When are Google data useful to nowcast GDP? An approach via pre-selection and shrinkage[1]

Laurent Ferrara*
Banque de France

Anna Simoni[2]
CREST, CNRS

January 14, 2019

Abstract

Nowcasting GDP growth is extremely useful for policy-makers to assess macroeconomic conditions in real-time. In this paper, we aim at nowcasting euro area GDP with a large database of Google search data. Our objective is to check whether this specific type of information can be useful to increase GDP nowcasting accuracy, and when, once we control for official variables. In this respect, we estimate shrunk bridge regressions that integrate Google data optimally screened through a targeting method, and we empirically show that this approach provides some gain in pseudo-real-time nowcasting of euro area GDP quarterly growth. Especially, we get that Google data bring useful information for GDP nowcasting for the four first weeks of the quarter when macroeconomic information is lacking. However, as soon as official data become available, their relative nowcasting power vanishes. In addition, a true real-time analysis confirms that Google data constitute a reliable alternative when official data are lacking.

Keywords: Nowcasting, Big data, Google search data, Sure Independence Screening, Ridge Regularization.

[1]We would like to thank Roberto Golinelli, Michele Lenza, Francesca Monti, Giorgio Primiceri, Simon Sheng, Hal Varian and the participants of the 10th ECB Conference on *Macro forecasting with large datasets* for useful comments. We would like to thank Per Nymand-Andersen (ECB) for sharing the Google dataset as well as Dario Buono and Rosa Ruggeri-Cannata (Eurostat) for sending the real-time euro area GDP data. We are grateful to Vivien Chbicheb for outstanding research assistance. A first version of this paper was circulated under the title: *Macroeconomic nowcasting with big data through the lens of a targeted factor model*. This paper represents the authors's personal opinions and does not necessarily reflect the view of the Banque de France.

*Banque de France, International Macroeconomics Division, 39 rue Croix des Petits Champs, Paris, France, e-mail: laurent.ferrara@banque-france.fr

[2]CREST, CNRS, École Polytechnique, ENSAE - 5, avenue Henry Le Chatelier, 91120 Palaiseau, France, e-mail: simoni.anna@gmail.com

1 Introduction

Big datasets are now widely used by practitioners for short-term macroeconomic forecasting and nowcasting purposes. In this paper, we ask the question whether such data are still useful when controlling for official variables, such as opinion surveys or production, generally used by forecasters. And if so, when exactly are those alternative data actually adding a gain in nowcasting accuracy, both in quasi and true real-time frameworks. In this respect, we focus on Google search data and assess their ability to provide useful information to nowcast the euro area quarterly GDP growth rate. We empirically show that they are indeed useful, but only when official data are not available to practitioners, that is during the first four weeks at the beginning of the quarter.

Nowcasting GDP growth is extremely useful for policy-makers to assess macroeconomic conditions in real-time. The concept of macroeconomic nowcasting has been popularized by many researchers (see e.g. Giannone et al. [2008]) and differs from standard forecasting approaches in the sense it aims at evaluating current macroeconomic conditions on a high-frequency basis. The idea is to provide policy-makers with a real-time evaluation of the state of the economy ahead of the release of official Quarterly National Accounts, that always come out with a delay. For example, the New York Fed and the Atlanta Fed have recently developed new tools in order to evaluate US GDP quarterly growth on a high-frequency basis[1]. The tool developed by the Atlanta Fed, referred to as *GDPNow*, is updated 6 to 7 times per month, while the NY Fed's tool is updated every Friday. With reference to countries other than the US, many papers have put forward econometric modelling to nowcast GDP growth in advanced countries (see among others Frale et al. [2010] or Kuzin et al. [2011] for the euro area, Aastveit and Trovik [2012] for Norway or Bragoli [2017] for Japan), as well as in emerging countries (see for example Modugno et al. [2016] for Turkey or Bragoli et al. [2015] for Brazil). Some researchers have also proposed approaches to nowcast economic output at a global level in order to assess on a regular high-frequency basis world economic conditions (see Ferrara and Marsilli [2018] or Golinelli and Parigi [2014]).

In the existing literature, GDP nowcasting tools integrate standard official macroeconomic information stemming, for instance, from National Statistical Institutes, Central Banks, International Organizations. Typically, three various sources of official data are considered: (i) hard data, like production, sales, employment, (ii) opinion surveys (households or companies are asked about their view on current and future economic conditions), and (iii) financial markets information (sometimes available on high fre-

[1]See the websites https://www.newyorkfed.org/research/policy/nowcast and https://www.frbatlanta.org/cqer/research/gdpnow.aspx

quency basis). However, more recently, a lot of emphasis has been put on the possible gain that forecasters can get from using alternative sources of high-frequency information, referred to as *Big Data* (see for example Varian [2014], Giannone et al. [2017] or Buono et al. [2018]). Various sources of *Big Data* have been used in the recent literature such as for example web scraped data, scanner data or satellite data. One of the main source of alternative data is Google search and seminal papers on the use of such data for forecasting are the ones by Choi and Varian [2009] and Choi and Varian [2012] (see also Scott and Varian [2015] who combine Kalman filters, spike-and-slab regression and model averaging to improve short-term forecasts).

Overall, empirical papers show evidence of some forecasting power for Google data, at least for some specific macroeconomic variables such as consumption (Choi and Varian [2012]), unemployment rate (D'Amuri and Marcucci [2012]), building permits (Coble and Pincheira [2017]) or car sales (Nymand-Andersen and Pantelidis [2018]). However, when correctly compared with other sources of information, the jury is still out on the gain that economists can get from using Google data for forecasting and nowcasting. For example, Vosen and Schmidt [2011] show that Google Trends data lead to an accuracy gain when compared with business surveys to forecast the annual growth rate of US household consumption. But some other papers tend to show that the gain in forecasting using Google data is very weak when other sources of information are accounted for in the analysis. For example, Goetz and Knetsch [2019] estimate German GDP using simultaneously both official and Google data on a monthly basis and show that adding Google data only leads to limited accuracy gains. However, they provide some evidence that those data can be a potential alternative to survey variables. We also refer to Li [2016] on this issue. Overall, it seems that Google data can be extremely useful when economist do not have access to information or when information is fragmented, as for example when dealing with emerging economies (see Carriere-Swallow and Labbe [2013]) or low-income developing countries (Narita and Yin [2018]).

In this paper, we estimate both pseudo real-time and true real-time nowcasts for the euro area quarterly GDP growth between 2014q1 and 2016q1 by plugging Google data into the analysis, in addition to official variables on industrial production and opinion surveys, commonly used as predictors for GDP growth. Google data are indexes of weekly volume changes of Google searches by keywords in the six main euro area countries about different topics which are gathered in 26 broad categories such as auto and vehicles, finance, food and drinks, real estate, etc. Those broad categories are then split into a total of 269 sub-categories per country, leading to a total of 1776 variables

for each country[2]. Our objective is to assess whether Google search data bring some gain in nowcasting accuracy and when. The approach that we carry out is deliberately extremely simple and relies on a bridge equation that integrates variables selected from a large set of Google data, as proposed by Angelini et al. [2011]. More precisely, we pre-select Google variables by targeting GDP growth in the vein of Bai and Ng [2008] but with a different approach. Pre-selection is implemented by using the Sure Independence Screening method put forward by Fan and Lv [2008] enabling to preselect the Google variables the most related to GDP growth before entering the bridge equation. After pre-selection we use Ridge regularization to estimate the bridge equation as the number of pre-selected variables may still be large.

Four main stylized facts come out from our empirical analysis. First, we point out the usefulness of Google search data for nowcasting euro area GDP for the first four weeks of the quarter when there is no available official information about the state of the economy. Indeed, we show that at the beginning of the quarter, Google data provide an accurate picture of the GDP growth rate. Against this background, this means that such data are a good alternative in the absence of official information and can be used by policy-makers. Second, we get that as soon as official data become available, that is starting from the fifth week of the quarter, then the gain from using Google data for GDP nowcasting rapidly vanishes. This result contributes to the debate on the use of big data for short-term macroeconomic assessment when controlling for standard usual macroeconomic information. Third, we show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy in terms of nowcasting accuracy. Indeed, this approach enables to retain only Google variables that have some link with the targeted variable. This result confirms previous analyses that have been carried when dealing with large datasets through dynamic factor models (see e.g. Bai and Ng [2008] or Schumacher [2010]). Finally, we carry out a true real-time analysis by nowcasting euro area GDP growth rate using the official Eurostat timeline and vintages of data. We show that the three previous results still hold in real-time, in spite of an expected increase in the size of errors, suggesting that Google search data can be effectively used in practice to help the decision-making process.

The rest of the paper is organized as follows. In Section 2 we describe the model we consider for nowcasting, the Sure Independence Screening (SIS) approach to pre-select the data, as well as the Ridge regularization. Section 3 describes the structure of the

[2]See for example Bontempi et al. [2018] for a detailed description of this dataset, in a different framework

Google search data used for nowcasting. The empirical results are presented in Section 4 and Section 5 concludes.

