Forecasting GDP growth rates in the United States and Brazil using Google Trends

International Journal of Forecasting 39 (2023) 1909–1924

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting GDP growth rates in the United States and Brazil using Google Trends

Evripidis Bantis *, Michael P. Clements, Andrew Urquhart

ICMA Centre, Henley Business School, University of Reading, United Kingdom

ARTICLE INFO

Keywords:
Big data
Google Search
Factor models
Nowcasting
Variable selection

ABSTRACT

In this paper we consider the value of Google Trends search data for nowcasting (and forecasting) GDP growth for a developed economy (the U.S.) and an emerging-market economy (Brazil). Our focus is on the marginal contribution of big data in the form of Google Trends data over and above that of traditional predictors, and we use a dynamic factor model to handle the large number of potential predictors and the "ragged-edge" problem. We find that factor models based on economic indicators and Google "categories" data provide gains compared to models that exclude this information. The benefits of using Google Trends data appear to be broadly similar for Brazil and the U.S., and depend on the factor model variable-selection strategy. Using more disaggregated Google Trends data than its "categories" is not beneficial.

1. Introduction

Macroeconomic nowcasting[1] has received much attention from policymakers and market practitioners who require an accurate reading of the state of the economy (recent, current, and prospective). Unlike financial variables collected at a higher frequency and published with little delay, key macroeconomic variables such as GDP are only available at lower frequencies, such as quarterly, and are generally only published with a significant delay. For example, in the U.S., the advance estimates of GDP and its components are only available a month after the reference quarter, and in some countries, the delays are longer. This means it may be possible to exploit higher-frequency indicators produced in a timely fashion to generate nowcasts and forecasts of macro-variables before the official estimates are released.

Traditionally, three sources of data have been considered for macroeconomic nowcasting: (i) hard indicators, such as retail sales and industrial production, (ii) surveys of opinions and intentions, and (iii) high-frequency financial market data. However, in recent years, due to computer technology advancements and the advent of online information-gathering services, alternative data sources have become available, usually referred to as big data.[2] A popular source of big data for short-term macroeconomic forecasting is Google Trends, which provides information about the frequency with which a particular

* Correspondence to: ICMA Centre, Henley Business School, University of Reading, Whiteknights, Reading, RG6 6BA, United Kingdom.
E-mail address: e.bantis@pgr.reading.ac.uk (E. Bantis).

[1] Nowcasting is defined as in Giannone, Reichlin, and Small (2008) and involves prediction estimates of the present, the near future, and the recent past. The term is a contraction of "now" and "forecasting" and originates in meteorology.

[2] The term "big data" was first used in the economics and econometrics literature in Diebold (2003) (for more information regarding the origins of this term, see Diebold (2021)). IBM classifies big data into four categories (the "Four 'V's"): volume, variety, velocity, and veracity. Types of big data include social networks, traditional business systems, and the Internet of Things (Kapetanios, Papailias, et al., 2018). Doornik and Hendry (2015) distinguish three shapes in numerical big data: "tall" datasets, where there are not so many variables, N, but many observations, T, with $T \gg N$; "fat" datasets, in which the number of variables exceeds the number of observations, $N \gg T$; and "huge" datasets, where there many variables and many observations, that is, extremely large N and T.

https://doi.org/10.1016/j.ijforecast.2022.10.003

term is searched. Google Search data may contain insights into consumers' and other agents' plans and intentions, and perhaps especially consumer spending. Consumers may seek information on Google's search engine before making economic decisions regarding purchases, for example. Consequently, Google Search data may constitute a valuable source of information for nowcasting macro-variables.

The above developments set the scene for the current paper. Our aim is to determine whether high-dimensional datasets, such as Google Search series, contain additional predictive power over and above that contained in traditional higher-frequency data sources, and our modeling and forecasting strategy is designed to address this question. It is worth noting that we are interested in the marginal additional benefit of Google Search data because the analyst will typically have access to both sources of information. Due to the large number of explanatory variables that are under consideration, we use the dynamic factor model (DFM) framework of Giannone et al. (2008).

Several papers have employed Google Search data to forecast specific macroeconomic variables such as private consumption (Vosen & Schmidt, 2011), unemployment and employment rates (Borup & Schütte, 2022; Choi & Varian, 2009; D'Amuri & Marcucci, 2017), and price levels (Seabold & Coppola, 2015). However, only a few papers have examined the usefulness of Google data in forecasting the overall economic activity, such as Ferrara and Simoni (2019) and Götz and Knetsch (2019). In particular, Götz and Knetsch (2019) found that Google Search data can lead to more accurate GDP growth forecasts for the German economy, while Google data works better as an alternative to survey indicators rather than in addition to them. Ferrara and Simoni (2019) concluded that Google Trends are particularly valuable in nowcasting the eurozone's GDP during the first four weeks of the quarter, since during that time there is a lack of information about variables related to the economy. Nevertheless, when official variables become available, the forecasting power of Google Trends data disappears. Overall, findings suggest that Google Search data may constitute a fruitful set of information for nowcasting or short-term forecasting of macroeconomic variables.

However, a critical question we address is whether these findings for advanced economies, such as Germany and the euro area, are replicated for emerging economies. There are opposing reasons suggesting that Google Search data might be more or less valuable for less developed economies. In such economies, traditional data sources may be of a lower quality, or information may not be available or may be fragmented. This suggests that Google Trends data may fill a void and be more valuable. Against this, the Google Trends data may themselves be less useful if lower rates of internet usage make the information less representative.

Our use of the DFM may improve upon the simple bridge equations employed in Götz and Knetsch (2019) and Ferrara and Simoni (2019). Although the choice between these models remains an empirical question, DFMs may be more suited to nowcasting, since they can read the flow of data in real time and effectively cope with non-synchronous data releases ("ragged edges").

Factor models have become a workhorse model at central banks and other institutions for short-term forecasting, due to this ability to deal with large "ragged-edged" datasets and mixed frequencies of monthly predictors and quarterly GDP rates. A parsimonious structure is achieved by summarizing the information of the many data releases with a few common factors. Regarding the estimation method of the DFM, the two-step estimator of Doz, Giannone, and Reichlin (2011) is employed, where in the first step, model parameters are estimated by principal components using a standardized balanced dataset; and in the second step, the Kalman filter is used to update the estimates using an unbalanced dataset. Many studies have used factor models to forecast macroeconomic variables: *inter alia*, Stock and Watson (2002) and Giannone et al. (2008) for the United States, Schumacher (2010) for Germany, Barhoumi, Darné, and Ferrara (2010) for France, Schiavoni, Palm, Smeekes, and van den Brakel (2021) for the Netherlands, Jansen, Jin, and de Winter (2016) for the euro area, Panagiotelis, Athanasopoulos, Hyndman, Jiang, and Vahid (2019) for Australia, Caruso (2018) for Mexico, Bragoli and Fosten (2018) for India, Dahlhaus, Guénette, and Vasishtha (2017) for BRIC economies, and Luciani, Pundit, Ramayandi, and Veronese (2018) for Indonesia. For further discussions regarding theoretical aspects, estimation techniques, and empirical applications of DFMs, see Bai, Ng, et al. (2008b), Stock and Watson (2011), and Doz and Fuleky (2020).

It is well established that including the largest available dataset in a forecasting context does not necessarily lead to more accurate predictions. As Boivin and Ng (2006) suggest, including more series to estimate common factors could be less beneficial for forecasting in cases where idiosyncratic components are cross-correlated. Furthermore, factors that are not "targeted" on the variable to be forecast may not perform well. Bai and Ng (2008a) provide evidence that it may be sensible to pre-select variables from the dataset prior to calculating the factors. Kim and Swanson (2018) and Cepni, Güney, and Swanson (2019) demonstrate the efficacy of dimension-reduction and shrinkage methods for forecasting for the U.S. and emerging economies.[3]

Informed by these findings, we use a number of variable selection and shrinkage methods to select variables that are relevant for nowcasting GDP growth rates. These are the elastic net, the least absolute shrinkage and selection operator (LASSO), and the adaptive LASSO. To avoid "look-ahead" bias,[4] we do not apply the variable selection methods to the entire sample period. Instead, first, we extract targeted predictors by using only information available during the in-sample period (2005–2014); and second, we compute targeted predictors by employing a most-recent-performance (MRP) method, where every time a balanced dataset is updated with new information,

[3] For further discussions of dimension-reduction methods and forecasting, see Schumacher (2010) and Bulligan, Marcellino, and Venditti (2015), *inter alia*.

[4] We wish to avoid using information that would not be known at the time the forecast is made, as this would tend to exaggerate forecast performance. Throughout, we use a real-time approach to respect the out-of-sample nature of forecasting.

the "best" variable selection method is selected according to local out-of-sample performance, i.e., the variable selection method which produces the lowest average RMSFEs in a series of one-step-ahead forecasts over the last four quarters.

It should be stressed that we focus on forecasting GDP growth rates for two reasons. First, GDP is typically available four weeks after the end of the quarter in developed economies (e.g., in the United Kingdom and the United States), while in the case of emerging economies, it requires more than eight weeks (e.g., in Brazil and India). Taking into account that GDP is usually used as the measure of activity in an economy, and that it is closely followed by policymakers and market practitioners, timely and accurate estimations are of utmost importance.

We choose to study Brazil because it typifies the usual challenges with developing economies, in that the data can be of lower quality than in developed economies, and soft indicators such as opinion surveys are often unavailable or published with significant lags. Hence, Google series can be used as a supplementary source of data to measure consumers' interest in certain keywords, since they are published in real time and are considered of high quality. In addition, Brazil contains some important features and consists of an interesting case to study the predictive ability of Google Trends data within an emerging-market environment. Google is the dominant search engine based on market share, while the internet penetration rate is considered satisfactory, especially compared to other developing economies.[5]

The results of our forecasting exercises provide a number of insights. First, as most of the literature suggests, DFMs successfully incorporate new information as it becomes available, with forecast errors tending to decrease as we move from forecasting to nowcasting and backcasting. Second, for the U.S., estimates of factor models outperform a simple autoregressive benchmark at all horizons, while for Brazil, they primarily outperform the benchmark at nowcasting and backcasting. Third, factor models that utilize both economic indicators and Google Trends categories outperform by far the benchmark in both countries, establishing the value of big data in the form of Google Trends data for now(fore)casting GDP growth. Fourth, benefits from performing variable selection before the computation of common factors tend to arise mainly at one-quarter-ahead forecast horizons ($h = 1$), and their performance decreases as we incorporate more data. In addition, we observe more gains from pre-selecting predictors when we use both economic indicators and Google data, rather than when we employ only economic indicators. The gains from including Google Trends data when the model already includes economic indicators are found to depend on the variable-selection strategy. Google Search data provide gains when we construct factor models based on economic indicators and the main Google Trends categories. But further disaggregation of the Google Trends data to consider the sub-categories is not helpful from a forecasting perspective. Finally, our findings do not strongly point in one direction in response to the question of whether Google data are more or less useful in an emerging economy such as Brazil (relative to the U.S.).

The remainder of this paper is organized as follows: Section 2 discusses the methodological framework, and Section 3 briefly describes the dataset. Section 4 presents the nowcasting design and empirical results. Section 5 summarizes the main findings of this paper. An Online Appendix is available which provides additional details on the empirical exercise.

2. Methodological framework

This section analyzes the methodological framework employed in this paper. Section 2.1 presents the dynamic factor model (DFM) proposed by Giannone et al. (2008), and Section 2.2 presents the main benchmark model. Finally, Section 2.3 briefly discusses the variable selection methods used to construct targeted predictors before computing common factors.

2.1. Dynamic factor model

In this paper, a DFM is employed to forecast real GDP growth rates, as proposed by Giannone et al. (2008). Dynamic factor models summarize the information contained in the set of predictors using a few latent common factors. In particular, if we assume that n corresponds to the cross-sectional dimension of the dataset and T to the number of observations, the aim of a DFM is to separate each observation of a series, say X_t, into two orthogonal unobserved components. The first component, the *common component*, captures the cross-sectional comovements across series and is assumed to be a linear function of a few, r, latent common factors, with $r \ll n$. The second component, the *idiosyncratic component*, captures variable-specific features and is assumed to be serially and cross-sectionally uncorrelated with the common factors. To summarize, the DFM can be written as

$$X_t = \Lambda f_t + \varepsilon_t \tag{1}$$

where $X_t = (X_{1t}, \dots, X_{nt})'$ is an $(n \times 1)$ stationary process of n variables with $t = 1, 2, \dots, T$ observations, Λ is an $(n \times r)$ matrix of factor loadings, $f_t = (f_{1t}, \dots, f_{rt})'$ is an $(r \times 1)$ stationary process of common factors, and $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{nt})'$ is an $(n \times 1)$ stationary process of idiosyncratic errors. The product Λf_t in Eq. (1) denotes the common component of X_t. Common factors (f_t) and idiosyncratic component (ε_t) are considered to be orthogonal, that is, $E(f_t \varepsilon_s') = 0$ for any t and s. Factors can be modeled as a vector autoregressive (VAR) process of order p:

$$f_t = \Phi_1 f_{t-1} + \cdots + \Phi_p f_{t-p} + B u_t, \\ u_t \sim WN\left(0, I_q\right) \tag{2}$$

[5] The most obvious choices for a leading emerging economy were China, India, Russia, and South Africa. China and Russia were discarded because Google is not the dominant search engine in these countries, and India and South Africa were rejected because internet penetration was deemed relatively low (below 50% in both countries; Statista, 2018). By contrast, Google is the dominant search engine in Brazil (97% market share; Statista, 2018) and the internet penetration rate is around 70% (Statista, 2018).

where $\Phi_1, \ldots, \Phi_p$ is an $(r \times r)$ matrix of autoregressive parameters, B is an $(r \times q)$ matrix of full rank q, and u_t is a q-dimensional white noise process of the shocks to factors. The idiosyncratic component is assumed to be orthogonal to common shocks. For more details, see Doz et al. (2011).

The system of Eqs. (1) and (2) can be cast in a state-space representation, in which Eq. (1) represents the *measurement equation* and describes the relationship between the observed predictor X_t and the unobserved common factor f_t, while Eq. (2) denotes the *state equation* and explains how the unobserved factors are generated from their lags and innovations.

The DFM, as specified in Giannone et al. (2008), follows a two-step approach. Firstly, preliminary estimations of factor loadings ($\hat{\Lambda}$) and common factors ($\hat{f}_t$) are derived by principal components, by using a standardized, balanced dataset. Then, the autoregressive coefficients ($\hat{\Phi}_j, j = 1, \ldots, p$) are derived by using the estimated factors, $\hat{f}_t$ in a VAR(p) model. Secondly, the Kalman filter and Kalman smoother are employed to re-estimate the unobserved factors using the unbalanced dataset. In order to deal with the "ragged-edge" problem (i.e., missing observations at the end of the sample) the variance of the idiosyncratic component is set to infinity when X_t is not available. The factors are then projected to the future ($\hat{f}_{T+h|T}$) by recursively estimating a VAR model of order p, as shown by Eq. (2). Forecasts of the monthly factors are aggregated into quarterly frequency by employing the transformation from Mariano and Murasawa (2003) and therefore a forecast of GDP growth rate is a linear function of the projected common factors and can be estimated via OLS:

$$\hat{y}^Q_{T+h|T} = \hat{\beta}_0 + \hat{\beta}'\hat{f}^Q_{T+h|T} \tag{3}$$

where $\hat{f}^Q_{T+h|T}$ denotes the $r \times 1$ vector of quarterly factors.

2.2. Benchmark model

As an additional way of evaluating the performance of the factor models, we use a simple autoregression as the benchmark:

$$y^Q_t = \mu + \sum_{i=1}^{p} \phi_i y^Q_{t-i} + u^Q_t \tag{4}$$

where y^Q_t denotes the quarterly growth rate of GDP, μ is a constant term, ϕ_i denotes the autoregressive parameters, u^Q_t is an error term, and the model order p is chosen so the error term is approximately white noise. As noted in Section 1, the key comparisons are between the different factor models, for example, between those which exploit Google Trends data and those which do not. However, an AR model serves as a check on whether any information can improve on simply using lags of the variable itself. In nowcasting, we would expect that information pertaining to the quarter we are forecasting would prove beneficial.

2.3. Variable selection methods

We use the DFM to capture the information in large sets of variables for forecasting/nowcasting. As shown by Bai and Ng (2008a) and others, the forecasting performance of factor models can often be improved by a judicious choice of variables from which to extract factors, including choosing targeted predictors, i.e., choosing variables that are correlated with the target variable. Götz and Knetsch (2019) show the value of this approach for forecasting GDP with Google Trends data. Other studies which support the pre-selection of variables prior to factor estimation include Schumacher (2010), who found that least angle regression with the elastic net is useful. Kim and Swanson (2018) also found in favor of combination models, i.e., factor models combined with variable selection models.

We consider a number of dimension-reduction methods prior to the estimation of the factors. These are the elastic net (ENET), the least absolute shrinkage and selection operator (LASSO), and an adaptive version of the LASSO (AdaLASSO). These techniques, which have been extensively used in the aforementioned papers, impose a penalty term and result in the coefficients on some putative explanatory variables being set to zero. Factors can then be estimated for the set of variables that attracted non-zero coefficients.

LASSO was introduced by Tibshirani (1996). We determine the LASSO regularization parameter (which controls the amount of shrinkage) using the BIC, as supported by the simulation evidence in Smeekes and Wijler (2018). A modified version of the LASSO estimator was later introduced, called the adaptive LASSO, which may be preferable in some circumstances (see Zou (2006)). ENET is another modification that might work well when the number of candidate variables exceeds the number of observations T (the LASSO can only choose up to T variables), or when there are predictors with high pairwise coefficients: see Zou and Hastie (2005). More information concerning technical details of these methods is provided in the Online Appendix.

3. Data description

This paper aims to produce forecasts, nowcasts, and backcasts of GDP growth rates for Brazil and the United States using large datasets of traditional economic indicators and Google Trends data. Figure B.1 shows the evolution of GDP growth rates. This section describes both types of data and the main categories, and analyzes the advantages and issues that arise using Google Search data.

3.1. Economic indicators

The main dataset consists of 96 and 115 economic indicators for Brazil and the United States, respectively, spread over ten groups: *Economic Activity, External Sector, Government Sector, Housing Market, Labor Market, Leading Indicators, Monetary Sector, Prices, Retail Sector, and Survey Indicators*. All economic indicators have been downloaded from Bloomberg's *Key Economic Indicators* category and cover the period from January 2005 to September 2019. Table 1 provides a brief summary about the number of selected economic indicators by category, while a complete list of variables with descriptions, publication

Table 1
Summary of economic indicators by category.

	Brazil	United States
Economic Activity	11	13
External Sector	10	9
Government Sector	10	9
Housing Market	2	10
Labor Market	2	19
Leading Indicators	2	19
Monetary Sector	19	10
Prices	25	6
Retail Sector	13	6
Survey Indicators	2	14
Total	96	115

delays, and transformations applied to make them stationary is available in the Online Appendix, Table B.4 and Table B.5.

Variables from consumer and producer prices categories were merged under the *Prices* category while the personal and monetary sectors were merged under the *Monetary* category. Although financial variables could provide information in a timely manner, they were discarded because their volatile nature may incorporate significant noise in our model. As Bańbura, Giannone, Modugno, and Reichlin (2013) show, moreover, financial indicators tend to have a limited role in nowcasting the overall economic activity when a large set of economic variables is included. It should be noted that all variables were standardized by subtracting the mean and dividing by the standard deviation, as a means to avoid overweighting of predictors with high variances when deriving the factors.

3.2. Google trends

The aim of Google Trends is to provide data about the frequency that a particular keyword is searched over the total search volume, on a specific period, in a given geographical region. Google provides an index instead of the actual search volume numbers because of privacy reasons. The index awards a score that ranges between 0 and 100. A value of 0 implies that there were not enough data for this query, while a value of 100 indicates the peak popularity of the search term.

Google trends data are available without any publication delay and show the popularity of internet users' searches in real time. The Google Trends data that we employ in this study span from January 2005 to September 2019 on a monthly frequency. Google classifies search queries into 25 main categories (Table 2), and within each category, there is a further division of 272 subcategories in total. For instance, the query "Debt Management" would be allocated to the category of "Credit and Lending", which is a subcategory of "Finance". As the Google data are not seasonally adjusted, we take the annual growth rates. Also, two abrupt breaks in the Google Trends series occur due to improvements in geographical assignments and data collection systems in January 2011 and January 2016. We adjust the data for these breaks by multiplying the post-break observations by the ratio of the local averages of the observations before and after the break.

From the literature emerges that there are plenty of advantages to using data like Google Trends for forecasting purposes: among others, more timely forecasts; easy data access and collection, as well as the ease of data management and treatment; the high possibility that similar data will be available in the future; and the good quality in data (Buono, Mazzi, Kapetanios, Marcellino, & Papailias, 2017). Furthermore, in developing economies, traditional economic indicators, such as opinion surveys, are sometimes unavailable or published with significant lags, and are often considered to be of lower quality compared to developed economies. Thus, Google data can be used as an alternative source of data to predict consumers' behavior, since they are published in real time and the data quality is high. Thus, in cases where there are high internet penetration rates and Google's search engine market share is dominant, Google data can be used as a proxy for consumer sentiment.

On the other hand, several issues have been identified when incorporating Google Trends data in a nowcasting framework. Firstly, according to Seabold and Coppola (2015), data from Google Trends are not the same over time, but instead, historical data from day to day can be different. This means that the sampling methodology that Google uses incorporates measurement error into the series. For a specific keyword on a particular day, Google provides precisely the same series, but for that specific keyword on different days, the results are not the same (Carrière-Swallow & Labbé, 2013).[6] Secondly, Google Trends series may exhibit strong seasonal components. Thirdly, there is a lack of knowledge of how Google treats and processes data. For example, queries are grouped in Google Trends categories by using a natural-language algorithm whose details are unknown (Kapetanios et al., 2018). Additionally, there are some drawbacks when Google Search data are used in emerging countries. According to Carrière-Swallow and Labbé (2013), there is no clear evidence on whether the internet has integrated into buyers' decision processes in developing economies. Moreover, a substantial proportion of household consumption consists of non-discretionary expenditure, where there is no need for a thorough search on Google before buying these products.

The dataset of this study covers Google Search series for both the United States and Brazil. Considering that each query's meaning can change over time, it seems logical to incorporate the main Google Search categories into our analysis. Additionally, we aim to shed light on the potential usefulness of disaggregate Google series, and therefore we also utilize the Google Search subcategories,

[6] To deal with this issue, D'Amuri and Marcucci (2017) take the simple average of Google Trends data for a specific keyword extracted from two different IP addresses and on 12 different days. Medeiros and Pires (2021) investigate the constantly changing Google Trends samples and highlight the importance of taking averages of several different samples to improve the consistency of the data series. In our setting, we rely on single downloads, since we deal only with main categories and subcategories in which cross-correlations between different samples are always above 0.99.

Table 2
Google Trends: Main categories.

Arts & Entertainment	Autos & Vehicles	Beauty & Fitness
Books & Literature	Business & Industrial	Computer & Electronics
Finance	Food & Drink	Games
Health	Hobbies & Leisure	Home & Garden
Internet & Telecom	Jobs & Education	Law & Government
News	Online Communities	People & Society
Pets & Animals	Real Estate	Reference
Science	Shopping	Sports
Travel		

leading to a total of 297 Google series for each country. This paper's primary purpose is to explore whether data from Google Trends can provide prediction gains in forecasting GDP growth rates.

4. Nowcasting exercise

This section presents the nowcasting exercise and empirical results. Section 4.1 describes the nowcasting design, the different types of datasets, and the mechanics of variable selection methods. Section 4.2 exhibits the main empirical results, and Section 4.3 attempts to provide an answer regarding the potential significance of variable and model selection methods. Section 4.4 focuses on the forecast benefits that arise by including Google data in the factor model and finally, Section 4.5 presents the indicators chosen by the variable selection model.

4.1. Setup

The forecasting performance of the factor model is assessed by a pseudo-real-time out-of-sample exercise. We take into consideration the publication delay of every variable in an attempt to avoid using data that would not have been available at the time of the forecast. However, due to the irregular publication pattern of the variables, we follow the approach of Giannone et al. (2008) and assume that publication delays are constant during the evaluation period. This assumption is not too unrealistic, since variation in publication delays of the variables are only minor. Also, our approach does not take into account data revisions, since real-time vintages for all the predictors in our dataset are not available. However, differences in data revisions are considered idiosyncratic, and therefore they do not affect the computation of common factors (Giannone et al., 2008).

For the United States, a sequence of seven predictions is produced for each quarter in the out-of-sample period, obtained in consecutive months. In particular, we generate three monthly one-quarter-ahead forecasts ($h = 1$), three monthly current-quarter nowcasts ($h = 0$), and one monthly backcast ($h = -1$; i.e., forecasts for the preceding quarter). For Brazil, we generate a sequence of eight predictions, i.e., three monthly forecasts, three monthly nowcasts, and two monthly backcasts, since the publication delay of GDP for Brazil is around eight weeks, while that for the United States is approximately four weeks.

Table 3 presents an example of timing predictions for Brazil for the first quarter of 2019. The first, the second,

Table 3
Timing of forecast exercise for the first quarter of 2019.

	Forecast type	Month	Forecast made on last day of
1	One quarter ahead	1	October
2		2	November
3		3	December
4	Nowcast	1	January
5		2	February
6		3	March
7	Backcast	1	April
8		2	May

and the third forecasts are made at the end of October, November, and December of 2018. Thereafter, three nowcasts are estimated throughout the reference quarter at the end of January, February, and March 2019. Finally, we backcast the 2019:Q1 at the end of April and May, and in early June, the first official estimate for the GDP is being released.

GDP growth rate predictions are estimated recursively, where the first sample begins in January 2005 and ends in January 2014, and the last sample begins in January 2005 and ends in August 2019. Hence, the out-of-sample evaluation period is 2014:Q2 to 2019:Q2, i.e., 21 quarters. For the factor model specification, we optimize the number of factors and shocks at every step of our forecasting process by employing information criteria from Bai and Ng (2002) and Bai and Ng (2007).

Let us assume that we have three types of data at our disposal: economic indicators (*EI*), Google Trends categories (*GTC*), and Google Trends subcategories (*GTS*), denoted by X_{EI}, X_{GTC}, and X_{GTS}, respectively. Thus, as we have shown in Section 2.1, our aim is to forecast GDP growth rates, $y^Q_{T+h|T}$, based on direct factor forecasts, $f_{T+h|T}$. If we denote the available information set by Ω_T,

$$Proj[f_{T+h}|\Omega_T] \tag{5}$$

and the information set is defined as

$$\Omega_T = \{X_{t,j},\ t = 1, \dots, T, \text{and } j \in \{EI, GTC, GTS\}\} \tag{6}$$

The goal of this paper is to evaluate whether Google Trends data can convey additional predictive power over and above that contained in traditional economic indicators. Thus, we first compute the common factors based only on economic indicators, and then we add Google Trends categories and subcategories to assess their marginal contribution. In particular, we first forecast common factors (f_{T+h}) based on economic indicators:

$$Proj[f_{T+h}|\Omega_T] \tag{7a}$$

$$\Omega_T = \{X_{t,j},\ t = 1, \ldots, T,\ j \in \{EIs\}\} \tag{7b}$$

Then we add Google Trends categories:

$$Proj[f_{T+h}|\Omega_T] \tag{8a}$$

$$\Omega_T = \{X_{t,j},\ t = 1, \ldots, T,\ j \in \{EI, GTC\}\} \tag{8b}$$

And finally, we add Google Trends subcategories:

$$Proj[f_{T+h}|\Omega_T] \tag{9a}$$

$$\Omega_T = \{X_{t,j},\ t = 1, \ldots, T,\ j \in \{EI, GTC, GTS\}\} \tag{9b}$$

Common factors are estimated based on datasets without variable pre-selection as well as based on targeted predictors. This paper uses three variable selection methods to construct a set of targeted predictors: the elastic net, the LASSO, and the adaptive LASSO. To facilitate this, we follow two methods. First, targeted predictors are extracted using only information available in the in-sample period, i.e., 2005–2014. Second, we attempt to dynamically adjust the set of targeted predictors throughout the out-of-sample period by re-estimating the variable selection models each time we have new information and a balanced dataset. Additionally, our aim is not only to dynamically update the set of targeted predictors but also to select the "best" variable selection method according to their local out-of-sample performance; that is, we select the model which produces the lowest average RMSFE in a series of four one-step-ahead forecasts over the last four quarters. We call this approach the most-recent-performance (MRP) method.

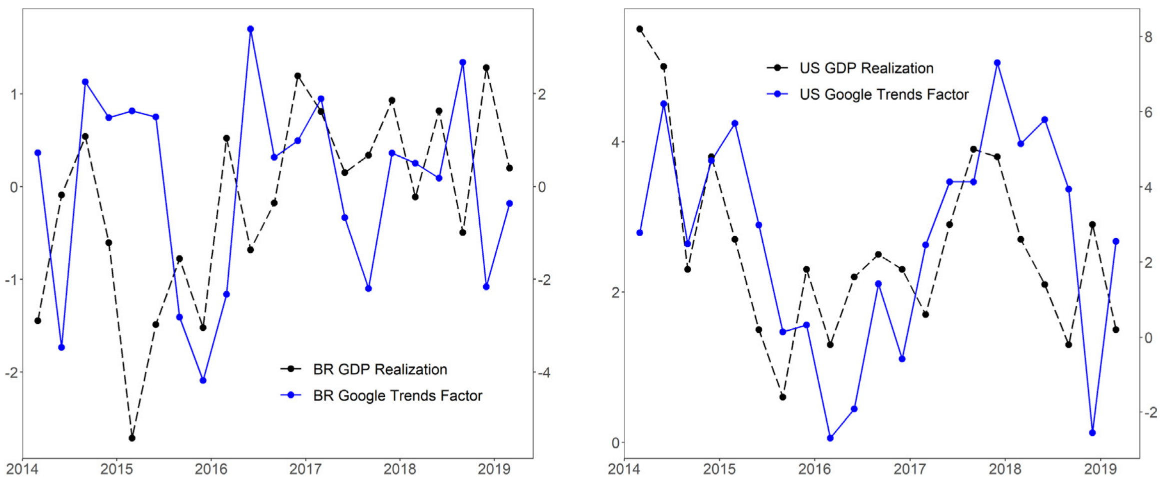
At each forecast origin, we forecast common factors based on observations of the predictors available at that period of time. For example, when our aim is to compute one-quarter-ahead forecasts of 2019:Q1, we compute a series of three monthly forecasts in the last day of October, November, and December based only on information that is available on those dates, that is, 2005:M1–2018:M10, 2005:M1–2018:M11, and 2005:M1–2018:M12, respectively. Specifically, we provide below a step-by-step analysis of our modeling framework in the case of the U.S. when the forecast origin is on the last day of October 2018 and when the MRP method is employed before estimating the DFM:

1. Undertake variable selection using X predictors for the period 2005:M1–2018:M7 (July 2018 is the latest available balanced dataset) and obtain $X_s, s = LASSO, AdaLASSO, ENET$.
2. Conduct a series of four one-step-ahead out-of-sample forecasts for $y^Q_{T+1|T}$, where $T = 2017$:Q2, 2017:Q3, 2017:Q4, and 2018:Q1 using X_s with a DFM (the steps of the DFM are explained below). The set of predictors that produce the best RMSFE is selected and denoted by X^*.
3. Given the selected variables X^*_t, where $t =$ 2005:M1–2018:M10, separate each observation into two orthogonal unobserved components:
$$X^*_t = \Lambda f_t + \varepsilon_t \tag{10}$$
4. Use PCA to obtain preliminary estimations of factor loadings $(\hat{\Lambda})$ and common factors $(\hat{f}_t)$ using a standardized and balanced dataset $(\overline{X}^*_t)$, where $t =$ 2005:M1–2018:M7:
$$(\hat{f}_t, \hat{\Lambda}) = \underset{f_t, \Lambda}{\arg\min}(NT)^{-1} \sum_{t=1}^{T} (\overline{X}^*_t - \Lambda f_t)'(\overline{X}^*_t - \Lambda f_t) \tag{11}$$
5. Use Kalman filtering and smoothing techniques to re-estimate the factors for the unbalanced dataset (2005:M1–2018:M10). To deal with the ragged edges, the variance of the idiosyncratic component is specified as follows:
$$E[\varepsilon^2_t] = \begin{cases} \psi, & \text{if } X^*_t \text{ is available} \\ \infty, & \text{if } X^*_t \text{ is not available} \end{cases} \tag{12}$$
and therefore no weight will be placed on missing observations in the estimation of common factors.
6. Estimate the VAR on the period $T =$ 2005:M1–2018:M10, and forecast with the VAR one step ahead, i.e. for 2018:M11:
$$\hat{f}_{T+1|T} = \hat{\Phi}_1 \hat{f}_T + \cdots + \hat{\Phi}_p \hat{f}_{T-p+1} \tag{13}$$
where T refers to 2018:M10. The two-step-ahead forecast of 2018:M12 is given by:
$$\hat{f}_{T+2|T} = \hat{\Phi}_1 \hat{f}_{T+1|T} + \hat{\Phi}_2 \hat{f}_T + \cdots + \hat{\Phi}_p \hat{f}_{T-p+2} \tag{14}$$
for $p > 1$ and so on. That is, the forecasts are generated iteratively.
7. Forecasts of the factors at the quarterly frequency are calculated from the forecasts of the months using the Mariano and Murasawa transformation.[7]
8. After the quarterly frequency of common factors is computed, bridge equations can be estimated (see for example, Giannone et al. (2008)), since the target variable and the predictors have the same frequency, for the period 2005:Q1 to 2018:Q2. These are used to forecast the target variable for periods 2018:Q3 to 2019:Q1:
$$y^Q_{T+h|T} = \hat{\beta}_0 + \hat{\beta}' \hat{f}^Q_{T+h|T} \tag{15}$$
9. Estimate the AR benchmark using data from 2005:Q1 to 2018:Q2, and generate a three-step-ahead forecast of the target variable for 2019:Q1:
$$Y_{T+3|T} = \hat{\alpha}_0 + \hat{\alpha} Y_T \tag{16}$$

Afterwards, we again update the model on the last day of November 2018 and perform exactly the same steps by incorporating the latest available data:

1. Undertake variable selection using X predictors for the period 2005:M1–2018:M8 (August 2018 is the latest available balanced dataset observed at the end of November 2018) and obtain X_s.

7 Other methods exist in the literature when dealing with mixed frequencies for forecasting. For further details, see Foroni and Marcellino (2014).



(a) Brazil (b) United States

Fig. 1. Percent change of GDP and Google Trends common factor.
Notes: The left-hand (right-hand) plot shows the evolution of GDP growth rates (quarter-on-quarter) and the estimated first common factor based on Google Trends categories for Brazil (United States).

2. Conduct a series of four one-step-ahead out-of-sample forecasts for $y^Q_{T+1|T}$, where T = 2017:Q3, 2017:Q4, 2018:Q1, 2018:Q2 using a DFM. The set of predictors that produce the best-RMSFE is selected and denoted by X^*.
3. Given the selected variables X^*_t, where t = 2005:M1–2018:M11, separate each observation into two orthogonal unobserved components, i.e. the common and idiosyncratic components.
4. Use PCA to obtain preliminary estimations of factor loadings and common factors using a standardized and balanced dataset ($\overline{X}^*_t$), where t = 2005:M1–2018:M8.
5. Use Kalman filtering and smoothing techniques to re-estimate the factors using the unbalanced dataset (2005:M1–2018:M11) and deal with the ragged edges.
6. Estimate the VAR on the period 2005:M1–2018:M11, and forecast with the VAR for periods 2018:M12–2019:M3.
7. Transform the monthly forecasts of the factors to quarterly using the Mariano and Murasawa transformation.
8. Estimate bridge equations for the period 2005:Q1 to 2018:Q3 and use these to forecast the target variable for the period 2018:Q4 to 2019:Q1.
9. Estimate the AR benchmark using data from 2005:Q1 to 2018:Q3, and generate a two-step-ahead forecast for the target variable for 2019:Q1.

Finally, the information criteria from Bai and Ng (2002) are employed to determine the number of common factors:

$$IC(r) = lnV_r(\hat{f}_t, \hat{\Lambda}) + rg(N, T) \tag{17}$$

where $V_r(\hat{f}_t, \hat{\Lambda})$ is the residual sum of squares objective function in (11), in which r common factors are computed as the principal components and where $g(N, T)$ is the penalty term. Although typically the Bai and Ng (2002) IC_{p2} : $g(N, T) = [(N + T)/NT]ln[min(N, T)]$ penalty function is utilized, we also employ the IC_{p1} : $g(N, T) = [(N+T)/NT]ln[NT/(N+T)]$ as a robustness check.

4.2. Forecasting performance of factor models

To assess the forecasting performance of the factor model, a pseudo-real-time out-of-sample forecasting exercise is employed for the period of 2014:Q2 to 2019:Q2, and a sequence of eight (seven) predictions is computed for each quarter for the case of Brazil (the United States). We measure the forecast accuracy with the root mean squared forecast error (RMSFE). Also, the statistical significance of forecast accuracy improvements is assessed by implementing the Diebold and Mariano (1995) test where under the null hypothesis, the two models have the same forecast accuracy. We conduct the Diebold–Mariano test against a one-sided alternative that factor models generate more accurate predictions than the benchmark. However, it is worth noting that these results should be treated with caution because the number of forecast errors for a given horizon is relatively small.

Fig. 1 plots GDP growth against the first common factor based only on Google Trends categories for Brazil (left) and the United States (right) during the out-of-sample period. It is evident that for the United States, the Google Trends factor tracks GDP growth rates quite well, while for Brazil it seems the factor is able to capture the main trends of GDP but with a slight delay. Therefore, it is reasonable to expect that Google Trends series are a more promising source of data for the United States than for Brazil.

Table 4 summarizes the empirical results of the forecasting performance of DFMs for Brazil and the United States. To quantify the forecast accuracy, we compare the

Table 4
Forecast evaluation of factor models (IC_{p2}).

	Forecast ($h = 1$)			Nowcast ($h = 0$)			Backcast ($h = -1$)	
	1	2	3	1	2	3	1	2
Brazil								
Economic indicators								
All	1.44	1.31	1.24	1.06	**0.72****	**0.55****	0.65*	0.59*
LASSO	**1.06**	1.04	**0.90***	0.94	0.91	0.68**	0.71*	0.72*
AdaLASSO	1.19	1.13	0.99	1.02	1.00	0.77**	0.78*	0.81*
ENET	1.22	1.11	0.98	0.98	0.85*	0.65**	0.64*	0.62*
MRP	1.07	**0.95***	0.97	**0.87***	0.73**	0.55**	**0.53***	**0.52***
Economic indicators and Google Trends categories								
All	1.52	1.40	1.37	1.23	0.80**	**0.46****	**0.49***	0.52*
LASSO	**1.07**	**0.88**	**0.86***	**0.90***	**0.74***	0.63**	0.65*	0.61*
AdaLASSO	1.29	1.28	1.03	1.11	1.04	0.68**	0.72*	0.72*
ENET	1.22	1.09	1.01	1.02	0.89*	0.70**	0.68*	0.65*
MRP	1.21	1.02	1.00	0.99	0.79**	0.54**	0.49*	**0.50***
Economic indicators, Google Trends categories, and Google Trends subcategories								
All	1.82	1.85	1.63	1.53	1.33	0.89	0.88	0.81
LASSO	1.26	1.22	1.14	1.17	1.09	0.83	0.88	0.87
AdaLASSO	1.45	1.45	1.28	1.47	1.37	1.04	1.17	1.09
ENET	1.28	1.17	**1.09**	0.99	**0.76****	**0.56****	**0.52***	**0.48***
MRP	**1.18**	**1.15**	1.15	**0.98**	1.03	0.56**	0.57*	0.55*
United States								
Economic indicators								
All	0.86	0.93	0.91*	0.83	0.84	0.76	**0.76**	
LASSO	0.80	**0.88**	**0.84****	**0.81**	0.80*	**0.75***	0.77	
AdaLASSO	0.84	0.90	0.94	0.91	0.86	0.83	0.81	
ENET	**0.79**	0.91	0.86*	0.82	0.81*	0.76*	0.76	
MRP	0.84	0.90*	0.89*	0.83	**0.78***	0.79*	0.79	
Economic indicators and Google Trends categories								
All	0.86	0.95	0.90*	0.83	0.77*	**0.68***	**0.68***	
LASSO	0.80	**0.84***	**0.81***	**0.76***	0.76*	0.73*	0.74	
AdaLASSO	0.81	0.87	0.89*	0.87	0.81*	0.77*	0.77	
ENET	**0.78**	0.85*	0.83*	0.79	0.80*	0.76*	0.74	
MRP	0.79	0.91	0.82*	0.77*	**0.75***	0.75*	0.72	
Economic indicators, Google Trends categories, and Google Trends subcategories								
All	0.84	0.92	0.85*	0.81	0.81	0.79*	**0.77**	
LASSO	**0.83**	0.89	0.86*	0.82	0.81	**0.78***	0.77	
AdaLASSO	0.85	0.89	**0.83****	0.85	0.82*	0.82*	0.80	
ENET	0.83	**0.89**	0.83**	**0.80***	**0.80***	0.79*	0.79	
MRP	0.83	0.91	0.87*	0.81	0.83	0.81	0.78	

Notes: Entries in this table show root mean squared forecast errors (RMSFEs) of dynamic factor models (DFMs) relative to the AR(1) benchmark. Therefore, entries lower than one suggest that a particular prediction based on a DFM is more accurate compared to the AR(1) benchmark. Numbers in bold indicate predictions with the lowest relative RMSFE for each forecast horizon, within a given DFM specification. Grey boxes denote the lowest error measure achieved by the competing models for a given country and forecast horizon. Numbers followed by asterisks (**5% level, *10% level) are significantly superior to the AR(1) benchmark based on the Diebold–Mariano test.

RMSFE of factor models relative to a simple AR(1) benchmark. Subsequently, we report comparisons between factor models with and without Google Trends categories. Results are reported for three monthly forecasts ($h = 1$), three monthly nowcasts ($h = 0$), and two (one) monthly backcasts ($h = -1$) for Brazil (the United States). For each country, figures in bold specify the RMSFE-"best" models across all dimension-reduction models, for a given set of predictors and forecast horizon. Grey boxes indicate the RMSFE-"best" models across all groups of predictors and dimension-reduction methods, for a given forecast horizon and country.

The results reveal several interesting insights. First, relative RMSFEs are consistently lower than unity for the United States, which implies that forecasts from DFMs are more accurate than those produced by the AR(1) benchmark during all prediction horizons; that is, they outperform, on average, by 12.8% in forecasting, by 18.8% in nowcasting, and by 21.8% in backcasting. For Brazil, factor models outperform the benchmark mostly in nowcasting and backcasting horizons, while for higher horizons, forecast gains tend to disappear. Specifically, factor models underperform the benchmark in forecasting on average by 10.7%, but they outperform in nowcasting and backcasting by 18.2% and 34.2%, respectively.

Second, the improvement in terms of forecast accuracy is noticeable for both Brazil and the United States, and the relative RMSFEs decrease as we move through

Table 5
Summary of the RMSFE-"best" variable selection methods.

	All	LASSO	AdaLASSO	ENET	MRP
Brazil					
Model 1	2	2	0	0	4
Model 2	2	5	0	0	1
Model 3	0	0	0	5	3
Total-BR	4	7	0	5	8
United States					
Model 1	1	4	0	1	1
Model 2	2	3	0	1	1
Model 3	1	2	1	3	0
Total – U.S.	4	9	1	5	2
Total	**8**	**16**	**1**	**10**	**10**

Notes: *Model 1* refers to factor models based on economic indicators, *Model 2* denotes models based on economic indicators and Google Trends categories, and *Model 3* represents specifications based on economic indicators, Google Trends categories, and Google Trends subcategories.

the prediction period from the forecast to the nowcast and backcast. This highlights the importance of updating the model with new information as we approach the date when GDP is published. It seems that in the case of Brazil, benefits from incorporating more information are more significant compared to the United States, since relative RMSFEs decrease at a higher rate, but this could be partially explained by the fact that for the United States, factor models during forecasting horizons perform much better compared to Brazil. Also, the volatile nature of Brazilian GDP could weigh negatively on long-horizon forecasts.

Third, factor models that employ economic indicators and Google Trends categories tend to have the greatest performance against the benchmark for both countries, as shown by the grey boxes in Table 4. In particular, in the case of Brazil, in five out of eight horizons, the "globally best" methods are those that employ both economic indicators and Google Trends categories. In the United States, this pattern is even more potent: in all horizons, models that utilize both economic indicators and Google Trends categories generate the "globally best" RMSFE results. In both cases, the LASSO approach produces the lowest relative RMSFE in the second and third round of forecasting and the first round of nowcasting, while the factor model without variable pre-selection reports the lowest relative forecast error in the final step of nowcasting and the first step of backcasting.

However, the variable selection method that produces the "locally best" RMSFE results (see entries presented in bold in Table 4) is not consistent either across forecast horizons or across countries. Thus, it would be a tough assignment to select a particular dimension-reduction approach *a priori*. Table 5 provides a summary regarding the RMSFE-"best" variable selection methods. The MRP approach fares quite well for Brazil by generating the lowest forecast error in eight out of 24 cases (recall that we have eight horizons and three different initial sets of predictors), while in the United States, the LASSO method wins in nine out of 21 cases. In total, the LASSO attains the top rank in 16 cases out of 45.

Thus, our results are consistent with the majority of literature (Bańbura et al., 2013; Giannone et al., 2008) and highlight the strong performance of factor models relative to autoregressive models, especially in nowcasting and backcasting, as they can exploit more information each time new data are released. Also, the findings imply that factor models based on both economic indicators and Google Search categories generate significant forecast gains, since, in most cases, the null hypothesis of equal forecast accuracy is rejected (notice the plethora of numbers in Table 4 that are in grey boxes and signed with asterisks, meaning Diebold–Mariano test rejections) and produce the lowest forecast errors.

Finally, it is worth noting that the specification of the DFM and the decision on how to select the number of common factors might significantly affect the results. In the main set of findings, the IC_{p2} information criterion from Bai and Ng (2002) is employed. To enhance the generalizability of the results, we also utilize the IC_{p1} information criterion (see Table B.3). As can be observed from Table 4 and Table B.3, the results are relatively stable when using different information criteria.

4.3. Forecasting performance of factor models with targeted predictors

To quantify the importance of including targeted predictors in a factor model, we compare the performance of factor models combined with variable selection methods to factor models without variable pre-selection. For example, when we use only economic indicators and Google Trends categories, each variable selection model (such as LASSO and elastic net) is compared to the factor model which uses all economic indicators and Google Trends categories.

Table 6 shows the corresponding relative RMSFEs and points to several interesting results. First, it is evident that forecast gains that arise from factor models with variable selection tend to decrease as we incorporate more information and the prediction horizon shortens; that is, relative RMSFEs tend to increase as we move from forecasting to nowcasting and backcasting. In particular, in Brazil's case, when only economic indicators are included, during the forecasting horizon, variable selection methods outperform the DFM without targeted predictors on average by 20.9%, while during the nowcasting and backcasting period underperform by 7.1% and 7.4%, respectively. Similarly, for the United States, during forecasting, variable selection methods outperform on average by 3.73%, while during nowcasting and backcasting, the simple factor model without variable pre-selection outperforms on average by 0.52% and 3.54%, respectively.

When Google Trends categories are included, variable selection methods outperform the factor models that include all predictors in forecasting horizons by 24.5% for Brazil and by 7.61% for the United States. However, in nowcasting and backcasting predictions, the factor model without pre-selection outperforms factor models with targeted predictors on average by 0.5% and 24.43% for Brazil and by 2.13% and 9.15% for the United States, respectively. For Brazil, only when Google Trends subcategories are incorporated, factor models with targeted

Table 6
Forecast evaluation of factor models with targeted predictors.

	Forecast ($h = 1$)			Nowcast ($h = 0$)			Backcast ($h = -1$)	
	1	2	3	1	2	3	1	2
Brazil								
Economic indicators								
LASSO	**0.74**	0.79	**0.73**	0.89	1.27	1.23	1.10	1.20
AdaLASSO	0.82	0.87	0.80	0.96	1.40	1.40	1.21	1.37
ENET	0.85	0.85	0.79	0.92	1.19	1.19	0.99	1.04
MRP	0.74	**0.72**	0.78	**0.82**	**1.02**	**1.00**	**0.82**	**0.87**
Economic indicators and Google Trends categories								
LASSO	**0.70**	**0.63**	**0.63**	**0.73**	**0.92**	1.36	1.33	1.18
AdaLASSO	0.85	0.92	0.75	0.90	1.30	1.48	1.48	1.37
ENET	0.80	0.78	0.74	0.83	1.10	1.52	1.40	1.24
MRP	0.79	0.73	0.73	0.80	0.98	**1.17**	**1.01**	**0.97**
Economic indicators, Google Trends categories, and Google Trends subcategories								
LASSO	0.69	0.66	0.70	0.77	0.82	0.94	1.00	1.07
AdaLASSO	0.79	0.78	0.79	0.96	1.03	1.17	1.33	1.34
ENET	0.70	0.63	**0.67**	0.64	**0.58**	**0.63**	**0.58**	**0.59**
MRP	**0.65**	**0.62**	0.71	**0.64**	0.77	0.63	0.65	0.67
United States								
Economic indicators								
LASSO	0.94	**0.94**	**0.93**	**0.98**	0.96	**0.99**	1.02	
AdaLASSO	0.98	0.96	1.03	1.10	1.03	1.08	1.08	
ENET	**0.92**	0.97	0.95	0.99	0.97	1.00	**1.01**	
MRP	0.98	0.96	0.98	1.00	**0.94**	1.04	1.04	
Economic indicators and Google Trends categories								
LASSO	0.93	**0.88**	**0.90**	**0.92**	0.99	**1.07**	1.09	
AdaLASSO	0.94	0.91	1.00	1.05	1.05	1.13	1.13	
ENET	**0.91**	0.89	0.93	0.95	1.04	1.10	1.09	
MRP	0.92	0.96	0.92	0.93	**0.98**	1.09	**1.06**	
Economic indicators, Google Trends categories, and Google Trends subcategories								
LASSO	**0.98**	0.97	1.01	1.01	1.00	**0.99**	**1.00**	
AdaLASSO	1.01	0.97	**0.98**	1.05	1.01	1.04	1.03	
ENET	0.99	**0.97**	0.98	**1.00**	**0.99**	1.00	1.03	
MRP	0.99	0.99	1.02	1.00	1.03	1.03	1.02	

Notes: Entries in this table denote relative root mean squared forecast errors (RMSFEs) of dynamic factor models (DFMs) based on different variable selection methods to RMSFEs of a factor model without a variable pre-selection. Therefore, entries lower than one suggest that a particular factor model with targeted predictors is more accurate compared to factor model without variable pre-selection, for a given forecast horizon. Numbers in bold indicate model specifications with the "locally best" RMSFEs, for a given forecast horizon.

predictors tend to perform better than a factor model without pre-selection during all horizons; that is, they outperform by 30.2% in forecasting, by 21.0% in nowcasting, and by 9.7% in backcasting. For the United States, when we add Google Trends subcategories, we have the same pattern as before, where variable selection methods outperform, by 1.25% on average, only in forecasting horizons. In contrast, factor models without variable selection outperform factor models with pre-selection in nowcasting and backcasting horizons on average by 1.25% and 2.03%, respectively.

Interesting conclusions are derived if we analyze the individual performance of variable selection methods in factor models compared to those without variable pre-selection. In the case of Brazil, when only economic indicators are included in the model, the MRP approach outperforms in all horizons the factor model without variable pre-selection by 25.2%, 7.5%, and 15.7%, respectively, whereas, in the United States, the LASSO approach outperforms only in forecast and nowcast horizons by 6.5% and 2.5%, respectively.

When we add Google Trends categories, the LASSO works best for Brazil in forecasting and nowcasting and outperforms the factor model without variable selection by 34.8% and 9%, respectively, while it underperforms during backcasting by 5.1%. A similar pattern is observed for the United States, where the LASSO model outperforms the corresponding factor model without targeted predictors during forecasting and nowcasting, on average by 9.6% and 1.3%, whereas it underperforms in backcasting on average by 8.5%.

Finally, when Google Trends subcategories are included, the MRP approach performs better for Brazil in forecasting and outperforms the factor model without targeted predictors on average by 34.3%, while the elastic net works best during nowcasting and backcasting and outperforms on average by 38.5% and 41.5%, respectively. For the United States, the elastic net performs better in forecasting and nowcasting, where it outperforms on average by 2.4% and 0.7%, respectively, while in backcasting, the LASSO approach works best but produces almost the same performance compared to the factor model that includes all predictors.

Figure B.2 shows diagrammatically the forecast gains that arise when constructing factor models with targeted predictors. In particular, it shows the root mean squared forecast errors of the benchmark (AR(1) model) and the estimated factor models with and without variable preselection. As can be seen, the benchmark performs relatively well only in the case of Brazil at the $h = 1$ horizon, and especially when economic indicators and Google Trends categories and subcategories are included in the model (Figure B.2e). The factor model estimated using all predictors outperforms the rest of the models mainly at backcast horizons (i.e., $h = -1$) for both countries. It is also evident that constructing targeted predictors before the computation of common factors generates some forecast gains at the forecast and early nowcast horizons in both cases. Moreover, the volatility of the forecast errors displays interesting insights, as it seems that in the case of Brazil, RMSFEs exhibit significantly higher volatility compared to the case of the United States, in which forecast errors conducted from the models are much more stable.

Thus, benefits from pre-selecting predictors before constructing the factor model tend to arise mainly at one-quarter-ahead forecast horizons ($h = 1$), and the most significant gains from targeted predictors appear when we employ economic indicators, Google Trends categories, and Google Trends subcategories in the case of Brazil, while the most significant gains for the United States arise when we utilize economic indicators and Google Trends categories. Also, when analyzing the variable selection methods individually, the LASSO approach provides significant benefits when economic indicators and Google Trends categories are included, while the elastic net provides forecast improvements when economic indicators, Google Trends categories, and Google Trends subcategories are included.

4.4. Forecasting performance of factor models with Google Trends

In this section, we explicitly measure forecast gains that arise from the inclusion of Google Trends data by comparing the performance of each model specification that incorporates Google series (e.g., LASSO, elastic net) with the same method without Google series (e.g., LASSO, elastic net).

Table 7 directly quantifies the forecast benefits of including Google Trends data when the model already has access to information on economic conditions. For Brazil, when we include Google Trends categories, there are some gains at nowcasting and backcasting horizons. The LASSO approach also outperforms the corresponding LASSO approach without Google series during the second and the third step in forecasting. However, the value of Google data is mainly in backcasting, where it works well for all the methods other than the elastic net method (ENET), although the improvements are not always statistically significant. Interestingly, when we add Google Trends subcategories, forecast gains tend to disappear during all prediction horizons, except in ENET, which improves the model in nowcasting and backcasting horizons. On average, factor models based on Google Trends subcategories underperform in all horizons by 17.5%, 21.3%, and 15.8%, respectively.

Moving to the United States, a similar pattern is observed. Forecast gains appear only when Google Trends categories are included, and when we add the subcategories, the forecast benefits vanish. Specifically, models that incorporate Google Trends categories outperform the corresponding models without Google series during forecasting, nowcasting, and backcasting on average by 3.1%, 5.1%, and 5.4%, respectively. On the other hand, models that employ Google categories and subcategories underperform their benchmarks by 1.9%, 4.6%, and 5.5%, respectively. For the U.S., the relative gains are the greatest for MRP across all horizons from including Google Trends, although MRP is only statistically significantly better at the longest forecast horizon.

Figure B.3 shows graphically the RMSFEs of the three factor models that employ economic indicators, economic indicators, and Google Search categories; and economic indicators, Google Search categories, and Google Search subcategories. In almost all cases, the Google-based factor models produce a lower forecast error compared to the factor models that exclude Google Search data as shown by the dashed lines. The only exception is in the case of Brazil in one-quarter-ahead forecasts and early nowcasts.

Figure B.4 attempts to characterize the common factors by showing the coefficient of determination (R^2) of the regressions of individual predictors against each of the three common factors over the full sample period. The individual predictors are grouped by category. Generally speaking, the first factor for Brazil is related with *Prices* and for the U.S. with *Leading* and *Survey* indicators. The second factor loads on *EconomicActivity* and *Retail* indicators in both countries, but factor loadings also spread out to other categories. Finally, the third factor primarily reflects *Google Trends Categories*, meaning that Google Search series indeed have a significant presence in the factor model.

Overall, forecast gains when including Google data appear when we incorporate only broad Google Trends categories. When we include the subcategories, forecast gains vanish. This suggests that subcategories might repackage information already captured in the main Google Trends categories, and any further disaggregation does not appear to improve forecast performance. Also, on average, Google data seem to improve the performance of factor models to a similar extent in the United States and in Brazil. For both countries, the improvements depend on the variable-selection strategy underpinning the factor model, and are not always statistically significant.

A priori, one might suppose that Google data would be more informative about consumer behavior when the level of *discretionary* consumption is higher, and more consumers use Google, which would favor the United States. Discretionary consumption and internet penetration rates are both higher in the United States: U.S. consumers are more likely to use Google searches to inform their decisions. However, the U.S. also has high-quality alternative sources of information — the economic indicators, which are valuable in predicting the course of the

Table 7
Forecast evaluation of factor models with Google Trends.

	Forecast ($h=1$)			Nowcast ($h=0$)			Backcast ($h=-1$)	
	1	2	3	1	2	3	1	2
Brazil								
Economic indicators and Google Trends categories								
All	1.06	1.07	1.10	1.16	1.12	**0.84***	**0.75***	0.88
LASSO	1.00	**0.84***	**0.95**	**0.96**	**0.82***	0.93	0.90	**0.86***
AdaLASSO	1.09	1.13	1.04	1.09	1.04	0.89	0.92	0.88
ENET	**1.00**	0.98	1.03	1.04	1.04	1.08	1.05	1.05
MRP	1.13	1.08	1.03	1.14	1.08	0.98	0.92	0.98
Economic indicators, Google Trends categories, and Google Trends subcategories								
All	1.27	1.42	1.32	1.45	1.85	1.62	1.36	1.37
LASSO	1.19	1.17	1.26	1.24	1.20	1.23	1.23	1.21
AdaLASSO	1.22	1.28	1.29	1.44	1.36	1.36	1.49	1.34
ENET	**1.04**	**1.05**	**1.11**	**1.01**	**0.90**	**0.85***	**0.80***	**0.77***
MRP	1.10	1.22	1.19	1.12	1.40	1.02	1.07	1.06
United States								
Economic Indicators and Google Trends categories								
All	1.04	1.02	0.99	1.00	**0.92**	**0.90**	**0.90**	
LASSO	0.96	0.94	0.97	**0.94**	0.95	1.00	0.98	
AdaLASSO	0.94	0.97	0.98	0.95	0.94	0.93	0.94	
ENET	1.00	**0.93**	0.96	1.04	1.01	1.02	1.01	
MRP	**0.91**	0.94	**0.89***	0.96	0.96	0.94	0.92	
Economic indicators, Google Trends categories, and Google Trends subcategories								
All	1.04	0.98	**0.93***	**0.97**	**0.97**	1.03	1.02	
LASSO	1.02	1.00	1.01	1.11	1.06	1.07	1.04	
AdaLASSO	**0.98**	1.00	1.04	1.05	1.02	**0.99**	**1.01**	
ENET	1.10	**0.98**	1.00	1.11	1.06	1.05	1.04	
MRP	1.04	1.00	1.05	1.23	1.23	1.03	1.06	

Notes: Entries in this table denote relative root mean squared forecast errors (RMSFEs) of factor models constructed with economic indicators and Google Search series compared to factor models based only on economic indicators. Therefore, entries lower than one suggest that a particular DFM specification that utilizes Google data next to economic indicators is more accurate compared to factor models based on economic indicators of the same variable selection method, for a given forecast horizon. Numbers in bold indicate model specifications with the "locally best" RMSFEs, for a given forecast horizon. Grey boxes denote the lowest error measure achieved by the competing models for a given country and forecast horizon. Numbers followed by asterisks (**5% level, *10% level) are significantly superior to the DFM based on only economic indicators according to the Diebold–Mariano test.

economy. We hypothesize that Google data might have an edge for predicting specific categories of expenditure, such as consumer durable expenditure, although we do not consider this here.

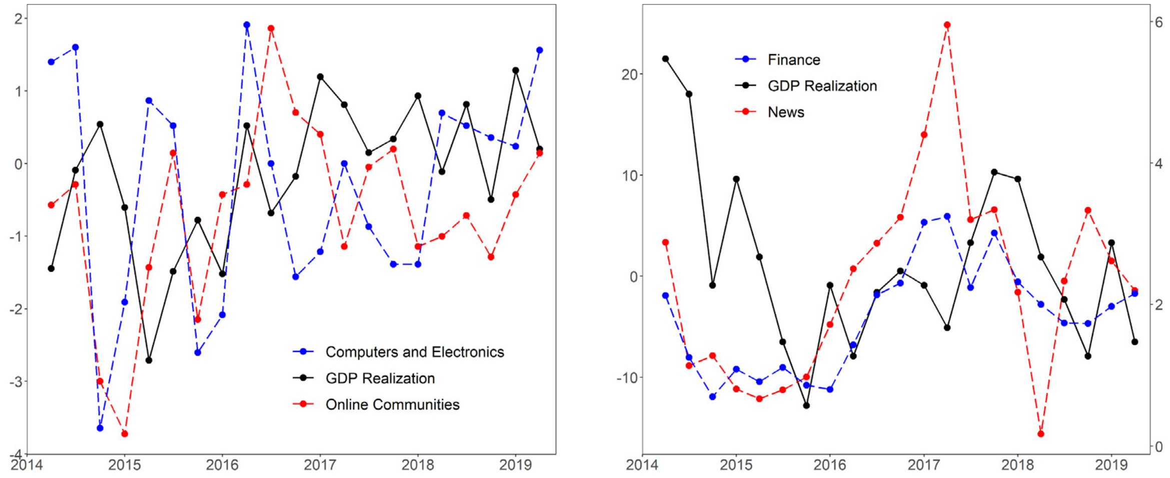
Finally, although our sample of forecast errors is necessarily short (because of the availability of Google Trends data), we provide some rudimentary analysis of whether forecast performance changes over time. Table B.1 and Table B.2 in the Online Appendix split the forecast evaluation period into two, 2014–2016 and 2017–2019, and show broadly similar patterns of results to Table 7. For example, Google Trends data generate relative improvements on both sub-periods using LASSO for Brazil.

4.5. Selected predictors and Google Trends categories

In the final part of our analysis, we focus on the predictors selected by the most-recent-performance method. Figure B.5 shows the importance of each variable group for a given model for each country. The values in the graphs are scaled such that they sum to one. A first glimpse at the graphs reveals that when only economic indicators are included, the composition of predictors remains relatively stable during the out-of-sample period. However, it is clear that when Google Search series are incorporated, the share of each group of predictors becomes quite volatile, especially when all Google Search series are included, i.e., Google Trends categories and subcategories.

It is worth mentioning that in the case of Brazil, the dataset with targeted predictors is being updated 23 times through the out-of-sample period, while in the case of the United States, the dataset is being updated 45 times. The difference in the number of times the datasets are updated is explained by the fact that the publication delay for Brazilian GDP (around eight weeks) is two times longer than the publication delay of the United States GDP (around four weeks). Thus, in the United States, we have a more frequently updated balanced dataset and can re-estimate the variable selection methods more often.

Moreover, interesting insights are gained when analyzing the individual Google Search categories that are selected by the MRP method. Table 8 exhibits the 10 most frequent Google series for each country. Somewhat surprisingly, the most frequently selected category for Brazil is Online Communities, selected in 96% of cases. This category includes queries related to community websites and social networks. The next most frequently chosen categories are Computers & Electronics and Business &



(a) Brazil (b) United States

Fig. 2. Percent change of GDP and Google Trends categories.
Notes: The left-hand (right-hand) plot shows the evolution of GDP growth rates (quarter-on-quarter) and the two most frequent Google Trends categories selected by the MRP method for Brazil (the United States).

Table 8
Google Trends categories selected.

Brazil		United States	
Category	Frequency	Category	Frequency
Online Communities	96%	News	100%
Computers & Electronics	91%	**Finance**	**96%**
Business & Industrial	91%	**Home & Garden**	**96%**
Sports	**87%**	Reference	93%
Health	87%	Jobs & Education	91%
Finance	**83%**	**Sports**	**89%**
Internet & Telecom	83%	Science	89%
Games	74%	Hobbies & Leisure	87%
Home & Garden	**74%**	Food & Drink	87%
Shopping	74%	Beauty & Fitness	82%

Notes: Bold entries denote common Google Trends categories for both countries.

Industrial, which it seems logical to connect with overall economic activity, as they incorporate queries such as computer hardware, consumer electronics, agriculture, and construction.

Turning to the United States, the most frequently selected category is News, which includes queries such as business news, local and world news, and politics, and is selected in all cases. Next, Finance is selected in 96% of cases and incorporates search terms that are directly associated with the state of the economy, such as investing, lending, and insurance. Finally, the third most frequently selected category is Home & Garden, which contains searches related to home improvement, appliances, and furnishings. Fig. 2 shows the evolution of GDP growth rates and the two most frequently selected Google Search categories for each country.

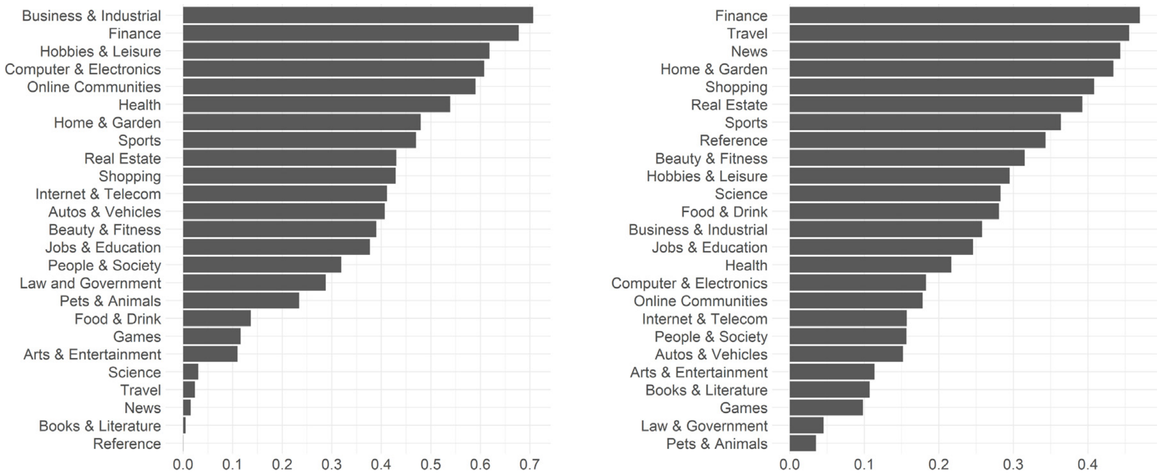
It is noteworthy that only three categories are common in both lists with categories that are most frequently selected by the MRP method: Sports, Finance, and Home & Garden. Clearly, these categories are related to the services sector. It is reasonable to expect that the majority of Google Trends categories and common categories are connected with services, since the tertiary sector contributes the most in both countries.[8] The fact that in the case of Brazil there are a few categories related to the manufacturing sector highlights the structural differences between the economies. Brazil is heavily dependent on the primary and secondary sectors, and this is reflected in the frequent usage of categories like Computers & Electronics and Business & Industrial. On the other hand, the U.S. economy has a relatively larger services sector, and intuitively, search terms related to News and Finance are the most frequently used.

Finally, Fig. 3 shows the contribution of each Google Trends category to the first Google Trends common factor by regressing the common factor against each Google category and then extracting the R^2. Categories like Business & Industrial, Finance, and Hobbies & Leisure seem to make a significant contribution to the Brazilian Google Trends factor, while categories like Finance, Travel, and News appear to make a substantial contribution to the U.S. common factor.

5. Concluding remarks

Although many studies have investigated the potential usefulness of big data for forecasting specific macroeconomic variables such as unemployment and inflation, only a few have focused on overall economic activity — GDP. This paper contributes to the literature by exploring whether Google Search data complements more traditional economic indicators to provide forecast gains in a nowcasting exercise.

8 The services sector consists of around 77% of GDP in the United States and 63% in Brazil (Statista, 2018).



(a) Brazil (b) United States

Fig. 3. Contribution of Google Trends categories to the Google factor.
Notes: The left-hand (right-hand) plot shows the contribution of Google Trends categories to the first Google Trends common factor computed for Brazil (the United States). The contribution of each variable is measured by the coefficient of determination (R^2) extracted from regressions of the first Google Trends common factor against the individual Google Trends categories over the entire sample.

In a pseudo-real-time framework, we estimated dynamic factor models to nowcast GDP growth rates for Brazil and the United States, from 2014 to 2019. We considered the efficacy of several variable selection methods, including the elastic net, the least absolute shrinkage and selection operator (LASSO), and an adaptive version of the LASSO. This provided evidence on whether forecast gains arise from estimating factor models on targeted predictors. Additionally, we constructed sets of targeted predictors from what we call the most-recent-performance approach. Every time the balanced dataset of predictors is updated with new information, we choose the variable selection method that produces the "best" local out-of-sample performance.

There were a number of findings. Firstly, factor models effectively incorporated the new information published within the reference quarter. Factor models outperformed the AR(1) benchmark at all horizons for the United States, while for Brazil, they outperformed the benchmark mostly at nowcasting and backcasting horizons.

Secondly, we found that factor models that utilize both economic indicators and Google Trends categories had the best performance against the autoregressive model. Variable selection methods worked best at forecast horizons, with diminishing performance as additional information was included (now- and back-casting). Perhaps unsurprisingly, the largest gains from variable selection occurred when both Google data and economic indicators were allowed, rather than when the information set was restricted to economic indicators.

Thirdly, only the main Google Trends categories were found to have the potential to predict GDP growth rates, since when we added the subcategories to the model, forecast gains vanished. This suggests that the information contained in subcategories is already captured by the main categories.

Finally, our results do not clearly suggest that Google data were more valuable for one country than the other. For both countries, their value depended on the chosen variable selection method.

Although our main conclusions regarding the overall usefulness of Google Trends data in forecasting economic activity are in line with Ferrara and Simoni (2019) and Götz and Knetsch (2019), they contrast with the former study with respect to the horizons at which Google Search series are useful. However, our results corroborate the findings of both papers concerning the benefits of a variable-selection step for Google Trends data.

Finally, there are potential limitations and extensions. First, our pseudo-real-time framework: Although data revisions might only have a small effect on the computation of common factors, the use of real-time vintages of data for GDP growth may affect the findings. However, Bernanke and Boivin (2003) and Clements (2016) find that the relative forecasting performance of factor models and AR models in real-time and pseudo-out-of-sample exercises is similar. Nevertheless, this may depend on the nature of the revisions: see Clements and Galvão (2019) for a recent review of data revisions and forecasting. A possible extension would be to consider the conceptually distinct effects of model re-estimation and the newly released data points when a nowcast is revised as more data become available. This exercise might usefully be disaggregated across different blocks of variables.

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

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2022.10.003.

References

Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191–221.
Bai, J., & Ng, S. (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics*, *25*(1), 52–60.
Bai, J., & Ng, S. (2008a). Forecasting economic time series using targeted predictors. *Journal of Econometrics*, *146*(2), 304–317.
Bai, J., Ng, S., et al. (2008b). Large dimensional factor analysis. *Foundations and Trends in Econometrics* , *3*(2), 89–163.
Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow. In *Handbook of Economic Forecasting (Vol. 2)* (pp. 195–237). Elsevier.
Barhoumi, K., Darné, O., & Ferrara, L. (2010). Are disaggregate data useful for factor analysis in forecasting French GDP? *Journal of Forecasting*, *29*(1–2), 132–144.
Bernanke, B. S., & Boivin, J. (2003). Monetary policy in a data-rich environment. *Journal of Monetary Economics*, *50*(3), 525–546.
Boivin, J., & Ng, S. (2006). Are more data always better for factor analysis? *Journal of Econometrics*, *132*(1), 169–194.
Borup, D., & Schütte, E. C. M. (2022). In search of a job: Forecasting employment growth using Google Trends. *Journal of Business & Economic Statistics*, *40*(1), 186–200.
Bragoli, D., & Fosten, J. (2018). Nowcasting Indian GDP. *Oxford Bulletin of Economics and Statistics*, *80*(2), 259–282.
Bulligan, G., Marcellino, M., & Venditti, F. (2015). Forecasting economic activity with targeted predictors. *International Journal of Forecasting*, *31*(1), 188–206.
Buono, D., Mazzi, G. L., Kapetanios, G., Marcellino, M., & Papailias, F. (2017). Big data types for macroeconomic nowcasting. *Eurostat Review on National Accounts and Macroeconomic Indicators*, *1*(2017), 93–145.
Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting*, *32*(4), 289–298.
Caruso, A. (2018). Nowcasting with the help of foreign indicators: The case of Mexico. *Economic Modelling*, *69*, 160–168.
Cepni, O., Güney, I. E., & Swanson, N. R. (2019). Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes. *International Journal of Forecasting*, *35*(2), 555–572.
Choi, H., & Varian, H. (2009). *Predicting initial claims for unemployment benefits* (pp. 1–5). Google Inc.
Clements, M. P. (2016). Real-time factor model forecasting and the effects of instability. *Computational Statistics & Data Analysis*, *100*, 661–675.
Clements, M. P., & Galvão, A. B. (2019). Data revisions and real-time forecasting. In *The Oxford Research Encyclopedia of Economics and Finance*. Oxford University Press.
Dahlhaus, T., Guénette, J. D., & Vasishtha, G. (2017). Nowcasting BRIC+M in real time. *International Journal of Forecasting*, *33*(4), 915–935.
D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801–816.
Diebold, F. X. (2003). Big data dynamic factor models for macroeconomic measurement and forecasting.
Diebold, F. X. (2021). What's the big idea? "Big Data" and its origins. *Significance*, *18*(1), 36–37.
Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.
Doornik, J. A., & Hendry, D. F. (2015). Statistical model selection with "Big Data". *Cogent Economics & Finance*, *3*(1), Article 1045216.
Doz, C., & Fuleky, P. (2020). Dynamic factor models. *Macroeconomic Forecasting in the Era of Big Data*, 27–64.
Doz, C., Giannone, D., & Reichlin, L. (2011). A two-step estimator for large approximate dynamic factor models based on Kalman filtering. *Journal of Econometrics*, *164*(1), 188–205.
Ferrara, L., & Simoni, A. (2019). When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage.
Foroni, C., & Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates. *International Journal of Forecasting*, *30*(3), 554–568.
Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, *55*(4), 665–676.
Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for German GDP. *International Journal of Forecasting*, *35*(1), 45–66.
Jansen, W. J., Jin, X., & de Winter, J. (2016). Forecasting and nowcasting real GDP: Comparing statistical models and subjective forecasts. *International Journal of Forecasting*, *32*(2), 411–436.
Kapetanios, G., Papailias, F., et al. (2018). Big data & macroeconomic nowcasting: Methodological review. In *Economic statistics centre of excellence*. National Institute of Economic and Social Research.
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.
Luciani, M., Pundit, M., Ramayandi, A., & Veronese, G. (2018). Nowcasting Indonesia. *Empirical Economics*, *55*(2), 597–619.
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.
Medeiros, M. C., & Pires, H. F. (2021). The proper use of Google Trends in forecasting models. ArXiv preprint arXiv:2104.03065.
Panagiotelis, A., Athanasopoulos, G., Hyndman, R. J., Jiang, B., & Vahid, F. (2019). Macroeconomic forecasting for Australia using a large number of predictors. *International Journal of Forecasting*, *35*(2), 616–633.
Schiavoni, C., Palm, F., Smeekes, S., & van den Brakel, J. (2021). A dynamic factor model approach to incorporate big data in state space models for official statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *184*(1), 324–353.
Schumacher, C. (2010). Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters*, *107*(2), 95–98.
Seabold, S., & Coppola, A. (2015). *Nowcasting prices using google trends: An application to central america*. The World Bank.
Smeekes, S., & Wijler, E. (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting*, *34*(3), 408–430.
Stock, J. H., & Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.
Stock, J. H., & Watson, M. W. (2011). Dynamic Factor Models, chapter 2. *The Oxford Handbook of Economic Forecasting* (pp. 35–59). Oxford University Press,
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.
Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google Trends. *Journal of Forecasting*, *30*(6), 565–578.
Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.