NOVA IMS Information Management School

# MDSAA

Master Degree Program in

**Data Science and Advanced Analytics**

## Beyond Econometrics: Using Google Trends and Social Media Data to Forecast Unemployment

OECD analysis of accuracy gains and robustness of predictions

Pedro Sancho Vivas de Castro

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

[this page should not be included in the digital version. Its purpose is only for the printed version]

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# BEYOND ECONOMETRICS: USING GOOGLE TRENDS AND SOCIAL MEDIA TO FORECAST UNEMPLOYMENT

by

Pedro Sancho Vivas de Castro

Dissertation presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Business Analytics

**Supervisor:** Bruno Damásio

February 2023

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

*Lisbon, 26 of February 2023*

# ABSTRACT

Google Trends has been used for less than two decades in academia to forecast outcomes, using various techniques. While most research has focused on developed countries, there are clear information gaps that have not been fully addressed. Previous studies in this field indicate that non-linear algorithms with feature set selection while using a large set of queries can yield better results across more countries. However, it is unlikely that these methods will be widely and rapidly adopted given the skills required.

Therefore, the objective of this research is to explore whether the abundance of digital data sources, specifically Google searches, can aid agents as institutions and policy makers in their modeling efforts. The aim is to fill the gap in analysis for less influential countries and explore whether the use of Google searches data can be extended to multiple countries using a simple and agile methodology based on a widely used statistics-based modeling approach (ARIMAX). For this use we selected unemployment rate as the variable of interest.

However, our findings show that only 30% of countries had promising results using Google-augmented ARIMAs. Thus, more computationally intensive empirical strategies would be needed to extract more predictive power out of Google queries information pool for unemployment rate modelling.



# KEYWORDS

Google Trends; Unemployment; Time series forecasting; Information gaps

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

**OECD** Organisation for Economic Co-operation and Development

**IMF** International Monetary Fund

**EAP** Economically active population

# 1. INTRODUCTION

Unemployment rate has been a critical variable analyzed by several entities in every society. It is easy to capture the reasoning of that, since it relates to the state of the productive forces, the ability to export and several other aspects of our society today. Despite digitalization of many services in our day and age, as well as mechanization of many productive processes, labor in its general sense is very intertwined with the level of well-being of one's society. So much so, that the Federal Reserve (Fed) tracks down this macroeconomic variable to a point that, among other functions, the institution manages US money supply to achieve maximum employment. These other attention points are: the protection of price stability as well as moderate long-term interest rates, which could be conflicting in some cases.

Given that, especially when any economy experiences an exogenous shock, it is quite difficult for agents to forecast unemployment using econometrics-based framework, such as ARIMAX, and in general professional agents are equipped to use those tools and have been using them for the past decades. Additionally, some agents do incorporate exogenous variables in their modelling of macroeconomic variables, and this is commonly expressed by the X term in the acronym ARIMAX.

On the other hand, the influx of social media and the level of availability of its data emerged recently in academia as a true possibility of incrementing many specifications and frameworks. General, the new use of this data has led to novel research in various fields such as Economics, like in D'Amuri and Marcucci (2017), and even Epidemiology, like Ginsberg et al (2019). Additionally, there are interesting use cases in different fields, seen in Mellon (2014) for Political Science, more specifically on how Google data can be used to represent general public's views instead of expensive usual surveys, but also present in Prei et al. (2013) for Finance, specifically focused quantifying trading behaviors using Google queries. Moreover, when looking at the use of social media data for macroeconomic variables, there are some promising findings in a growing landscape of research. On that note, there are the first highlights done by Askitas and Zimmerman (2009), using four simple keywords from Google Trends data to help track German unemployment rate. More recently, Borup and Schütte (2020), showed that for the United States case, both at country and state-level, non-linear algorithms, like Random Forest, using a large panel of keywords from Google Trends, can augment forecasts for unemployment to a greater extent than a large panel of algorithms already implemented by this research field.

Askitas and Zimmerman (2009) inspired by the work on predicting flu done by Ginsberg et al (2009), tried to augment unemployment forecast using social media decoupling the "social media" component of job search into 4 components. These 4 components aimed to capture different dynamics that can drive the unemployment rate in any economy: (I) when people are trying to contact the unemployment office; (II) when people are trying to check how is the state of unemployment; (III) when people are trying to contact human resources consultants; (IV) when people are searching for the websites where job listings are posted.

There are various reasons why other researchers selected specific keywords to capture job market dynamics. For instance, there are country specific dynamics that could possibly discard the need of the same components used by Askitas and Zimmerman over different ones, when analyzing other countries with different circumstances. Apart from that, the proposed approach by these researchers of using an error-correction model specification, proposed by Engle and Granger (1987), needs

assumptions such as cointegration of the variables analyzed and therefore is a more restrictive approach than the one selected for the research presented in this document (ARIMAX specification).

More recently, the works of Elliot et al (2013 and 2015) are also influential to next contributions on this research field due to the possibility of integrating multiple weak predictors with a new method based on complete subset regressions. This approach can even provide a frontier to analyze a possible trade-off between bias and variance of forecast errors of the different combinations of forecast from the subset. Furthermore, for US case, Elliot shows that both univariate regressions like ARIMA and dynamic factor approach, such as specification with principal components from a larger feature set, produce less accurate point forecasts than this novel approach.

Borup and Schütte (2020) conducted a thorough analysis of complex specifications, using large feature sets (information exhaustive) to improve accuracy in predicting unemployment. They analyzed five different algorithms and found that Random Forests provide the best results for various forecast horizons at both country and state levels.

On the other hand, more agile approaches such as D'Amuri and Marcucci (2017) use simple keywords from Google Trends data to track US unemployment rate and achieved more accurate forecasts than US' Survey of Professional Forecasters. Therefore, the two researchers aim to produce specifications that produce consistently more accurate forecasts without increasing its complexity such as the novel approaches produced by Elliot et al (2013 and 2015) as well as the ones tested by Borup and Schütte (2020) later.

The above suggests that utilizing pre-COVID Google Trends data can enhance the accuracy of US unemployment rate forecast. As several studies using different methods and evaluation data have found similar results.

Despite that, we have few results produced for European countries, and they are restricted to less recent research for Germany, France, and Spain. Moreover, very little research has been conducted with emerging and middle-income relevant countries at the global context. The highlight of this was Narita and Yin (2021) contribution to narrow information gaps for low-income countries forecasts using Google data.

Another example of how crucial it is to advance this research field for a different range of countries, is the involvement of institutions such as the International Monetary Fund (IMF) (Narita and Yin, 2021), to fund and allocate human resources in these regions. Non-developed countries typically face a data scarcity challenge in informing public policy. This limitation hinders the evaluation of financial viability of projects. The possibility of augmenting the modelling of macroeconomic variables that very commonly are assumptions to these evaluation efforts, holds great promise in addressing this challenge. The ultimate aim of this endeavor is, therefore, to explore whether the abundance of new digital data sources, specifically Google searches, can aid policy makers, agents, and institutions in their modeling efforts.

This research aims to fill the gap in analysis for less influential countries and explore whether the use of Google searches data can be extended to multiple countries using a simple and agile methodology based on a widely used statistics-based modeling approach (ARIMAX).

The approach follows the benchmark result of D'Amuri and Marcucci (2017), but using up-to-date data, using a robust data collection procedure for Google searches data, seen in Medeiros and Pires (2020), for a widely known and influential panel of countries named Organisation for Economic Co-operation and Development (OECD).

The assessment of the predictability gains that this method could provide will be addressed in two ways, one related to the accuracy gains of the forecast with the exogenous component, the other one checking if the augmented model produces statistically significant different forecasts than the benchmark ARIMA. Apart from that, the out-of-sample data gives the capacity to do the assessment procedure for a broad range of forecast horizons (from one month to 12 months, inclusive), with a good number of datapoints, with at least 20 instances, for 12-month horizon.

Furthermore, this research can demonstrate the potential of Google searches as an incremental source of data and explore whether other data sources, such as Twitter data, can also yield information gains, like the promising results for inflation expectation modelling done by Angelico et al (2022).

Besides, a simple modeling approach is utilized to ensure the research can be widely adopted by professionals and the public sector. Since the most promising results so far were related to computationally heavy approaches, from data selection to the algorithm definition as well, which could deter its adoption. This ARIMAX approach with easily accessible data could provide an easy to implement tool for companies and government bodies to conduct macroeconomic forecasts essential to their daily activities.

The remainder of this document is organized as follows: Section 2 will present the literature review; Section 3 will address the methodology used; Section 4 will discuss the results; Section 5 will display the conclusions from the study; Section 6 will address the limitations from the conducted study as well as some recommendations for future works; Section 7 will present the References used on this study.

# 2. LITERATURE REVIEW

Section 2 will provide context of the research field and the current state of work published in academic journals. There will also be analysis of the main contributions to the literature. Finally, it will outline the aims and proposed contributions of this research to the field.

## 2.1. RESEARCH CONTEXT

The relevance of accurate forecasts for instructing state officials and relevant agents in firms has been linked with statistics and mathematics at least since early 20th century, as discussed in Tsay (2000).

At first, Academic research gathered enough knowledge acumen that these recommendations could be done through prolific researchers placed at very influential institutions, like Jan Tinbergen, one of the fathers of Econometrics, working at Dutch Statistical Office (1929-1945), at the League of Nations (1936-1938) and at Netherlands Bureau for Economic Policy Analysis (1945-1955). Nowadays, university-instructed forecasters populate relevant positions in every financial institution, government body and firms from different sizes in our economies, mainly due to the widespread use of the influential work in Econometrics on the 20th and 21st century.

The adoption of these tools for predictions was very intertwined with the digitalization of business models in the economy as well as the growth of new capabilities to use and gather information through computers and the internet. Additionally, various issues in our daily lives can be solved through using personal computers internet, and generally, companies store footprints of these operations. Naturally, this could be used for a variety of things: from understanding behavior patterns, associating different phenomena for anticipating events, or even aggregating information for better predictions.

This is such a strong trend that, just alike government body officials paid attention at the dawn of 20th century, major international institutions, such as the IMF, have funded more analysis on the relevance of forecasting using this digital footprint. Specially for the case of developing countries, this research field can be one of the instruments to better define economic policy in tight national budgets, and in that way help the desired catch-up of living standards to those of developed countries. Caveats may arise from lower adoption rates of computers, and even electrification of these less affluent countries. Therefore, it becomes relevant to investigate what circumstances of developed countries are associated with gains on using population's digital footprint on forecasts.

## 2.2. STATE OF THE RESEARCH FIELD

Researchers have been investigating the possible augmentation of forecasts with multiple things, and naturally with the increasing adoption of internet across the globe, this research field is expected to gravitate towards exploring this informational pool.

Despite the recency of most studies, there is instrumental body of research already produced that has gradually led to an adoption outside academia. Given this nuance, in this specific research the goal is to further investigate findings for a broad of range of countries through an extensive period of

assessment. Therefore, a broad overview of the current state of the most recent and relevant studies on this specific field is needed to ground our perspective over this research important aspects, namely methodology, modelling decision, assessment criteria and other things. This overview will be focused on contributions made to forecast macroeconomic variables and similar interesting variables, given that they highly influence business decisions.

The most recent and very promising finding is the contribution by Angelico et al (2022) which tried to augment inflation expectations forecasts. They used tweets from Twitter related to a certain number of keywords and topics. Twitter is a social media platform that allows people to produce small commentaries about any topic, people can link their commentaries (tweets) to other commentaries from themselves and others. Given this short description, it is easy to imagine that such a platform, if highly adopted by citizens could be a forum to discuss, vent, protest about several aspects on people's life and that seems to be the case for Italy and inflation expectations. This type of investigation has been used for different variables, such as to predict quarterly earnings surprises, when earnings releases differ from equity analyst's earnings forecasts, as well as for future stock returns from stock opinions in Twitter, as seen in Chen et al. (2014). Other fields of research have also resorted to this data, for example Golovchenko et al. (2020) paper about misinformation campaigns in Twitter, or studies conducted by Hinduva Watch on the impact of the spread of hate speech, specifically on members of minority and marginalized communities in India given their faith[1].

For the latter mentioned studies, that were conducted in regions with less economic resources, the low cost of accessing Twitter data was particularly relevant. However, it is important to note that due to governance changes that took place in 2022 at Twitter company what previously was seamless, with regards to *tweets* data retrieval, now has some additional challenges[2]. While initially there were no costs to researchers for data retrievals up to 10 million tweets per month and 50 requests per 15 minutes to their application, now it costs at least, 100 dollars monthly for a *basic access*, which could be particularly harmful for researchers based on developing countries. Secondly, this *basic access* full capabilities were not disclosed yet, and it is not expected to have a quota as large as the previous one for academic research users.

Angelico et al. (2022) used novel data mining techniques, specifically topic modeling on Twitter, developed by Alvarez-Melis and Saveski (2016), to focus on keywords related to inflation. They found that incorporating Twitter usage information pool led to enhancements in inflation forecast accuracy. Apart from that, this Twitter-enhanced prediction has only been verified for Italy's case. Therefore, despite OECD countries having higher adoption to internet, and consequently, to this social media platform, it could still be the case that it would not hold the same explanatory power as a highly developed like Italy.

Even though the social media platform provides an interface called TweetAPI that gives us easier access to this information pool, the necessary step to produce components given by Twitter's usage through

[1] As seen in February 2023, a broad discussion regarding this paragraph topic: https://techcrunch.com/2023/02/14/twitters-restrictive-api-may-leave-researchers-out-in-the-cold/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuYmluZy5jb20v&guce_referrer_sig=AQAAAN2oZl6i1GaJDpm4nB9RAztCwiOVx5jlIYoKPersTkXOrYaAf7xHRXCBWG4Dn9-i6oSwUg2AxQ0c1iwERJWgefk_bn6DadZdO-eCVytDKuPBSJ5aD1I75pKJum2HjFglv2D_0BDcsXcDX9wwmVAdRCmiZ2j17z0gHh4Al36IELMU

[2] As seen in February 2023, in Twitter website: https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

topic modelling is very computationally daunting (both time- and resource-wise). This concern has been addressed when Alvarez-Melis and Saveski (2016) proposed a novel topic modelling technique and investigated through multiple prisms the most used frameworks for this task: Latent Dirichlet Allocation, from Blei et al. (2003) and Author Topic Model (ATM), from Rosen-Zvi et al. (2005).

Before Alvarez-Melis and Saveski (2016), different proposals, like Hong and Davison (2010) and Mehtora et al. (2013), which revolved around aggregating related *tweets* (*pooling*), were produced to address the specificities of topic modelling in Twitter. Despite interesting results, new issues came to light, like higher training times and the produced aggregations commonly had heterogenous topics, which is clearly undesired result. In 2016, Alvarez-Melis and Saveski proposed a more appropriate technique to *microblogging* platforms such as Twitter, given the fact that their information usually has more noise and is composed of very short commentaries. The issue revolves around the fact the two most used constructs are designed for extracting per-document statistics, and for the outputs to be consistent they cannot as short as *tweets*.

Despite very interesting results while using *tweets* and Twitter information pool, the body of literature is not as extensive as the one when considering the Google information pool, which will be presented more thoroughly in the next section. Some findings use US Google usage to predict several different variables as well as other researchers have used some countries in EU Google usage with promising results.

In retrospect, the first contributions to the field of *Google-augmented macroeconometrics* were mostly using simple keyword selection from Google Trends, which is a website that allows you to interact with how people are searching for any term for a specific location over a certain period. Alongside this keyword selection pattern, these researchers generally used simple algorithms highly used on the job market and academia as mentioned previously, like ARIMA, as in Askitas and Zimmerman (2009); D'Amuri and Marcucci (2017) and others.

In contrast, new contributions are generally going in the direction of computationally heavy modelling decisions as well as large information pools to extract as much sensible data as possible from Google Trends. This extensive modelling decision is generally preferred when researchers select large number of keywords, so that the complex interaction of these keywords and the variable of interest can be correctly captured, as seen in the contributions of Bai and Ng (2008), Elliot et al (2013 and 2015), Borup and Schütte (2020) and others.

However, we have observed that there is very little literate on this field using data from developing and low-income countries. The most notable one was produced by Narita and Yin (2021) addressing information gaps and how nowcasting could be enhanced by aggregating Google Trends data specifically for these low-income countries.

Other findings using regression models and allowing different treatments over the Google queries data were analyzed, with special attention to Damásio and Nicolau (2014 and 2020). On that note, Damásio et al. (2018) narrowed information gaps with regards to the understanding of its economies in the *periphery* of European Union (EU), and shaped some decisions taken in this study and brought light to a more diverse and inclusive panel of countries so that dynamics that are partly understood for the most developed countries could be explored for other populations. Also Martins and Damásio's work about post-Great Financial Crisis reform, commonly considered as objective and ideal treatments, and

their fit to Portugal's case at that time, was analyzed. In summary, it is important to note how this type of research can shed light to some strong preconceived notions that could be, despite previous perceived consensus, very detrimental to the welfare of any population. And for this reason it is important that the field of econometrics provides tools that enlarge our understanding of relevant macroeconomic variables, such as unemployment, and how they interact with each other.

This theme of narrowing information gaps that do impact societies' welfare can also go into using data mining techniques to explore patterns in public activities, such as public procurement, as seen in Lyra et al. (2021), Curado et al. (2021) and Lyra et al. (2022). These works shed light on behavior patterns when placing bids, conducts that indicate possible frauds, collusions and eventually, in the worst-case scenario, corruption from public officials as well. However, some of these data-driven methods, for instance, network analysis, can be difficult to implement in large and diverse territories, such as the 184 municipalities analyzed for the State of Ceará in Brazil on Lyra et al. (2021). Despite that, their works shows how state-level officials should use either through interacting with researchers, or even through upskilling alongside universities so that these new body of research can instruct better procurement activity.

## 2.3. ANALYSIS OVER HISTORIC CONTRIBUTIONS TO THIS RESEARCH FIELD

Ginsberg et al. (2009) produced an interesting finding that flu epidemics can be predicted using these searches data and that, which kickstarted this research field of associating keyword searches with macroeconomic variables, initially unemployment rates, as this research will focus on as well. They did this with the release of Google Trends first version in 2008, initially named Google Insights, with available data from user searches starting in 2004. This epidemiology study influenced several researchers in Econometrics and Machine Learning to demonstrate through different fashions the power of this new source data for augmenting predictions.

The first important contribution to the body of research came from Askitas and Zimmerman (2009), partly influenced by Ginsberg earlier work, addressing the issue of tapping into one source of information that the internet provided us for augmenting predictions on unemployment rate, which is Google searches in this case. Their exercise extended to monthly unemployment rate for Germany, with special attention to how this augmentation could help us predict economic behavior under changing circumstances, like the Great Financial Crisis, for instance.

These researchers constructed different models using up to four components that ideally captured different dynamics that can possibly explain the unemployment rate: (I) when people are trying to contact the unemployment office; (II) when people are trying to check how is the state of unemployment; (III) when people are trying to contact human resources consultants; (IV) when people are searching for the websites where job listings are posted. Despite the interesting results, the reasoning behind using four keywords and choosing specifically those terms revolves around the simplicity argument, a parsimonious approach to the modeling decision something that could also be viewed on choosing an error-correction model specification algorithm, based on Engle and Granger seminal work in 1987.

Many contributions would stem from this first finding, but we highlight research conducted in 2012 by Google's Chief Economist Hal Varian and another Google employee, Hyunyoung Choi, a scholar from UC Berkeley. They disclosed several exercises with accuracy gains for predictions over product sales using their brand name as search keywords from Google for simple constructs, such as fixed-effects and univariate models. With the goal of familiarizing econometricians and statisticians with Google Trends data, foster a new myriad of forecasting exercises and showcase some possibilities of use cases with special attention to what nowadays is called *nowcasting*.

In a discussion paper for the Bank of Israel, Suhoy (2010) proposed an interesting two-step procedure to be able to assess multiple keywords from Google Searches without using Principal Component Analysis -based frameworks, like those proposed by Schmidt and Vosen (2009) and Kholodilin et al. (2011), for increasing predictions of private consumption. Here the nowcasting exercise comes up again, which is discovering real-time macroeconomic variables, given the fact that their true value disclosure may be subject to lags. The process revolves initially around some regression analysis between the variable of interest and the possible candidates, filtering out non-positive slope predictors for the last available month prior to the nowcast. Secondly, it proceeds to assess mutual correlation between the keyword candidates by trying all combinations in multivariate regression analysis. Lastly, through Akaike Information Criteria the best subset of keywords is selected. Despite the simplicity of the proposal, the number of regressions increases tremendously with the number of keywords being analyzed, which could a downside of the proposal.

Bai and Ng (2008) proposed a distinct approach to data-rich environments like Google Searches, emphasizing the preselection of informative predictors for the variable of interest. In that study, the researchers investigate the use of factor forecasting method for inflation modelling, while checking if non-linearity of their principal components aggregates predictive power and how the predictors may change given the forecast horizon. Apart from verifying that these two hypotheses are not falsified, all forecast horizons benefit from fewer estimated factors that are informative and as expected that some variables, like interest rates, have systematic predictive power for inflation at all forecast horizons.

Another possibility to use many web-queries is to follow the Bagging approach for inflation forecasting as stated by Inoue and Killian (2008), that uses multivariate forecasting with hard-thresholding statistically significant predictors at 1% level, bootstrapping the dataset for each multivariate forecast exercise and averaging the forecasts. Further analysis can be done with regards to autocorrelation avoidance of the forecasts averaged. However, the researchers highlight those other methods, that could be considered simpler can achieve similar prediction gains, as well as saying that the body of literature developed some frameworks that generate similar augmentations to forecasts.

One last approach on the use of large-dimensional datasets of predictors is the one present in Elliot et al. (2015), which the main aspect is its applicability when the number of predictors is higher than the sample size. Using Monte Carlo simulations, given many predictors, some of them with low predictive power, researchers were able to present interesting bias-variance trade-off with their approach, named Complete Subset Regression. It seems to be the case that this setup may be able to extract predictive power from cross-section of the predictors, unlike other simpler methods.

D'Amuri and Marcucci (2017) verify that simple Google keyword searches intensity may help predict US unemployment rate. They perform this by comparing important economic indicators, like economic surveys for consumers and employers, or the mostly used initial claims statistic for monthly

predictions. Additionally, quarterly forecasts using Google Searches data is compared with survey of professional forecasters and standard benchmark models for quarterly unemployment forecasts. Lastly, it seems that in most exercises Google-augmented models slightly outperform their competitors, although this advantage seems to be most prevalent in more volatile periods.

Borup and Schütte (2020) showed through some exercises that using a large panel of keywords with a nonlinear model can greatly exploit the information pool from Google Searches and led them to outperform large set of benchmarks, including previous constructs using Google Searches. Despite narrowing their exercise to US data, these researchers drilled down their analysis to industry-level unemployment, and state-level indicator. Their best construct revolves around the Random Forest algorithm, which is an ensemble learning method that makes several decision trees to perform bootstrapping aggregation and at the same time randomly select the features used in training the decision trees at each iteration. Additionally, they assessed individually their large number of keywords and show that some keywords could be almost as good as using the large panel of keywords, but this could only be defined ex-post, and therefore their proposition treats for this nuance by always extracting some information from very strong predictors within a large number of uninformative keywords. Lastly, the use of variable selection together with a high-dimensional information pool cannot be used as the sole reason why Google Searches beats macroeconomic and survey-based models, since it seems that the heterogeneity of information that comes from different search terms included at different time is what explains most of the accuracy gains.

While using multiple frequency data, Fondeur and Karamé (2013) showed that simple models that included Google Searches queries had better performance over those that do use that data, and that was specially the case for youth unemployment data. In addition, they constructed an approach that used unobserved components, with a modified Kalman filter, to use the full weekly series data provided by Google. Also, despite having interesting results for the general working force, in this study, the most promising ones were achieved for the population of 15- to 24-years old, which could be accounted for the fact that young people use more internet in general, including job searches.

Vicente et al (2015) produced an ARIMAX construct enriched with demand and supply components, where an employment confidence indicator and Google Searches data, respectively, where used as proxies for this analysis. The Spanish unemployment, differently than most research studies until that date, present a high rate and a more volatile behavior, apart from that, little research was produced using European countries data, and even less to countries that with some frequency endure periods of sharp increases in unemployment. Interestingly, despite being a developed country, the forecast gains generated by the proposed model also poses an interesting question: Does that in some way indicate that these same simple constructs, that use Google Searches information pool, can provide similar gains to a variety of more volatile countries, like low- and medium-income countries?

## 2.4. PURPOSE OF THE STUDY

Given the fact that this research field has a good body of literature that supports using Google Searches data as augmentation for modelling macroeconomic data, specifically unemployment rate, in specifics cases like: (i) developed countries; (ii) younger cohorts; (iii) volatile periods, like the Great Financial Crisis. Despite that, it is clear that little research was produced for low- and medium-income countries,

like in Narita and Yin (2021), and given the fact that the globe had a pandemic that imposed a very volatile period for most economies, it would also be interesting to reassess in a more general manner developed countries, both that previously were the focus of the research literature and those that were not.

The goal is to have a broad assessment of the explanatory power of Google Searches data in forecasting unemployment rate for various countries. The simple modelling decision derived from the previous notion that is important to test something that would be easily adopted by agents and institutions. By choosing OECD countries as its focus points, it will be very interesting to address its countries given its diversity, although countries in OECD tend to be at least medium-income countries, with most developed countries being present on this group.

Moreover, the assessment would allow us to see if there is a general pattern linked to internet usage, level of adoption of technologies, state research and development budget and, in general, how different factors could help us understand why some countries easily could benefit from adopting Google data to better predict their unemployment rate. Additionally, by using an extended dataset, and having different informative periods, like the COVID pandemic and the Great Financial Crisis, as well as specific downturns every country faced, is promising to further help the robustness of the results.

Finally, it will be assessed if there is any predictability gain on using components generated by a simple keyword, *jobs* in each language, from Google Searches data. Also, it will also be checked if the forecast that incorporates this information pool from Google queries and the standard ARIMA for modelling unemployment rate are statistically different.

## 2.5. INNOVATIVE INSIGHTS

This research can bring attention to an important source of information that is derived from our digital footprint today. The amount of our daily life that is intertwined with the internet is very high currently, despite that very little has had any impact in our modelling exercises and how forecasts are given. Therefore, given that most of the current research using Google queries data, which is the focus of our study, is narrowed to developed countries, in either North America or Europe, it is very important that this research field could be developed in a broader way.

Such analysis will reinforce the attention brought to this theme, as well as solidify our current perspective that there is a lot of predictive power in this information pool. Despite that, it is important to note that the selected approach will mostly give strength to a parsimonious wide adoption that could be a way to spread this use of internet data to augment forecasts. Since the modelling choice is based on widely adopted paradigm, SARIMAX, and presents a robust way to select and use simple keywords from Google Searches data, this will be clear sign on a possible low-hanging fruit to model macroeconomic indicators.

However, it could be the case that the results could further clarify that in order to use these types of data into macroeconomic modelling exercises, it is needed to assess and trim a very large panel of information pool. Additionally, use a more complex paradigm, like the ones brought up by machine learning practitioners, such as nonlinear ones, like Random Forests. Lastly, it could even be the case

that even larger dataset should be assessed, like Twitter data, with its multiple discussion topics by various commentators, just like Angelico et al. (2022) did it for Bank of Italy inflation expectation modelling exercise.

# 3. METHODOLOGY

In this section, the focus will be to address the methodology decisions of this research study. First, some diagnosis over why we selected a certain dataset over others and briefly describe data circumstances that needed to be overcome. Apart from, some description over general aspects of the selected countries data for the unemployment forecasting exercise, there will be some explanation over how the data gathering procedure was done, with special attention to how Google queries data was retrieved in order the generate consistent representation of the phenomena that is being used.

Second, some analysis over the algorithm decision as well as the modelling procedure conducted will be showed as to inform the rationale behind them. Additionally, how the model specification was performed for each country, how the forecast assessment was dealt with, will also be disclosed at this part.

Third, there will be some explanation over why other societal indicators, like internet usage, or research and development government annual expenditures, were used to bring light to forecast results. Moreover, what is the reasoning behind using this data as a possible guidance to illuminate the different results that may arise from using Google data for augmenting unemployment rate forecasts.

## 3.1. DATA SELECTION

Our panel of countries that will be assessed whether Google data can augment their unemployment rate forecast is the OECD which is a very influential group of 38 countries. Some of their characteristics are having high standards of living, high- and mid-level income per habitant and a large amount of global population. Also, the relevance of this can be seen by the fact that all of the G7 is present in the selected dataset for this study, a group of *the seven most advanced economies* as considered by the IMF.

Additionally, what is most interesting to our study: having some diversity of their economies, such that most of the countries that are already present in this research field are present, high-income countries, like France and the US, as well as medium-income countries like Colombia and Mexico. Thus, these standard of living divisions can be conducted into an even finer granularity, since OECD despite its similarities will have Colombia, Costa Rica and Mexico that can be defined as medium-income countries, but historically had less than half of the average GDP per capita of the OECD group, which shows how much income inequality can be present even when analyzing only these group of well-off countries.

Given this fact, there is an opportunity to broadly address why the OECD was created in the first place. At first, the OECD was created to adapt a previous institution that, after the 2nd World War, aimed to manage US and Canadian economic aid to reconstruct Europe, in the context of the Marshall Plan. However, as time went by this institution naturally gravitated towards helping European countries on economic policy matter, including being a participant on the talks that eventually led to the European Union. As time went by, this notion of helping economies prepare economic reforms, from countries that were opening their economies like Poland, Hungary and former Czechoslovakia in 1990, to

countries that were trying to further integrate themselves into global value chains in the 2000s like South Korea and Mexico. Therefore, these panel of groups as time went by naturally encompassed more diversity due to the adaptation of the group core goal, something that can be seen by the dispersion of OECD countries throughout the globe, as seen in Figure 3.1 (Founding members are displayed in dark blue, and other member are displayed in light blue).

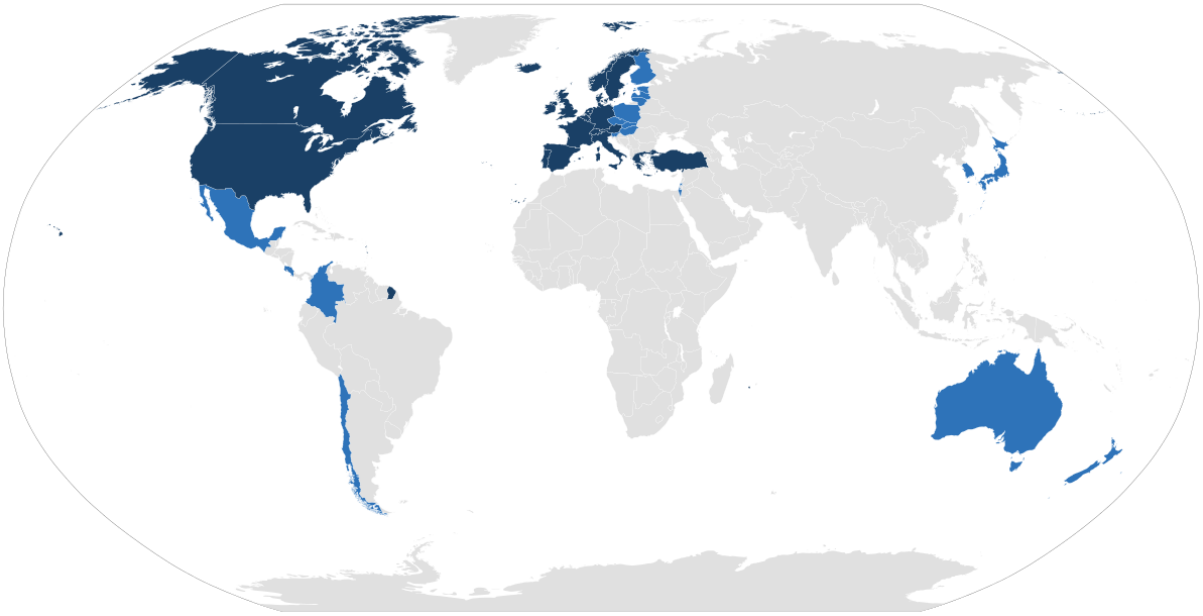

Figure 1 - OECD countries – Source: OECD (2022)

Moreover, there is some illustration to be done on these different aspects of these countries that could impact the augmentation results while using Google queries for unemployment rate modelling. Thus, some overview of job market differences, each countries internet adoption levels but also data regarding unemployment aids or even population percentage that did not finish secondary school. Somehow, it is reasonable to assume that these indicators will help elucidate the results of the augmentation for each country, since it makes sense that they affect transposition of the job market dynamics to the internet, more specifically for this study, Google searches patterns.

It is important to note that these socio-economic data over OECD countries was gathered from OECD Data website in 2022, and the graphs display latest available data for each country and each indicator. In that case, most countries are displaying year-end figures for 2021 and others latest available data from 2020.

Given different legislation that these countries have, especially those that do affect job market dynamics, such as labor law and imposed minimum wage is reasonable to suspect this affect how workers interact with the perspective and actual employment status changes. To highlight these differences, at Figure 2, there is some overview of unemployment benefits given in every OECD country. Additionally, such variance will likely affect the interaction between these job market dynamics and its respective digital footprint in Google data.

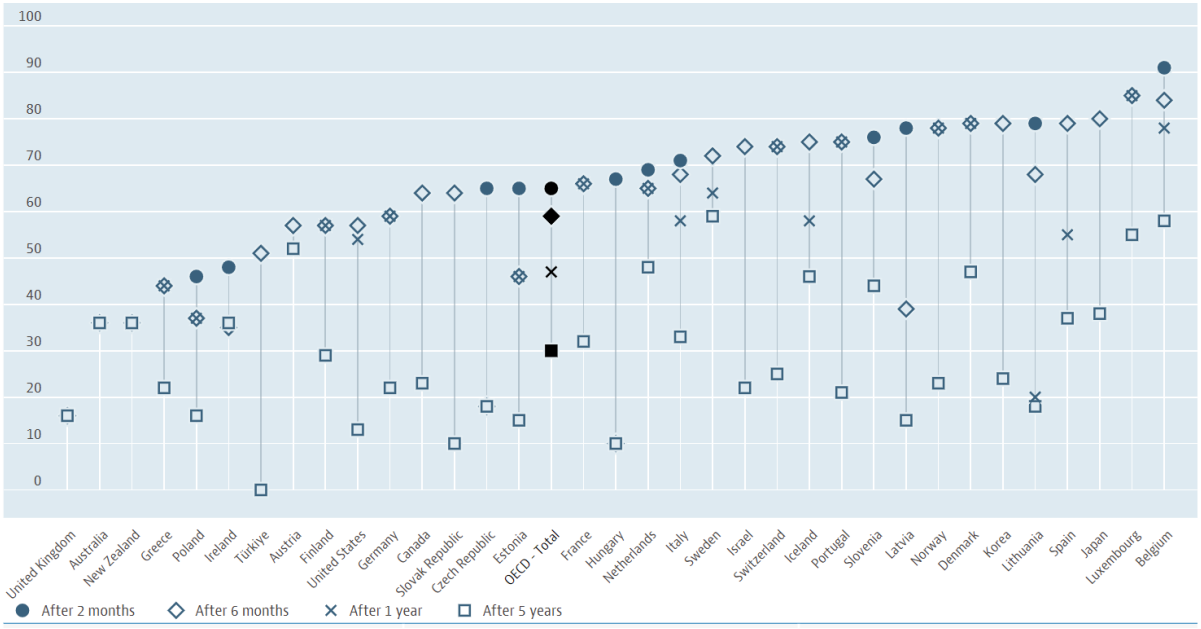


Figure 2 - Benefits in unemployment, % of previous in-work income – Source: OECD (2022)

Other aspect that may affect the adoption level of job searches being conducted in Google, and therefore presented in Google queries information pool, is productive forces differences, particularly labor productivity discrepancies within OECD population. On this note, as initially addressed by Mincer (1958) and Schultz (1960), human capital has direct relationship to income distribution and economic growth for any country, and naturally education levels for any population will have direct association to the impact of this difference in productive forces. Also, there is micro evidence for R&D expenditures for firms through government funds having impacts on R&D, in general, as seen in Paredes et al. (2022). Therefore, Figure 3 will aim to highlight profound differences for economic active populations within OECD, given the percentage of people that did not complete high school.

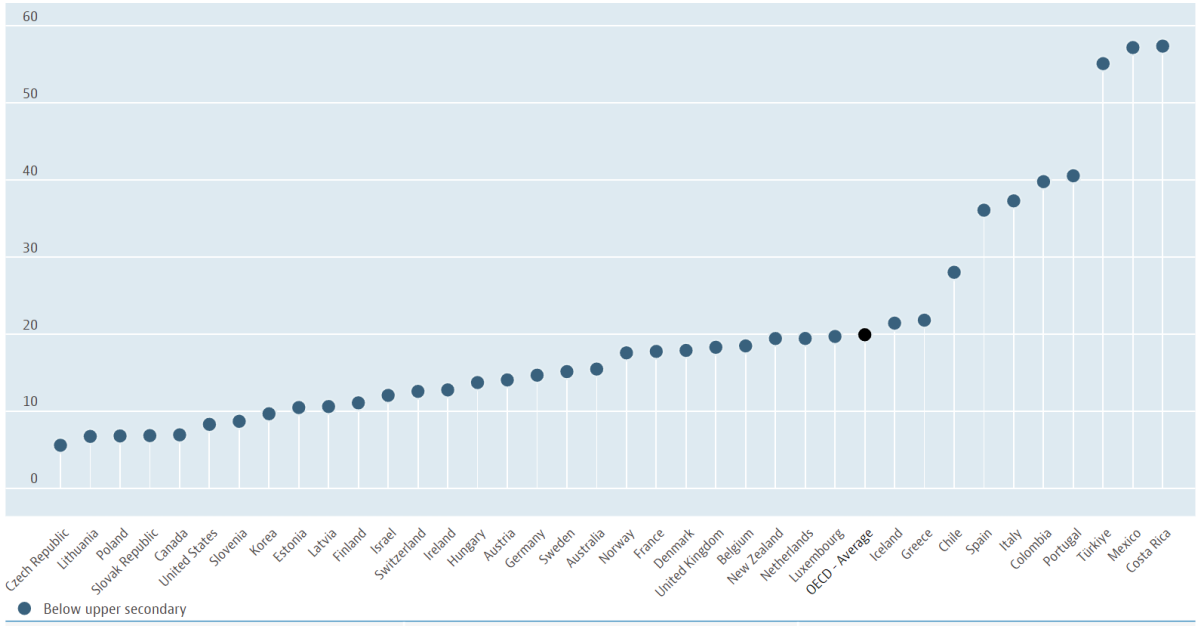


Figure 3 - % of economically active population with below upper secondary education level - Source: OECD (2022)

On another note, the production structure of these countries can be very different as well. OECD can encompass both *the club of most industrialized countries* (G7) as well as countries that, despite integrated into industrial global value chains, are secondary players in both technology adoption and relevance. Additionally, government bodies spend very different level of their annual budgets into research and development as seen in Figure 4.

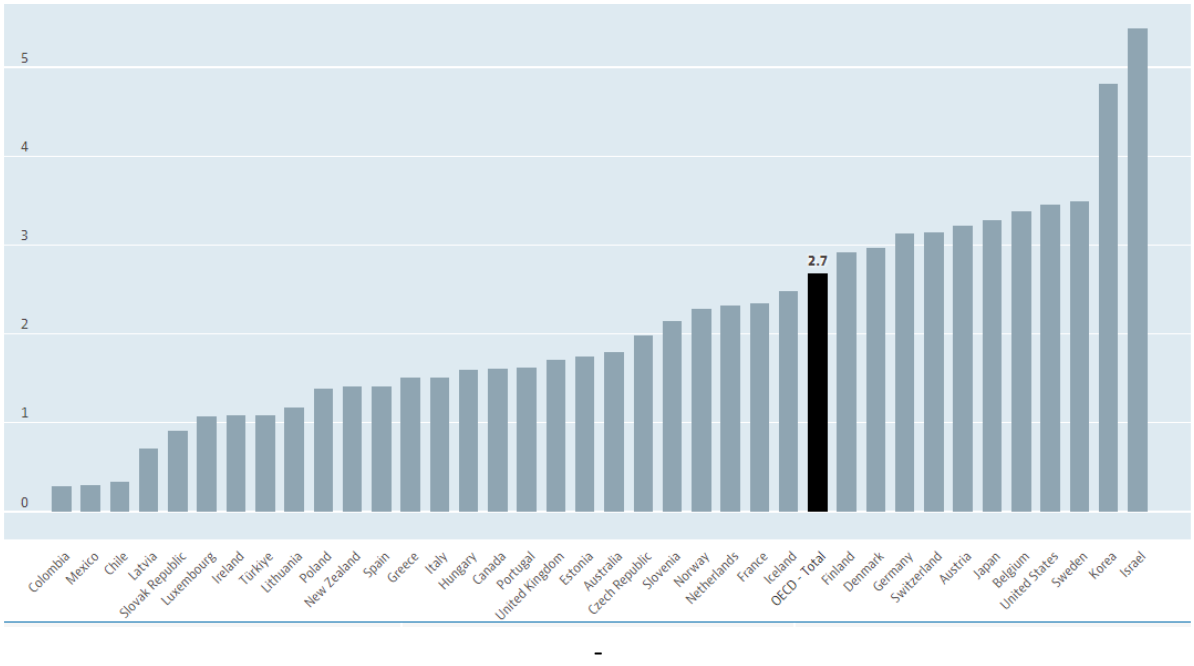


Figure 4 - Gross domestic spending on R&D, % of GDP - Source: OECD (2021)

Therefore, these are some aspects brought to highlight the motivation around selecting this large panel of countries, with a lot of diversity in different aspects of any economy, so that the study shows some depth over the results. Some of these characteristics will be used somehow to understand different results and significance of the augmentation proposal that this research will focus on.

Given that the area of augmenting macroeconomic variables, mostly highlights the most advanced economies of the world it was very natural that if the hypothesis tested wants to be extrapolated for as much circumstances as the internet adoption and population's behavior regarding the job market allows it, there was clear need to address a more diverse panel of countries. Apart from that, the large panel of countries help us to avoid cherry-picking results so that a clear narrative over to why certain countries seems to have greater augmentation and why others did not display the same results. Lastly, it could be the case that unclear results for this exercise will show that simple Google queries are not really a low-hanging fruit for modelling macroeconomic variables and more data intensive and computationally heavy constructs are needed to extract the explanatory power present in this new information pool.

## 3.2. DATA DEPICTION

In this part, the focus will be on addressing in a general manner the main characteristics of the dataset used for this study. Firstly, the dataset can be divided into two parts: (i) 29 time series unemployment

rate data regarding all countries that are members of OECD; (ii) 29 time series *jobs* keyword searches for the same group of countries. Naturally, the quality of the dataset will be different for each time series, especially with regards to OECD data, and when does each country started disclosing their unemployment rate, with a similar methodology as their peers. Also, Google searches data can be very volatile for the same keyword in different retrieval trials, as seen in Cebrián and Domenech (2022), and that will need some procedure to address in a sensible manner, as suggested by Medeiros and Pires (2020).

With respects to unemployment rate data, most countries will have available data starting in August 2003, with exception of four countries: (i) Israel/IL: January 2012; (ii) Costa Rica/CR: August 2010; (iii) Colombia/CO: January 2007; (iv) Turkey/TR: January 2005. Given that, and the fact that Google Searches data starts in January 2004, countries will be assessed, if possible, from January 2004 until March 2021, month of the retrieval of Google data. For these cases, where less data is available, the analysis will not be conducted since the goal of the study is to have a broad panel of countries with the rationale of having the same empirical strategy with the same amount of data.

Therefore, the rest of countries from OECD will have their analysis with this most extensive time period described previously. These countries are Australia, Austria, Belgium, Canada, Czech Republic, Denmark, Estonia, Finland, France, Great Britain, Germany, Greece, Hungary, Iceland, Ireland, Lithuania, Luxembourg, Mexico, Netherlands, Norway, Italy, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden, Switzerland, and United States.

On another note, Google queries data have some characteristics that should be described, so that the understanding of why the data gathering procedure of this data must follow a certain method can be perceived. The data is disclosed in a weekly datapoints for keywords that have at least a certain amount of search tries (the value is not disclosed by Google) and reflect the prevalence of these keyword query relative to the total number of searches for that week. Lastly, the data depiction is scaled to the maximum value over the period displayed. However, it is important note that this Google data could not be representative of internet job searches, because different countries in different points in time will show different adoption levels of job searches in internet through Google, as seen in Foundeur and Karamé (2013), that in France Google had a stable 90% market share for a number of years, but the United States of America had the search engine market changing over time as Google had 60% of market share in 2006 and 70% in 2010.

Apart from those aspects, as Cebrián and Domenech (2022) thoroughly analyzed it, inadvertently using Google searches data could lead analysts, researchers, public policy makers and other professionals to claim strong predictability gains without the expected robustness in their statements. On that note, using a keyword that is defined by an entity, which is an abstraction that refers to single semantic unit, like a concept, such as jobs, is important to avoid problems such as polysemic terms (when different concepts are expressed using a single term). Apart from that, there are other issues related to small prevalence of this search term over the analyzed horizon that leads to potential completeness and validity issues, with regards to data quality, but nevertheless, both these problems can be dealt with during data preparation before using this data source. Nevertheless, the reasonable level of dissimilarity between data retrievals (using the Pearson correlation coefficient), when performed using the same period, for the same region and the same keyword in different days leads us to proceed with using this data source with extra care, as suggested by the findings of Medeiros and Pires (2020).

Lastly, some data over some characteristics that could lead up to more or less adoption of Google searches as a tool for job market decisions and its overall dynamics will be used for some exploratory data analysis. The goal of this analysis is to highlight common characteristics of certain countries that could help us understand the different in augmentation gains by using this internet queries component when trying to forecast unemployment rate. The following indicators were selected: (i) percentage of economically active population that has below upper secondary education level; (ii) mobile broadband subscriptions per capita; (iii) gross domestic research and development expenditure as a percentage of each country's GDP; (iv) % of households with internet access; (v) immediate unemployment benefits as % of in-work income. Later, there will be some assessment on why these indicators can influence the amount of activities societies perform online and therefore how this could be linked to different results in the augmentation proposal of this study.

## 3.3. DATA GATHERING PROCEDURES

With regards to OECD countries unemployment data, it was easily gathered using simple search terms in OECD official website, specifically placed at their Data tab. Also, countries data regarding research and development, internet usage and other important societies characteristics were also retrieved in a similar fashion in OECD website.

As far as Google searches data gathering was conducted, a Python script was created, that accessed Google Trends API and retrieved sequentially for each OECD country, for previously described period, the keyword *jobs* in its respective native language. Apart from that, a Bash script was created so that an automated procedure could be set up in order to execute this Python Script for several days, in order to retrieve multiple samples from the Google searches information pool related to this keyword for every country.

On that note, this multiple retrievals of the same keyword for every country, for the same period, was critical to ensure that we could reconstruct consistent estimates of the true value of this keyword for every country in this given period, as shown in Medeiros and Pires (2020) work. This necessity relates to the fact that in order to have something that truly relates to the Data Generation Process of the google query prevalence some data retrievals must be performed since at every retrieval Google provides a sample of the true dataset, so that it can be easily accessible for each user that requests its data. Lastly, the use of this process automation logic for the data gathering had an intention of efficiency gains in this part of the study as well as to be able to produce consistent estimates of the real data.

## 3.4. EMPIRICAL STRATEGY

### 3.4.1. Algorithm selection

Given the wide adoption of econometric modelling in public policy analysis and in forecasting practices among different industries, as highlighted in previous parts of this study, there was clear preference on assessing the augmentation gains when using an algorithm that different agents could easily incorporate these new findings. This notion of knowledge acumen among sectors on how to

incorporate new components to statistics-based forecasting practices led us greatly narrow down our algorithms choices.

Furthermore, among the possible candidates the lack of need to have some structural modelling, and therefore hypothesis, over the variables treated among this research study led us to choose a general algorithm based on auto-regressive integrated moving average models, with its respective seasonal and exogenous terms, as commonly named SARIMAX.

This seasonal $\mathrm{ARIMA}(p,d,q)(P,D,Q)_m$ process can be described by:

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t$$

Where $\Phi(z)$ and $\Theta(z)$ are polynomials of orders *P* and *Q* respectively, each containing no roots inside the unit circle. If $c \neq 0$, there is an implied polynomial of order *d + D* in the forecast function. The main task in automatic ARIMA forecasting is selecting an appropriate model order, that is the values *p*, *q*, *P*, *Q*, *D*, *d*, which will be discussed later. Additionally, *m* is the seasonal frequency of the seasonal component.

### 3.4.2. Modelling approach

Given that, in order to explore most of the information pool of a given keyword from Google queries, some summary statistics, like mean, standard deviation, skewness, of the Google keyword time series was calculated so that these could eventually amplify the forecasting gains of Google searches data when compared to not using it. These process of creating new time series out of an original one, is commonly called feature engineering, and it is mostly used in machine learning approaches, nevertheless, the goal was to narrow down the augmentation proposition to simple keywords and a simple algorithm, but without neglecting different aspects that can be extracted from the original time series with some time series manipulation.

These manipulations somehow are not that different than the benchmark implementation of SARIMAX, without the exogenous term, in the sense that it uses the original time series lagged values, as well as moving averages, which uses a linear combination of error terms. Thus, the selected specification is defined based on how the data can be best fit to these modelling approach.

Therefore, the two competing models for each country, a linear model using an exhaustive approach to adding exogenous components versus a linear model that uses the *best fit* model using the same algorithm without exogenous components.

## 3.5. Model specification and forecast procedure

### 3.5.1. Model specification

In order to implement these comparisons, the study needed to specify two models for each country, the Google data enriched one and the simple *best fit* ARIMA. This was done using an auto-ARIMA procedure in a stepwise algorithm manner, as proposed by Hyndman and Khandakar (2008), e.g. to

check some plausible combinations of a SARIMAX implementation, verify Akaike Information Criteria (AIC) to every combination of each term of this algorithm and return the specification that best fits the training data for each model. Besides that, in all specification's stationarity is verified using a unit root test, leading to accept only those that are strictly invertible.

More specifically, at first four possible models are calculated, given an $\mathrm{ARIMA}(p,d,q)(P,D,Q)_m$:

(i) $\mathrm{ARIMA}(2,d,2)(1,D,1)_{12}$
(ii) $\mathrm{ARIMA}(0,d,0)(0,D,0)_{12}$
(iii) $\mathrm{ARIMA}(1,d,0)(1,D,0)_{12}$
(iv) $\mathrm{ARIMA}(0,d,1)(0,D,1)_{12}$

If $d + D \leq 1$, these models are fitted with $c \neq 0$. Otherwise, we set $c = 0$. Given these four models, the one with the smallest AIC value is selected for now.

Then, thirteen variations of *this current specification* are evaluated in the following manner: (i) Where one of *p*, *q*, *P* and *Q* is allowed to vary by ±1 from the *current specification*; (II) Where *p* and *q* both vary by ±1 from the *current specification*; (III) where *P* and *Q* both vary by ±1 from the *current specification*; (IV) where the constant *c* is included if the *current specification* has $c = 0$ or excluded if the current model has $c \neq 0$. Given that, whenever a model with lower AIC is found, it becomes the new "current" model, and the procedure is repeated. This process finishes when we cannot find a model close to the current model with lower AIC.

Additionally, there are several restrictions on the fitted specification to avoid problems with convergence or near unit-roots. These constraints are outlined below: (I) The values of *p* and *q* are not allowed to exceed specified upper bounds (with default values of 5 in each case); (II) The values of *P* and *Q* are not allowed to exceed specified upper bounds (with default values of 2 in each case); (III) We reject any specification which is "close" to non-invertible or non-causal. Specifically, we compute the roots of $\Phi(B^m)\phi(B)$ and $\Theta(B^m)\theta(B)$. If either have a root that is smaller than 1.001 in absolute value, the specification is rejected; (IV) If there are any errors arising in the non-linear optimization routine used for estimation, the specification is rejected. The rationale here is that any specification that is difficult to fit is probably not a good model for the data. The algorithm is guaranteed to return a valid specification because the specification space is finite and at least one of the starting specifications will be accepted (the specification with no AR or MA parameters). The selected specification is used to produce forecasts. Given this, the data used for verifying the best model specification would be 185 datapoints, more than 15 years of data.

### 3.5.2. Forecast procedure

For the forecast procedure, we define a simple cut-off criterion that would split the total amount of datapoints, and by doing so, would have most of the dataset to assess the best specification for each model and a small percentage of the dataset, so that we can assess how the augmented forecasts performed versus the standard SARIMA implementation. In this case, 85% of the datapoints were used for training the model, and 15% of the datapoints were used for testing data, e.g. unseen data that will be used for assessment and robustness checks of the proposed exogenous Google based components.

Additionally, we used the testing data in a rolling window fashion so that for the countries that there is data starting in 2004, there will be forecasts ranging for 31 one-month forward predictions to 20 12-month forward predictions. On that note, the same procedure will be conducted for those models that use Google data and the standard auto-regressive moving average models so that with these same point forecasts we can in for each country assess both accuracy gains, with respects to MAE and RMSE, and if these forecasts are statistically different. Given that, forecasts will be assessed from, at best, almost three years of data to approximately two years of data, on the worst scenario.

In order to conduct this robustness check of any augmentation that could be generated by Google based components to a standard econometrics model, it is going to be used a commonly used test to assess statistical difference in forecasts named Diebold-Mariano test. This test main motivation is for empirical applications, where usually there are competing models for generating forecasts, given a variable of interest. Given that, by having the actual values of the unseen data, e.g. the testing data, and two competing models forecasts, the test assess if these two forecasts have equal prediction power (for reference, check below equations that describe how test is conducted). Therefore, we are going to present for different significance levels (0.01, 0.05 and 0.10, respectively depicted alongside other metrics with ***, ** and *, respectively) if these forecasts can be stated as different alongside the metrics used to analyze prediction errors. Given the previous explanation of how the tested is going to be conducted, below are the steps performed for each country to do this hypothesis test:

- Actual values: $\{y_t : t = 1, \dots, T\}$
- Two forecasts: $\{\hat{y}_{1t} : t = 1, \dots, T\}$ and $\{\hat{y}_{2t} : t = 1, \dots, T\}$
- Forecast error: $e_{it} = \hat{y}_{it} - y_t \ , \ i = 1 \ and \ 2$
- Loss function: $g(e_{it}) = e_{it}^{\ 2}$
- Difference between forecasts: $d_t = g(e_{1t}) - g(e_{2t})$
- $Diebold - Mariano$ test: $H_0 : E(d_t) = 0 \quad and \ H_1 : E(d_t) \neq 0$

Different loss functions can be used based on what type of modelling exercise the forecaster is trying to perform, like an absolute error loss function or even *asymmetric* loss functions. However, there are certain conditions that this function must abide: (i) takes value 0 when no error is made; (ii) can only have positive values; (iii) increases as errors become larger. The selection of the square-error loss function in this study, besides being typically used, is that it has a direct association with a common metric used assess forecast accuracy, root mean squared-error, that will be used later for this type of forecast assessment. Additionally, one caveat that this loss function has, that could be seen a reason not to use it, is that it equally considers underpredictions or overpredictions of the target variable, nevertheless, this symmetric treatment for negative and positive forecast errors poses no harm in unemployment forecast for most use cases.

### 3.6. REASONING ON COUNTRY SPECIFIC SOCIETAL DATA SELECTION

There was a selection of five socio-economic variables that could be linked to greater adoption of internet job searches by the workforce. The data will later be used for a clustering exercise that will help us assess if these socio-economic variables could be associated with different accuracy gains made possible by using Google information in forecasting unemployment.

It is important to address why such data were selected in the first place: (i) % of GDP spent on R&D; (ii) Unemployment benefits; (iii) mobile internet access points; (iv) households with internet access; (v) % of economically active population with below upper-secondary level of education. The reasoning behind a general selection of these variables was that, in general, they can be associated with any population's internet usage, regardless of if this relation is about job searches or any other aspect of daily life.

However, each of these indicators could have different specific reasons, some are more immediate, like unemployment benefits and how they are linked to job search dynamics in general (including internet job searches), or even less educated workers and how they generally face digital inequality, as in Wilhelm (2004). Thus, workers that are generally low-skilled workers, due to their low education level (Holzer 1996), and that integrate a workforce in a country that has low investments on research and development[3], will expectedly have a greater difficulty on making a transition to using the internet and their different platforms for doing any activity, including job searching.

[3] See in Mark Warschauer's book published in 2003, named *Technology and Social Inclusion: Rethinking the Digital Divide*

# 4. RESULTS AND DISCUSSION

Section 4 will present how the forecast were assessed, which robustness check was used and how the accuracy gains were evaluated. Other than that, some cluster analysis over societal indicators from the panel of countries will be presented to further facilitate any discrepant results from the augmentation exercise for different countries. But lastly, the google searches data augmentation proposition will be analyzed in the final part of this section.

## 4.1. ASSESSMENT OF FORECASTS AND AUGMENTATION PROPOSITION

In this part of the study, the focus will be the forecast results for all countries, that was previously described in the methodology section. In that sense, firstly, we will analyze countries that seem to benefit from having Google data in their specification using simple keywords and a nimble and widely adopted modelling choice (SARIMAX). In order to do so, we are going to present mean absolute error percentage and root mean squared error decreases, which, if positive, corresponds to lower errors for the google-enriched model versus the standard SARIMA benchmark model. Apart from that, it will be provided through number superscripts the Diebold-Mariano results for each forecast period.

