Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Words or numbers? Macroeconomic nowcasting with textual and macroeconomic data

Tingguo Zheng [a,b,c], Xinyue Fan [b], Wei Jin [d,*], Kuangnan Fang [c]

[a] Center for Macroeconomic Research, Xiamen University, China
[b] Wang Yanan Institute for Studies in Economics, Xiamen University, China
[c] Department of Statistics and Data Science, School of Economic, Xiamen University, China
[d] Penghua Fund Management Co., Ltd., China

ARTICLE INFO

Keywords:
Nowcasting
Textual data
Macroeconomic data
Machine learning
Latent Dirichlet allocation

ABSTRACT

This paper performs the nowcasting of GDP growth rate and inflation expectation in China with traditional macroeconomic and novel textual data estimated by the latent Dirichlet allocation (LDA) model. We combine the MIDAS model with various machine learning techniques to handle the mixed-frequency and high-dimensional problems. Our empirical findings are threefold. First, we collected 866234 articles published over 20 years of Chinese economic newspapers. We systemically decomposed the textual data into news attention time series, which provide narrative descriptions of the economic and social conditions. Second, news attention data can provide similar or even better precision for nowcast, especially for inflation expectation compared with traditional macroeconomic data. Random forest delivers the most accurate forecast among the three machine learning methods, even for longer horizons. Thirdly, the most informative predictors for the nowcast align with existing literature, and news attention variables provide narrative realism for the forecast targets.

1. Introduction

Significant advancements in computing technology has enabled economists to handle large and complex data over the past decades. The rich information in these data are important for macroeconomic nowcasting (Bok et al., 2018; Kim & Swanson, 2018). Since most quarterly-released macroeconomic data in China are published with a half-month delay, subject to several revisions after the first release, an accurate early estimate will provide valuable insight for monitoring the state of the economy.

Nowcast, a terminology borrowed from meteorology, is intrinsically the short-term forecast of low-frequency data with predictors of various frequencies (Bańbura et al., 2013; Giannone et al., 2008). The key to obtaining accurate nowcasting results is properly handling mixed-frequency data and efficiently exploiting the information content embedded in big data. In this paper, we employed traditional monthly macroeconomic and novel news attention data derived from textual data in Chinese business-related newspapers to evaluate their performance in macroeconomic nowcasting.

As mentioned, two main impediments in macroeconomic nowcasting are handling the mixed-frequency data problem and efficiently incorporating high-dimensional predictors. Firstly, to address the mixed-frequency data problems, some studies, such as (Bańbura et al., 2013; Giannone et al., 2008; Mariano & Murasawa, 2003, 2010; Schorfheide & Song, 2015), treat the low-frequency data as high-frequency data with periodic missing values, and estimate the latent low-frequency data with Kalman filters. Other researchers, such as Andreou et al. (2010), Babii et al. (2021), Clements and Galvão (2008), Foroni

[*] Corresponding author.
E-mail addresses: zhengtg@xmu.edu.cn (T. Zheng), xinyue@stu.xmu.edu.cn (X. Fan), jinwei@phfund.com.cn (W. Jin), xmufkn@xmu.edu.cn (K. Fang).

https://doi.org/10.1016/j.ijforecast.2023.05.006
0169-2070/

et al. (2015), Ghysels et al. (2007), Kuzin et al. (2012), applied the mixed-frequency data sampling (MIDAS) method to mitigate the frequency mismatch problem. The MIDAS model is a reduced-form method that aggregates the high-frequency data with certain polynomial functions (such as the exponential Almon or Beta function) to transform the high-frequency data accordingly, resulting in a regression problem estimated with the nonlinear least square (NLS) method. Secondly, to efficiently incorporate the rich information content in high-dimensional predictors, many researchers have attempted to apply Bayesian shrinkage techniques (Giannone et al., 2021; Mogliani & Simoni, 2021) or machine learning methods (Medeiros et al., 2021; Siliverstovs, 2017) to reduce the number of parameters in the model. Compared with the mixed-frequency state-space model, the MIDAS model is more suitable for incorporating high-dimensional data. The unrestricted-MIDAS (U-MIDAS) model proposed by Foroni et al. (2015) treats each lag variable with equal weight. The model is linear and can be estimated with OLS, which is more efficient and stable than NLS. Besides its simplicity and efficiency, the U-MIDAS model with many predictors can be estimated with state-of-art machine learning methods.

Nevertheless, one shortcoming of U-MIDAS is that it assigns equal weights to different lags of high frequency. Still, the more recent data release is more informative for forecasting the near future. To deal with the inappropriate weights in U-MIDAS, Babii et al. (2021) uses the orthogonal polynomials to aggregate the different lags of high frequency data. Orthogonal polynomial functions are parameter-free and possess flexible shapes, which preserve the realistic weighting scheme while preventing the usage of nonlinear estimation techniques.

So far, nowcasting research has mainly relied on traditional macroeconomic or financial time series, such as industrial production or term spread in the treasury market. One prominent example of nowcasting under a big-data environment is performed by the Federal Reserve Bank of New York (Bok et al., 2018), which uses 36 predictors with different frequencies in their dynamic factor model. However, traditional macroeconomic data suffered from publication lags and data revisions. For example, the first release of monthly data is usually arranged 15 days after the month-end and is often followed by several data revisions. In the era of big data, nontraditional data showed great potential for nowcasting, such as textual (Bybee et al., 2021; Ellingsen et al., 2021; Thorsrud, 2018), GPS tracking (Moriwaki, 2020) or payment data (Barnett et al., 2016; Galbraith & Tkacz, 2018). These alternative data are immune from publication lag and thus more timely than traditional macroeconomic data. Including alternative data will tremendously increase the model's flexibility but bring extra obstacles for model estimation and evaluation.

In this paper, we perform nowcasts of two key low-frequency indicators, GDP growth rate and inflation expectations in China. This is performed using traditional macroeconomic and novel news attention data derived from newspaper textual data and by combining the MIDAS model with state-of-art machine learning techniques to handle the mixed-frequency high-dimensional data. Our contributions are threefold.

We are the first to collect 20 years of full-text data from three leading Chinese business-related newspapers. We estimate an LDA model to decompose the unstructured textual data into structured multivariate time series. The estimated news attention time series is the narrative reflection of the real world, and the most significant events can cause changes in news attention time series. Unlike traditional macroeconomic data, news attention data is not plagued by publication lag. These news attention data will facilitate further studies in the macroeconomics and financial fields.

Secondly, we combine the MIDAS weighting scheme in Babii et al. (2021) and three machine learning methods, i.e., sparse-group LASSO, random forest regression, and principal component regression. Since we have many predictors with 120 news attention variables and 159 macroeconomic variables, the MIDAS models suffer from the curse of dimensionality. We adopted the weighting scheme in Babii et al. (2021) enables us to include more predictors while preventing over-fitting. The nowcasting results indicate that macroeconomic and news attention data showed different predictive power for GDP growth and inflation expectations. Specifically, news attention yields comparable precision for nowcasting GDP growth compared with macroeconomic data, but it is much more accurate than macroeconomic data when nowcasting inflation expectation. We conjecture that the news attention data is more influential for individuals and thus affect their expectation and then decisions (Shiller, 2017) since the inflation expectation is a survey-based measure. The conditional predictive ability test confirmed that news attention data dominate macroeconomic data in turbulent periods (high economic policy uncertainty, low coincident index, and extreme inflation). For the three machine learning methods, random forest performs best for news attention data, and principal component regression performs better when combining macroeconomic data due to its strong factor structure (Stock & Watson, 2002). The excellent performance of random forest regression can be attributed to its ability to allow nonlinearity and ensemble schemes, thus more flexible than the linear models.

Thirdly, when exploiting which predictors are critical for nowcasting, we find that news attention and macroeconomic data deliver interpretable results. For example, when nowcasting GDP growth, macroeconomic data in "Output & Income" and "Financial Markets" categories and news attention data in "Macroeconomic" and "Policies & Reformation" categories showed strong predictive power. As for nowcasting inflation expectation, macroeconomic variables in the "Prices" and "Consumption" categories and news attention variables in "Civil Life" and "Consumption" exhibit unneglectable information contents. Unsurprisingly, macroeconomic variables possess predictive power for GDP growth and inflation expectation. Still, news attention variables also stand out when we include the whole data set (both macroeconomic and news attention data). This finding implies that the news attention data contain incremental information beyond macroeconomic variables; it should be a complementary but not substitution of traditional macroeconomic data.

In addition, our study has some connections with Ellingsen et al. (2021), but there are also noticeable differences. Firstly, our forecast targets include quarterly "hard" and "soft" indicators, which helps to compare the news media's impact on both indicators. And the low-frequency nature of our forecast targets makes it more sensible and plausible in mixed-frequency nowcasting. Secondly, the important topics in the out-of-sample are quite different. In our results, the most important news topic attention data for GDP nowcasts align more with the actual Chinese context. Thirdly, as in their work, we use the sparse-group-LASSO method instead of plain vanilla LASSO. Our approach guarantees variable selection consistency, and the weighting function in MIDAS models is more realistic than in U-MIDAS. Finally, we provide solid evidence of the conditional predictive power of news attention data. Their analysis was conducted from a descriptive perspective, and recursive model combinations may be slow to adapt to sudden changes in economic conditions, especially during COVID-19. We can address this issue with the help of conditional superior predictive ability (CSPA, Li et al., 2022).

The rest of this paper is organized as follows: Section 2 introduces the text and macroeconomic data we used in nowcasting and briefly discusses their difference. Section 3 briefly introduces the MIDAS model and machine learning techniques we used in this paper. Section 4 introduces the nowcasting experiment settings and presents the nowcasting results. Section 5 discusses the importance of each predictor in nowcasting procedures and compares the difference between macro data and news attention data for nowcasting different macroeconomic data. Section 6 concludes the paper.

2. Data

This section discussed the data, both news and macro, we used in the nowcasting task. We describe how to transform the unstructured, high-dimensional text data into structured and relatively low-dimensional time series data, the collection of macroeconomic data, and the nexus between news data and macroeconomic data, representing soft and hard information.

2.1. News attention data

We collect the full-text data of 866234 news articles published in the three most prevailing business-focused newspapers written in Chinese: *Economic Daily*, *The Economic Observers* and *21st Century Business Herald* for the period from January 2000 to June 2021. These three newspapers are the leading publishers of business-related news.[1]

Text data is infamous for its unstructured and high-dimensional nature, preventing its direct use in econometric modeling. Several procedures must be implemented before any further explorations. First, we prepossess the textual data by segmentation,[2] remove the stop-words and use TF-IDF to filter out uninformative terms to enhance the interpretability of LDA results.[3] We do not elaborate on these procedures to save space since they are standard in NLP literature. Interested readers can find the details of the preprocessing procedures in Appendix A. Besides the cleaned textual data, we need to pin down the number of topics K in the corpus. The optimal K can be determined by cross-validation or perplexity score of the model (Blei et al., 2003; Bybee et al., 2021). We choose a wide range of possible topic numbers from 50 to 150 with increments of 10, estimate these models, and choose $K = 120$ as the best model since it minimizes the perplexity score.

After the cleaning procedures, we fit the cleaned textual data with the latent Dirichlet allocation (LDA hereafter) model. The LDA model was proposed by Blei et al. (2003) as an unsupervised natural language processing apparatus that shared many similarities with the dynamic factor model, the workhorse for macroeconometrics (Stock & Watson, 2016). LDA clusters the words and phrases based on their word counts and summarizes the semantic content into a handful of interpretable topics. There are two outputs of the LDA model: topic-word distribution $\boldsymbol{\varphi}$ and document-topic distribution $\boldsymbol{\theta}$. The first distribution describes what a particular topic is talking about. Every unique word and phrase is assigned a numerical weight to measure the relative importance of describing this topic. Besides, this distribution can be useful to manually label the topics.[4] The second output of the LDA model is the document's topic distribution, which is a K-dimensional vector with each element that measures the editors' attention weight allocated to each topic; a larger value means this topic draws more attention in that article. After estimating the LDA model, we

[1] One may argue that the prosperity of mobile Internet and self-media in China after 2010 may abate the influence of traditional newspapers. We consider this not problematic for our application because, first, compared to self-media, traditional newspapers are restricted by higher moral and legal standards, and the information published by newspapers is more authentic than self-media. Second, the material information is published both in newspapers and self-media. In other words, self-media is a supplement, not a replacement for traditional newspapers. Thirdly, readers might be concerned that media bias or media censorship in the newspaper will undermine the empirical results. Huang and Luk (2020) discussed this potential problem by calculating the sentiment distribution toward certain events in the economic field conveyed by newspapers published in mainland China (faced with media censorship) and newspapers published in Hong Kong (free from media censorship). They conclude that media censorship does not significantly affect the subjectivity of news reports.

[2] Unlike English-based textual data, Chinese words and phrases are separated by blanks. The extra step of segmentation is necessary. We use "jieba", a Python package and a customized dictionary specially designed for Chinese business-related newspapers to accomplish this task.

[3] We filter out the words whose TF-IDF is less than the 20% quantile in the corpus. Details about the preprocessing and construction of customized dictionaries can be found in Appendix A.1. Note that the TF-IDF is used for term filtering to enhance the interpretability of LDA results, but the input for the LDA model is still the document-term matrix (Doc2Bow).

[4] Since LDA is unsupervised, it will not provide the label of topics, but the labels can be summarized from the word distribution of topics (Bybee et al., 2021; Thorsrud, 2018).

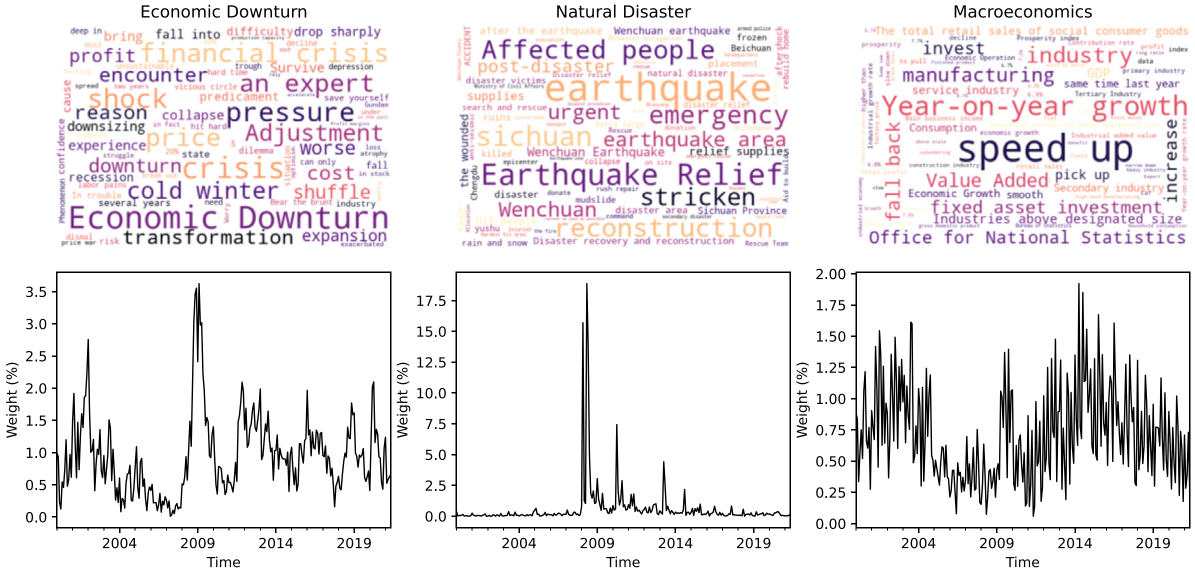


Fig. 1. Word distribution of topics and monthly topic distributions.
Note: This figure depicts three out of 120 topics given by the estimated LDA model. The upper row is the word distribution of topics (The words are translated into English), while the lower row is the topic distribution in the distribution in monthly frequency.

aggregate all the articles in each day as a new document and calculate its topic distribution. Since we specify the nowcasting model in monthly frequency, we calculate the monthly news attention data by averaging the daily news attention data within each month. Further details about the structure of the LDA model and estimation procedures can be found in Appendix A.1.

Fig. 1 provides a glance at the result given by the LDA model; the words are translated into English. We chose three out of the 120 topics estimated from the LDA model. The upper row represents the word distribution of topics, while the lower row represents its distribution in each month.[5] The first topic, "Crisis" spiked during several notable periods, such as the Great Financial Crisis in 2008 and the outbreak of the COVID-19 pandemic at the end of 2019. Words like "Downturn", "Crisis", "Decline", and "Shocks" possess large weights in the word distribution. The second one, "Natural Disaster" featuring "Disaster Area", "Earthquake", and "Rescue" as keywords, remains flat during routine periods but soars during the great earthquake in Sichuan Province, China, among several following rare disasters. The third topic, labeled as "Macroeconomics", loaded heavily on keywords like "Growth", "Year-on-Year Growth", and "Fixed Asset Investment", exhibits an apparent seasonal pattern. The news attention on this topic is likely to increase when the macroeconomic data is released.

Following Bybee et al. (2021) and Ellingsen et al. (2021), we perform Agglomerative Hierarchical Clustering (AHC hereafter) on the word distribution of topics φ to group 120 topics into 15 categories. AHC computes the pairwise distances between each topic and sequentially combines two topics into a high-order one until only one topic is left. Appendix A.3 shows the group structure and detailed descriptions.

2.2. Macroeconomic data and nowcast targets

Following McCracken and Ng (2016), we collect 159 monthly macroeconomic time series from the eight categories as (i) labor; (ii) real estate; (iii)income and output; (iv) consumption; (v) money market; (vi) interest and exchange rates; (vii) price; (viii) financial market and (ix) others. Since there is no comprehensive real-time macroeconomic data vintage in China, we only use the final vintage.[6] We transform all the macroeconomic data to achieve stationarity as suggested in McCracken and Ng (2016); the details about macroeconomic data and transformation methods can be found in Appendix B.

We choose GDP growth and inflation expectation as nowcast targets. GDP growth is the most critical indicator for monitoring the state of the national economy. The Department of Statistics in China releases the first estimate of the quarterly GDP level 15 days after the quarter ends. We use the year-on-year GDP growth rate divided by four as the average annual GDP growth rate, and data is obtained in the China Economic Information Network (CEIN) database. The inflation expectation data is quarterly published by People's Bank of China (PBC) based on the nationwide survey for depositors of commercial banks and released ten days after each quarter ends.[7]

[5] In the word cloud figure, the font size reflects the weight of each word; we choose the first 100 words with the largest weights to represent the distribution.

[6] Final vintage would not reflect the data revision, but we take the publication lags into consideration; that is, our method is pseudo-real-time.

[7] The questionnaires used by PBC request the depositors to answer the following question: "How do you feel about the price level over the next quarter?" with three options: "Go up", "Stay the same", or "Go down". PBC calculates the percentage of depositors choosing the first option as the estimate of inflation expectation.

