Big data: The prospects for public statistics?

Didier Blanchet,
Director of the economic studies, French National Institute of Statistics and Economic Studies (INSEE),
&
Pauline Givord,
Head of SSPLab, INSEE

Abstract:
Public statistics draw on a wide variety of data. National offices of statistics try to extract from data the information that, relevant for discussions on social issues, is as comparable as possible over time and space. Are big data going to upend this work? Efforts are being made to find ways to make big data and public statistics complementary — a quest that implies identifying the real comparative advantages of big data. Three examples illustrate this: the price index, the analysis of the economic situation, and the production of experimental statistics for filling gaps in existing data or on questions (*e.g.*, the digital economy, the sharing economy, or the monitoring of sustainable development) not adequately addressed by ordinary data collection methods.

Are big data going to radically alter the way of producing the statistics used as an input in public debates and the implementation of economic and social policies? At present, the work of producing them mainly relies on national offices of statistics, which draw from quite diverse, often huge sources: censuses, surveys, registries, administrative sources (chiefly social and fiscal). Surveys and censuses follow protocols that, as stable as possible, ensure the coherence of results over time. To their advantage, administrative sources limit the burden of obtaining answers from persons in a survey. Since their contents are not directly formatted to the needs of statistics however, processing the data from them calls for intense work to curate and consolidate them.

The core of the public statistician's trade is to collect, process and structure data from these sources. This work is organized under international agreements and regulations and, in Europe, subjected to peer review procedures. Representatives from other institutes regularly inspect each national office of statistics, the goal being to guarantee both the quality and independence of the statistics produced.[1]

By recalling how statistics are currently produced, we can better formulate the questions now being raised by the processing of big data (BLANCHET & GIVORD 2017). Big data refer, above all, to the masses of data generated by expanding digital technology: information directly available on the Web, the traces left there by cybernauts, data on transactions, and, too, the data recorded on networks or by sensors (*e.g.*, via mobile telephones or satellites). Given their volume and the fact that they are often poorly structured, specific problems crop up in processing them. Extracting the pertinent information from big data requires substantial investments, which can soon turn out to be obsolete given the rapid changes in digital technology and its uses.

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references.

Furthermore, big data are often generated from the private sector's activities, and are held by the firms that generate them. Should we expect that public statistics will be sidelined because of a radical alteration of how economic and social information is produced and diffused? We imagine the risks: the proliferation of rival information using data that has been neither harmonized nor stabilized over time, the absence of a guarantee of neutrality when these data are processed and circulated (without surveillance procedures of the sort required for public statistics). The right approach for national offices of statistics to take is to explore complementarities between these new data sources and those that already exist. How to combine the two? How can these offices gradually integrate big data in their own work of processing information?

The consumer price index, an example

Measuring prices illustrates many of these problems. At present, prices are mainly monitored by directly collecting them at points of purchase in retail outlets. To its advantage, this method, though unwieldy and expensive, works for all types of goods. To measure the consumer price index, the French National Institute of Statistics and Economic Studies (INSEE: Institut National de la Statistique et des Études Économiques) records approximately 200,000 prices every month in nearly 30,000 retail outlets.

Breakthroughs in digital technology offer two new data collection methods. The first is to gather in real time prices from e-business platforms on the Internet. The MIT Billion Prices Project (BPP) — with no relation to official statistics — scrapes data from the Web on prices. It arose out of a dispute about the official statistics for measuring inflation in Argentine in 2007. Mistrust has long been directed at price indices. In Argentine, measurements from independent local authorities and from studies conducted by economists corroborated this mistrust: the official inflation rate was around 7% per year whereas the estimates from alternative sources hovered around 20%. Scraping data from the websites of major retailers confirmed this difference and also proved the reliability of this data collection method. Born from this experiment, the BPP was launched in 2008 as an academic project with the objective of covering as many countries as possible. By 2010, it had reached the target (symbolized by its name) of one billion prices per year. The change of scale required raising funds, whence the creation of a firm[2] that monitors 15 million products in 900 retail outlets in 50 countries (CAVALLO & RIGOBON 2016).

The expansion of the project to other countries has turned up a reassuring result: the failure observed in the Argentinian case is an exception. For the United States and the eurozone, the BPP and official price indices have proven concordant, in particular about the very low level of inflation during recent years. While this finding provides comfort for traditional data collection methods, it might also be an argument for replacing them with the new one. This is not the direction taken by the majority of national offices of statistics, even though some offices are studying the possibility of data scraping. In France, prices for shipping and air transportation are now retrieved from the Web.

