Technological Forecasting & Social Change xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

Combining official and Google Trends data to forecast the Italian youth unemployment rate

Alessia Naccarato[a,*], Stefano Falorsi[b], Silvia Loriga[b], Andrea Pierini[a]

[a] Department of Economics, Roma Tre University, Via Silvio D'Amico 77, 00145 Rome, Italy
[b] ISTAT - Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Rome, Italy

ARTICLE INFO

Keywords:
Labour force survey
Google Trends query share
ARIMA model
VAR model

ABSTRACT

The increased availability of online information in recent years has aroused interest in the possibility of deriving indications for many kinds of phenomena. In the more specific economic and statistical context, numerous studies suggest the use of online search data to improve the forecasting and nowcasting of official economic indicators with a view to increasing the promptness of their circulation. The purpose of this work is to investigate if the use of big data can improve the forecasting of the youth unemployment rate – estimated in Italy on a monthly basis by the Italian National Institute of Statistics – by means of time series models. The time series used are those of the Google Trends query share for the keyword *offerte di lavoro* (job offers) and the official labour force survey data for the Italian youth unemployment rate since 2004. Two different models are estimated: an ARIMA model using only the official youth unemployment rate series and a VAR model combining the former series with the Google Trends query share. The results show that the use of Google Trends information leads to an average decrease in the forecast error.

1. Introduction

The paper addresses a problem that is closely connected with the forecasts of official statistical indicators produced monthly by national departments of statistics in all the European countries based on data gathered through sample surveys. For all these indicators, including the youth unemployment rate, it is useful to have forecasts that give some idea of the variations to be expected in the short term, not only because they can serve as preliminary estimates of the phenomenon examined but also because they can be incorporated into econometric models to provide indications of the prospects for the economic system as a whole. The Italian National Institute of Statistics (ISTAT) produces such forecasts by means of statistical models that make no use of information from external sources and rely only on the data provided by its own sample surveys.

The purpose of the experiment presented in this paper is to ascertain if the combined use of information from Google Trends (GT) and the ISTAT labour force survey makes any improvement to the forecasting of the youth unemployment rate in terms of the accuracy and timeliness of the estimates.

Many scholars and researchers have become aware in recent years that the vast amount of information to be derived from the mass of online search data available could prove useful in the study of social phenomena (Askitas and Zimmermann, 2015; Chamberlin, 2010; Choi and Varian, 2012; Daas et al., 2015; Einav and Levin, 2014; Guzman, 2011; Marchetti et al., 2015). This is because the information that people disclose about their needs using the Internet (Ettredge et al., 2005) can shed light on the variability of numerous phenomena under examination. A review of the application of these types of data in different fields can be found in Hassani and Silva (2015).

With particular reference to Google Trends data, it is easy to state that the relevance and potential of this tool lies in their characteristic of containing a huge amount of information that is easily accessible yet difficult to obtain in another way. This represents a great opportunity especially for statistical purposes, and the interest of the scientific community is evident by the amount of work proposed in literature over the past 10 years. Only a few of the most recent studies are involved in small area estimation problems (Porter et al., 2013; Rao and Molina, 2015, p. 159–170) in the spread of infectious diseases (Anggraeni and Aristiani, 2017; Teng et al., 2017), in numerous economic contexts such as consumer purchases (Schmidt and Vosen, 2013), housing market (Limnios and You, 2016), tourist inflows (Dinis et al., 2017), forecasting energy (Hassani and Silva, 2016), credit developments (Burdeau and Knitzler, 2017), and social phenomena (Nghiem et al., 2016; Nixon, 2016; Parker et al., 2017).

In the official statistics field, some studies have been carried out to

* Corresponding author.
E-mail addresses: alessia.naccarato@uniroma3.it (A. Naccarato), stfalors@istat.it (S. Falorsi), loriga@istat.it (S. Loriga), andrea.pierini@uniroma3.it (A. Pierini).

https://doi.org/10.1016/j.techfore.2017.11.022
Received 28 November 2016; Received in revised form 28 October 2017; Accepted 14 November 2017

assess if online search data – in particular Google Trends data – can be used to facilitate the estimation of the phenomena of interest for the national statistical institutes or to produce additional information (Charles-Edwards, 2016; D'Alò et al., 2015; Falorsi et al., 2017; Kristoufek et al., 2016).

Regarding the phenomenon studied in this work, numerous authors suggest the use of Internet data to forecast unemployment (Anvik and Gjelstad, 2010; Askitas and Zimmermann, 2009; D'Amuri and Marcucci, 2017; Falorsi et al., 2015; McLaren and Shanbhogue, 2011), and the results show that they can indeed be regarded as useful in the estimation procedure.

The monthly estimates of Italian employment and unemployment published by the ISTAT have been accompanied regularly in the mass media by a particular focus on youth unemployment, the levels of which rose steadily during the recent economic crisis and are markedly above the European average, hence the interest in providing more timely and reliable information regarding this aspect.

Moreover, the official estimate of the youth unemployment rate in Italy is based on a monthly sample survey and generally becomes available about 30 days after the end of the month in question due to the time required to gather and process the relevant data. It could prove to be useful in many situations to have reliable indications of the results of an ongoing survey before all of its phases have been completed. In other words, the need may arise for a preview or nowcast (Choi and Varian, 2012) of the youth unemployment rate for which the survey is being carried out in the same period. The problem therefore is to produce estimates that are available while the actual survey is still under way – or in any case in a short space of time – to provide reliable information on the indicator in the current month and its variation with respect to the previous one (Ayoubkhani, 2012; D'Alò et al., 2006).

It should be noted that for Italy, as well as the other European countries, the information required by the European Community for purposes of economic analysis is vast and drawn from a variety of large-scale surveys based both on samples and on censuses. These requirements also are laid down in the European Community regulations on short-term business statistics in force since August 2005. The same date saw the launch of the Action Plan on Economic and Monetary Union (EMU) Statistical Requirements by Eurostat and the Central European Bank with the involvement of the European national departments of statistics in an effort to reduce the time required for the production and circulation of the most important indicators that are essential to the short-term analysis of the European economy.

The purpose of this work is to ascertain if the use of information from Google Trends together with data from the time series of the ISTAT labour force survey can improve the accuracy and timeliness of the youth unemployment rate forecast (Barreira et al., 2013; Fondeur and Karamè, 2013) and make it possible to provide an "immediate" forecast for use, even if only as a pointer, while awaiting the official figures.

Google Trends information is subjected to no form of quality control and, in our opinion, can be used solely as a sort of "snapshot", providing indications about the phenomenon of interest and its evolution in time and space. As C.F. Citro (2014) observes:

> Statistical agencies need, above all, sources of data that cover a known population with error properties that are reasonably well understood and that are not likely to change under their feet – characteristics that are not inherent in such data sources as autonomous interactions with websites on the Internet. There are, however, at least two ways in which household survey-based statistical agency programs could obtain an "edge" from non-traditional sources: one is to improve timeliness for preliminary estimates of key statistics; and the other is to provide leading indicators of social change (e.g., the emergence of new occupations and fields of training) that alert statistical agencies to needed changes in their concepts and measures.

Goel et al. (2010) provide a useful survey of the work undertaken in this area and describe some of the limitations of web search data. As they point out, such data are readily obtainable and often helpful in making forecasts but may not provide dramatic increases in predictability.

As the use of the Internet for job seeking is more widespread among the younger population, the trend in the monthly number of job searches made through Google can provide useful information for forecasting the youth unemployment rate. The basic idea is that an increase in this rate is accompanied by an increase in the number of young people seeking employment opportunities through online searches. As shown in Section 2.1, this idea is borne out by the results of a specific query in ISTAT's monthly survey on the labour force.

Google Trends data corresponding to the keyword "*offerte di lavoro*" (job offers) are used in this paper as auxiliary information to forecast the Italian youth unemployment rate. They belong to the sphere of what is known in the literature as "big data" (Ceron et al., 2014; Choi and Varian, 2012; Einav and Levin, 2014), which have the characteristic of being immediately available and capable of supplying an up-to-date picture of social and economic phenomena. As Chamberlin (2010) writes:

> As this search data from Google Trends is available in real time, any significant relationship could be potentially exploited for nowcasting. Google Trends data may also be used informally to pick up on changing patterns and rising trends in search queries and the implications they have for types of economic activity and spending.

Moreover, Google Trends data generally can be used at a very low cost, which constitutes a further reason for the interest in the sphere of official statistical surveys.

However, there are some aspects in the use of GT data processing that cannot be left out. Among these, one of the most important is the choice of keywords to be used for the selection of data sets. Different keywords lead to different results; however, if the selection criterion derives from an adequate knowledge of the phenomenon under study and if you carry out controls on a large number of keywords and their combinations, you can obtain useful results.

In extracting information from large amounts of data, the errors that can be made are multiple (Braaksma and Zeelemberg, 2015; Nuti et al., 2014; Zeelemberg and Braaksma, 2017, p. 274–296) and many are the consequences of an inappropriate choice of keywords. The most probable error, especially in the event that GT information is used to make forecasts by means of regression models – as is the case in this paper – is to run into spurious correlations. The GT series generally is extracted on the basis of the correlation that it shows with the phenomenon under study. If two or more phenomena are statistically correlated, however, it does not necessarily mean that there is a direct cause-effect link between them because this correlation can be completely random (i.e. spurious) or dependent on a further variable in common.

To reduce this risk, different keywords or combinations of keywords are chosen, the variability of the series extracted from Google Trends is studied as keywords inserted in the query shares change, and finally for the extracted series the statistical significance of the correlation between these and the variable being studied is verified.

To evaluate the utility of GT data in forecasting the Italian youth unemployment rate, a comparison was developed between results obtained by means of an ARIMA model (Box et al., 2007), making no use of auxiliary GT information, and a vector autoregressive (VAR) model (Lutkephol, 2007), harnessing the correlation of official and Internet search data.

The ARIMA methodology is the one already employed by the ISTAT to obtain forecasts in the sphere of national accounts and indicators of trends in the Italian economy. These forecasts are always short term (no longer than 3 months) and provide information on the expected tendency of the economic indicator in a given month of the year with respect to its value as published based on survey data 2 or 3 months earlier. In other words, the purpose of the forecasts is to provide