Ellingsen, Jon; Larsen, Vegard Høghaug; Thorsrud, Leif Anders

Working Paper

News media vs. FRED-MD for macroeconomic forecasting

Working Paper, No. 14/2020

Provided in Cooperation with:
Norges Bank, Oslo

Suggested Citation: Ellingsen, Jon; Larsen, Vegard Høghaug; Thorsrud, Leif Anders (2020) : News media vs. FRED-MD for macroeconomic forecasting, Working Paper, No. 14/2020, ISBN 978-82-8379-168-6, Norges Bank, Oslo, https://hdl.handle.net/11250/2690107

This Version is available at:
https://hdl.handle.net/10419/246117

WORKING PAPER

News media vs. FRED-MD for macroeconomic forecasting

NORGES BANK RESEARCH

14 | 2020

JON ELLINGSEN
VEGARD H. LARSEN
LEIF ANDERS THORSRUD

NORGES BANK

Working papers fra Norges Bank, fra 1992/1 til 2009/2 kan bestilles over e-post:
FacilityServices@norges-bank.no

Fra 1999 og senere er publikasjonene tilgjengelige på www.norges-bank.no

Working papers inneholder forskningsarbeider og utredninger som vanligvis ikke har fått sin endelige form. Hensikten er blant annet at forfatteren kan motta kommentarer fra kolleger og andre interesserte. Synspunkter og konklusjoner i arbeidene står for forfatternes regning.

Working papers from Norges Bank, from 1992/1 to 2009/2 can be ordered by e-mail
FacilityServices@norges-bank.no

Working papers from 1999 onwards are available on www.norges-bank.no

Norges Bank's working papers present research projects and reports (not usually in their final form) and are intended inter alia to enable the author to benefit from the comments of colleagues and other interested parties. Views and conclusions expressed in working papers are the responsibility of the authors alone.

ISSN 1502-8190 (online)
ISBN 978-82-8379-168-6 (online)

News media vs. FRED-MD for macroeconomic forecasting*

Jon Ellingsen† Vegard H. Larsen‡ Leif Anders Thorsrud§

October 8, 2020

Abstract

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 U.S. 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.

JEL-codes: C53, C55, E27, E37

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

*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 one anonymous referee, 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.

†Norges Bank and Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School. Email: Jon.Ellingsen@bi.no

‡Norges Bank and Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School. Email: Vegard-Hoghaug.Larsen@norges-bank.no

§Centre for Applied Macroeconomics and Commodity Prices, BI Norwegian Business School. Email: leif.a.thorsrud@bi.no

1 Introduction

During the last decades advances in econometric techniques have substantially improved short-term forecasting performance in economics (Stock and Watson (2002), Ghysels et al. (2004), Giannone et al. (2008), Aastveit et al. (2014)). 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 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 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 and 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 and 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. (2020)). 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, e.g., politics and uncertainty (Baker et al. (2016)), making it a useful supplement for capturing the complexity of expectations (Sims (2003)).[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

[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)).

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, e.g., business cycle and monetary policy analysis (Larsen and Thorsrud (2019), Thorsrud (2018), Hansen et al. (2018), Hansen and McMahon (2016)).[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 U.S. 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 (Ghysels et al. (2004), Foroni et al. (2015)) 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 and 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

[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 U.S.

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. (2020)), 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 towards 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, 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 U.S., and the U.K., 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.

By using ML techniques to form predictions our analysis also relates to recent research

by Medeiros et al. (2019). Whereas they 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.

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, e.g., Carroll (2003), Nimark and Pitschner (2019), and Larsen et al. (2020), 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. (2020) 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, i.e., 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 U.S. 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 B.

The cleaned text corpus is decomposed into news topics using a Latent Dirichlet Allocation (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 (Hansen and McMahon (2016), Larsen (2017), Larsen and Thorsrud (2017), Hansen et al. (2018), and Dybowski and Adämmer (2018)).

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 B 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, e.g., 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 2.1 illustrates the output from the above steps for six of the 80 topics. A full list of the estimated topics is given in Table A.2, in Appendix A. The LDA topic distributions are illustrated using word clouds. A bigger font illustrates

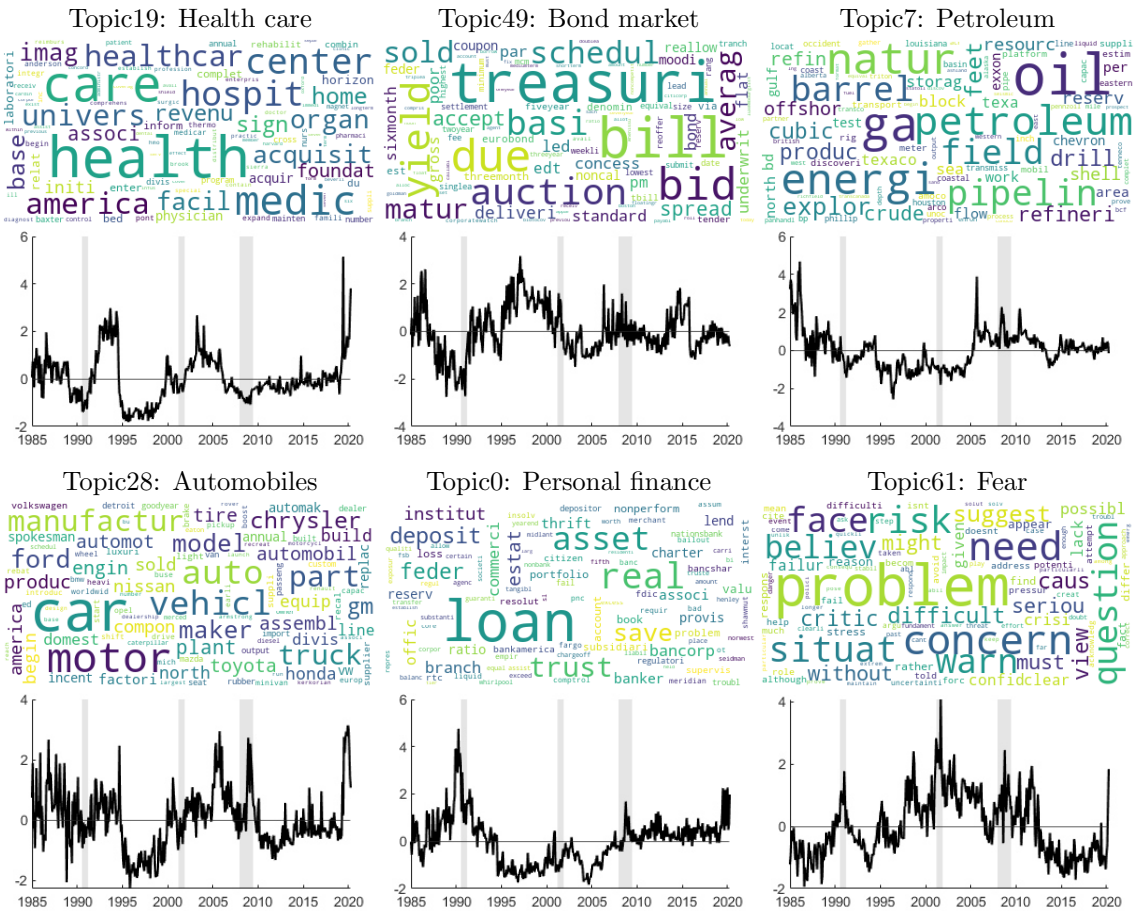


Figure 2.1. Topic distributions and time series. For each topic, the size of a word in the word cloud reflects the probability of this word occurring in the topic. Each word cloud only contains a subset of all the most important words in the topic distribution. Topic labels are subjectively given. The topic time series are linearly detrended and normalized. January 1985 - April 2020.

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 A.1, in Appendix A, 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 and 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). This dataset is one of the most widely used in the newer forecasting literature and contains well over 100 monthly (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 A.1 in Appendix A provide 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. I.e., the real-time *FRED-MD* dataset is unbalanced and contains so-called ragged-edges.[3]

Both the outcome variables and the *FRED-MD* variables are collected to span the same time period as the news data, i.e., January 1985 to April 2020.

2.3 Descriptive statistics

In terms of simple descriptive statistics, and using the final vintage of the *FRED-MD* data, Figure 2.2 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.

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 2.2. 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,

[3]If a given variable does not exist for a particular vintage, 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.

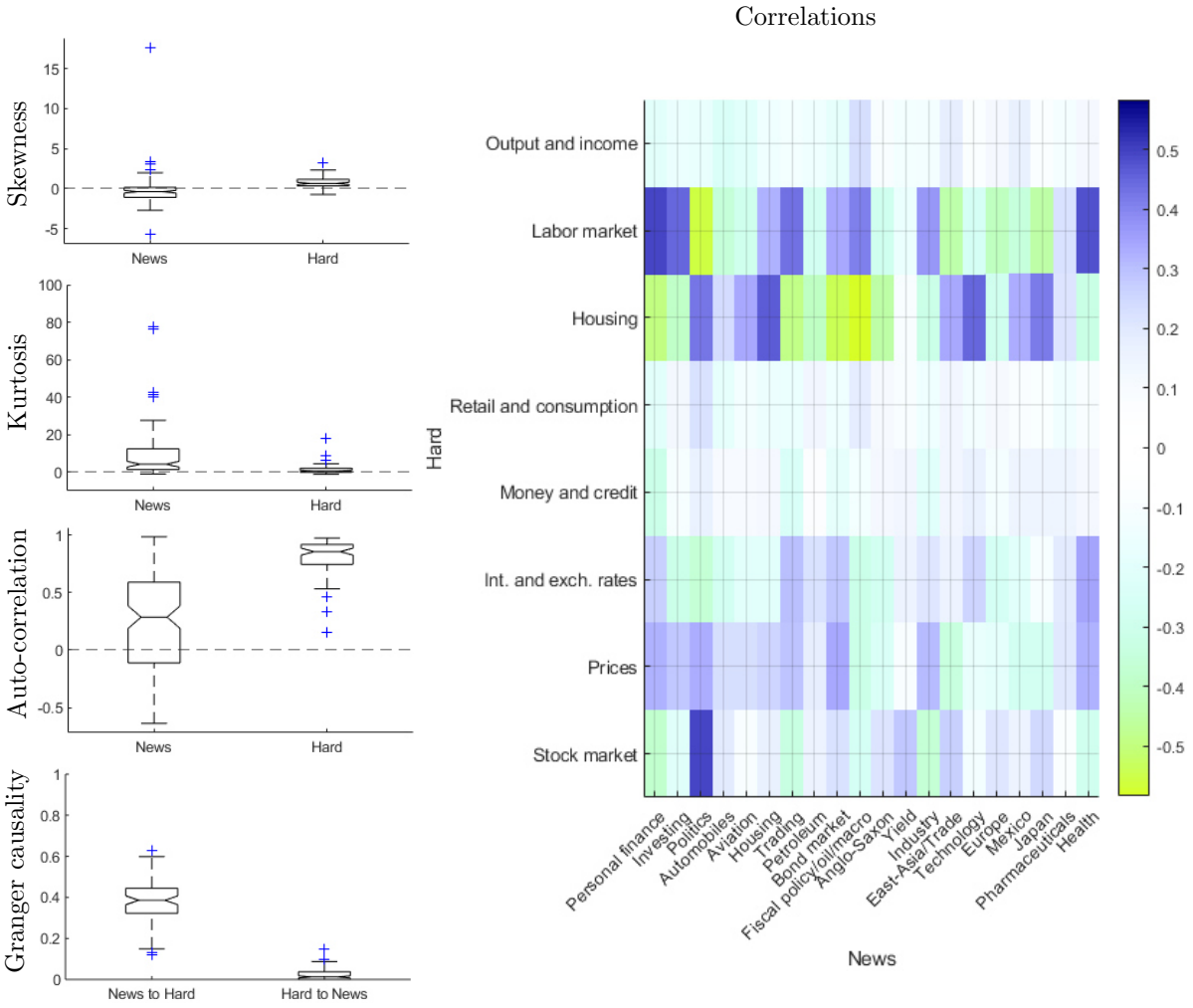


Figure 2.2. Descriptive statistics. The box plots report skewness, kurtosis and the first-order auto-correlation, where the skewness and kurtosis of the normal distribution is defined to be 0. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th 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 A.1 in Appendix A).

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 A.1 and Figure A.1, in Appendix A. 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.[4] For example, on average a news topic Granger causes almost 40 percent of the hard-based variables, whereas a hard-based variable at best Granger causes less than 20 percent 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.

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 (Ghysels et al. (2004), Foroni et al. (2015)).

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

$$y_{t+h}^L = a_h + \sum_{j=0}^{p} \beta_{j,h} L^{j/m} x_t + \varepsilon_{t+h}^L, \tag{3.1}$$

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 (Ghysels et al. (2004), Clements and Galvão (2008), Foroni et al. (2015)). When the set of predictors is low dimensional, estimation can

[4]To handle the high-dimensionality of the problem, the group LASSO (Yuan and Lin (2006)) is used to estimate a Directed Cyclical Graph (DAG), and from that summarize the Granger causality statistics (Lozano et al. (2009), Shojaie and 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 2.2, two-way predictive relationships, i.e., when both the news- and hard-based variable Granger cause each other, are not counted.

be done by Ordinary Least Squares (OLS). Here, where the set of predictors is large, this is not feasible. For this reason (3.1) is estimated using three different approaches; Least Absolute Shrinkage and Selection Operator (LASSO; Tibshirani (1996)); Random Forest (RF; Breiman (2001)); Principal component analysis (PCA) on the predictor set coupled with OLS on a factor augmented version of (3.1). 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 Machine Learning (ML).

In the interest of preserving space, a description of each estimation method is relegated to Appendix C. In short, we use 5-fold 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 three factors 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; $h = 0$, i.e., the current quarter), one- (H1; $h = 1$), and two-quarter ahead (H2; $h = 2$) horizons are produced. Since (3.1) 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 prediction when using M1 data, but a one-quarter ahead prediction using M2 and M3 data. 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 (3.1) 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.[5]

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, i.e., within each recursion and with the appropriate vintage of data, to avoid look-ahead biases. For the same reason, and because it is very 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.

We focus on point forecasting and use Root Mean Squared Errors (RMSE) 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, i.e., the release containing data for 2020Q1 and the first COVID-19 economic effects in the U.S., 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.

4.1 The value of news

Figure 4.1 summarizes our main results. The left column of the figure reports a scatter plot of the RMSE of each model and forecasting horizon, highlighting the overall performance of the news- relative to the hard-based approaches. The second column of the figure

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

statistically compares the best performing news- and hard-based models across forecasting horizons and months, while the third column reports CSPED plots and nowcasting performance for these best performing models.

First, as seen from column two in Figure 4.1, 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 7 out of 9 cases the news-based predictions are the best performing predictors for *Consumption*. The news-based predictions are also significantly different (at the 90 percent 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 news-based predictions tend to have the lowest RMSE.

Second, in terms of models, the news-based predictors work best together with the RF method, which is the best performing news-based model in 85 percent of the cases. For comparison, the RF method is the best performing model in less than 50 percent 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. As seen from the scatter plots in column one in Figure 4.1, it is also a general pattern across all outcome variables that the news-based predictions are more sensitive to method used to produce the predictions. I.e, the variance in model performance is larger for the news-based predictions than it is for the hard-based predictions. The exception to this general pattern is for the hard-based *Consumption* nowcasts (H0), where the LASSO method stands out as particularly good (for M2 or M3).

Third, looking at the CSPED plots, where 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*. 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 (Giannone et al. (2008),

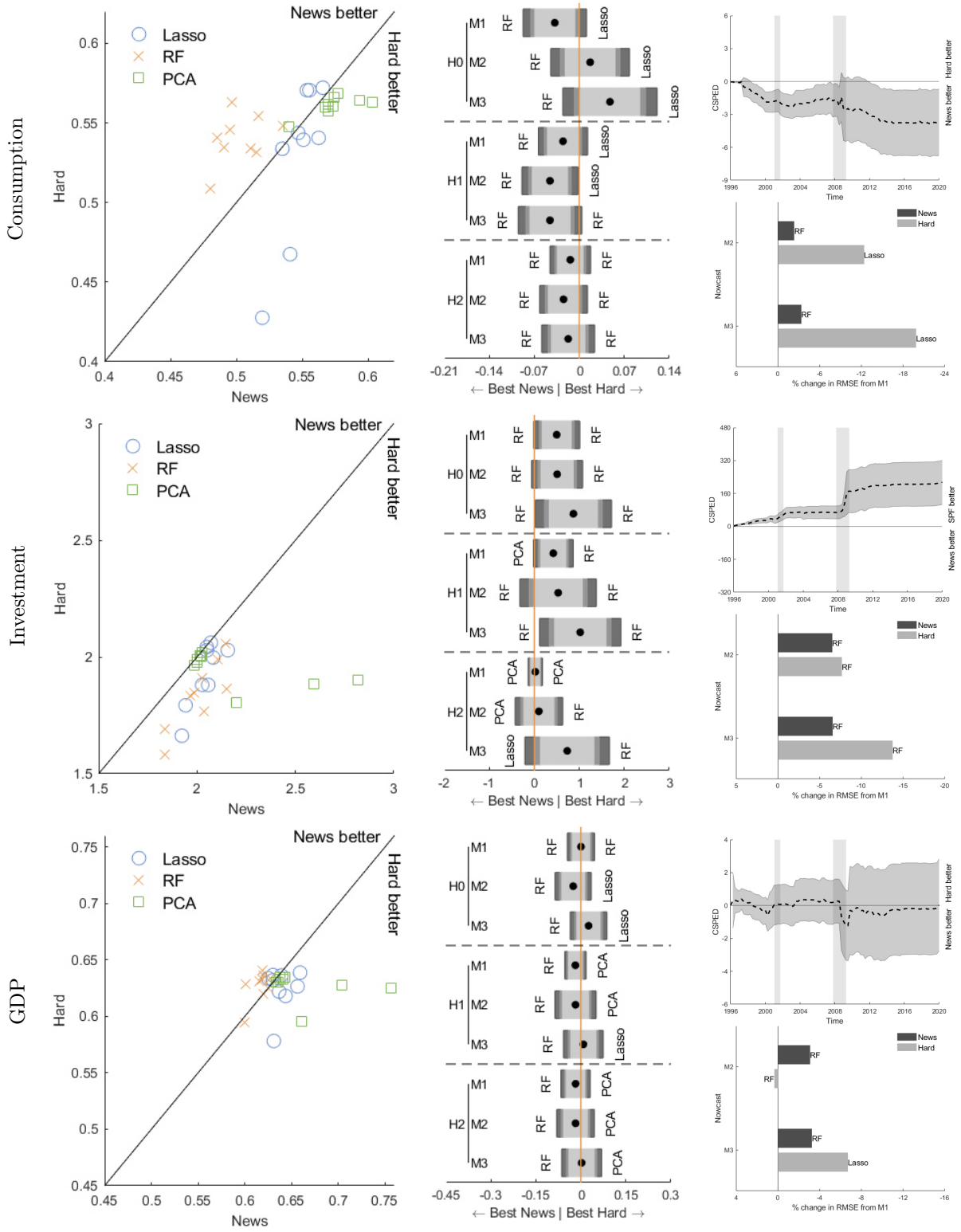


Figure 4.1. Root mean squared errors, cumulative squared prediction error differences and nowcasting. The evaluation sample is 1996Q1-2020Q1. In columns two and three of the figure the best performing news- and hard-based models are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and 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.

Banbura et al. (2011), Aastveit et al. (2014)), namely that predictive performance improves as more hard-based information becomes available throughout the quarter. For the *Consumption* nowcasts produced using the LASSO, for example, the improvement in RMSE from M1 to M3 is roughly 20 percent. 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 4.1, 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.[6]

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, i.e., 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

$$\hat{y}_{t+h} = \sum_{i=1}^{N} w^o_{ith} \hat{y}_{i,t+h}, \tag{4.1}$$

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

$$\boldsymbol{w}^o_{th} = argmin_w \sum_{r=1}^{t-h} (y_{r+h} - \boldsymbol{w}\hat{y}_{r+h})^2, \tag{4.2}$$

which is estimated using OLS under the restriction that the weights are positive and sum to unity.[7]

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

[7] Formally, the weights are optimal in population only to the extent that the joint distribution of outcomes

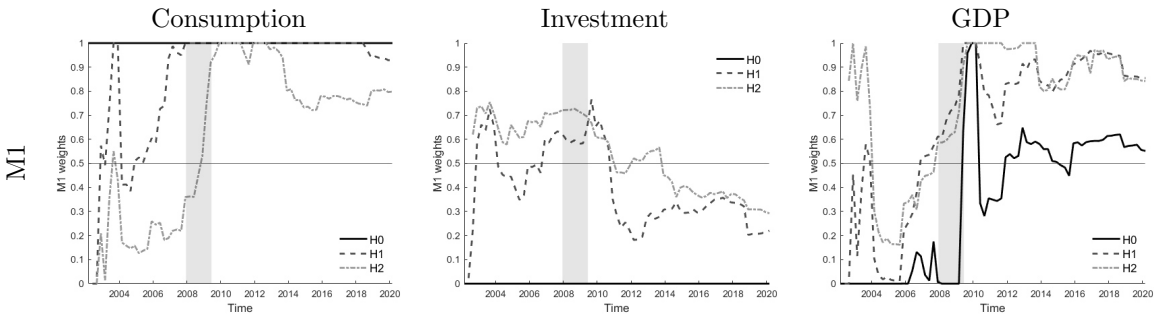


Figure 4.2. Optimal combination and weights. The evaluation sample is 2002Q1-2020Q1, and the weights attached to the news-based models are summed.

24 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 D.1, in Appendix D, report the same type of statistics as in Figure 4.1, 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 D.1), but the best performing models still tend to be hard-based only (column two in Figure D.1). As seen from Figure D.2, in Appendix D, 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 4.2 illustrates how the optimal weights attached to the news-based models vary through time. In the interest of readability and preserving space, the weights 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 percent for a bigger part of the sample irrespective of the forecasting horizon. Moreover, even for *Investment* the news-based

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)).

predictions receive a weight above 20 percent 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 4.3, 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.[8] 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, i.e., 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.[9]

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 percent probability that a news-based *Consumption* predictor will be in the 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 percent 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 2.1 and discussed in Section 2.3. Relatedly, and as pointed out by Larsen and Thorsrud (2018), the particular topic composition of a given

[8] 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.

[9] 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.

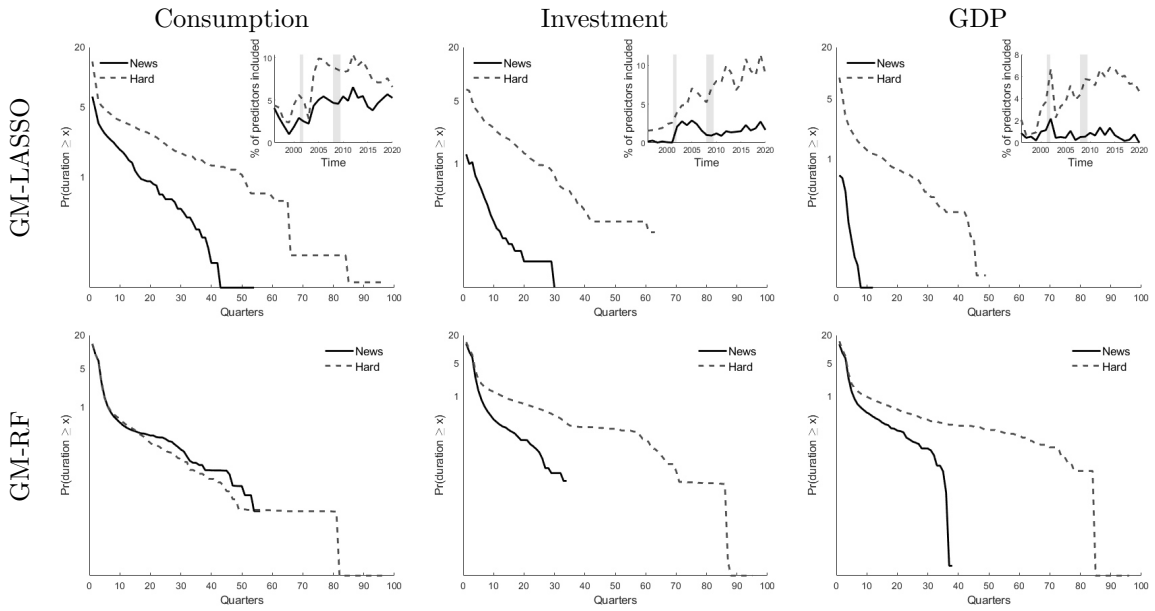


Figure 4.3. Dynamic sparsity and predictor importance. The first row displays the average duration of the predictors and the sparsity (aggregated into average yearly observations) implied by the GM-LASSO. The duration is computed as the probability that a predictor is used by the LASSO when making forecasts in more than x consecutive quarters. The second row shows the average duration of the predictors using the GM-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.