For goods however, INSEE has been using another method for massively collecting electronic data, namely from the cash register slips recorded for payments in stores. To their advantage, these slips contain information about both the price and the quantity purchased — a direct source of the two types of information needed to build the consumer price index. In France, the use of the information from this source ensued from a program launched by INSEE in 2015 as an experiment but that should be full-scale by 2020. It now has a solid legislative framework; the act of law for a digital republic has laid down the conditions under which big retail outlets may make this type of data available.[3] Public statistics will be using data from retail outlets in the

[2] www.pricestats.com

[3] Act n°2016-1321 of 7 October 2016 for a "digital republic" available at https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id.

context of a stable contractual relationship with them — without scraping the data like the BPP. The trend represented by the switch to cash register data mirrors, a few decades later, what happened in the case of using data from administrative sources. Obtaining access to these sources did not occur immediately (and some public offices of statistics have less access to such sources than is the case nowadays in France).

"Nowcasting" and the illusion of velocity

The example of measuring prices illustrates several of the *v*'s that are often mentioned to define big data.

Let us first mention the VOLUME of big data, since much is expected of the availability of data more granular than what is contained in manual records.

VARIETY tends to imply the technique for "webscraping". This *v* is, in the case under discussion, a drawback rather than an asset. The BPP accomplished the feat of apparently producing relevant statistics from quite disparate information gathered from e-business websites. However cash register data have the advantage of being closer to the format of traditional data sources, even though we should not underestimate the cost of applying a single format to data coming from several information systems.

VERACITY might be mentioned even though it does not provide a comparative advantage. The prices collected from store shelves, from e-business websites or from cash registers are, in each case, as "true" as those collected by the other methods. The only additional advantage for cash register data is that they report the discounts offered when purchases are made in brick-and-mortar stores. On the contrary, scraped data have the disadvantage of not covering all goods or all sales methods, since scraping is restricted to online sales.

What about VELOCITY, a *v* touted by the BPP? Indeed, gathering prices on line should make it possible to detect sudden increases or decreases in nearly real time. In France however, this advantage holds for a very limited period, since the first estimates of a given month's consumer price index are released once the month has ended. Moreover, very few changes are made to them before definitive publication in the middle of the month.

Can gains in velocity be more decisive in other segments of business forecasting? At present, the analysis of the macroeconomic situation relies on a chained series of ever more granular sources. Qualitative business surveys and quantitative indices of production or sales are published monthly. They are the main sources for short-term analyses prior to the use of administrative sources and more extensive surveys for eventually establishing detailed annual accounts. Since the start of 2016, the first quarterly estimates are released one month after the quarter ends. The period is, therefore, very short for statistics that can be, and inevitably are, revised but that have the advantage of being grounded on stable protocols and data representative of the whole economy.

Is it possible to do any better by using information extracted from big data? Several experiments have tried to draw information from cybernauts' behavior patterns, such as the frequency of searches using key words on Google or the tonality of exchanges on the social media — the intent being to detect social or economic trends in nearly real time. Underlying these experiments is the hypothesis that search behaviors are predictive of the order of magnitude of what is to being sought. For example, the frequency of searches using the words "job" or "unemployment insurance" is probably correlated with the situation in the labor market; and searches for information on given consumer goods or services, with eventual purchases of them. This approach obviously reminds us of how web data have recently been used to predict voting patterns, the hypothesis being that trends observed on the Internet could predict the vote more reliably than traditional public opinion polls. As we know, the results have been quite ambivalent. This approach sometimes does much better than the usual methods but also, sometimes, much more poorly — the risk being that successes as well as failures are but happenstances. The same

mishap has occurred with another use promoted by Google, namely searches for medical terms as a way of monitoring in real time the influenza epidemic in the United States. This experiment was pursued for a while and then abandoned (LAZER *et al*. 2013).

In general, studies in economics have shown that such techniques produce at best "marginal" information compared with the contents of business surveys (BORTOLI & COMBES 2014). The data obtained by using them make a substantial contribution only in countries that do not have well-developed services for providing economic statistics.