Furthermore, our analysis will be split into three groups of countries given their augmentation results: (i) the *best candidates;* (ii) the possible use cases; (iii) the poor cases. The first category results somehow endorse the use of Google information pool, since linear models with some simple robust data gathering procedure will have increased predictability by using these exogenous terms, over standard implementations. The second group results could be placed on a grey area, where despite accuracy gains, there a few forecast periods that past the robustness check, and therefore with a little bit more attention into the keyword selection and/or the algorithm choice would possibly be sufficient to find promising results as the first group. On the hand, the third group, which comprises most OECD countries displayed poor results, since in general, none of the forecast periods passed any significance level when testing it via Diebold-Mariano test, and some countries did not present much predictability gains as well.

With the regards to the first group, these group of countries displayed interesting results that adhere to the notion that very low effort endeavors into using Google searches data can be instrumental to find new predictability power for simple econometrics modelling widely used by economic agents, researchers and policy makers. Since the augmentation proposal revolves around the simplicity of modelling and use of new datasets, accuracy gains over various forecasts periods statistically different than the SARIMA competing model provides some evidence over wide adoption of this source data for unemployment modelling.

Concerning the second group, these countries show results that clearly depict some possibility of extracting better predictions for unemployment rate. On that note, our assessments made clear that a simple algorithm with just a single keyword, like jobs, will not be sufficient to widely produce statistically different forecasts that produce better estimates for this macroeconomic indicator. As it is going to be depicted for every country that fits this case, some forecast periods do pass our robustness check and present better forecasts than the benchmark specification, however, it is never the case for more than two periods out of twelve forecast periods possible. Additionally, as other results from this

research field shows, like in Borup and Schütte (2020)**,** other non-linear models and more broad keyword selection could clearly put these countries on generating robust predictions that consistently beat this competing model. Therefore, for these countries it seems that the information pools could be sufficient with computationally more expensive algorithms and/or using a wider range of search terms from Google queries data, but with this nimble augmentation proposal it is not quite enough.

Lastly, with respects to the third group, these large group of countries results gives some reality check over the reasonable amount of research that poses a lot of promising use cases for macroeconomic forecasting that not necessarily are transposable for wide adoption by agents across the globe. Despite OECD being somehow the club of countries, that has societies that live in either good or great standards of living, it holds some representability over the whole globe. Additionally, this panel of countries is more broad, diverse and a better representation than the countries analyzed in the past by the research field. Therefore, unfortunately, it seems that for most countries Google queries data will not be a *low-hanging fruit*, as expected from reading interesting results for United States, especially if it is considered implementation could easily be adopted by financial institutions, companies and other entities. However, different reasons could be behind this result, from inadequacy for certain countries of using jobs as a keyword, or less adoption of agents of job market search on the internet, specifically through Google search, to a more exhaustive methodology that could account for the predictive power being dispersed in a lot of keywords for certain countries, or just a few for others.

### 4.1.1. Augmented proposal: *best candidates*

The countries considered best candidates for providing evidence for using simple keyword selection from Google searches data and augmenting standard econometrics modelling with it are: Estonia (EE), Austria (AT), Latvia (LV), Finland (FI) and Spain (ES). These five countries share some similarities over some important societies' indicators, such as education level and internet access, as seen in Figures 5, 6 and 7.

In the clusters identified in the last part of Section 3, there are some confirmations of the similarities of these countries, given the fact that three out of the five countries in Cluster 0 are placed in this group. In bellow figures, these countries are Estonia (in purple), Austria (in red) and Finland (in blue). The other countries are Latvia (in orange) and Spain (in green), both presents in different clusters.

Lastly, the figures present statistics for year-end 2021 OECD data for mobile internet usage per 100 inhabitants, percentage of economic active population that holds bellow upper-secondary level education and unemployment benefits because these were the three characteristics out of five socio-economic indicators that were used for the clustering exercise on the panel of countries.

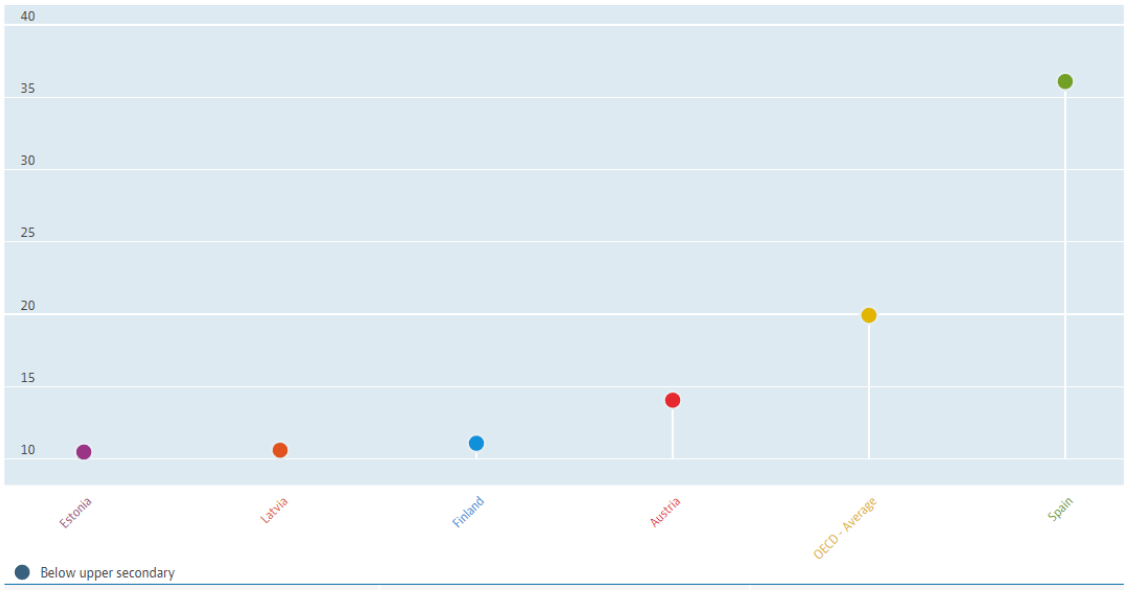


Figure 5 - % of economically active population with below upper-secondary level of education - Source: OECD (2022)

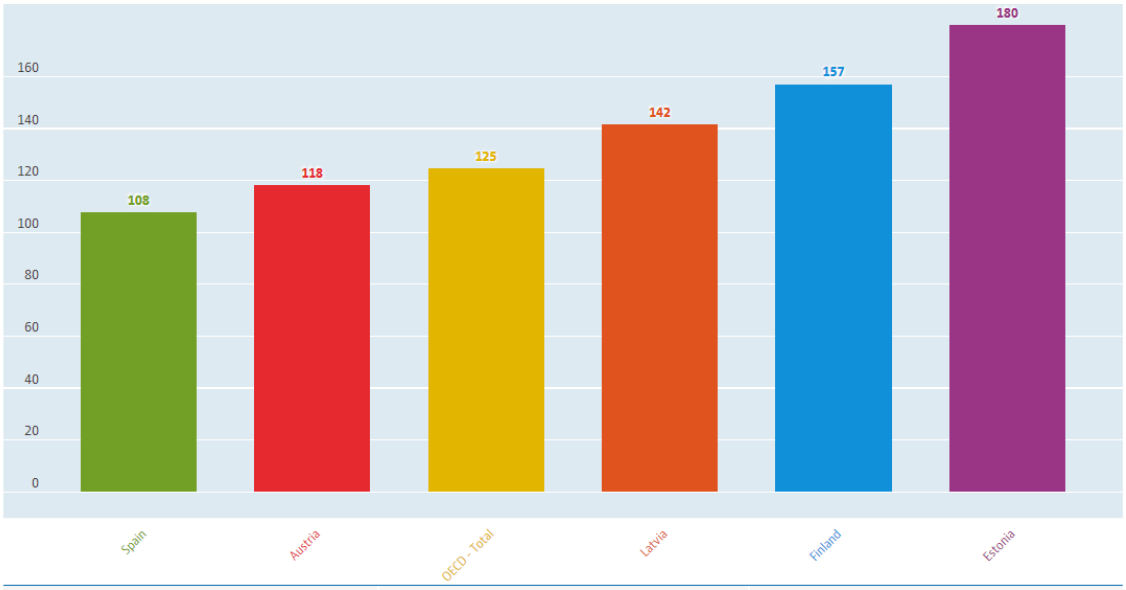


Figure 6 - Mobile internet access point per 100 habitants - Source: OECD (2022)

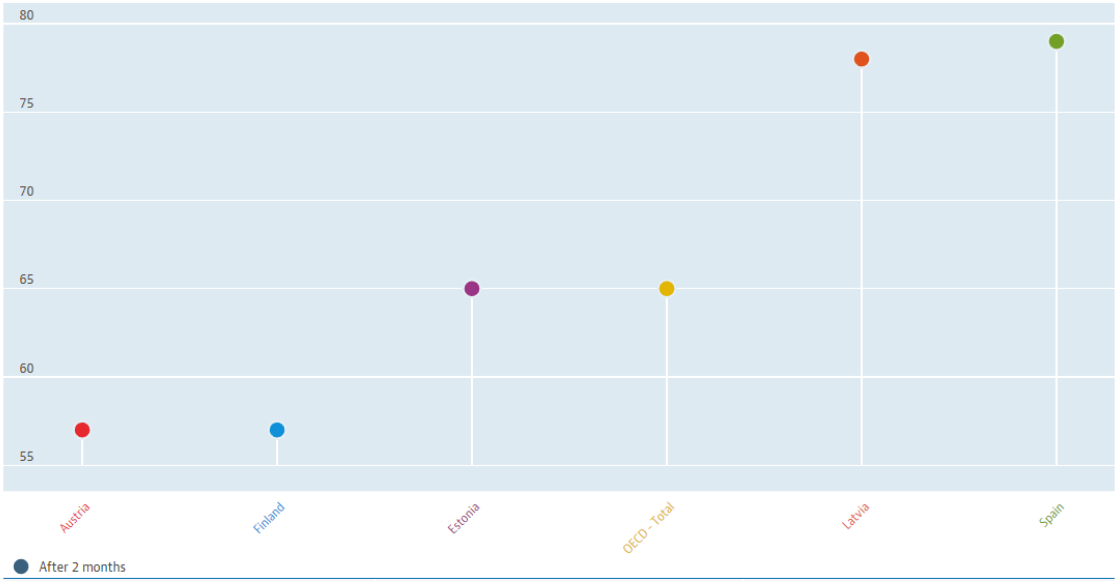


Figure 7 - Unemployment benefits as a % of previous in-work income 2 months after employment status change – Source: OECD (2022)

These indicators clearly show a socio-economic pattern for indicators that could play a role on job searches on the internet specially for Estonia, Finland and Austria. Thus, it makes some sense that these groups belong to a single cluster due to their similarities. However, it is important to note that, Latvia is closer than Austria to the centroid value of some parameters of this clusters giving us a hint why it displayed such results for our augmentation proposition. Lastly, Spain is one of the countries that was studied earlier by this research field, these previous findings, found in Vicente and Menéndez (2015), lead us to understand why it displayed interesting results, despite being clearly the least similar country as far as these indicators can point out.

### *Estonia*

Below, it will be presented the augmentation results for Estonia's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 1 - Estonia augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | -15.8 | -21.4 |
| 2 | -23.0 | -27.0 |
| 3 | -0.7 | -6.3 |
| 4 | 20.3* | 7.0* |
| 5 | 29.4*** | 15.0*** |
| 6 | 32.4*** | 22.0*** |
| 7 | 34.1*** | 27.5*** |
| 8 | 42.7*** | 32.3*** |
| 9 | 43.8*** | 38.1*** |
| 10 | 44.9*** | 41.0*** |
| 11 | 46.5*** | 41.4*** |
| 12 | 48.7*** | 43.6*** |

Estonia is a very interesting case, because it is one of the most digital countries in the world. Their internet adoption level goes as far as being able to file for internet citizenship, and solve various issues related to bureaucratic things, that in many countries it would be necessary to visit a notary or a public institution.

Given that, when you look at accuracy gains prism, from one-month ahead forecast to three-months ahead forecast Estonia results are worse predictions compared to the benchmark model. However, from four-months ahead forecast to 12-months ahead forecast there is strong gains in accuracy, specially in MAE, and these forecasts are different then the competing model, mostly in the lowest level of significance (8 forecast periods with 0.01).

### *Austria*

Below, it will be presented the augmentation results for Austria's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 2 - Austria augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 21.6 | 19.5 |
| 2 | 20.0 | 21.9 |
| 3 | 35.6* | 34.0* |
| 4 | 37.4** | 32.6** |
| 5 | 34.4** | 28.8** |
| 6 | 33.4** | 28.5** |
| 7 | 29.7** | 27.3** |
| 8 | 26.3*** | 27.3*** |
| 9 | 36.3*** | 32.6*** |
| 10 | 41.9*** | 35.9*** |
| 11 | 46.2*** | 39.0*** |
| 12 | 49.9*** | 42.5*** |

Austria augmentation analysis starts to reveal an interesting pattern, that will later be seen in other cases, that as the forecast period increases, e.g. the forecasts are given to increasingly more months forward, there seems to be a trend of increased accuracy for Google augmented forecasts. Additionally, it seems that the robustness of the predictions increases as well as Diebold-Mariano tests are rejected for lower significance levels as forecasts periods increase. Given these two associated characteristics for Austria's results, it seems that agents would highly benefit from using Google information pool to augment their unemployment forecast.

Our results present some promising grounds for further investigation of different approaches to using these data, like different keyword selection or even using the same dataset of exogenous terms from a single keyword with a non-linear model. However, despite these aforementioned trends, both root mean squared error (RMSE) and mean absolute error (MAE) present un unsteady increase as the forecast increases, as an example, refer to the MAE by the increase from three-month forward forecast (35.6%) to four-month forwards (37.4%), and then a steady decrease from that forecast to eight-month forward forecasts (26.3%), to get back to a steady increase until the last one (12-month forward forecasts, with 49.9%).

***Latvia***

Below, it will be presented the augmentation results for Latvia's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 3 - Latvia augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 1.2 | 15.2 |
| 2 | 11.9 | 21.2 |
| 3 | 17.7 | 18.9 |
| 4 | 15.8 | 19.3 |
| 5 | 20.5 | 22.2 |
| 6 | 20.5 | 25.8 |
| 7 | 26.3* | 28.7* |
| 8 | 32.3** | 32.8** |
| 9 | 37.9*** | 36.6*** |
| 10 | 41.5*** | 40.0*** |
| 11 | 45.6*** | 43.4*** |
| 12 | 48.7** | 46.6** |

Latvia results are interesting, since the augmentation accuracy gains increases as forecast period increases, despite this gain in predictability from one-month ahead forecast to six-months ahead forecasts the augmented model produce statistically indifferent forecasts compared to the standard model. However, from seven-month forward forecast until 12-month forward a similar pattern of increasing forecast accuracy appears when analyzing the difference between the two forecasts, it becomes increasingly less likely that these difference in accuracy is due to pure luck. On this note, seven-months ahead forecasts present 28.7% less root mean squared error with a significance level of 0.10, for eight-months ahead forecasts there was 32.8% less root mean squared error versus the competing model forecasts but with a more robust significance level of 0.05, and lastly, from nine-month forward forecast to 12-month forward forecasts produced increasingly better predictions, respectively, from 36.6% to 46.6% less RMSE, with a significance level of 0.01.

Given the current case of Latvia's augmentation proposition, it is reasonable to speculate that a little bit more attention to the keyword selection, with regards to gathering the multiple job market dynamics that influence job status changes, like Askitas and Zimmerman (2009) did, would most likely generate more forecast periods with even higher accuracy gains, while being statistically different than the benchmark specification, as Estonia's and Austria's cases.

### *Finland*

Below, it will be presented the augmentation results for Finland's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 4 - Finland augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 17.7 | 26.1 |
| 2 | -2.8 | 5.1 |
| 3 | 1.1 | 13.7 |
| 4 | 13.4 | 11.3 |
| 5 | 11.2 | 10.1 |
| 6 | 12.3 | 11.2 |
| 7 | 14.6* | 9.9* |
| 8 | 19.9*** | 15.1*** |
| 9 | 19.7* | 14.6* |
| 10 | 17.1*** | 13.3*** |
| 11 | 20.0** | 16.8** |
| 12 | 21.4* | 18.2* |

Finland results are like Austria's in a certain aspect, the fact that the accuracy gains are unsteady. Though, there are 6 forecast horizons where Google-augmented forecasts are more accurate than the competing model and the null hypothesis of the Diebold-Mariano test is rejected at one or several usual significance levels. Additionally, the first 6 forecast horizon fail to reject the null hypothesis of Diebold-Mariano test, therefore it cannot be said that the augmented forecasts really bring a different information pool, since the forecasts from the two models could be generated by the same distribution.

Different than other countries from this specific group, the unsteadiness of Finland results does not hold itself only to MAE or RMSE, but also to the significance level that the last 6 forecast horizons that are good candidates to endorse the use of Google enriched model specifications. Thus, as Table 4 references, 7-, 9- and 12-steps ahead forecasts reject the null hypothesis at 0.10, however, 8 and 10-steps ahead rejects at 0.01, and lastly, 11-steps ahead forecasts have 0.05 significance level.

### *Spain*

Below, it will be presented the augmentation results for Spain's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 5 - Spain augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 6.9 | 8.8 |
| 2 | 15.5 | 7.1 |
| 3 | 19.7 | 15.0 |
| 4 | 18.2 | 17.6 |
| 5 | 23.5 | 21.9 |
| 6 | 22.8 | 22.7 |
| 7 | 25.0 | 24.0 |
| 8 | 24.7 | 25.2 |
| 9 | 23.5** | 26.6** |
| 10 | 24.1** | 28.5** |
| 11 | 19.1** | 29.0** |
| 12 | 21.2** | 31.1** |

As demonstrated by Vicente and Menéndez (2015), general Google augmentations using simple constructs could help professional forecasters model Spanish unemployment rate specially when close to periods where the economy faced an exogenous shock. Given that, it is interesting that Spain these displayed promising results with this empirical strategy although that was the case most specifically to forecasts that are more in the future from 3 quarters forward to 4 quarters forward (9-, 10-, 11- and 12-month forward unemployment forecasts).

On Spain's case, as far as root mean squared error (RMSE) goes, there was some consistency over the results. Given that, percentage accuracy gains slowly increased as forecasts were more in the future and starting from 3 quarters forward, all Google-augmented forecasts outperformed with more than 25% error reduction, while being statistically different then their benchmark at 0.05 significance level.

Significant error reductions from 26.6% to 31.1% decrease poses a good case to further investigate if other constructs could lead to significant greater improvements. Despite that, there is a point on the adoption of this type of construct, just like for the other countries in this group, since they can produce consistently robust forecasts that outperform generally used competing models, mostly based on econometric modelling of the unemployment time series.

### 4.1.2. Augmented proposal: possible use cases

Here, it will be highlighted countries with results that are potential use cases of our augmentation proposition for simple econometrics forecasting of unemployment rate. These countries will have at least, 2 forecast periods with greater accuracy than the competing model with statistically different forecasts. Additionally, as seen previously, some countries were focus of studies in this research area with promising results although using this empirical strategy their results were not as promising as previously found, like United States and Italy.

*Iceland*

Below, it will be presented the augmentation results for Iceland's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 6 - Iceland augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 22.3*** | 19.7*** |
| 2 | 36.5*** | 24.6*** |
| 3 | 32.1* | 16.1* |
| 4 | 20.3 | 7.5 |
| 5 | 13.9 | 2.4 |
| 6 | 7.6 | -0.3 |
| 7 | 1.5 | -0.1 |
| 8 | 1.5 | 2.7 |
| 9 | -0.2 | 5.5 |
| 10 | 0.1 | 9.1 |
| 11 | 0.7 | 12.2 |
| 12 | 1.7 | 14.8 |

Iceland's results are the most divergent ones when compared to the best candidates for use cases, and that is the case since it displayed their best results for short-term forecasts. On Table 6, when using Google-augmented model, for 1-month forward forecast MAE decreased 22.3% for a significance level of 0.01, as well as 36.5% decrease for the same metric and significance level. Lastly, one-quarter forward the decrease was 32.1% with the significance level of 0.10, and all other forecasts were not statistically different than the benchmark specification.

There is a very interesting nuance that Iceland discloses for our findings in this research, is that even though some countries could display similar promising results when assessing the same data with a different empirical strategy, the pattern of their results could be very different than what was mainly found. On that note, observe that most countries that manage to extract better predictive explanatory power of this new information pool did so as forecasts were more in the future, having a better and more robust forecasts generally for the latter quarters. However, divergently than this pattern, with this simple empirical strategy Iceland displayed shallow but interesting results for very short-term forecasts and that could be the case for other countries that displayed poor results, but with a different keyword selection or with models that could extract non-linear relations between the feature set.

*Italy*

Below, it will be presented the augmentation results for Italy's case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 7 - Italy augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 18.6 | 32.2 |
| 2 | 17.9 | 32.1 |
| 3 | 20.0 | 30.3 |
| 4 | 24.6 | 31.6 |
| 5 | 27.8 | 32.7 |
| 6 | 23.1 | 31.4 |
| 7 | 19.3 | 30.4 |
| 8 | 15.3 | 29.2 |
| 9 | 17.5* | 33.0* |
| 10 | 20.3** | 36.5** |
| 11 | 20.4 | 40.1 |
| 12 | 23.1 | 42.2 |

For Italy, there were only two forecast periods with accuracy gains that produced statistically different forecasts when compared to the standard specification, which were 9-month and 10-month forward forecasts. Thus, these forecasted, respectively, enabled a 33% and 36.5% reduction in RMSE.

With a different, more specific keyword *oferta di lavoro*, which relates to job offers in English, Naccarato et al. (2018) had promising results when verifying for a similar augmentation proposition when forecasting young unemployment rate for Italy. On another note, Angelico et al. (2022) when modelling inflation expectations alongside widespread survey data used Twitter information pool to augment forecasts for Italy case with promising effects. Therefore, alongside with the results found at this table, it could be the case that with small tweaks on the empirical strategy that is a great amount of explanatory power to be extracted from Google information for unemployment rate forecasting.

### *United States*

Below, it will be presented the augmentation results for United States case. For different forecast periods, the possible accuracy gains alongside the robustness check given by Diebold-Mariano's test over the Google-augmented specification and the benchmark.

Table 8 - United States augmentation results: accuracy gains and robustness

| Step ahead forecasts | Augmentation MAE % decrease | Augmentation RMSE % decrease |
|---|---|---|
| 1 | 67.9 | 81.4 |
| 2 | 54.2 | 61.9 |
| 3 | 63.4 | 69.6 |
| 4 | 65.8 | 73.5 |
| 5 | 67.3 | 77.5 |
| 6 | 70.4 | 80.7 |
| 7 | 71.2 | 83.0 |
| 8 | 72.4 | 84.9 |
| 9 | 73.4* | 86.3* |
| 10 | 74.4** | 87.4** |
| 11 | 75.0 | 88.4 |
| 12 | 75.8 | 89.2 |

For United States, despite a lot of previous findings already addressed in the literature review section of this study, its results just passed our criteria to consider it a possible use case. Here, there was a statistically significant decrease of RMSE of 86.3% and 87.4% for 9- and 10-month forward forecasts, respectively. It seems that simple empirical strategies are unable to extract much predictive power out of Google queries data, unlike sophisticated constructs like the ones recently tested by Borup and Schütte (2022).

### 4.1.3. Augmented proposal: poor use cases

On this part we are going to broadly assess countries that displayed forecasts that are either less performant than their benchmark models, or countries that despite having better forecasts with Google information components could not produce statistically different forecasts than the standard ARIMA specification. These countries are: Slovakia (SK), Slovenia (SI), Sweden (SE), Portugal (PT), Poland (PL), Norway (NO), Netherlands (NL), Mexico (MX), Luxembourg (LU), Lithuania (LT), Ireland (IE), Hungary (HU), Greece (GR), Great Britain (GB), France (FR), Denmark (DE), Germany (DE), Czech Republic (CZ), Canada (CA), Belgium (BE) and Australia (AU).

Given the fact that most OECD countries are within this group, there is a reasonable point on how much this Google information pool cannot be treated as easily adopted augmentation given the current most used modelling practices in our economy. Additionally, despite having countries that are clearly more prone to job search within the internet, and consequently in Google, that are present in *the best candidates* group like Estonia, Finland and Austria, two countries from the same cluster displayed poor results (Ireland and Poland), regardless of their socio-economic status and the impact of it for the augmentation proposal.

Another case, that could be highlighted is cluster 1, that held only three countries that are generally considered very similar: Spain, Italy and Portugal. However, Spain ended up in the *best candidates* group, Italy on the *possible use cases* group and Portugal on the *poor use cases* group. Given the fact

that, the body of literature has some interesting results for the use of Google information for Italy and Spain, these findings are somewhat a confirmation of the potential gains on using this information pool. Additionally, despite being from the same cluster, there is a pattern for the socio-economic indicators that can be seen in bellow figures that could play role on how often these populations do job searches through the internet.

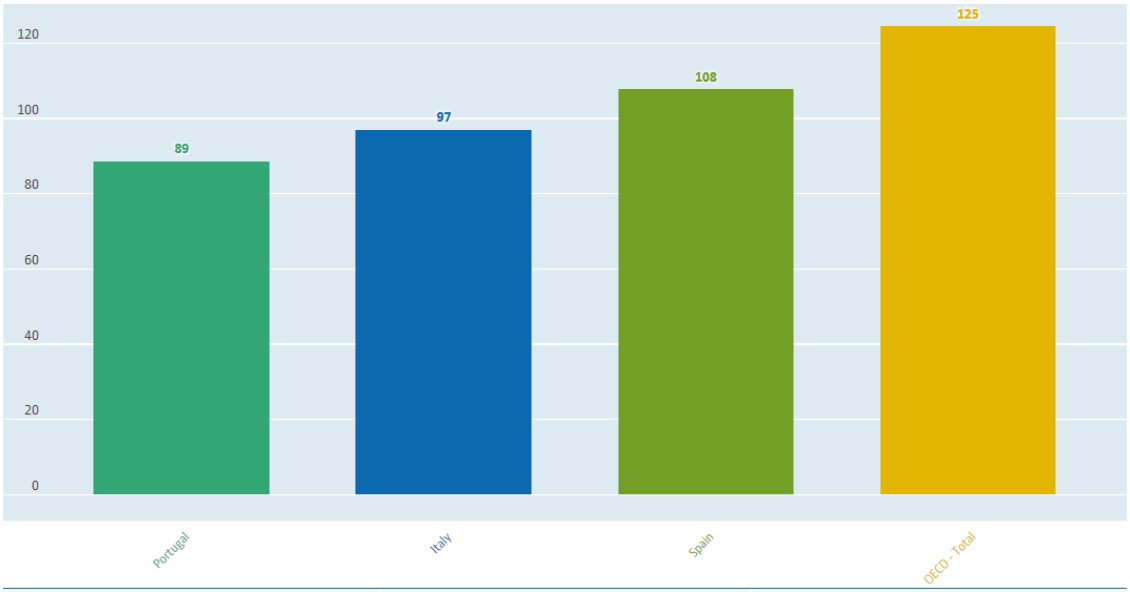


Figure 8 - Mobile broadbands per 100 inhabitants for Cluster 1 countries - Source: OECD (2022)

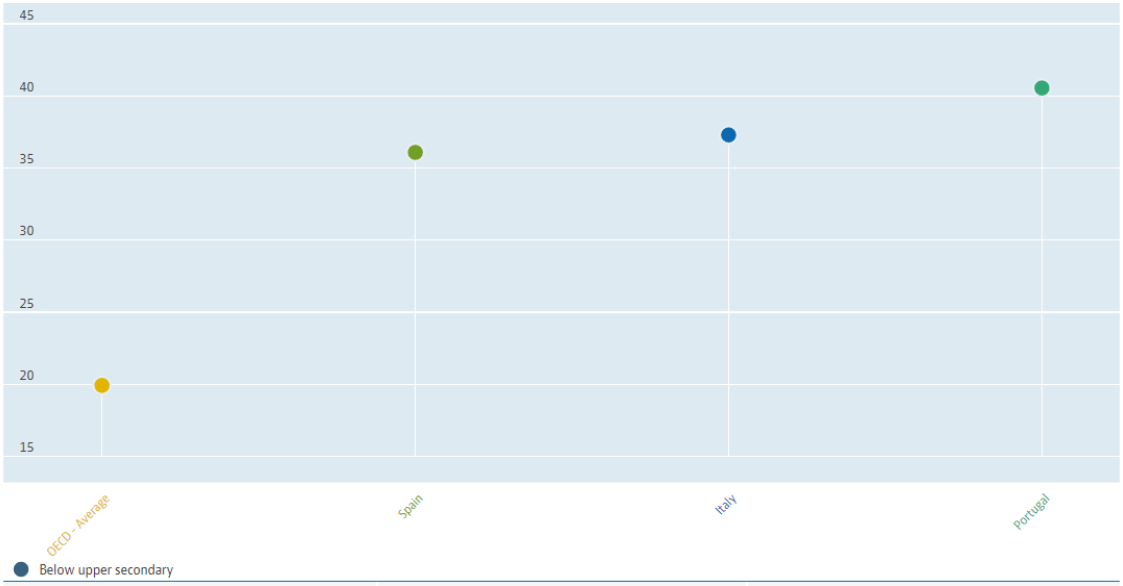


Figure 9 - % of economically active population with below upper-secondary level of education for Cluster 1 countries - Source: OECD (2022)

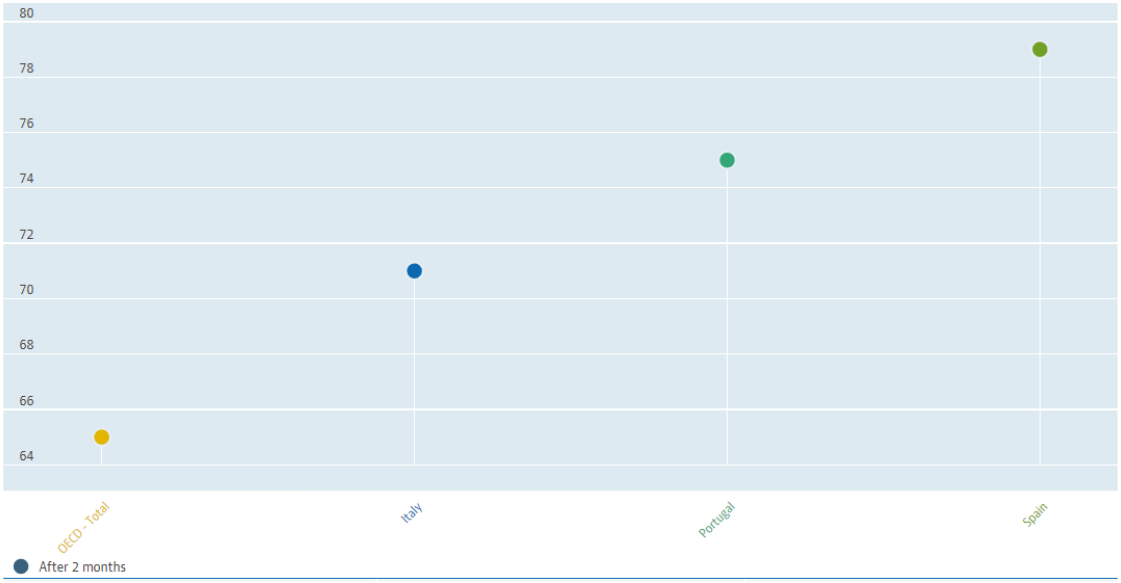


Figure 10 - Unemployment benefits as a % of previous in-work income for Cluster 1 countries - Source: OECD (2022)

Countries like France and Germany are on this group of *poor use cases*, despite having some findings from this research field that shows that Google information could help on better forecasts. Fondeur and Karamé (2013) found using a model that allowed multi-frequency data inputs, and therefore using Google searches weekly time series on its full potential that simple keywords could help predict global unemployment rate and even better so France's young population unemployment rate. Also, Askitas and Zimmermann (2009) showed similar results while using a larger set of keywords for forecasting Germany's unemployment rate and a simpler model than the French researchers. In both cases, the empirical strategy was more sophisticated, and therefore less prone to rapid adoption outside of academia. Apart from that, the amount of datapoints used for both studies were relatively small compared to this study, given the proximity to the Google Trends start date data release (January 2004).

### 4.2. INTERNET JOB SEARCH PRONENESS: CLUSTER ANALYSIS

Firstly, five indicators were selected so that from the OECD countries different groups of countries could adopt internet-related activities with different propensities and that could be associated with these indicators. Furthermore, in order to let the selected dataset eventually displayed these dissimilarities, an unsupervised learning approach was selected: a clustering analysis of these five indicators. Additionally, it could be the case that by using five indicators, this analysis could be impacted by what is commonly addressed as *curse of dimensionality*, due to the five dimensions that the clustering would need to use for finding groups.

For the sake of simplicity, the clustering exercise will be conducted in a two-step fashion: (i) a dimensionality reduction technique will be used to select the three original variables that are mostly

associated with the first three principal components of the full dataset; (ii) a K-means algorithm will be used for defining *k* clusters given different propensity scores presented in below graphs.

Given that, the first approach to eventually reduce the number of dimensions used is to analyze to correlation between the variables, and that, despite mobile internet access points and households with internet access had 32% correlation. Thus, as discussed by Mossberger and Tolbert (2003), low-skilled workers could be less likely to exposed to internet usage on their jobs but nevertheless could be exposed to it through a mobile, or even a household access. However, even with research and development budget as % of GDP had 55% correlation with household internet access, in order to select a more intelligible number of indicators, such as three, a principal component analysis should be conducted.

Therefore, five linear combinations out of the five original variables are constructed, so that the most amount of variance is explained by each principal component. Then, the correlation matrix is recalculated using the three first principal components, those that hold most of the variance of the original dataset (79% of total variance), and the five original variables. The selection criteria were those three variables that could best approximate these three principal components, which were: unemployment benefits (as a % of previous in-work income), number of mobile broadbands per 100 inhabitants and percentage of economically active population with below upper-secondary education level.

*Algorithm and number of clusters definition*

There was a need to group OECD countries, based on the aforementioned socio-economic indicators, in a sensible manner. Therefore, clustering analysis was the selected as a way to perform a data modelling procedure that could objectively extract groups out of the panel of countries. Given that, with a parsimonious approach in mind, there was a clear preference for *hard* cluster specifications over *fuzzy* cluster specifications, which relates to records belonging to different clusters with different levels of affinity.

For that reason, K-Means algorithm was used on this analysis due to being widely adopted as general-purpose clustering algorithm. On that note, the main reasons it is viewed in that manner are: (i) Minimize within-cluster heterogeneity; (ii) Guarantee that groups of separate instances have equal variance. In general, the algorithm selects centroids for each cluster that best approximate their constituents, and at the same time have greater outside cluster heterogeneity.

In order to define how many clusters will be selected for the OECD countries, a structured procedure was performed: (i) Firstly, it was calculated the K-means implementation, for K=[2, 10] 100 times-average and different measures for assessing cluster quality were calculated as well (see graphs below); (ii) Secondly, it was analyzed which K would yield a parsimonious fit to the data while having a strong basis for its selection given the below graphs.

Thus, for the left graph that displays Calinski-Harabasz score and Homogeneity score the higher the value, a better fit to the data the cluster configuration would have. Additionally, for Bayesian Information Criteria (BIC) and Akaike Information Criteria the lower the score, a greater intra-cluster homogeneity and outside cluster heterogeneity. However, it is important to note that, when trying to

define the cluster optimal value different criteria could play a role; the K that gives the best marginal gain; the K that is on point of *diminished returns*, commonly considered the elbow of the plot; as well as having a parsimonious number of clusters, since as K increases the interpretability of the cluster configuration diminishes.

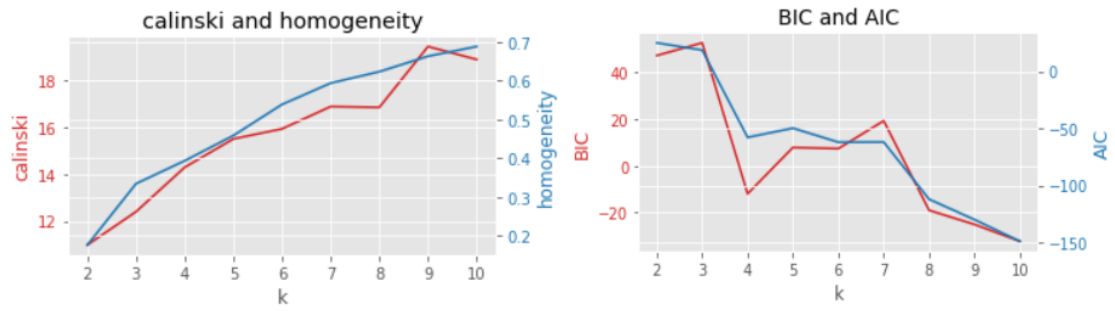


Figure 11 - Scores for defining number of clusters: Calinski-Harabasz and Homogeneity scores / Bayesian Information Criteria (BIC) and Akaike Information Criteria (AIC)

Thus, for the left graph that displays Calinski-Harabasz score and Homogeneity score the higher the value, a better fit to the data the cluster configuration would have. Additionally, for Bayesian Information Criteria (BIC) and Akaike Information Criteria the lower the score, a greater intra-cluster homogeneity and outside cluster heterogeneity. However, it is important to note that, when trying to define the cluster optimal value different criteria could play a role; the K that gives the best marginal gain; the K that is on point of *diminished returns*, commonly considered the elbow of the plot; as well as having a parsimonious number of clusters, since as K increases the interpretability of the cluster configuration diminishes.

Given that, 4 clusters appeared to be the most appropriate choice and we are going to present some of the characteristics that constituted its definition in some graphs below, alongside their grouping based on the augmentation proposition results.

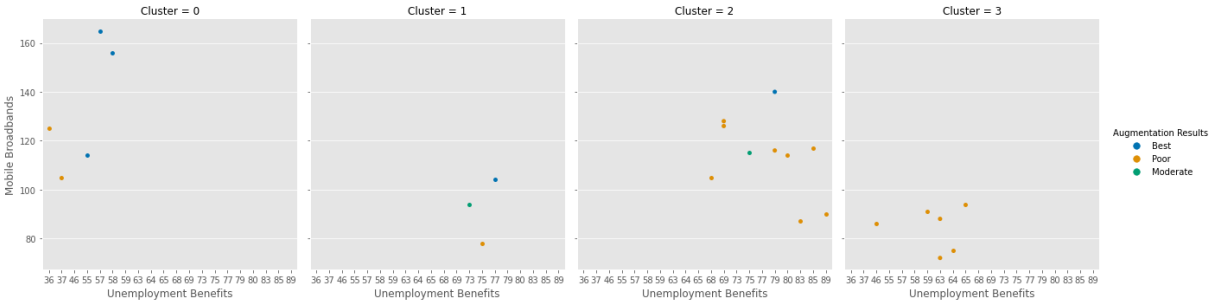


Figure 12 - Unemployment benefits, % of previous in-work income, and Mobile broadbands per 100 inhabitants for every cluster

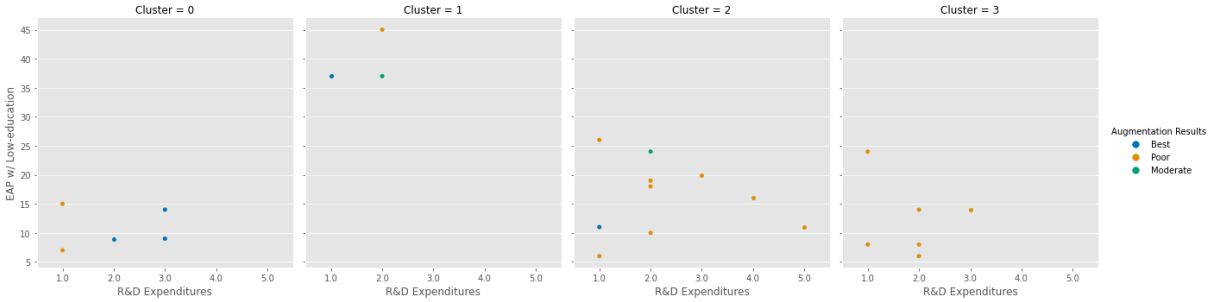


Figure 13 - Research and development expenditures, % of GDP, and Economically active population (EAP) with at most below upper-secondary education level

By using the median of the clusters, the following summary of the above graphs can be stated:

Cluster 0 has the following characteristics, in general, highest accessibility over internet, lowest unemployment benefits as percentage of previous in-work income, 2nd highest research and development expenditures and lowest percentage of economically active population that holds, at most, below upper-secondary level of education. An elucidative example of this country is Finland.

Cluster 1 has the following characteristics, in general, 2nd lowest accessibility over internet, 2nd highest unemployment benefits as percentage of previous in-work income, lowest research and development expenditures and highest percentage of economically active population that holds, at most, below upper-secondary level of education. An elucidative example of this country is Italy.

Cluster 2 has the following characteristics, in general, 2nd highest accessibility over internet, highest unemployment benefits as percentage of previous in-work income, highest research and development expenditures and 2nd highest percentage of economically active population that holds, at most, below upper-secondary level of education. An elucidative example of this country is Sweden.

Cluster 3 has the following characteristics, in general, lowest accessibility over internet, 2nd lowest unemployment benefits as percentage of previous in-work income, 2nd lowest research and development expenditures and 2nd lowest percentage of economically active population that holds, at most, below upper-secondary level of education. An elucidative example of this country is Hungary.

Lastly, if there is any indication of socio-economic characteristics that leads to considering including Google components in unemployment forecasting would be the ones presented by Cluster 0. With regards to a single socio-economic indicator that seems to be associated with higher and more robust accuracy gains would be mobile broadbands per 100 inhabitants, and more broadly socio-economic indicators that depict internet usage by any population.

# 5. CONCLUSION, LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Section 5 revolves around how likely it is that Google Trends data could be a *panacea*, how societies circumstances could play a role on the explaining the likelihood of an easy augmentation through Google data usage for modelling the job market. On top of that, how data mining and machine learning seems to be a way too fully exploit this new information pool from the digital footprint societies leave on the Internet.

## 5.1. CONCLUSION

### Google Trends as a panacea

As thoroughly described in this literature review and during our explanation on the methodology used on this study, Google Trends could be instrumental to narrow information gaps for some macroeconomic variables, as seen in Narita et al. (2021), and also augment common forecasted indicators, such as unemployment rate.

However, our empirical strategy and data selection aimed to propose a simple augmentation to commonly used models so that, if robust results were found across the whole panel of countries, Google Trends information pool could rapidly be exploited. The notion that Google Trends could be a low-hanging fruits for general forecasters seemed very much possible given the current state of this research field, nevertheless, our results posed a pinch of salt into this findings in on country or another, like United States and France.

It seems to be that Google Trends are no *panacea* for generating better forecasts, since while controlling for statistically different better forecasts were only able to bet industry plain and simple ARIMA specifications for eight countries out of the 29 countries. Additionally, three out of these eight countries are possible good candidates of a similar empirical strategy used in this analysis, however, small tweaks will be needed since most of forecast periods did not produce robust forecasts with greater accuracy as described in the previous section.

Thus, five countries are strong candidates to wide adoption of Google information data to enrich their modelling practices with regards to unemployment rate. Most of these countries are similar in many characteristics that could play a role on moving part of job search dynamics to the internet, but some countries that displayed interesting results are not that homogenous on those criteria.

