Big Data and official statistics:

local experiences and international initiatives

Big Data e statistiche ufficiali:

esperienze locali e iniziative internazionali

Barteld Braaksma, Piet Daas, May Offermans, Marco Puts and Martijn Tennekes[1]

Abstract Big Data is an interesting new source for official statistics, which promises abundant opportunities but also implies non-trivial challenges. The paper illustrates these with concrete examples. A particular characteristic of many Big Data sources is that their nature makes them very well suited for cross-border and multidisciplinary approaches.

Abstract *I Big Data costituiscono un'interessante nuova fonte per le statistiche ufficiali che da un lato offre grandi opportunità ma dall'altro comporta sfide non banali. Il lavoro illustra questi aspetti attraverso esempi concreti mettendo in evidenza la caratteristica peculiare di molti Big Data, insita nella loro stessa natura, di ben adattarsi ad approcci sia al confine tra discipline sia multidisciplinari.*

Key words: Big Data, official statistics, examples, international collaboration.

1 Introduction

According to some futurologists, Big Data will reshape the whole of society in unprecedented and unpredictable ways [10]. The 'Data Deluge' will change business models for both government bodies and private enterprises. It will also change their mutual relations and comparative advantages. Even if this view is too extreme, it is

[1] All authors, Statistics Netherlands. Contact: Barteld Braaksma, bbka@cbs.nl

unlikely that Big Data will not have any impact at all. This applies equally to the world of official statistics. The question to be addressed in this paper is therefore: *How will the arrival of Big Data affect official statistics?*

In order to discuss this question, the paper identifies a number of challenges that are particularly relevant for official statistics. These challenges are illustrated by local Dutch experiences based on concrete Big Data experiments and then reviews two international initiatives. The paper ends with some concluding remarks.

2 Big Data challenges for official statistics

In order to appreciate the challenges in Big Data for official statistics, it is important to have a clear picture of its role and in particular its data collection strategies.

2.1 What is official statistics?

Official statistics is a public good that informs many important decisions of government bodies, politicians, enterprises and individuals. It is a critical resource for a modern democratic society. In Europe, it is to a large extent produced by the European Statistical System (ESS) in which national statistical institutes and Eurostat cooperate. The ESS mission statement reads:

“We provide the European Union, the world and the public with independent high quality information on the economy and society on European, national and regional levels and make the information available to everyone for decision-making purposes, research and debate.”

Most of official statistics is produced according to a pre-set working programme. Its concepts are heavily standardised and output requirements are internationally harmonised, often governed by binding regulations. The European Statistics Code of Practice [7] provides principles and indicators to safeguard correct functioning of the ESS, while observing high quality standards. Among them is the principle that official statistics should be based on sound scientific principles. A cornerstone of official statistics is the absolute respect for confidentiality: data collected for statistical purposes (especially on identifiable individual persons or enterprises) may never be disclosed and may never be used for other purposes.

2.2 Source data for official statistics

Official statistics must be based on observations: often raw data that needs further processing and is honed to produce accurate, reliable, robust and timely information. Although model assumptions are used from time to time, in most cases it is desired to stay as close as possible to the actually observed data.

For many years, producers of official statistics have relied on their own data collection, using paper questionnaires, face-to-face and telephone interviews, or (somewhat less traditional) web surveys. This classical approach originates from the era of data scarcity, when official statistics institutes were among the few organisations that could gather data and disseminate information. A main advantage of the survey-based approach is that it gives full control over questions asked and populations studied. A big disadvantage is that it is rather costly and burdensome, for the surveying organisation and the respondents, respectively.

More recently, statistical institutes have started to use administrative (mostly government) registers as additional sources. Using such secondary sources reduces control over the available data, and the administrative population often does not exactly match the statistical one. However, these data are much cheaper to obtain than conducting a survey as they are already present. In some countries, the access and use of secondary sources is regulated by law.

Big Data sources offer even less control. They typically consist of ‘organic’ data [8] collected by others, who have a non-statistical purpose for their data. For example, a statistical organization might want to use retail transaction data to provide prices for their Consumer Price Index statistics, while the data generator sees it as a way to track inventories and sales.

2.3 Challenges

Several challenges have been identified, that need to be addressed when starting to use Big Data for official statistics [3,4,11,16]. Below we enumerate the main ones.

Access. Statistical institutes typically do not own Big Data sources. A first challenge thus is to obtain access to relevant sources. This implies agreements with data owners and processors, who have their own concerns regarding costs, confidentiality and other issues. However, they might also benefit from cooperating with statistical agencies. Terms and conditions have to be negotiated that are acceptable to both official statisticians and data providers.

