GOOGLE TRENDS FOR NOWCASTING QUARTERLY HOUSEHOLD CONSUMPTION EXPENDITURE

Andrea Fasulo, Alessio Guandalini, Marco D. Terribili

Introduction

Italian National Institute of Statistics (Istat) carries out an important sample survey about Italian households expenditures on goods and services; the observed expenditure is obviously considered in the computation of the aggregates on which national accounts are based, first and foremost the Gross Domestic Product (GDP). For this reason, in the calculation of expenditure estimates, the quickness represents an important issue: estimating these indicators in a swifter way would involve a remarkable improvement of the process which leads to the final statistical indicator. Exploiting innovative data sources potentiality can allow to introduce forecasting and nowcasting in the estimation process.

Big data exploitation is actually a reality in the official statistics field: web data is under study for the production of official statistics indicators, such as unemployment rate, forecasted using data from the web (D'Alo *et al.*, 2015), and consumer price index, calculated using prices collected using webscraping techniques (Polidoro *et al.*, 2015).

The "3Vs" (volume, velocity, variety) definition model for describing big data can lead to better estimates which exploit their informative potentiality, also in complex fields as small area estimation (Falorsi *et al.*, 2017) or forecasting.

One of the most innovative data source used for this purpose is Google Trends, a web facility provided by the web search, which allows every user to evaluate how the research flows of a set of keywords change during an observation period. This longitudinal data can increase the predictive power of a time series baseline model. Google Trend data has many important assets: they are weekly update, they can be referred to one or several keywords and they can be circumscribed in the space (referring to a country or to a region) and in the time (referring to a period).

The paper is organized as follow: in Section 1 the household budget survey carried out by the Italian National Institute of Statistics is described. In Section 2 the Google trend tool is presented and the keywords chosen for catching the trends of expenditure are listed. In section 3 the methodology used for the estimation of

model for nowcasting and the results are discussed. Finally, in Section 4 conclusions and further research.

1. The household budget survey

The household budget survey (HBS)[1] supplies information on the evolution of level and composition of household consumption expenditure, according to their main socioeconomic and territorial characteristics. It provides the official estimates of relative and absolute poverty in Italy and the measure of inflation by household expenditure classes.

The HBS is a continuous sample survey carried out by ISTAT in each week of the whole year. The sample is a two-stage sample of municipalities and household. The municipalities are stratified by typology[2] and population size within regions. The size of the yearly sample consists in around 500 municipalities and 20 thousand households.

The survey collects data on all expenditures incurred by resident household to purchase goods and services exclusively devoted to household consumption. The data are collected through three questionnaires:

- *internal interview*, in which the main socio-demographic characteristics of the household, each member and the house are collected; furthermore, the possession of durable goods, expenditure for transport and communication are reported;
- *daily diary*, in which all the members indicate the expenses incurred during a period of 14 days, the amount of consumed or donated self-produced goods and the place of purchase of goods and services;
- *final interview*, in which information on other expenditure are gathered.

The survey provides data on the household consumption expenditure for different consumption purposes (ECOICOP)[3], geographical areas and household typologies. The yearly estimates are released in July of each year and are related at the consumption expenditure of the previous year $(t-1)$. Furthermore, quarterly estimates are provided to Nation Account for compiling the national accounts

[1] See http://siqual.istat.it/SIQual/visualizza.do?id=0021002.
[2] The municipalities are classified in Metropolitan area, neighborhood of metropolitan area and municipalities with more than 50 thousand inhabitants, municipalities with less than 50 thousand of inhabitants not included in the neighborhood of metropolitan area.
[3] Classification of individual consumption by purpose. For further information see https://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=5.

statistics. This data are available at 45/50 days after the end of the quarter, then with a quarter of delay.

2. Google Trend

The time series query share extracted from Google Trends is used as an auxiliary variable to improve the model-based nowcast.

