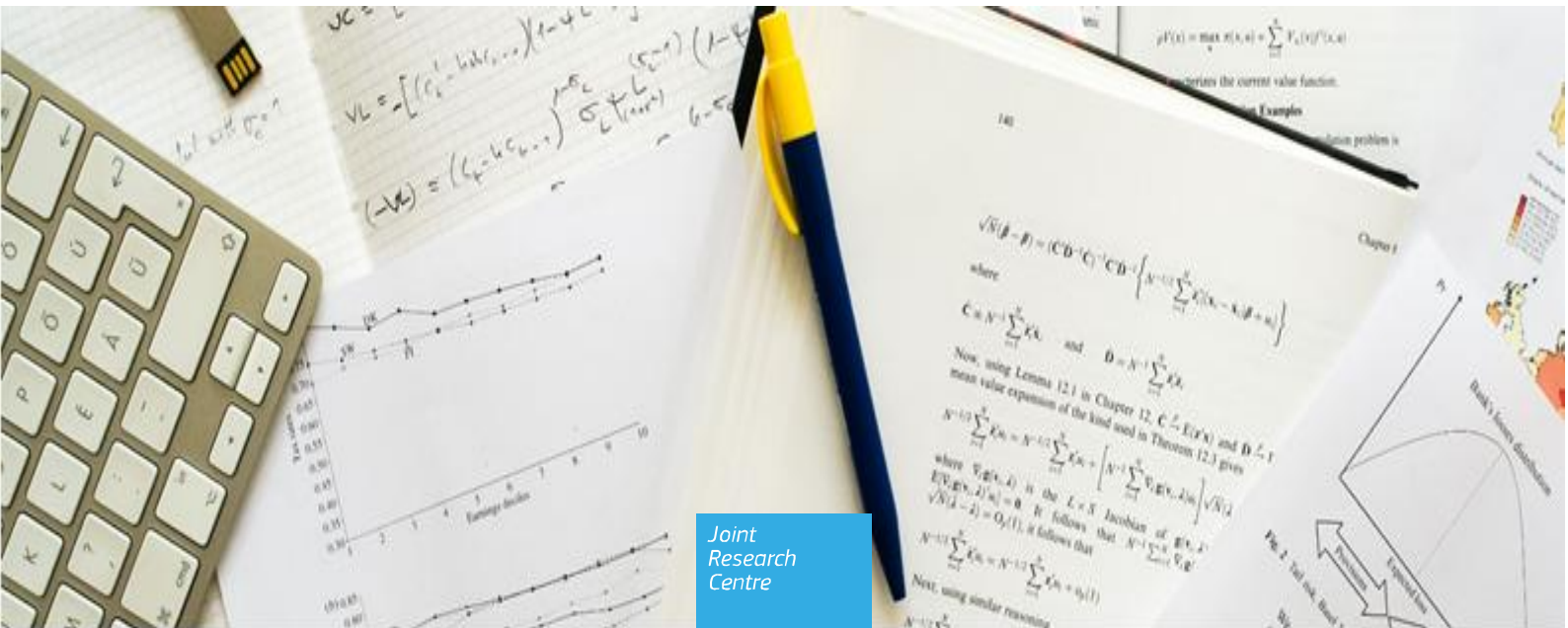
European Commission

JRC TECHNICAL REPORT

Googling Unemployment During the Pandemic: Inference and Nowcast Using Search Data

Giulio Caperna
Marco Colagrossi
Andrea Geraci
Gianluca Mazzarella

JRC Working Papers in Economics and Finance 2020/04



This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

Contact information
Name: Marco Colagrossi
Address: European Commission, Joint Research Centre (JRC), Directorate I - Competences, Unit I.1 - Monitoring, Indicators and Impact Evaluation, Via Enrico Fermi 2749, TP 361, 21027 Ispra (VA), Italy
Email: marco.colagrossi@ec.europa.eu
Tel.: +39 0332 78 9526