2 Methodology

2.1 The nowcasting approach

In order to get GDP nowcasts, we focus on linear bridge equations that link quarterly GDP growth rates and monthly economic variables. The classical bridging approach is based on linear regressions of quarterly GDP growth on a small set of key monthly indicators as for example in Diron [2008]. In our exercise, in addition to those monthly variables, we also consider Google data, available at a higher frequency, and we aim at assessing their nowcasting power. More precisely, Google data are available on a weekly basis, providing thus additional information when official information is not yet available. Even if Google data are not on average extremely correlated with the GDP growth rate, we are going to show that they still provide accurate GDP nowcasts if conveniently treated.

Therefore, we assume that we have three types of data at disposal: *soft* data, such as opinion surveys, *hard* data, such as industrial production or sales, and data stemming from Google search machines. Let t denote a given quarter of interest identified by its last month, for example the first quarter of 2005 is dated by $t = \text{March2005}$. A general model to nowcast the growth rate of any macroeconomic series of interest Y_t for a specific quarter t is the following, for $t = 1, \ldots, T$:

$$Y_t = \beta_0 + \beta_s' x_{t,s} + \beta_h' x_{t,h} + \beta_g' x_{t,g} + \varepsilon_t, \qquad \mathbf{E}[\varepsilon_t | x_{t,s}, x_{t,h}, x_{t,g}] = 0, \tag{2.1}$$

where $x_{t,s}$ is the N_s-vector containing *soft* variables, $x_{t,h}$ is the N_h-vector containing *hard* variables, $x_{t,g}$ is the N_g-vector of variables coming from Google search and ε_t is an unobservable shock. In our empirical analysis Y_t is the quarterly GDP growth rate of the euro area. Because variables $x_{t,s}$, $x_{t,h}$ and $x_{t,g}$ are sampled over different frequencies (monthly *vs* weekly), the relevant dataset for calculating the nowcast evolves within the quarter. By denoting with $x_{t,j}^{(w)}$, $j \in \{s, h, g\}$, the j-th series released at week $w = 1, \ldots, 13$ of quarter t, we denote the relevant information set at week w of a quarter t by

$$\Omega_t^{(w)} := \{x_{t,j}^{(w)}, j \in \{s, h, g\} \text{ such that } x_{t,j} \text{is released at } w\}.$$

For simplicity, we keep in $\Omega_t^{(w)}$ only the observations relative to the current quarter t

and do not consider past observations. While $x_{t,g}$ is in $\Omega_t^{(w)}$ for every $w = 1, \ldots, 13$, the other variables are in the relevant information set only for the weeks corresponding to their release and so the dataset is unbalanced.

To explicitly account for the different frequencies of the variables, we replace model (2.1) by a model for each week w such that:

$$\begin{aligned} \widehat{Y}_{t|w} &= \mathbf{E}[Y_t|\Omega_t^{(w)}], \qquad t = 1, \ldots, T \quad \text{and} \quad w = 1, \ldots, 13 \\ \text{and} \quad \mathbf{E}[Y_t|\Omega_t^{(w)}] &= \beta_{0,w} + \beta'_{s,w} x_{t,s}^{(w)} + \beta'_{h,w} x_{t,h}^{(w)} + \beta'_{g,w} x_{t,g}^{(w)} \end{aligned} \tag{2.2}$$

where $x_{t,j}^{(w)} = 0$ if $x_{t,j}^{(w)} \notin \Omega_t^{(w)}$. For instance, as the first observation of industrial production relative to the current quarter t is only released in week 9, then we set $x_{t,h}^{(w)} = 0$ for every $w = 1, \ldots, 8$. The bridge equation (2.2) exploits weekly information to obtain more accurate nowcasts of quarterly GDP growth.

2.2 Pre-selection of Google data

The recent literature on nowcasting and forecasting with large datasets comes to the conclusion that using the largest available dataset is not necessarily the optimal approach when aiming at nowcasting a specific macroeconomics variable such as GDP, at least in terms of nowcasting accuracy. For example, against the background of bridge equations augmented with dynamic factors, Barhoumi et al. [2010] empirically show that factors estimated on a small database lead to competitive results for nowcasting French GDP compared with the most disaggregated data. From a theoretical point of view, Boivin and Ng [2006] suggest that larger databases lead to poor forecast when idiosyncratic errors are cross-correlated or when the forecasting power comes from a factor that is dominant in a small database but is dominated in a larger dataset. An empirical way to circumvent this issue is to target more accurately the variable to be nowcast. For example, Bai and Ng [2008] show that forming targeted predictors enables to improve the accuracy of inflation forecasts while Schumacher [2010] shows that targeting German GDP within a dynamic factor model is a performing strategy.

In this respect, all the categories and subcategories in the Google search data are not necessarily correlated with the GDP growth that we want to nowcast. Therefore, using all the variables in the Google search dataset is not necessarily a good strategy because one would pay the price of dealing with ultra-high dimensionality without increasing the nowcasting accuracy as measured by the Mean Squared Forecasting Error (MSFE). For this reason we consider a pre-selection procedure before using data for nowcasting, that

is, a procedure enabling to select a subset of the variables in the Google search dataset that are the most relevant for GDP growth nowcasting. In a second step, we will use a Ridge regularization to estimate model (2.2) by using the selected subset of Google data, which is the most "related" with the variable Y_t and which captures much of the variability in GDP growth. As explained in Section 2.3 below, a regularization technique is required because the number of selected variables can still be large while not ultra-high.

While in our empirical analysis we have tried several pre-selection procedures, it turns out that the innovative approach put forward by Fan and Lv [2008] appears to provide interesting and intuitive results. This approach is referred to as *Sure Independence Screening*, or SIS hereafter. Sure screening refers to the property that *all important variables survive after applying a variable screening procedure with probability tending to 1* (see Fan and Lv [2008], p. 853). The basic idea of this approach is based on correlation learning and relies on the fact that only the variables with the highest absolute correlation should be used in modelling.

Let us start from a standard linear regression equation with only the standardized N_g Google variables as explanatory variables, that is $\beta_0 = \beta_s = \beta_h = 0$ in equation (2.1). Let $M^* = \{1 \leq j \leq N_g : \beta_{g,j} \neq 0\}$ be the true sparse model with non-sparsity size $s = |M^*|$. The other $N_g - s$ variables can also be correlated with Y via linkage to the predictors contained in the true sparse model. Let Y denote the T-vector of quarterly GDP growth: $Y = (Y_1, \ldots, Y_T)'$. We compute $\omega = (\omega_1, \ldots, \omega_{N_g})'$, the vector of marginal correlations of predictors with the response variable Y_t, such as

$$\omega = \overline{X}_g' Y, \tag{2.3}$$

where $\overline{X}_g$ is the $T \times N_g$ matrix of average Google data where the average is taken over each quarter and that then has been centered and standardized columnwise. The average over each quarter is taken to make the weekly Google data comparable to the quarterly GDP growth data in terms of frequency. For any given $\lambda \in]0, 1[$, the N_g componentwise magnitudes of the vector ω are sorted in a decreasing order and we define a submodel M_λ such as: $M_\lambda = \{1 \leq j \leq N_g : |\omega_j| \text{ is among the first } [\lambda T] \text{ largest of all } \}$, where $[\lambda T]$ denotes the integer part of λT. Since only the order of componentwise magnitudes of ω is used, this procedure is invariant under scaling and thus it is identical to selecting predictors using their correlations with the response. This approach is an easy way to filter out Google variables with the weaker correlations with GDP growth rate so that we are left with $d = [\lambda T] < T$ Google variables. An important feature of the SIS procedure is that it uses each covariate $x_{t,g,j}$ independently as a predictor to decide how useful it

is for predicting Y_t.

This method is desirable because it has the sure screening property, that is, with probability tending to one, all the important variables in the true model are retained after applying this method. In fact, Fan and Lv [2008, Theorem 1] show that under Normality of ε_t and other conditions (see Fan and Lv [2008, Conditions 1-4]) the sure screening property holds, namely for a given λ:

$$P(M^* \subset M_\lambda) \to 1$$

as $N_g \to \infty$. In particular, SIS can reduce the dimension to $[\lambda T] = O(T^{1-\theta}) < T$ for some $\theta > 0$ and the reduced model M_λ still contains all the variables in the true model M^* with a probability converging to one.

In the following, we write $X_{t,M_\lambda} = [1, x'_{t,s}, x'_{t,h}, x'_{t,g,M_\lambda}]'$, where $x_{t,g,M_\lambda} = \{x_{t,g,j}; j \in M_\lambda\}$ is the vector containing only the selected Google variables. Moreover, for a vector $\beta \in \mathbb{R}^p$ and a set $M \subset \{1, \dots, p\}$ we write M^c for the complement of M in $\{1, \dots, p\}$ and $\beta_M = \{\beta_j; j \in M\}$. The empirical choice of the hyperparameter λ is discussed in subsection 3.3.

2.3 Ridge regression

Google search data have an extremely large dimension, with the number of variables much larger than the number of observations (i.e. $N_g >> T$, sometimes referred to as *fat datasets*). Therefore, when using Google search data for nowcasting one has to deal with such high dimensionality. Even after implementing the pre-selection described in subsection 2.2, the number of Google variables remains large compared to the time dimension T. Therefore, one needs to use a machine learning technique suitable to treat fat datasets.

One of the most popular ways to deal with a large number of covariates and possibly problems of multicollinearity is the Ridge regression. Let $\beta = (\beta_0, \beta'_s, \beta'_h, \beta'_g)'$. Ridge regression estimates equation (2.1) by minimizing a penalized residuals sum of squares where the penalty is given by the Euclidean squared norm $\|\cdot\|$:

$$\widehat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta'_s x_{t,s} - \beta'_h x_{t,h} - \beta'_g x_{t,g} \right)^2 + \alpha \|\beta\|^2 \right\},$$

where $\alpha > 0$ is regularization parameter that tunes the amount of shrinkage. The estimated coefficients in $\widehat{\beta}$ are then shrunk towards zero. By using model (2.2) for each

week w and the pre-selection methodology in Section 2.2 we can compute the Ridge estimator after model selection:

$$\widehat{\beta}^{(w)}_{Ridge} := \arg \min_{\beta;\beta_{g,j}=0,j\in M^c_\lambda} \left\{ \frac{1}{T}\sum_{t=1}^{T} \left(Y_t - \beta_0 - \beta_s' x^{(w)}_{t,s} - \beta_h' x^{(w)}_{t,h} - \beta_g' x^{(w)}_{t,g} \right)^2 + \alpha \|\beta\|^2 \right\}.$$

Equivalently, we can write $\widehat{\beta}^{(w)}_{Ridge} = (\widehat{\beta}^{(w)'}_{Ridge,M_\lambda}, \widehat{\beta}^{(w)'}_{Ridge,M^c_\lambda})'$ where

$$\widehat{\beta}^{(w)}_{Ridge,M_\lambda} = \left(\frac{1}{T}\sum_{t=1}^{T} X_{t,M_\lambda} X'_{t,M_\lambda} + \alpha I \right)^{-1} \frac{1}{T}\sum_{t=1}^{T} X'_{t,M_\lambda} Y_t, \qquad \widehat{\beta}^{(w)}_{Ridge,M^c_\lambda} = 0$$

and I is the $|M_\lambda|$-dimensional identity matrix. This is the estimator we are going to use in our empirical analysis. Even if it depends on α in a crucial way, we leave implicit this dependence. The empirical choice of the hyperparameter α is a crucial issue because it has an important impact on the nowcasting accuracy. We discuss this choice in Section 3.3

3 Design of the empirical analysis

This section first describes the data used in the empirical analysis. Then, it describes how to deal with the various reporting lags. Finally, we propose a way to select both hyperparameters λ and α involved in the estimation procedure.

3.1 Data

Our objective in this paper is to assess the role of Google data for nowcasting the euro area GDP, especially to assess (i) if these big data are relevant when there is no official data available for the forecaster and (ii) to what extent these data provide useful information when official data become available. In this respect, the variable Y_t in model (2.1)-(2.2) that we target is the quarterly growth rate of the real euro area GDP, stemming from Eurostat. The official data that we consider are of two kinds: industrial production for the euro area as a whole provided by Eurostat, which is a global measure of hard data and is denoted by IP_t, and a composite index of opinion surveys from various sectors computed by the European Commission (the so called *euro area Sentiment Index*) denoted by S_t.

Our big dataset covers Google searches for the six main euro area countries: Belgium, France, Germany, Italy, Netherlands and Spain. We have at disposal a total of $N_g = 1776$ variables, corresponding to 26 categories and 296 subcategories for each country. Google search data are data related to queries performed with Google search. The data are indexes of weekly volume changes of Googles queries grouped by category and by country. Data are normalized at 1 at the first week of January 2004 which is the first week of availability of these data. Then, the following values indicate the deviation from the first value. However, there is no information about the search volume. Google data are weekly data that are received and made available by the European Central Bank every Tuesday. Original data are not seasonally adjusted, thus we take the growth rate over 52 weeks to eliminate the seasonality within the data.[3]

We use data from 20 March 2005 (twelfth week of the first quarter) until 29 March 2016 (thirteen week of the first quarter). We split the sample in two parts and use data starting from the first week of January 2014 for the out-of-sample analysis.

3.2 Dealing with various reporting lags

An important feature of all these data is that they are released with various reporting lags, leading thus to non-balanced information dataset at each point in time. In the literature, this issue is refereed to as *ragged-edge database* (see Angelini et al. [2011]). For instance, Google search data are weekly data available every Tuesday, while the soft and hard data are monthly data available at the end of every month and at the middle of the third month of the quarter, respectively. Treating weekly data is particularly challenging as the number of entire weeks present in every quarter is not always the same, and a careful analysis has to be done when incorporating these data. In addition, there is a frequency mismatch in the data as the explained variable is quarterly and the explanatory variables are either weekly (Google data) or monthly (hard and soft variables). In order to account for the various frequencies and the timing at which the predictive variables become available, we adopt the strategy to consider a different model for every week of the quarter as described in Section 2.1. Thus we end up with thirteen models given in equation (2.2), each model including the variables available at this date.

As regards the dates of availability, we mimic the exact release dates as published by

[3]Applying this standard seasonal filter eliminates a large part of seasonal effects in the dataset. We also test for outliers in our study, but dropping detected outliers does not seem to improve nowcasting accuracy. Obviously this question needs to be tackle in more details in further research.

Eurostat. This means that the first survey of the quarter, referring to the first month, typically arrives in week 5. Then, the second survey of the quarter, related to the second month, is available in week 9. Industrial production for the first month of the quarter is only available about 45 days after the end of the reference month, that is generally in week 11. Finally, the last survey, related to the third month of the quarter, is available in week 13. A scheme of the release timeline is presented in Figure 1.

By denoting with $x_{t,g,M_\lambda,w}$ the vector of pre-selected variables from the Google search data for week w of quarter t we construct the variable $x_{t,g}^{(w)}$ in equation (2.2) as the average of the vector of selected Google variables up to the w-th week: $x_{t,g}^{(w)} = \sum_{v \leq w} x_{t,g,M_\lambda,v}$. That is, take for instance $w = 3$ (*i.e.* Model 3 which is used at week 3), then $x_{t,g}^{(3)}$ is equal to $(x_{t,g,M_\lambda,1} + x_{t,g,M_\lambda,2} + x_{t,g,M_\lambda,3})/3$.[4] The other variables in equation (2.2) denote, respectively: Y_t the euro area GDP growth rate, $x_{t,s}^{(w)}$ the monthly data from surveys, available at the end of each month, and $x_{t,h}^{(w)}$ denotes the growth rate of the index of industrial production, available about 45 days after the end of the reference month. Because of the frequency mismatch within the whole dataset, the thirteen models include a different number of predictors, as we have explained above. As regards the survey, $x_{t,s}^{(w)}$, and the industrial production, $x_{t,h}^{(w)}$, we impose the following specific structure which mimics the data release explained above, and that will be used throughout our exercise. The variable $x_{t,s}^{(w)}$ is not present in models 1 to 4 because the survey is not available in the first four weeks of the quarter, so that $\beta_{t,s}^{(1)} = \beta_{t,s}^{(2)} = \beta_{t,s}^{(3)} = \beta_{t,s}^{(4)} = 0$. Then, for models 5 to 8, $x_{t,s}^{(w)}$ is the value of the survey for the first month of the quarter: $x_{t,s}^{(w)} = S_{t,1}$

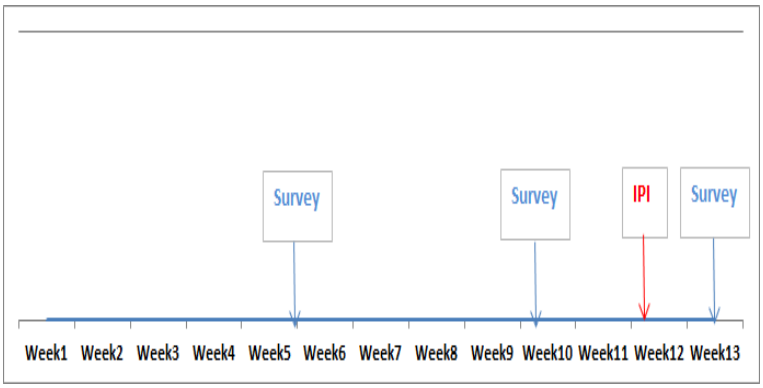

Figure 1: Timeline of data release in the pseudo real-time exercise within the quarter.