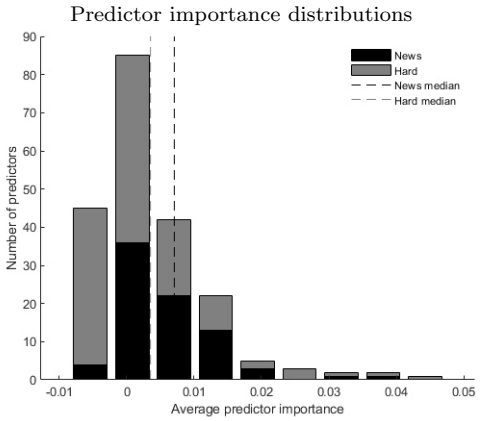
story at a given point in time, might very well be unique, but the topics that the narrative constitute are potentially shared by many other stories at different time periods and with different weighting. Thus, how the topics operate together to form narratives change and evolve over time to a much larger 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 4.4 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 D.3 in Appendix D.

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 pro-

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

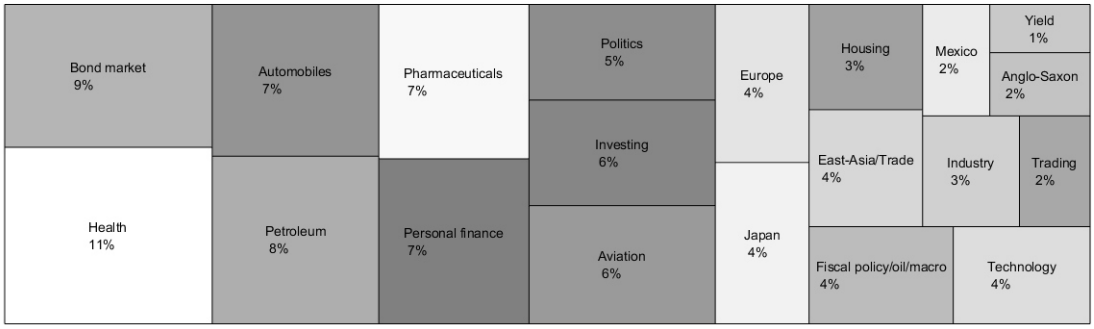


Figure 4.4. GM-RF and predictor importance for *Consumption*. The table reports the top 10 most important predictors on average across the sample, while the histogram reports the empirical distribution of the average predictor importance statistics for the news- and hard-based datasets as a whole. In 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 A.1 in Appendix A). The graph then illustrates the average importance of predictors within each group, where the size of the rectangles represent the group's relative weight.

cess (Larsen et al. (2020)). Moreover, in line with the sparsity statistics discussed above, the upper right histogram in Figure 4.4 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 4.4 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 A.1 in Appendix A, 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 U.S. consumption context.

4.4 How good are the predictions?

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.

The SPF is the oldest quarterly survey of macroeconomic forecasts in the U.S., 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". We use the mean forecasts from the survey, and transform them to quarterly (log) percentage growth rates.

As seen from Table 4.1, the 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 4.5, 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.[10]

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 D.4 in Appendix D). One reason for this is likely that the best performing individual models benefit from having longer time-spans of data available

[10]Results comparing the best hard-based predictions to the simple model-based benchmarks and the SPF are reported in Table D.1 in Appendix D. The overall pattern is very much similar to that described above.

Table 4.1. Relative RMSE scores. The best news-based models are compared to an auto-regressive model (AR), a constant growth rate model (RW) and the 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 RMSE. Significant differences in forecasting performance (marked in gray) are calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). *, **, and *** denote the 10%, 5%, and 1% significance level, respectively.

		H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
Consumption	AR	0.87***	0.89***	0.86**	0.91***	0.90***	0.88***	0.94***	0.89***	0.90***
	RW	0.80**	0.85***	0.46	0.83*	0.86***	0.47	0.88**	0.82*	0.49
	SPF	1.20	1.17	1.16	1.17	1.12	1.11	1.14**	1.09	1.10
Investment	AR	0.93	0.93	0.88	0.95	1.00	0.97	1.00	1.02	1.05
	RW	0.75	0.78	0.69*	0.77	0.84	0.76	0.73*	0.79	0.74
	SPF	1.54***	1.44***	1.43***	1.50***	1.47***	1.51***	1.39**	1.38**	1.40**
GDP	AR	0.98	0.97	0.93	0.97	1.00	0.97	0.97	0.97	0.99
	RW	0.80**	0.86**	0.61	0.80**	0.89*	0.64	0.78**	0.85	0.62
	SPF	1.39**	1.35**	1.35**	1.29*	1.29**	1.30**	1.14	1.13	1.15*

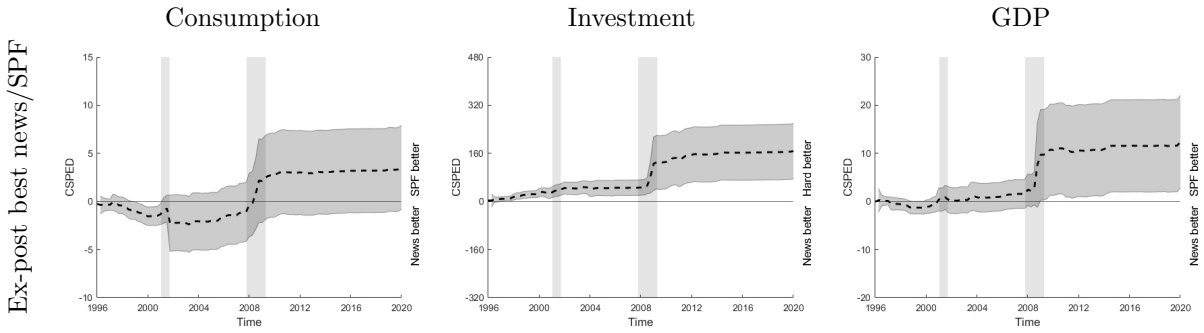


Figure 4.5. SFP and ex-post best news-based forecasts. The graphs compare nowcasting performance (H0), and to align informational available at the time of prediction between the SPF and the model-based forecasts, predictions produced in M2 are used. An upward slope means that the hard economic data outperforms the news data, while the gray band is the equivalent of 90% two-sided levels, based on the Diebold-Mariano test statistic.

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 (3.1) does not change our main qualitative conclusions. I.e., the 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 D.2 in Appendix D). These results are well in line with those presented in Figure 4.2, 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 D.5 and D.6, in Appendix D, 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 D.3 in Appendix D).[11] 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. (2017) and Ardia et al. (2019)).

Finally, using 80 news topics as predictors 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, i.e., 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 D.4 in Appendix D).

[11] In short, for each day and topic, the article that is best explained by each topic is identified and its tone computed, i.e., 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.

5 Conclusion

Decades of research have investigated how hard economic data best can be used for macroeconomic forecasting, i.e., 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 U.S. 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. Still, there are 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. The horse-race has just begun.

References

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

Ardia, D., K. Bluteau, and K. Boudt (2019). Questioning the news about economic growth: Sparse forecasting using thousands of news-based sentiment values. *International Journal of Forecasting 35*(4), 1370 – 1386.

Baffigi, A., R. Golinelli, and G. Parigi (2004). Bridge models to forecast the Euro area GDP. *International Journal of Forecasting 20*(3), 447 – 460.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics 131*(4), 1593–1636.

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

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

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM 55*, 77–84.

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

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees*. CRC Press.

Bybee, L., B. T. Kelly, A. Manela, and D. Xiu (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.

Chang, J., S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty,

C. Williams, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems 22*, pp. 288–296. Cambridge, MA: The MIT Press.

Clements, M. P. and A. B. Galvão (2008). Macroeconomic forecasting with mixed-frequency data. *Journal of Business & Economic Statistics 26*(4), 546–554.

Croushore, D. and T. Stark (2001). A real-time data set for macroeconomists. *Journal of Econometrics 105*(1), 111 – 130. Forecasting and empirical methods in finance and macroeconomics.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–63.

Dybowski, T. and P. Adämmer (2018). The economic effects of u.s. presidential tax communication: Evidence from a correlated topic model. *European Journal of Political Economy 55*, 511 – 525.

Foroni, C., M. Marcellino, and C. Schumacher (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.

Gentzkow, M., B. Kelly, and M. Taddy (2019). Text as Data. *Journal of Economic Literature 57*(3), 535–74.

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

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Griffiths, T. L. and M. Steyvers (2004). Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America 101*(Suppl 1), 5228–5235.

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

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios, and S. Kapadia (2020). Making Text Count: Economic Forecasting Using Newspaper Text. Bank of England working papers 865, Bank of England.

Kuzin, V., M. Marcellino, and C. Schumacher (2011). Midas vs. mixed-frequency VAR: Nowcasting GDP in the euro area. *International Journal of Forecasting 27*(2), 529–542.

Larsen, V. H. (2017). Components of uncertainty. Working Paper 2017/5, Norges Bank.

Larsen, V. H. and L. A. Thorsrud (2017). Asset returns, news topics, and media effects. Working Paper 2017/17, Norges Bank.

Larsen, V. H. and L. A. Thorsrud (2018). Business Cycle Narratives. Working Paper 2018/03, Norges Bank.

Larsen, V. H. and L. A. Thorsrud (2019). The value of news for economic developments. *Journal of Econometrics 210*(1), 203–218.

Larsen, V. H., L. A. Thorsrud, and J. Zhulanova (2020). News-driven inflation expectations and information rigidities. *Journal of Monetary Economics* (Forthcoming).

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014). The parable of google flu: Traps in big data analysis. *Science 343*(6176), 1203–1205.

Lozano, A. C., N. Abe, Y. Liu, and S. Rosset (2009). Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics 25*(12), i110–i118.

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

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

Murtagh, F. and P. Legendre (2014). Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? *Journal of Classification 31*(3), 274–295.

Nimark, K. P. and S. Pitschner (2019). News media and delegated information choice. *Journal of Economic Theory 181*, 160 – 196.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2017). Measuring News Sentiment. Working Paper Series 2017-1, Federal Reserve Bank of San Francisco.

Shojaie, A. and G. Michailidis (2010). Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics 26*(18), i517–i523.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50*(3), 665 – 690.

Stark, T. and D. Croushore (2002, December). Forecasting with a real-time data set for macroeconomists. *Journal of Macroeconomics 24*(4), 507–531.

Stark, T. and D. Croushore (2019). Fifty Years of the Survey of Professional Forecasters. Federal Reserve Bank of Philadelphia Economic Insights 2019Q4, FRB Philadelphia.

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

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–62.

Stock, J. H. and M. W. Watson (2003). Forecasting output and inflation: The role of asset prices. *Journal of Economic Literature 41*(3), 788–829.

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

Timmermann, A. (2006). Forecast combinations. Volume 1 of *Handbook of Economic Forecasting*, Chapter 4, pp. 135–196. Amsterdam, North Holland: Elsevier.

Ulbricht, D., K. A. Kholodilin, and T. Thomas (2017). Do media data help to predict german industrial production? *Journal of Forecasting 36*(5), 483–496.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Appendices for online publication

Appendix A Data

Table A.1. *FRED-MD* variables. The "ID" column denotes the variable's identity number in the *FRED-MD* database. The column denoted "TC" defines how the series, x, are transformed, with: (1) no transformation, (2) Δx_t, (3) $\Delta^2 x_t$, (4) $log(x_t)$, (5) $\Delta log(x_t)$, (6) $\Delta^2 log(x_t)$ and (7) $\Delta(x_t/x_{t-1}-1)$. Finally, the "Lag" column lists the typical release lag structure. E.g., the *Real Personal Income* statistic for the current quarter is typically released with a 1 month lag. The *FRED-MD* database is updated in real time and can be downloaded from https://research.stlouisfed.org/econ/mccracken/fred-databases/, and historical vintages of the data can be obtained from the same source.

Variable	ID	TC	Lag	Group
Real Personal Income	1	5	1	Output and income
Real personal income ex transfer receipts	2	5	1	Output and income
IP Index	6	5	1	Output and income
IP: Final Products and Nonindustrial Supplies	7	5	1	Output and income
IP: Final Products (Market Group)	8	5	1	Output and income
IP: Consumer Goods	9	5	1	Output and income
IP: Durable Consumer Goods	10	5	1	Output and income
IP: Nondurable Consumer Goods	11	5	1	Output and income
IP: Business Equipment	12	5	1	Output and income
IP: Materials	13	5	1	Output and income
IP: Durable Materials	14	5	1	Output and income
IP: Nondurable Materials	15	5	1	Output and income
IP: Manufacturing (SIC)	16	5	1	Output and income
IP: Residential Utilities	17	5	1	Output and income
IP: Fuels	18	5	1	Output and income
Capacity Utilization: Manufacturing	20	2	1	Output and income
Help-Wanted Index for United States	21	2	2	Labor market
Ratio of Help Wanted/No. Unemployed	22	2	2	Labor market
Civilian Labor Force	23	5	1	Labor market
Civilian Employment	24	5	1	Labor market
Civilian Unemployment Rate	25	2	1	Labor market
Average Duration of Unemployment (Weeks)	26	2	1	Labor market
Civilians Unemployed - Less Than 5 Weeks	27	5	1	Labor market
Civilians Unemployed for 41760 Weeks	28	5	1	Labor market
Civilians Unemployed - 15 Weeks & Over	29	5	1	Labor market
Civilians Unemployed for 15-26 Weeks	30	5	1	Labor market
Civilians Unemployed for 27 Weeks and Over	31	5	1	Labor market
Initial Claims	32	5	1	Labor market
All Employees: Total nonfarm	33	5	1	Labor market
All Employees: Goods-Producing Industries	34	5	1	Labor market
All Employees: Mining and Logging: Mining	35	5	1	Labor market
All Employees: Construction	36	5	1	Labor market
All Employees: Manufacturing	37	5	1	Labor market
All Employees: Durable goods	38	5	1	Labor market
All Employees: Nondurable goods	39	5	1	Labor market
All Employees: Service-Providing Industries	40	5	1	Labor market
All Employees: Trade, Transportation & Utilities	41	5	1	Labor market
All Employees: Wholesale Trade	42	5	1	Labor market
All Employees: Retail Trade	43	5	1	Labor market
All Employees: Financial Activities	44	5	1	Labor market
All Employees: Government	45	5	1	Labor market

Continued on next page

Table A.1 – *Continued from previous page*

Variable	ID	TC	Lag	Group
Avg Weekly Hours : Goods-Producing	46	1	1	Labor market
Avg Weekly Overtime Hours : Manufacturing	47	2	1	Labor market
Avg Weekly Hours : Manufacturing	48	1	1	Labor market
Avg Hourly Earnings : Goods-Producing	127	6	1	Labor market
Avg Hourly Earnings : Construction	128	6	1	Labor market
Avg Hourly Earnings : Manufacturing	129	6	1	Labor market
Housing Starts: Total New Privately Owned	50	4	1	Housing
Housing Starts, Northeast	51	4	1	Housing
Housing Starts, Midwest	52	4	1	Housing
Housing Starts, South	53	4	1	Housing
Housing Starts, West	54	4	1	Housing
New Private Housing Permits (SAAR)	55	4	1	Housing
New Private Housing Permits, Northeast (SAAR)	56	4	1	Housing
New Private Housing Permits, Midwest (SAAR)	57	4	1	Housing
New Private Housing Permits, South (SAAR)	58	4	1	Housing
New Private Housing Permits, West (SAAR)	59	4	1	Housing
Real personal consumption expenditures	3	5	1	Consumption, orders and invent.
Real Manu. and Trade Industries Sales	4	5	2	Consumption, orders and invent.
Retail and Food Services Sales	5	5	1	Consumption, orders and invent.
New Orders for Durable Goods	65	5	1	Consumption, orders and invent.
New Orders for Nondefense Capital Goods	66	5	1	Consumption, orders and invent.
Unlled Orders for Durable Goods	67	5	1	Consumption, orders and invent.
Total Business Inventories	68	5	2	Consumption, orders and invent.
Total Business: Inventories to Sales Ratio	69	2	2	Consumption, orders and invent.
Consumer Sentiment Index	130	2	1	Consumption, orders and invent.
M1 Money Stock	70	6	1	Money and credit
M2 Money Stock	71	6	1	Money and credit
Real M2 Money Stock	72	5	1	Money and credit
St. Louis Adjusted Monetary Base	73	6	1	Money and credit
Total Reserves of Depository Institutions	74	6	1	Money and credit
Reserves Of Depository Institutions	75	7	1	Money and credit
Commercial and Industrial Loans	76	6	1	Money and credit
Real Estate Loans at All Commercial Banks	77	6	1	Money and credit
Total Nonrevolving Credit	78	6	3	Money and credit
Nonrevolving consumer credit to Personal Income	79	2	3	Money and credit
MZM Money Stock	131	6	1	Money and credit
Consumer Motor Vehicle Loans Outstanding	132	6	2	Money and credit
Total Consumer Loans and Leases Outstanding	133	6	2	Money and credit
Securities in Bank Credit at All Commercial Banks	134	6	1	Money and credit
Eective Federal Funds Rate	84	2	0	Interest and exchange rates
3-Month AA Financial Commercial Paper Rate	85	2	0	Interest and exchange rates
3-Month Treasury Bill:	86	2	0	Interest and exchange rates
6-Month Treasury Bill:	87	2	0	Interest and exchange rates
1-Year Treasury Rate	88	2	0	Interest and exchange rates
5-Year Treasury Rate	89	2	0	Interest and exchange rates
10-Year Treasury Rate	90	2	0	Interest and exchange rates
Moody's Seasoned Aaa Corporate Bond Yield	91	2	0	Interest and exchange rates
Moody's Seasoned Baa Corporate Bond Yield	92	2	0	Interest and exchange rates
3-Month Commercial Paper Minus FEDFUNDS	93	1	0	Interest and exchange rates
3-Month Treasury C Minus FEDFUNDS	94	1	0	Interest and exchange rates
6-Month Treasury C Minus FEDFUNDS	95	1	0	Interest and exchange rates
1-Year Treasury C Minus FEDFUNDS	96	1	0	Interest and exchange rates
5-Year Treasury C Minus FEDFUNDS	97	1	0	Interest and exchange rates
10-Year Treasury C Minus FEDFUNDS	98	1	0	Interest and exchange rates
Moody's Aaa Corporate Bond Minus FEDFUNDS	99	1	0	Interest and exchange rates
Moody's Baa Corporate Bond Minus FEDFUNDS	100	1	0	Interest and exchange rates

Continued on next page

Table A.1 – *Continued from previous page*

Variable	ID	TC	Lag	Group
Trade Weighted U.S. Dollar Index: Major Currencies	101	5	0	Interest and exchange rates
Switzerland / U.S. Foreign Exchange Rate	102	5	0	Interest and exchange rates
Japan / U.S. Foreign Exchange Rate	103	5	0	Interest and exchange rates
U.S. / U.K. Foreign Exchange Rate	104	5	0	Interest and exchange rates
Canada / U.S. Foreign Exchange Rate	105	5	0	Interest and exchange rates
PPI: Finished Goods	106	5	1	Prices
PPI: Finished Consumer Goods	107	5	1	Prices
PPI: Intermediate Materials	108	5	1	Prices
PPI: Crude Materials	109	5	1	Prices
Crude Oil, spliced WTI and Cushing	110	5	1	Prices
PPI: Metals and metal products	111	5	1	Prices
CPI : All Items	113	5	1	Prices
CPI : Apparel	114	5	1	Prices
CPI : Transportation	115	5	1	Prices
CPI : Medical Care	116	5	1	Prices
CPI : Commodities	117	5	1	Prices
CPI : Durables	118	5	1	Prices
CPI : Services	119	5	1	Prices
CPI : All Items Less Food	120	5	1	Prices
CPI : All items less shelter	121	5	1	Prices
CPI : All items less medical care	122	5	1	Prices
Personal Cons. Expend.: Chain Index	123	5	1	Prices
Personal Cons. Exp: Durable goods	124	5	1	Prices
Personal Cons. Exp: Nondurable goods	125	5	1	Prices
Personal Cons. Exp: Services	126	5	1	Prices
S&P's Common Stock Price Index: Composite	80	5	0	Stock market
S&P's Common Stock Price Index: Industrials	81	5	0	Stock market
S&P's Composite Common Stock: Dividend Yield	82	2	0	Stock market
S&P's Composite Common Stock: Price-Earnings Ratio	83	5	0	Stock market
VXO	135	1	0	Stock market

Table A.2. US news topics. Subjective labeling and the most important words with weights.

