Benefits and Challenges of using Big Data for Official Statistics

Ronald Jansen*, Ivo Havinga and Shaswat Sapkota
United Nations Statistics Division, New York, United States
jansen1@un.org, havinga@un.org and sapkota@un.org

Abstract

Big Data—data sources with a high volume, velocity, and variety—have become increasingly available driven by the general use and vast penetration of electronic devices and the rapid advancements in information technology, and, correspondingly, new tools and methods are being developed to uncover meaningful information from them. Big Data originates from a wide range of sources such as mobile phone, social media, electronic commercial transaction, sensor networks, smart meters, GPS tracking devices, or satellite images, which collectively touch upon almost all aspects of our daily lives; and Big Data is regarded as potentially providing great support to official statistics, and to the delivery of meaningful indicators for monitoring the post-2015 development agenda. However, there are several challenges that first need to be tackled before benefits of Big Data can be realized. This paper will systematically explore those challenges. Drawing from the United Nations Survey on Big Data Strategy and Project Inventory, and other relevant literature, the paper will highlight key obstacles that national statistical systems face to use Big Data for official statistics. Some of the challenges are: representativeness, biases, modelling and other methodological issues, access to data, privacy, transparency and management of public trust, and IT and financial challenges to implementing a Big Data project. The paper will also explore some of the ways—for example, public-private partnerships and legal frameworks—in which these challenges can be addressed. The UN Working Group on Big Data for Official Statistics has the mandate to provide a global work programme on Big Data, to promote practical use of Big Data sources, foster communication and build public trust in use of private Big Data for official statistics. As such, the paper will highlight its work in the current status of the mentioned challenges.

Keywords: mobile phone data, social media data, data access, UN Global Working Group on Big Data

1. Introduction

Big Data—data sources characterized by high volume, velocity, and variety—have become increasingly available. This is driven through two key processes: increased digitization of data including generation of new digital data, and significant reduction in costs of data collection, storage and analysis (OECD, 2013; United Nations, 2014)[1]. Big Data originates from a wide range of sources such as mobile phone, social media, electronic commercial transactions, sensor networks, smart meters, GPS tracking device, or satellite image, and touch upon almost all aspects of our day-to-day lives. Big Data sources have shown great potential to improve generation of official statistics, and numerous existing and planned initiatives continue to explore this data source further. The potential of this data source resides in the timely availability of large volumes of data and their low cost of generation, especially compared to national household and business surveys. Additionally, in the era of declining responses to surveys, Big Data can add great value in production of official statistics by reducing costs and decreasing response burden. However, to take advantage of Big Data for official statistics, there are several issues and challenges that need to be addressed.

[1] See also http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

This paper will provide a brief overview of the potential benefits and the challenges of using Big Data, and suggest some ways in which the challenges can be addressed. The next section will highlight, using recent examples, some of the potential benefits that have been realized by incorporating Big Data in production of official statistics. The following section will then discuss some of the challenges of using Big Data drawing from the UNSD/UNECE Survey on organizational context and individual projects of Big Data[2], various meetings and reports on the topic. That will be followed by suggestions on addressing these challenges.

2. Potential benefits of using Big Data sources for official statistics

The International Conference[3] on Big Data for Official Statistics held in Beijing, China, in October 2014, showcased a considerable number of Big Data projects that demonstrate how Big Data sources can improve official statistics. These cases not only touched upon different types of data sources such as mobile phone, GPS records, satellite imagery, geo-spatial information, and social media, but they also came from different parts of the world, including both developed and developing countries. Some of them are presented here as concrete examples to demonstrate that Big Data can be used to strengthen official statistics. Other examples that demonstrate the potential of Big Data sources for official statistics are also highlighted in this section. For instance, mobile phone data can provide great benefits in generating relevant official statistics for tourism, poverty mapping, and tracking of mobility patterns in case of disease outbreaks. Mobile phone and other location devices are particularly attractive in these contexts because of their widespread use, including in the developing world, and due to the availability of both active and passive positioning data (United Nations, 2015). For example, Statistics Netherlands has been using mobile phone data to estimate daytime population, and road sensor data to generate a wide range of traffic statistics. Similarly, the Italian statistical office has been exploring the use of mobile phone location data to study mobility of its citizens. Eurostat[4] conducted a feasibility study on using mobile positioning data for tourism statistics and found that mobile positioning data can supplement current official tourism indicators mandated by European regulation.

Several studies by UN Global Pulse and its partners have shown tremendous potential for further use of mobile phone data for humanitarian purposes. Global Pulse demonstrated[5] how seasonal mobility and livelihood patterns in Senegal could be studied using mobile phone data, which could then be used for humanitarian early warning mechanisms in times of exposure to shocks. Further, in a study[6] conducted in East Africa by the World Food Programme (WFP) and Global Pulse, high correlations were found between airtime credit purchases and survey results referring to consumption of several food items, such as vitamin-rich vegetables, meat and cereals. The study highlighted that airtime credit purchases could serve as a proxy indicator for food spending. In another study that involved the Government of Mexico, WFP and Telefonica Research, mobile phone data were found to be highly representative to estimate population mobility at times of flooding[7]. Mobile phone data could be used as a proxy indicator of population movement in areas where other data sources were not available or reliable.

Other Big Data sources, such as satellite imagery and geospatial data, have great potential to provide more frequent and timelier data at a disaggregated level, particularly in agricultural and environmental statistics. The Australian Bureau of Statistics is developing a methodology for predicting crop yields using satellite imagery data. This could complement and partly replace the existing surveys for

[2] See http://unstats.un.org/unsd/statcom/doc15/BG-BigData.pdf
[3] See http://unstats.un.org/unsd/trade/events/2014/beijing/default.asp
[4] See http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf
[5] See http://www.unglobalpulse.org/sites/default/files/UNGP%20Case%20Study_D4D%20Mobility_2015.pdf
[6] See http://unglobalpulse.org/sites/default/files/Topups_Food%20Security_WFP_Final.pdf
[7] See http://unglobalpulse.org/sites/default/files/Mobile_flooding_WFP_Final.pdf

measuring agricultural crop production. Likewise, China is implementing, in certain regions, major statistical surveys with remote sensing and satellite imagery data, and Columbia is using satellite imagery to complement the information in areas not surveyed on its National Agricultural Census and to support agricultural statistics broadly. Additional studies have shown that satellite imagery and remote sensing have also potential to derive estimates for economic activity, particularly in countries where economic data are not easily available (Vernon et al, 2012; Chen & Nordhaus, 2010) .

Social media data contain rich, detailed information on human behaviour and expressions, and that makes them highly suitable for exploring wide ranging issues. For example, Statistics Netherlands analysed publicly available social media messages to estimate consumer sentiments. The social media estimates of consumer sentiment were of similar quality as the traditional survey based estimates, and can be produced earlier, more frequently and at a lower cost. Other statistical offices, for instance Italy and China, are also using social media data to support labour statistics and improving monthly predictions and refining existing estimates (United Nations, 2015). Global Pulse worked with the Government of Indonesia and the Korea Advanced Institute of Science and Technology on a study[8] that examined how Twitter data could be used to 'nowcast' or provide real time food prices. It showed that price estimates based on public tweets were very close to the official market prices. Such tweets based estimates could therefore provide early warning for unexpected price spike at much lower costs than traditional data collection. The examples show that Big Data sources can strengthen production of official statistics. They can help improve the timeliness and frequency of official statistics, decrease costs associated with administering surveys and eventually reduce the response burden. These benefits are particularly important in the post-2015 development era, when the new development goals will place huge demands on the statistical systems. However, there are several challenges that need to be addressed before benefits of Big Data can be fully realized. These challenges have been identified at the Big Data Conference in Beijing, as well as through the UNSD/UNECE survey, and they represent themes that cut across the different sources of Big Data.

3. Challenges of using Big Data

(i) Access to datasets

The UNSD/UNECE Big Data Project Survey identified limited access to potential datasets as the biggest challenge for most Big Data projects. Many Big Data sources are generated and owned by private entities, making access difficult. Government bodies do not have the same authority or legislative framework to access such privately held data as to access for instance administrative data held by other Government agencies. In addition, mobile phone data or data from social networks contain personal information, which private companies are reluctant to share unless guarantees are in place that such access will not backfire on the company or impose considerable costs. An associated challenge of working with data producers from the private sector is that it can result in a loss of control by the statistical office on such data source, if the private producer decides to, for business reasons, change the definitions of data, collect different data, or altogether stop data collection (Landefeld, 2014).

(ii) Data quality and new methodologies

A second challenge in using Big Data concerns its quality and relevance for official statistics. This concern stems from the fact that Big Data are collected for non-statistical purposes and they usually do not meet the official statistical standards in terms of representativeness, concepts, and definitions (Landefeld, 2014). They are also not generated from instruments and methods that make it ready for consumption and dissemination as official statistics. The massive size and large dimensionality of data

[8] See http://www.unglobalpulse.org/nowcasting-food-prices

makes analysis of Big Data hard and conclusions derived from them not as robust (AAPOR, 2015). For example, the City of Boston had developed an application that used data from a smartphone's accelerometer and GPS to collect data on road conditions, including potholes, and report those back to the city's Public Work's department. However, the team identified that because the poor and elderly are less likely to carry a smartphone or download the app, that its release could have the effect of systematically directing city's resources to affluent neighbourhoods where smartphone-owners lived (United States, Whitehouse, 2014).This suggests the need for developing new methodologies and quality frameworks—a fact clearly reflected in the UNSD/UNECE survey where more than two-thirds of the respondents explained that they do not yet have a defined quality assessment framework for Big Data sources or output of analyses.

(iii) Privacy and trust

Privacy and the risk of disclosure of confidential information is another challenge facing the use of Big Data for official statistics. While survey data faces privacy considerations too, the issue of privacy becomes much more salient while using Big Data because it engages private data providers and usage of highly sensitive information that was collected for non-statistical purpose. Violation of privacy can have many serious consequences for the individual, particularly if it concerns sensitive areas such as healthcare and financial information; it could lead to discrimination against individuals and groups (US Whitehouse, 2014). This makes managing perceptions of risk as important as managing actual risks of disclosure. Additionally, the three Vs—velocity, variety, and volume—make Big Data very vulnerable for risk of disclosure of confidential data, especially when combined with other data sources. For instance, a person with knowledge of an individual's zip code, birthdate and sex can re-identify 80 percent of Netflix users (Ohm, 2010). UNSD/UNECE survey revealed that most institutes use the same privacy framework that applies to traditional statistics and very few have a privacy framework only for Big Data.

(iv) Financial issues and new IT requirements

Financial issues, particularly for improving IT infrastructure to handle and source Big Data are also a challenge for government offices. Private companies seem to be increasingly recognizing the value of the data they hold and are increasing the prices too (United Nations, 2014). Additionally hiring IT staff, data analysts and statisticians who are capable of working with Big Data requires significant additional investments. Finally, integrating Big Data sources into the existing data production processes require a restructuring of the statistical offices and this change needs to be effectively managed. The vast majority of respondents of the UNSD/UNECE survey indicated that so far they have relied more on providing training to existing staff than hire staff with new skills sets. The survey also indicated that only very few countries have defined a Big Data strategy or identified how Big Data will be integrated into existing work processes.

4. Addressing the Challenges to unleash the Benefits

(i) Quality framework

The aforementioned challenges can be addressed in a number of ways. Established quality frameworks can help ascertain Big Data source's adherence to the strict quality standard that is expected from official statistics. National statistical offices already make use of data quality frameworks—primarily developed for handling survey or administrative data—and they should be extended to Big Data too. The UNECE Big Data Quality task team reviewed several existing quality frameworks with respect to their applicability to Big Data and developed an updated framework for assessing Big Data sources'

quality. The resulting Big Data Quality Framework[9] consists of three hyper-dimensions. These are (a) source of the data, (b) metadata, and (c) the data itself. For each of the hyper-dimensions, there are quality dimensions with different factors to consider. For example, for the "data" hyper-dimension, validity is one of the quality dimensions, for which coherence between processes and methods, and observed data values are factors to be considered. Similarly, for the "source" hyper-dimension, institutional and business environment is a quality dimension, for which sustainability of the data provider is one of the factor to consider. Each of the factors has indicators, which makes it easier to systematically gauge the quality of a Big Data source. This data quality framework is closely aligned with the stages of the General Statistical Business Process Model (GSBPM)—input, throughput, and output—and covers the entire chain of the business process of statistical production.

(ii) Public-private partnerships

Partnerships and collaboration with private data providers can help gain access to Big Data sources. There are benefits for both the private data providers and the statistical agencies in collaborating to use Big Data for official statistics, and any basis of partnership should aim to leverage this overlap of interest. A successful public-private collaboration requires, in addition, transparency between the NSO and private data producer about data collection and estimation methods, and clear and strong rules to protect confidentiality and the proprietary nature of the private data (Landefeld, 2014). Articulating key principles that guide future collaboration, identifying good practices, and creating templates for data protocols, Memorandum of Understandings, non-disclosure agreements, and other contractual documents can help maintain predictability, improve trust, and strengthen effective partnerships. Mutually beneficial partnerships with other entities such as academia, research institutes, civil society organizations, to mention a few, are also important to realize benefits of Big Data. These partnerships can be crucial in mobilizing resources once the issue of access has been addressed. Partnerships with research universities, for example, can enable research in new methodologies, when existing NSO resources do not allow for doing so or when the specific skill sets are lacking. Initiatives such as data philanthropy (Stempeck, 2014)[10] can provide innovative modalities for such partnerships. Other partnerships with universities, for instance, 'college to government service' type internship programmes allow statistical offices to attract qualified talent to work in this area[11].

(iii) Maintaining privacy and confidentiality

To protect confidentiality of Big Data sources, legal rules, frameworks and policies on privacy need to be updated to reflect the specific Big Data characteristics. UNECE Big Data Task Team on Privacy has several recommendations[12]. Database activities and usage need to be constantly monitored. There should be good internal controls to help restrict the amount of access available to any given use within an organization. Similarly access to confidential data should be provided on the basis of principle of least privilege—provide only the minimum necessary level of access rights a user or role requires to complete their job. Technologies that encrypt data should be used extensively to avoid unwanted access, and statistical disclosure limitation should be applied extensively. A risk and consequence based approach where the security provided is commensurate with the risk and consequence of disclosure can help determine the correct level of access restriction and application of security technology.

Since statistical organizations need to manage perceptions of risk to secure continued cooperation of data providers, aforementioned disclosure techniques should be supplemented with legal frameworks that stipulate standards of security for sensitive data, and sanctions for non-compliance. An example of

[9] See http://www1.unece.org/stat/platform/display/bigdata/2014+Project
[10] See also http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience
[11] See also http://www-01.ibm.com/software/data/demystifying-big-data/
[12] http://www1.unece.org/stat/platform/display/bigdata/Deliverable+3%3A+Big+Data+Privacy

this is the Health Insurance Portability and Accountability Act, which strictly protects the confidentiality and security of healthcare information in the United States. Such regulations build confidence and trust with both the data providers and as well as the individuals whose personal data is being stored and transmitted. Since the users provide information about themselves to private data providers for specific purposes, clarification should be made if this consent extends to using their data for official purposes as well. Regulations/legislations that provide users a better understanding of what their data is being used for before it is being collected is crucial.

The GSMA guidelines[13] on the protection of privacy in the use of mobile phone data for responding to the Ebola outbreak is a good example privacy standards that has been initiated by the industry itself. It stipulates that mobile phone numbers of subscribers making and receiving the calls will be anonymized, that such data will not be transferred outside of the operator's system and premises and all analyses will be conducted in their premises too, and that no analyses will single out identifiable individuals. In addition, the guideline states that only the output of the analyses, and not real data, will be provided to relevant aid and government agencies.

A report by the Whitehouse (2014) of the United States identified advancing a consumer privacy bill of right, passing a national legislation on data breach, and amending existing privacy right legislations to reflect reality of the new Big Data sources, as some of the policy recommendations pertinent for protecting privacy. Some of the foundations of a consumer privacy bill of right would include the right to individual control of personal data, right to easily understand information about privacy and security practices, right to expect that personal data will be used in ways that are consistent with the context in which the consumers provided them, right to access and correct personal data, among others. Broad in scope, such measures would advance protection of privacy considerably.

(iv) Communication and advocacy

Using Big Data for official statistics will require the statistics community to understand the different aspects of Big Data. The community needs to know how Big Data sources can add value to official statistics, what the sources of Big Data are, and what Big Data can accomplish and what it cannot. Leaders of the national statistics office as well as decision/policy makers in key ministries need to be convinced of the benefit in using Big Data for official statistics. To that end, communication and advocacy is crucial. Effective communication can help raise awareness of relevant issues surrounding Big Data and can help start dialogue/processes to address challenges that exist. Communication is also crucial to develop alliances and partnerships to fully utilize benefits of Big Data for official statistics. Strategic communication that focuses on key themes and messages, and uses targeted information materials is important. Highlighting case studies of successful use of Big Data in official statistics, along with lessons learned, can help build confidence and reduce reticence to use, and investment of resources in expanding the use of Big Data for official statistics.

5. Conclusions

Big Data can contribute significantly to official statistics by complementing traditional sources of data such as surveys and administrative data. Many case studies and pilots conducted all over the world have helped demonstrate promising benefits of using Big Data. At the same time, these projects have also raised the statistical community's awareness of what Big Data can realistically accomplish, and what the challenges of using it for official purposes are. These challenges can be addressed, and more work in this rapidly evolving field will provide the official statistical community with insights necessary to use this data source to continue providing reliable, timely, and policy-relevant data.

[13] http://www.gsma.com/mobilefordevelopment/gsma-guidelines-on-the-protection-of-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-ebola-outbreak

International cooperation and collaboration is crucial for the further development of successful applications of Big Data for official statistics. The Global Working Group on Big Data for Official Statistics[14], established by the UN Statistical Commission at its recent 46th session[15], provides strategic vision and promotes practical use of big data sources, capacity-building, training and sharing of experiences. The group are developing pilot studies using both of the Big Data sources discussed in this paper as well as addressing the challenge of access to data and partnership, in particular, investigate the possibility of establishing umbrella agreements on access to data with companies operating globally.

References

Chen, X. & Nordhaus, W. (2010), Using luminosity data as a proxy for economic statistics. *NBER Working Paper No. 16317.*

Landefeld, S. (2014), *Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges and other Issues.* Prepared for the International Conference on Big Data for Official Statistics, Beijing, October 2014

OECD, (2013). Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by Big Data. *OECD Digital Economy Papers, No. 222*, OECD Publishing, Paris.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57(6), pp. 1701-1808.

Stempeck, M. (2014), *Sharing Data is a Form of Corporate Philanthropy.* Harvard Business Review, 24 July 2014

United Nations (2014), *Big Data and modernization of statistical systems. Report of the Secretary General.* Document E/CN.3/2014/11

United Nations (2015), *Report of the Global Working Group on Big Data for official statistics. Note by the Secretary-General.* Document E/CN.3/2015/4

United States, Whitehouse (2014), Executive Office of the President, *Big Data: Seizing Opportunities, Preserving Values.* May 2014

Vernon Henderson J, Storeygard A & Weil D. (2012). Measuring Economic Growth from Outer Space. *American Economic Review, American Economic Association*, 102(2), pp. 994-1028.

[14] See http://unstats.un.org/unsd/trade/bigdata/

[15] See http://unstats.un.org/unsd/statcom/sc2015.htm