Distr.
GENERAL

Working Paper
11 April 2013

ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE (ECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (iv): Collaboration

Big Data (and official statistics)*

Prepared by Piet Daas and Mark van der Loo, Statistics Netherlands, the Netherlands

I. Introduction

1. In our modern world more and more data are generated on the web and produced by sensors in the ever growing number of electronic devices surrounding us. The amount of data and the frequency at which they are produced have led to the concept of 'Big Data'. This concept can be defined as:

> *Big data are data sources that can be –generally– described as: "high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making."*

2. Big data is characterized as data sets of increasing volume, velocity and variety; the 3 V's. Big data is often largely unstructured, meaning that it has no pre-defined data model and/or does not fit well into conventional relational databases. Apart from generating new commercial opportunities in the private sector, Big data is also potentially very interesting as an input for official statistics; either for use on its own, or in combination with more traditional data sources such as sample surveys and administrative registers. However, harvesting the information from Big data and incorporating it into a statistical production process is not easy.

II. Experiences at Statistics Netherlands

3. At Statistics Netherlands several Big Data case studies were performed. Data sources studied as potential input for statistics were: a) traffic loop detection data, b) mobile phone data, and c) social media messages. The findings are briefly described below.

* A considerable part of this paper is based on the working paper 'Big Data and its Potential Use by the Statistical Community' written for the High-level group for the modernisation of statistical production and services meeting by Michael Glasson (Australia), Julie Trepanier (Canada), Vincenzo Patruno (Italy), Piet Daas (Netherlands), Michail Skaliotis (Eurostat) and Anjum Khan (UNECE).

A. Traffic loop detection data

4. In the Netherlands, approximately 80 million traffic loop detection records are generated a day (Daas *et al.*, 2013). This data can be used as a source of information for traffic and transport statistics and potentially also for statistics on other economic phenomena. The data is provided at a very detailed level. 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.

5. The downside of this source is that it seriously 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 yet. Fortunately, the first can be correct by imputing the absent data with data that is reported by the same location during a 5-min interval before or after that minute (Daas *et al.*, 2013). 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 a year more then 2000 loops were added.

6. A considerable part of the loops are able to discern vehicles in various length classes, enabling the differentiation between cars and trucks. This is illustrated in Figure 1. In this figure, for the whole of the Netherlands, normalized profiles are shown for 3 classes of vehicles. The vehicles were differentiated in three length categories: small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter), and large (> 12.2 meter). The results after correction for missing data were used. Because the small vehicle category comprised around 75% of all vehicles detected, compared to 12% for the medium-sized and 13% for the large vehicles, the normalized results for each category are shown.

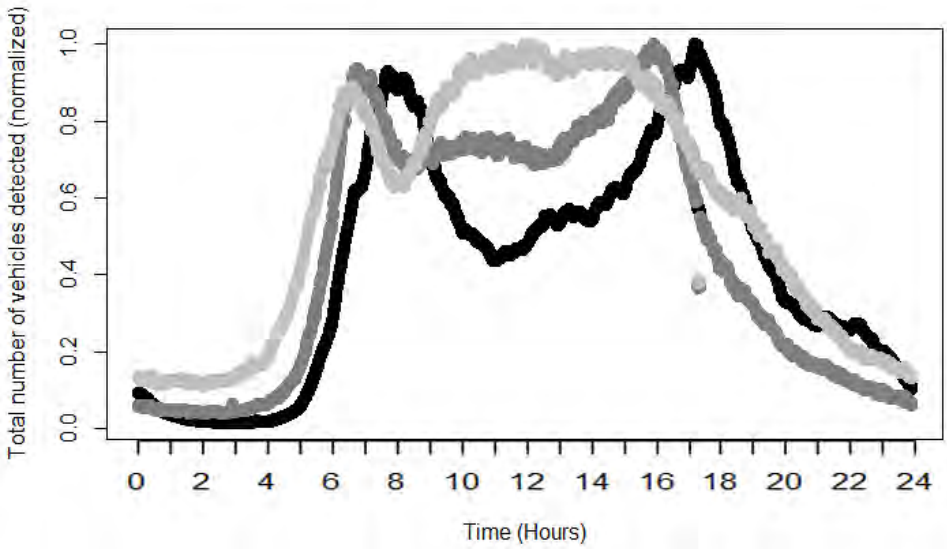

Figure 1. Normalized number of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown in black, dark grey and grey, respectively. Profiles are normalized to more clearly reveal the differences in driving behaviour.

7. The profiles clearly reveal differences in the driving behaviour of the vehicle classes. The small vehicles have clear morning and evening rush-hour peaks at 8 am and 5 pm respectively. The medium-sized vehicles have both an earlier morning and evening rush hour peak, at 7 am and 4 pm respectively. The large vehicle category has a clear morning rush hour peak around 7 am and displays a more distributed driving behaviour during the remainder of the day. After 3 pm the number of large vehicles gradually declines. Most remarkable is the decrease in the relative number of medium-sized and large vehicles detected at 8 am, during the morning rush hour peak of the small vehicles. This may be caused by a deliberate action of the drivers of the medium-sized and large vehicles of wanting to avoid the morning rush hour peak of the small vehicles.

8. At the most detailed level, that of individual loops, the number of vehicles detected demonstrates (highly) volatile behaviour, indicating the need for a more statistical approach (Daas *et al.*, 2013). Harvesting the vast amount of information from the data is a major challenge for statistics. Making full use of this information would result in speedier and more robust statistics on traffic and more detailed information of the traffic of large vehicles which is very likely indicative of changes in economic development.

B. Mobile phone location data

9. The use of mobile phones nowadays is ubiquitous. People often carry phones with them and use their phones throughout the day. Instrumental for the infrastructure enabling the coverage for mobile phones, are mobile phone masts/towers, called ‘sites’ in the industry. Those sites are located at strategic points, covering as wide an area as possible.

10. Much of the activity that is associated with handling the phone traffic, i.e. handling the localisation of mobile phones, optimizing the capacity of a site is stored by the mobile phone company. So mobile phone companies record data that are very closely associated with behaviour of people; behaviour that is of interest to statistical agencies. Obvious examples are behaviour regarding tourism, mobility, commuting and transport. The destinations and residences of people during day-time are topics of various surveys. Using data from mobile phone companies we should be able to provide additional and more detailed insight on the whereabouts and the activity of mobile phone users.

11. For our research we obtained a dataset from a mobile telecommunication provider containing records of all call-events (speech-calls and text messages) on their network in the Netherlands for a time period of two weeks. Each record contains information about the time and serving antenna of a call-event and an (scrambled version of the) identification number of the phone. This study revealed several uses for official statistics, such as economic activity, tourism, population density to mobility and road use (De Jonge *et al.*, 2012).

C. Social media messages

12. Around 1 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 has the potential of being a data source as people voluntarily share information, discuss topics of interest, and contact family and friends. To respond to whether social media is an interesting data source for statistics, Dutch social media messages were studied by Statistics Netherland from two perspectives: content and sentiment.

13. Studies of the content of Dutch Twitter messages (the predominant public social media message in the Netherlands at the time of the study) revealed that nearly 50% of messages were composed of 'pointless babble'. The remainder predominantly discussed spare time activities (10%), work (7%), media (TV & radio; 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 approaches.

14. Determination of the sentiment in social media messages revealed a very interesting potential use of this data source for statistics. The sentiment in Dutch social media messages was found to be highly correlated with Dutch consumer confidence; in particular with the sentiment towards the economic situation (Figure 2). The latter relation was stable on a monthly and on a weekly basis. Daily figures, however, displayed highly volatile behaviour (Daas *et al.*, 2013). This highlights that it is possible to produce weekly indicators for consumer confidence and could be produced on the first working day following the week studied, demonstrating the ability to deliver quick results.

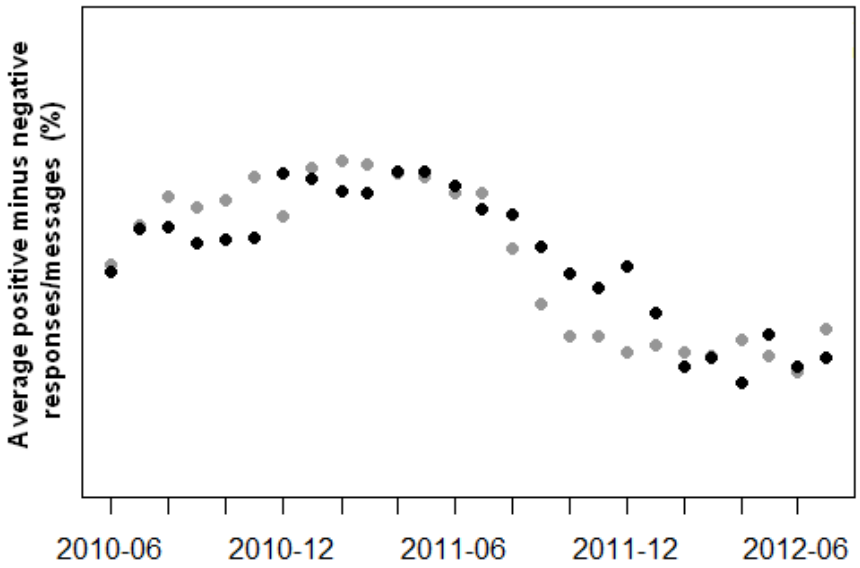

Figure 2. Dutch consumer confidence (grey) and the overall sentiment in Dutch social media messages on a monthly basis (black). The social media sentiment in December months is considerably more positive compared to the sentiment in the months before and after. This is caused by the positive 'merry Christmas' and 'happy new year' messages send during this month.

III. Future plans

A. Challenges identified

15. Our studies and the High level working group paper (2013) revealed several challenges/issues that need to be addressed. These fall into the following categories:

(a) *Legislative*, i.e. with respect to the access and use of data. The right to access admin data, established in principle by the law, is not adequately supported by specific obligations for Big data. Many potential Big Data sources are collected by non-governmental organisations or are 'freely' available on the web; situations that may not be covered by existing legislation.

