Distr. GENERAL
03 October 2013
WP 31
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on Statistical Data Collection
(Geneva, Switzerland, 25-27 September 2013)

Topic (v): Integration and management of new data sources

BIG DATA, BIG IMPACT?

Working Paper

Prepared by Peter Struijs and Piet Daas, Statistics Netherlands[1]

I. Introduction

1. Big Data sources create a number of opportunities for National Statistical Institutes (NSIs). They may be used to substitute or supplement more traditional data sources, such as questionnaires and administrative data sources, for already existing statistics. There may also be opportunities in respect of new statistical output. Their use, however, poses a number of challenges. The paper describes the opportunities and challenges posed by Big Data and discusses the strategic and policy questions that arise from them.

2. Big Data seems to be a hype. According to Google Trends, in August last year it overtook Open Data as a search term. Does that mean that the attention to Big Data will diminish after a peak? Maybe the term Big Data will fade after some time, but as an important phenomenon it will most probably last, because it is based on various trends in society. First of all, there is a seemingly exponential growth of data. Data are registered by networks of sensors, camera's, public administrations, banks, enterprises, mobile networks, satellites, drones, social networks, internet sites, etc. (Mayer-Schönberger and Cukier, 2013).

3. Another trend is the increasing accessibility of data, often for free. Information on the internet and on social media is widely accessible, data are increasingly available in the form of open data and open source, and there are private enterprises that have a business model that rests on making data public, for instance Google. The shape of the information is also changing. There is more interaction and visualisation, messages tend to get shorter, and the share of traditional tables in the provision of information seems to be shrinking. Of course, ever faster technological developments are behind many such trends. Applications become smarter, for instance in recognizing faces or using speech interfaces. Cloud computing leads to data accessibility from any place at any time.

4. As a side effect, public opinion on privacy and confidentiality seems to be in flux. On the one hand, privacy seems to be ever more under pressure when public safety or commercial interests are perceived to be at stake, and young people who have grown up using social networks tend to consider

[1] The views expressed in this paper are those of the authors and do not necessarily reflect the position of Statistics Netherlands.

privacy less important than the elderly. On the other hand, there seems to be a growing general awareness of possible privacy implications of the ubiquity of data, resulting in a more critical attitude towards the unquestioned processing of data by anyone. Anyway, the understanding for the need of data collection by organizations is decreasing, especially if such data are already registered elsewhere.

5. Some of these trends are particularly important for NSIs. They are faced with more and more potential data sources, whereas the modalities for their use are changing. New opportunities and challenges abound. At the same time, the position of NSIs in the information society is becoming less evident, even though their institutional setting is stable for the time being. Other providers of information on the society pop up everywhere. They are often very quick and perceived as knowledgeable. Because of this, society becomes less dependent on information from NSIs. There are alternatives, for instance, to official price indices. Apart from the many practical questions, Big Data is bound to have an impact on NSIs at the strategic level.

II. Big Data Characteristics

6. What is Big Data, how can it be defined? And why is there suddenly so much talk about Big Data? Moore's Law stems from 1965, and the volume of data has been increasing for many decades. What threshold has been passed? Apparently, none. The emergence of the concept of Big Data appears to result from qualitative changes induced by changes in data quantity and public availability. We seem to have reached a point where the traditional way of making statistics does not provide the answers to the new questions that arise – or not fast enough.

7. For statistical purposes, a recent UNECE document gives the following definition of Big Data: "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" (UNECE, 2013). That is, Big Data is Big Data *sources*. However, this definition is not precise enough to decide in concrete cases whether the data source belongs to Big Data or not.

8. Rather than trying – possibly in vain – to give a more precise definition, it may help to mention aspects of Big Data sources that are regarded as characteristic for such sources by many statisticians, and to supplement this by mentioning examples of data sources that many statisticians consider Big Data sources. In this way, a picture of Big Data can be obtained that is clear enough to allow making progress without being stuck in discussions on definition. These can be found in abundance on the internet.

9. Volume, velocity and variety, the three V's, are among the most important characteristics, although they may not apply all three at the same time. In fact, there exist pretty high-volume traditional data sources that are generally not considered to be Big Data. Other characteristics often mentioned are the novelty of the data source, the dynamics of its population, the need to use new methodological approaches, the essentially new character of the resulting information, the possible need to process the data at the source, the unstructured nature of the data, the reference of the data to events, the circumstance that the data are often a by-product of the principal activity of an organization, and their physical distribution over several databases or points of measurement. These characteristics do support the assumption that the emergence of the concept of Big Data has to do with the qualitative changes that come with quantitative ones.