### Country differences in augmentation proposition: circumstances could play a role?

It is quite clear that Cluster 0 countries is associated with ease of adoption of our augmentation proposition, given the fact that most of its countries were placed on the *best candidates* group. These countries have, in general, a high-skilled workforce, are relatively very much connected to the internet, while living in states that provide high investments in research and development. Regardless of any

causal effect, their workforce generally experiences a strong change in income when becoming unemployed.

An interesting pattern could also be observed by the three countries that constitute the Cluster 1, since one of them is placed on the *best candidates* group, another one is on the *possible use cases* group and the last one on the *poor use case* group. The country within this cluster that provides higher investments in research and development, as well as have higher prevalence of internet accessibility and has lower percentage of low-skilled workers, as defined by Holzer (1996), yielded the best results and alternatively, the poorest results were displayed by the country with relatively the inverse characteristics.

OECD countries in summary do share a lot of similarities and therefore have similar standards of living. However, their differences could be associated with the need to use a more computationally intensive empirical strategy in order to exploit the new information pool that Google queries can provide. Borup and Schütte (2022) provides some insights for the United States, analyzing each state individually, that with a very large number of keywords gathered and a non-linear algorithm despite different circumstances for each state Google queries can robustly augment forecasts for unemployment modelling. Lastly, Estonia's case of being one of the most digitalized countries shows how some country circumstances can lead to an easier rote to adopt any new information pool that becomes available for professional forecasters looking for more predictive power over variable of interests.

**Machine learning as a way to exploit new information pool**

As of now, Elliot et al. (2015) findings and many others display us that with "*enough steam power*" Google Trends can be instrumental for professional forecasters. However, public policymakers, economists and analysts in general are not at par with the cutting-edge algorithms that are being widely used by Machine Learning researchers in academia. Given that, our empirical strategy aimed to check if globally there could a low-hanging fruit for professional forecasters with the use of Google searches data, which does not seem to be the case.

About 30% of OECD countries results endorse the adoption of Google-augmented specifications for widely adopted algorithms, such as ARIMA. Additionally, when using empirical strategies with daunting implementation procedures for any forecaster, but that select, despite socio-economic dissimilarities, different set of keywords that hold explanatory power in Google queries data, shows the current trade-off this research area currently has.

Nowadays, it does not seem that the job market has the maturity to adopt non-linear algorithms, like Random Forest, as well as data mining procedures to gather more than 100 keywords, while following a robust procedure like in Medeiros and Pires (2020). Furthermore, it seems that the explanatory power is coming from a large set of keywords and a framework to do feature selection, while not holding the most interesting features fixed for every horizon forecasted. Lastly, this nuance seems to be the biggest takeaway from current state of academia with regards to the computationally daunting models.

Works upscaling Machine Learning for different topics have been popping out in academia, such as spatial analysis of injuries in Vaz et al. (2021), or even spatial analysis of wealth in a city in Vaz et al.

(2021). Studies like these could inform public officials in different tasks like urban planning, policymaking with respects to income distribution, and other sensitive areas that state officials must grapple nowadays. Given that, the wide adoption of Machine Learning (ML) toolset seems to be nearing its *tipping point*, when most forecasters that are not versed in ML frameworks will be consistently beat by their peers that have upskilled themselves into it.

## 5.2. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Firstly, it will be presented the constraints of econometrics-based approaches to augment macroeconomic modelling in a more general manner. Secondly, how nimble modelling decisions seem to impact robustness and accuracy gains for augmentations based on Google data. Thirdly, how a broad study like this one using also youth data could further elucidate the limitations around this empirical strategy, and not country characteristics or lack of predictive power in the keyword selection.

### Econometrics based approach and its limitations

On this study, our rationale was to use one of the most common algorithms (ARIMA) used for forecasting procedures, given the fact that most statisticians and econometricians are familiar with it.

However, this algorithm due to its linearity “force” any process a practitioner is trying to model solely due to its construct. Additionally, any additional exogenous component that could be used on its specification will not be removed or included with any standard procedure. On that note, this modelling approach can allow, at most, some regularization techniques that are not feature selection *per se*, despite providing some of the benefits of it, and can end up removing components that are important for forecasts *x* steps forward, but not so much for forecasts *y* steps forward.

### Agile versus slow approaches to augmenting forecasts

Another limitation was that the Google keyword used was "jobs”, given its variants for every country, and it could be the case that some countries it held some insight to the dynamics desired and other was not the case. Here, there is the clear *trade-off* of having an agile procedure, that is computationally- and time-friendly, and having an exhaustive bag of words that could be very slow to gather, because there is the need to gather it several times to guarantee consistency of the keywords).

*Simple keyword query - current status*

This study provides an overview of the current status of potential accuracy gains while using simple keywords from Google. The broad nature of our selection of countries, the assessment that revolved around different forecasting periods with a robustness check can provide some nuance to new researchers when deciding their focus point.

Furthermore, there is still room to cover when verifying the potential of single keyword for augmentations. Two main points can be addressed in next studies: (i) Information gaps in low-income

countries: the necessary focus in some of these countries, ideally in large set of countries; (ii) Using an algorithm that could better use the information of one single keyword, like Random Forest.

*Random Forest with a simple keyword query*

The Random Forest (RF) algorithm implementation could, in theory, extract more predictive power through, at least, these procedures: (i) Some feature engineering on the single keyword from Google data; (ii) Feature sampling on these Google-constructed components. To have this for all countries analyzed in this study would be very interesting to assess simple keywords full potential for augmentations.

Furthermore, different algorithm, like XGBoost could produce even better results due to their construction. However, in general the enforcement of these procedures alongside the algorithm characteristics that allows for extracting non-linear associations between the regressor variables and the variable of interest could help researchers validate the augmentation power of Google data.

Despite the caveat that this algorithm, apart from data practitioners, is not so widely used and known outside academia, interesting findings like in Richardson and Mulder (2018) show that Machine Learning models outperform statistics-based models, like ARIMA, in most cases. With respects to RF, the ability to prune the feature set to features with highest cost-benefit, as well as to assess different specifications with larger feature sets, since it can easily be specified with less amount of regressors through recursive feature extraction are also important key points to extract the most of this different empirical strategy.

*Random Forest with very large pool of keywords*

Borup and Schütte (2022) proposition depicts what nowadays could be the most exhaustive assessment of the potential use of Google data for unemployment forecasting. Their study displayed results both to the US and to the states individually but did not discuss why different states had different accuracy gains results. The focus was to verify for a data-rich environment, a large set of keywords, which modelling approach could extract the most explanatory power out of the information pool.

Their study consolidated an empirical strategy that should be tested to a wide variety of countries, alongside with an analysis of socio-economic indicators of each country. This recommendation could be very instrumental to understand why different countries have greater accuracy gains when using models that exploit the most out of Google trends.

**Youth data**

Lastly, a very interesting analysis could be to assess for a large panel of countries what is the difference between augmenting unemployment forecasting and young unemployment forecasting. Promising findings were done by Fondeur and Karamé (2013) and Naccarato et al. (2018), however their focus point were France and Italy, respectively, and by having various countries it could be assessed the reasons why young unemployment rate had greater accuracy gains as seen on their results.

# 6. REFERENCES


Alvarez-Melis, D., & Saveski, M. (2016). Topic modeling in Twitter: Aggregating tweets by conversations. In Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016 (pp. 519–522). AAAI Press.

Angelico, C., Marcucci, J., Miccoli, M., & Quarta, F. (2022). Can we measure inflation expectations using Twitter? Journal of Econometrics, 228(2), 259–277. https://doi.org/10.1016/j.jeconom.2021.12.008

Askitas, N., & Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly, 55(2), 107–120. https://doi.org/10.3790/aeq.55.2.107

Bai, J., & Ng, S. (2008). Forecasting economic time series using targeted predictors. Journal of Econometrics, 146(2), 304–317. https://doi.org/10.1016/j.jeconom.2008.08.010

Borup, D., & Schütte, E. C. M. (2022). In Search of a Job: Forecasting Employment Growth Using Google Trends. Journal of Business and Economic Statistics, 40(1), 186–200. https://doi.org/10.1080/07350015.2020.1791133

Cebrián, E., & Domenech, J. (2022). Is Google Trends a quality data source? Applied Economics Letters. https://doi.org/10.1080/13504851.2021.2023088

Chen, H., De, P., Hu, Y., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. Review of Financial Studies, 27(5), 1367–1403. https://doi.org/10.1093/rfs/hhu001

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends. Economic Record, 88(SUPPL.1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Curado, A., Damasio, B., Encarnacão, S., Candia, C., & Pinheiro, F. L. (2021). Scaling behavior of public procurement activity. PLoS ONE, 16(12 December). https://doi.org/10.1371/journal.pone.0260806

Damásio, B., & Nicolau, J. (2014). Combining a regression model with a multivariate Markov chain in a forecasting problem. Statistics and Probability Letters, 90(1), 108–113. https://doi.org/10.1016/j.spl.2014.03.026

Damásio, B., & Nicolau, J. (2020). Time inhomogeneous multivariate Markov chains: detecting and testing multiple structural breaks occurring at unknown.

Damásio, B., Louçã, F., & Nicolau, J. (2018). The changing economic regimes and expected time to recover of the peripheral countries under the euro: A nonparametric approach. Physica A: Statistical Mechanics and Its Applications, 507, 524–533. https://doi.org/10.1016/j.physa.2018.05.089

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. International Journal of Forecasting, 33(4), 801–816. https://doi.org/10.1016/j.ijforecast.2017.03.004

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business and Economic Statistics, 13(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

Elliott, G., Gargano, A., & Timmermann, A. (2013). Complete subset regressions. Journal of Econometrics, 177(2), 357–373. https://doi.org/10.1016/j.jeconom.2013.04.017

Elliott, G., Gargano, A., & Timmermann, A. (2015). Complete subset regressions with large-dimensional sets of predictors. Journal of Economic Dynamics and Control, 54, 86–110. https://doi.org/10.1016/j.jedc.2015.03.004

Engle, R. F., Granger, C. W. J. (1987). CO-INTEGRATION AND ERROR CORRECTION: REPRESENTATION, ESTIMATION, AND TESTING. Source: Econometrica Econometrica, 55(2), 251–276. Retrieved from http://www.jstor.org

Fan, J., Lv, J., & Qi, L. (2011). Sparse high-dimensional models in economics. Annual Review of Economics. https://doi.org/10.1146/annurev-economics-061109-080451

Fondeur, Y., & Karamé, F. (2013). Can Google data help predict French youth unemployment? Economic Modelling, 30(1), 117–125. https://doi.org/10.1016/j.econmod.2012.07.017

Francesco, D. (2009). Predicting unemployment in short samples with internet job search query data. MPRA Working Paper, (18403). Retrieved from http://mpra.ub.uni-muenchen.de/18403/1/MPRA_paper_18403.pdf

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457(7232), 1012–1014. https://doi.org/10.1038/nature07634

Golovchenko, Y., Buntain, C., Eady, G., Brown, M. A., & Tucker, J. A. (2020). Cross-Platform State Propaganda: Russian Trolls on Twitter and YouTube during the 2016 U.S. Presidential Election. International Journal of Press/Politics, 25(3), 357–389. https://doi.org/10.1177/1940161220912682

Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. Consumer Price Inflation. Journal of the American Statistical Association, 103(482), 511–522. https://doi.org/10.1198/016214507000000473

Lyra, M. S., Curado, A., Damásio, B., Bação, F., & Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. Applied Network Science, 6(1). https://doi.org/10.1007/s41109-021-00418-y

Lyra, M. S., Damásio, B., Pinheiro, F. L., & Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. Applied Network Science, 7(1), 83.

Holzer, H. J. (1996). What employers want: Job prospects for less-educated workers. Russell Sage Foundation.

Hong, L., & Davison, B. D. (2010, July). Empirical study of topic modeling in twitter. In Proceedings of the first workshop on social media analytics (pp. 80-88).

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. Journal of Statistical Software, 27(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Kholodilin, K. A., Podstawski, M., Siliverstovs, B., & Bürgi, C. (2011). Google Searches as a Means of Improving the Nowcasts of Key Macroeconomic Variables. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1507084

Martins, D., & Damásio, B. (2020). One Troika fits all? Job crash, pro-market structural reform and austerity-driven therapy in Portugal. Empirica, 47(3), 495–521. https://doi.org/10.1007/s10663-019-09433-w

Medeiros, M. C., & Pires, H. F. (2021). The proper use of google trends in forecasting models. arXiv preprint. https://arxiv.org/pdf/2104.03065.pdf

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In SIGIR 2013 - Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 889–892). https://doi.org/10.1145/2484028.2484166

Mincer, J. (1958). Investment in Human Capital and Personal Income Distribution. Journal of Political Economy, 66(4), 281–302. https://doi.org/10.1086/258055

Mossberger, K., Tolbert, C. J., & Stansbury, M. (2003). Virtual inequality: Beyond the digital divide. Georgetown University Press.

Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. Technological Forecasting and Social Change, 130, 114–122. https://doi.org/10.1016/j.techfore.2017.11.022

Narita, F., & Yin, R. (2021). In Search of Information: Use of Google Trends' Data to Narrow Information Gaps for Low-Income Developing Countries. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3333732

Paredes, A., Mendonça, J., Bação, F., & Damásio, B. (2022). Does R&D tax credit impact firm behaviour? Micro evidence for Portugal. Research Evaluation, 31(2), 226–235. https://doi.org/10.1093/reseval/rvac002

Rosen-Zvi, M., Griffiths, T., Smyth, P., & Steyvers, M. (2005). Learning author topic models from text corpora. Journal of Machine Learning Research, V(October), 1–38. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.7284&rep=rep1&type=pdf http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Learning+Author-Topic+Models+from+Text+Corpora%7B%5C#%7D0

Schultz, T. W. (1960). Capital formation by education. Journal of political economy, 68(6), 571-583. https://www.journals.uchicago.edu/doi/10.1086/258393

Tsay, R. S. (2000). Time series and forecasting: Brief history and future research. Journal of the American Statistical Association, 95(450), 638–643. https://doi.org/10.1080/01621459.2000.10474241

Vaz, E., Bação, F., Damásio, B., Haynes, M., & Penfound, E. (2021). Machine learning for analysis of wealth in cities: A spatial-empirical examination of wealth in Toronto. Habitat International, 108. https://doi.org/10.1016/j.habitatint.2021.102319

Vaz, E., Cusimano, M. D., Bação, F., Damásio, B., & Penfound, E. (2021). Open data and injuries in urban areas—A spatial analytical framework of Toronto using machine learning and spatial regressions. PLoS ONE, 16(3 March). https://doi.org/10.1371/journal.pone.0248285

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? Technological Forecasting and Social Change, 92, 132–139. https://doi.org/10.1016/j.techfore.2014.12.005

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. Journal of Forecasting, 30(6), 565–578. https://doi.org/10.1002/for.1213


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa