A Babel of Web-Searches: Googling Unemployment During the Pandemic

Giulio Caperna, Marco Colagrossi*, Andrea Geraci, and Gianluca Mazzarella

European Commission, DG Joint Research Centre

Abstract

Researchers are increasingly exploiting web-searches to study phenomena for which timely and high-frequency data are not readily available. We propose a data-driven procedure which, exploiting machine learning techniques, solves the issue of identifying the list of queries linked to the phenomenon of interest, even in a cross-country setting. Queries are then aggregated in an indicator which can be used for causal inference. We apply this procedure to construct a search-based unemployment index and study the effect of lock-downs during the covid-19 pandemic. In a Difference-in-Differences analysis, we show that the indicator rose significantly and persistently in the aftermath of lock-downs.

Keywords: Unemployment; nowcast; random forest; covid-19; Google Trends; Difference-in-Differences.

JEL: E24; C53; C82.

*Corresponding author: marco.colagrossi@ec.europa.eu.

We thank Claudio Deiana, Massimiliano Ferraresi, Francesco Panella, Paolo Paruolo and audience at seminar series of the Joint Research Centre of the European Commission for valuable comments. Opinions expressed herein are those of the authors only and do not reflect the views of, or involve any responsibility for, the institutions to which they are affiliated. Any errors are the fault of the authors only.

1 Introduction

Starting with the seminal contribution of Choi & Varian (2012), Google search data have been increasingly used in various fields of the economic literature. Web searches proved useful to forecast and nowcast a variety of economic indicators.[1] Further, they have been used in financial studies (e.g., Da et al., 2011; Preis et al., 2013; Vlastakis & Markellos, 2012); to understand tourism flows (Siliverstovs & Wochner, 2018); to gauge the consequences of racial animus on black candidates in the US presidential elections (Stephens-Davidowitz, 2014); to measure the effect of news coverage on the degree of online popularity and radicalisation of the Al-Quaeda terrorist group (Jetter, 2019); and to estimate the impact of the advertised degree of "greenness" on house prices (Zheng et al., 2012).

Google searches are particularly attractive in those contexts in which data about the phenomenon of interest are either not available or available at a low time-frequency. Further, compared to surveys, Google searches are less sensitive to the small-sample bias (Baker & Fradkin, 2017). This two features made web searches an ideal source of data for researcher during the covid-19 pandemic. For example, Brodeur et al. (2020) and Fetzer et al. (2020) use Google Trends data to investigate the impact of lock-downs on, respectively, well-being and economic anxiety. Brunori & Resce (2020) show instead how web queries related to symptoms can be used to monitor the diffusion of the virus.

The use of online searches crucially hinges on their association with the underlying phenomenon of interest. This, in turn, translates into the researchers' ability to identify the most relevant set of queries in a given language and institutional context. This task is particularly challenging in a cross-country setting, where finding an ad-hoc list of keywords is either costly (in terms of time) or not feasible (due to language barriers).

In this paper, we propose a data-driven procedure to retrieve, validate and identify a

[1] In Section 3) we provide an overview of this literature.

set of Google Trends queries which are linked to an underlying economic phenomenon of interest. This set of queries can then be combined to construct an indicator which can, in turn, be used for causal inference. We apply this procedure to estimate the impact of containment measures on unemployment during the covid-19 pandemic in the EU27.

There is already a growing literature investigating the economic impact of the covid-19 pandemic, and unemployment in particular.[2] There are indeed already signs of unprecedented demand for unemployment benefits in the US (Aaronson et al., 2020; Goldsmith-Pinkham & Sojourner, 2020; Kahn et al., 2020) and the number of unemployed people in the OECD area increased by 18 million in April alone.[3] Further, the impact on seasonal activities (e.g., tourism and agriculture) on which several EU countries depends heavily might be particularly severe.[4] Finally, the sudden lock-down of non-essential activities might cast worries on the liquidity of many SMEs, which represent 99.8% of all enterprises in the EU non-financial business sector (NFBS) and employ more than 65% of the workers in non-NFBS activities (Hope et al., 2019).

Since timely and high-frequency administrative data on unemployment are not available in the EU, we use daily web search data from Google Trends.[5] We present a simple conceptual framework linking unemployment-related web searches to current unemployment levels and expectations. We face the challenge of identifying the correct set of keywords for each EU country. Google Trends topics, which are aggregations of differ-

[2]Scholars are investigating the consequences of the evolution of the contagion and mitigation policies on the economy as a whole (e.g., Akira Toda, 2020; Baker et al., 2020; Jones et al., 2020; Ludvigson et al., 2020; Kahn et al., 2020; Stock, 2020), the impact on financial markets and their stability (e.g., Boot et al., 2020; Ramelli & Wagner, 2020) as well as its cost in terms of inequality (e.g., Adams-Prassl et al., 2020; Alon et al., 2020; Coronini-Cronberg et al., 2020) and overall well-being (Hamermesh, 2020).

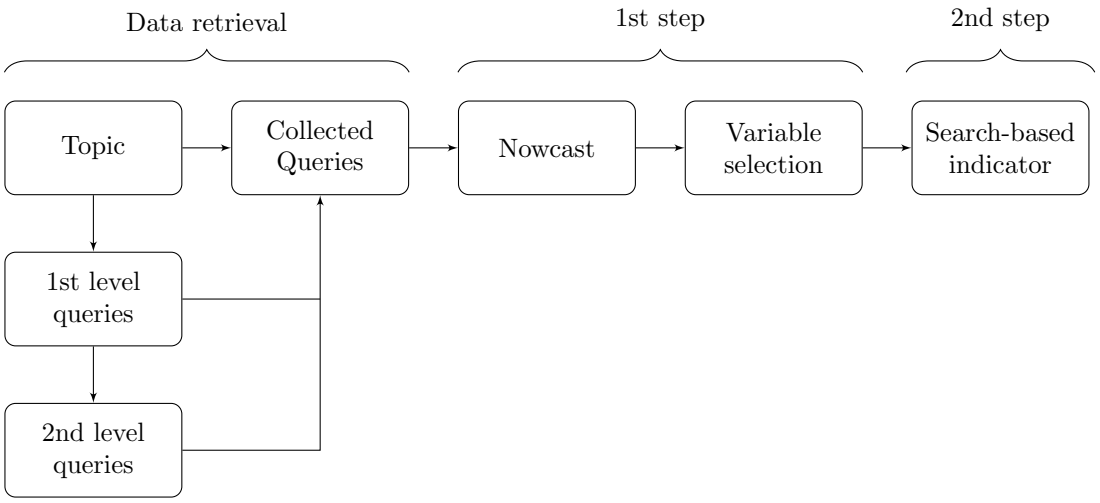
[3]The loss seems to have been particularly severe among youth and women – see Unemployment Rates, OECD - Updated: June 2020, available at:
https://www.oecd.org/newsroom/unemployment-rates-oecd-update-june-2020.htm

[4]See "Tourism and transport in 2020 and beyond", Brussels, 13.5.2020 COM(2020) 550 final, available at https://www.europeansources.info/record/tourism-and-transport-in-2020-and-beyond/

[5]The literature on the labour market impacts of the pandemic and subsequent containment measures has so far focused on single countries, mostly the US, (e.g., Aaronson et al., 2020; Amburgey et al., 2020; Baert et al., 2020; Şahin et al., 2020; Goldsmith-Pinkham & Sojourner, 2020; Kahn et al., 2020) or few selected countries (Adams-Prassl et al., 2020).

ent queries belonging to the same semantic concept, being language-independent, are the ideal candidates for this purpose. However, the algorithm generating topics is Google's proprietary information, thus a black-box to researchers. In this paper, we propose to use the topic *unemployment* to collect, for each country, the entire set of language-specific associated queries in a given time-span (1st level queries) and all the top queries linked to the latter (2nd level queries). Then, as aforementioned, we develop an ad-hoc two-step procedure to construct a search-based unemployment indicator – see Figure 1.

Figure 1: From Google topics to search-based indicators: a two-step procedure



Note: two-step procedure flowchart. Details about data retrieval are outlined in Section 2. The nowcast and variable selection methods (first step) as well as the construction of the indicator (second step) are discussed in Section 3.

In the first step, we nowcast, separately for each country, the monthly unemployment rate time-series using the Search Volume Index (SVI hereafter, see Section 2) of the collected queries. We show that nowcasting unemployment using the topic alone does not provide a statistically significant improvement over what a simple auto-regressive model would predict for the vast majority of the countries considered (Section 3). Instead, once we add all the queries linked to the topic and perform variable selection using random forest-based methods, the predictive accuracy increases significantly in almost all coun-

tries.

In the second step, we select the country-specific queries that best predict the unemployment rate and aggregate them to create a daily indicator of unemployment-related searches. The indicator is built, separately for each country, as the linear projection of the daily SVI of the topic on the daily SVIs of the set of best predictors.

Finally, we use the search-based indicator as the dependent variable in a Difference-in-Differences (DiD) analysis. Following the lock-down measures imposed by some EU governments to limit the spread of the SARS-CoV-2 virus, unemployment-related searches rose by roughly 30% compared to their pre-pandemic average. The higher level of searches persists throughout the lock-down period. Finally, we provide evidence suggesting that announcements of fiscal stimuli by EU Governments are perceived as signals of a worsening economic scenario.

Importantly, the data-driven procedure outlined in this paper is not only relevant in the context of the covid-19 pandemic and unemployment. It could be easily adapted to study a variety of events, policies and economic indicators.

The remainder of this paper is structured as follows: Section 2 briefly introduces Google search data. Section 3 describes our two-step procedure. Section 4 shows the results of the DiD using the indicator of unemployment-related searches. Section 5 concludes.

2 Google searches

Google Trends (https://trends.google.com/trends/) provides access to the search requests made to the Google search engine by its users. In particular, Google Trends contains a random sample representative of all queries that Google handles daily.[6] Search

[6]Google excludes from the sampling queries made by very few people; duplicate searches – i.e., queries made by the same individual over a short period; queries containing special characters; and illegal search activities, such as automated searches performed by bots.

results are normalized to the time and location of a query. By time range (either daily, weekly or monthly) and geography (either country or NUTS-2 level), each data point is divided by the total searches to obtain relative popularity. The resulting numbers are then scaled on a range of 0 to 100 based on a query's proportion to all searches on all queries. Following the literature, we refer to this quantity as the SVI.

Google Trends returns the SVI of either queries or topics. The former are the actual search queries input by users on the Google search engine. Topics are instead aggregations of different queries that could be assigned to a particular semantic domain (in our case, unemployment). Aggregation is done by Google using semantic integration algorithms in the context of the Google knowledge graph.[7]

Topics provide few advantages over simple queries. First, since topics are language-independent, it is possible to use them to perform a cross-country analysis, whereas the same does not apply to keywords. Evidence shows that search terms related to the same topic vary across countries due to cultural and institutional differences (Bousquet et al., 2017). Further, searches linked to topics might vary across time. This is particularly true for searches related to unemployment, which might depend on the name and the seasonality of particular policies in place in any given country. All queries broadly related to a topic are then linked to it independently from the spelling and the wording of the associated queries. In addition, Google Trends also returns the top-25 (when available) queries and topics related to any given topic or query. Top queries and topics are queries (or topics) that are most frequently searched by users within the same session for any given time and geography.

Recently, Google Trends topics have been used by (Brodeur et al., 2020) to estimate the impact of lock-downs on well-being. Fetzer et al. (2020) instead uses topics to measure the degree of economic anxiety during the pandemic. We take a different approach and exploit

[7] Topics were introduced by Google in late 2013 for the US and in the following years for EU countries. See https://developers.google.com/knowledge-graph for additional information.

the topic both for its SVI (as done in the recent literature) and to retrieve associated queries in their native languages.

We collect the monthly SVI for the topic "unemployment" for each country for the period January 2015–December 2019. We then collect, for the same period, the monthly SVI of the "level-1" queries (i.e., the top-25 related search terms associated with the topic) and the monthly SVI for "level-2" queries (i.e., the 10-top related search terms associated with level-1 queries). For the DiD (Section 4) we instead retrieve the daily SVI of both the topic and the subset of queries we identify as the best predictors of unemployment in each country (Section 3) from the 13th of January to the 9th of May 2020.[8]

Of course, Google searches have some limitations. While 90% of EU27 household have internet access, younger individuals are more likely to use the internet than the elderly. Further, access to the internet is not random with respect to socio-economic status.[9] While the former is a lesser concern in our case, as we do not expect the elderly to look for unemployment related-queries given that they are likely to be retired, the latter might impact our results. In particular, if low socio-economic status individuals are excluded from the queries sample, both the nowcast and the event-study analyses could be downward biased.

3 From Google queries to an indicator of unemployment

In the last decade Google search data have been used to forecast and nowcast different macroeconomic indicators. Götz & Knetsch (2019) use Google data to forecast German

[8]We chose the 13th of January as the starting date because (i) it is past the Christmas' holidays period, which might influence online search behaviour but (ii) it is before the events and the lock-down of Wuhan (23rd January) which might have influenced individuals' economic expectations.

[9]See Eurostat, Digital Economy and Society Data https://ec.europa.eu/eurostat/web/digital-economy-and-society/data/database.

GDP, Vosen & Schmidt (2011) and Vosen & Schmidt (2012) focus on forecasting consumption in, respectively, US and Germany. Focusing on financial markets, Da et al. (2015) use Google search data to build an investment sentiment index to predict different US aggregate market indices, while Hamid & Heiden (2015) create a proxy for investors attention to predict stock market volatility. In a recent contribution, Koop & Onorante (2019) show how Google search data can be used to improve nowcast of different macroeconomic variables in the context of dynamic model selection. Finally, focusing on unemployment, D'Amuri & Marcucci (2017) assess the performance of Google search data related to job-search in forecasting US monthly unemployment rate. Fondeur & Karamé (2013) and Smith (2016) perform a similar forecasting exercise focusing, respectively, on France and UK. The consensus in this literature is that the inclusion of Google search data leads to significant improvements in model accuracy, especially for nowcasts and short-term forecasts.

In the first step of the proposed procedure, we follow this literature and perform a nowcast exercise of the monthly unemployment rate time series for each EU27 country from January 2015 to March 2020. Although this exercise is of interest in itself, we use it here to identify the queries that best predict the unemployment rate in each EU27 country.

To understand the relationship between Google searches and unemployment, we start with a simple and stylized conceptual framework. We assume an economy in which, at any given time, the amount of unemployed individuals is given by:

$$\begin{aligned} U_t &= U_{t-1} - O_{t-1,t} + I_{t-1,t} \\ &= U_{t-1} - O_{t-1,t} + \widetilde{\delta}_{t-1,t} E_{t-1}, \end{aligned} \tag{1}$$

where $O_{t-1,t}$ and $I_{t-1,t}$ represent, respectively, the outflows and inflows from and in unemployment. $\widetilde{\delta}_{t-1,t}$ is the true probability of employed individuals E in time $t-1$ to become

unemployed at time t. We then assume the existence of a latent variable ω_t^* representing the volume of online activities related to unemployment at time t:

$$\begin{aligned} \omega_t^* &= \tau U_t + \phi E_t + \eta_t \\ &= \tau U_t + \tau(\widetilde{\delta}_{t,t+1} + \epsilon_t) E_t + \eta_t, \end{aligned} \tag{2}$$

where τ is the volume of online activities performed by the average unemployed individual to retrieve unemployment-related information. We assume that also employed individuals engage in such activities. Their volume ϕ is the same of unemployed individuals, τ, scaled by their (subjective) expectation of becoming unemployed in the next period (δ_t). The relationship between the expectation and the true probability is given by the error model $\delta_t = \widetilde{\delta}_{t,t+1} + \epsilon_t$. Finally, η_t is a residual term capturing online behaviour of those neither in employment nor unemployment.

In this simple representation, the volume of online activities related to unemployment carries information about the level of unemployment at time t – via τU_t – and $t+1$ – via $\tau(\widetilde{\delta}_{t,t+1} + \epsilon_t) E_t$. We proxy ω_t^* with Google searches related to unemployment.

The first challenge is to define the set of Google search queries of interest. D'Amuri & Marcucci (2017) exploit the use of logical operators in the Google Trend platform, and identify the SVI associated to all queries containing the word "jobs". Fondeur & Karamé (2013) use the single term "emploi". Smith (2016) uses a different approach based on the root term "redundancy". The root query is used to obtain the associated queries, and the relative volume data are aggregated using weights to produce a composite "Google Redundancy Index". Borup et al. (2020) show that using a set of queries rather than a single one improves out-of-sample prediction of unemployment growth in the US.

An ad-hoc choice of keywords is not feasible in our context since it would require the identification of the words which semantically define the unemployment concept in each European country. We exploit the Google topic *unemployment* to retrieve, separately

for each country, the top-25 level-1 queries and the top-10 level-2 queries in the original language in the period January 2015 - December 2019. This data-driven approach is similar to the use of a list of root keywords in Da et al. (2015) and Smith (2016) to retrieve the associated queries. Our root, however is not a single keyword or a list of keywords, but the language-independent topic.

After retrieving the full list of associated queries, we extract their SVI in the interval January 2015–March 2020, as well as the SVI of the topic itself.[10] We retrieve monthly Google search data to match the EU unemployment rate time series available from Eurostat (*ei_lmhr_m*).

The number of associated keywords retrieved in each country, after removing duplicates, varies from 3 (Estonia) to 178 (Italy), with a mean of 80 and a median of 85.[11] For each country we estimate different nowcast models which can be summarized as:

$$u_t = f_h(\boldsymbol{K}_t, \boldsymbol{K}_{t-1}, u_{t-h}, u_{t-h-1}) + u_t, \quad h = 1, 2, 3, \tag{3}$$

where u_t is the log-difference of the unemployment rate between month t and month $t-1$, $\boldsymbol{K}_t$ is a P_c-vector comprising the log-differences of the monthly SVI for the P keywords retrieved for country c, including the SVI of the topic ($k1$ hereafter). $\boldsymbol{K}_{t-1}$ is simply the lag of $\boldsymbol{K}_t$. Finally each model includes two lags of the dependent variable: u_{t-h}, and u_{t-h-1}. Since the nowcasting equations embed also lags of the dependent variable, we considered three different horizons (i.e., $h = 1, 2, 3$) corresponding to the last date for which information on unemployment is available.[12]

The models considered differ by the target function f_h, which maps the available information at time t to the dependent variable, as well as the number of keywords

[10] Notice that we only retrieve the keywords associated with the topic until the end of 2019 to avoid covid-19 related keywords. However, we track the SVI of of the selected keywords until March 2020.

[11] For Luxembourg and Malta we were not able to retrieve any associated query.

[12] The maximum value considered ($h = 3$) is the maximum time lag between the release of official Eurostat statistics on unemployment and the availability of contemporaneous data on Google searches.

included in $\boldsymbol{K}_t$, and $\boldsymbol{K}_{t-1}$. More specifically, we consider five different models.

LM.1, our benchmark, is a classical linear AR model which makes no use of Google search data. LM.2 is a linear model where only $k1$ is included in $\boldsymbol{K}_t$ and $\boldsymbol{K}_{t-1}$. RF.1 uses a Random Forest algorithm including the same covariates used in LM.2. RF.2 is a Random Forest where $\boldsymbol{K}_t$ and $\boldsymbol{K}_{t-1}$ include the SVI of all the retrieved keywords for country c plus the SVI of $k1$. RF.3 is a Random Forest model including a subset of the keywords used in RF.2. The subset is identified using the Boruta variable selection method (Kursa et al., 2010; Stoppiglia et al., 2003).[13]

In most of the countries considered, the dimension of the time series is quite small with respect to the number of predictors, a high-dimensional context with $T << P$. As an example, we retrieved 178 keywords in Italy compared with 63 data-points in the unemployment rate monthly time series. Medeiros et al. (2019) explore the performance of different machine learning methods in a forecast race targeted at predicting US inflation using a wide set of covariates. The authors show that, in a data rich environment, the Random Forest algorithm outperforms all the considered alternative high-dimensional models (including the linear LASSO and RIDGE regressions), as well as state-of-the-art dynamic factor models widely used in time series modelling. Therefore, we select Random Forest – an ensemble learning method based on a collection of regression trees introduced by Breiman (2001) – to estimate the target function f_h.

We evaluate the performance of each model using Pseudo-Out-of-Sample prediction (POOS hereafter) based on a rolling window framework with increasing length starting from the first 36 months. The procedure can be summarized as follows: a) the models are trained using the first 36 observations; b) the trained models are used to obtain the prediction for the 37^{th} month; c) the models are then re-trained using the first 37

[13] A brief description of the Random Forest algorithm and of the Boruta variable selection method is provided in the supplementary online material B.1. For a detailed description of Random Forest see Hastie et al. (2009). Results are robust to a different variable selection method – VSURF (Genuer et al., 2010) — and are available upon request.

observations and predictions for the 38^{th} are computed. The entire procedure is iterated separately for each country until month $T-1$.