GT is a web tool that provides sample time series computed as ratio between the number of queries for a keywords at time t weighted with the total volume of searches at time t and the maximum number of queries for the keywords throughout the period considered, weighted with the volume of total searches.

Actually, GT supplies a query index data for a group of ten keywords at most. Therefore, a group of keywords has been selected according to the topics in the HBS interviews. Then, among these, ten keywords with the most powerful trends have been included in the query index data. The list of the keywords used is as follow:

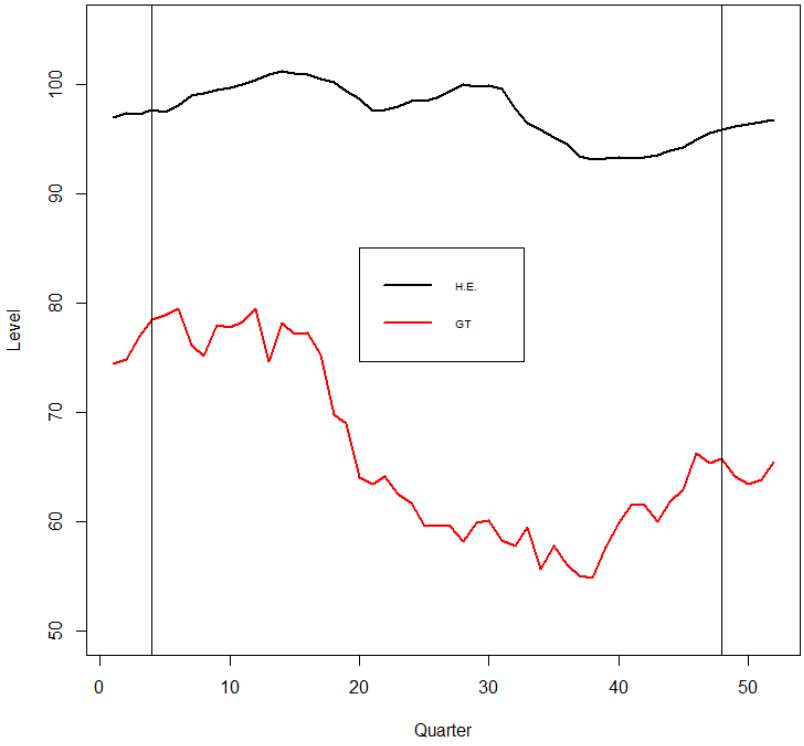
- *abbigliamento* (clothing);
- *bollette* (bills);
- *finanziamento* (financing);
- hotel;
- IKEA;
- *orologi* (watches)*;*
- *prestito* (loan);
- *preventivo* (estimate);
- *ristrutturazione* (refurbishment);
- *voli* (flights).

3. Methodology

In this section, we will describe the relevant methodological statistical approach used for the study. Our statistical model is implemented in R software. The paper focused on the study of nowcasting quarterly households consumption expenditures, using the seasonally adjusted data sample series from first quarter of 2004 to fourth quarter of 2016. Furthermore, GT data for the keywords described in the section 2 is used for the query index data. The GT series has been seasonally adjusted using the X-13ARIMA-SEATS method. Figure 1 shows the structural

trends for the two series, the household consumption expenditure (black line) and the GT series (red line). The time series are index number and the same base is used (3rd quarter of 2006). A very simple baseline forecasting model has been studied, based on the series from first quarter of 2005 to fourth quarter of 2015, modelling the disturbance terms as an Auto Regressive Integrated Moving Average (ARIMA) model with order (1,1,0).

Figure 1 – *Level of household consumption expenditure (H.E.) vs. Google Trend series level (GT). 1st quarter of 2005 – 4th quarter of 2015, Italy.*



Note: time series level are index number and the same base is used (3rd quarter of 2006).

We next add to the baseline model the information coming from GT. The GT predictive power has been validated by comparing the results obtained with the baseline model and other models that take into account important leading indicators of the household consumption series (Vosen & Schmidt, 2011a, 2011b).

In particular has been added information about GDP, interest rate of the Italian Ordinary Treasury Bills (OTB) and Consumer Confidence index (CC). All the auxiliary variables have been tested with different lag structures considering information at the time t, $t-1$, $t-2$, $t-3$ and $t-4$.

The goodness of fit of the models were carried out studying both the residual correlation and distribution by means of the Ljung-Box (8 lags) and the Jarque-Bera test. To evaluate the explanatory power of the different external variables considered in the application and so to compare the quality of each model has been considered also the AIC, BIC and log-likehood values.

The 1-step ahead nowcast of the four quarters of the 2016 has been carried out using a rolling regression procedure for all the models and the results has been compared by means of two classic indicators of forecast accuracy, the Relative Root Mean Squared Error (RRMSE) and the Mean Percentage Error (MPE). The indicators are formulated as follows:

$$RRMSE = 100\frac{\sqrt{\frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T}}}{\bar{y}}$$

$$MPE = \frac{100}{T}\sum_{t=1}^{T}\frac{(y_t - \hat{y}_t)}{y_t}$$

where T are the quarters available, y_t and $\hat{y}_t$ are respectively the actual estimates for the survey and the forecast value at the time t. All the models that will be compared in the next section are summarized in the list below:

- Model 1: the ARIMA benchmark model;
- Model 2: the augmented benchmark model including GT;
- Model 3: the augmented benchmark model including GDP;
- Model 4: the augmented benchmark model including OTB;
- Model 5: the augmented benchmark model including CC.

A significance comparison will be carried out through the Harvey-Leybourne-Newbold (1997) modification of the Diebold-Mariano (1995) test statistic for equal forecast accuracy. This aim of the test is to understand if the differences between the MSEs are statistically significant.

4. Results

The case-study follow the steps identified by the Box-Jeckins approach (Box & Jenkins, 1970). The figure 1 shows an absence of stationarity for the household expenditure consumption series. The presence of a unit root is confirmed by the augmented Dickey-Fuller test (p-value=0.9) and the KPSS stationarity test (p-value<0.01), so a first difference has been made on the data. The model that best fits the correlation and auto-correlation plot is an ARIMA (1,0,0).

The goodness-of-fit of the models is shown in Table 2, in which Model 1 refers to the benchmark model, Model 2 utilizes the lagged GT series, while Model 3-5 include the lagged series respectively of GDP, OTB, and CC.

Table 1 – *Goodness-of-fit results.*

Variables	Model 1	Model 2	Model 3	Model 4	Model 5
AR (1)	0.67***	0.65***	0.68***	0.66***	0.66***
GT $(t-1)$		0.05 **			
GDP $(t-3)$			-0.05		
OTB$(t-3)$				0.002	
CC$(t-3)$					0.0007
AIC	54,50	52,68	52.59	52,08	57,36
BIC	58,22	51,16	57,94	57,43	62,91
Log-Lik	-25,28	-23,34	-23,35	-23,04	-25,68
Residual Diagnostics					
Mean	-0,0030	-0,0004	-0,0010	-0,0080	-0,0040
Skewness	-0,96	-0,87	-0,92	-0,9	-0,96
Kurtosis	1,87	1,47	1,8	2,35	2,04
Jb normal test	0.0003	0,001	0,000009	0,0000004	0,000001
LB test (lag 8)	>0.05	>0.05	>0.05	>0.05	>0.05

Significance codes: 0 ”***” 0.001 ”**” 0.01 ”*” 0.05 ”.” 0.1 ” ” 1

In all the models the AR component is significant and with positive values, while just the GT auxiliary variables shows a significant coefficient (p=<0.05). Model 2 shows very good results in terms of AIC, BIC and log-likehood. Very good performance are provided by Model 4 too, but the OTB coefficient estimate is not significant (p>0.1). Not good results are showed from all the models for the residuals distribution. In fact all the models refuse the Jarque-Bera null hypothesis

of normality (p-value<0.1). Looking the results for the Ljung-Box test is possible to accept for all the models the null hypothesis that the residual are independently distributed, so the correlations in the data, up to 8 lags, are supposed to be 0.

