Statistical Journal of the IAOS 31 (2015) 139–143
DOI 10.3233/SJI-150903
IOS Press

Editorial

Part 1: Introduction

This June the Journal is divided into three parts: Part 1 begins the Journal, as has now become standard, with an interview. The interviewee is Shigeru Ishakawa from Japan, the current President of the International Association for Official Statistics. A picture of him follows below.

Part 2 of the June issue consists of a special section on BIG DATA. The Journal should have more directly on BIG DATA or on related BIG DATA submissions to publish going forward. So look for them.

Part 3 completes this issue with a wide selection of papers. More is said on them later; but first we describe the BIG DATA revolution and then provide information about the individual papers.

Part 2: Big Data: Submissions

By Fride Eeg-Henriksen and Peter Hackl

Big Data is that notion which at present is probably most often referenced in the context of information science and information technology; this extraordinary interest or possible hype also affects official statistics. This is due to two factors:

- Big Data is a synonym for the availability of huge and growing amounts of digital information from all areas of human life.
- This deluge of information is a promise for seeing and understanding better and more in detail the reality and the relations which are ruling our world.

The notion big data

In spite of the wide interest in and the great popularity of Big Data, no clear and commonly accepted definition of the notion Big Data could be established so far [3]. Modern technological, social and economic developments including the growth of smart devices and infrastructure, the growing availability and efficiency of the internet, the appeal of social networking sites and the prevalence and ubiquity of IT systems are resulting in the generation of huge streams of digital data. The complexities of the structure and dynamic of corresponding datasets, the challenges in developing the suitable software tools for data analytics, generally the diversity of potentials in making use of the masses of available data make it difficult to find a suitable and generally applicable definition. The often mentioned characterization of Big Data by 3 – or more – Vs (volume, velocity, variety – as well as veracity and value), does not capture the enormous scope of the corresponding data sets and the extensive potentials of making use of these data. A highly relevant aspect is that Big Data are so large and complex that traditional database management tools and data processing applications are not feasible and efficient means. This is illustrated by a look at the categories of data sources which typically are seen in the context of Big Data: Such data sources may be

- Administrative, e.g., medical records, insurance records, bank records.
- Commercial transactions, e.g., credit card transactions, scanners in supermarkets.
- Sensors, e.g., satellite imaging, environmental sensors, road sensors.
- Tracking devices, e.g., tracking data from mobile telephones, GPS.

- Tracks of human behaviour, e.g., online searches, online page viewing.
- Documentation of opinion, e.g., comments posted in social media.

Big data and official statistics

For official statistics, some of these sources can be, or are hoped to be used as alternative or supplemental sources of data. In order to fulfil the obligations imposed by the statistical programme, the NSIs collect data in censuses or surveys, or they use data from administrative sources. The tendency to reduce the response burden to businesses and households and the growing demand for new statistical products let the NSIs look out for new sources of data. The increasing diversity and availability of administrative data are gaining relevance for the statistical production. But also other data sources as mentioned above are potentially very interesting as an input for official statistics. The use of such data may reduce the production time and costs of statistics, another fact that increases the attraction of these data sources.

The interest in using the mentioned data sources for the production of official statistics started about half a decade ago. Following a request of the participants at a High-Level Seminar on Streamlining Statistical Production and Services in 2012, the report "What does 'Big data' mean for official statistics?" [5] outlines the opportunities and challenges that Big Data poses for official statistics. In response to this report and following the proposal of a task team composed of representatives of 13 national and international statistics organizations, the Big Data Project [6] was established. The report "How big is Big Data?" [7] is a valuable and up-to-date description of the potential role of Big Data in official statistics, in particular of the challenges and requirements in terms of statistical methods including quality issues, of information technology, and of competencies and skills of the staff. In 2014, the United Nations Statistics Division established the UN Global Working Group with eight task teams on various topics including training and capacity building, mobile phone data, satellite imagery, and social media data [9]. Eurostat has been involved in all these activities from the beginning. The NSIs of a number of countries were pioneers in investigating the potentials of Big Data. The most prominent example is probably the Australian ABS; see Tam and Clarke (2015).

Evidence of the enormous interest of official statistics in Big Data is its being featured at various conferences, workshops, and other events during the recent years. Examples are the Eurostat Big Data Event in Rome (2014), the International Conference on Big Data for Official Statistics in Beijing (2014), and the UNECE NTTS 2015 Satellite Workshop on Big Data in Brussels (2015). Papers related to and reports on Big Data issues are playing a prominent role in events like the DGINS 2013 in The Hague, the Eurostat conference Quality 2014 (Q2014) in Vienna, the IAOS Conference 2014 (IAOS2014) in Da Nang, the UN Statistical Commission in 2015, and others. Many of the contributions are dealing with conceptual or strategic issues. However, reports on – mainly experimental – Big Data projects demonstrate how Big Data can be used in official statistics and what methodological and technological problems have to be solved.

Experiences and challenges

A closer look at these projects indicates that the statistical methods and the IT tools to be used in dealing with data from the typical Big Data sources are specific for the statistical product in mind. In the following areas, experiences in the use of Big Data in official statistics are available:

- ICT usage statistics.
- Price statistics.
- Labour market statistics.
- Tourism statistics.
- Traffic and transport statistics.
- Agricultural censuses and surveys.

Within the Big Data Project [6], several Big Data projects have been conducted by the NSIs from participating countries like the Netherlands, Italy, UK, Ireland, Australia, and Slovenia.

Data sources are the internet in the context of ICT usage, price and labour market statistics, mobile positioning data for tourism statistics, traffic loop detection sensors for traffic and transport statistics, and remote sensing from satellites for agricultural statistics.

Massive technological issues are related to the use of the internet as data source [1]. The amount of relevant data is usually huge and potentially distributed over a vast number of websites. This means that tools for retrieving the relevant websites are needed as well as tools for collecting the relevant data; web crawler and web scraper are the names of such tools, respectively. For dealing with the huge size of Big Data sources, special programming environments have been developed: e.g., Map-Reduce is a programming tool and

an associated environment for processing and generating large data sets. Programming frameworks like the open-source system Hadoop allow writing programs for processing Map-Reduce problems across huge datasets using a large number of computers and producing the output file within the file system named HDFS (Hadoop Distributed File System). Hadoop is well suited for long-running batch processes, such as data mining; tools like Big Query allow ad hoc queries that require quick results. The enormous challenges of Big Data to the information technology have the consequence that IT issues and – IT experts – are dominant in the discussion of Big Data. The use of Big Data in official statistics also needs adaptations in the statistical methodology. A key methodological concern in the context of internet and mobile positioning data is the representativity of the resulting statistics: Does the data selection mechanism allow stating the population for which the statistical product is representative, and does this population coincide with the target population for which the statistical product is designed? If not, how can the statistical product be interpreted? Other methodological issues concern the quality assessment of data and statistical products, the combination of data from different sources, the volatility of data sources over time, privacy concerns, confidentiality, and others. The methodological problems are specific for the data sets and have to be solved individually for each data set. The reports on the mentioned projects cover the methodological issues in a rather general mode. The representativity and also other quality aspects of Big Data-based statistical products are the crucial aspects for their trustworthiness. The use of Big Data in official statistics needs new skills and competencies. A survey among statistical organizations [8] indicates that about 37% work with Big Data, and other 43% plan to do that in the near future. Whereas most respondents say that their staff is familiar at an intermediate or advanced level with IT tools like Java, SAS, SQL databases, and R, no or only basic skills are said to be available in tools like Map Reduce and Hadoop. This result indicates both the strong interest of statistical organizations in Big Data and the need to enhance the competencies and skills in order to integrate the new potentials in the daily life of statistical organizations. Training courses within NSIs or by institutions like Eurostat as well as practical projects will help to build capacities in IT technologies but also in statistical methodology.

Big Data is supposed to create new opportunities within the area of dissemination of statistics, an area not yet highlighted much neither in the discussions so far nor in the articles of our special section. The increased interest for visualization of statistics is an important aspect of this. New possibilities for analysis and visualization certainly also creates challenges for capacity building in national and international statistical agencies.

The scope of this special section

This special section presents an overview article as well as experiences from pioneering projects indicating areas where Big Data sources may prove suitable as substitutes for traditional data sources or may allow the production of new statistics.

Steven Vale's "International Collaboration to Understand the Relevance of Big Data for Official Statistics" gives an account of the Big Data Project organized in response to the High-Level Group for the Modernisation of Statistical Production and Services [5]. The paper describes the goals and priorities of the project, the establishment of a computing environment, called "sandbox", for administrating and analysing large-scale datasets, and gives an overview of results. The high relevance of the project and its results are due to fact that seven teams have been engaged in practical experimental work in areas like consumer price index, mobile telephone data, smart meters, traffic loops, job portals, web scraping, and social media.

In the programmes of the Eurostat conference Quality 2014 (Q2014) and the IAOS Conference 2014 (IAOS2014) in Da Nang, a number of reports on concrete pioneering projects are given. The other articles are based on papers which have been presented at one of these two conferences.

One article reports experiences in the production of consumer price indices. "Web scraping techniques to collect data on consumer electronics and airfares for Italian HICP compilation" by Riccardo Giannini and his co-authors from ISTAT show in detail how web scraping techniques can be used to collect price data for consumer electronics and airfares. "The production of salary profiles of ICT professionals: Moving from structured database to big data analytics" by Ramachandran Ramasamy from the National ICT Association of Malaysia reports about the production of salary profiles based on data from a private sector online job registration system. The statistics are provided at a high level of disaggregation. The article discusses in detail issues of quality including consistency and stability in trends.

Two other contributions are dealing with methodological aspects. "Remake/Remodel – should big data change the modelling paradigm in official statistics?" by Barteld Braaksma and Kees Zeelenberg from Statistics Netherlands discuss the use of models for assessing and improving the representativity of Big Data sources. The paper sketches possible applications. The paper "Quality indicators for statistics based on multiple sources" by Mihaela Agafiţei and her co-authors from Eurostat analyses the appropriateness of standard quality measures for multiple source statistics and suggests improvements the merits of which have to be investigated in further work.

Conclusions

A general conclusion from the set of articles in this Special Section can be drawn as follows: The feasibility and the potentials of using Big Data in official statistics have to be assessed from case to case. In some areas the use of Big Data sources has already proved to be feasible. The choice of the appropriate IT technology and statistical methods must be specific for each situation. Also issues like the representativity and the quality of the resulting statistics, or the confidentiality and the risk of disclosure of personal data need to be assessed individually for each case. There is no doubt that Big Data will have a place in the future of official statistics, helping to reduce costs and burden on respondents. However, major efforts will be necessary to establish the routine wise use of Big Data, and new approaches will be needed for assessing all aspects of quality.

Part 3: Regular Submissions

By Greta Cherry

This is the called the Big Data issue. Those are the papers featured in Section 2 above. This June issue could also have been called the Rio World Statistical Congress (WJC) issue, since the WSC will be held in Rio de Janeiro, Brazil, in July.

Anyway, the papers chosen for the regular submissions were partly selected to help prepare for the 60^{th} WSC. For example, the Brazilian Census Paper by A.D. da Silva, M.P.S. de Freitas and D.G.C. Pessoa leads off this part of the June Issue.

In keeping with featuring timely articles, we next include a paper by M. Wolcott, C. Graham, C. Thompson and M. Tran on the Philippine Typhoon using a still novel data collection tool, twitter. This twitter paper continues our practice of publishing the volunteer work of the international Statistics without Borders (SwB). By the way, SwB just won the Di Vinci Prize for their outstanding work. Expect more SwB Journal papers in future issues in this series.

Another article on fabrication or curbstoning is found in this issue. This one is by Arthur Kennickell, entitled *Curbstoning and Culture.* Curbstoning, as a source of survey error, has been characterized as a "dirty secret." This series was begun with the *Identification of Partial Falsifications and Survey Data* paper in the September IAOS Journal.

At the time we did not realize how frequently new papers on data fabrication would emerge. But there are at least two more articles already scheduled for the September issue and likely more after that in later issues, as a workshop is being planned next year on this topic.

In this June issue we also conclude the remaining papers from the 59^{th} WSC held Hong Kong. These are by A. Wallgren, P. Stender, and J. Pannekoek.

A pair of papers follows on a topic that breaks our hearts. These articles are by M. Price and P. Ball, entitled *Selection bias and the statistical patterns of mortality in conflict*, plus a paper by Hasan Abu-Libdeh, entitled *Role of official statistics in situations of conflict and non-conflict*. Sadly, statistically measuring man's inhumanity to his fellow man is a role for Official Statisticians. These papers are part of a series that may continue indefinitely, despite our wishes to the contrary.

The next three papers begin with a modest-sized fine paper with a long title by G. Neideck, P. Siu and A. Waters, *Meeting national information needs on homelessness: Partnerships in developing, collecting and reporting homelessness services statistics*. There are then papers by Ilka Willand, *Beyond traditional customer surveys: The reputation analysis* and by Paul Ross, entitled *Understanding customer needs*. It is hard to believe that Dr. Ross is 88 years old.

Now we end this our largest entirely open issue with three remaining paper. The first two are survey papers: A. Persson, E. Elvers, A. Björnram and J. Erikson's paper, *A strategy to test questionnaires at a national statistical office*; and Wright, *An empirical examination of the relationship between nonresponse rate and nonresponse bias.* The last paper concludes our series on the Indigenous at least for now with the T. Kukutai and M. Walter article, *Recognition and Indigenizing Official Statistics: Reflections from Aotearoa New Zealand and Australia.*

References

[1] G. Barcaroli et al., Dealing with Big Data for Official Statistics: IT Issues. Meeting on the Management of Statistical Information Systems (MSIS 2014).

[2] Eurostat (2013), Scheveningen Memorandum on Big Data and Official Statistics. http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf.

[3] C. Reimsbach-Kounatze, (2015), The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis", OECD Digital Economy Papers, No. 245, OECD Publishing. http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

[4] S.-M. Tam and F. Clarke, Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics. IAOS 2014 Conference, Da Nang, Viet Nam, 2014.

[5] UNECE (2013), What does "big data" mean for official statistics? Report of the High-Level Group for the Modernisation of Statistical Production and Services (HLG). http://www1.unece.org/stat/platform/display/hlgbas.

[6] UNECE (2014a), Final project proposal: The Role of Big Data in the Modernisation of Statistical Production. http://www1.unece.org/stat/platform/display/bigdata/2014+Project.

[7] UNECE (2014b), How big is Big Data? Exploring the role of Big Data in Official Statistics. Report on the Virtual Sprint workshop. http://www1.unece.org/stat/platform/display/ bigdata/How+big+is+Big+Data.

[8] UNECE (2014c), Questionnaire about the skills necessary for people working with Big Data in the Statistical Organisations. Report from Oct 2014. http://www1.unece.org/stat/platform.

[9] UNECE (2014d), Report of the Global Working Group on Big Data for Official Statistics. Note by the Secretary-General to the 46th UNSC. http://unstats.un.org/unsd/statcom/doc15.