Privacy. Privacy protection of individuals is an imperative, but familiar approaches do not always work when dealing with Big Data. Moreover, when the legal situation is not clear statisticians may have to fall back on ethical principles. Of critical importance is the public perception of any use of Big Data: this has a direct impact on trust in official statistics. Concerns have been heightened by the revelations that intelligence agencies are among the most active Big Data users.

Methodology. Many Big Data sources are composed of event-driven observational data which are not designed for data analysis. They lack well-defined target populations, data structures and quality guarantees. This makes it hard to apply traditional statistical methods, based on sampling theory. For example, assessing selectivity issues may prove problematic. Since many Big Data sources are text-based, the need to extract information from text increases. This calls for text mining and machine learning techniques, yet unfamiliar to official statisticians.

Interpretation. Extracting statistical meaning from Big Data sources is not easy. A tweet, a phone call or a car passing a detection loop all relate to persons, but how to interpret these signals is far from obvious. For example, the interpretation of mobile phone data is hampered by several issues: people may carry multiple phones or none, children use phones registered to their parents, phones may be switched off, etcetera. For social media messages, similar issues may arise when trying to identify characteristics of their authors. Remedies like deriving the gender and age of Twitter users from their choice of words appear feasible; but a lot still needs to be done.

Technology. An obvious challenge is the processing, storage and transfer of large data sets. Technological advances may partly solve these issues. Having data processed at the source, preventing the transfer of large data sets and the duplication of storage, may also be considered. The technological challenges include security mechanisms, which makes for example cheap cloud-based solutions less attractive.

Continuity. Typically, official statistics take the form of time series. For many users the continuity of these series is of the utmost importance. Many Big Data sources, however, have only recently emerged, are ever evolving and may disappear as quickly as they rise (remember MySpace?). This poses a risk for continuity.

3 Big Data research at Statistics Netherlands

In this chapter, we discuss three examples of Big Data research conducted at Statistics Netherlands. Other Big Data related examples at Statistics Netherlands include internet robots, scanner data and satellite images. Further opportunities like analysing financial transactions information are currently being studied, the first challenge often being to get access to the data.

Most of the above examples are still in the research phase, apart from scanner data which is in production for ten years now. Internet robots for the housing market are on the verge of being implemented in production.

Note that administrative data is usually not considered as Big Data, but in fact the larger administrative sources like the population register, VAT data and wages and salaries records could be included due to their size.

Statistics Netherlands' Innovation Programme [2] supported the experimental work described below, for example by establishing a dedicated Big Data Lab with powerful PCs and other facilities.

3.1 Traffic detection loops

In the Netherlands, approximately 100 million traffic detection loop records are generated a day. More specifically, for more than 12,000 detection loops on Dutch roads, the number of passing cars in various length classes is available on a minute-by-minute basis. The data are collected and stored by the National Data Warehouse for Traffic Information (NDW, www.ndw.nu/en/); a government body which provides the data free of charge to Statistics Netherlands. Since the data cannot be related back to individual vehicles, privacy concerns do not apply such that this data set lends itself very well to many experiments.

An issue is that this source suffers from under-coverage and selectivity. The number of vehicles detected is not available for every minute and not all (important) Dutch roads have detection loops. Fortunately, the first can be corrected by imputing the absent data with data that is reported by the same loop during a 5-minute interval before or after that minute. Coverage is improving over time. Gradually more and more roads have detection loops, enabling a more complete coverage of the most important Dutch roads. In one year more than 2000 loops were added.

A considerable part of the loops discern length classes, enabling the differentiation between cars and trucks. This is illustrated in Figure 1. In this figure, for the whole of the Netherlands, aggregated profiles are shown for three length classes. The results after correction for missing data were used. Because the small vehicles comprised around 75% of all vehicles detected, compared to 12% for the medium-sized and 13% for the large vehicles, normalized results are shown.

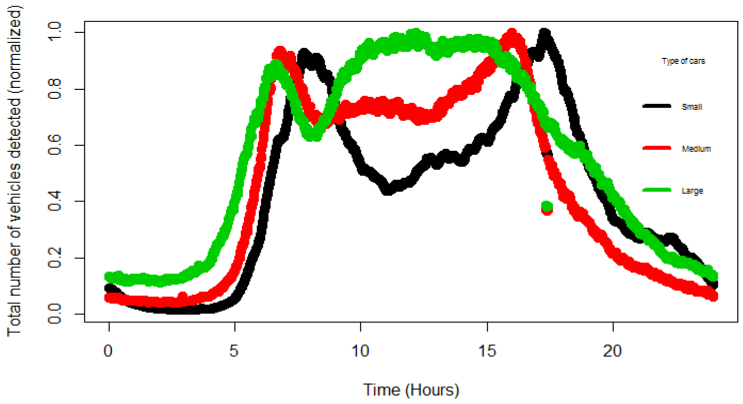

Figure 1. Normalized number of vehicles detected in three length classes on December 1st, 2011 after correcting for missing data. Small, medium-sized and large vehicles are shown in black, red and green, respectively. Profiles are normalized

The profiles clearly reveal differences in driving behaviour. Small cars have morning and evening rush-hour peaks at 8 am and 5 pm, respectively. For medium-sized vehicles, both peaks appear an hour earlier. The large vehicle category has a clear morning rush hour peak around 7 am, but displays a more distributed driving behaviour during the remainder of the day. After 3 pm the number of large vehicles gradually declines. Remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am [3,12].

At the level of individual loops, the number of detected vehicles demonstrates highly volatile behaviour, see Figure 2 below. This calls for a more refined approach which was done by applying signal analysis techniques. Harvesting the vast amount of data is a major challenge for statistics; but it could result in speedier and more robust traffic statistics, including more detailed information on selected groups. This is also likely indicative of changes in economic activity in a broader sense.

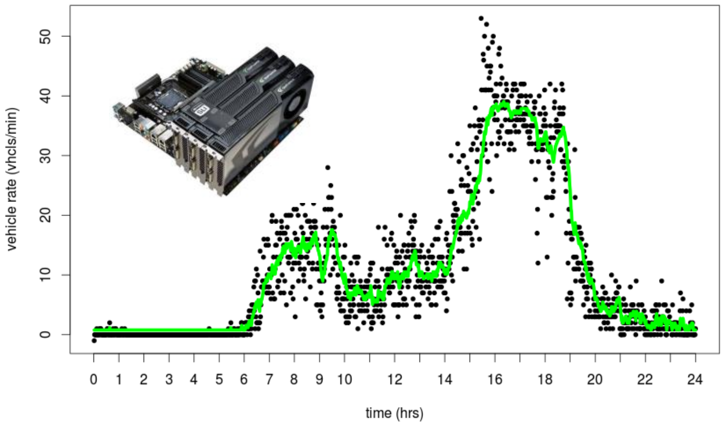

Figure 2. Result (green line) of application of a Bayesian recursive filter to raw data (black dots) from a single detection loop, assuming that they obey a Poisson distribution. Computations were done on a high-end PC with a GeForce GTX680 graphics card (inset), whose parallel processing capabilities speeded up processing time tremendously.

3.2 Mobile phone metadata

Nowadays, people carry mobile phones with them everywhere and use their phones throughout the day. To manage the phone traffic, a lot of data needs to be processed by mobile phone companies. This data is very closely associated with behaviour of people; behaviour that is of interest for official statistics. For example, the traffic is relayed through geographically distributed phone masts, which enables determination of the location of phone users. The relaying mast, however, may change several times during a call: nontrivial location algorithms are needed.

Through a three-party contract, Statistics Netherlands got access to call detail records (CDR) data from Vodafone, which has a marker share of approximately one third of the Dutch mobile phone market. The CDR data amount to 115 million records a day and contain information on both Dutch and roaming users of the Vodafone network. The anonymized CDR microdata were processed by a specialized intermediate company, Mezuro, according to queries specified by Statistics Netherlands. Only aggregate results were forwarded to Statistics

Netherlands which moreover were required to satisfy agreed privacy protection rules. Several uses for official statistics were studied, including inbound tourism [9] and daytime population [13], see Figure 3. The ‘daytime whereabouts’ is a topic about which so far very little is known due to lack of sources; in contrast to the ‘night-time population’ based on official (residence) registers.

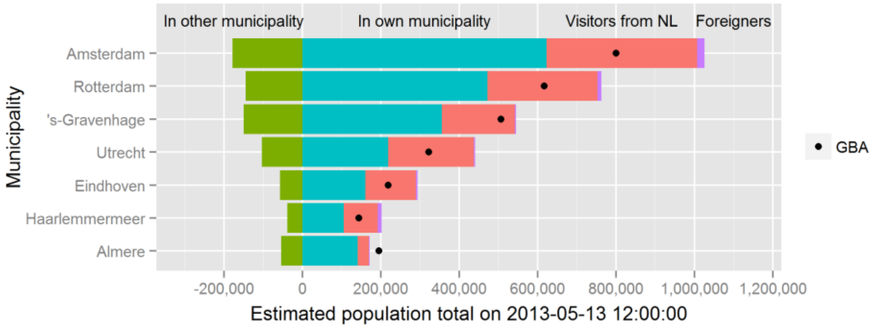

Figure 3. Daytime population at noon on a May Monday for the five largest Dutch municipalities, a typical working municipality (Haarlemmermeer, where Schiphol Airport is located) and a typical commuter municipality (Almere). Colours indicate the number of people that leave, stay or enter. Dots indicate the officially registered population.

3.3 Social media messages

More than three million public social media messages are produced on a daily basis in the Netherlands. These messages are available to anyone with internet access. Social media is a data source where people voluntarily share information, discuss topics of interest, and contact family and friends. To find out whether social media is an interesting data source for statistics, Dutch social media messages were studied from two perspectives: content and sentiment. The social media source data were provided by the company Coosto (www.coosto.com/uk/), which routinely collects all Dutch social media messages and assigns sentiment scores, among other things.

Studies of the content of Dutch Twitter messages (the predominant public social media message in the Netherlands at the time) revealed that nearly 50% of those messages were composed of 'pointless babble'. The remainder predominantly discussed spare time activities (10%), work (7%), media (5%) and politics (3%). Use of these, more serious, messages was hampered by the less serious 'babble' messages. The latter also negatively affected text mining studies.

The sentiment in Dutch social media messages was found to be highly correlated with Dutch consumer confidence, see Figure 4. Facebook gave the best overall results. The observed sentiment was stable on a monthly and weekly basis, but daily figures displayed highly volatile behaviour [5]. Thus it might become possible to produce useful weekly sentiment indicators, even on the first working day after the week studied.

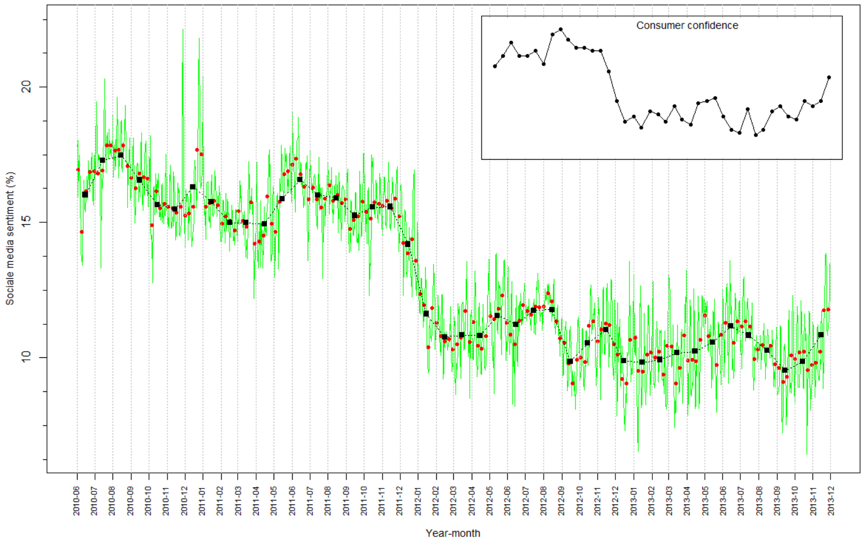

Figure 4. Development of daily, weekly and monthly aggregates of social media sentiment from June 2010 until November 2013, in green, red and black, respectively. The insert shows the development of consumer confidence for the same period.

4 International collaboration

One distinguishing aspect of many Big Data sources is that they are not confined to national borders. They are driven by the same technology everywhere, owned by multinational corporations or describe global phenomena. This presents unique opportunities for collaboration across borders and across domains.

Below we present two examples that show how the international statistical community has responded to these opportunities. It should be noted that these examples are not exhaustive: more initiatives have already started, like a Big Data Collaboratory (see sloddo.wordpress.com) organised by Goldsmiths College in London, in which representatives from several domains (including official statistics) work together with university researchers to identify common ground and areas of mutual interest.

4.1 The Scheveningen Memorandum

On 27 September 2013, the 99th DGINS[1] conference adopted the Scheveningen Memorandum on Big Data and Official Statistics [6]. This happened after a full day of presentations and discussions, in which all heads of European national statistical institutes and of a number of international statistical organisations participated.

The adoption of the Scheveningen Memorandum was a clear signal that the importance of Big Data has been recognized by the Official Statistics community. The Memorandum encourages the European Statistical System and its partners to effectively examine the potential of Big Data sources. Moreover, it was agreed to follow up the implementation of the Scheveningen Memorandum by adopting an action plan and roadmap which is currently being elaborated. An important element to gather views on possible implementation elements was the international Big Data Event in Rome on 31 March and 1 April 2014, sponsored by Eurostat.

