The use of web activity evidence to increase the timeliness of official statistics indicators

Fernando Reis¹

Pedro Ferreira²

Vittorio Perduca³

Abstract

The official statistics community is reacting to the challenges and opportunities offered by big data. At European level the heads of the national statistical institutes and of Eurostat agreed on a memorandum of understanding to tackle together big data. One of the big data sources available to official statistics is the electronic traces left by users while they use web services. Many of these services provide data based on these traces, either at real time or with very small time lags. As many human activities measured by official statistics are closely related to people’s behaviour online, this data on people’s web activity offers the potential to produce

¹ European Commission (Eurostat), Joseph Bech building, 5 rue Alphonse Weicker, L-2721 – Luxembourg, Luxembourg. E-mail: fernando.reis@ec.europa.eu

² European Commission (Eurostat), Joseph Bech building, 5 rue Alphonse Weicker, L-2721 – Luxembourg, Luxembourg. E-mail: pedro-jorge.martins-ferreira@ec.europa.eu

³ Laboratoire MAP5, Université Paris Descartes and CNRS, Sorbonne Paris Cité, 45 rue des Saints Pères 75006 – Paris, France, Email: vittorio.perduca@parisdescartes.fr (V.P. worked at this project while with the European Commission - Eurostat)
predictions for socio-economic indicators with the purpose to increase the timeliness of the
statistics. Many experiments have being carried out recently and there is evidence that these
predictions can be made. However, having good prediction models is not enough to produce
official statistics figures. If we want to assess the possibility of using this type of big data
source then we need to reflect on the transparency, continuity, quality and potential to be in-
tegrated with official statistics traditional methods, as well study in more detail the relation
between web activity and the phenomena we want to predict.

Keywords: Big data, modernisation, web, prediction, flash estimates

1. Introduction

Big data brought to the attention of the official statistics community the existence of inumer-
ous new sources of data which can be potentially used in the production of statistics. One of
those sources is the traces left by users of web services, which activity is related to other
facets of their social live and which are measured by official statistics. For example, when
confronted with the loss of their job, users search for information on new jobs online, consult
employment related websites and post about it in their Facebook page or Twitter.

This data about users’ web activity is potentially available very quickly because web services
are pure electronic services completely supported by IT systems and with very high levels of
automation. These data is automatically stored in databases supporting the web services or
in the log files of the web servers. Some of this data is public (e.g. Twitter) or is made availa-
ble (in aggregated form) by the web services themselves (e.g. Google).

There is already some experience in using this web activity data to predict official statistics
socio-economic indicators, like flu incidence, unemployment and tourism and migration flows.
Some statistical offices have also conducted some experiments.
In this paper we show that it is relatively easy to integrate some data on web search activity of users to increase the precision of simple predictive models, in the case of unemployment. However, if official statistics are going to make use of web activity data to produce flash estimates of socio-economic indicators then it should not do it by simply reproducing what others can do but instead do it making use of its specific comparative advantages. In order to integrate this type of source in the production of flash estimates of official socio-economic indicators, statistical offices need to address several challenges. The experiences so far provide important lessons to help address those challenges.

The paper is organized as follows: Section 2 summarizes the opportunity and challenges posed by big data to official statistics and describes the actions undertaken by the European Statistical System; Section 3 describes previous work by researchers and official statisticians about the prediction of socio-economic indicators based on web-activity; as an example of such applications, Section 4 describes a toy model for improving the timeliness of unemployment statistics based on both official data and Google Trend data; Section 5 illustrates the experience developed by Eurostat in the production of flash estimates based on secondary data and how this could be beneficial in developing new statistical products based on big data; Section 6 outlines a program for introducing web-activity data in the production of flash estimates.
2. The European Statistical System response to the big data challenge

2.1. Big Data, New Data

After centuries as the, first sole and then main, collector of data on the society and the economy, the monopoly of the statistical public authorities is over. Data is now all around us. What once was a scarce, and expensive to collect, resource is now abundantly available.

