Identifying the Predictive Power of FED Minutes

Luiz Renato Lima[*†] Lucas Lúcio Godeiro[*] Mohammed Mohsin[*]

November 20, 2018

Abstract

This paper proposes a novel method to extract the most predictive information from FED minutes. Instead of considering a dictionary (set of words) with a fixed content, we construct a dictionary whose content is allowed to change over time. Specifically, we utilize machine learning to identify the most predictive words (the most predictive content) of a given minute and use them to derive new predictors. We show that the new predictors improve real time forecasts of output growth by a statistically significant margin, suggesting that the combination of supervised machine learning and text regression can be interpreted as a powerful device for out-of-sample macroeconomic forecasting.

Keywords: Text Regression ; Supervised Machine Learning; Elastic Net; Central Bank Communication; Forecasting, Real Time.

JEL Classification codes: C53, C55, E37, E47.

[*]Department of Economics, The University of Tennessee, Knoxville.
[†]Corresponding author e-mail: llima@utk.edu

1 Introduction

According to Gentzkow et al. (2017), the information encoded in text is a rich complement to the more structured kinds of data traditionally used in empirical research. Indeed, in recent years, we have seen an intense use of textual data in different areas of research. The idea consist of transforming strings into numeric variables, and then use it as predictors in different models. Several studies have already explored this additional source of information. In the finance literature, Garcia (2013) studies the effect of sentiment on asset prices during the 20th century (1905 to 2005). The author uses texts from the New York Times to construct sentiment variables using textual analysis. He concludes that the predictability of stock returns using news content is particularly strong during recessions. Engelberg and Parsons (2011) compare the behavior of investors with access to different media coverage of the same information event. They focus on the relation between local media coverage and local stock portfolio trading volume and find that local media coverage is a strong predictor of local trading returns.

In the the economics literature, Dossani (2018) analyzes how the tone of central Bank press conferences impacts risk premia in the currency market. He measures the tone as the difference between the number of hawkish and dovish phrases made during a press conference. He used four currency future contracts traded on the Chicago Mercantile Exchange (CME) and found that implied risk aversion increases when Central Banks are hawkish and decreases when Central Banks are dovish. Other examples where textual data is used for macroeconomic analysis includes Armesto et al. (2009), Boukus and Rosenberg (2006), Cecchetti et al. (2003), among others.[1]

A common assumption among the above studies is that the content of the dictionary (set of words) used to construct sentiment/information indexes is constant over time and subjectively chosen by the user. In practice, this implies that, given a fixed dictionary, we compute the frequency that each word (or combination of words) appears on texts and use this information to construct predictors (sentiment/information indexes) that can be used for forecasting or impulse-response analysis. The assumption of time-invariant dictionary is particularly difficult to hold in documents that introduce new words over time or if the vocabulary used in periods of recession differs from the one used in periods of economic expansions. Even if the vocabulary were the same across time, the predictive power of some words can vary, but the existing literature does not account for such an effect and therefore the resulting predictors do not reflect the most

[1] Wright (2012) and Altavilla and Giannone (2017) study the effects of news about the monetary policy on the yield curve, but they do not rely on text mining. They find that news affects market agents' expectations about corporate and Treasury bond yields.

predictive textual information encountered in documents at a given time.

In this paper, we aim at allowing the content of dictionaries to vary over time, making it entirely determined by the predictive power of its words. The goal is, therefore, to maximize the predictive power of the dictionary. We also propose new predictors that rely on time-varying dictionaries and show that they have more predictive power than the ones constructed from time-invariant dictionaries. Our methodology can be summarized by three steps: (i) In the first step, a vector of time series X_t is created, where each element shows time series observations of the frequency in that each word (or combination of words) appears on the FED minute up to time t. Thus, this step transforms words into numerical values without using a pre-specified (fixed) dictionary. This numerical representation is high dimensional and sparse, so dimensionality reduction must be employed in the next step; (ii) in the second step, we use supervised machine learning (SML) to select the most predictive time series (words) $X_t^* \subset X_t$. This is the numerical representation of our most predictive dictionary which is then used to construct a predictor (or set of predictors), $D_t = g(X_t^*)$. For instance, the function $g(.)$ can select only the k^{th} most predictive words; it can also represent common factors (computed by principal components) of words in X_t^*; or we can even let $g(.)$ represent an index of sentiment based on the most predictive dictionary X_t^*. In section 3, several possibilities for $g(.)$ are considered. Finally, in the third step, we use the predictor(s) D_t to make out-of-sample forecasts of our target variable. This 3-step procedure is repeated recursively towards the end of the sample (recursive out-of-sample forecasting), implying that the content of the most predictive dictionary changes over time.[2]

Our paper is related, in motivation, to recent works by Thorsrud (2018) and Hansen et al. (2017) who used Latent Dirichlet Allocation (LDA) to address limitations imposed by pre-specified dictionaries. The main difference between our approach and the one based on LDA is that the latter is an unsupervised machine learning (UML) technique whereas our approach relies on supervised machine learning. As explained in Chakraborty and Joseph (2017), the difference between the two approaches is the existence of a target variable y. Supervised learning is the classical case of modelling y using inputs (predictors) x in that we choose a learning algorithm that will fit the target using the given input features. In the unsupervised learning, there is no target y and the algorithms merely aim to find structure in the data, for example by grouping observations, or grouping words in topics as done by the LDA technique. Recursive estimation is a key component to generate time-varying dictionaries. Indeed, our methodology consists of

[2]In SML, model validation is essential to guarantee model's out-of-sample performance. We discuss model validation in detail in Section 3.

choosing recursively a set of words that minimizes a strictly convex loss function, implying that only the most predictive words will be included in the dictionary at each time t. As shown by Thorsrud (2018), the LDA approach is very compute-intensive and updating recursively the training sample to re-estimate the topic model is unfeasible.[3] Our methodology, on the other hand, allows for low-cost recursive updating of the training sample, making re-estimation of the forecasting model computationally feasible.[4]

The methodology proposed in this paper has several innovations. Firstly, the words (numerical representation) contained in X_t^*, rather than being chosen by the user, are selected based on their predictive power. This is carried out by applying supervised machine learning in the second step to select the most predictive words. Second, contrary to the existing methodology where the content of the dictionary is fixed over time implying that the time variation in the predictor D_t is solely explained by the frequency that each pre-specified word appears on a document, the new methodology also allows for time variation in the content of the dictionary. We show that this additional innovation has potential to improve out-of-sample macroeconomic forecasting substantially. In this paper, we use FED minute based predictors to forecast real time output growth.[5] Central banks rely on output forecasts to make decisions about changes in the monetary policies whereas the private sector uses output growth forecasts to make decision about investment, marketing and risk management. A large pool of output growth forecasting models has been suggested recently by the literature, which includes simple linear autoregressive (AR) models as well as judgmental forecasts and more sophisticated methods based on the dynamic stochastic general equilibrium (DSGE) model.

To be clear, the main methodological contribution of this paper is not proposing an ultimate forecasting model for output growth but rather combining new developments from text regression and supervised machine learning in a manner that is novel for identifying the most predictive information from FED minutes. In order to illustrate the predictive power of FED minutes, we choose a simple linear *AR(2)* to represent the benchmark forecasting model. Our choice is motivated by the findings of Chauvet and Potter (2013) that compared the performance of many econometrics models used to forecast real-time US output growth from 1992 to 2010 and concluded that a simple *AR(2)* model has the best forecasting performance during expansion periods and performs relatively well during recessions. Their findings support early evidence of Nelson (1972) that output growth forecasts from simple autoregressive models

[3]This restriction led Thorsrud (2018) to hold the training sample constant over time.

[4]Another caveat with re-estimating the LDA recursively is the lack of identifiability, that is, topic estimates cannot be combined across samples for an analysis that relies on the content of specific topics (Thorsrud, 2018, p. 22).

[5]Using real time rather than revised GDP data implies that we are considering solely the information that was available at the time the forecast was being made. Thus, we are reproducing the forecasting problem in real time.

are hard to beat. For this reason, we choose a simple *AR(2)* model as the strongest benchmark to be outperformed by the proposed method. Our results indicate that the textual data found on FED minutes contains non-trivial predictive power but fixed dictionaries are not the best approach to identify such an information. Indeed, when we add fixed-dictionary based predictors to the benchmark *AR(2)*, we found no improvement in the real time forecasts, suggesting that predictors computed from a fixed dictionary does not add predictive power to the benchmark model. However, when such predictors are computed from time-varying dictionaries, we found a strong accuracy gain relative to the nested *AR(2)* model. We also report that the proposed FED minute based forecasts outperform the Green Book forecasts suggesting that the information generated during FOMC meetings does contain non-trivial predictive power.

The choice of the SML (shrinkage) method also matters. We show that time-varying dictionaries that rely on elastic net to select the most predictive words have more predictive power than the ones that rely on LASSO or ridge regression. An explanation to this result is that ridge regression does not perform model selection (it does not shrink coefficients to zero) and, unlike *LASSO*, the number of predictors (words) selected by elastic net is not bounded by the sample size, T. Moreover, elastic net is also robust to group effects, which is particularly important to our analysis because many words appearing on the FED minutes, such as inflation and unemployment, are highly correlated.[6] Thus, elastic net not only selects the most predictive words, but also guarantees that equally predictive words will not be randomly thrown out just because they are highly correlated with other predictive words. In the forecasting literature, a short list of papers that successfully employed elastic net includes Bai and Ng (2008), Li et al. (2015), Li (2015) and Lima et al. (2018). Our results suggest that macroeconomic prediction could be improved through the combination of textual data and supervised machine learning in a way that only the most predictive words would be considered for forecasting and impulse response analysis. It has significant implication for those interested in evaluating the effects of monetary policy on the economy. Identifying monetary policy shocks remains to be an important challenge due to inconclusive evidences reported by leading researchers. See Christiano et al. (1999) and Ramey (2016) for detailed survey. The text based predictors identified in this study could potentially capture true monetary policy shocks to evaluate the effects of monetary policy on output.

The remainder of this paper is organized as follows. Section 2 introduces the channel through which information from FED minutes affect future values of output growth. Section 3 introduces text regression

[6] words are positively (negatively) correlated when the number of times they appear in a document are positively (negatively) correlated across time.

and explains how we use machine learning to identify a time-varying dictionary. Section 4 is used to define the predictors we are going to use in our forecasting exercise. Section 5 introduces our empirical analysis, including a full description of the dataset, the forecasting models and the methods used to evaluate the out-of-sample forecasts. Section 6 presents the main results followed by concluding remarks in section 7.

2 Central Bank's Actions, Term Structure and Output Growth

In this section we explain the channel through which the information contained in FED minutes affect economic growth. Identifying useful indicators to predict future economic activities is an important quest for policy makers and various investors. The main rationale explored in this study is that central bank's tone regarding the current and expected future monetary policies (as demonstrate in FED meetings and announcements) has significant explanatory power in forecasting future output growth.

Before we highlight the underlying transmission mechanism, one needs to recall the main objectives of the central bank. Through monetary policy actions, the FED wishes to achieve sustainable economic growth and price stability. The effectiveness of various monetary policy actions ultimately depends on the Fed's ability to successfully control private sector's expectations about future interest rates and inflation. The long term expected real interest rate is the single most relevant variable for the firms in their investment decisions. This is also true for the households in their spending decisions on durable consumption. So, for the FED to control the long term real interest rate, it should be able to control long term nominal interest rate or long term expectations of inflation rate or both. Through monetary policy announcements in terms of intensity and directions (tone), the central bank is able to control both the short term and long term interest rates (yield spread or term structure) that ultimately affects economic activities.

Many empirical researches in the past have identified the slope of the yield spread as a good predictor of output growth. The yield spread measures the gap between long and short-term interest rates. Under efficient market conditions, the spread gap is expected to contain private sector's expectations regarding future economic conditions as well as future government policies. Usually, long-term bonds have higher yields than short-term bonds indicating an upward sloping yield curve. Similarly, an inverted or flat yield curve indicates that current short term interest rates to be higher than long term rates. A positively sloped yield curve with larger spread tends to predict faster output growth in the future. An inverted or flat yield curve, on the other hand, raises concerns because it indicates an economic downturn and signals

an economic recession. Important studies that tested this channel to forecast economic growth (mostly economic recessions) include Ahrens (2002), Ang et al. (2006), Bernard and Gerlach (1998), Duarte et al. (2005), Estrella (2005), Estrella and Hardouvelis (1991), Estrella and Mishkin (1998), Plosser and Rouwenhorst (1994), Stock and Watson (1989), and others.

It is clear that the central bank conducts monetary policy to affect short term nominal interest rates, inflation rate and the long term real interest rates to achieve their objectives. The effects of conventional monetary policy shocks (i.e. the growth rate of money supply or inflation targeting) on the term structure of interest rates is well known. See Mansoorian and Mohsin (2004) for details. In this study, we are interested in evaluating the effects of the tone of the monetary authority. Often the FED communicates its desired future policy changes through FOMC meetings and press conferences. Since early 2000s the central bank began using this forward guidance as an effective policy tool. Through this policy tool the Fed expects to communicate with the private sector regarding their future course of actions. With strong credibility, the central bank could effectively and successfully influence the private sector's expectation about future (Kydland and Prescott, 1977) and, therefore, affect the future values of output growth. Following the great recession of 2007-2008 when federal funds rates remain close to zero, the FED's communications through their monthly policy statements proved to be very effective in managing expectations. Effectively, the FED communications control the term structure of real interest rates in the economy. Impulse response analysis provided by Lucca and Moench (2015), and by ourselves at the end of this paper, demonstrated that FED's tone in fact significantly affect yield spreads.