4.2 The HLG project on Big Data

In 2010, the United Nations Economic Commission for Europe (UN/ECE) set up the High-Level Group for the Modernisation of Statistical Production and Services (HLG) to oversee and coordinate international work relating to statistical modernisation. In October 2013, the HLG decided to sponsor a project on Big Data [15]. This project is a collaborative approach to the new challenges of using Big Data sources for official statistics. The HLG project started in March 2014 with a ‘virtual sprint’ which resulted in a paper [14] that equally enumerates the opportunities, challenges, issues and questions that the use of Big Data by statistical organizations raises. The paper proposes that statistical agencies regard Big Data as:

Data that is difficult to collect, store or process within the conventional systems of statistical organizations. Either, their volume, velocity, structure or variety requires the adoption of new statistical software processing techniques and/or IT infrastructure to enable cost-effective insights to be made.

The paper also proposes to break down the many different types of Big Data sources into three main categories, each of which brings a different set of considerations for a statistical organization: human-sourced information, process-mediated data and machine-generated data; the examples presented above each represent one of these categories.

The work of the HLG project continues during 2014 and includes, among other things, the creation of a sandbox environment for practical Big Data experiments.

[1] Annual Conference of Directors General of National Statistical Institutes

5 Concluding remarks

The arrival of Big Data presents new opportunities for official statistics, but also poses new challenges. In fact, the notion of Big Data is a somewhat arbitrary moniker for a rather heterogeneous set of new data and information sources. They turn up in many different guises and have many different characteristics.

At this point in time, only limited experience is available to estimate the impact of Big Data on official statistics. To some extent a parallel to the introduction of survey sampling might apply [1]. When official statistics took shape in the middle of the 19th Century, at first only census-type enumerative approaches were considered valid. Around 1895, the first ideas were formulated for sample-based statistics, but it took several decades before the now dominant paradigm of survey sampling was accepted and firmly established.

All in all, the big challenges identified above will need to be addressed. In particular there is a need for new legislation, statisticians with new skill sets (‘data scientists’), new methodologies and appropriate computational facilities. The nature of Big Data points towards intensified international cooperation between data providers, scientists and official statistics.

References

1. Bethlehem, J.G. The rise of survey sampling. Statistics Netherlands Discussion Paper 09015 (2009). Available here
2. Braaksma, B. and Verbruggen, M. Innovation at Statistics Netherlands. Paper presented at the Conference on New Techniques and Technologies for Statistics, Brussels (2013). Available here
3. Daas, P.J.H. and Van der Loo, M.P.J. Big Data (and official statistics). Paper for the 2013 Meeting on the Management of Statistical Information Systems, Paris–Bangkok (2013). Available here
4. Daas, P.J.H. and Puts, M.J.H. Big Data as a source of statistical information. The Survey Statistician **69**, 22–31 (2014)
5. Daas, P.J.H. and Puts, M.J.H. Social media sentiment and consumer confidence. Paper for the Workshop on using Big Data for Forecasting and Statistics, Frankfurt (2014). Available here
6. DGINS. Scheveningen Memorandum on Big Data and official statistics (2013). Available here
7. Eurostat. European Statistics Code of Practice (2014). Available here
8. Groves, R.M. Three eras of survey research, Public Opinion Quarterly **75**, 861–871 (2011)
9. Heerschap, N.M., Ortega Azurduy, S.A., Priem, A.H. and Offermans, M.P.W. Innovation of tourism statistics through the use of new Big Data sources. Paper prepared for the Global Forum on Tourism Statistics, Prague (2014). Available here
10. Mayer-Schönberger, V. and Cukier, K. Big Data: A revolution that will transform how we live, work, and think. London, John Murray Publishers (2013)
11. Scannapieco, M., Virgillito, A. and Zardetto, D. Placing Big Data in official statistics: a big challenge? Paper presented at the Conference on New Techniques and Technologies for Statistics, Brussels (2013). Available here
12. Struijs, P. and Daas, P.J.H. Big Data, big impact? Paper presented at the Seminar on Statistical Data Collection, Geneva (2013). Available here
13. Tennekes, M. and Offermans, M.P.W. Daytime population estimations based on mobile phone metadata. Paper prepared for the Joint Statistical Meetings, Boston (2014). Forthcoming
14. UNECE. How big is Big Data? Exploring the role of Big Data in official statistics. Draft for public review (2014). Available here
15. UNECE. The role of Big Data in the modernisation of statistical production. Project proposal (2014). Available here
16. UNECE. What does ‘Big Data’ mean for official statistics? Paper prepared on behalf of the High-Level Group for the Modernisation of Statistical Production and Services (2013). Available here