Big data means first and utmost new data sources for official statistics composed of new types of data and with characteristics different than those of traditional data sources. Adding to the traditional quantitative measurements and qualitative characteristics of individuals and enterprises, big data brings the recognition that value may be found in any type of data. This includes network data (e.g. social networks and mobile phone communications), text (e.g. Twitter), pictures, sound and video. Web activity evidence includes the traces left by the users of web services which are registered in the log files of the web servers (sometimes compiled in aggregated form and made available by the providers) and the information (normally textual) entered by users which is available in the websites.

These new data sources present particular challenges to official statistics. Firstly, sometimes the organisations that hold the data are outside the jurisdiction of the statistical authorities (e.g. when they are foreign corporations, such as Google and Facebook). Secondly, the order of magnitude of the data which can be collected by the national statistical institutes from

\[^4\] Parts of this chapter were taken from the text of one of the authors in a paper (Reis, De- munter, “Eurostat Work in the Area of Big Data and the Scheveningen Memorandum”) submitted to the International Journal of Geographical Information Science special issue titled "Geographies of Mobility: Applications of Location Based Data".
the data holders is much higher than in traditional data collections. This has two consequences. On one hand, it is not reasonable anymore to leave the burden of the compilation and transmission of the data solely to the data provider. On the other hand, the noise to information ratio increases significantly. Thirdly, in some cases the data which is of interest of the statistical authority has a commercial value for the data provider, being in some cases the core of its business model (e.g. Google and Facebook).

2.2. The opportunities offered by big data

These new data sources offer several opportunities for official statistics. Many big data sources consist of very large datasets which can be used by NSI to provide statistics with a much higher detail than it’s possible with traditional statistical production methods. This higher detail refers not only to regional detail, but also to the production of statistics for very small groups of the population for which it has not been possible to serve by official statistics.

Another opportunity is the possibility of using data which is already available, potentially at a very low cost compared to the cost of traditional methods. This is not to say that big data sources are costless. As mentioned before, some datasets may be so large that it’s not reasonable anymore to leave the burden of statistical data provision to a few data providers.

The most relevant opportunity for us in this paper is the possibility of having access to data shortly after the event to which it refers has occurred. This happens because normally big data sources originate in automated systems and therefore the time lag of data collection is non-existent. In the case of web services the activity of the users in the website is automatically registered either in databases or in log files of the web server. In the case where web services provide data originating in these users activities, they can do it very fast (see further on the example of Google).
2.3. The Impact on Official Statistics Production

The swing from mostly designing primary production to the reuse of secondary sources will inevitably require transformations in National Statistical Institutes (NSI). Firstly, it requires a change in how statistical production is organised and in the skills set of official statisticians. From sole designers of single purpose atomic statistical production systems for particular statistical products statisticians will need to become designers of statistical products targeted to answer specific needs of the society or of policy makers based on a multitude of data sources. This is a change which is already happening for several other reasons. The use of administrative sources has increased significantly in the last decades, so the use of secondary sources is not new to NSI. The need for modernisation of the statistical production systems (in order to increase efficiency and flexibility) has also already initiated a movement towards an integration of statistical production across different statistical domains.

Secondly, big data may eventually bring new tasks and responsibilities to NSI. In particular, official statistics may assume the role of guarantee of quality of statistics produced from big data sources, either by itself or by other entities, through a mechanism of accreditation and certification ([2]).

2.4. The Analytical Challenge

However, big data is more than just new data. It represents a change in attitude towards data. While some private companies build complete business models based on the commercial exploration of data (e.g. Google, Facebook), others look for ways to monetise the data that in some cases have been in the company for some time. This dynamism in search for innovative ways to explore data together with several decades of development of data analysis tools and methods and the spectacular increase in data availability (or new possibilities to capture data) resulted in the appearance of new data products based on more or less sophisticated analytics, most notably predictive analytics.
In this world of big data so impregnated of analytics, official statistics cannot avoid looking as suffering from an analytical deficit. Therefore, big data also presents official statistics with the challenge of presenting statistical users with these new statistical products which they start to get used to enjoy from elsewhere.

The type of statistical product under discussion in this paper is one example of such new analytical products. Based on the much higher timeliness of some new sources based on web activity of individuals, there’s the potential to use predictive models to provide the users with flash estimates of traditional socio-economic indicators at near real time.

2.5. The Scheveningen Memorandum and its Follow-up

Recognising the change in the conditions and environment in which official statistics operate, the international community of official statisticians is reacting.

The UNECE High-Level Group for the Modernisation of Statistical Production and Services has identified, in its strategic vision (UNECE, 2010), the creation of new statistical products based on an active exploration of these new data sources as a key element of the modernisation of official statistics (High-Level Group for the Modernisation of Statistical Production and Services, 2011).

Recognising the strategic importance of big data for the European Statistical System, the Directors General of the European National Statistical Institutes agreed on a memorandum addressing big data which was formally adopted by the ESSC in Scheveningen in September 2013.

The Scheveningen Memorandum recognises that the increasing level of digitisation of society, and the consequent digital footprint people leave, offers an opportunity for the compilation of statistics and that it should be incorporated in the conceptual design of official statistics. In particular, it provides alternatives to deal with the current challenges faced by official statis-
tics, such as the decreasing response rates and the need to increase the overall efficiency of statistical production systems.

However, the Scheveningen Memorandum also recognises that the use of big data poses challenges to the ESS. Therefore, it calls for an examination of the potential of big data sources and the development of an official statistics big data strategy and roadmap. In order to draw such strategy and roadmap, Eurostat has set up a task-force composed of people from Eurostat, National Statistical Institutes, other international organisations and academia.

Although it is easy enough to recognise that big data will have potentially a large impact, at this stage it is not easy to see in detail what big data means for official statistics. New data sources may be available to produce statistics, but there are a large number of possible ones and each one of them seem to have their own particularities.

The strategy envisaged by the task-force set up by Eurostat is characterised by three elements. Firstly, start by piloting specific applications of big data sources to produce statistics traditionally in the scope of NSI. These pilots should demonstrate the potential of big data and provide experience which can be used to clarify what big data means for official statistics. Secondly, the adoption of a roadmap in three time horizons to organise the action plan: short-term, medium-term and long-term. The pilots would then be part of the short-term strategy. Thirdly, review the roadmap based on the lessons learned from the pilots and the methodological and technical developments in big data.
3. Experiences so far, including in official statistics

The use of web activity data to predict a socio-economic indicator was suggested as early as 2005 by [17] for the unemployment rate. Based on the idea that a significant proportion of job related information gathering is conducted via the Internet, the authors studied the relation in the U.S. between WordTracker’s Top 500 Keywords Report data (accessible in http://www.top-keywords.com/longterm.html as of September 2014) and the monthly unemployment levels published by the Bureau of Labor Statistics. The study concluded that there was a positive significant association between the search engine keyword usage data and the official unemployment figures. The study, however, did not attempt to predict unemployment using web search data, but simply established the correlation between the two data sources.

3.1. Google Trends

In 2006, Google launched “Google Trends” (see [16] for an example of the announcement in the online media), a service providing data on how often specific search terms are entered into the search engine of the company over a certain period of time. The tool was initially used to identify trending terms, i.e. terms for which a consistently increasing number of searches was observed ([18]). However, the high timeliness of Google Trends soon fuelled a significant number of studies dedicated to the use of the source to predict socio-economic indicators with the purpose of obtaining results quicker than the publishing of official statistics authorities.

Google itself published in 2009 in its research blog one of the first attempts to predict socio-economic indicators based on Google Trends data ([12]). The paper used search data to produce short-term predictions for several indicators, car sales, retail sales, home sales and number of visitors. It concluded that for simple auto-regressive time-series models the introduction of search data as predictors increased their precision in short-term predictions. In
addition to the lagged predictors, contemporaneous search data was used to predict the indicators. As Google Trends search data is released with a very high timeliness, a few days after the reference period, such models would allow making predictions practically for the present time.

Other studies have also used Google Trends data to produce predictions of those same indicators and other several ones. Between other indicators, we can find influenza epidemics ([14]), unemployment ([10], [9], [25]), and private consumption ([15], [13], [22]).

3.2. Lessons from Google Flu Trends

Based on studies like this and on others focusing on the use of web activity for influenza surveillance, Google launched, in 2008, Google Flu Trends ([14]). It used aggregated Google search data to estimate current flu activity for the United States with a higher timeliness than the official indicator from the Centers for Disease Control and Prevention (CDC).

The experience of Google Flu Trends provides several lessons about the use of search data for producing flash estimates in a domain of official statistics. For most of the time between 2009 and 2013 GFT performed well ([6]). However, in 2009 it failed to accurately estimate the official figures from the CDC by underestimating the incidence of flu, an event attributed to changes in people’s search behaviour and which led to a revision of GFT algorithms ([21]). In 2013, Nature reported that for the 2012/2013 flu season peak the GFT estimate was almost the double of the CDC later released figure ([1]). The possible cause pointed out was the widespread media coverage of that year’s severe flu season.

This generated to some extent a backlash against big data (see [24] for an example). The exaggeration by some of the potential of pure data-driven applications based on very large datasets created the eagerness of others to jump on the first sign of problems into a discourse about the limitations of big data. However, as shown in [4] there are possible im-
provements to the GFT prediction model which would have prevented the errors occurred. This is part of the process of building a reliable statistical product and GFT was probably still not ready “for production”. The lesson is that releasing such a product before it is properly matured can destroy prematurely its reputation. Another lesson is “big data hubris” ([4]), the belief that big data by itself will replace all traditional data collection. The key to extract value from big data for official statistics is its integration in multi-source statistical production systems.

Part of the difficulties of GFT in predicting flu prevalence was because of the frequent changes to the search algorithms implemented by Google engineers which had an impact on the results returned to the users and on how they perform multiple searches. This instability of the source data of the prediction model changed the validity of the model and would require its dynamic calibration ([4]).

Another lesson to be taken from the GFT experience is the need for transparency and replicability. Google did not release all the details of GFT. For example, the search terms used are not known. Transparency is one of the fundamental principles of official statistics (vide principle 3 in [23]). It is required for the correct interpretation of the official statistics by the users, including researchers who may want to assess those statistics to build their research on them. Replicability is also important at this stage where NSI can learn from each other’s experiences.

GFT and the other examples of application mentioned in the previous section are based on Google Trends (GT), which is an index computed from the individual search queries performed by users. Google does not provide access to the data on the individual search queries. The indexes are created based on a sample of individual search queries which changes every day ([19]). As a consequence, GT presents slightly different results depending on the day its data is collected and presents an additional source of uncertainty, sampling error.
(others are the percentage of people who use web searching, the percentage of those who use Google services and the relationship between search behaviour and the phenomenon analysed). Another undesirable characteristic of GT is that the sampling methodology is not revealed by Google, which effectively creates a black box.

3.3. Other sources of web activity data

Web search data, Google Trends in particular, is not the only source of online activity which has been used to predict socio-economic indicators. Twitter and Wikipedia page views have also been used in the prediction of socio-economic indicators.

The number of Wikipedia page views has been used in [5] to predict influenza-like illness in the United States. In comparison to Google Flu Trends (GFT), the prediction model developed was better in some situations but in others it was not. The prediction model based on Wikipedia page views was able to identify the peak week of the flu season more accurately than GFT. However, in 4 out of 6 flu seasons the prediction of GFT was closer to the official figures than the model based on Wikipedia.

One example of the use of Twitter for the prediction of official statistics is [7]. In this study, international and internal migration patterns were estimated from geolocated data for about 500,000 Twitter users. It concluded that the methods developed there can be used to predict turning points in migration trends and to improve the understanding of the relationship between internal and international migration.

3.4. Experiences in official statistics

NSI have also started exploring the use of web activity evidence for the prediction of socio-economic indicators.
The CBS has studied the relation between monthly consumer confidence and the sentiment in public Facebook messages and in Twitter messages [20]. It concluded that because of the timeliness of the social media data and the quickness it can be processed, a prediction of the official consumer confidence could be published before the official figures and at a higher frequency.

ISTAT used Google Trends data to predict one month ahead the number of people seeking for a job as estimated by the labour force survey ([8]).

4. A very simple example of an application with Google Trends

In this chapter we try to show how simple it can be to integrate Google Trends (GT) into a prediction model and still get significant improvements in prediction accuracy.

We present a simple example of applying GT time series to improve the prediction of unemployment statistics in two countries, France and Italy. Here prediction refers to the prediction of present (‘nowcasting’), as meant in [11]. Indeed, the models discussed in this work are based on the seminal papers [12], [10] and [11], where GT data are used to improve simple predictive models.

4.1. Models

We consider two models:

1) The baseline is a simple autoregressive model in which unemployment at month $t$ is predicted using unemployment at month $t-1$:

$$y_t = a + b \log y_{t-1} + e_t$$
where $y_t$ is unemployment at month $t$, $a$ and $b$ are coefficients to estimate, and $e_t$ is an error term.

2) The alternative model is the baseline model adjusted for query terms $q_i$:

$$y_t = a + b_0 y_{t-1} + \sum_i b_i q_{i,t}$$

where $a$ and $b_i$ are coefficients and $q_{i,t}$ is the search volume of the query term $q_i$ at time $t$.

As detailed below, we arbitrarily chose query terms which we believe a priori people search on Google when unemployed.

For France we considered the following three query terms:

- ‘pole emploi’ is the French governmental agency which registers unemployed people, helps them find jobs and provides them with financial aid;

- ‘indemnité’ refers to allocations;

- ‘etre au chomage’ is a query we believe unemployed people search in order to find useful resources for improving their condition.

For Italy, we considered four query terms:

- ‘impiego’, that is ‘job’;

- ‘offerte lavoro’, that is ‘job vacancies’;

- ‘curriculum’ is a term people looking for jobs might search in order to find useful hints and improve the chances that their resume retains employers’ attention;

- ‘infojobs’ refers to a popular website consulted in Italy for job hunting.
4.2. Data

All time-series were downloaded on 16 July 2014. Official data consisted in non-seasonal adjusted monthly unemployment data downloaded from Eurostat public database.

For France, GT data for the three terms were downloaded from:

- www.google.fr/trends/explore#q=pole%20emploi&geo=FR&cmpt=q
- www.google.fr/trends/explore#q=%27indemnit%C3%A9%20chomage%27&geo=FR&cmpt=q
- www.google.fr/trends/explore#q=%27etre%20au%20chomage%27&geo=FR&cmpt=q

Weekly data for terms ‘pole emploi’ and ‘indemnité’ were aggregated on a monthly basis. Only months for which data were available in the whole four datasets were kept for further analysis. These include 63 months from March 2009 to May 2014.

For Italy, indexed data for the four terms were downloaded from:

- www.google.fr/trends/explore#cat=0-958-60&q=impiego&geo=IT&cmpt=q
- www.google.fr/trends/explore#cat=0-958-60&q=%27offerte%20lavoro%27&geo=IT&cmpt=q
- www.google.fr/trends/explore#cat=0-958-60&q=curriculum&geo=IT&cmpt=q
- www.google.fr/trends/explore#cat=0-958-60&q=infojobs&geo=IT&cmpt=q

Only months for which data were available in the whole four datasets were kept for further analysis. These include 77 months from January 2008 to May 2014.

4.3. Results for France

In the results that follow, all computations were performed in R.

For each month $t$ after August 2011 we fitted the two models on all previous months only (that is from August 2011 to $t-1$) and predicted the unemployment levels at month $t$. 
Figure 1 shows the results. The adjusted model fits the actual data slightly better than the simple AR model, as shown by the averages of the absolute values of the relative prediction errors (the so called Mean Average Error): \( \text{MAE}_{\text{AR model}} = 2.5\% \) and \( \text{MAE}_{\text{Adjusted AR model}} = 2.4\% \). The Pearson correlation coefficients are \( r_{\text{AR model}} = 0.88 \) and \( r_{\text{Adjusted AR model}} = 0.90 \).

Figure 2 shows the relative errors for the two models: clearly the adjusted model outperforms the simple AR model after a few months. This is probably due to the fact that the adjusted model has more coefficients to estimate and therefore more observations (i.e. months) are required. Errors have seasonal patterns (figure not shown), indicating that both models need drastic improvements.

4.4. Results for Italy

Also in the case of Italy, accounting for query terms improves the performance of the prediction baseline model.

The Mean Average Error \( \text{MAE}_{\text{AR model}} = 6.3\% \) \( (r_{\text{AR model}} = 0.93) \) and \( \text{MAE}_{\text{Adjusted AR model}} = 4.7\% \) \( (r_{\text{Adjusted AR model}} = 0.97) \), see Figure 3.

This is confirmed by the relative errors shown in Figure 4.

5. Experience with flash indicators in Eurostat

5.1. What is the euro area HICP flash estimate?

The euro area HICP (Harmonised Index of Consumer Prices) flash estimates broken down in main components is a statistical product that is produced every month and is one of the most notable indicators produced at Eurostat. At the end of each month (if the end of the month falls during a weekend, then it is published on the following working day) estimates of what
inflation was during that month are published. Since September 2012 Eurostat have been regularly publishing flash estimates not only for the all-items but also for main components. In September 2014 three extra main components were added to the existing set of flash estimates, making the set the following: ‘all-items’, ‘food’, ‘processed food’, ‘unprocessed food’, ‘non-energy industrial goods’, ‘energy’, ‘services’, ‘all-items except energy’, ‘all-items except energy and food’ and ‘all-items except energy and unprocessed food’.

Inflation flash estimates are important indicators for the general public, financial markets but most importantly for the European Central Bank. In fact, flash estimates were a request from ECB in order to have the most updated measurement of inflation by the time of the ECB Governing Council Meeting, responsible for formulating the euro area monetary policy.

When producing such an important indicator, extra care on the quality in the broader sense is needed. Accuracy is only a part of the quality but punctuality is also very relevant. In addition to being able to release the flash estimates in the dates specified in advance, it is also very important not to miss a publication. Once the regular production has started, it can’t stop.

5.2. How is it produced?

The euro area HICP flash estimate combines early information sent by some Member States with forecasted data for the remaining ones. In most of the cases, ‘early information’ is preliminary estimates based on collected prices that will be part of the final HICP data set but taken at a very early stage of the production process, e.g. data not completely validated, no quality adjustment made, etc. Since preliminary data is based on the same collected prices than the final HICP indices, it is not a surprise that it is very accurate. In fact, it has been proved that preliminary data is much more accurate than any model based forecast. Therefore, preliminary data is always preferred.
Preliminary data might be biased so the flash estimate procedure developed at Eurostat corrects whenever possible with a calibration procedure developed for this specific purpose.

Unfortunately, not all the countries can provide preliminary data on time: for these countries it is necessary to forecast the missing data.

Different main components of inflation have very distinct stochastic behaviours and some of them are very volatile and hard to predict. As such, each component is treated separately and any auxiliary data that can improve the forecast are taken into account. The auxiliary data used by the flash estimate is the energy prices of the Weekly Oil Bulletin, produced by the Directorate General for Energy of the European Commission (DG ENER), an administrative data source.

Due to short time period to produce the flash estimate, usually no more than 3 hours, an automatic forecasting tool was developed at Eurostat.

5.3. The flash estimates as a comprehensive example for the use of Big Data in official statistics

The euro area HICP flash estimate doesn't make use of Big Data. However, the need to use an administrative data source to overcome a coverage problem (euro area is not totally covered by the preliminary data) can serve as an example to show one possible use of Big Data in official statistics regular production.

The auxiliary data used in the flash estimate is very useful due to several factors:

- It is cheap: it is not Eurostat's data so resources were not spent on data collection, compilation, etc. Eurostat just take it as it is;

- It is regular: each week DG ENER publishes an update on the energy prices;
- It is easily available: the data is freely available on the web for anyone who wants to use it.

Another important aspect is that the Weekly Oil Bulletin was meant to another purpose than the euro area HICP flash estimates: its main purpose is to improve the transparency of oil prices and to strengthen the internal market. Nevertheless, the data is now used as well to improve the regular production of inflation indicators, an application which was probably not foreseen by the time DG ENER implemented this data collection.

However, the use of this administrative data source is only possible due to two very important facts:

(1) Data is regularly available, with no interruptions. It is a very important aspect since the flash estimate ‘once started cannot be stopped’. Eurostat cannot afford to tell to its users that such an important indicator is not available for that time period because an alternative data source was not available. Furthermore, even if the availability is granted by DG ENER (there is even a legal act that obliges Member States to report energy prices: Council Decision of 22 April 1999) by some reason is not available, there are contingency plans, e.g. Europe Brent Spot prices;

(2) There is a strong but more important stable correlation between some HICP main components and the administrative data source. This is also a very important aspect because Eurostat cannot afford to produce statistics with certain accuracy and realize that in a couple of months the accuracy has deteriorated so much that would jeopardize its release.

There is another aspect about this administrative data source that is very important to the successful use in the flash estimate. Weekly oil bulletin is about reference prices of energy products, which is very much related to the price that an average consumer pays. So, when
using this administrative data source there was almost no risk of confusing the signal with
noisy data that only apparently could be related with the HICP indices. This might not be the
case when we’re speaking about a big data source and / or the connection between the two
data sources is not so clear.

Making the parallel between the use of an administrative data in the production of official sta-
tistics and the potential use of big data, one can conclude the following:

- There might be out there lots of data which were produced for many other purposes
  than official statistics that turn out to be an important part of the production process of
  official statistics. We, official statisticians, can only be encouraged and motivated to
  search for it;

- At the same time we, as official statisticians, must keep our feet on the ground and be
  very selective when incorporating an unconventional data source in the official statistics
  production. Before incorporating that extra data source, it is important to be able to an-
  swer two important questions:
    - Will the Big Data source be available in the future so I will have some guarantees
      that I can release an official statistic without being forced to stop it after a couple
      of releases?
    - What I’m extracting from the huge data available is really a signal or is it just noise?
      And if it is a signal, is it measuring the phenomena that I want?

6. A programme for the introduction of web activity data in the pro-
duction of flash estimates

As it has been shown in this paper, it is not difficult to use web activity data (in this case
Google Trends) to improve the predictions of at least very simple time-series models. It was
also shown that there is plenty of literature showing cases where baseline models are improved by using this big data source, even if this literature is unbalanced.

However, the use of a source such as Google Trends for the regular production of flash estimates of official statistics poses particular challenges which we need to address. What are then the steps we need to take to integrate web activity data sources in the production of official flash estimates?

6.1. A balanced study on the use of web activity data sources for prediction

As pointed out by [3], normally research results are presented when the use of web activity data successfully improves nowcasting, but when it is not successful results are not disseminated. For this reason a reading of the literature on this subject provides an unbalanced overview of the potential in general of this type of data for the prediction of socio-economic indicators.

Finding positive results is by itself relevant information inviting more research on the subject. However, in order to have a more accurate idea of the potential of this type of data which can guide further investment in National Statistical Institutes, a balanced study is required. Some studies like the ones presented in [3] and [19] provide a more balanced overview by including more than one country and more than one indicator. The next step should be to launch a larger scale balanced study including several socio-economic indicators and several countries following the same approach, which would report on both positive and negative results so an overall assessment can be made.

6.2. Diversification and assessment of the data sources on web activity

Big data sources, in particular web activity data sources, present some challenges to the principles which guide official statistics (here we will follow the European Statistics Code of Practice - CoP). As secondary external sources, they are out of the control of the NSI. In the
case of the traditional sources, the NSI either have full control in the case of surveys or it has some degree of influence, depending on the country, as it is the case of administrative records. That lack of control presents several risks.

Firstly, there is the risk that the data source is a black box. NSI make an effort to document as completely as possible the production process of official statistics. This transparency is required to keep the level of trust of the society and of political stakeholders in official statistics. However, in the case of big data sources held by private entities it might not be possible to guarantee the same level of transparency. It is expectable that in some cases the disclosure of the data processing in the web service may put the data supplier in a competitive disadvantage in its market.

Secondly, unless the NSI audits thoroughly the data processing of the web service, it cannot guarantee that the source was not subject to manipulation, regardless of that manipulation having taken place or not. A meticulous audit might not be possible (if the data supplier is outside the jurisdiction of the statistical authority) or very expensive.

Thirdly, the data source may be subject to frequent breaks in series. The data processing procedures of web services are designed according to the business needs of the entity running the services and these may change over time. As pointed out in [4], this has been the case of Google, which since the launch of Google Trends in 2006 has made several revisions to its algorithms which impacted the data made available via Google Trends.

Fourthly, there’s a risk of lack of continuity of source as the NSI is unable to guarantee that the source will be available for as long as it is necessary. The usefulness of the data coming from particular web services, such as a search engine, depends directly on its popularity which changes over time. The availability of the source can also be broken by technological changes which are not under the control of the NSI.
Some of these risks can be mitigated by the combined use of several web activity data sources in the prediction models. It would reduce the influence of individual data sources, which the NSI do not control, in the predicted values and provide some guarantee that the official flash estimated was not tampered. The diversity of sources would also allow the building of contingency plans for the eventual lack of continuity of some of the sources. For example, in the case of an employment rate flash estimate, a possible source, besides the ones already mentioned in this paper, can be visits to employment related websites.

We also need to assess how and how frequently the prediction models should be reviewed in order to accommodate breaks in series.

Finally, the establishment of procedures for the accreditation and certification of big data sources for official statistics ([2]) should be set up to assure the transparency and quality of the sources.

6.3. Integration of web activity data with traditional official statistics sources

Several examples of prediction of socio-economic indicators were given in this paper and most of them were not done by official statistical offices (national statistical institutes and European and international statistical offices). A legitimate question is why should the official statistical offices do it themselves if others can do it.

In this paper we do not try to answer this question. What we argue is that if official statistics provide flash estimates of socio-economic indicators using predictive models based on web activity data, then it should not do it by simply reproducing what others can do but instead do it making use of its specific comparative advantages.

The most obvious comparative advantage of official statistical offices is that they are the ones who produce the official indicators and therefore are the ones in the best position to know its specificities and in some cases have provisional data (as is the case of the inflation
flash estimate) which might also be introduced in the model. Another comparative advantage is the experience in conducting surveys and, in the particular case of NSIs, the fact that they have massive data collection systems.

Therefore, official statistical offices should integrate the production of flash estimates in their regular statistical production systems. This means making use of possibly more detailed information about the indicators than what is published. Surveys might also be adjusted so they provide information which helps make use of web activity evidence or correct for the bias present in the big data sources.

6.4. Research on the relation between web activity and the phenomena being predicted

The robustness of the predictions of the predictive models based on web activity data can only really be assured if there is a good understanding of the relation between the phenomenon being predicted and the web activity of individuals. Therefore, a programme for the introduction of this type of source in the production of flash estimates needs to be accompanied by the research of this subject.

6.5. Joint effort on the development of appropriate prediction models

Although in this paper we focus on the challenges of using web activity data in the prediction of socio-economic indicators, the development of appropriate prediction models is also very important. The model we present in this paper is very simple and serves only the purpose of illustrating how simple it can be to improve the accuracy of the predictions with web search
data from Google Trends. The use of such data in flash estimates would require the development of more sophisticated models, possibly including other variables.

In order to assure transparency, the development of such “production models” should be done in an open way in discussion with stake-holders, such as the policy makers in the European Commission and in the European Central Bank in the case of Europe, and also between statistical offices and involving academic researchers. Other reasons to do it collectively are the need to assess the international comparability and because there are lessons to be learned from each other.
References


27


[24] T. Harford, Big Data: are We Making a Big Mistake, Financial Times Magazine (2014), http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz2xlNF6ljV, last accessed on 30 September 2014
Figure captions

Figure 1: Prediction at month t done by fitting the models only on data at previous months.

Figure 2: Deviation (actual value – predicted value)/actual value; models fitted on all previous months only.

Figure 3: Prediction at month t done by fitting the models only on data at previous months.

Figure 4: Deviation (actual value – predicted value)/actual value; models fitted on all previous months only.
Figure 1
Figure 2

Deviation of predicted values from Eurostat data, France
Figure 3

Predictions based on all previous data points

- Eurostat data
- AR model
- AR model adjusted on google trends

Unemployment (1000 persons)

Time

2010M02 2010M10 2011M02 2012M02 2012M10 2013M02 2013M06 2014M02
Figure 4

Deviation of predicted values from Eurostat data, Italy