Id		Top words (word probability)
0	Personal finance	loan 0.074, real 0.058, asset 0.055, trust 0.052, save 0.047, deposit 0.037, estat 0.036
1	Civil commissions	file 0.114, commiss 0.095, sec 0.046, partner 0.032, partnership 0.031, propos 0.024, outstand 0.021
2	Communication	spokesman 0.068, comment 0.065, today 0.05, declin 0.033, didnt 0.029, immedi 0.026, sourc 0.024
3	Income	incom 0.065, cash 0.06, expens 0.041, asset 0.038, tax 0.027, loss 0.024, account 0.024
4	Defense contracts	contract 0.086, receiv 0.052, guilder 0.026, engin 0.024, equip 0.022, defens 0.021, forc 0.02
5	Economic analysis	analyst 0.033, think 0.033, dont 0.026, look 0.024, much 0.021, peopl 0.018, want 0.018
6	Stock market	analyst 0.112, volum 0.075, estim 0.064, cent 0.061, averag 0.052, compar 0.041, daili 0.037
7	Petroleum	oil 0.096, ga 0.094, natur 0.038, energi 0.031, petroleum 0.026, pipelin 0.022, field 0.021
8	Mortgage market	mortgag 0.083, home 0.065, associ 0.04, loan 0.026, certif 0.023, feder 0.021, base 0.015
9	Oil trading	crude 0.021, trader 0.019, fuel 0.018, oil 0.018, cargo 0.017, brent 0.015, mt 0.015
10	Data release	period 0.105, earlier 0.085, rose 0.072, compar 0.068, figur 0.065, half 0.044, latest 0.043
11	Analysis	deal 0.025, big 0.019, much 0.013, analyst 0.011, bare 0.011, come 0.011, move 0.01
12	Agriculture	food 0.051, paper 0.024, brand 0.019, produc 0.016, agricultur 0.012, mill 0.012, tobacco 0.011
13	Korea	south 0.062, won 0.051, north 0.05, korea 0.041, electron 0.031, buy 0.031, lead 0.019
14	Manufacturing	project 0.099, ventur 0.079, joint 0.071, construct 0.056, steel 0.043, build 0.043, plant 0.03
15	Money market	central 0.088, money 0.061, repurchas 0.03, discount 0.027, rupiah 0.024, call 0.024, liquid 0.022
16	Restructuring	cost 0.111, program 0.093, reduc 0.073, restructur 0.044, reduct 0.034, improv 0.031, effect 0.022
17	Safety	peopl 0.021, caus 0.016, ship 0.016, fire 0.015, area 0.014, damag 0.013, spokesman 0.01
18	Options trading	right 0.1, option 0.077, warrant 0.052, class 0.037, exercis 0.035, outstand 0.026, sharehold 0.025
19	Health care	health 0.065, care 0.044, medic 0.042, center 0.038, hospit 0.027, america 0.022, healthcar 0.021
20	Utilities	power 0.101, electr 0.082, util 0.055, energi 0.036, plant 0.031, public 0.022, nuclear 0.021
21	Canada	canada 0.103, canadian 0.093, quebec 0.017, today 0.016, toronto 0.014, provinc 0.013, ontario 0.013
22	China	china 0.084, hong 0.055, kong 0.053, hk 0.033, taiwan 0.03, singapor 0.028, chines 0.022
23	Macroeconomics	growth 0.085, economi 0.056, econom 0.056, inflat 0.031, rise 0.021, slow 0.02, forecast 0.02
24	US states	pacif 0.041, texa 0.033, west 0.03, california 0.029, san 0.025, southern 0.022, calif 0.018
25	UK	pound 0.099, uk 0.086, london 0.041, british 0.04, penc 0.029, sterl 0.021, england 0.018
26	Credit	debt 0.068, moodi 0.041, sp 0.037, senior 0.029, agenc 0.028, standard 0.021, poor 0.02
27	Composite index	index 0.101, gain 0.052, fell 0.031, volum 0.03, rose 0.024, higher 0.024, drop 0.022
28	Automobiles	car 0.054, motor 0.046, auto 0.041, vehicl 0.03, part 0.024, manufactur 0.024, truck 0.023
29	Trading	trader 0.079, dealer 0.037, session 0.025, open 0.023, earli 0.021, buy 0.02, higher 0.019
30	Russia	russia 0.017, soviet 0.016, russian 0.015, militari 0.013, republ 0.012, israel 0.011, peac 0.011
31	Business growth	competit 0.032, posit 0.027, growth 0.027, expand 0.023, strategi 0.022, opportun 0.019, grow 0.018
32	Yield	na 0.071, contact 0.048, form 0.041, issuer 0.034, symbol 0.031, editori 0.028, type 0.024
33	Futures contracts	futur 0.107, contract 0.093, cent 0.037, crude 0.03, oil 0.026, gasolin 0.022, full 0.021
34	Insurance	insur 0.127, life 0.05, premium 0.025, base 0.018, subsidiari 0.016, chase 0.013, benefit 0.012
35	Technology	comput 0.05, softwar 0.032, technolog 0.028, ibm 0.016, avail 0.013, applic 0.011, machin 0.01
36	Germany	mark 0.106, german 0.074, germani 0.065, deutsch 0.05, ag 0.039, bundesbank 0.034, europ 0.028
37	Mexico	peso 0.076, mexico 0.052, de 0.028, mexican 0.026, philippin 0.018, brazil 0.018, sa 0.017
38	Japan	yen 0.141, japan 0.124, japanes 0.068, tokyo 0.036, fiscal 0.035, pretax 0.024, ministri 0.022
39	Negotiation	possibl 0.037, decis 0.035, discuss 0.028, review 0.023, consid 0.022, negoti 0.022, regard 0.018
40	Financing	debt 0.107, credit 0.101, financ 0.085, loan 0.037, borrow 0.037, payment 0.036, facil 0.028
41	Currencies	dollar 0.177, late 0.069, currenc 0.066, mark 0.06, york 0.045, yen 0.034, around 0.028
42	Disclosure	term 0.122, privat 0.076, disclos 0.064, subsidiari 0.053, distribut 0.044, sign 0.034, acquir 0.033
43	Regulations	requir 0.047, allow 0.036, regul 0.032, rule 0.031, can 0.018, limit 0.017, law 0.017
44	Stock options	prefer 0.062, dividend 0.061, convert 0.053, debentur 0.045, amount 0.039, due 0.039, outstand 0.039
45	France	franc 0.173, french 0.049, swiss 0.034, de 0.023, sa 0.023, baht 0.021, belgian 0.018
46	The Middle East	oil 0.03, iraq 0.026, countri 0.019, gulf 0.019, opec 0.018, minist 0.016, saudi 0.016
47	Aviation	airlin 0.059, air 0.041, transport 0.03, passeng 0.021, flight 0.02, carrier 0.019, mile 0.018
48	Chemical industry	chemic 0.056, manufactur 0.052, technolog 0.034, materi 0.032, environment 0.025, plant 0.022
49	Bond market	treasuri 0.051, bill 0.046, due 0.045, bid 0.044, auction 0.043, yield 0.043, basi 0.037
50	Leadership	execut 0.099, chief 0.097, offic 0.065, chairman 0.065, vice 0.041, name 0.035, director 0.035
51	Monetary policy	fed 0.059, reserv 0.054, feder 0.054, polici 0.049, monetari 0.028, inflat 0.024, eas 0.022
52	Retail	store 0.084, retail 0.061, open 0.026, chain 0.016, restaur 0.015, entertain 0.013, depart 0.011
53	Investment banking	firm 0.058, initi 0.048, underwrit 0.043, brother 0.034, public 0.034, merril 0.03, morgan 0.03
54	Results	quarter 0.19, loss 0.095, incom 0.082, cent 0.07, revenu 0.064, charg 0.027, fiscal 0.024
55	Australia	stake 0.094, australian 0.04, australia 0.039, rais 0.026, sold 0.025, bought 0.023, control 0.022
56	Telecommunication	commun 0.07, network 0.048, telephon 0.037, telecommun 0.035, cabl 0.027, telecom 0.02, bell 0.019
57	Media	street 0.054, wall 0.047, news 0.046, jone 0.042, publish 0.036, newspap 0.031, journal 0.024
58	Fiscal policy	tax 0.09, budget 0.076, cut 0.06, deficit 0.057, fiscal 0.053, spend 0.041, balanc 0.022
59	Natural resources	gold 0.072, mine 0.058, metal 0.032, resourc 0.031, ounc 0.02, rand 0.019, ton 0.018
60	Pharmaceuticals	drug 0.04, pharmaceut 0.018, research 0.015, test 0.015, studi 0.014, patient 0.013, approv 0.012
61	Fear	problem 0.041, concern 0.024, need 0.018, can 0.015, risk 0.015, situat 0.013, question 0.012
62	US politics	hous 0.056, clinton 0.029, bill 0.029, administr 0.028, senat 0.027, committe 0.027, congress 0.023
63	Commercial papers	pc 0.336, commerci 0.034, six 0.031, major 0.03, corpor 0.029, repres 0.028, paper 0.027
64	Mergers and acquis.	acquisit 0.075, transact 0.069, merger 0.067, approv 0.059, complet 0.049, acquir 0.045
65	Regions	state 0.155, citi 0.051, york 0.047, counti 0.038, author 0.036, depart 0.023, municip 0.018
66	Employment conditions	union 0.076, employe 0.051, work 0.046, worker 0.044, job 0.038, strike 0.034, employ 0.03
67	Institutional investing	fund 0.228, institut 0.042, equiti 0.034, asset 0.032, portfolio 0.023, deriv 0.021, money 0.021
68	Scandinavia	signific 0.042, headquart 0.039, figur 0.038, histori 0.034, kronor 0.033, currenc 0.03
69	Elections	parti 0.047, elect 0.035, minist 0.031, polit 0.028, prime 0.027, vote 0.024, support 0.017
70	SEC investigations	investig 0.028, charg 0.026, former 0.018, alleg 0.014, case 0.013, offic 0.012, depart 0.012
71	Trade	foreign 0.091, import 0.076, export 0.068, countri 0.05, world 0.041, domest 0.035, region 0.029
72	Italy	american 0.15, lire 0.042, itali 0.03, italian 0.026, trillion 0.024, express 0.017, great 0.015
73	EU	meet 0.039, european 0.031, countri 0.029, talk 0.025, minist 0.024, econom 0.023, eu 0.019
74	Data revisions	improv 0.041, declin 0.038, higher 0.036, margin 0.035, strong 0.034, demand 0.033, perform 0.027
75	Statistics	data 0.048, adjust 0.045, rise 0.031, rose 0.028, season 0.027, revis 0.024, survey 0.024
76	Board of Directors	board 0.111, sharehold 0.091, propos 0.068, meet 0.067, approv 0.043, vote 0.03, director 0.029
77	Real estate	properti 0.058, hotel 0.028, game 0.028, land 0.027, casino 0.024, ringgit 0.019, park 0.014
78	Litigation	court 0.067, file 0.028, claim 0.027, settlement 0.025, suit 0.022, bankruptci 0.02, appeal 0.019
79	Government debt	bond 0.203, yield 0.054, treasuri 0.044, futur 0.029, late 0.026, spread 0.023, basi 0.022

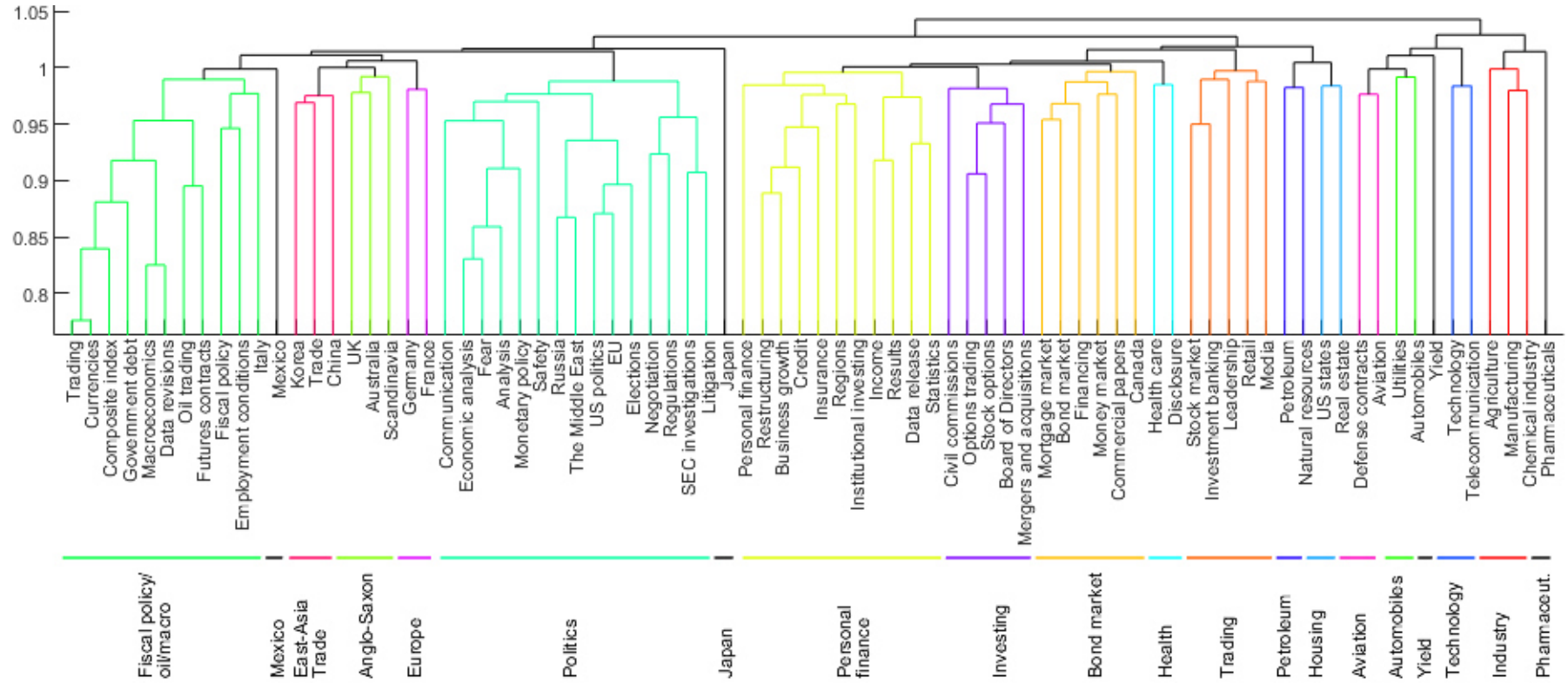


Figure A.1. A dendrogram of the news topic decomposition. The 80 topics listed in Table A.2 might not have been given names by us that are intuitive and easy to understand. As an alternative, to help interpretation, one could interpret each topic as belonging to clusters of higher order abstractions, such as politics, technology, etc.. The figure, reporting a dendrogram of the news topic decomposition, is one way of formally doing this that is often applied in the NLP literature, and by, e.g., Bybee et al. (2019) in economics. The graph is constructed using a hierarchical agglomerative clustering algorithm (Murtagh and Legendre (2014)) on the LDA word distribution output. The connectivity of topics in the dendrogram is data driven, while the cutoff into 20 clusters, and the names given to the clusters, are subjectively given.

Appendix B Feature selection and the LDA

We apply three feature selection steps on the news data corpus prior to estimation. First, a stop-word list is employed. This is a list of common words not expected to have any information relating to the subject of an article. Examples of such words are *the*, *is*, *are*, and *this*. Next, an algorithm known as stemming is run. The objective of this algorithm is to reduce all words to their respective word stems. A word stem is the part of a word that is common to all of its inflections. An example is the word *effective* whose stem is *effect*. Finally, a measure called *tf-idf*, which stands for term frequency - inverse document frequency, is calculated. This measures how important all the words in the complete corpus are in explaining single articles. The more often a word occurs in an article, the higher the *tf-idf* score of that word. On the other hand, if the word is common to all articles, meaning the word has a high frequency in the whole corpus, the lower that word's *tf-idf* score will be. 160.000 of the stems with the highest *tf-idf* score are kept, and used as the final corpus.

The LDA is implemented on the cleaned corpus. More formally, denote the whole corpus as M distinct documents (articles), $N = \sum_{m=1}^{M} N_m$ the total number of words in all documents, and K the total number of latent topics. Letting bold-font variables denote the vector version of variables, the distribution of topics for a document is given by $\boldsymbol{\theta}_m$, while the distribution of words for each topic is determined by $\boldsymbol{\varphi}_k$. Both $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$ are assumed to have conjugate Dirichlet distributions with hyper-parameters (vectors) α and β, respectively. Then, each document consists of a repeated choice of topics $Z_{m,n}$ and words $W_{m,n}$, drawn from the Multinomial distribution using $\boldsymbol{\theta}_m$ and $\boldsymbol{\varphi}_k$.

In the LDA, the joint distribution of all known and hidden variables given the hyper-parameters is:

$$P(\boldsymbol{W}_m, \boldsymbol{Z}_m, \boldsymbol{\theta}_m, \boldsymbol{\Phi}; \alpha, \beta) = \overbrace{\underbrace{\prod_{n=1}^{N_m} P(W_{m,n}|\boldsymbol{\varphi}_{z_{m,n}}) P(Z_{m,n}|\boldsymbol{\theta}_m)}_{\text{word plate}} \cdot P(\boldsymbol{\theta}_m; \alpha)}^{\text{document plate (1 document)}} \cdot \underbrace{P(\boldsymbol{\Phi}; \beta)}_{\text{topic plate}} \tag{B.1}$$

where $\boldsymbol{\Phi} = \{\boldsymbol{\varphi}_k\}_{k=1}^{K}$ is a $(K \times V)$ matrix, and V is the size of the vocabulary. The unknown distributions in (B.1) can be estimated using many different methods. We refer to Larsen and Thorsrud (2019) or Thorsrud (2018) for a more thorough technical description of how the LDA is implemented here, but note that symmetric Dirichlet priors are used for α and β which are defined as a function of the number of topics and unique words: $\alpha = 50/K$ and $\beta = 200/N$. These values are in essence the same as advocated by Griffiths and Steyvers (2004).

Appendix C Predictive models

The LASSO augments the standard linear regression model with a penalization term on the absolute value of the coefficients (Tibshirani (1996)). This introduces sparsity since some coefficients will be set equal to zero. Accordingly, the LASSO performs both variable selection and regularization of the parameters, and thus works in high-dimensional settings as here.

Formally, the LASSO solves the penalized least squares problem

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{argmin} \frac{1}{T-p+1} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1, \tag{C.1}$$

where $\lambda \geq 0$ is the penalty parameter deciding the amount of regularization, $\mathbf{y} = (y_{p+h}, \ldots, y_{T+h})'$ is a $(T-p+1) \times 1$ vector of the response variable and

$$\mathbf{X} = \begin{bmatrix} x_{p,1} & x_{p-1,1} & \ldots & x_{1,1} & \ldots & x_{p,N} & x_{p-1,N} & \ldots & x_{1,N} \\ x_{p+1,1} & x_{p,1} & \ldots & x_{2,1} & \ldots & x_{p+1,N} & x_{p,N} & \ldots & x_{2,N} \\ \vdots & & & & \ddots & & & & \\ x_{T,1} & \ldots & \ldots & \ldots & \ldots & \ldots & x_{T-1,N} & \ldots & x_{T,N} \end{bmatrix} \tag{C.2}$$

is a $(T-p+1) \times pN$ matrix of predictors h periods lagged behind $\mathbf{y}$ where p denotes the number of lags of the predictors and h denotes the forecast horizon. Following common practice, we use 5-fold cross validation and mean squared error (MSE) loss to tune the regularization parameter (λ), and all variables in (C.1) are standardized prior to estimation to make estimation invariant to scale.

The RF is an ensemble method proposed by Breiman (2001). The RF method builds on regression trees, but reduces the variance of single regression trees by combining them using bootstrap aggregation (bagging) when forming predictions. Conceptually, a regression tree is a non-parametric model that estimates an unknown non-linear function of covariates to form a prediction by recursively partioning the covariate space (Breiman et al. (1984)). Thus, the RF method can handle both high-dimensional predictive problems as well as incorporate non-linear relationships.

More formally, and following Hastie et al. (2001), a regression tree recursively applies binary partitions on the predictor space. The predictors and the splitting points s are chosen at each step of the algorithm to minimize the sum of squares. The binary partition gives the following two half-planes

$$R_1(j,s) = \{\mathbf{x}_t | x_{tj} \leq s\} \quad \text{and} \quad R_2(j,s) = \{\mathbf{x}_t | x_{tj} > s\}, \tag{C.3}$$

where j and s are chosen subject to

$$\underset{j,s}{min} \Big[\underset{c_1}{min} \sum_{\mathbf{x}_t \in R_1(j,s)} (y_t - c_1)^2 + \underset{c_2}{min} \sum_{\mathbf{x}_t \in R_2(j,s)} (y_t - c_2)^2 \Big], \tag{C.4}$$

and the solution to the inner minimization problem is

$$\hat{c}_1 = ave(y_t|\mathbf{x}_t \in R_1(j,s)) \quad \text{and} \quad \hat{c}_2 = ave(y_t|\mathbf{x}_t \in R_2(j,s)). \tag{C.5}$$

Hence, by applying a "greedy" approach, looping through all the j predictors and potential splitting points s, minimization of (C.4) is obtained.

A downside with regression trees is that they can easily be grown too big and thus over-fit the data in sample. This naturally degrades their out-of-sample predictive performance. For this reason, the RF combines many individual regression trees using bagging together with the random subspace method, and works in the following way. First, the dataset is split into B bootstrap samples with the same length as the original dataset (bagging), and only a random subset of $m < j$ predictors are considered (random subspace). Next, a regression tree is fit on this dataset. Aggregation is done by averaging the predictions from all the B trees grown. A nice feature of this approach is that in each bootstrap sample a natural candidate for validation sample exists, namely the observations not used in the bootstrap sample. This is called the out-of-bag-sample. Accordingly, all the trees are grown to minimize the prediction error in these samples, reducing the over-fit problem associated with using only one (big) regression tree. Following conventional practice, we set $B = 500$ and consider $1/3$ of the predictors in each bootstrap sample.

Finally, the factor augmented MIDAS builds on a simple Principal Component Analysis (PCA) of the predictor set, and then uses the estimated factors as predictors in (3.1). The factors are obtained from the minimization problem

$$\underset{\mathbf{F},\mathbf{\Lambda}}{min}\ V(\mathbf{\Lambda},\mathbf{F}) \quad \text{s.t.} \quad N^{-1}\mathbf{\Lambda}'\mathbf{\Lambda} = \mathbf{I} \quad \text{and} \quad \mathbf{\Sigma_F} \text{ diagonal}, \tag{C.6}$$

where $\mathbf{F}$ and $\mathbf{\Lambda}$ contain the factors and factor loadings, respectively, and $V(\mathbf{\Lambda},\mathbf{F}) = \frac{1}{NT}\sum_{t=1}^{T}(\mathbf{X}_t - \mathbf{\Lambda}\mathbf{F}_t)'(\mathbf{X}_t - \mathbf{\Lambda}\mathbf{F}_t)$. Thus, as the number of factors is much smaller than the number of predictors in $\mathbf{X}_t$, the usage of the factor-augmented approach offers substantial dimension reduction and facilitates estimating (3.1) using OLS.

Appendix D Additional results

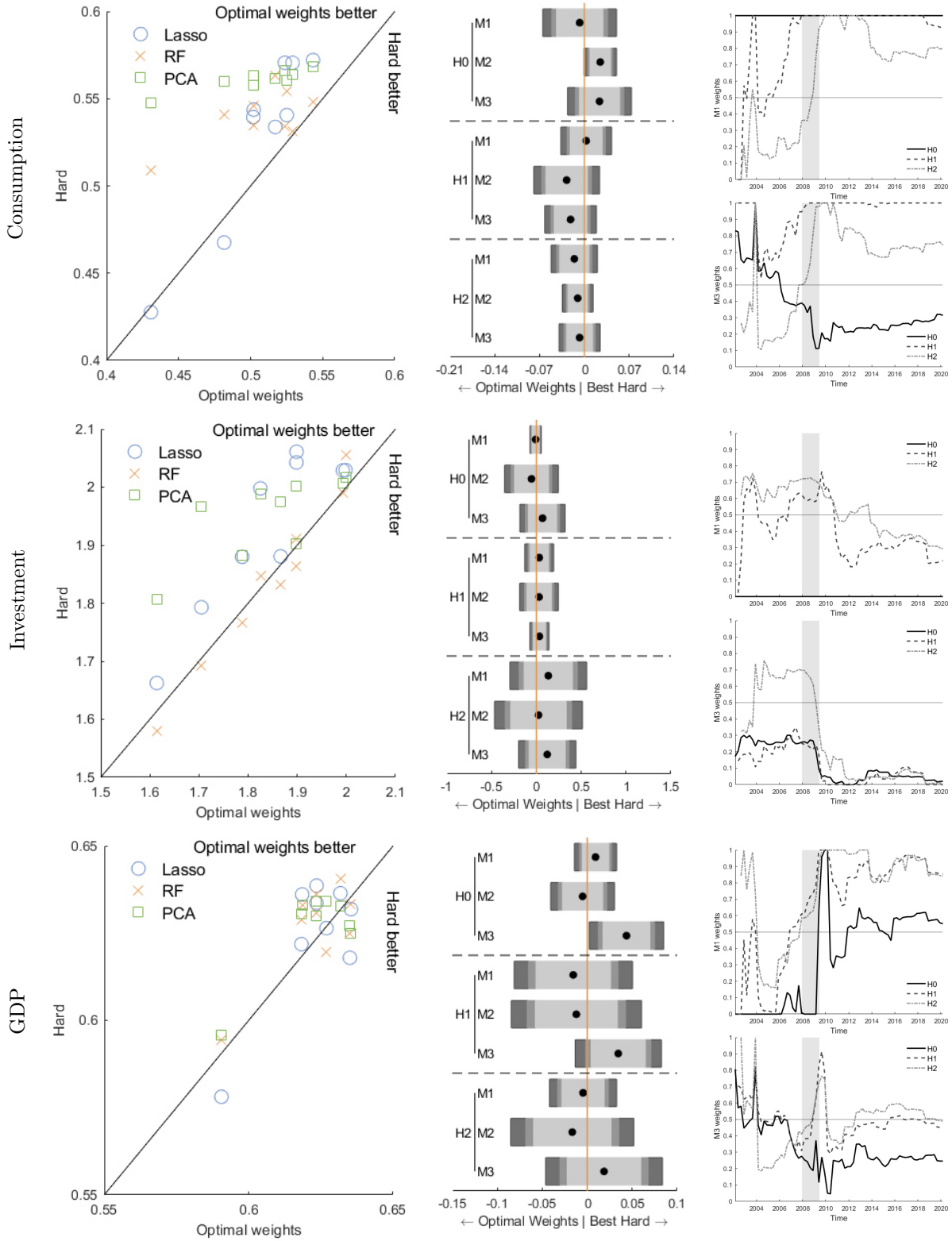


Figure D.1. Optimal combination, root mean squared errors and weights. The evaluation sample is 2002Q1-2020Q1. In column two of the figure the optimal combination and the best performing hard-based models are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). Color shadings illustrate 99%, 95% and 90% confidence bands.

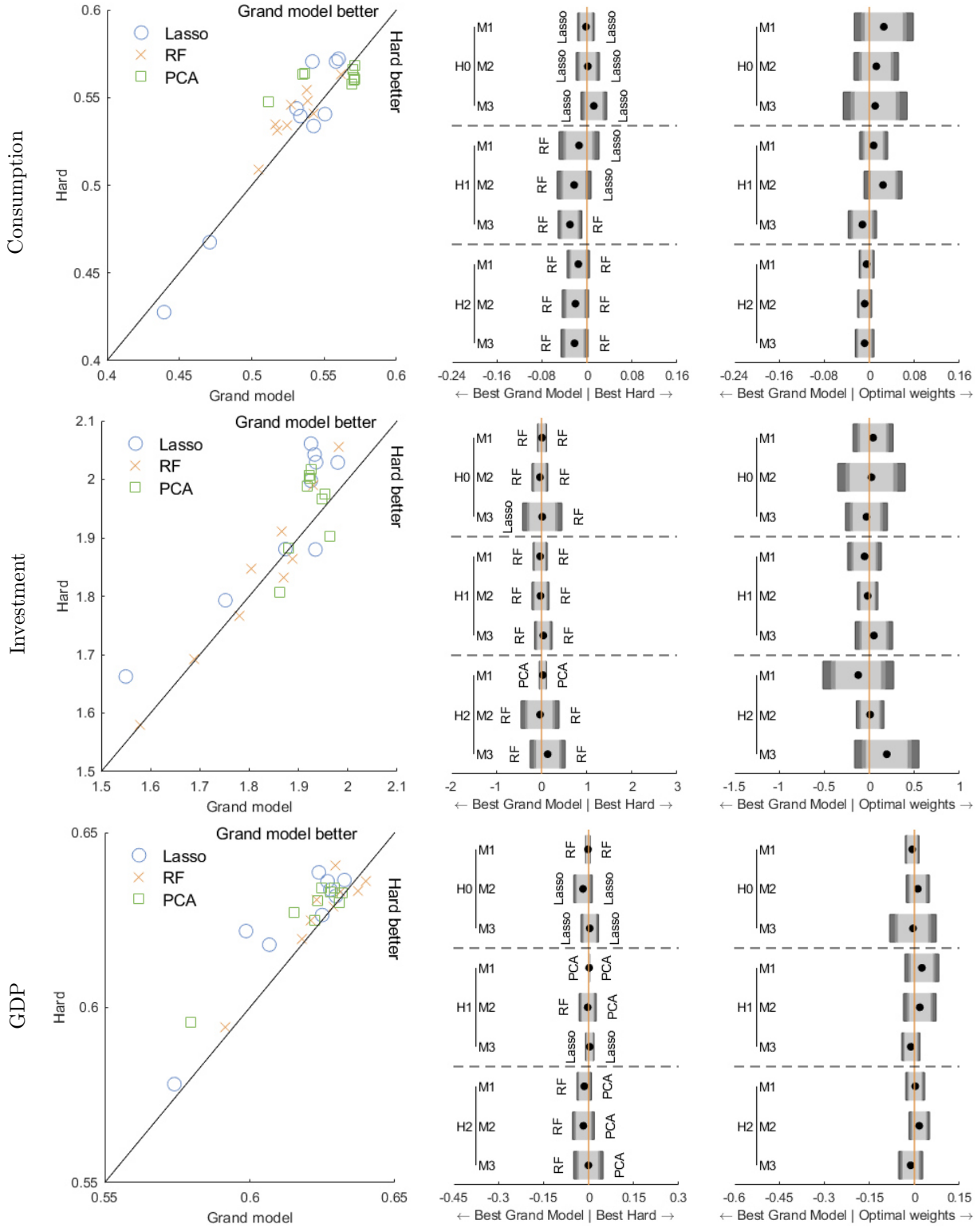
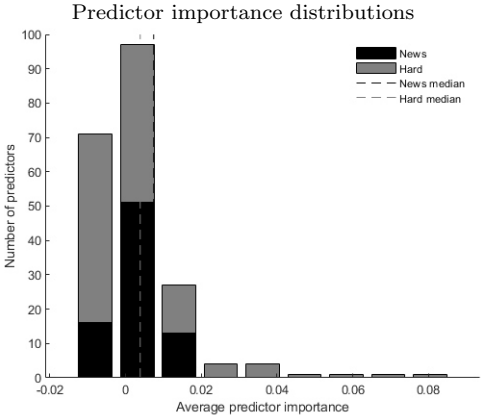


Figure D.2. Variable combination and root mean squared errors. The evaluation sample is 2002Q1-2020Q1. In column two of the figure the best variable combination and the best performing hard-based models are compared across forecasting horizons and months. In column three of the figure the best variable combination and the optimal combination model is compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). Color shadings illustrate 99%, 95% and 90% confidence bands.

Investment and top 10 predictors

Rnk	Type	Group	Name
1	Hard	Labor market	All Employees: Goods-Producing Industries
2	Hard	Labor market	Avg Weekly Hours : Manufacturing
3	Hard	Labor market	All Employees: Durable goods
4	Hard	Labor market	All Employees: Wholesale Trade
5	Hard	Labor market	Avg Weekly Hours : Goods-Producing
6	Hard	Labor market	All Employees: Total nonfarm
7	Hard	Labor market	All Employees: Manufacturing
8	Hard	Int. and exch.	3-Month Commercial Paper Minus FEDFUNDS
9	Hard	Retail and cons.	Total Business Inventories
10	Hard	Output and inc.	IP: Durable Materials



GDP and top 10 predictors

Rnk	Type	Group	Name
1	Hard	Labor market	All Employees: Retail Trade
2	Hard	Labor market	All Employees: Total nonfarm
3	News	Bond market	Topic49: Bond market
4	Hard	Output and inc.	IP: Durable Materials
5	Hard	Labor market	All Employees: Service-Providing Industries
6	News	Personal finance	Topic0: Personal finance & Over
7	Hard	Labor market	All Employees: Goods-Producing Industries
8	Hard	Labor market	All Employees: Financial Activities
9	Hard	Labor market	All Employees: Construction
10	Hard	Labor market	Help-Wanted Index for United States

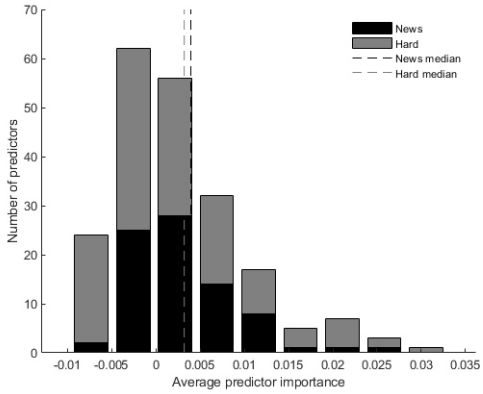


Figure D.3. GM-RF and predictor importance for *Investment* and *GDP*. The table reports the top 10 most important predictors on average across the sample. The histogram reports the empirical distribution of the average predictor importance statistics.

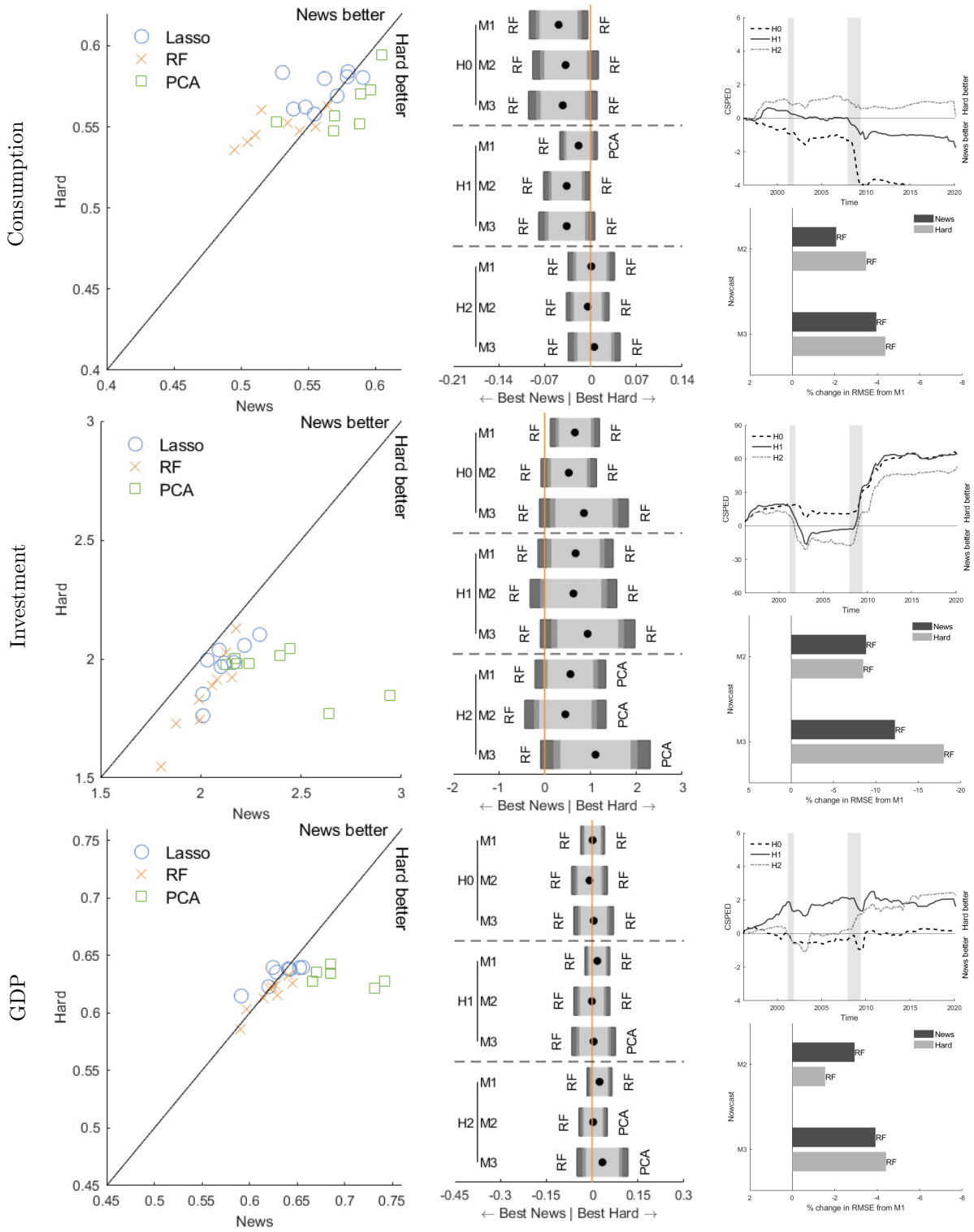


Figure D.4. Rolling estimation window. Root mean squared errors, cumulative squared prediction error differences and nowcasting. The evaluation sample is 1996Q1-2020Q1. In columns two and three of the figure the best performing news- and hard-based models are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and 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.

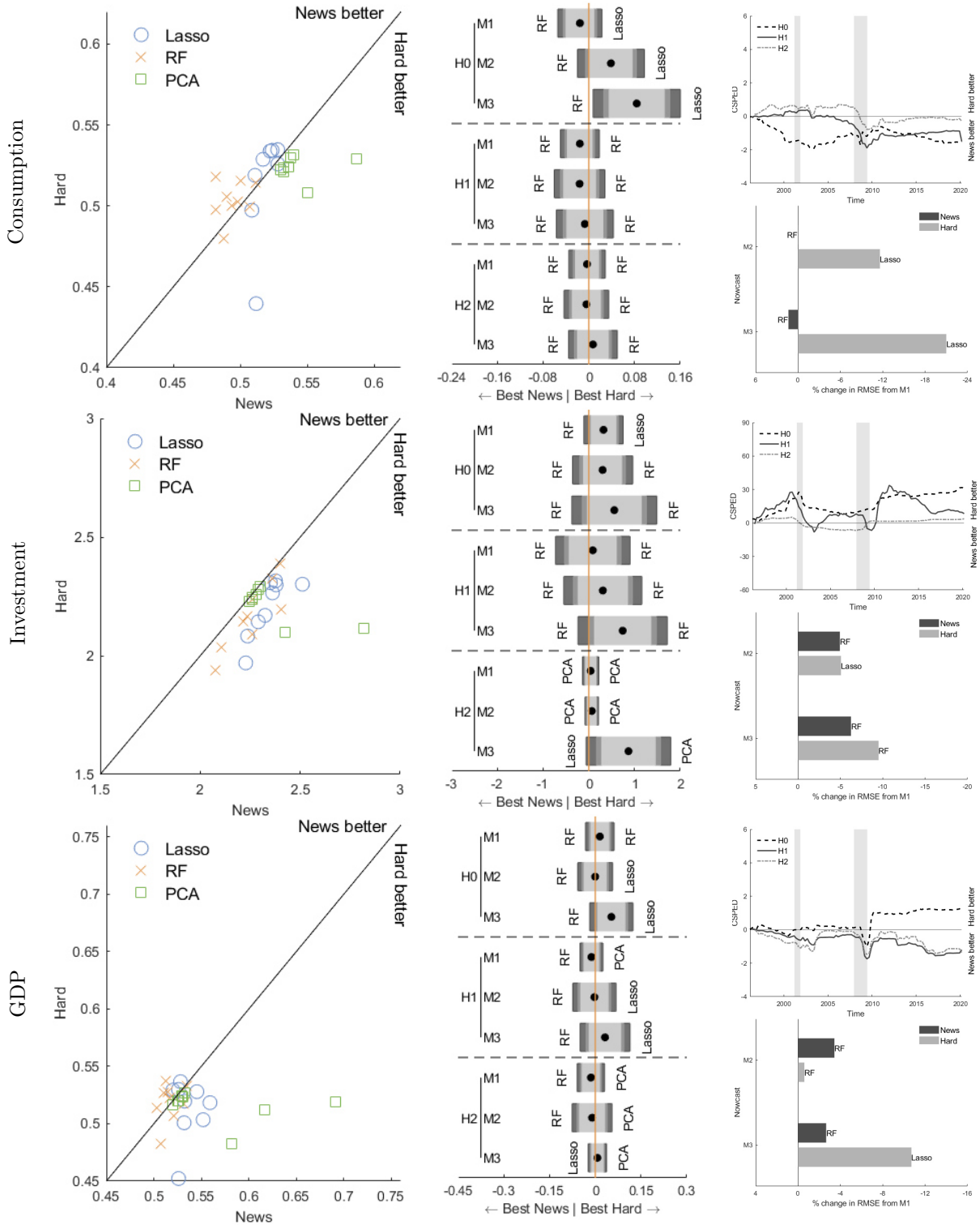


Figure D.5. First release. Root mean squared errors, cumulative squared prediction error differences and nowcasting. The evaluation sample is 1996Q1-2020Q1. In columns two and three of the figure the best performing news- and hard-based models are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and 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.

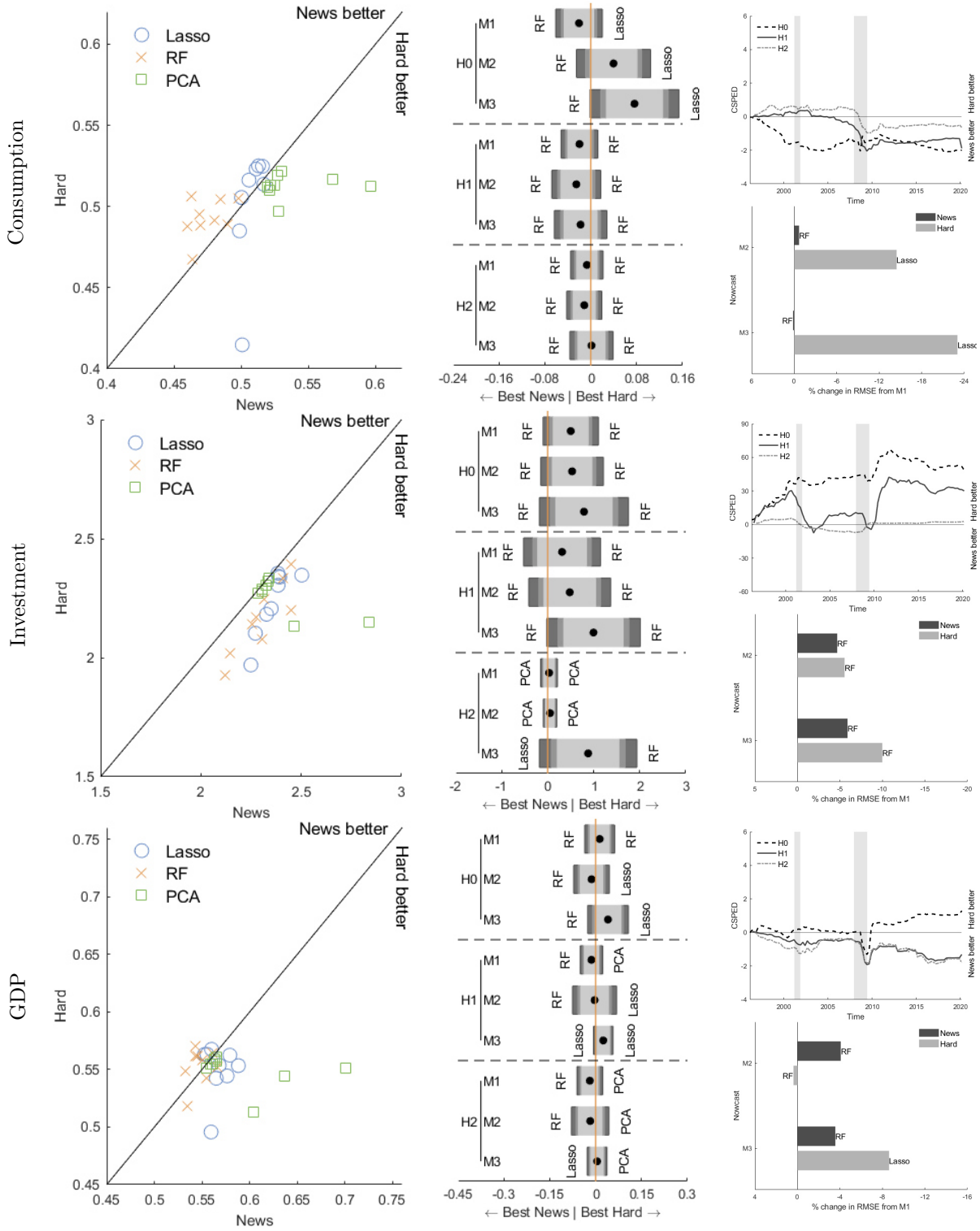


Figure D.6. Second release. Root mean squared errors, cumulative squared prediction error differences and nowcasting. The evaluation sample is 1996Q1-2020Q1. In columns two and three of the figure the best performing news- and hard-based models are compared across forecasting horizons and months. The bar plot reports differences in forecasting performance calculated using the Diebold-Mariano test (Diebold and 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.

Table D.1. Relative RMSE scores. The best hard-based models are compared to an auto-regressive model (AR), a constant growth rate model (RW) and the 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 hard-based model has the lowest RMSE. Significant differences in forecasting performance (marked in gray) are calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). *, **, and *** denote the 10%, 5%, and 1% significance level, respectively.

		H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
Consumption	AR	0.94	0.85**	0.77***	0.95*	0.98	0.96	0.97***	0.93***	0.92***
	RW	0.86	0.82***	0.41	0.87	0.94	0.51	0.90*	0.86	0.51
	SPF	1.29**	1.13	1.03	1.22	1.22	1.21	1.17**	1.14*	1.13*
Investment	AR	0.87	0.85	0.75**	0.90	0.93	0.84	0.99	1.01	0.95
	RW	0.70*	0.72*	0.59**	0.73	0.78	0.66*	0.72*	0.78	0.68**
	SPF	1.43**	1.32**	1.24**	1.42**	1.37**	1.31**	1.38***	1.36***	1.28**
GDP	AR	0.98	1.01	0.90**	0.99	1.02	0.96	1.00	1.00	0.98
	RW	0.80*	0.89	0.59	0.82*	0.91	0.63	0.80**	0.87	0.62
	SPF	1.39**	1.40**	1.30*	1.32**	1.32**	1.29**	1.16*	1.16*	1.15*

Table D.2. Relative RMSE scores. The optimal combination strategy is compared to an equal and inverse-MSE weighting scheme. The evaluation sample is 2002Q1-2020Q1. A value less than 1 indicates that the optimal combination strategy has the lowest RMSE. Significant differences in forecasting performance (marked in gray) are calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). *, **, and *** denote the 10%, 5%, and 1% significance level, respectively.

		H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
Consumption	Equal weights	0.94**	0.92*	0.88**	0.95**	0.93**	0.95	0.97	0.95*	0.98
	MSE weights	0.94**	0.93	0.90**	0.96**	0.93**	0.95	0.97	0.96*	0.98
Investment	Equal weights	0.98	0.94	0.94	0.99	0.97	0.94**	1.03**	1.03*	0.97
	MSE weights	0.98	0.95	0.95	0.99	0.97	0.95**	1.03**	1.03*	0.99
GDP	Equal weights	1.01	1.01	1.01	0.99	1.00	1.02	1.01	1.00	1.01
	MSE weights	1.01	1.01	1.01	0.99	1.00	1.02	1.01	1.00	1.01

Table D.3. Relative RMSE scores. The best news model with frequencies is compared to the best news model with only tone and tone interacted with frequencies. The evaluation sample is 1996Q1-2020Q1. A value less than 1 indicates that the model with frequencies has the lowest RMSE. Significant differences in forecasting performance (marked in gray) are calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). *, **, and *** denote the 10%, 5%, and 1% significance level, respectively.

		H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
Consumption	Tone	0.90***	0.90***	0.90**	0.91***	0.90***	0.89***	0.96*	0.93**	0.93**
	Tone adjusted	0.94***	0.95**	0.95*	0.95**	0.94***	0.93***	0.98	0.96*	0.96
Investment	Tone	1.01	0.99	0.99	1.00	1.01	1.01	1.00	1.00	1.01
	Tone adjusted	1.00	1.03	1.04	1.01	1.03	1.02	1.00	1.00	1.01
GDP	Tone	0.98	0.96	0.97	0.98	0.99	0.99	0.99	1.01	1.01
	Tone adjusted	1.00	1.00	1.00	0.98	0.98	0.99	0.98	1.00	1.02

Table D.4. Relative RMSE scores. The best news model with 80 topics is compared to the best news model with 40 and 120 topics. The evaluation sample is 1996Q1-2020Q1. A value less than 1 indicates that the model with 80 topics has the lowest RMSE. Significant differences in forecasting performance (marked in gray) are calculated using the Diebold-Mariano test (Diebold and Mariano (1995)). *, **, and *** denote the 10%, 5%, and 1% significance level, respectively.

		H0			H1			H2		
		M1	M2	M3	M1	M2	M3	M1	M2	M3
Consumption	40 topics	0.96**	0.96**	0.99	0.98	0.96	0.96	0.99	0.96*	0.99
	120 topics	0.96**	0.95**	0.96	0.99	0.99	0.96	1.00	0.97*	0.98
Investment	40 topics	1.00	1.00	1.00	1.03	1.03	1.06**	1.00	1.01	1.03
	120 topics	1.00	1.02	1.02	1.01	1.01	1.04**	0.99	1.00	1.05
GDP	40 topics	1.02	1.02	1.02	0.99	1.00	0.99	1.00	1.01	1.01
	120 topics	1.00	0.99	1.00	1.01	1.00	1.00	1.00	1.01	1.00