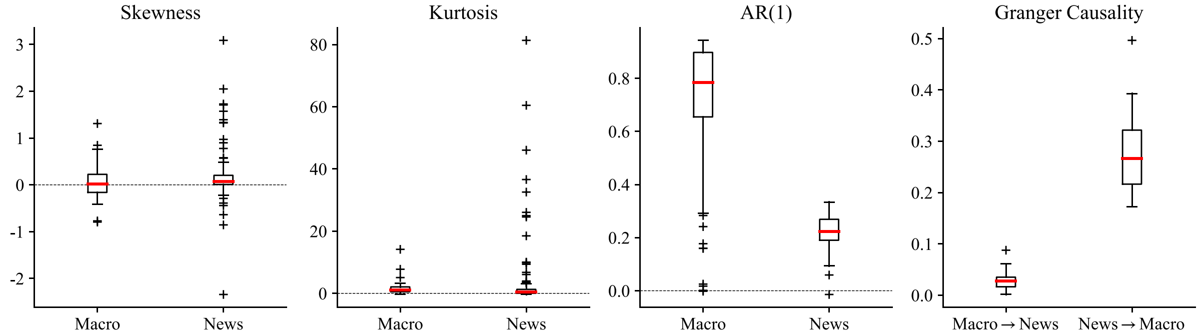


Fig. 2. Summary Statistics for Macro Data and News Attention Data.
Note: This figure reports skewness, excess kurtosis, first-order autocorrelation coefficient, and percentage of Granger causality for macro data and news attention data. The top and bottom edges of the box illustrate the 25th and 75th percentiles, respectively. The whiskers extend beyond the box, excluding the outliers, while the "+" symbols indicate the outliers. The red line in the middle of the boxes shows the median value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The inflation expectation data can be found on the official website of PBC.

2.3. The nexus between news data and macroeconomic data

Unlike macro data, which has been studied extensively in the literature, news attention data is more alien to researchers. Before further studies, we first explore the statistical properties of news attention data and compare them with macroeconomic data.

Fig. 2 visualizes the summary statistics for standardized macroeconomic and news attention data. We compute the skewness, excess kurtosis, and first-order auto-correlation coefficients and perform a pairwise Granger causality test between news attention and macroeconomic data.[8] From the figure, we can observe three differences between the two datasets. First, news attention data exhibit more dispersed skewness and larger excess kurtosis than macro data, meaning that news attention data has many extreme values and is thus noisier than macroeconomic data. Secondly, macro data have larger auto-correlation coefficients than news attention data on average, meaning the persistence is stronger for macro data. This finding is intuitive because news attention data is more timely than macroeconomic data. It can react quickly to exogenous shocks. Thirdly, news attention data Granger causes more macro data on average, which shows the leading property for news attention data.

Fig. 3 visualizes the factor structure in both macroeconomic and news attention data. We perform principal component analysis on both datasets and calculate the percentage of variance explained by each principal component. From the figure, we can see that the first five principal components (PCs) of macroeconomics explain more than 50% of the total variation, and the scree plot exhibits an exponentially decaying pattern, meaning that the comovement phenomenon is more evident for macroeconomic data (Stock & Watson, 2002). Compared with macroeconomic data, the factor structure is weaker for news attention data, and the scree plot is hyperbolically decreasing; the first five PCs only explained about 30% of the total variation. These results indicate that the information in news attention data is more idiosyncratic than macroeconomic data.

3. Methodology

This section briefly introduces the econometric and machine learning methods we used in the nowcasting exercises. We combine the MIDAS technique and three popular machine learning methods to handle the mixed-frequency and high-dimensional problems.

3.1. Mixed-frequency data sampling (MIDAS) model

Pioneered by Ghysels et al. (2004), the MIDAS model has become the paradigm for macroeconomic nowcasting. The MIDAS model aggregates the high-frequency predictors with certain weighting functions to predict the low-frequency variables. The MIDAS specification avoids computational demanding Kalman filter procedure and possible misspecification of the joint distribution, which can be problematic in the mixed-frequency state-space model.

The key element in MIDAS modeling is the weighting function. Popular choices include the exponential Almon function (Ghysels et al., 2007), Beta function (Andreou et al., 2010), and equal weighting (Foroni et al., 2015). Equal weighting schemes, also known as U-MIDAS, prevent the usage of nonlinear least square estimation and can be easily estimated by OLS. A weak spot for U-MIDAS is the unrealistic equally weighting scheme for more recent and relatively staled information. To preserve the linear property of U-MIDAS, Babii et al. (2021) employed parameter-free weighting functions to aggregate the high-frequency variables. They choose orthogonal polynomials, Legendre polynomials, for example, in a certain order to replace the exponential Almon function. We adopt this parametrization in our setting.

[8] The pairwise Granger causality tests are performed with bivariate VAR systems, whose lag order is chosen by Bayesian information criterion (BIC). We calculate the Wald-type Granger causality test statistics; if the test statistics are higher than the 5% critical value, Granger causality is confirmed. We take the average along the *cause* variables to compute the percentage of variables that can be Granger caused by each *cause* variables.

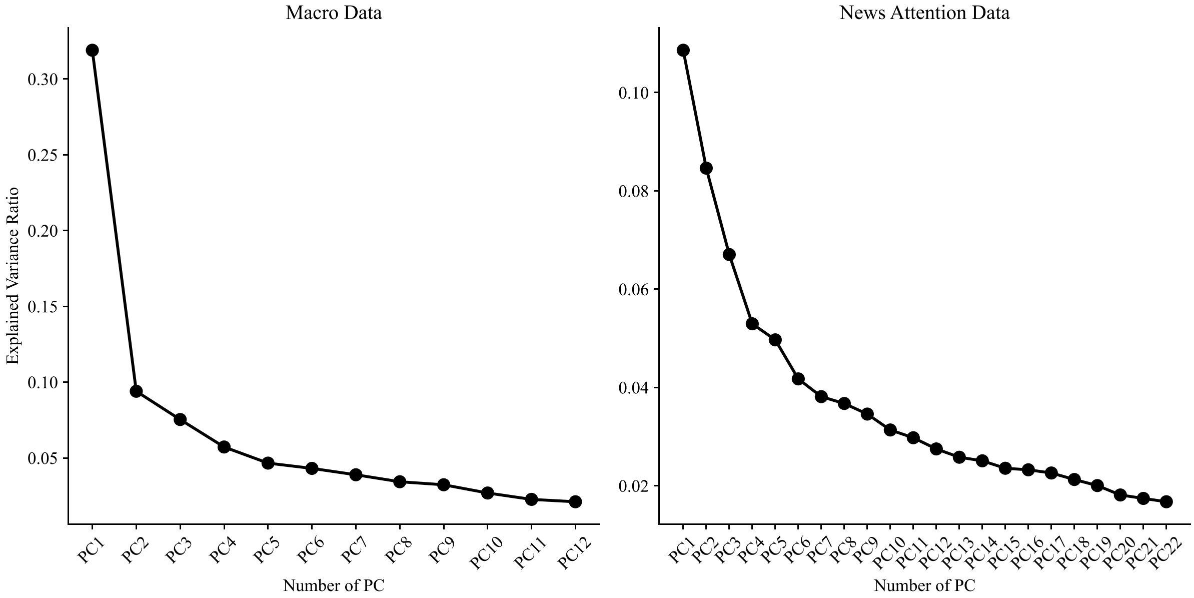


Fig. 3. Ratios of variance explained by each principal component of macroeconomic and news attention data.
Note: This figure demonstrates the variance-explained ratio by each component for macro data (left panel) and news attention data (right panel). The horizontal axis represents principal components, and the vertical axis represents each component's explained variance ratio.

Assuming that the nowcast target $\{y_t,\ t \in [T]\}$ is observed once in every period t, and there are K potential predictors, $\{x_{t-(j-1)/m,\ k},\ j \in [m],\ k \in [K],\ t \in [T]\}$, which can be observed m times in each period t,[9] where $[p] = \{1, 2, \ldots, p\}$ for $p \in \mathbf{N}$. The general model form for nowcasting can be expressed as:

$$y_t = g\left(\sum_{k=1}^{K} \psi(L^{1/m}; \beta_k) x_{t,k}\right) + \epsilon_t, \tag{1}$$

The lag polynomial can be written as:

$$\psi(L^{1/m}; \beta_k) x_{t,k} = \frac{1}{m} \sum_{j=1}^{m} \omega\left(\frac{j-1}{m}; \beta_k\right) x_{t-(j-1)/m,k}, \tag{2}$$

where β_k is a L—dimensional vector of coefficients with $L \leq m$ and $\omega : [0, 1] \times \mathbf{R}^L \to \mathbf{R}$ is some weighting function. The weighting function can be approximated as follows:

$$\omega(u; \beta_k) \approx \sum_{l=1}^{L} \beta_{k,l} w_l(u), \ \ u \in [0, 1], \tag{3}$$

where $\{w_l : l \in [L]\}$ is a collection of functions, called the dictionary. Babii et al. (2021) suggested that orthogonal polynomials reduce multi-collinearity and lead to better finite sample performance. In this paper, we use the Legendre polynomial as a weighting function. The first five orders of the Legendre polynomial can be written as:

$$w_l(x) = \begin{cases} 1 & l = 0 \\ x & l = 1 \\ \frac{1}{2}\left(3x^2 - 1\right) & l = 2 \\ \frac{1}{2}\left(5x^3 - 3x\right) & l = 3 \\ \frac{1}{8}\left(35x^4 - 30x^2 + 3\right) & l = 4 \\ \frac{1}{8}\left(63x^5 - 70x^2 + 15x\right) & l = 5 \end{cases} \tag{4}$$

Finally, let $\boldsymbol{y} = (y_1, \ldots, y_T)'$ be the predict target, and $\boldsymbol{X} = (Z_1 W, \ldots, Z_K W)$ is the aggregated high-frequency predictors, where $Z_k = \left(x_{k,t-(j-1)/m}\right)_{t\in[T],\ j\in[m]}$ is a $T \times m$ matrix of high-frequency predictors for $k \in [K]$, and $W = [w_l\left((j-1)/m\right)/m]_{j\in[m],l\in[L]}$ is an $m \times L$ matrix of weights. The general form of the nowcasting model can be written as

$$\boldsymbol{y} = g\left(\boldsymbol{X}\right) + \boldsymbol{\epsilon}. \tag{5}$$

where $g(\cdot)$ is a general function, which can be parametric or nonparametric.[10] More detail about the model can be found in Babii et al. (2021).

3.2. Machine learning methods for high-dimensional data

The inclusion of high-frequency predictors increases the dimension of data hugely. Under the big data setting, the number of parameters is larger than the effective sample size and even exceeds it. Due to the scarcity of macroeconomic data, we need an econometric model that can handle high-dimensional data. In this paper,

[9] In our application, $m = 3$. But the weighting scheme allows different frequencies in the predictors.

[10] When combined with sg-LASSO or PCR, the general functional form is (1) when combined with RF, the general functional form is (5) since it is nonparametric.

we focus on three canonical methods in machine learning, sparse-group LASSO, random forest regression, and principal component regression, representing variable selection, ensemble method, and dimension reduction methods. In this section, we provide some intuitive discussion for these methods; further details and tuning protocols can be found in Appendix C.

The Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani (1996) is one of the most widely applied variable selection techniques. LASSO penalizes the objective function with ℓ_1 norm of parameters, which will shrink the coefficient of irrelevant variables to exact zero, thus achieving variable selection. A well-known shortcoming of LASSO is that it only selects one variable among a group of correlated variables and drops the rest, which is more pronounced combined with U-MIDAS (Babii et al., 2021). This defect can be solved by introducing group structure into LASSO objective function, thus group LASSO (Yuan & Lin, 2006). Group LASSO shrinks the variables in the same group; if a group of variables is considered irrelevant, this group will be dropped altogether. A further extension of group LASSO is sparse group LASSO (sg-LASSO hereafter) proposed by Simon et al. (2013), which combines LASSO penalty and group LASSO penalty, sg-LASSO can introduce inter-group sparsity and intra-group sparsity thus enhances the interpretability of variable selection. Babii et al. (2021) proved the validity of sg-LASSO MIDAS estimator for α-mixing time series data in finite sample.

The sg-LASSO estimator $\hat{\boldsymbol{\beta}}$ solves the following problem:

$$\min_{\beta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + 2\lambda \left\{\alpha|\boldsymbol{\beta}|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_{2,1}\right\} \tag{6}$$

where $\|\boldsymbol{\beta}\|_{2,1} = \sum_{G\in\mathcal{G}} |\boldsymbol{\beta}_G|_2$ is the group LASSO penalty and $\mathcal{G}$ is the group structure specified by econometrician. The hyper-parameter λ controls the degree of overall shrinkage while α controls the weight of the LASSO penalty. Following Bai and Ng (2008), we choose λ and α by searching over grids of predefined hyper-parameters and find the combination of λ and α which minimizes the Bayesian information criterion (BIC).

The second method we adopted is random forest regression (RF hereafter) proposed by Breiman (2001). RF reduces the variance of regression trees by bootstrap aggregate (bagging) several regression trees. The regression tree is a nonparametric model approximating an unknown nonlinear function with the local average. As the base learner in RF, the regression tree is building up by successively partitioning the covariate space (Breiman, 1996). The splitting point is chosen at each splitting step based on minimizing the mean squared error. The partition is stopped until some criteria are met.[11] The regression tree's terminal node (also known as leaves) represents the dependent data's fitted value.

Since the regression tree model tends to overfit, random forest ensemble several regression trees specified on a bootstrap sample of the original data. Since our data are time series, we choose block bootstrap protocol. For each sample, $b \in [B]$, a regression tree with K_b leaves is estimated for a randomly selected subset of the original regressors. Then we take the average of the forecast produced by each tree as the final forecast. Following Medeiros et al. (2021), we halt the partitioning process until only five observations are in each leaf. The bootstrap proportion at each split step is set to 1/3. The number of the bootstrap samples $B = 500$. We also try several alternative combinations of these hyper-parameters, and the result remains reasonably stable.

The last machine learning technique for high dimensional data is principal components regression (PCR hereafter). The intuition of PCR is to regress the dependent variable on a handful of common components instead of the whole predictors to avoid overfitting. The common components capture the comovements in the predictors and filter out the idiosyncratic noises. Stock and Watson (2002) applied PCR on a large panel of 215 macroeconomic time series in the US for forecasting.

Since we are dealing with mixed-frequency data, the aggregated predictor matrix in Eq. (1) contains lags of predictors. Taking principal components on $\boldsymbol{X}$ is not appropriate. Instead, following Marcellino and Schumacher (2010), we conduct PCR-MIDAS in a two-step manner. First, we perform eigenvalue decomposition of the covariance matrix of the standardized high-frequency predictors and calculate the principal components. The optimal number of components is determined by the information criterion proposed by Bai and Ng (2008). Second, with estimated principal components F_t, the model can be written as:

$$y_t = \beta_0 + \sum_{p=1}^{P} \frac{1}{m} \sum_{j=1}^{m} \sum_{l=1}^{L} \beta_{p,l} w_l \left(\frac{j-1}{m}\right) F_{t-(j-1)/m,p} + \epsilon_t \tag{7}$$

where P is the number of principal components included, and $F_{t,p}$ is the pth principal component at time t. Coefficients in Eq. (7), β_0 and $\{\beta_{p,l}\}_{p\in[P],\ l\in[L]}$ for can be estimated with ordinary least square method.

4. Nowcast results

In this section, we first introduce the nowcast setups and then present the nowcasting results both unconditionally and conditionally.

4.1. Nowcasting setup

The nowcast exercises are performed in a recursive fashion (expanding window). We use the samples between January 2000 and December 2010 as an initial estimation and add one-quarter of observation at each step. The first forecast is made for 2011Q1 and results in 42 out-of-sample forecasts. We choose three nowcast horizons and two forecast horizons. In nowcasting, horizons $h = 0, 1$ or 2 stand for one-, two-, or three-month-ahead estimates. Furthermore, we also conduct long-horizon forecasts with $h = 3$ and $h = 6$; they

[11] The criteria may be only one observation left on the terminal nodes or exceeding the prespecified depth (Hastie et al., 2009).

are one- and two-quarter ahead forecasts.[12] We use the autoregressive (AR) model as a benchmark, whose order is determined by BIC in real-time.

Due to the publication lag for macro data, we add one more lag to macro data except for data from the financial market and interest and exchange rate categories.[13] On the contrary, news attention data are immune from publication lag, so we do not add additional lag to news attention data. To avoid looking-ahead bias, we use the text data from January 2000 to December 2010 to estimate the LDA model, and once the model is estimated, we can perform the inference for the text data in the remaining samples. Since estimating the LDA model is quite time-consuming, we do not recursively update the LDA model as in Ellingsen et al. (2021).

4.2. Unconditional evaluation

Firstly, we present the root of mean squared error (RMSE) for the nowcast results. Since we have two sets of predictors, three models, and five forecasting horizons, we will yield 30 RMSE. Fig. 4 illustrates the comparison of forecasting performances between news attention data and macroeconomic data. The horizontal and vertical axes represent the RMSE of forecasts produced by macroeconomic and news attention data, respectively. Different markers represent three different models. The dotted line across the diagonal is the equal-RMSE line. Markers located below the line indicate news attention data achieve more precious forecasts for particular methods and horizons and vice-versa.[14]

There are several noteworthy findings in Fig. 4. First, when nowcasting GDP growth rate, two sets of predictors achieve similar accuracy since they are located near the 45° line. However, news attention data performed much better when nowcasting inflation expectations. Since inflation expectation is based on the survey of households who use news media as information delegation (Nimark & Pitschner, 2019), it is not surprising that news attention data achieve superior performances when nowcasting inflation expectations since the households are more inclined to be affected by information from the newspaper. Secondly, forecasts produced by the RF cluster are in the northwest corner of both panels, indicating that the forecast performances are better than the two methods. The nonlinearity and bagging lend flexibility and guarded against overfitting are the main reasons for its excellent performance (Medeiros et al., 2021). Thirdly, principal components regression works better with macroeconomic data since almost all its forecasts are above the 45° line. This can be supported by the stronger factor structure in macroeconomic data compared to news attention data.

In addition, we calculate the mean RMSE reduction across three models when new monthly predictors become available. We calculate the decrease of RMSE when we shorten the horizon for nowcasts by one month. Fig. 5 presents the results. For example, the two leftmost bars in the left panel of Fig. 5 represent the RMSE reduction when we can use the information released in the first month in each quarter (moving from $h = 3$ to $h = 2$). The figure shows that when new monthly macroeconomic variables became available, the RMSE reduction was much larger than news attention data. It is reasonable that macroeconomic data is more material and accumulates richer information than the already wide-spreading news attention data.

Finally, we calculate the cumulative differences of square prediction error (CDSPE hereafter) between macroeconomic and news attention data for each model. The CDSPE can be calculated as follows:

$$\mathrm{CDSPE}_t = \sum_{\tau=1}^{t} \left[e^2_{\tau,\mathrm{macro},i} - e^2_{\tau,\mathrm{news},i}\right] \tag{8}$$

where $e^2_{\tau,\mathrm{macro}}$ and $e^2_{\tau,\mathrm{news}}$ are the squared forecasting error using macroeconomic data and news attention data at time τ with model i ($i =$ sg-LASSO, PCR or RF) respectively.

The diagram of CDSPE is an informative tool to trace the forecasting performance along the time dimension. When the CDSPE curve shows an upward trend, it implies that news attention data is getting more accurate than macroeconomic data since it produces smaller squared errors. Fig. 6 visualizes the time trend of CDSPE three models. For GDP growth, we can observe a significant upward spike at the end of 2019, which coincides with the outbreak of the COVID-19 pandemic. Since the lockdown policy in China impedes routine economic activities and heightens uncertainty, the superiority of news attention data becomes obvious. After the initial outbreak of COVID-19 in Wuhan province, the Chinese government enforced effective strategies against the virus's spread immediately, and the CDSPE curve eventually went down. Furthermore, the magnitude of jumps is different among the three models. From the left panel of Fig. 6, we can see that RF-MIDAS had the largest jump at the beginning of the COVID-19 pandemic, while PCR-MIDAS had the smallest increase in CDSPE. This result indicates that the nonlinearity may be more important during turbulent periods. As for inflation expectation, the CDSPE curve shows an upward trend most times because news attention data is more effective when nowcasting inflation expectation as we already demonstrated in Fig. 4.

To prove the value of big data and machine learning algorithms, we perform the forecasting comparison test of Diebold and Mariano (2002) and choose the autoregressive (AR) model as the benchmark. Furthermore, we merged macroeconomic and news attention data to a large panel and then repeated the nowcasting exercises.

[12] For example, when we nowcast the GDP growth rate of the first quarter of 2011 and $h = 0$ ($h = 1$), we use the news attention variable and financial variables until March 2011 (February 2011), macroeconomic variable until February 2011 (January 2011).

[13] The typical schedule for macroeconomic data publication in China is the 15th day of the following month. For example, the industrial production for March 2011 was published on April 15th, 2011. Due to this convention, when forecasting the GDP growth or inflation expectation for 2011Q1, the latest available macro data are up to February 2011.

[14] Since the COVID-19 pandemic caused severe disruption for the macro economy, we repeat the out-of-sample forecast based on the pre-COVID period (2010Q1 to 2019Q4) and perform the out-of-sample test. Details can be found in Section D.1.

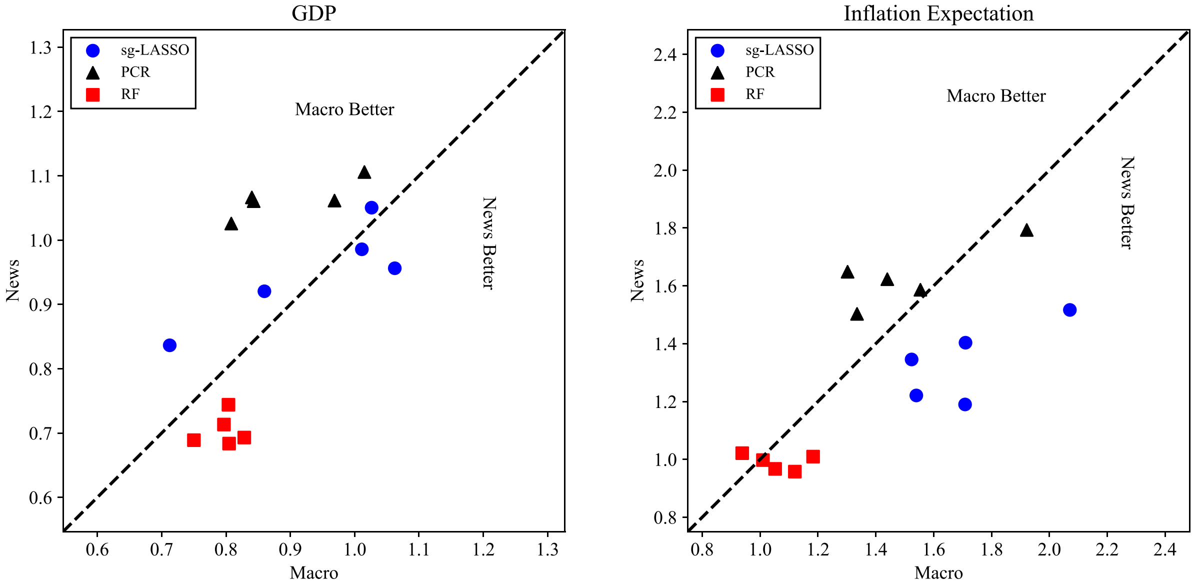


Fig. 4. Forecasting comparison.
Note: This figure presents the news attention and macro data forecast performance. The horizontal axis indicates the RMSE of forecasts using news attention data, while the vertical axis indicates the RMSE of forecasts using macro data and vice versa. The dashed line across the diagonal indicates the equal RMSE for two forecasts. Different symbols indicate different forecasting methods.

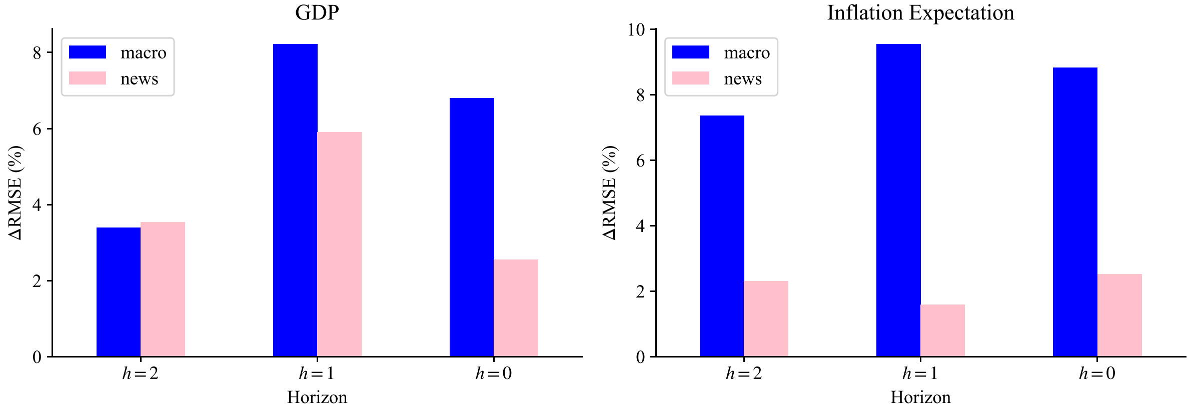


Fig. 5. Accuracy gain from intra-quarter information release.
Note: This figure presents the reduction of RMSE when new monthly data become available. For example, when we move from $h = 3$ to $h = 2$, the data for the first month in the quarter is known as ready to use for nowcasting. The percentage of RMSE decreased can be calculated.

We labeled the merged data set as All-In-One (AIO) data and also performed the Diebold–Mariano test against AR benchmark.

Table 1 reports the results of the Diebold–Mariano test. The numbers in the table are test statistics; the "***", "**", and "*" in the superscripts stand for the rejection of the null hypothesis of equal predictive power at the significant level of 1%, 5%, and 10%, respectively. The following facts emerged from the table: (1) For GDP growth, almost all the model-predictor combinations can beat the AR benchmark except for PCR-News forecasts, which can be attributed to the weak factor structure demonstrated in Fig. 3. Sg-LASSO-Macro and PCR-AIO PCR-News forecasts failed to beat the simple AR model for inflation expectations. (2) The RF model delivers more accurate results than two other machine learning methods, especially for longer horizons (one and two quarters ahead forecasts). The bagging procedure and nonlinearity of the RF model are the key reasons behind this (Medeiros et al., 2021). (3) News attention data provide incremental information when nowcasting GDP when using sg-LASSO and RF models but failed to enhance the more extended horizon forecasts, which can be concluded by comparing AIO-based and Macro-based forecasts (first two rows for each method). Compared with macro-based forecasts, AIO-based forecasts generally produced larger DM statistics for $h = 0$, 1 and 2, and this means news attention data

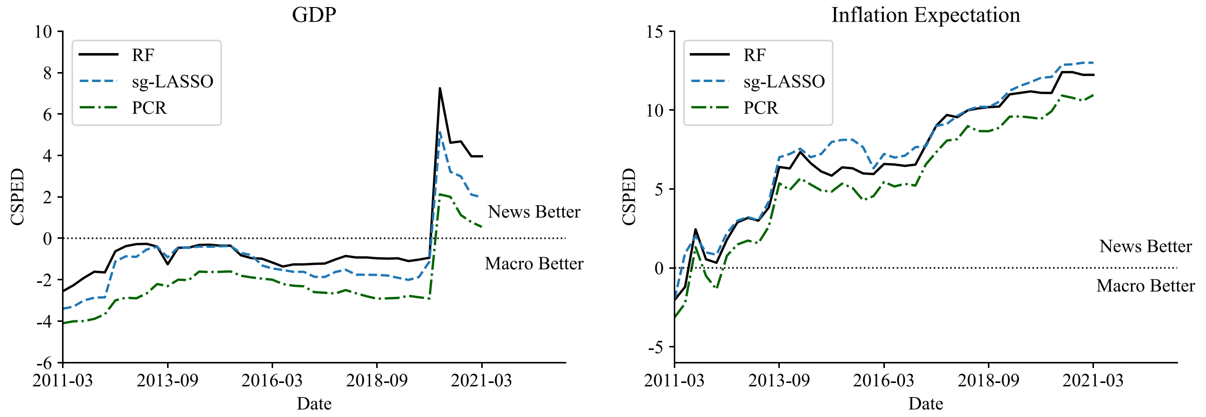


Fig. 6. Cumulative differences of squared prediction error (CDSPE).
Note: This figure presents cumulative squared prediction error (CDSPE) differences between macroeconomic and news attention data with different models. An upward trend in CDSPE indicates better forecasting power of news attention data and vice versa.

Table 1
Diebold–Mariano test for nowcasting results.

Method		Panel A: GDP				
		$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 6$
sg-LASSO	AIO	3.496***	2.920***	2.227**	2.042**	1.416
	Macro	3.118***	2.633***	2.258**	2.286**	1.592
	News	2.193**	2.012**	1.579	2.352**	1.439
PCR	AIO	2.304**	2.503**	2.483**	1.549	0.163
	Macro	2.495**	1.897*	2.147**	1.649*	0.972
	News	1.139	0.516	0.922	0.276	0.153
RF	AIO	2.313**	1.954*	1.805*	1.660*	1.514
	Macro	2.929***	2.869***	2.837***	2.797***	3.005***
	News	3.346***	3.009***	3.015***	2.915***	2.200**
Method		Panel B: Inflation Expectation				
		$h = 0$	$h = 1$	$h = 2$	$h = 3$	$h = 6$
sg-LASSO	AIO	1.539	0.599	2.147**	−0.280	−0.240
	Macro	−0.901	−1.015	−1.423	−1.617	−2.453**
	News	4.153***	3.813***	3.854***	3.580***	3.543***
PCR	AIO	−0.559	−1.516	−2.294**	−3.781***	−2.837***
	Macro	2.045**	1.895*	0.798	0.391	0.266
	News	−0.468	−0.382	−1.585	−1.469	−2.318**
RF	AIO	5.825***	5.958***	6.283***	6.848***	5.020***
	Macro	5.321***	4.356***	4.061***	3.656***	2.663***
	News	5.104***	5.281***	5.833***	5.919***	6.125***

Notes: This table reports the results of Diebold and Mariano (2002) test for forecasting GDP growth (Panel A) and inflation expectation (Panel B) with different methods, predictors and horizons. "sg-LASSO", "PCR" and "RF" denotes sparse-group LASSO, principal components regression and random forest with mixed-frequency data structures. Forecast horizons range from nowcast ($h = 0$) to two-quarters ahead forecast ($h = 6$). Predictors are macro data (Macro), news attention data (News) and the merged dataset of macro and news attention data (AIO). Numbers in the table are the value of Diebold–Mariano test statistics with benchmark as AR model.

***Stand for the rejection of null hypothesis of equal predictive power at the significant level of 1%.

**Stand for the rejection of null hypothesis of equal predictive power at the significant level of 5%.

*Stand for the rejection of null hypothesis of equal predictive power at the significant level of 10%.

can improve the nowcasting performance but provide less noticeable improvement during the long run.

Since the Diebold–Mariano test is only capable of pairwise comparison, we further validate our results by conducting the model confidence set (MCS) test proposed by Hansen et al. (2011), designed for multiple model comparison. Table 2 reports the results. The numbers in the table are the p-values for the MCS test, which can be interpreted as the probability that a certain model is included in a superior model set, i.e., the better model has a higher p-value. We construct MCS for each horizon based on the t_{max} statistics described in Hansen et al. (2011).

The results of MCS further validate the former results. Firstly, at each horizon, RF models achieve larger p-values for most cases, especially for inflation expectations. The p-values for using AIO and news attention data are approaching 1, which indicates it is the best model for forecasting inflation expectations even in the longer horizons. Secondly, for forecasting GDP growth with an

Table 2
Model Confidence Set for nowcasting results.

Horizon	Panel A: GDP								
	AIO			Macro			News		
	sg-LASSO	PCR	RF	sg-LASSO	PCR	RF	sg-LASSO	PCR	RF
$h = 0$	0.89	0.11	0.02	**0.89**	0.18	**0.89**	0.55	0.00	**0.89**
$h = 1$	0.52	0.06	0.00	0.34	**0.89**	0.18	0.15	0.00	0.72
$h = 2$	**0.50**	0.06	0.01	0.47	0.18	0.46	0.17	0.00	0.47
$h = 3$	0.11	0.06	0.01	0.18	0.51	0.33	0.28	0.00	**0.56**
$h = 6$	0.18	0.03	0.00	0.20	0.04	0.09	0.06	0.00	**0.89**
Horizon	Panel B: Inflation Expectation								
	AIO			Macro			News		
	sg-LASSO	PCR	RF	sg-LASSO	PCR	RF	sg-LASSO	PCR	RF
$h = 0$	0.10	0.18	0.99	0.01	0.83	**1.00**	0.35	0.00	**1.00**
$h = 1$	0.11	0.23	**1.00**	0.01	0.76	0.62	0.97	0.12	0.80
$h = 2$	0.82	0.07	**1.00**	0.00	0.37	0.27	0.72	0.01	**1.00**
$h = 3$	0.06	0.00	**1.00**	0.01	0.23	0.33	0.20	0.03	**1.00**
$h = 6$	0.30	0.36	0.91	0.06	0.14	0.03	0.12	0.00	**1.00**

Notes: This table reports the results of model confidence set test of Hansen et al. (2011) for forecasting GDP growth (Panel A) and inflation expectation (Panel B) for different models, horizons and predictors. "sg-LASSO", "PCR" and "RF" denotes sparse-group LASSO, principal components regression and random forest with mixed-frequency data structures. Forecast horizons range from nowcast ($h = 0$) to two-quarters ahead forecast ($h = 6$). Predictors are macro data (Macro), news attention data (News) and the All-In-One data (AIO) which is the merged dataset of macro and news attention data (AIO). The numbers in the table are the $p-$ values of MCS tests, larger value means the forecasts are more likely to be included in the superior forecast set. The largest value in each row is marked as bold.

AIO predictor, sg-LASSO produces larger p-values than the other methods. Since there are 265 variables in the AIO dataset, sparsity-inducing methods can result in a parsimonious model, thus improving the out-of-sample forecasting performances. Finally, comparing the nowcasting result of GDP growth using AIO data and macroeconomic data, we can see that the p-values for horizon $h = 0, 1$ and 2 for AIO is larger than macroeconomic data, which proved the extra information contents in news attention data, i.e., including news attention data will facilitate the nowcasting performance in the short run.

4.3. Conditional evaluation

Traditional forecast evaluation test focus on the performance of two models on *average* but are non-informative about the *conditional* performances. In the real world, we may want to know *ex-ante* which model is more suitable under the current circumstances. For example, the long memory pattern of some time series is more pronounced during the tranquil period while compromised during market turmoil, so models with large persistency will produce more reasonable forecasts during tranquility and maybe be undermined during turmoil.

To cope with the conditional evaluation of forecasting results, Li et al. (2022) developed a nonparametric test for conditional predictive ability comparison: the conditional superiority predictive ability (CSPA) test. They define the difference of loss function for two forecasts as:

$$Y_t \equiv L(F_t, F_t^1) - L(F_t, F_t^0) \tag{9}$$

where F_t is the true value, F_t^0 and F_t^1 are the forecasts given by the benchmark and alternative models, respectively. The null hypothesis of the CSPA test can be written as follows:

$$h = \mathbb{E}_0^{\text{CSPA}}\left[Y_t | X_t = x\right] > 0 \text{ for } x \in \mathcal{X} \tag{10}$$

where X_t is the conditioning variable, which can be economic uncertainty or proxy for economic fundamentals, and $\mathcal{X}$ is the collection of all the possible values for X_t. The main obstacle for CSPA is to infer the full functional path for test statistic for each $X_t = x$. Li et al. (2022) bypass this problem by approximating the functional path by basis polynomial of the conditioning variable X_t, then perform bootstrap to obtain the confidence interval.

We adopted the testing procedure of Li et al. (2022) and chose economic policy uncertainty (Huang & Luk, 2020), the leading index downloaded from the CEIC database and inflation rate as conditioning variables. These three conditioning variables are monthly data, so we take the average within each quarter to match the frequency of the targets. Besides, we only use the results produced by the RF model averaged over the nowcast horizons ($h = 0, 1$ and 2) and chose nowcasts made by macroeconomic data as benchmark F^0. The solid black lines and red dash lines in Fig. 7 visualized the conditional expectation of loss function differences and its 95% confidence intervals, respectively, and the horizontal axis is the quantiles of conditioning variables. Intuitively, if any part of the red dash line drops below 0, then we can reject the null hypothesis, which is forecasts made by macroeconomic data dominate the forecasts made by news attention data for all the possible values, at 95% significant level.

There are several noteworthy findings in Fig. 7. Firstly, panels (a) and (b) illustrate the CSPA test result of GDP growth rate conditioning on economic policy uncertainty (EPU) and leading index (LI), respectively. The tests showed that the relative performance of macroeconomic data is inferior compared with news attention data when EPU is high and LI is low. Heightened EPU and sunken LI indicate economic downturn or recession (Baker et al.,

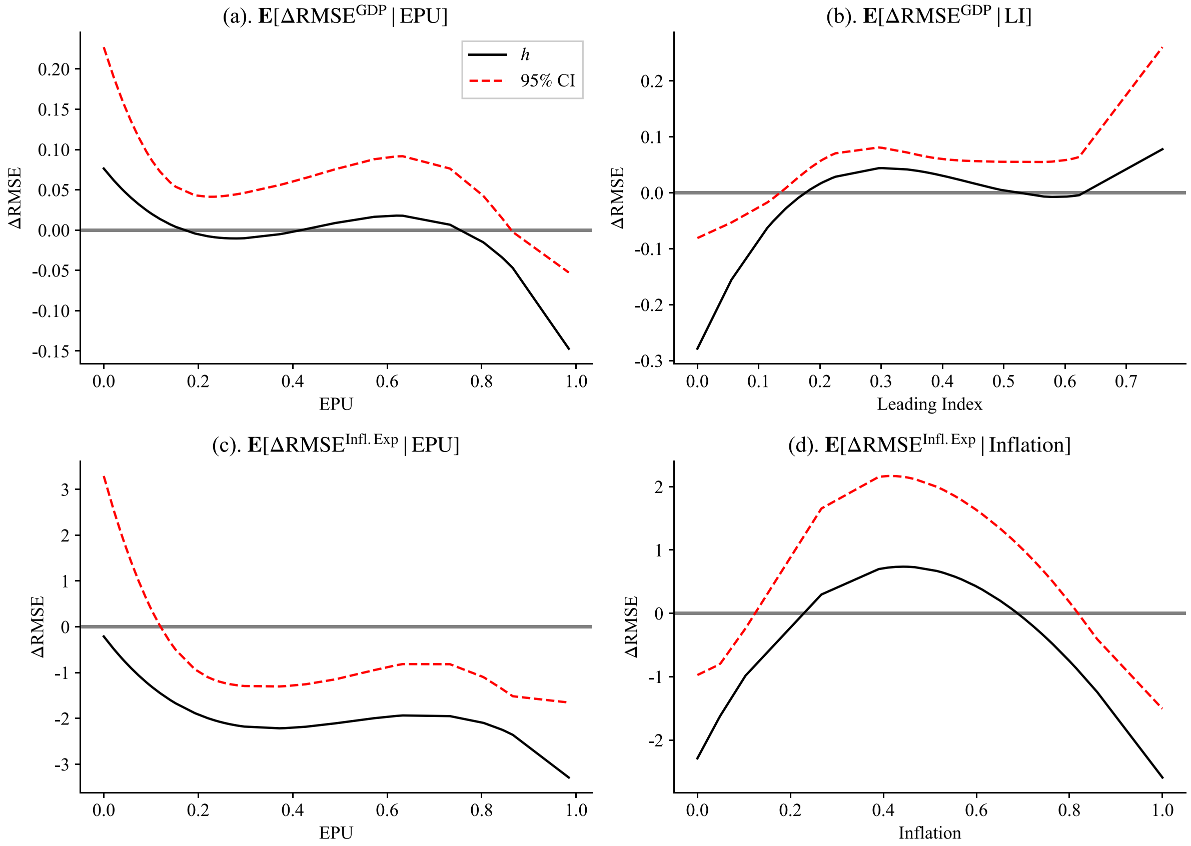


Fig. 7. Conditional superior predictive ability.
Note: This figure illustrate the conditional loss function differences as in Li et al. (2022). The solid black line is the conditional loss function difference $\mathbb{E}[Y_t|X_t = x]$. The red dash line is the upper 95% significant level. The horizontal axis is the quantiles of conditioning variables. Any part of the 95% confidence interval dropping under zero indicates the rejection of the null hypothesis, showing that macroeconomic data dominates over news attention data for every possible value in the conditioning variables.

2016). In other words, news attention data is more appropriate to nowcast GDP growth during the turmoil. This finding is supported by Nimark and Pitschner (2019). Firstly, the information delegation role of news media is state-dependent and more pronounced during the economic downturn. Secondly, the "news selection function" for editors of the newspapers is somewhat asymmetric, i.e., editors tend to report more negative news when the economy is trapped in recession since they are more eye-grabbing, which is also supported by the research of Shiller (2017). Secondly, when conditioning on EPU, nowcasts of inflation expectation showed a similar pattern as in panel (a) besides most of the red dash line located under 0, which indicates news attention data outperform macroeconomic data in most situations. When taking the current inflation rate as the state variable, the 95% confidence interval exhibits a hump shape. In other words, news attention data achieves more accurate forecasts under more extreme inflation rates.

5. Exploring the predictor contributions

The previous results show that both news attention data and macroeconomic data present satisfying performance in macroeconomic nowcasting. Since machine learning methods are often stigmatized as "black box" models, we further explore which predictor or group of predictors contribute to nowcasting. Due to the unsatisfying performance of PCR, we focus on the sg-LASSO and RF model in this section. Although sg-LASSO and RF can deal with high dimensional data, they are intrinsically different. sg-LASSO is a variable selection method, while RF incorporates bootstrapping and ensemble to guard against overfitting. To measure which predictor is informative in nowcasting, we calculate two measures. For sg-LASSO, we compute the ratio of active predictors (whose coefficients are nonzero) in out-of-sample. We use impurity-based feature scores for the RF models to measure predictor importance.

5.1. Does news attention data provide any incremental information?

Nowcasting evaluation in Section 4 revealed that news attention data provide incremental information beyond macroeconomic data in the short run. We further demonstrate this by exploring the percentage of active variables in sg-LASSO.

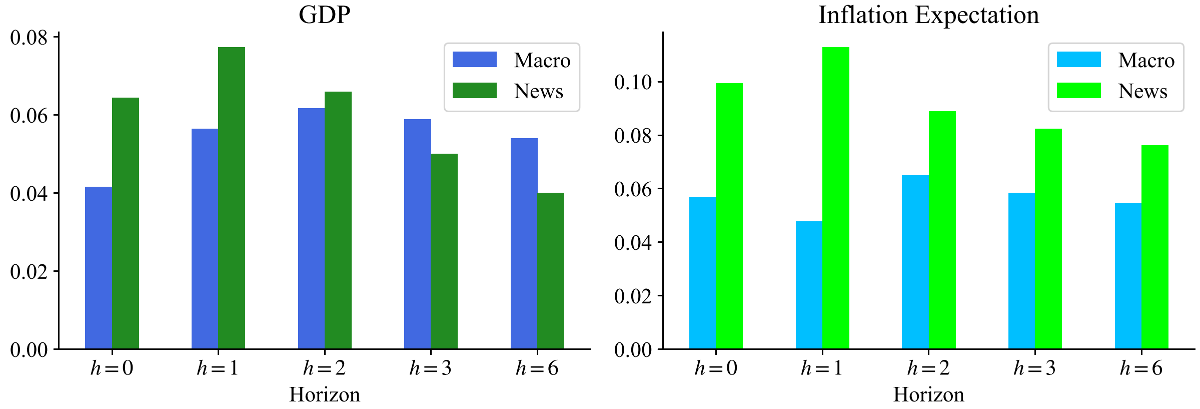


Fig. 8. Ratio of active variables in macroeconomic data and news attention data.
Note: This figure illustrates the ratio of active numbers for macroeconomic data and news attention data when nowcast GDP growth rate (left panel) and inflation expectations (right panel) when using sg-LASSO and AIO predictors.

Specifically, we count the number of variables with nonzero coefficients for news attention and macroeconomic data in each out-of-sample step, divide by the number of total variables in each dataset, and finally, take the average across time.[15] We also calculate the share of active predictors in each out-of-sample step; the results can be found in figure D.4.

Fig. 8 illustrates the results. The left bar denotes the share of active predictors in macroeconomic data at each horizon, and the right bar denotes the share of active predictors in news attention data. For both GDP growth and inflation expectation, the total number of active predictors decreases as the horizon increases, but there are some distinct features for GDP growth and inflation expectation. From the left panel, we can see that the percentage of active predictors is higher for news attention data for the shorter horizon, such as $h = 0, 1$ and 2. As the horizon increase, the share of active predictors in macroeconomic data exceeds news attention data. This finding implied that news attention data are more informative in the shorter horizon, while macroeconomic data seems more suitable for forecasting in longer horizons. As for inflation expectation, news attention data dominate macroeconomic data across every horizon. This is plausible since news attention data alone can achieve better precision than macroeconomic data.

5.2. Predictor contribution and narrative realism

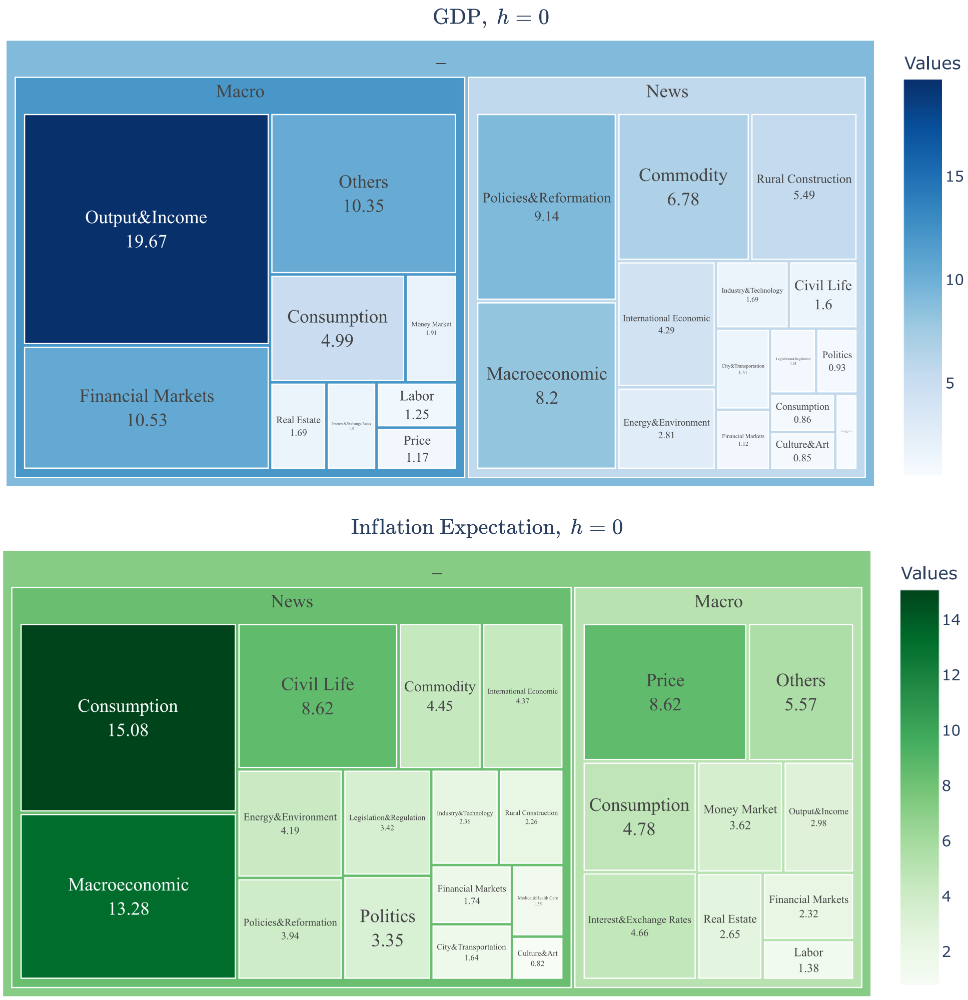
As a variable selection method, sg-LASSO can only tell us which variable is selected. It has nothing to say about the relative importance of predictors. In contrast, the byproduct of the RF model, the impurity-based feature score, can be a proxy of predictor importance measure. In detail, when the bth tree in the random forest is grown, the out-of-bag (OOB) samples are passed down this tree, and prediction accuracy can be obtained. Then the values of jth predictor are randomly permuted in the OOB sample, and the accuracy is again computed. The decrease in accuracy due to the random permutation is averaged over all the trees and normalized by the standard deviation of the whole ensemble trees, that is, the measure of the importance of jth predictor in RF. Intuitively, if this measure is high for a particular predictor, it means that this predictor will significantly reduce the prediction error, and thus, this predictor is vital for forecasting.

Fig. 9 provides a complete picture of the predictor importance in the AIO dataset.[16] The outer rectangles represent different types of data (news attention or macroeconomic), the inner rectangles denote the predictor groups illustrated in Section 2, group names, and the value of predictor importance average over the group are reported in the center of each inner rectangles. The area and darkness of each rectangle represent the magnitudes. Several findings can be concluded from the figure. (1) For GDP growth, news attention and macroeconomic data play a similar role in nowcasting since the area of two outer rectangles is roughly the same. As for inflation expectation, news attention data have more significant predictor importance. (2) The predictor groups which are important for nowcasting are reasonable. For example, when forecasting GDP growth, “Output&Income” and “Financial Market” in the macroeconomic category and “Rural Construction” and “Macroeconomic” in the news attention category are crucial. It is not surprising that these macroeconomic data show predictive power, but some news attention data also play unneglectable roles. The group “Rural Construction” contains reports about infrastructure and poverty elimination projects which are the main focus of the Chinese government. In nowcasting inflation expectation, news reports about “Consumption” and “Macroeconomic” and macroeconomic data in “Price”, “Consumption” and “Interest&Exchange Rates” groups are highlighted, which also provide narrative aspects for the nowcasts results.

Fig. 9 only illustrate *average* predictor importance in out-of-sample periods. It is well known that predictability can be time-varying, and the latent driving force may

[15] More precisely, since we deal with mixed-frequency data with MIDAS model and treat the lag of each variable as a natural group, if there are any nonzero coefficients in a certain group, we label this variable is selected by sg-LASSO.

[16] We only report $h = 0$ in the article; other horizons can be found in Appendix D.


GDP, $h = 0$

Inflation Expectation, $h = 0$

Fig. 9. Treemaps for predictor importance in AIO dataset.
Note: This figure demonstrates the importance of predictors in the AIO dataset. Different types of data are annotated in the outer rectangle. The predictor's importance can be inferred from the area and darkness of the inner rectangles. The group name and numeric predictor importance value are annotated in each inner rectangle's center.

exhibit structural breaks. For this reason, we also calculate the predictor importance at each out-of-sample step. Fig. 10 visualize the results. We only report $h = 0$ to save space; additional results can be found in Appendix D.

We can draw several conclusions from Fig. 10. Firstly, the predictor importance is time-varying. For instance, the predictor importance of news attention data in "Policies & Reformations" is gradually increasing after 2012. This is intuitive since Chairman Xi Jinping was inaugurated in 2012 and implemented policies against corruption and poverty, accelerating market reform. Another noticeable finding is that the predictor importance of news attention data from the "Medical & Health Care" group spiked during the end of 2019, which coincided with the outbreak of the COVID-19 pandemic in China. Still, the predictor importance of this group is neglectable before 2019.

Secondly, predictor importance of some predictors is persistent over time. For example, the predictor importance of "Input&Output" for GDP growth rate remains significant over time, as well as "Price" in the macroeconomic data category and "Consumption" in the news attention category. All these groups are either constitutional or critical for the formation process of the targets. Finally, comparing the panel (a) (panel (b)) and panel (c) (panel (d)) in Fig. 10, we can observe that more cells are darker at each time, i.e., the cross-sectional distribution of predictor importance is more disperse for inflation expectation. This indicates that many predictors can influence agents' expectation formation process.

6. Conclusion

In this paper, we forecast GDP growth and inflation expectation in China using macroeconomic and news attention data from 866234 articles published in three Chinese

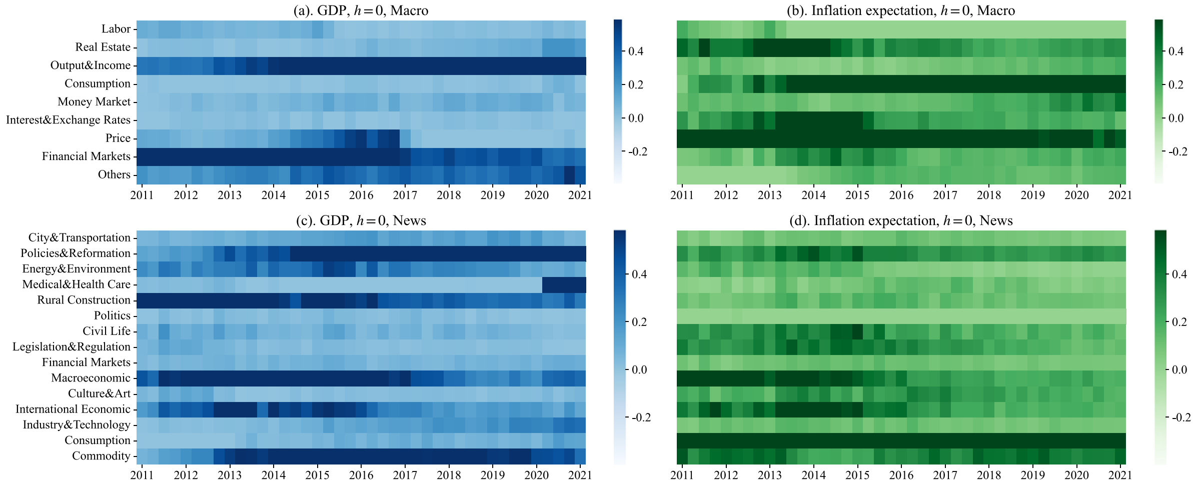


Fig. 10. Predictor importance in RF across time.
Note: This figure visualizes the predictor importance for macroeconomic and news attention data across out-of-sample periods. The darkness of the cells indicates the magnitude of the predictor group at each time.

business newspapers. By applying the MIDAS technique and machine learning methods, we can effectively incorporate high-frequency predictors while handling the high-dimensional problem.

The main findings of this paper can be concluded as follows. (i) News attention time series provide a narrative perspective of the real economy; counterparts of significant events can be found in the news attention time series. (ii) Combining the MIDAS model and machine learning techniques, we find that news attention data deliver comparable or even better precision in nowcasting GDP growth and inflation expectation. Specifically, news attention data is more informative when nowcasting inflation expectation and performs better when combined with the RF method or in turbulent periods. (iii) When inspecting which predictor contributes to nowcast, we find that the importance of some predictors is time-varying while some are persistent in the out-of-sample periods. Predictor importance aligns with economic theory and reveals the narrative of news attention data. The results of this paper confirmed the value of textual data for macroeconomic research and provided evidence for the narrative economics pioneered by Shiller (2017).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the editor, the associate editor, and two anonymous reviewers for useful comments and suggestions which led us to improve the paper. The authors would like to acknowledge that this work is supported by the Humanities and Social Science Fund of Ministry of Education of the People's Republic of China under Grant No. 22JJD790050, the National Natural Science Foundation of China, General Program under Grant No. 71973110, the National Natural Science Foundation of China, Key Program under Grant No. 72033008, the Fundamental Research Funds for the Central Universities under Grant No. 20720191072, the Statistical Science Research Program of China under Grant No. 2022LZ37 and No. 2022LZ06, the Chinese National Funding of Social Sciences under Grant No. 21&ZD146, and the Cultivation Program of Financial Security Collaborative Innovation Center, Southwestern University of Finance and Economics under Grant No. JRXTP202202.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2023.05.006.

References

Andreou, E., Ghysels, E., & Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, *158*(2), 246–261.

Babii, A., Ghysels, E., & Striaukas, J. (2021). Machine learning time series regressions with an application to nowcasting. *Journal of Business & Economic Statistics*, 1–23.

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

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

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Nowcasting and the real-time data flow. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting. Vol. 2* (pp. 195–237). Elsevier.

Barnett, W., Chauvet, M., Leiva-Leon, D., & Su, L. (2016). *Nowcasting nominal GDP with the credit-card augmented divisia monetary aggregates*: *Working Papers of UoK 201605*, University of Kansas, Department of Economics.

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

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, *10*(1), 615–643.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

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

Bybee, L., Kelly, B. T., Manela, A., & Xiu, D. (2021). *Business news and business cycles*: *Working Paper*.

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

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

Ellingsen, J., Larsen, V. H., & Thorsrud, L. A. (2021). News media versus FRED-MD for macroeconomic forecasting. *Journal of Applied Econometrics*.

Foroni, C., Marcellino, M., & Schumacher, C. (2015). Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 57–82.

Galbraith, J. W., & Tkacz, G. (2018). Nowcasting with payments system data. *International Journal of Forecasting*, *34*(2), 366–376.

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

Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90.

Giannone, D., Lenza, M., & Primiceri, G. E. (2021). Economic predictions with big data: The illusion of sparsity. *Econometrica*, *89*(5), 2409–2437.

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

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. New York: Springer.

Huang, Y., & Luk, P. (2020). Measuring economic policy uncertainty in China. *China Economic Review*, *59*, Article 101367.

Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354.

Kuzin, V., Marcellino, M., & Schumacher, C. (2012). Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics*, *28*(3), 392–411.

Li, J., Liao, Z., & Quaedvlieg, R. (2022). Conditional superior predictive ability. *Review of Economic Studies*, *89*(2), 843–875.

Marcellino, M., & Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, *72*(4), 518–550.

Mariano, R. S., & Murasawa, Y. (2003). A new coincident index of business cycles based on monthly and quarterly series. *Journal of Applied Econometrics*, *18*(4), 427–443.

Mariano, R. S., & Murasawa, Y. (2010). A coincident index, common factors, and monthly real GDP. *Oxford Bulletin of Economics and Statistics*, *72*(1), 27–46.

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

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.

Mogliani, M., & Simoni, A. (2021). Bayesian MIDAS penalized regressions: Estimation, selection, and prediction. *Journal of Econometrics*, *222*(1), 833–860.

Moriwaki, D. (2020). Nowcasting unemployment rates with smartphone GPS data. In *Multiple-aspect analysis of semantic trajectories* (pp. 21–33). Cham: Springer International Publishing.

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

Schorfheide, F., & Song, D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, *33*(3), 366–380.

Shiller, R. J. (2017). Narrative economics. *American Economic Review*, *107*(4), 967–1004.

Siliverstovs, B. (2017). Short-term forecasting with mixed-frequency data: a MIDASSO approach. *Applied Economics*, *49*(13), 1326–1343.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group Lasso. *Journal of Computational and Graphical Statistics*, *22*(2), 231–245.

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

Stock, J., & Watson, M. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics* (pp. 415–525). Elsevier.

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, *38*(2), 393–409.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

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