predictions produced in M2 are used. An upward slope means that the SPF outperforms the news data, while the gray band is the equivalent of 90% two-sided levels, based on the Diebold-Mariano test statistic (Diebold & Mariano, 1995)

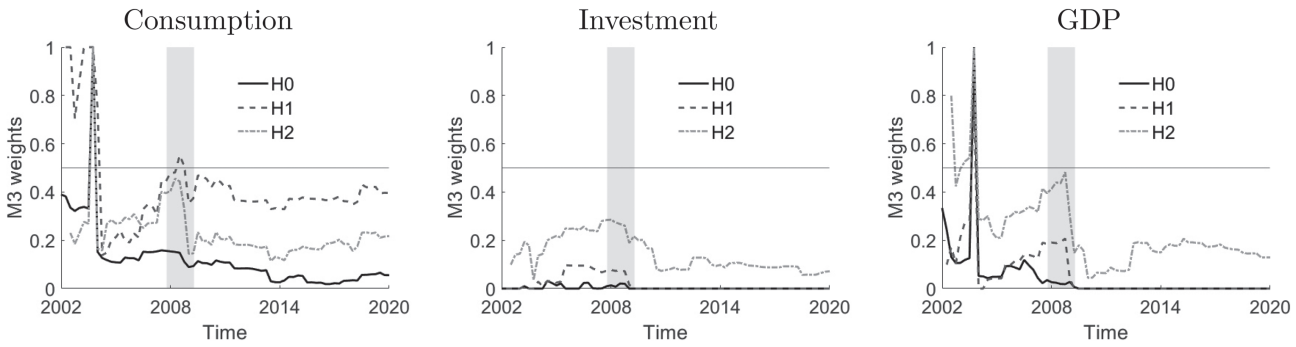


FIGURE 9 Optimal combination and weights. All the news- and hard-based forecasts as well as the Survey of Professional Forecasters (SPF) forecasts are included in the regression. The evaluation sample is 2002Q1-2020Q1, and the weights attached to the news-based models are summed

The SPF is the oldest quarterly survey of macroeconomic forecasts in the United States and is currently conducted by the Federal Reserve Bank of Philadelphia. According to Stark and Croushore (2019) it " … has become the gold standard for evaluating forecasts or comparing forecasting models", and Pesaran and Weale (2006) provide an overview of the usage of this type of soft data for capturing expectations and forecasting. Here we use the mean forecasts from the survey, and transform them to quarterly (log) percentage growth rates.

As seen from Table 1, the (ex post) best news-based *Consumption* predictions outperform the simple model-based benchmarks. Except for in a few cases, the differences in predictive performance are also statistically significant. The news-based *Investment* and *GDP* predictions tend to have a lower RMSE than the benchmark models, but these differences are less significant. In contrast, the SPF predictions have a lower RMSE than the news-based ones across both forecasting horizons, months, and variables. However, for *Consumption*, the differences in performance between the SPF and news-based approach are not significant. In fact, as illustrated in Figure 8, which reports the CSPED between the SPF forecasts and the best news-based forecasts, using H0 and M2, the better SPF score is almost entirely due to the GR period which naturally favors subjective and adaptive predictions over model-based predictions capturing averages over a longer timespan.[11]

Since accessing and using new and alternative data sources can be costly, the above results might suggest that using the news-based predictions might be less attractive from a cost benefit perspective. After all, the publicly available survey data gives very good predictions. Still, to take into account the more realistic setting where practitioners might utilize both the survey data and the news- and hard-based predictions in real time, Figure 9 reports the recursively estimated weights attached to the new-based models when performing the same type of combination scheme as described in Section 4.2. In particular, we allow all three news- and model-based predictions to enter the forecast combination scheme together with the three hard-based ones as well as the SPF predictions. Thus, in total, we give weight to seven different predictions. In line with the results reported above, the news-based predictions get very little weight for *Investment* and *GDP*. In contrast, the news-based predictions get between 10% and 40% of the weight on average (depending on the forecasting horizon) for *Consumption*, and only for H2 does the Great Recession seem to strongly tilt the weights away from news. As such, the news-based predictions seem to contain supplementary information even to the soft SPF data.

