# Tracking socio-economic activities in European countries with unconventional data

Marco Colagrossi
European Commission, Joint Research Centre (JRC)
Ispra (VA), Italy
marco.colagrossi@ec.europa.eu

Sergio Consoli
European Commission, Joint Research Centre (JRC)
Ispra (VA), Italy
sergio.consoli@ec.europa.eu

Francesco Panella
European Commission, Joint Research Centre (JRC)
Ispra (VA), Italy
francesco.panella@ec.europa.eu

Luca Barbaglia
European Commission, Joint Research Centre (JRC)
Ispra (VA), Italy
luca.barbaglia@ec.europa.eu

## ABSTRACT

This contribution shows our ongoing work aimed at monitoring societal issues and economic activities (e.g., industrial production, unemployment, loneliness, cultural participation) across EU member states mining unconventional data sources to complement official statistics. Considered unconventional data sources include the Global Dataset of Events, Language and Tone (GDELT), Google Search data, and Dow Jones Data, News and Analytics (DNA). We show an early experiment aiming at nowcasting unemployment in Germany, Spain, France, and Italy, demonstrating the added value of these data both for scholars and policymakers.

## CCS CONCEPTS

• **Applied computing** → *Forecasting*; *Economics*; • **Human-centered computing** → *Collaborative and social computing*; • **Information systems** → *Business intelligence*; *Data mining*;

## KEYWORDS

alternative (big) datasets, text analysis, social media

**ACM Reference Format:**
Marco Colagrossi, Sergio Consoli, Francesco Panella, and Luca Barbaglia. 2022. Tracking socio-economic activities in European countries with unconventional data. In *Conference on Information Technology for Social Good (GoodIT'22), September 7–9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3524458.3547242

## 1 INTRODUCTION AND BACKGROUND

Rapid advancements in Information and Communications Technology have produced an exponential growth in the amount of collected and available information. The *Big Data* era [35] also affected economic and social sciences. From forecasting economic and social phenomena with just a few observations and variables, scholars can now profit from plenty. Questions that could be previously answered only with a significant delay can now be addressed nearly in real-time, albeit posing challenges from a technical perspective – for instance, handling, processing, linking and analyzing this vast amount of data. Against this background, data science and artificial intelligence [34], deep machine learning [23, 32], and the increasing availability of performing hardware (e.g., cloud computing infrastructures, GPUs [40]) are playing a crucial role. This has been particularly relevant during the COVID-19 pandemic [22, 45], when information from unconventional data sources (e.g., news, search and social media data) has been used to integrate and augment the information produced by national and international statistical agencies [7]. In general, the evolution of these technologies has generated insights and contributed to the development of several decision-making instruments helping investors in making decisions as well as policymakers in designing policy interventions with the potential of fostering economic growth and societal well-being.

Traditional forecasting models usually adopt a mixed frequency approach which bridges information from high frequency social and economic indexes (e.g., weekly initial claims) with low-frequency variables, for instance, quarterly GDP [29]. Yet, traditional methods present two main drawbacks when applied to large data sets. First, they cannot directly handle (a large amount of) unstructured data. Second, even if augmented with new predictors obtained from alternative (big) data sources, the relationship across variables is often assumed to be linear, which is not the case for the majority of the real-world cases [2, 20].

In this context, big data applications can potentially deal with these issues [6]. Machine learning algorithms can extract novel insights from unstructured information and account for non-linear dynamics across socio-economic variables. They can grasp hidden knowledge even when the number of features under analysis is larger than the available observations (i.e., high dimensional data), which occurs increasingly often in social and economic environments. Differently from traditional time-series techniques, machine learning methods have no ex-ante assumptions about the stochastic process underlying the state of the economy or society. For instance, in deep learning highly non-linear data are modelled by deriving or learning directly from the data the order of non-linearity.

Examples of big data sources that can potentially be useful for socio-economic analyses include administrative data (e.g., tax and hospital records), commercial data sets (e.g., consumer panels,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*GoodIT'22, September 7–9, 2022, Limassol, Cyprus*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9284-6/22/09…$15.00
https://doi.org/10.1145/3524458.3547242

credit/debit card transactions), and textual data (e.g., social media, web searches, news data). In some cases, the data sets are structured and ready for analysis, while in other cases, for instance text, the data is unstructured and requires some preliminary steps to extract and organize the relevant information.

These opportunities and challenges are inspiring the research activities at the European Commission's Competence Center on Composite Indicators and Scoreboards (COIN)[1] at the Joint Research Centre (JRC)[2]. Currently on-going work, described in this contribution, aims at tracking socio-economic activities by obtaining policy-relevant insights from data sets which are considered unconventional in social sciences as well as stimulating the adoption of cutting-hedge modeling technologies in the EU intuitions.

Our recent work aims at monitoring European citizens' interest and opinions about their alignment with the ongoing EC work programme. To support the EC in this effort, it is possible to create, for instance, composite indicators to monitor EU citizens' interest in topics related to the major EC policy goals over time, and across the member states and their regions. In the context of the European Green Deal, we use web searches related to (i) citizens' green behaviour in the context of specific policy areas, such as their searches related to mobility, energy, waste, and nutrition; and (ii) citizens' green awareness, such as those about pollution, clean energy, ecosystems and climate [41]. In addition, we aim at incorporating and aggregating also information extracted from news media (i.e., emotions and sentiment), which has shown already great potential [4, 5, 14–16]. News media might be used as a powerful additional feature in forecasting applications since they timely describe current events, represent the updated expectations of economic agents, and significantly influence investors' perception and policymakers' decisions.

The remainder of this contribution describes our work-in-progress on the use of unconventional data for socio-economic analyses. We began by discussing the data sets currently adopted, detailing the steps needed to properly handle, process, and extract useful information from each of these data sources (Section 2). We then describe the Business Intelligence (BI) dashboards produced to interact and visualize the signals drawn from the data (Section 3). Section 4 then shows an application using unconventional data for socio-economic analyses focusing on predicting the unemployment rate in France, Germany, Italy and Spain. In Section 5 we provide our conclusions and an overlook on the future work.

# 2 UNCONVENTIONAL DATA FOR SOCIO-ECONOMIC ANALYSIS

## 2.1 Global Dataset of Events, Language and Tone (GDELT)

GDELT[3] is an open, big data platform of meta-information extracted from broadcast, print, and web news collected worldwide and translated nearly in real-time into English from over 65 different languages [30, 33]. Extracted and processed information are stored in different databases, with the most comprehensive one being the GDELT Global Knowledge Graph (GKG). GKG is a news-level data set starting from February 2015 freely available to users through custom REST APIs.[4] Each news in GDELT GKG provides information on people, locations and organizations mentioned in the text of the article and retrieves counts, quotes, images and themes using a number of popular topical taxonomies. The output of GDELT processing is updated on the its website every 15 minutes.[5] In terms of volume, GKG analyses over 88 million articles a year and more than 150,000 news outlets.

The huge amount of unstructured documents coming from GDELT has been re-engineered and stored into an ad-hoc Elasticsearch infrastructure [24, 38]. Elasticsearch is a popular and efficient document-store built on the Apache Lucene search library[6], providing real-time searches and analytics for different types of complex data structures, like text, numerical data, or geospatial data, that have been serialized as JSON documents. For our applications, we use the GKG themes to filter out news related to certain social or economic topics (e.g., "economy", "industrial production", "unemployment", "inflation", "capital market", "cultural activities", "housing market", "international trade", "monetary policy", "loneliness"), limiting only to the news of the European country we are interested about. After this processing, we compute as output the following measures: *(i) Article Tone* - score between $-1$ and $+1$ expressing whether a certain message conveys a positive or negative sentiment with respect to a certain topic in the text, calculated by GDELT[7] by averaging the tone scores of the terms contained in the text using the VADER sentiment dictionary[8], a popular lexicon-based approach. *(ii) Topic Popularity Rate* - number of articles referred to the searched topic (i.e., "cultural employment") normalized by the total number of available articles in the period of interest.

Our extractions from GDELT start in February 2015 until the current day, given we perform regular monthly updates. The main steps of the extraction algorithm are the following:

1. Specify the topic of interest, providing a curated list of representative keywords.
2. Specify the country to focus (one among all EU-27+UK countries).
3. Skip articles too short from the analysis (minimum word count = 500).
4. Specify the interval period for the extraction. It can be *day*, *week*, *month*, *quarter*, *year*.
5. For consistency, consider only the journal outlets that are present within the entire period of extraction, from the start to the end.
6. Compute the Tone score and Topic Popularity Rate by averaging the obtained measures for the selected articles by the period of extraction.
7. Data treatment, like smoothing and variable standardization.

Suppose, for example, that you want to extract the news whose main themes are related to the *Unemployment* topic and focus on

[1] European Commission's Competence Center on Composite Indicators and Scoreboards (COIN): https://composite-indicators.jrc.ec.europa.eu/.

[2] The Joint Research Centre (JRC) of the European Commission (EC): https://ec.europa.eu/info/departments/joint-research-centre_en.

[3] GDELT website: https://blog.gdeltproject.org/

[4] See https://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

[5] See http://data.gdeltproject.org/gdeltv2/lastupdate.txt for the English data, while http://data.gdeltproject.org/gdeltv2/lastupdate-translation.txt for the translated data.

[6] Apache Lucene search library: https://lucene.apache.org/

[7] https://blog.gdeltproject.org/vader-sentiment-lexicon-now-available-in-gcam/.

[8] VADER: https://github.com/cjhutto/vaderSentiment.

*Italy*. First the user has to select a list of representative keywords for the topic (Step 1). For instance, for the *Unemployment* topic: *unemployment, unemployment rate, job opportunities, job market,* .... Second, the country to focus on, that is, *Italy* for this specific example (Step 2). Third, the period of extraction, from 2015 to 2021 (Step 3). The list of curated keywords specified by the user is further extended programmatically by means of synonyms, which are computed by using the Sense2Vec python library[9], an extension to the popular Word2Vec model [36] that learns the semantic similarities across all word vectors within the Reddit comments[10] from 2015 to 2019. The obtained list of synonyms for the *Unemployment* topic include concepts like *labour market, wage growth, unemployment benefits, joblessness,* .... After we get the complete list of input keywords, we match only the articles from GDELT such that the GDELT topics are related to one of the selected themes of interest. To obtain this match, we rely on Word Embedding [28], a popular language model for the representation of words for text analysis. We use the pre-trained Word Embeddings from the GloVe model [37], an unsupervised learning algorithm for obtaining word vector representations trained on Wikipedia and Gigawords. We are then able to select the GDELT themes which are closer to the input themes list by picking up only those GDELT themes whose cosine similarity values with respect to the word vectors of the input keywords are higher than a threshold, meaning they are semantically similar. In our case a threshold equal to 0.7 has been experimentally set.

After this selection procedure, to obtain a pool of news that is not too heterogeneous in length, we consider only articles that are at least 500 words long (Step 3). Once collected the relevant news data, we map this information to the relevant interval period (e.g. *day*, Step 4). For consistency of the analysis, we also consider only the outlets that are present within the entire period of extraction, i.e. from the start to the end (Step 5). We are then able to calculate *Articles Tone* score and *Topic Popularity Rate* by averaging the obtained measures from GDELT for the selected articles by the period of extraction (Step 6). We allow users, as an option, to perform some additional data treatment steps (Step 7), like, for instance, smoothing of the time series via Moving Average, focusing more on the trend of the variable, and variable standardization, to map the different variables on the same scale by scaling them to the standard normal distribution (subtracting the mean and dividing by the standard deviation of the data sample).

## 2.2 Google Search data

Starting with the seminal contribution of Choi and Varian [13], Google Search data have been used to proxy a variety of economic indicators, for instance, to forecast German GDP [25]; to forecast consumption in the US [42] and Germany [43]; to build an investment sentiment index to predict different US aggregate market indices [17] and stock market volatility [26]; to improve nowcasting performances of different macroeconomic variables in the context of dynamic model selection [27]; to assess how online job-searches can improve forecasting the US monthly unemployment rate [18]; to understand tourism flows [39]; and also to estimate the impact of the advertised degree of "greenness" on house prices [46]. Web searches are particularly attractive in those contexts in which data are not available, available at a low frequency, or available with some delay. Furthermore, compared to surveys, web searches are less sensitive to small-sample biases [3]. These characteristics have made Google Search data an ideal source of information for researchers during the COVID-19 pandemic. Scholars have used searches to predict the number of unemployment insurance (UI) claims in the US [1, 8, 21, 31]. Other applications have used web searches to investigate the impact of lock-downs on well-being or economic anxiety [10, 19]. Brunori and Resce [11] showed instead how web queries related to symptoms can be used to monitor the diffusion of the virus. Work at the European Commission's JRC investigated whether searches could be used to monitor citizens' interests in three main fields related to the pandemic crisis: health, the economy and social isolation.[11]

The use of web 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 each language and institutional context. This task is particularly challenging in a cross-country setting, where finding the correct queries is either costly (in terms of time) or not feasible (due to language barriers). Recent applications [10, 12, 41] began exploiting Google Trends topics, which are aggregations of different queries belonging to the same semantic concept.[12] Topics provide some advantages over queries. First, each topic has an unique identifier, making topics language-independent. Therefore, it is possible to perform cross-country analyses without translating or identifying the different set of queries belonging to the same semantic concept across different countries and languages. This is particularly relevant as search queries related to the same semantic concept vary across countries due to cultural and institutional differences [9]. Second, queries linked to the same semantic concept may vary over time. Topics include (or exclude) the relevant set of queries over time. Finally, all queries broadly related to a topic are then linked to it independently from the spelling and the wording of the associated queries.

Google Search data are available through Google Trends.[13] Google Trends returns the Search Volume Index (SVI) of both queries and topics. Results are normalized to the time and location of a query. By time range (either daily, weekly or monthly) and geography (either country or ISO 3166-2), each data point is divided by the total searches to obtain relative popularity. Depending on the type of access the resulting numbers are then (i) scaled on a range of 0 to 100 based on a query's proportion to all searches on all queries (if the Google Trends end-point is used); or (ii) multiplied by 10 million (if the private non-commercial Google Trends API is used). In both cases, numbers are calculated on a uniformly distributed random

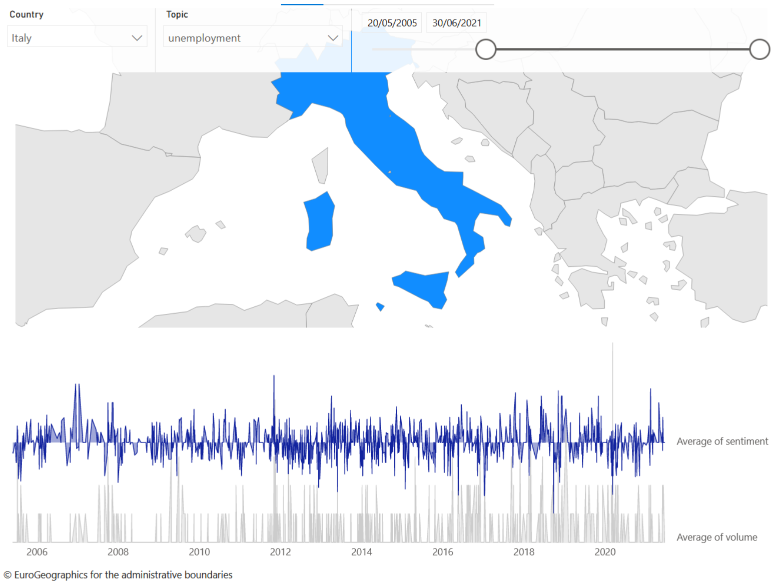
[9] Sense2Vec library: https://pypi.org/project/sense2vec/.

[10] Reddit: https://www.reddit.com/.

[11] See https://knowledge4policy.ec.europa.eu/projects-activities/tracking-eu-citizens%E2%80%99-concerns-using-google-search-data_en.

[12] Topics are Google Knowledge Graph entities (https://developers.google.com/knowledge-graph). Until late 2013, Knowledge Graph entities matched those in the Wikidata knowledge base (https://www.wikidata.org/), where it is still possible to search for a topic and use the ID from its Freebase identifier (https://www.wikidata.org/wiki/Property:P646) to denote all searches that were classified to be about this topic. However, starting in 2014, not all Google Knowledge Graph entities are available through Wikidata.

[13] https://trends.google.com/trends/.



Country: Italy

Topic: unemployment

20/05/2005 30/06/2021

© EuroGeographics for the administrative boundaries

**Figure 1: Example of the Power BI dashboard developed for the DNA data and showing the processed information (i.e., news sentiment and volume of extracted news) for the *Unemployment* news topic in *Italy* between 2005 and 2021.**

sample of Google web searches done since 2004, updated once a day, thus there may be some variance between similar requests. Finally, Google also provides 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.

## 2.3 Dow Jones Data, News and Analytics (DNA)

Another unconventional data that we consider are newspaper articles. Real-time news contain information about the mood of economic and financial agents and can be used to build a proxy of the current state of the economy. We obtain news data from the Dow Jones Data, News and Analytics (DNA) platform.[14] In particular, we download news articles published by Thomson Reuters News from January 1988 to June 2021, summing to a total of approximately 7.4 million items. The information provided is the publication date, the title and body of the article. The newspaper articles deal with a wide set of topics, ranging from financial matters to macro-economic announcement or political implications on national economies. Note that the articles included in DNA might overlap with the information set available in GDELT, as described in Section 2.1. While GDELT provide open-access to a list of pre-calculated tone and popularity scores, the added value of DNA consists of granting proprietary access to the raw full texts, thus allowing the researcher to optimally choose the text mining strategy to analyze the article information.

We use this news data set to build a set of real-time economic indicators for the EU27 countries and the UK. Following Barbaglia et al. [5] and Consoli et al. [14], we built fine-grained aspect-based sentiment indicators about a number of topics of interest. The sentiment indicators are (i) aspect-based, meaning that they are computed only about the specific topic of interest, and (ii) fine-grained, that is, they are bound in [-1, +1]. We build indicators using as topic-of-interest the same keywords extracted from the World Bank Ontology as presented in Section 2.1: this way we aim to obtain indicators comparable to the GDELT ones, with the advantage of having a longer time series. The sentiment indicators are computed for the EU27 countries by applying a filter on a direct mention of each country in the newspaper article in the analysis.

We report two different time-series for each topic and country in analysis. The first one is the *sentiment*, which is obtained by taking the daily average of the sentiment scores of all the sentences that talk about that specific topic in that country. Lower-frequency aggregations are computed by averaging at a monthly or quarterly frequency. The second one is the *volume*, which is the number of sentences that talk about that specific topic-country. While the latter provides an idea of the popularity of a certain topic in a country, the former explores whether the news deals with that subject with a favourable or pessimistic tone.

# 3 BUSINESS INTELLIGENCE DASHBOARDS

To provide intuitive and user-friendly yet extensive access to the data analysed, an interactive dashboard has been developed, relying on the Microsoft Power BI infrastructure.[15] This choice has been made after identifying the key needs of the project: ease of access, streamlined data update process, and simplicity of operations in the hosting infrastructure. The first priority aims to provide the target users with an uncomplicated navigation experience, focusing especially on the delivery of a simplified and intuitive representation of the observed phenomena. The second priority responds to the potentially close-to-real-time nature of the data: as the project can develop into a tracking tool which accounts for frequent changes

[14] DNA platform: https://www.dowjones.com/dna/.

[15] Microsoft Power BI: https://powerbi.microsoft.com/.

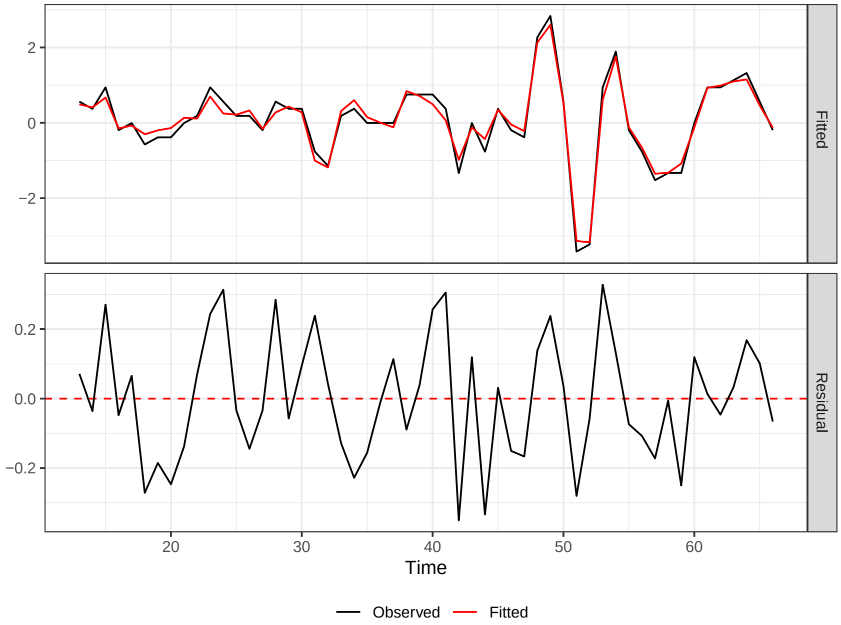


**Figure 2: In-sample fit and residuals from a sparse VAR to target unemployment rate in Italy.**

in sentiment variables and trends, this choice aims at reducing the effort and time needed to perform each update, with the advantage of reducing the gap between the data gathering and the publication. The third priority aims at harmoniously integrating the output of the project in the web outlets of our institution, thus reducing the need for ad-hoc development.

The published visualization consists of three separate dashboards, each focusing on one of the data sets analyzed, along with short introductory texts guiding the users in the exploration of the analysis. Each dashboard allows users to choose which data to visualize by filtering the country, topic and time-frame to be explored. The example illustrated in Figure 1 shows the information computed for the DNA data (i.e., news sentiment and volume of extracted news) around the *Unemployment* news topic between May 2005 and June 2021 for *Italy*. If no country filter is applied to the dashboard, it provides a comparative overview of the sentiment (or search volumes) across European countries, along with its EU27 average over time. The dashboard, although still at an early stage, can be publicly accessed by visiting: https://knowledge4policy.ec.europa.eu/composite-indicators/socioeconomic-tracker_en.

## 4 APPLICATION OF THE METHODOLOGY FOR THE FORECASTING OF UNEMPLOYMENT RATES

As an empirical exercise, we consider the case of forecasting the *unemployment rate* in France, Germany, Italy and Spain using additional regressors from the proposed dashboard. Timely and reliable forecasts of the unemployment figures play a relevant role in planning policies in support to the most vulnerable [12]. Given the delay and infrequent publication of official figures from statistical agencies, the importance of reliable unconventional indicators is even more prominent in times of high uncertainty. If an unexpected shock hits the labour markets, policymakers are in need of timely estimates of the magnitude of the (potential) slowdown to design an adequate policy response. The COVID-19 pandemic has emphasized the importance that timely indicators can play at times of unanticipated slowdowns of economic activities.

We obtain monthly harmonized unemployment rate figures from Eurostat.[16] We take the one and twelve months differences from the unemployment rate series in order to remove any trend or seasonal components. As it concerns the independent variables, we look for unemployment indicators in the proposed dashboard and select the features where there is an explicit mention of the target variable. In particular, we select (i) the article tone of the unemployment measure from the GDELT data set, and (ii) the unemployment sentiment indicators based on DNA newspaper data. We impose stationarity on the data by removing any seasonal component and by automatic differentiating until each feature passes a univariate unit-root test.

The final data set that we consider is averaged at the monthly frequency and it ranges from February 2015 to June 2021. Given that we have a short time series relative to the number of parameters to be estimated, we choose as a forecasting tool the sparse Vector AutoRegressive (VAR) model with a lag-based hierarchical penalty by Wilms et al. [44]. The VAR is an effective tool when forecasting multiple time series simultaneously, as it is in our application where we jointly model unemployment rates in four countries. The hierarchical penalty induces sparsity in the estimation of the autoregressive parameters and imposes structure on the selection of the number of lags. The result is a model whose estimation is feasible also in the presence of high-dimensional data, that is, when the time series

[16] Harmonized unemployment rate: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ei_lmhr_m&lang=en.

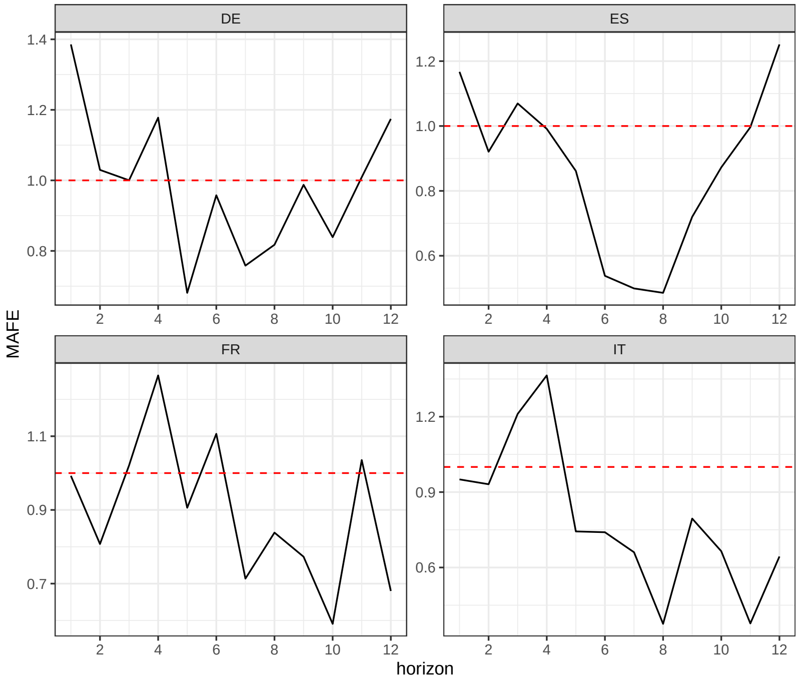


**Figure 3: MAFE by country relative to model without unconventional data. Values below 1 indicate a better performance than the benchmark model.**

length is short relative the number of number of unknowns. We tune the parameters with time series cross-validation and take 12 as the maximum number of lags. We consider two models, namely (i) a sparse VAR without additional regressors, and (ii) a sparse VAR with exogenous variables (VAR-X) where we include the unconventional series as additional regressors. The former is a nested model of the latter, thus providing us an appropriate benchmark to evaluate the added value of the proposed unconventional indicators.

| Horizon | DE | ES | FR | IT |
|---|---|---|---|---|
| 1 | 1.39 | 1.17 | 0.99 | 0.95 |
| 2 | 1.03 | 0.92 | 0.81 | 0.93 |
| 3 | 1.00 | 1.07 | 1.02 | 1.21 |
| 4 | 1.18 | 0.99 | 1.27 | 1.36 |
| 5 | 0.68 | 0.86 | 0.91 | 0.74 |
| 6 | 0.96 | 0.54 | 1.11 | 0.74 |
| 7 | 0.76 | 0.50 | 0.71 | 0.66 |
| 8 | 0.82 | 0.49 | 0.84 | 0.38 |
| 9 | 0.99 | 0.72 | 0.77 | 0.79 |
| 10 | 0.84 | 0.87 | 0.59 | 0.66 |
| 11 | 1.01 | 1.00 | 1.04 | 0.38 |
| 12 | 1.17 | 1.25 | 0.68 | 0.64 |

**Table 1: MSFE by country and forecast horizon relative to the model without unconventional data. Values below 1 indicate a better performance than the benchmark model.**

Figure 2 reports the in-sample fit and residuals estimated by the sparse VAR-X model with the unconventional indicators as exogenous regressors. From a qualitative view point, we observe that model achieves a high degree of in-sample fit and the residuals seem well-behaved. Similar conclusions can be derived for the analogous graph from the nested benchmark, which we do not report for brevity. In order to quantitatively assess the added-value of the proposed unconventional indicators in terms of in-sample fit, we compare the adjusted-$R^2$ of the two competing models. The sparse VAR-X with unconventional data provides an overall 13% gain in terms of adjusted-$R^2$ with respect to the benchmark, thus suggesting the proposed indicators to be useful measures to explain dynamics in unemployment rate.

Are the proposed unconventional indicators useful at predicting future values of the unemployment rate? To answer this question, we run a pseudo out-of-sample forecasting exercise. We consider an expanding window setting with out-of-sample dates starting in January 2019 until the end of the sample. For each monthly window, we estimate the sparse VAR-X with unconventional data and the nested benchmark, and produce a set of direct forecasts starting from 1 to 12 months ahead horizons. Table 1 reports reduction in median squared forecast error (MSFE) by forecast horizon and country attained by the sparse VAR-X model relative to the nested benchmark. Values below one indicate that the model with the proposed indicators attains a more accurate forecast than the model without. For each country, the proposed sparse VAR-X model delivers more accurate predictions in the majority of the forecast horizons, thus indicating the additional predictive content of the proposed measures with respect to unemployment rate dynamics. We observe that there is some heterogeneity in the performance across horizons: for instance, in all countries the proposed unconventional measures do not seem to convey useful information when forecasting 3 and 4 months ahead, while there is convincing evidence of their added-value when considering horizons longer than 5 months. The forecast gains associated to the use of the proposed measures can be large, reaching a maximum of 62% (i.e., a relative MSFE of 0.38) in Italy when forecasting at 8 and 11 months ahead horizons. Similar dynamics can be observed in Figure 3, where we report the evolution of the median absolute forecast error (MAFE) across forecast horizons, thus confirming the robustness of our conclusions.

## 5 CONCLUSIONS AND OUTLOOK

In this contribution we have presented our work-in-progress related to the development of a methodology for building alternative economic and social indicators from various unconventional data sets, including GDELT, Google Search, and DNA. The currently on-going project aims at providing intuitive and user-friendly access to the data analysed by using interactive BI dashboards, as well as producing improved nowcasting and forecasting methods to analyse various socio-economic measures for countries in the EU. We have reported some preliminary results on the application of this methodology for predicting the unemployment rate in Germany, Spain, France, and Italy. Overall, although preliminary, the obtained results look promising. This use case reveals indeed initial good performance of the methodology, suggesting the validity of the approach. Using the information extracted from these unconventional data sets within an opportunely trained VAR model, we have been able to achieve improved forecasting results. We believe that these new measures are able to capture and predict changes in economy and society especially in periods of turmoil. In the future we are considering expanding the set of unconventional data by including other interesting data sets with potential high impact, like, for instance, Twitter and Facebook data.

## REFERENCES

[1] D. Aaronson, S. A. Brave, R. Butters, D. W. Sacks, and B. Seo. 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. https://www.chicagofed.org/~/media/publications/working-papers/2020/wp2020-10-pdf.pdf

[2] S. B. Aruoba, F. X. Diebold, and C. Scotti. 2009. Real-Time Measurement of Business Conditions. *Journal of Business & Economic Statistics* 27, 4 (2009), 417–427.

[3] S. R. Baker and A. Fradkin. 2017. The impact of unemployment insurance on job search: Evidence from Google search data. *Review of Economics and Statistics* 99, 5 (2017), 756–768.

[4] L. Barbaglia, S. Consoli, and S. Manzan. 2021. Forecasting GDP in Europe with textual data. *Available at SSRN* 3898680 (2021), 1–38.

[5] L. Barbaglia, S. Consoli, and S. Manzan. 2022. Forecasting with Economic News. *Journal of Business & Economic Statistics* (in press) (2022), 1–12. https://doi.org/10.1080/07350015.2022.2060988

[6] L. Barbaglia, S. Consoli, S. Manzan, D. Reforgiato Recupero, M. Saisana, and L. Tiozzo Pezzoli. 2021. Data Science Technologies in Economics and Finance: A Gentle Walk-In. In *Data Science for Economics and Finance: Methodologies and Applications.* Springer Nature, Switzerland AG, 1–17.

[7] L. Barbaglia, L. Frattarolo, L. Onorante, F. Pericoli, M. Ratto, and L. Tiozzo Pezzoli. 2022. Testing Big Data in a Big Crisis: Nowcasting under COVID-19. *Working paper available at SSRN* 4066479 (2022), 38 pages.

[8] D. Borup, D. E. Rapach, and E. C. M. Schütte. 2021. Now-and backcasting initial claims with high-dimensional daily internet search-volume data. *CREATES Research Papers* 2021-02 (2021), 1–52.

[9] J. Bousquet, I. Agache, J. M. Anto, K. C. Bergmann, C. Bachert, I. Annesi-Maesano, P. J. Bousquet, G. D'Amato, P. Demoly, G. De Vries, et al. 2017. Google Trends terms reporting rhinitis and related topics differ in European countries. *Allergy* 72, 8 (2017), 1261–1266.

[10] A. Brodeur, A. E. Clark, S. Flèche, and N. Powdthavee. 2021. COVID-19, Lockdowns and Well-Being: Evidence from Google Trends. *Journal of Public Economics* 193 (2021), 104346.

[11] P. Brunori and G. Resce. 2020. *Searching for the peak Google Trends and the Covid-19 outbreak in Italy.* Technical Report. IRIS - Università degli Studi del Molise, Italy. https://ssrn.com/abstract=3569909

[12] G. Caperna, M. Colagrossi, A. Geraci, and G. Mazzarella. 2022. A babel of websearches: Googling unemployment during the pandemic. *Labour Economics* 74 (2022), 102097.

[13] H. Choi and H. Varian. 2012. Predicting the present with Google Trends. *Economic record* 88 (2012), 2–9.

[14] S. Consoli, S. Barbaglia, and S. Manzan. 2022. Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. *Knowledge-Based Systems* 247 (2022), 108781. https://doi.org/10.1016/j.knosys.2022.108781

[15] S. Consoli, L.T. Pezzoli, and E. Tosetti. 2021. Emotions in macroeconomic news and their impact on the European bond market. *Journal of International Money and Finance* 118 (2021), 102472.

[16] S. Consoli, L. Tiozzo Pezzoli, and E. Tosetti. 2022. Neural forecasting of the Italian sovereign bond market with economic news. *Journal of the Royal Statistical Society. Series A: Statistics in Society* (in press) (2022), 1–28.

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

[18] F. D'Amuri and J. Marcucci. 2017. The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting* 33, 4 (2017), 801–816.

[19] T. Fetzer, L. Hensel, J. Hermle, and C. Roth. 2020. Coronavirus perceptions and economic anxiety. *Review of Economics and Statistics* 103, 5 (2020), 1–36.

[20] D. Giannone, L. Reichlin, and D. Small. 2008. Nowcasting: The Real-Time Informational Content of Macroeconomic Data. *Journal of Monetary Economics* 55, 4 (2008), 665–676.

[21] P. Goldsmith-Pinkham and A. Sojourner. 2020. *Predicting Initial Unemployment Insurance Claims Using Google Trends.* Technical Report. Yale School of Management. https://paulgp.github.io/GoogleTrendsUINowcast/google_trends_UI.html

[22] J. W. Goodell. 2020. COVID-19 and finance: Agendas for future research. *Finance Research Letters* 35 (2020), 101512.

[23] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning.* MIT Press, US.

[24] C. Gormley and Z. Tong. 2015. *Elasticsearch: The definitive guide.* O' Reilly Media, United States.

[25] T. B. Götz and T. A. Knetsch. 2019. Google data in bridge equation models for German GDP. *International Journal of Forecasting* 35, 1 (2019), 45–66.

[26] A. Hamid and M. Heiden. 2015. Forecasting volatility with empirical similarity and Google Trends. *Journal of Economic Behavior & Organization* 117 (2015), 62–81.

[27] G. Koop and L. Onorante. 2019. Macroeconomic Nowcasting Using Google Probabilities. *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part A (Advances in Econometrics)* 40 (2019), 17–40.

[28] M.J. Kusner, Y. Sun, N.I. Kolkin, and K.Q. Weinberger. 2015. From word embeddings to document distances. In *32nd International Conference on Machine Learning (ICML'15)*, Vol. 2. ACM, United States, 957–966.

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

[30] H. Kwak and J. An. 2014. *A First Look at Global News Coverage of Disasters by Using the GDELT Dataset.* Springer International Publishing, Cham, 300–308.

[31] W. D. Larson and T. M. Sinclair. 2021. Nowcasting unemployment insurance claims in the time of COVID-19. *International Journal of Forecasting* 38, 2 (2021), 635–647.

[32] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444.

[33] K. Leetaru and P. A. Schrodt. 2013. *GDELT: Global Data on Events, Location and Tone.* Technical Report. KOF Working Papers, 1979-2012.

[34] T. Marwala. 2013. *Economic modeling using Artificial Intelligence methods.* Springer, Switzerland.

[35] V. Marx. 2013. The Big Challenges of Big Data. *Nature* 498 (2013), 255–260.

[36] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS 2013).* ACM, United States, 3111–3119.

[37] J. Pennington, R. Socher, and C.D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference.* ACL, United States, 1532–1543.

[38] N. Shah, D. Willick, and V. Mago. 2022. A framework for social media data analytics using Elasticsearch and Kibana. *Wireless Networks* 28 (2022), 1179–1187.

[39] B. Siliverstovs and D. S. Wochner. 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 (2018), 1–23.

[40] M. Taddy. 2019. *Business Data Science: Combining Machine Learning and Economics to optimize, automate, and accelerate business decisions.* McGraw-Hill, United States.

[41] Alberti V, Caperna G, Colagrossi M, Geraci A, Mazzarella G, Panella F, and Saisana M. 2021. *Tracking EU Citizens? Interest in EC Priorities Using Online Search Data - The European Green Deal.* Publications Office of the European Union, Luxembourg (Luxembourg). https://doi.org/10.2760/18216(online)

[42] S. Vosen and T. Schmidt. 2011. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting* 30, 6 (2011), 565–578.

[43] S. Vosen and T. Schmidt. 2012. A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters* 19, 7 (2012), 683–687.

[44] I. Wilms, S. Basu, J. Bien, and D. S. Matteson. 2021. Sparse identification and estimation of large-scale vector autoregressive moving averages. *J. Amer. Statist. Assoc.* (in press) (2021), 1–12. https://doi.org/10.1080/01621459.2021.1942013

[45] D. Zhang, M. Hu, and Q. Ji. 2020. Financial markets under the global pandemic of COVID-19. *Finance Research Letters* 36 (2020), 101528.

[46] S. Zheng, J. Wu, M. E. Kahn, and Y. Deng. 2012. The nascent market for "green" real estate in Beijing. *European Economic Review* 56, 5 (2012), 974–984.