10. In the context of MSIS (Management of Statistical Information Systems), in which the UNECE, Eurostat and the OECD cooperate, an inventory is being made of the use of Big Data by NSIs and international statistical organizations[2]. Examples of Big Data sources so far identified are traffic detection records, mobile phone location data, social media messages, internet traffic flow data, satellite images (remote sensing), commercial and financial transaction data, and internet sites (scanning by internet robots).

[2] The inventory has not been made public yet.

11. Before discussing the challenges posed by Big Data in chapter IV, three examples of experiments with Big Data sources done by Statistics Netherlands are presented in the next chapter. The examples concern the use of traffic loop detection records (i.e., information from a network of sensors under the roads) for traffic statistics, the use of mobile phone location data for statistics on the so-called daytime population and related phenomena, and the use of social media messages as an indicator of social sentiment. The contents of chapter III was published earlier in a paper by Piet Daas and Mark van der Loo (Daas and Van der Loo, 2013).

III. Three Examples

A. Traffic loop detection data

12. In the Netherlands, approximately 100 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.

13. 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 corrected by imputing the absent data with data that is reported by the same location during a 5-minutes 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 one year more than 2000 loops were added.

14. 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.

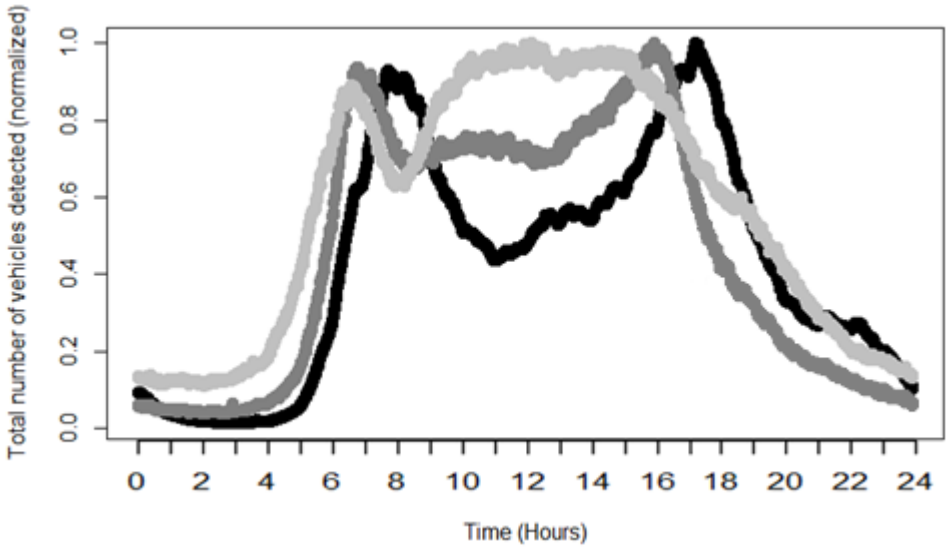

Figure 1. Normalized number of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown in black, dark grey and grey, respectively. Profiles are normalized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

15. 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, wanting to avoid the morning rush hour peak of the small vehicles.

16. 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 in general and will provide more detailed information of the traffic of large vehicles. This is very likely indicative of changes in economic development.

B. Mobile phone location data

17. 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.

18. Much of the activity that is associated with handling the phone traffic, that is, handling the localisation of mobile phones and 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 NSIs. Obvious examples are behaviour regarding tourism, mobility, commuting and transport. The destinations and residences of people during daytime are also topics of various surveys. Data from mobile phone companies could provide additional and more detailed insight on the whereabouts and the activity of its users, which may be indicative for the behaviour of people in general.

19. 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 was studied. There are about 35 million records a day. Each record contains information about the time and serving antenna of a call-event and a (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

20. Around one 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.

21. 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 those 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 studies.

22. Determination of the sentiment in social media messages revealed a very interesting potential use of this data 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. It also revealed that such an indicator could be produced on the first working day following the week studied, demonstrating the ability to deliver quick results.

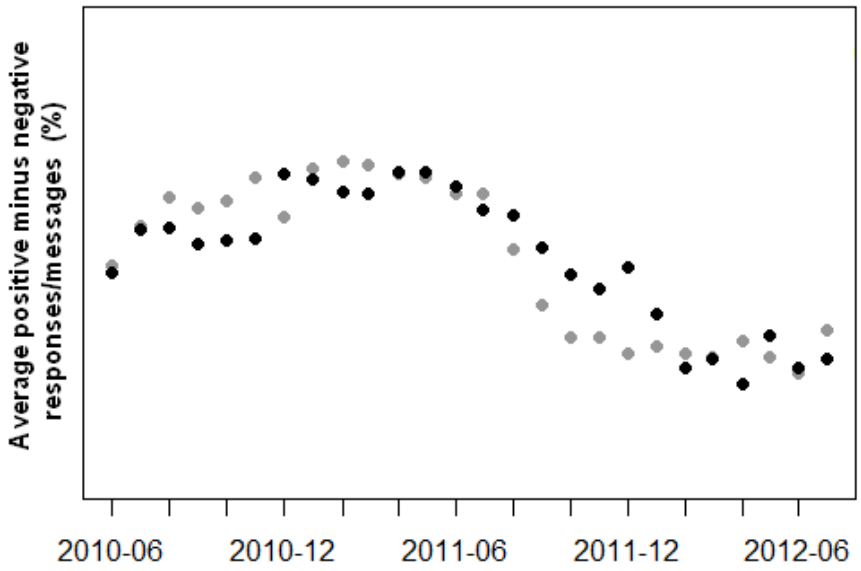

Figure 2. Dutch consumer confidence (grey) and the overall sentiment in Dutch social media messages on a monthly basis (black). The social media sentiment in December months is considerably more positive compared to the sentiment in the months before and after. This is caused by the positive 'merry Christmas'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

IV. Big Data Issues

A. Positioning of the NSI

23. There are many practical issues concerning Big Data, but there is also an important strategic one: What position does the NSI want to occupy in the future information society? There are two reasons to ask this question. The first is that we are witnessing a change from a situation in which NSIs have to deal with scarcity of data to one of data abundance. NSIs are used to the need to conduct surveys and obtain their data from respondents. The growing availability of public registrations, however, means that data can be obtained without imposing a burden on individuals and businesses. Since any action by persons or businesses – transactions, movements, communication, social and business activities – nowadays leaves digital traces in one way or another, ever increasing amounts of data are becoming available. And, contrary to survey data, these data are not available exclusively for NSIs. As such, the NSI is becoming less unique as a user of data and a provider of information and has to reassess its position.

24. The second reason to look at the position of NSIs has to do with the fact that many Big Data sources do not have a deliberate design. Traditional administrative registers have a well-defined target population, variables, structure and (administrative) quality. They also have an explicit legal basis. But what design is behind Twitter messages, commercial websites or mobile phone traffic? Traditionally, the value added of NSIs is largely based on their unique ability to relate data from different sources, the data often referring to known target populations. However, for Big Data sources, populations can often not be specified, let alone related to other sources. How can NSIs ensure that they still have a unique added value in the future?

25. An obvious question is: To what extent can Big Data sources be used for the production and improvement of current statistics? But the more important positioning questions pertain to the output

side of NSIs. What new output can an NSI produce by using new data sources, and what value has to be attached to output coherence in the future? NSIs may even ask more fundamental questions. Does the NSI want to continue making statistics for which there is a market alternative? Would it be better to shift the role of the NSI from producing statistical information towards validating information produced by others? Can it assume new roles, based on its institutional position and the knowledge it has accumulated?

26. In order to answer these strategic questions, one has to have a better idea of the possibilities and limitations of deploying Big Data sources. So let's have a look at the other issues.

B. Statistical output

27. Big Data sources may be used to replace or supplement other sources for the production of statistics that are part of the existing statistics programme. Some of the examples mentioned earlier have such potential, for instance the use of traffic loop detection data for traffic statistics, or the use of social media messages for composing a consumer confidence index. Such applications may lead to a reduction of the administrative burden on persons and businesses, or reduce other costs. If possible obstacles regarding data access, methodology, privacy, etc., can be overcome, their business case is clear.

28. Many new statistics can be conceived, based on Big Data sources, and the possibilities are growing. For instance, statistics on daytime populations can, as we have seen, be derived from mobile phone location data. If new statistics do not lead to a reduction in costs or administrative burden, their business case must be based on the value of the data for the users of these statistics. For example, information on daytime populations is valuable for those responsible for public order and safety, or for emergency arrangements. It is conceivable that new users can be persuaded to financially support the production of such statistics, or related research.

29. The distinction between the use of Big Data for existing statistics versus new statistics is not clear-cut. Statistics based on Big Data may measure something that is slightly different from what has been measured traditionally. In some cases the frequency may be improved, as might be the case with the measurement of social media sentiment, or the level of detail improved, as with mobile phone location data. In fact, the NSI has to find a new and richer optimum for its output, given the increased availability of data sources. This implies that the NSI must also be prepared to drop elements from its existing programme or make small adjustments. At the European level, this may also be true. European regulations stipulating the production of statistics can be very detailed and based on the assumption of using specific data sources. In such cases some flexibility may be called for.

30. Although the definition of Big Data refers to data sources, it is important to be output oriented when it comes to policy decisions. Competitors are increasingly relevant to NSIs, and can also be used as an inspiration for doing things better, for instance in respect of visualisation, interaction with customers and speed. It is also important to be aware of the information needs of actual and potential users of statistics. There is a risk that only the needs of current customers are known to the NSI. Especially if positioning of the NSI is an issue, the external needs must be known.

C. Statistical methodology

31. What exactly is the meaning and relevance of the data found in Big Data sources, from a user's perspective? What does the number of searches on an internet search engine reveal, or the sentiment observed in social media, or the number of mobile phones connected to a site? The interpretation of Big Data can be a big methodological problem. Moreover, meaning and relevance are user and use dependant.

32. Most statistics aim at giving information about populations of persons or businesses, or other relevant sets, such as goods imported or sold. However, the population covered by Big Data may be unclear. Mobile phones may be carried by others than the owner, vehicles passing a detection loop may be private or company vehicles, and what do we know about the population using social media? And

how do these populations change over time? How stable are they? In some cases it may be possible to obtain background variables, such as for credit card data, while in other cases background variables may be estimated. For instance, the choice of wording is correlated with age and sex of the user of social media. Text mining may become more important in the age of Big Data.

33. Not knowing the composition of the populations included in Big Data leads to the question what to do when sampling theory cannot be applied. What to do if one does not know for what part of the population the dataset is representative? More fundamentally, one may wonder whether sampling theory deserves being the default approach to statistics in the age of Big Data. Maybe more model-based approaches need to be applied. Examples are probabilistic modelling, Bayesian methods, multilevel approaches, statistical-learning methods and occupancy models, such as those used in measuring wild animal populations. Econometric models can also be considered. Then the measured phenomena are leading, and research may be aimed at relating them to information already known. Big Data might truly be causing a paradigm shift.

34. Quality considerations are generally important when deciding on the methods used. But the definition of quality has its own problems, especially if modelling approaches become more prominent. As mentioned earlier, coherence of the information disseminated by an NSI used to be a strength. When deciding on methodological issues, this aspect of quality may deserve special attention, especially since competitors may not be in a position to provide information on the quality and the relationship of the data with other information that is publicly available.

D. Statistical process

35. When using Big Data, the design of the statistical process needs special attention. It may be difficult to receive and process really high volume datasets, especially if the second “V” of Big Data, velocity, applies. There may be technical solutions for this, albeit possibly at additional costs, but another solution is conceivable: perhaps the data can remain at the source. Maybe it is possible to arrange that queries are done by the source holder in the source, for instance that data is first aggregated or sampled prior to submission. But such solutions themselves entail other issues. Are the results reproducible, can they still be linked to data available at the NSI?

36. As is the case with more traditional statistical processing, the data may be processed physically at the NSI offices or elsewhere, in various arrangements. Although it entails its own issues, cloud computing may be considered if some conditions are met (see below). Another possibility is to collaborate with another party, for instance a research institute with facilities for Big Data processing. This may be beneficial to all parties involved, since knowledge and experience can be shared. Whatever arrangement is entered, it is important to be aware of possible risks, for instance in respect of the continuity of the partnership and public trust.

37. The continuity and possible volatility of the Big Data source also deserve consideration. Social media, for instance, seems to have an ever shorter lifecycle. As a consequence, the use of Big Data requires a more flexible set-up of production processes, with a short time-to-market. Not only data collection has to be flexible, but also data processing further in the production chain. More generally, NSIs that start using Big Data may have to adapt or even reconsider their enterprise architecture.

38. Cost considerations are of course important when designing the statistical process. In some cases Big Data have to be paid for, although NSIs are not used to doing this. But using Big Data may also save costs, for instance if a labour intensive survey can be replaced by completely automated data collection. The benefits for data providers in the form of response burden reduction and for data users in the form of better statistical services may even be more important considerations. The availability of Big Data sources requires a new optimisation effort, aimed at getting the best set of statistical services, given the potential data sources, the demand for statistical information and budgetary constraints and possibilities.

E. Privacy and security

39. Looking at privacy and security, there are several issues, real or perceived, that may impede using Big Data. Data ownership and copyright may be an issue, and the purpose for which data are registered. Even if data are publicly accessible, for instance on websites or as social media messages that do not have access restrictions, questions of ownership and purpose of publication can be raised. Internet robots cause a burden on the providers of the sites, and in some cases site owners prefer sending data directly to the NSI. And even if there are no legal impediments, the perception of the public may be a factor to take into account. These concerns have to be taken seriously.

40. Fortunately, there are measures NSIs can take to overcome at least some of the obstacles. In some cases the use of informed consent may be a solution. If the NSI can offer a reduction of the response burden, this can be very helpful, also in getting the support of the general public. For the long run changes in legislation may be considered, to ensure continuous data access. But it remains important to stay in line with public opinion, because credibility and public trust are important assets of NSIs.

41. As to cloud computing, a few sensible rules may be used. Cloud services include the provision of computing resources, platforms and IT applications via the internet. At the present legal and technical state of the art, it is advisable in general not to host sensitive and critical data and processes in the cloud. The user of the cloud service remains responsible for the security and privacy of the data, and public trust depends on this. It is also advisable to use data encryption where possible.

F. Other issues

42. For NSIs that want to make Big Data a serious part of their business, governance may become an issue. Because of the important strategic aspects of Big Data, this subject should get attention at the highest management level of the NSI. Setting priorities, creating favourable conditions for using Big Data and taking related budget decisions would be tasks for the strategic level, as would be the making of policy choices. The issues mentioned above require a number of policy decisions that influence various parts of the organization. The organization's CIO (Chief Information Officer) would likely have an important say in the way the NSI deals with Big Data.

43. An issue not discussed so far is human capital. In order to work with Big Data, specific technical skills are needed, such as advanced computing skills, a fair command of math and statistics, modelling skills and data engineering skills. But equally important are the mental orientation and behavioural skills of the staff. Working with Big Data requires an open mind-set and the ability not to see all problems *a priori* in terms of sampling theory. For this type of staff the term data scientist has been coined. However, it is not evident that the culture of NSIs can smoothly absorb this type of professionals. A way to deal with this cultural issue is to create one or more kernels of data scientists working with Big Data, and let these kernels grow, which will be a natural process if they are successful.

V. The Future

44. In this paper many questions have been raised. Importantly, the answers to these questions depend not only on what is possible, but also on what is strategically desirable. In what position does the NSI want to be in the information market in ten years' time? What types of output does it want to produce, what role does it want to play in a society in which information is a commodity supplied by a large number of businesses and institutions, often with high speed but of unknown quality? The answers to such questions will be crucial to the effect Big Data will have on official statistics.

45. It is obvious that Big Data will have a big impact on the statistical community and on official statistics. The specifics of this impact will only gradually become clear, but some features are already visible or foreseeable. NSIs will be subject to more competition from actors outside the traditional statistical institutions. They will adapt their way of making statistics and find a new balance, making use of the new possibilities Big Data offers. This may require a paradigm shift from a survey oriented to a more secondary data focussed orientation, in which model-based approaches are the norm. Data scientists will get an important role, and the organizational culture of NSIs will change accordingly.

Anyway, it is encouraging to see that Big Data is an area in which international cooperation gets increasing attention. Together, the statistical community can face the future with confidence – provided there is a willingness to adapt.

VI. References

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

Daas, P.J.H., and Van der Loo, M (2013). *Big Data (and official statistics)*. Paper for the 2013 Meeting on the Management of Statistical Information Systems (MSIS 2013).

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

Mayer-Schönberger, and Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think.* Eamon Dolan/Houghton Mifflin Harcourt, 2013.

UNECE (2013). *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. 10 March 2013.