[4]In our empirical analysis, we also test models that do not use the average over weeks of Google search data as explanatory variables, but instead, Google search data for each new weeks is considered as the variable for the quarter. Results clearly point that models integrating averaged Google search data give smaller Mean Squared Forecasting Errors than models that do not use the averaged Google search data.

where $S_{t,i}$ denotes the variable S_t referring to the i-th month of quarter t. In models 9 to 12, $x_{t,s}^{(w)}$ will be equal to the average of the survey data available at the end of the first and second month of the quarter: $x_{t,s}^{(w)} = (S_{t,1} + S_{t,2})/2$. Last, in model 13, $x_{t,s}^{(w)}$ is the average of the survey data over the quarter: $x_{t,s}^{(w)} = (S_{t,1} + S_{t,2} + S_{t,3})/3$. Similarly, the variable $x_{t,h}^{(w)}$ is not present in models 1 to 10 (so that $\beta_{t,h}^{(1)} = \ldots = \beta_{t,h}^{(10)} = 0$) and in models 11 to 13, $x_{t,h}^{(w)}$ will be the value of the growth rate of the index of industrial production IP_t for the first month of the quarter.

The idea of having thirteen models is that a researcher will use one of these thirteen models to nowcast the current-quarter values of Y_t depending on the current week of the quarter. For instance, to nowcast the current-quarter value of Y_t at the end of week 2, the Model $w = 2$ will be used. In Table 1 in the Annex, we give the thirteen models based on equation (2.2) and we denote them by $M1, \ldots, M13$.

One of the main issue in the literature on big data is to know whether and when such alternative data are able to bring an additional gain with respect to standard types of variables, like hard and soft data. To contribute to the existing literature on this issue, we have also estimated nowcasting models without including the vector of variables selected from the Google search data. That is, these models only include as predictors the survey and the growth rate of the index of industrial production (i.e. $\beta_{t,g}^{(w)} = 0$ in equation (2.2)). We have in total four such models, one for each release of data of these two variables within the quarter, denoted $NoGoogle_1, \ldots, NoGoogle_4$ in Table 2 in the Annex, that will be used for comparison purposes.

An additional issue with the reporting lags concerns the release of GDP figures. In fact, the first GDP assessment is generally released about 45 days after the end of the reference quarter, but sometimes the delay may be longer. For instance, GDP figures for the first quarter of 2014 were only released on the 4^{th} of June 2014. For this reason if one wants to nowcast in real-time GDP growth for 2014q2 it is not possible to use the fitting computed with the data available up to 2014q1 because one does not observe the GDP for 2014q1. Instead, one has to use the estimated parameters computed with the data available up to 2013q4. Because of this, we impose a gap of two quarters between the sample used for fitting the model (training sample) and the sample used for the out-of-sample analysis (test data). For coherence, we use this structure in both the pseudo-real-time and the true real-time analysis.

Another issue concerns the inclusion of lagged GDP among the explanatory variables. Because of the delay in the release of the GDP we cannot include the lagged GDP

as explanatory variable in every nowcasting model. In addition to this, the GDP is not released at a fixed date, meaning that the release is different at every period (every quarter and every year). For these reasons we have not included the lagged GDP among the explanatory variables in the thirteen models (2.2) for the pseudo-real-time analysis. On the other hand, for the true real-time analysis we have exploited the additional information arising from lagged GDP and have included it among the explanatory variables when it is available. We provide in Figure 3 in the Annex an overview of the dates at which specific GDP figures are released as well as the indication of the time from which we can include the lagged GDP among the explanatory variables and the arrival times of new vintages. We have used this calendar to construct our real-time analysis.

In fact we carry out two types of real-time analysis: (I) a true real-time analysis which includes the lagged GDP growth when it is available, and (II) a true real-time analysis which does not include the lagged GDP growth. The latter is meant for comparison with the pseudo-real-time analysis which does neither include lagged GDP values.

3.3 Selection of the tuning parameters λ and α

To construct our estimator of the thirteen models (2.2) two tuning parameters, or hyperparameters, have to be fixed: λ and α. We select them by using a data-driven method based on a grid-search procedure over the last training period. We select a pair of values for (λ, α) for each model and for each nowcasting period. Hence, in total we have $13 * 9 = 117$ values for the pair (λ, α).

The empirical choice of the hyperparameter α is a crucial issue because it has an important impact on the nowcasting accuracy. Ideally, we would like to choose a value for α for which the MSFE is as small as possible. Therefore, we follow Li [1986, 1987] and use the Generalized cross-validation (GCV) technique to select α. This technique has recently been used by Carrasco and Rossi [2016] in a forecasting setting. The idea is to select the value of α that minimizes the following quantity:

$$\widehat{Q}_T^{(w)}(\alpha) = \frac{T^{-1}\sum_{t=1}^{T}(Y_t - X_t'\widehat{\beta}_{Ridge}^{(w)})^2}{\left(1 - T^{-1}tr(\widehat{R}_T(\alpha))\right)^2}$$

where $tr(\cdot)$ denotes the trace operator and $\widehat{R}_T(\alpha)$ is given by

$$\widehat{R}_T(\alpha) = X_{M_\lambda}\left(T^{-1}\sum_{t=1}^{T} X_{t,M_\lambda}X_{t,M_\lambda}' + \alpha I\right)^{-1} T^{-1}\sum_{t=1}^{T} X_{t,M_\lambda}'$$

and $X_{M_\lambda} = [X_{1,M_\lambda}, \ldots, X_{T,M_\lambda}]$. In our analysis we minimize $\widehat{Q}_T^{(w)}(\alpha)$ over a grid of 31 equispaced values in $[0.09, 1.1]$ denoted by $\mathcal{A}$, so that $\widehat{\alpha}_T^{(w)} = \arg\min_{\alpha \in \mathcal{A}} \widehat{Q}_T^{(w)}(\alpha)$.

For λ we consider a grid of 99 equispaced values in $(0, 1]$, denoted by Λ. The selection is made sequentially: for each value of λ in the grid we select for α in model w the value $\widehat{\alpha}_T^{(w)}$ that solves $\widehat{\alpha}_T^{(w)} := \arg\min_{\alpha \in \mathcal{A}} \widehat{Q}_T^{(w)}(\alpha)$. This is done for each of the thirteen models and for each nowcasting period by using only the training sample corresponding to the specific nowcasting period we are considering. We notice that T depends on the nowcasting period.

Once a value $\widehat{\alpha}_T^{(w)}$ is selected for each value of λ in the grid, we select the value of λ that minimizes the MSFE for the GDP growth of the last quarter of the training sample obtained by using the selected $\widehat{\alpha}_T^{(w)}$. That is, if T denotes the last quarter of the training sample, then we select for λ in model w the value $\widehat{\lambda}_T^{(w)} = \arg\min_{\lambda \in \Lambda}(Y_T - X'_{T,M_\lambda} \widehat{\beta}^{(w)}_{Ridge,M_\lambda}(\widehat{\alpha}_T^{(w)}))^2$.

4 Empirical Results

In this section we present the results of our empirical exercises aiming at nowcasting the euro area GDP growth using various types of data sources. This section is split into three parts. First, we look at the accuracy gains stemming from using Google data when controlling for standard official macroeconomic data, by comparing nowcasts obtained with and without such data, in a pseudo real-time exercise. Then we look at the effects of pre-selecting Google data before estimating Ridge regressions. Third, we perform a true real-time analysis.

4.1 Is there a gain from using Google data, and when ?

In this subsection we compare the evolution over the quarter of weekly Root MSFEs (RMSFE) stemming from the nowcasting models, with and without Google search data. We do this exercise in pseudo-real time, that is, by using historical data but by accounting for their ragged-edge nature. To evaluate the impact of Google search data on current-quarter nowcasts of the GDP growth, we make two types of comparisons. First, we estimate the thirteen nowcasting models by using only Google data, that is, $x_{t,s}^{(w)} = x_{t,h}^{(w)} = 0$ for every $w = 1, \ldots, 13$ in Equation (2.2). Second, to assess the marginal gain of integrating Google data, we compare the four models that only ac-

count for hard and soft data (i.e. without Google data) with the corresponding models given by (2.2) accounting for the full set of information (Google, Survey and Industrial Production). More precisely, we directly compare the following pairs of models: $(NoGoogle_1, M5)$, $(NoGoogle_2, M9)$, $(NoGoogle_3, M11)$, and $(NoGoogle_4, M13)$. The results of these comparisons are reported in Figure 2 below and in Panel 1 of Figure 8 in the Annex. The estimation has been conducted by using Ridge regularization coupled with the SIS pre-selection approach as described in Sections 2.2 and 2.3.

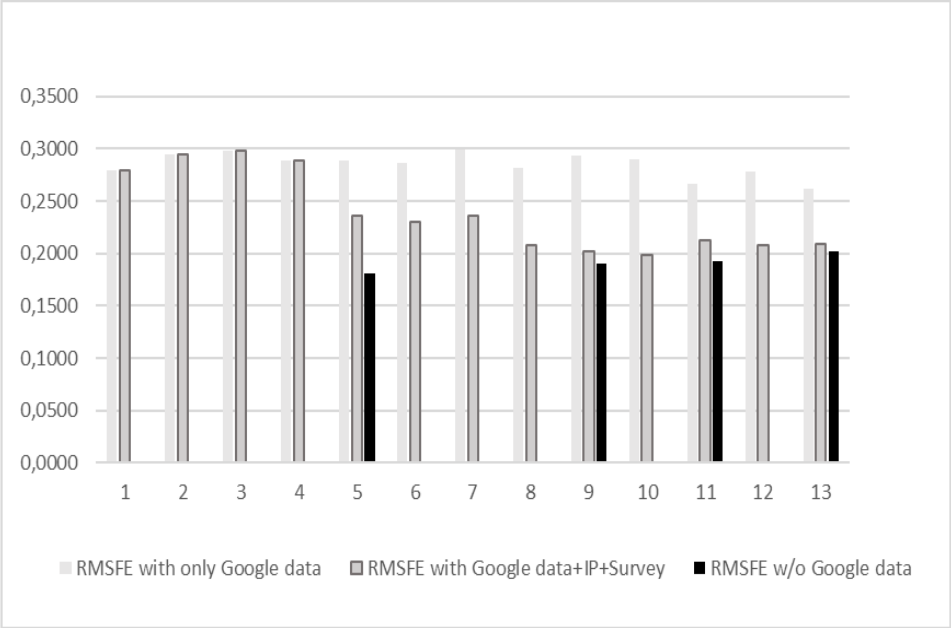

Figure 2: The importance of Google data. Pseudo-real-time analysis with pre-selection of Google data. RMSFEs from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t and Google data) (in gray), (iii) models with only official variables $NoGoogle_1$ - $NoGoogle_4$ (in black).

The first striking feature that we observe in Figure 2 is the downward sloping evolution of RMSFEs stemming from the models with full information (Google, Industrial Production and Survey) over the quarter. This is in line with what could be expected from nowcasting exercises when integrating more and more information throughout the quarter (see Angelini et al. [2011]). When using Google information only (light gray bars), we still observe a decline but to a much lower extent and the RMSFEs stay above 0.25 even at the end of the quarter. However, when focusing on the beginning of the quarter, models that only integrate Google information provide reasonable RMSFEs that do not exceed 0.30 (see Figure 2). This result shows that Google search data possess a informational content that can be valuable for nowcasting GDP growth for the first four weeks

of the quarter, when there is no other available official information about the current state of the economy. However, it is striking to see that when information about the first survey of the quarter arrives, that is in week 5, the model that only incorporates Google data clearly suddenly underperfoms. Looking at Table 8, we see that the RMSFE goes from 0.2887 in week 4 when only Google information is used to 0.2361 in week 5 when the full information model is used. In addition, we note that a simple model, only accounting for official hard and soft information, leads to a much lower RMSFE in week 5 (equal to 0.1807, see Panel 1 of Figure 8 in the Annex). Comparing black bars and gray bars in Figure 2 clearly shows evidence that there is no additional gain of adding Google data to the model starting from week 5; a simple model with only hard and soft information cannot indeed be outperformed.

4.2 Is it worth to pre-select Google data?

As mentioned in Section 2.2, the literature suggests that it could be useful to first pre-select a sub-sample of Google data before estimating the thirteen models (2.2). In this respect, various approaches have been put forward in order to target *ex ante* the variable of interest (see *e.g.* Bai and Ng [2008] or Schumacher [2010], against the background of bridge equations augmented with dynamic factors). In this section we present the performance of a Ridge regularization approach for nowcasting GDP growth when it is coupled with the SIS pre-selection method of Fan and Lv [2008] described in Section 2.2, compared with a standard Ridge regularization approach without any pre-selection.

The idea of the SIS pre-selection method is to identify *ex ante* specific Google variables, among the initial large dataset, that have the highest absolute correlation with the targeted variable, namely the GDP growth rate. First, let us have a look at the relationship between the number of selected variables through the SIS procedure and the absolute correlation between each Google variable and the GDP growth rate at the same quarter. We recall that for the Google variables we take the average over each quarter, see Section 2.2. This relationship is described in Figure 3. We clearly observe an inverse non-linear relationship, with a kind of plateau starting from an absolute correlation of about 0.25. Indeed, most of Google variables present an absolute correlation with current GDP growth rate lower than 0.30. Thus it seems useful to only focus on a core dataset with the highest correlations.

We then analyse the performances in terms of RMSFEs from bridge regressions that use the SIS pre-selection approach associated with Ridge regularization. Figure 4 presents

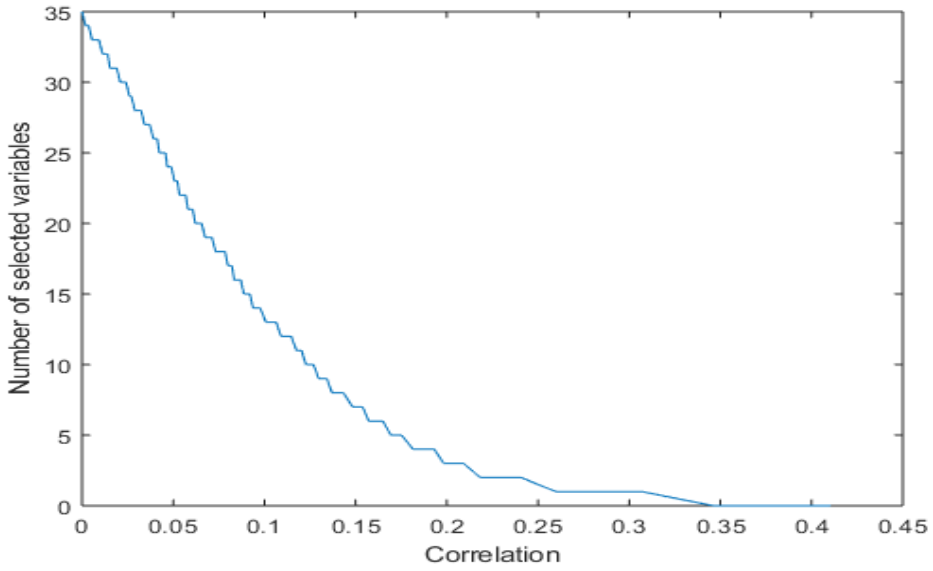

Figure 3: Plot of the number of selected Goggle variables by the SIS method versus the correlation with current GDP (computed for the first training sample).

the evolution over the 13 weeks of the quarter of RMSFEs stemming from bridge models estimated using Google search data and Ridge regression coupled or not with the SIS pre-selection approach. We clearly see that the SIS pre-selection approach (gray bars, similar to the gray bars in Figure 2) allows for an overall improvement in nowcasting accuracy. A striking result is that the RMSFE is lower for all the weeks when the SIS pre-selection approach is used. Moreover, when pre-selection is implemented, RMSFEs evolve over the quarter in a more smoother way. For example, without any pre-selection, we observe that in week 6 the RMSFE jumps to 0.3829, from 0.3239 in week 5. The overall gain underlines the need for pre-selecting data using a targeted approach. Panel 2 in Figure 8 in the Annex reports the exact values of the RMSFEs with and without pre-selection.

4.3 A true real-time analysis

In this subsection, we carry a true real-time analysis by using vintages of data for GDP and industrial production[5] and by accounting for the observed timeline of data release as provided by Eurostat. As regards the dates of the GDP release, there is a large heterogeneity from one period to the other. When available, we also include the lagged GDP growth among the explanatory variables of the nowcasting models. Figure

[5]Survey data are generally not revised.

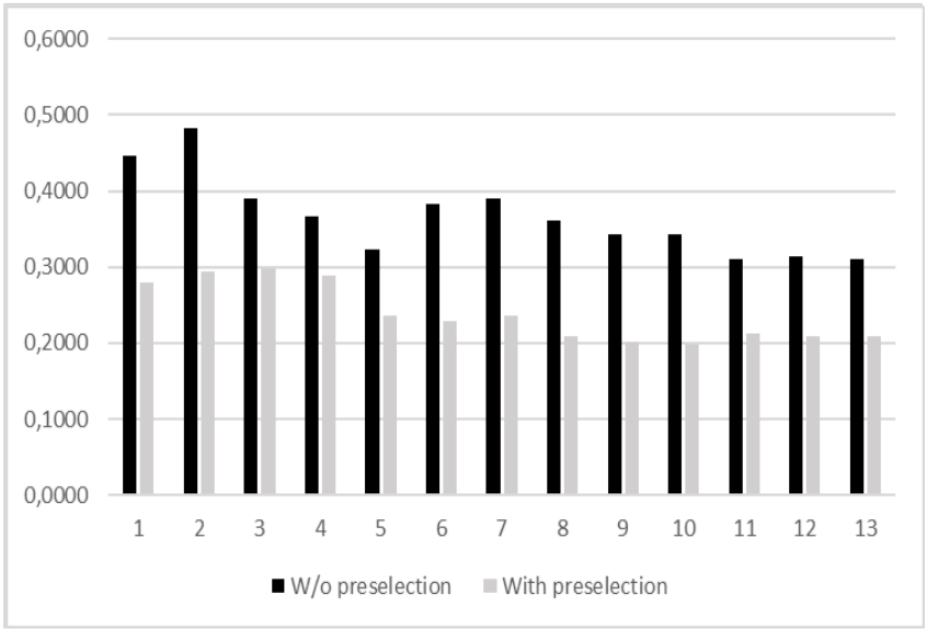

Figure 4: Pseudo-real-time: is it worth to preselect? Evolution over the 13 weeks of the quarter of the RMSFEs stemming from bridge models using Google data, S_t and IP_t estimated from Ridge regularization with and without SIS pre-selection approaches.

3 in the Annex gives the exact weeks in the out-of-sample period 2014q1-2016q1 where the lagged GDP growth is included in the real-time analysis.

In Figure 5 we show that pre-selecting Google data is still worth in real-time. Indeed, RMSFEs obtained from models integrating pre-selected Google data are systematically lower, for all weeks, than those obtained without any pre-selection. The corresponding RMSFE values are reported in Panel 3 of Table 8.

In Figure 6, we show the impact of Google search data on GDP growth nowcasting accuracy in the context of a true real-time nowcasting analysis. The corresponding RMSFE values are reported in Panel 4 of Table 8. Similarly to the pseudo real-time exercise, we get that during the first 4 weeks of the quarters, when only Google information is available, RMSFEs are quite reasonable. This fact is reassuring about the real-time use of Google search data when nowcasting GDP. However, starting from week 5, as soon as the first survey of quarter is released, the marginal gain of using Google data instantaneously vanishes. [6]

[6]There is an exception in week 11, where it is surprising to note that the integration of surveys, past GDP value and industrial production tend to suddenly increase the RMSFEs, in opposition to what can be expected from previous empirical results. This stylized has to be further explored.

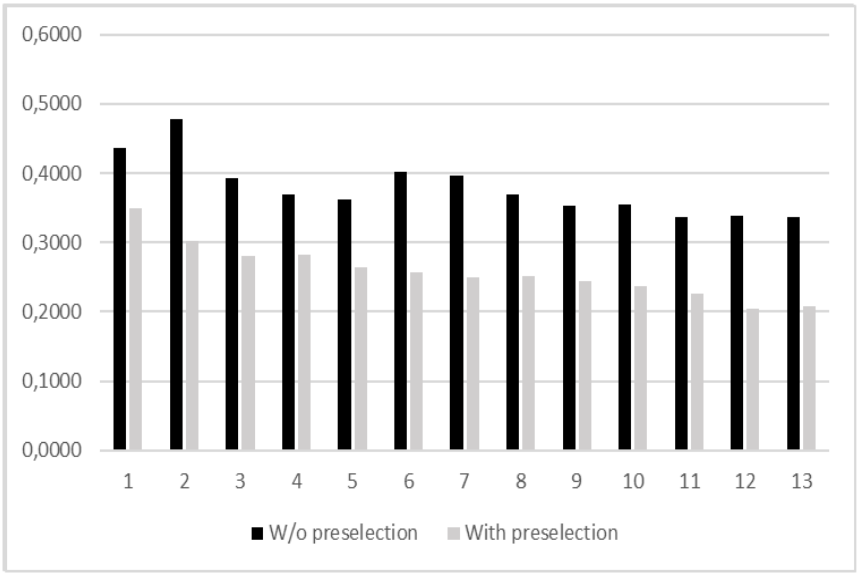

Figure 5: True real-time analysis: is it worth to preselect? Evolution over the 13 weeks of the quarter of the RMSFEs stemming from bridge models using Google data, S_t, IP_t, and lagged GDP growth estimated from Ridge regularization with and without SIS pre-selection approaches. The models include the lagged GDP growth when it is available.

Finally, in order to compare the results of the real-time analysis with the ones from the pseudo-real-time analysis, we compute GDP growth nowcasts without including the lagged GDP growth among the explanatory variables. The results are given in Figure 7. The corresponding RMSFE values are reported in Panel 5 of Table 8. We see that both analyses lead to a similar shape in the evolution of RMSFEs within the quarter, although, as expected, the uncertainty around weekly nowcasts is a bit higher in real time.

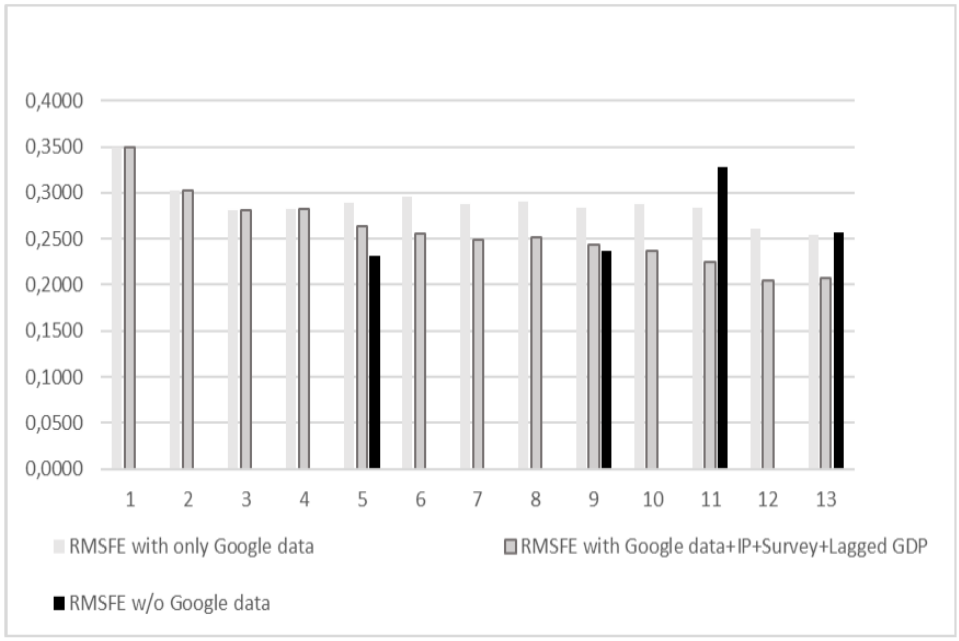

Figure 6: The importance of Google data. True Real-time analysis with pre-selection of Google data. RMSFE from: (i) models M1 - M13 with only variables extracted from Google data (in light gray), (ii) models M1 - M13 with all the variables (S_t, IP_t, $laggedGDP$ and Google data) (in gray), (iii) models $NoGoogle_1$ - $NoGoogle_4$ (in black).

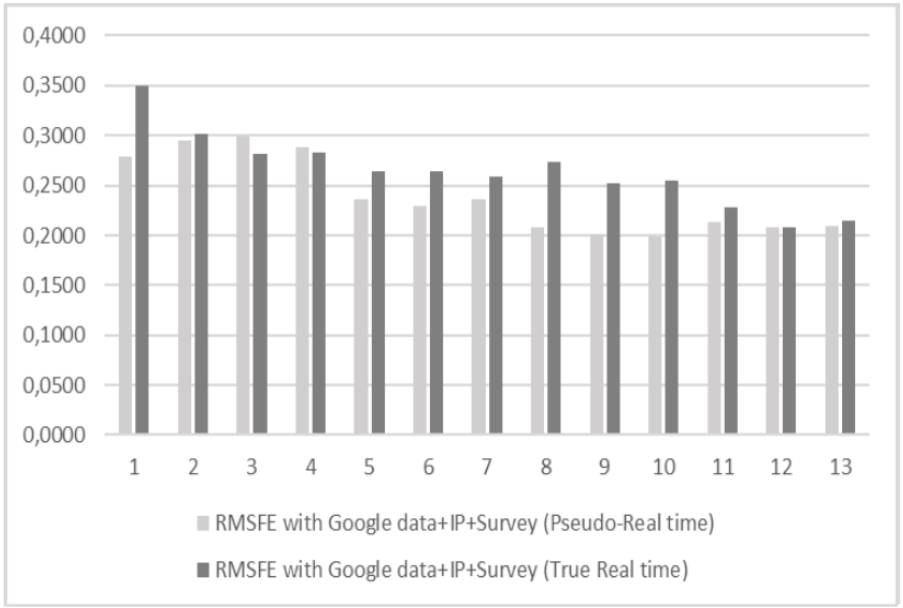

Figure 7: Pseudo-Real-time versus True Real-time analysis (with pre-selection). Comparison of RMSFEs within the quarter from pseudo-real-time (in light gray) and true real-time (in gray) analysis. The true real-time analysis does not include lagged GDP growth among the explanatory variables.

5 Conclusions

In this paper we consider the use of Google search data to nowcast the euro area GDP growth rate. Our main objective is to evaluate the usefulness of Google search data for nowcasting when official data are not available, against the background of a pseudo real-time analysis. Because Google search data are high dimensional, in the sense that the number of variable is large compared to the time series dimension, there is a price to pay for using them: first, we need to reduce their dimensionality from ultra-high to high by using a screening procedure, and second we need to use a regularized estimator to deal with the pre-selected variables. Our second objective is to perform a true real-time analysis and assess the validity of our results in this framework.

Four salient facts emerge from our empirical analysis. First, against the background of a pseudo real-time analysis, we point out the usefulness of Google search data in nowcasting euro area GDP growth rate for the first four weeks of the quarter when there is no information about the state of the economy. We show that at the beginning of the quarter, Google data provide an accurate picture of the GDP growth rate.

Second, as soon as official data become available, that is starting from week 5 with the release of opinion surveys, then the relative nowcasting power of Google data instantaneously vanishes.

Third, we show that pre-selecting Google data before entering the nowcasting models appears to be a pertinent strategy in terms of nowcasting accuracy. Especially, we implement the Sure Independent Screening approach put forward by Fan and Lv [2008] enabling to retain only the Google variables that are the most correlated with the targeted variable, that is GDP growth rate. This result confirms previous results obtained with bridge equations augmented with dynamic factor (see e.g. Bai and Ng [2008] or Schumacher [2010]).

Finally, we show when using Google search data in the context of a true real-time analysis, the three previous salient facts remain valid. This result argues in favor of the use of Google search data at the beginning of the quarter, when there is no official information available, for real-time policy-making.

References

K. Aastveit and T. Trovik. Nowcasting norwegian GDP: the role of asset prices in a small open economy. *Empirical Economics*, 42(1):95–119, 2012.

E. Angelini, G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Ruenstler. Short-term forecasts of euro area gdp growth. *Economic Journal*, 14:C25–C44, 2011.

J. Bai and S. Ng. Forecasting economic time series using targeted predictors. *Journal of Econometrics*, 146(2):304 – 317, 2008. Honoring the research contributions of Charles R. Nelson.

K. Barhoumi, O. Darne, and L. Ferrara. Are disaggregate data useful for forecasting french gdp with dynamic factor models ? *Journal of Forecasting*, 29(1-2):132–144, 2010.

J. Boivin and S. Ng. Are more data always better for factor analysis? *Journal of Econometrics*, 132:169–194, 2006.

M.-E. Bontempi, M. Frigeri, R. Golinelli, and M. Squadrini. Uncertainty, perception and internet. Technical report, 2018.

D. Bragoli. Nowcasting the japanese economy. *International Journal of Forecasting*, 33 (2):390–402, April 2017.

D. Bragoli, L. Metelli, and M. Modugno. The importance of updating: Evidence from a Brazilian nowcasting model. *OECD Journal: Journal of Business Cycle Measurement and Analysis*, 2015(1):5–22, 2015.

D. Buono, G. Kapetanios, M. Marcellino, G. L. Mazzi, and F. Papailias. Big data econometrics: Nowcasting and early estimates. Technical Report 82, Working Paper Series, Universita Bocconi, 2018.

M. Carrasco and B. Rossi. In-sample inference and forecasting in misspecified factor models. *Journal of Business & Economic Statistics*, 34(3):313–338, 2016.

Y. Carriere-Swallow and F. Labbe. Nowcasting with google trends in an emerging market. *Journal of Forecasting*, 32(4):289–298, 2013.

H. Choi and H. Varian. Predicting initial claims for unemployment insurance using google trends. *Google Technical Report*, 2009.

H. Choi and H. Varian. Predicting the present with google trends. *Google Technical Report*, 2012.

D. Coble and P. Pincheira. Nowcasting building permits with google trends. MPRA Paper 76514, University Library of Munich, Germany, 2017.

F. D'Amuri and J. Marcucci. The predictive power of google searches in forecasting unemployment. Temi di discussione (Economic working papers) 891, Bank of Italy, Economic Research and International Relations Area, 2012.

M. Diron. Short-term forecasts of euro area real gdp growth: An assesment of real-time performance based on vintage data. *Journal of Forecasting*, 27:371–390, 2008.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society B*, 70:849–911, 2008.

L. Ferrara and C. Marsilli. Nowcasting global economic growth: A factor-augmented mixed-frequency approach. *The World Economy*, 2018.

C. Frale, M. Marcellino, G. L. Mazzi, and T. Proietti. Euromind: a monthly indicator of euro area economic conditions. *Journal of the Royal Statistical Society, Series A*, 174(2):439–470, 2010.

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

D. Giannone, M. Lenza, and G. Primiceri. Economic predictions with big data: The illusion of sparsity. *mimeo*, 2017.

T. Goetz and T. Knetsch. Google data in bridge equation models for german gdp. *International Journal of Forecasting*, 35(1):45–66, 2019.

R. Golinelli and G. Parigi. Tracking world trade and GDP in real time. *International Journal of Forecasting*, 30(4):847–862, 2014.

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

K.-C. Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

X. Li. Nowcasting with big data : Is google useful in the presence of other information? *mimeo*, 2016.

M. Modugno, B. Soybilgen, and E. Yazgan. Nowcasting Turkish GDP and news decomposition. *International Journal of Forecasting*, 32(4):1369–1384, 2016.

F. Narita and R. Yin. In search for information: Use of google trends' data to narrow information gaps for low-income developing countries. Technical Report WP/18/286, IMF Working Paper, 2018.

P. Nymand-Andersen and E. Pantelidis. Google econometrics: Nowcasting euro area car sales and big data quality requirements. Technical report, European Central Bank, 2018.

C. Schumacher. Factor forecasting using international targeted predictors: The case of German GDP. *Economics Letters*, 107(2):95–98, May 2010.

S. Scott and H. Varian. Bayesian variable selection for nowcasting economic time series. In A. Goldfarb, S. Greenstein, and C. Tucker, editors, *Economic Analysis of the Digital Economy*, pages 119–135. NBER, 2015.

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

S. Vosen and T. Schmidt. Forecasting private consumption: Survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6):565–578, 2011.

Annex

Model	Equation	Predictors
M1	$Y_t = \beta_{0,1} + \beta'_{g,1} x^{(1)}_{t,g} + \epsilon^{(1)}_t$	$x^{(1)}_{t,g} = x_{t,g,1}$
M2	$Y_t = \beta_{0,2} + \beta'_{g,2} x^{(2)}_{t,g} + \epsilon_t$	$x^{(2)}_{t,g} = \frac{x_{t,g,1}+x_{t,g,2}}{2}$
M3	$Y_t = \beta_{0,3} + \beta'_{g,3} x^{(3)}_{t,g} + \epsilon_t$	$x^{(3)}_{t,g} = \frac{x_{t,g,1}+x_{t,g,2}+x_{t,g,3}}{3}$
M4	$Y_t = \beta_{0,4} + \beta'_{g,4} x^{(4)}_{t,g} + \epsilon_t$	$x^{(4)}_{t,g} = \frac{x_{t,g,1}+...+x_{t,g,4}}{4}$
M5	$Y_t = \beta_{0,5} + \beta_{s,5} x^{(5)}_{t,s} + \beta'_{g,5} x^{(5)}_{t,g} + \epsilon_t$	$x^{(5)}_{t,g} = \frac{x_{t,1}+...+x_{t,5}}{5}$, $x^{(5)}_{t,s} = S_{t,1}$
M6	$Y_t = \beta_{0,6} + \beta_{s,6} x^{(6)}_{t,s} + \beta'_{g,6} x^{(6)}_{t,g} + \epsilon_t$	$x^{(6)}_{t,g} = \frac{x_{t,1}+...+x_{t,6}}{6}$, $x^{(6)}_{t,s} = S_{t,1}$
M7	$Y_t = \beta_{0,7} + \beta_{s,7} x^{(7)}_{t,s} + \beta'_{g,7} x^{(7)}_{t,g} + \epsilon_t$	$x^{(7)}_{t,g} = \frac{x_{t,1}+...+x_{t,7}}{7}$, $x^{(7)}_{t,s} = S_{t,1}$
M8	$Y_t = \beta_{0,8} + \beta_{s,8} x^{(8)}_{t,s} + \beta'_{g,8} x^{(8)}_{t,g} + \epsilon_t$	$x^{(8)}_{t,g} = \frac{x_{t,1}+...+x_{t,8}}{8}$, $x^{(8)}_{t,s} = S_{t,1}$
M9	$Y_t = \beta_{0,9} + \beta_{s,9} x^{(9)}_{t,s} + \beta'_{g,9} x^{(9)}_{t,g} + \epsilon_t$	$x^{(9)}_{t,g} = \frac{x_{t,1}+...+x_{t,9}}{9}$, $x^{(9)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$
M10	$Y_t = \beta_{0,10} + \beta_{s,10} x^{(10)}_{t,s} + \beta'_{g,10} x^{(10)}_{t,g} + \epsilon_t$	$x^{(10)}_{t,g} = \frac{x_{t,1}+...+x_{t,10}}{10}$, $x^{(10)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$
M11	$Y_t = \beta_{0,11} + \beta_{s,11} x^{(11)}_{t,s} + \beta_{h,11} x^{(11)}_{t,h} + \beta'_{g,11} x^{(11)}_{t,g} + \epsilon_t$	$x^{(11)}_{t,g} = \frac{x_{t,1}+...+x_{t,11}}{11}$, $x^{(11)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$, $x^{(11)}_{t,h} = IP_{t,1}$
M12	$Y_t = \beta_{0,12} + \beta_{s,12} x^{(12)}_{t,s} + \beta_{h,12} x^{(12)}_{t,h} + \beta'_{g,12} x^{(12)}_{t,g} + \epsilon_t$	$x^{(12)}_{t,g} = \frac{x_{t,1}+...+x_{t,12}}{12}$, $x^{(12)}_{t,s} = \frac{S_{t,1}+S_{t,2}}{2}$, $x^{(12)}_{t,h} = IP_{t,1}$
M13	$Y_t = \beta_{0,13} + \beta_{s,13} x^{(13)}_{t,s} + \beta_{h,13} x^{(13)}_{t,h} + \beta'_{g,13} x^{(13)}_{t,g} + \epsilon_t$	$x^{(13)}_{t,g} = \frac{x_{t,1}+...+x_{t,13}}{13}$, $x^{(13)}_{t,s} = \frac{S_{t,1}+...+S_{t,3}}{3}$, $x^{(13)}_{t,h} = IP_{t,1}$

Table 1: Equations of the 13 models ($M1, \ldots, M13$) used to nowcast GDP growth over each quarter. Equations include the variables pre-selected from Google data as well as information stemming from surveys (S_t) and industrial production (IP_t). $S_{t,i}$ denotes the variable surveys S_t referring to the i-th month of the current-quarter t and $IP_{t,i}$ denotes the growth rate of the industrial production available at the 11^{th} week of the current-quarter t and referring to the i-th month of the current-quarter t.

Model	Equation	Predictors
$NoGoogle_1$	$Y_t = \beta_{0,1} + \beta_{s,1} x_{t,s}^{(1)} + \epsilon_t$	$x_{t,s}^{(1)} = S_{t,1}$
$NoGoogle_2$	$Y_t = \beta_{0,2} + \beta_{s,2} x_{t,s}^{(2)} + \epsilon_t$	$x_{t,s}^{(2)} = \frac{S_{t,1}+S_{t,2}}{2}$
$NoGoogle_3$	$Y_t = \beta_{0,3} + \beta_{s,3} x_{t,s}^{(3)} + \beta_{h,3} x_{t,h}^{(3)} + \epsilon_t$	$x_{t,s}^{(3)} = \frac{S_{t,1}+S_{t,2}}{2}$, $x_{t,h}^{(3)} = IP_{t,1}$
$NoGoogle_4$	$Y_t = \beta_{0,4} + \beta_{s,4} x_{t,s}^{(4)} + \beta_{h,4} x_{t,h}^{(3)} + \epsilon_t$	$x_{t,s}^{(4)} = \frac{S_{t,1}+...+S_{t,3}}{3}$, $x_{t,h}^{(4)} = IP_{t,1}$

Table 2: Equations of the four models used to nowcast GDP growth without the variables extracted from Google data. $S_{t,i}$ denotes the variable surveys S_t referring to the i-th month of the current-quarter t and $IP_{t,i}$ denotes the growth rate of the industrial production available at the 11^{th} week of the current-quarter t and referring to the i-th month of the current-quarter t.

2 lags between estimation period and forecasting period

Last Training Period	Nowcasting Period	1st GDP Vintage which contains the last GDP in the Training sample	Lagged GDP	week of new Vintage
2013Q3	2014Q1	08/04/2014	no	
2013Q4	2014Q2	08/04/2014	no	
		08/04/2014	no	
		15/04/2014	no	3rd week
		04/06/2014	yes	10th week
2014Q1	2014Q3	02/07/2014	no	
2014Q2	2014Q4	01/10/2014	no	
		21/10/2014	no	4th week
		14/11/2014	yes	7th week
		09/12/2014	yes	11th week
2014Q3	2015Q1	09/12/2014	no	
		17/03/2015	yes	12th week
2014Q4	2015Q2	17/03/2015	no	
		02/06/2015	yes	10th week
2015Q1	2015Q3	02/06/2015	no	
		30/07/2015	no	4th week
		09/09/2015	yes	11th week
		24/09/2015	yes	13th week
2015Q2	2015Q4	24/09/2015	no	
		13/11/2015	yes	7th week
		08/12/2015	yes	11th week
2015Q3	2016Q1	08/12/2015	no	
		12/02/2016	yes	6th week
		16/02/2016	yes	7th week
		08/03/2016	yes	10th week

Table 3: Timeline of GDP release in real-time within the quarter. The first column gives the last period used for the in-sample analysis (training sample), the second column indicates the nowcasting period, the third column indicates the date of the first vintage which contains the GDP growth in the last period of the training sample (indicated in the first column), the fourth columns indicates whether a lagged GDP growth is available to be included among the explanatory variables (the corresponding date and week of availability are given in the third and fifth columns, respectively). Finally, the fifth column gives the week, and so the model, corresponding to the date in the third column.

Panel 1 - Pseudo real time: The importance of Google data (with preselection)													
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Google Data + Survey + IP	0,2793	0,2945	0,2985	0,2887	0,2361	0,2296	0,2362	0,2083	0,2019	0,1985	0,2127	0,2082	0,2086
Google Data	0,2793	0,2945	0,2985	0,2887	0,2887	0,2861	0,2993	0,2811	0,2929	0,2894	0,2658	0,2779	0,2612
No Google Data					0,1807				0,1897		0,1928		0,2017

Panel 2 - Pseudo real time: is it worth to preselect?													
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Ridge + SIS Preselection	0,2793	0,2945	0,2985	0,2887	0,2361	0,2296	0,2362	0,2083	0,2019	0,1985	0,2127	0,2082	0,2086
Ridge	0,4467	0,4816	0,3897	0,3659	0,3239	0,3829	0,3901	0,3609	0,3427	0,3422	0,3103	0,3142	0,3111

Panel 3 - True real time: is it worth to preselect?													
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Ridge + SIS Preselection	0,3496	0,3018	0,2812	0,2828	0,2636	0,2561	0,2494	0,2520	0,2438	0,2375	0,2252	0,2045	0,2078
Ridge	0,4357	0,4785	0,3935	0,3700	0,3628	0,4025	0,3970	0,3700	0,3535	0,3540	0,3360	0,3391	0,3372

Panel 4 - True real time: The importance of Google data (with preselection)													
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Google Data + Survey + IP + GDPlag	0,3496	0,3018	0,2812	0,2828	0,2636	0,2561	0,2494	0,2520	0,2438	0,2375	0,2252	0,2045	0,2078
Google Data	0,4357	0,4785	0,3935	0,3700	0,3628	0,4025	0,3970	0,3700	0,3535	0,3540	0,3360	0,3391	0,3372
No Google Data					0,2320				0,2365		0,3283		0,2576

Panel 5 - Pseudo real time vs. True real time (with preselection)													
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13
Pseudo real time	0,2793	0,2945	0,2985	0,2887	0,2361	0,2296	0,2362	0,2083	0,2019	0,1985	0,2127	0,2082	0,2086
True real time (w/o GDPlag)	0,3496	0,3018	0,2812	0,2828	0,2636	0,2640	0,2593	0,2740	0,2525	0,2549	0,2281	0,2082	0,2146

Figure 8: RMSFE from different models: "Google Data + Survey + IP" refers to models M1 - M13 with all the variables: S_t, IP_t and Google data, "Google Data" refers to models M1 - M13 with only variables extracted from Google data, "No Google Data" refers to models $NoGoogle_1$ - $NoGoogle_4$, "Ridge + SIS Preselection" refers to model (2.2) estimated with pre-selected variables from Google data and Ridge regularization, "Ridge" refers to model (2.2) estimated with Ridge regularization without pre-selection, "Google Data + Survey + IP + GDPlag" refers to models M1 - M13 with all the variables: S_t, IP_t, Google data and the lagged GDP growth when it is available, "True real time (w/o GDPlag)" refers to models M1 - M13 estimated without including the lagged GDP growth in true real time.