Experimental statistics to fill data gaps

Tapping new data sources, such as cash registers, seems to hold a potential not for replacing existing data collection methods but for completing information in fields not yet fully covered by public statistics. One of these fields is the digital economy. Harmonized European surveys are already offering information on how firms and households use digital devices and equipment. Nonetheless, we could learn more about how firms are swept up in the digital economy by drawing on the contents of corporate websites or on what is said about firms in the online press or on the social media. Public-private partnerships have conducted experiments of this sort in the United Kingdom and Netherlands (NATHAN & ROSSO 2013, OOSTROM *et al*. 2016). Banking data also has a potential as a source of information on the income derived from new forms of work related, in part, to the digital economy (FARRELL & GREIG 2016). The analysis of consumer behavior toward online services can also help us gauge the monetary value of electronic services (COHEN *et al*. 2016, BRYNJOLFSSON *et al*. 2018).

Explorations are being carried out in fields other than econometrics in the strict sense of the word, in particular to assess the distance from the sustainable development goals set by United Nations Agenda 2030 in September 2015. Satellite data could be used to provide a granular description of the soil (type of crops, wetlands, etc.) and land (urbanization, compacting). This description could conceivably be used to make indicators corresponding to objectives for the conservation of ecosystems or for food security via sustainable agriculture.[4]

In France, experiments are under way in three fields. Data are being processed: from platforms offering rentals between private persons in order to obtain fuller statistics on tourism (FRANCESCHI 2017); from the registration of mobile telephones in order to obtain data by local area for completing information from administrative sources and the census on the country's social geography; and from online job offers in order to better describe the labor market. Although these experimental statistics will produce new information about topics that are emerging or need to be better covered, they cannot immediately be placed on the same level as the regular statistics produced over long periods. These new data sources must be tested case by case. Extracting stable, conceptually coherent data is not to be taken for granted, even less so when they are not structured. Big data have an incontrovertible potential, and public offices of statistics are trying to benefit from it — but far from the myth of a universal, low-cost response to the demand for ever more rapid, more reliable and more numerous statistics.

[4] SOES (2009), *CORINE Land Cover France Guide d'utilisation*, Service de l'Observation et des Statistiques, Commissariat Général au Développement Durable, Ministry of the Environment. See too the report by a UN work group steered by the Australian Bureau of Statistics, https://unstats.un.org/bigdata/taskteams/satellite/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf.

References

BLANCHET D. & GIVORD P. (2017) "Données massives, statistique publique et mesure de l'économie", *L'Économie française* (Paris: INSEE Références), pp. 59-77.

BORTOLI C. & COMBES S. (2015) "Apports de Google Trends pour prévoir la conjoncture: des pistes limitées", *Note de conjoncture* (INSEE), March, pp. 43-56.

BRYNJOLFSSON E., EGGERS F. & GANNAMAMENI A. (2018) "Using massive online choice experiments to measure changes in well-being", *NBER Working Paper*, 24514, 75p. Available at: http://www.nber.org/papers/w24514.

CAVALLO A. & RIGOBON R. (2016) "The Billion Prices Project: using online prices for measurement and research", *Journal of Economic Perspectives*, 30(2), pp. 151-178.

COHEN P., HAHN R., HALL J., LEVITT S. & METCALFE R. (2016) "Using big data to estimate consumer surplus: The case of Uber", *NBER Working Paper*, 22627, 43p. Available at: http://www.nber.org/papers/w22627.

FARRELL D.& GREIG F. (2016) "Paychecks, paydays and the online platform economy: Big data on income volatility", JPMorgan Chase Institute, 44p. Available at: https://www.jpmorganchase.com/corporate/institute/document/jpmc-institute-volatility-2-report.pdf.

FRANCESCHI P. (2017) "Les logements touristiques de particuliers proposés par Internet", *INSEE Analyse*, 33. Available at: https://www.insee.fr/fr/statistiques/2589218.

LAZER D., KENNEDY R., KING G. & VESPIGNANI A. (2014) "The parable of Google Flu: Traps in big data analysis", *Science*, 343(6176), pp. 1203-1205.

NATHAN M., ROSSO A., GATTEN T., MAJMUDAR P. & MITCHELL A. (2013) *Measuring the UK's digital economy with big data*, report of the National Institute of Economci and Social Research, 43p.. Avaiable via: https://www.niesr.ac.uk/sites/default/files/publications/SI024_GI_NIESR_Google_Report12.pdf.

OOSTROM L. WALKER A.N., STAATS B., SLOOTBEEK-VAN LAAR M., AZURDUY S.O. & ROOIJAKKERS B. (2016) "Measuring the Internet economy in the Netherlands: A big data analysis", *CBS Working Paper*, 14, 58p. Available at: https://www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf.