ROMA TRE
UNIVERSITÀ DEGLI STUDI

COLLANA DEL
DIPARTIMENTO DI ECONOMIA

USING GOOGLE TREND DATA TO PREDICT THE ITALIAN UNEMPLOYMENT RATE

Stefano Falorsi - Alessia Naccarato - Andrea Pierini

ISSN 2279-6916 Working papers
(Dipartimento di Economia Università degli studi Roma Tre) (online)

Working Paper n° 203, 2015

I Working Papers del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all'approvazione del Comitato Scientifico.

Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall'art. l del D.L.L. 31.8.1945, n. 660 e successive modifiche.

Copie della presente pubblicazione possono essere richieste alla Redazione.

esemplare fuori commercio
ai sensi della legge 14 aprile 2004 n.106

REDAZIONE:
Dipartimento di Economia
Università degli Studi Roma Tre
Via Silvio D'Amico, 77 - 00145 Roma
Tel. 0039-06-57335655 fax 0039-06-57335771
E-mail: dip_eco@uniroma3.it
http://dipeco.uniroma3.it

ROMA TRE
UNIVERSITÀ DEGLI STUDI

DIPARTIMENTO DI ECONOMIA

USING GOOGLE TREND DATA TO PREDICT THE ITALIAN UNEMPLOYMENT RATE

Stefano Falorsi - Alessia Naccarato - Andrea Pierini

Using Google Trend Data to predict the Italian Unemployment Rate[1]

S. Falorsi[*]*, A. Naccarato,*[**] *A. Pierini*[**]

Abstract
The increased availability of online information in recent years has aroused interest as to the possibility of deriving indications on phenomena under studies. In the more specifically economic and statistical context, numerous studies suggest the use of online search data to improve the nowcasting and forecasting of the official economic indicators with a view to increasing the promptness of their circulation. In the same way, this paper puts forward a model for multiple time series that harnesses cointegration of the official time series of the Italian unemployment rate and the series of the Google Trend job offers query share to nowcast the monthly unemployment rate. *Nowcasting* is to be understood here as estimating the monthly unemployment rate for the month in which official survey is actually under way. The aim is thus to assess whether the use of Internet search data can improve the nowcasting of the economic indicator considered.

Keywords multivariate time series analysis, preliminary estimates, online search data

JEL CODE: C130, C320, C530

Introduction

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 can prove useful in the study of social phenomena (Askitas and Zimmermann, 2015). This is so because the information that people disclose about their needs through use of the Internet (Ettredge et al., 2005) can shed light on the variability of numerous phenomena under examination.

To this end, many authors (Choi and Varian, 2009a, 2009b; D'Amuri, 2009; Della Penna and Huang, 2009; Guzman, 2011; McLaren and Shanbhougue, 2011; Vosen and Schimdt, 2011, 2012a, 2012b; Wu and Brynjolfsson, 2010; Goel et al., 2010) propose the use of online search data to improving the nowcasting and forecasting of official economic indicators, which are normally published a certain time after the period to which they refer. In particular, Choi and Varian (2009a) introduce the term "nowcasting" or "predicting the present" in suggesting the use of Internet search data to improve short-term forecasts of economic indicators.

Consider, for example, the monthly surveys on prices, inflation and industrial production, for which the gathering of data takes place all through the month in question. The data quality control procedures and calculation of estimates are then carried out at the end of the month after

[1] Thanks are due to Prof. Luciano Pieraccini, for valuable suggestions on this article and for his constructive comments

[*] ISTAT – Italian National Statistical Institute
[**] Department of Economics, Roma Tre University

receiving the questionnaires completed by the sample units, and the results are therefore generally made available with a some delay with respect to the end of the month of reference.

It can 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 of the indicator for which the survey is being carried out in the same period.

The problem is therefore to produce estimates that are available while the actual survey is still under way – or in any case in a short space of time – so as to provide reliable information on the indicator at the current month and its variation with respect to the previous one (D'Alò et al., 2006; Naccarato et al., 2006; Barrow, 2004; Brown et al., 2009). It should be noted that for Italy as well as the other European countries, the information required by the European Community for the purpose of economic analysis is vast and drawn from a variety of large-scale surveys based on both samples and censuses. These requirements are also laid down in the EC regulation on short-term business statistics in force as from August 2005, which date saw the launch of the Action Plan on 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 essential to analysis of the short-term trend of the European economy.

Google Trend data is used in this paper as auxiliary information to nowcast the Italian rate of unemployment. Numerous authors suggest the use of Internet data to forecast unemployment (Anvik Gjelstad, 2010; Askitas and Zimmermann, 2009; D'Amuri, 2009; D'Amuri and Marcucci, 2009; Fondeur and Karamè, 2013; McLaren and Shanbhougue, 2011; Shuoy, 2009) and the results, mostly obtained by the use of models for univariate time series, show that they can be regarded as useful in the estimation procedure.

The methodology put forward here is that of models for multiple time series with a view to exploiting cointegration of the official time series of unemployment rate and the Google Trend series for the query share "*offerte di lavoro*" (job offers).

This approach can be regarded as highly economical as it entails no recourse to additional resources but is based essentially on identification of the statistical models that produce the best nowcast of the unemployment rate solely on the basis of the information available by a certain date.

The auxiliary information used, namely the Google Trend series is located in the sphere of what is known in the literature as Big Data (Einav and Levin, 2014; Daas et al., 2013; Ceron et al.,

2013; Choi And Varian, 2009b; Vanderkam et al., 2011; Mohebbi et al., 2011; Chamberlin, 2010). In addition to their size, these data have the characteristic of being immediately available and are capable of supplying an up-to-date representation of social and economic phenomena. Moreover, they can generally be used at very low cost, which constitutes a further reason for interest in the sphere of official statistical surveys.

If they are to be used effectively, however, it is necessary to develop *ad hoc* methodologies making it possible to extract from an enormous mass of information the data relevant to the phenomenon under examination in compliance with the quality standards laid down by the official departments of statistics.

Given that the GT information is subjected to no form of quality control, it is our view that it can be used solely as a sort of "snapshot" providing indications about the phenomenon of interest and its evolution in time and space.

In other words, the direct inclusion of information on the phenomenon of interest from sources outside the official surveys can create major problems as regards the quality and accuracy of estimates. On the other hand, however, taking advantage of information about the spatial and temporal dynamics of a phenomenon closely related to the one of interest and readily available at little cost may well offer a good opportunity to produce preliminary estimates that can be corrected as necessary when all the phases of the official survey have been completed. 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 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.

Our aim is to nowcast the unemployment rate, for which purpose use is made of a model for cointegrated time series (Engle and Granger, 1987) known in the literature as the Vector Error Correction model (VEC). The time series considered are the monthly unemployment rate in Italy (LF) produced by the ISTAT- Italian National Statistical Institute and the query share series (GT) produced by Google Trend for the keyword "*offerte di lavoro*" (job offers).

If *t* is the month of the Labour Force survey (Istat, 2014), it is reasonable to assume that at the end of the month *t* of reference – once the data drawn directly from the sample units has been gathered – a further period of time is required in order to produce an estimate of the unemployment rate. This is generally made available about one month after *t* to which it refers, as it is known that the phases subsequent to the gathering of data from the sample units can take a considerable amount of time. The aim of producing a nowcast estimate of the unemployment rate at time *t* may arise from the need to carry out analyses of trends and variations or from requirements of economic planning before all the phases of the official survey have been completed (D'Elia, 2014). The nowcast estimate, which must in any case comply with the fixed levels of accuracy, can then be corrected and/or replaced once all the phases of the official survey have been completed.

This paper develops a nowcast of the unemployment rate by estimating a VEC model based on cointegration between LF series and the G T series observed up to the time t-1.

The cointegration between the two series, i.e. the fact that they present the same trend in the long run, constitutes an element of improvement of the model in terms of the accuracy with respect to the use of models for univariate time series (D'Amuri and Marcucci, 2009; Vicente et al., 2015, Bacchini et al., 2014). The latter are in fact models that take into consideration solely the autoregressive and/or purely explanatory effects of the phenomena in question and ignore any cross-correlation or cointegration of more than one time series.

Excluding *a priori*, in the phase of model selection, the existence of any component of cross-correlation and/or cointegration between the series can result in a major loss of information and the choice of an unsuitable model.

The cointegration formally expresses the way in which two or more time series evolve together in the long period, and is therefore a necessary representation of the relationship between the series. It is necessary because it is precisely the observation of reality that makes it natural to expect, as in our case, a connection between job offers query share and the unemployment rate disseminated by Istat through a sample-based survey. In other words, it is natural to expect that the two time series, considered separately in their temporal evolution, will provide information on one another, and this characteristic, once ascertained by means of suitable tests, must therefore be formally represented in the model through the cointegration.

In order to ascertain the utility of using the GT information, the results thus obtained were compared with those based exclusively on the official time series for the monthly unemployment rate.

The results obtained by means of the VEC model were compared with those of an autoregressive integrated moving average (ARIMA) model for the unemployment rate series (Chadvick and Sengul, 2012).

The paper is organised as follows. Section 2 presents the data used and their temporal characteristics, section 3 the VEC model adopted and the statistical tests used to ascertain its validity, section 4 the results obtained in terms of a nowcast of the unemployment rate, section 5 the comparison with the benchmark model, and section 6 some concluding observations.

2. Description of the series

The aim of the paper is to develop obtain a nowcast of the monthly unemployment rate in Italy (D'Amuri Marcucci, 2009; Askitas and Zimmermann, 2009; Choi and Varian, 2009; Barreira et al., 2013; Vicente et al., 2015) using the cointegration with the GT series.

Produced on a monthly basis since 2004 by the Istat Forze Lavoro survey, this indicator is generally made available about a month after the period to which it refers. It may therefore become usefull to develop at time *t*, while the monthly survey is still being carried out, a forecast of the unemployment rate available by the end of the same month.

The nowcast does not therefore replace the estimate produced by the survey but provides a preview of the indicator in a shorter time serving only to suggest the possible trends.

The input data used in the application presented here are the monthly time series LF produced by Istat in the period January 2004 – September 2014 and the monthly time series GT.

The weekly GT data were aggregated into monthly averages according to the LF survey scheme of allocation of weeks of investigation to months.

D'Amuri (2009) and D'Amuri and Marcucci (2009) justify the use of the keyword "Offerte di lavoro" for its popularity among competitor job-search related keywords. Infact it is possible to verify its popularity comparing the relative incidences of it *vs* other search terms. Furthermore, its broad definition makes it ideally robust against irrelevant strong variations connected to demand or supply side of labor force of specific subgroups of job seekers. On the contrary it has to take into account that not all workers have access to the internet, nonetheless this is a minor issue, given the increasing use of internet as a job search method. Furthermore, the two LF montly series

refered respectively to the overall sample and to sub-sample of internet job seekers, show strong similarities.

Both the series considered, LF and GT, were seasonally adjusted by means of an ARIMA (1.0.1) procedure to eliminate fluctuations of a seasonal character of the phenomena considered. Reference will therefore be made solely to the seasonally adjusted series from now on.

3. The VEC model for the unemployment rate nowcast

The availability of a time series for a phenomenon of interest naturally prompts the idea of predicting its future values on the basis of the information as to its past realizations. This is the idea underlying the well-known class of autoregressive models (Hamilton, 1994, pp. 43–71). In cases where more than one time series is available and it is reasonable to assume that they present a quite similar variability, it is necessary to make use of the models for multiple time series known in the literature as vector autoregressive model (VAR) (Lutkephol, 2007). It is in fact important to consider not only the autoregressive effects of each time series but also the effects due to cross-correlation between the different series.

The case in point is precisely one of multiple time series, as the time series of the unemployment rate and of the "job offers" query share are both available and present very similar characteristics of variability (fig. 1). Estimation of the parameters of the model defining each time series must therefore take place simultaneously, i.e. by means of a vectorial model.

The basic idea is obviously that an increase in the unemployment rate is accompanied by an increase in the number of people seeking employment opportunities also through online searches. It is also true that a decrease in unemployment may prompt some people to seek new opportunities on the Internet even though they have jobs. In other words, it is reasonable to assume that the temporal dynamics are closely linked.

In general, $\mathrm{Y_t} = (\mathrm{Y_{1t}},\ \mathrm{Y_{2t}}, \dots, \mathrm{Y_{Kt}})'$ is defined as a vector of K random variables for the time $\mathrm{t} = 1, 2, \dots, \mathrm{T}$ and it is assumed that this verifies the following equation of the Vector Autoregressive Model of order p (VAR(p))

$$Y_t = \mu + A_1 Y_{t-1} + A_2 Y_{t-2} + \cdots + A_p Y_{t-p} + \varepsilon_t \qquad (1)$$

where A_i $(i = 1, \dots, p)$ are matrices of coefficients, μ is a vector of constants and $\varepsilon_t \sim N(0, \Sigma)$.

In other words, it is assumed that the K series making up the random vector $\mathrm{Y_t}$ are correlated with one another. Equation (1) implies that a structure of correlation to the various temporal instants $t-h\ \ (h=1,\dots,p)$ exists between the K components of the random vector $\mathrm{Y_t}$.

In cases where the series are also cointegrated, it is necessary to reformulate (1) in order to make the relations of cointegration explicit, i.e. to go from estimation of a VAR model to estimation of a vector error correction (VEC) model (Engle and Granger, 1987).

It should be recalled that $\mathrm{Y_t}$ is said to be cointegrated if there exists at least one vector β_i such that $\beta_i'\,Y_t$ is stationary, i.e. has its first two moments constant and finite in time. The vector β_i is called the vector of cointegration. If there are r linearly independent vectors β_i such that $\beta_i'\,Y_t$ is stationary, then $\mathrm{Y_t}$ is said to be cointegrated of rank r. In this case, equation (1) must be reformulated so as to take the cointegration into account and thus becomes:

$$\Delta Y_t = \mu + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \cdots + \Gamma_p \Delta Y_{t-p} + \Pi_1 Y_{t-1} + \varepsilon_t \qquad (2)$$

where $\Gamma_i, (i=1,\dots,p)$ are the autoregressive effects and $\Pi_1 Y_{t-1}$ represents the long-period relation, i.e. the combined effect of the relations of cointegration.

Equation (2) indicates that the variation of $\mathrm{Y_t}$ between time $t-1$ and t is given for each of the components of the random vector $\mathrm{Y_t}$ by the linear combination of the autoregressive effects of cross-correlation until time $t-p$ and by the effects of cointegration at time $t-1$.

Equation (2) indicates that the variation of $\mathrm{Y_t}$ between time $t-1$ and time t is given for each of the components of the random vector $\mathrm{Y_t}$ by the linear combination of the autoregressive effects of cross-correlation until time $t-p$ and by the effects of cointegration at time $t-1$.

The autoregressive effects of cross-correlation are regarded as generally transitory (in this short period sense), as they disappear with an increase in the time lag considered, while the effects of cointegration are to be considered permanent (in this long-period sense), i.e. they always exist, thus defining way the structure that temporally links the time series under examination.

Since just two time series are considered here, reference will be made from now on to the two-dimensional random vector $Y_t = (Y_{1t}, Y_{2t})$ where Y_{1t} indicates the time series LF and Y_{2t} GT.

The first step in the estimation of the model (2) is deciding on the VAR order in equation (1): the order selection is a trade-off between two aspects: the correspondence of the model to the data observed (measured by means of the log-likelihood value) and the efficiency of the

estimates, assessed by bearing in mind that a model with fewer parameters produces a more efficient estimate (Cappuccio and Orsi, 2005).

The next step to estimate the model (2) is testing the hypothesis of cointegration (Johansen, 1991). In our case the VAR order is equal to 2 and the series are cointegrated of order 1. Equation (2) therefore becomes

$$\Delta Y_t = \mu + \Gamma_1 \Delta Y_{t-1} + \Pi_1 Y_{t-1} + \varepsilon_t \qquad (3)$$

Equation (3) indicates that the variation in the unemployment rate Y_{1t} between $t-1$ and t is a function of the information on it available until $t-1$ and is a function of GT serie (Y_{2t}) until $t-1$. The basic idea of the VEC model (3) is that exists a condition of equilibrium between the $K=2$ components of the random vector Y_t

$$\beta' Y_t = \beta_1 Y_{1t} + \beta_2 Y_{2t} = 0 \qquad (4)$$

from which

$$Y_{1t} = -\left(\frac{\beta_1}{\beta_2}\right) Y_{2t} \qquad (5)$$

In actual application, condition (4) is not exactly satisfied, in general there is an accidental shock z_t such that $\beta_1 Y_{1t} + \beta_2 Y_{2t} = z_t$ and (5) thus becomes

$$Y_{1t} = {(z_t - \beta_2 Y_{2t})}\Big/{\beta_1} \qquad (6)$$

Bearing in mind (6), we shall now consider the matrix Π_1, which represents the parameters related to the effects of cointegration in (2). It is a square matrix of non-full rank r . Every matrix of this type (Searle 1982) can always be factorized as

$$\Pi_{1\,(K,K)} = \alpha_{(K,r)} \beta'_{(r,K)} \qquad (7)$$

The non-uniqueness of factorization (7) (Searle 1982) gives rise to the problem of the non-uniqueness of the cointegration relations (Lutkephol, 2007, p. 249). It is therefore necessary to impose constraints on the elements of the vector $\beta = \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}$ in order to obtain estimates of the

model's parameters. The solution most frequently adopted in the literature (Lutkepohl p. 250) is to make the first r components of the vector β equal to 1.

In our case, $r = 1$. As a result $\beta_1 = 1$ and (6) becomes

$$Y_{1t} = z_t - \beta_2 Y_{2t} \qquad (8)$$

The above equation (8) indicates the existence for every t of a "misalignment" between Y_{1t} and Y_{2t} given by z_t. Putting forward a model of the VEC type therefore means assuming that this misalignment – occurring prior to t – is involved together with the autoregressive effects in determine the variation between $t-1$ and t of both the components of the random vector.

4. Results

The procedure leading to the estimation of the VEC model for the two series LF and GT can be shown in the phases described below.

First, an order of lag was identified for the VAR model. The choice of an order equal to 2 (VAR(2)) was made by comparing VAR models with different orders of lag on the basis of a trade-off between the number of unknown parameters and the value of log-likelihood. In our case, the model chosen presents a log-likelihood value equal to -786.855, a number with 10 unknown parameters and an AIC value equal to 1598.66 (Akaike, 1974).

Actually, the optimal order of the VAR model established only on the basis of the Akaike criterion (AIC=1562,98), would have been equal to 5, which would have entailed a number of parameters equal to 12 and a log-likelihood value of -759.489. Since the values of log-likelihood and AIC are similar between the two models (VAR(2) and VAR(5)), the one with fewer parameters is preferred for reasons of parsimony.

Once the autoregressive order had been established, the hypothesis of cointegration was verified by means of the Johansen test (1995) based on the rank r of the matrix Π_1. It is a sequential test, which starts by ascertaining whether the rank of the matrix Π_1 is equal to 0. If this hypothesis is rejected, others are put forward in sequence until a certain value of the rank of the matrix is accepted.

Tab. 1 shows the results of the test carried out on the matrix Π in (3). The first column shows the hypotheses tested, the second one the values of the Johansen statistics, the third the value of the Johansen statistics distribution corresponding to the 95th percentile

Table 1 - Johansen Cointegration Test

NullHypothesis	Johansen Statistics (JS)	95° percentile(*)
$r = 0$	26,50	17.95
$r \leq 1$	5,85	8.18

(*) refers to the 95th percentile of the tabulated Johansen distribution.

While the test leads to rejection of the hypothesis $r = 0$ since the JS value is higher than that of the 95th percentile, the hypothesis that $r = 1$ cannot be rejected (the JS value is lower than that of the 95th percentile).

It is therefore concluded that the series LF and GT are cointegrated of order 1 and the estimate of maximum likelihood of the parameters of equation (3) are shown in Tab. 2; it has to notice that the elements of Π_1 are indicate with α and β because of (7).

Table 2 - Maximum Likelihood Estimation of parameters of the VEC Model (3)

Parameters	Estimates	Level of Significance
μ_1	0,2829	0,6022
μ_2	-0,3079	0,5353
γ_{11}	-0,4519	0,0000
γ_{12}	-0,0579	0,4613
γ_{21}	0,0753	0,4626
γ_{22}	-0,1475	0,1072
β_1	1	
β_2	4,332	
α_1	-0,0043	0,0133
α_2	0,0449	0,0055

It should be observed that the parameters γ_{11} and γ_{22}, which represent the effects of the autoregressive component of order 1 for each of the two series, both prove to be significant. On the contrary, the extradiagonal elements γ_{12} and γ_{21} representing the cross-correlation of the two series do not prove significant. The choice of a vectorial model appears appropriate nevertheless, as the parameters in α and β associated with the cointegration component do prove significant.

It is hardly necessary to point out that the significance of the parameter β_2 was previously ascertained on the basis of the cointegration test shown in the Tab. 1. In particular, in the case under examination ($K = 2,\ r(\Pi_1) = 1$) the null hypothesis thus tested is $\beta_2 \neq 0$.

The test is not performed for the parameter β_1 because it is constrained equal to 1 in order to solve the problem of the non-uniqueness of the solution (see section 3).

5. Nowcasting the unemployment rate

The goal of our work was to obtain a nowcast of the unemployment rate for the month in which the official sample-based survey is still under way. For this reason, the VEC model was estimated excluding the last observation, corresponding to September 2014, which was then forecast by means of the model.

The results obtained are compared with those provided by a model that makes no use of GT auxiliary information.

In other words, the VEC model estimated on the two series LF and GT is compared with a model based solely on time data regarding the unemployment rate.

An ARIMA (1, 1, 2) model, selected on the basis of the AIC criterion from among the possible univariate models for the series LF, was adopted for the latter. If B is the backward operator and $BY_{1t} = Y_{1t-1}$, the ARIMA model employed is

$$(1 - B)(1 - a_1 B)Y_{1t} = (1 + b_1 B + b_2 B^2)\varepsilon_t \qquad (9)$$

where a_1 is the coefficient of the autoregressive component a (AR) and b_1 and b_2 are the coefficients of the moving-average (MA) component.

Equation (9), estimated on the LF series of the unemployment rate, proves to be

$$(1 - B)(1 - 0{,}562B)Y_{1t} = (1 - 1{,}066B + 0{,}459B^2) \qquad (9a)$$

with a value of log-likelihood equal to -406,27, and a value of the AIC index equal to 820.544.

Table 3 shows and compares the results of the nowcast obtained with both models (3) and (9a).

Table 3 - Actual and predicted values of the unemployment rate for the month of September 2014 as estimated with the VEC and ARIMA models

Model	Actual	Nowcast	Lower bound	Upperbound	MSE

VEC	7,192	5,390	-6,221	17,001	3,247
ARIMA	7,192	2,776	-8,828	14,381	19,499

The mean square error (MSE) of the estimate obtained with the VEC model is lower than that obtained with the ARIMA model (with a gain equal to 83% = (19.499-3.247)/19.499). In other terms the nowcast obtained with the help of the GT series is more accurate than the one obtained by means of the official time series alone.

In order to evaluate the nowcasting ability of the two models over a longer period a rolling regression (D'Amuri, 2009) was carried out.

This procedure consists of performing month after month the nowcasts of the series of interest and inserting the value actually observed each time before going on to the next.

The first two years (from January 2004 to December 2005) of LF and GT series were first considered in order to estimate the two models. A nowcast for the month of January 2006 was then performed by means of the fitted models. This procedure was repeated inserting the real value each time month after month and performing the following month's nowcast for each of the months from February 2006 to September 2014 (approximately 100). In 63% of the months, the absolute error (i.e. the difference between the nowcast and actual value) for the VEC model proves lower than that obtained with the ARIMA model. Figure 3 shows the series observed and the series nowcast by the two models for the period of interest and Figure 4 shows the series of the absolute error for the two model.

In order to obtain a global evaluation for the comparison between the proposed model and the benchmark one, the mean absolute relative error (MARE) was estimated for each of the two models on the basis of the results of the rolling regression

$$MARE = \frac{\sum_{i=1}^{T} \left| \frac{\hat{y}_i - y_i}{y_i} \right|}{T} \times 100$$

where $\hat{y}_i$ is the nowcast estimate of the rate of unemployment in month i and y_i the estimate carried out by means of the official survey for the same month.

As shown in Figure 1, both the series initially present a falling trend that then begins to rise. It was therefore decided to identify the month in which this change becomes statistically significant by means of an appropriate test of hypotheses; in particular the applied test for structural change (Zeileis et al., 2002) identified a significant structural break in January 2007 (Figure 5).

Then the MARE index was calculated both on all the observations of the period considered and by splitting the series into two blocks (Table 4). The results leads to the conclusion that use of the GT series produces more accurate estimates in short-period forecasting.

Table 4 – Mean Absolute Relative Error

	MARE	
	VEC	ARIMA
Jan 2004–Jan 2007	0,4877421	0,6573869
Jan 2007–Sept 2014	1,130606	1,245186
Jan 2004 – Sept 2014	1,213644	1,217137

6. Conclusions

The interval between the conclusion of the official survey gathering information on the monthly unemployment rate and the calculation of the rate itself can give rise to difficulties in the publication of the data by the national statistical offices and slow down both communication of the said data to the main national and international centre of statistical production and assessment useful for economic policies subsequent to normative intervention on the labour market. Hence the interest of official statistics in the nowcasting of the unemployment rate.

It can therefore prove useful in order to bridge the short lag between the end of the official survey and the calculation of the unemployment rate temporarily and hence make immediate or simultaneous appraisals of the phenomenon possible.

In recent years the enormous availability of Internet search data has prompted researchers to examine the possibility of harnessing them in order to forecast economic indicators. This is due not only to the fact that such data are plentiful, digitally organized, accessible and economical but also and above all to the fact that – since use of the Internet has been common everyday practice for many years now – this information provides an “instant” snapshot of the trends of individuals behaviour that may be partial but reflects the reality.

It therefore appears reasonable to seek indications from Internet search data when examining economic phenomena whose variability has an almost simultaneous effect also on the variability of associated phenomena of social nature.

This paper examines whether the Google Trend series on the amount of “job offers” handled through the Internet can prove useful in nowcasting the unemployment rate. Its conclusion is that the GT series provides a real and immediate image of the level of the

unemployment rate and therefore constitutes a useful tool with a view to obtaining a reliable nowcast of this indicator if used together with the official time series.

The results obtained suggest that this approach can be adopted, as the model put forward furnishes more accurate estimates than those of a model that makes no use of the GT series and is instead based solely on the standard data produced by the Istituto Nazionale di Statistica.

Two final remarks should be made in this connection. First, the GT series plays a part in the nowcasting process because of the choice to harness its cointegration with the LF series, which gives a formal representation of the fact that the "movement in unison" of the two series is a statistically significant and lasting characteristic and therefore highly informative. Second, it is necessary to recall that the model proposed was applied to the unemployment rate in Italy, which presents situations of marked variability, especially in 2014, making it difficult to isolate the erratic component effectively from the structural one due to the economic crisis that has hit Italy harder than other European countries in the last few years (*Nota mensile sull'economia italiana*: http://www.istat.it/it/archivio/144185). These characteristics generally make it more difficult to model and forecast the unemployment rate, above all in the short period.

In the light of these considerations, it appears possible to conclude that the use of Internet data – together with a suitable forecasting model – is a good way to obtain reliable nowcasts of the unemployment rate for use until the official monthly estimate has been produced.

It should finally be pointed out that, given the nature of the GT and the LF data, the procedure put forward here is replicable for small-scale contexts such as particular categories of labour and/or specific territorial and temporal spheres.

References

[1] Akaike H., 1974 , A new look at the statistical model identification, IEEE Transactions on Automatic Control, Volume 19, Issue 6, pp. 716–723.
[2] Askitas N., Zimmermann K. F., 2015, The Internet as a Data Source for Advancement in Social Sciences, ForschungsinstitutzurZukunft der Arbeit, Institute for the Study of Labor, IZA DP No.8899.
[3] Askitas N., Zimmermann K. F., 2009, Google Econometrics and Unemployment Forecasting, ForschungsinstitutzurZukunft der Arbeit, Institute for the Study of Labor, Discussion paper n°4201.
[4] Anvik C., Gjelstad K., 2010, Just Google it. Forecasting Norwegian unemployment figures with web queries, Working Paper No. 11, Centre for Research in Economics and Management.
[5] Bacchini F., D'Alò M., Falorsi S., Fasulo A., PappalardoC., 2014, Does Google index improve the forecast of Italian labour market?, Proceedings of 47th Scientific Meeting of the Italian Statistical Society.

[6] Barreira N., Godinho P., Melo P., 2013, Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends, NETNOMICS: Economic Research and Electronic Networking, Volume 14, Issue 3, pp. 129-165.
[7] Barrow R., 2004, Using Administrative data in Short term Statistics: Sub annual industry surveys at Statistics New Zeland, OECD, STES Timeliness Framework: Preliminary Estimates Based on Statistical Models, avaliable online at http://www.oecd.org/std/fin-stats/32108022.pdf.
[8] Brown G., Buccellato T., Chamberlin G., Chowdhury S. D., Youll R., 2009, Undestanding the quality of early estimates of Gross Domestic Product, Office for National Statistics, avaliable online at file:///Users/ /Downloads/QualityofEarlyGDP_tcm77-168233.pdf
[9] Cappuccio N., Orsi R., Econometria, 2005, Ed. Il Mulino, Bologna, Italy.
[10] Casella G., Berger R. L., 2002, Statistical Inference, 2nd Edition, Duxbury Advanced Series, Thomson Learning, US.
[11] Ceron, A., Curini, L., Iacus, S.M., Porro, G., 2013, Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to Italy and France, New Media & Society, SAGE Journals.
[12] Chadwick M. G., Sengul G., 2012, Nowcasting Unemployment Rate in Turkey : Let's Ask Google, avaliable from https://ideas.repec.org/p/tcb/wpaper/1218.html.
[13] Chamberlin G., 2010, Googling the present, Economic&Labour Market Review, Volume 4, Issue 12, pp. 59-95.
[14] Choi H., Varian H., 2009a, Predicting the present with google trends, Draft Date April 10, avaliable online at file:
[15] Choi H., Varian H., 2009b, Predicting inizial claims for unemployment insurance using Google Trends. Technical Report, Google, avaliable online at file
http://research.google.com/archive/papers/initialclaimsUS.pdf
[16] Citro C. F., 2014, From multiple modes for surveys to multiple data sources for estimate, Survey Methodology, Volume 40, Issue 2, pp. 137-161.
[17] Daas P.J.H., Puts M.J., Buelens B., Van Den Hurk P. A. M., 2013, Big Data and Official Statistics, Paper for the 2013 New Techniques and Technologies for Statistics conference, Brussels, Belgium, avaliable online at file: www.unece.org/.../stats/.../2013/Topic_4_Daas.pdf
[18] D'Alò M., Gismondi R., Naccarato A., Solari F., 2006, Estimation in Repeated Business Surveys using Preliminary Sample Data, In: Proceeding of the XLIII Scientific Meeting of the Italian Statistical Society. p. 333-339, Torino, June 14-16, 2006.
[19] D'Amuri F., 2009, Predicting unemployment in short samples with internet job search query data, MPRA - Munich Personal RePEC Archive, n°18403.
[20] D'Amuri F., Marcucci J., 2009, Google it! Forecasting the US unemployment rate with a Google Job Search Index. MPRA Paper No. 18732.
[21] D'Elia E., 2014, Predictions vs. Preliminary Sample Estimates: The Case of Eurozone Quarterly GDP, Journal of Official Statistics. Volume 30, Issue 3, pp. 499-520.
[22] Della Penna N., Huang H., 2009, Constructing consumer sentiment index for US using Google searches, Working papers n°26, University of Alberta, Department of Economics.
[23] Engle R. F., Granger C. W. J., 1987, Co-integration and error correction: Representation, estimation and testing, Econometrica, Volume 55, Issue 2, pp. 251–276.
[24] Einav L., Levin J., 2014, Economics in the age of big data, Science, Volume 346, n° 6210.
[25] Ettredge M., Gerdes J., Karuga G., 2005, Using Web-based search data to predict macroeconomics statistics, Communication of the ACM, Volume 48, Issue 11, pp. 87-92.
[26] Fondeur Y., Karamè F., 2013, Can Google data help predict French youth unemployment? Economic Modelling, Volume 30, pp. 117-125.

[27] Goel S., Hofman J. M., Lahaie S., Pennoc D. M., Watts J. D., 2010, Predicting consumer behavior with Web search, PNAS – Proceedings of the National Accademy of Sciences of the United State of America, Volume 107, n. 41.
[28] Guzman G., 2011, Internet search behavior as an economic forecasting tool: the case of inflation expectations. J.Econ. Soc. Meas., Volume 36, Issue 3, pp. 119-167.
[29] ISTAT – Istituto Nazionale di Statistica, 2014, Nota mensile sull'andamento dell'economia italiana, 11/14, avaliable online from http://www.istat.it/it/archivio/144185.
[30] Johansen S., 1991, Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, Econometrica, Volume 59, Issue 6, pp. 1551–1580.
[31] Johansen S., 1995, Likelihood-Based Inference in Cointegrated Vector Autoregressive Models, Advanced Text in Econometrics, Clarendon Press Oxford.
[32] Hamilton J. D., 1994, Time Series Analysis, Princeton University Press.
[33] Lutkephol H., 2007, New Introduction to Multiple Time Series Analysis, Springer, Berlin.
[34] Mohebbi M., Vanderkam D., Kodysh J., 2011, Google correlate whitepaper, avaliable online from http://www.google.com/trends/correlate/whitepaper.pdf
[35] McLaren N., Shanbhogue R., 2011, Using internet search data as economic indicators, Bank of England Quarterly Bulletin, avaliable from http://www.bankofengland.co.uk/publications/Documents/quarterlybulletin/qb110206.pdf
[36] Naccarato A., Pallara A., Solari F., 2006, A dynamic linear model for preliminary estimation with an application to the Italian industrial turnover, in "Metodi statistici per l'integrazione di dati da fonti diverse", Eds. Liseo B., Montanari G. E., Torelli N., Ed. F. Angeli, pp. 369-379.
[37] Searle, S.R., 1982, Matrix Algebra Useful for Statistics, John Wileym New York.
[38] Shuoy T., 2009, Query indices and a 2008 downturn: Israeli data. Technical Report, Bank of Israel, avaliable online from http://www.bankisrael.gov.li/deptdata/mehkar/paper/dp0906e.pdf.
[39] Vanderkam D., Schonberger R., Rowley H., Kumar S., 2011, Nearest Neighbor Search in Google Correlate, avaliable online from http://www.google.com/trends/correlate/nnsearch.pdf
[40] Vicente M. R., Lopez-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 & Social Change, Volume 92, Issue 3, pp. 132-139.
[41] Vosen S., Schmidt T., 2011, Forecasting private consumption: survey-based indicators vs. Google Trends, J. Forecast., Volume 30, Issue 6, pp. 565-578.
[42] Vosen S., Schmidt T., 2012a, A monthly consumption indicator for Germany based on Internet search query data, Appl. Econ. Lett., Volume 19, Issue 7, pp. 683-687.
[43] Vosen S., Schmidt T., 2012b, Using Internet Data to Account for Special Events in Economic Forecasting, Social Science Research Network, available online from http://papers.ssrn.com/sol3/papers.cfm?abstract_jd=2200402.
[44] Wu L., Brynjolfsson E., 2010, The future of prediction: how Google searches foreshadow housing prices and sales, Technical Report, MIT, available online from http://www.nber.org/confer/2009/PRf09/Wu_Brynjolfsson.pdf
[45] Zeileis A., Leisch F., Hornik K., Kleiber C., 2002, An R Package for Testing for Structural Change in Linear Regression Models, Journal of Statistical Software, Volume 7, Issue 2, pp. 1–38.

Figures

Figure 1. Trend of the monthly unemployment rate (red) and the Google Trend job offers series (blue) in Italy, January 2004 – September 2014.

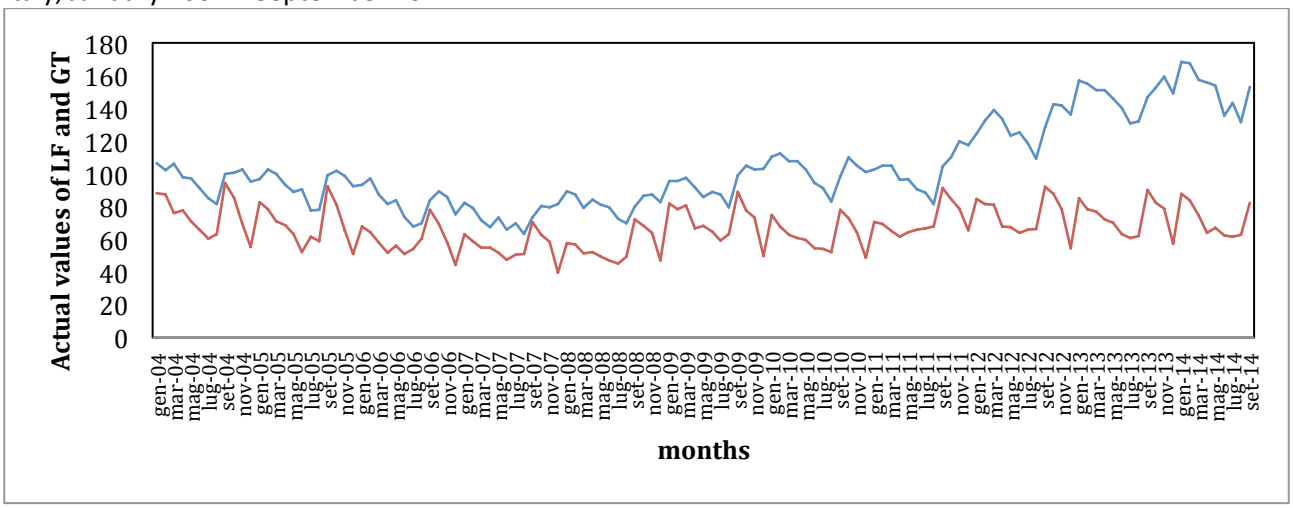

Figure 2. Destaginalized observed values (blue) and estimated values (red) with the VEC model for the unemployment rate series in Italy, January 2004 – September 2014

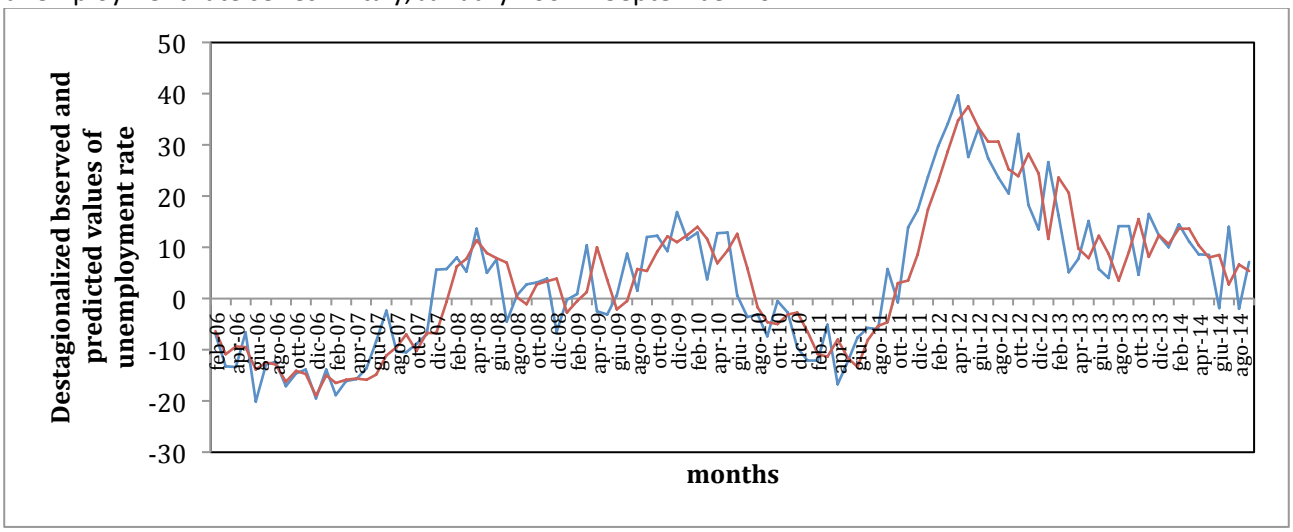

Figure 3. Rolling regression output - destagionalized observed values (blue) of the unemployment rate and predicted values with the VEC model (red) and the ARIMA model (green), January 2006 – September 2014

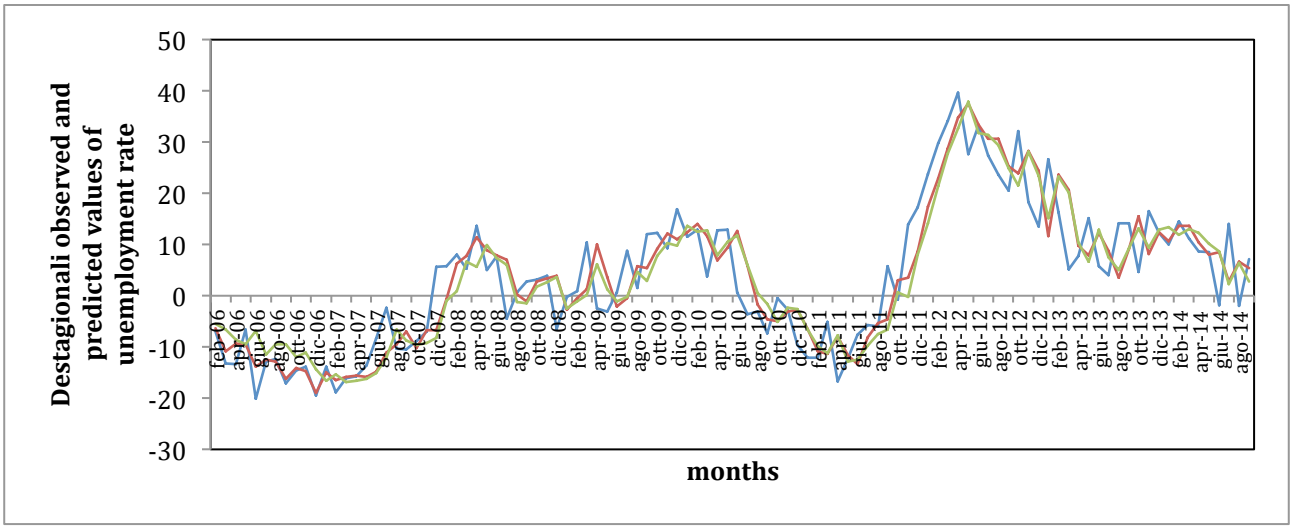

Figure 4. Absolute value of differences in between the observed values and the predicted values with VEC model (blue) and ARIMA model (red)

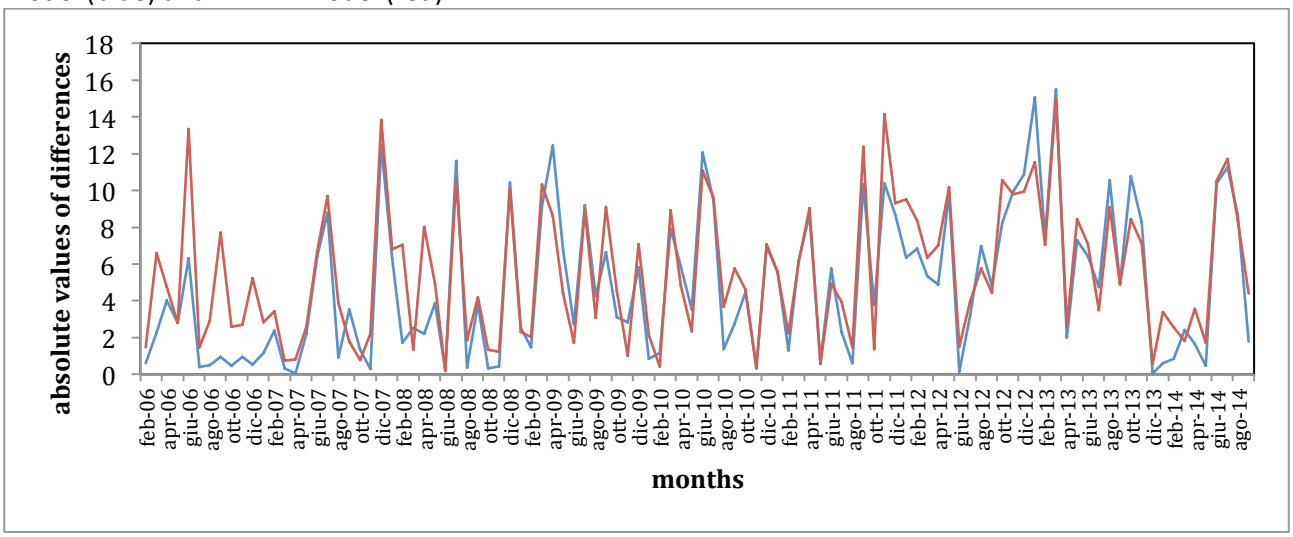

Figure 5. Empirical fluctuations for the VEC model, estimates the identification of a structural break,

January 2006 – September 2014