DISCOVER SOCIETY
MEASURED – FACTUAL – CRITICAL

WHAT DOES BIG DATA MEAN FOR OFFICIAL STATISTICS?

By discoversociety July 30, 2015 3 Comments Articles, Issue 23

Rob Kitchin (National University of Ireland, Maynooth)

For the past couple of centuries National Statistical Institutions (NSIs) have produced a range of official statistics using two main sources of data: surveys which they conduct, and public sector administrative data. The generation of big data across a number of domains has the potential to be a significant disruptive innovation to the work of NSIs and the production of official statistics, providing new sources of highly temporal and widely sampled data that might supplement, improve or replace existing datasets and statistics, or provide entirely new statistical outputs (Florescu *et al*. 2014).

For example, mobile phone data might be used in the production of tourism statistics; web scraped data relating to real estate used to help to calculate property price statistics, or employment opportunities used to help to calculate labour/employment statistics; social media data used to calculate sentiment towards different issues, health and wellbeing statistics, or consumer confidence; sensor data used for traffic and pollution statistics; smart meter data used for energy statistics; satellite images for land use, agriculture and environment statistics; and supermarket scanners used for price or household consumption statistics (ESSC 2014).

Importantly, big data offer the opportunity to produce more timely official statistics, drastically reducing their processing and calculation, and to do so on a rolling basis (Eurostat 2014). For example, rather than it taking several weeks to produce quarterly statistics (such as GDP), it might take a few minutes or hours, with the results being released daily. In this sense, big data offers the possibility for 'nowcasting' – the prediction of the present (Choi and Varian 2011: 1). Moreover, since big data tend to be exhaustive to a system, rather than sampled (i.e., it is a count of all cars on the network; the prices of all houses for sale; all the smart meters on a network; all the transactions at a checkout; all the land in a jurisdiction), they have strong population and spatial coverage at the level of the individual.

Further, big data tend to be direct measurements of a phenomena, and provide a reflection of actual transactions, interactions and behaviour, unlike surveys which reflect what people say they do or think. In the developing world, where the resourcing of NSIs has sometimes been limited and traditional surveys are often affected by external factors (e.g., political pressure, war, etc), big data are seen as a means of filling basic gaps in official statistics. An additional advantage is that big data offers the possibility to add significant value to official statistics at marginal cost, given the data are already being produced by third parties (Struijs *et al*. 2014).

Not unsurprisingly, given these qualities, big data has captured the interest of NSIs and related agencies such as Eurostat, the European Statistical System, United Nations Economic Commission for Europe (UNECE), and the United Nations Statistical Division (UNSD). In 2013 the Heads of the NSIs of the EU signed the Scheveningen Memorandum to examine the use of big data in official statistics and to formulate a roadmap for their incorporation into their workflow. In 2014 the UNSD established a Global Working Group on Big Data for

Official Statistics (comprising of representatives from 28 developed and developing countries). The approach adopted has been one of collaboration between NSIs, trying to develop a common strategic and operational position, including the creation of a big data 'sandbox' environment to experiment with big data and associated methods, techniques, models, software, equipment and resourcing.

In 2014, approximately 40 statisticians/data scientists from 25 different organisations were working with the sandbox (Dunne 2014).

What these organisations are discovering is that whilst big data offers a number of opportunities for NSIs, they also offer a series of challenges and risks that are not easy to handle and surmount. A primary issue is gaining access to the data. Although some big data are produced by public agencies, such as weather data, some website and administrative systems, and some transport data, much big data are presently generated by private interests such as mobile phone, social media, utility, financial and retail companies. Big data are valuable commodities to these companies, either providing a resource that generates competitive advantage or constituting a key product. Gaining access to such data requires NSIs to form binding strategic partnerships with relevant companies or creating/altering legal instruments (such as Statistics Acts) to compel companies to provide such data and neither approach will be easy to negotiate or implement.

Once data has been sourced, it needs to be assessed for its suitability for producing official statistics, a purpose for which it has not been generated. A key issue in this respect is the representativeness of the data. NSIs carefully set their sampling frameworks and parameters, whereas big data although exhaustive are generally not representative of an entire population given they only relate to whomever uses a service. For example, credit card data only relates to those that possess a credit card and social media data only relates to those using that platform, which in both cases are stratified by social class and age (and in the latter case also includes many anonymous and bot accounts). Further, NSIs spend a great deal of effort in establishing the quality and parameters of their datasets with respect to veracity (accuracy, fidelity), uncertainty, error, bias, reliability, and calibration, and documenting the provenance and lineage of a dataset, whereas these are largely unknown with most big data.

Once the suitability of the data is established, an assessment needs to be made as to the technological feasibility regarding transferring, storing, cleaning, checking, and linking big data, and conjoining the data with established existing official statistical datasets. Moreover, it needs to be established whether big data processing and analysis can be integrated into existing workflows and how big data infrastructures are aligned with existing infrastructure. In particular, there is a real challenge of developing techniques for dealing with streaming data, such as processing such data on the fly (spotting anomalies, sampling/filtering for storage) (Scannapieco *et al.* 2013), and in producing new methodological techniques and analytics for making sense of large, dynamic datasets.

Given these uncertainties and challenges, it is clear that there are a number of risks associated with using big data for official statistics. A key risk is gaining access to the necessary data and maintaining continuity of access. NSIs have little control or mandate with respect to big data held by private entities, nor is there an assurance that the company and its data will exist into the future. If access is denied in the future, or the data production is terminated, then there is a significant risk to data continuity and time-series datasets, especially if existing systems have been replaced by the new big data solution.

Moreover, in partnering with third parties NSIs lose overall control of generation, sampling, and data processing and have limited ability to shape the data produced, especially in cases where the data are the exhaust of a system that are being significantly repurposed (Landefeld 2014). This raises an additional question concerning the management of quality assurance and risks damaging a NSI's reputation as a fair, impartial, objective, neutral provider of high quality outputs. Further, partnering with a commercial third party and using their data to compile official statistics exposes the reputation of a NSI to that of the partner. A scandal with respect to data security and privacy breaches, for example, may well reflect onto the NSI. Such breaches also become a concern for NSI's themselves, with big data increasing the challenge of securing data by providing new types of systems and databases, and new flows of data between institutions. As the Wikileaks and Snowden scandals and other data breaches have demonstrated, public trust in state agencies and their handling and use of personal data has already been undermined. A similar scandal with respect to a NSI could be highly damaging. Similarly, given big data is being repurposed, often without the explicit consent of those the data represent, there is the potential for a public backlash and resistance to their re-use.

There is also a risk related to competition and privatisation. If NSIs choose to ignore or dismiss big data for compiling useful statistical data then it is highly likely that private data companies will fill the gap, generating the data either for free distribution (e.g. Google Trends) or for sale. They will do so in a timeframe far quicker (near real-time) than NSIs are presently working, perhaps sacrificing some degree of veracity for timeliness, creating the potential for lower quality but more timely data to displace high quality, slower data (Eurostat 2014). Data brokers are already taking official statistical data and using them to create new derived data, combining them with private data, and providing valued-added services such as data analysis. They are also producing alternative datasets, registers and services, combining multiple commercial and public datasets to produce their own private databanks from which they can produce a multitude of statistics and new statistical products (Kitchin 2014). The danger for NSIs is that their role as the predominant provider of official statistics will diminish or that their services are privatised like other parts of the public sector; the danger for the public is that current official statistics are replaced by more timely but less stable and weaker quality products.

The growing generation of big data presents NSIs with a set of opportunities, challenges and risks. Whilst some statisticians at NSIs are cautious about embracing big data, worried about their effect on the quality of the official statistics, others are enthusiastic about the data deluge and the potential for new, improved and more timely outputs. Given the uncertainties, the current approach being taken by NSIs seems sensible: working together to test the suitability of big data for official statistics, assess the implications to their practices and workflows, and to develop a coordinated, strategic response. Indeed, whilst it is good to embrace new innovations, there are a still many open issues that require much thinking, debate, negotiation, and resolution to ensure that any use of big data improves official statistics rather than weakening them.

References:

Choi, H. and Varian, H. (2011) **Predicting the present with Google Trends**. Google Research.
Dunne, J. (2014) Big data …. now playing at "the sandbox". Paper presented at The International Association for Official Statistics 2014 conference, 8-10 October, Da Nang, Vietnam.
ESSC (2014) ***ESS Big Data Action Plan and Roadmap 1.0***. European Statistical System Committee, 26th September 2014.
Eurostat (2014) ***Big data – an opportunity or a threat to official statistics?*** Paper presented at the Conference of European Statisticians, 62nd plenary session, Paris, 9-11 April 2014.
Florescu, D., Karlberg, M., Reis, F., Del Castillo, P.R., Skaliotis, M. and Wirthmann, A. (2014) **Will 'big data' transform official statistics?**
Kitchin, R. (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. Sage, London.
Landefeld, S. (2014) **Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues**. Discussion Paper at International Conference on Big Data for Official Statistics, Beijing, China, 28-30 Oct 2014.
Scannapieco, M., Virgillito, A. and Zardetto, D. (2013) **Placing Big Data in Official Statistics: A Big Challenge?** Paper presented at New Techniques and Technologies in Statistics.
Struijs, P., Braaksma, B. and Daas, PJH. (2014) Official statistics and Big Data. *Big Data & Society* 1(1): 1–6.

Rob Kitchin *is an ERC Advanced Investigator on The Programmable City project and a Professor at the National Institute for Regional and Spatial Analysis at the National University of Ireland Maynooth. He is Principal Investigator for two data infrastructures – the All-Island Research Observatory and the Digital Repository of Ireland – and is author or editor of 22 books.* This article is a shortened and modified version of a paper forthcoming in the *Statistical Journal of the International Association of Official Statistics* titled 'The opportunities, challenges and risks of big data for official statistics'. A preprint version is available **here**. The research for this paper was provided by a European Research Council Advanced Investigator Award, 'The Programmable City' (ERC-2012-AdG-323636).

3 COMMENT RESPONSES

LEAVE A COMMENT

Comment:

Your comment..

Name: E-mail: Website:

Nickname.. E-mail.. Website..

Send

Notify me of follow-up comments by email.

Notify me of new posts by email.