The above discussion suggests that FED's tone can predict output growth through changes in the yield spread. In this paper, we propose using a time-varying dictionary approach to identify the tone of the central bank by correctly selecting the most predictive words from FED minutes. We will construct new predictors out of the time-varying dictionary and will use them to forecast the US output growth. The main idea of this new approach is that one can maximize the predictive power of the dictionary by allowing its content to change over time. In the next section, we briefly explain how text regression and machine learning can be combined to construct time-varying dictionaries.

3 Text Regression

Text regression is a branch of econometrics which consists in transforming string into numerical variables. As documented in Gentzkow et al. (2017), there has been an increasing interest in the information encoded

in text as a rich complement to the more structured kinds of data traditionally used in research. In this section we will explain how textual data enters our database as well as how regularization methods, such as elastic net, can be used to compute a time-varying dictionary.

3.1 Textual Data.

In this section, we explain how textual data enters the database. Our initial goal is to transform text into numerical data. In order to perform this task we first save all downloaded minutes into what the literature calls "a corpus", which is a collection of written texts, i.e., a set of FED minutes. Prior to perform any word counting, we preprocess the raw text in several steps. The purpose is to reduce the vocabulary to a set of words that are mostly meaningful. Following Hansen et al. (2017), we first identify collocations or sequence of words that have a specific meaning. For example, "federal fund", "financial market", "labor market" correspond to a single economic concept but it is composed of two separate words. Thus, we identify collocations by using the taggers proposed by Toutanova et al. (2003) and create a single term for two-word sequences whose frequency is above 100. We also removed punctuation, stop words, etc. This cleaning process was carried out through the following commands in the programming language R:

tolower: Since R is case sensitive, we used the command "tolower" to classify words with lower and upper case letters (such as house and House) as equal words.

removePunctuation: This command was used to remove punctuation.

removenumbers: We used this command to remove numbers.

stopwords: We used this command to remove stop words such as the, that, which, what, etc.

stripWhitespace: This command is used to collapse multiple space so that "economy " and "economy" are treated as having the same meaning.

stemming: This function is used to guarantee that our database will include only "stems". For instance, words as "Economy" and "Economics" are counted as "Econom". If both appear in a document, their sum is showed. Otherwise, a single word is counted. In this paper we consider only stemmed words. Notice that stemming identifies the linguistic root of a term which means that the outcome of stemming is not necessarily an English word.

Finally, as in Hansen et al. (2017) , we follow the idea of Blei and Lafferty (2009) and rank the remaining words using their term frequency-inverse document frequency (tf-idf), which is a measure that punishes words that are both rare and too frequent. We dropped all terms ranked 1040 or lower, which

still left us with a large amount of words per time period.

In our output growth forecasting study we divide total sample of $T = R + P$ observations into in-sample and out-of-sample portions. The in-sample observations span 1 to R, whereas the out-of-sample observations span $R + 1$ through T for a total of P h-step-ahead out-of-sample forecasts. There will be a corpus for each quarter s, $s = 1, ..., t - h$ with $t = R, ..., T - h$ where $h \geq 0$ is the forecast (nowcast) horizon.[7] After the text preprocessing described in the above paragraph, we make a word counting on all minutes available at quarter s, resulting into new time series $X_{j,t} = \{X_{j,s}\}_{s=1}^{t}$ for each word $j = 1, ..., p$. Thus, our word counting generates new time series that contains observations up to time (forecast origin) t, with $t = R, ..., T - h$.

In order to avoid problems related to the estimation of common factors by principal components, we normalize the new time series by using their respective historical average and standard deviation. Specifically, we calculate $(X_{j,t} - \mu_j) / \sigma_j$ where μ_j and σ_j are the historical mean and standard deviation respectively of the word j computed using observations up to time t. Thus, before making a new forecast, we count words (or combination of words) without imposing any pre-specified dictionary and normalize the resulting time series.[8] Figure (1) exhibits a time series corresponding to a normalized counting of the collocation (two-word term) "economic activity". The time series ranges from 1958.1 to 2017.2 and the gray bars indicate periods of economic recession. It is interesting to notice that the term "economic activity" was largely employed during FOMC meetings that occurred in recessions periods (especially in 1975 and 2008). This suggests that such a two-term word may be especially useful to forecasting output growth during recessions. In our empirical section, we show that this term is indeed selected as an important predictor of output growth during recession times.

{ Place Figure 1 here. }

Our final step is to collect all normalized time series into a $p \times 1$ vector $X_t = (X_{1,t}, ..., X_{p,t})'$ where the dimension of the vector X_t reflects the number of terms appearing on the minutes, i.e., $dim\,(X_t) = p$. In our forecasting exercise, the dimension (p) of the vector X_t will be quite large and it can even be larger than the sample size, i.e., $p >> T$. This creates a high-dimensional problem in the sense that the number of time series (predictors of output growth) can be much larger than sample size, implying that we cannot

[7] A corpus at quarter s will include all FED minutes from that quarter.

[8] Normalization implies that if a word does not appear in the minutes during quarter s, then it will receive a value $(-\mu_j) / \sigma_j$. Notice, however, that our preprocessing of raw texts use the term frequency-inverse document frequency $(tf - idf)$ to remove from our sample words that are rare. This avoids the occurrence of observations that are almost always equal to $(-\mu_j) / \sigma_j$.

estimate any prediction equation without first performing a model selection. In this paper, we address this model selection problem by using regularization methods such as the elastic net. Since each time series represents a term (word), our recursive model selection procedure will end up selecting a time-varying dictionary (i.e., a vector of time series that best predicts output growth).

3.2 Constructing Time-Varying Dictionaries

In this section we want to construct a time-varying dictionary, that is, a vector $X_t^* \subset X_t$ that contains only the most predictive words from the FED minutes. We implement this step by applying elastic net to the following linear prediction equation

$$y_{t+h} = W_t'\boldsymbol{\beta_h} + X_t'\phi_h + \epsilon_{t+h} \tag{1}$$

where $h \geq 0$ is the forecasting horizon and $\widehat{\beta}_h$ and $\widehat{\phi}_h$ are estimated by minimizing the following objective function

$$\min_{\boldsymbol{\beta_h},\boldsymbol{\phi_h}} \sum_t (y_{t+h} - W_t'\boldsymbol{\beta_h} - X_t'\boldsymbol{\phi_h})^2 + \lambda_1 \parallel \boldsymbol{\phi_h} \parallel_{\ell_1} + \lambda_2 \parallel \boldsymbol{\phi_h} \parallel_{\ell_2} \tag{2}$$

where W_t is a $k \times 1$ vector of pre-determined predictors, such as lags of y_t as well as traditional predictors from structured data; X_t is the $p \times 1$ vector defined previously, and $\|\cdot\|_{\ell_1}$ and $\|\cdot\|_{\ell_2}$ are the ℓ_1 and ℓ_2 norm, respectively. Equation (2) is estimated recursively toward the end of the sample, that is, we regress the observations y_{s+h} on the predictors W_s and X_s for $s = 1, ..., t-h$ and $t = R, ..., T-h$[9].

In statistics literature, a combination of ℓ_1 norm constraint and ℓ_2 norm constraint is known as "elastic-net" (Zou and Hastie, 2005). Since we are not considering a fixed dictionary, the dimension of X_t will be very large, i.e. $p >> T$, and for this reason ϕ_h will be the only coefficients penalized in (2). The choice of the most predictive words depend on the values of the tuning parameters λ_1 and λ_2. In this paper, we estimate the tuning parameters by using the the cross-validation procedure suggested in the GLMNET R package (linear regression section) developed by Trevor Hastie and Junyang Qian.[10] This cross-validation procedure performs an ex ante selection of the tuning parameters which is essential to avoid overfitting. In particular, we employ cross-validations for dependent data as in Elliott and Timmermann (2013) and (Diebold and Shin, 2018, p.15). When $\lambda_1 = \lambda_2 = 0$ then the objective function (2) becomes equal to

[9] We end at $T-h$ because we need to use observation T to evaluate forecasts made at $T-h$

[10] The link can be found here GLMNET.

the usual sum of the squared residuals. When $\lambda_2 = 0$, the problem is the ℓ_1-norm constrained one (the so called LASSO estimator) whereas $\lambda_1 = 0$ corresponds to ridge regression. Ridge regression does not do model selections because it does not shrink coefficients toward zero.

Elastic net corresponds to the case in that both λ_1 and λ_2 are positive. In this case, the coefficients $\widehat{\phi}_h$ are shrunk toward zero in two different ways, promoting both sparsity and stability. In other words, the elastic net is an estimator that shrinks all parameter estimates towards zero, while penalizing the smaller and larger parameters more. It is important to note that this is achieved in a way that reduces the mean-squared-error (MSE) of both parameter estimator and forecasting model, suggesting that we can use elastic net to select the most predictive time series (words) X_t^*. Specifically, the time series included in X_t^* are the ones whose estimated coefficients ϕ_h in Equation (2) are different from zero. Since estimation takes place recursively over $t = R, ..., T - h$, different time series (words) $X_t^* \subset X_t$ will be selected over time. This is what we call time-varying dictionary. It is important to notice that the content of the time-varying dictionary also depends on the forecast horizon h, but we omit this extra notation for ease of exposition. In practice, this approach allows the econometrician to take advantage of using the most predictive dictionary (X_t^*) at each forecast origin t to make h-step-ahead forecasts.

Although *LASSO* is successful at variable selection, in the particular case in that the number of predictors is larger than the sample size, T, as in our empirical exercise, *LASSO* selects at most T predictors before it saturates. For this reason, Zou and Hastie (2005) described the combination of ℓ_1-constraint and ℓ_2-constraint as "a stretchable fishing net that retains all the big fish." Moreover, in a regression problem, if the ith column and the jth column of X_t are highly correlated and both independent variables (words) are important, regression with only ℓ_1-norm constraint tends to assign a large estimate to one of $\phi_{i,h}$ and $\phi_{j,h}$ randomly, and set the other to zero (Efron et al., 2004; Zou and Hastie, 2005). But with an additional ℓ_2 norm constraint, regression with both constraints tends to produce similar estimates of $\phi_{i,h}$ and $\phi_{j,h}$ while maintaining sparsity. This is particularly important for our analysis because many words appearing on the FED minutes are highly correlated.[11] Thus, elastic net not only selects the most predictive words, but also guarantees that equally predictive words will not be randomly thrown out just because they are highly correlated with other predictive words.

The next step is to construct FED minute based predictors $D_t = g(X_t^*)$, where X_t^* was selected as above. We finally add the selected FED based predictor to a linear prediction equation and use it to make

[11] see footnote (6).

h-step-ahead recursive forecasts. In other words, we consider the following prediction equation.

$$y_{t+h} = W_t'\boldsymbol{\beta}_h + D_t'\phi_h + \epsilon_{t+h} \tag{3}$$

and compute direct h-step-ahead forecasts as:

$$f_{t+h,t} = W_t'\widehat{\boldsymbol{\beta}}_h + D_t'\widehat{\phi}_h \tag{4}$$

where $f_{t+h,t}$ is the h-step-ahead forecast of y_{t+h} made at time $t = R, ..., T-h$; $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{\phi}_h$ are OLS estimates. Recall that we divide the total sample of $T = R + P$ observations into in-sample and out-of-sample portions. The in-sample observations span 1 to R, whereas the out-of-sample observations span $R+1$ through T for a total of P h-step-ahead out-of-sample forecasts. Hence, for each forecast origin $t = R, ..., T-h$, we estimate the coefficients $\widehat{\boldsymbol{\beta}}_h$ and $\widehat{\phi}_h$ by regressing the observations y_{s+h} on predictors W_s and D_s for $s = 1, ..., t-h$ (recursive forecasting scheme), and compute $f_{t+h,t}$ by evaluating equation (3) at the t^{th} observation . In what follows, we describe the FED based predictors used in the empirical section of this paper.

4 FED minute Based Predictors.

We use the selected time series (words) X_t^* to construct predictors of the form $D_t = g(X_t^*)$. We consider several options.

4.1 time-varying dictionary

$(D_{1,t})$: Our first set of predictors correspond to common factors of X_t^*, which are estimated by principal components using data $X_{k,s}^*$ up to forecast origin t, where $k = 1, ..., K$ are the most predictive time series with $K << p$.[12] In other words, the principal components estimators are defined as:

$$(\Lambda_t, F_t) = \arg\min\{\lambda_k, f_s\} \frac{1}{Kt} \sum_{k=1}^{K} \sum_{s=1}^{t} \left(X_{k,t}^* - \lambda_k' f_s\right)^2 \tag{5}$$

where $\Lambda_t = (\lambda_{1,t}, ..., \lambda_{K,t})'$ is a $K \times r$ matrix of factor loadings and the corresponding r common

[12]Notice that $X_{k,s}^*$ is the kth element of the vector X_s^*.

factors are collected in the $t \times r$ matrix $F_t = (f_{1,t}, ..., f_{t,t})'$ where $f_{s,t}$ denotes the sth observation on the $r \times 1$ vector of common factors estimated using data up to time t. The double index (s, t) is on purpose to make explicit the information that the r common factors are estimated using information up to the forecast origin t. This way, we can use the observations on the r common factors available at the forecast origin t to make h-step-ahead forecasts. This approach is standard in the forecasting literature (see, for instance, (Gonçalves et al., 2017)).

