2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8576

Using big data in official statistics: Why? When? How? What for?

Mazzi, Gian Luigi
Technical Director, GOPA, Luxembourg

Abstract

This paper analyses the potential usefulness of big data in official statistics starting from four key questions such as Why? When? How? and What for - should we use big data in official statistics? To derive some answers related to empirical cases. This paper presents a big data classification by types, which is then used to identify how big data can answer to specific information needs in key policy areas. Based on the findings of these investigations, some very provisional and subjective answers to the questions raised above are derived.

Keywords: *Big data, nowcasting, indicators, policy areas.*

This work is licensed under a Creative Commons License CC BY-NC-ND 4.0

1. Introduction

Whether or not big data should be used in official statistics and how far we should eventually go in this direction still remain open issues for official statisticians, data scientists and analysts. It is a fact that policy makers claim for more and more information which cannot always be found in official statistics. Big data have the potential to meet policy makers' needs but, due to their nature and characteristics they face official statisticians with new challenges.

The two main obstacles for the regular use of big data in official statistics, are constituted by the evidence that they are often based on non-requested information (not collected within a robust sampling frame) and by their often unstructured nature. There are still non or partially answered questions for their use in producing official statistics such as: Why should we use big data to produce official statistics? When big data could be useful? How big data should be used to produce statistical indicators? What big data should be used for?

In this paper, we are trying to provide some answers to the questions above without pretending to be neither exhaustive nor conclusive, but aiming to provide an additional contribution to the ongoing debate. To achieve this objective, we are showing how big data could be useful in providing relevant information for key economic and socioeconomics policies, also showing advantages and drawbacks with respect to traditional sources of information.

The paper is structured as follows: Section 2 will present a big data typology which will be the basis of our investigation in the rest of the paper; Section 3 will relate the various big data types to the information needs in designing and monitoring some relevant policies; and Section 4: will conclude by providing some tentative answers to the above raised questions.

2. Big data typology

Several ways of classifying and characterizing the big data ecosystem have been proposed in the literature. Probably the most known is the so-called 4V which identify big data according to 4 main characteristics: volume, velocity, variety and veracity. Alternatively the UNECE proposes the following classification in 3 groups of the big data ecosystem: human sourced information including social networks, traditional business systems and internet of things.

With a totally new focus, mainly oriented to big data modeling, Dornik and Hendry (2015) proposed a big data classification according to the size of the data set: tall (not many variables but many observation), fat (many variables and few observations) and huge (many variables and many observations). None of the above classifications/characterizations are

really fully satisfactory since they don't emphasize enough the crucial aspects represented by the very different sources originating big data. In this respect, we are proposing a so-called big data typology already proposed in Buono et al. (2017) which distinguishes big data into10 main types presented in the table below:

Table 1: Big data typologies

	Type	Main Utilisation
1	Financial market data	Macroeconomics, financial sector monitoring
2	Electronic payments data	Macroeconomics, inflation, consumers behavior
3	Mobile phone data	Labour market, sustainable development
4	Sensor data and the Internet of Things	Sustainable development, urban and environmental monitoring
5	Satellite image data	Sustainable development, economic growth and land utilisation
6	Scanner prices data	Macroeconomics, inflation, consumers behaviour
7	Online prices data	Macroeconomics, inflation, consumers behaviour
8	Online search data	Macroeconomics, sustainable development, human behavior
9	Textual data	Human sentiments, confidence, uncertainty
10	Social media data	Macroeconomics, sustainable development, human behavior

In the table above the 10 main big data types have been associated to policies and related statistical areas. In this way we have highlighted the potential usefulness of big data in relation to various statistical areas, either as a complement of traditional data sources or as an alternative to provide reliable data and to fill existing gaps. This link between big data types and statistical information needed to design and implement key macroeconomics and socioeconomics policies will be further addressed in the next session.

3. Big data contribution in designing and implementing key policies

Official statisticians are supposed to answer to policy needs providing all necessary information for the implementation, follow up and monitoring of policy actions. They

provide a reliable set of statistical indicators based on traditional sources of information, such as census, sampling/surveys, and administrative data, integrated and complemented by grossing up and estimation techniques. Unfortunately not necessarily official statistics meet policy maker needs especially in terms of timeliness, relevance and ability to describe some complex phenomena.

The size of such a gap between policy makers' needs and available data can vary from country to country and also over the time. As described further, extracting information available within the big data ecosystem can help in filling the gap between the policy makers demand and the official statistics supply.

3.1 Macroeconomic growth and stability policies

When looking at macroeconomic policies, the relevance of available official statistics is quite good especially in developed countries. On the other hand, the timeliness and the frequencies at which data can be available do not necessarly meet policy maker expectations. In this respect, big data can contribute to both, increasing the timeliness of macroeconomics aggregates (i.e. GDP and consumption) and to provide higher frequency estimates of the same variables or even of inflation.

3.1.1 Economic growth

In this context financial and electronic payment data have proven to produce good results either in estimating real time economic growth or in deriving higher frequency proxies on GDP or consumption Galbraith and Tkacz (2007) Stock and Watson (2002a), Giannone, Reichlin and Small (2008) and Aprigliano et al (2016). In particular to obtain higher frequency estimates such us at weekly and even daily frequency, it is necessary to build up a quite complex modeling structure combining a data selection or data reduction tool adapted to large scale data set with mixed frequency models such as UMIDAS or MIDAS.

Also, online search data and in particular Google search data pre-synthetized within the Google Trend application provide good quality nowcasting and advanced estimates for macroeconomics variables as shown by Koop and Onorante (2013), who use the Google Trend information to select the most appropriate nowcasting model. Alternatively, several authors such as Baldacci et al (2016) and Buono et al (2018) use Google Trend data as regressors in a generalized regression model.

3.1.2 Inflation

Inflation usually measured by the Harmonized Index of consumer process (HICP) is timely available and produced at monthly frequency, meeting most of the policy makers requests. Nevertheless, starting with the paper of Silver and Heravi (2001), a large literature on the use of scanner data to estimate inflation has demonstrated how this alternative source of

information can produce additional information not really present in the HICP data. In particular, they can provide higher frequency estimates of the inflation, more detailed information at product level as well as indications on retailers and consumers behavior. As a last consideration, we have to say that scanner prices data enable us to considerably reduce the burden on retailers and consumers and to reduce the production costs of inflation data. This is the main reason for which several countries such as the Netherlands and Luxembourg are planning to use in an extensive way scanner data for compiling their HICP.

3.1.3 Additional big data related information

As we mentioned in 3.1.2. scanner price data can provide useful insights on the retailers and consumers behavior. Furthermore, electronic payment data can help in better understanding the reaction of consumers to unexpected or exceptional events as shown by Galbraith and Tkacz (2011). By means of text mining and text analytics technics it is also possible to derive measures of the economic uncertainty based on textual information available in newspapers and other media. Examples of such uncertainty measures have been proposed by Baker, Bloom and Davis (2015) and Bacchini et al (2017). The use of textual information has also shown its relevance in calculating a daily business cycle indicator obtained by combining within a complex modeling structure, quarterly GDP data and daily textual information extracted from newspapers (see Thorsrud (2016). Finally, social network data can also provide useful information to study the changes in individual and collective mood or sentiment.

3.2 Labor market policies

They usually require both macro indicators such as employment/unemployment, job vacancies etc. and more and more micro indicators describing individual behavior or changes in the habits related to the employment status especially in developed countries. Official statistics provide a pretty detailed picture of the sector both at the macro and micro level. With some drawbacks related to a certain lack of timeliness and to the frequency at which different kind of information become available. The basis for labour market statistics is represented by the labour force surveys implemented in the large majority of the developed countries as well as in some of the emerging ones. The situation can be much more complex in developing countries or in underdeveloped ones where the lack of information can be really relevant to both at macro and micro level.

3.2.1 Employment and unemployment indicators

Those indicators are generally obtained by labor force surveys in developed countries being usually very reliable even if not necessarily timely available. Big data can help in increasing their timeliness either by using online search data (Google Trend), mobile phone

conversation and mobile phone positioning. As an example D'Amuri and Marcucci (2012) and Tuhkuri (2016) investigate the power of big data in nowcasting and forecasting Unemployment data by using Google Trend.

On the other hand Toole et al. (2015), forecasted the employment at regional and European countries level by using the call duration information and changing behavior in social communication related to the employment status. They use an innovative approach based on Bayesian classification models.

In emerging and especially developing and underdeveloped economies, the situation can be radically different, either because labor force surveys are not being yet implemented or because they are not producing fully reliable estimates. In this case, big data can become almost the primary source for providing information on the situation of unemployment and employment.

Finally, especially mobile phone data can help in providing more granular estimates of the employment/unemployment status in a fully consistent way with the aggregated data.

3.2.2 Additional information provided by big data sources

Thanks to the availability of big data sources, and especially mobile phone ones, it is also possible to derive very useful information on individual behavior in relation to the unemployment status Sundsøy et al. (2016). Finally Nomura et al (2017) show how using big data from online job search portals, in combination with data analytics tools can produce several aggregated and disaggregated indicators extremely useful in several areas related to labor market policies such as labor market monitoring and analysis, assessing demand for workforce skills, observing job-search behavior and improving skills matching, predictive analysis of skills demand and, finally experimental studies.

3.3 Sustainable Development Goals (SDGs)

In the context of the SDGs and related policy actions, the situation in terms of availability or traditional data source to support policy decision is much more complex than in the previous cases. Often, SDGs refer to complex phenomena such as poverty, well-being, social exclusion, etc., which are measured both by qualitative and quantitative indicators. Some weaknesses of the traditional information system and consequently of official statistics appeared since the beginning of the discussion of the SDGs. For this reason, the attention was moving to alternative data sources especially big data. This is one of the main reasons for which traditionally, the so-called big data revolution has been strongly associated to SDGs activities. Going into many details by considering individual goals and discussing how big data can contribute to their measurement and achievement goes largely out of the scope of this paper also taking into account space limitations.

Nevertheless, it is worth to say that, especially some ,big data categories, such us mobile phone calls and positioning, satellite images and IOT and social networks can be particularly relevant in the context of the SGDs while others such as financial market data , electronic payment data etc., are less helpful. Obviously, the lack of official statistics and the big data availability also strongly depend on the degree of country development especially with reference to underdeveloped and developing countries. Furthermore when looking at big data it is important to carefully analyze their spatial and cross-sectional coverage in order to avoid providing misleading information.

4. Conclusions

In this paper, we have briefly investigated the usefulness of various typologies of big data in answering to policy needs in different economic and socio economic areas. We can now provide some answers to the question formulated in section 1. We would like to stress that the answers provided here reflect personal opinions and experiences and won't pretend to be generally accepted.

Why to use big data? Because they contain an incredible and still largely unexploited amount of information of which statisticians and policy makers could benefit.

When using big data? They could be used whenever possible with traditional data source to circumvent: unsatisfactory timeliness, the lack of coverage/relevance of traditional data sources, impossibility of measuring some phenomena via surveys/sampling.

How to use big data? They should be used within a methodological sound and robust framework using advanced tools and methods especially designed to deal with specific big data features. They have also to be used in a careful manner meaning that big data should be used being aware of their limitation and drawbacks deriving from the way in which they are collected.

What should be big data used for? Increase data timeliness by means of nowcasting and advanced estimates made available already during the reference period as well as to produce high frequency estimate such as daily or weekly frequency; produce more reliable and more granular estimates of given phenomena; construct new indicators measuring phenomena for which traditional data source are weak or unavailable; provide indicators of mood, sentiment or individual and collective behavior.

In our opinion, the outcome of this paper shows quite clear that big data have the potential to complement and supplement traditional data sources (not to replace them in the production of official statistics).

References

Aprigliano, V., Ardizzi, G., Monteforte, L. (2016). Using the payment system data to forecast the Italian GDP, *Bank of Italy, Working Paper*.

Baker, S.R., Bloom, N., Davis, S.J. (2015). Measuring Economic Policy Uncertainty". *NBER Working Paper Series, Working Paper* 21633.

Bacchini, F., Bontempi, M.E., Golinelli, R., Jona-Lasinio, C. (2017). Shortand long-run heterogeneous investment dynamics". *Empirical Economics*, DOI: 10.1007/s00181-016-1211-4.

Baldacci, E., Buono, D., Kapetanios, G., Krische, S., Marcellino, M., Mazzi, G-L., Papailias, F. (2016). Big Data and Macroeconomic Nowcasting: From data access to modelling, *EUROSTAT Statistical Working Paper collection*.

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G.L., and Papailias, F. (2017), Big data types for macroeconomic nowcasting, *Eurona*, 94-145

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G.L., (2018) and Papailias big data ecnomotrecis nowcasting and early estimates, *forthcoming in Bidsa working paper series*.

D'Amuri, F., Marcucci, J. (2012). The Predictive Power of Google Searches in Predicting Unemployment. *Banca d'Italia Working Paper*, 891.

Doornik, J. A., Hendry, D. F. (2015). Statistical Model Selection with Big Data. *Cogent Economics & Finance*, 3(1), 2015.

Galbraith, J.W., Tkacz, G. (2007). Analyzing Economic E_ects of Extreme Events using Debit and Payments System Data". *CIRANO Scienti_c Series, Working Paper* 2011s-70.

Galbraith, J.W., Tkacz, G. (2011). Electronic Transactions as High-Frequency Indicators of Economic Activity". *Bank of Canada, Working Paper* 2007-58.

Giannone, D., Reichlin, L., Small, D. (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data", *Journal of Monetary Economics*, 55, 665-676.

Kapetanios, G., Marcellino, M., Papailias, F. (2017). Big Data and Macroeconomic Nowcasting, *Eurostat Working Paper*, ESTAT No 11111.2013.001- 2015.278.

Koop, G., Onorante, L. (2013). \Macroeconomic Nowcasting Using Google Probabilities". *European Central Bank Presentation*.

Nomura, S., Imaizumi, S., Areias, A., Yamauchi, F. (2017). Toward Labor Market Policy 2.0: The Potential for Using Online Job-Portal: Big Data to Inform Labor Market Policies in India1, *The World Bank, Policy Research Working Paper* 7966.

Silver, M., Heravi, S. (2001). Scanner Data and the Measurement of Ination. *The Economic Journal*, 111, F383-F404.

Stock, J., Watson, M. (2002a). Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, 297, 1167-1179.

Sundsøy P., Bjelland J., Reme B., Jahani E., Wetter E., Bengtsson L. (2016). Estimating individual employment status using mobile phone network data. arXiv:1612.03870 [cs.SI] (Dec 2016).

Thorsrud, L.A. (2016). \Words are the new numbers: A newsy coincident index of business cycles". *Norges Bank Working Paper Series, Working Paper* 21-2016.

Toole, J.L., Lin, Y.-R., Muehlegger, E., Shoag, D., Gonzalez, M.C., Lazer, D. (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface*, 2015 12 20150185.

Tuhkuri, J. (2016). Forecasting unemployment with google searches. *ETLA Working Paper* No 35.