Using Internet Search Data for Nowcasting Unemployment Rate in Ukraine

Introduction

Timely economic statistics is crucial for effective decision making, but it is released with a lag by national statistical services even in developed countries, to say nothing of emerging economies, which generally experience bigger delays in term of data availability. Consequently, we need alternative ways of discovering current state of an economy. One of the most promising approaches is to use Google search queries data that is freely available through Google Trends service. This service provides weekly reports on the volume of search queries made by people in a given country. Taking into consideration the fact that people are regularly looking for the information, which is necessary for economic decision-making, we could expect search queries data to be relevant for “predicting” at least some of the economic indicators.

Objectives

Accordingly, the research question of this study is: Can the current state of Ukrainian economy be predicted based on available historical data and web search queries? To answer this question, we will focus on nowcasting unemployment rate, which is one of the major macroeconomic indicators, as changes in employment is a good predictor of current stage of business cycle.

Practical contribution of research

Independence of central banks is one of the principal foundations of market economy. Central banks can change, for example, reserve requirements or interest rates without obtaining permission of any kind from government or parliament. It allows them to react quickly to sudden changes in economic situation. However, in order to take prompt and sensible decisions, it is essential to have timely economic statistics. The scarcity of reliable data in emerging and developing economies is a well-known problem. At the same time the growing number of internet users in these countries creates conditions for using on-line data as a valuable source of information.
about current consumer sentiments. Thus our project will contribute to developing new instruments for economic forecasting in emerging economies.

**Hypotheses**

Based on the research objectives, the following hypotheses are to be tested.

**Hypothesis 1.** The Internet penetration level in Ukraine is sufficient to generate relevant information about the current state of Ukrainian economy.

It is expected to find the evidence which confirms that Ukraine reached the necessary number of internet users.

**Hypothesis 2.** The job-search related queries can be used for predicting the current rate of unemployment in Ukraine.

It is expected to prove the hypothesis and find keywords (search queries) that help to detect turning points in employment trends and therefore can be used to nowcast short-term changes in unemployment rates in Ukraine.

**Literature review**

*Using Google Trends for forecasting and nowcasting.* Recent developments in information technologies have created instruments, which we could not dream about. The Internet produced the environment, where the information is abundant and every our action is carefully recorded and saved forever on hard drives of search engines and web-sites. This information can be particularly useful to gain an insight into consumer sentiments or to see the reaction of households to an economic event in real time.

The timely information is absolutely crucial for policymakers and now, as the existing researches suggest, they can have it at the tip of their fingers at any time. A new term “nowcast” has been coined, which is defined as a “forecast” of the current state of an economy or prediction of the present (Choi & Varian, 2009a). Nowcasting is based on using available statistical data, web search queries, and other exogenous variables.

One of the first studies that investigated using web search queries to predict macroeconomic
statistics was Ettredge et al. (2005). The article shows a positive and significant association between job-search related queries and the official unemployment data. Launching two services by Google that provided access to search data “Insights for Search” and “Google trends” triggered more research in this field, starting with Ginsberg et al (2008). It showed that Google based search data can be used for forecasting influenza epidemics.

In economics, the seminal work of Choi and Varian (2009a) gives examples of nowcasting for car and home sales in the US and for travel to Hong-Kong. The other their work (Choi & Varian, 2009b) suggests that initial claims for unemployment benefits in the US can be predicted based on Google Trends data. Kholodilin, Podstawski, and Silverstovs (2010) nowcasted year on year growth rate of monthly private consumption in the U.S. They prove that models with Google indicators provide significantly better forecast accuracy. Hand and Judge (2010) used Google Trends search information to forecast the cinema demand in UK. They argue Google trends data increase the accuracy of forecasting models. Vosen and Schmidt (2009) developed a monthly consumption indicator for Germany based on Google Trends data. The paper shows that the Google based indicator outperforms the survey-based indicators.

There are other studies, but we will focus only on two aspects, firstly, using Google Trends for predicting the unemployment rates and, secondly, existing evidence for its usage in emerging economies.

*Predicting the unemployment rate.* There is growing (substantial) number of articles on using Google Trends for predicting the unemployment rate (Table 1) in different countries.

Table 1 – Using Google Trends for predicting the unemployment rate

<table>
<thead>
<tr>
<th>The authors</th>
<th>Country</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choi and Varian (2009b)</td>
<td>The USA</td>
<td>Find positive correlation between initial claims for unemployment benefits and the search related to two Google Trends categories: “Jobs” and “Welfare &amp; Unemployment”. Using Google Trends data also significantly improves forecasting accuracy</td>
</tr>
<tr>
<td>Askitas and Zimmermann (2009)</td>
<td>Germany</td>
<td>The paper establishes strong correlations between search activities for certain keywords and unemployment rates in Germany</td>
</tr>
</tbody>
</table>
The models including Google Trends data perform better than the ones without it. The indicator constructed by D’Amuri reliably predicts unemployment rates and it proved to be more accurate than other standard indicators such as: employment expectations surveys and the industrial production index.

Google search contains useful information that can be used for predicting short-term changes in employment. Best models, which are based on Google Trends data, improve the one month ahead prediction accuracy compared to the baseline models up to 18% on average.

Google Index is the best indicator to predict the U.S. unemployment rate. The model included the Google search data outperforms the traditional ones as well as forecasts made by professional bodies, e.g. the Survey of Professional Forecasts.

The paper confirms nowcasting ability of Google Trends data and its significance in detecting turning points in macro-economic trends. In particular, the spread in search queries explains about 15% of total unemployed spread.

Internet search data contain relevant information for explaining changes in unemployment and can be used for predicting changes in unemployment.

Google search data enhance predictions and nowcasts for the 15-24 years unemployed people, both in level and accuracy.

The findings are promising but the question arises whether they can be applied to emerging economies. All countries analysed by the authors are the developed countries with a high level of the Internet penetration (Table 2), but emerging economies have not reached such levels yet. For instance, the Internet penetration level in Ukraine was about 34% in 2012.

Table 2 – Number of Internet users in 2012

<table>
<thead>
<tr>
<th>Country</th>
<th>Number of Internet users</th>
<th>Penetration, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>The USA</td>
<td>254,295,536</td>
<td>81.0</td>
</tr>
<tr>
<td>France</td>
<td>54,473,474</td>
<td>83.0</td>
</tr>
<tr>
<td>Germany</td>
<td>68,296,919</td>
<td>84.0</td>
</tr>
<tr>
<td>Norway</td>
<td>4,471,907</td>
<td>95.0</td>
</tr>
<tr>
<td>Italy</td>
<td>35,531,527</td>
<td>58.0</td>
</tr>
<tr>
<td>Belgium</td>
<td>8,559,449</td>
<td>82.0</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>54,861,245</td>
<td>87.0</td>
</tr>
<tr>
<td>World globally</td>
<td>-</td>
<td>36.0</td>
</tr>
<tr>
<td>Average level in developing countries</td>
<td>-</td>
<td>28.0</td>
</tr>
<tr>
<td>Ukraine</td>
<td>15,115,820</td>
<td>33.7</td>
</tr>
</tbody>
</table>

Source: “Percentage of Individuals using the Internet 2000-2012”, International Telecommunications Union (Geneva), June 2013, retrieved 22 June 2013
Nowcasting in Emerging Economies. To the best of our knowledge, there are only three papers, which look at using Google Trends index for improving accuracy of nowcasting in emerging economies.

First of all, it is an article “Query Indices and a 2008 Downturn: Israeli Data”, written by Suhoy (2009). The article investigated whether Google query indexes could have predicted the 2008 downturn in Israel in real-time and it appeared that, indeed, Google query indices may be helpful in drawing inferences about the state of current economic growth.

Secondly, Carriere-Swallow and Labbe (2011) in “Nowcasting with Google Trends in an Emerging Market” proved that Internet search can be an indication of automobile purchases and therefore consumption patterns in Chile.

Lastly, Chadwick, Sengul (2012) in “Nowcasting Unemployment Rate in Turkey: Let's Ask Google” successfully nowcasted monthly nonagricultural unemployment rate for Turkey and proved that models with Google search query data perform statistically better than the ones without them.

According to International Telecommunications Union data, The Internet penetration level in these countries was (in corresponding years): Israel (2008) – 59.4%; Chile (2010) – 45.0%; Turkey (2011) – 43.1%. This is significantly higher than in Ukraine (33.7% in 2012).

The project's contribution to the academic debate. Currently, there is no study, which defines necessary prerequisites for adopting approach discussed above to emerging economies. It is obvious that there are some practical limitations (e.g. minimum level of the Internet penetration or most internet users in a country must use Google as a default search engine), but they are not clearly set in the existing literature. Moreover, it is still unclear whether the results received for developed countries can be applied to emerging economies. Yes, there are some evidence for Israel, Chile, and Turkey, but we cannot find any paper on using on-line data for nowcasting in Ukraine or any other similar country with low rate of the Internet penetration. Thus our paper is going to address the issue.
Methodology

Data

Datasets for the research are proposed to use from the official unemployment statistics and from Google Trends service.

We will use raw data provided by the State Employment Service of Ukraine (http://www.dcz.gov.ua/control/en/index) and by the State Statistics Service of Ukraine (http://www.ukrstat.gov.ua). It is necessary to take data at least for 5 years to make the forecast adequate.

The next variable contains Google Trends Index. This Index shows the number of aggregated data on search volume for keywords “that receive a significant amount of traffic”.

Google Trends data does not report the raw level of queries for a given search term. Rather, it reports a query index. The query index starts with the query share: the total query volume for search term in a given geographic region divided by the total number of queries in that region at a point in time. The query share numbers are then normalized so that they start at 0 in 01.01.2004. Numbers at later dates indicated the percentage deviation from the query share on January 01.01.2004.

The choice of keywords is the key feature for the research. For our purposes proposed to try several keywords relative to Ukrainian labor market such as “state job center” (Russian «центр занятости», Ukrainian «центр зайнятості»), “vacancy” (Russian «вакансия», Ukrainian «вакансія»), “job search” (Russian «поиск работы», Ukrainian «пошук роботи») etc. Data are taken from the request that better represents the key fluctuations in the level of unemployment.

Theoretical model

The most appropriate model for nowcasting is autoregressive model (AR). This model is used in many financial sectors, where it is necessary to predict a variety of data. AR belongs to the class of regression methods. The model is based on the correlation of the reference random process at the current time with a finite or infinite number of samples in previous times. Thus, the AR
model describes the dependence between current time series level and its previous levels.

Autoregressive process of the p-order defined as follow:

\[ X_t = c + \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t, \]

where \( c \) – constant (usually it equals 0 to simplify the model);

\( \alpha_1 \ldots \alpha_p \) – model parameters (autoregression coefficients)

\( \varepsilon_t \) – “white noise”

This model has some limitations.

The first limitation refers to the selection of the factors that will be included in AR model. Every factor (variable) must lead to a substantial increase in the quality of the model. This value characterized by the coefficient of determination – R2. Thus the selection of factors in the model aims at minimizing the number of variables, because their unwarranted increase leads to difficulties in the interpretation of the model and reduce the reliability of the results.

The second – the key limitation – is the lack of multicollinearity of the independent variables. Multicollinearity – is the presence of a linear dependence between the independent variables (factors) of the regression model. Multicollinearity affects negatively on the quantitative characteristics of the constructed model or make it impossible to build.

**Estimation**

First of all, we must determine the order of the model, which is suitable for our data. For this purpose we will use correlograms analysis. We will construct the correlograms for zero and first differences of the time series, which contains official unemployment statistics. To determine the order, it is necessary to analyze the correlation values. In the research, the needed lag with the biggest correlation rate is taken.

After that we need to determine the autoregression coefficients. We can determine them as the coefficients of the linear regression:
\[ Y_t = c + ax_1 + ... + ax_n, \]

where the dependent variable – initial time series with unemployment rate, independent variables – Google Trends Index and unemployment rate with the lag.

Lastly, the AR model for our data can be formed. To estimate the quality rate of the model, we propose to use the coefficient of determination:

\[
R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},
\]

where \( \hat{y}_i \) – unemployment rate calculated by the model;

\( y_i \) – initial unemployment rate;

\( \bar{y} \) – the mean of the initial unemployment rate.

Expected results

We expect to find the keywords that help to nowcast short-term changes in unemployment rates in Ukraine using data from official statistics for the last available periods and Google Trends Index. There are a lot of unexpected fluctuations that current models can’t predict. That’s why using the AR model with such variables can help to predict turning points in the unemployment rate.

Bibliography

Available at SSRN: http://ssrn.com/abstract=1465341

Bughin, J. R. (2011). “Nowcasting” the Belgian Economy. Available at SSRN:
http://ssrn.com/abstract=1903791


Participants

Vladyslav Dombrovskyi – project leader, responsible for theoretical part

Olga Sinyavskaya – project participant, responsible for econometric part

Alternative sources of funding

There are no other alternative sources of project financing

Project timetable

July-September 2014 – Literature review;

August-October 2014 – Data collection and preliminary data analysis;

October-November 2014 – Interpretation of the preliminary results, first research draft;

December 2014 – Discussions of the obtained results; getting the feedback and remarks;

January-February 2015 – Final report preparation and policy recommendations development;

March 2015 – Submission of a journal paper to a referred journal.