We select the optimal number of factors, r, via the eigenvalue ratio approach developed by Ahn and Horenstein (2013). Then, we follow Bai and Ng (2008) to keep only the factors with p-value less than or equal to 0.01 in the prediction equation (3). For example, if the optimal number of factors suggested by the eigenvalue ratio approach is 3, then we include only the first three common factors in the predictive equation (3) and we keep them if their p-value is less than or equal to 0.01, otherwise we re-estimate the prediction equation by including only the significant common factor(s). Recall that X_t^* only includes the most predictive words and, therefore, this approach is close to the one developed by Bai and Ng (2008) which applied the same idea to structured data, i.e., they used elastic net to select the most predictive variables from a large set of covariates and then computed common factors of the selected variables. In this paper, a forecasting model that employs this set of predictors is going to be labeled as M_1.

$(D_{2,t})$: This set of predictors are computed from selected bigrams (tokens) rather than words. Bigrams are two-word combinations that are used to avoid semantic ambiguity of single words when they are taken out of context. Following Apel and Grimaldi (2012), we consider a list of nouns: (inflation, cyclical position, growth, price, wages, oil price, development, fund rate, interest rate, labor market, output growth); a list of dovish adjectives: (decreasing, decreased, slower, weaker, lower, weak,low, slow); and hawkish adjectives: (increasing, increased, faster, stronger, higher, fast, strong, high).

We combine the nouns and adjectives listed above, which results into 192 bigrams (96 hawkish and 96 dovish). Next, we collect the frequency by which these 192 bigrams appears on the FED minute at time t, giving rise to a vector, B_t, with 192 new time series. We replace X_t with B_t in equations (1) and use elastic net (equation 2) to select the most predictive bigrams B_t^*. Finally, we compute the common factors of B_t^* by using principal components. We select the optimal number of factors via the eigenvalue ratio approach developed by Ahn and Horenstein (2013) and keep only the factors with p-value less than or equal to 0.01 in the predictive equation. Notice that the main difference between $D_{1,t}$ and $D_{2,t}$ is that the former relies on selected words whereas the latter uses selected bigrams. In this paper, a forecasting model that employs this set of predictors is going to be labeled as M_2.

($D_{3,t}$): In order to compute this set of predictors, we split the same set of bigrams B_t used in $D_{2,t}$ into a positive (dovish) subset (nouns+dovish adjectives) and a negative (hawkish) one (nouns+hawkish adjectives). Then, we apply elastic net on both of them to select the most predictive positive bigrams, $B_t^{positive^*}$, and the most predictive negative bigrams, $B_t^{negative^*}$. Finally, we apply principal components to obtain a common factor for $B_t^{positive^*}$ and a common factor for $B_t^{negative^*}$. These two single common factors will be included in the predictive equation. Notice that, unlike M_1 and M_2, this approach does impose a restriction on the percentage of "hawkish" and "dovish" information loaded into the factors. In fact, factors computed using the positive bigrams are entirely loaded with "dovish" information whereas factors computed using the negative bigrams are only loaded with "hawkish" information. In this paper, a forecasting model that employs this set of predictors is going to be labeled as M_3.

4.2 fixed dictionary

($D_{4,t}$): Unlike the predictors defined previously ($D_{1,t}$ to $D_{3,t}$), the predictor we define here is based on a fixed dictionary proposed by Lucca and Trebbi (2009). They consider a list of "hawkish" and "dovish" words: **List of hawkish words** = {hawkish, tighten, hike, raise, increase, boost}; **List of dovish words** = {dovish, ease, cut, lower, decrease, loose}. We proceed by doing the word counting and then normalize the data. Then, we aggregate the dovish and hawkish bigrams frequency for time t and calculate the net index as in Apel and Grimaldi (2012):

$$D_{4,t} = \left[\left(\frac{hawk}{hawk + dove} \right) - \left(\frac{dove}{hawk + dove} \right) + 1 \right] \tag{6}$$

where hawk and dove are the number of hawkish and dovish phrases, and 1 is added to exclude negative numbers. This predictor can be interpreted as the net hawkishness of minute in quarter t. In this paper, a forecasting model that employs this predictor is going to be labeled as $M4$.

($D_{5,t}$): The predictors developed here also relies on a fixed dictionary, but we use elastic net to select the most predictive words from the fixed dictionary. In this way, this method can be interpreted as a compromise between time-varying and fixed dictionaries. In other words, we consider the same list of words as $D_{4,t}$, then we do the word counting and normalize the resulting time series leading us to two sets of time series: the set of hawkish $X_t^{hawkish}$ and X_t^{dovish}. The final step is to use elastic net to identify the most predictive words, that is, a subset of $X_t^{hawkish}$ and a subset of X_t^{dovish}, respectively. These two

subsets will be used as predictors. In this paper, a forecasting model that employs these predictors is going to be labeled as M_5.

4.3 other predictors

Although we appreciate the generality of elastic net, we also compare it with both ridge regression and LASSO, which are nested in elastic net. The idea here is to make the contribution clearer in terms of methodology. In other words, suppose we find that both ridge and LASSO outperform a fixed-dictionary forecasts; then there is value in shrinkage above and beyond the specific methodology chosen. We believe this would add contribution to the paper, which otherwise would be too dependent on the regularization technique chosen. For this reason, we consider the following two predictors.

$(D_{6,t})$: This predictor is identical to $D_{1,t}$ except that we use LASSO to select $X_t^* \in X_t$. Recall that LASSO is not robust to group effects and selects at most T predictors. In this paper, a forecasting model that employs these predictors is going to be labeled as M_6.

$(D_{7,t})$: This predictor is identical to $D_{1,t}$ except that we use ridge regression to shrink the coefficients on X_t. Recall that ridge does not do model selection because it does not shrink coefficients to zero. In this paper, a forecasting model that employs these predictors is going to be labeled as M_7.

We also consider a pure principal component (PCA) analysis on the set of the original words X_t. This exercise is useful to check if using supervised machine learning (elastic net, LASSO) to pre-select the most predictive words increase the predictive power of the resulting predictor. Thus, we consider the following predictor.

$(D_{8,t})$: This predictor is identical to $D_{1,t}$ except that we eliminate the first stage, that is, we do not use elastic net (or LASSO) to pre-select the most predictive words. Hence, the FED minute based predictors will correspond to common factors of X_t, rather than X_t^*. In this paper, a forecasting model that employs these predictors is going to be labeled as M_8.

It is important to notice that models M_i $i = 1, 6, 7$ use the same procedure to compute the final predictors, $D_{1,t}$, but they differ on the machine learning method used to select the most predictive words. While M_1 relies on elastic net, M_6 and M_7 rely on LASSO and ridge regression, respectively. Thus, in the empirical section, these models will be used to compare the ability of these methods to improve the predictive power of the time-varying dictionary.

5 Empirical Analysis

5.1 Data

We obtain quarterly real *GDP* data (Y_t) from the Federal Reserve Bank of St. Louis ranging from $1958Q1$ to $2017Q2$. Then we compute the annualized log change of *GDP* as $y_t = ln(\frac{Y_t}{Y_{t-1}}) * 400$. This is our target variable and there are 238 quarterly observations in total. The data from FED minutes are aggregated within each quarter and cover the same period $1958Q1 - 2017Q2$. Although it is possible to obtain observations of output growth and FED minutes since $1947.Q2$, the FED minutes released before $1958Q1$ were very limited in terms of communication, with an average of 15 pages per minute against an average of 60 pages after $1958Q1$. Thus, in order to maximize the likelihood of occurrence of words (or combination of words) per FED minute, we start our sample in $1958Q1$.

Given 236 observations[13] of *output* growth, we use the first 134 observations, from $1958Q3$ to $1991Q4$, as our initial estimation sample for the cases with $h = 0$ and $h = 1$. For longer forecast horizons $h = 3$ and $h = 6$, the initial estimation windows are shortened, ranging from $1958Q4 + h$ to $1992Q1 - h$. This implies that there will be exactly 102 out-of-sample forecasts for each forecast horizon h. The models are recursively estimated using only collected real time realizations of the series as released at each quarter to generate h-quarter-ahead real time forecasts. As explained by Chauvet and Potter (2013), all versions of the historical unrevised real time GDP series released each month are collected and archived by the Federal Reserve Bank of Saint Louis and the Federal Reserve Bank of Philadelphia. The quarterly real time database used in this paper consists of realizations, or quarterly vintages, of the series as they would have appeared in the end of each quarter from 1992Q1 to 2017Q2.

We downloaded the FOMC minutes from the Federal Reserve(FED) Website.[14] According to this website, the minutes of the last meeting in each quarter is released before the release of the corresponding GDP data (see Table 1 in the appendix). This is an important information because it will allow us to perform nowcasts as well as forecasts. In order words, we can use the FED minutes information available at time t to forecast output growth at time $t+h$, $h \geq 0$ where $h = 0$ corresponds to what we call nowcast. The next step is to import the FED minutes (in pdf format) to the software R Statistics. The package used to import these minutes is the "tm" package, which provides a function by which one can import pdf files

[13] Recall that we include 2 auto-regressive lags so the dependent variable starts at $1958Q3$.

[14] The minutes from 1993 to 2017 can be found in this link: https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm, and the minutes from 1936-1992 can be found in this link: https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm.

to R[15]. With the FED minutes in R, we can construct a corpus and apply the preprocessing of raw texts described in Section 3.1.

{ Place Table 1 here. }

5.2 The Forecasting Procedure

Chauvet and Potter (2013) compared the performance of many econometric models used to forecast the US output growth from 1992 to 2010. They reported that a simple *AR(2)* model outperforms more sophisticated models during periods of economic expansion and performs relatively well during recessions. Their finding supports early evidence of Nelson (1972) that forecasts from simple autoregressive models are hard to beat. Thus, based on this empirical evidence, we consider a *AR(2)* model as a benchmark to be beat, especially during periods of economic expansion. This benchmark corresponds to prediction equation (3) with $W_t = (y_t; y_{t-1})'$ and $\phi_h = 0$, where y is defined as the real time output growth and h is the forecast horizon. Thus, the h-step ahead (MSE) forecast of y_{t+h} conditional on the information available at time t, $f_{t+h,t}$, is represented by the conditional mean and can be computed as [16]

$$f_{t+h,t} = \widehat{c}_h + \widehat{\beta}_{1,h} y_t + \widehat{\beta}_{2,h} y_{t-1}$$

where t is the last in-sample observation (forecast origin) and $\left(\widehat{c}_h, \widehat{\beta}_{1,h}, \widehat{\beta}_{2,h}\right)$ are OLS estimates of y_{s+h} on predictors $y_s; y_{s-1}$ for $s = 1, ..., t-h$ with $t = R, ..., T-h$ (recursive forecasting scheme). The benchmark *AR(2)* is nested by the proposed forecasting models which are simply equation (3) with $W_t = (y_t; y_{t-1})'$ and $\phi_h \neq 0$. Thus, the h-step ahead (MSE) forecast of y_{t+h} conditional on the information available at time t is represented by the conditional mean, that is

$$f^i_{t+h,t} = \widehat{c}_h + \widehat{\beta}_{1,h} y_t + \widehat{\beta}_{2,h} y_{t-1} + \widehat{\phi}'_h D_{i,t}, \tag{7}$$

where $D_{i,t}$ $i = 1, ..., 8$ corresponds to one of the 8 sets of predictors generated in the previous section and $\left(\widehat{c}_h, \widehat{\beta}_{1,h}, \widehat{\beta}_{2,h}, \widehat{\phi}_{2,h}\right)$ are OLS estimates of y_{s+h} on predictors $(y_s, y_{s-1}, D_{i,s})$ for $s = 1, ..., t-h$ with $t = R, ..., T-h$. Thus, if the FED minutes have some predictive power, then the corresponding out-of-sample forecast should outperform the benchmark *AR(2)*.

[15] A quick tutorial can be found in the web page "Reading PDF files into R for text mining".

[16] Throughout this paper we assume that the target variable y_t is a covariance-stationary process.

If time-varying dictionaries are more predictive than fixed ones, then out-of-sample forecasts from model (7) with $D_{i,t}$ $i = 1, 2, 3, 6, 7$ should outperform the forecasts from model (7) with $D_{i,t}$ $i = 4, 5$. If elastic net does a better job selecting the words, then model (7) with $D_{i,t}$ $i = 1$ should outperform the out-of-sample forecasts from model (7) with $D_{i,t}$ $i = 6, 7$. Finally, if the pre-selection of words helps increase forecasting accuracy, then model (7) with $D_{i,t}$ $i = 1, 6, 7$ should outperform the forecasts from model (7) with $D_{i,t}$, $i = 8$.[17]. Thus, this framework is rich enough to investigate not only if the FED minutes contain strong predictive information but also if such a predictive power depends on both the usage of time-varying dictionaries and the method used to select the most predictive words.

Since the data on real time GDP is made available only after the release of the FED minute, we can use the FED minutes available at time t to nowcast the real time output growth, that is, we can compute

$$f^i_{t,t} = \widehat{c} + \widehat{\beta}_1 y_{t-1} + \widehat{\beta}_2 y_{t-2} + \widehat{\phi}' D_{i,t}, \tag{8}$$

where $\left(\widehat{c}, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\phi}'\right)$ are OLS estimates from a time series regression of y_s on y_{s-1}, y_{s-2} and $D_{i,s}$; $f^i_{t,t}$ is a nowcast of the real time output growth which depends on y_{t-1} and y_{t-2} as well as the value of the predictor(s) $D_{i,t}$ at the forecast origin t. Thus, a nowcast is a h-step ahead forecast with $h = 0$.

