

Does Google index improve the forecast of Italian labour market?

E' possibile migliorare le previsioni del mercato Italiano del lavoro attraverso i dati di Google trend?

Bacchini F., D'Alò M., Falorsi S., Fasulo A., Pappalardo C.

Abstract It is possible to exploit Google Trend Index (GTI) for different statistical purposes. At national level GTI allows to exploit the time series of query shares to improve the quality of estimates of short-term (monthly or quarterly) socio-economic indicators. In particular GTI may be used for improving the forecasting or nowcasting of such short-term indicators. In this paper the first results of an empirical study aimed to evaluate the use of GTI for the forecasting (one month ahead) of the number of people seeking for job from Labour Force Survey.

Abstract È possibile sfruttare Google Trend Index (GTI) a vari scopi statistici. A livello nazionale GTI permette di sfruttare la serie storica delle query share per migliorare la qualità delle stime di breve termine (mensile o trimestrale) di alcuni indicatori socio-economici. In particolare GTI può essere utilizzato per migliorare la previsione a breve e lungo termine di tali indicatori. In questo lavoro si mostrano i primi risultati di una ricerca empirica volta a valutare l'uso di GTI per la previsione (un mese prima), del numero di persone in cerca di occupazione usando dati dell'indagine sulle forze di lavoro.

Key words: Google, Big Data, forecast, labour market, autoregressive models, time series

1 Introduction

A stream of studies have considered the predictive capability of web search data, such as those made available by the Google Trends service, a real-time daily and weekly index of the volume of queries that users enter into Google. Choi and Varian introduced the term “predicting the present”, or “nowcasting” as opposed to “forecasting”, to advocate the tendency of web searches to correlate often with various economic indicators, which may reveal helpful for short-term economic prediction. The releases of indicators of socio-economic activity are typically available with a reporting lag of several weeks and are often revised a few months later. It would clearly be helpful to have more timely estimates of these indicators.

As a popular example in the field of epidemiology, Polgreen et al. [2008] and Ginsberg et al. [2009] showed that search data could help predict the incidence of influenza-like diseases. Intuitively, there is a close relationship between how many people search for flu-related topics and how many people actually have flu symptoms. Of course, not every person who searches for "flu" is actually sick, but a pattern emerges when all the flu-related search queries are aggregated. The authors compared the query counts with traditional flu surveillance systems and found that many search queries tend to be popular exactly when flu season is happening. By counting how often we see these search queries, it is possible estimate how much flu is circulating in different countries and regions around the world. This finding is important because traditional flu surveillance agencies deliver their estimates with a delay of weeks, while the web search-based service delivers basically daily, and timeliness is crucial to enable public health officials and health professionals to better respond to seasonal epidemics and pandemics.



Figure 1: United States flu activity. Blu: Google Flu Trends estimate; Orange: Influenza-like illness data (ILI) provided by the U.S. Centers for Disease Control

One of the first study that suggested how web search data are useful in forecasting economic statistics was Ettredge et al. [2005], which examined the U.S. unemployment rate. Since then there have been several papers that have examined web search data in various fields.

In economics, Choi and Varian [2009] described how to use Google Search Insights data to predict several economic metrics including initial claims for unemployment, automobile demand, and vacation destinations. Askitas and Zimmermann [2010], D'Amuri and Marcucci [2010, 2012], and Suhoy [2009] examined unemployment. Guzman [2011] has examined Google data as a predictor of inflation.

Radinsky et al. [2009], Huang and Penna [2009], and Preis et al. [2010] examine the use of search data for measuring consumer sentiment while Schmidt and Vosen [2009] and Lindberg [2011] examine retail sales and consumption metrics. Wu and Brynjolfsson [2010] examine housing data using longitudinal data extracted from Google Search Insights

Goel et al. [2010] provide a useful survey of work in this area and describe some of the limitations of web search data. As they point out, search data is easy to acquire and is often helpful in making forecasts, but may not provide dramatic increases in predictability.

2 The use of big data as auxiliary information for model based techniques for the production of provisional data and small area estimates

Some recent literature proposes the use of big data as external auxiliary information for the production of model based estimates obtained by large scale surveys on households and enterprises. We have discussed in this paper how to exploit the time series of data collected by web search services and made available through analytic interfaces like Google Trends (GT), with the aim to improve basic preliminary estimates for a set of strategic target parameters useful to economic analyses (Choi and Varian 2012).

In addition to provisional estimation, in Pratesi et al (2013) three possible approaches are considered for statistical use of big data: 1) to use big-data as auxiliary external information in the small area model; 2) to check and remove the self-selection bias from estimates obtained using big data; 3) to use big data to make comparisons with results obtained with small area estimation methods.

Istat is involved since several years in the study and application of predictive models aimed to increase the quality of preliminary estimates of business and socio-economic indicators and for small area estimation purposes.

Starting from early 2000, Istat has been producing preliminary estimates for economic surveys of businesses, and a significant contribution was provided within the EU funded project “Action Plan”. A major objective of ACTION PLAN, involving both Eurostat and the European Central Bank, was reducing time for production and dissemination of an integrated set of information necessary to study the trends of European economy in the short term period. Similarly to what was provided by other European countries, Italy had to provide information for such an integrated dataset, coming from both sampling and total surveys. Such duties were stated in the Regulation for Statistics on Economic Surveys for Business, dated back to August 2005. In this context, it was defined a body of statistical methods to be used for economic business indicators in different surveys. The principal Istat surveys for which experimental studies on the production of preliminary estimates were performed are: Retail sales survey, Wholesale survey and Prodcom survey.

Istat applied some time series models (Rao et al 1989) in which the estimator is expressed as the weighted mean between the final estimate $\theta_{(fe)t}$, referred to the previous time t , and the preliminary cross-section estimate $\theta_{((fe)t+1)}$ at time $(t+1)$ adjusted with the error relative to the previous period t .

The methodology developed is based on a state space model from the time series of preliminary and final estimates obtained by the survey. The methodology proposed by Choi and Varian use both the time series of final estimates obtained for the previous times together with the data obtained from Google Trend as auxiliary exogenous variables of the model. It was launched a research project aimed at integrating the methodology for improving Istat's preliminary estimates, with auxiliary information such as that available from GT.

In this respect it is recalled that Choi and Varian illustrated some simple forecasting methods that improve the basic autoregressive methods. In the cited article, the authors expose a methodology for the construction of early estimates based on the use of weekly data drawn from Google Trends (GT) in terms of query share, normalized to be zero on 1 January 2004. The article shows some interesting examples of the use of big-data in the context in question: (1) sales related to the category NAICS 441 "Motor vehicles and parts dealers" estimated by the Bureau of Census' advance on the basis of Monthly Retail Sales; (2) the data on "U.S. car and light truck sales by make" Automotive Monthly, (3) "new household for sale by price and stage of construction", produced Bureau of the Census, (4) "monthly visitor arrival by country / territory of residence and by entry mode".

As regards the methods of small area estimation, they are intended to improve the reliability of the direct estimates, based on the observed sample units in each domain. A domain is said "small" if the number of sample units that fall in it is not sufficient to produce reliable direct estimates and therefore publishable; this depends, of course, both on the number of units that fall in the domain and on the variability of the phenomenon investigated. Models for small area estimation are utilized in order to improve the reliability of direct estimates. These models "draw strength" from other areas and / or from previous occasions of investigation and use the known population totals of auxiliary variables available from sources external.

With regard to applications in small area estimation Istat (D'Alo et al. 2012; D'Alo et al. 2008) produces the annual average estimates of total occupied and the total number of people in search of employment at the level of local labour market areas. These aggregates are obtained by means of methods of small area estimation based on labour force survey micro data and auxiliary variables known from external sources. To ensure in time a high degree of accuracy to the data products, the statistical model applied, defined by an empirical predictor that exploits the spatial correlation between the area residuals, is constantly validated both on real data and through simulations on census information. We are currently studying extensions of the predictors that exploit both the temporal and spatial correlations of random effects that those type of semiparametric models. Some major projects conducted by Istat on the topic of small area estimation, regard the web system SMART (Small Area Estimation Tool: <http://smart.istat.it/smart/>) which extends the possibilities of estimation of aggregates on employment on areas that may be requested interactively

by users. Another project in this context is the project ESSnet on Small Area Estimators (Grant Agreement of 17 December 2009, duration 24 months), which had as its overall objective to disseminate, among statistical agencies of the European Statistical System, the knowledge on the methods of small area estimation and to provide tools and recommendations that favour the production of estimates through the use of such methods.

On the basis of what has been said Istat has launched an integrated research project aimed at evaluating the potential of big-data for the production of preliminary estimates and estimates for small areas. The project involves the actions mentioned below:

Action 1: (1) modify Istat's methodology to introduce GT auxiliary variables in the time-space model for provisional estimation (2) to study the variables available on Google Trend for the construction of advanced estimators of certain categories of products related to retail, wholesale and PRODCOM survey; (3) the time series of monthly ILO variables can be studied exploiting the GT weekly queries.

Action 2: (1) to study the variables available on Google Trend for the construction of estimates for small areas related to the labour market (employed, people looking for employment, etc. ..); (2) to study the predictive ability of these variables to produce estimates of the labour market (employed and unemployed) on a monthly quarterly regional and provincial level.

Action 3: extend the study conducted in Action 2 to other Istat surveys : (1) Survey trips and vacations, which produces quarterly data, (2) the survey Multipurpose aspects of daily life that produces annual data at the level of the region from 1993 on the socio-demographic variables and health (health services use variables of morbidity + body mass index).

Action 4: explore the use of mobile phone data as auxiliary data for the small-area estimation of poverty, or other specific dimensions of well-being. Besides aggregated data from web search services like GT, it is important to experiment also with mobile phone data, which ensure a finer spatio-temporal resolution (compared with aggregated GT data) and combine both mobility and social network information anchored to geographic location (residence/domicile of users, either directly available or estimated as the most frequent location of user). Action lines include: 1) the investigation of the best proxies based on GSM data for the various dimensions of well-being (especially considering the dimensions of BES – Benessere Equo e Sostenibile) with severe longitudinal studies, and 2) the construction of reliable small area estimates based on GSM data.

3 Results of the study

The overall objective of the project is to exploit the time series query share (qs) extracted from Google Trends (GT) as auxiliary variables to improve the estimates produced by ISTAT through the use of model based estimation methods (early estimates and small area estimation).

Two sets were downloaded from GT:

1. Data from the category 'Job'
2. Data from the search term 'Job offers'

Some results are shown below.

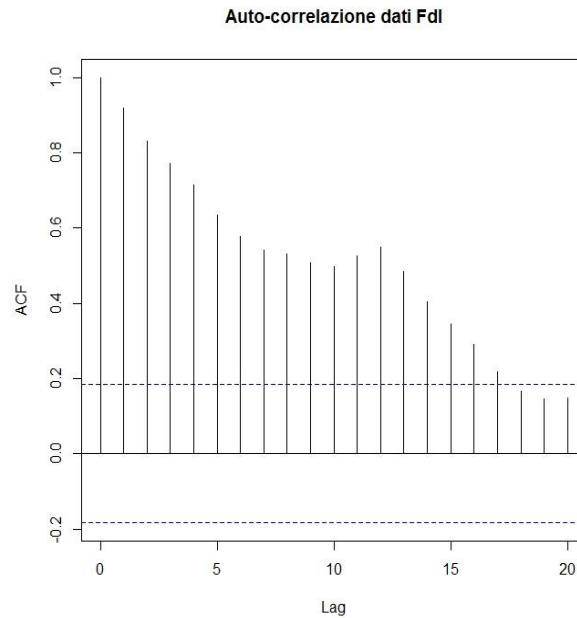


Figure 2: Choice of the time-lag for the AR benchmark model

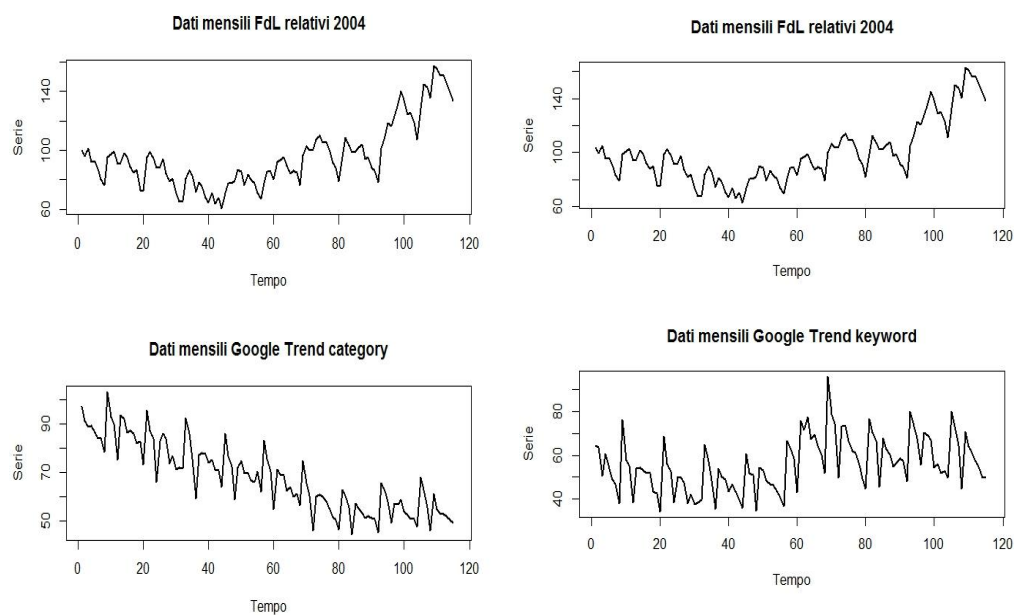


Figure 3: Graphical analysis on the trends of the time series GT and LB

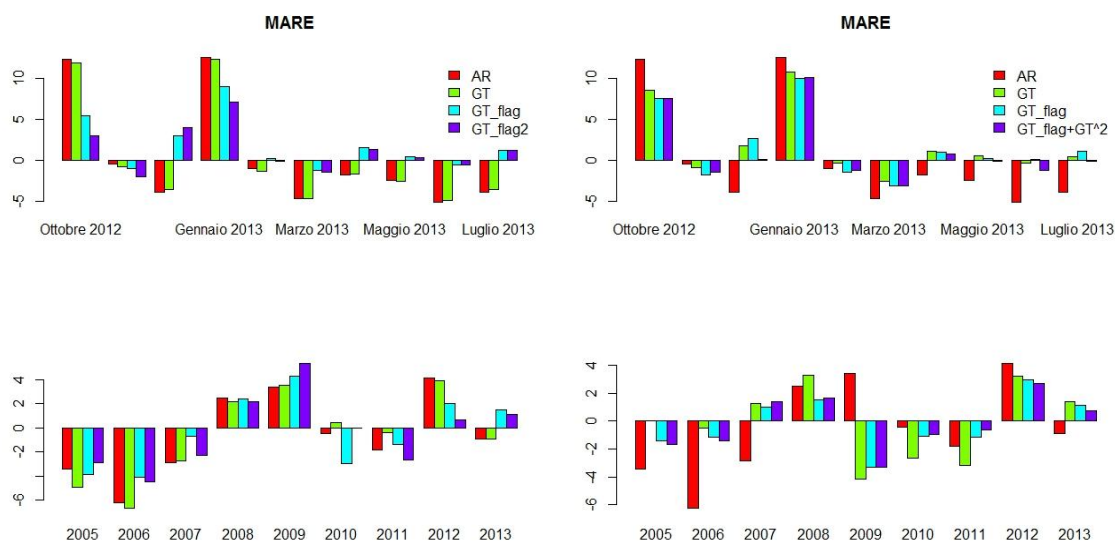


Figure 4: Analysis of alternative models and comparison with the benchmark model

References

1. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, et al. Detecting influenza epidemics using search engine query data. *Nature* 457: 1012–10155. (2008)
2. Choi, H. and Varian, H., Predicting the Present with Google Trends. *Economic Record*, 88: 2–9. doi: 10.1111/j.1475-4932.2012.00809.x (2012)
3. H. Choi and Hal Varian (2009) , “Predicting the present”, Google Inc., Draft Date April 10
4. Askitas, N. and Zimmermann, K. F. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55: 107–120 (2009)
5. Francesco D'Amuri, Juri Marcucci. The Predictive Power of Google Searches in Forecasting Unemployment. Banca D'Italia, Working Papers n. 891 (2012)
6. Pratesi M. , Pedreschi D., Giannotti F., Marchetti S., Salvati N. , Maggino F. (2013), “Small area model-based estimators using big data sources”, NTTS 2013; Brussels
7. D'Alò M., Gismondi R., Solari F., Naccarato A. (2006), Estimation in Repeated Business Surveys using Preliminary Sample Data, Atti della XLIII Riunione Scientifica SIS, 14-16 giugno, Torino
8. D'Alò M., Di Consiglio L., Falorsi S. Solari F., (2008), Small Area Estimation Methods for Socio-Economic Indicators in Households Surveys, *Rivista Internazionale di Scienze Sociali*, Università Cattolica del Sacro Cuore di Milano, Anno CXVI, Ottobre-Dicembre 2008
9. D'Alò M., Di Consiglio L., Falorsi S. Solari F., (2012), Use of spatial information in small area models for unemployment rate estimation at subprovincial areas in Italy, *Journal of the Indian Society of Agricultural Statistics*, Volume 66, December 2012, pp. 1-239