EU Science Hub
https://ec.europa.eu/jrc

JRC121050

PDF ISBN 978-92-76-19817-8 ISSN 2467-2203 doi:10.2760/142454

Luxembourg: Publications Office of the European Union, 2020

How to cite this report: Caperna, G., Colagrossi, M., Geraci, A. and Mazzarella, G., *Googling Unemployment During the Pandemic: Inference and Nowcast Using Search Data,* Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-19817-8, doi:10.2760/142454, JRC121050

Contents

Abstract

The economic crisis caused by the covid-19 pandemic is unprecedented in recent history. We contribute to a growing literature investigating the economic consequences of covid-19 by showing how unemployment-related online searches across the EU27 reacted to the introduction of lock-downs. We exploit Google Trends topics to retrieve over two thousand search queries related to unemployment in 27 countries. We nowcast the monthly unemployment rate in the EU Member States to assess the relationship between search data and the underlying phenomenon as well as to identify the keywords that improve predictive accuracy. Drawing from this finding, we use the set of best predictors in a Difference-in-Differences framework to document a surge of unemployment-related searches in the wake of lock-downs of about 30%. This effect persists for more than five weeks. We suggest that the effect is most likely due to an increase in unemployment expectations.

Keywords: Unemployment; nowcast; random forest; covid-19; Google Trends; Difference-in-Differences.
JEL: E24; C21; C53.

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

The economic crisis caused by the covid-19 pandemic is unprecedented in recent history. Never before, after the end of the Second World War, economic activities abruptly halted in most of the world's advanced economies. Predictions are gloomy, as the European Commission forecast that the EU GDP will shrink by about 7.7%, with the best-performing country (Poland) anticipating an economic contraction of about 4.3% (European Commission, 2020).

This study contributes to a growing literature investigating the economic impact of the covid-19 pandemic.[1] In particular, we look at the consequences of the pandemic (and the measures enacted to fight it) on unemployment, which is an exceptionally worrisome feature of this crisis. There are indeed already signs of unprecedented demand for unemployment benefits in the US (Aaronson et al., 2020; Goldsmith-Pinkham and Sojourner, 2020; Kahn et al., 2020) and of a loss of about 35 million jobs among OECD countries in March alone.[2] Further, the impact on seasonal activities – such as tourism and agriculture – on which several EU countries depends heavily might be particularly severe.[3] 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 EU28 non-financial business sector (NFBS) and employ more than 65% of the workers in non-NFBS activities (Hope et al., 2019).

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

We contribute to this literature in a variety of ways. We are the first, to the best of our knowledge, to provide a EU-wide perspective. Differently from the US case, timely and high-frequency administrative data on unemployment-related benefits are not available in the EU. We take a different approach and rely on unemployment-related online searches. In particular, to extend our analysis to all EU27 countries, we exploit Google Trends topics (as in Brodeur et al., 2020), which are an aggregation of different queries having the same semantic meaning – in our case, all queries linked to the broad concept of unemployment. We then adopt a novel approach, which has not yet been used in the literature. We exploit the topic "unemployment" to collect, for each country, all search queries linked to it (first level queries) and all the top queries linked to the latter (second level 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 3section). Instead, once we add all the keywords linked to the topic unemployment and perform variable selection using random forest-based methods, the predictive accuracy increases significantly in almost all countries.

Drawing from this finding, we select the queries that best predict the unemployment rate, separately for each country, and aggregate them to create a daily indicator of unemployment-related searches. Such an indicator, we argue, contains information on both the current unemployment rate and expectations of unemployment. We then perform 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 methodology and the approach outlined in this paper is not only relevant in the context of the covid-19 pandemic and unemployment. It could be adapted to study a variety of events, policies and economic indicators, making it a powerful tool to inform policymakers when reporting data are not readily available.

The remainder of this paper is structured as follows: Section 2section briefly introduces Google search data. Section 3section reviews the related literature on nowcasting and describes our methodology and results. Section 4section builds on the findings of the nowcasting exercise and shows the results of the DiD. Section 5section concludes.

[1] 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; Kahn et al., 2020; Ludvigson et al., 2020; Stock, 2020), the impact on financial markets and their stability (e.g., Boot et al., 2020; Ramelli and 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 (e.g., Brodeur et al., 2020; Fetzer et al., 2020; Hamermesh, 2020).

[2] The loss seems to have been particularly severe among youth and women – see Unemployment Rates, OECD - Updated: May 2020, available at `http://www.oecd.org/sdd/labour-stats/unemployment-rates-oecd-update-may-2020.htm`

[3] See "Tourism and transport in 2020 and beyond", Brussels, 13.5.2020 COM(2020) 550 final, available at `https://ec.europa.eu/info/sites/info/files/communication-commission-tourism-transport-2020-and-beyond_en.pdf`

2 Google searches

Google searches have been used in various fields of the economic literature. Starting with the seminal contribution of Choi and Varian (2012), Google search data proved useful to forecast (nowcast) a variety of economic indicators (see Section 3section for a brief review of the main contributions in this literature). Further, they have been used in financial studies (e.g., Da et al., 2011; Preis et al., 2013; Vlastakis and Markellos, 2012), to understand tourism flows (Siliverstovs and Wochner, 2018) and even the consequences of racial animus on black candidates in the US presidential elections (Stephens-Davidowitz, 2014). Importantly, compared to surveys, Google searches are less sensitive to the small-sample bias (Baker and Fradkin, 2017).

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.[4] Search 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 Search Volume Index (hereafter 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.[5]

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.

As mentioned in Section 1section, in this paper we exploit a variety of features of Google Trends, a novelty in this literature. First, we collect the monthly SVI for the topic "unemployment" starting in January 2015. Further, we collect, for the same period of time, the monthly SVI for "level-1" queries – i.e., the top-25 related search terms associated with the topic. Finally, we collect the monthly SVI for "level-2" queries – i.e., the 10-top related search terms associated with level-1 queries. This procedure is repeated separately for each country for the period January 2015 - December 2019. For the DiD (Section 4section) 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 3section) from the 13th of January to the 9th of May 2020.[6]

Of course, Google searches also present 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.[7] 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 Google searches and unemployment rate in the EU

In the last decade Google search data have been used to forecast and nowcast different macroeconomic indicators. Götz and Knetsch (2019) use Google data to forecast German GDP, Vosen and Schmidt (2011) and Vosen and 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 and Heiden (2015) create a proxy for investors attention to predict stock market volatility. In a recent contribution, Koop and Onorante (2019) show how Google search

[4] 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.

[5] 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.

[6] 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.

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

data can be used to improve nowcast of different macroeconomic variables in the context of dynamic model selection. Finally, focusing on unemployment, D'Amuri and Marcucci (2017) assess the performance of Google search data related to job-search in forecasting US monthly unemployment rate. Fondeur and 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.

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 with the aim of assessing if, and to which extent, Google search data related to unemployment are a good predictor of the true economic phenomenon.

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 – through τU_t – and $t+1$ – through $\tau(\widetilde{\delta}_{t,t+1} + \epsilon_t)E_t$. 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 and 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 and 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".

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 follow a different approach and 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.[8] 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.[9] 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

[8] 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

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

considered three different horizons (i.e., $h = 1, 2, 3$) corresponding to the last date for which information on unemployment is available.[10]

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 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).[11]

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 37th month; c) the models are then re-trained using the first 37 observations and predictions for the 38th 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 and Mariano, 1995) based on absolute deviations.[12] The aim of this test is to assess whether Google search data carry additional informational content.

Figure 1 summarizes the main findings. Table F.2 in Appendix 6section 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 (α = 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 sizeable 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 also 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 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 and 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.

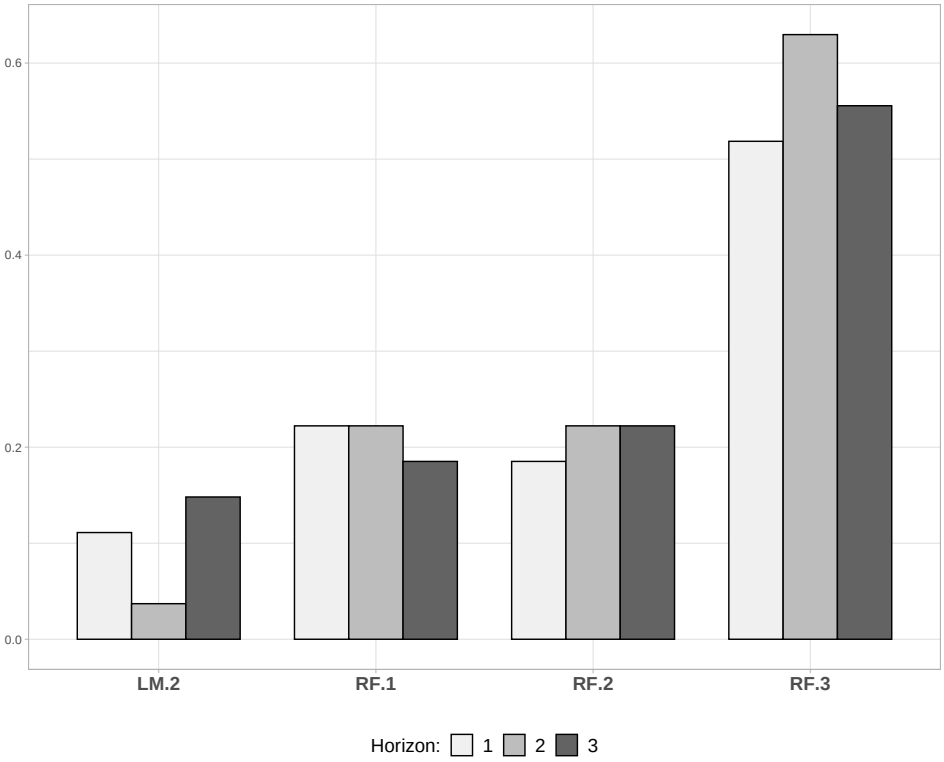
Overall, the results suggest that a subset of relevant keywords helps to improve nowcast accuracy with respect to the benchmark model. This is not the case for the topic alone. Combining the use of topics and the variable selection step in our nowcast framework presents two advantages. On the one hand, the use of a common Google topic allows to retrieve a broad set of keywords in a context of heterogeneous countries

[10] 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.

[11] A brief description of the Random Forest algorithm and of the Boruta variable selection method is provided in the supplementary online material 6.1. 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.

[12] 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 1: 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.

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. This is particularly important when Google Search data are used for causal inference (Section 4section). The use of topics alone could pose the risk of identifying spurious relationship not linked with the underlying phenomenon.

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

In this section, we investigate the effect of lock-down measures on unemployment-related online searches using DiD. Although the daily SVI of the topic ($k1$) is an ideal candidate for this task in a multi-country setup, we showed that the topic alone is a poor predictor of the unemployment rate. For this reason, we complement the analysis using, as dependent variable, an indicator based on the country-specific subset of best predictors identified in the previous section. The proposed indicator ($k1_u$) is a weighted linear combination of the SVIs of the selected keywords. The weights are obtained regressing, separately for each country, the daily SVI of the topic on the daily SVI of the selected keywords. Intuitively, this allows us to extract the component of the topic explained by the keywords that best predict the unemployment rate.

The data are complemented with information about Governments' announcements of measures to respond to the crisis. In particular, we focus on the lock-down measures enacted by EU governments as recorded by The Assessment Capacities Project (ACAPS).[13] 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.[14] Data are collected from the 13th of January to the 9th of May. We estimate the following regression:

[13] In Table 1 we show that results are robust to a different definition of lock-down recorded by the Blavatnik School of Government of the University of Oxford. See `https://github.com/OxCGRT/covid-policy-tracker`.

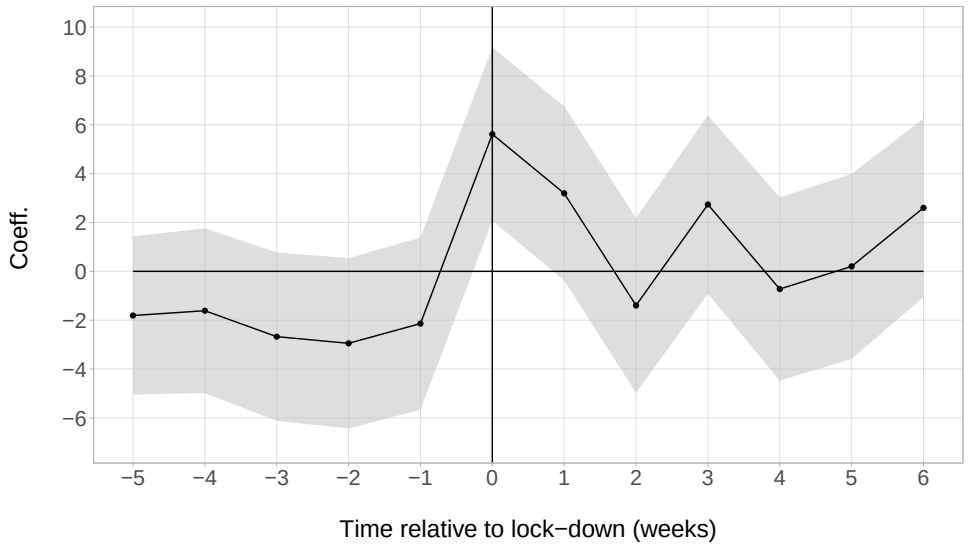
[14] Luxembourg and Portugal are then excluded from our sample due to the unavailability of related keywords. The dates considered are those of the measure's announcement: Austria 16-03; Belgium 18-03; Bulgaria 20-03; Cyprus 24-03; Croatia 18-04; Denmark 18-03; Estonia 30-03; France 17-03; Germany 21-03; 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.

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

where $y_{c,t}$ is either the daily SVI of the topic $k1$ or of the indicator $k1_u$ 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 2.

Figure 2: DiD coefficients for $k1$ and $k1_u$

(a) Topic ($k1$)



(b) Indicator ($k1_u$)

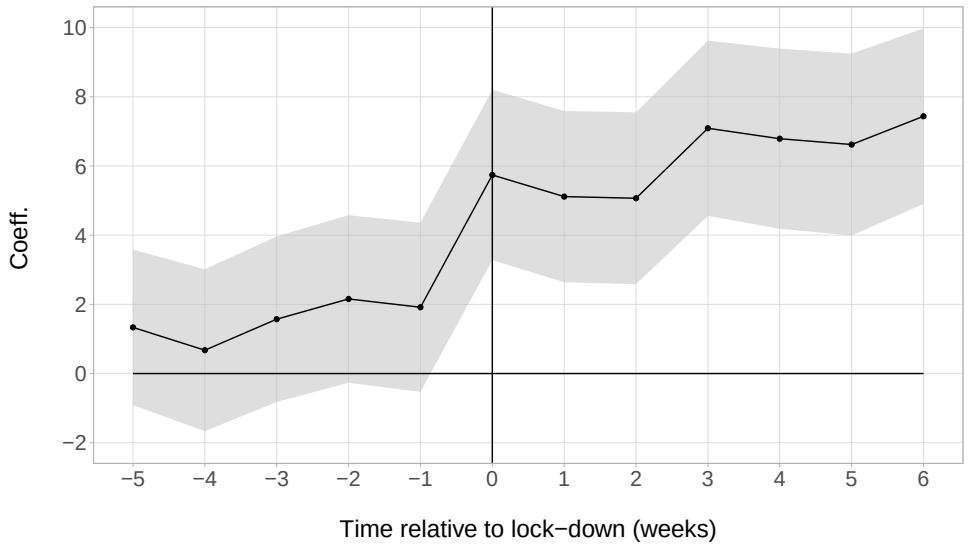


Figure 2 (a) and (b) shows, respectively, the results for $k1$ and $k1_u$. Both measures exhibit an increase of roughly 30% in the week of the announcement of the first introduction of the lock-down. While the effect on $k1$ is short-lived, the opposite is true for $k1_u$. The higher level of unemployment-related 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.

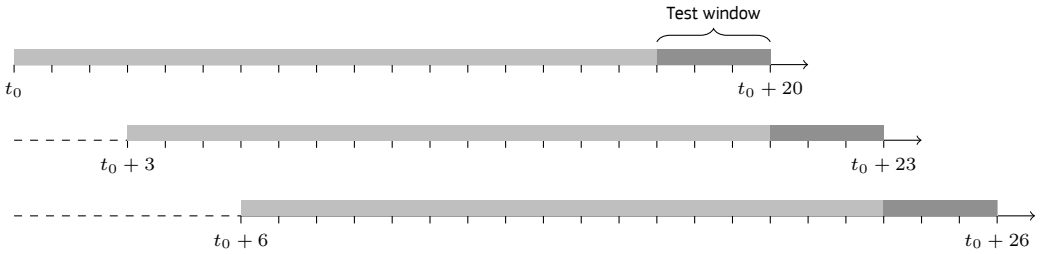
Our findings for the EU27 compare favourably to those by Aaronson et al. (2020) for the US. Aaronson et al. show that unemployment-related queries surged before the record increase in unemployment insurance claims,

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 the conceptual framework presented in Section 3section.

Lock-downs are not the only measures enacted by Governments that might have affected 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 the SVI of $k1_u$ exhibits a significant increase. We do so conducting country-specific rolling window event-studies. Starting from the first available date – 13th 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 May (see Figure 3).

Figure 3: 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 2: 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

We contribute to a fast-growing literature looking at the economic consequences of the covid-19 pandemic. More specifically, we focus on unemployment and investigate the response to lock-down measures enacted by Governments to fight the spread of the SARS-CoV-2 virus.

In the absence of timely and high-frequency EU-wide administrative data, we resort to online search activities. We first show how Google searches related to unemployment are linked with the underlying phenomenon – the unemployment rate in each EU27 country. Faced with the difficulty of a multi-language and multi-institutional context, we propose a procedure in which we exploit Google Trends topics to retrieve over two-thousand search queries related to unemployment in the EU27.

To test whether Google searches contain information on the underlying phenomenon, we nowcast the country-level unemployment rate time series. We show that the topic alone does not add a significant amount of information over a simple autoregressive model in most of the EU27 countries. Instead, Random Forest variable selection methods allow us to identify a set of keywords increasing predictive accuracy.

Based on this finding, we select the variables best-describing the unemployment rate and aggregate them to create a daily indicator of unemployment-related searches. 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 the light of our conceptual framework, we interpret this finding as an increase in unemployment expectations.

Importantly, the methodology described in this paper is not only relevant in the context of the covid-19 pandemic. It could be used to study a variety of events, policies and economic indicators, making it a powerful tool to inform policymakers when administrative data are not timely (and readily) available.

References

Aaronson, D., Brave, S. A., Butters, R., Sacks, D. W. and Seo, B., "Using the eye of the storm to predict the wave of covid-19 ui claims", Tech. Rep. 2020-10, Federal Reserve Bank of Chicago, 2020. URL https://www.chicagofed.org/~/media/publications/working-papers/2020/wp2020-10-pdf.pdf.

Adams-Prassl, A., Boneva, T., Golin, M. and Rauh, C., "Inequality in the impact of the coronavirus shock: Evidence from real time surveys", Tech. Rep. 13183, IZA Institute of Labor, 2020. URL http://ftp.iza.org/dp13183.pdf.

Akira Toda, A., "Susceptible-infected-recovered (sir) dynamics of covid-19 and economic impact", Tech. rep., arXiv:2003.11221, 2020. URL https://arxiv.org/pdf/2003.11221v2.pdf.

Alon, T. M., Doepke, M., Olmstead-Rumsey, J. and Tertilt, M., "The impact of covid-19 on gender equality", Tech. rep., National Bureau of Economic Research, 2020. URL https://www.nber.org/papers/w26947.

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

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

Baker, S. R., Bloom, N., Davis, S. J. and Terry, S. J., "Covid-induced economic uncertainty", Tech. rep., National Bureau of Economic Research, 2020. URL https://www.nber.org/papers/w26983.

Baker, S. R. and Fradkin, A., "The impact of unemployment insurance on job search: Evidence from google search data", *Review of Economics and Statistics*, Vol. 99, No 5, 2017, pp. 756–768.

Boot, A. W., Carletti, E., Kotz, H.-H., Krahnen, J. P., Pelizzon, L. and Subrahmanyam, M. G., "Corona and financial stability 3.0: Try equity-risk sharing for companies, large and small", Tech. Rep. 81, Leibniz Institute for Financial Research SAFE, 2020. URL http://hdl.handle.net/10419/215544.

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., "Google trends terms reporting rhinitis and related topics differ in european countries", *Allergy*, Vol. 72, No 8, 2017, pp. 1261–1266.

Breiman, L., "Random forests", *Machine learning*, Vol. 45, No 1, 2001, pp. 5–32.

Brodeur, A., Clark, A. E., Flèche, S., Powdthavee, N. et al., "Covid-19, lockdowns and well-being: Evidence from google trends", Tech. rep., Institute of Labor Economics (IZA), 2020. URL http://ftp.iza.org/dp13204.pdf.

Choi, H. and Varian, H., "Predicting the present with google trends", *Economic record*, Vol. 88, 2012, pp. 2–9.

Coronini-Cronberg, S., John Maile, E. and Majeed, A., "Health inequalities: the hidden cost of covid-19 in nhs hospital trusts?", *Journal of the Royal Society of Medicine*, Vol. 113, No 5, 2020, pp. 179–184.

Da, Z., Engelberg, J. and Gao, P., "In search of attention", *The Journal of Finance*, Vol. 66, No 5, 2011, pp. 1461–1499.

Da, Z., Engelberg, J. and Gao, P., "The sum of all fears investor sentiment and asset prices", *The Review of Financial Studies*, Vol. 28, No 1, 2015, pp. 1–32.

Degenhardt, F., Seifert, S. and Szymczak, S., "Evaluation of variable selection methods for random forests and omics data sets", *Briefings in bioinformatics*, Vol. 20, No 2, 2019, pp. 492–503.

Diebold, F. X. and Mariano, R. S., "Comparing predictive accuracy", *Journal of Business & Economic Statistics*, Vol. 13, No 3, 1995, pp. 253–263.

D'Amuri, F. and Marcucci, J., "The predictive power of google searches in forecasting us unemployment", *International Journal of Forecasting*, Vol. 33, No 4, 2017, pp. 801–816.

European Commission, "European economic forecast, spring 2020", Tech. rep., DG Economic and Financial Affairs, 2020. URL https://ec.europa.eu/info/sites/info/files/economy-finance/ip125_en.pdf.

Fetzer, T., Hensel, L., Hermle, J. and Roth, C., "Coronavirus perceptions and economic anxiety", Tech. rep., arXiv:2003.03848, 2020. URL https://arxiv.org/pdf/2003.03848.pdf.

Fondeur, Y. and Karamé, F., "Can google data help predict french youth unemployment?", *Economic Modelling*, Vol. 30, 2013, pp. 117–125.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C., "Variable selection using random forests", *Pattern recognition letters*, Vol. 31, No 14, 2010, pp. 2225–2236.

Goldsmith-Pinkham, P. and Sojourner, A., "Predicting initial unemployment insurance claims using google trends", Tech. rep., Yale School of Management, 2020. URL https://paulgp.github.io/GoogleTrendsUINowcast/google_trends_UI.html.

Götz, T. B. and Knetsch, T. A., "Google data in bridge equation models for german gdp", *International Journal of Forecasting*, Vol. 35, No 1, 2019, pp. 45–66.

Hamermesh, D. S., "Lock-downs, loneliness and life satisfaction", Tech. rep., National Bureau of Economic Research, 2020. URL https://www.nber.org/papers/w27018.

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

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

Hope, K. et al., "Annual report on european smes 2018/2019", Tech. rep., DG for Internal Market, Industry, Entrepreneurship and SMEs, 2019. URL DOI:10.2826/500457.

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

Kahn, L. B., Lange, F. and Wiczer, D. G., "Labor demand in the time of covid-19: Evidence from vacancy postings and ui claims", Tech. rep., National Bureau of Economic Research, 2020.

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

Kursa, M. B., Rudnicki, W. R. et al., "Feature selection with the boruta package", *Journal of Statistical Software*, Vol. 36, No 11, 2010, pp. 1–13.

Ludvigson, S. C., Ma, S. and Ng, S., "Covid19 and the macroeconomic effects of costly disasters", Tech. rep., National Bureau of Economic Research, 2020. URL https://www.nber.org/papers/w26987.

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

Preis, T., Moat, H. S. and Stanley, H. E., "Quantifying trading behavior in financial markets using google trends", *Scientific reports*, Vol. 3, 2013, p. 1684.

Ramelli, S. and Wagner, A. F., "Feverish stock price reactions to covid-19", Tech. rep., Centre for Economic Policy Research, 2020. URL https://cepr.org/active/publications/discussion_papers/dp.php?dpno=14511.

Şahin, A., Tasci, M. and Yan, J., "The unemployment cost of covid-19: How high and how long?", Tech. Rep. 2020-09, Federal Reserve Bank of Cleveland, 2020. URL https://doi.org/10.26509/frbc-ec-202009.

Siliverstovs, B. and Wochner, D. S., "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*, Vol. 145, 2018, pp. 1–23.

Smith, P., "Google's midas touch: Predicting uk unemployment with internet search data", *Journal of Forecasting*, Vol. 35, No 3, 2016, pp. 263–284.

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

Stock, J. H., "Data gaps and the policy response to the novel coronavirus", Tech. rep., National Bureau of Economic Research, 2020. URL https://www.nber.org/papers/w26902.

Stoppiglia, H., Dreyfus, G., Dubois, R. and Oussar, Y., "Ranking a random feature for variable and feature selection", *Journal of machine learning research*, Vol. 3, No Mar, 2003, pp. 1399–1414.

Vlastakis, N. and Markellos, R. N., "Information demand and stock market volatility", *Journal of Banking & Finance*, Vol. 36, No 6, 2012, pp. 1808–1821.

Vosen, S. and Schmidt, T., "Forecasting private consumption: survey-based indicators vs. google trends", *Journal of forecasting*, Vol. 30, No 6, 2011, pp. 565–578.

Vosen, S. and Schmidt, T., "A monthly consumption indicator for germany based on internet search query data", *Applied Economics Letters*, Vol. 19, No 7, 2012, pp. 683–687.

List of Tables

6 Appendix

Table F.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.

Supplementary Online Material

6.1 Random Forest and Variable Selection

6.1.1 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 $m = P/3$.

6.1.2 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).

GETTING IN TOUCH WITH THE EU

In person

All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

On the phone or by email

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

- at the following standard number: +32 22999696, or

- by electronic mail via: https://europa.eu/european-union/contact_en

FINDING INFORMATION ABOUT THE EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

EU publications

You can download or order free and priced EU publications from EU Bookshop at: https://publications.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

KJ-AE-20-008-EN-N

The European Commission's science and knowledge service

Joint Research Centre

JRC Mission

As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub

ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office of the European Union

doi:10.2760/142454

ISBN 978-92-76-19817-8