Having obtained the time series of POOS predictions for each country, we follow the literature and assess the accuracy of each model against our AR benchmark (LM.1) using the standard one-sided Diebold-Mariano (DM) test (Diebold & Mariano, 1995) based on absolute deviations.[14] The aim of this test is to assess whether Google search data carry additional informational content.

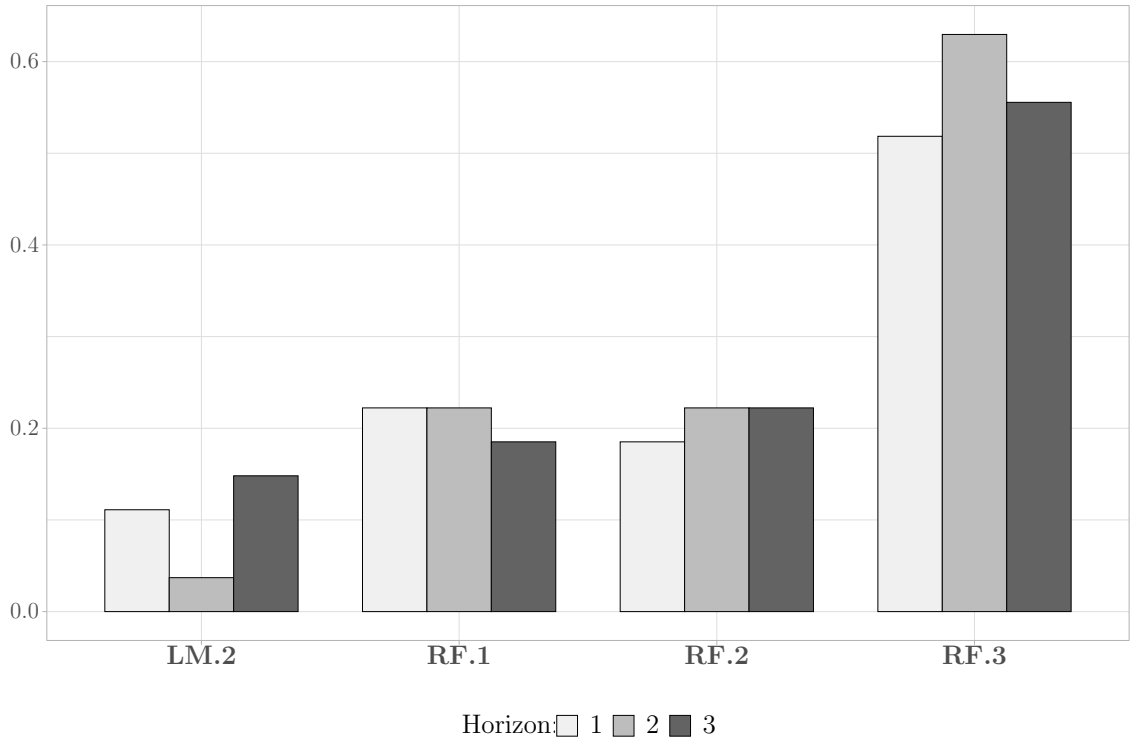
Figure 2 summarizes the main findings. Table A.2 in Appendix A contains the full set of results. Each bar represents the fraction of *DM-victories* of each model against the benchmark LM.1. across the countries considered. A model *wins* over the benchmark if its predictive accuracy is significantly higher ($\alpha = 0.1$).

The results of the comparison indicates that the usage of the SVI of $k1$ alone (LM.2) does not improve the accuracy of the AR model. A slight improvement is visible when $k1$ is used in a Random Forest rather than OLS (RF.1), suggesting that non-linearities are of some importance. Interestingly, the inclusion of the full set of associated keywords in RF.2 is not associated with an additional increase in performance with respect to RF.1. A sizable gain is instead visible when the Boruta variable selection method is used to select the list of relevant predictors to be used in the Random Forest – i.e., RF.3.

The introduction of a selection step in machine learning algorithms has two objectives. On the one hand it is aimed at reducing noise due to highly correlated or redundant predictors. On the other hand, the identification of relevant predictors is useful in itself for interpretation purposes. In our context, the selection step is a way to solve the problem of identifying the most relevant set of country-specific keywords. This is similar in spirit to the procedure adopted by Da et al. (2015) to construct their index of investor sentiment

[14]The choice of absolute deviations instead of the common squared deviations is driven by the scale of our response variable. The log-difference of monthly unemployment rate is close to the zero. Using absolute deviations implicitly assign the same weight to each error avoiding to reward those that are particularly small.

Figure 2: Comparing predictive accuracy of different models against the benchmark AR model with no Google search data



Note: Each bar represents the fraction of countries in which model i has a significantly higher predictive accuracy than the benchmark AR model considered, based on a one-sided Diebold-Mariano test. LM.2 is a linear model where only the SVI of $k1$ is added to the set of predictors. RF.1 is a Random Forest including the same covariates used in LM.2. RF.2 is a Random Forest where the SVI of all the retrieved keywords for each country is included plus the SVI of $k1$. RF.3 is a Random Forest model including a subset of the keywords used in RF.2, chosen using the Boruta algorithm.

starting from the volume of queries related to households economic concerns. Da et al. (2015) use as root a selected set of keywords taken from annotated dictionaries which express negative and positive economic sentiments. Götz & Knetsch (2019) also employ different variable selection methods to identify the set of keywords to be embedded in their GDP forecast models, including principal component analysis, partial least squares, LASSO and boosting.

Overall, the results suggest that a subset of relevant keywords helps to improve nowcast accuracy with respect to the benchmark model. This is not the case for the topic alone.

Combining the use of topics and the variable selection step in our nowcast framework presents two advantages. On the one hand, the use of a common Google topic allows to retrieve a broad set of keywords in a context of heterogeneous countries with different languages and institutions. On the other hand, the variable selection step allows us to identify the subset of keywords which are relevant for the underlying economic variable of interest.

Drawing from these results, in the last step of the proposed procedure, we construct the search-based unemployment indicator, $\widetilde{k1}$, as a weighted linear combination of the SVI of the country-specific subset of best predictors. Weights are obtained by projecting, separately for each country, the linear daily SVI of the topic on the daily SVI of the selected keywords. Intuitively, $\widetilde{k1}$ contains, in our application, the component of the topic explained by the keywords that best predict the unemployment rate.

4 Measuring the effect of lock-down measures on online search activities

We complement $\widetilde{k1}$ daily data with information about Governments' announcements of measures to respond to the covid-19 crisis. In particular, we focus on the lock-down measures enacted by EU governments as recorded by The Assessment Capacities Project (ACAPS). According to this definition, we identify 18 countries which enacted lock-down measures: Austria, Belgium, Bulgaria, Cyprus, Croatia, Denmark, Estonia, France, Germany, Greece, Hungary, Ireland, Italy, Lithuania, Luxembourg, Poland, Portugal and Spain.[15] Data, including the daily SVI of the best predictors, are collected from the 13th

[15]Luxembourg and Portugal are then excluded from our sample due to the unavailability of related keywords. The dates considered are those of the measure's announcement: Austria 16-03; Belgium 18-03; Bulgaria 20-03; Cyprus 24-03; Croatia 18-04; Denmark 18-03; Estonia 30-03; France 17-03; Germany 21-03; Greece 23-03; Hungary: 28-03; Ireland 28-03; Italy 08-03; Lithuania 27-05; Poland 24-03; Portugal 03-04; and Spain 16-03.

of January to the 9th of May.

Our DiD regression can be written as follows:

$$y_{c,t} = \alpha + \sum_{\tau=-5}^{5} \beta_\tau D_{c,w+\tau} + \beta_{\tau^+} D_{c,w+\tau^+} + \mu_c + \delta_t + \varepsilon_{c,t}, \tag{4}$$

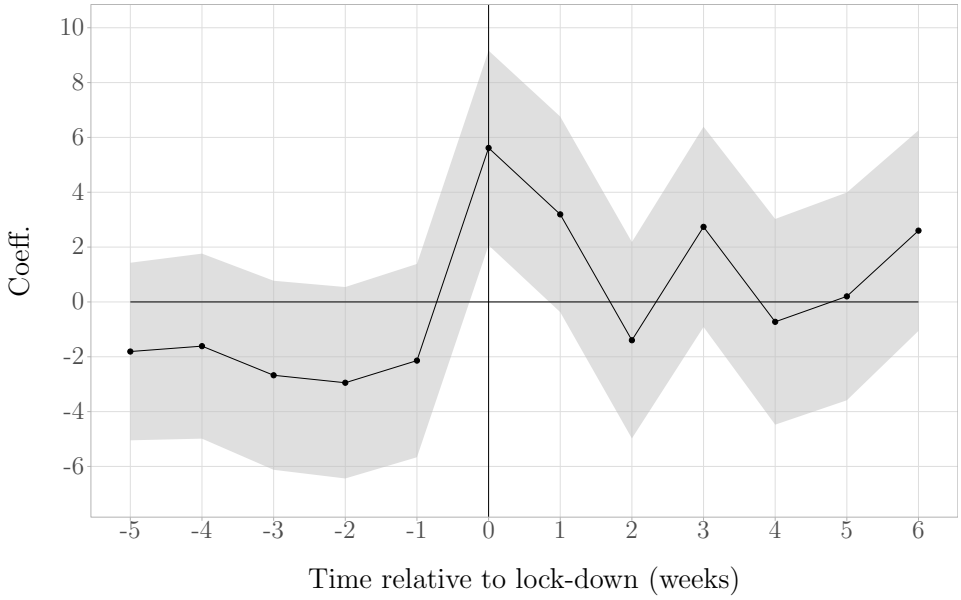
where the generic term $y_{c,t}$ corresponds to either $k1_{c,t}$, the daily SVI of the topic, or $\widetilde{k1_{c,t}}$, the daily indicator of unemployment-related searches in country c at time t; $D_{c,w+\tau}$ are 11 relative week dummies centered around the dates of lock-down. $D_{c,w+\tau^+}$ is a dummy for weeks greater than 5 which is added to avoid the latter being included in the baseline; μ_c are country fixed-effect and δ_t are date fixed-effect. The inclusion of a set of pre-lock-down dummies is used to provide evidence on the validity of the DiD identifying assumption. Estimates are reported in Figure 3.

Figure 3 (a) and (b) shows, respectively, the results for $k1$ and $\widetilde{k1}$. Both measures exhibit an increase of roughly 30% in the week of the announcement of the first introduction of the lock-down. While the effect on $k1$ is short-lived, the opposite is true for $\widetilde{k1}$. The higher level of unemployment-related queries persists throughout the lock-down period. Importantly, this indicates that the keywords that best predict the unemployment rate increased in the aftermath of the lock-down. Further, the coefficient on the pre-lock-down weeks suggest the absence of anticipatory effects, supporting the common trend assumption.

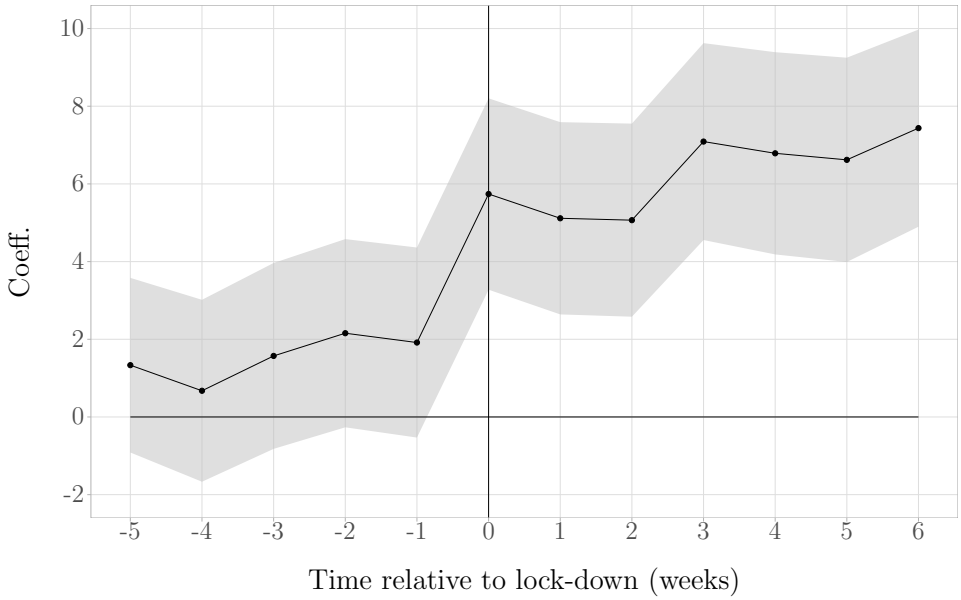
Our findings for the EU27 compare favourably to those by Aaronson et al. (2020) for the US. Aaronson et al. show that unemployment-related queries surged before the record increase in unemployment insurance claims, which peaked before the lock-down measures were implemented. The results suggest that measures introduced by Governments to contain the pandemic generated a negative effect on EU citizens' economic prospects. As it is unlikely that people lost their job immediately after lock-down measures are introduced, our results indicate an increase of unemployment expectations. This is consistent with

Figure 3: DiD coefficients for $k1$ and $\widetilde{k1}$

(a) Topic ($k1$)



(b) Indicator ($\widetilde{k1}$)



the conceptual framework presented in Section 3.

Lock-downs are not the only measures enacted by Governments that might have af-

fected individuals' unemployment expectations. In most EU countries Governments announced, either before or after the pandemic peaked, a variety of economic measures to contrast the worsening economic situation. This might confound the estimated effect of lock-down measures. Since the crisis evolved quite rapidly, these announcements are very close in time. As a consequence the time dynamics of their effects can not be separately identified in a multiple-treatments DiD framework.

To assess the robustness of our findings we first identify, separately for each country, all dates in which $\widetilde{k1}$ exhibits a significant increase. We do so conducting country-specific rolling window event-studies. Starting from the first available date – 13th of January – we consider a time window of 20 days and test whether there has been a statistically significant mean-shift if the last three days of the window. We then roll the time window three days forward and repeat the event-study until the last available date – 9th of May (see Figure 4).

Figure 4: Event-study with rolling windows

Finally, we pool together the results and test whether the significant increases detected are correlated with Governments' announcements. In particular, we focus on two broad sets of measures: fiscal stimuli for the whole economy and support to households either in the form of income support or debt relief. We estimate a linear probability model in which the dependent variable is a dummy which takes value one if a significant increase is detected at time t in country c, and zero otherwise. The set of covariates includes a dummy identifying the week of announcement of the lock-down; a dummy for the week

of announcement of any fiscal stimuli; and one for the week in which income support and debt relief measures are first announced. We also include country and time fixed effect. Results are presented in Table 1. The four columns are relative to different definitions of the time and test windows.

Table 1: Rolling windows event-study

	(1)	(2)	(3)	(4)
Lockdown	0.116**	0.125***	0.165***	0.090**
	(0.046)	(0.042)	(0.048)	(0.043)
Fiscal stimuli	0.046*	0.066***	0.069***	0.076***
	(0.024)	(0.024)	(0.026)	(0.025)
Income support	0.018	0.099**	0.039	-0.002
	(0.044)	(0.043)	(0.045)	(0.043)
Country FE	✓	✓	✓	✓
Day FE	✓	✓	✓	✓
Time window	20	15	20	15
Test window	3	3	7	7
N	2376	2520	2520	2520

Notes: *, **, and *** denote significance of the difference at the 10, 5, and 1 % level. The dependent variable is a dummy which takes value one if a significant increase is detected at time t in country c in country-specific rolling-windows event-studies.

Results confirm the findings of Figure 3: the introduction of lock-down measures increased the volume of unemployment-related searches. Interestingly, a similar effect is shown for the announcement of fiscal stimuli, while no robust effect is found for income support and debt relief measures. These findings suggest that the announcements of fiscal stimuli are perceived as signals of a deteriorating economic scenario, potentially worsening unemployment expectations.

5 Conclusion

Researchers are increasingly exploiting online search activities to study phenomena for which timely and high-frequency data are not readily available. In this paper, we propose

a data-driven procedure which solves the issue of identifying and combining the list of queries linked to the underlying phenomenon of interest. The resulting indicator can then be used for causal inference.

Exploiting Google Trends topics, we retrieve over two-thousand search queries related to unemployment in the EU27 in their native languages. Then, in the first step of the procedure, using machine learning techniques, we select the search queries that best predict unemployment in each EU country. In the second step, we combine these queries and create a search-based unemployment indicator.

Finally, using a DiD approach, we show that, in the aftermath of lock-downs, such indicator rose by about 30% compared to the pre-pandemic average. This effect is persistent over time. In light of a simple conceptual framework, we interpret this finding as an increase in unemployment expectations.

Importantly, the procedure described in this paper is not only relevant in the context of unemployment nor restricted to the case of the covid-19 pandemic. It could be used to study a variety of events, policies and economic indicators, especially when administrative or survey data are not timely available and/or comparable. In particular, the procedure perfectly fits scenarios in which Google Trends data are used in a multi-language and multi-institutional context. Further, while we use the obtained indicator as a dependent variable, it can be also used on the right-hand side of the estimating equation.

References

Aaronson, D., Brave, S. A., Butters, R., Sacks, D. W., & Seo, B. (2020). *Using the Eye of the Storm to Predict the Wave of Covid-19 UI Claims*. Technical Report 2020-10 Federal Reserve Bank of Chicago. URL: https://www.chicagofed.org/~/media/publications/working-papers/2020/wp2020-10-pdf.pdf.

Adams-Prassl, A., Boneva, T., Golin, M., & Rauh, C. (2020). *Inequality in the impact of the coronavirus shock: Evidence from real time surveys*. Technical Report 13183 IZA Institute of Labor. URL: http://ftp.iza.org/dp13183.pdf.

Akira Toda, A. (2020). *Susceptible-Infected-Recovered (SIR) Dynamics of COVID-19 and Economic Impact*. Technical Report arXiv:2003.11221. URL: https://arxiv.org/pdf/2003.11221v2.pdf.

Alon, T. M., Doepke, M., Olmstead-Rumsey, J., & Tertilt, M. (2020). *The impact of COVID-19 on gender equality*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w26947.

Amburgey, A., Birinci, S. et al. (2020). The effects of covid-19 on unemployment insurance claims. *Economic Synopses*, *9*.

Baert, S., Lippens, L., Moens, E., Sterkens, P., & Weytjens, J. (2020). *How do we think the COVID-19 crisis will affect our careers (if any remain)?*. Technical Report 520 Global Labor Organization (GLO). URL: http://hdl.handle.net/10419/215884.

Baker, S. R., Bloom, N., Davis, S. J., & Terry, S. J. (2020). *Covid-induced economic uncertainty*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w26983.

Baker, S. R., & Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from google search data. *Review of Economics and Statistics*, *99*, 756–768.

Boot, A. W., Carletti, E., Kotz, H.-H., Krahnen, J. P., Pelizzon, L., & Subrahmanyam, M. G. (2020). *Corona and Financial Stability 3.0: Try equity-risk sharing for companies, large and small*. Technical Report 81 Leibniz Institute for Financial Research SAFE. URL: http://hdl.handle.net/10419/215544.

Borup, D., Christian, E., & Schütte, M. (2020). In search of a job: Forecasting employment growth using google trends. *Journal of Business & Economic Statistics*, (pp. 1–38).

Bousquet, J., Agache, I., Anto, J. M., Bergmann, K. C., Bachert, C., Annesi-Maesano, I., Bousquet, P. J., D'Amato, G., Demoly, P., De Vries, G. et al. (2017). Google trends terms reporting rhinitis and related topics differ in european countries. *Allergy*, *72*, 1261–1266.

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

Brodeur, A., Clark, A. E., Flèche, S., Powdthavee, N. et al. (2020). *COVID-19, Lockdowns and Well-Being: Evidence from Google Trends*. Technical Report Institute of Labor Economics (IZA). URL: http://ftp.iza.org/dp13204.pdf.

Brunori, P., & Resce, G. (2020). *Searching for the peak Google Trends and the Covid-19 outbreak in Italy*. Technical Report. URL: https://ssrn.com/abstract=3569909.

Choi, H., & Varian, H. (2012). Predicting the present with google trends. *Economic record*, *88*, 2–9.

Coronini-Cronberg, S., John Maile, E., & Majeed, A. (2020). Health inequalities: the hidden cost of covid-19 in nhs hospital trusts? *Journal of the Royal Society of Medicine*, *113*, 179–184.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, *66*, 1461–1499.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all fears investor sentiment and asset prices. *The Review of Financial Studies*, *28*, 1–32.

Degenhardt, F., Seifert, S., & Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, *20*, 492–503.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*, 253–263.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of google searches in forecasting us unemployment. *International Journal of Forecasting*, *33*, 801–816.

Fetzer, T., Hensel, L., Hermle, J., & Roth, C. (2020). *Coronavirus perceptions and economic anxiety*. Technical Report arXiv:2003.03848. URL: https://arxiv.org/pdf/2003.03848.pdf.

Fondeur, Y., & Karamé, F. (2013). Can google data help predict french youth unemployment? *Economic Modelling*, *30*, 117–125.

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, *31*, 2225–2236.

Goldsmith-Pinkham, P., & Sojourner, A. (2020). *Predicting Initial Unemployment Insurance Claims Using Google Trends*. Technical Report Yale School of Management. URL: https://paulgp.github.io/GoogleTrendsUINowcast/google_trends_UI.html.

Götz, T. B., & Knetsch, T. A. (2019). Google data in bridge equation models for german gdp. *International Journal of Forecasting*, *35*, 45–66.

Hamermesh, D. S. (2020). *Lock-downs, loneliness and life satisfaction*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w27018.

Hamid, A., & Heiden, M. (2015). Forecasting volatility with empirical similarity and google trends. *Journal of Economic Behavior & Organization*, *117*, 62–81.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hope, K. et al. (2019). *Annual Report on European SMEs 2018/2019*. Technical Report DG for Internal Market, Industry, Entrepreneurship and SMEs. URL: DOI:10.2826/500457.

Jetter, M. (2019). The inadvertent consequences of al-qaeda news coverage. *European Economic Review*, *119*, 391–410.

Jones, C. J., Philippon, T., & Venkateswaran, V. (2020). *Optimal mitigation policies in a pandemic: Social distancing and working from home*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w26984.

Kahn, L. B., Lange, F., & Wiczer, D. G. (2020). *Labor Demand in the time of COVID-19: Evidence from vacancy postings and UI claims*. Technical Report National Bureau of Economic Research.

Koop, G., & Onorante, L. (2019). Macroeconomic nowcasting using google probabilities. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A (Advances in Econometrics)*, *40*, 17–40.

Kursa, M. B., Rudnicki, W. R. et al. (2010). Feature selection with the boruta package. *Journal of Statistical Software*, *36*, 1–13.

Ludvigson, S. C., Ma, S., & Ng, S. (2020). *Covid19 and the macroeconomic effects of costly disasters*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w26987.

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

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports*, *3*, 1684.

Ramelli, S., & Wagner, A. F. (2020). *Feverish stock price reactions to covid-19*. Technical Report Centre for Economic Policy Research. URL: https://cepr.org/active/publications/discussion_papers/dp.php?dpno=14511.

Şahin, A., Tasci, M., & Yan, J. (2020). *The Unemployment Cost of COVID-19: How High and How Long?*. Technical Report 2020-09 Federal Reserve Bank of Cleveland. URL: https://doi.org/10.26509/frbc-ec-202009.

Siliverstovs, B., & Wochner, D. S. (2018). Google trends and reality: Do the proportions match?: Appraising the informational value of online search behavior: Evidence from swiss tourism regions. *Journal of Economic Behavior & Organization*, *145*, 1–23.

Smith, P. (2016). Google's midas touch: Predicting uk unemployment with internet search data. *Journal of Forecasting*, *35*, 263–284.

Stephens-Davidowitz, S. (2014). The cost of racial animus on a black candidate: Evidence using google search data. *Journal of Public Economics*, *118*, 26–40.

Stock, J. H. (2020). *Data gaps and the policy response to the novel coronavirus*. Technical Report National Bureau of Economic Research. URL: https://www.nber.org/papers/w26902.

Stoppiglia, H., Dreyfus, G., Dubois, R., & Oussar, Y. (2003). Ranking a random feature for variable and feature selection. *Journal of machine learning research*, *3*, 1399–1414.

Vlastakis, N., & Markellos, R. N. (2012). Information demand and stock market volatility. *Journal of Banking & Finance*, *36*, 1808–1821.

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

Vosen, S., & Schmidt, T. (2012). A monthly consumption indicator for germany based on internet search query data. *Applied Economics Letters*, *19*, 683–687.

Zheng, S., Wu, J., Kahn, M. E., & Deng, Y. (2012). The nascent market for "green" real estate in beijing. *European Economic Review*, *56*, 974–984.

Table A.2: Results of the Diebold-Mariano test for equal predictive accuracy comparing different nowcasting models against the benchmark AR model

	LM.2			**RF.1**			**RF.2**			**RF.3**		
Country	h=1	h=2	h=3	h=1	h=2	h=3	h=1	h=2	h=3	h=1	h=2	h=3
AT	0.002	-0.001	0	-0.019***	0.005	0.008	-0.004	-0.002	0.007	-0.016***	0.007	0.014
BE	0.007	0.008	0.009	0.002	0.002	0.002	0	-0.002	-0.002	-0.004*	-0.006**	-0.004*
BG	-0.002	-0.001	0.001	-0.003	-0.006*	-0.004	0.004	-0.002	-0.001	-0.003	-0.01**	-0.01*
CY	-0.005	-0.001	-0.001	-0.025**	-0.016*	-0.024**	-0.009	-0.008	-0.006	-0.03**	-0.021**	-0.033***
CZ	-0.003	-0.002	0.001	0.001	-0.008	0.003	-0.005	-0.018**	-0.011**	-0.011	-0.004	0.006
DE	0.001	-0.002	-0.001	-0.001	-0.003	-0.003	-0.002	-0.004*	-0.003	-0.018***	-0.017***	-0.019***
DK	0	-0.003***	-0.005***	0.003	0	0.001	-0.002	0.003	0.003	0	0	0
EE	0.002	-0.001	0	-0.005	-0.003	0	0	-0.001	-0.001	-0.006	-0.004	-0.005
ES	-0.002	-0.002	-0.002	-0.004**	-0.003*	-0.004*	-0.001	-0.001	-0.001	-0.004**	-0.003*	-0.003
FI	0.001	-0.009	-0.015**	0	0.003	0.009	-0.004	-0.002	-0.004**	-0.015*	-0.016*	-0.02**
FR	0.003	0.003	0.003	-0.008**	-0.002	-0.009***	-0.005***	-0.004***	-0.002	-0.01***	-0.01***	-0.01***
GR	-0.002*	-0.003	-0.001	-0.001	-0.004*	-0.002	0	0.001	0.005	-0.009**	-0.009*	-0.005
HR	0	-0.006	-0.008*	0.001	-0.007*	-0.013**	0.008	-0.004	-0.009**	-0.001	-0.015**	-0.019***
HU	0.001	0	-0.001	0.001	-0.001	-0.001	0.001	0	0	-0.007**	-0.006***	-0.009***
IE	0.002	0.005	0.006	0	0.007	0.003	0.002	0.001	0.001	-0.005*	-0.004	-0.004
IT	-0.006	-0.005	-0.008	-0.009	-0.004	-0.009	-0.012**	-0.01*	-0.016**	-0.016**	-0.016**	-0.02**
LT	0.002	0.001	-0.001	0.006	0.005	0.008	-0.007**	0	0.003	-0.017***	-0.014***	-0.009*
LU	-0.002*	-0.002	-0.004**	0.003	0.004	0.003	0.003	0.004	0.003	0.003	0.004	0.003
LV	0.001	0.001	0	0	-0.001	0	0	-0.004	-0.004	-0.004	-0.008*	-0.007*
MT	0	0	0	0.001	-0.002	-0.001	0.001	-0.002	-0.001	0.001	-0.002	-0.001
NL	-0.003	-0.004	-0.002	-0.012***	-0.007	-0.006	-0.014***	-0.011**	-0.011***	-0.018***	-0.016***	-0.016***
PL	-0.001	-0.001	-0.001	-0.003	-0.001	-0.004*	0.001	-0.002	-0.003	-0.005	-0.007**	-0.009***
PT	0.001	0.002	0	0.002	0.004	0	0.004	0.003	-0.001	0.003	-0.002	-0.01***
RO	0.003	-0.001	-0.001	0.001	0.004	0.006	0	0.002	0.002	-0.005	-0.004	-0.004
SE	0.005	0.004	0.002	0.001	0	-0.001	-0.002	0.001	-0.002	-0.007	-0.011*	-0.011
SI	0.001	0.001	0.002	0.004	0.002	0.001	0.006	0.006	0.004	-0.005	-0.001	-0.004
SK	-0.002*	-0.001	-0.001	-0.002*	-0.002*	-0.001	-0.004***	-0.002***	-0.003***	-0.006***	-0.005***	-0.007***

Notes: Each cell represents, separately for country and each horizon, the difference $g(e_{mod.i}) - g(e_{LM.1})$. The loss function used is the absolute deviation, i.e. $g(e_{mod.i}) = E(|y_t - \widehat{y_{t,mod.i}}|)$. *, **, and *** denote significance of the difference at the 10, 5, and 1 percent level, computed according to the one-sided Diebold-Mariano test for predictive accuracy. Missing values for RF.2 and RF.3 are due to the impossibility to retrieve additional related queries for the relative countries. LM.2 is a linear model where only the SVI of $k1$ is added to the set of predictors. RF.1 is a Random Forest including the same covariates used in LM.2. RF.2 is a Random Forest where the SVI of all the retrieved keywords for each country is included plus the SVI of $k1$. RF.3 is a Random Forest model including a subset of the keywords used in RF.2, chosen using the Boruta algorithm.

B Supplementary Online Material

B.1 Random Forest and Variable Selection

B.1.1 Regression Trees and Random Forest

Random Forest as a learning method was developed by Breiman (2001) to reduce the variance of regression trees. Regression trees are non-linear and non-parametric predictive models in which the space defined by the covariates is split in sub-regions. Predictions are then computed as the sample mean of the dependent variable across the observations in the sub-regions.

The partition of the space defined by the covariates is obtained recursively. In the trivial case of a single covariate x, finding the best possible split means finding the value k such that the prediction error in the two sub-regions defined by $x < k$ and $x > k$ is minimized according to some loss function – e.g., the sum of squared errors. When the number of predictors is greater than one – i.e., $X = (x_1, x_2, ..., x_P)$ – at each step all the possible predictors and splitting values are considered, and the best split is based on the combination of the predictor-splitting value which minimize the prediction error. Once the first best split is found, the resulting sub-region is re-split iteratively using the same procedure.

The final structure of the partitions resemble the one of a tree in which the splitting nodes are the start of the branches, and the final node are the leaves. In this context, the choice of the stopping rule is crucial. On the one hand, growing a tree *too deep* might result in overfitting, hence noisy out-of-sample predictions. On the other hand, a small tree might not capture non-linearities in the relationship between the dependent variable and the covariates.

Single regression trees present an important limitation: they are extremely prone to overfitting (Hastie et al., 2009). Small changes in the data can cause large changes in

the estimated model. Random Forest was developed to reduce the variance of regression trees. It does so by considering a collection of trees (a forest), each estimated on a bootstrap sample of the original training data. Bootstrapping is not the only source of randomness. At each step of the process only a subset of the predictors, typically $P/3$, are used as potential candidates for splitting. Once B regression trees are grown, the final predictions are computed averaging the predictions of each tree. Averaging across bootstrapped trees allows to grow trees deep (typically the number of observations in the *leaves* is 5), without the risk of overfitting.

Another advantage of Random Forest is that there are very few parameters to tune, namely the number of trees of the forest B, and the number of variables considered at each step m. In our application we estimated all Random Forest models using the **R** package **randomForest**, setting $B = 5000$, and $m = P/3$.

B.1.2 Variable importance and selection

Ensemble methods, like Random Forest, are often regarded as black boxes. This is due to the implicit trade off between variance reduction (which enhances prediction accuracy) and interpretability. One important feature of Random Forest, however, is the possibility to use the B bootstrapped trees to estimate the predictive importance of the covariates used. This information can then be used for interpretation purposes. As an example, Medeiros et al. (2019) use Random Forest variable importance to show that one possible explanation for the better performance of the algorithm is its ability to capture the importance of predictors which are neglected by other linear and non-linear methods.

In this article, Random Forest uses Out-of-Bag (hereafter OOB) randomization (permutation method) to compute the importance of predictors, as described in Hastie et al. (2009). Observations are OOB in the b^{th} tree (out of the B trees of the forest) if they are excluded from the training set in that specific tree due to bootstrapping. Since each observation in the data is OOB in a fraction of the B trees (tipically $B/3$), these fraction

of trees can be used to compute the average prediction for the entire set of observations. The difference between OOB realizations and predictions across the B trees can then be used to compute the OOB error, an estimate of the true test error and a measure of predictive accuracy. Additionally, in order to compute a measure of variable importance based on prediction accuracy another step is needed. The values of the covariates used to split are randomly permuted in each split. The OOB error rate is then re-computed using the randomly permuted version of the covariates used. The diference between the two OOB error rates is then used to assess the loss in accuracy due to the random permutation the covariates. This is done separately for each of the covariates used to split the B trees. Intuitively if a covariate is important in terms of prediction accuracy, permuting at random its values should induce an increase in the OOB error rate. The average loss of accuracy due to the permutation is computed across all trees for each covariate, and is used as a measure of variable importance.

It is important to stress that variable importance measures are not used by the Random Forest algorithm in any of its steps, at least in the original definition of the algorithm. However, there are a number of contributions in the scientific literature on Random Forest which propose different methods to use variable importance measures as a way to identify the most relevant features to be included in the model. Here we focus on the Boruta algorithm.

Boruta aims at reducing the effect of random associations and fluctuations among the observed variables. The Boruta algorithm is made by nine main steps. In the first (and most meaningful) two, the algorithm copies the existing variables and rearranges the values of the copies by permuting the original values, disrupting any pre-existent relationship with the response variable. The following steps then select the predictors outperforming the randomised variables, until a stop criterion is reached. An extended description can be found in Kursa et al. (2010), while a comparison of variable selection methods is presented by Degenhardt et al. (2019).