(b) *Privacy*, i.e. managing public trust and acceptance of data re-use and its link to other sources. Privacy is generally defined as the right of individuals to control or influence what information related to them may be disclosed. The problem with Big data is that the users of services and devices generating the data are most likely unaware that they are doing so, and/or what it can be used for. The data would become even bigger if they are pooled, as would the privacy concerns.

(c) *Financial*, i.e. potential costs of sourcing data vs. benefits. There is likely to be a cost to acquire Big data, especially Big data held by the private sector and especially if legislation is silent on the financial modalities surrounding acquisition of external data.

(d) *Management*, e.g. policies and directives about the management and protection of the data. Big data for official statistics means more information coming to NSI's that is subject to policies and directives on the management and protection of the information to which NSI's must adhere. Long-term stability may be a problem when using Big Data. Typically, statistics for policy making and evaluation are required for extended periods of time, often covering many years. Many big data sources have only recently been 'established'.

(e) *Methodological*, i.e. data quality and suitability of statistical methods. When more and more data are being analysed traditional statistical methods, developed for the very thorough analysis of small samples, run into trouble; in the most simple case there just not fast enough. Since text is an essential part of many Big Data sources, the need to extract information from text increases. Also, the subpopulations covered by Big Data sources studied are not the target populations for official statistics. Therefore such data are likely to be selective, not

representative of a relevant target population. Assessing representativity of Big Data may prove problematic, as often there are no characteristics readily available to conduct such comparison. Next, including the information content of Big Data sources in the statistical production process (often without unique statistical ID keys) makes integration challenging.

(f) *Technological*, i.e. issues related to information technology. Dedicated and specialized computing infrastructures are required to cope with Big Data to enable processing and speed up analysis of large amounts of data. Certainly for the exploratory phase, during which the content and structure of Big Data sets has to be understood, fast technology certainly speeds up this process and more quickly enable the revelation of their use for statistics.

16. Overall it can be stated that the work described above revealed that there is a need for new legislation (enabling access to Big Data), persons with new skills (statistical aware 'data scientists'; Loukides, 2011), new methods (specifically tailored to large data files and fast) and computational facilities that enable the speedy analysis of large data files while ensuring privacy (Daas, 2012).

B. Vision

17. The Official Statistics community is only scratching the surface when it comes to exploring the opportunities offered by Big Data. Moreover, at this moment, research activities related to Big Data are limited to isolated initiatives at some NSI's. In our opinion, the methodological and technological challenges mentioned above should be met in a Big Data research programme. Such a research programme should provide guidance and financial instruments for the following research.

18. Such a research programme should include the following topics:

(a) Experimentation with Big Data sources by setting up a number of pilot projects in selected statistical areas. These pilots will provide guidelines for the effective use of Big data for purposes of official statistics. Important research areas include:
- combining Big Data with traditional data sources (survey, administrative);
- replacing traditional data sources, i.e. decreasing administrative burden;
- opportunities for new output;
- opportunities for faster or real-time statistics production.

(b) Development of new exploration and analysis methods, specific for the study of huge volumes of data, in the context of official statistics.

(c) Further experimentation with High Performance Computing technologies which are essential for the processing of huge volumes of data.

(d) Collaboration with third parties such as universities or IT/consulting companies with experience in the statistical analysis of large data sources.

19. Big Data is a highly multidisciplinary field requiring subject matter knowledge, strong math skills as well as strong programming skills. To ensure a speedy progress, research subprojects should be performed by small, highly skilled and dedicated teams covering such expertise. Moreover, because of the multidisciplinary character of the research programme, guidance could be provided by a steering committee composed of experts in various statistical fields.

20. The strategic contributions of the above research programme consist of the knowledge and experience gained in applications of Big Data for Official Statistics as well as breeding a number of 'data scientists' with a strong knowledge of Official Statistics. Such data scientists will be an indispensable part of NSI's human capital in the near future.

References

Daas, P. (2012) *Big Data and official statistics*. Sharing Advisory Board, Software Sharing Newsletter 7, 2-3. Located at: http://www1.unece.org/stat/platform/download/attachments/ 22478904/issue+7.pdf

Daas, P.J.H., Puts, M.J., Buelens, B., van den Hurk, P.A.M. (2013) *Big Data and Official Statistics*. Paper for the 2013 NTTS conference, Brussels, Belgium. Located at: http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf

De Jonge, E., van Pelt, M., Roos, M. (2012) *Time patterns, geospatial clustering and mobility statistics based on mobile phone network data.* Discussion paper 201214, Statistics Netherlands. Located at: http://www.cbs.nl/NR/rdonlyres/010F11EC-AF2F-4138-8201-2583D461D2B6/0/201214x10pub.pdf

High level working group (2013) *Big Data and its Potential Use by the Statistical Community*. Working paper for the High-level group for the modernisation of statistical production and services meeting, d.d. March 10. Located at: http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184

Loukides, M. (2011) *What is Data Science?* O'Reilly Radar report, O'Reilly Media Inc. Located at: http://cdn.oreilly.com/radar/2010/06/What_is_Data_Science.pdf