The four quarters of the 2016 have been nowcasted with the five models. Table 3 shows the comparison of the forecast made by means of the RRMSE and MPE indicators. Very interesting results are obtained by Model 2 and Model 5 both in terms of RRMSE and MPE. A direct comparison between the baseline model and the GT model shows that using GT information the Mean Squared Error reduce by 14%. It's important to underline that just the models 1-2 present statistically significant coefficients.

Table 2 – *RRMSE and MPE for 1-step ahead prediction.*

	Model 1	Model 2	Model 3	Model 4	Model 5
RRMSE (%)	0,068	0,063	0,080	0,079	0,060
MPE (%)	0.058	0.054	0.069	0.056	0.049

Significance level is determined in Table 3 through the Harvey-Leybourne-Newbold modification of the Diebold-Mariano test. The table shows that although the Google model shows almost always better results of all the other indicators, the Diebold-Mariano statistics are not significant.

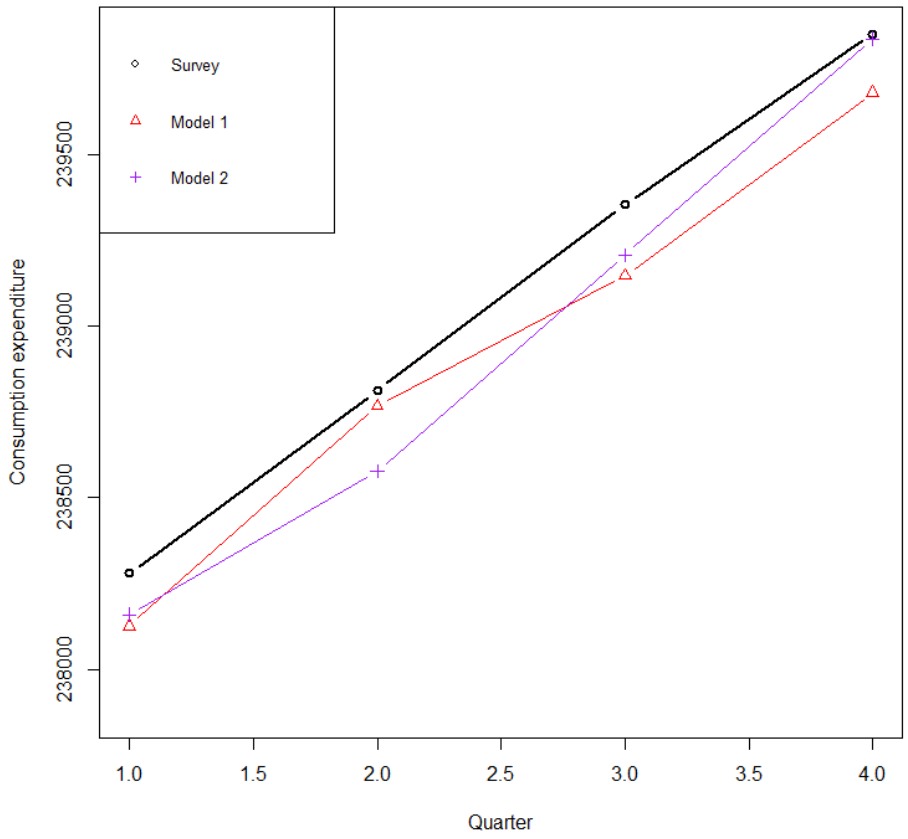
Table 3 – *Diebold-Mariano test.*

	Model 2/Model 3	Model 2/Model 4	Model 2/Model 5
DM test	1.19	-0.14	0.35

Significance codes: ”***” 0.01 ”**” 0.05 ”*” 0.1 ” ” 1

The household expenditure consumption series with the nowcast estimates obtained using Model 1 and Model 2 for 1-step ahead prediction are plotted in Figure 2. Both the models underestimate the true value in all the four quarters studied and the model including GT information outperforms the benchmark model over three quarters of four.

Figure 2 – *1-step ahead nowcast estimates. Four quarters of the 2016, Italy.*



5. Conclusions and further research

Google trends data exploitation which has been shown in this paper represents a promising approach for nowcasting quarterly households expenditures estimation. The proposed models give good results both in terms of predictive capability and in terms of goodness of fit.

Nowcasting would guarantee to fill the informative gap between the issue of estimations and the reference period, providing reliable and useful estimates to the users in a least time.

Further developments to this paper could be a deeper study of the analyzed keywords of which google trends data have been downloaded and exploited, in order to further improve the predictive models.

Moreover applying this method to other statistical field (mainly Gross Domestic Product and Consumer Confidence Indicator) can be interesting and it may lead to results as good as those presented in the present paper.

References

ASKITAS N., ZIMMERMANN K.F. 2009. Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, Vol. 55, No. 2, pp. 107-120.

BORTOLI C., COMBES S. 2015. Contribution from Google Trends for forecasting the short-term economic outlook in France: limited avenues. *Insee – Conjuncture in France*, March 2015, pp. 43-55.

BOX G.E.P., JENKINS G.M. 1970, Time Series Analysis, *Forecasting and Control. Holden Day*, San Francisco

D'ALÒ M., FALORSI S., FASULO A. 2015, Monthly unemployment rate prediction with Google Trends data: does Google search data improve the nowcast of Italian labour market?. *Big Data and the complexity of Labour Market Policies: New Approaches in Regional and Local Labour Market Monitoring for Reducing Skills Mismatches*, Larsen C., Rand S., Schmid A., Mezzanzanica M., Dusi S. (Eds.), Reiner Hamp Verlag. 99-114.

FALORSI S., FASULO A., NACCARATO A., PRATESI M. 2017, 61st World Statistics Congress ISI 2017, *Small Area model for Italian regional monthly estimates of young unemployed using Google Trends Data* https://www.researchgate.net/publication/320554956_Small_Area_model_for_Italian_regional_monthly_estimates_of_young_unemployed_using_Google_Trends_Data

POLIDORO F., GIANNINI R., LO CONTE R., MOSCA S., ROSSETTI F. 2015. Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation. *Statistical Journal of the IAOS*, Vol. 31, no. 2, pp. 165-176, 2015

VOSEN S., SCHMIDT T. 2011a. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, Vol. 30, No. 6, pp. 565-578.

VOSEN S., SCHMIDT T. 2011b. A monthly consumption indicator for Germany based on internet search query data. *Ruhr Economic Papers*, No. 208. ISSN: 1864-4872 (online).

SUMMARY

Google Trends for Nowcasting Quarterly Household Expenditure

During last years, several studies focused on the predictive capability of web data to forecast statistical indicators. Google Trends is a free web tool that quantify search-term volume on the search engine. The aim of this work is to forecast the household expenditures for consumption in Italy, using Google Trends related to particular expenditure keywords. Several prediction models have been tested, also including relevant leading indicators correlated to the household expenditures behavior, on a time series survey data from 2004 to 2016. The ARIMA model has been chosen and models with different lag structures have been tested. The model comparison shows that the results including Google data outperform those of both benchmark and augmented models.

Andrea FASULO, Italian National Institute of Statistics, fasulo@istat.it
Alessio GUANDALINI, Italian National Institute of Statistics, alessio.guandalini@istat.it
Marco D. TERRIBILI, Italian National Institute of Statistics, terribili@istat.it