As Chauvet and Potter (2013) pointed out, a judgmental forecast occurs when forecasts obtained from econometric models are adjusted by their users. Some of the judgments rely on subjective information extracted from FED minutes or (and) other sources of information about the future course of the economy. Since our approach relies on information from FED minutes to improve forecasts from an *AR(2)* model, we could expect that survey forecasts, such as the Blue Chip[18], might have additional information that would improve the out-of-sample forecasts of output growth. We test such an hypothesis in this empirical section of this paper. Thus, besides the forecasting models proposed in this paper, we also included a judgmental forecast from the Blue Chip Indicators (BC). We take the BC forecasts individually but we also add it to an AR(2) model, giving rise to what we label AR(2)-BC model.[19] Finally, we consider an equal-weight forecast combination (FC) of all FED minute based forecasts, which comprises forecasts

[17]Recall that models $D_{i,t}$ $i = 1, 6, 7$ use the same procedure to compute the final predictors, they only differ on the machine learning method used to select the most predictive words.

[18]The Blue Chip Indicators is a poll of around top 50 forecast economists from banks, manufacturing industries, brokerage firms, and insurance companies. The poll has been conducted since 1976 and comprises several macro series, including GDP growth.

[19]There is a Blue Chip forecast for each forecasting horizon considered in this paper. Moreover, BC forecasts are released more than once during the same quarter. We always use the last release of the quarter, which typically occurs 10 days after the end of the quarter.

from models $M_i, i = 1, ..., 8$. Overall, our set of forecasting models includes $f^j_{t+h,t} \in (M_i,\ BC,$ AR(2)-BC, $FC)$

5.3 Forecast Evaluation

To evaluate the performance of different models in forecasting (nowcasting) output growth rates, we compute the root mean squared forecast error (*RMSFE*) of each model $f^j_{t+h,t}$, relative to the benchmark *AR(2)* model, $f_{t+h,t}$. The *RMSFE* is calculated as follows:

$$RMSFE^h_j = \frac{\sqrt{\sum_{t=1}^{P}(y_{t+h} - f^j_{t+h,t})^2}}{\sqrt{\sum_{t=1}^{P}(y_{t+h} - f_{t+h,t})^2}} \tag{9}$$

where P is the number of h-step-ahead out-of-sample (*OOS*) forecasts, $h \geq 0$. If the value of $RMSFE_j$ is lower than 1, then model j outperforms the benchmark *AR(2)* model in terms of *RMSFE*, producing better output growth forecasts (nowcasts). To test whether model j produces significantly better forecasts, we test the null hypothesis of equal predictability proposed by Clark and West (2007). We choose this test because the benchmark *AR(2)* is nested by all other forecasting models[20]. For autocorrelated forecast errors ($h > 1$), we use Newey and West (1986) to estimate an autocorrelation consistent standard errors. Finally, recent findings by Gonçalves et al. (2017) justifies using the usual critical values when testing for equal predictability with estimated factors in the larger, nesting model.

It is worthwhile mentioning again that the main methodological contribution of this paper is not proposing an ultimate forecasting model for output growth but rather combining new developments from text regression and supervised machine learning in a manner that is novel for identifying the most predictive information from FED minutes. We are aware that other forecasting models, such as the ones that include different lag structures for D_t in prediction equations (7 and 8), could generate more accurate forecasts of real time output growth. Moreover, FED based predictors developed in this paper could also be used to increase accuracy of existing forecasting models such as the MIDAS models proposed by Andreou et al. (2013), Carriero et al. (2015) and Marcellino et al. (2016). In this case, we would be using both high-frequency and textual information for predicting real time output growth. We present our main results in the next section.

[20]The test by Diebold and Mariano (2002) is designed to compare non-nested models. If the forecasting models are nested, then the DM test may be undersized under the null and may have low power under the alternative hypothesis.

6 Results

In this section we report our results for the full out-of-sample period, 1992Q1-2017Q2, as well as for two sub periods. The first one includes only quarters for which the actual GDP growth was negative: we used the NBER classification of recession from which the periods 2001Q1-2001Q4 and 2007Q4-2009Q2 were classified as recessions; the second sub period includes quarters for which the GDP growth was positive (expansion period). According, to Chauvet and Potter (2013), forecasts from *AR(2)* are especially good to forecast output growth during expansion periods. Hence, in order to conclude that the information from FED minutes contains predictive power, the proposed forecasting models should outperform the benchmark *AR(2)* not only when it under performs (recession periods) but also when it performs the best (expansion periods).

Table 2 displays the values of the RMSFE (equation 9) relative to the benchmark *AR(2)* as well as the result of the CW test for the null hypothesis of equal model accuracy. The fixed dictionary based forecasts (models 4 and 5) do not seem to add much predictive power to the benchmark *AR(2)* as their relative RMSFE are mostly not statistically different from 1. Based on this result, we might be led to conclude that the information contained in the FED minutes does not help forecast output growth. However, the main contribution of this paper is to show that fixed dictionaries does not allow us to extract the best from the FED minutes since the vocabulary used by central banks are likely changing over time. Indeed, Table 2 also shows that time-varying dictionary based forecasts (M_i, $i = 1, 2, 3, 6, 7.$) outperform the benchmark *AR(2)* by a good margin and the result of the CW test indicates the rejection of the null hypothesis of equal accuracy at the usual 10% significance level. Model M_1 stands out as the best FED minute based forecast.

Although using LASSO and ridge regression to select the most predictive words (models M_6 and M_7) improves forecasting accuracy relative to the benchmark, they do not improve over the corresponding model that uses elastic net (model M_1). Thus, allowing for a time-varying dictionary improves forecasting accuracy but the method employed to select the most predictive words also matters. Results for model M_8 show the importance of pre-selecting the most predictive words before computing the common factors. Indeed, Table 2 shows that models $M_i, i = 1, 2, 3, 6, 7$ outperform model M_8 suggesting that the pre-selection of words help increase forecasting accuracy. Finally, the Blue Chip (BC) and AR(2)-BC forecasts does not perform well relatively to the benchmark, suggesting that, if we are able to construct time-varying dictionaries, the information extracted from FED minutes may have more predictive power

than the information used by individual forecasters. These findings represents the first empirical evidence about the predictive content of the FED minutes and points out the importance of combining machine learning techniques (elastic net) and textual data to improve performance of macroeconomic forecasting models.

Table 3 shows that some striking results are observed when we divide the full out-of-sample period across business cycles phases. The performance of all models is better in forecasting expansions than recessions. All models are less accurate to forecast output growth during recessions but such a loss is lower in the models that use FED minute based predictors computed from time-varying dictionaries (M_i, $i = 1, 2, 3.$). This result suggests that information from FED minutes can also be used to minimize the accuracy loss of forecasting models during recession times as long as we allow for time-varying dictionaries. Notice that models that use LASSO and ridge regression (M_6 and M_7) do not outperform model M_1 during either expansion or recession periods, suggesting again that the method used to pre-select the words also matters a lot. Tables 2 and 3 also suggest that combining forecasts from models that incorporate FED minute information, M_i, $i = 1, 2, ...8$, produces sizable accuracy gains relative to a benchmark *AR(2)* model, especially during the expansion periods. The analysis of the CW test also suggests that the simple equal-weighted forecast combination cannot be statistically worse than the *AR(2)* benchmark. These results support the conclusion that if we are able to extract the most predictive information from the minutes, then the FED minute based forecasts will outperform the benchmark AR(2) exactly during the periods in that it has been unbeatable.

Finally, Tables 4 and 5 reports the results for our nowcasting exercise. The main motivation for this exercise relies on the fact that actual GDP growth in quarter t is released with a delay. Recall that a nowcast is computed using equation (8) and the benchmark nowcast uses the same equation without predictors $D_{i,t}$. Tables 4 and 5 confirm that FED minutes contain predictive power when a time-varying dictionary is accounted for. We also found that the equal-weight combination of nowcasting models works very well.

The results reported in this section lend support to the conclusion that if we allow for time-varying vocabulary, then text information extracted from FED minutes by using elastic net can be used to forecast (nowcast) output growth. In the next section, we explain why the proposed forecasting approach performs so well.

{ Place Table 2 here. }

{ Place Table 3 here. }

{ Place Table 4 here. }

{ Place Table 5 here. }

6.1 Explaining the Success of the FED minute Based Forecasts

In this section, we decompose the mean-square forecast error (*MSFE*) into two parts: the forecast variance and the squared forecast bias. We calculate the *MSFE* of any forecast $f^j_{t+h,t} \in (M_i,\ BC,$ AR(2)-BC, $FC)$ as $\frac{1}{P}\sum_t \left(y_{t+h} - f^j_{t+h,t}\right)^2$ and the unconditional forecast variance as $\frac{1}{P}\sum_t \left(f^j_{t+h,t} - \frac{1}{P}\sum_t f^j_{t+h,t}\right)^2$, where P is the total number of out-of-sample forecasts. The squared forecast bias is computed as the difference between *MSFE* and forecast variance (Elliott and Timmermann, 2013; Rapach and Strauss, 2010) and Lima and Meng (2017).

Figures 2-5 depict the relative forecast variance and squared forecast bias of all forecasting models $M_i\ i = 1, ...,$; Blue Chip (BC); The $AR(2) - BC$; and equal-weight forecast combination (FC) for the full out-of-sample period. The relative forecast variance (squared bias) is calculated as the difference between the forecast variance (squared bias) of the jth model and the forecast variance (squared bias) of the benchmark *AR(2)* model. Hence, the value of relative forecast variance (squared bias) for the *AR(2)* is necessarily equal to zero. Each point on the solid line represents a forecast with the same *MSFE* as the *AR(2)*; points to the right of the line are forecasts outperformed by the *AR(2)*, and points to the left represent forecasts that outperform the *AR(2)*. Finally, both forecast variance and squared forecast bias are measured in the same scale so that it is possible to determine the trade-off between variance and bias of each forecasting model.

Since the *AR(2)* model is a (nested) parsimonious version of M_i and $AR(2) - BC$, it will have the lowest variance among all of them, but will be biased. Figures 2-5 show the (*MSFE*) decomposition for various forecast horizons $h = 1, 3, 6,$ and nowcasting ($h = 0$). Figures 2-5 confirm that the benchmark *AR(2)* forecast has the lowest variance but it is biased. Therefore, the only channel by which a forecasting model could outperform the benchmark is through a reduction of the forecast bias. A good model will be the one that can substantially reduce bias without increasing variance too much. The results displayed on those figures show that forecasting models that rely on fixed dictionaries, M_i, $i = 4, 5$ are unable to reduce bias substantially and, for this reason, they fail to outperform the benchmark *AR(2)* by a large margin. On the other hand, we can see that the good performance of the forecasting models that rely on

time-varying dictionaries, M_i, $i = 1, 2, 3, 6, 7$, stems from their ability to reduce substantially the squared forecast bias without increasing their forecast variance too much. Figures 2-5 shows that this finding occurs at all forecasting horizons.

The preceding discussion offers an explanation to the results presented in the previous section. The information contained in the FED minutes are useful to forecast real time output growth because it reduces forecast bias substantially, relative to the benchmark *AR(2)*, without increasing forecasting variance too much. This result relies on our ability to use machine learning (elastic net) to identify the most predictive words from the FED minutes. It also helps us understand why the popular text regression approach based on fixed dictionaries is not useful for macroeconomic forecasting. Indeed, forecasting models based on fixed dictionaries fail to reduce forecasting bias substantially and for this reason they hardly or never outperform the benchmark *AR(2)*.

Table (6) helps us understand why a time-varying dictionary should be interpreted as a powerful device for macroeconomic forecasting. It displays a list with some of the most predictive words during the out-of-sample period. This classification is based on the absolute value of the coefficients on X_t, estimated by elastic net. The largest coefficients are used to identify the most predictive words.[21]. The first conclusion from Table (6) is that some words (such as "economic activity") have predictive power during recession periods but is less selected during expansion times. Other words (such as "oil") is less selected during recession periods. Finally, the stemmed word "recoveri" has strong predictive power throughout the out-of-sample period. This analysis suggests that the information contained in the FED minutes are indeed useful for macroeconomic forecasting, but the predictive power of the words are quite different over different periods. Thus, the success of a forecasting model that utilizes information from FED minutes depends on its ability to identify the most predictive words over time, and this is exactly what is done by the method proposed in this paper.

{ Place Table 6 here. }

6.2 Green Book forecasts and the FED minutes based forecasts

A natural question is whether the FED minutes have value above and beyond the staff's forecast (Green Book forecast). There are several reasons to conclude that the FED minutes provide additional information that is useful for macroeconomic forecasting. In order to elaborate on this point, we describe below

[21] This classification method works because all time series X_t are measured at the same standard normal scale.

some stages of a FOMC (Federal Open Market Committee) meeting:[22]

stage i. A New York FED official presents financial and foreign exchange market developments, and staff answer questions on these financial conditions.

stage ii. There is an economic situation discussion:

a. Board of Governors's staff present the Green Book forecasts;

b. There are a series of questions on the staff presentation;

c. FOMC members present their views on the economic outlook.

These discussions bring out a new information set (opinions and comments) that has not been yet incorporated in the Green Book forecasts. There is also a monetary policy strategy discussion where more information is brought out.

stage iii. Monetary policy strategy discussion:

a. The board's director of monetary affairs presents a variety of monetary policy alternatives;

b. a potential round of staff questions;

c. the chairman and the other FOMC members discuss their policy preferences.

v. The FOMC votes on the policy decision.

As we can see from the above description, new information is generated through discussions that occur during a FOMC meeting. To help the minute writers, FOMC meetings are recorded and the tapes are used later as a source of information to write a minute. If such information is useful for economic forecasting, then the FED minute based forecasts proposed in this paper should outperform the Green Book forecasts. Table 7 reports the RMSFE relative to the benchmark AR(2) model. We report the out-of-sample performance of the Green Book and the forecasts obtained using model M_1 for $h = 0, 1, 3, 6$.[23] Notice that the Green Book forecast is published with a 5 year delay and for this reason we end our out-of-sample analysis in 2012Q4.[24] The main conclusion from Table 7 is that both Green Book and M_1 model outperform the AR(2) benchmark for $h = 0$ (nowcasting), but only the FED minute based forecast (M_1) outperforms the benchmark by a large margin for longer horizons. We take this result as a strong evidence that information generated during a FOMC meeting is useful for out-of-sample forecasting.

[22] We extracted this information from Hansen et al. (2017).

[23] We only consider the Green book forecasts presented in the last meeting of each quarter.

[24] This explains why the results reported for the M_1 in Table 7 are different from the ones previously reported in Tables 2 and 4.

6.3 Impulse Response Function Analysis

In this section, we show some empirical evidence of the transmission mechanism from FED's communication shocks to future values of the yield spread. This empirical finding is identified through the estimation of a structural impulse-response function (Lucca and Trebbi (2009)). We compute yield spread as the difference between the 10-year treasury yield and the 3-month treasury yield whose data are collected from the FRED database available at the FED-Saint Louis (T10Y3M).

We estimate a VAR model with variables ordered as {hawkish factor, dovish factor, spread}, where hawkish (dovish) factor is the common factor for the hawkish (dovish) bigrams selected by approach M_3. Recall that these factors only load hawkish (dovish) information and, therefore, they can be used in our impulse response analysis to identify the effect of a shock to FED's hawkish (dovish) communication on future values of yield spread. We expect that shocks in the hawkish factor would raise the yield spread (positive relationship) whereas shocks in the dovish factor would decrease the yield spread (negative relationship).

In order to estimate the structure of impulse-response function, we apply a short-run restriction (Choleski factorization) where, given the ordering of the variables in the VAR model, the contemporaneous effect of the yield spread on the two factors is zero. The data on yield spread ranges from 1984Q1 to 2017Q2. We also multiply the yield spread by 100 so that the response to one-standard deviation shock in the factors is expressed in basis points. Finally, the VAR model was estimated with 2 lags (Schwarz information criterion) and we used bootstrap with 100 runs to estimate a confidence interval for the impulse-response function.

Figures (6) and (7) show responses of the yield spread to unexpected unit standard deviation shocks in the hawskish and dovish factors. A positive shock in the hawkish factor is subsequently followed by a yield spread increase. The yield spread displays a hump-shaped response to a shock in the hawkish factor with a peak of about 15 basis points after 5 quarters. On the other hand, a positive shock in the dovish factor (Figure (7)) is subsenquetly followed by a yield spread decrease with a peak of about 25 basis points after 5 quarters. The red lines shown in Figure (6) and (7) are two-standard error bootstrapped confidence bands. The main conclusion from this impulse-response analysis is that shocks to FED's communication is transmitted to the yield spread with a peak of this transmission occurring at 5 quarters. This finding helps us understand why the FED based predictors proposed in this paper are useful for macroeconomic forecasting.

{ Place Figure 6 here. }

{ Place Figure 7 here. }

7 Conclusion

Words matter. The central bank's communications could indicate the expected monetary policy directions of the economy. The monetary authority with strong credibility is capable of controlling the expectations of private agents. Words used in FED minutes, thus, have market moving power. The channel through which this result operates is that positive shocks to hawkish (dovish) communication lead to an increase (decrease) in the yield spread and therefore are interpreted as contractionary (expansionary) monetary policy shocks. A simple impulse-response analysis conducted in this paper lends empirical support to this channel.

Applied researchers, policy makers and various investors are always in search for useful indicators to successfully predict future economic activities. The forecasting literature has shown that a simple parsimonious time series *AR(2)* model often outperforms more complex macro models in forecasting economic growth, especially during periods of economic expansions. Interestingly, we propose augmenting the simple *AR(2)* model by including predictors computed from a time-varying dictionary and we find that this new model has superior predictive power. This was not the case when we used predictors constructed from fixed dictionaries instead. As explained earlier, the time varying dictionary is more capable in capturing the tone of the central bank at any given time and, therefore, can be used to construct better predictors of the future economic activities.

Our results suggest that the combination of supervised machine learning (elastic net) and textual data has potential to identify document based predictors that are more powerful than the counterpart constructed from fixed dictionaries. Our findings show that information contained in FED minutes has predictive power and that, for this reason, FED minute based forecasts outperform forecasts that rely on other information sets, such as Blue Chip and Green Book forecasts. It, thus, has significant implication for empirical researchers in evaluating the effects of monetary policy on the economy. Identifying monetary policy shocks remains to be an important challenge among researchers. The identified predictors in a time varying setting, as we show in this study, could potentially capture true monetary policy shocks. Future research should explore this channel to evaluate the effects of monetary policy on output and other leading macro aggregates.

References

S. C. Ahn and A. R. Horenstein. Eigenvalue ratio test for the number of factors. *Econometrica*, 81(3): 1203–1227, 2013.

R. Ahrens. Predicting recessions with interest rate spreads: a multicountry regime-switching analysis. *Journal of international Money and Finance*, 21(4):519–537, 2002.

C. Altavilla and D. Giannone. The effectiveness of non-standard monetary policy measures: Evidence from survey data. *Journal of Applied Econometrics*, 32(5):952–964, 2017.

E. Andreou, E. Ghysels, and A. Kourtellos. Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251, 2013.

A. Ang, M. Piazzesi, and M. Wei. What does the yield curve tell us about gdp growth? *Journal of econometrics*, 131(1-2):359–403, 2006.

M. Apel and M. Grimaldi. The information content of central bank minutes. *Sveriges Riksbank Working Paper Series*, (261), 2012.

M. T. Armesto, R. HERNÁNDEZ-MURILLO, M. T. Owyang, and J. Piger. Measuring the information content of the beige book: A mixed data sampling approach. *Journal of Money, Credit and Banking*, 41(1):35–55, 2009.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304–317, 2008.

H. Bernard and S. Gerlach. Does the term structure predict recessions? the international evidence. *International Journal of Finance & Economics*, 3(3):195–215, 1998.

D. M. Blei and J. D. Lafferty. Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC, 2009.

E. Boukus and J. V. Rosenberg. The information content of fomc minutes. *New York Fed*, 2006.

A. Carriero, T. E. Clark, and M. Marcellino. Realtime nowcasting with a bayesian mixed frequency model with stochastic volatility. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178 (4):837–862, 2015.

S. G. Cecchetti et al. What the fomc says and does when the stock market booms. In *Asset Prices and Monetary Policy, Proceedings of the Research Conference of the Reserve Bank of Australia*, pages 77–96, 2003.

C. Chakraborty and A. Joseph. Machine learning at central banks. 2017.

M. Chauvet and S. Potter. Forecasting output. In *Handbook of Economic Forecasting*, volume 2, pages 141–194. Elsevier, 2013.

L. J. Christiano, M. Eichenbaum, and C. L. Evans. Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, 1:65–148, 1999.

T. E. Clark and K. D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of econometrics*, 138(1):291–311, 2007.

F. X. Diebold and R. S. Mariano. Comparing predictive accuracy. *Journal of Business & economic statistics*, 20(1):134–144, 2002.

F. X. Diebold and M. Shin. Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. Technical report, National Bureau of Economic Research, 2018.

A. Dossani. *Essays on Inference from Option Markets*. PhD thesis, UC San Diego, 2018.

A. Duarte, I. A. Venetis, and I. Paya. Predicting real growth and the probability of recession in the euro area using the yield spread. *International Journal of Forecasting*, 21(2):261–277, 2005.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32 (2):407–499, 2004.

G. Elliott and A. Timmermann. *Handbook of economic forecasting*. Elsevier, 2013.

J. E. Engelberg and C. A. Parsons. The causal impact of media in financial markets. *The Journal of Finance*, 66(1):67–97, 2011.

A. Estrella. Why does the yield curve predict output and inflation? *The Economic Journal*, 115(505): 722–744, 2005.

A. Estrella and G. A. Hardouvelis. The term structure as a predictor of real economic activity. *The journal of Finance*, 46(2):555–576, 1991.

A. Estrella and F. S. Mishkin. Predicting us recessions: Financial variables as leading indicators. *Review of Economics and Statistics*, 80(1):45–61, 1998.

D. Garcia. Sentiment during recessions. *The Journal of Finance*, 68(3):1267–1300, 2013.

M. Gentzkow, B. T. Kelly, and M. Taddy. Text as data. Working Paper 23276, National Bureau of Economic Research, March 2017. URL http://www.nber.org/papers/w23276.

S. Gonçalves, M. W. McCracken, and B. Perron. Tests of equal accuracy for nested models with estimated factors. *Journal of Econometrics*, 198(2):231–252, 2017.

S. Hansen, M. McMahon, and A. Prat. Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870, 2017.

F. E. Kydland and E. C. Prescott. Rules rather than discretion: The inconsistency of optimal plans. *Journal of political economy*, 85(3):473–491, 1977.

J. Li. Sparse and stable portfolio selection with parameter uncertainty. *Journal of Business & Economic Statistics*, 33(3):381–392, 2015.

J. Li, I. Tsiakas, and W. Wang. Predicting exchange rates out of sample: Can economic fundamentals beat the random walk? *Journal of Financial Econometrics*, 13(2):293–341, 2015.

L. R. Lima and F. Meng. Out-of-sample return predictability: A quantile combination approach. *Journal of Applied Econometrics*, 32(4):877–895, 2017.

L. R. Lima, F. Meng, and L. Godeiro. Quantile forecasting with mixed-frequency data. *International Journal of Forecasting*, Forthcoming, 2018.

D. O. Lucca and E. Moench. The pre-fomc announcement drift. *The Journal of Finance*, 70(1):329–371, 2015.

D. O. Lucca and F. Trebbi. Measuring central bank communication: an automated approach with application to fomc statements. Technical report, National Bureau of Economic Research, 2009.

A. Mansoorian and M. Mohsin. Monetary policy in a cash-in-advance economy: employment, capital accumulation, and the term structure of interest rates. *Canadian Journal of Economics/Revue canadienne d'économique*, 37(2):336–352, 2004.

M. Marcellino, M. Porqueddu, and F. Venditti. Short-term gdp forecasting with a mixed-frequency dynamic factor model with stochastic volatility. *Journal of Business & Economic Statistics*, 34(1):118–127, 2016.

C. R. Nelson. The prediction performance of the frb-mit-penn model of the us economy. *The American Economic Review*, 62(5):902–917, 1972.

W. K. Newey and K. D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix, 1986.

C. I. Plosser and K. G. Rouwenhorst. International term structures and real economic growth. *Journal of monetary economics*, 33(1):133–155, 1994.

V. A. Ramey. Macroeconomic shocks and their propagation. In *Handbook of Macroeconomics*, volume 2, pages 71–162. Elsevier, 2016.

D. E. Rapach and J. K. Strauss. Bagging or combining (or both)? an analysis based on forecasting us employment growth. *Econometric Reviews*, 29(5-6):511–533, 2010.

J. H. Stock and M. W. Watson. New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 4:351–394, 1989.

L. A. Thorsrud. Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, (just-accepted):1–35, 2018.

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

J. H. Wright. What does monetary policy do to long-term interest rates at the zero lower bound? *The Economic Journal*, 122(564), 2012.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

Appendix

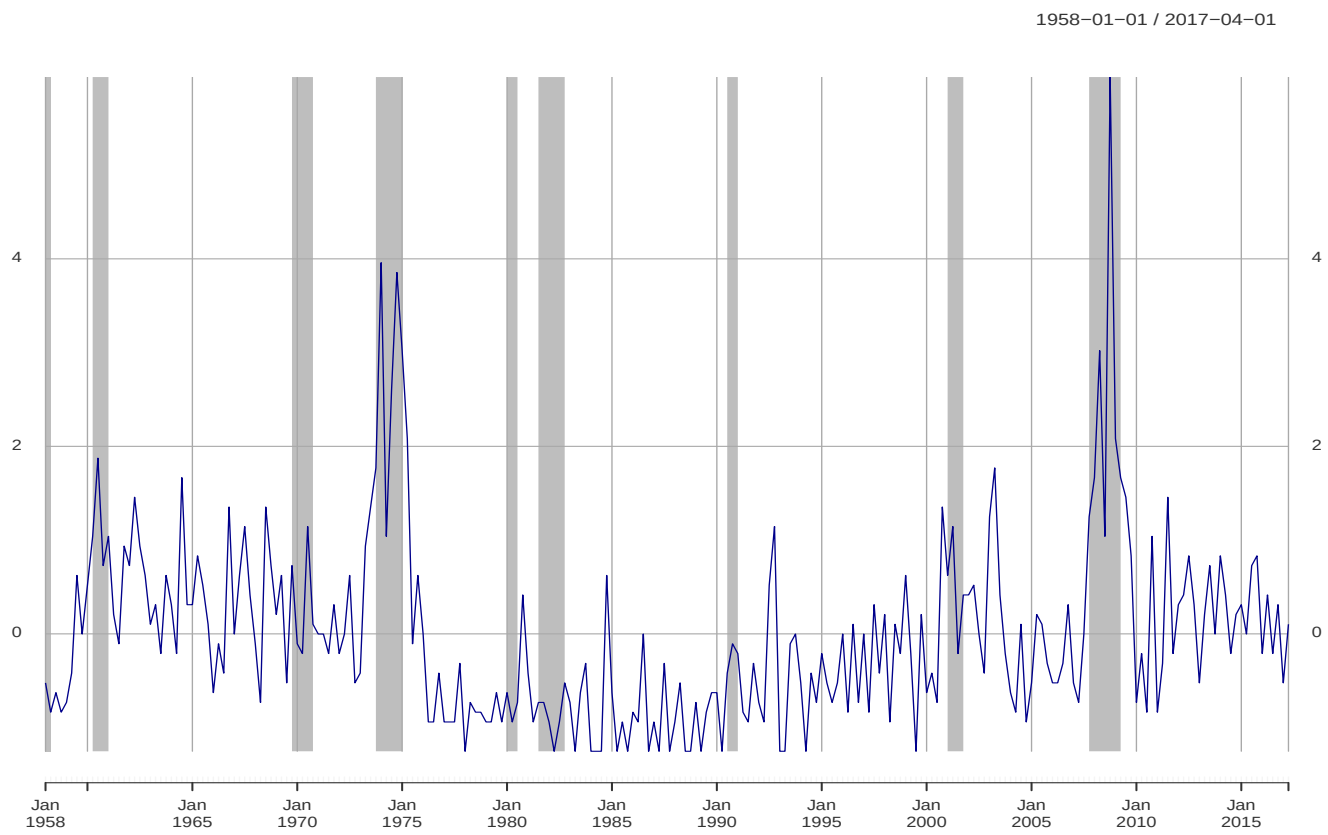

Figure 1: Time series Plot of term "economic activity", 1958(Q1)-2017(Q2).

This figure shows the time series plot of the normalized count value for the collocation term "economic activity" in the period 1958(Q1)-2017(Q2). The shaded area represents the recession dates according to NBER.

Table 1: GDP and FED minute release dates

Quarter	Output Release date	FED Minutes release date
1992Q1	29-May-92	22-May-92
1992Q2	27-Aug-92	21-Aug-92
1992Q3	27-Oct-92	9-Oct-92
1992Q4	27-Feb-93	5-Feb-93
1993Q1	28-May-93	21-May-93
1993Q2	29-Jul-93	12-Jul-93
1993Q3	27-Oct-93	8-Oct-93
1993Q4	1-Mar-94	7-Feb-94
1994Q1	27-May-94	20-May-94
1994Q2	29-Jul-94	11-Jul-94
1994Q3	28-Oct-94	13-Oct-94
1994Q4	27-Jan-95	16-Jan-95
1995Q1	31-May-95	1-May-95
1995Q2	28-Jul-95	10-Jul-95
1995Q3	27-Oct-95	17-Oct-95
1995Q4	23-Feb-96	5-Feb-96
1996Q1	30-May-96	24-May-96
1996Q2	1-Aug-96	5-Jul-96
1996Q3	27-Nov-96	15-Nov-96
1996Q4	28-Feb-97	7-Feb-97
1997Q1	26-May-97	23-May-97
1997Q2	31-Jul-97	7-Jul-97
1997Q3	26-Nov-97	14-Nov-97
1997Q4	27-Feb-98	6-Feb-98
1998Q1	27-May-98	22-May-98
1998Q2	31-Jul-98	6-Jul-98
1998Q3	30-Oct-98	19-Oct-98
1998Q4	26-Feb-99	5-Feb-99

1999Q1	27-May-99	21-May-99
1999Q2	29-Jul-99	15-Jul-99
1999Q3	29-Oct-99	8-Oct-99
1999Q4	28-Jan-00	13-Jan-00
2000Q1	27-Apr-00	10-Apr-00
2000Q2	28-Jul-00	14-Jul-00
2000Q3	27-Oct-00	6-Oct-00
2000Q4	31-Jan-01	6-Jan-01
2001Q1	27-Apr-01	13-Apr-01
2001Q2	27-Jul-01	12-Jul-01
2001Q3	31-Oct-01	5-Oct-01
2001Q4	30-Jan-02	11-Jan-02
2002Q1	26-Apr-02	3-Apr-02
2002Q2	31-Jul-02	12-Jul-02
2002Q3	31-Oct-02	10-Oct-02
2002Q4	31-Jan-03	11-Jan-03
2003Q1	25-Apr-03	4-Apr-03
2003Q2	28-Aug-03	15-Aug-03
2003Q3	30-Oct-03	2-Oct-03
2003Q4	30-Jan-04	5-Jan-04
2004Q1	26-May-04	6-May-04
2004Q2	31-Aug-04	12-Aug-04
2004Q3	30-Nov-04	11-Nov-04
2004Q4	28-Jan-05	4-Jan-05
2005Q1	28-Apr-05	12-Apr-05
2005Q2	29-Jul-05	21-Jul-05
2005Q3	28-Oct-05	11-Oct-05
2005Q4	27-Jan-06	3-Jan-06
2006Q1	28-Apr-06	18-Apr-06
2006Q2	28-Jul-06	20-Jul-06
2006Q3	27-Oct-06	11-Oct-06

2006Q4	31-Jan-07	3-Jan-07
2007Q1	27-Apr-07	11-Apr-07
2007Q2	27-Jul-07	19-Jul-07
2007Q3	31-Oct-07	9-Oct-07
2007Q4	30-Jan-08	2-Jan-08
2008Q1	30-Apr-08	8-Apr-08
2008Q2	31-Jul-08	16-Jul-08
2008Q3	30-Oct-08	7-Oct-08
2008Q4	30-Jan-09	6-Jan-09
2009Q1	29-Apr-09	8-Apr-09
2009Q2	31-Jul-09	15-Jul-09
2009Q3	29-Oct-09	14-Oct-09
2009Q4	29-Jan-10	6-Jan-10
2010Q1	30-Apr-10	10-Apr-10
2010Q2	30-Jul-10	14-Jul-10
2010Q3	29-Oct-10	12-Oct-10
2010Q4	28-Jan-11	4-Jan-11
2011Q1	28-Apr-11	5-Apr-11
2011Q2	27-Jul-11	12-Jul-11
2011Q3	27-Oct-11	12-Oct-11
2011Q4	27-Jan-12	3-Jan-12
2012Q1	27-Apr-12	3-Apr-12
2012Q2	27-Jul-12	11-Jul-12
2012Q3	26-Oct-12	4-Oct-12
2012Q4	30-Jan-13	3-Jan-13
2013Q1	26-Apr-13	10-Apr-13
2013Q2	31-Jul-13	10-Jul-13
2013Q3	7-Nov-13	9-Oct-13
2013Q4	30-Jan-14	8-Jan-14
2014Q1	30-Apr-14	9-Apr-14
2014Q2	30-Jul-14	9-Jul-14

2014Q3	30-Oct-14	8-Oct-14
2014Q4	30-Jan-15	7-Jan-15
2015Q1	29-Apr-15	8-Apr-15
2015Q2	30-Jul-15	8-Jul-15
2015Q3	29-Oct-15	8-Oct-15
2015Q4	29-Jan-16	6-Jan-16
2016Q1	28-Apr-16	8-Apr-16
2016Q2	28-Jul-16	8-Jul-16
2016Q3	28-Oct-16	8-Oct-16
2016Q4	27-Jan-17	6-Jan-17
2017Q1	28-Apr-17	5-Apr-17
2017Q2	28-Jul-17	5-Jul-17

Table 1 shows the date of the first release of quarterly GDP data and the release date of minutes corresponding to the last FOMC meeting of each quarter.

Table 2: RMSFEs for $h = 1, 3, 6$ during full out-of-sample period. Relative to the Benchmark *AR(2)*.

Models	h=1	h=3	h=6
Model 1	**0.748***	**0.722****	**0.707****
Model 2	0.837*	0.792**	0.771*
Model 3	0.852*	0.811**	0.802**
Model 4	1.033	1.005***	1.009
Model 5	0.994	1.020	0.988
Model 6	0.861**	0.883*	0.847**
Model 7	0.904***	0.894**	0.865*
Model 8	0.976*	0.962**	0.956*
Blue Chip	0.947**	1.011	0.985
AR(2) Blue Chip	0.995**	0.992	0.977***
Forecast Combination	0.874*	0.832**	0.811*

Table 2 shows the *RMSFE* of forecasting models $f^j_{t+h,t}$ relative to the benchmark *AR(2)* for the full out-of-sample period. A value less than 1 indicates that $f^j_{t+h,t}$ outperforms *AR(2)*. $(*)$, $(**)$ and $(***)$ denote that the difference between the mean-squared-error (MSE) of model $f^j_{t+h,t}$ and the benchmark

is statistically significant at the 1% , 5% and 10% level, respectively, using Clark and West (2007) test. For autocorrelated forecast errors ($h > 1$), we use Newey and West (1986) to estimate standard errors. Recent findings by Gonçalves et al. (2017) justifies using the usual critical values when testing for equal predictability with estimated factors in the larger, nesting model.

Table 3: RMSFEs for $h = 1, 3, 6$ during expansion and recession sub periods. Total and relative to the Benchmark *AR(2)*

	h=1		h=3		h=6	
Models	Expansion	Recession	Expansion	Recession	Expansion	Recession
AR2	1.825	3.474	1.938	4.369	2.007	4.431
Model 1	1.348	2.161	1.701	2.320	1.571	2.743
Relative	**0.739***	**0.622***	0.878**	**0.531***	**0.783****	**0.619***
Model 2	1.527	2.737	1.748	3.154	1.639	3.185
Relative	0.837*	0.788**	0.902***	0.722*	0.817**	0.719**
Model 3	1.693	2.709	1.649	3.534	1.623	3.451
Relative	0.928***	0.780**	**0.851****	0.771*	0.809**	0.779*
Model 4	1.914	3.274	1.952	4.408	2.038	4.434
Relative	1.048	0.942	1.007***	1.009	1.015	1.000
Model 5	1.823	3.043	1.995	4.070	1.974	4.229
Relative	0.999*	0.876**	1.029	0.931***	0.983	0.954*
Model 6	1.600	3.123	1.688	4.159	1.686	4.205
Relative	0.877**	0.899**	0.871	0.952**	0.840**	0.949
Model 7	1.655	2.918	1.688	4.172	1.790	4.001
Relative	0.909**	0.840**	0.871*	0.955	0.892**	0.903**
Model 8	1.790	3.240	1.880	4.094	1.937	4.107
Relative	0.980	0.932*	0.970*	0.937**	0.965*	0.926**
Blue Chip	1.727	2.619	1.982	4.191	1.985	4.171
Relative	0.946**	0.754**	1.022	0.959**	0.989**	0.941**
AR(2) Blue Chip	1.817	3.312	1.933	4.134	1.979	4.214
Relative	0.995	0.953	0.997	0.946	0.986	0.950
Forecast Combination	1.655	2.674	1.614	3.372	1.565	3.416
Relative	0.907*	0.771***	0.833*	0.772**	0.780*	0.771**

Table 3 shows the absolute *RMSFE* and the *RMSFE* of forecasting models $f^j_{t+h,t}$ relative to the benchmark *AR(2)* for recession and expansion sub periods. A value less than 1 indicates that $f^j_{t+h,t}$ outperforms *AR(2)*. $(*)$, $(**)$ and $(***)$ denote that the difference between the mean-squared-error (MSE) of model $f^j_{t+h,t}$ and the benchmark is statistically significant at the 1% , 5% and 10% level, respectively, using Clark and West (2007) test. For autocorrelated forecast errors ($h > 1$), we use Newey and West (1986) to estimate standard errors. Recent findings by Gonçalves et al. (2017) justifies using the usual critical values

when testing for equal predictability with estimated factors in the larger, nesting model.

Table 4: RMSFEs for $h = 0$ during full out-of-sample period. Relative to the Benchmark *AR(2)*.

Models	Full Sample
Model 1	$\mathbf{0.702^{*}}$
Model 2	0.772^{*}
Model 3	0.808^{**}
Model 4	1.027
Model 5	1.014
Model 6	0.814^{**}
Model 7	0.866^{**}
Model 8	0.989^{**}
Blue Chip	0.921^{***}
AR(2) Blue Chip	0.863^{**}
Forecast Combination	0.862^{*}

Table 4 shows the *RMSFE* of nowcasting models $f^{j}_{t+h,t}$ relative to the benchmark *AR(2)* for the full out-of-sample period. A value less than 1 indicates that $f^{j}_{t+h,t}$ outperforms *AR(2)*. $(*)$, $(**)$ and $(***)$ denote that the difference between the mean-squared-error (MSE) of model $f^{j}_{t+h,t}$ and the benchmark is statistically significant at the 1% , 5% and 10% level, respectively, using Clark and West (2007) test.

Table 5: RMSFEs for $h = 0$ during expansion and recession sub periods. Total and relative to the Benchmark *AR(2)*.

Models	Expansion	Recession
AR2	1.825	3.474
Model 1	1.297	2.199
Relative	**0.711***	**0.633***
Model 2	1.595	2.602
Relative	0.874**	0.749*
Model 3	1.646	2.678
Relative	0.902**	0.771**
Model 4	1.984	3.192
Relative	1.087	0.918**
Model 5	1.856	3.469
Relative	1.017	0.998
Model 6	1.538	2.681
Relative	0.843**	0.772**
Model 7	1.587	2.845
Relative	0.870**	0.819**
Model 8	1.812	3.279
Relative	0.992***	0.943*
Blue Chip	1.684	3.025
Relative	0.923**	0.871**
AR(2) Blue Chip	1.658	2.852
Relative	0.909**	0.821*
Forecast Combination	1.569	2.838
Relative	0.860*	0.817*

Table 5 shows the *RMSFE* of nowcasting models $f^j_{t+h,t}$ relative to the benchmark *AR(2)* for recession and expansion sub periods. A value less than 1 indicates that $f^j_{t+h,t}$ outperforms *AR(2)*. $(*)$, $(**)$ and $(***)$ denote that the difference between the mean-squared-error (MSE) of model $f^j_{t+h,t}$ and the benchmark is statistically significant at the 1% , 5% and 10% level, respectively, using Clark and West (2007) test.

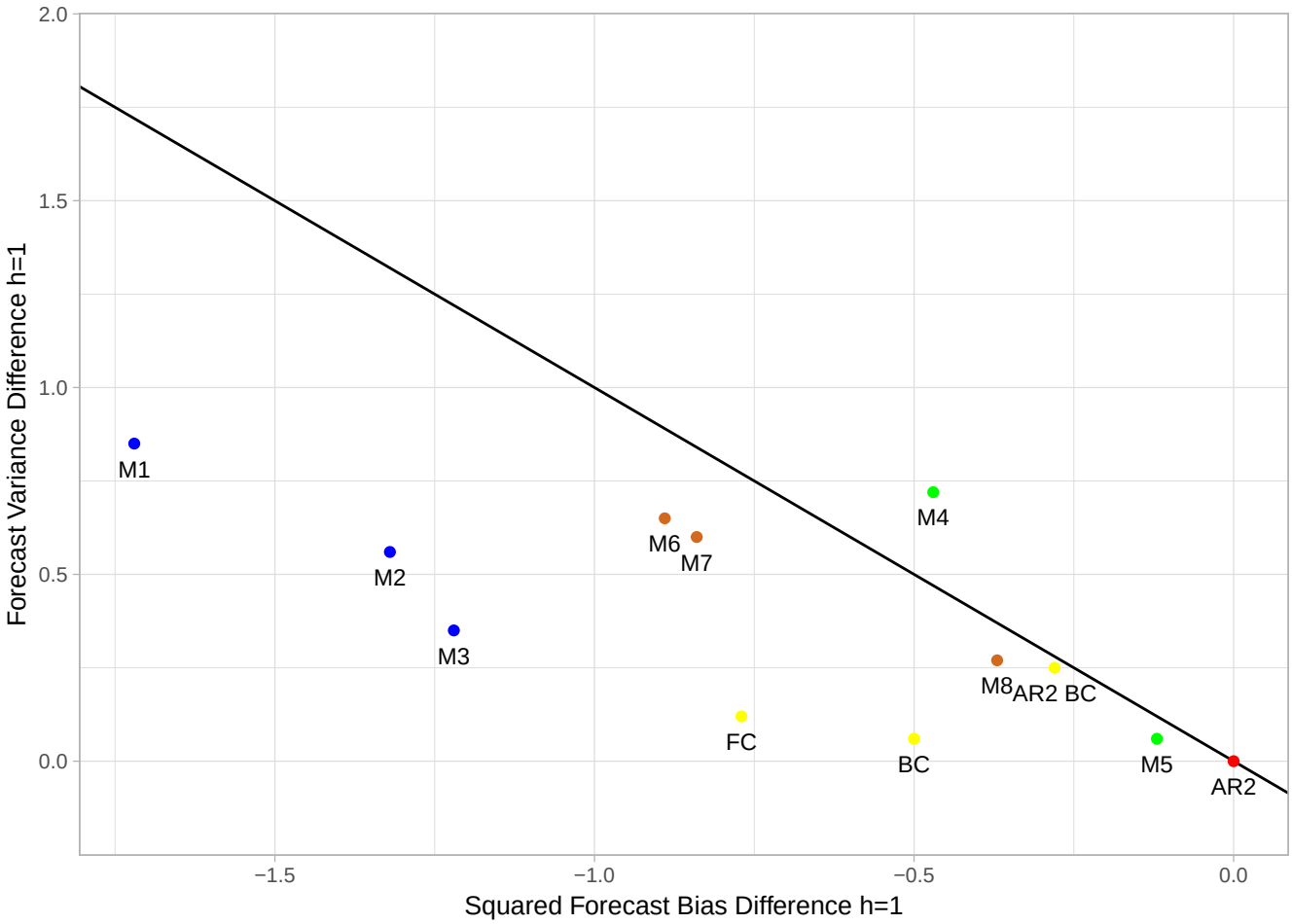

Figure 2: Scatter Plot of relative forecast variance and squared forecast bias, 1992Q1-2017Q2.

The y-axis and x-axis represent relative forecast variance and squared forecast bias for $h = 1$, calculated as the difference between the forecast variance (squared bias) of $f^j_{t+h,t}$ and the forecast variance (squared bias) of the *AR(2)*. Each point on the dotted line represents a forecast with the same *MSFE* as the *AR2*; points to the right are forecasts outperformed by the *AR(2)*, and points to the left represent forecasts that outperform the *AR(2)*. In blue we have time-varying dictionaries based models; in green, fixed dictionaries based models; in yellow, the Blue Chip, AR2 Blue Chip and Forecasts Combination; in orange other predictors based models; and in red we have the benchmark *AR(2)* Model.

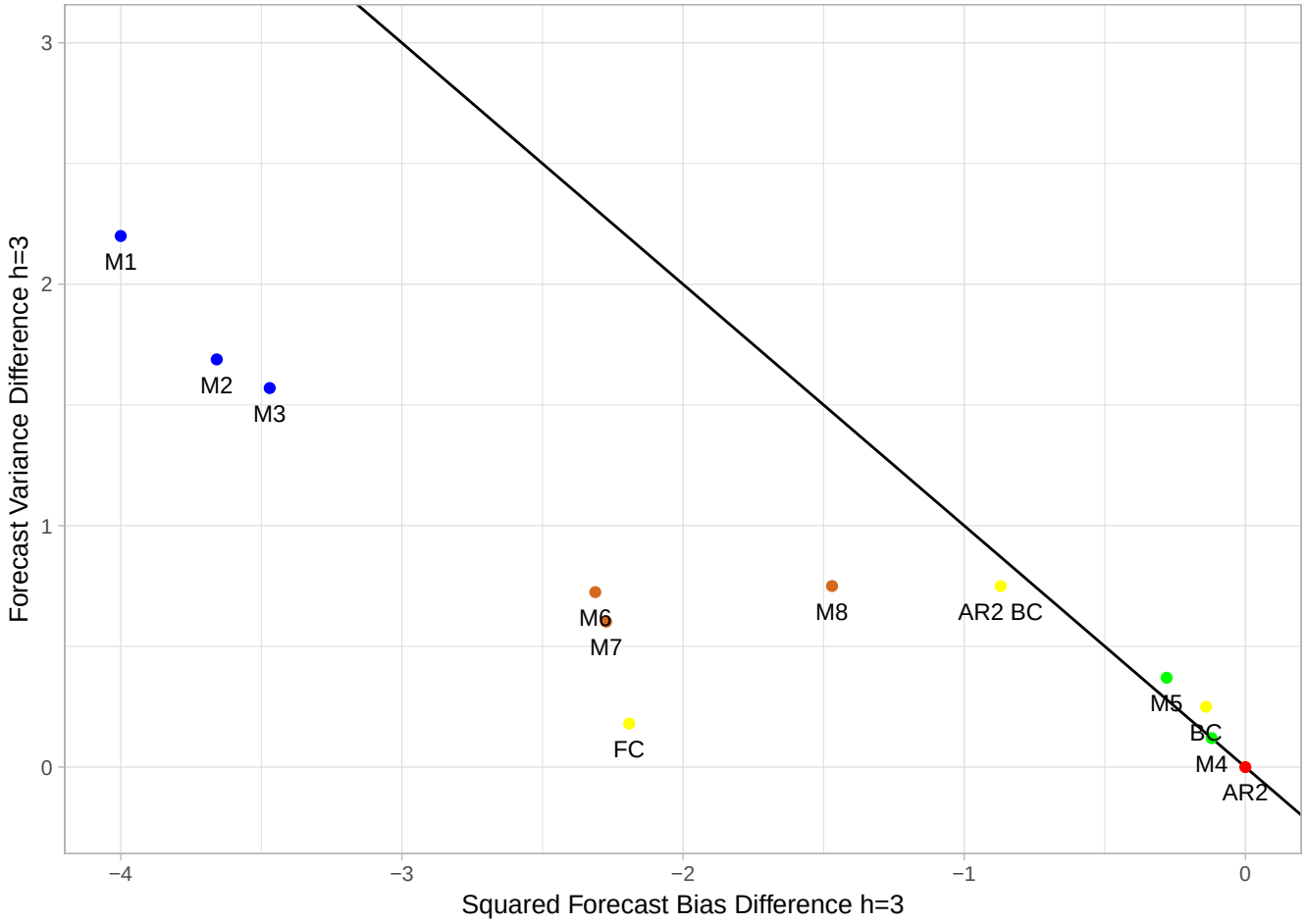

Figure 3: Scatter Plot of relative forecast variance and squared forecast bias, 1992Q1-2017Q2.

The y-axis and x-axis represent relative forecast variance and squared forecast bias for $h = 3$, calculated as the difference between the forecast variance (squared bias) of $f^j_{t+h,t}$ and the forecast variance (squared bias) of the *AR(2)*. Each point on the dotted line represents a forecast with the same *MSFE* as the *AR2*; points to the right are forecasts outperformed by the *AR(2)*, and points to the left represent forecasts that outperform the *AR(2)*. In blue we have time-varying dictionaries based models; in green, fixed dictionaries based models; in yellow, the Blue Chip, AR2 Blue Chip and Forecasts Combination; in orange other predictors based models; and in red we have the benchmark *AR(2)* Model.

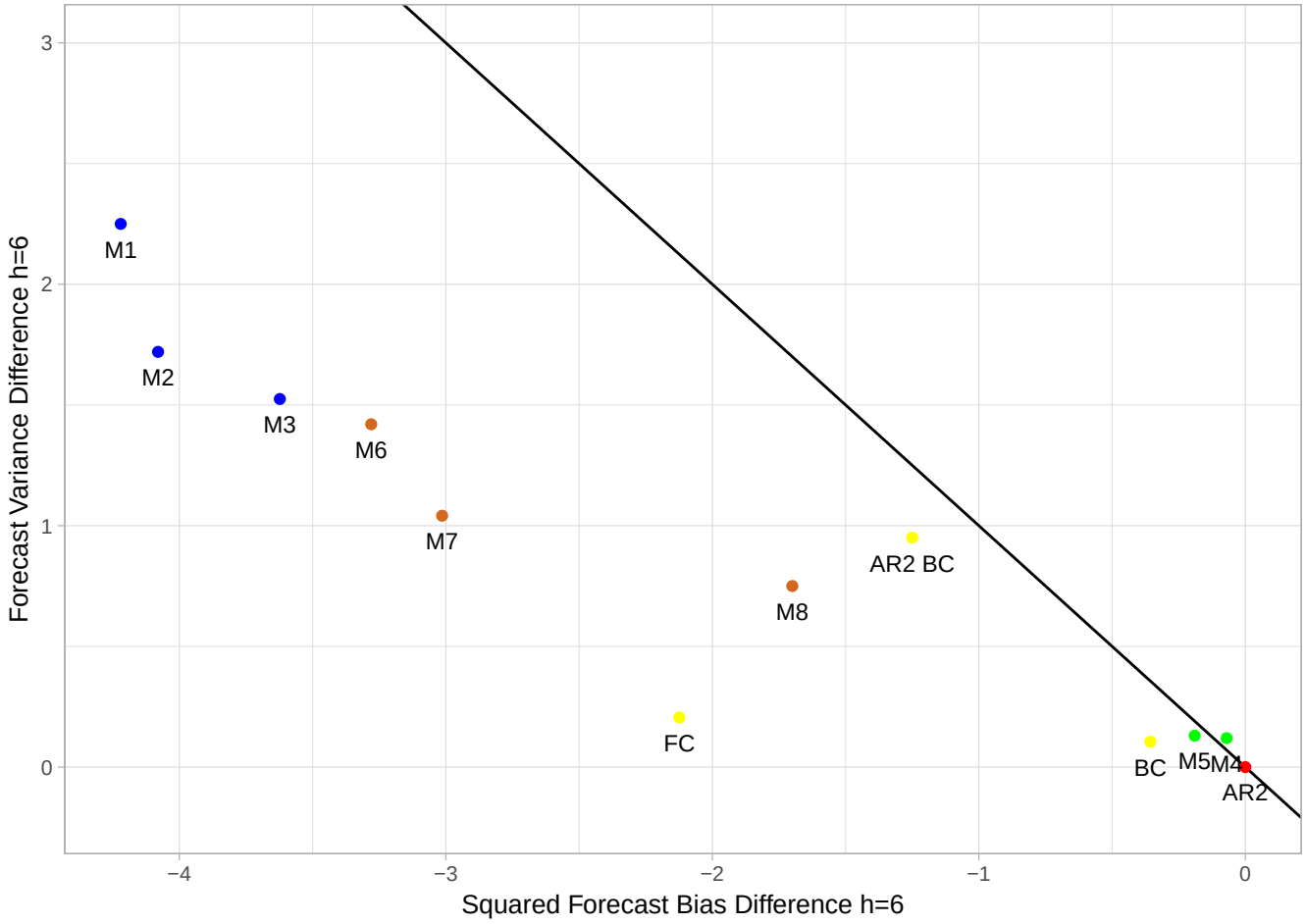

Figure 4: Scatter Plot of relative forecast variance and squared forecast bias, 1992Q1-2017Q2.

The y-axis and x-axis represent relative forecast variance and squared forecast bias for $h = 6$, calculated as the difference between the forecast variance (squared bias) of $f^j_{t+h,t}$ and the forecast variance (squared bias) of the *AR(2)*. Each point on the dotted line represents a forecast with the same *MSFE* as the *AR2*; points to the right are forecasts outperformed by the *AR(2)*, and points to the left represent forecasts that outperform the *AR(2)*. In blue we have time-varying dictionaries based models; in green, fixed dictionaries based models; in yellow, the Blue Chip, AR2 Blue Chip and Forecasts Combination; in orange other predictors based models; and in red we have the benchmark *AR(2)* Model.

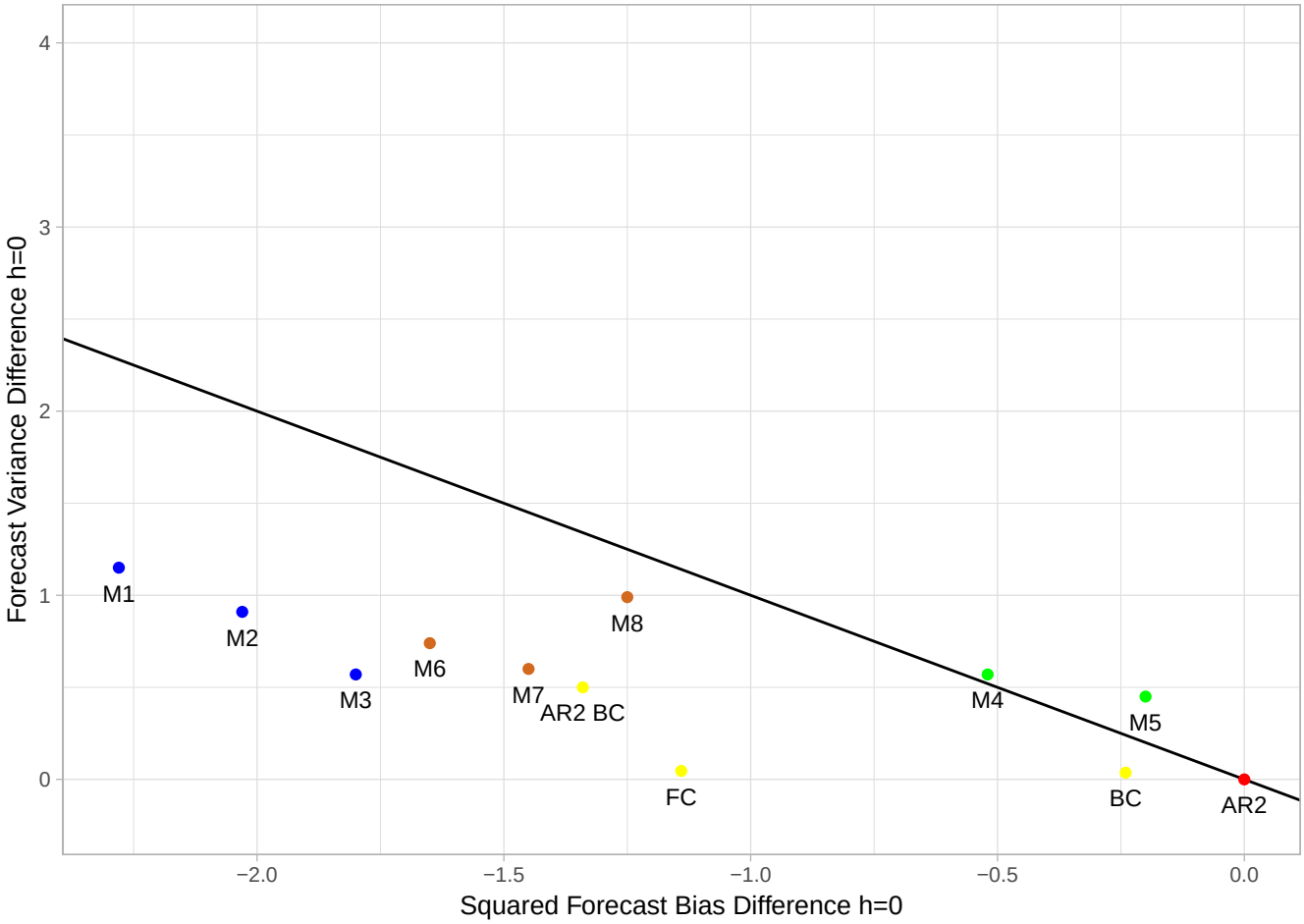

Figure 5: Scatter Plot of relative nowcast variance and squared nowcast bias, 1992Q1-2017Q2

The y-axis and x-axis represent relative forecast variance and squared forecast bias for $h = 0$, calculated as the difference between the forecast variance (squared bias) of $f^j_{t+h,t}$ and the forecast variance (squared bias) of the *AR(2)*. Each point on the dotted line represents a forecast with the same *MSFE* as the *AR2*; points to the right are forecasts outperformed by the *AR(2)*, and points to the left represent forecasts that outperform the *AR(2)*. In blue we have time-varying dictionaries based models; in green, fixed dictionaries based models; in yellow, the Blue Chip, AR2 Blue Chip and Forecasts Combination; in orange other predictors based models; and in red we have the benchmark *AR(2)* Model.

Table 6: Most selected (stemmed) words.

Term	Full Sample(%)	Expansion(%)	Recession(%)
recoveri	95.10	94.51	100.00
economic activity	60.78	56.04	100.00
weak	29.41	26.37	54.55
oil	25.49	27.47	9.09
upward pressure	18.63	16.48	36.36

Table 6 shows the average relative frequency of words classified as top-5 in terms of predictive power for all out-of-sample (sub)periods and $h = 1$. We rank the predictive power of words by using the absolute values of the coefficients on X_t, estimated by elastic net. The 5 largest coefficients correspond to the 5 most predictive words. Notice that "recoveri" is a stemmed word, not a typo.

Table 7: Green Book versus FED Minute Based Forecasts

Model	h=0	h=1	h=3	h=6
Green Book Fed	0.732^{*}	0.901^{**}	1.041^{***}	1.047
M1 Elastic Net	0.714^{*}	0.759^{**}	0.738^{*}	0.732^{*}

Table 7 shows the *RMSFE* relative to bechmark AR(2). Since the Green Book forecast is released with a 5-year delay, the out-of-sample period is from 1992-Q1 to 2012-Q4. We only consider the Green book forecasts presented in the last meeting of each quarter. For comparison, we report the results for the M_1 model.

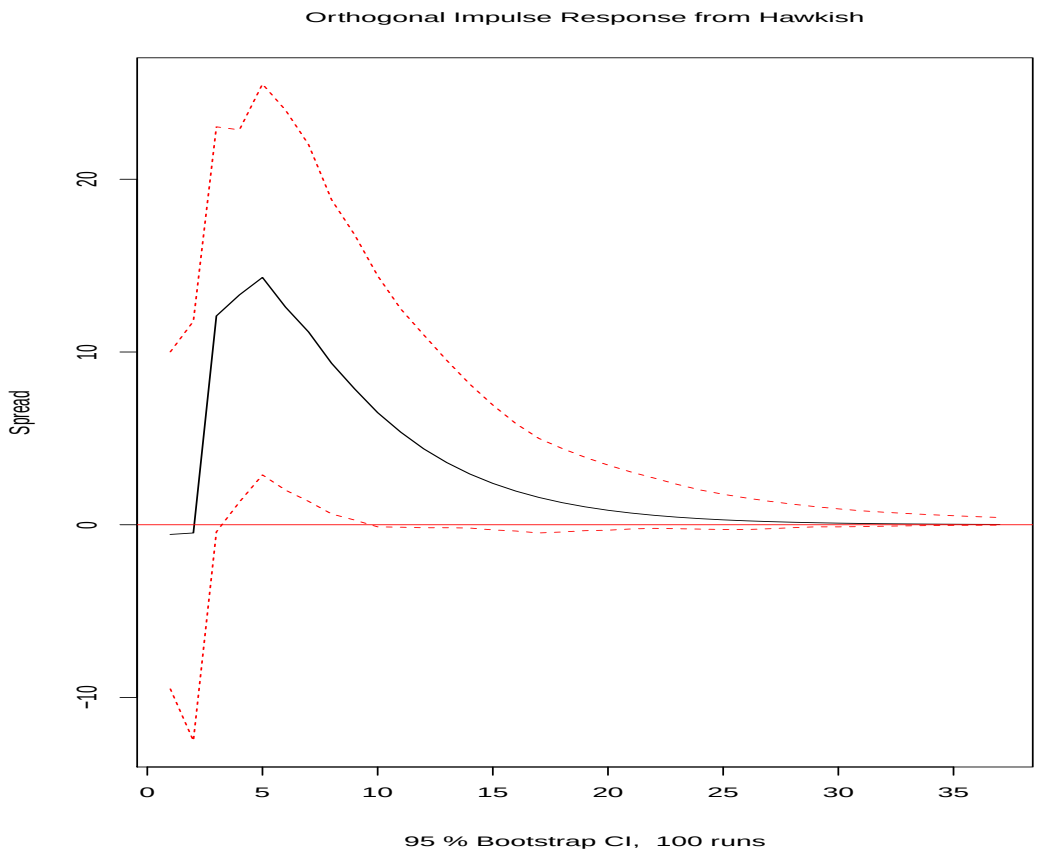

Figure 6: Response of the Yield Spread (10-year and 3-month Treasury bonds) to a shock in the Hawkish Factor

We estimate a VAR model ordered as: { ,Hawkish Factor,Dovish Factor, Spread}. Our sample analysis begins in 1984Q1 and ends in 2017Q2. The Factors comes from Methodology 3. We select the most predictive Hawkish and Dovish bigrams using Elastic Net, then we use PCA to compute a set of hawkish and Dovish Factors. Next, we pick the first Hawkish Factor and the first Dovish Factor. The spread is the difference between the 10-year treasury yield and 3-month treasury yield (T10Y3M). In ( ) we have the FRED ticker of time series. Based on the AIC, we include 2 lags in the VAR.

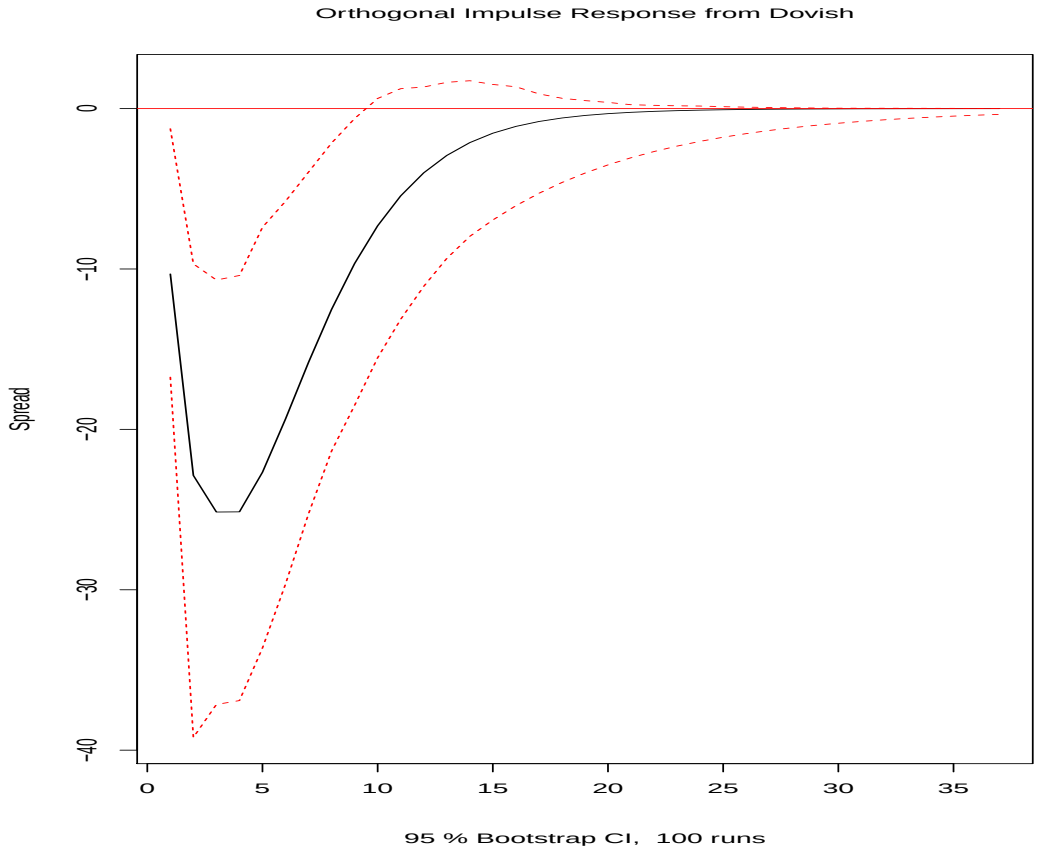

Figure 7: Response of the Yield Spread (10-year and 3-month Treasury bonds) to a shock in the Dovish Factor

We estimate a VAR model ordered as: { Hawkish Factor,Dovish Factor, Spread}. Our sample analysis begins in 1984Q1 and ends in 2017Q2. The Factors comes from Methodology 3. We select the most predictive Hawkish and Dovish bigrams using Elastic Net, then we use PCA to compute a set of hawkish and Dovish Factors. Next, we pick the first Hawkish Factor and the first Dovish Factor.The spread is the difference between the 10-year treasury yield and 3-month treasury yield (T10Y3M). In ( ) we have the FRED ticker of time series. Based on the AIC, we include 2 lags in the VAR.