4.5 | Robustness and additional results

Our main conclusions are robust along a number of dimensions. To better capture potential structural changes in the data, and their joint distribution, across time, we have experimented with using a rolling window when estimating the individual models and doing the OOS analysis. The main conclusions regarding the news versus hard predictor dichotomy continue to hold when doing so, but the average absolute performance becomes slightly worse (Figure SD4). One reason for this is likely that the best performing individual models benefit from having longer time-spans of data available for estimation rather than shorter windows. Moreover, experimenting with a richer lag structure, allowing for up to 12 monthly time lags, in the underlying MIDAS model in (1) does not change our main qualitative conclusions. That is, the

[11] Results comparing the best hard-based predictions to the simple model-based benchmarks and the SPF are reported in Table SD1. The overall pattern is very much similar to that described above.

cross-validation techniques used when estimating the different models automatically picks up the relevant lag structure, which then is, or falls below, six as in our benchmark specification.

In terms of producing combined predictions, simple equal and inverse-MSE weights are often used and perform well in empirical settings (Timmermann, 2006). Here, the optimal combination scheme outperforms the two simpler alternatives in terms of *Consumption* predictions, and to some extent also in terms of *Investment*predictions. For *GDP*, the three combination schemes perform very much the same (Table SD2). These results are well in line with those presented in Figure 5, where the optimal weights varied substantially across the sample and were far from equal for *Consumption* and *Investment*, but closer to equal for *GDP*.

Because of data revisions in quarterly National Account Statistics, a key issue in OOS experiments is the choice of the "actual" outcome variable and vintage. Stark and Croushore (2002) discuss three alternatives: the most recent data vintage, the last vintage before a structural revision, and finally the estimate released a fixed period of time after the first release. In the main analysis we have used the first of these three alternatives. As a robustness check we show in Figures SD5 and SD6 that the main conclusions in terms of the news- versus hard-based datasets hold when evaluating the predictions against both the first and second release of the data. Still, there are clear patterns in the results showing that the news-based predictions are relatively better at predicting the final release of the outcome data rather than the preliminary ones.

Results presented in Thorsrud (2018) highlight how adjusting the topic times series with the positive or negative tone of news reporting increases their correlation with the (Norwegian) business cycle. In the main analysis, we have not worked with tone adjusted topic time series. However, following the same dictionary-based adjustment procedure as described in Thorsrud (2018) the news-based predictive performance actually becomes worse for *Consumption* when considering only the tone of reporting, or the tone interacted with the topic frequencies, while the results for *GDP* and *Investment* remain relatively unaffected (Table SD3).[12] One potential reason for this, as also noted by Thorsrud (2018), is that the tone-adjustment procedure is very simplistic and dependent on the exact dictionary used to define positive and negative words. We leave it to future research to investigate whether predictive performance could be improved using more sophisticated and robust methods to extract sentiment (see e.g., Shapiro et al., 2020 and Ardia et al., 2019).

In terms of the number of news topics to extract, using 80 topics was motivated by two factors. First, this was the choice showing the best statistical results in Larsen and Thorsrud (2019) and Thorsrud (2018) (on a similar corpus). Second, it is our experience that with a substantially higher number of topics, each topic starts to become highly event specific, that is, there are signs of over-fitting the corpus. Conversely, extracting substantially fewer topics results in too general topics making narrative interpretation more difficult. Here, re-doing the OOS analysis using either 40 or 120 estimated news topics does not alter our main qualitative conclusions regarding news- versus hard-based data. However, in line with the conjectures made above, the 80 topic case seems to perform marginally better than using either 40 or 120 estimated news topics (Table SD4).

Finally, to avoid potential look-ahead biases, the LDA model was estimated on the initial training sample only, that is, all the news after 1995Q4 is classified OOS using the topic distributions learned from the 1985Q1 to 1995Q4 sample. This is of course a very restrictive usage of the LDA. Still, when estimating the LDA using the full sample of news data, and then simply truncating the resulting news topics time series for the OOS forecasting experiment, we observe that the results from Section 4.1 remain qualitative robust (Figure SD7). As also commented on in Larsen and Thorsrud (2019), this suggests that the issue related to look-ahead biases likely is not very empirically important.

Still, since we do not observe large improvements for the news-based models when using topics estimated on the full sample, it also suggests that using topics learned using the whole sample might give just as representative/unrepresentative topics for the OOS evaluation sample as those learned using only the initial training sample. Indeed, informal evidence obtained when trying to label and categorize the topics estimated on the full sample suggest that this might be the case, where either the word clouds themselves or the categorization of topics into broader clusters becomes less intuitive. Figure SD8 documents this more formally by showing a CSPED plot comparing the news-based results using topics estimated either on the training sample or using the whole sample. As seen in the figure, using the former outperforms the latter in the beginning of the evaluation sample, while the reverse happens toward the latter half of the evaluation sample. In particular, the large relative improvements during the Great Recession and the early Covid-19 episode stand out.

[12] In short, for each day and topic, the article that is best explained by each topic is identified and its tone computed, that is, whether the news is more positive than negative. This is done using an external word list, the Harvard IV-4 Psychological Dictionary, and simple word count differences. Then, the topic frequencies are simply multiplied by their respective tone.

We conjecture that further improvements in the news-based predictions can be obtained in future research taking aboard the computational cost of re-estimating the topic model for each new vintage of data, potentially allowing for a rolling estimation window, or using topic models featuring dynamic properties (Blei & Lafferty, 2006).

5 | CONCLUSION

Decades of research have investigated how hard economic data best can be used for macroeconomic forecasting, that is, which datasets and variables are informative, which models work, etc. Much less is known about the value of alternative data sources, such as news and text.

This article contributes to a fast growing economic literature using text as data for economic analysis and forecasting. In particular, entertaining a unique dataset of 22.5 million news articles from the *Dow Jones Newswires Archive*, we perform an in depth out-of-sample forecasting comparison study with what has become the "industry standard" in the newer forecasting literature, namely the *FRED-MD*dataset.

Prior to estimation, the unstructured and high-dimensional textual data is transformed into time series objects using an unsupervised topic model which is both widely used, simple and transparent, and delivers interpretable outputs. Next, real time and truly out-of-sample predictions are formed using off the shelf, but state-of-the-art, Machine Learning and econometric forecasting techniques.

Our evaluation, focusing on predicting US GDP, consumption and investment growth, strongly suggest that the news data contains information not captured by the hard economic indicators, and that news is particularly informative for forecasting consumption developments. There are also clear patterns in the results suggesting that news data performs relatively better for one- and two-quarter ahead predictions than for nowcasting, and that the news-based predictions tend to improve upon the predictions made using hard economic indicators in times of economic turmoil, such as during and after the Great Recession. Finally, we document that the narrative realism of the news-based approach is good, and that the news-based predictors are more short-lived and sparse relative to the hard-based predictors.

These results are all new in the literature and establish several "stylized facts" about the value of hard-based relative to news-based data for macroeconomic forecasting. From a practitioners perspective accessing and using new and alternative data sources can be costly. Whether the potential gains in forecasting performance documented here outweigh the costs will depend on the loss function in each particular case. Using other soft data, in form of the SPF, we find little evidence suggesting that the news-based predictions constructed here are superior. Still, our results indicate that the news topics contain supplementary information also to this type of data. Moreover, the news-based data is available at a much higher frequency which might be beneficial around economic turning points. There are also many avenues for future research in terms of how textual data can be decomposed into useful time series objects, and how to model these types of data relative to conventional economic time series. As such, the horse-race has just begun.

ACKNOWLEDGEMENTS

This article should not be reported as representing the views of Norges Bank. The views expressed are those of the authors and do not necessarily reflect those of Norges Bank. We thank the editor and two anonymous referees, Hilde C. Bjørnland, Martin Blomhoff Holm, and Felix Kapfhammer for valuable comments. Comments from conference participants at BI Norwegian Business School and SNDE 2020 also helped improve the paper. This work is part of the research activities at the Centre for Applied Macroeconomics and Commodity Prices (CAMP) at the BI Norwegian Business School. We are grateful to the *Dow Jones Newswires Archive* for sharing their data with us for this research project.

ORCID

Jon Ellingsen https://orcid.org/0000-0002-1840-5458
Vegard H. Larsen https://orcid.org/0000-0002-0419-3028
Leif Anders Thorsrud https://orcid.org/0000-0002-5115-4806

REFERENCES

Aastveit, K. A., Gerdrup, K. R., Jore, A. S., & Thorsrud, L. A. (2014). Nowcasting GDP in real time: A density combination approach. *Journal of Business & Economic Statistics*, *32*(1), 48–68. https://doi.org/10.1080/07350015.2013.844155

Ardia, D., Bluteau, K., & Boudt, K. (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting*, *35*(4), 1370–1386. https://doi.org/10.1016/j.ijforecast.2018.10.010

Babii, A., Ghysels, E. & Striaukas, J. (2020). Machine learning time series regressions with an application to nowcasting. arXiv preprint arXiv:2005.14057.

Baffigi, A., Golinelli, R., & Parigi, G. (2004). Bridge models to forecast the Euro area GDP. *International Journal of Forecasting*, *20*(3), 447–460. https://doi.org/10.1016/S0169-2070(03)00067-0

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221. https://doi.org/10.1111/1468-0262.00273

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. https://doi.org/10.1093/qje/qjw024

Banbura, M., Giannone, D., & Reichlin, L. (2011). Nowcasting. In *The Oxford Handbook of Economic Forecasting (Oxford Handbooks in Economics)*. New York: Oxford University Press.

Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks: Handbook. *Centre for Central Banking Studies*, *33*, 1–19.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*, 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 113–120). New York, NY, USA: ICML'06.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bybee, L., Kelly, B. T., Manela, A. & Xiu, D. (2019). The structure of economic news. Available at SSRN 3446225.

Carroll, C. D. (2003). Macroeconomic Expectations of Households and Professional Forecasters. *The Quarterly Journal of Economics*, *118*(1), 269–298. https://doi.org/10.1162/00335530360535207

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22* (pp. 288–296). Cambridge, MA: The MIT Press.

Clements, M. P., & Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data. *Journal of Business & Economic Statistics*, *26*(4), 546–554. https://doi.org/10.1198/073500108000000015

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, *105*(1), 111–130.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Dybowski, T., & Adämmer, P. (2018). The economic effects of u.s. presidential tax communication: Evidence from a correlated topic model. *European Journal of Political Economy*, *55*, 511–525. https://doi.org/10.1016/j.ejpoleco.2018.05.001

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (midas): Midas regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 57–82. https://doi.org/10.1111/rssa.12043

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Ghysels, E., Santa-Clara, P. & Valkanov, R. (2004). The midas touch: Mixed data sampling regression models.

Giannone, D., Reichlin, L., & Simonelli, S. (2009). Nowcasting euro area economic activity in real time: The role of confidence indicators. *National Institute Economic Review*, *210*, 90–97. https://doi.org/10.1177/0027950109354413

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676. https://doi.org/10.1016/j.jmoneco.2008.05.010

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*(S1), 114–133.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801–870. https://doi.org/10.1093/qje/qjx045

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G. & Kapadia, S. (2020). Making Text Count: Economic Forecasting Using Newspaper Text. Bank of England working papers 865, Bank of England. https://doi.org/10.2139/ssrn.3610770.

Kuzin, V., Marcellino, M., & Schumacher, C. (2011). Midas vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting*, *27*(2), 529–542. https://doi.org/10.1016/j.ijforecast.2010.02.006

Larsen, V. H. (2021). COMPONENTS OF UNCERTAINTY. *International Economic Review*, *62*(2), 769–788. http://doi.org/10.1111/iere.12499

Larsen, V. H., & Thorsrud, L. A. (2017). Asset returns, news topics, and media effects. Working Paper 2017/17, Norges Bank. https://doi.org/10.2139/ssrn.3057950.

Larsen, V. H., & Thorsrud, L. A. (2018). Business Cycle Narratives. Working Paper 2018/03, Norges Bank. https://doi.org/10.2139/ssrn.3130108.

Larsen, V. H., & Thorsrud, L. A. (2019). The value of news for economic developments. *Journal of Econometrics*, *210*(1), 203–218. https://doi.org/10.1016/j.jeconom.2018.11.013

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2021). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics*, *117*, 507–520. https://doi.org/10.1016/j.jmoneco.2020.03.004

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lozano, A. C., Abe, N., Liu, Y., & Rosset, S. (2009). Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, *25*(12), i110–i118. https://doi.org/10.1093/bioinformatics/btp199

McCracken, M. W., & Ng, S. (2016). Fred-md: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics*, *34*(4), 574–589. https://doi.org/10.1080/07350015.2015.1086655

Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*, 1–22.

Nimark, K. P., & Pitschner, S. (2019). News media and delegated information choice. *Journal of Economic Theory*, *181*, 160–196. https://doi.org/10.1016/j.jet.2019.02.001

Pesaran, M. H., & Weale, M. (2006). Chapter 14 survey expectations. In *Volume 1 of Handbook of Economic Forecasting* (pp. 715–776). Elsevier.

Shapiro, A. H., Sudhof, M. & Wilson, D. J. (2020). Measuring news sentiment. Journal of Econometrics. https://doi.org/10.1016/j.jeconom.2020.07.053

Shojaie, A., & Michailidis, G. (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, *26*(18), i517–i523. https://doi.org/10.1093/bioinformatics/btq377

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. https://doi.org/10.1016/S0304-3932(03)00029-1

Stark, T., & Croushore, D. (2002, December). Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics*, *24*(4), 507–531. https://doi.org/10.1016/S0164-0704(02)00062-9

Stark, T. & Croushore, D. (2019). Fifty years of the survey of professional forecasters. Federal Reserve Bank of Philadelphia Economic Insights 2019Q4, FRB Philadelphia.

Stock, J. H., & Watson, M. W. (1989). New indexes of coincident and leading economic indicators. In O. J. Blanchard & F. Stanley (Eds.), *NBER Macroeconomics Annual*, NBER Chapters (pp. 351–394). Cambridge, MA: The MIT Press.

Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162. https://doi.org/10.1198/073500102317351921

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature*, *41*(3), 788–829. https://doi.org/10.1257/jel.41.3.788

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*, 1–17.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B: Methodological*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Timmermann, A. (2006). Forecast combinations. In *Volume 1 of Handbook of Economic Forecasting* (pp. 135–196). Amsterdam, North Holland: Elsevier.

Ulbricht, D., Kholodilin, K. A., & Thomas, T. (2017). Do media data help to predict german industrial production? *Journal of Forecasting*, *36*(5), 483–496. https://doi.org/10.1002/for.2449

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3

Wu, H. D., McCracken, M. W., & Saito, S. (2004). Economic communication in the'lost decade': News coverage and the japanese recession. *Gazette (Leiden, Netherlands)*, *66*(2), 133–149. https://doi.org/10.1177/0016549204041474

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ellingsen, J., Larsen, V. H., & Thorsrud L. A. (2021) News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*, 1–19. https://doi.org/10.1002/jae.2859